We thank the reviewers for their valuable comments and appreciate that all three reviewers acknowledge that we honestly discuss limitations of directly applying our mainly theoretical insights (which two reviewers find stimulating).

Reviewer 1: "The authors do not provide a way [...] to check how much confounding still exists or what to do about it; or what assumptions are required for their method to completely identify the causal estimands." This would indeed be desirable but it's probably one of the hardest problems of causal inference (if confounders are unobserved!).

"In that case there is no need for confounding adjustment in the asymptotic regime because the expected value of $\mathbf{Zc} = 0$." Note that the parameters M and $\mathbf{c}$ are only drawn *once* for every data set. If $\mathbf{c}$ is drawn from a Gaussian with high variance, it induces arbitrarily large confounding bias. Note, as an aside, that population Ridge and Lasso maximize $p(\mathbf{a}|\Sigma_{\mathbf{XX}}, \Sigma_{\mathbf{X}Y})$ regardless of the distribution of M, it could even be any fixed matrix.

"There is a bit of a discordance across some of the theoretical results [...] In section 5, however, they switch gears to talk about the finite sample efficiency of the Y estimate." This may be a misunderstanding too. Section 5 is not about finite sample effects. It just motivates our learning theory via the analogy between inferring the interventional loss from the observations loss (our goal) to inferring the population loss from the empirical loss (the usual goal of statistical learning theory).

"One paradoxical finding that the authors arrive at is that when $\mathbf{X} = \mathbf{Z}$, the "deconfounding" task becomes harder. This is a bit surprizing since this describes the case where there is no hidden confounding (all the unobserved confounders are observed if X is observed)." The case $\mathbf{X} = \mathbf{Z}$ is not trivial for two reasons. First, we do not assume that the observer *knows* the mixing matrix M. Second, even if he/she does, it is still unclear what part of the correlations are due to the causal vector $\mathbf{a}$ and what part due to the confounding part $\mathbf{c}$ (the joint covariance matrix of $\mathbf{X}$ and $\mathbf{Z}$ is degenerate).

"... I believe the statement on line 139 about the distribution unregularized regression vector is not accurate in the general case." Note that the statement relies on our set of assumptions, for which it should be true.

Yes, figure 2, 2nd from the left, shows that cross-validation is not better than no regularization because the bias is dominated by confounding and finite sample effects are almost irrelevant.

Consistency statements were not in our focus. On the one hand, the tight analogy between both scenarios suggests that those could be transferred across them without novel insights. On the other hand, we had to drop some further theoretical results already to satisfy the space constraints. Our focus is the analogy.

Reviewer 2: We can briefly comment on *why* combining finite sample with confounding is theoretically challenging. Regarding related work, we will try to also add a few words, although it's not clear what to cut out instead. If the reviewers feel that some other explanations are too lengthy, we are open to suggestions. We will improve readability of the figures and their explanations in the text and also the abstract.

Derivations 93-98: $\hat{E} \sim \mathcal{N}(0, \sigma_E^2 I)$ by assumption. Thus, $\widehat{\Sigma_{\mathbf{X}E}} = \hat{\mathbf{X}}^T \hat{E} \sim \mathcal{N}(0, \sigma^2 \hat{\mathbf{X}}^T \hat{\mathbf{X}}) = \mathcal{N}(0, \sigma_E^2 \widehat{\Sigma_{\mathbf{XX}}})$. For scenario 2: $\mathbf{c} \sim \mathcal{N}(0, \sigma_c^2 I)$ and thus $\Sigma_{\mathbf{X}E} = M^T \mathbf{c} \sim \mathcal{N}(0, \sigma_c^2 M^T M) = \mathcal{N}(0, \sigma_c^2 \Sigma_{\mathbf{XX}})$. Every point in figure 2 represents one of the 100 runs. Unregularized RSE means RSE with OLS estimator (which coincides with β in the population limit).

Reviewer 3: "Does the adjusted ridge penalty always end up being larger than the cross-validation chosen one?... " Interesting question. For strong confounding, CV certainly under-regularized for causal purposes. In the regime where the bias of $\mathbf{a}$ is dominated by finite sampling we would rather trust CV instead of claiming that it over-regularizes just because our adjusted penalty regularizes less. After all, estimation of β remains a bottleneck. We agree with the intuition that sparse models behave even better for Lasso, and some preliminary experiments confirmed that intuition, but we didn't elaborate on it. We also thought about using information criteria for selecting λ, but for scenario 2 this would require the unobserved number ℓ. We also derived statements on selecting λ by cross-validation across *different environments*, but we had to drop that due to lack of space.

"Should we expect nonlinear models to also require stronger regularization?" We believe so. The following high level view may help: (1) Our influence of the confounder depends on multiple independently drawn parameters (the entries of $\mathbf{c}$). (2) Due to a concentration of measure effect, the regression loss is likely to be close to its average over the distribution of parameters (all $\mathbf{c}$) – uniformly over $\mathcal{F}$ if $\mathcal{F}$ is 'small'. (3) In our case, the average loss over all distribution of parameters coincides with the interventional loss. Thus, the loss for a typical draw of $\mathbf{c}$ is close to the interventional loss. Nonlinear models satisfying (1) can easily satisfy (2) if the regression loss depends *weakly* on each of the confounding parameters. Then, Efron-Stein entails the same concentration of measure. How to achieve (3) in nonlinear models is less obvious, however.

We have assumed $n > d$ just to get a concise description (e.g. unique minimum in (2)), the analogy between scenario 1 and 2 still holds for $n < d$ and $\ell < d$.