Maximum Likelihood Estimation is All You Need for Well-Specified Covariate Shift

Jiawei Ge[*†] **Shange Tang** [*†] **Jianqing Fan**[†] **Cong Ma**[‡] **Chi Jin**[§]

Abstract

A key challenge of modern machine learning systems is to achieve Out-of-Distribution (OOD) generalization—generalizing to target data whose distribution differs from that of source data. Despite its significant importance, the fundamental question of "what are the most effective algorithms for OOD generalization" remains open even under the standard setting of covariate shift. This paper addresses this fundamental question by proving that, surprisingly, classical Maximum Likelihood Estimation (MLE) purely using source data (without any modification) achieves the *minimax* optimality for covariate shift under the *well-specified* setting. This result holds for a very large class of parametric models, including but not limited to linear regression, logistic regression, and phase retrieval, and does not require any boundedness condition on the density ratio. This paper further complement the study by delving into the *misspecified setting* where the Maximum Weighted Likelihood Estimator (MWLE) emerges as minimax optimal in specific scenarios.

1 Introduction

Distribution shift, where the distribution of test data (target data) significantly differs from the distribution of training data (source data), is commonly encountered in practical machine learning scenarios (Zou et al., 2018; Ramponi & Plank, 2020; Guan & Liu, 2021). A central challenge of modern machine learning is to achieve Out-of-Distribution (OOD) generalization, where learned models maintain good performance in the target domain despite the presence of distribution shifts. To address this challenge, a variety of algorithms and techniques have been proposed, including vanilla empirical risk minimization (ERM) (Vapnik, 1999; Gulrajani & Lopez-Paz, 2020), importance weighting (Shimodaira, 2000; Huang et al., 2006; Cortes et al., 2010b; Cortes & Mohri, 2014), learning invariant representations (Ganin et al., 2016; Arjovsky et al., 2019; Wu et al., 2019; Rosenfeld et al., 2020), distributionally robust optimization (DRO) (Sagawa et al., 2019), etc. See the recent survey (Shen et al., 2021) for more details. These results claim the effectiveness of the corresponding proposed algorithms in different regimes. This leads to a natural question:

What are the most effective algorithms for OOD generalization?

In this paper, we answer the aforementioned question in a standard and fundamental setting of OOD generalization—covariate shift. Under covariate shift, the marginal distributions of the input covariates X vary between the source and target domains, while the conditional distribution of output given covariates $Y \mid X$ remains the same across domains. While covariate shift has been

[*] equal contribution

[†] Department of Operations Research and Financial Engineering, Princeton University; {jg5300,shangetang,jqfan}@princeton.edu

[‡] Department of Statistics, University of Chicago; congm@uchicago.edu

[§] Department of Electrical and Computer Engineering, Princeton University; chij@princeton.edu

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

extensively studied under the non-parametric setups recently (Kpotufe & Martinet, 2018; Pathak et al., 2022; Ma et al., 2023), what is the optimal algorithm for learning parametric models under covariate shift—an arguably more fundamental setup—remains open.

This paper considers learning a model from a known parametric model class under covariate shift. We discuss learning in both well-specified and mis-specified settings. Well-specified setting refers to the problems where the true conditional distribution of $Y \mid X$ lies in the given parametric model class, while the mis-specified setting refers to the remaining problems. This paper puts a special emphasis on the well-specified problems, as they become increasingly more relevant in modern learning applications. These applications use large-scale models with an enormous number of parameters, which are highly expressive and often make the settings "approximately" well-specified.

Our contributions. Below we summarize the results and contributions of this paper:

1. For a large set of well-specified covariate shift problems, we prove that the classical Maximum Likelihood Estimation (MLE) — which is computed purely based on source data without using any target data — is minimax optimal up to a universal constant. The optimality holds even if the learner has access to the distribution of covariate X in the target domain, and our results imply the minimax optimal algorithm need not to use that information. Our results suggest that additional techniques, such as importance reweighting, are not necessary for well-specified covariate shift.
2. The aforementioned result applies to a very rich set of well-specified covariate shift problems, and in particular does not require any boundedness condition on the density ratio. We illustrate the generality of our results by instantiating our framework and calculating the minimax optimal rates for three concrete examples: linear regression, logistic regression and phase retrieval.
3. We further establish the non-asymptotic upper bound for the Maximum Weighted Likelihood Estimator (MWLE) provided the bounded likelihood ratio condition is met. Under certain mis-specified problems, MWLE is shown to be minimax optimal.

1.1 Related work

The statistical study of covariate shift under parametric models can be dated back to Shimodaira (2000), which established the asymptotic normality of MWLE and pointed out that vanilla MLE is asymptotically optimal among all the weighted likelihood estimators when the model is well-specified. However, no finite sample guarantees were provided, and the optimality of MLE is only proved within the restricted class of weighted likelihood estimators. In contrast, this paper establishes non-asymptotic results and proves the optimality of MLE among all possible estimators under well-specified models. Cortes et al. (2010a) studied the importance weighting under the statistical learning framework and gave a non-asymptotic upper bound for the generalization error of the weighted estimator. However, their rate scales as $\mathcal{O}(1/\sqrt{n})$ compared to our rate $\mathcal{O}(1/n)$, where n is the sample size. A recent line of work also provide non-asymptotic analyses for covariate shift under well-specified setting, however they focus on linear regression or a few specific models which are more restrictive than our setting: Mousavi Kalan et al. (2020) introduces a statistical minimax framework and provides lower bounds for OOD generalization in the context of linear and one-hidden layer neural network regression models. When applied to covariate shift, their lower bounds are loose and no longer minimax optimal. Lei et al. (2021) considers the minimax optimal estimator for linear regression under fixed design, the estimator they proposed is not MLE and is much more complicated in certain regimes. Finally, Zhang et al. (2022) considers covariate shift in linear regression where the learner can have access to a small number of target labels, this is beyond the scope of this paper, where we focus on the classical covariate shift setup in which target labels are not known.

Another line of work focuses on well-specified nonparametric models under covariate shift. Kpotufe & Martinet (2018) presented minimax results for nonparametric classification problem, which was controlled by a transfer-exponent that measures the discrepancy between source and target. Inspired by the aforementioned work, Pathak et al. (2022) studied nonparametric regression problem over the class of Hölder continuous functions with a more fine-grained similarity measure. When considering reproducing kernel Hilbert space (RKHS), Ma et al. (2023) showed kernel ridge regression (KRR) estimator with a properly chosen penalty is minimax optimal for a large family of RKHS when the likelihood ratio is uniformly bounded, and a reweighted KRR using truncated likelihood ratios

is minimax optimal when the likelihood ratio has a finite second moment. Later, Wang (2023) proposed a learning strategy based on pseudo-labels. When the likelihood ratio is bounded, their estimator enjoyed the optimality guarantees without prior knowledge about the amount of covariate shift. Although these works focused on covariate shift problems, they considered nonparametric setting, and hence are not directly comparable to our work. As an example, Ma et al. (2023) showed that MLE (empirical risk minimization in their language) is provably suboptimal for addressing covariate shift under nonparametric RKHS assumptions. In contrast, we show that MLE is optimal for covariate shift for a well-specified parametric model. We highlight that our non-asymptotic rate is instance dependent in the sense that it depends on the source and target distributions. This is in contrast to prior work (e.g. Ma et al. (2023), Kpotufe & Martinet (2018)) that consider the worst-case scenario over certain classes of source-target pairs (e.g., bounded density ratios).

A crucial part of this work is analyzing MLE, which is a dominant approach in statistical inference. There exists a variety of work studying the behavior of MLE under the standard no-distribution-shift setting. It is well known that MLE is asymptotically normal (Casella & Berger, 2021) with the inverse of Fisher information as the asymptotic variance. Cramér (1946); Rao (1992) established the famous Cramer-Rao bound for unbiased estimators, which also showed that no consistent estimator has lower asymptotic mean squared error than the MLE. White (1982) gave the asymptotic distribution of MLE under the mis-specified setting. More recently, non-asymptotic behaviours of MLE are studied under certain models. Bach (2010); Ostrovskii & Bach (2021) established the non-asymptotic error bound for MLE in logistic regression using self-concordance. This line of work does not consider covariate shift, which is an indispensable part of this paper.

Notations. Throughout the paper, we use c to denote universal constants, which may vary from line to line.

2 Background and Problem Formulation

In this section, we provide background on the problem of learning under covariate shift. We also review the widely adopted maximum likelihood estimator.

2.1 Covariate shift and excess risk

Let $X \in \mathcal{X}$ be the covariates and $Y \in \mathcal{Y}$ be the response variable that we aim to predict. In a general out-of-distribution (OOD) generalization problem, we have two domains of interest, namely a source domain S and a target domain T. Each domain is associated with a data generating distribution over (X, Y): $\mathbb{P}_S(X, Y)$ for the source domain and $\mathbb{P}_T(X, Y)$ for the target domain. Given n i.i.d. labeled samples $\{(x_i, y_i)\}_{i=1}^n \sim \mathbb{P}_S(X, Y)$ from the source domain, the goal of OOD generalization is to learn a prediction rule $X \to Y$ that performs well in the target domain. In this paper, we focus on the covariate shift version of the OOD generalization problem, in which the marginal distributions $\mathbb{P}_S(X)$ and $\mathbb{P}_T(X)$ of the covariates could differ between the source and target domains, while the conditional distribution $Y \mid X$ is assumed to be the same on both domains.

More precisely, we adopt the notion of excess risk to measure of the performance of an estimator under covariate shift. Let $\mathcal{F} := \{f(y \mid x; \beta) \mid \beta \in \mathbb{R}^d\}$ be a parameterized function class to model the conditional density function $p(y \mid x)$ of $Y \mid X$. A typical loss function is defined using the negative log-likelihood function:

$$\ell(x, y, \beta) := -\log f(y \mid x; \beta).$$

The excess risk at β is then defined as

$$R(\beta) := \mathbb{E}_T\left[\ell(x, y, \beta)\right] - \inf_{\beta} \mathbb{E}_T\left[\ell(x, y, \beta)\right], \tag{1}$$

where the expectation $\mathbb{E}_T$ is taken over $\mathbb{P}_T(X, Y)$. When the model is well-specified, i.e., when the true density $p(y \mid x) = f(y \mid x; \beta^\star)$ for some $\beta^\star$, we have $\inf_\beta \mathbb{E}_T[\ell(x, y, \beta)] = \mathbb{E}_T[\ell(x, y, \beta^\star)]$. As a result, we evaluate the loss at β against the loss at the true parameter $\beta^\star$. In contrast, in the case of mis-specification, i.e., when $p(y \mid x) \notin \mathcal{F}$, the loss at β is compared against the loss of the best fit in the model class.

2.2 Maximum likelihood estimation

In the no-covariate-shift case, maximum likelihood estimation (MLE) is arguably the most popular approach. Let

$$\ell_n(\beta) := \frac{1}{n}\sum_{i=1}^n \ell(x_i, y_i, \beta) \tag{2}$$

be the empirical negative log-likelihood using the samples $\{(x_i, y_i)\}_{i=1}^n$ from the source domain. The vanilla MLE is defined as

$$\beta_{\mathsf{MLE}} := \arg\min_{\beta\in\mathbb{R}^d} \ell_n(\beta). \tag{3}$$

To ease presentations later, we would also recall the classical notion of Fisher information—an important quantity to measure the difficulty of parameter estimation. The Fisher information evaluated at β on source and target is defined as

$$\mathcal{I}_S(\beta) := \mathbb{E}_{x\sim\mathbb{P}_S(X), y\,|\,x\sim f(y\,|\,x;\beta)}\left[\nabla^2\ell(x,y,\beta)\right],$$
$$\mathcal{I}_T(\beta) := \mathbb{E}_{x\sim\mathbb{P}_T(X), y\,|\,x\sim f(y\,|\,x;\beta)}\left[\nabla^2\ell(x,y,\beta)\right].$$

Here, the gradient and Hessian are taken with respect to the parameter β.

3 Well-Specified Parametric Model under Covariate Shift

In this section, we focus on covariate shift with a well-specified model, that is, the true conditional distribution falls in our parametric function class. We assume there exists some $\beta^\star$ such that $p(y\,|\,x) = f(y\,|\,x;\beta^\star)$, and denote the excess risk evaluated at β under true model parameter $\beta^\star$ as $R_{\beta^\star}(\beta)$, i.e.,

$$R_{\beta^\star}(\beta) := \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X)\\ y|x\sim f(y|x;\beta^\star)}}[\ell(x,y,\beta)] - \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X)\\ y|x\sim f(y|x;\beta^\star)}}[\ell(x,y,\beta^\star)]. \tag{4}$$

While the objective of MLE (cf. (3)) is not an unbiased estimate of the risk under the target domain, we will show in this section that MLE is in fact optimal for addressing covariate shift under well-specified models.

3.1 Upper bound for MLE

In this subsection, we establish a non-asymptotic upper bound for MLE under generic assumptions on the model class.

Assumption A. *We make the following assumptions on the model class $\mathcal{F}$:*

A.1 There exist B_1, B_2, $N(\delta)$, and absolute constants c, γ such that for any fixed matrix $A \in \mathbb{R}^{d\times d}$, any $\delta \in (0,1)$, and any $n > N(\delta)$, with probability at least $1-\delta$:

$$\|A\left(\nabla\ell_n(\beta^\star) - \mathbb{E}[\nabla\ell_n(\beta^\star)]\right)\|_2 \le c\sqrt{\frac{V\log\frac{d}{\delta}}{n}} + B_1\|A\|_2\log^\gamma\left(\frac{B_1\|A\|_2}{\sqrt{V}}\right)\frac{\log\frac{d}{\delta}}{n}, \tag{5}$$

$$\left\|\nabla^2\ell_n(\beta^\star) - \mathbb{E}[\nabla^2\ell_n(\beta^\star)]\right\|_2 \le B_2\sqrt{\frac{\log\frac{d}{\delta}}{n}}, \tag{6}$$

where $V = n\cdot\mathbb{E}\|A(\nabla\ell_n(\beta^\star) - \mathbb{E}[\nabla\ell_n(\beta^\star)])\|_2^2$ is the variance.

A.2 There exists some constant $B_3 \ge 0$ such that $\|\nabla^3\ell(x,y,\beta)\|_2 \le B_3$ for all $x \in \mathcal{X}_S \cup \mathcal{X}_T, y \in \mathcal{Y}, \beta \in \mathbb{R}^d$, where $\mathcal{X}_S$ (resp. $\mathcal{X}_T$) is the support of $\mathbb{P}_S(X)$ (resp. $\mathbb{P}_T(X)$).

A.3 The empirical loss $\ell_n(\cdot)$ defined in (2) has a unique local minimum in $\mathbb{R}^d$, which is also the global minimum.

Several remarks on Assumption A are in order. Assumption A.1 is a general version of Bernstein inequality (when $\gamma = 0$ it reduces to classical Bernstein inequality), which gives concentration on gradient and Hessian. This assumption is naturally satisfied when the gradient and Hessian are

bounded (see Proposition F.2 for details). Assumption A.2 requires the third order derivative of log-likelihood to be bounded, which is easy to satisfy (e.g., linear regression satisfies this assumption with $B_3 = 0$). Assumption A.3 ensures the MLE is unique, which is standard in the study of the behaviour of MLE. We can see that it naturally applies to traditional convex losses. It is worth noting that our general theorem can also be applied under a relaxed version of Assumption A.3, which will be shown in Theorem A.5. In Section A, we will see that Assumption A is mild and easily satisfied for a wide range of models.

Now we are ready to present the performance upper bound for MLE under covariate shift.

Theorem 3.1. *Suppose that the model class $\mathcal{F}$ satisfies Assumption A. Let $\mathcal{I}_T := \mathcal{I}_T(\beta^\star)$ and $\mathcal{I}_S := \mathcal{I}_S(\beta^\star)$. For any $\delta \in (0,1)$, if $n \geq c\max\{N^\star \log(d/\delta), N(\delta)\}$, then with probability at least $1-2\delta$, we have*

$$R_{\beta^\star}(\beta_{\mathsf{MLE}}) \leq c\frac{\mathsf{Tr}\left(\mathcal{I}_T\mathcal{I}_S^{-1}\right)\log\frac{d}{\delta}}{n}$$

for an absolute constant c. Here $N^\star := Poly(d, B_1, B_2, B_3, \|\mathcal{I}_S^{-1}\|_2, \|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-1}\mathcal{I}_T^{\frac{1}{2}}\|_2^{-1})$.

For an exact characterization of the threshold $N^\star$, one can refer to Theorem C.1 in the appendix.

Theorem 3.1 gives a non-asymptotic upper bound for the excess risk of MLE: when the sample size exceeds a certain threshold of $\max\{N^\star \log(d/\delta), N(\delta)\}$, MLE achieves an instance dependent risk bound $\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})/n$. It is worth noting that our analysis does not require boundedness on the density ratios between the target and source distributions (as have been assumed in prior art (Ma et al., 2023)), which yields broader applicability. In Section A, we will instantiate our generic analysis on three different examples: linear regression, logistic regression and phase retrieval.

3.2 Minimax lower bound

We turn to the complementary question regarding the fundamental limit of covariate shift under well-specified models.

Assumption B. *Let $\beta_0 \in \mathbb{R}^d$ and $B > 0$. We make the following assumptions on the model class $\mathcal{F}$:*

B.1 Assumption A.2 holds.

B.2 There exist some constants $L_S, L_T \geq 0$ such that for any $\beta_1, \beta_2 \in \mathbb{B}_{\beta_0}(B)$:

$$\|\mathcal{I}_S(\beta_1) - \mathcal{I}_S(\beta_2)\|_2 \leq L_S\|\beta_1 - \beta_2\|_2,$$
$$\|\mathcal{I}_T(\beta_1) - \mathcal{I}_T(\beta_2)\|_2 \leq L_T\|\beta_1 - \beta_2\|_2.$$

B.3 For any $\beta^\star \in \mathbb{B}_{\beta_0}(B)$, the excess risk $R_{\beta^\star}(\beta)$ defined in (4) is convex in $\beta \in \mathbb{R}^d$.

B.4 We assume $\mathcal{I}_S(\beta)$ and $\mathcal{I}_T(\beta)$ are positive definite for all $\beta \in \mathbb{B}_{\beta_0}(B)$.

Assumption B.2 essentially requires the Fisher information will not vary drastically in a small neighbourhood of β_0. This assumption is easy to hold when the fisher information has certain smoothness (e.g., in linear regression, the fisher information does not change when β varies). Since Assumption B is a slight variant of Assumption A, both assumptions are often satisfied simultaneously for a wide range of models, as we will show in Section A.

Theorem 3.2. *Suppose the model class $\mathcal{F}$ satisfies Assumption B. As long as $n \geq N_0$, we have*

$$\inf_{\hat{\beta}} \sup_{\beta^\star \in \mathbb{B}_{\beta_0}(B)} \mathsf{Tr}\left(\mathcal{I}_T(\beta^\star)\mathcal{I}_S^{-1}(\beta^\star)\right)^{-1} \mathbb{E}_{\substack{x_i \sim \mathbb{P}_S(X) \\ y_i|x_i \sim f(y|x;\beta^\star)}} \left[R_{\beta^\star}(\hat{\beta})\right] \geq \frac{1}{50n},$$

where $N_0 := Poly(d, B^{-1}, B_3, L_S, L_T, \|\mathcal{I}_S(\beta_0)\|_2, \|\mathcal{I}_T(\beta_0)\|_2, \|\mathcal{I}_S(\beta_0)^{-1}\|_2, \|\mathcal{I}_T(\beta_0)^{-1}\|_2)$.

For an exact characterization of the threshold N_0, one can refer to Theorem C.4 in the appendix.

Comparing Theorem 3.1 and 3.2, we can see that, under Assumptions A and B, then for large enough sample size n, $\mathsf{Tr}\left(\mathcal{I}_T(\beta^\star)\mathcal{I}_S^{-1}(\beta^\star)\right)/n$ exactly characterizes the fundamental hardness of covariate shift under well-specified parametric models. It also reveals that vanilla MLE is minimax optimal under this scenario. To gain some intuitions, $\mathcal{I}_S^{-1}$ captures the variance of the parameter estimation, and $\mathcal{I}_T$ measures how the excess risk on the target depends on the estimation accuracy of the parameter. Therefore what really affects the excess risk (on target) is the accuracy of estimating the parameter, and vanilla MLE is naturally the most efficient choice.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4(none):384 – 414, 2010. doi: 10.1214/09-EJS521. URL https://doi.org/10.1214/09-EJS521.

Liyuan Cao. Some useful expected values with multivariate normal distribution and uniform distribution on sphere. https://coral.ise.lehigh.edu/lic314/files/2020/02/MVNuseful.pdf, February 2020.

George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2021.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010a. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/59c33016884a62116be975a9bb8257e3-Paper.pdf.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010b.

Harald Cramér. Mathematical methods of statistics. 1946. URL https://api.semanticscholar.org/CorpusID:122802041.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pp. 1233–1242. PMLR, 2017.

Richard D Gill and Boris Y Levit. Applications of the van trees inequality: a bayesian cramér-rao bound. *Bernoulli*, pp. 59–79, 1995.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. 2011.

Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference On Learning Theory*, pp. 1882–1886. PMLR, 2018.

Qi Lei, Wei Hu, and Jason D Lee. Near-optimal linear regression under distribution shift. *arXiv preprint arXiv:2106.12108*, 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.

Dmitrii M Ostrovskii and Francis Bach. Finite-sample analysis of m-estimators using self-concordance. 2021.

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*, pp. 17517–17530. PMLR, 2022.

Alan Ramponi and Barbara Plank. Neural unsupervised domain adaptation in nlp—a survey. *arXiv preprint arXiv:2006.00632*, 2020.

C. Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. 1992. URL https://api.semanticscholar.org/CorpusID:117034671.

Alessandro Rinaldo. Advanced statistical theory lecture 13: February 26, February 2018.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

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. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(00)00115-4. URL https://www.sciencedirect.com/science/article/pii/S0378375800001154.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Kaizheng Wang. Pseudo-labeling for kernel ridge regression under covariate shift. *arXiv preprint arXiv:2302.10160*, 2023.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pp. 1–25, 1982.

Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International conference on machine learning*, pp. 6872–6881. PMLR, 2019.

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*, pp. 3794–3820. PMLR, 2022.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 289–305, 2018.

A Applications

In this section, we illustrate the broad applicability of our framework by delving into three distinct statistical models, namely linear regression, logistic regression and phase retrieval. For each model, we will demonstrate the validity of the assumptions, and give the explicit non-asymptotic upper bound on the vanilla MLE obtained by our framework as well as the threshold of sample size needed to obtain the upper bound.

A.1 Linear regression

In linear regression, we have $Y = X^T\beta^\star + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,1)$ and $\varepsilon \perp\!\!\!\perp X$. The corresponding negative log-likelihood function (i.e. the loss function) is given by

$$\ell(x, y, \beta) := \frac{1}{2}(y - x^T\beta)^2.$$

We assume $X \sim \mathcal{N}(0, I_d)$ on the source domain and $X \sim \mathcal{N}(\alpha, \sigma^2 I_d)$ on the target domain.

Proposition A.1. *The aforementioned linear regression model satisfies Assumption A and B with $\gamma = 1$, $N(\delta) = d\log(1/\delta)$, $B_1 = c\sqrt{d}$, $B_2 = c\sqrt{d}$, $B_3 = 0$ and $L_S = L_T = 0$. Moreover, we have $\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1}) = \|\alpha\|_2^2 + \sigma^2 d$.*

By Theorem 3.1 and Theorem 3.2, since Assumption A and B are satisfied, we immediately demonstrate the optimality of MLE under linear regression. The following theorem gives the explicit form of excess risk bound by applying Theorem 3.1:

Theorem A.2. *For any $\delta \in (0,1)$, if $n \geq \mathcal{O}(N \log \frac{d}{\delta})$, then with probability at least $1-2\delta$, we have*

$$R_{\beta^\star}(\beta_{\mathsf{MLE}}) \leq c\frac{\left(\|\alpha\|_2^2 + \sigma^2 d\right)\log\frac{d}{\delta}}{n},$$

where $N := d\left(1 + \frac{\|\alpha\|_2^2 d + \sigma^2 d}{\|\alpha\|_2^2 + \sigma^2 d}\right)^2$.

Remark (Excess risk). Regarding the upper bound of the excess risk, we categorize it into two scenarios: large shift and small shift. In the small shift scenarios (i.e., $\|\alpha\|_2^2 \leq \sigma^2 d$), the result is the same as that in scenarios without any mean shift, with a rate of $\sigma^2 d/n$. On the other hand, in the large shift scenarios (i.e., $\|\alpha\|_2^2 \geq \sigma^2 d$), the upper bound of the excess risk increases with the mean shift at a rate of $\|\alpha\|_2^2/n$.

Remark (Threshold N). For a minor mean shift, specifically when $\|\alpha\|_2 = c\sigma$ for a given constant c, the threshold is $N = d$. This aligns with the results from linear regression without any covariate shift. On the other hand, as the mean shift increases (i.e., $|\alpha|_2 = \sigma d^k$ for some $0 < k < 1/2$), the threshold becomes $N = d^{4k+1}$, increasing with the growth of k. In scenarios where the mean shift significantly surpasses the scaling shift, denoted as $\alpha \geq \sigma\sqrt{d}$, the threshold reaches $N = d^3$.

A.2 Logistic regression

In the logistic regression, the response variable $Y \in \{0,1\}$ obeys

$$\mathbb{P}(Y = 1 \,|\, X = x) = \frac{1}{1 + e^{x^T\beta^\star}},\ \mathbb{P}(Y = 0 \,|\, X = x) = \frac{1}{1 + e^{-x^T\beta^\star}}.$$

The corresponding negative log-likelihood function (i.e. the loss function) is given by

$$\ell(x, y, \beta) := \log(1 + e^{x^T\beta}) - y(x^T\beta).$$

We assume $X \sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))$ on the source domain and $X \sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d})) + v$ on the target domain, where $\mathcal{S}^{d-1}(\sqrt{d}) := \{x \in \mathbb{R}^d \mid \|x\|_2 = \sqrt{d}\}$. In the following, we will give the upper bound of the excess risk for MLE when $v = r\beta_\perp^\star$, where $\beta_\perp^\star$ represents a vector perpendicular to $\beta^\star$ (i.e., $\beta_\perp^{\star T}\beta^\star$=0). Without loss of generality, we assume $\|\beta^\star\|_2 = \|\beta_\perp^\star\|_2 = 1$.

Proposition A.3. *The aforementioned logistic regression model satisfies Assumption A and B with $\gamma = 0$, $N(\delta) = 0$, $B_1 = c\sqrt{d}$, $B_2 = cd$, $B_3 = (\sqrt{d} + r)^3$, $L_S = d^{1.5}$ and $L_T = (\sqrt{d} + r)^3$. Moreover, we have $\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1}) \asymp d + r^2$.*

By Theorem 3.1 and Theorem 3.2, since Assumption A and B are satisfied, we immediately demonstrate the optimality of MLE under logistic regression. The following theorem gives the explicit form of excess risk bound by applying Theorem 3.1:

Theorem A.4. *For any $\delta \in (0,1)$, if $n \geq \mathcal{O}(N \log \frac{d}{\delta})$, then with probability at least $1-2\delta$, we have*

$$R_{\beta^\star}(\beta_{\mathsf{MLE}}) \leq c\frac{(d+r^2)\log\frac{d}{\delta}}{n},$$

where $N := d^4(1+r^6)$.

Remark (Excess risk). The bound on the excess risk incorporates a r^2 term, which is a measurement of the mean shift. This is due to the fact that the MLE does not utilize the information that $v^T\beta^\star = 0$. Therefore, $v^T\beta_{\mathsf{MLE}}$ is not necessarily zero, which will lead to an additional bias. Similar to linear regression, we can categorize the upper bound of the excess risk into two scenarios: large shift ($r \geq \sqrt{d}$) and small shift ($r \leq \sqrt{d}$).

Remark (Threshold N). We admit that the N here may not be tight, as we lean on a general framework designed for a variety of models rather than a specific one.

A.3 Phase retrieval

As we have mentioned, our generic framework can also be applied to the scenarios where some of the assumptions are relaxed. In this subsection, we will further illustrate this point by delving into the phase retrieval model.

In the phase retrieval, the response variable $Y = (X^T\beta^\star)^2 + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,1)$ and $\varepsilon \perp\!\!\!\perp X$. We assume $\mathbb{P}_S(X)$ and $\mathbb{P}_T(X)$ follow the same distribution as that in the logistic regression model (i.e., Section A.2). Note that both the phase retrieval model and the logistic regression model belong to generalized linear model (GLM), thus they are expected to have similar properties. However, given the loss function $\ell(x,y,\beta) := \frac{1}{2}\left(y - (x^T\beta)^2\right)^2$, it is obvious that Assumption A.3 is not satisfied, since if β is a global minimum of ℓ_n, $-\beta$ is also a global minimum. The following theorem shows that we can still obtain results similar to logistic regression though Assumption A.3 fails to hold.

Theorem A.5. *For any $\delta \in (0,1)$, if $n \geq \mathcal{O}(N \log \frac{d}{\delta})$, then with probability at least $1-2\delta$, we have*

$$R_{\beta^\star}(\beta_{\mathsf{MLE}}) \leq c\frac{(d+r^2)\log\frac{d}{\delta}}{n},$$

where $N := d^8(1+r^8)$.

B Mis-Specified Parametric Model under Covariate Shift

One potential "criticism" against MLE in the covariate shift setting is that the empirical negative log-likelihood is not a faithful estimate of the out-of-distribution generalization performance, i.e., $\mathbb{E}_T[\ell(x,y,\beta)]$. In light of this, a weighted version of MLE is proposed. Let $w(x) := d\mathbb{P}_T(x)/d\mathbb{P}_S(x)$ be the density ratio function and

$$\ell_n^w(\beta) := \frac{1}{n}\sum_{i=1}^n w(x_i)\ell(x_i,y_i,\beta) \tag{7}$$

be the weighed loss. Then the maximum weighted likelihood estimator is defined as

$$\beta_{\mathsf{MWLE}} := \arg\min_{\beta\in\mathbb{R}^d} \ell_n^w(\beta). \tag{8}$$

It is easy to see that the weighted loss is an unbiased estimate of $\mathbb{E}_T[\ell(x,y,\beta)]$.

In the case of model mis-specification, we still employ a parameterized function class $\mathcal{F} := \{f(y\,|\,x;\beta)\,|\,\beta\in\mathbb{R}^d\}$ to model the conditional density function of $Y\,|\,X$. However, the true density $p(y\,|\,x)$ might not be in $\mathcal{F}$. As we previously showed, under a well-specified parametric model, the vanilla MLE is minimax optimal up to constants. However, when the model is mis-specified, the classical MLE may not necessarily provide a good estimator.

Proposition B.1. *There exist certain mis-specified scenarios such that classical MLE is not consistent, whereas MWLE is.*

Proposition B.1 illustrates the necessity of adaptation under model mis-specification since the classical MLE asymptotically gives the wrong estimator. In this section, we study the non-asymptotic property of MWLE. Let $\mathcal{M}$ be the model class of the ground truth $Y \mid X$, and $M \in \mathcal{M}$ be the ground truth model for $Y \mid X$.

We denote the optimal fit on target as

$$\beta^\star(M) := \arg\min_\beta \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X)\\ y|x\sim M}}[\ell(x,y,\beta)].$$

The excess risk evaluated at β is then given by

$$R_M(\beta) = \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X)\\ y|x\sim M}}\left[\ell(x,y,\beta)\right] - \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X)\\ y|x\sim M}}\left[\ell(x,y,\beta^\star(M))\right]. \tag{9}$$

B.1 Upper bound for MWLE

In this subsection, we establish the non-asymptotic upper bound for MWLE, as an analog to Theorem 3.1. We make the following assumption which is a modification of Assumption A.

Assumption C. *We assume the function class $\mathcal{F}$ satisfies the follows:*

C.1 There exists some constant $W > 1$ such that the density ratio $w(x) \le W$ for all $x \in \mathcal{X}_S \cup \mathcal{X}_T$.

C.2 There exist B_1, B_2 and $N(\delta)$, and absolute constants c, γ such that for any fixed matrix $A \in \mathbb{R}^{d\times d}$, any $\delta \in (0,1)$, and any $n > N(\delta)$, with probability at least $1-\delta$:

$$\|A\left(\nabla\ell_n^w(\beta^\star(M)) - \mathbb{E}[\nabla\ell_n^w(\beta^\star(M))]\right)\|_2 \le c\sqrt{\frac{V\log\frac{d}{\delta}}{n}} + WB_1\|A\|_2\log^\gamma\left(\frac{WB_1\|A\|_2}{\sqrt{V}}\right)\frac{\log\frac{d}{\delta}}{n},$$

$$\left\|\nabla^2\ell_n^w(\beta^\star(M)) - \mathbb{E}[\nabla^2\ell_n^w(\beta^\star(M))]\right\|_2 \le WB_2\sqrt{\frac{\log\frac{d}{\delta}}{n}},$$

where $V = n\cdot\mathbb{E}\|A(\nabla\ell_n^w(\beta^\star(M)) - \mathbb{E}[\nabla\ell_n^w(\beta^\star(M))])\|_2^2$ is the variance.

C.3 Assumption A.2 holds.

C.4 There exists $N'(\delta)$ such that for any $\delta \in (0,1)$ and any $n \ge N'(\delta)$, with probability at least $1-\delta$, the empirical loss $\ell_n^w(\cdot)$ defined in (7) has a unique local minimum in $\mathbb{R}^d$, which is also the global minimum.

Assumption C.1 is a density ratio upper bound (not required for analyzing MLE), which is essential for the analysis of MWLE. Assumption C.2 is an analog of Assumption A.1, in the sense that the empirical loss ℓ_n is replaced by its weighted version ℓ_n^w. Assumption C.4 is a weaker version of Assumption A.3 in the sense that it only requires ℓ_n^w has a unique local minimum with high probability. This is due to the nature of reweighting: when applying MWLE, $w(x_i)$ can sometimes be zero, which lead to the degeneration of ℓ_n^w (with a small probability). Therefore we only require the uniqueness of local minimum holds with high probability.

To state our non-asymptotic upper bound for MWLE, we define the following "weighted version" of Fisher information:

$$G_w(M) := \mathbb{E}_{\substack{x\sim\mathbb{P}_S(X)\\ y|x\sim M}}\left[w(x)^2\nabla\ell(x,y,\beta^\star(M))\nabla\ell(x,y,\beta^\star(M))^T\right],$$

$$H_w(M) := \mathbb{E}_{\substack{x\sim\mathbb{P}_S(X)\\ y|x\sim M}}\left[w(x)\nabla^2\ell(x,y,\beta^\star(M))\right] = \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X)\\ y|x\sim M}}\left[\nabla^2\ell(x,y,\beta^\star(M))\right].$$

Theorem B.2. *Suppose the function class $\mathcal{F}$ satisfies Assumption C. Let $G_w := G_w(M)$ and $H_w := H_w(M)$. For any $\delta \in (0,1)$, if $n \ge c\max\{N^\star\log(d/\delta), N(\delta), N'(\delta)\}$, then with probability at least $1-3\delta$, we have*

$$R_M(\beta_{\mathsf{MWLE}}) \le c\frac{\mathsf{Tr}\left(G_wH_w^{-1}\right)\log\frac{d}{\delta}}{n}$$

for an absolute constant c. Here $N^\star := \mathit{Poly}(W, B_1, B_2, B_3, \|H_w^{-1}\|_2, \mathsf{Tr}(G_wH_w^{-2}), \mathsf{Tr}(G_wH_w^{-2})^{-1})$.

For an exact characterization of the threshold $N^\star$, one can refer to Theorem E.1 in the appendix.

Compared with Theorem 3.1, Theorem B.2 does not require well-specification of the model, demonstrating the wide applicability of MWLE. The excess risk upper bound can be explained as follows: note that $\mathsf{Tr}(G_w H_w^{-1})$ can be expanded as $\mathsf{Tr}(H_w H_w^{-1} G_w H_w^{-1})$. As shown by Shimodaira (2000), the term $\sqrt{n}(\beta_{\mathsf{MWLE}} - \beta^\star)$ converges asymptotically to a normal distribution, denoted as $\mathcal{N}(0, H_w^{-1} G_w H_w^{-1})$. Thus, the component $H_w^{-1} G_w H_w^{-1}$ characterizes the variance of the estimator, corresponding to the $\mathcal{I}_S^{-1}$ term in Theorem 3.1. Additionally, the excess risk's dependence on the parameter estimation is captured by H_w as a counterpart of $\mathcal{I}_T$ in Theorem 3.1.

However, to establish Theorem B.2, it is necessary to assume the bounded density ratio, which does not appear in Theorem 3.1. Moreover, when the model is well-specified, by Cauchy-Schwarz ineqaulity, we have $\mathsf{Tr}(G_w H_w^{-1}) \geq \mathsf{Tr}(\mathcal{I}_T \mathcal{I}_S^{-1})$, which implies the upper bound for MWLE is larger than the vanilla MLE. This observation aligns with the results presented in Shimodaira (2000), which point out that when the model is well specified, MLE is more efficient than MWLE in terms of the asymptotic variance.

B.2 Optimality of MWLE

To understand the optimality of MWLE, it is necessary to establish a matching lower bound. However, deriving a lower bound similar to Theorem 3.2, which holds for any model classes that satisfies certain mild conditions, is challenging due to hardness of capturing the difference between $\mathcal{M}$ and $\mathcal{F}$. As a solution, we present a lower bound tailored for certain model classes and data distributions in the following.

Theorem B.3. *There exist $\mathbb{P}_S(X) \neq \mathbb{P}_T(X)$, a model class $\mathcal{M}$ and a prediction class $\mathcal{F}$ satisfying Assumption C such that when n is sufficiently large, we have*

$$\inf_{\hat{\beta}} \sup_{M \in \mathcal{M}} \mathsf{Tr}\left(G_w(M) H_w^{-1}(M)\right)^{-1} \mathbb{E}_{\substack{x_i \sim \mathbb{P}_S(X) \\ y_i | x_i \sim M}} \left[R_M(\hat{\beta})\right] \gtrsim \frac{1}{n}. \tag{10}$$

By Theorem B.2, the excess risk of MWLE is upper bounded by $\mathsf{Tr}(G_w H_w^{-1})/n$. Therefore, Theorem B.3 shows that there exists a non-trivial scenario where MWLE is minimax optimal.

Notice that Theorem B.3 presents a weaker lower bound compared to Theorem 3.2. The lower bound presented in Theorem B.3 holds only for certain meticulously chosen $\mathbb{P}_S(X), \mathbb{P}_T(X)$, model class $\mathcal{M}$ and prediction class $\mathcal{F}$. In contrast, the lower bound in Theorem 3.2 applies to any $\mathbb{P}_S(X), \mathbb{P}_T(X)$, and class $\mathcal{F}$ that meet the required assumptions.

C Proofs for Section 3

C.1 Proofs for Theorem 3.1

The detailed version of Theorem 3.1 is stated as the following.

Theorem C.1. *Suppose that the model class $\mathcal{F}$ satisfies Assumption A. Let $\mathcal{I}_T := \mathcal{I}_T(\beta^\star)$ and $\mathcal{I}_S := \mathcal{I}_S(\beta^\star)$. For any $\delta \in (0,1)$, if $n \geq c \max\{N^\star \log(d/\delta), N(\delta)\}$, then with probability at least $1 - 2\delta$, we have*

$$R_{\beta^\star}(\beta_{\mathsf{MLE}}) \leq c \frac{\mathsf{Tr}\left(\mathcal{I}_T \mathcal{I}_S^{-1}\right) \log \frac{d}{\delta}}{n}$$

for an absolute constant c. Here

$$N^\star := (1 + \tilde{\kappa}/\kappa)^2 \cdot \max\left\{\tilde{\kappa}^{-1} \alpha_1^2 \log^{2\gamma}\left((1 + \tilde{\kappa}/\kappa)\tilde{\kappa}^{-1}\alpha_1^2\right),\ \alpha_2^2,\ \tilde{\kappa}(1 + \|\mathcal{I}_T^{\frac{1}{2}} \mathcal{I}_S^{-1} \mathcal{I}_T^{\frac{1}{2}}\|_2^{-2}) \alpha_3^2\right\},$$

where $\alpha_1 := B_1 \|\mathcal{I}_S^{-1}\|_2^{1/2}$, $\alpha_2 := B_2 \|\mathcal{I}_S^{-1}\|_2$, $\alpha_3 := B_3 \|\mathcal{I}_S^{-1}\|_2^{3/2}$,

$$\kappa := \frac{\mathsf{Tr}(\mathcal{I}_T \mathcal{I}_S^{-1})}{\|\mathcal{I}_T^{\frac{1}{2}} \mathcal{I}_S^{-1} \mathcal{I}_T^{\frac{1}{2}}\|_2}, \quad \tilde{\kappa} := \frac{\mathsf{Tr}(\mathcal{I}_S^{-1})}{\|\mathcal{I}_S^{-1}\|_2}.$$

For proving Theorem C.1, we first state two main lemmas. Informally speaking, Lemma C.2 and Lemma C.3 capture the distance between β_{MLE} and $\beta^\star$ under different measurements.

Lemma C.2. *Suppose Assumption A holds. For any $\delta \in (0,1)$ and any $n \geq c\max\{N_1\log(d/\delta), N(\delta)\}$, with probability at least $1-\delta$, we have $\beta_{\mathsf{MLE}} \in \mathbb{B}_{\beta^\star}(c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}})$ for some absolute constant c. Here*

$$N_1 := \max\Bigg\{B_2^2\|\mathcal{I}_S^{-1}\|_2^2, B_3^2\|\mathcal{I}_S^{-1}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1}), \left(\frac{B_1^2B_2\|\mathcal{I}_S^{-1}\|_2^3\log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_S^{-1})}\right)^{\frac{2}{3}}, \left(\frac{B_1^3B_3\|\mathcal{I}_S^{-1}\|_2^4\log^{3\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_S^{-1})}\right)^{\frac{1}{2}}, \frac{B_1^2\|\mathcal{I}_S^{-1}\|_2^2\log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_S^{-1})}\Bigg\}.$$

Lemma C.3. *Suppose Assumption A holds. For any $\delta \in (0,1)$ and any $n \geq c\max\{N_1\log(d/\delta), N_2\log(d/\delta), N(\delta)\}$, with probability at least $1-2\delta$, we have*

$$\|\mathcal{I}_T^{\frac{1}{2}}(\beta_{\mathsf{MLE}}-\beta^\star)\|_2^2 \leq c\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}$$

for some absolute constant c. Here N_1 is defined in Lemma C.2 and

$$N_2 := \max\Bigg\{\left(\frac{B_2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2, \left(\frac{B_3\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2, \left(\frac{B_1^2B_2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^2\log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^{\frac{2}{3}}, \left(\frac{B_1^3B_3\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^3\log^{3\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^{\frac{1}{2}}, \frac{B_1^2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2\log^{2\gamma}(\kappa^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\Bigg\}.$$

The proofs for Lemma C.2 and C.3 are delayed to the end of this subsection. With these two lemmas, we can now state the proof for Theorem C.1.

Proof of Theorem C.1. By Assumption A.2, we can do Taylor expansion w.r.t. β as the following:

$$\begin{aligned}R_{\beta^\star}(\beta_{\mathsf{MLE}}) &= \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X)\\ y|x\sim f(y|x;\beta^\star)}}[\ell(x,y,\beta_{\mathsf{MLE}})-\ell(x,y,\beta^\star)]\\ &\leq \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X)\\ y|x\sim f(y|x;\beta^\star)}}[\nabla\ell(x,y,\beta^\star)]^T(\beta_{\mathsf{MLE}}-\beta^\star)\\ &\quad+\frac{1}{2}(\beta_{\mathsf{MLE}}-\beta^\star)^T\mathcal{I}_T(\beta_{\mathsf{MLE}}-\beta^\star)+\frac{B_3}{6}\|\beta_{\mathsf{MLE}}-\beta^\star\|_2^3.\end{aligned}$$

Applying Lemma C.2 and C.3, we know for any δ and any $n \geq c\max\{N_1\log(d/\delta), N_2\log(d/\delta), N(\delta)\}$, with probability at least $1-2\delta$, we have

$$(\beta_{\mathsf{MLE}}-\beta^\star)^T\mathcal{I}_T(\beta_{\mathsf{MLE}}-\beta^\star) \leq c\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}$$

and

$$\|\beta_{\mathsf{MLE}}-\beta^\star\|_2 \leq c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}}.$$

Also notice that, $\mathbb{E}_{\substack{x\sim\mathbb{P}_T(X)\\ y|x\sim f(y|x;\beta^\star)}}[\nabla\ell(x,y,\beta^\star)]=0$. Therefore, with probability at least $1-2\delta$, we have

$$R_{\beta^\star}(\beta_{\mathsf{MLE}}) \leq \frac{c}{2}\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}+\frac{c^3}{6}B_3\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5}$$

for any δ and any $n \geq c\max\{N_1\log(d/\delta), N_2\log(d/\delta), N(\delta)\}$. If we further assume $n \geq c(\frac{B_3\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})})^2\log(d/\delta)$, it then holds that

$$R_{\beta^\star}(\beta_{\mathsf{MLE}}) \leq c\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}.$$

Note that

$$
\begin{aligned}
&\max\left\{N_1, N_2, \left(\frac{B_3\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2\right\} \\
&= \max\left\{B_2^2\|\mathcal{I}_S^{-1}\|_2^2, B_3^2\|\mathcal{I}_S^{-1}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1}), \left(\frac{B_1^2B_2\|\mathcal{I}_S^{-1}\|_2^3\log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_S^{-1})}\right)^{\frac{2}{3}}, \left(\frac{B_1^3B_3\|\mathcal{I}_S^{-1}\|_2^4\log^{3\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_S^{-1})}\right)^{\frac{1}{2}},\right. \\
&\quad \frac{B_1^2\|\mathcal{I}_S^{-1}\|_2^2\log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_S^{-1})}, \left(\frac{B_2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2, \left(\frac{B_3\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2, \\
&\quad \left(\frac{B_1^2B_2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^2\log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^{\frac{2}{3}}, \left(\frac{B_1^3B_3\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^3\log^{3\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^{\frac{1}{2}}, \\
&\quad \left.\frac{B_1^2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2\log^{2\gamma}(\kappa^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}, \left(\frac{B_3\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2\right\} \\
&= \max\left\{\alpha_2^2, \tilde{\kappa}\alpha_3^2, \alpha_1^{4/3}\alpha_2^{2/3}\tilde{\kappa}^{-2/3}\log^{4\gamma/3}(\tilde{\kappa}^{-1/2}\alpha_1), \alpha_1^{3/2}\alpha_3^{1/2}\tilde{\kappa}^{-1/2}\log^{3\gamma/2}(\tilde{\kappa}^{-1/2}\alpha_1), \alpha_1^2\tilde{\kappa}^{-1}\log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1),\right. \\
&\qquad \alpha_2^2(\tilde{\kappa}/\kappa)^2, \alpha_3^2\tilde{\kappa}^3/\kappa^2, \alpha_1^{4/3}\alpha_2^{2/3}\kappa^{-2/3}\log^{4\gamma/3}(\tilde{\kappa}^{-1/2}\alpha_1), \alpha_1^{3/2}\alpha_3^{1/2}\kappa^{-1/2}\log^{3\gamma/2}(\tilde{\kappa}^{-1/2}\alpha_1), \\
&\qquad \left.\alpha_1^2\kappa^{-1}\log^{2\gamma}(\kappa^{-1/2}\alpha_1), \alpha_3^2\tilde{\kappa}^3\kappa^{-2}\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-1}\mathcal{I}_T^{\frac{1}{2}}\|_2^{-2}\right\} \\
&\le \max\left\{\tilde{\kappa}^{-1}\alpha_1^2\log^{2\gamma}\left((1+\tilde{\kappa}/\kappa)\tilde{\kappa}^{-1}\alpha_1^2\right), \kappa^{-1}\alpha_1^2\log^{2\gamma}\left((1+\tilde{\kappa}/\kappa)\tilde{\kappa}^{-1}\alpha_1^2\right), \alpha_2^2, (\tilde{\kappa}/\kappa)^2\alpha_2^2,\right. \\
&\qquad \left.\tilde{\kappa}\alpha_3^2, (\tilde{\kappa}^3/\kappa^2)\alpha_3^2, \tilde{\kappa}^3\kappa^{-2}\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-1}\mathcal{I}_T^{\frac{1}{2}}\|_2^{-2}\alpha_3^2\right\} \\
&\le (1+\tilde{\kappa}/\kappa)^2\cdot\max\{\tilde{\kappa}^{-1}\alpha_1^2\log^{2\gamma}\left((1+\tilde{\kappa}/\kappa)\tilde{\kappa}^{-1}\alpha_1^2\right), \alpha_2^2, \tilde{\kappa}(1+\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-1}\mathcal{I}_T^{\frac{1}{2}}\|_2^{-2})\alpha_3^2\} \\
&=: N^\star.
\end{aligned}
$$

To summarize, for any δ, any $n \ge c\max\{N^\star\log(d/\delta), N(\delta)\}$, with probability at least $1-2\delta$, we have

$$
R_{\beta^\star}(\beta_{\mathsf{MLE}}) \le c\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}.
$$

□

In the following, we prove Lemma C.2 and C.3.

Proof of Lemma C.2

Proof of Lemma C.2. For notation simplicity, we denote $g := \nabla\ell_n(\beta^\star) - \mathbb{E}[\nabla\ell_n(\beta^\star)]$. Note that

$$
\begin{aligned}
V &= n\cdot\mathbb{E}[\|A(\nabla\ell_n(\beta^\star) - \mathbb{E}[\nabla\ell_n(\beta^\star)])\|_2^2] \\
&= n\cdot\mathbb{E}[\nabla\ell_n(\beta^\star)^TA^TA\nabla\ell_n(\beta^\star)] \\
&= n\cdot\mathbb{E}[\mathsf{Tr}(A\nabla\ell_n(\beta^\star)\nabla\ell_n(\beta^\star)^TA^T)] \\
&= \mathsf{Tr}(A\mathcal{I}_SA^T).
\end{aligned}
$$

By taking $A = \mathcal{I}_S^{-1}$ in Assumption A.1, for any δ, any $n > N(\delta)$, we have with probability at least $1-\delta$:

$$\|\mathcal{I}_S^{-1} g\|_2 \le c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}} + B_1\|\mathcal{I}_S^{-1}\|_2 \log^{\gamma}\left(\frac{B_1\|\mathcal{I}_S^{-1}\|_2}{\sqrt{\mathsf{Tr}(\mathcal{I}_S^{-1})}}\right)\frac{\log\frac{d}{\delta}}{n}$$

$$= c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}} + B_1\|\mathcal{I}_S^{-1}\|_2 \log^{\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)\frac{\log\frac{d}{\delta}}{n}, \tag{11}$$

$$\left\|\nabla^2\ell_n(\beta^\star) - \mathbb{E}[\nabla^2\ell_n(\beta^\star)]\right\|_2 \le B_2\sqrt{\frac{\log\frac{d}{\delta}}{n}}. \tag{12}$$

Let event $A := \{(11), (12) \text{ holds}\}$. Under the event A, we have the following Taylor expansion:

$$\begin{aligned}
\ell_n(\beta) - \ell_n(\beta^\star) &\overset{\text{by Assumption A.2}}{\le} (\beta-\beta^\star)^T\nabla\ell_n(\beta^\star) + \frac{1}{2}(\beta-\beta^\star)^T\nabla^2\ell_n(\beta^\star)(\beta-\beta^\star) + \frac{B_3}{6}\|\beta-\beta^\star\|_2^3 \\
&\overset{\nabla\ell(\beta^\star)=0}{=} (\beta-\beta^\star)^T g + \frac{1}{2}(\beta-\beta^\star)^T\nabla^2\ell_n(\beta^\star)(\beta-\beta^\star) + \frac{B_3}{6}\|\beta-\beta^\star\|_2^3 \\
&\overset{\text{by (12)}}{\le} (\beta-\beta^\star)^T g + \frac{1}{2}(\beta-\beta^\star)^T\mathcal{I}_S(\beta-\beta^\star) + B_2\sqrt{\frac{\log\frac{d}{\delta}}{n}}\|\beta-\beta^\star\|_2^2 + \frac{B_3}{6}\|\beta-\beta^\star\|_2^3 \\
&\overset{\Delta_\beta := \beta-\beta^\star}{=} \Delta_\beta^T g + \frac{1}{2}\Delta_\beta^T\mathcal{I}_S\Delta_\beta + B_2\sqrt{\frac{\log\frac{d}{\delta}}{n}}\|\Delta_\beta\|_2^2 + \frac{B_3}{6}\|\Delta_\beta\|_2^3 \\
&= \frac{1}{2}(\Delta_\beta - z)^T\mathcal{I}_S(\Delta_\beta - z) - \frac{1}{2}z^T\mathcal{I}_S z + B_2\sqrt{\frac{\log\frac{d}{\delta}}{n}}\|\Delta_\beta\|_2^2 + \frac{B_3}{6}\|\Delta_\beta\|_2^3
\end{aligned} \tag{13}$$

where $z := -\mathcal{I}_S^{-1} g$. Similarly

$$\ell_n(\beta) - \ell_n(\beta^\star) \ge \frac{1}{2}(\Delta_\beta - z)^T\mathcal{I}_S(\Delta_\beta - z) - \frac{1}{2}z^T\mathcal{I}_S z - B_2\sqrt{\frac{\log\frac{d}{\delta}}{n}}\|\Delta_\beta\|_2^2 - \frac{B_3}{6}\|\Delta_\beta\|_2^3. \tag{14}$$

Notice that $\Delta_{\beta^\star+z} = z$, by (11) and (13), we have

$$\begin{aligned}
\ell_n(\beta^\star+z) - \ell_n(\beta^\star) &\le -\frac{1}{2}z^T\mathcal{I}_S z + B_2\sqrt{\frac{\log\frac{d}{\delta}}{n}}\left(c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}} + B_1\|\mathcal{I}_S^{-1}\|_2\log^{\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)\frac{\log\frac{d}{\delta}}{n}\right)^2 \\
&\quad + \frac{B_3}{6}\left(c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}} + B_1\|\mathcal{I}_S^{-1}\|_2\log^{\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)\frac{\log\frac{d}{\delta}}{n}\right)^3 \\
&\le -\frac{1}{2}z^T\mathcal{I}_S z + 2c^2B_2\mathsf{Tr}(\mathcal{I}_S^{-1})(\frac{\log\frac{d}{\delta}}{n})^{1.5} + 2B_1^2B_2\|\mathcal{I}_S^{-1}\|_2^2\log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^{2.5} \\
&\quad + \frac{2}{3}c^3B_3\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5} + \frac{2}{3}B_1^3B_3\|\mathcal{I}_S^{-1}\|_2^3\log^{3\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^3,
\end{aligned} \tag{15}$$

where we use the fact that $(a+b)^n \le 2^{n-1}(a^n+b^n)$ in the last inequality. For any $\beta \in \mathbb{B}_{\beta^\star}(3c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}})$, by (14), we have

$$\begin{aligned}
\ell_n(\beta) - \ell_n(\beta^\star) &\ge \frac{1}{2}(\Delta_\beta - z)^T\mathcal{I}_S(\Delta_\beta - z) - \frac{1}{2}z^T\mathcal{I}_S z \\
&\quad - 9c^2B_2\mathsf{Tr}(\mathcal{I}_S^{-1})(\frac{\log\frac{d}{\delta}}{n})^{1.5} - \frac{9}{2}c^3B_3\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5}.
\end{aligned} \tag{16}$$

(16) - (15) gives

$$\begin{aligned}
\ell_n(\beta) - \ell_n(\beta^\star + z) \geq & \frac{1}{2}(\Delta_\beta - z)^T \mathcal{I}_S (\Delta_\beta - z) \\
& - \Bigg(9c^2 B_2 \mathsf{Tr}(\mathcal{I}_S^{-1}) (\frac{\log \frac{d}{\delta}}{n})^{1.5} + \frac{9}{2} c^3 B_3 \mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5} (\frac{\log \frac{d}{\delta}}{n})^{1.5} \\
& \quad + 2c^2 B_2 \mathsf{Tr}(\mathcal{I}_S^{-1}) (\frac{\log \frac{d}{\delta}}{n})^{1.5} + 2B_1^2 B_2 \|\mathcal{I}_S^{-1}\|_2^2 \log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1) (\frac{\log \frac{d}{\delta}}{n})^{2.5} \\
& \quad\quad + \frac{2}{3} c^3 B_3 \mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5} (\frac{\log \frac{d}{\delta}}{n})^{1.5} + \frac{2}{3} B_1^3 B_3 \|\mathcal{I}_S^{-1}\|_2^3 \log^{3\gamma}(\tilde{\kappa}^{-1/2}\alpha_1) (\frac{\log \frac{d}{\delta}}{n})^3 \Bigg) \\
= & \frac{1}{2}(\Delta_\beta - z)^T \mathcal{I}_S (\Delta_\beta - z) \\
& - \Bigg(11c^2 B_2 \mathsf{Tr}(\mathcal{I}_S^{-1}) (\frac{\log \frac{d}{\delta}}{n})^{1.5} + \frac{31}{6} c^3 B_3 \mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5} (\frac{\log \frac{d}{\delta}}{n})^{1.5} \\
& \quad + 2B_1^2 B_2 \|\mathcal{I}_S^{-1}\|_2^2 \log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1) (\frac{\log \frac{d}{\delta}}{n})^{2.5} + \frac{2}{3} B_1^3 B_3 \|\mathcal{I}_S^{-1}\|_2^3 \log^{3\gamma}(\tilde{\kappa}^{-1/2}\alpha_1) (\frac{\log \frac{d}{\delta}}{n})^3 \Bigg)
\end{aligned} \tag{17}$$

Consider the ellipsoid

$$\begin{aligned}
\mathcal{D} := \Bigg\{ \beta \in \mathbb{R}^d \Bigg| & \frac{1}{2}(\Delta_\beta - z)^T \mathcal{I}_S (\Delta_\beta - z) \\
& \leq 11c^2 B_2 \mathsf{Tr}(\mathcal{I}_S^{-1}) (\frac{\log \frac{d}{\delta}}{n})^{1.5} + \frac{31}{6} c^3 B_3 \mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5} (\frac{\log \frac{d}{\delta}}{n})^{1.5} \\
& \quad + 2B_1^2 B_2 \|\mathcal{I}_S^{-1}\|_2^2 \log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1) (\frac{\log \frac{d}{\delta}}{n})^{2.5} + \frac{2}{3} B_1^3 B_3 \|\mathcal{I}_S^{-1}\|_2^3 \log^{3\gamma}(\tilde{\kappa}^{-1/2}\alpha_1) (\frac{\log \frac{d}{\delta}}{n})^3 \Bigg\}.
\end{aligned}$$

Then by (17), for any $\beta \in \mathbb{B}_{\beta^\star}(3c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1}) \log \frac{d}{\delta}}{n}}) \cap \mathcal{D}^C$,

$$\ell_n(\beta) - \ell_n(\beta^\star + z) > 0. \tag{18}$$

Notice that by the definition of $\mathcal{D}$, using $\lambda_{\min}^{-1}(\mathcal{I}_S) = \|\mathcal{I}_S^{-1}\|_2$, we have for any $\beta \in \mathcal{D}$,

$$\begin{aligned}
\|\Delta_\beta - z\|_2^2 \leq & 22c^2 B_2 \|\mathcal{I}_S^{-1}\|_2 \mathsf{Tr}(\mathcal{I}_S^{-1}) (\frac{\log \frac{d}{\delta}}{n})^{1.5} + \frac{31}{3} c^3 B_3 \|\mathcal{I}_S^{-1}\|_2 \mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5} (\frac{\log \frac{d}{\delta}}{n})^{1.5} \\
& + 4B_1^2 B_2 \|\mathcal{I}_S^{-1}\|_2^3 \log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1) (\frac{\log \frac{d}{\delta}}{n})^{2.5} + \frac{4}{3} B_1^3 B_3 \|\mathcal{I}_S^{-1}\|_2^4 \log^{3\gamma}(\tilde{\kappa}^{-1/2}\alpha_1) (\frac{\log \frac{d}{\delta}}{n})^3.
\end{aligned}$$

Thus for any $\beta \in \mathcal{D}$, we have

$$\begin{aligned}
\|\Delta_\beta\|_2^2 \leq & 2(\|\Delta_\beta - z\|_2^2 + \|z\|_2^2) \\
\overset{\text{by}(11)}{\leq} & 44c^2 B_2 \|\mathcal{I}_S^{-1}\|_2 \mathsf{Tr}(\mathcal{I}_S^{-1}) (\frac{\log \frac{d}{\delta}}{n})^{1.5} + \frac{62}{3} c^3 B_3 \|\mathcal{I}_S^{-1}\|_2 \mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5} (\frac{\log \frac{d}{\delta}}{n})^{1.5} \\
& + 8B_1^2 B_2 \|\mathcal{I}_S^{-1}\|_2^3 \log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1) (\frac{\log \frac{d}{\delta}}{n})^{2.5} + \frac{8}{3} B_1^3 B_3 \|\mathcal{I}_S^{-1}\|_2^4 \log^{3\gamma}(\tilde{\kappa}^{-1/2}\alpha_1) (\frac{\log \frac{d}{\delta}}{n})^3 \\
& + 4c^2 \frac{\mathsf{Tr}(\mathcal{I}_S^{-1}) \log \frac{d}{\delta}}{n} + 4B_1^2 \|\mathcal{I}_S^{-1}\|_2^2 \log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1) (\frac{\log \frac{d}{\delta}}{n})^2.
\end{aligned}$$

To guarantee $\frac{\mathsf{Tr}(\mathcal{I}_S^{-1}) \log \frac{d}{\delta}}{n}$ is the leading term, we only need $\frac{\mathsf{Tr}(\mathcal{I}_S^{-1}) \log \frac{d}{\delta}}{n}$ to dominate the rest of the terms. Hence, if we further have $n \geq cN_1 \log(d/\delta)$, it then holds that

$$\|\Delta_\beta\|_2^2 \leq 9c^2 \frac{\mathsf{Tr}(\mathcal{I}_S^{-1}) \log \frac{d}{\delta}}{n},$$

i.e., $\beta \in \mathbb{B}_{\beta^\star}(3c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}})$. Here

$$N_1 := \max\left\{B_2^2\|\mathcal{I}_S^{-1}\|_2^2, B_3^2\|\mathcal{I}_S^{-1}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1}), \left(\frac{B_1^2B_2\|\mathcal{I}_S^{-1}\|_2^3\log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_S^{-1})}\right)^{\frac{2}{3}}, \left(\frac{B_1^3B_3\|\mathcal{I}_S^{-1}\|_2^4\log^{3\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_S^{-1})}\right)^{\frac{1}{2}}, \frac{B_1^2\|\mathcal{I}_S^{-1}\|_2^2\log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_S^{-1})}\right\}.$$

In other words, we show that $\mathcal{D} \subset \mathbb{B}_{\beta^\star}(3c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}})$ when $n \geq c\max\{N_1\log(d/\delta), N(\delta)\}$. Recall that by (18), we know that for any $\beta \in \mathbb{B}_{\beta^\star}(3c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}}) \cap \mathcal{D}^C$,

$$\ell_n(\beta) - \ell_n(\beta^\star + z) > 0.$$

Note that $\beta^\star + z \in \mathcal{D}$. Hence there is a local minimum of $\ell_n(\beta)$ in $\mathcal{D}$. By Assumption A.3, we know that the global minimum of $\ell_n(\beta)$ is in $\mathcal{D}$, i.e.,

$$\beta_{\mathsf{MLE}} \in \mathcal{D} \subset \mathbb{B}_{\beta^\star}(3c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}}).$$

□

Proof of Lemma C.3

Proof of Lemma C.3. Let $E := \{\beta_{\mathsf{MLE}} \in \mathcal{D} \subset \mathbb{B}_{\beta^\star}(c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}})\}$. For any δ and any $n \geq c\max\{N_1\log(d/\delta), N(\delta)\}$, by the proof of Lemma C.2, we have $\mathbb{P}(E) \geq 1-\delta$.

By taking $A = \mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-1}$ in Assumption A.1, for any δ, any $n > N(\delta)$, we have with probability at least $1-\delta$:

$$\begin{aligned}\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-1}g\|_2 &\leq c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1}\mathcal{I}_T)\log\frac{d}{\delta}}{n}} + B_1\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-1}\|_2\log^{\gamma}\left(\frac{B_1\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-1}\|_2}{\sqrt{\mathsf{Tr}(\mathcal{I}_S^{-1}\mathcal{I}_T)}}\right)\frac{\log\frac{d}{\delta}}{n} \\ &\leq c\sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1}\mathcal{I}_T)\log\frac{d}{\delta}}{n}} + B_1\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-1}\|_2\log^{\gamma}(\kappa^{-1/2}\alpha_1)\frac{\log\frac{d}{\delta}}{n}. \qquad (19)\end{aligned}$$

We denote $E' := \{(19)\text{ holds}\}$. For any δ and any $n \geq c\max\{N_1\log(d/\delta), N(\delta)\}$, we have $\mathbb{P}(E\cap E') \geq 1-2\delta$.

Under $E\cap E'$, $\beta_{\mathsf{MLE}} \in \mathcal{D}$, i.e.,

$$\begin{aligned}&\frac{1}{2}(\Delta_{\beta_{\mathsf{MLE}}} - z)^T\mathcal{I}_S(\Delta_{\beta_{\mathsf{MLE}}} - z) \\ &\leq 11c^2B_2\mathsf{Tr}(\mathcal{I}_S^{-1})(\frac{\log\frac{d}{\delta}}{n})^{1.5} + \frac{31}{6}c^3B_3\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5} \\ &\quad + 2B_1^2B_2\|\mathcal{I}_S^{-1}\|_2^2\log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^{2.5} + \frac{2}{3}B_1^3B_3\|\mathcal{I}_S^{-1}\|_2^3\log^{3\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^3.\end{aligned}$$

In other words,

$$\begin{aligned}&\|\mathcal{I}_S^{\frac{1}{2}}(\Delta_{\beta_{\mathsf{MLE}}} - z)\|_2^2 \\ &\leq 22c^2B_2\mathsf{Tr}(\mathcal{I}_S^{-1})(\frac{\log\frac{d}{\delta}}{n})^{1.5} + \frac{31}{3}c^3B_3\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5} \\ &\quad + 4B_1^2B_2\|\mathcal{I}_S^{-1}\|_2^2\log^{2\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^{2.5} + \frac{4}{3}B_1^3B_3\|\mathcal{I}_S^{-1}\|_2^3\log^{3\gamma}(\tilde{\kappa}^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^3 \qquad (20)\end{aligned}$$

Thus we have

$$
\begin{aligned}
&\|\mathcal{I}_T^{\frac{1}{2}}(\beta_{\mathsf{MLE}}-\beta^\star)\|_2^2\\
&=\|\mathcal{I}_T^{\frac{1}{2}}\Delta_{\beta_{\mathsf{MLE}}}\|_2^2\\
&=\|\mathcal{I}_T^{\frac{1}{2}}(\Delta_{\beta_{\mathsf{MLE}}}-z)+\mathcal{I}_T^{\frac{1}{2}}z\|_2^2\\
&\le 2\|\mathcal{I}_T^{\frac{1}{2}}(\Delta_{\beta_{\mathsf{MLE}}}-z)\|_2^2+2\|\mathcal{I}_T^{\frac{1}{2}}z\|_2^2\\
&=2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}(\mathcal{I}_S^{\frac{1}{2}}(\Delta_{\beta_{\mathsf{MLE}}}-z))\|_2^2+2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-1}g\|_2^2\\
&\le 2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{\frac{1}{2}}(\Delta_{\beta_{\mathsf{MLE}}}-z)\|_2^2+2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-1}g\|_2^2\\
&\overset{\text{by(20)and(19)}}{\le} 4c^2\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}\\
&+44c^2B_2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})(\frac{\log\frac{d}{\delta}}{n})^{1.5}+\frac{62}{3}c^3B_3\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5}\\
&+8B_1^2B_2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^2\log^{2\gamma}(\tilde\kappa^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^{2.5}+\frac{8}{3}B_1^3B_3\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^3\log^{3\gamma}(\tilde\kappa^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^3\\
&\quad+4B_1^2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2\log^{2\gamma}(\kappa^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^2
\end{aligned}
$$

To guarantee $\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}$ is the leading term, we only need $\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}$ to dominate the rest of the terms. Hence, if we further have $n\ge cN_2\log(d/\delta)$, we have

$$
\|\mathcal{I}_T^{\frac{1}{2}}(\beta_{\mathsf{MLE}}-\beta^\star)\|_2^2\le 9c^2\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}.
$$

Here

$$
\begin{aligned}
N_2:=\max\Bigg\{&\left(\frac{B_2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2,\left(\frac{B_3\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2,\\
&\left(\frac{B_1^2B_2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^2\log^{2\gamma}(\tilde\kappa^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^{\frac{2}{3}},\left(\frac{B_1^3B_3\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^3\log^{3\gamma}(\tilde\kappa^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^{\frac{1}{2}},\\
&\frac{B_1^2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2\log^{2\gamma}(\kappa^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\Bigg\}.
\end{aligned}
$$

To summarize, we show that for any $\delta\in(0,1)$ and any $n\ge c\max\{N_1\log(d/\delta),N_2\log(d/\delta),N(\delta)\}$, with probability at least $1-2\delta$, we have

$$
\|\mathcal{I}_T^{\frac{1}{2}}(\beta_{\mathsf{MLE}}-\beta^\star)\|_2^2\le 9c^2\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}.
$$

□

C.2 Proofs for Theorem 3.2

The detailed version of Theorem 3.2 is stated as the following.

Theorem C.4. *Suppose the model class $\mathcal{F}$ satisfies Assumption B. Then we have*

$$
\begin{aligned}
&\inf_{\hat\beta}\sup_{\beta^\star\in\mathbb{B}_{\beta_0}(B)}\mathsf{Tr}\left(\mathcal{I}_T(\beta^\star)\mathcal{I}_S^{-1}(\beta^\star)\right)^{-1}\mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X)\\ y_i|x_i\sim f(y|x;\beta^\star)}}\left[R_{\beta^\star}(\hat\beta)\right]\\
&\ge\frac{1}{16}\cdot\frac{1}{2n+\frac{\pi^2d}{R_1^2}\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-2}(\beta_0)\right)\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right)^{-1}},
\end{aligned}
$$

where

$$
R_1:=\frac{1}{4}\sqrt{\frac{\lambda_{\min}(\mathcal{I}_T(\beta_0))}{\lambda_{\max}(\mathcal{I}_T(\beta_0))}}\cdot\min\left\{\frac{\lambda_{\min}^2(\mathcal{I}_S(\beta_0))}{4L_S\lambda_{\max}(\mathcal{I}_S(\beta_0))},\frac{\lambda_{\min}(\mathcal{I}_T(\beta_0))}{4B_3+2L_T},B\right\}.
$$

We first present some useful lemmas that will be used in the proof of Theorem C.4.

Lemma C.5. *Under Assumptions A.2, B.2 and B.3, we can choose $R_0 \le B$ such that for any $\beta, \beta^\star \in \mathbb{B}_{\beta_0}(R_0)$:*

$$\frac{1}{2}\cdot \mathcal{I}_S^{-1}(\beta_0) \preceq \mathcal{I}_S^{-1}(\beta) \preceq 2\cdot \mathcal{I}_S^{-1}(\beta_0), \tag{21}$$

$$\frac{1}{2}\cdot \mathcal{I}_T(\beta_0) \preceq \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X)\\ y|x\sim f(y|x;\beta^\star)}}\left[\nabla^2 \ell(x,y,\beta)\right] \preceq 2\cdot \mathcal{I}_T(\beta_0). \tag{22}$$

We can further choose $R_1 \le R_0$ such that for any $\beta^\star \in \mathbb{B}_{\beta_0}(R_1), \beta \notin \mathbb{B}_{\beta_0}(R_0)$: $R_{\beta^\star}(\beta) \ge R_{\beta^\star}(\beta_0)$.

Taking $\beta^\star = \beta$, Lemma C.5 (22) implies for any $\beta \in \mathbb{B}_{\beta_0}(R_0)$:

$$\frac{1}{2}\cdot \mathcal{I}_T(\beta_0) \preceq \mathcal{I}_T(\beta) \preceq 2\cdot \mathcal{I}_T(\beta_0). \tag{23}$$

Lemma C.6. *Let $C_{\beta_0}(B) := \{\beta \in \mathbb{R}^d \mid \beta - \beta_0 \in [-B, B]^d\}$ be a cube around β_0. For any $\beta_0 \in \mathbb{R}^d$ and $B > 0$, there exists a prior density $\lambda(\beta)$ supported on $C_{\beta_0}(B)$ such that for any estimator $\hat{\beta}$, we have*

$$\begin{aligned}
&\mathbb{E}_{\beta^\star\sim\lambda(\beta)}\mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X)\\ y_i|x_i\sim f(y|x;\beta^\star)}}\left[(\hat{\beta}-\beta^\star)^T\mathcal{I}_T(\beta_0)(\hat{\beta}-\beta^\star)\right]\\
&\ge \frac{\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right)^2}{n\mathbb{E}_{\beta^\star\sim\lambda(\beta)}\left[\mathsf{Tr}\left(\mathcal{I}_S^{-1}(\beta_0)\mathcal{I}_S(\beta^\star)\mathcal{I}_S^{-1}(\beta_0)\mathcal{I}_T(\beta_0)\right)\right] + \frac{\pi^2}{B^2}\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-2}(\beta_0)\right)}
\end{aligned}$$

The proofs for the above lemmas are delivered to the end of this subsection. With Lemma C.5 and Lemma C.6 in hand, we are now ready to prove Theorem C.4.

Proof of Theorem C.4. For any estimator $\hat{\beta}$, we define

$$\hat{\beta}^p := \begin{cases} \hat{\beta} & \hat{\beta} \in \mathbb{B}_{\beta_0}(R_0)\\ \beta_0 & \hat{\beta} \notin \mathbb{B}_{\beta_0}(R_0). \end{cases}$$

By Lemma C.5, for any $\beta^\star \in \mathbb{B}_{\beta_0}(R_1)$, we have $R_{\beta^\star}(\hat{\beta}) \ge R_{\beta^\star}(\hat{\beta}^p)$. We then have

$$\begin{aligned}
&\inf_{\hat{\beta}}\sup_{\beta^\star\in\mathbb{B}_{\beta_0}(B)}\mathsf{Tr}\left(\mathcal{I}_T(\beta^\star)\mathcal{I}_S^{-1}(\beta^\star)\right)^{-1}\mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X)\\ y_i|x_i\sim f(y|x;\beta^\star)}}\left[R_{\beta^\star}(\hat{\beta})\right]\\
&\ge \inf_{\hat{\beta}}\sup_{\beta^\star\in\mathbb{B}_{\beta_0}(R_1)}\mathsf{Tr}\left(\mathcal{I}_T(\beta^\star)\mathcal{I}_S^{-1}(\beta^\star)\right)^{-1}\mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X)\\ y_i|x_i\sim f(y|x;\beta^\star)}}\left[R_{\beta^\star}(\hat{\beta})\right]\\
&\ge \inf_{\hat{\beta}^p}\sup_{\beta^\star\in\mathbb{B}_{\beta_0}(R_1)}\mathsf{Tr}\left(\mathcal{I}_T(\beta^\star)\mathcal{I}_S^{-1}(\beta^\star)\right)^{-1}\mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X)\\ y_i|x_i\sim f(y|x;\beta^\star)}}\left[R_{\beta^\star}(\hat{\beta}^p)\right]\\
&\ge \inf_{\hat{\beta}\in\mathbb{B}_{\beta_0}(R_0)}\sup_{\beta^\star\in\mathbb{B}_{\beta_0}(R_1)}\mathsf{Tr}\left(\mathcal{I}_T(\beta^\star)\mathcal{I}_S^{-1}(\beta^\star)\right)^{-1}\mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X)\\ y_i|x_i\sim f(y|x;\beta^\star)}}\left[R_{\beta^\star}(\hat{\beta})\right],
\end{aligned} \tag{24}$$

where the first inequality follows from the fact that $R_1 \le R_0 \le B$, the second inequality follows from $R_{\beta^\star}(\hat{\beta}) \ge R_{\beta^\star}(\hat{\beta}^p)$, and the third inequality follows from $\hat{\beta}^p \in \mathbb{B}_{\beta_0}(R_0)$. For any $\beta^\star \in \mathbb{B}_{\beta_0}(R_1) \subseteq \mathbb{B}_{\beta_0}(R_0)$, by (21) and (23), we have

$$\mathcal{I}_T(\beta^\star) \preceq 2\mathcal{I}_T(\beta_0), \quad \mathcal{I}_S^{-1}(\beta^\star) \preceq 2\mathcal{I}_S^{-1}(\beta_0),$$

which implies

$$\mathsf{Tr}\left(\mathcal{I}_T(\beta^\star)\mathcal{I}_S^{-1}(\beta^\star)\right)^{-1} \ge \frac{1}{4}\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right)^{-1}. \tag{25}$$

Combine (24) and (25), we have

$$\begin{aligned}
&\inf_{\hat{\beta}}\sup_{\beta^\star\in\mathbb{B}_{\beta_0}(B)}\mathsf{Tr}\left(\mathcal{I}_T(\beta^\star)\mathcal{I}_S^{-1}(\beta^\star)\right)^{-1}\mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X)\\ y_i|x_i\sim f(y|x;\beta^\star)}}\left[R_{\beta^\star}(\hat{\beta})\right]\\
&\ge \frac{1}{4}\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right)^{-1}\inf_{\hat{\beta}\in\mathbb{B}_{\beta_0}(R_0)}\sup_{\beta^\star\in\mathbb{B}_{\beta_0}(R_1)}\mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X)\\ y_i|x_i\sim f(y|x;\beta^\star)}}\left[R_{\beta^\star}(\hat{\beta})\right].
\end{aligned} \tag{26}$$

By Taylor expansion, for any $\hat{\beta} \in \mathbb{B}_{\beta_0}(R_0), \beta^\star \in \mathbb{B}_{\beta_0}(R_1)$, we have

$$\begin{aligned} R_{\beta^\star}(\hat{\beta}) &= R_{\beta^\star}(\beta^\star) + (\hat{\beta}-\beta^\star)^T \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\nabla \ell(x,y,\beta^\star)] \\ &\quad + \frac{1}{2}(\hat{\beta}-\beta^\star)^T \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} \left[\nabla^2 \ell(x,y,\tilde{\beta})\right](\hat{\beta}-\beta^\star) \\ &= \frac{1}{2}(\hat{\beta}-\beta^\star)^T \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} \left[\nabla^2 \ell(x,y,\tilde{\beta})\right](\hat{\beta}-\beta^\star) \end{aligned}$$

for some $\tilde{\beta} \in \mathbb{B}_{\beta_0}(R_0)$. By Lemma C.5 (22), it then holds that

$$R_{\beta^\star}(\hat{\beta}) \geq \frac{1}{4}(\hat{\beta}-\beta^\star)^T \mathcal{I}_T(\beta_0)(\hat{\beta}-\beta^\star). \tag{27}$$

By (26) and (27), we then have

$$\begin{aligned} &\inf_{\hat{\beta}} \sup_{\beta^\star\in\mathbb{B}_{\beta_0}(B)} \mathsf{Tr}\left(\mathcal{I}_T(\beta^\star)\mathcal{I}_S^{-1}(\beta^\star)\right)^{-1} \mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X) \\ y_i|x_i\sim f(y|x;\beta^\star)}} \left[R_{\beta^\star}(\hat{\beta})\right] \\ &\geq \frac{1}{16}\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right)^{-1} \inf_{\hat{\beta}\in\mathbb{B}_{\beta_0}(R_0)} \sup_{\beta^\star\in\mathbb{B}_{\beta_0}(R_1)} \mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X) \\ y_i|x_i\sim f(y|x;\beta^\star)}} \left[(\hat{\beta}-\beta^\star)^T \mathcal{I}_T(\beta_0)(\hat{\beta}-\beta^\star)\right] \\ &\geq \frac{1}{16}\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right)^{-1} \inf_{\hat{\beta}\in\mathbb{B}_{\beta_0}(R_0)} \sup_{\beta^\star\in C_{\beta_0}(\frac{R_1}{\sqrt{d}})} \mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X) \\ y_i|x_i\sim f(y|x;\beta^\star)}} \left[(\hat{\beta}-\beta^\star)^T \mathcal{I}_T(\beta_0)(\hat{\beta}-\beta^\star)\right], \end{aligned} \tag{28}$$

where the last inequality follows from the fact that $C_{\beta_0}(\frac{R_1}{\sqrt{d}}) \subseteq \mathbb{B}_{\beta_0}(R_1)$. By Lemma C.6, there exists a prior density $\lambda(\beta)$ supported on $C_{\beta_0}(\frac{R_1}{\sqrt{d}})$ such that for any estimator $\hat{\beta}$, we have

$$\begin{aligned} &\mathbb{E}_{\beta^\star\sim\lambda(\beta)} \mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X) \\ y_i|x_i\sim f(y|x;\beta^\star)}} \left[(\hat{\beta}-\beta^\star)^T \mathcal{I}_T(\beta_0)(\hat{\beta}-\beta^\star)\right] \\ &\geq \frac{\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right)^2}{n\mathbb{E}_{\beta^\star\sim\lambda(\beta)}\left[\mathsf{Tr}\left(\mathcal{I}_S^{-1}(\beta_0)\mathcal{I}_S(\beta^\star)\mathcal{I}_S^{-1}(\beta_0)\mathcal{I}_T(\beta_0)\right)\right] + \frac{\pi^2 d}{R_1^2}\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-2}(\beta_0)\right)} \\ &\geq \frac{\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right)^2}{2n\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right) + \frac{\pi^2 d}{R_1^2}\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-2}(\beta_0)\right)}. \end{aligned}$$

Here the last inequality uses the fact that for any $\beta^\star \in C_{\beta_0}(\frac{R_1}{\sqrt{d}}) \subseteq \mathbb{B}_{\beta_0}(R_0)$, by Lemma C.5 (21), we have $\mathcal{I}_S^{-1}(\beta_0) \preceq 2\mathcal{I}_S^{-1}(\beta^\star)$, which implies

$$\begin{aligned} \mathbb{E}_{\beta^\star\sim\lambda(\beta)}\left[\mathsf{Tr}\left(\mathcal{I}_S^{-1}(\beta_0)\mathcal{I}_S(\beta^\star)\mathcal{I}_S^{-1}(\beta_0)\mathcal{I}_T(\beta_0)\right)\right] &\leq \mathbb{E}_{\beta^\star\sim\lambda(\beta)}\left[\mathsf{Tr}\left(2\mathcal{I}_S^{-1}(\beta^\star)\mathcal{I}_S(\beta^\star)\mathcal{I}_S^{-1}(\beta_0)\mathcal{I}_T(\beta_0)\right)\right] \\ &= 2\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right). \end{aligned}$$

We then conclude for any estimator $\hat{\beta}$

$$\begin{aligned} &\sup_{\beta^\star\in C_{\beta_0}(\frac{R_1}{\sqrt{d}})} \mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X) \\ y_i|x_i\sim f(y|x;\beta^\star)}} \left[(\hat{\beta}-\beta^\star)^T \mathcal{I}_T(\beta_0)(\hat{\beta}-\beta^\star)\right] \\ &\geq \mathbb{E}_{\beta^\star\sim\lambda(\beta)} \mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X) \\ y_i|x_i\sim f(y|x;\beta^\star)}} \left[(\hat{\beta}-\beta^\star)^T \mathcal{I}_T(\beta_0)(\hat{\beta}-\beta^\star)\right] \\ &\geq \frac{\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right)^2}{2n\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right) + \frac{\pi^2 d}{R_1^2}\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-2}(\beta_0)\right)}. \end{aligned} \tag{29}$$

Combine (28) and (29), we have

$$\begin{aligned}
&\inf_{\hat{\beta}} \sup_{\beta^\star \in \mathbb{B}_{\beta_0}(B)} \mathsf{Tr}\left(\mathcal{I}_T(\beta^\star)\mathcal{I}_S^{-1}(\beta^\star)\right)^{-1} \mathbb{E}_{\substack{x_i \sim \mathbb{P}_S(X) \\ y_i | x_i \sim f(y|x;\beta^\star)}} \left[R_{\beta^\star}(\hat{\beta})\right] \\
&\geq \frac{1}{16}\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right)^{-1} \cdot \frac{\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right)^2}{2n\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right) + \frac{\pi^2 d}{R_1^2}\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-2}(\beta_0)\right)} \\
&= \frac{1}{16} \cdot \frac{1}{2n + \frac{\pi^2 d}{R_1^2}\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-2}(\beta_0)\right)\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right)^{-1}}.
\end{aligned}$$

Thus we prove Theorem C.4. □

In the following, we prove Lemma C.5 and Lemma C.6.

Proofs for Lemma C.5

Proof of Lemma C.5. We choose

$$R_0 := \min\left\{\frac{\lambda_{\min}^2(\mathcal{I}_S(\beta_0))}{4L_S\lambda_{\max}(\mathcal{I}_S(\beta_0))}, \frac{\lambda_{\min}(\mathcal{I}_T(\beta_0))}{4B_3 + 2L_T}, B\right\}, \ R_1 := \frac{1}{4}\sqrt{\frac{\lambda_{\min}(\mathcal{I}_T(\beta_0))}{\lambda_{\max}(\mathcal{I}_T(\beta_0))}} \cdot R_0.$$

In the sequel, we will show the aforementioned choices of R_0 and R_1 satisfy the conditions outlined in Lemma C.5.

First of all, we show (21) holds. Fix any $\beta \in \mathbb{B}_{\beta_0}(R_0)$. By Assumption B.2, we have

$$\|\mathcal{I}_S(\beta) - \mathcal{I}_S(\beta_0)\|_2 \leq L_S\|\beta - \beta_0\|_2 \leq L_S R_0,$$

which implies

$$\|\mathcal{I}_S^{-1}(\beta) - \mathcal{I}_S^{-1}(\beta_0)\|_2 \leq \|\mathcal{I}_S^{-1}(\beta_0)\|_2 \cdot \|\mathcal{I}_S(\beta) - \mathcal{I}_S(\beta_0)\|_2 \cdot \|\mathcal{I}_S^{-1}(\beta)\|_2 \leq \frac{L_S R_0}{\lambda_{\min}(\mathcal{I}_S(\beta_0))\lambda_{\min}(\mathcal{I}_S(\beta))}.$$

By Weyl's inequality (Lemma 2.2 in Chen et al. (2021)), we have

$$|\lambda_{\min}(\mathcal{I}_S(\beta)) - \lambda_{\min}(\mathcal{I}_S(\beta_0))| \leq \|\mathcal{I}_S(\beta) - \mathcal{I}_S(\beta_0)\|_2 \leq L_S R_0.$$

Note that

$$R_0 \leq \frac{\lambda_{\min}^2(\mathcal{I}_S(\beta_0))}{4L_S\lambda_{\max}(\mathcal{I}_S(\beta_0))} \leq \frac{\lambda_{\min}(\mathcal{I}_S(\beta_0))}{2L_S}.$$

Thus we have

$$\lambda_{\min}(\mathcal{I}_S(\beta)) \geq \lambda_{\min}(\mathcal{I}_S(\beta_0)) - L_S R_0 \geq \frac{1}{2}\lambda_{\min}(\mathcal{I}_S(\beta_0)),$$

which implies

$$\|\mathcal{I}_S^{-1}(\beta) - \mathcal{I}_S^{-1}(\beta_0)\|_2 \leq \frac{L_S R_0}{\lambda_{\min}(\mathcal{I}_S(\beta_0))\lambda_{\min}(\mathcal{I}_S(\beta))} \leq \frac{2L_S R_0}{\lambda_{\min}^2(\mathcal{I}_S(\beta_0))} \leq \frac{1}{2\lambda_{\max}(\mathcal{I}_S(\beta_0))}. \tag{30}$$

Then for any $x \in \mathbb{R}^d$, we have

$$\begin{aligned}
x^T\left(\mathcal{I}_S^{-1}(\beta) - \frac{1}{2}\mathcal{I}_S^{-1}(\beta_0)\right)x &= \frac{1}{2}x^T\mathcal{I}_S^{-1}(\beta_0)x + x^T\left(\mathcal{I}_S^{-1}(\beta) - \mathcal{I}_S^{-1}(\beta_0)\right)x \\
&\geq \frac{\|x\|_2^2}{2\lambda_{\max}(\mathcal{I}_S(\beta_0))} - \|x\|_2^2 \cdot \|\mathcal{I}_S^{-1}(\beta) - \mathcal{I}_S^{-1}(\beta_0)\|_2 \\
&= \|x\|_2^2\left(\frac{1}{2\lambda_{\max}(\mathcal{I}_S(\beta_0))} - \|\mathcal{I}_S^{-1}(\beta) - \mathcal{I}_S^{-1}(\beta_0)\|_2\right) \\
&\geq 0,
\end{aligned}$$

where the last inequality follows from (30). Thus we conclude $\mathcal{I}_S^{-1}(\beta) \succeq \frac{1}{2}\mathcal{I}_S^{-1}(\beta_0)$. Similarly, we can show that $\mathcal{I}_S^{-1}(\beta) \preceq 2\mathcal{I}_S^{-1}(\beta_0)$. As a result, we show that (21) holds.

Next, we show (22) holds. Fix any $\beta^\star, \beta \in \mathbb{B}_{\beta_0}(R_0)$. By Assumption A.2, for any $x \in \mathcal{X}, y \in \mathcal{Y}$, we have

$$\|\nabla^2 \ell(x, y, \beta) - \nabla^2 \ell(x, y, \beta^\star)\|_2 \leq B_3 \|\beta - \beta^\star\|_2 \leq 2B_3 R_0,$$

which implies

$$\begin{aligned}
&\left\| \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\nabla^2 \ell(x, y, \beta)] - \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\nabla^2 \ell(x, y, \beta^\star)] \right\|_2 \\
&\quad \leq \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\|\nabla^2 \ell(x, y, \beta) - \nabla^2 \ell(x, y, \beta^\star)\|_2] \leq 2B_3 R_0. \qquad (31)
\end{aligned}$$

By Assumption B.2, we have

$$\|\mathcal{I}_T(\beta^\star) - \mathcal{I}_T(\beta_0)\|_2 \leq L_T \|\beta^\star - \beta_0\|_2 \leq L_T R_0 \qquad (32)$$

Thus, by (31) and (32), we have

$$\begin{aligned}
&\left\| \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\nabla^2 \ell(x, y, \beta)] - \mathcal{I}_T(\beta_0) \right\|_2 \\
&\leq \left\| \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\nabla^2 \ell(x, y, \beta)] - \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\nabla^2 \ell(x, y, \beta^\star)] \right\|_2 + \|\mathcal{I}_T(\beta^\star) - \mathcal{I}_T(\beta_0)\|_2 \\
&\leq (2B_3 + L_T) R_0 \\
&\leq \frac{1}{2} \lambda_{\min}(\mathcal{I}_T(\beta_0)),
\end{aligned}$$

where the last inequality follows from the choice of R_0. Consequently, for any $x \in \mathbb{R}^d$, we have

$$\begin{aligned}
&x^T \left(\mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\nabla^2 \ell(x, y, \beta)] - \frac{1}{2}\mathcal{I}_T(\beta_0) \right) x \\
&= \frac{1}{2} x^T \mathcal{I}_T(\beta_0) x + x^T \left(\mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\nabla^2 \ell(x, y, \beta)] - \mathcal{I}_T(\beta_0) \right) x \\
&\geq \frac{1}{2} \|x\|_2^2 \lambda_{\min}(\mathcal{I}_T(\beta_0)) - \|x\|_2^2 \left\| \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\nabla^2 \ell(x, y, \beta)] - \mathcal{I}_T(\beta_0) \right\|_2 \\
&\geq \frac{1}{2} \|x\|_2^2 \lambda_{\min}(\mathcal{I}_T(\beta_0)) - \frac{1}{2} \|x\|_2^2 \lambda_{\min}(\mathcal{I}_T(\beta_0)) = 0.
\end{aligned}$$

We then conclude $\mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\nabla^2 \ell(x, y, \beta)] \succeq \frac{1}{2}\mathcal{I}_T(\beta_0)$. Similarly, we can show that $\mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\nabla^2 \ell(x, y, \beta)] \preceq 2\mathcal{I}_T(\beta_0)$. Thus we show that (22) holds.

Finally, we need to show that for any $\beta^\star \in \mathbb{B}_{\beta_0}(R_1), \beta \notin \mathbb{B}_{\beta_0}(R_0)$: $R_{\beta^\star}(\beta) \geq R_{\beta^\star}(\beta_0)$. Fix any $\beta^\star \in \mathbb{B}_{\beta_0}(R_1), \beta \notin \mathbb{B}_{\beta_0}(R_0)$. We denote

$$\beta' := \{\lambda\beta + (1-\lambda)\beta^\star \mid \lambda \in [0,1]\} \cap \{\beta' \mid \|\beta' - \beta_0\|_2 = R_0\}.$$

By the choice of R_1, we know that $R_1 \leq R_0/2$, which implies

$$\|\beta' - \beta^\star\|_2 \geq \|\beta' - \beta_0\|_2 - \|\beta_0 - \beta^\star\|_2 \geq R_0 - R_1 \geq \frac{R_0}{2}. \qquad (33)$$

By convexity of $R_{\beta^\star}(\cdot)$ assumed in Assumption B.3 and $R_{\beta^\star}(\beta) \geq R_{\beta^\star}(\beta^\star)$, we have $R_{\beta^\star}(\beta) \geq R_{\beta^\star}(\beta')$. Thus, we obtain

$$\begin{aligned}
R_{\beta^\star}(\beta) - R_{\beta^\star}(\beta^\star) &\geq R_{\beta^\star}(\beta') - R_{\beta^\star}(\beta^\star) \\
&\overset{\text{Taylor}}{=} \frac{1}{2}(\beta' - \beta^\star)^T \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\nabla^2 \ell(x, y, \tilde{\beta})](\beta' - \beta^\star) \\
&\overset{\text{by (22)}}{\geq} \frac{1}{4}(\beta' - \beta^\star)^T \mathcal{I}_T(\beta_0)(\beta' - \beta^\star) \\
&\geq \frac{1}{4}\lambda_{\min}(\mathcal{I}_T(\beta_0))\|\beta' - \beta^\star\|_2^2 \\
&\overset{\text{by (33)}}{\geq} \frac{R_0^2}{16}\lambda_{\min}(\mathcal{I}_T(\beta_0)).
\end{aligned} \tag{34}$$

Note that

$$\begin{aligned}
R_{\beta^\star}(\beta_0) - R_{\beta^\star}(\beta^\star) &\overset{\text{Taylor}}{=} \frac{1}{2}(\beta_0 - \beta^\star)^T \mathbb{E}_{\substack{x\sim\mathbb{P}_T(X) \\ y|x\sim f(y|x;\beta^\star)}} [\nabla^2 \ell(x, y, \tilde{\beta})](\beta_0 - \beta^\star) \\
&\overset{\text{by (22)}}{\leq} (\beta_0 - \beta^\star)^T \mathcal{I}_T(\beta_0)(\beta_0 - \beta^\star) \\
&\leq \lambda_{\max}(\mathcal{I}_T(\beta_0))\|\beta_0 - \beta^\star\|_2^2 \\
&\leq R_1^2 \lambda_{\max}(\mathcal{I}_T(\beta_0)) \\
&= \frac{R_0^2}{16}\lambda_{\min}(\mathcal{I}_T(\beta_0)),
\end{aligned} \tag{35}$$

where the last equation follows from the choice of R_1. By (34) and (35), we obtain $R_{\beta^\star}(\beta) \geq R_{\beta^\star}(\beta_0)$. Thus, we finish the proof of Lemma C.5. □

Proofs for Lemma C.6

Proof of Lemma C.6. Let $\beta_0 = [\beta_{0,1}, \ldots, \beta_{0,d}]^T$, $\beta = [\beta_1, \ldots, \beta_d]^T$ and

$$f_i(x) := \frac{\pi}{4B} \cos\left(\frac{\pi}{2B}(x - \beta_{0,i})\right), \; i = 1, \ldots, d.$$

We define the prior density as

$$\lambda(\beta) := \begin{cases} \Pi_{i=1}^d f_i(\beta_i) & \beta \in C_{\beta_0}(B) \\ 0 & \beta \notin C_{\beta_0}(B) \end{cases},$$

which is supported on $C_{\beta_0}(B)$. In the sequel, we will show this prior density satisfies the condition outlined in Lemma C.6.

For notation simplicity, we denote

$$A = (A_{ij}) := \mathcal{I}_T^{-1}(\beta_0), \; C = (C_{ij}) := \mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0).$$

By multivariate van Trees inequality (Theorem 1 in Gill & Levit (1995)), for any estimator $\hat{\beta}$, we have

$$\begin{aligned}
&\mathbb{E}_{\beta^\star\sim\lambda(\beta)} \mathbb{E}_{\substack{x_i\sim\mathbb{P}_S(X) \\ y_i|x_i\sim f(y|x;\beta^\star)}} \left[(\hat{\beta} - \beta^\star)^T \mathcal{I}_T(\beta_0)(\hat{\beta} - \beta^\star)\right] \\
&\geq \frac{\mathsf{Tr}\left(\mathcal{I}_T(\beta_0)\mathcal{I}_S^{-1}(\beta_0)\right)^2}{n\mathbb{E}_{\beta^\star\sim\lambda(\beta)}\left[\mathsf{Tr}\left(\mathcal{I}_S^{-1}(\beta_0)\mathcal{I}_S(\beta^\star)\mathcal{I}_S^{-1}(\beta_0)\mathcal{I}_T(\beta_0)\right)\right] + \tilde{\mathcal{I}}(\lambda)},
\end{aligned} \tag{36}$$

where

$$\tilde{\mathcal{I}}(\lambda) = \int_{C_{\beta_0}(B)} \left(\sum_{i,j,k,\ell} A_{ij}C_{ik}C_{j\ell} \frac{\partial}{\partial\beta_k}\lambda(\beta) \frac{\partial}{\partial\beta_\ell}\lambda(\beta)\right) \frac{1}{\lambda(\beta)} d\beta.$$

By the choice of $\lambda(\beta)$, we have

$$\begin{aligned}
&\int_{C_{\beta_0}(B)} \left(\sum_{\substack{i,j,k,\ell \\ k\neq\ell}} A_{ij}C_{ik}C_{j\ell} \frac{\partial}{\partial\beta_k}\lambda(\beta)\frac{\partial}{\partial\beta_\ell}\lambda(\beta) \right) \frac{1}{\lambda(\beta)} d\beta \\
&= \int_{C_{\beta_0}(B)} \sum_{\substack{i,j,k,\ell \\ k\neq\ell}} A_{ij}C_{ik}C_{j\ell} f_k'(\beta_k) f_\ell'(\beta_\ell) \Pi_{i\neq k,\ell} f_i(\beta_i) d\beta \\
&= \sum_{\substack{i,j,k,\ell \\ k\neq\ell}} A_{ij}C_{ik}C_{j\ell} \int_{C_{\beta_0}(B)} f_k'(\beta_k) f_\ell'(\beta_\ell) \Pi_{i\neq k,\ell} f_i(\beta_i) d\beta \\
&= 0.
\end{aligned}$$

Here the last equation follows from the fact

$$\int_{\beta_{0,k}-B}^{\beta_{0,k}+B} f_k'(\beta_k) d\beta_k = \int_{\beta_{0,\ell}-B}^{\beta_{0,\ell}+B} f_\ell'(\beta_\ell) d\beta_\ell = 0.$$

Note that

$$\begin{aligned}
&\int_{C_{\beta_0}(B)} \left(\sum_{\substack{i,j,k,\ell \\ k=\ell}} A_{ij}C_{ik}C_{j\ell} \frac{\partial}{\partial\beta_k}\lambda(\beta)\frac{\partial}{\partial\beta_\ell}\lambda(\beta) \right) \frac{1}{\lambda(\beta)} d\beta \\
&= \sum_{i,j,k} A_{ij}C_{ik}C_{jk} \int_{C_{\beta_0}(B)} \frac{(f_k'(\beta_k))^2}{f_k(\beta_k)} \Pi_{i\neq k} f_i(\beta_i) d\beta \\
&= \sum_{i,j,k} A_{ij}C_{ik}C_{jk} \int_{\beta_{0,k}-B}^{\beta_{0,k}+B} \frac{(f_k'(\beta_k))^2}{f_k(\beta_k)} d\beta_k \\
&= \frac{\pi^2}{B^2} \sum_{i,j,k} A_{ij}C_{ik}C_{jk} \\
&= \frac{\pi^2}{B^2} \mathsf{Tr}(ACC^T).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\tilde{\mathcal{I}}(\lambda) &= \int_{C_{\beta_0}(B)} \left(\sum_{\substack{i,j,k,\ell \\ k\neq\ell}} A_{ij}C_{ik}C_{j\ell} \frac{\partial}{\partial\beta_k}\lambda(\beta)\frac{\partial}{\partial\beta_\ell}\lambda(\beta) \right) \frac{1}{\lambda(\beta)} d\beta \\
&\quad + \int_{C_{\beta_0}(B)} \left(\sum_{\substack{i,j,k,\ell \\ k=\ell}} A_{ij}C_{ik}C_{j\ell} \frac{\partial}{\partial\beta_k}\lambda(\beta)\frac{\partial}{\partial\beta_\ell}\lambda(\beta) \right) \frac{1}{\lambda(\beta)} d\beta \\
&= \frac{\pi^2}{B^2} \mathsf{Tr}(ACC^T) \\
&= \frac{\pi^2}{B^2} \mathsf{Tr}\left(\mathcal{I}_T(\beta_0) \mathcal{I}_S^{-2}(\beta_0) \right). \qquad (37)
\end{aligned}$$

Combine (36) and (37), we prove Lemma C.6.

□

D Proofs for Section A

D.1 Proofs for Proposition A.1 and Theorem A.2

Proof. For our linear regression model,

$$\ell(x, y, \beta) = \frac{1}{2}\log(2\pi) + \frac{1}{2}(y - x^T\beta)^2.$$

The convexity of ℓ in β immediately implies Assumption B.3. We then have

$$\begin{aligned}
&\nabla\ell(x, y, \beta) = -x(y - x^T\beta),\\
&\nabla^2\ell(x, y, \beta) = xx^T,\\
&\nabla^3\ell(x, y, \beta) = 0,\\
&\mathcal{I}_S = \mathbb{E}_{x\sim\mathbb{P}_S(X)}[xx^T] = I_d,\\
&\mathcal{I}_T = \mathbb{E}_{x\sim\mathbb{P}_T(X)}[xx^T] = \alpha\alpha^T + \sigma^2 I_d.
\end{aligned}$$

Therefore Assumption B.2 is satisfied with $L_S = L_T = 0$ and Assumption B.4 trivially holds. Note that $\nabla\ell(x_i, y_i, \beta^\star) = -x_i\varepsilon_i$. Since $\|x_i\|_2$ is $\sqrt{d}$-subgaussian and $|\varepsilon_i|$ is 1-subgaussian, by Lemma 2.7.7 in Vershynin (2018), it holds that $\|x_i\|_2|\varepsilon_i|$ is $\sqrt{d}$-subexponential random variable. Thus $\|A\nabla\ell(x_i, y_i, \beta^\star)\|_2$ is $\|A\|_2\sqrt{d}$-subexponential random variable.

Then, by Lemma F.1 with $u_i = A(\nabla\ell(x_i, y_i, \beta^\star) - \mathbb{E}[\nabla\ell(x_i, y_i, \beta^\star)]) = A\nabla\ell(x_i, y_i, \beta^\star)$, $V = \mathbb{E}[\|u_i\|_2^2] = n\cdot\mathbb{E}\|A(\nabla\ell_n(\beta^\star) - \mathbb{E}[\nabla\ell_n(\beta^\star)])\|_2^2$, $\alpha = 1$ and $B_u^{(\alpha)} = c\sqrt{d}\|A\|_2$, we have for any matrix $A \in \mathbb{R}^{d\times d}$, and any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\|A\left(\nabla\ell_n(\beta^\star) - \mathbb{E}[\nabla\ell_n(\beta^\star)]\right)\|_2 \le c\left(\sqrt{\frac{V\log\frac{d}{\delta}}{n}} + \sqrt{d}\|A\|_2\log(\frac{\sqrt{d}\|A\|_2}{\sqrt{V}})\frac{\log\frac{d}{\delta}}{n}\right),$$

which satisfies the gradient concentration in Assumption A.1 with $B_1 = c\sqrt{d}$ and $\gamma = 1$.

Note that $x_i \sim \mathcal{N}(0, I_d)$. Thus, by Theorem 13.3 in Rinaldo (2018), for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have

$$\begin{aligned}
\|\nabla^2\ell_n(\beta^\star) - \mathbb{E}[\nabla^2\ell_n(\beta^\star]\|_2 &= \left\|\frac{1}{n}\sum_{i=1}^n x_ix_i^T - I_d\right\|_2\\
&\le c\left(\sqrt{\frac{d\log(1/\delta)}{n}} + \frac{d\log(1/\delta)}{n}\right)\\
&\le 2c\sqrt{\frac{d\log(1/\delta)}{n}},
\end{aligned}$$

where the last inequality holds if $n \ge \mathcal{O}(d\log\frac{1}{\delta})$. Hence linear regression model satisfies the matrix concentration in Assumption A.1 with $B_2 = c\sqrt{d}$, $N(\delta) = d\log\frac{1}{\delta}$. Since $\nabla^3\ell \equiv 0$, we know Assumption A.2 holds with $B_3 = 0$.

Note that

$$\nabla^2\ell_n(\beta) = \frac{1}{n}\sum_{i=1}^n x_ix_i^T = \frac{1}{n}X^TX,$$

where $X := [x_1, \ldots, x_n]^T$. Given that $\{x_i\}_{i=1}^n$ are i.i.d $\mathcal{N}(0, I_d)$, it follows that X is almost surely full rank when $n \ge d$. Hence, when $n \ge d$, we have

$$\nabla^2\ell_n(\beta) = \frac{1}{n}\sum_{i=1}^n x_ix_i^T = \frac{1}{n}X^TX \succ 0.$$

Consequently, $\ell_n(\cdot)$ is strictly convex and thus satisfies Assumption A.3. Finally, Theorem A.2 follows directly from Theorem 3.1 with $\gamma = 1$, $B_1 = c\sqrt{d}, B_2 = c\sqrt{d}, B_3 = 0$, $N(\delta) = d\log\frac{1}{\delta}$, $\mathcal{I}_S = I_d$ and $\mathcal{I}_T = \alpha\alpha^T + \sigma^2 I_d$.

□

D.2 Proofs for Proposition A.3 and Theorem A.4

Proof. In the following, we will show the logistic regression model satisfies Assumptions A and B. For logistic regression, the loss function is defined as

$$\ell(x, y, \beta) = \log(1 + e^{x^T\beta}) - y(x^T\beta).$$

We then have

$$\begin{aligned}
\nabla\ell(x, y, \beta) &= \frac{x}{1 + e^{-x^T\beta}} - xy, \\
\nabla^2\ell(x, y, \beta) &= \frac{xx^T}{2 + e^{-x^T\beta} + e^{x^T\beta}}, \\
\nabla^3\ell(x, y, \beta) &= \frac{e^{-x^T\beta} - e^{x^T\beta}}{(2 + e^{-x^T\beta} + e^{x^T\beta})^2} \cdot x \otimes x \otimes x.
\end{aligned}$$

Here $\otimes$ represents the tensor product and $x \otimes x \otimes x \in \mathbb{R}^{d\times d\times d}$ with $(x \otimes x \otimes x)_{ijk} = x_i x_j x_k$. The convexity of ℓ in β immediately implies Assumption B.3; Assumption B.4 trivially holds. Note that on source domain $\|x\|_2 = \sqrt{d}$ and $|y| \le 1$. Hence we have for any (x, y) on source domain:

$$\begin{aligned}
\|\nabla\ell(x, y, \beta^\star)\|_2 &= \left\| \frac{x}{1 + e^{-x^T\beta^\star}} - xy \right\|_2 \le \left\| \frac{x}{1 + e^{-x^T\beta^\star}} \right\|_2 + \|xy\|_2 \le \|x\|_2 + \|x\|_2 = 2\sqrt{d}, \\
\|\nabla^2\ell(x, y, \beta^\star)\|_2 &= \left\| \frac{xx^T}{2 + e^{-x^T\beta^\star} + e^{x^T\beta^\star}} \right\|_2 \le \|xx^T\|_2 \le \|x\|_2^2 \le d.
\end{aligned}$$

By Lemma F.1 with $u_i = A(\nabla\ell(x_i, y_i, \beta^\star) - \mathbb{E}[\nabla\ell(x_i, y_i, \beta^\star)]) = A\nabla\ell(x_i, y_i, \beta^\star)$, $V = \mathbb{E}[\|u_i\|_2^2]$, $\alpha = +\infty$, $B_u^{(\alpha)} = 2\sqrt{d}\|A\|_2$, we have for any matrix $A \in \mathbb{R}^{d\times d}$, and any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\|A\left(\nabla\ell_n(\beta^\star) - \mathbb{E}[\nabla\ell_n(\beta^\star)]\right)\|_2 \le c\left(\sqrt{\frac{V\log\frac{d}{\delta}}{n}} + \frac{\sqrt{d}\|A\|_2\log\frac{d}{\delta}}{n}\right),$$

which satisfies the gradient concentration in Assumption A.1 with $B_1 = c\sqrt{d}$ and $\gamma = 0$. By matrix Hoeffding inequality, logistic regression model satisfies the matrix concentration in Assumption A.1 with $B_2 = cd$. We conclude that logistic regression model satisfies Assumption A.1 with $N(\delta) = 0$, $B_1 = c\sqrt{d}$, $\gamma = 0$, $B_2 = cd$.

Note that for x on source domain, we have $\|x\|_2 \le \sqrt{d}$; for x on target domain, we have $\|x\|_2 \le \sqrt{d} + r$. Thus, it holds that

$$\|\nabla^3\ell(x, y, \beta)\|_2 = \left\| \frac{e^{-x^T\beta} - e^{x^T\beta}}{(2 + e^{-x^T\beta} + e^{x^T\beta})^2} \cdot x \otimes x \otimes x \right\|_2 \le_{(i)} \|x \otimes x \otimes x\|_2 \le \|x\|_2^3 \le (\sqrt{d} + r)^3.$$

Here (i) uses the fact that

$$\left| \frac{e^{-x^T\beta} - e^{x^T\beta}}{(2 + e^{-x^T\beta} + e^{x^T\beta})^2} \right| \le \frac{e^{-x^T\beta} + e^{x^T\beta}}{(2 + e^{-x^T\beta} + e^{x^T\beta})^2} \le \frac{1}{2 + e^{-x^T\beta} + e^{x^T\beta}} \le 1.$$

Hence logistic regression satisfies Assumptions A.2 with $B_3 = (\sqrt{d} + r)^3$. Notice that this also implies Assumption B.2: By definition,

$$\mathcal{I}_S(\beta) := \mathbb{E}_{x\sim\mathbb{P}_S(X)}[\nabla^2\ell(x, y, \beta)],$$

therefore

$$\begin{aligned}
\|\mathcal{I}_S(\beta_1) - \mathcal{I}_S(\beta_2)\| &= \|\mathbb{E}_{x\sim\mathbb{P}_S(X)}[\nabla^2\ell(x, y, \beta_1) - \nabla^2\ell(x, y, \beta_2)]\| \\
&\le \mathbb{E}_{x\sim\mathbb{P}_S(X)}[\|\nabla^2\ell(x, y, \beta_1) - \nabla^2\ell(x, y, \beta_2)\|] \\
&\le (\sqrt{d})^3\|\beta_1 - \beta_2\|.
\end{aligned}$$

Similarly

$$\|\mathcal{I}_T(\beta_1) - \mathcal{I}_T(\beta_2)\| \le (\sqrt{d}+r)^3\|\beta_1 - \beta_2\|.$$

These inequlities shows that logistic regression model satisfies Assumption B.2 with $L_S = d^{1.5}$ and $L_T = (\sqrt{d}+r)^3$. Note that

$$\nabla^2 \ell_n(\beta) = \frac{1}{n}\sum_{i=1}^n \nabla^2 \ell(x_i, y_i, \beta) = \frac{1}{n}\sum_{i=1}^n \frac{x_i x_i^T}{2 + e^{-x_i^T\beta} + e^{x_i^T\beta}} = \frac{1}{n} X^T A X,$$

where $X := [x_1, \ldots, x_n]^T \in \mathbb{R}^{n\times d}$ and $A := \mathsf{diag}(1/(2+e^{-x_i^T\beta}+e^{x_i^T\beta})) \succ 0$. When $n \ge d$, X is full rank (i.e., $\mathsf{rank}(X) = d$) almost surely, consequently, $\ell_n(\cdot)$ is strictly convex and thus satisfies Assumption A.3.

By Theorem 3.1, we have when $n \ge \mathcal{O}(N^\star \log \frac{d}{\delta})$,

$$R_{\beta^\star}(\beta_{\mathsf{MLE}}) \lesssim \frac{\mathsf{Tr}\left(\mathcal{I}_T\mathcal{I}_S^{-1}\right)\log\frac{d}{\delta}}{n}.$$

Here

$$N^\star := (1+\tilde\kappa/\kappa)^2 \cdot \max\left\{\tilde\kappa^{-1}\alpha_1^2 \log^{2\gamma}\left((1+\tilde\kappa/\kappa)\tilde\kappa^{-1}\alpha_1^2\right),\ \alpha_2^2,\ \tilde\kappa(1+\|\mathcal{I}_T^{\frac12}\mathcal{I}_S^{-1}\mathcal{I}_T^{\frac12}\|_2^{-2})\alpha_3^2\right\},$$

where $\alpha_1 := B_1\|\mathcal{I}_S^{-1}\|_2^{0.5}$, $\alpha_2 := B_2\|\mathcal{I}_S^{-1}\|_2$, $\alpha_3 := B_3\|\mathcal{I}_S^{-1}\|_2^{1.5}$,

$$\kappa := \frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}{\|\mathcal{I}_T^{\frac12}\mathcal{I}_S^{-1}\mathcal{I}_T^{\frac12}\|_2},\ \tilde\kappa := \frac{\mathsf{Tr}(\mathcal{I}_S^{-1})}{\|\mathcal{I}_S^{-1}\|_2}.$$

Now it remains to calculate the quantities $N^\star$ and $\mathsf{Tr}\left(\mathcal{I}_T\mathcal{I}_S^{-1}\right)$ for this instance, where the crucial part is to identify what are $\mathcal{I}_S$ and $\mathcal{I}_T$. The following two lemmas give the characterization of $\mathcal{I}_S$ and $\mathcal{I}_T$.

Lemma D.1. *Under the conditions of Theorem A.4, we have $\mathcal{I}_S = U\mathsf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_2)U^T$ and $\mathcal{I}_T = U\mathsf{diag}(\lambda_1, \lambda_2 + r^2\lambda_3, \lambda_2, \ldots, \lambda_2)U^T$ for an orthonormal matrix U. Where*

$$\lambda_1 := \mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{(\beta^{\star T}x)^2}{2+\mathsf{exp}(\beta^{\star T}x)+\mathsf{exp}(-\beta^{\star T}x)}],$$
$$\lambda_2 := \mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{(\beta_\perp^{\star T}x)^2}{2+\mathsf{exp}(\beta^{\star T}x)+\mathsf{exp}(-\beta^{\star T}x)}],$$
$$\lambda_3 := \mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{1}{2+\mathsf{exp}(\beta^{\star T}x)+\mathsf{exp}(-\beta^{\star T}x)}].$$

Lemma D.2. *Under the conditions of Theorem A.4, there exist absolute constants $c, C, c' > 0$ such that $c < \lambda_1, \lambda_2, \lambda_3 < C$, for $d \ge c'$.*

The proofs for these two lemmas are in the next section. With Lemma D.1, we have $\mathcal{I}_T\mathcal{I}_S^{-1} = U\mathsf{diag}(1, 1+r^2\frac{\lambda_3}{\lambda_2}, \ldots, 1)U^T$, $\mathcal{I}_S^{-1} = U\mathsf{diag}(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_2})U^T$. By Lemma D.2, since $\lambda_1, \lambda_2, \lambda_3 = O(1)$, we have $\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1}) = d + r^2\frac{\lambda_3}{\lambda_2} \asymp d + r^2$, $\|\mathcal{I}_T\mathcal{I}_S^{-1}\|_2 = 1 + r^2\frac{\lambda_3}{\lambda_2} \asymp 1 + r^2$. Similarly $\mathsf{Tr}(\mathcal{I}_S^{-1}) = \lambda_1^{-1} + (d-1)\lambda_2^{-1} \asymp d$, $\|\mathcal{I}_S^{-1}\|_2 = \max\{\lambda_1^{-1}, \lambda_2^{-1}\} \asymp 1$. Also recall that $B_1 = \sqrt{d}, B_2 = d, B_3 = (\sqrt{d}+r)^3$, plug in all those quantities we have $\kappa = \frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}{\|\mathcal{I}_T\mathcal{I}_S^{-1}\|_2} \asymp \frac{d+r^2}{1+r^2}$, $\tilde\kappa = \frac{\mathsf{Tr}(\mathcal{I}_S^{-1})}{\|\mathcal{I}_S^{-1}\|_2} \asymp d$, $\alpha_1 = B_1\|\mathcal{I}_S^{-1}\|_2^{0.5} \asymp \sqrt{d}$, $\alpha_2 = B_2\|\mathcal{I}_S^{-1}\|_2 \asymp d$, $\alpha_3 = B_3\|\mathcal{I}_S^{-1}\|_2^{1.5} \asymp (\sqrt{d}+r)^3$. Therefore we have when $n \ge \mathcal{O}(N^\star \log\frac{d}{\delta})$,

$$R_{\beta^\star}(\beta_{\mathsf{MLE}}) \lesssim \frac{\mathsf{Tr}\left(\mathcal{I}_T\mathcal{I}_S^{-1}\right)\log\frac{d}{\delta}}{n} \asymp \frac{(d+r^2)\log\frac{d}{\delta}}{n},$$

where

$$
\begin{aligned}
N^\star &= (1+\tilde{\kappa}/\kappa)^2 \cdot \max\left\{\tilde{\kappa}^{-1}\alpha_1^2 \log^{2\gamma}\left((1+\tilde{\kappa}/\kappa)\tilde{\kappa}^{-1}\alpha_1^2\right),\ \alpha_2^2,\ \tilde{\kappa}(1+\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-1}\mathcal{I}_T^{\frac{1}{2}}\|_2^{-2})\alpha_3^2\right\} \\
&\asymp \left(1+\frac{d+r^2 d}{d+r^2}\right)^2 \cdot \max\left\{1, d^2, d(1+(1+r^2)^{-2})(\sqrt{d}+r)^6\right\} \\
&= \left(1+\frac{d+r^2 d}{d+r^2}\right)^2 \cdot d(\sqrt{d}+r)^6.
\end{aligned}
$$

When $r \lesssim 1$, $N^\star \asymp d^4$. When $1 \lesssim r \lesssim \sqrt{d}$, $N^\star \asymp r^4 d^4$. When $\sqrt{d} \lesssim r$, $N^\star \asymp r^6 d^3$. □

D.2.1 Proofs for Lemma D.1 and D.2

The intuition of proving Lemma D.1 and D.2 is that, when d is large, distribution $\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))$ behaves similar to distribution $\mathcal{N}(0, I_d)$ which has good properties (isotropic, independence of each entry, etc.)

Proof of Lemma D.1. By definition,

$$
\mathcal{I}_S := \mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{xx^T}{2+\exp(\beta^{\star T}x)+\exp(-\beta^{\star T}x)}]
$$

Let $z \sim \mathcal{N}(0, I_d)$, then x and $z\frac{\sqrt{d}}{\|z\|_2}$ have the same distribution. Therefore

$$
\begin{aligned}
\mathcal{I}_S &= \mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{xx^T}{2+\exp(\beta^{\star T}x)+\exp(-\beta^{\star T}x)}] \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{zz^T\frac{d}{\|z\|_2^2}}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}] \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{(\beta^\star\beta^{\star T}+U_\perp U_\perp^T)zz^T\frac{d}{\|z\|_2^2}}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}]
\end{aligned}
$$

where $[\beta^\star, U_\perp] \in \mathbb{R}^{d\times d}$ is a orthogonal basis.

With this expression, we first prove $\beta^\star$ is an eigenvector of $\mathcal{I}_S$ with corresponding eigenvalue λ_1.

$$
\begin{aligned}
\mathcal{I}_S\beta^\star &= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{(\beta^\star\beta^{\star T}+U_\perp U_\perp^T)zz^T\frac{d}{\|z\|_2^2}}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}]\beta^\star \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\beta^\star\beta^{\star T}zz^T\frac{d}{\|z\|_2^2}\beta^\star}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}] \\
&+ \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{U_\perp U_\perp^T zz^T\frac{d}{\|z\|_2^2}\beta^\star}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}] \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{(\beta^{\star T}z)^2\frac{d}{\|z\|_2^2}}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}]\beta^\star \\
&+ \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{U_\perp U_\perp^T zz^T\frac{d}{\|z\|_2^2}\beta^\star}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}] \\
&= \lambda_1\beta^\star + \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{U_\perp U_\perp^T zz^T\frac{d}{\|z\|_2^2}\beta^\star}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}].
\end{aligned}
$$

Therefore we only need to prove

$$\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{U_\perp U_\perp^T z z^T \frac{d}{\|z\|_2^2}\beta^\star}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}]=0.$$

In fact,

$$\begin{aligned}
&\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{U_\perp^T z z^T \frac{d}{\|z\|_2^2}\beta^\star}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}]\\
=&\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{d}{\|z\|_2^2}(U_\perp^T z)(z^T\beta^\star)}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}]\\
=&\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{d}{|A|^2+\|B\|^2}AB}{2+\exp(A\cdot\frac{\sqrt{d}}{\sqrt{|A|^2+\|B\|^2}})+\exp(-A\cdot\frac{\sqrt{d}}{|A|^2+\|B\|^2})}]
\end{aligned}$$

where we let $A:=z^T\beta^\star$, $B:=U_\perp^T z$. Notice that by the property of $z\sim\mathcal{N}(0,I_d)$, A and B are independent. Also, B is symmetric, i.e., B and $-B$ have the same distribution. Therefore

$$\begin{aligned}
&\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{d}{|A|^2+\|B\|^2}AB}{2+\exp(A\cdot\frac{\sqrt{d}}{\sqrt{|A|^2+\|B\|^2}})+\exp(-A\cdot\frac{\sqrt{d}}{|A|^2+\|B\|^2})}]\\
\overset{\text{replace } B \text{ by } -B}{=}&\ \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{-\frac{d}{|A|^2+\|B\|^2}AB}{2+\exp(A\cdot\frac{\sqrt{d}}{\sqrt{|A|^2+\|B\|^2}})+\exp(-A\cdot\frac{\sqrt{d}}{|A|^2+\|B\|^2})}]\\
=&-\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{d}{|A|^2+\|B\|^2}AB}{2+\exp(A\cdot\frac{\sqrt{d}}{\sqrt{|A|^2+\|B\|^2}})+\exp(-A\cdot\frac{\sqrt{d}}{|A|^2+\|B\|^2})}],
\end{aligned}$$

which implies

$$\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{U_\perp U_\perp^T z z^T \frac{d}{\|z\|_2^2}\beta^\star}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}]=0.$$

Next we will prove that for any $\beta_\perp$ such that $\|\beta_\perp\|_2=1$, $\beta^{\star T}\beta_\perp=0$, $\beta_\perp$ is an eigenvector of $\mathcal{I}_S$ with corresponding eigenvalue λ_2. Let $[\beta_\perp, U]$ be an orthogonal basis ($\beta^\star$ is the first column of U).

$$\begin{aligned}
\mathcal{I}_S\beta_\perp&=\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{(\beta_\perp\beta_\perp^T+UU^T)zz^T\frac{d}{\|z\|_2^2}}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}]\beta_\perp\\
&=\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\beta_\perp\beta_\perp^T zz^T\frac{d}{\|z\|_2^2}\beta_\perp}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}]\\
&+\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{UU^T zz^T\frac{d}{\|z\|_2^2}\beta_\perp}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}]\\
&=\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{(\beta_\perp^T z)^2\frac{d}{\|z\|_2^2}}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}]\beta_\perp\\
&+\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{UU^T zz^T\frac{d}{\|z\|_2^2}\beta_\perp}{2+\exp(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\exp(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}]\\
&=\lambda_2\beta_\perp+0\\
&=\lambda_2\beta_\perp
\end{aligned}$$

Here

$$\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{UU^Tzz^T\frac{d}{\|z\|_2^2}\beta_\perp}{2+\mathsf{exp}(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})+\mathsf{exp}(-\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})}]=0$$

because of a similar reason as in the previous part.

For $\mathcal{I}_T$, the proving strategy is similar. For $x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))+v$ on the target domain, where $v=r\beta_\perp^\star$, let $w=x-v=x-r\beta_\perp^\star$, then $w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))$. Let $z\sim\mathcal{N}(0,I_d)$, then w and $z\frac{\sqrt{d}}{\|z\|_2}$ have the same distribution. We have

$$\begin{aligned}\mathcal{I}_T&=\mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))+v}[\frac{xx^T}{2+\mathsf{exp}(\beta^{\star T}x)+\mathsf{exp}(-\beta^{\star T}x)}]\\&=\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{(w+v)(w+v)^T}{2+\mathsf{exp}(\beta^{\star T}(w+v))+\mathsf{exp}(-\beta^{\star T}(w+v))}]\\&\overset{v^T\beta^\star=0}{=}\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{ww^T+wv^T+vw^T+vv^T}{2+\mathsf{exp}(\beta^{\star T}w)+\mathsf{exp}(-\beta^{\star T}w)}]\end{aligned}$$

Therefore

$$\begin{aligned}\mathcal{I}_T\beta^\star&=\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{ww^T+wv^T+vw^T+vv^T}{2+\mathsf{exp}(\beta^{\star T}w)+\mathsf{exp}(-\beta^{\star T}w)}]\beta^\star\\&\overset{v^T\beta^\star=0}{=}\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{ww^T}{2+\mathsf{exp}(\beta^{\star T}w)+\mathsf{exp}(-\beta^{\star T}w)}]\beta^\star\\&=\mathcal{I}_S\beta^\star\\&=\lambda_1\beta^\star,\end{aligned}$$

where the last line follows from the previous proofs. Similarly, for any $\tilde{\beta_\perp}$ such that $\|\tilde{\beta_\perp}\|_2=1$, $\beta_\perp^{\star T}\tilde{\beta_\perp}=0$,

$$\begin{aligned}\mathcal{I}_T\tilde{\beta_\perp}&=\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{ww^T+wv^T+vw^T+vv^T}{2+\mathsf{exp}(\beta^{\star T}w)+\mathsf{exp}(-\beta^{\star T}w)}]\tilde{\beta_\perp}\\&\overset{v^T\tilde{\beta_\perp}=0}{=}\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{ww^T}{2+\mathsf{exp}(\beta^{\star T}w)+\mathsf{exp}(-\beta^{\star T}w)}]\tilde{\beta_\perp}\\&=\mathcal{I}_S\tilde{\beta_\perp}\\&=\lambda_2\tilde{\beta_\perp}.\end{aligned}$$

For $\beta_\perp^\star$,

$$\begin{aligned}\mathcal{I}_T\beta_\perp^\star&=\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{ww^T+wv^T+vw^T+vv^T}{2+\mathsf{exp}(\beta^{\star T}w)+\mathsf{exp}(-\beta^{\star T}w)}]\beta_\perp^\star\\&=\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{ww^T}{2+\mathsf{exp}(\beta^{\star T}w)+\mathsf{exp}(-\beta^{\star T}w)}]\beta_\perp^\star\\&+\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{wv^T}{2+\mathsf{exp}(\beta^{\star T}w)+\mathsf{exp}(-\beta^{\star T}w)}]\beta_\perp^\star\\&+\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{vw^T}{2+\mathsf{exp}(\beta^{\star T}w)+\mathsf{exp}(-\beta^{\star T}w)}]\beta_\perp^\star\\&+\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{vv^T}{2+\mathsf{exp}(\beta^{\star T}w)+\mathsf{exp}(-\beta^{\star T}w)}]\beta_\perp^\star\\&:=I_1+I_2+I_3+I_4.\end{aligned}$$

As in the previous proofs,

$$I_1=\mathcal{I}_S\beta_\perp^\star=\lambda_2\beta_\perp^\star.$$

$$
\begin{aligned}
I_2 &= \mathbb{E}_{w\sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{wv^T}{2+\exp(\beta^{\star T}w)+\exp(-\beta^{\star T}w)}]\beta_\perp^\star \\
&\overset{v=r\beta_\perp^\star}{=} r\mathbb{E}_{w\sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{w\beta_\perp^{\star T}\beta_\perp^\star}{2+\exp(\beta^{\star T}w)+\exp(-\beta^{\star T}w)}] \\
&\overset{\|\beta_\perp^\star\|=1}{=} r\mathbb{E}_{w\sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{w}{2+\exp(\beta^{\star T}w)+\exp(-\beta^{\star T}w)}] \\
&= 0.
\end{aligned}
$$

where the last lines follows from w is symmetric and $\frac{w}{2+\exp(\beta^{\star T}w)+\exp(-\beta^{\star T}w)}$ is a odd function of w.

$$
\begin{aligned}
I_3 &= \mathbb{E}_{w\sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{vw^T}{2+\exp(\beta^{\star T}w)+\exp(-\beta^{\star T}w)}]\beta_\perp^\star \\
&\overset{v=r\beta_\perp^\star}{=} r\mathbb{E}_{w\sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{\beta_\perp^\star w^T\beta_\perp^\star}{2+\exp(\beta^{\star T}w)+\exp(-\beta^{\star T}w)}] \\
&= r\mathbb{E}_{w\sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{w^T\beta_\perp^\star}{2+\exp(\beta^{\star T}w)+\exp(-\beta^{\star T}w)}]\beta_\perp^\star \\
&= 0.
\end{aligned}
$$

where the last lines follows from w is symmetric and $\frac{w^T\beta_\perp^\star}{2+\exp(\beta^{\star T}w)+\exp(-\beta^{\star T}w)}$ is a odd function of w.

$$
\begin{aligned}
I_4 &= \mathbb{E}_{w\sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{vv^T}{2+\exp(\beta^{\star T}w)+\exp(-\beta^{\star T}w)}]\beta_\perp^\star \\
&\overset{v=r\beta_\perp^\star}{=} r^2\mathbb{E}_{w\sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{\beta_\perp^\star\beta_\perp^{\star T}\beta_\perp^\star}{2+\exp(\beta^{\star T}w)+\exp(-\beta^{\star T}w)}] \\
&\overset{\|\beta_\perp^\star\|=1}{=} r^2\mathbb{E}_{w\sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{1}{2+\exp(\beta^{\star T}w)+\exp(-\beta^{\star T}w)}]\beta_\perp^\star \\
&= r^2\lambda_3\beta_\perp^\star.
\end{aligned}
$$

Combine the calculations of I_1, I_2, I_3, I_4, we have

$$
\begin{aligned}
\mathcal{I}_T\beta_\perp^\star &= I_1+I_2+I_3+I_4 \\
&= \lambda_2\beta_\perp^\star + r^2\lambda_3\beta_\perp^\star \\
&= (\lambda_2+r^2\lambda_3)\beta_\perp^\star.
\end{aligned}
$$

In conclusion, we have $\mathcal{I}_S = U\mathsf{diag}(\lambda_1,\lambda_2,\ldots,\lambda_2)U^T$ and $\mathcal{I}_T = U\mathsf{diag}(\lambda_1,\lambda_2+r^2\lambda_3,\lambda_2,\ldots,\lambda_2)U^T$ for an orthonormal matrix U, where $U=[\beta^\star,\beta_\perp^\star,\cdots]$. □

Proof of Lemma D.2. Recall the definition of $\lambda_1,\lambda_2,\lambda_3$:

$$
\begin{aligned}
\lambda_1 &:= \mathbb{E}_{x\sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{(\beta^{\star T}x)^2}{2+\exp(\beta^{\star T}x)+\exp(-\beta^{\star T}x)}] = \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{d}{\|z\|_2^2}(\beta^{\star T}z)^2}{2+\exp(\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)+\exp(-\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)}], \\
\lambda_2 &:= \mathbb{E}_{x\sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{(\beta_\perp^{\star T}x)^2}{2+\exp(\beta^{\star T}x)+\exp(-\beta^{\star T}x)}] = \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{d}{\|z\|_2^2}(\beta_\perp^{\star T}z)^2}{2+\exp(\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)+\exp(-\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)}], \\
\lambda_3 &:= \mathbb{E}_{x\sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\frac{1}{2+\exp(\beta^{\star T}x)+\exp(-\beta^{\star T}x)}] = \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{1}{2+\exp(\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)+\exp(-\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)}].
\end{aligned}
$$

Next we will show that there exists constants $c, C, c' > 0$ such that when $d \geq c'$, we have $c \leq \lambda_1 \leq C$. The proofs for λ_2 and λ_3 are similar. Notice that, when d is large, $\frac{d}{\|z\|_2^2}$ concentrates around 1. If we replace $\frac{d}{\|z\|_2^2}$ by 1 in the above expressions, we have

$$
\lambda_1 \approx \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{(\beta^{\star T}z)^2}{2+\exp(\beta^{\star T}z)+\exp(-\beta^{\star T}z)}]
$$

Since $\beta^{\star T} z \sim \mathcal{N}(0,1)$ when $z \sim \mathcal{N}(0, I_d)$ and $\|\beta^\star\| = 1$, we have

$$\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{(\beta^{\star T} z)^2}{2+\exp(\beta^{\star T} z)+\exp(-\beta^{\star T} z)}] = \mathbb{E}_{y\sim\mathcal{N}(0,1)}[\frac{y^2}{2+\exp(y)+\exp(-y)}]$$

which is a absolute constant greater than zero and not related to d. Following this intuition, we can bound λ_1 as the following. We first state the concentration of the norm of $\mathcal{N}(0, I_d)$. By Vershynin (2018) (3.7),

$$\mathbb{P}(|\|z\| - \sqrt{d}| \geq t) \leq 2e^{-4ct^2} \tag{38}$$

for some absolute constant $c > 0$. Take $t = \frac{\sqrt{d}}{2}$, we have

$$\mathbb{P}(\frac{\|z\|}{\sqrt{d}} \notin [\frac{1}{2}, \frac{3}{2}]) \leq 2e^{-cd}.$$

With this concentration, we do the following truncation:

$$\begin{aligned}
\lambda_1 &= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{d}{\|z\|_2^2}(\beta^{\star T} z)^2}{2+\exp(\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T} z)+\exp(-\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T} z)}] \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{d}{\|z\|_2^2}(\beta^{\star T} z)^2}{2+\exp(\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T} z)+\exp(-\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T} z)}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\in[\frac{1}{2},\frac{3}{2}]}] \\
&+ \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{d}{\|z\|_2^2}(\beta^{\star T} z)^2}{2+\exp(\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T} z)+\exp(-\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T} z)}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\notin[\frac{1}{2},\frac{3}{2}]}] \\
&:= J_1 + J_2.
\end{aligned}$$

For J_2, it is obvious that

$$0 \leq J_2 \leq \frac{d}{4}\mathbb{P}(\frac{\|z\|}{\sqrt{d}} \notin [\frac{1}{2}, \frac{3}{2}]) \leq \frac{d}{2}e^{-cd}. \tag{39}$$

For upper bound of J_1,

$$\begin{aligned}
J_1 &= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{d}{\|z\|_2^2}(\beta^{\star T} z)^2}{2+\exp(\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T} z)+\exp(-\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T} z)}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\in[\frac{1}{2},\frac{3}{2}]}] \\
&\leq \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{4(\beta^{\star T} z)^2}{4}] = 1.
\end{aligned}$$

Therefore

$$\begin{aligned}
\lambda_1 &= J_1 + J_2 \\
&\leq 1 + \frac{d}{2}e^{-cd}.
\end{aligned}$$

It's obvious that there exists an absolute constant c' such that when $d \geq c'$, $\lambda_1 \leq 2$.

For lower bound of J_1, we have

$$
\begin{aligned}
J_1 &= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{d}{\|z\|_2^2}(\beta^{\star T}z)^2}{2+\exp(\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)+\exp(-\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\in[\frac{1}{2},\frac{3}{2}]}] \\
&\geq \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{4}{9}(\beta^{\star T}z)^2}{2+\exp(2\beta^{\star T}z)+\exp(-2\beta^{\star T}z)}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\in[\frac{1}{2},\frac{3}{2}]}] \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{4}{9}(\beta^{\star T}z)^2}{2+\exp(2\beta^{\star T}z)+\exp(-2\beta^{\star T}z)}] - \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{4}{9}(\beta^{\star T}z)^2}{2+\exp(2\beta^{\star T}z)+\exp(-2\beta^{\star T}z)}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\notin[\frac{1}{2},\frac{3}{2}]}] \\
&\geq \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{4}{9}(\beta^{\star T}z)^2}{2+\exp(2\beta^{\star T}z)+\exp(-2\beta^{\star T}z)}] - \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{4}{9}(\beta^{\star T}z)^2}{4}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\notin[\frac{1}{2},\frac{3}{2}]}] \\
&\geq \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\frac{4}{9}(\beta^{\star T}z)^2}{2+\exp(2\beta^{\star T}z)+\exp(-2\beta^{\star T}z)}] - \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\|z\|_2^2}{9}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\notin[\frac{1}{2},\frac{3}{2}]}] \\
&= \mathbb{E}_{y\sim\mathcal{N}(0,1)}[\frac{\frac{4}{9}y^2}{2+\exp(2y)+\exp(-2y)}] - \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\|z\|_2^2}{9}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\notin[\frac{1}{2},\frac{3}{2}]}] \\
&:= c_1 - \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\|z\|_2^2}{9}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\notin[\frac{1}{2},\frac{3}{2}]}]
\end{aligned}
$$

Notice that here c_1 is a positive constant not related to d. For the second term,

$$
\begin{aligned}
&\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\|z\|_2^2}{9}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\notin[\frac{1}{2},\frac{3}{2}]}] \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\|z\|_2^2}{9}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\leq\frac{1}{2}}] + \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{\|z\|_2^2}{9}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\geq\frac{3}{2}}] \\
&\leq \frac{d}{36}\mathbb{P}(\frac{\|z\|}{\sqrt{d}}\leq\frac{1}{2}) + \frac{1}{9}\int_{\frac{9}{4}d}^{\infty}\mathbb{P}(\|z\|_2^2\geq t)\mathrm{d}t + \frac{1}{9}\cdot\frac{9}{4}d\mathbb{P}(\|z\|_2^2\geq\frac{9}{4}d) \\
&\overset{\text{by (38)}}{\leq} \frac{d}{36}2e^{-cd} + \frac{1}{9}\int_{\frac{9}{4}d}^{\infty}\mathbb{P}(\|z\|_2^2\geq t)\mathrm{d}t + \frac{d}{4}2e^{-cd} \\
&\overset{t=d(y+1)^2}{\leq} de^{-cd} + \frac{1}{9}\int_{\frac{1}{2}}^{\infty}2d(y+1)\mathbb{P}(\|z\|_2\geq\sqrt{d}+\sqrt{d}y)\mathrm{d}y \\
&\overset{\text{by (38)}}{\leq} de^{-cd} + \frac{1}{9}\int_{\frac{1}{2}}^{\infty}2d(y+1)2e^{-4cdy^2}\mathrm{d}y \\
&\leq de^{-cd} + 2d\int_{\frac{1}{2}}^{\infty}ye^{-4cdy^2}\mathrm{d}y \\
&\leq de^{-cd} + \frac{1}{4c}e^{-cd}
\end{aligned}
$$

Combine this inequality and previous inequalities of J_1 and J_2, we have

$$
\begin{aligned}
\lambda_1 &= J_1 + J_2 \\
&\geq c_1 - de^{-cd} - \frac{1}{4c}e^{-cd}
\end{aligned}
$$

Therefore it's obvious that there exists an absolute constant c' such that when $d\geq c'$, $\lambda_1\geq\frac{c_1}{2}$.

The proof for λ_2 is almost the same, the only difference is that in the numerator, we replace $\beta^{\star T}z$ by $\beta_\perp^{\star T}z$. The proof for λ_3 is even simpler. For upper bound,

$$
\lambda_3 = \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{1}{2+\exp(\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)+\exp(-\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)}] \leq \frac{1}{4}.
$$

For lower bound,

$$\begin{aligned}
\lambda_3 &= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{1}{2+\exp(\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)+\exp(-\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)}] \\
&\geq \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{1}{2+\exp(\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)+\exp(-\frac{\sqrt{d}}{\|z\|_2}\beta^{\star T}z)}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\in[\frac{1}{2},\frac{3}{2}]}] \\
&\geq \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{1}{2+\exp(2\beta^{\star T}z)+\exp(-2\beta^{\star T}z)}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\in[\frac{1}{2},\frac{3}{2}]}] \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{1}{2+\exp(2\beta^{\star T}z)+\exp(-2\beta^{\star T}z)}] - \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\frac{1}{2+\exp(2\beta^{\star T}z)+\exp(-2\beta^{\star T}z)}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\notin[\frac{1}{2},\frac{3}{2}]}] \\
&= c_2 - \frac{1}{4}\mathbb{P}(\frac{\|z\|}{\sqrt{d}}\notin[\frac{1}{2},\frac{3}{2}]) \\
&\geq c_2 - \frac{1}{2}e^{-cd}.
\end{aligned}$$

Therefore there exists constant c' such that when $d \geq c'$, $\lambda_3 \leq \frac{c_2}{2}$. □

D.3 Proofs for Theorem A.5

In this section, our objective is to establish the upper bound of MLE for the phase retrieval model. A direct application of Theorem 3.1 is impractical, as Assumption A.3 is not met; notably, both $\beta^\star, -\beta^\star$ serve as global minimums of population loss. To circumvent the issue of non-unique global minimums, we employ a methodology similar to that used in proving Theorem 3.1, though with a slightly refined analysis.

Proof of Theorem A.5. In the sequel, we will use the same notations as in the proof of Theorem 3.1. Even though the global minimum of population loss for the phase retrieval model isn't unique, meaning it could be either $\beta^\star$ or $-\beta^\star$, we can still show that the MLE falls into a small ball around either $\beta^\star$ or $-\beta^\star$.

Lemma D.3. *Under the settings of Theorem A.5, if $n \geq \mathcal{O}(d^4 \log\frac{d}{\delta})$, then with probability at least $1-\delta$, we have*

$$\min\{\|\beta_{\mathsf{MLE}} - \beta^\star\|_2, \|\beta_{\mathsf{MLE}} + \beta^\star\|_2\} \lesssim \sqrt{\frac{d^2\log\frac{d}{\delta}}{n}}.$$

Without loss of generality, in the sequel, we consider $n \geq \mathcal{O}(d^4\log\frac{d}{\delta})$ and assume

$$\|\beta_{\mathsf{MLE}} - \beta^\star\|_2 \lesssim \sqrt{\frac{d^2\log\frac{d}{\delta}}{n}}, \tag{40}$$

which implies $\beta_{\mathsf{MLE}} \in \mathbb{B}_{\beta^\star}(1)$.

Recall that for the phase retrieval model,

$$\ell(x,y,\beta) = \frac{1}{2}\log(2\pi) + \frac{1}{2}\left(y - (x^T\beta)^2\right)^2.$$

It then holds that

$$\begin{aligned}
\nabla\ell(x,y,\beta) &= 2(x^T\beta)^3 x - 2(x^T\beta)yx, \\
\nabla^2\ell(x,y,\beta) &= 6(x^T\beta)^2xx^T - 2yxx^T, \\
\nabla^3\ell(x,y,\beta) &= 12(x^T\beta)x\otimes x\otimes x.
\end{aligned}$$

Note that for $Y = (X^T\beta^\star)^2 + \varepsilon$, we have $\nabla\ell(X,Y,\beta^\star) = -2(X^T\beta^\star)X\varepsilon$. Therefore (recall that $\|\beta^\star\| = 1$) $\|\nabla\ell(x_i,y_i,\beta^\star)\|$ is $2d$-subgaussian, by Lemma F.1, we have for any δ, with probability at least $1-\delta$,

$$\|\mathcal{I}_S^{-1}g\|_2 \lesssim \sqrt{\frac{\mathsf{Tr}(\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}} + d\|\mathcal{I}_S^{-1}\|\sqrt{\log\frac{d^2\|\mathcal{I}_S^{-1}\|^2}{\mathsf{Tr}(\mathcal{I}_S^{-1})}}\frac{\log\frac{d}{\delta}}{n}. \tag{41}$$

Which can be viewed as setting $B_1 = d$ and $\gamma = \frac{1}{2}$ in Assumption A.1. Hence $\beta^\star + z = \beta^\star - \mathcal{I}_S^{-1} g \in \mathbb{B}_{\beta^\star}(1)$ when $n \geq \mathcal{O}(\max\{\mathsf{Tr}(\mathcal{I}_S^{-1}) \log \frac{d}{\delta}, d\|\mathcal{I}_S^{-1}\|_2 \sqrt{\log \frac{d^2\|\mathcal{I}_S^{-1}\|^2}{\mathsf{Tr}(\mathcal{I}_S^{-1})}} \log \frac{d}{\delta}\})$.

We then show the concentration inequality for the Hessian matrix. Note that

$$\nabla^2 \ell_n(\beta^\star) = \frac{1}{n}\sum_{i=1}^n \nabla^2 \ell(x_i, y_i, \beta^\star) = \frac{4}{n}\sum_{i=1}^n (x_i^T \beta^\star)^2 x_i x_i^T - \frac{2}{n}\sum_{i=1}^n \varepsilon_i x_i x_i^T.$$

Since $\|(x^T\beta^\star)^2 x x^T\| \leq d^2$, by matrix Hoeffding, with probability at least $1-\delta$, we have

$$\mathbb{E}_{\mathbb{P}_S}[(x^T\beta^\star)^2 x x^T] - d^2\sqrt{\frac{8\log\frac{d}{\delta}}{n}} I_d \preceq \frac{1}{n}\sum_{i=1}^n (x_i^T\beta^\star)^2 x_i x_i^T \preceq \mathbb{E}_{\mathbb{P}_S}[(x^T\beta^\star)^2 x x^T] + d^2\sqrt{\frac{8\log\frac{d}{\delta}}{n}} I_d \tag{42}$$

Moreover, by matrix Chernoff bound, with probability at least $1-\delta$, we have

$$-d\sqrt{\frac{8\log\frac{d}{\delta}}{n}} I_d \preceq -\frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i x_i^T \preceq d\sqrt{\frac{8\log\frac{d}{\delta}}{n}} I_d. \tag{43}$$

Combine (42) and (43), we obtain

$$\nabla^2 \ell(\beta^\star) - 6d^2\sqrt{\frac{8\log\frac{d}{\delta}}{n}} I_d \preceq \nabla^2 \ell_n(\beta^\star) \preceq \nabla^2 \ell(\beta^\star) + 6d^2\sqrt{\frac{8\log\frac{d}{\delta}}{n}} I_d, \tag{44}$$

which can be viewed as setting $B_2 = d^2$ in (12).

For any $\beta \in \mathbb{B}_{\beta^\star}(1)$, we have

$$\|\nabla^3 \ell(x, y, \beta)\|_2 = 12\|(x^T\beta) x \otimes x \otimes x\| \leq 24(\sqrt{d} + r)^4.$$

Thus, we can view as if this model satisfies $B_3 = (\sqrt{d}+r)^4$ in Assumption A.2.

Then same as (15) we have with probability $1-\delta$,

$$\begin{aligned}\ell_n(\beta^\star + z) - \ell_n(\beta^\star) \leq &-\frac{1}{2} z^T \mathcal{I}_S z + 2c^2 B_2 \mathsf{Tr}(\mathcal{I}_S^{-1})(\frac{\log\frac{d}{\delta}}{n})^{1.5} + 2B_1^2 B_2 \|\mathcal{I}_S^{-1}\|_2^2 \log(\tilde{\kappa}^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^{2.5} \\ &+ \frac{2}{3} c^3 B_3 \mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5} + \frac{2}{3} B_1^3 B_3 \|\mathcal{I}_S^{-1}\|_2^3 \log^{1.5}(\tilde{\kappa}^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^3,\end{aligned} \tag{45}$$

By Lemma D.3, we have (40). Then same as (16) we have with probability at least $1-\delta$,

$$\begin{aligned}\ell_n(\beta_{\mathsf{MLE}}) - \ell_n(\beta^\star) \geq &\frac{1}{2}(\Delta_{\beta_{\mathsf{MLE}}} - z)^T \mathcal{I}_S (\Delta_{\beta_{\mathsf{MLE}}} - z) - \frac{1}{2} z^T \mathcal{I}_S z \\ &- \mathcal{O}(B_2 d^2 (\frac{\log\frac{d}{\delta}}{n})^{1.5} + B_3 d^3 (\frac{\log\frac{d}{\delta}}{n})^{1.5}).\end{aligned} \tag{46}$$

Consequently, by (45), (46) and the fact that $\ell_n(\beta_{\mathsf{MLE}}) - \ell_n(\beta^\star + z) \leq 0$, we have

$$\begin{aligned}(\Delta_{\beta_{\mathsf{MLE}}} - z)^T \mathcal{I}_S (\Delta_{\beta_{\mathsf{MLE}}} - z) \leq \mathcal{O}\Big(&B_2 \mathsf{Tr}(\mathcal{I}_S^{-1})(\frac{\log\frac{d}{\delta}}{n})^{1.5} + B_1^2 B_2 \|\mathcal{I}_S^{-1}\|_2^2 \log(\tilde{\kappa}^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^{2.5} \\ &+ B_3 \mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5} + B_1^3 B_3 \|\mathcal{I}_S^{-1}\|_2^3 (\log(\tilde{\kappa}^{-1/2}\alpha_1))^{1.5}(\frac{\log\frac{d}{\delta}}{n})^3 \\ &+ B_2 d^2 (\frac{\log\frac{d}{\delta}}{n})^{1.5} + B_3 d^3 (\frac{\log\frac{d}{\delta}}{n})^{1.5}\Big)\end{aligned}$$

Then, same as the proof of Lemma C.3, we further have for any δ, with probability at least $1-2\delta$,

$$
\begin{aligned}
&(\beta_{\mathsf{MLE}}-\beta^\star)^T\mathcal{I}_T(\beta_{\mathsf{MLE}}-\beta^\star)\\
&\lesssim \frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}\\
&+\mathcal{O}\Bigg(B_2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})(\frac{\log\frac{d}{\delta}}{n})^{1.5}+B_1^2B_2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^2\log(\tilde{\kappa}^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^{2.5}\\
&\quad+B_3\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5}+B_1^3B_3\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^3(\log(\tilde{\kappa}^{-1/2}\alpha_1))^{1.5}(\frac{\log\frac{d}{\delta}}{n})^3\\
&\quad+B_2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2d^2(\frac{\log\frac{d}{\delta}}{n})^{1.5}+B_3\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2d^3(\frac{\log\frac{d}{\delta}}{n})^{1.5}\\
&\quad+B_1^2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2\log(\kappa^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^2\Bigg)\\
&=\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}\\
&+\mathcal{O}\Bigg(d^2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})(\frac{\log\frac{d}{\delta}}{n})^{1.5}+d^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^2\log(\tilde{\kappa}^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^{2.5}\\
&\quad+(\sqrt{d}+r)^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5}+d^3(\sqrt{d}+r)^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^3(\log(\tilde{\kappa}^{-1/2}\alpha_1))^{1.5}(\frac{\log\frac{d}{\delta}}{n})^3\\
&\quad+d^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2(\frac{\log\frac{d}{\delta}}{n})^{1.5}+d^3(\sqrt{d}+r)^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2(\frac{\log\frac{d}{\delta}}{n})^{1.5}\\
&\quad+d^2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2\log(\kappa^{-1/2}\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^2\Bigg)
\end{aligned}
\tag{47}
$$

To guarantee $\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}$ is the leading term, we only need $n\geq\mathcal{O}(N_1\log\frac{d}{\delta})$, where

$$
\begin{aligned}
N_1:=\max\Bigg\{&\left(\frac{d^2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2,\left(\frac{d^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^2\log(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^{\frac{2}{3}},\left(\frac{(\sqrt{d}+r)^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2,\\
&\left(\frac{d^3(\sqrt{d}+r)^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^3(\log(\tilde{\kappa}^{-1/2}\alpha_1))^{1.5}}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^{\frac{1}{2}},\left(\frac{d^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2,\left(\frac{d^3(\sqrt{d}+r)^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2,\\
&\frac{d^2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2\log(\kappa^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\Bigg\}.
\end{aligned}
$$

That is, for any δ, when $n\geq\mathcal{O}(\max\{d^4,\mathsf{Tr}(\mathcal{I}_S^{-1}),d\|\mathcal{I}_S^{-1}\|_2\log^{0.5}(\tilde{\kappa}^{-\frac{1}{2}}\alpha_1),N_1\}\log\frac{d}{\delta})$, with probability $1-2\delta$,

$$
(\beta_{\mathsf{MLE}}-\beta^\star)^T\mathcal{I}_T(\beta_{\mathsf{MLE}}-\beta^\star)\lesssim\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}.
$$

Then following the proof of Theorem 3.1, do Taylor expansion w.r.t. β as the following:

$$
\begin{aligned}
R_{\beta^\star}(\beta_{\mathsf{MLE}})&=\mathbb{E}_{\substack{x\sim\mathbb{P}_T(X)\\ y|x\sim f(y|x;\beta^\star)}}[\ell(x,y,\beta_{\mathsf{MLE}})-\ell(x,y,\beta^\star)]\\
&\leq\mathbb{E}_{\substack{x\sim\mathbb{P}_T(X)\\ y|x\sim f(y|x;\beta^\star)}}[\nabla\ell(x,y,\beta^\star)]^T(\beta_{\mathsf{MLE}}-\beta^\star)\\
&\quad+\frac{1}{2}(\beta_{\mathsf{MLE}}-\beta^\star)^T\mathcal{I}_T(\beta_{\mathsf{MLE}}-\beta^\star)+\frac{B_3}{6}\|\beta_{\mathsf{MLE}}-\beta^\star\|_2^3.\\
&\leq\frac{c}{2}\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}+\frac{c^3}{6}d^3(\sqrt{d}+r)^4(\frac{\log\frac{d}{\delta}}{n})^{1.5}.
\end{aligned}
$$

with probability at least $1-2\delta$. If we further assume $n \geq \mathcal{O}((\frac{d^3(\sqrt{d}+r)^4}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})})^2 \log\frac{d}{\delta})$, it then holds that

$$R_{\beta^\star}(\beta_{\mathsf{MLE}}) \leq c\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n}.$$

Therefore we conclude that for any δ, when $n \geq \mathcal{O}(N\log\frac{d}{\delta})$, with probability at least $1-2\delta$,

$$R_{\beta^\star}(\beta_{\mathsf{MLE}}) \leq c\frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})\log\frac{d}{\delta}}{n},$$

where

$$\begin{aligned}
N &:= \max\{d^4, \mathsf{Tr}(\mathcal{I}_S^{-1}), d\|\mathcal{I}_S^{-1}\|_2\log^{0.5}(\tilde{\kappa}^{-\frac{1}{2}}\alpha_1), N_1, (\frac{d^3(\sqrt{d}+r)^4}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})})^2\} \\
&= \max\Bigg\{\left(\frac{d^2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2, \left(\frac{d^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^2\log(\tilde{\kappa}^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^{\frac{2}{3}}, \left(\frac{(\sqrt{d}+r)^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\mathsf{Tr}(\mathcal{I}_S^{-1})^{1.5}}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2, \\
&\left(\frac{d^3(\sqrt{d}+r)^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2^3(\log(\tilde{\kappa}^{-1/2}\alpha_1))^{1.5}}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^{\frac{1}{2}}, \left(\frac{d^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2, \left(\frac{d^3(\sqrt{d}+r)^4\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}\right)^2, \\
&\frac{d^2\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-\frac{1}{2}}\|_2^2\|\mathcal{I}_S^{-1}\|_2\log(\kappa^{-1/2}\alpha_1)}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}, d^4, \mathsf{Tr}(\mathcal{I}_S^{-1}), d\|\mathcal{I}_S^{-1}\|_2\log^{0.5}(\tilde{\kappa}^{-\frac{1}{2}}\alpha_1), (\frac{d^3(\sqrt{d}+r)^4}{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})})^2\Bigg\}.
\end{aligned}$$

Now it remains to calculate N and $\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})$. Similar to logistic regression (see Lemma D.1 and D.2), we have the following two lemmas that characterize $\mathcal{I}_S$ and $\mathcal{I}_T$.

Lemma D.4. *Under the conditions of Theorem A.5, we have $\mathcal{I}_S = U\mathsf{diag}(\lambda_1,\lambda_2,\ldots,\lambda_2)U^T$ and $\mathcal{I}_T = U\mathsf{diag}(\lambda_1,\lambda_2+r^2\lambda_3,\lambda_2,\ldots,\lambda_2)U^T$ for an orthonormal matrix U. Where*

$$\begin{aligned}
\lambda_1 &:= 4\mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[(\beta^{\star T}x)^4], \\
\lambda_2 &:= 4\mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[(\beta^{\star T}x)^2(\beta_\perp^{\star T}x)^2], \\
\lambda_3 &:= 4\mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[(\beta^{\star T}x)^2].
\end{aligned}$$

Lemma D.5. *Under the conditions of Theorem A.5, there exist absolute constants $c, C, c' > 0$ such that $c < \lambda_1, \lambda_2, \lambda_3 < C$, for $d \geq c'$.*

The proofs for these two lemmas are in the next section. With Lemma D.4, we have $\mathcal{I}_T\mathcal{I}_S^{-1} = U\mathsf{diag}(1, 1+r^2\frac{\lambda_3}{\lambda_2},\ldots,1)U^T$, $\mathcal{I}_S^{-1} = U\mathsf{diag}(\frac{1}{\lambda_1},\frac{1}{\lambda_2},\ldots,\frac{1}{\lambda_2})U^T$. By Lemma D.5, since $\lambda_1,\lambda_2,\lambda_3 = O(1)$, we have $\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1}) = d+r^2\frac{\lambda_3}{\lambda_2} \asymp d+r^2$, $\|\mathcal{I}_T\mathcal{I}_S^{-1}\|_2 = 1+r^2\frac{\lambda_3}{\lambda_2} \asymp 1+r^2$. Similarly $\mathsf{Tr}(\mathcal{I}_S^{-1}) = \lambda_1^{-1}+(d-1)\lambda_2^{-1} \asymp d$, $\|\mathcal{I}_S^{-1}\|_2 = \max\{\lambda_1^{-1},\lambda_2^{-1}\} \asymp 1$, $\alpha_1 = B_1\|\mathcal{I}_S^{-1}\|_2^{\frac{1}{2}} \asymp d$. Plug in these quantities, recall

$$\kappa := \frac{\mathsf{Tr}(\mathcal{I}_T\mathcal{I}_S^{-1})}{\|\mathcal{I}_T^{\frac{1}{2}}\mathcal{I}_S^{-1}\mathcal{I}_T^{\frac{1}{2}}\|_2} \asymp \frac{d+r^2}{1+r^2}$$

we have

$$\begin{aligned}
N &= \max\Bigg\{d^6\kappa^{-2}, d^{\frac{8}{3}}\kappa^{-\frac{2}{3}}\log^{\frac{2}{3}}(\tilde{\kappa}^{-1/2}\alpha_1), d^3(\sqrt{d}+r)^8\kappa^{-2}, d^{\frac{3}{2}}(\sqrt{d}+r)^2\kappa^{-\frac{1}{2}}\log^{\frac{3}{4}}(\tilde{\kappa}^{-1/2}\alpha_1), d^8\kappa^{-2}, d^6(\sqrt{d}+r)^8\kappa^{-2}, \\
&\quad d^2\kappa^{-1}\log(\kappa^{-1/2}\alpha_1), d^4, d, d\log^{\frac{1}{2}}(\tilde{\kappa}^{-1/2}\alpha_1), d^6(\sqrt{d}+r)^8\kappa^{-2}\|\mathcal{I}_T\mathcal{I}_S^{-1}\|^{-2}\Bigg\} \\
&\overset{1\leq\kappa\leq d}{=} \max\left\{d^6(\sqrt{d}+r)^8\kappa^{-2}, d^6(\sqrt{d}+r)^8\kappa^{-2}\|\mathcal{I}_T\mathcal{I}_S^{-1}\|^{-2}\right\} \\
&\overset{\|\mathcal{I}_T\mathcal{I}_S^{-1}\|\asymp 1+r^2\geq 1}{=} d^6(\sqrt{d}+r)^8\kappa^{-2} \\
&\asymp \frac{d^6(\sqrt{d}+r)^8(1+r^2)^2}{(d+r^2)^2} \asymp d^6(d+r^2)^2(1+r^2)^2
\end{aligned}$$

We can see that when $r \leq 1$, $N \asymp d^8$. When $1 \leq r \leq \sqrt{d}$, $N \asymp d^8r^4$. When $r \geq \sqrt{d}$, $N \asymp d^6r^8$.

□

D.3.1 Proof of Lemma D.3

In the following, we prove Lemma D.3. The intuition is that, although ℓ is not convex in β, ℓ is quadratic in $M := \beta\beta^T$.

Proof of Lemma D.3. With a little bit abuse of notation, for matrix $M \in \mathbb{R}^{d\times d}$, we denote

$$\ell(x, y, M) := \frac{1}{2}(y - \langle xx^T, M\rangle)^2.$$

Under the case where $M = \beta\beta^T$, we have

$$\ell(x, y, M) := \frac{1}{2}(y - \langle xx^T, \beta\beta^T\rangle)^2 = \frac{1}{2}(y - (x^T\beta)^2)^2 = \ell(x, y, \beta).$$

We further denote

$$\ell_n(M) := \frac{1}{n}\sum_{i=1}^n \ell(x_i, y_i, M) = \frac{1}{2n}\sum_{i=1}^n (y_i - \langle x_i x_i^T, M\rangle)^2.$$

and $M^\star := \beta^\star\beta^{\star T}$.

It then holds that

$$\nabla\ell_n(M^\star) = -\frac{1}{n}\sum_{i=1}^n \mathsf{vec}(x_i x_i^T)\varepsilon_i,\ \nabla^2\ell_n(M^\star) = \frac{1}{n}\sum_{i=1}^n \mathsf{vec}(x_i x_i^T)\mathsf{vec}(x_i x_i^T)^T,\ \nabla^3\ell_n(M) = 0.$$

Denote $\Sigma_S := \mathbb{E}_{x\sim\mathbb{P}_S(X)}[\mathsf{vec}(xx^T)\mathsf{vec}(xx^T)^T]$, then by Lemma F.1 with $V = \mathsf{Tr}(\Sigma_S)$, $\alpha = 2$, $B_u^\alpha = cd$ for some absolute constants c, c', we have with probability at least $1-\delta$,

$$\|\nabla\ell_n(M^\star)\|_2 \le c'\left(\sqrt{\frac{\mathsf{Tr}(\Sigma_S)\log\frac{d}{\delta}}{n}} + d(\log\frac{c^2d^2}{\mathsf{Tr}(\Sigma_S)})^{\frac{1}{2}}\frac{\log\frac{d}{\delta}}{n}\right). \tag{48}$$

By matrix Hoeffding, we have with probability at least $1-\delta$,

$$\Sigma_S - d^2\sqrt{\frac{8\log\frac{d}{\delta}}{n}}I_d \preceq \nabla^2\ell_n(M^\star) \preceq \Sigma_S + d^2\sqrt{\frac{8\log\frac{d}{\delta}}{n}}I_d. \tag{49}$$

Before conducting further analysis, we need some characterizations of Σ_S. By the definition of Σ_S, we can see that the $((i,j),(k,l))$ entry of Σ_S is $\mathbb{E}_{X\sim\mathbb{P}_S(X)}[X_iX_jX_kX_l]$. Since X is symmetric and isotropic, we have

$$\mathbb{E}_{X\sim\mathbb{P}_S(X)}[X_iX_jX_kX_l] = \begin{cases}\mathbb{E}_{X\sim\mathbb{P}_S(X)}[X_i^2X_k^2] & \text{if } i=j, k=l \text{ and } i\neq k\\ \mathbb{E}_{X\sim\mathbb{P}_S(X)}[X_i^2X_j^2] & \text{if } \{i,j\}=\{k,l\} \text{ and } i\neq j\\ \mathbb{E}_{X\sim\mathbb{P}_S(X)}[X_i^4] & \text{if } i=j=k=l\\ 0 & \text{Otherwise}\end{cases}$$

For the calculation of moments, using (3a) in Cao (2020) with $a = (1, 0, \cdots, 0)^T$ and $\epsilon = \frac{1}{\sqrt{d}}X$, we have $\mathbb{E}_{X\sim\mathbb{P}_S(X)}[X_1^4] = \frac{3d}{d+2}$, $\mathbb{E}_{X\sim\mathbb{P}_S(X)}[X_1^2X_2^2] = \frac{d}{d+2}$. Since X is isotropic, we have

$$(\Sigma_S)_{((i,j),(k,l))} = \begin{cases}\frac{d}{d+2} & \text{if } i=j, k=l \text{ and } i\neq k\\ \frac{d}{d+2} & \text{if } \{i,j\}=\{k,l\} \text{ and } i\neq j\\ \frac{3d}{d+2} & \text{if } i=j=k=l\\ 0 & \text{Otherwise}\end{cases} \tag{50}$$

Therefore

$$\mathsf{Tr}(\Sigma_S) = \sum_{i,j}\mathbb{E}[X_i^2X_j^2] = d(d-1)\frac{d}{d+2} + d\frac{3d}{d+2} = d^2. \tag{51}$$

The following lemma characterizes the "minimum eigenvalue" of Σ_S on a special subspace, which will be useful in our analysis.

Lemma D.6. *For any vector $a = (a_{ij})_{(i,j)\in[d]\times[d]} \in \mathbb{R}^{d^2}$ satisfies $a_{ij} = a_{ji}$,*

$$a^T \Sigma_S a \geq \frac{2d}{d+2}\|a\|_2^2.$$

Proof.

$$\begin{aligned}
a^T\Sigma_S a &= \sum_{i,j,k,l} a_{ij}a_{kl}(\Sigma_S)_{((i,j),(k,l))} \\
&\overset{\text{by (50)}}{=} \frac{d}{d+2}\Big(\sum_{i\neq j} a_{ij}^2 + \sum_{i\neq j} a_{ij}a_{ji} + \sum_{i\neq j} a_{ii}a_{jj} + 3\sum_i a_{ii}^2\Big) \\
&\overset{a_{ij}=a_{ji}}{=} \frac{d}{d+2}\Big(2\sum_{i\neq j} a_{ij}^2 + \sum_{i\neq j} a_{ii}a_{jj} + 3\sum_i a_{ii}^2\Big) \\
&= \frac{d}{d+2}\Big(2\Big(\sum_{i\neq j} a_{ij}^2 + \sum_i a_{ii}^2\Big) + \Big(\sum_{i\neq j} a_{ii}a_{jj} + \sum_i a_{ii}^2\Big)\Big) \\
&= \frac{d}{d+2}\Big(2\|a\|_2^2 + \Big(\sum_i a_{ii}\Big)^2\Big) \\
&\geq \frac{2d}{d+2}\|a\|_2^2.
\end{aligned}$$

□

With Lemma D.6 and (51), we are now able to prove Lemma D.3. By Taylor expansion, we have for $M = \beta\beta^T$, $M^\star = \beta^\star\beta^{\star T}$, with probability at least $1-\delta$,

$$\begin{aligned}
\ell_n(M) - \ell_n(M^\star) &\overset{\nabla^3\ell_n \equiv 0}{=} \mathsf{vec}(M-M^\star)^T\nabla\ell_n(M^\star) + \frac{1}{2}\mathsf{vec}(M-M^\star)^T\nabla^2\ell_n(M^\star)\mathsf{vec}(M-M^\star) \\
&\overset{\text{by (48),(49)}}{\geq} -c'\|M-M^\star\|_F\left(\sqrt{\frac{\mathsf{Tr}(\Sigma_S)\log\frac{d}{\delta}}{n}} + d(\log\frac{c^2d^2}{\mathsf{Tr}(\Sigma_S)})^{\frac{1}{2}}\frac{\log\frac{d}{\delta}}{n}\right) \\
&\quad + \frac{1}{2}\mathsf{vec}(M-M^\star)^T\Sigma_S\mathsf{vec}(M-M^\star) - \|M-M^\star\|_F^2 d^2\sqrt{\frac{8\log\frac{d}{\delta}}{n}} \\
&\overset{\text{by Lemma D.6 and (51)}}{\geq} \left(\frac{d}{d+2} - d^2\sqrt{\frac{8\log\frac{d}{\delta}}{n}}\right)\|M-M^\star\|_F^2 - c''\left(\sqrt{\frac{d^2\log\frac{d}{\delta}}{n}} + d\frac{\log\frac{d}{\delta}}{n}\right)\|M-M^\star\|_F \\
&\geq \frac{1}{2}\|M-M^\star\|_F^2 - c''\left(\sqrt{\frac{d^2\log\frac{d}{\delta}}{n}} + d\frac{\log\frac{d}{\delta}}{n}\right)\|M-M^\star\|_F
\end{aligned}$$

when $n \geq \mathcal{O}(d^4\log\frac{d}{\delta})$.

We denote $M_{\mathsf{MLE}} := \beta_{\mathsf{MLE}}\beta_{\mathsf{MLE}}^T$. Note that $\ell_n(M_{\mathsf{MLE}}) - \ell_n(M^\star) = \ell_n(\beta_{\mathsf{MLE}}) - \ell_n(\beta^\star) \leq 0$. Thus we have

$$\frac{1}{2}\|M_{\mathsf{MLE}} - M^\star\|_F^2 - c''\left(\sqrt{\frac{d^2\log\frac{d}{\delta}}{n}} + d\frac{\log\frac{d}{\delta}}{n}\right)\|M_{\mathsf{MLE}} - M^\star\|_F \leq 0,$$

which implies

$$\|M_{\mathsf{MLE}} - M^\star\|_F \lesssim \left(\sqrt{\frac{d^2\log\frac{d}{\delta}}{n}} + d\frac{\log\frac{d}{\delta}}{n}\right) \lesssim \sqrt{\frac{d^2\log\frac{d}{\delta}}{n}}.$$

Thus so far we have shown, if $n \geq \mathcal{O}(d^4 \log \frac{d}{\delta})$, then with probability at least $1-\delta$, we have

$$\|M_{\mathsf{MLE}} - M^\star\|_F \lesssim \sqrt{\frac{d^2 \log \frac{d}{\delta}}{n}}.$$

By Lemma 6 in Ge et al. (2017), we further have

$$\min\{\|\beta_{\mathsf{MLE}} - \beta^\star\|_2, \|\beta_{\mathsf{MLE}} + \beta^\star\|_2\} \lesssim \frac{1}{\|\beta^\star\|_2}\|M_{\mathsf{MLE}} - M^\star\|_F \lesssim \sqrt{\frac{d^2 \log \frac{d}{\delta}}{n}}.$$

□

D.3.2 Proofs for Lemma D.4 and D.5

The proofs for Lemma D.4 and D.5 are similar to proofs for Lemma D.1 and D.2.

Proof of Lemma D.4. By definition,

$$\mathcal{I}_S := 4\mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[xx^T(x^T\beta^\star)^2]$$

Let $z \sim \mathcal{N}(0, I_d)$, then x and $z\frac{\sqrt{d}}{\|z\|_2}$ have the same distribution. Therefore

$$\begin{aligned}
\mathcal{I}_S &= 4\mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[xx^T(x^T\beta^\star)^2] \\
&= 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[zz^T\frac{d}{\|z\|_2^2}(\beta^{\star T}z\cdot\frac{\sqrt{d}}{\|z\|_2})^2] \\
&= 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\beta^\star\beta^{\star T} + U_\perp U_\perp^T)zz^T(\beta^{\star T}z)^2\frac{d^2}{\|z\|_2^4}]
\end{aligned}$$

where $[\beta^\star, U_\perp] \in \mathbb{R}^{d\times d}$ is a orthogonal basis.

With this expression, we first prove $\beta^\star$ is an eigenvector of $\mathcal{I}_S$ with corresponding eigenvalue λ_1.

$$\begin{aligned}
\mathcal{I}_S\beta^\star &= 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\beta^\star\beta^{\star T} + U_\perp U_\perp^T)zz^T(\beta^{\star T}z)^2\frac{d^2}{\|z\|_2^4}]\beta^\star \\
&= 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\beta^\star\beta^{\star T}zz^T(\beta^{\star T}z)^2\frac{d^2}{\|z\|_2^4}\beta^\star] \\
&+ 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[U_\perp U_\perp^T zz^T(\beta^{\star T}z)^2\frac{d^2}{\|z\|_2^4}\beta^\star] \\
&= 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\beta^{\star T}z)^4\frac{d^2}{\|z\|_2^4}]\beta^\star \\
&+ 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[U_\perp U_\perp^T zz^T(\beta^{\star T}z)^2\frac{d^2}{\|z\|_2^4}\beta^\star] \\
&= \lambda_1\beta^\star + 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[U_\perp U_\perp^T zz^T(\beta^{\star T}z)^2\frac{d^2}{\|z\|_2^4}\beta^\star].
\end{aligned}$$

Therefore we only need to prove

$$\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[U_\perp U_\perp^T zz^T(\beta^{\star T}z)^2\frac{d^2}{\|z\|_2^4}\beta^\star] = 0.$$

In fact,

$$\begin{aligned}
&\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[U_\perp^T zz^T(\beta^{\star T}z)^2\frac{d^2}{\|z\|_2^4}\beta^\star] \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(U_\perp^T z)(\beta^{\star T}z)^3\frac{d^2}{\|z\|_2^4}] \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\frac{d}{|A|^2+\|B\|^2})^2A^3B]
\end{aligned}$$

where we let $A := z^T\beta^\star$, $B := U_\perp^T z$. Notice that by the property of $z \sim \mathcal{N}(0, I_d)$, A and B are independent. Also, B is symmetric, i.e., B and $-B$ have the same distribution. Therefore

$$\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[U_\perp U_\perp^T z z^T (\beta^{\star T} z)^2 \frac{d^2}{\|z\|_2^4}\beta^\star] = \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\frac{d}{|A|^2+\|B\|^2})^2 A^3 B] = 0.$$

Next we will prove that for any $\beta_\perp$ such that $\|\beta_\perp\|_2 = 1$, $\beta^{\star T}\beta_\perp = 0$, $\beta_\perp$ is an eigenvector of $\mathcal{I}_S$ with corresponding eigenvalue λ_2. Let $[\beta_\perp, U]$ be an orthogonal basis ($\beta^\star$ is the first column of U).

$$\begin{aligned}
\mathcal{I}_S\beta_\perp &= 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\beta_\perp\beta_\perp^T + UU^T) z z^T (\beta^{\star T} z)^2 \frac{d^2}{\|z\|_2^4}]\beta_\perp \\
&= 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\beta_\perp\beta_\perp^T z z^T (\beta^{\star T} z)^2 \frac{d^2}{\|z\|_2^4}\beta_\perp] \\
&+ 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[UU^T z z^T (\beta^{\star T} z)^2 \frac{d^2}{\|z\|_2^4}\beta_\perp] \\
&= 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\beta_\perp^T z)^2 (\beta^{\star T} z)^2 \frac{d^2}{\|z\|_2^4}]\beta_\perp \\
&+ 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[UU^T z z^T (\beta^{\star T} z)^2 \frac{d^2}{\|z\|_2^4}\beta_\perp] \\
&= \lambda_2\beta_\perp + 0 \\
&= \lambda_2\beta_\perp
\end{aligned}$$

Here

$$4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[UU^T z z^T (\beta^{\star T} z)^2 \frac{d^2}{\|z\|_2^4}\beta_\perp] = 0$$

because of a similar reason as in the previous part.

For $\mathcal{I}_T$, the proving strategy is similar. For $x \sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d})) + v$ on the target domain, where $v = r\beta_\perp^\star$, let $w = x - v = x - r\beta_\perp^\star$, then $w \sim \mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))$. Let $z \sim \mathcal{N}(0, I_d)$, then w and $z\frac{\sqrt{d}}{\|z\|_2}$ have the same distribution. We have

$$\begin{aligned}
\mathcal{I}_T &= 4\mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))+v}[x x^T (x^T\beta^\star)^2] \\
&= 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[(w+v)(w+v)^T((w+v)^T\beta^\star)^2] \\
&\overset{v^T\beta^\star=0}{=} 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[(ww^T + wv^T + vw^T + vv^T)(w^T\beta^\star)^2]
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathcal{I}_T\beta^\star &= 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[(ww^T + wv^T + vw^T + vv^T)(w^T\beta^\star)^2]\beta^\star \\
&\overset{v^T\beta^\star=0}{=} 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[ww^T(w^T\beta^\star)^2]\beta^\star \\
&= \mathcal{I}_S\beta^\star \\
&= \lambda_1\beta^\star,
\end{aligned}$$

where the last line follows from the previous proofs. Similarly, for any $\tilde{\beta_\perp}$ such that $\|\tilde{\beta_\perp}\|_2 = 1$, $\beta_\perp^{\star T}\tilde{\beta_\perp} = 0$,

$$\begin{aligned}
\mathcal{I}_T\tilde{\beta_\perp} &= 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[(ww^T + wv^T + vw^T + vv^T)(w^T\beta^\star)^2]\tilde{\beta_\perp} \\
&\overset{v^T\tilde{\beta_\perp}=0}{=} 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[ww^T(w^T\beta^\star)^2]\tilde{\beta_\perp} \\
&= \mathcal{I}_S\tilde{\beta_\perp} \\
&= \lambda_2\tilde{\beta_\perp}.
\end{aligned}$$

For $\beta_\perp^\star$,

$$\begin{aligned}\mathcal{I}_T\beta_\perp^\star &= 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[(ww^T+wv^T+vw^T+vv^T)(w^T\beta^\star)^2]\beta_\perp^\star\\ &= 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[ww^T(w^T\beta^\star)^2]\beta_\perp^\star + 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[wv^T(w^T\beta^\star)^2]\beta_\perp^\star\\ &+ 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[vw^T(w^T\beta^\star)^2]\beta_\perp^\star + 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[vv^T(w^T\beta^\star)^2]\beta_\perp^\star\\ &:= I_1+I_2+I_3+I_4.\end{aligned}$$

As in the previous proofs,

$$I_1 = \mathcal{I}_S\beta_\perp^\star = \lambda_2\beta_\perp^\star.$$

$$\begin{aligned}I_2 &= 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[wv^T(w^T\beta^\star)^2]\beta_\perp^\star\\ &\overset{v=r\beta_\perp^\star}{=} 4r\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[w(\beta_\perp^{\star T}\beta_\perp^\star)(w^T\beta^\star)^2]\\ &\overset{\|\beta_\perp^\star\|=1}{=} 4r\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[w(w^T\beta^\star)^2]\\ &= 0.\end{aligned}$$

where the last lines follows from w is symmetric and $w(w^T\beta^\star)^2$ is a odd function of w.

$$\begin{aligned}I_3 &= 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[vw^T(w^T\beta^\star)^2]\beta_\perp^\star\\ &\overset{v=r\beta_\perp^\star}{=} 4r\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\beta_\perp^\star w^T\beta_\perp^\star(w^T\beta^\star)^2]\\ &= 4r\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[(w^T\beta_\perp^\star)(w^T\beta^\star)^2]\beta_\perp^\star\\ &= 0.\end{aligned}$$

where the last lines follows from w is symmetric and $(w^T\beta_\perp^\star)(w^T\beta^\star)^2$ is a odd function of w.

$$\begin{aligned}I_4 &= 4\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[vv^T(w^T\beta^\star)^2]\beta_\perp^\star\\ &\overset{v=r\beta_\perp^\star}{=} 4r^2\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\beta_\perp^\star\beta_\perp^{\star T}\beta_\perp^\star(w^T\beta^\star)^2]\\ &\overset{\|\beta_\perp^\star\|=1}{=} 4r^2\mathbb{E}_{w\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[\beta_\perp^\star(w^T\beta^\star)^2]\\ &= r^2\lambda_3\beta_\perp^\star.\end{aligned}$$

Combine the calculations of I_1, I_2, I_3, I_4, we have

$$\begin{aligned}\mathcal{I}_T\beta_\perp^\star &= I_1+I_2+I_3+I_4\\ &= \lambda_2\beta_\perp^\star + r^2\lambda_3\beta_\perp^\star\\ &= (\lambda_2+r^2\lambda_3)\beta_\perp^\star.\end{aligned}$$

In conclusion, we have $\mathcal{I}_S = U\mathsf{diag}(\lambda_1,\lambda_2,\ldots,\lambda_2)U^T$ and $\mathcal{I}_T = U\mathsf{diag}(\lambda_1,\lambda_2+r^2\lambda_3,\lambda_2,\ldots,\lambda_2)U^T$ for an orthonormal matrix U, where $U=[\beta^\star,\beta_\perp^\star,\cdots]$. □

Proof of Lemma D.5. Recall the definition of $\lambda_1,\lambda_2,\lambda_3$:

$$\begin{aligned}\lambda_1 &:= 4\mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[(\beta^{\star T}x)^4] = 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\beta^{\star T}z)^4\frac{d^2}{\|z\|_2^4}],\\ \lambda_2 &:= 4\mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[(\beta^{\star T}x)^2(\beta_\perp^{\star T}x)^2] = 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\beta^{\star T}z)^2(\beta_\perp^{\star T}z)^2\frac{d^2}{\|z\|_2^4}],\\ \lambda_3 &:= 4\mathbb{E}_{x\sim\mathsf{Uniform}(\mathcal{S}^{d-1}(\sqrt{d}))}[(\beta^{\star T}x)^2] = 4\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\beta^{\star T}z)^2\frac{d}{\|z\|_2^2}].\end{aligned}$$

Next we will show that there exists constants $c, C, c' > 0$ such that when $d \geq c'$, we have $c \leq \lambda_1 \leq C$. The proofs for λ_2 and λ_3 are similar. 38 With this concentration, we do the following truncation:

$$\begin{aligned}\frac{1}{4}\lambda_1 &= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\beta^{\star T}z)^4\frac{d^2}{\|z\|_2^4}]\\ &= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\beta^{\star T}z)^4\frac{d^2}{\|z\|_2^4}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\in[\frac{1}{2},\frac{3}{2}]}] + \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\beta^{\star T}z)^4\frac{d^2}{\|z\|_2^4}\mathbb{I}_{\frac{\|z\|}{\sqrt{d}}\notin[\frac{1}{2},\frac{3}{2}]}]\\ &:= J_1+J_2.\end{aligned}$$

For J_2, it is obvious that

$$0 \le J_2 \le d^2 \mathbb{P}(\frac{\|z\|}{\sqrt{d}} \notin [\frac{1}{2}, \frac{3}{2}]) \le 2d^2 e^{-cd}. \tag{52}$$

For upper bound of J_1,

$$\begin{aligned} J_1 &= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\beta^{\star T} z)^4 \frac{d^2}{\|z\|_2^4} \mathbb{I}_{\frac{\|z\|}{\sqrt{d}} \in [\frac{1}{2}, \frac{3}{2}]}] \\ &\le \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[16(\beta^{\star T} z)^4] = 48. \end{aligned}$$

Therefore

$$\frac{1}{4}\lambda_1 = J_1 + J_2 \le 48 + 2d^2 e^{-cd}.$$

It's obvious that there exists an absolute constant c' such that when $d \ge c'$, $\frac{1}{4}\lambda_1 \le 50$.

For lower bound of J_1, we have

$$\begin{aligned} J_1 &= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\beta^{\star T} z)^4 \frac{d^2}{\|z\|_2^4} \mathbb{I}_{\frac{\|z\|}{\sqrt{d}} \in [\frac{1}{2}, \frac{3}{2}]}] \\ &\ge \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[(\frac{2}{3})^4 (\beta^{\star T} z)^4] = (\frac{2}{3})^4 \cdot 3. \end{aligned}$$

Therefore

$$\frac{1}{4}\lambda_1 = J_1 + J_2 \ge (\frac{2}{3})^4 \cdot 3$$

Therefore it's obvious that there exists an absolute constant c' such that when $d \ge c'$, $\frac{1}{4}\lambda_1 \ge \frac{1}{2}$. The proofs for λ_2 and λ_3 are almost the same.

□

E Proofs for Section B

E.1 Poofs for Proposition B.1

Proof. We consider the case where $Y = X^2 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,1)$, $\varepsilon \perp\!\!\!\perp X$, and we have $X \sim \mathcal{N}(-10,1)$ on the source domain and $X \sim \mathcal{N}(10,1)$ on the target domain. Then the optimal linear fit on the target is given by

$$\beta^\star = \arg\min_{\beta\in\mathbb{R}} \mathbb{E}_{(x,y)\sim\mathbb{P}_T(X,Y)} \left[(y - x\beta)^2\right] = \left(\mathbb{E}_{x\sim\mathcal{N}(10,1)}[x^2]\right)^{-1} \mathbb{E}_{x\sim\mathcal{N}(10,1)}[x^3] > 0.$$

However, the linear fit learned via classical MLE asymptotically behaves as

$$\begin{aligned} \beta_{\mathsf{MLE}} &= \arg\min_{\beta\in\mathbb{R}} \frac{1}{2n} \sum_{i=1}^n (y_i - x_i\beta)^2 = \left(\frac{1}{n}\sum_{i=1}^n x_i^2\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n x_i y_i\right) \\ &\xrightarrow{n\to\infty} \left(\mathbb{E}_{x\sim\mathcal{N}(-10,1)}[x^2]\right)^{-1} \mathbb{E}_{x\sim\mathcal{N}(-10,1)}[x^3] < 0. \end{aligned}$$

Hence, the classical MLE losses consistency. For MWLE, we have

$$\begin{aligned} \beta_{\mathsf{MWLE}} &= \arg\min_{\beta\in\mathbb{R}} \frac{1}{2n} \sum_{i=1}^n w(x_i)(y_i - x_i\beta)^2 \\ &= \left(\frac{1}{n}\sum_{i=1}^n w(x_i) x_i^2\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n w(x_i) x_i y_i\right) \xrightarrow{n\to\infty} \beta^\star, \end{aligned}$$

which asymptotically provides a good estimator. □

E.2 Proofs for Theorem B.2

The detailed version of Theorem B.2 is stated as the following.

Theorem E.1. *Suppose the function class $\mathcal{F}$ satisfies Assumption C. Let $G_w := G_w(M)$ and $H_w := H_w(M)$. For any $\delta \in (0,1)$, if $n \geq c\max\{N^\star \log(d/\delta), N(\delta), N'(\delta)\}$, then with probability at least $1-3\delta$, we have*

$$R_M(\beta_{\mathsf{MWLE}}) \leq c\frac{\mathsf{Tr}\left(G_w H_w^{-1}\right)\log\frac{d}{\delta}}{n}$$

for an absolute constant c. Here

$$N^\star := W^2 \cdot \max\{\lambda^{-1}\tilde{\alpha}_1^2 \log^{2\gamma}(W^2\lambda^{-1}\tilde{\alpha}_1^2), \tilde{\alpha}_2^2, \lambda\tilde{\alpha}_3^2\},$$

where $\tilde{\alpha}_1 := B_1\|H_w^{-1}\|_2^{0.5}$, $\tilde{\alpha}_2 := B_2\|H_w^{-1}\|_2$, $\tilde{\alpha}_3 := B_3\|H_w^{-1}\|_2^{1.5}$, and $\lambda := \mathsf{Tr}(G_w H_w^{-2})/\|H_w^{-1}\|_2$.

The proofs for Theorem E.1 is similar to proofs for Theorem C.1. For notation simplicity, through out the proofs for Theorem E.1, let $\beta^\star := \beta^\star(M)$, $H_w := H_w(M)$, $G_w := G_w(M)$. We first state two main lemmas, which capture the distance between β_{MWLE} and $\beta^\star$ under different measurements.

Lemma E.2. *Suppose Assumption C holds. For any $\delta \in (0,1)$ and any $n \geq c\max\{N_1\log(d/\delta), N(\delta), N'(\delta)\}$, with probability at least $1-2\delta$, we have $\beta_{\mathsf{MWLE}} \in \mathbb{B}_{\beta^\star}(c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-2})\log\frac{d}{\delta}}{n}})$ for some absolute constant c. Here*

$$N_1 := \max\Bigg\{W^2B_2^2\|H_w^{-1}\|_2^2, W^2B_3^2\mathsf{Tr}(G_wH_w^{-2})\|H_w^{-1}\|_2^2, \left(\frac{W^3B_1^2B_2\|H_w^{-1}\|_2^3\log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)}{\mathsf{Tr}(G_wH_w^{-2})}\right)^{\frac{2}{3}},$$
$$\left(\frac{W^4B_1^3B_3\|H_w^{-1}\|_2^4\log^{3\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)}{\mathsf{Tr}(G_wH_w^{-2})}\right)^{\frac{1}{2}}, \frac{W^2B_1^2\|H_w^{-1}\|_2^2\log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)}{\mathsf{Tr}(G_wH_w^{-2})}\Bigg\}.$$

Lemma E.3. *Suppose Assumption C holds. For any $\delta \in (0,1)$ and any $n \geq c\max\{N_1\log(d/\delta), N_2\log(d/\delta), N(\delta), N'(\delta)\}$, with probability at least $1-3\delta$, we have*

$$\|H_w^{\frac{1}{2}}(\beta_{\mathsf{MWLE}} - \beta^\star)\|_2^2 \leq c\frac{\mathsf{Tr}(G_wH_w^{-1})\log\frac{d}{\delta}}{n}.$$

for some absolute constant c. Here N_1 is defined in Lemma E.2 and

$$N_2 := \max\Bigg\{\left(\frac{WB_2\mathsf{Tr}(G_wH_w^{-2})}{\mathsf{Tr}(G_wH_w^{-1})}\right)^2, \left(\frac{WB_3\mathsf{Tr}(G_wH_w^{-2})^{1.5}}{\mathsf{Tr}(G_wH_w^{-1})}\right)^2, \left(\frac{W^3B_1^2B_2\|H_w^{-1}\|_2^2\log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)}{\mathsf{Tr}(G_wH_w^{-1})}\right)^{\frac{2}{3}},$$
$$\left(\frac{W^4B_1^3B_3\|H_w^{-1}\|_2^3\log^{3\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)}{\mathsf{Tr}(G_wH_w^{-1})}\right)^{\frac{1}{2}}, \frac{W^2B_1^2\|H_w^{-1}\|_2\log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)}{\mathsf{Tr}(G_wH_w^{-1})}\Bigg\}.$$

The proofs for Lemma E.2 and E.3 are delayed to the end of this subsection. With these two lemmas, we can now state the proof for Theorem E.1.

Proof of Theorem E.1. By Assumption C.1 and C.3, we can do Taylor expansion w.r.t. β as the following:

$$\begin{aligned}
R_M(\beta_{\mathsf{MWLE}}) &= \mathbb{E}_{(x,y)\sim\mathbb{P}_T(x,y)}\left[\ell(x,y,\beta_{\mathsf{MWLE}}) - \ell(x,y,\beta^\star)\right] \\
&\leq \mathbb{E}_{(x,y)\sim\mathbb{P}_T(x,y)}[\nabla\ell(x,y,\beta^\star)]^T(\beta_{\mathsf{MWLE}} - \beta^\star) \\
&\quad + \frac{1}{2}(\beta_{\mathsf{MWLE}} - \beta^\star)^T H_w(\beta_{\mathsf{MWLE}} - \beta^\star) + \frac{WB_3}{6}\|\beta_{\mathsf{MWLE}} - \beta^\star\|_2^3.
\end{aligned}$$

Applying Lemma E.2 and E.3, we know for any δ and any $n \geq c\max\{N_1\log(d/\delta), N_2\log(d/\delta), N(\delta), N'(\delta)\}$, with probability at least $1-3\delta$, we have

$$(\beta_{\mathsf{MWLE}} - \beta^\star)^T H_w(\beta_{\mathsf{MWLE}} - \beta^\star) \leq c\frac{\mathsf{Tr}(G_wH_w^{-1})\log\frac{d}{\delta}}{n}$$

and

$$\|\beta_{\mathsf{MWLE}} - \beta^\star\|_2 \le c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-2})\log\frac{d}{\delta}}{n}}.$$

Also notice that, $\mathbb{E}_{(x,y)\sim\mathbb{P}_T(x,y)}[\nabla\ell(x,y,\beta^\star)] = 0$. Therefore, with probability at least $1-3\delta$, we have

$$R_M(\beta_{\mathsf{MWLE}}) \le \frac{c}{2}\frac{\mathsf{Tr}(G_w H_w^{-1})\log\frac{d}{\delta}}{n} + \frac{c^3}{6}WB_3\mathsf{Tr}(G_w H_w^{-2})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5}.$$

If we further have $n \ge c(\frac{WB_3\mathsf{Tr}(G_w H_w^{-2})^{1.5}}{\mathsf{Tr}(G_w H_w^{-1})})^2\log(d/\delta)$, it then holds that

$$R_M(\beta_{\mathsf{MWLE}}) \le c\frac{\mathsf{Tr}(G_w H_w^{-1})\log\frac{d}{\delta}}{n}.$$

Note that

$$\begin{aligned}
&\max\left\{N_1, N_2, \left(\frac{WB_3\mathsf{Tr}(G_w H_w^{-2})^{1.5}}{\mathsf{Tr}(G_w H_w^{-1})}\right)^2\right\}\\
&= \max\left\{W^2B_2^2\|H_w^{-1}\|_2^2,\ W^2B_3^2\mathsf{Tr}(G_w H_w^{-2})\|H_w^{-1}\|_2^2,\ \left(\frac{W^3B_1^2B_2\|H_w^{-1}\|_2^3\log^{2\gamma}(W\lambda^{-1/2}\tilde\alpha_1)}{\mathsf{Tr}(G_w H_w^{-2})}\right)^{\frac{2}{3}},\right.\\
&\left.\left(\frac{W^4B_1^3B_3\|H_w^{-1}\|_2^4\log^{3\gamma}(W\lambda^{-1/2}\tilde\alpha_1)}{\mathsf{Tr}(G_w H_w^{-2})}\right)^{\frac{1}{2}},\ \frac{W^2B_1^2\|H_w^{-1}\|_2^2\log^{2\gamma}(W\lambda^{-1/2}\tilde\alpha_1)}{\mathsf{Tr}(G_w H_w^{-2})}\right\}\\
&= W^2\cdot\max\{\tilde\alpha_2^2, \lambda\tilde\alpha_3^2, \tilde\alpha_1^{4/3}\tilde\alpha_2^{2/3}\lambda^{-2/3}\log^{4\gamma/3}(W\lambda^{-1/2}\tilde\alpha_1), \tilde\alpha_1^{3/2}\tilde\alpha_3^{1/2}\lambda^{-1/2}\log^{3\gamma/2}(W\lambda^{-1/2}\tilde\alpha_1), \lambda^{-1}\tilde\alpha_1^2\log^{2\gamma}(W\lambda^{-1/2}\tilde\alpha_1)\\
&\le W^2\cdot\max\{\lambda^{-1}\tilde\alpha_1^2\log^{2\gamma}(W^2\lambda^{-1}\tilde\alpha_1^2), \tilde\alpha_2^2, \lambda\tilde\alpha_3^2\}\\
&=: N^\star.
\end{aligned}$$

Here the first equation follows from the fact that
$\mathsf{Tr}(G_w H_w^{-2}) = \mathsf{Tr}(H_w^{-1/2}G_w H_w^{-1/2}H_w^{-1}) \le \|H_w^{-1}\|_2\mathsf{Tr}(H_w^{-1/2}G_w H_w^{-1/2}) = \|H_w^{-1}\|_2\mathsf{Tr}(G_w H_w^{-1})$.
To summarize, for any $\delta\in(0,1)$ and any $n \ge c\max\{N^\star\log(d/\delta), N(\delta), N'(\delta)\}$, with probability at least $1-3\delta$, we have

$$R_M(\beta_{\mathsf{MWLE}}) \le c\frac{\mathsf{Tr}(G_w H_w^{-1})\log\frac{d}{\delta}}{n}.$$

□

In the following, we prove Lemma E.2 and E.3.

Proof of Lemma E.2

Proof of Lemma E.2. For notation simplicity, we denote $g := \nabla\ell_n^w(\beta^\star) - \mathbb{E}_{\mathbb{P}_S}[\nabla\ell_n^w(\beta^\star)]$. Note that

$$\begin{aligned}
V &= n\cdot\mathbb{E}[\|A(\nabla\ell_n^w(\beta^\star) - \mathbb{E}[\nabla\ell_n^w(\beta^\star)])\|_2^2]\\
&= n\cdot\mathbb{E}[\nabla\ell_n^w(\beta^\star)^T A^T A\nabla\ell_n^w(\beta^\star)]\\
&= n\cdot\mathbb{E}[\mathsf{Tr}(A\nabla\ell_n^w(\beta^\star)\nabla\ell_n^w(\beta^\star)^T A^T)]\\
&= \mathsf{Tr}(AG_wA^T).
\end{aligned}$$

By taking $A = H_w^{-1}$ in Assumption C.2, for any δ and any $n > N(\delta)$, we have with probability at least $1-\delta$:

$$\begin{aligned}
\|H_w^{-1}g\|_2 &\le c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-2})\log\frac{d}{\delta}}{n}} + WB_1\|H_w^{-1}\|_2\log^\gamma\left(\frac{WB_1\|H_w^{-1}\|_2}{\sqrt{\mathsf{Tr}(G_w H_w^{-2})}}\right)\frac{\log\frac{d}{\delta}}{n}\\
&= c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-2})\log\frac{d}{\delta}}{n}} + WB_1\|H_w^{-1}\|_2\log^\gamma(W\lambda^{-1/2}\tilde\alpha_1)\frac{\log\frac{d}{\delta}}{n} \qquad (53)
\end{aligned}$$

$$\left\|\nabla^2\ell_n^w(\beta^\star) - \mathbb{E}[\nabla^2\ell_n^w(\beta^\star)]\right\|_2 \le WB_2\sqrt{\frac{\log\frac{d}{\delta}}{n}}. \qquad (54)$$

Let event $\tilde{A} := \{(53), (54) \text{ holds}\}$ and $\tilde{A}' := \{\ell_n^w(\cdot) \text{ has a unique local minimum, which is also global minimum}\}$. By Assumption C.2 and Assumption C.4, it then holds for any δ and any $n \geq \max\{N(\delta), N'(\delta)\}$ that $\mathbb{P}(\tilde{A} \cap \tilde{A}') \geq 1 - 2\delta$. Under the event $\tilde{A} \cap \tilde{A}'$, we have the following Taylor expansion:

$$\begin{aligned}
\ell_n^w(\beta) - \ell_n^w(\beta^\star) &\overset{\text{by Assumption C.1, C.3}}{\leq} (\beta - \beta^\star)^T \nabla \ell_n^w(\beta^\star) + \frac{1}{2}(\beta - \beta^\star)^T \nabla^2 \ell_n^w(\beta^\star)(\beta - \beta^\star) + \frac{WB_3}{6}\|\beta - \beta^\star\|_2^3 \\
&\overset{\mathbb{E}_{\mathbb{P}_S}[\nabla \ell_n^w(\beta^\star)]=0}{=} (\beta - \beta^\star)^T g + \frac{1}{2}(\beta - \beta^\star)^T \nabla^2 \ell_n^w(\beta^\star)(\beta - \beta^\star) + \frac{WB_3}{6}\|\beta - \beta^\star\|_2^3 \\
&\overset{\text{by (54)}}{\leq} (\beta - \beta^\star)^T g + \frac{1}{2}(\beta - \beta^\star)^T H_w(\beta - \beta^\star) + WB_2\sqrt{\frac{\log\frac{d}{\delta}}{n}}\|\beta - \beta^\star\|_2^2 + \frac{WB_3}{6}\|\beta - \beta^\star\|_2^3 \\
&\overset{\Delta_\beta := \beta - \beta^\star}{=} \Delta_\beta^T g + \frac{1}{2}\Delta_\beta^T H_w \Delta_\beta + WB_2\sqrt{\frac{\log\frac{d}{\delta}}{n}}\|\Delta_\beta\|_2^2 + \frac{WB_3}{6}\|\Delta_\beta\|_2^3 \\
&= \frac{1}{2}(\Delta_\beta - z)^T H_w(\Delta_\beta - z) - \frac{1}{2}z^T H_w z + WB_2\sqrt{\frac{\log\frac{d}{\delta}}{n}}\|\Delta_\beta\|_2^2 + \frac{WB_3}{6}\|\Delta_\beta\|_2^3
\end{aligned} \tag{55}$$

where $z := -H_w^{-1} g$. Similarly

$$\ell_n^w(\beta) - \ell_n^w(\beta^\star) \geq \frac{1}{2}(\Delta_\beta - z)^T H_w(\Delta_\beta - z) - \frac{1}{2}z^T H_w z - WB_2\sqrt{\frac{\log\frac{d}{\delta}}{n}}\|\Delta_\beta\|_2^2 - \frac{WB_3}{6}\|\Delta_\beta\|_2^3. \tag{56}$$

Notice that $\Delta_{\beta^\star + z} = z$, by (53) and (55), we have

$$\begin{aligned}
&\ell_n^w(\beta^\star + z) - \ell_n^w(\beta^\star) \\
&\leq -\frac{1}{2}z^T H_w z + WB_2\sqrt{\frac{\log\frac{d}{\delta}}{n}}\left(c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-2})\log\frac{d}{\delta}}{n}} + WB_1\|H_w^{-1}\|_2 \log^\gamma(W\lambda^{-1/2}\tilde{\alpha}_1)\frac{\log\frac{d}{\delta}}{n}\right)^2 \\
&\quad + \frac{WB_3}{6}\left(c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-2})\log\frac{d}{\delta}}{n}} + WB_1\|H_w^{-1}\|_2 \log^\gamma(W\lambda^{-1/2}\tilde{\alpha}_1)\frac{\log\frac{d}{\delta}}{n}\right)^3 \\
&\leq -\frac{1}{2}z^T H_w z + 2c^2 WB_2 \mathsf{Tr}(G_w H_w^{-2})(\frac{\log\frac{d}{\delta}}{n})^{1.5} + 2W^3 B_1^2 B_2\|H_w^{-1}\|_2^2 \log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log\frac{d}{\delta}}{n})^{2.5} \\
&\quad + \frac{2}{3}c^3 WB_3 \mathsf{Tr}(G_w H_w^{-2})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5} + \frac{2}{3}W^4 B_1^3 B_3\|H_w^{-1}\|_2^3 \log^{3\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log\frac{d}{\delta}}{n})^3.
\end{aligned} \tag{57}$$

For any $\beta \in \mathbb{B}_{\beta^\star}(3c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-2})\log\frac{d}{\delta}}{n}})$, by (56), we have

$$\begin{aligned}
\ell_n^w(\beta) - \ell_n^w(\beta^\star) &\geq \frac{1}{2}(\Delta_\beta - z)^T H_w(\Delta_\beta - z) - \frac{1}{2}z^T H_w z \\
&\quad - 9c^2 WB_2 \mathsf{Tr}(G_w H_w^{-2})(\frac{\log\frac{d}{\delta}}{n})^{1.5} - \frac{9}{2}c^3 WB_3 \mathsf{Tr}(G_w H_w^{-2})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5}.
\end{aligned} \tag{58}$$

(58) - (57) gives

$$\begin{aligned}
&\ell_n^w(\beta) - \ell_n^w(\beta^\star + z) \\
&\geq \frac{1}{2}(\Delta_\beta - z)^T H_w(\Delta_\beta - z) \\
&- \Big(11c^2 WB_2 \mathsf{Tr}(G_w H_w^{-2})(\frac{\log\frac{d}{\delta}}{n})^{1.5} + \frac{31}{6}c^3 WB_3 \mathsf{Tr}(G_w H_w^{-2})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5} \\
&\quad + 2W^3 B_1^2 B_2\|H_w^{-1}\|_2^2 \log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log\frac{d}{\delta}}{n})^{2.5} + \frac{2}{3}W^4 B_1^3 B_3\|H_w^{-1}\|_2^3 \log^{3\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log\frac{d}{\delta}}{n})^3\Big)
\end{aligned} \tag{59}$$

Consider the ellipsoid

$$\mathcal{D} := \Bigg\{ \beta \in \mathbb{R}^d \,\Bigg|\, \frac{1}{2}(\Delta_\beta - z)^T H_w (\Delta_\beta - z) \leq 11c^2 W B_2 \mathsf{Tr}(G_w H_w^{-2}) (\frac{\log \frac{d}{\delta}}{n})^{1.5} + \frac{31}{6} c^3 W B_3 \mathsf{Tr}(G_w H_w^{-2})^{1.5} (\frac{\log \frac{d}{\delta}}{n})^{1.5} + 2W^3 B_1^2 B_2 \|H_w^{-1}\|_2^2 \log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log \frac{d}{\delta}}{n})^{2.5} + \frac{2}{3} W^4 B_1^3 B_3 \|H_w^{-1}\|_2^3 \log^{3\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log \frac{d}{\delta}}{n})^3 \Bigg\}$$

Then by (59), for any $\beta \in \mathbb{B}_{\beta^\star}(3c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-2}) \log \frac{d}{\delta}}{n}}) \cap \mathcal{D}^C$, we have

$$\ell_n^w(\beta) - \ell_n^w(\beta^\star + z) > 0. \tag{60}$$

Notice that by the definition of $\mathcal{D}$, using $\lambda_{\min}^{-1}(H_w) = \|H_w^{-1}\|_2$, we have for any $\beta \in \mathcal{D}$,

$$\begin{aligned}\|\Delta_\beta - z\|_2^2 \leq{}& 22c^2 \|H_w^{-1}\|_2 W B_2 \mathsf{Tr}(G_w H_w^{-2}) (\frac{\log \frac{d}{\delta}}{n})^{1.5} + \frac{31}{3} c^3 \|H_w^{-1}\|_2 W B_3 \mathsf{Tr}(G_w H_w^{-2})^{1.5} (\frac{\log \frac{d}{\delta}}{n})^{1.5} \\ &+ 4\|H_w^{-1}\|_2 W^3 B_1^2 B_2 \|H_w^{-1}\|_2^2 \log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log \frac{d}{\delta}}{n})^{2.5} \\ &+ \frac{4}{3} \|H_w^{-1}\|_2 W^4 B_1^3 B_3 \|H_w^{-1}\|_2^3 \log^{3\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log \frac{d}{\delta}}{n})^3.\end{aligned}$$

Thus for any $\beta \in \mathcal{D}$,

$$\begin{aligned}\|\Delta_\beta\|_2^2 \leq{}& 2(\|\Delta_\beta - z\|_2^2 + \|z\|_2^2) \\ \overset{\text{by}(53)}{\leq}{}& 44c^2 \|H_w^{-1}\|_2 W B_2 \mathsf{Tr}(G_w H_w^{-2}) (\frac{\log \frac{d}{\delta}}{n})^{1.5} + \frac{62}{3} c^3 \|H_w^{-1}\|_2 W B_3 \mathsf{Tr}(G_w H_w^{-2})^{1.5} (\frac{\log \frac{d}{\delta}}{n})^{1.5} \\ &+ 8\|H_w^{-1}\|_2 W^3 B_1^2 B_2 \|H_w^{-1}\|_2^2 \log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log \frac{d}{\delta}}{n})^{2.5} \\ &+ \frac{8}{3} \|H_w^{-1}\|_2 W^4 B_1^3 B_3 \|H_w^{-1}\|_2^3 \log^{3\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log \frac{d}{\delta}}{n})^3 \\ &+ 4c^2 \mathsf{Tr}(G_w H_w^{-2}) \frac{\log \frac{d}{\delta}}{n} + 4W^2 B_1^2 \|H_w^{-1}\|_2^2 \log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log \frac{d}{\delta}}{n})^2.\end{aligned}$$

To guarantee $\mathsf{Tr}(G_w H_w^{-2}) \frac{\log \frac{d}{\delta}}{n}$ is the leading term, we only need $\mathsf{Tr}(G_w H_w^{-2}) \frac{\log \frac{d}{\delta}}{n}$ to dominate the rest of the terms. Hence, if we further have $n \geq cN_1 \log(d/\delta)$, it then holds that

$$\|\Delta_\beta\|_2^2 \leq 9c^2 \mathsf{Tr}(G_w H_w^{-2}) \frac{\log \frac{d}{\delta}}{n},$$

i.e., $\beta \in \mathbb{B}_{\beta^\star}(3c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-2}) \log \frac{d}{\delta}}{n}})$. Here

$$\begin{aligned}N_1 := \max\Bigg\{ & W^2 B_2^2 \|H_w^{-1}\|_2^2, W^2 B_3^2 \mathsf{Tr}(G_w H_w^{-2}) \|H_w^{-1}\|_2^2, \left(\frac{W^3 B_1^2 B_2 \|H_w^{-1}\|_2^3 \log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)}{\mathsf{Tr}(G_w H_w^{-2})}\right)^{\frac{2}{3}}, \\ & \left(\frac{W^4 B_1^3 B_3 \|H_w^{-1}\|_2^4 \log^{3\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)}{\mathsf{Tr}(G_w H_w^{-2})}\right)^{\frac{1}{2}}, \frac{W^2 B_1^2 \|H_w^{-1}\|_2^2 \log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)}{\mathsf{Tr}(G_w H_w^{-2})} \Bigg\}.\end{aligned}$$

In other words, we show that $\mathcal{D} \subset \mathbb{B}_{\beta^\star}(3c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-2}) \log \frac{d}{\delta}}{n}})$. Recall that by (60), we know that for any $\beta \in \mathbb{B}_{\beta^\star}(3c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-2}) \log \frac{d}{\delta}}{n}}) \cap \mathcal{D}^C$,

$$\ell_n^w(\beta) - \ell_n^w(\beta^\star + z) > 0.$$

Note that $\beta^\star + z \in \mathcal{D}$. Hence there is a local minimum of $\ell_n^w(\beta)$ in $\mathcal{D}$. Under the event $\tilde{A}'$, we know that the global minimum of $\ell_n^w(\beta)$ is in $\mathcal{D}$, i.e.,

$$\beta_{\mathsf{MWLE}} \in \mathcal{D} \subset \mathbb{B}_{\beta^\star}(3c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-2})\log\frac{d}{\delta}}{n}}).$$

□

Proof of Lemma E.3

Proof of Lemma E.3. Let $\tilde{E} := \{\beta_{\mathsf{MWLE}} \in \mathcal{D} \subset \mathbb{B}_{\beta^\star}(\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-2})\log\frac{d}{\delta}}{n}})\}$. Then by the proof of Lemma E.2, for any $\delta \in (0,1)$ and any $n \geq c\max\{N_1\log(d/\delta), N(\delta), N'(\delta)\}$, we have $\mathbb{P}(\tilde{E}) \geq 1-2\delta$.

By taking $A = H_w^{-\frac{1}{2}}$ in Assumption C.2, for any $\delta \in (0,1)$ and any $n \geq N(\delta)$, with probability at least $1-\delta$, we have:

$$\begin{aligned}
\|H_w^{-\frac{1}{2}} g\|_2 &\leq c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-1})\log\frac{d}{\delta}}{n}} + WB_1\|H_w^{-\frac{1}{2}}\|_2 \log^\gamma\left(\frac{WB_1\|H_w^{-\frac{1}{2}}\|_2}{\sqrt{\mathsf{Tr}(G_w H_w^{-1})}}\right)\frac{\log\frac{d}{\delta}}{n} \\
&\leq c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-1})\log\frac{d}{\delta}}{n}} + WB_1\|H_w^{-\frac{1}{2}}\|_2 \log^\gamma\left(\frac{WB_1\|H_w^{-\frac{1}{2}}\|_2}{\sqrt{\mathsf{Tr}(G_w H_w^{-2})\|H_w^{-1}\|_2^{-1}}}\right)\frac{\log\frac{d}{\delta}}{n} \\
&= c\sqrt{\frac{\mathsf{Tr}(G_w H_w^{-1})\log\frac{d}{\delta}}{n}} + WB_1\|H_w^{-\frac{1}{2}}\|_2 \log^\gamma(W\lambda^{-1/2}\tilde{\alpha}_1)\frac{\log\frac{d}{\delta}}{n} \qquad (61)
\end{aligned}$$

We denote $\tilde{E}' := \{(61)\text{ holds}\}$. Then for any δ and any $n \geq c\max\{N_1(M)\log(d/\delta), N(\delta), N'(\delta)\}$, we have $\mathbb{P}(\tilde{E}\cap\tilde{E}') \geq 1-3\delta$.

Under $\tilde{E}\cap\tilde{E}'$, $\beta_{\mathsf{MWLE}} \in \mathcal{D}$, i.e.,

$$\begin{aligned}
&\frac{1}{2}(\Delta_{\beta_{\mathsf{MWLE}}} - z)^T H_w(\Delta_{\beta_{\mathsf{MWLE}}} - z) \\
&\leq 11c^2 WB_2\mathsf{Tr}(G_w H_w^{-2})(\frac{\log\frac{d}{\delta}}{n})^{1.5} + \frac{31}{6}c^3 WB_3\mathsf{Tr}(G_w H_w^{-2})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5} \\
&\quad + 2W^3B_1^2B_2\|H_w^{-1}\|_2^2\log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log\frac{d}{\delta}}{n})^{2.5} + \frac{2}{3}W^4B_1^3B_3\|H_w^{-1}\|_2^3\log^{3\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log\frac{d}{\delta}}{n})^3.
\end{aligned}$$

In other words,

$$\begin{aligned}
&\|H_w^{\frac{1}{2}}(\Delta_{\beta_{\mathsf{MWLE}}} - z)\|_2^2 \\
&\leq 22c^2 WB_2\mathsf{Tr}(G_w H_w^{-2})(\frac{\log\frac{d}{\delta}}{n})^{1.5} + \frac{31}{3}c^3 WB_3\mathsf{Tr}(G_w H_w^{-2})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5} \\
&\quad + 4W^3B_1^2B_2\|H_w^{-1}\|_2^2\log^{2\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log\frac{d}{\delta}}{n})^{2.5} + \frac{4}{3}W^4B_1^3B_3\|H_w^{-1}\|_2^3\log^{3\gamma}(W\lambda^{-1/2}\tilde{\alpha}_1)(\frac{\log\frac{d}{\delta}}{n})^3. \qquad (62)
\end{aligned}$$

Thus we have

$$
\begin{aligned}
&\|H_w^{\frac{1}{2}}(\beta_{\mathsf{MWLE}}-\beta^\star)\|_2^2\\
&=\|H_w^{\frac{1}{2}}\Delta_{\beta_{\mathsf{MWLE}}}\|_2^2\\
&=\|H_w^{\frac{1}{2}}(\Delta_{\beta_{\mathsf{MWLE}}}-z)+H_w^{\frac{1}{2}}z\|_2^2\\
&\le 2\|H_w^{\frac{1}{2}}(\Delta_{\beta_{\mathsf{MWLE}}}-z)\|_2^2+2\|H_w^{\frac{1}{2}}z\|_2^2\\
&=2\|H_w^{\frac{1}{2}}(\Delta_{\beta_{\mathsf{MWLE}}}-z))\|_2^2+2\|H_w^{-\frac{1}{2}}g\|_2^2\\
&\overset{\text{by(62)and(61)}}{\le} 4c^2\frac{\mathsf{Tr}(G_wH_w^{-1})\log\frac{d}{\delta}}{n}\\
&+44c^2WB_2\mathsf{Tr}(G_wH_w^{-2})(\frac{\log\frac{d}{\delta}}{n})^{1.5}+\frac{62}{3}c^3WB_3\mathsf{Tr}(G_wH_w^{-2})^{1.5}(\frac{\log\frac{d}{\delta}}{n})^{1.5}\\
&\quad+8W^3B_1^2B_2\|H_w^{-1}\|_2^2\log^{2\gamma}(W\lambda^{-1/2}\tilde\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^{2.5}+\frac{8}{3}W^4B_1^3B_3\|H_w^{-1}\|_2^3\log^{3\gamma}(W\lambda^{-1/2}\tilde\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^3\\
&\quad+4W^2B_1^2\|H_w^{-1}\|_2\log^{2\gamma}(W\lambda^{-1/2}\tilde\alpha_1)(\frac{\log\frac{d}{\delta}}{n})^2 \qquad (63)
\end{aligned}
$$

To guarantee $\frac{\mathsf{Tr}(G_wH_w^{-1})\log\frac{d}{\delta}}{n}$ is the leading term, we only need $\frac{\mathsf{Tr}(G_wH_w^{-1})\log\frac{d}{\delta}}{n}$ to dominate the rest of the terms. Hence, if we further have $n\ge cN_2\log(d/\delta)$, we have

$$\|H_w^{\frac{1}{2}}(\beta_{\mathsf{MWLE}}-\beta^\star)\|_2^2\le 9c^2\frac{\mathsf{Tr}(G_wH_w^{-1})\log\frac{d}{\delta}}{n}.$$

Here

$$
\begin{aligned}
N_2:=\max\Bigg\{&\left(\frac{WB_2\mathsf{Tr}(G_wH_w^{-2})}{\mathsf{Tr}(G_wH_w^{-1})}\right)^2,\left(\frac{WB_3\mathsf{Tr}(G_wH_w^{-2})^{1.5}}{\mathsf{Tr}(G_wH_w^{-1})}\right)^2,\left(\frac{W^3B_1^2B_2\|H_w^{-1}\|_2^2\log^{2\gamma}(W\lambda^{-1/2}\tilde\alpha_1)}{\mathsf{Tr}(G_wH_w^{-1})}\right)^{\frac{2}{3}},\\
&\left(\frac{W^4B_1^3B_3\|H_w^{-1}\|_2^3\log^{3\gamma}(W\lambda^{-1/2}\tilde\alpha_1)}{\mathsf{Tr}(G_wH_w^{-1})}\right)^{\frac{1}{2}},\frac{W^2B_1^2\|H_w^{-1}\|_2\log^{2\gamma}(W\lambda^{-1/2}\tilde\alpha_1)}{\mathsf{Tr}(G_wH_w^{-1})}\Bigg\}.
\end{aligned}
$$

To summarize, we show that for any δ and any $n\ge c\max\{N_1\log(d/\delta),N_2\log(d/\delta),N(\delta),N'(\delta)\}$, with probability at least $1-3\delta$, we have

$$\|H_w^{\frac{1}{2}}(\beta_{\mathsf{MWLE}}-\beta^\star)\|_2^2\le 9c^2\frac{\mathsf{Tr}(G_wH_w^{-1})\log\frac{d}{\delta}}{n}.$$

□

E.3 Proofs for Theorem B.3

Proof of Theorem B.3. For any $W>1$, we construct $\mathbb{P}_S(X)$, $\mathbb{P}_T(X)$, $\mathcal{M}$ and $\mathcal{F}$ as follows. We define $\mathbb{P}_T(X):=\mathsf{Uniform}(\mathbb{B}(1))$ and $\mathbb{P}_S(X):=\mathsf{Uniform}(\mathbb{B}(W^{\frac{1}{d}}))$, where $\mathbb{B}(1)$ and $\mathbb{B}(W^{\frac{1}{d}})$ are d-dimensional balls centered around the original with radius 1 and $W^{\frac{1}{d}}$, respectively. For notation simplicity, we denote $Q:=\mathbb{B}(1)$ and $P:=\mathbb{B}(W^{\frac{1}{d}})$ in the following. The density ratios is then given by

$$w(x):=\frac{d\mathbb{P}_T(x)}{d\mathbb{P}_S(x)}=\begin{cases}W & x\in Q\\ 0 & x\notin Q\end{cases},$$

which is upper bounded by W. We further have

$$\mathcal{I}_S(\beta)=\mathbb{E}_{x\sim\mathbb{P}_S(X)}[xx^T]=\frac{W^{\frac{2}{d}}}{3d}I_d\succ 0,\ \mathcal{I}_T(\beta)=\mathbb{E}_{x\sim\mathbb{P}_T(X)}[xx^T]=\frac{1}{3d}I_d\succ 0.$$

Let $\mathcal{F}:=\{f(y\,|\,x;\beta)\,|\,\beta\in\mathbb{R}^d\}$ be the linear regression class, i.e., $-\log f(y\,|\,x;\beta)=(\log 2\pi)/2+(y-x^T\beta)^2/2$. We assume the true conditional distribution belongs to a class $\mathcal{M}$ that is defined as

$$\mathcal{M}:=\left\{Y\,|\,X\ \mathsf{s.t}\ p(y\,|\,x)=f(y\,|\,x;\beta_1^\star)\mathbf{1}_{\{x\in Q\}}+f(y\,|\,x;\beta_2^\star)\mathbf{1}_{\{x\in P\setminus Q\}},\beta_1^\star,\beta_2^\star\in\mathbb{B}_{\beta_0}(B)\right\}$$

for some $\beta_0 \in \mathbb{R}^d$ and $B > 0$. We utilize the function class $\mathcal{F}$ to approximate the true conditional density function, which subsequently results in model mis-specification. In the sequel, we will show the lower bound of excess risk for any estimators under this model class $\mathcal{M}$.

Fix any ground truth model $M \in \mathcal{M}$, that is, we are assuming the true conditional distribution follows the form:

$$p(y \,|\, x) = f(y \,|\, x; \beta_1^\star)\mathbf{1}_{\{x \in Q\}} + f(y \,|\, x; \beta_2^\star)\mathbf{1}_{\{x \in P \setminus Q\}},$$

where $\beta_1^\star$ and $\beta_2^\star$ are arbitrarily chosen fixed points from $\mathbb{B}_{\beta_0}(B)$. Note that the model is actually well-specified on the target domain. Hence the optimal fit on the target is given by

$$\beta^\star(M) = \arg\min_\beta \mathbb{E}_{(x,y)\sim\mathbb{P}_T(X,Y)}[\ell(x, y, \beta)] = \beta_1^\star.$$

For linear regression, it is easy to verify that Assumption B.2, B.3 and B.4 hold. Let R_0 and R_1 be the parameters chosen by Lemma C.5. Then similar to the proofs of Theorem 3.2, we have

$$\begin{aligned}
&\inf_{\hat{\beta}} \sup_{M\in\mathcal{M}} \mathbb{E}_{(x_i,y_i)\sim\mathbb{P}_S(X,Y)} \left[R_M(\hat{\beta})\right] \\
&= \inf_{\hat{\beta}} \sup_{\beta_1^\star,\beta_2^\star\in\mathbb{B}_{\beta_0}(B)} \mathbb{E}_{(x_i,y_i)\sim\mathbb{P}_S(X,Y)} \left[R_{\beta_1^\star}(\hat{\beta})\right] \\
&\geq \inf_{\hat{\beta}} \sup_{\beta_1^\star,\beta_2^\star\in\mathbb{B}_{\beta_0}(R_1)} \mathbb{E}_{(x_i,y_i)\sim\mathbb{P}_S(X,Y)} \left[R_{\beta_1^\star}(\hat{\beta})\right] \\
&\geq \inf_{\hat{\beta}\in\mathbb{B}_{\beta_0}(R_0)} \sup_{\beta_1^\star,\beta_2^\star\in\mathbb{B}_{\beta_0}(R_1)} \mathbb{E}_{(x_i,y_i)\sim\mathbb{P}_S(X,Y)} \left[R_{\beta_1^\star}(\hat{\beta})\right] \\
&\geq \frac{1}{4} \inf_{\hat{\beta}\in\mathbb{B}_{\beta_0}(R_0)} \sup_{\beta_1^\star,\beta_2^\star\in\mathbb{B}_{\beta_0}(R_1)} \mathbb{E}_{(x_i,y_i)\sim\mathbb{P}_S(X,Y)} \left[(\hat{\beta} - \beta_1^\star)^T \mathcal{I}_T(\beta_0)(\hat{\beta} - \beta_1^\star)\right] \\
&\geq \frac{1}{4} \inf_{\hat{\beta}\in\mathbb{B}_{\beta_0}(R_0)} \sup_{\beta_1^\star,\beta_2^\star\in C_{\beta_0}(\frac{R_1}{\sqrt{d}})} \mathbb{E}_{(x_i,y_i)\sim\mathbb{P}_S(X,Y)} \left[(\hat{\beta} - \beta_1^\star)^T \mathcal{I}_T(\beta_0)(\hat{\beta} - \beta_1^\star)\right] \\
&= \frac{1}{4} \inf_{\hat{\beta}\in\mathbb{B}_{\beta_0}(R_0)} \sup_{[\beta_1^{\star T},\beta_2^{\star T}]\in C_{[\beta_0^T,\beta_0^T]}(\frac{R_1}{\sqrt{d}})} \mathbb{E}_{(x_i,y_i)\sim\mathbb{P}_S(X,Y)} \left[(\hat{\beta} - \beta_1^\star)^T \mathcal{I}_T(\beta_0)(\hat{\beta} - \beta_1^\star)\right] \qquad (64)
\end{aligned}$$

By Theorem 1 in Gill & Levit (1995) (multivariate van Trees inequality) with $\psi(\beta_1^\star, \beta_2^\star) = \beta_1^\star$, $C(\beta_1^\star, \beta_2^\star) \equiv C := [WI_d, 0] \in \mathbb{R}^{d\times 2d}$ and $B(\beta_1^\star, \beta_2^\star) \equiv B := \mathcal{I}_T^{-1}(\beta_0)$, we have for any estimator $\hat{\beta}$ and good prior density λ that supported on $C_{[\beta_0^T,\beta_0^T]}(\frac{R_1}{\sqrt{d}})$,

$$\mathbb{E}_{[\beta_1^{\star T},\beta_2^{\star T}]\sim\lambda} \mathbb{E}_{(x_i,y_i)\sim\mathbb{P}_S(X,Y)} \left[(\hat{\beta} - \beta_1^\star)^T \mathcal{I}_T(\beta_0)(\hat{\beta} - \beta_1^\star)\right] \geq \frac{(Wd)^2}{2nWd + \tilde{\mathcal{I}}(\lambda)},$$

where

$$\tilde{\mathcal{I}}(\lambda) = \int_{C_{[\beta_0^T,\beta_0^T]}(\frac{R_1}{\sqrt{d}})} \left(\sum_{i,j,k,\ell} B_{ij} C_{ik} C_{j\ell} \frac{\partial}{\partial\tilde{\beta}_k}\lambda(\tilde{\beta}) \frac{\partial}{\partial\tilde{\beta}_\ell}\lambda(\tilde{\beta}) \right) \frac{1}{\lambda(\tilde{\beta})} d\tilde{\beta}.$$

Let $\tilde{\beta}_0 = [\beta_{0,1}, \ldots, \beta_{0,d}, \beta_{0,1}, \ldots, \beta_{0,d}]^T$, $\tilde{\beta} = [\beta_1, \ldots, \beta_{2d}]^T$ and

$$f_i(x) := \frac{\pi\sqrt{d}}{4R_1} \cos\left(\frac{\pi\sqrt{d}}{2R_1}(x - \tilde{\beta}_{0,i}) \right), \; i = 1, \ldots, 2d.$$

We define the prior density as

$$\lambda(\tilde{\beta}) := \begin{cases} \Pi_{i=1}^{2d} f_i(\beta_i) & \tilde{\beta} \in C_{[\beta_0^T,\beta_0^T]}(\frac{R_1}{\sqrt{d}}) \\ 0 & \tilde{\beta} \notin C_{[\beta_0^T,\beta_0^T]}(\frac{R_1}{\sqrt{d}}) \end{cases}.$$

Then following the same argument as in the proof of Lemma C.6, we have

$$\tilde{\mathcal{I}}(\lambda) = \frac{\pi^2 d}{R_1^2} \mathsf{Tr}(BCC^T) = \frac{\pi^2 W^2 d}{R_1^2} \mathsf{Tr}(\mathcal{I}_T^{-1}(\beta_0)).$$

As a result, for any estimator $\hat{\beta}$, we have

$$\begin{aligned}&\mathbb{E}_{[\beta_1^{\star T},\beta_2^{\star T}]\sim\lambda}\mathbb{E}_{(x_i,y_i)\sim\mathbb{P}_S(X,Y)}\left[(\hat{\beta}-\beta_1^\star)^T\mathcal{I}_T(\beta_0)(\hat{\beta}-\beta_1^\star)\right]\\&\geq\frac{(Wd)^2}{2nWd+\frac{\pi^2W^2d}{R_1^2}\mathsf{Tr}(\mathcal{I}_T^{-1}(\beta_0))},\end{aligned}$$

which implies

$$\begin{aligned}&\sup_{[\beta_1^{\star T},\beta_2^{\star T}]\in C_{[\beta_0^T,\beta_0^T]}(\frac{R_1}{\sqrt{d}})}\mathbb{E}_{(x_i,y_i)\sim\mathbb{P}_S(X,Y)}\left[(\hat{\beta}-\beta_1^\star)^T\mathcal{I}_T(\beta_0)(\hat{\beta}-\beta_1^\star)\right]\\&\geq\mathbb{E}_{[\beta_1^{\star T},\beta_2^{\star T}]\sim\lambda}\mathbb{E}_{(x_i,y_i)\sim\mathbb{P}_S(X,Y)}\left[(\hat{\beta}-\beta^\star)^T\mathcal{I}_T(\beta_0)(\hat{\beta}-\beta^\star)\right]\\&\geq\frac{(Wd)^2}{2nWd+\frac{\pi^2W^2d}{R_1}\mathsf{Tr}(\mathcal{I}_T^{-1}(\beta_0))}.\end{aligned}\tag{65}$$

Combine (64) and (65), we have

$$\inf_{\hat{\beta}}\sup_{M\in\mathcal{M}}\mathbb{E}_{(x_i,y_i)\sim\mathbb{P}_S(X,Y)}\left[R_M(\hat{\beta})\right]\geq\frac{1}{4}\cdot\frac{(Wd)^2}{2nWd+\frac{\pi^2W^2d}{R_1}\mathsf{Tr}(\mathcal{I}_T^{-1}(\beta_0))}\gtrsim\frac{Wd}{n}$$

when n is sufficiently large.

Recall that

$$H_w(M)=\mathbb{E}_{(x,y)\sim\mathbb{P}_T(X,Y)}\left[\nabla^2\ell(x,y,\beta^\star(M))\right]=\mathbb{E}_{(x,y)\sim\mathbb{P}_T(X,Y)}\left[\nabla^2\ell(x,y,\beta_1^\star)\right]=\mathcal{I}_T(\beta_1^\star).$$

and by the definition of $w(x)$, we further have

$$\begin{aligned}G_w(M)&=\mathbb{E}_{(x,y)\sim\mathbb{P}_S(X,Y)}\left[w(x)^2\nabla\ell(x,y,\beta^\star(M))\nabla\ell(x,y,\beta^\star(M))^T\right]\\&=\mathbb{E}_{(x,y)\sim\mathbb{P}_T(X,Y)}\left[w(x)\nabla\ell(x,y,\beta^\star(M))\nabla\ell(x,y,\beta^\star(M))^T\right]\\&=W\mathbb{E}_{(x,y)\sim\mathbb{P}_T(X,Y)}\left[\nabla\ell(x,y,\beta^\star(M))\nabla\ell(x,y,\beta^\star(M))^T\right]\\&=W\mathbb{E}_{(x,y)\sim\mathbb{P}_T(X,Y)}\left[\nabla\ell(x,y,\beta_1^\star)\nabla\ell(x,y,\beta_1^\star)^T\right]\\&=W\mathcal{I}_T(\beta_1^\star).\end{aligned}$$

Therefore $\mathsf{Tr}(G_w(M)H_w(M)^{-1})=Wd$, which gives the desired result. What remains is to verify that $\mathcal{M}$ satisfies Assumption C.1, C.2, C.3 and C.4. Assumption C.1 is trivially satisfied. For Assumption C.2 and C.3, notice that

$$\begin{aligned}\nabla\ell(x,y,\beta)&=-x(y-x^T\beta),\\\nabla^2\ell(x,y,\beta)&=xx^T,\\\nabla^3\ell(x,y,\beta)&=0.\end{aligned}$$

and

$$w(x):=\frac{d\mathbb{P}_T(x)}{d\mathbb{P}_S(x)}=\begin{cases}W & x\in Q\\0 & x\notin Q\end{cases},$$

By the definition of $\mathcal{M}$, we can write the distribution of y as

$$y_i=\begin{cases}x_i^T\beta_1^\star+\epsilon_i & x_i\in Q\\x_i^T\beta_2^\star+\epsilon_i & x_i\notin Q\end{cases},$$

where ϵ_i is a $\mathcal{N}(0,1)$ noise independent of all x_i's. Therefore let $u_i:=Aw(x_i)\nabla\ell(x_i,y_i,\beta^\star(M))$, we have

$$u_i=\begin{cases}-WAx_i\epsilon_i & x_i\in Q\\0 & x_i\notin Q\end{cases},$$

which indicates that $\|u_i\|$ is $\|A\|W$-subgaussian. Therefore by Lemma F.1, the vector concentration in Assumption C.2 is satisfied with $\gamma = 0.5$, $B_1 = 1$. For the matrix concentration, notice that

$$w(x_i)\nabla^2\ell(x_i, y_i, \beta^\star(M)) = \begin{cases} Wx_ix_i^T & x_i \in Q \\ 0 & x_i \notin Q \end{cases},$$

therefore my matrix Hoeffding, $\|w(x_i)\nabla^2\ell(x_i, y_i, \beta^\star(M))\|_2 \le W$, thus the matrix concentration in Assumption C.2 is satisfied with $B_2 = 1$. Further more, $N(\delta) = 0$ is enough for satisfying Assumption C.2.

Assumption C.3 is satisfied with $B_3 = 0$ since $\nabla^3\ell(x, y, \beta) = 0$.

For Assumption C.4, we can prove that it is satisfied with $N'(\delta) = \max\{8W \log \frac{1}{\delta}, 2dW\}$. This is because,

$$\begin{aligned}
\mathbb{P}(\nabla^2\ell_n^w(\beta) \succ 0 \text{ for all } \beta) &= \mathbb{P}(\frac{W}{n}\sum_{i=1}^n x_ix_i^T\mathbb{I}_{x_i\in Q} \succ 0) \\
&\ge \mathbb{P}(\#\{x_i \in Q\} > d) \\
&= 1 - \mathbb{P}(\#\{x_i \in Q\} \le d) \\
&\overset{\text{by Chernoff bound}}{\ge} 1 - \mathsf{exp}(-\frac{\mu}{2}(1 - \frac{d}{\mu})^2) \\
&\ge 1 - \delta,
\end{aligned}$$

where $\mu := \frac{n}{W}$, and the last inequality hold when $n \ge N'(\delta)$. Therefore when $n \ge N'(\delta)$, with probability at least $1 - \delta$, ℓ_n^w is strictly convex, therefore has a unique local minimum which is also the global minimum. □

F Auxiliaries

In this section, we present several auxiliary lemmas and propositions.

F.1 Concentration for gradient and Hessian

The following lemma gives a generic version of Bernstein inequality for vectors.

Lemma F.1. *Let $u, u_1, \cdots, u_n$ be i.i.d. mean-zero random vectors. We denote $V = \mathbb{E}[\|u\|_2^2]$ and*

$$B_u^{(\alpha)} := \inf\{t > 0 : \mathbb{E}[\mathsf{exp}(\|u\|^\alpha/t^\alpha)] \le 2\}, \quad \alpha \ge 1.$$

Suppose $B_u^{(\alpha)} < \infty$ for some $\alpha \ge 1$. Then there exists an absolute constant $c > 0$ such that for all $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\left\|\frac{1}{n}\sum_{i=1}^n u_i\right\|_2 \le c\left(\sqrt{\frac{V\log\frac{d}{\delta}}{n}} + B_u^{(\alpha)}\left(\log\frac{B_u^{(\alpha)}}{\sqrt{V}}\right)^{1/\alpha}\frac{\log\frac{d}{\delta}}{n}\right).$$

Proof. See Proposition 2 in Koltchinskii et al. (2011) for the proof. □

The following proposition shows that when gradient and Hessian are bounded or sub-Gaussian (sub-exponential), Assumption A.1 is naturally satisfied.

Proposition F.2. *If $\|\nabla\ell(x_i, y_i, \beta^\star)\|_2 \le b_1$ for all $i \in [n]$, then the vector concentration (5) is satisfied with $B_1 = b_1$ and $\gamma = 0$. Alternatively, if $\|\nabla\ell(x_i, y_i, \beta^\star)\|_2$ is b_1-subgaussian, then (5) is satisfied with $B_1 = b_1$ and $\gamma = 1/2$. When $\|\nabla\ell(x_i, y_i, \beta^\star)\|_2$ is b_1-subexponential, then (5) is satisfied with $B_1 = b_1$ and $\gamma = 1$. For the Hessian concenntration, if $\|\nabla^2\ell(x_i, y_i, \beta^\star)\|_2 \le b_2$ for all $i \in [n]$, then (6) is satisfied with $B_2 = b_2$.*

Proof. The vector concentration (5) is a direct proposition of Lemma F.1. The Hessian concentration (6) is a direct consequence of matrix Hoeffiding inequality. □