

A Reproductive Kernel Hilbert Space and Kernel Mean Matching

A.1 RKHS

Let the tuple $(\mathcal{F}, \kappa, \mathcal{Z})$ denote a **reproductive kernel Hilbert space** (RKHS) \mathcal{F} on the sample space \mathcal{Z} with kernel κ . The RKHS is a Hilbert space of functions $f : \mathcal{Z} \rightarrow \mathbb{R}$, on which the inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ satisfies the reproducing property:

$$\langle f, \kappa(z, \cdot) \rangle_{\mathcal{F}} = f(z), \quad \forall f \in \mathcal{F}, z \in \mathcal{Z}.$$

That said, the evaluation of function f on a single point z can be viewed as an inner product between function f and the evaluation operator $\kappa(z, \cdot)$. We define the **feature map** $\phi : \mathcal{Z} \rightarrow \mathcal{F}$ by $\phi[z] := \kappa(z, \cdot)$. Throughout this part, we will use the square bracket $[\cdot]$ to denote mappings to functional spaces. Based on the reproducing property, the feature map satisfies $\langle \phi[z], \phi[z'] \rangle_{\mathcal{F}} = \kappa(z, z')$ for all $z, z' \in \mathcal{Z}$.

The **kernel mean embedding** of a probability distribution \mathcal{P}_z [41], denoted by $\mu_{\mathcal{F}}[\cdot]$, is a mapping from the space of all the probability distributions on \mathcal{Z} to \mathcal{F} . The mapping is given by

$$\mu_{\mathcal{F}}[\mathcal{P}_z] := \mathbb{E}_{Z \sim \mathcal{P}_z} [\phi[Z]].$$

The kernel κ is called *characteristic* if the kernel mean embedding is injective: for $\mathcal{P}_z \neq \mathcal{P}'_z$, it holds that $\mu_{\mathcal{F}}[\mathcal{P}_z] \neq \mu_{\mathcal{F}}[\mathcal{P}'_z]$. And following [19], the operator $\mu_{\mathcal{F}}[\cdot]$ is bijective if κ is a universal kernel in the sense of [35]. The core idea of the **kernel mean matching** is estimate \mathcal{P}_z with a $\hat{\mathcal{P}}_z$ satisfying $\mu_{\mathcal{F}}[\mathcal{P}_z] \approx \mu_{\mathcal{F}}[\hat{\mathcal{P}}_z]$.

A.2 Cross-covariance Operator

Consider the joint random variable $(Z, C) \in \mathcal{Z} \times \mathcal{C}$. Define two RKHS's by $(\mathcal{F}, \kappa, \mathcal{Z})$ and $(\mathcal{G}, \eta, \mathcal{C})$ respectively. Let \mathcal{P} denote the joint distribution of (Z, C) and $\mathcal{P}_z, \mathcal{P}_c$ their marginal distributions respectively, then the **cross-covariance operator** $\mathcal{A}_{Z,C}$ (with the dependency on \mathcal{P} omitted) is defined as [1]

$$\mathcal{A}_{Z,C} := \mathbb{E}_{(Z,C) \sim \mathcal{P}} [\phi[Z] \otimes \psi[C]] - \mathbb{E}_{Z \sim \mathcal{P}_z} [\phi[Z]] \otimes \mathbb{E}_{C \sim \mathcal{P}_c} [\psi[C]].$$

In the following, we will omit the explicit distribution of the random variable in the subscript of \mathbb{E} and replace expressions like $\mathbb{E}_{(Z,C) \sim \mathcal{P}}$ with $\mathbb{E}_{Z,C}$. The operator $\mathcal{A}_{Z,C}[\cdot]$ can be viewed as a mapping from \mathcal{G} to \mathcal{F} in the following way: by noting that $\langle \psi[C], g \rangle_{\mathcal{G}} = g(C)$ for any $g \in \mathcal{G}$, we can define $\mathcal{A}_{Z,C}[g]$ as

$$\mathcal{A}_{Z,C}[g] := \mathbb{E}_{Z,C} [g(C) \cdot \phi[Z]] - \mathbb{E}_C [g(C)] \cdot \mathbb{E}_Z [\phi[Z]].$$

For any functions $f \in \mathcal{F}$ and $g \in \mathcal{G}$, the cross-covariance operator has the following property:

$$\langle f, \mathcal{A}_{Z,C}[g] \rangle_{\mathcal{F}} = \mathbb{E}_{Z,C} [f(Z) \cdot g(C)] - \mathbb{E}_Z [f(Z)] \cdot \mathbb{E}_C [g(C)],$$

which exactly corresponds to the covariance between $f(Z)$ and $g(C)$. The **conditional embedding operator** $\mathcal{U}_{Z|C}$ is a mapping from \mathcal{G} to \mathcal{F} such that, for any $c \in \mathcal{C}$, the follow equation holds:

$$\mathcal{U}_{Z|C}[\psi(c)] = \mu_{\mathcal{F}}[\mathcal{P}_{z|c}]. \quad (13)$$

In other words, $\mathcal{U}_{Z|C}$ maps the feature map $\psi(c)$ to the kernel mean embedding of the conditional distribution $Z|C = c$. Following [34], if the cross-covariance operator $\mathcal{A}_{C,C}$ is invertible, by defining

$$\mathcal{U}_{Z|C} := \mathcal{A}_{Z,C} \mathcal{A}_{C,C}^{-1}$$

the equation (13) is satisfied. To see this, it is sufficient to show that $\langle f, \mathcal{A}_{Z,C} \mathcal{A}_{C,C}^{-1} [\psi[c]] \rangle_{\mathcal{F}} = \langle f, \mu_{\mathcal{F}}[\mathcal{P}_{z|c}] \rangle_{\mathcal{F}}$ for all $f \in \mathcal{F}$, and this holds following the derivations below,

$$\begin{aligned} \langle f, \mu_{\mathcal{F}}[\mathcal{P}_{z|c}] \rangle_{\mathcal{F}} &= \mathbb{E}_{z|c} [f(Z)|c] \\ &= \langle \mathbb{E}_{z|c} [f(Z)|C], \psi[c] \rangle_{\mathcal{G}} \\ &= \langle \mathcal{A}_{C,C} [\mathbb{E}_{z|c} [f(Z)|C]], \mathcal{A}_{C,C}^{-1} [\psi[c]] \rangle_{\mathcal{G}} \\ &= \langle \mathcal{A}_{C,Z} [f], \mathcal{A}_{C,C}^{-1} [\psi[c]] \rangle_{\mathcal{G}} \\ &= \langle f, \mathcal{A}_{Z,C} \mathcal{A}_{C,C}^{-1} [\psi[c]] \rangle_{\mathcal{F}}. \end{aligned}$$

Equation (13) directly implies the following property, which serves as the key step of the KMM procedure in the density ratio estimation method of [41]:

$$\mathcal{U}_{Z|C}[\mu_{\mathcal{G}}[\mathcal{P}_c]] = \mu_{\mathcal{F}}[\mathcal{P}_z].$$

A.3 Empirical Estimations

In this section, we briefly outline how to estimate the quantities above using i.i.d. samples $(z_i, c_i)_{i=1}^N$ drawn from \mathcal{P} . By Mercer's theorem, the feature map $\phi[z]$ can be represented as a column vector in a (possibly infinite-dimensional) Hilbert space. We use $\phi[z]\psi[c]^\top$ to denote the outer product $\phi[z] \otimes \psi[c]$. Define the matrices $\Phi := (\phi[z_1], \dots, \phi[z_N])$ and $\Psi := (\psi[c_1], \dots, \psi[c_N])$. Further define $\mathbf{K}, \mathbf{H} \in \mathbb{R}^{N \times N}$ with $K_{i,j} = \kappa(z_i, z_j)$ and $H_{i,j} = \eta(c_i, c_j)$. The empirical estimators are indicated by adding a hat symbol $\hat{\cdot}$ to the original quantities, with their explicit expressions shown below:

$$\begin{aligned}\mu_{\mathcal{F}}[\mathcal{P}_z] &\approx \hat{\mu}_{\mathcal{F}} := \frac{1}{N} \sum_{i=1}^N \phi(z_i), \\ \mu_{\mathcal{G}}[\mathcal{P}_c] &\approx \hat{\mu}_{\mathcal{G}} := \frac{1}{N} \sum_{i=1}^N \psi(c_i), \\ \mathcal{A}_{Z,C} &\approx \hat{\mathcal{A}}_{Z,C} := \frac{1}{N} (\Phi - \hat{\mu}_{\mathcal{F}} \mathbf{1}^\top) (\Psi - \hat{\mu}_{\mathcal{G}} \mathbf{1}^\top)^\top \\ &= \frac{1}{N} \Phi \left(I - \frac{1}{N} \mathbf{1} \mathbf{1}^\top \right) \Psi^\top, \\ \mathcal{A}_{C,C}^{-1} &\approx \hat{\mathcal{A}}_{C,C}^{-1} := N \cdot \Psi \mathbf{H}^{-1} \left(I - \frac{1}{N} \mathbf{1} \mathbf{1}^\top \right)^{-1} \mathbf{H}^{-1} \Psi^\top, \\ \mathcal{U}_{Z|C} &\approx \hat{\mathcal{U}}_{Z|C} := \Phi \mathbf{H}^{-1} \Psi^\top.\end{aligned}$$

B Illustrative Example

Consider the illustrative example:

$$\min_x \text{VaR}_\alpha^{\mathcal{Q}}(c \cdot x | z) \quad \text{s.t.} \quad -1 \leq x \leq 1, \quad (14)$$

where $\alpha \in (0.5, 1)$ and $c = z + \epsilon$. In the training distribution \mathcal{P} , the z and ϵ are independent and respectively follow $\mathcal{N}(0, \sigma_1^2)$ and $\mathcal{N}(0, \sigma_2^2)$. We list the explicit expressions for the following probability distributions:

$$\begin{aligned}\mathcal{P}_z &= \mathcal{N}(0, \sigma_1^2), \quad \mathcal{P}_c = \mathcal{N}(0, \sigma_1^2 + \sigma_2^2), \\ \mathcal{P}_{c|z} &= \mathcal{N}(z, \sigma_2^2), \quad \mathcal{P}_{z|c} = \mathcal{N}\left(\frac{\sigma_1^2 \cdot c}{\sigma_1^2 + \sigma_2^2}, \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right), \\ (c, z) &\sim \mathcal{P} = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 \end{pmatrix}\right).\end{aligned}$$

We describe the explicit form of \mathcal{Q} under the following two distribution shift scenarios:

- Covariate shift: let s represent the extent of covariate shift. The distribution of z is shifted from $\mathcal{P}_z = \mathcal{N}(0, \sigma_1^2)$ to $\mathcal{Q}_z = \mathcal{N}(s, \sigma_1^2)$, while the conditional distribution of $c|z$ remains to be $\mathcal{N}(z, \sigma_2^2)$. The joint distribution of (c, z) is shifted to $\mathcal{Q} = \mathcal{N}\left(\begin{pmatrix} s \\ s \end{pmatrix}, \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 \end{pmatrix}\right)$.
- Label shift: let s represent the extent of label shift. The distribution of c is shifted from $\mathcal{P}_c = \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$ to $\mathcal{Q}_c = \mathcal{N}(s, \sigma_1^2 + \sigma_2^2)$, while the conditional distribution $z|c$ remains to be $\mathcal{N}\left(\frac{\sigma_1^2 \cdot c}{\sigma_1^2 + \sigma_2^2}, \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)$. It follows that the distribution of z in \mathcal{Q} is shifted to $\mathcal{N}\left(\frac{\sigma_1^2 \cdot s}{\sigma_1^2 + \sigma_2^2}, \sigma_1^2\right)$, and the distribution of ϵ is shifted to $\mathcal{N}\left(\frac{\sigma_2^2 \cdot s}{\sigma_1^2 + \sigma_2^2}, \sigma_2^2\right)$. The joint distribution of (c, z) is shifted to $\mathcal{Q} = \mathcal{N}\left(\begin{pmatrix} s \\ \frac{\sigma_1^2 \cdot s}{\sigma_1^2 + \sigma_2^2} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 \end{pmatrix}\right)$.

The solution to (14) admits a simple form under the idealized setting, as given in (9) and restated below:

$$x^* = \begin{cases} 1, & \text{VaR}_\alpha^{\mathcal{Q}}(c|z) \leq 0, \\ -1, & \text{VaR}_{1-\alpha}^{\mathcal{Q}}(c|z) \geq 0, \\ 0, & \text{otherwise,} \end{cases} \quad (15)$$

The probability of yielding an over-conservative solution is

$$\mathbb{P}(x^* = 0) = \mathbb{P}_{z \sim \mathcal{Q}_z}(\text{VaR}_{1-\alpha}(c|z) \leq 0 \leq \text{VaR}_\alpha(c|z)). \quad (16)$$

Specifically, in the covariate shift setting, the probability equals

$$\mathbb{P}(x^* = 0) = \Phi\left(\frac{\sigma_2}{\sigma_1}\Phi^{-1}(\alpha) - \frac{s}{\sigma_1}\right) - \Phi\left(-\frac{\sigma_2}{\sigma_1}\Phi^{-1}(\alpha) - \frac{s}{\sigma_1}\right),$$

and in the label shift setting, the probability also equals

$$\mathbb{P}(x^* = 0) = \Phi\left(\frac{\sigma_2}{\sigma_1}\Phi^{-1}(\alpha) - \frac{s}{\sigma_1}\right) - \Phi\left(-\frac{\sigma_2}{\sigma_1}\Phi^{-1}(\alpha) - \frac{s}{\sigma_1}\right).$$

Worst-case Approach Alternative to our density ratio-based approach is the worst-case approach. Construct the worst-case ball:

$$\mathcal{B}(R) := \left\{ \mathcal{N}\left(r, \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 \end{pmatrix}\right) : \|r\| \leq R \right\}.$$

Following the discussions in the main context, the objective of the worst-case approach is given by (11) and restated below:

$$\min_x \max_{\mathcal{Q} \in \mathcal{B}(R)} \text{VaR}_\alpha^\mathcal{Q}(c \cdot x|z) \quad \text{s.t.} \quad -1 \leq x \leq 1.$$

The probability of yielding an over-conservative solution is

$$\mathbb{P}(x^* = 0) = \mathbb{P}_{z \sim \mathcal{Q}_z}(\text{VaR}_{1-\alpha}(c|z) \leq 0 \leq \overline{\text{VaR}}_\alpha(c|z)).$$

In the covariate shift setting, we have $R = \sqrt{2}s$ in order that $\mathcal{B}(R)$ covers \mathcal{Q} . In the label shift setting, $R = \sqrt{1 + \frac{\sigma_1^4}{(\sigma_1^2 + \sigma_2^2)^2}} \cdot s$. The explicit form of the conservative probability is

$$\begin{aligned} \mathbb{P}(x^* = 0) &= \Phi\left(\frac{\sigma_2}{\sigma_1}\Phi^{-1}(\alpha) + \frac{\sqrt{2} \cdot s}{\sigma_1} - \frac{s}{\sigma_1}\right) \\ &\quad - \Phi\left(-\frac{\sigma_2}{\sigma_1}\Phi^{-1}(\alpha) - \frac{\sqrt{2} \cdot s}{\sigma_1} - \frac{s}{\sigma_1}\right) \end{aligned}$$

for the covariate shift setting and

$$\begin{aligned} \mathbb{P}(x^* = 0) &= \Phi\left(\frac{\sigma_2}{\sigma_1}\Phi^{-1}(\alpha) + \sqrt{\frac{1}{\sigma_1^2} + \frac{\sigma_1^2}{(\sigma_1^2 + \sigma_2^2)^2}} \cdot s - \frac{\sigma_1 \cdot s}{\sigma_1^2 + \sigma_2^2}\right) \\ &\quad - \Phi\left(-\frac{\sigma_2}{\sigma_1}\Phi^{-1}(\alpha) - \sqrt{\frac{1}{\sigma_1^2} + \frac{\sigma_1^2}{(\sigma_1^2 + \sigma_2^2)^2}} \cdot s - \frac{\sigma_1 \cdot s}{\sigma_1^2 + \sigma_2^2}\right) \end{aligned}$$

for the label shift setting.

Figure 1 is plotted by setting $\sigma_1 = \sigma_2 = 1$.

C More Experiments and Experiment Details

We consider several implementations of the three components: \hat{f} , \hat{h} and \hat{w} for our Algorithm 1. For the expectation predictor \hat{f} which predicts $\mathbb{E}[c|z]$ with z , we consider using Lasso regression, random forest, and neural network. On different training datasets, the regularization parameter λ of the Lasso regression is selected as the optimal parameter in range $[0, 4]$; for the random forest searches the best number of trees in range $[100, 1000]$ and the best depth in range $[10, 80]$; the neural network has a single middle layer with 16 neurons, with the learning rate set to 0.01 and the training epochs set to 500. For the quantile predictor \hat{h} , we implement linear quantile regression, gradient boosting regression, and neural network, all trained by minimizing the quantile loss. For the density estimator, we mainly consider estimators for the covariate shift setting, including the trivial estimator, the kernel mean matching method introduced

in [16], and the probabilistic classification method introduced in [5]. The computational complexity of the kernel mean matching method scales at least quadratically with the data size, so we only sample 200 data from the training distribution and 200 data from the test distribution to run the algorithm.

The performance metrics we consider include the coverage rate and the α -quantile of the objective $c^\top x$. Specifically, we evaluate the marginal and conditional coverage rates of the uncertainty sets generated from different implementations of our algorithm, and how different they are from the target coverage rate. The α -quantile of the objective, denoted by $\text{VaR}_\alpha^\mathcal{Q}(c^\top x|z)$, is evaluated by generating 100 samples of c from the underlying conditional distribution $\mathcal{Q}_{c|z}$ and then calculate the empirical α -quantile of $c^\top x$.

C.1 Additional Experiments on the Simple Example

This section presents additional experimental results for the simple optimization problem discussed in Section 5. In Figure 4, we experiment with different choices of \hat{f} and \hat{h} , and different covariate dimensions d . Specifically, the implementations of \hat{f} and \hat{h} follow the setup of Figure 2, and the d 's we consider include $d = 2, 4, 8$. Both the mean squared error and the quantile loss are evaluated on the test dataset, with the c values known. The results indicate that the algorithm tends to perform better, as reflected by a lower conservative probability, when the prediction models \hat{f} and \hat{h} have lower prediction errors. Further, for both \hat{f} and \hat{h} , neural networks (NN) generally achieve strong predictive performance.

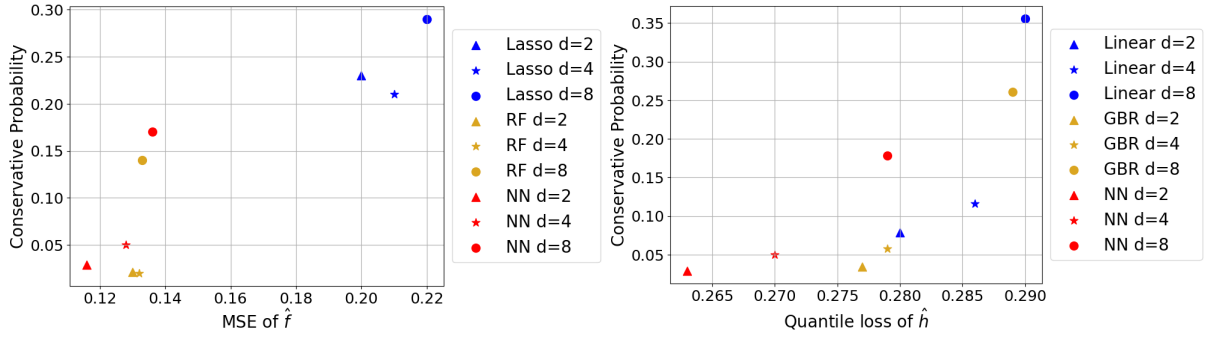


Figure 4: The quality of the predictors affects the performance of Algorithm 1. In the left figure, the x-axis represents the mean squared error (MSE) of \hat{f} on the test dataset, while in the right figure, the x-axis represents the quantile loss of \hat{h} . For both metrics, lower values indicate higher predictor quality. The y-axis shows the probability of obtaining an over-conservative solution at $x^* = 0$, where a lower probability indicates better algorithm performance. Both figures illustrate that improved prediction models enhance the performance of the algorithm.

C.2 Shortest Path Problem and Fractional Knapsack Problem

We test our algorithm on two practical LP settings. The risk level α is fixed to 0.8 and we consider the covariate shift of the test distribution.

The *shortest path problem* seeks a path between two vertices in a weighted graph such that the total accumulated cost along the path's edges is minimized. Specifically, we consider the shortest path problem on a 5×5 grid with 25 nodes and 40 edges. The objective is to find the shortest path from the top-left node to the bottom-right node. For $i = 0, \dots, 4$ and $j = 0, \dots, 4$, let (i, j) denote the node on the i -th row and j -th column, with the top-left node located at the coordinate $(0, 0)$ and the bottom-right node at the coordinate $(5, 5)$. The tuple $((i, j), (i', j'))_e$ denotes the edge between node (i, j) and (i', j') . Let V denote the set of all nodes and E denote all edges. For $((i, j), (i', j'))_e \in E$, use $c_{(i, j), (i', j')}$ to represent the cost of the edge $((i, j), (i', j'))_e$ (for the undirected graph that we consider, set $c_{(i, j), (i', j')} = c_{(i', j'), (i, j)}$). The decision variables are $x_{(i, j), (i', j')}$ for all $((i, j), (i', j'))_e \in E$, where $x_{(i, j), (i', j')} = 1$ denotes a directed path segment from the node (i, j) to the node (i', j') . The shortest path problem then has the following

risk-sensitive LP formulation:

$$\begin{aligned} \min_x \text{VaR}_\alpha & \left(\sum_{((i,j),(i',j')) \in E} c_{(i,j),(i',j')} \cdot x_{(i,j),(i',j')} \right) \\ \text{s.t.} \quad & \sum_{(i',j') : ((i,j),(i',j')) \in E} (x_{(i,j),(i',j')} - x_{(i',j'),(i,j)}) = \begin{cases} 1 & (i,j) = (0,0) \\ -1 & (i,j) = (4,4) \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (17)$$

Extending to the OOD robust formulation, we assume that the dynamic of the cost vector $c \in \mathbb{R}^{40}$ is controlled by the covariate $z \in \mathbb{R}^d$ (we fix $d = 10$ in this experiment) through

$$\mathcal{P}_{c|z} \sim \left(\left(\frac{1}{\sqrt{d}} \Theta z + 3 \right)^5 + 1 \right) \circ \epsilon,$$

where $\Theta \in \mathbb{R}^{40 \times d}$ is a 0-1 matrix with each entry generated independently from a Bernoulli(0.5) distribution. The Θ matrix is fixed once it is generated. The \circ symbol denotes an elementwise multiplication. Each element of the random vector $\epsilon \in \mathbb{R}^{40}$ is generated independently from $\text{Uniform}(\frac{3}{4}, \frac{5}{4})$. In the training data, the covariate z is generated from $\mathcal{P}_z = \mathcal{N}(0, \mathbf{I}_d)$, and in the test data z is generated from $\mathcal{Q}_z = \mathcal{N}(\mathbf{1}_d, \mathbf{I}_d)$. The objective of the OOD formulation replaces the $\text{VaR}_\alpha(\cdot)$ part in (17) with $\text{VaR}_\alpha^\mathcal{Q}(\cdot)$.

The *fractional knapsack problem* models the case where customers select items that maximize their utility, under a budget constraint. We consider a simple setting with 20 items. The decision variables $x = (x_1, \dots, x_{20})$ denote the fractions (in range $[0, 1]$) of items to purchase, $c \in \mathbb{R}^{20}$ denote the item utilities, $p \in \mathbb{R}^{20}$ denote the price of items, and $B > 0$ the total budget. The fractional knapsack problem has the following risk-sensitive LP formulation:

$$\begin{aligned} \min_x \text{VaR}_\alpha & (-c^\top x) \\ \text{s.t.} \quad & p^\top x \leq B \\ & x_i \in [0, 1], \quad i = 1, \dots, 20. \end{aligned} \quad (18)$$

The OOD robust formulation considers covariate $z \in \mathbb{R}^d$ (with $d = 10$) and assumes the conditional distribution of $c|z$ to be

$$\mathcal{P}_{c|z} \sim (\Theta z)^2 \circ \epsilon,$$

where $\Theta \in \mathbb{R}^{20 \times d}$ is a 0-1 matrix with each entry generated independently from a Bernoulli(0.5) distribution. Each element of $\epsilon \in \mathbb{R}^{20}$ is generated independently from $\text{Uniform}(\frac{4}{5}, \frac{6}{5})$. The training distribution of the covariate z is $\mathcal{P}_z = \mathcal{N}(0, \mathbf{I}_d)$ and the test distribution is $\mathcal{Q}_z = \mathcal{N}(\mathbf{1}_d, \mathbf{I}_d)$.

In Figure 5, we compare our algorithm against multiple benchmarks on the two LP settings above. As there is no existing algorithm that handles covariate shift in the risk-sensitive setting, we just implement the existing benchmark methods and evaluate them in the test environment. The ‘‘Ellipsoid’’ method ignores the contextual information and calibrates the ellipsoid to achieve an empirical coverage rate of α on the training samples. The ‘‘DCC’’ and ‘‘IDCC’’ algorithms are proposed in [9]. The ‘‘kNN’’ algorithm is a conditional robust optimization method proposed in [29]. The ‘‘Ours-Trivial’’ and ‘‘Ours’’ both implement our Algorithm 1 (with \hat{f} and \hat{h} being neural networks). ‘‘Ours-Trivial’’ sets a trivial density ratio estimator $\hat{w} \equiv 1$, while ‘‘Ours’’ uses the probabilistic classification method (which is shown to enjoy the best performance based on the previous experiments) to estimate \hat{w} . The result demonstrates that our algorithm generally outperforms the rest benchmarks, and leveraging the density ratio information further improves the performance on the test dataset.

D Proof of Theorems

D.1 Proof of Theorem 1

Let $N := |\mathcal{D}_2|$ and define $x_i := (c_i, z_i)$ for $i = 1, \dots, N$. The final η term produced from Algorithm 1 is a function of (x_1, \dots, x_N) , which we define below as the $\hat{\eta}(\cdot)$ function

$$\hat{\eta}(x_1, \dots, x_N) := \min \left\{ \eta \geq 0 : \sum_{i=1}^N w(x_i) \cdot \mathbb{1}\{|\hat{f}(z_i) - c_i| \leq \eta \cdot \hat{h}(z_i)\} \geq \alpha \cdot \sum_{i=1}^N w(x_i) \right\}.$$

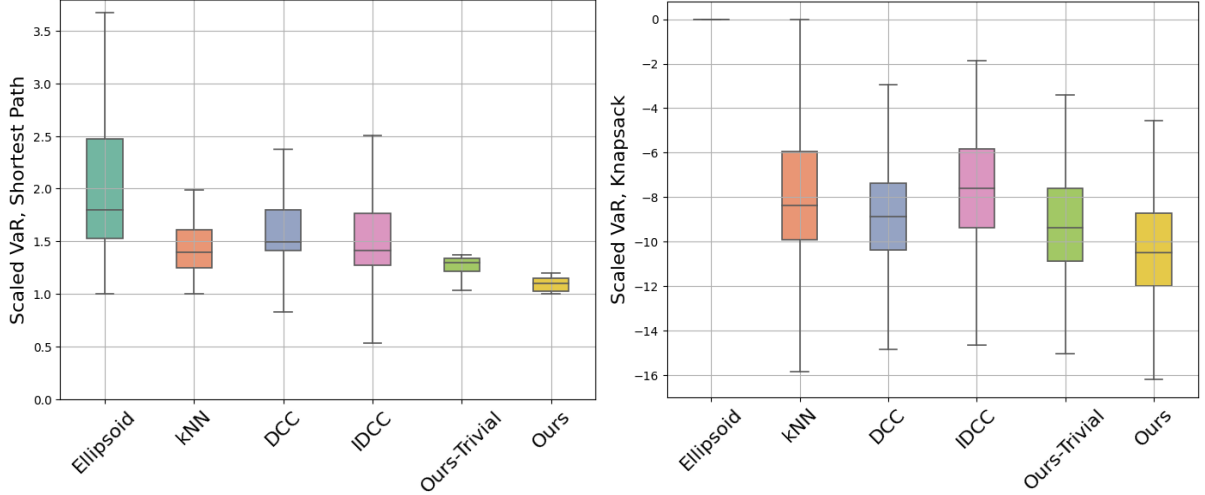


Figure 5: The scaled VaR for the objective in the shortest path and fractional knapsack problems. The x-axis lists the algorithms we test. The “Ellipsoid”, “kNN”, “DCC” and “IDCC” algorithms are introduced in the experiment descriptions. “Ours-Trivial” implements our algorithm but uses a trivial density ratio estimator $\hat{w} \equiv 1$. “Ours” implements our algorithm and uses the probabilistic classification method to estimate \hat{w} . Amongst the benchmark algorithms, our proposed Algorithm 1 generally achieves the lowest VaR.

For each x_i , let $\tilde{\eta}(x_i)$ denote the minimum η such that $[\hat{f}(z_i) - \eta \cdot \hat{h}(z_i), \hat{f}(z_i) + \eta \cdot \hat{h}(z_i)]$ covers c_i . The $\tilde{\eta}(\cdot)$ function is formally defined below

$$\tilde{\eta}(x_i) := \min \left\{ \eta \geq 0 : |\hat{f}(z_i) - c_i| \leq \eta \cdot \hat{h}(z_i) \right\}.$$

Under this new notation, the form of $\hat{\eta}(\cdot)$ can be formulated as

$$\hat{\eta}(x_1, \dots, x_N) = \min \left\{ \eta \geq 0 : \sum_{i=1}^N w(x_i) \cdot \mathbb{1}\{\eta \geq \tilde{\eta}(x_i)\} \geq \alpha \cdot \sum_{i=1}^N w(x_i) \right\}.$$

Further, if the density ratio estimate in Algorithm 1 is perfect, then the event $\{c_{\text{new}} \in \mathcal{U}_\alpha(z_{\text{new}})\}$ is equivalent to $\{\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N)\}$. We restate Theorem 1 below and provide a proof.

Theorem D.1. *Under Assumption 1, suppose the density ratio estimate is perfect, i.e., $\hat{w}(c, z) = w(c, z) = q(c, z)/p(c, z)$, then the uncertainty set $\mathcal{U}_\alpha(z)$ generated by Algorithm 1 satisfies the following coverage guarantee,*

$$\left| \mathbb{P}(\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N)) - \alpha \right| \leq \frac{1}{N+1} \cdot \frac{\bar{w}}{\underline{w}}$$

where the probability on the left-hand-side is with respect to $x_{\text{new}} \sim \mathcal{Q}$ and $x_1, \dots, x_N \sim \mathcal{P}$.

Proof. Since the distributions \mathcal{P} and \mathcal{Q} are continuous, almost surely the samples $x_1, \dots, x_N, x_{\text{new}}$ are mutually distinct. Let $\{\cdot\}$ denote an unordered set (e.g. $\{x_1, \dots, x_N\}$ denotes an unordered set containing distinct elements x_1, \dots, x_N). Then the following equation holds:

$$\begin{aligned} & \mathbb{P}(\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N)) \\ &= \int_{\{a_1, \dots, a_{N+1}\}} \mathbb{P}\left(\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N) \mid \{x_1, \dots, x_n, x_{\text{new}}\} = \{a_1, \dots, a_{N+1}\}\right) \\ & \quad \cdot \mathbb{P}\left(\{x_1, \dots, x_n, x_{\text{new}}\} = \{a_1, \dots, a_{N+1}\}\right). \end{aligned} \tag{19}$$

The integration is over all possible sets of $N+1$ distinct elements, denoted by $\{a_1, \dots, a_{N+1}\}$. The remaining part of the proof uniformly bounds the $\mathbb{P}\left(\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N) \mid \{x_1, \dots, x_n, x_{\text{new}}\} = \{a_1, \dots, a_{N+1}\}\right)$ term for all sets $\{a_1, \dots, a_{N+1}\}$.

Given a set $\{a_1, \dots, a_{N+1}\}$, let E denote the event that $\{x_1, \dots, x_n, x_{\text{new}}\} = \{a_1, \dots, a_{N+1}\}$, and let E_i denote the event that $x_{\text{new}} = a_i$ and $\{x_1, \dots, x_N\} = \{a_1, \dots, a_{N+1}\} \setminus \{a_i\}$. The $\mathbb{P}\left(\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N) \mid E\right)$ term can be decomposed by the following chain of equations:

$$\begin{aligned}
 & \mathbb{P}\left(\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N) \mid E\right) \\
 &= \sum_{i=1}^{N+1} \mathbb{P}(E_i \mid E) \cdot \mathbb{P}\left(\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N) \mid E_i\right) \\
 &\stackrel{(1)}{=} \sum_{i=1}^{N+1} \mathbb{P}(E_i \mid E) \cdot \mathbb{1}\left\{\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N) \mid E_i\right\} \\
 &\stackrel{(2)}{=} \sum_{i=1}^{N+1} \frac{w(a_i)}{\sum_{j=1}^{N+1} w(a_j)} \cdot \mathbb{1}\left\{\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N) \mid E_i\right\},
 \end{aligned} \tag{20}$$

where (1) is from the fact that $\hat{\eta}(x_1, \dots, x_N)$ is invariant to the permutation of (x_1, \dots, x_N) , and (2) is by that

$$\begin{aligned}
 \mathbb{P}(E_i) &= N! \cdot q(a_i) \cdot \prod_{j \neq i} p(a_j) \\
 &= N! \cdot w(a_i) \cdot \prod_{j=1}^{N+1} p(a_j),
 \end{aligned}$$

and $\mathbb{P}(E) = \sum_{i=1}^{N+1} \mathbb{P}(E_i)$.

We now characterize the terms $\mathbb{1}\left\{\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N) \mid E_i\right\}$. Without loss of generality, let $\tilde{\eta}(a_1) < \tilde{\eta}(a_2) \dots < \tilde{\eta}(a_{N+1})$, then

$$\begin{aligned}
 & \mathbb{1}\left\{\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N) \mid E_i\right\} \\
 &= \mathbb{1}\left\{\tilde{\eta}(a_i) \leq \hat{\eta}(\{a_1, \dots, a_{N+1}\} \setminus \{a_i\})\right\} \\
 &= \mathbb{1}\left\{\sum_{j=1}^{i-1} w(a_j) < \alpha \cdot \sum_{j \neq i} w(a_j)\right\}.
 \end{aligned} \tag{21}$$

The last line of (21) has the following bounds:

$$\mathbb{1}\left\{\sum_{j=1}^i w(a_j) < \alpha \cdot \sum_{j=1}^{N+1} w(a_j)\right\} \leq \mathbb{1}\left\{\sum_{j=1}^{i-1} w(a_j) < \alpha \cdot \sum_{j \neq i} w(a_j)\right\} \leq \mathbb{1}\left\{\sum_{j=1}^{i-1} w(a_j) < \alpha \cdot \sum_{j=1}^{N+1} w(a_j)\right\}. \tag{22}$$

Combining (20), (21) and (22), the $\mathbb{P}\left(\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N) \mid \{x_1, \dots, x_n, x_{\text{new}}\} = \{a_1, \dots, a_{N+1}\}\right)$ term is uniformly bounded:

$$\begin{aligned}
 & \mathbb{P}\left(\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N) \mid E\right) \\
 &\geq \sum_{i=1}^{N+1} \frac{w(a_i)}{\sum_{j=1}^{N+1} w(a_j)} \cdot \mathbb{1}\left\{\sum_{j=1}^i w(a_j) < \alpha \cdot \sum_{j=1}^{N+1} w(a_j)\right\} \\
 &\geq \alpha - \frac{\bar{w}}{\sum_{j=1}^{N+1} w(a_j)} \geq \alpha - \frac{1}{N+1} \cdot \frac{\bar{w}}{\underline{w}}.
 \end{aligned}$$

$$\begin{aligned}
 & \mathbb{P}\left(\tilde{\eta}(x_{\text{new}}) \leq \hat{\eta}(x_1, \dots, x_N) \mid E\right) \\
 & \leq \sum_{i=1}^{N+1} \frac{w(a_i)}{\sum_{j=1}^{N+1} w(a_j)} \cdot \mathbb{1}\left\{\sum_{j=1}^{i-1} w(a_j) < \alpha \cdot \sum_{j=1}^{N+1} w(a_j)\right\} \\
 & \leq \alpha + \frac{\bar{w}}{\sum_{j=1}^{N+1} w(a_j)} \leq \alpha + \frac{1}{N+1} \cdot \frac{\bar{w}}{\underline{w}}.
 \end{aligned}$$

Plugging the bounds above into (19) finishes the proof. \square

D.2 Proof of Corollary 1

By defining the estimated test distribution $\hat{\mathcal{Q}}$ as a distribution with the following density function:

$$\hat{q}(c, z) = \frac{\hat{w}(c, z) \cdot p(c, z)}{\int_{(c, z)} \hat{w}(c, z) \cdot p(c, z) dz dc},$$

we could directly apply Theorem 1 to see that

$$\left| \mathbb{P}_{\hat{\mathcal{Q}}}(c_{\text{new}} \in \mathcal{U}_{\alpha}(z_{\text{new}})) - \alpha \right| \leq \frac{1}{|\mathcal{D}_2| + 1} \cdot \frac{\bar{w}}{\underline{w}}, \quad (23)$$

where the probability $\mathbb{P}_{\hat{\mathcal{Q}}}$ is with respect to $(c_{\text{new}}, z_{\text{new}}) \sim \hat{\mathcal{Q}}$ and $\mathcal{D}_2 \sim \mathcal{P}$. From the definition of the total variation distance, the following inequality holds:

$$\left| \mathbb{P}_{\hat{\mathcal{Q}}}(c_{\text{new}} \in \mathcal{U}_{\alpha}(z_{\text{new}})) - \mathbb{P}(c_{\text{new}} \in \mathcal{U}_{\alpha}(z_{\text{new}})) \right| \leq D_{\text{TV}}(\mathcal{Q}, \hat{\mathcal{Q}}), \quad (24)$$

where \mathbb{P} is with respect to $(c_{\text{new}}, z_{\text{new}}) \sim \mathcal{Q}$ and $\mathcal{D}_2 \sim \mathcal{P}$. Combining (23) and (24) gives the result of Corollary 1, which we restate below:

$$\left| \mathbb{P}(c_{\text{new}} \in \mathcal{U}_{\alpha}(z_{\text{new}})) - \alpha \right| \leq \frac{1}{|\mathcal{D}_2| + 1} \cdot \frac{\bar{w}}{\underline{w}} + D_{\text{TV}}(\mathcal{Q}, \hat{\mathcal{Q}}).$$