Result

Global setting of experiments:
start_recording_step = 1000
stop_recording_step = 15000
n_rollout = 5

1. Bias and Variance analysis

   (a) Parametrized Model

   From the nature of the parametrized model, we can obtain a analytical solution of the policy gradient by solving the stationary distribution of (s, a, s) triple from the parameter π in the parametrized model. From this "true value", we can measure the bias of the estimator from each method. For it's not making straightforward sense to measure a bias for vector, we used cosine similarity as the directional bias, and which is more coming into handy when working on actual optimization problem with used with gradient descent.

   Table 1: Directional Bias of each Estimate vs the True Value
   (measured by cosine similarity)

Benchmark	-0.10798715
Variant Control: Causality	-0.105943285
Variant Control: Baseline	0.015888477
IHP1	0.8471538

   Note that this the "directional bias" of estimates from all benchmark methods are subject to be changed by different initialization of model's parameter π. But the pattern that these benchmark happened to have a similar direction with the "True Value" is completely random and of a low ratio (9 times/60 experiments). But the $ihp1$ performs a consistent low directional bias (57 times/60 experiments)

   To measure the variance of the policy gradient, we provided multiple pespectives: in table 2, the variance is analyzed from the following angles:

   - tr(cov): the trace of the covariance matrix i.e. the general variance
   - mean(l2-norm): the magnitude of the policy gradient estimation
   - std(l2-norm): the variance from the magnitude perspective
   - tr(cov(normed)): the trace of covariance matrix of normalized gradient, the variance from the direction perspective

   From the measurement above, the variance of ihp1 method overperformed all the opponents from different magnitudes – the advantage is often more than a order of 10.

Table 2: Variance analysis on each estimates: parameter model

	tr(cov)	mean(l2_norm)	std(l2_norm)	tr(cov(normalized))
Benchmark	2.2811e+1	4.5671e+0	9.0563e-1	1.0310e+0
Variant Control: Causality	6.3045e+0	2.5340e+0	3.9562e-1	9.5016e-1
Variant Control: Baseline	4.7190e-2	1.7521e-1	1.4286e-1	9.8227e-1
IHP1	**1.3853e-4**	**3.8117e-2**	**4.2127e-3**	**7.9728e-2**

For we have the analytical solution of the policy gradient, we can provide a more straightforward perspective about the bias and variance entry-wisely by comparing the distribution of each estimate with the true value on each specific dimension. Thus we have figure 1, which gives a clear illustration to the advantage of *ihp*1 method.

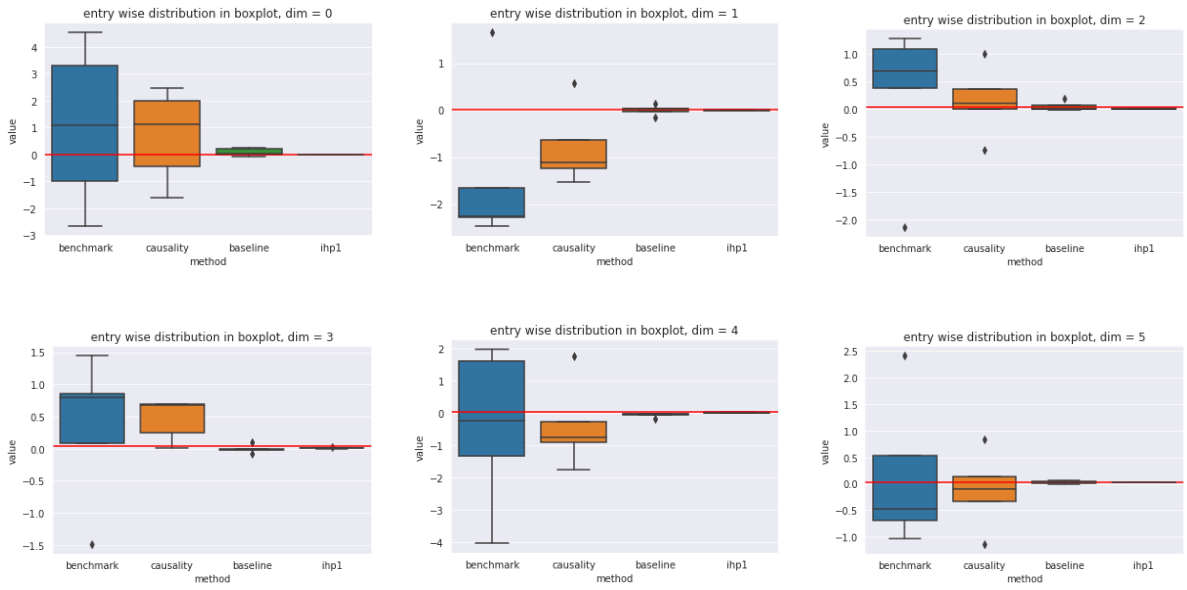


Figure 1: Dimension-wise Boxplot, red line is the theoretical true value

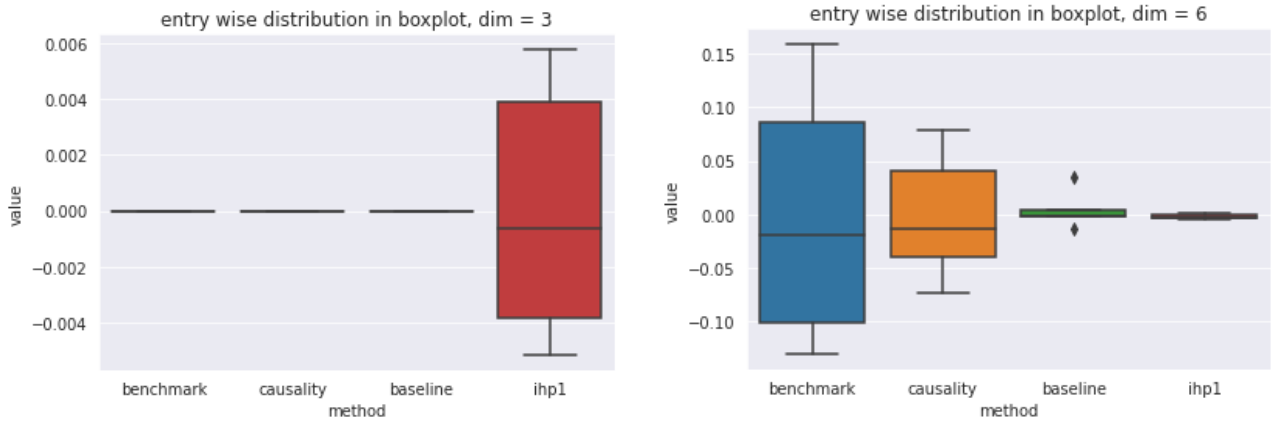
(b) Neural Network Based Model

When working with Neural Network based model, we can no longer obtain the closed form of the policy gradient. But other measurements remains valid and conveys same information as the parametrized model. See table 3 and figure 2.

What should be noticed is that in the neural-net-based model, there are some dimension whose estimate from Benchmark, Variant Control: Causality, and Variant Control: Baseline are all zero, and therefore have 0 variance(see Figure 2a). However, one should also noticed that the magnitude of value on vertical axis is way smaller, i.e. it doesn't introduce much variance to the variance of ihp. Meanwhile, in the experiments with different neural network based models, the number of zero entries are never more than 25% from our observation.

Table 3: Variance analysis on each estimates: Neural-Network-Based model

	tr(cov)	mean(l2_norm)	std(l2_norm)	tr(cov(normalized))
Benchmark	3.0473e-1	5.0623e-1	1.6671e-1	1.1509e+0
Variant Control: Causality	1.3137e-1	2.9891e-1	1.5419e-1	9.2969e-1
Variant Control: Baseline	6.1244e-3	6.0158e-2	5.1098e-2	1.1904e+0
IHP1	**2.7747e-4**	**1.8444e-2**	**2.0253e-3**	**7.9622e-1**



(a) The dimension with zero terms

(b) The dimension doesn't have zero terms

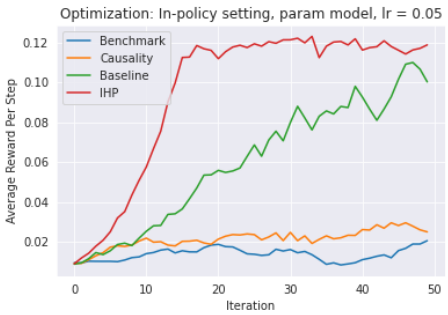
Figure 2: Dimension-wise Boxplot, the left one stand for the dimension with all benchmarks are 0, the right one stand for the dimension benchmarks are not 0

2. Optimization performance

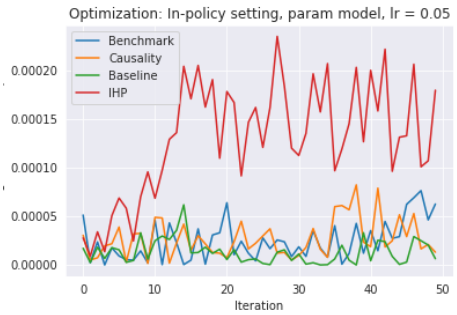
   Global Setting:
   normalized_gradient = $True$
   learning_rate_for_normalization = 0.05

   (a) In-policy Optimization During experiments with different model parameter π initialization and learning rate (0.1, 0.05, 0.025, 0.01), $ihp1$ performed an consistent advantage (19/20 times) in both average return (discount constant = 1) and discounted return (discount constant = 0.95) case. See figure 3.

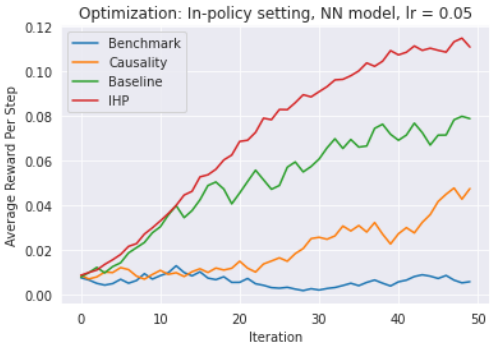
   (b) Off-policy Optimization During experiments with different initialization and different learning rate(0.1,0.05,0.025,0.01), $ihp1$ performed an consistent advantage (for 20/20 times) in both average return(See figure 4a,4c) and discounted return case(see figure 4b, 4d).



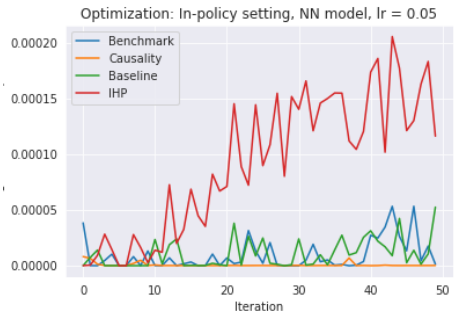
(a) parameter model



(b) parameter model with discounting

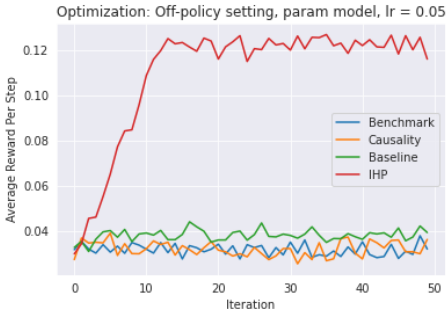


(c) NN model

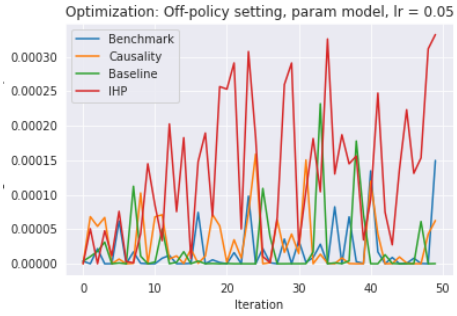


(d) NN model with discounting

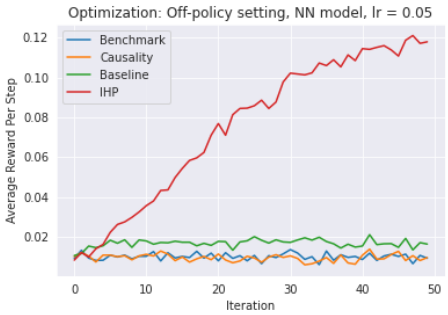
Figure 3: In-Policy Optimization



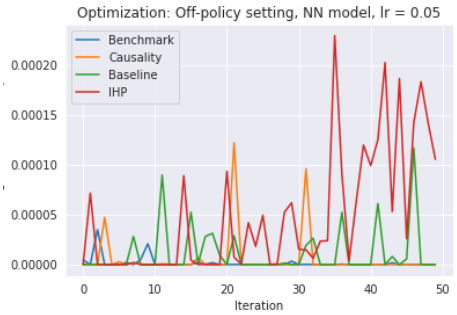
(a) parameter model



(b) parameter model with discounting



(c) NN model



(d) NN model with discounting

Figure 4: Off-Policy Optimization