Evaluating Robustness to Dataset Shift via Parametric Robustness Sets

Nikolaj Thams[*]
Dept. of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark
thams@math.ku.dk

Michael Oberst[*]
CSAIL & IMES
MIT
Cambridge, MA
moberst@mit.edu

David Sontag
CSAIL & IMES
MIT
Cambridge, MA
dsontag@csail.mit.edu

Abstract

We give a method for proactively identifying small, plausible shifts in distribution which lead to large differences in model performance. These shifts are defined via parametric changes in the causal mechanisms of observed variables, where constraints on parameters yield a "robustness set" of plausible distributions and a corresponding worst-case loss over the set. 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].

[*]Equal Contribution, order determined by coin flip. Code is available at this link.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

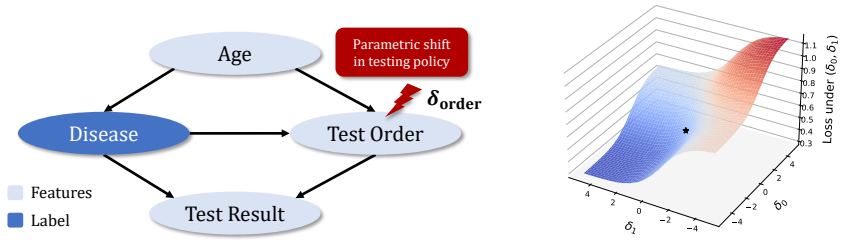


Figure 1: (Left) Causal graph for Example 1, with a shift in conditional testing rates, parameterized by δ_{order}. (Right) We illustrate a shift using $s(Y; \delta_{\text{order}}) = \delta_1 \cdot Y + \delta_0(1 - Y)$, where $\delta_{\text{order}} = (\delta_0, \delta_1)$. Here we plot the (non-concave) 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 laboratory test has been ordered (O), and test results (L) if a test was 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 $\mathbb{P}(O = 1|A, Y) = \sigma(\eta(A, Y))$, where σ is the sigmoid function and $\eta(A, Y)$ is the log-odds. In Figure 1 (right), we show the loss under a set of new distributions parameterized by $\delta = (\delta_0, \delta_1)$, where we modify $\mathbb{P}_\delta(O = 1|A, Y) = \sigma(\eta(A, Y) + s(Y; \delta))$ 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 Figure 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. When Δ is a single quadratic constraint, we can approximate the general non-convex optimization problem in Equation (2) with a particular 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 (i.e., conditional distributions) 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 substantially more reliable (see Section 4).

1.1 Related Work

Distributionally robust optimization/evaluation: Distributionally robust optimization (DRO) seeks to learn models that minimize objectives like Equation (1) with respect to the model [Duchi and Namkoong, 2021, Duchi et al., 2020, Sagawa et al., 2020]. We focus on proactive worst-case evaluation of a fixed model, not optimization, similar to Subbaswamy et al. [2021], Li et al. [2021], but we also differ in our **definition of the set of plausible future distributions** $\mathcal{P}$, often called an "uncertainty set" in the optimization literature. Prior work often defines these sets using distributional distances (such as f-divergences): For instance, Joint DRO [Duchi and Namkoong, 2021] allows for shifts in the entire joint distribution (i.e., all distributions in an f-divergence ball around $\mathbb{P}(X, Y)$), which may be overly conservative. Marginal DRO [Duchi et al., 2020] considers shifts in a marginal distribution (e.g., $\mathbb{P}(X)$), while assuming that the remaining conditionals (e.g., $\mathbb{P}(Y \mid X)$) are fixed. However, this assumption is not applicable in all scenarios: In Example 1, for instance, this assumption does not hold for a shift in testing policy. Conditional shifts are considered in recent work that focuses on evaluation [Subbaswamy et al., 2021], using worst-case conditional subpopulations. However, choosing a plausible size of conditional subpopulation is often non-obvious. In Appendix D we give a simple lab-testing example where taking worst-case 20% conditional subpopulations corresponds to an implausible shift: Healthy patients are always tested, and sick patients never tested.

In contrast, our approach uses explicit parametric perturbations to define shifts, as opposed to distributional distances or subpopulations. In addition, our approach allows for shifts in multiple marginal or conditional distributions simultaneously: In Example 1, for instance, we could model a simultaneous change in both the marginal distribution of age $\mathbb{P}(A)$, as well as the conditional distribution of lab testing $\mathbb{P}(O \mid A, Y)$, leaving other factors unchanged.

Causality-motivated methods for learning robust models: Several approaches proactively specify shifting causal mechanisms/conditional distributions, and then seek to learn predictors that have good performance under arbitrarily large changes in these mechanisms [Subbaswamy et al., 2019, Veitch et al., 2021, Makar et al., 2022, Puli et al., 2022]. Other approaches use environments [Magliacane et al., 2018, Rojas-Carulla et al., 2018, Arjovsky et al., 2019] or identity indicators [Heinze-Deml and Meinshausen, 2021] to learn models that rely on invariant conditional distributions.

However, when shifts are not arbitrarily strong, causality-motivated predictors can be overly conservative. In Example 1, a model that ignores all test-related features (and only uses age as a predictor) is a particularly simple example of a causality-motivated predictor, with invariant risk over changes in testing policy. Closer to our setting is a line of work that considers bounded mechanism changes in linear causal models [Rothenhäusler et al., 2021, Oberst et al., 2021], where estimation of the worst-case loss enables learning of worst-case optimal models. Our work can be seen as extending this idea to more general non-linear causal models, where we focus on evaluation rather than optimization.

Evaluating out-of-distribution performance with unlabelled samples: A recent line of work has focused on predicting model performance in out-of-distribution settings, where unlabelled data is available from the target distribution [Garg et al., 2022, Jiang et al., 2022, Chen et al., 2021]. In contrast, our method operates using only samples from the original source distribution, and seeks to estimate the worst-case loss over a set of possible target distributions.

In Appendix F we give a more detailed discussion of these approaches and others.

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 use $\mathbb{P}(\mathbf{V})$ to denote the probability density/mass function in the training distribution. We also refer to $\mathbb{P}$ as simply "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 the log-partition function.

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

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

be a factorization, where $Z_i, U_j, V_j \subseteq \mathbf{V}$ are possibly overlapping (or empty) sets of variables, where $\mathbb{P}(V_j \mid \varnothing) := \mathbb{P}(V_j)$. For each W_i we assume Z_i is known and $\mathbb{P}(W_i|Z_i)$ satisfies Definition 2.

If $\mathbb{P}(\mathbf{V})$ factorizes according to a DAG $\mathcal{G}$, the factorization in Assumption 1 is always satisfied by $Z_i = \mathrm{PA}_{\mathcal{G}}(W_i)$. While we assume data is generated according to Equation (4), we do not require knowledge of the full distribution, but only the conditionals 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 which satisfies $s(Z; 0) = 0$ for all Z. For $\mathbb{P}(W|Z)$ satisfying Equation (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 $\eta_\delta(Z) := \eta(Z) + s(Z; \delta)$. Note that this differs from Equation (3) in that $\eta(Z)$ is replaced by $\eta_\delta(Z)$.

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 patients 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[2] 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 available, causal knowledge helps identify which factors in the joint distribution are subject to shifts (e.g., $\mathbb{P}(O \mid Y, A)$ in Example 1), and which remain stable. It is worth noting, however, that our methodology can be used to model any change in distribution that satisfies Assumption 1, including choices of "non-causal" factorizations and shifting factors. For example, in the context of Example 1, we could choose the factorization $\mathbb{P}(Y)\mathbb{P}(O \mid Y)\mathbb{P}(L, A \mid O, Y)$, and model a change only in the conditional $\mathbb{P}(O \mid Y)$ while keeping other factors unchanged. This shift is not interpretable as a change in causal mechanisms: The shifted distribution would imply a change in the marginal distribution of age, which should be unaffected by a real-world change in laboratory testing. Nonetheless, we can still estimate a worst-case loss over such non-causal shifts in distribution. In short, our machinery can model shifts in non-causal conditionals (for example because the causal structure is unknown), though the resulting shifted distribution is not interpretable as a plausible shift in the ground-truth data generating mechanism.

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 estimators can exhibit high variance for shifts that produce large density ratios (see Appendix C.5 for an example), and 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,* $\mathrm{supp}(\mathbb{P}) = \mathrm{supp}(\mathbb{P}_\delta)$ *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 G, 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

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

[2] 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 $\eta(Z)$.

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

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

Equation (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

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

where $\mathbb{E}[\ell]$ denotes the loss in the training distribution and $\mathrm{SG}^1, \mathrm{SG}^2$ 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

$$\mathrm{SG}^1 := \nabla_\delta \mathbb{E}_\delta[\ell]\big|_{\delta=0} \qquad \text{and} \qquad \mathrm{SG}^2 := \nabla_\delta^2 \mathbb{E}_\delta[\ell]\big|_{\delta=0}.$$

Equation (7) is a local approximation of the loss, whose approximation error we bound in Theorem 2, with smaller approximation error for smaller shifts.[3] 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 $\mathrm{cov}(A, B|C) := \mathbb{E}[(A - \mathbb{E}[A|C])(B - \mathbb{E}[B|C])|C]$.

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*

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

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

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

where $D_{i,k} := \nabla_{\delta_i}^k s_i(Z_i; \delta_i)|_{\delta=0}$, 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*

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

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) follows a normal distribution with mean μ and variance σ^2, and consider a shift in the mean (without changing lab testing). 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 $\mathrm{SG}^1 = \mathrm{cov}\left(\ell, A\right)/\sigma$ and $\mathrm{SG}^2 = \mathrm{cov}\left(\ell, (A - \mathbb{E}[A])^2\right)/\sigma^2$.

[3] 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].

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 estimates of the gradient and Hessian, we 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 $E_{\delta,\text{Taylor}}$ be the population Taylor estimate (Equation (7)) and let $\sigma(M)$ denote the largest absolute value of the eigenvalues of a matrix M. Then*

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

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

To build intuition, in Appendix C.8 we give a scenario where this bound can be simplified. In particular, we consider a "covariate shift" setting [Quiñonero-Candela et al., 2008] where X is standard Gaussian, $Y = f_0(X) + \epsilon$ with a noise term independent of X and we consider a shift δ in the mean of X. When evaluating a predictor $f(X)$ with the loss ℓ being the squared error, the bound in Theorem 2 depends on how the modelling error $g(X) = f_0(X) - f(X)$ behaves over the domain. In particular, the bound scales as the supremum (over $t \in [0,1]$) of $\sqrt{\mathrm{var}(g(X + t\cdot\delta)^2 - g(X)^2)}$. As a simple corollary, if our predictor is off by an additive constant factor, $f = f_0 + C$, then the bound is zero, and the approximation is exact for any δ. On the other hand, if the squared modelling error $g(X)^2$ at one point X tends to be a poor predictor of the squared modelling error at another point $X + t\cdot\delta$, then this variance will be large, and the approximation will be loose.

In exchange for considering a second-order approximation of the loss, we gain two benefits: Variance reduction and tractable optimization. First, the variance of $\hat{E}_{\delta,\text{Taylor}}$ is $O(\|\delta\|^4)$ for large $\|\delta\|$, while the variance of $\hat{E}_{\delta,\text{IS}}$ can be much larger: We give a simple case in Appendix C.6 where $\mathrm{var}(\hat{E}_{\delta,\text{Taylor}}) = O(\delta^4)$ while $\mathrm{var}(\hat{E}_{\delta,\text{IS}}) = O(\delta^2 \exp(\delta^2))$. 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

$$\sup_{\|\delta\|_2\leq\lambda} \mathbb{E}[\ell] + \delta^\top \mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \mathrm{SG}^2 \delta. \tag{9}$$

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'[4] 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 $\|\delta\| \leq \lambda$. In particular, the strong duality of Equation (9) also holds when $\|\delta\|_2 \leq \lambda$ is replaced by any single quadratic constraint $\delta^\top A\delta + \delta^\top b \leq \lambda$, allowing for e.g., larger shifts in some directions than in others.

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

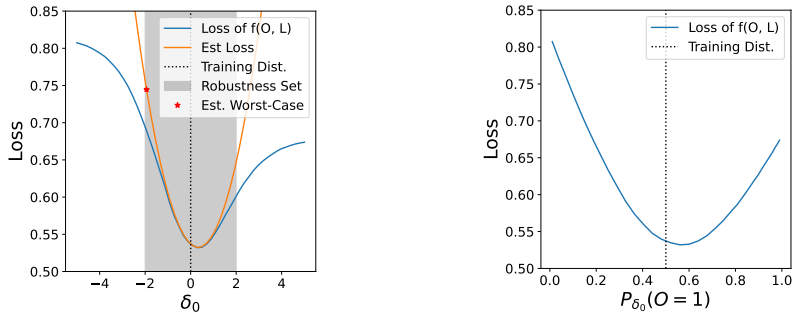


Figure 3: The blue line gives the (unobserved) cross-entropy loss under parametric shifts, plotted with respect to the parameter δ_0 (left) and the resulting change in the marginal laboratory testing rate (right). We also provide the quadratic approximation (orange line), estimated using validation data,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

4 Experiments

4.1 Illustrative example: Laboratory testing

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

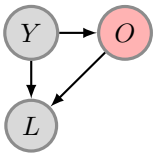


Figure 2

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

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 Figure 2. The predictive model $f(O, L)$ 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 uses a logistic regression model trained on 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 Figure 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 Figure 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 Figure 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 allow both δ_0 and δ_1 to vary, and compare our approach to that of worst-case $(1 - \alpha)$ conditional 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.

4.2 Detecting sensitivity to non-causal correlations

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

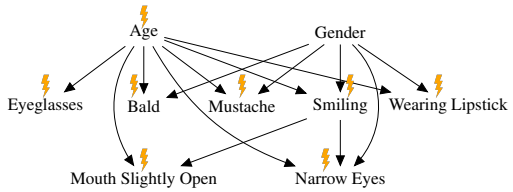


Figure 4: Causal graph over attributes in the synthetic CelebA dataset, where lightning bolts indicate changes in mechanisms. All of these attributes are causal parents of the image X (not shown here), which is generated by a GAN conditioned on these attributes.

[Liu et al., 2015], which contains images of faces and binary attributes (e.g., glasses, beard, etc.) encoding several features whose correlations may be unstable (e.g., the relation between gender and being bald). We consider the task of predicting gender (Y) from images of faces (X), and assess sensitivity to a shift in the distributions of attributes ($\mathbf{W}$).[5]

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]. 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, $\delta \in \mathbb{R}^{31}$. Due to the synthetic nature of our setup, we can simulate from $\mathbb{P}_\delta(X, \mathbf{W}, Y)$ to evaluate the ground-truth impact of this shift, simulating first from the shifted attribute distribution, and then simulating images from the GAN conditional on those attributes. We use the 0/1 loss $\ell = \mathbf{1}\{f(X) \neq Y\}$, and constrain δ by $\|\delta\|_2 \leq \lambda = 2$.

Comparing importance sampling and Taylor across multiple simulations: We simulate $K = 100$ validation sets from $\mathbb{P}$, in each estimating the worst-case shifts δ_{Taylor} (via the approach in Section 3.3) and δ_{IS}, where the latter corresponds to minimizing $\hat{E}_{\delta,\text{IS}}$ 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 $\mathbb{P}_{\delta_{\text{Taylor}}}$, to compare the two shifts. First, we demonstrate that the Taylor approach finds more impactful shifts, when searching over the space of small, bounded shifts considered here. In Table 1 (right), we compare the average drop in accuracy using the Taylor shifts (3.8%) and the IS shifts (2.2%). In Figure 5 (right) we plot the differences in test accuracy $\mathbb{E}_{\delta_{\text{Taylor}}}[\mathbf{1}\{f(X) = Y\}] - \mathbb{E}_{\delta_{\text{IS}}}[\mathbf{1}\{f(X) = Y\}]$, where the Taylor approach finds a more impactful shift in 96% of cases. Second, the Taylor approach has an average run-time of $0.01s$, versus $2.14s$ for the IS approach. Third, when **only** used to evaluate the shift δ_{Taylor}, the IS estimator is comparable to the Taylor estimator, with a near-identical average bias (shown in Table 1 (right)) and RMSE (0.0191 and 0.0192 respectively). Finally, however, in Table 1 (right) we observe that $\hat{E}_{\delta_{\text{IS}},\text{IS}}$ is strongly biased in predicting $\mathbb{E}_{\delta_{\text{IS}}}$, yielding a mean absolute prediction error (MAPE) of 0.069 (not shown in the table). This can be contrasted with a MAPE of 0.015 when using $\hat{E}_{\delta_{\text{Taylor}},\text{Taylor}}$ to predict $\mathbb{E}_{\delta_{\text{Taylor}}}$. This may suggest that optimizing the IS objective is prone to "overfitting", choosing a sub-optimal δ from a region of the search space that has high variance. Here, where $\lambda = 2$, the drop in accuracy is relatively mild for the shifts found by both approaches. In Appendix E.4 we show that larger values of λ correspond to more substantial drops in accuracy (e.g., an average drop of 23% for $\lambda = 8$ using the Taylor approach).

Examining a single shift: To illustrate the type of shift found by our approach, we consider the δ_{Taylor} (over the K runs) which yields the $\mathbb{P}_\delta$ with median test accuracy. We display the largest components of that δ in Table 1 (left). Among others, this shift entails a 5% increase in the probability of an older woman being bald, and a 5% decrease in the probability of a young woman wearing lipstick. This suggests that the learned classifier f relies on these associations in the images for prediction. We validate that this shift leads to a decrease in accuracy of around 3.8%, using simulated data from $\mathbb{P}_\delta$. To validate that this drop in accuracy is a non-trivial occurrence, we simulate $K = 400$ random shifts

[5]We do not endorse gender classification as an inherently worthwhile task. Nonetheless, gender classification is commonly studied in the context of understanding the implicit biases of machine learning models [Buolamwini and Gebru, 2018, Schwemmer et al., 2020], and we consider the task with that context in mind.

Table 1: (Left) Top 5 components (by magnitude) of the example shift vector $\delta \in \mathbb{R}^{31}$ where $\mathbb{P}$ and $\mathbb{P}_\delta$ denote conditional probabilities. The full example shift vector can be found in Appendix E.2. (Right) Taylor and IS estimates vs. true accuracy for the δ_{Taylor} found by the Taylor approach, and IS estimate vs. true accuracy for the δ_{IS} found by the IS approach. Averages are taken over 100 simulations.

Conditional		δ_i	$\mathbb{P}$	$\mathbb{P}_\delta$
Bald	| Female, Old	0.899	0.047	0.109
Bald	| Male, Young	-0.800	0.378	0.214
Bald	| Male, Old	-0.680	0.622	0.455
Wearing Lipstick	| Female, Young	-0.618	0.924	0.868
Wearing Lipstick	| Female, Old	-0.543	0.953	0.921

Metric	Example δ	Avg.
Original acc. ($\mathbb{E}[\mathbf{1}\{f(X) = Y\}]$)		0.912
Acc. under Taylor shift ($\mathbb{E}_{\delta_{\text{Taylor}}}[\mathbf{1}\{f(X) = Y\}]$)	0.874	0.874
IS est. of acc. under Taylor shift ($\hat{E}_{\delta_{\text{Taylor}},\text{IS}}$)	0.829	0.863
Taylor est. of acc. under Taylor shift ($\hat{E}_{\delta_{\text{Taylor}},\text{Taylor}}$)	0.844	0.863
Acc. under IS shift ($\mathbb{E}_{\delta_{\text{IS}}}[\mathbf{1}\{f(X) = Y\}]$)		0.889
IS est. of acc. under IS shift ($\hat{E}_{\delta_{\text{IS}},\text{IS}}$)		0.821

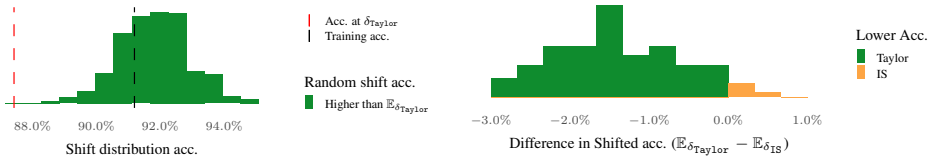


Figure 5: (Left) Model accuracy at randomly drawn shifts. (Right) Difference in accuracy in the worst-case shifts identified by Taylor and importance sampling approaches. The Taylor method identifies a more adversarial shift than importance sampling in 96% of simulations (green).

δ_k where $\|\delta_k\| = \lambda$ and evaluate the model accuracy in $\mathbb{P}_{\delta_k}$ (Figure 5, left). As expected, the chosen δ yields a lower accuracy (red line) than all of the random shifts.

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.

Of course, our method is not without limitations. 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 small shifts in distribution.

Acknowledgements

We thank Jonas Peters, Chandler Squires, and Stefan Hegselmann for helpful feedback and discussion, and Irene Chen and Christina X Ji for providing comments on an earlier draft. MO and DS were supported in part by Office of Naval Research Award No. N00014-21-1-2807. NT was supported by a research grant (18968) from VILLUM FONDEN.

References

A. Ahmadi-Javid. Entropic Value-at-Risk: A new coherent risk measure. *Journal of optimization theory and applications*, 155(3):1105–1123, Dec. 2012.

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. *arXiv preprint arXiv:2107.10356*, 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.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 2018.

M. Chen, K. Goel, N. S. Sohoni, F. Poms, K. Fatahalian, and C. Re. Mandoline: Model evaluation under distribution shift. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1617–1629. PMLR, 2021.

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.

S. Garg, S. Balakrishnan, Z. C. Lipton, B. Neyshabur, and H. Sedghi. Leveraging unlabeled data to predict Out-of-Distribution performance. In *ICLR*, 2022.

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.

C. Heinze-Deml and N. Meinshausen. Conditional variance penalties and domain shift robustness. *Machine learning*, 110(2):303–348, Feb. 2021.

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.

Y. Jiang, V. Nagarajan, C. Baek, and J. Zico Kolter. Assessing generalization of SGD via disagreement. In *ICLR*, 2022.

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.

H. Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41 (4):1248–1275, Nov. 2016.

M. Li, H. Namkoong, and S. Xia. Evaluating model performance under worst-case subpopulations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

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. Makar, B. Packer, D. Moldovan, D. Blalock, Y. Halpern, and A. D'Amour. Causally motivated shortcut removal using auxiliary labels. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 739–766. PMLR, 2022.

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.

A. Puli, L. H. Zhang, E. K. Oermann, and R. Ranganath. Out-of-distribution generalization in the presence of Nuisance-Induced spurious correlations. In *International Conference on Learning Representations*, 2022.

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. *Journal of machine learning research: JMLR*, 19(36):1–34, 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.

C. Schwemmer, C. Knight, E. D. Bello-Pardo, S. Oklobdzija, M. Schoonvelde, and J. W. Lockhart. Diagnosing gender bias in image recognition systems. *Socius : sociological research for a dynamic world*, 6, Jan. 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, P. Schulam, and S. Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

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.

Veitch, D'Amour, Yadlowsky, and Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *Advances in neural information processing systems*, 34, Dec. 2021.

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 G 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 include a link to the code on the first page of the paper.
   (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 100 random seeds, and report results that capture the resulting variation in Figure 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 a link to our code, but do not include any other assets.
   (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]

Appendix

This appendix is structured as follows:

- In Appendix A, we provide details on the synthetic lab testing example, including how we generate the loss landscape in Figure 1 (right).
- In Appendix B, we provide a "user's guide" to defining and interpreting parametric shifts, including worked examples for many common conditional distributions, as well as guidance on how to define and interpret the shift functions $s(Z; \delta)$.
- In Appendix C, we provide additional details on the worst-case optimization problem, as well as comparisons of the reweighting-based approach to the Taylor approximation approach. We also demonstrate that the quadratic approximation is exact, for particularly simple structural causal models.
- In Appendix D, we compare our approach to that of worst-case conditional subpopulation shifts, in the context of a simpler laboratory testing example where we can explicitly compute the worst-case conditional subpopulations. Here, we demonstrate that our approach can capture more realistic intuition regarding which shifts are plausible in practice.
- In Appendix E, we give additional experimental details, as well as illustrative samples from the generative model, for the CelebA experiment described in Section 4.
- In Appendix F, we give an extended discussion of related work.
- In Appendix G, we give proofs for all the results in the main paper.

A Details of Figure 1

In Figure 1 (right), we consider the following, artificial, generative model, which resembles the setup in Section 4.1, but with the addition of age as a continuous variable.

$$\begin{aligned}
\text{Age} &\sim \mathcal{N}(0, 0.5^2) \\
\mathbb{P}(\text{Disease} = 1 | \text{Age}) &= \text{sigmoid}(0.5 \cdot \text{Age} - 1) \\
\mathbb{P}(\text{Order} = 1 | \text{Disease, Age}) &= \text{sigmoid}(2 \cdot \text{Disease} + 0.5 \cdot \text{Age} - 1) \\
\text{Test Result} | \text{Order} = 1, \text{Disease} &\sim \mathcal{N}(-0.5 + \text{Disease}, 1)
\end{aligned}$$

where if Order $= 0$, the test result is a placeholder value of zero. In Figure 1 (right), we consider a simple predictive model: If lab tests are not available (Order $= 0$), this model predicts disease based on an unregularized logistic regression model, which uses age to predict disease. If a lab test is available, then it uses both age and the lab test for prediction. This model is trained on 100,000 samples from the training distribution. To construct the loss landscape shown in Figure 1 (right), we first observe that

$$\mathbb{P}(O = 1 | \text{Disease}, \text{Age}) = \text{sigmoid}(\eta(\text{Disease}, \text{Age})),$$

where

$$\eta(\text{Disease}, \text{Age}) = 2 \cdot \text{Disease} + 0.5 \cdot \text{Age} - 1.$$

We construct shifts using the shift function $s(\text{Disease}, \text{Age}; \delta) = \delta_0 \cdot (1 - \text{Disease}) + \delta_1 \cdot \text{Disease}$, and for a grid of values for $(\delta_0, \delta_1) \in [-5, 5]^2$ we consider perturbed distributions with a different conditional distribution of testing,

$$\mathbb{P}_\delta(O = 1 | \text{Disease, Age}) = \text{sigmoid}\left(\eta(\text{Disease}, \text{Age}) + \delta_0 \cdot (1 - \text{Disease}) + \delta_1 \cdot \text{Disease}\right),$$

but where all other parts of the generative model are fixed. For each value of $(\delta_0, \delta_1) \in [-5, 5]^2$, we draw 10,000 samples from the corresponding distribution, and compute the negative log-likelihood of the original predictive model under this new distribution. The resulting surface is plotted in Figure 1 (right).

Table 2: Examples of conditional exponential family distributions.

Distribution	Parameter space	Sufficient statistic	Inverse parameter map
Binary(p)	$\eta(Z) \in \mathbb{R}$	$T(W) = W$	$p(W = 1\|Z) = \text{sigmoid}(\eta(Z))$
Categorical($p_1, \ldots, p_k$)	$\eta(Z) \in \mathbb{R}^k$	$[T(W)]_i = \mathbf{1}\{W = i\}$	$\mathbb{P}(W = i\|Z) = [\text{softmax}(\eta(Z))]_i$
Poisson(λ)	$\eta(Z) \in \mathbb{R}$	$T(W) = W$	$\lambda = \exp(\eta(Z))$
Gaussian(μ, σ^2)	$\eta(Z)_1 \in \mathbb{R}, \eta(Z)_2 < 0$	$T(W) = (W, W^2)$	$\mu(Z) = -\frac{\eta(Z)_1}{2\eta(Z)_2}, \sigma^2(Z) = -\frac{1}{2\eta(Z)_2}$
Gamma(α, β)	$\eta(Z)_1 > -1, \eta(Z)_2 < 0$	$T(W) = (\log W, W)$	$\alpha(Z) = \eta(Z)_1 + 1, \beta(Z) = -\eta(Z)_2$

B A user's guide to defining parametric shifts

In this section, we discuss practical considerations in designing parametric shift functions for different distributions.

- In Appendix B.1, we give examples of conditional exponential families, illustrative shift functions, and how to interpret them.
- In Appendix B.2, we formalize the idea that one can choose shift functions which depend on additional variables, other than the causal parents of a variable W_i.
- In Appendix B.3 we give guidance on how to define shift functions when the parameters $\eta(Z)$ are constrained to lie in a particular domain, which is relevant for considering shifts such as changing the variance of a conditional Gaussian.

B.1 Conditional exponential family models and interpretations of shifts

In this section, we give examples of exponential families and their sufficient statistics, and discuss design considerations in specifying the shift function $s(Z; \delta)$. Here, we restrict attention to shifts in a single variable, for ease of notation. In Table 2 we give examples of conditional exponential families, along with their typical parameterizations. In the examples below, we review how shift functions $s(Z; \delta)$ impact these parameters, and how they can also be interpreted on the scale of more commonly considered parameters (e.g., conditional means and variances).

Example B.1 (Log-odds shift in a binary variable)**.** Consider the distribution of a binary variable W conditioned on variables Z. Without loss of generality, we can write that

$$\mathbb{P}(W = 1|Z) = \sigma(\eta(Z))$$

where σ is the sigmoid function, and $\eta(Z)$ is an arbitrary measurable function of Z, taking on values in the extended real line $\eta(Z) \in \mathbb{R} \cup \{-\infty, +\infty\}$. This can be written in canonical form as

$$\mathbb{P}(W|Z) = \exp\left\{\eta(Z) \cdot W - \log(1 + \exp^{\eta(Z)})\right\}$$

where $\eta(Z)$ is the canonical parameter (the log-odds ratio), $T(W) = W$ is the sufficient statistic, and $h(\theta) = \log(1 + \exp^{\eta(Z)})$ is the normalizing constant. We can consider shifts $\eta_\delta(Z) := \eta(Z) + \delta$, yielding the new conditional distribution

$$\mathbb{P}_\delta(W = 1|Z) = \sigma(\eta(Z) + \delta),$$

which is well-defined for any $\delta \in \mathbb{R}$.

Here, we note that these shifts occur on the "natural" parameter scale $\eta(Z)$ (e.g., the log-odds), which at first glance may seem difficult to interpret: Why should we care about changes on the log-odds scale, instead of on the original probability scale? In addition to mathematical convenience, we argue that in some settings, working with natural parameters is advantageous for retaining a common scale across across multiple variables.

For instance, consider shifts in the two independent variables W_1 and W_2, where $V_i \sim \text{Bernoulli}(p_i)$, with $p_1 = 10^{-4}$ and $p_2 = 0.6$. Suppose we wished to consider an additive shift on the probability scale, e.g., $p_1' = p_1 + 0.1, p_2' = p_2 + 0.1$. Setting aside the inconvenience that we need to ensure $p_1', p_2' \in [0, 1]$, we argue that these shifts are not truly of a comparable scale. In particular, this shift in p_1 may seem implausible in magnitude, while the same shift in p_2 seems more reasonable. On the other hand, an additive shift in the log-odds captures some aspect of this idea.

Of course, there is some flexibility to incorporate prior expectations of shifts in absolute probabilities. For instance, in binary variable with no causal parents, we can always construct a one-to-one map of δ to a change in the marginal probability. For conditional shifts, we can similarly construct a one-to-one map between the value of δ in a shift $s(Z;\delta) = \delta$ and the resulting marginal probability of W_i, as formalized below.

Proposition B.1. *Consider a binary random variable W with conditional distribution*

$$\mathbb{P}_\delta(W = 1|Z) = \sigma(\eta(Z) + \delta)$$

for an arbitrary measurable function $\eta(Z)$ whose range is the extended real numbers $\eta(Z) \in \mathbb{R} \cup \{+\infty, -\infty\}$. Let $p_+ := \mathbb{P}(\eta(Z) = +\infty)$, $p_- := \mathbb{P}(\eta(Z) = -\infty)$, and assume that $p_+ + p_- < 1$. Then, the marginal probability

$$p_\delta = \mathbb{P}_\delta(W = 1)$$

is a strictly monotonically increasing function of $\delta \in \mathbb{R}$ whose range is $(p_+, 1 - p_-)$,

Proposition B.1 states that, for any achievable marginal probability $p_\delta = \mathbb{P}_\delta(W = 1)$, there exists a unique value of δ that achieves this probability. Because this relationship is strictly monotonic, we can hope to efficiently find such a value by e.g., binary search. In the laboratory testing example of Example 1, this would allow us to specify a plausible strength for the conditional shift δ in terms of an impact on the overall testing rate, e.g., modelling a scenario where the testing rate decreases from 20% to 15%.

Similar to the binary case, we can (if desired) directly parameterize shifts in terms of the conditional mean of a Gaussian distribution, as illustrated in Example B.2, which operates on the scale of $\mu(Z)$ alone.

Example B.2 (Mean shift in a conditional Gaussian)**.** Consider the distribution of a multi-variate Gaussian variable W conditioned on a binary variable Z, where we write

$$p(w|z) \overset{(d)}{=} \mathcal{N}(w; \mu(z), \Sigma(z))$$

where $\mathcal{N}(w; \mu(z), \Sigma(z))$ denotes the Gaussian density with mean $\mu(z)$ and covariance $\Sigma(z)$. This can be written as an exponential family model with natural parameters $\eta(Z) = [\Sigma(Z)^{-1}\mu(Z), -\frac{1}{2}\Sigma(Z)^{-1}]$ and sufficient statistic $T(W) = [W, WW^\top]$. Here, a shift in the mean can be parameterized by $s(Z;\delta) = [\Sigma(Z)^{-1}\delta, 0]$, such that

$$p_\delta(w|z) \overset{(d)}{=} \mathcal{N}(w; \mu(z) + \delta, \Sigma(z)).$$

However, shifts of the same magnitude in the conditional mean may not be comparable. Suppose that

$$\mathbb{P}(W|Z = 0) \overset{(d)}{=} \mathcal{N}(0, 1) \qquad \text{and} \qquad \mathbb{P}(W|Z = 1) \overset{(d)}{=} \mathcal{N}(0, 0.001),$$

such that $\delta = 1$ in Example B.2 corresponds to

$$\mathbb{P}_{\delta=1}(W|Z = 0) \overset{(d)}{=} \mathcal{N}(1, 1) \qquad \text{and} \qquad \mathbb{P}_{\delta=1}(W|Z = 1) \overset{(d)}{=} \mathcal{N}(1, 0.001).$$

While it may seem plausible that the mean of $W|Z = 0$ can increase by 1, it may seem unrealistic for $W|Z = 1$. Here, it may be more reasonable to consider a different parameterization of $s(Z;\delta)$, where the impact of the shift in a direction is proportional to the variance in that direction; we discuss this in the next example.

Example B.3 (Variance-scaled mean shift in a conditional Gaussian)**.** Consider the distribution of a multi-variate Gaussian variable W conditioned on variables Z, where we write

$$p(w|z) \overset{(d)}{=} \mathcal{N}(w; \mu(z), \Sigma(z))$$

where $\mathcal{N}(w; \mu(z), \Sigma(z))$ denotes the Gaussian density with mean $\mu(z)$ and covariance $\Sigma(z)$. This can be written as an exponential family model with natural parameters $\eta(Z) = [\Sigma(Z)^{-1}\mu(Z), -\frac{1}{2}\Sigma(Z)^{-1}]$ and sufficient statistic $T(W) = [W, WW^\top]$. Here, a shift in the mean can be parameterized by $s(Z;\delta) = [\delta, 0]$, such that

$$p_\delta(w|z) \overset{(d)}{=} \mathcal{N}(w; \mu(z) + \delta^\top \Sigma(Z), \Sigma(z)).$$

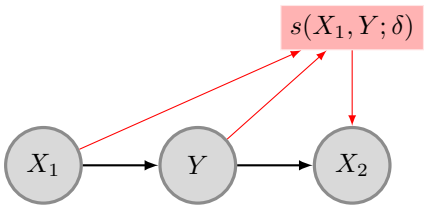


Figure 6: Illustrative example of an intervention $s(X_1, Y; \delta)$, and modified causal graph, which creates a dependence between X_1 and X_2 that bypasses Y.

In Example B.3, the parameter δ has a different interpretation, as a variance-scaled mean-shift. If W is one-dimensional, we can see that this becomes

$$p_\delta(w|z) \overset{(d)}{=} \mathcal{N}(w; \mu(z) + \delta\sigma^2(Z), \sigma^2(z)).$$

As we demonstrate in Appendix C.2, this particular example of a parameterization has other benefits: For instance, for estimation of shift gradients and Hessians at $\delta = 0$ can be done without knowledge of $\Sigma(Z)$.

B.2 Adding causal edges to the graph

In Section 2, we consider the case where the shift function $s(Z; \delta)$ alters a conditional $\mathbb{P}(W|Z)$ by a shift function $s(Z; \delta)$. We now discuss shift functions that use a larger set Z'. In particular, we consider the setting where Z represents the parents in a graph $\mathcal{G}$ (that is, $Z := \mathrm{PA}_{\mathcal{G}}(W)$), and consider shift functions that correspond to adding additional parents in that causal graph. Our definitions and results immediately extend to measuring the impact of shifts that **add edges** to the graph, in the form of shift functions that depend on non-descendants of W.

Building intuition with a simple example: To build intuition, consider the causal graph given in Figure 6. We consider a shift in X_2, with a shift function which depends not only on the causal parent Y, but also on X_1. Suppose that the distribution $\mathbb{P}(X_2|Y)$ is a conditional exponential family, given by

$$\mathbb{P}(X_2|Y) = g(X_2)\exp(\eta(Y)^\top T(X_2) - h(\eta(Y))).$$

Using that $X_2 \perp\!\!\!\perp X_1|Y$, we have $\mathbb{P}(X_2|Y) = \mathbb{P}(X_2|Y, X_2)$, and the joint probability factorizes as

$$\mathbb{P}(X_1, X_2, Y) = \mathbb{P}(X_2|Y)\mathbb{P}(Y|X_1)\mathbb{P}(X_1) = \mathbb{P}(X_2|Y, X_1)\mathbb{P}(Y|X_1)\mathbb{P}(X_1).$$

This enables us to consider $Z = (Y, X_1)$ as the conditioning set in the context of Assumption 1. This is useful, because it allows us to consider shift functions that depend on Z, which includes X_1 in addition to Y. The δ-perturbation of this conditional distribution under the shift function $s(Y, X_1; \delta)$ is given by

$$\mathbb{P}_\delta(X_2|Y, X_1) = g(X_2)\exp\Big(\{\eta(Y) + s(Y, X_1; \delta)\}^\top T(X_2) - h\big(\eta(Y) + s(Y, X_1; \delta)\big)\Big),$$

and we can observe that under both graphs, the distribution factorizes in the same fashion, where

$$\mathbb{P}_\delta(X_1, X_2, Y) = \mathbb{P}_\delta(X_2|Y, X_1)\mathbb{P}(Y|X_1)\mathbb{P}(X_1),$$

keeping the same convention that $s(Y, X_1; \delta = 0) = 0$, such that $\mathbb{P}_0 = \mathbb{P}$. This is one example of how our results can be applied with shift functions that effectively add edges to the causal graph. Of course, not all edges are permitted, so we give a more general treatment below.

General guidelines for adding edges: Allowing for the use of non-causal parents in the shift functions is straightforward, and can be done safely as follows, without violating Assumption 1: Given knowledge of the directed acyclic graph $\mathcal{G}$ which generates the observed distribution $\mathbb{P}$, we can **add** edges to the graph, as long as they do not create cycles.

Formally, let $\mathcal{G} = (\mathbf{V}, E)$ denote the causal DAG which generates the distribution $\mathbb{P}$, where $\mathbf{V}$ denotes variables and E denotes the set of edges, where we denote a directed edge by $e = (V_i, V_j)$, going from V_i to V_j. Let $\mathcal{G}' = (\mathbf{V}', E')$ denote another DAG (of our creation) with the constraint that we can only add edges, and that the graph must remain acyclic, such that $E' \supseteq E$, and $\mathbf{V}' = \mathbf{V}$.

For any variable $W_i \in \mathbf{V}$, this implies that $\mathrm{PA}_{\mathcal{G}'}(W_i) \supseteq \mathrm{PA}_{\mathcal{G}}(W_i)$. Moreover, any new causal parent V_i of W_i in $\mathcal{G}'$ must have been a non-descendant of W_i in the original graph, as otherwise the graph $\mathcal{G}'$ would have a cycle from $W_i \to V_i \to W_i$. For ease of notation, let $N(W_i) := \mathrm{PA}_{\mathcal{G}'}(W_i) \setminus \mathrm{PA}_{\mathcal{G}}(W_i)$ denote the set of new causal parents of W_i in $\mathcal{G}'$. For any variable W_i such that $N(W_i) \neq \varnothing$, we can write that

$$W_i \perp\!\!\!\perp_{\mathcal{G}} N(W_i) | \, \mathrm{PA}_{\mathcal{G}}(W_i) \tag{10}$$

by the rules of d-separation [Pearl, 2009]. As in Assumption 1, we use $\mathbf{W} = \{W_1, \ldots, W_m\}$ to denote the set of variables to be intervened upon, and accordingly will assume that in the causal graph $\mathcal{G}'$, we have not added new parents to any other variables, i.e., $N(V_i) = \varnothing$ for any $V_i \subsetneq \mathbf{W}$.

By Equation (10), we can write that the distribution $\mathbb{P}$ factorizes as

$$\mathbb{P}(\mathbf{V}) = \left(\prod_{W_i \in \mathbf{W}} \mathbb{P}(W_i | \, \mathrm{PA}_{\mathcal{G}'}(W_i)) \right) \prod_{V_i \in \mathbf{V} \setminus \mathbf{W}} \mathbb{P}(V_i | \, \mathrm{PA}_{\mathcal{G}}(V_i))$$

because $\mathbb{P}(W_i | \, \mathrm{PA}_{\mathcal{G}'}(W_i)) = \mathbb{P}(W_i | \, \mathrm{PA}_{\mathcal{G}}(W_i)$, and if $\mathbb{P}(W_i | \, \mathrm{PA}_{\mathcal{G}}(W_i))$ is a conditional exponential family satisfying Definition 2, then $\mathbb{P}(W_i | \, \mathrm{PA}_{\mathcal{G}}(W_i))$ also satisfies this definition, where the function $\eta(\mathrm{PA}_{\mathcal{G}}(W_i), N(W_i))$ is constant with respect to fluctuation in the variables $N(W_i)$. Thus, taking $Z_i := \mathrm{PA}_{\mathcal{G}'}(W_i)$ as the conditioning set satisfies Assumption 1, and the rest of our results hold, where the corresponding δ-perturbations in Definition 4 are given by

$$\mathbb{P}_\delta(\mathbf{V}) = (\prod_{W_i \in \mathbf{W}} \mathbb{P}_{\delta_i}(W_i | \, \mathrm{PA}_{\mathcal{G}'}(W_i))) \prod_{V_i \in \mathbf{V} \setminus \mathbf{W}} \mathbb{P}(V_i | \, \mathrm{PA}_{\mathcal{G}}(V_i))$$

with shift function $s_i(\mathrm{PA}_{\mathcal{G}'}(W_i); \delta_i)$ that are parametric functions of causal parents in the modified graph $\mathcal{G}'$.

B.3 Domain-preserving parameterizations of shift

For both of the examples considered above, we did not need to restrict the magnitude of the additive change to $\eta(Z)$. However, in some cases, such as changing the variance of a conditional Gaussian, we have the restriction that $\eta_\delta(Z) = \eta(Z) + s(Z; \delta)$ must lie in the proper domain, e.g., we cannot consider a shift which causes the conditional variance to become negative. For a conditional Gaussian, we can consider unrestricted shifts in $\eta(Z)_1$, which controls the mean, because the mean has unrestricted domain. On the other hand, $\eta(Z)_2 = (-2\sigma^2(Z))^{-1}$ controls the variance, and must remain negative, such that $\eta(Z)_2 + s(Z; \delta)_2 < 0$ for the shifts we consider.

This can be resolved in one of two ways. First, one can consider parameterizations of $s(Z; \delta)$ which are guaranteed to preserve the correct domain with an additional constraint on the values of δ, such as the multiplicative shift below, which is sign-preserving for $\delta > -1$

$$\eta_\delta(Z)_2 = \eta(Z)_2 + \underbrace{\delta \eta(Z)_2}_{s(Z;\delta)} = (1 + \delta)\eta(Z)_2.$$

To handle the general case, at the expense of some additional complexity in the gradients of $s(Z; \delta)$, one can define the shifts as follows for parameters $\eta(Z)$ that have a lower bound L, with an equivalent formulation for shifts where the parameters have an upper bound, for any desired shift function $s'(Z; \delta)$

$$\eta(Z) + \underbrace{s'(Z; \delta) \cdot \mathrm{sigmoid}(\gamma \cdot [(\eta(Z) + s'(Z; \delta)) - (L + \epsilon)])}_{s(Z;\delta)}$$

where $\mathrm{sigmoid}(\gamma \cdot (x - (L + \epsilon)))$ is a smooth relaxation of the indicator function $\mathbf{1}\{x > L + \epsilon\}$, for a sufficiently large temperature parameter $\gamma > 0$ and a small $\epsilon > 0$. This transformation preserves the twice-differentiable nature of $s(Z; \delta)$. In practice, however, we typically evaluate the gradient of $s(Z; \delta)$ at $\delta = 0$, where $\eta(Z)$ does not lie at the boundary of allowable parameter space, such that we can consider simpler parameterizations like

$$\eta(Z) + \underbrace{s'(Z; \delta) \cdot \mathbf{1}\{\eta(Z) + s'(Z; \delta) > L + \epsilon\}}_{s(Z;\delta)}$$

as long as ϵ is taken sufficient small such that $\eta(Z) > L + \epsilon$ almost everywhere in $\mathbb{P}$.

C Considerations and additional results for evaluation of the worst-case loss

In this section, we present additional results on the Taylor approximation and compare how the Taylor approximation compares to the reweighting approach in evaluation and worst-case optimization of the shifted loss.

- In Appendix C.1 we give a full treatment of how shift gradients and Hessians are estimated from samples, following Theorem 1.
- In Appendix C.2, we demonstrate in some cases, one does not need to estimate all of $\eta(Z)$, but only the parts of $\eta(Z)$ that is shifting.
- In Appendix C.3, we demonstrate that the second-order Taylor expansion is exact in a linear-Gaussian setting, which gives a conceptual connection between this work and that of Anchor Regression [Rothenhäusler et al., 2021], which considered a restricted type of additive shift intervention in a globally linear structural causal model.
- In Appendix C.4, we work out the expression for the shift gradient and Hessian when we condition on binary variables.
- In Appendices C.5 to C.7, we provide experiments that compare the variance of the importance sampling estimate $\hat{E}_{\delta,\text{IS}}$ (see Equation (6)) to the variance of the Taylor estimate $\hat{E}_{\delta,\text{Taylor}}$ (see Equation (7)) of the loss in a shifted distribution.
- In Appendix C.8, we consider the bound in Theorem 2 in a covariate shift setting, and give an explicit expression for this under additional assumptions.

C.1 Algorithm for Estimation of Shift Gradients and Hessians

Here, we recall the form of the shift gradients and Hessians in Theorem 1, and demonstrate how to compute them in practice using a set of auxiliary regression functions fit to the validation data.

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 $\epsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}[T(W_i)|Z_i]$.

Notation and Dimensions: Let $\mathbf{W} = \{W_1, \ldots, W_m\}$ denote the set of m intervened variables, and let $\mathbf{Z} = \{Z_1, \ldots, Z_m\}$ denote the conditioning sets. Note that for a single $W_i \in \mathbb{R}^{d_{W_i}}$, we will generally have it that $Z_i \in \mathbb{R}^{d_Z}$, where d_W is the dimension of W (typically 1) and d_Z is the number of conditioning variables, and when considering n samples, W_i will be a matrix in $\mathbb{R}^{n \times d_W}$, and Z_i will be a matrix $\mathbb{R}^{n \times d_Z}$. The sufficient statistic $T_i(W_i)$ maps from $\mathbb{R}^{d_W}$ to $\mathbb{R}^{d_T}$, where d_T is the dimension of the sufficient statistic. For many common distributions, $T_i(W_i) = W_i$, the identity function. For others, like the conditional multi-variate Gaussian, $T_i(W_i) = [W_i, W_i W_i^\top]$, where $W \in \mathbb{R}^{d_W}$ and $W_i W_i^\top \in \mathbb{R}^{d_W \times d_W}$. In these cases, we squeeze $T_i(W_i)$ to be a single vector, so in this case $d_T = d_W + d_W^2$.

Auxiliary models: To estimate the shift gradients and Hessians, we first learn auxiliary predictive models, which are required for computing the relevant conditional covariances. For simplicity, we do not consider sample-splitting in the algorithm given below, but one could employ sample-splitting to learn these predictive models on an independent validation sample.

- For each W_i, we learn $\hat{\mu}_{W_i}(Z_i)$ as a regression model for $\mathbb{E}[T_i(W_i)|Z_i]$. Because $T_i(W_i)$ may have multiple dimensions, this is a function from $\mathbb{R}^{d_Z}$ to $\mathbb{R}^{d_T}$.

- For each conditioning set Z_i, we learn $\hat{\mu}_\ell(Z_i)$ as a regression model for $\mathbb{E}[\ell|Z_i]$. Because the loss is one-dimensional, this is a function from $\mathbb{R}^{d_Z}$ to $\mathbb{R}$.

We then construct the following, which are defined for each data point in the sample.

- For each W_i, we construct $\hat{\epsilon}_{T_i|Z_i} := T_i(W_i) - \hat{\mu}_{W_i}(Z_i)$, which is a vector of length d_{T_i}.
- For each conditioning set Z_i, for the loss ℓ, we construct $\hat{\epsilon}_{\ell|Z_i} := \ell - \hat{\mu}_\ell(Z_i)$, which is a real number.
- For each conditioning set Z_i, we compute $D_{i,1}(Z_i)$ as $\nabla_{\delta_i} s_i(Z_i;\delta_i)\big|_{\delta=0}$, which is a matrix of size $d_T \times d_{\delta_i}$, and a function of Z_i that we can evaluate on each sample.
- For each conditioning set Z_i, we compute $D_{i,2}(Z_i)$ as $\nabla^2_{\delta_i} s_i(Z_i;\delta_i)\big|_{\delta=0}$, which is a tensor of size $d_T \times d_{\delta_i} \times d_{\delta_i}$, and a function of Z_i that we can evaluate on each sample.

Estimating shift gradients The shift gradient and Hessian in Theorem 1 are expressed as conditional covariance. Since $\mathbb{E}[\mathrm{cov}(A, B|C)] = \mathbb{E}[\epsilon_{A|C}\epsilon_{B|C}]$ where $\epsilon_{A|C} := A - \mathbb{E}[A|C]$ and $\epsilon_{B|C} := B - \mathbb{E}[B|C]$, we can use the estimated conditional means above, to compute the shift gradient and Hessian. Suppose that we observe N samples, $n \in \{1, \dots, N\}$. For each index $i \in [m] := \{1, \dots, m\}$,

$$\hat{\mathrm{SG}}^1_i = \frac{1}{N}\sum_{n=1}^{N} \hat{\epsilon}^{(n)}_{\ell|Z_i} \cdot D_{i,1}(Z_i^{(n)})^\top \hat{\epsilon}^{(n)}_{T_i|Z_i}$$

which yields a vector of length d_{δ_i}, and these are concatenated together for each i to yield the entire shift gradient. The shift Hessian is constructed block-wise, for each index $i, j \in [m] \times [m]$ as follows: If $i = j$, then we construct the corresponding $d_{\delta_i} \times d_{\delta_i}$ block as

$$\hat{\mathrm{SG}}^2_{i,i} = \frac{1}{N}\sum_{n=1}^{N} \hat{\epsilon}^{(n)}_{\ell|Z_i} \cdot \left[\left(D_{i,1}(Z_i^{(n)})^\top \hat{\epsilon}^{(n)}_{T_i|Z_i}\right)^{\otimes 2} - D_{i,2}(Z_i^{(n)})^\top \hat{\epsilon}_{T_i|Z_i}\right]$$

where $v^{\otimes 2}$ denotes the outer product so that $v^{\otimes 2} = vv^\top$, and the transpose of $D_{i,2}$ refers to a transpose which has dimension $d_{\delta_i} \times d_{\delta_i} \times d_T$. On the other hand, if $i \neq j$ we have

$$\hat{\mathrm{SG}}^2_{i,j} = \frac{1}{N}\sum_{n=1}^{N} (\ell^{(n)} - \bar{\ell}) \cdot \left(D_{i,1}(Z_i^{(n)})^\top \hat{\epsilon}^{(n)}_{T_i|Z_i}\right)\left(D_{j,1}(Z_j^{(n)})^\top \hat{\epsilon}^{(n)}_{T_j|Z_j}\right)^\top$$

where $\bar{\ell}$ is the average value of ℓ in the validation sample.

C.2 Shifts where estimating all of $\eta(Z)$ is not necessary for estimating shift gradient and Hessian

The following example shows that when a shift occurs in an exponential conditional distribution with parameter $\eta(Z)$, we do not necessarily need to model all of $\eta(Z)$ in order to compute the shift gradient and Hessian. In particular, we only need to model the parts of $\eta(Z)$ that shift. This is different from estimating the shifted loss using importance sampling, where $\eta(Z)$ needs to be evaluated to evaluate Equation (5).

Example C.1. Consider the distribution of W conditioned on variables Z that is a multi-variate Gaussian variable,

$$W|Z = \mathcal{N}(\mu(Z), \Sigma(Z)),$$

for unknown functions μ, Σ. The sufficient statistic for the multivariate Gaussian distribution is $T(W) = (W, WW^\top)$ and the canonical parameter is $\eta(Z) = (\Sigma(Z)^{-1}\mu(Z), -\frac{1}{2}\Sigma(Z)^{-1})$.[6] The first component of $\eta(Z)$ is a signal-to-variance ratio and the second is the inverse covariance matrix. For a shift $(\delta, 0)$ that only affects the first component, we show that we do not need to model $\Sigma(Z)$, but only $\mu(Z)$. This is beneficial, since estimating a conditional covariance from data can be challenging, especially if W is high-dimensional.

[6] Or, more formally, $T(W) = \big(W, \mathrm{vec}(WW^\top)\big)$ and $\eta(Z) = \big(\sigma(Z)^{-1}\mu(Z), -\frac{1}{2}\,\mathrm{vec}(\mu(Z))\big)$, where vec denotes the vectorization operation. For a detailed walk through of the exponential family parameterization of multivariate Gaussian distributions, see https://maurocamaraescudero.netlify.app/post/multivariate-normal-as-an-exponential-family-distribution/.

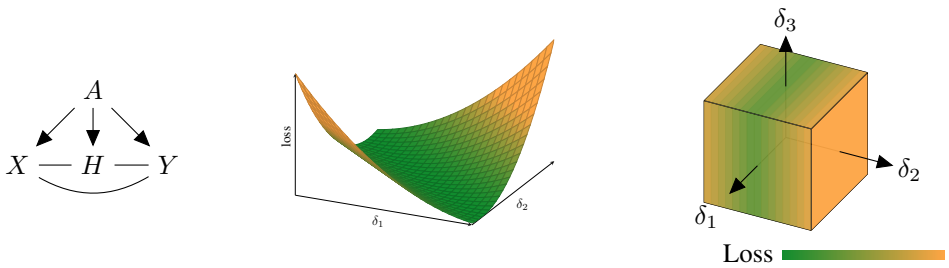


Figure 7: (Left) Graphical model assumed by Equation (11). The undirected edges represent either any directed configuration of directed edges or the dependence structures arising due to an acyclic SCM [Bongers et al., 2021]. (Middle) Plotting $\mathbb{E}_\delta[(Y - \gamma^\top X)^2]$ as a function of $\delta \in \mathbb{R}^2$ for a fixed predictor γ. (Right) Plotting $\mathbb{E}_\delta[(Y - \gamma^\top X)^2]$ as a function of $\delta \in \mathbb{R}^3$, with the loss indicated by the color. The loss only varies with changes in δ_2 (corresponding in Lemma C.1 to $v_\gamma \propto (0, 1, 0)^\top$).

For $\delta \in \mathbb{R}^{d_W}$, let $s(Z; \delta) = (\delta, 0)^\top$, and suppose that we wish to estimate $\mathbb{E}_\delta[\ell]$ using Equation (7). The derivative of s is given by

$$D_1 = \nabla^2_\delta s(Z; \delta) = \left(\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \right),$$

where the first block is a $d_W \times d_W$ diagonal matrix, and the second is a $d_W \times d_W^2$ matrix of zeros. The second derivative of s is $D_2 = 0$. Hence, using Theorem 1, the shift gradient is

$$\mathrm{SG}^1 = \mathbb{E}[D_1 \operatorname{cov}(\ell, (W, WW^\top)|Z)] = \mathbb{E}[\operatorname{cov}(\ell, W|Z)],$$

and

$$\begin{aligned} \mathrm{SG}^2 &= \mathbb{E}\left[D_1 \operatorname{cov}(\ell, \left(W - \mathbb{E}[W|Z], WW^\top - \mathbb{E}[WW^\top|Z]^{\otimes 2} \right)|Z) D_1^\top \right] \\ &= \mathbb{E}\left[\operatorname{cov}(\ell, \left(W - \mathbb{E}[W|Z] \right)^{\otimes 2}|Z) \right]. \end{aligned}$$

Conditional covariances can be computed by only residualizing one of the variables: $\mathbb{E}[\operatorname{cov}(A, B|C)] = \mathbb{E}[A(B - \mathbb{E}[B|C])]$. Thus, if we only residualize ℓ, we get

$$\mathrm{SG}^1 = \mathbb{E}[(\ell - \mathbb{E}[\ell|Z])W] \qquad \text{and} \qquad \mathrm{SG}^2 = \mathbb{E}[(\ell - \mathbb{E}[\ell|Z]) \cdot (W - \mu(Z))^{\otimes 2}].$$

Therefore, given data from $\mathbb{P}$, we can estimate the shift gradients by plugging in estimators $\hat{\mu}(Z)$ of $\mathbb{E}[W|Z]$ and $\hat{L}(Z)$ of $\mathbb{E}[\ell|Z]$. It follows that we do not need to model $\Sigma(Z)$ in order to estimate the shift gradients and Hessian at $\delta = 0$.

The story is different for a reweighting based estimator that seeks to estimate $\mathbb{E}_\delta[\ell]$ using importance sampling (see Section 3.1), where the weights are given by

$$w_{\eta,\delta}(Z) = (W - \mu(Z))^\top \delta - \tfrac{1}{2} \delta^\top \Sigma(Z) \delta,$$

and hence estimating $w_{\eta,\delta}(Z)$ requires estimation of $\Sigma(Z)$.

C.3 The quadratic approximation is exact, for mean shifts in linear models

We now consider data generated by a linear model, and show that the shifted loss is a quadratic function of δ, meaning that the Taylor approximation $E_{\delta,\text{Taylor}}$ is globally exact. Suppose that data is sampled from a linear structural causal model, and a shift in mean occurs in an variable A that does not have any causal parents. In particular, let A have a normal distribution with mean μ and finite variance and let

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + MA + \epsilon. \tag{11}$$

This is the model assumed by Rothenhäusler et al. [2021], and the corresponding graphical model is shown in Figure 7 (left). We consider the linear predictor $f_\gamma(X) = \gamma^\top X$ and the mean squared loss $\ell(f_\gamma(X), Y) = (Y - f(X))^2$. Due to the linearity of the model, the loss under a mean shift in A is quadratic [Rothenhäusler et al., 2021].

Lemma C.1. *Suppose $A \sim \mathcal{N}(\mu, \Sigma)$ and that (X, Y, H) are generated according to Equation (11). For $\gamma \in \mathbb{R}^{d_X}$ define $\ell := (Y - \gamma^\top X)^2$. Then there exist $v_\gamma, u_{\mu,\gamma} \in \mathbb{R}^{d_A}$ such that for all shifts $\delta \in \mathbb{R}^{d_A}$:*

$$\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell] + \delta^\top u_{\mu,\gamma} + \tfrac{1}{2}\delta^\top v_\gamma v_\gamma^\top \delta,$$

where $\mathbb{E}_\delta$ corresponds to taking the mean in the distribution where $A \sim \mathcal{N}(\mu + \delta, \Sigma)$. Further $u_{\mu,\gamma} = 0$ if $\mu = 0$.

Proposition C.1 elicits two properties of this linear model: First the loss is described by a quadratic function globally, i.e. also for very large δ. In Figure 7 (middle), we plot $\mathbb{E}_\delta[\ell]$ as a function of δ. We observe a 'valley' in the loss, in which the expected loss does not at all change with δ. This is a consequence of Lemma C.1, and particularly that if δ is orthogonal to both $u_{\mu,\gamma}$ and v_γ then $\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell]$. In higher dimensions $d_A > 2$, since $v_\gamma v_\gamma^\top$ has rank 1, the 'valley' persists in that the loss does not grow at all in $d_A - 2$ dimensions (or $d_A - 1$ if A has mean $\mu = 0$), see Figure 7 (right).

We now show that coefficients in the quadratic form in Lemma C.1 is equal to the shift gradient and Hessian. We use that the Gaussian distribution with known variance Σ can be parameterized as an exponential family with sufficient statistic $T(A) = \Sigma^{-1}A$ and parameter $\eta = \mu$.[7]

Proposition C.1. *Suppose $A \sim \mathcal{N}(\mu, \Sigma)$ and that (X, Y, H) are generated according to Equation (11). Then the shift gradient and Hessian are given by*

$$\mathrm{SG}^1 = \mathrm{cov}(\ell, \Sigma^{-1}A) \qquad \textit{and} \qquad \mathrm{SG}^2 = \mathrm{cov}(\ell, \Sigma^{-1}(A-\mu)(A-\mu)^\top \Sigma^{-\top})$$

and the loss under a mean shift of δ in A is given by

$$\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell] + \delta^\top \mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \mathrm{SG}^2\, \delta,$$

where $\ell := (Y - \gamma^\top X)^2$ and $\mathbb{E}_\delta$ corresponds to taking the mean in the distribution where $A \sim \mathcal{N}(\mu + \delta, \Sigma)$.

This elicits a connection to anchor regression [Rothenhäusler et al., 2021]: Under the generative model Equation (11) and using the quadratic loss $\ell = (Y - \gamma^\top X)^2$ for $\gamma \in \mathbb{R}^{d_X}$, they show that for any $\lambda \geq 0$, the worst-case loss $\mathbb{E}_\delta[\ell]$ over a set $\Delta = \{\delta | \delta\delta^\top \preceq \lambda \mathbb{E}[AA^\top]\}$ equals the objective $\ell_{\mathrm{AR}} = \mathbb{E}[\ell] + \lambda \mathbb{E}[\mathbb{E}[Y - \gamma^\top X | A]^2]$, which is computable from the observed distribution.

Because of Proposition C.1, ℓ_{AR} *also* equals the solution of the optimization problem Equation (9) over the constraint set Δ. Therefore minimizing the anchor regression objective over γ or minimizing Equation (9) over γ will lead to the same estimator. Since our proposed Taylor approximation in Equation (9) does not assume linearity, one could use the approximation to extend the rationale of anchor regression of minimizing the worst-case loss to non-linear models. This however comes at the cost of not optimizing the exact worst-case loss, but rather an approximation, whose quality is given by Theorem 2. Further, this would involving a minimax problem, minimizing Equation (9) over models f, and there are questions, such as convexity and tractability, which would need to be solved.

C.4 Estimating the shift gradient and Hessian for conditional on binary variables

To build intuition for the shift gradient and Hessian, we here give an example where we condition on variables Z that take a finite number of values and write out explicit expressions for the shift gradient and Hessian. However, we emphasize, that in most practical scenarios, one will not have to work out the shift gradient and Hessian explicitly, but can simply estimate them as covariances from the data (Theorem 1).

Example C.2 (Shift Function of Discrete Parents)**.** Consider a conditional distribution $W|Z$ where Z takes values in a finite set $\mathcal{Z}$. This is for instance the case if $Z = (Z_1, \ldots, Z_d)$ where each Z_i is binary, so $|\mathcal{Z}| = 2^d$. Instead of a shift $\eta(Z) + \delta$, where the parameter increases by the same

[7] It can also be parameterized as $T(A) = \Sigma^{-1/2}A, \eta = \Sigma^{-1/2}\mu$, which would yield the same result.

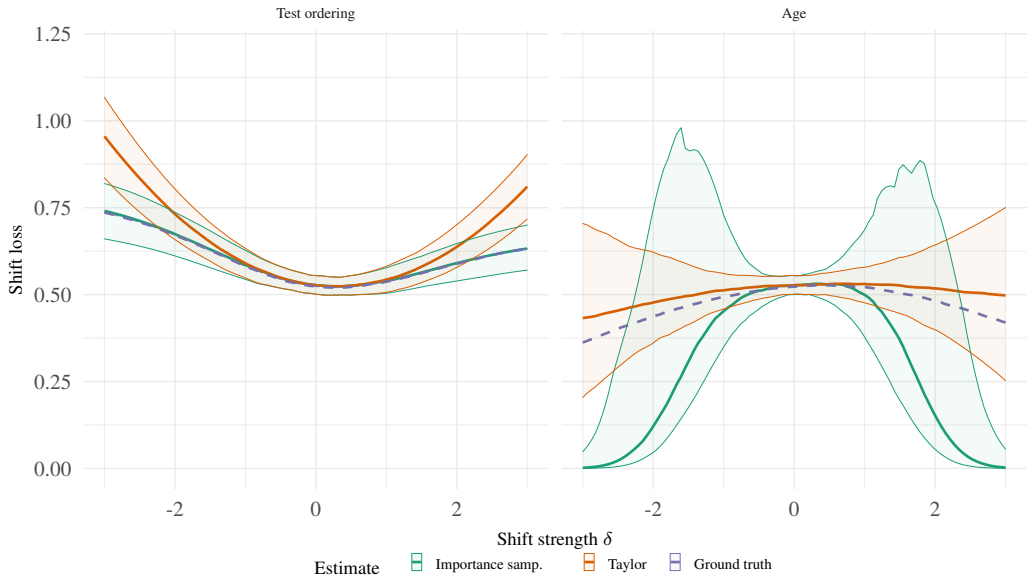


Figure 8: We plot the mean and confidence intervals of $\hat{E}_{\delta,\text{Taylor}}$ and $\hat{E}_{\delta,\text{IS}}$ when the shifted loss as in the lab test ordering example Example 1. (Left) We consider a shift in the logits of ordering lab tests from $\eta(Z)$ to $\eta(Z) + \delta_0$. (Right) We consider a shift in the mean of Age. In the observed distribution $\eta = \mu/\sigma = 0$ and we shift to a mean of $\eta = \delta$.

amount for all values of Z, we may consider a shift $\eta(Z) + s(Z;\delta)$ where $s(Z;\delta) = \sum_{z\in\mathcal{Z}} \delta_z 1_{Z=z}$, meaning that the shift is different in each category Z. Since $\eta(Z)$ only takes a finite number of variables, this shift corresponds to an arbitrary change in $\eta(Z)$.

$s(Z;\delta)$ is a differentiable function in δ, and if $d_T = 1$ the shift gradient is a (1×2^d)-row vector, $\nabla_\delta s(Z;\delta) = (1_{Z=z})_{z\in\mathcal{Z}}$, and the shift Hessian vanishes, $\nabla^2_\delta s(Z;\delta) = 0$. Enumerating $\mathcal{Z} = \{1,\dots,2^d\}$, the i'th entry in the shift gradient becomes

$$(\text{SG}^1)_i = \mathbb{E}\left[1_{Z=i}\operatorname{cov}\left(\ell, T(W)\middle|Z\right)\right] = \mathbb{P}(Z=i)\operatorname{cov}(\ell, T(W)|Z=i),$$

and the i,j'th entry of the shift Hessian becomes 0 if $j \neq i$ and else

$$(\text{SG}^2)_{i,i} = \mathbb{E}\left[1_{Z=i}\operatorname{cov}_\delta\left(\ell, \epsilon^{\otimes 2}_{T|Z}\middle|Z\right)\right] = \mathbb{P}(Z=i)\operatorname{cov}(\ell, \epsilon^{\otimes 2}_{T|Z}|Z=i).$$

Consider for example the case where both W and Z are binary. Then $T(W) = W$ and $s(Z;\delta) = 1_{Z=0}\delta_0 + 1_{Z=1}\delta_1$ and $s^{(1)} = (1_{Z=0}, 1_{Z=1})$ and $s^{(2)} = 0$. The conditional covariance can be evaluated by residualizing only one of the variables, $\mathbb{E}[\operatorname{cov}(A,B|C)] = \mathbb{E}[A(B - \mathbb{E}[B|C])]$, so we can chose to residualize only W (for SG^1) or $(W - \mathbb{E}[W|Z=i])^2$ (for SG^2). Finally, if we let $p_i = \mathbb{P}(W=1|Z=i)$ and use that $\mathbb{E}[W|Z=i] = p_i$ and $\mathbb{E}[(W-p_i)^2|Z=i] = \operatorname{var}(W|Z=i) = p_i(1-p_i)$, we get that

$$\text{SG}^1 = \mathbb{E}\left[\begin{pmatrix} p_0 \cdot \ell \cdot (W - p_0) \\ p_1 \cdot \ell \cdot (W - p_1) \end{pmatrix}\right],$$

and

$$\text{SG}^2 = \mathbb{E}\left[\begin{pmatrix} \ell p_0\{(W-p_0)^2 - p_0(1-p_0)\} & 0 \\ 0 & \ell p_1\{(W-p_1)^2 - p_1(1-p_1)\} \end{pmatrix}\right].$$

C.5 Comparison of variance of reweighting and Taylor estimates in the lab ordering example

To compare the bias and variance of the Taylor and the importance sampling estimates of the shifted loss, we simulate data from the following, artificial, generative model (which is the same generative

model that was used to construct the loss landscape in Figure 1 (right)).

$$\begin{aligned}
\text{Age} &\sim \mathcal{N}(0, 0.5^2) \\
\mathbb{P}(\text{Disease} = 1|\text{Age}) &= \text{sigmoid}(0.5 \cdot \text{Age} - 1) \\
\mathbb{P}(\text{Order} = 1|\text{Disease, Age}) &= \text{sigmoid}(2 \cdot \text{Disease} + 0.5 \cdot \text{Age} - 1) \\
\text{Test Result}|\text{Order} = 1, \text{Disease} &\sim \mathcal{N}(-0.5 + \text{Disease}, 1)
\end{aligned}$$

where if Order $= 0$, the test result is a placeholder value of zero.

We consider either a shift in the logits of ordering lab tests $\eta_\delta(Z) = \eta(Z) + \delta$ (Figure 8 left) or a mean shift in the Gaussian distribution of age $\eta_\delta = \delta$ (Figure 8 right). For each δ in a grid, we compute estimates $\hat{E}_{\delta,\text{IS}}$ and $\hat{E}_{\delta,\text{Taylor}}$ of the loss under a shift of size δ, We repeat this $n = 1{,}000$ times, and plot the mean and point-wise prediction intervals (the pointwise 0.05 and 0.95 quantiles) for $\hat{E}_{\delta,\text{IS}}$ and $\hat{E}_{\delta,\text{Taylor}}$. We also simulate ground truth data from $\mathbb{P}_\delta$, to compute the actual loss under shift.

For shifts in the binary variable (Figure 8, left), both estimates capture the loss well for small shifts, but as δ gets larger, the quadratic approximation increasingly deviates from the true mean; the importance sampling estimate remains very close to the ground truth shifted loss. On the contrary, for the Gaussian mean shift (Figure 8, right), the importance sampling weights are ill-behaved, and the variance dramatically increases as δ becomes larger. This supports the intuition, that while importance sampling tends to work well for binary variables, the variance can be large in continuous distributions, such as the Gaussian distribution.

C.6 Comparison of theoretical variance of reweighting and Taylor estimates

Example C.3. To demonstrate the reduction in variance obtained from using the Taylor approximation of the importance weights, we consider a simple example where $\mathbb{P}(X) \sim \mathcal{N}(0, 1)$ and $\mathbb{P}_\delta(X) \sim \mathcal{N}(\delta, 1)$ and we wish to estimate $\mathbb{E}_\delta[\ell(X)]$ for some loss function $\ell(X)$.[8] The importance sampling weights are given by $w_\delta(X) = \exp(-\frac{1}{2}\delta^2 + X \cdot \delta)$, and the shift gradient and Hessians are $\text{SG}^1 = \mathbb{E}[\ell(X)X]$ and $\text{SG}^2 = \mathbb{E}[\ell(X)X^2]$.

Therefore samples $X_1, \ldots, X_n$ from $\mathbb{P}$ consider the estimators, for any loss function $\ell(X)$, two estimators of $\mathbb{E}_\delta[\ell]$ are

$$\hat{\mu}_{\text{IS}} = \frac{1}{n}\sum_{i=1}^{n} w_\delta(X_i)\ell(X_i) \quad \text{and} \quad \hat{\mu}_{\text{Taylor}} = \frac{1}{n}\sum_{i=1}^{n} \ell(X_i) + \delta \cdot \ell(X_i)X_i + \tfrac{1}{2}\delta^2\ell(X_i)X_i^2,$$

and the variances of the estimators are

$$\begin{aligned}
\text{var}(\hat{\mu}_{\text{IS}}) &= \frac{\mathbb{E}[\{\ell(X + 2\delta)\}^2]}{n}\exp(\delta^2) \\
\text{var}(\hat{\mu}_{\text{Taylor}}) &= \frac{\text{var}\left(\ell(X) + \delta X\ell(X) + \frac{1}{2}\delta^2 X^2\ell(X)\right)}{n}.
\end{aligned}$$

The variance of $\hat{\mu}_{\text{Taylor}}$ grows like δ^4 and the variance of $\hat{\mu}_{\text{IS}}$ grows exponentially fast (unless $\mathbb{E}[\{\ell(X + 2\delta)\}^2]$ also diminishes exponentially fast, which is generally not the case), and so except for small δ, the variance of the importance sampling estimator will be orders of magnitude larger than the variance of the estimator using the Taylor approximation. While, $\hat{\mu}_{\text{IS}}$ is an unbiased estimator of $\mathbb{E}_\delta[\ell(X)]$ and $\hat{\mu}_{\text{Taylor}}$ is a biased, the overall mean squared error will be smaller for the Taylor approximation, unless the bias of the Taylor approximation also grows exponentially.

For the sake of analysis, consider the simple example $\ell(X) = X$. In this case, the Taylor estimate is unbiased because $\mathbb{E}_\delta[X] = \delta$ is a linear function of δ, so the quadratic approximation is adequate. Further, the variances are given by

$$\text{var}(\hat{\mu}_{\text{IS}}) = \frac{\exp(\delta^2)(1 + 4\delta^2) - \delta^2}{n} \quad \text{and} \quad \text{var}(\hat{\mu}_{\text{Taylor}}) = \frac{1 + 5\delta^2 + \frac{15}{4}\delta^4}{n}.$$

In particular, the variance of the importance sampling estimate grows like $\exp(\delta^2)$ while that of the Taylor estimate grows like δ^4.

[8] In practice one would not use importance sampling estimation for such a simple shift, but use other approaches, such as analytically work out an estimate of $\mathbb{E}_\delta[\ell]$.

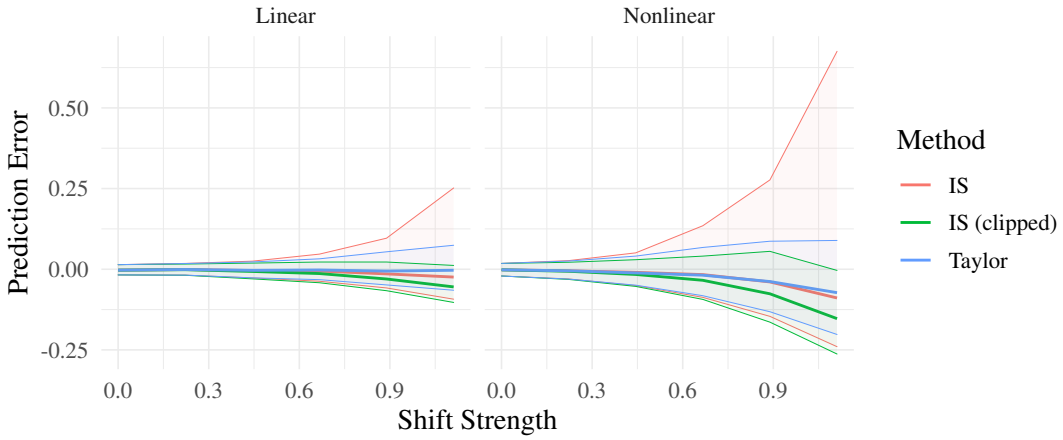


Figure 9: Median and quantiles of the error in predicting $\mathbb{E}_\delta[\ell]$ under a shift δ.

C.7 Comparison of variance of reweighting and Taylor estimates in a simple synthetic example

In this experiment, we compare the variance of importance sampling and Taylor estimates in a simple synthetic example. We simulate data from $\mathbb{P}$ where $X \in \mathbb{R}^3$ and $Y \in \mathbb{R}^1$ depend either linearly or quadratically on $W \in \mathbb{R}^3$,

$$W \sim \mathcal{N}(0, \mathrm{Id}_3) \qquad \text{and} \qquad \begin{pmatrix} X \\ Y \end{pmatrix} = (\mathrm{Id}_4 - B)^{-1} M(W + \alpha(W \odot W) + \epsilon),$$

where $\odot$ refers to entrywise multiplication, $\epsilon \sim \mathcal{N}(0, \mathrm{Id}_4)$, α is either 0 (linear) or $\frac{1}{2}$ (nonlinear) and

$$B := \begin{pmatrix} 2 & 1 & 0 & 1 \\ 2 & 2 & 0 & 3 \\ 3 & 3 & 0 & 2 \\ 4 & 2 & 4 & 0 \end{pmatrix} \qquad \text{and} \qquad M := \begin{pmatrix} 2 & 1 & 0 \\ 2 & 1 & 1 \\ 2 & 2 & 0 \\ 4 & 1 & 1 \end{pmatrix}.$$

On the simulated data from $\mathbb{P}$, we then fit a linear predictor $f(X)$ of Y, and consider a shift in the mean of W from $\mathbb{P}(W) \sim \mathcal{N}(0, \mathrm{Id}_3)$ to $\mathbb{P}_\delta(W) \sim \mathcal{N}(\delta, \mathrm{Id}_3)$, where $\delta = [s, s, s]^\top$ for some shift strength $s > 0$. We then compute the shift gradient $\mathrm{SG}^1 = \mathrm{cov}(\ell, W)$ and Hessian $\mathrm{SG}^2 = \mathrm{cov}(\ell, WW^\top)$, and approximate $\mathbb{E}_\delta[\ell]$ by $\hat{E}_{\delta,\text{Taylor}}$ (see Equation (7)). In the linear data, the Taylor approximation is exact (see Appendix C.3), such that any prediction error can be attributed to finite-sample fluctuation, whereas both model misspecification and finite-sample fluctuation contribute to the error in the nonlinear setting.

Similarly, we estimate $\mathbb{E}_\delta[\ell]$ by importance sampling, $\mathbb{E}_\delta[\ell] = \mathbb{E}[w_\delta(W)\ell] \approx \frac{1}{n}\sum w_\delta(W)\ell$, where $w_\delta(W) = \frac{\mathbb{P}_\delta(W)}{\mathbb{P}(W)} = \delta^\top W - \frac{1}{2}\delta^\top\delta$, and compare this to ground truth data sampled from $\mathbb{P}_\delta$; we do the same for an importance sampling estimator with weights 'clipped' at the 99% quantile.

We compare the predicted loss $\mathbb{E}_\delta[\ell]$ by actually simulating data from $\mathbb{P}_\delta$ and evaluating $\mathbb{E}_\delta[\ell]$ (where ℓ is still the model trained on data from $\mathbb{P}$). We then compute the prediction error, as the difference $\mathbb{E}_\delta[\ell] - \hat{E}_{\delta,\text{Taylor}}$ or $\mathbb{E}_\delta[\ell] - \hat{E}_{\delta,\text{IS}}$.

For a number of different shift strengths s, we repeat this procedure $M = 1{,}000$ times, and in Figure 9 we plot the median and a confidence interval defined by the 2.5 and the 97.5% quantiles of the prediction error.

In the linear case, bot the importance sampling and the Taylor approximation retains a median error close to 0, with the variance of $\hat{E}_{\delta,\text{IS}}$ being larger than $\hat{E}_{\delta,\text{Taylor}}$. The clipped importance sampling estimate has a smaller variance than that of ordinary importance sampling, though the median deviates further from 0, and the variance is not smaller than that of the Taylor estimate.

In the non-linear cases, all three models underestimate the shifted loss. For $\hat{E}_{\delta,\text{Taylor}}$, this happens because as the mean of W shift, the mean shift is amplified by the non-linearity, such that the

quadratic approximation of the loss is an underestimate. While the variance of the clipped importance sampling is smaller than the variance of the ordinary importance sampling estimate and comparable to the variance of the Taylor estimate, this prediction is further from 0 than the Taylor estimate.

Since importance sampling methods are known to produce very large outliers, the use of the median and quantiles, as opposed to the mean an confidence intervals based on the standard deviation, is favouring importance sampling; the Taylor method looks even more favourable if we instead plot the mean and standard deviations.

C.8 The bound in Theorem 2 under covariate shift

The bound in Theorem 2 is in a general form that applies to any shift in the CEF framework. In concrete cases, the bound can be made simpler, as we now demonstrate.

Suppose that X is a covariate that is Gaussian distributed $\mathcal{N}(0,1)$. Also consider a prediction target $Y \coloneqq f_0(X) + \epsilon$ for some function f_0 and noise variable ϵ that is independent of X.

Suppose we consider a predictor $\hat{Y} = f(X)$ and apply our proposed methodology to estimate the mean squared prediction error when predicting $Y \approx f(X)$ under a mean shift of size $\delta \in \mathbb{R}$ to X. When we only consider shifts in the mean (and not the variance), the sufficient statistic is $T(X) = X$. We can use Theorem 2 to bound the prediction error. In this setting,

$$\ell = (Y - \hat{Y})^2 = (f_0(X) - f(X) + \epsilon)^2 \quad \text{and} \quad \epsilon_{t\cdot\delta T} = X - t \cdot \delta,$$

such that the bound in Theorem 2 becomes

$$\begin{aligned}
&\left|\mathbb{E}_\delta[\ell] - E_{\delta,\text{Taylor}}\right| \\
&\leq \tfrac{1}{2} \sup_{t\in[0,1]} \Big| \operatorname{cov}_{t\cdot\delta}\left((f_0(X) - f(X) + \epsilon)^2, (X - t\cdot\delta)^2\right) \\
&\qquad - \operatorname{cov}\left((f_0(X) - f(X) + \epsilon)^2, (f_0(X) - f(X) + \epsilon)^2, X^2\right)\Big| \cdot \delta^2.
\end{aligned}$$

The subscript $\operatorname{cov}_{t\cdot\delta}$ indicates that the covariance is taken in the distribution $\mathcal{N}(t\cdot\delta, 1)$; instead we can write this in the observed distribution, and add $t\cdot\delta$ to X. Further, the terms relating to ϵ disappear, as they are independent of X. Thus, if we define the modelling error $g(x) = f_0(x) - f(x)$, we can write

$$\left|\mathbb{E}_\delta[\ell] - E_{\delta,\text{Taylor}}\right| \leq \tfrac{1}{2} \sup_{t\in[0,1]} \left|\operatorname{cov}\left(g(X + t\cdot\delta)^2 - g(X)^2, X^2\right)\right| \cdot \delta^2.$$

We can bound the covariance using the inequality $\operatorname{cov}(A,B) \leq \sqrt{\operatorname{var}(A)\operatorname{var}(B)}$,

$$\left|\mathbb{E}_\delta[\ell] - E_{\delta,\text{Taylor}}\right| \leq \tfrac{1}{2} \sup_{t\in[0,1]} \left|\sqrt{\operatorname{var}\left((g(X + t\cdot\delta)^2 - g(X)^2\right)}\right| \cdot \left|\sqrt{\operatorname{var}(X^2)}\right| \cdot \delta^2.$$

The first term on the right hand side is the variance of the difference of approximation error in X and in $X + t\delta$. If we are willing to make assumptions on the quality of the approximation f, we can simplify this further. For example, we can assume that $|g(x)^2 - g(y)^2| \leq C \cdot |x - y|^2$, meaning that the squared error of $f_0(x) - f(x)$ does not change faster than quadratically in x. In that case, we get

$$\left|\mathbb{E}_\delta[\ell] - E_{\delta,\text{Taylor}}\right| \leq \tfrac{1}{2} C \left|\sqrt{\operatorname{var}(X^2)}\right| \cdot \delta^4.$$

In some cases, one can sharpen this bound by using prior knowledge about the data generating mechanism (for example, the data generating function f_0 may be bounded).

D Limitations of worst-case conditional subpopulation shift for defining plausible robustness sets

For the example in Section 4.1, we can contrast the type of shift we consider with the worst-case $(1-\alpha)$-conditional subpopulation shift considered by Subbaswamy et al. [2021].

In this section, we will make the following points: First, worst-case conditional $(1-\alpha)$-subpopulation shifts can be too pessimistic, with even moderate values of α leading to implausible conditional distributions. Second, we will argue that parametric robustness sets enable more fine-grained control over the set of plausible shifts, leading to more informative estimates of worst-case risk. Overall, we argue that the two approaches are complementary, with different strengths.

Before we proceed, we define a conditional $(1-\alpha)$ subpopulation shift. A $(1-\alpha)$ subpopulation shift in the conditional distribution $\mathbb{P}(O|Y)$ is defined by a weighting function $h : \mathcal{O} \times \mathcal{Y} \mapsto [0,1]$, which has the property that $\mathbb{E}[h(O,Y)|Y] = 1-\alpha$ for all values of Y. This can be used to construct a worst-case objective, which measures the worst-case loss under such a shift:

$$\begin{aligned} \sup_{h:\{0,1\}^2 \mapsto [0,1]} \quad & \frac{1}{(1-\alpha)} \mathbb{E}[h(O,Y)\mu(O,Y)] \\ \text{s.t.} \quad & \mathbb{E}[h(O,Y)|Y=y] = 1-\alpha, \quad \text{for } y \in \{0,1\} \end{aligned} \tag{12}$$

where $\mu(O,Y) := \mathbb{E}[\ell(Y,f)|O,Y]$, for a predictor f and loss ℓ. This has the effect of leaving the distribution $\mathbb{P}(Y)$ untouched, while changing the conditional distribution $\mathbb{P}(O|Y)$. Throughout this section, we will use the same predictor $f(O,L)$ described in Section 4.1. The rest of this section is structured as follows:

In Appendix D.1, we derive the feasible set of conditional distributions $\mathbb{P}(O|Y)$ implicitly considered by this objective in the simple generative model of Section 4.1, which only involves variables O, L and Y. We do so by showing that (for discrete O, Y), maximizing Equation (12) over h is equivalent to solving a linear program, where we can characterize the constraints on h exactly, and translate them into constraints on $\mathbb{P}(O=1|Y=1), \mathbb{P}(O=1|Y=0)$. Here, we show that the resulting feasible set is quite large, even for moderately large subpopulations. In particular, whenever $(1-\alpha) < \min\{\mathbb{P}(O=1|Y=0), \mathbb{P}(O=0|Y=1)\}$, all conditional distributions are possible.

In Appendix D.2, we derive the value of h that maximizes Equation (12), and show that, as we vary α, the worst-case shift is always in the same "direction" probability space: Healthy patients $(Y=0)$ are tested more, and sick patients $(Y=1)$ are tested less, and for $\alpha < 0.27$, the worst-case subpopulation shift is the (unrealistic) scenario where healthy patients are always tested, and sick patients are never tested.

In Appendix D.3, we illustrate how this type of behavior can be avoided with our approach. We first give a parameterized shift function $s(Z;\delta_0,\delta_1)$ such that we can reach any conditional distribution of $\mathbb{P}(O|Y)$, for sufficiently large values of δ_0, δ_1. We then demonstrate how an iterative process might play out with domain experts, where we consider different constraint sets until we find a constraint set that contains plausible shifts.

D.1 Feasible conditional subpopulations in Section 4.1

For the simple example in Section 4.1, we give a self-contained derivation of the feasible region for $1-\alpha$ conditional subpopulations in the distribution $\mathbb{P}(O|Y)$. The advantage of working with this simple generative model is that the conditional distribution can be described by only two numbers, $\mathbb{P}(O=1|Y=1)$ and $\mathbb{P}(O=1|Y=0)$, and so we can visualize the resulting conditional distribution.

Because O, Y are discrete, the worst-case subpopulation in this simple example can be solved via a linear program, for a fixed α. We have an optimization problem in two variables, since $h_{11}\mathbb{P}(O=1|Y=1) + h_{01}\mathbb{P}(O=0|Y=1) = 1-\alpha$, and likewise for h_{10}, h_{00}, where $h_{ij} = h(O=i, Y=j)$. We also have the constraint that each variable must live in $[0,1]$. Meanwhile, the loss to maximize is a linear function, as an expectation of $\mathbb{E}[h(O,Y)\mu(O,Y)]$, where $\mu(O,Y)$ takes on four possible values, where we write $p_{ij} = \mathbb{P}(O=i|Y=j)$, and μ_{ij} similarly.

$$\begin{aligned} \max_{h \in \mathbb{R}^{2\times 2}} \quad & h_{00}\mu_{00} + h_{10}\mu_{10} + h_{01}\mu_{01} + h_{11}\mu_{11} \\ \text{s.t.,} \quad & h_{11}p_{11} + h_{01}(1-p_{11}) = 1-\alpha \\ & h_{10}p_{10} + h_{00}(1-p_{10}) = 1-\alpha \\ & 0 \le h_{ij} \le 1, \forall i,j \end{aligned} \tag{13}$$

This linear program is simple enough to solve by hand, and we will do here to build intuition. In this section, we begin by characterizing the feasible region of h, and then translating that into a feasible region for $\mathbb{P}_h(O|Y)$, which we can plot in two dimensions.

Characterizing feasible values of h: Here, we focus on characterizing the feasible set that h can lie in, as a way of characterizing the feasible set for $\mathbb{P}(O|Y)$. From the constraints, we can write that

$$h_{11}p_{11} + h_{01}(1 - p_{11}) = 1 - \alpha \qquad \Longrightarrow \qquad h_{01} = \frac{1 - \alpha - h_{11}p_{11}}{1 - p_{11}}$$

$$h_{10}p_{10} + h_{00}(1 - p_{10}) = 1 - \alpha \qquad \Longrightarrow \qquad h_{00} = \frac{1 - \alpha - h_{10}p_{10}}{1 - p_{10}}$$

There are only two constraints on h_{11}: Those directly imposed by $0 \leq h_{11} \leq 1$, and those which are imposed by the equality constraint with h_{01} and the fact that $0 \leq h_{01} \leq 1$. For the latter, with some algebra we can write that

$$0 \leq \frac{1 - \alpha - h_{11}p_{11}}{1 - p_{11}} \leq 1 \qquad \Longrightarrow \qquad \frac{p_{11} - \alpha}{p_{11}} \leq h_{11} \leq \frac{1 - \alpha}{p_{11}}$$

So that the constraints on h_{11} become

$$\max\left\{0, \frac{p_{11} - \alpha}{p_{11}}\right\} \leq h_{11} \leq \min\left\{1, \frac{1 - \alpha}{p_{11}}\right\} \tag{14}$$

which recovers our intuition that if $\alpha = 0$, it must be that $h_{11} = 1$ and $h_{01} = 1$.

Bounding feasible values of $\mathbb{P}_h(O|Y)$ The parameters h can be understood as importance weights whose expectation is $1 - \alpha$ instead of 1, that reweight $\mathbb{P}$ to a new distribution $\mathbb{P}_h$ when appropriately normalized. To compute conditional probabilities $\mathbb{P}_h(O = i|Y = j)$ under the new distribution, we can compute the expectation of $\mathbf{1}\{O = i, Y = j\}$, and normalize by $\mathbb{P}(Y = j)$.

$$\mathbb{P}_h(O = i, Y = j) = \frac{1}{1 - \alpha}\mathbb{E}[h(O, Y)\mathbf{1}\{O = i, Y = j\}] = \frac{h_{ij}}{1 - \alpha}\mathbb{P}(O = i, Y = j)$$

$$\Longrightarrow \mathbb{P}_h(O = i|Y = j) = \frac{h_{ij}}{1 - \alpha}\mathbb{P}(O = i|Y = j)$$

where the implication follows from the fact that $\mathbb{P}_h(Y) = \mathbb{P}(Y)$. This allows us to translate bounds on h_{ij} directly into bounds on $\mathbb{P}_h(O = i|Y = j)$. Making use of Equation (14), we can write that

$$\max\left\{0, \frac{p_{11} - \alpha}{p_{11}}\right\} \cdot \frac{p_{11}}{1 - \alpha} \leq \mathbb{P}_h(O = 1|Y = 1) \leq \min\left\{1, \frac{1 - \alpha}{p_{11}}\right\} \cdot \frac{p_{11}}{1 - \alpha}$$

which yields

$$\max\left\{0, \frac{p_{11} - \alpha}{1 - \alpha}\right\} \leq \mathbb{P}_h(O = 1|Y = 1) \leq \min\left\{\frac{p_{11}}{1 - \alpha}, 1\right\}$$

We can apply a similar logic to h_{10}, which is identical except for p_{11} being replaced by p_{10}, yielding

$$\max\left\{0, \frac{p_{10} - \alpha}{1 - \alpha}\right\} \leq \mathbb{P}_h(O = 1|Y = 0) \leq \min\left\{\frac{p_{10}}{1 - \alpha}, 1\right\}$$

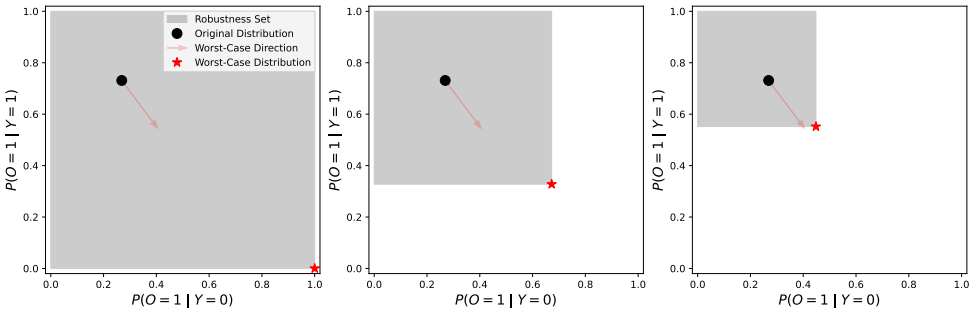
Visualizing the constraint set: Figure 10 gives feasible conditional distributions under different values of α. We can observe that when $\alpha = 0.8$, all conditional distributions are feasible, including the distribution where $\mathbb{P}(O = 1|Y = 0) = 1$ and $\mathbb{P}(O = 1|Y = 1) = 0$, representing the case where every healthy patient gets tested, and no sick patients receive a test. This is generally possible in this example whenever $1 - \alpha < \min\{\mathbb{P}(O = 1|Y = 0), \mathbb{P}(O = 0|Y = 1)\}$, as it permits the following subpopulation function, which yields this result.

$$h(O = o, Y = y) = \frac{1 - \alpha}{\mathbb{P}(O = o|Y = y)}\mathbf{1}\{o \neq y\}$$

D.2 Worst-case conditional subpopulation shifts

Given the constraint set which describes the feasible set of conditional distributions under the $(1 - \alpha)$-conditional subpopulation objective, we can derive the worst-case conditional distribution. Here, since Y, O are both binary, the expected loss under a new distribution $\mathbb{P}_h$ is given by

$$\mathbb{E}_h[\ell] = \sum_{y,o} \mu(o, y)\mathbb{P}_h(O = o|Y = y)\mathbb{P}(Y = y)$$



(a) $(1-\alpha) = 0.2$ (b) $(1-\alpha) = 0.4$ (c) $(1-\alpha) = 0.6$

Figure 10: Feasible sets, worst-case directions, and worst-case solutions for a $(1-\alpha)$ subpopulation shift in the conditional distribution $\mathbb{P}(O|Y)$ for differing values of α. Worst-case directions are computed using Equation (15), as unit-norm vectors re-scaled to fit in the plot, and the colored dots give the worst-case solutions, all of which lie in the lower-right corner of the constraint set. The original conditional distribution is given by the black dot.

which we can write in terms of the constrained probabilities $\mathbb{P}_h$ as follows, where $q_{11} := \mathbb{P}_h(O = 1|Y = 1)$ and $q_{10} := \mathbb{P}_h(O = 1|Y = 0)$

$$\mathbb{P}(Y = 1)[\mu(1,1)q_{11} + \mu(0,1)(1 - q_{11})] + \mathbb{P}(Y = 0)[\mu(1,0)q_{10} + \mu(0,0)(1 - q_{10})]$$

which also gives us a direction in which the loss is maximized, since the loss is given by

$$\mathbb{E}_h[\ell] = q_{11} \cdot \mathbb{P}(Y = 1) \cdot (\mu(1,1) - \mu(0,1)) + q_{10}\mathbb{P}(Y = 0) \cdot (\mu(1,0) - \mu(0,0)) + C \tag{15}$$

where $C = \mathbb{P}(Y = 1)\mu(0,1) + \mathbb{P}(Y = 0)\mu(0,0)$. Since q_{11}, q_{10} can be optimized independently, the worst-case solution is given by taking the maximum value of q_{11} if $\mu(1,1) > \mu(0,1)$ and the minimum value if $\mu(1,1) < \mu(0,1)$, and likewise taking the maximum value of q_{10} if $\mu(1,0) > \mu(0,0)$, and the minimum value otherwise. If $\mu(1,1) = \mu(0,1)$ or $\mu(1,0) = \mu(0,0)$, then the objective is unaffected by the choice of q_{11} or q_{10} respectively.

Visualizing the worst-case conditional distributions The worst-case directions on the probability scale, and the resulting worst-case conditional distribution obtained by solving Equation (13), are given in Figure 10. The red line arrow visualizes the direction from Equation (15), and the worst-case distribution is the point which is furthest in this direction in the constraint set. Here, we are finding the worst-case accuracy of the same predictive model $f(O, L)$ described in Section 4.1. We can observe that the worst-case loss is obtained by seeking to reverse the correlation between Y and O, decreasing the probability that a sick patient $(Y = 1)$ gets a test ordered, and increasing the probability that a healthy patient $(Y = 0)$ gets a test ordered.

D.3 Iterating with domain experts to define realistic parametric robustness sets

In the previous sections, we saw that $(1-\alpha)$-conditional subpopulation shift does not always produce realistic worst-case conditional distributions. Moreover, given only the parameter α, there is limited ability to control the nature of the resulting worst-case conditional distribution $\mathbb{P}(O|Y)$. In this section, we contrast this limitation with the finer-grained control enabled by considering parametric robustness sets. In particular, we argue that parametric shifts allow for end-users to customize robustness sets, ruling out shifts that represent unrealistic changes.

In practice, we imagine that the following iterative process could be a useful tool in model development: (i) Define a class of shifts with an appropriate $s(Z; \delta)$ and constraint set Δ, and search for a worst-case shift δ. (ii) Present to domain experts **both** the worst-case shift δ (in terms of summary statistics of the resulting distribution $\mathbb{P}_\delta$) alongside the associated estimate of the worst-case loss. For instance, report both the worst-case loss, as well as corresponding rate of testing among sick and healthy patients. (iii) If the shift itself is unrealistic, further the constrain parameter set or shift function, and repeat the process.

In Figure 11, we give a concrete example. Each sub-figure shows the set of conditional probability distributions $\mathbb{P}(O|Y)$ that can be represented by a shift of $(\delta_0, \delta_1) \in \Delta_0 \times \Delta_1$, along with the

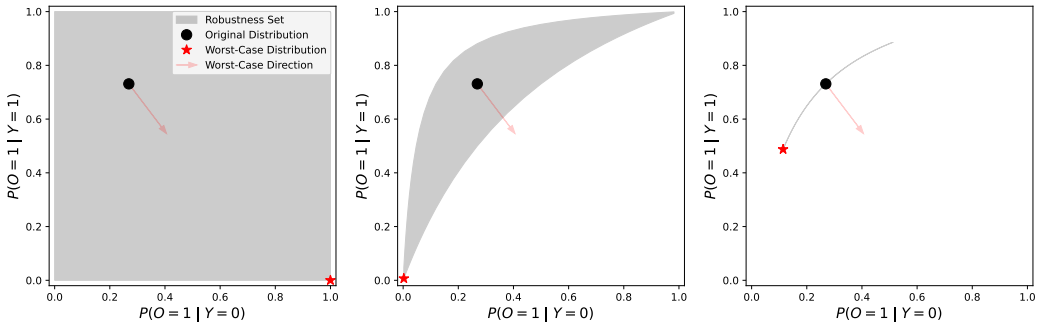


Figure 11: Each figure shows the set of conditional probability distributions ("CPDs") $\mathbb{P}(O|Y)$ that can be represented by a shift of $(\delta_0, \delta_1) \in \Delta_0 \times \Delta_1$, along with the worst-case distribution (given by the red star) for the 0–1 loss. In this example, the expected loss under $\mathbb{P}_\delta$ is a linear function of the two conditional probabilities (see Appendix D.2), where the loss increases along the red arrow. (a) captures (nearly) all conditional probability distributions, with Δ_0, Δ_1 unconstrained. (b) shows a set of CPDs with Δ_0 unconstrained, and $\Delta_1 = [-1, 1]$, with resulting worst-case accuracy of 50%. (c) shows a more restrictive set of shifts, where $\Delta_0 = [-1.05, 1.05], \Delta_1 = \{0\}$. The worst-case accuracy in this case is 69%, comparable to the accuracy of 75% on the original distribution.

worst-case conditional distribution (given by the red star) for the 0–1 loss. Recall that we use the shift function $s(Y; \delta) = \delta_0 + \delta_1 Y$, where δ_0 controls a general increase or decrease in testing, while δ_1 controls a shift in the testing rate for only sick patients, and allows for a different change in the testing rate of sick vs healthy patients.

Iteration 1: We might imagine starting with a relatively unconstrained robustness set, where δ_0 and δ_1 are unconstrained. Figure 11a shows the resulting robustness set of conditional distributions, and finds a shift with with a worst-case accuracy of 16%, compared to accuracy of 75% on the original distribution. However, the corresponding δ-perturbation $\mathbb{P}_\delta$ is unrealistic, where all healthy patients (and no sick patients) are tested. Luckily, because we have parameterized the shift, we can constrain the robustness set to exclude these types of results.

Iteration 2: A benefit of our approach is that we can refine the robustness set, with this type of feedback in mind. In Figure 11b, we restrict the support of δ_1 to $[-1, 1]$, to avoid large changes in the relative probability of testing sick vs healthy patients. Here, the resulting worst-case accuracy is much higher (50%), but the corresponding worst-case conditional probability distribution is perhaps still unrealistic: No patients undergo laboratory testing at all!

Iteration 3: Finally, we consider only shifts that affect all patients in a similar way, generally raising or lowering the conditional probability of a lab test, represented by shifts in δ_0 alone. This may correspond to a more realistic scenario where (in a new hospital) laboratory testing use is more or less constrained. Additionally, we can specify that this shift should decrease testing rates by at most 20%, which translates directly into a lower-bound on δ_0.[9] Figure 11c shows the resulting robustness set of distributions, where the worst-case shift may seem more plausible: A reduction in testing rates for both populations. The worst-case accuracy in this case is 69%, comparable to the accuracy of 75% on the original distribution.

E CelebA: Experiment details and additional results

In this section, we give details of the computer vision experiment in Section 4.2.

E.1 Details for the experiment

Creating the training distribution To construct the training distribution $\mathbb{P}$, we use the conditional GAN in Kocaoglu et al. [2018]. In particular, we use their CausalBEGAN, which is an extends the

[9] In Proposition B.1, we prove that for binary random variables with a shift $\eta(Z) + \delta$, there is a one-to-one mapping between a new marginal distribution ($\mathbb{P}(O = 1)$ in this case) and the value of the parameter δ.

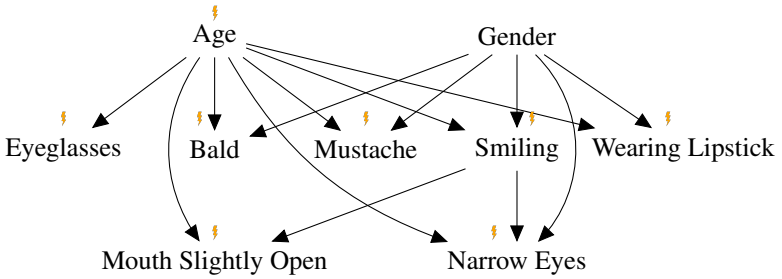


Figure 12: Causal graph over attributes, where lightning bolts indicate changes in mechanisms. Also displayed in Figure 4.

boundary equillibrium GAN [Berthelot et al., 2017] to also take attributes as inputs. We train the CausalBEGAN using the default hyper parameters in the implementation provided by Kocaoglu et al. [2018], available under the MIT license. The model is trained for 250,000 iterations on a single GPU, taking around approximately 16 hours.

Similar to Kocaoglu et al. [2018], we use the CelebA dataset [Liu et al., 2015], which contains approximately 200,000 images of faces, along 40 binary attributes. Of those, we use the following 9 attributes {Male, Young, Wearing Lipstick, Bald, Mustache, Eyeglasses, Narrow Eyes, Smiling, Mouth Slightly Open}. The CelebA dataset is licensed for non-commercial research purposes only, and consists of publicly available images of celebrities, which were collected from the internet. Although the data set has been widely used, Liu et al. [2015] do not make any mention of consent by the individuals to have the images included in the data set, and it is therefore likely that those celebrities did not provide consent.

Training distribution over attributes For the training distribution, we simulate binary attributes according to the structural causal model in Figure 4 (for convenience also copied to Figure 12), where the model parameters are

$$
\begin{aligned}
\mathbb{P}(\text{Young} = 1) &= \sigma(0.0) \\
\mathbb{P}(\text{Male} = 1) &= \sigma(0.0) \\
\mathbb{P}(\text{Eyeglasses} = 1 | \text{Young}) &= \sigma(0.0 - 0.4 \cdot \text{Young}) \\
\mathbb{P}(\text{Bald} = 1 | \text{Young, Male}) &= \sigma(-3.0 + 3.5 \cdot \text{Male} - \text{Young}) \\
\mathbb{P}(\text{Mustache} = 1 | \text{Young, Male}) &= \sigma(-2.5 + 2.5 \cdot \text{Male} - \text{Young}) \\
\mathbb{P}(\text{Smiling} = 1 | \text{Young, Male}) &= \sigma(0.25 - 0.5 \cdot \text{Male} + 0.5 \cdot \text{Young}) \\
\mathbb{P}(\text{Wearing Lipstick} = 1 | \text{Young, Male}) &= \sigma(3.0 - 5.0 \cdot \text{Male} - 0.5 \cdot \text{Young}) \\
\mathbb{P}(\text{Mouth Slightly Open} = 1 | \text{Young, Smiling}) &= \sigma(-1.0 + 0.5 \cdot \text{Young} + \text{Smiling}) \\
\mathbb{P}(\text{Narrow Eyes} = 1 | \text{Male, Young, Smiling}) &= \sigma(-0.5 + 0.3 \cdot \text{Male} + 0.2 \cdot \text{Young} + \text{Smiling}),
\end{aligned}
$$

where each variable either takes the value 0 or 1 and σ indicates the sigmoid. To generate data, we first simulate attributes from this binary Bayesian network, which we then pass as inputs to the GAN to simulate images (in addition to the random noise used by the GANs to simulate different images). In Figures 13 and 14, we plot examples of the training images that were generated.

Predictive model We simulate a training set of 12,000 attribute-image pairs, and a validation set of 2,000 pairs. The training set is used to fit a classifier f, and the validation set is used for model selection. To build a classifier f, we use the ResNet-50 [He et al., 2016] model implemented in the python package `torch`. We add a final fully connected layer to adapt the ResNet model to a binary classification task, and fine-tune the model on the training data by (only) learning the weights and bias of the final layer. The model is trained using the negative log-likelihood criterion and an ADAM optimizer. The model is trained for 25 epochs and we select the model which after a full epoch had the best validation set performance. Given the learned model f, we simulate a separate validation dataset of $n = 1{,}000$ samples, and make model predictions $f(X)$. We then compute the model accuracy as $\ell = \mathbf{1}\{f(X) = Y\}$, which is the input to computing the shift gradient and Hessian.

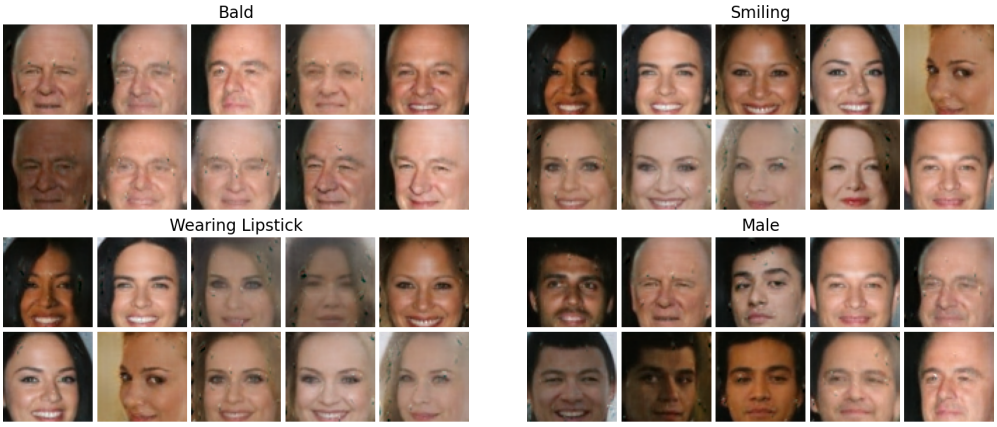


Figure 13: Examples of images from the training distribution $\mathbb{P}$. Each of the four groups (Bald, Smiling, Wearing Lipstick, Male) show training images who have that characteristic.

Estimation of shifted loss We apply the methods in Section 3.2 to estimate the worst-case shift to the distribution $\mathbb{P}$ (given by the binary probabilities above). For each conditional $\mathbb{P}(W_i | \mathrm{PA}(W_i))$, we consider a shift $\eta_{\delta_i}(\mathrm{PA}(W_i)) = \eta(\mathrm{PA}(W_i)) + \sum_{z \in \mathcal{Z}} \mathbf{1}\{\mathrm{PA}(W_i) = z\} \delta_i$, which corresponds to arbitrarily shifting the conditional distribution (see Appendix C.4). For example, for $W_i = \text{Bald}$, where $\eta(\text{Young}, \text{Male}) = -3.0 + 3.5 \cdot \text{Male} - 1.0 \cdot \text{Young}$, the shift would be

$$\eta_{\delta_{\text{Bald}}}(\text{Young}, \text{Male}) = \eta(\text{Young}, \text{Male}) + \begin{cases} \delta_{\text{Bald},0}, & \text{Young} = 0, \text{Male} = 0 \\ \delta_{\text{Bald},1}, & \text{Young} = 0, \text{Male} = 1 \\ \delta_{\text{Bald},2}, & \text{Young} = 1, \text{Male} = 0 \\ \delta_{\text{Bald},3}, & \text{Young} = 1, \text{Male} = 1. \end{cases} \tag{16}$$

For each W_i, this means that δ_i is $\mathbb{R}^{2^{|\mathrm{PA}(W_i)|}}$, and in total $\delta = (\delta_1, \ldots, \delta_8) \in \mathbb{R}^{31}$ (we do not consider shifts in the distribution of gender, since this is the label we are predicting).

We compute the shift gradient and Hessian using Theorem 1. In particular, since W_i is binary, the sufficient statistic is $T(W_i) = W_i$, so the shift gradients and Hessians given by Appendix C.4. See Appendix C.1 for a detailed walk through of computing the shift gradient and Hessian from a sample.

For any given δ, the shifted distribution of W_i is given by $\mathbb{P}_\delta(W_i = 1 | \mathrm{PA}(W_i)) = \sigma(\eta_{\delta_i})$, where η_{δ_i} is computed similar to Equation (16), and σ is the sigmoid function. Then the importance sampling weights are given by

$$w_\delta = \prod_{i=1}^{8} \frac{\sigma(\eta_{\delta_i}(\mathrm{PA}(W_i)))}{\sigma(\eta(\mathrm{PA}(W_i))}.$$

Using these weights, for any δ, we can estimate $\mathbb{E}_\delta[\ell]$ by $\hat{E}_{\delta,\text{IS}}$ and $\hat{E}_{\delta,\text{Taylor}}$ using Equations (6) and (8), respectively.

E.2 Full table of worst-case shift in Section 4.2

In Section 4.2, we find the worst-case shift δ, and display the 5 largest components. In Table 3, we display the full vector $\delta \in \mathbb{R}^{31}$, sorted by absolute value of the size of the component.

E.3 Sample images from training distribution in Section 4.2

In Figure 13, for the 4 attributes {Bald, Smiling, Wearing Lipstick, Male}, we display images generated from the training distribution $\mathbb{P}$ (i.e. by the GAN) with that particular attribute. In Figure 14 we show 10 randomly drawn images from the training distribution $\mathbb{P}$ as well as the test distribution $\mathbb{P}_\delta$ corresponding to the worst-case δ found in Section 4.2.

Conditional	δ_i
Bald | Male= 0, Young= 0	0.899
Bald | Male= 1, Young= 1	-0.800
Bald | Male= 1, Young= 0	-0.680
Wearing Lipstick | Male= 0, Young= 1	-0.618
Wearing Lipstick | Male= 0, Young= 0	-0.543
Eyeglasses | Young= 1	0.507
Mustache | Male= 1, Young= 0	-0.476
Mustache | Male= 0, Young= 0	0.449
Mustache | Male= 1, Young= 1	-0.415
Eyeglasses | Young= 0	0.399
Smiling | Male= 0, Young= 0	-0.261
Wearing Lipstick | Male= 1, Young= 0	0.205
Narrow Eyes | Male= 0, Smiling= 0, Young= 0	0.192
Mouth Slightly Open | Smiling= 1, Young= 1	0.191
Smiling | Male= 1, Young= 0	0.183
Narrow Eyes | Male= 1, Smiling= 1, Young= 1	0.179
Mouth Slightly Open | Smiling= 0, Young= 1	-0.153
Mustache | Male= 0, Young= 1	0.133
Bald | Male= 0, Young= 1	0.128
Mouth Slightly Open | Smiling= 1, Young= 0	-0.127
Narrow Eyes | Male= 0, Smiling= 1, Young= 0	-0.125
Wearing Lipstick | Male= 1, Young= 1	0.123
Narrow Eyes | Male= 1, Smiling= 1, Young= 0	-0.117
Narrow Eyes | Male= 0, Smiling= 0, Young= 1	0.106
Young | No parents	0.092
Narrow Eyes | Male= 0, Smiling= 1, Young= 1	0.057
Narrow Eyes | Male= 1, Smiling= 0, Young= 1	-0.050
Narrow Eyes | Male= 1, Smiling= 0, Young= 0	-0.039
Mouth Slightly Open | Smiling= 0, Young= 0	0.028
Smiling | Male= 1, Young= 1	0.028
Smiling | Male= 0, Young= 1	0.017

Table 3: Worst case shift in the $\delta \in \mathbb{R}^{31}$ identified by the Taylor approach in Section 4.2. Each entry corresponds to a shift in a conditional distribution given a particular outcome, and the squared sum of the entries equal $\lambda^2 = 4$.

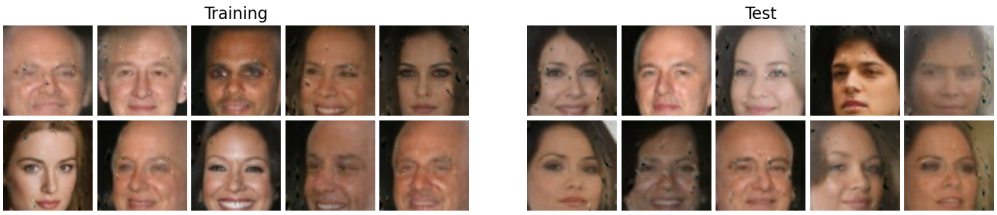


Figure 14: Examples of images from the training distribution $\mathbb{P}$ and the test distribution $\mathbb{P}_\delta$ that is characterized by the worst-case shift δ, see Figure 4.

E.4 Impact of changing λ

The shift considered in the main text yields a relatively small drop in accuracy. To demonstrate that larger drops in accuracy are possible, we repeated our experimental setup over the same 100 initial validation datasets, while varying the size of the constraint $\|\delta\|_2 \leq \lambda$. We report results in Table 4 for $\lambda \in [2, 4, 6, 8, 10]$, where $\lambda = 2$ corresponds to the setting of Table 1 (right).

Table 4: Performance of the Taylor and IS approaches over different values of λ, where $\lambda = 2$ corresponds to the setting of Table 1 (right). Averages taken over 100 simulations.

	$\lambda = 2$	$\lambda = 4$	$\lambda = 6$	$\lambda = 8$	$\lambda = 10$
Original Acc. ($\mathbb{E}[\mathbf{1}\{f(X) = Y\}]$)	0.912	0.912	0.912	0.912	0.912
Acc. under Taylor shift ($\mathbb{E}_{\delta_{\text{Taylor}}}[\mathbf{1}\{f(X) = Y\}]$)	0.874	0.812	0.736	0.681	0.648
IS est. of acc. under Taylor shift ($\hat{E}_{\delta_{\text{Taylor}},\text{IS}}$)	0.863	0.795	0.715	0.658	0.625
Taylor est. of acc. under Taylor shift ($\hat{E}_{\delta_{\text{Taylor}},\text{Taylor}}$)	0.863	0.798	0.711	0.601	0.466
Acc. under IS shift ($\mathbb{E}_{\delta_{\text{IS}}}[\mathbf{1}\{f(X) = Y\}]$)	0.889	0.830	0.746	0.670	0.596
IS est. of acc. under IS Shift ($\hat{E}_{\delta_{\text{IS}},\text{IS}}$)	0.821	0.670	0.463	0.264	0.130

Recall that we have two complementary goals: First, we would like to **find** a shift that results in a large drop in accuracy. Second, we would like to reliably **evaluate** the impact of the shift that we find, using only the training data. These two goals can be tackled with different approaches, such as using the Taylor approximation to find a shift, but using importance sampling (IS) to estimate the loss under that shift. Table 4 allows us to compare three different strategies: (i) using the Taylor approximation for both finding and evaluating the shift, (ii) using IS for both finding and evaluating, and (iii) using Taylor to find, but IS to evaluate the shift.

From Table 4, we can observe that **using Taylor to find, but IS to evaluate, consistently performs best in terms of reliable evaluation** (i.e., predicting the shifted accuracy), across all values of λ. For $\lambda = 2$, the bias in evaluation is 1% (predicting 86% vs ground truth of 87% on average), and for $\lambda = 10$, the bias of this approach is still only 2% (predicting 63% vs ground truth of 65% on average). In contrast, for $\lambda = 10$, the first strategy (using Taylor to find and evaluate) over-predicts the impact by 18%, and the second strategy (using IS to find and evaluate) over-predicts the impact by 47%.

This strategy **also tends to find the most impactful shifts, for moderate values of** λ. For $\lambda \leq 6$, the shifts found by the Taylor approach are more impactful than those found by the IS approach. Moreover, the drop in accuracy remains substantial (e.g., a drop of around 17% at $\lambda = 6$). For $\lambda > 6$, the story is more subtle: The third approach (using IS to find and evaluate shifts) finds more impactful shifts, but (as noted previously) dramatically over-estimates their impact.

F Relationship to other approaches

In this section, we give a more detailed discussion of how our work relates to other approaches for evaluation of distributional robustness and learning of robust models. Much of the content from Section 1.1 is duplicated here, but expanded upon to include other relevant work and detailed discussion.

Distributionally Robust Optimization/Evaluation with divergence measures: Distributionally robust optimization (DRO) seeks to learn models that minimize objectives of the form of Equation (1) [Duchi and Namkoong, 2021, Duchi et al., 2020, Sagawa et al., 2020]. We focus on proactive worst-case evaluation of a fixed model, not optimization, similar to Subbaswamy et al. [2021], Li et al. [2021], but major differences between our work and prior work lie in the **definition of the set of plausible future distributions** $\mathcal{P}$, often called an "uncertainty set" in the optimization literature, where the goal is to specify a set that captures expected shifts, without being overly conservative.

Shifts in $\mathbb{P}(X, Y)$: A conservative approach is to include all joint distributions $\mathbb{P}(X, Y)$ within a certain neighborhood of the training distribution. Many coherent risk measures can be written as a worst-case loss of this form. For instance, the Entropic Value-at-Risk (EVaR), with confidence level $1 - \alpha$, corresponds to the worst-case loss over a set of distributions $\mathcal{P} = \{P \ll P_0 : D_{KL}(P\|P_0) \leq$

$-\ln \alpha\}$, where P_0 is the original distribution [Ahmadi-Javid, 2012]. Similarly, the Conditional Value-at-Risk (CVaR) with parameter α can be seen as the worst-case loss over an uncertainty set obtained from a limiting f-divergence (see Example 3 of Duchi and Namkoong [2021]), including all α-fractions of the original distribution. These measures are appealing, in that they are straightforward to compute, but can be very conservative.

Indeed, such measures often reduce to only considering the distribution of the loss itself. CVaR, for instance is equivalent to sorting the training examples by their loss, and taking the average loss of the top α-fraction. To illustrate these limitations, it is straightforward to see that, using the 0-1 loss and a classifier with 80% accuracy, the worst-case loss under both of these measures is 1.0 for any $\alpha \leq 0.2$. This is intuitive for CVaR (since over 20% of samples are misclassified in the original distribution), and follows for EVaR from the fact that the binary distribution with probability $q = 1$ has a KL-divergence to the original distribution $p = 0.2$ of $-\ln 0.2$.

Lam [2016] consider a more general problem of estimating the worst-case performance of stochastic systems over infinitesimal changes in distribution, measured by Kullback-Leibler divergence. Their approach is applicable beyond machine-learning settings, and generalizes to e.g., worst-case waiting times in a queueing system. They demonstrate that for a sufficiently small neighborhood of distributions, this worst-case performance can be well-approximated by a Taylor expansion whose coefficients can be estimated from the original distribution.

Shifts in $\mathbb{P}(X)$ alone: Partially due to this overly-conservative behavior, there has been a line of work incorporating additional restrictions on the allowable shift (i.e., adding more assumptions). For instance, Duchi et al. [2020] considers learning predictive models that optimize a worst-case loss similar to CVaR (a "worst-case subpopulation shift"), but where only $\mathbb{P}(X)$ is allowed to change, and $\mathbb{P}(Y \mid X)$ is assumed to be constant. For similar shifts, Li et al. [2021] considers only the task of evaluation, but provides a novel estimation procedure with dimension-free finite-sample guarantees. However, many real-world shifts do not fit this framework: In Example 1, both $\mathbb{P}(X)$ and $\mathbb{P}(Y \mid X)$ are changing, where $X = (A, O, L)$, as a result of a shift in $\mathbb{P}(O \mid Y, A)$.

Shifts in a conditional distribution: Closer to our work is Subbaswamy et al. [2021] who consider evaluating the loss under worst-case changes in a conditional distribution, but while we consider parametric shifts, they estimates the loss under worst-case $(1 - \alpha)$ conditional subpopulation shifts. However, it is not obvious how to choose an appropriate level of α: in some settings, seemingly plausible values of α (e.g., a 20% subpopulation) correspond to entirely implausible shifts. We give a simple lab-testing example in Appendix D, where the worst-case subpopulation is one where healthy patients are always tested, and sick patients never tested.

In contrast to these methods, our approach uses explicit parametric perturbations to define shifts, as opposed to distributional distances or subpopulations. In addition, our approach allows for shifts in multiple marginal or conditional distributions simultaneously: In Example 1, for instance, we can model a simultaneous change in both the marginal distribution of age, as well as the conditional distribution of lab testing, while other conditionals are unchanged. Our main requirement is that each shifting distribution is exponential family, and that the shift can be represented via the natural parameters: For continuous variables this is a non-trivial restriction, but for discrete variables it is true by definition.

Causality-motivated methods for learning robust models: Several approaches seek to learn models that perform well under arbitrarily large causal interventions (which result in arbitrary changes in selected conditional distributions). Several approaches proactively specify shifting mechanisms/conditional distributions, and then seek to learn predictors that have good performance under arbitrarily large changes in these mechanisms [Subbaswamy et al., 2019, Veitch et al., 2021, Makar et al., 2022, Puli et al., 2022]. Less relevant to our setting are approaches that use auxiliary information, such as environments [Magliacane et al., 2018, Rojas-Carulla et al., 2018, Arjovsky et al., 2019] or identity indicators [Heinze-Deml and Meinshausen, 2021] to learn models that rely on invariant conditional distributions. The worst-case optimality of these approaches is often restricted to cases where the shifts are arbitrarily large: In Example 1, this would correspond to minimizing worst-case loss over all possible lab testing policies.

However, when the causal interventions (i.e., changes in causal mechanisms) are bounded (i.e., not arbitrary), then these approaches are not necessarily optimal. Closest to our work in motivation is prior work on robustness to bounded shift interventions in linear causal models [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.

Our work can serve as an aid to deploying these causality-motivated methods in a few ways, by comparing their worst-case performance under bounded shifts: First, our work can inform whether such methods should be deployed at all, as for sufficiently small shifts, it may be the case that standard training yields better performance. Second, our work can inform hyperparameter selection for several of these approaches, which include regularization terms that implicitly trade off between robustness and in-distribution performance. More broadly, our approach is useful for probing (and comparing) the reliability of specific learned models under shift, regardless of the algorithm that produced them.

Evaluating out-of-distribution performance with unlabelled samples: A recent line of work has focused on predicting model performance in out-of-distribution settings, where unlabelled data is available from the target distribution [Garg et al., 2022, Jiang et al., 2022, Chen et al., 2021]. In contrast, our method operates using only samples from the original source distribution, and seeks to estimate the worst-case loss over a set of possible target distributions.

G Proofs

G.1 Proof of Proposition 1

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

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

Proof. By Definition 4 and Assumption 1, we have that

$$\mathbb{P}_\delta(\mathbf{V}) = \prod_{i=1}^m \mathbb{P}_{\delta_i}(W_i|Z_i) \prod_{V_j \in \mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_j|U_j)$$

$$\mathbb{P}(\mathbf{V}) = \prod_{i=1}^m \mathbb{P}(W_i|Z_i) \prod_{V_j \in \mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_j|U_j).$$

It follows that the supports of $\mathbb{P}_\delta$ and $\mathbb{P}$ are the same: Since the exponential family density is given by the base measure $g_i(W_i)$ times a exponential term (which is always strictly positive), and since the terms $\prod_{V_j \in \mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_j|U_j)$ are shared between $\mathbb{P}_\delta$ and $\mathbb{P}$, their supports agree.

To get the density ratio, we take the ratio of $\mathbb{P}_\delta(\mathbf{V})$ and $\mathbb{P}(\mathbf{V})$, and the terms $V_j \in \mathbf{V} \setminus \mathbf{W}$ cancel:

$$\begin{aligned} w_\delta(\mathbf{V}) &= \frac{\mathbb{P}_\delta(\mathbf{V})}{\mathbb{P}(\mathbf{V})} \\ &= \prod_{i=1}^m \frac{\mathbb{P}_{\delta_i}(W_i|Z_i)}{\mathbb{P}(W_i|Z_i)}. \end{aligned}$$

By Definition 4 and Assumption 1, each $\mathbb{P}_{\delta_i}(W_i|Z_i)$ is a δ_i-perturbation around the CEF distribution $\mathbb{P}(W_i|Z_i)$, so plugging in the exponential family densities, we get

$$\begin{aligned} w_\delta(\mathbf{V}) &= \prod_{i=1}^m \frac{g(W_i)\exp\left(\{\eta_i(Z_i) + s_i(Z_i;\delta_i)\}^\top T_i(W_i) - h_i(\eta_i(Z_i) + s_i(Z_i;\delta_i))\right)}{g(W_i)\exp\left(\eta_i(Z_i)^\top T_i(W_i) - h_i(\eta_i(Z_i))\right)} \\ &= \prod_{i=1}^m \exp\left(s_i(Z_i;\delta_i)T_i(W_i) - h_i(\eta_i(Z_i) + s_i(Z_i;\delta_i)) + h_i(\eta_i(Z_i))\right) \\ &= \exp\left(\sum_{i=1}^m s_i(Z_i;\delta_i)T_i(W_i)\right) \exp\left(\sum_{i=1}^m h_i(\eta_i(Z_i)) - h_i(\eta_i(Z_i) + s_i(Z_i;\delta_i))\right). \end{aligned}$$

□

G.2 Proof of Theorem 1

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 $\epsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}[T(W_i)|Z_i]$.

Proof. For simplicity throughout, we use $h_i^{(1)}$ to denote the gradient of the log-partition function $\nabla h_i(\cdot)$ with respect to the arguments, which is a column vector of length d_{T_i}, and we use $h_i^{(2)}$ to denote the Hessian $\nabla^2 h_i(\cdot)$, which is a matrix of size $d_{T_i} \times d_{T_i}$. We also use $\eta_{\delta_i}(z_i)$ as short-hand for $\eta_i(z_i) + s_i(z_i; \delta_i)$.

Shift Gradient: By Definition 4, the probability density / mass function $\mathbb{P}_\delta$ factorizes as follows, where $\delta = (\delta_1, \ldots, \delta_m)$

$$\mathbb{P}_\delta(\mathbf{V}) = \left(\prod_{W_i \in \mathbf{W}} \mathbb{P}_{\delta_i}(W_i|Z_i)\right)\left(\prod_{V_i \in \mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_i|\operatorname{PA}(V_i))\right), \tag{17}$$

and the gradient with respect to shift parameters δ_i is given by

$$\nabla_{\delta_i} p_\delta(v) = p_\delta(v) \nabla_{\delta_i} \log p_\delta(v) = p_\delta(v) \nabla_{\delta_i} \log p_{\delta_i}(w_i|z_i)$$

where the last equality follows from additivity of the log-likelihood in the conditionals, the factorization above, and the fact that δ_i only enters into the given conditional distribution. Given the assumed form of $\log p_{\delta_i}(w_i|z_i)$ given in Definition 3, we can observe that

$$\begin{aligned} \nabla_{\delta_i} \log p_{\delta_i}(w_i|z_i) &= \nabla_{\delta_i}\left[(\eta_i(z_i) + s_i(z_i;\delta_i))^\top T_i(w_i) - h_i(\eta(z_i) + s_i(z_i;\delta_i))\right] \\ &= (\nabla_{\delta_i} s_i(z_i;\delta_i))^\top T_i(w_i) - (\nabla_{\delta_i} s_i(z_i;\delta_i))^\top \nabla h_i(\eta(z_i) + s_i(z_i;\delta_i)) \\ &= (\nabla_{\delta_i} s_i(z_i;\delta_i))^\top (T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))) \end{aligned} \tag{18}$$

where $\nabla_{\delta_i} s_i(z_i;\delta_i) \in \mathbb{R}^{d_{T_i} \times d_{\delta_i}}$, and $\nabla h_i(\eta(z_i) + s_i(z_i;\delta_i))$ is the gradient of the function $h_i : \mathbb{R}^{d_{T_i}} \to \mathbb{R}$, which is a column vector of length d_{T_i}. It follows from known properties of the log-partition function [Wainwright et al., 2008, Proposition 3.1], that $h_i^{(1)}(\eta_{\delta_i}(z_i)) = \mathbb{E}_\delta[T_i(W_i)|z_i]$. This gives us that

$$\begin{aligned} \nabla_{\delta_i}\mathbb{E}_\delta[\ell] &= \mathbb{E}_\delta\left[\ell \cdot (\nabla_{\delta_i} s_i(Z_i;\delta_i))^\top (T_i(W_i) - \mathbb{E}_\delta[T_i(W_i)|Z_i])\right] \\ &= \mathbb{E}_\delta\left[(\nabla_{\delta_i} s_i(Z_i;\delta_i))^\top \mathbb{E}_\delta[\ell \cdot (T_i(W_i) - \mathbb{E}_\delta[T_i(W_i)|Z_i])|Z_i]\right] \\ &= \mathbb{E}_\delta\left[(\nabla_{\delta_i} s_i(Z_i;\delta_i))^\top \operatorname{cov}_\delta(\ell, T_i(W_i)|Z_i)\right], \end{aligned}$$

where the second equality follows from the tower property and Z_i-measurability of $\nabla_{\delta_i} s_i(Z_i;\delta_i)$, and the final equality follows from the definition of the conditional covariance. This expression, evaluated at $\delta = 0$, gives us the desired result, that

$$\mathrm{SG}^1_i := \nabla_{\delta_i}\mathbb{E}_\delta[\ell]\big|_{\delta=0} = \mathbb{E}\left[D_{i,1}^\top \operatorname{cov}(\ell, T_i(W_i)|Z_i)\right],$$

where $D_{i,1} = \nabla_{\delta_i} s_i(Z_i, \delta_i)|_{\delta=0}$. The result follows from the definition that gradients are taken entry-wise, giving $\mathrm{SG}^1 = (\mathrm{SG}^1_1, \ldots, \mathrm{SG}^1_m) \in \mathbb{R}^{d_{\delta_1} + \cdots d_{\delta_m}}$.

Shift Hessian (Diagonal): For the shift Hessian, we first compute the diagonal entries of $\nabla^2_\delta \mathbb{E}_\delta[\ell]|_{\delta=0}$, which are blocks of size $\mathbb{R}^{d_{\delta_i} \times d_{\delta_i}}$. We begin by computing the Hessian of the likelihood.

$$
\begin{aligned}
&\nabla^2_{\delta_i} p_\delta(v) \\
&= \nabla_{\delta_i}\Big(p_\delta(v)\nabla_{\delta_i}\log p_{\delta_i}(w_i|z_i)\Big) \\
&= p_\delta(v)\Big((\nabla_{\delta_i}\log p_{\delta_i}(w_i|z_i))^{\otimes 2} + \nabla^2_{\delta_i}\log p_{\delta_i}(w_i|z_i)\Big) \\
&= p_\delta(v)\Big(\{\nabla_{\delta_i} s_i(z_i;\delta_i)\}^\top \big(T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))\big)^{\otimes 2}\{\nabla_{\delta_i} s_i(z_i;\delta_i)\} \\
&\qquad - \{\nabla^2_{\delta_i} s_i(z_i;\delta_i)\}^\top (T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))) \\
&\qquad - \{\nabla_{\delta_i} s_i(z_i;\delta_i)\}^\top h_i^{(2)}(\eta_{\delta_i}(z_i))\{\nabla_{\delta_i} s_i(z_i;\delta_i)\}\Big), \\
&= p_\delta(v)\Big(\{\nabla_{\delta_i} s_i(z_i;\delta_i)\}^\top \Big(\big(T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))\big)^{\otimes 2} - h_i^{(2)}(\eta_{\delta_i}(z_i))\Big)\{\nabla_{\delta_i} s_i(z_i;\delta_i)\} \\
&\qquad - \{\nabla^2_{\delta_i} s_i(z_i;\delta_i)\}^\top \big(T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))\big)\Big)
\end{aligned}
$$

where we use the notation $v^{\otimes 2} := vv^\top$, and we note that $\nabla^2_{\delta_i} s(z_i;\delta_i)$ is a tensor of size $d_{T_i} \times d_{\delta_i} \times d_{\delta_i}$, and $\{\nabla^2_{\delta_i} s_i(z_i;\delta_i)\}^\top h_i^{(1)}(\cdot)$ is a matrix of size $d_{\delta_i} \times d_{\delta_i}$, where the (m,n)'th entry is $\{\frac{\partial}{\partial \delta_{im}}\frac{\partial}{\partial \delta_{in}} s(z_i;\delta_i)\}^\top h^{(1)}(\cdot)$.

Now, using the fact that $h^{(1)}(\eta_{\delta_i}(z)) = \mathbb{E}_\delta[T_i(W_i)|z_i]$ and $h^{(2)}(\eta_{\delta_i}(z_i)) = \mathrm{var}_\delta[T_i(W_i)|z_i]$ [Wainwright et al., 2008, Proposition 3.1], and the definition $\epsilon_{T_i|Z_i} = T_i(W_i) - \mathbb{E}_\delta[T_i(W_i)|Z_i]$, we obtain

$$
\begin{aligned}
&\nabla^2_{\delta_i}\mathbb{E}_\delta[\ell] \\
&= \mathbb{E}_\delta\left[\ell \cdot \{\nabla_{\delta_i} s_i(Z_i;\delta_i)\}^\top \Big(\epsilon^{\otimes 2}_{T|Z_i} - \mathrm{var}_\delta(T_i(W_i)|Z_i)\Big)\{\nabla_{\delta_i} s_i(Z_i;\delta_i)\}\right] \\
&\quad - \mathbb{E}_\delta\left[\ell \cdot \{\nabla^2_{\delta_i} s_i(Z_i;\delta_i)\}^\top \epsilon_{T_i|Z_i}\right] \\
&= \mathbb{E}_\delta\left[\{\nabla_{\delta_i} s_i(Z_i;\delta_i)\}^\top \mathrm{cov}_\delta\left(\ell, \epsilon^{\otimes 2}_{T_i|Z_i}\middle| Z_i\right)\{\nabla_{\delta_i} s_i(Z_i;\delta_i)\}\right] \\
&\qquad - \mathbb{E}_\delta\left[\ell \cdot \{\nabla^2_{\delta_i} s_i(Z_i;\delta_i)\}^\top \epsilon_{T_i|Z_i}\right]
\end{aligned}
$$

which gives the desired result when we evaluate at $\delta = 0$.

Shift Hessian (Off-Diagonal) For $i \neq j$, we have that

$$
\begin{aligned}
&\nabla_{\delta_i}\nabla_{\delta_j} p_\delta(v) \\
&= \nabla_{\delta_i}(p_\delta(v)\nabla_{\delta_j}\log p_{\delta_j}(w_j|z_j)) \\
&= \nabla_{\delta_i}(p_\delta(v)\nabla_{\delta_j}\log p_{\delta_j}(w_j|z_j)) \\
&= p_\delta(v)\nabla_{\delta_i}\log p_{\delta_i}(w_i|z_i)\big(\nabla_{\delta_j}\log p_{\delta_j}(w_j|z_j)\big)^\top \\
&= p_\delta(v)\Big(\{\nabla_{\delta_i} s_i(z_i;\delta_i)\}^\top (T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i)))\Big) \\
&\qquad\qquad \Big(\{\nabla_{\delta_j} s_j(z_j;\delta_j)\}^\top (T_j(w_j) - h_j^{(1)}(\eta_{\delta_j}(z_j)))\Big)^\top
\end{aligned}
$$

where the third line follows from the fact that $\nabla_{\delta_i}(\nabla_{\delta_j}\log p_{\delta_j}(w_j|z_j)) = 0$, and the last line follows from the derivation of the gradient of the log-likelihood in Equation (18). We can again use the fact that $h_i^{(1)}(\eta_{\delta_i}(Z_i)) = \mathbb{E}_\delta[T_i(W_i)|Z_i]$ and the shorthand $\epsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}_\delta[T_i(W_i)|Z_i]$ to write

that

$$\nabla_{\delta_i}\nabla_{\delta_j}\mathbb{E}_\delta[\ell]$$
$$= \mathbb{E}_\delta\Bigg[\ell\cdot\{\nabla_{\delta_i}s_i(z_i;\delta_i)\}^\top\Big((T_i(w_i)-h_i^{(1)}(\eta_{\delta_i}(z_i)))\Big)$$
$$\Big((T_j(w_j)-h_j^{(1)}(\eta_{\delta_j}(z_j)))\Big)^\top\{\nabla_{\delta_j}s_j(z_j;\delta_j)\}\Bigg]$$

and when we evaluate this expression at $\delta=0$, we obtain

$$\nabla_{\delta_i}\nabla_{\delta_j}\mathbb{E}_\delta[\ell]\big|_{\delta=0} = \mathbb{E}\left[\ell\cdot D_{i,1}^\top\epsilon_{T_i|Z_i}(\epsilon_{T_j|Z_j})^\top D_{j,1}\right] = \mathrm{cov}(\ell, D_{i,1}^\top\epsilon_{T_i|Z_i}\epsilon_{T_j|Z_j}^\top D_{j,1}).$$

Where the last equality follows because $\mathbb{E}[D_{i,1}^\top\epsilon_{T_i|Z_i}\epsilon_{T_j|Z_j}^\top D_{j,i}]=0$. To see this, note that one of W_i, W_j must be a non-descendant of the other, and we will assume without loss of generality that W_j is a non-descendant of W_i in the causal graph consistent with the factorization given in Equation (17), which implies that Z_j (the parents of W_j in the underlying graph) are also non-descendants of W_i.Thus, $W_i \perp\!\!\!\perp (W_j, Z_j)|Z_i$, because (W_j, Z_j) are both non-descendants of W_i. Then, observe that $D_{i,1}$ is a function of Z_i, and $\epsilon_{T_i|Z_i}$ is a variable with zero-mean conditioned on Z_i. Thus, $\mathbb{E}[D_{i,1}^\top\epsilon_{T_i|Z_i}|Z_i]=0$, for all Z_i. Moreover, given Z_i, we have that $D_{i,1}^\top\epsilon_{T_i|Z_i}$ is independent of $D_{j,1}^\top\epsilon_{T_j|Z_j}$. As a result, we can write that

$$\begin{aligned}\mathbb{E}[D_{i,1}^\top\epsilon_{T_i|Z_i}\epsilon_{T_j|Z_j}^\top D_{j,1}] &= \mathbb{E}[\mathbb{E}[D_{i,1}^\top\epsilon_{T_i|Z_i}\epsilon_{T_j|Z_j}^\top D_{j,1}|Z_i]]\\ &= \mathbb{E}[\mathbb{E}[D_{i,1}^\top\epsilon_{T_i|Z_i}|Z_i]\mathbb{E}[\epsilon_{T_j|Z_j}^\top D_{j,1}|Z_i]]\\ &= \mathbb{E}[0\cdot\mathbb{E}[\epsilon_{T_j|Z_j}^\top D_{j,1}|Z_i]]\\ &= 0\end{aligned}$$

□

G.3 Proof of 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*

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

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

Proof. We have $\nabla_\delta s(Z;\delta)=\nabla_\delta\delta=1$ and $\nabla_\delta^2 s(Z;\delta)=\nabla_\delta^2\delta=0$. The result now follows from Theorem 1. □

G.4 Proof of Theorem 2

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 (Equation (7)) and let $\sigma(M)$ denote the largest absolute value of the eigenvalues of a matrix M. Then*

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

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

Proof. The expectation is continuous and twice-differentiable with respect to δ, because of the smoothness of the exponential family in the parameter, the fact that the shift function s is twice-differentiable, and because the support does not change. Thus, applying Taylors remainder theorem to the function $t\mapsto\mathbb{E}_{t\cdot\delta}[\ell]$, it follows that there exist a $t_0\in[0,1]$ such that

$$\mathbb{E}_{1\cdot\delta}[\ell]-\mathbb{E}_{0\cdot\delta}[\ell]-\left(\tfrac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=0} = \left(\tfrac{1}{2}\tfrac{\mathrm{d}^2}{\mathrm{d}^2t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=t_0}. \tag{19}$$

We have $\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=0} = \mathrm{SG}^1$ and by the same arguments (see the proof of Theorem 1), it follows that $\left(\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}^2 t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=t_0} = \delta^\top \operatorname{cov}_{t_0\cdot\delta}(\ell, \epsilon^{\otimes 2}_{t_0\cdot\delta,T|Z})\delta$. Plugging this in, and subtracting $\frac{1}{2}\delta^\top \mathrm{SG}^2\,\delta$ on both sides of Equation (19) yields

$$\begin{aligned}\left|\mathbb{E}_\delta[\ell] - E_{\delta,\text{Taylor}}\right| &= \tfrac{1}{2}\left|\delta^\top\left(\operatorname{cov}_{t_0\cdot\delta}(\ell, \epsilon^{\otimes 2}_{t_0\cdot\delta,T|Z}) - \operatorname{cov}(\ell, \epsilon^{\otimes 2}_{0,T|Z})\right)\delta\right| \\ &\leq \tfrac{1}{2}\sup_{t\in[0,1]}\left|\delta^\top\left(\operatorname{cov}_{t\cdot\delta}(\ell, \epsilon^{\otimes 2}_{t\cdot\delta,T|Z}) - \operatorname{cov}(\ell, \epsilon^{\otimes 2}_{0,T|Z})\right)\delta\right|.\end{aligned}$$

Let $K := \left(\operatorname{cov}_{t\cdot\delta}(\ell, \epsilon^{\otimes 2}_{t\cdot\delta,T|Z}) - \operatorname{cov}(\ell, \epsilon^{\otimes 2}_{0,T|Z})\right)$. Since K is symmetric and real valued, it is diagonalizeable, $K = U^\top \Lambda U$ for an orthonormal matrix U and diagonal matrix $\Lambda = \operatorname{diag}(\alpha_1, \ldots, \alpha_d)$. We then have

$$\begin{aligned}|\delta^\top K\delta| &= |\delta^\top U^\top \Lambda U\delta| \\ &= |(\Lambda^{1/2}U\delta)^\top(\Lambda^{1/2}U\delta)| \\ &= \|\Lambda^{1/2}U\delta\|_2^2 \\ &\leq \|\Lambda^{1/2}\|_2^2\|U\delta\|_2^2 \\ &= \sigma(K)\|\delta\|_2^2,\end{aligned}$$

where $\Lambda^{1/2} = \operatorname{diag}(\sqrt{\alpha_1}, \ldots, \sqrt{\alpha_d})$, $\|\cdot\|_2$ denotes the supremum-norm when applied to matrices and the 2-norm when applied to vectors and $\|U\delta\|_2 = \|\delta\|_2$ because $\|U\delta\|_2^2 = \delta^\top U^\top U\delta = \delta^\top\delta = \|\delta\|_2^2$, using orthonormality of U. Plugging in this inequality, we get that

$$\left|\mathbb{E}_\delta[\ell] - E_{\delta,\text{Taylor}}\right| \leq \tfrac{1}{2}\sup_{t\in[0,1]}\sigma\left(\operatorname{cov}_{t\cdot\delta}(\ell, \epsilon^{\otimes 2}_{t\cdot\delta,T|Z}) - \operatorname{cov}(\ell, \epsilon^{\otimes 2}_{0,T|Z})\right)\|\delta\|_2^2,$$

which concludes the proof. □

G.5 Proof of Proposition B.1

Proposition B.1. *Consider a binary random variable W with conditional distribution*

$$\mathbb{P}_\delta(W = 1|Z) = \sigma(\eta(Z) + \delta)$$

for an arbitrary measurable function $\eta(Z)$ whose range is the extended real numbers $\eta(Z) \in \mathbb{R}\cup\{+\infty, -\infty\}$. Let $p_+ := \mathbb{P}(\eta(Z) = +\infty)$, $p_- := \mathbb{P}(\eta(Z) = -\infty)$, and assume that $p_+ + p_- < 1$. Then, the marginal probability

$$p_\delta = \mathbb{P}_\delta(W = 1)$$

is a strictly monotonically increasing function of $\delta \in \mathbb{R}$ whose range is $(p_+, 1 - p_-)$,

Proof. Let F denote the event that $\eta(Z)$ is finite (i.e., $\eta(Z) \notin \{-\infty, +\infty\}$). Under F, the conditional probability function $\sigma(\eta(Z) + \delta)$ is a strictly monotonically increasing function of δ, and if $\eta(Z) \in \{-\infty, +\infty\}$, then the conditional probability is a constant function of δ (zero or one, respectively). Hence, we can write that

$$\mathbb{P}_\delta(W = 1) = \mathbb{P}_\delta(W = 1|F)(1 - p_+ - p_-) + p_+$$

and by assumption, $1 - p_+ - p_- > 0$. The marginal probability $\mathbb{P}_\delta(W = 1|F)$ is a strictly monotonically increasing function of δ, with a limit of 1 as $\delta \to \infty$, and a limit of 0 as $\delta \to -\infty$. As a result, it is bounded in $(p_+, 1 - p_-)$. □

G.6 Proof of Lemma C.1

Lemma C.1. *Suppose $A \sim \mathcal{N}(\mu, \Sigma)$ and that (X, Y, H) are generated according to Equation (11). For $\gamma \in \mathbb{R}^{d_X}$ define $\ell := (Y - \gamma^\top X)^2$. Then there exist $v_\gamma, u_{\mu,\gamma} \in \mathbb{R}^{d_A}$ such that for all shifts $\delta \in \mathbb{R}^{d_A}$:*

$$\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell] + \delta^\top u_{\mu,\gamma} + \tfrac{1}{2}\delta^\top v_\gamma v_\gamma^\top\delta,$$

where $\mathbb{E}_\delta$ corresponds to taking the mean in the distribution where $A \sim \mathcal{N}(\mu+\delta, \Sigma)$. Further $u_{\mu,\gamma} = 0$ if $\mu = 0$.

Proof. It follows from Equation (11) that one can write $(X^\top, Y^\top, H^\top) = (1-B)^{-1}(MA+\epsilon)$, and for a given γ, there exist b_γ, κ_γ such that $Y - \gamma^\top X = b_\gamma^\top A + \kappa_\gamma^\top \epsilon$ [Rothenhäusler et al., 2021]. In $\mathbb{P}_\delta$, we can write $A = \mu + \delta + \epsilon_A$, where $\epsilon_A \sim \mathcal{N}(0, \Sigma)$, for all values of μ and δ. Plugging this in yields

$$\begin{aligned}
\mathbb{E}_\delta[(Y-\gamma^\top X)^2] &= \mathbb{E}_\delta[(b_\gamma^\top A + \kappa_\gamma^\top \epsilon)^2] \\
&= \mathbb{E}_\delta[(b_\gamma^\top(\mu+\delta+\epsilon_A) + \kappa_\gamma^\top \epsilon)^2] \\
&= \mathbb{E}[(b_\gamma^\top(\mu+\epsilon_A) + \kappa^\top \epsilon)^2] + (2b_\gamma^\top \mu)\delta^\top b_\gamma + \delta^\top b_\gamma b_\gamma^\top \delta \\
&= \mathbb{E}[(Y-\gamma^\top X)^2] + (2b_\gamma^\top \mu)\delta^\top b_\gamma + \delta^\top b_\gamma b_\gamma^\top \delta.
\end{aligned}$$

where we do not put a subscript on the expectation in the third line because it is taking expectations over ϵ_A and ϵ, both which do not depend on the choice of μ and δ. The statement of the lemma follows by letting $u_{\mu,\gamma} = 2b_\gamma^\top \mu$ and $v_\gamma = \sqrt{2} b_\gamma$. □

G.7 Proof of Proposition C.1

Proposition C.1. *Suppose $A \sim \mathcal{N}(\mu, \Sigma)$ and that (X, Y, H) are generated according to Equation (11). Then the shift gradient and Hessian are given by*

$$\mathrm{SG}^1 = \mathrm{cov}(\ell, \Sigma^{-1}A) \qquad \textit{and} \qquad \mathrm{SG}^2 = \mathrm{cov}(\ell, \Sigma^{-1}(A-\mu)(A-\mu)^\top \Sigma^{-\top})$$

and the loss under a mean shift of δ in A is given by

$$\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell] + \delta^\top \mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \mathrm{SG}^2\, \delta,$$

where $\ell := (Y - \gamma^\top X)^2$ and $\mathbb{E}_\delta$ corresponds to taking the mean in the distribution where $A \sim \mathcal{N}(\mu+\delta, \Sigma)$.

Proof. Similar to Lemma C.1, we rewrite $Y - \gamma^\top X = b_\gamma^\top A + \kappa^\top \epsilon$, and by rewriting $A = \mu + \delta + \epsilon_A$, where $\epsilon_A \sim \mathcal{N}(0, \Sigma)$, we obtain

$$\begin{aligned}
\mathbb{E}_\delta[(Y-\gamma^\top X)^2] = \mathbb{E}(b_\gamma^\top(\mu+\epsilon_A) + \kappa^\top \epsilon)^2 &\qquad (20)\\
+ (2b_\gamma^\top \mu)\delta^\top b &\qquad (21)\\
+ \delta^\top b b^\top \delta. &\qquad (22)
\end{aligned}$$

We recognize that Equation (20) equals $\mathbb{E}(Y - \gamma^\top X)^2$. Similarly, we now show that Equations (21) and (22) match the shift gradients (multiplied appropriately with δ).

First, we assume that $\Sigma = \mathrm{Id}$. Since A is a Gaussian with (known) mean Id, the sufficient statistic is $T(A) = A$. Hence, according to Theorem 1, we can compute the shift gradient as

$$\mathrm{SG}^1 = \mathrm{cov}(A, \ell) = \mathrm{cov}(A, (Y-\gamma^\top X)^2) = \mathrm{cov}(A, (b_\gamma^\top A)^2).$$

We can calculate the i'th entrance of this vector as:

$$\begin{aligned}
\mathrm{SG}^1 = \mathrm{cov}(A_i, (b_\gamma^\top A)^2) &= \mathrm{cov}(A_i - \mu_i, (b_\gamma^\top A)^2)) \\
&= \mathrm{cov}(A_i - \mu_i, b_{\gamma,i}^2 A_i^2 + 2\sum_{j \neq i} b_i b_j A_i A_j) \\
&= b_{\gamma,i}^2 \,\mathrm{cov}(A_i - \mu_i, A_i^2) + 2b_{\gamma,i} \sum_{j \neq i} b_j \,\mathrm{cov}(A_i - \mu_i, A_i A_j),
\end{aligned}$$

where in the first equality we use that subtracting a constant doesn't change the covariance, and we use independence of A_i from $A_j A_{j'}$ when $i \notin \{j, j'\}$. Using the assumption that A_i has unit variance, we now get that

$$\begin{aligned}
\mathrm{cov}(A_i - \mu_i, A_i^2) &= \mathbb{E}[A_i^3 - \mu_i A_i^2] = (\mu_i^3 + 3\mu_i) - \mu_i(\mu_i^2 + 1) = 2\mu_i \\
\mathrm{cov}(A_i - \mu_i, A_i A_j) &= \mathbb{E}[A_i^2 - A_i \mu_i]\mathbb{E}[A_j] = (\mu_i^2 + 1 - \mu_i^2)\mu_j = \mu_j.
\end{aligned}$$

By plugging in, we obtain

$$\begin{aligned}\mathrm{SG}^1(\mu_i) &= 2b_{\gamma,i}^2\mu_i + 2b_{\gamma,i}\sum_{j\neq i} b_j\mu_j \\ &= 2b_{\gamma,i}b_\gamma^\top\mu.\end{aligned}$$

Since this was element-wise, we obtain that the full vector is $\mathrm{SG}^1 = 2b_\gamma b_\gamma^\top\mu$, which, when multiplied with δ yields Equation (21).

We compute SG^2 similarly. The diagonal entries are given by

$$\begin{aligned}\mathrm{SG}^2_{i,i} &= \mathrm{cov}((A_i-\mu_i)^2, (b_\gamma^\top A)^2) \\ &= \mathrm{cov}((A_i-\mu_i)^2, b_{\gamma,i}^2A_i^2 + b_{\gamma,i}\sum_{j\neq i} b_{\gamma,j}A_iA_j) \\ &= b_{\gamma,i}^2\,\mathrm{cov}((A_i-\mu_i)^2, A_i^2) + b_{\gamma,i}\sum_{j\neq i} b_{\gamma,j}\,\mathrm{cov}((A_i-\mu_i)^2, A_iA_j).\end{aligned}$$

Because $\Sigma = \mathrm{Id}$, the second through fourth moments of A_i are given by $\mathbb{E}[A_i^2] = \mu_i^2+1$, $\mathbb{E}[A_i^3] = \mu_i^3+3\mu_i$ and $\mathbb{E}[A_i^4] = \mu_i^4+6\mu_i^2+3$. Using this, we get

$$\begin{aligned}\mathrm{cov}((A_i-\mu_i)^2, A_i^2) &= \mathbb{E}[A_i^4 - 2\mu_iA_i^3 + \mu_i^2A_i^2] - \mathbb{E}[(A_i-\mu_i)^2]\mathbb{E}[A_i^2] \\ &= (\mu_i^4+6\mu_i^2+3) - 2\mu_i(\mu_i^3+3\mu_i) + \mu_i^2(\mu_i^2+1) - 1\cdot(\mu_i^2+1) \\ &= 2,\end{aligned}$$

and for $j\neq i$:

$$\begin{aligned}\mathrm{cov}((A_i-\mu_i)^2, A_iA_j) &= \mathrm{cov}((A_i-\mu_i)^2, (A_i-\mu_i)A_j) + \mathrm{cov}((A_i-\mu_i)^2, \mu_iA_j) \\ &= \mathrm{cov}((A_i-\mu_i)^2, (A_i-\mu_i)A_j) \\ &= \mathbb{E}[(A_i-\mu_i)^3]\mathbb{E}[A_j] - \mathbb{E}[(A_i-\mu_i)^2]\mathbb{E}[(A_i-\mu_i)]\mathbb{E}[A_j] \\ &= 0-0,\end{aligned}$$

using linearity of the covariance, that $A_i \perp\!\!\!\perp A_j$ and that the first and third moments are zero for a centered Gaussian $A_i-\mu_i$. Plugging this in, we get that the diagonal entries are given by

$$\mathrm{SG}^2_{i,i} = 2b_{\gamma,i}^2.$$

We can compute the off-diagonal entries similarly. For $i\neq j$, we have:

$$\begin{aligned}\mathrm{SG}^2_{i,j} = \mathrm{cov}\Bigg(&(A_i-\mu_i)(A_j-\mu_j), \\ &b_{\gamma,i}^2A_i^2 + b_{\gamma,j}^2A_j^2 + 2b_{\gamma,i}b_{\gamma,j}A_iA_j + 2\sum_{v\notin\{i,j\}} b_{\gamma,i}b_{\gamma,v}A_iA_v + b_{\gamma,j}b_{\gamma,v}A_jA_v\Bigg).\end{aligned} \tag{23}$$

Using the independence of A_i and A_j, we have

$$\begin{aligned}&\mathrm{cov}((A_i-\mu_i)(A_j-\mu_j), A_i^2) \\ &= \mathbb{E}[A_i^2(A_i-\mu_i)]\underbrace{\mathbb{E}[A_j-\mu_j]}_{=0} - \underbrace{\mathbb{E}[A_i-\mu_i]}_{=0}\mathbb{E}[A_j-\mu_j]\mathbb{E}[A_i^2] \\ &= 0,\end{aligned}$$

and similarly $\mathrm{cov}((A_i-\mu_i)(A_j-\mu_j), A_j^2) = 0$. Using the same reasoning, for $v\notin\{i,j\}$

$$\begin{aligned}&\mathrm{cov}((A_i-\mu_i)(A_j-\mu_j), A_iA_v) \\ &= \mathbb{E}[(A_i-\mu_i)A_i]\mathbb{E}[A_j-\mu_j]\mathbb{E}[A_v] - \mathbb{E}[(A_i-\mu_i)]\mathbb{E}[A_i]\mathbb{E}[A_j-\mu_j]\mathbb{E}[A_v] \\ &= 0,\end{aligned}$$

and the same for $\mathrm{cov}((A_i - \mu_i)(A_j - \mu_j), A_j A_v)$. Finally, we have

$$\begin{aligned}
&\mathrm{cov}((A_i - \mu_i)(A_j - \mu_j), A_i A_j) \\
&= \mathbb{E}[(A_i - \mu_i)A_i]\mathbb{E}[(A_j - \mu_j)A_j] - \mathbb{E}[(A_i - \mu_i)]\mathbb{E}[A_i]\mathbb{E}[(A_j - \mu_j)]\mathbb{E}[A_j] \\
&= \mathbb{E}[(A_i - \mu_i)A_i]\mathbb{E}[(A_j - \mu_j)A_j] \\
&= \mathbb{E}[A_i^2 - \mu_i A_i]\mathbb{E}[A_j^2 - \mu_j A_j] \\
&= [(\mu_i^2 + 1) - \mu_i^2][[(\mu_j^2 + 1) - \mu_j^2]] \\
&= 1.
\end{aligned}$$

Plugging into Equation (23), we get that

$$\mathrm{SG}^2_{i,j} = 2b_{\gamma,i}b_{\gamma,j},$$

and hence for both diagonal and off-diagonal entries, $\mathrm{SG}^2_{i,j} = 2b_{\gamma,i}b_{\gamma,j}$, implying that

$$\mathrm{SG}^2 = 2b_\gamma b_\gamma^\top.$$

In particular $\frac{1}{2}\delta^\top \mathrm{SG}^2\, \delta$ matches Equation (22).

Finally, we consider the case $\Sigma \neq \mathrm{Id}$. Let $\Sigma^{-1/2}$ be the 'square-root' of Σ^{-1}, such that $\Sigma^{-1/2}\Sigma^{-\top/2}$ (where the latter denotes $(\Sigma^{-1/2})^\top$.[10]

The sufficient statistics for the mean in a multivariate Gaussian distribution with known variance is given by $T(A) = \Sigma^{-1}A$. We then have

$$\begin{aligned}
\mathrm{SG}^1 &= \mathrm{cov}(\Sigma^{-1}A, (b_\gamma^\top A)^2) \\
&= \Sigma^{-1/2}\,\mathrm{cov}(\Sigma^{-1/2}A, ((\Sigma^{1/2}b_\gamma)^\top \Sigma^{-1/2}A)^2) \\
&= \Sigma^{-1/2}\,\mathrm{cov}_{\tilde{\mu}}(\tilde{A}, (\tilde{b}_\gamma^\top \tilde{A})^2),
\end{aligned}$$

where $\tilde{A} = \Sigma^{-1/2}A =\sim \mathcal{N}(\tilde{\mu}, \mathrm{Id})$, $\tilde{\mu} = \Sigma^{-1/2}\mu$ and $\tilde{b}_\gamma = \Sigma^{1/2}b_\gamma$. In particular, since $\tilde{A}$ has unit variance, we can use the above derivations to obtain

$$\mathrm{SG}^1 = 2\Sigma^{-1/2}(\tilde{b}_\gamma \tilde{b}_\gamma^\top \tilde{\mu}) = 2b_\gamma b_\gamma^\top \mu.$$

In particular, the first shift gradient is the when $\Sigma \neq \mathrm{Id}$ as when $\Sigma = \mathrm{Id}$. Similarly,

$$\begin{aligned}
\mathrm{SG}^2 &= \mathrm{cov}(\Sigma^{-1}(A - \mu)(A - \mu)^\top \Sigma^{-\top}, (b_\gamma^\top A)^2) \\
&= \mathrm{cov}(\Sigma^{-1/2}\Sigma^{-1/2}(A - \mu)(A - \mu)^\top \Sigma^{-\top/2}\Sigma^{-\top/2}, (\Sigma^{1/2}b_\gamma)^\top \Sigma^{-1/2}A)^2) \\
&= \Sigma^{-1/2}\,\mathrm{cov}_{\tilde{\mu}}((\tilde{A} - \tilde{\mu})(\tilde{A} - \tilde{\mu})^\top, (\tilde{b}_\gamma^\top \tilde{A})^2)\Sigma^{-\top/2} \\
&= \Sigma^{-1/2} 2\tilde{b}_\gamma \tilde{b}_\gamma^\top \Sigma^{-\top/2} \\
&= 2b_\gamma b_\gamma^\top.
\end{aligned}$$

Hence, also when $\Sigma \neq \mathrm{Id}$, the terms of Equations (21) and (22) matches the expression given by SG^1 and SG^2. This concludes the proof. □

[10] Formally, if $\Sigma^{-1} = U\Lambda U^\top$ where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{d_A})$, define $\Sigma^{-1/2} := U\,\mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_{d_A}})$.