Evaluating Robustness to Dataset Shift via Parametric Robustness Sets

Anonymous Authors[1]

Abstract

We give a method for proactively identifying small, plausible shifts in distribution which lead to large differences in model performance. To ensure that these shifts are plausible, we parameterize them in terms of interpretable changes in causal mechanisms of observed variables. This defines a parametric robustness set of plausible distributions and a corresponding worst-case loss. We construct a local approximation to the loss under shift, and show that problem of finding worst-case shifts can be efficiently solved.

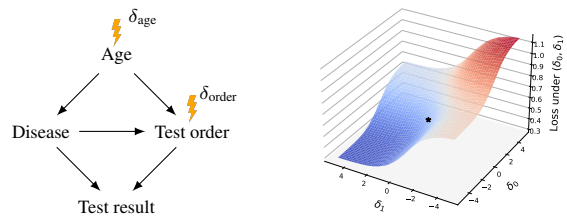

Figure 1. (Left) Causal graph for Example 1.1. Our approach allows for simultaneous shifts in age and test ordering, parameterized by $\delta_{\text{age}}, \delta_{\text{order}}$. (Right) We illustrate a shift in testing rates, using $s(Y; \delta_{\text{order}}) = \delta_1 \cdot Y + \delta_0(1 - Y)$, where $\delta_{\text{order}} = (\delta_0, \delta_1)$. We plot the landscape of the expected cross-entropy loss of a fixed model over distributions parameterized by (δ_0, δ_1), with the training distribution given as the black star.

1. Introduction

Predictive models may perform poorly outside of the training distribution, a problem broadly known as dataset shift (Quiñonero-Candela et al., 2008). In this paper, our goal is to proactively understand the sensitivity of a predictive model to dataset shift, using only data from the training distribution. Formally, for a model $f(X)$ trained on data from $\mathbb{P}(X, Y)$, with loss function $\ell(f(X), Y)$, we seek to understand the loss of the model under a set of *plausible* future distributions $\mathcal{P}$. We seek to evaluate the worst-case loss over $\mathcal{P}$,

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

and provide an interpretable description of a distribution P which maximizes this objective. To illustrate, we use a running example inspired by Subbaswamy et al. (2021).

Example 1.1 (Shifts in laboratory testing)**.** We seek to classify disease (Y) based on the age (A) of a patient, whether a lab test has been ordered (O), and test results (L) if ordered. The performance of a predictive model may be sensitive to changes in testing policies, as the *fact that a test has been ordered* itself is predictive of disease. Figure 1 (left) gives a plausible causal relationship between variables. Let $\mathbb{P}(O|A, Y) = \sigma(\eta(A, Y))$, where σ is the sigmoid function and $\eta(A, Y)$ is the log-odds. In Figure 1 (right), we show the loss under a set of new distributions parameterized by $\delta = (\delta_0, \delta_1)$, where we modify $\mathbb{P}_\delta(O|A, Y) = \sigma(\eta(A, Y) + s(Y; \delta))$ for a *shift function* $s(Y; \delta) = \delta_1 \cdot Y + \delta_0 \cdot (1 - Y)$, which modifies the log-odds of testing for both sick and healthy patients. If δ_0, δ_1 are unconstrained, the worst-case occurs when all healthy patients are tested, and no sick patients are tested.

We address two challenges: The first (Section 2) is to define a set of possible distributions $\mathcal{P}$ such that each distribution $P \in \mathcal{P}$ satisfies two desiderata: First, they should be *causally interpretable and simple to specify*, without placing unnecessary restrictions on the data-generating process. Second, they should be *realistic*, which often entails bounding the magnitude of the shift. We construct causally interpretable shifts by defining perturbed distributions $\mathbb{P}_\delta$ using changes in causal mechanisms, parameterized by a finite-dimensional parameter δ. Our main requirement is that the shifting mechanisms follow a conditional exponential family distribution. We also demonstrate that constraining δ can ensure that shifts are realistic: The unconstrained worst-case shift in Example 1.1 is implausible, where all healthy patients (and no sick patients) are tested. Equation (1) becomes

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

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

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

The second challenge (Section 3) is evaluation of the expected loss under shift, as well as finding the worst-case shift. Under our definition of shifts, we show that the test distribution can always be seen as a reweighting of the training distribution, allowing for reweighting approaches, such as importance sampling, to estimate the expected loss under shifts. While this is practical for some distribution shifts, for others, importance sampling can lead to extreme variance in estimation. Further, finding the worst-case shift using a reweighted objective involves maximization over a non-concave objective (see Figure 1), a problem that is generally NP-hard. We derive a second-order approximation to the expected loss under shift, and show how it can be estimated without the use of reweighting. For quadratic constraints Δ, we can approximate the general non-convex optimization problem in Equation 2 with a non-convex, quadratically constrained quadratic program (QCQP) for which efficient solvers exist (Conn et al., 2000, Section 7). We bound the approximation error of this surrogate objective, and show experimentally that it is well-behaved.

2. Defining parametric robustness sets

Notation: Let $\mathbf{V}$ denote all observed variables, where $(X, Y) \subseteq \mathbf{V}$ for features X and labels Y, and let $\mathbb{P}(\mathbf{V})$ denote the training distribution. $\mathbb{E}[\cdot]$ and $\mathrm{cov}(\cdot,\cdot)$ refer to the mean and covariance in $\mathbb{P}$, and for a shifted distribution $\mathbb{P}_\delta$ (Definition 2.1) we use $\mathbb{E}_\delta[\cdot]$, $\mathrm{cov}_\delta(\cdot,\cdot)$. For a random variable Z, we use $\mathcal{Z}$ to denote the space of realizations, and d_Z for dimension e.g., $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$. For a set of random variables $\mathbf{V} = \{V_1, \ldots, V_d\}$, we use V_i to denote an individual element, and use $\mathrm{PA}_\mathcal{G}(V_i)$ to denote the set of parents in a directed acyclic graph (DAG) $\mathcal{G}$, omitting the subscript when otherwise clear.

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

Definition 2.1. A *parameterized robustness set around* $\mathbb{P}(\mathbf{V})$ is a family of distributions $\mathcal{P}$ with elements $\mathbb{P}_\delta(\mathbf{V})$ indexed by $\delta \in \Delta \subseteq \mathbb{R}^{d_\delta}$, with $0 \in \Delta$, and $\mathbb{P}_0(\mathbf{V}) = \mathbb{P}(\mathbf{V})$.

We give examples shortly that satisfy this general definition. To construct such a robustness set, we consider distributions $\mathbb{P}_\delta$ that differ from $\mathbb{P}$ in one or more conditional distributions (Assumption 2.3). We require that the relevant conditional distributions can be described by an exponential family.

Definition 2.2 (Conditional exponential family (CEF) distribution)**.** $\mathbb{P}(W|Z)$ is a conditional exponential family distribution if there exists a function $\eta(Z) : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_T}$ such that the conditional probability density (for continuous W) or probability mass function (for discrete W) is given by

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

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

Definition 2.2 extends to marginal distributions where $Z = \varnothing$ and $\eta(Z)$ is a constant function.

Example 1.1 (Continued)**.** Suppose the probability of ordering a test (O) depends on age (A) and disease (Y), such that $\mathbb{P}(O = 1|A, Y) = \sigma(\eta(A, Y))$, where σ is the sigmoid, and η is an arbitrary function. Here, Definition 2.2 is satisfied with $W = O$, $Z = (A, Y)$, and sufficient statistic $T(O) = O$.

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

Assumption 2.3 (Factorization into CEF distributions)**.** Let $\mathbf{W} = \{W_1, \ldots, W_m\} \subseteq \mathbf{V}$ be a *intervention set* of variables and let

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

be a factorization, where $Z_i, U_j, V_j \subseteq \mathbf{V}$ are possibly overlapping sets of variables. We assume for each W_i that Z_i is known and that $\mathbb{P}(W_i|Z_i)$ satisfies Definition 2.2.

If $\mathbb{P}(\mathbf{V})$ factorizes according to a DAG $\mathcal{G}$, the factorization in Assumption 2.3 is always satisfied by $Z_i = \mathrm{PA}_\mathcal{G}(W_i)$. Here, we require limited knowledge of the underlying graph, and only need to know the parents $\mathrm{PA}(W_i)$ for the variables W_i that shift. We now define parametric perturbations and give the general form of the robustness sets that we consider, involving simultaneous perturbations to multiple W_i.

Definition 2.4 (Parameterized shift functions and δ-perturbations)**.** Let $s(Z; \delta) : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_T}$ be a *parameterized shift function* with parameters $\delta \in \Delta \subseteq \mathbb{R}^{d_\delta}$ which is twice-differentiable with respect to δ and satisfy $s(Z; 0) = 0$ for all Z. For $\mathbb{P}(W|Z)$ satisfying Equation (3), we refer to

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

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

Example 1.1 (Continued)**.** A model developer may be concerned about a uniform change in testing rates across all types of patients. This can be modelled by choosing $s(Z; \delta) = \delta$, for $\delta \in \mathbb{R}$, an additive intervention on the log-odds scale. A separate change in testing rates for sick and healthy could instead be modeled using $s(Z; \delta) = \delta_0(1-Y) + \delta_1 Y$, using $\delta \in \mathbb{R}^2$. This reasoning extends readily to more complex shifts (e.g., allowing for age-specific changes in testing rates, with a non-linear dependence on age), as long as $s(Z; \delta)$ remains a parametric function.

While the shift function $s(Z;\delta)$ is parametric, $\eta(Z)$ is unconstrained in Definitions 2.2 and 2.4. Note that this formulation includes multiplicative shifts $\eta_\delta(Z) = (1+\delta)\eta(Z)$ by letting $s(Z;\delta) = \delta \cdot \eta(Z)$.

Definition 2.5 (CEF parameterized robustness set)**.** For a distribution $\mathbb{P}$ and intervention set $\mathbf{W} = \{W_1, \ldots, W_m\} \subseteq \mathbf{V}$ satisfying Assumption 2.3, let each $\mathbb{P}_{\delta_i}(W_i|Z_i)$ be a δ_i-perturbation (Definition 2.4) of $\mathbb{P}(W_i|Z_i)$. Then

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

is called a δ-perturbation of $\mathbb{P}(\mathbf{V})$, and the robustness set $\mathcal{P}$ consists of all $\mathbb{P}_\delta$ for $\delta \in \Delta_1 \times \cdots \Delta_m$.

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

Remark 2.6 (Causal Interpretation of Shifts). If the DAG $\mathcal{G}$ represents a causal graph (Pearl, 2009), then $\mathbb{P}_\delta$ can interpreted as a change in causal mechanisms. We see this as an important perspective for interpreting and specifying shifts, but our methods do not require a causal interpretation.

3. Evaluation of the worst-case loss

For a fixed predictor and loss function, we can use data from $\mathbb{P}(\mathbf{V})$ to estimate the expected loss $\mathbb{E}_\delta[\ell] := \mathbb{E}_\delta[\ell(f(X), Y)]$ for a fixed δ, and estimate the worst-case loss over all δ of bounded magnitude.

Remark 3.1. The methods here can be used with an arbitrary predictor f and loss function $\ell := \ell(f(X), Y)$. We do not even require access to the original predictor f. Both methods here simply treat ℓ as a random variable in $\mathbb{P}$, for which we have samples from the training distribution.

3.1. Modelling shifted losses using reweighting

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

Proposition 3.2. *For any $\mathbb{P}_\delta(\mathbf{V}), \mathbb{P}(\mathbf{V})$ that satisfy Definition 2.5, $\mathrm{supp}(\mathbb{P}) = \mathrm{supp}(\mathbb{P}_\delta)$ and the density ratio $w_\delta := \mathbb{P}_\delta/\mathbb{P}$ is given by*

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

Example 1.1 (Continued)**.** Suppose we perturb the probability of ordering a test O given age A and disease Y with shift function $s(Y;\delta) = \delta_0(1-Y) + \delta_1 Y$, independently changing the conditional probability of testing for healthy and sick patients. Here, the density ratio is given by

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

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

$$\mathbb{E}_\delta[\ell] \approx \hat{E}_{\delta,\mathrm{IS}} := \frac{1}{n}\sum_{j=1}^n \hat{w}_\delta(\mathbf{V}^j)\ell(\mathbf{V}^j). \tag{5}$$

In practice, Equation (5) can have high variance when density ratios are large, and maximizing this equation with respect to δ is a general non-convex optimization problem, which is generally NP-hard to solve.

3.2. Approximating the shifted loss for exponential family models

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

$$\mathbb{E}_\delta[\ell] \approx \mathbb{E}[\ell] + \delta^\top \mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \mathrm{SG}^2\,\delta, \tag{6}$$

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

Definition 3.3 (Shift gradient and Hessian)**.** For a parametric shift satisfying Definition 2.1 where $\delta \mapsto \mathbb{E}_\delta[\ell]$ is twice-differentiable, we denote the shift gradient SG^1 and Hessian SG^2 as $\mathrm{SG}^1 := \nabla_\delta \mathbb{E}_\delta[\ell]\big|_{\delta=0}$ and $\mathrm{SG}^2 := \nabla^2_\delta \mathbb{E}_\delta[\ell]\big|_{\delta=0}$.

Equation (6) is a local approximation of the loss, whose approximation error we bound in Theorem B.2, with smaller approximation error for smaller shifts. For $\mathbb{P}_\delta$ satisfying Definition 2.5, SG^1 and SG^2 can be computed as expectations in the training distribution, without estimation of density ratios. Recall that the conditional covariance is given by $\mathrm{cov}(A, B|C) := \mathbb{E}[(A - \mathbb{E}[A|C])(B - \mathbb{E}[B|C])|C]$.

Theorem 3.4 (Simple shift in a single variable)**.** *Assume the setup of Theorem B.1, restricted to a shift in a single variable W, and that $s(Z;\delta) = \delta$.*

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

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

In Theorem B.1, we state a general form of Theorem 3.4, which allows for shifts in multiple variables and for arbitrary shift functions $s(Z;\delta)$.

Example 1.1 (Continued)**.** Suppose that age (A), which has no causal parents, follows a normal distribution with mean μ and variance σ^2, and that we wish to consider a shift in the mean. We can parameterize $\mathbb{P}(A)$ as an exponential family with parameter $\eta = \mu/\sigma$ and sufficient statistic $T(A) = A/\sigma$. Here, $s(\delta) = \delta$ implies a shift in the mean of δ standard deviations $\eta_\delta = \eta + s(\delta) = (\mu + \sigma\delta)/\sigma$, and we can write that $\mathrm{SG}^1 = \mathrm{cov}\left(\ell, A\right)/\sigma$ and $\mathrm{SG}^2 = \mathrm{cov}\left(\ell, (A - \mathbb{E}[A])^2\right)/\sigma^2$.

To estimate the shift gradient and Hessian from a sample from $\mathbb{P}$, we fit models $\hat{\mu}_\ell(Z) \approx \mathbb{E}[\ell|Z]$ and $\hat{\mu}_W(Z) \approx \mathbb{E}[T(W)|Z]$ and compute residuals on these predictions, which permits estimation of the gradient and Hessian as a sample average of residuals. Using these, we can estimate the expected loss as

$$\mathbb{E}_\delta[\ell] \approx \hat{E}_{\delta,\text{Taylor}} := \hat{\mathbb{E}}[\ell] + \delta^\top \hat{\mathrm{SG}}^1 + \tfrac{1}{2}\delta^\top \hat{\mathrm{SG}}^2 \delta. \quad (7)$$

Here, there are two sources of error: Finite-sample error, due to the estimates of $\mathrm{SG}^1, \mathrm{SG}^2$, as well as approximation error; in Theorem B.2 we give a bound on the latter. In exchange for considering a second-order approximation of the loss, we gain two benefits: Variance reduction and tractable optimization. First, as $\mathrm{SG}^1, \mathrm{SG}^2$ are not functions of δ, the variance of $\hat{E}_{\delta,\text{Taylor}}$ is $O(\|\delta\|^4)$, while the variance of $\hat{E}_{\delta,\text{IS}}$ can be much larger (see Section 4). Second, maximizing $\hat{E}_{\delta,\text{Taylor}}$ over the set $\|\delta\| \leq \lambda$ can be solved in polynomial time by exploiting the quadratic structure (see Section 3.3), while maximizing $\hat{E}_{\delta,\text{IS}}$ over the constraints is generally hard, and may be infeasible in high dimensions.

3.3. Identifying worst-case parametric shifts

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

$$\sup_{\|\delta\|_2 \leq \lambda} \delta^\top \mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \mathrm{SG}^2 \delta. \quad (8)$$

Since SG^2 is generally not negative definite, the maximization objective is non-concave. However, this particular problem is an instance of the 'trust region problem' which is well-studied in the optimization literature (Conn et al., 2000), and can be solved in polynomial time (Pólik & Terlaky, 2007, Section 8.1).

4. Experiments

Synthetic Example: Lab Tests: To compare the bias and variance of the Taylor and the importance sampling estimates of the shifted loss, we simulate synthetic data from

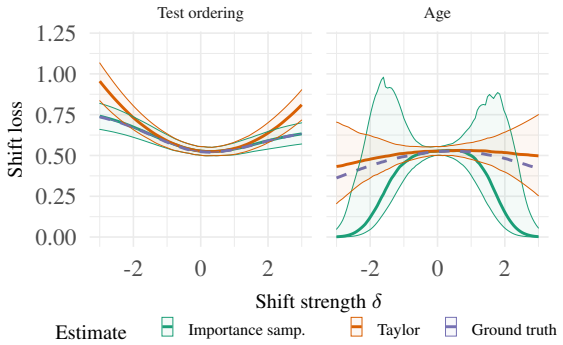

Figure 2. Estimates of the shifted loss in Section 4 using either the importance sampling or the Taylor estimate. For each shift δ we plot the median, and the 0.05 and 0.95 quantiles for each approach. The dashed line indicates simulated ground truth (indistinguishable from the importance sampling in the left plot).

Example 1.1. We consider either a shift in the logits of ordering lab tests (Figure 2 left) or a mean shift in the Gaussian distribution of age (Figure 2 right). We compute estimates $\hat{E}_{\delta,\text{IS}}$ and $\hat{E}_{\delta,\text{Taylor}}$ of the loss under a shift of size δ, and compare this to ground truth data simulated from $\mathbb{P}_\delta$. Additional details are given in Appendix C.

For shifts in binary test ordering, both estimates capture the loss well for small shifts, but as δ gets larger, the quadratic approximation increasingly deviates from the true mean. For the Gaussian mean shift, the importance sampling weights quickly become ill-behaved, and the variance dramatically increases as δ grows. 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.

Finding worst-case shifts in computer vision, given image attributes: In Appendix A we demonstrate the application of the approach outlined in Section 3.3, to find bounded worst-case shifts in a gender classification task using a synthetic variant of the CelebA dataset. Here, we uncover (and validate) sensitivity to interpretable changes in distribution, such as reduced rates of wearing lipstick among young women.

5. Conclusion

We argue for considering parametric shifts in distribution, to evaluate model performance under a set of changes that are interpretable and controllable. For parametric shifts in conditional exponential family distributions, we derive a local second-order approximation to the loss under shift. This approximation enables the use of efficient optimization algorithms (to find the worst-case shift), and empirically provides realistic estimates of the resulting loss.

References

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

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

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

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

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

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

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

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

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

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

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

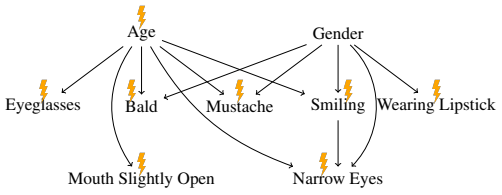

Figure 3. Causal graph over attributes, where lightning bolts indicate changes in mechanisms.

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

A. Experiment: Detecting sensitivity to spurious correlations

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

Setup: To obtain ground-truth shifts in distribution, we generate synthetic datasets of faces using CausalGAN (Kocaoglu et al., 2018), trained on the CelebA data. We simulate attributes following the causal graph in Figure 3, and then simulate images from the GAN conditioned on those attributes. We draw a training sample from this distribution $\mathbb{P}$, and fit a gender classifier $f(X)$ using the image data alone, by finetuning a pretrained ResNet50 classifier (Hu et al., 2018), with a test accuracy of 0.919. Each attribute W_i is binary, so we consider shifts in the log-odds $\eta_i(Z_i)$ of each attribute W_i given parents Z_i. Here, we use a maximally flexible shift function $s_i(Z_i; \delta_i) = \sum_{z \in \mathcal{Z}_i} \delta_{i,z} \mathbf{1}\{Z_i = z\}$, such that for $Z_i \in \{0,1\}^k$ there are 2^k parameters. Across all intervened variables, $\delta \in \mathbb{R}^{31}$. Per Section 3.3, we identify the shift δ causing the largest drop in accuracy, and compute $\hat{E}_{\delta,\text{Taylor}}$ and $\hat{E}_{\delta,\text{IS}}$ to estimate accuracy under that shift. Due to the synthetic nature of our setup, we can simulate from $\mathbb{P}_\delta(X, \mathbf{W}, Y)$ to estimate the ground-truth impact of this shift, by simulating from the shifted attribute distribution, and then simulating images from the GAN conditional

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

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

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

Figure 4. (Top) Model accuracy at randomly drawn shifts. (Bottom) Difference in accuracy in the worst-case shifts identified by Taylor and importance sampling approaches. The Taylor method identifies a more adversarial shift than importance sampling in 73% of simulations (green).

on those attributes.

Finding and validating a single, high impact shift: Using a validation sample from $\mathbb{P}$, we estimate the shift gradient and Hessian (Theorem B.1). Solving the quadratic optimization problem in Section 3.3, we find the worst-case shift δ such that $\|\delta\| \leq \lambda = 2$. We display the largest components of δ in Table 1 (top). Among others, this shift entails a 10% increase in the probability of a young man wearing lipstick, and a similar decrease for young women. This suggests that the learned classifier f relies on this non-causal association in the images for prediction. We validate that this shift leads to a measurable decrease in accuracy, from 92% to 89%, using simulated data from $\mathbb{P}_\delta$. To validate that this shift is indeed a worst-case shift, we simulate $K = 400$ random shifts δ_k and evaluate the model accuracy in $\mathbb{P}_{\delta_k}$ (Figure 4, top). As expected, our chosen shift δ (red line) is in the left tail of the distribution. We compare the ground-truth accuracy under this shift (89%) to the original estimates $\hat{E}_{\delta,\text{Taylor}} = 0.878$ and $\hat{E}_{\delta,\text{IS}} = 0.792$ (Table 1), and observe that both correctly predict a drop in accuracy, although $\hat{E}_{\delta,\text{IS}}$ overestimates the size of the drop.

Comparing importance sampling and Taylor across multiple simulations: We simulate $K = 400$ validation sets from $\mathbb{P}$, in each estimating the worst-case shifts δ_{Taylor} and δ_{IS}, where the latter corresponds to maximizing $\hat{E}_{\delta,\text{IS}}$ using a standard non-convex solver from the `scipy` library (Virtanen et al., 2020). We simulate ground truth data from $\mathbb{P}_{\delta_{\text{IS}}}$ and $\mathbb{P}_{\delta_{\text{Taylor}}}$, and in Figure 4 (bottom) we plot the differences $\mathbb{E}_{\delta_{\text{Taylor}}}[\ell] - \mathbb{E}_{\delta_{\text{IS}}}[\ell]$, showing that in 73% of cases, the Taylor method finds a more impactful shift. Moreover, the average run-time for the Taylor approach is $0.02s$ while that of the importance sampling approach is $2.52s$. Finally, the optimal value of the $\hat{E}_{\delta,\text{Taylor}}$ objective tends to a reasonably accurate estimate of the shifted accuracy, while the optimal value of the $\hat{E}_{\delta,\text{IS}}$ objective (both on the validation set) is a poor predictor, with a mean absolute prediction error (in predicting $\mathbb{E}_\delta[\ell]$) of 0.176 for $\hat{E}_{\delta,\text{IS}}$ and 0.014 for $\hat{E}_{\delta,\text{Taylor}}$.

B. Additional theorems

Theorem B.1 (Shift gradients and Hessians as covariances)**.** *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy Definition 2.5, 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 $\text{SG}^1 = (\text{SG}^1_1, \ldots, \text{SG}^1_m) \in \mathbb{R}^{d_\delta}$ where*

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

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

$$\mathbb{E}\left[D_{i,1}^\top \operatorname{cov}\left(\ell,\, \epsilon_{T_i|Z_i}\epsilon_{T_i|Z_i}^\top | Z_i\right) D_{i,1}\right] - \mathbb{E}\left[\ell \cdot D_{i,2}^\top \epsilon_{T|Z}\right]$$

and the (i,j)th off-diagonal block of size $d_{\delta_i} \times d_{\delta_j}$ equals

$$\operatorname{cov}(\ell,\; D_{i,1}^\top \epsilon_{T_i|Z_i}\epsilon_{T_j|Z_j}^\top D_{j,1})$$

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

Theorem B.2. *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy the conditions of Theorem B.1, with a shift in a single variable W, where $s(Z; \delta) = \delta$. Let $E_{\delta,\text{Taylor}}$ be the population Taylor estimate (Equation (6)) and let $\sigma(M)$ denote the largest absolute*

value of the eigenvalues of a matrix M. Then

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

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

C. Simulation details for Section 4

The data in Section 4 (and also the loss land-scape in Figure 1 (right)), was simulated using the following generative model

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

where σ denotes the sigmoid function and if Order $= 0$, the test result is a placeholder value of zero.

For the lab test ordering, we consider the shift $\eta_\delta(Z) = \eta(Z) + \delta$ (i.e. a linear shift in the logits of test ordering) and for age, we consider a shift in the marginal mean $\eta_\delta = \delta$. To construct the plot, we simulate data $n = 1{,}000$ times, and for each dataset and each δ in a grid, we compute estimates $\hat{E}_{\delta,\text{IS}}$ and $\hat{E}_{\delta,\text{Taylor}}$. We then plot the mean and point-wise 90% prediction intervals for $\hat{E}_{\delta,\text{IS}}$ and $\hat{E}_{\delta,\text{Taylor}}$.

D. Proofs

D.1. Proof of Proposition 3.2

Proposition 3.2. *For any $\mathbb{P}_\delta(\mathbf{V}), \mathbb{P}(\mathbf{V})$ that satisfy Definition 2.5, $\mathrm{supp}(\mathbb{P}) = \mathrm{supp}(\mathbb{P}_\delta)$ and the density ratio $w_\delta := \mathbb{P}_\delta/\mathbb{P}$ is given by*

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

Proof. By Definition 2.5 and Assumption 2.3, we have that

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

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

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

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

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

By Definition 2.5 and Assumption 2.3, each $\mathbb{P}_{\delta_i}(W_i|Z_i)$ is a δ_i-perturbation around the CEF distribuition $\mathbb{P}(W_i|Z_i)$, so plugging in the exponential family densities, we get

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

□

D.2. Proof of Theorem B.1

Theorem B.1 (Shift gradients and Hessians as covariances)**.** *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy Definition 2.5, with intervened variables $\mathbf{W} = \{W_1, \ldots, W_m\}$ and shift functions $s_i(Z_i;\delta_i)$, where $\delta = (\delta_1, \ldots, \delta_m)$. Then the shift gradient is given by $\mathrm{SG}^1 = (\mathrm{SG}^1_1, \ldots, \mathrm{SG}^1_m) \in \mathbb{R}^{d_\delta}$ where*

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

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

$$\mathbb{E}\left[D_{i,1}^\top \mathrm{cov}\left(\ell,\, \epsilon_{T_i|Z_i}\epsilon_{T_i|Z_i}^\top|Z_i\right)D_{i,1}\right] - \mathbb{E}\left[\ell\cdot D_{i,2}^\top\epsilon_{T|Z}\right]$$

and the (i,j)th off-diagonal block of size $d_{\delta_i}\times d_{\delta_j}$ equals

$$\mathrm{cov}(\ell,\ D_{i,1}^\top\epsilon_{T_i|Z_i}\epsilon_{T_j|Z_j}^\top D_{j,1})$$

where $D_{i,k} := \nabla^k_{\delta_i} s_i(Z_i;\delta_i)|_{\delta=0}$, is the gradient of the shift function for $k = 1$, and the Hessian for $k = 2$. Here,

$T_i(W_i)$ is the sufficient statistic of $\mathbb{P}(W_i|Z_i)$ and $\epsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}[T(W_i)|Z_i]$.

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

Shift Gradient: By Definition 2.5, the probability density / mass function $\mathbb{P}_\delta$ factorizes as follows, where $\delta = (\delta_1, \ldots, \delta_m)$

$$\mathbb{P}_\delta(\mathbf{V}) = \left(\prod_{W_i \in \mathbf{W}} \mathbb{P}_{\delta_i}(W_i|Z_i)\right)\left(\prod_{V_i \in \mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_i|\,\mathrm{PA}(V_i))\right), \tag{9}$$

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 2.4, we can observe that

$$\begin{aligned}
&\nabla_{\delta_i}\log p_{\delta_i}(w_i|z_i)\\
&= \nabla_{\delta_i}\big[(\eta_i(z_i) + s_i(z_i;\delta_i))^\top T_i(w_i)\\
&\qquad - h_i(\eta(z_i) + s_i(z_i;\delta_i))\big]\\
&= (\nabla_{\delta_i} s_i(z_i;\delta_i))^\top T_i(w_i)\\
&\qquad - (\nabla_{\delta_i} s_i(z_i;\delta_i))^\top \nabla h_i(\eta(z_i) + s_i(z_i;\delta_i))\\
&= (\nabla_{\delta_i} s_i(z_i;\delta_i))^\top (T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i)))
\end{aligned} \tag{10}$$

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

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

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

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

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

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.

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

where we use the notation $v^{\otimes 2} := vv^\top$, and we note that $\nabla_{\delta_i}^2 s(z_i;\delta_i)$ 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 $h^{(2)}(\eta_{\delta_i}(z_i)) = \operatorname{var}_\delta[T_i(W_i)|z_i]$ (Wainwright et al., 2008, Proposition 3.1), and the definition $\epsilon_{T_i|Z_i} = T_i(W_i) -$

$\mathbb{E}_\delta[T_i(W_i)|Z_i]$, we obtain

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

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

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

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

where the third line follows from the fact that $\nabla_{\delta_i}(\nabla_{\delta_j}\log p_{\delta_j}(w_j|z_j)) = 0$, and the last line follows from the derivation of the gradient of the log-likelihood in Equation (10). We can again use the fact that $h^{(1)}_i(\eta_{\delta_i}(Z_i)) = \mathbb{E}_\delta[T_i(W_i)|Z_i]$ and the shorthand $\epsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}_\delta[T_i(W_i)|Z_i]$ to write that

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

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

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

Where the last equality follows because $\mathbb{E}[D^\top_{i,1}\epsilon_{T_i|Z_i}\epsilon^\top_{T_j|Z_j}D_{j,i}] = 0$. 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 (9), 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, $\mathbb{E}[D^\top_{i,1}\epsilon_{T_i|Z_i}|Z_i] = 0$, for all Z_i. Moreover, given Z_i, we have that $D^\top_{i,1}\epsilon_{T_i|Z_i}$ is independent of $D^\top_{j,1}\epsilon_{T_j|Z_j}$. As a result, we can write that

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

□

D.3. Proof of Theorem 3.4

Theorem 3.4 (Simple shift in a single variable)**.** *Assume the setup of Theorem B.1, restricted to a shift in a single variable W, and that $s(Z;\delta) = \delta$.*

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

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 $\nabla^2_\delta s(Z;\delta) = \nabla^2_\delta\delta = 0$. The result now follows from Theorem B.1. □

D.4. Proof of Theorem B.2

Theorem B.2. *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy the conditions of Theorem B.1, with a shift in a single variable W, where $s(Z;\delta) = \delta$. Let $E_{\delta,\textit{Taylor}}$ be the population Taylor estimate (Equation (6)) and let $\sigma(M)$ denote the largest absolute value of the eigenvalues of a matrix M. Then*

$$\begin{aligned}
&\left|\mathbb{E}_\delta[\ell] - E_{\delta,\textit{Taylor}}\right| \leq \\
&\quad \tfrac{1}{2}\sup_{t\in[0,1]}\sigma\Big(\mathrm{cov}_{t\cdot\delta}(\ell, \epsilon_{t\cdot\delta,T|W}\epsilon^\top_{t\cdot\delta,T|W}) - \\
&\qquad\qquad \mathrm{cov}(\ell, \epsilon_{0,T|W}\epsilon^\top_{0,T|W})\Big)\cdot\|\delta\|^2,
\end{aligned}$$

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

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

$$\mathbb{E}_{1\cdot\delta}[\ell] - \mathbb{E}_{0\cdot\delta}[\ell] - \left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\bigg|_{t=0} = \left(\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}^2 t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\bigg|_{t=t_0}. \tag{11}$$

We have $\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=0} = \mathrm{SG}^1$ and by the same arguments (see the proof of Theorem B.1), it follows that $\left(\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}^2 t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=t_0} = \delta^\top \operatorname{cov}_{t_0\cdot\delta}(\ell, \epsilon^{\otimes 2}_{t_0\cdot\delta,T|W})\delta$. Plugging this in, and subtracting $\frac{1}{2}\delta^\top \mathrm{SG}^2\, \delta$ on both sides of Equation (11) yields

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

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

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

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

$$\begin{aligned}
&\left|\mathbb{E}_\delta[\ell] - \kappa(\delta)\right| \\
&\leq \frac{1}{2}\sup_{t\in[0,1]} \sigma\left(\operatorname{cov}_{t\cdot\delta}(\ell, \epsilon^{\otimes 2}_{t\cdot\delta,T|W}) - \operatorname{cov}(\ell, \epsilon^{\otimes 2}_{0,T|W})\right)\|\delta\|_2^2,
\end{aligned}$$

which concludes the proof. □