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.

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

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,

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

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).

16

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 $\eta_\delta(Z) \coloneqq \eta(Z) + \delta$, 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.

17

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

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

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

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

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

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

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

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

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

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

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

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

18
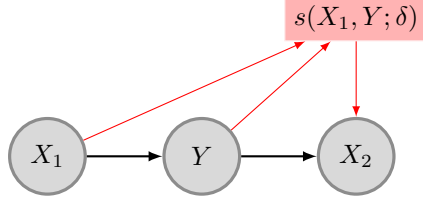
Figure 6: Illustrative example of an intervention $s(X_1, Y; \delta)$, and modified causal graph, which creates a dependence between X_1 and X_2 that bypasses Y.

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

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

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

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

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 $\mathbf{V}' = \mathbf{V}$.

19

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 $\mathbb{P}$.

20

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 $\mathrm{SG}^1 = (\mathrm{SG}_1^1, \ldots, \mathrm{SG}_m^1) \in \mathbb{R}^{d_\delta}$ where*

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

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 $D_{i,k} := \nabla_{\delta_i}^k s_i(Z_i; \delta_i)|_{\delta=0}$, 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 $\epsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}[T(W_i)|Z_i]$.

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 $\mathbb{R}^{d_T}$.

21

- 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 $\mathbb{E}[\text{cov}(A, B|C)] = \mathbb{E}[\epsilon_{A|C}\epsilon_{B|C}]$ 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 $i \in [m] := \{1, \dots, m\}$,

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

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

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

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

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

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, $T(W) = \big(W, \text{vec}(WW^\top)\big)$ and $\eta(Z) = \big(\sigma(Z)^{-1}\mu(Z), -\frac{1}{2}\text{vec}(\mu(Z))\big)$, 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/.

22
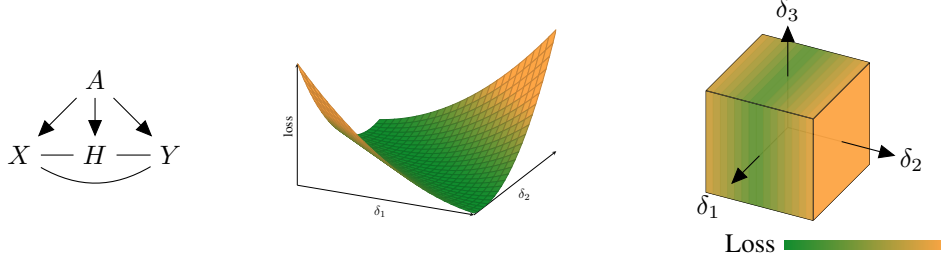
Figure 7: (Left) Graphical model assumed by Equation (11). The undirected edges represent either any directed configuration of directed edges or the dependence structures arising due to an acyclic SCM [Bongers et al., 2021]. (Middle) Plotting $\mathbb{E}_\delta[(Y - \gamma^\top X)^2]$ as a function of $\delta \in \mathbb{R}^2$ 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

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

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

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

$$= \mathbb{E}\left[\operatorname{cov}(\ell, \left(W - \mathbb{E}[W|Z] \right)^{\otimes 2}|Z) \right].$$

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 $E_{\delta,\mathrm{Taylor}}$ 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

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

23

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*

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

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.

24

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

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

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

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

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

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

and

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

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)).

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

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 $\mathbb{P}_\delta(X) \sim$

25
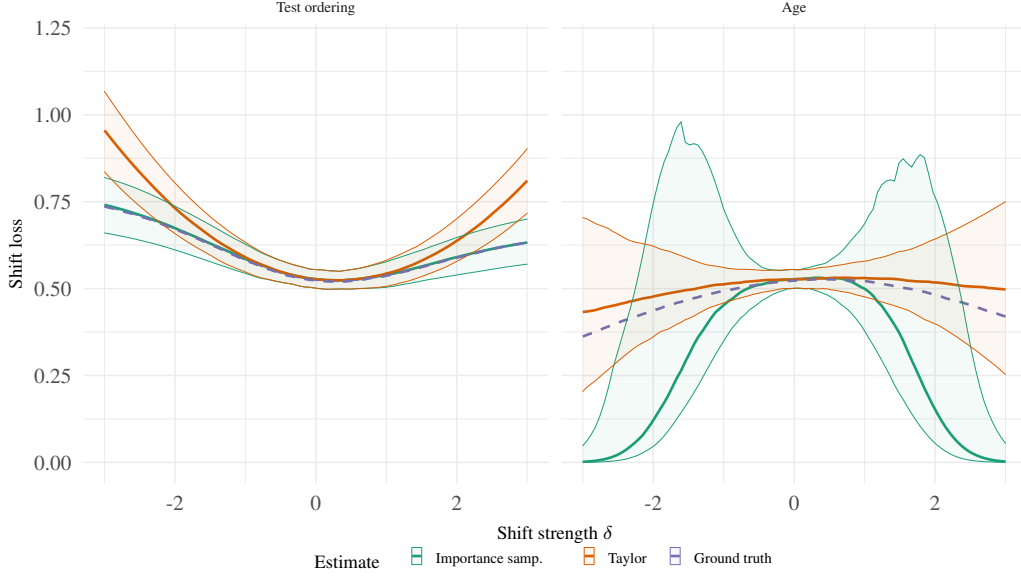
Figure 8: We plot the mean and confidence intervals of $\hat{E}_{\delta,\text{Taylor}}$ and $\hat{E}_{\delta,\text{IS}}$ when the shifted loss as in the lab test ordering example Example 1. (Left) We consider a shift in the logits of ordering lab tests from $\eta(Z)$ to $\eta(Z) + \delta_0$. (Right) We consider a shift in the mean of Age. In the observed distribution $\eta = \mu/\sigma = 0$ and we shift to a mean of $\eta = \delta$.

$\mathcal{N}(\delta, 1)$ 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

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

and the variances of the estimators are

$$\text{var}(\hat{\mu}_{\text{IS}}) = \frac{\mathbb{E}[\{\ell(X + 2\delta)\}^2]}{n}\exp(\delta^2)$$

$$\text{var}(\hat{\mu}_{\text{Taylor}}) = \frac{\text{var}\left(\ell(X) + \delta X\ell(X) + \frac{1}{2}\delta^2 X^2\ell(X)\right)}{n}.$$

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 $\mathbb{E}_\delta[\ell]$.

26
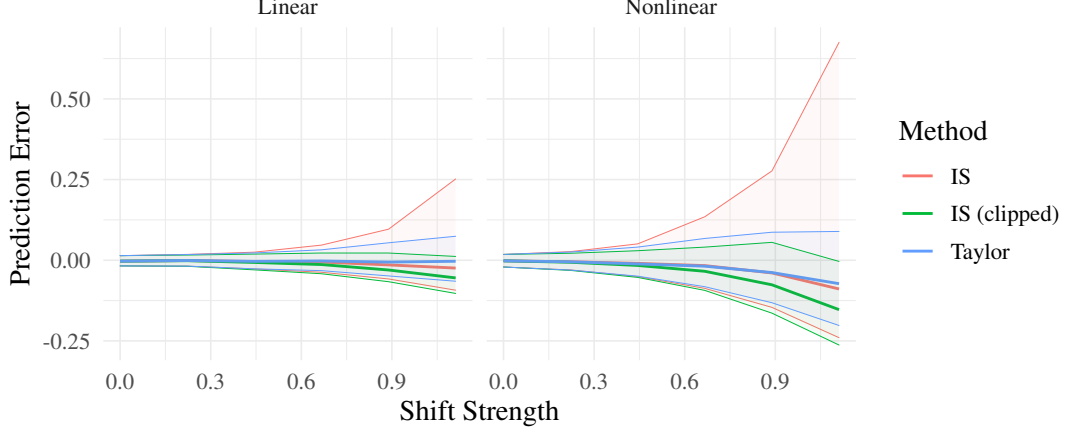
Figure 9: Median and quantiles of the error in predicting $\mathbb{E}_\delta[\ell]$ under a shift δ.

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 $W \in \mathbb{R}^3$,

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

where $\odot$ refers to entrywise multiplication, $\epsilon \sim \mathcal{N}(0, \text{Id}_4)$, 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 $\mathbb{P}(W) \sim \mathcal{N}(0, \text{Id}_3)$ to $\mathbb{P}_\delta(W) \sim \mathcal{N}(\delta, \text{Id}_3)$, where $\delta = [s, s, s]^\top$ for some shift strength $s > 0$. We then compute the shift gradient $\text{SG}^1 = \text{cov}(\ell, W)$ and Hessian $\text{SG}^2 = \text{cov}(\ell, WW^\top)$, 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, both 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

27

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 $Y := f_0(X) + \epsilon$ 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

$$\left| \mathbb{E}_\delta[\ell] - E_{\delta,\text{Taylor}} \right|$$

$$\leq \tfrac{1}{2} \sup_{t \in [0,1]} \Bigg| \text{cov}_{t \cdot \delta} \left((f_0(X) - f(X) + \epsilon)^2, (X - t \cdot \delta)^2 \right)$$

$$- \text{cov} \left((f_0(X) - f(X) + \epsilon)^2, (f_0(X) - f(X) + \epsilon)^2, X^2) \right) \Bigg| \cdot \delta^2.$$

The subscript $\text{cov}_{t \cdot \delta}$ 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

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

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

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

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

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

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].

28

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:

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

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.

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

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.

29

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 $0 \le h_{11} \le 1$, and those which are imposed by the equality constraint with h_{01} and the fact that $0 \le h_{01} \le 1$. For the latter, with some algebra we can write that

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

So that the constraints on h_{11} become

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

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

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

which yields

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

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

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

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.

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

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

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

30
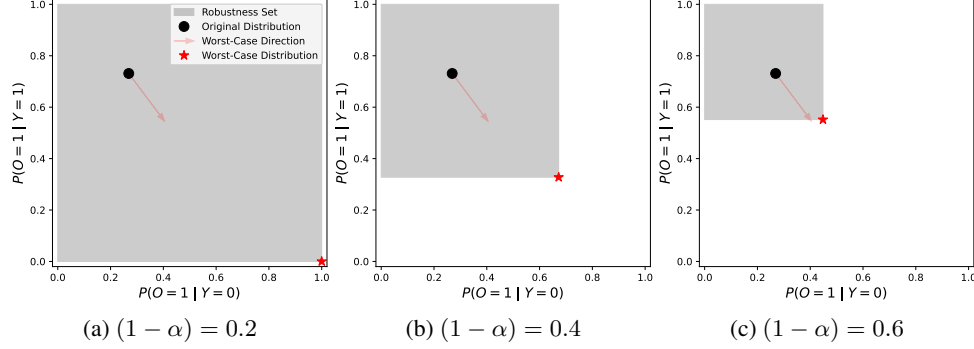
Figure 10: Feasible sets, worst-case directions, and worst-case solutions for a $(1 - \alpha)$ subpopulation shift in the conditional distribution $\mathbb{P}(O|Y)$ for differing values of α. Worst-case directions are computed using Equation (15), as unit-norm vectors re-scaled to fit in the plot, and 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

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

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

31
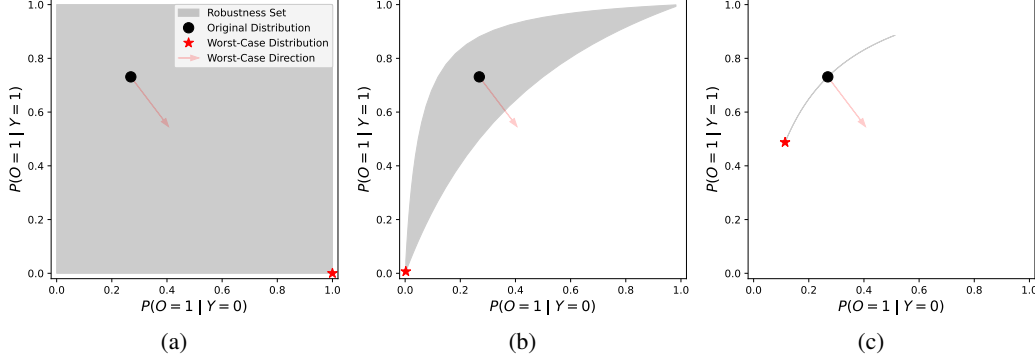
Figure 11: Each figure shows the set of conditional probability distributions ("CPDs") $\mathbb{P}(O|Y)$ that can be represented by a shift of $(\delta_0, \delta_1) \in \Delta_0 \times \Delta_1$, along with the worst-case distribution (given by the red star) for the 0–1 loss. In this example, the expected 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.

32
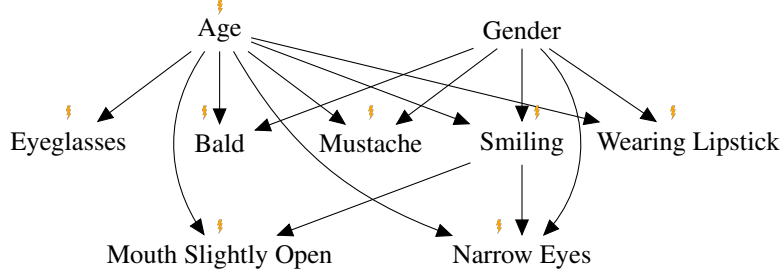
Figure 12: Causal graph over attributes, where lightning bolts indicate changes in mechanisms. Also displayed in Figure 4.

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

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

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.

33

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 $W_i = \mathrm{Bald}$, where $\eta(\mathrm{Young}, \mathrm{Male}) = -3.0 + 3.5 \cdot \mathrm{Male} - 1.0 \cdot \mathrm{Young}$, the shift would be

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

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 $\hat{E}_{\delta,\mathrm{IS}}$ and $\hat{E}_{\delta,\mathrm{Taylor}}$ 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.

34

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 $\lambda^2 = 4$.
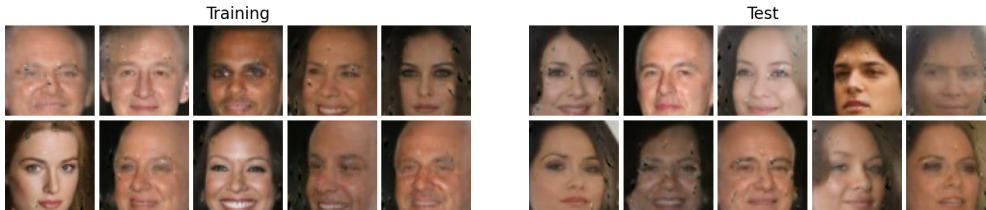
Figure 14: Examples of images from the training distribution $\mathbb{P}$ and the test distribution $\mathbb{P}_\delta$ that is characterized by the worst-case shift δ, see Figure 4.

35

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 $\mathcal{P} = \{P \ll P_0 : D_{KL}(P\|P_0) \leq$

36

$-\ln \alpha\}$, 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 $\alpha \le 0.2$. 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]. Other approaches 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, worst-case optimality under arbitrarily large shifts would correspond to minimizing the worst-case loss under 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.,

37

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. We discuss this point in more detail in Appendix F.1 below.

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.

F.1 The importance of considering restricted shifts in causal mechanisms

In Figure 15 we revisit Example 1, adopting the perspective of a model developer, who is aware that laboratory testing policies (i.e., $P(O \mid A, Y)$) may change. As this change may impact the correlation between laboratory testing features (O, L) and the label Y, how should the model developer proceed?

From a causal perspective, one way to approach model development is to learn a predictive model that is "causal" in the sense that it only relies on the causal parents of the label Y. In this example, A is the full set of causal parents of Y, and the conditional distribution $P(Y = 1 \mid A)$ does not change under changes in laboratory testing policy. This conditional distribution is an example of an "invariant" conditional distribution [Rojas-Carulla et al., 2018], reflecting the unchanging causal mechanisms that generate Y which are not affected by changes in laboratory testing policy. With this in mind, we consider the choice between two models:

- **Age-based model**: $f(A) \approx P(Y = 1 \mid A)$, predicting disease using age alone.[10]
- **Full model**: $f(A, O, L) \approx P(Y = 1 \mid A, O, L)$, predicting disease using all features.

We now demonstrate the utility of incorporating additional knowledge, considering not only "what" can change (i.e., $P(O \mid A, Y)$), but also considering "how" and "how much" it can change, and translating that knowledge into a quantitative comparison between these modelling choices. The question of "how" corresponds to our choice of shift function, and "how much" corresponds to our choice of constraints on shift parameters. We consider changes in testing that correspond to a uniform increase/decrease in testing rates, parameterized as

$$P_\delta(O = 1 \mid A, Y) = \text{sigmoid}(\eta(A, Y) + \delta) \tag{17}$$

Other details of the underlying distribution are given in Appendix A.

In Figure 15 (right), we plot the loss of each model under distributions[11] that correspond to different choices of δ, and observe that despite having invariant performance, the age-based model only out-performs the full model under substantial changes in testing policy. In this case, the model $f(A)$ (throwing away laboratory testing information) yields better performance if testing rates drop substantially, but for a large set of changes in testing rates, the full model $f(A, O, L)$ is superior.

Considering the worst-case performance of each model can guide model selection. If a substantial change in testing rates is not plausible (which can be expressed as constraints on δ), and the worst-case loss (over plausible changes) of $f(A, O, L)$ is lower than that of $f(A)$, the model developer may decide to use the full model $f(A, O, L)$ in any case.

---

[10]Details of how the full model $f(A, O, L)$ is trained are described in Appendix A. The model $f(A)$ is trained using unregularized logistic regression. Both models are trained on data drawn from the original distribution, where the marginal testing rate is 50%.

[11]In this case, every choice of δ maps to a unique marginal testing rate in the distribution P_δ (see Proposition B.1), so we plot the loss as a function of testing rate, instead of δ directly.

38
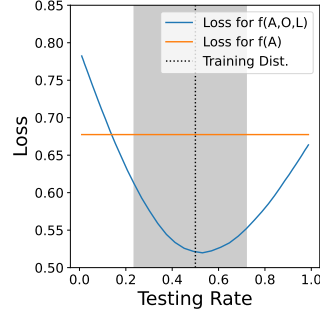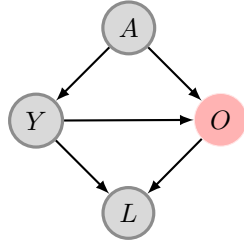
Figure 15: (Left) Causal graph for Example 1, where the variables are $Y \in \{0,1\}$ for the label (Disease), $A \in \mathbb{R}$ for Age, $O \in \{0,1\}$ for whether a laboratory test is ordered (Test Order), and $L \in \mathbb{R}$ for the lab result (Test Result), if available. (Right) Using the same generative model as in Appendix A, we contrast the performance of the full model $f(A, O, L)$ and a model $f(A)$ that only uses age, across distributions which differ in testing rates according to $P_\delta(O = 1 \mid A, Y) = \text{sigmoid}(\eta(A, Y) + \delta)$. Comparing performance on a range of distributions where we vary δ, we observe that $f(A)$ has invariant loss, but $f(A, O, L)$ has better performance for a wide range of shifts δ. In particular, if we compare the worst-case loss under shifts $|\delta| \leq 1.5$ (corresponding to marginal testing rates in the grey region), we can observe that the worst-case loss of $f(A, O, L)$ is lower than that of $f(A)$.

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,* $\text{supp}(\mathbb{P}) = \text{supp}(\mathbb{P}_\delta)$ *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:

$$w_\delta(\mathbf{V}) = \frac{\mathbb{P}_\delta(\mathbf{V})}{\mathbb{P}(\mathbf{V})}$$

$$= \prod_{i=1}^m \frac{\mathbb{P}_{\delta_i}(W_i | Z_i)}{\mathbb{P}(W_i | Z_i)}.$$

39

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

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

$$= \prod_{i=1}^{m} \exp\left(s_i(Z_i;\delta_i)T_i(W_i) - h_i(\eta_i(Z_i) + s_i(Z_i;\delta_i)) + h_i(\eta_i(Z_i))\right)$$

$$= \exp\left(\sum_{i=1}^{m} s_i(Z_i;\delta_i)T_i(W_i)\right)\exp\left(\sum_{i=1}^{m} h_i(\eta_i(Z_i)) - h_i(\eta_i(Z_i) + s_i(Z_i;\delta_i))\right).$$

$\square$

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 $\mathrm{SG}^1 = (\mathrm{SG}_1^1, \ldots, \mathrm{SG}_m^1) \in \mathbb{R}^{d_\delta}$ where*

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

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 $D_{i,k} := \nabla_{\delta_i}^k s_i(Z_i;\delta_i)|_{\delta=0}$, 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 $\epsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}[T(W_i)|Z_i]$.

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 $\delta = (\delta_1, \ldots, \delta_m)$

$$\mathbb{P}_\delta(\mathbf{V}) = \left(\prod_{W_i \in \mathbf{W}} \mathbb{P}_{\delta_i}(W_i|Z_i)\right)\left(\prod_{V_i \in \mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_i|\mathrm{PA}(V_i))\right), \tag{18}$$

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

$$\nabla_{\delta_i}\log p_{\delta_i}(w_i|z_i) = \nabla_{\delta_i}\left[(\eta_i(z_i) + s_i(z_i;\delta_i))^\top T_i(w_i) - h_i(\eta(z_i) + s_i(z_i;\delta_i))\right]$$

$$= (\nabla_{\delta_i}s_i(z_i;\delta_i))^\top T_i(w_i) - (\nabla_{\delta_i}s_i(z_i;\delta_i))^\top \nabla h_i(\eta(z_i) + s_i(z_i;\delta_i))$$

$$= (\nabla_{\delta_i}s_i(z_i;\delta_i))^\top(T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))) \tag{19}$$

40

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

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

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

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

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 $\mathrm{SG}^1 = (\mathrm{SG}_1^1, \ldots, \mathrm{SG}_m^1) \in \mathbb{R}^{d_{\delta_1} + \cdots d_{\delta_m}}$.

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

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

where we use the notation $v^{\otimes 2} := vv^\top$, and we note that $\nabla_{\delta_i}^2 s(z_i; \delta_i)$ is a tensor of size $d_{T_i} \times d_{\delta_i} \times d_{\delta_i}$, and $\{\nabla_{\delta_i}^2 s_i(z_i; \delta_i)\}^\top h_i^{(1)}(\cdot)$ 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

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

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

41

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

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

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 (19). 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

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

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 (18), 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

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

$\square$

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*

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

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. $\square$

42

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

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

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

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

$$\mathbb{E}_{1 \cdot \delta}[\ell] - \mathbb{E}_{0 \cdot \delta}[\ell] - \left(\tfrac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}_{t \cdot \delta}[\ell] \right) \bigg|_{t=0} = \left(\tfrac{1}{2} \tfrac{\mathrm{d}^2}{\mathrm{d}^2 t} \mathbb{E}_{t \cdot \delta}[\ell] \right) \bigg|_{t=t_0}. \tag{20}$$

We have $\left(\tfrac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}_{t \cdot \delta}[\ell] \right) \bigg|_{t=0} = \mathrm{SG}^1$ and by the same arguments (see the proof of Theorem 1), it follows that $\left(\tfrac{1}{2} \tfrac{\mathrm{d}^2}{\mathrm{d}^2 t} \mathbb{E}_{t \cdot \delta}[\ell] \right) \bigg|_{t=t_0} = \delta^\top \mathrm{cov}_{t_0 \cdot \delta}(\ell, \epsilon_{t_0 \cdot \delta, T|Z}^{\otimes 2}) \delta$. Plugging this in, and subtracting $\tfrac{1}{2} \delta^\top \mathrm{SG}^2 \delta$ on both sides of Equation (20) yields

$$\left| \mathbb{E}_\delta[\ell] - E_{\delta, \text{Taylor}} \right| = \tfrac{1}{2} \left| \delta^\top \left(\mathrm{cov}_{t_0 \cdot \delta}(\ell, \epsilon_{t_0 \cdot \delta, T|Z}^{\otimes 2}) - \mathrm{cov}(\ell, \epsilon_{0,T|Z}^{\otimes 2}) \right) \delta \right|$$

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

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

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

where $\Lambda^{1/2} = \mathrm{diag}(\sqrt{\alpha_1}, \ldots, \sqrt{\alpha_d})$, $\|\cdot\|_2$ 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

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

which concludes the proof. $\qquad\square$

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 $(p_+, 1 - p_-)$,

43

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 $(p_+, 1 - p_-)$. $\qquad\square$

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 $v_\gamma = \sqrt{2}b_\gamma$. $\qquad\square$

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*

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

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

$$\mathbb{E}_\delta[(Y - \gamma^\top X)^2] = \mathbb{E}(b_\gamma^\top(\mu + \epsilon_A) + \kappa^\top \epsilon)^2 \qquad (21)$$

$$+ (2b_\gamma^\top \mu)\delta^\top b \qquad (22)$$

$$+ \delta^\top bb^\top \delta. \qquad (23)$$

We recognize that Equation (21) equals $\mathbb{E}(Y - \gamma^\top X)^2$. Similarly, we now show that Equations (22) and (23) match the shift gradients (multiplied appropriately with δ).

44

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:

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

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

$$\mathrm{cov}(A_i - \mu_i, A_i^2) = \mathbb{E}[A_i^3 - \mu_i A_i^2] = (\mu_i^3 + 3\mu_i) - \mu_i(\mu_i^2 + 1) = 2\mu_i$$

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

By plugging in, we obtain

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

Since this was element-wise, we obtain that the full vector is $\mathrm{SG}^1 = 2 b_\gamma b_\gamma^\top \mu$, which, when multiplied with δ yields Equation (22).

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

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

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

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

and for $j \neq i$:

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

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

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

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

$$\mathrm{SG}_{i,j}^2 = \mathrm{cov}\bigg((A_i - \mu_i)(A_j - \mu_j), \tag{24}$$

$$b_{\gamma,i}^2 A_i^2 + b_{\gamma,j}^2 A_j^2 + 2 b_{\gamma,i} b_{\gamma,j} A_i A_j + 2 \sum_{v \notin \{i,j\}} b_{\gamma,i} b_{\gamma,v} A_i A_v + b_{\gamma,j} b_{\gamma,v} A_j A_v \bigg).$$

45

Using the independence of A_i and A_j, we have

$$\text{cov}((A_i - \mu_i)(A_j - \mu_j), A_i^2)$$
$$= \mathbb{E}[A_i^2(A_i - \mu_i)] \underbrace{\mathbb{E}[A_j - \mu_j]}_{=0} - \underbrace{\mathbb{E}[A_i - \mu_i]}_{=0} \mathbb{E}[A_j - \mu_j]\mathbb{E}[A_i^2]$$
$$= 0,$$

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

$$\text{cov}((A_i - \mu_i)(A_j - \mu_j), A_i A_v)$$
$$= \mathbb{E}[(A_i - \mu_i)A_i]\mathbb{E}[A_j - \mu_j]\mathbb{E}[A_v] - \mathbb{E}[(A_i - \mu_i)]\mathbb{E}[A_i]\mathbb{E}[A_j - \mu_j]\mathbb{E}[A_v]$$
$$= 0,$$

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

$$\text{cov}((A_i - \mu_i)(A_j - \mu_j), A_i A_j)$$
$$= \mathbb{E}[(A_i - \mu_i)A_i]\mathbb{E}[(A_j - \mu_j)A_j] - \mathbb{E}[(A_i - \mu_i)]\mathbb{E}[A_i]\mathbb{E}[(A_j - \mu_j)]\mathbb{E}[A_j]$$
$$= \mathbb{E}[(A_i - \mu_i)A_i]\mathbb{E}[(A_j - \mu_j)A_j]$$
$$= \mathbb{E}[A_i^2 - \mu_i A_i]\mathbb{E}[A_j^2 - \mu_j A_j]$$
$$= [(\mu_i^2 + 1) - \mu_i^2][[(\mu_j^2 + 1) - \mu_j^2]]$$
$$= 1.$$

Plugging into Equation (24), we get that

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

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

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

In particular $\frac{1}{2}\delta^\top \text{SG}^2 \delta$ matches Equation (23).

Finally, we consider the case $\Sigma \neq \text{Id}$. Let $\Sigma^{-1/2}$ be the 'square-root' of Σ^{-1}, such that $\Sigma^{-1/2}\Sigma^{-\top/2}$ (where the latter denotes $(\Sigma^{-1/2})^\top$.[12]

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

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

where $\tilde{A} = \Sigma^{-1/2}A =\sim \mathcal{N}(\tilde{\mu}, \text{Id})$, $\tilde{\mu} = \Sigma^{-1/2}\mu$ 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

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

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

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

Hence, also when $\Sigma \neq \text{Id}$, the terms of Equations (22) and (23) matches the expression given by SG^1 and SG^2. This concludes the proof. $\qquad\square$

---

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

46