Online Supplement of "Decision-Driven Regularization: A Blended Model for Learning and Optimization"

EC1. Omitted Proofs

In this segment, we present all deferred proofs from the main text.

EC1.1. Proof of Claim 1 in Illustration 1

The least squares minimization problem for OLS has the solution: $\boldsymbol{W}^{=}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{Z})$, where $\boldsymbol{X} = \left[\tilde{\boldsymbol{x}}_1^\top\ \tilde{\boldsymbol{x}}_2^\top\ \tilde{\boldsymbol{x}}_3^\top\ \tilde{\boldsymbol{x}}_4^\top\right]^\top$ and $\boldsymbol{Z} = (\tilde{\boldsymbol{z}}_1^\top, \tilde{\boldsymbol{z}}_2^\top, \tilde{\boldsymbol{z}}_3^\top, \tilde{\boldsymbol{z}}_4^\top)^\top$. It is easy to verify that these equations return $\hat{\boldsymbol{W}}^{\text{SLO}} = \check{\boldsymbol{W}} = \begin{bmatrix} 1 & 0 \\ 0 & 1.5 \end{bmatrix}$. Hence, SLO picks route b if and only if $1.5\hat{x}_2 \le \hat{x}_1$.

To verify the conclusion for ILO, note that for any given feature vector $\hat{\boldsymbol{x}}$ and coefficient matrix $\boldsymbol{W}$, route b is chosen if and only if $(\boldsymbol{W}\hat{\boldsymbol{x}})_1 - (\boldsymbol{W}\hat{\boldsymbol{x}})_2 = w_{11}\hat{x}_1 + w_{12}\hat{x}_2 - w_{21}\hat{x}_1 - w_{22}\hat{x}_2 \ge 0$. For convenience, let $\bar{w} := w_{11} - w_{21}$ and $\underline{w} := w_{22} - w_{12}$. Hence, the optimal decision $\boldsymbol{y}^\star$ will pick route b if and only if $\underline{w}\hat{x}_2 \le \bar{w}\hat{x}_1$.

For ILO, in order to get the minimum objective, ILO chooses $\hat{\boldsymbol{W}}^{\text{ILO}}$ such that the smaller of the two components of $\tilde{\boldsymbol{z}}_n$ is chosen. Specifically, to minimize ILO loss, we need to pick route b for data points 1, 3 and 4, but pick route a for data point 2. Hence, optimal $\hat{\boldsymbol{W}}^{\text{ILO}}$ is attained when

$$\begin{aligned} \bar{w} - \underline{w} &\ge 0, \\ 2\bar{w} - \underline{w} &\ge 0, \\ \bar{w} - 2\underline{w} &\le 0, \\ 3\bar{w} - 2\underline{w} &\ge 0. \end{aligned}$$

Subtracting the LHS of the first inequality from the third one, we have $\bar{w} - \underline{w} - (\bar{w} - 2\underline{w}) \ge 0$, i.e., $\underline{w} \ge 0$. Then the above four inequalities are equivalent to the condition $1 \le \bar{w}/\underline{w} \le 2$. In other words, any matrix $\hat{\boldsymbol{W}}^{\text{ILO}}$ satisfying $1 \le \bar{w}/\underline{w} \le 2$, $\underline{w} \ge 0$ is optimal for the ILO loss function. Thus, ILO chooses route b if and only if $\hat{x}_2 \le (\bar{w}/\underline{w})\hat{x}_1$ for some $1 \le \bar{w}/\underline{w} \le 2$ that is degenerate and indifferentiable in-sample under the ILO loss. □

EC1.2. Proof of Proposition 1

$$\begin{aligned} &\mathbb{E}_{\mathcal{D}^N}\left[\left(\boldsymbol{y}^\top\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}[\boldsymbol{z}] - \boldsymbol{y}^\top\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{w}})\right)^2\right] \\ =&\mathbb{E}_{\mathcal{D}^N}\left[\left(\boldsymbol{y}^\top g(\boldsymbol{x}) - \boldsymbol{y}^\top\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{w}})\right)^2\right] \\ =&\mathbb{E}_{\mathcal{D}^N}\left[\boldsymbol{y}^\top g(\boldsymbol{x})\right]^2 - 2\mathbb{E}_{\mathcal{D}^N}\left[\boldsymbol{y}^\top g(\boldsymbol{x})\boldsymbol{y}^\top\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{w}})\right] + \mathbb{E}_{\mathcal{D}^N}\left[(\boldsymbol{y}^\top\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{w}}))^2\right] \\ =&[\boldsymbol{y}^\top g(\boldsymbol{x})]^2 - 2\boldsymbol{y}^\top g(\boldsymbol{x})\mathbb{E}_{\mathcal{D}^N}[\boldsymbol{y}^\top\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{w}})] + \mathbb{E}_{\mathcal{D}^N}\left[(\boldsymbol{y}^\top\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{w}}))^2\right] - (\mathbb{E}_{\mathcal{D}^N}\left[\boldsymbol{y}^\top\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{w}})\right])^2 + (\mathbb{E}_{\mathcal{D}^N}\left[\boldsymbol{y}^\top\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{w}})\right])^2 \\ =&\left[\boldsymbol{y}^\top g(\boldsymbol{x}) - \mathbb{E}_{\mathcal{D}^N}\boldsymbol{y}^\top\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{w}})\right]^2 + \mathbb{E}_{\mathcal{D}^N}\left[(\boldsymbol{y}^\top\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{w}}))^2\right] - (\mathbb{E}_{\mathcal{D}^N}\left[\boldsymbol{y}^\top\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{w}})\right])^2. \end{aligned}$$

where the first equality holds due to $\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}[\boldsymbol{z}] = g(\boldsymbol{x})$, the third equality holds because $\boldsymbol{y}$ and $\boldsymbol{x}$ are deterministic with respect to the data $\mathcal{D}^N$ and the last equality comes from rearranging terms.

EC1.3. Proof of Proposition 2

Let (α_1, α_2) be the dual variables, then the Lagrangian relaxation of Problem (4) is given by:

$$\begin{aligned}&\min_{\boldsymbol{y}\in\mathcal{Y}} \quad \Big\{\boldsymbol{y}^\top \boldsymbol{f}(\tilde{\boldsymbol{x}},\boldsymbol{w}) + \alpha_1\Big(\boldsymbol{y}^\top\tilde{\boldsymbol{z}} - \boldsymbol{y}^\top\boldsymbol{f}(\tilde{\boldsymbol{x}},\boldsymbol{w}) - \gamma\Big) + \alpha_2\Big(-\boldsymbol{y}^\top\tilde{\boldsymbol{z}} + \boldsymbol{y}^\top\boldsymbol{f}(\tilde{\boldsymbol{x}},\boldsymbol{w}) - \gamma\Big)\Big\}\\ =&\min_{\boldsymbol{y}\in\mathcal{Y}} \quad \Big\{(\alpha_1-\alpha_2)\boldsymbol{y}^\top\tilde{\boldsymbol{z}} + (1-\alpha_1+\alpha_2)\boldsymbol{y}^\top\boldsymbol{f}(\tilde{\boldsymbol{x}},\boldsymbol{w}) - \gamma(\alpha_1+\alpha_2)\Big\}.\end{aligned}$$

Setting $\mu = \alpha_1 - \alpha_2$, we recover the statement of the Proposition 2, where the term associated with γ is dropped as it do not influence the choice of decisions $\boldsymbol{y}$.

EC1.4. Proof of Proposition 3

When $c(\boldsymbol{y};\boldsymbol{z}) = \boldsymbol{y}^\top\boldsymbol{z}$ and $\mathcal{Y} = \{\boldsymbol{A}\boldsymbol{y} \le \boldsymbol{b}, \boldsymbol{y} \ge \boldsymbol{0}\}$, the inner supreme problem of the DDR becomes

$$\begin{aligned}&\max_{\boldsymbol{y}_n\ge 0} \boldsymbol{y}_n^\top\big(-\mu\tilde{\boldsymbol{z}}_n - (1-\mu)\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\big)\\ &\text{s.t. } \boldsymbol{A}\boldsymbol{y}_n \le \boldsymbol{b},\end{aligned}$$

which is a linear programming problem and the dual of it is

$$\begin{aligned}&\min_{\boldsymbol{\kappa}_n\ge 0} \boldsymbol{\kappa}_n^\top\boldsymbol{b}\\ &\text{s.t. } \boldsymbol{A}^\top\boldsymbol{\kappa}_n \ge -\mu\tilde{\boldsymbol{z}}_n - (1-\mu)\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}).\end{aligned}$$

□

EC1.5. Motivation for the loss function uncertainty set

The following illustration gives some explanation about why this geometry for the uncertainty set makes sense.

Illustration 1 [as seen in Zhu et al. (2022)]. *Suppose $L(\boldsymbol{w}) = -2\log\Big(\prod_n p_{\boldsymbol{z}|\boldsymbol{x}}(\boldsymbol{z}_n;\boldsymbol{x}_n,\boldsymbol{w})\Big)$ was chosen as the log-likelihood function, where $p_{\boldsymbol{z}|\boldsymbol{x}}$ is the density of $\boldsymbol{z}|\boldsymbol{x}$. Then the uncertainty set $\mathcal{U}(\rho) := \{\boldsymbol{w} : L(\boldsymbol{w}) - L(\tilde{\boldsymbol{w}}) \le \rho\}$ reduces to the set of all weights $\boldsymbol{w}$, which likelihood ratio against $\tilde{\boldsymbol{w}}$ is no greater than some bound:*

$$\mathcal{U}(\rho) := \left\{\boldsymbol{w} : LR(\tilde{\boldsymbol{w}};\boldsymbol{w}) := \frac{\prod_n p_{\boldsymbol{z}|\boldsymbol{x}}(\boldsymbol{z}_n;\boldsymbol{x}_n,\tilde{\boldsymbol{w}})}{\prod_n p_{\boldsymbol{z}|\boldsymbol{x}}(\boldsymbol{z}_n;\boldsymbol{x}_n,\boldsymbol{w})} \le e^{\rho/2}\right\}. \tag{EC1.1}$$

Suppose we interpret the null hypothesis as $\mathcal{H}_0 : \tilde{\boldsymbol{w}} = \tilde{\boldsymbol{w}}$ and the alternative hypothesis as $\mathcal{H}_1 : \tilde{\boldsymbol{w}} = \boldsymbol{w} \ne \tilde{\boldsymbol{w}}$. Then Neyman-Pearson Lemma gives that there is some confidence level $\alpha(\rho)$, corresponding to $e^{\rho/2}$, under which, the likelihood ratio test grants the highest power, that is, probability of rejecting $\mathcal{H}_0$. In other words, this uncertainty set is equivalent to considering the maximal set of all weights $\boldsymbol{w}$ that, if true, have any chance at all of rejecting $\tilde{\boldsymbol{w}}$ under the significance level corresponding to $e^{\rho/2}$, given the existing dataset.

EC1.6. Proof of Theorem 1

Before proving Theorem 1, we first prove the following lemma.

LEMMA 1. *The valuation function $v_\mu(\boldsymbol{w})$ is concave in $\boldsymbol{w}$.*

Proof. Under the assumption that $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{w})$ is concave in $\boldsymbol{w}$ for all $\boldsymbol{x}\in\mathcal{X}$, we see that $\boldsymbol{y}^\top\boldsymbol{f}(\boldsymbol{x},\boldsymbol{w})$ is convex in $\boldsymbol{y}$ and concave in $\boldsymbol{w}$ for all $\boldsymbol{x}$. It follows that $\boldsymbol{y}^\top\tilde{\boldsymbol{z}}$ is convex in $\boldsymbol{y}$. Now for any $\mu\in[0,1]$, it is easy to see that the objective function $\mu\boldsymbol{y}^\top\tilde{\boldsymbol{z}}+(1-\mu)\boldsymbol{y}^\top\boldsymbol{f}(\boldsymbol{x};\boldsymbol{w})$ is convex in $\boldsymbol{y}$ and concave in $\boldsymbol{w}$. The concavity of $v_\mu(\boldsymbol{w})$ over $\boldsymbol{w}$ arises from the infimum operator over $\boldsymbol{y}$. □

Now we are ready to prove Theorem 1. For DDR, we have

$$\begin{aligned}&\min_{\boldsymbol{w}}\ \Big\{L(\boldsymbol{w})-\lambda\frac{1}{N}\sum_{n\in[N]}\min_{\boldsymbol{y}_n\in\mathcal{Y}}\Big\{\mu\boldsymbol{y}_n^\top\tilde{\boldsymbol{z}}_n+(1-\mu)\boldsymbol{y}_n^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big\}\Big\} \qquad \text{(EC1.2)}\\ &=\min_{\boldsymbol{w}}\ \Big\{\frac{1}{N}\sum_{n\in[N]}\max_{\boldsymbol{y}_n\in\mathcal{Y}}\Big\{\ell\big(\tilde{\boldsymbol{z}}_n;\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\big)-\lambda\Big(\mu\boldsymbol{y}_n^\top\tilde{\boldsymbol{z}}_n+(1-\mu)\boldsymbol{y}_n^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big)\Big\}\Big\}.\end{aligned}$$

In what follows, we denote $\sigma=L(\tilde{\boldsymbol{w}})$, which is a known constant given the data. For Robust-DDR, we can write the Lagrangian of it as

$$\begin{aligned}&\max_{\boldsymbol{y}_n\in\mathcal{Y},\,n\in[N],\alpha\geq0}\ \min_{\boldsymbol{w}}\quad \alpha\Big(L(\boldsymbol{w})-\sigma-\rho\Big)-\frac{1}{N}\sum_{n\in[N]}\Big[\mu\boldsymbol{y}_n^\top\tilde{\boldsymbol{z}}_n+(1-\mu)\boldsymbol{y}_n^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big]\\ =\ &\min_{\boldsymbol{w}}\ \max_{\boldsymbol{y}_n\in\mathcal{Y},\,n\in[N],\alpha\geq0}\quad \alpha\Big(L(\boldsymbol{w})-\sigma-\rho\Big)-\frac{1}{N}\sum_{n\in[N]}\Big[\mu\boldsymbol{y}_n^\top\tilde{\boldsymbol{z}}_n+(1-\mu)\boldsymbol{y}_n^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big] \qquad \text{(EC1.3)}\\ =\ &\min_{\boldsymbol{w}}\frac{1}{N}\sum_{n\in[N]}\max_{\boldsymbol{y}_n\in\mathcal{Y},\alpha\geq0}\quad \Big\{\alpha\ell\big(\tilde{\boldsymbol{z}}_n;\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\big)-\mu\boldsymbol{y}_n^\top\tilde{\boldsymbol{z}}_n-(1-\mu)\boldsymbol{y}_n^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big\}-\alpha(\sigma+\rho).\end{aligned}$$

Notice that $\boldsymbol{w}=\tilde{\boldsymbol{w}}$ is always a feasible solution by assumption. Moreover, $\rho>0$, hence $\boldsymbol{w}=\tilde{\boldsymbol{w}}$ is an interior point in the feasibility region. Additionally, the objective function of the inner maximization problem permits convexity in $\boldsymbol{w}$ and concavity in $(\boldsymbol{y}_n,\alpha)$. This achieves Slater's condition, and therefore strong duality holds. It is easy to see that the solution of Problem (EC1.2) coincides with that of Problem (EC1.3) when letting $\alpha^\star=1/\lambda$.

REMARK 1. Here, we make a quick comment about the limiting case $\mu=1$. Notice that if $\mu=1$, the regularization term does not involve the weights $\boldsymbol{w}$ in any form, hence the argmin obtained would coincide with $\arg\min_{\boldsymbol{w}} L(\boldsymbol{w})$, which is just $\hat{\boldsymbol{w}}$. However, it is not clear, that given a fixed λ bounded away from 0, that as $\mu\nearrow 1$, it is necessary for $\boldsymbol{w}(\mu)\to\hat{\boldsymbol{w}}$ uniformly. This is because the behaviour of μ is asymptotic at $\mu=1$; beyond $\mu=1$, there are no consistent ways for ensuring that Lemma 1 holds. Indeed, we see from our numerical simulations that as μ gets closer to 1, we do not recover $\hat{\boldsymbol{w}}$, as long as λ does not also uniformly decrease to 0. □

EC1.7. Proof of Proposition 4 and Corollary 1

We prove both the proposition and the corollary simultaneously.

DEFINITION 1 (ROBUSTNESS DDR). The Robustness Decision-driven Regularization (RDDR, for short) model is defined as the following problem for $\mu\in[0,1)$:

$$\max\quad \rho$$

$$\text{s.t.} \quad \frac{1}{N}\sum_{n\in[N]}\Big[\mu \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n + (1-\mu)\boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big] \le \tau \qquad \forall \boldsymbol{w}\in\mathcal{U}(\rho) \qquad \textbf{(RDDR)}$$
$$\rho>0;\, \boldsymbol{y}_n\in\mathcal{Y} \qquad \forall n\in[N],$$

Note that when $\mu=0$, it recovers the JERO-Like model. Once again, denote $\sigma = L(\tilde{\boldsymbol{w}})$. Now for the robust counterpart of the first constraint, it can be written as:

$$\begin{aligned}\max_{\boldsymbol{w}} \quad & \frac{1}{N}\sum_{n\in[N]}\Big[\mu \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n + (1-\mu)\boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big]\\ \text{s.t.} \quad & L(\boldsymbol{w}) \le \sigma+\rho,\end{aligned}$$

the dual of which is

$$\min_{\alpha\ge 0}\max_{\boldsymbol{w}} \quad \frac{1}{N}\sum_{n\in[N]}\Big[\mu \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n + (1-\mu)\boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big] + \alpha\Big(\sigma+\rho-L(\boldsymbol{w})\Big).$$

Note that $\boldsymbol{w}=\tilde{\boldsymbol{w}}$ is feasible by definition. Moreover, since $\rho>0$, $\boldsymbol{w}=\tilde{\boldsymbol{w}}$ is an interior point. Hence, Slater's condition is achieved, and strong duality holds. Further, $\rho>0$ suffices to consider only $\alpha>0$. Hence, using $\alpha = 1/\beta$ where $\beta>0$,

$$\begin{aligned}&\min_{\alpha>0}\max_{\boldsymbol{w}} \quad \frac{1}{N}\sum_{n\in[N]}\Big[\mu \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n + (1-\mu)\boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big] + \alpha(\sigma+\rho-L(\boldsymbol{w})) \le \tau\\ \Leftrightarrow &\min_{\frac{1}{\beta}}\max_{\boldsymbol{w}} \quad \beta\frac{1}{N}\sum_{n\in[N]}\Big[\mu \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n + (1-\mu)\boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big] + \sigma+\rho-L(\boldsymbol{w}) - \beta\tau \le 0\end{aligned}$$

it follows that

$$\rho \le \max_{\frac{1}{\beta}}\min_{\boldsymbol{w}} -\beta\frac{1}{N}\sum_{n\in[N]}\Big[\mu \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n + (1-\mu)\boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big] - \sigma + L(\boldsymbol{w}) + \beta\tau.$$

This allows us to reformulate the 1 problem as:

$$\max_{\boldsymbol{y}_n\in\mathcal{Y},\, n\in[N],\, \frac{1}{\beta}}\ \min_{\boldsymbol{w}} -\beta\frac{1}{N}\sum_{n\in[N]}\Big[\mu \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n + (1-\mu)\boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big] - \sigma + L(\boldsymbol{w}) + \beta\tau.$$

On the other hand, by Theorem 1, DDR can be reformulated as:

$$\max_{\boldsymbol{y}_n\in\mathcal{Y},\, n\in[N]}\ \min_{\boldsymbol{w}} \Big\{L(\boldsymbol{w}) + \lambda\frac{1}{N}\sum_{n\in[N]}\Big[\tau - \mu \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n + (1-\mu)\boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big]\Big\}.$$

It follows that by setting $\lambda=\beta^\star$—the optimal β attained under the reformulated **RDDR**—the solution of **RDDR** coincides with that of DDR. □

REMARK 2. Theorem 1 guarantees that, out-of-sample, if the optimal objective value of Robust-DDR is $\Gamma^\star$, then $v_\mu(\check{\boldsymbol{w}}) \le \Gamma^\star$, with probability $\mathbb{P}[\check{\boldsymbol{w}}\in\mathcal{U}(\rho)]$, is related to the power of the learning model if L is the log-likelihood function. This is almost as good as saying that the out-of-sample costs are upper bounded with high probability, until we recall that the valuation function is only a surrogate for the true cost function.

EC1.8. Proof of Theorem 2

Define

$$\mathcal{Z}^\star(\boldsymbol{w}) = \left\{ \check{\boldsymbol{z}} := \left\{\check{\boldsymbol{z}}_n \in \mathbb{R}^s\right\}_{n\in[N]} \;\middle|\; \begin{array}{ll} \exists\, \boldsymbol{y}_n \neq \boldsymbol{y}^\star(\check{\boldsymbol{z}}_n),\ n\in[N]: & \\ \left|\dfrac{1}{N}\displaystyle\sum_{n\in[N]}\left[\ell(\check{\boldsymbol{z}}_n;\tilde{\boldsymbol{z}}_n) - \ell(\check{\boldsymbol{z}}_n;\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}))\right]\right| \leq t & \\ \left|\boldsymbol{y}_n^\top\check{\boldsymbol{z}}_n - \boldsymbol{y}_n^\top\tilde{\boldsymbol{z}}_n\right| \leq \phi & \forall n\in[N] \\ \left|\boldsymbol{y}_n^\top\check{\boldsymbol{z}}_n - \boldsymbol{y}_n^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\right| \leq \psi & \forall n\in[N] \\ \left|\boldsymbol{y}^\star(\check{\boldsymbol{z}}_n)^\top\check{\boldsymbol{z}}_n - \boldsymbol{y}^\star(\check{\boldsymbol{z}}_n)^\top\tilde{\boldsymbol{z}}_n\right| \leq \phi & \forall n\in[N] \\ \left|\boldsymbol{y}^\star(\check{\boldsymbol{z}}_n)^\top\check{\boldsymbol{z}}_n - \boldsymbol{y}^\star(\check{\boldsymbol{z}}_n)^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\right| \leq \psi & \forall n\in[N] \\ \left|\boldsymbol{y}^\star(\check{\boldsymbol{z}}_n)^\top\check{\boldsymbol{z}}_n - \boldsymbol{y}^\star(\tilde{\boldsymbol{z}}_n)^\top\tilde{\boldsymbol{z}}_n\right| \leq \eta & \forall n\in[N] \end{array} \right\}.$$

Notice that $\mathcal{Z}^\star(\boldsymbol{w})$ is just $\mathcal{Z}(\boldsymbol{w})$, where we preserve the global condition and then demand that good estimates of the cost function are satisfied on two decision points, $\boldsymbol{y}^\star(\boldsymbol{z}_n)$ and $\boldsymbol{y}_n$, which will be the optimal recourse solution obtained for the optimization problem at the end. As such, $\mathcal{Z}^\star(\boldsymbol{w})$ relaxes the conditions for $\boldsymbol{z}_n$, thus $\mathcal{Z}^\star(\boldsymbol{w}) \supseteq \mathcal{Z}(\boldsymbol{w})$, and the supremum over the latter is bounded above by the supremum of the former. Henceforth, we are interested in bounding the supremum over $\mathcal{Z}^\star(\boldsymbol{w})$.

For ease of notations, hereinafter in this proof, we let $\hat{\boldsymbol{y}}_n \in \arg\min_{\boldsymbol{y}\in\mathcal{Y}} \boldsymbol{y}^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})$, $\check{\boldsymbol{y}}_n \in \arg\min_{\boldsymbol{y}\in\mathcal{Y}} \boldsymbol{y}^\top\check{\boldsymbol{z}}_n$, and $\tilde{\boldsymbol{y}}_n \in \arg\min_{\boldsymbol{y}\in\mathcal{Y}} \boldsymbol{y}^\top\tilde{\boldsymbol{z}}_n$ for $n\in[N]$. In what ensues, we shall bound the supremum by decomposing $\mathcal{Z}^\star(\boldsymbol{w})$ into two steps. Denote

$$\mathcal{Z}_0(\boldsymbol{w}) = \left\{ \check{\boldsymbol{z}} := \left\{\check{\boldsymbol{z}}_n \in \mathbb{R}^s\right\}_{n\in[N]} \;\middle|\; \begin{array}{ll} \left|\check{\boldsymbol{y}}_n^\top\check{\boldsymbol{z}}_n - \check{\boldsymbol{y}}_n^\top\tilde{\boldsymbol{z}}_n\right| \leq \phi & \forall n\in[N] \\ \left|\check{\boldsymbol{y}}_n^\top\check{\boldsymbol{z}}_n - \check{\boldsymbol{y}}_n^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\right| \leq \psi & \forall n\in[N] \\ \left|\check{\boldsymbol{y}}_n^\top\tilde{\boldsymbol{z}}_n - \tilde{\boldsymbol{y}}_n^\top\tilde{\boldsymbol{z}}_n\right| \leq \eta & \forall n\in[N] \end{array} \right\},$$

and

$$\mathcal{Z}_1(\boldsymbol{w}) = \left\{ \check{\boldsymbol{z}} := \left\{\check{\boldsymbol{z}}_n \in \mathbb{R}^s\right\}_{n\in[N]} \;\middle|\; \begin{array}{ll} \exists\, \boldsymbol{y}_n \neq \check{\boldsymbol{y}}_n,\ n\in[N]: & \\ \left|\boldsymbol{y}_n^\top\check{\boldsymbol{z}}_n - \boldsymbol{y}_n^\top\tilde{\boldsymbol{z}}_n\right| \leq \phi & \forall n\in[N] \\ \left|\boldsymbol{y}_n^\top\check{\boldsymbol{z}}_n - \boldsymbol{y}_n^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\right| \leq \psi & \forall n\in[N] \\ \left|\dfrac{1}{N}\displaystyle\sum_{n\in[N]}\left[\ell(\check{\boldsymbol{z}}_n;\tilde{\boldsymbol{z}}_n) - \ell(\check{\boldsymbol{z}}_n;\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}))\right]\right| \leq t & \end{array} \right\},$$

such that $\mathcal{Z}^\star(\boldsymbol{w}) = \mathcal{Z}_0(\boldsymbol{w}) \cap \mathcal{Z}_1(\boldsymbol{w})$. In other words, the worst case regret is:

$$\begin{aligned} &\max_{\check{\boldsymbol{z}}_n\in\mathcal{Z}(\boldsymbol{w})} \left\{\frac{1}{N}\sum_{n\in[N]}\left[\hat{\boldsymbol{y}}_n^\top\check{\boldsymbol{z}}_n - \check{\boldsymbol{y}}_n^\top\check{\boldsymbol{z}}_n\right]\right\} \\ \leq &\max_{\check{\boldsymbol{z}}_n\in\mathcal{Z}^\star(\boldsymbol{w})} \left\{\frac{1}{N}\sum_{n\in[N]}\left[\hat{\boldsymbol{y}}_n^\top\check{\boldsymbol{z}}_n - \check{\boldsymbol{y}}_n^\top\check{\boldsymbol{z}}_n\right]\right\} \\ = &\max_{\check{\boldsymbol{z}}_n\in\mathcal{Z}_0(\boldsymbol{w})\cap\mathcal{Z}_1(\boldsymbol{w})} \left\{\frac{1}{N}\sum_{n\in[N]}\left[\hat{\boldsymbol{y}}_n^\top\check{\boldsymbol{z}}_n - \check{\boldsymbol{y}}_n^\top\check{\boldsymbol{z}}_n\right]\right\} \\ = &\max_{\check{\boldsymbol{z}}_n\in\mathcal{Z}_1(\boldsymbol{w})} \left\{\frac{1}{N}\sum_{n\in[N]}\left[\hat{\boldsymbol{y}}_n^\top\check{\boldsymbol{z}}_n - \check{\boldsymbol{y}}_n^\top\check{\boldsymbol{z}}_n\right]\right\} \end{aligned} \tag{EC1.4}$$

$$\begin{aligned}
\text{s.t. } & \left|\check{\boldsymbol{y}}_n^\top \check{\boldsymbol{z}}_n - \check{\boldsymbol{y}}_n^\top \tilde{\boldsymbol{z}}_n\right| \le \phi && \forall n \in [N];\\
& \left|\check{\boldsymbol{y}}_n^\top \check{\boldsymbol{z}}_n - \check{\boldsymbol{y}}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\right| \le \psi && \forall n \in [N];\\
& \left|\check{\boldsymbol{y}}_n^\top \tilde{\boldsymbol{z}}_n - \tilde{\boldsymbol{y}}_n^\top \tilde{\boldsymbol{z}}_n\right| \le \eta && \forall n \in [N].
\end{aligned}$$

Here, we consider the Lagrangian relaxation of (EC1.4). Let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}^1, \boldsymbol{\beta}^2)$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_n^1, \boldsymbol{\gamma}_n^2)$ be the Lagrange multipliers. Then,

$$\begin{aligned}
&\max_{\check{\boldsymbol{z}} \in \mathcal{Z}^\star(\boldsymbol{w})} \left\{ \frac{1}{N} \sum_{n\in[N]} \left[\hat{\boldsymbol{y}}_n^\top \check{\boldsymbol{z}}_n - \check{\boldsymbol{y}}_n^\top \check{\boldsymbol{z}}_n \right] \right\} \\
&\le \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}\ge \boldsymbol{0}} \max_{\check{\boldsymbol{z}} \in \mathcal{Z}_1(\boldsymbol{w})} \left\{ \frac{1}{N} \sum_{n\in[N]} \Big[\hat{\boldsymbol{y}}_n^\top \check{\boldsymbol{z}}_n - \check{\boldsymbol{y}}_n^\top \check{\boldsymbol{z}}_n \right. \qquad \text{(EC1.5)}\\
&\quad + \alpha_n^1 \Big(\phi - \check{\boldsymbol{y}}_n^\top \check{\boldsymbol{z}}_n + \check{\boldsymbol{y}}_n^\top \tilde{\boldsymbol{z}}_n \Big) + \alpha_n^2 \Big(\phi + \check{\boldsymbol{y}}_n^\top \check{\boldsymbol{z}}_n - \check{\boldsymbol{y}}_n^\top \tilde{\boldsymbol{z}}_n \Big) \\
&\quad + \beta_n^1 \Big(\psi - \check{\boldsymbol{y}}_n^\top \check{\boldsymbol{z}}_n + \check{\boldsymbol{y}}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}) \Big) + \beta_n^2 \Big(\psi + \check{\boldsymbol{y}}_n^\top \check{\boldsymbol{z}}_n - \check{\boldsymbol{y}}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}) \Big) \\
&\quad \left. + \gamma_n^1 \Big(\eta - \check{\boldsymbol{y}}_n^\top \tilde{\boldsymbol{z}}_n + \tilde{\boldsymbol{y}}_n^\top \tilde{\boldsymbol{z}}_n \Big) + \gamma_n^2 \Big(\eta + \check{\boldsymbol{y}}_n^\top \tilde{\boldsymbol{z}}_n - \tilde{\boldsymbol{y}}_n^\top \tilde{\boldsymbol{z}}_n \Big) \Big] \right\} \\
&= \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}\ge \boldsymbol{0}} \max_{\check{\boldsymbol{z}} \in \mathcal{Z}_1(\boldsymbol{w})} \left\{ \frac{1}{N} \sum_{n\in[N]} \Big[\hat{\boldsymbol{y}}_n^\top \check{\boldsymbol{z}}_n + (\alpha_n^2 - \alpha_n^1 + \beta_n^2 - \beta_n^1 - 1) \check{\boldsymbol{y}}_n^\top \check{\boldsymbol{z}}_n \right. \\
&\quad + (\alpha_n^1 - \alpha_n^2 - \gamma_n^1 + \gamma_n^2) \check{\boldsymbol{y}}_n^\top \tilde{\boldsymbol{z}}_n + (\beta_n^2 - \beta_n^1) \Big(\hat{\boldsymbol{y}}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}) - \check{\boldsymbol{y}}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}) \Big) \\
&\quad \left. - (\beta_n^2 - \beta_n^1) \hat{\boldsymbol{y}}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}) + (\gamma_n^1 - \gamma_n^2) \tilde{\boldsymbol{y}}_n^\top \tilde{\boldsymbol{z}}_n + (\alpha_n^1 + \alpha_n^2)\phi + (\beta_n^1 + \beta_n^2)\psi + (\gamma_n^1 + \gamma_n^2)\eta \right] \Big\} \\
&\le \max_{\check{\boldsymbol{z}} \in \mathcal{Z}_1(\boldsymbol{w})} \left\{ \frac{1}{N} \sum_{n\in[N]} \Big[\hat{\boldsymbol{y}}_n^\top \check{\boldsymbol{z}}_n - 2\hat{\boldsymbol{y}}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}) \Big] + a_0 \right\} \qquad \text{(EC1.6)}\\
&\le \max_{\check{\boldsymbol{z}} \in \mathcal{Z}_1(\boldsymbol{w});\ \boldsymbol{y}_n \in \mathcal{Y}, n\in[N]} \left\{ \frac{1}{N} \sum_{n\in[N]} \Big[\boldsymbol{y}_n^\top \check{\boldsymbol{z}}_n - 2\boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}) \Big] \right\} + a_0, \qquad \text{(EC1.7)}
\end{aligned}$$

where $a_0 = \phi + 2\psi + \eta + \dfrac{1}{N} \sum_{n\in[N]} \tilde{\boldsymbol{y}}_n^\top \tilde{\boldsymbol{z}}_n$. Here, inequality (EC1.5) follows from weak duality. Inequality (EC1.6) follows from two facts. First, we make the following choice of dual variables

$$\left(\alpha_n^1, \alpha_n^2, \beta_n^1, \beta_n^2, \gamma_n^1, \gamma_n^2\right) = \Big(1, 0, 0, 2, 1, 0\Big) \quad n \in [N].$$

Second, note that $\hat{\boldsymbol{y}}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}) - \check{\boldsymbol{y}}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}) \le 0$, as $\hat{\boldsymbol{y}}_n \in \arg\min_{\boldsymbol{y}_n} \boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})$ by definition. Finally, inequality (EC1.7) follows trivially from the fact that $\hat{\boldsymbol{y}}_n \in \mathcal{Y}$ for all $n \in [N]$.

Having relaxed $\hat{\boldsymbol{y}}_n$ to $\boldsymbol{y}_n$, we once again perform another Lagrangian relaxation, this time on $\mathcal{Z}_1(\boldsymbol{w})$. Let $\bar{\boldsymbol{\alpha}} = (\bar{\boldsymbol{\alpha}}^1, \bar{\boldsymbol{\alpha}}^2)$, $\bar{\boldsymbol{\beta}} = (\bar{\boldsymbol{\beta}}^1, \bar{\boldsymbol{\beta}}^2)$, $\bar{\boldsymbol{\gamma}} = (\bar{\gamma}^1, \bar{\gamma}^2)$ be the Lagrange multipliers. It follows that:

$$\begin{aligned}
&\max_{\check{\boldsymbol{z}} \in \mathcal{Z}_1(\boldsymbol{w});\ \boldsymbol{y}_n \in \mathcal{Y}, n\in[N]} \left\{ \frac{1}{N} \sum_{n\in[N]} \Big[\boldsymbol{y}_n^\top \check{\boldsymbol{z}}_n - 2\boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}) \Big] \right\} \\
&\le \min_{\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\gamma}}\ge \boldsymbol{0}\ \check{\boldsymbol{z}}_n;} \max_{\boldsymbol{y}_n \in \mathcal{Y}, n\in[N]} \left\{ \frac{1}{N} \sum_{n\in[N]} \Big\{ \boldsymbol{y}_n^\top \check{\boldsymbol{z}}_n - 2\boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}) \right. \qquad \text{(EC1.8)}\\
&\quad + \bar{\alpha}_n^1 \Big(\phi - \boldsymbol{y}_n^\top \check{\boldsymbol{z}}_n + \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n \Big) + \bar{\alpha}_n^2 \Big(\phi + \boldsymbol{y}_n^\top \check{\boldsymbol{z}}_n - \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n \Big) \\
&\quad \left. + \bar{\beta}_n^1 \Big(\psi - \boldsymbol{y}_n^\top \check{\boldsymbol{z}}_n + \boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}) \Big) + \bar{\beta}_n^2 \Big(\psi + \boldsymbol{y}_n^\top \check{\boldsymbol{z}}_n - \boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w}) \Big) \right\}
\end{aligned}$$

$$+\bar{\gamma}^1\Big(t-\frac{1}{N}\sum_{n\in[N]}\big[\ell(\tilde{\boldsymbol{z}}_n;\check{\boldsymbol{z}}_n)-\ell\big(\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w});\check{\boldsymbol{z}}\big)\big]\Big)+\bar{\gamma}^2\Big(t+\frac{1}{N}\sum_{n\in[N]}\big[\ell(\tilde{\boldsymbol{z}}_n;\check{\boldsymbol{z}}_n)-\ell\big(\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w});\check{\boldsymbol{z}}\big)\big]\Big)\Bigg\}$$

$$=\min_{\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\gamma}}\geq\mathbf{0}}\ \max_{\check{\boldsymbol{z}}_n;\ \boldsymbol{y}_n\in\mathcal{Y},n\in[N]}\quad\Bigg\{\frac{1}{N}\sum_{n\in[N]}\Big[(1-\bar{\alpha}_n^1+\bar{\alpha}_n^2-\bar{\beta}_n^1+\bar{\beta}_n^2)\boldsymbol{y}_n^\top\check{\boldsymbol{z}}_n+(\bar{\alpha}_n^1-\bar{\alpha}_n^2)\boldsymbol{y}_n^\top\tilde{\boldsymbol{z}}_n$$

$$+(\bar{\beta}_n^1-\bar{\beta}_n^2-2)\boldsymbol{y}_n^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})+(\bar{\gamma}^1-\bar{\gamma}^2)\big(\ell\big(\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w});\check{\boldsymbol{z}}_n\big)-\ell(\tilde{\boldsymbol{z}}_n;\check{\boldsymbol{z}}_n)\big)\Big]\Bigg\}$$

$$+\frac{1}{N}\sum_{n\in[N]}\Big[(\bar{\alpha}_n^1+\bar{\alpha}_n^2)\phi+(\bar{\beta}_n^1+\bar{\beta}_n^2)\psi\Big]+(\bar{\gamma}^1+\bar{\gamma}^2)t$$

$$\leq\bar{\lambda}\frac{1}{N}\sum_{n\in[N]}\ell\big(\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w});\tilde{\boldsymbol{z}}_n\big)+\frac{1}{N}\sum_{n\in[N]}\max_{\boldsymbol{y}_n\in\mathcal{Y}}\Big[-\mu\boldsymbol{y}_n^\top\tilde{\boldsymbol{z}}_n-(1-\mu)\boldsymbol{y}_n^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big]+a_1,\tag{EC1.9}$$

where $a_1=\phi+2\psi+t$. The inequality (EC1.8) is once again due to the weak duality, and inequality (EC1.9) is because of the triangle inequality assumption on $\ell(\boldsymbol{u};\boldsymbol{v})$ and the following choice of dual variables,

$$\big(\bar{\alpha}_n^1,\bar{\alpha}_n^2,\bar{\beta}_n^1,\bar{\beta}_n^2,\bar{\gamma}^1,\bar{\gamma}^2\big)=\Big(\frac{1-\mu}{2},\frac{1+\mu}{2},\frac{3+\mu}{2},\frac{1-\mu}{2},\frac{1+\bar{\lambda}}{2},\frac{1-\bar{\lambda}}{2}\Big)\quad n\in[N]$$

for some $\mu\in[-1,1)$ and $\bar{\lambda}\in(0,1]$.

Therefore, we have

$$\max_{\check{\boldsymbol{z}}\in\mathcal{Z}(\boldsymbol{w})}\Bigg\{\frac{1}{N}\sum_{n\in[N]}\Big[\boldsymbol{y}^\star(\hat{\boldsymbol{z}}_n)^\top\check{\boldsymbol{z}}_n-\min_{\boldsymbol{y}_n\in\mathcal{Y}}\check{\boldsymbol{y}}_n^\top\check{\boldsymbol{z}}_n\Big]\Bigg\}$$

$$\leq\max_{\check{\boldsymbol{z}}\in\mathcal{Z}_1(\boldsymbol{w});\ \boldsymbol{y}_n\in\mathcal{Y},n\in[N]}\quad\Bigg\{\frac{1}{N}\sum_{n\in[N]}\Big[\boldsymbol{y}_n^\top\check{\boldsymbol{z}}_n-2\boldsymbol{y}_n^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big]+a_0\Bigg\}$$

$$\leq\bar{\lambda}\frac{1}{N}\sum_{n\in[N]}\ell\big(\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w});\tilde{\boldsymbol{z}}_n\big)-\frac{1}{N}\sum_{n\in[N]}\min_{\boldsymbol{y}_n\in\mathcal{Y}}\Big[\mu\boldsymbol{y}_n^\top\tilde{\boldsymbol{z}}_n+(1-\mu)\boldsymbol{y}_n^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big]+a_0+a_1=\frac{1}{\lambda}V(\boldsymbol{w})+a,$$

with $\lambda=1/\bar{\lambda}$ and $a:=a_0+a_1=2\phi+4\psi+\eta+t+\frac{1}{N}\sum_{n\in[N]}\tilde{\boldsymbol{y}}_n^\top\tilde{\boldsymbol{z}}_n$. Notice here that $\frac{1}{N}\sum_{n\in[N]}\tilde{\boldsymbol{y}}_n^\top\tilde{\boldsymbol{z}}_n$ is evaluated only over the training data $\tilde{\boldsymbol{z}}_n$ and hence is a constant given the training dataset.

Hence, in particular, if we choose $\boldsymbol{w}=\hat{\boldsymbol{w}}^{\mathrm{DDR}}$, we have

$$\max_{\check{\boldsymbol{z}}\in\mathcal{Z}(\boldsymbol{w}^{\mathrm{DDR}})}\Bigg\{\frac{1}{N}\sum_{n\in N}\Big[\boldsymbol{y}^\star(\hat{\boldsymbol{z}}_n)^\top\check{\boldsymbol{z}}_n-\min_{\boldsymbol{y}_n\in\mathcal{Y}}\boldsymbol{y}_n^\top\check{\boldsymbol{z}}_n\Big]\Bigg\}\leq\frac{1}{\lambda}V(\hat{\boldsymbol{w}}^{\mathrm{DDR}})+a.$$

This completes the proof. □

EC1.9. Proof of Proposition 5

When $L(\boldsymbol{w})=(1-\mu)\frac{1}{N}\sum_{n\in[N]}\boldsymbol{y}^\star(\tilde{\boldsymbol{z}}_n)^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})$, $\lambda=1$, and $\mu=-1$, DDR becomes

$$\min_{\boldsymbol{w}}\ (1-\mu)\frac{1}{N}\sum_{n\in[N]}\boldsymbol{y}^\star(\tilde{\boldsymbol{z}}_n)^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})-\frac{1}{N}\sum_{n\in[N]}\min_{\boldsymbol{y}_n\in\mathcal{Y}}\Big\{-\boldsymbol{y}_n^\top\tilde{\boldsymbol{z}}_n+2\boldsymbol{y}_n^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n;\boldsymbol{w})\Big\},$$

which is exactly the SPO+ model. □

REMARK 3. SPO+ occurs when $\mu=-1$. In this case, DDR reduces to, for some $\lambda\geq0$:

$$\min_{\boldsymbol{w}}\ L(\boldsymbol{w})-\lambda\Bigg(\frac{1}{N}\sum_{n\in[N]}\min_{\boldsymbol{y}_n\in\mathcal{Y}}\Big\{2\boldsymbol{y}^\top\boldsymbol{f}(\tilde{\boldsymbol{x}}_n,\boldsymbol{w})-\boldsymbol{y}^\top\tilde{\boldsymbol{z}}_n\Big\}\Bigg).\tag{EC1.10}$$

One way to understand the motivation of such a model, is to notice that the expression within the infimum has the following decomposition into the estimated cost and its estimation error when compared against the empirical cost:

$$\underbrace{\boldsymbol{y}^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n, \boldsymbol{w}) - \boldsymbol{y}^\top \tilde{\boldsymbol{z}}}_{\text{estimation error}} + \underbrace{\boldsymbol{y}^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n, \boldsymbol{w})}_{\text{best estimate}}. \tag{EC1.11}$$

We remark here that if we let $\phi = 0, \eta = 0$ and t, ψ to be sufficiently large such that the first and the third conditions in $\mathcal{Z}(\boldsymbol{w})$ no longer exist, *i.e.*,

$$\mathcal{Z}(\boldsymbol{w}) = \left\{ \boldsymbol{z} := \left\{\boldsymbol{z}_n \in \mathbb{R}^s\right\}_{n\in[N]} \;\middle|\; \begin{array}{ll} \forall\, \boldsymbol{y}_n \in \mathcal{Y},\ n \in [N]: & \\ \left|\boldsymbol{y}_n^\top \boldsymbol{z}_n - \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n\right| = 0 & \forall n \in [N] \\ \left|\boldsymbol{y}^\star(\boldsymbol{z}_n)^\top \tilde{\boldsymbol{z}}_n - \boldsymbol{y}^\star(\tilde{\boldsymbol{z}}_n)^\top \tilde{\boldsymbol{z}}_n\right| = 0 & \forall n \in [N] \end{array} \right\},$$

`Regret-DDR` is exactly `SPO`. Furthermore, if one follows the proof for Theorem 2 with these specifications of the parameters, then one will also recover the construction of `SPO+` from `SPO`. This once again reinforces the notion that the construction of `DDR` as an approximation to `Regret-DDR` is as canonical as the construction of `SPO+` as an approximation to `SPO`.

EC2. Sub-optimality of the SLO approach and relationship of DDR to other regularizers

EC2.1. Sub-optimality of the SLO approach

To better understand why directly minimizing prediction error is sub-optimal, consider the toy knapsack problem of picking smallest cost $\boldsymbol{z}$, where cost function is $c(\boldsymbol{y}; \boldsymbol{z}) = \boldsymbol{y}^\top \boldsymbol{z}$, the feasible set is $\mathcal{Y} = \{\boldsymbol{y} \in [0,1]^N, \boldsymbol{y}^\top \boldsymbol{1} = 1\}$, and $\boldsymbol{z}$ is generated from some true linear relationship $\boldsymbol{z} = \check{\boldsymbol{w}}^\top \boldsymbol{x} + \boldsymbol{\epsilon}$, with true parameter $\check{\boldsymbol{w}}$. Typically, prediction error is measured globally. Thus, the residuals $\boldsymbol{\epsilon}$ of a model that minimizes global prediction error, will be smaller over less extreme-valued regions of $\boldsymbol{x}$ and vice versa, as illustrated by the dark blue shaded confidence interval in Figure EC2.1 below. However, in the knapsack problem, we are concerned about extreme values of $\boldsymbol{z}$, which occur over the region of extreme values of $\boldsymbol{x}$ where the prediction error is currently largest. In other words, more generally, the cost minimization problem induces a region $\overline{\mathcal{X}} \subset \mathcal{X}$ over context $\boldsymbol{x}$, corresponding to a region $\overline{\mathcal{Z}} \subset \mathcal{Z}$ of the uncertainty $\boldsymbol{z}$, where optimal decisions to the cost minimization problem are more likely to occur than not. Moreover, this region $\overline{\mathcal{X}}$ is often extreme in relation to the support $\mathcal{X}$ as optimization, by definition, seeks extreme results. If one is able to improve the predictive accuracy over this region $\overline{\mathcal{Z}}$, then the optimization model is better able to differentiate between alternative decisions $\boldsymbol{y}$. Thus, a model with confidence interval indicated by the broken line in Figure EC2.1 is expected to have higher decision power than its shaded blue counterpart, as it incurs smaller prediction error over the region that matters—the small values of $\boldsymbol{z}$. It may have a larger overall prediction error, but that results from sacrificing prediction accuracy over non-critical regions of $\mathcal{X}$—that lead to large values of $\boldsymbol{z}$—in favour of $\overline{\mathcal{X}}$.

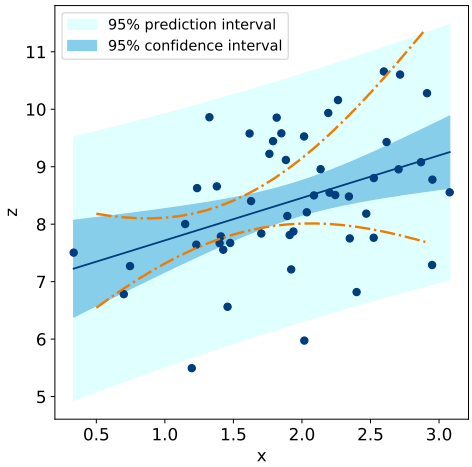

Figure EC2.1 One-dimensional illustrative of centred (dark blue shaded area) and non-centred (broken orange line) confidence intervals

EC2.2. Relationship to other regularizers

In the DDR model, the regularizer $R_\mu(\boldsymbol{w}) := -\frac{1}{N}\sum_n \nu_\mu^n(\boldsymbol{w})$ is chosen as the sample average approximation of the valuation function $\nu_\mu(\boldsymbol{w})$. Here, we shorthand $\nu_\mu^n(\boldsymbol{w})$ to mean the valuation function evaluated on the single data point indexed by n, $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{z}}) = (\tilde{\boldsymbol{x}}_n, \tilde{\boldsymbol{z}}_n)$. In both Lasso and Ridge regressions, the regularizer is chosen as some norm: $R(\boldsymbol{w}) = ||\boldsymbol{w}||_q$. Nonetheless, recall that each norm can be represented as the optimizer of some optimization problem. By the definition of dual norm, we have

$$R(\boldsymbol{w}) = ||\boldsymbol{w}||_q = \max_{\boldsymbol{\xi}} \sum_{i\in[p]}\sum_{j\in[d]} w_{ij}\xi_{i,j}, \quad \text{s.t. } \boldsymbol{\xi} \in \Xi := \left\{\boldsymbol{\xi} \in \mathbb{R}^{p\times d} : ||\boldsymbol{\xi}||_{q_*} \le 1\right\} \tag{EC2.12}$$

where $||\cdot||_{q_*}$ represents the dual norm of $||\cdot||_q$, and $\frac{1}{q} + \frac{1}{q_*} = 1$.

Consider the case where the learner is linear, *i.e.*, $\tilde{\boldsymbol{z}} := \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{x}$. In the case where $\mu = 0$, the decision-driven regularizer is

$$\begin{aligned} R_0(\boldsymbol{w}) &= \frac{1}{N}\sum_n \nu_0^n(\boldsymbol{w}) = \max_{y_n\in\mathcal{Y}} -\frac{1}{N}\sum_n \boldsymbol{y}_n^\top \boldsymbol{w}^\top \boldsymbol{x}_n \\ &= \max_{y_n\in\mathcal{Y}} \sum_{i\in[p]}\sum_{j\in[d]} w_{ij}\left(-\frac{1}{N}\sum_n x_{n,i}y_{n,j}\right). \end{aligned}$$

Thus, we may express $R_0(\boldsymbol{w})$ as the following:

$$R_0(\boldsymbol{w}) = \max_{\boldsymbol{\xi}} \sum_{i\in[p]}\sum_{j\in[d]} w_{ij}\xi_{i,j}, \quad \text{s.t. } \boldsymbol{\xi} \in \Xi_0^N := \left\{\boldsymbol{\xi} \in \mathbb{R}^{p\times d} : \exists \boldsymbol{y}_n \in \mathcal{Y}, n\in[N], \xi_{i,j} = -\frac{1}{N}\sum_n x_{n,i}y_{n,j}\right\} \tag{EC2.13}$$

that depends on the dataset $\mathcal{D}^N$. Denote X^N as the $p \times N$ matrix that collects all p-dimensional feature vectors in the dataset of N points, and denote $\mathcal{Y}^N := \mathcal{Y} \times \cdots \times \mathcal{Y}$ as the product space of $\mathcal{Y}$. Then, we can notice that $\Xi_0^N = \frac{1}{N}X^{N\top}\mathcal{Y}^N := \{\frac{1}{N}X^{N\top}\boldsymbol{y} : \boldsymbol{y} \in \mathcal{Y}^N\}$, which is the linear transformation of the feasibility space by the data matrix X^N.

The forms of (EC2.12) and (EC2.13) allow us to draw the connection between decision-driven regularization and norm-based regularizers like Lasso and Ridge regressions. We wish to analyze this relationship by assuming that we are currently at the OLS solution, *i.e.*, $\boldsymbol{w} = \hat{\boldsymbol{w}}^{\text{OLS}}$, and investigating what impact adding either of the regularizers has on the performance of the solution – specifically, how the base solution $\hat{\boldsymbol{w}}^{\text{OLS}}$, results in different weights, such as $\hat{\boldsymbol{w}}^{\text{Lasso}}$ in the case of Lasso. Here, we examine locally around the neighbourhood of OLS weights, $\hat{\boldsymbol{w}}^{\text{OLS}}$, the directions of improving (decreasing) $L(\boldsymbol{w}) + \lambda R(\boldsymbol{w})$ for different regularizers $R(\cdot)$. As $\hat{\boldsymbol{w}}^{\text{OLS}}$ is the minimizer of $L(\boldsymbol{w})$ (hence gradient of L at $\boldsymbol{w} = \hat{\boldsymbol{w}}^{\text{OLS}}$ is 0), the action of adding $R(\boldsymbol{w})$ is to draw the solution in the direction of the gradient of $R(\boldsymbol{w})$. Closer examination of problems (EC2.12) and (EC2.13) reveals that these directions are approximated by the optimal solutions of $\boldsymbol{\xi}$ at $\boldsymbol{w} = \hat{\boldsymbol{w}}^{\text{OLS}}$. This is because $\boldsymbol{\xi}$ are the dual variables in the convex conjugate forms of $R(\boldsymbol{w})$. In other words, around the neighborhood of base solution $\hat{\boldsymbol{w}}^{\text{OLS}}$, the base solution is pushed along the optimal $\boldsymbol{\xi}$'s to problems (EC2.12) and (EC2.13) at $\boldsymbol{w} = \hat{\boldsymbol{w}}^{\text{OLS}}$ towards the norm-based regularizer and DDR weights, respectively. However, problems (EC2.12) and (EC2.13) have the same objective, and hence the same optimizing direction, but with different feasibility sets Ξ and Ξ_0^N. Thus, the eventual optimal solution of $\boldsymbol{\xi}$ depends on how different these feasibility sets are.

Our goal is to examine the impact on the performance of the eventual decisions, of pulling the OLS weights in the directions of the solutions to (EC2.12) and (EC2.13). If the solution $\boldsymbol{\xi}$ to problem (EC2.13) is a local improving direction for the performance, *i.e.*, that DDR improves on the performance of OLS on the eventual cost function, and if the angle between the two solutions $\boldsymbol{\xi}$ to problem (EC2.12) and (EC2.13) is fairly large, then we can conclude that the direction in which norm-based regularizers pull the OLS weights towards, *i.e.*, the solution to problem (EC2.12), may very likely be a local worsening direction for the performance of the eventual cost function.

This actually turns out to be the case in our simulations. We see that DDR improves over OLS, and that the angles of the local directions from DDR to Lasso and Ridge regressions are around 60°. In Table EC2.1, we present the summary statistics of these angles amongst 100 simulations. These angles are significantly large—in other words, the regularizers in Lasso and Ridge regressions are pulling the OLS weights away in very different directions from the decision-driven regularizer. As a consequence, we discover, ironically, that while Lasso and Ridge regressions both improve the prediction accuracy over OLS, they both lead to significantly poorer performance on the cost function.

Table EC2.1 Angles of local directions of Lasso and Ridge to DDR in 100 simulations

Angles to DDR	Lasso	Ridge
Mean	82°	72°
10th Percentile	80°	72°
90th Percentile	85°	73°

In conclusion, what we have explained here is that blindly adopting regularizers, even if for the purposes of improving prediction accuracy, might not lead to better performance. In fact, we can potentially

see the opposite, where they lead to poorer performance, pulling the optimal solution away from the direction of improving performance.

EC3. Pseudo-code for Solving RDDR

COROLLARY 1. *Under the same assumptions as in Theorem 1, then for all $\mu \in [0,1)$ and τ for which $\exists \boldsymbol{y}_n \in \mathcal{Y}$, $n \in [N]$ such that the target is attained $\frac{1}{N}\sum_{n\in[N]}\Big[\mu \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n + (1-\mu)c\big(\boldsymbol{y}_n; \boldsymbol{f}(\tilde{\boldsymbol{x}}_n, \tilde{\boldsymbol{w}})\big)\Big] \leq \tau$, under learning optimal weights $\tilde{\boldsymbol{w}} = \arg\min L(\boldsymbol{w})$, there exists some $\lambda := \lambda(\tau) \geq 0$ for which the solutions of* DDR *and the worst-case weights attained under optimal uncertainty set size $\rho^\star$ from* **RDDR** *coincide.*

In short, we seek the best radius of the uncertainty set ρ, via bisection search, wherein we solve a feasibility sub-problem in each iteration. Here, we present the pseudo-code for the algorithm, however, for more details, the reader can be directed to Zhu et al. (2022), as the procedure for solving the model follows identically.

Algorithm 1 RDDR

Input τ and $\tilde{\boldsymbol{z}}$.
Initialization: $\rho_1 = 0, \rho_2 = \bar{\rho}$, where $\bar{\rho}$ is a sufficiently large number.
while $\rho_2 - \rho_1 > \epsilon$ **do**
  $\rho := (\rho_1 + r_2)/2$
  Solve the subproblem **Sub−RDDR**, obtain optimal value $\delta^\star$ and optimal decisions $\boldsymbol{y}_n^\star$.
  if $\delta^\star > 0$ **then**
    $\rho_2 = \rho$
  else
    $\rho_1 = \rho$
  end if
end while
Output: Optimal $\rho^\star = (\rho_1 + \rho_2)/2$ and optimal decision $\boldsymbol{y}_n^{\text{RDDR}} = \boldsymbol{y}_n^\star$.
Input Optimal $\rho = \rho^\star$ and optimal decision $\boldsymbol{y} = \boldsymbol{y}_n^\star$
Do Solve the Problem (**ROBUSTW**), and obtain $\boldsymbol{w}^\star$
Output Worst scenario $\hat{\boldsymbol{w}}^{\text{RDDR}} = \boldsymbol{w}^\star$

In this algorithm, the sub-problem is described as:

$$\begin{aligned} \min_{t} \quad & t - \tau && \textbf{(Sub−RDDR)} \\ \text{s.t.} \quad & \frac{1}{N}\sum_{n\in[N]}\Big[\mu \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n + (1-\mu)\boldsymbol{y}_n^\top \boldsymbol{f}(\tilde{\boldsymbol{x}}_n; \boldsymbol{w})\Big] \leq t, \qquad \forall \boldsymbol{w} \in \mathcal{U}(\rho) \\ \text{where} \quad & \mathcal{U}(\rho) := \{\boldsymbol{w} : L(\boldsymbol{w}) - L(\bar{\boldsymbol{w}}) \leq \rho\}. \end{aligned}$$

We finally utilize $\rho^\star$ and $\boldsymbol{y}^\star$ derived from the overarching problem to solve for the worst case $\hat{\boldsymbol{w}}^{\text{RDDR}}$. This is done via,

$$\arg\max_{\boldsymbol{w}\in\mathcal{U}(\rho^\star)} \frac{1}{N}\sum_{n\in[N]}\Big[\mu\, c(\boldsymbol{y}_n^\star; \tilde{\boldsymbol{z}}_n) + (1-\mu)\, c\big(\boldsymbol{y}_n^\star; \boldsymbol{f}(\tilde{\boldsymbol{x}}_n; \boldsymbol{w})\big)\Big] \qquad \textbf{(ROBUSTW)}$$

EC4. Additional Simulation Results

In this Appendix, we present further results from the numerical illustration that were omitted from the main text for brevity. This includes the robustness of our findings to different parameters and a short study into the case when the network size is varied.

EC4.1. A Closer Look at Over-fitting in the Tree-based SLO Benchmarks

In §4.1 of the main text, Table 2 reports the in- and out-of-sample prediction accuracy of the SLO benchmarks (OLS, RF, XGBoost) and DDR under their default settings, which points to potential over-fitting in the two tree-based learners. To further elucidate the extent to which such over-fitting depends on model complexity, Table EC4.2 refines the entries for RF and XGBoost by sweeping the maximum tree depth across $\{2, 4, 6, 8, \text{Default}\}$, while keeping OLS and DDR unchanged as references. We can see that changing the hyperparameters might not resolve over-fitting, but is rather an inherent feature of the learner.

Table EC4.2 Prediction accuracy of SLO models in terms of root mean square error (RMSE) across multiple datasets

Models	In-sample RMSE			Out-of-sample RMSE		
	$5^{\text{th}}\%$-tile	Mean	$95^{\text{th}}\%$-tile	$5^{\text{th}}\%$-tile	Mean	$95^{\text{th}}\%$-tile
OLS	3.8393	3.9416	4.0682	4.1279	4.1846	4.2395
DDR	3.8418	3.9440	4.0705	4.1286	4.1830	4.2352
RF(Maximum depth = 2)	4.0068	4.1499	4.2974	4.3693	4.4779	4.5875
RF(Maximum depth = 4)	3.36	3.4682	3.5785	4.3047	4.3904	4.4816
RF(Maximum depth = 6)	2.4572	2.6123	2.782	4.3191	4.3939	4.4766
RF(Maximum depth = 8)	1.8514	1.9727	2.1223	4.338	4.4143	4.504
RF(Default)	1.5872	1.6531	1.7039	4.3479	4.4255	4.5155
XGBoost(Maximum depth = 2)	3.924	4.0634	4.221	4.3788	4.4943	4.6057
XGBoost(Maximum depth = 4)	3.38	3.5202	3.6778	4.4097	4.517	4.6255
XGBoost(Maximum depth = 6)	2.9835	3.114	3.2322	4.4822	4.5833	4.6913
XGBoost(Maximum depth = 8)	2.8146	2.9393	3.0729	4.5155	4.6173	4.7142
XGBoost(Default)	2.9835	3.1140	3.2322	4.4822	4.5833	4.6913

EC4.2. Robustness of Our Model Across Various Settings

This section presents a series of numerical experiments designed to evaluate the robustness of the proposed `DDR` model under varying conditions. Specifically, we investigate the influence of training data size, feature dimensionality, noise level, and model misspecification. To assess the impact of training data size, the sample size N is varied across the set $\{50, 100, 200, 500, 1000\}$, while all other parameters are held constant at their baseline values. Likewise, the number of features p is selected from $\{1, 3, 5, 7, 10, 15\}$. To evaluate the effect of model misspecification, the parameter β is varied over the set $\{0.4, 0.8, 1.0, 1.4, 1.6, 2.0, 3.0, 4.0\}$. Finally, the noise level $\bar{\epsilon}$ is examined by selecting values from $\{0.25, 0.5, 0.75, 1.0\}$. The comparison between `DDR` and OLS is summarized in Figure EC4.2.

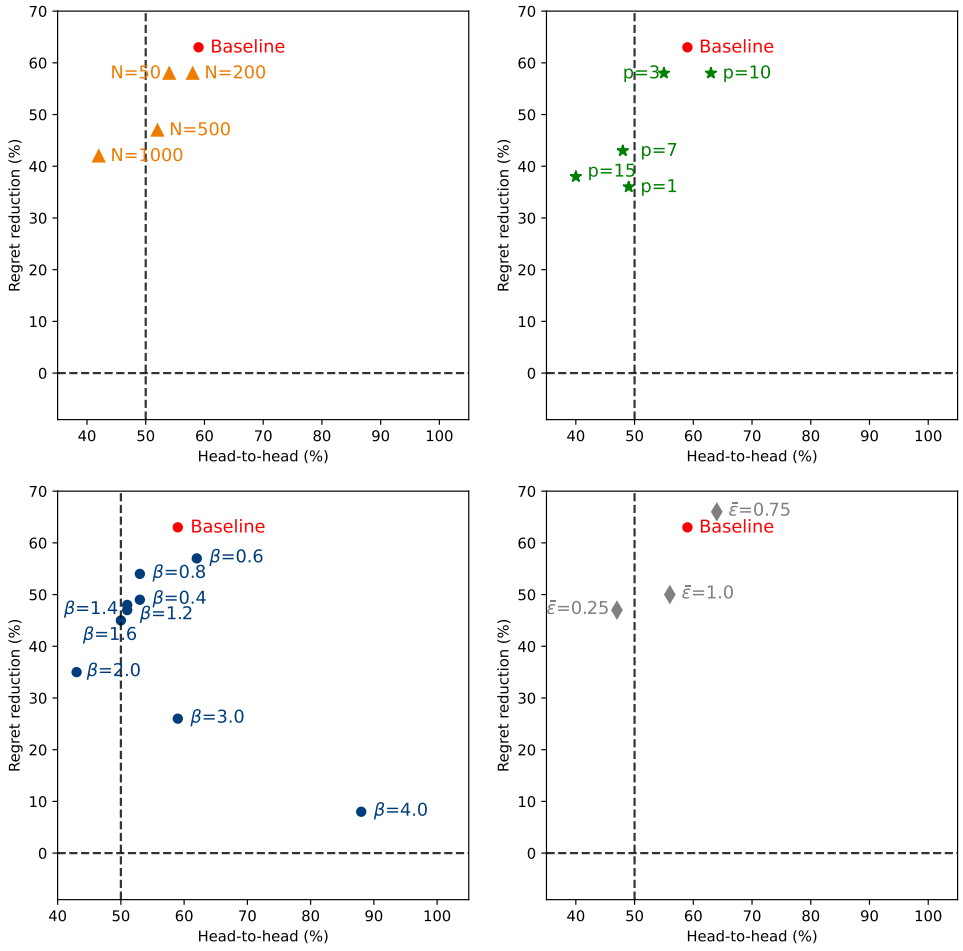

Figure EC4.2 Performance of DDR against OLS under 3×3 grid and $\mu = 0.75$ and $\lambda = 0.8$ under various settings

From Figure EC4.2, we first observe that the DDR model consistently attains superior regret performance, as indicated by positive regret reduction across all evaluated settings. This finding highlights the advantage of employing the DDR model. However, it is also apparent that in certain scenarios, the head-to-head performance of the DDR model does not surpass that of OLS. For instance, when assessing the impact of training data size, the DDR model exhibits improved head-to-head performance across all sample sizes except at $N = 1000$. This outcome is anticipated, as the OLS estimator's accuracy improves with larger sample sizes, thereby reducing the performance gap between DDR and OLS in such cases. Nevertheless, Figure EC4.2 demonstrates that our DDR model outperforms in terms of both regret reduction and head-to-head metrics in the majority of cases.

Similarly, the comparison between the DDR and SPO+ approaches is presented in Figure EC4.3. The results indicate that the DDR model consistently outperforms SPO+ across all settings in terms of both regret reduction and head-to-head performance metrics, thereby demonstrating the superiority of the DDR model over the SPO+ baseline.

EC5. DDR Tree

While we have elected to use a linear learner in our numerical results, DDR can more generally be applied to other learners. To illustrate this, in this section, we consider a tree-based learner in DDR, and term it as DDR tree.

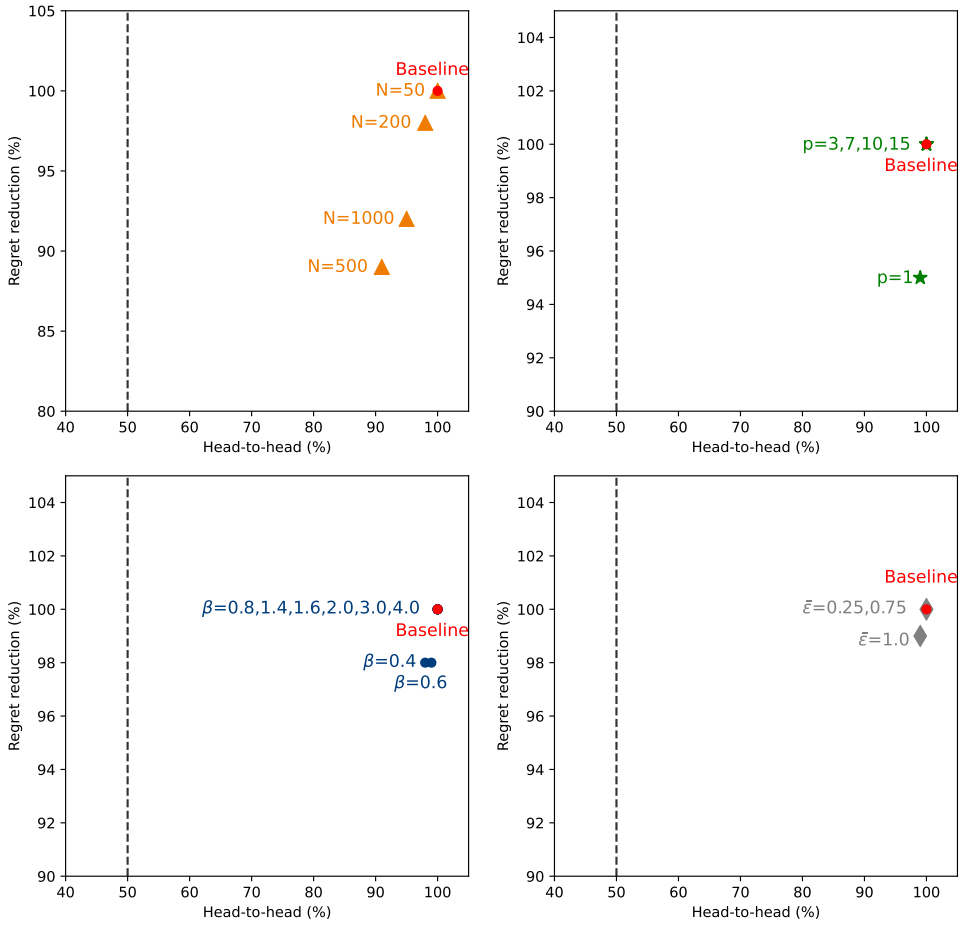

Figure EC4.3 Performance of DDR against SPO+ under 3×3 grid and $\mu = 0.75$ and $\lambda = 0.8$ under various settings

Consider a decision tree with K leaf nodes, where the predictions $\hat{\boldsymbol{z}}$ are constrained to be equal if they fall into the same leaf node. We model this by defining R_k the set of samples assigned to leaf node k, and predictions $\hat{\boldsymbol{z}}_k$ as being indexed by the leaf nodes $k \in [K]$. We write the DDR loss objective function as:

$$L_{\text{DDR}}(R_k, \hat{\boldsymbol{z}}_k) = \sum_{k\in[K]} \sum_{n\in R_k} \|\tilde{\boldsymbol{z}}_n - \hat{\boldsymbol{z}}_k\|_2^2 + \frac{\lambda}{N} \sum_{k\in[K]} \sum_{n\in R_k} \max_{\boldsymbol{y}_n\in\mathcal{Y}} \left\{ -\mu \boldsymbol{y}_n^\top \tilde{\boldsymbol{z}}_n - (1-\mu)\boldsymbol{y}_n^\top \hat{\boldsymbol{z}}_k \right\}$$

where the first term represents the prediction accuracy and the second term represents the decision cost. Here, sets R_k and predictions $\hat{\boldsymbol{z}}_k$ are the parameters ("weights") of the tree to be learned and determined.

Using Proposition 3, the DDR loss function can be further reformulated as:

$$L_{\text{DDR}}(R_k, \hat{\boldsymbol{z}}_k) = \frac{\lambda}{N} \sum_{k\in[K]} \sum_{n\in R_k} \left(\frac{N}{\lambda} \|\tilde{\boldsymbol{z}}_n - \hat{\boldsymbol{z}}_k\|_2^2 + \min_{\boldsymbol{A}^\top \boldsymbol{\kappa}_n \geq -\mu\tilde{\boldsymbol{z}}_n - (1-\mu)\hat{\boldsymbol{z}}_k} \boldsymbol{\kappa}_n^\top \boldsymbol{b} \right).$$

Accordingly, the prediction cost vectors for the leaf nodes, given the partitioning of observations into leaf nodes, *i.e.*, the sets R_k, are determined by solving the following constrained optimization problem:

$$\hat{\boldsymbol{z}}^{\text{DDR}} = \underset{\hat{\boldsymbol{z}}_k\,:\,\forall k\in[K]}{\arg\min} \left\{ \frac{\lambda}{N} \sum_{k\in[K]} \sum_{n\in R_k} \left(\frac{N}{\lambda} \|\tilde{\boldsymbol{z}}_n - \hat{\boldsymbol{z}}_k\|_2^2 + \boldsymbol{\kappa}_n^\top \boldsymbol{b} \right) \middle| \boldsymbol{A}^\top \boldsymbol{\kappa}_n \geq -\mu\tilde{\boldsymbol{z}}_n - (1-\mu)\hat{\boldsymbol{z}}_k, \forall n \in R_k, k \in [K] \right\}$$

If the sets R_k are known, the above optimization problem is easy. Instead, we propose the following two methodologies to optimize for R_k.

EC5.1. Recursive Partitioning Approach.

Emulating the CART algorithm, we construct the decision tree via a greedy recursive partitioning strategy. Let x_{nj} denote the j^{th} feature component of the n^{th} observation in the training data. Starting from the entire training dataset, we consider binary splits defined by pairs (j, s), where j indicates the selected feature and s denotes the split threshold. This split partitions the data into two disjoint subsets:

$$R_1(j,s) = \{n \in [N] \,|\, x_{nj} \leq s\} \quad \text{and} \quad R_2(j,s) = \{n \in [N] \,|\, x_{ij} > s\}.$$

The optimal first split of the decision tree is determined by selecting the pair (j, s) that minimizes the total L_{DDR} loss over both subsets:

$$\begin{array}{ll} \min\limits_{\hat{\boldsymbol{z}}_1, \hat{\boldsymbol{z}}_2} & \dfrac{\lambda}{N}\left[\displaystyle\sum_{n\in R_1}\left(\frac{N}{\lambda}\|\tilde{\boldsymbol{z}}_n - \hat{\boldsymbol{z}}_1\|_2^2 + \boldsymbol{\kappa}_n^\top \boldsymbol{b}\right) + \sum_{n\in R_2}\left(\frac{N}{\lambda}\|\tilde{\boldsymbol{z}}_n - \hat{\boldsymbol{z}}_2\|_2^2 + \boldsymbol{\kappa}_n^\top \boldsymbol{b}\right)\right] \\ \text{s.t.} & \boldsymbol{A}^\top \boldsymbol{\kappa}_n \geq -\mu\tilde{\boldsymbol{z}}_n - (1-\mu)\hat{\boldsymbol{z}}_k, \qquad \forall n \in R_k, k = 1,2 \end{array}$$

Once the optimal split is selected, the procedure is then recursively applied greedily to obtain binary splits in the resulting leaves until the stopping criteria is met. The prediction cost vectors $\hat{\boldsymbol{z}}^{\text{DDR}}$ for each leaf node are then optimized simultaneously, after the complete tree structure has been determined.

EC5.2. Integer Programming Approach.

Motivated by recent advances in decision tree construction through optimization techniques (**?**), DDR tree can be solved as an integer programming model. Specifically, we aim to construct a decision tree based on a pre-specified tree depth H and a minimum number of training observations N_{min} required in each leaf node.

Note that not all leaf nodes are required to be active (*i.e.*, contain training observations), and similarly, not all branch nodes need to be active splits (*i.e.*, nodes that partition the data). Denote the index sets of leaf and branch nodes by $\mathcal{T}_K = \{1, \ldots, K\}$ and $\mathcal{T}_B = \{1, \ldots, B\}$, respectively. Let $g_k = \mathbb{I}\{\text{leaf } k \text{ is not empty}\}$ and $d_t = \mathbb{I}\{\text{branch node } t \text{ is an active split}\}$ be binary variables that indicate the activation status of the corresponding nodes.

Each node t is associated with a split defined by decision variables $\boldsymbol{a}_t \in \{0,1\}^p$ and $b_t \in [0,1]$, where $\boldsymbol{a}_t$ indicates which feature component is involved with the split, and b_t indicates the split point, assuming that all feature components have been normalized to the interval $[0,1]$. Note that $\boldsymbol{a}_t$ and b_t are the decision variables.

Let $p(t)$ denote the parent node of t, and $A_L(t)$ the set of left ancestor nodes of node t, *i.e.*, the set of ancestors of t whose left branch has been followed on the path from the root to t. Similarly, we let $A_R(t)$ as the set of right ancestor nodes of t.

To ensure sufficient separation between data points in the splitting process, we define the minimum resolution for each feature dimension. Let $\varepsilon_j = \min\{x_j^{(n+1)} - x_j^{(n)} \,\big|\, x_j^{(n+1)} \neq x_j^{(n)}, n \in [N-1]\}$ is the smallest nonzero difference between observed value of feature component j, where $x_j^{(n)}$ is the n-th largest value in the j-th feature. In addition, we denote $\varepsilon_{max} = \max_j \varepsilon_j$.

The integer programming formulation of the DDR decision tree is given as follows:

$$
\begin{aligned}
\min\ & \frac{\lambda}{N}\sum_{k\in[K]}\sum_{n\in[N]} r_{nk}\Big(\frac{N}{\lambda}\|\tilde{\boldsymbol{z}}_n-\hat{\boldsymbol{z}}_k\|_2^2+\boldsymbol{\kappa}_n^\top\boldsymbol{b}\Big)\\
\text{s.t}\quad & \boldsymbol{A}^\top\boldsymbol{\kappa}_n\geq -\mu\tilde{\boldsymbol{z}}_n-(1-\mu)\hat{\boldsymbol{z}}_k+(1-r_{nk})M, && \forall n\in[N],k\in[K]\\
& \sum_{k\in[K]} r_{nk}=1, && \forall n\in[N]\\
& r_{nk}\leq g_k, && \forall n\in[N],l\in\mathcal{T}_K\\
& \sum_{n\in[N]} r_{nk}\geq N_{min}g_k, && \forall l\in\mathcal{T}_K\\
& \boldsymbol{a}_m^\top\tilde{\boldsymbol{x}}_n\geq b_m-(1-r_{nk}), && \forall l\in\mathcal{T}_K,n\in[N],m\in A_R(k)\\
& \boldsymbol{a}_m^\top(\tilde{\boldsymbol{x}}_n+\boldsymbol{\varepsilon})\leq b_m+(1+\varepsilon_{max})(1-r_{nk}), && \forall l\in\mathcal{T}_K,n\in[N],m\in A_L(k)\\
& \sum_{j\in[p]} a_{jt}=d_t, && \forall t\in\mathcal{T}_B\\
& 1-d_t\leq b_t\leq 1, && \forall t\in\mathcal{T}_B\\
& d_t\leq d_{p(t)}, && \forall t\in\mathcal{T}_B\setminus\{1\}\\
& a_{jt},d_t\in\{0,1\}, && \forall j\in[p],t\in\mathcal{T}_B\\
& r_{nk},g_k\in\{0,1\}, && \forall n\in[N],l\in\mathcal{T}_K
\end{aligned}
$$

The objective function and the first four constraints form the DDR optimization problem, while the remaining constraints are adapted from the requirements of the optimal classification tree. For more details about the tree constraints, the reader can consult **?**.

The objective function $\frac{\lambda}{N}\sum_{k\in[K]}\sum_{n\in[N]} r_{nk}\Big(\frac{N}{\lambda}\|\tilde{\boldsymbol{z}}_n-\hat{\boldsymbol{z}}_k\|_2^2+\boldsymbol{\kappa}_n^\top\boldsymbol{b}\Big)$ can be further reformulated as the following inequalities:

$$
\begin{aligned}
&\frac{\lambda}{N}\sum_{k\in[K]}\sum_{n\in[N]} r_{nk}\Big(\frac{N}{\lambda}\varrho_{nk}+\boldsymbol{\kappa}_n^\top\boldsymbol{b}\Big)\\
&(\varrho_{nk}+1)^2\geq(\varrho_{nk}-1)^2+4\varpi^2\\
&\varpi^2\geq\|\tilde{\boldsymbol{z}}_n-\hat{\boldsymbol{z}}_k\|_2^2
\end{aligned}
$$

Moreover, the bilinear terms $r_{nk}\varrho_{nk}$ and $r_{nk}\kappa_n$ can be linearized by introducing additional variables and applying standard linearization techniques, such as McCormick envelopes.