SUPPLEMENT TO "A SELF-PENALIZING OBJECTIVE FUNCTION FOR SCALABLE INTERACTION DETECTION"

By Keli Liu[†] and Feng Ruan[†,*]

University of California, Berkeley[*]

APPENDIX A: A ROADMAP TO THE APPENDIX

The organization of the Appendix is as follows.

1. Section B lists the common notation and assumptions in the Appendix.
2. Section C–Section F prepares the basic materials.
   - Section C starts from the very basics—the definition of the signal and noise variables.
   - Section D studies how to choose f in the metric learning objective so it can perform non-parametric variable selection. It turns out that the sufficient and necessary condition is that f' is strictly completely monotone.
   - Section E shows that the masking phenomenon always exists for the metric learning objective no matter how carefully you choose the function f and no matter how well you optimize the (non-convex) population metric learning objective. The justification is through a detailed characterization of the population landscape for two concrete statistical models. The masking phenomenon implies that naive maximization of metric learning objective leads to inconsistent estimate of the signal variables.
   - Section F shows how the ultimate metric learning algorithm leverages the iterative reweighing technique to resolve the inconsistency of the naive maximization of the metric learning objective; the final output of the ultimate metric learning algorithm includes the signals and excludes the noise variables.
3. Section G–Section J constitutes the main mathematical tools for the theoretical study of the metric learning algorithms.
   - Section G shows that, on population, the gradient of the objective with respect to noise variable is always non-positive, and its magnitude is lower bounded by the value of the objective function

itself. The implication is that, even without explicit ℓ_1 regularization, the metric learning objective (assuming having at least one signal variable) can *self-penalize* the noise variables.

- Section H studies the general population metric learning objective where the kernel in the objective can be of ℓ_1 and ℓ_2 type.
- Section I studies a special class of population objective where the kernel is of ℓ_1 type. Compared with those in Section H, the result and proof in Section I rely on a special property of the ℓ_1 type kernel, that is, its Fourier transform is heavy-tailed (Cauchy density function). It is important to notice that the same property does not hold for the ℓ_2 type kernel, whose Fourier transform is light-tailed (Gaussian density). The result in Section I has substantial consequence on the recovery guarantees of the signal variables in high dimension (Section N).
- Section J studies the uniform convergence of the empirical objective and its gradient to the population ones. An important consequence of these results is that all properties derived on population (Section F—Section I) continue to hold on the empirical counterparts after taking into account the sampling errors.

4. Section K–Section N presents the proof of the main results of the paper. All these results only assume the optimization procedure reach a stationary point of the (non-convex) metric learning objective.
   - Section K–L shows that in low dimension, even without explicit regularization, the metric learning algorithm, using either ℓ_1 or ℓ_2 kernel, recovers the signal and removes the the noise variables.
   - Section M shows that in high dimension, with explicit ℓ_1 regularization, the metric learning algorithm throw away noise variables.
   - Section N shows that in high dimension, the metric learning algorithm that uses ℓ_1 type kernel manages to recover the signal variables under three different nonparametric models (with different sample complexities), namely, the main effect, pure interaction and hierarchical interaction model. The proof uses heavily the special property of the ℓ_1 type kernel studied in Section I.
5. Section O contains the supporting lemma.

APPENDIX B: COMMON NOTATION AND ASSUMPTIONS

B.1. Notation. In addition to the notation in the main text, the following notation is used throughout the appendix.

We frequently use the shorthand $\mathbf{d}_i = |X_i - X_i'|^q$ and $\mathbf{d} = (\mathbf{d}_1, \mathbf{d}_2, \dots \mathbf{d}_p)$. We use ℓ to denote the objective function with explicit ℓ_1 penalization:

$$\ell(\beta; \mathbb{Q}) = F(\beta; \mathbb{Q}) - \lambda \|\beta\|_1 ,$$

and use $\mathcal{B} = \{\beta \in \mathbb{R}_+^p : \|\beta\|_1 \leq b\}$ to denote the ℓ_1-constraint set.

Let $w_A(x, y) \equiv 1 - \mathbb{P}(Y = y \mid X_A = x_A)$ be the weight function associated with any subset $A \subseteq [p]$. We introduce the notation $\mathbb{Q}^A$ to denote the unique probability distribution such that $dQ^A(x, y) \propto dP(x, y) \cdot w_A(x, y)$, i.e.,

$$\frac{dQ^A}{dP}(x, y) = \frac{w_A(x, y)}{\int w_A(x, y) dP}.$$

B.2. A List of Common Assumptions. This section lists the common assumptions that are used throughout the appendix. These are indeed the assumptions (A1)-(A3) and (B1)-(B2) in the main text.

(A1) The function $f \in \mathcal{C}^\infty(\mathbb{R}_+)$ satisfies f' is strictly completely monotone. Moreover $q = 1$ or $q = 2$.
(A2) For some constant $M > 0$, $\|X\|_\infty \leq M$ almost surely under $\mathbb{P}$.
(A3) Imperfect classification: for some constant $\varrho > 0$,

$$\mathbb{E}\left[\mathbb{P}\left(Y = 1 | X_S\right) | Y = 0\right] > \varrho, \quad \mathbb{E}\left[\mathbb{P}\left(Y = 0 | X_S\right) | Y = 1\right] > \varrho.$$

(B1) For some constant $\zeta > 0$, $\mathbb{E}[|X_j - X_j'|] \geq \zeta$ for $j \notin S$. Here the expectation is taken under $\mathbb{P}$.
(B2) In the case where $q = 2$, f' has an analytical extension on the complex plane $\mathbb{C}$ such that $|f'(z)| \leq A(1 + |z|)^N e^{B|\mathrm{Re}(z)|}$ for some $A, B, N < \infty$.

We refer the readers to Section 4 of the main text for the explanations of the assumptions.

APPENDIX C: BASICS ON MODELS

This section discusses the very basics of the modeling. Proposition 1 shows that the definition of signal and noise variables (see Section 1.3 in the main text) are well-posed.

PROPOSITION 1. *There exists a unique subset $S \subseteq [p]$ with the following three properties: (i) $Y \mid X \sim Y \mid X_S$, (ii) $X_S \perp X_{S^c}$ and (iii) there is no strict subset $A \subsetneq S$ which satisfies (i) and (ii).*

PROOF. First we prove existence. Start with $S = \{1, \dots, p\}$ and note that it trivially satisfies (i) and (ii). If no strict subset of $\{1, \dots, p\}$ satisfies (i)

and (ii), then S satisfies (iii) also and we are done. Otherwise if a strict subset $A \subsetneq S$ satisfies (i) and (ii), set S equal to A. Repeat this process until we arrive at a set S for which there is no strict subset that satisfies (i) and (ii). This process terminates in at most p steps and the S returned by the process satisfies (i), (ii), (iii).

Next, we prove uniqueness. Suppose there exist subsets A, B satisfying (i), (ii) and (iii). By (i), we have $Y|X_A = Y|X_B$ and therefore for all $t \in \mathbb{R}$,

$$\mathbb{E}[e^{itY}|X_A] = \mathbb{E}[e^{itY}|X_B]. \tag{1}$$

Taking the conditional expectation w.r.t X_A on both sides yields

$$\mathbb{E}[e^{itY}|X_A] = \mathbb{E}[\mathbb{E}[e^{itY}|X_B]|X_A] = \mathbb{E}[\mathbb{E}[e^{itY}|X_B]|X_{A\cap B}] = \mathbb{E}[e^{itY} \mid X_{A\cap B}]$$

where the second equality comes from the fact that $X_{A\setminus B} \perp X_B$ since B satisfies (ii) and the third equality comes from the tower property of conditional expectation. Thus, we have shown for all $t \in \mathbb{R}$,

$$\mathbb{E}[e^{itY}|X_A] = \mathbb{E}[e^{itY}|X_{A\cap B}].$$

This means $Y|X_{A\cap B} = Y|X_A = Y|X$. Moreover, we have

$$P(X) = P(X_B)P(X_{B^c}) = P(X_{A\cap B})P(X_{B\setminus A})P(X_{B^c})$$

where the first equality is from $X_B \perp X_{B^c}$ and the second equality is from $X_A \perp X_{A^c}$. Thus $X_{A\cap B} \perp X_{(A\cap B)^c}$. Hence, we have shown $A \cap B$ is a subset that satisfies (i) and (ii). Since A, B satisfy (iii), it implies $A = A \cap B = B$. This proves the uniqueness. □

APPENDIX D: PROPERTIES OF F NEEDED TO DETECT INTERACTIONS

This section presents the proof of Theorem 1 in the main text, showing that a necessary and sufficient condition for f to detect all possible interactions is that f' needs to be strictly completely monotone.

D.1. Proof of Theorem 1. The proof is divided into two parts.

Proof of Part 1 of Theorem 1. Consider the XOR signal of order s:

$$Y = \operatorname{sign}(X_1 X_2 \cdots X_s) \quad \text{where } X_j \text{ i.i.d } \mathbb{Q}\left(X_j = \pm\frac{1}{2}\right) = \frac{1}{2}$$

As f satisfies the Axiom 1 and 2, it means that for the above XOR signal, we have for all $\beta \in \mathbb{R}^s_+$ of full support (i.e., $\operatorname{supp}(\beta) = [s]$) and $c \in \mathbb{R}$,

$$\mathbb{E}_{B-W}\left[f(c + \left\|X - X'\right\|^q_{q,\beta})\right] > 0.$$

Now, a simple evaluation of the expectation gives that for all β of full support

$$\begin{aligned}&\mathbb{E}_{B-W}\left[f(c+\left\|X-X'\right\|_{q,\beta}^{q})\right]\\&=\frac{1}{2^{s-1}}\sum_{u\in\{0,1\}^s}(-1)^{\mathrm{par}(u)-1}\cdot f\Big(c+\sum_{j=1}^{s}\beta_j u_j\Big)>0\end{aligned}\tag{2}$$

where $\mathrm{par}(u)$ is defined as the parity of number of 1s in u, i.e.,

$$\mathrm{par}(u)=\begin{cases}1 & \text{if } \#\{i:u_i=1\} \text{ is odd}\\ 0 & \text{if } \#\{i:u_i=1\} \text{ is even}\end{cases}.$$

The mid-term of equation (2) is the s-th order finite difference of the function f, a quantity that measures the moduli of local smoothness of the function f [4]. Lemma D.1 connects the s-th order finite difference to the s-th order derivative when f is smooth. The proof of Lemma D.1 is deferred to Section D.2.

LEMMA D.1. *Let $f:\mathbb{R}\to\mathbb{R}$ be a smooth function that has continuous differentiable derivatives up to order s. Let $\{a_{j,1}\}_{j\in[s]}$ and $\{a_{j,0}\}_{j\in[s]}$ be two sequences where $a_{j,i}\in\mathbb{R}$ for $j\in[s]$ and $i\in\{0,1\}$. Consider*

$$\overline{F}(x;s):=\sum_{u\in\{0,1\}^s}(-1)^{\mathrm{par}(u)-1}\cdot f\Big(x+\sum_{j=1}^{s}a_{j,u_j}\Big).$$

Then we have the integral representation for $\overline{F}(x;s)$:

$$\overline{F}(x;s)=(-1)^{s-1}\prod_{j=1}^{s}(a_{j,1}-a_{j,0})\int_0^1\int_0^1\dots\int_0^1 f^{(s)}\Big(x+\sum_{j=1}^{s}a_j(t_j)\Big)dt_1dt_2\dots dt_j$$

where the function $a_j:[0,1]\to\mathbb{R}$ is defined by

$$a_j(t)=ta_{j,1}+(1-t)a_{j,0}.$$

Now we apply Lemma D.1 to the sequence $a_{j,1}=\beta_j$, $a_{j,0}=0$. From equation (2), we get for all β of full support (i.e. $\beta_j>0$ for $j\in[s]$)

$$\begin{aligned}&\mathbb{E}_{B-W}\left[f(c+\left\|X-X'\right\|_{q,\beta}^{q})\right]\\&=(-1)^{s-1}\cdot\Big(\prod_{j\in[s]}\beta_j\Big)\cdot\int_0^1\int_0^1\dots\int_0^1 f^{(s)}\Big(c+\sum_{j=1}^{s}\beta_j t_j\Big)dt_1dt_2\dots dt_s>0.\end{aligned}$$

Starting from below we set $\beta_j = \varepsilon$ for some $\varepsilon > 0$. As a result, we obtain

$$(-1)^{s-1} \int_0^1 \int_0^1 \dots \int_0^1 f^{(s)}\Big(c + \varepsilon \sum_{j=1}^{s} t_j\Big) dt_1 dt_2 \dots dt_s > 0. \tag{3}$$

We wish to mention that inequality (3) holds for all $\varepsilon > 0$ and $s \in \mathbb{N}$.

Now we prove that for all $s \in \mathbb{N}$, the following holds:

$$(-1)^{(s-1)} f^{(s)}(c) \geq 0 \quad \text{for all } c \geq 0. \tag{4}$$

Indeed, suppose $f^{(s)}(c) \neq 0$ for some $c \geq 0$. By continuity of $x \mapsto f^{(s)}(x)$ (since $f \in \mathcal{C}^\infty(\mathbb{R}_+)$ by assumption), we can choose $\varepsilon > 0$ sufficiently small such that $f^{(s)}(c + \varepsilon \sum_{j=1}^{s} t_j)$ has the same sign as $f^{(s)}(c)$ for all $t_j \in [0, 1]$. Fix this $\varepsilon > 0$. Now we can see from equation (3) that this immediately implies that $(-1)^{(s-1)} f^{(s)}(c) > 0$. This proves the desired equation (4).

Next, we show the inequality in equation (4) must be strict. Suppose on the contrary $f^{(s)}(c) = 0$ for some $c \geq 0$ and $s \in \mathbb{N}$. We divide our discussion into two cases.

1. In the first case, we assume s is odd. Now equation (4) implies $f^{(s)}$ is nonnegative and non-increasing on $\mathbb{R}_+$. Hence $f^{(s)}(c) = 0$ implies $f^{(s)}(x) = 0$ for all $x \geq c$. This contradicts equation (3).
2. In the second case, we assume s is even. Now equation (4) implies $f^{(s)}$ is nonpositive and non-decreasing on $\mathbb{R}_+$. Hence $f^{(s)}(c) = 0$ implies $f^{(s)}(x) = 0$ for $x \geq c$. This contradicts equation (3).

As a summary, we have shown the inequality in equation (4) must be strict. This proves that f' has to be strictly completely monotone.

Proof of Part 2 of Theorem 1. Let $f \in \mathcal{C}^\infty(\mathbb{R}_+)$ be such that f' is strictly completely monotone. It suffices to prove for $q \in \{1, 2\}$:

$$\mathbb{E}_{B-W}[f(\|X - X'\|_q^q)] \geq 0$$

for all distributions on (X, Y) with equality if and only if $X \perp Y$. We introduce the following definition of conditionally positive (negative) (semi)-definite kernels.

Definition D.1. *A continuous function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is said to be conditionally positive semi-definite if*

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j K(x_i, x_j) \geq 0. \tag{5}$$

for all $N \in \mathbb{N}$, all pairwise distinct values $x_1, x_2, \ldots, x_N \in \mathbb{R}^d$ and all $\alpha \in \mathbb{R}^N$ satisfying $\sum_{i=1}^N \alpha_i = 0$. The function f is said to be conditionally positive definite if Eq (5) is strict unless α is zero. Conditionally negative semi-definite and conditionally negative definite functions are defined analogously.

A fundamental result due to Székely and Rizzo [14] shows that

$$\mathbb{E}_{B-W}\left[K\left(X, X'\right)\right] \geq 0$$

whenever K is conditionally negative definite and moreover the equality is strict when X is not independent of Y. Hence it suffices to prove that $K(X, X') = f(\|X - X'\|_q^q)$ is conditionally negative definite when $q \in \{1, 2\}$.

- For the case $q = 2$, Schoenberg [12] shows that when f' is strictly completely monotone, then the function $K(X, X') = f(\|X - X'\|_2^2)$ is conditionally negative definite (see also [10]).
- For the case $q = 1$, Bernstein's theorem for completely monotone functions [11] shows that any function $f \in \mathbb{C}^\infty(\mathbb{R}_+)$ whose derivative f' is strictly completely monotone admits the following Lévy–Khintchine representation

  $$f(x) = a + bx + \int_0^\infty (1 - e^{-tx})\mu(dt).$$

  where $a \in \mathbb{R}$, $b \geq 0$ and μ is a non-negative measure on $\mathbb{R}_+$ satisfying (i) $\mu(0, \infty) > 0$ and (ii) $\int_0^\infty \min\{1, x\}\, \mu(dx) < \infty$. Therefore,

  $$f(\left\|x - x'\right\|_1) = a + b\left\|x - x'\right\|_1 + \int_0^\infty \left(1 - e^{-t \cdot \|x - x'\|_1}\right) d\mu(t) \tag{6}$$

  Note that both $(x, x') \mapsto \|x - x'\|_1$ and $(x, x') \mapsto -e^{-\|x - x'\|_1}$ are conditionally negative definite (see [12]). This shows that $K(x, x') = f(\|x - x'\|_1)$, as an weighted average of conditionally negative definite kernels (note the total weights are positive since $\mu((0, \infty)) > 0$ is strict), must be also conditionally negative definite.

D.2. Proof of Lemma D.1. The proof is by induction. Consider $s = 1$.

$$\begin{aligned}\overline{F}(x; 1) &= f(x + a_1(1)) - f(x + a_1(0)) \\ &= \int_0^1 f^{(1)}(x + a_1(t_1))dt_1 \cdot (a_{1,1} - a_{1,0}).\end{aligned}$$

Assume the integral representation holds for $s \leq k-1$. Consider $s = k$. Use the notation $f_x(\cdot)$ to represent the function $f(x+\cdot)$. For $s = k$, we compute

(7)

$$\begin{aligned}
\overline{F}(x;k) &= \sum_{u\in\{0,1\}^k} (-1)^{\mathrm{par}(u)-1} \cdot f_x\Big(\sum_{j=1}^{k} a_{j,u_j}\Big) \\
&= \sum_{u\in\{0,1\}^k, u_k=1} (-1)^{\mathrm{par}(u)-1} \cdot f_x\Big(\sum_{j=1}^{k-1} a_{j,u_j} + a_{k,1}\Big) \\
&\quad + \sum_{u\in\{0,1\}^k, u_k=0} (-1)^{\mathrm{par}(u)-1} \cdot f_x\Big(\sum_{j=1}^{k-1} a_{j,u_j} + a_{k,0}\Big) \\
&= \sum_{u\in\{0,1\}^{k-1}} (-1)^{\mathrm{par}(u)} \cdot f_x\Big(\sum_{j=1}^{k-1} a_{j,u_j} + a_{k,1}\Big) \\
&\quad + \sum_{u\in\{0,1\}^{k-1}} (-1)^{\mathrm{par}(u)-1} \cdot f_x\Big(\sum_{j=1}^{k-1} a_{j,u_j} + a_{k,0}\Big) \\
&= (-1)\cdot \sum_{u\in\{0,1\}^{k-1}} (-1)^{\mathrm{par}(u)-1} \cdot \left(f_x\Big(\sum_{j=1}^{k-1} a_{j,u_j} + a_{k,1}\Big) - f_x\Big(\sum_{j=1}^{k-1} a_{j,u_j} + a_{k,0}\Big)\right) \\
&= (-1)\cdot \sum_{u\in\{0,1\}^{k-1}} (-1)^{\mathrm{par}(u)-1} \cdot \int_0^1 f_x^{(1)}\Big(\sum_{j=1}^{k-1} a_{j,u_j} + a_k(t_k)\Big) dt_k \cdot (a_{k,1} - a_{k,0}) \\
&= (-1)\cdot \int_0^1 \sum_{u\in\{0,1\}^{k-1}} (-1)^{\mathrm{par}(u)-1} \cdot f_x^{(1)}\Big(\sum_{j=1}^{k-1} a_{j,u_j} + a_k(t_k)\Big) dt_k \cdot (a_{k,1} - a_{k,0})
\end{aligned}$$

By induction hypothesis for $s = k-1$, we have the representation

$$\begin{aligned}
&\sum_{u\in\{0,1\}^{k-1}} (-1)^{\mathrm{par}(u)-1} f_x^{(1)}\Big(\sum_{j=1}^{k-1} a_{j,u_j} + a_k(t_k)\Big) \\
&= (-1)^{k-1} \cdot \int_0^1 \int_0^1 \cdots \int_0^1 f_x^{(k)}\Big(\sum_{j=1}^{k-1} a_j(t_j)\Big) dt_1 dt_2 \ldots dt_{k-1} \cdot \prod_{j=1}^{k-1} (a_{j,1} - a_{j,0}).
\end{aligned}$$

Now substitute it into the last line of equation (7). We obtain

$$\overline{F}(x;k) = (-1)^k \cdot \int_0^1 \int_0^1 \cdots \int_0^1 f_x^{(k)}\Big(\sum_{j=1}^{k-1} a_j(t_j)\Big) dt_1 dt_2 \ldots dt_k \cdot \prod_{j=1}^{k} (a_{j,1} - a_{j,0}),$$

This gives the desired integral representation.

APPENDIX E: THE MASKING PHENOMENON: LANDSCAPE ANALYSIS OF OBJECTIVE FUNCTION

This section studies the masking phenomenon of the metric learning objective $F(\beta;\mathbb{Q})$. Proposition 2 and Proposition 2 show that it is possible to construct some distribution $\mathbb{Q}$ such that the support of any stationary point of $F(\beta;\mathbb{Q})$ is a strict subset of the signal set S. This implies that naive maximization of the objective $F(\beta;\mathbb{Q})$ can't recover the signal set S.

The organization of the section is as follows. Section E.1 presents the proof of Proposition 2. Section E.2 presents Proposition 2 (which complements the example in Section 2.2 of the main text). Section E.3 presents the proof of Proposition 2.

E.1. Proof of Proposition 2. We consider the model

$$\begin{aligned}
&\mathbb{Q}(Y=1)=\mathbb{Q}(Y=0)=\frac{1}{2}, \quad X_1 \perp X_2 \mid Y\\
&\mathbb{Q}\big(X_j=\pm\frac{1}{2} \mid Y=1\big)=\frac{1}{2}(1\pm\delta_j), \quad j=1,2\\
&\mathbb{Q}\big(X_j=\pm\frac{1}{2} \mid Y=0\big)=\frac{1}{2}(1\mp\delta_j), \quad j=1,2
\end{aligned}$$

where the parameters $\delta_1,\delta_2\in(0,1)$ are to be determined. Denote

$$\mathcal{D}_r=\{\beta\in\mathbb{R}^2_+ : \max\{\beta_1,\beta_2\}\le r\}.$$

Below we show we can carefully choose the parameters $1>\delta_1>\delta_2>0$ such that for all large enough r, $\beta=(r,0)$ is the unique stationary point w.r.t $\mathcal{D}_r$.

We first show for any $\delta_1>\delta_2$ and any $r>0$, any stationary point $\beta\in\mathcal{D}_r$ must satisfy $\beta_1=r$. To see this, it suffices to show the objective is strictly increasing w.r.t β_1. Indeed, we prove for all $\beta\in\mathcal{D}_r$,

$$\frac{\partial}{\partial\beta_1}F(\beta;\mathbb{Q})>0 \tag{8}$$

Below we prove equation (8) holds for all $\beta\in\mathcal{D}_r$. Indeed, by definition,

$$\begin{aligned}
\frac{\partial}{\partial\beta_1}F(\beta;\mathbb{Q}) &= \mathbb{E}_{B-W}\left[|X_1-X_1'|f'(\beta_1|X_1-X_1'|+\beta_2|X_2-X_2'|)\right]\\
&= \frac{1}{2}(\delta_1^2-\delta_2^2)(f'(\beta_1)-f'(\beta_1+\beta_2))+\delta_1^2 f'(\beta_1+\beta_2)
\end{aligned} \tag{9}$$

Now, by assumption $\delta_1>\delta_2>0$, and $f'(\beta_1)\ge f'(\beta_1+\beta_2)>0$ since f' is strictly completely monotone. As a result, equation (9) implies the desired

equation (8). As mentioned before, this implies for all $\delta_1 > \delta_2 > 0$, all $r > 0$, any stationary point $\beta \in \mathcal{D}_r$ must satisfy $\beta_1 = r$.

Next, we show for any $\delta_1 > \delta_2$, $\beta^* = (r, 0)$ is the unique global maximum in $\mathcal{D}_r$ for all large enough r. Since $F(\beta; \mathbb{Q})$ is always strictly increasing w.r.t β_1, it suffices to prove for all large enough r, the inequality below

$$F(\beta^*; \mathbb{Q}) \geq F(\beta; \mathbb{Q}) \tag{10}$$

holds for any β of the form $\beta = (r, c)$ where $c \in [0, r]$ and additionally with equality if and only if $c = 0$. We evaluate $F(\beta; \mathbb{Q})$ at $\beta = (r, c)$

$$\begin{aligned} &F(\beta; \mathbb{Q}) \mid_{\beta=(r,c)} \\ &= \mathbb{E}_{B-W}\left[f(r|X_1 - X_1'| + c|X_2 - X_2'|)\right] \\ &= \frac{1}{2}\Big[(\delta_1^2 + \delta_2^2)f(0) + (\delta_1^2 + \delta_2^2)f(r + c) + (\delta_1^2 - \delta_2^2)f(r) + (\delta_2^2 - \delta_1^2)f(c)\Big]. \end{aligned}$$

Substitute the above expression into (10). It suffices to prove for all large enough r, the inequality below holds for all $c \in [0, r]$

$$(\delta_1^2 - \delta_2^2)(f(c) - f(0)) \geq (\delta_1^2 + \delta_2^2)(f(r + c) - f(r)) \tag{11}$$

with equality if and only if $c = 0$.

Below we prove this result. By assumption $f' \in \mathcal{C}^\infty(\mathbb{R}_+)$ is strictly completely monotone satisfying $f'(\infty) = 0$. By Lemma H.1, we obtain that

$$f(x) = a - \int_0^\infty e^{-tx}\mu(dt).$$

where $a \in \mathbb{R}$ and μ is a non-negative finite measure on $[0, \infty)$ that satisfies $|\mu| = \mu(\mathbb{R}_+) > 0$. As a result, we have that

$$\begin{aligned} f(r + c) - f(r) &= \int_0^\infty e^{-tr}\left(1 - e^{-tc}\right)\mu(dt) \\ f(c) - f(0) &= \int_0^\infty \left(1 - e^{-tc}\right)\mu(dt). \end{aligned} \tag{12}$$

Now that $t \mapsto e^{-tr}$ is decreasing and $t \mapsto 1 - e^{-tc}$ is increasing. By covariance inequality (Lemma O.6), we obtain (recall $|\mu| = \mu([0, \infty)) > 0$)

$$\int_0^\infty e^{-tr}\left(1 - e^{-tc}\right)\mu(dt) \leq \frac{1}{|\mu|}\int_0^\infty e^{-tr}\mu(dt) \cdot \int_0^\infty (1 - e^{-tc})\mu(dt)$$

which, in view of the identity (12), is equivalent to

$$f(r + c) - f(r) \leq \frac{1}{|\mu|}\int_0^\infty e^{-tr}\mu(dt) \cdot (f(c) - f(0)). \tag{13}$$

Note that equation (13) holds for all $r, c > 0$. Now, by dominated convergence theorem, $\int_0^\infty e^{-tr}\mu(dt) \to 0$ as $r \to \infty$. Hence, equation (13) implies that the desired inequality (11) must hold for all large enough r, and with equality if and only if $c = 0$. This proves that for any fixed $\delta_1 > \delta_2$, $\beta^* = (r, 0)$ is the unique global maximum in $\mathcal{D}_r$ for all large enough r.

Finally, we prove that, under the assumption that f' is strictly completely monotone, and that f' is not slowly varying at ∞, we can carefully choose $\delta_1 > \delta_2 > 0$ such that $\beta^* = (r, 0)$ is the unique stationary point in $\mathcal{D}_r$ for all large enough r. Indeed, let $\delta_1 > \delta_2 > 0$. By the first part, we know there exists some $r_0 \equiv r_0(\delta_1, \delta_2) < \infty$ such that for all $r > r_0$, any stationary point $\beta \in \mathcal{D}_r$ must satisfy $\beta = (r, c)$ for some $c \in [0, r]$. Now, our strategy is to carefully design $\delta_1 > \delta_2 > 0$ such that for all large enough r, any point $\beta = (r, c)$ where $c > 0$ can't be stationary. The high level idea is to prove that under the appropriate choice of the parameters δ_1, δ_2, we have for all large enough r, the objective is strictly decreasing w.r.t β_2, i.e., for all $c \in (0, r]$

$$\frac{\partial}{\partial \beta_2} F(\beta; \mathbb{Q}) \mid_{\beta=(r,c)} < 0. \tag{14}$$

Below we show we can achieve this goal.

To see this, we start by evaluating the gradient w.r.t β_2 at $\beta = (r, c)$

$$\frac{\partial}{\partial \beta_2} F(\beta; \mathbb{Q}) \mid_{\beta=(r,c)} = \frac{1}{2}\left[(\delta_1^2 + \delta_2^2) f'(r + c) - (\delta_1^2 - \delta_2^2) f'(c)\right].$$

By simple algebraic manipulation, we obtain the bound

$$\frac{\partial}{\partial \beta_2} F(\beta; \mathbb{Q}) \mid_{\beta=(r,c)} \leq \frac{1}{2}(\delta_1^2 + \delta_2^2) \cdot \left(\sup_{c \in [0,r]} \frac{f'(r + c)}{f'(c)} - \frac{\delta_1^2 - \delta_2^2}{\delta_1^2 + \delta_2^2} \right) \tag{15}$$

Here comes the key. We show at the end of the section that any completely monotone function f' with $f'(\infty) = 0$ must satisfy

$$\sup_{c \in [0,r]} \frac{f'(r + c)}{f'(c)} = \frac{f'(2r)}{f'(r)}. \tag{16}$$

Since f' is not slowly varying, we can always pick $\alpha < 1$ and $1 > \delta_1 > \delta_2 > 0$ such that

$$\limsup_{r \to \infty} \frac{f'(2r)}{f'(r)} < \alpha < \frac{\delta_1^2 - \delta_2^2}{\delta_1^2 + \delta_2^2}. \tag{17}$$

Pick any $\delta_1, \delta_2, \alpha$ that satisfies equation (17). By equation (15) and equation (16), we conclude that for the δ_1, δ_2 we pick, equation (14) holds for all

$c \in (0, r]$ when r is large enough. As mentioned earlier, this proves that for the δ_1, δ_2 we pick, the only stationary point in $\mathcal{D}_r$ is $\beta^* = (r, 0)$ when r is sufficiently large.

To complete the proof, here we show the deferred claim at equation (16). It suffices to prove the mapping $c \mapsto g(c) := \frac{f'(r+c)}{f'(c)}$ is monotonically increasing. Since $f' \in \mathcal{C}^\infty(\mathbb{R}_+)$ is strictly completely monotone with $f'(\infty) = 0$, Lemma H.1 implies the following representation of f':

$$f'(x) = \int_0^\infty te^{-tx}\mu(dt). \tag{18}$$

where μ is a non-negative finite measure on $[0, \infty)$. Let $0 \le c_1 < c_2 \le r$. Below we show $g(c_1) \le g(c_2)$, or equivalently,

$$\log f'(r + c_1) + \log f'(c_2) \le \log f'(r + c_2) + \log f'(c_1). \tag{19}$$

To see this, $x \to f'(x)$ is log-convex since $x \to te^{-tx}$ is log-convex [3]. Now equation (19) is merely a consequence of the majorization inequality [8].

E.2. Proposition 2. Proposition 2 makes precise the qualitative statements described in the example in Section 2.2.

We start by recalling the setup in the example. Consider the following additive main effect model:

$$\begin{aligned}
&\mathbb{Q}(Y = 1) = \mathbb{Q}(Y = 0) = \frac{1}{2}, \quad X_1 \perp X_2 \perp \ldots \perp X_s | Y \\
&\mathbb{Q}(X_j = \pm\frac{1}{2} \mid Y = 1) = \frac{1}{2}(1 \pm \delta_j) \quad \text{for } j \in [s] \\
&\mathbb{Q}(X_j = \pm\frac{1}{2} \mid Y = 0) = \frac{1}{2}(1 \mp \delta_j) \quad \text{for } j \in [s].
\end{aligned}$$

In above $\delta_j > 0$ is the signal size for the feature X_j. For convenience, we reparametrize the signal size as $\rho_j = (1 + \delta_j^2)/(1 - \delta_j^2)$. We assume $\rho_1 > \rho_2 \ldots > \rho_s > 0$, or equivalently $\delta_1 > \delta_2 > \ldots > \delta_s > 0$. Fix $f(x) = -e^{-x}$ and $q > 0$. Proposition 2 below gives a full description on all the stationary points of the objective

$$F(\beta; \mathbb{Q}) = \mathbb{E}_{B-W}\left[F(\left\|X - X'\right\|_{q,\beta}^q)\right]$$

with respect to the box-constraint set $\mathcal{D}_r = \{\beta : 0 \le \beta_j \le r\}$ when r is large. To conveniently state the result, we introduce the following definition.

Definition E.1 (Classification of Stationary Points). *We make the below definition on stationary points β of $F(\beta; \mathbb{Q})$ w.r.t the constraint $\mathcal{D}_r$.*

- *A stationary point* β *is called* regular *if* $\max_{i\in\mathrm{supp}(\beta)}\beta_i < r$.
- *A stationary point* β *is called* irregular *if* $\max_{i\in\mathrm{supp}(\beta)}\beta_i = r$.
- *A stationary point* β *is called* abnormal *if it is* irregular *with* $\beta_{a_k} = r > \max_{i\neq k}\beta_{a_i}$. *Here* $\mathrm{supp}(\beta) = \{a_1,\dots,a_k\}$ *where* $a_1 < \dots < a_k$.

We briefly explain the motivation for the definition. Proposition 2 shows that a stationary point of $F(\beta;\mathbb{Q})$ w.r.t the constraint $\mathcal{D}_r$ can be either regular or abnormal, and that existence of an *abnormal* stationary point is due to the boundary effect caused by the box constraint of the optimization (an abnormal stationary point β has the largest coordinate β_{a_k} for the weakest signal X_{a_k}).

We also introduce the notation $\bar\rho_B = (\prod_{j\in B}\rho_j)^{\frac{1}{|B|}}$ for any subset $B\subseteq[s]$. Hence $\bar\rho_B$ is the geometric average of the signal size ρ_j over $j\in B$.

PROPOSITION 2. *For all large enough* r[1], *the stationary points of* $F(\beta;\mathbb{Q})$ *with respect to* $\mathcal{D}_r = \{\beta : 0\le\beta_j\le r\}$ *has the below characterization:*

1. *There exists one and only one* 1*-sparse stationary point at* $(r,0,\dots,0)$, *which is also the (unique) global maximum.*
2. *There does not exist any* 2*-sparse stationary point.*
3. *There may exist* k*-sparse stationary points for any* $k\ge 3$. *Pick any set* $A = \{a_1,\dots,a_k\}\subseteq[s]$ *with* $|A| = k\ge 3$ *and* $a_1 > \dots > a_k$.
   (a) *There is a* regular *stationary point* β *with* $\mathrm{supp}(\beta) = A$ *only if*

$$\min_{j\in A}\rho_j^2 \ge (\bar\rho_A)^{\frac{k}{k-1}} \ge \max_{j\in[s]}\rho_j. \tag{20}$$

   Conversely, if the inequality (20) *strictly holds, then there is a* regular *stationary point* β *with* $\mathrm{supp}(\beta) = A$.
   (b) *There is an* irregular *stationary point* β *with* $\mathrm{supp}(\beta) = A$ *only if*

$$\min_{j\in A\setminus\{a_k\}}\rho_j^2 > \big(\bar\rho_{A\setminus\{a_k\}}\big)^{\frac{k-1}{k-2}}\cdot\rho_{a_k}^{-\frac{1}{k-2}} > \max_{j\in[s]\setminus\{a_k\}}\rho_j. \tag{21}$$

   Conversely, if the inequality (21) *strictly holds, then there is an* irregular *stationary point* β *with* $\mathrm{supp}(\beta) = A$, *and moreover that stationary point* β *must also be* abnormal.

E.3. Proof of Proposition 2. We start with Lemma E.1, which gives the expression of the gradient of the objective $F(\beta;\mathbb{Q})$ under the above model. The proof of Lemma E.1 is deferred into Section E.3.4.

[1]There exists $r_0 < \infty$ such that the statement of Proposition 2 holds for all $r\ge r_0$.

LEMMA E.1. *For any $j \in [s]$, we have the expression of the gradient:*

$$\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) = \omega_j e^{-\beta_j} \cdot \prod_{k \neq j} \left((1-\omega_k) + \omega_k e^{-\beta_k} \right) - (1-\omega_j) e^{-\beta_j} \cdot \prod_{k \neq j} \left(\omega_k + (1-\omega_k) e^{-\beta_k} \right).$$

In above, $\omega_j = \frac{1}{2}(1+\delta_j^2)$ for all $j \in [s]$.

Since the statement of Proposition 2 contains three separate parts, we present the proof of them in the three subsections below.

E.3.1. *Proof of Part 1 of Proposition 2.* The organization of the proof can be decomposed into the following three steps.

- First we show that $\beta = r\mathbf{e}_1 = (r, 0, \ldots, 0)$ is the unique global maximum for large enough r.
- Next we show that $\beta = r\mathbf{e}_1$ is the only stationary point of the form $\beta = c\mathbf{e}_1$ when r is large enough.
- Finally, we show no 1-sparse stationary point can take the form of $c\mathbf{e}_j$ for $j \neq 1$. This holds for all $r > 0$.

First, we start by proving that $\beta = r\mathbf{e}_1$ is the unique global maximum of $F(\beta; \mathbb{Q})$ over the constraint set $\mathcal{D}_r$ when r is large enough. Let us first compute the objective function (recall $\omega_j = \frac{1}{2}(1+\delta_j^2)$)

$$\begin{aligned} F(\beta; \mathbb{Q}) &= -\prod_j \mathbb{E}_B \left[e^{-\beta_j |X_j - X_j'|} \right] + \prod_j \mathbb{E}_W \left[e^{-\beta_j |X_j - X_j'|} \right] \\ &= \prod_j \left(\omega_j + (1-\omega_j) e^{-\beta_j} \right) - \prod_j \left(1 - \omega_j + \omega_j e^{-\beta_j} \right). \end{aligned} \tag{22}$$

Writing $\beta^* = r\mathbf{e}_1$, we obtain $F(\beta^*; \mathbb{Q}) = (2\omega_1 - 1)(1 - e^{-r})$. Below, we prove that for large enough r, the inequality below holds for all $\beta \in \mathcal{D}_r$

$$F(\beta^*; \mathbb{Q}) \geq F(\beta; \mathbb{Q}) \tag{23}$$

with an equality if and only if $\beta = \beta^*$. To see this, we find it convenient to introduce an algebraic transformation $x_j = e^{-\beta_j} \in [e^{-r}, 1]$. Recall the expression in equation (22). Define the function $x \mapsto G(x)$ by

$$G(x) = \prod_j \left(\omega_j + (1-\omega_j) x_j \right) - \prod_j \left(1 - \omega_j + \omega_j x_j \right).$$

Hence, $G(x) = F(\beta; \mathbb{Q})$ when x is the transformation of β ($x_j = e^{-\beta_j}$). It suffices to prove that for the point $x^* = e^{-r}\mathbf{e}_1$ (which is the transformation of β^*), we have the inequality below holds for all x such that $x_j \in [e^{-r}, 1]$

$$G(x^*) = (2\omega_1 - 1)(1 - e^{-r}) \geq G(x) \tag{24}$$

and this inequality becomes an equality if and only if $x = x^*$.

Below we prove this. Although the proof is elementary, it is technically non-trivial. The first step is to show that the maximum of $G(x)$ over $\mathcal{E}_r$ must be attained at some $x \in \text{ext}(\mathcal{E}_r)$ where $\text{ext}(\mathcal{E}_r) = \{x : x_j \in \{e^{-r}, 1\}\}$. To see this, we use the method of adjustment[13]. Indeed, for any $x \notin \text{ext}(\mathcal{E}_r)$, we can always pick a variable $x_j \in (e^{-r}, 1)$. Since $G(x)$ is linear in x_j given all other variables x_{-j}, we can always increase $G(x)$ by fixing x_{-j} and adjusting x_j to one of the boundary values e^{-r} and 1. The principle of method of adjustment then says the maximum of $G(x)$ must be attained at some $x \in \text{ext}(\mathcal{E}_r)$. As a result of this fact, it suffices to prove for all large enough r, the inequality below holds for all $x \in \text{ext}(\mathcal{E}_r)$

$$G(x^*) = (2\omega_1 - 1)(1 - e^{-r}) \geq G(x). \tag{25}$$

In addition, we need to show the point $x \in \text{ext}(\mathcal{E}_r)$ that attains the maximum $\max_{x \in \text{ext}(\mathcal{E}_r)} G(x)$ is unique, and that point is $x = x^*$.

Checking inequality (25) holds for large r is the second step. Here we note that given any x, G (as a function of ω) is monotonically increasing w.r.t ω_j. Now we introduce the definition of $x \mapsto \bar{G}(x)$, where we replace all the ω_j ($j \geq 2$) in the definition of G by ω_2:

$$\begin{aligned}\bar{G}(x) = (\omega_1 + (1-\omega_1)x_1) \cdot \prod_{j>1} (\omega_2 + (1-\omega_2)x_j) \\ - ((1-\omega_1) + \omega_1 x_1) \cdot \prod_{j>1} ((1-\omega_2) + \omega_2 x_j)\end{aligned}$$

One can show that $G(x) \leq \bar{G}(x)$ using the assumption $\omega_2 \geq \omega_j$ for all $j \geq 2$. Additionally, $G(x^*) = \bar{G}(x^*)$. As a result, it suffices to prove for large enough r, the inequality $\bar{G}(x^*) \geq \bar{G}(x)$ holds for all $x \in \mathcal{E}_r$. Now, for any $x \in \mathcal{E}_r$, if we denote $k = |\{j \geq 2 : x_j = e^{-r}\}|$, then we have that $\bar{G}(x) = H(k)$ where

$$\begin{aligned}H(k) = \left(\omega_1 + (1-\omega_1)e^{-r}\right)\left(\omega_2 + (1-\omega_2)e^{-r}\right)^k \\ - \left((1-\omega_1) + \omega_1 e^{-r}\right) \cdot \left((1-\omega_2) + \omega_2 e^{-r}\right)^k.\end{aligned}$$

In particular $\bar{G}(x^*) = H(0)$. Thus, it amounts fo check for the mapping $k \mapsto H(k)$ where $k \in \mathbb{N}$, $H(k)$ attains its unique maximum at $k = 0$ when

r is sufficiently large. With some legwork, we can easily show this by first proving that $H(1) < H(0)$ for all large enough r, and then showing that $k \mapsto H(k)$ is decreasing as long as $H(1) < H(0)$ and $\omega_1 + (1-\omega_1)e^{-r} < 1$. Consequently, we can show easily that the mapping $k \mapsto H(k)$ has its unique global maximum attained at $k = 0$ when r is large enough. From our previous (lengthy) discussions, we know that this implies that $\beta = r\mathbf{e}_1$ is the unique global maximum of $F(\beta; \mathbb{Q})$ over $\mathcal{D}_r$ for all sufficiently large r.

Next, we prove that $\beta = r\mathbf{e}_1$ is the unique stationary point of the form $\beta = c\mathbf{e}_1$ for large r. To do so, our strategy is to prove that $\frac{\partial}{\partial \beta_1} F(\beta; \mathbb{Q}) > 0$ at $\beta = c\mathbf{e}_1$ for all $c > 0$. To see this, Lemma E.1 implies for any $c > 0$

$$\frac{\partial}{\partial \beta_1} F(\beta; \mathbb{Q}) \mid_{\beta = c\mathbf{e}_1} = (2\omega_1 - 1)e^{-c} > 0.$$

Hence $\beta = c\mathbf{e}_1$ can be stationary only if $\beta = r\mathbf{e}_1$.

Finally, we prove that $\beta = c\mathbf{e}_j$ can't be stationary for $c > 0$ and $j \neq 1$. We show that $\frac{\partial}{\partial \beta_1} F(\beta; \mathbb{Q}) \mid_{\beta = c\mathbf{e}_j} > 0$ and thereby such a point can't be stationary. Once again using Lemma E.1, and noticing that $\omega_1 > \omega_j > \frac{1}{2}$, we obtain

$$\frac{\partial}{\partial \beta_1} F(\beta; \mathbb{Q}) \mid_{\beta = c\mathbf{e}_j} = \omega_1 - \omega_j + (\omega_1 + \omega_j - 1)e^{-c} > 0,$$

Thus $\beta = c\mathbf{e}_j$ can't be stationary for all $c > 0$ and $j \neq 1$.

E.3.2. *Proof of Part 2 of Proposition 2.* The proof is by contradiction. Let's assume $\beta = c_1\mathbf{e}_{j_1} + c_2\mathbf{e}_{j_2}$ for $j_1 < j_2$ is a stationary point. As before, using Lemma E.1, and using the assumption that $\omega_{j_1} > \omega_{j_2} > \frac{1}{2}$, we obtain

$$\frac{\partial}{\partial \beta_{j_1}} F(\beta; \mathbb{Q}) = e^{-c_1} \cdot \left((\omega_{j_1} - \omega_{j_2}) + (\omega_{j_1} + \omega_{j_2} - 1) \cdot e^{-c_2} \right) > 0.$$

Thus, if $\beta = c_1\mathbf{e}_{j_1} + c_2\mathbf{e}_{j_2}$ is a stationary point, it must be $c_1 = r$.

Now, using Lemma E.1 again, we obtain at $\beta = r\mathbf{e}_{j_1} + c_2\mathbf{e}_{j_2}$

$$\frac{\partial}{\partial \beta_{j_2}} F(\beta; \mathbb{Q}) = e^{-c_2} \cdot \left((\omega_{j_2} - \omega_{j_1}) + (\omega_{j_1} + \omega_{j_2} - 1) \cdot e^{-r} \right)$$

Note then this is negative for large r since $\omega_{j_1} > \omega_{j_2}$ by assumption. Thus, for large enough r, any stationary point $\beta = c_1\mathbf{e}_{j_1} + c_2\mathbf{e}_{j_2}$ with $c_1 = r$ must have $c_2 = 0$, and thus can't be 2-sparse.

This proves that no 2-sparse stationary points exists for large enough r.

E.3.3. *Proof of Part 3 of Proposition 2.* Let $A = \{a_1, a_2, \ldots, a_k\}$ with $a_1 < a_2 < \ldots < a_k$. Assume β is a stationary point with $\text{supp}(\beta) = A$. We decompose $A = A_1 \cup A_2$ where $A_1 = \{i \in A : \beta_i < r\}$ and $A_2 = \{i \in A : \beta_i = r\}$. Write $\beta = \sum_{l=1}^{k} \beta_l \mathbf{e}_{a_l}$. Note that β is stationary if and only if

$$\begin{aligned}
\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) &= 0 \quad \text{for } j \in A_1, \\
\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) &\geq 0 \quad \text{for } j \in A_2, \\
\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) &\leq 0 \quad \text{for } j \in A^c.
\end{aligned}$$

Using Lemma E.1, we obtain equivalent expressions of above

$$\rho_j = \prod_{l \neq j, l \in A} \rho_l^* \quad \text{for } j \in A_1 \tag{26}$$

$$\rho_j \geq \prod_{l \neq j, l \in A} \rho_l^* \quad \text{for } j \in A_2 \tag{27}$$

$$\rho_j \leq \prod_{l \in A} \rho_l^* \quad \text{for } j \in A^c \tag{28}$$

where we use the notation

$$\begin{aligned}
\rho_j &= \frac{\omega_j}{1 - \omega_j}, \quad \omega_j = \frac{1}{2}(1 + \delta_j^2) \\
\rho_j^* &= \frac{\omega_j^*}{1 - \omega_j^*}, \quad \omega_j^* = \omega_j(1 - \phi_j) + (1 - \omega_j)\phi_j, \ \phi_j = \frac{1}{1 + e^{\beta_j}}
\end{aligned}$$

The constraint $\beta_j \in [0, r]$ is equivalent to $\frac{\omega_j + e^{-r}}{1 + e^{-r}} \geq \omega_j^* \geq \frac{1}{2}$, or equivalently

$$1 < \rho_j^* \leq \rho_j + e^{-r}(1 + \rho_j) \tag{29}$$

Now we have the system of equations (26)–(28) w.r.t the unknown variables ρ_j^* along with the constraint (29). Each solution of the system of equations corresponds to one stationary point. We note that the solution of the system is a tuple $(\{\rho_j^*\}_{j \in A}, A_1, A_2)$. Sometimes, we write $(\{\rho_j^*\}_{j \in A}, A_1, A_2) = (\{\rho_j^*(r)\}_{j \in A}, A_1(r), A_2(r))$ to emphasize its dependence on r.

In the first part of the proof, we prove the following results.

1. The set $A_2(r)$ must be either $A_2(r) = \emptyset$ or $A_2(r) = \{a_k\}$ for large enough r.

2. There exists some $r_0 < \infty$ such that the following holds: assuming there exists a solution $(\{\rho_j^*(r)\}_{j\in A}, A_1(r), A_2(r))$ for the system of equations (26)–(28) with $A_2(r) = \emptyset$ for some $r > r_0$, then it is necessary that the following condition holds:

$$\min_{j\in A} \rho_j^2 \geq (\bar{\rho}_A)^{\frac{k}{k-1}} \geq \max_{j\in[s]} \rho_j. \tag{30}$$

   By our definition, the stationary point $\beta = \beta(r)$ corresponding to such a solution $(\{\rho_j^*(r)\}_{j\in A}, A_1(r), A_2(r))$ is a *regular* stationary point.
3. There exists some $r_0 < \infty$ such that the following holds: assuming there exists a solution $(\{\rho_j^*(r)\}_{j\in A}, A_1(r), A_2(r))$ for the system of equations (26)–(28) with $A_2(r) = \{a_k\}$ for some $r > r_0$, then it is necessary that the following condition holds:

$$\min_{j\in A\setminus\{a_k\}} \rho_j^2 \geq \left(\bar{\rho}_{A\setminus\{a_k\}}\right)^{\frac{k-1}{k-2}} \cdot \rho_{a_k}^{-\frac{1}{k-2}} \geq \max_{j\in[s]\setminus\{a_k\}} \rho_j. \tag{31}$$

   By our definition, the stationary point $\beta = \beta(r)$ corresponding to such a solution $(\{\rho_j^*(r)\}_{j\in A}, A_1(r), A_2(r))$ is an *irregular* stationary point.

Below we prove the above results. Let $\tilde{A}_1$ and $\tilde{A}_2$ be two sets such that $\tilde{A}_1 = A_1(r)$ and $\tilde{A}_2 = A_2(r)$ hold for some sequence $r = r_n$ where r_n tends to infinity. Fix this sequence. Let $\tilde{\rho}_j$ be (one of) the accumulation point of the sequence $\{\rho_j^*(r_n)\}_{n\in\mathbb{N}}$ (note the existence of the accumulation point is guaranteed since $\{\rho_j^*(r_n)\}_{n\in\mathbb{N}}$ is uniformly bounded). Since the system of equations (26), (26), (28) hold for $(\{\rho_j^*(r)\}_{j\in A}, A_1(r), A_2(r))$ for $r = r_n$, it also holds for the limit $(\{\tilde{\rho}_j\}_{j\in A}, \tilde{A}_1, \tilde{A}_2)$, i.e.,

$$\rho_j = \prod_{l\neq j, l\in A} \tilde{\rho}_l \quad \text{for } j \in \tilde{A}_1 \tag{32}$$

$$\rho_j \geq \prod_{l\neq j, l\in A} \tilde{\rho}_l \quad \text{for } j \in \tilde{A}_2 \tag{33}$$

$$\rho_j \leq \prod_{l\in A} \tilde{\rho}_l \quad \text{for } j \in A^c \tag{34}$$

where the constraint becomes (compare it with equation (29))

$$\begin{cases} \tilde{\rho}_j = \rho_j & \text{for } j \in \tilde{A}_2 \\ 1 \leq \tilde{\rho}_j \leq \rho_j & \text{for } j \notin \tilde{A}_2 \end{cases} \tag{35}$$

According to equations (32) and (33) and the constraint equation (35), we know that $\rho_j \geq \tilde{\rho}_l$ for any $j, l \in A$ and $l \neq j$. In particular, this means that

$\min_{j\neq l}\rho_j \geq \tilde{\rho}_l$ for $l \in A$. As $\rho_l = \tilde{\rho}_l$ for $l \in \tilde{A}_2$, this implies that $\tilde{A}_2$ is either empty or a singleton, and further $\tilde{A}_2 = \{a_k\}$ when it is a singleton. Now we divide our discussion into two cases:

1. In the case where $\tilde{A}_2 = \emptyset$, one can solve that
$$\tilde{\rho}_j = \rho_j^{-1} \cdot (\bar{\rho}_A)^{\frac{k}{k-1}}.$$
Checking the constraint gives the necessary condition (30).
2. In the second case where $\tilde{A}_2 = \{a_k\}$, one can solve for $j \in A$, $j \neq a_k$
$$\tilde{\rho}_j = \rho_j^{-1} \cdot \rho_{a_k}^{-\frac{1}{k-2}} \cdot (\bar{\rho}_{A\setminus\{a_k\}})^{\frac{k-1}{k-2}}.$$
Checking the constraint gives the necessary condition (31).

In summary, this proves the aforementioned claim.

Next, we argue the following claim.

1. Suppose the following strengthening of condition (30) holds
$$\min_{j\in A}\rho_j^2 > (\bar{\rho}_A)^{\frac{k}{k-1}} > \max_{j\in[s]}\rho_j. \tag{36}$$
then the system of equations (26)–(28) has a solution with $A_2 = \emptyset$. This holds for all $r > 0$.
2. Suppose the following strengthening of condition (31) holds
$$\min_{j\in A\setminus\{a_k\}}\rho_j^2 > \left(\bar{\rho}_{A\setminus\{a_k\}}\right)^{\frac{k-1}{k-2}} \cdot \rho_{a_k}^{-\frac{1}{k-2}} > \max_{j\in[s]\setminus\{a_k\}}\rho_j. \tag{37}$$
then the system of equations (26)–(28) has a solution with $A_2 = \{a_k\}$. This holds for all sufficiently large r.

The above claim is easy to prove.

1. Suppose condition (36) holds. Then we can take $A_1 = a$, $A_2 = \emptyset$, and set for $j \in A$
$$\rho_j^* = \rho_j^{-1} \cdot (\bar{\rho}_A)^{\frac{k}{k-1}}.$$
It is easy to check that the above ρ_j^* satisfy all the constraints.
2. Suppose condition (37) holds. Then we can take $A_1 = A\setminus\{a_k\}$, $A_2 = \{a_k\}$, and set $\rho_{a_k}^* = \rho_{a_k} + e^{-r}(1+\rho_j)$ and for $j \in A_1$,
$$\rho_j^* = \rho_j^{-1} \cdot (\rho_{a_k}^*)^{-\frac{1}{k-2}} \cdot (\bar{\rho}_{A\setminus\{a_k\}})^{\frac{k-1}{k-2}}$$
It is easy to verify that the above ρ_j^* thus defined satisfy all the constraints for all sufficiently large r.

The proof is thus complete.

E.3.4. *Proof of Lemma E.1.* The proof is simply computation. First, we have by definition

$$\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) = \mathbb{E}_{B-W}\left[|X_j - X_j'| e^{-\sum_k \beta_k |X_k - X_k'|}\right]. \tag{38}$$

Now we evaluate the RHS. Consider the case $\mathbb{E}_B[\cdot] = \mathbb{E}[\cdot | Y \neq Y']$. By conditional independence of X_k given Y, we have

$$\begin{aligned}
&\mathbb{E}_B\left[|X_j - X_j'| e^{-\sum_k \beta_k |X_k - X_k'|}\right] \\
&= \mathbb{E}_B\left[\mathbb{E}\left[|X_j - X_j'| e^{-\beta_j |X_j - X_j'|} \mid Y, Y'\right] \cdot \mathbb{E}\left[e^{-\sum_{k\neq j} \beta_k |X_k - X_k'|} \mid Y, Y'\right]\right]
\end{aligned}$$

Note the conditional expectation given Y, Y' are the same for $(Y, Y') = (0, 1)$ and $(Y, Y') = (1, 0)$. Thus, we have on $Y \neq Y'$

$$\begin{aligned}
\mathbb{E}\left[|X_j - X_j'| e^{-\beta_j |X_j - X_j'|} | Y, Y'\right] &= \omega_j e^{-\beta_j}. \\
\mathbb{E}\left[e^{-\beta_k |X_k - X_k'|} | Y, Y'\right] &= (1 - \omega_k) + \omega_k e^{-\beta_k}.
\end{aligned}$$

Substituting these expressions back into our formula yields

$$\mathbb{E}_B\left[|X_j - X_j'| e^{-\sum_k \beta_k |X_k - X_k'|}\right] = \omega_j e^{-\beta_j} \cdot \prod_{k\neq j} \left((1 - \omega_k) + \omega_k \cdot e^{-\beta_k}\right).$$

Similarly, we can solve the case $\mathbb{E}_W[\cdot] = \mathbb{E}[\cdot | Y = Y']$:

$$\mathbb{E}_W\left[|X_j - X_j'| e^{-\sum_k \beta_k |X_k - X_k'|}\right] = (1 - \omega_j) e^{-\beta_j} \cdot \prod_{k\neq j} \left(\omega_k + (1 - \omega_k) e^{-\beta_k}\right).$$

Substituting the expressions into equation (38) yields the desired result.

APPENDIX F: PROPERTIES OF ALGORITHMS ON POPULATION

F.1. Proof of Proposition 1. The proof is divided into two parts. For notational simplicity, we denote $B = \text{supp}(\beta)$.

1. Theorem 1 says $F(\beta) > 0$ only if X_B is not independent of Y.
2. Let A be any strict subset $A \subsetneq B$ such that $X_A \perp X_{B\setminus A}$. We prove that $Y|X_B \not\sim Y|X_A$. Suppose on the contrary that $Y|X_B \sim Y|X_A$. This says that if we denote $\tilde{X} = X_B$, then we have $Y|\tilde{X} \sim Y|\tilde{X}_A$ and $\tilde{X}_A \perp \tilde{X}_{A^c}$. Here comes to the key of the proof. Since $F(\beta) > 0$ by assumption, Proposition 3 implies that for any variable $j \in B\setminus A$:

$$\frac{\partial}{\partial \beta_j} F(\beta) < 0. \tag{39}$$

This means that we can strictly increase the objective by decreasing β_j for any variable $j \in B \backslash A$, which contradicts the assumption that β is a local maximum! Thus we must have $Y|X_B \not\sim Y|X_A$.

F.2. Proof of Proposition 3. The proof is based on routine calculation. Let P and Q denote the density functions of the distributions $\mathbb{P}$ and $\mathbb{Q}$ w.r.t some base measure μ (say $\mu = \frac{1}{2}(\mathbb{P} + \mathbb{Q})$). We start from the definition for the rebalancing distribution $\mathbb{Q}$:

$$Q(x_{\hat{S}}, x_{\hat{S}^c}, y) = \frac{1}{Z} \cdot P(x_{\hat{S}}, x_{\hat{S}^c}, y) \cdot P(1 - y \mid x_{\hat{S}}) \tag{40}$$

where Z is the normalization factor. We now prove the two claims (a)-(b) of Proposition 3 in the below paragraphs.

(a) Integratation over $x_{\hat{S}^c}$ on both sides of equation (40) gives

$$Q(x_{\hat{S}}, y) = \frac{1}{Z} \cdot P(x_{\hat{S}}, y) \cdot P(1 - y \mid x_{\hat{S}}) \tag{41}$$

$$= \frac{1}{Z} \cdot P(x_{\hat{S}}) \cdot P(y \mid x_{\hat{S}}) \cdot P(1 - y \mid x_{\hat{S}}). \tag{42}$$

Note y takes value only from $\{0, 1\}$. Thus, equation (42) shows the joint distribution $Q(x_{\hat{S}}, y)$ is in fact independent of the value $y \in \{0, 1\}$. This proves that $X_{\hat{S}} \perp Y$ under $\mathbb{Q}$ and moreover that $\mathbb{Q}(Y = \pm 1) = \frac{1}{2}$.

(b) We use equation (40) and (41) to obtain

$$Q(x_{\hat{S}^c} \mid x_{\hat{S}}, y) = \frac{Q(x_{\hat{S}}, x_{\hat{S}^c}, y)}{Q(x_{\hat{S}}, y)} = \frac{P(x_{\hat{S}}, x_{\hat{S}^c}, y)}{P(x_{\hat{S}}, y)} = P(x_{\hat{S}^c} \mid x_{\hat{S}}, y).$$

This proves that the conditional distribution of $X_{\hat{S}^c}|X_{\hat{S}}, Y$ is the same under $\mathbb{P}$ and $\mathbb{Q}$ as desired.

(c) We can W.L.O.G assume $\hat{S} \subseteq S$ since the weight satisfies

$$P(1 - y \mid x_{\hat{S}}) = P(1 - y \mid x_{\hat{S} \cap S}).$$

As $Y|X_S = Y|X$ and $X_S \perp X_{S^c}$ under $\mathbb{P}$, we have $X_{S^c} \perp (Y, X_S)$ under P. In particular, the density P factorizes into products: $P(y, x) = P(y, x_S)P(x_{S^c})$. Consequently, if we substitute it into equation (40), we obtain the expression

$$Q(x_{\hat{S}}, x_{\hat{S}^c}, y) = \frac{1}{Z} \cdot \big(P(x_S, y) \cdot P(1 - y \mid x_{\hat{S}})\big) \cdot P(x_{S^c})$$

which shows that the density Q factorizes into products of functions of (X_S, Y) and of X_{S^c}. This implies that $X_{S^c} \perp (Y, X_S)$ under $\mathbb{Q}$. Thus $Y|X = Y|X_S$ and $X_S \perp X_{S^c}$ under $\mathbb{Q}$.

F.3. Proof of Proposition 4. We divide our proof into two parts.

1. Suppose $Y \mid X_S \not\sim Y \mid X_{\hat{S}}$. Below we show that $\text{supp}(\beta)$ is a subset of $S \backslash \hat{S}$ for any local maximum β. Our proof proceeds in three steps.
   - First, we show that $F(\beta; \mathbb{P}^w) > 0$. Let $\tilde{\beta}(\varepsilon) = \beta + \varepsilon \mathbf{1}_S$ for $\varepsilon > 0$. Note $X_S \not\perp Y$ (since $Y \mid X_S \not\sim Y \mid X_{\hat{S}}$). Since $\text{supp}(\tilde{\beta}(\varepsilon)) \supseteq S$, Theorem 1 immediate implies that $F(\tilde{\beta}(\varepsilon); \mathbb{P}^w) > 0$ for any $\varepsilon > 0$. This implies that $F(\beta; \mathbb{P}^w) > 0$ since β is a local maximum.
   - Second, we show that $\text{supp}(\beta) \cap (S \backslash \hat{S}) \neq \emptyset$. Note that
     - $X_{\text{supp}(\beta)} \not\perp Y$. Indeed, we have $F(\beta; \mathbb{P}^\omega) > 0$ since the first step. Thus $X_{\text{supp}(\beta)} \not\perp Y$ by Proposition 1.
     - $X_{S^c \cup \hat{S}} \perp Y$ under $\mathbb{P}^w$ thanks to Proposition 3.

     The above points show that $\text{supp}(\beta)$ can't be a subset of $S^c \cup \hat{S}$. Equivalently, this means that $\text{supp}(\beta) \cap (S \backslash \hat{S}) \neq \emptyset$.
   - Lastly, we show that $\text{supp}(\beta) \cap S^c = \emptyset$. It suffices to prove that $\text{supp}(\beta) = \text{supp}(\beta) \cap S$. Proposition 3 shows that S^c remains the noise under $\mathbb{P}^w$, i.e., $Y|X = Y|X_S$ and $X_S \perp X_{S^c}$ under $\mathbb{P}^w$. Thus $Y|X_{\text{supp}(\beta)} \sim Y|X_{\text{supp}(\beta) \cap S}$ under $\mathbb{P}^w$. As β is the local maximum, the second part of Proposition 1 shows that $\text{supp}(\beta) \cap S = \text{supp}(\beta)$ must be true. This proves $\text{supp}(\beta) \cap S^c = \emptyset$.
2. Suppose $Y|X_S \sim Y|X_{\hat{S}}$. By definition of $\mathbb{P}^w$, we have
$$P^w(x_{\hat{S}}, x_{\hat{S}^c}, y) = \frac{1}{Z} \cdot P(x, y) \cdot P(1 - y \mid x_{\hat{S}})$$
where Z is the normalization factor. Notice that
$$P(x, y) = P(y \mid x_S) P(x_S) P(x_{S^c}) = P(y \mid x_{\hat{S}}) P(x_S) P(x_{S^c}),$$
where the last identity uses $Y|X_S \sim Y|X_{\hat{S}}$. Hence we have
$$P^w(x_{\hat{S}}, x_{\hat{S}^c}, y) = \frac{1}{Z} \cdot P(1 - y \mid x_{\hat{S}}) P(y \mid x_{\hat{S}}) P(x_S) P(x_{S^c}) \tag{43}$$
Note y takes value only from $\{0, 1\}$. Thus, equation (43) shows that the joint distribution $P^w(x_{\hat{S}}, x_{\hat{S}^c}, y)$ is independent of the value $y \in \{0, 1\}$, which implies that $X \perp Y$ under P^w. Hence $F(\beta; \mathbb{P}^w) \equiv 0$ for all β.

F.4. Proof of Proposition 5. In fact, Proposition 5 is an immediate consequence of Proposition 4. By Proposition 4, (i) the population Algorithm 1 never selects any noise variable from S^c, and (ii) the algorithm always adds new signal variables from S as long as $Y|X_S \not\sim Y|X_{\hat{S}}$. Hence, Algorithm 1 terminates in finite iterations, and outputs a set $\hat{S}$ that satisfies the desired properties (i) $Y|X_S \sim Y|X_{\hat{S}}$ and (ii) $\hat{S} \subseteq S$.

APPENDIX G: POPULATION GRADIENT ON NOISE VARIABLES

This section studies the gradient of the *population* objective with respect to noise variables. The results in the section serve as the foundation on establishing the false discovery guarantee of the metric learning algorithm (see the proof of Theorem 2 and Theorem 3 in Section K and Section M).

G.1. Main Results. Below we present the main results of the section. The result of the section only applies to $\mathbb{Q}$ that is balanced: $\mathbb{Q}(Y = 0) = \mathbb{Q}(Y = 1) = \frac{1}{2}$. Let's consider the population objective function

$$F(\beta; \mathbb{Q}) = \mathbb{E}_{B-W}\left[f(\langle \beta, \mathbf{d} \rangle)\right]. \tag{44}$$

Proposition 3 shows that the gradient with respect to noise variable is non-positive. The proof of Proposition 3 is deferred to Section G.2.

PROPOSITION 3. *Assume the following assumption.*

- $Y \mid X = Y \mid X_S$ *and* $X_S \perp X_{S^c}$ *under* $\mathbb{Q}$.
- *Let* $f \in \mathcal{C}^\infty(\mathbb{R}_+)$ *be such that* f' *is strictly completely monotone.*
- *The choice* $q \in \{1, 2\}$.

Then we have for any $\beta \in \mathbb{R}_+^p$ *and any* $j \in S^c$,

$$\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) \leq 0.$$

The inequality is strict if the below two conditions are satisfied:

- $F(\beta; \mathbb{Q}) > 0$
- *the random variable* X_j *is not degenerate.*

While Proposition 3 shows that *qualitatively* the gradient with respect to noise variables is negative, Proposition 4 makes one step further, showing that *quantitatively* the absolute value of the (negative) gradient is lower bounded by the (square of the) objective function. We emphasize that Proposition 4 is the key technical result that leads to the *self-penalizing property* of the metric learning algorithm (see Theorem 2). The proof of Proposition 4 is deferred to Section G.3.

PROPOSITION 4. *Assume the following assumptions.*

- $Y \mid X = Y \mid X_S$ *and* $X_S \perp X_{S^c}$ *under* $\mathbb{Q}$.
- *Let* $f \in \mathcal{C}^\infty(\mathbb{R}_+)$ *be such that* f' *is strictly completely monotone.*
- *The choice* $q \in \{1, 2\}$.

- *For some $M < \infty$, $|X|_\infty \leq M$ almost surely under $\mathbb{Q}$.*
- *In the case where $q = 2$, f' has an analytical extension on the complex plane $\mathbb{C}$ such that $|f'(z)| \leq A(1 + |z|)^N e^{B|\Re z|}$ for some $A, B, N < \infty$.*

Let $\mathcal{B} = \left\{\beta \in \mathbb{R}_+^p : \|\beta\|_1 \leq b\right\}$. Then there exists a constant $c > 0$ depending only on b, M, q, f such that for any $\beta \in \mathcal{B}$ and any noise variable $j \in S^c$,

$$\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) \leq -c \cdot \mathbb{E}\big[\mathbf{d}_j\big] \cdot (F(\beta; \mathbb{Q}))^2 \leq 0.$$

G.2. Proof of Proposition 3. We compute the gradient w.r.t β_j:

$$\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) = \mathbb{E}_{B-W}\left[\mathbf{d}_j \cdot f'\left(\langle \beta, \mathbf{d}\rangle\right)\right]. \tag{45}$$

We condition on X_{S^c}, X'_{S^c} so that $\mathbf{d}_{S^c}$ may be treated as constant for $j \in S^c$. Note that

$$\begin{aligned}\mathbb{E}_B\left[\mathbf{d}_j \cdot f'\left(\langle \beta, \mathbf{d}\rangle\right)\right] =& \mathbb{E}\left[\mathbf{d}_j \cdot \mathbb{E}_B[f'(\langle \beta_{S^c}, \mathbf{d}_{S^c}\rangle + \langle \beta_S, \mathbf{d}_S\rangle) \mid X_{S^c}, X'_{S^c}]\right] \\ \overset{(i)}{\leq}& \mathbb{E}\left[\mathbf{d}_j \cdot \mathbb{E}_W[f'(\langle \beta_{S^c}, \mathbf{d}_{S^c}\rangle + \langle \beta_S, \mathbf{d}_S\rangle) \mid X_{S^c}, X'_{S^c}]\right] \\ =& \mathbb{E}_W\left[\mathbf{d}_j \cdot f'\left(\langle \beta, \mathbf{d}\rangle\right)\right]\end{aligned} \tag{46}$$

For (i), note that $X_S \perp X_{S^c}$ and $Y|X = Y|X_S$ imply $X_{S^c} \perp X_S|Y$ under $\mathbb{Q}$. So the distribution of $\mathbf{d}_S$ is unaffected by conditioning on X_{S^c}, X'_{S^c}. Then (i) follows from the fact that

$$\mathbb{E}_B\left[f'\left(c + \langle \beta_S, \mathbf{d}_S\rangle\right)\right] \geq \mathbb{E}_W\left[f'\left(c + \langle \beta_S, \mathbf{d}_S\rangle\right)\right]$$

for all $c > 0$ since f' is completely monotone (see discussion of translation invariance in Section 2.1). Finally, we notice that if $F(\beta; \mathbb{Q}) > 0$, then $X_{\text{supp}(\beta)}$ is not independent of Y, and if X_j is non-degenerate, then $\mathbf{d}_j > 0$ with positive probability, in which case inequality (i) becomes strict since f' is a strictly completely monotone function.

G.3. Proof of Proposition 4. In the proof, we use the notation $F(\beta; f, \mathbb{Q}) = F(\beta; \mathbb{Q})$ to emphasize the dependence of $F(\beta; \mathbb{Q})$ on f.

Our first step is to derive the expression for the gradient of the $F(\beta; f, \mathbb{Q})$. According to equations (45) and (46) in the proof of Proposition 3, we have the following characterization on the gradient of $F(\beta; f, \mathbb{Q})$. For any $\beta \in \mathcal{B}$ and $j \in S^c$, we have the identity

$$\frac{\partial}{\partial \beta_j} F(\beta; f, \mathbb{Q}) = \mathbb{E}\left[\mathbf{d}_j \cdot \mathbb{E}_{B-W}[f'(\langle \beta, \mathbf{d}\rangle) \mid X_{S^c}, X'_{S^c}]\right]. \tag{47}$$

The key part to establish Proposition 4 is to upper bound the RHS of the above expression, and in particular, the conditional expectation inside the RHS of the above expression. The bound is formally summarized in the following claim, whose proof is deferred into Section G.4.

CLAIM 1. *Under the Assumption of Proposition 4, there exists some constant $c > 0$ that depends only on b, M, q, f such that for any $\beta \in \mathcal{B}$:*

$$\mathbb{E}_{B-W}[f'(\langle \beta, \mathbf{d} \rangle) \mid X_{S^c}, X'_{S^c}] \leq -c \cdot (F(\beta; f, \mathbb{Q}))^2. \tag{48}$$

Let $c > 0$ be any constant such that equation (48) holds for all $\beta \in \mathcal{B}$. Substituting equation (48) into equation (47) yields the desired Proposition 4.

G.4. Proof of Claim 1. The proof is divided into two steps.

- In the first step, we prove a reduction argument, showing that it suffices to prove Claim 1 under the additional assumption that $f'(\infty) = 0$.
- In the second step, we prove Claim 1 holds under the additional assumption that the function f satisfies $f'(\infty) = 0$.

G.4.1. *Step 1: reduction argument.* To see this how this reduction argument works, we introduce the auxiliary function ($f'(\infty)$ is well-defined since f' is completely monotone—see Lemma H.1):

$$\overline{f}(x) = f(x) - f'(\infty)x.$$

Note that $\overline{f}$ satisfies $\overline{f}'(\infty) = 0$ by assumption. Here comes the key argument—we show that by moving from f to $\overline{f}$, the LHS of equation (48) remains the same, while the RHS of equation (48) decreases. More precisely, we prove

1. The LHS of equation (48) remains the same after we substitute the function f by the auxiliary function $\overline{f}$, i.e.,

$$\mathbb{E}_{B-W}[f'(\langle \beta, \mathbf{d} \rangle) \mid X_{S^c}, X'_{S^c}] = \mathbb{E}_{B-W}[\overline{f}'(\langle \beta, \mathbf{d} \rangle) \mid X_{S^c}, X'_{S^c}] \tag{49}$$

2. The RHS of equation (48) decreases after we substitute the function f by the auxiliary function $\overline{f}$, i.e.,

$$F(\beta; f, \mathbb{Q}) \geq F(\beta; \overline{f}, \mathbb{Q}). \tag{50}$$

Here is a quick proof of equations (49) and (50).

1. Equation (49) follows from the fact that the functions $\overline{f}'(x)$ and $f'(x)$ are off by a constant, i.e., $\overline{f}'(x) = f'(x) - f'(\infty)$.

2. To prove equation (50), we start with the following identity:

$$F(\beta; f, \mathbb{Q}) = F(\beta; \overline{f}, \mathbb{Q}) + f'(\infty) \cdot \langle \beta, \mathbb{E}_{B-W}[\mathbf{d}] \rangle. \tag{51}$$

Now we prove that $f'(\infty) \geq 0$ and $\mathbb{E}_{B-W}[\mathbf{d}] \in \mathbb{R}^p_+$. To see this,

- $f'(\infty) \geq 0$ since f' is completely monotone.
- We prove $\mathbb{E}_{B-W}[\mathbf{d}_i] \geq 0$ for all $i \in [p]$. In the case where $q = 1$, $\mathbb{E}_{B-W}[\mathbf{d}_i] = \mathbb{E}_{B-W}[|X_i - X_i'|] \geq 0$ (note that $f(x) = \sqrt{x}$ is completely monotone, and the result follows from Theorem 1). In the case where $q = 2$, a simple calculation yields that $\mathbb{E}_{B-W}[\mathbf{d}_i] = \mathbb{E}_{B-W}[(X_i - X_i')^2] = (\mathbb{E}_{B-W}[X_i])^2 \geq 0$.

Now equation (50) follows from equation (51).

With equations (49) and (50) at hand, it suffices to prove that Claim 1 holds for the function $\overline{f}$. Now the function $\overline{f}$ satisfies all the assumptions stated in the Claim 1 with the additional property $\overline{f}'(\infty) = 0$. One can verify this via

- $\overline{f}'$ is completely monotone (since f' is completely monotone).
- In the case when $q = 2$, $\overline{f}'$ has an analytical extension on the complex plane $\mathbb{C}$ such that $|\overline{f}'(z)| \leq A(1 + |z|)^N e^{B|\Re z|}$ for some $A, B, N < \infty$ (since in the case when $q = 2$, by assumption f has an analytical extension with $|f'(z)| \leq A(1 + |z|)^N e^{B|\Re z|}$ for some $A, B, N < \infty$).

This proves the reduction—we only need to prove that Claim 1 holds under the additional assumption that $f'(\infty) = 0$.

G.4.2. *Step 2: main argument.* Below we prove Claim 1 holds under the additional assumption that $f'(\infty) = 0$. The proof follows from a series of technical inequalities that are detailed in Lemma G.1-G.4 below. For simplicity of the statement, we introduce the notational shorthand $\overline{M} = (2M)^q$.

LEMMA G.1. *Assume the following assumptions:*

- $Y \mid X = Y \mid X_S$ *and* $X_S \perp X_{S^c}$ *under* $\mathbb{Q}$.
- *The function* $f \in \mathcal{C}^\infty(\mathbb{R}_+)$ *and* f' *is completely monotone.*
- *For some* $M < \infty$, $|X|_\infty \leq M$ *almost surely under* $\mathbb{Q}$.

Then the following inequality holds for all $\beta \in \mathcal{B}$*:*

$$\mathbb{E}_{B-W}[f'(\langle \beta, \mathbf{d} \rangle) \mid X_{S^c}, X'_{S^c}] \leq \mathbb{E}_{B-W}\left[f'(\langle \beta_S, \mathbf{d}_S \rangle + \overline{M} b)\right].$$

The proof of Lemma G.1 is deferred into Section G.5.1.

LEMMA G.2. *Assume* f' *is completely monotone and* $f'(\infty) = 0$.

1. *In the case where $q = 1$, we have*

$$\mathbb{E}_{B-W}\left[f'(\langle\beta_S, \mathbf{d}_S\rangle + \overline{M}b)\right] \leq \frac{f^{|S|+1}(\overline{M}b)}{f^{|S|+1}(0)} \cdot \mathbb{E}_{B-W}\left[f'(\langle\beta_S, \mathbf{d}_S\rangle)\right].$$

2. *In the case where $q = 2$, if assuming further that f' has an analytical extension on the complex plane $\mathbb{C}$ such that $|f'(z)| \leq A(1+|z|)^N e^{B|\Re z|}$ for some $A, B, N < \infty$, then we have*

$$\mathbb{E}_{B-W}\left[f'(\langle\beta_S, \mathbf{d}_S\rangle + \overline{M}b)\right] \leq e^{-B\overline{M}b} \cdot \mathbb{E}_{B-W}\left[f'(\langle\beta_S, \mathbf{d}_S\rangle)\right].$$

The proof of Lemma G.2 is deferred into Section G.5.2.

LEMMA G.3. *Assume the following assumptions:*

- *Assume f' is completely monotone on $\mathbb{R}_+$ and $f'(\infty) = 0$.*
- *For some $M < \infty$, $|X|_\infty \leq M$ almost surely under $\mathbb{Q}$.*

Then the following inequality holds for all $\beta \in \mathcal{B}$:

$$\mathbb{E}_{B-W}\left[f'(\langle\beta_S, \mathbf{d}_S\rangle)\right] \leq -\frac{1}{4\overline{M}b|f(0) - f(\infty)|} \cdot \left(\mathbb{E}_{B-W}\left[f(\langle\beta_S, \mathbf{d}_S\rangle)\right]\right)^2.$$

The proof of Lemma G.3 is deferred into Section G.5.3.

LEMMA G.4. *Assume f' is completely monotone and $f'(\infty) = 0$. Then the following inequality holds for all $\beta \in \mathcal{B}$:*

$$\mathbb{E}_{B-W}\left[f(\langle\beta_S, \mathbf{d}_S\rangle)\right] \geq \mathbb{E}_{B-W}\left[f(\langle\beta, \mathbf{d}\rangle)\right] \geq 0.$$

The proof of Lemma G.4 is deferred into Section G.5.4.

As a summary, Lemma G.1–G.4 immediately imply the desired Claim 1 holds under the additional assumption that $f'(\infty) = 0$.

G.4.3. *Summary.* The desired Claim 1 now follows from the results in the above two subsections G.4.1 and G.4.2.

G.5. Proof of Lemma G.1-G.4.

G.5.1. *Proof of Lemma G.1.* Introduce the function

$$G(x) = \mathbb{E}_{B-W}\left[f'(\langle\beta_S, \mathbf{d}_S\rangle + x)\right]. \tag{52}$$

The desired Lemma G.1 follows from the following two claims.

(i) We have the representations that hold for all $\beta \in \mathcal{B}$

$$\begin{aligned} \mathbb{E}_{B-W}[f'(\langle \beta, \mathbf{d}\rangle) \mid X_{S^c}, X'_{S^c}] &= G(\langle \beta_{S^c}, \mathbf{d}_{S^c}\rangle) \\ \mathbb{E}_{B-W}\left[f'(\langle \beta_S, \mathbf{d}_S\rangle + \overline{M}b)\right] &= G(\overline{M}b) \end{aligned} \tag{53}$$

(ii) The function $x \mapsto G(x)$ is monotonically increasing.

The above claims (i) and (ii) immediately imply the desired Lemma G.1 since $\langle \beta_{S^c}, \mathbf{d}_{S^c}\rangle \le \overline{M}b$ holds for all $\beta \in \mathcal{B}$ (by Hölder's inequality).

Below we prove the aforementioned two claims. To see claim (i), we note that $X_{S^c} \perp X_S|Y$ since $X_S \perp X_{S^c}$ and $Y|X = Y|X_S$ by assumption. Hence, the distribution of $\mathbf{d}_S$ is unaffected by conditioning on X_{S^c}, X'_{S^c}. This proves the first equation in (53). The second equation in (53) follows by definition. This proves claim (i). To see claim (ii), we take the derivative of $G(x)$ and obtain that $G'(x) = \mathbb{E}_{B-W}[f''(\langle \beta_S, \mathbf{d}_S\rangle + x)]$. Since $-f''$ is completely monotone (since f' is completely monotone by assumption), this gives $G'(x) \ge 0$ for all x and hence $x \mapsto G(x)$ is increasing. This proves the claim (ii).

G.5.2. *Proof of Lemma G.2.* By Lemma H.1, we have for some nonnegative measure μ on $\mathbb{R}_+$ such that for all integers $k \ge 1$ (note that we have used the assumption $f'(\infty) = 0$ here):

$$f^{(k)}(x) = (-1)^{(k-1)} \cdot \int_0^\infty t^k e^{-tx} \mu(dt) \tag{54}$$

As a result, we can use Fubini's theorem to obtain for all $x \ge 0$:

$$\mathbb{E}_{B-W}\left[f'(\langle \beta_S, \mathbf{d}_S\rangle + x)\right] = \int \mathbb{E}_{B-W}\left[e^{-t\|X_S - X'_S\|^q_{q,\beta_S}}\right] t e^{-tx} \mu(dt). \tag{55}$$

1. Consider the first case where $q = 1$. We aim to prove the inequality

$$-\mathbb{E}_{B-W}\left[f'(\langle \beta_S, \mathbf{d}_S + \overline{M}b\rangle)\right] \ge \frac{f^{|S|+1}(\overline{M}b)}{f^{|S|+1}(0)} \cdot \left(-\mathbb{E}_{B-W}\left[f'(\langle \beta_S, \mathbf{d}_S\rangle)\right]\right). \tag{56}$$

The core technique of the proof is to decouple the two integrands in the RHS of equation (55) using the covariance inequality (see Lemma O.6). Indeed, the covariance inequality states that any pair of monotonically decreasing functions g_1, g_2 and any nonnegative measure $\tilde{\mu}$ must satisfy

$$\int g_1(t) g_2(t) \tilde{\mu}(dt) \ge \frac{1}{|\tilde{\mu}|} \int g_1(t) \tilde{\mu}(dt) \cdot \int g_2(t) \tilde{\mu}(dt). \tag{57}$$

Below we use equation (55) to rewrite the LHS of equation (56) into

$$-\mathbb{E}_{B-W}\left[f'(\langle \beta_S, \mathbf{d}_S\rangle + \overline{M}b)\right] = \int g_1(t) g_2(t) \tilde{\mu}(dt) \tag{58}$$

for appropriately chosen monotonic functions g_1, g_2 and measure $\tilde{\mu}$, and then we will see how covariance inequality leads to the desired bound in equation (56). First we pick $g_1(t) = e^{-t\overline{M}b}$ which is monotonically decreasing. Next, we construct $g_2(t)$ in a careful way. By Lemma H.2, if we denote $q_t(\omega) = \frac{t}{\omega^2+t^2}$, we have the identity (the notation $\phi_{i,S}(\omega_S)$ for $i = 0, 1$ stands for $\phi_{i,S}(\omega_S) = \mathbb{E}[e^{i\langle\omega_S, X_S\rangle} \mid Y = i]$, see Section H for details)

$$\begin{aligned} &- \mathbb{E}_{B-W}\left[e^{-t\left\|X_S - X_S'\right\|_{1,\beta_S}}\right] \\ &= \int |\phi_{0,S}(t\omega_S) - \phi_{1,S}(t\omega_S)|^2 \cdot \prod_{k\in S} q_{\beta_k}(\omega_k) d\omega_S \\ &= \int |\phi_{0,S}(\omega_S) - \phi_{1,S}(\omega_S)|^2 \cdot \prod_{k\in S} q_{t\beta_k}(\omega_k) d\omega_S. \end{aligned} \tag{59}$$

where the last equality follows from change of variables. Let $g_2(t)$ be

$$g_2(t) = \int |\phi_{0,S}(\omega_S) - \phi_{1,S}(\omega_S)|^2 \cdot \prod_{k\in S} \left(\frac{1}{t} \cdot q_{t\beta_k}(\omega_k)\right) d\omega_S. \tag{60}$$

According to equation (59), we have the identity

$$-\mathbb{E}_{B-W}\left[e^{-t\left\|X_S - X_S'\right\|_{1,\beta_S}}\right] = t^{|S|} g_2(t).$$

In addition, the function $t \to g_2(t)$ is monotonically decreasing since $t \to \frac{1}{t} \cdot q_{t\beta}(\omega) = \frac{1}{\pi} \cdot \frac{\beta}{\omega^2 + t^2\beta^2}$ is monotonically decreasing. Finally, we pick $\tilde{\mu}$ to be $d\tilde{\mu} = t^{|S|+1} d\mu$. One can then easily verify that equation (58) holds for the functions g_1, g_2 and the measure $\tilde{\mu}$ that we pick. As a consequence of the covariance inequality, we obtain the lower bound

$$-\mathbb{E}_{B-W}\left[f'(\langle\beta_S, \mathbf{d}_S\rangle + x)\right] \geq \frac{1}{|\tilde{\mu}|} \cdot \int g_1(t)\tilde{\mu}(dt) \cdot \int g_2(t)\tilde{\mu}(dt). \tag{61}$$

Now we evaluate the RHS. By equation (54) and (55), we obtain

$$\begin{aligned} |\tilde{\mu}| &= \int t^{|S|+1}\mu(dt) = |f^{|S|+1}(0)|. \\ \int g_1(t)\tilde{\mu}(dt) &= \int e^{-t\overline{M}b} t^{|S|+1}\mu(dt) = |f^{|S|+1}(\overline{M}b)| \\ \int g_2(t)\tilde{\mu}(dt) &= \mathbb{E}_{B-W}\left[f'(\langle\beta_S, \mathbf{d}_S\rangle)\right]. \end{aligned} \tag{62}$$

Substituting equations (62) into equation (61) immediately yields the desired equation (56). This completes the proof.

2. Consider the second case where $q = 2$. The Schwartz–Paley–Wiener theorem implies that μ is compactly supported, with $\text{supp}(\mu) \subseteq [0, B]$. By equation (55), it is immediate that

$$\mathbb{E}_{B-W}\left[f'(\langle \beta_S, \mathbf{d}_S\rangle + \overline{M}b)\right] \leq e^{-B\overline{M}b} \cdot \mathbb{E}_{B-W}\left[f'(\langle \beta_S, \mathbf{d}_S\rangle)\right]. \tag{63}$$

The desired result follows from equation (56) and equation (63).

G.5.3. *Proof of Lemma G.3.* By Lemma H.1, we have for some non-negative measure μ on $\mathbb{R}_+$ (note the assumption $f'(\infty) = 0$),

$$f(x) = f(\infty) - \int_0^\infty e^{-tx}\mu(dt) \quad \text{and} \quad f'(x) = \int_0^\infty te^{-tx}\mu(dt) \tag{64}$$

As a result, we can use Fubini's theorem to write

$$\begin{aligned} \mathbb{E}_{B-W}\left[f'(\langle \beta_S, \mathbf{d}_S\rangle)\right] &= \int_0^\infty t \cdot \mathbb{E}_{B-W}\left[e^{-t\langle \beta_S, \mathbf{d}_S\rangle}\right]\mu(dt). \\ \mathbb{E}_{B-W}\left[f(\langle \beta_S, \mathbf{d}_S\rangle)\right] &= -\int_0^\infty \mathbb{E}_{B-W}\left[e^{-t\langle \beta_S, \mathbf{d}_S\rangle}\right]\mu(dt). \end{aligned} \tag{65}$$

Let $\delta > 0$ be a constant to be determined. We apply Markov's inequality to the first equation of (65), and we derive for all $\delta > 0$

$$\begin{aligned} &\mathbb{E}_{B-W}\left[f'(\langle \beta_S, \mathbf{d}_S\rangle)\right] \leq \delta \cdot \int_\delta^\infty \mathbb{E}_{B-W}\left[e^{-t\langle \beta_S, \mathbf{d}_S\rangle}\right]\mu(dt) \\ &= -\delta \cdot \left(\mathbb{E}_{B-W}\left[f(\langle \beta_S, \mathbf{d}_S\rangle)\right] + \int_0^\delta \mathbb{E}_{B-W}\left[e^{-t\langle \beta_S, \mathbf{d}_S\rangle}\right]\mu(dt)\right). \end{aligned} \tag{66}$$

Now we upper bound the integral in the last line of equation (66). Using the fact that $\langle \beta_S, \mathbf{d}_S\rangle \leq \overline{M}b$ and the elementary inequality $|e^{-x} - 1| \leq |x|$ for all $x \geq 0$, we obtain for all $\delta > 0$,

$$\begin{aligned} \left|\int_0^\delta \mathbb{E}_{B-W}\left[e^{-t\langle \beta_S, \mathbf{d}_S\rangle}\right]\mu(dt)\right| &= \left|\int_0^\delta \mathbb{E}_{B-W}\left[\left(e^{-t\langle \beta_S, \mathbf{d}_S\rangle} - 1\right)\right]\mu(dt)\right| \\ &\leq \int_0^\delta 2\overline{M}bt\mu(dt) \leq 2\overline{M}b|f(0) - f(\infty)|\delta. \end{aligned}$$

Substituting the bound into equation (66) yields for all $\delta > 0$

$$\mathbb{E}_{B-W}\left[f'(\langle \beta_S, \mathbf{d}_S\rangle)\right] \leq -\delta \cdot \left(\mathbb{E}_{B-W}\left[f(\langle \beta_S, \mathbf{d}_S\rangle)\right] - 2\overline{M}b|f(0) - f(\infty)|\delta\right).$$

Optimizing $\delta > 0$ on the RHS yields the final bound

$$\mathbb{E}_{B-W}\left[f'(\langle \beta_S, \mathbf{d}_S\rangle)\right] \leq -\frac{1}{4\overline{M}b|f(0) - f(\infty)|} \cdot \left(\mathbb{E}_{B-W}\left[f(\langle \beta_S, \mathbf{d}_S\rangle)\right]\right)^2.$$

G.5.4. *Proof of Lemma G.4.* This is implied by Lemma H.3.

APPENDIX H: PROPERTIES OF $F(\beta;\mathbb{Q})$—GENERAL RESULT

The result of this section applies to the objective $F(\beta;\mathbb{Q})$ when $q=1$ or $q=2$. We recall the definition of $F(\beta;\mathbb{Q})$:

$$F(\beta;\mathbb{Q}) = \mathbb{E}_{B-W}\left[f\big(\|X-X'\|_{q,\beta}^q\big)\right] = \mathbb{E}_{B-W}\left[f\big(\langle\beta,\mathbf{d}\rangle\big)\right].$$

The section assumes $\mathbb{Q}$ is balanced: $\mathbb{Q}(Y=0)=\mathbb{Q}(Y=1)=\frac{1}{2}$. This assumption assures the following identity that holds for any function h,

$$\begin{aligned}&\mathbb{E}_{B-W}\left[h(X,X')\right]\\ &=\frac{1}{2}\int h(x,x')(Q_0(dx)-Q_1(dx))(Q_0(dx')-Q_1(dx')),\end{aligned}$$

where $\mathbb{Q}_0,\mathbb{Q}_1$ denote the conditional distribution $X|Y=0$ and $X|Y=1$.

H.1. Notation. Let $\mathbb{Q}_0$ and $\mathbb{Q}_1$ denote the conditional distribution of X given $Y=0$ and $Y=1$. Write

$$\phi_0(\omega)=\mathbb{E}_0\left[e^{i\langle\omega,X\rangle}\right] \quad\text{and}\quad \phi_1(\omega)=\mathbb{E}_1\left[e^{i\langle\omega,X\rangle}\right].$$

We use Q_β to denote the following function (depending on the choice of q):

1. $Q_\beta(\omega)$ is the Cauchy density with scale β when $q=1$.

$$Q_\beta(\omega)=\frac{1}{\pi}\frac{\beta}{\omega^2+\beta^2}.$$

2. $Q_\beta(\omega)$ is the Gaussian density with scale β when $q=2$.

$$Q_\beta(\omega)=\frac{1}{\sqrt{2\pi\beta^2}}e^{-\frac{\omega^2}{2\beta^2}}.$$

H.2. Main Result. Lemma H.1 below gives a useful characterization of the function f whose derivative f' is strictly completely monotone with $f'(\infty)=0$. We defer the proof of Lemma H.1 into Section H.3.

LEMMA H.1. *Assume Assumption (A1). Then for some scalar $a\in\mathbb{R}$ and some non-negative measure μ on $[0,\infty)$ with $\mu(\mathbb{R}_+)>0$, we have the representation:*

$$f(x)=a+f'(\infty)x-\int_0^\infty e^{-tx}\mu(dt), \tag{67}$$

where $f'(\infty) = \lim_{x\to\infty} f'(x)$. *Moreover, we have that*

$$\begin{aligned} f'(x) &= f'(\infty) + \int_0^\infty te^{-tx}\mu(dt) \\ f^{(k)}(x) &= (-1)^{k-1}\cdot\int_0^\infty t^k e^{-tx}\mu(dt) \quad \text{for } k\geq 2,\ k\in\mathbb{N}. \end{aligned} \tag{68}$$

Lemma H.2 below gives useful representations of the objective $F(\beta;\mathbb{Q})$. We defer the proof of Lemma H.2 in Section H.4.

LEMMA H.2. *Assume Assumption (A1). Then, for some non-negative measure* μ *on* $(0,\infty)$ *with* $\mu(\mathbb{R}_+) > 0$, *we have the representations:*

$$F(\beta;\mathbb{Q}) = F_1(\beta;\mathbb{Q}) + F_2(\beta;\mathbb{Q}) \tag{69}$$

where $F_1(\beta;\mathbb{Q})$ *and* $F_2(\beta;\mathbb{Q})$ *are defined by*

$$\begin{aligned} F_1(\beta;\mathbb{Q}) &= f'(\infty)\cdot\mathbb{E}_{B-W}\left[\left\|X-X'\right\|_{q,\beta}^q\right] \\ F_2(\beta;\mathbb{Q}) &= -\int \mathbb{E}_{B-W}\left[e^{-t\|X-X'\|_{q,\beta}^q}\right]\mu(dt) \\ &= \iint |\phi_0(t\omega)-\phi_1(t\omega)|^2\cdot\prod_{k=1}^p Q_{\beta_k}(\omega_k)d\omega\mu(dt). \end{aligned} \tag{70}$$

In above, the measure μ *also satisfies*

$$\begin{aligned} f'(x) &= f'(\infty) + \int_0^\infty te^{-tx}\mu(dt) \\ f^{(k)}(x) &= (-1)^{k-1}\cdot\int_0^\infty t^k e^{-tx}\mu(dt) \quad \text{for } k\geq 2,\ k\in\mathbb{N}. \end{aligned} \tag{71}$$

In addition, we have for $\beta\in\mathbb{R}_+^p$ *and* $j\in[p]$,

$$\frac{\partial}{\partial\beta_j}F(\beta;\mathbb{Q}) = \frac{\partial}{\partial\beta_j}F_1(\beta;\mathbb{Q}) + \frac{\partial}{\partial\beta_j}F_2(\beta;\mathbb{Q})$$

where we have

$$\frac{\partial}{\partial\beta_j}F_1(\beta;\mathbb{Q}) = f'(\infty)\cdot\mathbb{E}_{B-W}\left[|X_j-X_j'|^q\right]$$

and we have for all $\beta\in\mathbb{R}_+^p$

$$\frac{\partial}{\partial\beta_j}F_2(\beta;\mathbb{Q}) = \int \mathbb{E}_{B-W}\left[t|X_j-X_j'|^q e^{-t\|X-X'\|_{q,\beta}^q}\right]\mu(dt)$$

and for $\beta \in \mathbb{R}_+^p$ *with* $\beta_j > 0$,

$$\frac{\partial}{\partial \beta_j} F_2(\beta; \mathbb{Q}) = \iint |\phi_0(t\omega) - \phi_1(t\omega)|^2 \cdot \prod_{k \neq j} Q_{\beta_k}(\omega_k) \cdot \frac{\partial}{\partial \beta_j} Q_{\beta_j}(\omega_j) d\omega \mu(dt).$$

Lemma H.3 below essentially says that at any fixed $\beta \in \mathbb{R}_+^p$, throwing away noise variables (by setting the coordinates on S^c to be 0) increases the objective value. We defer the proof of Lemma H.3 into Section H.5.

LEMMA H.3. *Let Assumption (A1) hold. Assume* $f'(\infty) = 0$. *Let* $\beta \in \mathbb{R}_+^p$ *and define* $\bar{\beta}$ *to be the vector such that* $\bar{\beta}_S = \beta_S$ *and* $\bar{\beta}_{S^c} = 0$. *Then we have*

$$F(\bar{\beta}; \mathbb{Q}) \geq F(\beta; \mathbb{Q}).$$

H.3. Proof of Lemma H.1. As f' is strictly completely monotone, Bernstein's theorem for completely monotone functions [11] shows that f admits the Lévy–Khintchine representation

$$f(x) = a + cx - \int_0^\infty e^{-tx} \mu(dt). \tag{72}$$

where $c \geq 0$ and μ is a non-negative finite measure on $[0, \infty)$ with $\mu((0, \infty)) > 0$, $\int_0^\infty (t \wedge 1) \mu(dt) < \infty$. Applying dominated convergence theorem, we obtain

$$f'(x) = c + \int_0^\infty t e^{-tx} \mu(dt). \tag{73}$$

Again, by dominated convergence theorem, we derive $f'(\infty) = c$. Substituting it back into equation (72) gives the desired equation (67). Applying dominated convergence theorem to equation (67) yields equation (68).

H.4. Proof of Lemma H.2. By Lemma H.1, we have for some scalar $a \in \mathbb{R}$ and non-negative finite measure μ on $[0, \infty)$ satisfying $\mu(\mathbb{R}_+) > 0$,

$$f(x) = a + f'(\infty)x - \int_0^\infty e^{-tx} \mu(dt).$$

As a result, we obtain the identity

$$f\left(\|X - X'\|_{q,\beta}^q \right) = a + f'(\infty) \|X - X'\|_{q,\beta}^q - \int_0^\infty e^{-t\|X - X'\|_{q,\beta}^q} d\mu(t)$$

and therefore, Fubini's theorem yields $F(\beta; \mathbb{Q}) = F_1(\beta; \mathbb{Q}) + F_2(\beta; \mathbb{Q})$ where

$$\begin{aligned} F_1(\beta; \mathbb{Q}) &= f'(\infty) \cdot \mathbb{E}_{B-W} \left[\|X - X'\|_{q,\beta}^q \right] \\ F_2(\beta; \mathbb{Q}) &= -\int_0^\infty \mathbb{E}_{B-W} \left[e^{-t\|X - X'\|_{q,\beta}^q} \right] d\mu(t). \end{aligned} \tag{74}$$

Now, we prove the second expression of $F_2(\beta;\mathbb{Q})$, i.e., the last identity in equation (70). To do so, we use the fact that the Fourier transform of the Laplace function is Cauchy density (for $q=1$), and the Fourier transform of the Gaussian density function is still the Gaussian density function (for $q=2$). This gives for $q\in\{1,2\}$

$$e^{-t\|X-X'\|_{q,\beta}^q}=\int_{\mathbb{R}^d} e^{-it\langle\omega,X-X'\rangle}\cdot\prod_{j=1}^p Q_{\beta_j}(\omega_j)d\omega. \tag{75}$$

where we recall Q_β is the Cauchy density with scale β. Substituting equation (75) into equation (74), we obtain

$$\begin{aligned} F_2(\beta;\mathbb{Q}) &= -\int_0^\infty \mathbb{E}_{B-W}\left[\int_{\mathbb{R}^d} e^{-it\langle\omega,X-X'\rangle}\cdot\prod_{j=1}^p Q_{\beta_j}(\omega_j)d\omega\right]d\mu(t)\\ &= -\int_0^\infty\int_{\mathbb{R}^d}\mathbb{E}_{B-W}\left[e^{-it\langle\omega,X-X'\rangle}\right]\cdot\prod_{j=1}^p Q_{\beta_j}(\omega_j)d\omega d\mu(t), \end{aligned} \tag{76}$$

where the second identity follows from Fubini's theorem. Now that X, X' are independent copies, and recall we use $\mathbb{Q}_0$ and $\mathbb{Q}_1$ to denote the conditional distribution $X|Y=0$ and $X|Y=1$.Therefore,

$$\begin{aligned} &\mathbb{E}_{B-W}\left[e^{-it\langle\omega,X-X'\rangle}\right]\\ &= -\frac{1}{2}\iint e^{-it\langle\omega,x-x'\rangle}(Q_0(dx)-Q_1(dx))(Q_0(dx')-Q_1(dx'))\\ &= -\frac{1}{2}\left|\int e^{-it\langle\omega,x\rangle}(dQ_0(x)-dQ_1(x))\right|^2 = -\frac{1}{2}\left|\phi_0(t\omega)-\phi_1(t\omega)\right|^2. \end{aligned} \tag{77}$$

Substituting equation (77) into equation (76) yields the second expression in equation (70) as desired. Up to here, we have proved equations (69) and (70).

Finally, using dominated convergence theorem, it is easy to show that the expressions on the gradients of $F(\beta;\mathbb{Q})$. We omit the details.

H.5. Proof of Lemma H.3. Lemma H.2 shows that for some nonnegative finite measure μ with $\mu(\mathbb{R}_+)>0$, we have the representation

$$F(\beta;\mathbb{Q})=-\int\mathbb{E}_{B-W}\left[e^{-t\|X-X'\|_{q,\beta}^q}\right]\mu(dt) \tag{78}$$

Recall $X_S \perp X_{S^c} \mid Y$ (since $X_S \perp X_{S^c}$ and $Y \mid X = Y \mid X_{S^c}$). Hence

$$\begin{aligned}
&\mathbb{E}_{B-W}\left[e^{-t\|X-X'\|_{q,\beta}^q}\right] \\
&= \mathbb{E}_{B-W}\left[e^{-t\left\|X_S-X'_S\right\|_{q,\beta_S}^q} \cdot e^{-t\left\|X_{S^c}-X'_{S^c}\right\|_{q,\beta_S^c}^q}\right] \\
&= \mathbb{E}_{B-W}\left[\mathbb{E}\left[e^{-t\left\|X_S-X'_S\right\|_{q,\beta_S}^q} \mid Y, Y'\right] \cdot \mathbb{E}\left[e^{-t\left\|X_{S^c}-X'_{S^c}\right\|_{q,\beta_S^c}^q} \mid Y, Y'\right]\right] \\
&= \mathbb{E}_{B-W}\left[e^{-t\left\|X_S-X'_S\right\|_{q,\beta_S}^q}\right] \cdot \mathbb{E}\left[e^{-t\left\|X_{S^c}-X'_{S^c}\right\|_{q,\beta_S^c}^q}\right].
\end{aligned}$$

Substituting the expression into equation (78), we obtain

$$\begin{aligned}
F(\beta;\mathbb{Q}) &= -\int \mathbb{E}\left[e^{-t\left\|X_{S^c}-X'_{S^c}\right\|_{1,\beta_S^c}}\right] \cdot \mathbb{E}_{B-W}\left[e^{-t\left\|X_S-X'_S\right\|_{1,\beta_S}}\right] \mu(dt) \\
&\le -\int \mathbb{E}_{B-W}\left[e^{-t\left\|X_S-X'_S\right\|_{1,\beta_S}}\right] \mu(dt) = F(\bar{\beta};\mathbb{Q}).
\end{aligned}$$

This proves the desired Lemma H.3.

APPENDIX I: PROPERTIES OF THE OBJECTIVE $F(\beta;\mathbb{Q})$—THE ℓ_1 CASE

The result of this section only applies to the objective $F(\beta;\mathbb{Q})$ when $\mathbf{q} = \mathbf{1}$. We recall the definition of $F(\beta;\mathbb{Q})$:

$$F(\beta;\mathbb{Q}) = \mathbb{E}_{B-W}\left[f\left(\left\|X - X'\right\|_{1,\beta}\right)\right] = \mathbb{E}_{B-W}\left[f\left(\langle \beta, \mathbf{d}\rangle\right)\right].$$

The section assumes $\mathbb{Q}$ is balanced: $\mathbb{Q}(Y=0) = \mathbb{Q}(Y=1) = \frac{1}{2}$.

I.1. Notation. The notation of this section follows the notation in Section H. Let $\mathbb{Q}_0$ and $\mathbb{Q}_1$ denote the conditional distribution of X given $Y=0$ and $Y=1$. Write

$$\phi_0(\omega) = \mathbb{E}_0\left[e^{i\langle\omega,X\rangle}\right] \quad \text{and} \quad \phi_1(\omega) = \mathbb{E}_1\left[e^{i\langle\omega,X\rangle}\right].$$

We use Q_β to denote the Cauchy density with scale β:

$$Q_\beta(\omega) = \frac{1}{\pi}\frac{\beta}{\omega^2+\beta^2}.$$

For any subset $A \subseteq [p]$, we introduce (with notation abuse)

$$F_A(\beta_A;\mathbb{Q}) = \mathbb{E}_{B-W}\left[f\left(\left\|X_A - X'_A\right\|_{1,\beta_A}\right)\right].$$

$$\phi_{0,A}(\omega_A) = \mathbb{E}_0\left[e^{i\langle\omega_A,X_A\rangle}\right] \quad \text{and} \quad \phi_{1,A}(\omega_A) = \mathbb{E}_1\left[e^{i\langle\omega_A,X_A\rangle}\right].$$

I.2. Main Result. Lemma I.1 below shows $F(\beta;\mathbb{Q})$ satisfies a self-bounding property. We defer the proof of Lemma I.1 in Section I.3.

LEMMA I.1. *Let Assumption (A1) hold. Then $F(\beta;\mathbb{Q})$ satisfies a self-bounding property: for any $\beta,\beta' \in \mathbb{R}^p$ such that $0 \le \beta_j \le \beta'_j$ for all $j \in [p]$,*

$$F(\beta;\mathbb{Q}) \ge F(\beta';\mathbb{Q}) \cdot \prod_{j=1}^{p} \left(\frac{\beta_j}{\beta'_j}\right). \tag{79}$$

Lemma I.2 below studies the derivative of $F(\beta;\mathbb{Q})$ when the signal is a pure interaction. We defer the proof of Lemma I.2 in Section I.4.

LEMMA I.2. *Let Assumption (A1) hold. Assume that $\|X\|_\infty \le M$ almost surely under $\mathbb{Q}$ for some $M < \infty$. Assume the signal is a pure interaction: $X_A \perp Y$ for any strict subset $A \subsetneq S$. Then, for any $\beta \in \mathcal{B}$ and any signal variable $j \in S$*

$$\frac{\partial}{\partial \beta_j} F(\beta;\mathbb{Q}) = \frac{1}{\beta_j} \cdot F(\beta;\mathbb{Q}) - R(\beta;\mathbb{Q}), \tag{80}$$

where the remainder term $R(\beta;\mathbb{Q})$ satisfies

$$0 \le R(\beta;\mathbb{Q}) \le \pi \cdot (8M)^{|S|+1} \cdot f^{(|S|+1)}(0) \cdot \prod_{k\in S} \beta_k.$$

Lemma I.3 below provides a lower bound of $F(\beta;\mathbb{P})$ using $F_S(\beta_S;\mathbb{P})$. To better appreciate Lemma I.3, the reader should compare it with Lemma H.3 where we derive an upper bound of $F(\beta;\mathbb{P})$ using $F_S(\beta_S;\mathbb{P})$ for all $\beta \in \mathbb{R}^p_+$. We defer the proof of Lemma I.3 into Section I.5.

LEMMA I.3. *Let Assumption (A1) hold. Assume that $\|X\|_\infty \le M$ almost surely under $\mathbb{Q}$ for some $M < \infty$. Then, for any $\beta \in \mathcal{B}$:*

$$F(\beta;\mathbb{Q}) \ge \frac{|f^{|S|}(2Mb)|}{|f^{|S|}(0)|} \cdot F_S(\beta_S;\mathbb{Q}). \tag{81}$$

Lemma I.4 below derives a lower bound for the derivative of $F(\beta;\mathbb{Q})$—this lower bound is particularly useful in the study of hierarchical interaction recovery (it is one of the key lemma that gives the signal term). We provide the proof of Lemma I.4 into Section I.6.

LEMMA I.4. *Assume Assumption (A1). Assume $Y \perp X_A$ for some $A \subseteq [p]$. For any β such that* $\mathrm{supp}(\beta) \subseteq A$, *and variable $j \in A^c$, we have*

$$\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) \geq \frac{1}{\tau} \cdot F(\beta_\tau; \mathbb{Q}),$$

where β_τ is the following vector that differs from β only at coordinate j:

$$(\beta_\tau)_i = \begin{cases} \tau & \text{if } i = j \\ \beta_i & \text{if } i \neq j. \end{cases}$$

I.3. Proof of Lemma I.1. Lemma H.2 shows the existence of a non-negative measure μ on $[0, \infty)$ such that $F(\beta; \mathbb{Q}) = F_1(\beta; \mathbb{Q}) + F_2(\beta; \mathbb{Q})$ where

(82)
$$F_1(\beta; \mathbb{Q}) = f'(\infty) \cdot \mathbb{E}_{B-W}\left[\left\|X - X'\right\|_{1,\beta}\right]$$
$$F_2(\beta; \mathbb{Q}) = \iint \left| \int e^{-it\langle \omega, x\rangle} (dP_0(x) - dP_1(x)) \right|^2 \cdot \prod_{j=1}^p Q_{\beta_j}(\omega_j)\, d\omega d\mu(t).$$

Now we show both $F_1(\beta; \mathbb{Q})$ and $F_2(\beta; \mathbb{Q})$ satisfy the self bounding property.

Showing that $F_1(\beta; \mathbb{Q})$ satisfies the self-bounding property is simple. A direct computation gives for any vector $\beta, \beta' \in \mathbb{R}^p_+$ with $0 \leq \beta_j \leq \beta'_j$

$$F_1(\beta; \mathbb{Q}) \geq F_1(\beta'; \mathbb{Q}) \cdot \prod_{j=1}^p \left(\frac{\beta_j}{\beta'_j}\right). \tag{83}$$

Showing that $F_2(\beta; \mathbb{Q})$ satisfies the self-bounding property requires a little bit more thinking. The key observation is that $\beta \to Q_\beta(\omega)$ satisfies a self-bounding property: for any scalars β, β' with $\beta \leq \beta'$,

$$Q_\beta(\omega) = \frac{1}{\pi} \frac{\beta}{w^2 + \beta^2} \geq \frac{1}{\pi} \frac{\beta}{w^2 + (\beta')^2} = Q_{\beta'}(w) \cdot \frac{\beta}{\beta'}.$$

Hence, for any vectors $\beta, \beta' \in \mathbb{R}^p$ with $0 \leq \beta_j \leq \beta'_j$,

$$\prod_{j=1}^p Q_{\beta_j}(\omega_j) \geq \prod_{j=1}^p Q_{\beta'_j}(\omega_j) \cdot \prod_{j=1}^p \left(\frac{\beta_j}{\beta'_j}\right).$$

As such, we see from equation (82) that $F_2(\beta; \mathbb{Q})$ must satisfy

$$F_2(\beta; \mathbb{Q}) \geq F_2(\beta'; \mathbb{Q}) \cdot \prod_{j=1}^p \left(\frac{\beta_j}{\beta'_j}\right)$$

for any vectors $\beta, \beta' \in \mathbb{R}^p$ satisfying $0 \leq \beta_j \leq \beta'_j$. This completes the proof.

I.4. Proof of Lemma I.2. Lemma H.2 shows the existence of a non-negative measure μ on $[0,\infty)$ such that $F(\beta;\mathbb{Q}) = F_1(\beta;\mathbb{Q}) + F_2(\beta;\mathbb{Q})$ where

$$F_1(\beta;\mathbb{Q}) = f'(\infty)\cdot \mathbb{E}_{B-W}\left[\left\|X - X'\right\|_{1,\beta}\right]. \tag{84}$$

$$F_2(\beta;\mathbb{Q}) = \iint \left|\int e^{-it\langle\omega,x\rangle}(dP_0(x) - dP_1(x))\right|^2 \cdot \prod_{j=1}^{p} Q_{\beta_j}(\omega_j)\, d\omega d\mu(t).$$

Below we show that for any signal variable $j \in S$, we have

$$\begin{aligned} &\frac{\partial}{\partial\beta_j}F_1(\beta;\mathbb{Q}) = \frac{1}{\beta_j}F_1(\beta;\mathbb{Q}).\\ &\frac{\partial}{\partial\beta_j}F_2(\beta;\mathbb{Q}) = \frac{1}{\beta_j}F_2(\beta;\mathbb{Q}) - R(\beta;\mathbb{Q})\\ &\text{where}\quad R(\beta;\mathbb{Q}) \le \pi\cdot|f^{(|S|+1)}(0)|\cdot(8M)^{|S|+1}\cdot\prod_{k\in S}\beta_k. \end{aligned} \tag{85}$$

Equation (85) in conjunction with the identity $F(\beta;\mathbb{Q}) = F_1(\beta;\mathbb{Q}) + F_2(\beta;\mathbb{Q})$ immediately yield the desired Lemma I.2.

Now we prove equation (85) holds for any signal variable $j \in S$. First of all, we can easily derive the first identity of equation (85). Indeed, for $j \in S$

$$\frac{\partial}{\partial\beta_j}F_1(\beta;\mathbb{Q}) = \frac{1}{\beta_j}F_1(\beta;\mathbb{Q}). \tag{86}$$

To see this, recall our assumption that X_S is a pure interaction. Hence, in the case where $|S| \ge 2$, $F_1(\beta;\mathbb{Q}) \equiv 0$ since $X_i \perp Y$ for any $i \in [p]$, for which equation (86) trivially follows. Similarly, in the case where $|S| = 1$, $F_1(\beta;\mathbb{Q})$ is a linear function of β_j, and hence equation (86) trivially holds.

Next, we show that for any $j \in S$,

$$\begin{aligned} &\frac{\partial}{\partial\beta_j}F_2(\beta;\mathbb{Q}) = \frac{1}{\beta_j}F_2(\beta;\mathbb{Q}) - R(\beta;\mathbb{Q})\\ &\text{where}\quad R(\beta;\mathbb{Q}) \le \pi\cdot|f^{(|S|+1)}(0)|\cdot(8M)^{|S|+1}\cdot\prod_{k\in S}\beta_k. \end{aligned} \tag{87}$$

The proof of this part is non-trivial. To start with, Lemma H.2 gives the expression for the gradient of $F_2(\beta;\mathbb{Q})$: for any $\beta \in \mathbb{R}^p_+$ with $\beta_j > 0$, we have

$$\frac{\partial}{\partial\beta_j}F_2(\beta;\mathbb{Q}) = \iint |\phi_0(t\omega) - \phi_1(t\omega)|^2 \cdot \frac{\partial}{\partial\beta_j}Q_{\beta_j}(\omega_j)\cdot\prod_{k\neq j}Q_{\beta_k}(\omega_k)d\omega\mu(dt). \tag{88}$$

Since $Q_\beta(\omega) = \frac{\beta}{\beta^2+\omega^2}$ is the Cauchy density function, a simple calculation gives for any scalar $\beta > 0$, $\omega \geq 0$ the identity below:

$$\frac{\partial}{\partial\beta} Q_\beta(\omega) = \frac{1}{\beta} \cdot Q_\beta(\omega) - \pi \cdot Q_\beta^2(\omega).$$

Substitute it into equation (88). We obtain for all β with $\beta_j > 0$,

(89)
$$\frac{\partial}{\partial\beta_j} F_2(\beta; \mathbb{Q}) = \frac{1}{\beta_j} \cdot F_2(\beta; \mathbb{Q}) - R(\beta; \mathbb{Q})$$
$$\text{where} \quad R(\beta; \mathbb{Q}) = \pi \cdot \iint |\phi_0(t\omega) - \phi_1(t\omega)|^2 \cdot Q_{\beta_j}^2(\omega_j) \cdot \prod_{k\neq j} Q_{\beta_k}(\omega_k) d\omega \mu(dt)$$

Now we upper bound $R(\beta; \mathbb{Q})$. Lemma I.5 bounds $|\phi_0(\omega) - \phi_1(\omega)|$.

LEMMA I.5. *Under the assumption of Lemma I.2, we have for all ω,*

$$|\phi_0(\omega) - \phi_1(\omega)| \leq 2 \prod_{j\in S} (|M\omega_j| \wedge 2)$$

Lemma I.5 shows $|\phi_0(t\omega) - \phi_1(t\omega)| \leq 2 \cdot \prod_{j\in S} (tM|\omega_j| \wedge 2)$ for any $t > 0$. As a result, we obtain that

(90)
$$\int |\phi_0(t\omega) - \phi_1(t\omega)|^2 \cdot Q_{\beta_j}^2(\omega_j) \cdot \prod_{k\neq j} Q_{\beta_j}(\omega_j) d\omega$$
$$\leq 2 \cdot \int \prod_{k\in S} (tM|\omega_k| \wedge 2)^2 \cdot Q_{\beta_j}^2(\omega_j) \cdot \prod_{k\neq j} Q_{\beta_j}(\omega_j) d\omega$$
$$= 2 \cdot \int (tM|\omega_j| \wedge 2)^2 \cdot Q_{\beta_j}^2(\omega_j) d\omega_j \cdot \prod_{k\neq j, k\in S} \int (tM|\omega_k| \wedge 2)^2 \cdot Q_{\beta_k}(\omega_k) d\omega_k,$$

Lemma I.6 upper bounds the integrals. We defer its proof to Section I.8.

LEMMA I.6. *We have for any $\alpha, \beta > 0$,*

$$\int (\alpha |\omega| \wedge 2)^2 \cdot Q_\beta(\omega) d\omega \leq 8\alpha\beta,$$
$$\int (\alpha |\omega| \wedge 2)^2 \cdot Q_\beta^2(\omega) d\omega \leq 4\alpha^2\beta.$$

We apply Lemma I.6 to equation (90), and we derive

$$\int |\phi_0(t\omega) - \phi_1(t\omega)|^2 \cdot Q_{\beta_j}^2(\omega_j) \cdot \prod_{k \neq j} Q_{\beta_j}(\omega_j) d\omega \leq (8tM)^{|S|+1} \cdot \prod_{k \in S} \beta_k.$$

Substituting the above bound into equation (89), we obtain that

$$R(\beta; \mathbb{Q}) \leq \pi \cdot (8M)^{|S|+1} \prod_{k \in S} \beta_k \cdot \int t^{|S|+1} d\mu(t). \tag{91}$$

Now that Lemma H.2 shows $\int t^k d\mu(t) = (-1)^{k-1} f^{(k)}(0)$ for all $k \geq 2$. Thus

$$R(\beta; \mathbb{Q}) \leq \pi \cdot |f^{(|S|+1)}(0)| \cdot (8M)^{|S|+1} \cdot \prod_{k \in S} \beta_k.$$

I.5. Proof of Lemma I.3. Lemma H.2 shows the existence of a non-negative measure μ on $[0, \infty)$ such that $F(\beta; \mathbb{Q}) = F_1(\beta; \mathbb{Q}) + F_2(\beta; \mathbb{Q})$ where

$$\begin{aligned} F_1(\beta; \mathbb{Q}) &= f'(\infty) \cdot \mathbb{E}_{B-W} \left[\left\| X - X' \right\|_{1,\beta} \right]. \\ F_2(\beta; \mathbb{Q}) &= - \int \mathbb{E}_{B-W} \left[e^{-t \| X - X' \|_{1,\beta}} \right] \mu(dt). \end{aligned} \tag{92}$$

Now it suffices to show that both $F_1(\beta; \mathbb{Q})$ and $F_2(\beta; \mathbb{Q})$ satisfy the bound:

$$F_i(\beta; \mathbb{Q}) \geq \frac{|f^{|S|}(2Mb)|}{|f^{|S|}(0)|} \cdot F_{i,S}(\beta_S; \mathbb{Q}) \quad \text{for } i = 1, 2. \tag{93}$$

In above, the definition of $F_{i,S}(\beta_S; \mathbb{Q})$ is analogous to that of $F_S(\beta_S; \mathbb{Q})$:

$$\begin{aligned} F_{1,S}(\beta; \mathbb{Q}) &= f'(\infty) \cdot \mathbb{E}_{B-W} \left[\left\| X_S - X'_S \right\|_{1,\beta_S} \right]. \\ F_{2,S}(\beta; \mathbb{Q}) &= - \int \mathbb{E}_{B-W} \left[e^{-t \| X_S - X'_S \|_{1,\beta_S}} \right] \mu(dt). \end{aligned}$$

Showing that $F_1(\beta; \mathbb{Q})$ satisfies equation (93) is simple. A direct computation shows that for any $\beta \in \mathbb{R}_+^p$:

$$F_1(\beta; \mathbb{Q}) \geq F_{1,S}(\beta_S; \mathbb{Q}) \geq \frac{|f^{|S|}(2Mb)|}{|f^{|S|}(0)|} \cdot F_{1,S}(\beta_S; \mathbb{Q}). \tag{94}$$

Showing that $F_2(\beta;\mathbb{Q})$ satisfies equation (93) requires a little bit more thinking. We start with the following identity

$$\begin{aligned}
&\mathbb{E}_{B-W}\left[e^{-t\|X-X'\|_{1,\beta}}\right] \\
&= \mathbb{E}_{B-W}\left[e^{-t\left\|X_S-X'_S\right\|_{1,\beta_S}} \cdot e^{-t\left\|X_{S^c}-X'_{S^c}\right\|_{1,\beta_S^c}}\right] \\
&= \mathbb{E}_{B-W}\left[\mathbb{E}\left[e^{-t\left\|X_S-X'_S\right\|_{1,\beta_S}} \mid Y, Y'\right] \cdot \mathbb{E}\left[e^{-t\left\|X_{S^c}-X'_{S^c}\right\|_{1,\beta_S^c}} \mid Y, Y'\right]\right] \\
&= \mathbb{E}_{B-W}\left[e^{-t\left\|X_S-X'_S\right\|_{1,\beta_S}}\right] \cdot \mathbb{E}\left[e^{-t\left\|X_{S^c}-X'_{S^c}\right\|_{1,\beta_S^c}}\right].
\end{aligned}$$

where in the second identity, we use the fact that $X_S \perp X_{S^c} \mid Y$ (since $X_S \perp X_{S^c}$ and $Y \mid X = Y \mid X_{S^c}$), and in the last identity, we use the fact that $X_{S^c} \perp Y$. Substituting the expression into equation (92), we obtain

$$F_2(\beta;\mathbb{Q}) = -\int \mathbb{E}\left[e^{-t\left\|X_{S^c}-X'_{S^c}\right\|_{1,\beta_S^c}}\right] \cdot \mathbb{E}_{B-W}\left[e^{-t\left\|X_S-X'_S\right\|_{1,\beta_S}}\right] \mu(dt) \tag{95}$$

Below we show how equation (95) implies that $F_2(\beta;\mathbb{Q})$ satisfies the equation (93). The key is to decouple the two integrands in equation (95) using the following covariance inequality: for any function g_1, g_2 that is monotonically decreasing, and any non-negative measure $\tilde{\mu}$, we have

$$\int g_1(t)g_2(t)\tilde{\mu}(dt) \geq \frac{1}{|\tilde{\mu}|}\int g_1(t)\tilde{\mu}(dt)\int g_2(t)\tilde{\mu}(dt)$$

We apply the covariance inequality to appropriately chosen functions g_1, g_2 and measure $\tilde{\mu}$. We first choose the function $g_1(t)$ to be

$$g_1(t) = \mathbb{E}\left[e^{-t\left\|X_{S^c}-X'_{S^c}\right\|_{1,\beta_S^c}}\right]. \tag{96}$$

It is clear that $t \to g_1(t)$ is monotonically decreasing. Next we introduce $g_2(t)$. By Lemma H.2, we have the identity

$$\begin{aligned}
-\mathbb{E}_{B-W}\left[e^{-t\left\|X_S-X'_S\right\|_{1,\beta_S}}\right] &= \int |\phi_{0,S}(t\omega_S) - \phi_{1,S}(t\omega_S)|^2 \cdot \prod_{k\in S} Q_{\beta_k}(\omega_k)d\omega_S \\
&= \int |\phi_{0,S}(\omega_S) - \phi_{1,S}(\omega_S)|^2 \cdot \prod_{k\in S} Q_{t\beta_k}(\omega_k)d\omega_S.
\end{aligned}$$

where the second identity follows from the change of variables. Let $g_2(t)$ be

$$g_2(t) = \int |\phi_{0,S}(\omega_S) - \phi_{1,S}(\omega_S)|^2 \cdot \prod_{k\in S}\left(\frac{1}{t}\cdot Q_{t\beta_k}(\omega_k)\right)d\omega_S.$$

Then $t \to g_2(t)$ is monotonically decreasing since $t \to \frac{1}{t} \cdot Q_{t\beta}(\omega) = \frac{1}{\pi} \cdot \frac{\beta}{\omega^2 + t^2\beta^2}$ is monotonically decreasing. In addition, we have the identity

$$-\mathbb{E}_{B-W}\left[e^{-t\left\|X_S - X_S'\right\|_{1,\beta_S}}\right] = t^{|S|} g_2(t). \tag{97}$$

Finally, we choose the non-negative measure $\tilde{\mu}$ by $d\tilde{\mu} = t^{|S|} d\mu$. Now, we apply the covariance inequality to the functions g_1, g_2 and the measure $\tilde{\mu}$, and since equations (95), (96) and (97), we obtain

$$F_2(\beta; \mathbb{Q}) = \int g_1(t) g_2(t) \tilde{\mu}(dt) \geq \frac{1}{|\tilde{\mu}|} \int g_1(t) \tilde{\mu}(dt) \cdot \int g_2(t) \tilde{\mu}(dt). \tag{98}$$

Now we evaluate the terms on the RHS. First, by Lemma H.2, we have

$$|\tilde{\mu}| = \int t^{|S|} \mu(dt) = |f^{|S|}(0)| \tag{99}$$

Next, we have the lower bound

$$\begin{aligned} \int g_1(t) \tilde{\mu}(dt) &= \int \mathbb{E}\left[e^{-t\left\|X_S - X_S'\right\|_{1,\beta_S}}\right] t^{|S|} \mu(dt) \\ &= \mathbb{E}\left[f^{|S|}\left(\left\|X_S - X_S'\right\|_{1,\beta_S}\right)\right] \geq |f^{|S|}(2Mb)|, \end{aligned} \tag{100}$$

where the second step uses Lemma H.2, and the last step utilizes the fact that f is strictly completely monotone, and that $\|X_S - X_S'\|_{1,\beta_S} \leq 2Mb$ since $\beta \in \mathcal{B}$. Finally, we have the identity

$$\begin{aligned} \int g_2(t) \tilde{\mu}(dt) &= \int t^{|S|} g_2(t) \mu(dt) \\ &= -\int \mathbb{E}_{B-W}\left[e^{-t\left\|X_S - X_S'\right\|_{1,\beta_S}}\right] \mu(dt) = F_{2,S}(\beta_S; \mathbb{Q}). \end{aligned} \tag{101}$$

Substituting equations (99), (100) and (101) into equation (98) yields

$$F_2(\beta; \mathbb{Q}) \geq \frac{|f^{|S|}(2Mb)|}{|f^{|S|}(0)|} \cdot F_{2,S}(\beta_S; \mathbb{Q}).$$

This proves that $F_2(\beta; \mathbb{Q})$ satisfies the equation (93). The proof of Lemma I.3 is thus complete.

I.6. Proof of Lemma I.4. The proof starts from the identity:

$$\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) = \lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \cdot \left(F(\beta_\varepsilon; \mathbb{Q}) - F(\beta; \mathbb{Q}) \right). \tag{102}$$

Now we evaluate the RHS. First, since $\operatorname{supp}(\beta) \subseteq A$ and $Y \perp X_A$, we have

$$F(\beta; \mathbb{Q}) = \mathbb{E}_{B-W}\left[f(\left\| X - X' \right\|_{1,\beta}) \right] = \mathbb{E}_{B-W}\left[f(\left\| X_A - X'_A \right\|_{1,\beta_A}) \right] = 0 \tag{103}$$

Next, Lemma I.1 shows when $0 < \varepsilon \leq \tau$,

$$F(\beta_\varepsilon; \mathbb{Q}) \geq \frac{\varepsilon}{\tau} \cdot F(\beta_\tau; \mathbb{Q}). \tag{104}$$

Substitute equations (103) and (104) into equation (102). We obtain

$$\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) \geq \frac{1}{\tau} \cdot F(\beta_\tau; \mathbb{Q}).$$

I.7. Proof of Lemma I.5. Note that $Y \perp X_{S^c}$ and $X_S \perp X_{S^c} \mid Y$ (since $X_S \perp X_{S^c}$ and $Y \mid X = Y \mid X_S$ by assumption). Hence, we have

$$\begin{aligned} |\phi_0(\omega) - \phi_1(\omega)| &= \left| \mathbb{E}_0 e^{i\langle \omega, X \rangle} - \mathbb{E}_1 e^{i\langle \omega, X \rangle} \right| \\ &= \left| \mathbb{E} e^{i\langle \omega_{S^c}, X_{S^c} \rangle} \cdot (\mathbb{E}_0 - \mathbb{E}_1) e^{i\langle \omega_S, X_S \rangle} \right| \leq \left| (\mathbb{E}_0 - \mathbb{E}_1) e^{i\langle \omega_S, X_S \rangle} \right|. \end{aligned} \tag{105}$$

Now, we bound the RHS. The key idea is the identity below:

$$\begin{aligned} e^{i\langle \omega_S, X_S \rangle} =& R(X; \omega) + \sum_{A \subsetneq S} (-1)^{|S-A|-1} e^{i\langle \omega_A, X_A \rangle} \\ \text{where} \quad & R(X; \omega) = \prod_{j \in S} \left(e^{i\omega_j X_j} - 1 \right). \end{aligned} \tag{106}$$

The identity is obtained by simply multiplying out the terms in $R(X; \omega)$. By assumption, X_A has the same distribution under P_0 and P_1 whenever $A \subsetneq S$. So we have for $A \subsetneq S$,

$$\mathbb{E}_0\left[e^{i\langle \omega_A, X_A \rangle} \right] = \mathbb{E}_1\left[e^{i\langle \omega_A, X_A \rangle} \right]. \tag{107}$$

Now, in view of Eq (106) and Eq (107), we have the identity:

$$\mathbb{E}_0\left[e^{i\langle \omega_S, X_S \rangle} \right] - \mathbb{E}_1\left[e^{i\langle \omega_S, X_S \rangle} \right] = \mathbb{E}_0\left[R(X; \omega) \right] - \mathbb{E}_1\left[R(X; \omega) \right], \tag{108}$$

i.e. all terms of the form $e^{i\langle \omega_A, X_A\rangle}$ cancel in the difference for $A \subsetneq S$. Now, note the following elementary inequality that holds for any $x \in \mathbb{R}$,

$$|e^{ix} - 1| \le |x| \wedge 2.$$

Since $|X|_\infty \le M$ by assumption, we have for all ω,

$$|R(X;\omega)| \le \prod_{j\in S}(|\omega_j X_j| \wedge 2) \le \prod_{j\in S}(|M\omega_j| \wedge 2). \tag{109}$$

As a direct consequence of equations (105), (108) and (109), we obtain

$$|\phi_0(\omega) - \phi_1(\omega)| \le |\mathbb{E}_0[R(X;\omega)] - \mathbb{E}_1[R(X;\omega)]| \le 2\prod_{j\in S}(|M\omega_j| \wedge 2).$$

This completes the proof of Lemma I.5.

I.8. Proof of Lemma I.6. First, we prove the first inequality.

$$\begin{aligned}
&\int (\alpha\,|\omega| \wedge 2)^2 \cdot Q_\beta(\omega)d\omega = 2\cdot\int_0^\infty (\alpha\,|\omega| \wedge 2)^2 \cdot Q_\beta(\omega)d\omega \\
&= 2\int_0^{2/\alpha} \alpha^2\omega^2 \cdot \frac{\beta}{\beta^2+\omega^2}d\omega + 8\int_{2/\alpha}^\infty \frac{\beta}{\beta^2+\omega^2}d\omega \\
&\le 2\int_0^{2/\alpha} \alpha^2\beta\; d\omega + 8\int_{2/\alpha}^\infty \frac{\beta}{\omega^2}d\omega = 8\alpha\beta.
\end{aligned} \tag{110}$$

Next, we prove the second inequality. Note

$$\begin{aligned}
&\int (\alpha\,|\omega| \wedge 2)^2 \cdot Q_\beta^2(\omega)d\omega = 2\cdot\int_0^\infty (\alpha\,|\omega| \wedge 2)^2 \cdot Q_\beta^2(\omega)d\omega \\
&\le 2\int_0^{\beta} \alpha^2\omega^2 \cdot \frac{\beta^2}{(\omega^2+\beta^2)^2}d\omega + 2\int_\beta^\infty \alpha^2\omega^2 \cdot \frac{\beta^2}{(\omega^2+\beta^2)^2}d\omega \\
&\le 2\int_0^{\beta} \alpha^2 d\omega + 2\alpha^2\int_\beta^\infty \frac{\beta^2}{\omega^2}d\omega = 4\alpha^2\beta.
\end{aligned} \tag{111}$$

This completes the proof of Lemma I.6.

APPENDIX J: UNIFORM CONVERGENCE RESULTS

J.1. Main Results. In this section, we study the uniform convergence of the empirical objective $F(\beta;\mathbb{Q}_n)$ to population objective $F(\beta;\mathbb{Q})$ over the constraint set $\beta \in \mathcal{B}$, where we recall the definition $\mathcal{B} = \{\beta \in \mathbb{R}_+^p : \|\beta\|_1 \le b\}$. Here we restrict $\mathbb{Q}$ to be a reweighting distribution, i.e., $\mathbb{Q} = \mathbb{Q}^w$ for some

nonnegative weight function w, where formally $\mathbb{Q}^w$ is defined as the unique probability measure that satisfies $dQ^w(x,y) \propto dP(x,y) \cdot w(x,y)$. It is helpful for the reader to keep in mind, for the purpose of studying metric learning algorithm, the underlying distribution $\mathbb{Q}$ of interest is the reweighting distribution $\mathbb{Q} = \mathbb{Q}^A (= \mathbb{Q}^{w_A})$ where $A \subseteq [p]$.

The rest of the section is organized as follows.

- We start with a general result (Proposition 5) that establishes the uniform convergence of $F(\beta; \mathbb{Q}_n)$ to $F(\beta; \mathbb{Q})$ for a generic reweighting distribution $\mathbb{Q} = \mathbb{Q}^w$ that satisfies the assumption Q (see below for the definition of assumption Q). We wish that assumption Q clarifies the essential property for the distribution $\mathbb{Q}$ to satisfy the uniform convergence guarantee.
- We next show in Proposition 6 that any reweighting distribution $\mathbb{Q} = \mathbb{Q}^A$ where $A \subseteq [p]$ satisfies the assumption Q.
- Finally, we combine the above results to establish the uniform convergence of $F(\beta; \mathbb{Q}_n)$ to $F(\beta; \mathbb{Q})$ for any reweighting distribution $\mathbb{Q} = \mathbb{Q}^A$.

Here is the abstract assumption on the reweighting distribution $\mathbb{Q}$ that we need in order to establish the uniform convergence result.

ASSUMPTION Q. *Let $\mathbb{Q} = \mathbb{Q}^w$ be the (reweighting) distribution associated with some weight function $w(x,y)$. Assume the weight function w satisfies*

- *The weight function is bounded: $0 \le w(X,Y) \le 1$ under $\mathbb{P}$.*
- *The class label is balanced after reweighting:*

$$\mathbb{Q}(Y=0) = \mathbb{Q}(Y=1).$$

- *For some constant $\varrho > 0$, we have that*

$$\mathbb{E}_B[w(X,Y)w(X',Y')] \ge \varrho^2,$$
$$\mathbb{E}_W[w(X,Y)w(X',Y')] \ge \varrho^2.$$

Here the expectation $\mathbb{E}$ is taken under measure $\mathbb{P}$.

Notation. We wish to emphasize that throughout the section, the notation $\mathbb{E}$ always stands for the expectation under the common probability measure $\mathbb{P}$ (and not under the probability measure $\mathbb{Q}$!). The reason is that we wish to state and prove our results under a common measure $\mathbb{P}$ so that we can easily relate results for different measures $\mathbb{Q}', \mathbb{Q}'', \ldots$. The way to denote the expectation under a reweighting disribution $\mathbb{Q} = \mathbb{Q}^w$ is to use the notation $\mathbb{E}^w$, where w is the weight function associated with $\mathbb{Q}$.

Proposition 5 is the main result of the section, which establishes a uniform convergence result for a generic distribution $\mathbb{Q}$ satisfying assumption Q.

PROPOSITION 5. *Assume Assumptions (A1)-(A2). Then we have for some constant $C > 0$ that depends only on $b, M, \varrho, f(0), f'(0), f''(0)$ such that for any probability distribution $\mathbb{Q}$ satisfying Assumption Q, we have*

1. *For any $t > 0$, with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$ under $\mathbb{P}$,*

$$\sup_{\beta \in \mathcal{B}} |F(\beta; \mathbb{Q}_n) - F(\beta; \mathbb{Q})| \leq C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t).$$

2. *For any $t > 0$, with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$ under $\mathbb{P}$,*

$$\sup_{j \in [p]} \sup_{\beta \in \mathcal{B}} \left| \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}_n) - \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) \right| \leq C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t).$$

Proposition 5 provides a uniform convergence result for a generic reweighting distribution $\mathbb{Q}$ that satisfies Assumption Q. Below we study the uniform convergence for some specific choices of $\mathbb{Q}$. For any subset $A \subseteq [p]$, recall the reweighing distribution $\mathbb{Q}^A$ where $dQ^A(x, y) \propto dP(x, y) \cdot w_A(x, y)$ and where the weighting function w_A is defined by

$$w_A(x, y) = 1 - \mathbb{P}(Y = y \mid X_A = x_A).$$

Proposition 6 shows that the reweighting distribution $\mathbb{Q}^A$ satisfies Assumption Q under Assumption (A3). We defer the proof to Section J.3.

PROPOSITION 6. *Assume Assumption (A3). Then for any set $A \subseteq [p]$, the reweighting function $w_A(x, y) = 1 - \mathbb{P}(Y = y | X_A = x)$ satisfies*

1. *The weight function is bounded: $0 \leq w_A(X, Y) \leq 1$ under $\mathbb{P}$.*
2. *The class label is balanced after reweighting:*

$$\mathbb{Q}^A(Y = 0) = \mathbb{Q}^A(Y = 1)$$

3. *For the constant $\varrho > 0$ in the statement of Assumption (A3), we have*

$$\begin{aligned} &\mathbb{E}_B \left[w_A(X, Y) \, w_A\left(X', Y'\right) \right] \geq \varrho^2, \\ &\mathbb{E}_W \left[w_A(X, Y) \, w_A\left(X', Y'\right) \right] \geq \varrho^2. \end{aligned} \tag{112}$$

In above, the expectation $\mathbb{E}$ is taken under the probability measure $\mathbb{P}$.

Proposition 5 and Proposition 6 immediately yield Corollary J.1 below.

COROLLARY J.1. *Assume Assumptions (A1)-(A3). There exists a constant $C > 0$ that depends only on $b, M, \varrho, f(0), f'(0), f''(0)$ such that for any subset $A \subseteq [p]$, we have*

1. *For any $t > 0$, with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$ under $\mathbb{P}$,*

$$\sup_{\beta \in \mathcal{B}} \left| F(\beta; \mathbb{Q}_n^A) - F(\beta; \mathbb{Q}^A) \right| \leq C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t).$$

2. *For any $t > 0$, with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$ under $\mathbb{P}$,*

$$\sup_{j \in [p]} \sup_{\beta \in \mathcal{B}} \left| \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}_n^A) - \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}^A) \right| \leq C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t).$$

J.2. Proof of Proposition 5. Let $\mathbb{Q}$ be any probability measure that satisfies Assumption Q. Let $w(x, y) \propto \frac{dQ}{dP}(x, y)$ denote the weight function associated with $\mathbb{Q}$ so that $\mathbb{Q} = \mathbb{Q}^w$ and that weight function w satisfies the conditions in the Assumption Q.

To start with, we compute

$$\begin{aligned} |F(\beta; \mathbb{Q}_n) - F(\beta; \mathbb{Q})| &= \left| (\hat{\mathbb{E}}_{n,B-W}^w - \mathbb{E}_{B-W}^w)[f(\langle \beta, \mathbf{d} \rangle)] \right| \\ \left| \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}_n) - \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) \right| &= \left| (\hat{\mathbb{E}}_{n,B-W}^w - \mathbb{E}_{B-W}^w)[\mathbf{d}_j \cdot f'(\langle \beta, \mathbf{d} \rangle)] \right| \end{aligned}$$

By triangle inequality, we obtain

$$\begin{aligned} |F(\beta; \mathbb{Q}_n) - F(\beta; \mathbb{Q})| &\leq \bar{\varepsilon}_{n,B}(\beta) + \bar{\varepsilon}_{n,W}(\beta) \\ \left| \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}_n) - \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) \right| &\leq \varepsilon_{n,B,j}(\beta) + \varepsilon_{n,W,j}(\beta) \end{aligned} \tag{113}$$

where we define the empirical deviations

$$\begin{aligned} \bar{\varepsilon}_{n,B}(\beta) &= \left| (\hat{\mathbb{E}}_{n,B}^w - \mathbb{E}_B^w)[f(\langle \beta, \mathbf{d} \rangle)] \right| \\ \bar{\varepsilon}_{n,W}(\beta) &= \left| (\hat{\mathbb{E}}_{n,W}^w - \mathbb{E}_W^w)[f(\langle \beta, \mathbf{d} \rangle)] \right| \\ \varepsilon_{n,B,j}(\beta) &= \left| (\hat{\mathbb{E}}_{n,B}^w - \mathbb{E}_B^w) \left[\mathbf{d}_j \cdot f'(\langle \beta, \mathbf{d} \rangle) \right] \right|. \\ \varepsilon_{n,W,j}(\beta) &= \left| (\hat{\mathbb{E}}_{n,W}^w - \mathbb{E}_W^w) \left[\mathbf{d}_j \cdot f'(\langle \beta, \mathbf{d} \rangle) \right] \right|. \end{aligned} \tag{114}$$

Following equation (113), the key to the proof is to provide high probability upper bounds onto the following (random) quantities

$$\sup_{\beta\in\mathcal{B}} \bar{\varepsilon}_{n,B}(\beta), \quad \sup_{\beta\in\mathcal{B}} \bar{\varepsilon}_{n,W}(\beta), \quad \sup_{j\in[p]}\sup_{\beta\in\mathcal{B}} \varepsilon_{n,B,j}(\beta), \quad \sup_{j\in[p]}\sup_{\beta\in\mathcal{B}} \varepsilon_{n,W,j}(\beta).$$

Lemma J.1 does this technical work, whose proof, which is based on the empirical process theory, is deferred to Section J.2.1.

LEMMA J.1. *Assume Assumptions (A1)–(A2). Assume $w(x,y)$ satisfies*

- *The weight function is $0 \le w(X,Y) \le 1$.*
- *For some constant $\varrho > 0$, we have that*

$$\mathbb{E}_B[w(X,Y)w(X',Y')] \ge \varrho^2,$$
$$\mathbb{E}_W[w(X,Y)w(X',Y')] \ge \varrho^2.$$

Let g, h be two functions that satisfy

- *$x \mapsto g(x)$ is Lipschitz with Lipschitz constant L.*
- *$x \mapsto |g(x)|$ is upper bounded by G for all $x \in [0, 2Mb]$.*
- *$x \mapsto |h(x)|$ is upper bounded by H for all $x \in [0, M]$.*

Consider the two random functions:

$$\begin{aligned} \delta_{n,W,j}(\beta) &= \left|(\hat{\mathbb{E}}^w_{n,W} - \mathbb{E}^w_W)\left[g(\langle\beta, \mathbf{d}\rangle)\cdot h(\mathbf{d}_j)\right]\right| \\ \delta_{n,B,j}(\beta) &= \left|(\hat{\mathbb{E}}^w_{n,W} - \mathbb{E}^w_W)\left[g(\langle\beta, \mathbf{d}\rangle)\cdot h(\mathbf{d}_j)\right]\right| \end{aligned} \tag{115}$$

There exists some constant $C > 0$ that depends only on G, H, L, b, M, ϱ such that we have with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$,

$$\sup_{\beta\in\mathcal{B}} \left(\delta_{n,B,j}(\beta) + \delta_{n,W,j}(\beta)\right) \le C\cdot\sqrt{\frac{\log p}{n}}\cdot(1+t). \tag{116}$$

We apply Lemma J.1 to two specific groups of choice of (g, h).

- We first specify $g(x) = f(x)$ and $h(x) \equiv 1$. Note that f is $f'(0)$ Lipschitz since f' is completely monotone. Lemma J.1 implies with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$ under $\mathbb{P}$,

$$\sup_{\beta\in\mathcal{B}} \left(\bar{\varepsilon}_{n,B}(\beta) + \bar{\varepsilon}_{n,W}(\beta)\right) \le C\cdot\sqrt{\frac{\log p}{n}}\cdot(1+t),$$

  where $C > 0$ is a constant that depends only on $b, M, \varrho, f(0), f'(0)$.

- We next specify $g(x) = f'(x)$ and $h(x) = x$. Note that f' is $|f''(0)|$ Lipschitz since f' is completely monotone. Lemma J.1 implies with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$ under $\mathbb{P}$,

$$\sup_{j\in[p]}\sup_{\beta\in\mathcal{B}}\left(\varepsilon_{n,B,j}\left(\beta\right)+\varepsilon_{n,W,j}\left(\beta\right)\right)\leq C\cdot\sqrt{\frac{\log p}{n}}\cdot(1+t).$$

  where $C > 0$ is a constant that depends only on $b, M, \varrho, f'(0), f''(0)$.

Proposition 5 now follows straightforwardly from the previous discussions.

J.2.1. *Proof of Lemma J.1.* For simplicity, we slightly modify the definition of the empirical average $\hat{\mathbb{E}}^w_{n,B}$ and $\hat{\mathbb{E}}^w_{n,W}$ in the proof below. Concretely, we define for any function $h(x, x', y, y')$

(117)
$$\hat{\mathbb{E}}_{n,B}[h(X,X',Y,Y')] = \frac{\sum_{i\neq i'} w(X_i,Y_i)w(X_{i'},Y_{i'})h(X_i,X_{i'},Y_i,Y_{i'})\mathbf{1}_{Y_i\neq Y_{i'}}}{\sum_{i\neq i'} w(X_i,Y_i)w(X_{i'},Y_{i'})\mathbf{1}_{Y_i\neq Y_{i'}}}$$
$$\hat{\mathbb{E}}_{n,W}[h(X,X',Y,Y')] = \frac{\sum_{i\neq i'} w(X_i,Y_i)w(X_{i'},Y_{i'})h(X_i,X_{i'},Y_i,Y_{i'})\mathbf{1}_{Y_i=Y_{i'}}}{\sum_{i\neq i'} w(X_i,Y_i)w(X_{i'},Y_{i'})\mathbf{1}_{Y_i=Y_{i'}}}.$$

The only difference between the new definition above and the original definition in the main text is that, while the original definition sums over all possible tuples (i, i') on the RHS where $1 \leq i, i' \leq n$, the new definition only sums over all possible *distinct* tuple (i, i') where $1 \leq i, i' \leq n$ and $i \neq i'$. One important claim is that such modification doesn't change qualitatively the uniform convergence result (and it can only change the numerical constants in the high probability bound). The reason why this is true is largely due to the fact that the weight $w(x, y)$ and the function $h(x, x', y, y') = g(\langle\beta, \mathbf{d}\rangle) \cdot h(\mathbf{d}_j)$ of interest is bounded—$0 \leq w(x, y) \leq 1$ and $|h(x, x', y, y')| \leq GH$ by assumption—so summing over all tuples (i, i') versus summing over all *distinct* tuples (i, i') do not make a real difference. The formal proof of this claim is tedious (simple, lengthy, but not insightful) and thus omitted.

In the proof below, we will work under this definition of $\hat{\mathbb{E}}^w_{n,B}$ and $\hat{\mathbb{E}}^w_{n,W}$, and still the random quantities of interest are defined by

(118)
$$\delta_{n,W,j}(\beta) = \left|\left(\hat{\mathbb{E}}^w_{n,W} - \mathbb{E}^w_W\right)\left[g(\langle\beta,\mathbf{d}\rangle)\cdot h(\mathbf{d}_j)\right]\right|$$
$$\delta_{n,B,j}(\beta) = \left|\left(\hat{\mathbb{E}}^w_{n,W} - \mathbb{E}^w_W\right)\left[g(\langle\beta,\mathbf{d}\rangle)\cdot h(\mathbf{d}_j)\right]\right|$$

We prove for any fixed $j \in [p]$, with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$,

$$\sup_{\beta \in \mathcal{B}} \delta_{n,B,j}(\beta) \leq C \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t). \tag{119}$$

holds for some large constant $C > 0$ depending on G, H, L, b, M, ϱ. An analogous high probability bound also holds for $\delta_{n,W,j}(\beta)$. Together, with a union bound, this proves Lemma J.1 as desired.

Now we show equation (119) holds with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$. Fix $j \in [p]$. We introduce notational shorthand to simplify the proof below. We use $Z_i = (X_i, Y_i)$ to denote the i.i.d data pair. Introduce

$$\begin{aligned} r_\beta(z_i, z_i') &= g\left(\langle \beta, \mathbf{d} \rangle\right) \cdot h(\mathbf{d}_j) \\ w(z_i, z_i') &= 1_{y_i \neq y_{i'}} w(z_i) w(z_{i'}) \\ q_\beta(z_i, z_i') &= r_\beta(z_i, z_i') w(z_i, z_i') \end{aligned}$$

where $w(z_i)$ is the weight. Let $\hat{\mathbb{E}}_n$ denote the empirical average over all the $n(n-1)$ distinct tuples (i_1, i_2). By definition of $\hat{\mathbb{E}}^w_{n,B}, \mathbb{E}^w_B$,

$$\hat{\mathbb{E}}^w_{n,B}[h_\beta(z, z')] = \frac{\hat{\mathbb{E}}_n[q_\beta(z,z')]}{\hat{\mathbb{E}}_n[w(z,z')]} \quad \text{and} \quad \mathbb{E}^w_B[h_\beta(z,z')] = \frac{\mathbb{E}[q_\beta(z,z')]}{\mathbb{E}[w(z,z')]}.$$

Hence, we have the identity

$$\sup_{\beta \in \mathcal{B}} \delta_{n,B,j}(\beta) = \sup_{\beta \in \mathcal{B}} \left| \frac{\hat{\mathbb{E}}_n[q_\beta(z,z')]}{\hat{\mathbb{E}}_n[w(z,z')]} - \frac{\mathbb{E}[q_\beta(z,z')]}{\mathbb{E}[w(z,z')]} \right|. \tag{120}$$

Now we use the elementary inequality: for $a, a', b, b' \in \mathbb{R}$,

$$\left| \frac{b}{a} - \frac{b'}{a'} \right| \leq \frac{1}{|aa'|} \left(|a - a'||b'| + |b - b'||a'| \right)$$

Applying this inequality to equation (120), we obtain the bound

$$\sup_{\beta \in \mathcal{B}} \delta_{n,B,j}(\beta) \leq \frac{1}{\Gamma_n} \left(\Delta_{1,n} + \Delta_{2,n} \right) \tag{121}$$

where

$$\begin{aligned} \Gamma_n &= \hat{\mathbb{E}}_n[w(z,z')] \cdot \mathbb{E}[w(z,z')] \\ \Delta_{1,n} &= \mathbb{E}[w(z,z')] \cdot \sup_{\beta \in \mathcal{B}} \left| (\hat{\mathbb{E}}_n - \mathbb{E})[q_\beta(z,z')] \right| \\ \Delta_{2,n} &= \left| (\hat{\mathbb{E}}_n - \mathbb{E})[w(z,z')] \right| \cdot \sup_{\beta \in \mathcal{B}} \mathbb{E}[q_\beta(z,z')] \end{aligned}$$

To obtain high probability upper bound on $\sup_{\beta\in\mathcal{B}}\delta_{n,B,j}(\beta)$, it suffices to derive a high probability lower bound on Γ_n and upper bound on $\Delta_{1,n}, \Delta_{2.n}$. The following three technical lemma are useful towards this end.

LEMMA J.2. *We have with probability at least* $1-e^{-4t^2}$*:*

$$\left|(\hat{\mathbb{E}}_n-\mathbb{E})[w(z,z')]\right| \le \sqrt{\frac{1}{n}}\cdot t.$$

PROOF. Note that the weights $0\le w\le 1$ by assumption. The lemma is a consequence of Hoeffding's inequality for U-statistics (see Lemma O.1). □

LEMMA J.3. *We have with probability at least* $1-e^{-n\varrho^2/32}$*:*

$$\hat{\mathbb{E}}_n[w(z,z')] \ge \frac{1}{2}\varrho.$$

PROOF. By assumption, $0\le w\le 1$ and $\mathbb{E}[w(z,z')]\ge\varrho$. The lemma is a consequence of Bernstein's inequality for U-statistics (see Lemma O.1). □

LEMMA J.4. *For any* $\beta\in\mathcal{B}$*, we have*

$$\sup_{z,z'}\left|q_\beta(z,z')\right| \le GH.$$

PROOF. Assumption (A2) gives $0\le \mathbf{d}_j\le 2M$ for all $j\in[p]$, and moreover gives $0\le\langle\beta,\mathbf{d}\rangle\le 2Mb$ for all $\beta\in\mathcal{B}$ by Hölder's inequality. Assumption Q gives $0\le w\le 1$. Thus, we have for all $\beta\in\mathcal{B}$,

$$\sup_{z,z'}|q_\beta(z,z')| \le GH.$$

This completes the proof of Lemma J.4. □

LEMMA J.5. *We have with probability at least* $1-p^{-t^2}$*:*

$$\sup_{\beta\in\mathcal{B}}\left|(\hat{\mathbb{E}}_n-\mathbb{E})[q_\beta(z,z')]\right| \le C\sqrt{\frac{\log p}{n}}\cdot(1+t).$$

In above C *is a constant depending only on* b, M, L, G, H*.*

We provide the proof of Lemma J.5 into Section J.2.2.

Now we give high probability upper bound on $\sup_{\beta\in\mathcal{B}}\delta_{n,B,j}(\beta)$ using equation (121). By Lemma J.2, Lemma J.4 and Lemma J.5, the bound below

$$\Delta_{1,n}\le C\cdot\sqrt{\frac{\log p}{n}}\cdot(1+t),\quad \Delta_{2,n}\le C\cdot\sqrt{\frac{\log p}{n}}t$$

holds with probability at least $1 - 2p^{-t^2}$, where $C > 0$ is some constant depending only on b, M, L, G, H. Recall Assumption Q: $\mathbb{E}[w(z, z')] \geq \varrho^2$. Thus Lemma J.3 further implies

$$\Gamma_n \geq \mathbb{E}[w(z, z')] \cdot \hat{\mathbb{E}}_n[w(z, z')] \geq \frac{1}{2}\varrho^4.$$

with probability at least $1 - e^{-n\varrho^2/32}$. Substitute the high probability bounds into equation (121). We get with probability at least $1 - 2p^{-t^2} - e^{-n\varrho^2/32}$,

$$\sup_{\beta \in \mathcal{B}} \delta_{n,B,j}(\beta) \leq C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t).$$

where $C > 0$ is a constant that depends on b, M, ϱ, L, G, H. This completes the proof of Lemma J.1.

J.2.2. *Proof of Lemma J.5.* Let's denote the random variable

$$W = \sup_{\beta \in \mathcal{B}} \left| (\hat{\mathbb{E}}_n - \mathbb{E})[q_\beta(z, z')] \right|.$$

We view $W \equiv W(Z_{1:n})$ as a function of the i.i.d data pair $Z_i = (X_i, Y_i)$. Lemma J.4 implies that this function is of bounded difference with bound $2GH/n$, i.e., for any $Z_{1:n}$ and $Z'_{1:n}$ differing in only one coordinate,

$$\left| W(Z_{1:n}) - W(Z'_{1:n}) \right| \leq \frac{2GH}{n}.$$

Thus, McDiarmid's bounded difference inequality [9] gives

$$W \leq \mathbb{E}[W] + 2GH \cdot \sqrt{\frac{\log p}{n}} \cdot t \tag{122}$$

holds with probability at least $1 - p^{-t^2}$.

Now, we bound $\mathbb{E}[W]$. On a high level, our idea works as follows: viewing W as the suprema of some empirical process, we symmetrize W and bound the resulting Rademacher complexity with the Ledoux-Talagrand contraction inequality [6]. Formally, we decompose

$$q_\beta(z, z') = \phi\left(\langle \beta, \mathbf{d} \rangle, z, z'\right) + \bar{\phi}\left(z, z'\right),$$

where $(z, z') \mapsto \bar{\phi}(z, z')$ and $(u, z, z') \mapsto \phi(u, z, z')$ are defined by

$$\begin{aligned} \bar{\phi}\left(z, z'\right) &= w(z, z') \cdot g(0) \cdot h(\mathbf{d}_j) \\ \phi(u, z, z') &= w(z, z') \cdot (g(u) - g(0)) \cdot h(\mathbf{d}_j). \end{aligned}$$

We emphasize that the definition of the two functions ϕ and $\bar{\phi}$ are independent of $\beta \in \mathcal{B}$. By triangle inequality, we can upper bound

$$\mathbb{E}[W] \leq \mathbb{E}[W_1] + \mathbb{E}[W_2] \tag{123}$$

where W_1 and W_2 are the supremum of some empirical process:

$$W_1 = \left|(\hat{\mathbb{E}}_n - \mathbb{E})[\bar{\phi}(z, z')]\right| \quad \text{and} \quad W_2 = \sup_{\beta \in \mathcal{B}} \left|(\hat{\mathbb{E}}_n - \mathbb{E})[\phi(\langle \beta, \mathbf{d} \rangle, z, z')]\right|.$$

Below we bound $\mathbb{E}[W_1]$ and $\mathbb{E}[W_2]$ separately. Our major technique is to use symmetrization argument followed by Hoeffding [5] and Ledoux-Talagrand contraction inequality [6]. As both W_1 and W_2 involve averages of *dependent* random variables, standard symmetrization argument does not immediately apply, for which reason we adapt a decoupling technique that is due to Hoeffding [5] to overcome this technical difficulty. We introduce the notation.

- Let $\sigma_{i,i'}$ be independent Rademacher random variables.
- Let $\bar{\phi}_{(i,i')} = \bar{\phi}(z_i, z_i')$ and $\phi_{(i,i')}(u) = \phi(u, z_i, z_i')$.
- Let $\mathcal{I} = \{(i, i') \mid i \neq i', 1 \leq i, i' \leq n\}$. A simple combinatorial argument shows that we can decompose $\mathcal{I} = \cup_{j=1}^{I} \mathcal{I}_j$ where $I \leq n$, $|\mathcal{I}_j| \geq \lfloor \frac{n}{2} \rfloor$, and where for any two different tuples $(i_1, i_2), (i_3, i_4) \in \mathcal{I}_j$ where $j \in [I]$, we have $i_k \neq i_l$ for $1 \leq k < l \leq 4$. For each $j \in [I]$, let $\hat{\mathbb{E}}_{n,j}$ denote the empirical average over the distinct tuples $(i_1, i_2) \in \mathcal{I}_j$.

Part 1: Bound on $\mathbb{E}[W_1]$. As $\mathcal{I} = \cup_{j=1}^{I} \mathcal{I}_j$, we have by triangle inequality

$$\mathbb{E}[W_1] \leq \frac{1}{I} \sum_{j=1}^{I} \mathbb{E}\left[\left|(\hat{\mathbb{E}}_{n,j} - \mathbb{E})[\bar{\phi}(z, z')]\right|\right] \leq \max_{j \in [I]} \mathbb{E}\left[\left|(\hat{\mathbb{E}}_{n,j} - \mathbb{E})[\bar{\phi}(z, z')]\right|\right] \tag{124}$$

Now for each $j \in [I]$, $\hat{\mathbb{E}}_{n,j}[\bar{\phi}(z, z')]$ is the average of independent random variables. Invoking the standard symmetrization argument, we obtain

$$\mathbb{E}[W_1] \leq 2 \max_{j \in [I]} \mathbb{E}\left\{\mathbb{E}_\sigma\left[\frac{1}{|\mathcal{I}_j|}\left|\sum_{(i,i') \in \mathcal{I}_j} \sigma_{i,i'} \bar{\phi}_{i,i'}\right| \mid Z\right]\right\}. \tag{125}$$

Note then $\sup_{i,i'} |\bar{\phi}_{i,i'}| \leq GH$ since we have by assumption $0 \leq w(z, z') \leq 1$, $|g(0)| \leq G$ and $\sup_{x \in [0,M]} |h(x)| \leq H$. As $\{\sigma_{i,i'}\}_{i \neq i'}$ are independent 1-subgaussian random variables, we obtain the bound for all j:

$$\mathbb{E}_\sigma\left[\frac{1}{|\mathcal{I}_j|}\left|\sum_{i \neq i'} \sigma_{i,i'} \bar{\phi}_{i,i'}\right| \mid Z\right] \leq GH \cdot \sqrt{\frac{1}{|\mathcal{I}_j|}} \leq 2GH \cdot \sqrt{\frac{1}{n}}. \tag{126}$$

Substituting equation (126) into equation (125) yields the final bound

$$\mathbb{E}[W_1] \leq 4GH \cdot \sqrt{\frac{1}{n}}. \tag{127}$$

Part 2: Bound on $\mathbb{E}[W_2]$. Similar to equation (124), we have

$$\mathbb{E}[W_2] \leq \max_{j\in[I]} \sup_{\beta\in\mathcal{B}} \left|(\hat{\mathbb{E}}_{n,j} - \mathbb{E})[\phi(\langle\beta, \mathbf{d}\rangle, z, z')]\right|.$$

Now for each $j \in I$, $\hat{\mathbb{E}}_{n,j}[\phi(\langle\beta, \mathbf{d}\rangle, z, z')]$ is the average of independent random variables. Invoking the standard symmetrization argument, we obtain

$$\mathbb{E}[W_2] \leq 2 \max_{j\in[I]} \mathbb{E}\left\{\mathbb{E}_\sigma\left[\sup_{\beta\in\mathcal{B}}\left|\frac{1}{|\mathcal{I}_j|}\sum_{(i,i')\in\mathcal{I}_j} \sigma_{i,i'} \cdot \phi_{i,i'}\left(\langle\beta, \mathbf{d}\rangle\right)\right| | Z\right]\right\}. \tag{128}$$

The inner expectation is over $\sigma_{i,i'}$ with the data Z (all data z_i) fixed while the outer expectation is over Z. Now for any fixed z_i, z_i', since we have

$$\left|\frac{d}{du}\phi(u, z_i, z_i')\right| = \left|w(z_i, z_i') \cdot h(\mathbf{d}_j) \cdot g'(u)\right| \leq LH.$$

the mapping $u \to \phi_{i,i'}(\cdot, z_i, z_i')$ is Lipschitz with constant LH. Note further that $\phi_{i,i'}(0, z_i, z_i') = 0$. Hence, conditional on $\mathbf{d}$, we may apply the Ledoux-Talagrand contraction [6] inequality to obtain the following upper bound

$$\begin{aligned}&\mathbb{E}_\sigma\left[\sup_{\beta\in\mathcal{B}}\left|\frac{1}{|\mathcal{I}_j|}\sum_{(i,i')\in\mathcal{I}_j} \sigma_{i,i'} \cdot \phi_{i,i'}\left(\langle\beta, \mathbf{d}\rangle\right)\right| | Z\right] \\ &\leq 2LH \cdot \mathbb{E}_\sigma\left[\sup_{\beta\in\mathcal{B}}\left|\frac{1}{|\mathcal{I}_j|}\sum_{(i,i')\in\mathcal{I}_j} \sigma_{i,i'} \langle\beta, \mathbf{d}\rangle\right| | Z\right].\end{aligned} \tag{129}$$

To further bound the RHS, we notice the two key facts

- $\{\sigma_{i,i'}\}_{i\neq i'}$ are mutually independent 1-subgaussian random variables.
- The ℓ_∞ bound on $\|\mathbf{d}\|_\infty \leq M$ which is due to Assumption (A2).

Using the above facts, we can easily obtain the further bound

$$
\begin{aligned}
&\mathbb{E}_\sigma\left[\sup_{\beta\in\mathcal{B}}\left|\frac{1}{|\mathcal{I}_j|}\sum_{(i,i')\in\mathcal{I}_j}\sigma_{i,i'}\left\langle\beta,\mathbf{d}\right\rangle\right||Z\right] \\
&=\mathbb{E}_\sigma\left[\sup_{\beta\in\mathbb{R}_+^p:\|\beta\|_1\le b}\left|\left\langle\beta,\frac{1}{|\mathcal{I}_j|}\sum_{(i,i')\in\mathcal{I}_j}\sigma_{i,i'}\mathbf{d}\right\rangle\right||Z\right] \qquad (130)\\
&=b\cdot\mathbb{E}_\sigma\left[\left|\frac{1}{|\mathcal{I}_j|}\sum_{(i,i')\in\mathcal{I}_j}\sigma_{i,i'}\cdot\mathbf{d}\right|_\infty|\mathbf{d}\right]\le 2Mb\sqrt{\frac{\log p}{|\mathcal{I}_j|}},
\end{aligned}
$$

Consequently, equations (128)–(130) yield the final bound

$$
\mathbb{E}[W_2]\le 4MbLH\cdot\sqrt{\frac{\log p}{n}}. \qquad (131)
$$

Summary. Back to equation (122). We use equations (123), (127) and (131) to conclude the following: for any $t>0$, with probability at least $1-p^{-t^2}$.

$$
Q\le 8(2MbLH+GH)\cdot\sqrt{\frac{\log p}{n}}\cdot(1+t) \qquad (132)
$$

J.3. Proof of Proposition 6. The first part of Proposition 6 is trivial. The second part of Proposition 6 follows from the identity

$$
\mathbb{Q}^A(Y=1)=\mathbb{E}[w_A(X,1)w_A(X,0))]=\mathbb{Q}^A(Y=0),
$$

and this identity can be proven easily by using the law of iterated expectation. Below, we prove the third part of Proposition 6. That is, under Assumption (A3), we have equation (112) holds.

Write $\pi(X_A)=\mathbb{P}(Y=1\mid X_A)$. Recall that $Y\mid X=Y\mid X_S$. The proof is based on a fundamental inequality that holds for any set A:

$$
\mathbb{E}[\pi(X_A)\mid Y=0]\ge\mathbb{E}[\pi(X_S)\mid Y=0]. \qquad (133)
$$

To obtain this result, we note first for any set A:

$$
\begin{aligned}
\mathbb{E}[\pi(X_A)\mid Y=0]&=\frac{1}{\mathbb{P}(Y=0)}\mathbb{E}[\pi(X_A)\mathbf{1}_{Y=0}]\\
&=\frac{1}{\mathbb{P}(Y=0)}\mathbb{E}[\pi(X_A)(1-\pi(X_A))] \qquad (134)\\
&=\frac{1}{\mathbb{P}(Y=0)}\mathbb{E}[\mathrm{Var}(Y\mid X_A)].
\end{aligned}
$$

Next, for any set A, we have

$$\text{Var}(Y \mid X_A) \geq \mathbb{E}[\text{Var}(Y \mid X_{S\cup(A\setminus S)}) \mid X_A] = \mathbb{E}[\text{Var}(Y \mid X_S) \mid X_A]. \tag{135}$$

where the inequality follows from the conditional variance decomposition and the equality from $Y|X = Y|X_S$. Substituting it into equation (134), we obtain the desired inequality (133)

$$\begin{aligned}\mathbb{E}[\pi(X_A) \mid Y = 0] &= \frac{1}{\mathbb{P}(Y=0)}\mathbb{E}[\text{Var}(Y \mid X_A)] \\ &\geq \frac{1}{\mathbb{P}(Y=0)}\mathbb{E}[\text{Var}(Y \mid X_S)] = \mathbb{E}[\pi(X_S) \mid Y = 0]\end{aligned}$$

In the same way, one can show for any set A,

$$\mathbb{E}[(1-\pi(X_A)) \mid Y = 1] \geq \mathbb{E}[(1-\pi(X_S)) \mid Y = 1]$$

This shows for any set A:

$$\begin{aligned}\mathbb{E}_B[w_A(X,Y)w_A(X',Y')] &= \mathbb{E}[\pi(X_A)(1-\pi(X'_A)) \mid Y = 0, Y' = 1] \\ &= \mathbb{E}[\pi(X_A) \mid Y = 0] \cdot \mathbb{E}[(1-\pi(X_A)) \mid Y = 1] \\ &\geq \mathbb{E}[\pi(X_S) \mid Y = 0] \cdot \mathbb{E}[(1-\pi(X_S)) \mid Y = 1] \\ &\geq \varrho^2,\end{aligned}$$

where the first inequality uses our previous results. One can prove analogously the bound $\mathbb{E}_W[w_A(X,Y)w_A(X',Y')] \geq \varrho^2$.

APPENDIX K: PROOF OF NO FALSE DISCOVERY IN LOW DIMENSION

This section presents the proof of Theorem 2. The fundamental mathematical tool that underlies the proof of Theorem 2 is Proposition 4 (for false positive control) and Proposition 5 (for recovery). Proposition 4 basically shows that the gradient with respect to a noise variable X_j where $j \in S^c$ is negative, and its magnitude is lower bounded by the (square of the) objective value. This causes the *self-penalization*—the larger the objective value, the stronger the penalization on the noise variable. The recovery guarantee is a careful extension of the population recovery result in Proposition 5.

K.1. Notation. Consider the metric learning algorithm. Let $\mathbb{Q}^{(1)}, \mathbb{Q}^{(2)}, \ldots, \mathbb{Q}^{(k)}, \ldots$ denote the sequence of the weighting distribution, and $\hat{S}^{(1)}, \hat{S}^{(2)}, \ldots, \hat{S}^{(k)}, \ldots$ denote the set of variables selected by the algorithm through the

iterations. By convention, we define $\hat{S}^{(0)} = \emptyset$. Note then $\hat{S} = \bigcup_k \hat{S}^{(k)}$ is the final output of the algorithm.

On population, we use $\{\beta^{(m;\mathbb{Q}^{(k)})}\}_{m\in\mathbb{N}}$ to denote the inner-loop projected gradient ascent iterates that solve the population maximization problem

$$\max_{\beta\in\mathcal{B}} F(\beta;\mathbb{Q}^{(k)})$$

We use $\beta^{(*;\mathbb{Q}^{(k)})}$ to denote the accumulation point that's returned from the gradient ascent inner-loop. This means in particular on population

$$\hat{S}^{(k+1)} = \hat{S}^{(k)} \cup \operatorname{supp}(\beta^{(*;\mathbb{Q}^{(k)})}).$$

In finite case, we use $\{\beta^{(m;\mathbb{Q}_n^{(k)})}\}_{m\in\mathbb{N}}$ to denote the inner-loop projected gradient ascent iterates that solve the empirical maximization problem

$$\max_{\beta\in\mathcal{B}} F(\beta;\mathbb{Q}_n^{(k)})$$

We use $\beta^{(*;\mathbb{Q}_n^{(k)})}$ to denote the accumulation point that's returned from the gradient ascent inner-loop. This means in particular in finite case

$$\hat{S}^{(k+1)} = \hat{S}^{(k)} \cup \operatorname{supp}(\beta^{(*;\mathbb{Q}_n^{(k)})}).$$

We sometimes drop the dependence of the gradient ascent iterates on the probability measure $\mathbb{Q}$. This means that we may refer $\beta^{(m)}$ to $\beta^{(m;\mathbb{Q}^{(k)})}$, and refer $\beta^{(*)}$ to $\beta^{(*;\mathbb{Q}^{(k)})}$ when the context is clear.

K.2. Proof of Theorem 2. The key to the proof is the following proposition, whose proof is deferred into Section K.3.

PROPOSITION 7. *Let $A \subseteq S$. Consider the optimization problem:*

$$\max_{\beta:\beta\in\mathcal{B}} F(\beta;\mathbb{Q}_n^A). \tag{136}$$

Let $\{\beta^{(m)}\}_{m\in\mathbb{N}}$ denote the projected gradient ascent iterates with initialization at $\beta^{(0)}$ and stepsize α. Then there exists some constant $C > 0$ that depends only on $b, M, q, \varrho, f(0), f'(0), f''(0)$ such that the following holds: for any $t > 0$, any choice of the stepsize $\alpha \le \frac{1}{C\cdot p}$, and any choice of the threshold

$$\gamma > C \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t), \tag{137}$$

we have with probability at least $1 - 2(p^{-t^2} + e^{-n\varrho^2/32})$ such that at least one of the following two happens:

- *the initialization condition* $(F(\beta^{(0)}; \mathbb{Q}_n^A))^2 > \gamma$ *fails*
- *any accumulation point* β^* *of the projected gradient iterates satisfies*

$$\text{supp}(\beta^{(*)}) \subseteq S. \tag{138}$$

We are now ready to show that $\hat{S} \subseteq S$ holds with high probability. Fix $t > 0$ in the statement of Theorem 2. Let $\mathcal{E}_A$ denote the event that's stated in Proposition 7. Let $\mathcal{E} = \cap_{A:A\subseteq S}\mathcal{E}_A$ denote the event where all events $\mathcal{E}_A$ where $A \subseteq S$ happens. By Proposition 7 and the union bound, we understand with appropriate choice of constant $C > 0$ in the definition of $\mathcal{E}$,

$$\mathbb{P}(\mathcal{E}) \geq 1 - 2^{s+1}(p^{-t^2} + e^{-n\varrho^2}). \tag{139}$$

Below, we show on the event $\mathcal{E}$ that the following happens:

$$\hat{S}^{(k)} \subseteq S \ \text{ holds for all } k \in \mathbb{N}. \tag{140}$$

Note we define by convention $\hat{S}^{(k)} = \hat{S}$ for any $k > T$ where T is the total number of iterations after which the algorithm halts. Our proof is based on induction on $m \in \mathbb{N}$. Below we assume we are on the event $\mathcal{E}$.

- The base case where $m = 0$ trivially holds since $\hat{S}_0 = \emptyset$.
- Assume the induction hypothesis holds for $m = k$, i.e., $\hat{S}^{(k)} \subseteq S$. Consider the case where $m = k + 1$. We aim to show that $\hat{S}^{(k+1)} \subseteq S$. Below we divide our discussion into two cases:
  (i) If the initialization condition $(F(\beta^{(0)}; \mathbb{Q}_n^{(k)}))^2 > \gamma$ holds, then we run the gradient ascent to solve the maximization problem
  $$\max_{\beta\in\mathcal{B}} F(\beta; \mathbb{Q}_n^{(k)}).$$
  By definition of the event $\mathcal{E}$, we know that the output from the gradient ascent algorithm must satisfy
  $$\text{supp}\left(\beta^{(*;\mathbb{Q}_n^{(k)})}\right) \subseteq S.$$
  Consequently, this shows that, the output $\hat{S}^{(k+1)}$ satisfies
  $$\hat{S}^{(k+1)} = \hat{S}^{(k)} \cup \text{supp}\left(\beta^{(*;\mathbb{Q}_n^{(k)})}\right) \subseteq S.$$
  (ii) If the initialization condition $(F(\beta^{(0)}; \mathbb{Q}_n^{(k)}))^2 > \gamma$ fails, then the algorithm halts at this iteration. As a result, we obtain
  $$\hat{S}^{(k+1)} = \hat{S} \subseteq S.$$

Summarizing the above discussions, we have shown that $\hat{S}^{(k+1)} \subseteq S$. This proves that the hypothesis holds for $m = k + 1$.

As such, we have shown that $\hat{S} \subseteq S$ on the event $\mathcal{E}$. This happens with probability at least $1 - 2^{s+1}(p^{-t^2} + e^{-n\varrho^2/32})$, thanks to equation (167).

K.3. Proof of Proposition 7. The key to the proof is to establish the following result.

CLAIM 2. *There exists some constant $C > 0$ that depends only on the parameters $b, M, q, \varrho, f(0), f'(0), f''(0)$ such that the following holds: for any $t > 0$, for any choice of the stepsize $\alpha \leq \frac{1}{C \cdot p}$, and any threshold γ satisfying*

$$\gamma \geq C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t), \tag{141}$$

we have for some constant $\varepsilon > 0$ independent of m (but can be dependent on the rest of the parameters, e.g., p, n, α, C) such that, with probability at least $1 - 2(p^{-t^2} + e^{-n\varrho^2/32})$, the gradient ascent iterate satisfies

$$\beta_j^{(m+1)} \leq \left(\beta_j^{(m)} - \varepsilon\right)_+ \quad \text{for all } j \in S^c \text{ and all } m \in \mathbb{N} \tag{142}$$

as long as the initialization condition $(F(\beta^{(0)}; \mathbb{Q}_n^A))^2 > \gamma$ holds.

Once we can show this, then it is immediate that there exists a constant $C > 0$ such that for any $t > 0$, any threshold γ satisfying equation (141), one of the following two events must happen

- The initialization condition $(F(\beta^{(0)}; \mathbb{Q}_n^A))^2 > \gamma$ fails
- The initialization condition $(F(\beta^{(0)}; \mathbb{Q}_n^A))^2 > \gamma$ holds. Then by the claim 2, we know with probability at least $1 - 2(p^{-t^2} + e^{-n\varrho^2/32})$ any accumulation point $\beta^{(*)}$ of the iterate $\{\beta^{(m)}\}_{m \in \mathbb{N}}$ must satisfy

$$\operatorname{supp}(\beta^{(*)}) \subseteq S. \tag{143}$$

Below we prove Claim 2. Our strategy is to first prove that the Claim holds on population ($n = \infty$), and then extend the result to finite sample case ($n < \infty$) by standard techniques on concentration and perturbation.

Population Analysis $n = \infty$. Below we prove that equation (142) holds with probability one for some constant $\varepsilon > 0$ when the constant $C > 0$ (stated in Claim 2) is sufficiently large. The key to showing this is Proposition 4.

By Proposition 4, we have for some constant $c > 0$ that depends only on f, q, M, b, ζ such that for any $\beta \in \mathbb{R}^p_+$ and any variable $j \in S^c$,

$$\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}^A) \leq -c \cdot (F(\beta; \mathbb{Q}^A))^2. \tag{144}$$

We wish to emphasize that, in using Proposition 4, we implicitly use the fact that our choice of the weighting distribution $\mathbb{Q}^A$ maintains the distributional property $Y \mid X = Y \mid X_S$ and $X_S \perp X_{S^c}$ thanks to Proposition 3. Note that equation (144) basically says that the gradient with respect to a noise variable X_j at any β is negative, and, moreover the absolute value of the negative gradient is lower bounded by the square of the objective. In conjunction with the projection Lemma O.2, this implies for the same constant $c > 0$, we have

$$\begin{aligned} \beta_j^{(m+1)} &= \Pi_{\mathcal{B}} \left(\beta_j^{(m)} + \alpha \cdot \frac{\partial}{\partial \beta_j} F(\beta_j^{(m)}; \mathbb{Q}^A) \right) \\ &\leq \left(\beta_j^{(m)} - \alpha \cdot c \cdot (F(\beta^{(m)}; \mathbb{Q}^A))^2 \right)_+ \end{aligned} \tag{145}$$

To show the desired result, i.e., equation (142) holds for some constant $\varepsilon > 0$ independent of $m \in \mathbb{N}$, it remains to show the lower bound

$$F(\beta^{(m)}; \mathbb{Q}^A) \geq F(\beta^{(0)}; \mathbb{Q}^A) \quad \text{for all } m \in \mathbb{N}. \tag{146}$$

Indeed, once we have equations (145) and (146), then it is immediate that equation (142) holds for the constant $\varepsilon = \alpha \cdot c \cdot F(\beta^{(0)}; \mathbb{Q}^A)^2 \geq \alpha c \gamma > 0$.

Below we prove equation (146) holds when the constant $C > 0$ (stated in Claim 2) is sufficiently large. To do so, we invoke the very basic property of the projected gradient ascent algorithm—that is, the objective value increases monotonically along the iterates when the stepsize is sufficiently small. More precisely, by Lemma O.4, it suffices to show the stepsize $\alpha \leq 1/L$, where L is the Lipschitz constant of the gradient of the objective (with respect to $\|\cdot\|_2$ norm). Lemma K.1 upper bounds $L \leq |f''(0)| \overline{M} p$ where $\overline{M} = (2M)^q$. Hence, the stepsize $\alpha \leq \frac{1}{C \cdot p} \leq 1/L$ when $C > |f''(0)| \bar{M}$, and therefore $m \mapsto F(\beta^{(m)}; \mathbb{Q}^A)$ is monotonically increasing thanks to Lemma O.4. This proves equation (146).

Finite Sample Analysis $n < \infty$. Here we extend the above analysis to the finite sample case where $n < \infty$. The proof here is similar to that of $n = \infty$—the major difference is that we need to substitute the population objective and gradient by the empirical ones. For this reason, the key to extending the proof to finite case $n < \infty$ is to bound the difference between the population

and empirical objectives and gradients uniformly over $\beta \in \mathcal{B}$. Fortunately, Corollary J.1 provides such high probability bounds.

According to Corollary J.1, we have for some constant $\bar{C} > 0$ depending only on the parameters $b, M, \varrho, f(0), f'(0), f''(0)$ such that

- we have with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$,

$$\sup_{\beta \in \mathcal{B}} \left| F(\beta; \mathbb{Q}_n^A) - F(\beta; \mathbb{Q}^A) \right| \leq \bar{C} \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t) \tag{147}$$

- we have with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$,

$$\sup_{j \in [p]} \sup_{\beta \in \mathcal{B}} \left| \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}_n^A) - \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}^A) \right| \leq \bar{C} \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t). \tag{148}$$

Let Λ denote the event on which both equations (147) and (148) hold. We prove that equation (142) holds on the event Λ, provided that $C > 0$ (stated in Claim 2) is sufficiently large, and γ satisfyies equation (137).

Since the proof is essentially the same as before, we only briefly describe the major steps, leaving the details to the reader. In the discussions below, we assume we are on the event Λ. First, equation (147) implies that

$$\sup_{\beta \in \mathcal{B}} \left| F^2(\beta; \mathbb{Q}_n^A) - F^2(\beta; \mathbb{Q}^A) \right| \leq 2\bar{C} \cdot |f(0)| \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t) \tag{149}$$

as both $F(\beta; \mathbb{Q}_n^A)$ and $F(\beta; \mathbb{Q}^A)$ are uniformly bounded by $|f(0)|$ when $\beta \in \mathcal{B}$. Consequently, by equations (144), and (148) and (149) and the triangle inequality, we can obtain for all $\beta \in \mathcal{B}$ and all $j \in S^c$:

$$\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}_n^A) \leq -c \cdot (F(\beta; \mathbb{Q}_n^A))^2 + C' \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t) \tag{150}$$

where we define the constant $C' = (1 + c) \cdot \bar{C} \cdot |f(0)| > 0$. Next, by equation (150), and the projection Lemma O.2, we can obtain for all $t \in \mathbb{N}$ and $j \in S^c$:

$$\beta_j^{(m+1)} \leq \left(\beta_j^{(m)} - \alpha \cdot \left(c \cdot (F(\beta^{(t)}; \mathbb{Q}_n^A))^2 - C' \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t) \right) \right)_+ \tag{151}$$

Finally, by using the same strategy as proving equation (145), we obtain

$$F(\beta^{(m)}; \mathbb{Q}_n^A) \geq F(\beta^{(0)}; \mathbb{Q}_n^A) \quad \text{for all } m \in \mathbb{N} \tag{152}$$

when the constant $C > 0$ (stated in Claim 2) is sufficiently large. Hence, equations (151) and (152) together imply that, on the event Λ, the desired equation (142) holds for the constant

$$\varepsilon = \alpha \cdot \left(c \cdot (F(\beta^{(0)}; \mathbb{Q}_n^A))^2 - C' \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t) \right). \tag{153}$$

Since $(F(\beta^{(0)}; \mathbb{Q}_n^A))^2 \geq \gamma$ by assumption, the constant $\varepsilon \geq \frac{1}{2}\alpha c\gamma > 0$ if the condition $\gamma > C\sqrt{\frac{\log p}{n}}(1+t)$ holds and $C > 0$ is sufficiently large.

LEMMA K.1. *Let $f \in \mathcal{C}^\infty(\mathbb{R}_+)$ be such that f' is completely monotone. Assume that $\|X\|_\infty \leq M$ under $\mathbb{Q}$. Then, both the population objective $F(\beta; \mathbb{Q})$ and the empirical objective $F(\beta; \mathbb{Q}_n)$ has Lipschitz gradient with Lipschitz constant $L \leq |f''(0)|\overline{M}p$ where $\overline{M} = (2M)^q$.*

PROOF. We only prove the result for the population objective $F(\beta; \mathbb{Q})$ (the proof for $F(\beta; \mathbb{Q}_n)$ is essentially the same). To start with, note for $j \in [p]$,

$$\frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}) = \mathbb{E}_{B-W}^w \left[\mathbf{d}_j \cdot f'(\langle \beta, \mathbf{d} \rangle) \right].$$

As $\|\mathbf{d}\|_\infty \leq \overline{M}$, we obtain that $\beta \mapsto \langle \beta, \mathbf{d} \rangle$ is $\overline{M}\sqrt{p}$ Lipschitz. Since f' is completely monotone, it is Lipschitz with constant $|f''(0)|$. Consequently, we have shown that $\beta \mapsto \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q})$ is $|f''(0)|\overline{M}\sqrt{p}$ Lipschitz, and hence $\beta \mapsto \nabla F(\beta; \mathbb{Q})$ is Lipschitz with Lipschitz constant $L \leq |f''(0)|\overline{M}p$. □

K.4. Discussion on Early Stopping of Gradient Ascent. The proof of Theorem 2 suggests that we can still achieve the high probability no-false-positive guarantees if we modify the algorithm to perform an early stopping on the inner-loop gradient ascent iterates. Indeed, the proof of Theorem 2, and in particular, the proof of Claim 2 shows that with high probability the inner-loop gradient ascent iterates satisfy for all noise variable j:

$$\beta_{j+1}^{(m)} \leq \left(\beta_j^{(m)} - c \cdot \alpha \cdot \sqrt{\frac{\log p}{n}} \right)_+$$

for some constant $c > 0$ that is independent of p, n, provided that $\alpha \leq \frac{1}{C \cdot p}$ for some large constant $C > 0$ independent of p. Hence, if the initialization $\beta^{(0)}$ has coordinates on the order of $\frac{1}{p}$ (say $\beta^{(0)} = \frac{b}{p}\mathbf{1}$) and the stepsize $\alpha = \Omega(\frac{1}{p})$, then with constant number of iteration m' (here constant means the number

of iteration m' is independent of p), the gradient ascent iteration will reach a point where $\beta_j^{m'} = 0$ for all noise variable j. Performing such early stopping will significantly reduce the computation cost, while maintaining the statistical no-false-positive control. One thing that really deserves attention is that there is also the signal recovery guarantees when the stepsize $\alpha = \Omega(\frac{1}{p})$ (see the low-dimensional recovery result; Theorem 4).

APPENDIX L: PROOF OF RECOVERY GUARANTEE IN LOW DIMENSION

This section presents the proof of Theorem 4. The proof is based on a careful (and not trivial) extension of the proof of the recovery guarantee of the population algorithm studied in Proposition 5.

L.1. Notation. This section uses exactly the same notation as appeared in the previous Section K (see Section K.1).

L.2. Proof of Theorem 4. We prove that, with probability at least $1 - 2^s(p^{-t^2} + e^{-n\varrho^2/32})$, the output $\hat{S}$ satisfies $Y \mid X_{\hat{S}} = Y \mid X_S$. The key to the proof is the following proposition, whose proof is deferred into Section L.3.

PROPOSITION 8. *Let $A \subseteq S$ be such that $Y \mid X_A \neq Y \mid X_S$. Consider the optimization problem:*

$$\max_{\beta:\beta\in\mathcal{B}} F(\beta; \mathbb{Q}_n^A).$$

Let $\{\beta^{(m)}\}_{m\in\mathbb{N}}$ denote the projected gradient ascent iterates with initialization at $\beta^{(0)}$ and stepsize α. Then there exists some constant $C > 0$ that depends only on the parameters $b, M, q, \varrho, f(0), f'(0), f''(0)$ such that the following holds: for any $t > 0$, any choice of the stepsize $\alpha \leq \frac{1}{C\cdot p}$, and any choice of the threshold γ that satisfies

$$\gamma \geq C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t),$$

then if the following initialization condition holds (on population):

$$\left(F(\beta^{(0)}; \mathbb{Q}_A)\right)^2 > \gamma + C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t).$$

then we have with probability $1 - (p^{-t^2} + e^{-n\varrho^2/32})$ the following two happen:

- *the initialization condition $(F(\beta^{(0)}; \mathbb{Q}_n^A))^2 > \gamma$ holds*

- *any accumulation point β^* of the projected gradient iterates satisfies*

$$\text{supp}(\beta^{(*)}) \backslash A \neq \emptyset. \tag{154}$$

We are now ready to show that $\hat{S}$ satisfies $Y \mid X_{\hat{S}} = Y \mid X_S$ with high probability. Fix $t > 0$ in the statement of Theorem 4. Let $\mathcal{E}_A$ and $\mathcal{E}'_A$ denote the event that's stated in Proposition 7 and Proposition 8 respectively. Let $\mathcal{E} = \cap_{A:A\subseteq S}\mathcal{E}_A$ and $\mathcal{E}' = \cap_{A:A\subseteq S}\mathcal{E}'_A$. By Proposition 7 and Proposition 8 and the union bound, we understand with the appropriate choice of constant $C > 0$ in the definition of $\mathcal{E}, \mathcal{E}'$, we have

$$\mathbb{P}(\mathcal{E} \cap \mathcal{E}') \geq 1 - 2^{s+2}(p^{-t^2} + e^{-n\varrho^2}). \tag{155}$$

Fix the constant $C > 0$. Suppose the following condition holds:

$$\inf_{A:A\subseteq S, Y|X_A \neq Y|X_S} \left(F(\beta^{(0)}; \mathbb{Q}_A)\right)^2 \geq \gamma + C \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t).$$

Below we show under this condition, and on the event $\mathcal{E} \cap \mathcal{E}'$, the algorithm does not halt until it finds $\hat{S}$ such that $Y \mid X_{\hat{S}} = Y \mid X_S$. Our reasoning is based on the following points.

- By the same analysis in the proof of Theorem K (see Section K.2), we understand on the event $\mathcal{E}$, the algorithm does not over-select noise variables, i.e.,
$$\hat{S}^{(k)} \subseteq S \text{ holds for all } k \in \mathbb{N}.$$
- Now, suppose we are at the iteration k and $Y \mid \hat{S}^{(k)} \neq Y \mid S$. Now we show the algorithm does not halt at this iteration and $\hat{S}^{(k+1)} \supsetneq \hat{S}^{(k)}$. Indeed, since $\hat{S}^{(k)} \subseteq S$ and $Y \mid \hat{S}^{(k)} \neq Y \mid S$, by definition we know that on the event $\mathcal{E}'$, the initialization condition $(F(\beta^{(0)}; \mathbb{Q}_n^{(k)}))^2 > \gamma$ holds, and moreover the gradient ascent iterate returns $\beta^{(*;\mathbb{Q}_n^{(k)})}$ with
$$\text{supp}\left(\beta^{(*;\mathbb{Q}_n^{(k)})}\right) \backslash \hat{S}^{(k)} \neq \emptyset.$$
Thus the algorithm does not halt at the iteration, and moreover,
$$\hat{S}^{(k+1)} = \hat{S}^{(k)} \cup \text{supp}\left(\beta^{(*;\mathbb{Q}_n^{(k)})}\right) \supsetneq \hat{S}^{(k)}.$$

As such, we have shown that $\hat{S} \subseteq S$ on the event $\mathcal{E}$, and moreover, the output $\hat{S}$ must satisfy $Y \mid \hat{S} = Y \mid S$. This proves the recovery result in Theorem 4.

L.3. Proof of Proposition 8. Our strategy is to first prove that the proposition holds on population ($n = \infty$), and then extend the result to finite sample case ($n < \infty$) by standard concentration inequalities and perturbation arguments.

Population Analysis: $n = \infty$. Let $A \subseteq S$ be such that $Y \mid X_A \neq Y \mid X_S$. We prove that with probability one, any accumulation point $\beta^{(*)}$ of the gradient iterates must satisfy $\text{supp}(\beta^{(*)}) \backslash A \neq \emptyset$. To prove this, we notice the following basic facts.

- By assumption, the inequality below holds

$$(F(\beta^{(0)}; \mathbb{Q}_A))^2 > \gamma. \tag{156}$$

  Hence, the algorithm passes the initialization condition.
- Mimic the proof of equation (146), one can show that if the stepsize $\alpha \leq \frac{1}{Cp}$ for some sufficiently large constant $C > 0$, then

$$F(\beta^{(m)}; \mathbb{Q}_A) \geq F(\beta^{(0)}; \mathbb{Q}_A) \quad \text{for all } m \in \mathbb{N}.$$

  In particular, this shows that any accumulation point $\beta^{(*)}$ of the gradient iterates must satisfy

$$F(\beta^{(*)}; \mathbb{Q}_A) \geq F(\beta^{(0)}; \mathbb{Q}_A) > 0. \tag{157}$$

- By Proposition 3, we know that $X_A \perp Y$ under $\mathbb{Q}_A$. Hence,

$$F(\beta; \mathbb{Q}_A) = 0 \text{ when } \text{supp}(\beta) \subseteq A. \tag{158}$$

Equations (157) and (158) immediately show that $\text{supp}(\beta^{(*)}) \backslash A$ is not empty.

Finite Sample Analysis: $n < \infty$. Here we extend the above analysis to the finite sample situation where $n < \infty$. The proof here is essentially the same as that of $n = \infty$—the only change we make is to substitute the population objective by the empirical one. The main technical tool that we need is Corollary J.1, which provides a high probability bound on the difference between the empirical and population objective uniformly over $\beta \in \mathcal{B}$.

Again, we start with any set $A \subseteq S$ such that $Y \mid X_A \neq Y \mid X_S$. By Corollary J.1, for some constant $\bar{C}$ that depends only on $b, M, \varrho, f(0), f'(0), f''(0)$, we have with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$,

$$\sup_{\beta \in \mathcal{B}} \left| F(\beta; \mathbb{Q}_n^A) - F(\beta; \mathbb{Q}_A) \right| \leq \bar{C} \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t) \tag{159}$$

Let Λ denote the event that equation (159) holds. It suffices to prove that, on the event Λ, any accumulation point $\beta^{(*)}$ of the gradient iterates must satisfy $\mathrm{supp}(\beta^{(*)})\backslash A \neq \emptyset$. To prove this, we mimic what we have done before. In the discussion below, we assume that we are on the event Λ.

- We check whether the algorithm passes the initialization condition:

$$(F(\beta^{(0)}; \mathbb{Q}_n^A))^2 \geq \gamma. \tag{160}$$

Note that, since both $\beta \mapsto F(\beta; \mathbb{Q}_n^A)$ and $\beta \mapsto F(\beta; \mathbb{Q}_A)$ are uniformly bounded by $|f(0)|$ when $\beta \in \mathcal{B}$, equation (159) immediately implies that

$$\sup_{\beta \in \mathcal{B}} \left| F^2(\beta; \mathbb{Q}_n^A) - F^2(\beta; \mathbb{Q}_A) \right| \leq 2\bar{C} \cdot |f(0)| \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t). \tag{161}$$

By definition of Λ and triangle inequality, a sufficient condition is

$$(F(\beta^{(0)}; \mathbb{Q}_A))^2 \geq \gamma + 2\bar{C} \cdot |f(0)| \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t). \tag{162}$$

where $\bar{C} > 0$ is the constant that appears in the equation (159).

- Again, mimic the proof of equation (146), one can show that if the stepsize $\alpha \leq \frac{1}{(C \cdot p)}$ for some sufficiently large constant $C > 0$, then

$$F(\beta^{(m)}; \mathbb{Q}_n^A) \geq F(\beta^{(0)}; \mathbb{Q}_n^A) \quad \text{for all } m \in \mathbb{N}.$$

In particular, this shows that any accumulation point $\beta^{(*)}$ of the gradient iterates must satisfy

$$(F(\beta^{(*)}; \mathbb{Q}_n^A))^2 \geq (F(\beta^{(0)}; \mathbb{Q}_n^A))^2.$$

By triangle inequality, and using the definition of Λ (cf. equation (161)) again, we obtain that any accumulation point $\beta^{(*)}$ must satisfy

$$(F(\beta^{(*)}; \mathbb{Q}_A))^2 \geq \gamma - 2\bar{C} \cdot |f(0)| \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t)$$

where $\bar{C} > 0$ is the constant that appears in the equation (159). This shows that if $\gamma > 2\bar{C} \cdot |f(0)| \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t)$, then on the event Λ,

$$F(\beta^{(*)}; \mathbb{Q}_A) > 0. \tag{163}$$

- By Proposition 3, we know that $X_A \perp Y$ under $\mathbb{Q}_A$. Hence,

$$F(\beta; \mathbb{Q}_A) = 0 \text{ when } \operatorname{supp}(\beta) \subseteq A. \tag{164}$$

Summarizing the above discussions, we have seen that if the stepsize $\alpha \leq \frac{1}{Cp}$, and if the threshold γ satisfies

$$\gamma > C \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t)$$

for some large constant $C > 0$, then equations (160), (163) and (164) hold conditional on the event Λ. In other words, when the constant $C > 0$ stated in Proposition 8 is sufficiently large, then $\operatorname{supp}(\beta^{(*)}) \backslash A \neq \emptyset$, conditional on the event Λ. Recall $\mathbb{P}(\Lambda) \geq 1 - p^{-t^2} - e^{-n\varrho^2/32}$. This completes the proof.

APPENDIX M: PROOF OF FALSE POSITIVE CONTROL IN HIGH-DIMENSION

This section presents the proof of Theorem 3. The fundamental mathematical tool that underlies the proof is Proposition 3 and Corollary J.1. Proposition 3 shows on population the gradient with respect to noise variable is negative—and thereby there is no false discovery. Corollary J.1 shows that the empirical gradient is uniformly close to the population gradient. Therefore, with an explicit ℓ_1 penalty, we can easily translate the no false positive result from the population to the empirical objective, which is the content of Theorem 3.

M.1. Notation. Consider the metric learning algorithm. Let $\mathbb{Q}^{(1)}, \mathbb{Q}^{(2)}, \ldots, \mathbb{Q}^{(k)}, \ldots$ denote the sequence of the weighting distribution, and $\hat{S}^{(1)}, \hat{S}^{(2)}, \ldots, \hat{S}^{(k)}, \ldots$ denote the set of variables selected by the algorithm through the iterations. By convention, we define $\hat{S}^{(0)} = \emptyset$. Note then $\hat{S} = \bigcup_k \hat{S}^{(k)}$ is the final output of the algorithm. We use $\{\beta^{(m;\mathbb{Q}^{(k)})}\}_{m\in\mathbb{N}}$ to denote the inner-loop projected gradient ascent iterates that solve the maximization problem

$$\max_{\beta \in \mathcal{B}} \ell(\beta; \lambda, \mathbb{Q}_n^{(k)})$$

where we recall the definition of the regularized objective (for all $\mathbb{Q}$)

$$\ell(\beta; \lambda, \mathbb{Q}) = F(\beta; \mathbb{Q}) - \lambda \|\beta\|_1 .$$

We use $\beta^{(*;\mathbb{Q}^{(k)})}$ to denote the accumulation point that's returned from the gradient ascent inner-loop. This means in particular we have

$$\hat{S}^{(k+1)} = \hat{S}^{(k)} \cup \operatorname{supp}(\beta^{(*;\mathbb{Q}^{(k)})}).$$

We sometimes drop the dependence of the gradient ascent iterates on the probability measure $\mathbb{Q}$. This means that we may refer $\beta^{(m)}$ to $\beta^{(m;\mathbb{Q}^{(k)})}$, and refer $\beta^{(*)}$ to $\beta^{(*;\mathbb{Q}^{(k)})}$ when the context is clear.

M.2. Proof of Theorem 3.

M.2.1. *Organization of the Proof.* On a high level, the proof has two parts.

- In the first part, we show that with high probability $\hat{S} \subseteq S$.
- In the second part, we show that with high probability the algorithm terminates in finite time.

We detail the two parts of the proofs in subsection M.2.2 and M.2.3 below.

M.2.2. *Proof of Part 1: $\hat{S} \subseteq S$.* In this section, we show that $\hat{S} \subseteq S$ with probability at least $1 - 2^{s+1}(p^{-t^2} + e^{-n\varrho^2/32})$. The key to the proof is the following proposition 3, whose proof is deferred to section M.3.1.

PROPOSITION 9. *Let $A \subseteq S$. Consider the optimization problem*

$$\mathcal{O}_{n,\lambda,A} : \underset{\beta \in \mathcal{B}}{\text{maximize}}\, \ell(\beta; \lambda, \mathbb{Q}_n^A). \tag{165}$$

Then there exists some constant $C > 0$ that depends only on the parameters $b, M, q, \varrho, f(0), f'(0), f''(0)$ such that for any $t > 0$ and any penalty

$$\lambda > C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t),$$

with probability at least $1 - (p^{-t^2} + e^{-n\varrho^2/32})$, we have any stationary point β of the optimization problem $\mathcal{O}_{n,\lambda,A}$ (cf. equation (165)) must satisfy

$$\text{supp}(\beta) \subseteq S. \tag{166}$$

We are now ready to show that $\hat{S} \subseteq S$ with high probability. This is achieved by showing that with high probability $\hat{S}^{(k)} \subseteq S$ for all $k \in \mathbb{N}$. Let $\mathcal{E}_A$ denote the event that's stated in Proposition 9. Let $\mathcal{E} = \cap_{A:A\subseteq S}\mathcal{E}_A$ denote the event where all events $\mathcal{E}_A$ where $A \subseteq S$ happens. By Proposition 9 and the union bound, we understand with appropriate choice of constant $C > 0$ in the definition of $\mathcal{E}$,

$$\mathbb{P}(\mathcal{E}) \geq 1 - 2^{s+1}(p^{-t^2} + e^{-n\varrho^2/32}). \tag{167}$$

Below, we show on the event $\mathcal{E}$ that the following happens:

$$\hat{S}^{(k)} \subseteq S \text{ holds for all } k \in \mathbb{N}. \tag{168}$$

Note we define by convention $\hat{S}^{(k)} = \hat{S}$ for any $k > T$ where T is the total number of iterations after which the algorithm halts. Our proof is based on induction on $m \in \mathbb{N}$. Below we assume we are on the event $\mathcal{E}$.

- The base case where $k = 0$ trivially holds since $\hat{S}_0 = \emptyset$.
- Assume the induction hypothesis holds for k, i.e., $\hat{S}^{(k)} \subseteq S$. Consider the case where $k+1$. We aim to show that $\hat{S}^{(k+1)} \subseteq S$. Note that if we run the gradient ascent to solve the maximization problem

$$\mathcal{O}_{n,\lambda,\hat{S}^{(k)}} : \max_{\beta \in \mathcal{B}} \ell_n(\beta; \lambda, \mathbb{Q}^{(k)})$$

the solution $\beta^{(*;\mathbb{Q}^{(k)})}$ we obtain must be a stationary point of the problem when the stepsize $\alpha \leq 1/(Cp)$ for sufficiently large C (This is a consequence of Lemma K.1 and Lemma O.4; Lemma K.1 shows that $\ell_n(\beta; \lambda, \mathbb{Q}^{(k)})$ is $L = |f''(0)|\overline{M}p$ smooth where $\overline{M} = (2M)^q$, and Lemma O.4 shows that any accumulation point of gradient ascent with stepsize $\alpha \leq 1/L$ must be stationary). Since $\hat{S}^{(k)} \subseteq S$, by definition of the event $\mathcal{E}$, we have

$$\operatorname{supp}\left(\beta^{(*;\mathbb{Q}^{(k)})}\right) \subseteq S.$$

Consequently, this shows that, the output $\hat{S}^{(k+1)}$ satisfies

$$\hat{S}^{(k+1)} = \hat{S}^{(k)} \cup \operatorname{supp}\left(\beta^{(*;\mathbb{Q}^{(k)})}\right) \subseteq S.$$

This proves that the hypothesis also holds for $m = k+1$.

As such, we have shown that $\hat{S} \subseteq S$ on the event $\mathcal{E}$. This happens with probability at least $1 - 2^{s+1}(p^{-t^2} + e^{-n\varrho^2/32})$, thanks to equation (167).

M.2.3. *Proof of Part 2: Termination in Finite Time.* In this section, we show that with high probability the algorithm terminates in finite time. The key to the proof is the following Proposition 10, whose proof is deferred into Section M.3.2.

PROPOSITION 10. *Let $A \subseteq S$. Consider the optimization problem*

$$\mathcal{O}_{n,\lambda,A} : \operatorname*{maximize}_{\beta \in \mathcal{B}} \ell(\beta; \lambda, \mathbb{Q}_n^A). \tag{169}$$

Then there exists some constant $C > 0$ that depends only on the parameters $b, M, q, \varrho, f(0), f'(0), f''(0)$ such that for any $t > 0$ and any penalty

$$\lambda > C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t),$$

with probability at least $1 - (p^{-t^2} + e^{-n\varrho^2/32})$, we have any stationary point β of the optimization problem $\mathcal{O}_{n,\lambda,A}$ (cf. equation (169)) must satisfy

$$\text{either} \quad \beta = 0 \quad \text{or} \quad \operatorname{supp}(\beta) \backslash A \neq \emptyset. \tag{170}$$

We are now ready to show that the algorithm terminates in finite time. The basic intuition is that with high probability at any iteration $k \in \mathbb{N}$, the solution $\beta^{(*,\mathbb{Q}^{(k)})}$ must fall into one of the following two cases:

- $\beta^{(*,\mathbb{Q}^{(k)})} = 0$. This is the case when the algorithm halts.
- $\operatorname{supp}(\beta^{(*,\mathbb{Q}^{(k)})}) \backslash \hat{S}^{(k)} \neq \emptyset$. This is the case where the algorithm finds at least one new variable, i.e., $\hat{S}^{(k+1)} \supseteq \hat{S}^{(k)}$.

Let $\mathcal{E}_A$ and $\mathcal{E}'_A$ denote the event that's stated in Proposition 9 and Proposition 10 respectively. Let $\mathcal{E} = \cap_{A:A\subseteq S}\mathcal{E}_A$ and $\mathcal{E}' = \cap_{A:A\subseteq S}\mathcal{E}'_A$. By Proposition 7 and Proposition 8 and the union bound, we understand with the appropriate choice of constant $C > 0$ in the definition of $\mathcal{E}, \mathcal{E}'$, we have

$$\mathbb{P}(\mathcal{E} \cap \mathcal{E}') \geq 1 - 2^{s+2}(p^{-t^2} + e^{-n\varrho^2/32}). \tag{171}$$

Below we show on the event $\mathcal{E} \cap \mathcal{E}'$ the algorithm terminates in finite time. Our reasoning is based on the following points.

- By the analysis in the previous part, we understand on the event $\mathcal{E}$, the algorithm does not over-select noise variables, i.e.,

  $$\hat{S}^{(k)} \subseteq S \quad \text{holds for all } k \in \mathbb{N}.$$

- Now, suppose we are at the iteration k. By the previous point, on the event $\mathcal{E}$, $\hat{S}^{(k)} \subseteq S$. By the definition of $\mathcal{E}'$, we know that the gradient ascent iterate returns $\beta^{(*;\mathbb{Q}^{(k)})}$ that satisfies

  $$\text{either} \quad \beta^{(*;\mathbb{Q}^{(k)})} = 0 \quad \text{or} \quad \operatorname{supp}\left(\beta^{(*;\mathbb{Q}^{(k)})}\right) \backslash \hat{S}^{(k)} \neq \emptyset. \tag{172}$$

  Thus, we have two cases based on the value of $\beta^{(*;\mathbb{Q}^{(k)})}$.
  - $\beta^{(*;\mathbb{Q}^{(k)})} = 0$. Then the algorithm halts at this kth iteration.

– $\beta^{(*;\mathbb{Q}^{(k)})} \neq 0$. In this case, the algorithm adds at least one new variable. Indeed, by equation (172), we have

$$\hat{S}^{(k+1)} = \hat{S}^{(k)} \cup \operatorname{supp}\left(\beta^{(*;\mathbb{Q}^{(k)})}\right) \supsetneq \hat{S}^{(k)}.$$

This shows that the algorithm terminates in at most $|S|$ steps.

As such, we have shown that the algorithm terminates in finite time on the event $\mathcal{E} \cap \mathcal{E}'$. This happens with probability at least $1 - 2^{s+2}(p^{-t^2} + e^{-n\varrho^2/32})$, thanks to equation (171).

M.3. Proof of the Propositions 9-10.

M.3.1. *Proof of Proposition 9.* The key to the proof is to establish the following result.

Claim 3. *Let $A \subseteq S$. There exists some constant $C > 0$ depending only on $b, M, q, \varrho, f(0), f'(0), f''(0)$ such that for any $t > 0$ and any penalty*

$$\lambda > C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t), \tag{173}$$

we have with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$, the inequality

$$\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}_n^A) < 0$$

holds simultaneously for any $\beta \in \mathcal{B}$ and any index $j \in S^c$.

Claim 3 says with high probability, any $\beta \in \mathcal{B}$ with $\beta_{S^c} \neq 0$ cannot be a stationary point of the optimization problem—decreasing the value of β_j where $j \in \operatorname{supp}(\beta) \cap S^c$ strictly increases the objective $\ell(\beta; \lambda, \mathbb{Q}_n^A)$. Thus, proving the claim immediately leads to the desired Proposition 9.

Now we prove Claim 3. Our proof of Claim 3 proceeds as follows. We first prove the Claim holds on population ($n = \infty$). Then, we extend the Claim, showing it also holds in finite sample ($n < \infty$) by using standard concentration and perturbation arguments. We note for any $\beta \in \mathcal{B}$ the gradient takes the form

$$\begin{aligned} \frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A) &= \mathbb{E}_{B-W}^{w_A}\left[\mathbf{d}_j \cdot f'\left(\langle \beta, \mathbf{d} \rangle\right)\right] - \lambda. \\ \frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}_n^A) &= \hat{\mathbb{E}}_{n,B-W}^{w_A}\left[\mathbf{d}_j \cdot f'\left(\langle \beta, \mathbf{d} \rangle\right)\right] - \lambda. \end{aligned} \tag{174}$$

Population Analysis $n = \infty$. Here we show that Claim 3 holds in population. Proposition 3 is crucial towards this end. By Proposition 3, we obtain for any $\beta \in \mathbb{R}^p_+$ and $j \in S^c$,

$$\frac{\partial}{\partial \beta_j} F(\beta; \lambda, \mathbb{Q}^A) = \mathbb{E}^{w_A}_{B-W} \left[\mathbf{d}_j \cdot f' \left(\langle \beta, \mathbf{d} \rangle \right) \right] \leq 0.$$

We wish to emphasize that, in using Proposition 3, we implicitly use the fact that our choice of the weighting distribution $\mathbb{Q}^A$ maintains the distributional property $Y \mid X = Y \mid X_S$ and $X_S \perp X_{S^c}$ thanks to Proposition 3.

Now, using the first equation in equations (174), we immediately obtain

$$\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A) = \mathbb{E}^{w_A}_{B-W} \left[\mathbf{d}_j \cdot f' \left(\langle \beta, \mathbf{d} \rangle \right) \right] - \lambda \leq -\lambda < 0. \tag{175}$$

This proves Claim 3 holds on population ($n = \infty$).

Finite Sample Analysis $n < \infty$. Here we prove Claim 3 also holds in finite sample scenario $n < \infty$. The proof is essentially the same as before, and the only difference is to replace the population gradient by the empirical one. Hence, the key to the proof is to bound the difference between the quantities

$$\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A_n) \quad \text{and} \quad \frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A)$$

over all possible values of $\beta \in \mathcal{B}$ and all possible variable $j \in S^c$. Corollary J.1 provides such a high probability bound. Indeed, it shows that for some constant $C > 0$ depending only on $b, M, \varrho, f(0), f'(0), f''(0)$, we have with probability at least $1 - (p^{-t^2} + e^{-n\varrho^2/32})$,

$$\sup_{j \in [p]} \sup_{\beta \in \mathcal{B}} \left| \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}^A_n) - \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}^A) \right| \leq C \sqrt{\frac{\log p}{n}} (1 + t). \tag{176}$$

Let's denote Λ to be the event where equation (176) holds. Using equations (174), we immediately obtain that, on the same event Λ,

$$\sup_{j \in [p]} \sup_{\beta \in \mathcal{B}} \left| \frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A_n) - \frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A) \right| \leq C \sqrt{\frac{\log p}{n}} (1 + t). \tag{177}$$

Now assume the penalty λ is greater than the RHS of equation (177), i.e.,

$$\lambda > C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t).$$

By equations (175), (177) and the triangle inequality, on the event Λ, we have for any $\beta \in \mathcal{B}$ and any $j \in S^c$:

$$\begin{aligned}
&\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}_n^A) \\
&\leq \frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A) + \left| \frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}_n^A) - \frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A) \right| \\
&\leq -\lambda + C \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t) < 0.
\end{aligned}$$

Recall $\mathbb{P}(\Lambda) \geq 1 - p^{-t^2} - e^{-n\varrho^2/32}$. This shows Claim 3 holds in finite case.

Summary. As mentioned earlier, Claim 3 implies Proposition 9 as desired.

M.3.2. *Proof of Proposition 10.* The key to the proof is to establish the following result.

CLAIM 4. *Let $A \subseteq S$. There exists some constant $C > 0$ depending only on $b, M, q, \varrho, f(0), f'(0), f''(0)$ such that for any $t > 0$ and any penalty*

$$\lambda > C \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t), \tag{178}$$

we have with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$, the inequality

$$\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}_n^A) < 0$$

holds simultaneously for any $\beta \in \mathcal{B}$ with $\operatorname{supp}(\beta) \subseteq A$ and any index $j \in A$.

Claim 4 says with high probability, any $0 \neq \beta \in \mathcal{B}$ with $\operatorname{supp}(\beta) \subseteq A$ cannot be a stationary point of the optimization problem—decreasing the value of β_j where $j \in \operatorname{supp}(\beta) \cap A$ strictly increases the objective $\ell(\beta; \lambda, \mathbb{Q}_n^A)$. Thus, proving the claim immediately leads to the desired Proposition M.3.2.

Now we prove Claim 4. Our proof of Claim 4 proceeds as follows. We first prove the Claim holds on population ($n = \infty$). Then, we extend the Claim, showing it also holds in finite sample ($n < \infty$) by using standard concentration and perturbation arguments. We note any $\beta \in \mathcal{B}$ has gradients

$$\begin{aligned}
\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A) &= \mathbb{E}_{B-W}^{w_A} \left[\mathbf{d}_j \cdot f'\left(\langle \beta, \mathbf{d} \rangle \right) \right] - \lambda. \\
\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}_n^A) &= \hat{\mathbb{E}}_{n,B-W}^{w_A} \left[\mathbf{d}_j \cdot f'\left(\langle \beta, \mathbf{d} \rangle \right) \right] - \lambda.
\end{aligned} \tag{179}$$

Population Analysis $n = \infty$. Here we show that Claim 4 holds in population. By Proposition 3, $X_A \perp Y$ under $\mathbb{Q}^A$. Hence, we have for all $\beta \in \mathcal{B}$ with $\text{supp}(\beta) \subseteq A$ and all index $j \in A$:

$$\begin{aligned}\frac{\partial}{\partial \beta_j}\ell(\beta;\lambda;\mathbb{Q}^A) &= \mathbb{E}^{w_A}_{B-W}\left[\mathbf{d}_j \cdot f'(\left\|X - X'\right\|_{q,\beta})\right] - \lambda \\ &= \mathbb{E}^{w_A}_{B-W}\left[\mathbf{d}_j \cdot f'(\left\|X_A - X'_A\right\|_{q,\beta_A})\right] - \lambda \\ &= -\lambda < 0,\end{aligned} \tag{180}$$

This proves Claim 4 holds on population ($n = \infty$).

Finite Sample Analysis $n < \infty$. Here we prove Claim 4 also holds in finite sample scenario $n < \infty$. The proof is essentially the same as before, and the only difference is to replace the population gradient by the empirical one. Hence, the key to the proof is to bound the difference between the quantities

$$\frac{\partial}{\partial \beta_j}\ell(\beta;\lambda,\mathbb{Q}^A_n) \quad \text{and} \quad \frac{\partial}{\partial \beta_j}\ell(\beta;\lambda,\mathbb{Q}^A)$$

over all possible values of $\beta \in \mathcal{B}$ and all possible variable $j \in S^c$. Corollary J.1 provides such a high probability bound. Indeed, it shows that for some constant $C > 0$ depending only on $b, M, \varrho, f(0), f'(0), f''(0)$, we have with probability at least $1 - (p^{-t^2} + e^{-n\varrho^2/32})$,

$$\sup_{j\in[p]}\sup_{\beta\in\mathcal{B}}\left|\frac{\partial}{\partial \beta_j}F(\beta;\mathbb{Q}^A_n) - \frac{\partial}{\partial \beta_j}F(\beta;\mathbb{Q}^A)\right| \le C\sqrt{\frac{\log p}{n}}(1+t). \tag{181}$$

Let's denote Λ to be the event where equation (181) holds. Using equations (179), we immediately obtain that, on the same event Λ,

$$\sup_{j\in[p]}\sup_{\beta\in\mathcal{B}}\left|\frac{\partial}{\partial \beta_j}\ell(\beta;\lambda,\mathbb{Q}^A_n) - \frac{\partial}{\partial \beta_j}\ell(\beta;\lambda,\mathbb{Q}^A)\right| \le C\sqrt{\frac{\log p}{n}}(1+t). \tag{182}$$

Now assume the penalty λ is greater than the RHS of equation (182), i.e.,

$$\lambda > C \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t).$$

By equations (180), (182) and the triangle inequality, on the event Λ, we have for for all $\beta \in \mathcal{B}$ with $\text{supp}(\beta) \subseteq A$ and all index $j \in A$:

$$\begin{aligned}
&\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}_n^A) \\
&\leq \frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A) + \left| \frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}_n^A) - \frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A) \right| \\
&\leq -\lambda + C \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t) < 0.
\end{aligned}$$

Recall $\mathbb{P}(\Lambda) \geq 1 - p^{-t^2} - e^{-n\varrho^2/32}$. This shows Claim 4 holds in finite case.

Summary. As mentioned, Claim 4 implies Proposition 10 as desired.

M.4. Discussion on Early Stopping of Gradient Ascent. The proof of Theorem 3 suggests that we can still achieve the high probability no-false-positive guarantees if we modify the algorithm to perform an early stopping on the inner-loop gradient ascent iterates. Indeed, the proof of Theorem 3, and in particular, the result in Claim 3 shows that with high probability the inner-loop gradient ascent iterates satisfy for all noise variable j:

$$\beta_{j+1}^{(m)} \leq \left(\beta_j^{(m)} - \alpha\lambda \right)_+$$

provided that $\lambda \geq C \cdot \sqrt{\frac{\log p}{n}}$ for some constant $C > 0$ independent of p. Hence, if the initialization $\beta^{(0)}$ has coordinates on the order of $\frac{1}{p}$ (say $\beta^{(0)} = \frac{b}{p}\mathbf{1}$) and the stepsize $\alpha = \Omega(\frac{1}{p})$, then with constant number of iteration m' (here constant means the number of iteration m' is independent of p), the gradient ascent iteration will reach a point where $\beta_j^{m'} = 0$ for all noise variable j. Performing such early stopping will significantly reduce the computation cost, while maintaining the statistical guarantee on no-false-positive control. One thing that really worth attention is that there is also the signal recovery guarantees when the stepsize $\alpha = \Omega(\frac{1}{p})$ (see the high-dimensional recovery results; Theorem 5—7).

APPENDIX N: PROOFS OF RECOVERY IN HIGH DIMENSION

This section presents the proof of all the recovery results in Section 5 in the main text. The roadmap of the section is as follows.

1. Section N.1 presents the proof of Theorem 5, showing that metric learning recover main effect signals w.h.p with $n \sim \log p$ samples.

2. Section N.2 presents a more general recovery result on the main effect signals, namely, Proposition 11, that generalizes Theorem 5. As a comparison, while Theorem 5 assumes conditional independence between signal variables, Proposition 11 places no assumptions on the dependence structure on the signal variables.
3. Section N.3 presents the proof of Theorem 6, showing that metric learning algorithm recovers s order pure interaction w.h.p with $n \sim p^{2(s-1)} \log p$ sample; the recovery assumes a computational budget that is linear in the total variables p. Section N.4 discusses how a slight modification of the original metric learning algorithm manages to detect s order pure interaction with $n \sim p^{2(s-s_0)+} \log p$ samples, under the assumption of a larger computational budget $O(p^{s_0})$.
4. Section N.5 presents the proof of Theorem N.5, showing that a slight modification of the original metric learning (which we call hierarchical metric learning, see Algorithm 1) recovers the signals under hierarchical model with $n \sim \log p$ samples. Section N.6 shows such modification is necessary from a computational perspective (by analyzing the landscape of the hierarchical model).
5. Section N.7 presents the proof of Proposition 6, showing that the statistical information in the objective and gradient for ℓ_1 type kernel is much stronger than that for the ℓ_2 type kernel.
6. Section N.8 presents the proof of a technical result used in the proof of Theorem 6.

N.1. Proof of Theorem 5 (Recovery of Main Effects). We will show with probability at least $1-2^s(p^{-t^2}+e^{-n\varrho^2/32})$, the algorithm successfully selects all the signal variables within the set $S(\lambda)$ before it terminates. Together with Theorem 3, which shows the algorithm with high probability does not over-select any noise variable, we obtain Theorem 5 as desired.

Below we show the algorithm outputs $\hat{S}$ that contains $S(\lambda)$ with probability at least $1-2^s(p^{-t^2}+e^{-n\varrho^2/32})$. Denote the optimization problem

$$\mathcal{O}_{n,\lambda,\mathbb{Q}} : \underset{\beta\in\mathcal{B}}{\text{maximize}}\, \ell(\beta;\lambda;\mathbb{Q}_n).$$

For any $A \subseteq [p]$, solving $\mathcal{O}_{n,\lambda,\mathbb{Q}}$ for $\mathbb{Q} = \mathbb{Q}^A$ with gradient ascent will always return a stationary point of $\mathcal{O}_{n,\lambda,\mathbb{Q}}$ when the stepsize $\alpha \le 1/(Cp)$ for the constant $C = |f''(0)|\overline{M}$ where $\overline{M} = 2M$ (thanks to Lemma K.1 and O.4).

Here is the key to the proof: with probability at least $1-2^s(p^{-t^2}+e^{-n\varrho^2/32})$, $0 \in \mathcal{B}$ can't be stationary for the optimization problem $\mathcal{O}_{n,\lambda,\mathbb{Q}^{w_{\hat{S}}}}$ unless $\hat{S}$ contains $S(\lambda)$. Once we prove the claim, by Proposition 10 and Theorem 3, we know with high probability the algorithm will proceed until $\hat{S}$ contains

$S(\lambda)$. Below we will first prove the key claim holds on population ($n = \infty$), and then prove it also holds on finite sample ($n < \infty$) using standard concentration and perturbation techniques.

Population case: $n = \infty$. Write $A \equiv \hat{S}^{(k)}$, the variables being selected by iteration k. Suppose A does not contain the set $S(\lambda)$. Now we prove $0 \in \mathcal{B}$ can't be stationary with respect to the optimization problem $\mathcal{O}_{\infty,\lambda,\mathbb{Q}^A}$.

Pick a signal variable $j \in S(\lambda)\backslash A$. We compute the derivative of the objective function $\ell(\beta; \lambda, \mathbb{Q}^A)$ with respect to variable j at $\beta = 0$:

$$\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A) \mid_{\beta=0} = f'(0) \cdot \mathbb{E}_{B-W}^{w_A} \left[\left| X_j - X_j' \right| \right] - \lambda. \tag{183}$$

According to Proposition 3, the rebalancing procedure keeps the law of $X_{A^c} \mid Y$ before and after reweighting. In particular, $X_j \mid Y$ has the same law under $\mathbb{Q}^A$ and $\mathbb{P}$. Therefore, we have

$$\mathbb{E}_{B-W}^{w_A} \left[\left| X_j - X_j' \right| \right] = \mathbb{E}_{B-W} \left[\left| X_j - X_j' \right| \right]. \tag{184}$$

By equations (183) and (184), and the fact that $j \in S(\lambda, t)$, we obtain

$$\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A) \mid_{\beta=0} = f'(0) \cdot \mathbb{E}_{B-W} \left[\left| X_j - X_j' \right| \right] - \lambda > 0, \tag{185}$$

where the last inequality uses the definition of the set $S(\lambda)$. This shows that $\beta = 0$ can't be stationary for the optimization problem $\mathcal{O}_{\infty,\lambda,\mathbb{Q}^A}$ unless A contains $S(\lambda)$. As a result, the algorithm does not stop at iteration k unless $\hat{S}^{(k)} \supseteq S(\lambda)$. Consequently, this implies that $\hat{S} \supseteq S(\lambda)$.

Finite Sample case: $n < \infty$. The proof is essentially the same as in the population case $n = \infty$. The major difference is that we replace the population gradient $\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q})$ by the empirical gradient $\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}_n)$ over all possible reweighting distributions $\mathbb{Q} = \mathbb{Q}^A$ where A is a subset of S (recall the algorithm does not over-select any noise variable with high probability). Our proof uses the high probability bound (Corollary J.1) to control the difference between the population and empirical gradients.

Below we give the details of the proof. We introduce two events that are of crucial importance.

- Let $\overline{\mathcal{E}}$ be the event that is stated in Theorem 3—the event on which the algorithm does not over-select noise variables. By Theorem 3, $\overline{\mathcal{E}}$ happens with probability at least $1 - 2^s(p^{-t^2} + e^{-n\varrho^2/32})$.

- Let $\mathcal{E}'$ be the event for which we have the high probability bound on the empirical and population gradients
(186)
$$\left|\frac{\partial}{\partial\beta_j}\ell(\beta;\lambda,\mathbb{Q}_n^A)\mid_{\beta=0}-\frac{\partial}{\partial\beta_j}\ell(\beta;\lambda,\mathbb{Q}^A)\mid_{\beta=0}\right|\le C\cdot\sqrt{\frac{\log p}{n}}\cdot(1+t)$$
holds simultaneously for all set $A\subseteq S$ and all variables $j\in[p]$. By Corollary J.1 and union bound, we can choose appropriate constant $C>0$ such that $\mathcal{E}'$ happens with probability at least $1-2^s(p^{-t^2}+e^{-n\varrho^2/32})$. Here $C>0$ depends only on $b,M,\varrho,f(0),f'(0),f''(0)$.

Let $\mathcal{E}=\overline{\mathcal{E}}\cap\mathcal{E}'$. Let the constant C in the definition of $S(\lambda)$ be the same constant C appeared in equation (186). Assume $\lambda>C\cdot\sqrt{\frac{\log p}{n}}\cdot(1+t)$. We show on event $\mathcal{E}$, the algorithm selects $S(\lambda)$ before it terminates. To see this, let k denote the iteration, and $A\equiv\hat{S}^{(k)}$ denote the set of variables selected by iteration k. Suppose A does not contain the set $S(\lambda)$. Now we prove $0\in\mathcal{B}$ can't be stationary with respect to the optimization $\mathcal{O}_{n,\lambda,\mathbb{Q}^A}$. Indeed, pick any signal variable $j\in S(\lambda,t)\backslash A$. By equations (185) and (186), we have

$$\begin{aligned}
&\frac{\partial}{\partial\beta_j}\ell(\beta;\lambda,\mathbb{Q}_n^A)\mid_{\beta=0}\\
&\ge\frac{\partial}{\partial\beta_j}\ell(\beta;\lambda,\mathbb{Q}^A)\mid_{\beta=0}-\left|\frac{\partial}{\partial\beta_j}\ell(\beta;\lambda,\mathbb{Q}_n^A)\mid_{\beta=0}-\frac{\partial}{\partial\beta_j}\ell(\beta;\lambda,\mathbb{Q}^A)\mid_{\beta=0}\right|\\
&=|f'(0)|\cdot\left|\mathbb{E}_{B-W}\left[\left|X_j-X_j'\right|\right]\right|-C\cdot\sqrt{\frac{\log p}{n}}\cdot(1+t)-\lambda>0.
\end{aligned}$$

where the last inequality uses the definition of $S(\lambda)$. As a result, we have shown on event $\mathcal{E}$, the algorithm does not stop at iteration k unless the set $\hat{S}^{(k)}\supseteq S(\lambda)$. In other words, on event $\mathcal{E}$, the output of the algorithm must satisfy $\hat{S}\supseteq S(\lambda)$. Since the event $\mathcal{E}$ happens with probability at least $1-2^{s+1}(p^{-t^2}+e^{-n\varrho^2/32})$, this concludes the proof for the finite case $n<\infty$.

N.2. Recovery of Main Effects under Dependence. Proposition 11 generalizes Theorem 5. It shows that metric learning can w.h.p. recover signal variables that contribute additional explanatory power after accounting for the other variables (and the explanatory power has to be significant enough in the finite sample case; cf. equation (187)).

PROPOSITION 11. *Assume (A1)-(A3). Let $q=1$. There exists some constant $C>0$ depending only on $f(0),f'(0),f''(0),M,b,\nu$ such that the following holds: for any $t>0$, any initialization $\beta^{(0)}$, and any stepsize α and*

penalty λ *satisfying*

$$\lambda \geq C \cdot \sqrt{\frac{\log p}{n}} \cdot (1+t) \quad \textit{and} \quad \alpha \leq \frac{1}{C \cdot p},$$

with probability at least $1 - 2^{s+2}(p^{-t^2} + e^{-n\varrho^2/32})$*, the metric learning algorithm outputs a set* $\hat{S}$ *such that* $S \supseteq \hat{S} \supseteq S(\lambda)$*, where*

$$S(\lambda) := \left\{ j : \min_{A \subseteq S \setminus \{j\}} \text{SIGNAL}(\{j\} \mid A) \geq 2\lambda \right\} \tag{187}$$
$$\textit{where} \quad \text{SIGNAL}(\{j\} \mid A) = f'(0) \cdot \mathbb{E}^{(A)}_{B-W} \left[\left| X_j - X'_j \right| \right]$$

N.2.1. *Proof of Proposition 11.* The proof of Proposition 11 is very similar to that of Theorem 5. Our proof strategy is essentially the same: we show $0 \in \mathcal{B}$ can't be the stationary point with high probability unless $\hat{S} \supseteq S(\lambda)$. Below we sketch out the proof.

First, we consider the population case where $n = \infty$. Let $A \equiv \hat{S}^{(k)}$, the variables being selected at iteration k. Suppose A does not contain $S(\lambda)$. Now we prove $0 \in \mathcal{B}$ can't be the stationary point (and so the algorithm does not stop). Pick a signal variable $j \in S(\lambda) \backslash A$. The derivative of the objective function $\ell(\beta; \lambda, \mathbb{Q}^A)$ with respect to variable j at $\beta = 0$ gives exactly the same equation (183) as before. Without the conditional independence assumption, we no longer have equation (184) to hold. Yet we can still derive an analogous version of equation (185) with only equation (183) at hand—the key is to exploit this new (and different) definition of $S(\lambda)$. In fact, we have for any $j \in S(\lambda)$,

$$\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A) \mid_{\beta=0} = f'(0) \cdot \mathbb{E}^{(w_A)}_{B-W} \left[\left| X_j - X'_j \right| \right] - \lambda > 0, \tag{188}$$

where the inequality holds since the definition of $S(\lambda)$. As such, we have proven the desired claim, that is, $\beta = 0$ can't be stationary unless the set of selected variables A contains the set $S(\lambda)$.

Next, we extend the result to the finite sample case where $n < \infty$. The proof is essentially the same as that appeared in the proof of Theorem 5. We omit the details for the interested readers.

N.3. Proof of Theorem 6 (Recovery of Pure Interactions). Throughout the proof, we assume W.L.O.G that for the constants λ, C (the constant C is to be determined from the proof)

$$\text{SIGNAL}(S) \geq \text{NOISE}(\lambda; C) = 2\lambda + \frac{C}{p^s}.$$

We will show with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$, the algorithm successfully selects all the pure interaction signal variables S. Together with Theorem 3, which shows the algorithm with high probability does not over-select any noise variable, we obtain Theorem 6 as desired.

Below we show the metric learning algorithm selects all the signal variables S in a single iteration, i.e., $\hat{S}^{(1)} = S$, with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$. In the first iteration, the algorithm aims to solve the optimization

$$\mathcal{O}_{n,\lambda} : \underset{\beta \in \mathcal{B}}{\text{maximize}}\, \ell(\beta; \lambda; \mathbb{Q}_n)$$

Here $\mathbb{Q}$ is a reweighted version of $\mathbb{P}$, i.e., $dQ(x, y) \propto dP(x, y) \cdot P(1-y)$. Note that the conditional distribution of $X|Y$ is the same under $\mathbb{P}$ and $\mathbb{Q}$. Yet the marginal distribution of Y is not the same. In fact, $\mathbb{Q}$ is always balanced, i.e., $\mathbb{Q}(Y = 0) = \mathbb{Q}(Y = 1) = \frac{1}{2}$, while $\mathbb{P}$ is not necessarily balanced.

Recall that we solve the optimization using projected gradient ascent: formally, with initialization $\beta^{(0)} = \frac{b}{p}\mathbf{1}$, we update

$$\begin{aligned} \beta^{(m+1)} &= \Pi_{\mathcal{B}}\left(\beta^{(m+1/2)}\right) \\ \text{where} \quad \beta^{(m+1/2)} &= \beta^{(m)} + \alpha \cdot \nabla \ell(\beta^{(m)}; \lambda; \mathbb{Q}_n). \end{aligned}$$

The key to the proof is to establish Proposition 12, whose proof is pretty technical, and is deferred to Section N.8.

PROPOSITION 12. *Assume the same assumption as in Theorem 6. Assume the constant $C > 0$ stated in Theorem 6 is large enough. Then, we have with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$, the gradient iterates $\beta^{(m)}$ satisfy*

$$\min_{i \in S} \beta_i^{(m)} \geq \zeta \equiv \frac{b}{8p} > 0 \quad \textit{for all } m \in \mathbb{N}. \tag{189}$$

This proposition essentially says that with high probability at any iteration, the coordinates of the gradient iterates $\beta_S^{(m)}$ is bounded away from 0. As a result, this means that with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$, any accumulation point $\beta^{(*)}$ of the iterates must satisfy $\beta_i^{(*)} \neq 0$ for all $i \in S$. Thus, the metric learning algorithm must select the entire signal set S in its first iteration. This completes the proof.

N.4. Tradeoffs between statistical and computational complexity in interaction search. The sample size requirement, $n \sim p^{2(s-1)} \log p$, for the recovery of a pure order s interaction is quite high (see Theorem 6).

We emphasize though that this is the sample size requirement when we restrict ourselves to a computation budget of $O(p)$. If we are willing to increase our computational budget—by trying many different initial starting points when maximizing $F_n(\beta) - \lambda|\beta|_1$—we can reduce the necessary sample size. For example consider searching for pairwise interactions ($s = 2$) but suppose we only have $n \sim p \log p$ samples. To guarantee high probability of finding a pure 2-way interaction (assuming it exists), we proceed by partitioning $\{1, \ldots, p\}$ into $\sqrt{p}$ subsets of size $\sqrt{p}$: $\{A_l\}_{l=1}^{\sqrt{p}}$. For every pair of partitions, A_l and A_m, we perform metric screening using only the variables in A_l and A_m. Since there are only $2\sqrt{p}$ variables in A_l and A_m, we initialize at $\beta_j = \frac{1}{2\sqrt{p}}$. By Theorem 6, if the pure 2-way interaction is among any of the variables in $A_l \cup A_m$, we will recover it with high probability given $n \sim p \log p$ samples. But by construction, every pair of variables (j, k) appears in one of the A_l, A_m pairs. The computation cost of running metric screening on each pair A_l, A_m is $O(\sqrt{p})$ and since there are p pairs of partitions, the overall cost is $p^{3/2}$. So by increasing our computation budget from p to $p^{3/2}$, we can find a 2-way interaction with high probability using only $n \sim p \log p$ samples. This computation is cheaper than an exhaustive search which costs p^2. In general, if we dedicate a budget of p^{s_0} towards finding a pure order s interaction, the sample size requirement is $p^{2(s-s_0)_+} \log p$.

Algorithm 1 Hierarchical Metric Screening

Require: $\lambda \geq 0$, $\tau > 0$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $y \in \{0, 1\}^n$

Ensure: Initialize $\hat{S} = \emptyset$ and the weight w by

$$w_i = \begin{cases} \#\{y_i = 0\}/n & \text{if } \ y_i = 1 \\ \#\{y_i = 1\}/n & \text{if } \ y_i = 0 \end{cases} \qquad \text{for } i = 1, 2, \ldots, n.$$

1: **while** $\hat{S}$ not converged **do**

2: Initialize $\beta_j = \tau$ for $j \in \hat{S}$ and $\beta_j = 0$ for $j \notin \hat{S}$

3: Run projected gradient ascent (with stepsize α, and initialization $\beta^{(0)}$) to solve

$$\max_{\beta \in \mathcal{B}: \beta_{\hat{S}} = \tau \mathbf{1}_{\hat{S}}} F_n(\beta; w) - \lambda \cdot |\beta|_1.$$

Update $\hat{S} = \hat{S} \cup \text{supp}(\beta)$ where β is any stationary point found by the iterates[2].

4: Estimate $\mathbb{P}(Y|X_{\hat{S}})$ and update the weight w by

$$w_i \propto \begin{cases} \widehat{\mathbb{P}}(Y = 0|x_{i,\hat{S}}) & \text{if } y_i = 1 \\ \widehat{\mathbb{P}}(Y = 1|x_{i,\hat{S}}) & \text{if } y_i = 0 \end{cases} \quad \text{for } i = 1, 2, \ldots, n.$$

5: **end while**

N.5. Proof of Theorem 7 (Recovery of Hierarchical Interactions). We will show with probability at least $1 - 2^s(p^{-t^2} + e^{-n\varrho^2/32})$, the algorithm (Algorithm 1) successfully selects all the signal variables within the set $S(\lambda, \tau) \equiv S_{l(\lambda,\tau)}$ before it terminates. Together with Theorem 3, which shows the algorithm with high probability does not over-select any noise variable, we obtain Theorem 7 as desired.

Below we show the algorithm selects all variables within $S(\lambda, \tau)$ with probability at least $1 - 2^s(p^{-t^2} + e^{-n\varrho^2/32})$. Denote the optimization problem

$$\begin{aligned} \mathcal{O}_{n,\lambda,\mathbb{Q};A,\tau} : \underset{\beta\in\mathcal{B}}{\text{maximize}} \quad & \ell(\beta;\lambda;\mathbb{Q}_n) \\ \text{subject to} \quad & \beta_A = \tau \mathbf{1}_A. \end{aligned}$$

For any $A \subseteq [p]$, solving $\mathcal{O}_{n,\lambda,\mathbb{Q};A,\tau}$ for $\mathbb{Q} = \mathbb{Q}^A$ with gradient ascent will always return a stationary point of $\mathcal{O}_{n,\lambda,\mathbb{Q};A,\tau}$ when the stepsize $\alpha \le 1/(Cp)$ where $C = |f''(0)|\overline{M}$ and $\overline{M} = 2M$ (thanks to Lemma K.1 and O.4). As a result, the hierarchical metric learning algorithm iteratively finds a stationary point $\beta_{\hat{S}}$ of the optimization $\mathcal{O}_{n,\lambda,\mathbb{Q}^{w_{\hat{S}}},\hat{S},\tau}$ and updates $\hat{S}$ by $\hat{S} \cup \text{supp}(\beta_{\hat{S}})$. The algorithm terminates when $\text{supp}(\beta_{\hat{S}}) \subseteq \hat{S}$.

The key to the proof is show the following claim: denoting $\beta_0^{\tau,A} \in \mathbb{R}^p$ to be the vector whose coordinate is τ in A, and 0 in A^c, i.e.,

$$(\beta_0^{\tau,A})_i = \begin{cases} \tau & \text{if } i \in A \\ 0 & \text{if } i \notin A \end{cases}$$

then with probability at least $1-2^s(p^{-t^2}+e^{-n\varrho^2/32})$, $\beta_0^{\tau,A}$ can't be stationary with respect to the optimization problem $\mathcal{O}_{n,\lambda,\mathbb{Q}^A;A,\tau}$ for any set $A \subseteq S$ that does not contain $S(\lambda, \tau)$. In particular, with the same probability, any stationary point β of $\mathcal{O}_{n,\lambda,\mathbb{Q}^A;A,\tau}$ will satisfy $\text{supp}(\beta)\backslash A \neq \emptyset$. Together with Theorem 3, which shows the algorithm does not over-select any noise variable with high probability, this means the algorithm will not terminate unless $\hat{S}$ contains $S(\lambda, \tau)$. Below we prove this claim. Our strategy is to first show the claim holds on population ($n = \infty$), and then extends it to finite case ($n < \infty$) using standard concentration and perturbation techniques.

Population case: $n = \infty$. Let $A \subseteq S$ be any set that does not contain $S(\lambda, \tau)$. We show $\beta_0^{\tau,A}$ can't be stationary of the problem $\mathcal{O}_{\infty,\lambda,\mathbb{Q}^A,A,\tau}$.

[2]Technically, β is defined to be any accumulation point of the iterates since there is no prior knowledge that the algorithm will converge to a stationary point. Technically, it is this definition that's used in the proof.

Let l be the largest integer such that $S_{l-1} \subseteq A$. Let $j \in [p]$ be such that $\{j\} = S_l \backslash S_{l-1}$. Note $j \in S(\lambda, \tau)$ since A does not contain $S(\lambda, \tau)$. We compute the derivative of $\ell(\beta; \lambda, \mathbb{Q}^A)$ with respect to variable j at $\beta = \beta_0^{\tau,A}$:

$$\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A) \mid_{\beta=\beta_0^{\tau,A}} = \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}^A) \mid_{\beta=\beta_0^{\tau,A}} - \lambda \tag{190}$$

Now we show the above partial gradient is positive. Note the following two important observations. First, the rebalancing procedure makes $X_A \perp Y$ after rebalancing by Proposition 3. Second, we have $w_A = w_{S_{j-1}}$ since $\mathbb{P}(Y \mid X_A) = \mathbb{P}(Y \mid X_{S_{j-1}})$ by definition of hierarchical interaction. Let $\beta_{0;j}^{\tau,A}$ be the vector whose coordinate is τ in $A \cup \{j\}$, and 0 in $(A \cup \{j\})^c$. In view of these two observations, Lemma I.4 immediately implies

$$\begin{aligned} \frac{\partial}{\partial \beta_j} F(\beta; \mathbb{Q}^A) \mid_{\beta=\beta_0^{\tau,A}} &\geq \frac{1}{\tau} \cdot F(\beta; \mathbb{Q}^A) \mid_{\beta=\beta_{0;j}^{\tau,A}} \\ &= \frac{1}{\tau} \cdot \mathbb{E}_{B-W}^{w_A} \left[f\big(\tau \cdot \big\| X_{A\cup\{j\}} - X'_{A\cup\{j\}} \big\|_1\big) \right] \\ &= \frac{1}{\tau} \cdot \mathbb{E}_{B-W}^{w_{S_{j-1}}} \left[f\big(\tau \cdot \big\| X_{A\cup\{j\}} - X'_{A\cup\{j\}} \big\|_1\big) \right], \end{aligned} \tag{191}$$

Substitute the bound (191) into equation (190), we obtain

$$\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}^A) \mid_{\beta=\beta_0^{\tau,A}} \geq \frac{1}{\tau} \cdot \mathbb{E}_{B-W}^{w_{S_{j-1}}} \left[f\big(\tau \cdot \big\| X_{A\cup\{j\}} - X'_{A\cup\{j\}} \big\|_1\big) \right] - \lambda > 0, \tag{192}$$

where the last inequality is due to the fact that $j \in S(\lambda, \tau)$. As a consequence, increasing β_j at $\beta_0^{\tau,A}$ would increase the objective. Hence $\beta_0^{\tau,A}$ isn't a stationary point with respect to the optimization $\mathcal{O}_{\infty,\lambda,\mathbb{Q}^A,A,\tau}$.

Finite Sample case: $n < \infty$. The proof is essentially the same as in the population case $n = \infty$. The major difference is that we replace the population gradient $\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q})$ by the empirical gradient $\frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}_n)$ over all possible reweighting distributions $\mathbb{Q} = \mathbb{Q}^A$ where A is a subset of S (recall the algorithm does not over-select any noise variable with high probability). The technique we use is the high probability bound (cf. Corollary J.1) that controls the difference between the population and empirical gradients.

Below we give the details of the proof. We introduce two events that are of crucial importance.

1. Let $\overline{\mathcal{E}}$ be the event that is stated in Theorem 3—the event on which the algorithm does not over-select noise variables. By Theorem 3, $\overline{\mathcal{E}}$ happens with probability at least $1 - 2^s(p^{-t^2} + e^{-n\varrho^2/32})$.

2. Let $\mathcal{E}'$ be the event for which we have the high probability bound on the empirical and population gradients

$$\left|\frac{\partial}{\partial\beta_j}\ell(\beta;\lambda,\mathbb{Q}_n^A)-\frac{\partial}{\partial\beta_j}\ell(\beta;\lambda,\mathbb{Q}^A)\right|\le C\cdot\sqrt{\frac{\log p}{n}}\cdot(1+t) \tag{193}$$

holds simultaneously for all $\beta\in\mathcal{B}$, all set $A\subseteq S$ and all variables $j\in[p]$. By Corollary J.1 and union bound, we know for some large constant $C>0$ that depends only on $b,M,\nu,f(0),f'(0),f''(0)$, the event $\mathcal{E}'$ happens with probability at least $1-2^s(p^{-t^2}+e^{-n\varrho^2/32})$.

Now, let $\mathcal{E}=\overline{\mathcal{E}}\cap\mathcal{E}'$. Let the constant C in the definition of $S(\lambda,\tau)$ be the same constant C appeared in equation (186). Assume $\lambda>C\cdot\sqrt{\frac{\log p}{n}}\cdot(1+t)$. We show on event $\mathcal{E}$, $\beta_0^{\tau,A}$ can't be stationary of the problem $\mathcal{O}_{\infty,\lambda,\mathbb{Q}^A,A,\tau}$ for any set $A\subseteq S$ that does not contain $S(\lambda,\tau)$.

To see this, let A be any set that does not contain $S(\lambda,\tau)$. Let l be the largest integer such that $S_{l-1}\subseteq A$. Let j be such that $\{j\}=S_l\backslash S_{l-1}$. By equations (192) and (193), we have on event $\mathcal{E}$,

$$\begin{aligned}
&\frac{\partial}{\partial\beta_j}\ell(\beta;\lambda,\mathbb{Q}_n^A)\mid_{\beta=\beta_0^{\tau,A}}\\
&\ge\frac{\partial}{\partial\beta_j}\ell(\beta;\lambda,\mathbb{Q}^A)\mid_{\beta=\beta_0^{\tau,A}}-\left|\frac{\partial}{\partial\beta_j}\ell(\beta;\lambda,\mathbb{Q}_n^A)\mid_{\beta=\beta_0^{\tau,A}}-\frac{\partial}{\partial\beta_j}\ell(\beta;\lambda,\mathbb{Q}^A)\mid_{\beta=\beta_0^{\tau,A}}\right|\\
&\ge\frac{1}{\tau}\cdot\mathbb{E}_{B-W}^{w_{S_{j-1}}}\left[f\left(\tau\cdot\left\|X_{A\cup\{j\}}-X'_{A\cup\{j\}}\right\|_1\right)\right]-\lambda-C\sqrt{\frac{\log p}{n}}\cdot(1+t)>0
\end{aligned}$$

where the last inequality uses the definition of $S(\lambda,\tau)$ and the fact that $j\in S(\lambda,\tau)$. As a result, we have shown on event $\mathcal{E}$, $\beta_0^{\tau,A}$ can't be stationary of the problem $\mathcal{O}_{\infty,\lambda,\mathbb{Q}^A,A,\tau}$ for any set $A\subseteq S$ that does not contain $S(\lambda,\tau)$. Note the event $\mathcal{E}$ happens with probability at least $1-2^{s+1}(p^{-t^2}+e^{-n\varrho^2/32})$ by the union bound. This concludes the proof for the finite case $n<\infty$.

N.6. Landscape of $F(\beta;\mathbb{Q})$ Under a Hierarchical Interaction. Hierarchical signals can be tricky to recover[3]. To see why, consider the following situation. X_1 and X_2 are involved in a hierarchical interaction where $X_1\not\perp Y$ and $X_2\perp Y$. In the first round of metric screening, suppose we select X_1 but not X_2. After rebalancing, we know $X_1\perp Y$. As the following example demonstrates, it is possible to simultaneously have $X_2\perp Y$ after rebalancing. In this case, X_1 and X_2 are involved in a *pure* interaction after rebalancing and we know that pure interactions *cannot* be detected with

[3]Their recovery is not guaranteed by either Theorem 5 or Corollary 11.

$n \sim \log p$ samples. To bypass this difficulty, the modified Algorithm 1 is used to recover hierarchical interactions.

Example: Consider the following model: assume Y is balanced

$$\mathbb{P}(Y=1) = \mathbb{P}(Y=0) = \frac{1}{2}.$$

The first signal X_1 has its conditional distribution $X_1|Y$ given by

$$\mathbb{P}(X_1 = \pm\frac{1}{2} \mid Y=1) = \frac{1}{2}(1 \pm \delta_1),$$
$$\mathbb{P}(X_1 = \pm\frac{1}{2} \mid Y=0) = \frac{1}{2}(1 \mp \delta_1).$$

The second signal X_2 has its conditional distribution $X_2|X_1, Y$ given by

$$\mathbb{P}\left(X_2 = \pm\frac{1}{2} \mid X_1 = +\frac{1}{2}, Y=1\right) = \frac{1}{2}(1 \pm p(\delta_1, \delta_{12}))$$
$$\mathbb{P}\left(X_2 = \pm\frac{1}{2} \mid X_1 = +\frac{1}{2}, Y=0\right) = \frac{1}{2}(1 \mp \delta_{1,2})$$
$$\mathbb{P}\left(X_2 = \pm\frac{1}{2} \mid X_1 = -\frac{1}{2}, Y=1\right) = \frac{1}{2}(1 \mp \delta_{12})$$
$$\mathbb{P}\left(X_2 = \pm\frac{1}{2} \mid X_1 = -\frac{1}{2}, Y=0\right) = \frac{1}{2}(1 \mp p(\delta_1, \delta_{12}))$$

where we adopt the parametrization

$$p(\delta_1, \delta_{12}) = \frac{1}{2} \cdot \delta_{12} \cdot \frac{1-\delta_1}{1+\delta_1}.$$

The strength of the marginal signal between X_1 and Y is captured by $\delta_1 > 0$ and the strength of the interaction signal by $\delta_{12} > 0$. When $\delta_1 = 0$ and $\delta_{12} = 1$ this reduces to the XOR signal, a pure interaction.

The landscape of $F(\beta; \mathbb{Q}^{(0)})$ under this statistical model has the following characteristic:

- If $\delta_1 > \frac{1}{2}$, $F(\beta; \mathbb{Q}^{(0)})$ has a single maximum at $\beta = (\infty, 0)$[4]; this is also the case if $\delta_1 < \frac{1}{2}$ but $\delta_1 > \delta_{12}$. The main effect will mask the interaction and only X_1 will be selected in the first round of metric screening.

[4]This means that $(r, 0)$ is the maximum of $F(\beta; \mathbb{Q}^{(0)})$ over $\mathcal{D}_r$ for all sufficiently large r. Here $\mathcal{D}_r = \{(\beta_1, \beta_2) : 0 \le \beta_i \le r\}$.

- If $\delta_1 < \frac{1}{2}$ and $\delta_1 < \delta_{12}$, $F(\beta; \mathbb{Q}^{(0)})$ has a single local maximum at $\beta = (\infty, \infty)$[5]. Both X_1 and X_2 would be selected in this case.

When the main effect is strong, the first round of metric screening selects X_1 but not X_2. One can show that after rebalancing, $X_1 \perp Y$ and $X_2 \perp Y$. Therefore, in the second round of metric screening, the signal is a pure interaction. In particular, $F(\beta; \mathbb{Q}^{(1)})$ will have one stationary point at $(0, 0)$ (and this stationary point will be with high probability a local maximum in the empirical penalized objective $\ell(\beta; \lambda, \mathbb{Q}_n^{(1)}) = F(\beta; \mathbb{Q}_n^{(1)}) - \lambda \|\beta\|_1$). An algorithmic challenge to evade this bad stationary point thus appears. ♣

N.7. Proof of Proposition 6.

Proof of Part 1. The proof builds on the fundamental result in Lemma I.1–Lemma I.3. Below we fix $\mathbb{P}_0$ and the sequence $\{\mathbb{Q}_p\}_{p\in\mathbb{N}}$ where $\mathbb{Q}_p \in \mathcal{Q}(p, M, S, \mathbb{P}_0)$. In the proof, we omit the dependence of F on f, q (so for example, we use the notation $F(\bar\beta; \mathbb{Q}_p)$ to refer to $F(\bar\beta; \mathbb{Q}_p, f, 1)$).

- Consider first the objective value $F(\bar\beta; \mathbb{Q}_p)$. Lemma I.3 shows

$$F(\bar\beta; \mathbb{Q}_p) \geq \frac{f^{|S|}(2Mb)}{f^{|S|}(0)} \cdot F_S(\bar\beta_S; \mathbb{Q}_p), \tag{194}$$

where $F_S(\bar\beta_S; \mathbb{Q}_p) = \mathbb{E}_{B-W}\Big[f(\|X_S - X_S'\|_{1,\bar\beta_S})\Big]$. By Lemma I.1, F_S satisfies the self-bounding property, i.e., for any β, β' satisfies $\beta_i \leq \beta_i'$,

$$F_S(\beta_S; \mathbb{Q}_p) \geq F_S(\beta_S'; \mathbb{Q}_p) \cdot \prod_{i\in S} \left(\frac{\beta_i}{\beta_i'}\right).$$

In particular, if $p \geq b$ (so that $\bar\beta_i \leq 1$ for all $i \in [p]$), then we have

$$F_S(\bar\beta_S; \mathbb{Q}_p) \geq \left(\prod_{i\in S} \bar\beta_i\right) \cdot F_S(\mathbf{1}_S; \mathbb{Q}_p) \tag{195}$$

Note then $F_S(\mathbf{1}_S; \mathbb{Q}_p) = F_S(\mathbf{1}_S; \mathbb{P}_0)$ since the marginal distribution of $(X_S, Y) \sim \mathbb{P}_0$ when $(X, Y) \sim \mathbb{Q}_p$. As a result, we obtain for $p \geq b$,

$$F(\bar\beta; \mathbb{Q}_p) \geq \frac{f^{|S|}(2Mb)}{f^{|S|}(0)} \cdot F_S(\mathbf{1}_S; \mathbb{P}_0) \cdot \prod_{i\in S} \bar\beta_i. \tag{196}$$

[5]This means that (r, r) is the maximum of $F(\beta; \mathbb{Q}^{(0)})$ over $\mathcal{D}_r$ for all sufficiently large r. Here $\mathcal{D}_r = \{(\beta_1, \beta_2) : 0 \leq \beta_i \leq r\}$.

Recall $\bar{\beta}_i = t/p$ for all $i \in S$. Equation (196) immediately implies

$$F(\bar{\beta}; \mathbb{Q}_p) \geq c \cdot p^{-|S|} \quad \text{for all } p \geq b,$$

where the constant $c = |f^{|S|}(2Mb)/f^{|S|}(0)| \cdot F_S(\mathbf{1}_S; \mathbb{P}_0) \cdot t^{|S|} > 0$ depends only on $f, M, b, \mathbb{P}_0, S, t$. This proves the first line of equation (9).

- Consider next the gradient $\frac{\partial}{\partial \beta_j} F(\bar{\beta}; \mathbb{Q}_p)$ where $j \in S$. Lemma I.2 gives

$$\frac{\partial}{\partial \beta_j} F(\bar{\beta}; \mathbb{Q}_p) = \bar{\beta}_j^{-1} \cdot F(\bar{\beta}; \mathbb{Q}_p) - R(\bar{\beta}; \mathbb{Q}_p), \tag{197}$$

where the remainder term $R(\bar{\beta}; \mathbb{Q}_p)$ satisfies

$$0 \leq R(\bar{\beta}; \mathbb{Q}_p) \leq \pi \cdot (8M)^{|S|+1} \cdot f^{(|S|+1)}(0) \cdot \prod_{k \in S} \bar{\beta}_k. \tag{198}$$

Substitute equations (196) and (198) into equation (197). This gives

$$\frac{\partial}{\partial \beta_j} F(\bar{\beta}; \mathbb{Q}_p) \geq \prod_{i \in S, i \neq j} \bar{\beta}_i \cdot \left(c' - C \cdot \bar{\beta}_j \right). \tag{199}$$

where $c' = \frac{f^{|S|}(2Mb)}{f^{|S|}(0)} \cdot F(\mathbf{1}_S; \mathbb{P}_0) > 0$ and $C = \pi \cdot (8M)^{|S|+1} \cdot f^{(|S|+1)}(0) > 0$ are constants that depend only on $f, M, b, \mathbb{P}_0, S, t$. Recall $\bar{\beta}_i = t/p$ for all $i \in S$. Hence, equation (199) implies that for all large enough p:

$$\frac{\partial}{\partial \beta_j} F(\bar{\beta}; \mathbb{Q}_p) \geq \bar{c} \cdot p^{-(|S|-1)}$$

where $\bar{c} = \frac{1}{2} c' \cdot t^{|S|} = \frac{1}{2} \frac{f^{|S|}(2Mb)}{f^{|S|}(0)} \cdot F(\mathbf{1}_S; \mathbb{P}_0) \cdot t^{|S|} > 0$. Again, $\bar{c}$ depends only on $f, M, b, \mathbb{P}_0, S, t$. This proves the second line of equation (9).

Proof of Part 2. Fix $l \in \mathbb{N}$. W.L.O.G we assume $l \geq 2$. We construct the distribution $\mathbb{P}_0$ as follows.

1. First, we set the marginal distribution of Y to be $\mathbb{P}_0(Y = \pm 1) = \frac{1}{2}$.
2. Next, we set the conditional distribution of X_S given Y. More precisely, we let $X_S \mid Y = 1 \sim P_{+1}$ and $X_S \mid Y = -1 \sim P_{-1}$ where $\mathbb{P}_+$ and $\mathbb{P}_-$ are any two distinct distributions supported on $[-M, M]^{|S|}$ that satisfy
   - For any strict subset $A \subsetneq S$, the distribution of X_A under $\mathbb{P}_{+1}$ is the same as the distribution of X_A under $\mathbb{P}_{-1}$.
   - $\mathbb{P}_{+1}$ and $\mathbb{P}_{-1}$ have the same moments up to $2l$-th order, i.e., for any $\alpha_S \in \mathbb{N}^{|S|}$ with $\sum_{i \in S} \alpha_i \leq 2l$, we have

$$\mathbb{E}_{+1} \left[\prod_{i \in S} X_i^{\alpha_i} \right] = \mathbb{E}_{-1} \left[\prod_{i \in S} X_i^{\alpha_i} \right].$$

The existence of such distributions $\mathbb{P}_{+1}, \mathbb{P}_{-1}$ is proven in Lemma O.5.

By construction, it is easy to see that $\mathbb{P}_0$ exhibits a pure interaction: $X_A \perp Y$ for all strict subset $A \subsetneq S$ under $\mathbb{P}_0$. Fix this $\mathbb{P}_0$. Pick any sequence $\{\mathbb{Q}_p\}_{p\in\mathbb{N}}$ where $\mathbb{Q}_p \in \mathcal{Q}(p, M, S, \mathbb{P}_0)$. We show that when $q = 2$, the objective function $F(\beta; \mathbb{Q}_p)$ satisfies the following property:

- There exists a constant $C_1 > 0$ depending only on f, M, b, t, l such that

$$F(\bar{\beta}; \mathbb{Q}_p, f, 2) \le C_1 \cdot p^{-l} \text{ for all } p \in \mathbb{N}. \tag{200}$$

- There exists a constant $C_2 > 0$ depending only on f, M, b, t, l such that

$$\frac{\partial}{\partial \beta_j} F(\bar{\beta}; \mathbb{Q}_p, f, 2) \le C_2 \cdot p^{-l} \text{ for all } j \in S,\ p \in \mathbb{N}. \tag{201}$$

The proof of the statements (200) and (201) are simlar. For space considerations, we only detail the proof of equation (200) below.

The key idea to the proof is to perform a careful Taylor expansion. For notational simplicity, we denote $Z_{\bar{\beta}}(X) = \|X - X'\|^2_{2,\bar{\beta}}$. By definition,

$$F(\bar{\beta}; \mathbb{Q}_p, f, 2) = \mathbb{E}_{B-W}\left[f\big(Z_{\bar{\beta}}(X)\big)\right] \tag{202}$$

Clearly we can expand $Z_{\bar{\beta}}(X) = Z_{\bar{\beta}}(X_S) + Z_{\bar{\beta}}(X_{S^c})$ where

$$Z_{\bar{\beta}}(X_S) = \left\|X_S - X'_S\right\|^2_{2,\bar{\beta}_S} \text{ and } Z_{\bar{\beta}}(X_{S^c}) = \left\|X_{S^c} - X'_{S^c}\right\|^2_{2,\bar{\beta}_{S^c}}.$$

Recall the following version of Taylor's intermediate theorem. For any smooth function $g \in \mathcal{C}^\infty[0,\infty)$, any scalars $c \in \mathbb{R}, x \in \mathbb{R}_+$ and any $l \in \mathbb{N}$,

$$\left|g(c+x) - \sum_{k<l} \frac{g^{(k)}(c)}{k!} x^k\right| \le \frac{1}{l!} \cdot \sup_{\xi\in[c,c+x]} |g^{(l)}(\xi)| \cdot x^l.$$

By specifying $g \equiv f$, $c = Z_{\bar{\beta}}(X_S)$ and $x = Z_{\bar{\beta}}(X_{S^c})$, and using the fact that $\sup_{x\in\mathbb{R}} |f^{(l)}(x)| = |f^{(l)}(0)|$ (since f' is completely monotone), we get

$$\left|f\big(Z_{\bar{\beta}}(X)\big) - H_{\bar{\beta}}(X)\right| \le R_{\bar{\beta}}(X). \tag{203}$$

In above, $H_{\bar{\beta}}(X)$ and $R_{\bar{\beta}}(X)$ are defined by

$$H_{\bar{\beta}}(X) = \sum_{k<l} \frac{1}{k!} \cdot f^{(k)}(Z_{\bar{\beta}}(X_{S^c})) \cdot Z_{\bar{\beta}}(X_S)^k$$

$$R_{\bar{\beta}}(X) = \frac{1}{l!} \cdot |f^{(l)}(0)| \cdot |Z_{\bar{\beta}}(X_S)|^l$$

Using equations (202), (203) and triangle inequality, we immediately obtain

$$\left|F(\bar{\beta};\mathbb{Q}_p,f,2)-\mathbb{E}_{B-W}\left[H_{\bar{\beta}}(X)\right]\right|\leq \mathbb{E}_B\left[R_{\bar{\beta}}(X)\right]+\mathbb{E}_W\left[R_{\bar{\beta}}(X)\right]. \tag{204}$$

Here comes the two crucial observations.

- Under the distribution $\mathbb{Q}_p$, we have that

  $$\mathbb{E}_{B-W}\left[H_{\bar{\beta}}(X)\right]=0. \tag{205}$$

  Indeed, as $(X_S,Y)\perp X_{S^c}$ since S is the signal set, we obtain

  $$\mathbb{E}_{B-W}\left[H_{\bar{\beta}}(X)\right]=\frac{1}{k!}\cdot\mathbb{E}\left[f^{(k)}(Z_{\bar{\beta}}(X_{S^c}))\right]\cdot\mathbb{E}_{B-W}\left[Z_{\bar{\beta}}(X_S)^k\right] \tag{206}$$

  Now that $Z_{\bar{\beta}}(X_S)^k$ is a polynomial of X_S of order at most $2k\leq 2l$. Moreover, by construction of $\mathbb{P}_0$, the conditional distribution of X_S given $Y=1$ and of X_S given $Y=-1$ have the same moments up to order $2l$. As a result, this shows that for all $k\leq l$,

  $$\mathbb{E}_{B-W}\left[Z_{\bar{\beta}}(X_S)^k\right]=0. \tag{207}$$

  Now equations (206) and equation (207) yield the desired (205).
- Under the distribution $\mathbb{Q}_p$, we have the following bound on the RHS of equation (204): for the constant $C=\frac{2}{l!}\cdot|f^{(l)}(0)|\cdot(2|S|Mt)^l$,

  $$\mathbb{E}_B\left[R_{\bar{\beta}}(X)\right]+\mathbb{E}_W\left[R_{\bar{\beta}}(X)\right]\leq C\cdot p^{-l}. \tag{208}$$

  By construction, $\|X\|_\infty\leq M$. Thus we have almost surely

  $$0\leq Z_{\bar{\beta}}(X_S)=\frac{t}{p}\left\|X_S-X'_S\right\|_2^2\leq 2M|S|\cdot\frac{t}{p}.$$

  Recall the definition $R_{\bar{\beta}}(X)=\frac{|f^{(l)}(0)|}{l!}\cdot|Z_{\bar{\beta}}(X_S)|^l$. This proves that we have almost surely

  $$0\leq R_{\bar{\beta}}(X)\leq\frac{1}{l!}\cdot|f^{(l)}(0)|\cdot(2|S|Mt)^l\cdot p^{-l}.$$

  This desired equation (208) now thus follows.

Recall equation (204). The previous observations leads to the bound

$$\left|F(\bar{\beta};\mathbb{Q}_p,f,2)\right|\leq C\cdot p^{-l}$$

for the constant $C=\frac{2}{l!}\cdot|f^{(l)}(0)|\cdot(2|S|Mt)^l$ that depends only $f,M,b,|S|,t,l$. Hence, we have shown the desired statement at equation (200). As discussed before, the proof of the statement at equation (201) is essentially the same, and is thus omitted for space consideration.

N.8. Proof of Proposition 12. We first prove Proposition 12 holds on population ($n = \infty$), and then prove this also holds in finite sample ($n < \infty$) using standard concentration and perturbation techniques.

N.8.1. *Population case:* $n = \infty$. We prove via induction that the following inequality holds for all $m \in \mathbb{N}$

$$\min_{i \in S} \beta_i^{(m)} \geq \zeta \quad \text{and} \quad \max_{i \in S^c} \left(\beta_i^{(m)} - \beta_i^{(m-1)} \right) \leq 0. \tag{209}$$

Clearly, this implies Proposition 12 as desired.

The base case $m = 0$ trivially holds since $\beta^{(0)} = \frac{b}{p}\mathbf{1}$.

Suppose the induction hypothesis, i.e., equation (209) holds for m. Below we prove that it also holds for $m + 1$. We proceed our proof in two steps.

- In the first step, we prove that

  $$\max_{i \in S^c} \left(\beta_i^{(m+1)} - \beta_i^{(m)} \right) \leq 0. \tag{210}$$

  The key here is Proposition 3, which shows the gradient with respect to noise variables is non-positive, i.e., for $\beta \in \mathcal{B}$ and $i \in S^c$,

  $$\frac{\partial}{\partial \beta_i} \ell(\beta; \lambda, \mathbb{Q}) \leq 0. \tag{211}$$

  Hence, we derive for $i \in S^c$,

  $$\beta_i^{(m+1/2)} = \beta_i^{(m)} + \alpha \cdot \frac{\partial}{\partial \beta_i} \ell(\beta; \lambda, \mathbb{Q}) \leq \beta_i^{(m)}. \tag{212}$$

  To pass the bound from the intermediate $\beta^{(m+1/2)}$ to the final iterate $\beta^{(m+1)}$, we use the projection Lemma O.2. According to Lemma O.2, there exists a nonnegative scalar $\gamma^{(m)} \geq 0$ such that for $i \in S^c$

  $$\beta_i^{(m+1)} = \Pi_{\mathcal{B}}\left(\beta_i^{(m+1/2)} \right) = \left(\beta_i^{(m+1/2)} - \gamma^{(m)} \right)_+.$$

  Using this identity, we obtain immediately that: for $i \in S^c$,

  $$\beta_i^{(m+1)} \leq \left(\beta_i^{(m+1/2)} \right)_+ \leq \beta_i^{(m)}.$$

  This gives equation (210) as desired. We finish the first step.
- In the second step, we prove that

  $$\min_{i \in S} \beta_i^{(m+1)} \geq \zeta. \tag{213}$$

  Fix $i \in S$. We need to prove $\beta_i^{(m+1)} \geq \zeta$. Below we divide our discussion into two cases based on the size of $\beta_i^{(m)}$. Let $\Delta \equiv \frac{1}{2s}$.

(a) First, we consider the scenario where $\beta_i^{(m)} \geq \zeta(1+\Delta)$. The key observation is that the gradient $\nabla \ell(\beta;\lambda;w)$ is bounded in ℓ_∞ norm. Indeed, using $\|X\|_\infty \leq M$ and $\sup_x |f'(x)| \leq |f'(0)|$,

$$\begin{aligned} &\|\nabla \ell(\beta;\lambda;w)\|_\infty \\ &= \Big\| \mathbb{E}^w_{B-W}[|X_i - X_i'| \cdot f'(\|X - X'\|_{\beta,1})] \Big\|_\infty \leq \overline{M}|f'(0)| \end{aligned} \tag{214}$$

where $\overline{M} = 2M$. After all, we immediately obtain

$$\big\|\beta^{(m+1/2)} - \beta^{(m)}\big\|_\infty = \alpha \big\|\nabla \ell(\beta;\lambda;w)\big\|_\infty \leq \alpha \overline{M}|f'(0)|. \tag{215}$$

Now we pass this bound from $\beta^{(m+1/2)}$ to $\beta^{(m+1)}$. To do so, we use the projection Lemma O.3. By Lemma O.3, we have

$$\big\|\beta^{(m+1)} - \beta^{(m)}\big\|_\infty \leq 2 \cdot \big\|\beta^{(m+1/2)} - \beta^{(m)}\big\|_\infty. \tag{216}$$

With equations (215) and (216), we use triangle inequality to get

$$\beta_i^{(m+1)} \geq \beta_i^{(m)} - \big\|\beta^{(m+1)} - \beta^{(m)}\big\|_\infty \geq \beta_i^{(m)} - 2\alpha \overline{M}|f'(0)|.$$

Note $\beta_i^{(m)} \geq \zeta(1+\Delta)$ by assumption. Hence, for C large enough such that $2\alpha M|f'(0)| \leq \zeta\Delta$ (recall $\alpha \leq \frac{1}{C(p \cdot s)}$), we have

$$\beta_i^{(m+1)} \geq \zeta.$$

This proves that equation (213) holds in the first scenario.

(b) Next, we consider the scenario where $\zeta(1+\Delta) \geq \beta_i^{(m)} \geq \zeta$. The key here is the gradient $\frac{\partial}{\partial \beta_i}\ell(\beta;\lambda;w)$ is strictly positive which is due to $i \in S$. To see this, Lemma I.2 first shows for $i \in S$,

$$\begin{aligned} \frac{\partial}{\partial \beta_i}\ell(\beta;\lambda,\mathbb{Q}) &= \frac{\partial}{\partial \beta_i}F(\beta;\mathbb{Q}) - \lambda \\ &= \frac{1}{\beta_i} \cdot F(\beta;\mathbb{Q}) - R_{i,\infty}(\beta;\mathbb{Q}) - \lambda, \end{aligned} \tag{217}$$

where the remainder term $R_{i,\infty}(\beta;\mathbb{Q})$ satisfies

$$0 \leq R_{i,\infty}(\beta;\mathbb{Q}) \leq \overline{C} \cdot \prod_{k \in S} \beta_k. \tag{218}$$

In above, the constant $\overline{C} \leq |f^{s+1}(0)| \cdot (8\pi M)^{s+1}$. Further, Lemma I.1 and Lemma I.3 show that for the constant $c = b^{-s} \cdot \frac{f^s(2Mb)}{f^s(0)} > 0$,

$$F(\beta;\mathbb{Q}) \geq c \cdot F(b\mathbf{1}_S;\mathbb{Q}) \cdot \prod_{k \in S} \beta_k. \tag{219}$$

From equations (217)—(219), we obtain for $\beta \in \mathcal{B}$ and $i \in S$,

$$\frac{\partial}{\partial \beta_i} \ell(\beta; \lambda, \mathbb{Q}) \geq \left(c \cdot F(b\mathbf{1}_S; \mathbb{Q}) - \overline{C} \cdot \beta_i \right) \cdot \prod_{k \in S, k \neq i} \beta_k - \lambda. \tag{220}$$

Now we apply the bound (220) to $\beta = \beta^{(m)}$. Notice the below two observations. First, by induction hypothesis, we have

$$\prod_{k \in S, k \neq i} \beta_k^{(m)} \geq \zeta^{s-1}$$

Next, because we assume $\beta_i^{(m)} \leq \zeta(1 + \Delta) \leq 2 \cdot \zeta$, we have when $\text{SIGNAL}(S) \geq C \cdot \frac{1}{p^s}$ for sufficiently large C, the following bound

$$c \cdot F(b\mathbf{1}_S; \mathbb{Q}) \geq 2\overline{C} \cdot \beta_i^{(m)}.$$

Now, applying equation (220) to $\beta = \beta^{(m)}$, we obtain for $i \in S$

$$\frac{\partial}{\partial \beta_i} \ell(\beta^{(m)}; \lambda, \mathbb{Q}) \geq \frac{1}{2} \cdot c \cdot F(b\mathbf{1}_S; \mathbb{Q}) \cdot \zeta^{s-1} - \lambda > 0 \tag{221}$$

where the last inequality is due to our assumption that $\text{SIGNAL}(S) > 2\lambda$. As a consequence of equation (221), we obtain

$$\beta_i^{(m+1/2)} = \beta_i^{(m)} + \alpha \cdot \frac{\partial}{\partial \beta_i} \ell(\beta^{(m)}; \lambda, \mathbb{Q}) \geq \beta_i^{(m)} \geq \zeta.$$

Now we pass the result from $\beta_i^{(m+1/2)}$ to $\beta_i^{(m+1)}$. To do so, we use the projection Lemma O.2. Indeed, According to Lemma O.2,

$$\beta_i^{(m+1)} = \Pi_{\mathcal{B}}\left(\beta_i^{(m+1/2)} \right) = \left(\beta_i^{(m+1/2)} - \gamma^{(m)} \right)_+.$$

where the scalar $\gamma^{(m)} \geq 0$ is defined by

$$\gamma^{(m)} = \inf \left\{ \gamma \geq 0 : \sum_i \left(\beta_i^{(m+1/2)} - \gamma \right)_+ \leq b \right\}. \tag{222}$$

We will prove the below technical inequality (223) in Section N.8.3:

$$\gamma^{(m)} \leq \alpha \cdot \frac{\partial}{\partial \beta_i} \ell(\beta^{(m)}; \lambda, \mathbb{Q}). \tag{223}$$

As a consequence of inequality (223), we obtain

$$\beta_i^{(m+1)} = \left(\beta_i^{(m+1/2)} - \gamma^{(m)} \right)_+ \geq \beta_i^{(m)} \geq \zeta. \tag{224}$$

This proves equation (213), as desired.

N.8.2. *Finite Sample case:* $n < \infty$. The proof is effectively the same as in the population case $n = \infty$. The only essential difference is that in finite case ($n < \infty$), we are using the empirical gradient $\nabla \ell(\beta; \lambda, \mathbb{Q}_n)$ rather than the population gradient $\nabla \ell(\beta; \lambda, \mathbb{Q})$.

The way to resolve this is to give a high probability upper bound on the difference between the empirical and population gradient. In fact, from Corollary J.1, we know for some constant $C > 0$ depending only on b, M, f, we have for all $t > 0$, with probability at least $1 - p^{-t^2} - e^{-n\varrho^2/32}$

$$\sup_{j \in [p]} \sup_{\beta \in \mathcal{B}} \left| \frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}_n) - \frac{\partial}{\partial \beta_j} \ell(\beta; \lambda, \mathbb{Q}) \right| \leq C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t).$$

Now, for any $t > 0$, define $\bar{\lambda}(t)$ by

$$\bar{\lambda}(t) = \lambda + C \cdot \sqrt{\frac{\log p}{n}} \cdot (1 + t).$$

By replacing λ by $\bar{\lambda}(t)$, and $\nabla \ell(\beta; \lambda, \mathbb{Q})$ by $\nabla \ell(\beta; \lambda, \mathbb{Q}_n)$ in the proof of the population ($n = \infty$) and by copying the rest verbatim, we obtain a proof for the finite case $n < \infty$. The details are omitted for space considerations.

N.8.3. *Deferred proof of Inequality* (223). Below we show the deferred technical inequality (223). For notational simplicity, we denote

$$\Delta^{(m)} = \alpha \cdot \nabla \ell(\beta^{(m)}; \lambda, \mathbb{Q})$$

Thus $\beta^{(m+1/2)} = \beta^{(m)} + \Delta^{(m)}$. Suppose on the contrary that equation (223) fails, i.e., $\gamma^{(m)} > \Delta_i^{(m)} > 0$. By definition of $\gamma^{(m)}$ (cf. equation (222)), we get

$$\sum_k \left(\beta_k^{(m)} + \Delta_k^{(m)} - \Delta_i^{(m)}\right)_+ = \sum_k \left(\beta_k^{(m+1/2)} - \Delta_i^{(m)}\right)_+ > b. \tag{225}$$

Note $\Delta_k^{(m)} \leq 0$ for $k \in S^c$ by equation (211) and $\beta_k^{(m)} \leq \frac{b}{p}$ for $k \in S^c$ by equation (212). Since $b \geq \sum_k \beta_k^{(m)}$, equation (225) gives

$$\sum_{k \in S} \left(\beta_k^{(m)} + \Delta_k^{(m)} - \Delta_i^{(m)}\right)_+ > \max\left\{ \sum_{k \in S} \beta_k^{(m)}, \frac{bs}{p} \right\}. \tag{226}$$

Now $\left\|\Delta^{(m)}\right\|_\infty \leq 2\alpha M |f'(0)| \leq \zeta \Delta \leq \zeta/2$ by equation (214). Hence, by induction hypothesis, we have for $k \in S$, $\beta_k^{(m)} + \Delta_k^{(m)} - \Delta_i^{(m)} \geq 0$. Thus, equation (226) implies

$$\frac{1}{|S|} \sum_{k \in S} \left(\Delta_i^{(m)} - \Delta_k^{(m)}\right) < 0. \tag{227}$$

Using equation (217), equation (227) is equivalent to

$$\frac{1}{|S|}\sum_{k\in S}\left(\frac{1}{\beta_i^{(m)}}-\frac{1}{\beta_k^{(m)}}\right)\cdot F(\beta^{(m)};\mathbb{Q})-\frac{1}{|S|}\sum_{k\in S}R_{k,\infty}(\beta^{(m)};\mathbb{Q})-\lambda<0. \tag{228}$$

Below we show equation (228) can't happen, which leads to a contradiction. Thus, the desired technical inequality (223) holds.

To do so, we bound each term on the LHS of equation (228). First, by equation (218), the second term satisfies

$$\frac{1}{|S|}\sum_{k\in S}R_{k,\infty}(\beta^{(m)};\mathbb{Q})\le\overline{C}\cdot\prod_{k\in S}\beta_k^{(m)}. \tag{229}$$

To bound the first term, we note first from equation (219) that

$$F(\beta_k^{(m)};\mathbb{Q})\ge c\cdot F(b\mathbf{1}_S;\mathbb{Q})\cdot\prod_{k\in S}\beta_k^{(m)}>0. \tag{230}$$

Next, we prove the technical inequality

$$\frac{1}{\beta_i^{(m)}}\ge 2\cdot\frac{1}{|S|}\cdot\sum_{k\in S}\frac{1}{\beta_k^{(m)}}. \tag{231}$$

This is true because of the following points

- $\min_{k\in S}\beta_k^{(m)}\ge\zeta$ by induction hypothesis
- $\beta_i^{(m)}\le\zeta(1+\Delta)$ by assumption
- $\sum_{k\in S}\beta_k^{(m)}\ge\frac{b}{2p}\cdot s=4\zeta s$. This is true since equation (226), the fact that $\Delta_i^{(m)}\ge 0$ and that $\sum_{k\in S}|\Delta_k^{(m)}|\le\zeta\Delta s\le\frac{b}{2p}\cdot s$.

Now we are ready to use equations (229), (230) and (231) to give a lower bound on the LHS of (228). Formally, denote the LHS term of equation (228) by Γ. The previous results imply

$$\Gamma\ge\left(\frac{1}{2}\cdot c\cdot F(b\mathbf{1}_S;\mathbb{Q})-\overline{C}\cdot\beta_i^{(m)}\right)\cdot\prod_{k\in S,k\neq i}\beta_k^{(m)}-\lambda. \tag{232}$$

Note $\frac{1}{2}\cdot c\cdot F(b\mathbf{1}_S;\mathbb{Q})-\overline{C}\cdot\beta_i^{(m)}\ge\frac{1}{4}\cdot c\cdot F(b\mathbf{1}_S;\mathbb{Q})$ when the constant C is large enough (recall $\beta_i^{(m)}\le 2\cdot\zeta$ and $\mathrm{SIGNAL}(S)\ge C\cdot\frac{1}{p^s}$). Moreover, $\beta_k^{(m)}\ge\zeta$ for all $k\in S$. Therefore, we obtain

$$\Gamma\ge\frac{1}{4}\cdot c\cdot F(b\mathbf{1}_S;\mathbb{Q})\cdot\zeta^{s-1}-\lambda>0$$

where the last inequality is due to $\mathrm{SIGNAL}(S)>2\lambda$. This contradicts equation (228). The proof is now complete.

APPENDIX O: SUPPORTING LEMMA

O.1. Concentration Inequality for U-statistics. We state the Hoeffding's and Bernstein's inequality for U-statistics [5, 1].

LEMMA O.1. *Let $X_1, X_2, \ldots, X_n$ be i.i.d random variables taking values in $\mathbb{R}$. Let h be a measurable function of m variables. The U-statistics of order m and kernel h is defined by*

$$U_m(h) = \frac{(n-m)!}{n!} \sum_{(i_1,i_2,\ldots,i_m)\in I_n^m} h\left(X_{i_1}, X_{i_2}, \ldots, X_{i_m}\right).$$

where

$$I_n^m = \{(i_1, i_2, \ldots, i_m) : 1 \le i_j \le n, i_j \ne i_k \text{ if } j \ne k\}.$$

Assume for some $M, \sigma > 0$, we have $|h(X_1, X_2, \ldots, X_m)| \le M$ almost surely and $\mathbb{E}[h^2(X_1, X_2, \ldots, X_m)] \le \sigma^2$. Then,

1. *(Hoeffding's inequality) We have for any $t > 0$,*

$$\mathbb{P}\left(|U_m(h) - \mathbb{E}U_m(h)| \ge t\right) \le 2\exp\left(-\frac{nt^2}{mM^2}\right).$$

2. *(Bernstein's inequality) We have for any $t > 0$,*

$$\mathbb{P}\left(|U_m(h) - \mathbb{E}U_m(h)| \ge t\right) \le 2\exp\left(-\frac{nt^2}{2m(Mt + \mathbb{E}[h^2])}\right).$$

O.2. Projection onto ℓ_1 ball. Fix $b > 0$. As usual, we denote the polytope

$$\mathcal{B} = \min\left\{\beta \in \mathbb{R}_+^p : \mathbf{1}^T\beta \le b\right\}.$$

Lemma O.2 gives a characterization of the projection onto $\mathcal{B}$.

LEMMA O.2. *Let $\beta \in \mathbb{R}^p$. Its projection $\tilde{\beta} = \Pi_{\mathcal{B}}(\beta)$ satisfies*

$$\tilde{\beta} = (\beta - \gamma)_+$$

where $\gamma \ge 0$ is defined by

$$\gamma = \inf\{\gamma \ge 0 : \sum_{i\in[p]} (\beta_i - \gamma)_+ \le b\}. \tag{233}$$

PROOF. By definition, $\tilde{\beta} = \Pi_{\mathcal{B}}(\beta)$ is the solution of the convex optimization problem:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \frac{1}{2}\|x-\beta\|_2^2 \\ \text{subject to} \quad & x \geq 0, \mathbf{1}^T x \leq b. \end{aligned}$$

The KKT condition of the optimization problem is, for some $\mu \in \mathbb{R}^p, \gamma \in \mathbb{R}$,

$$\begin{aligned} \tilde{\beta} - \beta + \gamma\mathbf{1} - \mu = 0. \\ \gamma(\mathbf{1}^T\tilde{\beta} - b) = 0, \mu^T\tilde{\beta} = 0 \\ \tilde{\beta} \geq 0, \mathbf{1}^T\tilde{\beta} \leq b. \\ \gamma \geq 0, \mu \geq 0. \end{aligned} \tag{234}$$

Thus $\tilde{\beta} = \beta - \gamma\mathbf{1} + \mu \geq 0$, $\mu \geq 0$, and $\tilde{\beta}^T\mu = 0$. From this, we derive

$$\tilde{\beta} = (\beta - \gamma\mathbf{1})_+.$$

To determine the value of γ, we note, $\gamma \geq 0$, $\mathbf{1}^T\tilde{\beta} \leq b$, $\gamma(\mathbf{1}^T\tilde{\beta} - b) = 0$ where $\tilde{\beta} = (\beta - \gamma\mathbf{1})_+$. Hence, γ must be the smallest nonnegative number such that $\mathbf{1}^T(\beta - \gamma\mathbf{1})_+ \leq b$. This proves the desired Lemma O.2.

□

LEMMA O.3. *Let $\bar{\beta} \in \mathcal{B}$ and $\beta \in \mathbb{R}^p$, we have*

$$\left\|\Pi_{\mathcal{B}}(\beta) - \bar{\beta}\right\|_\infty \leq 2 \cdot \left\|\beta - \bar{\beta}\right\|_\infty.$$

PROOF. Let $\tilde{\beta} = \Pi_{\mathcal{B}}(\beta)$. From Lemma O.2, we know that $\tilde{\beta} = (\beta - \gamma)_+$ for the $\gamma \geq 0$ which is defined by the equation (233). This implies

$$\|\tilde{\beta} - \bar{\beta}\|_\infty \leq \left\|\beta - \bar{\beta}\right\|_\infty + \gamma. \tag{235}$$

Below we show $\gamma \leq \gamma_0 := \left\|\beta - \bar{\beta}\right\|_\infty$. Indeed, $(\beta_i - \gamma_0)_+ \leq \bar{\beta}_i$ by triangle inequality. Thus we have

$$\sum_i (\beta_i - \gamma_0)_+ \leq \sum_i \bar{\beta}_i \leq b.$$

According to the definition of γ, this implies $\gamma \leq \gamma_0 = \left\|\beta - \bar{\beta}\right\|_\infty$. Substituting it into equation (235) yields the desired Lemma O.3. □

O.3. Basic Properties on Projected Gradient Ascent. The following result is standard in nonlinear optimization [2, Prop 2.3.2].

LEMMA O.4. *Consider the (non-convex) optimization problem*

$$\begin{aligned} &\text{maximize } J(\beta) \\ &\text{subject to } \beta \in \mathcal{C}. \end{aligned}$$

Assume the following assumptions on J and $\mathcal{C}$:

- *The gradient $x \mapsto \nabla J(x)$ is L-Lipschitz on $\mathcal{C}$, i.e.,*

$$\left\|\nabla J(\beta) - \nabla J(\beta')\right\|_2 \leq L \left\|\beta - \beta'\right\|_2 \quad \text{for any } \beta, \beta' \in \mathcal{C}.$$

- *The constraint set $\mathcal{C}$ is convex.*

Consider the projected gradient ascent algorithm with stepsize α:

$$\beta^{(k+1)} = \Pi_{\mathcal{C}}\left(\beta^{(k)} + \alpha \nabla J(\beta^{(k)})\right).$$

Let the stepsize $\alpha \leq 1/L$. Then we have

1. *The mapping $k \mapsto J(\beta^{(k)})$ is increasing. In particular, we have,*

$$J(\beta^{(k))}) \geq J(\beta^{(0)}) \quad \text{for all } k \in \mathbb{N}.$$

2. *Any accumulation point β^{∞} of $\{\beta^{(k)}\}_{k\in\mathbb{N}}$ is a stationary point, i.e.,*

$$\langle \nabla J(\beta^{\infty}), \beta' - \beta^{\infty} \rangle \leq 0 \text{ for any } \beta' \in \mathcal{C}.$$

O.4. Construction of two distinct distributions with matching moments. Lemma O.5 constructs two distinct multivariate distributions of supported on the same compact interval $[-1,1]^m$ that share the same mixed moments (up to l) and same marginal distribution for any strict subset of variables. The proof adapts a classical argument (see e.g., [7]).

LEMMA O.5. *Fix $m, l \in \mathbb{N}$. Let $X = (X_1, \ldots, X_m) \in \mathbb{R}^m$. There exist two distinct probability distributions $\mathbb{P}_+, \mathbb{P}_-$ supported on $[-1,1]^m$ such that*

1. *For any strict subset $A \subsetneq [m]$, the distribution of X_A under $\mathbb{P}_+$ is the same as the distribution of X_A under $\mathbb{P}_-$.*
2. *For any $\alpha = (\alpha_1, \ldots \alpha_m) \in \mathbb{N}^m$ with $\sum_{i=1}^m \alpha_i \leq l$, we have*

$$\mathbb{E}_+\left[\prod_{i=1}^m X_i^{\alpha_i}\right] = \mathbb{E}_-\left[\prod_{i=1}^m X_i^{\alpha_i}\right].$$

PROOF. The proof of Lemma O.5 is based on the Hahn-Banach Theorem and the Riesz representation theorem. Below for simplicity, we only present the proof for the case where $m = 2$. The argument for $m > 2$ is essentially the same as the argument for $m = 2$ presented below.

Let $m = 2$. Consider the space $\mathcal{D} = C([-1, 1]^2)$ of continuous real-valued functions on the box $[-1, 1]^2$ with uniform norm $\|\cdot\|_\infty$. Let us denote

$$f_{\alpha_1,\alpha_2}(X) = X_1^{\alpha_1} X_2^{\alpha_2}$$

for any $\alpha = (\alpha_1, \alpha_2) \in \mathbb{N}^2$. We denote $\mathcal{P} \subseteq \mathcal{D}$ to be the linear subspace spanned by $\{f_\alpha\}_{\alpha\in\mathcal{A}}$ where

$$\mathcal{A} = \{\alpha \in \mathbb{N}^2 : \alpha_1\alpha_2 = 0\} \cup \{\alpha \in \mathbb{N}^2 : \alpha_1 + \alpha_2 \le l\}.$$

and denote $\mathcal{F}$ to be the linear subspace spanned by $\bar{f} \equiv f_{l+1,l+1}$ and $\mathcal{P}$. Let T be the following linear functional defined on $\mathcal{F}$: we define $T(c\bar{f} + f) = c$ for any $f \in \mathcal{P}$ and $c \in \mathbb{R}$. The BLT theorem in functional analysis says that T has a continuous extension on $\overline{\mathcal{F}}$, which is the closure of $\mathcal{F}$ under the $\|\cdot\|_\infty$ norm. Clearly, the norm of the linear functional T on $\overline{\mathcal{F}}$ is positive (and is in fact equal to 1), and T vanishes on the closed subspace $\overline{\mathcal{P}}$, the closure of $\mathcal{P}$ under $\|\cdot\|_\infty$.

Now by the Hahn-Banach theorem, T has a continuous extension to the whole space $\mathcal{D}$ without changing its norm. For simplicity, we also call this extension T. It then follows from the Riesz representation theorem that for some (non-degenerate) Borel signed measure τ, we have for each $g \in \mathcal{D}$,

$$T(g) = \iint_{[-1,1]^2} g(x)\tau(dx).$$

Now the Hahn-Jordan decomposition shows that there exist two positive measures τ_+ and τ_- such that $\tau = \tau_+ - \tau_-$. Define $\mathbb{P}_+$ and $\mathbb{P}_-$ to be the probability measures normalized from τ_+ and τ_- respectively. The fact that $T(f) = 0$ for any $f \in \overline{\mathcal{P}}$ immediately implies for any $\alpha \in \mathcal{A}$, we have

$$\iint x_1^{\alpha_1} x_2^{\alpha_2} \mathbb{P}_+(dx) = \iint x_1^{\alpha_1} x_2^{\alpha_2} \mathbb{P}_-(dx),$$

and therefore for any $\alpha \in \mathcal{A}$, we have

$$\mathbb{E}_+ \left[X_1^{\alpha_1} X_2^{\alpha_2}\right] = \mathbb{E}_- \left[X_1^{\alpha_1} X_2^{\alpha_2}\right].$$

As both $\mathbb{P}_+$ and $\mathbb{P}_-$ are defined on compact domain, the fact that X_1 (and X_2) has the same moments under $\mathbb{P}_+$ and $\mathbb{P}_-$ imply that X_1 (and X_2) has the same marginal distribution under $\mathbb{P}_+$ and $\mathbb{P}_-$. In addition, it also shows that the vector $X = (X_1, X_2)$ has the same mixed moments up to l. Finally, the fact that $T \neq 0$ implies that $\mathbb{P}_+$ and $\mathbb{P}_-$ must be distinct. □

O.5. Covariance inequality. The covariance of two monotone functions of X is always positive.

LEMMA O.6. *For any function g_1, g_2 that is monotonically increasing (or decreasing), and any non-negative measure $\tilde{\mu}$ with $|\tilde{\mu}| < \infty$, we have*

$$\int g_1(t)g_2(t)\tilde{\mu}(dt) \geq \frac{1}{|\tilde{\mu}|}\int g_1(t)\tilde{\mu}(dt)\int g_2(t)\tilde{\mu}(dt).$$

REFERENCES

[1] M. A. Arcones. A Bernstein-type inequality for u-statistics and u-processes. *Statistics & probability letters*, 22(3):239–247, 1995.

[2] D. P. Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

[3] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] Z. Ditzian and V. Totik. *Moduli of smoothness*, volume 9. Springer Science & Business Media, 2012.

[5] W. Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[6] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[7] O. Lepski, A. Nemirovski, and V. Spokoiny. On estimation of the ℓ_r norm of a regression function. *Probability theory and related fields*, 113(2):221–253, 1999.

[8] A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: theory of majorization and its applications*, volume 143. Springer, 1979.

[9] C. McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[10] C. A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. In *Approximation theory and spline functions*, pages 143–145. Springer, 1984.

[11] R. L. Schilling, R. Song, and Z. Vondracek. *Bernstein functions: theory and applications*, volume 37. Walter de Gruyter, 2012.

[12] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[13] Y. Su and B. Xiong. *Methods and Techniques for Proving Inequalities: In Mathematical Olympiad and Competitions*, volume 11. World Scientific Publishing Company, 2015.

[14] G. J. Székely and M. L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.

UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
E-MAIL: keli.liu25@gmail.com
ruanfeng2124@gmail.com