1. SUPPLEMENTARY MATERIAL

1.1. HYPERPARAMETER SETUP

Since our Bayesian Warped GP consists of various components, there are a couple of hyperparameters that can be optimized, see Table (1). We represent the bijective mapping g_ϕ of the Normalizing flow by a neural spline flow transform with element-wise (referred to as bins) rational conditional spline functions, to represent the conditional distributions in the causal model. Each conditional spline transform consists of a dense neural network with three Bayesian linear layers and a RELU activation function. In this neural network the hidden dimensions (`hidden dims`) need to be set (implemented with Pyro (Bingham et al., 2019)). Furthermore, the spline is defined in a bounding box (`bounds`), which should cover the range of input data, for details see Durkan et al. (2019). To relax this requirement, we normalize the input data.

According to our variational inference scheme, we can optimize further parameters affecting the training: the number of monte carlo samples (`S`) to be drawn, the prior variance (`prior var`), the learning rate (`lr`), and the training steps (`steps`). We optimize these variables in the seven variable setup to minimize the MMD on a held-out validation dataset of size 250 (generated from the ground truth SCM). For the three variable setup we optimized the hyperparameters w.r.t. the cost due to time-constraints. In both cases, we used the BOHB (Bayesian Optimization algorithm using Hyperband) algorithm (Falkner et al., 2018) for optimization. More specifically, we used the python ray-tune package of Liaw et al. (2018) as implementation of BOHB.

Table 1: Optimal Hyperparameters found with BOHB on a validation set for each SCM and classifier setting.

	LINEAR SCM	NON-LINEAR SCM	NON-ADDITIVE	LINEAR LOG. REGR.	NON-LINEAR LOG. REGR.	RANDOM FOREST
`bounds`	6	1	10	27	3	21
`hidden dims`	10	13	40	2	6	27
`lr`	0.03	0.03	0.01	0.04	0.008	0.05
`steps`	5719	5719	4501	6982	6198	4956
`S`	15	21	20	31	24	21
`prior var`	0.1	0.1	0.05	0.03	0.01	0.02

1.2. INTERVENTIONAL DISTRIBUTION

In this section, we provide additional plots, illustrating the properties of the different models visually. Due to the complex yet low-dimensional setting, the non-additive SCM of the three-variable model (see Tab.1) is of particular interest. Since this was visually not the case for the other SCMs, we do not explicitly show them. In Fig. 1 we plotted the ground truth distribution (see Fig. 1a) and in Fig. 1b the corresponding distribution as modeled by the BW-GP. To generate samples from the different models, we generated samples from the ground truth model for the variable of the root-node X_1. Using the different models for the conditional distributions of children given the parents, we generated the remaining variables X_2, X_3 according to the causal graph. As also indicated by the small MMD-Values (cf. Tab.1), the BW-GP also visually fits the ground truth data much better than the GP (see Fig. 2a), which learns two Gaussian distributions for the multimodal distribution. While the CVAE in Fig. 2b fits the data also better than the GP, it exhibits a higher variance than the BW-GP, which is also reflected by a slightly larger MMD-value.

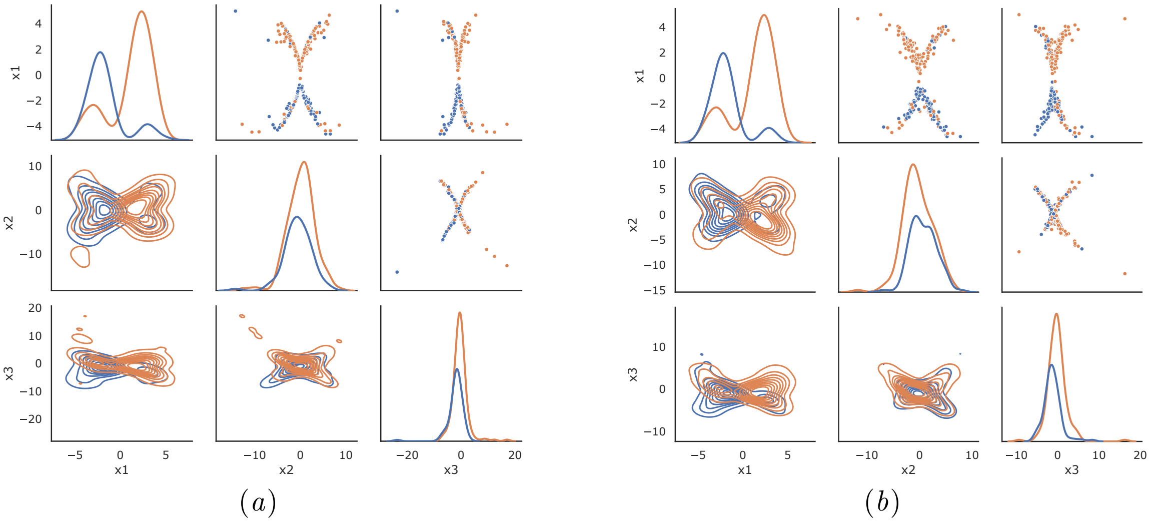

Figure 1: In (a) we plot the ground truth and in (b) the Bayesian Warped Gaussian Process model. The coloring corresponds to the classes the classifier yields in the recourse task (blue are the negatively and orange the positive classified points).

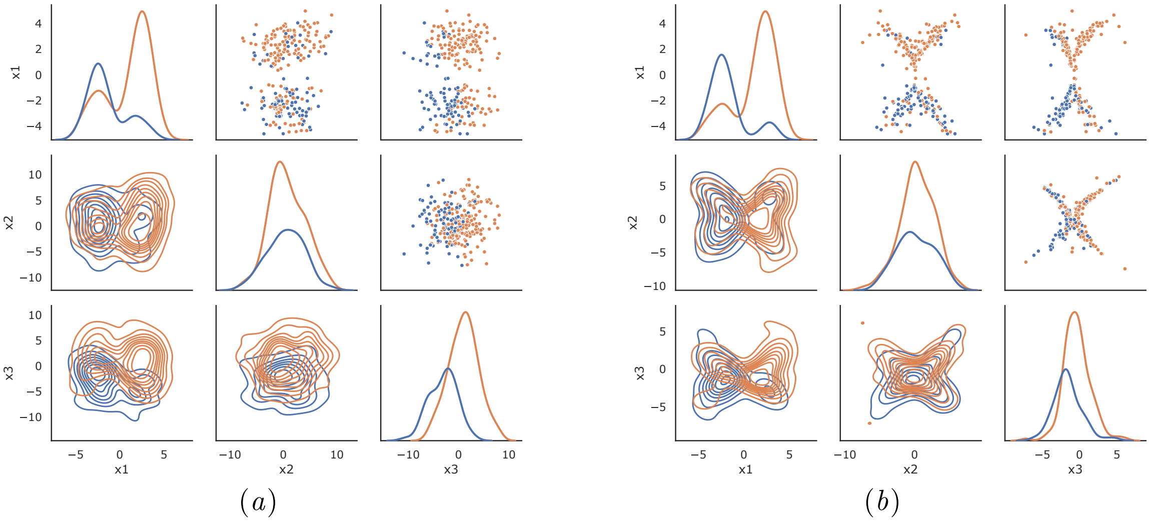

Figure 2: In (a) the Gaussian Process is plotted and in (b) the CVAE model. The coloring corresponds to the classes the classifier yields in the recourse task (blue are the negatively and orange the positive classified points).

References

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20, 2019.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: robust and efficient hyperparameter optimization at scale. *CoRR*, abs/1807.01774, 2018.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.