Appendices to *"Optimal Robust Policy for Feature-based newsvendor"*

Appendix A: Proofs for Section 3.1

The following Lemma EC.1 is a direct consequence of the strong duality result in Wasserstein distributionally robust optimization (e.g., Gao and Kleywegt (2022), Mohajerin Esfahani and Kuhn (2018), Blanchet and Murthy (2019)). To ease the notation, in the sequel we denote $\Psi_f(x,z) := \Psi(f(x),z)$ and $\Psi_{\widehat{f}}(\widehat{x},z) = \Psi(\widehat{f}(\widehat{x}),z)$.

LEMMA EC.1. *For each $f \in \mathcal{F}$, the inner primal problem*

$$v_P^f := \sup_{\mathbb{P}\in\mathcal{P}_1(\mathcal{X}\times\mathcal{Z})} \left\{ \mathbb{E}_{(X,Z)\sim\mathbb{P}}[\Psi_f(X,Z)] : \mathcal{W}(\mathbb{P},\widehat{\mathbb{P}}) \le \rho \right\}.$$

is equal to the following inner dual problem,

$$v_D^f := \inf_{\lambda\ge 0} \left\{ \lambda\rho + \mathbb{E}_{(\widehat{X},\widehat{Z})\sim\widehat{\mathbb{P}}} \left[\sup_{(x,z)\in\mathcal{X}\times\mathcal{Z}} \left\{ \Psi_f(x,z) - \lambda(\|x-\widehat{X}\| + |z-\widehat{Z}|) \right\} \right] \right\}.$$

Note that $v_P^f = v_D^f$ can be infinite if $\limsup_{(x,z)\to\infty} \Psi(f(x),z) = \infty$, but in this case f cannot be a minimizer of (P).

Proof of Lemma 1. Denote $y_k = \widehat{f}(\widehat{x}_k)$, and we define

$$\begin{aligned}
A_{kk}(x) &:= y_k, \qquad k = 1,\dots,K,\\
A_{jk}(x) &:= \frac{\|x-\widehat{x}_k\|}{\|x-\widehat{x}_j\| + \|x-\widehat{x}_k\|} y_j + \frac{\|x-\widehat{x}_j\|}{\|x-\widehat{x}_j\| + \|x-\widehat{x}_k\|} y_k, \qquad j \ne k,\\
A^+(x) &:= \min_{1\le k\le K} \max_{1\le j\le K} A_{jk}(x), \qquad A^-(x) := \max_{1\le k\le K} \min_{1\le j\le K} A_{jk}(x).
\end{aligned}$$

In Figure EC.1, we plot the graph of the function A_{12} when $K=2$ (left) and A_{12}, A_{23}, A_{13} when $K=3$ (right), in the case $\mathcal{X} = \mathbb{R}^2$. Same as the setting of Figure 2, when $K=2$ the Shapley policy is A_{12}, and when $K=3$ the Shapley policy is the middle one among A_{12}, A_{13} and A_{23}, which is rendered with a mesh in this figure.

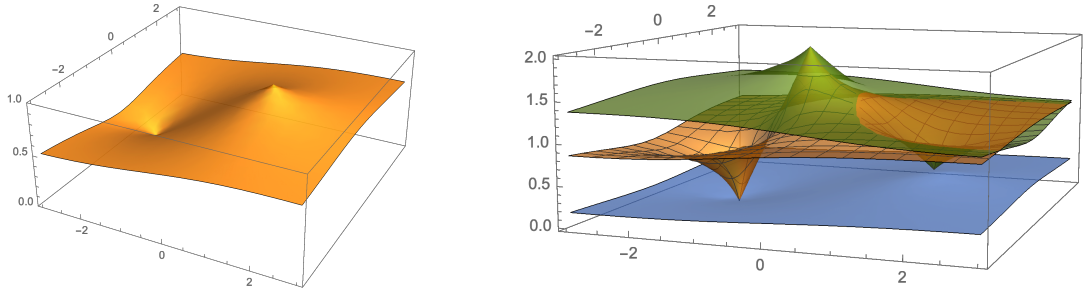

Figure EC.1 Graph of the Shapley policy $y = f(x)$ when $K = 2, 3$, $x \in \mathbb{R}^2$

We claim A^+ and A^- both satisfy the four properties. First, we show they are indeed extensions. Fix $\ell \in 1,\dots,K$, then $A_{j\ell}(\widehat{x}_\ell) = A_{\ell j}(\widehat{x}_\ell) = y_\ell$ for every j. This implies

$$\begin{aligned}
A^+(\widehat{x}_\ell) &= \min_{1\le k\le K} \max_{1\le j\le K} A_{jk}(\widehat{x}_\ell) \le \max_{1\le j\le K} A_{j\ell}(\widehat{x}_\ell) = y_\ell,\\
A^-(\widehat{x}_\ell) &= \max_{1\le k\le K} \min_{1\le j\le K} A_{jk}(\widehat{x}_\ell) \ge \min_{1\le j\le K} A_{j\ell}(\widehat{x}_\ell) = y_\ell.
\end{aligned}$$

However, $A^+ \geq A^-$, so in fact $A^+(\widehat{x}_\ell) = A^-(\widehat{x}_\ell) = y_\ell$, that is, A^+ and A^- interpolate given data.

Next we show the boundedness and the Lipschitzness. It suffices to show them for each A_{jk}, because both bounds are compatible with min max operations. Because $A_{jk}(x)$ is just an interpolation between y_j and y_k, clearly we have $\min\{y_j, y_k\} \leq A_{jk}(x) \leq \max\{y_j, y_k\}$. As for the Lipschitz bound of A_{jk}, when $j = k$, $A_{kk} \equiv y_k$ are constant functions, so they always satisfy the Lipschitz bound. When $j \neq k$, fix $x, x' \in \mathcal{X}$, and we denote

$$d_{xj} = \|\widehat{x}_j - x\|, \quad d_{x'j} = \|\widehat{x}_j - x'\|, \quad d_{xk} = \|\widehat{x}_k - x\|, \quad d_{x'k} = \|\widehat{x}_k - x'\|, \quad d_{xx'} = \|x - x'\|, \quad d_{jk} = \|\widehat{x}_j - \widehat{x}_k\|.$$

Then

$$\begin{aligned}
A_{jk}(x) - A_{jk}(x') &= \frac{d_{xk}}{d_{xj} + d_{xk}} y_j + \frac{d_{xj}}{d_{xj} + d_{xk}} y_k - \frac{d_{x'k}}{d_{x'j} + d_{x'k}} y_j - \frac{d_{x'j}}{d_{x'j} + d_{x'k}} y_k \\
&= (y_j - y_k) \left(\frac{d_{xk}}{d_{xj} + d_{xk}} - \frac{d_{x'k}}{d_{x'j} + d_{x'k}} \right) \\
&= (y_j - y_k) \left(\frac{d_{xk} d_{x'j} + d_{xk} d_{x'k} - d_{xj} d_{x'k} - d_{xk} d_{x'k}}{(d_{xj} + d_{xk})(d_{x'j} + d_{x'k})} \right) \\
&= (y_j - y_k) \left(\frac{d_{xk} d_{x'j} - d_{xk} d_{xj} + d_{xj} d_{xk} - d_{xj} d_{x'k}}{(d_{xj} + d_{xk})(d_{x'j} + d_{x'k})} \right) \\
&= (y_j - y_k) \left(\frac{d_{xk}(d_{x'j} - d_{xj}) + d_{xj}(d_{xk} - d_{x'k})}{(d_{xj} + d_{xk})(d_{x'j} + d_{x'k})} \right).
\end{aligned}$$

By triangular inequality,

$$\begin{aligned}
|A_{jk}(x) - A_{jk}(x')| &\leq |y_j - y_k| \left(\frac{d_{xk} d_{xx'} + d_{xj} d_{xx'}}{(d_{xj} + d_{xk})(d_{x'j} + d_{x'k})} \right) \\
&\leq |y_j - y_k| \left(\frac{d_{xx'}}{d_{x'j} + d_{x'k}} \right) \\
&\leq |y_j - y_k| \left(\frac{d_{xx'}}{d_{jk}} \right) \\
&= \frac{|y_j - y_k|}{\|\widehat{x}_j - \widehat{x}_k\|} \|x - x'\|.
\end{aligned}$$

Thus if we denote $L := \max_{j \neq k} \dfrac{|y_j - y_k|}{\|\widehat{x}_j - \widehat{x}_k\|}$ to be the discrete Lipschitz constant of the given data, then all the A_{jk} are L-Lipschitz in x, so there min and max are also L-Lipschitz in x.

It remains to prove (5), which is to show that $y = A^+(x), A^-(x)$ satisfy the following condition for every k:

$$\Phi(y) - \|x - \widehat{x}_k\| \leq \max_{j=1,\ldots,K} \left\{ \Phi(y_j) - \|\widehat{x}_j - \widehat{x}_k\| \right\} =: M_k. \tag{Mk}$$

We first claim that $y = A_{jk}(x)$ satisfy the bound (Mk). By convexity of Φ,

$$\begin{aligned}
\Phi(A_{jk}(x)) - \|x - \widehat{x}_k\| &= \Phi\left(\frac{\|x - \widehat{x}_k\|}{\|x - \widehat{x}_j\| + \|x - \widehat{x}_k\|} y_j + \frac{\|x - \widehat{x}_j\|}{\|x - \widehat{x}_j\| + \|x - \widehat{x}_k\|} y_k \right) - \|x - \widehat{x}_k\| \\
&\leq \frac{\|x - \widehat{x}_k\|}{\|x - \widehat{x}_j\| + \|x - \widehat{x}_k\|} \Phi(y_j) + \frac{\|x - \widehat{x}_j\|}{\|x - \widehat{x}_j\| + \|x - \widehat{x}_k\|} \Phi(y_k) - \|x - \widehat{x}_k\|.
\end{aligned}$$

Using definition of M_k in (Mk),

$$\Phi(y_j) \leq M_k + \|\widehat{x}_j - \widehat{x}_k\|, \qquad \Phi(y_k) \leq M_k + \|\widehat{x}_k - \widehat{x}_k\| = M_k,$$

Plug in into the above inequality,

$$\begin{aligned}\Phi(A_{jk}(x)) - \|x-\widehat{x}_k\| &\le M_k + \frac{\|x-\widehat{x}_k\|}{\|x-\widehat{x}_j\|+\|x-\widehat{x}_k\|}\|\widehat{x}_j-\widehat{x}_k\| - \|x-\widehat{x}_k\| \\ &= M_k + \|x-\widehat{x}_k\|\left(\frac{\|\widehat{x}_j-\widehat{x}_k\|}{\|x-\widehat{x}_j\|+\|x-\widehat{x}_k\|}-1\right) \le M_k.\end{aligned}$$

In the last step we used the triangle inequality $\|\widehat{x}_j-\widehat{x}_k\| \le \|x-\widehat{x}_j\|+\|x-\widehat{x}_k\|$. Fixing a k in $1,\dots,K$, then from the definition we can find j_1, j_2 such that $A_{j_1k}(x) \le A^-(x) \le A^+(x) \le A_{j_2k}(x)$. Because in a closed interval $[A_{j_1k}(x), A_{j_2k}(x)]$ convex function Φ can only attain maximum at the endpoints due to the maximum principle, we conclude that

$$\Phi(A^-(x)), \Phi(A^+(x)) \le \max\{\Phi(A_{j_1k}(x)), \Phi(A_{j_2k}(x))\}.$$

As a result, since both $y = A_{j_1k}(x)$ and $y = A_{j_2k}(x)$ satisfy (Mk), we conclude that $y = A^+(x), A^-(x)$ also satisfy (Mk), which implies (5).

The existence of the saddle point is proved in Lemma EC.2 below. It is purely algebraic and it makes use of the Shapley's theorem. □

LEMMA EC.2. *With the same notation as Lemma 1, $A^+ = A^-$.*

Proof. We fix an x and then omit the x in $A_{jk}(x)$ to ease the notation. Denote the symmetric matrix $A = (A_{jk})_{jk}$, and we want to show that for this matrix, $\min_k \max_j A_{jk} = \max_j \min_k A_{jk}$, i.e., a saddle point exists. By Theorem 2.1 in Shapley (1964), to show the existence of a saddle point for A, it is sufficient to show that any 2×2 submatrix of A has a saddle point.

For a general 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, we claim that if it has no saddle point then it has the *diagonal dominant property*: the two elements on one diagonal is strictly greater than the two elements on the other diagonal. To show this claim, we note that the maximin being different from the minimax means

$$(a \wedge c) \vee (b \wedge d) \ne (a \vee b) \wedge (c \vee d).$$

Without loss of generality, assume a is the smallest entry. Then the above inequality is simplified to

$$b \wedge d \ne b \wedge (c \vee d).$$

If $d \ge c$, then both sides equals to $b \wedge d$. If $d \ge b$, then both sides equals to b. Hence the above inequality can only hold if $d < c$ and $d < b$, which implies a, d are both strictly smaller than b, c.

Applying to our case, we need to show that for any i, j, k, l, the matrix

$$B = \begin{pmatrix} A_{ik} & A_{jk} \\ A_{i\ell} & A_{j\ell} \end{pmatrix}$$

doesn't have the diagonal dominant property. Recall that A_{jk} is an interpolation of y_j and y_k, so it is important to compare the values of y_i, y_j, y_k, y_ℓ. Without loss of generality, assume $y_i \ge y_j$, $y_k \ge y_\ell$, and assume $y_i \ge y_k$ using the transpose symmetry. There are three possibilities:

(I) $y_i \ge y_k \ge y_\ell > y_j$.

(II) $y_i \ge y_k \ge y_j \ge y_\ell$.

(III) $y_i \ge y_j > y_k \ge y_\ell$.

In case (I), $A_{ik} \ge y_k \ge A_{jk}$, $A_{j\ell} \le y_\ell \le A_{il}$, so neither diagonal can dominate the other. In case (II), $A_{ik} \ge y_k \ge A_{jk} \ge y_j \ge A_{j\ell}$, again neither diagonal can dominate the other. The third case needs a further discussion.

Since A^+, A^- are both extension of $\widehat{f}$, they always agree on $\widehat{\mathcal{X}}$, so we may assume $x \ne \widehat{x}_i, \widehat{x}_j, \widehat{x}_k, \widehat{x}_\ell$ thus

$$d_i = \|x - \widehat{x}_i\|, \qquad d_j = \|x - \widehat{x}_j\|, \qquad d_k = \|x - \widehat{x}_k\|, \qquad d_\ell = \|x - \widehat{x}_\ell\|$$

are all positive. Recall that $A_{jk} = \frac{d_k}{d_j+d_k} y_j + \frac{d_j}{d_j+d_k} y_k$. We prove by contradiction and assume B has the diagonal dominant property. If the main diagonal is strictly greater than the off diagonal, then

$$A_{ik}, A_{j\ell} > A_{jk}, A_{i\ell}.$$

For instance, from $A_{ik} > A_{jk}$ we have

$$\begin{aligned}
\frac{d_k}{d_i+d_k} y_i + \frac{d_i}{d_i+d_k} y_k &> \frac{d_k}{d_j+d_k} y_j + \frac{d_j}{d_j+d_k} y_k \\
\frac{d_k}{d_i+d_k} (y_i - y_k) + y_k &> \frac{d_k}{d_j+d_k} (y_j - y_k) + y_k \\
\frac{d_k}{d_i+d_k} (y_i - y_k) &> \frac{d_k}{d_j+d_k} (y_j - y_k) \\
(d_j+d_k)(y_i - y_k) &> (d_i+d_k)(y_j - y_k). \qquad \text{(EC.1)}
\end{aligned}$$

Similarly, from $A_{ik} > A_{i\ell}, A_{j\ell} > A_{jk}, A_{j\ell} > A_{i\ell}$ we conclude

$$\begin{aligned}
(d_i+d_k)(y_i - y_\ell) &> (d_i+d_\ell)(y_i - y_k), \\
(d_j+d_\ell)(y_j - y_k) &> (d_j+d_k)(y_j - y_\ell), \\
(d_i+d_\ell)(y_j - y_\ell) &> (d_j+d_\ell)(y_i - y_\ell).
\end{aligned}$$

Note that in case (III), every term in (EC.1) and the above three inequalities is positive. So if we multiply four inequalities, we would reach a contradiction.

If the main diagonal is strictly smaller than the off diagonal, then all the inequalities above flip sign, and we would still reach a contradiction. In conclusion, B never has the diagonal dominant property. In other words, B admits a saddle point. By Theorem 2.1 in Shapley (1964), A also admits a saddle point, therefore $A^+ = A^-$. □

Proof of Theorem 1. To show the direction $v_D \ge v_{\hat{D}}$, note that $v_{\hat{D}}$ can be written with $f \in \mathcal{F}$ instead of $\widehat{f} \in \widehat{\mathcal{F}}$:

$$v_{\hat{D}} = \inf_{f \in \mathcal{F}} \inf_{\lambda \ge 0} \left\{ \lambda\rho + \mathbb{E}_{\widehat{\mathbb{P}}} \left[\sup_{z \in \mathcal{Z}} \max_{x \in \widehat{\mathcal{X}}} \left\{ \Psi(f(x), z) - \lambda \|x - \widehat{X}\| - \lambda |z - \widehat{Z}| \right\} \right] \right\}$$

since the target function depends only on the value of the restriction $\widehat{f} = f|_{\widehat{\mathcal{X}}}$. From this, $v_D \ge v_{\hat{D}}$ is straightforward because we are restricting the set over which the supremum of x is taken.

To show $v_D \le v_{\hat{D}}$, we let f be the Shapley extension of a given $\widehat{f} \in \widehat{\mathcal{F}}$ as defined in Lemma 1, and split into two cases: $\lambda > 0$ and $\lambda = 0$. If $\lambda > 0$, by Lemma 1 we know that f satisfies (5). In particular, If we choose $\Phi(y) = \frac{1}{\lambda} \Psi(y, z) - |z - \widehat{Z}|$, then

$$\sup_{x \in \mathcal{X}} \left\{ \Psi(f(x), z) - \lambda \|x - \widehat{X}\| \right\} \le \max_{x \in \widehat{\mathcal{X}}} \left\{ \Psi(\widehat{f}(x), z) - \lambda \|x - \widehat{X}\| \right\}, \qquad \text{for all } \hat{X} \in \widehat{\mathcal{X}}, z \in \mathcal{Z}.$$

Consequently,

$$\lambda\rho + \mathbb{E}_{\widehat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}}\sup_{x\in\mathcal{X}}\left\{\Psi(f(x),z) - \lambda\|x-\widehat{X}\| - \lambda|z-\widehat{Z}|\right\}\right]$$
$$\le \lambda\rho + \mathbb{E}_{\widehat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}}\max_{x\in\widehat{\mathcal{X}}}\left\{\Psi(\widehat{f}(x),z) - \lambda\|x-\widehat{X}\| - \lambda|z-\widehat{Z}|\right\}\right].$$

If $\lambda = 0$, by Lemma 1 we know that the range of f is $[\min\widehat{f}, \max\widehat{f}]$. Since $\Psi(\cdot,z)$ is convex, the supremum of $\Psi(f(\cdot),z)$ is attained at the extreme points, so

$$\mathbb{E}_{\widehat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}}\sup_{x\in\mathcal{X}}\{\Psi(f(x),z)\}\right] \le \mathbb{E}_{\widehat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}}\max_{x\in\widehat{\mathcal{X}}}\left\{\Psi(\widehat{f}(x),z)\right\}\right].$$

Therefore, taking infimum in $\lambda \ge 0$ gives

$$\inf_{\lambda\ge 0}\left\{\lambda\rho + \mathbb{E}_{\widehat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}}\sup_{x\in\mathcal{X}}\left\{\Psi(f(x),z) - \lambda\|x-\widehat{X}\| - \lambda|z-\widehat{Z}|\right\}\right]\right\}$$
$$\le \inf_{\lambda\ge 0}\left\{\lambda\rho + \mathbb{E}_{\widehat{\mathbb{P}}}\left[\sup_{z\in\mathcal{Z}}\max_{x\in\widehat{\mathcal{X}}}\left\{\Psi(\widehat{f}(x),z) - \lambda\|x-\widehat{X}\| - \lambda|z-\widehat{Z}|\right\}\right]\right\}.$$

The left dominates v_D, so taking the inf over $\widehat{f}\in\widehat{\mathcal{F}}$ on the right gives $v_D \le v_{\hat{D}}$, which completes the proof of $v_D = v_{\hat{D}}$. Note that the above also proves that if $\widehat{f}^*$ is a minimizer of $v_{\hat{D}}$, then the Shapley extension of $\widehat{f}^*$ is a minimizer of v_D, also a minimizer of v_P by Lemma EC.1. □

Appendix B: Proofs for Section 4.1

LEMMA EC.3. *For each fixed $\lambda \ge b\vee h$ and $\widehat{f}\in\widehat{\mathcal{F}}$, we can always find a $\frac{\lambda}{b\vee h}$-Lipschitz policy $\widehat{g}\in\widehat{\mathcal{F}}$ satisfying $\underline{z}_k \le \widehat{g}(\widehat{x}_k) \le \overline{z}_k$, where*

$$\underline{z}_k := \min_{(x,z)\in\operatorname{supp}\widehat{\mathbb{P}}}\{z + \|x-\widehat{x}_k\|\}, \qquad \overline{z}_k := \max_{(x,z)\in\operatorname{supp}\widehat{\mathbb{P}}}\{z - \|x-\widehat{x}_k\|\},$$

such that for all $(\widehat{x},\widehat{z})\in\operatorname{supp}\widehat{\mathbb{P}}$,

$$\max_{x\in\widehat{\mathcal{X}}}\{\Psi(\widehat{g}(x),\widehat{z}) - \lambda\|x-\widehat{x}\|\} \le \max_{x\in\widehat{\mathcal{X}}}\left\{\Psi(\widehat{f}(x),\widehat{z}) - \lambda\|x-\widehat{x}\|\right\}.$$

Proof of Lemma EC.3. Denote $y_k = \widehat{f}(\widehat{x}_k)$. First, we show that if $\widehat{f}$ is "not Lipschitz enough" at some point, in the sense that its local Lipschitz constant at a point is too large, then we can reduce it by modifying $\widehat{f}$. Suppose there exists j_0, k_0 such that

$$y_{j_0} \ge y_{k_0} + L\|\widehat{x}_{k_0} - \widehat{x}_{j_0}\| := \tilde{y}_{j_0}$$

for some constant $L > 0$ to be specify later. We claim that replacing decision y_{j_0} by $\tilde{y}_{j_0}$ will not deteriorate the objective value, in the sense that for any $(\widehat{x},\widehat{z})$ in the support of $\widehat{\mathbb{P}}$,

$$h(\tilde{y}_{j_0} - \widehat{z})_+ + b(\widehat{z} - \tilde{y}_{j_0})_+ - \lambda\|\widehat{x}_{j_0} - \widehat{x}\| \le \max_{j=1,\dots,K}\left\{h(y_j - \widehat{z})_+ + b(\widehat{z} - y_j)_+ - \lambda\|\widehat{x}_j - \widehat{x}\|\right\} =: \text{RHS}.$$

We consider two cases. If $\tilde{y}_{j_0} \ge \widehat{z}$, then

$$h(\tilde{y}_{j_0} - \widehat{z})_+ + b(\widehat{z} - \tilde{y}_{j_0})_+ - \lambda\|\widehat{x}_{j_0} - \widehat{x}\| = h(\tilde{y}_{j_0} - \widehat{z}) - \lambda\|\widehat{x}_{j_0} - \widehat{x}\|$$
$$\le h(y_{j_0} - \widehat{z}) - \lambda\|\widehat{x}_{j_0} - \widehat{x}\| \le \text{RHS}.$$

If $\tilde{y}_{j_0} < \widehat{z}$, then

$$\begin{aligned} h(\tilde{y}_{j_0} - \widehat{z})_+ + b(\widehat{z} - \tilde{y}_{j_0})_+ - \lambda\|\widehat{x}_{j_0} - \widehat{x}\| &= b(\widehat{z} - \tilde{y}_{j_0}) - \lambda\|\widehat{x}_{j_0} - \widehat{x}\| \\ &= b(\widehat{z} - y_{k_0} - L\|\widehat{x}_{k_0} - \widehat{x}_{j_0}\|) - \lambda\|\widehat{x}_{j_0} - \widehat{x}\| \\ &= b(\widehat{z} - y_{k_0}) - bL\|\widehat{x}_{k_0} - \widehat{x}_{j_0}\| - \lambda\|\widehat{x}_{j_0} - \widehat{x}\|. \end{aligned}$$

If we choose any $L \geq \frac{\lambda}{b}$, then

$$\begin{aligned} h(\tilde{y}_{j_0} - \widehat{z})_+ + b(\widehat{z} - \tilde{y}_{j_0})_+ - \lambda\|\widehat{x}_{j_0} - \widehat{x}\| &\leq b(\widehat{z} - y_{k_0}) - \lambda\|\widehat{x}_{k_0} - \widehat{x}_{j_0}\| - \lambda\|\widehat{x}_{j_0} - \widehat{x}\| \\ &\leq b(\widehat{z} - y_{k_0}) - \lambda\|\widehat{x}_{k_0} - \widehat{x}\| \leq \text{RHS}. \end{aligned}$$

This completes the proof of the claim.

Now we make use of the above claim recursively.

`Step 0.` Denote $y_k^{(1)} = y_k$ for every $k \in [K]$.

`Step 1.` Pick $k_1 \in \arg\min_{k\in[K]}\{y_k^{(1)}\}$, and define $y_j^{(2)} = y_j^{(1)} \wedge (y_{k_1}^{(1)} + L\|\widehat{x}_j - \widehat{x}_{k_1}\|)$ for every $j \in [K]$.

`Step 2.` Pick $k_2 \in \arg\min_{k\in[K]\setminus\{k_1\}}\{y_k^{(2)}\}$, and define $y_j^{(3)} = y_j^{(2)} \wedge (y_{k_2}^{(2)} + L\|\widehat{x}_j - \widehat{x}_{k_2}\|)$ for every $j \in [K]$.

`Step 3.` Pick $k_3 \in \arg\min_{k\in[K]\setminus\{k_1,k_2\}}\{y_k^{(3)}\}$, and define $y_j^{(4)} = y_j^{(3)} \wedge (y_{k_3}^{(3)} + L\|\widehat{x}_j - \widehat{x}_{k_3}\|)$ for every $j \in [K]$

...

The above process terminates after `Step K-1`. We have a sequence of policies $\widehat{f}^{(m)}$ defined by $\widehat{f}^{(m)}(\widehat{x}_k) = y_k^{(m)}$. According to the previous claim, each step does not deteriorate the objective value: for any $1 \leq m \leq K-1$,

$$\max_{x\in\widehat{\mathcal{X}}}\left\{\Psi(\widehat{f}^{(m+1)}(x), \widehat{z}) - \lambda\|x - \widehat{x}\|\right\} \leq \max_{x\in\widehat{\mathcal{X}}}\left\{\Psi(\widehat{f}^{(m)}(x), \widehat{z}) - \lambda\|x - \widehat{x}\|\right\}.$$

It is easy to conclude that our selection has the following properties:

(I) It is decreasing: $\widehat{f}^{(m+1)} \leq \widehat{f}^{(m)}$.

(II) The sequence $y_{k_m}^{(m)}$ is increasing in m, that is,

$$y_{k_1}^{(1)} \leq y_{k_2}^{(2)} \leq y_{k_3}^{(3)} \leq \cdots \leq y_{k_K}^{(K)}.$$

This is because $y_{k_{m+1}}^{(m)} \geq y_{k_m}^{(m)}$ since k_m is the argmin in step k, and by definition we have

$$y_{k_{m+1}}^{(m+1)} = y_{k_{m+1}}^{(m)} \wedge (y_{k_m}^{(m)} + L\|\widehat{x}_{k_m} - \widehat{x}_{k_{m+1}}\|) \geq y_{k_m}^{(m)}.$$

(III) The above increasing order implies the value at $\widehat{x}_{k_m}$ stops decreasing after step m:

$$y_{k_1}^{(1)} = y_{k_1}^{(2)} = \cdots = y_{k_1}^{(K)}, \qquad y_{k_2}^{(2)} = y_{k_2}^{(3)} = \cdots = y_{k_2}^{(K)}, \qquad y_{k_3}^{(3)} = y_{k_3}^{(4)} = \cdots = y_{k_3}^{(K)}, \qquad \ldots$$

again following the definition. Therefore $\widehat{f}^{(K)}(\widehat{x}_{k_m}) = y_{k_m}^{(m)}$.

Combine the above three properties, we have for any $m < n$,

$$y_{k_m}^{(m)} \leq y_{k_n}^{(n)} \leq y_{k_n}^{(m+1)} \leq y_{k_m}^{(m)} + L\|\widehat{x}_{k_n} - \widehat{x}_{k_m}\|.$$

Now we define $\tilde{f} = \widehat{f}^{(K)}$, then it is L-Lipschitz. A similar argument works for $L \geq \frac{\lambda}{h}$, so we can pick $L = \frac{\lambda}{b} \wedge \frac{\lambda}{h} = \frac{\lambda}{b\vee h}$, and by the above construction $\tilde{f}$ is $\frac{\lambda}{b\vee h}$-Lipschitz and satisfy

$$\max_{j=1,\ldots,K}\left\{h(\tilde{f}(\widehat{x}_j) - \widehat{z})_+ + b(\widehat{z} - \tilde{f}(\widehat{x}_j))_+ - \lambda\|\widehat{x}_j - \widehat{x}\|\right\} \leq \max_{j=1,\ldots,K}\left\{h(y_j - \widehat{z})_+ + b(\widehat{z} - y_j)_+ - \lambda\|\widehat{x}_j - \widehat{x}\|\right\}$$

for all $(\widehat{x},\widehat{z}) \in \operatorname{supp}\widehat{\mathbb{P}}$, that is,

$$\max_{x\in\widehat{\mathcal{X}}}\left\{\Psi(\tilde{f}(x),\widehat{z}) - \lambda\|x-\widehat{x}\|\right\} \le \max_{x\in\widehat{\mathcal{X}}}\left\{\Psi(\widehat{f}(x),\widehat{z}) - \lambda\|x-\widehat{x}\|\right\}.$$

Now we deal with upper and lower bound. By the first part of this proof we can assume without loss of generality that $\widehat{f}$ is already $\frac{\lambda}{b\vee h}$-Lipschitz to begin with. Define $\tilde{y}_j = \overline{z}_j \wedge y_j$ for every j, we claim that for any $(\widehat{x},\widehat{z})$ in the support of $\widehat{\mathbb{P}}$,

$$\max_{j=1,\dots,K}\left\{h(\tilde{y}_j-\widehat{z})_+ + b(\widehat{z}-\tilde{y}_j)_+ - \lambda\|\widehat{x}_j-\widehat{x}\|\right\} \le \max_{j=1,\dots,K}\left\{h(y_j-\widehat{z})_+ + b(\widehat{z}-y_j)_+ - \lambda\|\widehat{x}_j-\widehat{x}\|\right\}.$$

Indeed, if $\tilde{y}_j = y_j$, then we are not changing anything, directly we have

$$h(\tilde{y}_j-\widehat{z})_+ + b(\widehat{z}-\tilde{y}_j)_+ - \lambda\|\widehat{x}_j-\widehat{x}\| = h(y_j-\widehat{z})_+ + b(\widehat{z}-y_j)_+ - \lambda\|\widehat{x}_j-\widehat{x}\| \le \text{RHS}.$$

When $\tilde{y}_j = \overline{z}_j \le y_j$, we split to two cases. On the one hand, if $\widehat{z} \le \tilde{y}_j$, then

$$h(\tilde{y}_j-\widehat{z})_+ + b(\widehat{z}-\tilde{y}_j)_+ - \lambda\|\widehat{x}_j-\widehat{x}\| \le h(\tilde{y}_j-\widehat{z}) - \lambda\|\widehat{x}_j-\widehat{x}\| \le h(y_j-\widehat{z}) - \lambda\|\widehat{x}_j-\widehat{x}\| \le \text{RHS}.$$

On the other hand, if $\widehat{z} \ge \tilde{y}_j$, using $\tilde{y}_j = \overline{z}_j \ge \widehat{z} - \|\widehat{x}-\widehat{x}_j\|$ by the definition of $\overline{z}$, so

$$h(\tilde{y}_j-\widehat{z})_+ + b(\widehat{z}-\tilde{y}_j)_+ - \lambda\|\widehat{x}_j-\widehat{x}\| = b(\widehat{z}-\tilde{y}_j) - \lambda\|\widehat{x}_j-\widehat{x}\| \le b\|\widehat{x}-\widehat{x}_j\| - \lambda\|\widehat{x}_j-\widehat{x}\| \le 0 \le \text{RHS}.$$

Here we used that $\lambda \ge b$, and the right hand side is always nonnegative because it is nonnegative when $\widehat{x}_j = \widehat{x}$. In conclusion, for every scenario we have

$$h(\tilde{y}_j-\widehat{z})_+ + b(\widehat{z}-\tilde{y}_j)_+ - \lambda\|\widehat{x}_j-\widehat{x}\| \le \max_{j=1,\dots,K}\left\{h(y_j-\widehat{z})_+ + b(\widehat{z}-y_j)_+ - \lambda\|\widehat{x}_j-\widehat{x}\|\right\}.$$

Take maximum on the left over j completes the proof of the claim.

Similarly we can let $\tilde{\tilde{y}}_j = \underline{z}_j \vee \tilde{y}_j$, and $\tilde{\tilde{y}}_j$ will satisfy

$$\max_{j=1,\dots,K}\left\{h(\tilde{\tilde{y}}_j-\widehat{z})_+ + b(\widehat{z}-\tilde{\tilde{y}}_j)_+ - \lambda\|\widehat{x}_j-\widehat{x}\|\right\} \le \max_{j=1,\dots,K}\left\{h(y_j-\widehat{z})_+ + b(\widehat{z}-y_j)_+ - \lambda\|\widehat{x}_j-\widehat{x}\|\right\}.$$

Now we define $\widehat{g}(\widehat{x}_j) = \underline{z}_j \vee (\overline{z}_j \wedge \tilde{f}(\widehat{x}_j))$ for every j, with $\tilde{f}$ being the $\frac{\lambda}{b\vee h}$-Lipschitz function define in the first part of the proof. Then $\widehat{g}$ satisfies

$$\max_{x\in\widehat{\mathcal{X}}}\left\{\Psi(\widehat{g}(x),\widehat{z}) - \lambda\|x-\widehat{x}\|\right\} \le \max_{x\in\widehat{\mathcal{X}}}\left\{\Psi(\tilde{f}(x),\widehat{z}) - \lambda\|x-\widehat{x}\|\right\} \le \max_{x\in\widehat{\mathcal{X}}}\left\{\Psi(\widehat{f}(x),\widehat{z}) - \lambda\|x-\widehat{x}\|\right\}.$$

Moreover, note that $\widehat{x}_k \mapsto \overline{z}_k$ and $\widehat{x}_k \mapsto \underline{z}_k$ are 1-Lipshitz since they are the max and min of a family of 1-Lipschitz function of $\widehat{x}_k$, and $1 \le \frac{\lambda}{b\vee h}$, so $\widehat{g}$ is $\frac{\lambda}{b\vee h}$-Lipschitz, and by definition $\underline{z}_k \le \widehat{g}(\widehat{x}_k) \le \overline{z}_k$. □

Proof of Proposition 1. We start by proving that

$$v_{\hat{D}} = \min_{\widehat{f}\in\widehat{\mathcal{F}},\,\lambda\ge b\vee h}\left\{\lambda\rho + \mathbb{E}_{\widehat{\mathbb{P}}}\left[\max_{x\in\widehat{\mathcal{X}}}\left\{\Psi_{\widehat{f}}(x,\widehat{Z}) - \lambda\|x-\widehat{X}\|\right\}\right]\right\}. \tag{EC.2}$$

To see this, consider maximizing over z first in the inner maximization of the dual problem

$$v_{\hat{D}} = \inf_{\widehat{f}\in\widehat{\mathcal{F}}}\inf_{\lambda\ge 0}\left\{\lambda\rho + \mathbb{E}_{\widehat{\mathbb{P}}}\left[\max_{x\in\widehat{\mathcal{X}}}\left\{\sup_{z\in\mathcal{Z}}\left\{\Psi(\widehat{f}(x),z) - \lambda|z-\widehat{Z}|\right\} - \lambda\|x-\widehat{X}\|\right\}\right]\right\}.$$

If $\lambda < b\vee h = b$, then the sup of

$$\Psi(\widehat{f}(x),z) - \lambda|z-\widehat{Z}| = h(\widehat{f}(x)-z)_+ + b(z-\widehat{f}(x))_+ - \lambda|z-\widehat{Z}|$$

will be infinity as $z \to \infty$. Therefore, in order to find the minimum over λ we can disregard this case and constrain $\lambda \geq b \vee h$. In this case, sup over z is $\Psi(\widehat{f}(x), \widehat{Z})$ attained at $z = \widehat{Z}$, which proves (EC.2).

For each $\lambda \geq b \vee h$, denote

$$\bar{\widehat{\mathcal{F}}} := \left\{ \widehat{g} \in \widehat{\mathcal{F}} : \underline{z}_k \leq \widehat{g}(\widehat{x}_k) \leq \bar{z}_k, \forall k \right\}, \qquad \widehat{\mathcal{F}}_\lambda := \left\{ \widehat{g} \in \widehat{\mathcal{F}} : \|\widehat{g}\|_{\mathrm{Lip}} \leq \frac{\lambda}{b \vee h} \right\}.$$

Then

$$v_{\hat{D}} = \inf_{\lambda \geq b \vee h} \inf_{\widehat{f} \in \bar{\widehat{\mathcal{F}}} \cap \widehat{\mathcal{F}}_\lambda} \left\{ \lambda \rho + \mathbb{E}_{\widehat{\mathbb{P}}} \left[\max_{x \in \mathcal{X}} \left\{ \Psi(\widehat{f}(x), \widehat{Z}) - \lambda \|x - \widehat{X}\| \right\} \right] \right\},$$

where we replace $\widehat{\mathcal{F}}$ by $\bar{\widehat{\mathcal{F}}} \cap \widehat{\mathcal{F}}_\lambda$ in (EC.2) using Lemma EC.3. For each $\widehat{f} \in \widehat{\mathcal{F}}_\lambda$, because $\|\Psi(\widehat{f}(\cdot), \widehat{z})\|_{\mathrm{Lip}} \leq \|\widehat{f}\|_{\mathrm{Lip}}(b \vee h) \leq \lambda$, the max over x is attained at $x = \widehat{X}$. Therefore,

$$v_{\hat{D}} = \inf_{\lambda \geq b \vee h} \inf_{\widehat{f} \in \bar{\widehat{\mathcal{F}}} \cap \widehat{\mathcal{F}}_\lambda} \left\{ \lambda \rho + \mathbb{E}_{\widehat{\mathbb{P}}} \left[\Psi(\widehat{f}(\widehat{X}), \widehat{Z}) \right] \right\}.$$

Now we switch the inf over λ and inf over $\widehat{f}$,

$$\begin{aligned} v_{\hat{D}} &= \inf_{\lambda \geq b \vee h} \inf_{\widehat{f} \in \bar{\widehat{\mathcal{F}}}, \|\widehat{f}\|_{\mathrm{Lip}} \leq \frac{\lambda}{b \vee h}} \left\{ \lambda \rho + \mathbb{E}_{\widehat{\mathbb{P}}} \left[\Psi(\widehat{f}(\widehat{X}), \widehat{Z}) \right] \right\} \\ &= \inf_{\widehat{f} \in \bar{\widehat{\mathcal{F}}}} \inf_{\lambda \geq b \vee h, \lambda \geq (b \vee h)\|\widehat{f}\|_{\mathrm{Lip}}} \left\{ \lambda \rho + \mathbb{E}_{\widehat{\mathbb{P}}} \left[\Psi(\widehat{f}(\widehat{X}), \widehat{Z}) \right] \right\} \\ &= \inf_{\widehat{f} \in \bar{\widehat{\mathcal{F}}}} \left\{ (b \vee h)(1 \vee \|\widehat{f}\|_{\mathrm{Lip}}) \rho + \mathbb{E}_{\widehat{\mathbb{P}}} \left[\Psi(\widehat{f}(\widehat{X}), \widehat{Z}) \right] \right\}. \end{aligned}$$

Using the change of variables $y_k = \widehat{f}(\widehat{x}_k)$, $k = 1, \ldots, K$, the above is equivalent to

$$v_{\hat{D}} = \inf_{y_k \in [\underline{z}_k, \bar{z}_k], 1 \leq k \leq K} \left\{ (b \vee h) \left(1 \vee \max_{i \neq j} \frac{|y_i - y_j|}{\|\widehat{x}_i - \widehat{x}_j\|} \right) \rho + \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \Psi(y_k, \widehat{z}_{ki}) \right\}.$$

This is an infimum over K variables which take values in closed intervals, a compact set, so the infimum is attained on a convex subset. Repeat the above argument with $\widehat{\mathcal{F}}$ in place of $\bar{\widehat{\mathcal{F}}}$, we conclude that $v_{\hat{D}} = v_{\hat{R}}$ and ($\widehat{\mathsf{D}}$) is equivalent to ($\widehat{\mathsf{R}}$). □

Appendix C: Additional Results for Section 4.2 and Proofs for Section 4.3

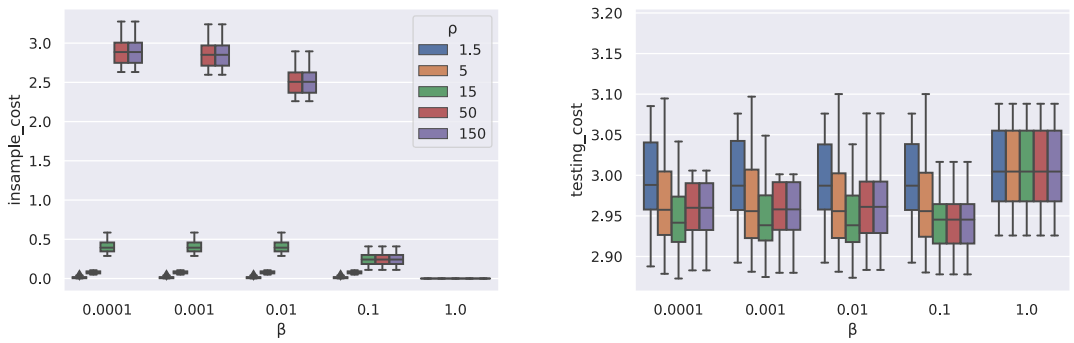

Figure EC.2 **Impact of hyperparameters ρ and β on the in-sample and out-of-sample performance**

In Figure EC.2, we illustrate the effect of different choices of hyperparameters on the model performance. We plot the in-sample costs and the out-of-sample costs in the case when $d = 5000$, $n = 100$, and $h = b = 1$, with various hyperparameters, under the same setup as in Section 5.1. The in-sample cost increases in the radius ρ and decreases in the distance weight β. As β becomes large ($\beta = 1$), in-sample cost reduces to zero and the policy becomes empirical risk minimization. This can be seen from the regularization point of view ($\widehat{\mathsf{R}}$): when we are indifferent in Lipschitz norms that are below a very high threshold—higher than the Lipschitz norm of the critical conditional quantile—the optimal policy is simply the unconditional quantile function. The out-of-sample performance seems less sensitive in β as long as it is sufficiently small (less than 0.1), but it has a clear preference in suitable tuning value of ρ. For large or small ρ, the policies are either too conservative or too restrictive, which lead to undesirable out-of-sample cost.

Proof for Proposition 3. First we give an upper bound for λ^* in terms of ρ, using the duality of the primal and the dual problem:

$$\lambda^* \rho \le v_D = v_P \le \max_{\mathcal{X}\times\mathcal{Z}} \Psi(f^*(x), z).$$

If the demand is bounded by $\bar{D}$, then $\mathcal{Z} \subset [0, \bar{D}]$ and subsequently the decision is also bounded by $f^*(x) \in [0, \bar{D}]$ for all x, by the boundedness of the Shapley extension. Therefore $|f^*(x) - z| \le \bar{D}$, thus $\Psi_{f^*}(x, z) \le (b \vee h)\bar{D}$, so

$$\|\Psi_{f^*}\|_{\mathrm{Lip}} = \lambda^* \le \frac{(b \vee h)\bar{D}}{\rho}.$$

Then the result follows from (Shalev-Shwartz and Ben-David 2014, Theorem 26.3 and Lemma 26.9). □

Appendix D: Proofs for Section 4.4

We recall and define $L = \|\widehat{f}^*\|_{\mathrm{Lip}}$ and

$$\begin{aligned} &\widehat{\mathcal{X}}_< := \{\widehat{x} \in \widehat{X} :\ \overline{q}(\widehat{x}) < \widehat{f}^*(\widehat{x})\}, \quad \widehat{\mathcal{X}}_> := \{\widehat{x} \in \widehat{X} :\ \underline{q}(\widehat{x}) > \widehat{f}^*(\widehat{x})\}, \\ &\widehat{\mathcal{X}}_= := \{\widehat{x} \in \widehat{X} :\ \underline{q}(\widehat{x}) \le \widehat{f}^*(\widehat{x}) \le \overline{q}(\widehat{x})\}, \quad \widehat{\mathcal{X}}_\ge := \widehat{\mathcal{X}}_> \cup \widehat{\mathcal{X}}_=, \quad \widehat{\mathcal{X}}_\le := \widehat{\mathcal{X}}_< \cup \widehat{\mathcal{X}}_=. \end{aligned} \tag{EC.3}$$

Proof of Proposition 4. For the first case we prove by constructing an optimal policy, and for the second case we prove by contradiction.

If condition (I) is satisfied, then an optimal policy can be constructed by the following algorithm.

`Step 1.` Define $y_k^* \leftarrow \overline{q}(\widehat{x}_k)$, $\forall k \in [K]$. By (I), we know that y_k^* satisfies

$$\underline{q}(\widehat{x}_j) - y_k^* \le \|\widehat{x}_j - \widehat{x}_k\|, \qquad \forall j, k \in [K]. \tag{EC.4}$$

Note that this also implies $\underline{q}(\widehat{x}_j) \le y_j^*$ for all j by setting $k = j$.

`Step 2.` Choose $k_1 \in \arg\min_k y_k^*$. For any $k \ne k_1$, denote $y_k^{(1)} = y_{k_1}^* + \|\widehat{x}_k - \widehat{x}_{k_1}\|$. Then for all j, k,

$$\underline{q}(\widehat{x}_j) - y_k^{(1)} = \underline{q}(\widehat{x}_j) - y_{k_1}^* - \|\widehat{x}_k - \widehat{x}_{k_1}\| \le \|\widehat{x}_j - \widehat{x}_{k_1}\| - \|\widehat{x}_k - \widehat{x}_{k_1}\| \le \|\widehat{x}_j - \widehat{x}_k\|.$$

This means that if we reassign values to $y_k^* \leftarrow y_k^* \wedge y_k^{(1)}$, $\forall k \ne k_1$, then (EC.4) would still hold. Note that since $y_k^{(1)} \ge y_{k_1}^*$, after reassignment $y_{k_1}^*$ is still the smallest among all y_k^*.

`Step 3.` Choose $k_2 \in \arg\min_{k \ne k_1} y_k^*$. For any $k \notin \{k_1, k_2\}$, denote

$$y_k^{(2)} = y_{k_2}^* + \|\widehat{x}_k - \widehat{x}_{k_2}\|,$$

and reassign $y_k^* \leftarrow y_k^* \wedge y_k^{(2)}$. Same as `Step 2`, (EC.4) still holds, and $y_{k_1}^* \le y_{k_2}^* \le y_k^*$ for all $k \notin \{k_1, k_2\}$.

`Step 4.` Repeat `Step 3`. Eventually we would have $y^*_{k_1} \le y^*_{k_2} \le \cdots \le y^*_{k_K}$, with (EC.4) still holds. Now for every $i < j$, we have

$$0 \le y^*_{k_j} - y^*_{k_i} \le y^{(i)}_{k_j} - y^*_{k_i} = \|\widehat{x}_{k_j} - \widehat{x}_{k_i}\|.$$

Therefore, define $\widehat{f}(\widehat{x}_k) := y^*_k$, then $\widehat{f}$ is a 1-Lipschitz function. In the above process y^*_k is decreasing its value, so $y^*_k \le \overline{q}(\widehat{x}_k)$ which is its initial value, and (EC.4) ensures $y^*_k \ge \underline{q}(\widehat{x}_k)$ by setting $j = k$.

Since $\underline{q} \le \widehat{f} \le \overline{q}$, $\widehat{f}$ is a conditional $\frac{b}{b+h}$-quantile, with $\|\widehat{f}\|_{\mathrm{Lip}} \le 1$. Then it must be a minimizer of $(\widehat{\mathsf{R}})$ for any $\rho \ge 0$, because it minimizes both terms. Any other optimal policy $\widehat{f}^*$ must also be a 1-Lipschitz quantile function to reach this minimum value. This completes the proof for the first part.

To see the second part, if condition (I) is not satisfied, then we claim that the optimizer $\widehat{f}^*$ must have Lipschitz constant $\|\widehat{f}^*\|_{\mathrm{Lip}} = L \ge 1$. Indeed, if $L < 1$, we can always adjust the value of $\widehat{f}^*$ to reduce costs in the second term of $v_{\widehat{R}}$ without paying more cost in the first term. So, the only possibility that $\|\widehat{f}^*\|_{\mathrm{Lip}} < 1$ is that the second term is already optimized, that is $\underline{q} \le \widehat{f}^* \le \overline{q}$. However, this would imply (I) holds, which is a contradiction.

Now we partition $\widehat{\mathcal{X}}$ according to (EC.3). First we fix $\widehat{x}_k \in \widehat{\mathcal{X}}_>$. Indeed, there must be $\widehat{x}_{j_1} \in \widehat{\mathcal{X}} \setminus \{\widehat{x}_k\}$ such that $\widehat{f}^*(\widehat{x}_k) - \widehat{f}^*(\widehat{x}_{j_1}) = L\|\widehat{x}_k - \widehat{x}_{j_1}\|$, otherwise we can increase the value of $\widehat{f}^*(\widehat{x}_k)$ and optimize the second term in $(\widehat{\mathsf{R}})$ without jeopardizing the first term. If $\widehat{x}_{j_1} \in \widehat{\mathcal{X}}_\le$, then the claim is proved. For the same reason, if $\widehat{x}_{j_1} \in \widehat{\mathcal{X}}_>$, then we can find $\widehat{x}_{j_2} \in \widehat{\mathcal{X}} \setminus \{\widehat{x}_k, \widehat{x}_{j_1}\}$ such that $\widehat{f}^*(\widehat{x}_{j_1}) - \widehat{f}^*(\widehat{x}_{j_2}) = L\|\widehat{x}_{j_1} - \widehat{x}_{j_2}\|$, thus $\widehat{f}^*(\widehat{x}_k) - \widehat{f}^*(\widehat{x}_{j_2}) = L\|\widehat{x}_k - \widehat{x}_{j_1}\| + L\|\widehat{x}_{j_1} - \widehat{x}_{j_2}\| \ge L\|\widehat{x}_k - \widehat{x}_{j_2}\|$, and here inequality sign must be equality because $\widehat{f}^*$ is L-Lipschitz. Note that this also shows that $(\widehat{x}_k, \widehat{f}^*(\widehat{x}_k))$, $(\widehat{x}_{j_1}, \widehat{f}^*(\widehat{x}_{j_1}))$ and $(\widehat{x}_{j_2}, \widehat{f}^*(\widehat{x}_{j_2}))$ are on the same straight line if $\|\cdot\| = \|\cdot\|_2$. If $\widehat{x}_{j_2} \in \widehat{\mathcal{X}}_\le$ then we finish the proof of the claim, otherwise $\widehat{x}_{j_3}$ can be found. Note that $\widehat{f}^*(\widehat{x}_k) > \widehat{f}^*(\widehat{x}_{j_1}) > \widehat{f}^*(\widehat{x}_{j_2}) > \dots$ is strictly decreasing, thus $\widehat{x}_k, \widehat{x}_{j_1}, \widehat{x}_{j_2} \dots$ are distinct. After finitely many steps, we must have $\widehat{x}_j \in \widehat{\mathcal{X}}_\le$ and $\widehat{f}^*(\widehat{x}_k) - \widehat{f}^*(\widehat{x}_j) = L\|\widehat{x}_k - \widehat{x}_j\|$ before we run out of points. □

Now we study the worst-case distribution under a optimal robust policy.

PROPOSITION EC.1 **(Worst-case Distribution)**. *Let $\widehat{f}^*$ be an in-sample optimal robust policy. Then there exists a worst-case distribution $\mathbb{P}^*$ of $(\widehat{\mathsf{P}})$ such that the following holds.*

(I) *If $\|\widehat{f}^*\|_{\mathrm{Lip}} \le 1$, then $\mathbb{P}^*$ perturbs $\widehat{\mathbb{P}}$ by moving $(\widehat{x}, \widehat{z})$ with $\widehat{z} \ge \widehat{f}^*(\widehat{x})$ toward $(\widehat{x}, z')$ for some $z' > \widehat{z}$.*

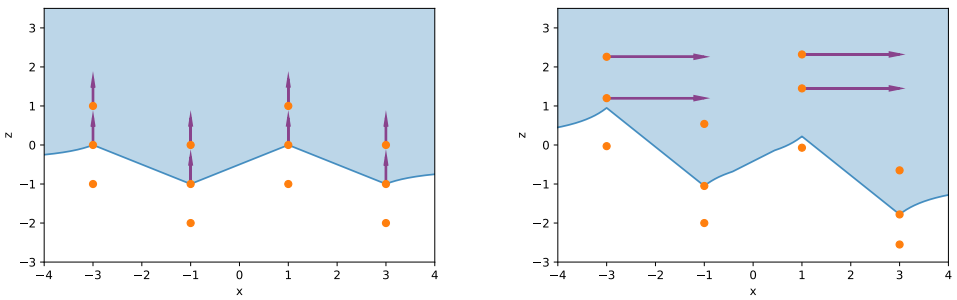

Figure EC.3 Transport map of worst-case distribution (purple arrows). When the optimal policy is 1-Lipschitz, worst-case distribution moves in z (left). Otherwise, worst-case distribution moves in x (right).

(II) *If* $\|\widehat{f}^*\|_{\mathrm{Lip}} > 1$, *then* $\mathbb{P}^*$ *perturbs* $\widehat{\mathbb{P}}$ *by moving each* $(\widehat{x},\widehat{z})$ *with* $\widehat{x} \in \widehat{\mathcal{X}}_>$ *and* $\widehat{z} \geq \widehat{f}^*(\widehat{x})$ *toward* $(x',\widehat{z})$ *for some* $x' \in \widehat{\mathcal{X}} \setminus \widehat{\mathcal{X}}_>$ *and* $\widehat{f}^*(\widehat{x}) - \widehat{f}^*(x') = \|\widehat{f}^*\|_{\mathrm{Lip}}\|\widehat{x} - x'\|$.

In Figure EC.3, we have the identical setting as in Figure 4. Above the graph of f^* is a blue shadow region representing $\{(x,z) : z \geq f^*(x)\}$, and $\mathbb{P}^*$ moves the probability mass in this region when backorder costs more than holding. In the left figure $\|\widehat{f}^*\|_{\mathrm{Lip}} < 1$, so it is more cost-efficient to move along the direction of z. In the right figure $\|\widehat{f}^*\|_{\mathrm{Lip}} > 1$, and it is more cost-efficient to move along the direction of x; since the worst-case distribution is for the in-sample problem, it perturbs x from one empirical value to another.

Proof of Proposition EC.1. To ease the notation we use f to represent $\widehat{f}^*$ in this proof.

(I) If $\|f\|_{\mathrm{Lip}} \leq 1$, then $\|\Psi_f\|_{\mathrm{Lip}} = b \vee h$. In this case we choose to transport Z instead of X. We define a transport map $T : \widehat{\mathcal{X}} \times \mathcal{Z} \to \widehat{\mathcal{X}} \times \mathcal{Z}$ by

$$T(x,z) := \begin{cases} (x, z+t) & z \geq f(x) \\ (x,z) & z < f(x) \end{cases}$$

for some t to be determined. Let $\mathbb{P} = T_{\#}\widehat{\mathbb{P}}$ be the push-forward of $\widehat{\mathbb{P}}$ via $T_{\#}\widehat{\mathbb{P}}$ defined by $\mathbb{P}[A] = \widehat{\mathbb{P}}[T^{-1}(A)]$ for every measurable set $A \subset \widehat{\mathcal{X}} \times \mathcal{Z}$, then

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}[\Psi_f] - \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi_f] &= \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi_f \circ T(X,Z)] - \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi_f(X,Z)] \\
&= \mathbb{E}_{\widehat{\mathbb{P}}}[\mathbf{1}_{\{Z \geq f(X)\}}(\Psi(f(X), Z+t) - \Psi(f(X),Z))] \\
&= \mathbb{E}_{\widehat{\mathbb{P}}}[\mathbf{1}_{\{Z \geq f(X)\}}(b(Z+t-f(X)) - b(Z-f(X)))] \\
&= bt\widehat{\mathbb{P}}[Z \geq f(X)] \\
&= b\mathbb{E}_{\widehat{\mathbb{P}}}[\|T(X,Z) - (X,Z)\|] \\
&\geq b\mathcal{W}_1(\mathbb{P}, \widehat{\mathbb{P}}).
\end{aligned}$$

By choosing $t = \rho / \widehat{\mathbb{P}}[Z \geq f(X)]$, we have $\mathcal{W}(\mathbb{P}, \widehat{\mathbb{P}}) = \rho$, so $\mathbb{P}$ is feasible, and $\mathbb{E}_{\mathbb{P}}[\Psi_f] = \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi_f] + b\rho$ is a worst case distribution. Note that the denominator $\widehat{\mathbb{P}}[Z \geq f(X)]$ is never zero, otherwise $\widehat{\mathcal{X}} = \widehat{\mathcal{X}}_<$ and $\widehat{\mathcal{X}}_{\geq} = \emptyset$ which contradicts with Proposition 4.

(II) If $\|f\|_{\mathrm{Lip}} = L > 1$, then $\|\Psi_f\|_{\mathrm{Lip}} = (b \vee h)L$. In this case we choose to transport X instead of Z. In order to find a worst case distribution, we are interested in how far this $\widehat{x}_j \in \widehat{\mathcal{X}}_{\leq}$ can be from $\widehat{x}_k \in \widehat{\mathcal{X}}_>$ in Proposition 4 (II). For every k we define

$$\tau(k) \in \operatorname*{arg\,min}_{j \neq k} \left\{ \widehat{f}^*(\widehat{x}_j) : \widehat{f}^*(\widehat{x}_k) - \widehat{f}^*(\widehat{x}_j) = L\|\widehat{x}_k - \widehat{x}_j\| \right\}, \qquad \Delta(\widehat{x}_k) := \|\widehat{x}_k - \widehat{x}_{\tau(k)}\| = \frac{1}{L}(\widehat{f}^*(\widehat{x}_k) - \widehat{f}^*(\widehat{x}_{\tau(k)})). \tag{EC.5}$$

Intuitively, whenever $\widehat{x}_k \in \widehat{\mathcal{X}}_>$, $\tau(k)$ specifies a moving direction from $\widehat{x}_k$ to $\widehat{x}_{\tau(k)}$, and $\Delta(\widehat{x}_k)$ denotes the moving distance.

We define a transport map $T : \widehat{\mathcal{X}} \times \mathcal{Z} \to \widehat{\mathcal{X}} \times \mathcal{Z}$ by

$$T(\widehat{x}_k, z) := \begin{cases} (\widehat{x}_{\tau(k)}, z) & x \in \mathcal{X}_>, z \geq f(x) \\ (\widehat{x}_k, z) & x \in \mathcal{X}_{\leq} \text{ or } z < f(x) \end{cases}.$$

This implies that for every k,

$$\begin{aligned}\Psi_f \circ T(\widehat{x}_k, z) - \Psi_f(\widehat{x}_k, z) &= \mathbf{1}_{\{\widehat{x}_k \in \mathcal{X}_>, z \geq \widehat{f}^*(\widehat{x}_k)\}} (b(z - \widehat{f}^*(\widehat{x}_{\tau(k)})) - b(z - \widehat{f}^*(\widehat{x}_k))) \\ &= b\mathbf{1}_{\{\widehat{x}_k \in \mathcal{X}_>, z \geq \widehat{f}^*(\widehat{x}_k)\}} L\|\widehat{x}_k - \widehat{x}_{\tau_k}\| \\ &= bL\mathbf{1}_{\{\widehat{x}_k \in \mathcal{X}_>, z \geq \widehat{f}^*(\widehat{x}_k)\}} \Delta(\widehat{x}_k) \\ &= bL\|T(\widehat{x}_k, z) - (\widehat{x}_k, z)\|\end{aligned}$$

Let $\tilde{\mathbb{P}} = T_{\#}\widehat{\mathbb{P}}$, then

$$\mathbb{E}_{\tilde{\mathbb{P}}}[\Psi_f] - \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi_f] = \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi_f \circ T(X, Z) - \Psi_f(X, Z)] = bL\mathbb{E}_{\widehat{\mathbb{P}}}[\|T(\widehat{x}_k, z) - (\widehat{x}_k, z)\|] \geq bL\mathcal{W}_1(\tilde{\mathbb{P}}, \widehat{\mathbb{P}}).$$

We will show that

$$\mathbb{E}_{\tilde{\mathbb{P}}}[\Psi_f] - \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi_f] \geq bL\rho. \tag{EC.6}$$

If this is true, we can construct a feasible $\mathbb{P}$ by a convex combination of the form $\mathbb{P} = \alpha\tilde{\mathbb{P}} + (1-\alpha)\widehat{\mathbb{P}}$, such that

$$\mathbb{E}_{\mathbb{P}}[\Psi_f] - \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi_f] = bL\rho \geq bL\mathcal{W}(\mathbb{P}, \widehat{\mathbb{P}}),$$

so $\mathbb{P}$ is a feasible worst case distribution. It is not hard to see that $\mathbb{P}$ is exactly ρ away from $\widehat{\mathbb{P}}$.

Recall Δ is defined in (EC.5). To prove (EC.6), we construct another solution $f^\varepsilon(\widehat{x}_k) := \widehat{f}^*(\widehat{x}_k) - \varepsilon\Delta(\widehat{x}_k)$ which means "ordering less" than $\widehat{f}^*$. We claim when ε is small, $\|f^\varepsilon\|_{\text{Lip}} = L - \varepsilon$. To see why this is true, we can consider only the pairs of points $\widehat{x}_k$ and $\widehat{x}_j$ with $\widehat{f}^*(\widehat{x}_k) - \widehat{f}^*(\widehat{x}_j) = L\|\widehat{x}_k - \widehat{x}_j\|$ since ε can be chosen sufficiently small. In this situation,

$$f^\varepsilon(\widehat{x}_k) - f^\varepsilon(\widehat{x}_j) = \widehat{f}^*(\widehat{x}_k) - \widehat{f}^*(\widehat{x}_j) - \varepsilon(\Delta(\widehat{x}_k) - \Delta(\widehat{x}_j)) = \widehat{f}^*(\widehat{x}_k) - \widehat{f}^*(\widehat{x}_j) - \frac{\varepsilon}{L}\left((\widehat{f}^*(\widehat{x}_k) - \widehat{f}^*(\widehat{x}_{\tau(k)})) - (\widehat{f}^*(\widehat{x}_j) - \widehat{f}^*(\widehat{x}_{\tau(j)}))\right).$$

It can be seen that $\widehat{f}^*(\widehat{x}_{\tau(k)}) \leq \widehat{f}^*(\widehat{x}_{\tau(j)})$ by the minimality of $\tau(k)$ (see the last paragraph in the proof of Proposition 4), hence

$$f^\varepsilon(\widehat{x}_k) - f^\varepsilon(\widehat{x}_j) \leq \widehat{f}^*(\widehat{x}_k) - \widehat{f}^*(\widehat{x}_j) - \frac{\varepsilon}{L}(\widehat{f}^*(\widehat{x}_k) - \widehat{f}^*(\widehat{x}_j)) = \left(1 - \frac{\varepsilon}{L}\right)(\widehat{f}^*(\widehat{x}_k) - \widehat{f}^*(\widehat{x}_j)) = (L - \varepsilon)\|\widehat{x}_k - \widehat{x}_j\|.$$

Therefore f^ε is $(L-\varepsilon)$-Lipschitz.

Since f minimizes $v_{\hat{R}}$, we have

$$\begin{aligned}(b \vee h)L\rho + \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi_f] &\leq (b \vee h)(L - \varepsilon)\rho + \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi_{f^\varepsilon}] \\ (b \vee h)\varepsilon\rho &\leq \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi_{f^\varepsilon} - \Psi_f].\end{aligned}$$

Here, "ordering less" means we need to pay more backorder cost and less holding cost, so

$$\Psi(f^\varepsilon(\widehat{x}_k), z) - \Psi(f(\widehat{x}_k), z) = \begin{cases} b\varepsilon\Delta(\widehat{x}_k), & \widehat{f}^*(\widehat{x}_k) \leq z, \\ -h\varepsilon\Delta(\widehat{x}_k), & \widehat{f}^*(\widehat{x}_k) > z. \end{cases}$$

Here we choose ε small such that $\widehat{f}^*(\widehat{x}_k) > z$ implies $\widehat{f}^*(\widehat{x}_k) > z + \varepsilon\Delta(\widehat{x}_k)$ for every $(\widehat{x}_k, z) \in \operatorname{supp}\widehat{\mathbb{P}}$. Now take the conditional expectation,

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\hat{Z}|\hat{X}}}[\Psi(f^\varepsilon(X), Z) - \Psi(f(X), Z) | X = \widehat{x}_k] = \varepsilon\Delta(\widehat{x}_k)\left(b\widehat{\mathbb{P}}[Z \geq \widehat{f}^*(\widehat{x}_k) | X = \widehat{x}_k] - h\widehat{\mathbb{P}}[Z < \widehat{f}^*(\widehat{x}_k) | X = \widehat{x}_k]\right).$$

If $\widehat{x}_k \in \mathcal{X}_{\leq}$, $\widehat{f}^*(\widehat{x}_k)$ is no less than the conditional quantile, so the above will be nonpositive. Thus

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[\Psi(f^{\varepsilon}(X),Z)-\Psi(f(X),Z)|X=\widehat{x}_k] \leq \begin{cases} b\varepsilon\Delta(\widehat{x}_k)\widehat{\mathbb{P}}[Z\geq \widehat{f}^*(\widehat{x}_k)|X=\widehat{x}_k], & \widehat{x}_k\in\mathcal{X}_{>}, \\ 0, & \widehat{x}_k\in\mathcal{X}_{\leq}. \end{cases}$$

Finally, take expectation in X, we have

$$(b\vee h)\varepsilon\rho \leq \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi_{f^{\varepsilon}}-\Psi_f] \leq b\varepsilon\mathbb{E}_{\widehat{\mathbb{P}}}[\Delta(X)\mathbf{1}_{\{X\in\mathcal{X}_{>},Z\geq f(X)\}}].$$

In particular, $\mathbb{E}_{\widehat{\mathbb{P}}}[\Psi_f\circ T-\Psi_f]=bL\mathbb{E}_{\widehat{\mathbb{P}}}[\Delta(X)\mathbf{1}_{\{X\in\mathcal{X}_{>},Z\geq f(X)\}}]\geq bL\rho$, which proves (EC.6). □

REMARK EC.1. Similarly, when $b<h$, the transport map should move $X\in\widehat{\mathcal{X}}_{<}$ in $\{Z\leq\widehat{f}^*(X)\}$ when $L>1$, and should decrease Z in $\{Z\leq\widehat{f}^*(X)\}$ when $L\leq 1$. Considering the demand must be nonnegative, in the case $L\leq 1$ we should use the transport map

$$T(x,z):=\begin{cases}(x,(z-t)_+) & z\geq\widehat{f}^*(x)\\ (x,z) & z\leq\widehat{f}^*(x)\end{cases}.$$

for some $t\geq 0$ if $\rho\leq\mathbb{E}_{\widehat{\mathbb{P}}}[Z\mathbf{1}\{Z<\widehat{f}^*(X)\}]$. From this proposition we can see that, when ρ is sufficiently small, the worst case distribution $\mathbb{P}$ is still supported in $\widehat{\mathcal{X}}\times\mathcal{Z}$.

Essentially the only place where $b\geq h$ really matters is the proof of (EC.2), where we send $z\to\infty$. When $b<h$, one would send $z\to-\infty$ instead, which would be absurd because $z\geq 0$. However, if we start with $\mathcal{Z}=\mathbb{R}$ instead of $\mathbb{R}_+$ in ($\widehat{\mathsf{P}}$):

$$v_{\hat{P}}=\min_{\widehat{f}:\widehat{\mathcal{X}}\to\mathbb{R}}\sup_{\mathbb{P}\in\mathcal{P}_1(\widehat{\mathcal{X}}\times\mathbb{R})}\left\{\mathbb{E}_{(X,Z)\sim\mathbb{P}}[\Psi_{\widehat{f}}(X,Z)]:\mathcal{W}(\mathbb{P},\widehat{\mathbb{P}})\leq\rho\right\},$$

we would have the same worst case distribution supported over $\{Z\geq 0\}$, so the value of $v_{\hat{P}}$ is going to be the same. Since the proof of $v_P=v_D=v_{\hat{D}}=v_{\hat{P}}$ in Theorem 1 doesn't rely on whether $\mathcal{Z}=\mathbb{R}$ or $\mathbb{R}_+$ because of the property (iii) in the Lemma 1, $\mathcal{Z}$ in (P), (D), ($\widehat{\mathsf{P}}$), ($\widehat{\mathsf{D}}$) can all be replaced by $\mathbb{R}$, and thus we can send $z\to-\infty$ to fix the proof for (EC.2).

Appendix E: Additional Numerical Results for Section 5.1

In this section, we provide additional numerical results for robustness check.

Methods	Hyperparameters
`Shapley`	radius and norm scaling parameter
`KO`	Bandwidth
`kNN`	No. of neighbours
`ERM2` (ℓ^1/ℓ^2)	Regularization parameter
`RandForest`	
`StochOptForest (apx-soln)`	None
`StochOptForest (apx-risk)`	

Table EC.1 Hyperparameters used in the methods

In Figure EC.4, we consider various noise distributions while fixing $n=100$, and $d=5000$. The model setup is the same as in Section 5.1 and the results are presented in Figures EC.4. We consider three different noise distributions for ε as considered by Zhu et al. (2012): (i) the Laplace distribution; (ii) the student-t

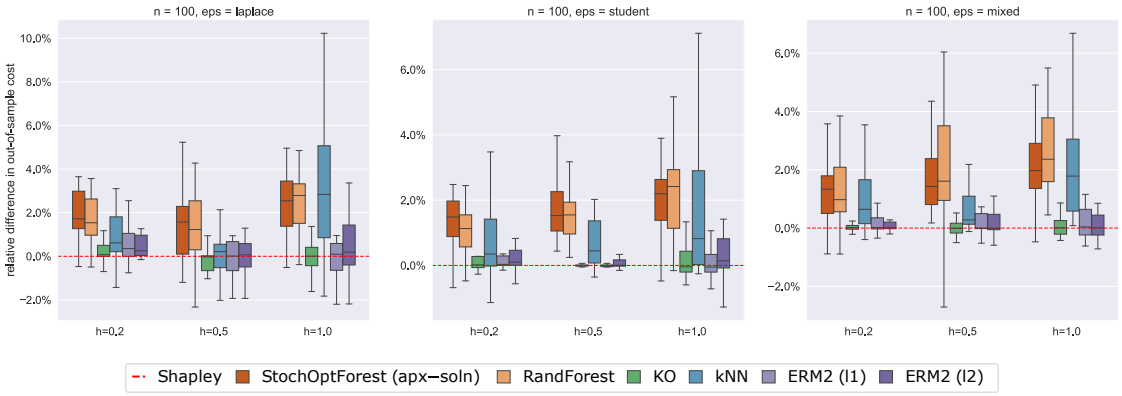

Figure EC.4 Boxplots of the out-of-sample performance for models with different noises

distribution with 3 degrees of freedom; and (iii) Gaussian mixture $0.8\mathcal{N}(0,1)+0.2\mathcal{N}(0,9)$. It can be seen that the observations that we made consistently hold for all three noise distributions.

In Figure EC.5, we consider a finer and larger grid for h while fixing $h+b=1$ to study the impact on the extremity of fractiles. Need to notice that the performance is not symmetric since the distribution of the non-linear model is asymmetric. We found that the performance of `Shapley` tends to be better than other non-parametric benchmarks (`KO`, k`NN` and forest-based methods) when the fractile becomes extreme, which demonstrate the advantages of a robust approach.

In Figure EC.6, to study the impact of sparsity, we consider a sparse model as in Zhu et al. (2012) where $\beta_0=(2,-2,-1,1,0,\ldots,0)^\top/\sqrt{10}$ is a d-dimensional vector, $d=5000$, and the covariate $X\in\mathbb{R}^d$ is a multivariate Gaussian random variable with mean zero and covariance matrix $(\sigma_{ij})_{d\times d}$ with $\sigma_{ij}=0.5^{|i-j|}$, and ε is the standard Gaussian distribution. We consider $n=10,100$ only because 1000 seems too large for this sparse model. Under this case, our `Shapley` policy still maintains competitive performance, although its advantage is no longer dominant.

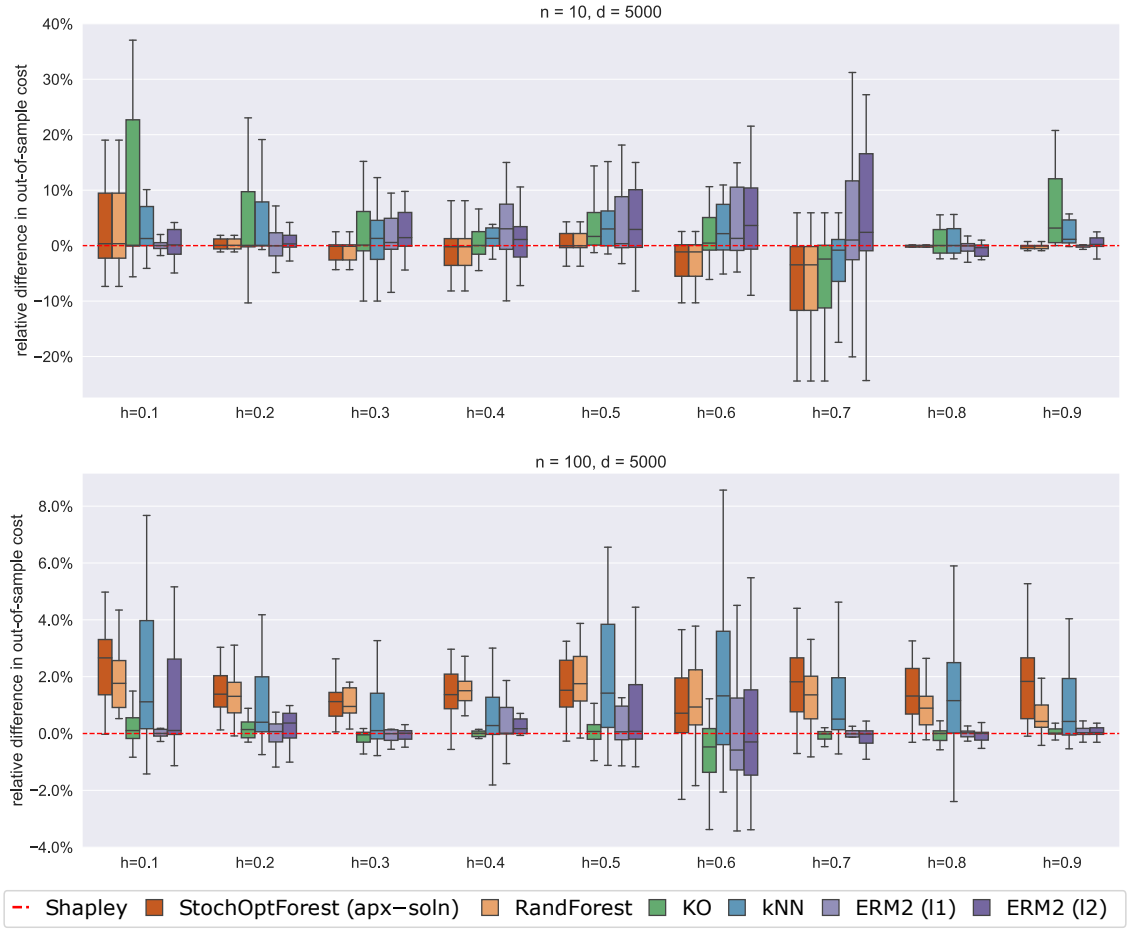

Figure EC.5 Boxplots of the out-of-sample performance for models with different h/b-ratios

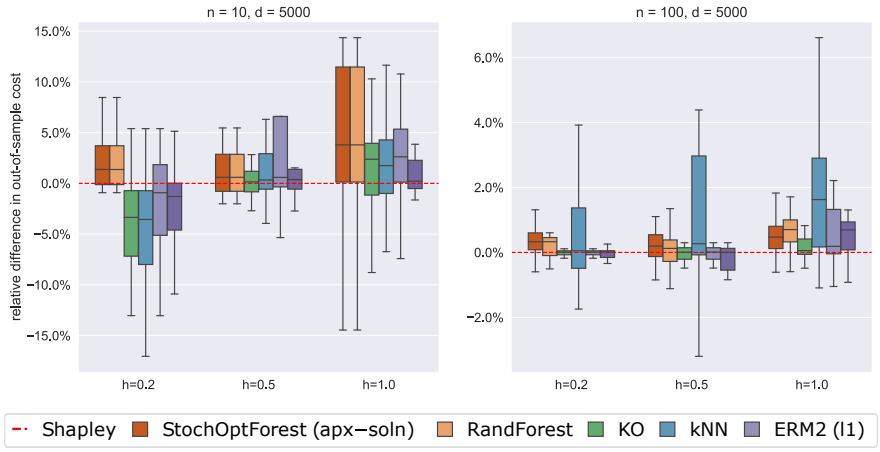

Figure EC.6 Differences in the out-of-sample performance between `Shapley` and other benchmarks in sparse case