We thank all reviewers for the positive reviews. Overall, there is a consensus among the reviewers that our work is deemed appropriate for publication at NeurIPS.

- R2 advocated that "*the paper has the potential of conveying the message of causality into the wider machine learning community and thereby trigger other ideas in this area.*"
- R3 highlighted the strengths of the paper as "*[c]oncrete analysis of dual formulation [and] [c]lear explanation of the roadmap towards the idea of the working algorithm.*"
- R4 wrote that "*[...] I find [the contribution] very interesting and definitely relevant for the NeurIPS community.*"

In the camera-ready version, we will take all comments into consideration. Below we respond to major concerns.

The interpretation of the do$(\cdot)$ operator (R3,R4). This concept is at the core of our paper, so we thank the reviewers for raising this point, which should have been made clearer in the initial submission. In our camera-ready version, we will provide additional discussion around this concept.

In Line 54 of our submission, $\text{do}(X = x)$ denotes a mathematical operator which simulates physical interventions by setting the value of X to x, while keeping the rest of the model unchanged [Pearl, 2009, Sec. 3.2.1]. In this work, we aim to estimate $\mathbb{E}[Y \mid \text{do}(X = x)]$ which is a conditional expectation computed w.r.t. the *interventional* distribution $P(Y \mid \text{do}(X = x))$. We can estimate $P(Y \mid \text{do}(X = x))$ if it is possible to manipulate X and then observe the resulting outcome Y. In Figure 1, for example, one may assign different levels of education to people and then observe the resulting levels of income when they enter the labor market. Unfortunately, such experiment is not always possible and we only have access to an *observational* distribution $P(Y \mid X = x)$, which can be different from $P(Y \mid \text{do}(X = x))$. In this example, the discrepancy results from the unobserved socioeconomic status, as illustrated in Figure 1.

Scalability of DualIV (R2). The scalability is an important aspect that was not adequately addressed in our submission. We thank the reviewer for pointing it out. We will discuss it in more detail in our camera-ready version.

To improve scalability of our method, we can employ the stochastic gradient descent (SGD) based algorithms similar to those proposed in Dai et al. [2017, Algorithm 1] (also cited in our submission) to solve the dual formulation, i.e., Eq. (7) in the submission. Based on SGD, other models such as deep neural networks can also be used to parametrize the function classes $\mathcal{F}$ and $\mathcal{U}$ in Eq. (7). Furthermore, we can improve the scalability of the kernelized DualIV algorithm (i.e., Algorithm 1 in our submission) by leveraging rich literature on large-scale kernel machines such as random Fourier feature (RFF) and Nyström method; see, e.g., Yang et al. [2012] and references therein.

Responses to reviewers' remaining questions.

- (R3) In Line 130, we stated that it is "cumbersome to solve (4) ..." because solving (4) requires *two-stage* estimation. In the first stage, $\mathbb{E}_{X|Z}[\cdot]$ must be estimated, which is challenging on its own because we only observe a single sample x_i from $P(X|Z = z_i)$ for each value z_i. The first-stage estimate is then used in the second stage to estimate $\mathbb{E}_{YZ}[\cdot]$ in Eq. (4). Our contribution is precisely to reformulate the problem such that one can solve the problem in a single step by estimating $\mathbb{E}_{XYZ}[\cdot]$ directly, as we did in Eq. (6).
- (R4) Page 7, Line 243: The notation $^\top$ translates to the inner product in RKHS $\mathcal{F}$ and $\mathcal{U}$. With slight abuse of notation, we define $\Phi := [\phi(x_1), \dots, \phi(x_n)]$ and $\Upsilon := [\varphi(y_1, z_1), \dots, \varphi(y_n, z_n)]$. Hence, we have

$$\Phi^\top \Phi = \begin{bmatrix} \langle \phi(x_1), \phi(x_1)\rangle_{\mathcal{F}} & \cdots & \langle \phi(x_n), \phi(x_1)\rangle_{\mathcal{F}} \\ \vdots & \ddots & \vdots \\ \langle \phi(x_1), \phi(x_n)\rangle_{\mathcal{F}} & \cdots & \langle \phi(x_n), \phi(x_n)\rangle_{\mathcal{F}} \end{bmatrix} = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_n, x_1) \\ \vdots & \ddots & \vdots \\ k(x_1, x_n) & \cdots & k(x_n, x_n) \end{bmatrix} = \mathbf{K}.$$

  The matrix $\Upsilon^\top \Upsilon$ is defined similarly. We will clarify this notation in our camera-ready version.

References

B. Dai, N. He, Y. Pan, B. Boots, and L. Song. Learning from Conditional Distributions via Dual Embeddings. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 1458–1467. PMLR, 2017.

J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

T. Yang, Y.-f. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems 25*, pages 476–484. 2012.