Evaluating Robustness to Dataset Shift via Parametric Robustness Sets

Anonymous Author(s)
Affiliation
Address
email

Abstract

We give a method for proactively identifying small, plausible shifts in distribution which lead to large differences in model performance. To ensure that these shifts are plausible, we parameterize them in terms of interpretable changes in causal mechanisms of observed variables. This defines a parametric robustness set of plausible distributions and a corresponding worst-case loss. While the loss under an individual parametric shift can be estimated via reweighting techniques such as importance sampling, the resulting worst-case optimization problem is non-convex, and the estimate may suffer from large variance. For small shifts, however, we can construct a local second-order approximation to the loss under shift and cast the problem of finding a worst-case shift as a particular non-convex quadratic optimization problem, for which efficient algorithms are available. We demonstrate that this second-order approximation can be estimated directly for shifts in conditional exponential family models, and we bound the approximation error. We apply our approach to a computer vision task (classifying gender from images), revealing sensitivity to shifts in non-causal attributes.

1 Introduction

Predictive models may perform poorly outside of the training distribution, a problem broadly known as dataset shift [Quiñonero-Candela et al., 2008]. In high-stakes applications, such as healthcare, it is important to understand the limitations of a model in advance [Finlayson et al., 2021]: given a model trained on data from one hospital, how will it perform under changes in the population of patients, in the incidence of disease, or in the treatment policy?

In this paper, our goal is to **proactively** understand the sensitivity of a predictive model to dataset shift, using only data from the training distribution. This requires domain knowledge, to specify what type of distributional changes are plausible. Formally, for a model $f(X)$ trained on data from $\mathbb{P}(X, Y)$, with loss function $\ell(f(X), Y)$, we seek to understand the loss of the model under a set of *plausible* future distributions $\mathcal{P}$. We seek to evaluate the worst-case loss over $\mathcal{P}$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(f(X), Y)], \tag{1}$$

and provide an interpretable description of a distribution P which maximizes this objective. If the value of the worst-case loss is low, this can build confidence prior to deployment, and otherwise, examining the worst-case distribution P can help identify weaknesses of the model. To illustrate, we use the following running example, inspired by Subbaswamy et al. [2021].

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

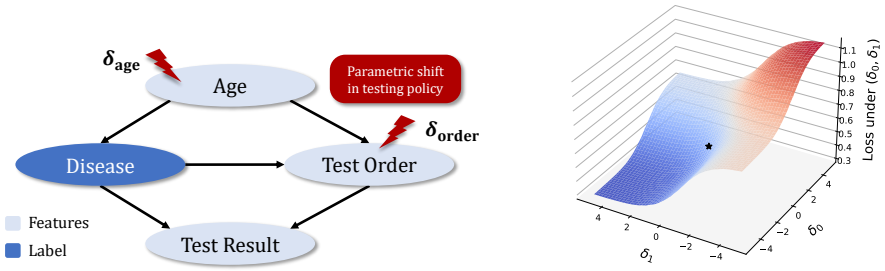

Figure 1: (Left) Causal graph for Example 1. Our approach allows for simultaneous shifts in age and test ordering, parameterized by $\delta_{\text{age}}, \delta_{\text{order}}$. (Right) We illustrate a shift only in testing rates, using $s(Y; \delta_{\text{order}}) = \delta_1 \cdot Y + \delta_0(1 - Y)$, where $\delta_{\text{order}} = (\delta_0, \delta_1)$. Here we plot the (non-concave) landscape of the expected cross-entropy loss of a fixed model over distributions parameterized by (δ_0, δ_1), with the training distribution given as the black star. Simulation details are given in Appendix A.

Example 1 (Changes in laboratory testing)**.** We seek to classify disease (Y) based on the age (A) of a patient, whether a lab test has been ordered (O), and test results (L) if ordered. The performance of a predictive model may be sensitive to changes in testing policies, as the *fact that a test has been ordered* itself is predictive of disease. Figure 1 (left) gives a plausible causal relationship between variables. Let $\mathbb{P}(O|A, Y) = \sigma(\eta(A, Y))$, where σ is the sigmoid function and $\eta(A, Y)$ is the log-odds. In Fig. 1 (right), we show the loss under a set of new distributions parameterized by $\delta = (\delta_0, \delta_1)$, where we modify $\mathbb{P}_\delta(O|A, Y) = \sigma(\eta(A, Y) + s(Y; \delta))$ for a *shift function* $s(Y; \delta) = \delta_1 \cdot Y + \delta_0 \cdot (1 - Y)$, which modifies the log-odds of testing for both sick and healthy patients. If δ_0, δ_1 are unconstrained, the worst-case occurs when all healthy patients are tested, and no sick patients are tested.

The first challenge is to define a set of possible distributions $\mathcal{P}$ such that each distribution $P \in \mathcal{P}$ satisfies two desiderata: First, they should be *causally interpretable and simple to specify*, without placing unnecessary restrictions on the data-generating process. Second, they should be *realistic*, which often entails bounding the magnitude of the shift. We construct causally interpretable shifts by defining perturbed distributions $\mathbb{P}_\delta$ using changes in causal mechanisms, parameterized by a finite-dimensional parameter δ. Our main requirement is that the shifting mechanisms follow a conditional exponential family distribution. For discrete variables, this places no restriction on $\mathbb{P}$: In Example 1, O is binary and the log-odds $\eta(A, Y)$ can be any function of A, Y. We also demonstrate that constraining δ can ensure that shifts are realistic: The unconstrained worst-case shift in Example 1 is implausible, where all healthy patients (and no sick patients) are tested. Equation (1) becomes

$$\sup_{\delta \in \Delta} \mathbb{E}_\delta[\ell(f(X), Y)], \tag{2}$$

where $\mathbb{E}_\delta$ is the expectation in the shifted distribution $\mathbb{P}_\delta$ and Δ is a bounded set of shifts.

The second challenge is evaluation of the expected loss under shift, as well as finding the worst-case shift. Under our definition of shifts, we show that the test distribution can always be seen as a reweighting of the training distribution, allowing for reweighting approaches, such as importance sampling, to estimate the expected loss under shifts. While this is practical for some distribution shifts, for others, importance sampling can lead to extreme variance in estimation. Further, finding the worst-case shift using a reweighted objective involves maximization over a non-concave objective (see Fig. 1), a problem that is generally NP-hard. We derive a second-order approximation to the expected loss under shift, and show how it can be estimated without the use of reweighting. For quadratic constraints Δ, we can approximate the general non-convex optimization problem in Equation 2 with a non-convex, quadratically constrained quadratic program (QCQP) for which efficient solvers exist [Conn et al., 2000, Section 7]. We bound the approximation error of this surrogate objective, and show in experiments that it tends to find impactful adversarial shifts.

Our contributions are as follows:

1. We provide a novel formulation of robustness sets which are defined using parametric shifts. This formulation only require that the shifting mechanisms can be modelled as a conditional exponential family (see Section 2).

2. We derive a second-order approximation to the expected loss and provide a bound on the approximation error. We show that this translates the general non-convex problem into a particular non-convex quadratic program, for which efficient solvers exist (see Section 3).
3. In a computer vision task, we find that this approach finds more impactful shifts than a reweighting approach, while taking far less time to compute, and that the resulting estimates of accuracy are an order of magnitude more reliable in predicting the shifted (see Section 4).

Related work: A recent line of work learns predictive models that minimize objectives similar to Eq. (1), a task known as distributionally robust optimization [Duchi and Namkoong, 2021, Duchi et al., 2020, Sagawa et al., 2020]. Beyond our difference in motivation (we consider evaluation of a fixed model, not optimization), prior work typically defines the robustness set using a notion of distance between distributions. In contrast, we consider sets of distributions arising from explicit parametric perturbations. Moreover, many of these approaches focus on changes in marginal distributions, while one of our primary motivations is handling conditional shifts. Closer to our work is Subbaswamy et al. [2021] who consider evaluating the loss under worst-case changes in a conditional distributions, but while we consider parametric shifts, they estimates the loss under worst-case $(1-\alpha)$ conditional subpopulation shifts. In some settings, such shifts may not represent plausible changes, as we demonstrate in in Appendix D, where (in a simplified lab-testing example) the worst-case subpopulation is one where healthy patients are always tested, and sick patients never tested. Prior work on robustness to parametric interventions has been restricted to linear causal models with additive shift interventions [Rothenhäusler et al., 2021, Oberst et al., 2021, Kook et al., 2022]. Our work can be seen as extending those ideas to general non-linear causal models, where our focus is on evaluation rather than learning robust models. A separate line of work attempts to learn models with optimal worst-case performance under arbitrary causal interventions of various forms [Magliacane et al., 2018, Rojas-Carulla et al., 2018, Arjovsky et al., 2019, Subbaswamy et al., 2022] while we focus on explicitly bounded interventions that have potentially restricted parameterizations.

2 Defining parametric robustness sets

Notation: Let $\mathbf{V}$ denote all observed variables, where $(X, Y) \subseteq \mathbf{V}$ for features X and labels Y, and let $\mathbb{P}(\mathbf{V})$ denote the training distribution. $\mathbb{E}[\cdot]$ and $\mathrm{cov}(\cdot,\cdot)$ refer to the mean and covariance in $\mathbb{P}$, and for a shifted distribution $\mathbb{P}_\delta$ (Definition 1) we use $\mathbb{E}_\delta[\cdot]$, $\mathrm{cov}_\delta(\cdot,\cdot)$. For a random variable Z, we use $\mathcal{Z}$ to denote the space of realizations, and d_Z for dimension e.g., $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$. For a set of random variables $\mathbf{V} = \{V_1, \ldots, V_d\}$, we use V_i to denote an individual element, and use $\mathrm{PA}_{\mathcal{G}}(V_i)$ to denote the set of parents in a directed acyclic graph (DAG) $\mathcal{G}$, omitting the subscript when otherwise clear.

We begin with a general definition of a parameterized robustness set of distributions $\mathcal{P}$.

Definition 1. A *parameterized robustness set around* $\mathbb{P}(\mathbf{V})$ is a family of distributions $\mathcal{P}$ with elements $\mathbb{P}_\delta(\mathbf{V})$ indexed by $\delta \in \Delta \subseteq \mathbb{R}^{d_\delta}$, with $0 \in \Delta$, where $\mathbb{P}_0(\mathbf{V}) = \mathbb{P}(\mathbf{V})$.

We give examples shortly that satisfy this general definition. To construct such a robustness set, we consider distributions $\mathbb{P}_\delta$ that differ from $\mathbb{P}$ in one or more conditional distributions (Assumption 1). We require that the relevant conditional distributions can be described by an exponential family.

Definition 2 (Conditional exponential family (CEF) distribution)**.** $\mathbb{P}(W|Z)$ is a conditional exponential family distribution if there exists a function $\eta(Z) : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_T}$ such that the conditional probability density (for continuous W) or probability mass function (for discrete W) is given by

$$\mathbb{P}(W|Z) = g(W) \exp\left(\eta(Z)^\top T(W) - h(\eta(Z))\right), \tag{3}$$

where $T(W)$ is a vector of sufficient statistics, $T(W) \in \mathbb{R}^{d_T}$, $g(\cdot)$ specifies the density of a base measure and $h(\eta(Z))$ is a normalizing constant.

Definition 2 does not restrict $\mathbb{P}(W|Z)$ for binary/categorical W, and captures a wide range of distributions, including the conditional Gaussian (see Appendix B.1 for other examples). Definition 2 extends to marginal distributions where $Z = \varnothing$ and $\eta(Z)$ is a constant function.

Example 1 (Continued)**.** Suppose the probability of ordering a test (O) depends on age (A) and disease (Y), such that $\mathbb{P}(O = 1|A, Y) = \sigma(\eta(A, Y))$, where σ is the sigmoid, and η is an arbitrary function. Here, Definition 2 is satisfied with $W = O$, $Z = (A, Y)$, and sufficient statistic $T(O) = O$.

We now state our main assumption, where we distinguish between the terms in the joint distribution of $\mathbb{P}$ that shift, which we will need to model, and those that remain fixed, which we do not.

Assumption 1 (Factorization into CEF distributions)**.** Let $\mathbf{W} = \{W_1, \ldots, W_m\} \subseteq \mathbf{V}$ be a "intervention set" of variables and let

$$\mathbb{P}(\mathbf{V}) = \left(\prod_{W_i \in \mathbf{W}} \underbrace{\mathbb{P}(W_i|Z_i)}_{\text{Conditionals that shift}} \right) \left(\prod_{V_j \in \mathbf{V} \setminus \mathbf{W}} \underbrace{\mathbb{P}(V_j|U_j)}_{\text{Conditionals that we do not model}} \right) \tag{4}$$

be a factorization, where $Z_i, U_j, V_j \subseteq \mathbf{V}$ are possibly overlapping sets of variables. We assume for each W_i that Z_i is known and that $\mathbb{P}(W_i|Z_i)$ satisfies Definition 2.

If $\mathbb{P}(\mathbf{V})$ factorizes according to a DAG $\mathcal{G}$, the factorization in Assumption 1 is always satisfied by $Z_i = \mathrm{PA}_{\mathcal{G}}(W_i)$. Here, we require limited knowledge of the underlying graph, and only need to know the parents $\mathrm{PA}(W_i)$ for the variables W_i that shift. In Appendix B.2 we show that we can also consider shifts that extend Z_i to include additional variables, subject to an acyclicity constraint. We now define parametric perturbations and give the general form of the robustness sets that we consider in this work, involving simultaneous perturbations to multiple W_i.

Definition 3 (Parameterized shift functions and δ-perturbations)**.** Let $s(Z; \delta) : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_T}$ be a *parameterized shift function* with parameters $\delta \in \Delta \subseteq \mathbb{R}^{d_\delta}$ which is twice-differentiable with respect to δ and satisfy $s(Z; 0) = 0$ for all Z. For $\mathbb{P}(W|Z)$ satisfying Eq. (3), we refer to

$$\mathbb{P}_\delta(W|Z) = g(W) \exp\left(\eta_\delta(Z)^\top T(W) - h(\eta_\delta(Z))\right)$$

as a δ-perturbation of $\mathbb{P}(W|Z)$ with shift function $s(Z; \delta)$, where $\eta_\delta(Z) := \eta(Z) + s(Z; \delta)$.

Example 1 (Continued)**.** A model developer may be concerned about a uniform change in testing rates across all types of patients. This can be modelled by choosing $s(Z; \delta) = \delta$, for $\delta \in \mathbb{R}$, an additive intervention on the log-odds scale. A separate change in testing rates for sick and healthy could instead be modeled using $s(Z; \delta) = \delta_0(1 - Y) + \delta_1 Y$, using $\delta \in \mathbb{R}^2$. This reasoning extends readily to more complex shifts (e.g., allowing for age-specific changes in testing rates, with a non-linear dependence on age), as long as $s(Z; \delta)$ remains a parametric function.

While the shift function $s(Z; \delta)$ is parametric, $\eta(Z)$ is unconstrained in Definitions 2 and 3. Note that this formulation includes multiplicative shifts $\eta_\delta(Z) = (1 + \delta)\eta(Z)$ by letting $s(Z; \delta) = \delta \cdot \eta(Z)$.

Definition 4 (CEF parameterized robustness set)**.** For a distribution $\mathbb{P}$ and intervention set $\mathbf{W} = \{W_1, \ldots, W_m\} \subseteq \mathbf{V}$ satisfying Assumption 1, let each $\mathbb{P}_{\delta_i}(W_i|Z_i)$ be a δ_i-perturbation (Definition 3) of $\mathbb{P}(W_i|Z_i)$. Then $\mathbb{P}_\delta(\mathbf{V}) = \left(\prod_{W_i \in \mathbf{W}} \mathbb{P}_{\delta_i}(W_i|Z_i)\right)\left(\prod_{V_j \in \mathbf{V} \setminus \mathbf{W}} \mathbb{P}(V_j|U_j)\right)$ is called a δ-perturbation of $\mathbb{P}(\mathbf{V})$, and the robustness set $\mathcal{P}$ consists of all $\mathbb{P}_\delta$ for $\delta \in \Delta_1 \times \cdots \Delta_m$.

To estimate the expected loss under $\mathbb{P}_\delta$, we will typically[1] need to estimate $\eta(Z_i)$ for each $W_i \in \mathbf{W}$. However, we make no distributional assumptions on the remaining variables $\mathbf{V} \setminus \mathbf{W}$. This is useful in applications such as computer vision, where we do not need to restrict the generative model of images given attributes (e.g., background, camera type, etc), but can still model the expected loss under changes in the joint distribution of those attributes.

Remark 1 (Causal Interpretation of Shifts). If the DAG $\mathcal{G}$ represents a causal graph [Pearl, 2009], then $\mathbb{P}_\delta$ can interpreted as a change in causal mechanisms. We see this as an important perspective for interpreting and specifying plausible shifts, but our methods do not require a causal interpretation.

[1] As a special case, in Appendix C.2, we show the second-order approximation (Theorem 1) can be estimated in the case of variance-scaled mean-shifts in a conditional Gaussian without estimation of all of $\eta(Z)$.

3 Evaluation of the worst-case loss

For a fixed predictor and loss function, we can use data from $\mathbb{P}(\mathbf{V})$ to estimate the expected loss $\mathbb{E}_\delta[\ell] := \mathbb{E}_\delta[\ell(f(X), Y)]$ for a fixed δ, and estimate the worst-case loss over all δ of bounded magnitude. In Section 3.1, we show that $\mathbb{P}_\delta$ shares support with $\mathbb{P}$, suggesting the use of reweighting estimators. However, these can exhibit high variance for shifts that produce large density ratios (see Appendix C.5 for an example), and moreover, maximizing a reweighted objective over δ is generally a non-convex problem. In Section 3.2 we derive an approximation to the expected loss under $\mathbb{P}_\delta$, yielding a tractable surrogate optimization problem under quadratic constraints such as $\|\delta\|_2 \leq \lambda$.

Remark 2. The methods here can be used with an arbitrary predictor f and loss function $\ell := \ell(f(X), Y)$. We do not even require access to the original predictor f. Both methods here simply treat ℓ as a random variable in $\mathbb{P}$, for which we have samples from the training distribution.

3.1 Modelling shifted losses using reweighting

The shifts defined in Section 2 share common support, with the following density ratio.

Proposition 1. *For any $\mathbb{P}_\delta(\mathbf{V}), \mathbb{P}(\mathbf{V})$ that satisfy Definition 4, $\operatorname{supp}(\mathbb{P}) = \operatorname{supp}(\mathbb{P}_\delta)$ and the density ratio $w_\delta := \mathbb{P}_\delta / \mathbb{P}$ is given by*

$$w_\delta(\mathbf{V}) = \exp\left(\sum_{i=1}^m s_i(Z_i;\delta_i)^\top T_i(W_i)\right) \exp\left(\sum_{i=1}^m h(\eta_i(Z_i)) - h(\eta(Z_i) + s_i(Z_i;\delta_i))\right).$$

The proof can be found in Appendix F, along with all proofs for all other claims.

Example 1 (Continued)**.** Suppose we perturb the probability of ordering a test O given age A and disease Y with shift function $s(Y;\delta) = \delta_0(1-Y) + \delta_1 Y$, independently changing the conditional probability of testing for healthy and sick patients. Here, the density ratio is given by

$$w_\delta(O, A, Y) = \exp(s(Y;\delta)\cdot O)\frac{1+\exp(\eta(A,Y))}{1+\exp(\eta(A,Y)+s(Y;\delta))}. \tag{5}$$

To model the loss $\mathbb{E}_\delta[\ell]$ using data from $\mathbb{P}$, we can consider an importance sampling (IS) estimator [Horvitz and Thompson, 1952, Shimodaira, 2000], observing that $\mathbb{E}_\delta[\ell] = \mathbb{E}[w_\delta(\mathbf{V})\cdot\ell]$. This requires estimation of the density ratio $w_\delta(\mathbf{V})$, and (given a sample $\{\mathbf{V}^j\}_{j=1}^n$ from $\mathbb{P}$) yields the estimator

$$\mathbb{E}_\delta[\ell] \approx \hat{E}_{\delta,\text{IS}} := \frac{1}{n}\sum_{j=1}^n \hat{w}_\delta(\mathbf{V}^j)\ell(\mathbf{V}^j). \tag{6}$$

In practice, Eq. (6) can have high variance when density ratios are large, and maximizing this equation with respect to δ is a general non-convex optimization problem, which is generally NP-hard to solve.

3.2 Approximating the shifted loss for exponential family models

We now propose an alternative approach for approximating the loss $\mathbb{E}_\delta[\ell]$. Recalling that $\mathbb{P}_{\delta=0} = \mathbb{P}$, we use a second-order Taylor expansion around the training distribution

$$\mathbb{E}_\delta[\ell] \approx \mathbb{E}[\ell] + \delta^\top \operatorname{SG}^1 + \tfrac{1}{2}\delta^\top \operatorname{SG}^2 \delta, \tag{7}$$

where $\mathbb{E}[\ell]$ denotes the loss in the training distribution and $\operatorname{SG}^1, \operatorname{SG}^2$ are defined as follows.

Definition 5 (Shift gradient and Hessian)**.** For a parametric shift satisfying Definition 1 where $\delta \mapsto \mathbb{E}_\delta[\ell]$ is twice-differentiable, we denote the *shift gradient* SG^1 and *shift Hessian* SG^2 as $\operatorname{SG}^1 := \nabla_\delta \mathbb{E}_\delta[\ell]\big|_{\delta=0}$ and $\operatorname{SG}^2 := \nabla_\delta^2 \mathbb{E}_\delta[\ell]\big|_{\delta=0}$.

Equation (7) is a local approximation of the loss, whose approximation error we bound in Theorem 2, with smaller approximation error for smaller shifts.[2] For $\mathbb{P}_\delta$ satisfying Definition 4, SG^1 and SG^2 can be computed as expectations in the training distribution, without estimation of density ratios. Recall that the conditional covariance is given by $\operatorname{cov}(A, B|C) := \mathbb{E}[(A - \mathbb{E}[A|C])(B - \mathbb{E}[B|C])|C]$.

[2] In Appendix C.3, we give an example of a linear-Gaussian generative model where this second-order expansion is exact, corresponding to the setting of anchor regression [Rothenhäusler et al., 2021].

Theorem 1 (Shift gradients and Hessians as covariances)**.** *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy Definition 4, with intervened variables $\mathbf{W} = \{W_1, \ldots, W_m\}$ and shift functions $s_i(Z_i; \delta_i)$, where $\delta = (\delta_1, \ldots, \delta_m)$. Then the shift gradient is given by $\mathrm{SG}^1 = (\mathrm{SG}^1_1, \ldots, \mathrm{SG}^1_m) \in \mathbb{R}^{d_\delta}$ where*

$$\mathrm{SG}^1_i = \mathbb{E}\left[D_{i,1}^\top \operatorname{cov}\left(\ell,\, T_i(W_i) \middle| Z_i\right)\right],$$

and the shift Hessian is a matrix of size $(d_\delta \times d_\delta)$, where the (i,j)th block of size $d_{\delta_i} \times d_{\delta_j}$ equals

$$\{\mathrm{SG}^2\}_{i,j} = \begin{cases} \mathbb{E}\left[D_{i,1}^\top \operatorname{cov}\left(\ell,\, \epsilon_{T_i|Z_i}\epsilon_{T_i|Z_i}^\top | Z_i\right) D_{i,1}\right] - \mathbb{E}\left[\ell \cdot D_{i,2}^\top \epsilon_{T|Z}\right] & i = j \\ \operatorname{cov}(\ell,\ D_{i,1}^\top \epsilon_{T_i|Z_i}\epsilon_{T_j|Z_j}^\top D_{j,1}) & i \neq j, \end{cases}$$

where $D_{i,k} := \nabla^k_{\delta_i} s_i(Z_i; \delta_i)|_{\delta=0}$, is the gradient of the shift function for $k = 1$, and the Hessian for $k = 2$. Here, $T_i(W_i)$ is the sufficient statistic of $\mathbb{P}(W_i|Z_i)$ and $\epsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}[T(W_i)|Z_i]$.

Theorem 1 handles arbitrary parametric shift functions in multiple variables, but for simple shift functions in a single variable, the notation simplifies substantially, as we show in Corollary 1.

Corollary 1 (Simple shift in a single variable)**.** *Assume the setup of Theorem 1, restricted to a shift in a single variable W, and that $s(Z; \delta) = \delta$. Then $D_1 = 1$, $D_2 = 0$, and*

$$\mathrm{SG}^1 = \mathbb{E}\left[\operatorname{cov}\left(\ell, T(W) \middle| Z\right)\right] \qquad \textit{and} \qquad \mathrm{SG}^2 = \mathbb{E}\left[\operatorname{cov}\left(\ell, \epsilon_{T|Z}\epsilon_{T|Z}^\top \middle| Z\right)\right],$$

where $T(W)$ is the sufficient statistic of W and $\epsilon_{T|Z} := T(W) - \mathbb{E}[T(W)|Z]$.

Example 1 (Continued)**.** Suppose that age (A), which has no causal parents, follows a normal distribution with mean μ and variance σ^2, and that we wish to consider a shift in the mean. We can parameterize $\mathbb{P}(A)$ as an exponential family with parameter $\eta = \mu/\sigma$ and sufficient statistic $T(A) = A/\sigma$. Here, $s(\delta) = \delta$ implies a shift in the mean of δ standard deviations $\eta_\delta = \eta + s(\delta) = (\mu + \sigma\delta)/\sigma$, and we can write that $\mathrm{SG}^1 = \operatorname{cov}(\ell, A)/\sigma$ and $\mathrm{SG}^2 = \operatorname{cov}\left(\ell, (A - \mathbb{E}[A])^2\right)/\sigma^2$.

To estimate the shift gradient and Hessian from a sample from $\mathbb{P}$, for each $i = 1, \ldots, m$ we fit models $\hat{\mu}_\ell(Z_i) \approx \mathbb{E}[\ell|Z_i]$ and $\hat{\mu}_{W_i}(Z_i) \approx \mathbb{E}[T_i(W_i)|Z_i]$ and compute residuals on these predictions, which permits estimation of the gradient/Hessian as a sample average of residuals. A detailed treatment is given in Appendix C.1. Using these, we can estimate the expected loss as

$$\mathbb{E}_\delta[\ell] \approx \hat{E}_{\delta,\text{Taylor}} := \hat{\mathbb{E}}[\ell] + \delta^\top \hat{\mathrm{SG}}^1 + \frac{1}{2}\delta^\top \hat{\mathrm{SG}}^2 \delta. \tag{8}$$

Here, there are two sources of error: Finite-sample error, due to the estimates of $\mathrm{SG}^1, \mathrm{SG}^2$, as well as approximation error. The latter is bounded by the norm of δ and a term that depends on the covariance between the loss and the deviations of the sufficient statistic from its shifted mean.

Theorem 2. *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy the conditions of Theorem 1, with a shift in a single variable W, where $s(Z; \delta) = \delta$. Let $E_{\delta,\textit{Taylor}}$ be the population Taylor estimate (Eq. (7)) and let $\sigma(M)$ denote the largest absolute value of the eigenvalues of a matrix M. Then*

$$\left|\mathbb{E}_\delta[\ell] - E_{\delta,\textit{Taylor}}\right| \leq \tfrac{1}{2} \sup_{t \in [0,1]} \sigma\Big(\operatorname{cov}_{t\cdot\delta}(\ell, \epsilon_{t\cdot\delta,T|W}\epsilon_{t\cdot\delta,T|W}^\top) - \operatorname{cov}(\ell, \epsilon_{0,T|W}\epsilon_{0,T|W}^\top)\Big) \cdot \|\delta\|^2,$$

where $T(W)$ is the sufficient statistic of $W|Z$ and $\epsilon_{t\cdot\delta,T|W} = T(W|Z) - \mathbb{E}_{t\cdot\delta}[T(W|Z)]$.

In exchange for considering a second-order approximation of the loss, we gain two benefits: Variance reduction and tractable optimization. First, as $\mathrm{SG}^1, \mathrm{SG}^2$ are not functions of δ, the variance of $\hat{E}_{\delta,\text{Taylor}}$ is $O(\|\delta\|^4)$, while the variance of $\hat{E}_{\delta,\text{IS}}$ can be much larger: We give a simple case in Appendix C.6 where $\operatorname{var}(\hat{E}_{\delta,\text{Taylor}}) = O(\delta^4)$ while $\operatorname{var}(\hat{E}_{\delta,\text{IS}}) = O(\delta^2 \exp(\delta^2))$. Second, maximizing $\hat{E}_{\delta,\text{Taylor}}$ over the set $\|\delta\| \leq \lambda$ can be solved in polynomial time by exploiting the quadratic structure, while maximizing $\hat{E}_{\delta,\text{IS}}$ over the constraints is generally hard, and may be infeasible in high dimensions.

3.3 Identifying worst-case parametric shifts

For $\lambda > 0$, we can locally approximate the worst-case loss over all distributions $\mathbb{P}_\delta$ where $\|\delta\|_2 \leq \lambda$ by finding the worst-case loss in the Taylor approximation

$$\sup_{\|\delta\|_2 \leq \lambda} \delta^\top \mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \mathrm{SG}^2\, \delta. \tag{9}$$

Since SG^2 is generally not negative definite, the maximization objective is non-concave. However, this particular problem is an instance of the 'trust region problem'[3] which is well-studied in the optimization literature [Conn et al., 2000], and can be solved in polynomial time by specialized algorithms (see Pólik and Terlaky [2007, Section 8.1] for an example). This follows from the fact that strong duality holds, so that the optimal solution δ^* can be characterized in terms of the Karush-Kuhn-Tucker conditions [Boyd and Vandenberghe, 2004, Section 5.2]. For this problem, we use the `trsapp` routine from NEWUOA [Powell, 2006], as implemented in the python package `trustregion`. Depending on the application and prior knowledge, one may choose constraint sets that differ from $\|\delta\| \leq \lambda$. The strong duality of Equation (9) holds for any quadratic constraint $\delta^\top A\delta + \delta^\top b \leq \lambda$, making it a general non-convex quadratically constrained quadratic program (QCQP), allowing for e.g., larger shifts in some directions than in others.

4 Experiments

4.1 Illustrative example: Laboratory testing

To build intuition, we illustrate our method in a simple case. Motivated by Example 1, we use a simple generative where lab tests are more likely to be ordered (O) for sick patients (Y), and lab values (L) are predictive of Y.

$$Y \sim \mathsf{Ber}(0.5) \qquad O|Y \sim \mathsf{Ber}(\sigma(\alpha + \beta Y)) \qquad L|Y, O = 1 \sim \mathcal{N}(\mu_y, 1)$$

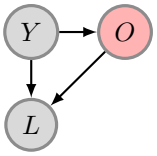

Figure 2

where $\mu_1 = 0.5, \mu_0 = -0.5$, and we initialize with $\alpha = -1$, $\beta = 2$, so that $\mathbb{P}(O = 1|Y = 0) \approx 0.27$ and $\mathbb{P}(O = 1|Y = 1) \approx 0.73$, and the marginal probability of test ordering is $\mathbb{P}(O = 1) = 0.5$. When $O = 0$, we set L to a dummy value of $L = 0$. The underlying causal graph is given in Fig. 2. The predictive model $f(O, L)$ for this example is trained on data from $\mathbb{P}$ to predict Y using all available features. If lab tests are not available ($O = 0$), this model predicts Y based on the observed likelihood of Y given $O = 0$, and otherwise predicts using a logistic regression model trained on the cases where $O = 1$ in the training data.

Defining a shift function: $\mathbb{P}(O|Y)$ is a conditional exponential family with $\eta(Y) = \alpha + \beta Y$. We consider the shift function $s(Y; \delta) = \delta_0 + \delta_1 Y$, where δ_0 models an overall change in testing rate, and δ_1 models an additional change in the likelihood of testing sick $(Y = 1)$ patients.

Estimating the impact of shift using quadratic approximation: To start, we keep $\delta_1 = 0$ fixed and vary only δ_0, which uniformly increases or decreases testing. In Fig. 3, we show the ground-truth cross-entropy loss of $f(O, L)$ under perturbed distributions $\mathbb{P}_{\delta_0}$. We observe that the **direction** of the shift matters: In Fig. 3, the model performance slightly increases under a small increase in testing rates, but degrades if testing increases too much; moreover, the loss under shift is generally asymmetric, as a decrease hurts more than an increase in testing. In Fig. 3 (left), we demonstrate the use of the quadratic approximation described in Section 3.2. For illustration, we consider a robustness set of $\delta_0 \in [-2, 2]$, and see that the predicted worst-case shift coincides with the actual worst-case shift, and that the quadratic approximation is accurate for smaller values of δ.

In Appendix D, we consider the case where δ_0 and δ_1 can both vary, and compare our approach to that of worst-case $(1 - \alpha)$ subpopulation shifts [Subbaswamy et al., 2021]. In the context of this example, we demonstrate that for any $1 - \alpha < 0.27$, the worst-case conditional subpopulation loss is achieved by having all healthy patients get tested, and no sick patients get tested. We contrast this with an iterative approach to designing constraints that is made possible by considering parametric shifts, where end-users can restrict the degree to which the shift differs across sick and healthy populations.

[3] Not to be confused with the 'trust region *method*', which repeatedly solves the trust region *problem*.

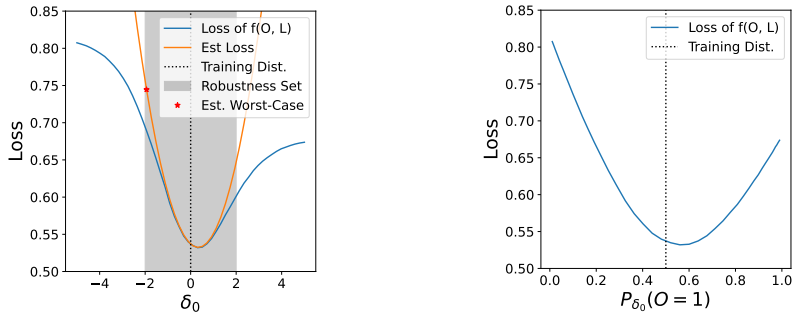

Figure 3: The blue line gives the (unobserved) cross-entropy loss under parametric shifts, plotted with respect to the parameter δ_0 (left) and the resulting change in the marginal laboratory testing rate (right). We also provide the quadratic approximation (orange line), estimated using validation data, and the predicted worst-case shift (red star) for $|\delta_0| < 2$ (region in grey).

4.2 Detecting sensitivity to non-causal correlations

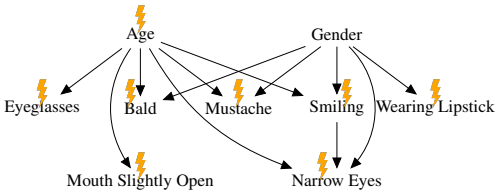

Figure 4: Causal graph over attributes, where lightning bolts indicate changes in mechanisms.

A predictive model may pick up on various problematic dependencies in the data that may not remain stable under dataset shift. To understand the impact of these dependencies, a model user may wish to understand which changes in distribution pose the greatest threats to model performance, and to measure the impact of these changes. To illustrate this use-case, we make use of the CelebA dataset [Liu et al., 2015], containing images of faces along and binary attributes (e.g., glasses, beard, etc.) encoding several non-causal features whose correlations may be unstable (e.g., the relation between gender and wearing lipstick). We consider the task of predicting gender (Y) from images of faces (X), and assess sensitivity to a shift in the distributions of attributes ($\mathbf{W}$).

Setup: To obtain ground-truth shifts in distribution, we generate synthetic datasets of faces using CausalGAN [Kocaoglu et al., 2018], trained on the CelebA data. We simulate attributes following the causal graph in Figure 4, and then simulate images from the GAN conditioned on those attributes. We draw a training sample from this distribution $\mathbb{P}$, and fit a gender classifier $f(X)$ using the image data alone, by finetuning a pretrained ResNet50 classifier [Hu et al., 2018], with a test accuracy of 0.919. Each attribute W_i is binary, so we consider shifts in the log-odds $\eta_i(Z_i)$ of each attribute W_i given parents Z_i. Here, we use a maximally flexible shift function $s_i(Z_i; \delta_i) = \sum_{z \in \mathcal{Z}_i} \delta_{i,z} \mathbf{1}\{Z_i = z\}$, such that for $Z_i \in \{0, 1\}^k$ there are 2^k parameters. Across all intervened variables, $\delta \in \mathbb{R}^{31}$. Per Section 3.3, we identify the shift δ causing the largest drop in accuracy, and compute $\hat{E}_{\delta,\text{Taylor}}$ and $\hat{E}_{\delta,\text{IS}}$ to estimate accuracy under that shift. Due to the synthetic nature of our setup, we can simulate from $\mathbb{P}_\delta(X, \mathbf{W}, Y)$ to estimate the ground-truth impact of this shift, by simulating from the shifted attribute distribution, and then simulating images from the GAN conditional on those attributes.

Finding and validating a single, high impact shift: Using a validation sample from $\mathbb{P}$, we estimate the shift gradient and Hessian (Theorem 1). Solving the quadratic optimization problem in Section 3.3, we find the worst-case shift δ such that $\|\delta\| \leq \lambda = 2$. We display the largest components of δ in Table 1 (left). Among others, this shift entails a 10% increase in the probability of a young man wearing lipstick, and a similar decrease for young women. This suggests that the learned classifier f relies on this non-causal association in the images for prediction. We validate that this shift leads to a measurable decrease in accuracy, from 92% to 89%, using simulated data from $\mathbb{P}_\delta$. To validate that this shift is indeed a worst-case shift, we simulate $K = 400$ random shifts δ_k and evaluate the model accuracy in $\mathbb{P}_{\delta_k}$ (Figure 5, left). As expected, our chosen shift δ (red line) is in the left tail of the distribution. We compare the ground-truth accuracy under this shift (89%) to the original

Conditional		δ_i	$\mathbb{P}$	$\mathbb{P}_\delta$
Wearing Lipstick	| Male, Young	0.998	0.08	0.182
Wearing Lipstick	| Female, Young	-0.992	0.92	0.819
Bald	| Female, Old	0.986	0.12	0.266
Wearing Lipstick	| Male, Old	0.499	0.12	0.182
Bald	| Male, Young	-0.471	0.12	0.078

Ground truth shift acc. ($\mathbb{E}_\delta[\ell_\gamma]$)	Original acc. ($\mathbb{E}[\ell_\gamma]$)
0.887	0.919
Taylor estimate ($\hat{E}_{\delta,\text{Taylor}}$)	**IS estimate ($\hat{E}_{\delta,\text{IS}}$)**
0.878	0.792

Table 1: (Left) Largest components of worst-case shift δ, where $\mathbb{P}$ and $\mathbb{P}_\delta$ denote conditional probabilities. (Right) Taylor and IS estimates of the loss $\mathbb{E}_\delta[\ell]$ at the worst-case shift δ.

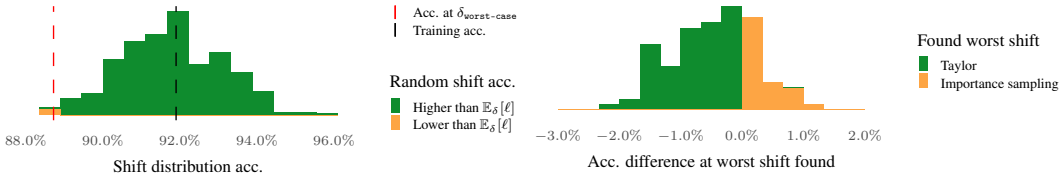

Figure 5: (Left) Model accuracy at randomly drawn shifts. (Right) Difference in accuracy in the worst-case shifts identified by Taylor and importance sampling approaches. The Taylor method identifies a more adversarial shift than importance sampling in 73% of simulations (green).

estimates $\hat{E}_{\delta,\text{Taylor}} = 0.878$ and $\hat{E}_{\delta,\text{IS}} = 0.792$ (Table 1), and observe that both correctly predict a drop in accuracy, although $\hat{E}_{\delta,\text{IS}}$ overestimates the size of the drop.

Comparing importance sampling and Taylor across multiple simulations: We simulate $K = 400$ validation sets from $\mathbb{P}$, in each estimating the worst-case shifts δ_{Taylor} and δ_{IS}, where the latter corresponds to maximizing $\hat{E}_{\delta,\text{IS}}$ using a standard non-convex solver from the `scipy` library [Virtanen et al., 2020]. We simulate ground truth data from $\mathbb{P}_{\delta_{\text{IS}}}$ and $\mathbb{P}_{\delta_{\text{Taylor}}}$, and in Figure 5 (right) we plot the differences $\mathbb{E}_{\delta_{\text{Taylor}}}[\ell] - \mathbb{E}_{\delta_{\text{IS}}}[\ell]$, showing that in 73% of cases, the Taylor method finds a more impactful shift. Moreover, the average run-time for the Taylor approach is $0.02s$ while that of the importance sampling approach is $2.52s$. Finally, the optimal value of the $\hat{E}_{\delta,\text{Taylor}}$ objective tends to a reasonably accurate estimate of the shifted accuracy, while the optimal value of the $\hat{E}_{\delta,\text{IS}}$ objective (both on the validation set) is a poor predictor, with a mean absolute prediction error (in predicting $\mathbb{E}_\delta[\ell]$) of 0.176 for $\hat{E}_{\delta,\text{IS}}$ and 0.014 for $\hat{E}_{\delta,\text{Taylor}}$.

5 Conclusion

We argue for considering parametric shifts in distribution, to evaluate model performance under a set of changes that are interpretable and controllable. For parametric shifts in conditional exponential family distributions, we derive a local second-order approximation to the loss under shift. This approximation enables the use of efficient optimization algorithms (to find the worst-case shift), and empirically provides realistic estimates of the resulting loss. In a computer vision task, this approach finds more impactful shifts (in far less time) than optimizing a reweighted objective, and the estimates of shifted accuracy under the chosen shift are substantially more reliable.

Limitations and potential for negative social impact: Our definition of parametric shifts and resulting approximation relies on the relevant mechanisms $\mathbb{P}(W|Z)$ being a conditional exponential family, and that the relevant variables are observed. As illustrated in our experiments, this can be used to model changes in the causal relationships **between** attributes of an image, but does not immediately extend to modelling changes in the distribution of images given a fixed set of attributes. As with any method that provides worst-case evaluation, there is potential for misuse and false confidence: If the specified shifts fail to capture important real-world changes, the resulting worst-case loss may be overly optimistic and misleading. Even if used correctly, our approach examines a narrow measure of model performance, and a small worst-case error should not be used to claim that a model is free of problematic behavior. For example, implicit dependence on certain attributes (e.g., race in medical imaging [Banerjee et al., 2021]) may be problematic based on ethical grounds, even if it does not lead to major issues with predictive performance under the small shifts in distribution.

References

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L.-C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang, P.-C. Kuo, M. P. Lungren, L. Palmer, B. J. Price, S. Purkayastha, A. Pyrros, L. Oakden-Rayner, C. Okechukwu, L. Seyyed-Kalantari, H. Trivedi, R. Wang, Z. Zaiman, H. Zhang, and J. W. Gichoya. Reading race: AI recognises patient's racial identity in medical images. July 2021.

D. Berthelot, T. Schumm, and L. Metz. BEGAN: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

A. R. Conn, N. I. Gould, and P. L. Toint. *Trust region methods*. SIAM, 2000.

J. Duchi, T. Hashimoto, and H. Namkoong. Distributionally robust losses for latent covariate mixtures. *arXiv preprint arXiv:2007.13982*, 2020.

J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

S. G. Finlayson, A. Subbaswamy, K. Singh, J. Bowers, A. Kupke, J. Zittrain, I. S. Kohane, and S. Saria. The clinician and dataset shift in artificial intelligence. *The New England journal of medicine*, 385(3):283–286, July 2021.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018.

L. Kook, B. Sick, and P. Bühlmann. Distributional anchor regression. *Statistics and Computing*, 32 (3):1–19, 2022.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

S. Magliacane, T. Van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

M. Oberst, N. Thams, J. Peters, and D. Sontag. Regularizing towards causal invariance: Linear models with proxies. In *International Conference on Machine Learning*, pages 8260–8270. PMLR, 2021.

J. Pearl. *Causality*. Cambridge University Press, 2009.

I. Pólik and T. Terlaky. A survey of the s-lemma. *SIAM review*, 49(3):371–418, 2007.

M. J. Powell. The NEWUOA software for unconstrained optimization without derivatives. In *Large-scale nonlinear optimization*, pages 255–297. Springer, 2006.

J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2008.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN 0378-3758.

A. Subbaswamy, R. Adams, and S. Saria. Evaluating model robustness and stability to dataset shift. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2611–2619. PMLR, 13–15 Apr 2021.

A. Subbaswamy, B. Chen, and S. Saria. The stability and accuracy tradeoff under dataset shift: A causal graphical analysis. *Journal of Causal Inference*, to appear, 2022.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272, 2020.

M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We clarify our contributions relative to prior work in Section 1.
   (b) Did you describe the limitations of your work? [Yes] As we introduce assumptions and our algorithm in Sections 2 and 3, we highlight what is required in order to use our approach. We summarize limitations explicitly in Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We do so explicitly in Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We clarify our main assumption and definitions in Section 2, and each theoretical statement refers back to the required assumptions.
   (b) Did you include complete proofs of all theoretical results? [Yes] Appendix F contains proofs of all our results.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will include code in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We specify all of these details in Appendix E for the CelebA experiment, and details of the laboratory testing synthetic experiment are described in Appendix A
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In our main experiment, we re-run for 400 random seeds, and report results that capture the resulting variation in Fig. 5.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We give the computation time for our main experiment in Section 4.2, with details of resources in Appendix E. The simpler synthetic examples (e.g., laboratory testing) are run on a personal laptop, not requiring extensive computation.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We make use of the CausalGAN code released by Kocaoglu et al. [2018], whom we cite. We also provide a citation for the CelebA dataset.
   (b) Did you mention the license of the assets? [Yes] See Appendix E
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include code in the supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The CelebA dataset is collected from publicly available images of celebrities on the internet, and we use it as an existing (and popular) research dataset. However, it is not clear that the individuals in question have provided consent, and we make note of this in Appendix E.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix E.
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]