SUPPLEMENTARY MATERIALS:GENERALIZED FORWARD SUFFICIENT DIMENSION REDUCTION FOR CLASSIFICATION

HARRIS QUACH AND BING LI

Contents

1. Preliminaries, Assumptions and the Conditional Risk 1
1.1. Preliminaries 1
1.2. Assumptions 2
1.3. Conditional Risk 4
2. Proofs for §2: Population Level Developments 5
2.1. Proofs of Propositions 5
2.2. Proofs of Theorems in §2 7
2.3. Lemmas for Theorem 2.1 10
3. Proofs for §3: Consistency of OPCG 17
3.1. Proof of Theorems in §3 17
3.2. Proof of Lemmas for Theorem 3.1 20
4. Computations for §4: Canonical Link and Cumulant Generating Functions 27
4.1. Multivariate Logistic Transformation 28
4.2. Adjacent-Categories Logit Link 28
4.3. Mean and Variance-Covariance of cumulative Variable Z_i 30
5. Fisher-Scoring Algorithm 31
5.1. OPCG 31
5.2. MADE 32
6. Theorems and Lemmas 33
6.1. Uniform Consistency for the mean with a parameter and index 33
6.2. Convergence Rates for Sums of Bounded Outer Products 42
6.3. Bai, Miao, Rao Lemma 44
References 47

1. Preliminaries, Assumptions and the Conditional Risk

1.1. **Preliminaries.** Let Y be a random vector taking values in $\Omega_Y \subset \mathbb{R}^m$ and X a continuous random vector taking values in $\Omega_X \subset \mathbb{R}^p$. Let θ be a parameter in $\Theta \subset \mathbb{R}^m$. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be an i.i.d. sample of (X, Y) and denote $W_i = (Y_i, X_i)$, $W = (X, Y)$

and $W_{1:n} = (W_1, ..., W_n)$. Let $\rho(Y, \theta)$ be a loss function and define the conditional risk associated with $\rho(Y, \theta)$ as $R(\theta, \chi) = E[\rho(Y, \theta)|X = \chi]$.

Suppose the predictor X relates to Y only through the parameter θ in a statistical model, i.e. $\theta(X)$ captures the entire relationship between X and Y. Using a linear approximation to $\theta(X)$ about some value χ, defined locally through the kernel $K(\cdot)$, i.e. $\theta(X) \approx a + B^\top(X - \chi)$, we get the empirical risk function

$$\hat{R}_n(a, B, \chi; W_{1:n}) = E_n K(h_n^{-1}(X - \chi))\rho(Y, a + B^\top(X - \chi))$$

where $h_n > 0$ is some bandwidth that depends on n. We are interested in estimating a, B by minimizing the objective function $\hat{R}_n(a, B, \chi; W_{1:n})$.

Let $b = \text{vec}(B^\top)$ and $c = (a^\top, b^\top)^\top \in \mathbb{R}^{m(p+1)}$ such that the parameter space of c is denoted by $\Theta_c \subset \mathbb{R}^{m(p+1)}$ For any vector t, Define $\nu(t) = (1, t^\top)^\top \otimes I_m$. Then

$$a + B^\top(X - \chi) = ((1, (X - \chi)^\top)^\top \otimes I_m)^\top c = [\nu(X - \chi)]^\top c,$$

and the empirical risk function becomes

$$\hat{R}_n(c, \chi; W_{1:n}) = E_n K(h_n^{-1}(X - \chi))\rho(Y, [\nu(X - \chi)]^\top c).$$

For convenience, we will drop $W_{1:n}$ from the notation and denote the empirical risk by $\hat{R}_n(c, \chi)$. Denote the expected risk function by

$$R_n(c, \chi) = EK(h_n^{-1}(X - \chi))\rho(Y, [\nu(X - \chi)]^\top c).$$

Let $\hat{c}(\chi)$ denote the minimizer of $\hat{R}_n(c, \chi)$, $c_n(\chi)$ the minimizer of $R_n(c, \chi)$, and let $c(\chi) = (\theta(\chi), \text{vec}(\partial\theta(\chi)/\partial\chi^\top)^\top)^\top \in \mathbb{R}^{(p+1)m}$. We will show uniform consistency of the estimates $\hat{c}(\chi)$ over $\chi \in \Omega_X$,

$$\sup_{\chi \in \Omega_X} \|\hat{c}(\chi) - c(\chi)\| \xrightarrow{a.s} 0 \quad as \quad n \longrightarrow \infty.$$

We will also derive the uniform rate of convergence for the estimate $\hat{c}(\chi)$. To compute the convergence rate, we upper bound the supremum above by a stochastic/variance and a deterministic/bias component:

$$\sup_{\chi \in \Omega_X} \|\hat{c}(\chi) - c(\chi)\| \leq \underbrace{\sup_{\chi \in \Omega_X} \|\hat{c}(\chi) - c_n(\chi)\|}_{variance/stochastic} + \underbrace{\sup_{\chi \in \Omega_X} \|c_n(\chi) - c(\chi)\|}_{bias/deterministic}.$$

We need to compute the uniform convergence rate of $\|c_n(\chi) - c(\chi)\|$ and $\|\hat{c}(\chi) - c_n(\chi)\|$.

1.2. **Assumptions.** We make some assumptions based on a general, strictly convex loss function $\rho(y, \theta)$. After we state the general assumptions, we show that Assumptions 1-6 made in the paper for a deviance criteria, associated with the linear exponential family, satisfy the general assumptions in this supplement.

Assumption A1. *The joint density $f(x, y)$ of (X, Y) is twice continuously differentiable with respect to x. The sample space $\Omega_X \subset \mathbb{R}^p$ is compact and X has finite moments.*

Assumption A2. *The canonical parameter θ is identifiable and the loss function $\rho(y,\theta)$ is strictly convex in θ, twice continuously differentiable in θ and has a unique minimum. Let the derivative be denoted by $g(y,\theta) = \partial\rho(y,\theta)/\partial\theta$.*

Assumption A3. *Furthermore, for each $\chi \in \Omega_X$, the conditional risk function $R(\theta,\chi) = E(\rho(y,\theta)|X=\chi)$ has a unique minimum over Θ, with $\theta(\chi)$ as the minimizer, where $\theta(\chi)$ is continuously differentiable. The parameter space $\Theta \subset \mathbb{R}^m$ and $\Theta_c \subset \mathbb{R}^{m(p+1)}$ are compact and convex.*

Assumption A4. *Derivatives and integrals in*

$$\frac{\partial}{\partial\theta}\int\int \rho(y,\theta)f(x,y)dydx \qquad \text{and} \qquad \frac{\partial}{\partial x}\int g(y,\theta)f(y|x)dy$$

are interchangeable.

Assumption A5. *The kernel K is symmetric with finite moments. Furthermore, $\int K(u)du = 1$, $\int K(u)uu^\top du = I_p$ and $K(u)$ is twice continuously differentiable.*

Assumption A6. *The bandwidth satisfies $h \propto n^{-\alpha}$ for $0 < \alpha \leq \frac{1}{p_0}$, where $p_0 > \frac{(p+4)s}{s-2}$ for $s > 2$. So that, as $n \to \infty$, we have $h \downarrow 0$, $\delta_{ph} = \sqrt{\frac{\log n}{h^p n}} \downarrow 0$, and $h^{-1}\delta_{ph} = \delta_{(p+2)h} \to 0$.*

Assumption A7. *Let $\|\cdot\|$ and $\|\cdot\|_F$ denote Euclidean and Frobenius norm respectively. For $h \in [0,1]$ and any compact set $A \subset \mathbb{R}^{m(p+1)}$, the following quantities are all finite:*

$$\sup_{h\in[0,1],c\in A,\chi\in\Omega_X}\int K(u)\left\|\nu(u)[\nu(u)^\top\otimes I_m]\frac{\partial}{\partial\theta^\top}\mathrm{vec}\left[\frac{\partial g(y,\nu(hu)^\top c)}{\partial\theta^\top}\right]\frac{\partial[\nu(hu)^\top c]}{\partial h}\right\|_F f(\chi+hu,y)dydu$$

$$\sup_{h\in[0,1],c\in A,\chi\in\Omega_X}\int K(u)\left\|\nu(u)\frac{\partial g(y,\nu(hu)^\top c)}{\partial\theta^\top}\nu(u)u^\top\frac{\partial f(\chi+hu,y)}{\partial x}\right\|_F dydu$$

$$\sup_{h\in[0,1],c\in A,\chi\in\Omega_X}\int K(u)\left\|\frac{\partial[\mathrm{vec}[\nu(u)\partial_2 g(y,\nu(hu)^\top c)\nu(u)^\top]f(\chi+hu,y)]}{\partial h\partial c^\top}\right\|_F dydu$$

$$\sup_{h\in[0,1],c\in A_1,\chi\in\Omega_X}\int K(u)\left\|\left[\left[\begin{pmatrix}0\\u^\top\end{pmatrix}\otimes I_m\right]^\top\otimes\nu(u)\right]\frac{\partial}{\partial\theta^\top}\mathrm{vec}\left[\frac{\partial g(y,\nu(hu)^\top c)}{\partial\theta^\top}\right]\nu(u)\right\|_F f(\chi+hu,y)dydu$$

$$\sup_{h\in[0,1],c\in A_1,\chi\in\Omega_X}\int K(u)\left\|\nu(u)\frac{\partial g(y,\nu(hu)^\top c)}{\partial\theta^\top}\left[\begin{pmatrix}0\\u^\top\end{pmatrix}\otimes I_m\right]\right\|_F f(\chi+hu,y)dydu$$

$$\sup_{h\in[0,1],c\in A_1,\chi\in\Omega_X}\int K(u)\left\|\nu(u)\frac{\partial g(y,\nu(hu)^\top c)u^\top}{\partial\theta^\top}\nu(u)\frac{\partial f(\chi+hu,y)}{\partial x}\right\|_F dydu$$

Note that since $S(h,c,\chi)$ is twice continuously differentiable, the above conditions ensure the derivatives with respect to c and h are absolutely integrable. The next set of quantities ensure the vector $S(h,c,\chi)$ and matrix $\partial_1 S(h,c,\chi)$ have integrable second moments. The

following quantities are all finite:

$$\sup_{h\in[0,1],c\in A,\chi\in\Omega_X}\int K(u)\left|g(y,\nu(hu)^\top c)^\top\nu(u)^\top\nu(u)g(y,\nu(hu)^\top c)\right|f(\chi+hu,y)dudy$$

$$\sup_{h\in[0,1],c\in A,\chi\in\Omega_X}\int K(u)\left\|\nu(u)g(y,\nu(hu)^\top c)g(y,\nu(hu)^\top c)^\top\nu(u)^\top\right\|_F f(\chi+hu,y)dudy$$

$$\sup_{c\in A}\sup_{h\in[0,1]}\sup_{\chi\in\Omega_X}\int K(u)\left\|\nu(u)\partial_2 g(y,\nu(hu)^\top c)\nu(u)^\top\nu(u)\partial_2 g(y,\nu(hu)^\top c)\nu(u)^\top\right\|_F f(\chi+hu,y)dudy.$$

Lastly, we assume that the following quantity is finite:

$$\sup_{h\in[0,1],c\in A,\chi\in\Omega_X}\left\|\left\{\int K(u)\nu(u)\frac{\partial g(y,\nu(hu)^\top c)}{\partial\theta^\top}\nu(u)f(\chi+hu,y)dydu\right\}^{-1}\right\|_F.$$

Assumption A8. *There exists a function $\tilde{g}_1(y)$ such that $\|g(y,\theta)\|\le\tilde{g}_1(y)$ where $E[\tilde{g}_1(Y)^{s_1}]<\infty$ for some $s_1>2$. There exists a function $\tilde{g}_2(y)$ such that $\|\partial g(y,\theta)/\partial\theta^\top\|_F\le\tilde{g}_2(y)$ where $E[\tilde{g}_2(Y)^{s_2}]<\infty$ for some $s_2>2$ and $\|\cdot\|_F$ refers to Frobenius norm. There exists a function $\tilde{g}_3(y)$ such that $\|\partial\text{vec}[\partial g(y,\theta)/\partial\theta^\top]/\partial\theta^\top\|_F\le\tilde{g}_3(y)$ where $E[\tilde{g}_3(Y)]<\infty$.*

Next we show the above assumptions hold for our applications in the paper.

Lemma 1.1. *Suppose Assumptions 1-6 in the paper hold. Then Assumptions A1-A8 above hold.*

Proof. Assumption 1-6 imply Assumptions A1-A6 with the loss function being $\rho(\theta,y)=\ell(\theta;y)$. The finite third moments of Y in Assumption 1 implies assumptions A7-A8. This is because the derivatives of the score are given by

$$\begin{aligned}g(y,\theta)=&-y+\partial b(\theta)/\partial\theta\\ \frac{\partial g(y,\theta)}{\partial\theta^\top}=&\partial^2 b(\theta)/\partial\theta\partial\theta^\top\\ \frac{\partial}{\partial\theta^\top}\text{vec}\left\{\frac{\partial g(y,\theta)}{\partial\theta^\top}\right\}=&\frac{\partial}{\partial\theta^\top}\text{vec}\left\{\partial^2 b(\theta)/\partial\theta\partial\theta^\top\right\}.\end{aligned}$$

So expectation of the square of $g(y,\theta)$ is bounded by finite third moments of Y, smoothness of the cumulant generating function, and compactness of Θ and Θ_c. Similarly, the derivatives of g are just continuous derivatives of the cumulant generating function, and so are bounded over compact parameter spaces. This boundedness means the quantities in A7 are bounded by quantities free of Y, so the finite moments of the kernel will suffice to show that A7 holds. Similarly, the boundedness implies A8 holds with the upper bound being the third moment of Y with some constant adjustment. ■

1.3. **Conditional Risk.** We proceed to prove the theorems and propositions in generality using Assumptions A1 - A8. The following Lemma gives an expression for the derivative of the minimizer of the conditional risk in Assumption A3, $\theta(\chi)$.

Lemma 1.2. *Under assumptions A1-A4, we have*

$$\frac{\partial\theta(\chi)}{\partial\chi^\top} = -\left\{E\left[\frac{\partial g(y,\theta(\chi))}{\partial\theta^\top}\middle|\chi\right]\right\}^{-1} E\left[g(y,\theta(\chi))\frac{\partial\log f(y|\chi)}{\partial\chi^\top}\middle|\chi\right]$$

where $g(y,\theta) = \partial\rho(y,\theta)/\partial\theta$.

Proof. Since $R(\theta,\chi)$ has unique a minimizer at $\theta(\chi)$, taking derivatives of the conditional risk gives

$$0 = \partial R(\theta(\chi),\chi)/\partial\theta = E[\partial\rho(y,\theta(\chi))/\partial\theta|X=\chi] = E[g(y,\theta(\chi))|X=\chi]$$

Since the above holds for all χ, taking the derivative with respect to χ will again yield 0,

$$\begin{aligned}
0 &= \frac{\partial}{\partial\chi^\top}E[g(y,\theta(\chi))|X=\chi] \\
&= \frac{\partial}{\partial\chi^\top}\int g(y,\theta(\chi))f(y|\chi)dy \\
&= \int\frac{\partial}{\partial\chi^\top}g(y,\theta(\chi))f(y|\chi)dy \\
&= \int\frac{\partial g(y,\theta(\chi))}{\partial\theta^\top}\frac{\partial\theta(\chi)}{\partial\chi^\top}f(y|\chi)dy + \int g(y,\theta(\chi))\frac{\partial f(y|\chi)}{\partial\chi^\top}\frac{1}{f(y|\chi)}f(y|\chi)dy \\
&= E\left\{\frac{\partial g(y,\theta(\chi))}{\partial\theta^\top}\middle|\chi\right\}\frac{\partial\theta(\chi)}{\partial\chi^\top} + E\left\{g(y,\theta(\chi))\frac{\partial\log f(y|\chi)}{\partial\chi^\top}\middle|\chi\right\}.
\end{aligned}$$

Solving for $\frac{\partial\theta(\chi)}{\partial\chi^\top}$ yields the desired result. ■

2. Proofs for §2: Population Level Developments

2.1. **Proofs of Propositions.**

2.1.1. *(Proof of Proposition 2.1: when $\mathscr{S}_{E(Y|X)} = \mathscr{S}_{Y|X}$.)*

Proof. Let $P_{\mathscr{S}}$ denote the projection onto a sufficient dimension reduction subspace $\mathscr{S}$. Suppose that Y depends on X only through $E(Y|X)$, i.e. $Y \perp\!\!\!\perp X|E(Y|X)$.

($\mathscr{S}_{E(Y|X)} \subset \mathscr{S}_{Y|X}$): Suppose $\mathscr{S}$ is an SDR subspace for $Y|X$. Then we have that $Y \perp\!\!\!\perp X|P_{\mathscr{S}}X \Rightarrow E(Y|X) = E(Y|P_{\mathscr{S}}X)$. That is, every reduction for the central space is also one for the central mean space, giving us

$$\begin{aligned}
&\{\mathscr{S} : Y \perp\!\!\!\perp E(Y|X)|P_{\mathscr{S}}X\} \supset \{\mathscr{S} : Y \perp\!\!\!\perp X|P_{\mathscr{S}}X\} \\
\Longrightarrow\ &\cap\{\mathscr{S} : Y \perp\!\!\!\perp E(Y|X)|P_{\mathscr{S}}X\} \subset \cap\{\mathscr{S} : Y \perp\!\!\!\perp X|P_{\mathscr{S}}X\} \\
\Longrightarrow\ &\mathscr{S}_{E(Y|X)} \subset \mathscr{S}_{Y|X},
\end{aligned}$$

where the intersection is over all respective SDR subspaces. The containment reverses since there are more reductions spaces for the conditional mean than for the whole pair (Y,X).

($\mathscr{S}_{E(Y|X)} \supset \mathscr{S}_{Y|X}$): Let β form a basis for $\mathscr{S}_{E(Y|X)}$. Then from the definition of $\mathscr{S}_{E(Y|X)}$, we have $E(Y|X) = E(Y|\beta^\top X)$. Since $Y \perp\!\!\!\perp X|E(Y|X)$, from the semi-graphoid axioms (Li, 2018, Ch. 2), we get

$$\begin{aligned} &Y \perp\!\!\!\perp X|E(Y|X) \equiv Y \perp\!\!\!\perp X|E(Y|\beta^\top X) \\ \Rightarrow &Y \perp\!\!\!\perp (X, \beta^\top X)|E(Y|\beta^\top X) \\ \Rightarrow &Y \perp\!\!\!\perp X|\{E(Y|\beta^\top X), \beta^\top X\} \equiv Y \perp\!\!\!\perp X|\beta^\top X, \end{aligned}$$

where the second line follows from $X \mapsto (X, \beta^\top X)$ being measurable. The last line follows from $\sigma(E(Y|\beta^\top X)) \subset \sigma(\beta^\top X)$, so $\sigma(E(Y|\beta^\top X), \beta^\top X) = \sigma(\beta^\top X)$. Therefore, any reduction for the central mean space is a reduction for the central space. This implies that the central space is a subset of the central mean space, concluding the proof. ■

2.1.2. *(Proof of Lemma 2.1: Unbiasedness and Exhaustiveness of OPCG).*

Proof. For this proof, we use the $\dot{b}^{-1}$ and $\ddot{b}^{-1}$ notation to indicate the first and second derivative of b^{-1}. For (a), under the canonical link for a GLM, when $X = x$,

$$\theta(x) = \dot{b}^{-1}(\mu(\beta^\top x)) = \dot{b}^{-1}[E(Y|\beta^\top x)]$$

and if $\theta(x)$ is differentiable, we have the canonical gradient as

$$\frac{\partial \theta(x)^\top}{\partial x} = \frac{\partial u^\top}{\partial x}\frac{\partial \mu(u)^\top}{\partial u}\ddot{b}^{-1}(\mu) = \beta\frac{\partial \mu(u)^\top}{\partial u}\ddot{b}^{-1}(\mu) \in span(\beta),$$

where $\frac{\partial \mu(u)^\top}{\partial u}\ddot{b}^{-1}(\mu) \in \mathbb{R}^{d\times m}$. Therefore, $\partial\theta(x)^\top/\partial x \in span(\beta)$, completing the proof of (a).

For (b), we want to show exhaustiveness of OPCG when $U = \beta^\top X$ has convex support. Denote $\Lambda = E\left(\frac{\partial\theta(X)^\top}{\partial x}\frac{\partial\theta(X)}{\partial x^\top}\right) \in \mathbb{R}^{p\times p}$, then showing exhaustiveness means showing following column spaces are equal $span(\Lambda) = span(\beta)$. Denoting $\theta(x) = \tilde{\theta}(\beta^\top x)$, unbiasedness in (a) gives us

$$\Lambda = E\left(\frac{\partial\theta(X)^\top}{\partial x}\frac{\partial\theta(X)}{\partial x^\top}\right) = \beta E\left(\frac{\partial\tilde{\theta}(\beta^\top X)^\top}{\partial u}\frac{\partial\tilde{\theta}(\beta^\top X)}{\partial u^\top}\right)\beta^\top,$$

implying $span(\Lambda) \subseteq span(\beta)$. We just need to show that $span(\Lambda)$ is not a proper subspace $\mathscr{S}_{E(Y|X)}$. So suppose $span(\Lambda) \subsetneq span(\beta)$. Then there exists $\alpha \neq 0 \in \mathbb{R}^p$ such that $\alpha \in span(\beta)$ and $\alpha \notin span(\Lambda)$. We can assume that $\alpha \perp span(\Lambda)$. Since $\frac{\partial\theta(x)^\top}{\partial x} \in span(\Lambda)$, we have that

$$0 = \alpha^\top\frac{\partial\theta^\top(x)}{\partial x} = \alpha^\top\beta\frac{\partial\mu^\top(u)}{\partial u}\ddot{b}^{-1}(\mu) \in \mathbb{R}^{d\times m},$$

for all $u \in supp(U)$, where $(\alpha^\top\beta)^\top = \gamma \neq 0 \in \mathbb{R}^d$ since $\alpha \neq 0 \in span(\beta)$ and $\ddot{b}^{-1}(\mu) \in \mathbb{R}^{d\times m} \neq 0$. Then we see that $\gamma^\top\frac{\partial\mu^\top(u)}{\partial u}\ddot{b}^{-1}(\mu) = 0$ for all $u \in supp(U)$.

Let $u_1, u_2 \in supp(U)$ such that the segment is $u_2 - u_1$ is parallel to $\gamma \in \mathbb{R}^d$. We take the derivative of $\mu(u)$ at any point along the line $u_2 - u_1$, say at the point $u_0 = (1-\varepsilon)u_1 + \varepsilon u_2$ for some $\varepsilon > 0$. Then we get

$$\frac{\partial \mu^\top((1-\varepsilon)u_1 + \varepsilon u_2)}{\partial \varepsilon} = (u_2 - u_1)^\top \frac{\partial \mu^\top((1-\varepsilon)u_1 + \varepsilon u_2)}{\partial u}.$$

This implies that taking the derivative of $\theta(x) = \frac{\partial b^{-1}[\mu\{u(x)\}]}{\partial \mu}$ with respect to u, at any point along the line $u_2 - u_1$, say at $u_0 = (1-\varepsilon)u_1 + \varepsilon u_2$, gives us

$$\begin{aligned}
\frac{\partial \theta^\top((1-\varepsilon)u_1 + \varepsilon u_2)}{\partial \varepsilon} =& \frac{\partial \dot{b}^{-\top}(\mu\{[(1-\varepsilon)u_1 + \varepsilon u_2]\})}{\partial \varepsilon} \\
=& (u_2 - u_1)^\top \frac{\partial \mu^\top(u)}{\partial u}\bigg|_{u_0} \ddot{b}^{-1}(\mu) \\
\equiv& \gamma^\top \frac{\partial \mu^\top(u)}{\partial u}\bigg|_{u_0} \ddot{b}^{-1}(\mu) = 0,
\end{aligned}$$

where the last line follows from γ being parallel to $u_1 - u_2$. This implies that $\theta(x)$ does not change in the direction of $u_2 - u_1$, which occurs only if the conditional mean function does not change along the direction of $u_2 - u_1$, since $\gamma \neq 0$ and $\ddot{b} \neq 0$. This contradicts $\mathscr{S}_{E(Y|X)}$ being a minimal dimension reduction space, since $\mathscr{S}_{E(Y|X)} - span(\gamma)$ is then an even smaller SDR subspace for $E(Y|X)$. Therefore, the assumption that $span(\Lambda) \subsetneq span(\beta)$ is false, completing the proof of (b). ■

2.2. **Proofs of Theorems in §2.** We start by showing that the bias term converges uniformly to zero and compute its uniform rate of convergence:

$$\sup_{\chi \in \Omega_X} \|c_n(\chi) - c(\chi)\| \longrightarrow 0.$$

This will be used in showing the uniform consistency of $\hat{c}_n(\chi)$ and computing its rate of convergence. Denote the parameter space of c by $\Theta_c \subset \mathbb{R}^{m(p+1)}$ and let $h_n = h$. Since R_n depends on n solely through the bandwidth parameter h, so will $c_n(\chi)$. For notational convenience, we express the objective function, $R_n(c, \chi)$ as $Q(h, c, \chi)$ where

$$Q(h, c, \chi) = EK(h^{-1}(X - \chi))\rho(Y, [\nu(X - \chi)]^\top c).$$

and its minimizer $c_n(\chi)$ over Θ_c by $c(h, \chi)$. Since

$$\lim_{n\to\infty} \sup_{\chi \in \Omega_X} \|c_n(\chi) - c(\chi)\| = \lim_{h \downarrow 0} \sup_{\chi \in \Omega_X} \|c(h, \chi) - c(\chi)\|$$

it suffices to show that, as $h \downarrow 0$,

$$\sup_{\chi \in \Omega_X} \|c(h, \chi) - c(\chi)\| \longrightarrow 0.$$

Define the diagonal matrix $D(h)$ by

$$D(h) = \begin{pmatrix} 1 & 0 \\ 0 & hI_p \end{pmatrix},$$

and let $u = h^{-1}(x - \chi)$, so that $x = \chi + hu$ and $dx = h^p du$. Recalling $g(y, \theta) = \partial\rho(y, \theta)/\partial\theta$, we get that $c(h, \chi)$ is a zero of the equation

$$S(h, c, \chi) = h^{-p}[D(h) \otimes I_m]^{-1}\partial_1 Q(h, c, \chi) = \int K(u)\nu(u)g(y, \nu(hu)^\top c)f(\chi + hu, y)dydu, \tag{1}$$

i.e. $S(h, c(h, \chi), \chi) = 0$, where ∂_i denotes the partial derivative with respect to the i^{th} argument, for $i = 1, 2, 3$, so that $S(h, c, \chi)$ is proportional to the partial derivative of $Q(h, c, \chi)$.

Let $a(h, \chi)$ and $b(h, \chi)$ denote the first m and last mp components of $c(h, \chi)$, respectively. The next theorem gives the uniform convergence rates of $a(h, \chi)$ and $b(h, \chi)$ as $h \to 0$. The proof will use a string of Lemmas that will be presented in section 2.3.

Theorem 2.1. *If Assumptions A1-A7 hold, then, for $h \in [0, 1]$, where $h \downarrow 0$,*

$$\sup_{\chi\in\Omega_X} \|a(h, \chi) - \theta(\chi)\| = O(h^2), \qquad \sup_{\chi\in\Omega_X} \|b(h, \chi) - \text{vec}(\partial\theta(\chi)/\partial\chi^\top)\| = O(h).$$

Proof. We follow Fan et al. (1995) and re-parametrise c in (1) by centering c at $c(0+, \chi)$, which is the limit of the minimizer $c(h, \chi)$, and then re-scaling the last mp components of $c - c(0+, \chi)$ by h. This re-parametrises c to $c_1 = [D(h) \otimes I_m](c - c(0+, \chi))$, and the minimum $c(h, \chi)$ to

$$c_1(h, \chi) = [D(h) \otimes I_m](c(h, \chi) - c(0+, \chi)) = \begin{pmatrix} a(h, \chi) - a(0+, \chi) \\ h[b(h, \chi) - b(0+, \chi)] \end{pmatrix} = \begin{pmatrix} a_1(h, \chi) \\ b_1(h, \chi) \end{pmatrix},$$

where $a_1(h, \chi)$ and $b_1(h, \chi)$ denote the first m and last mp components of $c_1(h, \chi)$, respectively, and $c_1(0+, \chi) = 0$ by construction. The purpose of this re-parametrisation is so all components of $c_1(h, \chi)$ converge at the same rate.

Taylor expanding the solution $c_1(h, \chi)$ to second order about $\varepsilon > 0$ and letting $\varepsilon \downarrow 0$, we get

$$c_1(h, \chi) = h\dot{c}_1(0+, \chi) + \frac{1}{2}h^2\ddot{c}_1(h^\dagger, \chi), \tag{2}$$

where $\dot{c}_1(\cdot, \chi)$ and $\ddot{c}_1(\cdot, \chi)$ refers to differentiation in h, and $h^\dagger \in (0, h)$. By Lemma 2.2, $\dot{a}_1(0+, \chi) = 0$ for all χ, so the first m entries of (2) are

$$a_1(h, \chi) = h\dot{a}_1(0+, \chi) + \frac{1}{2}h^2\ddot{a}_1(h^\dagger, \chi) = \frac{1}{2}h^2\ddot{a}_1(h^\dagger, \chi).$$

By Lemma 2.5, $\sup_{h\in[0,1]} \sup_{\chi\in\Omega_X} \|\ddot{c}_1(h, \chi)\| < \infty$, and so taking supremum over χ, we get ,

$$\sup_{\chi\in\Omega_X} \|a_1(h, \chi)\| \leq \frac{1}{2}h^2 \sup_{h\in[0,1]} \sup_{\chi\in\Omega_X} \|\ddot{a}_1(h^\dagger, \chi)\| = O(h^2).$$

Since $a_1(h,\chi) = a(h,\chi) - a(0+,\chi)$ by construction, and $a(0+,\chi) = \theta(\chi)$ by Lemma 2.1, we get

$$\sup_{\chi\in\Omega_X} \|a(h,\chi) - \theta(\chi)\| = O(h^2),$$

as desired.

The last mp entries of (2) are

$$b_1(h,\chi) = h\dot{b}_1(0+,\chi) + \frac{1}{2}h^2\ddot{b}_1(h^\dagger,\chi) = h[\mathrm{vec}(\partial\theta(\chi)/\partial\chi^\top) - b(0+,\chi)] + \frac{1}{2}h^2\ddot{b}_1(h^\dagger,\chi)$$

where the second equality follows from by Lemma 2.2. Since $b_1(h,\chi) = h[b(h,\chi) - b(0+,\chi)]$ by construction, we get

$$h[b(h,\chi) - b(0+,\chi)] = h[\mathrm{vec}(\partial\theta(\chi)/\partial\chi^\top) - b(0+,\chi)] + \frac{1}{2}h^2\ddot{b}_1(h^\dagger,\chi).$$

Re-arranging the above gives us

$$b(h,\chi) - \mathrm{vec}(\partial\theta(\chi)/\partial\chi^\top) = \frac{1}{2}h\ddot{b}_1(h^\dagger,\chi).$$

Taking supremum over $\chi\in\Omega_X$, we get by Lemma 2.5,

$$\sup_{\chi\in\Omega_X} \|b(h,\chi) - \mathrm{vec}(\partial\theta(\chi)/\partial\chi^\top)\| \le \frac{1}{2}h \sup_{h\in[0,1],\chi\in\Omega_X} \|\ddot{b}_1(h^\dagger,\chi)\| = O(h),$$

as desired. ■

Let $B(h,\chi) = \mathrm{mat}\{b(h,\chi)\} \in \mathbb{R}^{p\times d}$. We use the uniform fisher consistency above to show that the candidate matrices, and their corresponding eigenvectors, are also fisher consistent.

Theorem 2.2. *If Assumptions A1-A7 hold, then, for $h\in[0,1]$, where $h\downarrow 0$,*

(a) Then $\Lambda(h) = E\{B(h,X)B(h,X)^\top\}$ is fisher consistent for

$$\Lambda = E\left\{\frac{\partial\theta(X)^\top}{\partial\chi}\frac{\partial\theta(X)}{\partial\chi^\top}\right\},$$

and $\|\Lambda(h) - \Lambda\|_F = O(h)$.

(b) Let $\eta(h), \eta$ be matrices with columns comprised of the first d eigenvectors of $\Lambda(h), \Lambda$ respectively. Then

$$m\{\eta(h),\eta\} = \|\eta(h)\eta(h)^\top - \eta\eta^\top\|_F = O(h).$$

Proof. For (a), we have

$$\begin{aligned}\|\Lambda(h) - \Lambda\|_F =& \left\| E\{B(h,X)B(h,X)^\top\} - E\left\{\frac{\partial\theta(X)^\top}{\partial\chi}\frac{\partial\theta(X)}{\partial\chi^\top}\right\}\right\|_F \\ =& \left\| E\left\{B(h,X)B(h,X)^\top - \frac{\partial\theta(X)^\top}{\partial\chi}\frac{\partial\theta(X)}{\partial\chi^\top}\right\}\right\|_F\end{aligned}$$

$$\leq E\left\{\left\|B(h,X)B(h,X)^\top - \frac{\partial\theta(X)^\top}{\partial\chi}\frac{\partial\theta(X)}{\partial\chi^\top}\right\|_F\right\},$$

where the last inequality follows from Jensen's Inequality. Note that we can re-arrange the difference as

$$\begin{aligned}
&B(h,X)B(h,X)^\top - \frac{\partial\theta(X)^\top}{\partial\chi}\frac{\partial\theta(X)}{\partial\chi^\top}\\
=&\left\{B(h,X) - \frac{\partial\theta(X)^\top}{\partial\chi}\right\}\left\{B(h,X)^\top - \frac{\partial\theta(X)^\top}{\partial\chi}\right\}^\top + \left\{B(h,X) - \frac{\partial\theta(X)^\top}{\partial\chi}\right\}\frac{\partial\theta(X)}{\partial\chi^\top}\\
&+ \frac{\partial\theta(X)^\top}{\partial\chi}\left\{B(h,X) - \frac{\partial\theta(X)^\top}{\partial\chi}\right\}.
\end{aligned}$$

Then, with $\|\cdot\|_F = \|\text{vec}(\cdot)\|$, Theorem 2.1, smoothness of $\theta(\cdot)$ and compactness of Ω_X, we get

$$\begin{aligned}
&E\left\{\left\|B(h,X)B(h,X)^\top - \frac{\partial\theta(X)^\top}{\partial\chi}\frac{\partial\theta(X)}{\partial\chi^\top}\right\|_F\right\}\\
\leq&\left[\sup_{\chi\in\Omega_X}\left\|B(h,\chi) - \frac{\partial\theta(\chi)^\top}{\partial\chi}\right\|_F\right]^2 + 2\sup_{\chi\in\Omega_X}\left\|B(h,\chi) - \frac{\partial\theta(\chi)^\top}{\partial\chi}\right\|_F \times \sup_{\chi\in\Omega_X}\left\|\frac{\partial\theta(\chi)}{\partial\chi^\top}\right\|_F\\
\leq&O(h^2+h) = O(h),
\end{aligned}$$

which completes the proof for (a).

For (b), since $\|\Lambda(h) - \Lambda\|_F = O(h)$, by Lemma 6.3(b) of Bai, Miao and Rao, we get the final result

$$\|\eta(h)\eta(h)^\top - \eta\eta^\top\|_F \leq \sum_{k=1}^{d}\|\eta_k(h)\eta_k(h)^\top - \eta_k\eta_k^\top\|_F = O(h),$$

where $\eta_k(h)$ and η_k are the k^{th} columns of $\eta(h)$ and η, respectively. This completes the proof for determining the uniform rate of convergence for OPCG. ■

2.3. **Lemmas for Theorem 2.1.** Before proceeding to prove the requisite lemmas for Theorem 2.1, the re-parametrisation $c_1 = [D(h)\otimes I_m][c - c(0+,\chi)]$ implies

$$\phi(c_1) = c = [D(h)\otimes I_m]^{-1}c_1 + c(0+,\chi),$$

where $\dot\phi(c_1) = [D(h)\otimes I_m]^{-1}$. Then the objective function becomes

$$\begin{aligned}
Q(h,\phi(c_1),\chi) =& EK[h^{-1}(X-\chi)]\rho(Y,\nu(X-\chi)^\top\phi(c_1))\\
=& EK[h^{-1}(X-\chi)]\rho(Y,\nu(X-\chi)^\top c(0+,\chi) + \nu[h^{-1}(X-\chi)]^\top c_1),
\end{aligned}$$

where $c_1(h,\chi)$ minimizes $Q(h,\phi(c_1),\chi)$, since $c(h,\chi)$ minimizes $Q(c,h,\chi)$. By construction, we have $c_1(0+,\chi) = 0$ and $c_1(h,\chi)$ is a zero of

$$S(h,\phi(c_1),\chi) = \int K(u)\nu(u)g(y,\nu(hu)^\top\phi(c_1))f(\chi+hu,y)dydu$$

$$= \int K(u)\nu(u)g(y, \nu(hu)^\top c(0+,\chi) + \nu(u)^\top c_1) f(\chi + hu, y)dydu.$$

As in Fan et al. (1995), the components of $c_1(h,\chi)$ all converge in h at the same rate. Let $0_{k,l}, 1_{k,l} \in \mathbb{R}^{k\times l}$ be the $k \times l$ matrix of 0's and 1's, respectively. The next lemma gives $a(0+,\chi) = \theta(\chi)$.

Lemma 2.1. *If Assumptions A1-A6 hold, then for each χ, $a(0+,\chi) = \theta(\chi)$.*

Proof. Taking $h \downarrow 0$ in the equation $S(h, \phi(c_1(h,\chi)),\chi) = 0$, and recalling the conditions $\int K(u)du = 1$ and $c_1(0+,\chi) = 0$, we get

$$\begin{aligned}
0 &= S(0+, \phi(c_1(0+,\chi)),\chi) \\
&= \int K(u) \left(\begin{pmatrix} 1 \\ u \end{pmatrix} \otimes I_m \right) g(y, a(0+,\chi)) f(\chi, y)dydu \\
&= \int K(u) \begin{pmatrix} I_m \\ u \otimes I_m \end{pmatrix} du \int g(y, a(0+,\chi)) f(\chi, y)dy \\
&= \begin{pmatrix} I_m \\ 0_{mp,m} \end{pmatrix} \int g(y, a(0+,\chi)) f(\chi, y)dy.
\end{aligned}$$

The first m entries of the above equation are

$$0 = \int g(y, a(0+,\chi)) f(\chi, y)dy = f(\chi) \int g(y, a(0+,\chi)) f(y|\chi)dy. = f(\chi) E[g(Y, a(0+,\chi))|\chi]$$

Therefore, $a(0+,\chi)$ satisfies $E[g(Y, a(0+,\chi))|\chi] = 0$. But since $E[g(Y,\theta)|\chi]$ has a unique zero at $\theta(\chi)$, we conclude that $a(0+,\chi) = \theta(\chi)$. ■

The next lemma gives the limit of $\dot{a}_1(h,\chi)$ and $\dot{b}_1(h,\chi)$ as $h \downarrow 0$.

Lemma 2.2. *If Assumptions A1-A6 hold, then for each χ,*

$$\dot{a}_1(0+,\chi) = 0, \quad \text{and} \quad \dot{b}_1(0+,\chi) = \text{vec}(\partial\theta(\chi)/\partial\chi^\top) - b(0+,\chi).$$

Proof. Since $S(h, \phi(c_1(h,\chi)),\chi) = 0$ for all h, we have $\partial S(h, \phi(c_1(h,\chi)),\chi)/\partial h = 0$ for all h. Applying chain rule to $\partial S(h, \phi(c_1(h,\chi)),\chi)/\partial h$, and taking $h \downarrow 0$, gives us

$$\begin{aligned}
(3) \qquad 0 =& \lim_{h\downarrow 0} \frac{\partial S(h, \phi(c_1(h,\chi)),\chi)}{\partial h} \\
=& \partial_1 S(0+, \phi(c_1(0+,\chi)),\chi) + \lim_{h\downarrow 0} \partial_2 S(h, \phi(c_1(h,\chi)),\chi)[D(h) \otimes I_m]^{-1} \times \dot{c}_1(0+,\chi).
\end{aligned}$$

Using $\int K(u)du = 1$, $\int K(u)udu = 0$ and $\int K(u)uu^\top du = I_p$, we get

$$\begin{aligned}
\lim_{h\downarrow 0} \partial_2 S(h, \phi(c_1(h,\chi)),\chi)[D(h) \otimes I_m]^{-1} =& \int K(u)\nu(u) \frac{\partial g(y, a(0+,\chi))}{\partial\theta^\top} \nu(u)^\top f(\chi, y)dydu \\
=& I_{p+1} \otimes f(\chi) E\left(\frac{\partial g(y, a(0+,\chi))}{\partial\theta^\top} \middle| \chi \right).
\end{aligned}$$

Similarly, the term $\partial_1 S(0+, \phi(c_1(0+,\chi)),\chi)$ is computed as

$$\begin{aligned}\partial_1 S(0+,\phi(c_1(0+,\chi)),\chi) =& \int K(u)\nu(u)\frac{\partial g(y,a(0+,\chi))}{\partial\theta^\top}\left(\begin{pmatrix}0\\u^\top\end{pmatrix}\otimes I_m\right)f(\chi,y)dydu\times c(0+,\chi)\\ &+\int K(u)V(u)g(y,a(0+,\chi))\frac{\partial f(\chi,y)}{\partial x^\top}udydu\\ =&\begin{pmatrix}0_{m,1}\\ f(\chi)[I_p\otimes E(\partial g(y,a(0+,\chi))/\partial\theta^\top|\chi)]b(0+,\chi)\end{pmatrix}\\ &+\begin{pmatrix}0_{m,1}\\ f(\chi)\int[\partial\log f(y|\chi)/\partial\chi\otimes g(y,a(0+,\chi))]dy\end{pmatrix}.\end{aligned}$$

The first m components of (3) is

$$0 = E\left(\frac{\partial g(y,a(0+,\chi))}{\partial\theta^\top}\middle|\chi\right)\dot a_1(0+,\chi).$$

Since the loss function is assumed to be strictly convex, $E[\partial g(y,a(0+,\chi))/\partial\theta^\top|\chi]$ is positive-definite for all χ, which implies that, for all χ, $\dot a_1(0+,\chi)=0$.

Because $a\otimes b=\mathrm{vec}(ba^\top)$ for any vectors a and b, the last mp components of (3) are

$$\begin{aligned}0 =& f(\chi)\left(I_p\otimes E\left[\frac{\partial g(y,a(0+,\chi))}{\partial\theta^\top}\middle|\chi\right]\right)[b(0+,\chi)+\dot b_1(0+,\chi)]\\ &+f(\chi)\mathrm{vec}\left(E\left(g(y,a(0+,\chi))\frac{\partial\log f(\chi,y)}{\partial x^\top}\middle|\chi\right)\right).\end{aligned}$$

Since inverses of Kronecker products are Kronecker products of the inverses, we have

$$\begin{aligned}[b(0+,\chi)+\dot b_1(0+,\chi)] =& -\left(I_p\otimes E\left[\frac{\partial g(y,a(0+,\chi))}{\partial\theta^\top}\middle|\chi\right]^{-1}\right)\mathrm{vec}\left(E\left(g(y,a(0+,\chi))\frac{\partial\log f(\chi,y)}{\partial x^\top}\middle|\chi\right)\right)\\ =&\mathrm{vec}\left(-E\left[\frac{\partial g(y,a(0+,\chi))}{\partial\theta^\top}\middle|\chi\right]^{-1}E\left[g(y,a(0+,\chi))\frac{\partial\log f(y|\chi)}{\partial\chi^\top}\middle|\chi\right]\right)\\ =&\mathrm{vec}(\partial\theta(\chi)/\partial\chi^\top),\end{aligned}$$

where the last equality follows from Lemma 1.2. Re-arranging the last equality gives

$$\dot b_1(0+,\chi) = \mathrm{vec}(\partial\theta(\chi)/\partial\chi^\top) - b(0+,\chi),$$

which completes the proof. ■

The next lemma gives continuity of the solution $c_1(h,\chi)$.

Lemma 2.3. *Suppose Assumptions A1-A7 hold. Then, for $h\in[0,1]$ and $\chi\in\Omega_X$, $c_1(h,\chi)$ is continuous in h and χ. Furthermore, the image of $[0,1]\times\Omega_X$ under $c_1(h,\chi)$, denoted by A, is compact and $\sup_{h\in[0,1],\chi\in\Omega_X}\|c_1(h,\chi)\|<\infty$.*

Proof. Since $c_1(h,\chi)$ is a zero of $S(h,\phi(c_1),\chi)$, which is continuously differentiable in all its arguments, the implicit function theorem implies that $c_1(h,\chi)$ is continuous in h and χ. Let A be the continuous image of $[0,1]\times\Omega_X$ under $c_1(h,\chi)$. Since $[0,1]$ and Ω_X are compact, we get that A is also compact, since it is the continuous image of a compact set. The uniform boundedness follows from continuity of $c_1(h,\chi)$ over compact sets. ■

The next lemma gives uniform boundedness of $\dot{c}_1(h,\chi)$.

Lemma 2.4. *Suppose Assumptions A1-A7 hold. Then,* $\sup_{h\in[0,1],\chi\in\Omega_X}\|\dot{c}_1(h,\chi)\|<\infty$.

Proof. Since $S(h,\phi(c_1(h,\chi)),\chi)=0$ for all h, we have $\partial S(h,\phi(c_1(h,\chi)),\chi)/\partial h=0$ for all $h\in[0,1]$ and $\chi\in\Omega_X$. By chain rule, we have

$$\begin{aligned}0=&\frac{\partial S(h,\phi(c_1(h,\chi)),\chi)}{\partial h}\\ =&\partial_1 S(h,\phi(c_1(h,\chi)),\chi)+\partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h)\otimes I_m]^{-1}\dot{c}_1(h,\chi).\end{aligned}$$

Solving for $\dot{c}_1(h,\chi)$, we get

$$\dot{c}_1(h,\chi)=-\left\{\partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h)\otimes I_m]^{-1}\right\}^{-1}\partial_2 S(h,\phi(c_1(h,\chi)),\chi),$$

The two factors in $\dot{c}_1(h,\chi)$ above are

$$\begin{aligned}&\left\{\partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h)\otimes I_m]^{-1}\right\}^{-1}\\ =&\left\{\int K(u)\nu(u)\frac{\partial g(y,\nu(hu)^\top\phi(c_1(h,\chi)))}{\partial\theta^\top}\nu(u)f(\chi+hu,y)dydu\right\}^{-1}\\ \partial_1 S(h,\phi(c_1(h,\chi)),\chi)=&\int K(u)\nu(u)\frac{\partial g(y,\nu(hu)^\top\phi(c_1(h,\chi)))}{\partial\theta^\top}\left[\begin{pmatrix}0\\ u^\top\end{pmatrix}\otimes I_m\right]\\ &\times f(\chi+hu,y)dydu\;\; c(h,\chi)\\ &+\int K(u)\nu(u)g(y,\nu(hu)^\top\phi(c_1(h,\chi)))u^\top\frac{\partial f(\chi+hu,y)}{\partial x}dydu.\end{aligned}$$

By Lemma 2.3, the image of $[0,1]\times\Omega_X$ under $c_1(h,\chi)$, denoted by A, is compact. Since $c=\phi(c_1)$ is continuous, the image of A under ϕ, denoted by A_1 is also compact. Then, by Assumption A7, all the above integrals in $\dot{c}_1(h,\chi)$ are uniformly bounded over $h\in[0,1]$, $\chi\in\Omega_X$ and $\phi(c_1)=c\in A_1$. In particular, with Lemma 2.3, this implies

$$\sup_{h\in[0,1],\chi\in\Omega_X}\|\partial_1 S(h,\phi(c_1),\chi)\|\le\sup_{h\in[0,1],\chi\in\Omega_X,c\in A_1}\|\partial_1 S(h,c,\chi)\|<\infty.$$

Taking the norm and supremum over h and χ, we get

$$\begin{aligned}&\sup_{h\in[0,1],\chi\in\Omega_X}\|\dot{c}_1(h,\chi)\|\\ \le&\sup_{h\in[0,1],\chi\in\Omega_X,c\in A_1}\left\|\left\{\int K(u)\nu(u)\frac{\partial g(y,\nu(hu)^\top c)}{\partial\theta^\top}\nu(u)f(\chi+hu,y)dydu\right\}^{-1}\right\|_F\end{aligned}$$

$$\times \sup_{h\in[0,1],\chi\in\Omega_X,c\in A_1} \|\partial_1 S(h,c,\chi)\| < \infty,$$

completing the proof. ■

The next lemma gives uniform boundedness of $\ddot{c}_1(h,\chi)$.

Lemma 2.5. *If Assumptions A1-A7 hold, then* $\sup_{h\in[0,1],\chi\in\Omega_X} \|\ddot{c}_1(h,\chi)\| < \infty$.

Proof. Since $S(h,\phi(c_1(h,\chi)),\chi) = 0$ for all h, we have that $\partial^2 S(h,\phi(c_1(h,\chi)),\chi)/\partial h^2 = 0$ for all h and all $\chi \in \Omega_X$. Furthermore, because $S(h,\phi(c_1(h,\chi)),\chi)$ is a vector, $\partial S(h,\phi(c_1(h,\chi)),\chi)/\partial h$ is also a vector. This implies

$$\begin{aligned}
&\frac{\partial^2 S(h,\phi(c_1(h,\chi)),\chi)}{\partial h^2} \\
=&\frac{\partial}{\partial h}\text{vec}\left(\frac{\partial S(h,\phi(c_1(h,\chi)),\chi)}{\partial h}\right) \\
=&\frac{\partial}{\partial h}\text{vec}\Big(\partial_1 S(h,\phi(c_1(h,\chi)),\chi) + \partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h)\otimes I_m]^{-1}\dot{c}_1(h,\chi)\Big) \\
=&\frac{\partial}{\partial h}\Big(\partial_1 S(h,\phi(c_1(h,\chi)),\chi) + [\dot{c}_1(h,\chi)^\top \otimes I_{m(p+1)}]\text{vec}\{\partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h^{-1})\otimes I_m]\}\Big).
\end{aligned}$$

Carrying out the differentiation with respect to h on the right-hand side above via the chain rule, we get

$$\begin{aligned}
0 =&\frac{\partial^2 S(h,\phi(c_1(h,\chi)),\chi)}{\partial h^2} \\
=&\partial(\text{vec}\{\partial_1 S(h,\phi(c_1(h,\chi)),\chi)\})/\partial h \\
&+ \partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h^{-1})\otimes I_m]\ddot{c}_1(h,\chi) \\
&+ [\dot{c}_1(h,\chi)^\top \otimes I_{m(p+1)}]\partial(\partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h^{-1})\otimes I_m])/\partial h \\
=&\partial_2(\partial_1 S(h,\phi(c_1(h,\chi)),\chi))[D(h^{-1})\otimes I_m]\dot{c}_1(h,\chi) + \partial_1^2 S(h,\phi(c_1(h,\chi)),\chi) \\
&+ \partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h^{-1})\otimes I_m]\ddot{c}_1(h,\chi) \\
&+ [\dot{c}_1(h,\chi)^\top \otimes I_{m(p+1)}]\partial(\text{vec}\{\partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h^{-1})\otimes I_m]\})/\partial h.
\end{aligned}$$

Solving for $\ddot{c}_1(h,\chi)$, we get

$$\begin{aligned}
\ddot{c}_1(h,\chi) =& -\{\partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h^{-1})\otimes I_m]\}^{-1} \\
&\times\Big\{\partial_2(\partial_1 S(h,\phi(c_1(h,\chi)),\chi))[D(h^{-1})\otimes I_m]\dot{c}_1(h,\chi) + \partial_1^2 S(h,\phi(c_1(h,\chi)),\chi) \\
&[\dot{c}_1(h,\chi)^\top \otimes I_{m(p+1)}]\partial(\text{vec}\{\partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h^{-1})\otimes I_m]\})/\partial h\Big\}.
\end{aligned}$$

Before taking the norm and supremum of $\ddot{c}_1(h,\chi)$ over $h \in [0,1]$ and $\chi \in \Omega_X$, we need to show that the expressions involving h^{-1} are well-defined. Recall that

$$\partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h^{-1})\otimes I_m] = \int K(u)\nu(u)\frac{\partial g(y,\nu(hu)^\top\phi(c_1(h,\chi)))}{\partial\theta^\top}\nu(u)f(\chi+hu,y)dydu.$$

Then we have to show the following expressions are not proportional to h^{-1}:

$$\partial(\mathrm{vec}\{\partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h^{-1})\otimes I_m]\})/\partial h, \quad \partial_2(\partial_1 S(h,\phi(c_1(h,\chi)),\chi))[D(h^{-1})\otimes I_m].$$

For $\partial(\mathrm{vec}\{\partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h^{-1})\otimes I_m]\})/\partial h$, interchanging the integral and derivative, we get

$$\begin{aligned}
&\frac{\partial}{\partial h}\int K(u)\nu(u)\frac{\partial g(y,\nu(hu)^\top\phi(c_1(h,\chi)))}{\partial\theta^\top}\nu(u)f(\chi+hu,y)dydu\\
&=\int K(u)\nu(u)\frac{\partial}{\partial h}\left[\frac{\partial g(y,\nu(hu)^\top\phi(c_1(h,\chi)))}{\partial\theta^\top}\nu(u)f(\chi+hu,y)\right]dydu\\
&=\int K(u)\nu(u)[\nu(u)^\top\otimes I_m]\frac{\partial}{\partial\theta^\top}\mathrm{vec}\left[\frac{\partial g(y,\nu(hu)^\top\phi(c_1(h,\chi)))}{\partial\theta^\top}\right]\frac{\partial[\nu(hu)^\top\phi(c_1(h,\chi))]}{\partial h}\\
&\quad\times f(\chi+hu,y)dydu\\
&\quad+\int K(u)\nu(u)\frac{\partial g(y,\nu(hu)^\top\phi(c_1(h,\chi)))}{\partial\theta^\top}\nu(u)u^\top\frac{\partial f(\chi+hu,y)}{\partial x}dydu.
\end{aligned}$$

By Assumption A7, we have

$$\begin{aligned}
&\left\|\frac{\partial(\mathrm{vec}\{\partial_2 S(h,\phi(c_1(h,\chi)),\chi)[D(h^{-1})\otimes I_m]\})}{\partial h}\right\|_F\\
&\le\sup_{h\in[0,1],c\in A_1,\chi\in\Omega_X}\left\|\int K(u)\nu(u)[\nu(u)^\top\otimes I_m]\frac{\partial}{\partial\theta^\top}\mathrm{vec}\left[\frac{\partial g(y,\nu(hu)^\top c)}{\partial\theta^\top}\right]\frac{\partial[\nu(hu)^\top c]}{\partial h}\right.\\
&\quad\left.\times f(\chi+hu,y)dydu\right\|_F\\
&\quad+\sup_{h\in[0,1],c\in A_1,\chi\in\Omega_X}\left\|\int K(u)\nu(u)\frac{\partial g(y,\nu(hu)^\top c)}{\partial\theta^\top}\nu(u)u^\top\frac{\partial f(\chi+hu,y)}{\partial x}dydu\right\|_F<\infty.
\end{aligned}$$

For $\partial_2(\partial_1 S(h,\phi(c_1(h,\chi)),\chi))[D(h^{-1})\otimes I_m]$, recall

$$\begin{aligned}
\partial_1 S(h,c,\chi) &= \int K(u)\nu(u)\frac{\partial g(y,\nu(hu)^\top c)}{\partial\theta^\top}\left[\begin{pmatrix}0\\u^\top\end{pmatrix}\otimes I_m\right]cf(\chi+hu,y)dydu\\
&\quad+\int K(u)\nu(u)g(y,\nu(hu)^\top c)u^\top\frac{\partial f(\chi+hu,y)}{\partial x}dydu.
\end{aligned}$$

Now, taking the derivative with respect to c, we get

$$\partial_2(\partial_1 S(h,c,\chi))[D(h^{-1})\otimes I_m]$$

$$
\begin{aligned}
=&[c^\top \otimes I_{m(p+1)}] \int K(u) \left[\left[\begin{pmatrix} 0 \\ u^\top \end{pmatrix} \otimes I_m \right]^\top \otimes \nu(u) \right] \frac{\partial}{\partial \theta^\top} \mathrm{vec} \left[\frac{\partial g(y, \nu(hu)^\top c)}{\partial \theta^\top} \right] \nu(u) \\
&\times f(\chi + hu, y) dy du \\
&+ \int K(u) \nu(u) \frac{\partial g(y, \nu(hu)^\top c)}{\partial \theta^\top} \left[\begin{pmatrix} 0 \\ u^\top \end{pmatrix} \otimes I_m \right] f(\chi + hu, y) dy du \\
&+ \int K(u) \nu(u) \frac{\partial g(y, \nu(hu)^\top c) u^\top}{\partial \theta^\top} \nu(u) \frac{\partial f(\chi + hu, y)}{\partial x} dy du.
\end{aligned}
$$

Therefore, by Assumption A7, we have

$$
\begin{aligned}
&\|\partial_2(\partial_1 S(h, \phi(c_1(h,\chi)), \chi))[D(h^{-1}) \otimes I_m]\| \\
\leq &\sup_{c \in A_1} \|c^\top \otimes I_{m(p+1)}\|_F \times \sup_{h \in [0,1], c \in A_1, \chi \in \Omega_X} \left\| \int K(u) \left[\left[\begin{pmatrix} 0 \\ u^\top \end{pmatrix} \otimes I_m \right]^\top \otimes \nu(u) \right] \right. \\
&\left. \times \frac{\partial}{\partial \theta^\top} \mathrm{vec} \left[\frac{\partial g(y, \nu(hu)^\top c)}{\partial \theta^\top} \right] \nu(u) f(\chi + hu, y) dy du \right\|_F \\
&+ \sup_{h \in [0,1], c \in A_1, \chi \in \Omega_X} \left\| \int K(u) \nu(u) \frac{\partial g(y, \nu(hu)^\top c)}{\partial \theta^\top} \left[\begin{pmatrix} 0 \\ u^\top \end{pmatrix} \otimes I_m \right] f(\chi + hu, y) dy du \right\|_F \\
&+ \sup_{h \in [0,1], c \in A_1, \chi \in \Omega_X} \left\| \int K(u) \nu(u) \frac{\partial g(y, \nu(hu)^\top c) u^\top}{\partial \theta^\top} \nu(u) \frac{\partial f(\chi + hu, y)}{\partial x} dy du \right\|_F < \infty.
\end{aligned}
$$

Taking the norm of $\ddot{c}_1(h, \chi)$, we get

$$
\begin{aligned}
\|\ddot{c}_1(h,\chi)\| \leq &\left\| \left\{ \int K(u) \nu(u) \frac{\partial g(y, \nu(hu)^\top \phi(c_1(h,\chi)))}{\partial \theta^\top} \nu(u) f(\chi + hu, y) dy du \right\}^{-1} \right\|_F \\
&\times \left\{ O(1) \|\dot{c}_1(h,\chi)\| + \partial_1^2 S(h, \phi(c_1(h,\chi)), \chi) + O(1) \|\dot{c}_1(h,\chi)^\top \otimes I_{m(p+1)}\|_F \right\}.
\end{aligned}
$$

Since we saw the big-O terms are free of h , c, and χ, we take the supremum of $\|\ddot{c}_1(h,\chi)\|$ over h and χ, and appeal to Lemma 2.3, Lemma 2.4 and Assumption A7, to get

$$
\begin{aligned}
\sup_{h \in [0,1], \chi \in \Omega_X} \|\ddot{c}_1(h,\chi)\| \leq &\sup_{h \in [0,1], c \in A_1, \chi \in \Omega_X} \left\| \left\{ \int K(u) \nu(u) \frac{\partial g(y, \nu(hu)^\top c)}{\partial \theta^\top} \nu(u) f(\chi + hu, y) dy du \right\}^{-1} \right\|_F \\
&\times \left\{ O(1) \sup_{h \in [0,1], \chi \in \Omega_X} \|\dot{c}_1(h,\chi)\| + \sup_{h \in [0,1], c \in A_1, \chi \in \Omega_X} \|\partial_1^2 S(h, c, \chi)\| \right. \\
&\left. + O(1) \sup_{h \in [0,1], \chi \in \Omega_X} \|\dot{c}_1(h,\chi)^\top \otimes I_{m(p+1)}\|_F \right\} < \infty,
\end{aligned}
$$

which completes the proof. ■

3. Proofs for §3: Consistency of OPCG

3.1. **Proof of Theorems in §3.** In this section, we show uniform consistency of

$$\sup_{\chi\in\Omega_X} \|\hat{c}(\chi) - c(h,\chi)\| = o_{a.s}(1).$$

where we recall that $c(h,\chi) = c_n(\chi)$ is the minimizer of $Q(h,c,\chi) \equiv R_n(c,\chi)$ and $\hat{c}(\chi)$ is the minimizer of $\hat{R}_n(\chi)$. Let $\hat{S}_n(c,\chi) = h^{-p}[D(h^{-1})\otimes I_m]\partial_1\hat{R}_n(c,\chi)$. Then the minimizers $\hat{c}(\chi)$ and $c(h,\chi)$ are the zeroes of the equations

$$\begin{aligned}0 =& \hat{S}_n(\hat{c}(\chi),\chi) = h^{-p}[D(h^{-1})\otimes I_m]E_nK[(X-\chi)/h]\nu(X-\chi)g(Y,\nu(X-\chi)^\top\hat{c}(\chi)),\\ 0 =& S(c(h,\chi),h,\chi) = h^{-p}[D(h^{-1})\otimes I_m]EK[(X-\chi)/h]\nu(X-\chi)g(Y,\nu(X-\chi)^\top c(h,\chi)).\end{aligned}$$

The proof of the next theorem will use a string of lemmas that will be proven in section 3.2.

Theorem 3.1. *Suppose Assumptions A1-A8 hold. Then, as $n\to\infty$,*
(a) $\sup_{\chi\in\Omega_X}\|[D(h)\otimes I_m](\hat{c}(\chi)-c(h,\chi))\| = O_{a.s}(\delta_{ph})$,
(b) $\sup_{\chi\in\Omega_X}\|\hat{b}(\chi) - \mathrm{vec}(\partial\theta(\chi)/\partial\chi^\top)\| = O_{a.s}(h+h^{-1}\delta_{ph})$, *and*
(c) $\sup_{\chi\in\Omega_X}\|\hat{c}(\chi)-c(\chi)\| = o_{a.s}(1)$.

Proof.
We Taylor expand $\hat{S}_n(c,\chi)$ in c about $c_n(\chi)$, and evaluate at $\hat{c}(\chi)$, giving us

$$0 = \hat{S}_n(\hat{c}(\chi),\chi) = \hat{S}_n(c_n(\chi),\chi) + \partial_1\hat{S}_n(c^\dagger,\chi)(\hat{c}(\chi)-c_n(\chi)),$$

where $\|c^\dagger - c_n(\chi)\| \le \|\hat{c}(\chi)-c_n(\chi)\|$. Solving for $[D(h)\otimes I_m](\hat{c}(\chi)-c_n(\chi))$, we get

$$[D(h)\otimes I_m](\hat{c}(\chi)-c_n(\chi)) = -\{\partial_1\hat{S}_n(c^\dagger,\chi)[D(h^{-1})\otimes I_m]\}^{-1}\hat{S}_n(c_n(\chi),\chi).$$

Taking norm of both sides and observing that $E(\hat{S}_n(c_n(\chi),\chi)) = S_n(c_n(\chi),\chi) = 0$, we can upper bound the RHS as follows, where $\|\cdot\|$ refers to operator norm for matrices and euclidean otherwise,

$$\begin{aligned}&\|[D(h)\otimes I_m](\hat{c}(\chi)-c_n(\chi))\|\\ \le& \sup_{c\in\Theta_c}\|\{\partial_1\hat{S}_n(c,\chi)[D(h^{-1})\otimes I_m]\}^{-1}\|\|\hat{S}_n(c_n(\chi),\chi)\|\\ =& \sup_{c\in\Theta_c}\|\{\partial_1\hat{S}_n(c,\chi)[D(h^{-1})\otimes I_m]\}^{-1}\|\|\hat{S}_n(c_n(\chi),\chi) - S_n(c_n(\chi),\chi)\|\\ \le& \sup_{c\in\Theta_c}\|\{\partial_1\hat{S}_n(c,\chi)[D(h^{-1})\otimes I_m]\}^{-1}\|\sup_{c\in\Theta_c}\|\hat{S}_n(c,\chi) - S_n(c,\chi)\|.\end{aligned}$$

Since $E\partial_1\hat{S}_n(c,\chi) = \partial_1 S_n(c,\chi)$ by Assumption A4, the upper bound can be further decomposed into

$$\|[D(h)\otimes I_m](\hat{c}(\chi)-c_n(\chi))\|$$

$$\leq \sup_{c\in\Theta_c} \|\{\partial_1 \hat{S}_n(c,\chi)[D(h^{-1})\otimes I_m]\}^{-1}\| \sup_{c\in\Theta_c} \|\hat{S}_n(c,\chi) - S_n(c,\chi)\|$$
$$\leq \sup_{c\in\Theta_c} \|\{\partial_1 \hat{S}_n(c,\chi)[D(h^{-1})\otimes I_m]\}^{-1} - \{\partial_1 S_n(c,\chi)[D(h^{-1})\otimes I_m]\}^{-1}\| \sup_{c\in\Theta_c} \|\hat{S}_n(c,\chi) - S_n(c,\chi)\|$$
$$+ \sup_{c\in\Theta_c} \|\{\partial_1 S_n(c,\chi)[D(h^{-1})\otimes I_m]\}^{-1}\| \sup_{c\in\Theta_c} \|\hat{S}_n(c,\chi) - S_n(c,\chi)\|.$$

Taking supremum over $\chi \in \Omega_X$, we get

$$\sup_{\chi\in\Omega_X} \|[D(h)\otimes I_m](\hat{c}(\chi) - c_n(\chi))\| \leq T_1 T_2 + T_3 T_2,$$

where

$$T_1 = \sup_{\chi\in\Omega_X, c\in\Theta_c} \|\{\partial_1 \hat{S}_n(c,\chi)[D(h^{-1})\otimes I_m]\}^{-1} - \{\partial_1 S_n(c,\chi)[D(h^{-1})\otimes I_m]\}^{-1}\|$$
$$T_2 = \sup_{\chi\in\Omega_X, c\in\Theta_c} \|\hat{S}_n(c,\chi) - S_n(c,\chi)\|$$
$$T_3 = \sup_{\chi\in\Omega_X, c\in\Theta_c} \|\{\partial_1 S_n(c,\chi)[D(h^{-1})\otimes I_m]\}^{-1}\|$$

Since $S_n(c,\chi) = S(c,h,\chi)$, by Assumption A7,

$$T_3 \leq \sup_{h\in[0,1], c\in A, \chi\in\Omega_X} \left\| \left\{ \int K(u)\nu(u) \frac{\partial g(y,\nu(hu)^\top c)}{\partial \theta^\top} \nu(u) f(\chi + hu, y) dy du \right\}^{-1} \right\|_F < \infty.$$

Then, from Corollary 3.1 and Lemma 3.3, we get

$$\sup_{\chi\in\Omega_X} \|[D(h)\otimes I_m](\hat{c}(\chi) - c_n(\chi))\| \leq O_{a.s}(\delta_{ph}) O_{a.s}(\delta_{ph}) + O(1) O_{a.s}(\delta_{ph}) = O_{a.s}(\delta_{ph}),$$

which completes the proof of (a).

For (b), we can read off the last mp entries of $\|[D(h)\otimes I_m](\hat{c}(\chi) - c(h,\chi))\|$ to see that

$$\sup_{\chi\in\Omega_X} \|h\hat{b}(\chi) - b(h,\chi)\| = O_{a.s}(\delta_{ph}) \implies \sup_{\chi\in\Omega_X} \|\hat{b}(\chi) - b(h,\chi)\| = O_{a.s}(h^{-1}\delta_{ph}).$$

Combined with Theorem 2.1, we have

$$\sup_{\chi\in\Omega_X} \|\hat{b}(\chi) - \mathrm{vec}(\partial\theta(\chi)/\partial\chi^\top)\| \leq \sup_{\chi\in\Omega_X} \|\hat{b}(\chi) - b(h,\chi)\| + \sup_{\chi\in\Omega_X} \|b(h,\chi) - \mathrm{vec}(\partial\theta(\chi)/\partial\chi^\top)\|$$
$$= O_{a.s}(h + h^{-1}\delta_{ph}),$$

completing the proof of (b).

For (c), since $h^{-1}\delta_{ph} \to 0$ as $n \to \infty$ by Assumption A6, from parts (a) and (b) above, and Theorem 2.1, we get

$$\sup_{\chi\in\Omega_X} \|\hat{c}(\chi) - c(\chi)\| = \sup_{\chi\in\Omega_X} \left\| \begin{pmatrix} \hat{a}(\chi) - a(\chi) \\ \hat{b}(\chi) - b(\chi) \end{pmatrix} \right\| = \begin{pmatrix} O_{a.s}(h^2 + \delta_{ph}) \\ O_{a.s}(h + h^{-1}\delta_{ph}) \end{pmatrix} = o_{a.s}(1),$$

which completes the proof of (c). ■

The loss function $\rho(Y, \theta(X))$ is used to estimate the canonical parameter, $\theta(X)$, that summarizes the relation of interest between the predictor X and response Y. The Outer Product of Canonical Gradients (OPCG) estimator estimates the associated minimal SDR subspace by taking a spectral decomposition of the expected outer product of the gradients. We denote the true gradients by $B(\chi) = \partial\theta(\chi)^\top/\partial\chi$. Since $\hat{b}(\chi)$ is uniformly consistent for $\text{vec}(\partial\theta(\chi)/\partial\chi^\top)$, we know that $\hat{B}(\chi) = \text{mat}(\hat{b}(\chi))$ is uniformly consistent for $B(\chi)$. The candidate matrix for estimating the minimal sufficient dimension reduction subspace is then given by $\hat{\Lambda}_n = n^{-1}\sum_{j=1}^n \hat{B}(X_j)\hat{B}(X_j)^\top$.

Theorem 3.2. *Suppose the assumptions A1-A8 hold.*

(a) Then $\hat{\Lambda}_n = n^{-1}\sum_{j=1}^n \hat{B}(X_j)\hat{B}(X_j)^\top$ *consistently estimates* $\Lambda = E\{B(\chi)B(\chi)^\top\}$*, and*

$$\|\hat{\Lambda}_n - \Lambda\|_F = O_{a.s}(h + h^{-1}\delta_{ph} + \delta_n),$$

where $\delta_{ph} = (\log n/h^p n)^{1/2}$ *and* $\delta_n = (\log n/n)^{1/2}$.

(b) Let $\hat{\eta}, \eta \in \mathbb{R}^{p\times d}$ *be matrices with columns comprised of the first* d *eigenvectors of* $\hat{\Lambda}_n, \Lambda$ *respectively. Then*

$$m(\hat{\eta}, \eta) = \|\hat{\eta}\hat{\eta}^\top - \eta\eta^\top\|_F = O_{a.s}\left(h + h^{-1}\delta_{ph} + \delta_n\right)$$

where $\delta_{ph} = (\log n/h^p n)^{1/2}$ *and* $\delta_n = (\log n/n)^{1/2}$.

Proof. For (a), we have

$$\begin{aligned}
\|\hat{\Lambda}_n - \Lambda\|_F =& \left\| n^{-1}\sum_{j=1}^n \hat{B}(X_j)\hat{B}(X_j)^\top - EB(X)B(X)^\top \right\|_F \\
=& \left\| n^{-1}\sum_{j=1}^n \hat{B}(X_j)\hat{B}(X_j)^\top - E_nB(X)B(X)^\top + E_nB(X)B(X)^\top - EB(X)B(X)^\top \right\|_F \\
\leq& n^{-1}\sum_{j=1}^n \|\hat{B}(X_j)\hat{B}(X_j)^\top - B(X_j)B(X_j)^\top\|_F + \|E_n\{B(X)B(X)^\top - EB(X)B(X)^\top\}\|_F.
\end{aligned}$$

We use the uniform consistency by Theorem 3.1, and that $\|\cdot\|_F = \|\text{vec}(\cdot)\|$, to bound the first term above as follows:

$$\begin{aligned}
&\|\hat{B}(X_j)\hat{B}(X_j)^\top - B(X_j)B(X_j)^\top\|_F \\
=&\|[\{\hat{B}(X_j) - B(X_j)\} + B(X_j)][\{\hat{B}(X_j) - B(X_j)\} + B(X_j)]^\top - B(X_j)B(X_j)^\top\|_F \\
\leq&\|\{\hat{B}(X_j) - B(X_j)\}\{\hat{B}(X_j) - B(X_j)\}^\top\|_F + \|\{\hat{B}(X_j) - B(X_j)\}B(X_j)^\top\|_F \\
&+ \|B(X_j)\{\hat{B}(X_j) - B(X_j)\}^\top\|_F \\
\leq&\left(\sup_{\chi\in\Omega_X} \|\hat{B}(\chi) - B(\chi)\|_F\right)^2 + 2\sup_{\chi\in\Omega_X}\|B(\chi)\|_F \sup_{\chi\in\Omega_X}\|\hat{B}(\chi) - B(\chi)\|_F
\end{aligned}$$

$$= O_{a.s}[\{(h + h^{-1}\delta_{ph})\}^2] + O_{a.s}(h + h^{-1}\delta_{ph})$$
$$= O_{a.s}(h + h^{-1}\delta_{ph}).$$

Using Lemma 6.2 to bound the second term gives

$$\|E_n\{B(X)B(X)^\top - EB(X)B(X)^\top\}\|_F = O_{a.s}(\delta_n),$$

which gives the final result for (a),

$$\|\hat{\Lambda}_n - \Lambda\|_F = O_{a.s}(h + h^{-1}\delta_{ph} + \delta_n).$$

For (b), since $\|\hat{\Lambda}_n - \Lambda\|_F = O_{a.s}(h + h^{-1}\delta_{ph} + \delta_n)$, by Lemma 6.3(b) of Bai, Miao and Rao, we get the final result

$$\|\hat{\eta}\hat{\eta}^\top - \eta\eta^\top\|_F \leq \sum_{k=1}^{d} \|\hat{\eta}_k\hat{\eta}_k^\top - \eta_k\eta_k^\top\|_F = O(h + h^{-1}\delta_{ph} + \delta_n),$$

where $\hat{\eta}_k$ and η_k are the k^{th} columns of $\hat{\eta}$ and η, respectively. This completes the proof for determining the uniform rate of convergence for OPCG.

■

3.2. **Proof of Lemmas for Theorem 3.1.** To prove the string of lemmas for Theorem 3.1, we need the following proposition that shows our choice of bandwidth satisfies the second condition of Lemma 6.1.

Proposition 3.1. *Let $a_n = h^{p+k}$, where h satisfies Assumption A6 and $0 \leq k \leq 4$. Then, for $s > 2$, we have $a_n \downarrow 0$ and*

$$\frac{a_n^{s/(s-2)}n}{\log n} \to \infty, \quad as \quad n \to \infty.$$

Proof. By Assumption A6, $h = cn^{-\alpha}$ for $\alpha > 0$, and so $a_n \downarrow 0$ as $n \to \infty$. To compute the limit, note that

$$\frac{a_n^{s/(s-2)}n}{\log n} = \frac{[(cn^{-\alpha})^{p+k}]^{s/(s-2)}n}{\log n} = \frac{c^{(p+k)s/(s-2)}n^{-\alpha(p+k)s/(s-2)}n}{\log n} \propto \frac{n^{-\alpha(p+k)s/(s-2)+1}}{\log n}.$$

For $s > 2$, $0 < \alpha \leq 1/p_0$, and $p_0 > (p+4)s/(s-2)$, we get for $0 \leq k \leq 4$,

$$1 > 1 - \alpha\frac{(p+k)s}{(s-2)} \geq 1 - \frac{(p+k)s}{p_0(s-2)} > 0,$$

and so $n^{-\alpha(p+k)s/(s-2)+1}/\log n$ is monotonically diverging in n completing the proof. ■

In particular, Proposition 3.1 implies that h^{p+2} and h^{p+4} will satisfy assumption 2 of Lemma 6.1. The next lemma gives the uniform convergence of $\partial_1\hat{S}_n(c,\chi)[D(h^{-1}) \otimes I_m]$ in Theorem 3.1.

Lemma 3.1. *Suppose Assumptions A1-A8 hold. Then,*

$$\sup_{\chi\in\Omega_X, c\in\Theta_c} \|\partial_1\hat{S}_n(c,\chi)[D(h^{-1})\otimes I_m] - \partial_1 S_n(c,\chi)[D(h^{-1})\otimes I_m]\| = O_{a.s}(\delta_{ph}).$$

Proof. We appeal to Lemma 6.1 where we let $Z=(X,Y)$ and

$$m_n(c,\chi;Z) = h^2 K[(X-\chi)/h]\nu[(X-\chi)/h]\partial_2 g(Y,\nu(X-\chi)^\top c)\nu[(X-\chi)/h]^\top,$$

so that $\partial_1\hat{S}_n(c,\chi)[D(h^{-1})\otimes I_m] = h^{-p}h^{-2}E_n m_n(c,\chi;Z)$.

- For the first assumption, compactness of Ω_X, continuity of the kernel K, and the assumption $\|\partial_2 g(y,\theta)\|_F \le \|\tilde{g}_2(y)\|$ such that $E\|\tilde{g}_2(Y)^{s_2}\| < \infty$ for some $s_2 > 2$, we get that

$$\begin{aligned}
\|m_n(c,\chi;Z)\| =& h^2 K[(X-\chi)/h]\|\nu[(X-\chi)/h]\partial_2 g(Y,\nu(X-\chi)^\top c)\nu[(X-\chi)/h]^\top\|_F \\
\le& h^2 K[(X-\chi)/h]\|\nu[(X-\chi)]\|^2\|[D(h^{-1})\otimes I_m]\|_F^2\|\partial_2 g(Y,\nu(X-\chi)^\top c)\|_F \\
\le& h^2 C_K C_X^2(1+ph^{-2})m\tilde{g}_2(Y) \\
=& C_K C_X^2(h^2+p)m\tilde{g}_2(Y) \\
\le& C_K C_X^2(1+p)m\tilde{g}_2(Y),
\end{aligned}$$

  where C_K and C_X bounds the kernel and $\|\nu(X-\chi)\|_F$ respectively, and the norm of $\|[D(h^{-1})\otimes I_m]\|_F^2$ can be upper bounded by $(1+ph^{-2})m$, so the first assumption of Lemma 6.1 holds.

- For the second assumption of Lemma 6.1, since $m_n(c,\chi;Z)$ is a symmetric matrix, we just need to compute

$$\begin{aligned}
\sigma^2 =& \|E[m_n(c,\chi;Z)^\top m_n(c,\chi;Z)]\| \\
\le& \|Eh^4 K[(X-\chi)/h]^2\nu[(X-\chi)/h]\partial_2 g(Y,\nu(X-\chi)^\top c)\nu[(X-\chi)/h]^\top \\
& \times \nu[(X-\chi)/h]\partial_2 g(Y,\nu(X-\chi)^\top c)\nu[(X-\chi)/h]^\top\|_F \\
\le& C_K h^{p+4}\int K(u)\left\|\nu(u)\partial_2 g(y,\nu(hu)^\top c)\nu(u)^\top\nu(u)\partial_2 g(y,\nu(hu)^\top c)\nu(u)^\top\right\|_F f(\chi+huy)dudy \\
\le& C_K h^{p+4}\sup_{c\in\Theta_c, h\in[0,1], \chi\in\Omega_X}\int K(u)\left\|\nu(u)\partial_2 g(y,\nu(hu)^\top c)\nu(u)^\top\nu(u)\partial_2 g(y,\nu(hu)^\top c)\nu(u)^\top\right\|_F \\
& \times f(\chi+hu,y)dudy \\
\le& O(h^{p+4}),
\end{aligned}$$

  where the boundedness follows from Assumption A7. The second part of the assumption regarding h^{p+4} holds by Proposition 3.1.

- For the third assumption of Lemma 6.1, we need to check the Lipschitz condition in c for fixed χ and in χ for fixed c by Corollary 6.2. Since $m_n(c,\chi;Z)$ is matrix-valued,

we have

$$\|m_n(c,\chi;Z)\| \leq \|m_n(c,\chi;Z)\|_F = \|\text{vec}[m_n(c,\chi;Z)]\|,$$

where $\|\cdot\|$ refers to operator norm for matrices and euclidean norm for vectors. For Lipschitz in c, by the smoothness of g in its second argument, and the convexity of Θ_c, we can apply mean value theorem to get:

$$\begin{aligned}
\|m_n(c,\chi;Z) - m_n(c',\chi;Z)\| \leq& \|\text{vec}[m_n(c,\chi;Z)] - \text{vec}[m_n(c',\chi;Z)]\| \\
\leq& \|\partial \text{vec}[m_n(c^\dagger,\chi;Z)]/\partial c^\top\| \|c-c'\| \\
\leq& h^2 K[(X-\chi)/h] \|[\nu[(X-\chi)/h] \otimes \nu[(X-\chi)/h]]\|_F \\
&\times \|\partial \text{vec}[\partial_2 g(Y, [\nu(X-\chi)]^\top c^\dagger)]/\partial c^\top\|_F \|c-c'\| \\
\leq& h^2 C_K C_X^2 (1+ph^{-2}) m\tilde{g}_3(Y)\|c-c'\| \\
\leq& C_K C_X^2 (1+p) m\tilde{g}_3(Y)\|c-c'\| \\
\leq& h^{-1} C_K C_X^2 (1+p) m\tilde{g}_3(Y)\|c-c'\|,
\end{aligned}$$

where $c^\dagger$ lies between c and c', $h^{-1} \geq 1$, and Assumption A8 implies integrability of $\tilde{g}_3(Y)$.

For Lipschitz in χ, we can apply the same vectorization and mean value theorem argument as above to get:

$$\begin{aligned}
\|m_n(c,\chi;Z) - m_n(c,\chi';Z)\| \leq& \|\text{vec}[m_n(c,\chi;Z)] - \text{vec}[m_n(c,\chi';Z)]\| \\
\leq& \|\partial \text{vec}[m_n(c,\chi^\ddagger;Z)]/\partial \chi^\top\| \|\chi-\chi'\| \\
\leq& h^2 \|\partial \text{vec}[K[(X-\chi^\ddagger)/h]\nu[(X-\chi^\ddagger)/h] \\
&\times \partial_2 g(Y, \nu(X-\chi^\ddagger)^\top c)\nu[(X-\chi^\ddagger)/h]^\top]/\partial \chi^\top\|_F \|\chi-\chi'\| \\
\leq& h^{-1} C_{g,1}(Y)\|\chi-\chi'\|,
\end{aligned}$$

where $\chi^\ddagger$ lies between χ and χ', and the last inequality follows by Lemma 3.4 where $C_{g,1}(Y)$ being integrable.

Then, by Lemma 6.1 and Corollary 6.2, we get

$$\sup_{\chi\in\Omega_X, c\in\Theta_c} \|E_n m_n(c,\chi;Z) - E m_n(c,\chi;Z)\| = h^{p+4} O_{a.s}(\delta_{(p+4)h}).$$

Dividing above by $h^{(p+2)}$ on both sides, noting that $\delta_{(p+4)h} = h^{-2}\delta_{ph}$, and recalling that

$$\begin{aligned}
\partial_1 \hat{S}_n(c,\chi)[D(h^{-1}) \otimes I_m] =& h^{-(p+2)} E_n m_n(c,\chi;Z) \\
\partial_1 S_n(c,\chi)[D(h^{-1}) \otimes I_m] =& h^{-(p+2)} E m_n(c,\chi;Z),
\end{aligned}$$

we get

$$\sup_{\chi\in\Omega_X, c\in\Theta_c} \|\partial_1 \hat{S}_n(c,\chi)[D(h^{-1}) \otimes I_m] - \partial_1 S_n(c,\chi)[D(h^{-1}) \otimes I_m]\| = O_{a.s}(\delta_{ph}),$$

which completes the proof. ■

The next lemma and corollary gives the uniform convergence of T_1 in Theorem 3.1.

Lemma 3.2. *Let $\eta \in E$, where E is compact. Suppose a random invertible matrix $\hat{A}_n(\eta)$ and deterministic invertible matrix $A_n(\eta)$ satisfy*

$$\sup_{\eta\in E} \|\hat{A}_n(\eta) - A_n(\eta)\|_F = O_{a.s}(d_n),$$

where $d_n \to 0$. If $\sup_{\eta\in E} \|A_n(\eta)^{-1}\|_F = O(1)$, then

$$\sup_{\eta\in E} \|\hat{A}_n(\eta)^{-1} - A_n(\eta)^{-1}\|_F = O_{a.s}(d_n).$$

Proof. Since $\sup_{\eta\in E} \|\hat{A}_n(\eta) - A_n(\eta)\|_F = O_{a.s}(d_n)$, we have that $\hat{A}_n(\eta) = A_n(\eta) + D_n$, where $D_n = O_{a.s}(d_n)$ is meant entry-wise. Then,

$$\hat{A}_n(\eta)^{-1} =[A_n(\eta) + D_n]^{-1} = A_n(\eta)^{-1} + A_n(\eta)^{-1}D_n[A_n(\eta) + D_n]^{-1},$$

since, for invertible A and matrix D, we always have $[A + D]^{-1} = A^{-1} + A^{-1}D[A + D]^{-1}$. Because $d_n \to 0$, we have $\sup_{\eta\in E} \|[A_n(\eta) + D_n]^{-1}\|_F = O(1)$ and $\sup_{\eta\in E} \|A_n(\eta)^{-1}\|_F = O(1)$. This gives us

$$\sup_{\eta\in E} \|\hat{A}_n(\eta)^{-1} - A_n(\eta)^{-1}\|_F \leq \sup_{\eta\in E} \|A_n(\eta)^{-1}\|_F \times \|D_n\|_F \times \sup_{\eta\in E} \|[A_n(\eta) + D_n]^{-1}\|_F = O_{a.s}(d_n),$$

completing the proof. ■

Corollary 3.1. *Suppose Assumptions A1-A8 hold. Then,*

$$\sup_{\chi\in\Omega_X, c\in\Theta_c} \|\{\partial_1\hat{S}_n(c,\chi)[D(h^{-1}) \otimes I_m]\}^{-1} - \{\partial_1 S_n(c,\chi)[D(h^{-1}) \otimes I_m]\}^{-1}\|_F = O_{a.s}(\delta_{ph}).$$

Proof. We need to satisfy the assumptions of Lemma 3.2. By Lemma 3.1, we have

$$\sup_{\chi\in\Omega_X, c\in\Theta_c} \|\partial_1\hat{S}_n(c,\chi)[D(h^{-1}) \otimes I_m] - \partial_1 S_n(c,\chi)[D(h^{-1}) \otimes I_m]\|_F = O_{a.s}(\delta_{ph}),$$

where $\delta_{ph} \to 0$. We also have $\sup_{\chi\in\Omega_X, c\in\Theta_c} \|\{\partial_1 S_n(c,\chi)[D(h^{-1}) \otimes I_m]\}^{-1}\|_F = O(1)$ by Assumption A7. Then, by Lemma 3.2, we get

$$\sup_{\chi\in\Omega_X, c\in\Theta_c} \|\{\partial_1\hat{S}_n(c,\chi)[D(h^{-1}) \otimes I_m]\}^{-1} - \{\partial_1 S_n(c,\chi)[D(h^{-1}) \otimes I_m]\}^{-1}\|_F = O_{a.s}(\delta_{ph}),$$

completing the proof. ■

The next lemma gives the uniform convergence of T_3 in Theorem 3.1.

Lemma 3.3. *Suppose Assumptions A1-A8 hold. Then,*

$$\sup_{\chi\in\Omega_X, c\in\Theta_c} \|\hat{S}_n(c,\chi) - S_n(c,\chi)\| = O_{a.s}(\delta_{ph}).$$

Proof. We appeal to Lemma 6.1 where we let $Z = (X, Y)$ and

$$\begin{aligned} m_n(c, \chi; Z) =& hK[(X-\chi)/h][D(h^{-1}) \otimes I_m]\nu(X-\chi)g(Y, \nu(X-\chi)^\top c) \\ =& hK[(X-\chi)/h]\nu[(X-\chi)/h]g(Y, \nu(X-\chi)^\top c), \end{aligned}$$

so that $\hat{S}_n(c, \chi) = h^{-p}h^{-1}E_n m_n(c, \chi; Z)$.

- For the first assumption, compactness of Ω_X, continuity of the kernel K, and the assumption $|g(y, \theta)| \leq \tilde{g}_1(y)$ such that $E[\tilde{g}_1(Y)^{s_1}] < \infty$ for some $s_1 > 2$, we get that

$$\begin{aligned} \|m_n(c, \chi; Z)\| =& \|hK[(X-\chi)/h]\nu[(X-\chi)/h]g(Y, \nu(X-\chi)^\top c)\| \\ \leq& hC_K C_X(1 + ph^{-2})^{1/2}\sqrt{m}\tilde{g}_1(Y) \\ =& C_K C_X(h^2 + p)^{1/2}\sqrt{m}\tilde{g}_1(Y) \\ \leq& C_K C_X(1 + p)^{1/2}\sqrt{m}\tilde{g}_1(Y), \end{aligned}$$

  where C_K and C_X bounds the kernel and $\|\nu(X-\chi)\|$ respectively, and the norm of $\|[D(h^{-1}) \otimes I_m]\|_F$ can be upper bounded by $\sqrt{(1 + ph^{-2})m}$, so the first assumption of Lemma 6.1 holds.

- For the second assumption of Lemma 6.1, we need to compute

$$\begin{aligned} \sigma_1^2 =& |E[m_n(c, \chi; Z)^\top m_n(c, \chi; Z)]| \\ =& |Eh^2 K[(X-\chi)/h]^2 g(Y, \nu(X-\chi)^\top c)^\top \nu[(X-\chi)/h]^\top \nu[(X-\chi)/h]g(Y, \nu(X-\chi)^\top c)| \\ \leq& C_K h^{p+2} \int K(u) \left| g(y, \nu(hu)^\top c)^\top \nu(u)^\top \nu(u) g(y, \nu(hu)^\top c) \right| f(\chi + huy) du dy \\ \leq& C_K h^{p+2} \sup_{c \in \Theta_c, h \in [0,1], \chi \in \Omega_X} \int K(u) \left| g(y, \nu(hu)^\top c)^\top \nu(u)^\top \nu(u) g(y, \nu(hu)^\top c) \right| f(\chi + hu, y) du dy \\ \leq& O(h^{p+2}), \end{aligned}$$

  where the boundedness follows from Assumption A7. Similarly, we also need to compute

$$\begin{aligned} \sigma_2^2 =& |E[m_n(c, \chi; Z) m_n(c, \chi; Z)^\top]| \\ \leq& \|Eh^2 K[(X-\chi)/h]^2 \nu[(X-\chi)/h]g(Y, \nu(X-\chi)^\top c) g(Y, \nu(X-\chi)^\top c)^\top \nu[(X-\chi)/h]^\top \|_F \\ \leq& C_K h^{p+2} \int K(u) \left\| \nu(u) g(y, \nu(hu)^\top c) g(y, \nu(hu)^\top c)^\top \nu(u)^\top \right\|_F f(\chi + huy) du dy \\ \leq& C_K h^{p+2} \sup_{c \in \Theta_c, h \in [0,1], \chi \in \Omega_X} \int K(u) \left\| \nu(u) g(y, \nu(hu)^\top c) g(y, \nu(hu)^\top c)^\top \nu(u)^\top \right\|_F f(\chi + hu, y) du dy \\ \leq& O(h^{p+2}), \end{aligned}$$

  so that $\sigma^2 = \max\{\sigma_1^2, \sigma_2^2\} = O_{a.s}(h^{p+2})$. The second part of the assumption regarding h^{p+2} holds by Proposition 3.1.

- For the third assumption of Lemma 6.1, we need to check the Lipschitz condition in c for fixed χ and in χ for fixed c by Corollary 6.2. For Lipschitz in c, by the smoothness of g in its second argument, and the convexity of Θ_c, we can apply mean value theorem to get:

$$\begin{aligned}\|m_n(c,\chi;Z)-m_n(c',\chi;Z)\| \leq & \|\partial m_n(c^\dagger,\chi;Z)/\partial c^\top\|\|c-c'\| \\ \leq & C_K h\|\nu[(X-\chi)/h]\partial_2 g(Y,[\nu(X-\chi)]^\top c^\dagger)\nu(X-\chi)^\top\|_F\|c-c'\| \\ \leq & C_K C_X^2 h(1+ph^{-2})^{1/2}\sqrt{m}\|\partial_2 g(Y,[\nu(X-\chi)]^\top c^\dagger)\|_F\|c-c'\| \\ \leq & C_K h^{-1}\sqrt{(1+p)m}\tilde{g}_2(Y)\|c-c'\|\end{aligned}$$

where $c^\dagger$ lies between c and c', $h^{-1}\geq 1$ and Assumption A8 implies integrability of $\tilde{g}_2(Y)$.

For Lipschitz in χ, we again appeal to smoothness of g in θ and the convexity of Θ_c to apply mean value theorem to get:

$$\begin{aligned}&\|m_n(c,\chi;Z)-m_n(c,\chi';Z)\| \\ \leq & \|\partial m_n(c,\chi^\ddagger;Z)/\partial\chi^\top\|\|\chi-\chi'\| \\ = & h\|\partial[\mathrm{vec}\{K[(X-\chi^\ddagger)/h]\nu[(X-\chi^\ddagger)/h]g(Y,\nu(X-\chi^\ddagger)^\top c)\}]/\partial\chi^\top\|_F\|\chi-\chi'\| \\ \leq & h^{-1}C_{g,2}(Y)\|\chi-\chi'\|,\end{aligned}$$

where $\chi^\ddagger$ lies between χ and χ',and the last inequality follows by Lemma 3.4 with $C_{g,2}(Y)$ being integrable.

Then, by Lemma 6.1 and Corollary 6.2, we get

$$\sup_{\chi\in\Omega_X, c\in\Theta_c}\|E_n m_n(c,\chi;Z)-Em_n(c,\chi;Z)\| = h^{p+2}O_{a.s}(\delta_{(p+2)h}).$$

Note that $\delta_{(p+2)h} = h^{-1}\delta_{ph}$. Since $\hat{S}_n(c,\chi) = h^{-(p+1)}E_n m_n(c,\chi;Z)$ and $S_n(c,\chi) = h^{-(p+1)}Em_n(c,\chi;Z)$, dividing both sides by h^{p+1} gives

$$\sup_{\chi\in\Omega_X, c\in\Theta_c}\|\hat{S}_n(c,\chi)-S_n(c,\chi)\| = O_{a.s}(\delta_{ph}),$$

which completes the proof. ■

The next lemma shows that the derivatives used in the Lipschitz computations above are bounded, and so satisfy the conditions for Lemma 6.1.

Lemma 3.4. *Suppose Assumptions A1-A8 hold. Then, we have that for all $\chi\in\Omega_X$,*
(a) there exists $C_{g,1}(Y)$ such that $E[C_{g,1}(Y)]<\infty$ and

$$h^2\|\partial\mathrm{vec}[K[(X-\chi)/h]\nu[(X-\chi)/h]\partial_2 g(Y,\nu(X-\chi)^\top c)\nu(X-\chi)^\top]/\partial\chi^\top\|_F \leq h^{-1}C_{g,1}(Y)$$

(b) there exists $C_{g,2}(Y)$ such that $E[C_{g,2}(Y)]<\infty$ and

$$h\|\partial\mathrm{vec}[K[(X-\chi)/h]\nu[(X-\chi)/h]g(Y,\nu(X-\chi)^\top c)]/\partial\chi^\top\|_F \leq h^{-1}C_{g,2}(Y).$$

Proof. For (a), have

$$\begin{aligned}
&h\|\partial \text{vec}[K[(X-\chi)/h]\nu[(X-\chi)/h]\partial_2 g(Y,\nu(X-\chi)^\top c)\nu[(X-\chi)/h]^\top]/\partial\chi^\top\| \\
\leq& h\|[D(h^{-1})\otimes I_m]\|_F^2\left\|\frac{\partial \text{vec}[K[(X-\chi)/h]\nu(X-\chi)\partial_2 g(Y,\nu(X-\chi)^\top c)\nu(X-\chi)^\top]}{\partial\chi^\top}\right\| \\
=& h^{-1}(h^2+p)m\left\|\frac{\partial \text{vec}[K[(X-\chi)/h]\nu(X-\chi)\partial_2 g(Y,\nu(X-\chi)^\top c)\nu(X-\chi)^\top]}{\partial\chi^\top}\right\| \\
\leq& h^{-1}(1+p)m\left\|\frac{\partial \text{vec}[K[(X-\chi)/h]\nu(X-\chi)\partial_2 g(Y,\nu(X-\chi)^\top c)\nu(X-\chi)^\top]}{\partial\chi^\top}\right\|
\end{aligned}$$

Before proceeding with the calculations for the norm term above, recall that $\nu(X-\chi) = (1,(X-\chi)^\top)^\top \otimes I_m$, so that the Frobenius norm of the derivative of $\text{vec}[\nu(X-\chi)]$ with respect to χ is given by $\|\partial \text{vec}(\nu(X-\chi)^\top)/\partial\chi^\top\|_F = p^{1/2}$. Then the norm term is

$$\begin{aligned}
&\left\|\frac{\partial \text{vec}[K[(X-\chi)/h]\nu(X-\chi)\partial_2 g(Y,\nu(X-\chi)^\top c)\nu(X-\chi)^\top]}{\partial\chi^\top}\right\| \\
\leq&\left\| - h^{-1}\frac{\partial K(u)}{\partial u^\top}\text{vec}[\nu(X-\chi)\partial_2 g(Y,\nu(X-\chi)^\top c)\nu(X-\chi)^\top]\right\| \\
&+\left\|K[(X-\chi)/h][\nu(X-\chi)\partial_2 g(Y,\nu(X-\chi)^\top c)\otimes I_{m(p+1)}]\frac{\partial \text{vec}[\nu(X-\chi)]}{\partial\chi^\top}\right\|_F \\
&+\left\|K[(X-\chi)/h][\nu(X-\chi)\otimes\nu(X-\chi)]\frac{\partial \text{vec}[\partial_2 g(Y,\nu(X-\chi)^\top c)]}{\partial\theta^\top}\right. \\
&\left.\times[c^\top\otimes I_m]\frac{\partial \text{vec}(\nu(X-\chi)^\top)}{\partial\chi^\top}\right\|_F \\
&+\left\|K[(X-\chi)/h][I_{m(p+1)}\otimes\nu(X-\chi)\partial_2 g(Y,\nu(X-\chi)^\top c)]\frac{\partial \text{vec}[\nu(X-\chi)^\top]}{\partial\chi^\top}\right\|_F
\end{aligned}$$

Letting $C_{K,1}$ and C_c bound the derivative of the kernel and the parameter c respectively, and recalling $\|\text{vec}(\cdot)\| = \|\cdot\|_F$, we get

$$\begin{aligned}
\leq& h^{-1}C_{K,1}\|\nu(X-\chi)\|_F\partial_2 g(Y,\nu(X-\chi)^\top c)\|_F\left\|\nu(X-\chi)^\top\right\|_F \\
&+C_K\|\nu(X-\chi)\|_F\|\partial_2 g(Y,\nu(X-\chi)^\top c)\|\|I_{m(p+1)}\|_F\left\|\frac{\partial \text{vec}[\nu(X-\chi)]}{\partial\chi^\top}\right\|_F \\
&+C_K\|\nu(X-\chi)\|_F^2\left\|\frac{\partial \text{vec}[\partial_2 g(Y,\nu(X-\chi)^\top c)]}{\partial\theta^\top}\right\|_F\|c^\top\otimes I_m\|_F\left\|\frac{\partial \text{vec}(\nu(X-\chi)^\top)}{\partial\chi^\top}\right\|_F \\
&+C_K\|I_{m(p+1)}\|_F\|\nu(X-\chi)\|\|\partial_2 g(Y,\nu(X-\chi)^\top c)\|_F\left\|\frac{\partial \text{vec}[\nu(X-\chi)^\top]}{\partial\chi^\top}\right\|_F \\
\leq& h^{-1}C_{K,1}C_X^2\tilde{g}_2(Y)+C_KC_X\tilde{g}_2(Y)\sqrt{m(p+1)}p^{1/2}+C_KC_X^2\tilde{g}_3(Y)C_c\sqrt{m}p^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&+ C_K\sqrt{m(p+1)}C_X\tilde{g}_2(Y)p^{1/2}\\
\leq& h^{-1}[C_{K,1}C_X^2\tilde{g}_2(Y) + C_KC_X\tilde{g}_2(Y)\sqrt{m(p+1)}p^{1/2} + C_KC_X^2\tilde{g}_3(Y)C_c\sqrt{m}p^{1/2}\\
&+ C_K\sqrt{m(p+1)}C_X\tilde{g}_2(Y)p^{1/2}]\\
=& h^{-1}G_{g,1}(Y),
\end{aligned}
$$

where $G_{g,1}(Y) = C_{K,1}C_X^2\tilde{g}_2(Y) + C_KC_X\tilde{g}_2(Y)\sqrt{m(p+1)}p^{1/2} + C_KC_X^2\tilde{g}_3(Y)C_c\sqrt{m}p^{1/2} + C_K\sqrt{m(p+1)}C_X\tilde{g}_2(Y)p^{1/2}$, which is integrable by $\tilde{g}_2(Y)$ and $\tilde{g}_3(Y)$ being integrable by Assumption A8.

For (b), we have

$$
\begin{aligned}
&h\|\partial\{K[(X-\chi)/h]\nu[(X-\chi)/h]g(Y,\nu(X-\chi)^\top c)\}/\partial\chi^\top\|_F\\
\leq& h\left\|\frac{\partial K[(X-\chi)/h]}{\partial\chi^\top}\nu[(X-\chi)/h]g(Y,\nu(X-\chi)^\top c)\right\|_F\\
&+ h\left\|K[(X-\chi)/h]\{g(Y,\nu(X-\chi)^\top c)^\top\otimes I_{m(p+1)}\}[D(h^{-1})\otimes I_m]\frac{\partial\{\mathrm{vec}\{\nu(X-\chi)\}\}}{\partial\chi^\top}\right\|_F\\
&+ h\left\|K[(X-\chi)/h]\nu[(X-\chi)/h]\frac{\partial\{g(Y,\nu(X-\chi)^\top c)\}}{\partial\chi^\top}\right\|_F\\
\leq& h\left\|-h^{-1}\frac{\partial K(u)}{\partial u^\top}\right\|C_X(1+ph^{-2})^{1/2}\sqrt{m}\tilde{g}_1(Y)\\
&+ hC_K\tilde{g}_1(Y)\sqrt{m(p+1)}(1+ph^{-2})^{1/2}\sqrt{m}\left\|\frac{\partial\{\mathrm{vec}\{\nu(X-\chi)\}\}}{\partial\chi^\top}\right\|_F\\
&+ hC_KC_X(1+ph^{-2})^{1/2}\sqrt{m}\left\|\partial_2\{g(Y,\nu(X-\chi)^\top c)[c^\top\otimes I_m]\partial\mathrm{vec}(\nu(X-\chi)^\top)/\partial\chi^\top\right\|_F\\
\leq& h^{-1}C_{K,1}C_X(h^2+p)^{1/2}\sqrt{m}\tilde{g}_1(Y) + C_K\tilde{g}_1(Y)\sqrt{m(p+1)}(h^2+p)^{1/2}\sqrt{m}p^{1/2}\\
&+ C_KC_X(h^2+p)^{1/2}\sqrt{m}\tilde{g}_2(Y)C_c\sqrt{m}p^{1/2}\\
\leq& h^{-1}[C_{K,1}C_X(1+p)^{1/2}\sqrt{m}\tilde{g}_1(Y) + hC_K\tilde{g}_1(Y)m(p+1)p^{1/2} + hC_KC_XC_cp^{1/2}(1+p)^{1/2}m\tilde{g}_2(Y)]\\
\leq& h^{-1}C_{g,2}(Y),
\end{aligned}
$$

where $C_{g,2}(Y) = C_{K,1}C_X(1+p)^{1/2}\sqrt{m}\tilde{g}_1(Y) + C_K\tilde{g}_1(Y)m(p+1)p^{1/2} + C_KC_XC_cp^{1/2}(1+p)^{1/2}m\tilde{g}_2(Y)$ and is integrable by Assumption A8 on $\tilde{g}_1(Y)$ and $\tilde{g}_2(Y)$. This completes the proof of (b). ■

4. Computations for §4: Canonical Link and Cumulant Generating Functions

We need to construct initial values for the optimization algorithms used in estimating OPCG and MADE, such as Fisher-Scoring and the conjugate gradients algorithm used in

our paper. To do so, we apply the inverse canonical link transformation to the observed responses, as in Fisher-Scoring for GLM estimation McCullagh & Nelder (1989). We need the multivariate canonical link functions for the multinomial and adjacent-categories logit Agresti (2013, 2010), and derive them following McCullagh (1980).

4.1. **Multivariate Logistic Transformation.** Let $Y = (Y^1, ..., Y^{m-1})$ and $p = (p^1, ..., p^{m-1})$ be the unconstrained representation of a $multinomial\{k, (p^1, ..., p^{m-1}, 1 - \sum_{j=1}^{m-1} p^j)\}$ where $Y^m = k - \sum_{j=1}^{m-1} Y^j$. Note the mean of Y is p, and from standard computations, the negative log-likelihood of a multinomial implies the canonical link is $\theta = \log \frac{p}{1+\mathbf{1}^\top p}$. This gives us

$$e^\theta = \frac{p}{1 + \mathbf{1}^\top p} \Rightarrow p = e^\theta(1 - \mathbf{1}^\top p) = e^\theta - e^\theta(\mathbf{1}^\top p).$$

Re-arranging this and factoring out p gives us $e^\theta = (I_{m-1} + e^\theta \mathbf{1}^\top)p$. Employing the Sherman-Morrison-Woodbury formula results in

$$p = (I_{m-1} + e^\theta \mathbf{1}^\top)^{-1} e^\theta = e^\theta - \frac{e^\theta \mathbf{1}^\top e^\theta}{1 + \mathbf{1}^\top e^\theta} = e^\theta \left(1 - \frac{\mathbf{1}^\top e^\theta}{1 + \mathbf{1}^\top e^\theta}\right) \implies p = \frac{e^\theta}{1 + \mathbf{1}^\top e^\theta}.$$

Therefore, the logit and inverse logit transformations are $\theta = \log\{p/(1 + \mathbf{1}^\top p)\}$ and $p = e^\theta/(1 + \mathbf{1}^\top e^\theta)$, respectively, which are analogous to the scalar case.

4.2. **Adjacent-Categories Logit Link.** Suppose $\tilde{Y} \sim multinomial(k, p)$ and we associate $\tilde{Y}$ with $Y = (Y^1, ..., Y^m)$. When Y represents an ordinal variable, we transform Y to a cumulative variable $Z = R/k$, where $R^j = \sum_{l=1}^{j} Y^l$ for $j = 1, ..., m$. The variable Z can be interpreted as the empirical distribution over the categories for Y. The mean given is then the cumulative distribution over the m classes, and is denoted γ.

The multinomial density and log-likelihood for ordinal Y can be re-written as in McCullagh (1980):

$$f(y; p) \propto \prod_{j=1}^{m} (p^j)^{y^j} = \left[\left(\frac{\gamma^1}{\gamma^2}\right)^{R^1} \left(\frac{\gamma^2 - \gamma^1}{\gamma^2}\right)^{R^2 - R^1}\right] \left[\left(\frac{\gamma_2}{\gamma_3}\right)^{R_2} \left(\frac{\gamma_3 - \gamma_2}{\gamma_3}\right)^{R_3 - R_2}\right] \cdots$$
$$\cdots \left[\left(\frac{\gamma_{m-1}}{\gamma_m}\right)^{R_{m-1}} \left(\frac{\gamma_m - \gamma_{m-1}}{\gamma_m}\right)^{R_m - R_{m-1}}\right],$$
$$\ell(y; p) = \sum_{j=1}^{m-1} \left[R^j \log\left(\frac{\gamma^j}{\gamma^{j+1}}\right) + (R^{j+1} - R^j) \log\left(\frac{\gamma^{j+1} - \gamma^j}{\gamma^{j+1}}\right)\right]$$

Letting $Z^j = R^j/k$ for $j = 1, ..., m$, so that $Z^m = R^m/k = 1$, and defining

$$\phi^j = \log\left(\frac{\gamma^j}{\gamma^{j+1} - \gamma^j}\right), \quad g(\phi^j) = \log\left(\frac{\gamma^{j+1}}{\gamma^{j+1} - \gamma^j}\right),$$

we get the following relations:

$$\phi^j - g(\phi^j) = \log\left(\frac{\gamma^j}{\gamma^{j+1}}\right), \qquad \phi^j - g(\phi^{j-1}) = \log\left(\frac{\gamma^j - \gamma^{j-1}}{\gamma^{j+1} - \gamma^j}\right).$$

Then the log-likelihood can be re-expressed as

$$l(y;p) = k[Z^1\phi^1 + Z^2\{\phi^2 - g(\phi^1)\} + \cdots + Z^{m-1}\{\phi^{m-1} - g(\phi^{m-2})\} - g(\phi^{m-1})].$$

Letting the canonical parameter be $\theta = (\theta^1, ..., \theta^m)$, with entries

$$\theta^j = k[\phi^j - g(\phi^{j-1})] = k\log\left(\frac{\gamma^j - \gamma^{j-1}}{\gamma^{j+1} - \gamma^j}\right) = k\log(p^j/p^{j+1}),$$

for $j = 1, ..., m-1$. We refer to the canonical link that maps the conditional mean γ to θ as the Adjacent-Categories (Ad-Cat) transform or link function. Using the Ad-Cat link, we can write the log likelihood as

$$\ell(y;p) = k\left[Z^\top\theta - \{-\log(1-\gamma_{m-1})\}\right].$$

Setting $Z = k^{-1}U^\top Y$, where U is an upper triangular matrix of 1's, then $Y = kU^{-\top}Z$ and so the density with respect to Z is the same as Y with the jacobian term $k^{-(m-1)}$ being irrelevant for the log-likelihood. This means the log likelihood is now

$$\ell(Z;\gamma) = [Z^\top\theta - (-k\log(1 - e_{m-1}^\top\gamma))]$$

To check that the second term within the sum is indeed the log cumulant generating function, we need the relation between θ and γ, which we derive in the next section. Denoting

$$\exp\left\{\frac{\theta}{k}\right\} = \left(\frac{p^1}{p^2}, \frac{p^2}{p^3}, ..., \frac{p^{m-1}}{p^m}\right) = \left(\frac{\gamma^1 - \gamma^0}{\gamma^2 - \gamma^1}, \frac{\gamma^2 - \gamma^1}{\gamma^3 - \gamma^2}, ..., \frac{\gamma^{m-1} - \gamma^{m-2}}{\gamma^m - \gamma^{m-1}}\right),$$

we define the function $\tau(\theta) = (\tau^1(\theta), \tau^2(\theta), ..., \tau^{m-1}(\theta))$ with entries

$$\tau^j(\theta) = \sum_{r=1}^{j}\prod_{s=r}^{m-1}\exp\left\{\frac{\theta_s}{k}\right\} = \sum_{r=1}^{j}\exp\left\{\sum_{s=r}^{m-1}\frac{\theta_s}{k}\right\} = \frac{\gamma(\theta)^j}{1-\gamma(\theta)^{m-1}}.$$

With this function, we are able to define an inverse Ad-Cat transform from θ to the conditional mean γ as:

$$\gamma = \gamma(\theta) = \left[I_{m-1} - \frac{\tau(\theta)e_{m-1}^\top}{1 + e_{m-1}^\top\tau(\theta)}\right]\tau(\theta).$$

With the inverse Ad-Cat transform, we have that the log-likelihood has the following linear exponential family form

$$l(Z;\gamma) = [Z^\top\theta - [-k\log(1 - e_{m-1}^\top\gamma(\theta))]],$$

where the log-cumulant function is now given by $b(\theta) = -k\log(1 - e_{m-1}^\top\gamma(\theta))$.

4.3. **Mean and Variance-Covariance of cumulative Variable Z_i.** We derive the mean and variance-covariance of Z for the Adjacent-Categories Logit model for completeness. Derivations for the logistic model are more standard and omitted for brevity. Recall from the re-formulation of the density in the previous section, for $j < s$, we have $R^j|R^s \sim bin(R^s, (\gamma^1/\gamma^s, \gamma^2/\gamma^s, ..., 1))$. Then from iterative expectations, we get

$$E(Z^{m-1}|Z^m \equiv 1) = \frac{1}{k}E(R^{m-1}|Z^m \equiv 1) = \frac{1}{k}k\frac{\gamma^{m-1}}{\gamma^m} = \gamma^{m-1} \implies E(Z^{m-1}) = \gamma^{m-1}.$$

Then, with the above as a base case, we can proceed iteratively to obtain

$$E(Z^j|Z^{j+1}) = Z^{j+1}\frac{\gamma^j}{\gamma^{j+1}} \implies E(Z^j) = \gamma^j.$$

So the mean for Z is just the vector of cumulative probabilities, γ.

The variance can be computed similarly using the conditional binomial factors:

$$\begin{aligned}
Var(Z^{m-1}|Z^m) &= \frac{1}{k^2}Var(R^{m-1}|Z^m) = \frac{1}{k^2}R^m\frac{\gamma^{m-1}}{\gamma^m}\left(1 - \frac{\gamma^{m-1}}{\gamma^m}\right) \\
&= \frac{1}{k}\gamma^{m-1}\left(1 - \gamma^{m-1}\right) \\
Var(Z^{m-2}|Z^{m-1}) &= \frac{1}{k^2}Var(R^{m-2}|Z^{m-1}) = \frac{1}{k}Z^{m-1}\frac{\gamma^{m-2}}{\gamma^{m-1}}\left(1 - \frac{\gamma^{m-2}}{\gamma^{m-1}}\right)
\end{aligned}$$

And the conditional variance formula:

$$\begin{aligned}
Var(Z^{m-2}) &= E\{Var(Z^{m-2}|Z^{m-1})\} + Var\{E(Z^{m-2}|Z^{m-1})\} \\
&= \frac{1}{k}\frac{\gamma^{m-2}}{\gamma^{m-1}}\left(1 - \frac{\gamma^{m-2}}{\gamma^{m-1}}\right)E(Z^{m-1}) + Var(Z^{m-1})\frac{(\gamma^{m-2})^2}{(\gamma^{m-1})^2} \\
&= \frac{1}{k}\frac{\gamma^{m-2}}{\gamma^{m-1}}\left(1 - \frac{\gamma^{m-2}}{\gamma^{m-1}}\right)\gamma^{m-1} + \frac{1}{k}\gamma^{m-1}\left(1 - \gamma^{m-1}\right)\frac{(\gamma^{m-2})^2}{(\gamma^{m-1})^2} \\
&= \frac{1}{k}\left\{\frac{\gamma^{m-2}}{\gamma^{m-1}}\left(\gamma^{m-1} - \gamma^{m-2}\right) + \left(\gamma^{m-2} - \gamma^{m-1}\gamma^{m-2}\right)\frac{\gamma^{m-2}}{\gamma^{m-1}}\right\} \\
&= \frac{1}{k}\left\{\frac{\gamma^{m-2}}{\gamma^{m-1}}\left\{\left(\gamma^{m-1} - \gamma^{m-2}\right) + \left(\gamma^{m-2} - \gamma^{m-1}\gamma^{m-2}\right)\right\}\right\} \\
&= \frac{1}{k}\left\{\frac{\gamma^{m-2}}{\gamma^{m-1}}\left\{\gamma^{m-1} - \gamma^{m-1}\gamma^{m-2}\right\}\right\} \\
&= \frac{1}{k}\gamma^{m-2}\left\{1 - \gamma^{m-2}\right\} \\
\Rightarrow Var(Z^j) &= \frac{1}{k}\gamma^j\left\{1 - \gamma^j\right\}
\end{aligned}$$

Similarly, the covariance for $j < s$ is given by

$$Cov(Z^j, Z^s) = E(Z^jZ^s) - \gamma^j\gamma^s = E(Z^sE(Z^j|Z^s)) - \gamma^j\gamma^s$$

$$
\begin{aligned}
&= E\left((Z^s)^2\frac{\gamma^j}{\gamma^s}\right) - \gamma^j\gamma^s = \left(Var(Z^s) + (\gamma^s)^2\right)\frac{\gamma^j}{\gamma^s} - \gamma^j\gamma^s \\
&= \left(\frac{1}{k}\gamma^s\left\{1-\gamma^s\right\} + (\gamma^s)^2\right)\frac{\gamma^j}{\gamma^s} - \gamma^j\gamma^s \\
&= \left(\frac{1}{k}\gamma^j\left\{1-\gamma^s\right\} + \gamma^j\gamma^s\right) - \gamma^j\gamma^s \\
&= \frac{1}{k}\gamma^j(1-\gamma^s)
\end{aligned}
$$

And by symmetry, the covariance for $j > s$ is given by $Cov(Z^j, Z^s) = \frac{1}{k}\gamma^s(1-\gamma^j)$. Putting the variance and covariance into a matrix, we get

$$
\frac{1}{k}\begin{pmatrix}
\gamma^1\left\{1-\gamma^1\right\} & \gamma^1\left\{1-\gamma^2\right\} & \cdots & \gamma^1\left\{1-\gamma^{m-1}\right\} \\
\gamma^1\left\{1-\gamma^2\right\} & \gamma^2\left\{1-\gamma^2\right\} & \cdots & \gamma^2\left\{1-\gamma^{m-1}\right\} \\
\vdots & \vdots & \cdots & \vdots \\
\gamma^1\left\{1-\gamma^{m-1}\right\} & \gamma^2\left\{1-\gamma^{m-1}\right\} & \cdots & \gamma^{m-1}\left\{1-\gamma^{m-1}\right\}
\end{pmatrix}
$$

So the mean and variance covariance matrix of Z is given by

$$
E(Z) = \gamma \quad Var(Z) = \frac{1}{k}[\Gamma - \gamma\gamma^\top], \quad \text{where} \quad \Gamma = \begin{pmatrix}
\gamma^1 & \gamma^1 & \cdots & \gamma^1 \\
\gamma^1 & \gamma^2 & \cdots & \gamma^2 \\
\vdots & \vdots & \cdots & \vdots \\
\gamma^1 & \gamma^2 & \cdots & \gamma^{m-1}
\end{pmatrix}.
$$

5. Fisher-Scoring Algorithm

We can estimate OPCG and MADE, using their respective negative log-likelihoods as the objective, via a conventional Fisher-Scoring algorithm for GLMs, as opposed to the use of Grassmanian Optimization by Adragni (2018).

5.1. **OPCG.**

For OPCG, we minimize the full negative log-likelihood

$$
\ell(a_1, ..., a_n, B_1, ..., B_n; Y_{1:n}, X_{1:n}) = \sum_{j=1}^{n} \ell_j(a_j, B_j; Y_{1:n}, X_{1:n}),
$$

where the negative local log-likelihood $\ell_j(a_j, B_j; Y_{1:n}, X_{1:n})$ is

$$
\ell_j(a_j, B_j; Y_{1:n}, X_{1:n}) = \sum_{i=1}^{n} w_{ij}[-\{a_j + B_j^\top(X_i - X_j)\}^\top Y_i + b(a_j + B_j^\top(X_i - X_j))],
$$

Letting

$$
A_j = (a_j, B_j^\top)^\top \in \mathbb{R}^{(p+1)\times m}, \qquad c_j = vec(A_j) \in \mathbb{R}^{m(p+1)\times 1}, \qquad V_{ij} = (1, (X_i - X_j)^\top)^\top \in \mathbb{R}^{(p+1)\times 1},
$$

the local log-likelihood, score and observed information about X_j can be reframed as

$$\ell_j(c_j) = \sum_{i=1}^{n} w_{ij}\{c_j^\top [V_{ij} \otimes I_m] Y_i - b([V_{ij}^\top \otimes I_m] c_j)\}$$

$$S_j(c_j) = \sum_{i=1}^{n} w_{ij}\{[V_{ij} \otimes I_m] Y_i - [V_{ij} \otimes I_m]^\top \dot{b}\{(V_{ij} \otimes I_m)^\top c_j\}\}$$

$$J_j(c_j) = -\sum_{i=1}^{n} -w_{ij}(V_{ij}^\top \otimes I_m)^\top \frac{\partial}{\partial c_j \partial c_j^\top} b((V_{ij}^\top \otimes I_m)c_j)(V_{ij}^\top \otimes I_m)$$

Given an initial estimate for α_j, we then iterate $\hat{c}_j^{(r+1)} = \hat{c}_j^{(r)} + J_j^{-1}(\hat{c}_j^{(r)}) S_j(\hat{c}_j^{(r)})$ until convergence with respect to some criteria, such as relative euclidean distance the between iterations $||\hat{\alpha}_j^{(r+1)} - \hat{\alpha}_j^{(r+1)}||_2 / ||\hat{\alpha}_j^{(r)}||_2 < \varepsilon$, for some chosen tolerance $\varepsilon > 0$. We denote the iteration after convergence as $\hat{\alpha}_j^*$, and construct the estimate $\hat{A}_j = mat(\hat{\alpha}_j^*) \in \mathbb{R}^{(p+1)\times m}$, where the *mat* operation takes the first $p+1$ entries of $\hat{\alpha}_j^*$ as the first column, the next $p+1$ entries as the second column, etc. From $\hat{A}_j$, we extract the estimate for the canonical gradient as the $2, ..., p+1$ rows of $\hat{A}_j$. Using $\hat{B}_j$, we construct $\hat{\Lambda}_n = n^{-1} \sum_{j=1}^{n} \hat{B}_j \hat{B}_j^\top$ and take the d eigenvectors $\hat{\eta}_1, ..., \hat{\eta}_d$ corresponding to the largest d eigenvalues of $\hat{\Lambda}_n$ as our estimate for the SDR directions, $\hat{\beta}_{OPCG} = (\hat{\eta}_1, ..., \hat{\eta}_d) \in R^{p\times d}$.

Since we reformulate OPCG as fitting a GLM, we can construct the initial values for our procedure as we would for fitting a conventional GLM (McCullagh & Nelder, 1989). That is, we construct an initial value for $c_j \in \mathbb{R}^{(p+1)m}$ by constructing the initial estimator for $A_j \in \mathbb{R}^{(p+1)\times m}$ and vectorizing. For each j, we regress the matrix of link transformed responses $l(Y) \in \mathbb{R}^{m\times n}$ onto the predictors $V_j \in \mathbb{R}^{p\times n}$, whose columns are $V_{1j}, ..., V_{nj}$. The initial value of A_j is set as the least squares estimator $\hat{A}_j^{(0)} = (V_j^\top V_j)^{-1} V_j l(Y)^\top$, from which we construct the initial estimate for $\hat{c}_j^{(0)}$ as $vec(\hat{A}_j^{(0)})$.

5.2. **MADE.** A similar re-formulation can be done for MADE in order to estimate the SDR directions β directly. The main difference in procedure from OPCG is that MADE requires alternating between two GLM fitting problems, both of which can be solved using a Fisher-Scoring approach. For the presentation here, we abuse notation by letting $B_j \in \mathbb{R}^{d\times m}$. Like with OPCG, the minimization for MADE can be reduced to minimizing the negative local log-likelihood

$$\ell(a_1, ..., a_n, B_1, ..., B_n, \beta; Y_{1:n}, X_{1:n}) = \sum_{j=1}^{n} \ell_j(a_j, B_j, \beta; Y_{1:n}, X_{1:n}),$$

with the local negative log-likelihood $\ell_j(a_j, B_j, \beta; Y_{1:n}, X_{1:n})$ as

$$\ell_j(a_j, B_j, \beta) = \sum_{i=1}^{n} w_{ij}[\{a_j + B_j^\top \beta^\top (X_i - X_j)\}^\top Y_i - b(a_j + B_j^\top \beta^\top (X_i - X_j))].$$

To estimate β, we iterate between estimating a_j, B_j while holding β constant, and estimating β while fixing a_j, D_j at the most recent estimate. Both of these steps are conducted via Fisher-Scoring. Suppose β is fixed. Then we want to minimize the local log-likelihood with respect to a_j, B_j for $j = 1, ..., n$. Abusing notation from the previous section for OPCG, we let

$$A_j = (a_j, B_j^\top)^\top \in \mathbb{R}^{(d+1)\times m}, \;\; c_j = vec(A_j) \in \mathbb{R}^{m(d+1)\times 1}, \;\; V_{ij} = (1, \beta^\top(X_i - X_j)) \in \mathbb{R}^{(d+1)\times 1}.$$

Then this setup is equivalent to fitting a GLM as in OPCG in the previous section. Therefore, we can fit this iteration of MADE as an OPCG problem and construct initial values in the same way.

Once we obtain the estimates $\hat{a}_{1:n}, \hat{B}_{1:n}$, we pug them back into the negative log likelihood and fix them. The full local log-likelihood is then

$$\ell(\hat{a}_1, ..., \hat{a}_n, \hat{B}_1, ..., \hat{B}_n, \beta; Y_{1:n}, X_{1:n}) = \sum_{j=1}^{n} \ell_j(\hat{a}_j, \hat{B}_j, \beta; Y_{1:n}, X_{1:n}),$$

with the local negative log-likelihood $\ell_j(\hat{a}_j, \hat{B}_j, \beta; Y_{1:n}, X_{1:n})$ as

$$\ell_j(\hat{a}_j, \hat{B}_j, \beta) = \sum_{i=1}^{n} w_{ij}[\{\hat{a}_j + \hat{B}_j^\top \beta^\top (X_i - X_j)\}^\top Y_i - b(\hat{a}_j + \hat{B}_j^\top \beta^\top (X_i - X_j))].$$

Letting $\beta_v = vec(\beta^\top) \in \mathbb{R}^{dp}$, $\tilde{V}_{ij} = (X_i - X_j)^\top \otimes \hat{B}_j^\top \in \mathbb{R}^{m\times dp}$, and suppressing the fixed variables, the negative log-likelihood, score and information, as a function of β_v, can be expressed as

$$\begin{aligned}
\ell(\beta_v) &= \sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij}\{\beta_v^\top \tilde{V}_{ij}^\top Y_i - b(\hat{a}_j + \tilde{V}_{ij}\beta_v)\},\\
S(\beta_v) &= \sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij}\tilde{V}_{ij}^\top\{Y_i - \frac{\partial}{\partial \beta_v} b(\hat{a}_j + \tilde{V}_{ij}\beta_v)\},\\
J(\beta_v) &= -\sum_{j=1}^{n}\sum_{i=1}^{n} -w_{ij}\tilde{V}_{ij}^\top \frac{\partial b(\hat{a}_j + \tilde{V}_{ij}\beta_v)}{\partial \beta_v \partial \beta_v^\top}\tilde{V}_{ij}.
\end{aligned}$$

Given an initial estimate for β_v, we iterate $\hat{\beta}_v^{(r+1)} = \hat{\beta}_v^{(r)} + J^{-1}(\hat{\beta}_v^{(r)})S(\hat{\beta}_v^{(r)})$ until convergence according to some criteria. We denote the converged iterate result as $\hat{\beta}_v$ and set $\hat{\beta} = mat(\hat{\beta}_v)$. Holding β fixed at $\hat{\beta}$, we re-estimate $a_{1:n}, B_{1:n}$. This alternating estimation between these two steps continues until our estimates for $\hat{\beta}$ converges according to some criterion, such as in Frobenius norm. One option for the initial value of β is to use the OPCG estimate, $\hat{\beta}_{opcg}$.

6. Theorems and Lemmas

6.1. **Uniform Consistency for the mean with a parameter and index.**

Lemma 6.1. *Suppose $m_n(\eta, Z) \in \mathbb{R}^{d_1 \times d_2}$, $n = 1, 2 \ldots$, are matrix-valued measurable functions of Z, compactly supported for $\eta \in E \subset \mathbb{R}^d$. Suppose $\{Z_i : i = 1, \ldots, n\}$ is a random sample from Z. Furthermore, assume that the functions m_n satisfy the following conditions, where $\|\cdot\|$ denotes operator norm when referring to matrices and Euclidean norm when referring to vectors:*

(1) (Uniform Boundedness) Suppose

$$\sup_{\eta \in E} \|m_n(\eta, Z)\| \leq M(Z),$$

with $E(M^s(Z)) < \infty$ for some $s > 2$;

(2) (Uniformly Bounded Second Moment): Let

$$\sigma_1^2 = \sup_{\eta \in E} \|E[m_n(\eta, Z)^\top m_n(\eta, Z)]\|, \quad \sigma_2^2 = \sup_{\eta \in E} \|E[m_n(\eta, Z) m_n(\eta, Z)^\top]\|.$$

Suppose

$$\sigma^2 = \max(\sigma_1^2, \sigma_2^2) < a_n,$$

where $a_n \to 0$, and

$$\liminf_n \frac{a_n^{s/(s-2)} n}{\log n} = \liminf_n a_n^{2/(s-2)} \frac{a_n n}{\log n} > 0$$

(3) (Lipschitz for $\eta \in E$) For all $\eta \in E$,

$$\|m_n(\eta, Z) - m_n(\eta', Z)\| \leq \|\eta - \eta'\|^{c_1} n^{c_2} L(Z),$$

for some $c_1, c_2 > 0$, with $E|L(Z)| < \infty$;

Then, we have

$$\sup_{\eta \in E} \left\| E_n m_n(\eta, Z) - E m_n(\eta, Z) \right\| = O_{a.s.}\left(\sqrt{\frac{a_n \log n}{n}} \right)$$

Proof. Since E is compact, there exists a finite covering of E by N balls or radius r about η_k for $k = 1, .., N$. Then, by the triangle inequality, the supremum we are interested in is upper bounded by

$$\begin{aligned}
&\sup_{\eta \in E} \left\| n^{-1} \sum_i [m_n(\eta, Z_i) - E m_n(\eta, Z_i)] \right\| \\
\leq &\sup_{\eta \in \cup_k B_r(\eta_k)} \left\| n^{-1} \sum_i [m_n(\eta, Z_i) - E m_n(\eta, Z_i)] \right\| \\
= &\max_k \sup_{\eta \in B_r(\eta_k)} \left\| n^{-1} \sum_i \Big[m_n(\eta, Z_i) \pm m_n(\eta_k, Z_i) - E m_n(\eta, Z_i) \right. \\
&\left. \pm E m_n(\eta_k, Z_i) \Big] \right\|
\end{aligned}$$

$$
\begin{aligned}
&\leq \max_k \sup_{\eta \in B_r(\eta_k)} \left\{ \left\| n^{-1} \sum_i \left[m_n(\eta_k, Z_i) - E m_n(\eta_k, Z_i) \right] \right\| \right. \\
&\quad \left. + \left\| n^{-1} \sum_i \left[[m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] - E[m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] \right] \right\| \right\} \\
&\leq \max_{k=1,\ldots,N} \left\| n^{-1} \sum_i \left[m_n(\eta_k, Z_i) - E m_n(\eta_k, Z_i) \right] \right\| \\
&\quad + \max_k \sup_{\eta \in B_r(\eta_k)} \left\| n^{-1} \sum_i \left[[m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] - E[m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] \right] \right\| \\
(*) \qquad &= \max_{k=1,\ldots,N} R_{n,k,1} + \max_{k=1,\ldots,N} \sup_{\eta \in B_r(\eta_k)} R_{n,k,2}
\end{aligned}
$$

The strategy from this point is to truncate and apply Bernstein's to $\max_k R_{n,k,1}$ in order to determine the rate. The Lipschitz properties are used to control $\max_k \sup_{\eta \in B_r(\eta_k)} R_{n,k,2}$.

First Term: For the first term in $(*)$,

$$
\max_k R_{n,k,1} = \max_{k=1,\ldots,N} \left\| n^{-1} \sum_i \left[m_n(\eta_k, Z_i) - E m_n(\eta_k, Z_i) \right] \right\|.
$$

Define the following truncations of the random function $m_n(\eta_k, Z)$:

$$
\begin{aligned}
m_n^{(o)}(\eta_k, Z) &= m_n(\eta_k, Z) \mathbb{1}\{|M(Z)| \geq C_n\} \\
m_n^{(I)}(\eta_k, Z) &= m_n(\eta_k, Z) \mathbb{1}\{|M(Z)| < C_n\} \\
\xi_{k,i} &= m_n^{(I)}(\eta_k, Z_i) - E m_n^{(I)}(\eta_k, Z_i)
\end{aligned}
$$

for some constant C_n, which we will make explicit later. This gives us

$$
\begin{aligned}
&\max_k \left\| n^{-1} \sum_i m_n(\eta_k, Z_i) - E m_n(\eta_k, Z_i) \right\| \\
&\leq \max_k \left\| n^{-1} \sum_i m_n^{(o)}(\eta_k, Z_i) - E m_n^{(o)}(\eta_k, Z_i) \right\| + \max_k \left\| n^{-1} \sum_i \xi_{k,i} \right\| \\
(**) \qquad &\leq \max_k \left\| n^{-1} \sum_i m_n^{(o)}(\eta_k, Z_i) \right\| + \max_k \left\| n^{-1} \sum_i E m_n^{(o)}(\eta_k, Z_i) \right\| + \max_k \left\| n^{-1} \sum_i \xi_{k,i} \right\|
\end{aligned}
$$

For the last term in $(**)$, we have that $\xi_{k,i}$ is bounded by C_n, is zero mean $E\xi_{k,i} = 0$, and has

$$
\begin{aligned}
\|E(\xi_{k,i}^\top \xi_{k,i})\| = &\left\| E \left\{ \left[m_n(\eta_k, Z) \mathbb{1}\{|M(Z)| < C_n\} - E(m_n(\eta_k, Z) \mathbb{1}\{|M(Z)| < C_n\}) \right]^\top \right. \right. \\
&\left. \left. \times \left[m_n(\eta_k, Z) \mathbb{1}\{|M(Z)| < C_n\} - E(m_n(\eta_k, Z) \mathbb{1}\{|M(Z)| < C_n\}) \right] \right\} \right\|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left\| E\left\{ \left[m_n(\eta_k, Z) - E(m_n(\eta_k, Z)) \right]^\top \left[m_n(\eta_k, Z) - E(m_n(\eta_k, Z)) \right] \right\} \right\| \\
&\leq \| E\{ m_n(\eta_k, Z)^\top m_n(\eta_k, Z) \} \| < \sigma_1^2 \\
\| E(\xi_{k,i} \xi_{k,i}^\top) \| &= \left\| E\left\{ \left[m_n(\eta_k, Z) \mathbb{1}\{|M(Z)| < C_n\} - E(m_n(\eta_k, Z) \mathbb{1}\{|M(Z)| < C_n\}) \right] \right. \right. \\
&\quad \left. \left. \times \left[m_n(\eta_k, Z) \mathbb{1}\{|M(Z)| < C_n\} - E(m_n(\eta_k, Z) \mathbb{1}\{|M(Z)| < C_n\}) \right]^\top \right\} \right\| \\
&\leq \left\| E\left\{ \left[m_n(\eta_k, Z) - E(m_n(\eta_k, Z)) \right] \left[m_n(\eta_k, Z) - E(m_n(\eta_k, Z)) \right]^\top \right\} \right\| \\
&\leq \| E\{ m_n(\eta_k, Z) m_n(\eta_k, Z)^\top \} \| < \sigma_2^2,
\end{aligned}
$$

which, by assumption (2), implies that

$$
\sigma_{\xi_k}^2 = \max \left\{ \left\| \sum_i E(\xi_{k,i}^\top \xi_{k,i}) \right\|, \left\| \sum_i E(\xi_{k,i} \xi_{k,i}^\top) \right\| \right\} < n\sigma^2 < n a_n.
$$

By Bernstein's Inequality for rectangular matrices (Vershynin, 2018; Tropp, 2015), for any $\varepsilon_n > 0$,

$$
\begin{aligned}
P\left(\left\| n^{-1} \sum_i \xi_{k,i} \right\| > \varepsilon_n \right) &\leq P\left(\left\| \sum_i \xi_{k,i} \right\| > n\varepsilon_n \right) \\
&\leq 2(d_1 + d_2) \exp\left\{ -\frac{n^2 \varepsilon_n^2}{2\sigma_{\xi_k}^2 + (2/3) C_n n \varepsilon_n} \right\} \\
&\leq 2(d_1 + d_2) \exp\left\{ -\frac{n\varepsilon_n^2}{2a_n + (2/3) C_n \varepsilon_n} \right\},
\end{aligned}
$$

where the last inequality is independent of k and i.

So then, to obtain an almost sure convergence rate, we apply the first Borel-Cantelli Lemma to $\max_k \| n^{-1} \sum_i \xi_{k,i} \|$. For any $\varepsilon_n > 0$, we have

$$
\begin{aligned}
\sum_{n=1}^{\infty} P\left(\max_k \left\| n^{-1} \sum_i \xi_{k,i} \right\| > \varepsilon_n \right) &= \sum_{n=1}^{\infty} P\left(\bigcup_{k=1}^{N} \left\{ n^{-1} \sum_i \| \xi_{k,i} \| > \varepsilon_n \right\} \right) \\
&\leq \sum_{n=1}^{\infty} \sum_{k=1}^{N} P\left(n^{-1} \sum_i \| \xi_{k,i} \| > \varepsilon_n \right) \\
&\leq \sum_{n=1}^{\infty} N \max_k P\left(n^{-1} \sum_i \| \xi_{k,i} \| > \varepsilon_n \right)
\end{aligned}
$$

$$\leq 2(d_1+d_2)\sum_{n=1}^{\infty} N\exp\left\{-\frac{n\varepsilon_n^2}{2a_n+(2/3)C_n\varepsilon_n}\right\}.$$

In particular, if we select ε_n and C_n such that $C_n\varepsilon_n \asymp a_n$, and $n\varepsilon_n^2 \asymp a_n \log n$; then we get $\varepsilon_n \asymp (a_n \log n/n)^{1/2}$, $C_n \asymp (a_n n/\log n)^{1/2}$. Say $C_n\varepsilon_n = b_1 a_n$ and $n\varepsilon_n^2 = b_2 a_n \log n$ for $b_1, b_2 > 0$, then we get

$$\begin{aligned}2(d_1+d_2)\sum_{n=1}^{\infty} N\exp\left\{-\frac{n\varepsilon_n^2}{2a_n+(2/3)C_n\varepsilon_n}\right\} =& 2(d_1+d_2)\sum_{n=1}^{\infty} N\exp\left\{-\frac{b_2}{2+\frac{2}{3}b_1}\log n\right\}\\ \leq& 2(d_1+d_2)\sum_{n=1}^{\infty} N n^{-\frac{b_2}{2+\frac{2}{3}b_1}}.\end{aligned}$$

We will see later that the number of balls is dependent on n; in particular, $N = N(n) \propto n^{dc_2/c_1}(a_n \log n/n)^{-d/(2c_1)}$ for constants $c_1, c_2 > 0$. Then we need to pick $b_1, b_2 > 0$ large enough such that the series converges. That is, we need to pick $b_1, b_2 > 0$ such that

$$N(n) n^{-\frac{b_2}{2+\frac{2}{3}b_1}} < n^{-1}.$$

With this choice of b_1 and b_2, and letting $\varepsilon_n = b_2^{1/2}(a_n \log n/n)^{1/2}$, we get

$$\sum_{n=1}^{\infty} P\left(\max_k \left\| n^{-1}\sum_i \xi_{k,i}\right\| > b_2^{1/2}(a_n \log n/n)^{1/2}\right) \leq 2(d_1+d_2)\sum_{n=1}^{\infty} N(n) n^{-\frac{b_2}{2+\frac{2}{3}b_1}} < \infty.$$

By the first Borel-Cantelli Lemma, we get that the event

$$P\left(\left\{\omega : \max_k \left\| n^{-1}\sum_{i=1}^{n} \xi_{k,i}(\omega)\right\| > b_2^{1/2}(a_n \log n/n)^{1/2}\right\} \ i.o.\right) = 0,$$

meaning

$$\max_k \left\| n^{-1}\sum_i \xi_{k,i}\right\| = O_{a.s.}((a_n \log n/n)^{1/2}).$$

For the second term of $(**)$, we have that

$$\begin{aligned}\|E m_n^{(o)}(\eta_k, Z_i)\| \leq& E\|m_n^{(o)}(\eta_k, Z_i)\|\\ =& E\|m_n(\eta_j, Z_i)\|\mathbb{1}\{|M(Z_i)| \geq C_n\}\\ \leq& E[|M(Z_i)|\mathbb{1}\{|M(Z_i)| \geq C_n\}]\\ =& C_n^{-(s-1)} E|M(Z_i)|C_n^{s-1}\mathbb{1}\{|M(Z_i)| \geq C_n\}\\ \leq& C_n^{-(s-1)} E|M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \geq C_n\}\\ \leq& C_n^{1-s} E|M(Z_i)|^s,\end{aligned}$$

where $E|M(Z_i)|^s < \infty$ for $s > 2$ by assumption (2). So we need to show $C_n^{1-s} = O(\varepsilon_n)$. But since $C_n \asymp (a_n n/\log n)^{1/2}$, we have

$$\frac{C_n^{1-s}}{\varepsilon_n} = \frac{(a_n n/\log n)^{(1-s)/2}}{(a_n \log n/n)^{1/2}} = \left[\frac{(a_n n/\log n)^{(1-s)}}{(a_n \log n/n)}\right]^{1/2}.$$

Ignoring the root, we focus on

$$\frac{a_n^{(1-s)} n^{(1-s)}/(\log n)^{(1-s)}}{a_n \log n/n} = \frac{a_n^{(1-s)} n n^{(1-s)}}{a_n \log n (\log n)^{(1-s)}} = a_n^{-s}\left(\frac{n}{\log n}\right)^{2-s} = \left[\frac{a_n^{-s/(2-s)} n}{\log n}\right]^{2-s} = \left[\frac{a_n^{s/(s-2)} n}{\log n}\right]^{2-s}.$$

Since $s > 2$, and $\liminf_n \frac{a_n^{s/(s-2)} n}{\log n} > 0$ by assumption (2), the RHS is bounded above. Therefore, $C_n^{1-s} = O(\varepsilon_n)$, and consequently,

$$\|E m_n^{(o)}(\eta_k, Z_i)\| \le C_n^{1-s} E|M(Z_i)|^s = O((a_n \log n/n)^{1/2}) O(1) = O((a_n \log n/n)^{1/2}).$$

This implies that

$$\left\| n^{-1} \sum_i E m_n^{(o)}(\eta_k, Z_i) \right\| \le n^{-1} \sum_i \|E m_n^{(o)}(\eta_k, Z_i)\| = O((a_n \log n/n)^{1/2}).$$

Because the $C_n^{1-s} E|M(Z_i)|^s$ does not depend on k, taking maximum over k implies the second term in $(**)$ satisfies

$$\max_k \left\| n^{-1} \sum_i E m_n^{(o)}(\eta_j, Z_i) \right\| = O((a_n \log n/n)^{1/2}).$$

For the first term in $(**)$, we have

$$\begin{aligned}
\left\| n^{-1} \sum_i m_n^{(o)}(\eta_k, Z_i) \right\| &\le n^{-1} \sum_i \|m_n^{(o)}(\eta_k, Z_i)\| \\
&= n^{-1} \sum_i \|m_n(\eta_k, Z_i)\| \mathbb{1}\{|M(Z_i)| \ge C_n\} \\
&\le C_n^{1-s} n^{-1} \sum_i |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \ge C_n\}.
\end{aligned}$$

Since $C_n^{1-s} \asymp (a_n \log n/n)^{1/2}$, we need $n^{-1} \sum_i |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \ge C_n\} = O(1)$ almost surely. That is, for some $K > 0$,

$$P\left(\limsup_n n^{-1} \sum_i |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \ge C_n\}| > K \right) = 0.$$

Note that assumption (2) implies that $C_n \longrightarrow \infty$ by our previous calculations for the second term in $(**)$. Therefore, any fixed $C > 0$ and sufficiently large n,

$$n^{-1} \sum_{i=1}^n |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \ge C_n\} \le n^{-1} \sum_{i=1}^n |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \ge C\}.$$

Therefore,

$$\limsup_n \left| n^{-1} \sum_{i=1}^n |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \geq C_n\} \right| \leq \limsup_n \left| n^{-1} \sum_{i=1}^n |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \geq C\} \right|$$

And so, for given K, and any $C > 0$,

$$\begin{aligned} &P\left\{ \limsup_n \left| n^{-1} \sum_i |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \geq C_n\} \right| > K \right\} \\ &\quad \leq P\left\{ \limsup_n \left| n^{-1} \sum_i |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \geq C\} \right| > K \right\}. \end{aligned}$$

So we just need to show that for a fixed C, there is some K such that

$$P\left\{ \limsup_n \left| n^{-1} \sum_i |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \geq C\} \right| > K \right\} = 0.$$

But $E(|M(Z)|^s \mathbb{1}\{|M(Z)| \geq C\}) \leq E|M(Z)|^s < \infty$ by assumption (1), and, by the strong law of large numbers, we get

$$n^{-1} \sum_i |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \geq C\} \longrightarrow E(|M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \geq C\})$$

almost surely. Therefore, $n^{-1} \sum_i |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \geq C\} = O(1)$ almost surely, i.e. for some $K > 0$, $P\{\limsup_n |n^{-1} \sum_i |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \geq C\}| > K\} = 0$. Then, taking the maximum over j gives us

$$\begin{aligned} \max_j n^{-1} \left\| \sum_i m_n^{(o)}(\eta_j, Z_i) \right\| \leq & C_n^{1-s} n^{-1} \sum_i |M(Z_i)|^s \mathbb{1}\{|M(Z_i)| \geq C_n\} \\ = & O((a_n \log n/n)^{1/2}) O(1), \end{aligned}$$

since the RHS is independent of k. That is,

$$\max_j n^{-1} \left\| \sum_i m_n^{(o)}(\eta_j, Z_i) \right\| = O((a_n \log n/n)^{1/2}).$$

Second Term: For the second term in $(*)$, we appeal to the Lipschitz of m_n in η with integrable Lipschitz constant. Recall that $\max_{k=1,\ldots,N} \sup_{\eta \in B_r(\eta_k)} R_{n,k,2}$ is

$$\max_{k=1,\ldots,N} \sup_{\eta \in B_r(\eta_k)} \left\| n^{-1} \sum_i \left\{ [m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] - E[m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] \right\} \right\|.$$

By Lipschitz in η, we have that

$$\left\| n^{-1} \sum_i \left\{ [m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] - E[m_n(\eta, Z_i) - m_n(\eta_k, Z_i)] \right\} \right\|$$

$$\leq n^{-1}\sum_i \left\|m_n(\eta, Z_i) - m_n(\eta_k, Z_i)\right\| + n^{-1}\sum_i E\left\|m_n(\eta, Z_i) - m_n(\eta_k, Z_i)\right\|$$
$$\leq n^{-1}\sum_i |L(Z_i)| \times \|\eta - \eta_k\|^{c_1} n^{c_2} + n^{-1}\sum_i E|L(Z_i)| \times \|\eta - \eta_k\|^{c_1} n^{c_2}$$
$$\leq r^{c_1} n^{c_2} n^{-1}\sum_i |L(Z_i)| + r^{c_1} n^{c_2} E|L(Z)|.$$

And so

$$\max_{k=1,\ldots,N} \sup_{\eta \in B_r(\eta_k)} R_{n,k,2} \leq r^{c_1} n^{c_2}\left[n^{-1}\sum_i |L(Z_i)| + E|L(Z)|\right],$$

where the strong law ensures the sums converges almost surely, and by Assumption (3), the expectation is finite. So we just need to pick $r^{c_1} n^{c_2} = O(\varepsilon_n)$ almost surely.

We want the second term in $(*)$ to be $O(\varepsilon_n) = O((a_n \log n/n)^{1/2})$ in the end, so we can select r such that

$$r^{c_1} n^{c_2} \asymp \varepsilon_n \asymp (a_n \log n/n)^{1/2} = O((a_n \log n/n)^{1/2}),$$

in particular, we set $r = n^{-c_2/c_1}(a_n \log n/n)^{1/(2c_1)}$. We also need to ensure that our N balls of radius r covers of E. Suppose E has volume V. Then we need

$$V \leq N r^d = N n^{-dc_2/c_1}(a_n \log n/n)^{d/(2c_1)} \implies N(n) = n^{dc_2/c_1}(a_n \log n/n)^{-d/(2c_1)} V,$$

implying the number of balls N depends on n, $N = N(n)$. Putting together all the results, we conclude that

$$\sup_{\eta \in E}\left\|n^{-1}\sum_i [m_n(\eta, Z_i) - E m_n(\eta, Z_i)]\right\| \leq \max_{k=1,\ldots,N} R_{n,k,1} + \max_k \sup_{\eta \in B_r(\eta_k)} R_{n,k,2}$$
$$= O((a_n \log n/n)^{1/2}) \quad a.s.,$$

completing the proof. ■

The next Corollary shows the Lipschitz condition in Lemma 6.1 can be replaced by a component-wise Lipschitz condition.

Corollary 6.1. *Let $\eta = (\theta, \chi) \in E = \Theta \times \Omega_X$. Then Assumption (3) in Lemma 6.1 can be replaced with the following*

(3a) (Lipschitz for $\chi \in \Omega_X$) For all $\chi \in \Omega_X$,

$$\|m_n(\theta.\chi, Z) - m_n(\theta, \chi', Z)\| \leq \|\chi - \chi'\|^{c_1} n^{c_2} L_1(Z),$$

for some $c_1, c_2 > 0$, with $E|L_1(Z)| < \infty$;

(3b) (Lipschitz for $\theta \in \Theta$) For all $\theta \in \Theta$,

$$\|m_n(\theta, \chi, Z) - m_n(\theta', \chi, Z)\| \leq \|\theta - \theta'\|^{c_1'} n^{c_2'} L_2(Z),$$

for some $c_1', c_2' > 0$, with $E|L_2(Z)| < \infty$.

Proof. We just need to show that the two conditions are sufficient for bounding the second term in $(*)$. Recall that $\max_{k=1,\dots,N}\sup_{\eta\in B_r(\eta_k)} R_{n,k,2}$ is

$$\max_{k=1,\dots,N}\sup_{\eta\in B_r(\eta_k)}\left\|n^{-1}\sum_i\left\{[m_n(\eta,Z_i)-m_n(\eta_k,Z_i)]-E[m_n(\eta,Z_i)-m_n(\eta_k,Z_i)]\right\}\right\|,$$

where we have

$$\begin{aligned}&\left\|n^{-1}\sum_i\left\{[m_n(\eta,Z_i)-m_n(\eta_k,Z_i)]-E[m_n(\eta,Z_i)-m_n(\eta_k,Z_i)]\right\}\right\|\\ &\le n^{-1}\sum_i\left\|m_n(\eta,Z_i)-m_n(\eta_k,Z_i)\right\|+n^{-1}\sum_i E\left\|m_n(\eta,Z_i)-m_n(\eta_k,Z_i)\right\|.\end{aligned}$$

By Lipschitz in θ and χ, we get

$$\begin{aligned}&\|m_n(\eta,Z_i)-m_n(\eta_k,Z_i)\|\\ =&\|m_n(\theta,\chi,Z_i)-m_n(\theta_k,\chi_k,Z_i)\|\\ =&\|m_n(\theta,\chi,Z_i)\pm m_n(\theta_k,\chi,Z_i)\pm m_n(\theta,\chi_k,Z_i)\pm m_n(\theta_k,\chi_k,Z_i)-m_n(\theta_k,\chi_k,Z_i)\|\\ \le&\|m_n(\theta,\chi,Z_i)-m_n(\theta_k,\chi,Z_i)\|+\|m_n(\theta_k,\chi,Z_i)-m_n(\theta_k,\chi_k,Z_i)\|\\ &+\|m_n(\theta_k,\chi_k,Z_i)-m_n(\theta,\chi_k,Z_i)\|+\|m_n(\theta,\chi_k,Z_i)-m_n(\theta_k,\chi_k,Z_i)\|\\ \le&\|\theta-\theta_k\|^{c_1'}n^{c_2'}L_2(Z)+\|\chi-\chi_k\|^{c_1}n^{c_2}L_1(Z)+\|\theta-\theta_k\|^{c_1'}n^{c_2'}L_2(Z)+\|\chi-\chi_k\|^{c_1}n^{c_2}L_1(Z)\\ \le&2r^{c_1'}n^{c_2'}L_2(Z)+2r^{c_1}n^{c_2}L_1(Z),\end{aligned}$$

since $\|\eta-\eta_k\|\le r$ gives us $r^2\ge\|\eta-\eta_k\|^2=\|\theta-\theta_k\|^2+\|\chi-\chi_k\|^2$, implying $\|\chi-\chi_k\|\le r$ and $\|\theta-\theta_k\|\le r$. And so

$$\begin{aligned}\max_{k=1,\dots,N}\sup_{\eta\in B_r(\eta_k)}R_{n,k,2}\le&2r^{c_1'}n^{c_2'}\left[n^{-1}\sum_i|L_2(Z_i)|+E|L_2(Z)|\right]\\ &+2r^{c_1}n^{c_2}\left[n^{-1}\sum_i|L_1(Z_i)|+E|L_1(Z)|\right],\end{aligned}$$

where the strong law ensures the sums converges almost surely, and by assumption, the expectations are finite. So we need to pick r such that $r^{c_1}n^{c_2}=O(\varepsilon_n)$ and $r^{c_1'}n^{c_2'}=O(\varepsilon_n)$ almost surely. Without loss of generality, let $c_2\ge c_2'$ and let c_1 satisfy $r^{c_1}\ge r^{c_1'}$. Then we have

$$\max_{k=1,\dots,N}\sup_{\eta\in B_r(\eta_k)}R_{n,k,2}\le2r^{c_1}n^{c_2}\left[n^{-1}\sum_i|L_2(Z_i)|+E|L_2(Z)|+n^{-1}\sum_i|L_1(Z_i)|+E|L_1(Z)|\right],$$

where we set $r=n^{-c_2/c_1}(a_n\log n/n)^{1/(2c_1)}$ and $N\le n^{dc_2/c_1}(a_n\log n/n)^{-d/(2c_1)}V$, as we did originally in the Lemma, where V is the volume of E. This completes the proof of the Corollary. ■

The next corollary provides uniformity in one component of η, but pointwise in the other.

Corollary 6.2. *Let $\eta = (\theta, \chi) \in E = \Theta \times \Omega_X$ and replace Assumption (3) in Lemma 6.1 with Assumption (3b). Then*

$$\sup_{\theta \in \Theta} \left\| E_n m_n(\theta, \chi, Z_i) - E m_n(\theta, \chi, Z_i) \right\| = O_{a.s.}\left(\sqrt{\frac{a_n \log n}{n}} \right)$$

Proof. Letting $Z' = (\chi, Z)$ and applying Lemma 6.1 to the function $m_n(\theta, Z')$ gives us the result. ∎

6.2. **Convergence Rates for Sums of Bounded Outer Products.**

Lemma 6.2. *Suppose $X_1, ..., X_n$ are independent, random vectors in $\mathbb{R}^p$. Let $g : \mathbb{R}^p \to \mathbb{R}^{d_1 \times d_2}$ be a continuous function such that $\|g(X_i)\| \leq K$ almost surely, in operator norm, for all i. Then, we have that*

(a)

$$\left\| \frac{1}{n} \sum_i \{ g(X_i) g(X_i)^\top - \mu \} \right\| = O_p\left(n^{-1/2} \right).$$

(b)

$$\left\| \frac{1}{n} \sum_i \{ g(X_i) g(X_i)^\top - \mu \} \right\| = O_{a.s}\left(\sqrt{\frac{\log n}{n}} \right).$$

Proof. Let $Z_i = g(X_i) g(X_i)^\top - E\{ g(X) g(X)^\top \}$ so that Z_i is bounded by $2K^2$ almost surely, and is zero-mean with finite variance. For (a), fix $\varepsilon > 0$. To show bounded in probability, we want to find a constant $c > 0$ and $N > 0$ such that

$$P\left(\left\| \frac{1}{n} \sum_i Z_i \right\| \geq c \right) < \varepsilon, \ \ \forall n > N.$$

Then Matrix Chebyshev (Vershynin, 2018; Tropp, 2015) gives us

$$P\left(\left\| \frac{1}{n} \sum_i Z_i \right\| \geq c \right) < \frac{1}{n^2 c^2} E \left\| \sum_i Z_i \right\|^2 \leq \frac{1}{n c^2} E \| Z_1 \|^2.$$

So for fixed $\varepsilon > 0$, choosing $c = \{ E(\|Z_1\|^2) \}^{1/2} (n\varepsilon)^{-1/2}$ gives

$$P\left(\left\| \frac{1}{n} \sum_i Z_i \right\| \geq \left(\frac{E(\|Z_1\|^2)}{\varepsilon} \right)^{1/2} n^{-1/2} \right) < \varepsilon, \ \ \forall n,$$

which completes the proof.

For (b), we will appeal to Bernstein's Inequality for Matrices and Borel-Cantelli's First Lemma. For Bernstein's Inequality, we want to show that

$$P\left(\left\| \sum_i Z_i \right\| \geq c \sqrt{n \log n} \right) \to 0$$

where $Z_i = g(X_i)g(X_i)^\top - E\{g(X)g(X)^\top\}$, Z_i is bounded by $2K^2$ almost surely, and is zero-mean with finite variance. In particular,

$$\|Z_i\| \leq \|g(X_i)\|\|g(X_i)^\top\| + E(\|g(X)\|\|g(X)^\top\|) \leq 2K^2.$$

The total second moment is also finite since

$$\left\| \sum_{i=1}^n E(Z_i^2) \right\| \leq E\|g(X_i)g(X_i)^\top g(X_i)g(X_i)^\top\| \leq nK^4.$$

This implies that $\sigma^2 = \|\sum_{i=1}^n E(Z_i^2)\| \leq nK^4$, or $\sigma^2 = \|\sum_{i=1}^n E(Z_i^2)\| = nE(Z_1^2)$. A direct application of Matrix Bernstein (Vershynin, 2018; Tropp, 2015) gives

$$\begin{aligned} P\left(\left\| \sum_i Z_i \right\| \geq c\sqrt{n\log n} \right) \leq & 2(p+1)\exp\left\{ -\frac{c^2 n\log n/2}{n\sigma^2 + \frac{1}{3}K^2 c\sqrt{n\log n}} \right\} \\ = & 2(p+1)\exp\left\{ -\log n \frac{c^2/2}{K^4 + \frac{1}{3}K^2 c\sqrt{\frac{\log n}{n}}} \right\} \to 0, \quad n\to\infty, \end{aligned}$$

so we see that $\|\sum_i Z_i\| = O_p(\sqrt{n\log n})$, or $\|n^{-1}\sum_i Z_i\| = O_p(\sqrt{\log n/n})$. We need to use Borel-Cantelli to obtain the $\sqrt{n\log n}$ rate almost surely. For this, we note that

$$P\left(\left\| \sum_i Z_i \right\| \geq c\sqrt{n\log n} \right) \leq 2(p+1)\exp\left\{ -\log n \frac{c^2/2}{K^4 + \frac{1}{3}K^2 c\sqrt{\frac{\log n}{n}}} \right\} = 2(p+1)n^{-d_n},$$

where $d_0 = \frac{c^2}{2K^4 + \frac{2K^2 c}{3\sqrt{e}}} < d_n = \frac{c^2}{2K^4 + \frac{2K^2 c\sqrt{\log n}}{3\sqrt{n}}} \uparrow d = \frac{c^2}{2K^4}$ with $n \geq 1$. Then, summing over n, we get

$$\sum_{n=1}^\infty P\left(\left\| \sum_i Z_i \right\| \geq c\sqrt{n\log n} \right) \leq 2(p+1)\sum_{n=1}^\infty n^{-d_n} \leq 2(p+1)\sum_{n=1}^\infty n^{-d_0},$$

and so we just need to choose c such that the series converges. This can be achieved by choosing

$$2 < \frac{c^2}{2\sigma^2 + \frac{2Kc}{3\sqrt{e}}} \implies 0 < 3\sqrt{e}c^2 - 4Kc - 12\sqrt{e}\sigma^2 \implies c > \frac{2[K + \sqrt{K^2 + 9e\sigma^2}]}{3\sqrt{e}}.$$

Then Borel-Cantelli gives us $\|\sum_i Z_i\| = O_{a.s}(\sqrt{n\log n})$, so $\|n^{-1}\sum_i Z_i\| = O_{a.s}(\sqrt{\log n/n})$, completing the proof of (b).

■

6.3. **Bai, Miao, Rao Lemma.** This Lemma is from Bai et al. (1991)

Lemma 6.3.
Let $A = (a_{ik})$ and $B = (b_{ik})$ be two Hermitian (self-adjoint/symmetric) $p \times p$ matrices with spectral decomposition

$$A = \sum_{k=1}^{p} \delta_k u_k u_k^\top, \quad \delta_1 \geq \delta_2 \geq \cdots \geq \delta_p$$

$$B = \sum_{k=1}^{p} \lambda_k v_k v_k^\top, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$

where δ_k, λ_k and u_k, v_k are the respective eigenvalues corresponding orthogonal/orthonormal eigenvectors of A, B respectively. Let $\lambda_{n_{h-1}+1} = \cdots = \lambda_{n_h} = \tilde{\lambda}_h$ denote multiplicities of $\tilde{\lambda}_h$, for $h = 1, ..., s$. The s unique eigenvalues of B in descending order are $\tilde{\lambda}_1 > \tilde{\lambda}_2 > \cdots > \tilde{\lambda}_s$. Define $n_0 = 0 < n_1 < \cdots < n_s = p$ as a partition representing the increasing dimension of eigenspaces until the whole space is recovered.

If $|a_{ik} - b_{ik}| < c$ for all $i, k = 1, ..., p$, then there exists a constant M, independent of c, such that

(1) $|\delta_k - \lambda_k| < Mc, \quad k = 1, 2, ..., p$ (i.e. $\delta_k = \lambda_k + O(c)$), and

(2) letting $C^{(h)} = (C^{(h)}_{lk})$ with entries $|C^{(h)}_{lk}| \leq Mc$, for all $l, k = 1, .., p$, for all $h = 1, ..., s$, we have

$$\sum_{k=n_{h-1}+1}^{n_h} u_k u_k^\top = \sum_{k=n_{h-1}+1}^{n_h} v_k v_k^\top + C^{(h)}.$$

Remark 3.1 (from text): *Note that $\sum_{k=n_{h-1}+1}^{n_h} u_k u_k^\top$ and $\sum_{k=n_{h-1}+1}^{n_h} v_k v_k^\top$ are projection operators onto the subspaces spanned by the corresponding eigenvectors $\{u_i : i = n_{h-1}, ..., n_h\}$ and $\{v_i : i = n_{h-1}, ..., n_h\}$, so statement (b) is one about the closeness of the two eigenspaces of two matrices when these two matrices are close to each other (entry-wise).*

Remark: *For simplicity, let $O(c) = D$ refer to an $m \times n$ matrix $D = (d_{ik})$, where $|d_{ik}| \leq Mc$ for $i = 1, ..., m$ and $k = 1, .., n$. Then the lemma is saying that if $A = B + O(c)$ (i.e. the difference $A - B$ is bounded entry-wise to the order c), then the eigenvalues and eigenspaces of A, B are close to the same order c.*

Proof. For (1), we apply Von-Neumann's (Trace) Inequality (Lemma **??**), $|tr(AB)| \leq \sum_{i=1}^{p} \delta_i \lambda_i$, to the following

$$\begin{aligned}\sum_{i=1}^{p} (\delta_i - \lambda_i)^2 &= \sum_{i=1}^{p} \delta_i^2 + \sum_{i=1}^{p} \lambda_i^2 - 2\sum_{i=1}^{p} \delta_i \lambda_i \\ &= trAA^\top + trBB^\top - 2\sum_{i=1}^{p} \delta_i \lambda_i\end{aligned}$$

$$\begin{aligned}
&\leq tr(A^2 + B^2 - 2AB) \\
&= tr(A - B)(A - B) \\
&= \sum_{i,j}(a_{ij} - b_{ij}) = p^2c^2,
\end{aligned}$$

which implies $(\delta_i - \lambda_i)^2 < p^2c^2$, giving the first result with the constant being $M = p$.

For (2), we begin by representing the matrices A, B by their decomposition with respect to unique eigenvalues. From the first result (1), we have

$$A = \sum_{i=1}^{p} \delta_i u_i u_i^\top = \sum_{i=1}^{p} (\lambda_i + O(c)) u_i u_i^\top = \sum_{i=1}^{p} \lambda_i u_i u_i^\top + O(c) = \sum_{h=1}^{s} \tilde{\lambda}_h \sum_{i \in L_h} u_i u_i^\top + O(c),$$

where $L_h = \{n_{h-1} + 1, ..., n_h\}$ are the indices of eigenvectors for the h^{th} eigenspace of the matrix B. Then, from boundedness assumption $A = B + O(c)$, we get

$$\sum_{h=1}^{s} \tilde{\lambda}_h \sum_{i \in L_h} u_i u_i^\top + O(c) = A = B + O(c) = \sum_{h=1}^{s} \tilde{\lambda}_h \sum_{i \in L_h} v_i v_i^\top + O(c).$$

Therefore, $\sum_{h=1}^{s} \tilde{\lambda}_h \sum_{i \in L_h} u_i u_i^\top = \sum_{h=1}^{s} \tilde{\lambda}_h \sum_{i \in L_h} v_i v_i^\top + O(c)$. For simplicity of notation, let $P_h(A) \equiv \sum_{i \in L_h} u_i u_i^\top$ and $P_h(B) \equiv \sum_{i \in L_h} v_i v_i^\top$. Then we can write the result as $\sum_{h=1}^{s} \tilde{\lambda}_h P_h(A) = \sum_{h=1}^{s} \tilde{\lambda}_h P_h(B) + O(c)$. If $s = 1$, we have only one eigenspace with $h = s = 1$, and so

$$\sum_{h=1}^{s} \tilde{\lambda}_h P_h(A) = \sum_{h=1}^{s} \tilde{\lambda}_h P_h(B) + O(c) \Rightarrow \tilde{\lambda}_1 P_1(A) = \tilde{\lambda}_1 P_1(B) + O(c) \Rightarrow P_h(A) = P_h(B) + O(c),$$

so the second result (2) holds for $s = 1$.

Assume, for induction, that the result (2) is true for $s = t - 1$; that is, $P_h(A) = P_h(B) + O(c)$, for all $h = 1, ..., t - 1$. We will show that the result holds for $s = t$. By the induction assumption, we have

$$\begin{aligned}
&\sum_{h=1}^{t-1} \tilde{\lambda}_h P_h(A) = \sum_{h=1}^{t-1} \tilde{\lambda}_h P_h(B) + O(c) \\
\Rightarrow &\sum_{h=1}^{t-1} \tilde{\lambda}_h P_h(A) - \tilde{\lambda}_t \sum_{h=1}^{t-1} P_h(A) = \sum_{h=1}^{t-1} \tilde{\lambda}_h P_h(B) - \tilde{\lambda}_t \sum_{h=1}^{t-1} P_h(A) + O(c) \\
\Rightarrow &\sum_{h=1}^{t-1} (\tilde{\lambda}_h - \tilde{\lambda}_t) P_h(A) = \sum_{h=1}^{t-1} \tilde{\lambda}_h P_h(B) - \left(\tilde{\lambda}_t \sum_{h=1}^{t-1} P_h(B) + O(c) \right) + O(c) \\
\Rightarrow &\sum_{h=1}^{t-1} (\tilde{\lambda}_h - \tilde{\lambda}_t) P_h(A) = \sum_{h=1}^{t-1} (\tilde{\lambda}_h - \tilde{\lambda}_t) P_h(B) + O(c).
\end{aligned}$$

By applying the above to any $v \in P_t(B)$, which is orthogonal to $P_1(B), ..., P_{t-1}(B)$, we get

$$\sum_{h=1}^{t-1}(\tilde{\lambda}_h - \tilde{\lambda}_t)P_h(A)v = \sum_{h=1}^{t-1}(\tilde{\lambda}_h - \tilde{\lambda}_t)P_h(B)v + O(c) \Rightarrow \sum_{h=1}^{t-1}(\tilde{\lambda}_h - \tilde{\lambda}_t)P_h(A)v = O(c).$$

Because the unique eigenvalues are strictly decreasing, we have $(\tilde{\lambda}_h - \tilde{\lambda}_t) > 0$ for all $h = 1, .., s$. This implies $P_h(A)v = O(c)$ for all $v \in P_t(B)$, which means $\langle P_h(A), P_t(B)\rangle = O(c)$ for $h = 1, ..., t-1$.

Let U_1 be a basis for $\oplus_{h=1}^{t-1} P_h(A)$, U_2 a basis for $P_t(A)$, and denote $U = (U_1, U_2)$. Similarly, let V_1 be a basis for $\oplus_{h=1}^{t-1} P_h(B)$, V_2 a basis for $P_t(B)$, and denote $V = (V_1, V_2)$. Then $\langle P_h(A), P_t(B)\rangle = O(c)$ implies that $\langle U_1, V_2\rangle = O(c)$ and, by symmetry of the argument, $\langle V_1, U_2\rangle = O(c)$. Furthermore, since $V = (V_1, V_2)$ forms a basis, we can express U_2 in terms of the bases V_1, V_2 by $U_2 = V_1G_1 + V_2G_2$, $G_1 \in \mathbb{R}^{n_{t-1} \times (n_t - n_{t-1})}$ and $G_2 \in \mathbb{R}^{(n_t - n_{t-1}) \times (n_t - n_{t-1})}$.

By $U = (U_1, U_2)$ being an orthonormal basis, we get

$$I_p = UU^\top = (U_1, U_2)(U_1, U_2)^\top = U_1U_1^\top + U_2U_2^\top,$$

implying $U_2U_2^\top = I_p - U_1U_1^\top$. Then we have that

$$V_2^\top U_2U_2^\top V_2 = V_2^\top(I_p - U_1U_1^\top)V_2 = V_2^\top V_2 + O(c)O(c) = I_{p-n_{t-1}} + O(c^2).$$

By how we expressed U_2, we also get

$$O(c) = \langle V_1, U_2\rangle = \langle V_1, V_1G_1 + V_2G_2\rangle = I_{n_{t-1}}G_1 = G_1,$$

implying $G_1 = O(c)$, and so $U_2 = V_1O(c) + V_2G_2 + O(c) = V_2G_2 + O(c)$. Now note that

$$\begin{aligned}
G_2G_2^\top &= V_2^\top V_2G_2G_2^\top V_2^\top V_2 \\
&= V_2^\top(U_2 + O(c))(U_2^\top + O(c))V_2 \\
&= (V_2^\top U_2 + O(c))(U_2^\top V_2 + O(c)) \\
&= V_2^\top U_2U_2^\top V_2 + O(c) + O(c^2) \\
&= V_2^\top U_2U_2^\top V_2 + O(c) \\
&= I_{p-n_{t-1}} + O(c^2) + O(c) \\
\Rightarrow G_2G_2^\top &= I_{p-n_{t-1}} + O(c).
\end{aligned}$$

This gives us

$$\begin{aligned}
P_t(A) = U_2U_2^\top &= V_2G_2G_2^\top V_2^\top + O(c) + O(c^2) \\
&= V_2(I_{p-n_{t-1}} + O(c))V_2^\top + O(c) \\
&= V_2V_2^\top + O(c) \\
\Rightarrow P_t(A) &= P_t(B) + O(c),
\end{aligned}$$

showing the t^{th} eigenspaces are close. By the induction assumption, we can conclude that the hypothesis holds for $s = t$; that is, $P_h(A) = P_h(B) + O(c)$ for $h = 1, .., t$. Therefore, by induction, we have shown that result (2) holds, completeing the proof. ■

References

Adragni, K. P. (2018). Minimum average deviance estimation for sufficient dimension reduction. Journal of Statistical Computation and Simulation, 88(3), 411–431.

Agresti, A. (2010). Analysis of ordinal categorical data (Vol. 656). John Wiley & Sons.

Agresti, A. (2013). Categorical data analysis (3rd ed.). Wiley.

Bai, D. Z., Miao, Q. B., & Rao, R. C. (1991). Estimation of directions of arrival of signals: Asymptotic results. In S. Haykin (Ed.), Advances in spectrum analysis and array processing (Vol. II, p. 327-347). Englewood Cliffs, NJ.: Prentice Hall.

Fan, J., Heckman, N. E., & Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. Journal of the American Statistical Association, 90(429), 141–150.

Li, B. (2018). Sufficient dimension reduction: Methods and applications with r. CRC Press.

McCullagh, P. (1980). Regression models for ordinal data. Journal of the royal statistical society. Series B (Methodological), 109–142.

McCullagh, P., & Nelder, J. (1989). Generalized linear models, second edition. Taylor & Francis.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. arXiv preprint arXiv:1501.01571.

Vershynin, R. (2018). High-dimensional probability: An introduction with applications in data science (Vol. 47). Cambridge university press.