Beyond Bounds: Quantifying the Probability of Counterfactual Fairness

Taehan Kim[2] **Minyoung Cho**[1] **Sanghack Lee***[1]

[1]Graduate School of Data Science, Seoul National University, Seoul, Republic of Korea
[2]Korean National Police Agency, Republic of Korea

Abstract

Counterfactual fairness is a rigorous criterion for algorithmic decision-making but remains fundamentally unidentifiable from observational data alone. Existing partial identification methods address this by deriving bounds for fairness measures; however, these intervals are often wide, limiting their practical utility. To address this limitation, we propose a framework that quantifies the probability that a black-box model satisfies counterfactual fairness, under a stated prior over the structural causal models compatible with the observed data. By exploiting conditional independencies to reduce the parameter space, we demonstrate that the set of causal parameters compatible with the observed data forms a convex polytope. We further show that, given domain-specific priors on exogenous distributions, this prior-dependent probability can be estimated via Hit-and-Run Monte Carlo integration. Our approach provides a practical tool that complements worst-case bounds, offering a prior-dependent probabilistic summary for auditing model fairness.

1 INTRODUCTION

Machine learning techniques are increasingly applied in decision-making processes. Their widespread adoption has spurred a growing interest in ensuring that these systems make fair and transparent decisions [Caton and Haas, 2024]. Poorly designed models may inadvertently learn biases present in the training dataset, leading to unintended discrimination. These concerns about fairness in machine learning have prompted extensive research into defining [Dwork et al., 2012, Hardt et al., 2016, Kilbertus et al., 2017, Kusner et al., 2017], evaluating [Chouldechova, 2017, Speicher et al., 2018, Wu et al., 2019b, 2023], and ensuring fairness [Kamishima et al., 2012, Kleinberg et al., 2017, Corbett-Davies et al., 2023, Zemel et al., 2013, Zafar et al., 2017] in machine learning models.

Counterfactual fairness, introduced by Kusner et al. [2017], is a fairness criterion that formalizes discrimination within a causal framework. Unlike group-based metrics, it is an individual-level criterion [Silva, 2024] that compares the decision outcome of an individual's hypothetical 'what-if' situation with the observed factual outcome. Formally, a decision process is considered counterfactually fair if the outcome remains unchanged even when the sensitive attribute (such as gender, age, or race) of an individual has been hypothetically altered. For instance, a hiring decision made by an AI model is considered counterfactually fair with respect to the gender attribute if the outcome for a hired applicant would remain unchanged if the applicant's gender were different.

However, certifying the counterfactual fairness of opaque black-box models presents fundamental challenges because of the difficulty in inferring counterfactual quantities that are fundamentally unobservable. Indeed, *Pearl's Causal Hierarchy* [Pearl, 2009] establishes that many critical counterfactual distributions are theoretically unidentifiable, even with complete knowledge of observational and experimental distributions. In such cases, the counterfactual distribution is not uniquely determined. Consequently, establishing a single ground truth for fairness evaluation is generally not possible without further assumptions.

To address this unidentifiability, recent works have focused on partial counterfactual identification, which bounds counterfactual probabilities within a finite interval across all compatible Structural Causal Models (SCMs). Balke and Pearl [1994a] introduced a method for partial identification based on the discretization of exogenous variables in Instrumental Variable (IV) settings, which was later generalized to canonical SCMs by Zhang et al. [2022]. Building on this, Wu et al. [2019a,b] adopted discretization to derive bounds

*Corresponding author: `sanghack@snu.ac.kr`.

for counterfactual predictions, proposing these bounds as a measure for assessing counterfactual fairness. Related numerical, symbolic, and credal approaches to counterfactual bounding include Duarte et al. [2024], Sachs et al. [2023], and Zaffalon et al. [2024].

While these interval-based approaches provide theoretical bounds, they often yield ambiguous practical interpretations. Wide bounds can render the fairness assessment inconclusive, leaving decision-makers without a clear verdict. Motivated by this limitation, we propose a systematic framework to quantify the probability that a black-box model satisfies counterfactual fairness as a post-hoc auditing metric for a fixed predictor. Adopting the standard assumption of a known causal graph while allowing for arbitrary unobserved confounders [Wu et al., 2019b], we first parameterize the causal model using the canonical Structural Causal Model framework [Zhang et al., 2022] and identify the set of parameters that satisfies the fairness criterion. We show that the target probability can be estimated by Monte Carlo integration, incorporating domain-specific priors on the exogenous distribution. Our work complements interval-based approaches from prior studies by estimating a prior-dependent probability that a model is counterfactually fair.

Related Work Kusner et al. [2017] introduced the concept of counterfactual fairness and showed that the variables that are causally affected by the sensitive attribute are the primary factors hindering counterfactual fairness. Subsequent research has expanded on this work by decomposing counterfactual fairness into path-specific components [Zhang and Bareinboim, 2018], removing unfairness from data [Nabi and Shpitser, 2018], defining path-specific variations [Chiappa, 2019], or proposing new fairness notions based on counterfactuals [Coston et al., 2020, Mishler et al., 2021].

Additionally, several studies have focused on adapting data or generating counterfactual samples to train counterfactually fair models [Plečko and Meinshausen, 2020, Bothmann et al., 2023, Ma et al., 2022, 2023, Garg et al., 2019, Mostafazadeh Davani et al., 2021, Dinan et al., 2020, Maudslay et al., 2019, Cheong et al., 2023, Zuo et al., 2023], including path-specific probability-valued unfairness measures [Chikahara et al., 2021]. Our work is closely related to studies on the partial identification of counterfactual predictions. Wu et al. [2019a] pointed out that the unprotected attributes that causally affect the prediction cause unidentifiability of counterfactual quantities and proposed the counterfactual prediction bounds in cases without unobserved confounders. Later, Wu et al. [2019b] applied the response-variable framework from Balke and Pearl [1994a] to derive bounds for path-specific counterfactual predictions. Recent bounding frameworks further develop numerical, symbolic, and approximate bound computation for discrete counterfactual queries [Duarte et al., 2024, Sachs et al., 2023, Zaffalon et al., 2024]. Recently, Rho et al. [2025] proposed a Bayesian approach to bound counterfactual fairness measures. In contrast to training-time or posterior-bound approaches, our framework reports a prior-dependent probability of counterfactual fairness over observationally compatible SCMs for a fixed predictor.

Crucially, while these prior works concentrate on delineating the range of possible counterfactual outcomes, they do not quantify the likelihood of fairness within those bounds. Our research complements this line of work by estimating a prior-dependent probability that a model satisfies counterfactual fairness, providing an additional summary beyond intervals alone.

Relation to partial-identification bounds Our framework differs from partial-identification methods in the *type* of object it returns. Numerical, symbolic, and credal bounding methods—Duarte et al. [2024] via polynomial programs, Sachs et al. [2023] via symbolic bounds reducible to linear programs, and Zaffalon et al. [2024] via credal-network inference—all return an interval of possible values for a counterfactual quantity, answering "what are the smallest and largest counterfactual values consistent with the data?". In contrast, we place a prior over the polytope of compatible SCMs and integrate (Lemma 4), returning a probability measure that answers "how likely is the predictor counterfactually fair under the stated prior?". The two views are complementary: when a worst-case bound is wide, a prior-dependent probability measure quantifies how the risk of counterfactual unfairness varies across the values within that bound under the stated prior, and therefore provides an additional, complementary summary.

Contributions In summary, our main contributions are threefold. First, we introduce a novel metric that quantifies the probability of a black-box model satisfying counterfactual fairness, moving beyond worst-case intervals. Second, we demonstrate that the feasible parameter space of compatible canonical SCMs forms a convex polytope, and by exploiting conditional independencies via C-parameterization, we reduce the dimensionality of the inference problem. Third, we translate these findings into an executable auditing scheme, showing that standard Hit-and-Run MCMC can evaluate the fairness probability over this polytope.

2 PRELIMINARIES

In this section, we review fundamental concepts and notations used throughout the paper.

Notation We denote random variables by uppercase letters (e.g., X) and their realizations by lowercase letters (e.g., x). Sets of variables are denoted by boldface uppercase letters (e.g., $\mathbf{X}$), and their cardinality by $|\mathbf{X}|$. The domain of a variable X is denoted by Ω_X and its cardinality by $|\Omega_X|$. For finite domains, we enumerate values as $\{x^{(1)}, \ldots, x^{(|\Omega_X|)}\}$. We use standard graph-theoretic terminology (parents, ancestors, descendants). The indicator function $\mathbb{I}(E)$ equals 1 if E is true, and 0 otherwise.

Structural Causal Models (SCMs) We adopt the SCM framework [Pearl, 2009]. An SCM is a tuple $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$, where $\mathbf{U}$ is a set of variables exogenous to the model, and $\mathbf{V}$ is a set of endogenous variables whose values are determined by other variables in the model. $\mathbf{F}$ is a set of deterministic functions $\{f_V\}_{V \in \mathbf{V}}$, where each value $v = f_V(\mathbf{pa}_V, \mathbf{u}_V)$ is determined by values of its endogenous parents $\mathbf{Pa}_V \subseteq \mathbf{V} \setminus \{V\}$ and exogenous parents $\mathbf{U}_V \subseteq \mathbf{U}$. For a set of variables $\mathbf{W} \subseteq \mathbf{V}$, we denote the set $\bigcup_{V \in \mathbf{W}} \mathbf{Pa}_V$ and $\bigcup_{V \in \mathbf{W}} \mathbf{U}_V$ as $\mathbf{Pa_W}$ and $\mathbf{U_W}$ respectively. The distribution $P(\mathbf{U})$ defines the stochasticity of the model. Each SCM induces an acyclic directed mixed graph $\mathcal{G}$ where a directed edge $V_i \to V_j$ exists if $V_i \in \mathbf{Pa}_{V_j}$, and a bidirected edge $V_i \leftrightarrow V_j$ exists if $\mathbf{U}_{V_i}$ and $\mathbf{U}_{V_j}$ are statistically dependent, i.e., share a common exogenous source (unobserved confounding). Throughout, we do not restrict $\mathcal{G}$ to be simple: bidirected edges between any pair of endogenous variables, including between an observed covariate and the deployed predictor, are allowed and are central to the auditing setting (see Figure 1).

Interventions and Counterfactuals The SCM framework provides systematic methods for inferring interventional and counterfactual distributions. An intervention on a subset of variables $\mathbf{X}$, denoted by $do(\mathbf{x})$, is defined as replacing the original causal mechanisms $X = f_X(\mathbf{Pa}_X, \mathbf{U}_X)$ for $X \in \mathbf{X}$ with fixed values $x \in \mathbf{x}$. This transforms the original causal model $\mathcal{M}$ into a modified submodel $\mathcal{M}_\mathbf{x}$. Graphically, the submodel $\mathcal{M}_\mathbf{x}$ induces subgraph $\mathcal{G}_{\overline{\mathbf{X}}}$, where all incoming edges onto $X \in \mathbf{X}$ are deleted and $\mathbf{X}$ is fixed to the value $\mathbf{x}$. For a specific context (or "unit") $\mathbf{u}$, the *potential response* of $\mathbf{Y}$ under intervention $do(\mathbf{x})$ is denoted as $\mathbf{Y_x}(\mathbf{u})$. The *counterfactual variable* $\mathbf{Y_x}$ is the random variable $\mathbf{Y}$ generated in $\mathcal{M}_\mathbf{x}$. The event $\mathbf{Y_x} = \mathbf{y}$ (simply $\mathbf{y_x}$) can be interpreted as "$\mathbf{Y}$ being $\mathbf{y}$ had $\mathbf{X}$ been $\mathbf{x}$". For $\mathbf{Y}, \mathbf{X}, \mathbf{Z} \subseteq \mathbf{V}$, the counterfactual probability $P(\mathbf{y_x}, \mathbf{x'_z})$ is defined as below:

$$P(\mathbf{y_x}, \mathbf{x'_z}) = \int_{\Omega_\mathbf{U}} \mathbb{I}\big(\mathbf{Y_x}(\mathbf{u}) = \mathbf{y}\big) \cdot \mathbb{I}\big(\mathbf{X_z}(\mathbf{u}) = \mathbf{x}'\big)\, \mathrm{d}P(\mathbf{u}).$$

For a comprehensive treatment of SCMs, see [Pearl, 2009].

Bidirected edges in $\mathcal{G}$ partition $\mathbf{V}$ into disjoint sets called *C-components* [Tian and Pearl, 2002].

Definition 1 (C-component)**.** *For an arbitrary causal graph $\mathcal{G}$, a C-component $\mathbf{C}$ is the set of endogenous variables $\mathbf{C} \subseteq \mathbf{V}$ such that for any pair of variables $V_i, V_j \in \mathbf{C}$, V_i and V_j are connected by a bidirected path. A C-component $\mathbf{C}$ is maximal if there exists no other C-component that includes $\mathbf{C}$.*

Tian and Pearl [2002] introduced a factorization of probabilities based on C-components.

Definition 2 (Q-factor)**.** *For a given causal graph $\mathcal{G}$, observational distribution $P(\mathbf{V})$ and a set of endogenous variables $\mathbf{W}$, a Q-factor $Q[\mathbf{W}](\mathbf{w}, \mathbf{pa_W})$ is defined as follows:*

$$Q[\mathbf{W}](\mathbf{w}, \mathbf{pa_W}) = \sum_{\mathbf{u_W}} P(\mathbf{u_W}) \prod_{V_i \in \mathbf{W}} \mathbb{I}(f_{V_i}(\mathbf{pa}_{V_i}, \mathbf{u}_{V_i}) = v_i).$$

In this paper, we will shorten $Q[\mathbf{W}](\mathbf{w}, \mathbf{pa_W})$ to $Q[\mathbf{W}]$ for readability. Note that $Q[\mathbf{W}] = P(\mathbf{w} \mid do(\mathbf{pa_W}))$ by definition. Tian and Pearl [2002] showed that the observational distribution $P(\mathbf{V})$ can be decomposed into products of Q-factors of C-components, and those Q-factors can also be expressed as a product of conditional probabilities.

Lemma 1 (Q-factorization)**.** *Let $\{\mathbf{C}_1, \ldots \mathbf{C}_m\}$ be the set of all maximal C-components and let a topological order over $\mathbf{V}$ be $V_1 < \ldots < V_n$. Let $\mathbf{V}^{(i-1)}$ denote the set of variables preceding V_i in the topological order. Then,*

$$P(\mathbf{v}) = \prod_{i=1}^{m} Q[\mathbf{C}_i],$$
$$Q[\mathbf{C}_i] = \prod_{V_j \in \mathbf{C}_i} P(v_j \mid \mathbf{v}^{(j-1)}).$$

Canonical SCM and Partial Identification Counterfactual queries are often *non-identifiable* from observational data. Balke and Pearl [1994a] addressed this by discretizing exogenous variables according to their functional impact on endogenous variables, thereby bounding counterfactual probabilities. Zhang et al. [2022] generalized this approach to arbitrary causal graphs with discrete endogenous variables, introducing the *Canonical SCM*.

Definition 3 (Canonical Structural Causal Model)**.** *An SCM $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ is said to be a canonical SCM if*

1. *For every endogenous $V \in \mathbf{V}$, its values are contained in a discrete finite domain Ω_V. For every exogenous $U_V \in \mathbf{U}$, its values are contained in the domain $\Omega_{U_V} = \{u_V^{(1)}, \ldots, u_V^{(m_V)}\}$ where $m_V = |\Omega_V|^{|\Omega_{\mathbf{Pa}_V}|}$*
2. *For every endogenous $V \in \mathbf{V}$, its value is given as $v = g_V^{(u_V)}(\mathbf{pa}_V)$ where $g_V^{(u_V)} \in (\Omega_{\mathbf{Pa}_V} \mapsto \Omega_V)$.*

Zhang et al. [2022] proved that Canonical SCMs can represent *all* counterfactual distributions in discrete settings. Consequently, computing the tightest bounds for any counterfactual probability can be formulated as a polynomial optimization problem over the finite parameters of the Canonical SCM.

Counterfactual Fairness One prominent application of such counterfactual reasoning lies in algorithmic fairness. Kusner et al. [2017] first proposed the concept of a *counterfactually fair predictor*. This notion requires that the predictor's outcome $\hat{Y}$ remains unchanged in a counterfactual scenario where the sensitive attribute S is altered, given the observed features. For simplicity, throughout this paper, we assume S to be binary. Let $\mathbf{X} := \mathbf{V} \setminus \{S, \hat{Y}\}$ denote the observed non-sensitive covariates used in the fairness query. The realization $\mathbf{x}$ in Definition 4 is the full individual-level feature vector, not a strict subset.

Definition 4 (Counterfactually fair predictor). *Given a factual observation $\mathbf{X} = \mathbf{x}$ and $S = s$ with $P(\mathbf{x}, s) > 0$, the* ***counterfactual effect*** *of an intervention $S = \neg s$ on the predictor $\hat{Y} = \hat{y}$ is defined as follows:*

$$\mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y}) = P(\hat{y}_{\neg s} \mid \mathbf{x}, s) - P(\hat{y} \mid \mathbf{x}, s).$$

The predictor $\hat{Y}$ is said to be counterfactually fair if $\mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y}) = 0$ for all $\hat{y}, \mathbf{x}, s$ with $P(\mathbf{x}, s) > 0$.

Achieving exact counterfactual fairness requires the factual prediction probability $P(\hat{y} \mid \mathbf{x}, s)$ to be identical to the counterfactual probability $P(\hat{y}_{\neg s} \mid \mathbf{x}, s)$, which is often practically unachievable. To address this, Wu et al. [2019b] introduced τ-counterfactual fairness, a relaxed version that allows the probability of the prediction outcome $\hat{Y}$ to deviate from perfect counterfactual fairness by at most τ.

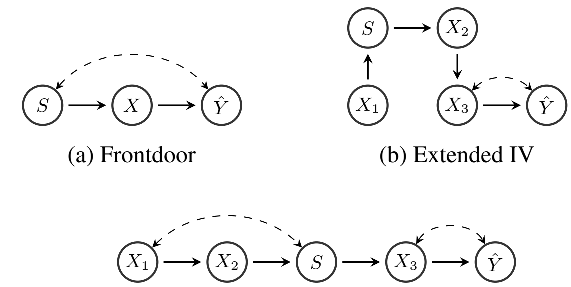
Definition 5 (τ-Counterfactually fair predictor). *A predictor $\hat{Y}$ is said to be τ-counterfactually fair if for all $\mathbf{x}, \hat{y}, s$ with $P(\mathbf{x}, s) > 0$:*

$$|\mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y})| \leq \tau.$$

Positivity Throughout, we adopt a strict positivity assumption on the observed covariates and sensitive attribute: $P(\mathbf{x}, s) > 0$ for every $\mathbf{x} \in \Omega_{\mathbf{X}}$ and $s \in \Omega_S$. This ensures that the conditional distributions appearing in our observational constraints are well defined and that every counterfactual query conditions on an event of positive probability. The assumption places no restriction on the predictor mechanism $\hat{Y}$, which may be randomized or deterministic.

In our setting, we allow bidirected edges involving the predictor $\hat{Y}$. Such edges arise from unobserved dependence between the deployed predictor and observed variables, such as unmeasured shared features or training-sample selection.

Example 1. *Consider the data-generating process in Figure 1a, where a deployed model predicts the employment success ($\hat{Y}$) of an applicant. Suppose the deployed model is trained on a richer profile of the applicant, while the auditor observes only a coarser set of variables, namely job-fitness (X) and education level (S). Unobserved residual factors (e.g., age, sex) may then influence both S and $\hat{Y}$, so $P(u_S, u_{\hat{Y}}) \neq P(u_S)\, P(u_{\hat{Y}})$. This unobserved dependence is represented by the bidirected edge $S \leftrightarrow \hat{Y}$. The $\mathbf{X} \leftrightarrow \hat{Y}$ patterns in Figure 1b and Figure 1c arise analogously.*



(c) Doubly-confounded chain

Figure 1: Causal diagrams of structural causal models of predictor $\hat{Y}$. Dashed bidirected edges encode unobserved confounding, including unobserved dependence involving the deployed predictor.

In this paper, we mainly focus on the fairness concepts in Definition 4 and Definition 5, but our proposed framework can be expanded to path-specific variations of counterfactual fairness [Wu et al., 2019b, Chiappa, 2019, Zhang and Bareinboim, 2018].

3 COUNTERFACTUAL PREDICTION PROBABILITY

In this section, we present a systematic framework to quantify the probability that a black-box model satisfies counterfactual fairness. Since the true causal mechanism is unidentifiable, relying solely on worst-case bounds ignores the relative plausibility of different causal models. To overcome this, we must rigorously define the space of all causal models compatible with the observed data and integrate domain-specific prior beliefs over this space.

This approach entails two main challenges: representational efficiency and geometric characterization. To address these challenges, we introduce a **C-parameterization** method in Section 3.1 to efficiently represent canonical SCMs by exploiting conditional independencies in the causal graph. Next, in Section 3.2, we investigate the geometry of the feasible model space. We demonstrate that this space forms a **convex polytope**, transforming the inference problem into an integration over a bounded linear constraint set.

3.1 PARAMETERIZATION OF CANONICAL SCM

A primary advantage of the canonical SCM framework is that it unifies the specification of structural functions $\mathbf{F}$ and the exogenous distribution $P(\mathbf{U})$. In this framework, a function $f_V \in \mathbf{F}$ is no longer treated as an unknown deterministic mechanism; rather, each value $u_V \in \Omega_{U_V}$ indexes one possible mapping from $\Omega_{\mathbf{Pa}_V}$ to Ω_V. Therefore, the uncertainty regarding the functional form of f_V is fully subsumed by the exogenous distribution $P(U_V)$. Since any two canonical SCMs compatible with the same causal graph and observational distribution differ only in their exogenous distributions, parameterizing a canonical SCM effectively reduces to characterizing $P(\mathbf{U})$.

Often, $P(\mathbf{U})$ is expressed as an $|\Omega_{\mathbf{U}}|$-dimensional joint probability vector lying on an $(|\Omega_{\mathbf{U}}|-1)$-dimensional simplex [Wu et al., 2019b]. While this parameterization simplifies observational and counterfactual probabilities into linear combinations of $P(\mathbf{U}=\mathbf{u})$, it fails to exploit the independence structure. Furthermore, it becomes computationally intractable as $|\mathbf{V}|$ and $|\Omega_{\mathbf{V}}|$ increase. To address these limitations, we employ a parameterization method that utilizes the factorization properties of the model.

Definition 6 (C-parameterization of a canonical SCM)**.** *Let $\mathcal{M}$ be a canonical SCM, and let $\mathcal{G}$ denote the causal graph induced by $\mathcal{M}$. Let $\mathcal{C}(\mathcal{G}) = \{\mathbf{C}_1, \dots, \mathbf{C}_m\}$ denote the set of all maximal C-components in $\mathcal{G}$, ordered by $\mathbf{I} = \{1, \dots, m\}$. Let $\{\mathbf{u}_{\mathbf{C}_i}^{(1)}, \dots, \mathbf{u}_{\mathbf{C}_i}^{(n_i)}\}$ be the set of all possible values of the exogenous $\mathbf{U}_{\mathbf{C}_i} = \bigcup_{V\in\mathbf{C}_i} \mathbf{U}_V$ ordered by $\mathbf{I}_i = \{1, \dots, n_i\}$ where $n_i = \prod_{V\in\mathbf{C}_i} |\Omega_{U_V}|$. Let $P_{\mathcal{M}}(\mathbf{U})$ denote the distribution of exogenous variable $\mathbf{U}$ of the model $\mathcal{M}$. The canonical SCM $\mathcal{M}$ is said to be in C-parameterized form if it is represented as follows:*

$$\theta_{\mathcal{M}} = [P_{\mathcal{M}}(\mathbf{U}_{\mathbf{C}_1} = \mathbf{u}_{\mathbf{C}_1}^{(1)}), \dots, P_{\mathcal{M}}(\mathbf{U}_{\mathbf{C}_m} = \mathbf{u}_{\mathbf{C}_m}^{(n_m)})]^\top.$$

We define the C-parameterized vector $\theta_{\mathcal{M}}$ as the concatenation of probability vectors corresponding to each maximal C-component. By decomposing the joint distribution, this method reduces the number of probability coordinates from $d_N = \prod_{i=1}^m n_i$ to $d = \sum_{i=1}^m n_i$, while encoding the independence between components.

As a concrete example, consider a canonical SCM $\mathcal{M}$, in which all variables $\{S, X, \hat{Y}\}$ are binary variables in $\{0,1\}$, inducing the causal graph in Figure 1a. The C-parameterized vector $\theta_{\mathcal{M}}$ has $d = |\Omega_{U_S}| \times |\Omega_{U_{\hat{Y}}}| + |\Omega_{U_X}| = 12$ probability coordinates:

$$\theta_{\mathcal{M}} = [P_{\mathcal{M}}(u_S^{(1)}, u_{\hat{Y}}^{(1)}), \dots, P_{\mathcal{M}}(u_S^{(2)}, u_{\hat{Y}}^{(4)}), \\ P_{\mathcal{M}}(u_X^{(1)}), \dots, P_{\mathcal{M}}(u_X^{(4)})]^\top.$$

3.2 EXPRESSING PROBABILITIES OF COUNTERFACTUAL PREDICTION INTERVALS

Building upon the C-parameterization, we formulate the probabilities of counterfactual prediction intervals. We systematically characterize the geometry of the feasible model space, integrate counterfactual constraints, and introduce a dimensionality reduction technique to improve computational feasibility.

Canonical hull and feasible parameter space To formulate the counterfactual prediction probability $P(\hat{y}_{\neg s} \mid \mathbf{x}, s)$, we first characterize the parameter set of all canonical SCMs compatible with the observational distribution $P(\mathbf{V})$ and causal graph $\mathcal{G}$.

Definition 7 (Canonical Hull over $\langle P(\mathbf{V}), \mathcal{G}\rangle$)**.** *For a given causal graph $\mathcal{G}$ and observational distribution $P(\mathbf{V})$, let $\mathbf{M}^{\langle P(\mathbf{V}),\mathcal{G}\rangle}$ be the set of all canonical SCMs compatible with $\langle P(\mathbf{V}), \mathcal{G}\rangle$. A canonical hull Θ over $\langle P(\mathbf{V}), \mathcal{G}\rangle$ is defined as follows:*

$$\Theta = \{\theta_{\mathcal{M}} \mid \mathcal{M} \in \mathbf{M}^{\langle P(\mathbf{V}),\mathcal{G}\rangle}\}.$$

A canonical hull is the set of C-parameterized vectors corresponding to observationally compatible canonical SCMs, which are mapped into $\mathbb{R}^d$ where $d = \sum_{\mathbf{C}_i} \prod_{V\in\mathbf{C}_i} |\Omega_{U_V}|$. Note that these models are observationally equivalent. This observational equivalence implies that the set Θ possesses the following property:

Lemma 2. *For any causal graph $\mathcal{G}$ and observational distribution $P(\mathbf{V})$, the canonical hull Θ over $\langle P(\mathbf{V}), \mathcal{G}\rangle$ is a convex polytope within an affine subspace of $\mathbb{R}^d$.*

Linear constraint representation of the canonical hull Lemma 2 establishes a canonical hull Θ can be represented as the solution set of linear constraints. To illustrate this, consider the causal graph shown in Figure 1b where all endogenous variables are binary. By Lemma 1, $P(\mathbf{v})$ is factorized into:

$$P(\mathbf{v}) = Q[X_1] \cdot Q[S] \cdot Q[X_2] \cdot Q[X_3, \hat{Y}],$$

where $Q[X_1] = P(x_1)$, $Q[S] = P(s \mid x_1)$, $Q[X_2] = P(x_2 \mid s)$, and $Q[X_3, \hat{Y}] = P(x_3 \mid x_2) \cdot P(\hat{y} \mid x_3, x_2)$. This expression follows from the topological-order factorization in Lemma 1 after applying the graph-implied conditional independencies for Figure 1b.

We define Θ as the set of parameter vectors $\theta_{\mathcal{M}} = [P_{\mathcal{M}}(u_{X_1}^{(1)}), \dots, P_{\mathcal{M}}(u_{X_3}^{(4)}, u_{\hat{Y}}^{(4)})] \in \mathbb{R}^{26}$ that satisfy the following linear constraints for all relevant endogenous value assignments and parent configurations. For brevity, following Wu et al. [2019b], we abbreviate the indicator function $\mathbb{I}\big(v = g_V^{(u_V)}(pa_V)\big)$ as $\mathbb{I}(v; pa_V, u_V)$. Then, $\forall \mathbf{U}_C \in \{\{U_{X_1}\}, \{U_S\}, \{U_{X_2}\}, \{U_{X_3}, U_{\hat{Y}}\}\}$,

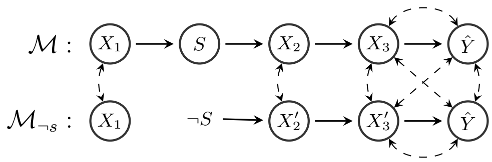


Figure 2: Twin network graph of the causal model in Figure 1b

$$\sum_{\mathbf{u}_C} P(\mathbf{u}_C) = 1, \quad 0 \le P(\mathbf{u}_C) \tag{1}$$

$$P(x_1) = \sum_{u_{X_1}} \mathbb{I}\left(x_1; \emptyset, u_{X_1}\right) \cdot P(u_{X_1}) \tag{2}$$

$$P(s \mid x_1) = \sum_{u_S} \mathbb{I}\left(s; x_1, u_S\right) \cdot P(u_S) \tag{3}$$

$$P(x_2 \mid s) = \sum_{u_{X_2}} \mathbb{I}\left(x_2; s, u_{X_2}\right) \cdot P(u_{X_2}) \tag{4}$$

$$\begin{aligned} P(\hat{y} \mid x_3, x_2) \cdot P(x_3 \mid x_2) \\ = \sum_{u_{\hat{Y}}, u_{X_3}} \mathbb{I}\left(\hat{y}; x_3, u_{\hat{Y}}\right) \cdot \mathbb{I}\left(x_3; x_2, u_{X_3}\right) \\ \cdot P(u_{\hat{Y}}, u_{X_3}). \end{aligned} \tag{5}$$

Eqs. (2)–(5) are derived from Definition 2. Since the system of linear equations above is underdetermined, the solution set Θ forms a convex polytope in $\mathbb{R}^{26}$.

Counterfactual interval constraints in parameter space Having defined the observational constraints, we now consider the subset of $\mathbf{M}^{\langle P(\mathbf{V}), \mathcal{G}\rangle}$ where the counterfactual prediction $P(\hat{y}_{\neg s} \mid \mathbf{x}, s)$ falls within a target interval $[l, r]$. Let $\Theta(\hat{y}, \mathbf{x}, s, l, r)$ denote the set of parameters $\theta_{\mathcal{M}}$ associated with the subset. With slight notational abuse, we use terms like $P_{\mathcal{M}}(\hat{y}_s)$, $\mathrm{CE}_{\mathcal{M}}(s, \neg s \mid \mathbf{x}; \hat{y})$ to denote the counterfactual quantity under a fixed causal model $\mathcal{M}$. The set $\Theta(\hat{y}, \mathbf{x}, s, l, r) \subseteq \Theta$ is formally defined as:

$$\Theta(\hat{y}, \mathbf{x}, s, l, r) = \{\theta_{\mathcal{M}} \in \Theta \mid l \le \mathrm{CE}_{\mathcal{M}}(s, \neg s \mid \mathbf{x}; \hat{y}) \le r\}. \tag{6}$$

Unlike the convex polytope Θ, the subset $\Theta(\hat{y}, \mathbf{x}, s, l, r)$ need not be convex, since it is the solution set of polynomial constraints. Returning to the extended IV example in Figure 1b, the set $\Theta(\hat{y}, \mathbf{x}, s, l, r)$ corresponds to the set of vectors $\theta_{\mathcal{M}}$ satisfying the observational constraints Eqs. (1)–(5) and the counterfactual constraint Eq. (6).

The counterfactual distribution $P_{\mathcal{M}}(\hat{y}_{\neg s} \mid \mathbf{x}, s)$ is the probability of the event $\hat{Y} = \hat{y}$ in the counterfactual model $\mathcal{M}_{\neg s}$ given factual observations $\mathbf{x}, s$ in model $\mathcal{M}$, where the factual and counterfactual worlds share the same exogenous context $\mathbf{U}$. To describe the process, we adopt the *twin network* framework [Balke and Pearl, 1994b] to integrate the SCMs $\mathcal{M}, \mathcal{M}_{\neg s}$ in a single causal graph. The twin network graph shown in Figure 2 implies that counterfactual quantity $P_{\mathcal{M}}(\hat{y}_{\neg s} \mid \mathbf{x}, s)$ in $\mathrm{CE}_{\mathcal{M}}(s, \neg s \mid \mathbf{x}; \hat{y})$ can be expressed as follows:

$$\begin{aligned} P_{\mathcal{M}}(\hat{y}_{\neg s} \mid \mathbf{x}, s) &= \frac{\sum_{x_2' x_3'} P_{\mathcal{M}}(\hat{y}_{\neg s}, x'_{2\neg s}, x'_{3\neg s}, x_1, x_2, x_3, s)}{P(x_1, x_2, x_3, s)} \\ &= \frac{\sum_{x_2' x_3'} Q_{\mathcal{M}}[\hat{Y}_{\neg s}, X_3, X'_{3\neg s}] \cdot Q_{\mathcal{M}}[X_2, X'_{2\neg s}] \cdot Q[X_1] \cdot Q[S]}{P(x_1, x_2, x_3, s)} \\ &= \frac{\sum_{x_2' x_3'} Q_{\mathcal{M}}[\hat{Y}_{\neg s}, X_3, X'_{3\neg s}] \cdot Q_{\mathcal{M}}[X_2, X'_{2\neg s}]}{P(x_2, x_3 \mid x_1, s)} \end{aligned} \tag{7}$$

The terms $Q_{\mathcal{M}}[\hat{Y}_{\neg s}, X_3, X'_{3\neg s}]$ and $Q_{\mathcal{M}}[X_2, X'_{2\neg s}]$ in Eq. (7) can be expanded as follows:

$$\begin{aligned} &Q_{\mathcal{M}}[\hat{Y}_{\neg s}, X_3, X'_{3\neg s}] \\ &= \sum_{u_{\hat{Y}}, u_{X_3}} \mathbb{I}\big(\hat{y}; x_3', u_{\hat{Y}}\big) \cdot \mathbb{I}\big(x_3; x_2, u_{X_3}\big) \\ &\qquad \cdot \mathbb{I}\big(x_3'; x_2', u_{X_3}\big) \cdot P_{\mathcal{M}}(u_{\hat{Y}}, u_{X_3}), \\ &Q_{\mathcal{M}}[X_2, X'_{2\neg s}] \\ &= \sum_{u_{X_2}} \mathbb{I}\big(x_2; s, u_{X_2}\big) \cdot \mathbb{I}\big(x_2'; \neg s, u_{X_2}\big) \cdot P_{\mathcal{M}}(u_{X_2}). \end{aligned} \tag{8}$$

Thus, $P(\hat{y}_{\neg s} \mid \mathbf{x}, s)$ involves the product of $P_{\mathcal{M}}(u_{X_2})$ and $P_{\mathcal{M}}(u_{X_3}, u_{\hat{Y}})$. Consequently, the counterfactual condition in Eq. (6) imposes a polynomial constraint on $\theta_{\mathcal{M}}$.

Probabilistic formulation over the model space To quantify the probability of fairness, we establish a probabilistic formulation over this parameter space. Given only the observational distribution $P(\mathbf{V})$ and the causal graph $\mathcal{G}$, any $\theta \in \Theta$ is a valid candidate for the true SCM parameters. To account for uncertainty regarding the true mechanism, one can formally define a probability space $\langle \Theta, \mathcal{F}_\Theta, \mathbb{P}_\Theta \rangle$, where Θ is the parameter space, $\mathcal{F}_\Theta$ is a σ-algebra on Θ, and $\mathbb{P}_\Theta$ is a probability measure defined on $\langle \Theta, \mathcal{F}_\Theta \rangle$. The probability that the counterfactual effect $\mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y})$ lies within $[l, r]$ is formulated as follows:

$$\mathbb{P}_\Theta\Big(l \le \mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y}) \le r\Big) = \int_{\Theta(\hat{y}, \mathbf{x}, s, l, r)} d\mathbb{P}_\Theta.$$

In the case of τ-counterfactual fairness, the set of all $\theta_{\mathcal{M}}$ such that $\hat{Y}$ is a τ-counterfactually fair predictor, denoted as $\Theta_{\mathrm{CF}}(\tau)$, is expressed as the intersection of $\Theta(\hat{y}, \mathbf{x}, s, -\tau, \tau)$ for all $(\hat{y}, \mathbf{x}, s) \in \Omega_{\hat{Y}} \times \Omega_{\mathbf{X}} \times \Omega_S$ with $P(\mathbf{x}, s) > 0$. The probability of $\hat{Y}$ being a τ-counterfactually fair predictor is formulated as follows:

$$\mathbb{P}_\Theta\Big(\max_{\substack{s, \mathbf{x}, \hat{y} \\ P(\mathbf{x}, s) > 0}} |\mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y})| \le \tau\Big) = \int_{\Theta_{\mathrm{CF}}(\tau)} d\mathbb{P}_\Theta,$$

$$\text{where} \quad \Theta_{\mathrm{CF}}(\tau) = \bigcap_{\substack{(\hat{y}, \mathbf{x}, s) \in \Omega_{\hat{Y}} \times \Omega_{\mathbf{X}} \times \Omega_S \\ P(\mathbf{x}, s) > 0}} \Theta\big(\hat{y}, \mathbf{x}, s, -\tau, \tau\big).$$

Dimension reduction via reduced C-parameterization Finally, to make the integration computationally feasible, we address the dimensionality of the parameter space. Although the constraint in Eq. (6) imposes a limitation on θ, it does not constrain the full dimensionality of Θ. In the previous example Figure 1b, as shown in Eqs. (7)–(8), the counterfactual constraint specifically restricts $P(u_{X_2})$ and $P(u_{X_3}, u_{\hat{Y}})$, while leaving $P(u_{X_1})$ and $P(u_S)$ unconstrained. This implies that computing the counterfactual effect does not require considering the dimensions of all u_V components. Let $\mathbf{de}_V$ denote the set of endogenous descendants of variable V, excluding V itself.

Definition 8 (Reduced C-parameterization)**.** *Considering the settings in Definition 6, let $\mathcal{C}^+ \subseteq \mathcal{C}(\mathcal{G})$ denote the set of all maximal C-components $\mathbf{C}_i$ such that $\mathbf{C}_i \cap \mathbf{de}_S \neq \emptyset$. For a canonical SCM $\mathcal{M} \in \mathbf{M}^{\langle P(\mathbf{V}),\mathcal{G}\rangle}$, the reduced C-parameter $\phi_\mathcal{M}$ is defined as the truncated vector of $\theta_\mathcal{M}$ such that for all $\mathbf{C}_i \in \mathcal{C}^+$,*

$$\phi_\mathcal{M} = [P_\mathcal{M}(\mathbf{U}_{\mathbf{C}_1} = \mathbf{u}^{(1)}_{\mathbf{C}_1}), \ldots, P_\mathcal{M}(\mathbf{U}_{\mathbf{C}_k} = \mathbf{u}^{(n_k)}_{\mathbf{C}_k})]^\top.$$

The parameter $\phi_\mathcal{M}$ is the reduced parameter vector of $\mathcal{M}$, which is the projection of $\theta_\mathcal{M}$ onto the counterfactually relevant coordinates $\mathbb{R}^{d_k}$, $d_k = \sum_{\mathbf{C} \in \mathcal{C}^+} |\Omega_{\mathbf{U}_\mathbf{C}}|$.

Proposition 3. *Let $\mathcal{C}^+ \subseteq \mathcal{C}(\mathcal{G})$ be the set of maximal C-components defined in Definition 8, and let $\pi : \Theta \to \Phi$ denote the coordinate projection onto the parameter blocks indexed by $\mathcal{C}^+$. Then for every fixed $(\hat{y}, \mathbf{x}, s)$ with $P(\mathbf{x}, s) > 0$, there exists a measurable map $g_{\hat{y},\mathbf{x},s} : \Phi \to \mathbb{R}$ such that*

$$\mathrm{CE}_\mathcal{M}(s, \neg s \mid \mathbf{x}; \hat{y}) = g_{\hat{y},\mathbf{x},s}(\phi_\mathcal{M}) \quad \textit{where} \quad \phi_\mathcal{M} = \pi(\theta_\mathcal{M})$$

Proposition 3 shows that the reduced C-parameter $\phi_\mathcal{M}$ is sufficient for calculating the counterfactual effect of a canonical SCM, and thus the dimension can be reduced. For example, consider a canonical SCM $\mathcal{M}$ in which all endogenous variables are binary, inducing the causal graph in Figure 1c. The components $[P(u^{(1)}_{X_1}, u^{(1)}_S), \ldots, P(u^{(4)}_{X_2})] \in \mathbb{R}^{12}$ are not included in the reduced C-parameter $\phi_\mathcal{M}$, thus the coordinate count $d = 28$ reduces to $d = 16$. By adopting the reduced C-parameter, we can define the dimension-reduced parameter sets Φ, $\Phi(\hat{y}, \mathbf{x}, s, l, r)$, $\Phi_{\mathrm{CF}}(\tau)$ as follows:

$$\Phi = \{\phi_\mathcal{M} \mid \mathcal{M} \in \mathbf{M}^{\langle P(\mathbf{V}),\mathcal{G}\rangle}\},$$
$$\Phi(\hat{y}, \mathbf{x}, s, l, r) = \{\phi_\mathcal{M} \in \Phi \mid l \leq \mathrm{CE}_\mathcal{M}(s, \neg s \mid \mathbf{x}; \hat{y}) \leq r\},$$
$$\Phi_{\mathrm{CF}}(\tau) = \bigcap_{\substack{(\hat{y},\mathbf{x},s) \in \Omega_{\hat{Y}} \times \Omega_\mathbf{X} \times \Omega_S \\ P(\mathbf{x},s)>0}} \Phi(\hat{y}, \mathbf{x}, s, -\tau, \tau).$$

Since the counterfactual effect depends solely on the reduced parameter ϕ, the target event can be evaluated on the reduced space. We write $\mathbb{P}_\Phi$ for the probability measure used on Φ; when a full-space prior $\mathbb{P}_\Theta$ is specified, $\mathbb{P}_\Phi$ may be taken as the pushforward $\pi_\# \mathbb{P}_\Theta$. We suggest modeling $P(\phi)$ as a uniform distribution over Φ, if no prior knowledge is provided, while noting that uniform prior on Φ is not necessarily induced by a uniform prior on Θ.

Lemma 4. *Given a causal graph $\mathcal{G}$, observational distribution $P(\mathbf{V})$, and a probability measure $\mathbb{P}_\Phi$ on Φ, for any $(\hat{y}, \mathbf{x}, s)$ with $P(\mathbf{x}, s) > 0$, the probability that $\mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y})$ lies within $[l, r]$ is expressed as:*

$$\mathbb{P}_\Phi\Big(l \leq \mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y}) \leq r\Big) = \int_{\Phi(\hat{y},\mathbf{x},s,l,r)} d\mathbb{P}_\Phi.$$

The probability that $\hat{Y}$ is a τ-counterfactually fair predictor is also expressed accordingly:

$$\mathbb{P}_\Phi\Big(\max_{\substack{s,\mathbf{x},\hat{y} \\ P(\mathbf{x},s)>0}} |\mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y})| \leq \tau\Big) = \int_{\Phi_{\mathrm{CF}}(\tau)} d\mathbb{P}_\Phi.$$

Lemma 4 demonstrates that the constraints on ϕ are sufficient to obtain the probability of the model's counterfactual prediction while the resulting probability remains prior-dependent. Detailed proofs of Lemma 2, Proposition 3 and Lemma 4 are provided in Appendix B.

4 ALGORITHMIC FRAMEWORK

Building upon these geometric insights, we now present the algorithmic formulation of our findings. In general, analytically computing these probabilities is infeasible, as it requires integrating a probability density over a high-dimensional polytope for which no closed-form solution exists. Therefore, we adopt an MCMC method for estimating the probabilities of counterfactual prediction intervals. The algorithm is intended as a post-hoc audit for a fixed black-box predictor $\hat{Y}$—which may be randomized (outputting a probability $P(\hat{Y}{=}\hat{y} \mid \cdot)$ from which the label is drawn, as in our synthetic studies) or deterministic (as in the LSAC study of Appendix D)—not as a training-time fairness objective.

Algorithm 1 outlines the procedure for estimating the probabilities of counterfactual prediction intervals using the Hit-and-Run algorithm [Smith, 1984, Lovász and Vempala, 2006]. Here, Φ^* denotes the subset of parameters satisfying the target counterfactual prediction interval: $\Phi(\hat{y}, \mathbf{x}, s, l, r)$ or $\Phi_{\mathrm{CF}}(\tau)$. We initialize the sampling process at a feasible point $\phi^{(0)} \in \Phi$. At each iteration t, we generate a random line passing through $\phi^{(t)}$ and identify the line segment formed by the intersection of this line and the polytope Φ. The direction d is sampled uniformly from the unit sphere lying within the affine subspace containing Φ. More specifically, d is obtained by projecting a standard normal vector $v \sim \mathcal{N}(0, I^{\dim(\phi)})$ onto the null space of the matrix A_1 (the coefficient matrix of the equality constraints) and normalizing it. The next state $\phi^{(t+1)}$ is then drawn from the density

Algorithm 1 Hit-and-Run Estimator

Input:
  Constraints on $\Phi : A_1\phi = b_1,\ A_2\phi \le b_2$
  Constraints on $\Phi^* : h(\phi) \le b_3$
  density function $f(\phi)$
  Number of iterations T
Output: A point estimate of $P(\phi \in \Phi^*)$
Initialize: $\phi^{(0)} \in \Phi$, $K \leftarrow 0$
for $t = 0$ to $T - 1$ **do**
  Sample random vector $v \sim \mathcal{N}(0, I)$
  Project v to nullspace of A_1 to get d
  Normalize $d \leftarrow d/\|d\|_2$
  Find $[\lambda_{\min}, \lambda_{\max}]$ such that $A_2(\phi^{(t)} + \lambda d) \le b_2$
  Define $g(\lambda) = f(\phi^{(t)} + \lambda d) \cdot \mathbb{I}(\lambda \in [\lambda_{\min}, \lambda_{\max}])$
  Sample $\lambda^* \sim P(\lambda) \propto g(\lambda)$
  Update $\phi^{(t+1)} \leftarrow \phi^{(t)} + \lambda^* d$
  if $h(\phi^{(t+1)}) \le b_3$ **then**
    $K \leftarrow K + 1$
return K/T

function $f(\phi) \propto \mathcal{P}(\phi)$ restricted to this line segment. Finally, we check whether the new sample $\phi^{(t+1)}$ falls within the target region Φ^*.

Sampling $\phi^{(t+1)}$ from the density function $f(\phi)$ restricted to the line segment can be performed by adopting an auxiliary MCMC method, such as the Metropolis-Hastings algorithm [Metropolis et al., 1953, Hastings, 1970] or slice sampling [Neal, 2003], depending on the characteristics of $f(\phi)$.

The output of Algorithm 1 is the ratio of K, the number of iterations in which the sample $\phi^{(t+1)}$ is included in Φ^*, to T, the total number of iterations. Under the standard stationarity and ergodicity conditions for the resulting Markov chain, the output K/T is a consistent estimator of the probability. We report this quantity as a prior-dependent audit statistic rather than prescribing a universal practitioner threshold.

Auditor protocol Because the audit statistic is prior-dependent, we suggest a pre-registered protocol that guards against prior-gaming. The auditor pre-declares (i) a prior family, (ii) the fairness threshold τ, and (iii) a swing tolerance Δ. The audit then reports the probability of τ-counterfactual fairness under the declared prior. If the estimate swings by more than Δ across the declared prior family (e.g., Dirichlet priors with $\alpha \in \{1, 5, 10\}$), or if the Hit-and-Run chains fail to converge (e.g., Gelman–Rubin $\hat{R} > 2$), the audit is declared *inconclusive* and falls back to the prior-free worst-case bound (e.g., the partial identification method from Zhang et al. [2022]). Consequently, an adversary who selects "the prior that helps" obtains a refusal rather than a favorable verdict, whereas genuinely conclusive cases are unaffected. This pre-registration is analogous to fixing a significance level before seeing the data.

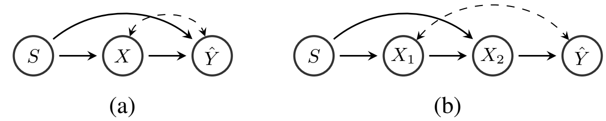


Figure 3: Causal diagrams of structural causal models of (a) dataset $\mathcal{D}_1$ and (b) dataset $\mathcal{D}_2$.

5 SIMULATION STUDIES

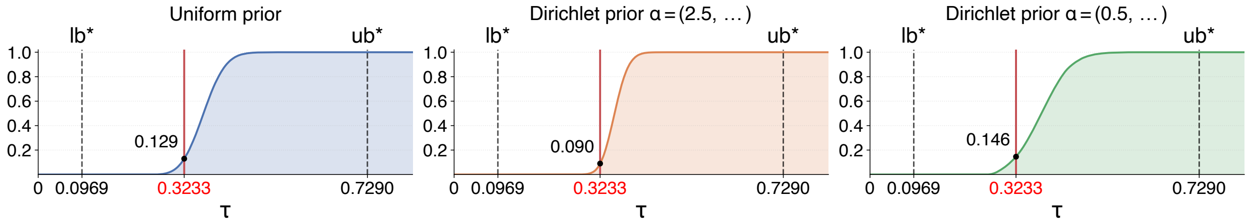
In this section, we present synthetic simulations,[1] illustrating both the consistency of our probabilistic metric with known ground truth and its behavior relative to existing bound-based methods.

Datasets To evaluate the validity of our proposed method, we generated two synthetic datasets, $\mathcal{D}_1$ and $\mathcal{D}_2$, based on distinct causal graphs in Figure 3. The datasets were generated under synthetic SCMs, where all endogenous variables $V \in \mathbf{V}$ are binary and all exogenous variables $U \in \mathbf{U}$ take values in $\mathbb{R}$. We constructed $\mathcal{D}_1$ and $\mathcal{D}_2$ by sampling 10^6 observations from each SCM. Ground truth counterfactuals were obtained by intervening to set $S = 1 - s$ while holding the same exogenous context fixed. Detailed specifications of the SCMs used for data generation are provided in Appendix C.

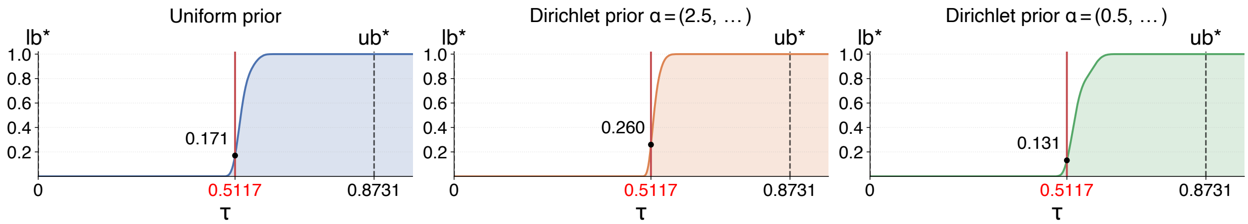
Hit-and-Run estimator details We implemented the Hit-and-Run sampling procedure outlined in Algorithm 1 using the Python `hopsy` package [Paul et al., 2024]. In both simulations, we drew a total of 10^5 valid samples from the reduced feasible polytope. For the starting point $\phi^{(0)}$, we used Chebyshev center of Φ—the center of the largest Euclidean ball inscribed in Φ—computed via `hopsy`. To mitigate autocorrelation among the generated samples, we discarded the first 10^4 iterations as a burn-in period and applied a thinning factor of 10^3. For the priors used in our simulations, $f(\phi)$ is either uniform over Φ or induced by a joint Dirichlet distribution over the relevant exogenous-variable blocks. For $\mathcal{D}_2$, the empirical equality-constraint right-hand sides were projected to a consistent right-hand side before sampling; this finite-sample correction is detailed in Appendix C.

Simulation 1 We empirically examined whether the probability estimated by Algorithm 1 is consistent with the ground truth SCM. Using 10^5 valid samples drawn by the Hit-and-Run sampler, we estimated the probability that an SCM compatible with $\mathcal{D}_1$ satisfies τ-counterfactual fairness, varying the threshold τ from 0 to 1. For the prior distribution of ϕ, we compared a uniform prior against joint Dirichlet priors. To establish a baseline, we evaluated our results against

[1] Code and result data for reproducing the simulations are available at `https://github.com/snu-causality-lab/measuring-counterfactual-fairness`.



(a) Simulation 1 Results



(b) Simulation 2 Results

Figure 4: Estimated probability of τ-counterfactual fairness for predictor $\hat{Y}$ on $\mathcal{D}_1$ and $\mathcal{D}_2$. Results are shown for Uniform, Dirichlet ($\alpha = 2.5$), and Dirichlet ($\alpha = 0.5$) priors. Dashed vertical lines indicate the bounds from Zhang et al. [2022] for $\mathcal{D}_1$ and the LP-based bounds from Wu et al. [2019b] for $\mathcal{D}_2$; red lines indicate the ground-truth $\max|\mathrm{CE}|$ values, 0.3233 and 0.5117, respectively.

the tightest bound $[lb^*, ub^*]$ of $\max|\mathrm{CE}|$, derived by the partial identification method of Zhang et al. [2022]. Theoretically, the probability of the SCM being τ-counterfactually fair must be exactly 0 when $\tau < lb^*$ and exactly 1 when $\tau \geq ub^*$. As illustrated in Figure 4a, the ground truth value of $\max|\mathrm{CE}|$ falls within the region where our estimated probability is strictly between 0 and 1. Furthermore, our estimates are consistent with the tight bounds $[0.0969, 0.7290]$ computed using the same partial-identification baseline, corroborating our proposed framework.

Simulation 2 To validate our estimator as a practical evaluation metric, we compared our probability measure against existing bound-based intervals on dataset $\mathcal{D}_2$. Similar to Simulation 1, we estimated the probability of τ-counterfactual fairness by drawing 10^5 samples under uniform and joint Dirichlet priors, and compared it with the bounds $[lb^*, ub^*]$ of $\max|\mathrm{CE}|$. Unlike for $\mathcal{D}_1$, applying the method of Zhang et al. [2022] to $\mathcal{D}_2$ yields a polynomial optimization problem. Thus, we adopted Wu et al. [2019b] as a computationally feasible baseline, in which the bounds are derived via linear programming by assuming all endogenous variables form a single maximal C-component. This LP interval should be interpreted as a feasible loose baseline for $\mathcal{D}_2$, rather than as a tight-bound comparison. As illustrated in Figure 4b, the baseline LP interval $[0, 0.8731]$ is broad, limiting its usefulness as a practical metric. In contrast, our estimated distribution is concentrated around the ground truth value of 0.5117 under the selected priors. This highlights that worst-case bounds are overly conservative as they are optimized over worst-case configurations. Consequently, when the chosen prior is defensible, our probabilistic metric offers a more informative, prior-dependent summary.

Our simulation studies are intended to validate our method under known ground truth, rather than to establish its general applicability. In Appendix D, we provide an auditing case study on the law school success example [Wightman, 1998, Kusner et al., 2017], where our proposed Hit-and-Run estimator is applied to a given predictor.

6 CONCLUSION

In this work, we introduced a novel probabilistic framework for auditing the counterfactual fairness of black-box models. Recognizing the practical limitations of interval-based partial identification, we shift the evaluation focus from worst-case bounds to quantifying the probability of fairness. We geometrically characterized the space of compatible canonical SCMs as a convex polytope and leveraged conditional independencies via C-parameterization to systematically reduce the dimensionality of the inference problem (Lemmas 2 and 4). This enabled a Hit-and-Run MCMC estimator for tractable probability estimation (Algorithm 1).

While our approach provides a rigorous metric for auditing fairness, it currently assumes discrete endogenous variables. Promising next steps include extending the geometric framework to continuous variables (e.g., via quantile discretization) and to path-specific fairness criteria, and scaling the integration to larger and denser causal graphs via more advanced sampling, building on the known polynomial mixing-time results for Hit-and-Run in suitably rounded convex bodies [Lovász and Vempala, 2006].

Acknowledgements

We thank anonymous reviewers for constructive comments to improve the manuscript. This work was supported by the NRF (RS-2023-00211904) grant funded by the Korean government.

References

Alexander Balke and Judea Pearl. Counterfactual probabilities: computational methods, bounds and applications. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, UAI'94, pages 46–54, San Francisco, CA, USA, 1994a. Morgan Kaufmann Publishers Inc. ISBN 1558603328.

Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI)*, volume 1, pages 230–237. AAAI Press, 1994b.

Ludwig Bothmann, Susanne Dandl, and Michael Schomaker. Causal fair machine learning via rank-preserving interventional distributions. In *Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with the 26th European Conference on Artificial Intelligence (ECAI 2023)*, volume 3523 of *CEUR Workshop Proceedings*, 2023.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7), 2024. doi: 10.1145/3616865.

Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Counterfactual fairness for facial expression recognition. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, volume 13805 of *Lecture Notes in Computer Science*, pages 245–261, Cham, 2023. Springer. doi: 10.1007/978-3-031-25072-9_16.

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

Yoichi Chikahara, Shinsaku Sakaue, Akinori Fujino, and Hisashi Kashima. Learning individually fair classifier with path-specific causal-effect constraint. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pages 145–153. PMLR, 2021.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312):1–117, 2023.

Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 582–593, 2020.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, 2020.

Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, 119(547):1778–1793, 2024. doi: 10.1080/01621459.2023.2216909.

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 (ITCS)*, pages 214–226. ACM, 2012.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 219–226, 2019.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57: 97–109, 1970. doi: 10.1093/biomet/57.1.97.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 35–50. Springer, 2012.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 656–666, 2017.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017. doi: 10.4230/LIPIcs.ITCS.2017.43.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

László Lovász and Santosh Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4):985–1005, 2006.

Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. Learning fair node representations with graph counterfactual fairness. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 695–703, 2022.

Jing Ma, Ruocheng Guo, Aidong Zhang, and Jundong Li. Learning for counterfactual fairness from observational data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1620–1630, 2023.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 386–400, 2021.

Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. Improving counterfactual generation for fair hate speech detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH)*, pages 92–101, 2021.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Radford M Neal. Slice sampling. *The annals of statistics*, 31(3):705–767, 2003.

Richard D Paul, Johann F Jadebeck, Anton Stratmann, Wolfgang Wiechert, and Katharina Nöh. hopsy — a methods marketplace for convex polytope sampling in python. *Bioinformatics*, 40(7):btae430, 07 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae430.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009. ISBN 9780521895606.

Drago Plečko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44, 2020.

Saeyoung Rho, Junzhe Zhang, and Elias Bareinboim. Partial identification approach to counterfactual fairness assessment. *arXiv preprint arXiv:2510.00163*, 2025.

Michael C. Sachs, Gustav Jonzon, Arvid Sjölander, and Erin E. Gabriel. A general method for deriving tight symbolic bounds on causal effects. *Journal of Computational and Graphical Statistics*, 32(2):567–576, 2023. doi: 10.1080/10618600.2022.2071905.

Ricardo Silva. Counterfactual fairness is not demographic parity, and other observations. *arXiv preprint arXiv:2402.02663*, 2024.

Robert L Smith. Efficient monte carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 2239–2248, 2018.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI)*, pages 567–573, 2002.

Linda F. Wightman. Lsac national longitudinal bar passage study. Lsac research report series, Law School Admission Council, 1998.

Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1438–1444, 2019a.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. PC-fairness: A unified framework for measuring causality-based fairness. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019b.

Yongkai Wu, Lu Zhang, and Xintao Wu. Fair machine learning through the lens of causality. In Sheng Li and Zhixuan Chu, editors, *Machine Learning for Causal Inference*, pages 103–135. Springer International Publishing, Cham, 2023. doi: 10.1007/978-3-031-35051-1_6.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970. PMLR, 2017.

Marco Zaffalon, Alessandro Antonucci, Rafael Cabañas, David Huber, and Dario Azzimonti. Efficient computation of counterfactual bounds. *International Journal of Approximate Reasoning*, 171:109111, 2024. doi: 10.1016/j.ijar.2023.109111.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28, pages 325–333, 2013.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26548–26558. PMLR, 17–23 Jul 2022.

Zhiqun Zuo, Mohammad Mahdi Khalili, and Xueru Zhang. Counterfactually fair representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Beyond Bounds: Quantifying the Probability of Counterfactual Fairness (Supplementary Material)

Taehan Kim[2] **Minyoung Cho**[1] **Sanghack Lee**[*1]

[1]Graduate School of Data Science, Seoul National University, Seoul, Republic of Korea
[2]Korean National Police Agency, Republic of Korea

A DISCUSSION AND LIMITATIONS

While our proposed framework provides a novel paradigm for auditing counterfactual fairness through a probabilistic lens, it is important to contextualize our methodological choices and discuss open challenges.

Computational Scalability and Baseline Comparisons A core theoretical contribution of this work is the geometric characterization of the parameter space and its subsequent dimensionality reduction via reduced C-parameterization. This structural reduction is precisely what makes the Hit-and-Run MCMC integration computationally viable. However, exploring high-dimensional convex polytopes intrinsically poses scalability challenges as the density of the causal graph and variable cardinalities grow.

In our evaluations, we contrast our probabilistic integration with existing partial identification bounds to highlight these pragmatic trade-offs. For relatively simple causal graphs, state-of-the-art bounding methods [Zhang et al., 2022] can provide tight intervals by solving polynomial optimization problems. Yet, as graph complexity increases, these optimizations quickly become theoretically and computationally intractable. In such regimes, relying on looser LP-based bounds or utilizing our proposed MCMC sampling—which yields a rich distributional understanding rather than mere worst-case endpoints—becomes a more practical alternative for fairness auditors. Hit-and-Run has polynomial mixing-time guarantees for suitably rounded convex bodies [Lovász and Vempala, 2006], but practical performance still depends on dimension, geometry, and the target density. Our empirical evaluation on fundamental synthetic structures aims to validate the correctness of our sampling geometry against known ground truths in a controlled setting. Scaling our integration approach to large-scale, real-world graphs via advanced sampling techniques or variational inference remains an important direction for future work.

The Role of Priors in Fairness Auditing Our probabilistic formulation computes the likelihood of counterfactual fairness conditioned on a prior distribution over the unidentifiable parameter space Φ. In our current implementation, we employ uniform and Dirichlet priors to model a state of minimal structural assumptions. We emphasize that this reliance on a prior is both a modeling choice and a source of sensitivity. In real-world applications (e.g., healthcare or criminal justice), domain experts often possess partial, substantive knowledge regarding plausible causal mechanisms. Our framework is explicitly designed to accommodate these domain-specific priors. By injecting expert knowledge, auditors can effectively narrow the canonical polytope to compute more contextually grounded and refined fairness probabilities. Because priors can also be misspecified or strategically chosen, practical audits should disclose the prior and report sensitivity across plausible prior classes. This capacity to translate qualitative expert intuition into quantitative fairness audits offers a practical complement to worst-case bounding methodologies.

Path-Specific Fairness and Granular Audits The present work focuses on quantifying global τ-counterfactual fairness to establish a foundational metric for overall black-box model behavior. Methodologically, because our convex polytope construction completely characterizes the ambiguity of all underlying structural equations, the framework naturally extends

*Corresponding author: `sanghack@snu.ac.kr`.

to more granular fairness concepts. For example, auditing for path-specific fairness (to disentangle direct versus indirect discrimination) ultimately queries the same unidentifiable mechanisms. Adapting our integration constraints to accommodate path-specific counterfactual queries represents a promising mathematical extension of the current pipeline, allowing for more nuanced legal and ethical auditing.

B PROOFS

In this section, we restate and prove Lemma 2, Proposition 3, and Lemma 4 from Section 3.

B.1 PROOF OF LEMMA 2

Lemma 2. *For any causal graph $\mathcal{G}$ and observational distribution $P(\mathbf{V})$, the canonical hull Θ over $\langle P(\mathbf{V}), \mathcal{G}\rangle$ is a convex polytope within an affine subspace of $\mathbb{R}^d$.*

Proof. We prove the lemma by showing that Θ is a bounded solution set of linear constraints.

For a causal graph $\mathcal{G}$, let $\mathcal{C}(\mathcal{G}) = \{\mathbf{C}_1, \dots \mathbf{C}_m\}$ denote the set of all maximal C-components in $\mathcal{G}$. Let n_i, m_i, l_i denote, respectively, the cardinalities $|\Omega_{\mathbf{C}_i}|, |\Omega_{\mathbf{Pa}_{\mathbf{C}_i}}|, |\Omega_{\mathbf{U}_{\mathbf{C}_i}}|$ of a canonical SCM corresponding to $\langle P(\mathbf{V}), \mathcal{G}\rangle$. Let $I_{\mathbf{C}_i}$ denote the selection matrix such that for a C-parameterized vector $\theta = [P(\mathbf{u}^{(1)}_{\mathbf{C}_1}), \dots, P(\mathbf{u}^{(l_m)}_{\mathbf{C}_m})]^\top \in \Theta, I_{\mathbf{C}_i} \cdot \theta = [P(\mathbf{u}^{(1)}_{\mathbf{C}_i}), \dots P(\mathbf{u}^{(l_i)}_{\mathbf{C}_i})]^\top$. Let $\mathbf{1}^n$ denote the vector $[1, \dots, 1]^\top \in \mathbb{R}^n$.

By Definition 6, for every $\mathbf{C}_i \in \mathcal{C}(\mathcal{G})$, θ satisfies the following constraints.

$$I_{\mathbf{C}_i} \cdot \theta \in [0,1]^{l_i}, \quad \left(\mathbf{1}^{l_i}\right)^\top \cdot I_{\mathbf{C}_i} \cdot \theta = 1 \tag{9}$$

By Definition 2 and Lemma 1, for all $\mathbf{C}_i \in \mathcal{C}(\mathcal{G})$, θ satisfies the following constraints.

$$\begin{bmatrix} \prod_{v \in \mathbf{c}_i^{(1)}} \mathbb{I}\big(v; \mathbf{pa}_V^{(1)}, \mathbf{u}^{(1)}_{\mathbf{C}_i}\big) & \dots & \prod_{v \in \mathbf{c}_i^{(1)}} \mathbb{I}\big(v; \mathbf{pa}_V^{(1)}, \mathbf{u}^{(l_i)}_{\mathbf{C}_i}\big) \\ \vdots & \ddots & \vdots \\ \prod_{v \in \mathbf{c}_i^{(n_i)}} \mathbb{I}\big(v; \mathbf{pa}_V^{(m_i)}, \mathbf{u}^{(1)}_{\mathbf{C}_i}\big) & \dots & \prod_{v \in \mathbf{c}_i^{(n_i)}} \mathbb{I}\big(v; \mathbf{pa}_V^{(m_i)}, \mathbf{u}^{(l_i)}_{\mathbf{C}_i}\big) \end{bmatrix} \cdot I_{\mathbf{C}_i} \cdot \theta = \begin{bmatrix} Q[\mathbf{C}_i](\mathbf{c}_i^{(1)}, \mathbf{pa}^{(1)}_{\mathbf{C}_i}) \\ \vdots \\ Q[\mathbf{C}_i](\mathbf{c}_i^{(n_i)}, \mathbf{pa}^{(m_i)}_{\mathbf{C}_i}) \end{bmatrix} \tag{10}$$

Let Ψ denote the solution set of constraints (9) and (10). Then, $\Theta \subseteq \Psi$. Since the set Ψ is bounded and defined by finitely many linear equality and inequality constraints, it is a convex polytope.

We now prove $\Psi \subseteq \Theta$. Take any $\psi \in \Psi$. Construct a canonical SCM on the same graph by using the canonical response-function domains and assigning the C-component exogenous distributions according to the corresponding blocks $I_{\mathbf{C}_i}\psi$. The constraint (9) makes each block a valid probability distribution, and independence across maximal C-components is exactly the C-parameterized exogenous factorization. Let $P_\psi(\mathbf{v}),\ Q_\psi[\mathbf{C}](\mathbf{c}, \mathbf{pa}_{\mathbf{C}})$ denote the observed probability and Q-factor of the canonical SCM respectively. Applying Lemma 1 gives

$$P_\psi(\mathbf{v}) = \prod_{\mathbf{C}_i \in \mathcal{C}(\mathcal{G})} Q_\psi[\mathbf{C}_i](\mathbf{c}_i, \mathbf{pa}_{\mathbf{C}_i}) = \prod_{\mathbf{C}_i \in \mathcal{C}(\mathcal{G})} Q[\mathbf{C}_i](\mathbf{c}_i, \mathbf{pa}_{\mathbf{C}_i}) = P(\mathbf{v})$$

for all $\mathbf{v}$. The constructed canonical SCM is compatible with $\langle P(\mathbf{V}), \mathcal{G}\rangle$, so $\Psi \subseteq \Theta$. Therefore Θ is a convex polytope. $\square$

B.2 PROOF OF PROPOSITION 3

Proposition 3. *Let $\mathcal{C}^+ \subseteq \mathcal{C}(\mathcal{G})$ be the set of maximal C-components defined in Definition 8, and let $\pi : \Theta \to \Phi$ denote the coordinate projection onto the parameter blocks indexed by $\mathcal{C}^+$. Then for every fixed $(\hat{y}, \mathbf{x}, s)$ with $P(\mathbf{x}, s) > 0$, there exists a measurable map $g_{\hat{y},\mathbf{x},s} : \Phi \to \mathbb{R}$ such that*

$$\mathrm{CE}_{\mathcal{M}}(s, \neg s \mid \mathbf{x}; \hat{y}) = g_{\hat{y},\mathbf{x},s}(\phi_{\mathcal{M}})$$
$$\textit{where} \quad \phi_{\mathcal{M}} = \pi(\theta_{\mathcal{M}})$$

Proof. Recall from Definition 4 that, for a canonical SCM $\mathcal{M}$ and fixed values $(\hat{y}, \mathbf{x}, s)$ with $P(\mathbf{x}, s) > 0$,

$$\mathrm{CE}_{\mathcal{M}}(s, \neg s \mid \mathbf{x}; \hat{y}) = P_{\mathcal{M}}(\hat{y}_{\neg s} \mid \mathbf{x}, s) - P_{\mathcal{M}}(\hat{y} \mid \mathbf{x}, s).$$

The factual term $P_{\mathcal{M}}(\hat{y} \mid \mathbf{x}, s)$ equals the observational probability $P(\hat{y} \mid \mathbf{x}, s)$ for every $\theta_{\mathcal{M}} \in \Theta$. It is therefore enough to prove that $P_{\mathcal{M}}(\hat{y}_{\neg s} \mid \mathbf{x}, s)$ depends only on $\pi(\theta_{\mathcal{M}})$.

Let $\mathcal{C}^- = \mathcal{C}(\mathcal{G}) \setminus \mathcal{C}^+$. Write $\theta_{\mathcal{M}} = (\theta^+_{\mathcal{M}}, \theta^-_{\mathcal{M}})$, where $\theta^+_{\mathcal{M}} = \pi(\theta_{\mathcal{M}})$ contains exactly the probability blocks for C-components in $\mathcal{C}^+$. By the C-parameterization, the exogenous distribution factorizes over maximal C-components:

$$P_{\mathcal{M}}(\mathbf{u}) = \prod_{\mathbf{C} \in \mathcal{C}(\mathcal{G})} P_{\mathcal{M}}(\mathbf{u}_{\mathbf{C}}).$$

By the twin-network representation, the numerator of the counterfactual term can be written as the following finite sum:

$$P_{\mathcal{M}}(\hat{y}_{\neg s}, \mathbf{x}, s) = \sum_{\mathbf{u}} \mathbb{I}\big(\mathbf{X}(\mathbf{u}) = \mathbf{x}, S(\mathbf{u}) = s\big) \mathbb{I}\big(\hat{Y}_{\neg s}(\mathbf{u}) = \hat{y}\big) \prod_{\mathbf{C} \in \mathcal{C}(\mathcal{G})} P_{\mathcal{M}}(\mathbf{u}_{\mathbf{C}}).$$

Fix a topological order compatible with $\mathcal{G}$ and sum out the exogenous blocks component by component. For any $\mathbf{C} \in \mathcal{C}^-$, no variable in $\mathbf{C} \setminus \{S\}$ is affected by the intervention $do(S = \neg s)$. If $S \in \mathbf{C}$, its counterfactual value is fixed to $\neg s$, while $\mathbf{u}_{\mathbf{C}}$ enters through the factual event $S(\mathbf{u}) = s$. On the support of the factual indicator $\mathbb{I}(\mathbf{X}(\mathbf{u}) = \mathbf{x}, S(\mathbf{u}) = s)$, the factual variables in $\mathbf{C} \cap \mathbf{X}$ are fixed to their observed values. If $\hat{Y} \in \mathbf{C}$, then $\hat{Y}_{\neg s}(\mathbf{u}) = \hat{Y}(\mathbf{u})$, and thus the counterfactual indicator also fixes $\hat{Y} = \hat{y}$. Conditional on the relevant parent values, the factors involving $\mathbf{u}_{\mathbf{C}}$ in the above finite sum have the form

$$\sum_{\mathbf{u}_{\mathbf{C}}} P_{\mathcal{M}}(\mathbf{u}_{\mathbf{C}}) \prod_{V \in \mathbf{C}} \mathbb{I}(c_V; \mathbf{pa}_V, \mathbf{u}_V) = Q_{\mathcal{M}}[\mathbf{C}](\mathbf{c}, \mathbf{pa}_{\mathbf{C}}).$$

Since $\theta_{\mathcal{M}} \in \Theta$, this factor equals the observational $Q[\mathbf{C}](\mathbf{c}, \mathbf{pa}_{\mathbf{C}})$ determined by $P(\mathbf{V})$ and therefore does not depend on $\theta^-_{\mathcal{M}}$. Iterating this summation over all components in $\mathcal{C}^-$ leaves a finite expression whose only nonconstant probability weights are the blocks $\{P_{\mathcal{M}}(\mathbf{u}_{\mathbf{C}}) : \mathbf{C} \in \mathcal{C}^+\}$. Hence $P_{\mathcal{M}}(\hat{y}_{\neg s}, \mathbf{x}, s)$ depends only on $\theta^+_{\mathcal{M}} = \pi(\theta_{\mathcal{M}})$, with all remaining terms fixed independently of $\theta^-_{\mathcal{M}}$. Since all domains are finite, this dependence is through a finite sum of products of coordinate projections of $\pi(\theta_{\mathcal{M}})$ and fixed constants, and is therefore measurable. The denominator $P_{\mathcal{M}}(\mathbf{x}, s) = P(\mathbf{x}, s)$ is observationally fixed and positive, and the factual term $P_{\mathcal{M}}(\hat{y} \mid \mathbf{x}, s) = P(\hat{y} \mid \mathbf{x}, s)$ is also observationally fixed. Therefore $\mathrm{CE}_{\mathcal{M}}(s, \neg s \mid \mathbf{x}; \hat{y})$ is a measurable function of $\pi(\theta_{\mathcal{M}})$. This induces the required measurable map $g_{\hat{y}, \mathbf{x}, s} : \Phi \to \mathbb{R}$. □

B.3 PROOF OF LEMMA 4

Lemma 4. *Given a causal graph $\mathcal{G}$, observational distribution $P(\mathbf{V})$, and a probability measure $\mathbb{P}_\Phi$ on Φ, for any $(\hat{y}, \mathbf{x}, s)$ with $P(\mathbf{x}, s) > 0$, the probability that $\mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y})$ lies within $[l, r]$ is expressed as:*

$$\mathbb{P}_\Phi\Big(l \leq \mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y}) \leq r\Big) = \int_{\Phi(\hat{y}, \mathbf{x}, s, l, r)} d\mathbb{P}_\Phi.$$

The probability that $\hat{Y}$ is a τ-counterfactually fair predictor is also expressed accordingly:

$$\mathbb{P}_\Phi\Big(\max_{\substack{s, \mathbf{x}, \hat{y} \\ P(\mathbf{x}, s) > 0}} |\mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y})| \leq \tau\Big) = \int_{\Phi_{\mathrm{CF}}(\tau)} d\mathbb{P}_\Phi.$$

Proof. Let $A_{\hat{y}, \mathbf{x}, s, l, r} = \{\phi \in \Phi : l \leq g_{\hat{y}, \mathbf{x}, s}(\phi) \leq r\}$. This set is measurable because $g_{\hat{y}, \mathbf{x}, s}$ is measurable. By Proposition 3, the corresponding full-space event is $\pi^{-1}(A_{\hat{y}, \mathbf{x}, s, l, r})$. If a full-space prior $\mathbb{P}_\Theta$ is specified and $\mathbb{P}_\Phi$ is taken as the pushforward $\pi_\# \mathbb{P}_\Theta$, then

$$\mathbb{P}_\Theta\big(\pi^{-1}(A_{\hat{y}, \mathbf{x}, s, l, r})\big) = \mathbb{P}_\Phi(A_{\hat{y}, \mathbf{x}, s, l, r}) = \int_{\Phi(\hat{y}, \mathbf{x}, s, l, r)} d\mathbb{P}_\Phi.$$

When $\mathbb{P}_\Phi$ is specified directly on the reduced space, as in the simulation priors, the right-hand side is the reduced-space audit probability under that chosen prior. For τ-counterfactual fairness, define

$$A_\tau = \bigcap_{\substack{(\hat{y}, \mathbf{x}, s) \in \Omega_{\hat{Y}} \times \Omega_{\mathbf{X}} \times \Omega_S \\ P(\mathbf{x}, s) > 0}} \{\phi \in \Phi : -\tau \leq g_{\hat{y}, \mathbf{x}, s}(\phi) \leq \tau\}.$$

The domains are finite, so A_τ is a measurable finite intersection and equals $\Phi_{\mathrm{CF}}(\tau)$. Integrating over A_τ gives the stated reduced-space expression for the probability of τ-counterfactual fairness. □

C SYNTHETIC DATASET SPECIFICATIONS

In this section, we provide detailed specifications of the synthetic datasets $\mathcal{D}_1$, $\mathcal{D}_2$ used in our simulation studies. Both datasets were generated in Python, with the underlying SCMs explicitly designed such that the counterfactual intervention on S induces a heterogeneous effect on the predictor $\hat{Y}$. The equality-constraint matrices used by the sampler are constructed programmatically from the corresponding Q-factor constraints, and the initial feasible point $\phi^{(0)}$ is set to the Chebyshev center computed by `hopsy`.

Let $\sigma(t) = 1/(1 + \exp(-t))$.

Data Generating Process for $\mathcal{D}_1$ The SCM used to generate $\mathcal{D}_1$ is parameterized as follows:

$$
\begin{aligned}
S &\sim \mathrm{Binomial}(1, 0.5) & U_{X\hat{Y}} &\sim \mathrm{Normal}(0,1)\\
X &\sim \mathrm{Binomial}(1, \rho_X) & \rho_X &= \sigma(-0.5 + 1.0U_{X\hat{Y}} + 1.0S + 0.5S\cdot U_{X\hat{Y}})\\
\hat{Y} &\sim \mathrm{Binomial}(1, \rho_{\hat{Y}}) & \rho_{\hat{Y}} &= \sigma(-1.0 - 1.0U_{X\hat{Y}} + 1.5X + 0.5S + 1.0S\cdot X - 0.5S\cdot U_{X\hat{Y}})
\end{aligned}
$$

Data Generating Process for $\mathcal{D}_2$ The SCM used to generate $\mathcal{D}_2$ is parameterized as follows:

$$
\begin{aligned}
S &\sim \mathrm{Binomial}(1, 0.5) & U_{X_1\hat{Y}}, U_{X_2} &\sim \mathrm{Normal}(0,1)\\
X_1 &\sim \mathrm{Binomial}(1, \rho_{X_1}) & \rho_{X_1} &= \sigma(-0.5 + 1.5U_{X_1\hat{Y}} - 2.5S\cdot U_{X_1\hat{Y}} + S)\\
X_2 &\sim \mathrm{Binomial}(1, \rho_{X_2}) & \rho_{X_2} &= \sigma(-1.0 + 0.5U_{X_2} + 2.0S - 1.5X_1 + 3S\cdot X_1)\\
\hat{Y} &\sim \mathrm{Binomial}(1, \rho_{\hat{Y}}) & \rho_{\hat{Y}} &= \sigma(-0.5 - 1.5U_{X_1\hat{Y}} + 3.5U_{X_1\hat{Y}}\cdot X_2 + 2.0X_2)
\end{aligned}
$$

The exogenous variables $U_{X_1\hat{Y}}$ and U_{X_2} are sampled independently. For $\mathcal{D}_2$, empirical right-hand sides of the equality constraints are projected by least squares to a consistent right-hand side before sampling to avoid finite-sample infeasibility. Specifically, for each reduced C-component block with empirical constraint vector $\hat{b}_j$ and constraint matrix A_j, we solve

$$
\tilde{\phi}_j \in \arg\min_{\phi_j \geq 0,\ \mathbf{1}^\top\phi_j = 1} \|A_j\phi_j - \hat{b}_j\|_2^2, \qquad \tilde{b}_j = A_j\tilde{\phi}_j,
$$

and use $\tilde{b}_j$ as the right-hand side for that block. This projection only corrects sampling noise in the empirical constraints; it does not change the causal graph or the target counterfactual query.

Explicit reduced constraint systems For concreteness, we give the explicit reduced equality and nonnegativity systems ($A_1\phi = b_1$, $A_2\phi \leq b_2$ in Lemma 2) used by the sampler. All endogenous variables are binary.

For $\mathcal{D}_1$ (edges $S \to X$, $S \to \hat{Y}$, $X \to \hat{Y}$, $X \leftrightarrow \hat{Y}$), the maximal C-components are $\{S\}$ and $\{X, \hat{Y}\}$. By the reduced C-parameterization (Definition 8) the block for $\{S\}$ is dropped, leaving the reduced parameter $\phi \in \mathbb{R}^{64}$ (4 response maps for X given S times 16 maps for $\hat{Y}$ given S, X). The system $A_1\phi = b_1$ has 9 rows: one observational constraint for each $(s, x, \hat{y}) \in \{0,1\}^3$, plus one simplex-normalization row $\mathbf{1}^\top\phi = 1$, so $A_1 \in \mathbb{R}^{9\times 64}$; nonnegativity is $A_2 = -I_{64}$, $b_2 = \mathbf{0}$.

For $\mathcal{D}_2$ (edges $S \to X_1$, $S \to X_2$, $X_1 \to X_2$, $X_2 \to \hat{Y}$, $X_1 \leftrightarrow \hat{Y}$), the C-components are $\{S\}$, $\{X_1, \hat{Y}\}$, and $\{X_2\}$. Dropping $\{S\}$ leaves $\phi = (\phi_{X_1\hat{Y}}, \phi_{X_2}) \in \mathbb{R}^{32}$ with $\phi_{X_1\hat{Y}} \in \mathbb{R}^{16}$ and $\phi_{X_2} \in \mathbb{R}^{16}$. The matrix A_1 is block-diagonal of size 26×32: the $\{X_1, \hat{Y}\}$ block contributes 17×16 rows (16 observational constraints over $(s, x_1, x_2, \hat{y}) \in \{0,1\}^4$ plus one normalization), and the $\{X_2\}$ block contributes 9×16 rows (8 constraints over $(s, x_1, x_2) \in \{0,1\}^3$ plus one normalization); again $A_2 = -I_{32}$, $b_2 = \mathbf{0}$. The indexing maps and the matrices A_1, b_1 for both datasets are produced by the constraint-construction routines in the released `simulation1.py` and `simulation2.py`.

D REAL-WORLD LSAC CASE STUDY

This section reports an additional real-world case study on the Law School Admissions Council (LSAC) data, following the law-school example studied by Kusner et al. [2017]. This case study illustrates how the same auditing pipeline can be applied to a fixed predictor on a real dataset and a causal graph.

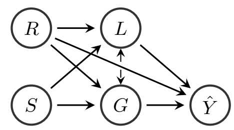


Figure 5: Audit graph for the LSAC case study. The variable $\hat{Y}$ is the output of the deterministic predictor $h(R, L, G)$.

Variables and preprocessing We use five binary variables for preprocessing and predictor training. Race is encoded as $R = 1$ for White and $R = 0$ for non-White; unlike the White/Black-only filter used in some LSAC analyses, the non-White group here pools the remaining recorded race categories after filtering. Sex is encoded as $S = 1$ for male and $S = 0$ for female. LSAT L, undergraduate GPA G, and first-year average Y are median-split, with $L = 1$, $G = 1$, and $Y = 1$ indicating values above the corresponding thresholds. The outcome Y is used only to train the predictor; the audit query below evaluates the fixed classifier $\hat{Y} = h(R, L, G)$ and conditions on (R, S, L, G). The resulting filtered dataset contains 21,791 observations. The median thresholds used by the run are LSAT > 37.000, UGPA > 3.300, and FYA > 0.090.

Audit graph and predictor We consider a hypothetical post-hoc auditing scenario in which Sex is treated as the sensitive attribute and the deployed model is trained under a fairness-through-unawareness design with respect to Sex. The purpose of the audit is to examine whether this predictor is nevertheless counterfactually fair with respect to Sex, since Sex may still affect the prediction indirectly through the measured covariates L and G.

The audit graph is described in Figure 5. In this graph, Sex is not a direct parent of the predictor, but it can affect the prediction through the measured covariates L and G. We train a logistic regression classifier to predict the binarized outcome Y from (R, L, G) only, and set $\hat{Y} = h(R, L, G)$ to be the hard binary prediction. The learned deterministic predictor is shown in Table 1.

Table 1: Learned deterministic predictor for the LSAC case study.

R	L	G	$h(R, L, G)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

This predictor achieves training accuracy 0.607 and prediction rate $P(\hat{Y} = 1) = 0.578$ on the binarized data. For $R = 0$, the learned deterministic predictor is identically zero over all (L, G) values; hence the Sex-flip counterfactual effect is structurally zero in the non-White cells under this fixed classifier. The nonzero violations reported below are therefore driven by the $R = 1$ cells.

Reduced C-parameterization The maximal C-components of the audit graph are $\{R\}$, $\{S\}$, $\{L, G\}$, and $\{\hat{Y}\}$. Since the only bidirected edge is $L \leftrightarrow G$, the only non-degenerate reduced C-parameter block that must be sampled is the block for $\{L, G\}$. The predictor $\hat{Y}$ is a fixed deterministic function $h(R, L, G)$, so the $\{\hat{Y}\}$ block is known and degenerate rather than sampled.

For the component $\{L, G\}$, we have $\mathrm{Pa}_L = \mathrm{Pa}_G = \{R, S\}$. Since all variables are binary, $|\Omega_{U_L}| = |\Omega_{U_G}| = 16$ and thus $|\Omega_{\mathbf{U}_{\{L,G\}}}| = 256$. We enumerate the values of $\mathbf{U}_{\{L,G\}}$ as follows:

$$\Omega_{\mathbf{U}_{\{L,G\}}} = \left\{ \mathbf{u}^{(1)}_{\{L,G\}}, \ldots, \mathbf{u}^{(256)}_{\{L,G\}} \right\}.$$

For each k, write $u_L^{(k)} \in \Omega_{U_L}$ and $u_G^{(k)} \in \Omega_{U_G}$ for the U_L and U_G elements of $\mathbf{u}^{(k)}_{\{L,G\}}$, respectively. The sampled reduced C-parameter is

$$\phi_{\mathcal{M}} = \left[P_{\mathcal{M}}(\mathbf{u}^{(1)}_{\{L,G\}}), \ldots, P_{\mathcal{M}}(\mathbf{u}^{(256)}_{\{L,G\}}) \right]^\top \in \mathbb{R}^{256}.$$

The sampled reduced space Φ consists of vectors of this form satisfying the following constraints:

$$\sum_{k=1}^{256} I(l;(r,s),u_L^{(k)})I(g;(r,s),u_G^{(k)})P_{\mathcal{M}}(\mathbf{u}_{\{L,G\}}^{(k)}) = P(l,g \mid r,s), \qquad (r,s,l,g) \in \{0,1\}^4,$$

$$\sum_{k=1}^{256} P_{\mathcal{M}}(\mathbf{u}_{\{L,G\}}^{(k)}) = 1,$$

$$P_{\mathcal{M}}(\mathbf{u}_{\{L,G\}}^{(k)}) \geq 0, \qquad k = 1,\ldots,256.$$

On the LSAC data all (R,S) strata have support; the resulting equality-constraint matrix has 17 rows and rank 13.

Sex-flip counterfactual query For a supported factual cell (r,s,l,g), the audit intervenes on Sex by setting $S = \neg s = 1-s$ while holding Race fixed at $R = r$. For notational brevity, write

$$L_k(r,s) = g_L^{(u_L^{(k)})}(r,s), \qquad G_k(r,s) = g_G^{(u_G^{(k)})}(r,s),$$

where $\mathbf{u}_{\{L,G\}}^{(k)} = (u_L^{(k)}, u_G^{(k)})$ indexes a response-function pair for the C-component $\{L,G\}$. Let $\phi_{\mathcal{M},k}$ be the kth coordinate of the reduced parameter $\phi_{\mathcal{M}}$ defined above. The sex-flip counterfactual prediction probability for the event $\hat{Y} = 1$ can be expressed as

$$P_{\mathcal{M}}(\hat{Y}_{\neg s} = 1 \mid r,s,l,g) = \frac{\sum_{k=1}^{256} \mathbb{I}\big(l;(r,s),u_L^{(k)}\big)\mathbb{I}\big(g;(r,s),u_G^{(k)}\big)h\big(r, L_k(r,\neg s), G_k(r,\neg s)\big)\phi_{\mathcal{M},k}}{P(l,g \mid r,s)}.$$

Because the factual predictor is deterministic, $P_{\mathcal{M}}(\hat{Y} = 1 \mid r,s,l,g) = h(r,l,g)$. Therefore, the counterfactual effect is

$$\mathrm{CE}_{\mathcal{M}}(s,\neg s \mid (r,l,g);1) = P_{\mathcal{M}}(\hat{Y}_{\neg s} = 1 \mid r,s,l,g) - h(r,l,g).$$

Since $\hat{Y}$ is binary, the counterfactual effect for $\hat{Y} = 0$ is the negative of the effect for $\hat{Y} = 1$. Hence the fixed predictor is τ-counterfactually fair in this audit if

$$\max_{\substack{r,s,l,g \\ P(r,s,l,g)>0}} |\mathrm{CE}_{\mathcal{M}}(s,\neg s \mid (r,l,g);1)| \leq \tau.$$

Sampling and diagnostics We ran Hit-and-Run on the 256-coordinate C-parameter polytope, whose affine dimension is 243 after the rank-13 equality constraints, for two priors: a uniform prior over the polytope and a Dirichlet prior with concentration parameter $\alpha = 2.5$. Each run used four chains, 50,000 retained samples per chain, burn-in 5,000, and thinning 1,000. We report split Gelman–Rubin $\hat{R}$ and effective sample size (ESS) diagnostics for the sampled coordinates, not direct diagnostics for the threshold events. The event probabilities in Table 2 are Monte Carlo estimates under the stated priors; in particular, a zero estimate means that no retained sample satisfied the threshold, not that infeasibility was certified by optimization.

Table 2: LSAC Sex-flip audit for a fixed deterministic predictor. Here $\widehat{P}_\pi(\tau)$ denotes the Monte Carlo estimate, under prior π, of $P_\pi(\max|\mathrm{CE}| \leq \tau)$. The table also reports prior-conditional summaries of $\max|\mathrm{CE}|$ and coordinate-wise MCMC diagnostics. The "Sample max" column is the largest value observed among retained MCMC samples, not an LP-certified upper bound.

Prior	$\widehat{P}_\pi(0.10)$	$\widehat{P}_\pi(0.25)$	$\widehat{P}_\pi(0.50)$	Mean $\max|\mathrm{CE}|$	Sample max	$\hat{R}_{\max}$ / $\mathrm{ESS}_{\min}$
Uniform	0.0000	0.0000	0.0085	0.7110	0.9446	1.052 / 149
Dirichlet ($\alpha = 2.5$)	0.0000	0.0000	0.0008	0.7100	0.9083	1.046 / 154

The largest witnessed counterfactual effects under both priors occurred in the cell $(R,S,L,G,\hat{Y}) = (1,0,0,0,0)$, that is, White, female, below-median LSAT, below-median UGPA, and predicted negative. This cell has empirical support 3519. The maximum witnessed values of $|\mathrm{CE}|$ were 0.9446 under the uniform prior and 0.9083 under the Dirichlet prior. A prior-free LP bound for the same cell gives $0.1562 \leq |\mathrm{CE}_\phi(1,0,0,0)| \leq 1.0000$ over the compatible polytope. Thus τ-counterfactual fairness is LP-certified to fail for $\tau < 0.1562$, while the upper bound remains too loose to summarize the typical compatible SCM. Under this audit graph, binarization, deterministic classifier, and pair of stated priors, these results illustrate that removing Sex from the predictor input does not, by itself, certify counterfactual fairness: Sex may still affect the prediction through the covariates L and G.