

Learning bounds for doubly-robust covariate shift adaptation

author names withheld

Editor: Under Review for ALT 2026

Abstract

Distribution shift between the training and test domains poses a key challenge for modern machine learning. An extensively studied instance is the *covariate shift*, where the marginal distribution of covariates differs across domains but the conditional distribution of outcome remains the same. The doubly-robust (DR) estimator, recently introduced by Kato et al. (2023), combines the density ratio estimation with a pilot regression model and manifests asymptotic normality and \sqrt{n} -consistency, even when the pilot estimates converge slowly. However, the prior arts has focused exclusively on establishing asymptotic results and has left open the question of non-asymptotic guarantees for the DR estimator.

This paper provides the first non-asymptotic learning bounds for the DR covariate shift adaptation. Our contributions are two-fold: (i) We establish *structure-agnostic* high-probability upper bounds on the excess target risk of the DR estimator that depend only on the L^2 -errors of the pilot estimates and the Rademacher complexity of the model class, without assuming specific procedures to obtain the pilot estimate, and (ii) under *well-specified parametric models*, we analyze the DR covariate shift adaptation based on modern techniques for non-asymptotic analysis of MLE, whose key terms governed by the Fisher information mismatch between the source and target distributions. Together, these findings bridge asymptotic efficiency results and a finite-sample out-of-distribution generalization bounds, providing a comprehensive theoretical underpinnings for the DR covariate shift adaptation.

Keywords: covariate shift, empirical risk minimization, importance-weighting, doubly-robust estimator

1. Introduction

Classical supervised learning assumes that the training and test data are drawn from the same distribution (Vapnik, 2013; Györfi et al., 2002). In practice, this assumption is rarely met. For instance, credit models are trained on approved customers but deployed on rejected applicants; medical imaging data vary across hospitals due to differences in equipment and protocols (Koh et al., 2021; Guan and Liu, 2021); and in natural language processing, labeled corpora such as the *Wall Street Journal*, differ sharply from domains like *arXiv* (Jiang and Zhai, 2007). For all these cases, distribution shift between training and test domains undermines predictive performance.

An important particular case of such a distribution shift is the *covariate shift* (Shimodaira, 2000; Quiñero-Candela et al., 2008; Pan and Yang, 2009), where the marginal distribution of covariates X varies across domains while the conditional distribution of $Y|X$ remains the same. Covariate shift is well-documented in healthcare (Wei et al., 2015; Hajiramezanali et al., 2018), image classification (Saenko et al., 2010), remote sensing Tuia et al. (2011), sentiment analysis Blitzer et al. (2007), and speech and language processing (Yamada et al., 2009; Hassan et al., 2013; Fei and Liu, 2015).

The covariate shift adaptation problem assumes access to labeled samples from a source domain and unlabeled covariates from a target domain, with the goal of learning a predictor with a desirable performance under the target distribution. This problem has been central to the literature of transfer

learning and domain adaptation (Sugiyama et al., 2007a,b; Sugiyama and Kawanabe, 2012; Pan and Yang, 2009; Kato et al., 2023), especially when the target labels are scarce or costly to obtain.

A core difficulty lies in estimating the covariate density ratio between the source and target domains. The standard approach – plugging-in an estimated covariate density ratio into an importance-weighted empirical risk minimization (Sugiyama et al., 2007a,b, 2008; Reddi et al., 2015) – turns out to be highly sensitive to the estimation errors of the density ratio and performs poorly unless the estimator converges at nearly parametric rates. To address this, Kato et al. (2023) proposes a *doubly-robust (DR) estimator*, which augments the importance-weighting with a pilot regression model and leverages double machine learning techniques (Chernozhukov et al., 2017, 2018, 2022, 2023; Foster and Syrgkanis, 2023). Their results establish the asymptotic normality and \sqrt{n} -consistency of their DR estimator under parametric models, even when the pilot estimates converge slowly.

Yet, the literature of covariate shift adaptation has centered exclusively on achieving asymptotic results. It remains unclear how the DR covariate shift adaptation performs in finite-sample regimes. This paper aims to close this gap. Our contributions can be summarized as follows:

- (i) **Structure-agnostic guarantees:** We first derive the first non-asymptotic upper bounds on the excess target risk for the DR estimator, depending only on a product of the statistical rates of convergence of the pilot estimates, without assumptions on how they are obtained.
- (ii) **Fast rates for parametric models:** By studying the DR estimator through the lens of modern non-asymptotic theory of maximum likelihood estimation (MLE), we prove that the estimator achieves a rate of convergence of the order $\mathcal{O}(1/n)$ under covariate shift.

Together, these results bridge asymptotic efficiency results and a finite-sample out-of-distribution (OOD) generalization bound, providing a comprehensive theoretical underpinning of the DR covariate shift adaptation.

1.1. Related works

Let us take a moment to discuss subsets of related works in covariate shift, doubly-robust estimation, and structure-agnostic estimation framework.

Covariate shift The study of covariate shift can be dated back to the seminal paper by Shimodaira (2000). This work investigates the impact of covariate shift under parametric models with the vanilla MLE and proposes an importance-weighting (IW) method, which has a prominent improvement if the underlying regression model is mis-specified. It also establishes the asymptotic normality for a weighted version of MLE under covariate shift, but no finite-sample guarantees are provided. Later, Sugiyama and Müller (2005) further extend this work by studying an unbiased estimator under the L^2 -generalization error. Motivated by these fundamental works, there has been a flurry of follow-up studies for parametric covariate shift. Mousavi Kalan et al. (2020) introduces a statistical minimax framework and gives lower bounds for out-of-distribution generalization under the regression models of linear and one-hidden layer neural networks. Lei et al. (2021) takes a closer investigation on the minimax optimal estimator for fixed-design linear regression under covariate shift. Zhang et al. (2022) studies linear models under covariate shift where the learner has access to a small amount of target labels. In contrast, this paper focuses on the problem of covariate shift where the learner has no access to target labels.

Beyond the cases of parametric covariate shift, Cortes et al. (2010) investigate the IW estimator under the framework of statistical learning and provide a non-asymptotic upper bound on the excess

target risk for the IW estimator. Also, there has been a line of recent works on well-specified non-parametric models under covariate shift. [Kpotufe and Martinet \(2021\)](#) consider the non-parametric classification problem over the class of Hölder continuous functions and provide a new fine-grained similarity measure. Within a focus on the class of Hölder continuous functions, [Pathak et al. \(2022\)](#) introduce a novel measure of distribution mismatch between the source and target domains. Under the setting of reproducing kernel Hilbert space (RKHS), [Ma et al. \(2023\)](#); [Gogolashvili et al. \(2023\)](#) establish the optimal learning rates of kernel ridge regression (KRR) estimation. In particular, [Ma et al. \(2023\)](#) prove that KRR estimator with a carefully selected regularization parameter is minimax optimal if the covariate density ratio is uniformly bounded, and a re-weighting version of the KRR estimator using truncated covariate density ratios is minimax-optimal if the covariate density ratio has a finite second-order moment. On the other hand, [Wang \(2023\)](#) suggests the strategy of learning a predictive model built upon *pseudo-labels*. As a final remark, over-parameterized models, such as high-dimensional models and classes of neural networks, under covariate shift has drawn increasing attention from the researchers ([Byrd and Lipton, 2019](#); [Hendrycks and Dietterich, 2019](#); [Hendrycks et al., 2021](#); [Tripuraneni et al., 2021](#))).

Doubly-robust (DR) estimation *Doubly-robust (DR) estimation* combines an outcome regression with a model for treatment or selection (e.g., the propensity score), guaranteeing its consistency if at least one is correctly specified. Its foundations lie in the seminal paper by [Robins et al. \(1994\)](#) on semi-parametric theory and influence functions, and were formalized for applications by [Bang and Robins \(2005\)](#). Some implementations include the *augmented inverse propensity weighting (AIPW)* ([Robins et al., 1994](#); [Robins and Rotnitzky, 1995](#); [Bang and Robins, 2005](#)) and the *target maximum likelihood estimation (TMLE)* ([Van Der Laan and Rubin, 2006](#); [Van der Laan et al., 2011](#)), both of which leverage influence functions to correct bias. A corpus of recent studies integrate modern ML techniques for flexible nuisance estimation by using Neyman orthogonalization and sample splitting ([Chernozhukov et al., 2017, 2018](#); [Van der Laan and Rose, 2018](#); [Kennedy, 2024](#)) for retaining valid inference. The DR estimation framework has expanded to settings such as difference-in-differences ([Sant’Anna and Zhao, 2020](#); [Ning et al., 2020](#)), instrumental variables ([Okui et al., 2012](#); [Lee et al., 2023](#)), and censored data ([Bai et al., 2013](#)). While the DR methods provide robustness and potential efficiency, they require careful handling of finite-sample bias ([Kang and Schafer, 2007](#); [Funk et al., 2011](#)), near-positivity violations ([Cole and Hernán, 2008](#)), and model diagnostics ([Bang and Robins, 2005](#); [Robins et al., 1994](#)), since the correctness of at least one nuisance estimate remains crucial.

Structure-agnostic estimation The *structure-agnostic estimation* framework stands for a class of statistical methods for estimating functionals or treatment effects without assuming any parametric or structural models for the underlying data generating process. [Balakrishnan et al. \(2023\)](#) establish fundamental limits for such functional estimation, characterizing the optimal rates achievable when only minimal assumptions – such as smoothness or boundedness – are imposed. [Jin and Syrgkanis \(2024\)](#) that the DR estimators both for the average treatment effect (ATE) and the average treatment effect on the treated (ATT) attain the minimax optimal rates under the structure-agnostic framework. Their findings underscore the effectiveness of the DR learning in causal inference, particularly when relying on flexible ML methods for nuisance estimation. [Jin et al. \(2025\)](#) further study the sensitivity of structure-agnostic estimation procedures to noise, highlight cases where standard estimators fail to achieve normality or efficiency. Lastly, [Bonvini et al. \(2024\)](#) extend the framework by formalizing the DR inference under smoothness conditions. Collectively, these works aim to construct a rigorous

framework for statistical estimation and inference that minimizes reliance on structural assumptions while achieving near-optimal statistical guarantees.

2. Problem formulation

Let \mathbb{X} denote the covariate space (feature space). We consider the *source distribution* $\mathbb{P} \in \Delta(\mathbb{X} \times \mathbb{R})$ and the *target distribution* $\mathbb{Q} \in \Delta(\mathbb{X} \times \mathbb{R})$. Also, let $\mathbb{P}_X \in \Delta(\mathbb{X})$ and $\mathbb{Q}_X \in \Delta(\mathbb{X})$ denote by the marginal distributions of X under \mathbb{P} and \mathbb{Q} , respectively. We further define $\mathbb{P}_{Y|X} : \mathbb{X} \rightarrow \Delta(\mathbb{R})$ and $\mathbb{Q}_{Y|X} : \mathbb{X} \rightarrow \Delta(\mathbb{R})$ to be the conditional laws of Y given X under \mathbb{P} and \mathbb{Q} :

$$\mathbb{P}_{Y|X}(\cdot | x) := \mathbb{P}(Y \in \cdot | X = x), \quad \mathbb{Q}_{Y|X}(\cdot | x) := \mathbb{Q}(Y \in \cdot | X = x).$$

Assumption 1 (Covariate shift model) For every $x \in \mathbb{X}$,

$$\mathbb{E}_{\mathbb{P}}[Y | X = x] = \mathbb{E}_{\mathbb{Q}}[Y | X = x]. \quad (1)$$

Thus, the two distributions share the same *Bayes regression function*,

$$f^*(x) := \mathbb{E}_{\mathbb{P}}[Y | X = x] = \mathbb{E}_{\mathbb{Q}}[Y | X = x], \quad x \in \mathbb{X}.$$

Here, we emphasize that Assumption 1 does not require $\mathbb{P}_{Y|X} = \mathbb{Q}_{Y|X}$; only their Bayes regression functions must coincide. In fact, this assumption is weaker compared to the classical covariate shift model (Shimodaira, 2000), which posits a full equality of the conditional distributions.

Observational data. We observe $n_{\mathbb{P}}$ labeled samples from the source distribution \mathbb{P} ,

$$\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} := \left(O_i^{\mathbb{P}} := (X_i^{\mathbb{P}}, Y_i^{\mathbb{P}}) : i \in [n_{\mathbb{P}}] \right) \sim \mathbb{P}^{\otimes n_{\mathbb{P}}},$$

and $n_{\mathbb{Q}}$ *unlabeled* target covariates,

$$\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} = \left(X_j^{\mathbb{Q}} : j \in [n_{\mathbb{Q}}] \right) \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}.$$

Hence, the labels are available only in the source domain.

Risk and excess risk. Given a function class $\mathcal{F} \subseteq (\mathbb{X} \rightarrow \mathbb{R})$, define the μ -risk $\mathcal{R}_{\mu} : \mathcal{F} \rightarrow \mathbb{R}_+$ by

$$\mathcal{R}_{\mu}(f) := \mathbb{E}_{(X,Y) \sim \mu} \left[\{Y - f(X)\}^2 \right], \quad \mu \in \Delta(\mathbb{X} \times \mathbb{R}).$$

For $\mu \in \Delta(\mathbb{X} \times \mathbb{R})$, let $f_{\mu}^* \in \operatorname{argmin} \{ \mathcal{R}_{\mu}(f) : f \in \mathcal{F} \}$ denote a μ -risk minimizer over the function class \mathcal{F} . The *excess μ -risk* is then defined by

$$\mathcal{E}_{\mu}(f) := \mathcal{R}_{\mu}(f) - \mathcal{R}_{\mu}(f_{\mu}^*), \quad f \in \mathcal{F}. \quad (2)$$

Goal: covariate shift adaptation. Our objective is to construct an estimator $\hat{f} \in \mathcal{F}$ that achieves small excess \mathbb{Q} -risk $\mathcal{E}_{\mathbb{Q}}(\hat{f})$ with high probability.

Covariate density ratio. A central quantity in the study of covariate shift is the *covariate density ratio* between the source and target distributions. We assume that the marginal distributions \mathbb{P}_X and \mathbb{Q}_X are absolutely continuous with respect to a σ -finite reference measure $\mu_{\mathbb{X}}$ on \mathbb{X} . Let

$$p_X := \frac{d\mathbb{P}_X}{d\mu_{\mathbb{X}}} : \mathbb{X} \rightarrow \mathbb{R}_+ \quad \text{and} \quad q_X := \frac{d\mathbb{Q}_X}{d\mu_{\mathbb{X}}} : \mathbb{X} \rightarrow \mathbb{R}_+$$

denote their respective densities with respect to $\mu_{\mathbb{X}}$. The covariate density ratio is then defined as

$$\rho^*(x) := \frac{q_X(x)}{p_X(x)}, \quad x \in \mathbb{X},$$

which is assumed to be finite everywhere throughout this paper.

3. Doubly-robust (DR) covariate shift adaptation

Re-weighting with respect to the source distribution \mathbb{P} yields an alternative expression of the \mathbb{Q} -risk as the ρ^* -weighted \mathbb{P} -risk:

$$\mathcal{R}_{\mathbb{Q}}(f) = \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\rho^*(X) \{Y - f(X)\}^2 \right], \quad f \in \mathcal{F}. \quad (3)$$

The *importance-weighting (IW)* estimator (Shimodaira, 2000) is obtained by minimizing the empirical variant of the ρ^* -weighted \mathbb{P} -risk (3) over \mathcal{F} . Its key limitation is the reliance on the knowledge of the *unknown* covariate density ratio $\rho^* : \mathbb{X} \rightarrow \mathbb{R}_+$: plug-in versions of the IW estimator using an estimate $\hat{\rho} : \mathbb{X} \rightarrow \mathbb{R}_+$ for the covariate density ratio ρ^* might have high variance and degrade its performance unless the estimation of the covariate density ratio is sufficiently accurate.

The *doubly-robust (DR) covariate shift adaptation* (Kato et al., 2023) augments the IW method with a pilot regression model, and then subtracts a squared-error correction term to cancel the leading error term incurred by the density ratio estimation. Given any pilot estimates $\hat{\rho} : \mathbb{X} \rightarrow \mathbb{R}_+$ and $\hat{f}_0 : \mathbb{X} \rightarrow \mathbb{R}$ for the covariate density ratio $\rho^* : \mathbb{X} \rightarrow \mathbb{R}_+$ and the shared Bayes regression function $f^* : \mathbb{X} \rightarrow \mathbb{R}$, respectively, we define the DR empirical risk $\hat{\mathcal{R}}_{\text{DR}} : \mathcal{F} \rightarrow \mathbb{R}$ by

$$\begin{aligned} \hat{\mathcal{R}}_{\text{DR}}(f) := & \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \hat{\rho}(X_i^{\mathbb{P}}) \left[\left\{ Y_i^{\mathbb{P}} - f(X_i^{\mathbb{P}}) \right\}^2 - \left\{ \hat{f}_0(X_i^{\mathbb{P}}) - f(X_i^{\mathbb{P}}) \right\}^2 \right] \\ & + \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \left\{ \hat{f}_0(X_j^{\mathbb{Q}}) - f(X_j^{\mathbb{Q}}) \right\}^2 \end{aligned} \quad (4)$$

and the DR estimator as

$$\hat{f}_{\text{DR}} \in \operatorname{argmin} \left\{ \hat{\mathcal{R}}_{\text{DR}}(f) : f \in \mathcal{F} \right\}. \quad (5)$$

Intuitively, the pilot regression model terms \hat{f}_0 makes the risk *orthogonal* to the first-order errors in $\hat{\rho}$ (and vice-versa), yielding stability even when the pilot estimates converge slowly.

Structure-agnostic estimation. In this section, the pilot estimates $\hat{\rho} : \mathbb{X} \rightarrow \mathbb{R}_+$ and $\hat{f}_0 : \mathbb{X} \rightarrow \mathbb{R}$ are viewed as *black-boxes*: the analysis only requires the pilot estimates to achieve certain statistical error rates, not how they are obtained. This *structure-agnostic estimation* framework (Balakrishnan et al., 2023; Jin and Syrgkanis, 2024; Kennedy et al., 2024; Bonvini et al., 2024; Jin et al., 2025) reflects practice, where the nuisance estimates $\hat{\rho}$ and \hat{f}_0 can be obtained by leveraging diverse modern ML methods (e.g., LASSO (Bickel et al., 2009; Wainwright, 2009), tree-based algorithms (Syrgkanis and Zampetakis, 2020; Wager and Athey, 2018), and deep neural nets (Chen and White, 1999; Schmidt-Hieber, 2020)). Later, our finite-sample guarantees will be directly stated in terms of their estimation errors.

4. Structure-agnostic learning bounds for DR covariate shift adaptation

This section aims to develop finite-sample structure-agnostic guarantees for the doubly-robust (DR) estimator. We first state the standing assumptions, introduce the complexity measure utilized in our analysis, and finally present a high-probability bound on the excess \mathbb{Q} -risk of the DR estimator (5) together with a concrete illustration based on classes of Frobenius-norm-bounded neural networks.

Throughout this section, we consider the structure-agnostic perspective that treat the given pilot estimates $(\hat{\rho}, \hat{f}_0)$ as black-boxes; our upper bounds depend only on their estimation errors measured by the mean-squared error with respect to \mathbb{P}_X .

4.1. Assumptions

We begin by introducing the minimal assumptions under which our non-asymptotic analysis holds.

Assumption 2 (Well-specified model) $f^* \in \mathcal{F}$.

Assumption 3 (Uniform boundedness) We have $\sup \{\|f\|_\infty : f \in \mathcal{F}\} \leq 1$ and $|Y| \leq 1$ almost surely under the source distribution \mathbb{P} and the target distribution \mathbb{Q} .

Assumption 4 The pilot estimates $\hat{\rho} : \mathbb{X} \rightarrow \mathbb{R}_+$ and $\hat{f}_0 : \mathbb{X} \rightarrow \mathbb{R}$ of the true covariate density ratio $\rho^* : \mathbb{X} \rightarrow \mathbb{R}_+$ and the shared Bayes regression function $f^* \in \mathcal{F}$, respectively, satisfy

$$\|\hat{\rho}\|_\infty \leq C_{\text{dr}} < +\infty \quad \text{and} \quad \|\hat{f}_0\|_\infty \leq C_{\text{rf}} < +\infty \quad (6)$$

for some universal constants $C_{\text{dr}}, C_{\text{rf}} \in (0, +\infty)$.

Remark 4.1 Note that the uniform boundedness assumption $\|\hat{\rho}\|_\infty \leq C_{\text{dr}} < +\infty$ on the black-box ML estimate $\hat{\rho} : \mathbb{X} \rightarrow \mathbb{R}_+$ is standard for the case of the *bounded* ground-truth covariate density ratio $\rho^* : \mathbb{X} \rightarrow \mathbb{R}_+$. In particular, estimation procedures built upon the density ratio matching under the Bregman divergence (Sugiyama et al., 2010, 2012) such as *least-squares importance fitting (LSIF)* (Kanamori et al., 2009), *kernel mean matching (KMM)* (Gretton et al., 2009), *kernel unconstrained LSIF (KuLSIF)* (Kanamori et al., 2012), *Kullback-Leibler importance estimation procedure (KLIEP)* (Sugiyama et al., 2008), logistic regression-based density ratio estimation methods (Sugiyama et al., 2010, 2012), and deep density ratio estimation (Kato and Teshima, 2021; Zheng et al., 2022), focus on the minimization of a specific empirical risk over a *uniformly bounded* hypothesis class.

4.2. Uniform convergence and Rademacher complexity guarantees

We now analyze the DR estimator (5) in finite-sample regimes via uniform convergence arguments. The key complexity measure is the Rademacher complexity of the f^* -shifted version of the function class $\mathcal{F} \subseteq (\mathbb{X} \rightarrow \mathbb{R})$:

$$\mathcal{F}^* := \{f - f^* : f \in \mathcal{F}\} \subseteq (\mathbb{X} \rightarrow \mathbb{R}).$$

We first recall the definition of the Rademacher complexity for completeness.

Definition 4.1 (Rademacher complexity) Given any function class $\mathcal{G} \subseteq (\mathbb{X} \rightarrow \mathbb{R})$, the *empirical Rademacher complexity* of \mathcal{G} with respect to n sample points $\mathbf{x}_{1:n} = (x_1, x_2, \dots, x_n) \in \mathbb{X}^n$ is

$$\widehat{\mathcal{R}}_n(\mathcal{G})(\mathbf{x}_{1:n}) := \mathbb{E}_{\boldsymbol{\sigma}_{1:n} \sim \text{Unif}(\{\pm 1\}^n)} \left[\sup \left\{ \left| \frac{1}{n} \sum_{i=1}^n \sigma_i g(x_i) \right| : g \in \mathcal{G} \right\} \right]. \quad (7)$$

The *Rademacher complexity* of \mathcal{G} with respect to a probability measure $\mu \in \Delta(\mathbb{X})$ is defined by

$$\begin{aligned} \mathcal{R}_n^\mu(\mathcal{G}) &:= \mathbb{E}_{\mathbf{X}_{1:n} \sim \mu^{\otimes n}} \left[\widehat{\mathcal{R}}_n(\mathcal{G})(\mathbf{X}_{1:n}) \right] \\ &= \mathbb{E}_{(\mathbf{X}_{1:n}, \boldsymbol{\sigma}_{1:n}) \sim \mu^{\otimes n} \otimes \text{Unif}(\{\pm 1\}^n)} \left[\sup \left\{ \left| \frac{1}{n} \sum_{i=1}^n \sigma_i g(X_i) \right| : g \in \mathcal{G} \right\} \right]. \end{aligned} \quad (8)$$

With these preliminary notions in hand, one can state a structure-agnostic high-probability upper bound on the excess \mathbb{Q} -risk of the DR estimator (5) that depends only on the $L^2(\mathbb{X}, \mathbb{P}_X)$ -errors of the pilot estimates and the Rademacher complexity of \mathcal{F}^* under \mathbb{P}_X and \mathbb{Q}_X .

Theorem 4.1 (Structure-agnostic upper bound I of the DR estimator) *With Assumptions 1–4, the doubly-robust (DR) estimator (5) achieves the \mathbb{Q} -estimation error*

$$\begin{aligned} \mathcal{E}_{\mathbb{Q}}(\hat{f}_{\text{DR}}) &= \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\left\{ \hat{f}_{\text{DR}}(X) - f^*(X) \right\}^2 \right] \\ &\leq 4 \|\hat{\rho} - \rho^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} \cdot \|\hat{f}_0 - f^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} \\ &\quad + 12(2 + C_{\text{rf}}) \log\left(\frac{3}{\delta}\right) \left(\frac{C_{\text{dr}}}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right) \\ &\quad + 4(1 + C_{\text{dr}})(2 + C_{\text{rf}}) \sqrt{2 \log\left(\frac{3}{\delta}\right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right)} \\ &\quad + 8(1 + C_{\text{dr}})(2 + C_{\text{rf}}) \sqrt{\log\left(\frac{3}{\delta}\right) \left(\frac{\mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P}_X}(\mathcal{F}^*)}{\sqrt{n_{\mathbb{P}}}} + \frac{\mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q}_X}(\mathcal{F}^*)}{\sqrt{n_{\mathbb{Q}}}} \right)} \\ &\quad + 8C_{\text{dr}}(1 + C_{\text{rf}}) \mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P}_X}(\mathcal{F}^*) + 8(3 + C_{\text{rf}}) \mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q}_X}(\mathcal{F}^*) \end{aligned} \quad (9)$$

with probability at least $1 - \delta$ under the probability measure $\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}$.

The proof of Theorem 4.1 is deferred to Appendix B.2. Let us make use of the notation

$$\text{Err}_{\rho} := \|\hat{\rho} - \rho^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} \quad \text{and} \quad \text{Err}_f := \|\hat{f}_0 - f^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)}.$$

The leading bias term in (9) can be written as the product $\text{Err}_{\rho} \cdot \text{Err}_f$. This observation leads us to the following two concrete implications:

- (I1) **Having just one good pilot estimate suffices.** Assuming either $\text{Err}_\rho = o(1)$ or $\text{Err}_f = o(1)$ as $\min\{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \rightarrow \infty$ and the remaining term is bounded, we have $\text{Err}_\rho \cdot \text{Err}_f = o(1)$. Hence, the DR estimator is still consistent even when one of the pilot estimates is inaccurate; this is the finite-sample manifestation of the *double robustness* phenomenon (Robins and Rotnitzky, 1995; Robins et al., 2008): the error of the one-step corrected estimators is upper bounded by a product of estimation errors of the underlying nuisance components. To put it another way, the DR covariate shift adaptation allows us to reduce the bias incurred by the estimation error of the covariate density ratio through the aforementioned double robustness property.
- (I2) **Rate multiplication.** Suppose $\text{Err}_\rho = \tilde{O}(n^{-\alpha})$ and $\text{Err}_f = \tilde{O}(n^{-\beta})$ with $n := \min\{n_{\mathbb{P}}, n_{\mathbb{Q}}\}$. Then, their product term scales as $\tilde{O}(n^{-(\alpha+\beta)})$. In contrast, the upper bound on the excess \mathbb{Q} -risk of the IW estimator depends additively on Err_ρ , and typically requires $\alpha \geq 1/2$ to be competitive. Thus, the DR estimator (5) outperforms the IW method whenever $\alpha + \beta > 1/2$.

Since we have trivial bounds $\mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P}_X}(\mathcal{F}^*) \leq 2$ and $\mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q}_X}(\mathcal{F}^*) \leq 2$, one can simplify the excess \mathbb{Q} -risk bound (9) in Theorem 4.1 of the DR estimator as follows: With Assumptions 1–4 in hand, it follows that the DR estimator (5) achieves

$$\begin{aligned} \mathcal{E}_{\mathbb{Q}}(\hat{f}_{\text{DR}}) &\lesssim \|\hat{\rho} - \rho^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} \cdot \|\hat{f}_0 - f^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} + \sqrt{\frac{\log(\frac{1}{\delta})}{n_{\mathbb{P}}}} + \sqrt{\frac{\log(\frac{1}{\delta})}{n_{\mathbb{Q}}}} \\ &\quad + \mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P}_X}(\mathcal{F}^*) + \mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q}_X}(\mathcal{F}^*) \end{aligned} \quad (10)$$

with probability at least $1 - \delta$.

4.3. An illustration with Frobenius-norm-bounded neural networks

Let $\mathbb{X} \subseteq \mathbb{R}^{n_0}$ be a bounded domain with $\sup\{\|\mathbf{x}\|_2 : \mathbf{x} \in \mathbb{X}\} \leq R$ for some radius $R \in (0, +\infty)$. We also consider a collection of 1-Lipschitz activation functions $\{\sigma_j \in (\mathbb{R} \rightarrow \mathbb{R}) : j \in \mathbb{N}\}$ that are *positive-homogeneous* (that is, $\sigma_j(\alpha t) = \alpha \sigma_j(t)$ for every $(\alpha, t) \in \mathbb{R}_+ \times \mathbb{R}$), and that are applied *element-wise*. We are interested in a class of real-valued neural networks of depth $d \in \mathbb{N}$ over the domain $\mathbb{X} \subseteq \mathbb{R}^{n_0}$ defined as

$$\mathcal{H}_d(\mathbb{X}; M_F) := \{\text{NN}_d(\cdot; \boldsymbol{\theta}) \in (\mathbb{X} \rightarrow \mathbb{R}) : \boldsymbol{\theta} \in \Theta(M_F)\}, \quad (11)$$

where $\boldsymbol{\theta} = (\mathbf{W}_1, \dots, \mathbf{W}_d) \in \prod_{j=1}^d \mathbb{R}^{n_j \times n_{j-1}}$ denotes the model parameter consists of d parameter matrices with $n_d = 1$, and the real-valued neural network of depth d $\text{NN}_d(\cdot; \boldsymbol{\theta}) : \mathbb{X} \rightarrow \mathbb{R}$ is defined to be

$$\text{NN}_d(\mathbf{x}; \boldsymbol{\theta}) := \mathbf{W}_d \sigma_{d-1}(\mathbf{W}_{d-1} \sigma_{d-2}(\dots \sigma_1(\mathbf{W}_1 \mathbf{x}) \dots)). \quad (12)$$

Here, $M_F : [d] \rightarrow \mathbb{R}_+$ specifies upper bounds on the Frobenius norm of the parameter matrices, and the parameter space $\Theta(M_F) \subseteq \prod_{j=1}^d \mathbb{R}^{n_j \times n_{j-1}}$ is given by

$$\Theta(M_F) := \left\{ \boldsymbol{\theta} = (\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_d) \in \prod_{j=1}^d \mathbb{R}^{n_j \times n_{j-1}} : \|\mathbf{W}_j\|_F \leq M_F(j), \forall j \in [d] \right\}.$$

A key example of the above construction are ReLU networks, where every $\sigma_j : \mathbb{R} \rightarrow \mathbb{R}$ corresponds to applying the ReLU activation function $\sigma(\cdot) := \max\{0, \cdot\} : \mathbb{R} \rightarrow \mathbb{R}_+$. With the class $\mathcal{H}_d(\mathbb{X}; M_F)$ in hand, we now introduce the function class of our interest. Let $\eta : \mathbb{R} \rightarrow [-1, 1]$ be an L -Lipschitz bounded activation function such that $\eta(0) = 0$, and define

$$\mathcal{F} := \{f(\cdot; \boldsymbol{\theta}) := \eta \circ \text{NN}_d(\cdot; \boldsymbol{\theta}) \in (\mathbb{X} \rightarrow [-1, 1]) : \boldsymbol{\theta} \in \Theta(M_F)\}. \quad (13)$$

For example, the *inverse tangent activation* $\frac{2}{\pi} \arctan(\cdot) : \mathbb{R} \rightarrow [-1, 1]$ satisfies the desired properties with $L = \frac{2}{\pi}$. One can show that the Rademacher complexity of the f^* -shift version of (13) with respect to any probability measure $\mu \in \Delta(\mathbb{X})$ is of order $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

Proposition 4.1 *The Rademacher complexity of the f^* -shifted version of the neural network class $\mathcal{F} \subseteq (\mathbb{X} \rightarrow [-1, 1])$ defined as (13), $\mathcal{F}^* := \mathcal{F} - \{f^*\}$, with respect to any given probability measure $\mu \in \Delta(\mathbb{X})$ is upper bounded as*

$$\mathcal{R}_n^\mu(\mathcal{F}^*) \leq \frac{2}{\sqrt{n}} \left\{ LR \left(1 + \sqrt{(2 \log 2) d}\right) \prod_{j=1}^d M_F(j) + \sqrt{\log 2} \right\} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right). \quad (14)$$

The proof of Proposition 4.1 can be found in Appendix B.1. With Proposition 4.1 in hand, one may conclude that the DR estimator (5) achieves the following excess \mathbb{Q} -risk bound when we select the hypothesis class $\mathcal{F} \subseteq (\mathbb{X} \rightarrow [-1, 1])$ of our interest as (13): with probability at least $1 - \delta$, one has

$$\mathcal{E}_{\mathbb{Q}}(\hat{f}_{\text{DR}}) \lesssim \|\hat{\rho} - \rho^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} \cdot \|\hat{f}_0 - f^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} + \sqrt{\frac{\log(\frac{1}{\delta})}{n_{\mathbb{P}}}} + \sqrt{\frac{\log(\frac{1}{\delta})}{n_{\mathbb{Q}}}}. \quad (15)$$

Remark 4.2 We now turn our attention to the following approach that utilizes the double/debiased machine learning (DML) method (Chernozhukov et al., 2017, 2018, 2022, 2023; Foster and Syrgkanis, 2023): We first split the observations $\mathcal{D} := (\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}})$ into two subgroups \mathcal{D}_1 and \mathcal{D}_2 with the equal size, and then we estimate the covariate density ratio $\rho^* : \mathbb{X} \rightarrow \mathbb{R}_+$ and the common Bayes regression function $f^* \in \mathcal{F}$ using the first subgroup \mathcal{D}_1 to compute a nuisance estimate $\hat{\rho} : \mathbb{X} \rightarrow \mathbb{R}_+$ and a pilot estimate $\hat{f}_0 : \mathbb{X} \rightarrow \mathbb{R}$. Several results from the literature of density ratio estimation propose algorithms achieving $\|\hat{\rho} - \rho^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} = \mathcal{O}_p\left(\min\{n_{\mathbb{P}}, n_{\mathbb{Q}}\}^{-\frac{1}{2+\gamma}}\right)$ as $\min\{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \rightarrow \infty$ for any constant $\gamma \in (0, 2)$ Kanamori et al. (2012); Kato and Teshima (2021). Therefore, if the pilot estimate $\hat{f}_0 : \mathbb{X} \rightarrow \mathbb{R}$ of $f^* \in \mathcal{F}$ is consistent under the source distribution \mathbb{P} with a rate

$$\|\hat{f}_0 - f^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} = \mathcal{O}_p\left(\min\{n_{\mathbb{P}}, n_{\mathbb{Q}}\}^{-\frac{\gamma}{2(2+\gamma)}}\right) \quad \text{as} \quad \min\{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \rightarrow \infty, \quad (16)$$

then the high-probability bound (15) on the \mathbb{Q} -risk for the DR estimator (5) together with the class (13) of Frobenius-norm-bounded neural networks (constructed using the second subgroup \mathcal{D}_2) gives

$$\mathcal{E}_{\mathbb{Q}}(\hat{f}_{\text{DR}}) \lesssim \sqrt{\frac{\log(\frac{1}{\delta})}{n_{\mathbb{P}}}} + \sqrt{\frac{\log(\frac{1}{\delta})}{n_{\mathbb{Q}}}}. \quad (17)$$

To summarize, as long as the pilot black-box estimate $\hat{f}_0 : \mathbb{X} \rightarrow \mathbb{R}$ for the Bayes regression function $f^* \in \mathcal{F}$ is consistent under the source distribution \mathbb{P} with the rate of convergence (16), we are able to enhance the pilot black-box estimate $\hat{f}_0 : \mathbb{X} \rightarrow \mathbb{R}$ of $f^* \in \mathcal{F}$ to an estimator achieving the rate of convergence (17) even if it is not consistent under the target distribution \mathbb{Q} .

5. Learning bounds for DR covariate shift adaptation: parametric models

This section focuses on the doubly-robust (DR) covariate shift adaptation when the underlying function class is finite-dimensional and well-specified. Our key takeaway messages are two-fold: (i) with parametric models, *fast* $1/n$ -type rates of convergence are attainable without assuming exact knowledge of the true covariate density ratio ρ^* ; and (ii) the DR estimator achieves these rates *regardless of the statistical accuracies* of the pilot estimates $(\hat{\rho}, \hat{f}_0)$.

Parametric model. Throughout this section, let us impose Assumptions 2 and 3 and then consider a d -dimensional parameterization

$$\mathcal{F} = \left\{ f(\cdot; \boldsymbol{\theta}) \in (\mathbb{X} \rightarrow [-1, 1]) : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d \right\}, \quad (18)$$

with the ground-truth parameter $\boldsymbol{\theta}^* \in \Theta$ such that $f^*(\cdot) = f(\cdot; \boldsymbol{\theta}^*) \in \mathcal{F}$. Given any pilot estimates $(\hat{\rho}, \hat{f}_0)$, the DR empirical risk specialized to the parameterized model (18) is

$$\begin{aligned} \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}) &:= \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \hat{\rho}(X_i^{\mathbb{P}}) \left\{ \ell(Y_i^{\mathbb{P}}, f(X_i^{\mathbb{P}}; \boldsymbol{\theta})) - \ell(\hat{f}_0(X_i^{\mathbb{P}}), f(X_i^{\mathbb{P}}; \boldsymbol{\theta})) \right\} \\ &\quad + \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \ell(\hat{f}_0(X_j^{\mathbb{Q}}), f(X_j^{\mathbb{Q}}; \boldsymbol{\theta})), \end{aligned} \quad (19)$$

where $\ell(a, b) := (b - a)^2$. We define the DR estimator specialized to the parametric model (18) as

$$\hat{\boldsymbol{\theta}}_{\text{DR}} \in \operatorname{argmin} \left\{ \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \right\}, \quad \hat{f}_{\text{DR}}(\cdot) := f(\cdot; \hat{\boldsymbol{\theta}}_{\text{DR}}) \in \mathcal{F}. \quad (20)$$

Regularity and landscape conditions. We first make the following smoothness assumptions customary in classical analysis of MLE (Lehmann, 1999; Le Cam, 1956; Cramér, 1999; Van der Vaart, 2000; Lehmann and Casella, 2006)

Assumption 5 (Smoothness assumptions) Suppose the parameter space $\Theta \subseteq \mathbb{R}^d$ is *star-shaped at center* $\boldsymbol{\theta}^* \in \Theta$, i.e., $[\boldsymbol{\theta}^*, \boldsymbol{\theta}] := \{\boldsymbol{\theta}^* + \lambda(\boldsymbol{\theta} - \boldsymbol{\theta}^*) : \lambda \in [0, 1]\} \subseteq \Theta$ for all $\boldsymbol{\theta} \in \Theta$, and

- (i) For each $x \in \mathbb{X}$, the function $\boldsymbol{\theta} \in \Theta \mapsto f(x; \boldsymbol{\theta}) \in [-1, 1]$ is three-times differentiable;
- (ii) There exist absolute constants $(b_1, b_2, b_3) \in (0, +\infty)^3$ such that

$$\|\nabla_{\boldsymbol{\theta}} f(x; \boldsymbol{\theta})\|_2 \leq b_1, \quad \|\nabla_{\boldsymbol{\theta}}^2 f(x; \boldsymbol{\theta})\|_{\text{op}} \leq b_2, \quad \text{and} \quad \|\nabla_{\boldsymbol{\theta}}^3 f(x; \boldsymbol{\theta})\|_{\text{op}} \leq b_3 \quad (21)$$

for every $(x, \boldsymbol{\theta}) \in \mathbb{X} \times \Theta$.

Assumption 6 (Benign landscape of the DR empirical risk) For any realization $(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}}) \in \mathbb{O}^{n_{\mathbb{P}}} \times \mathbb{X}^{n_{\mathbb{Q}}}$, the DR empirical risk $\widehat{\mathcal{R}}_{\text{DR}} : \Theta \rightarrow \mathbb{R}$ attains a unique local minimum, which is also the global minimum.

Note that Assumption 6 is satisfied, for example, if the population version of the DR empirical risk (19) is strongly convex in a neighborhood of $\theta^* \in \Theta$, and the Hessian of the DR empirical risk (19) uniformly concentrates on that neighborhood.

Now, we are ready to establish an improved structure-agnostic upper bound on the excess \mathbb{Q} -risk of the DR estimator (20) for parametrized hypothesis classes (18), which leads to faster rates of convergence. For ease of exposition, we would like to recall the classical notion of Fisher information, which plays a role as a key quantity to measure the difficulty of parameter estimation. The μ -Fisher information matrix evaluated at $\theta \in \Theta$ is defined as

$$\mathcal{I}_\mu(\theta) := \mathbb{E}_{(X,Y) \sim \mu} [\nabla_\theta^2 \ell(Y, f(X; \theta))], \quad \theta \in \Theta, \quad (22)$$

where $\mu \in \{\mathbb{P}, \mathbb{Q}\}$ and $\ell : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ is the squared error loss. Then, one can easily observe that

$$\mathcal{I}_\mu(\theta^*) := 2\mathbb{E}_{X \sim \mu_X} [\nabla_\theta f(X; \theta^*) \{\nabla_\theta f(X; \theta^*)\}^\top], \quad (23)$$

where $\mu_X(\cdot) := \mu(\cdot \times \mathbb{R}) \in \Delta(\mathbb{X})$ refers to the covariate marginal distribution of $\mu \in \{\mathbb{P}, \mathbb{Q}\}$. We now present our main result of this section, whose proof is deferred to Section B.3:

Theorem 5.1 (Informal, see Theorem B.1) *Consider the parametrized function class (18). Under Assumptions 1–6, there exists an absolute constant $K \in (0, +\infty)$ such that with probability at least $1 - 8\delta$ under the probability measure $\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}$,*

$$\begin{aligned} \mathcal{E}_{\mathbb{Q}}(\hat{\theta}_{\text{DR}}) &= \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\left\{ f(X; \hat{\theta}_{\text{DR}}) - f^*(X) \right\}^2 \right] \\ &\leq 18K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log\left(\frac{d}{\delta}\right) \left[\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\theta^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\theta^*) \right\}}{n_{\mathbb{P}}} + \frac{d}{n_{\mathbb{Q}}} \right], \end{aligned} \quad (24)$$

provided that $\min\{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \geq \bar{\kappa} \cdot \mathcal{N}^* \log\left(\frac{d}{\delta}\right)$ for some absolute constant $\bar{\kappa} \in (0, +\infty)$, where

$$\mathcal{N}^* = \text{poly} \left(d, \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\theta^*) \right\|_{\text{op}}, \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\theta^*) \mathcal{I}_{\mathbb{P}}(\theta^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\theta^*) \right\|_{\text{op}} \right).$$

Interpretations and implications Theorem 5.1 shows that, for well-specified parametric models, the DR estimator (20) achieves a fast and instance-dependent upper bound on the excess \mathbb{Q} -risk that decouples the contributions of the source and target samples to the bound:

$$\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\theta^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\theta^*) \right\}}{n_{\mathbb{P}}} : \text{contributed by the source data}, \quad \frac{d}{n_{\mathbb{Q}}} : \text{contributed by the target data},$$

up to logarithmic factors. Here, the trace factor $\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\theta^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\theta^*) \right\}$ quantifies the *Fisher information mismatch* between \mathbb{P} and \mathbb{Q} , and is the only way in which covariate shift affects the leading constant. Notably, the excess \mathbb{Q} -risk bound in Theorem 5.1 holds *without* access to the true covariate density ratio ρ^* , and is *independent of the statistical accuracies of the pilot estimates* $(\hat{\rho}, \hat{f}_0)$.

We also discuss some appealing attributes of DR covariate shift adaptation and its fast $1/n$ -type convergence guarantee (24) for well-specified parametric models provided in Theorem 5.1:

- *Fast rates of convergence under covariate shift without knowing ρ^** : The excess \mathbb{Q} -risk bound (24) of the DR estimator (20) matches the fast $1/n$ -rate behavior, where $n := \min\{n_{\mathbb{P}}, n_{\mathbb{Q}}\}$, known to be achievable in parametric models, yet it does so *without* requiring an exact knowledge (or a consistent estimate) of the covariate density ratio $\rho^* : \mathbb{X} \rightarrow \mathbb{R}_+$.
- *Pilot-agnostic tightness of the rates*: The rate of convergence for the DR estimator (20) from the excess \mathbb{Q} -risk bound (24) does not degrade with the quality of pilot estimates $(\hat{\rho}, \hat{f}_0)$; any black-box pilot estimates suffice.
- *No boundedness assumption on the covariate density ratio*: Unlike the prior works on covariate shift (e.g., (Cortes et al., 2010; Ma et al., 2023)), we make no boundedness assumptions on the covariate density ratio ρ^* , broadening applicability of our results.

It would be worth pointing out the trace factor $\text{Trace}\{\mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*)\mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*)\}$, which is different from the trace factors appeared in the excess \mathbb{Q} -risk bounds for the vanilla MLE and the weighted MLE of (Ge et al., 2024). However, on the closer look, (Ge et al., 2024) assumes the boundedness of the covariate density ratio, under which their excess \mathbb{Q} -risk bound for the weighted MLE (see Theorem 5.2 therein) can be translated to the same trace factor as in the bound (24) of Theorem 5.1.

6. Discussion

This paper establishes the first finite-sample guarantees for doubly-robust (DR) covariate shift adaptation, complementing the existing asymptotic analysis (Kato et al., 2023) and clarifying the role of pilot estimates, sample allocation, and parametric modeling for the Bayes regression function. The structure-agnostic upper bound (9) of the DR estimator (5) shows that the leading bias term scales as the product of statistical error rates of the pilot estimates, providing a non-asymptotic demonstration of the celebrated double robustness phenomenon (Robins and Rotnitzky, 1995; Robins et al., 2008): one consistent pilot estimate suffices to obtain the consistency of the one-step corrected estimators, and the joint improvement yields multiplicative gains. The decomposition of the DR empirical risk (4) highlights how the labeled source samples primarily benefit the pilot regression model, while the unlabeled target covariates strengthen the effect of the pilot estimate of the covariate density ratio, offering practical guidance on data collection under budget constraints in the target domain. Within well-specified parametric models, our analysis of the DR estimator (20) via modern techniques for finite-sample analysis of MLE yields a non-asymptotic fast $1/n$ -type convergence guarantee, which is independent of the statistical accuracies of pilot black-box estimates. In this result, the difficulty of learning a predictive model under covariate shift is quantified by the Fisher information mismatch between the source and target distributions. Together, the findings in this paper demonstrate that the DR covariate shift adaptation combines asymptotic efficiency results with strong finite-sample out-of-distribution generalization bounds.

References

- Xiaofei Bai, Anastasios A Tsiatis, and Sean M O’Brien. Doubly-robust estimators of treatment-specific survival distributions in observational studies with stratified sampling. *Biometrics*, 69(4):830–839, 2013.
- Sivaraman Balakrishnan, Edward H Kennedy, and Larry Wasserman. The fundamental limits of structure-agnostic functional estimation. *arXiv preprint arXiv:2305.04116*, 2023.
- Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.
- Peter J. Bickel, Ya’acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009. ISSN 00905364, 21688966. URL <http://www.jstor.org/stable/30243685>.
- John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447, 2007.
- Matteo Bonvini, Edward H Kennedy, Oliver Dukes, and Sivaraman Balakrishnan. Doubly-robust inference and optimality in structure-agnostic models with smoothness. *arXiv preprint arXiv:2405.08525*, 2024.
- Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International conference on machine learning*, pages 872–881. PMLR, 2019.
- Xiaohong Chen and Halbert White. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265, 2017.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL <https://doi.org/10.1111/ectj.12097>.
- Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022.
- Victor Chernozhukov, Whitney K Newey, and Rahul Singh. A simple and general debiased machine learning theorem with finite-sample guarantees. *Biometrika*, 110(1):257–264, 2023.
- Stephen R. Cole and Miguel A. Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664, 2008.
- Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

- Harald Cramér. *Mathematical methods of statistics*, volume 9. Princeton university press, 1999.
- John Duchi. Probability bounds. URL: http://www.cs.berkeley.edu/jduchi/projects/probability_bounds.pdf, 2009.
- Geli Fei and Bing Liu. Social media text classification under negative covariate shift. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2347–2356, 2015.
- Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.
- Mary J. Funk, Daniel Westreich, Charlotte Wiesen, Til Stürmer, Susan M. Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767, 2011.
- Jiawei Ge, Shange Tang, Jianqing Fan, Cong Ma, and Chi Jin. Maximum likelihood estimation is all you need for well-specified covariate shift. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=eoTCKKOGIs>.
- Davit Gogolashvili, Matteo Zecchin, Motonobu Kanagawa, Marios Kountouris, and Maurizio Filippone. When is importance weighting correction needed for covariate shift adaptation? *arXiv preprint arXiv:2303.04020*, 2023.
- Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.
- Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, Bernhard Schölkopf, et al. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.
- Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.
- László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.
- Ehsan Hajiramezanali, Siamak Zamani Dadaneh, Alireza Karbalayghareh, Mingyuan Zhou, and Xiaoning Qian. Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data. *Advances in Neural Information Processing Systems*, 31, 2018.
- Ali Hassan, Robert Damper, and Mahesan Niranjan. On acoustic emotion recognition: compensating for covariate shift. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1458–1468, 2013.
- Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

- Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021.
- Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in NLP. In Annie Zaenen and Antal van den Bosch, editors, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL <https://aclanthology.org/P07-1034/>.
- Jikai Jin and Vasilis Syrgkanis. Structure-agnostic optimality of doubly robust learning for treatment effect estimation. *arXiv preprint arXiv:2402.14264*, 2024.
- Jikai Jin, Lester Mackey, and Vasilis Syrgkanis. It’s hard to be normal: The impact of noise on structure-agnostic estimation. *arXiv preprint arXiv:2507.02275*, 2025.
- Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.
- Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86:335–367, 2012.
- Jin-Dong Y. Kang and Joseph L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539, 2007.
- Masahiro Kato and Takeshi Teshima. Non-negative bregman divergence minimization for deep direct density ratio estimation. In *International Conference on Machine Learning*, pages 5320–5333. PMLR, 2021.
- Masahiro Kato, Kota Matsui, and Ryo Inokuchi. Double debiased covariate shift adaptation robust to density-ratio estimation. *arXiv preprint arXiv:2310.16638*, 2023.
- Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of statistical methods for precision medicine*, pages 207–236, 2024.
- Edward H Kennedy, Sivaraman Balakrishnan, James M Robins, and Larry Wasserman. Minimax rates for heterogeneous causal effect estimation. *The Annals of Statistics*, 52(2):793–816, 2024.
- T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060 – 1077, 2005. doi: 10.1214/009117905000000044. URL <https://doi.org/10.1214/009117905000000044>.
- Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Bal-subramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.
- Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302 – 2329, 2011. doi: 10.1214/11-AOS894. URL <https://doi.org/10.1214/11-AOS894>.

- Samory Kpotufe and Guillaume Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021.
- Lucien Le Cam. On the asymptotic theory of estimation and testing hypotheses. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3, pages 129–157. University of California Press, 1956.
- Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.
- Youjin Lee, Edward H Kennedy, and Nandita Mitra. Doubly robust nonparametric instrumental variable estimators for survival outcomes. *Biostatistics*, 24(2):518–537, 2023.
- Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.
- Erich Leo Lehmann. *Elements of large-sample theory*. Springer, 1999.
- Qi Lei, Wei Hu, and Jason Lee. Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning*, pages 6164–6174. PMLR, 2021.
- Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in rkhs-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.
- Mohammadreza Mousavi Kalan, Zalan Fabian, Salman Avestimehr, and Mahdi Soltanolkotabi. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. *Advances in Neural Information Processing Systems*, 33:1959–1969, 2020.
- Yang Ning, Sida Peng, and Jing Tao. Doubly robust semiparametric difference-in-differences estimators with high-dimensional data. *arXiv preprint arXiv:2009.03151*, 2020.
- Ryo Okui, Dylan S Small, Zhiqiang Tan, and James M Robins. Doubly robust instrumental variable regression. *Statistica Sinica*, pages 173–205, 2012.
- Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.
- Reese Pathak, Cong Ma, and Martin Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, pages 17517–17530. PMLR, 2022.
- Joaquin Quiñero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 12 2008. ISBN 9780262255103. doi: 10.7551/mitpress/9780262170055.001.0001. URL <https://doi.org/10.7551/mitpress/9780262170055.001.0001>.
- Sashank Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift correction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

- James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, volume 2, pages 335–422. Institute of Mathematical Statistics, 2008.
- James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.
- James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.
- Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.
- Pedro HC Sant’Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of econometrics*, 219(1):101–122, 2020.
- Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020. doi: 10.1214/19-AOS1875. URL <https://doi.org/10.1214/19-AOS1875>.
- Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.
- Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.
- Masashi Sugiyama and Klaus-Robert Müller. Model selection under covariate shift. In *International Conference on Artificial Neural Networks*, pages 235–240. Springer, 2005.
- Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007a.
- Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20, 2007b.
- Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul Von Büna, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.
- Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density ratio estimation: A comprehensive review (statistical experiment and its related topics). , 1703:10–31, 2010.
- Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64:1009–1044, 2012.
- Vasilis Syrgkanis and Manolis Zampetakis. Estimation and inference with trees and forests in high dimensions. In *Conference on learning theory*, pages 3453–3454. PMLR, 2020.

- Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Overparameterization improves robustness to covariate shift in high dimensions. *Advances in Neural Information Processing Systems*, 34: 13883–13897, 2021.
- Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.
- Devis Tuia, E Pasolli, and William J Emery. Using active learning to adapt remote sensing image classifiers. *Remote Sensing of Environment*, 115(9):2232–2242, 2011.
- Mark J Van der Laan and Sherri Rose. *Targeted learning in data science*. Springer, 2018.
- Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.
- Mark J Van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.
- Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.
- Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.
- Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5): 2183–2202, 2009.
- Kaizheng Wang. Pseudo-labeling for kernel ridge regression under covariate shift. *arXiv preprint arXiv:2302.10160*, 2023.
- Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Health insurance market risk assessment: Covariate shift and k-anonymity. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 226–234. SIAM, 2015.
- Makoto Yamada, Masashi Sugiyama, and Tomoko Matsui. Covariate shift adaptation for semi-supervised speaker identification. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1661–1664. IEEE, 2009.
- Xuhui Zhang, Jose Blanchet, Soumyadip Ghosh, and Mark S Squillante. A class of geometric structures in transfer learning: Minimax bounds and optimality. In *International Conference on Artificial Intelligence and Statistics*, pages 3794–3820. PMLR, 2022.
- Siming Zheng, GUOHAO SHEN, Yuling Jiao, Yuanyuan Lin, and Jian Huang. An error analysis of deep density-ratio estimation with bregman divergence, 2022. URL <https://openreview.net/forum?id=df0BSd3tF9p>.

Appendix A. Preliminary facts

This section is devoted to collect a couple of useful preliminary facts for our analysis. The following contraction lemma is a modified version of *Theorem 4.12* of [Ledoux and Talagrand \(2013\)](#) that was established in [Duchi \(2009\)](#). See *Theorem 7* therein for the proof of Lemma 1.

Lemma 1 (The Ledoux-Talagrand contraction principle) *Let $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be a non-decreasing convex function, and $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$, $i \in [n]$, are L -Lipschitz continuous functions such that $\phi_i(0) = 0$. Then for any $T \subseteq \mathbb{R}^n$, we have*

$$\mathbb{E}_{\sigma_{1:n} \sim \text{Unif}(\{\pm 1\}^n)} \left[f \left\{ \frac{1}{2} \sup_{\mathbf{t}_{1:n} \in T} \left| \sum_{i=1}^n \sigma_i \phi_i(t_i) \right| \right\} \right] \leq \mathbb{E}_{\sigma_{1:n} \sim \text{Unif}(\{\pm 1\}^n)} \left[f \left(L \cdot \sup_{\mathbf{t}_{1:n} \in T} \left| \sum_{i=1}^n \sigma_i t_i \right| \right) \right].$$

In particular, if we let $f(t) = t$ for $t \in \mathbb{R}_+$, then we obtain

$$\mathbb{E}_{\sigma_{1:n} \sim \text{Unif}(\{\pm 1\}^n)} \left[\sup_{\mathbf{t}_{1:n} \in T} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \phi_i(t_i) \right| \right] \leq 2L \mathbb{E}_{\sigma_{1:n} \sim \text{Unif}(\{\pm 1\}^n)} \left[\sup_{\mathbf{t}_{1:n} \in T} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i t_i \right| \right]. \quad (25)$$

The following is a well-known standard deviation inequality for controlling the maxima of empirical processes; see *Theorem 1.1* in [Klein and Rio \(2005\)](#).

Lemma 2 (Classical Talagrand's concentration inequality) *Let $\mathcal{F} \subseteq (\mathbb{X} \rightarrow [-B, B])$ be a function class and $\mathbf{X}_{1:n} = (X_1, X_2, \dots, X_n) \sim \mathbb{P}^{\otimes n}$ for some $\mathbb{P} \in \Delta(\mathbb{X})$. We define*

$$Z := \sup \left\{ \left(\hat{\mathbb{P}} - \mathbb{P} \right) (f) := \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_{X \sim \mathbb{P}} [f(X)] : f(\cdot) \in \mathcal{F} \right\},$$

and $v^2 := \sup \{ \text{Var}_{X \sim \mathbb{P}} [f(X)] : f \in \mathcal{F} \}$, where $\hat{\mathbb{P}} := \frac{1}{n} \sum_{i=1}^n \delta_{X_i} \in \Delta(\mathbb{X})$ denotes the empirical measure for the n samples $\mathbf{X}_{1:n} \sim \mathbb{P}^{\otimes n}$. Then, it holds for every $x \in \mathbb{R}_+$ that

$$\mathbb{P} \{ Z > \mathbb{E}[Z] + x \} \leq \exp \left(- \frac{nx^2}{4B\mathbb{E}[Z] + 2v^2 + 3Bx} \right). \quad (26)$$

In particular, for any given $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ that

$$Z - \mathbb{E}[Z] \leq \frac{3B \log(\frac{1}{\delta})}{n} + 2\sqrt{\frac{B\mathbb{E}[Z] \log(\frac{1}{\delta})}{n}} + \sqrt{\frac{2v^2 \log(\frac{1}{\delta})}{n}} \quad (27)$$

under the probability measure $\mathbb{P}^{\otimes n}$.

Another key technical result is the following generic version of the Bernstein inequality for random vectors, which plays a significant role in establishing concentration properties of the gradient of the DR empirical risk (19) with respect to the parameter vector θ . Check Lemma 12 in Section D.1 for further details.

Lemma 3 Suppose $\mathbb{P} \in \Delta(\mathbb{R}^d)$ satisfies $\mathbb{E}_{\mathbf{X} \sim \mathbb{P}}[\mathbf{X}] = \mathbf{0}_d$ and $\mathcal{V} := \mathbb{E}_{\mathbf{X} \sim \mathbb{P}}[\|\mathbf{X}\|_2^2] < +\infty$. Define

$$\mathcal{B}(\alpha) := \inf \left\{ t \in (0, +\infty) : \mathbb{E}_{\mathbf{X} \sim \mathbb{P}} \left[\exp \left\{ \left(\frac{\|\mathbf{X}\|_2}{t} \right)^\alpha \right\} \right] \leq 2 \right\}, \quad \alpha \in [1, +\infty), \quad (28)$$

and assume that $\mathcal{B}(\alpha) < +\infty$ for some $\alpha \in [1, +\infty)$. Then, there is an absolute constant $C > 0$ such that for any given $\delta \in (0, 1)$, we have

$$\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \right\|_2 \leq C \left[\sqrt{\frac{\mathcal{V} \log(\frac{d}{\delta})}{n}} + \mathcal{B}(\alpha) \log^{\frac{1}{\alpha}} \left\{ \frac{\mathcal{B}(\alpha)}{\sqrt{\mathcal{V}}} \right\} \frac{\mathcal{V} \log(\frac{d}{\delta})}{n} \right] \quad (29)$$

under $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$, with probability at least $1 - \delta$.

We refer to *Proposition 2* in [Koltchinskii et al. \(2011\)](#) for the proof of Lemma 3.

Lastly, the following lemma provides a standard upper bound on the Rademacher complexity of finite hypothesis classes.

Lemma 4 Let $\mathcal{F} \subseteq (\mathbb{X} \rightarrow [-B, B])$ be a finite function class, i.e., $|\mathcal{F}| < +\infty$. Then, it holds that

$$\mathcal{R}_n^\mu(\mathcal{F}) \leq 2B \sqrt{\frac{\log(2|\mathcal{F}|)}{n}}, \quad (30)$$

for any probability measure $\mu \in \Delta(\mathbb{X})$.

Appendix B. Proofs for Section 3

B.1. Proof of Proposition 4.1

We first observe for any model parameter $\boldsymbol{\theta} = (\mathbf{W}_1, \dots, \mathbf{W}_d) \in \Theta(M_F)$ that

$$-\text{NN}_d(\mathbf{x}; \boldsymbol{\theta}) = -\text{NN}_d(\mathbf{x}; (\mathbf{W}_1, \dots, \mathbf{W}_d)) = \text{NN}_d(\mathbf{x}; (\mathbf{W}_1, \dots, -\mathbf{W}_d)), \quad \forall \mathbf{x} \in \mathbb{X},$$

together with $(\mathbf{W}_1, \dots, -\mathbf{W}_d) \in \Theta(M_F)$. This observation implies that

$$\mathcal{H}_d(\mathbb{X}; M_F) = -\mathcal{H}_d(\mathbb{X}; M_F) = \{-\text{NN}_d(\cdot; \boldsymbol{\theta}) \in (\mathbb{X} \rightarrow \mathbb{R}) : \boldsymbol{\theta} \in \Theta(M_F)\}. \quad (31)$$

With the observation (31) in hand, one can realize from *Theorem 1* in [Golowich et al. \(2018\)](#) that

$$\begin{aligned} & \mathcal{R}_n^\mu(\mathcal{H}_d(\mathbb{X}; M_F)) \\ &= \mathbb{E}_{(\mathbf{X}_{1:n}, \boldsymbol{\sigma}_{1:n}) \sim \mu^{\otimes n} \otimes \text{Unif}(\{\pm 1\}^n)} \left[\sup \left\{ \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \text{NN}_d(X_i; \boldsymbol{\theta}) \right| : \boldsymbol{\theta} \in \Theta(M_F) \right\} \right] \\ &\leq \frac{R \left(1 + \sqrt{(2 \log 2) d} \right) \prod_{j=1}^d M_F(j)}{\sqrt{n}}. \end{aligned} \quad (32)$$

On the other hand, by virtue of the Ledoux-Talagrand contraction principle (Lemma 1), we obtain that

$$\begin{aligned}
 & \mathcal{R}_n^\mu(\mathcal{F}) \\
 &= \mathbb{E}_{\mathbf{X}_{1:n} \sim \mu^{\otimes n}} \left[\widehat{\mathcal{R}}_n(\mathcal{F})(\mathbf{X}_{1:n}) \right] \\
 &= \mathbb{E}_{\mathbf{X}_{1:n} \sim \mu^{\otimes n}} \left[\mathbb{E}_{\boldsymbol{\sigma}_{1:n} \sim \text{Unif}(\{\pm 1\}^n)} \left[\sup \left\{ \left| \frac{1}{n} \sum_{i=1}^n \eta \{ \text{NN}_d(X_i; \boldsymbol{\theta}) \} \right| : \boldsymbol{\theta} \in \Theta(M_F) \right\} \right] \right] \\
 &\leq 2L \cdot \mathbb{E}_{\mathbf{X}_{1:n} \sim \mu^{\otimes n}} \left[\mathbb{E}_{\boldsymbol{\sigma}_{1:n} \sim \text{Unif}(\{\pm 1\}^n)} \left[\sup \left\{ \left| \frac{1}{n} \sum_{i=1}^n \text{NN}_d(X_i; \boldsymbol{\theta}) \right| : \boldsymbol{\theta} \in \Theta(M_F) \right\} \right] \right] \quad (33) \\
 &= 2L \cdot \mathbb{E}_{\mathbf{X}_{1:n} \sim \mu^{\otimes n}} \left[\widehat{\mathcal{R}}_n(\mathcal{H}_d(\mathbb{X}; M_F))(\mathbf{X}_{1:n}) \right] \\
 &= 2L \cdot \mathcal{R}_n^\mu(\mathcal{H}_d(\mathbb{X}; M_F)) \\
 &\stackrel{(a)}{\leq} \frac{2LR \left(1 + \sqrt{(2 \log 2) d} \right) \prod_{j=1}^d M_F(j)}{\sqrt{n}},
 \end{aligned}$$

where the step (a) follows from the upper bound (32) on the Rademacher complexity of $\mathcal{H}_d(\mathbb{X}; M_F)$. Hence, one can reveal that

$$\begin{aligned}
 \mathcal{R}_n^\mu(\mathcal{F}^*) &\stackrel{(b)}{\leq} \mathcal{R}_n^\mu(\mathcal{F}) + \mathcal{R}_n^\mu(\{f^*\}) \\
 &\stackrel{(c)}{\leq} \frac{2LR \left(1 + \sqrt{(2 \log 2) d} \right) \prod_{j=1}^d M_F(j)}{\sqrt{n}} + \frac{2\sqrt{\log 2}}{\sqrt{n}},
 \end{aligned}$$

which thus completes the proof of Proposition 4.1, where the step (b) holds by the triangle inequality and the step (c) invokes the bound (33) and the standard upper bound on the Rademacher complexity of finite hypothesis classes (see Lemma 4 for details).

B.2. Proof of Theorem 4.1

We first provide a formal definition of the *doubly-robust (DR) empirical risk* $\widehat{\mathcal{R}}_{\text{DR}} : \mathbb{O}^{n_P} \times \mathbb{X}^{n_Q} \rightarrow (\mathcal{F} \rightarrow \mathbb{R})$, where

$$\begin{aligned}
 \widehat{\mathcal{R}}_{\text{DR}}(\mathbf{o}_{1:n_P}, \mathbf{x}_{1:n_Q})(f) &:= \frac{1}{n_P} \sum_{i=1}^{n_P} \hat{\rho}(x_i^P) \left[\left\{ y_i^P - f(x_i^P) \right\}^2 - \left\{ \hat{f}_0(x_i^P) - f(x_i^P) \right\}^2 \right] \\
 &\quad + \frac{1}{n_Q} \sum_{j=1}^{n_Q} \left\{ \hat{f}_0(X_j^Q) - f(X_j^Q) \right\}^2, \quad (34)
 \end{aligned}$$

and define the *DR population risk* $\overline{\mathcal{R}} : \mathcal{F} \rightarrow \mathbb{R}$ by

$$\overline{\mathcal{R}}(f) := \mathbb{E}_{(\mathbf{O}_{1:n_P}^P, \mathbf{X}_{1:n_Q}^Q) \sim \mathbb{P}^{\otimes n_P} \otimes \mathbb{Q}_X^{\otimes n_Q}} \left[\widehat{\mathcal{R}}_{\text{DR}}(\mathbf{O}_{1:n_P}^P, \mathbf{X}_{1:n_Q}^Q)(f) \right], \quad f \in \mathcal{F}. \quad (35)$$

Here, we note that $\widehat{\mathcal{R}}_{\text{DR}} = \widehat{\mathcal{R}}_{\text{DR}}(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}}) : \mathcal{F} \rightarrow \mathbb{R}$ under $(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}}) \sim \mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}$. Then, one can decompose the DR population risk $\overline{\mathcal{R}} : \mathcal{F} \rightarrow \mathbb{R}$ as follows:

$$\begin{aligned}
& \overline{\mathcal{R}}(f) \\
&= \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\hat{\rho}(X) \{Y - f(X)\}^2 \right] - \mathbb{E}_{X \sim \mathbb{P}_X} \left[\hat{\rho}(X) \left\{ \hat{f}_0(X) - f(X) \right\}^2 \right] \\
&\quad + \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\left\{ \hat{f}_0(X) - f(X) \right\}^2 \right] \\
&= \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\rho^*(X) \{Y - f(X)\}^2 \right] \\
&\quad + \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left[\{Y - f(X)\}^2 - \left\{ \hat{f}_0(X) - f(X) \right\}^2 \right] \right] \\
&\stackrel{(a)}{=} \mathcal{R}_{\mathbb{Q}}(f) + \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left[\{Y - f(X)\}^2 - \left\{ \hat{f}_0(X) - f(X) \right\}^2 \right] \right],
\end{aligned} \tag{36}$$

where the step (a) follows due to the observation (3). The definition of the DR estimator (??) yields the following *basic inequality*: $0 \leq \widehat{\mathcal{R}}_{\text{DR}}(f) - \widehat{\mathcal{R}}_{\text{DR}}(\hat{f}_{\text{DR}})$ for every $f \in \mathcal{F}$. Thus, we have

$$\begin{aligned}
0 &\stackrel{(b)}{\leq} \widehat{\mathcal{R}}_{\text{DR}}(f^*) - \widehat{\mathcal{R}}_{\text{DR}}(\hat{f}_{\text{DR}}) \\
&= \left\{ \widehat{\mathcal{R}}_{\text{DR}}(f^*) - \overline{\mathcal{R}}(f^*) \right\} + \left\{ \overline{\mathcal{R}}(f^*) - \mathcal{R}_{\mathbb{Q}}(f^*) \right\} - \underbrace{\left\{ \mathcal{R}_{\mathbb{Q}}(\hat{f}_{\text{DR}}) - \mathcal{R}_{\mathbb{Q}}(f^*) \right\}}_{= \mathcal{E}_{\mathbb{Q}}(\hat{f}_{\text{DR}})} \\
&\quad - \left\{ \overline{\mathcal{R}}(\hat{f}_{\text{DR}}) - \mathcal{R}_{\mathbb{Q}}(\hat{f}_{\text{DR}}) \right\} - \left\{ \widehat{\mathcal{R}}_{\text{DR}}(\hat{f}_{\text{DR}}) - \overline{\mathcal{R}}(\hat{f}_{\text{DR}}) \right\},
\end{aligned} \tag{37}$$

where the step (b) holds by the well-specification assumption of the model $f^* \in \mathcal{F}$. It follows that

$$\begin{aligned}
\mathcal{E}_{\mathbb{Q}}(\hat{f}_{\text{DR}}) &= \mathcal{R}_{\mathbb{Q}}(\hat{f}_{\text{DR}}) - \mathcal{R}_{\mathbb{Q}}(f^*) \\
&\leq \underbrace{\left\{ \overline{\mathcal{R}}(f^*) - \mathcal{R}_{\mathbb{Q}}(f^*) \right\} - \left\{ \overline{\mathcal{R}}(\hat{f}_{\text{DR}}) - \mathcal{R}_{\mathbb{Q}}(\hat{f}_{\text{DR}}) \right\}}_{=: \text{(T1)}} \\
&\quad + \underbrace{\left\{ \widehat{\mathcal{R}}_{\text{DR}}(f^*) - \widehat{\mathcal{R}}_{\text{DR}}(\hat{f}_{\text{DR}}) \right\} - \left\{ \overline{\mathcal{R}}(f^*) - \overline{\mathcal{R}}(\hat{f}_{\text{DR}}) \right\}}_{=: \text{(T2)}}.
\end{aligned} \tag{38}$$

Bounding the term (T1): With the decomposition (38) in hand, let us first take a closer inspection on the first term (T1).

$$\begin{aligned}
\text{(T1)} &\stackrel{(c)}{=} \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left[\{Y - f^*(X)\}^2 - \left\{ \hat{f}_0(X) - f^*(X) \right\}^2 \right. \right. \\
&\quad \left. \left. - \left\{ Y - \hat{f}_{\text{DR}}(X) \right\}^2 + \left\{ \hat{f}_0(X) - \hat{f}_{\text{DR}}(X) \right\}^2 \right] \right] \\
&= 2\mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left\{ Y - \hat{f}_0(X) \right\} \left\{ \hat{f}_{\text{DR}}(X) - f^*(X) \right\} \right] \\
&= 2\mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left\{ f^*(X) - \hat{f}_0(X) \right\} \left\{ \hat{f}_{\text{DR}}(X) - f^*(X) \right\} \right],
\end{aligned}$$

where the step (c) utilizes the decomposition (36) of the DR population risk $\overline{\mathcal{R}} : \mathcal{F} \rightarrow \mathbb{R}$. Therefore, we have

$$\begin{aligned}
 (\text{T1}) &\leq 2 \left| \mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\} \{f^*(X) - \hat{f}_0(X)\} \{ \hat{f}_{\text{DR}}(X) - f^*(X) \} \right] \right| \\
 &\leq 2 \mathbb{E}_{X \sim \mathbb{P}_X} \left[|\hat{\rho}(X) - \rho^*(X)| \cdot |f^*(X) - \hat{f}_0(X)| \cdot |\hat{f}_{\text{DR}}(X) - f^*(X)| \right] \\
 &\stackrel{(d)}{\leq} 4 \left\{ \mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\}^2 \right] \right\}^{\frac{1}{2}} \left\{ \mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{f}_0(X) - f^*(X)\}^2 \right] \right\}^{\frac{1}{2}} \\
 &= 4 \|\hat{\rho} - \rho^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} \cdot \|\hat{f}_0 - f^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)},
 \end{aligned} \tag{39}$$

where the step (d) holds due to the Cauchy-Schwarz inequality together with the fact that $|\hat{f}_{\text{DR}}(x) - f^*(x)| \leq \|\hat{f}_{\text{DR}}\|_{\infty} + \|f^*\|_{\infty} \leq 2$, follows from Assumption 3.

Bounding the term (T2): With regards to the term (T2), we make use of tools from the empirical processes theory in order to establish its upper bound. First, we observe for any $f \in \mathcal{F}$ that

$$\begin{aligned}
 &\hat{\mathcal{R}}_{\text{DR}}(f^*) - \hat{\mathcal{R}}_{\text{DR}}(f) \\
 &= \frac{2}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \hat{\rho}(X_i^{\mathbb{P}}) \{f(X_i^{\mathbb{P}}) - f^*(X_i^{\mathbb{P}})\} \{Y_i^{\mathbb{P}} - \hat{f}_0(X_i^{\mathbb{P}})\} \\
 &\quad + \frac{2}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \{f(X_j^{\mathbb{Q}}) - f^*(X_j^{\mathbb{Q}})\} \{\hat{f}_0(X_j^{\mathbb{Q}}) - f^*(X_j^{\mathbb{Q}})\} \\
 &\quad - \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \{f(X_j^{\mathbb{Q}}) - f^*(X_j^{\mathbb{Q}})\}^2.
 \end{aligned} \tag{40}$$

With this observation in hand, it is seen that

$$\begin{aligned}
 (\text{T2}) &\leq \sup \left\{ \left| \left\{ \hat{\mathcal{R}}_{\text{DR}}(f^*) - \hat{\mathcal{R}}_{\text{DR}}(f) \right\} - \left\{ \overline{\mathcal{R}}(f^*) - \overline{\mathcal{R}}(f) \right\} \right| : f \in \mathcal{F} \right\} \\
 &\stackrel{(c)}{\leq} 2 \sup \left\{ \left| \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) \right| : \varphi(\cdot) \in \mathcal{F}^* \right\} + 2 \sup \left\{ \left| \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) \right| : \varphi \in \mathcal{F}^* \right\} \\
 &\quad + \sup \left\{ \left| \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) \right| : \varphi \in \mathcal{F}^* \right\} \\
 &= 2 \sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\
 &\quad + 2 \sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\
 &\quad + \sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\},
 \end{aligned} \tag{41}$$

where the step (e) follows due to the triangle inequality, $-\mathcal{F}^* := \{-\varphi : \varphi \in \mathcal{F}^*\}$, and the functions $\mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}} : (\mathbb{X} \rightarrow \mathbb{R}) \rightarrow (\mathbb{O}^{n_{\mathbb{P}}} \rightarrow \mathbb{R})$ and $\{\mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(l)} : (\mathbb{X} \rightarrow \mathbb{R}) \rightarrow (\mathbb{X}^{n_{\mathbb{Q}}} \rightarrow \mathbb{R}) : l \in [2]\}$ are defined as

$$\begin{aligned}\mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{o}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) &:= \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \hat{\rho}(x_i^{n_{\mathbb{P}}}) \varphi(x_i^{n_{\mathbb{P}}}) \left\{ y_i^{\mathbb{P}} - \hat{f}_0(x_i^{n_{\mathbb{P}}}) \right\} \\ &\quad - \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\hat{\rho}(X) \varphi(X) \left\{ Y - \hat{f}_0(X) \right\} \right], \\ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{x}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) &:= \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \varphi(x_j^{\mathbb{Q}}) \left\{ \hat{f}_0(x_j^{\mathbb{Q}}) - f^*(x_j^{\mathbb{Q}}) \right\} \\ &\quad - \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\varphi(X) \left\{ \hat{f}_0(X) - f^*(X) \right\} \right], \\ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{x}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) &:= \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \left\{ \varphi(x_j^{\mathbb{Q}}) \right\}^2 - \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\left\{ \varphi(X) \right\}^2 \right].\end{aligned}\tag{42}$$

If $\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}$ and $\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}$, then

- $\left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) = \left(\hat{\mathbb{P}} - \mathbb{P} \right) \left[\hat{\rho}(X) \varphi(X) \left\{ Y - \hat{f}_0(X) \right\} \right] : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\},$
- $\left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(1)}(f) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) = \left\{ \hat{\mathbb{Q}}_X - \mathbb{Q}_X \right\} \left[\varphi(X) \left\{ \hat{f}_0(X) - f^*(X) \right\} \right] : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\},$
- $\left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(f) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) = \left\{ \hat{\mathbb{Q}}_X - \mathbb{Q}_X \right\} \left[\left\{ \varphi(X) \right\}^2 \right] : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}$

are empirical processes indexed by $\varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*)$, where $\hat{\mathbb{P}} \in \Delta(\mathbb{X} \times \mathbb{R})$ and $\hat{\mathbb{Q}}_X \in \Delta(\mathbb{X})$ are the empirical distributions for the $n_{\mathbb{P}}$ labeled source samples $\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}$ and $n_{\mathbb{Q}}$ unlabeled target samples $\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}}$, respectively, i.e., $\hat{\mathbb{P}} := \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \delta_{(X_i^{\mathbb{P}}, Y_i^{\mathbb{P}})}$ and $\hat{\mathbb{Q}}_X := \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \delta_{X_j^{\mathbb{Q}}}$.

Control of the supremum of $\{\mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*)\}$: Firstly, we are in need of a delicate control of the expectation of the supremum of the empirical process

$$\left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}.$$

This goal can be achieved through the following lemma, whose proof is provided in Appendix C.1.

Lemma 5 *The expectation of the supremum of the empirical process $\{\mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*)\}$ is upper bounded by*

$$\mathbb{E}_{\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \leq 4C_{\text{dr}} (1 + C_{\text{rf}}) \mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P},X}(\mathcal{F}^*). \tag{43}$$

We then move on to a tight control of the size of

$$\left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} - \mathbb{E}_{\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}} \left[\left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right]$$

under $\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}$. This task can be settled via the following lemma, whose proof is deferred to Appendix C.2.

Lemma 6 *If $\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}$, then with probability at least $1 - \delta$, we have*

$$\begin{aligned}
 & \sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\
 & - \mathbb{E}_{\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\
 & \leq \frac{6C_{\text{dr}}(1 + C_{\text{rf}})}{n_{\mathbb{P}}} \log \left(\frac{1}{\delta} \right) + 2C_{\text{dr}}(1 + C_{\text{rf}}) \sqrt{\frac{2 \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{P}}}} \\
 & \quad + 4C_{\text{dr}}(1 + C_{\text{rf}}) \sqrt{\frac{2\mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P}X}(\mathcal{F}^*) \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{P}}}}.
 \end{aligned} \tag{44}$$

To finish up, we first denote the right-hand side of the inequality (44) from Lemma 6 by

$$\begin{aligned}
 \mathcal{B}_{\mathbb{P}}(\delta) &:= \frac{6C_{\text{dr}}(1 + C_{\text{rf}})}{n_{\mathbb{P}}} \log \left(\frac{1}{\delta} \right) + 2C_{\text{dr}}(1 + C_{\text{rf}}) \sqrt{\frac{2 \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{P}}}} \\
 & \quad + 4C_{\text{dr}}(1 + C_{\text{rf}}) \sqrt{\frac{2\mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P}X}(\mathcal{F}^*) \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{P}}}}.
 \end{aligned} \tag{45}$$

for ease of exposition. Then, with probability at least $1 - \delta$, one has

$$\begin{aligned}
 & \sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\
 & = \sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\
 & \quad - \mathbb{E}_{\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\
 & \quad + \mathbb{E}_{\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\
 & \stackrel{(f)}{\leq} \mathcal{B}_{\mathbb{P}}(\delta) + 4C_{\text{dr}}(1 + C_{\text{rf}}) \mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P}X}(\mathcal{F}^*),
 \end{aligned} \tag{46}$$

where the step (f) invokes Lemmas 5 and 6. For simplicity, we define the following event: for any $\delta \in (0, 1)$,

$$\begin{aligned}
 \mathcal{E}_{\mathbb{P}}(\delta) &:= \left\{ \left(\mathbf{o}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{x}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) \in \mathbb{O}^{n_{\mathbb{P}}} \times \mathbb{X}^{n_{\mathbb{Q}}} : \right. \\
 & \left. \sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{o}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \leq \mathcal{B}_{\mathbb{P}}(\delta) + 4C_{\text{dr}}(1 + C_{\text{rf}}) \mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P}X}(\mathcal{F}^*) \right\}.
 \end{aligned} \tag{47}$$

Then, the upper bound (46) implies

$$\left(\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}} \right) \{ \mathcal{E}_{\mathbb{P}}(\delta) \} \geq 1 - \delta \quad \text{for every } \delta \in (0, 1). \tag{48}$$

Control of the supremum of $\left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}$: Similar to the above argument for controlling the supremum of the empirical process $\left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}$, we first establish an upper bound on the expectation of the supremum of the empirical process

$$\left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}.$$

We provide a desired result in the following lemma, whose proof is postponed to Appendix C.3.

Lemma 7 *The expectation of the supremum of the empirical process $\left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}$ can be upper bounded by*

$$\mathbb{E}_{\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_Q}} \left[\sup \left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \leq 4(1 + C_{\text{rf}}) \mathcal{R}_{n_Q}^{\mathbb{Q},X}(\mathcal{F}^*). \quad (49)$$

Analogously, we now aim at a tight control of the size of

$$\sup \left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} - \mathbb{E}_{\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_Q}} \left[\sup \left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right]$$

under the data generating process $\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_Q}$. This goal can be achieved through the following lemma, whose detailed proof is provided in Appendix C.4.

Lemma 8 *If $\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_Q}$, then with probability at least $1 - \delta$, it holds that*

$$\begin{aligned} & \sup \left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\ & - \mathbb{E}_{\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_Q}} \left[\sup \left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\ & \leq \frac{6(1 + C_{\text{rf}})}{n_Q} \log \left(\frac{1}{\delta} \right) + 2(1 + C_{\text{rf}}) \sqrt{\frac{2 \log \left(\frac{1}{\delta} \right)}{n_Q}} + 4(1 + C_{\text{rf}}) \sqrt{\frac{2 \mathcal{R}_{n_Q}^{\mathbb{Q},X}(\mathcal{F}^*) \log \left(\frac{1}{\delta} \right)}{n_Q}}. \end{aligned} \quad (50)$$

For simplicity, we denote the right-hand side of the inequality (50) from Lemma 8 for any given $\delta \in (0, 1)$ as

$$\begin{aligned} \mathcal{B}_Q^{(1)}(\delta) &:= \frac{6(1 + C_{\text{rf}})}{n_Q} \log \left(\frac{1}{\delta} \right) + 2(1 + C_{\text{rf}}) \sqrt{\frac{2 \log \left(\frac{1}{\delta} \right)}{n_Q}} \\ &+ 4(1 + C_{\text{rf}}) \sqrt{\frac{2 \mathcal{R}_{n_Q}^{\mathbb{Q},X}(\mathcal{F}^*) \log \left(\frac{1}{\delta} \right)}{n_Q}}. \end{aligned} \quad (51)$$

Then, it holds with probability greater than $1 - \delta$ that

$$\begin{aligned} & \sup \left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\ & = \sup \left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\ & - \mathbb{E}_{\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_Q}} \left[\sup \left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\ & + \mathbb{E}_{\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_Q}} \left[\sup \left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\ & \stackrel{(g)}{\leq} \mathcal{B}_Q^{(1)}(\delta) + 4(1 + C_{\text{rf}}) \mathcal{R}_{n_Q}^{\mathbb{Q},X}(\mathcal{F}^*), \end{aligned} \quad (52)$$

where the step (g) holds due to Lemma 7 and 8. For the sake of conciseness, we define the following event: for any $\delta \in (0, 1)$,

$$\begin{aligned} \mathcal{E}_Q^{(1)}(\delta) &:= \left\{ \left(\mathbf{o}_{1:n_P}^{\mathbb{P}}, \mathbf{x}_{1:n_Q}^{\mathbb{Q}} \right) \in \mathbb{O}^{n_P} \times \mathbb{X}^{n_Q} : \right. \\ & \left. \sup \left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{x}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \leq \mathcal{B}_Q^{(1)}(\delta) + 4(1 + C_{\text{rf}}) \mathcal{R}_{n_Q}^{\mathbb{Q},X}(\mathcal{F}^*). \right\} \end{aligned} \quad (53)$$

Then, the upper bound (52) directly yields

$$\left(\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}\right) \left\{ \mathcal{E}_{\mathbb{Q}}^{(1)}(\delta) \right\} \geq 1 - \delta \quad \text{for every } \delta \in (0, 1). \quad (54)$$

Control of the supremum of $\left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}$: Akin to the above delicate control of the supremum of empirical processes, we upper bound the expectation of the supremum of the empirical process

$$\left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}.$$

It can be developed via the following lemma, with the detailed proof postponed to Appendix C.5.

Lemma 9 *The expectation of the supremum of the empirical process $\left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}$ has an upper bound*

$$\mathbb{E}_{\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \leq 16 \cdot \mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q},X}(\mathcal{F}^*). \quad (55)$$

As the next step, we now turn to a tight control of the size of

$$\sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} - \mathbb{E}_{\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right]$$

with $\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}$. The following lemma takes a step forward towards this goal, whose proof is deferred to Appendix C.6.

Lemma 10 *If $\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}$, then with probability at least $1 - \delta$, it holds that*

$$\begin{aligned} & \sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\ & - \mathbb{E}_{\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\ & \leq \frac{12}{n_{\mathbb{Q}}} \log \left(\frac{1}{\delta} \right) + 4 \sqrt{\frac{2 \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{Q}}}} + 16 \sqrt{\frac{\mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q},X}(\mathcal{F}^*) \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{Q}}}}. \end{aligned} \quad (56)$$

For ease of expression, we denote the right-hand side of the bound (56) in Lemma 10 by

$$\mathcal{B}_{\mathbb{Q}}^{(2)}(\delta) := \frac{12}{n_{\mathbb{Q}}} \log \left(\frac{1}{\delta} \right) + 4 \sqrt{\frac{2 \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{Q}}}} + 16 \sqrt{\frac{\mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q},X}(\mathcal{F}^*) \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{Q}}}}. \quad (57)$$

Then, one has with probability at least $1 - \delta$ that

$$\begin{aligned} & \sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\ & = \sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi(\cdot) \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\ & - \mathbb{E}_{\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\ & + \mathbb{E}_{\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\ & \stackrel{(h)}{\leq} \mathcal{B}_{\mathbb{Q}}^{(2)}(\delta) + 16 \cdot \mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q},X}(\mathcal{F}^*), \end{aligned} \quad (58)$$

where the step (h) invokes Lemmas 9 and 10. For the sake of simplicity, let us define the following event: for any $\delta \in (0, 1)$,

$$\begin{aligned} \mathcal{E}_{\mathbb{Q}}^{(2)}(\delta) &:= \left\{ \left(\mathbf{o}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{x}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) \in \mathbb{O}^{n_{\mathbb{P}}} \times \mathbb{X}^{n_{\mathbb{Q}}} : \right. \\ \sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{x}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} &\leq \mathcal{B}_{\mathbb{Q}}^{(2)}(\delta) + 16 \cdot \mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q},X}(\mathcal{F}^*). \end{aligned} \quad (59)$$

Then, the upper bound (59) gives

$$\left(\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}} \right) \left\{ \mathcal{E}_{\mathbb{Q}}^{(2)}(\delta) \right\} \geq 1 - \delta \quad \text{for every } \delta \in (0, 1). \quad (60)$$

Finally, it is time to put all pieces together in order to bound the term (T2) from our main bound (38). To this end, we introduce the event $\mathcal{E}(\delta) := \mathcal{E}_{\mathbb{P}}\left(\frac{\delta}{3}\right) \cap \mathcal{E}_{\mathbb{Q}}^{(1)}\left(\frac{\delta}{3}\right) \cap \mathcal{E}_{\mathbb{Q}}^{(2)}\left(\frac{\delta}{3}\right)$ for every $\delta \in (0, 1)$. By virtue of the union bound, the inequalities (48), (54), and (60) implies

$$\left(\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}} \right) \left\{ \mathcal{E}(\delta) \right\} = 1 - \left(\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}} \right) \left\{ (\mathbb{O}^{n_{\mathbb{P}}} \times \mathbb{X}^{n_{\mathbb{Q}}}) \setminus \mathcal{E}(\delta) \right\} \geq 1 - \delta.$$

On the other hand, by utilizing the definitions (47), (53), and (59) of the events $\mathcal{E}_{\mathbb{P}}(\delta)$, $\mathcal{E}_{\mathbb{Q}}^{(1)}(\delta)$, and $\mathcal{E}_{\mathbb{Q}}^{(2)}(\delta)$, respectively, it follows on the event $\mathcal{E}(\delta)$ that

$$\begin{aligned} \text{(T2)} &\stackrel{(i)}{\leq} 2 \sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\ &\quad + 2 \sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\ &\quad + \sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\ &\leq 2\mathcal{B}_{\mathbb{P}}\left(\frac{\delta}{3}\right) + 2\mathcal{B}_{\mathbb{Q}}^{(1)}\left(\frac{\delta}{3}\right) + \mathcal{B}_{\mathbb{Q}}^{(2)}\left(\frac{\delta}{3}\right) \\ &\quad + 8C_{\text{dr}}(1 + C_{\text{rf}}) \mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P},X}(\mathcal{F}^*) + 8(3 + C_{\text{rf}}) \mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q},X}(\mathcal{F}^*), \end{aligned} \quad (61)$$

where the step (i) holds by virtue of the inequality (41). Taking two bounds (39) and (61) collectively leads to the following upper bound on the excess Q-risk of the DR estimator (5): on the event $\mathcal{E}(\delta)$, which holds with probability at least $1 - \delta$ under the probability measure $\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}$, we obtain

$$\begin{aligned} \mathcal{E}_{\mathbb{Q}}(\hat{f}_{\text{DR}}) &\leq \text{(T1)} + \text{(T2)} \\ &\leq 4 \|\hat{\rho} - \rho^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} \cdot \|\hat{f}_0 - f^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} + 2\mathcal{B}_{\mathbb{P}}\left(\frac{\delta}{3}\right) + 2\mathcal{B}_{\mathbb{Q}}^{(1)}\left(\frac{\delta}{3}\right) + \mathcal{B}_{\mathbb{Q}}^{(2)}\left(\frac{\delta}{3}\right) \\ &\quad + 8C_{\text{dr}}(1 + C_{\text{rf}}) \mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P},X}(\mathcal{F}^*) + 8(3 + C_{\text{rf}}) \mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q},X}(\mathcal{F}^*) \\ &\leq 4 \|\hat{\rho} - \rho^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} \cdot \|\hat{f}_0 - f^*\|_{L^2(\mathbb{X}, \mathbb{P}_X)} + 12(2 + C_{\text{rf}}) \log\left(\frac{3}{\delta}\right) \left(\frac{C_{\text{dr}}}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right) \\ &\quad + 4(1 + C_{\text{dr}})(2 + C_{\text{rf}}) \sqrt{2 \log\left(\frac{3}{\delta}\right)} \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \\ &\quad + 8(1 + C_{\text{dr}})(2 + C_{\text{rf}}) \sqrt{\log\left(\frac{3}{\delta}\right)} \left(\frac{\mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P},X}(\mathcal{F}^*)}{\sqrt{n_{\mathbb{P}}}} + \frac{\mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q},X}(\mathcal{F}^*)}{\sqrt{n_{\mathbb{Q}}}} \right) \\ &\quad + 8C_{\text{dr}}(1 + C_{\text{rf}}) \mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P},X}(\mathcal{F}^*) + 8(3 + C_{\text{rf}}) \mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q},X}(\mathcal{F}^*), \end{aligned}$$

and this completes the proof of Theorem 4.1.

B.3. Proof of Theorem 5.1

To begin with, we introduce some key universal constants to formally state our improved structure-agnostic guarantee of the DR estimator (20) for parameterized hypothesis classes $\mathcal{F} \subseteq (\mathbb{X} \rightarrow [-1, 1])$:

$$\begin{aligned}\mathcal{B}_1 &:= 4(1 + C_{\text{dr}})(1 + C_{\text{rf}})b_1, \\ \mathcal{B}_2 &:= 8\sqrt{2} \cdot \max \left\{ C_{\text{dr}}(1 + C_{\text{rf}})b_2, b_1^2 + (1 + C_{\text{rf}})^2 b_2 \right\}, \\ \mathcal{B}_3 &:= \max \{ 2(1 + C_{\text{dr}})(1 + C_{\text{rf}}), 4 \} \cdot b_3 + 6b_1b_2.\end{aligned}\tag{62}$$

With the above conventions, the detailed version of Theorem 5.1 can be stated as follows:

Theorem B.1 (Structure-agnostic upper bound II of the DR estimator) *With the parameterized function class (18) and assumptions 1–6, the DR estimator (20) satisfies the following: there exists an absolute constant $K \in (0, +\infty)$ such that, with probability at least $1 - 8\delta$ under the probability measure $\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}$,*

$$\begin{aligned}\mathcal{E}_{\mathbb{Q}}(\hat{\boldsymbol{\theta}}_{\text{DR}}) &= \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\left\{ f(X; \hat{\boldsymbol{\theta}}_{\text{DR}}) - f^*(X) \right\}^2 \right] \\ &\leq 18K^2(1 + C_{\text{dr}})^2(1 + C_{\text{rf}})^2 \log \left(\frac{d}{\delta} \right) \left[\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}} + \frac{d}{n_{\mathbb{Q}}} \right],\end{aligned}\tag{63}$$

provided that $\min \{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \geq \bar{\kappa} \cdot \mathcal{N}^* \log \left(\frac{d}{\delta} \right)$, where $\mathcal{N}^* := \max \{ \mathcal{N}_1, \mathcal{N}_2 \}$ and

$$\begin{aligned}\bar{\kappa} &:= \max \left\{ \kappa, \{ 2\mathcal{B}_3 K(1 + C_{\text{dr}})(1 + C_{\text{rf}}) \}^2, 18 \{ K(1 + C_{\text{dr}})(1 + C_{\text{rf}}) \}^2 \right\}, \\ \kappa &:= \max \left\{ (260\mathcal{B}_2)^2, \left\{ \frac{860\mathcal{B}_3}{3} K(1 + C_{\text{dr}})(1 + C_{\text{rf}}) \right\}^2, \left\{ \frac{160\mathcal{B}_1^2 \mathcal{B}_2}{(1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2} \right\}^{\frac{2}{3}}, \right. \\ &\quad \left. \left\{ \frac{640\mathcal{B}_1^3 \mathcal{B}_3}{3(1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2} \right\}^{\frac{1}{2}}, \frac{80\mathcal{B}_1^2}{(1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2} \right\}, \\ \mathcal{N}_1 &:= \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^2 \cdot \max \left\{ 1, \text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) + \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}, \right. \\ &\quad \left[\min \left\{ \text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}, \text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\} \right\} \right]^{-\frac{2}{3}}, \\ &\quad \left. \left[\min \left\{ \text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}, \text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\} \right\} \right]^{-\frac{1}{2}} \right\}, \\ \mathcal{N}_2 &:= \max \left\{ \left[\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) + \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{\min \left\{ \text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}, d \right\}} \right]^2, \frac{\left[\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) + \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\} \right]^3}{\left[\min \left\{ \text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}, d \right\} \right]^2}, \right. \\ &\quad \left. \left[\frac{\left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^2}{\min \left\{ \text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}, d \right\}} \right]^{\frac{2}{3}}, \left[\frac{\left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^3}{\min \left\{ \text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}, d \right\}} \right]^{\frac{1}{2}}, \end{aligned}\tag{64}$$

$$\frac{\left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}}{\min \left\{ \text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}, d \right\}} \right\}.$$

Towards proving Theorem B.1, we first present a key technical lemma that plays a crucial role in the proof. Roughly speaking, the lemma below captures the distance between the DR estimate $\hat{\boldsymbol{\theta}}_{\text{DR}} \in \Theta$ and the ground-truth parameter $\boldsymbol{\theta}^* \in \Theta$ under different metrics.

Lemma 11 *With Assumptions 1–6, any $\delta \in (0, \frac{1}{8}]$ and $(n_{\mathbb{P}}, n_{\mathbb{Q}}) \in \mathbb{N} \times \mathbb{N}$ such that $\min \{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \geq \kappa \cdot \max \{\mathcal{N}_1, \mathcal{N}_2\} \log \left(\frac{d}{\delta} \right)$, where κ , \mathcal{N}_1 , and \mathcal{N}_2 are specified as (64), the following facts hold with probability at least $1 - 8\delta$ under $\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}$: for some universal constant $K \in (0, +\infty)$,*

(i) *we have $\hat{\boldsymbol{\theta}}_{\text{DR}} \in \mathbb{B}_{r(\delta)}(\boldsymbol{\theta}^*)$, where the radius $r(\delta) \in (0, +\infty)$ is given by*

$$r(\delta) := 3K(1 + C_{\text{dr}})(1 + C_{\text{rf}}) \sqrt{\log \left(\frac{d}{\delta} \right)} \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]. \quad (65)$$

(ii) *it holds that*

$$\left\| \mathcal{I}_{\mathbb{Q}}^{\frac{1}{2}}(\boldsymbol{\theta}^*) \left(\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^* \right) \right\|_2^2 \leq 9K^2(1 + C_{\text{dr}})^2(1 + C_{\text{rf}})^2 \log \left(\frac{d}{\delta} \right) \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{d}{n_{\mathbb{Q}}}} \right]^2. \quad (66)$$

For simplicity, let $\Lambda(\delta) \subseteq \mathbb{O}^{n_{\mathbb{P}}} \times \mathbb{X}^{n_{\mathbb{Q}}}$ denote the event for which Lemma 11 holds, which immediately gives us $\left(\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}} \right) \{ \Lambda(\delta) \} \geq 1 - 8\delta$ for any $\delta \in (0, \frac{1}{8}]$.

We embark on the proof of Theorem B.1 by doing a Taylor expansion with respect to as follows:

$$\begin{aligned} \mathcal{E}_{\mathbb{Q}}(\hat{\boldsymbol{\theta}}_{\text{DR}}) &= \mathbb{E}_{(X,Y) \sim \mathbb{Q}} \left[\ell(Y, f(X; \hat{\boldsymbol{\theta}}_{\text{DR}})) - \ell(Y, f(X; \boldsymbol{\theta}^*)) \right] \\ &= \left\{ \mathbb{E}_{(X,Y) \sim \mathbb{Q}} [\nabla_{\boldsymbol{\theta}} \ell(Y, f(X; \boldsymbol{\theta}^*))] \right\}^{\top} (\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^*) \\ &\quad + \frac{1}{2!} (\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^*)^{\top} \mathbb{E}_{(X,Y) \sim \mathbb{Q}} [\nabla_{\boldsymbol{\theta}}^2 \ell(Y, f(X; \boldsymbol{\theta}^*))] (\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^*) \\ &\quad + \frac{1}{3!} \left\langle \mathbb{E}_{(X,Y) \sim \mathbb{Q}} [\nabla_{\boldsymbol{\theta}}^3 \ell(Y, f(X; \tilde{\boldsymbol{\theta}}))] , (\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^*)^{\otimes 3} \right\rangle_{\text{F}} \\ &\stackrel{(a)}{=} \frac{1}{2!} (\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^*)^{\top} \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) (\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^*) \\ &\quad + \frac{1}{3!} \left\langle \mathbb{E}_{(X,Y) \sim \mathbb{Q}} [\nabla_{\boldsymbol{\theta}}^3 \ell(Y, f(X; \tilde{\boldsymbol{\theta}}))] , (\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^*)^{\otimes 3} \right\rangle_{\text{F}} \end{aligned} \quad (67)$$

for some $\tilde{\boldsymbol{\theta}} \in \left\{ (1 - \lambda)\boldsymbol{\theta}^* + \lambda\hat{\boldsymbol{\theta}}_{\text{DR}} : \lambda \in [0, 1] \right\}$, where the step (a) holds due to the following facts:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\ell(y, f(x; \boldsymbol{\theta})) &= 2\{f(x; \boldsymbol{\theta}) - y\} \nabla_{\boldsymbol{\theta}}f(x; \boldsymbol{\theta}), \\ \nabla_{\boldsymbol{\theta}}^2\ell(y, f(x; \boldsymbol{\theta})) &= 2\nabla_{\boldsymbol{\theta}}f(x; \boldsymbol{\theta}) \{\nabla_{\boldsymbol{\theta}}f(x; \boldsymbol{\theta})\}^\top + 2\{f(x; \boldsymbol{\theta}) - y\} \nabla_{\boldsymbol{\theta}}^2f(x; \boldsymbol{\theta}), \\ \nabla_{\boldsymbol{\theta}}^3\ell(y, f(x; \boldsymbol{\theta})) &= 2\nabla_{\boldsymbol{\theta}}^2f(x; \boldsymbol{\theta}) \otimes \nabla_{\boldsymbol{\theta}}f(x; \boldsymbol{\theta}) + 4\nabla_{\boldsymbol{\theta}}f(x; \boldsymbol{\theta}) \otimes \nabla_{\boldsymbol{\theta}}^2f(x; \boldsymbol{\theta}) \\ &\quad + 2\{f(x; \boldsymbol{\theta}) - y\} \nabla_{\boldsymbol{\theta}}^3f(x; \boldsymbol{\theta})\end{aligned}\tag{68}$$

Thus, it follows from the equation (67) that

$$\begin{aligned}\mathcal{E}_{\mathbb{Q}}\left(\hat{\boldsymbol{\theta}}_{\text{DR}}\right) &\leq \frac{1}{2}\left(\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^*\right)^\top \mathcal{I}_{\mathbb{Q}}\left(\boldsymbol{\theta}^*\right)\left(\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^*\right) \\ &\quad + \frac{1}{6}\left\|\mathbb{E}_{(X,Y)\sim\mathbb{Q}}\left[\nabla_{\boldsymbol{\theta}}^3\ell\left(Y, f\left(X; \tilde{\boldsymbol{\theta}}\right)\right)\right]\right\|_{\text{op}}\left\|\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^*\right\|_2^3 \\ &\stackrel{\text{(b)}}{\leq} \frac{1}{2}\left(\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^*\right)^\top \mathcal{I}_{\mathbb{Q}}\left(\boldsymbol{\theta}^*\right)\left(\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^*\right) + \frac{\mathcal{B}_3}{6}\left\|\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^*\right\|_2^3,\end{aligned}\tag{69}$$

where the step (b) follows due to the observation that the operator norm $\|\cdot\|_{\text{op}} : (\mathbb{R}^d)^{\otimes 3} \rightarrow \mathbb{R}_+$ is a convex function together with Jensen's inequality and the following bound: for any $\boldsymbol{\theta} \in \Theta$,

$$\begin{aligned}\left\|\nabla_{\boldsymbol{\theta}}^3\ell(Y, f(X; \boldsymbol{\theta}))\right\|_{\text{op}} &\leq 2(1 + |Y|)\left\|\nabla_{\boldsymbol{\theta}}^3f(X; \boldsymbol{\theta})\right\|_{\text{op}} + 6\left\|\nabla_{\boldsymbol{\theta}}f(X; \boldsymbol{\theta})\right\|_2\left\|\nabla_{\boldsymbol{\theta}}^2f(X; \boldsymbol{\theta})\right\|_{\text{op}} \\ &\stackrel{\text{(c)}}{\leq} 4b_3 + 6b_1b_2 \leq \mathcal{B}_3\end{aligned}$$

\mathbb{Q} -almost surely, where the step (c) holds by Assumption 3 and the part (ii) of Assumption 5. Therefore, while being conditioned on the event $\Lambda(\delta)$, we reach

$$\begin{aligned}\mathcal{E}_{\mathbb{Q}}\left(\hat{\boldsymbol{\theta}}_{\text{DR}}\right) &\leq \frac{9}{2}K^2(1 + C_{\text{dr}})^2(1 + C_{\text{rf}})^2\log\left(\frac{d}{\delta}\right)\left[\sqrt{\frac{\text{Trace}\left\{\mathcal{I}_{\mathbb{P}}\left(\boldsymbol{\theta}^*\right)\mathcal{I}_{\mathbb{Q}}^{-1}\left(\boldsymbol{\theta}^*\right)\right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{d}{n_{\mathbb{Q}}}}\right]^2 \\ &\quad + \frac{3}{2}\mathcal{B}_3K^3(1 + C_{\text{dr}})^3(1 + C_{\text{rf}})^3\log^{\frac{3}{2}}\left(\frac{d}{\delta}\right) \\ &\quad \cdot \left[\sqrt{\frac{\text{Trace}\left\{\mathcal{I}_{\mathbb{P}}\left(\boldsymbol{\theta}^*\right)\mathcal{I}_{\mathbb{Q}}^{-2}\left(\boldsymbol{\theta}^*\right)\right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace}\left\{\mathcal{I}_{\mathbb{Q}}^{-1}\left(\boldsymbol{\theta}^*\right)\right\}}{n_{\mathbb{Q}}}}\right]^3.\end{aligned}\tag{70}$$

At this point, one can observe that if $\min \{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \geq \bar{\kappa} \cdot \max \{\mathcal{N}_1, \mathcal{N}_2\} \log \left(\frac{d}{\delta} \right)$, then

$$\begin{aligned}
& \frac{3}{2} \mathcal{B}_3 K^3 (1 + C_{\text{dr}})^3 (1 + C_{\text{rf}})^3 \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \\
& \cdot \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^3 \\
& \leq \frac{9}{2} K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log \left(\frac{d}{\delta} \right) \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{d}{n_{\mathbb{Q}}}} \right]^2.
\end{aligned} \tag{71}$$

Hence, by taking two pieces (70) and (71) collectively, it holds that if $\min \{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \geq \bar{\kappa} \cdot \max \{\mathcal{N}_1, \mathcal{N}_2\} \log \left(\frac{d}{\delta} \right)$, then we have on the event $\Lambda(\delta)$ that

$$\begin{aligned}
& \mathcal{E}_{\mathbb{Q}}(\hat{\boldsymbol{\theta}}_{\text{DR}}) \\
& \leq 9K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log \left(\frac{d}{\delta} \right) \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{d}{n_{\mathbb{Q}}}} \right]^2 \\
& \stackrel{(d)}{\leq} 18K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log \left(\frac{d}{\delta} \right) \left[\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}} + \frac{d}{n_{\mathbb{Q}}} \right],
\end{aligned} \tag{72}$$

where the step (d) invokes the Cauchy-Schwarz inequality. Since $\left(\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}} \right) \{ \Lambda(\delta) \} \geq 1 - 8\delta$, the bound (72) on the excess \mathbb{Q} -risk of the DR estimator (20) holds with probability exceeding $1 - 8\delta$ under the probability measure $\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}$, which completes the proof of Theorem B.1.

Appendix C. Proof of auxiliary lemmas for the proof of Theorem 4.1

C.1. Proof of Lemma 5

By following the standard symmetrization argument from empirical processes theory, we find that

$$\begin{aligned}
& \mathbb{E}_{\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\
& = \mathbb{E}_{\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}} \left[\sup \left\{ \left| \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) \right| : \varphi \in \mathcal{F}^* \right\} \right] \\
& \leq 2 \mathbb{E}_{\left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \boldsymbol{\sigma}_{1:n_{\mathbb{P}}} \right) \sim \mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \text{Unif}(\{\pm 1\}^{n_{\mathbb{P}}})} \left[\sup \left\{ \left| \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \sigma_i \hat{\rho} \left(X_i^{\mathbb{P}} \right) \varphi \left(X_i^{\mathbb{P}} \right) \left\{ Y_i^{\mathbb{P}} - \hat{f}_0 \left(X_i^{\mathbb{P}} \right) \right\} \right| : \varphi \in \mathcal{F}^* \right\} \right].
\end{aligned} \tag{73}$$

To control the last term in the inequality (73), we leverage the Ledoux-Talagrand contraction principle, which is formally stated in Lemma 1. To this end, define the functions $\phi_i^{\mathbb{P}} : \mathbb{O}^{n_{\mathbb{P}}} \rightarrow (\mathbb{R} \rightarrow \mathbb{R})$, $i \in [n_{\mathbb{P}}]$, by

$$\phi_i^{\mathbb{P}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) (t) := \hat{\rho} \left(x_i^{\mathbb{P}} \right) \left\{ y_i^{\mathbb{P}} - \hat{f}_0 \left(x_i^{\mathbb{P}} \right) \right\} (t), \quad \forall t \in \mathbb{R}, \tag{74}$$

and let $\phi_i^{\mathbb{P}} := \phi_i^{\mathbb{P}}(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}) : \mathbb{R} \rightarrow \mathbb{R}$ for $\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}$ for simplicity. Let $\mathcal{A}_{\mathbb{P}} := \bigcap_{i=1}^{n_{\mathbb{P}}} \{|Y_i^{\mathbb{P}}| \leq 1\}$, which holds $\mathbb{P}^{\otimes n_{\mathbb{P}}}$ -almost surely. Then, one can easily find that the function $\phi_i^{\mathbb{P}} : \mathbb{O}^{n_{\mathbb{P}}} \rightarrow (\mathbb{R} \rightarrow \mathbb{R})$ turns out to be a $C_{\text{dr}}(1 + C_{\text{rf}})$ -Lipschitz continuous function so that $\phi_i^{\mathbb{P}}(0) = 0$ for every $i \in [n_{\mathbb{P}}]$ on the event $\mathcal{A}_{\mathbb{P}}$. Then, while being conditioned on $\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}$, we obtain by virtue of Lemma 1 that

$$\begin{aligned}
 & \mathbb{E}_{\boldsymbol{\sigma}_{1:n_{\mathbb{P}}} \sim \text{Unif}(\{\pm 1\}^{n_{\mathbb{P}}})} \left[\sup \left\{ \left| \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \sigma_i \hat{\rho}(X_i^{\mathbb{P}}) \varphi(X_i^{\mathbb{P}}) \{Y_i^{\mathbb{P}} - \hat{f}_0(X_i^{\mathbb{P}})\} \right| : \varphi \in \mathcal{F}^* \right\} \cdot \mathbb{1}_{\mathcal{A}_{\mathbb{P}}} \right] \\
 &= \mathbb{E}_{\boldsymbol{\sigma}_{1:n_{\mathbb{P}}} \sim \text{Unif}(\{\pm 1\}^{n_{\mathbb{P}}})} \left[\sup \left\{ \left| \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \sigma_i \hat{\rho}(X_i^{\mathbb{P}}) \varphi(X_i^{\mathbb{P}}) \{Y_i^{\mathbb{P}} - \hat{f}_0(X_i^{\mathbb{P}})\} \right| : \varphi \in \mathcal{F}^* \right\} \right] \cdot \mathbb{1}_{\mathcal{A}_{\mathbb{P}}} \\
 &= \mathbb{E}_{\boldsymbol{\sigma}_{1:n_{\mathbb{P}}} \sim \text{Unif}(\{\pm 1\}^{n_{\mathbb{P}}})} \left[\sup \left\{ \left| \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \sigma_i \phi_i^{\mathbb{P}}(t_i) \right| : \mathbf{t}_{1:n_{\mathbb{P}}} \in \mathcal{T}_{\mathbb{P}}(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}) \right\} \right] \cdot \mathbb{1}_{\mathcal{A}_{\mathbb{P}}} \quad (75) \\
 &\leq 2C_{\text{dr}}(1 + C_{\text{rf}}) \mathbb{E}_{\boldsymbol{\sigma}_{1:n_{\mathbb{P}}} \sim \text{Unif}(\{\pm 1\}^{n_{\mathbb{P}}})} \left[\sup \left\{ \left| \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \sigma_i t_i \right| : \mathbf{t}_{1:n_{\mathbb{P}}} \in \mathcal{T}_{\mathbb{P}}(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}) \right\} \right] \cdot \mathbb{1}_{\mathcal{A}_{\mathbb{P}}},
 \end{aligned}$$

where $\mathcal{T}_{\mathbb{P}} : \mathbb{O}^{n_{\mathbb{P}}} \rightarrow \mathcal{P}(\mathbb{R}^{n_{\mathbb{P}}})$ is defined as

$$\mathcal{T}_{\mathbb{P}}(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}) := \left\{ \left(f(x_i^{\mathbb{P}}) - f^*(x_i^{\mathbb{P}}) : i \in [n_{\mathbb{P}}] \right) : f \in \mathcal{F} \right\} \subseteq \mathbb{R}^{n_{\mathbb{P}}}.$$

Here, $\mathcal{P}(\mathbb{R}^{n_{\mathbb{P}}})$ stands for the power set of the $n_{\mathbb{P}}$ -dimensional Euclidean space $\mathbb{R}^{n_{\mathbb{P}}}$. By taking the bound (75) collectively into the inequality (73), we arrive at

$$\begin{aligned}
 & \mathbb{E}_{\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi)(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\
 &\leq 2\mathbb{E}_{(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \boldsymbol{\sigma}_{1:n_{\mathbb{P}}}) \sim \mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \text{Unif}(\{\pm 1\}^{n_{\mathbb{P}}})} \left[\sup \left\{ \left| \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \sigma_i \hat{\rho}(X_i^{\mathbb{P}}) \varphi(X_i^{\mathbb{P}}) \{Y_i^{\mathbb{P}} - \hat{f}_0(X_i^{\mathbb{P}})\} \right| : \varphi \in \mathcal{F}^* \right\} \cdot \mathbb{1}_{\mathcal{A}_{\mathbb{P}}} \right] \\
 &\leq 4C_{\text{dr}}(1 + C_{\text{rf}}) \mathbb{E}_{\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}} \left[\mathbb{E}_{\boldsymbol{\sigma}_{1:n_{\mathbb{P}}} \sim \text{Unif}(\{\pm 1\}^{n_{\mathbb{P}}})} \left[\sup \left\{ \left| \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \sigma_i t_i \right| : \mathbf{t}_{1:n_{\mathbb{P}}} \in \mathcal{T}_{\mathbb{P}}(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}) \right\} \right] \cdot \mathbb{1}_{\mathcal{A}_{\mathbb{P}}} \right] \\
 &= 4C_{\text{dr}}(1 + C_{\text{rf}}) \mathbb{E}_{\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}} \left[\widehat{\mathcal{R}}_{n_{\mathbb{P}}}(\mathcal{F}^*)(\mathbf{X}_{1:n_{\mathbb{P}}}^{\mathbb{P}}) \cdot \mathbb{1}_{\mathcal{A}_{\mathbb{P}}} \right] \\
 &= 4C_{\text{dr}}(1 + C_{\text{rf}}) \mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P}X}(\mathcal{F}^*),
 \end{aligned}$$

as desired.

C.2. Proof of Lemma 6

We begin the proof with the introduction of the function class $\{\theta_{\mathbb{P}}(\varphi) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*)\}$, where $\theta_{\mathbb{P}}(\varphi) : \mathbb{X} \times [-1, 1] \rightarrow \mathbb{R}$ is defined to be

$$\theta_{\mathbb{P}}(\varphi)(x, y) := \hat{\rho}(x) \varphi(x) \left\{ y - \hat{f}_0(x) \right\}, \quad \forall (x, y) \in \mathbb{X} \times [-1, 1].$$

Recall that the event $\mathcal{A}_{\mathbb{P}} = \bigcap_{i=1}^{n_{\mathbb{P}}} \{|Y_i^{\mathbb{P}}| \leq 1\}$ holds $\mathbb{P}^{\otimes n_{\mathbb{P}}}$ -almost surely, i.e., $(\mathbb{P}^{\otimes n_{\mathbb{P}}})(\mathcal{A}_{\mathbb{P}}) = 1$. We thus obtain that

$$\begin{aligned}
& \sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\
&= \sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \cdot \mathbb{1}_{\mathcal{A}_{\mathbb{P}}} \\
&= \sup \left\{ \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \theta_{\mathbb{P}}(\varphi) \left(O_i^{\mathbb{P}} \right) - \mathbb{E}_{O \sim \mathbb{P}} [\theta_{\mathbb{P}}(\varphi)(O)] : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \cdot \mathbb{1}_{\mathcal{A}_{\mathbb{P}}} \quad (76) \\
&= \sup \left\{ \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \theta_{\mathbb{P}}(\varphi) \left(O_i^{\mathbb{P}} \right) - \mathbb{E}_{O \sim \mathbb{P}} [\theta_{\mathbb{P}}(\varphi)(O)] : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}
\end{aligned}$$

$\mathbb{P}^{\otimes n}$ -almost surely. In light of the equation (76), it can be easily seen based on the assumption (6) that

- (i) $|\theta_{\mathbb{P}}(\varphi)(x, y)| \leq 2C_{\text{dr}}(1 + C_{\text{rf}})$ for every $(x, y, \varphi) \in \mathbb{X} \times [-1, 1] \times \{\mathcal{F}^* \cup (-\mathcal{F}^*)\}$.
- (ii) $\text{Var}_{(X, Y) \sim \mathbb{P}} [\theta_{\mathbb{P}}(\varphi)(X, Y)] \leq \mathbb{E}_{(X, Y) \sim \mathbb{P}} [\{\theta_{\mathbb{P}}(\varphi)(X, Y)\}^2] \leq 4C_{\text{dr}}^2(1 + C_{\text{rf}})^2$ for every $\varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*)$.

By virtue of the classical Talagrand's concentration inequality (Lemma 2) with parameters $(B, v^2) = (2C_{\text{dr}}(1 + C_{\text{rf}}), 4C_{\text{dr}}^2(1 + C_{\text{rf}})^2)$, it holds with probability at least $1 - \delta$ that

$$\begin{aligned}
& \sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\
& - \mathbb{E}_{\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\
& \leq \frac{6C_{\text{dr}}(1 + C_{\text{rf}})}{n_{\mathbb{P}}} \log \left(\frac{1}{\delta} \right) + 2C_{\text{dr}}(1 + C_{\text{rf}}) \sqrt{\frac{2 \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{P}}}} \quad (77) \\
& \quad + \frac{2}{\sqrt{n_{\mathbb{P}}}} \sqrt{2C_{\text{dr}}(1 + C_{\text{rf}}) \mathbb{E}_{\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{P}}}^{\mathbb{P}}(\varphi) \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \log \left(\frac{1}{\delta} \right)} \\
& \stackrel{(a)}{\leq} \frac{6C_{\text{dr}}(1 + C_{\text{rf}})}{n_{\mathbb{P}}} \log \left(\frac{1}{\delta} \right) + 2C_{\text{dr}}(1 + C_{\text{rf}}) \sqrt{\frac{2 \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{P}}}} + 4C_{\text{dr}}(1 + C_{\text{rf}}) \sqrt{\frac{2\mathcal{R}_{n_{\mathbb{P}}}^{\mathbb{P}}(\mathcal{F}^*) \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{P}}}},
\end{aligned}$$

where the step (a) invokes Lemma 5. We thus complete the proof of Lemma 6.

C.3. Proof of Lemma 7

In light of the standard symmetrization argument from empirical processes theory, we reveal that

$$\begin{aligned}
& \mathbb{E}_{\mathbf{X}_{1:n_Q}^Q \sim \mathbb{Q}_X^{\otimes n_Q}} \left[\sup \left\{ \mathbb{G}_{n_Q}^{Q, (1)}(\varphi) \left(\mathbf{X}_{1:n_Q}^Q \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\
&= \mathbb{E}_{\mathbf{X}_{1:n_Q}^Q \sim \mathbb{Q}_X^{\otimes n_Q}} \left[\sup \left\{ \left| \mathbb{G}_{n_Q}^{Q, (1)}(\varphi) \left(\mathbf{X}_{1:n_Q}^Q \right) \right| : \varphi \in \mathcal{F}^* \right\} \right] \quad (78)
\end{aligned}$$

$$\leq 2\mathbb{E}_{(\mathbf{X}_{1:n_Q}^Q, \sigma_{1:n_Q}) \sim \mathbb{Q}_X^{\otimes n_Q} \otimes \text{Unif}(\{\pm 1\}^{n_Q})} \left[\sup \left\{ \left| \frac{1}{n_Q} \sum_{j=1}^{n_Q} \sigma_j \varphi(X_j^Q) \{ \hat{f}_0(X_j^Q) - f^*(X_j^Q) \} \right| : \varphi \in \mathcal{F}^* \right\} \right].$$

We now focus on a tight control of the last term in the bound (78) via the Ledoux-Talagrand contraction principle (Lemma 1). Towards this end, we consider the functions $\alpha_j^Q : \mathbb{X}^{n_Q} \rightarrow [-2, 2] \rightarrow \mathbb{R}$ for $j \in [n_Q]$ defined as

$$\alpha_j^Q(\mathbf{x}_{1:n_Q}^Q)(t) := \left\{ \hat{f}_0(x_j^Q) - f^*(x_j^Q) \right\} t, \quad \forall t \in [-2, 2], \quad (79)$$

and simplify $\alpha_j^Q := \alpha_j^Q(\mathbf{X}_{1:n_Q}^Q) : [-2, 2] \rightarrow \mathbb{R}$, where $\mathbf{X}_{1:n_Q}^Q \sim \mathbb{Q}_X^{n_Q}$. Then, $\alpha_j^Q : [-2, 2] \rightarrow \mathbb{R}$ is an $(1 + C_{\text{rf}})$ -Lipschitz continuous function with $\alpha_j^Q(0) = 0$ for $j \in [n_Q]$. By applying the Ledoux-Talagrand contraction principle (Lemma 1), while being conditioned on $\mathbf{X}_{1:n_Q}^Q$, we now have

$$\begin{aligned} & \mathbb{E}_{\sigma_{1:n_Q} \sim \text{Unif}(\{\pm 1\}^{n_Q})} \left[\sup \left\{ \left| \frac{1}{n_Q} \sum_{j=1}^{n_Q} \sigma_j \varphi(X_j^Q) \{ \hat{f}_0(X_j^Q) - f^*(X_j^Q) \} \right| : \varphi \in \mathcal{F}^* \right\} \right] \\ &= \mathbb{E}_{\sigma_{1:n_Q} \sim \text{Unif}(\{\pm 1\}^{n_Q})} \left[\sup \left\{ \left| \frac{1}{n_Q} \sum_{j=1}^{n_Q} \sigma_j \alpha_j^Q(t_j) \right| : \mathbf{t}_{1:n_Q} \in \mathcal{T}_Q(\mathbf{X}_{1:n_Q}^Q) \right\} \right] \\ &\leq 2(1 + C_{\text{rf}}) \mathbb{E}_{\sigma_{1:n_Q} \sim \text{Unif}(\{\pm 1\}^{n_Q})} \left[\sup \left\{ \left| \frac{1}{n_Q} \sum_{j=1}^{n_Q} \sigma_j t_j \right| : \mathbf{t}_{1:n_Q} \in \mathcal{T}_Q(\mathbf{X}_{1:n_Q}^Q) \right\} \right], \end{aligned} \quad (80)$$

where $\mathcal{T}_Q : \mathbb{X}^{n_Q} \rightarrow \mathcal{P}(\mathbb{R}^{n_Q})$ is defined as

$$\mathcal{T}_Q(\mathbf{x}_{1:n_Q}^Q) := \left\{ \left(f(x_j^Q) - f^*(x_j^Q) : j \in [n_Q] \right) : f \in \mathcal{F} \right\} \subseteq \mathbb{R}^{n_Q}. \quad (81)$$

Here, $\mathcal{P}(\mathbb{R}^{n_Q})$ refers to the power set of the n_Q -dimensional Euclidean space \mathbb{R}^{n_Q} . By taking two pieces (78) and (80) collectively, we reach

$$\begin{aligned} & \mathbb{E}_{\mathbf{X}_{1:n_Q}^Q \sim \mathbb{Q}_X^{\otimes n_Q}} \left[\sup \left\{ \mathbb{G}_{n_Q}^{Q,(1)}(\varphi)(\mathbf{X}_{1:n_Q}^Q) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\ &\leq 4(1 + C_{\text{rf}}) \mathbb{E}_{(\mathbf{X}_{1:n_Q}^Q, \sigma_{1:n_Q}) \sim \mathbb{Q}_X^{\otimes n_Q} \otimes \text{Unif}(\{\pm 1\}^{n_Q})} \left[\sup \left\{ \left| \frac{1}{n_Q} \sum_{j=1}^{n_Q} \sigma_j t_j \right| : \mathbf{t}_{1:n_Q} \in \mathcal{T}_Q(\mathbf{X}_{1:n_Q}^Q) \right\} \right] \\ &= 4(1 + C_{\text{rf}}) \mathbb{E}_{\mathbf{X}_{1:n_Q}^Q \sim \mathbb{Q}_X^{\otimes n_Q}} \left[\hat{\mathcal{R}}_{n_Q}(\mathcal{F}^*)(\mathbf{X}_{1:n_Q}^Q) \right] \\ &= 4(1 + C_{\text{rf}}) \mathcal{R}_{n_Q}^Q(\mathcal{F}^*), \end{aligned}$$

which thus completes the proof of Lemma 7.

C.4. Proof of Lemma 8

Similar to the proof of Lemma 6, let us first introduce the function class $\left\{ \theta_Q^{(1)}(\varphi) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}$, where $\theta_Q^{(1)}(\varphi) : \mathbb{X} \rightarrow \mathbb{R}$ is a function defined as

$$\theta_Q^{(1)}(\varphi)(x) := \varphi(x) \left\{ \hat{f}_0(x) - f^*(x) \right\}, \quad \forall x \in \mathbb{X}.$$

Then, it follows that

$$\begin{aligned} & \sup \left\{ \mathbb{G}_{n_{\mathbb{Q}},(1)}^{\mathbb{Q}}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\ &= \sup \left\{ \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \theta_{\mathbb{Q}}^{(1)}(\varphi) \left(X_j^{\mathbb{Q}} \right) - \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\theta_{\mathbb{Q}}^{(1)}(\varphi)(X) \right] : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}. \end{aligned} \quad (82)$$

At this point, one can easily reveal based on the assumption 3 that

- (i) $\left| \theta_{\mathbb{Q}}^{(1)}(\varphi)(x) \right| \leq 2(1 + C_{\text{rf}})$ for every $(x, \varphi) \in \mathbb{X} \times \{\mathcal{F}^* \cup (-\mathcal{F}^*)\}$.
- (ii) $\text{Var}_{X \sim \mathbb{Q}_X} \left[\theta_{\mathbb{Q}}^{(1)}(\varphi)(X) \right] \leq \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\left\{ \theta_{\mathbb{Q}}^{(1)}(\varphi)(X) \right\}^2 \right] \leq 4(1 + C_{\text{rf}})^2$ for all $f \in \mathcal{F} \cup (-\mathcal{F})$.

The classical Talagrand's concentration inequality (Lemma 2) with parameters $(B, v^2) = (2(1 + C_{\text{rf}}), 4(1 + C_{\text{rf}})^2)$ together with the equation (82) tells us that with probability at least $1 - \delta$,

$$\begin{aligned} & \sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\ & - \mathbb{E}_{\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\ & \leq \frac{6(1 + C_{\text{rf}})}{n_{\mathbb{Q}}} \log \left(\frac{1}{\delta} \right) + 2(1 + C_{\text{rf}}) \sqrt{\frac{2 \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{Q}}}} \\ & \quad + \frac{2}{\sqrt{n_{\mathbb{Q}}}} \sqrt{2(1 + C_{\text{rf}}) \mathbb{E}_{\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(1)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \log \left(\frac{1}{\delta} \right)} \\ & \stackrel{(a)}{\leq} \frac{6(1 + C_{\text{rf}})}{n_{\mathbb{Q}}} \log \left(\frac{1}{\delta} \right) + 2(1 + C_{\text{rf}}) \sqrt{\frac{2 \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{Q}}}} + 4(1 + C_{\text{rf}}) \sqrt{\frac{2 \mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q},X}(\mathcal{F}^*) \log \left(\frac{1}{\delta} \right)}{n_{\mathbb{Q}}}}, \end{aligned} \quad (83)$$

where the step (a) follows by Lemma 7, and this finishes the proof of Lemma 8.

C.5. Proof of Lemma 9

Similar to the proofs for Lemmas 5 and 7, we embark on the proof with the standard symmetrization argument from empirical processes theory, which yields

$$\begin{aligned} & \mathbb{E}_{\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\ &= \mathbb{E}_{\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}} \left[\sup \left\{ \left| \mathbb{G}_{n_{\mathbb{Q}}}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) \right| : \varphi \in \mathcal{F}^* \right\} \right] \\ &\leq 2 \mathbb{E}_{(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}}, \sigma_{1:n_{\mathbb{Q}}}) \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}} \otimes \text{Unif}(\{\pm 1\}^{n_{\mathbb{Q}}})} \left[\sup \left\{ \left| \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \sigma_j \left\{ \varphi \left(X_j^{\mathbb{Q}} \right) \right\}^2 \right| : \varphi \in \mathcal{F}^* \right\} \right]. \end{aligned} \quad (84)$$

We now consider the function $\beta^{\mathbb{Q}} : [-2, 2] \rightarrow \mathbb{R}$ defined to be

$$\beta^{\mathbb{Q}}(t) := t^2, \quad \forall t \in [-2, 2].$$

It turns out that $\beta^{\mathbb{Q}} : [-2, 2] \rightarrow \mathbb{R}$ is a 4-Lipschitz continuous function so that $\beta^{\mathbb{Q}}(0) = 0$. Then, the Ledoux-Talagrand contraction principle (Lemma 1) tells us that while being conditioned on $\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}}$,

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{\sigma}_{1:n_{\mathbb{Q}}} \sim \text{Unif}(\{\pm 1\}^{n_{\mathbb{Q}}})} \left[\sup \left\{ \left| \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \sigma_j \left\{ \varphi \left(X_j^{\mathbb{Q}} \right) \right\}^2 \right| : \varphi \in \mathcal{F}^* \right\} \right] \\ &= \mathbb{E}_{\boldsymbol{\sigma}_{1:n_{\mathbb{Q}}} \sim \text{Unif}(\{\pm 1\}^{n_{\mathbb{Q}}})} \left[\sup \left\{ \left| \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \sigma_j \beta^{\mathbb{Q}}(t_j) \right| : \mathbf{t}_{1:n_{\mathbb{Q}}} \in \mathcal{T}_{\mathbb{Q}}(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}}) \right\} \right] \\ &\leq 8 \mathbb{E}_{\boldsymbol{\sigma}_{1:n_{\mathbb{Q}}} \sim \text{Unif}(\{\pm 1\}^{n_{\mathbb{Q}}})} \left[\sup \left\{ \left| \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \sigma_j t_j \right| : \mathbf{t}_{1:n_{\mathbb{Q}}} \in \mathcal{T}_{\mathbb{Q}}(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}}) \right\} \right], \end{aligned} \quad (85)$$

where the function $\mathcal{T}_{\mathbb{Q}} : \mathbb{X}^{n_{\mathbb{Q}}} \rightarrow \mathcal{P}(\mathbb{R}^{n_{\mathbb{Q}}})$ is previously defined as (81). Putting two pieces (84) and (85) together, we arrive at

$$\begin{aligned} & \mathbb{E}_{\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}} \left[\sup \left\{ \mathbb{G}_{n_{\mathbb{Q}},(2)}^{\mathbb{Q}}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\ &\leq 16 \cdot \mathbb{E}_{(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}}, \boldsymbol{\sigma}_{1:n_{\mathbb{Q}}}) \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}} \otimes \text{Unif}(\{\pm 1\}^{n_{\mathbb{Q}}})} \left[\sup \left\{ \left| \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \sigma_j t_j \right| : \mathbf{t}_{1:n_{\mathbb{Q}}} \in \mathcal{T}_{\mathbb{Q}}(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}}) \right\} \right] \\ &= 16 \cdot \mathbb{E}_{\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}} \left[\widehat{\mathcal{R}}_{n_{\mathbb{Q}}}(\mathcal{F}^*) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) \right] \\ &= 16 \cdot \mathcal{R}_{n_{\mathbb{Q}}}^{\mathbb{Q}}(\mathcal{F}^*), \end{aligned}$$

as desired.

C.6. Proof of Lemma 10

We begin the proof by introducing the function class $\left\{ \theta_{\mathbb{Q}}^{(2)}(\varphi) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}$, where $\theta_{\mathbb{Q}}^{(2)}(\varphi) : \mathbb{X} \rightarrow \mathbb{R}$ is a function defined as

$$\theta_{\mathbb{Q}}^{(2)}(\varphi)(x) := \{\varphi(x)\}^2, \quad \forall x \in \mathbb{X}.$$

Then, it is obvious that

$$\begin{aligned} & \sup \left\{ \mathbb{G}_{n_{\mathbb{Q}},(2)}^{\mathbb{Q}}(\varphi) \left(\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\ &= \sup \left\{ \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \theta_{\mathbb{Q}}^{(2)}(\varphi) \left(X_j^{\mathbb{Q}} \right) - \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\theta_{\mathbb{Q}}^{(2)}(\varphi)(X) \right] : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\}. \end{aligned} \quad (86)$$

Also, one can easily find based on Assumption 3 that

- (i) $\left| \theta_{\mathbb{Q}}^{(2)}(\varphi)(x) \right| \leq 4$ for every $(x, \varphi) \in \mathbb{X} \times \{\mathcal{F}^* \cup (-\mathcal{F}^*)\}$.
- (ii) $\text{Var}_{X \sim \mathbb{Q}_X} \left[\theta_{\mathbb{Q}}^{(2)}(\varphi)(X) \right] \leq \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\left\{ \theta_{\mathbb{Q}}^{(2)}(\varphi)(X) \right\}^2 \right] \leq 16$ for every $f \in \mathcal{F} \cup (-\mathcal{F})$.

By virtue of the classical Talagrand's concentration inequality (Lemma 2) with parameters $(B, v^2) = (4, 16)$ together with the equation (86), we establish with probability at least $1 - \delta$ that

$$\begin{aligned}
& \sup \left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \\
& - \mathbb{E}_{\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_Q}} \left[\sup \left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \\
& \leq \frac{12}{n_Q} \log \left(\frac{1}{\delta} \right) + 4 \sqrt{\frac{2 \log \left(\frac{1}{\delta} \right)}{n_Q}} \\
& \quad + \frac{4}{\sqrt{n_Q}} \sqrt{\mathbb{E}_{\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \sim \mathbb{Q}_X^{\otimes n_Q}} \left[\sup \left\{ \mathbb{G}_{n_Q}^{\mathbb{Q},(2)}(\varphi) \left(\mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) : \varphi \in \mathcal{F}^* \cup (-\mathcal{F}^*) \right\} \right] \log \left(\frac{1}{\delta} \right)} \\
& \stackrel{(a)}{\leq} \frac{12}{n_Q} \log \left(\frac{1}{\delta} \right) + 4 \sqrt{\frac{2 \log \left(\frac{1}{\delta} \right)}{n_Q}} + 16 \sqrt{\frac{\mathcal{R}_{n_Q}^{\mathbb{Q},X}(\mathcal{F}^*) \log \left(\frac{1}{\delta} \right)}{n_Q}},
\end{aligned}$$

where the step (a) follows due to Lemma 9. This ends the proof of Lemma 10.

Appendix D. Proof of auxiliary lemmas for the proof of Theorem B.1

D.1. Proof of Lemma 11

Before delving into the proof of our main Lemma (Lemma 11) that plays a key role in the proof for Theorem B.1, we first establish key concentration properties of the gradient and the Hessian matrix of the DR empirical risk (19) that holds under Assumptions 1–5, whose proofs are postponed to the final part of this subsection.

Lemma 12 (Concentration property of the gradient) *Given any $\mathbf{A} \in \mathbb{R}^{d \times d}$, there exists a universal constant $\mathcal{C}(\mathbf{A}) \in (0, +\infty)$ that satisfies the following concentration property of the gradient of the DR empirical risk (19) with respect to the parameter vector $\boldsymbol{\theta}$: for any $\delta \in (0, 1]$, it holds that*

$$\begin{aligned}
& \left\| \mathbf{A} \left\{ \nabla_{\boldsymbol{\theta}} \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_P}^{\mathbb{P}}, \mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) (\boldsymbol{\theta}^*) - \mathbb{E}_{(\mathbf{O}_{1:n_P}^{\mathbb{P}}, \mathbf{X}_{1:n_Q}^{\mathbb{Q}}) \sim \mathbb{P}^{\otimes n_P} \otimes \mathbb{Q}_X^{\otimes n_Q}} \left[\nabla_{\boldsymbol{\theta}} \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_P}^{\mathbb{P}}, \mathbf{X}_{1:n_Q}^{\mathbb{Q}} \right) (\boldsymbol{\theta}^*) \right] \right\} \right\|_2 \\
& \leq \mathcal{C}(\mathbf{A}) \left\{ \sqrt{\frac{\mathcal{V}_{\mathbb{P}}(\mathbf{A}) \log \left(\frac{2d}{\delta} \right)}{n_P}} + \sqrt{\frac{\mathcal{V}_{\mathbb{Q}}(\mathbf{A}) \log \left(\frac{2d}{\delta} \right)}{n_Q}} \right. \\
& \quad \left. + \underbrace{4(1 + C_{\text{dr}})(1 + C_{\text{rf}})b_1}_{= \mathcal{B}_1 \text{ (defined in (62))}} \|\mathbf{A}\|_{\text{op}} \cdot \log \left(\frac{2d}{\delta} \right) \left(\frac{1}{n_P} + \frac{1}{n_Q} \right) \right\}
\end{aligned} \tag{87}$$

with probability exceeding $1 - \delta$ under the data generating process $(\mathbf{O}_{1:n_P}^{\mathbb{P}}, \mathbf{X}_{1:n_Q}^{\mathbb{Q}}) \sim \mathbb{P}^{\otimes n_P} \otimes \mathbb{Q}_X^{\otimes n_Q}$, where the functions $\Phi_{\mathbb{P}} : \mathbb{X} \times \mathbb{R} \rightarrow \mathbb{R}^d$ and $\Phi_{\mathbb{Q}} : \mathbb{X} \rightarrow \mathbb{R}^d$ are defined as

$$\begin{aligned}
\Phi_{\mathbb{P}}(x, y) &:= 2\hat{\rho}(x) \left\{ \hat{f}_0(x) - y \right\} \nabla_{\boldsymbol{\theta}} f(x; \boldsymbol{\theta}^*) \quad \text{and} \\
\Phi_{\mathbb{Q}}(x) &:= 2 \left\{ f(x; \boldsymbol{\theta}^*) - \hat{f}_0(x) \right\} \nabla_{\boldsymbol{\theta}} f(x; \boldsymbol{\theta}^*),
\end{aligned} \tag{88}$$

respectively, and the quantities $\mathcal{V}_{\mathbb{P}}(\mathbf{A}) \in (0, +\infty)$ and $\mathcal{V}_{\mathbb{Q}}(\mathbf{A}) \in (0, +\infty)$ are defined by

$$\begin{aligned}\mathcal{V}_{\mathbb{P}}(\mathbf{A}) &:= \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\left\| \mathbf{A} \left\{ \Phi_{\mathbb{P}}(X, Y) - \mathbb{E}_{(X,Y) \sim \mathbb{P}} [\Phi_{\mathbb{P}}(X, Y)] \right\} \right\|_2^2 \right] \quad \text{and} \\ \mathcal{V}_{\mathbb{Q}}(\mathbf{A}) &:= \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\left\| \mathbf{A} \left\{ \Phi_{\mathbb{Q}}(X) - \mathbb{E}_{X \sim \mathbb{Q}_X} [\Phi_{\mathbb{Q}}(X)] \right\} \right\|_2^2 \right],\end{aligned}\tag{89}$$

respectively.

Lemma 13 (Concentration property of the Hessian) *The Hessian of the DR empirical risk (19) with respect to the parameter vector $\boldsymbol{\theta}$ has the following concentration property: for any $\delta \in (0, 1]$, it holds that*

$$\begin{aligned}& \left\| \nabla_{\boldsymbol{\theta}}^2 \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}^*) - \mathbb{E}_{(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}}) \sim \mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{n_{\mathbb{Q}}}} \left[\nabla_{\boldsymbol{\theta}}^2 \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}^*) \right] \right\|_{\text{op}} \\ & \leq \underbrace{8\sqrt{2} \cdot \max \{ C_{\text{dr}} (1 + C_{\text{rf}}) b_2, b_1^2 + (1 + C_{\text{rf}}) b_2 \}}_{= B_2 \text{ (defined in (62))}} \left\{ \sqrt{\frac{\log \left(\frac{4d}{\delta} \right)}{n_{\mathbb{P}}}} + \sqrt{\frac{\log \left(\frac{4d}{\delta} \right)}{n_{\mathbb{Q}}}} \right\}\end{aligned}\tag{90}$$

with probability at least $1 - \delta$ under the data generating process $(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}}) \sim \mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{n_{\mathbb{Q}}}$.

Lastly, it is straightforward to see that

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}) &= \frac{2}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \hat{\rho} \left(x_i^{\mathbb{P}} \right) \left\{ \hat{f}_0 \left(x_i^{\mathbb{P}} \right) - y_i^{\mathbb{P}} \right\} \nabla_{\boldsymbol{\theta}} f \left(x_i^{\mathbb{P}}; \boldsymbol{\theta} \right) \\ &\quad + \frac{2}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \left\{ f \left(x_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) - \hat{f}_0 \left(x_j^{\mathbb{Q}} \right) \right\} \nabla_{\boldsymbol{\theta}} f \left(x_j^{\mathbb{Q}}; \boldsymbol{\theta} \right), \\ \nabla_{\boldsymbol{\theta}}^2 \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}) &= \frac{2}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \hat{\rho} \left(x_i^{\mathbb{P}} \right) \left\{ \hat{f}_0 \left(x_i^{\mathbb{P}} \right) - y_i^{\mathbb{P}} \right\} \nabla_{\boldsymbol{\theta}}^2 f \left(x_i^{\mathbb{P}}; \boldsymbol{\theta} \right) \\ &\quad + \frac{2}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \left[\nabla_{\boldsymbol{\theta}} f \left(x_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) \left\{ \nabla_{\boldsymbol{\theta}} f \left(x_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) \right\}^{\top} + \left\{ f \left(x_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) - \hat{f}_0 \left(x_j^{\mathbb{Q}} \right) \right\} \nabla_{\boldsymbol{\theta}}^2 f \left(x_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) \right] \\ \nabla_{\boldsymbol{\theta}}^3 \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}) &= \frac{2}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \hat{\rho} \left(x_i^{\mathbb{P}} \right) \left\{ \hat{f}_0 \left(x_i^{\mathbb{P}} \right) - y_i^{\mathbb{P}} \right\} \nabla_{\boldsymbol{\theta}}^3 f \left(x_i^{\mathbb{P}}; \boldsymbol{\theta} \right) \\ &\quad + \frac{2}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \left[\nabla_{\boldsymbol{\theta}}^2 f \left(x_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) \otimes \nabla_{\boldsymbol{\theta}} f \left(x_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) + \nabla_{\boldsymbol{\theta}}^2 f \left(x_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) \otimes \nabla_{\boldsymbol{\theta}} f \left(x_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) \right. \\ &\quad \left. + \left\{ f \left(x_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) - \hat{f}_0 \left(x_j^{\mathbb{Q}} \right) \right\} \nabla_{\boldsymbol{\theta}}^3 f \left(x_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) \right].\end{aligned}\tag{91}$$

By making use of the observation (91), it follows that

$$\begin{aligned}& \mathbb{E} \left[\nabla_{\boldsymbol{\theta}} \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}^*) \right] \\ &= 2 \mathbb{E}_{X \sim \mathbb{P}_X} \left[\left\{ \hat{\rho}(X) - \rho^*(X) \right\} \left\{ \hat{f}_0(X) - f(X; \boldsymbol{\theta}^*) \right\} \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) \right], \\ & \mathbb{E} \left[\nabla_{\boldsymbol{\theta}}^2 \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}^*) \right] \\ &= 2 \mathbb{E}_{X \sim \mathbb{P}_X} \left[\left\{ \hat{\rho}(X) - \rho^*(X) \right\} \left\{ \hat{f}_0(X) - f(X; \boldsymbol{\theta}^*) \right\} \nabla_{\boldsymbol{\theta}}^2 f(X; \boldsymbol{\theta}^*) \right] + \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*),\end{aligned}\tag{92}$$

and while being conditioned on the event $\mathcal{A}_{\mathbb{P}} := \bigcap_{i=1}^{n_{\mathbb{P}}} \{|Y_i^{\mathbb{P}}| \leq 1\}$ that

$$\begin{aligned} \left\| \nabla_{\boldsymbol{\theta}}^3 \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}) \right\|_{\text{op}} &\leq \frac{2}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \|\hat{\rho}\|_{\infty} \left(1 + \|\hat{f}_0\|_{\infty} \right) \left\| \nabla_{\boldsymbol{\theta}}^2 f \left(X_i^{\mathbb{P}}; \boldsymbol{\theta} \right) \right\|_{\text{op}} \\ &\quad + \frac{2}{n_{\mathbb{Q}}} \sum_{i=1}^{n_{\mathbb{Q}}} \left\{ 3 \left\| \nabla_{\boldsymbol{\theta}} f \left(X_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) \right\|_2 \left\| \nabla_{\boldsymbol{\theta}}^2 f \left(X_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) \right\|_{\text{op}} \right. \\ &\quad \left. + \left(1 + \|\hat{f}_0\|_{\infty} \right) \left\| \nabla_{\boldsymbol{\theta}}^3 f \left(X_j^{\mathbb{Q}}; \boldsymbol{\theta} \right) \right\|_{\text{op}} \right\} \\ &\stackrel{(a)}{\leq} 2(1 + C_{\text{dr}})(1 + C_{\text{rf}})b_3 + 6b_1b_2 \leq \mathcal{B}_3 \end{aligned} \quad (93)$$

for every $\boldsymbol{\theta} \in \Theta$, where the step (a) comes from Assumption 4 and the part (ii) of Assumption 5.

With these key preliminary results in our hand, we are now ready to prove Lemma 11. Hereafter, we focus on the case where $\Theta = \mathbb{R}^d$ for simplicity of presentation. For any $\delta \in (0, 1)$ and any fixed matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, we define the events

$$\begin{aligned} \mathcal{E}_1(\delta; \mathbf{A}) &:= \left\{ \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) \in \mathbb{O}^{n_{\mathbb{P}}} \times \mathbb{X}^{n_{\mathbb{Q}}} : \right. \\ &\quad \left\| \mathbf{A} \left\{ \nabla_{\boldsymbol{\theta}} \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}^*) - \mathbb{E} \left[\nabla_{\boldsymbol{\theta}} \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}^*) \right] \right\} \right\|_2 \\ &\leq \mathcal{C}(\mathbf{A}) \left\{ \sqrt{\frac{\mathcal{V}_{\mathbb{P}}(\mathbf{A}) \log \left(\frac{d}{\delta} \right)}{n_{\mathbb{P}}}} + \sqrt{\frac{\mathcal{V}_{\mathbb{Q}}(\mathbf{A}) \log \left(\frac{d}{\delta} \right)}{n_{\mathbb{Q}}}} + \mathcal{B}_1 \|\mathbf{A}\|_{\text{op}} \cdot \log \left(\frac{d}{\delta} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right) \right\} \Bigg\}, \end{aligned} \quad (94)$$

and

$$\begin{aligned} \mathcal{E}_2(\delta) &:= \left\{ \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) \in \mathbb{O}^{n_{\mathbb{P}}} \times \mathbb{X}^{n_{\mathbb{Q}}} : \right. \\ &\quad \left\| \nabla_{\boldsymbol{\theta}}^2 \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}^*) - \mathbb{E} \left[\nabla_{\boldsymbol{\theta}}^2 \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}^*) \right] \right\|_{\text{op}} \\ &\leq \mathcal{B}_2 \left\{ \sqrt{\frac{\log \left(\frac{d}{\delta} \right)}{n_{\mathbb{P}}}} + \sqrt{\frac{\log \left(\frac{d}{\delta} \right)}{n_{\mathbb{Q}}}} \right\} \Bigg\}, \end{aligned} \quad (95)$$

so that $\left(\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}} \right) \{ \mathcal{E}_1(\delta; \mathbf{A}) \} \geq 1 - 2\delta$ for any $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\left(\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}} \right) \{ \mathcal{E}_2(\delta) \} \geq 1 - 4\delta$ for any given $\delta \in (0, \frac{1}{4}]$ due to Lemma 12 and 13. For simplicity, we employ the notation $\widehat{\mathcal{R}}_{\text{DR}} := \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) : \mathbb{R}^d \rightarrow \mathbb{R}$ for $\left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) \sim \mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}$ as well as $\mathbf{g} := \nabla_{\boldsymbol{\theta}} \widehat{\mathcal{R}}_{\text{DR}} (\boldsymbol{\theta}^*) - \mathbb{E} \left[\nabla_{\boldsymbol{\theta}} \widehat{\mathcal{R}}_{\text{DR}} (\boldsymbol{\theta}^*) \right]$ throughout this subsection. Owing to Assumption 5, it turns out for every $\boldsymbol{\theta} \in \mathbb{R}^d$ that while being conditioned on the event $\mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_2(\delta)$, where $\mathcal{A}_{\mathbb{P}} = \bigcap_{i=1}^{n_{\mathbb{P}}} \{|Y_i^{\mathbb{P}}| \leq 1\}$,

$$\begin{aligned} &\widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^*) \\ &\stackrel{(a)}{\leq} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\top} \nabla_{\boldsymbol{\theta}} \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^*) + \frac{1}{2!} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\top} \nabla_{\boldsymbol{\theta}}^2 \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^*) (\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{\mathcal{B}_3}{3!} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^3 \\ &\stackrel{(b)}{\leq} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\top} \mathbb{E} \left[\nabla_{\boldsymbol{\theta}} \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^*) \right] + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\top} \mathbf{g} + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\top} \mathbb{E} \left[\nabla_{\boldsymbol{\theta}}^2 \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^*) \right] (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \end{aligned}$$

$$\begin{aligned}
 & + \frac{\mathcal{B}_2}{2} \sqrt{\log \left(\frac{d}{\delta} \right)} \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 + \frac{\mathcal{B}_3}{6} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^3 \\
 & \stackrel{(c)}{=} 2\mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left\{ \hat{f}_0(X) - f(X; \boldsymbol{\theta}^*) \right\} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) \right] + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbf{g} \\
 & + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left\{ \hat{f}_0(X) - f(X; \boldsymbol{\theta}^*) \right\} \nabla_{\boldsymbol{\theta}}^2 f(X; \boldsymbol{\theta}^*) \right] (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\
 & + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) (\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{\mathcal{B}_2}{2} \sqrt{\log \left(\frac{d}{\delta} \right)} \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 + \frac{\mathcal{B}_3}{6} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^3,
 \end{aligned} \tag{96}$$

where the step (a) follows due to Taylor's theorem together with the fact (93), the step (b) invokes Lemma 13, and the step (c) holds by the observation (92). Thus, by letting $\Delta(\boldsymbol{\theta}) := \boldsymbol{\theta} - \boldsymbol{\theta}^* \in \mathbb{R}^d$, we obtain from the inequality (96) that on the event $\mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_2(\delta)$,

$$\begin{aligned}
 & \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^*) \\
 & \leq 2\mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left\{ \hat{f}_0(X) - f(X; \boldsymbol{\theta}^*) \right\} \{\Delta(\boldsymbol{\theta})\}^\top \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) \right] + \{\Delta(\boldsymbol{\theta})\}^\top \mathbf{g} \\
 & + \mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left\{ \hat{f}_0(X) - f(X; \boldsymbol{\theta}^*) \right\} \{\Delta(\boldsymbol{\theta})\}^\top \nabla_{\boldsymbol{\theta}}^2 f(X; \boldsymbol{\theta}^*) \Delta(\boldsymbol{\theta}) \right] \\
 & + \frac{1}{2} \{\Delta(\boldsymbol{\theta})\}^\top \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) \Delta(\boldsymbol{\theta}) + \frac{\mathcal{B}_2}{2} \sqrt{\log \left(\frac{d}{\delta} \right)} \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \|\Delta(\boldsymbol{\theta})\|_2^2 + \frac{\mathcal{B}_3}{6} \|\Delta(\boldsymbol{\theta})\|_2^3 \\
 & = \mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left\{ \hat{f}_0(X) - f(X; \boldsymbol{\theta}^*) \right\} \left[2\{\Delta(\boldsymbol{\theta})\}^\top \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) + \{\Delta(\boldsymbol{\theta})\}^\top \nabla_{\boldsymbol{\theta}}^2 f(X; \boldsymbol{\theta}^*) \Delta(\boldsymbol{\theta}) \right] \right] \\
 & + \frac{1}{2} \{\Delta(\boldsymbol{\theta}) - \mathbf{z}\}^\top \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) \{\Delta(\boldsymbol{\theta}) - \mathbf{z}\} - \frac{1}{2} \mathbf{z}^\top \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) \mathbf{z} \\
 & + \frac{\mathcal{B}_2}{2} \sqrt{\log \left(\frac{d}{\delta} \right)} \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \|\Delta(\boldsymbol{\theta})\|_2^2 + \frac{\mathcal{B}_3}{6} \|\Delta(\boldsymbol{\theta})\|_2^3,
 \end{aligned} \tag{97}$$

where $\mathbf{z} := -\mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \mathbf{g} \in \mathbb{R}^d$. Employing a similar argument, one can reveal for any $\boldsymbol{\theta} \in \mathbb{R}^d$ that while being conditioned on the event $\mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_2(\delta)$, we have

$$\begin{aligned}
 & \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^*) \\
 & \stackrel{(d)}{\geq} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla_{\boldsymbol{\theta}} \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^*) + \frac{1}{2!} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla_{\boldsymbol{\theta}}^2 \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^*) (\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \frac{\mathcal{B}_3}{3!} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^3 \\
 & \stackrel{(e)}{\geq} 2\mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left\{ \hat{f}_0(X) - f(X; \boldsymbol{\theta}^*) \right\} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) \right] + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbf{g} \\
 & + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left\{ \hat{f}_0(X) - f(X; \boldsymbol{\theta}^*) \right\} \nabla_{\boldsymbol{\theta}}^2 f(X; \boldsymbol{\theta}^*) \right] (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\
 & + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) (\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \frac{\mathcal{B}_2}{2} \sqrt{\log \left(\frac{d}{\delta} \right)} \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 - \frac{\mathcal{B}_3}{6} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^3 \\
 & = \mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left\{ \hat{f}_0(X) - f(X; \boldsymbol{\theta}^*) \right\} \left[2\{\Delta(\boldsymbol{\theta})\}^\top \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) + \{\Delta(\boldsymbol{\theta})\}^\top \nabla_{\boldsymbol{\theta}}^2 f(X; \boldsymbol{\theta}^*) \Delta(\boldsymbol{\theta}) \right] \right] \\
 & + \frac{1}{2} \{\Delta(\boldsymbol{\theta}) - \mathbf{z}\}^\top \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) \{\Delta(\boldsymbol{\theta}) - \mathbf{z}\} - \frac{1}{2} \mathbf{z}^\top \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) \mathbf{z} \\
 & - \frac{\mathcal{B}_2}{2} \sqrt{\log \left(\frac{d}{\delta} \right)} \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \|\Delta(\boldsymbol{\theta})\|_2^2 - \frac{\mathcal{B}_3}{6} \|\Delta(\boldsymbol{\theta})\|_2^3,
 \end{aligned} \tag{98}$$

where the step (d) invokes Taylor's theorem as well as the observation (93), and the step (e) follows due to Lemma 13 and the fact (92).

Now, we use Lemma 12 with $\mathbf{A} = \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*)$. While being conditioned on the event $\mathcal{E}_1(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*))$, we have

$$\begin{aligned}
\|\mathbf{z}\|_2 &= \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \mathbf{g} \right\|_2 \\
&\leq \mathcal{C} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\} \left[\sqrt{\frac{\mathcal{V}_{\mathbb{P}} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\} \log\left(\frac{d}{\delta}\right)}{n_{\mathbb{P}}}} + \sqrt{\frac{\mathcal{V}_{\mathbb{Q}} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\} \log\left(\frac{d}{\delta}\right)}{n_{\mathbb{Q}}}} \right. \\
&\quad \left. + \mathcal{B}_1 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}} \log\left(\frac{d}{\delta}\right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right) \right] \\
&\stackrel{(f)}{\leq} \mathcal{C} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\} \left[\sqrt{2} (1 + C_{\text{dr}}) (1 + C_{\text{rf}}) \sqrt{\log\left(\frac{d}{\delta}\right)} \right. \\
&\quad \cdot \left\{ \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right\} \\
&\quad \left. + \mathcal{B}_1 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}} \log\left(\frac{d}{\delta}\right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right) \right], \tag{99}
\end{aligned}$$

where the step (f) can be obtained by letting $\mathbf{A} = \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*)$ in the following facts: given any fixed matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, it holds that

$$\begin{aligned}
\mathcal{V}_{\mathbb{P}}(\mathbf{A}) &:= \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\left\| \mathbf{A} \left\{ \Phi_{\mathbb{P}}(X, Y) - \mathbb{E}_{(X,Y) \sim \mathbb{P}} [\Phi_{\mathbb{P}}(X, Y)] \right\} \right\|_2^2 \right] \\
&= \text{Trace} \left\{ \mathbf{A} \text{Cov}_{(X,Y) \sim \mathbb{P}} [\Phi_{\mathbb{P}}(X, Y)] \mathbf{A}^{\top} \right\} \\
&\leq \text{Trace} \left\{ \mathbf{A} \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\Phi_{\mathbb{P}}(X, Y) \left\{ \Phi_{\mathbb{P}}(X, Y) \right\}^{\top} \right] \mathbf{A}^{\top} \right\} \\
&= 4 \cdot \text{Trace} \left\{ \mathbf{A} \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\hat{\rho}^2(X) \left\{ \hat{f}_0(X) - Y \right\}^2 \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) \left\{ \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) \right\}^{\top} \right] \mathbf{A}^{\top} \right\} \tag{100} \\
&\stackrel{(g)}{\leq} 2C_{\text{dr}}^2 (1 + C_{\text{rf}})^2 \text{Trace} \left\{ \underbrace{\mathbf{A} \mathbb{E}_{X \sim \mathbb{P}_X} \left[2 \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) \left\{ \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) \right\}^{\top} \right]}_{= \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*)} \mathbf{A}^{\top} \right\} \\
&= 2C_{\text{dr}}^2 (1 + C_{\text{rf}})^2 \text{Trace} \left\{ \mathbf{A} \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathbf{A}^{\top} \right\},
\end{aligned}$$

and

$$\begin{aligned}
 \mathcal{V}_{\mathbb{Q}}(\mathbf{A}) &:= \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\|\mathbf{A} \{\Phi_{\mathbb{Q}}(X) - \mathbb{E}_{X \sim \mathbb{Q}_X} [\Phi_{\mathbb{Q}}(X)]\}\|_2^2 \right] \\
 &= \text{Trace} \left\{ \mathbf{A} \text{Cov}_{X \sim \mathbb{Q}_X} [\Phi_{\mathbb{Q}}(X)] \mathbf{A}^\top \right\} \\
 &\leq \text{Trace} \left\{ \mathbf{A} \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\Phi_{\mathbb{Q}}(X) \{\Phi_{\mathbb{Q}}(X)\}^\top \right] \mathbf{A}^\top \right\} \\
 &= 4 \cdot \text{Trace} \left\{ \mathbf{A} \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\left\{ f(X; \boldsymbol{\theta}^*) - \hat{f}_0(X) \right\}^2 \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) \{\nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*)\}^\top \right] \mathbf{A}^\top \right\} \quad (101) \\
 &\stackrel{(h)}{\leq} 2(1 + C_{\text{rf}})^2 \text{Trace} \left\{ \underbrace{\mathbf{A} \mathbb{E}_{X \sim \mathbb{Q}_X} \left[2 \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) \{\nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*)\}^\top \right]}_{= \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*)} \mathbf{A}^\top \right\} \\
 &= 2(1 + C_{\text{rf}})^2 \text{Trace} \left\{ \mathbf{A} \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) \mathbf{A}^\top \right\},
 \end{aligned}$$

where the steps (g) and (h) follow by Assumption 4 on the black-box ML estimates $\hat{\rho} : \mathbb{X} \rightarrow \mathbb{R}_+$ and $\hat{f}_0 : \mathbb{X} \rightarrow \mathbb{R}$. Now, let $K := \max \left\{ \mathcal{C} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}, \mathcal{C} \left\{ \mathcal{I}_{\mathbb{Q}}^{-\frac{1}{2}}(\boldsymbol{\theta}^*) \right\} \right\} \in (0, +\infty)$. By noticing that $\Delta(\boldsymbol{\theta}^* + \mathbf{z}) = \mathbf{z} = -\mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \mathbf{g}$, it follows from the inequality (97) that while being conditioned on the event $\mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_1(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*)) \cap \mathcal{E}_2(\delta)$,

$$\begin{aligned}
 &\widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^* + \mathbf{z}) - \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^*) \\
 &\leq \mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left\{ \hat{f}_0(X) - f(X; \boldsymbol{\theta}^*) \right\} \left[2 \{\Delta(\boldsymbol{\theta})\}^\top \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) + \{\Delta(\boldsymbol{\theta})\}^\top \nabla_{\boldsymbol{\theta}}^2 f(X; \boldsymbol{\theta}^*) \Delta(\boldsymbol{\theta}) \right] \right] \\
 &\quad - \frac{1}{2} \mathbf{z}^\top \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) \mathbf{z} + \frac{\mathcal{B}_2}{2} \sqrt{\log \left(\frac{d}{\delta} \right)} \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \|\mathbf{z}\|_2^2 + \frac{\mathcal{B}_3}{6} \|\mathbf{z}\|_2^3 \\
 &\stackrel{(i)}{\leq} \mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left\{ \hat{f}_0(X) - f(X; \boldsymbol{\theta}^*) \right\} \left[2 \{\Delta(\boldsymbol{\theta})\}^\top \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) + \{\Delta(\boldsymbol{\theta})\}^\top \nabla_{\boldsymbol{\theta}}^2 f(X; \boldsymbol{\theta}^*) \Delta(\boldsymbol{\theta}) \right] \right] \\
 &\quad - \frac{1}{2} \mathbf{z}^\top \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) \mathbf{z} + 2\mathcal{B}_2 \cdot K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \\
 &\quad \cdot \left\{ \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right\}^2 \quad (102) \\
 &\quad + \mathcal{B}_1^2 \mathcal{B}_2 \cdot K^2 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^2 \log^{\frac{5}{2}} \left(\frac{d}{\delta} \right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^2 \\
 &\quad + \frac{8}{3} \mathcal{B}_3 \cdot K^3 (1 + C_{\text{dr}})^3 (1 + C_{\text{rf}})^3 \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \left\{ \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right\}^3 \\
 &\quad + \frac{2}{3} \mathcal{B}_1^3 \mathcal{B}_3 \cdot K^3 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^3 \log^3 \left(\frac{d}{\delta} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^3,
 \end{aligned}$$

where the step (i) utilizes the bound (99) together with the following simple inequality:

$$(x + y)^n \leq 2^{n-1} (x^n + y^n), \quad \forall (x, y, n) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{N}.$$

On the other hand, for every $\boldsymbol{\theta} \in \mathbb{B}_{r(\delta)}(\boldsymbol{\theta}^*)$, one can observe by taking advantage of the lower bound (98) that while being on the event $\mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_2(\delta)$,

$$\begin{aligned} & \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^*) \\ & \geq \mathbb{E}_{X \sim \mathbb{P}_X} \left[\{\hat{\rho}(X) - \rho^*(X)\} \left\{ \hat{f}_0(X) - f(X; \boldsymbol{\theta}^*) \right\} \left[2 \{\Delta(\boldsymbol{\theta})\}^\top \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) + \{\Delta(\boldsymbol{\theta})\}^\top \nabla_{\boldsymbol{\theta}}^2 f(X; \boldsymbol{\theta}^*) \Delta(\boldsymbol{\theta}) \right] \right] \\ & \quad + \frac{1}{2} \{\Delta(\boldsymbol{\theta}) - \mathbf{z}\}^\top \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) \{\Delta(\boldsymbol{\theta}) - \mathbf{z}\} - \frac{1}{2} \mathbf{z}^\top \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) \mathbf{z} \\ & \quad - \frac{9}{2} \mathcal{B}_2 \cdot K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \\ & \quad \cdot \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^2 \\ & \quad - \frac{9}{2} \mathcal{B}_3 \cdot K^3 (1 + C_{\text{dr}})^3 (1 + C_{\text{rf}})^3 \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^3. \end{aligned} \quad (103)$$

Subtracting the bound (102) from (104) yields that on the event $\mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_1(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*)) \cap \mathcal{E}_2(\delta)$,

$$\begin{aligned} & \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^* + \mathbf{z}) \\ & \geq \frac{1}{2} \{\Delta(\boldsymbol{\theta}) - \mathbf{z}\}^\top \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) \{\Delta(\boldsymbol{\theta}) - \mathbf{z}\} \\ & \quad - \left[\frac{13}{2} \mathcal{B}_2 \cdot K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \right. \\ & \quad \cdot \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^2 \\ & \quad + \mathcal{B}_1^2 \mathcal{B}_2 \cdot K^2 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^2 \log^{\frac{5}{2}} \left(\frac{d}{\delta} \right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^2 \\ & \quad + \frac{43}{6} \mathcal{B}_3 \cdot K^3 (1 + C_{\text{dr}})^3 (1 + C_{\text{rf}})^3 \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \\ & \quad \cdot \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^3 \\ & \quad \left. + \frac{2}{3} \mathcal{B}_1^3 \mathcal{B}_3 \cdot K^3 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^3 \log^3 \left(\frac{d}{\delta} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^3 \right] \end{aligned} \quad (104)$$

for every $\boldsymbol{\theta} \in \mathbb{B}_{r(\delta)}(\boldsymbol{\theta}^*)$. At this point, we consider the d -dimensional ellipsoid

$$\begin{aligned}
 \Gamma(\delta) &:= \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \frac{1}{2} \{ \Delta(\boldsymbol{\theta}) - \mathbf{z} \}^\top \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) \{ \Delta(\boldsymbol{\theta}) - \mathbf{z} \} \right. \\
 &\leq \frac{13}{2} \mathcal{B}_2 \cdot K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \\
 &\quad \cdot \left[\sqrt{\frac{\text{Trace} \{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \}}{n_{\mathbb{Q}}}} \right]^2 \\
 &\quad + \mathcal{B}_1^2 \mathcal{B}_2 \cdot K^2 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^2 \log^{\frac{5}{2}} \left(\frac{d}{\delta} \right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^2 \\
 &\quad + \frac{43}{6} \mathcal{B}_3 \cdot K^3 (1 + C_{\text{dr}})^3 (1 + C_{\text{rf}})^3 \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \\
 &\quad \cdot \left[\sqrt{\frac{\text{Trace} \{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \}}{n_{\mathbb{Q}}}} \right]^3 \\
 &\quad \left. + \frac{2}{3} \mathcal{B}_1^3 \mathcal{B}_3 \cdot K^3 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^3 \log^3 \left(\frac{d}{\delta} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^3 \right\}. \tag{105}
 \end{aligned}$$

Then from the inequality (104), it follows for every $\boldsymbol{\theta} \in \mathbb{B}_{r(\delta)}(\boldsymbol{\theta}^*) \setminus \Gamma(\delta)$ that $\widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^* + \mathbf{z}) > 0$ on the event $\mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_1(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*)) \cap \mathcal{E}_2(\delta)$. Together with $\left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}} = \lambda_{\max} \{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \} =$

$\frac{1}{\lambda_{\min}\{\mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*)\}}$, we obtain for any $\boldsymbol{\theta} \in \Gamma(\delta)$ that

$$\begin{aligned}
& \|\Delta(\boldsymbol{\theta}) - \mathbf{z}\|_2^2 \\
& \leq 13\mathcal{B}_2 \cdot K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}} \log^{\frac{3}{2}}\left(\frac{d}{\delta}\right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \\
& \quad \cdot \left[\sqrt{\frac{\text{Trace}\left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace}\left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^2 \\
& \quad + 2\mathcal{B}_1^2 \mathcal{B}_2 \cdot K^2 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^3 \log^{\frac{5}{2}}\left(\frac{d}{\delta}\right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^2 \\
& \quad + \frac{43}{3} \mathcal{B}_3 \cdot K^3 (1 + C_{\text{dr}})^3 (1 + C_{\text{rf}})^3 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}} \log^{\frac{3}{2}}\left(\frac{d}{\delta}\right) \\
& \quad \cdot \left[\sqrt{\frac{\text{Trace}\left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace}\left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^3 \\
& \quad + \frac{4}{3} \mathcal{B}_1^3 \mathcal{B}_3 \cdot K^3 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^4 \log^3\left(\frac{d}{\delta}\right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^3.
\end{aligned} \tag{106}$$

Thus, the triangle inequality implies on the event $\mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right)$ that for any $\boldsymbol{\theta} \in \Gamma(\delta)$, one has

$$\begin{aligned}
 \|\Delta(\boldsymbol{\theta})\|_2^2 &\leq 2 \|\Delta(\boldsymbol{\theta}) - \mathbf{z}\|_2^2 + 2 \|\mathbf{z}\|_2^2 \\
 &\stackrel{(j)}{\leq} 26\mathcal{B}_2 \cdot K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}} \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \\
 &\quad \cdot \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^2 \\
 &\quad + 4\mathcal{B}_1^2 \mathcal{B}_2 \cdot K^2 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^3 \log^{\frac{5}{2}} \left(\frac{d}{\delta} \right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^2 \\
 &\quad + \frac{86}{3} \mathcal{B}_3 \cdot K^3 (1 + C_{\text{dr}})^3 (1 + C_{\text{rf}})^3 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}} \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \\
 &\quad \cdot \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^3 \\
 &\quad + \frac{8}{3} \mathcal{B}_1^3 \mathcal{B}_3 \cdot K^3 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^4 \log^3 \left(\frac{d}{\delta} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^3 \\
 &\quad + 8K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log \left(\frac{d}{\delta} \right) \\
 &\quad \cdot \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^2 \\
 &\quad + 4\mathcal{B}_1^2 \cdot K^2 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^2 \log^2 \left(\frac{d}{\delta} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^2,
 \end{aligned} \tag{107}$$

where the step (j) make use of the consequence (99) from Lemma 12 with $\mathbf{A} = \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*)$. In order to guarantee that

$$\log \left(\frac{d}{\delta} \right) \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^2$$

is the leading term, we only need it to dominate the remaining terms. In particular, if $\min \{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \geq \kappa \cdot \mathcal{N}_1 \log \left(\frac{d}{\delta} \right)$, then one can conclude from the bound (107) that on the event $\mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right)$,

$$\begin{aligned} \|\Delta(\boldsymbol{\theta})\|_2^2 &\leq 9K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log \left(\frac{d}{\delta} \right) \\ &\quad \cdot \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^2 \\ &= r^2(\delta) \end{aligned} \quad (108)$$

for every $\boldsymbol{\theta} \in \Gamma(\delta)$. To sum up, we have established the following conclusions so far:

(FACT A) On the event $\mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right) \cap \mathcal{E}_2(\delta)$, we have $\widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^* + \mathbf{z}) > 0$ for every $\boldsymbol{\theta} \in \mathbb{B}_{r(\delta)}(\boldsymbol{\theta}^*) \setminus \Gamma(\delta)$.

(FACT B) On the event $\mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right)$, we have $\Gamma(\delta) \subseteq \mathbb{B}_{r(\delta)}(\boldsymbol{\theta}^*)$ if $\min \{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \geq \kappa \cdot \mathcal{N}_1 \log \left(\frac{d}{\delta} \right)$.

Lastly, it is time to put everything (FACT A & B) together to establish the part (i) of Lemma 11. Towards this end, let us claim that on the event $\mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right) \cap \mathcal{E}_2(\delta)$, $\widehat{\mathcal{R}}_{\text{DR}} : \mathbb{R}^d \rightarrow \mathbb{R}$ attains a local minimum in the ellipsoid $\Gamma(\delta) \subseteq \mathbb{R}^d$ if $\min \{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \geq \kappa \cdot \mathcal{N}_1 \log \left(\frac{d}{\delta} \right)$. Due to its continuity together with the compactness of the d -dimensional closed ball $\mathbb{B}_{r(\delta)}(\boldsymbol{\theta}^*) \subseteq \mathbb{R}^d$, the empirical DR risk $\widehat{\mathcal{R}}_{\text{DR}} : \mathbb{B}_{r(\delta)}(\boldsymbol{\theta}^*) \rightarrow \mathbb{R}$ achieves a global minimum, and therefore it becomes a local minimum of $\widehat{\mathcal{R}}_{\text{DR}} : \mathbb{R}^d \rightarrow \mathbb{R}$. Let $\bar{\boldsymbol{\theta}} \in \arg\min \left\{ \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{B}_{r(\delta)}(\boldsymbol{\theta}^*) \right\}$. Then if $\bar{\boldsymbol{\theta}} \in \mathbb{B}_{r(\delta)}(\boldsymbol{\theta}^*) \setminus \Gamma(\delta)$, (FACT A) implies

$$\widehat{\mathcal{R}}_{\text{DR}}(\bar{\boldsymbol{\theta}}) > \widehat{\mathcal{R}}_{\text{DR}}(\boldsymbol{\theta}^* + \mathbf{z}) \stackrel{(k)}{\geq} \widehat{\mathcal{R}}_{\text{DR}}(\bar{\boldsymbol{\theta}}),$$

which yields a contradiction, where the step (k) follows since $\boldsymbol{\theta}^* + \mathbf{z} \in \Gamma(\delta) \subseteq \mathbb{B}_{r(\delta)}(\boldsymbol{\theta}^*)$, which holds by (FACT B). Hence, one can conclude that $\bar{\boldsymbol{\theta}} \in \Gamma(\delta)$ as desired. Assumption 6 implies that the global minimum of the empirical DR risk $\widehat{\mathcal{R}}_{\text{DR}} : \mathbb{R}^d \rightarrow \mathbb{R}$ belongs to the ellipsoid $\Gamma(\delta) \subseteq \mathbb{R}^d$ on the event $\mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right) \cap \mathcal{E}_2(\delta)$, i.e., on the event $\mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right) \cap \mathcal{E}_2(\delta)$,

$$\hat{\boldsymbol{\theta}}_{\text{DR}} \in \Gamma(\delta) \subseteq \mathbb{B}_{r(\delta)}(\boldsymbol{\theta}^*), \quad (109)$$

provided that $\min \{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \geq \kappa \cdot \mathcal{N}_1 \log \left(\frac{d}{\delta} \right)$.

In the sequel, we shall work with the high-probability event $\mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right) \cap \mathcal{E}_2(\delta)$ in order to further establish the part (ii) of Lemma 11. Because $\hat{\boldsymbol{\theta}}_{\text{DR}} \in \Gamma(\delta)$ while being on the event

$\mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right) \cap \mathcal{E}_2(\delta)$ provided that $\min \{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \geq \kappa \cdot \mathcal{N}_1 \log \left(\frac{d}{\delta} \right)$, it follows that

$$\begin{aligned}
 & \left\| \mathcal{I}_{\mathbb{Q}}^{\frac{1}{2}}(\boldsymbol{\theta}^*) \left\{ \Delta \left(\hat{\boldsymbol{\theta}}_{\text{DR}} \right) - \mathbf{z} \right\} \right\|_2^2 \\
 &= \left\{ \Delta \left(\hat{\boldsymbol{\theta}}_{\text{DR}} \right) - \mathbf{z} \right\}^{\top} \mathcal{I}_{\mathbb{Q}}(\boldsymbol{\theta}^*) \left\{ \Delta \left(\hat{\boldsymbol{\theta}}_{\text{DR}} \right) - \mathbf{z} \right\} \\
 &\leq 13\mathcal{B}_2 \cdot K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \\
 &\quad \cdot \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^2 \\
 &\quad + 2\mathcal{B}_1^2 \mathcal{B}_2 \cdot K^2 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^2 \log^{\frac{5}{2}} \left(\frac{d}{\delta} \right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^2 \\
 &\quad + \frac{43}{3} \mathcal{B}_3 \cdot K^3 (1 + C_{\text{dr}})^3 (1 + C_{\text{rf}})^3 \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \\
 &\quad \cdot \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^3 \\
 &\quad + \frac{4}{3} \mathcal{B}_1^3 \mathcal{B}_3 \cdot K^3 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^3 \log^3 \left(\frac{d}{\delta} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^3
 \end{aligned} \tag{110}$$

On the other hand, one can readily apply Lemma 12 by letting $\mathbf{A} = \mathcal{I}_{\mathbb{Q}}^{-\frac{1}{2}}(\boldsymbol{\theta}^*)$ to obtain

$$\begin{aligned}
 & \left\| \mathcal{I}_{\mathbb{Q}}^{\frac{1}{2}}(\boldsymbol{\theta}^*) \mathbf{z} \right\|_2^2 = \left\| \mathcal{I}_{\mathbb{Q}}^{-\frac{1}{2}}(\boldsymbol{\theta}^*) \mathbf{g} \right\|_2^2 \\
 &\leq c \left\{ \mathcal{I}_{\mathbb{Q}}^{-\frac{1}{2}}(\boldsymbol{\theta}^*) \right\}^2 \left[\sqrt{\frac{\nu_{\mathbb{P}} \left\{ \mathcal{I}_{\mathbb{Q}}^{-\frac{1}{2}}(\boldsymbol{\theta}^*) \right\} \log \left(\frac{d}{\delta} \right)}{n_{\mathbb{P}}}} + \sqrt{\frac{\nu_{\mathbb{Q}} \left\{ \mathcal{I}_{\mathbb{Q}}^{-\frac{1}{2}}(\boldsymbol{\theta}^*) \right\} \log \left(\frac{d}{\delta} \right)}{n_{\mathbb{Q}}}} \right. \\
 &\quad \left. + \mathcal{B}_1 \left\| \mathcal{I}_{\mathbb{Q}}^{-\frac{1}{2}}(\boldsymbol{\theta}^*) \right\|_{\text{op}} \log \left(\frac{d}{\delta} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right) \right]^2 \\
 &\stackrel{(1)}{\leq} K^2 \left[\sqrt{2} (1 + C_{\text{dr}}) (1 + C_{\text{rf}}) \sqrt{\log \left(\frac{d}{\delta} \right)} \left\{ \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{d}{n_{\mathbb{Q}}}} \right\} \right. \\
 &\quad \left. + \mathcal{B}_1 \left\| \mathcal{I}_{\mathbb{Q}}^{-\frac{1}{2}}(\boldsymbol{\theta}^*) \right\|_{\text{op}} \log \left(\frac{d}{\delta} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right) \right]^2
 \end{aligned} \tag{111}$$

$$\begin{aligned} &\leq 4K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log \left(\frac{d}{\delta} \right) \left\{ \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{d}{n_{\mathbb{Q}}}} \right\}^2 \\ &\quad + 2\mathcal{B}_1^2 \cdot K^2 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}} \log^2 \left(\frac{d}{\delta} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^2 \end{aligned}$$

on the event $\mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-\frac{1}{2}}(\boldsymbol{\theta}^*) \right)$, where the step (l) makes use of (100) and (101) with $\mathbf{A} = \mathcal{I}_{\mathbb{Q}}^{-\frac{1}{2}}(\boldsymbol{\theta}^*)$. Thus, we obtain by taking two pieces (110) and (111) collectively that while being on the event

$$\Lambda(\delta) := \left\{ \mathcal{A}_{\mathbb{P}} \cap \mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right) \cap \mathcal{E}_2(\delta) \right\} \cap \mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-\frac{1}{2}}(\boldsymbol{\theta}^*) \right),$$

one has

$$\begin{aligned} &\left\| \mathcal{I}_{\mathbb{Q}}^{\frac{1}{2}}(\boldsymbol{\theta}^*) \left(\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^* \right) \right\|_2^2 \\ &= \left\| \mathcal{I}_{\mathbb{Q}}^{\frac{1}{2}}(\boldsymbol{\theta}^*) \Delta \left(\hat{\boldsymbol{\theta}}_{\text{DR}} \right) \right\|_2^2 \\ &\stackrel{(m)}{\leq} 2 \left\| \mathcal{I}_{\mathbb{Q}}^{\frac{1}{2}}(\boldsymbol{\theta}^*) \left\{ \Delta \left(\hat{\boldsymbol{\theta}}_{\text{DR}} \right) - \mathbf{z} \right\} \right\|_2^2 + 2 \left\| \mathcal{I}_{\mathbb{Q}}^{\frac{1}{2}}(\boldsymbol{\theta}^*) \mathbf{z} \right\|_2^2 \\ &\leq 26\mathcal{B}_2 \cdot K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \\ &\quad \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^2 \\ &\quad + 4\mathcal{B}_1^2 \mathcal{B}_2 \cdot K^2 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^2 \log^{\frac{5}{2}} \left(\frac{d}{\delta} \right) \left(\frac{1}{\sqrt{n_{\mathbb{P}}}} + \frac{1}{\sqrt{n_{\mathbb{Q}}}} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^2 \\ &\quad + \frac{86}{3} \mathcal{B}_3 \cdot K^3 (1 + C_{\text{dr}})^3 (1 + C_{\text{rf}})^3 \log^{\frac{3}{2}} \left(\frac{d}{\delta} \right) \\ &\quad \left[\sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-2}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{Q}}}} \right]^3 \\ &\quad + \frac{8}{3} \mathcal{B}_1^3 \mathcal{B}_3 \cdot K^3 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}}^3 \log^3 \left(\frac{d}{\delta} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^3 \\ &\quad + 8K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log \left(\frac{d}{\delta} \right) \left\{ \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{d}{n_{\mathbb{Q}}}} \right\}^2 \\ &\quad + 4\mathcal{B}_1^2 \cdot K^2 \left\| \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\|_{\text{op}} \log^2 \left(\frac{d}{\delta} \right) \left(\frac{1}{n_{\mathbb{P}}} + \frac{1}{n_{\mathbb{Q}}} \right)^2 \end{aligned} \tag{112}$$

when $\min \{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \geq \kappa \cdot \mathcal{N}_1 \log \left(\frac{d}{\delta} \right)$, where the step (m) follows due to the triangle inequality. To guarantee that

$$\log \left(\frac{d}{\delta} \right) \left\{ \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{d}{n_{\mathbb{Q}}}} \right\}^2$$

becomes the leading term, it suffices to make it dominate the remaining terms. In particular, whenever $\min \{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \geq \kappa \cdot \max \{\mathcal{N}_1, \mathcal{N}_2\} \log \left(\frac{d}{\delta} \right)$, then it follows directly from the inequality (112) that while being on the event $\Lambda(\delta)$, we have

$$\begin{aligned} \left\| \mathcal{I}_{\mathbb{Q}}^{\frac{1}{2}}(\boldsymbol{\theta}^*) \left(\hat{\boldsymbol{\theta}}_{\text{DR}} - \boldsymbol{\theta}^* \right) \right\|_2^2 &\leq 9K^2 (1 + C_{\text{dr}})^2 (1 + C_{\text{rf}})^2 \log \left(\frac{d}{\delta} \right) \\ &\quad \cdot \left\{ \sqrt{\frac{\text{Trace} \left\{ \mathcal{I}_{\mathbb{P}}(\boldsymbol{\theta}^*) \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right\}}{n_{\mathbb{P}}}} + \sqrt{\frac{d}{n_{\mathbb{Q}}}} \right\}^2. \end{aligned} \quad (113)$$

Hence, while being on the event $\Lambda(\delta)$, the desired conclusions (i) and (ii) both hold if $\min \{n_{\mathbb{P}}, n_{\mathbb{Q}}\} \geq \kappa \cdot \max \{\mathcal{N}_1, \mathcal{N}_2\} \log \left(\frac{d}{\delta} \right)$. This completes the proof of Lemma 11 since

$$\begin{aligned} &(\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}) \{ \Lambda(\delta) \} \\ &= (\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}) \left\{ \mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right) \cap \mathcal{E}_2(\delta) \cap \mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-\frac{1}{2}}(\boldsymbol{\theta}^*) \right) \right\} \\ &\stackrel{(n)}{\geq} 1 - (\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}) \left\{ (\mathbb{O}^{n_{\mathbb{P}}} \times \mathbb{X}_{n_{\mathbb{Q}}}) \setminus \mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-1}(\boldsymbol{\theta}^*) \right) \right\} \\ &\quad - (\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}) \left\{ (\mathbb{O}^{n_{\mathbb{P}}} \times \mathbb{X}_{n_{\mathbb{Q}}}) \setminus \mathcal{E}_2(\delta) \right\} \\ &\quad - (\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}) \left\{ (\mathbb{O}^{n_{\mathbb{P}}} \times \mathbb{X}_{n_{\mathbb{Q}}}) \setminus \mathcal{E}_1 \left(\delta; \mathcal{I}_{\mathbb{Q}}^{-\frac{1}{2}}(\boldsymbol{\theta}^*) \right) \right\} \\ &\stackrel{(o)}{\geq} 1 - 8\delta, \end{aligned}$$

where the step (n) arises from the union bound, and the step (o) holds true due to Lemma 12 and 13.

D.1.1. PROOF OF LEMMA 12

For notational simplicity, let $\mathbf{g} := \nabla_{\boldsymbol{\theta}} \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}^*) - \mathbb{E} \left[\nabla_{\boldsymbol{\theta}} \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) (\boldsymbol{\theta}^*) \right]$ under $\left(\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}, \mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}} \right) \sim \mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{n_{\mathbb{Q}}}$. Then, we obtain from the equation (72) that

$$\mathbf{g} = \left(\widehat{\mathbb{P}} - \mathbb{P} \right) [\Phi_{\mathbb{P}}(X, Y)] + \left(\widehat{\mathbb{Q}}_X - \mathbb{Q}_X \right) [\Phi_{\mathbb{Q}}(X)] \quad (\mathbb{P}^{\otimes n_{\mathbb{P}}} \otimes \mathbb{Q}_X^{n_{\mathbb{Q}}})\text{-almost surely}, \quad (114)$$

where $\widehat{\mathbb{P}} \in \Delta(\mathbb{X} \times \mathbb{R})$ and $\widehat{\mathbb{Q}}_X \in \Delta(\mathbb{X})$ denote the empirical distributions for the $n_{\mathbb{P}}$ labeled source samples $\mathbf{O}_{1:n_{\mathbb{P}}}^{\mathbb{P}}$ and $n_{\mathbb{Q}}$ unlabeled target samples $\mathbf{X}_{1:n_{\mathbb{Q}}}^{\mathbb{Q}}$, respectively, i.e., $\widehat{\mathbb{P}} := \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \delta_{(X_i^{\mathbb{P}}, Y_i^{\mathbb{P}})}$ and $\widehat{\mathbb{Q}}_X := \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \delta_{X_j^{\mathbb{Q}}}$, and the functions $\Phi_{\mathbb{P}} : \mathbb{X} \times \mathbb{R} \rightarrow \mathbb{R}^d$ and $\Phi_{\mathbb{Q}} : \mathbb{X} \rightarrow \mathbb{R}^d$ are defined as (88).

Now, we fix any matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$. One can readily find that

$$\|\mathbf{A}\Phi_{\mathbb{P}}(X, Y)\|_2 = 2\hat{\rho}(X) \left| Y - \hat{f}_0(X) \right| \|\mathbf{A}\nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*)\|_2 \leq 2C_{\text{dr}}(1 + C_{\text{rf}}) b_1 \|\mathbf{A}\|_{\text{op}} \quad \mathbb{P}\text{-almost surely,}$$

which immediately yields

$$\begin{aligned} \mathcal{B}_{\mathbb{P}}(\alpha) &:= \inf \left\{ t \in (0, +\infty) : \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\exp \left\{ \left(\frac{\|\mathbf{A} \{ \Phi_{\mathbb{P}}(X, Y) - \mathbb{E}_{(X,Y) \sim \mathbb{P}} [\Phi_{\mathbb{P}}(X, Y)] \|_2}{t} \right)^\alpha \right\} \right] \leq 2 \right\} \\ &\leq \inf \left\{ t \in (0, +\infty) : \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\exp \left\{ \left(\frac{4C_{\text{dr}}(1 + C_{\text{rf}}) b_1 \|\mathbf{A}\|_{\text{op}}}{t} \right)^\alpha \right\} \right] \leq 2 \right\} \quad (115) \\ &\leq \mathcal{B}_1 \|\mathbf{A}\|_{\text{op}} (\log 2)^{-\frac{1}{\alpha}} \end{aligned}$$

for every $\alpha \in [1, +\infty)$. By virtue of Lemma 3, there exists an absolute constant $\mathcal{C}_{\mathbb{P}}(\mathbf{A}) \in (0, +\infty)$ such that for any $\delta \in (0, 1]$, it holds by utilizing the observation (115) that

$$\begin{aligned} \left\| \left(\hat{\mathbb{P}} - \mathbb{P} \right) [\mathbf{A}\Phi_{\mathbb{P}}(X, Y)] \right\|_2 &\leq \mathcal{C}_{\mathbb{P}}(\mathbf{A}) \left[\sqrt{\frac{\mathcal{V}_{\mathbb{P}}(\mathbf{A}) \log \left(\frac{2d}{\delta} \right)}{n_{\mathbb{P}}}} + \mathcal{B}_1 \|\mathbf{A}\|_{\text{op}} (\log 2)^{-\frac{1}{\alpha}} \right. \\ &\quad \cdot \log^{\frac{1}{\alpha}} \left\{ \frac{\mathcal{B}_1 \|\mathbf{A}\|_{\text{op}} (\log 2)^{-\frac{1}{\alpha}}}{\sqrt{\mathcal{V}_{\mathbb{P}}(\mathbf{A})}} \right\} \cdot \frac{\log \left(\frac{2d}{\delta} \right)}{n_{\mathbb{P}}} \left. \right] \quad (116) \end{aligned}$$

with probability exceeding $1 - \frac{\delta}{2}$ under $\mathbf{O}_{1:n_{\mathbb{P}}} \sim \mathbb{P}^{\otimes n_{\mathbb{P}}}$, for every $\alpha \in [1, +\infty)$. By taking $\alpha \rightarrow +\infty$ in the inequality (116), one can conclude that

$$\begin{aligned} \mathbb{P}^{\otimes n_{\mathbb{P}}} \left(\left\{ \left\| \left(\hat{\mathbb{P}} - \mathbb{P} \right) [\mathbf{A}\Phi_{\mathbb{P}}(X, Y)] \right\|_2 \leq \mathcal{C}_{\mathbb{P}}(\mathbf{A}) \right. \right. \\ \left. \left. \left\{ \sqrt{\frac{\mathcal{V}_{\mathbb{P}}(\mathbf{A}) \log \left(\frac{2d}{\delta} \right)}{n_{\mathbb{P}}}} + \mathcal{B}_1 \|\mathbf{A}\|_{\text{op}} \cdot \frac{\log \left(\frac{2d}{\delta} \right)}{n_{\mathbb{P}}} \right\} \right\} \right) \geq 1 - \frac{\delta}{2}, \quad (117) \end{aligned}$$

where $\mathcal{V}_{\mathbb{P}}(\mathbf{A}) := \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\left\| \mathbf{A} \{ \Phi_{\mathbb{P}}(X, Y) - \mathbb{E}_{(X,Y) \sim \mathbb{P}} [\Phi_{\mathbb{P}}(X, Y)] \} \right\|_2^2 \right]$.

On the other hand, it can be observed that

$$\|\mathbf{A}\Phi_{\mathbb{Q}}(X)\|_2 = 2 \left| f(X; \boldsymbol{\theta}^*) - \hat{f}_0(X) \right| \|\mathbf{A}\nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*)\|_2 \leq 2(1 + C_{\text{rf}}) b_1 \|\mathbf{A}\|_{\text{op}},$$

which directly implies

$$\begin{aligned} \mathcal{B}_{\mathbb{Q}}(\alpha) &:= \inf \left\{ t \in (0, +\infty) : \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\exp \left\{ \left(\frac{\|\mathbf{A} \{ \Phi_{\mathbb{Q}}(X) - \mathbb{E}_{X \sim \mathbb{Q}_X} [\Phi_{\mathbb{Q}}(X)] \|_2}{t} \right)^\alpha \right\} \right] \leq 2 \right\} \\ &\leq \inf \left\{ t \in (0, +\infty) : \mathbb{E}_{(X,Y) \sim \mathbb{P}} \left[\exp \left\{ \left(\frac{4(1 + C_{\text{rf}}) b_1 \|\mathbf{A}\|_{\text{op}}}{t} \right)^\alpha \right\} \right] \leq 2 \right\} \quad (118) \\ &= 4(1 + C_{\text{rf}}) b_1 \|\mathbf{A}\|_{\text{op}} \cdot (\log 2)^{-\frac{1}{\alpha}} \\ &\leq \mathcal{B}_1 \|\mathbf{A}\|_{\text{op}} (\log 2)^{-\frac{1}{\alpha}} \end{aligned}$$

for all $\alpha \in [1, +\infty)$. Applying Lemma 3, one can see that there exists a universal constant $\mathcal{C}_Q(\mathbf{A}) \in (0, +\infty)$ such that for any $\delta \in (0, 1]$, we have from the fact (118) that

$$\begin{aligned} \left\| \left(\widehat{Q}_X - Q_X \right) [\mathbf{A} \Phi_Q(X)] \right\|_2 &\leq \mathcal{C}_Q(\mathbf{A}) \left[\sqrt{\frac{\mathcal{V}_Q(\mathbf{A}) \log\left(\frac{2d}{\delta}\right)}{n_Q}} + \mathcal{B}_1 \|\mathbf{A}\|_{\text{op}} (\log 2)^{-\frac{1}{\alpha}} \right. \\ &\quad \cdot \log^{\frac{1}{\alpha}} \left\{ \frac{\mathcal{B}_1 \|\mathbf{A}\|_{\text{op}} (\log 2)^{-\frac{1}{\alpha}}}{\sqrt{\mathcal{V}_P(\mathbf{A})}} \right\} \cdot \frac{\log\left(\frac{2d}{\delta}\right)}{n_Q} \left. \right] \end{aligned} \quad (119)$$

with probability greater than $1 - \frac{\delta}{2}$ under $\mathbf{X}_{1:n_Q} \sim Q_X^{\otimes n_Q}$, for all $\alpha \in [1, +\infty)$. By taking $\alpha \rightarrow +\infty$ in the bound (119), it follows that

$$\begin{aligned} Q_X^{\otimes n_Q} \left(\left\{ \left\| \left(\widehat{Q}_X - Q_X \right) [\mathbf{A} \Phi_Q(X)] \right\|_2 \leq \mathcal{C}_Q(\mathbf{A}) \right. \right. \\ \left. \left. \left\{ \sqrt{\frac{\mathcal{V}_Q(\mathbf{A}) \log\left(\frac{2d}{\delta}\right)}{n_Q}} + \mathcal{B}_1 \|\mathbf{A}\|_{\text{op}} \cdot \frac{\log\left(\frac{2d}{\delta}\right)}{n_Q} \right\} \right\} \right) &\geq 1 - \frac{\delta}{2}, \end{aligned} \quad (120)$$

where $\mathcal{V}_Q(\mathbf{A}) := \mathbb{E}_{X \sim Q_X} \left[\|\mathbf{A} \{\Phi_Q(X) - \mathbb{E}_{X \sim Q_X} [\Phi_Q(X)]\}\|_2^2 \right]$.

Lastly, it is time to put all pieces together. By making use of the union bound together with two conclusions (117) and (120) and setting $\mathcal{C}(\mathbf{A}) := \max \{\mathcal{C}_P(\mathbf{A}), \mathcal{C}_Q(\mathbf{A})\} \in (0, +\infty)$, one has

$$\begin{aligned} &\left\| \mathbf{A} \left\{ \nabla_{\theta} \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_P}^P, \mathbf{X}_{1:n_Q}^Q \right) (\theta^*) - \mathbb{E}_{(\mathbf{O}_{1:n_P}^P, \mathbf{X}_{1:n_Q}^Q) \sim P^{\otimes n_P} \otimes Q_X^{n_Q}} \left[\nabla_{\theta} \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_P}^P, \mathbf{X}_{1:n_Q}^Q \right) (\theta^*) \right] \right\} \right\|_2 \\ &\stackrel{\text{a.s.}}{=} \left\| \left(\widehat{P} - P \right) [\mathbf{A} \Phi_P(X, Y)] + \left(\widehat{Q}_X - Q_X \right) [\mathbf{A} \Phi_Q(X)] \right\|_2 \\ &\stackrel{(a)}{\leq} \left\| \left(\widehat{P} - P \right) [\mathbf{A} \Phi_P(X, Y)] \right\|_2 + \left\| \left(\widehat{Q}_X - Q_X \right) [\mathbf{A} \Phi_Q(X)] \right\|_2 \\ &\leq \mathcal{C}(\mathbf{A}) \left\{ \sqrt{\frac{\mathcal{V}_P(\mathbf{A}) \log\left(\frac{2d}{\delta}\right)}{n_P}} + \sqrt{\frac{\mathcal{V}_Q(\mathbf{A}) \log\left(\frac{2d}{\delta}\right)}{n_Q}} + \mathcal{B}_1 \|\mathbf{A}\|_{\text{op}} \cdot \log\left(\frac{2d}{\delta}\right) \left(\frac{1}{n_P} + \frac{1}{n_Q} \right) \right\} \end{aligned}$$

with probability at least $1 - \delta$ under $(\mathbf{O}_{1:n_P}^P, \mathbf{X}_{1:n_Q}^Q) \sim P^{\otimes n_P} \otimes Q_X^{n_Q}$, where the step (a) follows due to the triangle inequality, as desired.

D.1.2. PROOF OF LEMMA 13

For notational simplicity, we let $\mathbf{H} := \nabla_{\theta}^2 \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_P}^P, \mathbf{X}_{1:n_Q}^Q \right) (\theta^*) - \mathbb{E} \left[\nabla_{\theta}^2 \widehat{\mathcal{R}}_{\text{DR}} \left(\mathbf{O}_{1:n_P}^P, \mathbf{X}_{1:n_Q}^Q \right) (\theta^*) \right]$ for $(\mathbf{O}_{1:n_P}^P, \mathbf{X}_{1:n_Q}^Q) \sim P^{\otimes n_P} \otimes Q_X^{n_Q}$. Then, one may express \mathbf{H} by using the fact (72) that

$$\mathbf{H} = \frac{1}{n_P} \sum_{i=1}^{n_P} \mathbf{U}_i^P + \frac{1}{n_Q} \sum_{j=1}^{n_Q} \mathbf{V}_j^Q, \quad (121)$$

where

$$\mathbf{U}_i^P := 2\hat{\rho} \left(X_i^P \right) \left\{ \hat{f}_0 \left(X_i^P \right) - Y_i^P \right\} \nabla_{\theta}^2 f \left(X_i^P; \theta^* \right)$$

$$\begin{aligned}
& -2\mathbb{E}_{(X,Y)\sim\mathbb{P}} \left[\hat{\rho}(X) \left\{ \hat{f}_0(X) - Y \right\} \nabla_{\boldsymbol{\theta}}^2 f(X; \boldsymbol{\theta}^*) \right], \\
\mathbf{V}_j^{\mathbb{Q}} := & 2 \left[\nabla_{\boldsymbol{\theta}} f(X_j^{\mathbb{Q}}; \boldsymbol{\theta}^*) \left\{ \nabla_{\boldsymbol{\theta}} f(X_j^{\mathbb{Q}}; \boldsymbol{\theta}^*) \right\}^{\top} \right. \\
& + \left\{ f(X_j^{\mathbb{Q}}; \boldsymbol{\theta}^*) - \hat{f}_0(X_j^{\mathbb{Q}}) \right\} \nabla_{\boldsymbol{\theta}}^2 f(X_j^{\mathbb{Q}}; \boldsymbol{\theta}^*) \left. \right] \\
& - 2\mathbb{E}_{X\sim\mathbb{Q}_X} \left[\nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) \left\{ \nabla_{\boldsymbol{\theta}} f(X; \boldsymbol{\theta}^*) \right\}^{\top} + \left\{ f(X; \boldsymbol{\theta}^*) - \hat{f}_0(X) \right\} \nabla_{\boldsymbol{\theta}}^2 f(X; \boldsymbol{\theta}^*) \right],
\end{aligned} \tag{122}$$

for each $(i, j) \in [n_{\mathbb{P}}] \times [n_{\mathbb{Q}}]$. At this moment, one can readily realize the following facts of the $d \times d$ random matrices $\{\mathbf{U}_i^{\mathbb{P}} : i \in [n_{\mathbb{P}}]\}$:

- The operator norm of $\mathbf{U}_i^{\mathbb{P}}$ can be bounded as

$$\begin{aligned}
& \left\| \mathbf{U}_i^{\mathbb{P}} \right\|_{\text{op}} \\
& \leq 2 \left\| \hat{\rho}(X_i^{\mathbb{P}}) \left\{ \hat{f}_0(X_i^{\mathbb{P}}) - Y_i^{\mathbb{P}} \right\} \nabla_{\boldsymbol{\theta}}^2 f(X_i^{\mathbb{P}}; \boldsymbol{\theta}^*) \right\|_{\text{op}} \\
& \quad + 2\mathbb{E}_{(X,Y)\sim\mathbb{P}} \left[\left\| \hat{\rho}(X) \left\{ \hat{f}_0(X) - Y \right\} \nabla_{\boldsymbol{\theta}}^2 f(X; \boldsymbol{\theta}^*) \right\|_{\text{op}} \right] \\
& \stackrel{(a)}{\leq} 4C_{\text{dr}} (1 + C_{\text{rf}}) \left\| \nabla_{\boldsymbol{\theta}}^2 f(X_i^{\mathbb{P}}; \boldsymbol{\theta}^*) \right\|_{\text{op}} \\
& \stackrel{(b)}{\leq} 4C_{\text{dr}} (1 + C_{\text{rf}}) b_2
\end{aligned} \tag{123}$$

for every $i \in [n_{\mathbb{P}}]$, where the step (a) holds \mathbb{P} -almost surely by Assumptions 3 and 4, and the step (b) comes from Assumption 5.

- Using the upper bound (123), one can obtain $(\mathbf{U}_i^{\mathbb{P}})^2 \preceq \sigma_{\mathbb{P}}^2 \mathbf{I}_d$ for all $i \in [n_{\mathbb{P}}]$ \mathbb{P} -almost surely, where

$$\sigma_{\mathbb{P}}^2 := 16C_{\text{dr}}^2 (1 + C_{\text{rf}})^2 b_2^2.$$

We can combine the above properties on the $d \times d$ random matrices $\{\mathbf{U}_i^{\mathbb{P}} : i \in [n_{\mathbb{P}}]\}$ together with the matrix Hoeffding inequality (*Theorem 1.3* in Tropp (2012)) to reach

$$\begin{aligned}
(\mathbb{P}^{\otimes n_{\mathbb{P}}}) \left(\left\{ \left\| \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \mathbf{U}_i^{\mathbb{P}} \right\|_{\text{op}} > t \right\} \right) & \leq 2d \exp \left(-\frac{n_{\mathbb{P}} t^2}{8\sigma_{\mathbb{P}}^2} \right) \\
& = 2d \exp \left\{ -\frac{n_{\mathbb{P}} t^2}{128C_{\text{dr}}^2 (1 + C_{\text{rf}})^2 b_2^2} \right\}
\end{aligned} \tag{124}$$

for any $t \in \mathbb{R}_+$. Thus, it follows for any $\delta \in (0, 1]$ that

$$\left\| \frac{1}{n_{\mathbb{P}}} \sum_{i=1}^{n_{\mathbb{P}}} \mathbf{U}_i^{\mathbb{P}} \right\|_{\text{op}} \leq 8\sqrt{2}C_{\text{dr}} (1 + C_{\text{rf}}) b_2 \sqrt{\frac{\log \left(\frac{2d}{\delta} \right)}{n_{\mathbb{P}}}} \tag{125}$$

with probability at least $1 - \delta$ under the probability measure $\mathbb{P}^{\otimes n_{\mathbb{P}}}$.

Likewise, one can make the following observations on the $d \times d$ random matrices $\{\mathbf{V}_j^{\mathbb{Q}} : j \in [n_{\mathbb{Q}}]\}$:

- The operator norm of $\mathbf{V}_j^{\mathbb{Q}}$ can be bounded as

$$\begin{aligned}
 \left\| \mathbf{V}_j^{\mathbb{Q}} \right\|_{\text{op}} &\leq 2 \left\| \nabla_{\boldsymbol{\theta}} f \left(X_j^{\mathbb{Q}}; \boldsymbol{\theta}^* \right) \left\{ \nabla_{\boldsymbol{\theta}} f \left(X_j^{\mathbb{Q}}; \boldsymbol{\theta}^* \right) \right\}^{\top} \right\|_{\text{op}} \\
 &\quad + 2 \left\| \left\{ f \left(X_j^{\mathbb{Q}}; \boldsymbol{\theta}^* \right) - \hat{f}_0 \left(X_j^{\mathbb{Q}} \right) \right\} \nabla_{\boldsymbol{\theta}}^2 f \left(X_j^{\mathbb{Q}}; \boldsymbol{\theta}^* \right) \right\|_{\text{op}} \\
 &\quad + 2 \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\left\| \nabla_{\boldsymbol{\theta}} f \left(X; \boldsymbol{\theta}^* \right) \left\{ \nabla_{\boldsymbol{\theta}} f \left(X; \boldsymbol{\theta}^* \right) \right\}^{\top} \right\|_{\text{op}} \right] \\
 &\quad + 2 \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\left\| \left\{ f \left(X; \boldsymbol{\theta}^* \right) - \hat{f}_0(X) \right\} \nabla_{\boldsymbol{\theta}}^2 f \left(X; \boldsymbol{\theta}^* \right) \right\|_{\text{op}} \right] \\
 &= 2 \left\| \nabla_{\boldsymbol{\theta}} f \left(X_j^{\mathbb{Q}}; \boldsymbol{\theta}^* \right) \right\|_2^2 \\
 &\quad + 2 \left| f \left(X_j^{\mathbb{Q}}; \boldsymbol{\theta}^* \right) - \hat{f}_0 \left(X_j^{\mathbb{Q}} \right) \right| \left\| \nabla_{\boldsymbol{\theta}}^2 f \left(X_j^{\mathbb{Q}}; \boldsymbol{\theta}^* \right) \right\|_{\text{op}} \\
 &\quad + 2 \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\left\| \nabla_{\boldsymbol{\theta}} f \left(X; \boldsymbol{\theta}^* \right) \right\|_2^2 \right] \\
 &\quad + 2 \mathbb{E}_{X \sim \mathbb{Q}_X} \left[\left| f \left(X; \boldsymbol{\theta}^* \right) - \hat{f}_0(X) \right| \left\| \nabla_{\boldsymbol{\theta}}^2 f \left(X; \boldsymbol{\theta}^* \right) \right\|_{\text{op}} \right] \\
 &\stackrel{(c)}{\leq} 4 \left\{ b_1^2 + (1 + C_{\text{rf}}) b_2 \right\}
 \end{aligned} \tag{126}$$

for every $j \in [n_{\mathbb{Q}}]$, where the step (c) holds due to Assumptions 3 and 4.

- Using the upper bound (126), one can obtain $\left(\mathbf{V}_j^{\mathbb{Q}} \right)^2 \preceq \sigma_{\mathbb{Q}}^2 \mathbf{I}_d$ for every $j \in [n_{\mathbb{Q}}]$, where

$$\sigma_{\mathbb{Q}}^2 := 16 \left\{ b_1^2 + (1 + C_{\text{rf}}) b_2 \right\}^2.$$

Making use of the above findings regarding the $d \times d$ random matrices $\left\{ \mathbf{V}_j^{\mathbb{Q}} : j \in [n_{\mathbb{Q}}] \right\}$, the matrix Hoeffding inequality then reveals that

$$\begin{aligned}
 \left(\mathbb{Q}_X^{\otimes n_{\mathbb{Q}}} \right) \left(\left\{ \left\| \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \mathbf{V}_j^{\mathbb{Q}} \right\|_{\text{op}} > t \right\} \right) &\leq 2d \exp \left(-\frac{n_{\mathbb{Q}} t^2}{8 \sigma_{\mathbb{Q}}^2} \right) \\
 &= 2d \exp \left\{ -\frac{n_{\mathbb{Q}} t^2}{128 \left\{ b_1^2 + (1 + C_{\text{rf}}) b_2 \right\}^2} \right\}
 \end{aligned} \tag{127}$$

for every $t \in \mathbb{R}_+$. The inequality (127) tells us for any $\delta \in (0, 1]$ that

$$\left\| \frac{1}{n_{\mathbb{Q}}} \sum_{j=1}^{n_{\mathbb{Q}}} \mathbf{V}_j^{\mathbb{Q}} \right\|_{\text{op}} \leq 8\sqrt{2} \left\{ b_1^2 + (1 + C_{\text{rf}}) b_2 \right\} \sqrt{\frac{\log \left(\frac{2d}{\delta} \right)}{n_{\mathbb{Q}}}} \tag{128}$$

with probability at least $1 - \delta$ under the probability measure $\mathbb{Q}_X^{\otimes n_{\mathbb{Q}}}$. By combining two inequalities (125) and (128) together with the union bound and replacing δ by $\frac{\delta}{2}$ completes the proof of Lemma 13.