Technical Report for *Heterogeneous Treatment Effects in Panel Data: Insights into the Healthy Incentives Program*

In this technical report, we present the omitted proofs for the results stated in the paper *Heterogeneous Treatment Effects in Panel Data: Insights into the Healthy Incentives Program*. In Section 1, we prove the estimation error bound stated in Theorem 2, and in Section 2, we prove the main lemmas used for the proof of Theorem 2. Finally, in Section 3, we provide the proofs of auxiliary lemmas used throughout the paper.

1 Proof of Theorem 2

We aim to bound $\left|\tau_i^d - \tilde{\tau}_i^\star\right|$. Throughout this section, we assume that the assumptions made in the statement of Theorem 2 are satisfied. To simplify notation, we define $\eta_i := \|\tilde{Z}_i\|_{\mathrm{F}}$. We have

$$\begin{aligned}
\left|\tau_i^d - \tilde{\tau}_i^\star\right| &\leq \left\|\tau^d - \tilde{\tau}^\star\right\| \\
&= \frac{1}{\eta_i}\left\|\eta_i\tau^d - \tau^\star\right\| \\
&\overset{(i)}{=} \frac{1}{\eta_i}\left\|D^{-1}(\Delta^2 + \Delta^3)\right\| \\
&\leq \frac{1}{\sigma_{\min}(D)\eta_i}\left(\left\|\Delta^2\right\| + \left\|\Delta^3\right\|\right),
\end{aligned} \tag{1}$$

where (i) is due to the error decomposition in Lemma 1 and the definition of τ^d.

Hence, we aim to upper bound $\|\Delta^2\|$ and $\|\Delta^3\|$ and lower bound $\sigma_{\min}(D)$. The main challenge lies in upper bounding $\|\Delta^3\|$. In fact, the desired bounds for $\sigma_{\min}(D)$ and $\|\Delta^2\|$ (presented in the following lemma) are obtained during the process of bounding $\|\Delta^3\|$.

Lemma 1.1 *Let D and Δ^2 be defined as in Lemma 1. We have $\sigma_{\min}(D) \geq \frac{c_s}{2\log n}$. Furthermore, we have the following for sufficiently large n:*

$$\left\|\Delta^2\right\| \lesssim \frac{\sigma^2 r^{1.5}\kappa n \log^{6.5}(n)}{\sigma_{\min}} + \log^{0.5}(n)\cdot\max_{i\in[k]}\frac{\left|\left\langle P_{\mathbf{T}^{\star\perp}}(\tilde{Z}_i), P_{\mathbf{T}^{\star\perp}}(E+\delta)\right\rangle\right|}{\|\tilde{Z}_i\|_{\mathrm{F}}}.$$

The proof of the above lemma is provided in Section 2.6.

We now discuss the strategy for bounding $\|\Delta^3\|$. In order to control $P_{\hat{\mathbf{T}}^\perp}(M^\star)$, we aim to show that the true counterfactual matrix $M^\star$ has tangent space close to that of $\hat{M}$. Because we introduced $\hat{m}$ in the convex formulation, which centers the rows of $\hat{M}$ without penalization from the nuclear norm term, we will focus on the projection of $M^\star$ that has rows with mean zero, $P_{\mathbf{1}^\perp}(M^\star)$. Recall that $P_{\mathbf{1}^\perp}(M^\star) = U^\star\Sigma^\star V^{\star\top}$ is its SVD, and define $X^\star := U^\star\Sigma^{\star 1/2}$ and $Y^\star := V^\star\Sigma^{\star 1/2}$. We similarly define these quantities for $\hat{M}$: $\hat{X} := \hat{U}\hat{\Sigma}^{1/2}$ and $\hat{Y} := \hat{V}\hat{\Sigma}^{1/2}$. We aim to show that we have $(\hat{X}, \hat{Y}) \approx (X^\star, Y^\star)$. This is stated formally in Lemma 1.2, which requires introducing a few additional notations first.

Let $g(M, \tau, m)$ denote the convex function that we are optimizing in (3). Recall that $(\hat{M}, \hat{\tau}, \hat{m})$ is a global optimum of function g. Following the main proof idea in Farias et al. (2021), we define

a non-convex proxy function f, which is similar to g, except M is split into two variables and expressed as $XY^\top$.

$$f(X,Y) = \min_{m\in\mathbb{R}^n,\tau\in\mathbb{R}^k} \frac{1}{2}\left\| O - XY^\top - m\mathbf{1}^\top - \sum_{i=1}^k \tau_i Z_i \right\|_{\mathrm{F}}^2 + \frac{1}{2}\lambda\langle X,X\rangle + \frac{1}{2}\lambda\langle Y,Y\rangle .$$

Our analysis will relate $(\hat{X},\hat{Y})$ to a specific local optimum of f, which we will show is close to $(X^\star,Y^\star)$. The local optimum of f considered is the limit of the gradient flow of f, initiated at $(X^\star,Y^\star)$, formally defined by the following differential equation:

$$\begin{cases} (\dot{X}^t,\dot{Y}^t) = -\nabla f(X^t,Y^t) \\ (X^0,Y^0) = (X^\star,Y^\star). \end{cases} \tag{2}$$

We define $H_{X,Y}$ as the rotation matrix that optimally aligns (X,Y) with $(X^\star,Y^\star)$. That is, letting $\mathcal{O}^{r\times r}$ denote the set of $r\times r$ orthogonal matrices, $H_{X,Y}$ is formally defined as follows:

$$H_{X,Y} := \operatorname*{argmin}_{R\in\mathcal{O}^{r\times r}} \|XR - X^\star\|_{\mathrm{F}}^2 + \|YR - Y^\star\|_{\mathrm{F}}^2 .$$

Define $F^\star$ to be the vertical concatenation of matrices $X^\star$ and $Y^\star$. That is, $F^\star = [(X^\star)^\top,(Y^\star)^\top]^\top$. We are now ready to present Lemma 1.2.

Lemma 1.2 *For sufficiently large n, the gradient flow of f, starting from $(X^\star,Y^\star)$, converges to (X,Y) such that $X=\hat{X}R$ and $Y=\hat{Y}R$ for some rotation matrix $R\in\mathcal{O}^{r\times r}$. Furthermore, $(X,Y)\in\mathcal{B}$, where*

$$\mathcal{B} := \left\{(X,Y) \mid \|XH_{X,Y} - X^\star\|_{\mathrm{F}}^2 + \|YH_{X,Y} - Y^\star\|_{\mathrm{F}}^2 \le \rho^2\right\}, \quad \rho := \frac{\sigma\sqrt{nr}\log^6(n)}{\sigma_{\min}}\|F^\star\|_{\mathrm{F}} .$$

Let us derive an upper bound on ρ. Because $\|X^\star\|_{\mathrm{F}} \le \|X^\star\|\sqrt{r} = \sqrt{\sigma_{\max} r}$ and $\|Y^\star\|_{\mathrm{F}} \le \sqrt{\sigma_{\max} r}$, we can conclude that $\|F^\star\|_{\mathrm{F}} \lesssim \sqrt{\sigma_{\max} r}$. Thus, we can upper bound ρ as follows:

$$\rho \lesssim \frac{\sigma\sqrt{nr}\log^6(n)}{\sigma_{\min}}\sqrt{\sigma_{\max} r} = \frac{\sigma\sqrt{nr}\log^6(n)}{\sigma_{\min}}\sqrt{\sigma_{\min}\kappa r} = \frac{\sigma r\kappa^{0.5}\sqrt{n}\log^6(n)}{\sqrt{\sigma_{\min}}}. \tag{3}$$

Using (3) and the assumption $\frac{\sigma\sqrt{n}}{\sigma_{\min}} \lesssim \frac{1}{\kappa^2 r^2\log^{12.5}(n)}$ provided in the theorem, we can further upper bound ρ as follows:

$$\rho \lesssim \frac{\sqrt{\sigma_{\min}}}{\log^{6.5}(n)\kappa}. \tag{4}$$

With Lemma 1.2 and the bound on ρ, we can complete the proof of the theorem as follows.

$$\begin{aligned} \left\|\Delta^3\right\| &\le \sqrt{k}\left\|\Delta^3\right\|_\infty \\ &\le \sqrt{k}\left\|P_{\hat{\mathbf{T}}^\perp}(M^\star)\right\|_{\mathrm{F}} \\ &= \sqrt{k}\left\|(I-\hat{U}\hat{U}^\top)(X^\star Y^{\star\top} + m^\star\mathbf{1}^\top)\left(I - \frac{\hat{r}\hat{r}^\top}{\hat{r}^\top\hat{r}}\right)(I-\hat{V}\hat{V}^\top)\right\|_{\mathrm{F}}, \end{aligned}$$

where $\hat{r} = (I - \hat{V}\hat{V}^\top)\mathbf{1}$, and the last line comes from the closed form of the projection derived in Lemma EC.3. We can use the fact that $\hat{V}^\top \mathbf{1} = 0$ from Claim EC.1 to simplify the expression. Note that with $\hat{V}^\top \mathbf{1} = 0$, $\hat{r}$ simply evaluates to $\mathbf{1}$.

$$\begin{aligned}\sqrt{k}\left\|P_{\hat{\mathbf{T}}^\perp}(M^\star)\right\|_{\mathrm{F}} &= \sqrt{k}\left\|(I - \hat{U}\hat{U}^\top)X^\star Y^{\star\top}\left(I - \frac{\mathbf{1}\mathbf{1}^\top}{T}\right)(I - \hat{V}\hat{V}^\top)\right\|_{\mathrm{F}} \\ &= \sqrt{k}\left\|(I - \hat{U}\hat{U}^\top)X^\star Y^{\star\top}(I - \hat{V}\hat{V}^\top)\right\|_{\mathrm{F}},\end{aligned}$$

where the last step is due to the fact that $X^\star Y^{\star\top}\mathbf{1} = 0$ because $X^\star Y^{\star\top} = P_{\mathbf{1}^\perp}(M^\star)$.

Finally, let (X, Y) be the limit of the gradient flow of f starting from $(X^\star, Y^\star)$. By Lemma 1.2, we have $X = \hat{X}R$ and $Y = \hat{Y}R$ for some rotation matrix $R \in \mathcal{O}^{r\times r}$. This, combined with the definition of $\hat{\mathbf{T}}$, implies $P_{\hat{\mathbf{T}}^\perp}(XA^\top) = P_{\hat{\mathbf{T}}^\perp}(BY^\top) = 0$ for any $A \in \mathbb{R}^{T\times r}$ and $B \in \mathbb{R}^{n\times r}$. Hence,

$$\begin{aligned}\left\|\Delta^3\right\| &\le \sqrt{k}\left\|(I - \hat{U}\hat{U}^\top)(XH_{X,Y} - X^\star)(YH_{X,Y} - Y^\star)^\top(I - \hat{V}\hat{V}^\top)\right\|_{\mathrm{F}} \\ &\le \sqrt{k}\left\|XH_{X,Y} - X^\star\right\|_{\mathrm{F}}\left\|YH_{X,Y} - Y^\star\right\|_{\mathrm{F}} \\ &\lesssim \sqrt{k}\rho^2 \\ &\lesssim \frac{\sigma^2 r^2 \kappa n \log^{12.5}(n)}{\sigma_{\min}},\end{aligned} \tag{5}$$

where the last step is due to the fact that $k = O(\log n)$.

Now that we have bounded $\sigma_{\min}(D)$, $\|\Delta^2\|$, and $\|\Delta^3\|$, we can plug these bounds into (1) to complete the proof of the theorem.

2 Proof of Lemma 1.2

Throughout this section, we assume that the assumptions made in the statement of Theorem 2 are satisfied. The proof of Lemma 1.2 is completed by combining the results of the following two lemmas.

Lemma 2.1 *Any point (X, Y) on the gradient flow of f starting from $(X^\star, Y^\star)$ satisfies $(X, Y) \in \mathcal{B}$.*

Lemma 2.2 *The limit (X, Y) of the gradient flow of f starting from $(X^\star, Y^\star)$ satisfies $X = \hat{X}R$ and $Y = \hat{Y}R$ for some rotation matrix $R \in \mathcal{O}^{r\times r}$.*

We note our methodology differs from that of Farias et al. (2021). Instead of analyzing a gradient descent algorithm, we analyze the gradient flow of the function f. This allows us to simplify the analysis by avoiding error terms due to the discretization of gradient descent.

In this section, we start by deriving the gradient of the function f and examining some properties of its gradient flow in Section 2.1. Then, we prove some technical lemmas. We extend Assumption 2 to a broader subset of matrices within the set $\mathcal{B}$ in Section 2.2, and establish bounds on the noise in Section 2.3. Finally, we complete the proofs of Lemma 2.1 in Section 2.4 and Lemma 2.2 in Section 2.5.

2.1 The gradient of the function f

Before we prove Lemma 2.1, we need to derive the gradient of the function f.

Define $P_{\mathbf{Z}}$ to be the projection operator that projects onto the subspace $\mathbf{Z}$. Note that in the definition of $f(X, Y)$, the quantities m and τ are chosen to minimize the distance between $O - XY^\top$ and this subspace, measured in terms of Euclidean norm. Hence, we can view m and τ as coordinates of the projection of $O - XY^\top$ onto this subspace. This observation gives us the following equivalent definition of $f(X, Y)$:

$$f(X,Y) = \frac{1}{2}\left\|P_{\mathbf{Z}^\perp}\left(O - XY^\top\right)\right\|_{\mathrm{F}}^2 + \frac{1}{2}\lambda\left\langle X, X\right\rangle + \frac{1}{2}\lambda\left\langle Y, Y\right\rangle.$$

Consider a single entry of the matrix X: X_{ij}. Note that the expression inside the Frobenius norm is linear in X_{ij}. Hence, we can rewrite $f(X, Y)$ as follows:

$$f(X,Y) = \frac{1}{2}\left\|X_{ij}A + B\right\|_{\mathrm{F}}^2 + \frac{1}{2}\lambda\left\langle X, X\right\rangle + \frac{1}{2}\lambda\left\langle Y, Y\right\rangle,$$

for some matrices A and B that do not depend on X_{ij} (but may depend on other entries of X, and Y).

Now we take the partial derivative of f with respect to X_{ij}. Let E_{ij} be the matrix where all elements are zero except for the element in the i-th row and j-th column, which is 1. Then,

$$\begin{aligned}\frac{\partial f}{\partial X_{ij}} &= \left\langle X_{ij}A + B, A\right\rangle + \lambda X_{ij} \\ &\overset{(i)}{=} \left\langle P_{\mathbf{Z}^\perp}\left(O - XY^\top\right), -P_{\mathbf{Z}^\perp}\left(E_{ij}Y^\top\right)\right\rangle + \lambda X_{ij} \\ &= -\left\langle P_{\mathbf{Z}^\perp}\left(O - XY^\top\right), E_{ij}Y^\top\right\rangle + \lambda X_{ij},\end{aligned}$$

where (i) is due to the fact that A and B were defined such that $X_{ij}A + B = P_{\mathbf{Z}^\perp}\left(O - XY^\top\right)$. Hence,

$$\begin{aligned}\frac{\partial f}{\partial X} &= -P_{\mathbf{Z}^\perp}\left(O - XY^\top\right)Y + \lambda X \\ &= -\left(O - XY^\top - m(X,Y)\mathbf{1}^\top - \sum_{i=1}^{k}\tau_i(X,Y)Z_i\right)Y + \lambda X,\end{aligned}$$

where $m(X,Y), \tau(X,Y) := \operatorname{argmin}_{m,\tau}\left\|O - XY^\top - m\mathbf{1}^\top - \sum_{i=1}^{k}\tau_i Z_i\right\|_{\mathrm{F}}^2$. We will write m and τ to represent $m(X,Y)$ and $\tau(X,Y)$ below for notational simplicity.

Using $O = X^\star Y^{\star\top} + m^\star\mathbf{1}^\top + \sum_{i=1}^{k}\tau_i^\star Z_i + \hat{E}$, we can simplify the gradient as follows:

$$\frac{\partial f}{\partial X} = \left(XY^\top - X^\star Y^{\star\top} + (m - m^\star)\mathbf{1}^\top + \sum_{i=1}^{k}(\tau_i - \tau_i^\star)Z_i - \hat{E}\right)Y + \lambda X.$$

Because m and τ are coordinates of the projection of $O - XY^\top$ onto the subspace $\mathbf{Z}$, we have

$$m\mathbf{1}^\top + \sum_{i=1}^{k}\tau_i Z_i = P_{\mathbf{Z}}\left(O - XY^\top\right).$$

Additionally, because $m^\star \mathbf{1}^\top + \sum_{i=1}^k \tau_i^\star Z_i = O - X^\star Y^{\star\top} - \hat{E}$, the right hand side quantity is also in the subspace $\mathbf{Z}$, so we have

$$m^\star \mathbf{1}^\top + \sum_{i=1}^k \tau_i^\star Z_i = P_{\mathbf{Z}} \left(O - X^\star Y^{\star\top} - \hat{E} \right).$$

Subtracting the two equations above, we have

$$(m - m^\star)\mathbf{1}^\top + \sum_{i=1}^k (\tau_i - \tau_i^\star) Z_i = P_{\mathbf{Z}} \left(X^\star Y^{\star\top} - XY^\top + \hat{E} \right). \tag{6}$$

Using (6), we can rewrite

$$\frac{\partial f}{\partial X} = \left(P_{\mathbf{Z}^\perp} \left(XY^\top - X^\star Y^{\star\top} - \hat{E} \right) \right) Y + \lambda X.$$

Similarly, we can derive

$$\frac{\partial f}{\partial Y} = \left(P_{\mathbf{Z}^\perp} \left(XY^\top - X^\star Y^{\star\top} - \hat{E} \right) \right)^\top X + \lambda Y.$$

2.1.1 Properties of the gradient flow

We now prove some properties of the gradient flow of the function f starting from the point $(X^\star, Y^\star)$. For simplicity of notation, we define

$$D := P_{\mathbf{Z}^\perp} \left(XY^\top - X^\star Y^{\star\top} - \hat{E} \right),$$

so we can simplify our gradients as follows:

$$\frac{\partial f}{\partial X} = DY + \lambda X \qquad \text{and} \qquad \frac{\partial f}{\partial Y} = D^\top X + \lambda Y. \tag{7}$$

Lemma 2.3 *Every point on the gradient flow of function $f(X, Y)$ starting from the point $(X^\star, Y^\star)$ satisfies $X^\top X = Y^\top Y$.*

Proof. Note that at the starting point $(X^\star, Y^\star)$, we have that $X^{\star\top} X^\star = \Sigma^\star = Y^{\star\top} Y^\star$ as desired.

Next, we examine the change in the value $\phi(t) := X^{t\top} X^t - Y^{t\top} Y^t$ as we move along the gradient flow defined by (2). For convenience, we omit the superscript t in X^t and Y^t. Using (2), we have

$$\begin{pmatrix} \dot{X} \\ \dot{Y} \end{pmatrix} = -\nabla f \begin{pmatrix} X \\ Y \end{pmatrix} = - \begin{pmatrix} DY + \lambda X \\ D^\top X + \lambda Y \end{pmatrix}.$$

The derivative of ϕ is

$$\begin{aligned} \phi'(t) &= \dot{X}^\top X + X^\top \dot{X} - \dot{Y}^\top Y - Y^\top \dot{Y} \\ &= -\left(DY + \lambda X\right)^\top X - X^\top (DY + \lambda X) + \left(D^\top X + \lambda Y\right)^\top Y + Y^\top (D^\top X + \lambda Y) \\ &= -2\lambda \left(X^\top X - Y^\top Y \right), \end{aligned}$$

showing that $\phi'(t) = -2\lambda \phi(t)$. As a result, recalling that $\phi(0) = 0$, we can solve the differential equation to obtain $\phi(t) = e^{-2\lambda t}\phi(0) = 0$ for all t. □

Lemma 2.4 *Every point on the gradient flow of function $f(X, Y)$ starting from the point $(X^\star, Y^\star)$ satisfies $Y^\top \mathbf{1} = 0$. Furthermore, if $U\Sigma V^\top$ is the SVD of $XY^\top$, then $V^\top \mathbf{1} = 0$.*

Proof. Using $X^\star Y^{\star\top}\mathbf{1} = 0$ (because $X^\star Y^{\star\top} = P_{\mathbf{1}^\perp}(M^\star)$), we have that

$$\left(X^\star Y^{\star\top}\mathbf{1}\right)^\top X^\star Y^{\star\top}\mathbf{1} = \mathbf{1}^\top V^\star \Sigma^{\star 2} V^{\star\top}\mathbf{1} = 0,$$

which can only happen if $V^{\star\top}\mathbf{1} = 0$. This implies that $Y^{\star\top}\mathbf{1} = 0$.

Using the same approach as the proof of Lemma 2.3, we examine the change in the value of $\phi(t) := {Y^t}^\top \mathbf{1}$ as we move along the gradient flow defined by (2). We omit the superscript t in Y^t for notational convenience. Using (2), we have $\dot{Y} = -D^\top X - \lambda Y$. Now we compute the derivative of $\phi(t)$:

$$\begin{aligned}
\phi'(t) &= \dot{Y}^\top \mathbf{1} \\
&= -X^\top D\mathbf{1} - \lambda Y^\top \mathbf{1} \\
&\overset{(i)}{=} -\lambda Y^\top \mathbf{1} \\
&= -\lambda\phi(t).
\end{aligned}$$

The equality (i) follows because D is the projection onto a space orthogonal to $\left\{\alpha\mathbf{1}^\top \mid \alpha \in \mathbb{R}^n\right\}$, effectively centering its rows. Because D has zero-mean rows, $D\mathbf{1} = 0$.

Note that $\phi(0) = Y^{\star\top}\mathbf{1} = 0$. Solving the differential equation, we have $\phi(t) = e^{-\lambda t}\phi(0) = 0$ for all t. Using the same logic we used to show $V^{\star\top}\mathbf{1} = 0$, we have that $Y^\top\mathbf{1}$ implies $V^\top\mathbf{1} = 0$. □

2.2 Extending assumptions to $\mathcal{B}$

We prove a lemma that extends Assumptions 2(a), 2(b), and 2(c) to a broader subset of matrices within the set $\mathcal{B}$, thereby expanding the applicability of the original assumptions.

We'll begin by proving a useful lemma that provides bounds for the singular values of matrices in the set $\mathcal{B}$. We show these values are within a constant factor of the singular value range of $X^\star$ and $Y^\star$, spanning from $\sqrt{\sigma_{\min}}$ to $\sqrt{\sigma_{\max}}$.

Lemma 2.5 *For large enough n and $(X, Y) \in \mathcal{B}$, we have the following for any $i \in [r]$:*

$$\sigma_i(X), \sigma_i(Y) \in \left[\frac{\sqrt{\sigma_{\min}}}{2}, 2\sqrt{\sigma_{\max}}\right].$$

Proof. The singular values of a matrix are not changed by right-multiplying by an orthogonal matrix. Hence, without loss of generality, we suppose that (X, Y) are rotated such that they are optimally aligned with $(X^\star, Y^\star)$; in other words, $H_{X,Y} = I$. Then, $(X, Y) \in \mathcal{B}$ gives us

$$\|X - X^\star\|_{\mathrm{F}} \le \rho \lesssim \frac{\sqrt{\sigma_{\min}}}{\log^{6.5}(n)\kappa},$$

where the last step is due to (4).

Then by Weyl's inequality for singular values, for any $i \in [r]$, we have

$$\sigma_i(X) \le \sigma_i(X^\star) + \|X - X^\star\| \le \sigma_1(X^\star) + \|X - X^\star\|_{\mathrm{F}} \le 2\sqrt{\sigma_{\max}}.$$

We also have

$$\sigma_i(X) \geq \sigma_i(X^\star) - \|X - X^\star\| \geq \sigma_r(X^\star) - \|X - X^\star\|_{\mathrm{F}} \geq \frac{\sqrt{\sigma_{\min}}}{2}.$$

The proof for $\sigma_i(Y)$ is identical. □

Now, we present the lemma which extends Assumption 2 to a broader subset of matrices in $\mathcal{B}$.

Lemma 2.6 *Suppose that $(X, Y) \in \mathcal{B}$ and that there exists a rotation matrix $H \in \mathcal{O}^{r\times r}$ such that (XH, YH) is along the gradient flow of function f starting from the point $(X^\star, Y^\star)$. Let m, τ denote the values that minimize $f(X, Y)$. Let $XY^\top = U\Sigma V^\top$ be the SVD of $XY^\top$. Let $\mathbf{T}_0$ be the span of the tangent space of $XY^\top$, and denote by $\mathbf{T}$ the span of $\mathbf{T}_0$ and $\left\{\alpha\mathbf{1}^\top \mid \alpha \in \mathbb{R}^n\right\}$. Define $\tilde{\Delta}^1 \in \mathbb{R}^k$ as the vector with components $\tilde{\Delta}^1_i = \left\langle Z_i, UV^\top \right\rangle$. Define $\tilde{D}$ to be the matrix with entries $\tilde{D}_{ij} = \langle P_{\mathbf{T}^\perp}(Z_i), P_{\mathbf{T}^\perp}(Z_j)\rangle$.*

Assume Assumption 2(a) holds, then for large enough n,

$$\|ZV\|_{\mathrm{F}}^2 + \left\|Z^\top U\right\|_{\mathrm{F}}^2 \leq 1 - \frac{c_{r_1}}{2\log(n)} \tag{8}$$

$$\left\|P_{\mathbf{T}_0{}^\perp}(Z)\right\|_{\mathrm{F}}^2 \geq \frac{c_{r_1}}{2\log(n)} \tag{9}$$

$$\|P_{\mathbf{T}^\perp}(Z) - P_{\mathbf{T}^{\star\perp}}(Z)\|_\star \lesssim \frac{\rho\sqrt{\kappa r}}{\sqrt{\sigma_{\min}}} \tag{10}$$

for any $Z \in \mathbf{T}$ that has $\|Z\|_{\mathrm{F}} = 1$.

Assume Assumption 2(b) holds, then for large enough n,

$$\|\tilde{D}^{-1}\| \left\|\tilde{\Delta}^1\right\| \sum_{i=1}^{k} \|P_{\mathbf{T}^\perp}(Z_i)\| \leq 1 - \frac{c_{r_2}}{2\log n}. \tag{11}$$

Assume Assumption 2(c) holds, then for large enough n,

$$\sigma_{\min}(\tilde{D}) \geq \frac{c_s}{2\log n}. \tag{12}$$

Proof of Lemma 2.6.. We will make use of the following claim throughout this proof:

Claim 2.7

$$\left\|U^\star U^{\star\top} - UU^\top\right\|_{\mathrm{F}}, \left\|V^\star V^{\star\top} - VV^\top\right\|_{\mathrm{F}}, \left\|U^\star V^{\star\top} - UV^\top\right\|_{\mathrm{F}} \lesssim \frac{\rho\sqrt{\kappa}}{\sqrt{\sigma_{\min}}} \lesssim \frac{1}{\log^{6.5}(n)}.$$

The proof of (8) and (9) follow the exact same logic as the proof of (68) and (69) in the proof of Lemma 13 in Farias et al. (2021).

Proof of (10)**.** We use Lemma EC.3 to compute the projections of the treatment matrices. Let $r = (I - VV^\top)\mathbf{1}$ and let $r^\star = (I - V^\star V^{\star\top})\mathbf{1}$. By Lemma 2.4, we have $V^\top\mathbf{1} = V^{\star\top}\mathbf{1} = 0$. Hence,

$r = r^\star = \mathbf{1}$, which allows us to simplify as follows:

$$
\begin{aligned}
&\|P_{\mathbf{T}^\perp}(Z) - P_{\mathbf{T}^{\star\perp}}(Z)\|_\star \\
&= \left\|(I - UU^\top)P_{\mathbf{1}^\perp}(Z)(I - VV^\top) - (I - U^\star U^{\star\top})P_{\mathbf{1}^\perp}(Z)(I - V^\star V^{\star\top})\right\|_\star \\
&\le \left\|(U^\star U^{\star\top} - UU^\top)P_{\mathbf{1}^\perp}(Z)(I - VV^\top)\right\|_\star + \left\|(I - U^\star U^{\star\top})P_{\mathbf{1}^\perp}(Z)(VV^\top - V^\star V^{\star\top})\right\|_\star \\
&\overset{(i)}{\lesssim} \left\|U^\star U^{\star\top} - UU^\top\right\|_{\mathrm{F}} \|P_{\mathbf{1}^\perp}(Z)\|_{\mathrm{F}} \sqrt{r} + \|P_{\mathbf{1}^\perp}(Z)\|_{\mathrm{F}} \left\|VV^\top - V^\star V^{\star\top}\right\|_{\mathrm{F}} \sqrt{r} \\
&\le \left\|U^\star U^{\star\top} - UU^\top\right\|_{\mathrm{F}} \sqrt{r} + \left\|VV^\top - V^\star V^{\star\top}\right\|_{\mathrm{F}} \sqrt{r} \\
&\lesssim \frac{\rho\sqrt{\kappa r}}{\sqrt{\sigma_{\min}}},
\end{aligned}
$$

where (i) is due to $\|A\|_\star \le \|A\|_{\mathrm{F}} \sqrt{rank(A)}$, the fact that projection matrices $(I - VV^\top)$ and $(I - U^\star U^{\star\top})$ have Frobenius norm at most 1. In particular for step (i), we note that the rank of the matrices inside the nuclear norm is $O(r)$ because $U^\star U^{\star\top}$, $UU^\top$, $V^\star V^{\star\top}$, and $VV^\top$ are all rank r matrices, and we have that $rank(A + B) \le rank(A) + rank(B)$ and $rank(AB) \le \min(rank(A), rank(B))$.

Proof of (12)**.** Define $\Delta^D := \tilde{D} - D^\star$ such that $\Delta^D_{ij} = \langle P_{\mathbf{T}^\perp}(Z_i), P_{\mathbf{T}^\perp}(Z_j)\rangle - \langle P_{\mathbf{T}^{\star\perp}}(Z_i), P_{\mathbf{T}^{\star\perp}}(Z_j)\rangle$ for $i, j \in [k]$. We upper bound the magnitude of the entries of Δ^D. For any $i \in [k]$ and $j \in [k]$, we have

$$
\begin{aligned}
\left|\Delta^D_{ij}\right| &= |\langle P_{\mathbf{T}^\perp}(Z_i), P_{\mathbf{T}^\perp}(Z_j)\rangle - \langle P_{\mathbf{T}^{\star\perp}}(Z_i), P_{\mathbf{T}^{\star\perp}}(Z_j)\rangle| \\
&\overset{(i)}{=} |\langle P_{\mathbf{T}^\perp}(Z_i), Z_j\rangle - \langle P_{\mathbf{T}^{\star\perp}}(Z_i), Z_j\rangle| \\
&= |\langle P_{\mathbf{T}^\perp}(Z_i) - P_{\mathbf{T}^{\star\perp}}(Z_i), Z_j\rangle| \\
&\overset{(ii)}{\le} \|P_{\mathbf{T}^\perp}(Z_i) - P_{\mathbf{T}^{\star\perp}}(Z_i)\|_{\mathrm{F}} \\
&= \left\|(I - UU^\top)P_{\mathbf{1}^\perp}(Z_i)(I - VV^\top) - (I - U^\star U^{\star\top})P_{\mathbf{1}^\perp}(Z_i)(I - V^\star V^{\star\top})\right\|_{\mathrm{F}} \\
&= \left\|(U^\star U^{\star\top} - UU^\top)P_{\mathbf{1}^\perp}(Z_i)(I - VV^\top) - (I - U^\star U^{\star\top})P_{\mathbf{1}^\perp}(Z_i)(VV^\top - V^\star V^{\star\top})\right\|_{\mathrm{F}} \\
&\le \left\|U^\star U^{\star\top} - UU^\top\right\|_{\mathrm{F}} + \left\|V^\star V^{\star\top} - VV^\top\right\|_{\mathrm{F}} \\
&\overset{(iii)}{\lesssim} \frac{1}{\log^{6.5}(n)},
\end{aligned}
\tag{13}
$$

The equality (i) is due to the fact that for any projection operator P and matrices A and B, we have $\langle P(A), B\rangle = \langle P(A), P(B)\rangle = \langle A, P(B)\rangle$, and (ii) is due to the Cauchy–Schwarz inequality and the fact that the $\|Z_i\|_{\mathrm{F}} = 1$, and (iii) follows from Claim 2.7.

By Weyl's inequality, we have $\sigma_{\min}(\tilde{D}) \ge \sigma_{\min}(D^\star) - \sigma_{\max}(\Delta^D)$. Now we upper bound the second term as follows:

$$
\sigma_{\max}(\Delta^D) = \left\|\Delta^D\right\| \le k \max_{l,m} \left|\Delta^D_{l,m}\right| = O\left(\frac{1}{\log^{5.5}(n)}\right). \tag{14}
$$

Putting everything together, for large enough n,

$$
\sigma_{\min}(\tilde{D}) \ge \frac{c_s}{\log n} - \sigma_{\max}(\Delta^D) \ge \frac{c_s}{2\log n}.
$$

Proof of (11)**.** We wish to bound $A := \left\|\tilde{D}^{-1}\right\| \left\|\tilde{\Delta}^1\right\| \sum_{i=1}^k \|P_{\mathbf{T}^\perp}(Z_i)\|$. This quantity can be restructured as

$$A = \left(\left\|D^{\star-1}\right\| \left\|\Delta^{\star 1}\right\| + E_1\right) \left(\sum_{i=1}^k \|P_{\mathbf{T}^{\star\perp}}(Z_i)\| + E_2\right), \tag{15}$$

where $D^\star$ and $\Delta^{\star 1}$ are defined just above Assumption 2, and E_1 and E_2 are defined as follows:

$$E_1 := \left\|\tilde{D}^{-1}\right\| \left\|\tilde{\Delta}^1\right\| - \left\|D^{\star-1}\right\| \left\|\Delta^{\star 1}\right\|$$
$$E_2 := \sum_{i=1}^k (\|P_{\mathbf{T}^\perp}(Z_i)\| - \|P_{\mathbf{T}^{\star\perp}}(Z_i)\|).$$

We bound these two quantities as follows.

Claim 2.8 *We have that* $|E_1| \lesssim \frac{1}{\log^{2.5} n}$ *and* $|E_2| \lesssim \frac{1}{\log^{5.5} n}$.

To bound A, we consider the following additional observations. Firstly, we have $\|D^{\star-1}\| \|\Delta^{\star 1}\| \lesssim \log^2 n$. This follows from the fact that $\sigma_{\min}(D^\star) \gtrsim \frac{1}{\log n}$, and $\Delta^{\star 1}$ is a vector of length k (where $k \lesssim \log n$) with entries that do not exceed 1. Moreover, we have $\sum_{i=1}^k \|P_{\mathbf{T}^{\star\perp}}(Z_i)\| \leq k \lesssim \log n$ because for every $i \in [k]$, $P_{\mathbf{T}^{\star\perp}}(Z_i)$ has Frobenius norm (and hence spectral norm) at most 1. Now, we are ready to plug all of this into (15) to bound A as follows.

$$\begin{aligned} A &\leq \left\|D^{\star-1}\right\| \left\|\Delta^{\star 1}\right\| \sum_{i=1}^k \|P_{\mathbf{T}^{\star\perp}}(Z_i)\| + O\left(\frac{1}{\log^{1.5} n}\right) \\ &\overset{(i)}{\leq} 1 - \frac{c_{r_2}}{\log n} + O\left(\frac{1}{\log^{1.5} n}\right) \\ &\leq 1 - \frac{c_{r_2}}{2\log n}, \end{aligned}$$

where (i) is due to Assumption 2(b). □

Proof of Claim 2.7. First, we note that we have

$$\begin{aligned} \left\|XY^\top - X^\star Y^{\star\top}\right\|_{\mathrm{F}} &= \left\|XH_{X,Y}(YH_{X,Y})^\top - X^\star Y^{\star\top}\right\|_{\mathrm{F}} \\ &\leq \|XH_{X,Y} - X^\star\|_{\mathrm{F}} \|Y\| + \|X^\star\| \left\|YH_{X,Y} - Y^\star\right\|_{\mathrm{F}} \\ &\lesssim \rho\sqrt{\sigma_{\max}}, \end{aligned}$$

where $\|Y\|$ is bounded by Lemma 2.5. Hence, we can make use of Theorem 2 from Yu et al. (2015), combined with the symmetric dilation trick from section C.3.2 of Abbe et al. (2020), to obtain the following. There exists an orthogonal matrix $O \in \mathbb{R}^{r\times r}$ such that

$$\|UO - U^\star\|_{\mathrm{F}} + \|VO - V^\star\|_{\mathrm{F}} \lesssim \frac{\left\|XY^\top - X^\star Y^{\star\top}\right\|_{\mathrm{F}}}{\sigma_r(X^\star Y^{\star\top}) - \sigma_{r+1}(X^\star Y^{\star\top})} \lesssim \frac{\rho\sqrt{\sigma_{\max}}}{\sigma_{\min}} = \frac{\rho\sqrt{\kappa}}{\sqrt{\sigma_{\min}}}.$$

Now we can obtained the desired bound as follows:

$$\begin{aligned}\left\|UV^\top - U^\star V^{\star\top}\right\|_{\mathrm{F}} &= \left\|UO(VO)^\top - U^\star V^{\star\top}\right\|_{\mathrm{F}} \\ &\le \|UO - U^\star\|_{\mathrm{F}} \|VO\| + \|VO - V^\star\|_{\mathrm{F}} \|U^\star\| \\ &\le \|UO - U^\star\|_{\mathrm{F}} + \|VO - V^\star\|_{\mathrm{F}} \\ &\lesssim \frac{\rho\sqrt{\kappa}}{\sqrt{\sigma_{\min}}} \\ &\lesssim \frac{1}{\log^{6.5} n} \qquad \text{using (4).}\end{aligned}$$

We can similarly bound both $\left\|UU^\top - U^\star U^{\star\top}\right\|_{\mathrm{F}}$ and $\left\|VV^\top - V^\star V^{\star\top}\right\|_{\mathrm{F}}$. □

Proof of Claim 2.8. $|E_2|$ can be simply bounded as follows:

$$|E_2| \le \sum_{i=1}^{k} \|P_{\mathbf{T}^\perp}(Z_i) - P_{\mathbf{T}^{\star\perp}}(Z_i)\|_{\mathrm{F}} \lesssim \frac{k}{\log^{6.5} n},$$

where the last bound was shown in (13). Because $k = O(\log n)$, we have the desired bound for E_2.

It remains to bound $|E_1|$:

$$\begin{aligned}|E_1| &= \left|\|D^{\star-1}\|\|\Delta^{\star 1}\| - \|\tilde{D}^{-1}\|\|\tilde{\Delta}^1\|\right| \\ &\le \|D^{\star-1}\| \left|\|\Delta^{\star 1}\| - \|\tilde{\Delta}^1\|\right| + \left|\|D^{\star-1}\| - \|\tilde{D}^{-1}\|\right| \|\tilde{\Delta}^1\| \\ &\le \underbrace{\|D^{\star-1}\|\|\Delta^{\star 1} - \tilde{\Delta}^1\|}_{A_1} + \underbrace{\|D^{\star-1} - \tilde{D}^{-1}\|\|\tilde{\Delta}^1\|}_{A_2}.\end{aligned}$$

Bounding A_1. Applying Assumption 2(c), we can bound A_1 as follows:

$$\begin{aligned}A_1 &\lesssim \log(n) \left\|\Delta^{\star 1} - \tilde{\Delta}^1\right\| \\ &\le \log(n)k \left\|\Delta^{\star 1} - \tilde{\Delta}^1\right\|_\infty \\ &= \log(n)k \max_{i\in[k]} \left|\left\langle Z_i, U^\star V^{\star\top} - UV^\top \right\rangle\right| \\ &\le \log(n)k \left\|U^\star V^{\star\top} - UV^\top\right\|_{\mathrm{F}} \\ &\lesssim \frac{1}{\log^{4.5}(n)},\end{aligned}$$

where the last step makes use of Claim 2.7.

Bounding A_2. Because $\tilde{D} = D^\star + \Delta^D = (I + \Delta^D D^{\star-1})D^\star$, we can write

$$\tilde{D}^{-1} = D^{\star-1} \sum_{k=0}^{\infty} (-\Delta^D D^{\star-1})^k.$$

Note that this quantity is well-defined because

$$\left\|\Delta^D D^{\star-1}\right\| \le \frac{\sigma_{\max}(\Delta^D)}{\sigma_{\min}(D^\star)} \lesssim \frac{1}{\log^{4.5} n} < 1,$$

where the last bound is due to (14) and Assumption 2(c). Now, we have

$$D^{\star-1} - \tilde{D}^{-1} = D^{\star-1}\Delta^D D^{\star-1} \sum_{k=0}^{\infty} (-\Delta^D D^{\star-1})^k.$$

We use this to bound A_2. Using (14) and Assumption 2(c), and the fact that $\|\tilde{\Delta}^1\| \leq k$ because it is a vector of length k with all components less than or equal to 1, we have

$$A_2 \leq \frac{\sigma_{\max}(\Delta^D)}{\sigma_{\min}(D^\star)^2} \cdot \frac{1}{1 - \frac{\sigma_{\max}(\Delta^D)}{\sigma_{\min}(D^\star)}} \cdot \|\tilde{\Delta}^1\| \lesssim \frac{1}{\log^{2.5} n}.$$

□

2.3 Discussion of $\hat{E}$

Recall that the matrix $\hat{E}$ consists of two components $\hat{E} = E + \delta$. E is a noise matrix, and δ is the matrix of the approximation error. Refer to Assumption 3 for our assumptions on E and δ.

We need to ensure that the error $\hat{E}$ does not significantly interfere with the recovery of the treatment effects. That is, we need to ensure that $\hat{E}$ is not confounded with the treatment matrices Z_i for $i \in [k]$. This condition is formalized and established in the following lemma.

Lemma 2.9 *Let $(X, Y) \in \mathcal{B}$ be along the gradient flow of function f starting from the point $(X^\star, Y^\star)$, and let m, τ denote the values that minimize $f(X, Y)$. Let $\mathbf{T}$ be the span of the tangent space of $XY^\top$ and $\left\{\alpha \mathbf{1}^\top \mid \alpha \in \mathbb{R}^n\right\}$. Then, we have*

$$\max_{i \in [k]} \left| \left\langle P_{\mathbf{T}^\perp}(Z_i), \hat{E} \right\rangle \right| \lesssim \sigma\sqrt{nr}.$$

Proof of Lemma 2.9. Fix any $i \in [k]$. We recall that $\hat{E} = E + \delta$. We note that $rank(P_{\mathbf{T}}(A)) \leq 2r+1$ for any matrix A due to the definition of $\mathbf{T}$. Then,

$$\begin{aligned}
\left| \left\langle P_{\mathbf{T}^\perp}(Z_i), \hat{E} \right\rangle \right| &\leq |\langle Z_i, E \rangle| + |\langle Z_i, \delta \rangle| + \left| \left\langle P_{\mathbf{T}}(Z_i), \hat{E} \right\rangle \right| \\
&\overset{(i)}{\lesssim} \sigma\sqrt{n} + \left| \left\langle P_{\mathbf{T}}(Z_i), \hat{E} \right\rangle \right| \\
&\overset{(ii)}{\leq} \sigma\sqrt{n} + \|\hat{E}\| \, \|P_{\mathbf{T}}(Z_i)\|_\star \\
&\overset{(iii)}{\leq} \sigma\sqrt{n} + (\|E\| + \|\delta\|)\sqrt{2r+1} \, \|P_{\mathbf{T}}(Z_i)\|_{\mathrm{F}} \\
&\overset{(iv)}{\lesssim} \sigma\sqrt{nr}.
\end{aligned}$$

Here, (i) is due to Assumption 3 $|\langle Z_i, E \rangle|, |\langle Z_i, \delta \rangle| \lesssim \sigma\sqrt{n}$; (ii) is due to Von Neumann's trace inequality; (iii) is due to the inequality $\|A\|_\star \leq \sqrt{rank(A)} \, \|A\|_{\mathrm{F}}$; and (iv) is due to the fact that $\|Z_i\|_{\mathrm{F}} = 1$, and our assumed upper bound on $\|E\|$ and $\|\delta\|$ from Assumption 3. □

We now argue that the assumptions on the noise matrix E from Assumption 3 are mild. The following lemma shows that under standard sub-Gaussianity assumptions, these are satisfied with very high probability.

Lemma 2.10 *Suppose that the entries of E are independent, zero-mean, sub-Gaussian random variables, and the sub-Gaussian norm of each entry is bounded by σ, and E is independent from Z_i for $i \in [k]$. Then, for n sufficiently large, with probability at least $1 - e^{-n}$, we have*

$$\|E\| \lesssim \sigma\sqrt{n} \quad \textit{and} \quad |\langle Z_i, E\rangle| \lesssim \sigma\sqrt{n}, \quad \forall i \in [k].$$

Proof of Lemma 2.10. We start by proving that with probability at least $1 - 2e^{-n}$, we have $\|E\| \lesssim \sigma\sqrt{n}$. We use the following result bounding the norm of matrices with sub-Gaussian entries.

Theorem 2.11 (Theorem 4.4.5, Vershynin (2018)) *Let A be an $m \times n$ random matrix whose entries A_{ij} are independent, mean zero, sub-Gaussian random variables with sub-Gaussian norm bounded by σ. Then, for any $t > 0$, we have*

$$\|A\| \lesssim \sigma(\sqrt{m} + \sqrt{n} + t)$$

with probability at least $1 - 2\exp(-t^2)$.

As a result of the above theorem, recalling that E is a $n \times T$ matrix with $T \lesssim n$ and using $t = \sqrt{2n}$, we have $\|E\| \lesssim \sigma\sqrt{n}$ with probability at least $1 - 2\exp(-2n)$.

We next turn to the second claim. The general version of Hoeffding's inequality states that: for zero-mean independent sub-Gaussian random variables $X_1, \ldots, X_n$, we have

$$\Pr\left[\left|\sum_{i=1}^{n} X_i\right| \geq t\right] \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^{n} \|X_i\|_{\psi_2}^2}\right),$$

for some absolute constant $c > 0$. Hence, for any $i \in [k]$:

$$\Pr\left[|\langle Z_i, E\rangle| \geq \sigma\sqrt{2n/c}\right] \leq 2\exp\left(-\frac{2\sigma^2 n}{\sigma^2 \|Z_i\|_{\mathrm{F}}^2}\right) \leq 2e^{-2n}.$$

Applying the Union Bound over all $i \in [k]$:

$$\Pr\left[\max_{i\in[k]} |\langle Z_i, E\rangle| \geq \sigma\sqrt{n/c}\right] \leq 2ke^{-2n} \lesssim 2\log(n)e^{-2n}.$$

Applying the Union Bound shows that the two desired claims hold with probability at least $1 - 2e^{-2n}(1 + \log n) \geq 1 - e^{-n}$ for sufficiently large n. □

2.4 Proof of Lemma 2.1

If the gradient flow of f started at $(X^\star, Y^\star)$ never intersects the boundary of $\mathcal{B}$, then we are done. For the rest of the proof, we will suppose that such an intersection exists. Fix any point (X, Y) at the intersection of the boundary of $\mathcal{B}$ and the gradient flow of f. We aim to show that at this boundary point, the inner product of the normal vector to the region (pointing to the exterior of the region) and the gradient of f is positive. This property implies that the gradient flow of f, initiated from any point inside $\mathcal{B}$, cannot exit $\mathcal{B}$. Proving this characteristic is sufficient to prove Lemma 2.1.

We denote by H the rotation matrix $H_{X,Y}$ that optimally aligns the fixed boundary point (X, Y) to $(X^\star, Y^\star)$. Consider

$$\mathcal{B}_H = \left\{ (X', Y') \mid \|X'H - X^\star\|_{\mathrm{F}}^2 + \|Y'H - Y^\star\|_{\mathrm{F}}^2 \le \rho^2 \right\}.$$

Note that $\mathcal{B}_H$ differs from $\mathcal{B}$ because H is fixed to be the rotation matrix associated with a given point (X, Y). $\mathcal{B}_H$ and $\mathcal{B}$ are tangent to each other at (X, Y), so the normal vector to $\mathcal{B}_H$ and $\mathcal{B}$ are co-linear at (X, Y). Thus, it suffices to show that the normal vector to $\mathcal{B}_H$ at (X, Y), and the gradient of f at (X, Y) have positive inner product.

To simplify our notation, we define

$$F := \begin{bmatrix} X \\ Y \end{bmatrix}, \qquad F^\star := \begin{bmatrix} X^\star \\ Y^\star \end{bmatrix}, \qquad \text{and} \qquad \frac{\partial f}{\partial F} := \begin{bmatrix} \frac{\partial f}{\partial X} \\ \frac{\partial f}{\partial Y} \end{bmatrix}.$$

Additionally, we define $\Delta_X := XH - X^\star$, $\Delta_Y := YH - Y^\star$, and $\Delta_{\mathrm{F}} := FH - F^\star$.

The normal vector to $\mathcal{B}$ at (X, Y) is simply the subgradient of $\|XH - X^\star\|_{\mathrm{F}}^2 + \|YH - Y^\star\|_{\mathrm{F}}^2$ at (X, Y), which is $2(FH - F^\star)H^\top$. Hence, we aim to show that the following inner product is positive:

$$\begin{aligned}\Gamma := \left\langle (FH - F^\star)H^\top, \frac{\partial f}{\partial F} \right\rangle = &\left\langle \Delta_X, P_{\mathbf{Z}^\perp}\left(XY^\top - X^\star Y^{\star\top} - \hat{E}\right) YH + \lambda XH \right\rangle \\ &+ \left\langle \Delta_Y, P_{\mathbf{Z}^\perp}\left(XY^\top - X^\star Y^{\star\top} - \hat{E}\right)^\top XH + \lambda YH \right\rangle,\end{aligned}$$

where we used the formula for the gradient of f from Eq (7). In the remainder of this subsection, we will use the following shorthand notations for simplicity: we denote XH, YH, and FH as X, Y, and F respectively. The above inner product Γ becomes:

$$\Gamma = \underbrace{\left\langle \Delta_X Y^\top + X\Delta_Y^\top, P_{\mathbf{Z}^\perp}\left(XY^\top - X^\star Y^{\star\top}\right) \right\rangle}_{A_3} - \underbrace{\left\langle \Delta_X, P_{\mathbf{Z}^\perp}(\hat{E})Y \right\rangle - \left\langle \Delta_Y, P_{\mathbf{Z}^\perp}(\hat{E})^\top X \right\rangle}_{A_1} + \underbrace{\langle \Delta_X, \lambda X \rangle + \langle \Delta_Y, \lambda Y \rangle}_{A_2}.$$

By Assumption 3(a), we have $\|\hat{E}\| \le \|E\| + \|\delta\| \lesssim \sigma\sqrt{n}$. In particular,

$$\begin{aligned}|A_1| &\lesssim \|\Delta_X\|_{\mathrm{F}} \|\hat{E}\| \|Y\|_{\mathrm{F}} + \|\Delta_Y\|_{\mathrm{F}} \|\hat{E}\| \|X\|_{\mathrm{F}} \\ &\lesssim \|\hat{E}\| \|\Delta_{\mathrm{F}}\|_{\mathrm{F}} \|F\|_{\mathrm{F}} \\ &\overset{(i)}{\lesssim} \sigma\sqrt{n} \|\Delta_{\mathrm{F}}\|_{\mathrm{F}} \|F^\star\|_{\mathrm{F}},\end{aligned}$$

where (i) is because $\|F\|_{\mathrm{F}} - \|F^\star\|_{\mathrm{F}} \le \|\Delta_{\mathrm{F}}\|_{\mathrm{F}}$ by the Triangle Inequality, and $(X, Y) \in \mathcal{B}$ gives $\|\Delta_{\mathrm{F}}\|_{\mathrm{F}} \le \rho \lesssim \|F^\star\|_{\mathrm{F}}$, where the last step is due to the assumed bound on $\frac{\sigma\sqrt{n}}{\sigma_{\min}}$.

Recalling $\lambda = \Theta\left(\sigma\sqrt{nr}\log^{4.5}(n)\right)$, we bound $|A_2|$:

$$\begin{aligned}|A_2| &\le \lambda \|\Delta_X\|_{\mathrm{F}} \|X\|_{\mathrm{F}} + \lambda \|\Delta_Y\|_{\mathrm{F}} \|Y\|_{\mathrm{F}} \\ &\le 2\lambda \|\Delta_{\mathrm{F}}\|_{\mathrm{F}} \|F\|_{\mathrm{F}} \\ &\lesssim \sigma\sqrt{nr}\log^{4.5}(n) \|\Delta_{\mathrm{F}}\|_{\mathrm{F}} \|F^\star\|_{\mathrm{F}}.\end{aligned}$$

Finally, we will lower bound A_3. For some $Z \in \mathbf{Z}$ with $\|Z\|_{\mathrm{F}} = 1$, we have the following:

$$\begin{aligned}
A_3 &= \left\langle \Delta_X Y^\top + X\Delta_Y^\top, P_{\mathbf{Z}^\perp}\left(XY^\top - X^\star Y^{\star\top}\right)\right\rangle \\
&= \left\langle \Delta_X Y^\top + X\Delta_Y^\top, P_{\mathbf{Z}^\perp}\left(\Delta_X Y^\top + X\Delta_Y^\top - \Delta_X\Delta_Y^\top\right)\right\rangle \\
&= \left\|P_{\mathbf{Z}^\perp}\left(\Delta_X Y^\top + X\Delta_Y^\top\right)\right\|_{\mathrm{F}}^2 - \underbrace{\left\langle \Delta_X Y^\top + X\Delta_Y^\top, P_{\mathbf{Z}^\perp}\left(\Delta_X\Delta_Y^\top\right)\right\rangle}_{B_0} \\
&= \left\|\Delta_X Y^\top + X\Delta_Y^\top\right\|_{\mathrm{F}}^2 - \left\|P_{\mathbf{Z}}\left(\Delta_X Y^\top + X\Delta_Y^\top\right)\right\|_{\mathrm{F}}^2 - B_0 \\
&= \left\|\Delta_X Y^\top + X\Delta_Y^\top\right\|_{\mathrm{F}}^2 - \left\langle Z, \Delta_X Y^\top + X\Delta_Y^\top\right\rangle^2 - B_0 \\
&\overset{(i)}{=} \|Z\|_{\mathrm{F}}^2 \left\|\Delta_X Y^\top + X\Delta_Y^\top\right\|_{\mathrm{F}}^2 - \left\langle P_{\mathbf{T}_0}(Z), \Delta_X Y^\top + X\Delta_Y^\top\right\rangle^2 - B_0 \\
&\geq \|Z\|_{\mathrm{F}}^2 \left\|\Delta_X Y^\top + X\Delta_Y^\top\right\|_{\mathrm{F}}^2 - \|P_{\mathbf{T}_0}(Z)\|_{\mathrm{F}}^2 \left\|\Delta_X Y^\top + X\Delta_Y^\top\right\|_{\mathrm{F}}^2 - B_0 \\
&= \left\|P_{\mathbf{T}_0{}^\perp}(Z)\right\|_{\mathrm{F}}^2 \left\|\Delta_X Y^\top + X\Delta_Y^\top\right\|_{\mathrm{F}}^2 - B_0 \\
&\overset{(ii)}{\geq} \frac{c_{r_1}}{2\log(n)} \left\|\Delta_X Y^\top + X\Delta_Y^\top\right\|_{\mathrm{F}}^2 - B_0,
\end{aligned}$$

where (i) is due to the fact that $\|Z\|_{\mathrm{F}} = 1$ and $\Delta_X Y^\top + X\Delta_Y^\top$ is already in the subspace $\mathbf{T}_0$ by the definition of $\mathbf{T}_0 = \left\{UA^\top + BV^\top \mid A \in \mathbb{R}^{T\times r}, B \in \mathbb{R}^{n\times r}\right\}$ as the tangent space of $XY^\top$; (ii) is due to Lemma 2.6. Note that Lemma 2.6 applies because $\mathbf{T}_0$ is the tangent space of $XY^\top$, and (X, Y) was fixed to be on the gradient flow. Furthermore, we can bound $|B_0|$ as follows:

$$\begin{aligned}
|B_0| &\leq \|\Delta_X\|_{\mathrm{F}} \|\Delta_Y\|_{\mathrm{F}} \left(\left\|\Delta_X Y^\top\right\|_{\mathrm{F}} + \left\|X\Delta_Y^\top\right\|_{\mathrm{F}}\right) \\
&\leq \|\Delta_X\|_{\mathrm{F}} \|\Delta_Y\|_{\mathrm{F}} \left(\|\Delta_X\|_{\mathrm{F}} \|Y\| + \|\Delta_Y\|_{\mathrm{F}} \|X\|\right) \\
&\overset{(i)}{\lesssim} \|\Delta_{\mathrm{F}}\|_{\mathrm{F}}^3 \|F^\star\| \\
&\leq \rho^2 \|\Delta_{\mathrm{F}}\|_{\mathrm{F}} \|F^\star\| \\
&\overset{(ii)}{\lesssim} \frac{\sigma^2 r^2 \kappa n \log^{12}(n)}{\sigma_{\min}} \|\Delta_{\mathrm{F}}\|_{\mathrm{F}} \|F^\star\| \\
&\lesssim \sigma\sqrt{n} \|\Delta_{\mathrm{F}}\|_{\mathrm{F}} \|F^\star\|,
\end{aligned}$$

where (i) makes use of Lemma 2.5, and (ii) follows from the bound on ρ from (3), and the last step follows from the bound on $\frac{\sigma\sqrt{n}}{\sigma_{\min}}$.

Now, we are ready to put everything together to bound the inner product:

$$\begin{aligned}
\Gamma &= A_3 - A_1 + A_2 \\
&\geq \frac{c_{r_1}}{2\log(n)} \left\|\Delta_X Y^\top + X\Delta_Y^\top\right\|_{\mathrm{F}}^2 - B_0 - A_1 + A_2
\end{aligned}$$

Putting previous bounds together, we have

$$|A_1| + |A_2| + |B_0| \lesssim \sigma\sqrt{nr}\log^{4.5}(n) \|\Delta_{\mathrm{F}}\|_{\mathrm{F}} \|F^\star\|_{\mathrm{F}}.$$

On the other hand, we can lower bound the positive term of Γ using the following lemma.

Lemma 2.12 *Consider a point (X, Y) on the gradient flow of function f starting from the point $(X^\star, Y^\star)$, such that $(X, Y) \in \mathcal{B}$. Then,*

$$\left\|\Delta_X Y^\top + X\Delta_Y^\top\right\|_{\mathrm{F}}^2 \geq \frac{\sigma_{\min}}{4}\left\|\Delta_{\mathrm{F}}\right\|_{\mathrm{F}}^2.$$

By Lemma 2.12 we have

$$\frac{c_{r_1}}{2\log(n)}\left\|\Delta_X Y^\top + X\Delta_Y^\top\right\|_{\mathrm{F}}^2 \geq \frac{c_{r_1}}{2\log(n)}\cdot\frac{\sigma_{\min}}{4}\left\|\Delta_{\mathrm{F}}\right\|_{\mathrm{F}}^2$$

We plug in $\left\|\Delta_{\mathrm{F}}\right\|_{\mathrm{F}} = \rho$ because we assumed that (X, Y) is at the border of region $\mathcal{B}$. Hence,

$$\begin{aligned}\frac{c_{r_1}}{2\log(n)}\left\|\Delta_X Y^\top + X\Delta_Y^\top\right\|_{\mathrm{F}}^2 &\gtrsim \frac{\sigma_{\min}}{\log(n)}\left\|\Delta_{\mathrm{F}}\right\|_{\mathrm{F}}\cdot\rho \\ &= \sigma\sqrt{nr}\log^5(n)\left\|\Delta_{\mathrm{F}}\right\|_{\mathrm{F}}\left\|F^\star\right\|_{\mathrm{F}}.\end{aligned}$$

Finally, for large enough n,

$$\Gamma \gtrsim \sigma\sqrt{nr}\log^5(n)\left\|\Delta_{\mathrm{F}}\right\|_{\mathrm{F}}\left\|F^\star\right\|_{\mathrm{F}} - \sigma\sqrt{nr}\log^{4.5}(n)\left\|\Delta_{\mathrm{F}}\right\|_{\mathrm{F}}\left\|F^\star\right\|_{\mathrm{F}} > 0.$$

Proof of Lemma 2.12. Claim 11 in Farias et al. (2021) states

$$\left\|\Delta_X Y^\top + X\Delta_Y^\top\right\|_{\mathrm{F}}^2 \geq \frac{\sigma_{\min}}{4}\left\|\Delta_{\mathrm{F}}\right\|_{\mathrm{F}}^2 - \frac{\sigma^2}{n^{13}}. \tag{16}$$

We can follow the steps of their proof to prove the desired statement. In particular, we note that the term $-\sigma^2/n^{13}$ in the right-hand side of Eq (16) comes from the fact that they use the bound

$$\left\|X^\top X - Y^\top Y\right\|_{\mathrm{F}} \lesssim \frac{\sigma}{\kappa n^{15}}.$$

Because we have, by Lemma 2.3 that $X^\top X = Y^\top Y$, we do not incur the term $-\sigma^2/n^{13}$ in our lower bound. □

2.5 Proof of Lemma 2.2

Let $g(M, \tau, m)$ denote the convex function that we are optimizing in (3). Recall that $(\hat{M}, \hat{\tau}, \hat{m})$ is a global optimum of function g. In this subsection, we prove Lemma 2.2, which relates $(\hat{M}, \hat{\tau}, \hat{m})$ to a local optimum of f.

We begin by proving the following useful lemma, which is similar to Lemma 20 from Chen et al. (2020), but allows X, Y to have different dimensions.

Lemma 2.13 *Consider matrices X and Y such that $X^\top X = Y^\top Y$. There is an SVD of $XY^\top$ denoted by $U\Sigma V^\top$ such that $X = U\Sigma^{1/2}R$ and $Y = V\Sigma^{1/2}R$ for some rotation matrix $R \in \mathcal{O}^{r\times r}$.*

Proof. Let $X = U_X\Sigma_X V_X^\top$ and $Y = U_Y\Sigma_Y V_Y^\top$ be their respective SVDs, ordering the diagonal components of Σ_X and Σ_Y by decreasing order. Then, $X^\top X = Y^\top Y$ implies $\Sigma_X = \Sigma_Y$ and that the singular subspaces of V_X and V_Y coincide. Hence, there exists an SVD decomposition of $Y = \tilde{U}_Y\Sigma_Y\tilde{V}_Y^\top$ such that $V_X = \tilde{V}_Y$. Then, $XY^\top = U_X\Sigma_X^2\tilde{U}_Y^\top$. This is an SVD of $XY^\top$, with

$U = U_X$, $\Sigma = \Sigma_X^2$, and $V = \tilde{U}_Y$. Substituting these quantities into the SVD of X and Y, we complete the proof that $X = U\Sigma^{1/2}R$ and $Y = V\Sigma^{1/2}R$, where $R = V_X = \tilde{V}_Y \in \mathcal{O}^{r\times r}$. □

Let (X, Y) represent the limit of the gradient flow of f from the initial point $(X^\star, Y^\star)$. Let m and τ be the values that minimize $f(X, Y)$. Furthermore, let the SVD of $XY^\top$ be denoted by $U\Sigma V^\top$.

By Lemma 2.3, we have that $X^\top X = Y^\top Y$. Then, Lemma 2.13 gives us

$$(I - UU^\top)X = 0 \qquad \text{and} \qquad (I - VV^\top)Y = 0. \tag{17}$$

We claim that to prove Lemma 2.2, it suffices to prove that $XY^\top = \hat{M}$. If we prove that $XY^\top = \hat{M}$, we would have by Lemma 2.13 $X = \hat{X}R$ and $Y = \hat{Y}R$ for some rotation matrix $R \in \mathcal{O}^{r\times r}$. This proves Lemma 2.2.

The proof that $XY^\top = \hat{M}$ consists of two parts. We will first establish that $(XY^\top, \tau, m)$ is also an optimal point of g by verifying the first order conditions of g are satisfied. We will then show that g has a *unique* optimal solution $(\hat{M}, \hat{\tau}, \hat{m})$. Putting these two parts together establishes that $XY^\top = \hat{M}$.

2.5.1 First order conditions of g are satisfied

We will first show that the following first order conditions of g are satisfied at $(XY^\top, \tau, m)$.

$$\left\langle Z_l, O - XY^\top - m\mathbf{1}^\top - \sum_{i=1}^k \tau_i Z_i \right\rangle = 0 \qquad \text{for } l = 1, 2, \ldots, k \tag{18a}$$

$$O - XY^\top - m\mathbf{1}^\top - \sum_{i=1}^k \tau_i Z_i = \lambda\left(UV^\top + W\right) \tag{18b}$$

$$U^\top W = 0 \tag{18c}$$

$$WV = 0 \tag{18d}$$

$$\|W\| \le 1 \tag{18e}$$

$$m = \frac{1}{T}\left(O - XY^\top - \sum_{i=1}^k \tau_i Z_i\right)\mathbf{1}. \tag{18f}$$

We select $W := \frac{1}{\lambda}\left(O - XY^\top - m\mathbf{1}^\top - \sum_{i=1}^k \tau_i Z_i\right) - UV^\top$. Note that (18a) and (18f) are automatically satisfied given the definition of τ and m, and (18b) is automatically satisfied by our choice of W.

To show (18c) and (18d), we use the fact that $\frac{\partial f}{\partial X} = \frac{\partial f}{\partial Y} = 0$:

$$\left(O - XY^\top - m\mathbf{1}^\top - \sum_{i=1}^k \tau_i Z_i\right) Y = \lambda X \quad \text{and} \quad \left(O - XY^\top - m\mathbf{1}^\top - \sum_{i=1}^k \tau_i Z_i\right)^\top X = \lambda Y$$

Now, by Lemma 2.13, we can decompose $X = U\Sigma^{1/2}R$ and $Y = V\Sigma^{1/2}R$ where R is a rotation matrix. Right-multiplying the above equations by $R^{-1}\Sigma^{-1/2}$ gives

$$\left(O - XY^\top - m\mathbf{1}^\top - \sum_{i=1}^k \tau_i Z_i\right) V = \lambda U \quad \text{and} \quad \left(O - XY^\top - m\mathbf{1}^\top - \sum_{i=1}^k \tau_i Z_i\right)^\top U = \lambda V.$$

Rearranging, the first equation shows that $WV = 0$ and the second shows $U^\top W = 0$.

The last step of the proof is to verify (18e). Using (18c) and (18d), we have

$$\begin{aligned} W &= (I - UU^\top)W(I - VV^\top) \\ &= \frac{1}{\lambda}(I - UU^\top)\left(O - m\mathbf{1}^\top - \sum_{i=1}^{k} \tau_i Z_i\right)(I - VV^\top), \end{aligned}$$

where the last line is obtained by plugging in our chosen value of W and using (17) to get rid of the $XY^\top$ term. Plugging in $O = X^\star Y^{\star\top} + m^\star \mathbf{1}^\top + \sum_{i=1}^{k} \tau_i^\star Z_i + \hat{E}$, we have

$$W = \frac{1}{\lambda}(I - UU^\top)\left(X^\star Y^{\star\top} + (m^\star - m)\mathbf{1}^\top + \sum_{i=1}^{k}(\tau_i^\star - \tau_i)Z_i + \hat{E}\right)(I - VV^\top).$$

We will use substitution to get rid of the $(m^\star - m)\mathbf{1}^\top$ term. Because we have $m^\star = \frac{M^\star \mathbf{1}}{T}$ and

$$m = \frac{1}{T}\left(O - XY^\top - \sum_{i=1}^{k} \tau_i Z_i\right)\mathbf{1} = \frac{1}{T}\left(M^\star - XY^\top + \sum_{i=1}^{k}(\tau_i^\star - \tau_i)Z_i + \hat{E}\right)\mathbf{1},$$

this implies that

$$(m^\star - m)\mathbf{1}^\top = \left(XY^\top - \sum_{i=1}^{k}(\tau_i^\star - \tau_i)Z_i - \hat{E}\right)\frac{\mathbf{1}\mathbf{1}^\top}{T} = P_{\mathbf{1}}\left(XY^\top - \sum_{i=1}^{k}(\tau_i^\star - \tau_i)Z_i - \hat{E}\right).$$

Substituting this expression into our expression for W, we have

$$W = \frac{1}{\lambda}(I - UU^\top)\left(X^\star Y^{\star\top} + \sum_{i=1}^{k}(\tau_i^\star - \tau_i)P_{\mathbf{1}^\perp}(Z_i) + P_{\mathbf{1}^\perp}(\hat{E})\right)(I - VV^\top)$$

where the $P_{\mathbf{1}}(XY^\top)$ term went away because $(I - UU^\top)X = 0$ by (17).

By Lemma 2.4, we have $V^\top \mathbf{1} = 0$. This allows us to simplify the closed form expression for the projection given in Lemma EC.3: $P_{\mathbf{T}^\perp}(A) = (I - UU^\top)P_{\mathbf{1}^\perp}(A)(I - VV^\top)$. Hence,

$$W = \frac{1}{\lambda}(I - UU^\top)\, X^\star Y^{\star\top}(I - VV^\top) + \frac{1}{\lambda}P_{\mathbf{T}^\perp}(\hat{E}) + \frac{1}{\lambda}\sum_{i=1}^{k}(\tau_i^\star - \tau_i)P_{\mathbf{T}^\perp}(Z_i).$$

We can upper bound its spectral norm as follows:

$$\lambda \|W\| \le \underbrace{\left\|(I - UU^\top)X^\star Y^{\star\top}(I - VV^\top)\right\|}_{A_1} + \underbrace{\left\|\hat{E}\right\|}_{A_2} + \underbrace{\|\tau^\star - \tau\| \sum_{i=1}^{k} \|P_{\mathbf{T}^\perp}(Z_i)\|}_{A_3}.$$

Now, we bound each of these terms separately.

Bounding A_1. By (17), we have $(I - UU^\top)X = 0$ and $(I - VV^\top)Y = 0$. Hence,

$$\begin{aligned}\left\|(I - UU^\top)X^\star Y^{\star\top}(I - VV^\top)\right\| &= \left\|(I - UU^\top)(XH_{X,Y} - X^\star)(YH_{X,Y} - Y^{\star\top})(I - VV^\top)\right\| \\ &\le \left\|XH_{X,Y} - X^\star\right\|_{\mathrm{F}} \left\|YH_{X,Y} - Y^\star\right\|_{\mathrm{F}} \\ &\lesssim \rho^2 \\ &\overset{(i)}{\lesssim} \frac{\sigma^2 r^2 \kappa n \log^{12}(n)}{\sigma_{\min}} \\ &\lesssim \sigma\sqrt{n},\end{aligned}$$

where (i) follows from the bound on ρ from (3), and the last step follows from the bound on $\frac{\sigma\sqrt{n}}{\sigma_{\min}}$.

Bounding A_2. A_2 is bounded by Assumption 3(a) which gives $\|\hat{E}\| \le \|E\| + \|\delta\| \lesssim \sigma\sqrt{n}$.

Bounding A_3. If $(XY^\top, \tau, m)$ satisfy (18a)–(18d) and (18f), then the following decomposition holds due to the same proof as in Lemma 1.

$$\tilde{D}\left(\tau - \tau^\star\right) = \lambda\tilde{\Delta}^1 + \tilde{\Delta}^2 + \tilde{\Delta}^3,$$

where $\tilde{D} \in \mathbb{R}^{k\times k}$ is the matrix with entries $\tilde{D}_{ij} = \langle P_{\mathbf{T}^\perp}(Z_i), P_{\mathbf{T}^\perp}(Z_j)\rangle$ and $\tilde{\Delta}^1, \tilde{\Delta}^2, \tilde{\Delta}^3 \in \mathbb{R}^k$ are vectors with components

$$\tilde{\Delta}^1_i = \left\langle Z_i, UV^\top\right\rangle, \quad \tilde{\Delta}^2_i = \left\langle Z_i, P_{\mathbf{T}^\perp}(\hat{E})\right\rangle, \quad \tilde{\Delta}^3_i = \left\langle Z_i, P_{\mathbf{T}^\perp}(M^\star)\right\rangle.$$

This leads us to have:

$$\begin{aligned}A_3 &= \left\|\tilde{D}^{-1}(\lambda\tilde{\Delta}^1 + \tilde{\Delta}^2 + \tilde{\Delta}^3)\right\| \sum_{i=1}^k \|P_{\mathbf{T}^\perp}(Z_i)\| \\ &\le \lambda\left\|\tilde{D}^{-1}\tilde{\Delta}^1\right\| \sum_{i=1}^k \|P_{\mathbf{T}^\perp}(Z_i)\| + \left(\left\|\tilde{D}^{-1}\tilde{\Delta}^2\right\| + \left\|\tilde{D}^{-1}\tilde{\Delta}^3\right\|\right) \sum_{i=1}^k \|P_{\mathbf{T}^\perp}(Z_i)\| \\ &\le \lambda\left(1 - \frac{c_{r_2}}{2\log n}\right) + \frac{2\log^2 n}{c_s}\left(\left\|\tilde{\Delta}^2\right\| + \left\|\tilde{\Delta}^3\right\|\right),\end{aligned}$$

where the last inequality made use of Lemma 2.6.

Claim 2.14 *We have*

$$\left\|\tilde{\Delta}^2\right\|, \left\|\tilde{\Delta}^3\right\| \lesssim \sigma\sqrt{nr}\log n.$$

With the above claim, we are ready to bound $\|W\|$.

$$\begin{aligned}\|W\| &\le \frac{1}{\lambda}\left(A_1 + A_2 + A_3\right) \\ &\le \frac{1}{\lambda}\left(A_1 + A_2 + \lambda\left(1 - \frac{c_{r_2}}{2\log n}\right) + \frac{2\log^2 n}{c_s}\left(\left\|\tilde{\Delta}^2\right\| + \left\|\tilde{\Delta}^3\right\|\right)\right).\end{aligned}$$

Hence, we conclude that

$$\|W\| \lesssim 1 - \frac{c_{r_2}}{2\log n} + O\left(\frac{\sigma\sqrt{nr}\log^3(n)}{\lambda}\right) \leq 1$$

for large enough n and $\lambda = \Theta\left(\sigma\sqrt{nr}\log^{4.5}(n)\right)$.

Proof of Claim 2.14.. *Bounding* $\left\|\tilde{\Delta}^2\right\|$. Using Lemma 2.9, we have

$$\left\|\tilde{\Delta}^2\right\| \leq k \max_{i\in[k]} \left|\left\langle P_{\mathbf{T}^\perp}(Z_i), \hat{E}\right\rangle\right| \lesssim \sigma\sqrt{nr}\log n.$$

Bounding $\left\|\tilde{\Delta}^3\right\|$. $\left\|\tilde{\Delta}^3\right\|$ can be bounded by following the exact same steps as the bound of $\|\Delta^3\|$ (5), replacing $\hat{\mathbf{T}}$, $\hat{U}$, and $\hat{V}$ with $\mathbf{T}$, U, and V, respectively. In particular, note that Lemma 2.4 gives that $V^\top \mathbf{1} = 0$ along the gradient flow that ends at X, Y, which allows the same simplifications.

$$\left\|\tilde{\Delta}^3\right\| \lesssim \frac{\sigma^2 r^2 \kappa n \log^{12.5}(n)}{\sigma_{\min}} \lesssim \sigma\sqrt{n},$$

where the last step used the assumed bound on $\frac{\sigma\sqrt{n}}{\sigma_{\min}}$. □

2.5.2 Function g has a unique minimizer

In this subsection, we aim to show that the convex function $g(M, \tau, m)$ has a unique minimizer. Throughout the proof, we fix a global minimum $(\hat{M}, \hat{\tau}, \hat{m})$ of g, such that $\hat{M} = XY^\top$, where (X, Y) is the limit of the gradient flow of f starting from $(X^\star, Y^\star)$. We have already showed in Section 2.5.1 that $(XY^\top, \tau, m)$, satisfies all of the first order conditions of g; hence $\hat{M} = XY^\top$ is indeed a minimizer of g. First note that up to a bijective change in variables, minimizing g is equivalent to minimizing the following function

$$\tilde{g}(N, \tau, m) = \frac{1}{2}\|O - N\|_{\mathrm{F}}^2 + \lambda \left\| N - \sum_i \tau_i Z_i - m\mathbf{1}^\top \right\|_\star.$$

Hence, it suffices to show that $\tilde{g}$ has a unique minimizer. This function is strictly convex in N, because of the term $\|O - N\|_{\mathrm{F}}^2$. We can then fix $\hat{N}$ to be the unique value of N that minimizes $\tilde{g}$. In particular, note that $\hat{N} = \hat{M} + \sum_i \hat{\tau}_i Z_i + \hat{m}\mathbf{1}^\top$. It only remains to show that the following convex optimization problem

$$\min_{\tau,m} \left\| \hat{N} - \sum_i \tau_i Z_i - m\mathbf{1}^\top \right\|_\star = \min_{\tau,m} \left\| \hat{M} - \sum_i \tau_i Z_i - m\mathbf{1}^\top \right\|_\star \tag{19}$$

has a unique minimizer. By the definition of $\hat{M} = XY^\top$, a minimum of the right-hand side of (19) is attained for $\tau = 0$ and $m = 0$ (otherwise, $\hat{M}$ wouldn't be an optimum of g). Now consider any other optimal solution to the problem in (19), that is $Z \in \mathbf{Z}$ such that $\|XY^\top\|_\star = \|XY^\top + Z\|_\star$. Write the SVD $XY\top = U\Sigma V^\top$. Recall that the subgradients of the nuclear norm are $\left\{UV^\top + W : U^\top W = 0, WV = 0, \|W\| \leq 1\right\}$. By the convexity of the nuclear norm, we have

$$\|XY^\top + Z\|_\star - \|XY^\top\|_\star \geq \left\langle Z, UV^\top + W\right\rangle, \tag{20}$$

for any matrix W with $U^\top W = 0$, $WV = 0$ and $\|W\| \leq 1$. Recall that $\mathbf{T}_0$ is defined to be the tangent space of $XY^\top$. Defining the SVD $P_{\mathbf{T}_0^\perp}(Z) = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$, we can take $W = \tilde{U}\tilde{V}^\top$, which gives

$$\|XY^\top + Z\|_\star - \|XY^\top\|_\star \geq \left\langle Z, UV^\top \right\rangle + \|P_{\mathbf{T}_0^\perp}(Z)\|_\star. \tag{21}$$

We now use the following lemma.

Claim 2.15 *Under the same assumptions as in Lemma 2.6, we have*

$$\|P_{\mathbf{T}_0^\perp}(Z)\|_\star > |\left\langle Z, UV^\top \right\rangle| \qquad \forall Z \in \mathbf{Z} \setminus \{0\}.$$

With this result, we have that if $Z \neq 0$, then $\|XY^\top + Z\|_\star > \|XY^\top\|_\star$, which contradicts the definition of Z. Hence, $Z = 0$ which ends the proof that g has a unique minimizer.

Proof of Claim 2.15. Up to changing Z into $-Z$, it suffices to show that for all non-zero $Z \in \mathbf{Z}$, we have $\|P_{\mathbf{T}_0^\perp}(Z)\|_\star > \left\langle Z, UV^\top \right\rangle$. First, using similar arguments as in (21) we show that for any matrices A and B such that $\langle A, B\rangle = 0$ with SVD $A = U_A \Sigma_A V_A^\top$, we have

$$\|A + B\|_\star - \|A\|_\star \geq \left\langle B, U_A V_A^\top \right\rangle = 0.$$

Hence, $\|A + B\|_\star \geq \|A\|_\star$. In particular, we can take $A = P_{\mathbf{T}^\perp}(Z)$ and $B = P_{\mathbf{T}_0^\perp}(Z) - P_{\mathbf{T}^\perp}(Z)$ since they are orthogonal to each other due to the fact that $\mathbf{T}^\perp \subset \mathbf{T}_0^\perp$. Then, we have

$$\|P_{\mathbf{T}_0^\perp}(Z)\|_\star \geq \|P_{\mathbf{T}^\perp}(Z)\|_\star.$$

Next, by Lemma 2.4 we have $V^\top \mathbf{1} = 0$, so that $\left\langle Z, UV^\top \right\rangle = \left\langle Z, P_{\mathbf{1}^\perp}(UV^\top) \right\rangle = \left\langle P_{\mathbf{1}^\perp}(Z), P_{\mathbf{1}^\perp}(UV^\top) \right\rangle$. The two previous steps essentially show that we can ignore the terms of the form $\alpha \mathbf{1}^\top$ within Z. Formally, it suffices to show that for any non-zero $Z \in span(Z_i, i \in [k])$, we have $\|P_{\mathbf{T}^\perp}(Z)\|_\star > \left\langle Z, P_{\mathbf{1}^\perp}(UV^\top) \right\rangle$. We decompose such a matrix as $Z = \sum_{i\in[k]} \alpha_i Z_i$. First, by (9) of Lemma 2.6, we have that $\|P_{\mathbf{T}^\perp}(Z)\|_{\mathrm{F}}^2 \geq \frac{c_{r_1}}{2\log(n)}\|Z\|_F^2 > 0$, which implies $\|P_{\mathbf{T}^\perp}(Z)\| > 0$. Then, with $\alpha := (\alpha_1, \ldots, \alpha_k) \neq 0$, we have

$$\begin{aligned}
\|P_{\mathbf{T}^\perp}(Z)\|_\star - \left\langle Z, P_{\mathbf{1}^\perp}(UV^\top) \right\rangle &\overset{(i)}{\geq} \frac{\|P_{\mathbf{T}^\perp}(Z)\|_{\mathrm{F}}^2}{\|P_{\mathbf{T}^\perp}(Z)\|} - \left\langle Z, P_{\mathbf{1}^\perp}(UV^\top) \right\rangle \\
&\overset{(ii)}{=} \frac{1}{\|P_{\mathbf{T}^\perp}(Z)\|} \left(\alpha^\top \tilde{D}\alpha - \alpha^\top \tilde{\Delta}^1 \cdot \|P_{\mathbf{T}^\perp}(Z)\| \right) \\
&= \frac{1}{\|P_{\mathbf{T}^\perp}(Z)\|} \left(\alpha^\top \tilde{D}\alpha - \alpha^\top \tilde{\Delta}^1 \cdot \left\| \sum_{i\in[k]} \alpha_i P_{\mathbf{T}^\perp}(Z_i) \right\| \right) \\
&\geq \frac{1}{\|P_{\mathbf{T}^\perp}(Z)\|} \underbrace{\left(\alpha^\top \tilde{D}\alpha - \alpha^\top \tilde{\Delta}^1 \cdot \|\alpha\| \sum_{i\in[k]} \|P_{\mathbf{T}^\perp}(Z_i)\| \right)}_{a},
\end{aligned}$$

where (i) is due to the fact that $\|P_{\mathbf{T}^\perp}(Z)\| > 0$ and the fact that the Frobenius norm of a matrix, squared, is the sum of the squares of the singular values of that matrix; (ii) uses the identity $\|P_{\mathbf{T}^\perp}(Z)\|_{\mathrm{F}}^2 = \left\langle \sum_{i\in[k]} \alpha_i P_{\mathbf{T}^\perp}(Z_i), \sum_{j\in[k]} \alpha_j P_{\mathbf{T}^\perp}(Z_j) \right\rangle = \alpha^\top \tilde{D}\alpha$.

By (12) of Lemma 2.6, the matrix $\tilde{D}$ is invertible. It is also symmetric by construction. We next define $\beta = \tilde{D}^{1/2}\alpha$. We obtain

$$\begin{aligned} a &= \|\beta\|^2 - \beta^\top \tilde{D}^{-1/2}\tilde{\Delta}^1 \cdot \|\tilde{D}^{-1/2}\beta\| \sum_{i\in[k]} \|P_{\mathbf{T}^\perp}(Z_i)\| \\ &\geq \|\beta\|^2 \left(1 - \|\tilde{D}^{-1/2}\| \cdot \|\tilde{\Delta}^1\| \cdot \|\tilde{D}^{-1/2}\| \sum_{i\in[k]} \|P_{\mathbf{T}^\perp}(Z_i)\|\right) \\ &\overset{(i)}{>} 0, \end{aligned}$$

where in (i) we used (11) of Lemma 2.6 together with the fact that $\beta \neq 0$. Combining the two previous inequalities shows that

$$\|P_{\mathbf{T}^\perp}(Z)\|_\star > \left\langle Z, P_{\mathbf{1}^\perp}(UV^\top)\right\rangle, \quad \forall Z \in span(Z_i, i \in [k]).$$

This ends the proof of the claim. □

2.6 Proof of Lemma 1.1

Lemma 1.1 *Let D and Δ^2 be defined as in Lemma 1. We have $\sigma_{\min}(D) \geq \frac{c_s}{2\log n}$. Furthermore, we have the following for sufficiently large n:*

$$\left\|\Delta^2\right\| \lesssim \frac{\sigma^2 r^{1.5}\kappa n \log^{6.5}(n)}{\sigma_{\min}} + \log^{0.5}(n) \cdot \max_{i\in[k]} \frac{\left|\left\langle P_{\mathbf{T}^{\star\perp}}(\tilde{Z}_i), P_{\mathbf{T}^{\star\perp}}(E+\delta)\right\rangle\right|}{\|\tilde{Z}_i\|_{\mathrm{F}}}.$$

Proof. By Lemma 1.2, for sufficiently large n, we can write $\hat{M} = XY^\top$ where (X, Y) is the limit of the gradient flow of f started at $(X^\star, Y^\star)$, and $(X, Y) \in \mathcal{B}$. Hence, the conditions for applying both Lemma 2.6 and Lemma 2.9 with $\hat{\mathbf{T}}$, the tangent space of $\hat{M}$, are satisfied. By (12) from Lemma 2.6, we have the desired bound on $\sigma_{\min}(D)$. Furthermore, we have

$$\left\|\Delta^2\right\| \leq \sqrt{k}\left\|\Delta^2\right\|_\infty \lesssim \sqrt{\log n}\left\|\Delta^2\right\|_\infty.$$

We aim to bound $\left\|\Delta^2\right\|_\infty = \max_{i\in[k]}\left|\left\langle P_{\hat{\mathbf{T}}^\perp}(Z_i), \hat{E}\right\rangle\right|$, where $\hat{E} = E + \delta$. We have

$$\begin{aligned} \left|\left\langle P_{\hat{\mathbf{T}}^\perp}(Z_i), \hat{E}\right\rangle\right| &\leq \left|\left\langle P_{\hat{\mathbf{T}}^\perp}(Z_i) - P_{\mathbf{T}^{\star\perp}}(Z_i), \hat{E}\right\rangle\right| + \left|\left\langle P_{\mathbf{T}^{\star\perp}}(Z_i), \hat{E}\right\rangle\right| \\ &\overset{(i)}{\leq} \left\|P_{\hat{\mathbf{T}}^\perp}(Z_i) - P_{\mathbf{T}^{\star\perp}}(Z_i)\right\|_\star (\|E\| + \|\delta\|) + \left|\left\langle P_{\mathbf{T}^{\star\perp}}(Z_i), \hat{E}\right\rangle\right| \\ &\overset{(ii)}{\lesssim} \frac{\rho\sqrt{\kappa r}}{\sqrt{\sigma_{\min}}}\sigma\sqrt{n} + \left|\left\langle P_{\mathbf{T}^{\star\perp}}(Z_i), \hat{E}\right\rangle\right| \\ &\overset{(iii)}{\lesssim} \frac{\sigma^2 r^{1.5}\kappa n \log^6(n)}{\sigma_{\min}} + \left|\left\langle P_{\mathbf{T}^{\star\perp}}(Z_i), P_{\mathbf{T}^{\star\perp}}(\hat{E})\right\rangle\right| \end{aligned}$$

where (i) is due to Von Neumann's trace inequality, (ii) is due to (10) and Assumption 3(a), and (iii) is due to (3). Hence,

$$\begin{aligned}\left\|\Delta^2\right\| &\lesssim \frac{\sigma^2 r^{1.5}\kappa n \log^{6.5}(n)}{\sigma_{\min}} + \log^{0.5}(n)\cdot \max_{i\in[k]} \left|\left\langle P_{\mathbf{T}^{\star\perp}}(Z_i), P_{\mathbf{T}^{\star\perp}}(\hat{E})\right\rangle\right| \\ &= \frac{\sigma^2 r^{1.5}\kappa n \log^{6.5}(n)}{\sigma_{\min}} + \log^{0.5}(n)\cdot \max_{i\in[k]} \frac{\left|\left\langle P_{\mathbf{T}^{\star\perp}}(\tilde{Z}_i), P_{\mathbf{T}^{\star\perp}}(E+\delta)\right\rangle\right|}{\|\tilde{Z}_i\|_{\mathrm{F}}}.\end{aligned}$$

□

3 Proofs of Auxiliary Lemmas

Proof of Lemma EC.1. We use ideas from the proof of the main theorem in Devroye (1977) and their notation. Because the i.i.d. samples $X_1,\ldots,X_n$ are drawn from a distribution with a density function, on an almost sure event $\mathcal{E} = \{X_{i_1 j} \neq X_{i_2 j}, i_1 \neq i_2 \in [n], j \in [d]\}$, they do not share any component in common. For $i \in [n]$, let $X_i = (X_{i1},\ldots,X_{id})$, and let $\mathcal{Y}$ denote the set of random vectors $Y_1,\ldots,Y_{n^d} \in \mathbb{R}^d$ that are obtained by considering all n^d vectors of the form $(X_{i_1 1},\ldots,X_{i_d d})$, where $(i_1,\ldots,i_d) \in \{1,\ldots,n\}^d$.

Because G_n is a staircase function with flat levels everywhere except at points in $\mathcal{Y}$, and G is monotonic, $|G_n(x_1,x_2) - G(x_1,x_2)|$ is maximized when x_1 and x_2 approach vectors that are in $\mathcal{Y}$. Under event $\mathcal{E}$, we have

$$\sup_{x_1,x_2\in\mathbb{R}^d} |G_n(x_1,x_2) - G(x_1,x_2)| \leq \sup_{i,j} |G_n(Y_i,Y_j) - G(Y_i,Y_j)| + \frac{2d}{n},$$

where the $\frac{2d}{n}$ is due to the fact that, for any i,j there may be up to $2d$ different points in $X_1,\ldots,X_n$ that share a component with Y_i or Y_j under the event $\mathcal{E}$. These $2d$ points lie on the perimeter of the hyper-rectangle between Y_i and Y_j and could be included or excluded from the count for G_n depending on whether x_1 approaches Y_i (and x_2 approaches Y_j) from the inside or outside of the hyper-rectangle.

Now, it remains to upper bound the right hand side of the above inequality. For any pair of indices $i,j \leq n^d$, there exists a subset of indices $\mathcal{I} \subseteq [n]$, corresponding to samples of X that have at least one component in common with either Y_i or Y_j. Hence, the samples indexed by $\mathcal{I}$ are not independent from both Y_i and Y_j. In order to apply Hoeffding's inequality to the samples, we sample $d' := |\mathcal{I}|$ additional i.i.d. random vectors $X_{n+1},\ldots,X_{n+d'}$ from the same distribution. These serve to 'substitute' those samples X_l for which $l \in \mathcal{I}$. Under $\mathcal{E}$, the number of indices in $\mathcal{I}$ satisfies $d' \leq 2d$, and we have

$$\begin{aligned}\sup_{i,j} |G_n(Y_i,Y_j) - G(Y_i,Y_j)| + \frac{2d}{n} &= \sup_{i,j} \left|\frac{1}{n}\sum_{k=1}^{n} I_{\{X_k\in[Y_i,Y_j]\}} - G(Y_i,Y_j)\right| + \frac{2d}{n} \\ &\leq \sup_{i,j} \left|\frac{1}{n}\sum_{k\in[n+d']\setminus\mathcal{I}} I_{\{X_k\in[Y_i,Y_j]\}} - G(Y_i,Y_j)\right| + \frac{1}{n}\left|\sum_{k\in\mathcal{I}} I_{X_k\in[Y_i,Y_j]} - \sum_{k=n+1}^{n+d'} I_{X_k\in[Y_i,Y_j]}\right| + \frac{2d}{n} \\ &\leq \sup_{i,j} \left|\frac{1}{n}\sum_{k\in[n+d']\setminus\mathcal{I}} I_{\{X_k\in[Y_i,Y_j]\}} - G(Y_i,Y_j)\right| + \frac{4d}{n}.\end{aligned}$$

The first inequality is due to the triangle inequality, and the second inequality follows because $|\mathcal{I}| = d' \leq 2d$.

By the independence of both Y_i and Y_j with X_k for $k \in [n + d'] \setminus \mathcal{I}$, we apply Hoeffding's inequality conditioned on the realizations of Y_i and Y_j to establish the following upper bound:

$$\Pr\left[\left|\frac{1}{n}\sum_{k\in[n+d']\setminus\mathcal{I}} I_{\{X_k\in[Y_i,Y_j]\}} - G(Y_i,Y_j)\right| + \frac{4d}{n} \geq \epsilon\right] \leq 2\exp\left(-2n\left(\epsilon - \frac{4d}{n}\right)^2\right).$$

Applying a Union Bound over all i, j, we obtain

$$\begin{aligned}\Pr\left[\sup_{x_1,x_2\in\mathbb{R}^d} |G_n(x_1,x_2) - G(x_1,x_2)| \geq \epsilon\right] &\leq 2n^{2d}\exp\left(-2n\left(\epsilon - \frac{4d}{n}\right)^2\right) \\ &\leq 2n^{2d}\exp\left(-2n\epsilon^2 + 16\epsilon d\right).\end{aligned}$$

This proves our first bound. To show the second bound, we note that the first line of the above inequality can be rewritten as

$$\Pr\left[\sup_{x_1,x_2\in\mathbb{R}^d} |G_n(x_1,x_2) - G(x_1,x_2)| \geq \frac{4d}{n} + \epsilon\right] \leq 2n^{2d}e^{-2n\epsilon^2}.$$

□

References

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452.

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. (2020). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121.

Devroye, L. P. (1977). A uniform bound for the deviation of empirical distribution functions. *Journal of Multivariate Analysis*, 7(4):594–597.

Farias, V., Li, A., and Peng, T. (2021). Learning treatment effects in panels with general intervention patterns. *Advances in Neural Information Processing Systems*, 34:14001–14013.

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

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323.