Interpolation and Regularization for Causal Learning Supplementary Materials

A Proof of Proposition 2.1

For the statistical risk, we first need one standard result about the distribution of a multivariate normal random variable conditioned on an affine function:

Lemma A.1. *Consider a multivariate normal random variable $X \sim \mathcal{N}(\mu, \Sigma)$ with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$. Then for any $A \in \mathbb{R}^{k\times d}$, $b \in \mathbb{R}^k$, and $y \in \mathbb{R}^k$ it holds*

$$X|(AX+b) = y \sim \mathcal{N}(\mu + \Sigma A^T(A\Sigma A^T)^+(y - A\mu - b), \Sigma - \Sigma A^T(A\Sigma A^T)^+ A\Sigma)\,.$$

In particular, if X is a standard normal random variable ($\Sigma = I_d$, $\mu = 0$) and $b = 0$, it is

$$X|AX = y \sim \mathcal{N}(A^T(AA^T)^+ y, I_d - A^T(AA^T)^+ A)$$

Proof. Let $Y = AX + b$. The joint distribution of X and Y is again a multivariate normal, because it can be written as an affine transformation of X:

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \underbrace{\begin{pmatrix} I_d \\ A \end{pmatrix}}_{=:A' \in \mathbb{R}^{(d+k)\times d}} X + \underbrace{\begin{pmatrix} 0_d \\ b \end{pmatrix}}_{=:b' \in \mathbb{R}^{d+k}} = A'X + b'\,,$$

which implies that

$$\begin{pmatrix} X \\ Y \end{pmatrix} = A'X + b' \sim \mathcal{N}(A'\mu + b', A'\Sigma(A')^T) = \mathcal{N}(\begin{pmatrix} \mu \\ A\mu + b \end{pmatrix}, \begin{pmatrix} \Sigma & \Sigma A^T \\ A\Sigma & A\Sigma A^T \end{pmatrix})\,.$$

The claim then follows from the standard formula for conditionals of multivariate normal distributions, which states that if $\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix})$, then

$$Z_1|Z_2 = z \sim \mathcal{N}(\mu_1 + \Sigma_{1,2}\Sigma_{2,2}^+(z - \mu_2), \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^+\Sigma_{2,1})\,.$$

□

Proposition 2.1 (Causal and Statistical Risk). *For any $\hat{\beta} \in \mathbb{R}^d$, the risks defined in Eq. (2) satisfy*

$$R^C(\hat{\beta}) = \|\hat{\beta} - \beta\|_\Sigma^2 + \tilde{\sigma}^2 + \|\Gamma\|_\Sigma^2 \quad \textit{and} \quad R^S(\hat{\beta}) = \|\hat{\beta} - \tilde{\beta}\|_\Sigma^2 + \tilde{\sigma}^2\,.$$

Proof. The key step for this proof is to characterize the distribution of y under the do-intervention $y|do(x)$ and the usual observational conditional $y|x$. We start with the proof for the causal risk under the do-intervention. Intervening on x under the causal model given by Eq. (1) corresponds to removing all arrows to x, which corresponds to the structural equations

$$z \sim \mathcal{N}(0, I_l)\,, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)\,, \quad y = x^T\beta + z^T\alpha + \varepsilon\,.$$

In this model, z acts as additional independent noise on y through $z^T\alpha \sim \mathcal{N}(0, \|\alpha\|^2)$, which implies that $y|do(x) \sim \mathcal{N}(x^T\beta, \|\alpha\|^2 + \sigma^2)$. Equivalently, $y|do(x)$ has the same distribution as $x^T\beta + \varepsilon'$ with $\varepsilon' \sim \mathcal{N}(0, \tilde{\sigma}^2 + \omega^2)$ because $\|\alpha\|^2 + \sigma^2 = \tilde{\sigma}^2 + \omega^2$. This lets us compute the causal risk of a linear predictor $\hat{\beta} \in \mathbb{R}^d$ as

$$\begin{aligned}
R^C(\hat{\beta}) &= \mathbb{E}_x \mathbb{E}_{y_0|do(x)} \left(x^T\hat{\beta} - y\right)^2 \\
&= \mathbb{E}_x \mathbb{E}_{\varepsilon'} \left(x^T\left(\hat{\beta} - \beta\right) - \varepsilon'\right)^2 \\
&= \mathbb{E}_x \left(x^T\left(\hat{\beta} - \beta\right)\right)^2 - 2\mathbb{E}_x\left[x^T\left(\hat{\beta} - \beta\right)\underbrace{\mathbb{E}_{\varepsilon'}\varepsilon'}_{=0}\right] + \mathbb{E}_x\mathbb{E}_{\varepsilon'}\left(\varepsilon'\right)^2 \\
&= \left\|\hat{\beta} - \beta\right\|_\Sigma^2 + \tilde{\sigma}^2 + \omega^2\,, && (\mathbb{E}_x xx^T = \Sigma)
\end{aligned}$$

which proves the claim for the causal risk. The proof for the statistical risk is analogous once we have characterized the conditional distribution $y|x$ under the causal model. Recall that $\Sigma = MM^T$, $\Gamma = M^{+T}\alpha$, and $\omega^2 = \|\Gamma\|_\Sigma^2$. We first observe that $x = Mz$ is a linear map of the Gaussian distribution $z \sim \mathcal{N}(0, I_l)$, for which Lemma A.1 yields

$$\begin{aligned} z|x &\sim \mathcal{N}(M^T(MM^T)^+x, I - M^T(MM^T)^+M) \\ \text{and therefore} \quad z^T\alpha|x &\sim \mathcal{N}(\alpha^T M^T(MM^T)^+x, \|\alpha\|^2 - \alpha^T M^T(MM^T)^+M\alpha) \\ &= \mathcal{N}(x^T\Gamma, \|\alpha\|^2 - \|\Gamma\|_\Sigma^2)\,, \end{aligned}$$

where the last equality used the identity

$$\alpha^T M^T(MM^T)^+M\alpha = \alpha^T M^+MM^TM^{+T}\alpha = \Gamma^T\Sigma\Gamma = \|\Gamma\|_\Sigma^2 = \omega^2\,.$$

Since $y = x^T\beta + z^T\alpha + \varepsilon$, it follows that

$$y|x \sim \mathcal{N}(x^T(\beta+\Gamma), \sigma^2 + \|\alpha\|^2 - \omega^2) = \mathcal{N}(x^T\tilde{\beta}, \tilde{\sigma}^2)\,,$$

which concludes the proof. □

B Proofs for Section 3.1

The bias-variance decomposition of the causal risk is based on the following general lemma:

Lemma B.1 (Bias-Variance Decomposition for General Norm). *Consider a random variable Z on $\mathbb{R}^d$, a constant $c \in \mathbb{R}^d$, and the general norm $\|x\|_A^2 = x^TAx$ for some positive-definite $A \in \mathbb{R}^{d\times d}$. Then we have the decomposition*

$$\mathbb{E}_Z \|Z - c\|_A^2 = \|\mathbb{E}Z - c\|_A^2 + \mathbb{E}_Z \|Z - \mathbb{E}_Z Z\|_A^2\,.$$

An alternative form of the variance term is given by $\mathbb{E}_Z \|Z - \mathbb{E}_Z Z\|_A^2 = \operatorname{Tr}[\operatorname{Cov} Z \cdot A]$.

Proof. Let $\mathbb{E} \coloneqq \mathbb{E}_Z$ and $\mu \coloneqq \mathbb{E}Z$. It is

$$\begin{aligned} \mathbb{E}\|Z - c\|_A^2 &= \mathbb{E}\|(Z-\mu) + (\mu - c)\|_A^2 \\ &= \mathbb{E}\|Z-\mu\|_A^2 + \mathbb{E}\|\mu - c\|_A^2 + 2\underbrace{\mathbb{E}(Z-\mu)^T}_{=0}A(\mu - c) \\ &= \mathbb{E}\|Z-\mu\|_A^2 + \mathbb{E}\|\mu - c\|_A^2\,, \end{aligned}$$

which proves the first part of the statement. For the second part, let $\Sigma_Z \coloneqq \mathbb{E}ZZ^T$ and denote the Hadamard product between matrices $A, B \in \mathbb{R}^{d\times d}$ by $(A \odot B)_{i,j} = A_{i,j}B_{i,j}$. It is

$$\begin{aligned} \mathbb{E}\|Z-\mu\|_A^2 &= \mathbb{E}Z^TAZ - 2\mathbb{E}Z^TA\mu + \mu^TA\mu \\ &= \sum_{i,j=1}^n (\Sigma_Z \odot A)_{i,j} - \mu^TA\mu \\ &= \operatorname{Tr}[\Sigma_Z \cdot A] - \mu^TA\mu && (\textstyle\sum_{i,j=1}^n (A\odot B)_{i,j} = \operatorname{Tr}(A\cdot B)) \\ &= \operatorname{Tr}[\Sigma_Z \cdot A] - \operatorname{Tr}\left[A\mu\mu^T\right] && (\operatorname{Tr}(ba^T) = a^Tb) \\ &= \operatorname{Tr}\left[(\Sigma_Z - \mu\mu^T)\cdot A\right] && (\operatorname{Tr}(B) = \operatorname{Tr}(B^T) \text{ and linearity of trace}) \\ &= \operatorname{Tr}[\operatorname{Cov} Z \cdot A]\,. && (\operatorname{Cov} Z = \mathbb{E}ZZ^T - \mu\mu^T) \end{aligned}$$

□

Proposition B.2 (Causal Bias-Variance Decomposition for the Ridge Estimator). *For any $\lambda > 0$, the expectation over the causal risk of the ridge regression estimator $\hat{\beta}_\lambda$ conditioned on X admits the bias-variance decomposition*

$$R_X^C(\hat{\beta}_\lambda) = \underbrace{\|\mathbb{E}_{Y|X}\hat{\beta}_\lambda - \beta\|_\Sigma^2}_{=:B_X^C(\hat{\beta}_\lambda)} + \underbrace{\mathbb{E}_{Y|X}\|\hat{\beta}_\lambda - \mathbb{E}_{Y|X}\hat{\beta}_\lambda\|_\Sigma^2}_{=:V_X^C(\hat{\beta}_\lambda)} + \tilde{\sigma}^2 + \|\Gamma\|_\Sigma^2\,, \tag{9}$$

where $B_X^C(\hat{\beta}_\lambda) = \|(I - (\hat{\Sigma}+\lambda I_d)\hat{\Sigma})\tilde{\beta} - \Gamma\|_\Sigma^2$ and $V_X^C(\hat{\beta}_\lambda) = \frac{\tilde{\sigma}^2}{n}\operatorname{Tr}[\hat{\Sigma}(\hat{\Sigma}+\lambda I_d)^{-2}\Sigma]$. The empirical covariance matrix of X is denoted by $\hat{\Sigma} \coloneqq X^TX/n$.

Proof. Recall that $R_X^C(\hat{\beta}_\lambda) = \mathbb{E}_{Y|X} \left\| \hat{\beta}_\lambda - \beta \right\|_\Sigma^2$. The first part of the statement follows directly from Lemma B.1 with $\hat{\beta}_\lambda$ as a random variable in $Y|X$ and β. The remainder of the proof consists of computing expectation and covariance of the ridge regression solution $\hat{\beta}_\lambda = \hat{\beta}_\lambda(X, Y)$ under the distribution $Y|X$. The samples (X, Y) are drawn from the observational distribution of the causal model defined in Eq. (1). As shown in the proof of Proposition 2.1, the corresponding conditional distribution is $y|x \sim \mathcal{N}(x^T\tilde{\beta}, \tilde{\sigma}^2)$. Since (X, Y) consist of independent draws, this implies $Y|X \sim \mathcal{N}(X\tilde{\beta}, \tilde{\sigma}^2 I_n)$. Together with $\hat{\beta}_\lambda = (X^TX + n\lambda I)^{-1}X^TY$ this yields

$$\begin{aligned}\hat{\beta}_\lambda|X &\sim \mathcal{N}((X^TX + n\lambda I)^{-1}X^TX\tilde{\beta}, (X^TX + n\lambda I)^{-1}X^T\tilde{\sigma}^2 I_n X(X^TX + n\lambda I)^{-1}) \\ &= \mathcal{N}(\left(\hat{\Sigma} + \lambda I_d\right)^{-1}\hat{\Sigma}\tilde{\beta}, \frac{\tilde{\sigma}^2}{n}\left(\hat{\Sigma} + \lambda I_d\right)^{-1}\hat{\Sigma}\left(\hat{\Sigma} + \lambda I_d\right)^{-1}).\end{aligned}$$

The characterizations of $B_X^C(\hat{\beta}_\lambda)$ and $V_X^C(\hat{\beta}_\lambda)$ then simply follow from plugging in expectation and covariance of $\hat{\beta}_\lambda$:

$$\begin{aligned}B_X^C(\hat{\beta}_\lambda) = \left\| \mathbb{E}_{Y|X}\hat{\beta}_\lambda - \beta \right\|_\Sigma^2 = \left\| \left(\hat{\Sigma} + \lambda I_d\right)^{-1}\hat{\Sigma}\tilde{\beta} - \beta \right\|_\Sigma^2 &= \|(I - \Pi_\lambda)(\beta + \Gamma) - \beta\|_\Sigma^2 \\ &= \|\Pi_\lambda\beta - (I - \Pi_\lambda)\Gamma\|_\Sigma^2\end{aligned}$$

and, using the alternate form of the variance term from Lemma B.1,

$$\begin{aligned}V_X^C(\hat{\beta}_\lambda) = \operatorname{Tr}\left[\operatorname{Cov}_{Y|X}\hat{\beta}_\lambda \cdot \Sigma\right] &= \operatorname{Tr}\left[\frac{\tilde{\sigma}^2}{n}\left(\hat{\Sigma} + \lambda I_d\right)^{-1}\hat{\Sigma}\left(\hat{\Sigma} + \lambda I_d\right)^{-1} \cdot \Sigma\right] \\ &= \frac{\tilde{\sigma}^2}{n}\operatorname{Tr}\left[\hat{\Sigma}\left(\hat{\Sigma} + \lambda I_d\right)^{-2}\Sigma\right],\end{aligned}$$

where the last equality used that $\left(\hat{\Sigma} + \lambda I_d\right)^{-1}$ commutes with $\hat{\Sigma}$. □

Theorem 2 (Limiting Causal Bias-Variance Decomposition for the Ridge Estimator). *Let $\|\beta\|^2 = r^2$, $\|\Gamma\|^2 = \omega^2$, $\langle\Gamma, \beta\rangle = \eta$, and $\sigma_\varepsilon^2 = \tilde{\sigma}^2$. Then as $n, d \to \infty$ such that $d/n \to \gamma \in (0, \infty)$, it holds almost surely in X for every $\lambda > 0$ that*

$$B_X^C(\hat{\beta}_\lambda) \to \mathcal{B}_\lambda^C := \omega^2 + \tilde{r}^2\lambda^2 m'(-\lambda) - 2(\omega^2 + \eta)\lambda m(-\lambda) \quad \textit{and} \tag{7}$$

$$V_X^C(\hat{\beta}_\lambda) \to \mathcal{V}_\lambda^C := \tilde{\sigma}^2\gamma(m(-\lambda) - \lambda m'(-\lambda)), \tag{8}$$

where $m(\lambda) = ((1 - \gamma - \lambda) - \sqrt{(1 - \gamma - \lambda)^2 - 4\gamma\lambda})/(2\gamma\lambda)$ and $\tilde{r}^2 = r^2 + \omega^2 + 2\eta$. Therefore $R_X^C(\hat{\beta}_\lambda) \to \mathcal{R}_\lambda^C := \mathcal{B}_\lambda^C + \mathcal{V}_\lambda^C + \tilde{\sigma}^2 + \omega^2$. The corresponding limiting quantities for the min-norm interpolator can be obtained by taking the limit $\lambda \to 0^+$ in equations (7) and (8), which yields

$$B_X^C(\hat{\beta}_0) \to \mathcal{B}_0^C = \begin{cases}\omega^2, & \gamma < 1 \\ \omega^2 + (r^2 - \omega^2)(1 - \frac{1}{\gamma}), & \gamma > 1\end{cases}, \quad V_X^C(\hat{\beta}_0) \to \mathcal{V}_0^C = \begin{cases}\tilde{\sigma}^2\frac{\gamma}{1-\gamma}, & \gamma < 1 \\ \tilde{\sigma}^2\frac{1}{\gamma-1}, & \gamma > 1\end{cases}.$$

Therefore $R_X^C(\hat{\beta}_0) \to \mathcal{R}_0^C = \mathcal{B}_0^C + \mathcal{V}_0^C + \tilde{\sigma}^2 + \omega^2$.

Proof. From Proposition B.2, the causal risk $R_X^C(\hat{\beta}_\lambda)$ can be decomposed as a sum of the causal bias $B_X^C(\hat{\beta}_\lambda)$, and causal variance $V_X^C(\hat{\beta}_\lambda)$. In what follows, we derive the limiting expressions for $B_X^C(\hat{\beta}_\lambda)$ and $V_X^C(\hat{\beta}_\lambda)$ to obtain the limiting causal risk for any $\gamma \in (0, \infty)$.

Limiting expressions for causal bias

$$\begin{aligned}B_X^C(\hat{\beta}_\lambda) &= \|\beta - \mathbb{E}_{|X}\hat{\beta}_\lambda\|_\Sigma^2 = \|\Pi_\lambda\beta - (I - \Pi_\lambda)\Gamma\|^2 \qquad (\Sigma = I) \\ &= \|\Pi_\lambda(\beta + \Gamma) - \Gamma\|^2 \\ &= \|\Pi_\lambda\tilde{\beta}\|^2 + \|\Gamma\|^2 - 2\langle\Gamma, \Pi_\lambda(\tilde{\beta})\rangle\end{aligned}$$

First, let us consider the sequence of functions given by

$$\begin{aligned}\|\Pi_\lambda \tilde{\beta}\|^2 &= \|(I - (\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma})\tilde{\beta}\|^2 \\ &= \left\|\lambda((\hat{\Sigma} + \lambda I)^{-1})\tilde{\beta}\right\|^2 \qquad \text{(Add and subtract } \lambda I\text{)} \\ &= \lambda^2 \tilde{\beta}^T(\hat{\Sigma} + \lambda I)^{-2}\tilde{\beta}^T \\ &= \lambda^2 \operatorname{Tr}\left[\tilde{\beta}\tilde{\beta}^T(\hat{\Sigma} + \lambda I)^{-2}\right]\end{aligned}$$

To derive the limiting expression for this sequence, we utilize the "derivative trick". This technique has been employed in a similar context in Dobriban et al. (2018). More generally similar terms (although not identical) often also arise in the analysis of the statistical of the ridge regression estimator and therefore one can find similar approaches to deriving the limiting expressions for such terms in the statistical analysis for ridge regression (for example, Hastie et al. (2022), Dobriban et al. (2018), and Dicker (2016)). Here, we include a self-contained proof of the result.

The idea relies on an application of Vitali's convergence theorem (see Bai et al. (2010, Lemma 2.14)) to obtain the limit of derivatives of a sequence of functions analytic on some domain $D \subset \mathbb{C}$ by the derivative of the limit of the sequence of functions. Observe that

$$\operatorname{Tr}\left[(\beta + \Gamma)(\beta + \Gamma)^T(\hat{\Sigma} + \lambda I)^{-2}\right] = \frac{\partial}{\partial \lambda} - \operatorname{Tr}\left[(\beta + \Gamma)(\beta + \Gamma)^T(\hat{\Sigma} + \lambda I)^{-1}\right]$$

By recognizing the quantity $(\hat{\Sigma} + \lambda I)^{-1}$ as the resolvent $Q(-\lambda)$, we can invoke the Marchenko-Pastur Theorem due to Marčenko et al. (1967) and Silverstein (1995) which states that the Stieltjes transform of the empirical distribution $\hat{m(z)}$ of eigenvalues of $\hat{\Sigma}$ converges almost surely to the Stieltjes transform $m(z)$ of the empirical spectral distribution given by the Marchenko-Pastur Law F for any $z \in \mathbb{C}/\mathbb{R}^+$. [2] That is, we have for all $\lambda > 0$,

$$\frac{1}{d}\operatorname{Tr}\left[(\hat{\Sigma} + \lambda I)^{-1}\right] \xrightarrow{a.s} m_F(-\lambda)$$

Rubio et al. (2011, Theorem 1) provide a generalization of this result which includes providing almost sure convergence of quadratic forms of resolvents of the form $u^T(\hat{\Sigma} - zI)v$ for sequences of vectors $\{u\}, \{v\}$ such that their outer product uv^T has a bounded trace norm for any $z \in \mathbb{C}/\mathbb{R}^+$. By this result, it is easy to verify that for any $\lambda > 0$,

$$\operatorname{Tr}\left[\tilde{\beta}\tilde{\beta}^T(\hat{\Sigma} + \lambda I)^{-1}\right] \xrightarrow{a.s} m_F(-\lambda)\tilde{r}^2$$

It is easy to see that the sequence of functions $\left\{f_d(\lambda) = \operatorname{Tr}\left[\tilde{\beta}\tilde{\beta}^T(\hat{\Sigma} + \lambda I)^{-1}\right]\right\}$ is analytic for $\lambda > 0$. Furthermore, for any $\lambda > 0$, the absolute value of the sequence of functions $\{f_d(\lambda)\}$ is uniformly bounded in d since

$$|f_d(\lambda)| \leq \operatorname{Tr}[\tilde{\beta}\tilde{\beta}^T]\frac{1}{\lambda} \leq \frac{\tilde{r}^2}{\lambda}$$

Therefore, by Vitali's convergence theorem, it holds (almost surely) that for every $\lambda > 0$, the derivatives of the sequence of functions $f_1, f_2, \cdots$ converges to the derivative of their limit and we have

$$\lambda^2 \operatorname{Tr}\left[\tilde{\beta}\tilde{\beta}^T(\hat{\Sigma} + \lambda I)^{-2}\right] \to \lambda^2\tilde{r}^2 m_F'(-\lambda),$$

where $m_F'(-\lambda)$ denotes the derivative of the Stieltjes transform of the Marchenko-Pastur Law evaluated at $-\lambda$.

To obtain the limiting function of the sequence $\langle \Gamma, \Pi_\lambda \tilde{\beta}\rangle$, observe that

$$\langle \Gamma, \Pi_\lambda \tilde{\beta}\rangle = \lambda\langle \Gamma, (\hat{\Sigma} + \lambda I)^{-1}\tilde{\beta}\rangle = \lambda \operatorname{Tr}[\tilde{\beta}\Gamma^T(\hat{\Sigma} + \lambda I)^{-1}] \xrightarrow{a.s} \lambda(\omega^2 + \eta)m_F(-\lambda),$$

[2] While the convergence result in Silverstein (1995) is stated for $z \in \mathbb{C}^+ = \{z = u + iv \in \mathbb{C} | Im(z) = v > 0\}$, it can be extended to $z \in \mathbb{C}/\mathbb{R}^+$ following standard arguments for convergence of sequences of analytic functions (see Hachem et al. (2007, Proposition 2.2)) via Vitali's convergence theorem or Montel's theorem. See Rubio et al. (2011, Proof of Theorem 1, Page 14) for an example of this argument.

where the limit is obtained by invoking Rubio et al. (2011, Theorem 1).

Therefore, we have that as $n, d \to \infty$ and $d/n \to \gamma$,

$$B_X^C(\hat{\beta}_\lambda) \xrightarrow{a.s} \omega^2 + \tilde{r}^2\lambda^2 m_F'(-\lambda) - 2(\omega^2 + \eta)\lambda m_F(-\lambda).$$

Limiting expressions for causal variance.

By recalling the expression for variance we have

$$\begin{aligned} V_X^C(\hat{\beta}_\lambda) &= \frac{\tilde{\sigma}^2}{n} \operatorname{Tr}\left[\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-2}\right] \\ &= \frac{\tilde{\sigma}^2}{n} \operatorname{Tr}\left[(\hat{\Sigma} + \lambda I - \lambda I)(\hat{\Sigma} + \lambda I)^{-2}\right] \\ &= \tilde{\sigma}^2 \frac{d}{n} \operatorname{Tr}\left[\frac{1}{d}(\hat{\Sigma} + \lambda I)^{-1} - \frac{1}{d}\lambda(\hat{\Sigma} + \lambda I)^{-2}\right] \end{aligned}$$

By Marchenko-Pastur Theorem (Marčenko et al., 1967; Silverstein, 1995), we already know that for any $\lambda > 0$

$$\operatorname{Tr}\left[\frac{1}{d}(\hat{\Sigma} + \lambda I)^{-1}\right] \to m_F(-\lambda)$$

Further, recognizing that

$$-\operatorname{Tr}\left[\frac{1}{d}(\hat{\Sigma} + \lambda I)^{-2}\right] = \frac{\partial}{\partial \lambda} \operatorname{Tr}\left[\frac{1}{d}(\hat{\Sigma} + \lambda I)^{-1}\right]$$

and that $|\operatorname{Tr}[\frac{1}{d}(\hat{\Sigma} + \lambda I)^{-1}]| \leq \frac{1}{\lambda}$, we can again invoke Vitali's convergence theorem to obtain the limit of the derivatives by taking the derivative of the limit to obtain

$$V_X^C(\hat{\beta}_\lambda) = \tilde{\sigma}^2\gamma(m_F(-\lambda) - \lambda m_F'(-\lambda)).$$

Marchenko-Pastur Law admits an explicit form under our model assumptions (see for example, (Bai et al., 2010, Page 52)) for any $z \in \mathbb{C}^+$ (which can be extended by analytic continuity arguments for any $z \in \mathbb{C}/\mathbb{R}^+$) and is given by

$$m_F(z) = \frac{1 - \gamma - z - \sqrt{(1 - \gamma - z)^2 - 4\gamma z}}{2\gamma z}.$$

Following arguments similar to Dobriban et al. (2018) and Hastie et al. (2022) for exchanging the limits $n, d \to \infty$ and $\lambda \to 0^+$, we can derive the limiting expressions for the causal bias and variance of the min-norm estimator.

□

C Asymptotics for the Statistical Risk

The following theorems describes the limiting expressions for the statistical risk analogue to the causal results from Theorem 3.1.

Theorem C.1 (Limiting Statistical Bias-Variance Decompositions). *Let $\hat{\beta}_0$ be the min-norm interpolator. Then as $n, d \to \infty$ such that $d/n \to \gamma \in (0, \infty)$, it holds almost surely in X that*

$$B_X^S(\hat{\beta}_0) \to \mathcal{B}_0^S = \begin{cases} 0, & \gamma < 1 \\ \tilde{r}^2(1 - \frac{1}{\gamma}), & \gamma > 1 \end{cases}, \quad V_X^S(\hat{\beta}_0) \to \mathcal{V}_0^S = \begin{cases} \tilde{\sigma}^2 \frac{\gamma}{1-\gamma}, & \gamma < 1 \\ \tilde{\sigma}^2 \frac{1}{\gamma - 1}, & \gamma > 1 \end{cases} \tag{10}$$

and therefore, $R_X^S(\hat{\beta}_0) \to \mathcal{R}_0^S = \mathcal{B}_0^S + \mathcal{V}_0^S + \tilde{\sigma}^2$.

For $\lambda > 0$ and the corresponding ridge regression estimator $\hat{\beta}_\lambda$, it holds almost surely in X that

$$B_X^S(\hat{\beta}_\lambda) \to \mathcal{B}_\lambda^S = \tilde{r}^2\lambda^2 m'(-\lambda), \quad V_X^S(\hat{\beta}_\lambda) \to \mathcal{V}_\lambda^S = \tilde{\sigma}^2\gamma(m(-\lambda) - \lambda m'(-\lambda)), \tag{11}$$

where $m(\lambda) = \frac{(1-\gamma-\lambda) - \sqrt{(1-\gamma-\lambda)^2 - 4\gamma\lambda}}{2\gamma\lambda}$. Therefore, $R_X^S(\hat{\beta}_\lambda) \to \mathcal{R}_\lambda^S = \mathcal{B}_\lambda^S + \mathcal{V}_\lambda^S + \tilde{\sigma}^2$.

Proof. As stated in the main paper, this result for the statistical model was already proven in Hastie et al. (2022). □

D Proof of Proposition 3.2

Proposition 3.2 (Causal Risk Increases with Confounding Strength). *Consider the family of causal models parameterized as in (1) that entail the same observational distribution. Let C_1 and C_2 be two such causal models with confounding strengths ζ_1 and ζ_2 and alignments η_1 and η_2 (defined in Theorem 3.1), respectively. Then for all $\lambda, \gamma \in (0, \infty)$,*

$$\zeta_1 > \zeta_2, \;\; \eta_1 \leq \eta_2 \implies \mathcal{R}_\lambda^{C_1} > \mathcal{R}_\lambda^{C_2}.$$

In particular, for any fixed η, the measure of confounding strength ζ establishes a strict ordering of causal models. This includes the ICM under which $\eta = 0$.

Proof. For any fixed $\lambda \in (0, \infty)$, the difference in limiting causal risks incurred by $\hat{\beta}_\lambda$ on causal models C_1 and C_2 is given by

$$\begin{aligned}
\mathcal{R}_1^C(\gamma, \lambda) - \mathcal{R}_2^C(\gamma, \lambda) &= 2\tilde{r}^2\big((\frac{\omega_1^2}{\tilde{r}^2} - \frac{\omega_2^2}{\tilde{r}^2}) - (\zeta_1 - \zeta_2)\lambda m(-\lambda)\big) \\
&= 2\tilde{r}^2\big((\zeta_1 - \zeta_2)(1 - \lambda m(-\lambda)) - (\eta_1 - \eta_2)\big) \\
&= 2\tilde{r}^2\big((\zeta_1 - \zeta_2)(1 - \lambda m(-\lambda)) - (\eta_1 - \eta_2)\big)
\end{aligned}$$

Since, as shown below, $(1 - \lambda m(-\lambda)) > 0$ for any $\lambda, \gamma \in (0, \infty)$, it holds that

$$\zeta_1 > \zeta_2, \;\; \eta_1 \leq \eta_2 \implies \mathcal{R}_1^C(\gamma, \lambda) > \mathcal{R}_2^C(\gamma, \lambda).$$

$$\begin{aligned}
1 - \lambda m(-\lambda) &= 1 - \frac{\gamma - 1 - \lambda + \sqrt{(1 + \lambda + \gamma)^2 - 4\gamma}}{2\gamma} \\
&= \frac{(1 + \gamma + \lambda) - \sqrt{(1 + \lambda + \gamma)^2 - 4\gamma}}{2\gamma} \\
&> 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{since } \gamma > 0)
\end{aligned}$$

□

E Proofs for Sections 4 and 5

We start with a technical lemma that we need in the proofs of the following theorems. It controls a function that appears in the derivative of the limiting causal riks $\partial_\lambda \mathcal{R}_\lambda^C$.

Lemma E.1. *For $\lambda \geq 0$ and $\gamma, S > 0$ consider the function*

$$f(\lambda, \gamma, S) = 2\gamma \frac{\lambda - S^{-1}\gamma}{(1 + \lambda + \gamma - \sqrt{(1 + \lambda + \gamma)^2 - 4\gamma})((1 + \lambda + \gamma)^2 - 4\gamma)}\,.$$

This function has the following properties

(i) *f is increasing in λ,*

(ii) $f(\lambda, \gamma, S) \xrightarrow[\lambda \to \infty]{} 1$, *and*

(iii) $f(\lambda, \gamma, S) \xrightarrow[\lambda \to 0]{} \begin{cases} -S^{-1}\frac{\gamma}{(\gamma-1)^2}, & \gamma < 1 \\ -\infty, & \gamma = 1 \\ -S^{-1}\frac{\gamma^2}{(\gamma-1)^2}, & \gamma > 1 \end{cases}.$

Proof. For readability, we use the shorthand notations $x = 1 + \lambda + \gamma$ and $\varphi = x^2 - 4\gamma$, under which f is given by

$$f(\lambda, \gamma, S) = 2\gamma \frac{\lambda - S^{-1}\gamma}{(x - \sqrt{\varphi})\varphi}\,.$$

(i) The partial derivative of f in λ is given by

$$\partial_\lambda f(\lambda,\gamma,S) = 2\gamma \frac{(x-\sqrt{\varphi})\varphi - (\lambda - S^{-1}\gamma)\left[(1-\frac{x}{\sqrt{\varphi}})\varphi + 2x(x-\sqrt{\varphi})\right]}{(x-\sqrt{\varphi})^2\varphi^2}$$
$$= \underbrace{\frac{2\gamma}{(x-\sqrt{\varphi})\varphi^2}}_{>0} \underbrace{\left[\varphi - (\lambda - S^{-1}\gamma)(2x-\sqrt{\varphi})\right]}_{=:g(\lambda)},$$

where the first fraction is positive because $\varphi > x^2$ and $x - \sqrt{\varphi} > 0$. It is therefore sufficient to show $g(\lambda) \geq 0$ for $\partial_\lambda f(\lambda,\gamma,S) \geq 0$. We first get rid of the S term via

$$g(\lambda) = \varphi - (\lambda - S^{-1}\gamma)\underbrace{(2x-\sqrt{\varphi})}_{\geq 0} \geq \varphi - \lambda(2x-\sqrt{\varphi})\,.$$

Finally, we lower bound $\sqrt{\varphi}$ in two different ways depending on γ. For $\gamma \leq 1$, it is $\varphi = (1+\lambda-\gamma)^2 + 4\gamma\lambda$ and therefore $\sqrt{\varphi} \geq 1+\lambda-\gamma = x - 2\gamma$. This yields

$$g(\lambda) \geq \varphi - \lambda(2x-\sqrt{\varphi}) \geq \varphi - \lambda(x+2\gamma) = (1-\gamma)\lambda + (\gamma-1)^2 \geq 0\,.$$

For $\gamma > 1$, it is $\varphi = (-1+\lambda+\gamma)^2 + 4\lambda$ and therefore $\sqrt{\varphi} \geq -1+\lambda+\gamma = x - 2$. This yields

$$g(\lambda) \geq \varphi - \lambda(2x-\sqrt{\varphi}) \geq \varphi - \lambda(x+2) = (\gamma-1)\lambda + (\gamma-1)^2 \geq 0\,.$$

In summary, we have shown $\partial_\lambda f(\lambda,\gamma,S) \geq g(\lambda) \geq 0$.

(ii) With the first order Taylor approximation $1 - \sqrt{1-h} = 1/2h + \mathcal{O}(h^2)$, we get

$$(x-\sqrt{\varphi})\varphi = \left(1 - \sqrt{1-\frac{4\gamma}{x^2}}\right) x\varphi = \left(\frac{2\gamma}{x^2} + \mathcal{O}(\lambda^{-4})\right) x\varphi = 2\gamma x + \mathcal{O}(\lambda^{-1}) = 2\gamma\lambda + \mathcal{O}(1)\,,$$

which yields

$$f(\lambda,\gamma,S) = 2\gamma\frac{\lambda - S^{-1}\gamma}{(x-\sqrt{\varphi})\varphi} = \frac{2\gamma\lambda - 2S^{-1}\gamma^2}{2\gamma\lambda + \mathcal{O}(1)} \xrightarrow[\lambda\to\infty]{} 1\,.$$

(iii) The denominator satisfies

$$(x-\sqrt{\varphi})\varphi \xrightarrow[\lambda\to 0]{} (1+\gamma-|\gamma-1|)(\gamma-1)^2 = \begin{cases} 2\gamma(\gamma-1)^2, & \gamma < 1 \\ 0, & \gamma = 1 \\ 2\gamma-1)^2, & \gamma > 1 \end{cases}.$$

Since $\lambda - S^{-1}\gamma \xrightarrow[\lambda\to 0]{} S^{-1}\gamma < 0$, the claim follows. □

Recall that the optimal causal regularization is defined as the minimizer of the causal risk $\lambda_C^*(\gamma) = \arg\inf_{\lambda\in(0,\infty)} \mathcal{R}_\lambda^C$. The following lemma distinguishes between three different regimes of the risk function $\mathcal{R}_\lambda^C$ depending on the confounding strength ζ.

Lemma E.2 (Regimes of the Optimal Causal Regularization). *For any causal model parameterized as in (1), we can distinguish the following regimes of $\lambda_C^*(\gamma)$:*

1. *The function $\lambda \mapsto \mathcal{R}_\lambda^C$ is increasing (which implies $\lambda_C^*(\gamma) = 0$), if and only if $\gamma \neq 1$ and*

$$\zeta \leq -\,\mathrm{SNR_S}^{-1}\,\frac{\gamma\max\{1,\gamma\}}{(1-\gamma)^2}\,.$$

2. *For any $\gamma > 0$, the function $\lambda \mapsto \mathcal{R}_\lambda^C$ is decreasing (which implies $\lambda_C^*(\gamma) = \infty$) if and only if $\zeta \geq 1$.*

3. *For any $\zeta \in \mathbb{R}$, $\gamma \in (0, \infty)$ which do not satisfy the conditions 1. or 2., it is $\lambda_C^*(\gamma) \in (0, \infty)$ and it $\lambda_C(\gamma)$ satisfies the critical point condition $\partial_\lambda \mathcal{R}_\lambda^C(\lambda_C^*(\gamma)) = 0$, or equivalently,*

$$0 = \lambda_C^*(\gamma) - {\mathrm{SNR}_\mathrm{S}}^{-1}\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda_C^*(\gamma) + \gamma - \sqrt{\varphi(\lambda_C^*(\gamma))}\right)\varphi(\lambda_C^*(\gamma))\,,$$

where $\varphi(\lambda) = (1 + \lambda + \gamma)^2 - 4\gamma$.

Proof. We use the shorthand notation $\varphi(\lambda) = (1 + \lambda + \gamma)^2 - 4\gamma$. Recall the confounding strength $\zeta = (r^2 + \eta)/\tilde{r}^2$ and the statistical signal-to-noise ratio $\mathrm{SNR}_\mathrm{S} = \tilde{r}^2/\tilde{\sigma}^2$. The derivative of the limiting causal risk $\mathcal{R}_\lambda^C$ in λ is given by

$$\partial_\lambda \mathcal{R}_\lambda^C = \frac{2\tilde{r}^2}{\varphi(\lambda)^{3/2}}\left(\lambda - {\mathrm{SNR}_\mathrm{S}}^{-1}\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda + \gamma - \sqrt{\varphi(\lambda)}\right)\varphi(\lambda)\right)$$

1. The first condition $\partial_\lambda \mathcal{R}_\lambda^C \geq 0$ for all $\lambda > 0$ can be equivalently rearranged for the confounding strength as

$$\zeta \leq 2\gamma \frac{\lambda - {\mathrm{SNR}_\mathrm{S}}^{-1}\gamma}{\left(1 + \lambda + \gamma - \sqrt{\varphi(\lambda)}\right)\varphi(\lambda)} = f(\lambda, \gamma, \mathrm{SNR}_\mathrm{S})\,,$$

   where f is the function investigated in Lemma E.1. This in turn is equivalent to taking the infimum over λ, which is given by Lemma E.1 as

$$\zeta \leq \inf_{\lambda > 0} f(\lambda, \gamma, \mathrm{SNR}_\mathrm{S}) = -{\mathrm{SNR}_\mathrm{S}}^{-1}\frac{\gamma \max\{1, \gamma\}}{(1 - \gamma)^2}.$$

   Note that for $\gamma = 1$ this infimum is $-\infty$, so the condition cannot be satisfied for any ζ.

2. The proof of the second claim is analogue to the first with the reverse inequality $\partial_\lambda \mathcal{R}_\lambda^C \leq 0$. Rearranging for ζ and using Lemma E.1 yields the equivalent condition

$$\zeta \geq \sup_{\lambda > 0} f(\lambda, \gamma, \mathrm{SNR}_\mathrm{S}) = 1\,.$$

3. For the third claim, assume that the pair of ζ and γ satisfies neither of the first points. We will use this to show that the derivative at 0 is negative $\partial_\lambda \mathcal{R}_\lambda^C(0) < 0$ and the derivative $\partial_\lambda \mathcal{R}_\lambda^C$ for sufficiently large λ is positive. This together then implies that the minimum $\lambda_C^*(\gamma)$ of the function $\mathcal{R}_\lambda^C$ is indeed attained at a finite value in $(0, \infty)$, and $\mathcal{R}_\lambda^C$ satisfies the critical point condition $\partial_\lambda \mathcal{R}_\lambda^C(\lambda_C^*(\gamma)) = 0$.

   For the derivative at 0, assume that the converse is true, that is, $\partial_\lambda \mathcal{R}_\lambda^C(0) \geq 0$. Rearranging this condition for ζ yields similarly to the first case of this lemma that $\zeta \leq f(0, \gamma, \mathrm{SNR}_\mathrm{S})$. However Lemma E.1 states that f is increasing in λ, which means that this condition already implies $\zeta \leq f(\lambda, \gamma, \mathrm{SNR}_\mathrm{S})$ for all λ. This means that the pair ζ, γ would satisfy the condition of the first case, which contradicts our assumption.

   For the behavior of large λ, observe that the sign of the derivative is determined by the sign of the term $\lambda - {\mathrm{SNR}_\mathrm{S}}^{-1}\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda + \gamma - \sqrt{\varphi(\lambda)}\right)\varphi(\lambda)$. As derived in the proof of Lemma E.1, we have the asymptotic behavior

$$\left(1 + \lambda + \gamma - \sqrt{\varphi(\lambda)}\right)\varphi(\lambda) = 2\gamma\lambda + \mathcal{O}(1)\,,$$

   which yields

$$\lambda - {\mathrm{SNR}_\mathrm{S}}^{-1}\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda + \gamma - \sqrt{\varphi(\lambda)}\right)\varphi(\lambda) = (1 - \zeta)\lambda + \mathcal{O}(1)\,.$$

   Since the pair ζ, γ does by assumption not satisfy the conditions of the second case, we have $\zeta < 1$, which means that the above term is eventually positive.

□

Theorem 4.1 (Optimal Regularization can be Negative). *For any causal model parameterized as in (1), the following cases distinguish between whether the min-norm interpolator is optimal or not.*

1. *For negative confounding strength $\zeta < 0$ the optimal causal regularization λ_C^* can be 0 or even negative. A necessary and sufficient condition for $\lambda_C^* \leq 0$ depends on the difference in causal and statistical signal-to-noise ratios and is given by*
$$\mathrm{SNR_C} - \mathrm{SNR_S} \geq \frac{\gamma \max\{1, \gamma\}}{(1-\gamma)^2}\,.$$
2. *For positive confounding strength $\zeta > 0$ the optimal causal regularization is positive $\lambda_C^* > 0$ and $\mathcal{R}_0^C > \mathcal{R}_{\lambda_C^*}^C$, hence regularization is beneficial. This includes the ICM.*

Proof. The first statement of the theorem is a special case of Theorem 5.2. The necessary and sufficient condition for $\lambda_C^* = 0$ stated there is equivalently reformulated as

$$\begin{aligned} & \zeta \leq -\,\mathrm{SNR_S}^{-1}\,\frac{\gamma \max\{1,\gamma\}}{(1-\gamma)^2} \\ \Leftrightarrow \quad & -\,\mathrm{SNR_S}\,\zeta \geq \frac{\gamma \max\{1,\gamma\}}{(1-\gamma)^2} \\ \Leftrightarrow \quad & \mathrm{SNR_C} - \mathrm{SNR_S} \geq \frac{\gamma \max\{1,\gamma\}}{(1-\gamma)^2}\,, \end{aligned}$$

where the last part used the equality $\mathrm{SNR_C} = (1-\zeta)\,\mathrm{SNR_S}$. The statement about negative λ_C^* refers to the fact that the derivative of the risk at 0 can be positive, that is, $\partial \mathcal{R}_\lambda^C(0) > 0$. This was shown in the proof of Lemma E.2 and suggests that without our restriction $\lambda_C^* \geq 0$, a negative value of λ would yield an even smaller risk.

For the second statement, observe that the condition $\zeta > 0$ implies the cases 2. or 3. from Lemma E.2. In particular, this implies $\lambda_C^* > 0$. The proof of Lemma E.2 showed that in both of these cases it holds $\partial_\lambda \mathcal{R}_\lambda^C(0) < 0$, which means that the causal limiting risk $\mathcal{B}_\lambda^C$ is strictly decreasing in a small neighborhood around 0. In particular, this implies that the minimal risk is strictly smaller than the risk at 0, that is, $\mathcal{R}_0^C > \mathcal{R}_{\lambda_C^*}^C$.

□

Theorem 5.1 (Optimal Statistical vs. Causal Regularization). *For any causal model parameterized as in (1), the condition $\zeta = 0$ defines a phase transition for the optimal regularization via*

$$\zeta < 0 \iff \lambda_C^* < \lambda_S^*, \qquad \zeta = 0 \iff \lambda_C^* = \lambda_S^*, \quad \textit{and} \quad \zeta > 0 \iff \lambda_C^* > \lambda_S^*.$$

In particular under the ICM, the optimal causal regularization λ_C^* is always strictly larger than the optimal statistical regularization λ_S^*, unless $\zeta = 0$, in which case they coincide.

Proof. Lemma E.2 distinguishes between three different regimes of ζ. The first two regimes yield

$$\zeta \leq -\,\mathrm{SNR_S}^{-1}\,\frac{\gamma \max\{1,\gamma\}}{(1-\gamma)^2} \implies \lambda_C^* = 0 \quad \text{and} \quad 1 \leq \zeta \implies \lambda_C^* = \infty\,.$$

Combined with $\lambda_S^* = \mathrm{SNR_S}^{-1}\,\gamma \in (0, \infty)$, these regimes agree with the claim in the theorem. It remains to show that the theorem also holds for the last regime $-\,\mathrm{SNR_S}^{-1}\,\frac{\gamma \max\{1,\gamma\}}{(1-\gamma)^2} < \zeta < 1$. In this regime according to Lemma E.2, the optimal causal regularization λ_C^* satisfies the critical point condition

$$\begin{aligned} & 0 = \lambda_C^* - \mathrm{SNR_S}^{-1}\,\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda_C^* + \gamma - \sqrt{\varphi(\lambda_C^*)}\right)\varphi(\lambda_C^*) \\ \Leftrightarrow \quad & \lambda_C^* - \lambda_S^* = \frac{\zeta}{2\gamma}\left(1 + \lambda_C^* + \gamma - \sqrt{\varphi(\lambda_C^*)}\right)\varphi(\lambda_C^*)\,. \end{aligned}$$

Since the term $1/(2\gamma)\left(1 + \lambda_C^* + \gamma - \sqrt{\varphi(\lambda_C^*)}\right)\varphi(\lambda_C^*)$ is positive, the sign of $\lambda_C^* - \lambda_S^*$ is determined by the sign of ζ as claimed in the theorem.

□

Theorem 5.2 (Increasing Confounding Strength Requires Stronger Regularization). *Consider the family of causal models parameterized as in (1) that entail the same observational distribution. The optimal causal regularization λ_C^* only depends on the confounding strength ζ and λ_C^* is an increasing function in ζ. More specifically, using $\varrho = -\,\mathrm{SNR_S}^{-1}\,\gamma \max\{1,\gamma\}/(1-\gamma)^2$:*

$$\varrho < \zeta < 1 \implies \lambda_C^* \in (0,\infty) \text{ with } \partial_\zeta \lambda_C^* > 0\,,$$

$\lambda_C^* = 0$ if $\zeta \leq \varrho$ and $\lambda_C^* = \infty$ for $\zeta \geq 1$.

Proof. The theorem follows directly from Lemma E.2, except for the statement about λ_C^* being strictly increasing in ζ. In the corresponding regime, Lemma E.2 states that λ_C^* satisfies the critical point condition $\partial_\lambda \mathcal{R}_\lambda^C(\lambda_C^*) = 0$, which we will use to show that the derivative of λ_C^* in ζ is strictly positive. For readability, we use the notation $x(\zeta) = 1 + \lambda_C^*(\zeta) + \gamma$ and $\varphi(\zeta) = x(\zeta)^2 - 4\gamma$. The optimal causal regularization $\lambda_C^*(\zeta)$ satisfies the critical point condition

$$0 = x(\zeta) - (1 + \gamma + \mathrm{SNR_S}^{-1}\,\gamma) - \frac{\zeta}{2\gamma}\left(x(\zeta) - \sqrt{\varphi(\zeta)}\right)\varphi(\zeta) =: g(x(\zeta),\zeta)\,.$$

Rearranging this equation yields

$$\frac{\zeta}{2\gamma}\left(x(\zeta) - \sqrt{\varphi(\zeta)}\right) = \frac{x(\zeta) - (1 + \gamma + \mathrm{SNR_S}^{-1}\,\gamma)}{\varphi(\zeta)}\,. \tag{12}$$

The partial derivatives of the function $g = g(x,\zeta)$ evaluated at $(x(\zeta),\zeta)$ are given by

$$\partial_\zeta g(x(\zeta),\zeta) = -\frac{1}{2\gamma}\left(x(\zeta) - \sqrt{\varphi(\zeta)}\right)\varphi(\zeta) < 0$$

and

$$\begin{aligned}
\partial_x g(x(\zeta),\zeta) &= 1 - \frac{\zeta}{2\gamma}\left[\left(1 - \frac{x(\zeta)}{\sqrt{\varphi(\zeta)}}\right)\varphi(\zeta) + 2x(\zeta)\left(x(\zeta) - \sqrt{\varphi(\zeta)}\right)\right] \\
&= 1 - \frac{\zeta}{2\gamma}\left(x(\zeta) - \sqrt{\varphi(\zeta)}\right)\left(2x(\zeta) - \sqrt{\varphi(\zeta)}\right) \\
&= 1 - \frac{x(\zeta) - (1 + \gamma + \mathrm{SNR_S}^{-1}\,\gamma)}{\varphi(\zeta)}\left(2x(\zeta) - \sqrt{\varphi(\zeta)}\right) && \text{(Using Eq. (12))} \\
&> 1 - \frac{x(\zeta) - 2\sqrt{\gamma}}{\varphi(\zeta)}\left(2x(\zeta) - \sqrt{\varphi(\zeta)}\right)\,. && (1 + \gamma + \mathrm{SNR_S}^{-1}\,\gamma > 2\sqrt{\gamma})
\end{aligned}$$

Since $\varphi(\zeta) = (x(\zeta) - 2\sqrt{\gamma})(x(\zeta) + 2\sqrt{\gamma}) < (x(\zeta) + 2\sqrt{\gamma})^2$, it further follows

$$\begin{aligned}
\partial_x g(x(\zeta),\zeta) &> 1 - \frac{x(\zeta) - 2\sqrt{\gamma}}{(x(\zeta) - 2\sqrt{\gamma})(x(\zeta) + 2\sqrt{\gamma})}\left(2x(\zeta) - (x(\zeta) + 2\sqrt{\gamma})\right) \\
&= 1 - \frac{x(\zeta) - 2\sqrt{\gamma}}{x(\zeta) + 2\sqrt{\gamma}} \\
&> 0\,.
\end{aligned}$$

With these results, we can take the derivative in ζ of the critical point condition $0 = g(x(\zeta),\zeta)$ and obtain

$$0 = \frac{\mathrm{d}}{\mathrm{d}\zeta} g(x(\zeta),\zeta) = \underbrace{\partial_x g(x(\zeta),\zeta)}_{>0} \cdot \frac{\mathrm{d}x}{\mathrm{d}\zeta}(\zeta) + \underbrace{\partial_\zeta g(x(\zeta),\zeta)}_{<0} \cdot 1\,,$$

which yields $0 < \frac{\mathrm{d}x}{\mathrm{d}\zeta}(\zeta) = \frac{\mathrm{d}\lambda_C^*}{\mathrm{d}\zeta}(\zeta)$. This implies that λ_C^* is increasing in ζ and concludes the proof. □

F Shift interventions.

F.1 Causal risk under relative interventions.

Here, we characterize the causal risk of any linear predictor under *relative* or *shift* interventions. Similar to the definition of causal risk under hard interventions, to isolate the effects of the choice of α on the risk, we draw perturbations from the marginal of x. Formally, intervening on x under the causal model given by Eq. (1) corresponds to the structural equations

$$z \sim \mathcal{N}(0, I_l)\,, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)\,, \quad \nu \sim \mathcal{N}(0, MM^T)\,, \quad x = Mz\,, \quad x' = x + \nu\,, \quad y = x'^T\beta + z^T\alpha + \varepsilon\,.$$

Similar to the proof of Proposition 2.1, the key step here is to characterize the distribution of y under the shift intervention $y|do(x' \coloneqq x + \nu)$ for some ν chosen independently of x.

This lets us compute the risk of a linear predictor $\hat{\beta} \in \mathbb{R}^d$ under a shift intervention as

$$\begin{aligned}
R^C(\hat{\beta}) &= \mathbb{E}_\nu \mathbb{E}_x \mathbb{E}_{y_0|do(x'=x+\nu)} \left(x^T\hat{\beta} - y\right)^2 \\
&= \mathbb{E}_\nu \mathbb{E}_{x,z,\epsilon} \left((\hat{\beta} - \beta)^T(x + \nu) + \alpha^T z + \epsilon\right)^2 \\
&= \mathbb{E}_\nu \left((\hat{\beta} - \beta)^T\nu\right)^2 + \mathbb{E}_x \mathbb{E}_{z,\epsilon|x} \left((\hat{\beta} - \beta)^T x + \alpha^T z + \epsilon\right)^2 \\
&= \left\|\hat{\beta} - \beta\right\|_\Sigma^2 + \left\|\hat{\beta} - \tilde{\beta}\right\|_\Sigma^2 + \tilde{\sigma}^2
\end{aligned}$$

To obtain the last equality, refer to the derivation of the statistical and causal risks in Proposition 2.1. The expected risk under conditioning of X is then given by

$$\mathbb{E}_{Y|X}\|\hat{\beta} - \beta\|_\Sigma^2 + \mathbb{E}_{Y|X}\|\hat{\beta} - \tilde{\beta}\|_\Sigma^2\,. \tag{13}$$

F.2 Asymptotics and Optimal Ridge Regularization.

The limiting risk of any ridge estimator can then be directly derived from Theorems 3.1 and C.1.

Theorem F.1 (Limiting Causal Risk of the Ridge Estimator Under Shift Interventions). *Let $\|\beta\|^2 = r^2$, $\|\Gamma\|^2 = \omega^2$, $\langle\Gamma, \beta\rangle = \eta$, and fix $\tilde{\sigma}^2$. Then as $n, d \to \infty$ such that $d/n \to \gamma \in (0, \infty)$, it holds almost surely in X for every $\lambda > 0$ that*

$$R_X^C(\hat{\beta}_\lambda) \to \mathcal{R}_\lambda^C = \omega^2 + 2\tilde{r}^2\lambda^2 m'(-\lambda) - 2(\omega^2 + \eta)\lambda m(-\lambda) + 2\tilde{\sigma}^2\gamma(m(-\lambda) - \lambda m'(-\lambda))\,,$$

where $m(\lambda) = ((1 - \gamma - \lambda) - \sqrt{(1 - \gamma - \lambda)^2 - 4\gamma\lambda})/(2\gamma\lambda)$ and $\tilde{r}^2 = r^2 + \omega^2 + 2\eta$. The corresponding limiting quantities for the min-norm interpolator can be obtained by taking the limit $\lambda \to 0^+$.

Lemma F.2 (Regimes of the Optimal Causal Regularization Under Shift Interventions). *For any causal model parameterized as in (1), we can distinguish the following regimes of $\lambda_C^*(\gamma)$:*

1. *The function $\lambda \mapsto \mathcal{R}_\lambda^{C_{soft}}$ is increasing (which implies $\lambda_{C_{soft}}^*(\gamma) = 0$), if and only if $\gamma \neq 1$ and*

$$\zeta \leq -2\,\mathrm{SNR_S}^{-1}\,\frac{\gamma \max\{1, \gamma\}}{(1 - \gamma)^2}\,.$$

2. *For any $\gamma > 0$, the function $\lambda \mapsto \mathcal{R}_\lambda^{C_{soft}}$ is decreasing (which implies $\lambda_{C_{soft}}^*(\gamma) = \infty$) if and only if $\zeta \geq 2$.*

3. *For any $\zeta \in \mathbb{R}$, $\gamma \in (0, \infty)$ which do not satisfy the conditions 1. or 2., it is $\lambda_{C_{soft}}^*(\gamma) \in (0, \infty)$ and it $\lambda_{C_{soft}}^*(\gamma)$ satisfies the critical point condition $\partial_\lambda \mathcal{R}_\lambda^{C_{soft}}(\lambda_{C_{soft}}^*(\gamma)) = 0$, or equivalently,*

$$0 = \lambda_{C_{soft}}^*(\gamma) - \mathrm{SNR_S}^{-1}\,\gamma - \frac{\zeta}{4\gamma}\left(1 + \lambda_{C_{soft}}^*(\gamma) + \gamma - \sqrt{\varphi(\lambda_{C_{soft}}^*(\gamma))}\right)\varphi(\lambda_{C_{soft}}^*(\gamma))\,,$$

where $\varphi(\lambda) = (1 + \lambda + \gamma)^2 - 4\gamma$.

Proof. We use the shorthand notation $\varphi(\lambda) = (1+\lambda+\gamma)^2 - 4\gamma$. Recall the confounding strength $\zeta = (r^2+\eta)/\tilde{r}^2$ and the statistical signal-to-noise ratio $\mathrm{SNR}_\mathrm{S} = \tilde{r}^2/\tilde{\sigma}^2$. The derivative of the limiting causal risk under shift interventions $\mathcal{R}_\lambda^{C_{\text{soft}}}$ in λ is given by

$$\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}} = \frac{2\tilde{r}^2}{\varphi(\lambda)^{3/2}} \left(\lambda - {\mathrm{SNR}_\mathrm{S}}^{-1} \gamma - \frac{\zeta}{4\gamma} \left(1+\lambda+\gamma - \sqrt{\varphi(\lambda)} \right) \varphi(\lambda) \right)$$

1. The first condition $\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}} \geq 0$ for all $\lambda > 0$ can be equivalently rearranged for the confounding strength as

$$\zeta \leq 4\gamma \frac{\lambda - {\mathrm{SNR}_\mathrm{S}}^{-1}\gamma}{\left(1+\lambda+\gamma-\sqrt{\varphi(\lambda)}\right)\varphi(\lambda)} = 2f(\lambda,\gamma,\mathrm{SNR}_\mathrm{S})\,,$$

   where f is the function investigated in Lemma E.1. This in turn is equivalent to taking the infimum over λ, which is given by Lemma E.1 as

$$\zeta \leq \inf_{\lambda>0} 2f(\lambda,\gamma,\mathrm{SNR}_\mathrm{S}) = -2\,{\mathrm{SNR}_\mathrm{S}}^{-1} \frac{\gamma \max\{1,\gamma\}}{(1-\gamma)^2}.$$

   Note that for $\gamma = 1$ this infimum is $-\infty$, so the condition cannot be satisfied for any ζ.

2. The proof of the second claim is analogue to the first with the reverse inequality $\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}} \leq 0$. Rearranging for ζ and using Lemma E.1 yields the equivalent condition

$$\zeta \geq \sup_{\lambda>0} 2f(\lambda,\gamma,\mathrm{SNR}_\mathrm{S}) = 2\,.$$

3. For the third claim, assume that the pair of ζ and γ satisfies neither of the conditions from above. We will use this to show that the derivative at 0 is negative $\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}}(0) < 0$ and the derivative $\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}}$ for sufficiently large λ is positive. This together then implies that the minimum $\lambda^*_{C_{\text{soft}}}(\gamma)$ of the function $\mathcal{R}_\lambda^{C_{\text{soft}}}$ is indeed attained at a finite value in $(0,\infty)$, and $\mathcal{R}_\lambda^{C_{\text{soft}}}$ satisfies the critical point condition $\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}}(\lambda^*_{C_{\text{soft}}}(\gamma)) = 0$.

   For the derivative at 0, assume that the converse is true, that is, $\partial_\lambda \mathcal{R}_\lambda^{C_{\text{soft}}}(0) \geq 0$. Rearranging this condition for ζ yields similarly to the first case of this lemma that $2\zeta \leq f(0,\gamma,\mathrm{SNR}_\mathrm{S})$. However Lemma E.1 states that f is increasing in λ, which means that this condition already implies $\zeta \leq 2f(\lambda,\gamma,\mathrm{SNR}_\mathrm{S})$ for all λ. This means that the pair ζ,γ would satisfy the condition of the first case, which contradicts our assumption.

   For the behavior of large λ, observe that the sign of the derivative is determined by the sign of the term $\lambda - {\mathrm{SNR}_\mathrm{S}}^{-1}\gamma - \frac{\zeta}{4\gamma}\left(1+\lambda+\gamma-\sqrt{\varphi(\lambda)}\right)\varphi(\lambda)$. As derived in the proof of Lemma E.1, we have the asymptotic behavior

$$\left(1+\lambda+\gamma-\sqrt{\varphi(\lambda)}\right)\varphi(\lambda) = 2\gamma\lambda + \mathcal{O}(1)\,,$$

   which yields

$$\lambda - {\mathrm{SNR}_\mathrm{S}}^{-1}\gamma - \frac{\zeta}{4\gamma}\left(1+\lambda+\gamma-\sqrt{\varphi(\lambda)}\right)\varphi(\lambda) = (1-\zeta/2)\lambda + \mathcal{O}(1)\,.$$

   Since the pair ζ,γ does by assumption not satisfy the conditions of the second case, we have $\zeta < 1$, which means that the above term is eventually positive.

□

Theorem F.3 (Optimal Causal Regularization Under Shift Interventions). *For any causal model parameterized as in (1),*

1. *If $\zeta \geq 0$, then the optimal causal regularization under shift interventions $\lambda^*_{C_{soft}}$ satisfies $\lambda^*_S \leq \lambda^*_{C_{soft}} \leq \lambda^*_C$.*

2. *If* $\zeta < 0$, *then* $\lambda_C^* \leq \lambda_{C_{soft}}^* \leq \lambda_S^*$.

Indeed, the optimal causal regularization under shift interventions satisfies $\lambda_{C_{soft}}^* = \lambda_S^* + (\lambda_C^* - \lambda_S^*)/2$.

Proof. Lemma E.2 distinguishes between three different regimes of ζ. The first two regimes yield

$$\zeta \leq -2\,{\mathrm{SNR}_{\mathrm{S}}}^{-1}\,\frac{\gamma \max\{1,\gamma\}}{(1-\gamma)^2} \implies \lambda_C^* = 0 \quad \text{and} \quad 2 \leq \zeta \implies \lambda_C^* = \infty\,.$$

Combined with $\lambda_S^* = {\mathrm{SNR}_{\mathrm{S}}}^{-1}\,\gamma \in (0,\infty)$, these regimes agree with the claim in the theorem. It remains to show that the theorem also holds for the last regime $-2\,{\mathrm{SNR}_{\mathrm{S}}}^{-1}\,\frac{\gamma \max\{1,\gamma\}}{(1-\gamma)^2} < \zeta < 2$. In this regime according to Lemma E.2, the optimal causal regularization λ_C^* satisfies the critical point condition

$$\begin{aligned} 0 &= \lambda_{C_{\mathrm{soft}}}^* - {\mathrm{SNR}_{\mathrm{S}}}^{-1}\,\gamma - \frac{\zeta}{4\gamma}\left(1 + \lambda_{C_{\mathrm{soft}}}^* + \gamma - \sqrt{\varphi(\lambda_{C_{\mathrm{soft}}}^*)}\right)\varphi(\lambda_{C_{\mathrm{soft}}}^*) \\ \Leftrightarrow \quad \lambda_{C_{\mathrm{soft}}}^* - \lambda_S^* &= \frac{\zeta}{4\gamma}\left(1 + \lambda_{C_{\mathrm{soft}}}^* + \gamma - \sqrt{\varphi(\lambda_{C_{\mathrm{soft}}}^*)}\right)\varphi(\lambda_{C_{\mathrm{soft}}}^*)\,. \end{aligned}$$

Similarly, we know from the proof of Theorem 5.1 λ_C^* satisfies

$$\begin{aligned} 0 &= \lambda_C^* - {\mathrm{SNR}_{\mathrm{S}}}^{-1}\,\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda_C^* + \gamma - \sqrt{\varphi(\lambda_C^*)}\right)\varphi(\lambda_C^*) \\ \Leftrightarrow \quad \lambda_C^* - \lambda_{C_{\mathrm{soft}}}^* &= \frac{\zeta}{4\gamma}\left(1 + \lambda_{C_{\mathrm{soft}}}^* + \gamma - \sqrt{\varphi(\lambda_{C_{\mathrm{soft}}}^*)}\right)\varphi(\lambda_{C_{\mathrm{soft}}}^*)\,. \end{aligned}$$

Since the term $1/(2\gamma)\left(1 + \lambda_{C_{\mathrm{soft}}}^* + \gamma - \sqrt{\varphi(\lambda_{C_{\mathrm{soft}}}^*)}\right)\varphi(\lambda_{C_{\mathrm{soft}}}^*)$ is positive, the sign of $\lambda_{C_{\mathrm{soft}}}^* - \lambda_S^*$ and $\lambda_C^* - \lambda_{C_{\mathrm{soft}}}^*$ is determined by the sign of ζ as claimed in the theorem. □

G Beyond Gaussianity

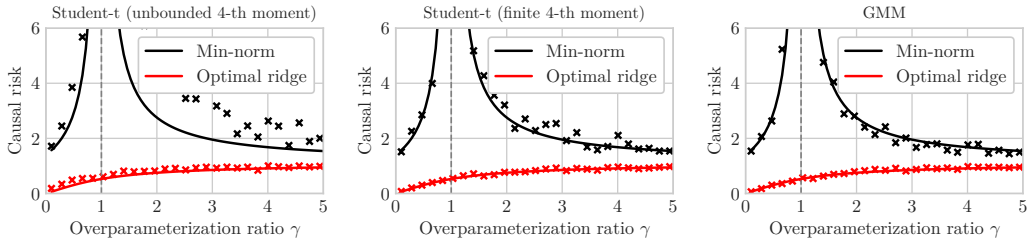

Figure 5: Causal risk of the minimum norm l_2 interpolator and the (causally)optimally regularized ridge regressor under a student-t distribution with unbounded 4th moments (3 degrees of freedom, left), a student-t distribution with bounded 4th moments (10 degrees of freedom, middle), a mixture of Gaussians (right). We choose the parameters $d = 300, l = 350$, statistical signal $\tilde{r}^2 = 5$, statistical noise $\tilde{\sigma}^2 = 1$, causal noise $\sigma^2 = .5$ and confounding strength $\zeta = 0.5$. For Gaussian mixtures, we consider a (centered and normalized) mixture of $k = 5$ Gaussians. Each individual mixture component has mean $\mu_i \sim \mathcal{N}(0_l, \frac{k^2}{(k-1)l} I_l)$ and identity covariance $\mathrm{Cov}_i = I_l$.

The analysis of this paper can be extended beyond the Gaussian setting by considering random variables generated by finite mixtures of Gaussians. The analysis can get considerably more technical and is left as future work, but we include a brief discussion here. Due to the Universality phenomenon in the high-dimensional limit, we believe that our limiting expressions (and the qualitative messages derived henceforth) would be rather robust to shifts in the marginal distribution as long as moments of order $(4 + \delta)$ for some $\delta > 0$ are bounded. We conducted experiments to verify this claim and the corresponding results can be found in Figure 5. These experiments compare our theoretically

derived asymptotic risks with finite-sample risks of the min-norm interpolator and causally optimally regularized ridge regressor. Instead of Gaussian confounders $z \sim \mathcal{N}(0, I_l)$, we only fix the first two moments 0 and I_l and generate z such that $\mathbb{E}[z] = 0$, $\mathrm{Cov}[z] = I$ from heavy-tailed multivariate t-distribution with different degrees of freedom, and finite mixture of Gaussians. Each plot shows the causal risk of min-norm interpolator and optimally regularized ridge regressor based on finite samples along with our theoretical asymptotic predictions. Our experiments show that, for distributions with finite 4th moments, the finite-sample risks closely match the theoretical results.