

757 A Additional Illustrations

758 In this section, we provide some useful illustrations. Figure 4 illustrates the corruption model
 759 described in Assumption 1. Figure 5 illustrates the linear maps $\mathbf{G}, \tilde{\mathbf{G}}$ used to generate the embeddings
 from observed data (according to the model in Fig 4). Figure 6 accompanies Remark A.1.

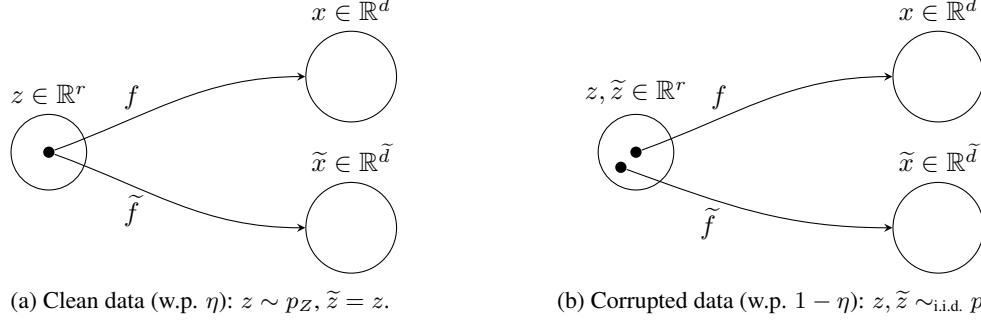


Figure 4: Model for stochastic corruptions. In this work, the forward maps f, \tilde{f} are linear (refer to Eq. (1)) and the latent distributions are Gaussians.

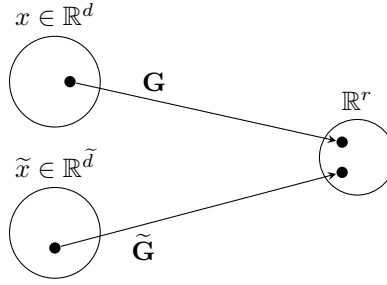


Figure 5: On seeing multimodal data (x, \tilde{x}) , linear maps $\mathbf{G}, \tilde{\mathbf{G}}$ (learnable parameters) create the embeddings that lie in \mathbb{R}^r (the knowledge of r , the true latent dimension, is assumed). The similarity is measured with the inner product $\langle \mathbf{G}x, \tilde{\mathbf{G}}\tilde{x} \rangle$.

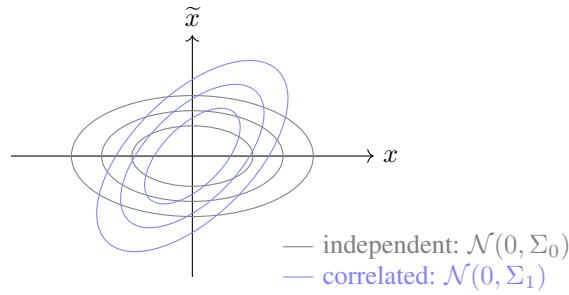


Figure 6: Illustration of the joint distribution of (x, \tilde{x}) . The overall distribution is a mixture of two zero mean Gaussians: the independent case (w.p. $1 - \eta$) and the correlated case (w.p. η).

760

761 **Remark A.1.** The distribution of $(x, \tilde{x}) \in \mathbb{R}^{d+\tilde{d}}$ from Section 3.1 is a mixture of two zero-mean
 762 Gaussians. With weight η , the covariance matrix is Σ_1 (for $c = 1$, i.e. the clean case). With weight
 763 $1 - \eta$, the covariance is Σ_0 (for $c = 0$). Figure 6 provides an illustration.

$$\Sigma_1 = \begin{bmatrix} \mathbf{U}\mathbf{U}^\top + \gamma^{-1}\mathbf{I}_d & \mathbf{U}\tilde{\mathbf{U}}^\top \\ \tilde{\mathbf{U}}\mathbf{U}^\top & \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top + \tilde{\gamma}^{-1}\mathbf{I}_{\tilde{d}} \end{bmatrix}, \quad \Sigma_0 = \begin{bmatrix} \mathbf{U}\mathbf{U}^\top + \gamma^{-1}\mathbf{I}_d & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top + \tilde{\gamma}^{-1}\mathbf{I}_{\tilde{d}} \end{bmatrix}.$$

764 B Background

765 This section covers some useful background concepts.

766 B.1 Measuring the distance between subspaces

767 The concept of principal angles provides a geometrically intuitive way to measure the closeness
768 between two subspaces. Let \mathcal{X} and \mathcal{Y} be two r -dimensional subspaces within a larger Euclidean
769 space \mathbb{R}^d . There exist r principal angles $0 \leq \theta_1 \leq \theta_2 \leq \dots \leq \theta_r \leq \pi/2$ that describe the relative
770 orientation of these subspaces.

- 771 • θ_1 represents the *smallest* possible angle between any two unit vectors $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.
- 772 • Subsequent angles θ_k capture the minimum angles within directions orthogonal to those
773 defining the previous angles $\theta_1, \dots, \theta_{k-1}$.
- 774 • The cosines $\cos(\theta_i)$ measure the alignment (1 means aligned, 0 means orthogonal within
775 that principal direction), while the sines $\sin(\theta_i)$ measure the separation or angle.

To aggregate this information into a single distance metric, we often use the frobenius norm of the
sine of the principal angles, denoted $\|\sin \Theta(\mathcal{X}, \mathcal{Y})\|_F$. It is defined as

$$\|\sin \Theta(\mathcal{X}, \mathcal{Y})\|_F = \sqrt{\sum_{i=1}^r \sin^2(\theta_i)} .$$

776 This metric provides an overall measure of the difference between the subspaces. It's zero if and only
777 if $\mathcal{X} = \mathcal{Y}$ (since all $\theta_i = 0$), and it increases as the subspaces diverge.

Computing this metric relies on matrix operations involving orthonormal bases for the subspaces. Let
 $\mathbf{X} \in \mathbb{R}^{d \times r}$ be a matrix whose columns form an orthonormal basis for \mathcal{X} (so $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_r$). Similarly,
let $\mathbf{Y} \in \mathbb{R}^{d \times r}$ be a matrix with orthonormal columns forming a basis for \mathcal{Y} . The distance metric
 $\|\sin \Theta(\mathcal{X}, \mathcal{Y})\|_F$ can be computed using \mathbf{X} and \mathbf{Y} via the following formula

$$\|\sin \Theta(\mathcal{X}, \mathcal{Y})\|_F = \|\mathbf{X}_\perp^\top \mathbf{Y}\|_F .$$

778 Here, \mathbf{X}_\perp is any $d \times (d-r)$ matrix such that its columns form an orthonormal basis for the orthogonal
779 complement of \mathcal{X} , denoted \mathcal{X}^\perp . This means that the combined matrix $[\mathbf{X} \ \mathbf{X}_\perp]$ must be a $d \times d$
780 orthogonal matrix. Notationally, we often just write $\|\sin \Theta(\mathbf{X}, \mathbf{Y})\|_F$ instead of using \mathcal{X}, \mathcal{Y} .

781 C Lemmas

782 This section presents Lemmas used in the proofs. The first three Lemmas are standard results in the
783 literature, and we include them without proof.

784 **Lemma 1** (Weyl's Inequality). *For matrices $A, B \in \mathbb{R}^{m \times n}$, let $p = \min(m, n)$ and let $\sigma_1(M) \geq$
785 $\sigma_2(M) \geq \dots \geq \sigma_p(M) \geq 0$ denote the singular values for $M \in \{A, B\}$. Then, for all $j = 1, \dots, p$,
786 it holds that*

$$|\sigma_j(A) - \sigma_j(B)| \leq \|A - B\|_2 .$$

787 **Lemma 2** (Wedin's Theorem). *Let $A, \hat{A} \in \mathbb{R}^{m \times n}$ be matrices of the same size. Let $r \leq \min(m, n)$
788 be the rank of both A, \hat{A} , and let the SVDs be $A = U \Sigma V^\top$ and $\hat{A} = \hat{U} \hat{\Sigma} \hat{V}^\top$. Let $\sigma_r(A) > 0$ denote
789 the r^{th} singular value of A , and assume $\sigma_r(A) > \|\hat{A} - A\|_2$. Then it holds that:*

$$\begin{aligned} \left\| \sin \Theta(\hat{U}, U) \right\|_F &\leq \frac{\|\hat{A} - A\|_F}{\sigma_r(A) - \|\hat{A} - A\|_2} , \\ \left\| \sin \Theta(\hat{V}, V) \right\|_F &\leq \frac{\|\hat{A} - A\|_F}{\sigma_r(A) - \|\hat{A} - A\|_2} . \end{aligned}$$

Lemma 3 (Whittle's Inequality). *Let X_1, X_2, \dots be a sequence of independent random variables
such that: (i) $\mathbb{E}[X_k] = 0$ for all $k \geq 1$, and (ii) the distribution of each X_k is symmetric about zero
(i.e., X_k and $-X_k$ have the same distribution). Let $S_n = \sum_{k=1}^n X_k$ be the partial sum (with $S_0 = 0$).*

If $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is a convex function such that $\phi(0) = 0$, then the sequence $\mathbb{E}[\phi(S_n)]$ is non-decreasing in n . That is, for all $n \geq 1$:

$$\mathbb{E}[\phi(S_n)] \geq \mathbb{E}[\phi(S_{n-1})] .$$

Lemma 4. Let X be a random variable with a log-concave density, mean μ_X , and variance σ_X^2 . It holds that

$$\mathbb{E}[X \mid X > \theta] \leq \theta + e \sigma_X , \quad \text{for } \theta \geq \mu_X .$$

Proof. Let $m(x) = \mathbb{E}[X - x \mid X > x]$ be the mean residual life function. We want to bound $\mathbb{E}[X \mid X > \theta] = \theta + m(\theta)$ for $\theta \geq \mu_X$. Due to log-concavity of X , $m(x)$ is non-increasing (see, eg, Bagnoli and Bergstrom [2], Theorem 6). Since $m(x)$ is non-increasing, $m(\theta) \leq m(\mu_X) = \mathbb{E}[X - \mu_X \mid X > \mu_X]$. We will now bound the conditional expectation for this case of $\theta = \mu_X$.

Let $Y = X - \mu_X$. Then $\mathbb{E}[Y] = 0$ and $\mathbb{V}(Y) = \sigma_X^2$. $m(\mu_X) = \mathbb{E}[Y \mid Y > 0] = \frac{\mathbb{E}[Y^+]}{\mathbb{P}(Y > 0)}$, where $Y^+ = \max(0, Y)$. We know $\mathbb{E}[Y^+] \leq \sqrt{\mathbb{E}[(Y^+)^2]} \leq \sqrt{\mathbb{E}[Y^2]} = \sigma_X$. As for the denominator, we know that for any random variable X with a log-concave density and mean μ_X , $\mathbb{P}(X \geq \mu_X) \geq 1/e$ (see, eg, Lovász and Vempala [20], Lemma 5.4). Thus, $m(\mu_X) \leq \frac{\sigma_X}{1/e} = e \sigma_X$. \square

Lemma 5. Let $x, y \in \mathbb{R}^d$ and $\tilde{x}, \tilde{y} \in \mathbb{R}^{\tilde{d}}$ be random vectors. Assume that the pair (x, \tilde{x}) is independent of the pair (y, \tilde{y}) . Let \mathbf{A} be a fixed $d \times \tilde{d}$ matrix and let $\theta \in \mathbb{R}$ be a scalar threshold. Define the events $C_x = \{x^\top \mathbf{A} \tilde{x} > \theta\}$ and $C_y = \{y^\top \mathbf{A} \tilde{y} > \theta\}$. Assume that these events have non-zero probability, i.e., $\mathbb{P}(C_x) > 0$ and $\mathbb{P}(C_y) > 0$. Then the conditional expectation of the outer product $x\tilde{y}^\top$ given both events C_x and C_y factorizes as follows:

$$\mathbb{E} [x\tilde{y}^\top \mid x^\top \mathbf{A} \tilde{x} > \theta, y^\top \mathbf{A} \tilde{y} > \theta] = \mathbb{E} [x \mid x^\top \mathbf{A} \tilde{x} > \theta] \cdot \left(\mathbb{E} [\tilde{y} \mid y^\top \mathbf{A} \tilde{y} > \theta] \right)^\top .$$

Proof. The definition of conditional expectation given multiple events is conditioning on their intersection. Here \mathbb{I} denotes the indicator function.

$$\mathbb{E} [x\tilde{y}^\top \mid C_x, C_y] = \mathbb{E} [x\tilde{y}^\top \mathbb{I}_{C_x \cap C_y}] = \frac{\mathbb{E} [x\tilde{y}^\top \mathbb{I}_{C_x \cap C_y}]}{\mathbb{P}(C_x \cap C_y)} .$$

The event C_x is determined solely by the random variables x and \tilde{x} . The event C_y is determined solely by the random variables y and \tilde{y} . By the initial assumption, the pair (x, \tilde{x}) is independent of the pair (y, \tilde{y}) . Therefore, the event C_x is independent of the event C_y . This implies $\mathbb{P}(C_x \cap C_y) = \mathbb{P}(C_x) \mathbb{P}(C_y)$. Hence the denominator factorizes (and is non-zero since $\mathbb{P}(C_x) > 0$ and $\mathbb{P}(C_y) > 0$).

Now consider the numerator. Since C_x and C_y are independent, $\mathbb{I}_{C_x \cap C_y} = \mathbb{I}_{C_x} \mathbb{I}_{C_y}$, which implies

$$\mathbb{E} [x\tilde{y}^\top \mathbb{I}_{C_x \cap C_y}] = \mathbb{E} [x\tilde{y}^\top \mathbb{I}_{C_x} \mathbb{I}_{C_y}] = \mathbb{E} [x \mathbb{I}_{C_x}] \cdot \mathbb{E} [\tilde{y} \mathbb{I}_{C_y}]^\top ,$$

again, due to independence of the pairs. Hence the numerator also factorizes. \square

Lemma 6. Let $x \in \mathbb{R}^d$ and $\tilde{x} \in \mathbb{R}^{\tilde{d}}$ be random vectors such that their joint distribution is a multivariate normal distribution with zero mean. Let \mathbf{A} be a fixed $d \times \tilde{d}$ matrix, and consider the conditioning event $\mathcal{R} = \{(x, \tilde{x}) \mid x^\top \mathbf{A} \tilde{x} > \theta\}$ for some threshold $\theta \in \mathbb{R}$. Assume that the probability of this event is non-zero, i.e., $\mathbb{P}(\mathcal{R}) > 0$. Then

$$\mathbb{E} [x \mid x^\top \mathbf{A} \tilde{x} > \theta] = \mathbf{0}_d .$$

Proof. Let $Z = (x, \tilde{x}) \in \mathbb{R}^{d+\tilde{d}}$. The joint probability density function of Z , denoted by $p(Z)$, corresponds to the $\mathcal{N}(0, \Sigma_{\text{joint}})$ distribution for some covariance matrix Σ_{joint} . The conditional expectation is defined as:

$$\mathbb{E} [x \mid x^\top \mathbf{A} \tilde{x} > \theta] = \mathbb{E} [x \mid Z \in \mathcal{R}] = \frac{\int_{\mathcal{R}} x p(Z) dZ}{\int_{\mathcal{R}} p(Z) dZ} = \frac{\int_{\mathcal{R}} x p(Z) dZ}{P(\mathcal{R})}$$

We focus on the numerator integral and show that it is zero owing to symmetry. First note that $p(Z)$ is symmetric around the origin. That is, $p(Z) = p(-Z)$ for all $Z \in \mathbb{R}^{d+\tilde{d}}$. Second, observe that under the transformation $Z \mapsto -Z$, the condition becomes $(-u)^\top \mathbf{A} (-\tilde{u}) > \theta$, which simplifies to $u^\top \mathbf{A} \tilde{u} > \theta$. Thus, the region \mathcal{R} is symmetric with respect to the origin: $Z \in \mathcal{R} \iff -Z \in \mathcal{R}$. \square

814 **Lemma 7.** Consider the random variable $z := uv$, where u, v are jointly Gaussian as

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \sigma_u^2 & \gamma \\ \gamma & \sigma_v^2 \end{pmatrix}\right), \quad \text{with } 0 < \sigma_u^2, \sigma_v^2, \text{ and } 0 \leq \gamma < \sigma_u \sigma_v.$$

815 Let $\{z_k\}_{k=1}^r$ be r independent copies. The conditional expectation is upper and lower bounded as

$$\begin{aligned} \mathbb{E}\left[z_i \mid \sum_{k=1}^r z_k > \theta\right] &\geq \max\left\{\gamma, \frac{\theta}{r}\right\} \text{ for all } \theta \in \mathbb{R}, \\ \mathbb{E}\left[z_i \mid \sum_{k=1}^r z_k > \theta\right] &\leq \max\left\{\gamma, \frac{\theta}{r}\right\} + e \sqrt{\frac{\sigma_u^2 \sigma_v^2 + \gamma^2}{r}} \text{ for } \theta \geq 0. \end{aligned}$$

816 For the specific case of $\gamma = 0$ (i.e. u, v independent) and $\theta = 0$, a stronger lower bound is

$$\mathbb{E}\left[z_i \mid \sum_{k=1}^r z_k > 0\right] \geq \frac{2}{\pi r} \sigma_u \sigma_v.$$

817 **Proof. Simplify the expression.** Observe that z_k are i.i.d. random variables. The expectation is
 818 $\mathbb{E}[z_k] = \mathbb{E}[uv] = \gamma$ (since $\mathbb{E}[u] = 0 = \mathbb{E}[v]$). Let $S = \sum_{k=1}^r z_k$, and let $p_S(\cdot)$ denote the PDF of S .
 819 The expectation is $\mathbb{E}[S] = r\gamma$, and the variance is $\mathbb{V}[S] = r(\sigma_u^2 \sigma_v^2 + \gamma^2)$.

820 Due to the symmetry among the i.i.d. variables z_k , the conditional expectation $\mathbb{E}[z_i \mid S > \theta]$ is the
 821 same for all $i \in \{1, \dots, r\}$. Let $Q(\theta) = \mathbb{E}[z_i \mid S > \theta]$. By linearity of expectation, we have

$$\begin{aligned} \mathbb{E}[S \mid S > \theta] &= \mathbb{E}\left[\sum_{k=1}^r z_k \mid S > \theta\right] = \sum_{k=1}^r \mathbb{E}[z_k \mid S > \theta] = r Q(\theta). \\ \implies Q(\theta) &= \frac{1}{r} \mathbb{E}[S \mid S > \theta]. \end{aligned} \tag{11}$$

822 **Proof of lower bounds: general case lower bound $\frac{\theta}{r}$.** Observe that

$$\begin{aligned} \mathbb{E}[S \mid S > \theta] &= \frac{\int_{\theta}^{\infty} s p_S(s) ds}{\int_{\theta}^{\infty} p_S(s) ds} \\ &\geq \frac{\int_{\theta}^{\infty} \theta p_S(s) ds}{\int_{\theta}^{\infty} p_S(s) ds} = \theta. \end{aligned} \tag{12}$$

823 Combining this with Eq. (11) shows the θ/r lower bound.

824 **Proof of lower bounds: general case lower bound γ .** For this, we show $\mathbb{E}[S \mid S > \theta]$ is non-
 825 decreasing in θ . Let $h(\theta) = \mathbb{E}[S \mid S > \theta]$. Using Eq. (12), its derivative is given by

$$\begin{aligned} h'(\theta) &= \frac{-\theta p_S(\theta) \int_{\theta}^{\infty} p_S(s) ds + p_S(\theta) \int_{\theta}^{\infty} s p_S(s) ds}{\mathbb{P}(S > \theta)^2} \\ &= \frac{p_S(\theta)}{\mathbb{P}(S > \theta)^2} \int_{\theta}^{\infty} \underbrace{(s - \theta)}_{\geq 0} p_S(s) ds \geq 0. \end{aligned} \tag{13}$$

826 Thus $\mathbb{E}[S \mid S > \theta]$ is non-decreasing in θ . In particular, $\mathbb{E}[S \mid S > \theta] \geq \mathbb{E}[S]$ (i.e. the unconditional
 827 limit in the limit $\theta \rightarrow -\infty$). Since $\mathbb{E}[S] = r\gamma$, using this in Eq. (11) shows the lower bound of γ .

828 **Proof of lower bounds: the specific case of $\gamma = 0$ and $\theta = 0$.** Since the distribution of z_k is
 829 symmetric around zero, the distribution of $S = \sum_k z_k$ is also symmetric around zero. Therefore,
 830 $\mathbb{P}(S > 0) = 1/2$. Using this, we get

$$\mathbb{E}[S \mid S > 0] = \frac{\int_0^{\infty} s p_S(s) ds}{\mathbb{P}(S > 0)} = 2 \int_0^{\infty} s p_S(s) ds. \tag{14}$$

Also, the expectation of the absolute value is $\mathbb{E}[|S|] = \int_{-\infty}^{\infty} |s| p_S(s) ds$. Due to symmetry (i.e. $p_S(-s) = p_S(s)$), we get

$$\mathbb{E}[|S|] = \int_{-\infty}^0 (-s) p_S(s) ds + \int_0^{\infty} s p_S(s) ds = 2 \int_0^{\infty} s p_S(s) ds . \quad (15)$$

Using Eq. (14) and Eq. (15), we get

$$\begin{aligned} \mathbb{E}[S | S > 0] &= \mathbb{E}[|S|] = \mathbb{E}\left[\left|\sum_{k=1}^r z_k\right|\right] \\ &\stackrel{(\dagger)}{\geq} \mathbb{E}[|z_1|] \\ &= \mathbb{E}[|u_1 v_1|] = \mathbb{E}[|u_1| |v_1|] = \mathbb{E}[|u_1|] \mathbb{E}[|v_1|] \quad (\text{using independence}) \\ &= \sigma_u \sigma_v \mathbb{E}[|a|]^2 = \frac{2}{\pi} \sigma_u \sigma_v . \quad (\text{for } a \sim \mathcal{N}(0, 1)) \end{aligned}$$

Eq (†) holds intuitively. To formally show it, we invoke Lemma 3 (Whittle's inequality) on the convex function $\phi(x) = |x|$. Using this with Eq. (11) gives the desired result.

Proof of the upper bound. The probability density function of $z = uv$ is given by

$$f_z(x) = \frac{1}{\pi \sigma_u \sigma_v \sqrt{1 - \rho^2}} \exp\left(\frac{\rho x}{\sigma_u \sigma_v (1 - \rho^2)}\right) K_0\left(\frac{|x|}{\sigma_u \sigma_v (1 - \rho^2)}\right) ,$$

where $\rho = \gamma/(\sigma_u \sigma_v)$ denotes the correlation factor. Note that $|\rho| < 1$ is ensured via $\gamma < \sigma_u \sigma_v$ in the lemma statement. The function $K_0(a|x|)$ is log-concave for $a > 0$. The term $\exp(bx)$ is log-linear (hence log-concave). The product of log-concave functions is log-concave. Thus, $f_z(x)$ is log-concave. Since S is a sum of r i.i.d. random variables with log-concave densities, S also has a log-concave density. We use Lemma 4 to get that $\mathbb{E}[S | S > \theta] \leq \theta + e \sqrt{r(\sigma_u^2 \sigma_v^2 + \gamma^2)}$ for $\theta \geq r\gamma$. For $\theta \in [0, r\gamma]$, we use the non-decreasing property of $\mathbb{E}[S | S > \theta]$ from Eq. (13). Plugging into Eq. (11) concludes the argument. \square

Lemma 8. Consider Gaussian random variables $x, y \in \mathbb{R}^r$, such that

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} a_x \mathbf{I}_r & a_{xy} \mathbf{I}_r \\ a_{xy} \mathbf{I}_r & a_y \mathbf{I}_r \end{pmatrix}\right), \quad \text{with } a_x, a_y > 0, a_{xy} \geq 0 .$$

For $\theta \in \mathbb{R}$, define $\mathbf{A}(\theta) := \mathbb{E}[xy^\top | x^\top y > \theta]$. It holds that $\mathbf{A}(\theta)$ satisfies

$$\mathbf{A}(\theta) = f(\theta) \mathbf{I}_r ,$$

where $f(\theta)$ is a scalar function of $\theta \in \mathbb{R}$, such that

$$\max\left\{a_{xy}, \frac{\theta}{r}\right\} + e \sqrt{\frac{a_x a_y + a_{xy}^2}{r}} \geq f(\theta) \geq \max\left\{a_{xy}, \frac{\theta}{r}\right\} .$$

In the special case of $a_{xy} = 0$, it further holds that $f(0) \geq 2\sqrt{a_x a_y}/\pi r$.

Proof. We first build an intuition for the quantity $\mathbf{A}(\theta) \in \mathbb{R}^{r \times r}$. For $\theta = -\infty$, $\mathbf{A}(\theta)$ becomes the unconditional expectation, which is $a_{xy} \mathbf{I}_r$ according to the given covariance structure. As θ increases in \mathbb{R} , we expect $\mathbf{A}(\theta)$ to increase.

$\mathbf{A}(\theta)$ is diagonal. We first show that $\mathbf{A}(\theta)$ is a diagonal matrix. The (i, j) -th entry is $\mathbf{A}(\theta)_{ij} = \mathbb{E}[x_i y_j | Z > \theta]$, where $Z = x^\top y = \sum_{l=1}^r x_l y_l$. Consider the transformation $T_i : \mathbb{R}^{2r} \rightarrow \mathbb{R}^{2r}$ that maps (x, y) to (x', y') where $x'_l = x_l$ for $l \neq i$, $x'_i = -x_i$, and $y'_l = y_l$ for $l \neq i$, $y'_i = -y_i$.

First, note that $Z' = \sum_{l \neq i} x_l y_l + (-x_i)(-y_i) = Z$. Hence the condition $Z > \theta$ is invariant under the transformation T_i . Second, due to independence and the block diagonal structure of the covariance, the overall joint density is a product of univariate Gaussians centered around zero. Due to the symmetry of a univariate Gaussian, the overall density is also invariant under T_i . Third, the entry $x_i y_j$ becomes $-x_i y_j$ under the transformation T_i . Due to this symmetry, we conclude that the off-diagonal entries are zero.

860 **All the diagonal entries of $\mathbf{A}(\theta)$ are equal by symmetry.** The diagonal entries are $\mathbf{A}(\theta)_{ii} =$
861 $\mathbb{E}[x_i y_i | Z > \theta]$. Let $Z_i = x_i y_i$, meaning $Z = \sum_{l=1}^r Z_l$. Due to the block diagonal structure on
862 (x, y) , each Z_i is independent and identically distributed. Hence, $\mathbf{A}(\theta)_{ii} = \mathbf{A}(\theta)_{jj}$ for any $i, j \in [r]$.
863 **Properties of $f(\theta)$.** From the above two steps, we conclude that $\mathbf{A}(\theta) = f(\theta) \mathbf{I}_r$ for some scalar
864 function $f : \mathbb{R} \rightarrow \mathbb{R}$. Using the trace trick, we see that

$$\begin{aligned} f(\theta) \cdot \text{Tr}(\mathbf{I}_r) &= \text{Tr}(\mathbb{E}[xy^\top | x^\top y > \theta]) \\ \implies f(\theta) &= \frac{1}{r} \mathbb{E}[x^\top y | x^\top y > \theta]. \end{aligned}$$

Since the covariances of x, y are scaled identity, each $x_i y_i, i \in [r]$ is identically distributed. This
865 distribution is akin to uv for $u \sim \mathcal{N}(0, a_x), v \sim \mathcal{N}(0, a_y)$ with $\text{Cov}(u, v) = a_{xy}$. Hence

$$f(\theta) = \mathbb{E} \left[u_1 v_1 \mid \sum_{i=1}^r u_i v_i > \theta \right],$$

866 for u_i, v_i i.i.d. according to the described distribution. Lemma 7 shows the required properties on
867 this conditional expectation, showing the desired inequalities in the statement of this lemma. \square

868 **Lemma 9.** Let $x \in \mathbb{R}^d$ and $\tilde{x} \in \mathbb{R}^{\tilde{d}}$ be jointly Gaussian vectors with mean zero and joint covariance
869 matrix Σ_{full} which is positive definite. Consider $\mathbf{M}_O, \mathbf{M}_T \in \mathbb{R}^{d \times \tilde{d}}$, and assume $\text{rank}(\mathbf{M}_O) \geq 2$.
870 Define $\Delta \mathbf{M} := \mathbf{M}_T - \mathbf{M}_O$, and for any matrix $\mathbf{A} \in \mathbb{R}^{d \times \tilde{d}}$, define $Y_{\mathbf{A}} := x^\top \mathbf{A} \tilde{x}$. For a real $\theta \geq 0$,

$$\begin{aligned} \Delta \mathbf{E}(\theta) &:= \left\| \mathbb{E}[x \tilde{x}^\top \mathbb{I}(Y_{\mathbf{M}_T} > \theta)] - \mathbb{E}[x \tilde{x}^\top \mathbb{I}(Y_{\mathbf{M}_O} > \theta)] \right\|_2, \\ \Delta P(\theta) &:= |\mathbb{P}\{Y_{\mathbf{M}_T} > \theta\} - \mathbb{P}\{Y_{\mathbf{M}_O} > \theta\}|. \end{aligned}$$

871 Then, there exist constants $C_{\mathbf{E}}(\theta, \Sigma_{\text{full}}, \mathbf{M}_O) > 0$ and $C_P(\theta, \Sigma_{\text{full}}, \mathbf{M}_O) > 0$ that depend on θ , the
872 covariance Σ_{full} , and \mathbf{M}_O , such that:

$$\begin{aligned} \Delta \mathbf{E}(\theta) &\leq C_{\mathbf{E}}(\theta, \Sigma_{\text{full}}, \mathbf{M}_O) \|\mathbf{M}_T - \mathbf{M}_O\|_2, \\ \Delta P(\theta) &\leq C_P(\theta, \Sigma_{\text{full}}, \mathbf{M}_O) \|\mathbf{M}_T - \mathbf{M}_O\|_2. \end{aligned}$$

873 *Proof sketch.* Let $D(\theta) = \mathbb{I}(Y_{\mathbf{M}_T} > \theta) - \mathbb{I}(Y_{\mathbf{M}_O} > \theta)$. If $D(\theta) \neq 0$, then θ must lie between $Y_{\mathbf{M}_O}$
874 and $Y_{\mathbf{M}_T}$. This implies $|Y_{\mathbf{M}_O} - \theta| \leq |Y_{\mathbf{M}_T} - Y_{\mathbf{M}_O}| = |Y_{\Delta \mathbf{M}}|$. Thus, we have the pointwise bound
875 $|D(\theta)| \leq \mathbb{I}(|Y_{\mathbf{M}_O} - \theta| \leq |Y_{\Delta \mathbf{M}}|)$. We first argue that there exists a constant $C_A(\theta) > 0$ such that

$$\mathbb{E}[\mathbb{I}(|Y_{\mathbf{M}_O} - \theta| \leq |Y_{\Delta \mathbf{M}}|)] \leq C_A(\theta) \mathbb{E}[|Y_{\Delta \mathbf{M}}|]. \quad (16)$$

876 Let $p_O(y)$ denote the density of the random variable $Y_{\mathbf{M}_O}$. If this density is bounded everywhere by
877 B , then $P(|Y_{\mathbf{M}_O}| \leq |Y_{\Delta \mathbf{M}}| \mid |Y_{\Delta \mathbf{M}}|) \lesssim B |Y_{\Delta \mathbf{M}}|$, leading to $\mathbb{E}[\mathbb{I}(|Y_{\mathbf{M}_O}| \leq |Y_{\Delta \mathbf{M}}|)] \lesssim B \mathbb{E}[|Y_{\Delta \mathbf{M}}|]$.
878 For the density $p_O(y)$ to be bounded, the main consideration point is $y = 0$. This is because the
879 random variable $Y_{\mathbf{M}_O}$ is a quadratic form as described, which can be singular at $y = 0$ (for example,
880 the degenerate case of $\mathbf{M}_O = \mathbf{0}$ causes this). It is known that $\text{rank}(\mathbf{M}_O) \geq 2$ is sufficient to ensure a
881 bounded density, leading to a finite B .

882 **1. Bounding $\Delta P(\theta)$:** First rewrite

$$\Delta P(\theta) = |\mathbb{E}[D(\theta)]| \leq \mathbb{E}[|D(\theta)|] \leq \mathbb{E}[\mathbb{I}(|Y_{\mathbf{M}_O} - \theta| \leq |Y_{\Delta \mathbf{M}}|)].$$

883 We use Eq. (16) and the fact that $|x^\top \Delta \mathbf{M} \tilde{x}| \leq \|\Delta \mathbf{M}\|_2 \|x\|_2 \|\tilde{x}\|_2$ holds pointwise. Since
884 $\mathbb{E}[\|x\|_2 \|\tilde{x}\|_2]$ is bounded by constants that depend on the covariance Σ_{full} , the upper bound on
885 $\Delta P(\theta)$ follows, with $C_P(\theta, \Sigma_{\text{full}}, \mathbf{M}_O)$ proportional to $C_A(\theta)$.

2. Bounding $\Delta \mathbf{E}(\theta)$: The matrix inside the norm defining $\Delta \mathbf{E}(\theta)$ is $\mathbb{E}[x \tilde{x}^\top D(\theta)]$. For an index $k \in$
 $[d], l \in [\tilde{d}]$, consider an element $|\mathbb{E}[x_k \tilde{x}_l D(\theta)]| \leq \mathbb{E}[|x_k \tilde{x}_l| |D(\theta)|]$. Using $|D(\theta)| \leq \mathbb{I}(|Y_{\mathbf{M}_O} - \theta| \leq$
 $|Y_{\Delta \mathbf{M}}|)$, we have

$$\mathbb{E}[|x_k \tilde{x}_l| |D(\theta)|] \leq \mathbb{E}[|x_k \tilde{x}_l| \mathbb{I}(|Y_{\mathbf{M}_O} - \theta| \leq |Y_{\Delta \mathbf{M}}|)].$$

886 Following a similar reasoning as the previous part and invoking Eq. (16), since the moments of any
887 Gaussian raised to a finite power is bounded, each entry in the matrix of $\Delta \mathbf{E}(\theta)$ is bounded. Hence
888 the upper bound on $\Delta \mathbf{E}(\theta)$ holds, again with the constant proportional to $C_A(\theta)$. \square

Lemma 10. Let $x_1, \dots, x_n \in \mathbb{R}^d$ be n i.i.d. random vectors drawn from a Gaussian distribution $\mathcal{N}(0, \Sigma)$, where Σ is a $d \times d$ positive definite covariance matrix, $d \geq 1, n \geq 1$. Let S be a random subset of indices $\{1, \dots, n\}$ generated by including each index $j \in \{1, \dots, n\}$ independently with probability $p \in (0, 1]$. Let $n_c = |S|$ denote the number of selected samples, and define the sample covariance matrix for $n_c > 0$ as $\hat{\Sigma}_{n_c} = (1/n_c) \sum_{i \in S} x_i x_i^\top$. For a failure probability $\delta \in (0, 1)$, assume that $np > 8 \log(2/\delta)$ holds. Then, with probability at least $1 - \delta$, both $n_c \geq np/2$ and the sample covariance matrix of the selected data satisfies:

$$\|\hat{\Sigma}_{n_c} - \Sigma\|_2 \lesssim \|\Sigma\|_2 \sqrt{\frac{d + \log \frac{1}{\delta}}{np}}.$$

889 *Proof.* Define $k_{\min} := \lceil np - \sqrt{2np \log(2/\delta)} \rceil$. Note that $k_{\min} \geq np/2$ due to the assumption. Let

$$\mathcal{F}_1 := \{n_c < k_{\min}\}, \quad \mathcal{F}_2 := \left\{ n_c \geq k_{\min} \text{ and } \|\hat{\Sigma}_{n_c} - \Sigma\|_2 > \|\Sigma\|_2 \sqrt{\frac{d + \log \frac{1}{\delta}}{k_{\min}}} \right\}.$$

890 denote the failure events. A union bound over the two failure probabilities will give the desired result.
891 Below we bound the individual failure probabilities.

892 **Bounding $\mathbb{P}(\mathcal{F}_1)$:** Define $\Delta_0 := \sqrt{2 \log(2/\delta)/(np)}$, so that $k_{\min} = \lceil (1 - \Delta_0)np \rceil$. Since we
893 assumed $np > 8 \log(2/\delta)$, $\Delta_0 < 0.5$. By a standard Chernoff bound for binomial distributions,
894 $\mathbb{P}(n_c < (1 - \Delta_0)np) \leq \exp(-np\Delta_0^2/2) = \exp(-\log(2/\delta)) = \delta/2$. Since $k_{\min} \geq (1 - \Delta_0)np$
895 (due to the ceil operation), it follows that $\mathbb{P}(\mathcal{F}_1) = \mathbb{P}(n_c < k_{\min}) \leq \mathbb{P}(n_c \leq (1 - \Delta_0)np) \leq \delta/2$.

Bounding $\mathbb{P}(\mathcal{F}_2)$: Using the law of total probability, we write

$$\mathbb{P}(\mathcal{F}_2) = \sum_{k=k_{\min}}^n \mathbb{P} \left(\left\| \frac{1}{k} \sum_{i \in S, |S|=k} x_i x_i^\top - \Sigma \right\|_2 > \|\Sigma\|_2 \sqrt{\frac{d + \log \frac{1}{\delta}}{k_{\min}}} \mid n_c = k \right) \mathbb{P}(n_c = k)$$

896 For any $k \geq k_{\min}$, we have $1/\sqrt{k} \leq 1/\sqrt{k_{\min}}$. Thus, for $k \geq k_{\min}$:

$$\begin{aligned} \mathbb{P} \left(\left\| \frac{1}{k} \sum x_i x_i^\top - \Sigma \right\|_2 > \|\Sigma\|_2 \sqrt{\frac{d + \log \frac{1}{\delta}}{k_{\min}}} \mid n_c = k \right) &\leq \\ \mathbb{P} \left(\left\| \frac{1}{k} \sum x_i x_i^\top - \Sigma \right\|_2 > \|\Sigma\|_2 \sqrt{\frac{d + \log \frac{1}{\delta}}{k}} \mid n_c = k \right). \end{aligned}$$

897 And the right hand side is bounded by $\delta/2$ owing to standard matrix concentration results. So,
898 $\mathbb{P}(\mathcal{F}_2) \leq \sum_{k=k_{\min}}^n (\delta/2) \mathbb{P}(n_c = k) \leq \delta/2$. \square

899 **D A proof of Corollary 1**

900 We present a proof of Corollary 1, which follows the proof presented in Nakada et al. [22] while
901 fixing some typos. Before diving into the proof, we make some remarks.

902 First, the result in Nakada et al. [22] is for a general covariance on the signal, Σ_z , and the noise, Σ_ξ ,
903 whereas our setting is more restricted from Assumptions 1 and 2. This restriction is required for the
904 analysis of filtering in Theorem 1. Second, the result in [22] is stated with probability $1 - O(1/n)$,
905 whereas we state it with probability $1 - \exp(-d)$. Due to this, Corollary 1 as stated does not have a
906 $\log n$ factor inside the square root, unlike Nakada et al. [22, Theorem 3.1].

907 Third, there is a small subtle difference in the setting of [22] and ours. We use η to denote the
908 fixed probability of clean samples in Assumption 1, whereas Nakada et al. [22] use η to denote the
909 fraction of clean samples in the *sampled* dataset, which is a random quantity. Using n_c to denote
910 the number of clean samples, we go through the additional step of controlling the error in $|n_c/n - \eta|$,
911 which scales as $1/\sqrt{n}$, since this source of error is 1-dimensional. Fourth, the result in Nakada et al.

[22] Theorem 3.1] is stated as $\min\{\sqrt{r}, \cdot\}$. While it is true that the $\sin \Theta$ metric can be at most \sqrt{r} , the final step in the proof is the application Lemma 2 which requires a condition that translates to $n \gtrsim (1/\eta^2) \max\{d, \tilde{d}\} (1 + \gamma^{-1})(1 + \tilde{\gamma}^{-1})$. And so this is how we state the result in Corollary 1, which makes the stated upper bound always smaller than \sqrt{r} .

For clarity, we write the algorithm:

Input. $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times \tilde{d}}$, $r \in \mathbb{Z}_+$, $\rho \in (0, \infty)$.

Output. $\mathbf{G}^\top \tilde{\mathbf{G}} \in \mathbb{R}^{d \times \tilde{d}}$ (with rank = r , since $\mathbf{G} \in \mathbb{R}^{r \times d}$, $\tilde{\mathbf{G}} \in \mathbb{R}^{r \times \tilde{d}}$) by minimizing Eq. (2).

Step 1: Reduction of loss. We show that

$$\mathcal{L}_0(\mathbf{G}, \tilde{\mathbf{G}}) = -\text{Tr}(\mathbf{G} \mathbf{S}_n \tilde{\mathbf{G}}^\top), \quad (17)$$

where \mathbf{S}_n denotes the cross covariance matrix of the data, given by (Eq. (4) rewritten)

$$\mathbf{S}_n = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) (\tilde{x}_i - \bar{\tilde{x}})^\top \in \mathbb{R}^{d \times \tilde{d}}.$$

Proof. Expand the LHS as

$$\begin{aligned} \mathcal{L}_0(\mathbf{G}, \tilde{\mathbf{G}}) &= \frac{1}{2n(n-1)} \left(\sum_{i=1}^n \left(\sum_{\substack{j=1 \\ j \neq i}}^n (s_{ij} - s_{ii}) + \sum_{\substack{j=1 \\ j \neq i}}^n (s_{ji} - s_{ii}) \right) \right) \\ &\stackrel{(a)}{=} \frac{1}{n(n-1)} \left(\sum_{i=1}^n \left(\sum_{\substack{j=1 \\ j \neq i}}^n (s_{ij} - s_{ii}) \right) \right) \\ &= \frac{1}{n(n-1)} \left(\sum_i \sum_{j \neq i} s_{ij} - (n-1) \sum_i s_{ii} \right) \\ &= \frac{1}{n(n-1)} \left(\sum_i \sum_{j \neq i} s_{ij} \right) - \frac{1}{n} \left(\sum_i s_{ii} \right), \end{aligned} \quad (\text{X})$$

where eq (a) holds because the overall sum over the $n \times n$ similarity matrix is the same whether done over rows or columns.

For the RHS, we first rewrite \mathbf{S}_n as

$$\begin{aligned} \mathbf{S}_n &= \frac{1}{n-1} \left(\sum_{i=1}^n x_i \tilde{x}_i^\top - n \bar{x} \bar{\tilde{x}}^\top \right) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n x_i \tilde{x}_i^\top \right) - \frac{1}{n(n-1)} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n \tilde{x}_i \right)^\top \\ &= \frac{1}{n-1} \left(\sum_i x_i \tilde{x}_i^\top \right) - \frac{1}{n(n-1)} \left(\sum_i x_i \tilde{x}_i^\top + \sum_i \sum_{j \neq i} x_i \tilde{x}_j^\top \right) \\ &= \frac{1}{n-1} \left(1 - \frac{1}{n} \right) \left(\sum_i x_i \tilde{x}_i^\top \right) - \frac{1}{n(n-1)} \left(\sum_i \sum_{j \neq i} x_i \tilde{x}_j^\top \right) \\ &= \frac{1}{n} \left(\sum_i x_i \tilde{x}_i^\top \right) - \frac{1}{n(n-1)} \left(\sum_i \sum_{j \neq i} x_i \tilde{x}_j^\top \right). \end{aligned}$$

925 Using the above, we rewrite the RHS as

$$\begin{aligned}
-\text{Tr}(\mathbf{G}\mathbf{S}_n\tilde{\mathbf{G}}^\top) &= -\text{Tr}\left(\frac{1}{n}\left(\sum_i \mathbf{G}x_i\tilde{x}_i^\top\tilde{\mathbf{G}}^\top\right) - \frac{1}{n(n-1)}\left(\sum_i \sum_{j \neq i} \mathbf{G}x_i\tilde{x}_j^\top\tilde{\mathbf{G}}^\top\right)\right) \\
&= \frac{1}{n(n-1)}\left(\sum_i \sum_{j \neq i} \text{Tr}(\mathbf{G}x_i\tilde{x}_j^\top\tilde{\mathbf{G}}^\top)\right) - \frac{1}{n}\left(\sum_i \text{Tr}(\mathbf{G}x_i\tilde{x}_i^\top\tilde{\mathbf{G}}^\top)\right) \\
&\quad \text{(Linearity of Trace)} \\
&= \frac{1}{n(n-1)}\left(\sum_i \sum_{j \neq i} \langle \mathbf{G}x_i, \tilde{\mathbf{G}}\tilde{x}_j \rangle\right) - \frac{1}{n}\left(\sum_i \langle \mathbf{G}x_i, \tilde{\mathbf{G}}\tilde{x}_i \rangle\right) \\
&\quad \text{(Cyclic nature of Trace)} \\
&= \frac{1}{n(n-1)}\left(\sum_i \sum_{j \neq i} s_{ij}\right) - \frac{1}{n}\left(\sum_i s_{ii}\right). \quad \text{(Definition of } s_{ij}\text{)}
\end{aligned}$$

926 Comparing the above to eq (X) concludes the proof. \square

927 **Step 2: Closed-form solution.** We show that (Eq. (5) rewritten)

$$\arg \min_{\mathbf{G}, \tilde{\mathbf{G}}} \mathcal{L}_\rho(\mathbf{G}, \tilde{\mathbf{G}}) = \left\{ (\mathbf{G}, \tilde{\mathbf{G}}) \mid \mathbf{G}^\top \tilde{\mathbf{G}} = \frac{1}{\rho} \text{SVD}_r(\mathbf{S}_n) \right\}.$$

928 Hence, even though the optimization problem is non-convex, there is a closed-form solution, and
929 no optimization analysis is needed. In particular, the right singular vectors of $\mathbf{G}, \tilde{\mathbf{G}}$ are determined
930 independent of the choice of ρ . This result is from Nakada et al. [22] Lemma 2.1].

931 *Proof.* Using Step 1's result, we can write

$$\min_{\mathbf{G}, \tilde{\mathbf{G}}} \mathcal{L}_\rho(\mathbf{G}, \tilde{\mathbf{G}}) \equiv \max_{\mathbf{G}, \tilde{\mathbf{G}}} \text{Tr}(\mathbf{G}\mathbf{S}_n\tilde{\mathbf{G}}^\top) - \frac{\rho}{2} \|\mathbf{G}^\top \tilde{\mathbf{G}}\|_F^2. \quad (18)$$

932 The objective can be rewritten as

$$\text{Tr}(\mathbf{G}\mathbf{S}_n\tilde{\mathbf{G}}^\top) - \frac{\rho}{2} \|\mathbf{G}^\top \tilde{\mathbf{G}}\|_F^2 = \frac{\rho}{2} \left(\left\| \frac{\mathbf{S}_n}{\rho} \right\|_F^2 - \left\| \mathbf{G}^\top \tilde{\mathbf{G}} - \frac{\mathbf{S}_n}{\rho} \right\|_F^2 \right).$$

933 The optimization variables appear only in the second term. Since $\text{rank}(\mathbf{G}^\top \tilde{\mathbf{G}}) = r$, by the
934 Eckart-Young-Minsky Theorem, the solution is given by the best rank r approximation of \mathbf{S}_n/ρ . \square

935 **Step 3: Relating error to op-norm concentration of \mathbf{S}_n .** We show the below, where \mathbf{S}_n concentrates
936 to $\mathbf{S} = \eta \mathbf{U}\tilde{\mathbf{U}}^\top$.

$$\|\text{SVD}_r(\mathbf{S}_n) - \mathbf{S}\| \leq 2 \|\mathbf{S}_n - \mathbf{S}\|. \quad (19)$$

937 *Proof.* By triangle inequality, we have

$$\|\text{SVD}_r(\mathbf{S}_n) - \mathbf{S}\| \leq \|\text{SVD}_r(\mathbf{S}_n) - \mathbf{S}_n\| + \|\mathbf{S}_n - \mathbf{S}\|.$$

938 And for the first term on the right hand side, we use

$$\begin{aligned}
\|\text{SVD}_r(\mathbf{S}_n) - \mathbf{S}_n\| &= \sigma_{r+1}(\mathbf{S}_n) \\
&\stackrel{(\dagger)}{\leq} \sigma_{r+1}(\mathbf{S}) + \|\mathbf{S}_n - \mathbf{S}\| \\
&\stackrel{(\dagger\dagger)}{\leq} \|\mathbf{S}_n - \mathbf{S}\|.
\end{aligned}$$

939 In Eq. (†), we used Lemma 1 and Eq. (††) holds because $\sigma_{r+1}(\mathbf{S}) = 0$, since \mathbf{S} is rank r . \square

940 **Step 4: Concentration of \mathbf{S}_n .** We show that the following holds:

$$\text{w.p. } 1 - \exp\left(-\Omega(\max\{d, \tilde{d}\})\right), \quad \|\mathbf{S}_n - \mathbf{S}\| \lesssim \sqrt{\frac{\max\{d, \tilde{d}\}(1 + \gamma^{-1})(1 + \tilde{\gamma}^{-1})}{n}}. \quad (20)$$

941 *Proof.* We start with the expansion of \mathbf{S}_n ,

$$\mathbf{S}_n = \frac{1}{n-1} \sum_{i=1}^n x_i \tilde{x}_i^\top - \frac{n}{n-1} \bar{x} \bar{x}^\top = \underbrace{\frac{1}{n} \sum_{i=1}^n x_i \tilde{x}_i^\top}_{\mathbf{S}_n^{(1)}} - \underbrace{\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n x_i \tilde{x}_j^\top}_{\mathbf{S}_n^{(2)}}.$$

942 The main term that dictates the convergence is $\mathbf{S}_n^{(1)}$. The term $\mathbf{S}_n^{(2)}$ concentrates around zero (since
 943 samples $i \neq j$, $i, j \in [n]$ are independent), and the rate of convergence is $\tilde{O}(1/n)$ due to averaging
 944 over n^2 terms, which is a higher order term. Let n_c be a random variable that denotes the number of
 945 clean data points. We expand the sum in $\mathbf{S}_n^{(1)}$ below.

$$\begin{aligned} \sum_{i=1}^n x_i \tilde{x}_i^\top &= \underbrace{\sum_{i=1}^{n_c} \mathbf{U} z_i \tilde{z}_i^\top \tilde{\mathbf{U}}^\top}_{J_1} + \underbrace{\sum_{i=n_c+1}^n \mathbf{U} z_i \tilde{z}_i^\top \tilde{\mathbf{U}}^\top}_{J_2} \\ &\quad + \underbrace{\sum_{i=1}^n \mathbf{U} z_i \tilde{\xi}_i^\top}_{K_1} + \underbrace{\sum_{i=1}^n \xi_i \tilde{z}_i^\top \tilde{\mathbf{U}}^\top}_{K_2} + \underbrace{\sum_{i=1}^n \xi_i \tilde{\xi}_i^\top}_{K_3}. \end{aligned}$$

946 We control the error in each term separately. For terms $J_2, K_{1:3}$, we need a result like Nakada et al.
 947 [22] Proposition C.1] in the simple case of $X \perp \tilde{X}$. For term J_1 , we need it for $X = \tilde{X}$.

948 The following two facts are going to be used multiple times. Here X, Y denote random quantities,
 949 and all others are fixed quantities (matrices/vectors).

$$\text{w.h.p. } \|X - A\| \leq E_A, \|Y - B\| \leq E_B \implies \text{w.h.p. } \|X + Y - (A + B)\| \leq E_A + E_B, \quad (21)$$

$$\text{w.h.p. } \|X - A\| \leq E_A \implies \text{w.h.p. } \|MXN - MAN\| \leq \|M\| \|N\| E_A. \quad (22)$$

950 For the independent terms $(J_2, K_{1:3})$, we will use the below generic result. For $\mathbb{R}^{d_x} \ni x \sim \mathcal{N}(0, \Sigma_x)$
 951 and $\mathbb{R}^{d_y} \ni y \sim \mathcal{N}(0, \Sigma_y)$ and N i.i.d. draws from both, we have the below result from the application
 952 of a Matrix-Bernstein result.

$$\text{w.p. } 1 - e^{-t}, \quad \left\| \frac{1}{N} \sum_{i=1}^N x_i y_i^\top \right\| \lesssim \sqrt{\frac{\|\Sigma_x\| \cdot \|\Sigma_y\|}{N}} (t + \log(d_x + d_y)). \quad (23)$$

953 For the dependent term (J_1) , we will use the below. Let $\mathbb{R}^{d_x} \ni x \sim \mathcal{N}(0, \Sigma_x)$ and N i.i.d. draws
 from this. This is also known in the literature, for e.g., Bunea and Xiao [4] Theorem 2.2].

$$\text{w.p. } 1 - e^{-t}, \quad \left\| \frac{1}{N} \sum_{i=1}^N x_i x_i^\top - \Sigma_x \right\| \lesssim \|\Sigma_x\| \sqrt{\frac{t + \log(d_x)}{N}}. \quad (24)$$

954 Note that the above two concentration results are tighter than Nakada et al. [22] Proposition C.1] by a
 955 factor of dimension, since the proposition has trace terms too, whereas only operator norms appear in
 956 the above two equations. This manifests in Corollary 1 as stated being tighter than Nakada et al. [22]
 957 Theorem 3.1] by a dimension factor inside the square root (since we avoided $\log n$ but did not incur
 958 an additional dimension due to the failure probability of $\exp(-d)$). Finally, since $n_c = \text{Bin}(n, \eta)$,
 959 the ratio n_c/n concentrates to η , with the error described by Hoeffding's inequality as

$$\mathbb{P}\left(\left|\frac{n_c}{n} - \eta\right| \geq \epsilon\right) \leq 2 \exp(-2n\epsilon^2). \quad (25)$$

Using these results, we bound the individual terms of deviation. We first bound the independent terms using Eq. (23) with $t := \max\{d, \tilde{d}\}$. The choice of N is given with each setting. With probability $1 - \exp(-\Omega(\max\{d, \tilde{d}\}))$, the following hold:

$$\left\| \frac{K_1}{n} \right\| \lesssim \sqrt{\frac{\|\Sigma_z\| \cdot \|\Sigma_\xi\| \cdot \max\{d, \tilde{d}\}}{n}} = \sqrt{\frac{\max\{d, \tilde{d}\} \tilde{\gamma}^{-1}}{n}}, \quad (N := n)$$

$$\left\| \frac{K_2}{n} \right\| \lesssim \sqrt{\frac{\|\Sigma_z\| \cdot \|\Sigma_\xi\| \cdot \max\{d, \tilde{d}\}}{n}} = \sqrt{\frac{\max\{d, \tilde{d}\} \gamma^{-1}}{n}}, \quad (N := n)$$

$$\left\| \frac{K_3}{n} \right\| \lesssim \sqrt{\frac{\|\Sigma_\xi\| \cdot \|\Sigma_\xi\| \cdot \max\{d, \tilde{d}\}}{n}} = \sqrt{\frac{\max\{d, \tilde{d}\} \gamma^{-1} \tilde{\gamma}^{-1}}{n}}, \quad (N := n)$$

$$\left\| \frac{J_2}{n} \right\| \lesssim \sqrt{1 - \frac{n_c}{n}} \cdot \sqrt{\frac{\|\Sigma_z\|^2 \cdot \max\{d, \tilde{d}\}}{n}} = \sqrt{1 - \frac{n_c}{n}} \cdot \sqrt{\frac{\max\{d, \tilde{d}\}}{n}}. \quad (N := n - n_c)$$

We now bound the dependent term using Eq. (24). We need some additional machinery to deal with the random denominator, which we capture in Lemma 10. The requirement of $np \gtrsim \log(1/\delta)$ in the lemma translates to $n \gtrsim \max\{d, \tilde{d}\}/\eta$, since we have $p := \eta$ and $\delta := \exp(-\max\{d, \tilde{d}\})$. As we will see later, step 5 of the proof requires $n \gtrsim \max\{d, \tilde{d}\}/\eta^2$, hence we will assume this requirement is satisfied. With probability $1 - \exp(-\Omega(\max\{d, \tilde{d}\}))$, it holds:

$$\begin{aligned} \left\| \frac{J_1}{n_c} - \mathbf{U} \tilde{\mathbf{U}}^\top \right\| &\lesssim \left\| \mathbf{U} \tilde{\mathbf{U}}^\top \right\| \cdot \sqrt{\frac{\max\{d, \tilde{d}\}}{n\eta}} \\ \Rightarrow \left\| \frac{J_1}{n} - \frac{n_c}{n} \mathbf{U} \tilde{\mathbf{U}}^\top \right\| &\lesssim \frac{n_c}{n} \sqrt{\frac{1}{\eta}} \cdot \sqrt{\frac{\max\{d, \tilde{d}\}}{n}} \\ \Rightarrow \left\| \frac{J_1}{n} - \eta \mathbf{U} \tilde{\mathbf{U}}^\top \right\| &\lesssim \frac{n_c}{n} \sqrt{\frac{1}{\eta}} \cdot \sqrt{\frac{\max\{d, \tilde{d}\}}{n}} + \left| \frac{n_c}{n} - \eta \right|. \end{aligned} \quad (26)$$

For the concentration of n_c/n , we use Eq. (25) to get that with probability $1 - \exp(-\Omega(\max\{d, \tilde{d}\}))$:

$$\left| \frac{n_c}{n} - \eta \right| \lesssim \sqrt{\frac{\max\{d, \tilde{d}\}}{n}}. \quad (27)$$

We now add all the error bounds. For the combined error from terms J_1 and J_2 , since $1 \leq \sqrt{a} + \sqrt{1-a} \leq \sqrt{2}$, the quantity $\sqrt{n_c/n}$ can be removed up to constants. Eq. (20) follows. \square

Step 5: Relating singular vector recovery error to operator norm concentration. We will apply Lemma 2 (a Davis-Kahan type result) to relate the $\sin \Theta$ metric to the operator norm. Combining Eqs. (20), (19) and (5), we get that with probability $1 - \exp(-\Omega(\max\{d, \tilde{d}\}))$:

$$\left\| \mathbf{G}^\top \tilde{\mathbf{G}} - \frac{\eta}{\rho} \mathbf{U} \tilde{\mathbf{U}}^\top \right\| \lesssim \frac{1}{\rho} \sqrt{\frac{\max\{d, \tilde{d}\} (1 + \gamma^{-1})(1 + \tilde{\gamma}^{-1})}{n}}. \quad (28)$$

The instantiation for Lemma 2 is as follows: $A = \frac{\eta}{\rho} \mathbf{U} \tilde{\mathbf{U}}^\top$, $\hat{A} = \mathbf{G}^\top \tilde{\mathbf{G}}$. Note that both A, \hat{A} are rank- r , and $\sigma_r(A) = \eta/\rho$. We get

$$\left\| \sin \Theta \left(\text{lsv}(\mathbf{G}^\top \tilde{\mathbf{G}}), \mathbf{U} \right) \right\|_F \leq \frac{\left\| \mathbf{G}^\top \tilde{\mathbf{G}} - \frac{\eta}{\rho} \mathbf{U} \tilde{\mathbf{U}}^\top \right\|_F}{\frac{\eta}{\rho} - \left\| \mathbf{G}^\top \tilde{\mathbf{G}} - \frac{\eta}{\rho} \mathbf{U} \tilde{\mathbf{U}}^\top \right\|_2}. \quad (29)$$

Now we will use three things. First, for the numerator, we use $\|M\|_F \leq \sqrt{\text{rank}(M)} \cdot \|M\|_2$. Second, for the denominator, we will need the additional condition of $n \gtrsim (1/\eta^2) \max\{d, \tilde{d}\} (1 + \gamma^{-1})(1 + \tilde{\gamma}^{-1})$ to ensure the second term is at most half of the first term. This appears to have been missed by [22].

979 Third, triangle inequality with the fact that $\left\| \sin \Theta \left(\text{lsv}(\mathbf{G}^\top \tilde{\mathbf{G}}), \text{rsv}(\mathbf{G}) \right) \right\|_F = 0$ gives the final
 980 result. To see this fact, write

$$\begin{aligned} \mathbf{G}^\top \tilde{\mathbf{G}} &= V_{\mathbf{G}} (\Sigma_{\mathbf{G}} U_{\mathbf{G}}^\top U_{\tilde{\mathbf{G}}} \Sigma_{\tilde{\mathbf{G}}}) V_{\tilde{\mathbf{G}}}^\top \\ &= V_{\mathbf{G}} P S Q^\top V_{\tilde{\mathbf{G}}}^\top. \end{aligned} \quad (\text{Using SVD of the middle component})$$

981 Using the uniqueness of SVD, we get that $\text{lsv}(\mathbf{G}^\top \tilde{\mathbf{G}}) = V_{\mathbf{G}} P$ and $\text{rsv}(\mathbf{G}^\top \tilde{\mathbf{G}}) = V_{\tilde{\mathbf{G}}} Q$. Since
 982 P, Q are just orthogonal transforms, the subspace spanned by $V_{\mathbf{G}}$ and $V_{\mathbf{G}} P$ are the same, implying
 983 $\left\| \sin \Theta(V_{\mathbf{G}}, V_{\mathbf{G}} P) \right\|_F = 0$ (and analogously for $V_{\tilde{\mathbf{G}}}$ and $V_{\tilde{\mathbf{G}}} Q$).

984 Combining Eqs. (28) and (29) gives the desired result. Since the upper bound is valid for recovery of
 985 both \mathbf{U} and $\tilde{\mathbf{U}}$, Corollary 1 as stated follows.

986 E A proof of Proposition 1

987 Consider the following construction for the hard problem instance (lower bound): (i) the latent
 988 dimension $r = 1$, and (ii) the noise $\tilde{\xi} = 0$ (i.e. $\tilde{\gamma} = \infty$), but $\xi \neq 0$ (i.e. γ is finite). This means the
 989 following proof recovers the $d\gamma^{-1}$ part from the $\max\{d\gamma^{-1}, \tilde{d}\tilde{\gamma}^{-1}\}$ term in Proposition 1. A similar
 990 argument can be made for the case when $\xi = 0, \tilde{\xi} \neq 0$, leading to the max over both errors.

991 Owing to $r = 1$, this becomes a 1-dimensional vector recovery problem. Let $\mathbf{u}, \tilde{\mathbf{u}} \in \mathbb{R}^d$ denote the
 992 vectors to recover. Upon seeing \mathbf{S}_n , there is no error in estimating $\tilde{\mathbf{u}}$ since $\tilde{\xi} = 0$, but there is error in
 993 estimating \mathbf{u} . To calculate this error, define \mathbf{u}_n to be the top-left singular vector of \mathbf{S}_n . Note that \mathbf{S}_n
 994 has only one non-zero singular value, since it fully lies on $\tilde{\mathbf{u}}$ in the right singular vector space (i.e.
 995 $\mathbf{S}_n \mathbf{v} = 0$ for any $\mathbf{v} \perp \tilde{\mathbf{u}}$). Hence

$$\mathbf{S}_n = \|\mathbf{S}_n\| \cdot \mathbf{u}_n \tilde{\mathbf{u}}^\top. \quad (30)$$

996 **Step 0. Writing down \mathbf{S}_n .**

$$\begin{aligned} \mathbf{S}_n &= \frac{1}{n-1} \sum_{i=1}^n x_i \tilde{x}_i^\top - \frac{1}{n(n-1)} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n \tilde{x}_i \right)^\top \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n x_i \tilde{x}_i^\top}_{\mathbf{S}_n^{(1)}} - \underbrace{\frac{1}{n(n-1)} \left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n x_i \tilde{x}_j^\top \right)}_{\mathbf{S}_n^{(2)}}. \end{aligned}$$

997 We expand $\mathbf{S}_n^{(1)}$ below, using n_c to denote the random variable denoting the clean samples. Note that
 998 $\mathbb{E} n_c = \eta n$. Similarly one can expand $\mathbf{S}_n^{(2)}$, however, the error of $\mathbf{S}_n^{(2)}$ will behave as $O(1/n)$ due to
 999 averaging over n^2 samples, which is a higher order term in the overall rate. That is, the behavior (in
 1000 the large n regime) will be largely dictated by $\mathbf{S}_n^{(1)}$.

$$\mathbf{S}_n^{(1)} = \frac{1}{n} \sum_{i=1}^n x_i \tilde{x}_i^\top = \frac{1}{n} \sum_{i=1}^{n_c} (z_i \mathbf{u} + \xi_i)(z_i \tilde{\mathbf{u}})^\top + \frac{1}{n} \sum_{i=n_c+1}^n (z_i \mathbf{u} + \xi_i)(\tilde{z}_i \tilde{\mathbf{u}})^\top.$$

1001 As for the expectations, they are given by:

$$\begin{aligned} \mathbb{E} [\mathbf{S}_n^{(1)}] &= \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^{n_c} z_i^2 \right] \mathbf{u} \tilde{\mathbf{u}}^\top = \frac{1}{n} \mathbb{E} [n_c] \mathbf{u} \tilde{\mathbf{u}}^\top = \eta \mathbf{u} \tilde{\mathbf{u}}^\top, \\ \mathbb{E} [\mathbf{S}_n^{(2)}] &= 0 \quad (\text{since all random quantities are zero-mean and independent}). \end{aligned}$$

1002 **Step 1. Decompose $\sin \theta$ metric.** Our goal is a high probability lower bound on $|\sin \theta(\mathbf{u}_n, \mathbf{u})|$,
 1003 where \mathbf{u}_n is the random quantity. Note that

$$|\sin \theta(\mathbf{u}_n, \mathbf{u})| = \|(\mathbf{I}_d - \mathbf{u} \mathbf{u}^\top) \mathbf{u}_n\|. \quad (31)$$

1004 To see this, note that $LHS = \sqrt{1 - (\mathbf{u}^\top \mathbf{u}_n)^2}$. Squaring both sides and expanding suffices.

1005 **Step 2. Compute the metric for this case.** Using Eq. (30) in Eq. (31), we can write

$$|\sin \theta(\mathbf{u}_n, \mathbf{u})| = \frac{\|(\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top) \mathbf{S}_n \tilde{\mathbf{u}}\|}{\|\mathbf{S}_n\|}. \quad (32)$$

1006 **Step 3. Computing the high probability bound.** We will give high probability lower bound on the
 1007 numerator and denominator of Eq. (32) separately.

1008 **Step 3.1. For the numerator:** We first expand $\mathbf{S}_n^{(1)}$ as

$$\begin{aligned} \mathbf{S}_n^{(1)} \tilde{\mathbf{u}} &= \frac{1}{n} \sum_{i=1}^{n_c} (z_i^2 \mathbf{u} + z_i \xi_i) + \frac{1}{n} \sum_{i=n_c+1}^n (z_i \tilde{z}_i \mathbf{u} + \tilde{z}_i \xi_i) \\ \implies (\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top) \mathbf{S}_n^{(1)} \tilde{\mathbf{u}} &= (\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top) \left(\frac{1}{n} \sum_{i=1}^{n_c} z_i \xi_i + \frac{1}{n} \sum_{i=n_c+1}^n \tilde{z}_i \xi_i \right) \\ &\stackrel{d}{=} (\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top) \left(\frac{1}{n} \sum_{i=1}^n z_i \xi_i \right). \end{aligned}$$

1009 Similarly, for $\mathbf{S}_n^{(2)}$ we have

$$\begin{aligned} (\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top) \mathbf{S}_n^{(2)} \tilde{\mathbf{u}} &= (\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top) \left(\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \tilde{z}_j \xi_i \right) \\ &\stackrel{d}{=} (\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top) \left(\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n z_j \xi_i \right). \end{aligned}$$

1010 Combining the two, we get

$$(\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top) \mathbf{S}_n \tilde{\mathbf{u}} = (\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top) \underbrace{\left(\frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z}) (\xi_i - \bar{\xi}) \right)}_{\mathbf{w}_n}.$$

1011 Now we want to compute a high confidence lower bound on the norm of the above. We first relate
 1012 $\|(\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top) \mathbf{w}_n\|$ to $\|\mathbf{w}_n\|$. This is because \mathbf{w}_n is spherically symmetric, and $(\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top)$ is a
 1013 rank- $(d-1)$ matrix with all non-zero eigenvalues equal to one. We get

$$\|(\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top) \mathbf{w}_n\| = \|\mathbf{w}_n\| \cdot \sqrt{1 - (\mathbf{u}^\top \hat{\mathbf{w}}_n)^2}.$$

1014 Now due to \mathbf{w}_n being spherically symmetric, $\|\mathbf{w}_n\|$ (the magnitude) and $\hat{\mathbf{w}}_n$ (the direction) are
 1015 independent random quantities. Further, $\hat{\mathbf{w}}_n$ is uniformly distributed on \mathbf{S}^{d-1} .

1016 For $\|\mathbf{w}_n\|$, we will use sharp Gaussian concentration. The intuition is that $\|\mathbf{w}_n\|$ cannot be too
 1017 smaller than $\sqrt{d\gamma^{-1}/n}$, for large d . Concretely, it holds that

$$\text{w.p. } 1 - \delta, \left\| \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z}) (\xi_i - \bar{\xi}) \right\| \geq \sqrt{\frac{\gamma^{-1}}{n}} \cdot \left(\sqrt{d} - \sqrt{2 \ln \frac{1}{\delta}} - \sqrt{2} \right). \quad (33)$$

1018 An appropriate choice of $\delta = \exp(-d/4)$, which results in

$$\text{w.p. } 1 - \exp(-d/4), \|\mathbf{w}_n\| \gtrsim \sqrt{\frac{d\gamma^{-1}}{n}}. \quad (34)$$

1019 For the second term (with the direction $\hat{\mathbf{w}}_n$), this will be at least $\Omega(1)$ with high probability, since
 1020 $\mathbf{u}^\top \hat{\mathbf{w}}_n$ will be large only with very small probability when then dimension d is big enough. Concretely,
 1021 it holds that

$$\text{w.p. } 1 - 2 \exp(-d/4), \quad \sqrt{1 - (\mathbf{u}^\top \hat{\mathbf{w}}_n)^2} \geq \sqrt{\frac{1}{2}}. \quad (35)$$

1022 Overall, for the numerator, we conclude that

$$\text{w.p. } 1 - c \exp(-d/4), \quad \text{Numerator} \gtrsim \sqrt{\frac{d \gamma^{-1}}{n}}. \quad (36)$$

1023 **Step 3.2. For the denominator:** We need a high confidence upper bound on $\|\mathbf{S}_n\|$. We can use
 1024 Matrix-Bernstein type analysis. Note that $\mathbb{E}[\mathbf{S}_n] = \eta \mathbf{u} \tilde{\mathbf{u}}^\top$. And the deviation is dominated by

$$\mathbf{S}_n - \mathbb{E}\mathbf{S}_n \approx \frac{1}{n} \sum_{i \in [n]} z_i \xi_i \tilde{\mathbf{u}}^\top + \frac{1}{n} \sum_{i \in [n(1-\eta)]} z_i \tilde{z}_i \mathbf{u} \tilde{\mathbf{u}}^\top.$$

1025 Again, the dominating term is the first one. This means that we only have to show high confidence
 1026 upper bound on $\|(1/n) \sum_i z_i \xi_i\|$, and hence the problem has reduced to vector concentration instead
 1027 of matrix concentration. Analogous to Eq. (33), one can show

$$\text{w.p. } 1 - \delta, \quad \left\| \frac{1}{n} \sum_{i=1}^n z_i \xi_i \right\| \leq \sqrt{\frac{\gamma^{-1}}{n}} \cdot \left(\sqrt{d} + \sqrt{2 \ln \frac{1}{\delta}} \right). \quad (37)$$

1028 Overall, using the triangle inequality, we have

$$\text{w.p. } 1 - \exp(-d/4), \quad \|\mathbf{S}_n\| \leq \underbrace{\|\mathbb{E}\mathbf{S}_n\|}_{=\eta} + 2\sqrt{\frac{d \gamma^{-1}}{n}}. \quad (38)$$

1029 **Step 4. Combined result:** From 3.1 and 3.2, for $n \geq 4d \gamma^{-1} / \eta^2$ (so the high-conf UB for $\|\mathbf{S}_n\|$ is 2η),

$$\text{w.p. } 1 - O(\exp(-d/4)), \quad |\sin \theta(\mathbf{u}_n, \mathbf{u})| \gtrsim \frac{1}{\eta} \sqrt{\frac{d \gamma^{-1}}{n}}. \quad (39)$$

1030 F Characterizing the score distribution of the oracle

1031 The bernoulli variable $c \in \{0, 1\}$ captures the status of clean/corrupted nature of a sample. We
 1032 first characterize the score distribution in both cases separately, and then create the relevant mixture
 1033 distribution using the proportions $\eta, 1 - \eta$ for clean, corrupted samples respectively.

1034 Before the calculations, we state some Lemmas that will be used.

1035 **Lemma 11.** *Let X be distributed as $\mathcal{N}(0, \Omega)$. For a fixed matrix \mathbf{A} , it holds:*

$$\begin{aligned} \mathbb{E}[X^\top \mathbf{A} X] &= \text{Tr}(\mathbf{A} \Omega), \\ \mathbb{V}[X^\top \mathbf{A} X] &= \frac{1}{2} \text{Tr} \left((\mathbf{A} + \mathbf{A}^\top) \Omega (\mathbf{A} + \mathbf{A}^\top) \Omega \right). \end{aligned}$$

1036 **Lemma 12.** *Let X be distributed as $\mathcal{N}(0, \Omega)$, and \tilde{X} be distributed as $\mathcal{N}(0, \tilde{\Omega})$. Let X, \tilde{X} be
 1037 independent of each other. For a fixed matrix \mathbf{A} , it holds:*

$$\begin{aligned} \mathbb{E}[X^\top \mathbf{A} \tilde{X}] &= 0, \\ \mathbb{V}[X^\top \mathbf{A} \tilde{X}] &= \text{Tr}(\Omega \mathbf{A} \tilde{\Omega} \mathbf{A}^\top). \end{aligned}$$

1038 Consider a block matrix \mathbf{X} given as below

$$\mathbf{X} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}.$$

1039 **Lemma 13.** For a block matrix \mathbf{X} given as above, it holds that

$$\text{Tr}(\mathbf{X}) = \text{Tr}(\mathbf{A}) + \text{Tr}(\mathbf{D}) .$$

1040 **Lemma 14.** For a block matrix \mathbf{X} given as above, with \mathbf{A}, \mathbf{D} are square matrices, it holds that

$$\mathbf{X}^2 = \begin{bmatrix} \mathbf{A}^2 + \mathbf{B}\mathbf{C} & \mathbf{A}\mathbf{B} + \mathbf{B}\mathbf{D} \\ \mathbf{C}\mathbf{A} + \mathbf{D}\mathbf{C} & \mathbf{C}\mathbf{B} + \mathbf{D}^2 \end{bmatrix} .$$

1041 **Case 0: Corrupted samples** ($c = 0$ case). Let $Z_0 \stackrel{d}{=} \{S(x, \tilde{x}; \mathbf{U}\tilde{\mathbf{U}}^\top) \mid c = 0\}$, with distribution \mathcal{D}_0 .
 1042 This (scalar) random variable is equivalent to $X^\top \mathbf{U}\tilde{\mathbf{U}}^\top \tilde{X}$, where X, \tilde{X} are independent and follow
 1043 $X \sim \mathcal{N}(0, \mathbf{U}\mathbf{U}^\top + \gamma^{-1} \mathbf{I}_d)$, $\tilde{X} \sim \mathcal{N}(0, \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top + \tilde{\gamma}^{-1} \mathbf{I}_{\tilde{d}})$. This is in-line with Remark A.1. We
 1044 invoke Lemma 12 to get the first two moments.

1045 1. Mean: 0.

1046 2. Variance: $r(1 + \gamma^{-1})(1 + \tilde{\gamma}^{-1})$.

$$\begin{aligned} \text{Variance} &= \text{Tr}\left((\mathbf{U}\mathbf{U}^\top + \gamma^{-1} \mathbf{I}_d) \mathbf{U}\tilde{\mathbf{U}}^\top (\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top + \tilde{\gamma}^{-1} \mathbf{I}_{\tilde{d}}) \tilde{\mathbf{U}}\mathbf{U}^\top\right) \\ &= \text{Tr}\left(\mathbf{U}^\top (\mathbf{U}\mathbf{U}^\top + \gamma^{-1} \mathbf{I}_d) \mathbf{U} \tilde{\mathbf{U}}^\top (\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top + \tilde{\gamma}^{-1} \mathbf{I}_{\tilde{d}}) \tilde{\mathbf{U}}\right) \\ &= \text{Tr}\left((\mathbf{I}_r + \gamma^{-1} \mathbf{I}_r) (\mathbf{I}_r + \tilde{\gamma}^{-1} \mathbf{I}_r)\right) . \end{aligned}$$

1047 3. Tails: Since X, \tilde{X} are independent, the tails are described by the quadratic form on two
 1048 independent Gaussians. This random variable is (i) symmetric, and (ii) uni-modal, and the
 1049 tails decay exponentially.

1050 **Case 1: Clean samples** ($c = 1$ case). Let $Z_1 \stackrel{d}{=} \{S(x, \tilde{x}; \mathbf{U}\tilde{\mathbf{U}}^\top) \mid c = 1\}$, with distribution \mathcal{D}_1 .
 1051 This random variable is equivalent to $X^\top \mathbf{B}X$, where $X = [x, \tilde{x}]^\top$ follows $X \sim \mathcal{N}(0, \Sigma_1)$ (refer
 1052 to Remark A.1); and \mathbf{B} is a block matrix given as below. We invoke Lemma 11 to get the first two
 1053 moments.

$$\mathbf{B} = \begin{bmatrix} \mathbf{0}_{d \times d} & \mathbf{U}\tilde{\mathbf{U}}^\top \\ \mathbf{0}_{\tilde{d} \times d} & \mathbf{0}_{\tilde{d} \times \tilde{d}} \end{bmatrix}_{(d+\tilde{d}) \times (d+\tilde{d})}$$

1054 1. Mean: r .

$$\begin{aligned} \text{Mean} &= \text{Tr}(\mathbf{B}\Sigma_1) \\ &= \text{Tr}\left(\begin{bmatrix} \mathbf{U}\mathbf{U}^\top & \\ & \mathbf{0} \end{bmatrix}\right) \\ &= \text{Tr}(\mathbf{U}\mathbf{U}^\top) = \text{Tr}(\mathbf{I}_r) = r . \end{aligned} \quad (\text{Using Lemma 13})$$

1055 2. Variance: $r + r(1 + \gamma^{-1})(1 + \tilde{\gamma}^{-1})$.

$$\begin{aligned} \text{Variance} &= \frac{1}{2} \text{Tr}\left((\mathbf{B} + \mathbf{B}^\top) \Sigma_1 (\mathbf{B} + \mathbf{B}^\top) \Sigma_1\right) \\ &= \frac{1}{2} \text{Tr}\left(\begin{bmatrix} \mathbf{U}\mathbf{U}^\top & \overbrace{\mathbf{U}\tilde{\mathbf{U}}^\top + \tilde{\gamma}^{-1} \mathbf{U}\tilde{\mathbf{U}}^\top}^{\mathbf{T}_1} \\ \underbrace{\tilde{\mathbf{U}}\mathbf{U}^\top + \gamma^{-1} \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top}_{\mathbf{T}_2} & \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top \end{bmatrix}\right)^2 \\ &= \frac{1}{2} \text{Tr}\left(\begin{bmatrix} \mathbf{U}\mathbf{U}^\top + \mathbf{T}_1 \mathbf{T}_2 & \\ & \mathbf{T}_2 \mathbf{T}_1 + \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top \end{bmatrix}\right) \quad (\text{Using Lemma 14}) \\ &= \text{Tr}(\mathbf{I}_r) + \text{Tr}(\mathbf{T}_1 \mathbf{T}_2) . \quad (\text{Using Lemma 13}) \end{aligned}$$

1056 3. **Tails:** Since X, \tilde{X} are dependent, the tails are described by the quadratic form on two
 1057 dependent Gaussians. The tails decay exponentially, and are described by the Hanson-
 1058 Wright inequality. A similar calculation as the variance provides the exact parameters, and
 1059 the inequality becomes:

$$\mathbb{P}(|Z_1 - \mathbb{E}Z_1| > t) \lesssim \exp \left(-c \min \left\{ \frac{2t^2}{r(1 + (1 + \gamma^{-1})(1 + \tilde{\gamma}^{-1}))}, \frac{\sqrt{2}t}{\sqrt{r(1 + (1 + \gamma^{-1})(1 + \tilde{\gamma}^{-1}))}} \right\} \right). \quad (40)$$

1060 G A proof of Theorem 1

1061 In this section, we present a proof of Theorem 1. We first define some additional notation. For the
 1062 orthonormal matrix \mathbf{U} , let $\mathbf{U}_\perp \in \mathbb{R}^{d \times (d-r)}$ denote the completion of the orthonormal basis. That
 1063 is, the matrix $\mathbf{U}_{\text{full}} = [\mathbf{U} \ \mathbf{U}_\perp] \in \mathbb{R}^{d \times d}$ is such that $\mathbf{U}_{\text{full}}^\top \mathbf{U}_{\text{full}} = \mathbf{I}_d = \mathbf{U}_{\text{full}} \mathbf{U}_{\text{full}}^\top$. Similarly define
 1064 $\tilde{\mathbf{U}}_\perp \in \mathbb{R}^{\tilde{d} \times (\tilde{d}-r)}$.

1065 Recall that we have n samples of the form $\{(x_i, \tilde{x}_i)\}_{i=1}^n$, i.i.d from the mixture distribution (with
 1066 $\eta, 1 - \eta$ ratios for clean, corrupted respectively). Let n_T samples be used to train the teacher, and
 1067 let $N = n_T - n$ samples be used to train the student. Let ρ_T, ρ be the respective regularization
 1068 parameters, and let $(\mathbf{G}_T, \tilde{\mathbf{G}}_T), (\mathbf{G}, \tilde{\mathbf{G}})$ denote the respective embedding matrices at the solution of
 1069 Eq. (2). Consider a general threshold $\theta \in \mathbb{R}$ that is used to filter the dataset based on the teacher
 1070 scores. Note that we have ensured that θ is independent of the N samples to be filtered, since it
 1071 depends only on the n_T samples used for teacher training. For the teacher, from Corollary 1, we know
 1072 that with probability $1 - \exp(-\Omega(\max\{d, \tilde{d}\}))$:

$$\left\| \mathbf{G}_T^\top \tilde{\mathbf{G}}_T - \frac{\eta}{\rho_T} \mathbf{U} \tilde{\mathbf{U}}^\top \right\| \leq \frac{1}{\rho_T} \sqrt{\frac{\max\{d, \tilde{d}\} \text{poly}(\gamma^{-1}, \tilde{\gamma}^{-1})}{n_T}}. \quad (41)$$

1073 Here $(\mathbf{G}_T, \tilde{\mathbf{G}}_T)$ are random quantities that depend on the n_T samples used. For the rest of the analysis,
 1074 we will assume them to be fixed (since they don't depend on the randomness of the remaining N
 1075 samples). Finally, we will give a high probability guarantee that will use the confidence bound in
 1076 Eq. (41) as one of the terms in the combined error bound.

1077 We now study the student with data filtering. Define $\mathbf{M}_T := \mathbf{G}_T^\top \tilde{\mathbf{G}}_T$, the matrix used for scoring
 1078 the samples. Also denote by $\mathbf{M}_O := (\eta/\rho_T) \mathbf{U} \tilde{\mathbf{U}}^\top$ the oracle scoring matrix (note the scaling factor).
 1079 From the teacher guarantee, it holds that $\mathbf{M}_T \rightarrow \mathbf{M}_O$ as $n_T \rightarrow \infty$. Recall that the scoring function is
 1080 $S(x, \tilde{x}; \mathbf{M}) = x^\top \mathbf{M} \tilde{x}$, and a sample (x, \tilde{x}) is selected/retained iff $S(x, \tilde{x}; \mathbf{M}_T) > \theta$.

1081 We define certain quantities that will be central to the analysis. Akin to Eq. (4), we define the
 1082 empirical cross-covariance matrix of the data *after selection* in Eq. (42). Let $n_{\text{sel}, T}(\theta)$ be the number
 1083 of samples selected, which is a random variable with $\mathbb{E}[n_{\text{sel}, T}(\theta)] = N P_T(\theta)$. Let $I_{\text{sel}, T}(\theta) \subseteq [N]$
 1084 denote the indices of the points selected. That is, $i \in I_{\text{sel}, T}(\theta) \iff S(x_i, \tilde{x}_i; \mathbf{M}_T) > \theta$. Similarly,
 1085 define $n_{\text{sel}, O}(\theta)$ and $I_{\text{sel}, O}(\theta)$. Construct the empirical cross-covariance matrix for the filtered dataset:

$$\mathbf{S}_{N, T}(\theta) := \frac{1}{n_{\text{sel}, T}(\theta) - 1} \underbrace{\sum_{i \in I_{\text{sel}, T}(\theta)} (x_i - \bar{x}(\theta)) (\tilde{x}_i - \bar{\tilde{x}}(\theta))^\top}_{\mathbf{Q}_{N, T}(\theta)}. \quad (42)$$

1086 To analyze its asymptotic limit, we define $\mathbf{S}(\theta)$ as the limit of the cross-covariance, for both the
 1087 teacher and the oracle. Similarly, let $P(\theta)$ denote the probability mass of data that is retained (also in
 1088 the limit of $n \rightarrow \infty$), for both the teacher and the oracle. These are described in Eqs (43), (44).

$$\mathbf{S}_T(\theta) = \mathbb{E} [x \tilde{x}^\top \mid S(x, \tilde{x}; \mathbf{M}_T) > \theta] \in \mathbb{R}^{d \times \tilde{d}}, \quad P_T(\theta) = \mathbb{P} \{S(x, \tilde{x}; \mathbf{M}_T) > \theta\}; \quad (43)$$

$$\mathbf{S}_O(\theta) = \mathbb{E} [x \tilde{x}^\top \mid S(x, \tilde{x}; \mathbf{M}_O) > \theta] \in \mathbb{R}^{d \times \tilde{d}}, \quad P_O(\theta) = \mathbb{P} \{S(x, \tilde{x}; \mathbf{M}_O) > \theta\}. \quad (44)$$

1089 Note that $\mathbf{S}_T(\theta), \mathbf{S}_O(\theta)$ are the limits of $\mathbf{S}_{N,T}(\theta), \mathbf{S}_{N,O}(\theta)$ as $N \rightarrow \infty$. The threshold $\theta \rightarrow -\infty$
 1090 recovers the no filtering case, i.e. both $\mathbf{S}_{N,T}(\theta), \mathbf{S}_{N,O}(\theta)$ approach \mathbf{S}_N . We will now follow proof
 1091 steps similar to Section D. Steps 1 and 2 hold for a general cross covariance matrix, and can be used
 1092 directly. Steps 3 and 4 are concerned with the limit of $\mathbf{S}_n(\theta)$ as $n \rightarrow \infty$, and how it concentrates
 1093 around the limit. These steps will change significantly. Finally, we will be able to reuse Lemma 2 for
 1094 step 5. We detail each of these proof steps below.

1095 **Step 1.** The unregularized contrastive loss objective on the $n_{\text{sel},T}(\theta)$ samples is equivalent to

$$\mathcal{L}_0(\mathbf{G}, \tilde{\mathbf{G}}) = -\text{Tr} \left(\mathbf{G} \mathbf{S}_{N,T}(\theta) \tilde{\mathbf{G}}^\top \right). \quad (45)$$

1096 This follows the exact same proof steps as in Section D.

1097 **Step 2.** The solution to the ρ -regularized minimization problem is given by

$$\arg \min_{\mathbf{G}, \tilde{\mathbf{G}}} \mathcal{L}_\rho(\mathbf{G}, \tilde{\mathbf{G}}) = \left\{ (\mathbf{G}, \tilde{\mathbf{G}}) \mid \mathbf{G}^\top \tilde{\mathbf{G}} = \frac{1}{\rho} \text{SVD}_r(\mathbf{S}_{N,T}(\theta)) \right\}. \quad (46)$$

1098 This also follows the exact same proof steps as in Section D.

1099 **Step 3.** This step changes from Section D. We use the following:

$$\|\text{SVD}_r(\mathbf{S}_{N,T}(\theta)) - \mathbf{S}_O(\theta)\| \leq \sigma_{r+1}(\mathbf{S}_O(\theta)) + 2\|\mathbf{S}_{N,T}(\theta) - \mathbf{S}_O(\theta)\|. \quad (47)$$

1100 By triangle inequality, we have

$$\|\text{SVD}_r(\mathbf{S}_{N,T}(\theta)) - \mathbf{S}_O(\theta)\| \leq \|\text{SVD}_r(\mathbf{S}_{N,T}(\theta)) - \mathbf{S}_{N,T}(\theta)\| + \|\mathbf{S}_{N,T}(\theta) - \mathbf{S}_O(\theta)\|.$$

1101 And for the first term on the right hand side, we use

$$\begin{aligned} \|\text{SVD}_r(\mathbf{S}_{N,T}(\theta)) - \mathbf{S}_{N,T}(\theta)\| &= \sigma_{r+1}(\mathbf{S}_{N,T}(\theta)) \\ &\leq^{(\dagger)} \sigma_{r+1}(\mathbf{S}_O(\theta)) + \|\mathbf{S}_{N,T}(\theta) - \mathbf{S}_O(\theta)\|, \end{aligned}$$

1102 where we used Lemma 1 in Eq (†).

1103 **Step 3'.** Analysis of $\mathbf{S}_O(\theta)$: The main difference in Eq. (19) and Eq. (47) is the term $\sigma_{r+1}(\mathbf{S}_O(\theta))$.
 1104 This additional step of the proof analyzes the properties of $\mathbf{S}_O(\theta)$. In particular, we will show that
 1105 $\mathbf{S}_O(\theta)$ is rank- r , and hence $\sigma_{r+1}(\mathbf{S}_O(\theta)) = 0$. Additionally, we establish upper and lower bounds on
 1106 the singular values of $\mathbf{S}_O(\theta)$ that will be used later in the proof. From Eq. (44), we simplify to write

$$\mathbf{S}_O(\theta) = \mathbb{E} \left[x \tilde{x}^\top \mid x^\top \mathbf{U} \tilde{\mathbf{U}}^\top \tilde{x} > \frac{\theta \rho_T}{\eta} \right],$$

1107 where (x, \tilde{x}) is drawn from the mixture model: $\eta \cdot \mathcal{N}(0, \Sigma_1) + (1 - \eta) \cdot \mathcal{N}(0, \Sigma_0)$. To simplify
 1108 notation, define $\tilde{\theta} := (\theta \rho_T)/\eta$. From the conditioning event, it seems that $\mathbf{U}^\top x$ and $\tilde{\mathbf{U}}^\top \tilde{x}$ is a good
 1109 ‘basis’ for a decomposition. Pre-multiply and post-multiply to recover this basis for the $x \tilde{x}^\top$ term
 1110 inside the expectation as

$$\begin{aligned} \mathbf{S}_O(\theta) &= \underbrace{\mathbf{U}_{\text{full}} \mathbf{U}_{\text{full}}^\top}_{=\mathbf{I}_d} \mathbb{E} \left[x \tilde{x}^\top \mid x^\top \mathbf{U} \tilde{\mathbf{U}}^\top \tilde{x} > \tilde{\theta} \right] \underbrace{\tilde{\mathbf{U}}_{\text{full}} \tilde{\mathbf{U}}_{\text{full}}^\top}_{=\mathbf{I}_{\tilde{d}}} \\ &= \mathbf{U}_{\text{full}} \mathbb{E} \left[\begin{pmatrix} \overbrace{(\mathbf{U}^\top x)(\tilde{\mathbf{U}}^\top \tilde{x})^\top}^{r \times r} & \overbrace{(\mathbf{U}^\top x)(\tilde{\mathbf{U}}_\perp^\top \tilde{x})^\top}^{r \times (\tilde{d}-r)} \\ \underbrace{(\mathbf{U}_\perp^\top x)(\tilde{\mathbf{U}}^\top \tilde{x})^\top}_{(d-r) \times r} & \underbrace{(\mathbf{U}_\perp^\top x)(\tilde{\mathbf{U}}_\perp^\top \tilde{x})^\top}_{(d-r) \times (\tilde{d}-r)} \end{pmatrix} \mid (\mathbf{U}^\top x)^\top (\tilde{\mathbf{U}}^\top \tilde{x}) > \tilde{\theta} \right] \tilde{\mathbf{U}}_{\text{full}}^\top. \end{aligned}$$

1111 Call the top left entry in this decomposition to be the ‘dominant’, and the other three as ‘non-
 1112 dominant’. We will show the non-dominant entries will be zero. The following reparametrization
 1113 makes things cleaner.

$$\mathbf{U}^\top x = z + \underbrace{\mathbf{U}^\top \xi}_{\varepsilon}, \quad \mathbf{U}_\perp^\top x = \underbrace{\mathbf{U}_\perp^\top \xi}_{\varepsilon_\perp}; \quad \tilde{\mathbf{U}}^\top \tilde{x} = \tilde{z} + \underbrace{\tilde{\mathbf{U}}^\top \tilde{\xi}}_{\tilde{\varepsilon}}, \quad \tilde{\mathbf{U}}_\perp^\top \tilde{x} = \underbrace{\tilde{\mathbf{U}}_\perp^\top \tilde{\xi}}_{\tilde{\varepsilon}_\perp}.$$

1114 Let's further simplify the expressions with another transformation. The subscripts S, N denote the
1115 signal (containing some noise) and noise part.

$$\underbrace{x_S}_{\in \mathbb{R}^r} \leftarrow z + \varepsilon, \quad \underbrace{x_N}_{\in \mathbb{R}^{d-r}} \leftarrow \varepsilon_\perp ; \quad \underbrace{\tilde{x}_S}_{\in \mathbb{R}^r} \leftarrow \tilde{z} + \tilde{\varepsilon}, \quad \underbrace{\tilde{x}_N}_{\in \mathbb{R}^{\tilde{d}-r}} \leftarrow \tilde{\varepsilon}_\perp .$$

1116 Due to the diagonal structure of $\Sigma_\xi, \Sigma_{\tilde{\xi}}$, we infer the distributions as

$$\varepsilon \sim \mathcal{N}\left(0, \frac{1}{\gamma} \mathbf{I}_r\right), \quad \varepsilon_\perp \sim \mathcal{N}\left(0, \frac{1}{\gamma} \mathbf{I}_{(d-r)}\right); \quad \tilde{\varepsilon} \sim \mathcal{N}\left(0, \frac{1}{\tilde{\gamma}} \mathbf{I}_r\right), \quad \tilde{\varepsilon}_\perp \sim \mathcal{N}\left(0, \frac{1}{\tilde{\gamma}} \mathbf{I}_{(\tilde{d}-r)}\right).$$

1117 And crucially, due to the diagonal structure of $\Sigma_\xi, \Sigma_{\tilde{\xi}}$, we infer that $\{\varepsilon, \varepsilon_\perp, \tilde{\varepsilon}, \tilde{\varepsilon}_\perp\}$ are all *mutually*
1118 *independent*, and independent of z, \tilde{z} . This entails that the transformed vector is Gaussian with mean
1119 zero and covariance given as below.

$$\begin{pmatrix} x_S \\ x_N \\ \tilde{x}_S \\ \tilde{x}_N \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} (1 + 1/\gamma) \mathbf{I}_r & \mathbf{0} & \mathbf{0} (\mathbf{I}_r) & \mathbf{0} \\ \cdot & (1/\gamma) \mathbf{I}_{(d-r)} & \mathbf{0} & \mathbf{0} \\ \cdot (\cdot) & \cdot & (1 + 1/\tilde{\gamma}) \mathbf{I}_r & \mathbf{0} \\ \cdot & \cdot & \cdot & (1/\tilde{\gamma}) \mathbf{I}_{(\tilde{d}-r)} \end{pmatrix}\right). \quad (48)$$

1120 The above is for the corrupted case (w.p. $1 - \eta$). In the clean case (w.p. η), the blue entries change to
1121 \mathbf{I}_r due to the relation of $z = \tilde{z}$. Our $\mathbb{E}[\cdot]$ notation includes the expectation over this randomness along
1122 with the randomness of x, \tilde{x} . Denote by Ω_0 and Ω_1 the covariances of the signal part, i.e. (x_S, \tilde{x}_S) in
1123 these two cases:

$$\Omega_0 := \begin{pmatrix} (1 + 1/\gamma) \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & (1 + 1/\tilde{\gamma}) \mathbf{I}_r \end{pmatrix}, \quad \Omega_1 := \begin{pmatrix} (1 + 1/\gamma) \mathbf{I}_r & \mathbf{I}_r \\ \mathbf{I}_r & (1 + 1/\tilde{\gamma}) \mathbf{I}_r \end{pmatrix}. \quad (49)$$

1124 Overall, under the transformation, the expectation simplifies to

$$\mathbf{S}_O(\theta) = \mathbf{U}_{\text{full}} \mathbb{E} \left[\begin{pmatrix} x_S \tilde{x}_S^\top & x_S \tilde{x}_N^\top \\ x_N \tilde{x}_S^\top & x_N \tilde{x}_N^\top \end{pmatrix} \middle| x_S^\top \tilde{x}_S > \tilde{\theta} \right] \tilde{\mathbf{U}}_{\text{full}}^\top. \quad (50)$$

1125 Due to x_N, \tilde{x}_N being independent of all other entries via Eq. (48), and since the conditioning event
1126 in Eq. (50) only involves x_S, \tilde{x}_S , we conclude that the non-dominant entries in the expectation will
1127 be *zero*. Hence we are left with the simplified rank- r form for the $d \times \tilde{d}$ matrix:

$$\begin{aligned} \mathbf{S}_O(\theta) &= \mathbf{U} \mathbb{E} \left[x_S \tilde{x}_S^\top \middle| x_S^\top \tilde{x}_S > \tilde{\theta} \right] \tilde{\mathbf{U}}^\top = \mathbf{U} \left(\eta \cdot \mathbb{E}_{(x_S, \tilde{x}_S) \sim \mathcal{N}(0, \Omega_1)} \left[x_S \tilde{x}_S^\top \middle| x_S^\top \tilde{x}_S > \tilde{\theta} \right] \right. \\ &\quad \left. + (1 - \eta) \cdot \mathbb{E}_{(x_S, \tilde{x}_S) \sim \mathcal{N}(0, \Omega_0)} \left[x_S \tilde{x}_S^\top \middle| x_S^\top \tilde{x}_S > \tilde{\theta} \right] \right) \tilde{\mathbf{U}}^\top. \end{aligned}$$

1128 We will now use Lemma 8 to simplify both the terms above. Note that Ω_1, Ω_0 satisfy the lemma's
1129 requirement of the block diagonal covariance.

$$\mathbf{S}_O(\theta) = \mathbf{U} \left(\eta f_1(\theta) \mathbf{I}_r + (1 - \eta) f_0(\theta) \mathbf{I}_r \right) \tilde{\mathbf{U}}^\top = (\eta f_1(\theta) + (1 - \eta) f_0(\theta)) \mathbf{U} \tilde{\mathbf{U}}^\top, \quad (51)$$

1130 where the following conditions hold on f_1, f_0 (converting back from $\tilde{\theta}$ to θ):

$$\begin{aligned} \max\{1, (\theta \rho_T)/\eta r\} + e \sqrt{((1 + \gamma^{-1})(1 + \tilde{\gamma}^{-1}) + 1)/r} &\geq f_1(\theta) \geq \max\{1, (\theta \rho_T)/\eta r\}, \\ \max\{0, (\theta \rho_T)/\eta r\} + e \sqrt{((1 + \gamma^{-1})(1 + \tilde{\gamma}^{-1}))/r} &\geq f_0(\theta) \geq \max\{0, (\theta \rho_T)/\eta r\}. \end{aligned}$$

1131 Using the above equations, and the special case of $\theta = 0$ in Lemma 8, we conclude:

$$f_1(0) \geq 1, \quad f_0(0) \geq \frac{2}{\pi r} \cdot \sqrt{(1 + \gamma^{-1})(1 + \tilde{\gamma}^{-1})}, \quad (52)$$

$$f_1\left(\frac{r\eta}{2\rho_T}\right) \geq 1, \quad f_0\left(\frac{r\eta}{2\rho_T}\right) \geq \frac{1}{2}. \quad (53)$$

1132 We will use these inequalities in step 5.

1133 **Step 4.** Concentration of $\mathbf{S}_{N,T}(\theta)$ to $\mathbf{S}_O(\theta)$: We break this into subparts as below.

1134 **Step 4.1.** Concentration of $\mathbf{S}_{N,T}(\theta)$ to $\mathbf{S}_T(\theta)$: Using the below substeps, we show that with probability
 1135 $1 - \exp(-\Omega(\max\{d, \tilde{d}\}))$:

$$\|\mathbf{S}_{N,T}(\theta) - \mathbf{S}_T(\theta)\| \leq \sqrt{\frac{\max\{d, \tilde{d}\} \text{poly}(\gamma^{-1}, \tilde{\gamma}^{-1})}{N P_T(\theta)}} + \tilde{O}\left(\frac{1}{N P_T(\theta)}\right). \quad (54)$$

1136 **Step 4.1.A.** Replacing the random denominator: We follow an argument similar to Eqs. (26)-(27).
 1137 The application of Lemma 10 is on the joint vector $[x, \tilde{x}]$, and we apply the lemma on both the
 1138 Gaussian components of the mixture. The failure probability is $\delta = \exp(-\max\{d, \tilde{d}\})$ and the
 1139 selection probability is $p = P_T(\theta)$. This allows us to deal with the deterministic quantity $N P_T(\theta)$ in
 1140 the denominator instead of the random $n_{\text{sel},T}(\theta)$. Here we use $\mathbf{L}(\theta)$ to denote the asymptotic limit of
 1141 $\frac{1}{N P_T(\theta) - 1} \mathbf{Q}_{N,T}(\theta)$, which we characterize in Step 4.1.C. With probability $1 - \exp(-\Omega(\max\{d, \tilde{d}\}))$:

$$\left\| \frac{1}{n_{\text{sel},T}(\theta) - 1} \mathbf{Q}_{N,T}(\theta) - \mathbf{L}(\theta) \right\| \leq \left\| \frac{1}{N P_T(\theta) - 1} \mathbf{Q}_{N,T}(\theta) - \mathbf{L}(\theta) \right\| + \sqrt{\frac{\max\{d, \tilde{d}\}}{N P_T(\theta)}}.$$

1142 **Step 4.1.B.** The centered vs un-centered version: We have that

$$\begin{aligned} \frac{1}{N P_T(\theta) - 1} \sum_{i \in I_{\text{sel},T}(\theta)} (x_i - \bar{x}(\theta)) (\tilde{x}_i - \bar{\tilde{x}}(\theta))^\top &= \\ \frac{1}{N P_T(\theta)} \sum_{i \in I_{\text{sel},T}(\theta)} x_i \tilde{x}_i^\top - \frac{1}{N P_T(\theta) (N P_T(\theta) - 1)} \sum_{i \in I_{\text{sel},T}(\theta)} \sum_{\substack{j \in I_{\text{sel},T}(\theta) \\ j \neq i}} x_i \tilde{x}_j^\top. \end{aligned}$$

1143 The second term on the right hand side concentrates to $\mathbb{E}[x \tilde{y}^\top \mid x^\top \mathbf{M}_T \tilde{x} > \theta, y^\top \mathbf{M}_T \tilde{y} > \theta]$, where
 1144 (x, \tilde{x}) and (y, \tilde{y}) are i.i.d. from the joint mixture distribution. This expectation is zero, which we
 1145 formally characterize in Lemmas 5 and 6. The rate of concentration is $\tilde{O}\left(\frac{1}{N P_T(\theta)}\right)$, due to averaging
 1146 over $(N P_T(\theta))^2$ terms, and is hence a higher order term.

1147 **Step 4.1.C.** Analysis of the fixed-denominator un-centered version: The selected samples satisfy
 1148 the property of being *i.i.d from the conditional law* of the selection rule. Using a Matrix-Bernstein
 1149 concentration result (Eqs. (23) and (24)), it follows that with probability $1 - \exp(-\Omega(\max\{d, \tilde{d}\}))$:

$$\left\| \frac{1}{N P_T(\theta)} \sum_{i \in I_{\text{sel},T}(\theta)} x_i \tilde{x}_i^\top - \mathbf{S}_T(\theta) \right\| \lesssim \sqrt{\frac{\max\{d, \tilde{d}\} \text{poly}(\gamma^{-1}, \tilde{\gamma}^{-1})}{N P_T(\theta)}}.$$

1150 **Step 4.2.** Error between teacher and oracle: We show that $\|\mathbf{S}_T(\theta) - \mathbf{S}_O(\theta)\|$ scales proportionally to
 1151 $\|\mathbf{M}_T - \mathbf{M}_O\|$, and the latter is precisely bounded by Eq. (41). To simplify the conditional expectation
 1152 in $\mathbf{S}_O(\theta)$, $\mathbf{S}_T(\theta)$, define $\mathbf{E}_O(\theta)$, $\mathbf{E}_T(\theta)$ as:

$$\mathbf{E}_O(\theta) := \mathbb{E}[x \tilde{x}^\top \mathbb{I}(x^\top \mathbf{M}_O \tilde{x} > \theta)] \iff \mathbf{S}_O(\theta) = \mathbf{E}_O(\theta) / P_O(\theta); \quad (55)$$

$$\mathbf{E}_T(\theta) := \mathbb{E}[x \tilde{x}^\top \mathbb{I}(x^\top \mathbf{M}_T \tilde{x} > \theta)] \iff \mathbf{S}_T(\theta) = \mathbf{E}_T(\theta) / P_T(\theta). \quad (56)$$

1153 where $\mathbb{I}(\cdot)$ denotes the indicator. Let $\Delta \mathbf{E}(\theta) := \mathbf{E}_T(\theta) - \mathbf{E}_O(\theta)$ and $\Delta P(\theta) := P_T(\theta) - P_O(\theta)$. Also
 1154 define $\Delta \mathbb{I}(\theta; x, \tilde{x}) := \mathbb{I}(x^\top \mathbf{M}_T \tilde{x} > \theta) - \mathbb{I}(x^\top \mathbf{M}_O \tilde{x} > \theta)$. Then, we write

$$\begin{aligned} \mathbf{S}_T(\theta) - \mathbf{S}_O(\theta) &= \frac{\mathbf{E}_T(\theta)}{P_T(\theta)} - \frac{\mathbf{E}_O(\theta)}{P_O(\theta)} \\ &= \frac{(\mathbf{E}_O(\theta) + \Delta \mathbf{E}(\theta)) P_O(\theta) - \mathbf{E}_O(\theta) (P_O(\theta) + \Delta P(\theta))}{P_T(\theta) P_O(\theta)} \\ &= \frac{\Delta \mathbf{E}(\theta)}{P_T(\theta)} - \frac{\Delta P(\theta)}{P_T(\theta)} \cdot \underbrace{\frac{\mathbf{E}_O(\theta)}{P_O(\theta)}}_{\mathbf{S}_O(\theta)}. \end{aligned}$$

$$\implies \|\mathbf{S}_T(\theta) - \mathbf{S}_O(\theta)\|_2 \leq \frac{1}{P_T(\theta)} (\|\Delta \mathbf{E}(\theta)\|_2 + |\Delta P(\theta)| \cdot \|\mathbf{S}_O(\theta)\|_2) .$$

1155 We will now bound $\|\Delta \mathbf{E}(\theta)\|_2$ and $|\Delta P(\theta)|$ in terms of $\|\mathbf{M}_T - \mathbf{M}_O\|_2$. Recall that (x, \tilde{x}) follow the
 1156 mixture distribution (Remark [A.1](#)). Decomposing the expectations and probabilities into respective
 1157 mixtures, we get

$$\begin{aligned} \Delta \mathbf{E}(\theta) &= \eta \mathbb{E}_{(x, \tilde{x}) \sim \mathcal{N}(0, \Sigma_1)} [x \tilde{x}^\top \Delta \mathbb{I}(\theta; x, \tilde{x})] + (1 - \eta) \mathbb{E}_{(x, \tilde{x}) \sim \mathcal{N}(0, \Sigma_0)} [x \tilde{x}^\top \Delta \mathbb{I}(\theta; x, \tilde{x})] , \\ \Delta P(\theta) &= \eta \mathbb{E}_{(x, \tilde{x}) \sim \mathcal{N}(0, \Sigma_1)} [\Delta \mathbb{I}(\theta; x, \tilde{x})] + (1 - \eta) \mathbb{E}_{(x, \tilde{x}) \sim \mathcal{N}(0, \Sigma_0)} [\Delta \mathbb{I}(\theta; x, \tilde{x})] . \end{aligned}$$

1158 From the above, since both $\eta, 1 - \eta$ are smaller than 1, we get that

$$\|\Delta \mathbf{E}(\theta)\|_2 \leq \|\Delta \mathbf{E}_1(\theta)\|_2 + \|\Delta \mathbf{E}_0(\theta)\|_2 , \quad |\Delta P(\theta)| \leq |\Delta P_1(\theta)| + |\Delta P_0(\theta)| ,$$

1159 where the subscripts 1, 0 denote the fully clean, corrupted cases respectively (i.e. $\eta = 1, \eta = 0$
 1160 respectively). Lemma [9](#) captures the general form of this, and we invoke this lemma on both the
 1161 clean data (with covariance Σ_1) and the noisy data (with covariance Σ_0). We get

$$\|\mathbf{S}_T(\theta) - \mathbf{S}_O(\theta)\|_2 \lesssim \frac{\|\mathbf{S}_O(\theta)\|_2}{P_T(\theta)} \|\mathbf{M}_T - \mathbf{M}_O\|_2 . \quad (57)$$

1162 **Step 4.3.** Analysis of $P_T(\theta)$ and $P_O(\theta)$: We first characterize the relationship between $P_T(\theta)$ and
 1163 $P_O(\theta)$. Using step 4.2, we have $P_T(\theta) \geq P_O(\theta) - |\Delta P(\theta)|$. In step 5 we will show that the latter
 1164 term will be small, and we will be able to use, for instance, $P_T(\theta) \geq (1/2)P_O(\theta)$.

1165 Next, we show that $P_O(\theta)$ is ‘large enough’ for the choices of $\theta \in \{0, r\eta/2\rho_T\}$, and we will use these
 1166 fixed points in Step 5. Recall from Section [6.2](#) due to the mixture distribution, the below holds. Here
 1167 we have accounted for the scaling factor in the definition of \mathbf{M}_O .

$$P_O(\theta) = \eta P_1\left(\frac{\theta\rho_T}{\eta}\right) + (1 - \eta) P_0\left(\frac{\theta\rho_T}{\eta}\right) . \quad (58)$$

1168 In step 5, we will consider the fixed points $\theta \in \{0, r\eta/2\rho_T\}$, and so we need lower bounds on
 1169 $P_0(0), P_0(r/2)$ and $P_1(0), P_1(r/2)$. We state them below:

$$P_0(0) \geq 0.5 , \quad P_1(0) \geq 1 - 1/e , \quad (59)$$

$$P_0(r/2) \geq 0 , \quad P_1(r/2) \geq 1 - 1/e . \quad (60)$$

1170 For $P_0(\cdot)$, we have lower bounds 0.5 (due to symmetry) and 0 (trivially). For $P_1(\cdot)$, we characterize
 1171 the tails in Eq. [\(40\)](#), which hold for $r \gtrsim 1 + (1 + \gamma^{-1})(1 + \tilde{\gamma}^{-1})$.

1172 **Step 5.** Final guarantee via application of Lemma [2](#): Using Eqs. [\(54\)](#) and [\(57\)](#) in Eq. [\(47\)](#) with
 1173 Eq. [\(46\)](#), and combining the guarantee from Eq. [\(41\)](#), with probability $1 - \exp(-\Omega(\max\{d, \tilde{d}\}))$:

$$\begin{aligned} \left\| \mathbf{G}^\top \tilde{\mathbf{G}} - \frac{1}{\rho} \mathbf{S}_O(\theta) \right\| &\lesssim \frac{1}{\rho} \left(\sqrt{\frac{\max\{d, \tilde{d}\} \text{poly}(\gamma^{-1}, \tilde{\gamma}^{-1})}{N P_T(\theta)}} \right. \\ &\quad \left. + \frac{\|\mathbf{S}_O(\theta)\|_2}{\rho_T P_T(\theta)} \sqrt{\frac{\max\{d, \tilde{d}\} \text{poly}(\gamma^{-1}, \tilde{\gamma}^{-1})}{n_T}} + \tilde{O}\left(\frac{1}{N P_T(\theta)}\right) \right) . \end{aligned}$$

1174 We set $n_T = n/2$, and so $N = n - n_T = n/2$ (as in Algorithm [1](#)). For ρ_T , we note that it can be
 1175 chosen arbitrarily large to reduce the second term in the error above. This is because any $\rho_T > 0$ will
 1176 allow the teacher parameters $\mathbf{G}_T, \tilde{\mathbf{G}}_T$ to recover the subspace spanned by $\mathbf{U}, \tilde{\mathbf{U}}$ respectively, but a
 1177 large choice of ρ_T will make the operator norm small. This does not cause the filtering to change,
 1178 since the threshold θ changes multiplicatively with ρ_T (effectively scaling the picture in Figure [2](#)).

1179 The condition of $n \gtrsim \frac{1}{\eta^2} \max\{d, \tilde{d}\} (1 + \gamma^{-1})(1 + \tilde{\gamma}^{-1})$ is inherited from Corollary [1](#) (to be able to
 1180 use eq [\(41\)](#)). The additional condition on n , from the application of Lemma [2](#) to the above equation
 1181 (similar to Eq. [\(29\)](#)), results in a larger factor than $1/\eta^2$, hence is already satisfied.

1182 Now we apply Lemma [2](#) on the above equation, and follow the argument similar to step 5 in Section [D](#).
 1183 An additional factor of \sqrt{r} appears due to the norm being the chordal distance (frobenius norm).

Using Eq. (51) and Eq. (58), we get that with probability $1 - \exp(-\Omega(\max\{d, \tilde{d}\}))$, the error $\text{ERR}(\mathbf{G}, \tilde{\mathbf{G}})$ is upper bounded (up to constants) by:

$$\frac{1}{[\eta f_1(\theta) + (1 - \eta) f_0(\theta)] \sqrt{\eta P_1(\theta_{\rho_T}/\eta) + (1 - \eta) P_0(\theta_{\rho_T}/\eta)}} \sqrt{\frac{r \max\{d, \tilde{d}\} \text{poly}(\gamma^{-1}, \tilde{\gamma}^{-1})}{n}}.$$

Finally, we plug in the values $\theta \in \{0, \eta^{r/2\rho_T}\}$ to recover the terms $T_0, T_{0.5}$ as stated in Theorem 1. Using Eq. (52) and (59), the scaling term of the error above becomes

$$\frac{1}{[\eta + (1 - \eta)(2/\pi r)] \cdot \sqrt{\eta(1 - 1/e) + (1 - \eta)(1/2)}} \lesssim r \quad \text{for any } \eta \in (0, 1].$$

Using Eq. (53) and (60), the scaling term of the error above becomes

$$\frac{1}{[\eta + (1 - \eta)(1/2)] \cdot \sqrt{\eta(1 - 1/e)}} \lesssim \frac{1}{\sqrt{\eta}}.$$

The above describes both regimes of behavior, and why an extra factor of r appears in the term T_0 , compared to the term $T_{0.5}$, in Theorem 1. This concludes the argument.

H Discussion on the potential of robust statistics for the analysis of filtering

An initial instinct based on Figure 2 is to use ideas from robust statistics. As discussed in Remark 6.1 we can expect \mathcal{D}_0 and \mathcal{D}_1 to be well-separated, which means there will exist some $\theta \in \mathbb{R}$ (a reasonable guess is $\theta \approx r/2$) such that the selected data is mostly clean. After filtering, the picture resembles the robust statistics setting: an α corruption on the clean distribution for some small α . This is a reasonable approach overall, but has two shortcomings. *First*, this approach will *not* achieve zero error as $n \rightarrow \infty$. We are shooting for $f(\eta) \cdot 1/\sqrt{n}$ which is better than $1/\sqrt{n} + g(\eta)$, since the latter is non-zero even when $n \rightarrow \infty$. This approach will end up getting the latter. This is because the canonical rate in robust statistics is $\sqrt{d/n_{\text{sel}}} + \alpha$. Under filtering, n_{sel} and α are functions of θ . One can determine the optimal θ to balance the tradeoff, but to get a final rate of the form $f(\eta) \cdot 1/\sqrt{n}$, this will require some *conditions* on n, η (possibly η bigger than a threshold, and n smaller than a threshold). Since our case has stochastic corruption which is weaker than adversarial corruption, we can expect to prove something for all n and all η . *Second*, this approach performs a “reductive” operation of treating data as only clean v/s corrupted, and assuming the corrupted part provides no signal. This is a closely linked argument to the first one above. The crucial observation is that the right tail of the corrupted data (i.e. \mathcal{D}_0 in Figure 2) actually provides ‘close to clean’ samples. This is because these just happened to be samples such that the z, \tilde{z} – albeit independently sampled in a high-dimensional space – happened to have a high inner product (small angle). Our adopted approach, based on the conditional properties of the Gaussian distribution, formalizes this intuition that the right tail of \mathcal{D}_0 also provides signal.