Online Appendix for "Optimize-via-Estimate: How to make finite-sample corrections for sampling uncertainty in data-driven optimization"

A. Deferred proofs

Here, we present the proofs that have been deferred from the main text.

A.1. Proposition 1

Proof of Proposition 1. Note that by the optimality of $\boldsymbol{x}_{\text{ETO}}(\boldsymbol{y}^\star)$,

$$\phi\left(\boldsymbol{x}_{\text{ETO}}(\boldsymbol{y}^\star);\hat{\boldsymbol{\theta}}(\boldsymbol{y}^\star)\right) \leq \phi\left(\boldsymbol{x}^\dagger_{\text{ETO}}(\boldsymbol{y});\hat{\boldsymbol{\theta}}(\boldsymbol{y}^\star)\right), \qquad \forall \boldsymbol{y}\in\mathcal{Y}.$$

As g is non-negative and the above result holds for all $\boldsymbol{y}$, we have that

$$\int_{\mathcal{Y}} \phi\left(\boldsymbol{x}_{\text{ETO}}(\boldsymbol{y}^\star);\hat{\boldsymbol{\theta}}(\boldsymbol{y}^\star)\right) g\left(\boldsymbol{y};\hat{\boldsymbol{\theta}}(\boldsymbol{y}^\star)\right) d\boldsymbol{y} \leq \int_{\mathcal{Y}} \phi\left(\boldsymbol{x}^\dagger_{\text{ETO}}(\boldsymbol{y});\hat{\boldsymbol{\theta}}(\boldsymbol{y}^\star)\right) g\left(\boldsymbol{y};\hat{\boldsymbol{\theta}}(\boldsymbol{y}^\star)\right) d\boldsymbol{y}.$$

However, $\bar{\boldsymbol{x}}_{\text{ETO}}(\boldsymbol{y}) = \boldsymbol{x}_{\text{ETO}}(\boldsymbol{y}^\star)$, thus,

$$\Phi\left[\bar{\boldsymbol{x}}_{\text{ETO}}(\cdot)\,\middle|\,\hat{\boldsymbol{\theta}}(\boldsymbol{y}^\star)\right] \leq \Phi\left[\boldsymbol{x}^\dagger_{\text{ETO}}(\cdot)\,\middle|\,\hat{\boldsymbol{\theta}}(\boldsymbol{y}^\star)\right].$$

Conversely, by definition of the optimality of $\boldsymbol{x}^\dagger_{\text{ETO}}$, we have

$$\Phi\left[\boldsymbol{x}^\dagger_{\text{ETO}}(\cdot)\,\middle|\,\hat{\boldsymbol{\theta}}(\boldsymbol{y})\right] \leq \Phi\left[\bar{\boldsymbol{x}}_{\text{ETO}}(\cdot)\,\middle|\,\hat{\boldsymbol{\theta}}(\boldsymbol{y})\right],$$

which completes the proof. □

A.2. Lemma 1

Proof of Lemma 1. Firstly, we show that Assumption 2 guarantees that for any given $\boldsymbol{\theta}\in\Theta$, there exists a minimizer of $\phi(\cdot;\boldsymbol{\theta})$. If $\mathcal{B}$ is bounded, the existence of the minimizer follows immediately. We focus on the case where $\mathcal{B}$ is unbounded. Suppose there does not exist a minimizer. Let $f^* := \inf_{\boldsymbol{x}\in\mathcal{B}} \phi(\boldsymbol{x};\boldsymbol{\theta})$. It is finite by assumption. Let $\{\boldsymbol{x}_k\}$ be a sequence in $\mathcal{B}$ such that $\phi(\boldsymbol{x}_k;\boldsymbol{\theta}) \to f^*$. We now have two mutually exclusive cases:

(i) $\underline{\sup_k \|\boldsymbol{x}_k\| = \rho < \infty}$: Define $\mathcal{B}_d := \{x\in\mathbb{R}^n : \|\boldsymbol{x}\| \leq \rho\}$. Then for all k, $\boldsymbol{x}_k \in \mathcal{B}\cap\mathcal{B}_d$ which is closed, bounded and convex. Thus, there exists some $x^*\in\mathcal{B}\cap\mathcal{B}_d$ that is a minimizer. However, as there already exists a sequence in $\mathcal{B}\cap\mathcal{B}_d$ that allows $\phi(\cdot;\boldsymbol{\theta})$ to get arbitrarily close to f^*, thus $\phi(\boldsymbol{x}^*;\boldsymbol{\theta}) = f^*$, which is a contradiction.

(ii) $\underline{\sup_k \|\boldsymbol{x}_k\| = \infty}$: This implies that $\|\boldsymbol{x}_k\| \to \infty$. Thus, we found a sequence where $\|\boldsymbol{x}_k\| \to \infty$, but $\phi(\boldsymbol{x}_k;\boldsymbol{\theta}) \to f^*$, contradicting the assumption that $\phi(\boldsymbol{x}_k;\boldsymbol{\theta}) \to \infty$.

Thus, $\phi(\boldsymbol{x};\boldsymbol{\theta})$ has a minimizer on $\mathcal{B}$.

Next, we show that $\phi(\boldsymbol{x};\boldsymbol{\theta})$ has a unique minimizer. Let $\boldsymbol{x}^*$ denote a minimizer of $\phi(\boldsymbol{x};\boldsymbol{\theta})$ and it is easy to verify that $\nabla\phi(\boldsymbol{x}^*;\boldsymbol{\theta})^\top(\boldsymbol{x}-\boldsymbol{x}^*)\geq 0$ for any $\boldsymbol{x}\in\mathcal{B}$ (Boyd and Vandenberghe 2004). Now, strict convexity implies that $\phi(\boldsymbol{x};\boldsymbol{\theta})>\phi(\boldsymbol{x}^*;\boldsymbol{\theta})+\nabla\phi(\boldsymbol{x}^*;\boldsymbol{\theta})^\top(\boldsymbol{x}-\boldsymbol{x}^*)\geq\phi(\boldsymbol{x}^*;\boldsymbol{\theta})$ for any $\boldsymbol{x}\in\mathcal{B}$ and $\boldsymbol{x}\neq\boldsymbol{x}^*$. Thus, $\boldsymbol{x}^*$ is a unique minimizer. □

A.3. Theorem 1

Proof of Theorem 1. Suppose there exists some $\boldsymbol{x}(\cdot)\in\bar{\mathcal{X}}$, that is uniformly optimal. Consider the shifted function $\check{\boldsymbol{x}}(\boldsymbol{y})=\boldsymbol{x}(\boldsymbol{y})+\epsilon\boldsymbol{r}(\boldsymbol{y})$ for some choice of continuous function $\boldsymbol{r}(\cdot)$ and $\epsilon\neq 0$. It is clear that $\check{\boldsymbol{x}}\in\bar{\mathcal{X}}$. For all $\theta\in\Theta$,

$$\Phi[\check{\boldsymbol{x}}(\cdot)|\boldsymbol{\theta}]=\Phi[\boldsymbol{x}(\cdot)|\boldsymbol{\theta}]+\epsilon\int_{\mathcal{Y}}\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}(\boldsymbol{y});\boldsymbol{\theta})^\top\boldsymbol{r}(\boldsymbol{y})g(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y}+O(\epsilon^2).$$

Suppose first that for some $\boldsymbol{\theta}\in\Theta$, $\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}(\boldsymbol{y});\boldsymbol{\theta})=\mathbf{0}$ for all $\boldsymbol{y}\in\mathcal{Y}$. Assumption 2 guarantees a unique minimizer for ϕ, say $\boldsymbol{x}^*(\boldsymbol{\theta})$. As $\boldsymbol{x}^*$ is a prescriptive solution, it is independent of $\boldsymbol{\theta}$. Thus, $\boldsymbol{x}(\boldsymbol{y})\equiv\boldsymbol{x}^*$, which implies that $\nabla_x\phi(\boldsymbol{x}(\boldsymbol{y});\boldsymbol{\theta})=\mathbf{0}$ for all $\boldsymbol{y}\in\mathcal{Y}$ and for all $\boldsymbol{\theta}\in\Theta$. For every $\boldsymbol{\theta}\in\Theta$, this induces the optimality condition that $\boldsymbol{x}(\boldsymbol{y})\equiv\boldsymbol{x}^*(\boldsymbol{\theta})$. However, by assumption, there are at least two different $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ where their minimas do not coincide $\boldsymbol{x}^*(\boldsymbol{\theta}_1)\neq\boldsymbol{x}^*(\boldsymbol{\theta}_2)$, leading to a contradiction.

Thus, we may now assume that for some $\boldsymbol{\theta}\in\Theta$, $\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}(\boldsymbol{y});\boldsymbol{\theta})\neq\mathbf{0}$ for some non-zero measure set of data $\boldsymbol{y}\in\mathcal{Y}$. Define

$$\boldsymbol{r}(\boldsymbol{y})=-\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}(\boldsymbol{y});\boldsymbol{\theta}),$$

for that particular $\boldsymbol{\theta}$. Such an $\boldsymbol{r}$ is continuous as ϕ is assumed to be continuously differentiable in the first argument, and $\boldsymbol{x}$ lies in the space of continuous functions. But this definition implies that

$$\int_{\mathcal{Y}}\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}(\boldsymbol{y});\boldsymbol{\theta})^\top\boldsymbol{r}(\boldsymbol{y})g(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y}=\int_{\mathcal{Y}}-\boldsymbol{r}(\boldsymbol{y})^\top\boldsymbol{r}(\boldsymbol{y})g(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y}<0,$$

with strict inequality as we had assume that $\boldsymbol{r}(\boldsymbol{y})\neq\mathbf{0}$ over a set of non-zero measure. This means that we have constructed a continous $\boldsymbol{r}(\cdot)$ for some $\boldsymbol{\theta}\in\Theta$ that leads to a negative integral, once again, contradicting the (uniform) optimality of $\boldsymbol{x}(\cdot)$. □

A.4. Theorem 2

Proof of Theorem 2a.

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}}[\phi(x_{\mathrm{OS}}(\boldsymbol{y});\theta)]&=\int_{\boldsymbol{y}\in\mathbb{R}_+^N}\left(\int_{z=0}^{\infty}C\left(x_{\mathrm{OS}}(\boldsymbol{y}),z\right)f(z;\theta)\mathrm{d}z\right)g\left(\boldsymbol{y};\theta\right)\mathrm{d}\boldsymbol{y}\\&=\int_{r=0}^{\infty}\int_{\{\boldsymbol{y}:\boldsymbol{y}^\top\mathbf{1}=r\}}\left(\int_{z=0}^{\infty}C\left(x_{\mathrm{OS}}(\boldsymbol{y}),z\right)f(z;\theta)\mathrm{d}z\right)g\left(\boldsymbol{y};\theta\right)J\mathrm{d}\boldsymbol{y}\,\mathrm{d}r,\end{aligned}$$

where we have mapped the space $\{\boldsymbol{y} \in \mathbb{R}_+^N\}$ to the space $\left\{(r, \boldsymbol{y}) \in \mathbb{R}_+^{N+1} : \boldsymbol{y}^\top \mathbf{1} = r\right\}$. As this is a linear projection, its Jacobian J is a constant. From here,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{y}}[\phi(x_{\mathrm{OS}}(\boldsymbol{y}); \theta)] &= \int_{r=0}^{\infty} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \left(\int_{z=0}^{\infty} C\left(x_{\mathrm{OS}}(r\boldsymbol{y}), z\right) f(z;\theta) \mathrm{d}z \right) g\left(r\boldsymbol{y}; \theta\right) r^{N-1} J \mathrm{d}\boldsymbol{y} \, \mathrm{d}r \\
&= \int_{r=0}^{\infty} r^{\eta} r^{N-1} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \left(\int_{z=0}^{\infty} C\left(x_{\mathrm{OS}}(\boldsymbol{y}), \frac{z}{r}\right) f(z;\theta) \mathrm{d}z \right) g\left(r\boldsymbol{y}; \theta\right) J \mathrm{d}\boldsymbol{y} \, \mathrm{d}r \\
&= \int_{r=0}^{\infty} r^{\eta} r^{N-1} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \left(\int_{s=0}^{\infty} C\left(x_{\mathrm{OS}}(\boldsymbol{y}), s\right) f(rs;\theta) r \mathrm{d}s \right) g\left(r\boldsymbol{y}; \theta\right) J \mathrm{d}\boldsymbol{y} \, \mathrm{d}r \\
&= \int_{r=0}^{\infty} r^{\eta} r^{N-1} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \left(\int_{s=0}^{\infty} C\left(x_{\mathrm{OS}}(\boldsymbol{y}), s\right) f\left(s; \frac{\theta}{r}\right) \mathrm{d}s \right) g\left(\boldsymbol{y}; \frac{\theta}{r}\right) r^{-N} J \mathrm{d}\boldsymbol{y} \, \mathrm{d}r.
\end{aligned}$$

Here, we perform the substitution $\vartheta = \theta / r$ ($d\vartheta = -dr\theta/r^2 = -dr\vartheta^2/\theta$):

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{y}}[\phi(x_{\mathrm{OS}}(\boldsymbol{y}); \theta)] =& - \int_{\vartheta=\infty}^{0} \left(\frac{\theta}{\vartheta}\right)^{\eta-1} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \left(\int_{s=0}^{\infty} C\left(x_{\mathrm{OS}}(\boldsymbol{y}), s\right) f\left(s; \vartheta\right) \mathrm{d}s \right) g\left(\boldsymbol{y}; \vartheta\right) J \mathrm{d}\boldsymbol{y} \frac{\theta}{\vartheta^2} \, \mathrm{d}\vartheta \\
=& \theta^{\eta} \int_{\vartheta=0}^{\infty} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \left(\int_{s=0}^{\infty} C\left(x_{\mathrm{OS}}(\boldsymbol{y}), s\right) f\left(s; \vartheta\right) \mathrm{d}s \right) g\left(\boldsymbol{y}; \vartheta\right) \frac{1}{\vartheta^{\eta+1}} J \mathrm{d}\boldsymbol{y} \, \mathrm{d}\vartheta.
\end{aligned}$$

Rearranging terms, we have that

$$\begin{aligned}
\text{LHS} =& \theta^{\eta} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \int_{s=0}^{\infty} C\left(x_{\mathrm{OS}}(\boldsymbol{y}), s\right) \left(\int_{\vartheta=0}^{\infty} f\left(s; \vartheta\right) g\left(\boldsymbol{y}; \vartheta\right) \frac{\mathrm{d}\vartheta}{\vartheta^{\eta+1}} \right) \mathrm{d}s J \mathrm{d}\boldsymbol{y} \\
=& \theta^{\eta} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \int_{s=0}^{\infty} C\left(x_{\mathrm{OS}}(\boldsymbol{y}), s\right) \frac{\int_{\vartheta=0}^{\infty} f\left(s; \vartheta\right) g\left(\boldsymbol{y}; \vartheta\right) \frac{\mathrm{d}\vartheta}{\vartheta^{\eta+1}}}{\int_{\alpha=0}^{\infty} g\left(\boldsymbol{y}; \alpha\right) \frac{\mathrm{d}\alpha}{\alpha^{\eta+1}}} \mathrm{d}s J \left(\int_{\alpha=0}^{\infty} g\left(\boldsymbol{y}; \alpha\right) \frac{\mathrm{d}\alpha}{\alpha^{\eta+1}} \right) \mathrm{d}\boldsymbol{y} \\
=& \theta^{\eta} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \int_{s=0}^{\infty} C\left(x_{\mathrm{OS}}(\boldsymbol{y}), s\right) \hat{f}_{\mathrm{OS}}(s, \boldsymbol{y}) \mathrm{d}s J \left(\int_{\alpha=0}^{\infty} g\left(\boldsymbol{y}; \alpha\right) \frac{\mathrm{d}\alpha}{\alpha^{\eta+1}} \right) \mathrm{d}\boldsymbol{y},
\end{aligned}$$

where we have abbreviated

$$\hat{f}_{\mathrm{OS}}\left(s; \boldsymbol{y}\right) = \int_{\vartheta=0}^{\infty} f(s;\vartheta) g\left(\boldsymbol{y}; \vartheta\right) \frac{1}{\vartheta^{\eta+1}} \, \mathrm{d}\vartheta \Big/ \int_{\alpha=0}^{\infty} g\left(\boldsymbol{y}; \alpha\right) \frac{1}{\alpha^{\eta+1}} \, \mathrm{d}\alpha \, .$$

But now, making the substitution $\beta = \theta / \alpha$ (Note: $d\beta = -d\alpha\theta/\alpha^2 = -d\alpha\beta^2/\theta$), we have

$$\begin{aligned}
\text{LHS} =& \theta^{\eta} \int_{\alpha=0}^{\infty} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \int_{s=0}^{\infty} C\left(x_{\mathrm{OS}}(\boldsymbol{y}), s\right) \hat{f}_{\mathrm{OS}}(s, \boldsymbol{y}) \mathrm{d}s g\left(\boldsymbol{y}; \alpha\right) \frac{1}{\alpha^{\eta+1}} J \mathrm{d}\boldsymbol{y} \, \mathrm{d}\alpha \\
=& - \theta^{\eta} \int_{\beta=\infty}^{0} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \int_{s=0}^{\infty} C\left(x_{\mathrm{OS}}(\boldsymbol{y}), s\right) \hat{f}_{\mathrm{OS}}(s, \boldsymbol{y}) \mathrm{d}s g\left(\boldsymbol{y}; \frac{\theta}{\beta}\right) \left(\frac{\beta}{\theta}\right)^{\eta+1} J \mathrm{d}\boldsymbol{y} \frac{\theta}{\beta^2} \, \mathrm{d}\beta \\
=& \int_{\beta=0}^{\infty} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \int_{s=0}^{\infty} C\left(x_{\mathrm{OS}}(\boldsymbol{y}), s\right) \hat{f}_{\mathrm{OS}}(s, \boldsymbol{y}) \mathrm{d}s g\left(\beta\boldsymbol{y}; \theta\right) \beta^{N} \beta^{\eta-1} J \mathrm{d}\boldsymbol{y} \, \mathrm{d}\beta \\
=& \int_{\beta=0}^{\infty} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \int_{s=0}^{\infty} C\left(\beta^{\gamma} x_{\mathrm{OS}}(\boldsymbol{y}), \beta s\right) \hat{f}_{\mathrm{OS}}(s, \boldsymbol{y}) \mathrm{d}s g\left(\beta\boldsymbol{y}; \theta\right) \beta^{N-1} J \mathrm{d}\boldsymbol{y} \, \mathrm{d}\beta \\
=& \int_{\beta=0}^{\infty} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \int_{s=0}^{\infty} C\left(\beta^{\gamma} x_{\mathrm{OS}}(\boldsymbol{y}), \beta s\right) \hat{f}_{\mathrm{OS}}(\beta s, \beta \boldsymbol{y}) \beta \mathrm{d}s g\left(\beta\boldsymbol{y}; \theta\right) \beta^{N-1} J \mathrm{d}\boldsymbol{y} \, \mathrm{d}\beta \qquad (\dagger) \\
=& \int_{\beta=0}^{\infty} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=1\}} \int_{s=0}^{\infty} C\left(x_{\mathrm{OS}}(\beta\boldsymbol{y}), s\right) \hat{f}_{\mathrm{OS}}(s, \beta \boldsymbol{y}) \mathrm{d}s g\left(\beta\boldsymbol{y}; \theta\right) \beta^{N-1} J \mathrm{d}\boldsymbol{y} \, \mathrm{d}\beta
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\beta=0}^{\infty} \int_{\{\boldsymbol{y}:\boldsymbol{y}^\top \mathbf{1}=\beta\}} \int_{s=0}^{\infty} C\left(x_{\mathrm{OS}}(\boldsymbol{y}), s\right) \hat{f}_{\mathrm{OS}}(s, \boldsymbol{y}) \mathrm{d}s g\left(\boldsymbol{y}; \theta\right) J \mathrm{d}\boldsymbol{y}\, \mathrm{d}\beta \\
&= \int_{\boldsymbol{y} \in \mathbb{R}_+^N} \underbrace{\left(\int_{s=0}^{\infty} C\left(x_{\mathrm{OS}}(\boldsymbol{y}), s\right) \hat{f}_{\mathrm{OS}}(s, \boldsymbol{y}) \mathrm{d}s \right)}_{(A)} g\left(\boldsymbol{y}; \theta\right) \mathrm{d}\boldsymbol{y},
\end{aligned}
$$

where in $(\dagger)$, we used the fact that $\beta \hat{f}_{\mathrm{OS}}(\beta s, \beta \boldsymbol{y}) = \hat{f}_{\mathrm{OS}}(s, \boldsymbol{y})$. Thus, it suffices to find $x_{\mathrm{OS}}(\cdot) \in \mathcal{X}_\gamma$ that minimizes (A). Due to homogeneity in C and $\hat{f}_{\mathrm{OS}}$, there must exist some the prescriptive solution $x_{\mathrm{OS}}(\cdot)$ minimizing (A) that is necessarily homogeneous. Moreover, uniform optimality holds as (A) is independent of θ. □

REMARK 4. Here, we present a slightly more elementary proof than what appears in Feng and Shanthikumar (2023), by avoiding the use of the prior and posterior interpretations of $\hat{f}_{\mathrm{OS}}$.

Proof of Theorem 2b. We first show that under the assumptions, $\frac{1}{a}\hat{\theta}_{\mathrm{MLE}}(a\boldsymbol{y}) = \hat{\theta}_{\mathrm{MLE}}(\boldsymbol{y})$. This holds because $g(a\boldsymbol{y}; \theta) = \frac{1}{a^N} g\left(\boldsymbol{y}; \frac{\theta}{a}\right)$. Thus, $\hat{\theta}_{\mathrm{MLE}}(a\boldsymbol{y}) = \arg\max_\theta g(a\boldsymbol{y}; \theta) = \arg\max_\theta g\left(\boldsymbol{y}; \frac{\theta}{a}\right) = a\hat{\theta}_{\mathrm{MLE}}(\boldsymbol{y})$.

Now, by definition, for any $\boldsymbol{y} \in \mathcal{Y}$,

$$
\begin{aligned}
x_{\mathrm{ETO}}(a\boldsymbol{y}) =& \arg\min_x \int_{z=0}^{\infty} C\left(x; z\right) f\left(z; \hat{\boldsymbol{\theta}}_{\mathrm{MLE}}(a\boldsymbol{y})\right)\, dz \\
=& \arg\min_x\, a^\eta \int_{z=0}^{\infty} \frac{1}{a^\eta} C\left(x; z\right) f\left(z; \hat{\theta}_{\mathrm{MLE}}(a\boldsymbol{y})\right)\, dz \\
=& \arg\min_x \int_{z=0}^{\infty} C\left(\frac{x}{a^\gamma}; \frac{z}{a}\right) f\left(z; \hat{\theta}_{\mathrm{MLE}}(a\boldsymbol{y})\right)\, dz \\
=& \arg\min_x\, a \int_{z=0}^{\infty} C\left(\frac{x}{a^\gamma}; z\right) f\left(az; \hat{\theta}_{\mathrm{MLE}}(a\boldsymbol{y})\right)\, dz \\
=& \arg\min_x \int_{z=0}^{\infty} C\left(\frac{x}{a^\gamma}; z\right) f\left(z; \frac{1}{a}\hat{\theta}_{\mathrm{MLE}}(a\boldsymbol{y})\right)\, dz \\
=& \arg\min_x \int_{z=0}^{\infty} C\left(\frac{x}{a^\gamma}; z\right) f\left(z; \hat{\theta}_{\mathrm{MLE}}(\boldsymbol{y})\right)\, dz \\
=& a^\gamma x_{\mathrm{ETO}}(\boldsymbol{y}).
\end{aligned}
$$

Thus, $x_{\mathrm{ETO}}(\boldsymbol{y}) \in \mathcal{X}_\gamma$. □

Proof of Theorem 2c. Suppose there exists a uniformly optimal prescriptive solution in the space of $\mathcal{X}_{\mathrm{aff}}$, denoted by $\boldsymbol{a}^{\star\top}\boldsymbol{y} + b^\star$. If $\boldsymbol{a}^\star = \mathbf{0}$, then Assumption 3 indicates that there are at least two $\theta_1 \neq \theta_2$ such that $\arg\min \phi(\cdot; \theta_1) \cap \arg\min \phi(\cdot; \theta_2) = \varnothing$. Thus, there cannot exist some $b^\star$ that is optimal for both θ_1 and θ_2.

If $\boldsymbol{a}^\star \neq \mathbf{0}$, then we first show that we must have $b^\star = 0$. Note that

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{y}}[\phi(\boldsymbol{a}^\top \boldsymbol{y} + b; \theta)] =& \int_{\boldsymbol{y} \in \mathbb{R}_+^N} \left(\int_{z=0}^{\infty} C\left(\boldsymbol{a}^\top \boldsymbol{y} + b, z\right) f(z; \theta) \mathrm{d}z \right) g\left(\boldsymbol{y}; \theta\right) \mathrm{d}\boldsymbol{y} \\
=& \alpha^\eta \int_{\boldsymbol{y} \in \mathbb{R}_+^N} \left(\int_{z=0}^{\infty} C\left(\frac{\boldsymbol{a}^\top \boldsymbol{y} + b}{\alpha}, \frac{z}{\alpha}\right) f(z; \theta) \mathrm{d}z \right) g\left(\boldsymbol{y}; \theta\right) \mathrm{d}\boldsymbol{y}
\end{aligned}
$$

$$
\begin{aligned}
&=\alpha^{\eta}\int_{\boldsymbol{y}\in\mathbb{R}_+^N}\left(\int_{z'=0}^{\infty}C\left(\frac{\boldsymbol{a}^\top\boldsymbol{y}+b}{\alpha},z'\right)f\left(z';\frac{\theta}{\alpha}\right)\mathrm{d}z'\right)g\left(\boldsymbol{y};\theta\right)\mathrm{d}\boldsymbol{y}\\
&=\alpha^{\eta-N+1}\int_{\boldsymbol{y}'\in\mathbb{R}_+^N}\left(\int_{z'=0}^{\infty}C\left(\boldsymbol{a}^\top\boldsymbol{y}'+\frac{b}{\alpha},z'\right)f\left(z';\frac{\theta}{\alpha}\right)\mathrm{d}z'\right)g\left(\boldsymbol{y}';\frac{\theta}{\alpha}\right)\mathrm{d}\boldsymbol{y}'\\
&=\alpha^{\eta-N+1}\mathbb{E}_{\boldsymbol{y}}\left[\phi\left(\boldsymbol{a}^\top\boldsymbol{y}+\frac{b}{\alpha};\frac{\theta}{\alpha}\right)\right],
\end{aligned}
$$

where the second equality is due to Assumption i, the third equality is using the substitution $z'=z/\alpha$ and Assumption ii, and the fourth equality is because of the substitution $\boldsymbol{y}'=\boldsymbol{y}/\alpha$ and $g(a\boldsymbol{y}';\theta)=\frac{1}{a^N}g\left(\boldsymbol{y}';\frac{\theta}{a}\right)$.

Suppose there exists a uniformly optimal prescriptive solution in the space of $\mathcal{X}_{\text{aff}}$, denoted by $\boldsymbol{a}^{\star\top}\boldsymbol{y}+b^\star$. Then we must have

$$
\arg\min_{\boldsymbol{a},b}\mathbb{E}_{\boldsymbol{y}}[\phi(\boldsymbol{a}^\top\boldsymbol{y}+b;\theta)]=\arg\min_{\boldsymbol{a},b}\mathbb{E}_{\boldsymbol{y}}\left[\phi\left(\boldsymbol{a}^\top\boldsymbol{y}+\frac{b}{\alpha};\frac{\theta}{\alpha}\right)\right]=\arg\min_{\boldsymbol{a},b}\mathbb{E}_{\boldsymbol{y}}\left[\phi\left(\boldsymbol{a}^\top\boldsymbol{y}+\frac{b}{\alpha};\theta\right)\right]
$$

for all θ, α, which implies that $b^\star=0$ and $\boldsymbol{a}^\star=\arg\min_{\boldsymbol{a}}\mathbb{E}_{\boldsymbol{y}}[\phi(\boldsymbol{a}^\top\boldsymbol{y};\theta)]$ is the OS solution.

To complete the proof, take any θ, and let $\beta=\arg\min_x\phi(x;\theta)$. By Assumption 2, this is a unique optima for the given θ. Let $\bar{x}(\boldsymbol{y})\equiv\beta$ for all $\boldsymbol{y}\in\mathcal{Y}$. Thus, for any other $\boldsymbol{x}(\cdot)\in\mathcal{X}_{\text{aff}}$ that is not a constant function, we have that

$$
\Phi[x(\cdot)|\theta]=\int_{\mathcal{Y}}\phi(x(\boldsymbol{y}),\theta)g(\boldsymbol{y},\theta)d\boldsymbol{y}>\int_{\mathcal{Y}}\phi(\beta,\theta)g(\boldsymbol{y},\theta)d\boldsymbol{y}=\Phi[\bar{x}(\cdot)|\theta],
$$

where the inequality holds because of the non-negativity of g and uniqueness of β. Since $\bar{x}(\cdot)\in\mathcal{X}_{\text{aff}}$, we have thus found some θ for which no non-constant policy can be uniformly optimal for $\mathcal{X}_{\text{aff}}$. However, we had also earlier shown that no constant policy can be uniformly optimal for $\mathcal{X}_{\text{aff}}$. □

A.5. Theorem 3

Proof of Theorem 3. Let $\boldsymbol{x}(\cdot)$ be prior optimal for u. Suppose it is not finite-sample Pareto optimal. Then $\Phi[\boldsymbol{x}(\cdot)|\boldsymbol{\theta}]>\Phi[\boldsymbol{x}'(\cdot)|\boldsymbol{\theta}]$ for all $\boldsymbol{\theta}\in\bar{\Theta}$ except on a set of measure greater than 0, implies that $\mathbb{E}_{u\sim\boldsymbol{\theta}}[\Phi[\boldsymbol{x}(\cdot)|\boldsymbol{\theta}]]>\mathbb{E}_{u\sim\boldsymbol{\theta}}[\Phi[\boldsymbol{x}'(\cdot)|\boldsymbol{\theta}]]$, contradicting the optimality of $\boldsymbol{x}(\cdot)$. □

A.6. Proposition 6

Note that

$$
\begin{aligned}
&\arg\min\int\int\phi(\boldsymbol{x}(\boldsymbol{y});\boldsymbol{\theta})g(\boldsymbol{y};\boldsymbol{\theta})u(\boldsymbol{\theta})d\boldsymbol{y}d\boldsymbol{\theta}=\arg\min\int\int\frac{\phi(\boldsymbol{x}(\boldsymbol{y});\boldsymbol{\theta})g(\boldsymbol{y};\boldsymbol{\theta})}{\varphi(\boldsymbol{\theta})}d\boldsymbol{y}d\boldsymbol{\theta}\\
&=\arg\min\int\left[\frac{\int\left(\phi(\boldsymbol{x}(\boldsymbol{y});\boldsymbol{\theta})-\varphi(\boldsymbol{\theta})\right)g(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y}}{\varphi(\boldsymbol{\theta})}+1\right]d\boldsymbol{\theta}\\
&=\arg\min\int\int\underbrace{\frac{\phi(\boldsymbol{x}(\boldsymbol{y});\boldsymbol{\theta})-\varphi(\boldsymbol{\theta})}{\varphi(\boldsymbol{\theta})}}_{:=R(\boldsymbol{x}(\boldsymbol{y});\boldsymbol{\theta})}g(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y}d\boldsymbol{\theta}=\arg\min\int\int R(\boldsymbol{x}(\boldsymbol{y});\boldsymbol{\theta})g(\boldsymbol{y};\boldsymbol{\theta})\underbrace{\frac{1}{|\bar{\Theta}|}}_{:=\tilde{u}(\boldsymbol{\theta})}d\boldsymbol{y}d\boldsymbol{\theta}.
\end{aligned}
$$

□

A.7. Theorem 5

As the proof is already provided in the main text, here, we put down two remarks about the proof.

REMARK 5. a) In the proof, if $\boldsymbol{x}(\cdot)$ is already sufficient-representable, then $\boldsymbol{x}(\cdot)$ has no component in $\boldsymbol{y}|\hat{\boldsymbol{\theta}}$, making ϕ constant in the argument of $\boldsymbol{y}|\hat{\boldsymbol{\theta}}$. This meets the equality conditions for Jensen's inequality, thus the construction $\tilde{\boldsymbol{x}}(\cdot)$ does not lead to a strictly better solution.

b) If $\mathcal{H} = \mathcal{M}(\mathbb{R}^N, \mathbb{R}^q)$, that is, that there are no distributional assumptions, then the order statistic is sufficient. However, it has dimensions N, so no reductions in dimensional complexity in $\boldsymbol{x}(\cdot)$ is obtained, though structurally we obtain an optimal prescriptive solution that is a function of the order statistic.

A.8. Theorem 6

Proof of Theorem 6. Suppose $\boldsymbol{x}_{\text{OVE}}$ is not optimal for (R-OVE). Hence, by Theorem 4, there exists some other prescriptive solution $\boldsymbol{X}$ such that

$$\int_{\hat{\Theta}} K(\hat{\boldsymbol{\theta}})\, \beth(\boldsymbol{X}(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}})\, d\hat{\boldsymbol{\theta}} - \int_{\hat{\Theta}} K(\hat{\boldsymbol{\theta}})\, \beth(\boldsymbol{x}_{\text{OVE}}(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}})\, d\hat{\boldsymbol{\theta}} < 0. \tag{25}$$

Note that by Theorem 5, it suffices to consider only $\boldsymbol{X} \in \mathcal{X}(\hat{\boldsymbol{\theta}})$. However, by change of variables from $(\hat{\boldsymbol{\theta}}, \boldsymbol{y}|\hat{\boldsymbol{\theta}})$ back to $\boldsymbol{y}$, we obtain that the LHS of the above is

$$\int_{\mathcal{Y}} \left[\beth(\boldsymbol{X}(\hat{\boldsymbol{\theta}}(\boldsymbol{y})), \hat{\boldsymbol{\theta}}(\boldsymbol{y})) - \beth(\boldsymbol{x}_{\text{OVE}}(\hat{\boldsymbol{\theta}}(\boldsymbol{y})), \hat{\boldsymbol{\theta}}(\boldsymbol{y}))\right] g_0(\boldsymbol{y}) d\boldsymbol{y}.$$

This, however is non-negative since $\boldsymbol{x}_{\text{OVE}}$ minimizes $\beth$ for all $\hat{\boldsymbol{\theta}} \in \hat{\Theta}$ almost surely in the class $\mathcal{X}(\hat{\boldsymbol{\theta}})$, and g_0 is non-negative. This contradicts (25). □

A.9. Corollary 2

Proof of Corollary 2. (i) We argue that $\beth$ obeys Assumption 2, thus the existence of a unique minimizer follows identically from Lemma 1. Strict convexity of $\beth(\cdot, \hat{\boldsymbol{\theta}})$ follows analogously to Proposition 5, whereas since ϕ is continuously differentiable, so is $\beth(\cdot, \hat{\boldsymbol{\theta}})$ for every $\hat{\boldsymbol{\theta}} \in \hat{\Theta}$. Similarly, $\beth(\cdot, \hat{\boldsymbol{\theta}})$ is bounded below: Denote $\nu = \int g_1(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) u(\boldsymbol{\theta}) d\boldsymbol{\theta}$, then $\beth(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) \geq \inf_{\boldsymbol{\theta} \in \Theta} \phi(\boldsymbol{x}; \boldsymbol{\theta}) \int g_1(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) u(\boldsymbol{\theta}) d\boldsymbol{\theta} \geq L\nu$ for all $\hat{\boldsymbol{\theta}} \in \hat{\Theta}$, since $\phi(\boldsymbol{x}; \boldsymbol{\theta}) \geq L$.

If $\mathcal{B}$ is bounded, then we are done. If not, we show that for any given $\hat{\boldsymbol{\theta}} \in \hat{\Theta}$, $\beth(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$. We assumed $\overline{\text{supp}(u)}$ is compact, thus it is closed and bounded. The conclusion follows as $\beth(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) = \int_{\overline{\text{supp}(u)}} \phi(\boldsymbol{x}; \boldsymbol{\theta}) g_1(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) u(\boldsymbol{\theta}) d\boldsymbol{\theta} \geq \inf_{\boldsymbol{\theta} \in \overline{\text{supp}(u)}} \phi(\boldsymbol{x}; \boldsymbol{\theta}) \int_{\overline{\text{supp}(u)}} g_1(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) u(\boldsymbol{\theta}) d\boldsymbol{\theta} \geq \nu Q(\boldsymbol{x}) \to \infty$ as $Q(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$.

(ii) Due to Theorem 6, if there exists a unique minimizer to $\beth$, then a prior optimal solution always exists. In the proof of Theorem 5, (11) holds with strict inequality for any function $\boldsymbol{x}(\cdot)$ that is not sufficient-representable, by Remark 5a). Thus, the prior optimal solution must be sufficient-representable. Moreover, under strict convexity, the proof of Corollary 2(i) ensures only one prescriptive solution that minimizes $\beth$. Thus, there is a unique sufficient-representable prescriptive solution that is prior optimal.

□

A.10. Further results on the relationship between prior and uniform optimality

Prior optimality is a weaker notion than uniform optimality, and conversely, the latter is stricter:

THEOREM 9 **(Conservativeness of uniform optimality)**. *Within class of functions $\bar{\mathcal{X}}$, $\boldsymbol{x}(\cdot)$ is uniformly optimal if and only if it is prior optimal for all priors $u \in \mathcal{M}(\Theta, \mathbb{R})$.*

Proof of Theorem 9. ($\Leftarrow$): Suppose that $\boldsymbol{x}$ is prior optimal for all priors on Θ. In particular, $\boldsymbol{x}$ would be prior optimal for $u \sim \delta(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}})$, the deterministic distribution that is equal to $\tilde{\boldsymbol{\theta}}$ w.p. 1 represented by the Dirac-delta distribution, for all $\tilde{\boldsymbol{\theta}} \in \Theta$. This, however, is just the definition of uniform optimality.

($\Rightarrow$): Suppose that $\boldsymbol{x}$ is such that for all other $\check{\boldsymbol{x}} \in \bar{\mathcal{X}}$, $\Phi[\boldsymbol{x}(\cdot)|\boldsymbol{\theta}] \leq \Phi[\check{\boldsymbol{x}}(\cdot)|\boldsymbol{\theta}]$ for all $\boldsymbol{\theta} \in \Theta$, then certainly, for any prior represented by u defined on Θ, we have that, since u is a density thus non-negative,

$$\int_{\Theta} \Phi[\boldsymbol{x}(\cdot)|\boldsymbol{\theta}] u(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq \int_{\Theta} \Phi[\check{\boldsymbol{x}}(\cdot)|\boldsymbol{\theta}] u(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

which is just the definition of prior optimality for prior u. □

COROLLARY 7. *Let u be a given prior with corresponding prior optimal solution $\boldsymbol{x}^*(\cdot)$. Suppose there exists some uniformly optimal prescriptive solution $\boldsymbol{x}^{uni}(\cdot)$ within a restricted class of functions $\bar{\mathcal{X}} \subsetneq \mathcal{X}$. Then $\boldsymbol{x}^*(\cdot)$ outperforms $\boldsymbol{x}^{uni}(\cdot)$:*

$$\Psi\left[\boldsymbol{x}^*(\cdot); u\right] \leq \Psi\left[\boldsymbol{x}^{uni}(\cdot); u\right], \tag{26}$$

with equality if and only if there exists a prior optimal solution in the restricted class $\boldsymbol{x}^* \in \bar{\mathcal{X}}$.

Proof of Corollary 7. Indeed,

$$\boldsymbol{x}^*(\cdot) \in \arg\min_{\boldsymbol{x}(\cdot) \in \mathcal{X}} \mathbb{E}_u\left[\Phi[\boldsymbol{x}(\cdot)|\boldsymbol{\theta}]\right],$$
$$\boldsymbol{x}^u(\cdot) \in \arg\min_{\boldsymbol{x}(\cdot) \in \bar{\mathcal{X}}} \mathbb{E}_u\left[\Phi[\boldsymbol{x}(\cdot)|\boldsymbol{\theta}]\right],$$

where the latter holds by Theorem 9. But $\bar{\mathcal{X}} \subsetneq \mathcal{X}$. □

REMARK 6. Notice that prior optimal solutions are unlikely the same across priors u, however, the uniformly optimal solution needs to remain fixed across all priors by Theorem 9.

A.11. Proposition 4

Proof of Proposition 4. Notice that

$$\begin{aligned}
\hat{f}_{\text{OS}}\left(s; \boldsymbol{y}\right) &= \int_{\vartheta=0}^{\infty} f(s;\vartheta) g_0\left(\boldsymbol{y}\right) g_1\left(\hat{\theta}(\boldsymbol{y}); \vartheta\right) \frac{1}{\vartheta^{\eta+1}} \,\mathrm{d}\vartheta \Big/ \int_{\alpha=0}^{\infty} g_0\left(\boldsymbol{y}\right) g_1\left(\hat{\theta}(\boldsymbol{y}); \alpha\right) \frac{1}{\alpha^{\eta+1}} \,\mathrm{d}\alpha \\
&= \int_{\vartheta=0}^{\infty} f(s;\vartheta) g_1\left(\hat{\theta}(\boldsymbol{y}); \vartheta\right) \frac{1}{\vartheta^{\eta+1}} \,\mathrm{d}\vartheta \Big/ \int_{\alpha=0}^{\infty} g_1\left(\hat{\theta}(\boldsymbol{y}); \alpha\right) \frac{1}{\alpha^{\eta+1}} \,\mathrm{d}\alpha\,,
\end{aligned}$$

since $g_0(\boldsymbol{y})$ is irrelevant to both integrals in the numerator and denominator. Thus, for each $\boldsymbol{y} \in \mathcal{Y}$,

$$\begin{aligned} x_{\mathrm{OS}}(\boldsymbol{y}) &= \arg\min_x \int_{s=0}^{\infty} C(x,s)\, \hat{f}_{\mathrm{OS}}(s;\boldsymbol{y}) \mathrm{d}s \\ &= \arg\min_x \int_{s=0}^{\infty} C(x,s) \int_{\vartheta=0}^{\infty} f(s;\vartheta) g_1\left(\hat{\theta}(\boldsymbol{y});\vartheta\right) \frac{1}{\vartheta^{\eta+1}}\, \mathrm{d}\vartheta \mathrm{d}s \\ &= \arg\min_x \int_{\vartheta=0}^{\infty} \left(\int_{s=0}^{\infty} C(x,s)\, f(s;\vartheta) \mathrm{d}s \right) g_1\left(\hat{\theta}(\boldsymbol{y});\vartheta\right) \frac{1}{\vartheta^{\eta+1}}\, \mathrm{d}\vartheta \\ &= \arg\min_x \int_{\vartheta=0}^{\infty} \phi(x,\vartheta)\, g_1\left(\hat{\theta}(\boldsymbol{y});\vartheta\right) \frac{1}{\vartheta^{\eta+1}}\, \mathrm{d}\vartheta = \arg\min_x \beth\left(x, \hat{\theta}(\boldsymbol{y})\right), \end{aligned}$$

where the denominator $\int_{\alpha=0}^{\infty} g_1\left(\hat{\theta}(\boldsymbol{y});\alpha\right) \frac{1}{\alpha^{\eta+1}}\, \mathrm{d}\alpha$ was dropped as it is irrelevant in the minimization of x. Thus, $x_{\mathrm{OS}}(\cdot) \in \mathcal{X}(\hat{\theta})$ and it is an OVE solution under the improper density $\frac{1}{\vartheta^{\eta+1}}$. □

We can actually verify this result in simulation here. Figure 9 below is taken from the simulations of the newsvendor problem when the prior is chosen as $u \propto 1/\theta^2$ (Note: $\eta = 1$ in the newsvendor problem). We can see just how close the regrets and the structure of the optimal solution are between OVE and OS, even though there is no particular restriction on OVE to belong to the class of linear solutions.

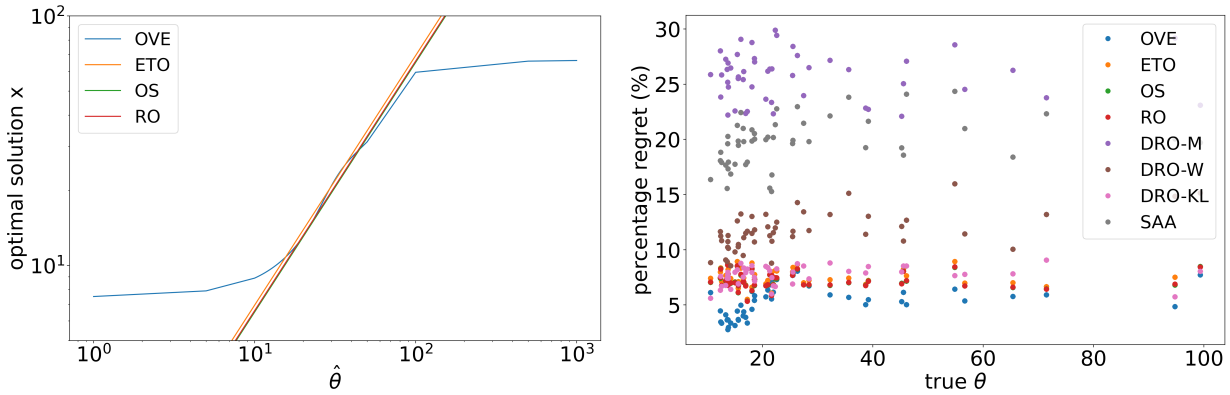


Figure 9 Log-log plot of the solution structure $x(\hat{\theta})$ and ex-ante percentage regret of different models in the prior $u \propto 1/\theta^2$

A.12. Theorem 7

For convenience, we will suppress the notation of $\boldsymbol{\theta}_0$ moving forward. Let $\boldsymbol{x}^*(\boldsymbol{\theta})$ denote the oracle solution for $\boldsymbol{\theta}$. Notice that $\Theta_0 = \{\boldsymbol{\theta}_0\}$ and so $\boldsymbol{x}(\boldsymbol{y}; t=0) \equiv \boldsymbol{x}^*$ for all data sets $\boldsymbol{y}$, and the oracle is constant in $\boldsymbol{y}$. On the other hand, $\Theta_1 = \bar{\Theta}$. Moreover, $\Theta_t \subset \Theta_{t'}$ if $t \le t'$.

Additionally, due to Theorem 5, $\boldsymbol{x}(\boldsymbol{y}; t) = \boldsymbol{x}(\hat{\boldsymbol{\theta}}; t)$ is sufficient-representable. By Remark 8, $\boldsymbol{x}(\hat{\boldsymbol{\theta}}; t)$ is unique for all $\hat{\boldsymbol{\theta}} \in \Theta$ and $t \in [0,1]$. $R(\boldsymbol{x}; \boldsymbol{\theta})$ being linear in $\phi(\boldsymbol{x}; \boldsymbol{\theta})$ retains its regularity, hence it is for all $\boldsymbol{\theta} \in \Theta$, convex in $\boldsymbol{x}$, twice continuously differentiable in $\boldsymbol{x}$ and has Hessian with eigenvalues bounded away from zero (Assumption 6).

Define

$$\tilde{\boldsymbol{F}}(\boldsymbol{x},t,\hat{\boldsymbol{\theta}}) := \mathbb{E}_{U[\Theta_t]}\left[\boldsymbol{\nabla}_{\boldsymbol{x}} R(\boldsymbol{x};\boldsymbol{\theta}) g_1(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})\right] = \frac{1}{|\Theta_t|}\int_{\Theta_t} \boldsymbol{\nabla}_{\boldsymbol{x}} R(\boldsymbol{x};\boldsymbol{\theta}) g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{27}$$

Note that this is well-defined as $\boldsymbol{\nabla}_{\boldsymbol{x}} R(\boldsymbol{x};\boldsymbol{\theta})$ exists by Assumption 6. Then for a given fixed $\hat{\boldsymbol{\theta}}$, $\left(\boldsymbol{x}(\hat{\boldsymbol{\theta}};t),t\right)$ is a path on which $\tilde{\boldsymbol{F}}(\boldsymbol{x},t,\hat{\boldsymbol{\theta}}) = 0$, by definition of the optimality of OVE.

To prove (18), we first prove a series of Lemmas:

LEMMA 2. *Under Assumptions 6 and 8, the prescriptive solution $x(\cdot;t)$ is a bounded function for all $t \in (0,1]$, that is the set $\mathcal{X}_t^* := \{\boldsymbol{x}(\hat{\boldsymbol{\theta}};t) : \hat{\boldsymbol{\theta}} \in \Theta\}$ is bounded. Moreover, there exists some constant $\bar{M} < +\infty$ such that for all $t \in (0,1]$ and all $\boldsymbol{\theta}_0 \in \bar{\Theta}$,*

$$\sup_{\hat{\boldsymbol{\theta}}} R\left(\boldsymbol{x}(\hat{\boldsymbol{\theta}};t);\boldsymbol{\theta}_0\right) \le \bar{M}.$$

Proof of Lemma 2. Assumption 5c) translates to R: There is also some function $\bar{Q}(\boldsymbol{x})$ such that $\bar{Q}(\boldsymbol{x}) \le \inf_{\boldsymbol{\theta}\in\Theta_t} R(\boldsymbol{x};\boldsymbol{\theta})$ for all $\boldsymbol{x} \in \mathcal{B}$ and $\boldsymbol{\theta} \in \Theta_t$ and $\bar{Q}(\boldsymbol{x}) \to +\infty$ as $\|\boldsymbol{x}\| \to \infty$, because $\varphi(\boldsymbol{\theta})$ is strictly positive for all $\boldsymbol{\theta} \in \bar{\Theta}$ and Θ_t is bounded and closed. Hence, if the set $\mathcal{X}_t^*$ is not bounded, then there is some sequence of $\hat{\boldsymbol{\theta}} \in \Theta$ such that $\bar{Q}(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t)) \to +\infty$. However,

$$\begin{aligned}
&\bar{Q}(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t))\mathbb{E}_{U(\Theta_t)}\left[g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\right] \\
&\le \inf_{\boldsymbol{\theta}\in\Theta_t} R(x(\hat{\boldsymbol{\theta}},t);\boldsymbol{\theta})\, \mathbb{E}_{U(\Theta_t)}\left[g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\right] = \mathbb{E}_{U(\Theta_t)}\left[\left(\inf_{\boldsymbol{\theta}\in\Theta_t} R(x(\hat{\boldsymbol{\theta}},t);\boldsymbol{\theta})\right) g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\right] \\
&\le \mathbb{E}_{U(\Theta_t)}\left[R(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t);\boldsymbol{\theta}) g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\right] \qquad (28) \\
&\le \mathbb{E}_{U(\Theta_t)}\left[R(\boldsymbol{x}^*(\boldsymbol{\theta}_0);\boldsymbol{\theta}) g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\right] \le \left(\max_{\boldsymbol{\theta}\in\Theta_t} R(\boldsymbol{x}^*(\boldsymbol{\theta}_0);\boldsymbol{\theta})\right) \mathbb{E}_{U(\Theta_t)}\left[g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\right], \qquad (29)
\end{aligned}$$

where (28) holds due to the non-negativity of g_1 and (29) holds by the optimality of $\boldsymbol{x}(\hat{\boldsymbol{\theta}},t)$ for the objective in (28). Furthermore, as Θ_t is closed, convex and compact for all $t > 0$, the maximum is attained in (29). But $\mathbb{E}_{U(\Theta_t)}\left[g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\right]$ is finite for all t. Hence,

$$\bar{Q}(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t)) \le \max_{\boldsymbol{\theta}\in\Theta_t} \phi(\boldsymbol{x}^*(\boldsymbol{\theta}_0),\boldsymbol{\theta}) < +\infty.$$

Thus $\mathcal{X}_t^*$ cannot be unbounded. □

REMARK 7. This Lemma guarantees that $\boldsymbol{x}(\hat{\boldsymbol{\theta}};t)$ remains bounded for all $\hat{\boldsymbol{\theta}} \in \Theta$. In general, solutions obtained under methods are not necessarily bounded, as illustrated in Figure 3. The key reason why this occurs is due to the averaging over θ in OVE. g_1 is vanishing as $\|\hat{\boldsymbol{\theta}}\| \to \infty$, thus the rarer $\hat{\boldsymbol{\theta}}$, the less $\sqsupset$ changes and correspondingly the less $\boldsymbol{x}(\hat{\boldsymbol{\theta}};t)$ changes. On the other hand, other methods are centered on $\hat{\boldsymbol{\theta}}$ and thus will grow with $\hat{\boldsymbol{\theta}}$.

LEMMA 3. *Let $\mathcal{H}_{\boldsymbol{xx}}\cdot$ denotes the Hessian matrix in $\boldsymbol{x}$ and*

$$\tilde{\boldsymbol{H}}(\boldsymbol{x},t,\hat{\boldsymbol{\theta}}) := \nabla_{\boldsymbol{x}}\tilde{\boldsymbol{F}}(\boldsymbol{x},t,\hat{\boldsymbol{\theta}}) = \mathbb{E}_{U[\Theta_t]}\left[\mathcal{H}_{\boldsymbol{xx}}R(\boldsymbol{x};\boldsymbol{\theta})g_1(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})\right].$$

Then under Assumptions 6 and 8, for all $t\in(0,1]$ and $\hat{\boldsymbol{\theta}}\in\Theta$,

i. $\boldsymbol{x}(\hat{\boldsymbol{\theta}};t)$ is differentiable in t, with derivative

$$\frac{\partial \boldsymbol{x}}{\partial t}(\hat{\boldsymbol{\theta}},t) = -\tilde{\boldsymbol{H}}(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t),t,\hat{\boldsymbol{\theta}})^{-1}\frac{\partial \tilde{\boldsymbol{F}}}{\partial t}(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t),t,\hat{\boldsymbol{\theta}}); \tag{30}$$

ii. $\boldsymbol{x}(\hat{\boldsymbol{\theta}};t)$ is differentiable in $\hat{\boldsymbol{\theta}}$, with derivative

$$\boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}}\boldsymbol{x}(\hat{\boldsymbol{\theta}},t) = -\tilde{\boldsymbol{H}}(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t),t,\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}}\tilde{\boldsymbol{F}}(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t),t,\hat{\boldsymbol{\theta}}). \tag{31}$$

REMARK 8. By Assumption 5′, there exists a unique solution $\boldsymbol{x}(\hat{\boldsymbol{\theta}};t)$ for all $\hat{\boldsymbol{\theta}}\in\Theta$ and for all $t\in[0,1]$. As such, we state this Lemma directly as derivatives, rather than as sub-gradients.

Proof of Lemma 3. We show the proof for part i., as part ii. is analogous. Note that for $\tilde{\boldsymbol{F}}$ as defined in (27), we can make a change of variables from $\boldsymbol{\theta}:\boldsymbol{\theta}\in\Theta_t\mapsto\boldsymbol{\theta}_0+t(\boldsymbol{\theta}-\boldsymbol{\theta}_0):\boldsymbol{\theta}\in\bar{\Theta}$, we have

$$\tilde{\boldsymbol{F}}(\boldsymbol{x},t,\hat{\boldsymbol{\theta}}) = \frac{1}{|\bar{\Theta}|}\int_{\bar{\Theta}}\nabla_{\boldsymbol{x}}\phi\big(\boldsymbol{x};\boldsymbol{\theta}_0+t(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\big)\,g_1\big(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0+t(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\big)\,d\boldsymbol{\theta}.$$

Hence, by Assumptions 6 and 8, $\tilde{\boldsymbol{F}}$ is continuously differentiable in $(\boldsymbol{x},t)$. As such,

$$\Delta\tilde{\boldsymbol{F}} = D_{\boldsymbol{x}}\tilde{\boldsymbol{F}}\,\Delta\boldsymbol{x} + \partial_t\tilde{\boldsymbol{F}}\,\Delta t + o\big((\|\Delta\boldsymbol{x}\|^2+|\Delta t|^2)^{1/2}\big),$$

where $D_{\boldsymbol{x}}\tilde{\boldsymbol{F}}$ denotes the Jacobian matrix of $\tilde{\boldsymbol{F}}$ with respect to $\boldsymbol{x}$. By the optimality conditions for $\boldsymbol{x}$, on the path $(\boldsymbol{x}(\hat{\boldsymbol{\theta}};t),t)$, we have $\tilde{\boldsymbol{F}}(\boldsymbol{x},t,\hat{\boldsymbol{\theta}})=\boldsymbol{0}$, that is $\Delta\tilde{\boldsymbol{F}}=\boldsymbol{0}$. Thus,

$$\boldsymbol{0} = D_{\boldsymbol{x}}\tilde{\boldsymbol{F}}\,\Delta\boldsymbol{x} + \partial_t\tilde{\boldsymbol{F}}\,\Delta t + o\big((\|\Delta\boldsymbol{x}\|^2+|\Delta t|^2)^{1/2}\big), \tag{32}$$

where

$$D_{\boldsymbol{x}}\tilde{\boldsymbol{F}}(\boldsymbol{x},t,\hat{\boldsymbol{\theta}}) = \tilde{\boldsymbol{H}}(\boldsymbol{x},t,\hat{\boldsymbol{\theta}}) := \frac{1}{|\bar{\Theta}|}\int_{\bar{\Theta}}\mathcal{H}_{\boldsymbol{xx}}\phi\big(\boldsymbol{x};\boldsymbol{\theta}_0+t(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\big)\,g_1\big(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0+t(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\big)\,d\boldsymbol{\theta}.$$

Since $\nabla^2_{\boldsymbol{xx}}\phi(\boldsymbol{x};\boldsymbol{\theta})\succ 0$ and $g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\geq 0$ and positive on a set of measure greater than zero, we have $D_{\boldsymbol{x}}\tilde{\boldsymbol{F}}$ is positive definite, and therefore invertible. Then by (32)

$$\lim_{\Delta t\to 0}\frac{\Delta\boldsymbol{x}}{\Delta t} = -\big(D_{\boldsymbol{x}}\tilde{\boldsymbol{F}}(\boldsymbol{x},t,\hat{\boldsymbol{\theta}})\big)^{-1}\partial_t\tilde{\boldsymbol{F}}(\boldsymbol{x},t,\hat{\boldsymbol{\theta}}),$$

the limit exists and $\boldsymbol{x}^*(\hat{\boldsymbol{\theta}},t)$ is differentiable in t. In particular, the derivative is given by (30). □

LEMMA 4. *There exists some radius W that is not a function of sample size N and data set $\boldsymbol{y}$; and there exists some constant $\bar{L}$ that is not a function of N, $\boldsymbol{y}$ or radius t such that*

$$||\boldsymbol{x}(\boldsymbol{y};t)-\boldsymbol{x}^*|| \le \bar{L}t, \qquad \forall t\in(0,1], \quad \forall \boldsymbol{y}\in B_W^{\hat{\theta}} := \left\{\boldsymbol{y}\,\middle|\, ||\hat{\boldsymbol{\theta}}(\boldsymbol{y})-\boldsymbol{\theta}_0|| \le W\right\}. \tag{33}$$

Proof of Lemma 4. To show result (33), we need to bound $\|\frac{\partial \boldsymbol{x}}{\partial t}(\hat{\boldsymbol{\theta}},t)\|$, whose expression is given by (30). Assumption 6 guarantees that $\mathcal{H}_{\boldsymbol{xx}}R(\boldsymbol{x};\boldsymbol{\theta})$ has eigenvalues bounded away from zero. Hence, we can bound the contributions from $\tilde{\boldsymbol{H}}(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t),t,\hat{\boldsymbol{\theta}})^{-1}$ by some fixed constant, $\bar{H}$ and it suffices to bound $\frac{\partial \tilde{\boldsymbol{F}}}{\partial t}(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t),t,\hat{\boldsymbol{\theta}})$.

Using the substitution $\boldsymbol{\theta}' = \boldsymbol{\theta}_0 + \frac{1}{t}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$, we can see that

$$\begin{aligned}
\tilde{\boldsymbol{F}}(\boldsymbol{x},t,\hat{\boldsymbol{\theta}}) &= \frac{1}{|\Theta_t|}\int_{\Theta_t} \boldsymbol{\nabla}_{\boldsymbol{x}} R(\boldsymbol{x};\boldsymbol{\theta}) g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \frac{1}{t^d|\bar{\Theta}|}\int_{\Theta} \boldsymbol{\nabla}_{\boldsymbol{x}} R(\boldsymbol{x};\boldsymbol{\theta}_0+t(\boldsymbol{\theta}'-\boldsymbol{\theta}_0)) g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0+t(\boldsymbol{\theta}'-\boldsymbol{\theta}_0)) t^d d\boldsymbol{\theta}' \\
&= \mathbb{E}_{U[\bar{\Theta}]}\left[\boldsymbol{\nabla}_{\boldsymbol{x}} R(\boldsymbol{x};\boldsymbol{\theta}_0+t(\boldsymbol{\theta}'-\boldsymbol{\theta}_0)) g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0+t(\boldsymbol{\theta}'-\boldsymbol{\theta}_0))\right].
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{\partial \tilde{\boldsymbol{F}}}{\partial t}(\boldsymbol{x},t,\hat{\boldsymbol{\theta}}) = \mathbb{E}_{U[\bar{\Theta}]}&\left[(\boldsymbol{\theta}'-\boldsymbol{\theta}_0)^\top \boldsymbol{\nabla}_{\boldsymbol{x}\boldsymbol{\theta}} R(\boldsymbol{x};\boldsymbol{\theta}_0+t(\boldsymbol{\theta}'-\boldsymbol{\theta}_0)) g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0+t(\boldsymbol{\theta}'-\boldsymbol{\theta}_0))\right] \\
&+ \mathbb{E}_{U[\bar{\Theta}]}\left[\boldsymbol{\nabla}_{\boldsymbol{x}} R(\boldsymbol{x};\boldsymbol{\theta}_0+t(\boldsymbol{\theta}'-\boldsymbol{\theta}_0))(\boldsymbol{\theta}'-\boldsymbol{\theta}_0)^\top \nabla_{\boldsymbol{\theta}} g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0+t(\boldsymbol{\theta}'-\boldsymbol{\theta}_0))\right].
\end{aligned}$$

Let δ be as defined in Assumption 5′. By Lemma 2, $\mathcal{X}_t^*$ is bounded for all $t\in(0,1]$. Hence, since $\boldsymbol{x}(\cdot;t)$ is a continuous function in $\hat{\boldsymbol{\theta}}$ (by Lemma 3, it is differentiable, thus continuous) and both $\boldsymbol{\nabla}_{\boldsymbol{x}} R(\cdot;\boldsymbol{\theta})$ and $\boldsymbol{\nabla}_{\boldsymbol{x}\boldsymbol{\theta}} R(\cdot;\boldsymbol{\theta})$ are assumed to be continuous by Assumption 6, we get that $\|\boldsymbol{\nabla}_{\boldsymbol{x}} R(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t);\boldsymbol{\theta})\|$ and $\|\boldsymbol{\nabla}_{\boldsymbol{x}\boldsymbol{\theta}} R(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t);\boldsymbol{\theta})\|$ are bounded over $\hat{\boldsymbol{\theta}}\in\Theta$, $\boldsymbol{\theta}\in\bar{\Theta}$ and $t\in[0,1]$. Correspondingly, let $R_1 := \sup_{\hat{\boldsymbol{\theta}}\in\Theta,\boldsymbol{\theta}\in\bar{\Theta}} \|\boldsymbol{\nabla}_{\boldsymbol{x}} R(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t);\boldsymbol{\theta}))\|$, and $R_2 := \sup_{\hat{\boldsymbol{\theta}}\in\Theta,\boldsymbol{\theta}\in\bar{\Theta}} \|\nabla_{\boldsymbol{x}\boldsymbol{\theta}} R(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t);\boldsymbol{\theta}))\|$. Moreover, by Assumption 8, both g_1 and $\boldsymbol{\nabla}_{\boldsymbol{\theta}} g_1$ are continuous and shrink as $\|\hat{\boldsymbol{\theta}}\|\to\infty$. Hence, since $\bar{\Theta}$ is closed, convex and compact,

$$\underline{G} := \sup_{\hat{\boldsymbol{\theta}}\in B_W^{\hat{\theta}},\boldsymbol{\theta}\in\bar{\Theta}} g_1(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}) < +\infty, \text{ and } \bar{G} := \sup_{\hat{\boldsymbol{\theta}}\in B_W^{\hat{\theta}},\boldsymbol{\theta}\in\bar{\Theta}} \|\boldsymbol{\nabla}_{\boldsymbol{\theta}} g_1(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})\| < +\infty.$$

Then,

$$\left\|\frac{\partial \tilde{\boldsymbol{F}}}{\partial t}(\boldsymbol{x}(\hat{\boldsymbol{\theta}},t),t,\hat{\boldsymbol{\theta}})\right\| \le \delta\left(\underline{G}R_2+\bar{G}R_1\right) < +\infty \quad \Rightarrow \quad \left\|\frac{\partial \boldsymbol{x}}{\partial t}(\hat{\boldsymbol{\theta}},t)\right\| \le \underbrace{\bar{H}\delta\left(\underline{G}R_2+\bar{G}R_1\right)}_{\bar{L}} < +\infty,$$

Thus,

$$\left\|\boldsymbol{x}(\hat{\boldsymbol{\theta}};t)-\boldsymbol{x}^*\right\| \le \int_0^t \|\partial_t x(\hat{\boldsymbol{\theta}};s)\| ds \le \bar{L}t,$$

□

Proof of Theorem 7 (resumed). By Lemma 4,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{y}}\left[R\left(\boldsymbol{x}(\boldsymbol{y};t);\boldsymbol{\theta}_0\right)\,|\,\boldsymbol{\theta}_0\right]\\
=&\mathbb{P}_{\boldsymbol{y}}\left[\boldsymbol{y}\in B_W^{\hat{\theta}}\right]\cdot\mathbb{E}_{\boldsymbol{y}}\left[R\left(\boldsymbol{x}(\boldsymbol{y};t);\boldsymbol{\theta}_0\right)\,\middle|\,\boldsymbol{\theta}_0,\boldsymbol{y}\in B_W^{\hat{\theta}}\right]+\mathbb{P}_{\boldsymbol{y}}\left[\boldsymbol{y}\notin B_W^{\hat{\theta}}\right]\cdot\mathbb{E}_{\boldsymbol{y}}\left[R\left(\boldsymbol{x}(\boldsymbol{y};t);\boldsymbol{\theta}_0\right)\,\middle|\,\boldsymbol{\theta}_0,\boldsymbol{y}\notin B_W^{\hat{\theta}}\right]\\
\leq&\mathbb{E}_{\boldsymbol{y}}\left[R\left(\boldsymbol{x}(\boldsymbol{y};t);\boldsymbol{\theta}_0\right)-\underbrace{R\left(\boldsymbol{x}^*;\boldsymbol{\theta}_0\right)}_{=0}\,\middle|\,\boldsymbol{\theta}_0,\boldsymbol{y}\in B_W^{\hat{\theta}}\right]+\mathbb{P}_{\boldsymbol{y}}\left[\boldsymbol{y}\notin B_W^{\hat{\theta}}\right]\cdot\underbrace{\sup_{\boldsymbol{y}}R\left(\boldsymbol{x}(\boldsymbol{y};t);\boldsymbol{\theta}_0\right)}_{\bar{M}\text{ from Lemma 2}}\\
\leq&\mathbb{E}_{\boldsymbol{y}}\left[||\nabla_{\boldsymbol{x}}R\left(\boldsymbol{x}^*;\boldsymbol{\theta}_0\right)||\cdot\|\boldsymbol{x}(\boldsymbol{y};t)-\boldsymbol{x}^*\|\,\middle|\,\boldsymbol{\theta}_0,\boldsymbol{y}\in B_W^{\hat{\theta}}\right]+o(\|\boldsymbol{x}(\boldsymbol{y};t)-\boldsymbol{x}^*\|)+\bar{M}\cdot\mathbb{P}_{\boldsymbol{y}}\left[\boldsymbol{y}\notin B_W^{\hat{\theta}}\right]\\
\leq&\underbrace{\bar{L}||\nabla_{\boldsymbol{x}}R\left(\boldsymbol{x}^*;\boldsymbol{\theta}_0\right)||}_{L}\,t+\underbrace{\bar{M}\cdot w(W,N)}_{M(N)}+o(t)\to 0\qquad\text{as}\qquad t\to 0
\end{aligned}$$

□

A.13. Theorem 8

Before we proceed to prove the theorem, we first present an important lemma used in the proof:

LEMMA 5. *For a given $\hat{\theta}\in\Theta$, suppose $x^*(\hat{\theta};w)$ is monotone and differentiable in w. Define*

$$F(w,t,\hat{\theta})=\mathbb{E}_{u_w\sim\theta}[\phi(x^*(\hat{\theta};t);\theta)g_1(\hat{\theta},\theta)]. \tag{34}$$

Then

i. For all $w\leq t$, $\frac{\partial F}{\partial z}\geq 0$; and,

ii. For all $w\geq t$, $\frac{\partial F}{\partial z}\leq 0$.

Proof of Lemma 5. We simply prove the version of the result for the first item i., as the proof of the other is analogous. Note that

$$\begin{aligned}
\frac{\partial F}{\partial t}&=\mathbb{E}_{u_w\sim\theta}\left[\frac{\partial\phi}{\partial x}(x^*(\hat{\theta};t);\theta)\frac{\partial x^*}{\partial t}(\hat{\theta};t)g_1(\hat{\theta},\theta)\right]\\
&=\underbrace{\frac{\partial x^*}{\partial t}(\hat{\theta};t)}_{(A)}\cdot\underbrace{\mathbb{E}_{u_w\sim\theta}\left[\frac{\partial\phi}{\partial x}(x^*(\hat{\theta};t);\theta)g_1(\hat{\theta},\theta)\right]}_{(B)}
\end{aligned}$$

where the last line holds due to the term (A) not involving θ which is averaged in the expectation. Also notice that when $t=w$, then term (B) is just 0 due to Theorem 6, that is, the optimality condition for x^*.

By assumption, x^* is also monotone. Suppose it is non-decreasing. Thus, $(A)\geq 0$. Then for $w\leq t$, $x^*(\hat{\theta};w)\leq x^*(\hat{\theta};t)$. Because ϕ is convex in x, $\frac{\partial\phi}{\partial x}$ is monotone non-decreasing for all θ. Thus, $\frac{\partial\phi}{\partial x}(x^*(\hat{\theta};t);\theta)\geq\frac{\partial\phi}{\partial x}(x^*(\hat{\theta};w);\theta)$. But now, due to the non-negativity of g_1,

$$\begin{aligned}
&\underbrace{\frac{\partial x^*}{\partial t}(\hat{\theta};t)}_{\geq 0}\cdot\mathbb{E}_{u_w\sim\theta}\left[\frac{\partial\phi}{\partial x}(x^*(\hat{\theta};t);\theta)g_1(\hat{\theta},\theta)\right]\\
\geq&\frac{\partial x^*}{\partial t}(\hat{\theta};t)\cdot\underbrace{\mathbb{E}_{u_w\sim\theta}\left[\frac{\partial\phi}{\partial x}(x^*(\hat{\theta};w);\theta)g_1(\hat{\theta},\theta)\right]}_{=0}=0.
\end{aligned}$$

Conversely, if x^* were monotone non-increasing, then $(A) \leq 0$. By similar logic, for $w \leq t$, $x^*(\hat{\theta}; w) \geq x^*(\hat{\theta}; t)$ and therefore, $\frac{\partial \phi}{\partial x}(x^*(\hat{\theta}; t); \theta) \leq \frac{\partial \phi}{\partial x}(x^*(\hat{\theta}; w); \theta)$. But now,

$$\begin{aligned}
&\underbrace{\frac{\partial x^*}{\partial t}(\hat{\theta}; t)}_{\leq 0} \cdot \mathbb{E}_{u_w \sim \theta}\left[\frac{\partial \phi}{\partial x}(x^*(\hat{\theta}; t); \theta) g_1(\hat{\theta}, \theta)\right] \\
\geq &\frac{\partial x^*}{\partial t}(\hat{\theta}; t) \cdot \underbrace{\mathbb{E}_{u_w \sim \theta}\left[\frac{\partial \phi}{\partial x}(x^*(\hat{\theta}; w); \theta) g_1(\hat{\theta}, \theta)\right]}_{=0} = 0.
\end{aligned}$$

The other case is analogous. This completes the proof. □

Proof of Theorem 8. We show the result for (19), as the result for (20) is fully analogous following from Lemma 5. Define the function

$$G(v, w) := \mathbb{E}_{u_v}[\Phi[x^*(\cdot; w)|\theta]].$$

Our goal is to prove that $G(v, w)$ is non-decreasing for all $w \geq v$. Due to the assumption that ϕ is continuously differentiable, G too is differentiable in w. Thus, it suffices to show that $\frac{\partial G}{\partial w}(v, w) \geq 0$ for all $w \geq v$.

By Theorem 4, $G(v, w) = \int_{\hat{\theta}} F(v, w, \hat{\theta}) K(\hat{\theta}) d\hat{\theta}$, and thus, as a result of Lemma 5, if indeed $x^*(\hat{\theta}; w)$ is monotone in w for every $\hat{\theta} \in \Theta$, then we are done, as

$$\frac{\partial G}{\partial w}(v, w) = \int_{\hat{\theta}} \frac{\partial F}{\partial w}(v, w, \hat{\theta}) K(\hat{\theta}) d\hat{\theta} \geq 0 \qquad \forall w \geq v$$

Thus, it suffices to prove that $x^*(\hat{\theta}; w)$ is monotone in w for all $\hat{\theta} \in \Theta$. Note that as a consequence of Lemma 3 from Theorem 7, $x^*(\hat{\theta}; w)$ is differentiable in w and moreover,

$$\frac{\partial x^*(\hat{\theta}; w)}{\partial w} = -\frac{\tilde{F}_z(x^*, z)}{\tilde{F}_x(x^*, z)} = \frac{-\frac{\partial \phi}{\partial x}(x^*(\hat{\theta}; w); \varsigma + w) g_1(\hat{\theta}, \varsigma + w) u(\varsigma + w)}{\mathbb{E}_{u_z \sim \theta}\left[\frac{\partial^2 \phi}{\partial x^2}(x^*(\hat{\theta}; w); \theta) g_1(\hat{\theta}, \theta)\right]} \tag{35}$$

Since $\frac{\partial^2 \phi}{\partial x^2}(x^*(\hat{\theta}; w); \theta) > 0$, it suffices to show that $\frac{\partial \phi}{\partial x}(x^*(\hat{\theta}; w); \varsigma + w)$ has the same sign for all w.

Suppose now that $\frac{\partial^2 \phi}{\partial x \partial \theta} \leq 0$. We aim to show that $\frac{\partial \phi}{\partial x}(x^*(\hat{\theta}; w); \varsigma + w) \leq 0$ for all w by contradiction. Suppose that $\frac{\partial \phi}{\partial x}(x^*(\hat{\theta}; w); \varsigma + w) > 0$ for some w. Since $\frac{\partial^2 \phi}{\partial x \partial \theta} \leq 0$, we have $\frac{\partial \phi}{\partial x}(x; \theta)$ is non-increasing in θ, which implies $\frac{\partial \phi}{\partial x}(x^*(\hat{\theta}; w); \theta) \geq \frac{\partial \phi}{\partial x}(x^*(\hat{\theta}; w); \varsigma + w) > 0$ for all $\theta \in [\varsigma, \varsigma + w]$. Therefore, we have

$$\tilde{F}(x, w) = \mathbb{E}_{u_w \sim \theta}\left[\frac{\partial \phi}{\partial x}(x^*(\hat{\theta}; w); \theta) g_1(\hat{\theta}, \theta)\right] > 0,$$

which contradicts the optimality of x^*.

Similarly, we can show that if $\frac{\partial^2 \phi}{\partial x \partial \theta} \geq 0$, then $\frac{\partial \phi}{\partial x}(x^*(\hat{\theta}; w); \varsigma + w) \geq 0$ for all w. As such, we have shown that x^* is monotone in w. This completes the proof. □

A.14. Corollary 5

Proof of Corollary 5. For (i) and (ii), conditions of Theorem 2 are met; (iii) holds because ϕ is composed of basic differentiable functions whose derivatives are well-defined and continuous for $\theta > 0$; (iv), as ϕ is convex, $\partial\phi/\partial x$ is monotone, with g_1 and u non-negative; (v) holds by observing that $\varphi(\theta) = \min_x \left\{xc - p\theta\left(1 - e^{-x/\theta}\right)\right\} = c\theta\left(\log(\frac{p}{c}) - \frac{p}{c} + 1\right)$, which is negative whenever $p > c$ and linear in θ. Thus, the scale-free prior is the Jeffrey's prior $u(\theta) \propto \theta^{-1}$. □

A.15. Proposition 7

Proof of Proposition 7. Let $x_{\text{RO}}(\boldsymbol{y})$ denote the parameter RO solution. The worst case parameter is always the smallest value in the uncertainty set and thus $x_{\text{RO}}(\boldsymbol{y}) = a(\boldsymbol{y})\log\rho$. Hence,

$$\phi(x_{\text{RO}};\theta) = p\theta\left(\rho^{-1}\frac{a(\boldsymbol{y})\log\rho}{\theta} - 1 + e^{-a(\boldsymbol{y})\log\rho/\theta}\right)$$

$$\text{and} \qquad \Phi[x_{\text{RO}}(\cdot)|\theta] = p\theta\left(\rho^{-1}\mathbb{E}_{\boldsymbol{y}}\left[\frac{a(\boldsymbol{y})\log\rho}{\theta}\right] - 1 + \mathbb{E}_{\boldsymbol{y}}\left[e^{-\frac{a(\boldsymbol{y})\log\rho}{\theta}}\right]\right).$$

Suppose we now consider the perturbed solution: $\bar{x}(\boldsymbol{y}) = a(\boldsymbol{y})\log\rho + \alpha\hat{\theta}(\boldsymbol{y})$, for some $\alpha > 0$. Then this leads to the ungeneralized cost:

$$\Phi[\bar{x}(\cdot)|\theta] = p\theta\left(\rho^{-1}\mathbb{E}_{\boldsymbol{y}}\left[\frac{a(\boldsymbol{y})\log\rho + \alpha\hat{\theta}}{\theta}\right] - 1 + \mathbb{E}_{\boldsymbol{y}}\left[e^{-\frac{a(\boldsymbol{y})\log\rho+\alpha\hat{\theta}}{\theta}}\right]\right).$$

So we have that

$$\begin{aligned}\Phi[\bar{x}(\cdot)|\theta] - \Phi[x_{\text{RO}}(\cdot)|\theta] &= p\theta\left(\rho^{-1}\mathbb{E}_{\boldsymbol{y}}\left[\frac{\alpha\hat{\theta}}{\theta}\right] + \mathbb{E}_{\boldsymbol{y}}\left[e^{-\frac{a(\boldsymbol{y})\log\rho}{\theta}}\left(e^{-\frac{\alpha\hat{\theta}}{\theta}} - 1\right)\right]\right)\\ &= p\theta\left(\rho^{-1}\alpha + \mathbb{E}_{\boldsymbol{y}}\left[e^{-\frac{a(\boldsymbol{y})\log\rho}{\theta}}\left(e^{-\frac{\alpha\hat{\theta}}{\theta}} - 1\right)\right]\right).\end{aligned}$$

Hence, we are asking if there exists some choice of ρ and α such that

$$\alpha < \rho\mathbb{E}_{\boldsymbol{y}}\left[e^{-\frac{a(\hat{\theta})\log\rho}{\theta}}\left(1 - e^{-\frac{\alpha\hat{\theta}}{\theta}}\right)\right], \qquad \forall\theta\in\mathbb{R}^+.$$

By restriction, $a(\boldsymbol{y}) \le \hat{\theta} \iff e^{-a(\boldsymbol{y})\log\rho/\theta} \ge e^{-\hat{\theta}\log\rho/\theta}$. Thus, it suffices to find some choice of ρ and α such that

$$\rho\mathbb{E}_{\boldsymbol{y}}\left[e^{-\frac{\hat{\theta}\log\rho}{\theta}}\left(1 - e^{-\frac{\alpha\hat{\theta}}{\theta}}\right)\right] - \alpha > 0, \qquad \forall\theta\in\mathbb{R}^+.$$

The moment generating function of the sample mean $\hat{\theta}$ under the exponential distribution with mean θ is $\mathbb{E}_{\boldsymbol{y}}[e^{t\hat{\theta}}] = (1 - t\theta/N)^{-N}$. Evaluating the above inequality, we get

$$\rho\left[\left(\frac{1}{1+\log\rho/N}\right)^N - \left(\frac{1}{1+(\log\rho+\alpha)/N}\right)^N\right] - \alpha > 0.$$

Notice that this inequality is independent of θ. Hence, should it hold for some value of ρ and α, then the conclusion would hold for all $\theta \in \mathbb{R}^+$.

Since $\alpha > 0$, by the mean value theorem, there exists some $\xi \in (\log\rho, \log\rho + \alpha)$ such that

$$\left(\frac{1}{1+\log\rho/N}\right)^N - \left(\frac{1}{1+(\log\rho+\alpha)/N}\right)^N = \alpha\left(\frac{1}{1+\xi/N}\right)^{N+1}.$$

Since

$$\rho\left[\left(\frac{1}{1+\log\rho/N}\right)^N - \left(\frac{1}{1+(\log\rho+\alpha)/N}\right)^N\right] - \alpha > 0 \quad \Longleftrightarrow \quad \rho\left(\frac{1}{1+\xi/N}\right)^{N+1} > 1,$$

and since, $\xi < \log\rho + \alpha$, it suffices to show that there exists some $\alpha > 0$ and $\rho > 1$ such that

$$\zeta(\rho,\alpha) := \rho\left(\frac{1}{1+(\log\rho+\alpha)/N}\right)^{N+1} > 1.$$

Note that ζ is differentiable, and therefore continuous, in $\rho > 1$. Moreover, since

$$\begin{aligned}\frac{\partial\zeta(\rho,\alpha)}{\partial\rho} &= \left(\frac{1}{1+(\log\rho+\alpha)/N}\right)^{N+1} + \rho\cdot\left[-(N+1)\left(\frac{1}{1+(\log\rho+\alpha)/N}\right)^{N+2}\right]\cdot\frac{1}{N\rho}\\ &= \left(1+\frac{\log\rho+\alpha}{N} - \rho(N+1)\frac{1}{N\rho}\right)\left(\frac{1}{1+(\log\rho+\alpha)/N}\right)^{N+2}\\ &= \frac{\log\rho+\alpha-1}{N}\left(1+\frac{\log\rho+\alpha}{N}\right)^{-(N+2)},\end{aligned}$$

when $\rho \geq e^{1-\alpha}$, *i.e.* $\log\rho + \alpha \geq 1$, then $\frac{\partial\zeta}{\partial\rho} \geq 0$ with equality only when $\log\rho + \alpha = 1$. Thus, ζ is continuous and increasing in ρ in the range $[e^{1-\alpha}, +\infty)$.

Now consider two points of ζ. First, for $\zeta(e^{1-\alpha},\alpha) = e^{1-\alpha}(1+1/N)^{-(N+1)}$, the famous inequality

$$\left(1+\frac{1}{N}\right)^N < e < \left(1+\frac{1}{N}\right)^{N+1}$$

permits $\zeta(e^{1-\alpha},\alpha) < e^{1-\alpha}e^{-1} = e^{-\alpha} < 1$; whereas for $\zeta(e^{N+1},\alpha) = e^{N+1}(1+(N+1+\alpha)/N)^{-(N+1)}$; note that

$$e^{N+1} > \left(1+\frac{N+1+\alpha}{N}\right)^{N+1} \quad \Longleftrightarrow \quad e > 1+\frac{N+1+\alpha}{N} = 2+\frac{1+\alpha}{N}.$$

Thus for any $\alpha < \bar{\alpha}(N) := e - 2 - 1/N$ (Note: $\bar{\alpha}(N) > 0$ for all $N \geq 2$), by the intermediate value theorem, there exists some $\bar{\rho}(N,\alpha)$ such that $\zeta(\bar{\rho}(N,\alpha),\alpha) = 1$. Then for some small $\epsilon > 0$, $\bar{\rho}(N) = \inf_{\alpha\in[\epsilon,\bar{\alpha}(N)]}\bar{\rho}(N,\alpha)$ is the $\bar{\rho}$ that we desire, whereupon for any $\rho \geq \bar{\rho}$, we can find an $\epsilon < \alpha < \bar{\alpha}(N)$ such that the constructed solution $\bar{x}$ uniformly dominates $\boldsymbol{x}_{\text{RO}}$ for all parameters $\theta \in \mathbb{R}^+$. □

REMARK 9. Technically, any Pareto optimal solution can also be written in the form $x(\boldsymbol{y}) = a(\boldsymbol{y})\log\rho$, simply by dividing it by the constant $\log\rho$. However, this does not guarantee that $a(\boldsymbol{y}) < \hat{\theta}(\boldsymbol{y})$, which is the point of the construction, to increase x_{RO} such that it exceeds $\hat{\theta}(\boldsymbol{y})\log\rho$ for some values of $\boldsymbol{y}$, and no longer remains a reasonably data-driven RO solution.

A.16. Corollary 6

Proof of Corollary 6. i. ϕ is a (second-order) polynomial in $\boldsymbol{x}$.

ii. First-order conditions on ϕ easily yields the desired expression.

iii. By Theorem 6, for a given prior u, $\boldsymbol{x}_{\text{OVE}}$ is the solution that point-wise minimizes $\beth(\boldsymbol{x},\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\theta}\sim u}\left[\phi(\boldsymbol{x}(\hat{\boldsymbol{\theta}});\boldsymbol{\theta})g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\right]$ over $\boldsymbol{x}\in\mathcal{B}$ for each $\hat{\boldsymbol{\theta}}\in\Theta$. As ϕ is convex, g_1 and u are non-negative, the minimizer is given by the first order condition.

iv. Consider the OVE-policy under the prior $u\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{V})$, we have

$$\begin{aligned}\beth(\boldsymbol{x},\hat{\boldsymbol{\theta}}) =&\mathbb{E}_{\boldsymbol{\theta}\sim u}\left[\phi(\boldsymbol{x}(\hat{\boldsymbol{\theta}});\boldsymbol{\theta})g_1(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\right]\\ =&\int_\Theta\left[-(1-\mathbf{1}^\top\boldsymbol{x})r_f-\boldsymbol{\theta}^\top\boldsymbol{x}+\frac{1}{2\tau}\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}\right]\exp\left(-\frac{N}{2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^\top\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\right)\\ &(2\pi)^{-\frac{S}{2}}|\boldsymbol{V}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^\top\boldsymbol{V}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right)d\boldsymbol{\theta}.\end{aligned}$$

First order condition gives

$$\begin{aligned}\nabla_{\boldsymbol{x}}\beth(\boldsymbol{x},\hat{\boldsymbol{\theta}}) =&\int_\Theta\left[\mathbf{1}r_f-\boldsymbol{\theta}+\frac{1}{\tau}\boldsymbol{\Sigma}\boldsymbol{x}\right]\exp\left(-\frac{N}{2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^\top\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\right)\\ &(2\pi)^{-\frac{S}{2}}|\boldsymbol{V}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^\top\boldsymbol{V}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right)d\boldsymbol{\theta}=\mathbf{0}.\end{aligned}$$

Then reorganize the above equations, we have

$$\begin{aligned}&\int_\Theta\left[\mathbf{1}r_f-\boldsymbol{\theta}+\frac{1}{\tau}\boldsymbol{\Sigma}\boldsymbol{x}\right]\underbrace{(2\pi)^{-S/2}\left|N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}\right|^{1/2}}_{(A)}\\ &\underbrace{\exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}-(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1})^{-1}(N\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}}+\boldsymbol{V}^{-1}\boldsymbol{\mu})\right)^\top\right.}_{(A)}\\ &\left.\underbrace{(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1})\left(\boldsymbol{\theta}-(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1})^{-1}(N\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}}+\boldsymbol{V}^{-1}\boldsymbol{\mu})\right)\right)}_{(A)} &&=\mathbf{0},\\ &\underbrace{\exp\left(-\frac{N}{2}\hat{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}}-\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{V}^{-1}\boldsymbol{\mu}\right.}_{(B)}\\ &\left.\underbrace{+\frac{1}{2}(N\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}}+\boldsymbol{V}^{-1}\boldsymbol{\mu})^\top(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1})^{-1}(N\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}}+\boldsymbol{V}^{-1}\boldsymbol{\mu})\right)}_{(B)}d\boldsymbol{\theta}\end{aligned}$$

where this holds because $|N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}|^{1/2}$ and $|\boldsymbol{V}^{-1}|^{1/2}$ are just constants.

Note that both $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ are symmetric and p.s.d., so $N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}$ too is symmetric and p.s.d.. Moreover, (B) no longer contains $\boldsymbol{\theta}$ and thus exits the integral and vanishes, while (A) simply turn out

to be the p.d.f. of a multi-variate normal distribution with mean $(N\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}}+\boldsymbol{V}^{-1}\boldsymbol{\mu})^\top(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1})^{-1}$ and covariance $(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1})^{-1}$. Thus, the first-order condition reduces to

$$\mathbf{1}r_f+\frac{1}{\tau}\boldsymbol{\Sigma}^\top\boldsymbol{x}-(N\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}}+\boldsymbol{V}^{-1}\boldsymbol{\mu})^\top(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1})^{-1}=\mathbf{0},$$

which gives $\boldsymbol{x}_{\text{OVE}}=\tau\boldsymbol{\Sigma}^{-1}((N\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}}+\boldsymbol{V}^{-1}\boldsymbol{\mu})^\top(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1})^{-1}-\mathbf{1}r_f)$.

v. It is easy to check that the relations hold.

□

A.17. Proposition 8

Proof of Proposition 8. The out-of-sample objective Φ is quadratic in $\hat{\boldsymbol{\theta}}$ and can be evaluated given $\boldsymbol{\theta}$ using the identity $\mathbb{E}\left[\boldsymbol{Y}^\top\boldsymbol{A}\boldsymbol{Y}\right]=\boldsymbol{\mu}^\top\boldsymbol{A}\boldsymbol{\mu}+\text{Tr}(\boldsymbol{A}\boldsymbol{\Sigma})$ if $\boldsymbol{Y}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Now, $\hat{\boldsymbol{\theta}}\sim\mathcal{N}\left(\boldsymbol{\theta},\frac{1}{N}\boldsymbol{\Sigma}\right)$. Now,

$$\begin{aligned}
&\Phi\left[\boldsymbol{x}_{\text{ETO}}\mid\boldsymbol{\theta}\right]-\Phi\left[\boldsymbol{x}_{\text{OVE}}\mid\boldsymbol{\theta}\right]=\mathbb{E}\left[\phi(\boldsymbol{x}_{\text{ETO}},\boldsymbol{\theta})-\phi(\boldsymbol{x}_{\text{OVE}},\boldsymbol{\theta})\mid\boldsymbol{\theta}\right]\\
=&\mathbb{E}\left[\left(\mathbf{1}r_f-\boldsymbol{\theta}\right)^\top\left(\boldsymbol{x}_{\text{ETO}}-\boldsymbol{x}_{\text{OVE}}\right)+\frac{1}{2\tau}\left(\boldsymbol{x}_{\text{ETO}}-\boldsymbol{x}_{\text{OVE}}\right)^\top\boldsymbol{\Sigma}\left(\boldsymbol{x}_{\text{ETO}}+\boldsymbol{x}_{\text{OVE}}\right)\,\middle|\,\boldsymbol{\theta}\right]\\
=&\mathbb{E}\left[-\tau\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\left(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{\text{OVE}}\right)+\frac{1}{2}\tau\left(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{\text{OVE}}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\hat{\boldsymbol{\theta}}+\hat{\boldsymbol{\theta}}_{\text{OVE}}\right)\,\middle|\,\boldsymbol{\theta}\right]\\
=&\mathbb{E}\left[\frac{1}{2}\tau\left(\hat{\boldsymbol{\theta}}+\hat{\boldsymbol{\theta}}_{\text{OVE}}-2\boldsymbol{\theta}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{\text{OVE}}\right)\,\middle|\,\boldsymbol{\theta}\right]\\
=&\mathbb{E}\left[\frac{1}{2}\tau\left(2\hat{\boldsymbol{\theta}}-2\boldsymbol{\theta}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{\text{OVE}}\right)\,\middle|\,\boldsymbol{\theta}\right]-\mathbb{E}\left[\frac{1}{2}\tau\left(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{\text{OVE}}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{\text{OVE}}\right)\,\middle|\,\boldsymbol{\theta}\right]\\
=&\tau\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^\top\boldsymbol{\Sigma}^{-1}\left(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{V}^{-1}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}\right)\,\middle|\,\boldsymbol{\theta}\right]\qquad(36)\\
&-\frac{\tau}{2}\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}\right)^\top\boldsymbol{V}^{-1}\left(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{\Sigma}^{-1}\left(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{V}^{-1}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}\right)\,\middle|\,\boldsymbol{\theta}\right]
\end{aligned}$$

where in (36), we substituted the form for $\hat{\boldsymbol{\theta}}_{\text{OVE}}$. Notice that

$$\hat{\boldsymbol{\Sigma}}:=\left(N\boldsymbol{V}+\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}\left(N\boldsymbol{V}+\boldsymbol{\Sigma}\right)^{-1}=\boldsymbol{V}^{-1}\left(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{\Sigma}^{-1}\left(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{V}^{-1}.$$

Thus, we can simplify

$$\begin{aligned}
&\Phi\left[\boldsymbol{x}_{\text{ETO}}\mid\boldsymbol{\theta}\right]-\Phi\left[\boldsymbol{x}_{\text{OVE}}\mid\boldsymbol{\theta}\right]\\
=&\tau\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^\top\left(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}\right)\boldsymbol{V}\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}\right)\,\middle|\,\boldsymbol{\theta}\right]-\frac{\tau}{2}\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}\right)^\top\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}\right)\,\middle|\,\boldsymbol{\theta}\right]\\
=&\tau\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}\right)^\top\left(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}\right)\boldsymbol{V}\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}\right)\,\middle|\,\boldsymbol{\theta}\right]+\tau\mathbb{E}\left[\left(\boldsymbol{\mu}-\boldsymbol{\theta}\right)^\top\left(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}\right)\boldsymbol{V}\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}\right)\,\middle|\,\boldsymbol{\theta}\right]\\
&-\frac{\tau}{2}\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}\right)^\top\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}\right)\,\middle|\,\boldsymbol{\theta}\right]\\
=&\tau\left(\boldsymbol{\theta}-\boldsymbol{\mu}\right)^\top\left(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}\right)\boldsymbol{V}\hat{\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}-\boldsymbol{\mu}\right)+\tau\,\text{Tr}\left(\left(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}\right)\boldsymbol{V}\hat{\boldsymbol{\Sigma}}\frac{\boldsymbol{\Sigma}}{N}\right)\\
&+\tau\left(\boldsymbol{\mu}-\boldsymbol{\theta}\right)^\top\left(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{V}^{-1}\right)\boldsymbol{V}\hat{\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}-\boldsymbol{\mu}\right)-\frac{\tau}{2}\left(\boldsymbol{\theta}-\boldsymbol{\mu}\right)^\top\hat{\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}-\boldsymbol{\mu}\right)-\frac{\tau}{2}\,\text{Tr}\left(\hat{\boldsymbol{\Sigma}}\frac{\boldsymbol{\Sigma}}{N}\right)\\
=&\frac{\tau}{2N}\,\text{Tr}\left(\hat{\boldsymbol{\Sigma}}\left(2\left(N\boldsymbol{V}+\boldsymbol{\Sigma}\right)-\boldsymbol{\Sigma}\right)\right)-\frac{\tau}{2}\left(\boldsymbol{\theta}-\boldsymbol{\mu}\right)^\top\hat{\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}-\boldsymbol{\mu}\right)\\
=&\frac{\tau}{2}\left[\frac{1}{N}\,\text{Tr}\left(\hat{\boldsymbol{\Sigma}}\left(2N\boldsymbol{V}+\boldsymbol{\Sigma}\right)\right)-\left(\boldsymbol{\theta}-\boldsymbol{\mu}\right)^\top\hat{\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}-\boldsymbol{\mu}\right)\right]
\end{aligned}$$

This recovers the result. □

A.18. Proposition 9

Proof of Proposition 9. Let $\mathcal{A}' := \left\{\boldsymbol{\theta} \,\middle|\, \boldsymbol{\theta}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\theta} \le \psi\right\}$ and $u' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{V})$. Then clearly, $\mathbb{P}_u[\boldsymbol{\theta} \in \mathcal{A}] = \mathbb{P}_{u'}[\boldsymbol{\theta} \in \mathcal{A}']$. Define $\mathcal{A}'' := \left\{\boldsymbol{\theta} \,\middle|\, \boldsymbol{\theta}^\top \boldsymbol{V}^{-1} \boldsymbol{\theta} \le \psi_V\right\}$ for some ψ_V. We seek the largest such ψ_V such that $\mathcal{A}'' \subseteq \mathcal{A}'$. Consider the problem:

$$\begin{aligned}
&\max \quad \psi_V \\
&\text{s.t.} \quad \left\{\boldsymbol{\theta} \,\middle|\, \boldsymbol{\theta}^\top \boldsymbol{V}^{-1} \boldsymbol{\theta} \le \psi_V\right\} \subseteq \left\{\boldsymbol{\theta} \,\middle|\, \boldsymbol{\theta}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\theta} \le \psi\right\} \\
=&\max \quad \psi_V \\
&\text{s.t.} \quad \max_{\boldsymbol{\theta}: \boldsymbol{\theta}^\top \boldsymbol{V}^{-1} \boldsymbol{\theta} \le \psi_V} \boldsymbol{\theta}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\theta} \le \psi \\
=&\max \quad \psi_V \\
&\text{s.t.} \quad \max_{\boldsymbol{\theta}: \|\boldsymbol{\theta}\| \le \psi_V^{1/2}} \boldsymbol{\theta}^\top \left(\frac{1}{\psi} \boldsymbol{V}^{1/2} \hat{\boldsymbol{\Sigma}} \boldsymbol{V}^{1/2}\right) \boldsymbol{\theta} \le 1
\end{aligned}$$

Then the radius of the largest circle that can be inscribed in an ellipsoid is precisely the shortest semi-axis of the ellipsoid, given by $\left(\bar{\psi}/\psi\right)^{-1/2}$. Thus, $\psi_V = \psi/\bar{\psi}$.

However, we compute that

$$\begin{aligned}
\mathbb{P}_u[\boldsymbol{\theta} \in \mathcal{A}] \ge& \mathbb{P}_{u'}[\boldsymbol{\theta} \in \mathcal{A}''] \\
=& \mathbb{P}_{\mathcal{N}(\mathbf{0}, \boldsymbol{I})}[\boldsymbol{\theta}^\top \boldsymbol{\theta} \le \psi_V] \\
=& \mathbb{P}\left[\chi^2(d) \le \psi_V\right] \\
=& \mathcal{F}_{\chi^2}\left(\psi_V; d\right)
\end{aligned}$$

In particular, if $\boldsymbol{V} = \Lambda \boldsymbol{\Sigma}$, then $\hat{\boldsymbol{\Sigma}} = \frac{1}{(N\Lambda+1)^2} \boldsymbol{\Sigma}^{-1}$ and $\psi = \frac{d(2N\Lambda+1)}{N(N\Lambda+1)^2}$, thus we can represent $\mathcal{A}' := \left\{\boldsymbol{\theta} \,\middle|\, \boldsymbol{\theta}^\top \boldsymbol{V}^{-1} \boldsymbol{\theta} \le \frac{d(2N\Lambda+1)}{N\Lambda}\right\}$. Hence,

$$\begin{aligned}
\mathbb{P}_{u'}[\boldsymbol{\theta} \in \mathcal{A}'] =& \mathbb{P}_{\mathcal{N}(\mathbf{0}, \boldsymbol{I})}\left[\boldsymbol{\theta}^\top \boldsymbol{\theta} \le \frac{d(2N\Lambda+1)}{N\Lambda}\right] \\
=& \mathcal{F}_{\chi^2}\left(2d + \frac{d}{N\Lambda}; d\right)
\end{aligned}$$

□

A.19. Pseudocode for solving general convex OVE objectives

Based on Corollary 3, solving for the prior optimal solution reduces to solving a convex optimization. We present a sample pseudocode on solving the convex optimization problem, using barrier method (see, for instance, Algorithm 11.1 in Boyd and Vandenberghe 2004).

Algorithm 2 General OVE algorithm

Require: $\hat{\boldsymbol{\theta}}$, Strictly feasible $\boldsymbol{x}^0$, initial barrier parameter $t^0 > 0$, barrier parameter multiplier $\kappa > 1$, tolerance $\epsilon > 0$, and a method (*e.g.*, Newton's method) for solving unconstrained convex optimization problems.

function OVE_OBJ($\boldsymbol{x}, \hat{\boldsymbol{\theta}}, t$):
    OVE_obj $\leftarrow t\sqsupset(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) + \sum_{i=1}^{I} -\log\left(-h_i(\boldsymbol{x})\right)$
end function
function OVESOLVE($\hat{\boldsymbol{\theta}}$):
    $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$
    $t \leftarrow t^0$
    while $I/t > \epsilon$ **do**
        Compute $\boldsymbol{x}^*$ by minimizing OVE_obj$(\boldsymbol{x}, \hat{\boldsymbol{\theta}}, t)$ using the given method, starting at $\boldsymbol{x}$.
        $\boldsymbol{x} \leftarrow \boldsymbol{x}^*$
        $t \leftarrow \kappa t$
    end while
return $\boldsymbol{x}$
end function

B. Further details of the numerical results for newsvendor example

In this Appendix, we provide further details on all numerical experiments performed for the newsvendor illustration.

Generating datasets: We chose $M = 50$, that is, the number of samples to draw from the prior, $N = 10$, that is, the number of observed demand data points per data set, and $\bar{N} = 200$, that is the number of different data sets we re-sampled in order to obtain a good spread over the sampling distribution. We chose $N = 10$ as differences between the models are sufficiently pronounced for clear comparisons. Findings are consistent even if N is increased (see discussion later and Appendix B.6 for sensitivity analysis under different N). When N is large, the number of data points is sufficient to estimate $\hat{\theta}$ to relatively high accuracy, thus all solutions begin to converge to the perfect information oracle. Notice that none of the benchmark models depend on M and $\bar{N}$. We chose $\bar{N}$ to be sufficiently large as we realized that there is reasonable amount of variation in performance across data sets (which also arises because the variation in $\hat{\theta}$ is large when N is small). In particular, we chose $\bar{N}$ to be large enough so that OS performs better than ETO on almost every θ, as is theoretically known. This would be a good indication that the variations have been sufficiently averaged away, and that happens roughly around $\bar{N} \geq 200$, which is what we have chosen. When $M = 20$, OVE already exhibits very clear distinction against the other benchmark models, but we chose $M = 50$ so

that we would obtain a better spread of test θ's, and to give ample chance for outliers to occur to test the generalizability of the models. Notice that for the prior, we had used different sets of ($M = 50$) points for computing the OVE solution in Algorithm 1 versus the out-of-sample performance in Algorithm 3, so as to ensure that OVE would be able to generalize regardless of the sample used for the prior.

In Algorithm 1, the upper and lower bounds for the bisection search in Algorithm 1 are given as $x_1 = 0$ and $x_2 = 2\hat{\theta}$, with automatic subsequent relaxation of x_2 if it is initially tested to be of the same sign as the solution generated slightly above x_1. Note that g_1 is likely a very small number, because it was originally the density of a high-dimensional integral, but is now defined only on a subspace subtended by $\hat{\theta}$. In the case of the newsvendor model, it would involve large divisions by θ_m^N and $e^{-N\hat{\theta}/\theta_m}$, thus numerical stability issues must be carefully circumvented, through the multiplication of a large constant S, which can be chosen as the circumstances dictate. For our case, we had used $\check{\theta}^N e^{N\hat{\theta}/\grave{\theta}}$, where $\check{\theta} = \min \theta_m$ and $\grave{\theta} = \max \theta_m$.

Within each experiment, for all the benchmarks, we used the same set of samples for the prior, and data sets. Only the data sets used for the cross-validation for the Wasserstein and KL divergence radii differ (see discussion in Appendix B.3).

B.1. Evaluating performance

In Algorithm 3, we present a sample pseudocode for evaluating the out-of-sample performance for a prescriptive solution, such as OVE, for the newsvendor problem with exponential demand.

B.2. Mathematical formulations for benchmark models

Here we clearly state the formulations for the benchmark models used in our numerical experiments, specific to the case of the newsvendor model.

Robust optimization model with θ as the uncertainty: The robust newsvendor problem can be expressed as

$$\min_{x\geq 0} \sup_{\theta\in\Omega} \{\phi(x;\theta)\} = \min_{x\geq 0} \sup_{\theta\in\Omega} \left\{xc - p\theta\left(1-e^{-x/\theta}\right)\right\}, \tag{37}$$

for some uncertainty set Ω. In the context of our setting, θ is the mean-parameter of the exponential distribution and is one-dimensional. Requiring Ω to be closed and convex, Ω will essentially be a closed interval – $\Omega = [\underline{\omega}, \bar{\omega}]$ containing $\hat{\theta}$. We specify these bounds as a fraction of $\hat{\theta}$, specifically, $\Omega = [0.95\,\hat{\theta}, 1.05\,\hat{\theta}]$.

Note that $\phi(x,\theta)$ is jointly convex in x and θ. Hence the worst-case θ must belong to the boundary of Θ, and the problem simplifies to

$$\min_{x\geq 0} \max\left\{xc - p\underline{\omega}\left(1-e^{-x/\underline{\omega}}\right), xc - p\bar{\omega}\left(1-e^{-x/\bar{\omega}}\right)\right\} = \min_{x\geq 0}\left\{xc - p\underline{\omega}\left(1-e^{-x/\underline{\omega}}\right)\right\}.$$

The equation above is due to the fact that $\frac{\partial}{\partial\theta}(xc - p\theta(1-e^{-x/\theta})) \leq 0$ for all $x \geq 0$ and $\theta > 0$. Thus, this problem has closed form solution $x^* = \underline{\omega}\log(p/c)$. In other words, 'the worst-case scenario is always when the demand is smaller than expected'.

Algorithm 3 Evaluating performance of OVE and benchmarks for exponential newsvendor

Require: $p, c, N, \mathcal{U}, \bar{N}$. Let $M = |\mathcal{U}|$. Functions from Algorithm 1. Let Solve be some function for obtaining the policy, *e.g.*, 'OVEsolve'.

function TRUE_COST(x, θ):
    True_cost $\leftarrow xc - p\theta\left(1 - e^{-x/\theta}\right)$
end function
for θ_m in $\mathcal{U}$ **do**
    for $i = 1$ to $\bar{N}$ **do**
        Sample $z_1, \ldots, z_N$ from $f(z; \theta_m)$.
        $\hat{\theta} \leftarrow \dfrac{1}{N}\sum_{n=1}^{N} z_n$
        $x \leftarrow$ Solve$(\hat{\theta})$
        $\phi[i] \leftarrow$ True_cost(x, θ_m)
    end for
    $\Phi[\theta_m] \leftarrow \dfrac{1}{\bar{N}}\sum_{i=1}^{\bar{N}} \phi[i]$
end for
return $\dfrac{1}{M}\sum_{m=1}^{M} \Phi[\theta_m]$

Distributionally robust optimization model with moment uncertainty: The distributionally robust optimization formulation for the newsvendor problem can be written as:

$$\min_{x \geq 0} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P} \sim z}\left[xc - p\min\{x, z\}\right], \tag{38}$$

for some ambiguity set $\mathcal{P}$. In the case of moments uncertainty, $\mathcal{P}$ is

$$\mathcal{P}_m = \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}_+) \;\middle|\; \begin{array}{l} z \sim \mathbb{P} \\ \mathbb{E}_{\mathbb{P}}[z] = \hat{z} \\ \mathbb{E}_{\mathbb{P}}\left[(z - \hat{z})^2\right] \leq \hat{\sigma}^2 \end{array} \right\},$$

with robust counterpart

$$\begin{array}{ll} \inf & \lambda\hat{z} + \beta\hat{\sigma}^2 + \gamma \\ \text{s.t.} & \lambda z + \beta(z - \hat{z})^2 + \gamma \geq xc - p\min\{x, z\}, \qquad \forall z \geq 0, \\ & \beta \geq 0,\ \lambda, \gamma \text{ free.} \end{array}$$

Distributionally robust optimization model with Wasserstein uncertainty set: We instead consider the ambiguity set $\mathcal{P}_W^r$ in (38),

$$\mathcal{P}_W^r = \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}_+) \;\middle|\; \begin{array}{l} z \sim \mathbb{P} \\ \Delta_W(\mathbb{P}, \hat{\mathbb{P}}) \leq r \end{array} \right\},$$

where $\Delta_W(\mathbb{P},\hat{\mathbb{P}})$ is the Type-1 Wasserstein distance defined on some norm $||\cdot||$, and the empirical distribution $\hat{\mathbb{P}}$ is given by $\hat{\mathbb{P}}[z=y_n]=1/N$, for all $n=1,\ldots,N$. Its robust counterpart is

$$\begin{aligned}
&\inf\ \lambda r+\frac{1}{N}\sum_{n=1}^{N}\beta_n\\
&\text{s.t. } \lambda\|z-y_n\|+\beta_n\geq xc-p\min\{x,z\}, \qquad \forall z\geq 0,\ n=1,\ldots,N,\\
&\quad \lambda\geq 0,\ \boldsymbol{\beta} \text{ free.}
\end{aligned}$$

In our case, we shall just consider the L_1-norm, and this problem can be easily solved as a linear program. We searched for the radius r via cross-validation and how this is done is explained in Appendix B.3.

Distributionally robust optimization model with Kullback-Leibler uncertainty set: We instead consider the ambiguity set $\mathcal{P}^r_{KL}$ in (38),

$$\mathcal{P}^r_{KL}=\left\{\mathbb{P}\in\mathcal{P}(\mathbb{R}_+)\ \middle|\ \begin{array}{l} z\sim\mathbb{P}\\ \Delta_{KL}(\mathbb{P},\hat{\mathbb{P}})\leq r\end{array}\right\},$$

and $\Delta_{KL}(\mathbb{P},\hat{\mathbb{P}})$ is the Kullback-Leibler divergence. Let $\hat{p}_n=\hat{\mathbb{P}}[z=y_n]$, for all $n=1,\ldots,N$. The corresponding robust counterpart is

$$\begin{aligned}
&\inf\ \lambda r+\sum_{n=1}^{N}\beta_n\hat{p}_n+\gamma\\
&\text{s.t. } \lambda\log\left(\lambda/\beta_n\right)+xc-p\min\{x,y_n\}\leq\lambda+\gamma, \qquad \forall n=1,\ldots,N,\\
&\quad \lambda\geq 0,\ \gamma \text{ free, } \boldsymbol{\beta}\geq\mathbf{0}.
\end{aligned}$$

We follow the convention that $0\log(0/t)=0$ if $t\geq 0$. This problem can be expressed as an exponential cone program, and we solved it using the exponential cone solver in MOSEK. We searched for the radius r via cross-validation and how this is done is explained in Appendix B.3.

DRO model with moment uncertainty on assumed exponential demand: We would have liked to examine DRO models with uncertain θ. In practice, there is only one data set – a single observation of θ. Only in the case of the moments uncertainty set, it is possible to estimate the variance of θ using the sample variance of the sample mean. However, this leads to a nonconvex formulation, thus it is not considered as a benchmark. We describe such a model below:

The distributionally robust optimization model assuming uncertainty in the unknown parameter θ can be expressed as

$$\min_{x\geq 0}\sup_{\mathbb{P}\in\mathcal{P}^e_m}\{\mathbb{E}_{\mathbb{P}}[\phi(x,\theta)]\}=\min_{x\geq 0}\sup_{\mathbb{P}\in\mathcal{P}^e_m}\left\{\mathbb{E}_{\mathbb{P}\sim\tilde{\theta}}\left[xc-p\tilde{\theta}\left(1-e^{-x/\tilde{\theta}}\right)\right]\right\},$$

where

$$\mathcal{P}_m^e = \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}_+) \;\middle|\; \begin{array}{l} \tilde{\theta} \sim \mathbb{P} \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{\theta}\right] = \hat{\theta} \\ \mathbb{E}_{\mathbb{P}}\left[\left(\tilde{\theta} - \hat{\theta}\right)^2\right] \leq s^2 \end{array} \right\},$$

and s^2 is the unbiased estimator for the variance of the sample mean $\hat{\theta}$. Unfortunately, this model is non-convex. As such, we do not consider it in our simulations.

B.3. Cross-validation for Wasserstein and KL radii

In our simulations, we aim to obtain the best parameter for the radii for the Wasserstein and KL divergence ambiguity sets to illustrate the limits of these approaches. In other words, the results presented for these two models are already conditional on having chosen the best possible radii $r = r(\boldsymbol{y}, \theta)$ per data set $\boldsymbol{y}$ and for each testing parameter θ where the performance is to be evaluated upon. As such, the procedure here would be possibly considered 'counterfactual' since in each instance, the decision-maker would only possess the training data set, and can only conduct cross-validation on that training data set. This is not even reasonable in some cases, such as in this simulation experiment where the number of data points in each data set is $N = 10$.

More specifically, we adopted the following procedure: For each **testing** parameter θ sampled under the testing prior, and for each data set $\boldsymbol{y}$ that is generated under parameter θ, the oracle's objective is $\phi(x; \theta)$. For each $r \in \{0, 0.05, 0.1, \ldots, 1.95, 2\}$ for the KL divergence and $r \in \{0, 0.05, 0.1, \ldots, 14.95, 15\}$ for the Wasserstein distance, we compute:

$$x^r \in \arg\min_x \sup_{\mathbb{P} \in \mathcal{P}^r(\boldsymbol{y})} \mathbb{E}_{\mathbb{P}}[C(x, z)],$$

where $\mathcal{P}^r(\boldsymbol{y})$ is the corresponding ambiguity set of radius r centred on the empirical distribution, here the training data set $\boldsymbol{y}$. We then pick the calibrated radius $\bar{r}$ as the clairvoyant radius $\bar{r} := \arg\min_r \phi(x^r; \theta)$. Based on this, no other radius r can achieve a better performance against the testing oracle.

Figures 10 and 11 below show a distribution for the optimal calibrated radius of uncertainty, that is drawn over all the testing parameters θ sampled from the testing prior and over all data sets sampled given that parameter. Notice that a significant proportion of the time, the optimal radius was zero. This happens because in many cases, the worst-case direction is opposite to the empirical distribution (which is the centre of the ambiguity set) as the true distribution. Therefore, solving the robust problem for any radius would perform poorer than the empirical distribution for that given data set.

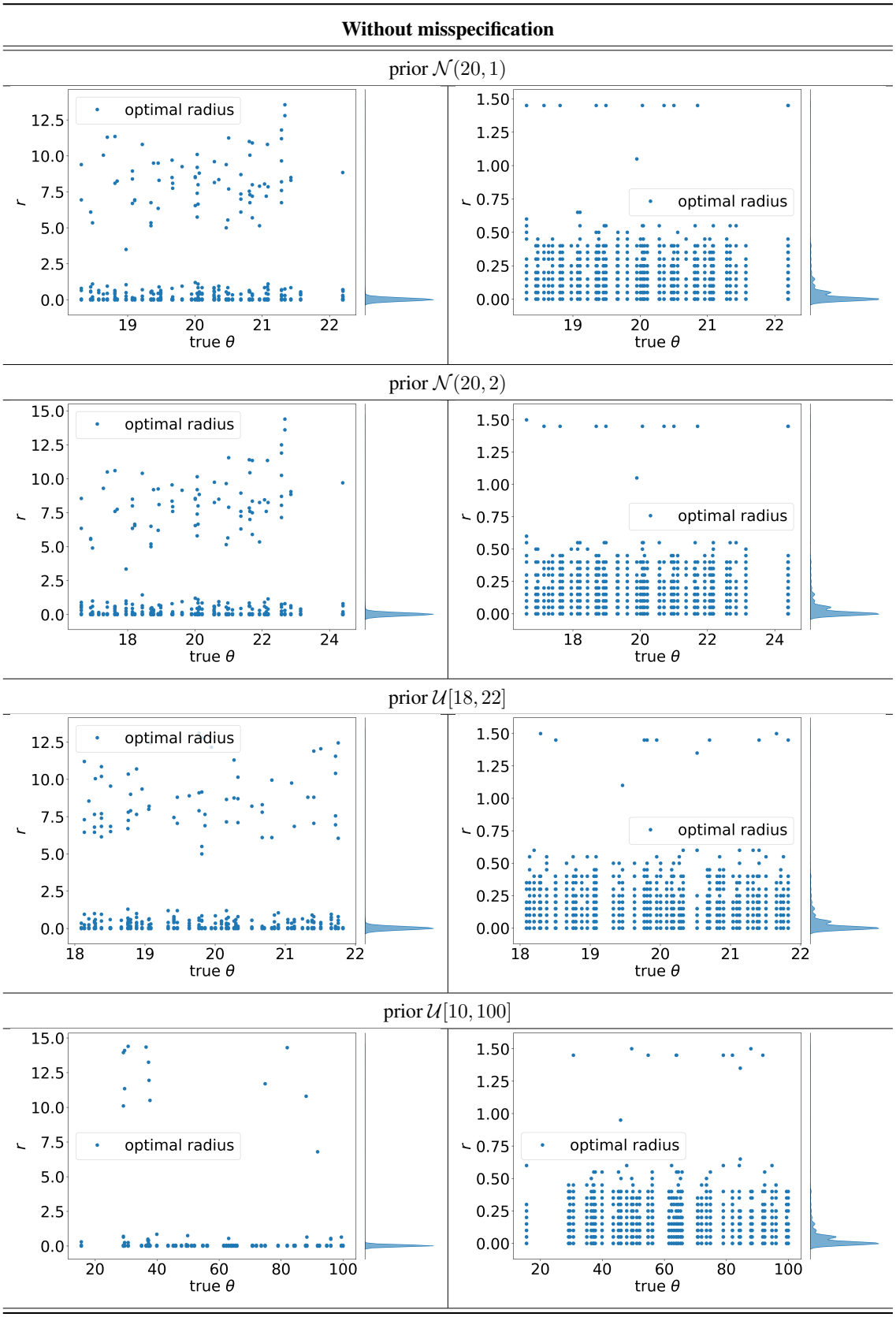


Figure 10 Cross validation for Wasserstein (left) and KL (right) radii under correct model specification

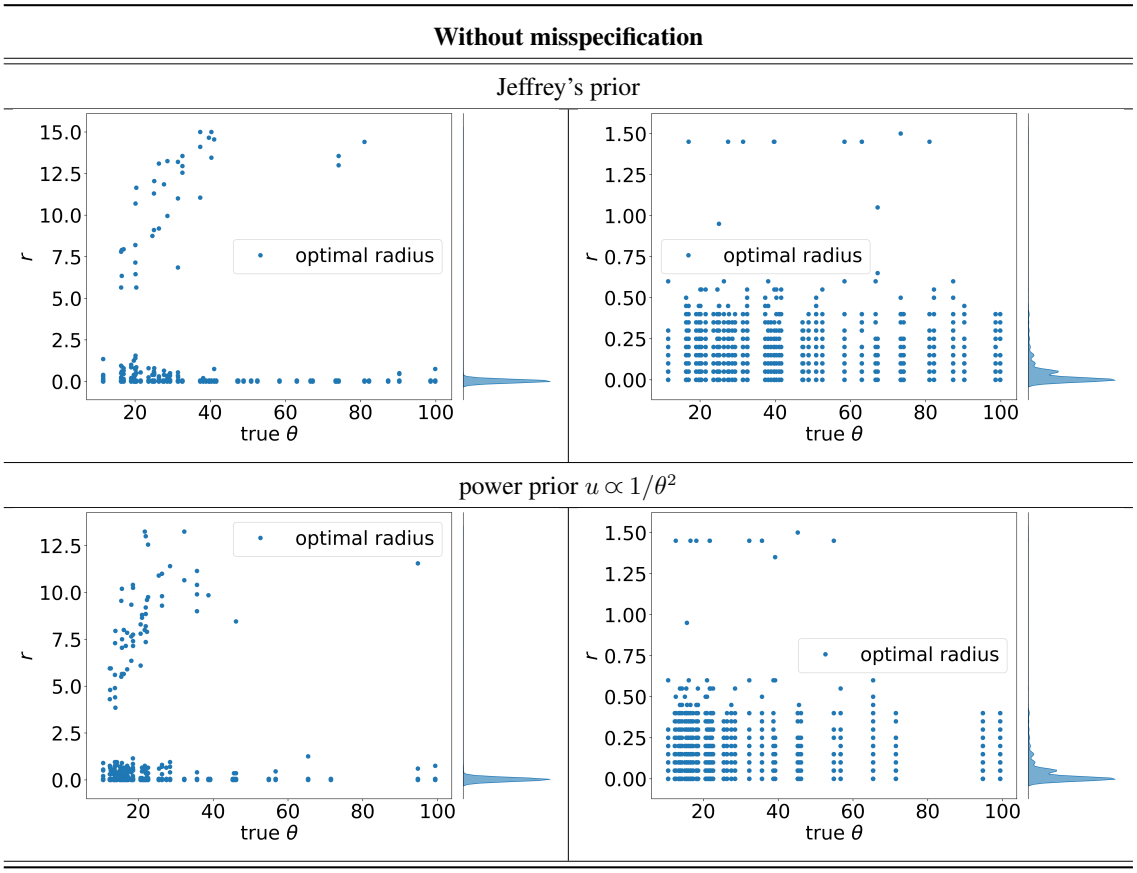


Figure 10 Cross validation for Wasserstein (left) and KL (right) radii under correct model specification (continued)

B.4. Stability to different priors

In Figure 12, we consider priors $u \sim \mathcal{N}(20, 2)$ (left) and $u \sim U[18, 22]$ (right). The same trends as what we have seen before hold, namely, that (i) OVE has the best performance and achieves almost zero regret against the oracle, (ii) the models assuming exponential demand perform better, and (iii) robust models are able to correct for prediction error.

In Figure 13, we compare the OVE solutions obtained over the three different priors. OVE solutions for different priors are pareto optimal. For example, when comparing the performance for priors $u \sim \mathcal{N}(20, 1)$ and $u \sim \mathcal{U}[18, 22]$, we see that there are regions of true parameter θ where one performs better than the other and vice versa. In other words, we cannot tell *a priori* which would perform better at any given θ in the intersection of their supports, even if on the overall we do see that the average regret incurred by the prior $u \sim \mathcal{U}[18, 22]$ is higher.

However, the wider the variance of the prior u, *i.e.*, the range of θ OVE accommodates, the poorer it performs on each specific θ, *e.g.*, $u \sim \mathcal{N}(20, 2)$ leads to an overall worse regret than $u \sim \mathcal{N}(20, 1)$. Moreover, in the interior regions of high density of $u \sim \mathcal{N}(20, 1)$, we can see that the performance of the solution generated by $u \sim \mathcal{N}(20, 2)$ is also poorer. Nonetheless, it might have better validity outside of this range. Already, we can see from the slope of the performance curve for $u \sim \mathcal{N}(20, 1)$ that it will likely

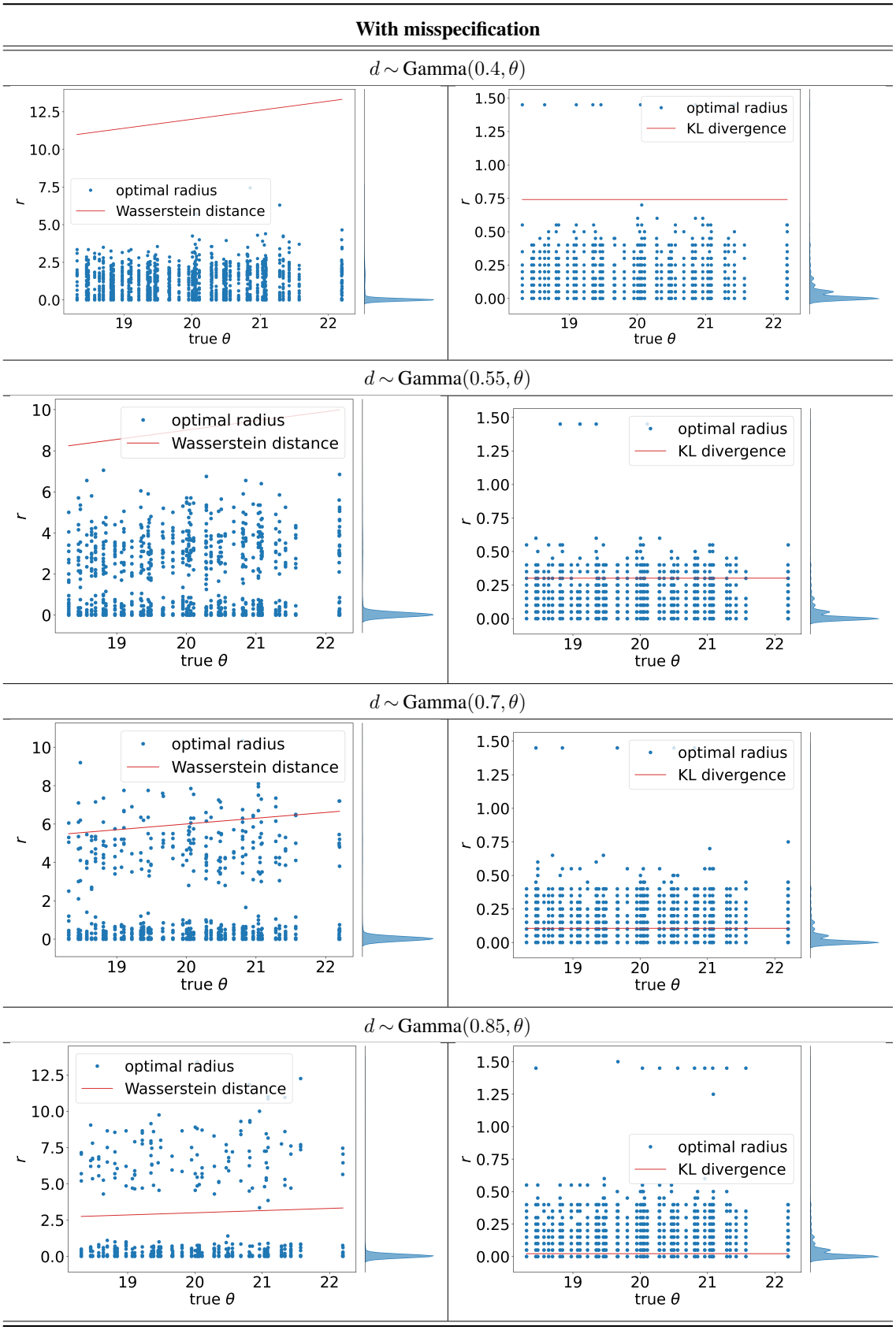


Figure 11 Cross validation for Wasserstein (left) and KL (right) radii under model misspecification

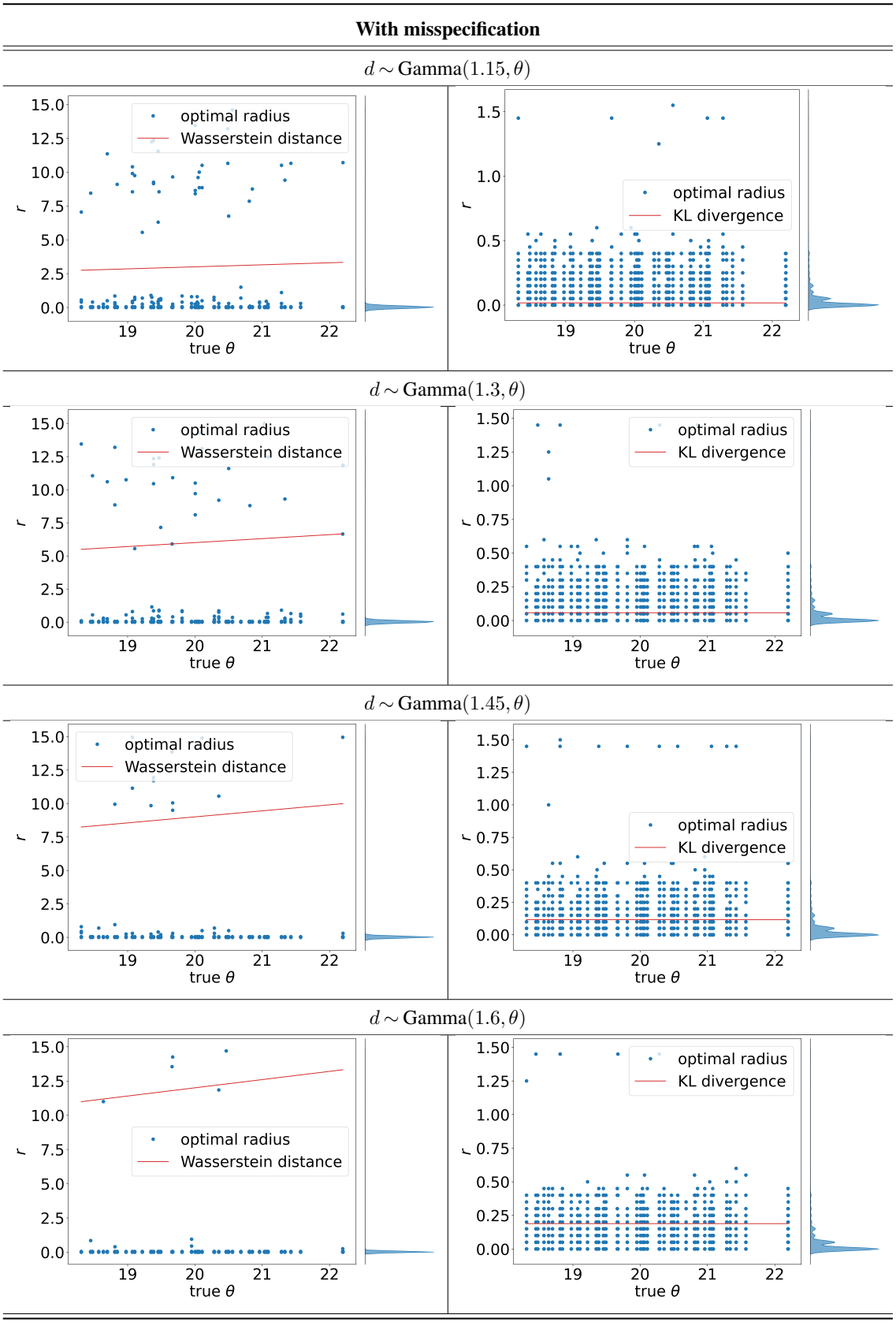


Figure 11 Cross validation for Wasserstein (left) and KL (right) radii under model misspecification (continued)

perform poorer than its wider variance counterpart for the regions of true θ that are further out. This is the specificity-sensitivity trade-off.

B.5. Uninformative priors

In Figure 14, we plot the ungeneralized profit and percentage regret of the various models evaluated (and with OVE computed) over the Jeffrey's prior and the uniform prior over the interval $[10, 100]$.

It is interesting to note that while the average performance of OVE is evidently better than the benchmarks, it does not uniformly dominate the models that assume exponential demand. In particular, its performance over smaller values of θ tend to be poorer. It is not certain why this might particularly be the case, though because of the large range of values of θ that need to be considered, there may be greater numerical instability that is creeping in via the g_1 term. In fact, the smooth shape of both the profit and regret function for the OVE solution that we have seen in earlier examples no longer appears.

B.6. Sensitivity analysis

In this segment, we investigate the stability of our results with respect to different parameter settings of N, the number of data points in the training sample, and p/c, the ratio of the selling to cost price. The average regret is summarized in Table 3. Here, we have fixed the rest of the parameters at the most initial setting, namely, the prior is fixed at $u \sim \mathcal{N}(20, 1)$.

Table 3 Average regret of different models under varying data size N and price to cost ratio p/c

	OVE	Exponential demand			Distribution free			
		ETO	RO	OS	SAA	DRO-M	DRO-KL	DRO-W
Stability tests on data size N								
$N = 5$	0.0118	0.912	0.864	0.837	1.86	2.26	0.805	1.57
$N = 10$	0.0116	0.473	0.453	0.452	1.20	1.60	0.487	0.717
$N = 20$	0.0113	0.244	0.238	0.237	0.579	1.25	0.266	0.429
$N = 30$	0.0111	0.160	0.159	0.157	0.367	1.12	0.182	0.293
Stability tests on price to cost ratio p/c								
$p/c = 1.5$	0.00396	0.164	0.155	0.153	0.559	2.87	0.175	0.447
$p/c = 2$	0.0116	0.473	0.453	0.452	1.20	1.60	0.487	0.717
$p/c = 3$	0.0291	1.18	1.15	1.15	1.79	1.08	0.996	1.79
$p/c = 4$	0.0463	1.87	1.85	1.85	2.72	2.18	1.41	2.72

We can see that the OVE solution continues to dominate the rest of the other models, and by the same margins as before, affirming the stability of our model with respect to changes in the parameters. As the number of data points N increases, the accuracy of the predicted estimator $\hat{\theta}$ increases and all of the models move closer to the perfect information oracle. In the same sense, when price to cost ratio p/c is larger,

the critical quantile is located at a more extreme part of the distribution, and thus with the same number of data points, is estimated with increasing uncertainty. Thus, the model regrets increase. Here, it is much clearer that the RO solution (being also linear in the estimate $\hat{\theta}$) is dominated by OS, though the difference is incredibly small.

Here, we present detailed ungeneralized performance plots for different relative price to cost settings (Figure 16), and for different data size (Figure 15).

B.7. Misspecification

Here we provide further discussion on the experiments on model misspecification. We consider the situation where the truth is Gamma(α, θ), but the exponential model Exp(θ) is assumed. When specifying the prior u, it is critical to realize that the true demand model has mean $\alpha\theta$, whereas the assumed distribution has mean θ. Correspondingly, the model would be mistaking $\alpha\hat{\theta}$ for $\hat{\theta}$, since the sufficient statistic in this case is the sample mean. Thus, if one assumes a prior for the Exponential distribution over the same range of θ that the Gamma distribution is sampled from, then the support of the prior would not even include a large proportion of very likely sample means $\hat{\theta}$ that can be sampled. This problem applies only to the OVE solution as none of the other models use any information regarding the potential range of θ.

To address this problem, we will use a different distribution for sampling the parameter θ for the Gamma distribution (denoted u_G) and the prior under which the OVE solution is solved (denoted u). Here, we take $u_G \sim \mathcal{N}(20, 1)$ and we use $u \sim \mathcal{N}(20\alpha, 1)$, to reflect the ability of the decision-maker to specify a prior for OVE in a way that is consistent with the data generated.

In Table 4, we present detailed results for the average profit and regret, and their extreme 10^{th} percentile values that are obtained from each experiment on different α's. More detailed plots of the ungeneralized profits and regrets are shown in Figures 17 and 18. They clearly show the migration of data-driven methods to a lower regret percentile, with the except of the DRO-Moments model, which consistently remains at the pre-misspecified level, for reasons as previously discussed, *i.e.*, that its extremal distribution remains disparate from the Gamma distribution.

C. Further details of the numerical results for portfolio optimization example

In this Appendix, we provide further details on all numerical experiments performed for the portfolio optimization illustration.

Parameter details: We construct a valid covariance matrix $\mathbf{\Sigma}$ by sampling a random matrix from a standard normal distribution and forming a symmetric matrix through its product with its transpose. To ensure numerical stability and maintain reasonable magnitudes, the matrix is scaled by a factor of 0.1. Diagonal

Distribution	Metric	**Average Profit**								Metric	**Regret**							
		OVE	ETO	OS	RO	DRO-M	DRO-W	DRO-KL	SAA		OVE	ETO	OS	RO	DRO-M	DRO-W	DRO-KL	SAA
Gamma(0.4,θ)	Average	0.446	0.149	0.251	0.233	-0.813	0.564	0.684	0.344	Average	0.443	0.812	0.684	0.707	2.022	0.29	0.139	0.567
	10^{th}%-tile	0.338	0.076	0.182	0.163	-0.973	0.518	0.643	0.28	90^{th}%-tile	0.542	0.904	0.762	0.788	2.202	0.32	0.154	0.643
Gamma(0.55,θ)	Average	1.61	1.265	1.361	1.344	0.195	1.465	1.6	1.176	Average	0.106	0.296	0.242	0.252	0.892	0.184	0.109	0.346
	10^{th}%-tile	1.423	1.145	1.246	1.228	-0.021	1.352	1.499	1.057	90^{th}%-tile	0.151	0.34	0.279	0.29	1.012	0.206	0.123	0.4
Gamma(0.7,θ)	Average	2.986	2.601	2.681	2.667	1.469	2.628	2.772	2.258	Average	0.027	0.152	0.126	0.13	0.521	0.143	0.096	0.264
	10^{th}%-tile	2.72	2.416	2.499	2.486	1.186	2.454	2.598	2.078	90^{th}%-tile	0.05	0.175	0.144	0.149	0.599	0.16	0.107	0.299
Gamma(0.85,θ)	Average	4.507	4.088	4.142	4.134	2.944	3.949	4.136	3.524	Average	0.006	0.098	0.086	0.088	0.35	0.129	0.088	0.223
	10^{th}%-tile	4.167	3.833	3.89	3.882	2.687	3.673	3.844	3.262	90^{th}%-tile	0.015	0.113	0.098	0.1	0.386	0.14	0.099	0.248
Gamma(1,θ)	Average	6.132	5.671	5.692	5.691	4.548	5.427	5.657	4.941	Average	0.002	0.077	0.074	0.074	0.26	0.117	0.079	0.196
	10^{th}%-tile	5.716	5.292	5.302	5.303	4.234	5.063	5.263	4.593	90^{th}%-tile	0.004	0.089	0.085	0.086	0.287	0.132	0.09	0.223
Gamma(1.15,θ)	Average	7.837	7.346	7.325	7.333	6.309	7.063	7.327	6.521	Average	0.004	0.066	0.069	0.068	0.198	0.102	0.069	0.171
	10^{th}%-tile	7.344	6.866	6.867	6.873	5.88	6.572	6.864	6.079	90^{th}%-tile	0.01	0.075	0.078	0.077	0.223	0.116	0.079	0.19
Gamma(1.3,θ)	Average	9.605	9.098	9.033	9.049	8.131	8.817	9.097	8.235	Average	0.008	0.061	0.067	0.066	0.161	0.089	0.061	0.15
	10^{th}%-tile	9.037	8.45	8.399	8.411	7.595	8.246	8.527	7.706	90^{th}%-tile	0.017	0.067	0.075	0.073	0.191	0.101	0.069	0.171
Gamma(1.45,θ)	Average	11.422	10.884	10.772	10.798	9.991	10.639	10.942	9.985	Average	0.013	0.06	0.069	0.067	0.137	0.081	0.055	0.137
	10^{th}%-tile	10.781	10.143	10.06	10.078	9.305	9.933	10.249	9.302	90^{th}%-tile	0.025	0.066	0.076	0.074	0.16	0.09	0.063	0.158
Gamma(1.6,θ)	Average	13.281	12.732	12.57	12.607	11.926	12.563	12.875	11.838	Average	0.018	0.059	0.071	0.068	0.119	0.072	0.049	0.125
	10^{th}%-tile	12.568	11.854	11.729	11.758	11.001	11.741	12.05	11.039	90^{th}%-tile	0.032	0.065	0.078	0.075	0.136	0.081	0.059	0.143

Table 4: Average and 10^{th}%-tile of average profit; and average and 90^{th}%-tile of regret percentile of different models

dominance is enforced by incrementing the diagonal elements by 0.01. Finally, we set $\boldsymbol{V} = \boldsymbol{\Sigma}$ in the experiment. The radii for the ellipsoidal uncertainty set, as well as the radii for the Wasserstein and KL divergence ambiguity sets, were determined through a grid search for the optimal radius, conducted over the range $\{0, 0.05, 0.1, \ldots, 1.95, 2\}$, while the radii for the budgeted uncertainty set are determined over the range $\{0, 0.05, 0.1, \ldots, 14.95, 15\}$. Further details on calibration of radius of uncertainty are provided in Figure 19.

C.1. Mathematical formulations for benchmark models

Here we clearly state the formulations for the benchmark models used in our numerical experiments, specific to the case of the portfolio optimization problems.

Robust optimization model with $\boldsymbol{\theta}$ as the uncertainty: The robust portfolio optimization problem can be expressed as

$$\min_{\boldsymbol{x}} \sup_{\boldsymbol{\theta}\in\Omega} \{\phi(\boldsymbol{x}, \boldsymbol{\theta})\} = \min_{\boldsymbol{x}} \sup_{\boldsymbol{\theta}\in\Omega} \left\{ -(1 - \mathbf{1}^\top \boldsymbol{x}) r_f - \boldsymbol{\theta}^\top \boldsymbol{x} + \frac{1}{2\tau} \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \right\}.$$

We consider the following uncertainty sets, which are ubiquitous in RO models.

- **Box Uncertainty Set.** Consider the box uncertainty set, in which Ω is given by

$$\Omega = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \;\middle|\; \begin{array}{l} \theta_i = \hat{\theta}_i + \xi_i \Delta\theta_i \;\forall i = 1, \cdots, d \\ \|\boldsymbol{\xi}\|_\infty \leq 1 \end{array} \right\},$$

  where $\Delta\theta_i = 0.05|\hat{\theta}_i|$. The robust portfolio optimization problem under this uncertainty set can be reformulated as

$$\min_{\boldsymbol{x}} \left\{ -(1 - \mathbf{1}^\top \boldsymbol{x}) r_f - \hat{\boldsymbol{\theta}}^\top \boldsymbol{x} + \sum_{i=1}^{d} |x_i \Delta\theta_i| + \frac{1}{2\tau} \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \right\}.$$

- **Ellipsoidal Uncertainty Set.** Consider the ellipsoidal uncertainty, in which Ω is given by

$$\Omega = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \;\middle|\; \begin{array}{l} \theta_i = \hat{\theta}_i + \xi_i \Delta\theta_i \;\forall i = 1, \cdots, d \\ \|\boldsymbol{\xi}\|_2 \leq r \end{array} \right\},$$

  where $\Gamma \geq 0$. The robust portfolio optimization problem under this uncertainty set can be reformulated as

$$\min_{\boldsymbol{x}} \left\{ -(1 - \mathbf{1}^\top \boldsymbol{x}) r_f - \hat{\boldsymbol{\theta}}^\top \boldsymbol{x} + r \sqrt{\sum_{i=1}^{d} (x_i \Delta\theta_i)^2} + \frac{1}{2\tau} \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \right\}.$$

- **Budgeted Uncertainty Set.** Consider the budgeted uncertainty set, in which Ω is given by

$$\Omega = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \;\middle|\; \begin{array}{l} \theta_i = \hat{\theta}_i + \xi_i \Delta\theta_i \;\forall i = 1, \cdots, d \\ \|\boldsymbol{\xi}\|_\infty \leq 1,\; \|\boldsymbol{\xi}\|_1 \leq r \end{array} \right\},$$

  where $\Gamma \geq 0$. The robust portfolio optimization problem under this uncertainty set can be reformulated as

$$\min_{\boldsymbol{x}, \boldsymbol{\beta}} \left\{ -(1 - \mathbf{1}^\top \boldsymbol{x}) r_f - \hat{\boldsymbol{\theta}}^\top \boldsymbol{x} + \sum_{i=1}^{d} |x_i \Delta\theta_i - \beta_i| + r \|\boldsymbol{\beta}\|_\infty + \frac{1}{2\tau} \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \right\}.$$

Distributionally robust optimization model with moment uncertainty: The distributionally robust optimization formulation for the robust portfolio optimization problem can be written as:

$$\min_{\boldsymbol{x}} \sup_{\mathbb{P}\in\mathcal{P}_m} \mathbb{E}_{\mathbb{P}\sim\boldsymbol{z}}\left[-(1-\mathbf{1}^\top\boldsymbol{x})r_f - \boldsymbol{z}^\top\boldsymbol{x} + \frac{1}{2\tau}\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}\right],$$

where the ambiguity $\mathcal{P}_m$ is given by

$$\mathcal{P}_m = \left\{ \mathbb{P}\in\mathcal{P}(\mathbb{R}^d) \;\middle|\; \begin{array}{l} \mathbb{E}_{\mathbb{P}}[\boldsymbol{z}]\in\Omega \\ \mathbb{E}_{\mathbb{P}}\left[(\boldsymbol{z}-\mathbb{E}_{\mathbb{P}}[\boldsymbol{z}])(\boldsymbol{z}-\mathbb{E}_{\mathbb{P}}[\boldsymbol{z}])^\top\right]=\boldsymbol{\Sigma} \end{array} \right\}.$$

Note that the corresponding robust counterpart reduces to the robust optimization problem we have mentioned before.

Distributionally robust optimization model with Wasserstein uncertainty set: We instead consider the ambiguity set $\mathcal{P}_W^r$ in the distributionally robust portfolio optimization problem,

$$\mathcal{P}_W^r = \left\{ \mathbb{P}\in\mathcal{P}(\mathbb{R}^d) \;\middle|\; \begin{array}{l} \boldsymbol{z}\sim\mathbb{P} \\ \Delta_W(\mathbb{P},\hat{\mathbb{P}})\le r \end{array} \right\},$$

where $\Delta_W(\mathbb{P},\hat{\mathbb{P}})$ is the Type-1 Wasserstein distance defined on some norm $||\cdot||$, and the empirical distribution $\hat{\mathbb{P}}$ is given by $\hat{\mathbb{P}}[\boldsymbol{z}=\boldsymbol{y}_n]=1/N$, for all $n=1,\ldots,N$. Its robust counterpart is

$$\begin{array}{ll} \inf & -(1-\mathbf{1}^\top\boldsymbol{x})r_f + \lambda r + \dfrac{1}{N}\displaystyle\sum_{n=1}^N \beta_n + \dfrac{1}{2\tau}\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x} \\ \text{s.t.} & \lambda\|\boldsymbol{z}-\boldsymbol{y}_n\| + \beta_n + \boldsymbol{z}^\top\boldsymbol{x}\ge 0, \qquad \forall \boldsymbol{z}, n=1,\cdots,N \\ & \lambda\ge 0,\ \boldsymbol{\beta} \text{ free}. \end{array}$$

Distributionally robust optimization model with Kullback-Leibler uncertainty set: We instead consider the ambiguity set $\mathcal{P}_{KL}^r$ in the distributionally robust portfolio optimization problem,

$$\mathcal{P}_{KL}^r = \left\{ \mathbb{P}\in\mathcal{P}(\mathbb{R}^d) \;\middle|\; \begin{array}{l} \boldsymbol{z}\sim\mathbb{P} \\ \Delta_{KL}(\mathbb{P},\hat{\mathbb{P}})\le r \end{array} \right\},$$

where $\Delta_{KL}(\mathbb{P},\hat{\mathbb{P}})$ is the Kullback-Leibler divergence. Let $\hat{p}_n=\hat{\mathbb{P}}[\boldsymbol{z}=\boldsymbol{y}_n]=1/N$, for all $n=1,\ldots,N$. Its robust counterpart is

$$\begin{array}{ll} \inf & -(1-\mathbf{1}^\top\boldsymbol{x})r_f + \lambda r + \displaystyle\sum_{n=1}^N \beta_n\hat{p}_n + \gamma + \dfrac{1}{2\tau}\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x} \\ \text{s.t.} & \lambda\log(\lambda/\beta_n) - \boldsymbol{y}_n^\top\boldsymbol{x}\le\lambda+\gamma, \qquad \forall n=1,\cdots,N \\ & \lambda\ge 0,\ \gamma \text{ free},\ \boldsymbol{\beta}\ge\mathbf{0}. \end{array}$$

C.2. Illustration with a uniform prior

Here, we illustrate the results using a uniform prior in the portfolio optimization setting for 10 stocks. We consider a box prior where each dimension is specified as $\theta_i \sim \mathcal{U}\left[\mu_i - \frac{1}{2}\sqrt{\lambda_{\max} \cdot \frac{\chi^2_{0.9}(10)}{N}},\ \mu_i + \frac{1}{2}\sqrt{\lambda_{\max} \cdot \frac{\chi^2_{0.9}(10)}{N}}\right]$, where μ is the prior mean, $\lambda_{\max}$ is the largest eigenvalue of Σ, and $\chi^2_{0.9}(10)$ denote the 90^{th} percentile of a chi-square distribution with 10 degrees of freedom. As with the normal prior, OVE does seem to dominate the benchmarks on every parameter value in the testing range. The performacnce of each model is summarized in Table 5 and Figure 20, while the calibration of radius of uncertainty is detailed in Figure 21.

Table 5 Average objective and regret in portfolio optimization under a uniform prior

Average objective							
Metrics	OVE	ETO	RO-box	Ro-el	RO-budg	DRO-W	DRO-KL
Average	-74.786	-74.677	-73.080	-74.704	-74.706	-74.701	-74.709
10^{th}%-tile	-159.164	-159.057	-156.900	-159.083	-159.085	-159.079	-159.079
Regret							
Metrics	OVE	ETO	RO-box	Ro-el	RO-budg	DRO-W	DRO-KL
Average	0.961%	1.229%	4.305%	1.161%	1.156%	1.170%	1.142%
90^{th}%-tile	2.044%	2.634%	8.562%	2.501%	2.505%	2.519%	2.442%

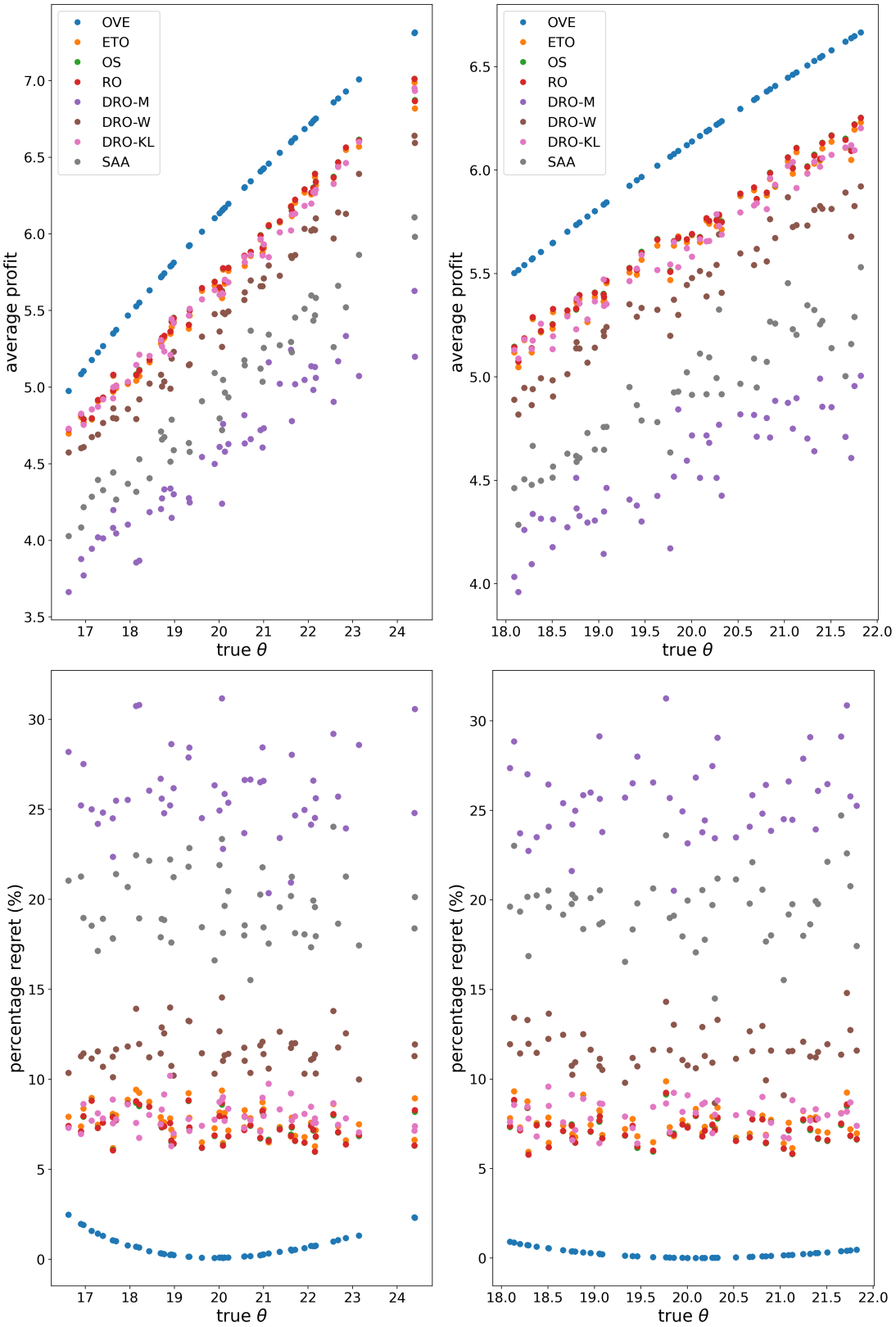


Figure 12 Average profit (top) and ex-ante percentage regret (bottom) of different models over the range of θ in the priors $u \sim \mathcal{N}(20, 2)$ (left) and $u \sim \mathcal{U}[18, 22]$ (right)

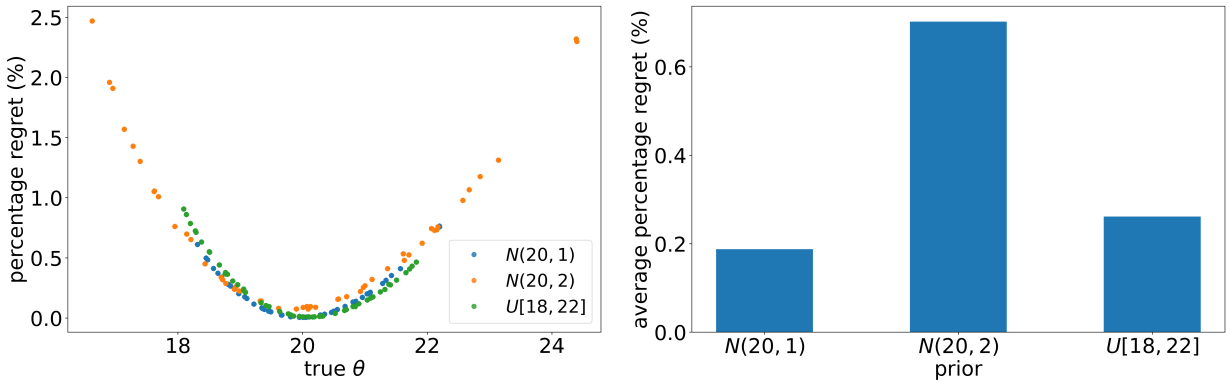


Figure 13 Ungeneralized (left) and averaged (right) ex-ante percentage regret of OVE for three priors

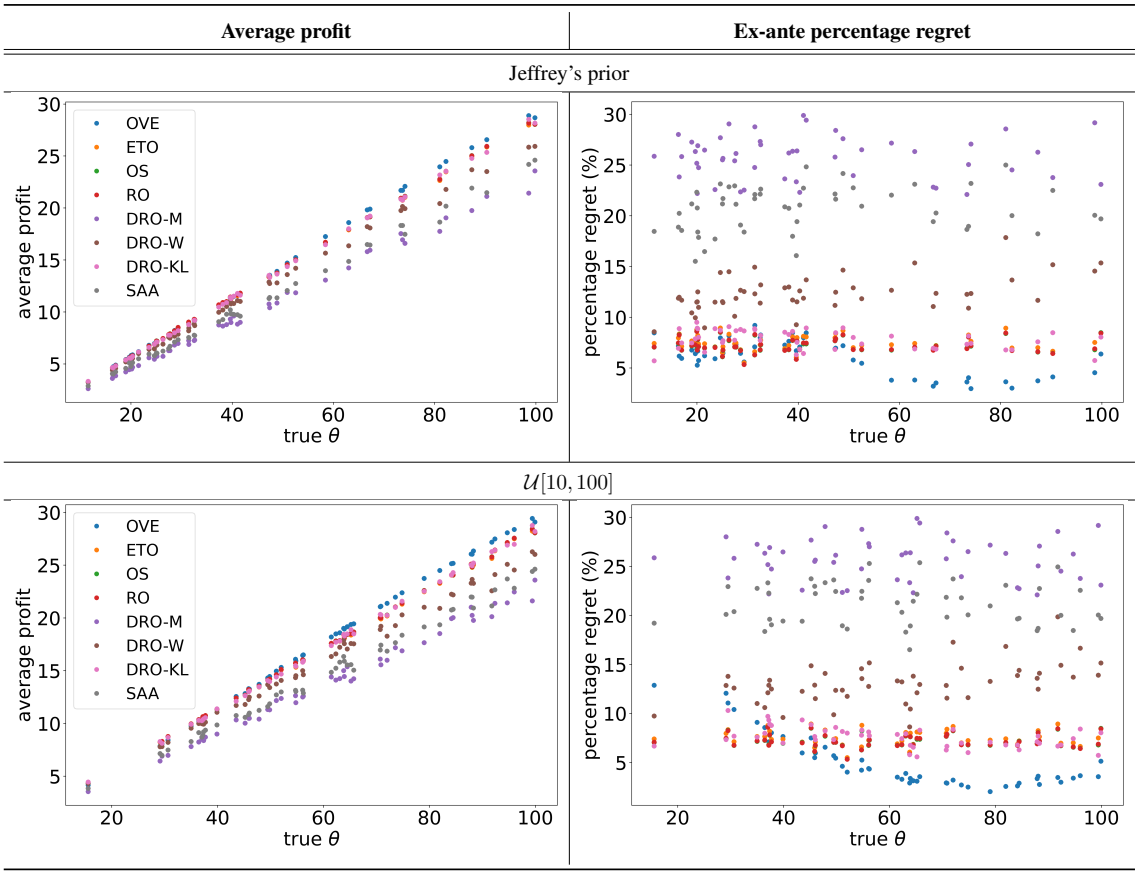


Figure 14 Average profit and ex-ante percentage regret of different models in the Jeffrey's prior and $\mathcal{U}[10, 100]$

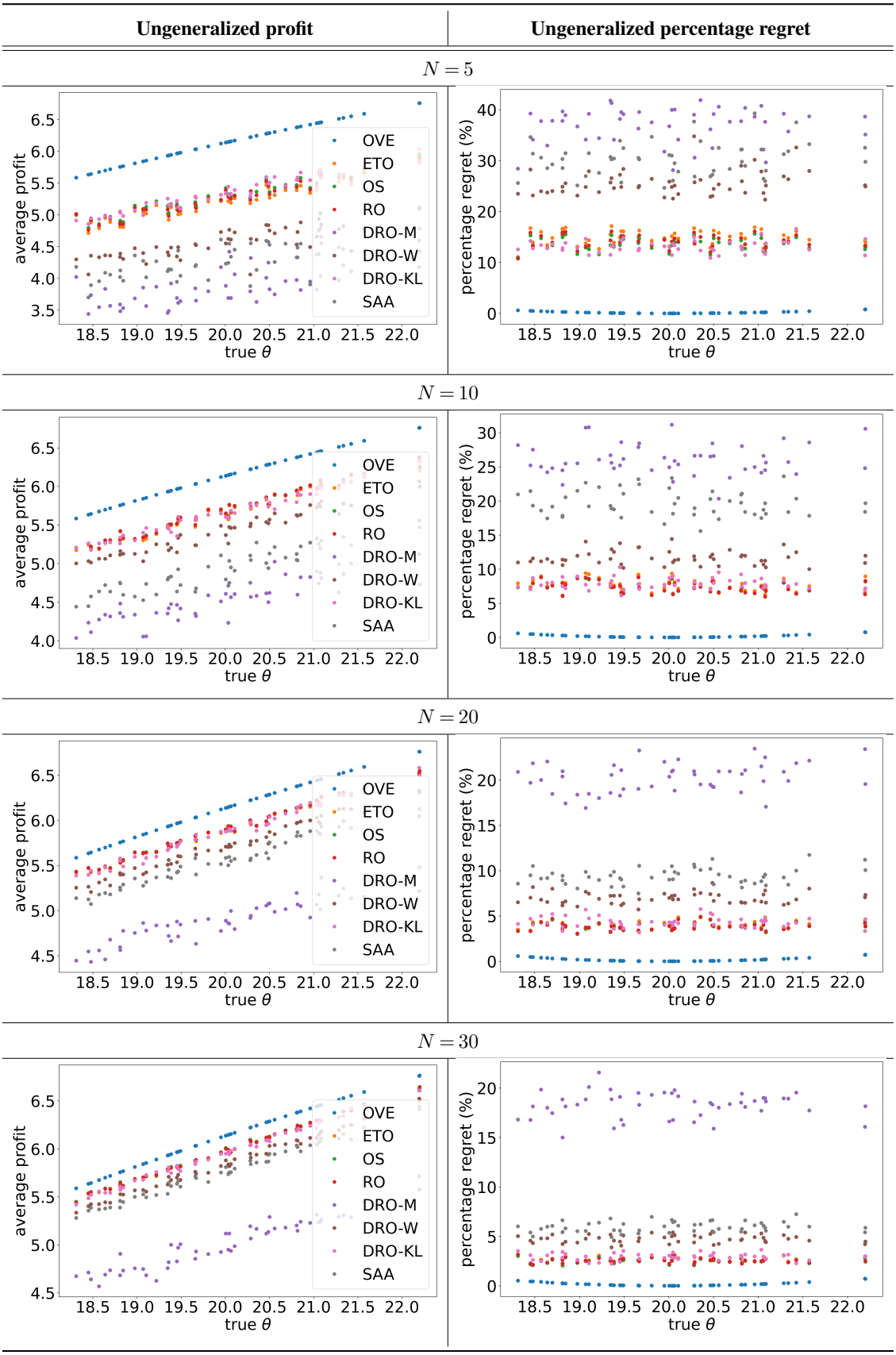


Figure 15 Ungeneralized profit and percentage regret as data size $N = 5, 10, 20, 30$ is varied

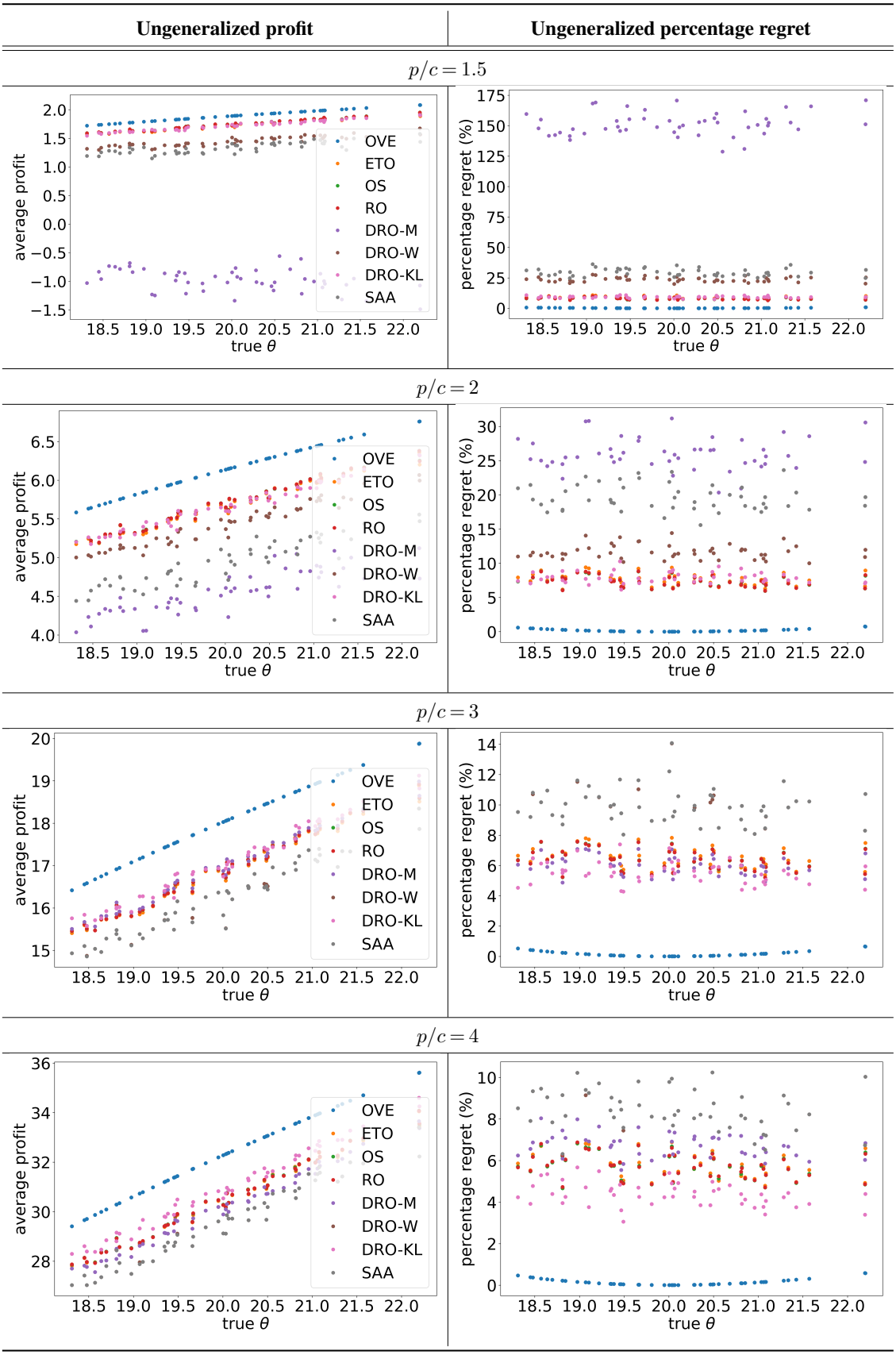


Figure 16 Ungeneralized profit and percentage regret as price to cost ratio $p/c = 1.5, 2, 3, 4$ is varied

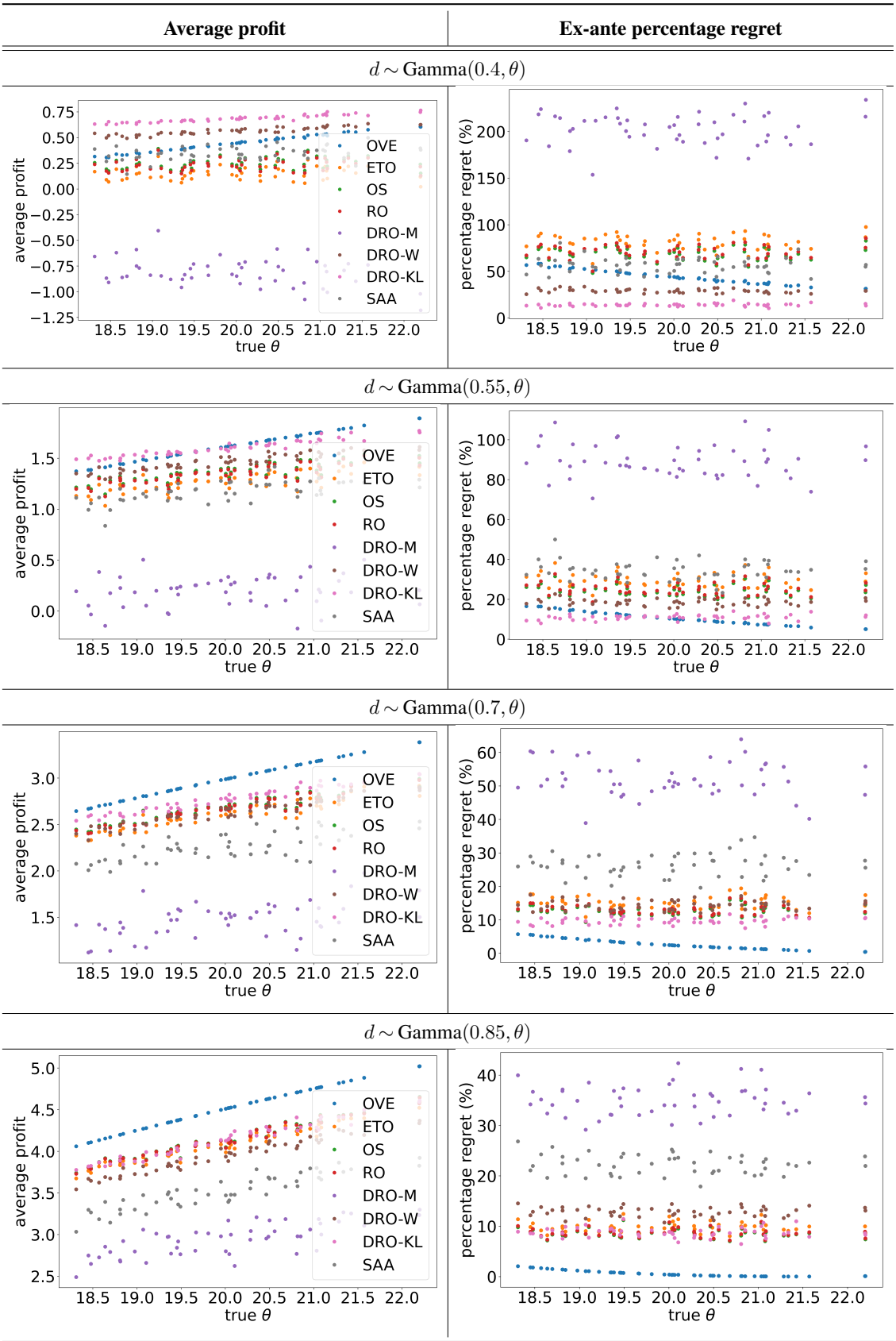


Figure 17 Average profit and ex-ante percentage regret of different models with shape parameters $\{0.4, 0.55, 0.7, 0.85\}$

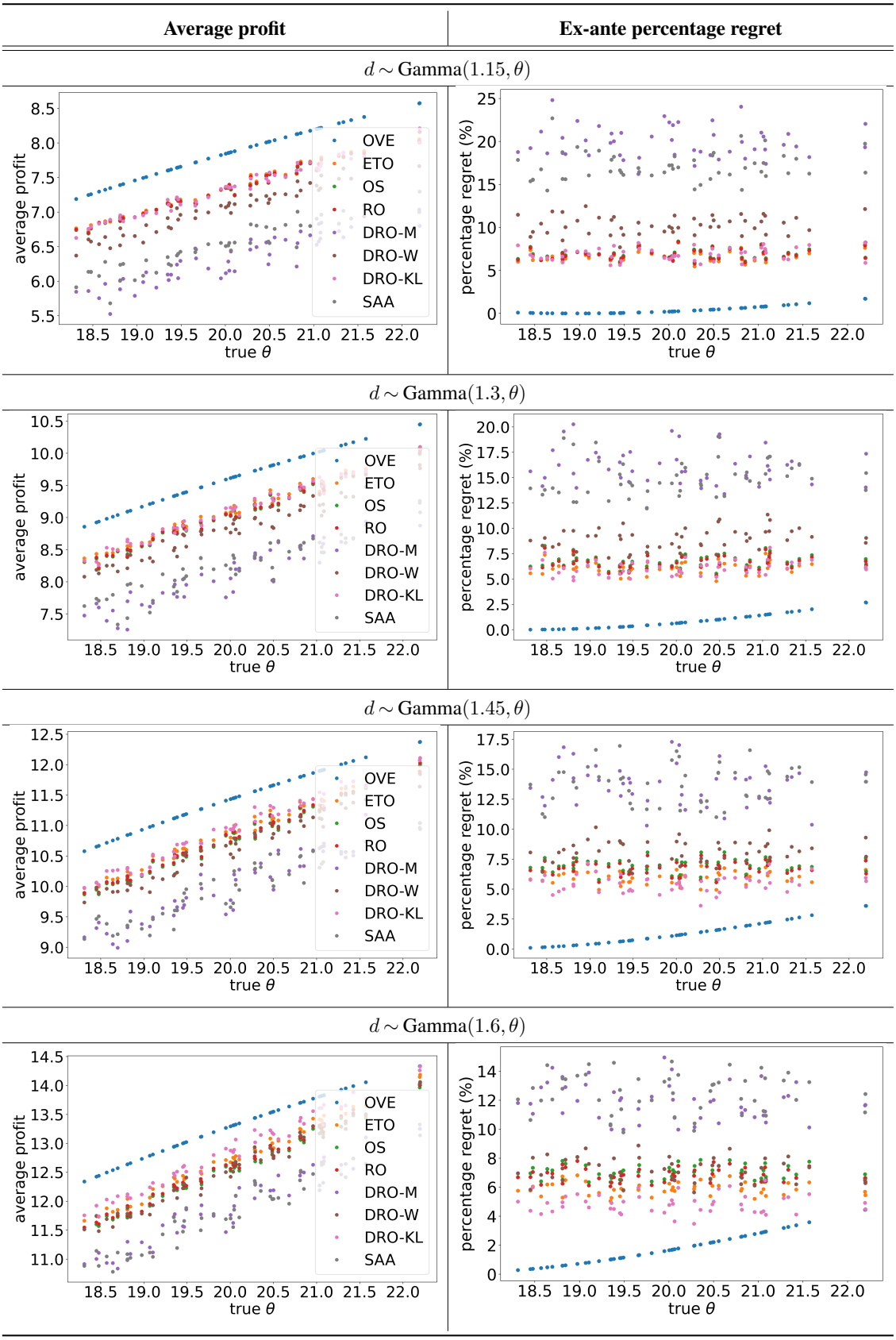


Figure 18 Average profit and ex-ante percentage regret of different models with shape parameters $\{1.15, 1.3, 1.45, 1.6\}$

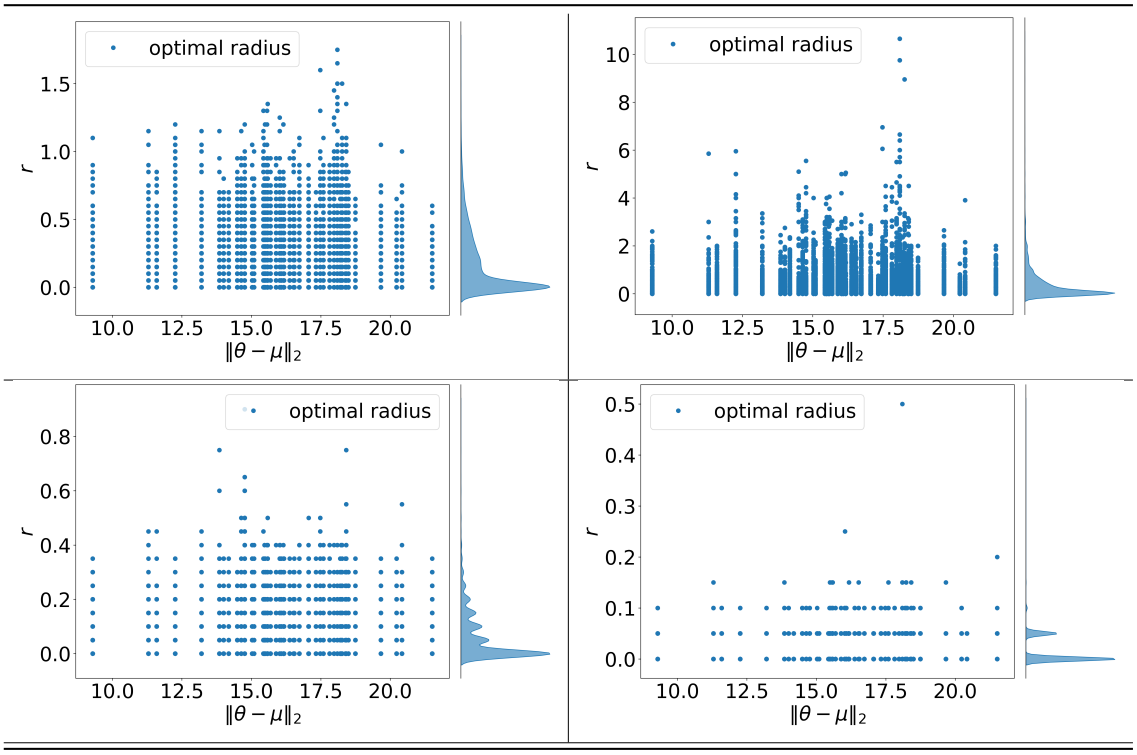


Figure 19 Calibration of radius of uncertainty in portfolio optimization under the prior $u \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{V})$: ellipsoidal (upper left), budgeted (upper right), Wasserstein (lower left), and KL (lower right).

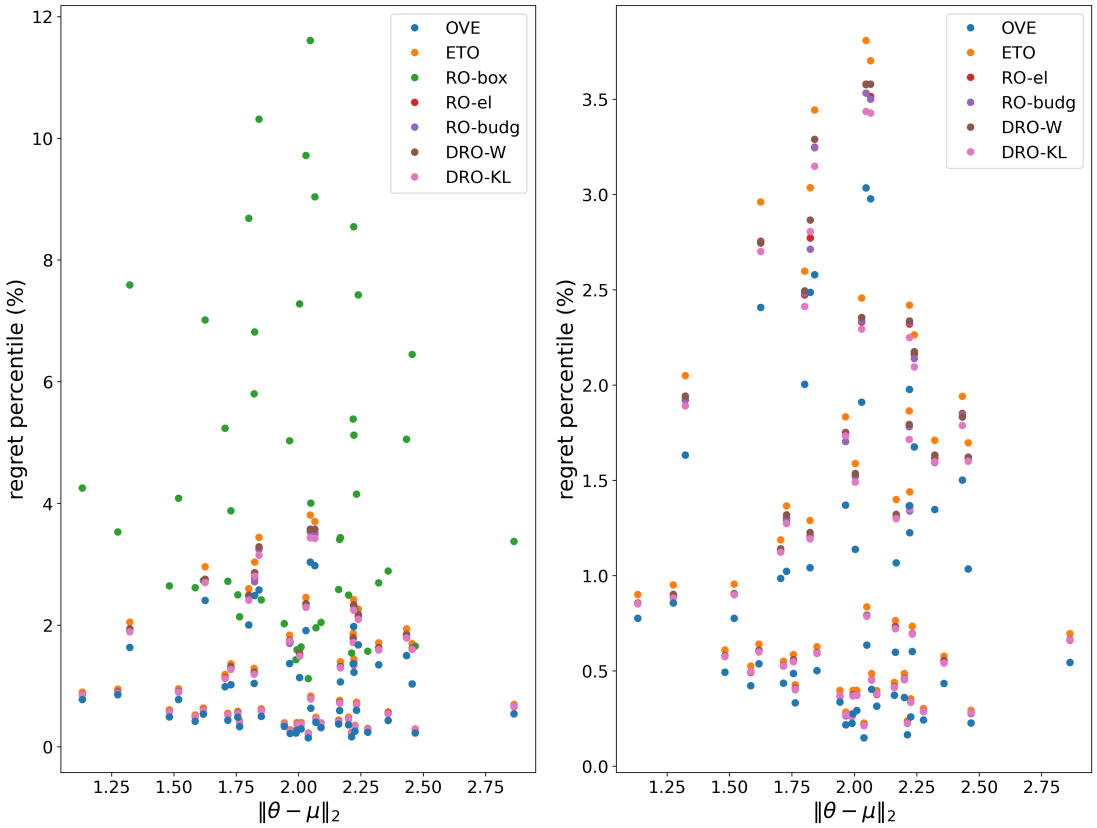


Figure 20 Percentage ex-ante regret of different models over the range of θ under the uniform prior, with the left panel including the RO model with a box uncertainty set (RO-box) and the right panel excluding it.

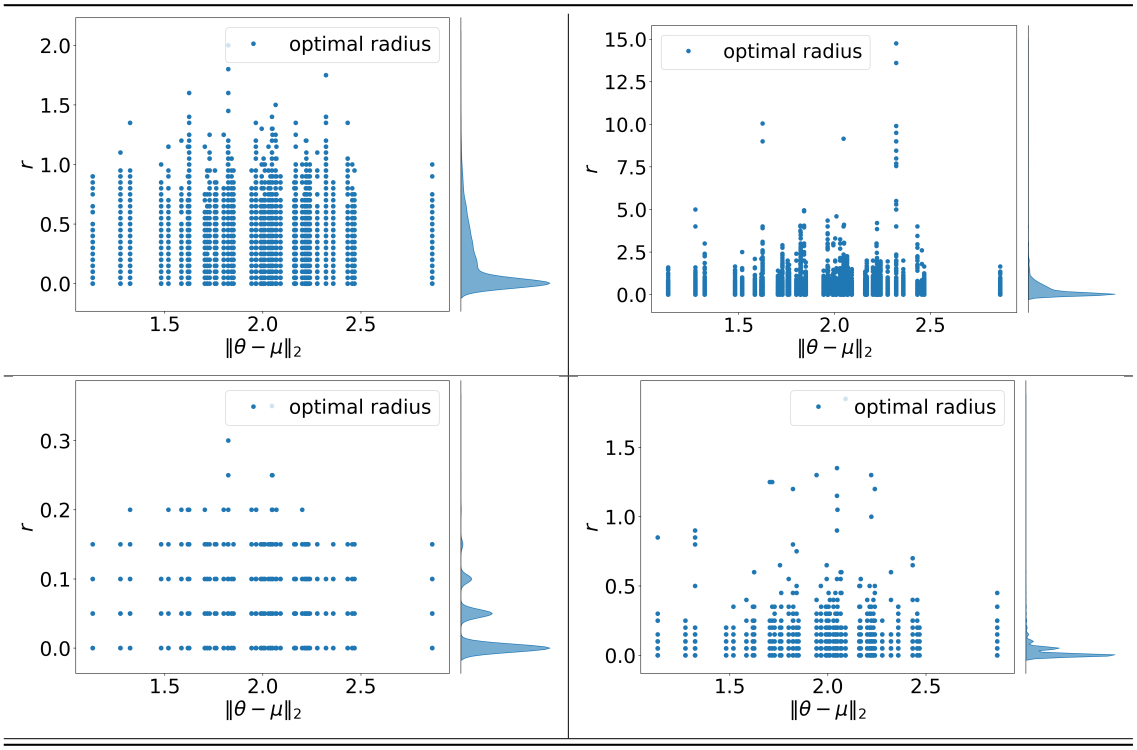


Figure 21 Calibration of radius of uncertainty in portfolio optimization under the uniform prior: ellipsoidal (upper left), budgeted (upper right), Wasserstein (lower left), and KL (lower right).