Table 1: Evaluation performance of different methods in two tasks, where boldfaced values mean **valid coverage** for $c(\hat{z})$ (i.e., within [0.85, 0.95]), IW and bias for each procedure are presented in unit $\times 10^3$ and each entry is averaged over 200 experiment repetitions. We do not list the evaluation procedure of Ridge in Regression problem and SAA in CVaR-Portfolio problem since they are parametric models and K-fold CV and plug-in achieve valid coverage guarantees for these models.

method	n	Plug-in		5-CV		10-CV		20-CV	
-	-	cov90	IW (bias)	cov90	IW (bias)	cov90	IW (bias)	cov90	IW (bias)
Regression Problem ($d_x = 10, d_y = 1$)									
Forest	1200	0.77	2.71 (0.40)	0.62	2.83 (-0.42)	0.65	2.81 (-0.48)	0.69	2.74 (-0.30)
	2400	0.77	1.77 (0.26)	0.66	1.83 (-0.37)	0.60	1.76 (-0.29)	0.64	1.76 (-0.23)
	4800	**0.86**	1.19 (0.15)	0.47	1.24 (-0.32)	0.56	1.26 (-0.38)	0.63	1.24 (-0.18)
	9600	**0.85**	0.73 (0.11)	0.42	0.76 (-0.28)	0.52	0.73 (-0.22)	0.51	0.74 (-0.17)
	19200	**0.86**	0.47 (0.07)	0.34	0.49 (-0.23)	0.42	0.48 (-0.13)	0.45	0.48 (-0.11)
kNN $n^{2/3}$	1200	0.81	2.48 (0.12)	0.76	2.57 (-0.46)	0.77	2.51 (-0.46)	0.76	2.48 (-0.21)
	2400	0.80	1.63 (0.08)	0.78	1.68 (-0.38)	0.72	1.62 (-0.29)	0.74	1.71 (-0.24)
	4800	0.84	1.11 (0.03)	0.66	1.14 (-0.37)	0.70	1.15 (-0.21)	0.68	1.14 (-0.18)
	9600	**0.88**	0.75 (0.04)	0.61	0.77 (-0.28)	0.69	0.70 (-0.15)	0.70	0.69 (-0.12)
	19200	**0.85**	0.46 (0.04)	0.49	0.47 (-0.23)	0.66	0.46 (-0.09)	0.68	0.46 (-0.04)
CVaR-Portfolio Optimization ($d_x = 5, d_y = 5$)									
kNN $n^{1/4}$	2400	0.00	0.172 (1.716)	0.76	0.337 (0.079)	0.82	0.328 (0.035)	**0.85**	0.329 (0.028)
	4800	0.00	0.126 (1.423)	0.74	0.221 (0.038)	0.79	0.218 (0.034)	0.80	0.217 (0.025)
	9600	0.00	0.094 (1.108)	0.66	0.147 (0.029)	0.72	0.146 (0.028)	0.76	0.144 (0.021)

Table 2: Evaluation performance of NCV versus CV for the forest learner in the regression problem, where boldfaced values mean **valid coverage** for $c(\hat{z})$ (i.e., within [0.85, 0.95]), IW and bias for each procedure are presented in unit $\times 10^2$ and each entry is averaged over 200 experiment repetitions.

n	Plug-in			5-CV			5-NCV			10-CV			10-NCV		
	cov90	IW	bias	cov90	IW	bias	cov90	IW	bias	cov90	IW	bias	cov90	IW	bias
10000	**0.88**	7.27	0.35	0.41	7.53	-4.36	0.81	15.06	-1.13	0.48	7.41	-2.65	0.80	13.86	-1.07
30000	**0.86**	3.73	0.22	0.38	3.82	-2.48	0.74	7.65	-0.56	0.29	3.80	-2.21	0.65	7.59	-1.15
100000	**0.90**	1.71	0.11	0.19	1.75	-2.25	0.53	3.51	-0.48	0.27	1.73	-1.26	0.56	3.45	-0.61
300000	**0.88**	0.92	0.08	0.13	0.93	-1.76	0.47	1.87	-0.28	0.26	0.92	-0.56	0.4	1.85	-0.31

*Plug-in, 5-CV, 10-CV correspond to the intervals (2) in our paper. For NCV, we use Algorithm 1 in [8] (following their practical MSE restriction in Section 4.3.2) and bias estimation in Appendix C in [8] to construct the interval for NCV (5-NCV and 10-NCV), i.e.,$(\widehat{Err}^{(NCV)} - \widehat{bias} - z_{1-\alpha/2}\sqrt{\widehat{MSE}}, \widehat{Err}^{(NCV)} - \widehat{bias} + z_{1-\alpha/2}\sqrt{\widehat{MSE}})$.

Table 3: Evaluation performance of **high-dimensional** ridge regression (100 repetitions)

(n, p)	Plug-in		5-CV		LOOCV	
	cov90	bias	cov90	bias	cov90	bias
(300, 200)	0.00	21.63	0.00	-48.72	0.59	-0.38
(900,600)	0.00	4.62	0.00	-8.45	0.51	-0.03
(1800, 1200)	0.00	2.81	0.00	-5.75	0.58	-0.01