Interpolation and Regularization for Causal Learning

Leena Chennuru Vankadara*
University of Tübingen

Luca Rendsburg*
University of Tübingen

Ulrike von Luxburg
University of Tübingen

Debarghya Ghoshdastidar
Technical University of Munich

Abstract

Recent work shows that in complex model classes, interpolators can achieve statistical generalization and even be optimal for statistical learning. However, despite increasing interest in learning models with good causal properties, there is no understanding of whether such interpolators can also achieve *causal generalization*. To address this gap, we study causal learning from observational data through the lens of interpolation and its counterpart—regularization. Under a simple linear causal model, we derive precise asymptotics for the causal risk of the min-norm interpolator and ridge regressors in the high-dimensional regime. We find a large range of behavior that can be precisely characterized by a new measure of *confounding strength*. When confounding strength is positive, which holds under independent causal mechanisms—a standard assumption in causal learning—we find that interpolators cannot be optimal. Indeed, causal learning requires stronger regularization than statistical learning. Beyond this assumption, when confounding is negative, we observe a phenomenon of self-induced regularization due to positive alignment between statistical and causal signals. Here, causal learning requires weaker regularization than statistical learning, interpolators can be optimal, and optimal regularization can even be negative.

1 Introduction

We consider the problem of learning the causal influence of multivariate covariates $x \in \mathbb{R}^d$ on a scalar target variable $y \in \mathbb{R}$ purely from observational data and under the presence of hidden confounders. Formally, given finite samples $\{(x_i, y_i)\}_{i=1}^n$ drawn independently and identically (i.i.d) from the joint *observational distribution* $p(x, y) = p(x)p(y|x)$, the goal of causal learning is to predict the effects on the target variable y under interventions on the covariates x. In other words, the goal is to learn a predictive model that minimizes the expected loss on a random draw from the *interventional distribution* $p_{do}(x, y) = p(x)p(y|do(x))$, which can be different from the observational distribution.

Recently, Janzing (2019) established a close analogy between statistical and causal learning (albeit under a highly constructed confounded model). As a consequence, Janzing (2019) suggested that standard statistical learning-theoretic techniques (such as norm-based regularization) may also help learn good causal models. However, the classical statistical principles of bias-variance trade-off have been challenged in recent years by highly complex classes of models that are trained to interpolate the data and yet achieve remarkable generalization properties across a broad range of problem domains (Zhang et al., 2021). A large volume of recent work suggests that interpolation can be compatible with and may even be necessary to achieve optimal statistical generalization in the high-dimensional regime (Belkin et al., 2018; Belkin et al., 2019a; Liang et al., 2020; Feldman, 2020). Despite the surge in interest, causal properties of such interpolators have not yet been explored. In this work, we

* denotes equal contribution.

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

consider a simple linear causal model in the high-dimensional regime ($n, d \to \infty, d/n \in \mathcal{O}(1)$) and ask: can interpolators achieve good causal generalization?

1.1 Motivation and Related Work

Resemblance between statistical and causal generalization Causal learning can be regarded as an instance of the general problem of learning under distribution shifts, where the training (observational) distribution is shifted from the test (interventional) distribution. In the framework of out-of-distribution generalization, an interesting proposition for causal learning arises from the following high-level idea. Observing small sample sizes may induce a similar bias as distribution shifts. Therefore, techniques for learning models with good *out-of-sample* generalization (such as regularization) may also help to learn models with good *out-of-distribution* generalization and vice-versa. The literature provides plentiful evidence to support this general principle for different classes of distribution shifts. For instance, under a broad class of distribution shifts, distributionally robust optimization is equivalent to norm-based regularization (Xu et al., 2009; Shafieezadeh Abadeh et al., 2015; Gao et al., 2017; Shafieezadeh-Abadeh et al., 2019; Blanchet et al., 2019; Kuhn et al., 2019). Analogously, distributionally robust optimization techniques are also employed for statistical learning under limited samples (Zhu et al., 2020). Particularly relevant to our work is Janzing (2019), which formally establishes a close analogy between "generalizing from *empirical to observational distributions*" and "generalizing from *observational to interventional distributions*" under a highly constructed confounding model. This analogy suggests that standard norm-based regularization such as lasso or ridge, typically used for statistical learning, may also help learn better causal models.

Interpolation can be compatible with statistical learning Explicit norm-based regularization techniques were initially motivated by the classical learning theory principle of bias-variance trade-off, which is characterized by a U-shaped generalization curve. This principle recommends to avoid interpolation and instead suggests to balance data fitting with the complexity of the hypothesis class. Recently, however, these classical principles have been challenged by deep learning models. Despite being highly complex with the ability to fit even random labels and often trained to interpolate the training data, they achieve state-of-the-art out-of-sample generalization across many domains (Zhang et al., 2021). A partial explanation is provided by the *double-descent* phenomenon (Belkin et al., 2019b; Belkin, 2021). Extending the generalization curve beyond the interpolation threshold reveals two regimes: the classical U-curve in the *underparameterized* regime and a decreasing curve in the *overparameterized* regime. This behaviour is not limited to deep neural networks, but extends to other settings such as random feature models and random forests (Belkin et al., 2019b; Hastie et al., 2022; Mei et al., 2021). Follow-up work suggests that in the overparameterized regime, interpolators can indeed achieve low statistical risk (Belkin et al., 2019a; Liang et al., 2020; Bartlett et al., 2020; Tsigler et al., 2020; Muthukumar et al., 2020).

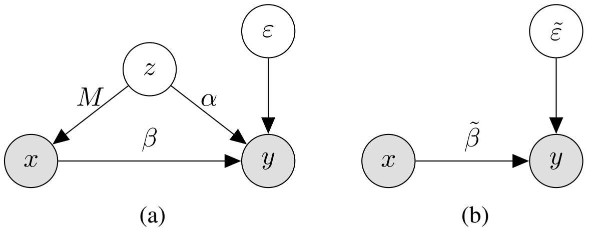
Is interpolation compatible with causal learning? On account of the parallels between statistical (out-of-sample) and causal (out-of-distribution) learning, it is therefore natural to ask: *can interpolators also learn good causal models?* One line of empirical work suggests that naively applying distributionally robust learning techniques such as importance reweighting or distributionally robust optimization approaches (which are equivalent to certain forms of regularization) offers vanishing benefits over empirical risk minimization in overparameterized model classes (Byrd et al., 2019; Sagawa et al., 2020; Gulrajani et al., 2021). However, there is also empirical evidence that suggests that augmenting such techniques with additional explicit norm-based regularization may help to learn distributionally robust models in the overparameterized regime (Sagawa et al., 2020; Donhauser et al., 2021). In the context of causal learning, it has been suggested that explicit regularization can be beneficial and might even need to be stronger than for statistical learning (Janzing, 2019; Vankadara et al., 2021). Existing work, therefore, remains unclear about the role of explicit regularization in causal learning, or correspondingly, whether interpolation is compatible with causal learning. In this work, we take a theoretical approach to systematically address these questions.

1.2 Our Contributions

We provide a first analysis of causal generalization from observational data in the modern, overparameterized and interpolating regime under a simple linear causal model. Specifically, we consider the interpolating minimum l_2 norm least-squares estimator and the family of regularized ridge regression estimators in the proportional asymptotic regime. We seek answers to the following questions: is there a regime where the optimal causal regularization parameter is 0, that is, can we observe *benign causal overfitting*? Furthermore, if the optimal causal regularization parameter is positive, how strongly do we need to regularize? How does the optimal causal regularization compare to the optimal statistical regularization? While our analysis is exhaustive, we emphasize the results under the assumption of independent causal mechanisms (Janzing et al., 2010), a standard assumption in causal learning.

- **Precise asymptotics of the causal risk (Section 3).** We provide precise asymptotics of the *causal risk* of the ridge regression estimator as well as the minimum l_2 norm interpolator in the high-dimensional setting: $n, d \to \infty, d/n \to \gamma \in (0, \infty)$. Our results confirm that, similar to the statistical setting, the causal generalization curve of the min-norm estimator exhibits the double-descent phenomenon. This is because the variance term diverges at the interpolation threshold and is decreasing in the overparameterized regime ($\gamma > 1$).
- **A measure of confounding strength ζ (Section 2.1).** We introduce a new measure of *confounding strength* ζ that quantifies the relative contribution of the "confounding signal" to the "causal signal". It can be interpreted as the strength of the distribution shift between the observational and interventional distributions. While ζ can take any real value in general, it is restricted to $[0, 1]$ under the assumption of independent causal mechanisms. There, it induces a strict, model-independent ordering of all causal models that entail the same observational distribution.
- **Benign causal overfitting (Section 4).** When the causal signal dominates the statistical signal ($\zeta < 0$), we observe a phenomenon of self-induced regularization due to the confounding signal. As a consequence, the optimal causal regularization can be 0 or negative even if the optimal statistical regularization is strictly positive. Under the assumption of independent causal mechanisms, however, we show that there is no benign causal overfitting. This is in contrast to benign statistical overfitting, which can occur in the highly underparameterized regime ($\gamma \to 0$).
- **Optimal causal vs. statistical regularization (Section 5).** We show that causal learning requires weaker regularization than statistical learning when the confounding strength ζ is negative. However, when $\zeta > 0$ and in particular under the principle of independent causal mechanisms, we show that causal learning requires stronger regularization than statistical learning. More specifically, the optimal causal regularization is strictly increasing in confounding strength.

2 Problem Setup



(a)

(b)

Figure 1: (*a*) Graphical model of the causal model defined in (1). (*b*) The usual statistical model. In both figures, observed random variables are shaded and unobserved variables are white.

We consider a linear causal model with parameters $M \in \mathbb{R}^{d \times l}$, $\alpha \in \mathbb{R}^l, \beta \in \mathbb{R}^d$ with $l \geq d$ and $\sigma^2 > 0$ described via the *structural equations*

$$z \sim \mathcal{N}(0, I_l)\,, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)\,, \quad x = Mz\,, \quad y = x^T\beta + z^T\alpha + \varepsilon\,. \tag{1}$$

The covariates $x \in \mathbb{R}^d$ and the target $y \in \mathbb{R}$ are *confounded* through $z \in \mathbb{R}^l$, which follows a standard normal distribution. This structure implies that $\mathbb{E}x = 0$ and the covariance of x is $\Sigma := \operatorname{Cov} x = MM^T$. A graphical representation of this causal model is given in Figure 1a. The observational joint distribution of this causal model is given by $p(x, y) = p(x)p(y|x)$, where $x \sim \mathcal{N}(0, \Sigma)$ and $y|x \sim \mathcal{N}(x^T\tilde{\beta}, \tilde{\sigma}^2)$. Here, the statistical parameter $\tilde{\beta} := \beta + \Gamma$ consists of the causal parameter β and a confounding parameter $\Gamma := M^{+T}\alpha$, where M^{+T} is shorthand for $(M^+)^T$ and M^+ denotes the Moore-Penrose inverse of M. The statistical noise is given by $\tilde{\sigma}^2 := \sigma^2 + \|\alpha\|^2 - \|\Gamma\|_\Sigma^2$, where $\|x\|_\Sigma^2 := x^T\Sigma x$ denotes the generalized norm. [1] Note that the observational distribution alone cannot distinguish the causal model from the one in Figure 1b. The

[1] Note that $\|\alpha\|^2 - \|\Gamma\|_\Sigma^2 = \|\alpha\|_{I-M^+M}^2 \geq 0$, where $I - M^+M$ is the orthogonal projection onto $\ker M$.

goal of statistical learning is to predict y after observing x, which is captured by the conditional distribution $p(y|x)$. In contrast, the goal of causal learning is to predict y after manipulating or intervening on x. This is formally captured by Pearl's do-calculus (Pearl, 2009), which describes interventions on random variables as a shift in the joint distribution. Intervening on x with the value x_0, denoted as $do(x = x_0)$, removes all arrows to x and *sets* $x = x_0$. In our causal model (1), the intervention $do(x = x_0)$ removes the arrow $z \to x$ and yields the updated structural causal equations

$$z \sim \mathcal{N}(0, I_l)\,, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)\,, \quad x = x_0\,, \quad y = x_0^T \beta + z^T \alpha + \varepsilon\,.$$

The corresponding distribution of y after intervening on x is therefore given by $y|do(x = x_0) \sim \mathcal{N}(x_0^T \beta, \tilde{\sigma}^2 + \|\Gamma\|_\Sigma^2)$. Since arbitrary interventions can introduce arbitrary distribution shifts, we consider the natural class of interventions drawn from the observational marginal distribution on x. This yields the interventional joint distribution $p_{do}(x, y) = p(x)p(y|do(x))$ with the slight abuse of notation $do(x)$ in which the random variable x and its value coincide.

Causal learning from observational data Assume we are given i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ from the observational joint distribution $p(x, y)$, which we collect in $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^n$. The usual statistical learning aims for the observational conditional $p(y|x)$, which means that train and test distributions coincide. Causal learning aims for the interventional conditional $p(y|do(x))$, a distribution shift problem for which train and test distributions differ. We define the corresponding *causal risk* R^C and *statistical risk* R^S of any linear regressor $\hat{\beta} \in \mathbb{R}^d$ under the squared loss as

$$R^C(\hat{\beta}) := \mathbb{E}_x \mathbb{E}_{y|do(x)} (x^T \hat{\beta} - y)^2 \quad \text{and} \quad R^S(\hat{\beta}) := \mathbb{E}_x \mathbb{E}_{y|x} (x^T \hat{\beta} - y)^2\,. \tag{2}$$

The following proposition (proven in Appendix A) characterizes the risks under the model (1).

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

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

Therefore, β is the optimal causal parameter and $\tilde{\beta}$ is the optimal statistical parameter. In the following, we simply refer to them as causal and statistical parameters.

2.1 A New Measure of Confounding Strength

Since the interventional distribution generally differs from the observational distribution, we require a measure that quantifies how this shift influences causal learning from observational data.

Signal-to-noise ratios (SNRs) Before we define our measure of confounding strength, we first define the statistical and causal signal-to-noise ratios, which help to intuitively understand our confounding strength measure. Recall that every causal model entails a statistical model since the causal parameter β and the confounding parameter Γ jointly specify the statistical parameter $\tilde{\beta} = \beta + \Gamma$. The statistical SNR is defined as usual by $\mathrm{SNR_S} := \|\tilde{\beta}\|^2/\tilde{\sigma}^2$. For the causal SNR, a natural notion would be $\|\beta\|^2/(\tilde{\sigma}^2 + \|\Gamma\|_\Sigma^2)$ if the learning algorithm had access to data from the interventional distribution $y|do(x) \sim \mathcal{N}(x^T\beta, \tilde{\sigma}^2 + \|\Gamma\|_\Sigma^2)$; but since we are constrained to data from the observational conditional $y|x \sim \mathcal{N}(x^T\tilde{\beta}, \tilde{\sigma}^2)$, the corresponding causal SNR, which quantifies the hardness of the learning problem, needs to take this into consideration. Accordingly, we consider the causal SNR as the ratio of the alignment between the statistical and causal parameters and the variance of the observational conditional. Formally, we define it as $\mathrm{SNR_C} := \langle \beta, \tilde{\beta} \rangle / \tilde{\sigma}^2$. In what follows, we therefore often refer to $\langle \beta, \tilde{\beta} \rangle$ as the *causal signal* and $\|\tilde{\beta}\|^2$ as the *statistical signal*. Correspondingly, we refer to $\langle \tilde{\beta} - \beta, \tilde{\beta} \rangle = \langle \Gamma, \tilde{\beta} \rangle$ as the *confounding signal*, which is the alignment between the confounding parameter Γ and the statistical parameter $\tilde{\beta}$.

Confounding strength Regression on observational data implicitly assumes that the interventional distribution coincides with the observational distribution, while it can be shifted in general. To quantify the impact of this distribution shift on the corresponding causal risk, we introduce a new *confounding strength measure* ζ. It measures the relative contribution of the confounding signal to the statistical signal and is defined by

$$\zeta := \frac{\langle \Gamma, \tilde{\beta} \rangle}{\langle \Gamma, \tilde{\beta} \rangle + \langle \beta, \tilde{\beta} \rangle} = \frac{\langle \Gamma, \tilde{\beta} \rangle}{\|\tilde{\beta}\|^2}\,. \tag{3}$$

Other notions of confounding strength are possible, but we will see later that this definition is well-suited to capture the shift strength for causal learning from observational data. Without further restrictions, ζ can take any value in $\mathbb{R}$. This measure divides the causal models into the following three regimes, depending on the relationship between causal and statistical signal:

- $\zeta \geq 1$: the causal signal $\langle \beta, \tilde{\beta} \rangle$ is non-positive, which implies that causal and statistical parameters are orthogonal or negatively aligned. Statistical learning is adversarial to causal learning.
- $0 < \zeta < 1$: causal and statistical parameters are positively aligned but the causal signal is weaker than the statistical signal $\|\tilde{\beta}\|^2$, for example $\beta = \tilde{\beta}/2$.
- $\zeta \leq 0$: the causal signal dominates the statistical signal, for example $\beta = 2\tilde{\beta}$.

The SNRs are related to the confounding strength measure via $\mathrm{SNR}_\mathrm{C} = (1-\zeta)\,\mathrm{SNR}_\mathrm{S}$. In particular, the causal signal decreases as the confounding strength increases.

The regime $0 \leq \zeta \leq 1$ is practically most relevant Causal learning often requires strong assumptions because causal models cannot be uniquely identified by their observational distribution. A standard assumption is the principle of independent causal mechanisms (ICM) (Janzing et al., 2010; Lemeire et al., 2013; Peters et al., 2017), which informally asserts that causal mechanisms share no information. In our causal model (1), a corresponding assumption could be that the causal mechanisms α and β are drawn from rotationally invariant distributions. This implies that $\langle \beta, \Gamma \rangle \to 0$ as $d \to \infty$, which in turn falls in the regime $0 \leq \zeta \leq 1$. While our following analysis covers all possible causal models, we pay special attention to this regime because it might be of highest practical relevance. Note that for $\langle \beta, \Gamma \rangle = 0$, our measure of confounding strength coincides with the measure $\zeta' = \|\Gamma\|^2/(\|\Gamma\|^2 + \|\beta\|^2)$ introduced by Janzing et al. (2017). It measures the relative contribution of causal and confounding signal in terms of lengths rather than inner products.

3 Causal and Statistical Risk of High-Dimensional Regression Models

Causal learning is extremely challenging, because it requires scarcely available interventional data or has to rely on other information such as exogenous (Rothenhäusler et al., 2021) or instrumental variables (Angrist et al., 1991). In our setting where only observational data are available, causal learning requires additional model assumptions. One such approach has been followed by the Concorr method (Janzing, 2019) which leverages the ICM assumption to make an improved choice of regularization parameter under a linear regression model. To fully characterize the effect of regularization on causal generalization, we consider two estimators for learning causal models from observational data $(X, Y) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$: the min-norm interpolator and ridge regressors. The *min-norm interpolator* is the minimum l_2 norm solution to the least squares regression problem

$$\hat{\beta}_0(X, Y) := \arg\min\{\|\hat{\beta}\|_2 : \hat{\beta} \in \arg\min_{\hat{\beta} \in \mathbb{R}^d} \|Y - X\hat{\beta}\|^2\}. \tag{4}$$

A closed form is given by $\hat{\beta}_0(X, Y) = (X^T X)^+ X^T Y$. For $\lambda > 0$, the *ridge regressor* solves

$$\hat{\beta}_\lambda(X, Y) := \arg\min_{\hat{\beta} \in \mathbb{R}^d} \frac{1}{n}\|Y - X\hat{\beta}\|^2 + \lambda\|\hat{\beta}\|^2\,, \tag{5}$$

which has the explicit solution $\hat{\beta}_\lambda(X, Y) = (X^T X + n\lambda I_d)^{-1} X^T Y$. The min-norm interpolator can be obtained as a limiting case from the ridge regression solution via $\hat{\beta}_0(X, Y) = \lim_{\lambda \to 0^+} \hat{\beta}_\lambda(X, Y)$. Whenever it is clear from the context, we drop the dependence of the predictors on X and Y.

Before proceeding with the analysis, we motivate the idea that appropriate regularization can help to learn causal models from purely observational data. To this end, we compare regularization chosen by statistical cross validation to regularization based on an *interventional validation set* in Figure 2. Since cross validation implicitly assumes that there is no confounding, it is close to Bayes optimal for $\zeta = 0$ when $n \gg d$. However, as confounding increases, it falls behind regularization based on the interventional validation set. The latter even yields Bayes optimal risk again in the purely confounded setting $\zeta = 1$, where the lack of causal signal $(\beta = 0)$ is encoded by infinite regularization. While we might not have access to an interventional validation set in practice, our theory will show that knowledge of confounding strength is sufficient for choosing appropriate regularization. Finally, we want to caution that even though regularization can help, it does not remove the hardness of causal learning. Reliable causal inference still requires stronger assumptions or additional data.

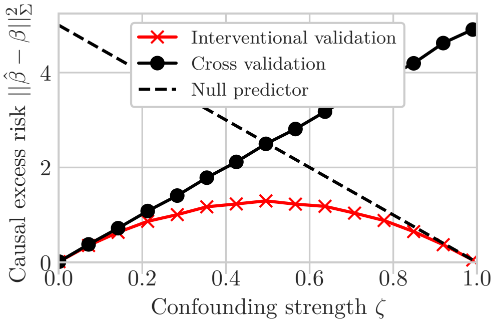


Figure 2: Causal excess risk of ridge predictors based on $n = 30,000$ samples from the observational distribution. Regularization is chosen either by cross validation or based on a validation set from the interventional distribution of same size. Each model has fixed dimensions $d = 300, l = 350$ and $\mathrm{SNR}_\mathrm{S} = 5$, but different underlying confounding strengths under the constraint $\langle \beta, \Gamma \rangle = 0$. The benefits of optimal regularization over cross validation increase with confounding strength.

3.1 Precise Asymptotics of the Causal and Statistical Risks

In this section, we provide precise asymptotics for the causal and statistical risks of the min-norm interpolator and ridge regression solutions in the high-dimensional regime. This regime is characterized by both $n, d \to \infty$ such that $d/n \to \gamma \in (0, \infty)$, where γ is called the *overparameterization ratio*. We distinguish between the *underparameterized regime* ($\gamma < 1$) and the *overparameterized regime* ($\gamma > 1$). All proofs for this section are deferred to Appendix B. Since the predictors $\hat{\beta} = \hat{\beta}(X, Y)$ are random variables in the training data X and Y, so is their corresponding causal risk. We consider the expectation of this risk under Y conditioned on X. According to Proposition 2.1, it is given by $R_X^C(\hat{\beta}) := \mathbb{E}_{Y|X} R^C(\hat{\beta}) = \mathbb{E}_{Y|X} \|\hat{\beta} - \beta\|_\Sigma^2 + \tilde{\sigma}^2 + \|\Gamma\|_\Sigma^2$. Due to its simple form, similar to the usual statistical risk, the causal excess risk can be decomposed into bias and variance:

$$\mathbb{E}_{Y|X} \|\hat{\beta} - \beta\|_\Sigma^2 = \underbrace{\|\mathbb{E}_{Y|X}\hat{\beta}_\lambda - \beta\|_\Sigma^2}_{=: B_X^C(\hat{\beta}_\lambda)} + \underbrace{\mathbb{E}_{Y|X}\|\hat{\beta}_\lambda - \mathbb{E}_{Y|X}\hat{\beta}_\lambda\|_\Sigma^2}_{=: V_X^C(\hat{\beta}_\lambda)} . \tag{6}$$

The next theorem is one of our main results. It gives a closed-form expression for the limiting causal bias and variance of the min-norm interpolator and ridge regression estimators. We make the simplifying assumption of isotropic covariance $\Sigma = I_d$. The proof relies on recent techniques from random matrix theory. It employs arguments similar to Dicker (2016), Dobriban et al. (2018), and Hastie et al. (2022) and can correspondingly be extended to arbitrary covariances under boundedness assumptions on the spectrum. We leave such extensions for future work and focus on thoroughly understanding the isotropic causal model, because it already exhibits rather rich behavior.

Theorem 3.1 (Limiting Causal Bias-Variance Decomposition for the Ridge Estimator). *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*

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

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

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 (7) and (8).

From these limiting expressions we can see that the causal risk curve of the min-norm interpolator exhibits the double descent phenomenon: it diverges at the interpolation threshold $\gamma = 1$ due to the variance term and decreases again for $\gamma > 1$. A corresponding visualization is given in Figure 4. Explicit regularization dampens the divergence of the variance term. While we are primarily interested in the causal risk, the corresponding statistical risk serves as a natural baseline. An analogue set of results for the statistical risk is given in Appendix C. These results have already been derived by Hastie et al. (2022) and can also be recovered as a special case of our causal results: for fixed statistical parameters $\tilde{\beta}$ and $\tilde{\sigma}^2$, the statistical risk coincides with the causal risk of an unconfounded causal model defined with $\beta = \tilde{\beta}$, $\sigma^2 = \tilde{\sigma}^2$, and $\alpha = 0$. In particular, the corresponding statistical limiting expressions are the same as in Theorem 3.1 after setting $\eta = \omega^2 = 0$.

Optimal statistical and causal regularization By directly optimizing the closed form expressions for limiting causal and statistical risks we can find the optimal causal and statistical regularization.

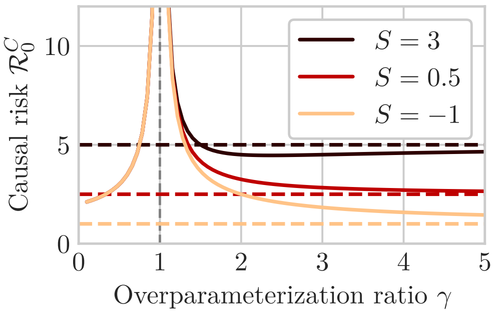


Figure 3: Limiting causal excess risk $\mathcal{R}_0^C$ (without the constant $\tilde{\sigma}^2 + \omega^2$) of the min-norm interpolator for different causal signal strengths S. Dashed lines are the corresponding null-risks ω^2, which are outperformed more often as S increases. For $\gamma < 1$, all three curves coincide.

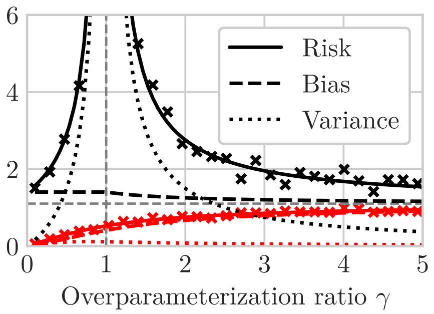


Figure 4: Limiting bias-variance decomposition and causal excess risk of the min-norm interpolator (black) and optimally regularized ridge regression (red). Crosses indicate finite-sample risks of $n = d/\gamma$ samples with $d = 300$. The finite risks are well-predicted by their theoretical limit.

For any $\gamma \in (0, \infty)$, the optimal statistical regularization $\lambda_S^*(\gamma) := \arg\inf_{\lambda \in (0,\infty)} \mathcal{R}_\lambda^S$ can be expressed in closed-form as $\lambda_S^*(\gamma) = \mathrm{SNR_S}^{-1} \gamma$. The closed-form expression for the optimal causal regularization parameter $\lambda_C^*(\gamma) := \arg\inf_{\lambda \in (0,\infty)} \mathcal{R}_\lambda^C$ is a root of a 4th order polynomial and as such considerably intricate. For readability, we do not include it here. We investigate the behavior of the optimal causal and statistical regularization in Section 4 and 5.

3.2 Basic Behavior of the Limiting Risk

We start to analyze the results by assessing the basic behavior of the limiting causal risk. The causal risk of the null estimator $\hat{\beta} = 0$ serves as a natural baseline to evaluate the performance of the the min-norm interpolator and the ridge regression estimators.

Regimes of the min-norm interpolator Theorem 3.1 characterizes the limiting causal risk of the min-norm interpolator. Its behavior is controlled by the causal signal-to-noise ratio, which we defined as $\mathrm{SNR_C} = (1 - \zeta)\,\mathrm{SNR_S}$. However, as we will later see, the causal risk of the min-norm interpolator can be lower than null risk when $\zeta < 0.5$. To distinguish the regimes of the min-norm interpolator, it is therefore convenient to consider the closely related quantity $S = (1 - 2\zeta)\,\mathrm{SNR_S}$. It distinguishes between three different regimes (visualized in Figure 3).

- For $S > 1$, the causal signal dominates the noise and the min-norm interpolator can perform better than null risk in both under- and overparameterized regime.
- For $0 \leq S \leq 1$, the causal signal is weaker than the noise. Only the underparameterized regime can beat the null risk, whereas the overparameterized regime is always worse.
- The previous two cases resemble the behavior of the statistical risk in the corresponding regimes of the statistical SNR. Contrary to the statistical risk, however, the causal risk admits a third regime $S < 0$. In this case, the min-norm interpolator always performs worse than null risk. Here, the causal signal $\langle \beta, \tilde{\beta} \rangle$ is dominated by the confounding signal $\langle \Gamma, \tilde{\beta} \rangle$, and interpolating the observational data overfits to the confounding.

Bias and variance The bias-variance decomposition of the causal risk given in Theorem 3.1 is visualized in Figure 4 for the min-norm interpolator and the optimally ridge-regularized regressor. The figure also shows the causal risk based on finite samples from the model, which is in high agreement with our asymptotic results. We compare the causal risk to the corresponding statistical risk. First note that the causal and statistical variance terms coincide exactly for both the min-norm interpolator and ridge regressors. This is because the variance term of the squared loss depends only on the variance in the training data, but not on the target parameter β or $\tilde{\beta}$. Since the training data are the same for both causal and statistical learning, the variance terms trivially coincide.

For the min-norm interpolator, as in the statistical case, the variance term is responsible for the double-descent behavior of the causal risk curve because it explodes at the interpolation threshold $\gamma = 1$ and decreases in the overparameterized regime $\gamma > 1$. In the statistical setting, the bias strictly increases in the overparameterized regime and, as a consequence, the best risk is always achieved in the underparameterized setting. In contrast, the causal bias of the min-norm interpolator can be decreasing in the overparameterized regime and therefore the optimal causal risk can be achieved in the highly overparameterized regime $\gamma \to \infty$. However, this only happens in the regime $S < 0$ where the risk of the min-norm interpolator is always worse than null risk.

Figure 4 shows the causal risk of the optimally regularized ridge regression estimator which trivially is always below that of the min-norm risk. Similar to the statistical setting, the corresponding generalization curve does not exhibit the double descent phenomenon. There are qualitatively different reasons for why regularization helps in statistical and causal learning. For both statistical and causal learning, regularization decreases the shared variance, which corresponds to the finite-sample error. However, while the statistical bias always increases with regularization, the causal bias can actually decrease. This implies that regularization not only helps with the finite-sample error, but can also reduce the error due to confounding.

Higher confounding implies higher causal risk for all λ So far, we have investigated the causal risk under a single causal model. Now we can compare different causal models using the confounding strength measure ζ introduced in Section 2.1. The next proposition shows that ζ governs the hardness of causal learning from observational data. Specifically, the causal risk of the ridge regression for any $\lambda \in (0, \infty)$ increases as the causal model becomes more confounded. A proof is given in Appendix D.

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 $\eta = 0$.

4 Benign Causal Overfitting

A large number of recent works suggest that minimum-norm interpolators can be optimal for statistical generalization (Belkin et al., 2018; Belkin et al., 2019a; Muthukumar et al., 2020). This phenomenon is often referred to as benign overfitting. Moreover, the optimal statistical generalization may even be achieved for negative regularization $\lambda < 0$ (Kobak et al., 2020; Bartlett et al., 2020; Tsigler et al., 2020). It is unclear, however, if such interpolators, which have implicit small-norm biases, can also be optimal when there is a shift between the training and test distributions. In particular, we ask: can the optimal causal regularization be 0 or even negative, that is, do we observe *benign causal overfitting?* To show that the optimal regularization can be negative, we simply show that the derivative of the causal risk at 0 is positive. We summarize our key findings in Theorem 4.1.

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.*

In the highly overparameterized regime ($\gamma \to \infty$), the benefit of explicit regularization vanishes and both the causal and statistical risks of the ridge regression estimator converge to their corresponding null risks, independently of the regularization. We do not refer to this as benign overfitting. However, we can observe benign causal overfitting when the causal SNR is larger than the statistical SNR ($\zeta < 0$), which happens when causal and statistical parameter are strongly aligned. This implies that the norm of the statistical parameter is smaller than the norm of the causal parameter. Consequentially, statistical regressors are implicitly biased towards solutions of smaller norm and causal learning exhibits self-induced regularization. Compare this to benign statistical overfitting, which happens for certain alignments between the regression parameter $\tilde{\beta}$ and the covariance matrix Σ. In our isotropic setting $\Sigma = I_d$, we can therefore never observe benign statistical overfitting, but we can observe benign causal overfitting. This phenomenon occurs in both the underparameterized as well as the overparameterized regime. The range of γ for which the optimal causal regularization is negative increases with the dominance of the causal signal over the statistical signal. As γ approaches the interpolation threshold, it becomes harder for the optimal causal regularization to be negative. When the causal SNR is smaller than the statistical SNR ($\zeta > 0$) and in particular under the ICM ($0 < \zeta \leq 1$), the optimal causal regularization is strictly positive and the benefit of explicit regularization does not vanish. This can be the case even when the optimal statistical regularization vanishes. To see this consider the statistical risk in the highly underparameterized regime $\gamma \to 0$. In this regime, the benefit of explicit regularization vanishes and the min-norm interpolator indeed achieves the optimal *statistical* risk. The optimal causal regularization is given explicitly by $\lambda_C^* = \zeta/(1-\zeta)$ for $0 \leq \zeta \leq 1$ and $\lambda_C^* = \infty$ for $\zeta > 1$. This is strictly positive and increasing in the confounding strength ζ, and in fact diverges as ζ approaches 1 (see Theorem 5.2).

5 On Optimal Regularization

In this section, we investigate two key questions which are natural in the context of our work. How does the optimal causal regularization λ_C^* compare to the optimal statistical regularization λ_S^*? What is the dependence of the optimal causal regularization λ_C^* on the confounding strength ζ?

Optimal statistical vs. causal regularization When the training and test distributions coincide, approaches such as cross-validation or information criteria (for example AIC or BIC) can be used to estimate the regularization parameter for optimal out-of-sample generalization. However, choosing the correct regularization parameter for causal learning can be challenging without interventional data. To understand the optimal causal regularization, it is natural to compare it to the optimal statistical regularization, which can usually be estimated from data. Interestingly, our analysis reveals that when confounding strength is positive $\zeta > 0$ and in particular under the ICM one needs to regularize more strongly for causal generalization than for statistical generalization. However, when the confounding strength is negative, that is, when the causal signal dominates the statistical signal, the optimal causal regularization λ_C^* can actually be smaller than the optimal statistical regularization λ_S^*. We formally present this result in Theorem 5.1.

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*

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

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.

Dependence on confounding strength ζ The problem of causal learning from observational data is a distribution shift problem where the distribution of the training data is shifted from that of the test distribution. As discussed earlier in Proposition 3.2, the confounding strength measure ζ quantifies the strength of this distribution shift. Therefore, we expect the optimal causal regularization to increase with confounding strength. Theorem 5.2 indeed confirms this intuition.

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 $\zeta \geq 1$.

6 Summary and Extensions

We characterize the role of explicit regularization for causal learning from observational data by computing the asymptotic risk of ridge-regularized regressors and the min-norm interpolator (Theorem 3.1). Under the principle of independent causal mechanisms (ICM), we find that causal learning requires stronger regularization than statistical learning (Theorem 5.1). A practical implication is that the regularization parameter for causal learning should be chosen larger than what is suggested by cross-validation. We can precisely state how much larger based on an estimate of confounding strength (Janzing et al., 2017; Janzing et al., 2018). Beyond ICM, we show that strong alignments between causal and statistical parameters can cause self-induced regularization and lead to benign causal overfitting (Theorem 4.1). One could consider generalizing our assumptions: arbitrary covariances, shifts in the marginal distributions of covariates, soft interventions, more complex hypothesis classes, or non-linear causal relationships. Since the linear model already exhibits rich behavior, we focus in this paper on understanding the simple setting. Below, we briefly discuss extensions of our analysis to causal learning under soft interventions, non-linearity, and non-Gaussianity.

Soft interventions It is not always appropriate to consider causal learning under hard interventions. Instead, it is often of interest to consider *soft interventions*. In these settings, the qualitative statements derived from our analysis still hold. To illustrate this, we consider the class of shift interventions where the structural dependence of the covariates x is not destroyed as in the case of hard interventions but the observed covariates are merely perturbed (i.e., interventions of the form $do(x := x + \nu)$). Then it turns out that Causal risk$_{\text{soft}}$ $=$ Causal risk$_{\text{hard}}$ $+$ Statistical risk. From our results, it then follows that under ICM, $\lambda^{\text{statistical}} \leq \lambda^{\text{causal}}_{\text{soft}} \leq \lambda^{\text{causal}}_{\text{hard}}$ This also supports our intuition since under soft interventions, we typically aim to achieve a tradeoff between statistical and causal predictability. We include a complete analysis under shift interventions in Appendix F.

Extensions to non-linear models It is feasible to extend the analysis to structural causal models that arise in a reproducing kernel Hilbert space corresponding to a positive definite kernel (i.e, where the best statistical model $\tilde{f}$ and the best causal model f are functions in some RKHS). There are two major technical challenges to deriving the theoretical analysis in such non-linear settings. Both are beyond what can be done in this paper and are left for future work, but we briefly outline them below.

1. **Extend the definition of confounding strength ζ beyond the linear setting.** Since such a definition is non-trivial already in the linear setting, it is challenging to meaningfully generalize this to the non-linear setting. However, under non-linear causal models in the RKHS, we can naturally extend this definition by replacing the Euclidean norms with functional norms in the RKHS. Generalizing the analysis beyond this setting would require further careful consideration.
2. **Derive limiting expressions for causal risk of regularized regressors in a non-linear hypothesis class.** In the case of kernel regression, this would still be feasible via recent random matrix theory results [27]. By optimizing the limiting expressions with respect to the regularization parameter, one can obtain the parameter that achieves the optimal causal risk and subsequently identify the relationship between optimal causal regularization and confounding strength.

Beyond Gaussianity The analysis can be extended beyond the Gaussian setting by considering random variables generated by finite mixtures of Gaussians. 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 Appendix G. They show that for distributions with finite 4th moments, the finite-sample risks of the min-norm interpolator and causally optimally regularized ridge regressor closely match the theoretically derived asymptotic risks.

Acknowledgments and Disclosure of Funding

This work has been supported by the German Research Foundation through the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1:390727645) and the BMBF Tübingen AI Center (FKZ: 01IS18039A). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Leena C Vankadara and Luca Rendsburg.

References

Angrist, Joshua D and Alan B Keueger (1991). “Does compulsory school attendance affect schooling and earnings?” *The Quarterly Journal of Economics*.

Bai, Zhidong and Jack W Silverstein (2010). *Spectral analysis of large dimensional random matrices*. Springer.

Bartlett, Peter L et al. (2020). “Benign overfitting in linear regression”. *Proceedings of the National Academy of Sciences*.

Belkin, Mikhail (2021). “Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation”. *Acta Numerica*.

Belkin, Mikhail, Siyuan Ma, and Soumik Mandal (2018). “To understand deep learning we need to understand kernel learning”. *International Conference on Machine Learning (ICML)*.

Belkin, Mikhail, Alexander Rakhlin, and Alexandre B Tsybakov (2019a). “Does data interpolation contradict statistical optimality?” *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Belkin, Mikhail et al. (2019b). “Reconciling modern machine-learning practice and the classical bias–variance trade-off”. *Proceedings of the National Academy of Sciences*.

Blanchet, Jose, Yang Kang, and Karthyek Murthy (2019). “Robust Wasserstein profile inference and applications to machine learning”. *Journal of Applied Probability*.

Byrd, Jonathon and Zachary Lipton (2019). “What is the effect of importance weighting in deep learning?” *International Conference on Machine Learning (ICML)*.

Dicker, Lee H (2016). “Ridge regression and asymptotic minimax estimation over spheres of growing dimension”. *Bernoulli*.

Dobriban, Edgar and Stefan Wager (2018). “High-dimensional asymptotics of prediction: Ridge regression and classification”. *The Annals of Statistics*.

Donhauser, Konstantin et al. (2021). “Interpolation can hurt robust generalization even when there is no noise”. *Advances in Neural Information Processing Systems (NeurIPS)*.

Feldman, Vitaly (2020). “Does learning require memorization? a short tale about a long tail”. *Symposium on Theory of Computing (STOC)*.

Gao, Rui, Xi Chen, and Anton J Kleywegt (2017). “Distributional robustness and regularization in statistical learning”. *arXiv preprint arXiv:1712.06050*.

Gulrajani, Ishaan and David Lopez-Paz (2021). “In Search of Lost Domain Generalization”. *International Conference on Learning Representations (ICLR)*.

Hachem, Walid, Philippe Loubaton, and Jamal Najim (2007). “Deterministic equivalents for certain functionals of large random matrices”. *The Annals of Applied Probability*.

Hastie, Trevor et al. (2022). “Surprises in high-dimensional ridgeless least squares interpolation”. *The Annals of Statistics*.

Janzing, Dominik (2019). “Causal Regularization”. *Advances in Neural Information Processing Systems (NeurIPS)*.

Janzing, Dominik and Bernhard Schölkopf (2010). “Causal Inference Using the Algorithmic Markov Condition”. *IEEE Transactions on Information Theory*.

– (2017). “Detecting Confounding in Multivariate Linear Models via Spectral Analysis”. *Journal of Causal Inference*.

– (2018). “Detecting non-causal artifacts in multivariate linear regression models”. *International Conference on Machine Learning (ICML)*.

Kobak, Dmitry, Jonathan Lomond, and Benoit Sanchez (2020). “The Optimal Ridge Penalty for Real-world High-dimensional Data Can Be Zero or Negative due to the Implicit Ridge Regularization.” *Journal of Machine Learning Research (JMLR)*.

Kuhn, Daniel et al. (2019). “Wasserstein distributionally robust optimization: Theory and applications in machine learning”. *Operations research & management science in the age of analytics*.

Lemeire, Jan and Dominik Janzing (2013). “Replacing Causal Faithfulness with Algorithmic Independence of Conditionals”. *Minds and Machines*.

Liang, Tengyuan and Alexander Rakhlin (2020). “Just interpolate: Kernel “ridgeless” regression can generalize”. *The Annals of Statistics*.

Marčenko, Vladimir A and Leonid Andreevich Pastur (1967). “Distribution of eigenvalues for some sets of random matrices”. *Mathematics of the USSR-Sbornik*.

Mei, Song and Andrea Montanari (2021). “The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve”. *Communications on Pure and Applied Mathematics*.

Muthukumar, Vidya et al. (2020). “Harmless interpolation of noisy data in regression”. *IEEE Journal on Selected Areas in Information Theory*.

Pearl, Judea (2009). “Causal inference in statistics: An overview”. *Statistics surveys*.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Rothenhäusler, Dominik et al. (2021). “Anchor regression: Heterogeneous data meet causality”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Rubio, Francisco and Xavier Mestre (2011). “Spectral convergence for a general class of random matrices”. *Statistics & probability letters*.

Sagawa, Shiori et al. (2020). “An investigation of why overparameterization exacerbates spurious correlations”. *International Conference on Machine Learning (ICML)*.

Shafieezadeh Abadeh, Soroosh, Peyman M Mohajerin Esfahani, and Daniel Kuhn (2015). “Distributionally robust logistic regression”. *Advances in Neural Information Processing Systems (NeurIPS)*.

Shafieezadeh-Abadeh, Soroosh, Daniel Kuhn, and Peyman Mohajerin Esfahani (2019). “Regularization via mass transportation”. *Journal of Machine Learning Research (JMLR)*.

Silverstein, Jack W (1995). “Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices”. *Journal of Multivariate Analysis*.

Tsigler, Alexander and Peter L Bartlett (2020). “Benign overfitting in ridge regression”. *arXiv preprint arXiv:2009.14286*.

Vankadara, Leena Chennuru et al. (2021). “Causal Forecasting: Generalization Bounds for Autoregressive Models”. *arXiv preprint arXiv:2111.09831*.

Xu, Huan, Constantine Caramanis, and Shie Mannor (2009). “Robustness and Regularization of Support Vector Machines.” *Journal of Machine Learning Research (JMLR)*.

Zhang, Chiyuan et al. (2021). “Understanding deep learning (still) requires rethinking generalization”. *Communications of the ACM*.

Zhu, Shixiang et al. (2020). “Distributionally Robust Weighted k-Nearest Neighbors”. *arXiv preprint arXiv:2006.04004*.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] While our analysis is exhaustive, we paid special attention to the additional ICM assumption described at the end of Section 2. Further possible generalizations of our model are discussed in Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Right before Section 3.1, we cautioned against using our results on optimal causal regularization as an excuse to dismiss the hardness of causal learning.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs are deferred to the appendix and referenced correspondingly in the main paper.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

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

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

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 $\mathbb{E}_Z \|Z - \mathbb{E}_Z Z\|_A^2 = \mathrm{Tr}\left[\mathrm{Cov}\, Z \cdot A\right]$.

Proof. Let $\mathbb{E} := \mathbb{E}_Z$ and $\mu := \mathbb{E}Z$. 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 $\Sigma_Z := \mathbb{E}ZZ^T$ 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

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

□

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 $V_X^C(\hat{\beta}_\lambda) = \frac{\tilde{\sigma}^2}{n}\mathrm{Tr}[\hat{\Sigma}(\hat{\Sigma} + \lambda I_d)^{-2}\Sigma]$. The empirical covariance matrix of X is denoted by $\hat{\Sigma} := X^TX/n$.

Proof. Recall that $R^C_X(\hat{\beta}_\lambda) = \mathbb{E}_{Y|X}\left\|\hat{\beta}_\lambda - \beta\right\|^2_\Sigma$. 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 $B^C_X(\hat{\beta}_\lambda)$ and $V^C_X(\hat{\beta}_\lambda)$ then simply follow from plugging in expectation and covariance of $\hat{\beta}_\lambda$:

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

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

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

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 $\sigma^2_{\tilde{\epsilon}} = \tilde{\sigma}^2$. 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*

$$\begin{aligned}B^C_X(\hat{\beta}_\lambda) &\to \mathcal{B}^C_\lambda := \omega^2 + \tilde{r}^2\lambda^2 m'(-\lambda) - 2(\omega^2 + \eta)\lambda m(-\lambda) \quad \textit{and} &(7)\\ V^C_X(\hat{\beta}_\lambda) &\to \mathcal{V}^C_\lambda := \tilde{\sigma}^2\gamma(m(-\lambda) - \lambda m'(-\lambda)), &(8)\end{aligned}$$

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 $R^C_X(\hat{\beta}_\lambda) \to \mathcal{R}^C_\lambda := \mathcal{B}^C_\lambda + \mathcal{V}^C_\lambda + \tilde{\sigma}^2 + \omega^2$. 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

$$B^C_X(\hat{\beta}_0) \to \mathcal{B}^C_0 = \begin{cases}\omega^2, & \gamma < 1 \\ \omega^2 + (r^2 - \omega^2)(1 - \frac{1}{\gamma}), & \gamma > 1\end{cases}, \quad V^C_X(\hat{\beta}_0) \to \mathcal{V}^C_0 = \begin{cases}\tilde{\sigma}^2\frac{\gamma}{1-\gamma}, & \gamma < 1 \\ \tilde{\sigma}^2\frac{1}{\gamma-1}, & \gamma > 1\end{cases}.$$

Therefore $R^C_X(\hat{\beta}_0) \to \mathcal{R}^C_0 = \mathcal{B}^C_0 + \mathcal{V}^C_0 + \tilde{\sigma}^2 + \omega^2$.

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

Limiting expressions for causal bias

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

First, let us consider the sequence of functions given by

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

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 $\widehat{m(z)}$ 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 $\lambda, \gamma \in (0, \infty)$,*

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

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

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

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

□

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

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

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 $\partial_\lambda \mathcal{R}^C_\lambda(\lambda^*_C(\gamma)) = 0$, or equivalently,*

$$0 = \lambda^*_C(\gamma) - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda^*_C(\gamma) + \gamma - \sqrt{\varphi(\lambda^*_C(\gamma))}\right)\varphi(\lambda^*_C(\gamma))\,,$$

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 $\mathrm{SNR_S} = \tilde{r}^2/\tilde{\sigma}^2$. The derivative of the limiting causal risk $\mathcal{R}^C_\lambda$ in λ is given by

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

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

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

   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

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

   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 $\partial_\lambda \mathcal{R}^C_\lambda \leq 0$. Rearranging for ζ and using Lemma E.1 yields the equivalent condition

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

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 $\partial_\lambda \mathcal{R}^C_\lambda(0) < 0$ and the derivative $\partial_\lambda \mathcal{R}^C_\lambda$ for sufficiently large λ is positive. This together then implies that the minimum $\lambda^*_C(\gamma)$ of the function $\mathcal{R}^C_\lambda$ is indeed attained at a finite value in $(0, \infty)$, and $\mathcal{R}^C_\lambda$ satisfies the critical point condition $\partial_\lambda \mathcal{R}^C_\lambda(\lambda^*_C(\gamma)) = 0$.

   For the derivative at 0, assume that the converse is true, that is, $\partial_\lambda \mathcal{R}^C_\lambda(0) \geq 0$. Rearranging this condition for ζ yields similarly to the first case of this lemma that $\zeta \leq f(0, \gamma, \mathrm{SNR_S})$. However Lemma E.1 states that f is increasing in λ, which means that this condition already implies $\zeta \leq f(\lambda, \gamma, \mathrm{SNR_S})$ 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 $\lambda - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{2\gamma}\left(1 + \lambda + \gamma - \sqrt{\varphi(\lambda)}\right)\varphi(\lambda)$. 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

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

   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

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

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

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

$\lambda_C^* = 0$ *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

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

Rearranging this equation yields

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

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

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

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 $y|do(x' := x+\nu)$ 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 $\mathrm{SNR_S} = \tilde{r}^2/\tilde{\sigma}^2$. The derivative of the limiting causal risk under shift interventions $\mathcal{R}_\lambda^{C_{\text{soft}}}$ in λ is given by

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

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

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

   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

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

   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

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

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 $2\zeta \leq f(0,\gamma,\mathrm{SNR_S})$. However Lemma E.1 states that f is increasing in λ, which means that this condition already implies $\zeta \leq 2f(\lambda,\gamma,\mathrm{SNR_S})$ 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 $\lambda - \mathrm{SNR_S}^{-1}\gamma - \frac{\zeta}{4\gamma}\left(1+\lambda+\gamma-\sqrt{\varphi(\lambda)}\right)\varphi(\lambda)$. 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

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

   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* $\lambda^*_C \le \lambda^*_{C_{soft}} \le \lambda^*_S$.

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

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

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 $-2\,\mathrm{SNR_S}^{-1}\frac{\gamma\max\{1,\gamma\}}{(1-\gamma)^2} < \zeta < 2$. In this regime according to Lemma E.2, the optimal causal regularization λ^*_C satisfies the critical point condition

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

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

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

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

G Beyond Gaussianity

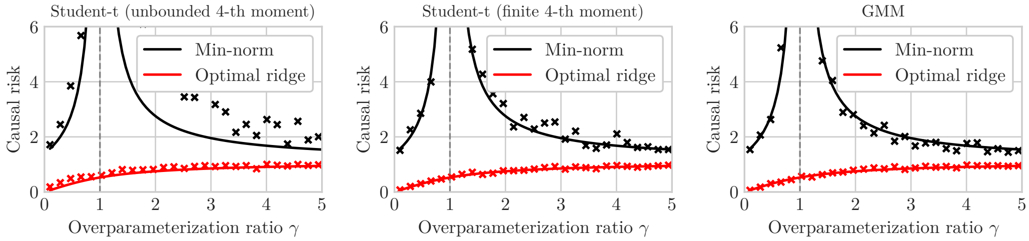


Figure 5: Causal risk of the minimum norm l_2 interpolator and the (causally)optimally regularized ridge regressor under a student-t distribution with unbounded 4th moments (3 degrees of freedom, left), a student-t distribution with bounded 4th moments (10 degrees of freedom, middle), a mixture of Gaussians (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.