COUNTERFACTUAL FAIRNESS AND UTILITY TRADE-OFFS: CHARACTERIZING THE PARETO FRONTIER

Anonymous authors
Paper under double-blind review

ABSTRACT

Artificial intelligence is increasingly utilized in high-stakes decision-making applications, and this trend is expected to continue growing, resulting in the fairness concerns becoming increasingly significant. While achieving fairness is an idealized goal, it often comes at a cost. From a machine learning perspective, fairness constraints can increase the prediction error and reduce the predictive utility of a model. Such losses in predictive accuracy may also harm individuals or society in some real-world applications. Therefore, it is necessary to consider the balance between fairness and utility by allowing some degree of fairness relaxation in such settings. The main challenge lies in how to achieve optimal trade-offs and how to define the trade-offs between fairness and utility. We study this problem under the framework of counterfactual fairness (CF). We formulate a quantifiable and interpretable measurement of fairness relaxation and characterize the Pareto-optimal predictor at each fairness level. With incomplete causal knowledge, we provide a theoretical analysis of the decomposition and variation of the excess risk bound. Synthetic experiments validate our theoretical results and semi-synthetic experiments based on real-world datasets further demonstrate the effectiveness of our method in more practical settings.

1 INTRODUCTION

Artificial intelligence based prediction systems are increasingly being used to inform decisions in high-stakes domains, including health care Rudin & Ustun (2018), Potash et al. (2015), Obermeyer et al. (2019), loan assessment Khandani et al. (2010), Liu et al. (2018), hiring Hoffman et al. (2018), and criminal justice Brennan et al. (2009), Van Dijck (2022). In these systems, predictions affect the access to treatment, credit, employment, and public resources. Fairness implies that a predictive system should not reproduce or amplify unjust disparities associated with sensitive attributes Barocas & Selbst (2016), Mehrabi et al. (2021). While some notions are formulated in terms of group-level statistical comparisons such as Demographic Parity Zemel et al. (2013) or Equalized Odds Hardt et al. (2016), fairness also has an individual dimension Dwork et al. (2012). In many applications, the concern is not only whether two groups receive similar predictions on average, but whether a particular individual would have received the same prediction had their sensitive attribute been different. This perspective motivates the use of causal reasoning, since such a question cannot be answered solely from observed group-level associations.

Causal approaches to fairness provide a framework for reasoning about the influence of sensitive attributes such as gender or ageChiappa (2019), Galhotra et al. (2022), Khademi et al. (2019). A prominent notion in this direction is counterfactual fairness introduced by Kusner et al. (2017). Counterfactual fairness (CF) provides an individual-level causal notion of fairness by requiring a prediction to remain unchanged when the sensitive attribute is counterfactually altered while the underlying characteristic is kept fixed. This invariance offers a principled way to reason about whether an individual would have received a different prediction under a different sensitive attribute.

Counterfactual fairness constraints, however, need not be costless Zafar et al. (2017), Agarwal et al. (2018). When the sensitive attribute or its causal descendants contain predictive information about the outcome, enforcing complete counterfactual invariance may increase prediction error. Such errors can themselves have meaningful consequences, for example by failing to identify a high-risk patient or incorrectly rejecting a qualified applicant. Hence, in applications where both predictive utility

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and fairness matter, we should balance and make trade-off Shi et al. (2026). The question is how predictive performance changes as fairness is gradually relaxed and how to make optimal trade-offs. Addressing this question requires that the relaxation should have a direct interpretation in terms of the counterfactual disparity being allowed, and the resulting predictor should be optimal at each prescribed fairness level.

Fairness–accuracy trade-offs have been studied in several group-level criteria Zhao & Gordon (2022), Xian et al. (2023), Xu & Strohmer (2023), Dehdashtian et al. (2024). Recent work on causal fairness also studies the utility costs of enforcing causal constraints at the level of causal pathways Plecko & Bareinboim (2025). Some of these works characterize Pareto optimal predictors under the fairness–accuracy trade-off, while others construct trade-off predictors through regularization or constraints without necessarily establishing optimality. For recent work on counterfactual fairness, Zhou et al. (2024) has investigated the optimal predictor under exact CF.

We study this problem by introducing an interpretable relaxation of counterfactual fairness. Let $\delta(u)$ denote the counterfactual gap of the Bayes predictor for an individual with latent state u, and let $\Delta_\phi(u)$ denote the corresponding gap under a predictor ϕ. We require

$$|\Delta_\phi(u)| \leq \epsilon|\delta(u)|, \qquad \epsilon \in [0,1].$$

The parameter ϵ therefore measures the fraction of the Bayes-optimal counterfactual gap that is permitted to remain. At $\epsilon = 0$, the predictor satisfies exact CF, while $\epsilon = 1$ recovers unconstrained Bayes prediction. For each intermediate value of ϵ, we minimize prediction risk subject to this constraint. Varying ϵ consequently traces the Pareto-optimal path between counterfactual fairness and predictive utility, rather than selecting a single operating point through an opaque regularization parameter.

We characterize this trade-off under both squared loss and cross-entropy loss. For squared loss, the optimal predictor admits a simple interpolation between the exact-CF predictor and the Bayes predictor, yielding an explicit characterization of the corresponding excess risk. For cross-entropy loss, the optimal predictor is characterized through a unique solution to a convex one-dimensional problem, which similarly describes the full fairness–utility path. We then move beyond the oracle setting, where both the Bayes predictor and the counterfactual mechanism are known, and study a plug-in construction based on estimated quantities.

Our contributions are as follows:

- We introduce an interpretable relaxation of counterfactual fairness in which ϵ directly controls the fraction of the unconstrained counterfactual gap that is retained, providing a common scale between exact CF and unconstrained prediction.
- We derive the optimal predictor at every relaxation level under squared and cross-entropy losses, thereby characterizing the Pareto frontier between predictive utility and counterfactual fairness.
- We analyze a practical plug-in version using an estimated Bayes predictor and counterfactual generator, and quantify how the two estimation errors affect excess risk along the trade-off path.

2 SETUP

We consider the framework of structural causal model (SCM) Pearl (2009). Let $A \in \{0,1\}$ represent a binary sensitive attribute with $\pi := \mathbb{P}(A = 1) \in (0,1)$, U represent the latent variable, X represent observed covariates, and Y represent the target variable.

A lowercase u denotes a realized value of U and represents the latent background characteristics of a single unit. Counterfactual reasoning compares this same unit under different interventions on A: the value u is held fixed, while the sensitive attribute is set to the intervened value. For such a unit, let $x_a(u) := F_X(u,a)$ denote its covariate value under $A \leftarrow a$, where F_X is the map from (U,A) to X. We consider the following assumption:

Assumption 1. *(i) A and U are independent. (ii) The map $u \mapsto x_a(u)$ is invertible given $A = a$. Equivalently, for any observed pair (x,a), there is a unique u such that $x = x_a(u)$.*

The independent assumption is standard in the fairness literature. The invertible assumption is primarily introduced to make individual-level counterfactuals theoretically tractable, and it has been adopted in recent work on counterfactual identification and fairness Nasr-Esfahany et al. (2023), Zhou et al. (2024), Guldogan et al. (2025). Intuitively, it can be viewed as a feature sufficiency condition: given A, the observed covariates X are sufficiently informative about the latent state relevant to counterfactual feature generation, so that the same (X, A) does not correspond to multiple indistinguishable latent states with different counterfactuals. However, we do not interpret the invertibility as a requirement of real data. When the assumption is violated and the counterfactual distribution is non-degenerate, the deterministic counterfactual used in our theory can instead be viewed as a tractable proxy for this distribution. Importantly, our experiments consider non-invertible data generating mechanisms and show that the proposed method remains effective beyond the idealized theoretical setting.

By assumption 1, we can define a counterfactual generating mechanism:

$$G^\star(x_a(u), a, 1-a) = x_{1-a}(u),$$

where counterfactual generator changes the intervention on sensitive attribute A while keeping the same underlying latent u fixed.

Counterfactual fairness. Let $\phi : (x, a) \mapsto \phi(x, a)$ be a predictor. Define

$$\phi_a(u) := \phi(x_a(u), a),$$
$$\Delta_\phi(u) := \phi_1(u) - \phi_0(u).$$

Following the notion of counterfactual fairness Kusner et al. (2017), a deterministic predictor is perfectly counterfactually fair if changing the sensitive attribute in this counterfactual comparison does not change the prediction. Counterfactual fairness implies

$$\Delta_\phi(U) = 0 \quad \text{almost surely.}$$

Prediction risk and Bayes optimality. Prediction accuracy is measured by the risk

$$R_\ell(\phi) := \mathbb{E}\big[\ell(\phi(X, A), Y)\big],$$

where ℓ is the loss function. Two main tasks are squared loss for regression problems and cross-entropy loss for classification problems, where

$$\ell_2(q, y) := (q - y)^2$$
$$\ell_{\mathrm{CE}}(q, y) := -y \log q - (1 - y) \log(1 - q).$$

For the two tasks, conditional risk $q \mapsto \mathbb{E}[\ell(q, Y) \mid U = u, A = a]$ has a unique minimizer, which is known as Bayes optimal predictor:

$$\mu_a(u) := \arg\min_q \mathbb{E}\big[\ell(q, Y) \mid U = u, A = a\big],$$

and we denote $\phi^\star$ as the Bayes optimal predictor, therefore $\phi^\star(x_a(u), a) = \mu_a(u)$. Under squared loss, $\mu_a(u) = \mathbb{E}[Y \mid U = u, A = a]$; under cross-entropy loss, $\mu_a(u) = \mathbb{P}(Y = 1 \mid U = u, A = a)$. The Bayes optimal predictor yields the gap

$$\delta(u) := \mu_1(u) - \mu_0(u).$$

This gap is nonzero when the Bayes optimal predictor is not counterfactually fair.

Fairness relaxation. The exact counterfactual fairness imposes the prediction gap $\Delta_\phi(u) = 0$, whereas the unconstrained Bayes optimal predictor has the gap $\delta(u)$. A natural pointwise relaxation is to allow the prediction gap to be no larger than an ϵ-fraction of the Bayes optimal gap:

$$|\Delta_\phi(u)| \le \epsilon |\delta(u)|, \quad \epsilon \in [0, 1].$$

The case $\epsilon = 0$ recovers exact CF, while $\epsilon = 1$ makes the Bayes optimal predictor feasible. For the squared loss and binary cross-entropy settings, the optimum under this inequality moves as close to the Bayes optimal predictor as the constraint permits. Therefore, the constraint is tight and reduced to a technically simpler equality constraint $\Delta_\phi(u) = \epsilon\, \delta(u)$. Further, we can define the relaxed counterfactual fairness constraint as

$$\Delta_\phi(U) = \epsilon\, \delta(U) \quad \text{almost surely.}$$

Smaller ϵ values impose a stronger counterfactual-fairness requirement. For each ϵ, denote the predictor ϕ^ϵ. The collection $\{\phi^\epsilon : \epsilon \in [0, 1]\}$ is the family of relaxed CF predictors. When the collection is Pareto optimal, the associated risk curve is the Pareto frontier for counterfactual fairness and predictive accuracy trade-offs.

3 RELAXED COUNTERFACTUAL FAIRNESS PREDICTORS

Among predictors satisfying the relaxed counterfactual fairness constraint $\Delta_\phi(U) = \epsilon\delta(U)$ almost surely, the optimum can be obtained via the pointwise minimization.

Proposition 3.1. *The global prediction risk* $R_\ell(\phi) = \mathbb{E}\big[\ell(\phi(X, A), Y)\big]$ *is expressed by*

$$R_\ell(\phi) = \mathbb{E}_U\Big[(1-\pi)\mathbb{E}\{\ell(\phi_0(U), Y) \mid U, A = 0\} + \pi\mathbb{E}\{\ell(\phi_1(U), Y) \mid U, A = 1\}\Big].$$

Fix $\epsilon \in [0, 1]$ *and define predictor* ϕ^ϵ*. For each* u*, let* $(\phi_0^\epsilon(u), \phi_1^\epsilon(u))$ *solve*

$$\begin{aligned} \min_{z_0, z_1} \quad & (1-\pi)\mathbb{E}\{\ell(z_0, Y) \mid U = u, A = 0\} + \pi\mathbb{E}\{\ell(z_1, Y) \mid U = u, A = 1\} \\ \textit{subject to} \quad & z_1 - z_0 = \epsilon\,\delta(u). \end{aligned} \tag{1}$$

Then ϕ^ϵ *minimizes* $R_\ell(\phi)$ *over all predictors satisfying* $\Delta_\phi(U) = \epsilon\delta(U)$ *almost surely.*

In particular, at $\epsilon = 0$, Proposition 3.1 gives the exact CF endpoint: $\phi_0^0(u) = \phi_1^0(u) = \phi^{\mathrm{cf}}(u)$. For squared loss and cross-entropy loss, $\phi^{\mathrm{cf}}(u) = (1-\pi)\mu_0(u) + \pi\mu_1(u)$, which can be interpreted as a barycenter, as it is the combination of Bayes optimal predictors weighted by group proportion.

Proposition 3.1 is the reduction from the search of predictor functions to the pointwise optimization.

3.1 REGRESSION TASK: SQUARED LOSS

For regression with squared loss, we take $\ell = \ell_2$ and denote $R_2 = R_{\ell_2}$. Let

$$\bar{\mu}(u) := (1-\pi)\mu_0(u) + \pi\mu_1(u).$$

The following theorem provides the ϵ-relaxation predictor and the corresponding excess risk.

Theorem 3.2. *For* $\epsilon \in [0, 1]$ *and each* u*, the unique minimizer in Proposition 3.1 is*

$$\phi_0^\epsilon(u) = \bar{\mu}(u) - \pi\epsilon\,\delta(u), \qquad \phi_1^\epsilon(u) = \bar{\mu}(u) + (1-\pi)\epsilon\,\delta(u).$$

The excess risk is

$$R_2(\phi^\epsilon) - R_2(\phi^\star) = \pi(1-\pi)(1-\epsilon)^2\mathbb{E}_U[\delta(U)^2].$$

This result implies $\phi^\epsilon = (1-\epsilon)\phi^{\mathrm{cf}} + \epsilon\phi^\star$, a simple geometric interpretation. Under squared loss, the entire relaxation path is the line segment in predictor space connecting the CF predictor ϕ^{cf} and the unconstrained Bayes optimal predictor $\phi^\star$. The parameter ϵ determines how far the predictor moves from the fair endpoint toward the Bayes optimal endpoint. The corresponding excess risk is squared cost at a fraction $1-\epsilon$ away from the Bayes endpoint.

3.2 CLASSIFICATION TASK: CROSS-ENTROPY LOSS

We assume a general interior condition for Bayes optimal predictor of success probability $\mu_a(u) \in (0, 1)$ to avoid certain singularity issues. Note that the pair fairness relaxation predictors are both bounded between the pair Bayes optimal predictors, thus they naturally lie in $(0, 1)$ region.

Theorem 3.3. *For* $\epsilon \in [0, 1]$ *and each* u*, the unique minimizer in Proposition 3.1 is*

$$\phi_0^\epsilon(u) = q_\epsilon(u), \qquad \phi_1^\epsilon(u) = q_\epsilon(u) + \epsilon\delta(u),$$

where $q_\epsilon(u)$ *is the unique solution of* $q \in (0, 1)$ *satisfying the first-order equation:*

$$(1-\pi)\frac{q - \mu_0(u)}{q(1-q)} + \pi\frac{q + \epsilon\delta(u) - \mu_1(u)}{(q + \epsilon\delta(u))(1 - q - \epsilon\delta(u))} = 0.$$

The excess risk is

$$\begin{aligned} R_{\ell_{\mathrm{CE}}}(\phi^\epsilon) - R_{\ell_{\mathrm{CE}}}(\phi^\star) = \mathbb{E}_U\Big[&(1-\pi)D_{\mathrm{KL}}\big(\mathrm{Ber}(\mu_0(U)) \,\|\, \mathrm{Ber}(\phi_0^\epsilon(U))\big) \\ &+ \pi D_{\mathrm{KL}}\big(\mathrm{Ber}(\mu_1(U)) \,\|\, \mathrm{Ber}(\phi_1^\epsilon(U))\big)\Big], \end{aligned}$$

where $\mathrm{Ber}(\cdot)$ *denotes the Bernoulli distribution and* $D_{\mathrm{KL}}(\cdot\|\cdot)$ *denotes Kullback–Leibler divergence.*

Unlike the squared-loss case, the cross-entropy solution does not admit a simple form with a clear geometric interpretation. The equation in Theorem 3.3 can be reduced to a cubic polynomial in q, and an algebraic solution can be obtained by Cardano's formula. We omit the explicit expression because it is lengthy and obscures the structure of the optimizer, providing little additional insight. The implicit form instead directly characterizes the optimizer as the unique root in the feasible interval and is also convenient for the subsequent stability analysis.

4 Estimation Construction

We present a plug-in method for practical implementation. The former derivations are written in terms of u for convenience, but in practice the observable factors are (x, a) rather than latent u. It is useful to rewrite the predictor as a function of observed variables:

$$\mu(x, a) := \phi^\star(x, a), \qquad \text{so that} \qquad \mu(x_a(u), a) = \mu_a(u).$$

Let $\hat{\mu}(x, a)$ estimate the unconstrained optimal predictor, and $\hat{G}(x, a, 1-a)$ estimate the counterfactual generator. Given an observed point (x, a), the pair estimated Bayes predictions are

$$\hat{\mu}_a(x, a) := \hat{\mu}(x, a), \qquad \hat{\mu}_{1-a}(x, a) := \hat{\mu}(\hat{G}(x, a, 1-a), 1-a).$$

The first term uses the observed covariates, while the second uses the estimated counterfactual covariates. Let $\hat{\delta}(x, a) := \hat{\mu}_1(x, a) - \hat{\mu}_0(x, a)$. For regression, define

$$\hat{\bar{\mu}}(x, a) := (1-\pi)\hat{\mu}_0(x, a) + \pi\hat{\mu}_1(x, a).$$

Then

$$\hat{\phi}_0^\epsilon(x, a) = \hat{\bar{\mu}}(x, a) - \pi\epsilon\hat{\delta}(x, a), \qquad \hat{\phi}_1^\epsilon(x, a) = \hat{\bar{\mu}}(x, a) + (1-\pi)\epsilon\hat{\delta}(x, a).$$

For binary classification, solve the first-order equation

$$(1-\pi)\frac{q - \hat{\mu}_0(x, a)}{q(1-q)} + \pi\frac{q + \epsilon\hat{\delta}(x, a) - \hat{\mu}_1(x, a)}{(q + \epsilon\hat{\delta}(x, a))(1 - q - \epsilon\hat{\delta}(x, a))} = 0$$

over $q \in (0, 1)$ and $q + \epsilon\hat{\delta}(x, a) \in (0, 1)$, and set

$$\hat{\phi}_0^\epsilon(x, a) = q, \qquad \hat{\phi}_1^\epsilon(x, a) = q + \epsilon\hat{\delta}(x, a),$$

In next section, we study how the estimation errors of $\hat{G}$ and $\hat{\mu}$ affect the predictors.

5 Limited causal knowledge and estimated predictors

The counterfactual generator $G^\star$ and the Bayes prediction function $\mu(x, a)$ may not be known in practice. We consider the following situations. First, μ is known but the counterfactual generator $G^\star$ is estimated. Second, both $G^\star$ and μ are estimated. The statements below give the resulting risk bounds directly, and the decomposition shows how they follow from the underlying errors.

5.1 Squared loss

Firstly, suppose μ is known and the estimated predictor uses estimated $\hat{G}$ instead of ground truth $G^\star$. The following result shows the excess risk caused by the estimated factors.

Proposition 5.1 (squared loss with estimated G)**.** *Assume each $x \mapsto \mu(x, a)$ is L-Lipschitz and*

$$\sup_{x,a} \left\| \hat{G}(x, a, 1-a) - G^\star(x, a, 1-a) \right\| \le \eta_G$$

and suppose μ is known. Then, for each $\epsilon \in [0, 1]$,

$$R_2(\hat{\phi}^\epsilon) - R_2(\phi^\star) \le \pi(1-\pi)(1-\epsilon)^2\mathbb{E}_U[\delta(U)^2] + \pi(1-\pi)(1-\epsilon)^2\Big\{2L\eta_G\mathbb{E}_U|\delta(U)| + L^2\eta_G^2\Big\}.$$

The first term on the right-hand side is exactly the oracle excess risk from Theorem 3.2. The second term is the additional error caused by using $\hat{G}$ to reconstruct the counterfactual covariates. The factor $1-\epsilon$ indicates that errors in the counterfactual generator matter most near the hard-fair endpoint and vanish at $\epsilon = 1$, where the unconstrained predictor does not use the counterfactual generator.

Next, considering the Bayes predictor is also estimated yields the following result.

Proposition 5.2 (squared loss with estimated G and μ)**.** *Assume the condition of Proposition 5.1 and*

$$\sup_{x,a} \left|\hat{\mu}(x,a) - \mu(x,a)\right| \le \eta_\phi,$$

Then, for each $\epsilon \in [0,1]$,

$$\begin{aligned} R_2(\hat{\phi}^\epsilon) - R_2(\phi^\star) \le\ & \pi(1-\pi)(1-\epsilon)^2 \mathbb{E}_U[\delta(U)^2] \\ & + \pi(1-\pi)(1-\epsilon)^2 \Big\{ 2L\eta_G \mathbb{E}_U|\delta(U)| + L^2\eta_G^2 \Big\} \\ & + 4\pi(1-\pi)(1-\epsilon)\eta_\phi \{\mathbb{E}_U|\delta(U)| + L\eta_G\} + \eta_\phi^2. \end{aligned}$$

The excess risk bound adds the prediction estimation error to the counterfactual reconstruction error. Unlike the error caused by generator estimation alone, this additional prediction error does not vanish at $\epsilon = 1$, because the unconstrained endpoint itself still needs to be estimated.

5.2 Cross-entropy loss

In classification task, the estimated predictor is obtained by solving the equation in Theorem 3.3, but with perturbed probability inputs. The following bounds use a local stability argument for the cross-entropy minimizer derivation. This requires excluding boundary effects near 0 and 1.

Assumption 2. *Fix* $\rho \in (0, 1/2)$. *For almost every* u*, the true probabilities, the estimated probabilities, and pairs of predictors for both the oracle and estimation setting lie in* $[\rho, 1-\rho]$.

This assumption is a standard clipping or interiority condition for cross-entropy losses. It ensures that the derivatives of the log-loss remain bounded uniformly over the oracle and estimated predictors. In the bounds below, all values depending on ρ from this boundary condition will be collected into a finite boundary-dependent factor C_ρ.

We first isolate the effect of estimating the counterfactual generator while keeping the Bayes prediction function known.

Proposition 5.3 (cross-entropy with estimated G)**.** *Assume Assumption 2. Suppose* μ *is known, each* $x \mapsto \mu(x,a)$ *is L-Lipschitz, and*

$$\sup_{x,a} \left\|\hat{G}(x,a,1-a) - G^\star(x,a,1-a)\right\| \le \eta_G.$$

Then, for each $\epsilon \in [0,1]$ *and some finite boundary-dependent factor* C_ρ *on* ρ,

$$R_{\ell_{\mathrm{CE}}}(\hat{\phi}^\epsilon) - R_{\ell_{\mathrm{CE}}}(\phi^\star) \le\ R_{\ell_{\mathrm{CE}}}(\phi^\epsilon) - R_{\ell_{\mathrm{CE}}}(\phi^\star) + C_\rho(1-\epsilon)^2 \Big\{ L\eta_G \mathbb{E}_U|\delta(U)| + L^2\eta_G^2 \Big\}.$$

The first term on the right-hand side is exactly the oracle excess risk in Theorem 3.3. The second term is the additional error caused by using $\hat{G}$ to reconstruct the counterfactual covariates. This term is multiplied by $(1-\epsilon)^2$, reflecting that generator error affects the predictor only through the remaining distance from the unconstrained Bayes endpoint. In particular, the generator-induced error vanishes at $\epsilon = 1$, where no counterfactual reconstruction is needed.

Further, We consider the Bayes optimal prediction function is also estimated.

Proposition 5.4 (cross-entropy with estimated G and estimated μ)**.** *Assume Assumption 2 and the conditions of Proposition 5.3. Suppose that*

$$\sup_{x,a} |\hat{\mu}(x,a) - \mu(x,a)| \le \eta_\phi.$$

Then, for each $\epsilon \in [0,1]$ and some finite boundary-dependent factor C_ρ on ρ,

$$R_{\ell_{\mathrm{CE}}}(\hat{\phi}^\epsilon) - R_{\ell_{\mathrm{CE}}}(\phi^\star) \le R_{\ell_{\mathrm{CE}}}(\phi^\epsilon) - R_{\ell_{\mathrm{CE}}}(\phi^\star) + C_\rho(1-\epsilon)^2\Big\{L\eta_G \mathbb{E}_U|\delta(U)| + L^2\eta_G^2\Big\} + C_\rho\Big[(1-\epsilon)\eta_\phi\{\mathbb{E}_U|\delta(U)| + L\eta_G\} + \eta_\phi^2\Big].$$

The bound in Proposition 5.4 shows a similar layer structure to the squared-loss result. The last layer is the additional error caused by the estimated μ.

6 Experiments

In this section, we validate our theoretical results and evaluate the proposed predictor. On synthetic experiment, we focus on validating the excess-risk characterization and studying how misspecification of the counterfactual generator affects the resulting trade-off frontier. On semi-synthetic experiment, we assess the effectiveness of our method in a more practical setting.

Prediction metric. We measure predictive performance by excess risk, which directly quantifies the predictive cost associated with a given level of counterfactual fairness relaxation. For a predictor $\hat{\phi}$ and factual observations $\{(x^{(i)}, a^{(i)}, y^{(i)})\}_{i=1}^n$, we report the empirical excess risk

$$\mathcal{E}_\ell(\hat{\phi}) = \frac{1}{n}\sum_{i=1}^n \left[\ell\Big(\hat{\phi}(x^{(i)}, a^{(i)}), y^{(i)}\Big) - \ell\Big(\mu(x^{(i)}, a^{(i)}), y^{(i)}\Big)\right]. \tag{2}$$

We use the known Bayes predictor μ in synthetic experiments and $\hat{\mu}$ in semi-synthetic experiments. For the loss function, we use squared loss for regression and cross-entropy loss for classification.

6.1 Synthetic experiments

The synthetic experiment is designed to validate and illustrate our theoretical results on how estimation error of the related factors affects the excess risk. For both regression and classification tasks, let

$$U \sim \mathrm{Unif}[b_1, b_2], \qquad A \sim \mathrm{Bernoulli}(\pi), \qquad X = \gamma U + \tau A.$$

Set the true generator $G^\star(x, a, 1-a) = x + \tau - 2\tau a$. We use $b_1 = -1$, $b_2 = 1$, $\pi = 1/2$, $\gamma = 1$ and $\tau = 1$. To introduce controlled misspecification, we replace it by

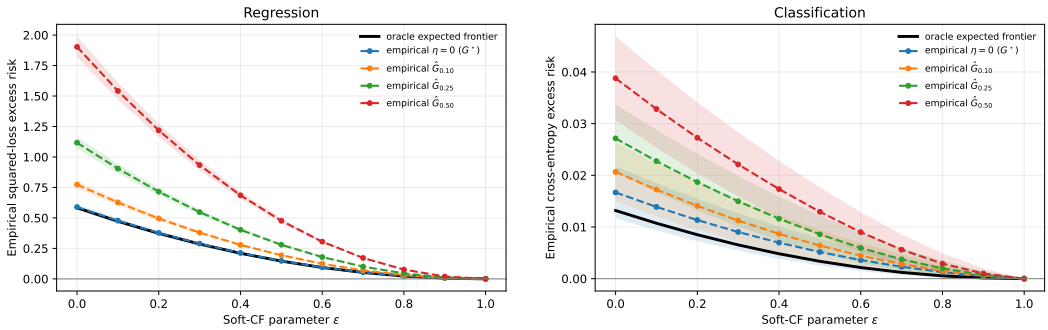
$$\hat{G}_\eta(x, a, 1-a) = x + (1+\eta)(1-2a), \qquad \eta \in \{0, 0.10, 0.25, 0.50\}.$$

Here $\eta = 0$ recovers $G^\star$, while positive η indicates the counterfactual covariate shift. In this section, we keep μ known to isolate the effect of counterfactual reconstruction; estimated μ leads to the additional prediction error as discussed in Section 5. Additional synthetic experiments with estimated $\hat{\mu}$ appear in Appendix B.2.

For regression, let $Y = X^2 + U + \xi$, where $\xi \sim N(0, 0.1^2)$ is independent noise. Then $\mu(x, a) = x^2 + x - a$ and $\delta(u) = 2u + 1$, where the signed gap is negative for $u < -0.5$ and positive for $u > -0.5$, so the experiment also checks that the predictors handle both directions of the Bayes gap. Theorem 3.2 gives the population oracle frontier $R_2(\phi^\epsilon) - R_2(\phi^\star) = \frac{7}{12}(1-\epsilon)^2$. This analytical solution helps validate the experiment output.

For classification, let $Y \mid X, A \sim \mathrm{Bernoulli}(\mu(X, A))$, with $\mu(x, a) = \sigma\{0.5(x^2 + x - a)\}$ and $\sigma(t) = (1 + e^{-t})^{-1}$.

Figure 1 uses $\epsilon \in \{0, 0.1, \ldots, 1\}$ and 30 repetitions of 1,000 fresh observations. As η increases, the excess risk rises systematically, with the effect being most pronounced under stricter fairness constraints. Moreover, the impact of generator misspecification diminishes as ϵ increases, consistent with the fact that relaxing counterfactual fairness reduces the predictor's dependence on counterfactual reconstruction. All curves converge to zero at $\epsilon = 1$, where the predictor reduces to the Bayes predictor and no longer depends on the generator.



(a) Regression: squared loss.

(b) Classification: cross-entropy.

Figure 1: Synthetic excess risk with known μ and misspecified $\hat{G}_\eta$. Black curves show the population oracle frontiers. Dashed curves show average over 30 independent samples of size 1,000; shaded bands indicate one sample standard deviation.

6.2 Semi-synthetic datasets

We consider the Law School Success dataset Wightman (1998) and the Adult Income dataset Asuncion et al. (2007). The prediction tasks are first-year grade prediction and binary income classification, respectively, with gender as the sensitive attribute. These experiments evaluate the trade-off predictors in more practical scenarios where both the prediction function and the counterfactual generator are estimated, without imposing the invertibility assumption used in the theoretical analysis.

Evaluating counterfactual fairness requires access to ground-truth counterfactuals. We therefore follow the semi-synthetic setup of Zhou et al. (2024) and Zuo et al. (2023). In this setup, a variational autoencoder (VAE) fitted to Law School and a disentangled causal effect VAE (DCEVAE) fitted to Adult define the data-generating mechanisms. Factual and counterfactual observations are generated by changing the sensitive attribute while retaining the same underlying latent variables. The resulting counterfactuals are ground truth under these mechanisms and are reserved for fairness evaluation. Downstream prediction instead uses a separately estimated counterfactual generator $\hat{G}$ and a prediction model $\hat{\mu}$ fitted to factual observations. Details are given in Appendix B.3.

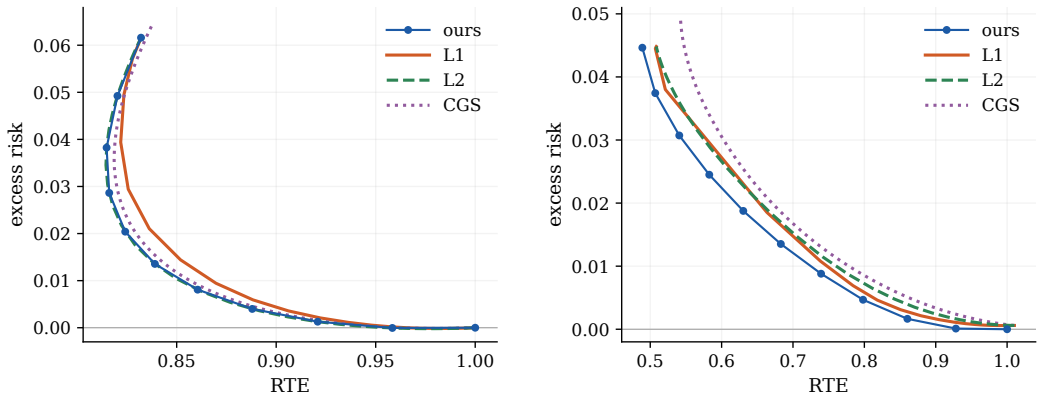
To measure the achieved fairness level of the predictor, we normalize the total effect criterion and define the relative total effect (RTE) as

$$\mathrm{RTE}(\hat{\varphi}) = \frac{\sum_{i=1}^{n} \left| \hat{\varphi}(x^{(i)}, a^{(i)}) - \hat{\varphi}\left(x^{(i)}_{1-a^{(i)}}, 1-a^{(i)}\right) \right|}{\sum_{i=1}^{n} \left| \hat{\mu}(x^{(i)}, a^{(i)}) - \hat{\mu}\left(x^{(i)}_{1-a^{(i)}}, 1-a^{(i)}\right) \right|}, \tag{3}$$

where $x^{(i)}_{1-a^{(i)}}$ is the ground-truth counterfactual of the ith test observation. Smaller RTE indicates a smaller counterfactual discrepancy, and $\hat{\mu}$ has RTE 1 and excess risk 0. Since estimation will make achieved RTE differ from the prescribed ϵ, we compare excess risk against achieved RTE.

We compare the proposed predictors with L1 and L2 regularization and Counterfactual Gap Shrinkage (CGS). All methods post-process the same $\hat{\mu}$ using the same estimated counterfactuals. We vary each method's relaxation or regularization parameter to obtain its trade-off curve. Experiments are repeated 5 times and metrics are averaged at the same parameter value.

Figure 2 show the excess risk curve on the Law School and Adult datasets. On Law School, our method overlaps with L2, consistent with their theoretical equivalence under squared loss, and generally achieves lower excess risk than L1 and CGS at comparable RTE levels. On Adult, where this equivalence no longer holds under cross-entropy loss, our method yields a more favorable trade-off than the competing methods over much of the common RTE range. Overall, these results show that the proposed method remains effective in more practical settings, with estimated Bayes



(a) Law School (b) Adult

Figure 2: Excess risk versus RTE. The horizontal axis shows the achieved relative total effect (RTE), and the vertical axis shows the excess risk. Our method is shown by the blue solid curve, L1 regularization by the orange solid curve, L2 regularization by the green dashed curve, and Counterfactual Gap Shrinkage (CGS) by the purple dotted curve.

predictors and counterfactual generators and under semi-synthetic data-generating mechanisms that do not satisfy the invertibility assumption.

7 Discussion

Existing fairness–utility trade-off methods often view the relaxation of fairness as allowing the predictor to depend more strongly on information related to the sensitive attribute. This intuition is natural: since the sensitive attribute may carry predictive information, allowing more of this information to enter the prediction can improve accuracy. However, such a perspective also raises several problems in quantification and interpretation. For example, it is unclear how to measure the amount of sensitive information that should be allowed in exchange for a given gain in predictive accuracy. Also, sensitive information is generally not decomposed into independent pieces, making it difficult to determine which part of the information should be relaxed first. Moreover, as the target accuracy level varies, the resulting relaxation strategy may not be monotone or structurally stable. Finally, such a strategy may not consistently account for both the direct effect of the sensitive attribute on the target and its indirect effect through the covariates.

In contrast, this paper approaches the trade-off from the perspective of prediction outcomes. Essentially, unfairness is reflected in the discrepancy between the predictions, and exact fairness aims to eliminate the discrepancy. Consequently, the proportion of retained disparity naturally serves as a metric to quantify the level of fairness. The proposed framework has practical implications for decision making. The Pareto frontier makes the predictive cost of requiring fairness explicit and measures the marginal accuracy gain by relaxing fairness. Moreover, the ϵ-budget makes the relaxation level quantifiable and controllable.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, provide feedback on mathematical derivations and proofs, support experimental design, and assist with the interpretation of experimental results. Generative AI tools were also used for code editing and debugging, literature search, and drafting and editing parts of the manuscript. We reviewed and verified all AI-assisted mathematical arguments, experimental procedures, code, citations, and written content. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018.

Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and behavior*, 36(1):21–40, 2009.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 7801–7808, 2019.

Sepehr Dehdashtian, Bashir Sadeghi, and Vishnu Naresh Boddeti. Utility-fairness trade-offs and how to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12037–12046, 2024.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R Varshney. Causal feature selection for algorithmic fairness. In *Proceedings of the 2022 International Conference on Management of Data*, pp. 276–285, 2022.

Ozgur Guldogan, Neeraj Sarna, Yuanyuan Li, and Michael Berger. Counterfactually fair conformal prediction. *arXiv preprint arXiv:2510.08724*, 2025.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Mitchell Hoffman, Lisa B Kahn, and Danielle Li. Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800, 2018.

Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, pp. 2907–2914, 2019.

Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

Hyemi Kim, Seungjae Shin, JoonHo Jang, Kyungwoo Song, Weonyoung Joo, Wanmo Kang, and Il-Chul Moon. Counterfactual fairness with disentangled causal effect variational autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8128–8136, 2021.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *International conference on machine learning*, pp. 25733–25754. PMLR, 2023.

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, 2019.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Drago Plecko and Elias Bareinboim. Fairness-accuracy trade-offs: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26344–26353, 2025.

Eric Potash, Joe Brew, Alexander Loewi, Subhabrata Majumdar, Andrew Reece, Joe Walsh, Eric Rozier, Emile Jorgenson, Raed Mansour, and Rayid Ghani. Predictive modeling for public health: Preventing childhood lead poisoning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2039–2047, 2015.

Cynthia Rudin and Berk Ustun. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, 48(5):449–466, 2018.

Enze Shi, Pankaj Bhagwat, Zhixian Yang, Linglong Kong, and Bei Jiang. Understanding fairness and prediction error through subspace decomposition and influence analysis. *Advances in Neural Information Processing Systems*, 38:99862–99887, 2026.

Gijs Van Dijck. Predicting recidivism risk meets ai act. *European Journal on Criminal Policy and Research*, 28(3):407–423, 2022.

Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

Ruicheng Xian, Lang Yin, and Han Zhao. Fair and optimal classification via post-processing. In *International conference on machine learning*, pp. 37977–38012. PMLR, 2023.

Shizhou Xu and Thomas Strohmer. Fair data representation for machine learning at the pareto frontier. *Journal of Machine Learning Research*, 24(331):1–63, 2023.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research*, 23(57):1–26, 2022.

Zeyu Zhou, Tianci Liu, Ruqi Bai, Jing Gao, Murat Kocaoglu, and David I Inouye. Counterfactual fairness by combining factual and counterfactual predictions. *Advances in Neural Information Processing Systems*, 37:47876–47907, 2024.

Zhiqun Zuo, Mahdi Khalili, and Xueru Zhang. Counterfactually fair representation. *Advances in neural information processing systems*, 36:12124–12140, 2023.

A PROOFS

A.1 PROOF OF PROPOSITION 3.1

We first rewrite the constrained risk in a form that separates over values of the latent variable U. For any predictor ϕ, by the tower property of conditional expectation:

$$\begin{aligned} R_\ell(\phi) &= \mathbb{E}\big[\ell(\phi(X, A), Y)\big] \\ &= \mathbb{E}\big[\ell(\phi_A(U), Y)\big] \\ &= \mathbb{E}_U\Big[\mathbb{E}\big\{\ell(\phi_A(U), Y) \mid U\big\}\Big]. \end{aligned}$$

Because A is independent of U and $\mathbb{P}(A = 1) = \pi$, we further obtain

$$\begin{aligned} \mathbb{E}\big\{\ell(\phi_A(U), Y) \mid U = u\big\} = (1-\pi)\mathbb{E}\{\ell(\phi_0(u), Y) \mid U = u, A = 0\} \\ + \pi\mathbb{E}\{\ell(\phi_1(u), Y) \mid U = u, A = 1\}. \end{aligned}$$

Thus

$$R_\ell(\phi) = \mathbb{E}_U\Big[(1-\pi)\mathbb{E}\{\ell(\phi_0(U), Y) \mid U, A = 0\} + \pi\mathbb{E}\{\ell(\phi_1(U), Y) \mid U, A = 1\}\Big].$$

Fix $\epsilon \in [0, 1]$. Let

$$\mathcal{F}_\epsilon := \Big\{\phi : \Delta_\phi(U) = \epsilon\delta(U) \text{ almost surely}\Big\}$$

be the class of predictors satisfying the relaxed counterfactual fairness constraint.

Now define the pointwise conditional objective

$$M_u(z_0, z_1) := (1-\pi)\mathbb{E}\{\ell(z_0, Y) \mid U = u, A = 0\} + \pi\mathbb{E}\{\ell(z_1, Y) \mid U = u, A = 1\}.$$

Then the global risk can be written as

$$R_\ell(\phi) = \mathbb{E}_U\big[M_U(\phi_0(U), \phi_1(U))\big].$$

Moreover, the global constraint

$$\Delta_\phi(U) = \epsilon\delta(U) \quad \text{almost surely}$$

Let $(\phi_0^\epsilon(u), \phi_1^\epsilon(u))$ solve the pointwise problem equation 1 for P_U-almost every u. By definition, for almost every u and for every pair (z_0, z_1) satisfying

$$z_1 - z_0 = \epsilon\delta(u),$$

we have

$$M_u(\phi_0^\epsilon(u), \phi_1^\epsilon(u)) \le M_u(z_0, z_1).$$

Now take any competing predictor $\phi \in \mathcal{F}_\epsilon$. Since $\phi \in \mathcal{F}_\epsilon$, its induced pair

$$(\phi_0(u), \phi_1(u))$$

satisfies

$$\phi_1(u) - \phi_0(u) = \epsilon\delta(u)$$

for P_U-almost every u. Therefore, applying the pointwise optimality inequality with

$$z_0 = \phi_0(u), \qquad z_1 = \phi_1(u),$$

gives

$$M_u(\phi_0^\epsilon(u), \phi_1^\epsilon(u)) \le M_u(\phi_0(u), \phi_1(u))$$

for P_U-almost every u. Integrating both sides with respect to P_U yields

$$\begin{aligned} R_\ell(\phi^\epsilon) &= \mathbb{E}_U\big[M_U(\phi_0^\epsilon(U), \phi_1^\epsilon(U))\big] \\ &\le \mathbb{E}_U\big[M_U(\phi_0(U), \phi_1(U))\big] \\ &= R_\ell(\phi). \end{aligned}$$

Since $\phi \in \mathcal{F}_\epsilon$ was arbitrary, ϕ^ϵ is a global minimizer of $R_\ell(\phi)$ over $\mathcal{F}_\epsilon$.

A.2 Proof of Theorem 3.2

Fix $\epsilon \in [0, 1]$. By Proposition 3.1, it is enough to solve the constrained problem pointwise for each value of u.

Under squared loss, the conditional risk admits the standard variance decomposition. For each $a \in \{0, 1\}$ and any scalar prediction z_a,

$$\mathbb{E}\{(z_a - Y)^2 \mid U = u, A = a\} = (z_a - \mu_a)^2 + \mathrm{Var}(Y \mid U = u, A = a),$$

where $\mu_a = \mathbb{E}[Y \mid U = u, A = a]$. The variance term does not depend on z_a. Hence, for the purpose of minimizing the pointwise objective in Proposition 3.1, we only need to minimize

$$(1 - \pi)(z_0 - \mu_0)^2 + \pi(z_1 - \mu_1)^2$$

subject to

$$z_1 - z_0 = \epsilon\delta.$$

Then

$$z_1 - \mu_1 = z_0 + \epsilon\delta - \mu_1 = z_0 - \mu_0 - (1 - \epsilon)\delta.$$

Let

$$r = z_0 - \mu_0.$$

The constrained pointwise objective becomes the one-dimensional quadratic function

$$F_u(r) = (1 - \pi)r^2 + \pi\{r - (1 - \epsilon)\delta\}^2.$$

Setting $F'_u(r) = 0$ yields

$$r^\epsilon = \pi(1 - \epsilon)\delta.$$

Thus

$$z_0^\epsilon = \mu_0 + \pi(1 - \epsilon)\delta.$$

The constraint $z_1 - z_0 = \epsilon\delta$ then gives

$$z_1^\epsilon = z_0^\epsilon + \epsilon\delta = \bar{\mu}(u) + (1 - \pi)\epsilon\delta(u).$$

Therefore, for each u,

$$\phi_0^\epsilon(u) = \bar{\mu}(u) - \pi\epsilon\delta(u), \qquad \phi_1^\epsilon(u) = \bar{\mu}(u) + (1 - \pi)\epsilon\delta(u).$$

The strict convexity of F_u also proves uniqueness of the pointwise minimizer.

Next, we verify the equivalent representation. For $a = 0$,

$$(1 - \epsilon)\phi_0^{\mathrm{cf}}(u) + \epsilon\phi_0^\star(u) = (1 - \epsilon)\bar{\mu}(u) + \epsilon\mu_0(u) = \bar{\mu}(u) - \epsilon\{\bar{\mu}(u) - \mu_0(u)\}.$$

Since

$$\bar{\mu}(u) - \mu_0(u) = \pi\delta(u),$$

we obtain

$$(1 - \epsilon)\phi_0^{\mathrm{cf}}(u) + \epsilon\phi_0^\star(u) = \bar{\mu}(u) - \pi\epsilon\delta(u) = \phi_0^\epsilon(u).$$

Similarly, for $a = 1$,

$$(1 - \epsilon)\phi_1^{\mathrm{cf}}(u) + \epsilon\phi_1^\star(u) = (1 - \epsilon)\bar{\mu}(u) + \epsilon\mu_1(u) = \bar{\mu}(u) + \epsilon\{\mu_1(u) - \bar{\mu}(u)\}.$$

Since

$$\mu_1(u) - \bar{\mu}(u) = (1 - \pi)\delta(u),$$

we get

$$(1 - \epsilon)\phi_1^{\mathrm{cf}}(u) + \epsilon\phi_1^\star(u) = \bar{\mu}(u) + (1 - \pi)\epsilon\delta(u) = \phi_1^\epsilon(u).$$

Hence

$$\phi^\epsilon = (1 - \epsilon)\phi^{\mathrm{cf}} + \epsilon\phi^\star$$

.

It remains to compute the excess risk. For any predictor ϕ, again using the conditional variance decomposition,

$$\begin{aligned} R_2(\phi) &= \mathbb{E}_U\Big[(1-\pi)\mathbb{E}\{(\phi_0(U)-Y)^2 \mid U, A=0\} + \pi\mathbb{E}\{(\phi_1(U)-Y)^2 \mid U, A=1\}\Big] \\ &= \mathbb{E}_U\Big[(1-\pi)\{(\phi_0(U)-\mu_0(U))^2 + \mathrm{Var}(Y \mid U, A=0)\} \\ &\qquad + \pi\{(\phi_1(U)-\mu_1(U))^2 + \mathrm{Var}(Y \mid U, A=1)\}\Big]. \end{aligned}$$

For the Bayes predictor $\phi^\star$, the squared-error terms vanish because $\phi_a^\star(U) = \mu_a(U)$. Therefore,

$$R_2(\phi) - R_2(\phi^\star) = \mathbb{E}_U\Big[(1-\pi)(\phi_0(U)-\mu_0(U))^2 + \pi(\phi_1(U)-\mu_1(U))^2\Big].$$

Applying this identity to ϕ^ϵ, we have

$$\phi_0^\epsilon(u) - \mu_0(u) = \bar{\mu}(u) - \pi\epsilon\delta(u) - \mu_0(u) = \pi(1-\epsilon)\delta(u),$$

and

$$\phi_1^\epsilon(u) - \mu_1(u) = \bar{\mu}(u) + (1-\pi)\epsilon\delta(u) - \mu_1(u) = -(1-\pi)(1-\epsilon)\delta(u).$$

Hence

$$\begin{aligned} R_2(\phi^\epsilon) - R_2(\phi^\star) &= \mathbb{E}_U\Big[(1-\pi)\pi^2(1-\epsilon)^2\delta(U)^2 + \pi(1-\pi)^2(1-\epsilon)^2\delta(U)^2\Big] \\ &= (1-\epsilon)^2\Big\{(1-\pi)\pi^2 + \pi(1-\pi)^2\Big\}\mathbb{E}_U[\delta(U)^2]. \end{aligned}$$

Since

$$(1-\pi)\pi^2 + \pi(1-\pi)^2 = \pi(1-\pi)\{\pi + (1-\pi)\} = \pi(1-\pi),$$

we conclude that

$$R_2(\phi^\epsilon) - R_2(\phi^\star) = \pi(1-\pi)(1-\epsilon)^2\mathbb{E}_U[\delta(U)^2].$$

A.3 Proof of Theorem 3.3

Under the assumption $\mu_a \in (0,1)$, the conditional distribution of Y given $U=u, A=a$ is Bernoulli with success probability μ_a. For any prediction probability $q \in (0,1)$, define the conditional cross-entropy risk

$$\ell_{\mathrm{CE}}(q, \mu_a) := -\mu_a \log q - (1-\mu_a)\log(1-q).$$

After substituting $z_0 = q$ and $z_1 = q + \epsilon\delta$, the pointwise objective in Proposition 3.1 becomes

$$F_u(q) = (1-\pi)\ell_{\mathrm{CE}}(q, \mu_0) + \pi\ell_{\mathrm{CE}}(q+\epsilon\delta, \mu_1).$$

We first show that this objective attains a unique minimum. Differentiate F_u. For a Bernoulli cross-entropy term,

$$\frac{\partial}{\partial q}\ell_{\mathrm{CE}}(q, \mu) = -\frac{\mu}{q} + \frac{1-\mu}{1-q} = \frac{q-\mu}{q(1-q)}.$$

Therefore,

$$F_u'(q) = (1-\pi)\frac{q-\mu_0}{q(1-q)} + \pi\frac{q+\epsilon\delta-\mu_1}{(q+\epsilon\delta)(1-q-\epsilon\delta)}.$$

Thus any interior minimizer must satisfy

$$(1-\pi)\frac{q-\mu_0}{q(1-q)} + \pi\frac{q+\epsilon\delta-\mu_1}{(q+\epsilon\delta)(1-q-\epsilon\delta)} = 0.$$

To prove uniqueness, compute the second derivative:

$$F_u''(q) = (1-\pi)\left\{\frac{\mu_0}{q^2} + \frac{1-\mu_0}{(1-q)^2}\right\} + \pi\left\{\frac{\mu_1}{(q+\epsilon\delta)^2} + \frac{1-\mu_1}{(1-q-\epsilon\delta)^2}\right\}.$$

Since $\mu_0, \mu_1 \in (0,1)$ and all feasible prediction probabilities lie in $(0,1)$, we have

$$F_u''(q) > 0.$$

Hence F_u is strictly convex. The minimizer is therefore unique, and because the minimum is attained in the interior, it is the unique solution of the first-order equation. Denote this solution by $q_\epsilon(u)$. Then the unique pointwise minimizer is

$$\phi_0^\epsilon(u) = q_\epsilon(u), \qquad \phi_1^\epsilon(u) = q_\epsilon(u) + \epsilon\delta(u).$$

It remains to compute the excess risk. For each $a \in \{0,1\}$ and each prediction probability $q_a \in (0,1)$,

$$\mathbb{E}\{\ell_{\mathrm{CE}}(q_a, Y) \mid U = u, A = a\} = -\mu_a \log q_a - (1-\mu_a)\log(1-q_a) = \ell_{\mathrm{CE}}(q_a, \mu_a).$$

The Bayes prediction under cross-entropy is $q_a = \mu_a$. Therefore,

$$\begin{aligned}
&\ell_{\mathrm{CE}}(q_a, \mu_a) - \ell_{\mathrm{CE}}(\mu_a, \mu_a) \\
&\quad = \mu_a \log \frac{\mu_a}{q_a} + (1-\mu_a)\log\frac{1-\mu_a}{1-q_a} \\
&\quad = D_{\mathrm{KL}}\big(\mathrm{Ber}(\mu_a) \,\|\, \mathrm{Ber}(q_a)\big).
\end{aligned}$$

Applying this identity with

$$q_0 = \phi_0^\epsilon(u), \qquad q_1 = \phi_1^\epsilon(u),$$

and averaging over U, we obtain

$$\begin{aligned}
R_{\ell_{\mathrm{CE}}}(\phi^\epsilon) - R_{\ell_{\mathrm{CE}}}(\phi^\star) = \mathbb{E}_U\Big[&(1-\pi) D_{\mathrm{KL}}\big(\mathrm{Ber}(\mu_0(U)) \,\|\, \mathrm{Ber}(\phi_0^\epsilon(U))\big) \\
&+ \pi D_{\mathrm{KL}}\big(\mathrm{Ber}(\mu_1(U)) \,\|\, \mathrm{Ber}(\phi_1^\epsilon(U))\big)\Big].
\end{aligned}$$

A.4 Proof of Proposition 5.1

For a latent individual u, define the counterfactual reconstruction errors

$$\zeta_1(u) := \mu(\hat{G}(x_0(u), 0, 1), 1) - \mu_1(u),$$

and

$$\zeta_0(u) := \mu(\hat{G}(x_1(u), 1, 0), 0) - \mu_0(u).$$

By the Lipschitz condition and the uniform generator error bound,

$$|\zeta_1(u)| \le L\eta_G, \qquad |\zeta_0(u)| \le L\eta_G.$$

Consider first the prediction for the observed state $A = 0$. The estimation uses the true factual value $\mu_0(u)$ and the reconstructed counterfactual value $\mu_1(u) + \zeta_1(u)$. Therefore

$$\hat{\delta}_0(u) = \{\mu_1(u) + \zeta_1(u)\} - \mu_0(u) = \delta(u) + \zeta_1(u).$$

The squared-loss predictor estimation gives

$$\hat{\phi}_0^\epsilon(u) = \mu_0(u) + \pi(1-\epsilon)\{\delta(u) + \zeta_1(u)\}.$$

Since the oracle predictor satisfies

$$\phi_0^\epsilon(u) = \mu_0(u) + \pi(1-\epsilon)\delta(u),$$

we have

$$\hat{\phi}_0^\epsilon(u) = \phi_0^\epsilon(u) + \pi(1-\epsilon)\zeta_1(u).$$

Similarly, for the observed state $A = 1$, we obtain

$$\hat{\phi}_1^\epsilon(u) = \phi_1^\epsilon(u) + (1-\pi)(1-\epsilon)\zeta_0(u).$$

Now compute the squared-loss excess risk. By variance decomposition, we have

$$R_2(\hat{\phi}^\epsilon) - R_2(\phi^\star) = \mathbb{E}_U\Big[(1-\pi)(\hat{\phi}_0^\epsilon(U) - \mu_0(U))^2 + \pi(\hat{\phi}_1^\epsilon(U) - \mu_1(U))^2\Big].$$

Therefore,

$$R_2(\hat{\phi}^\epsilon) - R_2(\phi^\star) = \pi(1-\pi)(1-\epsilon)^2\mathbb{E}_U\Big[\pi\{\delta(U)+\zeta_1(U)\}^2 + (1-\pi)\{\delta(U)-\zeta_0(U)\}^2\Big].$$

Expanding the square separates the oracle term from the generator-induced terms:

$$\begin{aligned}R_2(\hat{\phi}^\epsilon) - R_2(\phi^\star) =& \pi(1-\pi)(1-\epsilon)^2\mathbb{E}_U[\delta(U)^2] \\ &+ \pi(1-\pi)(1-\epsilon)^2\mathbb{E}_U\Big[2\delta(U)\{\pi\zeta_1(U)-(1-\pi)\zeta_0(U)\} \\ &\qquad + \pi\zeta_1(U)^2 + (1-\pi)\zeta_0(U)^2\Big].\end{aligned}$$

The first term is exactly oracle excess risk. For the second term, using

$$|\zeta_0(u)| \le L\eta_G, \qquad |\zeta_1(u)| \le L\eta_G,$$

we get

$$\begin{aligned}&2\delta\{\pi\zeta_1 - (1-\pi)\zeta_0\} + \pi\zeta_1^2 + (1-\pi)\zeta_0^2 \\ &\quad\le 2|\delta|\{\pi|\zeta_1| + (1-\pi)|\zeta_0|\} + \pi\zeta_1^2 + (1-\pi)\zeta_0^2 \\ &\quad\le 2L\eta_G|\delta| + L^2\eta_G^2.\end{aligned}$$

Averaging over U gives the result.

A.5 Proof of Proposition 5.2

Let $\hat{\phi}^\epsilon_G$ denote the predictor in Proposition 5.1, which uses the estimated generator $\hat{G}$ but the true Bayes prediction function μ. Let $\hat{\phi}^\epsilon$ denote the fully estimated predictor, which uses both $\hat{G}$ and $\hat{\mu}$.

From the proof of Proposition 5.1, define

$$r_a(u) := \hat{\phi}^\epsilon_{G,a}(u) - \mu_a(u), \qquad a \in \{0,1\}.$$

Then

$$r_0(u) = \pi(1-\epsilon)\{\delta(u)+\zeta_1(u)\},$$

and

$$r_1(u) = -(1-\pi)(1-\epsilon)\{\delta(u)-\zeta_0(u)\}.$$

Moreover,

$$|\zeta_0(u)| \le L\eta_G, \qquad |\zeta_1(u)| \le L\eta_G.$$

Now compare the fully estimated predictor $\hat{\phi}^\epsilon$ with $\hat{\phi}^\epsilon_G$. For the factual evaluations, define

$$\alpha_0(u) := \hat{\mu}(x_0(u),0) - \mu_0(u), \qquad \alpha_1(u) := \hat{\mu}(x_1(u),1) - \mu_1(u).$$

For the reconstructed counterfactual evaluations, define

$$\beta_1(u) := \hat{\mu}(\hat{G}(x_0(u),0,1),1) - \mu(\hat{G}(x_0(u),0,1),1),$$

and

$$\beta_0(u) := \hat{\mu}(\hat{G}(x_1(u),1,0),0) - \mu(\hat{G}(x_1(u),1,0),0).$$

By the uniform bound on $\hat{\mu}$,

$$|\alpha_0(u)|, |\alpha_1(u)|, |\beta_0(u)|, |\beta_1(u)| \le \eta_\phi.$$

For the state $A = 0$, we have

$$\hat{\phi}^\epsilon_0(u) - \hat{\phi}^\epsilon_{G,0}(u) = (1-\pi+\pi\epsilon)\alpha_0(u) + \pi(1-\epsilon)\beta_1(u).$$

The two coefficients sum to one. Hence

$$|\hat{\phi}^\epsilon_0(u) - \hat{\phi}^\epsilon_{G,0}(u)| \le \eta_\phi.$$

Similarly, for the state $A = 1$,

$$\hat{\phi}_1^\epsilon(u) - \hat{\phi}_{G,1}^\epsilon(u) = \{\pi + (1-\pi)\epsilon\}\alpha_1(u) + (1-\pi)(1-\epsilon)\beta_0(u),$$

Therefore,

$$|\hat{\phi}_1^\epsilon(u) - \hat{\phi}_{G,1}^\epsilon(u)| \le \eta_\phi.$$

Define

$$d_a(u) := \hat{\phi}_a^\epsilon(u) - \hat{\phi}_{G,a}^\epsilon(u), \qquad a \in \{0,1\}.$$

Then

$$|d_0(u)| \le \eta_\phi, \qquad |d_1(u)| \le \eta_\phi.$$

Since

$$\hat{\phi}_a^\epsilon(u) - \mu_a(u) = r_a(u) + d_a(u),$$

we have

$$\begin{aligned} R_2(\hat{\phi}^\epsilon) - R_2(\phi^\star) =& \mathbb{E}_U\Big[(1-\pi)\{r_0(U) + d_0(U)\}^2 + \pi\{r_1(U) + d_1(U)\}^2\Big] \\ =& R_2(\hat{\phi}_G^\epsilon) - R_2(\phi^\star) \\ &+ \mathbb{E}_U\Big[(1-\pi)\{2r_0(U)d_0(U) + d_0(U)^2\} + \pi\{2r_1(U)d_1(U) + d_1(U)^2\}\Big]. \end{aligned}$$

Using $|d_a(u)| \le \eta_\phi$, we obtain

$$\begin{aligned} R_2(\hat{\phi}^\epsilon) - R_2(\phi^\star) \le& R_2(\hat{\phi}_G^\epsilon) - R_2(\phi^\star) \\ &+ 2\eta_\phi \mathbb{E}_U\Big[(1-\pi)|r_0(U)| + \pi|r_1(U)|\Big] + \eta_\phi^2. \end{aligned}$$

It remains to bound the residual term. From the expressions for r_0 and r_1,

$$\begin{aligned} &(1-\pi)|r_0(u)| + \pi|r_1(u)| \\ &\quad = \pi(1-\pi)(1-\epsilon)\Big(|\delta(u) + \zeta_1(u)| + |\delta(u) - \zeta_0(u)|\Big) \\ &\quad \le 2\pi(1-\pi)(1-\epsilon)\{|\delta(u)| + L\eta_G\}. \end{aligned}$$

Therefore,

$$\mathbb{E}_U\Big[(1-\pi)|r_0(U)| + \pi|r_1(U)|\Big] \le 2\pi(1-\pi)(1-\epsilon)\{\mathbb{E}_U|\delta(U)| + L\eta_G\}.$$

Hence

$$\begin{aligned} R_2(\hat{\phi}^\epsilon) - R_2(\phi^\star) \le& R_2(\hat{\phi}_G^\epsilon) - R_2(\phi^\star) \\ &+ 4\pi(1-\pi)(1-\epsilon)\eta_\phi\{\mathbb{E}_U|\delta(U)| + L\eta_G\} + \eta_\phi^2. \end{aligned}$$

Finally, by Proposition 5.1,

$$\begin{aligned} R_2(\hat{\phi}_G^\epsilon) - R_2(\phi^\star) \le& \pi(1-\pi)(1-\epsilon)^2\mathbb{E}_U[\delta(U)^2] \\ &+ \pi(1-\pi)(1-\epsilon)^2\Big\{2L\eta_G\mathbb{E}_U|\delta(U)| + L^2\eta_G^2\Big\}. \end{aligned}$$

This completes the proof.

Lemma 1 (stability of cross-entropy predictor)**.** *Fix u and $\epsilon \in [0,1]$. For perturbations ξ_0, ξ_1, define*

$$\mu_0^\xi(u) := \mu_0(u) + \xi_0, \qquad \mu_1^\xi(u) := \mu_1(u) + \xi_1, \qquad \delta^\xi(u) := \mu_1^\xi(u) - \mu_0^\xi(u).$$

Let

$$(q_0^\epsilon(\xi_0, \xi_1; u), q_1^\epsilon(\xi_0, \xi_1; u))$$

be the cross-entropy predictors obtained by replacing $\mu_0(u), \mu_1(u)$ with $\mu_0^\xi(u), \mu_1^\xi(u)$, so that

$$q_1^\epsilon(\xi_0, \xi_1; u) - q_0^\epsilon(\xi_0, \xi_1; u) = \epsilon\delta^\xi(u).$$

Assume all probability inputs lie in $[\rho, 1-\rho]$. Then there exists a boundary-dependent factor C_ρ, depending only on ρ, such that

$$\begin{aligned} |q_0^\epsilon(0, \xi_1; u) - q_0^\epsilon(0, 0; u)| &\le C_\rho(1-\epsilon)|\xi_1|, \\ |q_1^\epsilon(\xi_0, 0; u) - q_1^\epsilon(0, 0; u)| &\le C_\rho(1-\epsilon)|\xi_0|. \end{aligned}$$

Moreover, for arbitrary perturbations ξ_0, ξ_1,

$$|q_a^\epsilon(\xi_0, \xi_1; u) - q_a^\epsilon(0, 0; u)| \le C_\rho(|\xi_0| + |\xi_1|), \qquad a \in \{0,1\}.$$

Proof. We prove the stability bound under the perturbation from counterfactual generator. We start with the state $A = 0$, where μ_0 is known. Let

$$\mu_{1,t}(u) := \mu_1(u) + t\xi_1, \qquad \delta_t(u) := \mu_{1,t}(u) - \mu_0(u), \qquad t \in [0,1],$$

and let $q(t)$ be the minimizer based on the inputs $\mu_0(u)$ and $\mu_{1,t}(u)$. Then

$$q(0) = q_0^\epsilon(0,0;u), \qquad q(1) = q_0^\epsilon(0,\xi_1;u).$$

The equation is

$$H(q(t),t) = 0,$$

where

$$\begin{aligned} H(q,t) :=& (1-\pi)\frac{q-\mu_0(u)}{q(1-q)} \\ &+ \pi\frac{q+\epsilon\delta_t(u)-\mu_{1,t}(u)}{\{q+\epsilon\delta_t(u)\}\{1-q-\epsilon\delta_t(u)\}}. \end{aligned}$$

Let

$$H_q(q,t) := \frac{\partial H}{\partial q}(q,t), \qquad H_t(q,t) := \frac{\partial H}{\partial t}(q,t).$$

Since $H(q(t),t)$ is identically zero as a function of t, the chain rule gives

$$0 = \frac{d}{dt}H(q(t),t) = H_q(q(t),t)q'(t) + H_t(q(t),t).$$

Hence

$$q'(t) = -\frac{H_t(q(t),t)}{H_q(q(t),t)}.$$

We now bound the denominator. Let

$$r(t) := q(t) + \epsilon\delta_t(u).$$

A direct calculation gives

$$\begin{aligned} H_q(q(t),t) =& (1-\pi)\left\{\frac{\mu_0(u)}{q(t)^2} + \frac{1-\mu_0(u)}{(1-q(t))^2}\right\} \\ &+ \pi\left\{\frac{\mu_{1,t}(u)}{r(t)^2} + \frac{1-\mu_{1,t}(u)}{(1-r(t))^2}\right\}. \end{aligned}$$

For any $s \in (0,1)$ and any $\mu \in [0,1]$,

$$\frac{\mu}{s^2} + \frac{1-\mu}{(1-s)^2} \geq \mu + (1-\mu) = 1.$$

Therefore,

$$H_q(q(t),t) \geq (1-\pi) + \pi = 1.$$

Thus the denominator is bounded away from zero.

We now bound the numerator. A direct calculation gives

$$H_t(q(t),t) = \pi\xi_1\left[\epsilon\ell''_{\mathrm{CE}}(r(t),\mu_{1,t}(u)) - \frac{1}{r(t)(1-r(t))}\right].$$

Using

$$\ell''_{\mathrm{CE}}(r,\mu) = \frac{\mu}{r^2} + \frac{1-\mu}{(1-r)^2},$$

we have the identity

$$\ell''_{\mathrm{CE}}(r,\mu) - \frac{1}{r(1-r)} = \frac{(\mu-r)(1-2r)}{r^2(1-r)^2}.$$

Hence

$$\begin{aligned} &\epsilon\ell''_{\mathrm{CE}}(r(t),\mu_{1,t}(u)) - \frac{1}{r(t)(1-r(t))} \\ &\quad = -(1-\epsilon)\frac{1}{r(t)(1-r(t))} + \epsilon\frac{(\mu_{1,t}(u)-r(t))(1-2r(t))}{r(t)^2(1-r(t))^2}. \end{aligned}$$

Because all relevant probabilities lie in $[\rho, 1-\rho]$,

$$\frac{1}{r(t)(1-r(t))} \le \frac{1}{\rho^2}, \qquad \frac{1}{r(t)^2(1-r(t))^2} \le \frac{1}{\rho^4}.$$

Moreover, applying the distance-to-Bayes argument above to the perturbed inputs $\mu_0(u)$ and $\mu_{1,t}(u)$ gives

$$|\mu_{1,t}(u) - r(t)| \le (1-\epsilon)|\mu_{1,t}(u) - \mu_0(u)| \le 1-\epsilon.$$

Therefore,

$$|H_t(q(t), t)| \le C_\rho(1-\epsilon)|\xi_1|.$$

Combining this with $H_q(q(t), t) \ge 1$ gives

$$|q'(t)| \le C_\rho(1-\epsilon)|\xi_1|.$$

Integrating from 0 to 1, we obtain

$$|q_0^\epsilon(0, \xi_1; u) - q_0^\epsilon(0, 0; u)| = |q(1) - q(0)| \le C_\rho(1-\epsilon)|\xi_1|.$$

The proof of

$$|q_1^\epsilon(\xi_0, 0; u) - q_1^\epsilon(0, 0; u)| \le C_\rho(1-\epsilon)|\xi_0|$$

is symmetric, with considering the state $A = 1$.

Finally, consider both perturbations ξ_0, ξ_1,

$$\mu_{0,t}(u) = \mu_0(u) + t\xi_0, \qquad \mu_{1,t}(u) = \mu_1(u) + t\xi_1,$$

the same implicit-differentiation argument gives

$$|q_0'(t)| \le C_\rho(|\xi_0| + |\xi_1|),$$

because the numerator derivative is bounded above by $C_\rho(|\xi_0| + |\xi_1|)$ and the denominator remains bounded below by 1. Note that factor C_ρ could be different from the one-side perturbation setting before. Integrating over $t \in [0, 1]$ yields

$$|q_0^\epsilon(\xi_0, \xi_1; u) - q_0^\epsilon(0, 0; u)| \le C_\rho(|\xi_0| + |\xi_1|).$$

The same bound for q_1^ϵ follows from the constraint

$$q_1^\epsilon(\xi_0, \xi_1; u) - q_0^\epsilon(\xi_0, \xi_1; u) = \epsilon\{(\mu_1(u) + \xi_1) - (\mu_0(u) + \xi_0)\}.$$

The proof is symmetric. □

A.6 Proof of Proposition 5.3

Let $\hat{\phi}_G^\epsilon$ denote the predictor which uses the estimated generator $\hat{G}$ but the true Bayes prediction function μ. For a latent individual u, define the counterfactual reconstruction errors

$$\zeta_1(u) := \mu(\hat{G}(x_0(u), 0, 1), 1) - \mu_1(u),$$

and

$$\zeta_0(u) := \mu(\hat{G}(x_1(u), 1, 0), 0) - \mu_0(u).$$

By the Lipschitz condition and the uniform generator error bound,

$$|\zeta_1(u)| \le L\eta_G, \qquad |\zeta_0(u)| \le L\eta_G.$$

For the observed state $A = 0$, the estimation uses the probability inputs

$$\mu_0(u), \quad \mu_1(u) + \zeta_1(u).$$

Therefore, by Lemma 1,

$$|\hat{\phi}_{G,0}^\epsilon(u) - \phi_0^\epsilon(u)| \le C_\rho(1-\epsilon)|\zeta_1(u)| \le C_\rho(1-\epsilon)L\eta_G.$$

Similarly, for the observed state $A = 1$, the estimation uses

$$\mu_0(u) + \zeta_0(u), \quad \mu_1(u),$$

and Lemma 1 gives

$$|\hat{\phi}^{\epsilon}_{G,1}(u) - \phi^{\epsilon}_1(u)| \le C_\rho(1-\epsilon)|\zeta_0(u)| \le C_\rho(1-\epsilon)L\eta_G.$$

Define

$$d_a(u) := \hat{\phi}^{\epsilon}_{G,a}(u) - \phi^{\epsilon}_a(u), \qquad a \in \{0,1\}.$$

Then

$$|d_a(u)| \le C_\rho(1-\epsilon)L\eta_G.$$

By Taylor's theorem and Assumption 2,

$$\begin{aligned}&\ell_{\mathrm{CE}}(\hat{\phi}^{\epsilon}_{G,a}(u), \mu_a(u)) - \ell_{\mathrm{CE}}(\phi^{\epsilon}_a(u), \mu_a(u)) \\ &\quad\le \left|\frac{\phi^{\epsilon}_a(u) - \mu_a(u)}{\phi^{\epsilon}_a(u)(1-\phi^{\epsilon}_a(u))}\right| |d_a(u)| + \frac{1}{2\rho^2} d_a(u)^2.\end{aligned}$$

By Lemma 1,

$$|\phi^{\epsilon}_a(u) - \mu_a(u)| \le (1-\epsilon)|\delta(u)|.$$

By Assumption 2,

$$\phi^{\epsilon}_a(u)(1-\phi^{\epsilon}_a(u)) \ge \rho^2.$$

Combining them gives

$$\ell_{\mathrm{CE}}(\hat{\phi}^{\epsilon}_{G,a}(u), \mu_a(u)) - \ell_{\mathrm{CE}}(\phi^{\epsilon}_a(u), \mu_a(u)) \le C_\rho(1-\epsilon)^2\{L\eta_G|\delta(u)| + L^2\eta_G^2\}.$$

Taking the weighted average over $a \in \{0,1\}$ and expectation over U,

$$R_{\ell_{\mathrm{CE}}}(\hat{\phi}^{\epsilon}_G) - R_{\ell_{\mathrm{CE}}}(\phi^{\epsilon}) \le C_\rho(1-\epsilon)^2\Big\{L\eta_G\mathbb{E}_U|\delta(U)| + L^2\eta_G^2\Big\}.$$

Finally,

$$R_{\ell_{\mathrm{CE}}}(\hat{\phi}^{\epsilon}_G) - R_{\ell_{\mathrm{CE}}}(\phi^{\star}) \le R_{\ell_{\mathrm{CE}}}(\phi^{\epsilon}) - R_{\ell_{\mathrm{CE}}}(\phi^{\star}) + C_\rho(1-\epsilon)^2\Big\{L\eta_G\mathbb{E}_U|\delta(U)| + L^2\eta_G^2\Big\}.$$

A.7 Proof of Proposition 5.4

Let $\hat{\phi}^{\epsilon}_G$ denote the estimated predictor in Proposition 5.3, which uses the estimated generator $\hat{G}$ but the true Bayes prediction function μ. Let $\hat{\phi}^{\epsilon}$ denote the fully estimated predictor, which uses both $\hat{G}$ and $\hat{\mu}$.

For the factual evaluations, define

$$\alpha_0(u) := \hat{\mu}(x_0(u), 0) - \mu_0(u), \qquad \alpha_1(u) := \hat{\mu}(x_1(u), 1) - \mu_1(u).$$

For the reconstructed counterfactual evaluations, define

$$\beta_1(u) := \hat{\mu}(\hat{G}(x_0(u), 0, 1), 1) - \mu(\hat{G}(x_0(u), 0, 1), 1),$$

and

$$\beta_0(u) := \hat{\mu}(\hat{G}(x_1(u), 1, 0), 0) - \mu(\hat{G}(x_1(u), 1, 0), 0).$$

By the uniform estimation bound on $\hat{\mu}$,

$$|\alpha_0(u)|, |\alpha_1(u)|, |\beta_0(u)|, |\beta_1(u)| \le \eta_\phi.$$

We first compare $\hat{\phi}^{\epsilon}$ with $\hat{\phi}^{\epsilon}_G$. For the state $A = 0$, the probability inputs change from

$$\mu_0(u), \qquad \mu_1(u) + \zeta_1(u)$$

to

$$\mu_0(u) + \alpha_0(u), \qquad \mu_1(u) + \zeta_1(u) + \beta_1(u).$$

By Lemma 1,

$$|\hat{\phi}^{\epsilon}_0(u) - \hat{\phi}^{\epsilon}_{G,0}(u)| \le C_\rho\eta_\phi.$$

Similarly,

$$|\hat{\phi}^{\epsilon}_1(u) - \hat{\phi}^{\epsilon}_{G,1}(u)| \le C_\rho\eta_\phi.$$

Define

$$d_a(u) := \hat{\phi}^{\epsilon}_a(u) - \hat{\phi}^{\epsilon}_{G,a}(u), \qquad a \in \{0,1\}.$$

Then

$$|d_a(u)| \le C_\rho \eta_\phi.$$

Next we control the derivative of the true cross-entropy risk at the intermediate predictor. From the proof of Proposition 5.3,

$$|\hat{\phi}^{\epsilon}_{G,a}(u) - \phi^{\epsilon}_a(u)| \le C_\rho(1-\epsilon)L\eta_G.$$

Since

$$|\phi^{\epsilon}_a(u) - \mu_a(u)| \le (1-\epsilon)|\delta(u)|.$$

Therefore,

$$|\hat{\phi}^{\epsilon}_{G,a}(u) - \mu_a(u)| \le C_\rho(1-\epsilon)\{|\delta(u)| + L\eta_G\}.$$

By Assumption 2,

$$\left|\frac{\hat{\phi}^{\epsilon}_{G,a}(u) - \mu_a(u)}{\hat{\phi}^{\epsilon}_{G,a}(u)(1-\hat{\phi}^{\epsilon}_{G,a}(u))}\right| \le C_\rho(1-\epsilon)\{|\delta(u)| + L\eta_G\}.$$

Similar to the proof of Proposition 5.3, with Taylor's theorem and Assumption 2, we further have

$$\begin{aligned}&\ell_{\mathrm{CE}}(\hat{\phi}^{\epsilon}_a(u), \mu_a(u)) - \ell_{\mathrm{CE}}(\hat{\phi}^{\epsilon}_{G,a}(u), \mu_a(u)) \\ &\quad\le C_\rho\Big[(1-\epsilon)\{|\delta(u)| + L\eta_G\}\eta_\phi + \eta_\phi^2\Big].\end{aligned}$$

Taking the weighted average over $a \in \{0,1\}$ and averaging over U, we get

$$R_{\ell_{\mathrm{CE}}}(\hat{\phi}^{\epsilon}) - R_{\ell_{\mathrm{CE}}}(\hat{\phi}^{\epsilon}_G) \le C_\rho\Big[(1-\epsilon)\eta_\phi\{\mathbb{E}_U|\delta(U)| + L\eta_G\} + \eta_\phi^2\Big].$$

Finally,

$$R_{\ell_{\mathrm{CE}}}(\hat{\phi}^{\epsilon}) - R_{\ell_{\mathrm{CE}}}(\phi^\star) = \{R_{\ell_{\mathrm{CE}}}(\hat{\phi}^{\epsilon}_G) - R_{\ell_{\mathrm{CE}}}(\phi^\star)\} + \{R_{\ell_{\mathrm{CE}}}(\hat{\phi}^{\epsilon}) - R_{\ell_{\mathrm{CE}}}(\hat{\phi}^{\epsilon}_G)\}.$$

Applying Proposition 5.3, we obtain

$$\begin{aligned}R_{\ell_{\mathrm{CE}}}(\hat{\phi}^{\epsilon}) - R_{\ell_{\mathrm{CE}}}(\phi^\star) \le& R_{\ell_{\mathrm{CE}}}(\phi^{\epsilon}) - R_{\ell_{\mathrm{CE}}}(\phi^\star) \\ &+ C_\rho(1-\epsilon)^2\Big\{L\eta_G\mathbb{E}_U|\delta(U)| + L^2\eta_G^2\Big\} \\ &+ C_\rho\Big[(1-\epsilon)\eta_\phi\{\mathbb{E}_U|\delta(U)| + L\eta_G\} + \eta_\phi^2\Big].\end{aligned}$$

B Experimental details and additional results

B.1 Synthetic experiments with misspecified counterfactual generator

We use the structural models in Section 6.1 and $\epsilon \in \{0, 0.1, \ldots, 1\}$. In each of 30 repetitions, we generate 1,000 independent observations and evaluate every (ϵ, η) combination on the same observations.

Tables 1 and 2 retain the numerical results underlying Figure 1. Each empirical entry is the mean (sample standard deviation) of the repetitions.

Table 1: Regression empirical excess risk: mean (standard deviation) over repetitions.

ϵ	Oracle	$\eta = 0$	$\eta = 0.10$	$\eta = 0.25$	$\eta = 0.50$
0.0	0.5833	0.5893 (0.0239)	0.7736 (0.0325)	1.1169 (0.0487)	1.9031 (0.0856)
0.1	0.4725	0.4775 (0.0192)	0.6268 (0.0261)	0.9049 (0.0391)	1.5418 (0.0689)
0.2	0.3733	0.3774 (0.0150)	0.4954 (0.0204)	0.7152 (0.0306)	1.2185 (0.0539)
0.3	0.2858	0.2891 (0.0113)	0.3795 (0.0154)	0.5478 (0.0231)	0.9332 (0.0408)
0.4	0.2100	0.2125 (0.0081)	0.2789 (0.0111)	0.4026 (0.0166)	0.6859 (0.0295)
0.5	0.1458	0.1477 (0.0055)	0.1939 (0.0075)	0.2798 (0.0113)	0.4766 (0.0200)
0.6	0.0933	0.0947 (0.0033)	0.1242 (0.0046)	0.1793 (0.0069)	0.3052 (0.0124)
0.7	0.0525	0.0534 (0.0018)	0.0700 (0.0024)	0.1010 (0.0036)	0.1719 (0.0066)
0.8	0.0233	0.0238 (0.0007)	0.0312 (0.0009)	0.0450 (0.0014)	0.0766 (0.0026)
0.9	0.0058	0.0060 (0.0002)	0.0079 (0.0002)	0.0114 (0.0003)	0.0193 (0.0005)
1.0	0.0000	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)

Table 2: Classification empirical excess risk: mean (standard deviation) over repetitions.

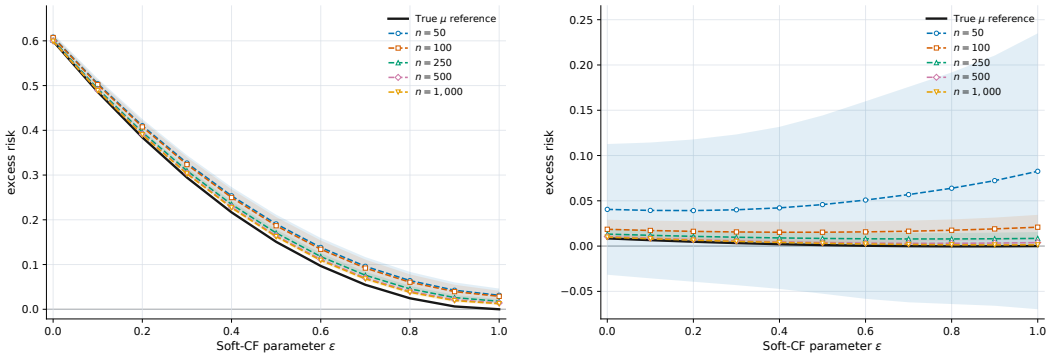
ϵ	Oracle	$\eta = 0$	$\eta = 0.10$	$\eta = 0.25$	$\eta = 0.50$
0.0	0.0132	0.0167 (0.0050)	0.0207 (0.0056)	0.0271 (0.0066)	0.0388 (0.0081)
0.1	0.0107	0.0139 (0.0046)	0.0172 (0.0051)	0.0228 (0.0060)	0.0328 (0.0075)
0.2	0.0085	0.0113 (0.0041)	0.0141 (0.0046)	0.0187 (0.0055)	0.0273 (0.0069)
0.3	0.0065	0.0090 (0.0036)	0.0112 (0.0041)	0.0150 (0.0049)	0.0221 (0.0062)
0.4	0.0048	0.0070 (0.0031)	0.0087 (0.0036)	0.0116 (0.0043)	0.0173 (0.0055)
0.5	0.0034	0.0052 (0.0026)	0.0064 (0.0030)	0.0086 (0.0036)	0.0129 (0.0047)
0.6	0.0022	0.0036 (0.0021)	0.0045 (0.0024)	0.0060 (0.0030)	0.0090 (0.0039)
0.7	0.0012	0.0023 (0.0016)	0.0028 (0.0018)	0.0037 (0.0022)	0.0056 (0.0030)
0.8	0.0005	0.0013 (0.0011)	0.0015 (0.0012)	0.0020 (0.0015)	0.0029 (0.0020)
0.9	0.0001	0.0005 (0.0005)	0.0006 (0.0006)	0.0007 (0.0008)	0.0010 (0.0010)
1.0	0.0000	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)

B.2 Synthetic experiments with estimated Bayes optimal prediction

For each training size $n \in \{50, 100, 250, 500, 1{,}000\}$, we fit $\hat{\mu}$ on training samples. An independent evaluation set of 1,000 observations is held fixed across training sizes, repetitions, and ϵ values. The true μ and evaluation responses are not used for fitting. The same fitted network supplies factual and counterfactual predictions for the entire ϵ path.

Both tasks use multilayer perceptron (MLP) with hidden layer $(20, 20)$ and tanh activations. The inputs are raw (X, A), without an X^2 feature, latent U, or additional standardization. Regression has a linear output and classification uses the class-1 probability.

The result is shown by Figure 3 and Tables 3 and 4. As n increases from 50 to 1,000, the mean endpoint excess risk decreases from 0.0313 to 0.0124 for regression and from 0.0826 to 0.0014 for classification. Unlike generator error alone, this cost need not vanish at the unrestricted endpoint, as reflected in Propositions 5.2 and 5.4. The small-sample classification curves have substantial variability.



(a) Regression: squared loss.

(b) Classification: cross-entropy.

Figure 3: Excess risk with estimated $\hat{\mu}$. Colored curves show means over 30 independently fitted networks; shaded bands show one sample standard deviation. The black "True μ reference" evaluates the oracle trade-off predictor on the same fixed test observations and labels.

Table 3: Estimated μ, regression: mean (standard deviation) over 30 training samples.

ϵ	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 1{,}000$
0.0	0.6083 (0.0082)	0.6073 (0.0062)	0.6014 (0.0023)	0.5998 (0.0012)	0.5991 (0.0007)
0.1	0.5042 (0.0117)	0.5022 (0.0101)	0.4930 (0.0037)	0.4899 (0.0020)	0.4888 (0.0015)
0.2	0.4103 (0.0150)	0.4076 (0.0136)	0.3957 (0.0052)	0.3915 (0.0031)	0.3900 (0.0025)
0.3	0.3268 (0.0177)	0.3234 (0.0164)	0.3095 (0.0065)	0.3044 (0.0039)	0.3027 (0.0034)
0.4	0.2537 (0.0196)	0.2498 (0.0184)	0.2345 (0.0074)	0.2287 (0.0046)	0.2268 (0.0040)
0.5	0.1908 (0.0208)	0.1867 (0.0195)	0.1706 (0.0079)	0.1645 (0.0049)	0.1624 (0.0045)
0.6	0.1382 (0.0211)	0.1342 (0.0197)	0.1178 (0.0080)	0.1116 (0.0051)	0.1095 (0.0047)
0.7	0.0960 (0.0207)	0.0921 (0.0190)	0.0761 (0.0077)	0.0701 (0.0049)	0.0680 (0.0046)
0.8	0.0641 (0.0195)	0.0605 (0.0175)	0.0455 (0.0070)	0.0401 (0.0046)	0.0380 (0.0044)
0.9	0.0425 (0.0176)	0.0395 (0.0153)	0.0261 (0.0060)	0.0214 (0.0041)	0.0195 (0.0041)
1.0	0.0313 (0.0151)	0.0290 (0.0124)	0.0178 (0.0048)	0.0142 (0.0035)	0.0124 (0.0036)

Table 4: Estimated μ, classification: mean (standard deviation) over 30 training samples.

ϵ	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 1{,}000$
0.0	0.0405 (0.0722)	0.0185 (0.0106)	0.0132 (0.0060)	0.0105 (0.0030)	0.0099 (0.0023)
0.1	0.0394 (0.0750)	0.0172 (0.0109)	0.0119 (0.0060)	0.0087 (0.0032)	0.0081 (0.0023)
0.2	0.0392 (0.0786)	0.0162 (0.0112)	0.0108 (0.0062)	0.0071 (0.0034)	0.0065 (0.0024)
0.3	0.0401 (0.0832)	0.0156 (0.0115)	0.0098 (0.0063)	0.0058 (0.0035)	0.0052 (0.0024)
0.4	0.0422 (0.0895)	0.0153 (0.0116)	0.0090 (0.0064)	0.0047 (0.0037)	0.0040 (0.0025)
0.5	0.0458 (0.0984)	0.0153 (0.0117)	0.0085 (0.0064)	0.0039 (0.0038)	0.0031 (0.0025)
0.6	0.0508 (0.1091)	0.0157 (0.0117)	0.0081 (0.0065)	0.0033 (0.0039)	0.0023 (0.0026)
0.7	0.0569 (0.1190)	0.0164 (0.0117)	0.0079 (0.0065)	0.0031 (0.0039)	0.0018 (0.0026)
0.8	0.0639 (0.1279)	0.0175 (0.0119)	0.0079 (0.0066)	0.0031 (0.0041)	0.0014 (0.0027)
0.9	0.0723 (0.1382)	0.0190 (0.0125)	0.0081 (0.0067)	0.0033 (0.0044)	0.0013 (0.0028)
1.0	0.0826 (0.1524)	0.0209 (0.0136)	0.0085 (0.0071)	0.0039 (0.0049)	0.0014 (0.0029)

B.3 SEMI-SYNTHETIC DATA GENERATION

We follow the generative-model-based construction in Zhou et al. (2024), using VAE-based mechanisms fitted to real data to obtain paired factual and counterfactual observations. The data-generation stage is distinct from the counterfactual-estimation stage used by the prediction methods.

Law School. The sensitive attribute is gender, the target is first-year grade, and the remaining variables include race, LSAT score, and undergraduate GPA. Following the construction in Zhou et al. (2024), the VAE has the structure

$$u \sim \mathrm{Enc}(x, a), \qquad x \sim \mathrm{Dec}_1(u, a), \qquad y \sim \mathrm{Dec}_2(u, x).$$

Once fitted, the model generates observations by independently sampling $u \sim N(0, I)$ and $a \sim \mathrm{Bernoulli}(\pi_{\mathrm{data}})$, where π_{data} is the empirical group proportion in the original data, and passing them through the decoders. The paired counterfactual is obtained under the same latent state with the sensitive attribute changed to $1 - a$. This construction supplies ground-truth counterfactuals for evaluation without making them available to the downstream fitting procedure.

Adult. The target indicates whether annual income exceeds \$50,000, and the sensitive attribute is gender. We follow the DCEVAE-based construction in Zhou et al. (2024), with the preprocessing and model settings adopted there from Zuo et al. (2023). DCEVAE Kim et al. (2021) uses separate latent components for features with different causal relationships to the sensitive attribute. One fitted DCEVAE defines the semi-synthetic data-generating mechanism, while a separate DCEVAE estimates counterfactuals from factual observations. The generating model's encoder includes the outcome, whereas the estimated generator's encoder does not. The ground-truth latent variables and paired counterfactuals are withheld from downstream fitting.

Prediction models. For Law School, $\hat{\mu}$ is a multilayer perceptron (MLP) regressor with hidden layers $(20, 20)$ and tanh activations. For Adult, it is an MLP classifier with one hidden layer of width 100 and ReLU activations. Each of the five repetitions uses a newly fitted MLP and the corresponding fitted counterfactual estimator.

B.4 COMPARED METHODS

L1/L2 counterfactual regularization. For a regularization parameter $\lambda \geq 0$, the regularized predictions solve

$$\min_{z_0, z_1} (1 - \pi)\ell(z_0, \hat{\mu}_0) + \pi\ell(z_1, \hat{\mu}_1) + \lambda|z_1 - z_0|^p, \qquad p \in \{1, 2\}. \tag{4}$$

We use $\lambda = 10^{-6+0.1k}$ for $k = 0, \ldots, 120$ in evaluation. Under squared loss, L2 coincides with the proposed predictor with

$$\lambda = \pi(1 - \pi)\frac{1 - \epsilon}{\epsilon}, \qquad 0 < \epsilon \leq 1,$$

Counterfactual Gap Shrinkage. This method considers the gap shrinkage:

$$m = \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}, \qquad z_a = m + \tau(\hat{\mu}_a - m), \qquad \tau \in [0, 1].$$

We use $\tau \in \{0, 0.01, \ldots, 1\}$ in evaluation.