mathematics

MDPI

Article

On a Low-Rank Matrix Single-Index Model

The Tien Mai

Department of Mathematical Sciences, Norwegian University of Science and Technology, 7034 Trondheim, Norway; the.t.mai@ntnu.no

Abstract: In this paper, we conduct a theoretical examination of a low-rank matrix single-index model. This model has recently been introduced in the field of biostatistics, but its theoretical properties for jointly estimating the link function and the coefficient matrix have not yet been fully explored. In this paper, we make use of the PAC-Bayesian bounds technique to provide a thorough theoretical understanding of the joint estimation of the link function and the coefficient matrix. This allows us to gain a deeper insight into the properties of this model and its potential applications in different fields.

Keywords: low-rank matrix; single-index model; PAC-Bayes bounds; optimal rate; oracle inequality

MSC: 62G05; 62C20

1. Introduction

In this study, we investigate a particular type of single-index model, where the response variable, denoted by Y, is a real number and the covariate matrix, represented by X, is a matrix of real numbers with dimensions of $d \times d$. The model is defined in Equation (1) as

$$Y = f^\star(\langle X, B^\star \rangle) + \epsilon \tag{1}$$

In this equation, $\langle X, B^\star \rangle = \text{trace}(X^\top B^\star)$ represents the inner product between matrices X and $B^\star$, where $B^\star$ is an unknown coefficient matrix with dimensions of $d \times d$. The link function $f^\star$ is an unknown univariate measurable function. The noise term, represented by ϵ, is assumed to have a mean of 0 and is independent of the covariate X.

In line with the recent research presented in [1,2], we make the assumption that the coefficient matrix $B^\star$ is a symmetric, low-rank matrix with $\text{rank}(B^\star) < d$. Additionally, in order to ensure the uniqueness of the model, we impose the condition that the Frobenius norm of $B^\star$ is equal to 1, i.e., $\|B^\star\|_F = 1$.

Previous studies have been conducted on a similar model to the one presented in this paper, where the unknown coefficient matrix $B^\star$ is assumed to have sparse elements. In particular, the work of [1] in the field of biostatistics has been used to examine the correlation between a response variable and the functional connectivity associated with a certain brain region. Additionally, recent research by [2] has focused on developing methods for estimating the unknown low-rank matrix $B^\star$ by using implicit regularization techniques.

The model discussed in this paper can be thought of as a nonparametric version of the trace regression model that has been previously proposed in the literature, specifically in the works in [3–5]. This trace regression model utilizes the identity function as the link function, and encompasses a diverse array of statistical models, including but not limited to reduced rank regression, matrix completion, and linear regression.

The single-index model is a versatile extension of the linear model, which offers a natural interpretation. This model only changes in the direction of the parameter (vector/matrix), and the nature of this change is depicted by the link function $f^\star$. This has been the subject of extensive research in the literature, with various studies exploring its

check for updates

Citation: Mai, T.T. On a Low-Rank Matrix Single-Index Model. *Mathematics* **2023**, *11*, 2065. https://doi.org/10.3390/math11092065

Academic Editors: Huiming Zhang and Ting Yan

Received: 17 March 2023
Revised: 25 April 2023
Accepted: 25 April 2023
Published: 26 April 2023

Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

applications and extensions in various fields. Examples of such works include [1,6–14]. These studies have demonstrated the versatility and utility of the single-index model in a wide range of contexts, making it a valuable tool for researchers in various fields.

Definition 1. *Let S_1^d denote the set of all symmetric matrix $B \in \mathbb{R}^{d\times d}$ such that $\|B\|_F = 1$.*

Given the covariates $\{X_i\}_{i=1}^n$, the response variables $\{Y_i\}_{i=1}^n$ are i.i.d. generated from model (1). We define the expected risk for any measurable $f : \mathbb{R} \to \mathbb{R}$ and $B \in S_1^d$ as

$$R(B, f) = \mathbb{E}\Big[(Y - f(\langle X, B\rangle)^2)\Big]$$

and denote the empirical counterpart of $R(f, B)$ by

$$r_n(B, f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\langle X_i, B\rangle)^2).$$

In this research, we examine the forecasting abilities of the model. More specifically, we consider a pair (f, B) to have comparable predictive performance to $(f^\star, B^\star)$ if the difference between $R(B, f)$ and $R(B^\star, f^\star)$ is minimal.

Our approach in this work is built on the PAC-Bayesian bound technique, which is a powerful tool for obtaining oracle inequalities bounds [15]. Similar to Bayesian analysis, one important aspect of a PAC-Bayesian bound is specifying a prior distribution over the parameter space. In our approach, we adopt the prior distribution for the link function from the reference [11], while the prior distribution for the matrix parameter B is inspired by the eigen decomposition of the matrix. The specifics of our approach and the details of the prior distributions we chose are discussed in the next section. The use of the PAC-Bayesian bound technique in combination with carefully chosen prior distributions allows us to obtain reliable and accurate estimates of the unknown parameters in our model.

2. Main Result

2.1. Method

We make an additional assumption in our model (1) that $\mathbb{E}[\epsilon|X] = 0$, and the following conditional moment assumptions on the noise ϵ are assumed.

Assumption 1. *We assume that there exist two constants $\sigma > 0$ and $L > 0$, such that for all integers $s \geq 2$,*

$$\mathbb{E}[|\epsilon|^s|X] \leq \frac{s!}{2}\sigma^2 L^{s-2}.$$

Remark 1. *The assumption stated above implies that the noise term in our model follows a subexponential distribution. This class of distributions includes, for example, Gaussian noise or bounded noise, as discussed in [16]. In simpler terms, this means that the noise term in our model is characterized by a rate of decay that is slower than that of an exponential distribution. This assumption is critical for the application of our approach, as it allows us to obtain accurate and reliable estimates of the unknown parameters under a wide range of noise conditions. This is an important consideration, as the presence of noise can have a significant impact on the accuracy of the estimates obtained from our model. By assuming that the noise follows a subexponential distribution, we can be confident that our estimates are robust to the presence of noise.*

In addition to the assumptions stated previously, it is also necessary to assume that the covariate matrix X is almost surely bounded by a constant. Additionally, the unknown link function $f^\star$ is also assumed to be bounded by some known positive constant. To make this more precise, we use the notation $\|X\|_\infty$ to represent its supremum norm and $\|f^\star\|_\infty$ to

denote its functional supremum norm over the interval $[-1,1]$. Based on these definitions, we make the following assumption:

Assumption 2. *We assume that* $\|X\|_\infty \leq 1$ *a.s. and* $\exists\, C \geq 1$, *such that* $\|f^\star\|_\infty \leq C$.

In order to present the technical proofs in the clearest and simplest manner, we did not attempt to find the best constant used in the proofs. Specifically, the condition that $C \geq 1$ is just convenient for the proofs in nature, and it could be eliminated by using $\max[C,1]$ in the proofs.

The link function $f^\star$ is approximately estimated through a given specific countable set of measurable functions (dictionary) $\{\varphi_k\}_{k=1}^\infty$. For this purpose, the set of finite linear combinations of functions from the dictionary is utilized, and we denote this vector space by $\mathcal{F}$. We assume that each element φ_k in the dictionary is defined on the interval $[-1,1]$ and takes values within the range $[-1,1]$.

Assumption 3. *For the sake of simplicity, we assume that the basic functions are differentiable and there exists some constant* $C_\phi > 0$, *such that*

$$\|\varphi_k'\|_\infty \leq kC_\phi.$$

An example of such a collection of functions is the system of non-normalized trigonometric functions, where

$$\varphi_1(t) = 1, \varphi_{2k}(t) = \cos(\pi k t), \varphi_{2k+1}(t) = \sin(\pi k t), k = 1,2,\ldots$$

satisfy this assumption. This assumption on the dictionary functions enables us to approximate the unknown link function $f^\star$ with a finite linear combination of these functions.

Our approach is inspired by the work of [11], where the authors explored the PAC-Bayesian approach in [15] for a sparse-vector single-index model. The method needs to first specify a distribution π on $\mathcal{S}_1^d \times \mathcal{F}$, similar to the prior distribution in Bayesian analysis. This prior distribution in our framework should enforce the characteristics of the underlying link function and the parameter matrix. In this work, we consider the following prior distribution:

$$\mathrm{d}\pi(B,f) = \mathrm{d}\mu(B)\mathrm{d}\nu(f),$$

in other words, it means that the prior distribution of the index matrix and the prior distribution over the link functions are assumed to be independent.

In this study, the matrix B is treated as a symmetric matrix and can be expressed in its eigen-decomposition form $B = U\Lambda U^\top$. The matrix U is an orthogonal matrix with $UU^\top = UU^{-1} = I_d$ (identity matrix of dimension $d \times d$), and the diagonal matrix Λ holds the corresponding eigenvalues $\lambda_1,\ldots,\lambda_d$. To enforce that $\|B\|_F = 1$, the sum of the squares of the eigenvalues λ_i must equal 1, as $\|B\|_F = \sqrt{\mathrm{trace}(B^2)}$ and $\mathrm{trace}(B^2) = \sum_{j=1}^d \lambda_i^2$. Additionally, the requirement of low-rankness on B means that most of the eigenvalues $\lambda_1,\ldots,\lambda_d$ are close to zero, with only a few being significantly larger.

With the goal of obtaining an appropriate low-rank-promoting prior for B, we propose the following approach. We simulate an orthogonal matrix V and simulate $(\gamma_1,\ldots,\gamma_d)$ from a Dirichlet distribution $Dir(\alpha_1,\ldots,\alpha_d)$. Put

$$B = V\mathrm{diag}(\gamma_1^{1/2},\ldots,\gamma_d^{1/2})V^\top.$$

To obtain an approximate low-rank matrix, we take all parameters of the Dirichlet distribution to be very close to 0, for example, by setting $\alpha_1 = \ldots = \alpha_d = 1/d$. It is worth noting that a typical drawing of the Dirichlet distribution leads to one of the γ_is being close to 1 and the others being close to 0. For more detailed discussions on how to choose the parameters for the Dirichlet distribution, one can refer to [17].

Now, we present a prior distribution on $\mathcal{F}$. We opted to use the prior introduced in [11]. With any integer M that $0 < M \leq n$, let us put

$$\mathcal{B}_M(c_\Lambda) = \left\{ (\beta_1, \ldots, \beta_M) \in \mathbb{R}^M : \sum_{s=1}^{M} s|\beta_s| \leq c_\Lambda \text{ and } \beta_M \neq 0 \right\}, \forall c_\Lambda > 0.$$

Now, we define $\mathcal{F}_M(c_\Lambda) \subset \mathcal{F}$ the image of $\mathcal{B}_M(c_\Lambda)$ by the function

$$\begin{array}{rccc} G_M & : \mathbb{R}^M & \to & \mathcal{F} \\ & (\beta_1, \ldots, \beta_M) & \mapsto & \sum_{j=1}^{M} \beta_j \varphi_j. \end{array}$$

Remark 2. *Corollary 1 (below) provides a discussion regarding the approximation of Sobolev spaces (see [18] by the set $\mathcal{F}_M(c_\Lambda)$), which become more accurate as M increases.*

Now, a prior distribution $\nu_M(\mathrm{d}f)$ is defined on the set $\mathcal{F}_M(C+1)$. This is performed by considering the image of the uniform measure on $\mathcal{B}_M(C+1)$ obtained through the function G_M. We consider the following choice for the prior distribution ν on $\mathcal{F}$

$$\mathrm{d}\nu(f) = \frac{\sum_{M=1}^{n} 10^{-M} \nu_M(\mathrm{d}f)}{1 - (\frac{1}{10})^n}. \tag{2}$$

The reason for choosing $C+1$ rather than C in the above definition of the prior distribution support is essentially for technical proof. This is to ensure that as soon as the underlying link function $f^\star$ belongs to $\mathcal{F}_n(C)$, there then exists a small ball around it that is contained in $\mathcal{F}_n(C+1)$. One could safely replace it by $C + a_n$, where $\{a_n\}_{n=1}^{\infty}$ is any positive sequence vanishing sufficiently slowly as $n \to \infty$.

Remark 3. *The integer M can be viewed as a measure of the "dimension" of the function f—the larger the M, the more complex the function—and the prior ν adapts again to the sparsity idea by penalizing large-dimensional functions f. The coefficient 10^{-M}, which appears in (2), shows that more complex models have a geometrically decreasing influence. Inspired from the practical results in [11], the value 10 is a random choice. This choice could be in general changed by another positive constant, but it requires more technical attention.*

2.2. The Proposed Estimator

Definition 2. *The Gibbs posterior distribution over $\mathcal{S}_1^d \times \mathcal{F}_n(C+1)$ is defined as*

$$\hat{\rho}_\lambda(B, f) = \frac{\exp[-\lambda r_n(B, f)] d\pi(B, f)}{\int \exp[-\lambda r_n(B, f)] d\pi(B, f)}.$$

Now, we define an estimator as follows. Let $\lambda > 0$ be a tuning parameter, or sometime called the inverse temperature parameter. Let $(\hat{B}_\lambda, \hat{f}_\lambda)$ be an estimator of $(B^\star, f^\star)$. It is simply achieved by a random draw from $\hat{\rho}_\lambda$, the Gibbs posterior distribution above.

2.3. Theoretical Results

As $\mathbb{E}[Y|X] = f^\star(\langle X, B^\star \rangle)$ almost surely, it is noted that for all $(B, f) \in \mathcal{S}_1^d \times \mathcal{F}_n(C+1)$,

$$\begin{aligned} R(B, f) - R(B^\star, f^\star) &= \mathbb{E}[Y - f(\langle X, B \rangle)]^2 - \mathbb{E}[Y - f^\star(\langle X, B^\star \rangle)]^2 \\ &= \mathbb{E}[f(\langle X, B \rangle) - f^\star(\langle X, B^\star \rangle)]^2 \end{aligned}$$

(Pythagoras theorem).

Definition 3. *For any positive integer $M \leq n$, we set*

$$(B_M^\star, f_M^\star) \in \arg \min_{(B,f)\in \mathcal{S}_1^d \times \mathcal{F}_M(C)} R(B,f).$$

Remark 4. *It is noted here that the infimum $f_M^\star$ is defined on $\mathcal{F}_M(C)$ for each value of M. However, the prior distribution is defined on a slightly larger set, that is, $\mathcal{F}_M(C+1)$.*

Let us define

$$w := 64(C+1)\max[L, C+1], \quad C_1 := 8[(C+1)^2 + \sigma^2].$$

The theoretical results in this work mainly come from the following theorem, the proof of which is provided in Section 3. It should be noted that throughout the paper, the phrase "with probability $1-\delta$" refers to the probability calculated with respect to both the distribution $\mathbf{P}^{\otimes n}$ of the data and the conditional Gibbs distribution $\hat{\rho}_\lambda$.

Theorem 1. *Assume that Assumptions 1 and 2 hold, with*

$$\lambda = \frac{n}{w + 2C_1}. \tag{3}$$

We have that, for all $\delta \in (0,1)$, with a probability of at least $1-\delta$,

$$R(\hat{B}_\lambda, \hat{f}_\lambda) - R(B^\star, f^\star) \leq \mathfrak{C} \inf_{1\leq M\leq n} \left\{ R(B_M^\star, f_M^\star) - R(B^\star, f^\star) + \frac{\log(n)(M + d\,\mathrm{rank}(B^\star) + d\log(d)) + \log\left(\frac{2}{\delta}\right)}{n} \right\},$$

where $\mathfrak{C} > 0$ is a constant depending only on L, σ, C, C_ϕ.

Remark 5. *As in practice, the value of w and C_1 are not known, and the theoretical value of λ cannot be used. However, it provides a good order to tune this parameter, for example, using cross-validation.*

Remark 6. *Theorem 1 can be interpreted in a straightforward manner. Essentially, it states that if there exists a "small" M and $\mathrm{rank}(B^\star)$ is small, such that the difference between $R(B_M^\star, f_M^\star)$ and $R(B^\star, f^\star)$ is minimal, then the difference between $R(\hat{B}_\lambda, \hat{f}_\lambda)$ and $R(B^\star, f^\star)$ will also be small in the order of $\log(n)/n$. On the other hand, if neither of these conditions are met, then the rate $M\log(n)/n$ or $\mathrm{rank}(B^\star)d\log(n)/n$ (or either) will start to dominate, thus resulting in a decrease in the general quality of the convergence rate.*

We can obtain a good convergence rate as soon as a low-rank assumption is considered. This is typically achievable when $B^\star$ is already low-rank or can be well approximated by a low-rank matrix. In the case that $f^\star$ is sufficiently regular, we can obtain a good approximation with a "small" M.

As shown in [11], when $f^\star$ belongs to a Sobolev space, we can derive a more specific nonparametric rate for the above theorem. For example, assume that $\{\varphi_k\}_{k=1}^\infty$ is the system of trigonometric functions and in addition that the link function $f^\star$ is in the following Sobolev ellipsoid space [18],

$$\mathcal{W}\left(k, \frac{6C^2}{\pi^2}\right) = \left\{ f \in L_2([-1,1]) : f = \sum_{j=1}^\infty \beta_j \varphi_j \text{ and } \sum_{j=1}^\infty j^{2k}\beta_j^2 \leq \frac{6C^2}{\pi^2} \right\}$$

where $k \geq 2$ is an unknown regularity parameter. In this context, the approximation set $\mathcal{F}_M(C+1)$ is in the following form:

$$\mathcal{F}_M(C+1) = \left\{ f \in L_2([-1,1]) : f = \sum_{s=1}^{M} \beta_s \varphi_s, \sum_{s=1}^{M} s|\beta_s| \leq C+1 \text{ and } \beta_M \neq 0 \right\}.$$

It should be noted that the results presented in this paper are in the so-called adaptive setting, where the regularity parameter k is not assumed to be known. However, in order to obtain these results, it is necessary to make an additional assumption.

Assumption 4. *We assume that the probability density of the random variable $\langle X, B^\star \rangle$ is defined on $[-1,1]$, and it is upper-bounded by a constant $A > 0$.*

Corollary 1. *Assume that Theorem 1 and additional Assumption 4 hold. Moreover, assume that $f^\star$ is in the Sobolev ellipsoid space $\mathcal{W}(k, 6C^2/\pi^2)$, where the regularity parameter $k \geq 2$ is unknown. The tuning parameter λ is as in (3). We have that for all $\delta \in (0,1)$ with a probability of at least $1-\delta$,*

$$R(\hat{B}_\lambda, \hat{f}_\lambda) - R(B^\star, f^\star) \leq \mathfrak{C}' \left\{ \left(\frac{\log(n)}{n} \right)^{\frac{2k}{2k+1}} + \frac{\log(n)(d\,\mathrm{rank}(B^\star) + d\log d) + \log\left(\frac{2}{\delta}\right)}{n} \right\}, \quad (4)$$

where $\mathfrak{C}' > 0$ is a constant depending only on L, C, σ, C_ϕ, A .

The proof for Corollary 1 follows a similar approach to that of Corollary 4 in [11], and thus, it is not included in this paper.

Remark 7. *From an asymptotic point of view, that d is fixed and $n \to \infty$, the leading rate on the right-hand side in the above Corollary is $(\log(n)/n)^{\frac{2k}{2k+1}}$. This is known to be the minimax rate of convergence up to a $\log(n)$ factor over a Sobolev class; see [18]. On the other hand, in a nonasymptotic setting where n is "small", we obtain the estimation rate $\mathrm{rank}(B^\star)d\log(n)/n$, which was also obtained by [2], and it is minimax optimal up to a logarithmic term, as in [3].*

From Theorem 1, it is actually possible to derive that the Gibbs posterior $\hat{\rho}_\lambda$ contracts around $(B^\star, f^\star)$ at the optimal rate.

Theorem 2. *Under the same assumptions for Theorem 1 and the same definition for λ, let ε_n be any sequence in $(0,1)$, such that $\varepsilon_n \to 0$ when $n \to \infty$. Define*

$$\mathcal{E}_n = \Bigg\{ (B,f) \in \mathcal{S}_1^d \times \mathcal{F}_n(C+1) : R(B,f) - R(B^\star, f^\star) \leq \mathfrak{C} \inf_{1 \leq M \leq n} \Bigg\{ R(B_M^\star, f_M^\star) - R(B^\star, f^\star) + \frac{\log(n)(M + \mathrm{rank}(B^\star)d + d\log d) + \log\left(\frac{2}{\varepsilon_n}\right)}{n} \Bigg\} \Bigg\}.$$

Then,

$$\mathbb{E}\left[\mathbb{P}_{(B,f)\sim\hat{\rho}_\lambda}((B,f) \in \mathcal{E}_n) \right] \geq 1 - \varepsilon_n \xrightarrow[n\to\infty]{} 1.$$

3. Proofs

For the sake of simplicity in the proofs, we put

$$R^\star := R(B^\star, f^\star), \quad r_n^\star := r_n(B^\star, f^\star).$$

We have that for each $f = \sum_{j=1}^M \beta_j \varphi_j \in \mathcal{F}_M(C+1)$, $\|f\|_\infty \leq \sum_{j=1}^M |\beta_j| \leq C+1$.

The following lemma, Lemma 1, is a Bernstein-type inequality [16] that is useful for our proofs. We denote by $(Z)_+$ the positive part of a random variable Z.

Lemma 1. *Let $Z_1, \ldots, Z_n$ be independent real-valued random variables. It is assumed that there exist two constants $v > 0, w > 0$ that for all integers $r \geq 2$, $\sum_{s=1}^n \mathbb{E}\left[(Z_s)_+^r\right] \leq \frac{r!}{2} v w^{r-2}$. We have that with $\zeta \in (0, 1/w)$,*

$$\mathbb{E} e^{\zeta \sum_{s=1}^n (Z_s - \mathbb{E} Z_s)} \leq e^{\frac{v\zeta^2}{2(1-w\zeta)}}.$$

Let $(A, \mathcal{A})$ be a measurable space and γ_1 and γ_2 be two probability measures on $(A, \mathcal{A})$. Denote by $\mathcal{K}(\gamma_1, \gamma_2)$ the Kullback–Leibler divergence of γ_1 with respect to γ_2. Lemma 2 is a classical result, and its proof can be found, for example, in [15], (page 4).

Lemma 2. *Let $(A, \mathcal{A})$ be a measurable space. For any probability measure ν on $(A, \mathcal{A})$ and any measurable function $g : A \to \mathbb{R}$, such that $\int (\exp \circ g) \mathrm{d}\nu < \infty$, we have*

$$\log \int (\exp \circ g) \mathrm{d}\nu = \sup_\kappa \left(\int g \mathrm{d}\kappa - \mathcal{K}(\kappa, \nu) \right), \tag{5}$$

where κ is a probability measure on $(A, \mathcal{A})$ and $\infty - \infty = -\infty$. In addition, when g is upper-bounded on the support of ν, the supremum in (5) is obtained by the Gibbs distribution g, given by

$$\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(a) = \frac{\exp(g(a))}{\int (\exp \circ g) \mathrm{d}\nu}, \quad a \in A.$$

Lemma 3. *We assume that Assumption 1 is satisfied. Put $w = 16(C+1)\max[L, 2(C+1)], C_1 := 8[(C+1)^2 + \sigma^2]$ and take $\lambda \in \left(0, \frac{n}{w + C_1}\right)$ and put*

$$\alpha = \left(\lambda - \frac{\lambda^2 C_1}{2n(1 - \frac{C_2 \lambda}{n})} \right) \quad \text{and} \quad \beta = \left(\lambda + \frac{\lambda^2 C_1}{2n(1 - \frac{C_2 \lambda}{n})} \right). \tag{6}$$

With $\delta \in (0,1)$ and any distribution $\hat{\rho}_\lambda \ll \pi$, we have that

$$\mathbb{E} \int \exp \left[\alpha(R(B,f) - R^\star) + \lambda(-r_n(B,f) + r_n^\star) - \log\left(\frac{d\hat{\rho}_\lambda}{d\pi}(B,f) \right) - \log \frac{2}{\delta} \right] d\hat{\rho}_\lambda(B,f) \leq \delta/2, \tag{7}$$

$$\mathbb{E} \sup_\rho \exp \left[\beta \left(- \int R(B,f) d\rho - R^\star \right) + \lambda \left(\int r_n(B,f) d\rho - r_n^\star \right) - \mathcal{K}(\rho, \pi) - \log \frac{2}{\delta} \right] \leq \delta/2, \tag{8}$$

Proof. Fix $B \in \mathcal{S}_1^d$ and $f \in \mathcal{F}_n(C+1)$. We start by using Lemma 1 with the following random variables:

$$T_i = -(Y_i - f(\langle X_i, B\rangle))^2 + (Y_i - f^\star(\langle X_i, B^\star\rangle))^2, \quad i = 1, \ldots, n.$$

Note that $T_i, i = 1, \ldots, n$ are independent, and we have that

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}T_i^2 &= \sum_{i=1}^n \mathbb{E}\Big\{ [2Y_i - f(\langle X_i, B\rangle) - f^\star(\langle X_i, B^\star\rangle)]^2 [f(\langle X_i, B\rangle) - f^\star(\langle X_i, B^\star\rangle)]^2 \Big\} \\
&= \sum_{i=1}^n \mathbb{E}\Big\{ [2\epsilon_i + f^\star(\langle X_i, B^\star\rangle) - f(\langle X_i, B\rangle)]^2 [f(\langle X_i, B\rangle) - f^\star(\langle X_i, B^\star\rangle)]^2 \Big\} \\
&\le \sum_{i=1}^n \mathbb{E}\Big\{ \Big[8\epsilon_i^2 + 8(C+1)^2\Big] [f(\langle X_i, B\rangle) - f^\star(\langle X_i, B^\star\rangle)]^2 \Big\}. \\
&\le 8\Big[(C+1)^2 + \sigma^2\Big] \sum_{i=1}^n \mathbb{E}[f(\langle X_i, B\rangle) - f^\star(\langle X_i, B^\star\rangle)]^2 := v,
\end{aligned}$$

where we set $C_1 := 8[(C+1)^2 + \sigma^2]$; and $v = nC_1[R(B,f) - R^\star]$.

Now, for all integers k greater than 3, we have that

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{E}\Big[(T_i)_+^k\Big] \\
&\le \sum_{i=1}^n \mathbb{E}\Big\{ |2Y_i - f(\langle X_i, B\rangle) - f^\star(\langle X_i, B^\star\rangle)|^k |f(\langle X_i, B\rangle) - f^\star(\langle X_i, B^\star\rangle)|^k \Big\} \\
&= \sum_{i=1}^n \mathbb{E}\Big\{ |2\epsilon_i + f^\star(\langle X_i, B^\star\rangle) - f(\langle X_i, B\rangle)|^k |f(\langle X_i, B\rangle) - f^\star(\langle X_i, B^\star\rangle)|^k \Big\} \\
&\le 2^{k-1} \sum_{i=1}^n \mathbb{E}\Big\{ \Big[2^k |\epsilon_i|^k + 2^k (C+1)^k\Big] 2^{k-2} (C+1)^{k-2} |f(\langle X_i, B\rangle) - f^\star(\langle X_i, B^\star\rangle)|^2 \Big\}.
\end{aligned}$$

In the last inequality, we used the fact that $|q+w|^k \le 2^{k-1}(|q|^k + |w|^k)$. We obtain that

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}\Big[(T_i)_+^k\Big] &\le \sum_{i=1}^n \Big[2^{2k-2} k! \sigma^2 L^{k-2} + 2^{2k-1}(C+1)^k\Big] 2^{k-2}(C+1)^{k-2} [R(B,f) - R^\star] \\
&= v \times \frac{\Big[2^{2k-2} k! \sigma^2 L^{k-2} + 2^{2k-1}(C+1)^k\Big] 2^{k-2}(C+1)^{k-2}}{[2(C+1)^2 + 4\sigma^2]} \\
&\le v \times \frac{k! 8^{k-2} \max\Big[L^{k-2}, 2^{k-2}(C+1)^{k-2}\Big] 2^{k-2}(C+1)^{k-2}}{2} := \frac{k!}{2} v w^{k-2},
\end{aligned}$$

with $w = 64(C+1)\max[L, C+1]$.

Thus, for any $\lambda \in (0, n/w)$, taking $\zeta = \lambda/n$, we apply Lemma 1 to obtain

$$\begin{aligned}
\mathbb{E}\exp[\lambda(R(B,f) - R^\star - r_n(B,f) + r_n^\star)] &\le \exp\left(\frac{v\lambda^2}{2n^2(1 - \frac{w\lambda}{n})}\right) \\
&= \exp\left(\frac{C_1[R(B,f) - R^\star]\lambda^2}{2n(1 - \frac{w\lambda}{n})}\right).
\end{aligned}$$

Therefore, we obtain, with the α given in (6),

$$\mathbb{E}e^{\alpha(R(B,f) - R^\star) + \lambda(-r_n(B,f) + r_n^\star) - \log\left(\frac{2}{\delta}\right)} \le \delta/2.$$

Next, integrating with respect to π and consequently using Fubini's theorem, we obtain

$$\mathbb{E}\int \exp\left[\alpha(R(B,f)-R^\star)+\lambda(-r_n(B,f)+r_n^\star)-\log(2/\delta)\right]\mathrm{d}\pi(B,f)\leq \delta/2.$$

To obtain (7), it is noted that for any measurable function h,

$$\int \exp[h(B,f)]d\pi = \int \exp\left[h(B,f)-\log\frac{d\hat{\rho}_\lambda}{d\pi}(B,f)\right]d\hat{\rho}_\lambda.$$

The proof for (8) is similar. More precisely, we apply Lemma 1 with $T_i = (Y_i - f(\langle X,B\rangle))^2 - (Y_i - f^\star(\langle X,B^\star\rangle))^2$. We obtain, for any $\lambda\in(0,n/w)$,

$$\mathbb{E}\exp[\lambda(r_n(B,f)+r_n^\star-R(B,f)+R^\star)]\leq \exp\left(\frac{v\lambda^2}{2n^2(1-\frac{w\lambda}{n})}\right).$$

By rearranging terms, using definition of β in (6), and multiplying both sides by $\delta/2$, we obtain

$$\mathbb{E}\exp\left[\beta(-R(B,f)+R^\star)+\lambda(r_n(B,f)-r_n^\star)-\log\frac{2}{\delta}\right]\leq \delta/2.$$

Integrating with respect to π and using Fubini's theorem, we obtain

$$\mathbb{E}\int \exp\left[\beta(-R(B,f)+R^\star)+\lambda(r_n(B,f)-r_n^\star)-\log\frac{2}{\delta}\right]\mathrm{d}\pi\leq \delta/2.$$

Now, Lemma 2 is applied to the integral, and this directly yields (8). □

Proof of Theorem 1. Recall that $\mathbf{P}^{\otimes n}$ stands for the distribution of the sample $\mathcal{D}_n$; the Equation (7) can be written conveniently as

$$\mathbb{E}_{\mathcal{D}_n\sim\mathbf{P}^{\otimes n}}\mathbb{E}_{(\hat{B},\hat{f})\sim\hat{\rho}_\lambda}\exp\Bigg[\alpha\Big(R(\hat{B},\hat{f})-R^\star\Big)+\lambda\Big(-r_n(\hat{B},\hat{f})+r_n^\star\Big)- \log\left(\frac{\mathrm{d}\hat{\rho}_\lambda}{\mathrm{d}\pi}(\hat{B},\hat{f})\right)-\log\frac{2}{\delta}\Bigg]\leq\delta/2,$$

Now, we use the standard Chernoff trick to transform an exponential moment inequality into a deviation inequality, i.e., using $\exp(\lambda x)\geq \mathbf{1}_{\mathbb{R}_+}(x)$. We obtain, with a probability of at least $1-\delta/2$ for any $\delta\in(0,1)$ and any distribution $\hat{\rho}_\lambda$,

$$R(\hat{B},\hat{f})-R^\star\leq\frac{\lambda}{\alpha}\left(r_n(\hat{B},\hat{f})-r_n^\star+\frac{\log\left(\frac{\mathrm{d}\hat{\rho}_\lambda}{\mathrm{d}\pi}(\hat{B},\hat{f})\right)+\log\left(\frac{2}{\delta}\right)}{\lambda}\right).$$

It is noted that we have

$$\begin{aligned}\log\left(\frac{\mathrm{d}\hat{\rho}_\lambda}{\mathrm{d}\pi}(\hat{B},\hat{f})\right)&=\log\left(\frac{\exp(-\lambda r_n(\hat{B},\hat{f}))}{\int\exp(-\lambda r_n(B,f))\mathrm{d}\pi}\right)\\&=-\lambda r_n(\hat{B},\hat{f})-\log\int e^{-\lambda r_n(B,f)}\mathrm{d}\pi;\end{aligned}$$

thus, we obtain, with a probability larger than $1-\delta/2$,

$$R(\hat{B},\hat{f}) - R^\star \leq \frac{1}{\alpha}\left(\log \int \exp(-\lambda r_n(B,f))\mathrm{d}\pi - \lambda r_n^\star + \log\left(\frac{2}{\delta}\right)\right).$$

Now, using Lemma 2, it yields that with a probability larger than $1-\delta/2$,

$$R(\hat{B},\hat{f}) - R^\star \leq \frac{\lambda}{\alpha}\left(\int r_n(B,f)\mathrm{d}\hat{\rho}_\lambda - r_n^\star + \frac{\mathcal{K}(\hat{\rho}_\lambda,\pi) + \log\left(\frac{2}{\delta}\right)}{\lambda}\right). \tag{9}$$

Now, from (8) with an application of the standard Chernoff trick, we obtain, with a probability larger than $1-\delta/2$ for any $\delta \in (0,1)$ and any distribution $\hat{\rho}_\lambda \ll \pi$,

$$\int r_n(B,f)\mathrm{d}\hat{\rho}_\lambda - r_n^\star \leq \frac{\beta}{\lambda}\left(\int R(B,f)\mathrm{d}\hat{\rho}_\lambda - R^\star\right) + \frac{\mathcal{K}(\hat{\rho}_\lambda,\pi) + \log\left(\frac{2}{\delta}\right)}{\lambda}. \tag{10}$$

Combining (9) and (10) with a union bound argument gives the bound, with a probability larger than $1-\delta$,

$$R(\hat{B},\hat{f}) - R^\star \leq \inf_{\rho}\left\{\frac{\beta}{\alpha}\left(\int R(B,f)\mathrm{d}\rho - R^\star\right) + 2\,\frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\delta}\right)}{\alpha}\right\}. \tag{11}$$

The final steps of the proof involve making the right-hand side of the inequality more explicit. To achieve this, we limit the infimum bound to a specific distribution. This allows us to have a more concrete understanding of the result and to explicitly obtain the error rate.

Put $B^\star = U\Lambda U^\top$ and let $r = \#\{i : \Lambda_i > \varepsilon\}$, with small $\varepsilon \in (0,1)$. Take

$$\mathrm{d}\rho_\eta^1 \propto \mathbf{1}(\forall i : |v_i - \Lambda_i| \leq \varepsilon; \forall i = 1,\ldots,r : \|u_i - U_i\|_F \leq \eta)\pi(du,dv)$$

For any positive integer $M \leq n$ and any $\eta,\gamma \in (0,1/n)$, let the probability measure $\rho_{M,\eta,\gamma}$ be defined by

$$\mathrm{d}\rho_{M,\eta,\gamma}(B,f) = \mathrm{d}\rho_\eta^1(B)\mathrm{d}\rho_{M,\gamma}^2(f),$$

with

$$\rho_{M,\gamma}^2(f) \propto \mathbf{1}_{[\|f-f_M^\star\|_M \leq \gamma]}\nu_M(f).$$

We denote for $f = \sum_{s=1}^M \beta_s\varphi_s \in \mathcal{F}_M(C+1)$, $\|f\|_M = \sum_{j=1}^M j|\beta_j|$.

Inequality (11) leads to

$$R(\hat{B},\hat{f}) - R^\star \leq \inf_{1\leq M\leq n}\inf_{\eta,\gamma>0}\left\{\frac{\beta}{\alpha}\left(\int R(B,f)\mathrm{d}\rho_{M,\eta,\gamma}(B,f) - R^\star\right) + 2\,\frac{\mathcal{K}(\rho_{M,\eta,\gamma},\pi) + \log\left(\frac{2}{\delta}\right)}{\alpha}\right\}. \tag{12}$$

To finish the proof, we have to control the different terms in (12). Note first that

$$\begin{aligned}\mathcal{K}(\rho_{M,\eta,\gamma},\pi) &= \mathcal{K}(\rho_\eta^1 \otimes \rho_{M,\gamma}^2, \mu\otimes\nu_M)\\ &= \mathcal{K}(\rho_\eta^1,\mu) + \mathcal{K}(\rho_{M,\gamma}^2,\nu_M) + \log\frac{1-(1/10)^n}{10^{-M}}.\end{aligned}$$

By technical Lemma 4, we know that

$$\mathcal{K}(\rho_\eta^1, \mu) \leq rd\log(16/\eta) + C_{D_1} d\log d(1+\log(2/\varepsilon)).$$

Additionally, by technical Lemma 10 in [11], we have that

$$\mathcal{K}(\rho_{M,\gamma}^2, \nu_M) = M\log\left(\frac{C+1}{\gamma}\right).$$

Bringing together all the parts, it arrives at

$$\mathcal{K}(\rho_{M,\eta,\gamma}, \pi) \leq rd\log(1/c) + C_{D_1} d\log d(1+\log(2/\delta)) + M\log\left(\frac{C+1}{\gamma}\right) + \log\frac{1}{10^{-M}}. \tag{13}$$

Finally, it remains to control the term $\int R(B,f)\mathrm{d}\rho_{M,\eta,\gamma}(B,f)$. To this aim, we write

$$\begin{aligned}
&\int R(B,f)\mathrm{d}\rho_{M,\eta,\gamma}(B,f)\\
&= \int \mathbb{E}\Big[(Y - f(\langle X,B\rangle))^2\Big]\mathrm{d}\rho_{M,\eta,\gamma}(B,f)\\
&= \int \mathbb{E}\big[(Y - f_M^\star(\langle X,B_M^\star\rangle) + f_M^\star(\langle X,B_M^\star\rangle) - f(\langle X,B_M^\star\rangle) +\\
&f(\langle X,B_M^\star\rangle) - f(\langle X,B\rangle))^2\big]\mathrm{d}\rho_{M,\eta,\gamma}(B,f)\\
&= R(B_M^\star,f_M^\star) + \int \mathbb{E}\Big[(f_M^\star(\langle X,B_M^\star\rangle) - f(\langle X,B_M^\star\rangle))^2 + (f(\langle X,B_M^\star\rangle) - f(\langle X,B\rangle))^2\\
&+ 2(Y - f_M^\star(\langle X,B_M^\star\rangle))(f_M^\star(\langle X,B_M^\star\rangle) - f(\langle X,B_M^\star\rangle)) +\\
&2(Y - f_M^\star(\langle X,B_M^\star\rangle))(f(\langle X,B_M^\star\rangle) - f(\langle X,B\rangle))\\
&+ 2(f_M^\star(\langle X,B_M^\star\rangle) - f(\langle X,B_M^\star\rangle))(f(\langle X,B_M^\star\rangle) - f(\langle X,B\rangle))\Big]\mathrm{d}\rho_{M,\eta,\gamma}(B,f)\\
&:= R(B_M^\star,f_M^\star) + \mathbf{A} + \mathbf{B} + \mathbf{C} + \mathbf{D} + \mathbf{E}.
\end{aligned}$$

Computation of C by Fubini's theorem:

$$\begin{aligned}
&\mathbf{C}\\
=&\mathbb{E}\left[\int 2(Y - f_M^\star(\langle X,B_M^\star\rangle))(f_M^\star(\langle X,B_M^\star\rangle) - f(\langle X,B_M^\star\rangle))\mathrm{d}\rho_{M,\eta,\gamma}(B,f)\right]\\
=&\mathbb{E}\left\{\int\left[2(Y - f_M^\star(\langle X,B_M^\star\rangle))\int(f_M^\star(\langle X,B_M^\star\rangle) - f(\langle X,B_M^\star\rangle))\mathrm{d}\rho_{M,\gamma}^2(f)\right]\mathrm{d}\rho_\eta^1(B)\right\}.
\end{aligned}$$

Using the triangle inequality, we obtain that for $f = \sum_{s=1}^M \beta_s\varphi_s$ and $f_M^\star = \sum_{s=1}^M(\beta_M^\star)_s\varphi_s$,

$$\sum_{j=1}^M j|\beta_j| \leq \sum_{j=1}^M j|\beta_j - (\beta_M^\star)_j| + \sum_{j=1}^M j|(\beta_M^\star)_j|.$$

Since $f_M^\star \in \mathcal{F}_M(C)$, and thus $\sum_{s=1}^M s|(\beta_M^\star)_s| \leq C$, as a consequence, $\sum_{s=1}^M s|\beta_s| \leq C+1$ as soon as $\|f - f_M^\star\|_M \leq 1$. This shows that the set

$$\left\{f = \sum_{j=1}^M \beta_j\varphi_j : \|f - f_M^\star\|_M \leq \gamma\right\}$$

is contained in the support of ν_M. In particular, this implies that $\rho_{M,\gamma}^2$ is centered at $f_M^\star$ and, consequently,

$$\int(f_M^\star(\langle X,B_M^\star\rangle) - f(\langle X,B_M^\star\rangle))\mathrm{d}\rho_{M,\gamma}^2(f) = 0.$$

This proves that $\mathbf{C} = 0$.
Control of A: Clearly,

$$\mathbf{A} \leq \int \sup_{y \in \mathbb{R}} ((f_M^\star(y) - f(y))^2 \mathrm{d}\rho_{M,\gamma}^2(f) \leq \gamma^2.$$

Control of B: We have

$$\begin{aligned}
\mathbf{B} &= \int \mathbb{E}\Big[(f(\langle X, B_M^\star\rangle) - f(\langle X, B\rangle))^2\Big] \mathrm{d}\rho_{M,\eta,\gamma}(B, f) \\
&\leq \int \mathbb{E}\Big[\big(C_\phi(C+1)(B_M^\star - B)X\big)^2\Big] \mathrm{d}\rho_\eta^1(B) \text{ (using the mean value theorem)} \\
&\leq C_\phi^2(C+1)^2 \mathbb{E}\Big[\|X\|_\infty^2\Big] \int \|B_M^\star - B\|_F^2 \mathrm{d}\rho_\eta^1(B) \quad \text{(by Assumption 4)}.
\end{aligned}$$

Using Lemma 6 from [19], we have that

$$\int \|B_M^\star - B\|_F^2 \mathrm{d}\rho_\eta^1(B) \leq (3dc + 2r\eta)^2.$$

Thus,

$$\mathbf{B} \leq C_\phi^2(C+1)^2(3dc + 2r\eta)^2.$$

Control of E: We have that

$$\begin{aligned}
|\mathbf{E}| &\leq 2 \int \mathbb{E}\Big[|f_M^\star(\langle X, B_M^\star\rangle) - f(\langle X, B_M^\star\rangle)||f(\langle X, B_M^\star\rangle) - f(\langle X, B\rangle)|\Big] \mathrm{d}\rho_{M,\eta,\gamma}(B, f) \\
&\leq 2 \int \mathbb{E}\Big[|f_M^\star(\langle X, B_M^\star\rangle) - f(\langle X, B_M^\star\rangle)|C_\phi(C+1)|(B_M^\star - B)X|\Big] \mathrm{d}\rho_{M,\eta,\gamma}(B, f) \\
&\leq 2\Big(\int \mathbb{E}(f_M^\star(\langle X, B_M^\star\rangle) - f(\langle X, B_M^\star\rangle))^2 \mathrm{d}\rho_{M,\eta,\gamma}(B, f)\Big)^{\frac{1}{2}} \\
&\Big(\int \mathbb{E}\big(C_\phi(C+1)(B_M^\star - B)X\big)^2 \mathrm{d}\rho_{M,\eta,\gamma}(B, f)\Big)^{\frac{1}{2}} \\
&\leq 2\Big(\gamma^2\Big)^{\frac{1}{2}}\Big(C_\phi^2(C+1)^2(3dc + 2r\eta)^2\Big)^{\frac{1}{2}} = 2C_\phi(C+1)\gamma(3d\varepsilon + 2r\eta).
\end{aligned}$$

Control of D: Finally,

$$\begin{aligned}
\mathbf{D} &= 2 \int \mathbb{E}[(Y - f_M^\star(\langle X, B_M^\star\rangle))(f(\langle X, B_M^\star\rangle) - f(\langle X, B\rangle))] \mathrm{d}\rho_{M,\eta,\gamma}(B, f) \\
&= 2 \int \mathbb{E}[(Y - f_M^\star(\langle X, B_M^\star\rangle))(f_M^\star(\langle X, B_M^\star\rangle) - f_M^\star(\langle X, B\rangle))] \mathrm{d}\rho_\eta^1(B) \\
&\quad (\text{since } \int f \mathrm{d}\rho_{M,\gamma}^2(f) = f_M^\star) \\
&= 2\mathbb{E}\Big[(Y - f_M^\star(\langle X, B_M^\star\rangle)) \int (f_M^\star(\langle X, B_M^\star\rangle) - f_M^\star(\langle X, B\rangle)) \mathrm{d}\rho_\eta^1(B)\Big] \\
&\leq 2\sqrt{\mathbb{E}\Big[\big(Y - f_M^\star(\langle X, B_M^\star\rangle)\big)^2\Big]} \sqrt{\mathbb{E}\Big[\int \big(f_M^\star(\langle X, B_M^\star\rangle) - f_M^\star(\langle X, B\rangle)\big) \mathrm{d}\rho_\eta^1(B)\Big]^2} \\
&= 2\sqrt{R(B_M^\star, f_M^\star)} \sqrt{\mathbb{E}\Big[\int \big(f_M^\star(\langle X, B_M^\star\rangle) - f_M^\star(\langle X, B\rangle)\big) \mathrm{d}\rho_\eta^1(B)\Big]^2}.
\end{aligned}$$

As we have that

$$|f_M^\star(\langle X, B_M^\star\rangle) - f_M^\star(\langle X, B\rangle)| \leq C_\phi(C+1)|\langle (B_M^\star - B)X\rangle| \leq C_\phi(C+1)\|B_M^\star - B\|_F,$$

it leads to

$$\left[\int (f_M^\star(\langle X, B_M^\star\rangle) - f_M^\star(\langle X, B\rangle))\mathrm{d}\rho_\eta^1(B)\right]^2 \leq C_\phi^2(C+1)^2\left[\int \|B_M^\star - B\|_F \mathrm{d}\rho_\eta^1(B)\right]^2 \leq C_\phi^2(C+1)^2(3dc+2r\eta)^2,$$

and therefore,

$$\mathbf{D} \leq 2C_\phi(C+1)(3dc+2r\eta)\sqrt{R(0,0)/2} \leq \sqrt{2}C_\phi(C+1)(3d\varepsilon+2r\eta)\sqrt{C^2+\sigma^2}.$$

Thus, taking $\eta = \gamma = \varepsilon = 1/n$ and assembling all the components, we obtain that

$$\mathbf{A}+\mathbf{B}+\mathbf{C}+\mathbf{D}+\mathbf{E} \leq \frac{\mathfrak{C}_1}{n},$$

where $\mathfrak{C}_1$ is a positive constant function of C, σ, and C_ϕ. Combining this inequality with (12) and (13) yields, with a probability larger than $1-\delta$,

$$R(\hat{B}_\lambda, \hat{f}_\lambda) - R^\star \leq \inf_{1\leq M\leq n}\left\{\frac{\beta}{\alpha}\left(R(B_M^\star, f_M^\star) - R^\star + \frac{\mathfrak{C}_1}{n}\right) + 2\,\frac{M\log((C+1)10n) + rd\log(16n) + C_{D_1}d\log d\log(2ne) + \log\left(\frac{2}{\delta}\right)}{\lambda}\right\}.$$

Finally, choosing $\lambda = \frac{n}{w+2C_1}$, it yields that there exists a constant $\mathfrak{C}_2 > 0$ depending only on L, σ, C, C_ϕ with a probability of at least $1-\delta$, such that

$$R(\hat{B}_\lambda, \hat{f}_\lambda) - R^\star \leq \mathfrak{C}_2 \inf_{1\leq M\leq n}\left\{R(B_M^\star, f_M^\star) - R^\star + \frac{M\log(10Cn) + rd\log(16n) + \mathfrak{C}_3 d\log d\log(2ne) + \log\left(\frac{2}{\delta}\right)}{n}\right\}.$$

This concludes the proof of Theorem 1. □

Lemma 4. *Let $r = \#\{i : \Lambda_i > \varepsilon\}$ with small $\varepsilon \in [0,1)$. Take*

$$d\rho_\eta^1 \propto \mathbf{1}(\forall i : |v_i - \Lambda_i| \leq \varepsilon; \forall i = 1,\ldots,r : \|u_i - U_i\|_F \leq \eta)\mu(du, dv)$$

Then,

$$\mathcal{K}(\rho_\eta^1, \mu) \leq rd\log(16/\eta) + \mathfrak{C}_3 d\log d\log(2e/\varepsilon)$$

where $\mathfrak{C}_3$ is a universal constant.

Proof. We have that

$$\mathcal{K}(\rho_\eta^1, \mu) = \log\frac{1}{\mu(\{u, v : \forall i : |v_i - \Lambda_i| \leq \varepsilon; \forall i = 1, r : \|u_i - U_i\|_F \leq \eta\})} = \log\frac{1}{\mu(\{\forall i = 1, r : \|u_{i\cdot} - U_{i\cdot}\|_F \leq \eta\})} + \log\frac{1}{\mu(\{\forall i : |v_i - \Lambda_i| \leq \varepsilon\})}.$$

The first log term

$$\pi(\{\forall i = 1, r : \|u_{i.} - U_{i.}\|_F \leq \eta\}) \geq \prod_{i=1}^{r} \left[\frac{\pi^{(d-1)/2}(\eta/2)^{d-1}}{\Gamma(\frac{d-1}{2}+1)} \Big/ \frac{2\pi^{(d+1)/2}}{\Gamma(\frac{d+1}{2})} \right]$$
$$\geq \left[\frac{\eta^{d-1}}{2^d \pi} \right]^r \geq \frac{\eta^{r(d-1)}}{2^{4rd}}.$$

Note the following for the above calculation: firstly, the distribution of the orthogonal vector is approximated by the uniform distribution on the sphere [20], and secondly, the probability is greater or equal to the volume of the (d-1)-"circle" with radius $c/2$ over the surface area of the d-"unit sphere".

It is noted that if $\gamma \sim Beta(a, b)$ (beta distribution), then $\gamma^{1/2}$ has the pdf as $f(\gamma) = 2\frac{\gamma^{2a-1}(1-\gamma^2)^{b-1}}{Be(a,b)}, 0 < \gamma < 1$ where $Be(a, b)$ is the beta function. The second log term in the Kullback–Leibler term with $a = \alpha_i, b = \sum_{i=1}^{d} \alpha_i - \alpha_i, \alpha_i = 1/d$ is

$$\pi(\{\forall i : |v_i - \Lambda_i| \leq \varepsilon\}) = \prod_{i=1}^{d} \int_{\max(\Lambda_i - \varepsilon, 0)}^{\min(\Lambda_i + \varepsilon, 1)} \frac{v_i^{2a-1}(1-v_i^2)^{b-1}}{2Be(a,b)} dv_i$$
$$\geq \prod_{i=1}^{d} \int_{0}^{\varepsilon} \frac{v_i^{2a-1}(1-v_i^2)^{b-1}}{2Be(a,b)} dv_i \geq \mathfrak{C}_3 (\varepsilon/2d)^d \, e^{-d \log d}.$$

The interval of integration contains at least an interval of length ε. Thus, we obtain

$$\mathcal{K}(\rho_\eta^1, \mu) \leq \log \frac{2^{4rd}}{\eta^{r(d-1)}} + \log \left(\frac{(2d)^d e^{d \log d}}{\mathfrak{C}_3 \varepsilon^d} \right) \leq rd \log(\frac{16}{\eta}) + \mathfrak{C}_3 d \log d \log(\frac{e2}{\varepsilon})$$

for some absolute numerical constant $\mathfrak{C}_3$ that does not depend on r, n or d. □

Proof of Theorem 2. We also apply Lemma 3, and focus on (7), applied to $\delta := \varepsilon_n$, that is

$$\mathbb{E} \int \exp \left[\alpha(R(B,f) - R^\star) + \lambda(-r_n(B,f) + r_n^\star) - \log \left(\frac{\mathrm{d}\hat{\rho}_\lambda}{\mathrm{d}\pi}(B,f) \right) - \right.$$
$$\left. \log \frac{2}{\varepsilon_n} \right] \mathrm{d}\hat{\rho}_\lambda(B,f) \leq \varepsilon_n/2$$

Using Chernoff's inequality, this leads to

$$\mathbb{E}\left[\mathbb{P}_{(B,f)\sim\hat{\rho}_\lambda}((B,f) \in \mathcal{A}_n) \right] \geq 1 - \frac{\varepsilon_n}{2}$$

where

$$\mathcal{A}_n = \left\{ (B,f) : \alpha\left(R(B,f) - R^\star\right) + \lambda\left(-r_n(B,f) + r_n^\star\right) \leq \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(B,f) \right] + \log \frac{2}{\varepsilon_n} \right\}.$$

From the definition of $\hat{\rho}_\lambda$, for $(B,f) \in \mathcal{A}_n$, we obtain that

$$\begin{aligned}\alpha\Big(R(B,f)-R^\star\Big) &\le \lambda\Big(r_n(B,f)-r_n^\star\Big)+\log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(B,f)\right]+\log\frac{2}{\varepsilon_n}\\ &\le -\log\int\exp[-\lambda r_n(B,f)]\pi(\mathrm{d}(B,f))-\lambda r_n^\star+\log\frac{2}{\varepsilon_n}\\ &= \lambda\Big(\int r_n(B,f)\hat{\rho}_\lambda(\mathrm{d}(B,f))-r_n^\star\Big)+\mathcal{K}(\hat{\rho}_\lambda,\pi)+\log\frac{2}{\varepsilon_n}\\ &= \inf_\rho\left\{\lambda\Big(\int r_n(B,f)\rho(\mathrm{d}(B,f))-r_n^\star\Big)+\mathcal{K}(\rho,\pi)+\log\frac{2}{\varepsilon_n}\right\}.\end{aligned}$$

Now, put

$$\mathcal{B}_n := \left\{\forall\rho,\beta\Big(-\int R(B,f)d\rho+R^\star\Big)+\lambda\Big(\int r_n d\rho-r_n^\star\Big)\le\mathcal{K}(\rho,\pi)+\log\frac{2}{\varepsilon_n}\right\}.$$

Using (8), we have that

$$\mathbb{E}\Big[\mathbf{1}_{\mathcal{B}_n}\Big]\ge 1-\frac{\varepsilon_n}{2}.$$

We now prove that if λ is such that $\alpha>0$,

$$\mathbb{E}\Big[\mathbb{P}_{(B,f)\sim\hat{\rho}_\lambda}((B,f)\in\mathcal{E}_n)\Big]\ge\mathbb{E}\Big[\mathbb{P}_{(B,f)\sim\hat{\rho}_\lambda}((B,f)\in\mathcal{A}_n)\mathbf{1}_{\mathcal{B}_n}\Big]$$

and, together with,

$$\begin{aligned}\mathbb{E}\Big[\mathbb{P}_{(B,f)\sim\hat{\rho}_\lambda}((B,f)\in\mathcal{A}_n)\mathbf{1}_{\mathcal{B}_n}\Big] &= \mathbb{E}\Big[(1-\mathbb{P}_{(B,f)\sim\hat{\rho}_\lambda}((B,f)\notin\mathcal{A}_n))(1-\mathbf{1}_{\mathcal{B}_n^c})\Big]\\ &\ge \mathbb{E}\Big[1-\mathbb{P}_{(B,f)\sim\hat{\rho}_\lambda}((B,f)\notin\mathcal{A}_n)-\mathbf{1}_{\mathcal{B}_n^c}\Big]\\ &\ge 1-\varepsilon_n\end{aligned}$$

leads to

$$\mathbb{E}\Big[\mathbb{P}_{(B,f)\sim\hat{\rho}_\lambda}((B,f)\in\mathcal{E}_n)\Big]\ge 1-\varepsilon_n.$$

To obtain that, assume that we are on the set $\mathcal{B}_n$, and let $(B,f)\in\mathcal{A}_n$. Then,

$$\begin{aligned}\alpha\Big(R(B,f)-R^\star\Big) &\le \inf_\rho\left\{\lambda\Big(\int r_n(B,f)\rho(\mathrm{d}(B,f))-r_n^\star\Big)+\mathcal{K}(\rho,\pi)+\log\frac{2}{\varepsilon_n}\right\}\\ &\le \inf_\rho\left\{\beta\Big(\int R(B,f)\rho(\mathrm{d}(B,f))-R^\star\Big)+2\mathcal{K}(\rho,\pi)+2\log\frac{2}{\varepsilon_n}\right\}\end{aligned}$$

that is,

$$R(B,f)-R^\star\le\inf_\rho\frac{\beta[\int Rd\rho-R^\star]+2\big[\mathcal{K}(\rho,\pi)+\log\frac{2}{\varepsilon}\big]}{\alpha}$$

We upper-bound the right-hand side similarly as in the proof of Theorem 1, which leads to $(B,f)\in\mathcal{E}_n$. □

4. Conclusions

In this paper, we conduct a theoretical study of a low-rank matrix single-index model. The model is used to estimate the link function and the coefficient matrix jointly. We leverage the PAC-Bayesian bounds technique to gain a deeper insight into the properties of this model and its potential applications. The study extends previous work in the field by considering a low-rank matrix, rather than a sparse vector, as the coefficient matrix. We also provide a detailed explanation of the choice of prior distributions for the link function and the coefficient matrix, which allows to obtain accurate and reliable estimates of the

unknown parameters. Overall, this study provides a thorough theoretical understanding of the low-rank matrix single-index model.

The focus of future research would center on executing the proposed approach. There are various possible avenues to explore. One of the promising approaches is to use the reversible jump Markov chain Monte Carlo method, which was successfully applied in the past to address the sparse vector single-index model, as documented in [11].

Funding: This research was funded by Norwegian Research Council grant number 309960 through the Centre for Geophysical Forecasting at NTNU.

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: The author is grateful to two anonymous reviewers for their expert analysis and helpful suggestions.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Weaver, C.; Xiao, L.; Lindquist, M.A. Single-index models with functional connectivity network predictors. *Biostatistics* **2021**, *24*, 52–67. [CrossRef] [PubMed]
2. Fan, J.; Yang, Z.; Yu, M. Understanding Implicit Regularization in Over-Parameterized Single Index Model. *J. Am. Stat. Assoc.* **2022**, 1–14. [CrossRef]
3. Rohde, A.; Tsybakov, A.B. Estimation of high-dimensional low-rank matrices. *Ann. Stat.* **2011**, *39*, 887–930. [CrossRef]
4. Koltchinskii, V.; Lounici, K.; Tsybakov, A.B. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Stat.* **2011**, *39*, 2302–2329. [CrossRef]
5. Zhao, J.; Niu, L.; Zhan, S. Trace regression model with simultaneously low rank and row (column) sparse parameter. *Comput. Stat. Data Anal.* **2017**, *116*, 1–18. [CrossRef]
6. Nelder, J.A.; Wedderburn, R.W. Generalized linear models. *J. R. Stat. Soc. Ser. A Gen.* **1972**, *135*, 370–384. [CrossRef]
7. Hardle, W.; Hall, P.; Ichimura, H. Optimal smoothing in single-index models. *Ann. Stat.* **1993**, *21*, 157–178. [CrossRef]
8. Ichimura, H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econom.* **1993**, *58*, 71–120. [CrossRef]
9. Jiang, B.; Liu, J.S. Variable selection for general index models via sliced inverse regression. *Ann. Stat.* **2014**, *42*, 1751–1786. [CrossRef]
10. Kong, E.; Xia, Y. Variable selection for the single-index model. *Biometrika* **2007**, *94*, 217–229. [CrossRef]
11. Alquier, P.; Biau, G. Sparse Single-Index Model. *JMLR* **2013**, *14*, 243–280.
12. Putra, I.; Dana, I.M. Study of Optimal Portfolio Performance Comparison: Single Index Model and Markowitz Model on LQ45 Stocks in Indonesia Stock Exchange. *Am. J. Humanit. Soc. Sci. Res.* **2020**, *3*, 237–244.
13. Pananjady, A.; Foster, D.P. Single-index models in the high signal regime. *IEEE Trans. Inf. Theory* **2021**, *67*, 4092–4124. [CrossRef]
14. Ganti, R.S.; Balzano, L.; Willett, R. Matrix completion under monotonic single index models. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
15. Catoni, O. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*; Institute of Mathematical Statistics Lecture Notes—Monograph Series 56; Institute of Mathematical Statistics: Beachwood, OH, USA, 2007; Volume 5544465.
16. Boucheron, S.; Lugosi, G.; Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*; Oxford University Press: Oxford, UK, 2013.
17. Wallach, H.; Mimno, D.; McCallum, A. Rethinking LDA: Why priors matter. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; Volume 22.
18. Tsybakov, A.B. *Introduction to Nonparametric Estimation*; Springer Series in Statistics; Springer: Berlin/Heidelberg, Germany, 2009. [CrossRef]
19. Mai, T.T.; Alquier, P. Pseudo-Bayesian quantum tomography with rank-adaptation. *J. Stat. Plan. Inference* **2017**, *184*, 62–76. [CrossRef]
20. Goldstein, S.; Lebowitz, J.L.; Tumulka, R.; Zanghì, N. Any orthonormal basis in high dimension is uniformly distributed over the sphere. *Ann. L'Institut Henri Poincaré Probab. Stat.* **2017**, *53*, 701–717. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.