Supplementary Materials to "Learning Optimal Distributionally Robust Individualized Treatment Rules"

Weibin Mo, Zhengling Qi and Yufeng Liu*

Supplementary Materials

S.1 Explicit Forms of the Power Uncertainty Set

In this section, we study the explicit forms of the power uncertainty set $\mathcal{P}_c^k(\mathbb{P})$ on certain parameteric families of distributions, and how they depend on the DR-constant c and the power k. We first examine the family of Bernoulli distributions and the normal distributions, and show that their power uncertainty sets depend on c and k differently. Then the general exponential family will be discussed.

Example S.1 (Bernoulli Distributional Ball)**.** Consider two Bernoulli distributions $\mathbf{Bernoulli}(p)$ and $\mathbf{Bernoulli}(q)$ for some $p, q \in [0,1]$. We have $\left\|\frac{\mathrm{d}\mathbf{Bernoulli}(q)}{\mathrm{d}\mathbf{Bernoulli}(p)}\right\|_{L^k(\mathbf{Bernoulli}(p))} = \left[p\left(\frac{q}{p}\right)^k + (1-p)\left(\frac{1-q}{1-p}\right)^k\right]^{1/k}$. If $p \leqslant q$, then the above becomes $\frac{q}{p}\left[p + (1-p)\left(\frac{p(1-q)}{q(1-p)}\right)^k\right]^{1/k} \in [(q/p) \times p^{1/k}, q/p]$. If $p \geqslant q$, then the above becomes $\frac{1-q}{1-p}\left[p\left(\frac{q(1-p)}{p(1-q)}\right)^k + 1 - p\right]^{1/k} \in \left[\frac{1-q}{1-p} \times (1-p)^{1/k}, \frac{1-q}{1-p}\right]$. As $k \to +\infty$, the above both approach to $\frac{q}{p} \vee \frac{1-q}{1-p}$. For fixed p and every $k \in [1, +\infty)$, we have

$$\mathcal{P}_c^k(\mathbf{Bernoulli}(p)) \supseteq \{\mathbf{Bernoulli}(q) : q \in [0,1],\ 1 - c(1-p) \leqslant q \leqslant cp\},$$

*Weibin Mo and Zhengling Qi are co-first authors for the paper. Weibin Mo is a Ph.D. student, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. E-mail: harrymok@email.unc.edu. Zhengling Qi is Assistant Professor, Department of Decision Sciences, George Washington University, Washington, D.C. 20052, USA. E-mail: qizhengling@gwu.edu. Yufeng Liu is Professor, Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Science, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, NC 27599, USA. E-mail: yfliu@email.unc.edu.

and

$$\mathcal{P}_c^k(\mathbf{Bernoulli}(p))^\complement \supseteq \left\{\mathbf{Bernoulli}(q) : q \in [0,1],\ q > \frac{cp}{p^{1/k}} \text{ or } q < 1 - \frac{c(1-p)}{(1-p)^{1/k}}\right\},$$

with the meaningful $c \leqslant \frac{1}{p\wedge(1-p)}$. In particular as the large enough k increases while $1 < c \leqslant \frac{1}{p\wedge(1-p)}$ is fixed, $\mathcal{P}_c^k(\mathbf{Bernoulli}(p))$ contains less Bernoulli distributions, down to that of success probabilities in $[1-c(1-p), cp]$ only.

Example S.2 (Normal Distributional Ball of Mean Shifts)**.** Consider two p-dimensional normal distributions $\mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p)$ and $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{I}_p)$ for some center parameter $\boldsymbol{\mu} \in \mathbb{R}^p$. The density ratio of $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{I}_p)$ *w.r.t.* $\mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p)$ is given by $\frac{\exp(-\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2/2)}{\exp(-\|\boldsymbol{x}\|_2^2/2)} = e^{-\|\boldsymbol{\mu}\|_2^2/2} \times e^{\boldsymbol{\mu}^\intercal \boldsymbol{x}}$. Then the L^k-norm of the density ratio under $\mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p)$ can be calculated analytically as $e^{-\|\boldsymbol{\mu}\|_2^2/2}\left(\int_{\mathbb{R}^p} e^{k\boldsymbol{\mu}^\intercal\boldsymbol{x}} \times (2\pi)^{-p/2} e^{-\|\boldsymbol{x}\|_2^2/2}\mathrm{d}\boldsymbol{x}\right)^{1/k} = e^{(k-1)\|\boldsymbol{\mu}\|_2^2/2}$. Then $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{I}_p) \in \mathcal{P}_c^k(\mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p))$ if and only if $e^{(k-1)\|\boldsymbol{\mu}\|_2^2/2} \leqslant c \Leftrightarrow \|\boldsymbol{\mu}\|_2^2 \leqslant \frac{2\log c}{k-1}$.

Note that the conclusion is presented in terms of the L^2-difference of the mean vectors $\|\boldsymbol{\mu}\|_2$ between two normal components. It can be extended to two p-dimensional normal distributions of the same covariance matrix: $\mathcal{N}_p(\boldsymbol{\mu}_1, \Sigma) \in \mathcal{P}_c^k(\mathcal{N}_p(\boldsymbol{\mu}_0, \Sigma))$ if and only if $\exp\left\{\frac{k-1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^\intercal\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)\right\} \leqslant c \Leftrightarrow (\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^\intercal\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0) \leqslant \frac{2\log c}{k-1}$. Then we have

$$\mathcal{P}_c^k(\mathcal{N}_p(\boldsymbol{\mu}_0, \Sigma)) \supseteq \left\{\mathcal{N}_p(\boldsymbol{\mu}, \Sigma) : \boldsymbol{\mu} \in \mathbb{R}^p,\ (\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\intercal\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) \leqslant \frac{2\log c}{k-1}\right\},$$

and

$$\mathcal{P}_c^k(\mathcal{N}_p(\boldsymbol{\mu}_0, \Sigma))^\complement \supseteq \left\{\mathcal{N}_p(\boldsymbol{\mu}, \Sigma) : \boldsymbol{\mu} \in \mathbb{R}^p,\ (\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\intercal\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) > \frac{2\log c}{k-1}\right\}.$$

In particular as k increases with $c > 1$ fixed, $\mathcal{P}_c^k(\mathcal{N}_p(\boldsymbol{\mu}_0, \Sigma))$ contains less normal distributions of covariance matrix Σ.

Example S.3 (Normal Distributional Ball of Covariance Scales)**.** Consider two p-dimensional normal distributions $\mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p)$ and $\mathcal{N}_p(\mathbf{0}_p, \sigma^2\mathbf{I}_p)$ for some scale parameter $\sigma^2 > 0$. The density ratio of $\mathcal{N}_p(\mathbf{0}_p, \sigma^2\mathbf{I}_p)$ *w.r.t.* $\mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p)$ is given by $\frac{\sigma^{-p}\exp\{-\|\boldsymbol{x}\|_2^2/(2\sigma^2)\}}{\exp(-\|\boldsymbol{x}\|_2^2/2)} = \sigma^{-p}e^{-(\sigma^{-2}-1)\|\boldsymbol{x}\|_2^2/2}$. Then the L^k-norm of the density ratio under $\mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p)$ can be calculated analytically as $\sigma^{-p}\left(\int_{\mathbb{R}^p} e^{-k(\sigma^{-2}-1)\|\boldsymbol{x}\|_2^2/2} \times (2\pi)^{-p/2}e^{-\|\boldsymbol{x}\|_2^2/2}\mathrm{d}\boldsymbol{x}\right)^{1/k} = \sigma^{-p}[k(\sigma^{-2}-1)+1]^{-p/(2k)}$, which is a nonlinear function in σ^2 ranging in $(0, k^\star)$ and attaining the minimum at $\sigma^2 = 1$. Then $\mathcal{N}_p(\mathbf{0}_p, \sigma^2\mathbf{I}_p) \in \mathcal{P}_c^k(\mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p))$ if and only if $\sigma^{-p}[k(\sigma^{-2}-1)+1]^{-p/(2k)} \leqslant c \Leftrightarrow \underline{\sigma}_k^2(c) \leqslant \sigma^2 \leqslant \bar{\sigma}_k^2(c)$ where $\underline{\sigma}_k^2(c) \in (0,1)$ and $\bar{\sigma}_k^2(c) \in (1, k^\star)$ are the unique roots solving the nonlinear equation $\sigma^{-p}[k(\sigma^{-2}-1)+1]^{-p/(2k)} = c \Leftrightarrow \sigma^{-2k} - c^{2k/p}[k(\sigma^{-2}-1)+1] \overset{t:=\sigma^{-2}-1}{=\!=} (t+1)^k - c^{2k/p}(kt+1) = 0$ on the interval $t \in (c^{2k^\star/p}-1, +\infty) \Leftrightarrow \sigma^2 \in (0, c^{-2k^\star/p})$ and $t \in (-1/k, 0) \Leftrightarrow \sigma^2 \in (1, k^\star)$ respectively. In particular as k increases while c is fixed, the lower root $\underline{\sigma}_k^2(c)$ increases to 1 while the upper root $\bar{\sigma}_k^2(c)$ decreases to 1, so that $\mathcal{P}_c^k(\mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p))$ contains fewer and fewer distributions of the form $\mathcal{N}_p(\mathbf{0}_p, \sigma^2\mathbf{I}_p)$ with $\sigma^2 \in [\underline{\sigma}_k^2(c), \bar{\sigma}_k^2(c)]$.

The result is general if the mean vector $\mathbf{0}_p$ is replaced by any vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and the covariance matrix $\mathbf{I}_p$ is replaced by some positive semi-definite matrix Σ:

$$\mathcal{P}_c^k(\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)) \supseteq \left\{\mathcal{N}_p(\boldsymbol{\mu}, \sigma^2\Sigma) : \underline{\sigma}_k^2(c) \leqslant \sigma^2 \leqslant \bar{\sigma}_k^2(c)\right\},$$

and

$$\mathcal{P}_c^k(\mathcal{N}_p(\boldsymbol{\mu}, \Sigma))^\complement \supseteq \left\{\mathcal{N}_p(\boldsymbol{\mu}, \sigma^2\Sigma) : \sigma^2 < \underline{\sigma}_k^2(c) \text{ or } \sigma^2 > \bar{\sigma}_k^2(c)\right\}.$$

As an extension of the Bernoulli and the normal distribution, we can also consider the mixture of two fixed normal components.

Lemma S.1 (Upper Bound of the Mixture ϕ-Divergence)**.** *Suppose $\mathbb{P}_0, \mathbb{P}_1$ are probability distributions, $p, q \in [0,1]$. Denote $\mathbb{P}_p := p\mathbb{P}_1 + (1-p)\mathbb{P}_0$, $\mathbb{P}_q := q\mathbb{P}_1 + (1-q)\mathbb{P}_0$. Let $\phi \in \Phi$ be a legitimate divergence function. Then*

$$D_\phi(\mathbb{P}_q \| \mathbb{P}_p) \leqslant D_\phi\big(q\mathbb{P}_1 \big\| (1-p)\mathrm{d}\mathbb{P}_0\big) + D_\phi\big((1-q)\mathbb{P}_0 \big\| p\mathbb{P}_1\big).$$

Proof.

$$\begin{aligned}
&D_\phi(\mathbb{P}_q \| \mathbb{P}_p) \\
=&\int \phi\left(\frac{q\mathrm{d}\mathbb{P}_1 + (1-q)\mathrm{d}\mathbb{P}_0}{p\mathrm{d}\mathbb{P}_1 + (1-p)\mathrm{d}\mathbb{P}_0}\right)[p\mathrm{d}\mathbb{P}_1 + (1-p)\mathrm{d}\mathbb{P}_0] \\
=&\int \phi\left(\frac{(1-p)\mathrm{d}\mathbb{P}_0}{p\mathrm{d}\mathbb{P}_1 + (1-p)\mathrm{d}\mathbb{P}_0} \times \frac{q\mathrm{d}\mathbb{P}_1}{(1-p)\mathrm{d}\mathbb{P}_0} + \frac{p\mathrm{d}\mathbb{P}_1}{p\mathrm{d}\mathbb{P}_1 + (1-p)\mathrm{d}\mathbb{P}} \times \frac{(1-q)\mathrm{d}\mathbb{P}_0}{p\mathrm{d}\mathbb{P}_1}\right)[p\mathrm{d}\mathbb{P}_1 + (1-p)\mathrm{d}\mathbb{P}_0] \\
\overset{\text{Jensen}}{\leqslant}&\int \phi\left(\frac{q\mathrm{d}\mathbb{P}_1}{(1-p)\mathrm{d}\mathbb{P}_0}\right)(1-p)\mathrm{d}\mathbb{P}_0 + \int \phi\left(\frac{(1-q)\mathrm{d}\mathbb{P}_0}{p\mathrm{d}\mathbb{P}_1}\right)p\mathrm{d}\mathbb{P}_1 \\
=&D_\phi\big(q\mathbb{P}_1 \big\| (1-p)\mathrm{d}\mathbb{P}_0\big) + D_\phi\big((1-q)\mathbb{P}_0 \big\| p\mathbb{P}_1\big).
\end{aligned}$$

□

Remark S.1. The conclusion can be stated in terms of the k-th moment of the density ratio. Suppose $\mathbb{P}_0 \ll \mathbb{P}_1$ and $\mathbb{P}_1 \ll \mathbb{P}_0$. Then

$$\left\|\frac{\mathrm{d}\mathbb{P}_q}{\mathrm{d}\mathbb{P}_p}\right\|_{L^k(\mathbb{P}_p)}^k \leqslant (1-p)\left(\frac{q}{1-p}\right)^k \left\|\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_0}\right\|_{L^k(\mathbb{P}_0)}^k + p\left(\frac{1-q}{p}\right)^k \left\|\frac{\mathrm{d}\mathbb{P}_0}{\mathrm{d}\mathbb{P}_1}\right\|_{L^k(\mathbb{P}_1)}^k.$$

Remark S.2 (Mixture Normal Distributional Ball)**.** Consider two mixture normal distributions $\mathsf{GMM}_p(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0; \Sigma) := p\mathcal{N}_d(\boldsymbol{\mu}_1, \Sigma) + (1-p)\mathcal{N}_d(\boldsymbol{\mu}_0, \Sigma)$ and $\mathsf{GMM}_q(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0; \Sigma) := q\mathcal{N}_d(\boldsymbol{\mu}_1, \Sigma) + (1-q)\mathcal{N}_d(\boldsymbol{\mu}_0, \Sigma)$ with the same components and different mixture probabilities $p, q \in [0,1]$. Example S.2, Lemma S.1 and Example S.1 together imply that

$$\begin{aligned}
&\left\| \frac{\mathrm{d}\mathsf{GMM}_q(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0; \Sigma)}{\mathrm{d}\mathsf{GMM}_p(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0; \Sigma)} \right\|_{L^k(\mathsf{GMM}_p(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0; \Sigma))} \\
&\leqslant \left[(1-p)\left(\frac{q}{1-p}\right)^k + p\left(\frac{1-q}{p}\right)^k \right]^{1/k} \exp\left\{ \frac{k-1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\intercal \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \right\} \qquad (1) \\
&\leqslant \left(\frac{q}{1-p} \vee \frac{1-q}{p} \right) \exp\left\{ \frac{k-1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\intercal \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \right\}.
\end{aligned}$$

Consequently, if $c \geqslant \exp\left\{\frac{k-1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\intercal \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right\}$, then $\mathcal{P}_c^k\big(\mathsf{GMM}_p(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0; \Sigma)\big)$ contains all those $\mathsf{GMM}_q(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0; \Sigma)$ with mixture probability q such that $1 - c\exp\left\{\frac{k-1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\intercal \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right\} p \leqslant q \leqslant c\exp\left\{\frac{k-1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\intercal \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right\}(1-p)$. However, since the inequality (1) applies the Jensen Inequality to $(\cdot)^k$, the right hand side can be loose when k is large.

Next we proceed to discuss the exponential family in its abstract canonical form. Depending on the growth of the log-partition function, the power divergence might or might not increase with the power k. And consequently when the distributional constant is held fixed, the power uncertainty set $\mathcal{P}_c^k(\mathbb{P})$ might or might not vanish.

Example S.4 (Canonical Exponential Family Distributional Ball)**.** Consider a canonical parameterized exponential family with density as $f(\boldsymbol{x}; \boldsymbol{\eta}) = h(\boldsymbol{x})\exp(\langle \boldsymbol{\eta}, \boldsymbol{x}\rangle - A(\boldsymbol{\eta}))$ where $\boldsymbol{\eta} \in \mathbb{R}^p$ is the canonical parameter, $A(\boldsymbol{\eta}) = \log \int h(\boldsymbol{x}) e^{\langle \boldsymbol{\eta}, \boldsymbol{x}\rangle} \mathrm{d}\boldsymbol{x}$ is the log-partition function. Note that $A(\cdot + \boldsymbol{\eta}_0) - A(\boldsymbol{\eta}_0)$ is the logarithm of the moment generating function of the sufficient statistic. Then for fixed $\boldsymbol{\eta}_1, \boldsymbol{\eta}_0 \in \mathbb{R}^p$,

$$\begin{aligned}
\left\| \frac{f(\cdot; \boldsymbol{\eta}_1)}{f(\cdot; \boldsymbol{\eta}_0)} \right\|_{L^k(\boldsymbol{\eta}_0)}^k &= e^{-k[A(\boldsymbol{\eta}_1) - A(\boldsymbol{\eta}_0)] - A(\boldsymbol{\eta}_0)} \int h(\boldsymbol{x}) e^{\langle k(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_0) + \boldsymbol{\eta}_0, \boldsymbol{x}\rangle} \mathrm{d}\boldsymbol{x} \\
&= \exp\Big(A[k(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_0) + \boldsymbol{\eta}_0] - k[A(\boldsymbol{\eta}_1) - A(\boldsymbol{\eta}_0)] - A(\boldsymbol{\eta}_0) \Big). \qquad (2)
\end{aligned}$$

Note that the relationship of (2) and k depends on the functional form of the log-partition function $A(\cdot)$. In Example S.2, $A(\boldsymbol{\eta}) = \boldsymbol{\eta}^\intercal \Sigma \boldsymbol{\eta} + \log\det(\Sigma)$ is a quadratic function in the scaled mean vector $\boldsymbol{\eta} = \Sigma^{-1}\boldsymbol{\mu}$ as the canonical parameter (where the covariance matrix Σ is assumed known and fixed), and hence (2) is a quadratic function in k in the exponential, which coincides with the conclusion from Example S.2 that the L^k-norm of the density ratio is exponentially linear in k. In Example S.1, the partition function $A(\eta) = \log(1+e^{\eta}) = \eta + \log(1+e^{-\eta})$ is at most linear in the log-odd $\eta = \log\left(\frac{p}{1-p}\right)$ as the canonical parameter. Then the L^k-norm of the density ratio should be bounded when k varies.

In general, the L^k-norm of the density ratio of distributions from the exponential family increases with k if A is super-linear: $\frac{A(\boldsymbol{\eta})}{\|\boldsymbol{\eta}\|} \to +\infty$ as $\|\boldsymbol{\eta}\| \to +\infty$.

S.2 Implementation Details

To practically optimize the DR-ITR ψ-risk $\mathcal{R}^k_{c,\psi}(f)$ based on the empirical data, we first estimate the CTE function $\widehat{C}_n(\cdot)$ using flexible nonparametric techniques. Then we replace the CTE function $C(\cdot)$ by its estimate $\widehat{C}_n(\cdot)$, and the population expectation $\mathbb{E}$ by its empirical version $\mathbb{E}_n$. We solve the following joint minimization problem based on the training data:

$$\min_{f\in\mathcal{F},\eta}\left\{c\left[\mathbb{E}_n\left([\widehat{C}_n(\boldsymbol{X})-\eta]_+^{k^\star}\frac{\psi[f(\boldsymbol{X})]}{2}+[-\widehat{C}_n(\boldsymbol{X})-\eta]_+^{k^\star}\frac{\psi[-f(\boldsymbol{X})]}{2}\right)\right]^{1/k^\star}+\eta\right\}.$$

In this section, we discuss more implementation details of $k < +\infty$ and $k = +\infty$.

S.2.1 Optimization when $k < +\infty$

When $k < +\infty$ and $k^\star > 1$, the $k^\star$-moment makes the direct optimization more challenging. To reduce the power $1/k^\star$, we introduce the auxiliary variable $\lambda \geqslant 0$ and consider $(\cdot)^{1/k^\star} = \inf_{\lambda\geqslant 0}\left(\frac{(\cdot)}{k^\star\lambda^{k^\star-1}}+\frac{\lambda}{k}\right)$, where due to the AM-GM Inequality, $\frac{1}{k^\star}\big(\frac{(\cdot)}{\lambda^{k^\star-1}}+\underbrace{\lambda+\cdots+\lambda}_{k^\star-1}\big) \geqslant (\cdot)^{1/k^\star}$ with equality if and only if $\lambda = (\cdot)^{1/k^\star} > 0$. Then we consider the following joint objective to minimize:

$$L(f,\eta,\lambda) := \frac{c}{k^\star\lambda^{k^\star-1}}\mathbb{E}_n\left([\widehat{C}_n(\boldsymbol{X})-\eta]_+^{k^\star}\frac{\psi[f(\boldsymbol{X})]}{2}+[-\widehat{C}_n(\boldsymbol{X})-\eta]_+^{k^\star}\frac{\psi[-f(\boldsymbol{X})]}{2}\right)+\frac{c\lambda}{k}+\eta. \quad (3)$$

Note that the joint objective (3) as multiple sum-products of DC functions is difference-of-convex in (f,η,λ), but the DC representation can be messy. Instead of using a direct DC algorithm, we apply the BSUM algorithm (Razaviyayn et al., 2013) to alternatively optimize over (η,λ) and f respectively, where the the upper-bound of the objective in f is a convex majorant. Specifically, we fix a small $\epsilon > 0$ and alternatively implement the following two steps:

Step I: For fixed $\widehat{f}_t$, we implement the t-th step optimization of $(\widehat{\eta}_t, \widehat{\lambda}_t)$ by solving

$$\begin{cases}\widehat{\eta}_t \in \operatorname*{argmin}_{\eta\in\mathbb{R}}\left\{c\left[\mathbb{E}_n\left(\frac{\psi[\widehat{f}_t(\boldsymbol{X})]}{2}[\widehat{C}_n(\boldsymbol{X})-\eta]_+^{k^\star}+\frac{\psi[-\widehat{f}_t(\boldsymbol{X})]}{2}[-\widehat{C}_n(\boldsymbol{X})-\eta]_+^{k^\star}\right)\right]^{1/k^\star}+\eta\right\} \\ \widehat{\lambda}_t := \left[\mathbb{E}_n\left(\frac{\psi[\widehat{f}_t(\boldsymbol{X})]}{2}[\widehat{C}_n(\boldsymbol{X})-\widehat{\eta}_t]_+^{k^\star}+\frac{\psi[-\widehat{f}_t(\boldsymbol{X})]}{2}[-\widehat{C}_n(\boldsymbol{X})-\widehat{\eta}_t]_+^{k^\star}\right)\right]^{1/k^\star}\vee\underline{\lambda}\end{cases}. \quad (4)$$

The objective in η is univariate and continuously differentiable and can be minimized by any univariate solver. The $\underline{\lambda} > 0$ is a prespecified small constant such that the updated $\widehat{\lambda}_t$ is trimmed at $\underline{\lambda}$ from below for better numerical stability.

Step II: For fixed $(\widehat{f}_t, \widehat{\eta}_t, \widehat{\lambda}_t)$, we solve the $(t+1)$-th step $\widehat{f}_{t+1}$ by minimizing the following convex upper-bound over $\mathcal{F}$:

$$\widetilde{L}(f; \widehat{f}_t, \widehat{\eta}_t, \widehat{\lambda}_t) := \mathbb{E}_n\Bigg(\frac{c}{2k^\star \widehat{\lambda}_t^{k^\star-1}}[+\widehat{C}_n(\boldsymbol{X}) - \widehat{\eta}_t]_+^{k^\star}\widetilde{\psi}[+f(\boldsymbol{X}); +\widehat{f}_t(\boldsymbol{X})] + \frac{c}{2k^\star \widehat{\lambda}_t^{k^\star-1}}[-\widehat{C}_n(\boldsymbol{X}) - \widehat{\eta}_t]_+^{k^\star}\widetilde{\psi}[-f(\boldsymbol{X}); -\widehat{f}_t(\boldsymbol{X})]\Bigg),$$

where given $u_0 \in \mathbb{R}$, $\widetilde{\psi}(\cdot; u_0)$ is a first-order convex majorant of ψ expanded at u_0:

$$\widetilde{\psi}(u; u_0) := \psi_+(u) - \psi_-(u_0) - \psi'_-(u_0)(u - u_0); \quad u \in \mathbb{R}.$$

In particular for fixed u_0, $\widetilde{\psi}$ satisfies: 1) the majorization $\widetilde{\psi}(u; u_0) \geqslant \psi(u)$ with equality if $u = u_0$; 2) the convexity of $\widetilde{\psi}(\cdot; u_0)$; and 3) the first-order condition $\widetilde{\psi}'(u; u_0) = \psi'_+(u) - \psi'_-(u_0)$ and $\widetilde{\psi}'(u_0; u_0) = \psi'(u_0)$, where $\widetilde{\psi}'(u; u_0)$ is taken over u. To organize the computation, define

$$Z_t^{(\pm)} := \frac{c}{2k^\star \widehat{\lambda}_t^{k^\star-1}}[\pm\widehat{C}_n(\boldsymbol{X}) - \widehat{\eta}_t]_+^{k^\star}; \quad S_t := Z_t^{(+)}\psi'_-[+\widehat{f}_t(\boldsymbol{X})] - Z_t^{(-)}\psi'_-[-\widehat{f}_t(\boldsymbol{X})]. \tag{5}$$

Then at the t-th step, we only need to keep track of $Z_t^{(\pm)}, S_t$ and minimize

$$\widetilde{L}(f; Z_t^{(\pm)}, S_t) := \mathbb{E}_n\left(Z_t^{(+)}\psi_+[+f(\boldsymbol{X})] + Z_t^{(-)}\psi_+[-f(\boldsymbol{X})] - S_t \times f(\boldsymbol{X})\right), \tag{6}$$

over $\mathcal{F}$. We summarize the algorithm for learning the DR-ITR when $k < +\infty$ in Algorithm 1.

Algorithm 1: Learning the DR-ITR ($k < +\infty$)

1 **Input:** Data $\{\boldsymbol{X}_i, \widehat{C}_n(\boldsymbol{X}_i)\}_{i=1}^n$, initial $\widehat{f}_0 \in \mathcal{F}$, $c \geqslant 1$, $\underline{\lambda} > 0$, and tolerance $\epsilon_{\text{tol}} > 0$.

2 Repeat for $t = 0, 1, \cdots$, do until $|\widehat{f}_{t+1} - \widehat{f}_t| \leqslant (|\widehat{f}_t| \vee 1)\epsilon_{\text{tol}}$:

3 Solve $(\widehat{\eta}_t, \widehat{\lambda}_t)$ by (4);

4 Update $(Z_t^{(\pm)}, S_t)$ as in (5);

5 Solve $\widehat{f}_{t+1}$ by optimizing the objective $\widetilde{L}(\cdot; Z_t^{(\pm)}, S_t)$ as in (6);

6 **Output:** $\widehat{f}_{t+1}$.

S.2.2 Optimization when $k = +\infty$

For $k = +\infty$ and $c > 1$, it is possible that the BSUM algorithm introduced in Algorithm 1 suffers potential convergence problems when the minimizer $\widehat{\eta}_t$ given $\widehat{f}_t$ in (4) is non-unique. Following Qi et al. (2019, Proposition 3.1), we see that the joint objective is minimized with respect to η at one of the $2n$ knots $\{\eta_j^*\}_{j=1}^{2n} := \{\pm\widehat{C}_n(\boldsymbol{X}_i)\}_{i=1}^{n}$. Then the joint minimization problem boils down to

$$\min_{f\in\mathcal{F}} \min_{1\leqslant j\leqslant 2n} \left\{L_j(f) := \frac{c}{2}\mathbb{E}_n\left([\widehat{C}_n(\boldsymbol{X}) - \eta_j^*]_+\psi[+f(\boldsymbol{X})] + [-\widehat{C}_n(\boldsymbol{X}) - \eta_j^*]_+\psi[-f(\boldsymbol{X})]\right) + \eta_j^*\right\}.$$

That is, the minimization with respect to η can attain at only finitely many candidates $\{\eta_j^*\}_{j=1}^{2n}$. For $1 \leqslant j \leqslant 2n$, we define the convex upper bound of L_j at f_0 as

$$\widetilde{L}_j(f; f_0) := \mathbb{E}_n\left(\frac{c}{2}[+\widehat{C}_n(\boldsymbol{X}) - \eta_j^*]_+\widetilde{\psi}[+f(\boldsymbol{X}); +f_0(\boldsymbol{X})] + \frac{c}{2}[-\widehat{C}_n(\boldsymbol{X}) - \eta_j^*]_+\widetilde{\psi}[-f(\boldsymbol{X}); -f_0(\boldsymbol{X})]\right),$$

where $\widetilde{\psi}$ is the first-order convex majorant of ψ as before. Then the previously discussed BSUM algorithm iteratively updates the following two steps: (I) for fixed $\widehat{f}_t$, solve for the t-th step $\widehat{j}_t \in \operatorname{argmin}_{1\leqslant j\leqslant 2n} L_j(\widehat{f}_t)$; (II) for fixed $(\widehat{f}_t, \widehat{j}_t)$, solve for the $(t+1)$-th step $\widehat{f}_{t+1}$ by minimizing $\widetilde{L}_{\widehat{j}_t}(\cdot; \widehat{f}_t)$. Notice that the non-uniqueness of the minimizer $\widehat{\eta}_t$ given $\widehat{f}_t$ now becomes the non-uniqueness of the index $\widehat{j}_t$.

To overcome the difficulty due to the non-uniqueness, Pang et al. (2016, Section 5) showed that the following two requirements should be met to ensure the convergence to stationarity: (1) minimizing the surrogate function $\widetilde{L}_{\widehat{j}_t}(\cdot; \widehat{f}_t)$ of the chosen index $\widehat{j}_t$ should let the true objective function L descend the most; (2) the most descent requirement (1) holds with respect to the indices chosen among the following ϵ-argmin index set for some fixed $\epsilon > 0$:

$$\mathscr{M}_\epsilon(\widehat{f}_t) := \left\{1 \leqslant j \leqslant 2n : L_j(\widehat{f}_t) \leqslant \min_{1\leqslant l\leqslant 2n} L_l(\widehat{f}_t) + \epsilon\right\}, \tag{7}$$

rather than the traditional argmin index set $\mathscr{M}_0(\widehat{f}_t)$. To avoid too many surrogate functions to be minimized at each step, Pang et al. (2016, Section 5.2) proposed to randomly choose $\widehat{j}_t \in \mathscr{M}_\epsilon(\widehat{f}_t)$ with a positive probability, so that at least for some positive chance the most descent index can be picked. To ensure the true objective is strictly decreasing, we accept the minimizer $\widetilde{f}_{t+1} \in \operatorname{argmin}_f \widetilde{L}_{\widehat{j}_t}(f; \widehat{f}_t)$ only when $\widetilde{L}_{\widehat{j}_t}(\widetilde{f}_{t+1}; \widehat{f}_t) \leqslant L(\widehat{f}_t)$, or equivalently,

$$L_{\widehat{j}_t}(\widehat{f}_t) - \min_{1\leqslant j\leqslant 2n} L_j(\widehat{f}_t) \leqslant \widetilde{L}_{\widehat{j}_t}(\widehat{f}_t; \widehat{f}_t) - \widetilde{L}_{\widehat{j}_t}(\widetilde{f}_{t+1}; \widehat{f}_t).$$

That is, the descent in terms of the surrogate objective $\widetilde{L}_{\widehat{j}_t}(\cdot; \widehat{f}_t)$ is no less than the excess value (up to ϵ) of the chosen $\widehat{j}_t$-th objective $L_{\widehat{j}_t}$ at $\widehat{f}_t$.

To organize the computation, we again define for $1 \leqslant j \leqslant 2n$ and f_0

$$Z_j^{(\pm)} := \frac{c}{2}[\pm \widehat{C}_n(\boldsymbol{X}) - \eta_j^*]_+; \quad S_j(f_0) := Z_j^{(+)}\psi'_-[+f_0(\boldsymbol{X})] - Z_j^{(-)}\psi'_-[-f_0(\boldsymbol{X})], \tag{8}$$

similarly as in (5), but with the index t replaced by j. Then at the t-th step, we first randomly pick $\widehat{j}_t \in \mathscr{M}_\epsilon(\widehat{f}_t)$ uniformly and keep the excess value $\epsilon_t := L_{\widehat{j}_t}(\widehat{f}_t) - \min_{1\leqslant j\leqslant 2n} L_j(\widehat{f}_t)$. Then we keep $Z_t := Z_{\widehat{j}_t}^{(\pm)}$ and $S_t := S_{\widehat{j}_t}(\widehat{f}_t)$ and minimize $\widetilde{L}(\cdot; Z_t^{(\pm)}, S_t)$ as in (6). Finally, we accept the minimizer $\widetilde{f}_{t+1} \in \operatorname{argmin}_f \widetilde{L}(f; Z_t^{(\pm)}, S_t)$ if $\widetilde{L}(\widehat{f}_t; Z_t^{(\pm)}, S_t) - \widetilde{L}(\widetilde{f}_{t+1}; Z_t^{(\pm)}, S_t) \geqslant \epsilon_t$. We summarize the algorithm for learning the DR-ITR when $k = +\infty$ in Algorithm 2.

Algorithm 2: Learning the DR-ITR ($k = +\infty$)

Input: Data $\{\boldsymbol{X}_i, \widehat{C}_n(\boldsymbol{X}_i)\}_{i=1}^n$, initial $\widehat{f}_0 \in \mathcal{F}$, $c > 1$, $\epsilon > 0$, and tolerance $\epsilon_{\text{tol}} > 0$.

For $t = 0, 1, \cdots$, do until $\|\widehat{f}_{t+1} - \widehat{f}_t\| \leqslant (\|\widehat{f}_t\| \vee 1)\epsilon_{\text{tol}}$:

 Choose $\widehat{j}_t \in \mathscr{M}_\epsilon(\widehat{f}_t)$ in (7) uniformly and randomly, and keep $\epsilon_t := L_{\widehat{j}_t}(\widehat{f}_t) - \min_{1\leqslant j\leqslant 2n} L(\widehat{f}_t)$;

 Update $Z_t^{(\pm)} = Z_{\widehat{j}_t}^{(\pm)}$ and $S_t = S_{\widehat{j}_t}(\widehat{f}_t)$ as in (8);

 Solve $\widetilde{f}_{t+1}$ by optimizing the objective $\widetilde{L}(\cdot; Z_t^{(\pm)}, S_t)$ as in (6);

 If $\widetilde{L}(\widehat{f}_t; Z_t^{(\pm)}, S_t) - \widetilde{L}(\widetilde{f}_{t+1}; Z_t^{(\pm)}, S_t) \geqslant \epsilon_t$, then set $\widehat{f}_{t+1} = \widetilde{f}_{t+1}$; otherwise, set $\widehat{f}_{t+1} = \widehat{f}_t$.

Output: $\widehat{f}_{t+1}$.

S.3 Technical Proofs

S.3.1 Proof of Lemma 2

(I) follows from direct calculation. Now we admit (I) and prove (II). First notice that

$$\begin{aligned}
\lambda\phi_k^\star(z/\lambda) &= \frac{(k-1)^{k^\star}/k}{\lambda^{1/(k-1)}}\left(z - \eta + \frac{\lambda}{k-1}\right)_+^{k^\star} - \frac{\lambda}{k}, \\
\nabla\phi_k^\star(z/\lambda) &= \frac{(k-1)^{1/(k-1)}}{\lambda^{1/(k-1)}}\left(z - \eta + \frac{\lambda}{k-1}\right)_+^{1/(k-1)}.
\end{aligned}$$

Now using the (6)-R.H.S., the Cressie-Read family defining worst-case expectation is further solved by

$$\begin{aligned}
&\min_{\lambda\geqslant 0, \eta\in\mathbb{R}} \; [(k-1)^{k^\star}/k] \times \lambda^{-1/(k-1)}\mathbb{E}_{\mathbb{P}}\left(Z - \eta + \tfrac{\lambda}{k-1}\right)_+^{k^\star} + \lambda\left(\rho - \tfrac{1}{k}\right) + \eta, \quad W^* = \tfrac{(k-1)^{1/(k-1)}}{(\lambda^*)^{1/(k-1)}}\left(Z - \eta^* + \tfrac{\lambda^*}{k-1}\right)_+^{1/(k-1)}, \\
\Leftrightarrow \; &\min_{\lambda\geqslant 0, \eta\in\mathbb{R}} \; [(k-1)^{k^\star}/k] \times \lambda^{-1/(k-1)}\mathbb{E}_{\mathbb{P}}(Z - \eta)_+^{k^\star} + \lambda\left(\rho + \tfrac{1}{k(k-1)}\right) + \eta, \quad W^* = \tfrac{(k-1)^{1/(k-1)}}{(\lambda^*)^{1/(k-1)}}(Z - \eta^*)_+^{1/(k-1)}.
\end{aligned} \tag*{1}$$

where (λ, η) can be optimized stagewise.

Fix $\eta \in \mathbb{R}$.

$$
\begin{aligned}
& [(k-1)^{k^\star}/k] \times \lambda^{-1/(k-1)} \mathbb{E}_{\mathbb{P}}(Z-\eta)_+^{k^\star} + \lambda \left(\rho + \frac{1}{k(k-1)}\right) \\
= \quad & (k-1) \times \lambda^{-1/(k-1)} \left(\frac{(k-1)^{1/(k-1)}}{k}\right) \mathbb{E}_{\mathbb{P}}(Z-\eta)_+^{k^\star} + \lambda \left(\rho + \frac{1}{k(k-1)}\right) \\
\geqslant \quad & k \left[\left(\frac{(k-1)^{1/(k-1)}}{k}\right)^{k-1} [\mathbb{E}_{\mathbb{P}}(Z-\eta)_+^{k^\star}]^{k-1} \left(\rho + \frac{1}{k(k-1)}\right)\right]^{1/k} \qquad \text{(by AM-GM Inequality)} \\
= \quad & [k(k-1)\rho + 1]^{1/k} [\mathbb{E}_{\mathbb{P}}(Z-\eta)_+^{k^\star}]^{1/k^\star}.
\end{aligned}
$$

Denote $c_k(\rho) := [k(k-1)\rho + 1]^{1/k}$. Then the objective in η becomes

$$
\min_{\eta \in \mathbb{R}} \quad c_k(\rho)[\mathbb{E}_{\mathbb{P}}(Z-\eta)_+^{k^\star}]^{1/k^\star} + \eta.
$$

S.3.2 Proof of Proposition 4

Define

$$
\widetilde{C}_{\eta,\lambda}^{(\pm)}(\boldsymbol{X}) := \frac{c}{k^\star \lambda^{k^\star - 1}} \mathbb{E}\left([\pm C(\boldsymbol{X}) - \eta]_+^{k^\star} \middle| \boldsymbol{X}\right), \quad \widetilde{C}_{\eta,\lambda}(\boldsymbol{X}) := \widetilde{C}_{\eta,\lambda}^{(+)}(\boldsymbol{X}) - \widetilde{C}_{\eta,\lambda}^{(-)}(\boldsymbol{X}).
$$

Then by conditioning on $\boldsymbol{X}$,

$$
\begin{aligned}
\mathcal{L}_c^k(f, \eta, \lambda) &= \mathbb{E}\left(\widetilde{C}_{\eta,\lambda}^{(+)}(\boldsymbol{X}) \mathbb{1}[f(\boldsymbol{X}) < 0] + \widetilde{C}_{\eta,\lambda}^{(-)}(\boldsymbol{X}) \mathbb{1}[f(\boldsymbol{X}) > 0]\right) + \frac{c\lambda}{k} + \eta, \\
\mathcal{L}_{c,\psi}^k(f, \eta, \lambda) &= \mathbb{E}\left(\widetilde{C}_{\eta,\lambda}^{(+)}(\boldsymbol{X}) \frac{\psi[f(\boldsymbol{X})]}{2} + \widetilde{C}_{\eta,\lambda}^{(-)}(\boldsymbol{X}) \frac{\psi[-f(\boldsymbol{X})]}{2}\right) + \frac{c\lambda}{k} + \eta.
\end{aligned}
$$

(I) (Fisher Consistency) Notice that for our ramp surrogate loss ψ, $f \geqslant 1$ implies that $\frac{\psi(f)}{2} = 0$, and $f \leqslant -1$ implies that $\frac{\psi(f)}{2} = 1$. Then without loss of generality, we might restrict to consider $f \in [-1, 1]$ for which $f = 1$ if and only if $\frac{\psi(f)}{2} = 0$ and $f = -1$ if and only if $\frac{\psi(f)}{2} = 1$. Then for fixed $\boldsymbol{x} \in \mathcal{X}$,

$$
\begin{aligned}
& \min_{f \in \{\pm 1\}} \left\{\widetilde{C}_{\eta,\lambda}^{(+)}(\boldsymbol{x}) \mathbb{1}(f < 0) + \widetilde{C}_{\eta,\lambda}^{(-)}(\boldsymbol{x}) \mathbb{1}(f > 0)\right\} \\
= & \min_{f \in [-1,1]} \left\{\widetilde{C}_{\eta,\lambda}^{(+)}(\boldsymbol{x}) \frac{\psi(f)}{2} + \widetilde{C}_{\eta,\lambda}^{(-)}(\boldsymbol{x}) \frac{\psi(-f)}{2}\right\} \\
= & \widetilde{C}_{\eta,\lambda}^{(+)}(\boldsymbol{x}) \wedge \widetilde{C}_{\eta,\lambda}^{(-)}(\boldsymbol{x})
\end{aligned}
$$

attained at the common function value $f_{\eta,\lambda}^*(\boldsymbol{x}) := \mathbf{sign}[\widetilde{C}_{\eta,\lambda}(\boldsymbol{x})]$. Define $\mathcal{L}_c^{k,*}(\eta, \lambda) :=$

$\mathcal{L}_c^k(f^*_{\eta,\lambda}, \eta, \lambda) = \mathbb{E}[\widetilde{C}^{(+)}_{\eta,\lambda}(\boldsymbol{X}) \wedge \widetilde{C}^{(-)}_{\eta,\lambda}(\boldsymbol{X})] + \frac{c\lambda}{k} + \eta$. Then

$$\min_{f:\mathcal{X}\to\{\pm 1\}} \mathcal{L}_c^k(f, \eta, \lambda) = \min_{f:\mathcal{X}\to[-1,1]} \mathcal{L}_{c,\psi}^k(f, \eta, \lambda) = \mathcal{L}_c^{k,*}(\eta, \lambda),$$
$$\operatorname*{argmin}_{f:\mathcal{X}\to\{\pm 1\}} \mathcal{L}_c^k(f, \eta, \lambda) = \operatorname*{argmin}_{f:\mathcal{X}\to[-1,1]} \mathcal{L}_{c,\psi}^k(f, \eta, \lambda) = f^*_{\eta,\lambda}(\boldsymbol{X}) \quad a.s.$$

(II) (Excess Risk) For fixed $f : \mathcal{X} \to \mathbb{R}$,

$$\mathcal{L}_c^k(f, \eta, \lambda) - \mathcal{L}_c^{k,*}(\eta, \lambda) = \mathbb{E}\left[\widetilde{C}_{\eta,\lambda}(\boldsymbol{X}) \times \left(\mathbb{1}[f(\boldsymbol{X}) < 0] - \mathbb{1}[f^*_{\eta,\lambda}(\boldsymbol{X}) < 0]\right) + \widetilde{C}^{(-)}_{\eta,\lambda}(\boldsymbol{X})\right],$$
$$\mathcal{L}_{c,\psi}^k(f, \eta, \lambda) - \mathcal{L}_c^{k,*}(\eta, \lambda) = \mathbb{E}\left[\widetilde{C}_{\eta,\lambda}(\boldsymbol{X}) \times \frac{\psi[f(\boldsymbol{X})] - \psi[f^*_{\eta,\lambda}(\boldsymbol{X})]}{2} + \widetilde{C}^{(-)}_{\eta,\lambda}(\boldsymbol{X})\right],$$

where the second equation follows from the fact that $\psi(u) + \psi(-u) \equiv 2$. For fixed $\boldsymbol{x} \in \mathcal{X}$, if $\widetilde{C}_{\eta,\lambda}(\boldsymbol{x}) > 0$, then $f^*_{\eta,\lambda}(\boldsymbol{x}) = 1$, and

$$\mathbb{1}[f(\boldsymbol{x}) < 0] - \mathbb{1}[f^*_{\eta,\lambda}(\boldsymbol{x}) < 0] = \mathbb{1}[f(\boldsymbol{x}) < 0] \leqslant 2 \times \frac{\psi[f(\boldsymbol{x})]}{2} = 2 \times \frac{\psi[f(\boldsymbol{x})] - \psi[f^*_{\eta,\lambda}(\boldsymbol{x})]}{2};$$

otherwise if $\widetilde{C}_{\eta,\lambda}(\boldsymbol{x}) < 0$, then $f^*_{\eta,\lambda}(\boldsymbol{x}) = -1$, and

$$2 \times \frac{\psi[f(\boldsymbol{x})] - \psi[f^*_{\eta,\lambda}(\boldsymbol{x})]}{2} = -\psi[-f(\boldsymbol{x})] \leqslant -\mathbb{1}[-f(\boldsymbol{x}) \leqslant 0] = \mathbb{1}[f(\boldsymbol{x}) < 0] - \mathbb{1}[f^*_{\eta,\lambda}(\boldsymbol{x}) < 0].$$

Therefore,

$$\mathcal{L}_c^k(f, \eta, \lambda) - \mathcal{L}_c^{k,*}(\eta, \lambda) \leqslant 2[\mathcal{L}_{c,\psi}^k(f, \eta, \lambda) - \mathcal{L}_c^{k,*}(\eta, \lambda)].$$

Finally, by rearranging $\mathcal{L}_c^{k^\star}(\eta, \lambda)$ to the same side and infimizing its $(\eta, \lambda) \in \mathbb{R} \times \mathbb{R}_+$, we have

$$\mathcal{L}_c^k(f, \eta, \lambda) \leqslant 2\mathcal{L}_{c,\psi}^k(f, \eta, \lambda) - \mathcal{L}_c^{k,*}(\eta, \lambda) \leqslant 2\mathcal{L}_{c,\psi}^k(f, \eta, \lambda) - \mathcal{R}_c^{k,*}$$
$$\Leftrightarrow \quad \mathcal{L}_c^k(f, \eta, \lambda) - \mathcal{R}_c^{k,*} \leqslant 2[\mathcal{L}_{c,\psi}^k(f, \eta, \lambda) - \mathcal{R}_c^{k,*}].$$

And by partially infimizing $(\eta, \lambda) \in \mathbb{R} \times \mathbb{R}_+$ on both sides, we have

$$\mathcal{R}_c^k(f) - \mathcal{R}_c^{k,*} \leqslant 2[\mathcal{R}_{c,\psi}^k(f) - \mathcal{R}_c^{k,*}].$$

S.3.3 Proof of Proposition 5

By Assumption 4, without loss of generality, we also assume that Assumptions 2 and 3 also hold for $\{\widehat{C}_n(\boldsymbol{X})\}_{n\in\mathbb{N}}$ uniformly.

First assume $k < +\infty$ and $k^\star > 1$. We first provide a few boundedenss results implied by Assumption 2. For $f : \mathcal{X} \to \mathbb{R}$, define

$$\eta_f^* := \operatorname*{argmin}_{\eta \in \mathbb{R}} \left\{ c \left[\mathbb{E} \left(\frac{\psi[f(\boldsymbol{X})]}{2} [C(\boldsymbol{X}) - \eta]_+^{k^\star} + \frac{\psi[-f(\boldsymbol{X})]}{2} [-C(\boldsymbol{X}) - \eta]_+^{k^\star} \right) \right]^{1/k^\star} + \eta \right\}, \tag{9}$$

$$\lambda_f^* := \left[\mathbb{E} \left(\frac{\psi[f(\boldsymbol{X})]}{2} [C(\boldsymbol{X}) - \eta_f^*]_+^{k^\star} + \frac{\psi[-f(\boldsymbol{X})]}{2} [-C(\boldsymbol{X}) - \eta_f^*]_+^{k^\star} \right) \right]^{1/k^\star}. \tag{10}$$

By Assumption 2 that $|C(\boldsymbol{X})| \leqslant M$, the optimal objective of (9) is bounded from above by $\min_{\eta \in \mathbb{R}} \{c(M - \eta)_+ + \eta\} = M$. And for any fixed $\eta \in \mathbb{R}$, the objective of (9) is bounded from below by $c(-M - \eta)_+ + \eta$. Then by the optimality of η_f^*, we have $c(-M - \eta_f^*)_+ + \eta_f^* \leqslant M \Leftrightarrow -\frac{c+1}{c-1} M \leqslant \eta_f^* \leqslant M$.

As for λ_f^*, since the optimal value of (9) is $c\lambda_f^* + \eta_f^*$, we have $c\lambda_f^* + \eta_f^* \leqslant M \Rightarrow \lambda_f^* \leqslant \frac{M - \eta_f^*}{c} \leqslant \frac{2M}{c-1}$. On the other hand, we need to elaborate more to give the lower bound (away from 0) on λ_f^*. The following lemma is a useful tool to motivate our analysis.

Lemma S.2. *Suppose Z is a bounded random variable, $k \geqslant 1$, $c \geqslant 1$. Define*

$$\eta^* := \operatorname*{argmin}_{\eta \in \mathbb{R}} \left\{ c[\mathbb{E}(Z - \eta)_+^k]^{1/k} + \eta \right\}.$$

Then $\mathbb{P}(Z \geqslant \eta^*) \geqslant c^{-k}$.

Proof. For $k = 1$, η^* as the VaR (Krokhmal, 2007) can be obtained by $\eta^* = \inf\{\eta \in \mathbb{R} : \mathbb{P}(Z \leqslant \eta) \geqslant 1 - c^{-1}\}$. Then for any $\epsilon > 0$, $\mathbb{P}(Z \leqslant \eta - \epsilon) < 1 - c^{-1} \Leftrightarrow \mathbb{P}(Z > \eta - \epsilon) \geqslant c^{-1}$. Let $\epsilon \to 0^+$ and by upper semi-continuity, we have $\mathbb{P}(Z \geqslant \eta^*) \geqslant c^{-1}$.

Suppose $k > 1$. If $\mathbb{P}(Z = \operatorname{ess.sup} Z) \geqslant c^{-k}$, then by $\eta^* \leqslant \operatorname{ess.sup} Z$, $\mathbb{P}(Z \geqslant \eta^*) \geqslant \mathbb{P}(Z = \operatorname{ess.sup} Z) \geqslant c^{-k}$ holds trivially. Now assume $\mathbb{P}(Z = \operatorname{ess.sup} Z) < c^{-k}$. By lower semi-continuity, there exists $\epsilon_0 > 0$, such that for any $0 \leqslant \epsilon \leqslant \epsilon_0$, $\mathbb{P}(Z \geqslant \operatorname{ess.sup} Z - \epsilon) < c^{-k}$. Then

$$c[\mathbb{E}(Z - \operatorname{ess.sup} Z + \epsilon)_+^k]^{1/k} + \operatorname{ess.sup} Z - \epsilon \leqslant c\epsilon\mathbb{P}(Z \geqslant \operatorname{ess.sup} - \epsilon)^{1/k} + \operatorname{ess.sup} Z - \epsilon < \operatorname{ess.sup} Z.$$

As a result, $\eta^* < \operatorname{ess.sup} Z - \epsilon_0$, hence

$$\mathbb{E}(Z - \eta^*)_+^k \geqslant (\epsilon_0/2)^k \mathbb{P}(Z \geqslant \eta^* + \epsilon_0/2) \geqslant (\epsilon_0/2)^k \mathbb{P}(Z \geqslant \operatorname{ess.sup} Z - \epsilon_0/2) > 0.$$

Finally, the first order condition for η^* is given by

$$-\frac{c\mathbb{E}(Z - \eta^*)_+^{k-1}}{\mathbb{E}[(Z - \eta^*)_+^k]^{1-1/k}} + 1 = 0 \quad \Leftrightarrow \quad \frac{\|(Z - \eta^*)_+\|_{L^{k-1}}}{\|(Z - \eta^*)_+\|_{L^k}} = c^{-\frac{1}{k-1}}.$$

On the other hand, by Hölder Inequality,

$$\mathbb{E}(Z-\eta^*)_+^{k-1} = \mathbb{E}[(Z-\eta^*)_+^{k-1}\mathbb{1}(Z \geqslant \eta^*)] \leqslant [\mathbb{E}(Z-\eta^*)_+^k]^{\frac{k-1}{k}}\mathbb{P}(Z \geqslant \eta^*)^{1/k}.$$

We have

$$c^{-\frac{1}{k-1}} = \frac{\|(Z-\eta^*)_+\|_{L^{k-1}}}{\|(Z-\eta^*)_+\|_{L^k}} \leqslant \mathbb{P}(Z \geqslant \eta^*)^{1/[k(k-1)]} \quad \Leftrightarrow \quad \mathbb{P}(Z \geqslant \eta^*) \geqslant c^{-k}.$$

□

Next, we introduce the sign variable $\zeta_\psi(f) \in \{\pm 1\}$ such that $\mathbb{P}[\zeta_\psi(f) = \pm 1|\boldsymbol{X}] = \frac{\psi[\pm f(\boldsymbol{X})]}{2}$. Then $\eta_f^* \in \operatorname{argmin}_{\eta\in\mathbb{R}}\left\{c\left(\mathbb{E}[C(\boldsymbol{X})\zeta_\psi(f)-\eta]_+^{k^\star}\right)^{1/k^\star}+\eta\right\}$. By Lemma S.2, we immediately have $\mathbb{P}[C(\boldsymbol{X})\zeta_\psi(f) \geqslant \eta_f^*] \geqslant c^{-k}$. Next by Assumption 3, $C(\boldsymbol{X})$ has uniformly bounded density h with respect to the Lebesgue measure. Then $C(\boldsymbol{X})\zeta_\psi(f)$ also has density $h_{\psi,f}(c) \leqslant h(c) \vee h(-c)$ with respect to the Lebesgue measure, and $h_{\psi,f}$ is uniformly bounded as well: $\|h_{\psi,f}\|_\infty \leqslant \|h\|_\infty < +\infty$. Then for any fixed $\underline{c} \leqslant \bar{c}$, we have $\mathbb{P}\{C(\boldsymbol{X})\zeta_\psi(f) \in [\underline{c},\bar{c}]\} \leqslant (\bar{c}-\underline{c})\|h\|_\infty$. In particular, for any $t>0$,

$$\mathbb{P}[C(\boldsymbol{X})\zeta_\psi(f) \geqslant \eta_f^* + t] \geqslant c^{-k} - t\|h\|_\infty.$$

In particular, by taking $t := 1/(2\|h\|_\infty c^k)$, we have

$$\begin{aligned}\lambda_f^* &= \left(\mathbb{E}[C(\boldsymbol{X})\zeta_\psi(f)-\eta_f^*]_+^{k^\star}\right)^{1/k^\star}\\ &\geqslant 1/(2\|h\|_\infty c^k)\mathbb{P}[C(\boldsymbol{X})\zeta_\psi(f) \geqslant \eta_f^* + 1/(2\|h\|_\infty c^k)]^{1/k^\star}\\ &\geqslant 1/(2\|h\|_\infty c^k)[c^{-k} - 1/(2\|h\|_\infty c^k) \times \|h\|_\infty]^{1/k^\star}\\ &= 1/(2^{(2k-1)/k}\|h\|_\infty c^{2k-1}) > 0,\end{aligned}$$

which decreases in c of order $c^{-(2k-1)}$. Note that the lower bound on λ_f^* depends on the order k, while its upper bound doesn't. In particular, as k increases, the vanishing rate of λ_f^* as $c \to +\infty$ gets faster.

We conclude the preceding boundedness results by denoting $\eta_f^* \in [\underline{\eta},\bar{\eta}] := \left[-\frac{c+1}{c-1}M, M\right]$ and $\lambda_f^* \in [\underline{\lambda},\bar{\lambda}] := \left[\frac{1}{2^{(2k-1)/k}\|h\|_\infty c^{2k-1}}, \frac{2M}{c-1}\right]$. Note that all those bounds above also hold when $\mathbb{E}$ is replaced by $\mathbb{E}_n$, $\frac{\psi(\cdot)}{2}$ is replaced by $\mathbb{1}(\cdot < 0)$, and $C(\cdot)$ is replaced by $\widehat{C}_n(\cdot)$. As an immediate result, we further have boundedness $\ell_c^k, \ell_{c,\psi}^k \in [\underline{\ell}_c^k, \overline{\ell}_c^k]$ where $\underline{\ell}_c^k := \frac{c}{k}\underline{\lambda} + \underline{\eta}$, and

$$\begin{aligned}
\overline{\ell}_c^k &:= \max_{(\eta,\lambda)\in\{\underline{\eta},\bar{\eta}\}\times\{\underline{\lambda},\bar{\lambda}\}} \left\{ \frac{c}{k^\star \lambda^{k^\star-1}}(M-\eta)^{k^\star} + \frac{c\lambda}{k} + \eta \right\} \\
&= \max \begin{cases} \frac{2}{k}\frac{c}{c-1}M + M, & \eta = \bar{\eta},\ \lambda = \bar{\lambda}; \\ \frac{2^{2k^\star-1/k^\star}\|h\|_\infty^{k^\star-1}}{k^\star}\frac{c^{2k^\star+2}}{(c-1)^{k^\star}}M^{k^\star} + \frac{1}{k2^{1+1/k^\star}\|h\|_\infty}\frac{1}{c^{2/(k^\star-1)}} - \frac{c+1}{c-1}M, & \eta = \underline{\eta},\ \lambda = \underline{\lambda}; \\ \frac{2}{k^\star}\frac{c^{k^\star+1}}{c-1}M + \frac{2}{k}\frac{c}{c-1}M - \frac{c+1}{c-1}M, & \eta = \underline{\eta},\ \lambda = \bar{\lambda}. \end{cases}
\end{aligned}$$

Notice that as $c, (c-1)^{-1}, M \to +\infty$, the leading order term is $\mathcal{O}\left(\frac{c^{2k^\star+2}}{(c-1)^{k^\star}}M^{k^\star}\right)$. To conclude all boundedness results, we introduce the joint parameter space

$$\theta := (f, \eta, \lambda) \in \Theta_n := \mathcal{F}_n \times \Pi_n \times \Lambda_n,$$

where $\mathcal{F}_n := \{f \in \mathcal{F} : \|f\|_{\mathcal{F}} \leqslant \gamma_n\}$, $\Pi_n := [\underline{\eta}, \bar{\eta}]$ and $\Lambda_n := [\underline{\lambda}, \bar{\lambda}]$. Moreover, we have

$$\left|\ell_c^k(\theta; \widehat{C}_n) - \ell_c^k(\theta; C)\right| \leqslant \underbrace{\frac{2c}{\underline{\lambda}^{k^\star-1}}(M-\underline{\eta})^{k^\star-1}}_{L_C} \left|\widehat{C}_n(\boldsymbol{X}) - C(\boldsymbol{X})\right|,$$

and

$$\left|\ell_{c,\psi}^k(\theta; \widehat{C}_n) - \ell_c^k(\theta; C)\right| \leqslant L_C \left|\widehat{C}_n(\boldsymbol{X}) - C(\boldsymbol{X})\right|.$$

In particular, $L_C = \frac{2^{2k^\star-1/k^\star}\|h\|_\infty^{k^\star-1}}{k^\star}\frac{c^{2k^\star+1}}{(c-1)^{k^\star-1}}M^{k^\star-1}$.

Next, we begin to prove the regret bound. Recall that the empirical minimizer is $\widehat{\theta}_n := (\widehat{f}_n, \widehat{\eta}_n, \widehat{\lambda}_n) \in \underset{(f,\eta,\lambda)\in\Theta_n}{\operatorname{argmin}}\ \mathbb{E}_n \ell_{c,\psi}^k(f, \eta, \lambda; \widehat{C}_n)$ where the distributions of (η, λ) can be constrained to $\Pi_n \times \Lambda_n = [\underline{\eta}, \bar{\eta}] \times [\underline{\lambda}, \bar{\lambda}]$ due to the previous boundedness. We also define the within-Θ_n oracle $\theta_\gamma^* := (f_\gamma^*, \eta_\gamma^*, \lambda_\gamma^*) \in \underset{(f,\eta,\lambda)\in\Theta_n}{\operatorname{argmin}}\ \mathcal{L}_{c,\psi}^k(f, \eta, \lambda)$. Then, by definition, we have $\mathcal{L}_{c,\psi}^k(\theta_\gamma^*) - \mathcal{R}_c^{k,*} = \mathcal{A}_c^k(\gamma_n)$. By Proposition 4, we have

$$\begin{aligned}
&\mathcal{L}_c^k(\widehat{\theta}_n) - \mathcal{R}_c^{k,*} \\
\leqslant& 2[\mathcal{L}_{c,\psi}^k(\widehat{\theta}_n) - \mathcal{R}_c^{k,*}] \\
=& 2\left(\mathcal{L}_{c,\psi}^k(\widehat{\theta}_n) - \mathbb{E}_n \ell_{c,\psi}^k(\widehat{\theta}_n; C)\right) + 2\mathbb{E}_n[\ell_{c,\psi}^k(\widehat{\theta}_n; C) - \ell_{c,\psi}^k(\widehat{\theta}_n; \widehat{C}_n)] + 2\left(\mathbb{E}_n \ell_{c,\psi}^k(\widehat{\theta}_n; \widehat{C}_n) - \mathcal{R}_c^{k,*}\right) \\
\leqslant& 2\sup_{\theta\in\Theta_n}(\mathbb{P} - \mathbb{P}_n)\ell_{c,\psi}^k(\theta; C) + 2L_C\|\widehat{C}_n - C\|_\infty + 2\left(\mathbb{E}_n \ell_{c,\psi}^k(\theta_\gamma^*; \widehat{C}_n) - \mathcal{R}_c^{k,*}\right) \\
\leqslant& 2\sup_{\theta\in\Theta_n}(\mathbb{P} - \mathbb{P}_n)\ell_{c,\psi}^k(\theta; C) + 4L_C\|\widehat{C}_n - C\|_\infty + 2\left(\mathbb{E}_n \ell_{c,\psi}^k(\theta_\gamma^*; C) - \mathcal{L}_{c,\psi}^k(\theta_\gamma^*)\right) + 2\mathcal{A}(\gamma_n) \\
\leqslant& 2\sup_{\theta\in\Theta_n}(\mathbb{P} - \mathbb{P}_n)\ell_{c,\psi}^k(\theta; C) + 4L_C\|\widehat{C}_n - C\|_\infty + 2\sup_{\theta\in\Theta_n}(\mathbb{P}_n - \mathbb{P})\ell_{c,\psi}^k(\theta; C) + 2\mathcal{A}(\gamma_n).
\end{aligned}$$

It follows standard routine to propose a Rademacher complexity bound. Fix $\delta > 0$. First by McDiarmid Inequality (Bartlett and Mendelson, 2002, Theorem 9), with probability $\geqslant 1-\delta$,

$$\sup_{\theta\in\Theta_n}(\mathbb{P}-\mathbb{P}_n)\ell^k_{c.\psi}(\theta;C) \leqslant \mathbb{E}\sup_{\theta\in\Theta_n}(\mathbb{P}-\mathbb{P}_n)\ell^k_{c,\psi}(\theta;C) + (\overline{\ell}^k_c - \underline{\ell}^k_c)\sqrt{\frac{\log(2/\delta)}{2n}},$$
$$\sup_{\theta\in\Theta_n}(\mathbb{P}_n-\mathbb{P})\ell^k_{c.\psi}(\theta;C) \leqslant \mathbb{E}\sup_{\theta\in\Theta_n}(\mathbb{P}_n-\mathbb{P})\ell^k_{c,\psi}(\theta;C) + (\overline{\ell}^k_c - \underline{\ell}^k_c)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Next we define the Rademacher complexity on Θ_n as follows:

$$\mathsf{R}_n(\Theta_n) := \mathbb{E}_{(\boldsymbol{X},\sigma)\sim\mathbb{P}}\sup_{\theta\in\Theta_n}\mathbb{E}_n[\sigma\ell^k_{c,\psi}(\theta;C)],$$

where σ is the Rademacher variable independent of $(\boldsymbol{X},A,Y)$ under $\mathbb{P}$. Then by standard symmetrization arguments, we have

$$\mathbb{E}\sup_{\theta\in\Theta_n}(\mathbb{P}-\mathbb{P}_n)\ell^k_{c,\psi}(\theta;C) \leqslant 2\mathsf{R}_n(\Theta_n), \quad \mathbb{E}\sup_{\theta\in\Theta_n}(\mathbb{P}_n-\mathbb{P})\ell^k_{c,\psi}(\theta;C) \leqslant 2\mathsf{R}_n(\Theta_n).$$

To obtain an error bound on $\mathsf{R}_n(\Theta_n)$, we decouple Θ_n by exploiting the ℓ^1-Lipschitzness of $\ell^k_{c,\psi}$. For ease of notation, we suppress the dependency on C in $\ell^k_{c,\psi}$. Note that for $\theta_i = (f_i,\eta_i,\lambda_i)$ $(i=1,2)$,

$$\begin{aligned}
&|\ell^k_{c,\psi}(\theta_1) - \ell^k_{c,\psi}(\theta_2)| \\
\leqslant&|\ell^k_{c,\psi}(f_1,\eta_1,\lambda_1) - \ell^k_{c,\psi}(f_1,\eta_1,\lambda_2)| + |\ell^k_{c,\psi}(f_1,\eta_1,\lambda_2) - \ell^k_{c,\psi}(f_1,\eta_2,\lambda_2)| + |\ell^k_{c,\psi}(f_1,\eta_2,\lambda_2) - \ell^k_{c,\psi}(f_2,\eta_2,\lambda_2)| \\
\leqslant&\frac{c}{k^\star}\left(\frac{2cM}{c-1}\right)^{k^\star}\left|\frac{1}{\lambda_1^{k^\star-1}} - \frac{1}{\lambda_1^{k^\star-1}}\right| + \frac{c}{k}|\lambda_1-\lambda_2| + \\
&\frac{c}{k^\star\underline{\lambda}^{k^\star-1}}\Bigg[\frac{\psi[+f(\boldsymbol{X})]}{2}\left|\left(+\widehat{C}_n(\boldsymbol{X})-\eta_1\right)_+^{k^\star} - \left(+\widehat{C}_n(\boldsymbol{X})-\eta_2\right)_+^{k^\star}\right| + \\
&\qquad\frac{\psi[-f(\boldsymbol{X})]}{2}\left|\left(-\widehat{C}_n(\boldsymbol{X})-\eta_1\right)_+^{k^\star} - \left(-\widehat{C}_n(\boldsymbol{X})-\eta_2\right)_+^{k^\star}\right|\Bigg] + |\eta_1-\eta_2| + \\
&\frac{c}{k^\star\underline{\lambda}^{k^\star-1}}\left(\frac{2cM}{c-1}\right)^{k^\star}|\psi[f_1(\boldsymbol{X})] - \psi[f_2(\boldsymbol{X})]| \\
\leqslant&L_\lambda|\lambda_1-\lambda_2| + L_\eta|\eta_1-\eta_2| + L_f|f_1(\boldsymbol{X}) - f_2(\boldsymbol{X})|,
\end{aligned}$$

where

$$\begin{cases}
L_\lambda := \frac{c}{k^\star}\left(\frac{2cM}{c-1}\right)^{k^\star} \times \frac{k^\star-1}{\underline{\lambda}^{k^\star}} + \frac{c}{k} & = \frac{2^{2k^\star+1}\|h\|_\infty^{k^\star}}{k}\frac{c^{\frac{(k^\star+1)(2k^\star-1)}{k^\star-1}}}{(c-1)^{k^\star}}M^{k^\star} + \frac{c}{k}; \\
L_\eta := \frac{c}{k^\star\underline{\lambda}^{k^\star-1}} \times k^\star\left(\frac{2cM}{c-1}\right)^{k^\star-1} + 1 & = \frac{2^{2k^\star-1/k^\star-1}\|h\|_\infty^{k^\star-1}}{k^\star}\frac{c^{2k^\star+1}}{(c-1)^{k^\star-1}}M^{k^\star-1} + 1; \\
L_f := \frac{c}{k^\star\underline{\lambda}^{k^\star-1}}\left(\frac{2cM}{c-1}\right)^{k^\star} \times 2 & = \frac{2^{2k^\star-1/k^\star+1}\|h\|_\infty^{k^\star-1}}{k^\star}\frac{c^{2k^\star+2}}{(c-1)^{k^\star}}M^{k^\star}.
\end{cases}$$

We Denote $L_\ell := L_f \vee L_\eta \vee L_\lambda$. Notice that the leading order term as $c, (c-1)^{-1}, M \to +\infty$ is $L_\lambda = \mathcal{O}\left(\frac{c^{\frac{(k^\star+1)(2k^\star-1)}{k^\star-1}}}{(c-1)^{k^\star}} M^{k^\star}\right)$. And we also define the marginal Rademacher complexities

$$\mathsf{R}_n(\mathcal{F}_n) := \mathbb{E}_{(\boldsymbol{X},\sigma)\sim\mathbb{P}} \sup_{f\in\mathcal{F}_n} \mathbb{E}_n[\sigma f(\boldsymbol{X})]; \quad \mathsf{R}_n(\Pi_n) := \mathbb{E}_\sigma \sup_{\eta\in\Pi_n} (\eta\mathbb{E}_n\sigma); \quad \mathsf{R}_n(\Lambda_n) := \mathbb{E}_\sigma \sup_{\lambda\in\Lambda_n} (\lambda\mathbb{E}_n\sigma).$$

Then by the multidimensional version (Qi et al., 2019, Lemma 3.1) of the Rademarcher complexity of the Lipschitz composition (Boucheron et al., 2005, Theorem 3.3), we have

$$\mathsf{R}_n(\Theta_n) \leqslant L_\ell[\mathsf{R}_n(\mathcal{F}_n) + \mathsf{R}_n(\Pi_n) + \mathsf{R}_n(\Lambda_n)],$$

where by Vapnik-Chervonenkis Inequality (Boucheron et al., 2005, Theorem 3.4), there exists a universal constant C_{VC} such that $\mathsf{R}_n(\Pi_n) \leqslant C_{\mathrm{VC}}\sqrt{2(|\overline{\eta}| \vee |\underline{\eta}|)/n}$ and $\mathsf{R}_n(\Lambda_n) \leqslant C_{\mathrm{VC}}\sqrt{2\overline{\lambda}/n}$, and by Bartlett and Mendelson (2002, Lemma 22), $\mathsf{R}_n(\mathcal{F}_n) \leqslant 2\sqrt{\gamma_n/n}$. Combining the above results, our regret bound becomes

$$\begin{aligned}\mathcal{L}_c^k(\widehat{f}_n, \widehat{\eta}_n, \widehat{\lambda}_n) - \mathcal{R}_c^{k,*} \leqslant & 8L_\ell\left(2\sqrt{\gamma_n/n} + C_{\mathrm{VC}}\sqrt{2(|\overline{\eta}| \vee |\underline{\eta}|)/n} + C_{\mathrm{VC}}\sqrt{2\overline{\lambda}/n}\right) \\ & + 4(\overline{\ell}_c^k - \underline{\ell}_c^k)\sqrt{\frac{\log(2/\delta)}{2n}} + 4L_C\|\widehat{C}_n - C\|_\infty + 2\mathcal{A}_c^k(\gamma_n).\end{aligned}$$

Finally by Assumption 5 that $\mathcal{A}_c^k(\gamma) = K_{\mathcal{A}}\gamma^\beta$, we choose $\gamma_n := n^{\frac{1}{2\beta+1}}$ to obtain the desired regret bound of rate $\mathcal{O}(n^{-\frac{\beta}{2\beta+1}})$ as $n \to \infty$, with the universal constant K_0 as

$$\begin{aligned}K_0 &= 8L_\ell\left(2 + C_{\mathrm{VC}}\sqrt{2}(|\overline{\eta}| \vee |\underline{\eta}|)^{1/2} + C_{\mathrm{VC}}\sqrt{2}\overline{\lambda}^{1/2}\right) + 2\sqrt{2}(\overline{\ell}_c^k - \underline{\ell}_c^k) + 2K_{\mathcal{A}} \\ &= \mathcal{O}\left(L_\ell[(|\overline{\eta}| \vee |\underline{\eta}|)^{1/2} + \overline{\lambda}^{1/2}] + \overline{\ell}_c^k - \underline{\ell}_c^k\right) \\ &= \mathcal{O}\left(\frac{c^{\frac{(k^\star+1)(2k^\star-1)}{k^\star-1}+\frac{1}{2}}}{(c-1)^{k^\star+1/2}} M^{k^\star+1/2}\right),\end{aligned}$$

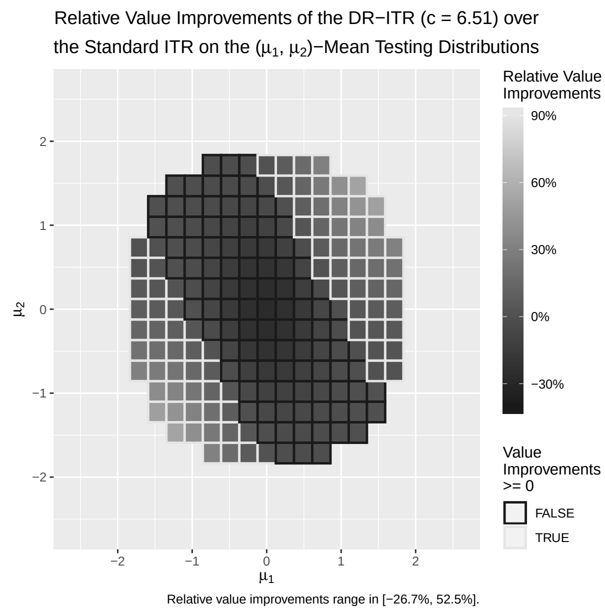
and $K_1 = 4L_C = \mathcal{O}\left(\frac{c^{2k^\star+1}}{(c-1)^{k^\star-1}} M^{k^\star-1}\right)$.

Consider the spectial case $k = +\infty$ and $k^\star = 1$. Consider η_f^* as in (9). Since for any $\eta \leqslant -M$, the objective (9) remains constant. Then we have $-M \leqslant \eta \leqslant M$. The regret bound analysis follows the same as above except that λ is redundant in $\ell_{c,\psi}^1$. For the bounds on $\ell_{c,\psi}^1$, have $\overline{\ell}_c^1 = (2c+1)M$ and $\underline{\ell}_c^1 = -M$. The Lipschitz constants are refined to be $L_C = 2c$, $L_\eta = c+1$, $L_f = 4cM$. And the final universal constants become

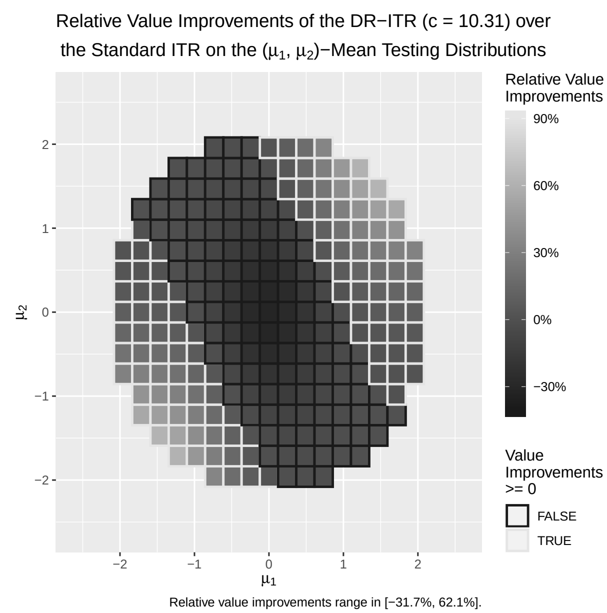
$$K_0 = \mathcal{O}\left(L_\ell(|\overline{\eta}| \vee |\underline{\eta}|)^{1/2} + \overline{\ell}_c^k - \underline{\ell}_c^k\right) = \mathcal{O}(cM^{3/2}); \quad K_1 = 8c = \mathcal{O}(c).$$

S.4 Additional Tables and Figures

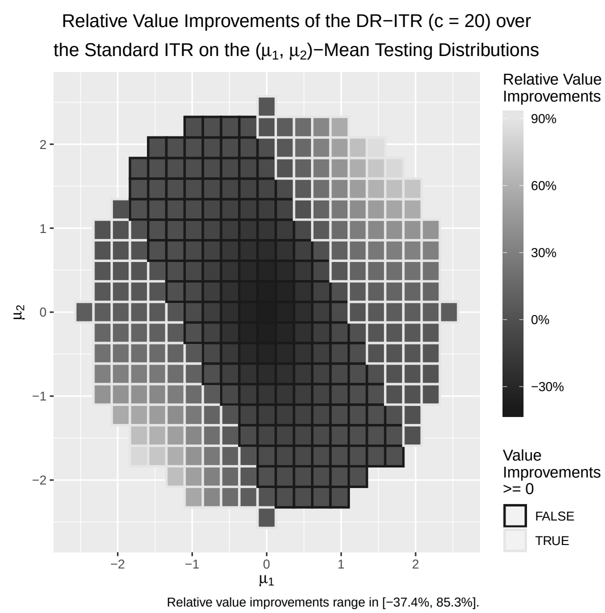
(a) $c = 2.71$



(b) $c = 6.51$



(c) $c = 10.31$



(d) $c = 20$

Figure S.1: Comparing the testing values of the DR-ITR for various c's with the standard ITR on testing distributions $\mathcal{N}_2\big(\boldsymbol{\mu}, \mathbf{I}_2\big)$ of means $\boldsymbol{\mu} \in \big\{(\mu_1, \mu_2)^\intercal \in \mathbb{R}^2 : \mu_1^2 + \mu_2^2 \leqslant 4\log 5\big\}$.

Table S.1: Relative Regrets (%) of Standard ITRs on Mean-Shifted Covariate Domains

μ_2 \ μ_1	−2.45	−1.96	−1.47	−0.979	−0.49	0	0.49	0.979	1.47	1.96	2.45
2.45	0	0	0	0	2	8	27	58	91	107	108
1.96	0	0	0	0	2	10	28	54	75	83	80
1.47	0	0	0	0	2	12	28	46	55	57	52
0.979	1	1	1	0	1	11	25	35	38	35	31
0.49	3	3	3	2	2	2	16	23	22	19	16
0	7	9	11	10	3	5	3	10	11	9	7
−0.49	16	19	22	23	17	3	1	2	3	3	3
−0.979	30	35	38	34	26	10	1	0	1	1	1
−1.47	52	57	55	45	27	12	2	0	0	0	0
−1.96	79	82	75	53	29	11	2	0	0	0	0
−2.45	108	107	91	58	27	9	2	0	0	0	0

[1] $\boldsymbol{\mu} = (\mu_1, \mu_2, 0, \cdots, 0)^\intercal$ with μ_1 in column and μ_2 in row is the testing covariate centroid.

[2] Relative regret(ITR) = [value(LB-ITR) − value(ITR)]/|value(LB-ITR)|

Table S.2: Misclassification Rates (%) of Standard ITRs on Mean-Shifted Covariate Domains

μ_2 \ μ_1	−2.45	−1.96	−1.47	−0.979	−0.49	0	0.49	0.979	1.47	1.96	2.45
2.45	1	1	2	3	4	6	10	18	30	43	53
1.96	2	3	5	7	8	10	13	20	29	38	44
1.47	3	6	10	13	15	16	19	23	28	33	35
0.979	6	10	16	20	23	25	26	27	28	27	26
0.49	9	15	22	27	30	32	32	30	27	23	19
0	13	19	26	31	34	35	34	30	26	19	13
−0.49	18	23	27	30	32	32	30	27	21	15	9
−0.979	26	27	28	27	26	25	23	20	16	11	6
−1.47	34	33	28	23	19	16	15	13	10	6	3
−1.96	44	38	29	20	14	10	8	7	5	3	2
−2.45	53	43	30	18	10	6	4	3	2	1	1

[1] $\boldsymbol{\mu} = (\mu_1, \mu_2, 0, \cdots, 0)^\intercal$ with μ_1 in column and μ_2 in row is the testing covariate centroid.

Table S.3: Relative Regrets (%) of RCT-DR-ITRs on Mean-Shifted Covariate Domains ($n_{\text{calib}} = 50$)

μ_2 \ μ_1	−2.45	−1.96	−1.47	−0.979	−0.49	0	0.49	0.979	1.47	1.96	2.45
2.45	0	0	0	1	3	8	16	19	16	10	6
1.96	0	0	1	1	4	11	19	21	15	10	5
1.47	0	0	1	2	4	14	23	23	15	8	4
0.979	0	0	1	2	6	15	24	22	14	7	3
0.49	1	2	2	3	7	9	18	18	11	5	2
0	1	3	7	9	8	16	9	10	7	3	1
−0.49	2	5	11	17	19	10	7	3	2	1	1
−0.979	3	7	14	21	23	14	5	2	1	0	0
−1.47	3	7	14	22	21	13	4	1	0	0	0
−1.96	5	9	15	21	19	10	3	1	0	0	0
−2.45	6	9	15	18	15	8	2	1	0	0	0

[1] $\boldsymbol{\mu} = (\mu_1, \mu_2, 0, \cdots, 0)^\intercal$ with μ_1 in column and μ_2 in row is the testing covariate centroid.

[2] Relative regret(ITR) = [value(LB-ITR) − value(ITR)]/|value(LB-ITR)|

Table S.4: Relative Regrets (%) of RCT-DR-ITRs on Mean-Shifted Covariate Domains ($n_{\text{calib}} = 100$)

μ_2 \ μ_1	−2.45	−1.96	−1.47	−0.979	−0.49	0	0.49	0.979	1.47	1.96	2.45
2.45	0	0	0	1	3	7	14	16	14	9	6
1.96	0	0	0	1	3	10	18	19	13	8	4
1.47	0	0	0	1	3	12	21	20	14	7	3
0.979	0	0	1	2	4	13	22	20	13	6	2
0.49	1	1	2	2	4	7	17	17	10	4	2
0	1	3	6	8	5	11	6	8	6	3	1
−0.49	2	4	10	16	17	7	4	2	2	1	1
−0.979	2	6	13	20	22	12	3	1	1	0	0
−1.47	3	7	13	20	20	12	3	1	0	0	0
−1.96	4	8	13	18	17	10	3	1	0	0	0
−2.45	5	8	14	16	13	7	2	0	0	0	0

[1] $\boldsymbol{\mu} = (\mu_1, \mu_2, 0, \cdots, 0)^\intercal$ with μ_1 in column and μ_2 in row is the testing covariate centroid.

[2] Relative regret(ITR) = [value(LB-ITR) − value(ITR)]/|value(LB-ITR)|

Table S.5: Relative Value Improvements (%) of RCT-DR-ITRs over Standard ITRs on Mean-Shifted Covariate Domains ($n_{\text{calib}} = 50$)

μ_2 \ μ_1	-2.45	-1.96	-1.47	-0.979	-0.49	0	0.49	0.979	1.47	1.96	2.45
2.45	0	0	0	-1	-1	1	11	40	75	98	102
1.96	0	0	-1	-1	-2	0	9	32	60	73	75
1.47	0	0	0	-2	-3	-3	6	23	40	49	48
0.979	0	0	0	-2	-4	-5	2	13	24	28	28
0.49	2	2	1	-2	-6	-7	-2	5	11	14	14
0	6	6	4	1	-5	-11	-6	0	4	6	5
-0.49	13	14	11	6	-2	-6	-5	-2	1	2	2
-0.979	27	29	24	13	2	-4	-4	-2	0	0	0
-1.47	48	49	41	23	6	-1	-3	-1	0	0	0
-1.96	74	73	60	33	10	0	-1	-1	0	0	0
-2.45	102	98	76	40	12	1	-1	-1	0	0	0

[1] $\boldsymbol{\mu} = (\mu_1, \mu_2, 0, \cdots, 0)^\intercal$ with μ_1 in column and μ_2 in row is the testing covariate centroid.

[2] Relative value improvement = difference of relative regrets.

Table S.6: Misclassification Rates (%) of RCT-DR-ITRs on Mean-Shifted Covariate Domains ($n_{\text{calib}} = 50$)

μ_2 \ μ_1	-2.45	-1.96	-1.47	-0.979	-0.49	0	0.49	0.979	1.47	1.96	2.45
2.45	1	2	3	4	5	7	12	16	20	19	15
1.96	2	3	6	7	10	12	15	20	21	20	15
1.47	3	7	11	14	17	19	22	24	24	21	15
0.979	6	11	17	22	26	28	29	30	27	21	14
0.49	9	15	23	30	34	35	35	33	28	21	13
0	11	19	27	34	37	39	37	33	27	19	11
-0.49	13	21	28	33	35	35	34	30	23	15	9
-0.979	14	21	27	29	29	28	25	22	17	11	6
-1.47	14	20	24	24	21	19	17	14	11	7	3
-1.96	15	19	21	20	15	12	9	8	6	3	2
-2.45	15	18	19	16	11	7	5	4	2	1	1

[1] $\boldsymbol{\mu} = (\mu_1, \mu_2, 0, \cdots, 0)^\intercal$ with μ_1 in column and μ_2 in row is the testing covariate centroid.

Table S.7: Misclassification Rates (%) of RCT-DR-ITRs on Mean-Shifted Covariate Domains ($n_{\text{calib}} = 100$)

μ_2 \ μ_1	−2.45	−1.96	−1.47	−0.979	−0.49	0	0.49	0.979	1.47	1.96	2.45
2.45	1	2	3	4	5	7	11	16	19	19	15
1.96	2	3	6	7	9	12	15	20	21	20	14
1.47	3	7	10	14	17	18	21	24	24	21	14
0.979	6	11	17	22	25	27	28	29	27	21	14
0.49	9	15	23	29	32	34	34	33	28	21	13
0	11	19	27	33	36	37	36	33	27	19	11
−0.49	12	21	28	32	34	34	33	29	23	15	9
−0.979	13	21	27	29	28	27	25	22	17	11	6
−1.47	14	20	24	24	21	18	16	14	11	7	3
−1.96	14	19	21	19	15	11	9	7	6	3	2
−2.45	15	18	19	16	11	7	5	3	2	1	1

[1] $\boldsymbol{\mu} = (\mu_1, \mu_2, 0, \cdots, 0)^\intercal$ with μ_1 in column and μ_2 in row is the testing covariate centroid.

Table S.8: Misclassification Improvements (%) of RCT-DR-ITRs over Standard ITRs on Mean-Shifted Covariate Domains ($n_{\text{calib}} = 50$)

μ_2 \ μ_1	−2.45	−1.96	−1.47	−0.979	−0.49	0	0.49	0.979	1.47	1.96	2.45
2.45	0	0	−1	−1	−1	−1	−1	2	10	24	38
1.96	0	0	−1	−1	−1	−2	−2	0	8	18	29
1.47	0	0	−1	−1	−2	−3	−3	−1	3	12	20
0.979	0	0	−1	−2	−3	−3	−3	−3	1	6	12
0.49	1	0	−1	−3	−3	−3	−3	−3	−1	2	6
0	2	0	−2	−3	−3	−4	−4	−3	−2	1	2
−0.49	6	3	−1	−3	−3	−3	−3	−3	−1	0	1
−0.979	12	7	1	−2	−3	−3	−2	−2	−1	0	0
−1.47	20	12	4	−1	−2	−2	−2	−1	−1	0	0
−1.96	29	18	8	0	−2	−2	−1	−1	0	0	0
−2.45	38	24	11	3	−1	−1	−1	−1	0	0	0

[1] $\boldsymbol{\mu} = (\mu_1, \mu_2, 0, \cdots, 0)^\intercal$ with μ_1 in column and μ_2 in row is the testing covariate centroid.

Table S.9: Testing Values (Standard Errors) on Mean-Shifted Covariate Domains ($n_{\text{calib}} = 50$)

μ_2 \ μ_1	type	0	0.734	1.469	1.958
1.958	LB-ITR	*2.333 (0.00244)*	*2.907 (0.011)*	*5.334 (0.0362)*	*9.27 (0.0154)*
	ℓ^1-PLS	**2.124** (0.0022)	2.235 (0.011)	3.613 (0.0505)	6.32 (0.103)
	RWL	2.067 (0.00125)	1.59 (0.0104)	0.7237 (0.0488)	0.2045 (0.108)
	Standard ITR	2.089 (0.00158)	1.735 (0.013)	1.348 (0.0595)	1.567 (0.13)
	RCT-ITR	1.913 (0.0082)	1.969 (0.026)	4.168 (0.034)	7.838 (0.0388)
	RCT-DR-ITR	2.085 (0.00444)	2.286 (0.0114)	4.545 (0.0255)	8.371 (0.0451)
	CTE-DR-ITR	2.098 (0.00348)	**2.304** (0.0106)	**4.551** (0.0238)	**8.459** (0.0424)
1.469	LB-ITR	*1.893 (0.00712)*	*2.627 (0.00656)*	*5.28 (0.0213)*	*9.379 (0.0128)*
	ℓ^1-PLS	1.667 (0.00307)	**2.021** (0.0076)	4.095 (0.0342)	7.573 (0.0706)
	RWL	1.655 (0.00131)	1.501 (0.0106)	1.798 (0.0472)	2.791 (0.102)
	Standard ITR	**1.674** (0.00152)	1.645 (0.0127)	2.377 (0.0553)	4.011 (0.119)
	RCT-ITR	1.414 (0.0094)	1.597 (0.025)	4.075 (0.0299)	8.022 (0.0334)
	RCT-DR-ITR	1.627 (0.00688)	1.987 (0.00997)	4.484 (0.0192)	8.611 (0.0285)
	CTE-DR-ITR	1.663 (0.00326)	1.997 (0.00992)	**4.55** (0.0163)	**8.686** (0.0269)
0.734	LB-ITR	*1.227 (0.00244)*	*2.144 (0.00609)*	*5.269 (0.00931)*	*9.608 (0.00898)*
	ℓ^1-PLS	1.094 (0.00418)	**1.676** (0.00442)	4.587 (0.0151)	8.8 (0.0314)
	RWL	1.168 (0.00134)	1.462 (0.00729)	3.357 (0.0344)	6.323 (0.0696)
	Standard ITR	**1.174** (0.00149)	1.553 (0.00806)	3.739 (0.0379)	7.06 (0.0763)
	RCT-ITR	0.7323 (0.011)	1.152 (0.021)	4.157 (0.0238)	8.534 (0.0299)
	RCT-DR-ITR	1.094 (0.00753)	1.651 (0.00675)	4.622 (0.0109)	9.036 (0.015)
	CTE-DR-ITR	1.152 (0.00292)	1.667 (0.00588)	**4.648** (0.0113)	**9.06** (0.0161)
0.000	LB-ITR	*0.9942 (0.00202)*	*1.774 (0.0034)*	*5.232 (0.00559)*	*9.767 (0.0068)*
	ℓ^1-PLS	0.8296 (0.00454)	1.648 (0.0036)	**4.914** (0.00501)	**9.476** (0.0103)
	RWL	**0.9457** (0.00126)	1.645 (0.00339)	4.494 (0.0165)	8.589 (0.0329)
	Standard ITR	0.9437 (0.00153)	1.679 (0.00336)	4.654 (0.017)	8.895 (0.0342)
	RCT-ITR	0.4303 (0.0109)	1.161 (0.0145)	4.518 (0.0172)	8.983 (0.034)
	RCT-DR-ITR	0.8374 (0.00821)	1.647 (0.00574)	4.868 (0.00797)	9.444 (0.00841)
	CTE-DR-ITR	0.9206 (0.00272)	**1.688** (0.00289)	4.888 (0.00698)	9.442 (0.00999)

[1] $\boldsymbol{\mu} = (\mu_1, \mu_2, 0, \cdots, 0)^\intercal$ with μ_1 in column and μ_2 in row is the testing covariate centroid.

[2] Values (larger the better) can be comparable for the same (μ_1, μ_2) but incomparable across different (μ_1, μ_2).

[3] LB-ITR maximizes the testing value function at (μ_1, μ_2) over the linear ITR class. The corresponding testing value is the best achievable among the linear ITR class. [4] RWL (Zhou et al., 2017) implements the same routine as Standard ITR except that $\widehat{C}_n(\boldsymbol{X}) = \widehat{Q}_n(\boldsymbol{X}, 1) - \widehat{Q}_n(\boldsymbol{X}, -1) + 2A[Y - \widehat{Q}_n(\boldsymbol{X}, A)]$. [5] RCT-ITR fits RWL on the calibrating RCT dataset directly.

Table S.10: Testing Values (Standard Errors) on Mean-Shifted Covariate Domains ($n_{\text{calib}} = 100$)

μ_2 \ μ_1	type	0	0.734	1.469	1.958
1.958	LB-ITR	*2.333 (0.00244)*	*2.907 (0.011)*	*5.334 (0.0362)*	*9.27 (0.0154)*
	ℓ^1-PLS	**2.124** (0.0022)	2.235 (0.011)	3.613 (0.0505)	6.32 (0.103)
	RWL	2.067 (0.00125)	1.59 (0.0104)	0.7237 (0.0488)	0.2045 (0.108)
	Standard ITR	2.089 (0.00158)	1.735 (0.013)	1.348 (0.0595)	1.567 (0.13)
	RCT-ITR	2.015 (0.00565)	**2.593** (0.0132)	**4.996** (0.0158)	**8.588** (0.0208)
	RCT-DR-ITR	2.109 (0.00342)	2.349 (0.00905)	4.62 (0.0219)	8.5 (0.0394)
	CTE-DR-ITR	2.099 (0.00392)	2.34 (0.00954)	4.602 (0.0215)	8.488 (0.0393)
1.469	LB-ITR	*1.893 (0.00712)*	*2.627 (0.00656)*	*5.28 (0.0213)*	*9.379 (0.0128)*
	ℓ^1-PLS	1.667 (0.00307)	2.021 (0.0076)	4.095 (0.0342)	7.573 (0.0706)
	RWL	1.655 (0.00131)	1.501 (0.0106)	1.798 (0.0472)	2.791 (0.102)
	Standard ITR	**1.674** (0.00152)	1.645 (0.0127)	2.377 (0.0553)	4.011 (0.119)
	RCT-ITR	1.54 (0.00529)	**2.286** (0.0129)	**4.846** (0.017)	8.713 (0.0183)
	RCT-DR-ITR	1.662 (0.00367)	2.044 (0.00721)	4.566 (0.0153)	8.711 (0.0254)
	CTE-DR-ITR	1.67 (0.00286)	2.044 (0.00818)	4.577 (0.0144)	**8.734** (0.0251)
0.734	LB-ITR	*1.227 (0.00244)*	*2.144 (0.00609)*	*5.269 (0.00931)*	*9.608 (0.00898)*
	ℓ^1-PLS	1.094 (0.00418)	1.676 (0.00442)	4.587 (0.0151)	8.8 (0.0314)
	RWL	1.168 (0.00134)	1.462 (0.00729)	3.357 (0.0344)	6.323 (0.0696)
	Standard ITR	**1.174** (0.00149)	1.553 (0.00806)	3.739 (0.0379)	7.06 (0.0763)
	RCT-ITR	0.8905 (0.00647)	1.651 (0.0138)	**4.701** (0.0168)	9.011 (0.013)
	RCT-DR-ITR	1.134 (0.00408)	1.662 (0.0065)	4.671 (0.00885)	9.094 (0.0122)
	CTE-DR-ITR	1.156 (0.00251)	**1.68** (0.00573)	4.699 (0.00824)	**9.132** (0.0112)
0.000	LB-ITR	*0.9942 (0.00202)*	*1.774 (0.0034)*	*5.232 (0.00559)*	*9.767 (0.0068)*
	ℓ^1-PLS	0.8296 (0.00454)	1.648 (0.0036)	4.914 (0.00501)	9.476 (0.0103)
	RWL	**0.9457** (0.00126)	1.645 (0.00339)	4.494 (0.0165)	8.589 (0.0329)
	Standard ITR	0.9437 (0.00153)	1.679 (0.00336)	4.654 (0.017)	8.895 (0.0342)
	RCT-ITR	0.6198 (0.00875)	1.388 (0.00857)	4.745 (0.00861)	9.376 (0.00737)
	RCT-DR-ITR	0.8879 (0.00506)	1.671 (0.00389)	4.901 (0.00451)	9.489 (0.0068)
	CTE-DR-ITR	0.925 (0.00233)	**1.689** (0.00262)	**4.916** (0.00496)	**9.508** (0.00626)

[1] $\boldsymbol{\mu} = (\mu_1, \mu_2, 0, \cdots, 0)^\intercal$ with μ_1 in column and μ_2 in row is the testing covariate centroid.

[2] Values (larger the better) can be comparable for the same (μ_1, μ_2) but incomparable across different (μ_1, μ_2).

[3] LB-ITR maximizes the testing value function at (μ_1, μ_2) over the linear ITR class. The corresponding testing value is the best achievable among the linear ITR class. [4] RWL (Zhou et al., 2017) implements the same routine as Standard ITR except that $\widehat{C}_n(\boldsymbol{X}) = \widehat{Q}_n(\boldsymbol{X}, 1) - \widehat{Q}_n(\boldsymbol{X}, -1) + 2A[Y - \widehat{Q}_n(\boldsymbol{X}, A)]$. [5] RCT-ITR fits RWL on the calibrating RCT dataset directly.

Table S.11: Testing Misclassification Rates (Standard Errors) on Mean-Shifted Covariate Domains ($n_{\text{calib}} = 50$)

μ_2 \ μ_1	type	0	0.734	1.469	1.958
1.958	LB-ITR	*0.05348 (0.000259)*	*0.0301 (0.000804)*	*0.02702 (0.0038)*	*0.02554 (0.00337)*
	ℓ^1-PLS	0.113 (0.000781)	**0.1625** (0.000913)	0.2239 (0.0015)	0.247 (0.00305)
	RWL	**0.09857** (0.000358)	0.1675 (0.000346)	0.3093 (0.00126)	0.4145 (0.00255)
	Standard ITR	0.0988 (0.000392)	0.1628 (0.000402)	0.29 (0.00163)	0.3802 (0.00322)
	RCT-ITR	0.1148 (0.00191)	0.1783 (0.00334)	0.2567 (0.00477)	0.2687 (0.00374)
	RCT-DR-ITR	0.118 (0.00135)	0.1785 (0.00196)	**0.2148** (0.00178)	**0.1997** (0.00192)
	CTE-DR-ITR	0.1142 (0.00114)	0.1879 (0.0021)	0.236 (0.00237)	0.209 (0.00201)
1.469	LB-ITR	*0.11 (0.00149)*	*0.05955 (0.000487)*	*0.0374 (0.00328)*	*0.03026 (0.00324)*
	ℓ^1-PLS	0.1904 (0.00113)	0.2229 (0.00132)	**0.2353** (0.0011)	0.2251 (0.00203)
	RWL	**0.1616** (0.000581)	0.2099 (0.000599)	0.2972 (0.00124)	0.3601 (0.00255)
	Standard ITR	0.1637 (0.00067)	**0.2066** (0.000681)	0.2781 (0.00153)	0.326 (0.00307)
	RCT-ITR	0.1875 (0.00248)	0.2381 (0.00365)	0.2895 (0.00471)	0.2744 (0.00324)
	RCT-DR-ITR	0.1927 (0.00205)	0.2306 (0.00196)	0.2437 (0.00199)	**0.2109** (0.00173)
	CTE-DR-ITR	0.181 (0.00132)	0.2373 (0.00221)	0.2514 (0.00208)	0.2155 (0.00168)
0.734	LB-ITR	*0.2575 (0.000703)*	*0.144 (0.00177)*	*0.07107 (0.00288)*	*0.04661 (0.00282)*
	ℓ^1-PLS	0.3275 (0.00147)	0.3291 (0.00165)	**0.273** (0.00104)	0.2085 (0.00091)
	RWL	**0.2764** (0.000746)	**0.2877** (0.000915)	0.2858 (0.000886)	0.2747 (0.00184)
	Standard ITR	0.283 (0.000914)	0.2898 (0.00109)	0.2747 (0.00101)	0.2519 (0.00205)
	RCT-ITR	0.333 (0.00275)	0.3537 (0.0036)	0.3333 (0.00393)	0.2615 (0.00234)
	RCT-DR-ITR	0.3178 (0.00237)	0.3203 (0.00214)	0.2778 (0.00192)	0.2102 (0.00128)
	CTE-DR-ITR	0.2974 (0.00129)	0.3147 (0.00189)	0.2771 (0.00173)	**0.2076** (0.00118)
0.000	LB-ITR	*0.3246 (0.000396)*	*0.2802 (0.0015)*	*0.1293 (0.00214)*	*0.08388 (0.00267)*
	ℓ^1-PLS	0.3988 (0.0016)	0.3649 (0.00139)	0.2742 (0.000873)	0.1875 (0.000467)
	RWL	**0.3358** (0.000755)	**0.3147** (0.000808)	0.2582 (0.000556)	0.2033 (0.000881)
	Standard ITR	0.3452 (0.000963)	0.3211 (0.001)	**0.2564** (0.000666)	0.1942 (0.000918)
	RCT-ITR	0.4085 (0.0025)	0.4158 (0.00234)	0.3261 (0.00214)	0.2349 (0.00169)
	RCT-DR-ITR	0.3864 (0.00274)	0.3529 (0.0021)	0.2726 (0.0015)	0.1889 (0.000857)
	CTE-DR-ITR	0.3575 (0.00126)	0.3345 (0.00123)	0.264 (0.00106)	**0.1848** (0.000668)

[1] $\boldsymbol{\mu} = (\mu_1, \mu_2, 0, \cdots, 0)^\intercal$ with μ_1 in column and μ_2 in row is the testing covariate centroid.

[2] LB-ITR maximizes the testing value function at (μ_1, μ_2) over the linear ITR class. The corresponding testing value is the best achievable among the linear ITR class. [3] RWL (Zhou et al., 2017) implements the same routine as Standard ITR except that $\widehat{C}_n(\boldsymbol{X}) = \widehat{Q}_n(\boldsymbol{X}, 1) - \widehat{Q}_n(\boldsymbol{X}, -1) + 2A[Y - \widehat{Q}_n(\boldsymbol{X}, A)]$. [4] RCT-ITR fits RWL on the calibrating RCT dataset directly.

Table S.12: Testing Values of RCT-DR-ITRs of Various k's on Mean-Shifted Covariate Domains ($n_{\text{calib}} = 50$)

μ_2 \ μ_1	k	0	0.734	1.47	1.96
1.96	1.25	2.08(0.004443)	2.25(0.01238)	4.4(0.03824)	8.17(0.07266)
	1.5	2.09(0.004052)	2.28(0.01154)	4.47(0.0317)	8.27(0.05863)
	2	2.09(0.004445)	2.29(0.01139)	4.54(0.02549)	8.37(0.04507)
	3	2.08(0.005431)	2.25(0.01187)	4.52(0.02422)	8.37(0.0428)
	∞	2.1(0.004169)	2.27(0.01313)	4.54(0.02419)	8.43(0.03522)
1.47	1.25	1.64(0.005444)	1.99(0.009954)	4.42(0.02606)	8.45(0.04875)
	1.5	1.64(0.005729)	2(0.009707)	4.42(0.02437)	8.52(0.04136)
	2	1.63(0.006885)	1.99(0.009965)	4.48(0.01924)	8.61(0.02852)
	3	1.64(0.006302)	1.98(0.01028)	4.47(0.01846)	8.63(0.02501)
	∞	1.64(0.006803)	1.98(0.01093)	4.51(0.01848)	8.63(0.02595)
0.734	1.25	1.11(0.006071)	1.64(0.006628)	4.58(0.01659)	8.95(0.02455)
	1.5	1.12(0.005547)	1.64(0.007019)	4.58(0.01508)	8.97(0.02298)
	2	1.09(0.007527)	1.65(0.006753)	4.62(0.01089)	9.04(0.01496)
	3	1.1(0.007473)	1.62(0.008308)	4.59(0.01228)	9.02(0.01563)
	∞	1.12(0.00672)	1.62(0.008311)	4.61(0.01417)	9.04(0.01468)
0	1.25	0.859(0.007158)	1.65(0.005616)	4.87(0.007131)	9.43(0.01052)
	1.5	0.859(0.007117)	1.64(0.006172)	4.88(0.006802)	9.43(0.0116)
	2	0.837(0.008205)	1.65(0.005744)	4.87(0.007969)	9.44(0.008415)
	3	0.854(0.007488)	1.64(0.006564)	4.86(0.006542)	9.46(0.007206)
	∞	0.888(0.005782)	1.64(0.005722)	4.85(0.008767)	9.45(0.008676)

[1] $\boldsymbol{\mu} = (\mu_1, \mu_2, 0, \cdots, 0)^\intercal$ with μ_1 in column and μ_2 in row is the testing covariate centroid. [2] Values (larger the better) can be comparable for the same (μ_1, μ_2) but incomparable across different (μ_1, μ_2).

Table S.13: Testing Values (Standard Errors) on Mixture of Subgroups ($n_{\text{calib}} = 50$)

	Testing Subgroup 1 Probability				
type	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
LB-ITR	*1.665 (0.0067)*	*1.537 (0.00618)*	*1.444 (0.00412)*	*1.545 (0.00537)*	*1.679 (0.00585)*
ℓ^1-PLS	1.182 (0.00191)	1.264 (0.0014)	**1.399** (0.000591)	**1.537** (0.000333)	1.624 (0.000781)
RWL	1.092 (0.00349)	1.194 (0.00265)	1.363 (0.00123)	1.535 (0.00046)	**1.64** (0.00114)
Standard ITR	1.143 (0.00434)	1.232 (0.00329)	1.383 (0.0015)	1.535 (0.000543)	1.632 (0.00142)
RCT-ITR	1.251 (0.0108)	1.116 (0.011)	1.046 (0.0108)	1.144 (0.0101)	1.275 (0.0102)
RCT-DR-ITR	**1.267** (0.0066)	**1.305** (0.00423)	1.395 (0.00256)	1.52 (0.00212)	1.614 (0.00234)
CTE-DR-ITR	1.16 (0.00409)	1.247 (0.00323)	1.388 (0.00137)	1.534 (0.00055)	1.628 (0.00149)

[1] Testing subgroup 1 probability = 0.75 is the same as the training one.

[2] Values (larger the better) can be comparable for the same subgroup 1 probability but incomparable across different subgroup 1 probabilities [3] LB-ITR maximizes the testing value function over the linear ITR class. The corresponding testing value is the best achievable among the linear ITR class. [4] RWL (Zhou et al., 2017) implements the same routine as Standard ITR except that $\widehat{C}_n(\boldsymbol{X}) = \widehat{Q}_n(\boldsymbol{X}, 1) - \widehat{Q}_n(\boldsymbol{X}, -1) + 2A[Y - \widehat{Q}_n(\boldsymbol{X}, A)]$. [5] RCT-ITR fits RWL on the calibrating RCT dataset directly.

Table S.14: Testing Values (Standard Errors) on Mixture of Subgroups ($n_{\text{calib}} = 100$)

	Testing Subgroup 1 Probability				
type	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
LB-ITR	*1.665 (0.0067)*	*1.537 (0.00618)*	*1.444 (0.00412)*	*1.545 (0.00537)*	*1.679 (0.00585)*
ℓ^1-PLS	1.182 (0.00191)	1.264 (0.0014)	1.399 (0.000591)	**1.537** (0.000333)	1.624 (0.000781)
RWL	1.092 (0.00349)	1.194 (0.00265)	1.363 (0.00123)	1.535 (0.00046)	**1.64** (0.00114)
Standard ITR	1.143 (0.00434)	1.232 (0.00329)	1.383 (0.0015)	1.535 (0.000543)	1.632 (0.00142)
RCT-ITR	**1.493** (0.00431)	**1.354** (0.00499)	1.25 (0.00489)	1.359 (0.0049)	1.5 (0.0046)
RCT-DR-ITR	1.284 (0.00654)	1.324 (0.00421)	**1.402** (0.00195)	1.524 (0.00191)	1.613 (0.00233)
CTE-DR-ITR	1.165 (0.00403)	1.247 (0.00305)	1.389 (0.00134)	1.535 (0.000584)	1.628 (0.00147)

[1] Testing subgroup 1 probability = 0.75 is the same as the training one.

[2] Values (larger the better) can be comparable for the same subgroup 1 probability but incomparable across different subgroup 1 probabilities [3] LB-ITR maximizes the testing value function over the linear ITR class. The corresponding testing value is the best achievable among the linear ITR class. [4] RWL (Zhou et al., 2017) implements the same routine as Standard ITR except that $\widehat{C}_n(\boldsymbol{X}) = \widehat{Q}_n(\boldsymbol{X}, 1) - \widehat{Q}_n(\boldsymbol{X}, -1) + 2A[Y - \widehat{Q}_n(\boldsymbol{X}, A)]$. [5] RCT-ITR fits RWL on the calibrating RCT dataset directly.

Table S.15: Testing Misclassification Rates (Standard Errors) on Mixture of Subgroups ($n_{\text{calib}} = 50$)

	Testing Subgroup 1 Probability				
type	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
LB-ITR	*0.06691 (0.0017)*	*0.1556 (0.0014)*	*0.2296 (0.00078)*	*0.153 (0.0012)*	*0.06668 (0.0015)*
ℓ^1-PLS	0.3059 (0.00044)	0.2775 (0.00027)	0.2291 (0.00016)	0.1789 (0.00041)	0.149 (0.00058)
RWL	0.3242 (0.00071)	0.2885 (4e-04)	0.2283 (0.00021)	**0.1664** (0.00069)	**0.1302** (0.00099)
Standard ITR	0.3103 (0.00097)	0.2785 (0.00058)	**0.2238** (0.00017)	0.1676 (0.00074)	0.1342 (0.0011)
RCT-ITR	**0.2472** (0.0027)	0.2822 (0.0025)	0.3001 (0.0022)	0.2763 (0.0023)	0.2436 (0.0026)
RCT-DR-ITR	0.2751 (0.0023)	**0.2614** (0.0013)	0.2266 (0.00052)	0.1809 (0.0012)	0.147 (0.0014)
CTE-DR-ITR	0.3068 (0.00093)	0.2759 (0.00059)	0.2242 (0.00019)	0.1701 (0.00074)	0.1379 (0.0011)

[1] Testing subgroup 1 probability = 0.75 is the same as the training one. [2] LB-ITR maximizes the testing value function over the linear ITR class. The corresponding testing value is the best achievable among the linear ITR class. [3] RWL (Zhou et al., 2017) implements the same routine as Standard ITR except that $\widehat{C}_n(\boldsymbol{X}) = \widehat{Q}_n(\boldsymbol{X}, 1) - \widehat{Q}_n(\boldsymbol{X}, -1) + 2A[Y - \widehat{Q}_n(\boldsymbol{X}, A)]$.
[4] RCT-ITR fits RWL on the calibrating RCT dataset directly.

References

Bartlett, P. L. and Mendelson, S. (2002), "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, 3, 463–482.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005), "Theory of classification: A survey of some recent advances," *ESAIM: Probability and Statistics*, 9, 323–375.

Krokhmal, P. A. (2007), "Higher moment coherent risk measures," *Quantitative Finance*, 7, 373–387.

Pang, J.-S., Razaviyayn, M., and Alvarado, A. (2016), "Computing B-stationary points of nonsmooth DC programs," *Mathematics of Operations Research*, 42, 95–118.

Qi, Z., Pang, J.-S., and Liu, Y. (2019), "Estimating Individualized Decision Rules with Tail Controls," *arXiv preprint arXiv:1903.04367*.

Razaviyayn, M., Hong, M., and Luo, Z.-Q. (2013), "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, 23, 1126–1153.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2017), "Residual weighted learning for estimating individualized treatment rules," *Journal of the American Statistical Association*, 112, 169–187.