A Causally Interpreting Distribution Shifts

Under certain conditions, shifts in a conditional distribution $P(W \mid Z)$ have an important interpretation as causal *policy interventions* or *process changes* (Pearl, 2009). That is, the effects of the shift corresponds to how the distribution would change under an intervention that changes the way W is generated. Formally, we have the following:

Proposition 2. *Suppose the data (X, Y) were generated by a structural causal model (SCM) with no unobserved confounders, respecting a causal directed acyclic graph (DAG) $\mathcal{G}$. Then, for a single variable W and set $Z = nd_{\mathcal{G}}(W)$ (non-descendants of W in $\mathcal{G}$), a policy shift in $P(W \mid Z)$ can be expressed as a policy intervention on the mechanism generating W which changes $P(W \mid pa_{\mathcal{G}}(W))$.*

Proof. Within the SCM, we have that W is generated by the structural equation $W = f_w(pa(W), \varepsilon_w)$ (where ε is a W-specific exogenous noise random variable). A policy intervention on W replaces this structural equation with a new function $g_w(pa(W), \varepsilon_w)$, which has the effect of changing $P(W \mid pa(W))$ to some new distribution $Q(W \mid pa(W)$. By the local Markov property, we have that $Q(W \mid pa(W)) = Q(W \mid nd(W))$. Thus, a shift from $P(W \mid nd(W))$ to $Q(W \mid nd(W))$ can be expressed as a policy intervention from f_w to g_w. □

This result means that in order to causally interpret distribution shifts, we need to adjust for (i.e., put into Z) variables that are relevant to the mechanism that generates W. Fortunately, we can place additional variables into Z so long as they precede W in a causal or topological order.

This result can be extended to the case in which the SCM contains unobserved variables. This is of practical importance because often we do not have all relevant variables recorded in the dataset (i.e., there may be unobserved confounders). In these cases, rather than a DAG, the SCM takes the graphical form of a causal acyclic directed mixed graph (ADMG) (Richardson et al., 2017). ADMGs have directed edges ($\to$) which represent direct causal influence (just like in DAGs), but also have bidirected edges ($\leftrightarrow$) which represent the existence of an unobserved confounder between the two endpoint variables.

We require one technical definition: we will define the Markov blanket of W in an ADMG to be $mb(W) = pa(dis(W)) \cup (dis(W) \setminus \{W\})$, where *dis* refers to the *district* of a variable and is the set of variables reachable through entirely bidirected paths.

Proposition 3. *Suppose the data (X, Y) were generated by a structural causal model (SCM), respecting a causal ADMG $\mathcal{G}$. Then, for a single variable W and set $Z = mb_{\mathcal{G}}(W)$, a policy shift in $P(W \mid Z)$ can be expressed as a policy intervention on the mechanism generating W which changes $P(W \mid mb_{\mathcal{G}}(W))$.*

Proof. The proof follows just as before, noting that the local Markov property in ADMGs states that a variable V is independent of all variables preceding V in a topological order conditional on $mb(V)$ (Richardson et al., 2017, Section 2.8.2). □

B Connection to distributionally robust optimization

In distributionally robust optimization (DRO), a model is trained to minimize loss on the worst-case distribution from within an uncertainty set of distributions. Here we show how Equation 1 of the main paper can equivalently be thought of a the expected loss under the worst-case distribution from within just such an uncertainty set. Formally, we define an uncertainty set or "ball" $\mathcal{P}_{\rho,Z,\mathcal{W}}$ of possible shifted distributions using a statistical divergence $D(\cdot \parallel \cdot)$ and radius ρ:

$$\mathcal{P}_{\rho,Z,\mathcal{W}} = \{Q : D(Q(W) \parallel P(W \mid Z)) \leq \rho\}. \tag{8}$$

Note that this uncertainty set depends explicitly on the value of Z. We are interested in the expected loss of the model when $P(W \mid Z)$ is replaced by $Q(W \mid Z) = Q_Z \in \mathcal{P}_{\rho,Z,\mathcal{W}}$ that maximizes expected loss, written

$$R_\rho(\mathcal{M}; P) = \mathbb{E}_P \left[\sup_{Q_Z \in \mathcal{P}_{\rho,Z,\mathcal{W}}} \mathbb{E}_{Q_Z}[\mu_0(W, Z) \mid Z] \right], \tag{9}$$

where, as in the main paper, $\mu_0(W, Z) = \mathbb{E}_P[\ell(Y, \mathcal{M}(X)) \mid W, Z]$ is the conditional expected loss given W and Z. By construction of $\mathcal{P}_{\rho,Z,\mathcal{W}}$, Q will never place positive weight on regions where P does not. Calculating R_ρ requires calculating the expected loss under various distributions Q. As in previous work on DRO and domain adaptation, we rely on *sample reweighting*, which allows us to estimate expectations under Q using samples from P. This is done by reweighting samples from P by the likelihood ratio q/p. Specifically, the expected loss under Q can be rewritten as

$$\mathbb{E}_Q[\mu_0(W, Z) \mid Z] = \mathbb{E}_P\left[\frac{q(W \mid Z)}{p(W \mid Z)}\mu_0(W, Z) \;\middle|\; Z\right]. \tag{10}$$

If Q is quite different from P, the variance of importance sampling can be high. This variance is naturally governed by ρ, which controls how different Q can be from P. In order to consider environments that look *very* different from P, a large test dataset may be needed.

To see how this formulation connects to Equation 1 of the main paper, we will make use of the following lemma which follows directly from Theorem 6 in Van Erven and Harremos (2014)

Lemma 4. *For probability measures P and Q defined with respect to the same base measure μ and with corresponding density functions p and q, $\sup_{A\in\mathcal{F}} \frac{Q(A)}{P(A)} \leq c$ if and only if $\frac{q}{p} \leq c$ almost everywhere.*

Using this lemma, we can rewrite Equation 1 as

$$\begin{aligned} &\sup_Q \;\mathbb{E}_P\left[\mathbb{E}_Q[\mu_0(X) \mid Z]\right] \\ &\text{s.t.} \quad \frac{q(W \mid Z)}{p(W \mid Z)} \leq \exp(\rho) \quad \textit{a.e.} \end{aligned} \tag{11}$$

This can, in turn, be rewritten as

$$\begin{aligned} &\sup_Q \;\mathbb{E}_P\left[\frac{q(W \mid Z)}{p(W \mid Z)}\mu_0(X)\right] \\ &\text{s.t.} \quad \frac{q(W \mid Z)}{p(W \mid Z)} \leq \exp(\rho) \quad \textit{a.e.} \end{aligned} \tag{12}$$

Define $\exp(\rho)h(w, z) = \frac{q(w|z)}{p(w|z)}$ and $\exp(\rho) = \frac{1}{1-\alpha}$. Then, the constraint in Equation 12, combined with the fact that p and q are both densities and thus are bounded below by zero, translates to $h : \mathcal{X} \to [0, 1]$. Further, the constraint that q must integrate to one (or, equivalently, $\mathbb{E}_P[q(W \mid Z)/p(W \mid Z) \mid Z] = 1$ almost everywhere) translates to the constraint $\mathbb{E}_P[h(W, Z) \mid Z] = 1 - \alpha$ almost everywhere. Finally, we can rewrite this optimization problem as

$$\begin{aligned} &\sup_{h:\mathcal{W}\times\mathcal{Z}\to[0,1]} \frac{1}{1-\alpha}\mathbb{E}_P\left[h(W, Z)\mu_0(W, Z)\right] \\ &\text{s.t.} \quad \mathbb{E}_P[h(W, Z) \mid Z] = 1 - \alpha \quad \textit{a.e.} \end{aligned} \tag{13}$$

C Derivation of Equation 3 of the main paper

To derive Equation 3 of the main paper, we will take the Lagrange dual of Equation 1. Recall that Equation 1 is given by

$$R_{\alpha,0} = \sup_{h:\mathcal{W}\times\mathcal{Z}\to[0,1]} \frac{1}{1-\alpha}\mathbb{E}_P\left[h(W, Z)\mu_0(W, Z)\right] \tag{14}$$

$$\text{s.t.} \quad \mathbb{E}_P[h(W, Z) \mid Z] = 1 - \alpha \quad \textit{a.e.} \tag{15}$$

Then the Lagrangian is given by

$$\mathcal{L}(h,\nu) = \frac{1}{1-\alpha}\mathbb{E}_P\left[h(W,Z)\mu_0(W,Z)\right] + \int \nu(z)(1-\alpha-\mathbb{E}_P[h(W,Z)\mid Z=z])\,dz, \tag{16}$$

where $\nu : \mathcal{Z} \to \mathbb{R}$ is the function of Lagrange multipliers. By recalling that p is the density function associated with P and by defining $\eta(z) = \nu(z)\frac{1-\alpha}{p(z)}$ we get

$$\begin{aligned}
\mathcal{L}(h,\eta) &= \frac{1}{1-\alpha}\mathbb{E}_P\left[h(W,Z)\mu_0(W,Z)\right] + \int_z p(z)\eta(z)(1-\frac{1}{1-\alpha}\mathbb{E}_P[h(W,Z)\mid Z=z]\,dz && (17)\\
&= \frac{1}{1-\alpha}\mathbb{E}_P\left[h(W,Z)\mu_0(W,Z)\right] + \mathbb{E}_P[\eta(Z)] - \frac{1}{1-\alpha}\mathbb{E}_P[h(W,Z)\eta(Z)] && (18)\\
&= \frac{1}{1-\alpha}\mathbb{E}_P\left[h(W,Z)(\mu_0(W,Z)-\eta(Z))\right] + \mathbb{E}_P[\eta(Z)]. && (19)
\end{aligned}$$

The Lagrange dual is then given by $\min_{\eta:\mathcal{Z}\to\mathbb{R}}\max_{h:\mathcal{Z}\times\mathcal{W}\to[0,1]}\mathcal{L}(h,\eta)$. To maximize h out of this equation, observe that the optimal solution occurs when h equals one whenever $\mu_0-\eta$ is positive and zero when $\mu_0-\eta$ is negative or $h(z,w) = \mathbb{I}(\mu_0(w,z) > \eta(z))$. Then observe that $\mathbb{I}(\mu_0(w,z) > \eta(z))(\mu_0(w,z)-\eta(z))$ can be rewritten as $(\mu_0(w,z)-\eta(z))_+$ where $(x)_+ = \max\{x,0\}$. Finally, because the original problem is a linear program, strong duality holds and we arrive at our final expression

$$R_{\alpha,0} = \min_{\eta:\mathcal{Z}\to\mathbb{R}} \frac{1}{1-\alpha}\mathbb{E}_P\left[(\mu_0(W,Z)-\eta(Z))_+\right] + \mathbb{E}_P[\eta(Z)] \tag{20}$$

D Proof of Theorem 1

In this section, we provide a proof of Theorem 1 in the main paper. This proof draws heavily on results from Chernozhukov et al. (2018) and Jeong and Namkoong (2020). For notational simplicity and consistency between our work and theirs, let $\theta_0 = R_{\alpha,0}$ be the target parameter. Algorithm 1 in the main paper is an instance of the DML2 algorithm from Chernozhukov et al. (2018) where the score function ψ is given by

$$\psi(O;\theta,\gamma) = \psi^b(O;\gamma) - \theta, \tag{21}$$

where

$$\psi^b(O;\gamma) = \frac{1}{\alpha}(\mu(W,Z)-\eta(Z))_+ + \eta(Z) + \frac{1}{\alpha}h(W,Z)(\ell(Y,\mathcal{M}(X))-\mu(W,Z)), \tag{22}$$

and where $O = (W,Z,V)$, $\gamma = (\mu,\eta)$, and $h = [\mu > \eta]$. In this proof, we will show that Assumptions 3.1 and 3.2 of Chernozhukov et al. (2018) are nearly satisfied and will fill in the gaps where they are not. We restate these assumptions here with some of the notation changed to match the notation used in this paper. First, some definitions: Let $c_0 > 0$, $c_1 > 0$, $s > 0$, $q > 2$ be some finite constants such that $c_0 \leq c$ and let $\{\delta_N\}_{N\geq 1}$ and $\{\Delta_N\}_{N\geq 1}$ be some positive constants converging to zero such that $\delta_N \geq N^{-1/2}$. Also, let $K \geq 2$ be some fixed integer, and let $\{\mathcal{P}_N\}_{N\geq 1}$ be some sequence of sets of probability distributions P of O on $\mathcal{O} = \mathcal{W}\times\mathcal{Z}\times\mathcal{V}$. Let T be a convex subset of some normed vector space repressenting the set of possible nuissance parameters (i.e., $\gamma \in T$). Finally, let $a \lesssim b$ denote that there exists a constant C such that $a \leq Cb$.

Assumption 2. (Assumption 3.1 from Chernozhukov et al. (2018)) For all $N \geq 3$ and $P \in \mathcal{P}_N$, the following conditions hold. (a) The true parameter value θ_0 obeys $\mathbb{E}_P[\psi(O;\theta_0,\gamma_0)] = 0$. (b) The score ψ can be written as $\psi(O;\theta,\gamma) = \psi^a(O;\gamma)\theta + \psi^b(O;\gamma)$. (c) The map $\gamma \mapsto \mathbb{E}_P[\psi(O;\theta,\gamma)]$ is twice continuously Gateaux-differentiable on T. (d) The score ψ obeys Neyman orthogonality. (e) The singular values of the matrix $J_0 = \mathbb{E}_P[\psi^a(O;\gamma_0)]$ are between c_0 and c_1.

Assumption 3. (Assumption 3.2 from Chernozhukov et al. (2018)) For all $N \geq 3$ and $P \in \mathcal{P}_N$, the following conditions hold. (a) Given a random subset I of $[N]$ of size $n = N/K$, the nuisance paramter estimator

$\hat{\gamma} = \hat{\gamma}((O_i)_{i\in I^c})$ belongs to the realization set $\mathcal{T}_N$ with probability at least $1-\Delta_N$, where $\mathcal{T}_N$ contains γ_0 and is constrained by the next conditions. (b) The following moment conditions hold:

$$\begin{aligned} m_N &= \sup_{\gamma\in\mathcal{T}_N} (\mathbb{E}_P[\|\psi(O;\theta_0,\gamma)\|^q])^{1/q} \leq c_1 \\ m'_N &= \sup_{\gamma\in\mathcal{T}_N} (\mathbb{E}_P[\|\psi^a(O;\gamma)\|^q])^{1/q} \leq c_1. \end{aligned}$$

(c) The following conditions on the statiestical rates r_N, r'_N, and λ'_N hold:

$$\begin{aligned} r_N &= \sup_{\gamma\in\mathcal{T}_N} \|\mathbb{E}_P[\psi^a(O;\gamma)] - \mathbb{E}_P[\psi^a(O;\gamma_0)]\| \leq \delta_N, \\ r'_N &= \sup_{\gamma\in\mathcal{T}_N} (\mathbb{E}_P[\|\psi(O;\theta,\gamma) - \mathbb{E}_P[\psi(O;\theta_0,\gamma_0)\|^2])^{1/2} \leq \delta_N, \\ \lambda'_N &= \sup_{r\in(0,1),\gamma\in\mathcal{T}_N} \|\partial_r^2 \mathbb{E}_P[\psi(O;\theta_0,\gamma_0 + r(\gamma-\gamma_0))]\| \leq \delta_N/\sqrt{N}. \end{aligned}$$

(d) The variance of the score ψ is non-degenerate: All eigenvalues of the matix $\mathbb{E}_P[\psi(O;\theta_0,\gamma_0)\psi(O;\theta_0,\gamma_0)']$ are bounded from below by c_0.

Here, we will show that all of these conditions are satisfied except for Assumption 2 (c) and, by extension the bound on λ'_N in Assumption 3. These two conditions are used in Chernozhukov et al. (2018) to prove that, for any sequence $\{P_N\}_{N\geq 1}$ such that $P_N \in \mathcal{P}_N$, the following holds for all $P_N \in \mathcal{P}_N$

$$\|R_{N,2}\| = \mathcal{O}_{P_N}(\delta_N/\sqrt{N}), \tag{23}$$

where

$$R_{N,2} = \frac{1}{K}\sum_k \mathbb{E}_{n,k}[\psi(O;\theta_0,\hat{\gamma}_k)] - \frac{1}{N}\sum_{i=1}^N \psi(W_i;\theta_0,\gamma_0), \tag{24}$$

and where $\mathbb{E}_{n,k}[\cdot] = \frac{1}{n}\sum_{i\in I_k}(\cdot)$ is the empirical expectation w.r.t. the k'th cross-validation fold. We will prove this using other means. First, we establish that all other conditions in Assumptions 2 and 3 hold for all $P_N \in \mathcal{P}_N$. For notational simplicity, we will drop the dependence of ℓ, γ, and h on O throughout. Additionally, denote by $\mathcal{E}_N$ the event that $\gamma_k \in \mathcal{T}_N$.

Proof of Assumption 2 (a) This holds trivially via the definitions of θ_0 and μ_0.

$$\mathbb{E}_{P_N}[\psi(O;\theta_0,\gamma_0)] = \mathbb{E}_{P_N}\left[\frac{1}{\alpha}(\mu_0-\eta_0)_+ + \eta_0 + \frac{1}{\alpha}h_0(\ell-\mu_0) - \frac{1}{\alpha}(\mu_0-\eta_0)_+ - \eta_0\right] \tag{25}$$

$$= \mathbb{E}_{P_N}\left[\frac{1}{\alpha}h_0(\ell-\mu_0)\right] = \mathbb{E}_{P_N}\left[\frac{1}{\alpha}h_0(\mu_0-\mu_0)\right] = 0 \tag{26}$$

Proof of Assumption 2 (b) This holds trivially with $\psi^a = -1$.

Proof of Assumption 2 (d) To show Neyman orthogonality of ψ, we must show that, for $P_N \in \mathcal{P}_N$, T the set of possible nuissance parameter values, and $\tilde{T} = \{\gamma - \gamma_0 : \gamma \in T\}$, the Gateaux derivative map $D_r : \tilde{T} \to \mathbb{R}$ exists for all $r \in [0,1)$ where

$$D_r[\gamma-\gamma_0] = \partial_r\left\{\mathbb{E}_{P_N}\left[\psi(O;\theta_0,\gamma_0 + r(\gamma-\gamma_0))\right]\right\},\ \gamma\in T,$$

and that $D_r[\gamma-\gamma_0]$ vanishes for $r=0$. For notational simplicity, let $\mu_r = \mu_0 - r(\mu-\mu_0)$, with analgous definitions for η_r and h_r. Then, using Danskin's theorem, $D_r[\gamma-\gamma_0]$ exists for $r\in[0,1)$ and is given by

$$\begin{aligned} D_r[\gamma-\gamma_0] =& \mathbb{E}_{P_N}\left[\frac{1}{\alpha}[\mu_r \geq \eta_r]((\mu-\mu_0) - (\eta-\eta_0)) + (\eta-\eta_0)\right. \\ & \left. + \frac{1}{\alpha}(h-h_0)(l-\mu_r) - \frac{1}{\alpha}h_r(\mu-\mu_0)\right] \end{aligned}$$

Finally, we have

$$\begin{aligned}
D_r[\gamma-\gamma_0]|_{r=0} &= \mathbb{E}_{P_N}\left[\frac{1}{\alpha}[\mu_0 \geq \eta_0]((\mu-\mu_0)-(\eta-\eta_0)) + (\eta-\eta_0) + \frac{1}{\alpha}(h-h_0)(l-\mu_0) - \frac{1}{\alpha}(h_0)(\mu-\mu_0)\right] \\
&= \mathbb{E}_{P_N}[\eta-\eta_0] - \frac{1}{\alpha}\mathbb{E}_{P_N}[h_0(\eta-\eta_0)] \\
&= \mathbb{E}_{P_N}[\eta-\eta_0] - \frac{1}{\alpha}\mathbb{E}_{P_N}[\mathbb{E}_{P_N}[h_0 \mid Z](\eta-\eta_0)] \\
&= \mathbb{E}_{P_N}[\eta-\eta_0] - \frac{1}{\alpha}\mathbb{E}_{P_N}[\alpha(\eta-\eta_0)] = 0
\end{aligned}$$

The second line follows from the definitions of $h_0 = [\mu_0 \geq \eta_0]$ and $\mu_0 = \mathbb{E}[\ell \mid W, Z]$. The final line follows from the constraint that $\mathbb{E}_{P_N}[h_0 \mid Z] = \alpha$ almost everywhere.

Proof of Assumption 2 (e) This hold trivially since $J_0 = \mathbb{E}_{P_N}[\psi^a(O;\gamma_0)] = -1$.

Proof of Assumption 3 (a) This holds by construction of $\mathcal{T}_N$ and Assumption 1.

Proof of Assumption 3 (b) The bound on m'_N holds trivially as $\psi^a(O;\gamma) = -1$. To bound m_N on the event $\mathcal{E}_N$, we begin by decomposing it using the triangle inequality as

$$\begin{aligned}
\|\psi(O;\theta_0,\gamma)\|_{P_N,q} &= \left\|\frac{1}{\alpha}(\mu-\eta)_+ + \eta + \frac{1}{\alpha}h(\ell-\mu)\right\|_{P_N,q} && (27) \\
&\leq \frac{1}{\alpha}\|(\mu-\eta)_+ + \alpha\eta\|_{P_N,q} + \frac{1}{\alpha}\|h(\ell-\mu)\|_{P_N,q} && (28)
\end{aligned}$$

Since $0 \leq h \leq 1$, and by the triangle inequality and Assumption 1, we have

$$\begin{aligned}
\|h(\ell-\mu)\|_{P_N,q} &\leq \|\ell-\mu\|_{P_N,q} && (29) \\
&= \|\ell-\mu_0+\mu_0-\mu\|_{P_N,q} && (30) \\
&\leq \|\ell-\mu_0\|_{P_N,q} + \|\mu_0-\mu\|_{P_N,q} && (31) \\
&\leq 2C, && (32)
\end{aligned}$$

where the fourth line follows from Assumption 1 (a). Next, we can bound $\|(\mu-\eta)_+ + \alpha\eta\|_{P_N,q}$ as

$$\begin{aligned}
\|(\mu-\eta)_+ + \alpha\eta\|_{P_N,q} &\leq \|(\mu-\eta)_+\|_{P_N,q} + \|\alpha\eta\|_{P_N,q} && (33) \\
&\leq \|\mu-\eta\|_{P_N,q} + \|\alpha\eta\|_{P_N,q} && (34) \\
&\leq \|\mu_0\|_{P_N,q} + \|\mu_0-\mu\|_{P_N,q} + (1+\alpha)(\|\eta_0\|_{P_N,q} + \|\eta_0-\eta\|_{P_N,q}) && (35) \\
&\leq (4+2\alpha)C && (36)
\end{aligned}$$

where the fourth line follows from Jensen's inequality and Assumption 1 (a). Thus, $\|\psi(O;\theta_0,\gamma)\|_{P_N,q} < \infty$ and Assumption 3 (b) holds.

Proof of Assumption 3 (c) The bound on r_N is trivially satisfied. Further,

$$\begin{aligned}
\|\psi(O;\theta_0,\gamma)-\psi(O;\theta_0,\gamma_0)\|_{P_N,2} &= \frac{1}{\alpha}\|\alpha(\eta-\eta_0) + \ell(h-h_0) + h_0\eta_0 - h\eta\|_{P_N,2} && (37) \\
&\leq \frac{1}{\alpha}\left(\alpha\|\eta-\eta_0\|_{P_N,2} + \|\ell(h-h_0)\|_{P_N,2} + \|h_0\eta_0 - h\eta\|_{P_N,2}\right) && (38) \\
&\leq \frac{1}{\alpha}(\alpha\delta_N + C\delta_N + \|h_0\eta_0 - h\eta\|_{P_N,2}) && (39)
\end{aligned}$$

where the first line follows from the definition of ψ and h, the second line follows from the triangle inequality, and the third line follows from the Assumption 1. Then, to bound $\|h_0\eta_0 - h\eta\|_{P_N,2}$, first observe that $\|h_0 - h\|_{P_N,2} =$

$\|[h_0 = 1, h = 0] + [h_0 = 0, h = 1]\|_{P_N,2}$ where $[\cdot]$ is the Iverson bracket. Then, we have

$$\begin{aligned}\|h_0\eta_0 - h\eta\|_{P_N,2} &= \|[h_0 = 1, h = 0]\eta_0 - [h_0 = 0, h = 1]\eta + [h_0 = 1, h = 0](\eta_0 - \eta)\|_{P_N,2} && (40)\\ &\leq \|([h_0 = 1, h = 0] - [h_0 = 0, h = 1])\eta_0 - [h_0 = 0, h = 1](\eta - \eta_0) + [h_0 = 1, h = 0](\eta_0 - \eta)\|_{P_N,2} && (41)\\ &\leq \|([h_0 = 1, h = 0] - [h_0 = 0, h = 1])\eta_0\|_{P_N,2} + \|[h_0 = 0, h = 1](\eta - \eta_0)\|_{P_N,2} && (42)\\ &\quad + \|[h_0 = 1, h = 0](\eta_0 - \eta)\|_{P_N,2} && (43)\\ &\leq C\|h_0 - h\|_{P_N,2} + C\delta_N + \delta_N && (44)\\ &\lesssim \delta_N, && (45)\end{aligned}$$

where the third line follows from the triangle inequality and the fourth line follows from Lemma 14 of Jeong and Namkoong (2020) and Assumption 1 (d) - (f). Finally, we have

$$\|\psi(O; \theta_0, \gamma) - \psi(O; \theta_0, \gamma_0)\|_{P_N,2} \leq \frac{1}{\alpha}(\alpha + 3C + 1)\delta_N \tag{46}$$

and thus the bound on r'_N. As discussed above the bound on $\lambda'_N = \sup_{r\in(0,1),\gamma\in\mathcal{T}_N} \|\partial_r^2 \mathbb{E}_{P_N}[\psi(O; \theta_0, \gamma_0 + r(\gamma - \gamma_0))]\|$ does *not* hold because ψ is not twice differentiable.

Proof of Equation A.6 from Chernozhukov et al. (2018) We have now shown that all parts of Assumptions 2 and 3 hold except for assumptions involving the second derivative of ψ. The assumptions regarding the second derivative of ψ are used by Chernozhukov et al. (2018) to show that, for all $P_N \in \mathcal{P}_N$

$$R_{N,2} = \frac{1}{K}\sum_k \mathbb{E}_{n,k}[\psi(O; \theta_0, \hat{\gamma}_k)] - \frac{1}{N}\sum_{i=1}^{N} \psi(O_i; \theta_0, \gamma_0) = \mathcal{O}_{P_N}(\delta_N/\sqrt{N}). \tag{47}$$

We will show this using similar arguments to those in Step 3 of the proof of Theorem 3.1 in Chernozhukov et al. (2018), but without relying on the second derivative of ψ. As in Chernozhukov et al. (2018), because the number of cross-validation folds K is a fixed integer, we need only show that

$$\mathbb{E}_{n,k}[\psi(O; \theta_0, \hat{\gamma}_k)] - \frac{1}{n}\sum_{i=1}^{n} \psi(O_i; \theta_0, \gamma_0) = \mathcal{O}_{P_N}(\delta_N/\sqrt{N}) \tag{48}$$

where $n = N/K$. Following Chernozhukov et al. (2018) this quantity can be bounded as

$$\left\|\mathbb{E}_{n,k}[\psi(O; \theta_0, \hat{\gamma}_k)] - \frac{1}{n}\sum_{i=1}^{n} \psi(O_i; \theta_0, \gamma_0)\right\| \leq \frac{\mathcal{I}_{3,k} + \mathcal{I}_{4,k}}{\sqrt{n}}, \tag{49}$$

where

$$\begin{aligned}\mathcal{I}_{3,k} &= \|\mathbb{G}_{n,k}[\psi(O; \theta_0, \hat{\gamma}_k)] - \mathbb{G}_{n,k}[\psi(O; \theta_0, \gamma_0)]\| && (50)\\ \mathcal{I}_{4,k} &= \sqrt{n}\|\mathbb{E}_{P_N}[\psi(O; \theta_0, \hat{\gamma}_k)] \mid (O_i)_{i\in I_k} - \mathbb{E}_{P_N}[\psi(O; \theta_0, \gamma_0)]\|, && (51)\end{aligned}$$

and where

$$\mathbb{G}_{n,k}[\phi(O)] = \frac{1}{\sqrt{n}}\sum_{i\in I_k}\left(\phi(O_i) - \int \phi(w)dP_N\right). \tag{52}$$

Chernozhukov et al. (2018) showed that $\mathcal{I}_{3,k} = \mathcal{O}_{P_N}(r'_N)$ (using only assumptions satisfied by Assumption 1) and thus, what remains to be shown is that $\mathcal{I}_{4,k} \leq \delta_N/\sqrt{N}$ which we do drawing on proofs in Jeong and Namkoong (2020). Define

$$f_k(r) = \mathbb{E}_{P_N}[\psi(O; \theta_0, \gamma_0 + r(\hat{\gamma}_k - \gamma_0)) \mid (O_i)_{i\in I_k^c}] - \mathbb{E}_{P_N}[\psi(O; \theta_0, \gamma_0)]. \tag{53}$$

Note that $|f_k(1)|$ is the quantity that we want to bound and $f_k(0) = 0$. Then, using the mean value theorem, for some $r^* \in (0,1)$, we have $|f_k(1)| = |f_k(0) + f_k'(r^*)| = |f_k'(r^*)| \leq \sup_r |f_k'(r)|$. Define $\hat\mu_r = \mu_0 + r(\hat\mu_k - \mu_0)$ with analogous definitions for $\hat\eta_r$ and $\hat h_r$. Then, for arbitrary $r \in (0,1)$, we bound $|f_k'(r)|$ as

$$
\begin{aligned}
|f_k'(r)| &= |\partial_r \mathbb{E}_{P_N}[\psi(O;\theta_0,\gamma_0 + r(\hat\gamma_k - \gamma_0)) \mid (O_i)_{i\in I_k^c}]| && (54)\\
&\leq \frac{1}{\alpha}\left(|\mathbb{E}_{P_N}[([\hat\mu_r > \hat\eta_r] - h_0)(\hat\mu_k - \mu_0)]| + |\mathbb{E}_{P_N}[(\alpha - [\hat\mu_r > \hat\eta_r])(\hat\eta_k - \eta_0)]| + 2\left|\mathbb{E}_{P_N}[(\hat h_k - h_0)(\hat\mu_k - \mu_0)]\right|\right) && (55)\\
&\leq \frac{1}{\alpha}\left(|\mathbb{E}_{P_N}[(\alpha - [\hat\mu_r > \hat\eta_r])(\hat\eta_k - \eta_0)]| + 3\left|\mathbb{E}_{P_N}[(\hat h_k - h_0)(\hat\mu_k - \mu_0)]\right|\right) && (56)\\
&\leq \frac{1}{\alpha}\left(\|\hat h_k - h_0\|_{P_N,1}\|\hat\eta_k - \eta_0\|_{P_N,\infty} + 3\|\hat h_k - h_0\|_{P_N,1}\|\hat\mu_k - \mu_0\|_{P_N,\infty}\right) && (57)\\
&\leq \frac{1}{\alpha}\|\hat h_k - h_0\|_{P_N,1}\left(\|\hat\eta_k - \eta_0\|_{P_N,\infty} + 3\|\hat\mu_k - \mu_0\|_{P_N,\infty}\right) && (58)
\end{aligned}
$$

where the second line follows from the triangle inequality, the third line follows from the obsrevation that $|[\hat\mu_r < \hat\eta_r] - h_0| \leq |\hat h_k - h_0|$, and the fourth line follows from this observation and applications of Jensen's inequality followed by Hölder's inequality. By Lemmas 13 and 14 of Jeong and Namkoong (2020) and Assumption 1 (f), we have $\|\hat\eta - \eta_0\|_{P_N,\infty} = \mathcal{O}(\|\hat\mu_k - \mu_0\|_{P_N,\infty})$ and $\|\hat h_k - h_0\|_{P_N,1} = \mathcal{O}(\delta_N N^{-1/6})$. Further, by Assumption 1 we have $\|\hat\mu_k - \mu_0\|_{P_N,\infty} = \mathcal{O}(\delta_N N^{-1/3})$. Thus, we have $|f_k'(r)| = \mathcal{O}(\delta_N N^{-1/2})$ and our proof is concluded.

E Handling discrete W

Algorithm 2: Discrete Worst-case Sampler

```
Input: Model $\mathcal{M}$, Dataset $\mathcal{D} = \{(w_i, z_i, v_i)\}_{i=1}^n$, noise parameter $\epsilon$, and $K$ cross-validation folds
       $I_k \subset \{1,\ldots,n\}$ and $I_k^c = \{1,\ldots,n\} \setminus I_k$
for $k = 1,\ldots,K$ do
    Estimate $\hat\mu_k \approx \mu_0$ using data in $I_k^c$
    Estimate $\hat\eta_k \approx \eta_0$ according to Eq. 4 using $\hat\mu_k + u_i$ and data in $I_k^c$
    for $i \in I_k$ do
        Let $\hat\mu_i = \hat\mu_k(w_i, z_i)$
        Let $\hat\eta_i = \hat\eta_k(z_i)$
        Let $\hat h_i = [\hat\mu_i + u_i > \hat\eta_i]$
    end
end
Let $\hat R_{\alpha,\epsilon} = \frac{1}{K}\sum_k \frac{1}{|I_k|}\sum_{i\in I_k} \frac{1}{1-\alpha}(\hat\mu_i + u_i - \hat\eta_i)_+ + \hat\eta_i$
        $+ \frac{1}{1-\alpha}\hat h_i(\ell(y_i, \mathcal{M}(x_i)) - \hat\mu_i)$
Result: $\hat R_\alpha$
```

When W contains only discrete variables, Assumption 1 (f) no longer holds. In such cases, we can retain the desirable theoretical properties of Theorem 1 at the cost of an arbitrarily small, user-controlled amount of bias, using a simple augmentation to the Worst-case Sampler in Algorithm 1 of the main paper. This augmentation works by adding a small amount of user-controlled noise to the μ_0 thereby smoothing the conditional distribution of μ_0 given Z. To derive this augmentation, first let $U \sim Unif(0,\epsilon)$ be a uniform random variable with support on $[0,\epsilon]$ such that $U \perp\!\!\!\perp \{W, Z, V\}$. Then, we can choose h to maximize the expected loss plus this extra noise term as

$$
\begin{aligned}
R_{\alpha,\epsilon,0} = &\sup_{h:[0,\epsilon]\times\mathcal{W}\times\mathcal{Z}\to[0,1]} \frac{1}{1-\alpha}\mathbb{E}_P\left[h(U,W,Z)\mu_{\epsilon,0}(U,W,Z)\right] && (59)\\
&\text{s.t.} \quad \mathbb{E}_P[h(U,W,Z) \mid Z] = 1-\alpha \quad a.e., && (60)
\end{aligned}
$$

where $\mu_{\epsilon,0}(U,W,Z) = \mu_0(W,Z) + U$. The corresponding change to the estimation algorithm is shown in Algorithm

2. This algorithm returns a consistent estimate for $R_{\alpha,\epsilon,0}$ as $\mu_{\epsilon,0}$ satisfies the conditions of Assumption 1 (f). In the following proposition, we show that the difference between $R_{\alpha,0}$ and $R_{\alpha,\epsilon,0}$ is bounded by ϵ.

Proposition 5. $|R_{\alpha,\epsilon,0} - R_{\alpha,0}| \leq \epsilon$

Proof. First, since U has support in the non-negatives, we have $R_{\alpha,\epsilon,0} \geq R_{\alpha,0}$. Next, let $S = \{\mathcal{W} \times \mathcal{Z} \to [0,1] : \mathbb{E}_P[h \mid Z] = 1 - \alpha \;\; a.e.\}$ and $\tilde{S} = \{[0,\epsilon] \times \mathcal{W} \times \mathcal{Z} \to [0,1] : \mathbb{E}_P[h \mid Z] = 1 - \alpha \;\; a.e.\}$. Then,

$$\begin{aligned}
R_{\alpha,\epsilon,0} &= \max_{\tilde{h} \in \tilde{S}} \frac{1}{1-\alpha} \mathbb{E}_P[\tilde{h}(\mu_0 + U)] && (61)\\
&\leq \max_{\tilde{h} \in \tilde{S}} \frac{1}{1-\alpha} \mathbb{E}_P[\tilde{h}(\mu_0 + \epsilon)] && (62)\\
&= \left(\max_{\tilde{h} \in \tilde{S}} \frac{1}{1-\alpha} \mathbb{E}_P[\tilde{h}\mu_0] \right) + \epsilon && (63)\\
&= \left(\max_{h \in S} \frac{1}{1-\alpha} \mathbb{E}_P[h\mu_0] \right) + \epsilon && (64)\\
&= R_{\alpha,0} + \epsilon. && (65)
\end{aligned}$$

□

F Experimental Details

F.1 Dataset

We loosely follow the setup of Giannini et al. (2019) in deriving the dataset for training sepsis diagnosis models. The dataset contains electronic health record data collected over four years at our institution's hospital (Hospital A). The dataset consists of 278,947 emergency department patient encounters. The prevalence of the target disease, sepsis (S), is 2.1%. 17 features pertaining to vital signs (V) (heart rate, respiratory rate, temperature), lab tests (L) (white blood cell count [wbc], lactate), and demographics (D) (age, gender) were extracted. For encounters that resulted in sepsis (i.e., positive encounters), physiologic data available up until sepsis onset time was used. For non-sepsis encounters, all data available until discharge was used. For each of the time-series physiologic features (V and L), min, max, and median summary features were derived. Unlike vitals, lab measurements are not always ordered (O) and are subject to missingness (lactate 89%, wbc 27%). To model lab missingness, missingness indicators (O) for the lab features were added. The evaluation dataset was created using a held-out sample of 10,000 patients. The remaining data was used to train the two models. As per its definition, qSOFA was computed from respiratory rate, systolic blood pressure, and glasgow coma score (gcs) (gcs and blood pressure were separately extracted for these patients) (Singer et al., 2016). Using existing standards, we remapped gcs to the Alert, Voice, Pain, Unresponsive (AVPU) score which is required to compute qSOFA (Gardner-Thorpe et al., 2006).

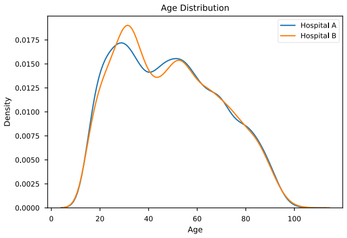

Figure 5: Age distributions at the two hospitals are very similar.

In Section 3.2.3 we use data from another Hospital B (within the same network as Hospital A used in the prior experiments). This dataset contains 96,574 patient encounters and has a sepsis prevalence of 2.8% (vs 2.1% at Hospital A). Turning to demographics, the population at Hospital A is 42% female while at Hospital B it is 39% female. Finally, Kernel Density Estimates (Fig 5) of the age distributions at the two hospitals were very similar. The missingness rates for lab orders were 28% wbc missingness (unchanged from Hospital A) and 77% lactate missingness (12% decrease in missingness). Thus, there is a sizeable increase in (lactate) test ordering patterns from Hospital A to Hospital B.

F.2 Models

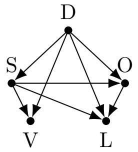

Figure 6: Posited DAG of the data generating process used to train the robust model. The robust model was trained to be stable to shifts in the policy for ordering lab tests (O).

The **classical** model was a Random Forest Classifier trained using the implementation in scikit-learn (Pedregosa et al., 2011). The hyperparameters were tuned via grid search 5-fold cross-validation (CV) on the training dataset. The resulting tuned parameters were: number of trees 1000, min samples in a leaf 1, max depth 15, and min samples per split 2.

The **robust** model was an implementation of the surgery estimator (Subbaswamy et al., 2019), a causal method for training models which make predictions that are stable to pre-specified shifts. Using the DAG in Fig 6, lab test ordering patterns are defined by the distribution $P(O|S, D)$. As opposed to the classical model, which models $P(S|V, D, O, L)$, the robust model models the *interventional* distribution $P(S|V, D, do(O), L)$ which considers a hypothetical intervention on test ordering patterns. The robust model was fit by inverse probability weighting (IPW). First, we fit a logistic regression model of $P(O|S, D)$ using the training data. Then, a Random Forest model with the full feature set was trained using sample weights $\frac{1}{P(o_i|s_i,d_i)}$. These weights create a training dataset in which lab orders are approximately randomized w.r.t. S and D. The tuned hyperparameters from the classical model were used as the hyperparameters for the robust model.

F.3 Estimating Worst-Case Risk

Estimating the Worst-Case risk using Algorithm 1 requires specifying the variable sets W and Z and fitting the models for the nuisance parameters μ_0 (the conditional expected loss) and η_0 (the conditional quantile of μ_0). To measure model robustness to changes in test ordering patterns we define $W = \{O\}$ and $Z = \{S, D\}$ (so that $P(W|Z)$ corresponds to test ordering policies). Since classification accuracy was the performance metric of interest, we chose $0 - 1$ loss as the loss function. Within Algorithm 1, 10-fold CV was used (i.e., $K = 10$).

To fit the conditional expected loss $\hat{\mu}_k$, we used the scikit-learn Kernel Ridge Regression implementation with the RBF kernel which minimizes ℓ_2 regularized mean squared error (MSE). Because W contained all discrete variables, we applied Algorithm 2 with noise $U \sim Unif(0, 1 \times 10^{-5})$. The bandwidth and regularization parameters were tuned using a nested 5-fold CV on each estimation fold k to which Algorithm 2 was applied. As $\hat{\mu}_k$ does not depend on α, it was not refit for different α values.

To model the conditional quantile function $\hat{\eta}_k$, we used ℓ_2 regularized quantile regression with a b-spline basis expansion. We used a quantile b-spline basis expansion for Age and added an interaction term between S and all other variables in Z (Age expansion and Gender). The regularization constant λ_k was chosen separately for each estimation fold k using a nested 5-fold cross-validation and grid search to produce the lowest mean absolute deviation.