Appendices

In the remainder of the paper, we give self-contained proofs of all results from the main text.

In Appendix A, we introduce some technical results that we will use in our analysis.

In Appendix B, we prove the main generalization bound (Theorem 1) and show its specialization to norm balls (Corollaries 1 to 3).

In Appendix C, we prove upper bounds on the norm of the minimal-norm interpolator for a general norm (Theorem 4), and show applications to the Euclidean case (Theorem 2).

In Appendix D, we show how to combine the previous sets of results to give risk guarantees for the minimal norm interpolators (Theorems 3 and 5). In particular, Appendix D.2.1 shows the equivalence of conditions for consistency in the Euclidean norm setting.

In Appendix E, we provide full theorem statements and proofs of the results on ℓ_1 interpolation (basis pursuit) mentioned in Section 6.

A Preliminaries

We will first give some general results useful to the rest of the proofs. Most are standard, but a few are variations on existing results.

Concentration of Lipschitz functions. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to the norm $\|\cdot\|$ if it holds for all $x, y \in \mathbb{R}^n$ that $|f(x) - f(y)| \leq L\|x - y\|$. We use the concentration of Lipschitz functions of a Gaussian.

Theorem 6 (van Handel 2014, Theorem 3.25)**.** *If f is L-Lipschitz with respect to the Euclidean norm and $Z \sim N(0, I_n)$, then*

$$\Pr(|f(Z) - \mathbb{E}\, f(Z)| \geq t) \leq 2e^{-t^2/2L^2}. \tag{23}$$

We also use a similar result for functions of a uniformly spherical vector (see Vershynin 2018, Theorem 5.1.4 and Exercise 5.1.12); we cite a result with sharp constant factor from Ledoux (1992).

Theorem 7 (Spherical concentration; Ledoux 1992)**.** *If f is L-Lipschitz with respect to the Euclidean norm and $Z \sim Uni(S^{n-1})$ where $S^{n-1} = \{u \in \mathbb{R}^n : \|u\| = 1\}$ is the unit sphere, $Uni(S^{n-1})$ is the uniform measure on the sphere, and $n \geq 3$, then*

$$\Pr(|f(Z) - \mathbb{E}\, f(Z)| \geq t) \leq 2e^{-(n-2)t^2/2L^2}. \tag{24}$$

The following lemma, which we will use multiple timues, says that a $o(n)$-dimensional subspace cannot align with a random spherically symmetric vector.

Lemma 1. *Suppose that S is a fixed subspace of dimension d in $\mathbb{R}^n$ with $n \geq 4$, P_S is the orthogonal projection onto S, and V is a spherically symmetric random vector (i.e. $V/\|V\|_2$ is uniform on the sphere). Then*

$$\frac{\|P_S V\|_2}{\|V\|_2} \leq \sqrt{d/n} + 2\sqrt{\log(2/\delta)/n}. \tag{25}$$

with probability at least $1 - \delta$. Conditional on this inequality holding, we therefore have uniformly for all $s \in S$ that

$$|\langle s, V\rangle| = |\langle s, P_S V\rangle| \leq \|s\|_2 \|P_S V\|_2 \leq \|s\|_2 \|V\|_2 \left(\sqrt{d/n} + 2\sqrt{\log(2/\delta)/n}\right). \tag{26}$$

Proof. This is trivial if $d \geq n$, since the left-hand side is at most 1. Thus assume without loss of generality that $d < n$. By symmetry, it suffices to fix S to be the span of basis vectors $e_1, \ldots, e_d$ and to bound $\|P_S V\|_2$ for V a uniformly random chosen vector from the unit sphere in $\mathbb{R}^n$. Recall that for any coordinate i, we have $\mathbb{E}\, V_i^2 = 1/n$ by symmetry among the coordinates and the fact that $\|V\|_2^2 = 1$ almost surely. The function $v \mapsto \|P_S v\|_2$ is a 1-Lipschitz function and $\mathbb{E}\, \|P_S V\|_2 \leq \sqrt{\mathbb{E}\, \|P_S V\|_2^2} = \sqrt{d/n}$, so by Theorem 7 above

$$\|P_S V\|_2 \leq \sqrt{d/n} + \sqrt{2\log(2/\delta)/(n-2)})$$

with probability at least $1 - \delta$. Using $n \geq 4$ gives the result. □

The concentration of the Euclidean norm of a Gaussian vector follows from Theorem 6; we state it explicitly below.

Lemma 2. *Suppose that $Z \sim N(0, I_n)$. Then*

$$\Pr(\big|\|Z\|_2 - \sqrt{n}\big| \geq t) \leq 4e^{-t^2/4}. \tag{27}$$

Proof. First we recall the standard fact (see e.g. Chandrasekaran et al. 2012) that

$$\sqrt{n} - 1 \leq \frac{n}{\sqrt{n+1}} \leq \mathbb{E}\,\|Z\|_2 \leq \sqrt{n}.$$

Because the norm is 1-Lipschitz, it follows from Theorem 6 that

$$\Pr(|\|Z\|_2 - \mathbb{E}\,\|Z\|_2| \geq t) \leq 2e^{-t^2/2}$$

so

$$\Pr(\big|\|Z\|_2 - \sqrt{n}\big| \geq t + 1) \leq 2e^{-t^2/2}.$$

Now using that $(t-1)^2 \geq t^2/2 - 1$ shows

$$\Pr(\big|\|Z\|_2 - \sqrt{n}\big| \geq t) \leq 2e^{-(t^2/2-1)/2} \leq 4e^{-t^2/4}. \qquad \square$$

Wishart concentration. We recall the notation for the Loewner order on symmetric matrices: $A \preceq B$ means that $B - A$ is positive semidefinite. Let $\sigma_{\min}(A)$ denote the minimum singular value of an arbitrary matrix A, and $\sigma_{\max}$ the maximum singular value. Similarly, let $\lambda_{\min}(A)$ denote the minimum eigenvalue. We use $\|A\|_{op} = \sigma_{\max}(A)$ to denote the operator norm of matrix A.

Theorem 8 (Vershynin 2010, Corollary 5.35)**.** *Let $n, N \in \mathbb{N}$. Let $A \in \mathbb{R}^{N \times n}$ be a random matrix with entries i.i.d. $N(0,1)$. Then for any $t > 0$, it holds with probability at least $1 - 2\exp(-t^2/2)$ that*

$$\sqrt{N} - \sqrt{n} - t \leq \sigma_{min}(A) \leq \sigma_{max}(A) \leq \sqrt{N} + \sqrt{n} + t. \tag{28}$$

Corollary 4. *Suppose $X_1, \ldots, X_n \sim N(0, \Sigma)$ are independent with $\Sigma : d \times d$ a positive semidefinite matrix, $t > 0$ and $n \geq 4(d + t^2)$. Let $\hat{\Sigma} = \frac{1}{n}\sum_i X_i X_i^T$ be the empirical covariance matrix. Then with probability at least $1 - \delta$,*

$$(1 - \epsilon)\Sigma \preceq \hat{\Sigma} \preceq (1 + \epsilon)\Sigma \tag{29}$$

with $\epsilon = 3\sqrt{d/n} + 3\sqrt{2\log(2/\delta)/n}$.

Proof. Let $X : n \times d$ be the random matrix with rows $X_1, \ldots, X_n$ so that $\hat{\Sigma} = \frac{1}{n} X^T X$. By equality in distribution, we can take $Z : n \times d$ to have $N(0,1)$ independent entries and write $X = Z\Sigma^{1/2}$ and

$$\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} = \frac{1}{m}\Sigma^{-1/2} X^T X \Sigma^{-1/2} = \frac{1}{n} Z^T Z.$$

By definition of singular values, from Theorem 8 the eigenvalues of $Z^T Z/n$ are bounded between $(1 - \sqrt{d/n} - \sqrt{t^2/n})^2$ and $(1 + \sqrt{d/n} + \sqrt{t^2/n})^2$. Since $1 - (1-x)^2 \leq (1+x)^2 - 1$, using the inequality $(1+x)^2 \leq 1 + 3x$ for $x \in [0,1]$, we have shown that

$$\|I - \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}\|_{op} \leq (1 + \sqrt{d/n} + \sqrt{t^2/n})^2 - 1 \leq 3\sqrt{d/n} + 3\sqrt{t^2/n}.$$

Rewriting and taking $t^2 = 2\log(2/\delta)$ gives the result. $\square$

Gaussian Minmax Theorem. The following result is Theorem 3 of Thrampoulidis et al. (2015), known as the Convex Gaussian Minmax Theorem or CGMT (see also Theorem 1 in the same reference). As explained there, it is a consequence of the main result of Gordon (1985), known as Gordon's Theorem or the Gaussian Minmax Theorem. Despite the name, convexity is only required for one of the theorem's conclusions.

Theorem 9 (Convex Gaussian Minmax Theorem; Gordon 1985; Thrampoulidis et al. 2015)**.** *Let $Z : n \times d$ be a matrix with i.i.d. $N(0,1)$ entries and suppose $G \sim N(0, I_n)$ and $H \sim N(0, I_d)$ are independent of Z and each other. Let S_w, S_u be compact sets and $\psi : S_w \times S_u \to \mathbb{R}$ be an arbitrary continuous function. Define the* Primary Optimization (PO) *problem*

$$\Phi(Z) := \min_{w \in S_w} \max_{u \in S_u} \langle u, Zw \rangle + \psi(w, u) \tag{30}$$

and the Auxiliary Optimization (AO) *problem*

$$\phi(G, H) := \min_{w \in S_w} \max_{u \in S_u} \|w\|_2 \langle G, u \rangle + \|u\|_2 \langle H, w \rangle + \psi(w, u). \tag{31}$$

Under these assumptions, $\Pr(\Phi(Z) < c) \leq 2 \Pr(\phi(G, H) \leq c)$ *for any* $c \in \mathbb{R}$.

Furthermore, if we suppose that S_w, S_u are convex sets and $\psi(w, u)$ is convex in w and concave in u, then $\Pr(\Phi(Z) > c) \leq 2 \Pr(\phi(G, H) \geq c)$.

In other words, the first conclusion says that high probability lower bounds on the auxiliary optimization $\phi(G, H)$ imply high probability lower bounds on the primary optimization $\Phi(Z)$. Importantly, this direction holds without any convexity assumptions. Under the additional convexity assumptions, the second conclusion gives a similar comparison of high probability upper bounds.

In our analysis, we need a slightly more general statement of the Gaussian Minmax Theorem than Theorem 9: we need the minmax formulation to include additional variables which only affect the deterministic term in the minmax problem. It's straightforward to prove this result by repeating the argument in Thrampoulidis et al. (2015); below we give an alternative proof which reduces to Theorem 9, by introducing extremely small extra dimensions to contain the extra variables. Intuitively, this works because the statement of the GMT allows for arbitrary continuous functions ψ, with no dependence on their quantitative smoothness.

Theorem 10 (Variant of GMT)**.** *Let $Z : n \times d$ be a matrix with i.i.d. $N(0,1)$ entries and suppose $G \sim N(0, I_n)$ and $H \sim N(0, I_d)$ are independent of Z and each other. Let S_W, S_U be compact sets in $\mathbb{R}^d \times \mathbb{R}^{d'}$ and $\mathbb{R}^n \times \mathbb{R}^{n'}$ respectively, and let $\psi : S_W \times S_U \to \mathbb{R}$ be an arbitrary continuous function. Define the* Primary Optimization (PO) *problem*

$$\Phi(Z) := \min_{(w,w') \in S_W} \max_{(u,u') \in S_U} \langle u, Zw \rangle + \psi((w, w'), (u, u')) \tag{32}$$

and the Auxiliary Optimization (AO) *problem*

$$\phi(G, H) := \min_{(w,w') \in S_W} \max_{(u,u') \in S_U} \|w\|_2 \langle G, u \rangle + \|u\|_2 \langle H, w \rangle + \psi((w, w'), (u, u')). \tag{33}$$

Under these assumptions, $\Pr(\Phi(Z) < c) \leq 2 \Pr(\phi(G, H) \leq c)$ *for any* $c \in \mathbb{R}$.

Proof. Let $\epsilon \in (0, 1)$ be arbitrary and

$$S_{W,\epsilon} := \{(w, \epsilon w') : (w, w') \in S_W\}, \quad S_{U,\epsilon} := \{(u, \epsilon u') : (u, u') \in S_U\}.$$

Define $\psi_\epsilon((w, w'), (u, u')) := \psi((w, \frac{1}{\epsilon} w'), (u, \frac{1}{\epsilon} u'))$ so that if $W = (w, \epsilon w')$ and $U = (u, \epsilon u')$, then $\psi_\epsilon(W, U) = \psi((w, w'), (u, u'))$. We also define $S_w = \{w \in \mathbb{R}^d : \exists w' \, s.t. \, (w, w') \in S_W\}$. The other sets $S_{w'}, S_u$ and $S_{u'}$ are defined similarly. It is clear that $S_w, S_{w'}, S_u, S_{u'}, S_{W,\epsilon}$ and $S_{U,\epsilon}$ are all still compact in their respective topology, and ψ_ϵ is continuous for every $\epsilon > 0$.

Let $Z' : (n + n') \times (d + d')$ be a matrix with i.i.d. $N(0,1)$ entries such that the top left $n \times d$ matrix is Z. Similarly, we define G' to be a $(n + n')$-dimensional Gaussian vector with independent coordinates such that the first n coordinates are G, and H' to be a $(d + d')$-dimensional Gaussian vector with independent coordinates such that the first d coordinates are H. Next, consider the augmented PO and AO:

$$\begin{aligned} \Phi_\epsilon(Z') &:= \min_{W \in S_{W,\epsilon}} \max_{U \in S_{U,\epsilon}} \langle U, Z'W \rangle + \psi_\epsilon(W, U) \\ \phi_\epsilon(G', H') &:= \min_{W \in S_{W,\epsilon}} \max_{U \in S_{U,\epsilon}} \|W\|_2 \langle G', U \rangle + \|U\|_2 \langle H', W \rangle + \psi_\epsilon(W, U) \end{aligned} \tag{34}$$

It is clear that for a small value of ϵ, the augmented problem will be close to the original problem. More precisely, for every $(w, w') \in S_W$ and $(u, u') \in S_U$

$$\begin{aligned} &|\langle (w, \epsilon w'), Z'(u, \epsilon u') \rangle - \langle w, Zu \rangle| \\ &\quad = |\epsilon \langle (0, w'), Z'(u, 0) \rangle + \epsilon \langle (w, 0), Z'(0, u') \rangle + \epsilon^2 \langle (0, w'), Z'(0, u') \rangle| \\ &\quad \leq \epsilon (R(S_w) + R(S_{w'}))(R(S_u) + R(S_{u'})) \|Z'\|_{op} = \epsilon A \|Z'\|_{op} \end{aligned} \tag{35}$$

where $A := (R(S_w) + R(S_{w'}))(R(S_u) + R(S_{u'}))$ is deterministic and does not depend on ϵ. Similarly, it is routine to check

$$\|w\|_2\langle G, u\rangle = \|w\|_2(\langle G', (u, \epsilon u')\rangle - \epsilon\langle G', (0, u')\rangle)$$
$$\|u\|_2\langle H, w\rangle = \|u\|_2(\langle H', (w, \epsilon w')\rangle - \epsilon\langle H', (0, w')\rangle)$$

so by the triangle inequality and Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\left|\|(w, \epsilon w')\|_2\langle G', (u, \epsilon u')\rangle - \|w\|_2\langle G, u\rangle\right| \\ &\quad\le \epsilon R(S_{w'})\|G'\|_2(R(S_u) + \epsilon R(S_{u'})) + \epsilon R(S_w)\|G'\|_2 R(S_{u'}) \le \epsilon A\|G'\|_2\end{aligned} \tag{36}$$

and

$$\begin{aligned}&\left|\|(u, \epsilon u')\|_2\langle H', (w, \epsilon w')\rangle - \|u\|_2\langle H, w\rangle\right| \\ &\quad\le \epsilon R(S_{u'})\|H'\|_2(R(S_w) + \epsilon R(S_{w'})) + \epsilon R(S_u)\|H'\|_2 R(S_{w'}) \le \epsilon A\|H'\|_2\end{aligned} \tag{37}$$

From (35), it follows that

$$|\Phi_\epsilon(Z') - \Phi(Z)| \le \epsilon A\|Z'\|_{op}. \tag{38}$$

Similarly, from (36) and (37), it follows that

$$|\phi_\epsilon(G', H') - \phi(G, H)| \le \epsilon A(\|G'\|_2 + \|H'\|_2). \tag{39}$$

Approximating the original PO and AO by (34) allows us to directly apply the Gaussian Minmax Theorem. For any $c \in \mathbb{R}$, we have

$$\begin{aligned}\Pr(\Phi(Z) < c) &\le \Pr(\Phi_\epsilon(Z') < c + \sqrt{\epsilon}) + \Pr(\epsilon A\|Z'\|_{op} > \sqrt{\epsilon}) \\ &\le 2\Pr(\phi_\epsilon(G', H') \le c + \sqrt{\epsilon}) + \Pr(\epsilon A\|Z'\|_{op} > \sqrt{\epsilon}) \\ &\le 2\Pr(\phi(G', H') \le c + 2\sqrt{\epsilon}) + 2\Pr\left(\epsilon A(\|G'\|_2 + \|H'\|_2) > \sqrt{\epsilon}\right) \\ &\quad + \Pr(\epsilon A\|Z'\|_{op} > \sqrt{\epsilon}) \\ &\le 2\Pr(\phi(G', H') \le c + 2\sqrt{\epsilon}) + 2\Pr\left(\|G'\|_2 > \frac{1}{2A\sqrt{\epsilon}}\right) \\ &\quad + 2\Pr\left(\|H'\|_2 > \frac{1}{2A\sqrt{\epsilon}}\right) + \Pr\left(\|Z'\|_{op} > \frac{1}{A\sqrt{\epsilon}}\right)\end{aligned}$$

where we used (38) in the first inequality, Theorem 9 in the second inequality, and (39) in the last inequality. This holds for arbitrary $\epsilon > 0$ and taking the limit $\epsilon \to 0$ shows the result, because the CDF is right continuous (Durrett 2019) and the remaining terms go to zero by standard concentration inequalities (Lemma 2 and Theorem 8). □

B Uniform Convergence Bounds

We will now prove the main generalization bound, as well as its special cases in norm balls and specifically Euclidean norm balls.

B.1 General case: Proof of Theorem 1

For convenience, we restate the definition of covariance splitting here:

Definition 2 (Covariance splitting)**.** Given a positive semidefinite matrix $\Sigma \in \mathbb{R}^{d\times d}$, we write $\Sigma = \Sigma_1 \oplus \Sigma_2$ if $\Sigma = \Sigma_1 + \Sigma_2$, each matrix is positive semidefinite, and their spans are orthogonal.

It follows from our definition that $\Sigma_1\Sigma_2 = 0$. Although our results in Appendix C requires this orthogonality condition (in particular, Lemma 8), we note that all of our results here in Appendix B continue to hold as long as $\Sigma = \Sigma_1 + \Sigma_2$ and both Σ_1, Σ_2 are positive semi-definite. To apply the Gaussian Minimax Theorem, we first formulate the generalization gap as an optimization problem in terms of a random matrix with $N(0, 1)$ entries.

Lemma 3. *Under the model assumptions in (1), let $\mathcal{K}$ be an arbitrary compact set and $\Sigma = \Sigma_1 \oplus \Sigma_2$. Define the primary optimization problem (PO) as*

$$\Phi := \max_{\substack{(w_1,w_2)\in\mathcal{S} \\ Z_1 w_1 + Z_2 w_2 = \xi}} \|w_1\|_2^2 + \|w_2\|_2^2 \tag{40}$$

where

$$\mathcal{S} = \{(w_1, w_2) : \exists w \in \mathcal{K}\, s.t.\, w_1 = \Sigma_1^{1/2}(w - w^*) \text{ and } w_2 = \Sigma_2^{1/2}(w - w^*)\} \tag{41}$$

and Z_1, Z_2 are both $n \times d$ random matrices with i.i.d. standard normal entries independent of ξ and each other. Then the generalization gap of interpolators is equal in distribution to the sum of the Bayes risk and the PO:

$$\max_{w\in\mathcal{K}, \hat{L}(w)=0} L(w) - \hat{L}(w) \overset{\mathcal{D}}{=} \sigma^2 + \Phi. \tag{42}$$

Proof. Recall that $L(w) = \sigma^2 + \|w - w^*\|_\Sigma^2$ and $\hat{L}(w) = 0$ is equivalent to $Y = Xw$. Observe that

$$X \overset{\mathcal{D}}{=} Z_1\Sigma_1^{1/2} + Z_2\Sigma_2^{1/2} \quad \text{and} \quad \|w\|_\Sigma^2 = \|w\|_{\Sigma_1}^2 + \|w\|_{\Sigma_2}^2$$

so we can decompose

$$\begin{aligned}
\max_{w\in\mathcal{K}, \hat{L}(w)=0} L(w) - \hat{L}(w) &= \sigma^2 + \max_{w\in\mathcal{K}, Y = Xw} \|w - w^*\|_\Sigma^2 \\
&= \sigma^2 + \max_{w\in\mathcal{K}, X(w-w^*)=\xi} \|w - w^*\|_\Sigma^2 \\
&\overset{\mathcal{D}}{=} \sigma^2 + \max_{\substack{w\in\mathcal{K}-w^* \\ (Z_1\Sigma_1^{1/2} + Z_2\Sigma_2^{1/2})w = \xi}} \|w\|_{\Sigma_1}^2 + \|w\|_{\Sigma_2}^2 = \sigma^2 + \Phi.
\end{aligned}$$

□

Lemma 4 (Application of GMT)**.** *In the same setting as Lemma 3, let $G \sim N(0, I_n)$, $H \sim N(0, I_d)$ be Gaussian vectors independent of Z_1, Z_2, ξ and each other. With the same definition of $\mathcal{S}$, define the auxiliary optimization problem (AO) as*

$$\phi := \max_{\substack{(w_1,w_2)\in\mathcal{S} \\ \|\xi - Z_1 w_1 - G\|w_2\|_2\|_2 \le \langle w_2, H\rangle}} \|w_1\|_2^2 + \|w_2\|_2^2 \tag{43}$$

Then it holds that

$$\Pr(\Phi > t \,|\, Z_1, \xi) \le 2\Pr(\phi \ge t \,|\, Z_1, \xi), \tag{44}$$

and taking expectations we have

$$\Pr(\Phi > t) \le 2\Pr(\phi \ge t). \tag{45}$$

Proof. By introducing Lagrange multipliers, we have

$$\begin{aligned}
\Phi &= \max_{(w_1,w_2)\in\mathcal{S}} \min_\lambda \|w_1\|_2^2 + \|w_2\|_2^2 + \langle\lambda, Z_2 w_2 - (\xi - Z_1 w_1)\rangle \\
&= \max_{(w_1,w_2)\in\mathcal{S}} \min_\lambda \langle\lambda, Z_2 w_2\rangle + \|w_1\|_2^2 + \|w_2\|_2^2 - \langle\lambda, \xi - Z_1 w_1\rangle.
\end{aligned}$$

By independence, the distribution of Z_2 remains the same after conditioning on Z_1 and ξ and the randomness in Φ comes solely from Z_2. Since the mapping from w to (w_1, w_2) is continuous and $\mathcal{K}$ is compact, $\mathcal{S}$ is compact. To apply Theorem 10, we can take $\psi(w_1, w_2, \lambda) = \|w_1\|_2^2 + \|w_2\|_2^2 - \langle\lambda, \xi - Z_1 w_1\rangle$, which is clearly continuous. The only challenge is that the domain of λ is not compact, but we can handle it by a truncation argument. Define

$$\Phi_r := \max_{(w_1,w_2)\in\mathcal{S}} \min_{\|\lambda\|\le r} \langle\lambda, Z_2 w_2\rangle + \|w_1\|_2^2 + \|w_2\|_2^2 - \langle\lambda, \xi - Z_1 w_1\rangle \tag{46}$$

and observe that $\Phi \le \Phi_r$, since the minimum in the definition of Φ_r ranges over a smaller set. The AO associated with Φ_r is

$$\begin{aligned}
\phi_r :&= \max_{(w_1,w_2)\in\mathcal{S}} \min_{\|\lambda\|\le r} \|w_2\|_2\langle G, \lambda\rangle + \|\lambda\|_2\langle H, w_2\rangle + \|w_1\|_2^2 + \|w_2\|_2^2 - \langle\lambda, \xi - Z_1 w_1\rangle \\
&= \max_{(w_1,w_2)\in\mathcal{S}} \min_{\|\lambda\|\le r} \|\lambda\|_2\langle H, w_2\rangle - \langle\lambda, \xi - Z_1 w_1 - G\|w_2\|_2\rangle + \|w_1\|_2^2 + \|w_2\|_2^2 \\
&= \max_{(w_1,w_2)\in\mathcal{S}} \min_{0\le\lambda\le r} \lambda\left(\langle H, w_2\rangle - \|\xi - Z_1 w_1 - G\|w_2\|_2\|_2\right) + \|w_1\|_2^2 + \|w_2\|_2^2.
\end{aligned} \tag{47}$$

We observe that the untruncated auxiliary problem ϕ from (43) has a completely analogous form:

$$\phi = \max_{(w_1,w_2)\in\mathcal{S}} \min_{\lambda\geq 0} \lambda\left(\langle H, w_2\rangle - \|\xi - Z_1 w_1 - G\|w_2\|_2\|_2\right) + \|w_1\|_2^2 + \|w_2\|_2^2.$$

This is because if $\langle H, w_2\rangle - \|\xi - Z_1 w_1 - G\|w_2\|_2\|_2 \geq 0$ then the minimum is achieved at $\lambda = 0$, and if w_1, w_2 do not satisfy the constraint then taking $\lambda \to \infty$ sends the minimum to $-\infty$. From this formulation, we see that $\phi \leq \phi_r \leq \phi_s$ for any $r \geq s \geq 0$ since the minimum is taken over a larger set as r grows, and is unconstrained in ϕ.

The proof that $\lim_{r\to\infty} \phi_r = \phi$ is an exercise in real analysis, which splits into two cases:

1. The auxiliary problem ϕ is infeasible. In this case, we know that for all $(w_1, w_2) \in \mathcal{S}$

   $$\langle H, w_2\rangle - \|\xi - Z_1 w_1 - G\|w_2\|_2\|_2 < 0.$$

   By compactness of $\mathcal{S}$ and continuity of the right hand side, there exists $\mu = \mu(\xi, Z_1, G, H) < 0$ (in particular, independent of r) such that

   $$\langle H, w_2\rangle - \|\xi - Z_1 w_1 - G\|w_2\|_2\|_2 \leq \mu.$$

   Therefore, we show

   $$\begin{aligned}\phi_r &\leq \max_{(w_1,w_2)\in\mathcal{S}} \min_{0\leq\lambda\leq r} \lambda\mu + \|w_1\|_2^2 + \|w_2\|_2^2 \\ &= r\mu + \max_{(w_1,w_2)\in\mathcal{S}} \|w_1\|_2^2 + \|w_2\|_2^2.\end{aligned}$$

   Since the second term is bounded and has no dependence on r, taking $r \to \infty$ we have $\phi_r \to -\infty$ as desired (since $\phi = -\infty$ by definition).

2. The auxiliary problem ϕ is feasible. In this case, we can let $(w_1(r), w_2(r)) \in \mathcal{S}$ be an arbitrary maximizer achieving the objective ϕ_r for each $r \geq 0$ by compactness. By compactness again, the sequence $(w_1(r), w_2(r))_{r=1}^{\infty}$ at positive integer values of r has a subsequential limit $(w_1(\infty), w_2(\infty)) \in \mathcal{S}$, i.e. this point satisfies $(w_1(\infty), w_2(\infty)) = \lim_{n\to\infty}(w_1(r_n), w_2(r_n))$ for some sequence r_n satisfying $r_n \geq n$.

   Suppose that $(w_1(\infty), w_2(\infty))$ does not satisfy the last constraint defining ϕ, then by continuity, there exists $\mu < 0$ and a sufficiently small $\epsilon > 0$ such that for all $\|w_1 - w_1(\infty)\|_2 \leq \epsilon$ and $\|w_2 - w_2(\infty)\|_2 \leq \epsilon$, we have

   $$\langle H, w_2\rangle - \|\xi - Z_1 w_1 - G\|w_2\|_2\|_2 \leq \mu.$$

   This implies that for sufficiently large n, we have

   $$\langle H, w_2(r_n)\rangle - \|\xi - Z_1 w_1(r_n) - G\|w_2(r_n)\|_2\|_2 \leq \mu$$

   and

   $$\begin{aligned}\phi_{r_n} &\leq r_n\mu + \|w_1(r_n)\|_2^2 + \|w_2(r_n)\|_2^2 \\ &\leq r_n\mu + \max_{(w_1,w_2)\in\mathcal{S}} \|w_1\|_2^2 + \|w_2\|_2^2\end{aligned}$$

   so $\phi_{r_n} \to -\infty$ – but this is impossible, since considering any feasible element of ϕ we can show that $\phi_{r_n} \geq 0$. By contradiction, we find that $(w_1(\infty), w_2(\infty))$ is feasible for ϕ.

   By taking $\lambda = 0$ in the definition of ϕ_r we have

   $$\phi_{r_n} \leq \|w_1(r_n)\|_2^2 + \|w_2(r_n)\|_2^2.$$

   By continuity, we show that

   $$\begin{aligned}\limsup_{n\to\infty} \phi_{r_n} &\leq \lim_{n\to\infty} \|w_1(r_n)\|_2^2 + \|w_2(r_n)\|_2^2 \\ &= \|w_1(\infty)\|_2^2 + \|w_2(\infty)\|_2^2 \leq \phi.\end{aligned}$$

   Since $\phi_{r_n} \geq \phi$, the limit of ϕ_{r_n} exists and equals ϕ. We can conclude that $\lim_{r\to\infty} \phi_r = \phi$ because ϕ_r is a monotone decreasing function of r.

By our version of the Gaussian Minmax Theorem, Theorem 10,

$$\Pr(\Phi_r > t|Z_1,\xi) = \Pr(-\Phi_r < -t|Z_1,\xi) \le 2\Pr(-\phi_r \le -t|Z_1,\xi) = 2\Pr(\phi_r \ge t|Z_1,\xi)$$

We introduce the negative signs here because we have originally a max-min problem instead of a min-max problem. This means the comparison theorem gives an upper bound, instead of a lower bound, on the quantity of interest.

Finally, we can conclude

$$\Pr(\Phi > t|Z_1,\xi) \le \inf_{r\ge 0} \Pr(\Phi_r > t|Z_1,\xi) \le 2 \inf_{r\ge 0} \Pr(\phi_r \ge t|Z_1,\xi) \le 2\Pr(\phi \ge t|Z_1,\xi).$$

where the last step uses continuity (from above) of probability measure and the fact that ϕ_r monotonically decreases to ϕ almost surely. □

Recall the definition of Gaussian width and radius:

Definition 1. The *Gaussian width* and the *radius* of a set $S \subset \mathbb{R}^d$ are

$$W(S) := \mathop{\mathbb{E}}_{H\sim N(0,I_d)} \sup_{s\in S} |\langle s, H\rangle| \quad \text{and} \quad \operatorname{rad}(S) := \sup_{s\in S} \|s\|_2.$$

It remains to analyze the auxiliary problem, which we do in the following lemma:

Lemma 5. *Let $\beta = 33\sqrt{\frac{\log(32/\delta)}{n}} + 18\sqrt{\frac{\operatorname{rank}(\Sigma_1)}{n}}$. If n is sufficiently large such that $\beta \le 1$, then with probability at least $1-\delta$, it holds that*

$$\phi \le \frac{1+\beta}{n}(W(\Sigma_2^{1/2}\mathcal{K}) + \operatorname{rad}(\Sigma_2^{1/2}\mathcal{K})\sqrt{2\log(16/\delta)} + \|w^*\|_{\Sigma_2}\sqrt{2\log(16/\delta)})^2 - \sigma^2. \tag{48}$$

Proof. For notational simplicity, define

$$\begin{aligned}
\alpha &:= 2\sqrt{\frac{\log(32/\delta)}{n}} \\
\gamma &:= 3\sqrt{\frac{\operatorname{rank}(\Sigma_1)}{n}} + 3\sqrt{\frac{2\log(16/\delta)}{n}} \\
\rho &:= \sqrt{\frac{\operatorname{rank}(\Sigma_1)+1}{n}} + 2\sqrt{\frac{\log(16/\delta)}{n}}.
\end{aligned}$$

By a union bound, the following collection of events, which together we call $\mathcal{E}$, occurs with probability at least $1-\delta$:

1. (Approximate orthogonality.) By Lemma 1, uniformly over all $w_1 \in \Sigma_1^{1/2}(\mathcal{K}-w^*)$ and $a \in \mathbb{R}$, it holds that

$$|\langle \xi a - Z_1 w_1, G\rangle| \le \|\xi a - Z_1 w_1\|_2 \|G\|_2 \rho \tag{49}$$

and

$$|\langle \xi, Z_1 w_1\rangle| \le \|\xi\|_2 \|Z_1 w_1\|_2 \rho. \tag{50}$$

2. (Approximate isometry.) By Corollary 4, uniformly over all $w_1 \in \Sigma_1^{1/2}(\mathcal{K}-w^*)$, it holds that

$$(1-\gamma)\|w_1\|_2^2 \le \frac{\|Z_1 w_1\|_2^2}{n} \le (1+\gamma)\|w_1\|_2^2. \tag{51}$$

3. (Typical norm of G and ξ.) By Lemma 2, it holds that

$$-\alpha \le \frac{1}{\sqrt{n}}\|G\|_2 - 1 \le \alpha \tag{52}$$

and

$$-\alpha\sigma \le \frac{1}{\sqrt{n}}\|\xi\|_2 - \sigma \le \alpha\sigma. \tag{53}$$

4. (Typical size of $\langle \Sigma_2^{1/2} w^*, H\rangle$.) By the standard Gaussian tail bound $\Pr(|Z| \geq t) \leq 2e^{-t^2/2}$, it holds that
$$|\langle \Sigma_2^{1/2} w^*, H\rangle| \leq \|w^*\|_{\Sigma_2} \sqrt{2\log(16/\delta)} \tag{54}$$
because the marginal law of $\langle \Sigma_2^{1/2} w^*, H\rangle$ is $N(0, \|w^*\|_{\Sigma_2}^2)$.

5. (Gaussian process concentration.) By Theorem 6, it holds that
$$\max_{w_2 \in \Sigma_2^{1/2}\mathcal{K}} |\langle w_2, H\rangle| \leq W(\Sigma_2^{1/2}\mathcal{K}) + \mathrm{rad}(\Sigma_2^{1/2}\mathcal{K})\sqrt{2\log(16/\delta)} \tag{55}$$
because $\max_{w_2 \in \Sigma_2^{1/2}\mathcal{K}} |\langle u_2, H\rangle|$ is a $\mathrm{rad}(\Sigma_2^{1/2}\mathcal{K})$-Lipschitz function of H.

From now on, the argument is conditional on the event $\mathcal{E}$ defined above. By squaring the last constraint in the definition of ϕ we see that
$$\begin{aligned}
\langle w_2, H\rangle^2 &\geq \|\xi - Z_1 w_1 - \|w_2\|_2 G\|_2^2 \\
&= \|\xi - Z_1 w_1\|_2^2 + \|w_2\|_2^2 \|G\|_2^2 - 2\langle \xi - Z_1 w_1, \|w_2\|_2 G\rangle \\
&\geq (1-\rho)[\|\xi - Z_1 w_1\|_2^2 + \|w_2\|_2^2 \|G\|_2^2]
\end{aligned}$$
where in the last line we used (49) and the AM-GM inequality ($ab \leq a^2/2 + b^2/2$). Rearranging gives the inequality
$$\begin{aligned}
\|w_2\|_2^2 &\leq \frac{(1-\rho)^{-1}\langle w_2, H\rangle^2 - \|\xi - Z_1 w_1\|_2^2}{\|G\|_2^2} \\
&\leq \frac{(1-\rho)^{-1}\langle w_2, H\rangle^2 - (1-\rho)[\|\xi\|_2^2 + \|Z_1 w_1\|_2^2]}{\|G\|_2^2} \\
&\leq \frac{(1-\rho)^{-1}\langle w_2, H\rangle^2 - (1-\rho)[\|\xi\|_2^2 + \|Z_1 w_1\|_2^2]}{(1-\alpha)^2 n} \\
&\leq -\frac{(1-\gamma)(1-\rho)}{(1-\alpha)^2}\|w_1\|_2^2 + \frac{(1-\rho)^{-1}\langle w_2, H\rangle^2 - (1-\rho)\|\xi\|_2^2}{(1-\alpha)^2 n}
\end{aligned}$$
where in the second inequality we used (50) and the AM-GM inequality again, in the third inequality we used (52) and in the last inequality we used (51). This shows
$$\begin{aligned}
(1-\gamma)(1-\rho)(\|w_1\|_2^2 + \|w_2\|_2^2) &\leq \frac{(1-\gamma)(1-\rho)}{(1-\alpha)^2}\|w_1\|_2^2 + \|w_2\|_2^2 \\
&\leq \frac{(1-\rho)^{-1}\langle w_2, H\rangle^2 - (1-\rho)\|\xi\|_2^2}{(1-\alpha)^2 n}.
\end{aligned}$$
Dividing through by the first two factors on the left hand side and plugging in (53) gives
$$\begin{aligned}
\|w_1\|_2^2 + \|w_2\|_2^2 &\leq \frac{(1-\rho)^{-2}\langle w_2, H\rangle^2 - \|\xi\|_2^2}{(1-\gamma)(1-\alpha)^2 n} \\
&\leq \frac{1}{(1-\gamma)(1-\alpha)^2(1-\rho)^2}\frac{\langle w_2, H\rangle^2}{n} - \frac{\sigma^2}{1-\gamma}.
\end{aligned}$$
We can simplify the first term by defining $\beta = (1-\gamma)^{-1}(1-\alpha)^{-2}(1-\rho)^{-2} - 1$ and the second term by observing $-\frac{\sigma^2}{1-\gamma} \leq -\sigma^2$. Finally, plugging into (43) gives
$$\begin{aligned}
\phi &\leq \max_{(w_1,w_2)\in S}(1+\beta)\frac{\langle w_2, H\rangle^2}{n} - \sigma^2 \\
&= \frac{1+\beta}{n}\max_{w_2 \in \Sigma_2^{1/2}(\mathcal{K} - w^*)} |\langle w_2, H\rangle|^2 - \sigma^2 \\
&\leq \frac{1+\beta}{n}\left(\max_{w_2 \in \Sigma_2^{1/2}\mathcal{K}} |\langle w_2, H\rangle| + |\langle \Sigma_2^{1/2} w^*, H\rangle|\right)^2 - \sigma^2
\end{aligned}$$

by the triangle inequality, and (48) follows by (54) and (55). To deduce the explicit bound for β, first use that

$$(1-\alpha)^2 = 1 - 2\alpha + \alpha^2 \geq 1 - 2\alpha$$

and similarly $(1-\rho)^2 \geq 1 - 2\rho$ to show

$$\frac{1}{(1-\gamma)(1-\alpha)^2(1-\rho)^2} \leq \frac{1}{(1-\gamma)(1-2\rho)(1-2\alpha)}.$$

If $\gamma, \rho < 1/2$, then

$$\begin{aligned}(1-\gamma)(1-2\alpha)(1-2\rho) &= 1 - \gamma - 2\alpha - 2\rho + 2\gamma\alpha + 2\gamma\rho + 4\alpha\rho - 4\gamma\alpha\rho \\ &\geq 1 - \gamma - 2\alpha - 2\rho - 4\gamma\alpha\rho > 1 - 2\gamma - 2\alpha - 2\rho.\end{aligned}$$

Provided that $2\gamma + 2\alpha + 2\rho < 1/2$ (which implies that $\gamma, \rho < 1/2$), we can use the inequality $(1-x)^{-1} \leq 1 + 2x$ for $x \in [0, 1/2]$ to show that

$$\frac{1}{(1-\gamma)(1-\alpha)^2(1-\rho)^2} \leq \frac{1}{1 - 2\gamma - 2\alpha - 2\rho} \leq 1 + 4\gamma + 4\alpha + 4\rho$$

and thus we can choose

$$\beta = 33\sqrt{\frac{\log(32/\delta)}{n}} + 18\sqrt{\frac{\operatorname{rank}(\Sigma_1)}{n}} \geq 4\gamma + 4\alpha + 4\rho \qquad \square$$

We are finally ready to prove our main generalization bound:

Theorem 1 (Main generalization bound)**.** *There exists an absolute constant $C_1 \leq 66$ such that the following is true. Under the model assumptions in (1), let $\mathcal{K}$ be an arbitrary compact set, and take any covariance splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$. Fixing $\delta \leq 1/4$, let $\beta = C_1\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\operatorname{rank}(\Sigma_1)}{n}}\right)$. If n is large enough that $\beta \leq 1$, then the following holds with probability at least $1 - \delta$:*

$$\sup_{w \in \mathcal{K}, \hat{L}(w)=0} L(w) \leq \frac{1+\beta}{n}\left[W(\Sigma_2^{1/2}\mathcal{K}) + \left(\operatorname{rad}(\Sigma_2^{1/2}\mathcal{K}) + \|w^*\|_{\Sigma_2}\right)\sqrt{2\log\left(\frac{32}{\delta}\right)}\right]^2.$$

Proof. By Lemmas 3 and 4, we show that for any t

$$\Pr\left(\max_{w \in \mathcal{K}, \hat{L}(w)=0} L(w) - \hat{L}(w) > t\right) = \Pr(\Phi > t - \sigma^2) \leq 2\Pr(\phi \geq t - \sigma^2).$$

By Lemma 5, the above is upper bounded by δ if we set $t - \sigma^2$ according to (48) with δ replaced by $\delta/2$. Observe that the σ^2 term cancels, and the proof is complete. $\square$

Remark 2 (Translation-invariant version)**.** Our generalization guarantee is stated in terms of $W(\cdot)$ and $\operatorname{rad}(\cdot)$, which are not translation-invariant. However, the generalization guarantee of Theorem 1 can be made translation invariant, e.g. replacing $W(\Sigma_2^{1/2}\mathcal{K})$ by $W(\Sigma_2^{1/2}(\mathcal{K} - a))$ for an arbitrary $a \in \mathbb{R}^d$, by recentering the problem before applying Theorem 1, i.e. by subtracting Xa from both sides of the interpolation constraint $Xw = Xw^* + \xi$.

We also note that in Theorem 1, there is no requirement that $w^* \in \mathcal{K}$, so the true function may not necessarily lie in the class even if there is no noise ($\sigma = 0$).

B.2 Specialization to General Norm Balls

For convenience, we restate the general definition of effective rank.

Definition 5. The effective $\|\cdot\|$-ranks of a covariance matrix Σ are given as follows. Let $H \sim N(0, I_d)$, and define $v^* = \arg\min_{v \in \partial\|\Sigma^{1/2}H\|_*} \|v\|_\Sigma$. Then

$$r_{\|\cdot\|}(\Sigma) = \left(\frac{\mathbb{E}\left\|\Sigma^{1/2}H\right\|_*}{\sup_{\|w\| \leq 1}\|w\|_\Sigma}\right)^2 \quad \text{and} \quad R_{\|\cdot\|}(\Sigma) = \left(\frac{\mathbb{E}\left\|\Sigma^{1/2}H\right\|_*}{\mathbb{E}\|v^*\|_\Sigma}\right)^2.$$

Applying Theorem 1 to an arbitrary norm ball yield the following:

Corollary 3. *There exists an absolute constant $C_1 \le 66$ such that the following is true. Under the model assumptions in (1), take any covariance splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$ and let $\|\cdot\|$ be an arbitrary norm. Fixing $\delta \le 1/4$, let $\gamma = C_1\left(\sqrt{\frac{\log(1/\delta)}{r_{\|\cdot\|}(\Sigma_2)}} + \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\mathrm{rank}(\Sigma_1)}{n}}\right)$. If $B \ge \|w^*\|$ and n is large enough that $\gamma \le 1$, then the following holds with probability at least $1-\delta$:*

$$\sup_{\|w\| \le B, \hat{L}(w)=0} L(w) \le (1+\gamma)\frac{\left(B \cdot \mathbb{E}\|\Sigma_2^{1/2} H\|_*\right)^2}{n}. \tag{9}$$

Proof. Let $\mathcal{K} = \{w : \|w\| \le B\}$ in Theorem 1. It is easy to see that

$$W(\Sigma_2^{1/2}\mathcal{K}) = \mathbb{E}\sup_{\|w\|\le B} |\langle \Sigma_2^{1/2} w, H\rangle| = \mathbb{E}\sup_{\|w\|\le B} |\langle w, \Sigma_2^{1/2} H\rangle| = B\,\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*$$

and

$$R(\Sigma_2^{1/2}\mathcal{K}) = \sup_{\|w\|\le B} \|\Sigma_2^{1/2} w\|_2 = B \sup_{\|w\|\le 1} \|w\|_{\Sigma_2}.$$

From our definition, it is clear that

$$r_{\|\cdot\|}(\Sigma) = \left(\frac{W(\Sigma^{1/2}\mathcal{K})}{R(\Sigma^{1/2}\mathcal{K})}\right)^2.$$

Observe that

$$\|w^*\|_{\Sigma_2} \le \|w^*\| \sup_{\|w\|\le 1} \|w\|_{\Sigma_2} \le B \sup_{\|w\|\le 1} \|w\|_{\Sigma_2} = R(\Sigma_2^{1/2}\mathcal{K}).$$

The two above equations imply

$$\begin{aligned}
&W(\Sigma_2^{1/2}\mathcal{K}) + \mathrm{rad}(\Sigma_2^{1/2}\mathcal{K})\sqrt{2\log(32/\delta)} + \|w^*\|_{\Sigma_2}\sqrt{2\log(32/\delta)} \\
\le\, &W(\Sigma_2^{1/2}\mathcal{K}) + 2\sqrt{2\log(32/\delta)}\,\mathrm{rad}(\Sigma_2^{1/2}\mathcal{K}) \\
=\, &W(\Sigma_2^{1/2}\mathcal{K}) + 2\sqrt{\frac{2\log(32/\delta)}{r_{\|\cdot\|}(\Sigma_2)}}\,W(\Sigma_2^{1/2}\mathcal{K}) \\
=\, &\left(1 + 2\sqrt{\frac{2\log(32/\delta)}{r_{\|\cdot\|}(\Sigma_2)}}\right)\left(B\,\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*\right).
\end{aligned}$$

Under our assumptions that $\gamma \le 1$ and $\delta \le 1/4$, using the inequality $(1+x)(1+y) \le 1 + x + 2y$ for $x \le 1$, it is routine to check that

$$(1+\beta)\left(1 + 2\sqrt{\frac{2\log(32/\delta)}{r_{\|\cdot\|}(\Sigma_2)}}\right)^2 \le 1 + \gamma.$$

Plugging into Theorem 1 concludes the proof. □

B.3 Special Case: Euclidean Norm

In the Euclidean setting, the effective ranks are defined as follows:

Definition 3 (Bartlett et al. 2020)**.** The *effective ranks* of a covariance matrix Σ are

$$r(\Sigma) = \frac{\mathrm{Tr}(\Sigma)}{\|\Sigma\|_{op}} \quad \text{and} \quad R(\Sigma) = \frac{\mathrm{Tr}(\Sigma)^2}{\mathrm{Tr}(\Sigma^2)}.$$

Due to the small difference between $r(\Sigma)$ and $r_{\|\cdot\|_2}(\Sigma)$, our generalization bound below requires a slightly different proof (see discussion in Section 5), but the proof strategies are exactly the same.

Corollary 2. *There exists an absolute constant $C_1 \le 66$ such that the following is true. Under (1), pick any split $\Sigma = \Sigma_1 \oplus \Sigma_2$, fix $\delta \le 1/4$, and let $\gamma = C_1\left(\sqrt{\frac{\log(1/\delta)}{r(\Sigma_2)}} + \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\mathrm{rank}(\Sigma_1)}{n}}\right)$. If $B \ge \|w^*\|_2$ and n is large enough that $\gamma \le 1$, the following holds with probability at least $1-\delta$:*

$$\sup_{\|w\|_2 \le B, \hat{L}(w)=0} L(w) \le (1+\gamma)\frac{B^2 \operatorname{Tr}(\Sigma_2)}{n}. \tag{5}$$

Proof. The proof is identical to Corollary 3 except for the inconsequential difference between $\mathbb{E}\,\|\Sigma_2^{1/2} g\|_2$ and $\operatorname{Tr}(\Sigma_2)^{1/2}$. It is easy to see that

$$W(\Sigma_2^{1/2}\mathcal{K}) \le B\operatorname{Tr}(\Sigma_2)^{1/2} \quad \text{and} \quad R(\Sigma_2^{1/2}\mathcal{K}) = B\|\Sigma_2\|_{op}^{1/2}.$$

By the same argument, we can show that $\|w^*\|_{\Sigma_2} \le R(\Sigma_2^{1/2}\mathcal{K})$ and

$$\begin{aligned}
&W(\Sigma_2^{1/2}\mathcal{K}) + \mathrm{rad}(\Sigma_2^{1/2}\mathcal{K})\sqrt{2\log(32/\delta)} + \|w^*\|_{\Sigma_2}\sqrt{2\log(32/\delta)} \\
&\le W(\Sigma_2^{1/2}\mathcal{K}) + 2\sqrt{2\log(32/\delta)}\,\mathrm{rad}(\Sigma_2^{1/2}\mathcal{K}) \\
&\le B\operatorname{Tr}(\Sigma_2)^{1/2} + 2\sqrt{2\log(32/\delta)}B\|\Sigma_2\|_{op}^{1/2} \\
&= \left(1 + 2\sqrt{\frac{2\log(32/\delta)}{r(\Sigma_2)}}\right) B\operatorname{Tr}(\Sigma_2)^{1/2}.
\end{aligned}$$

Plugging into Theorem 1 concludes the proof. □

Next, by choosing a particular covariance split, we prove the speculative bound from Zhou et al. (2020) when the features are Gaussian:

Corollary 1 (Proof of the speculative bound ($\star$) for Gaussian data)**.** *Fix any $\delta \le 1/4$. Under the model assumptions in (1) with $B \ge \|w^*\|_2$ and $n \gtrsim \log(1/\delta)$, for some $\gamma \lesssim \sqrt[4]{\log(1/\delta)/n}$, it holds with probability at least $1-\delta$ that*

$$\sup_{\|w\|_2 \le B, \hat{L}(w)=0} L(w) \le (1+\gamma)\frac{B^2 \operatorname{Tr}(\Sigma)}{n}. \tag{4}$$

Proof. By Theorem 1 and the same argument in proof of Corollary 2, we obtain

$$\begin{aligned}
\sup_{w \in \mathcal{K}, Y = Xw} L(w) &\le \frac{1+\beta}{n}\left(B\operatorname{Tr}(\Sigma_2)^{1/2} + B\|\Sigma_2\|_{op}^{1/2} \cdot 2\sqrt{2\log\left(\frac{32}{\delta}\right)}\right)^2 \\
&\le \frac{B^2}{n}(1+\beta)\left(\operatorname{Tr}(\Sigma)^{1/2} + \|\Sigma_2\|_{op}^{1/2} \cdot 6\sqrt{\log(1/\delta)}\right)^2.
\end{aligned}$$

Let Σ_1 contain the largest eigenvalues, then we have

$$\mathrm{rank}(\Sigma_1)\|\Sigma_2\|_{op} \le \operatorname{Tr}(\Sigma).$$

Plugging in the inequality shows

$$\sup_{w \in \mathcal{K}, Y = Xw} L(w) \le \frac{B^2 \operatorname{Tr}(\Sigma)}{n}(1+\beta)\left(1 + 6\sqrt{\frac{\log(1/\delta)}{\mathrm{rank}(\Sigma_1)}}\right)^2.$$

Therefore, we can pick $\gamma = (1+\beta)\left(1 + 6\sqrt{\frac{\log(1/\delta)}{\mathrm{rank}(\Sigma_1)}}\right)^2 - 1$ and it is clear that

$$\gamma \lesssim \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\mathrm{rank}(\Sigma_1)}{n}} + \sqrt{\frac{\log(1/\delta)}{\mathrm{rank}(\Sigma_1)}}$$

for sufficiently large n and $\mathrm{rank}(\Sigma_1)$. To balance the last two terms, we can pick a covariance split such that $\mathrm{rank}(\Sigma_1)$ is of order $[n\log(1/\delta)]^{1/2}$, which proves the $\sqrt[4]{\log(1/\delta)/n}$ rate. □

C Bounds on the Norm of the Minimal-Norm Interpolator

In this section, we will give bounds – again based on the Gaussian Minimax Theorem – for the norm of the minimal norm interpolator, first in general and then in the Euclidean case.

C.1 General Norms: Proof of Theorem 4

Similar to the analysis in the previous section, we first formulate the minimal norm as an optimization problem in terms of a random matrix with $N(0,1)$ entries. Next, we apply the Convex Gaussian Minimax Theorem.

Lemma 6. *Under the model assumptions in (1), let $\|\cdot\|$ be an arbitrary norm and $Z : n \times d$ be a matrix with i.i.d. $N(0,1)$ entries independent of ξ. Define the primary optimization problem (PO) as*

$$\Phi := \min_{Zw=\xi} \|\Sigma^{-1/2} w\|. \tag{56}$$

Then for any t, it holds that

$$\Pr\left(\min_{Xw=Y} \|w\| > t \right) \le \Pr\left(\|w^*\| + \Phi > t \right). \tag{57}$$

Proof. By equality in distribution, we can write $X = Z\Sigma^{1/2}$. By the triangle inequality and two changes of variables, we have

$$\begin{aligned} \min_{Xw=Y} \|w\| &= \min_{Xw=\xi} \|w + w^*\| \\ &\le \|w^*\| + \min_{Z\Sigma^{1/2}w=\xi} \|w\| \\ &= \|w^*\| + \min_{Zw=\xi} \|\Sigma^{-1/2} w\|. \end{aligned}$$

□

Lemma 7 (Application of CGMT)**.** *In the same setting as Lemma 6, let $G \sim N(0, I_n)$, $H \sim N(0, I_d)$ be Gaussian vectors independent of ξ and each other. Define the auxiliary optimization problem (AO) as*

$$\phi := \min_{\|\xi - \|w\|_2 G\|_2 \le \langle H, w \rangle} \|\Sigma^{-1/2} w\|. \tag{58}$$

Then it holds that

$$\Pr(\Phi > t \,|\, \xi) \le 2 \Pr(\phi \ge t \,|\, \xi), \tag{59}$$

and taking expectations we have

$$\Pr(\Phi > t) \le 2 \Pr(\phi \ge t). \tag{60}$$

Proof. By introducing Lagrange multipliers, we have

$$\begin{aligned} \Phi &= \min_w \max_\lambda \|\Sigma^{-1/2} w\| + \langle \lambda, Zw - \xi \rangle \\ &= \min_w \max_\lambda \langle \lambda, Zw \rangle + \|\Sigma^{-1/2} w\| - \langle \lambda, \xi \rangle. \end{aligned}$$

By independence, the distribution of Z remains the same after conditioning on ξ and the randomness in Φ comes solely from Z. Therefore, we can apply CGMT in Theorem 9 with $\psi(w, \lambda) = \|\Sigma^{-1/2} w\| - \langle \lambda, \xi \rangle$ because ψ is *convex-concave*, but we again have the technical difficulty that the domains of w and λ are not compact. To overcome this, we will use a double truncation argument. For any $r, t > 0$, we define

$$\Phi_r(t) := \min_{\|\Sigma^{-1/2} w\| \le 2t} \max_{\|\lambda\|_2 \le r} \langle \lambda, Zw \rangle + \|\Sigma^{-1/2} w\| - \langle \lambda, \xi \rangle \tag{61}$$

and the corresponding AO

$$\begin{aligned} \phi_r(t) :&= \min_{\|\Sigma^{-1/2} w\| \le 2t} \max_{\|\lambda\|_2 \le r} \|w\|_2 \langle G, \lambda \rangle + \|\lambda\|_2 \langle H, w \rangle + \|\Sigma^{-1/2} w\| - \langle \lambda, \xi \rangle \\ &= \min_{\|\Sigma^{-1/2} w\| \le 2t} \max_{\|\lambda\|_2 \le r} \|\lambda\|_2 \langle H, w \rangle - \langle \lambda, \xi - \|w\|_2 G \rangle + \|\Sigma^{-1/2} w\| \\ &= \min_{\|\Sigma^{-1/2} w\| \le 2t} \max_{0 \le \lambda \le r} \lambda \left(\langle H, w \rangle + \|\xi - \|w\|_2 G\|_2 \right) + \|\Sigma^{-1/2} w\|. \end{aligned} \tag{62}$$

Note that the optimization in $\Phi_r(t)$ and $\phi_r(t)$ now ranges over compact sets. We will also use an intermediate problem between Φ and $\Phi_r(t)$, defined as

$$\begin{aligned}\Phi(t) :&= \min_{\|\Sigma^{-1/2}w\|\le 2t} \max_{\lambda} \langle \lambda, Zw\rangle + \|\Sigma^{-1/2}w\| - \langle \lambda, \xi\rangle \\ &= \min_{\substack{Zw=\xi \\ \|\Sigma^{-1/2}w\|\le 2t}} \|\Sigma^{-1/2}w\|. \end{aligned} \tag{63}$$

We similarly define the intermediate AO as

$$\begin{aligned}\phi(t) :&= \min_{\|\Sigma^{-1/2}w\|\le 2t} \max_{\lambda\ge 0} \lambda\left(\langle H, w\rangle + \|\xi - \|w\|_2 G\|_2\right) + \|\Sigma^{-1/2}w\| \\ &= \min_{\substack{\|\xi-\|w\|_2 G\|_2 \le \langle -H, w\rangle \\ \|\Sigma^{-1/2}w\|\le 2t}} \|\Sigma^{-1/2}w\|. \end{aligned} \tag{64}$$

Compared to the definition of ϕ, we have $\langle -H, w\rangle$ instead of $\langle H, w\rangle$, but this difference is negligible because H is Gaussian. It can be easily seen that the event $\Phi > t$ is the same as $\Phi(t) > t$, and the same holds for ϕ and $\phi(t)$. It is also clear that $\phi(t) \ge \phi_r(t)$ and we can connect $\phi_r(t)$ with $\Phi_r(t)$ by CGMT. It remains to show that $\Phi_r(t) \to \Phi(t)$ as $r \to \infty$.

By definition, $\Phi_r(t) \le \Phi_s(t)$ for $r \le s$. We consider two cases:

1. $\Phi(t) = \infty$, i.e. the minimization problem defining $\Phi(t)$ is infeasible. In this case, we know that for all $\|\Sigma^{-1/2}w\| \le 2t$
$$\|Zw - \xi\|_2 > 0.$$
By compactness, there exists $\mu = \mu(Z, \xi) > 0$ (in particular, independent of r) such that
$$\|Zw - \xi\|_2 \ge \mu.$$
Therefore, considering λ along the direction of $Zw - \xi$ shows that
$$\Phi_r(t) = \min_{\|\Sigma^{-1/2}w\|\le 2t} \max_{\|\lambda\|_2 \le r} \langle \lambda, Zw - \xi\rangle + \|\Sigma^{-1/2}w\| \ge r\mu$$
so $\Phi_r(t) \to \infty$ as $r \to \infty$.

2. Otherwise $\Phi(t) < \infty$, i.e. the minimization problem defining $\Phi(t)$ is feasible. In this case, we can let $w(r)$ be an arbitrary minimizer achieving the objective $\Phi_r(t)$ for each $r \ge 0$ by compactness. By compactness again, the sequence $\{w(r)\}_{r=1}^{\infty}$ at positive integer values of r has a subsequential limit $w(\infty)$ such that $\|\Sigma^{-1/2}w(\infty)\| \le 2t$. Equivalently, there exists an increasing sequence r_n such that $\lim_{n\to\infty} w(r_n) = w(\infty)$.

   Suppose for the sake of contradiction that $Zw(\infty) \ne \xi$, then by continuity, there exists $\mu > 0$ and a sufficiently small $\epsilon > 0$ such that for all $\|w - w(\infty)\|_2 \le \epsilon$
$$\|Zw - \xi\|_2 \ge \mu.$$
This implies that for sufficiently large n, we have
$$\|Zw(r_n) - \xi\|_2 \ge \mu$$
and by the same argument as in the previous case
$$\Phi_{r_n}(t) = \max_{\|\lambda\|_2 \le r} \langle \lambda, Zw(r_n) - \xi\rangle + \|\Sigma^{-1/2}w(r_n)\| \ge r\mu$$
so $\Phi_{r_n} \to \infty$, but this is impossible since $\Phi_r(t) \le \Phi(t) < \infty$. By contradiction, it must be the case that $Zw(\infty) = \xi$. By taking $\lambda = 0$ in the definition of $\Phi_r(t)$, we have
$$\Phi_{r_n}(t) \ge \|\Sigma^{-1/2}w(r_n)\|.$$
By continuity, we show that
$$\lim\inf_{n\to\infty} \Phi_{r_n}(t) \ge \lim_{n\to\infty} \|\Sigma^{-1/2}w(r_n)\| = \|\Sigma^{-1/2}w(\infty)\| \ge \Phi(t).$$
Since $\Phi_{r_n}(t) \le \Phi(t)$, the limit of $\Phi_{r_n}(t)$ exists and equals $\Phi(t)$. We can conclude that $\lim_{r\to\infty} \Phi_r(t) = \Phi(t)$ because $\Phi_r(t)$ is an increasing function of r.

By the last part of Theorem 9 (the CGMT),

$$\Pr(\Phi_r(t) > t \,|\, \xi) \le 2\Pr(\phi_r(t) \ge t \,|\, \xi).$$

By continuity (from below) of the probability measure, and the fact that $\Phi_r(t)$ monotonically increases to $\Phi(t)$ almost surely, we can conclude

$$\begin{aligned}\Pr(\Phi > t \,|\, \xi) &= \Pr(\Phi(t) > t \,|\, \xi) \le \Pr\left(\cup_r \cap_{r' \ge r} \Phi_{r'}(t) > t \,|\, \xi\right) \\ &= \lim_{r\to\infty} \Pr\left(\cap_{r' \ge r} \Phi_{r'}(t) > t \,|\, \xi\right) = \lim_{r\to\infty} \Pr\left(\Phi_r(t) > t \,|\, \xi\right) \\ &\le 2 \lim_{r\to\infty} \Pr(\phi_r(t) \ge t \,|\, \xi) \le 2\Pr(\phi(t) \ge t \,|\, \xi) \\ &= 2\Pr(\phi \ge t \,|\, \xi).\end{aligned}$$

□

It remains to analyze the auxiliary problem, which we do in the following lemma:

Lemma 8. *For any covariance splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$, denote P as the orthogonal projection matrix onto the space spanned by Σ_2, and let $v^* = \arg\min_{v \in \partial \|\Sigma_2^{1/2} H\|_*} \|v\|_{\Sigma_2}$. Assume that there exists $\epsilon_1, \epsilon_2 \ge 0$ such that with probability at least $1 - \delta/2$,*

$$\|v^*\|_{\Sigma_2} \le (1 + \epsilon_1)\, \mathbb{E}\, \|v^*\|_{\Sigma_2} \tag{65}$$

and

$$\|Pv^*\|^2 \le 1 + \epsilon_2. \tag{66}$$

Let

$$\epsilon = 8n^{-1/2} + 28\sqrt{\frac{\log(32/\delta)}{n}} + 8\sqrt{\frac{\log(8/\delta)}{r_{\|\cdot\|}(\Sigma)}} + 2(1+\epsilon_1)^2 \frac{n}{R_{\|\cdot\|}(\Sigma_2)} + 2\epsilon_2.$$

If n and the effective ranks are sufficiently large such that $\epsilon \le 1$, then with probability at least $1 - \delta$, it holds that

$$\phi^2 \le (1 + \epsilon)\, \sigma^2 \frac{n}{(\mathbb{E}\, \|\Sigma_2^{1/2} H\|_*)^2} \tag{67}$$

Proof. For notational simplicity, we define

$$\begin{aligned}\alpha &= 2\sqrt{\frac{\log(32/\delta)}{n}} \\ \rho &= \sqrt{\frac{1}{n}} + 2\sqrt{\frac{\log(16/\delta)}{n}}.\end{aligned}$$

By a union bound, the following collection of events occurs with probability at least $1 - \delta/2$:

1. (Approximate Orthogonality.) By Lemma 1, it holds that
$$|\langle \xi, G \rangle| < \|\xi\|_2 \, \|G\|_2 \, \rho. \tag{68}$$
2. (Typical Norm of G and ξ.) By Lemma 2, it holds that
$$-\alpha \le \frac{1}{\sqrt{n}} \|G\|_2 - 1 \le \alpha \tag{69}$$
and
$$-\alpha\sigma \le \frac{1}{\sqrt{n}} \|\xi\|_2 - \sigma \le \alpha\sigma. \tag{70}$$
3. (Typical Norm of $\Sigma_2^{1/2} H$.) By Theorem 6, it holds that
$$\begin{aligned}\|\Sigma_2^{1/2} H\|_* &\ge \mathbb{E}\, \|\Sigma_2^{1/2} H\|_* - \sup_{\|u\| \le 1} \|u\|_{\Sigma_2} \sqrt{2\log(8/\delta)} \\ &= \left(1 - \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Sigma)}}\right) \mathbb{E}\, \|\Sigma_2^{1/2} H\|_*,\end{aligned} \tag{71}$$
because $\|\Sigma_2^{1/2} H\|_*$ is a $\sup_{\|u\| \le 1} \|u\|_{\Sigma_2}$-Lipschitz function of H.

By a change of variables, recall that

$$\phi := \min_{\left\|\xi - \|\Sigma^{1/2}w\|_2 G\right\|_2 \le \langle H, \Sigma^{1/2}w\rangle} \|w\|.$$

Equations (68) to (70) imply that

$$\begin{aligned}\left\|\xi - \|\Sigma^{1/2}w\|_2 G\right\|_2^2 &= \|\xi\|_2^2 - 2\langle \xi, G\rangle \|\Sigma^{1/2}w\|_2 + \|\Sigma^{1/2}w\|_2^2 \|G\|_2^2 \\ &\le (1+\rho)\left(\|\xi\|_2^2 + \|\Sigma^{1/2}w\|_2^2\|G\|_2^2\right) \\ &\le (1+\rho)(1+\alpha)^2 n(\sigma^2 + \|\Sigma^{1/2}w\|_2^2).\end{aligned}$$

To upper bound ϕ, it suffices to construct a w that satisfies the constraint. Consider w of the form $s(Pv^*)$, then $\Sigma^{1/2}w = s\Sigma_2^{1/2}v^*$. Plugging in, it suffices to choose s such that

$$(1+\rho)(1+\alpha)^2 n(\sigma^2 + s^2\|\Sigma_2^{1/2}v^*\|_2^2) \le s^2\langle H, \Sigma_2^{1/2}v^*\rangle^2 = s^2\|\Sigma_2^{1/2}H\|_*^2.$$

Solving for s, we can choose

$$s^2 = \sigma^2\left(\frac{\|\Sigma_2^{1/2}H\|_*^2}{(1+\rho)(1+\alpha)^2 n} - \|v^*\|_{\Sigma_2}^2\right)^{-1}$$

given that it is positive. By (65) and (71), we have

$$\begin{aligned}&\frac{\|\Sigma_2^{1/2}H\|_*^2}{(1+\rho)(1+\alpha)^2 n} - \|v^*\|_{\Sigma_2}^2 \\ &\ge \frac{(\mathbb{E}\,\|\Sigma_2^{1/2}H\|_*)^2}{(1+\rho)(1+\alpha)^2 n}\left(1 - \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Sigma)}}\right)^2 - (1+\epsilon_1)^2(\mathbb{E}\,\|v^*\|_{\Sigma_2})^2 \\ &= \frac{(\mathbb{E}\,\|\Sigma_2^{1/2}H\|_*)^2}{n}\left(\frac{1}{(1+\rho)(1+\alpha)^2}\left(1 - 2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Sigma)}}\right) - (1+\epsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)}\right).\end{aligned}$$

If $\alpha < 1$, then

$$\begin{aligned}(1+\rho)(1+\alpha)^2 &= (1+\rho)(1+2\alpha+\alpha^2) \\ &\le (1+\rho)(1+3\alpha) = 1 + 3\alpha + \rho + 3\alpha\rho \\ &\le 1 + 3\alpha + 4\rho\end{aligned}$$

and using the inequality $(1+x)^{-1} \ge 1 - x$, we show

$$\begin{aligned}\frac{1}{(1+\rho)(1+\alpha)^2} &\ge 1 - ((1+\rho)(1+\alpha)^2 - 1) \\ &\ge 1 - (3\alpha + 4\rho).\end{aligned}$$

Therefore, we can conclude that

$$\begin{aligned}&\frac{1}{(1+\rho)(1+\alpha)^2}\left(1 - 2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Sigma)}}\right) - (1+\epsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)} \\ &\ge (1 - (3\alpha+4\rho))\left(1 - 2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Sigma)}}\right) - (1+\epsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)} \\ &\ge 1 - (3\alpha+4\rho) - 2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Sigma)}} - (1+\epsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)} \ge 1 - \epsilon'\end{aligned}$$

where we define

$$\epsilon' = 4n^{-1/2} + 14\sqrt{\frac{\log(32/\delta)}{n}} + 4\sqrt{\frac{\log(8/\delta)}{r_{\|\cdot\|}(\Sigma)}} + (1+\epsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)}.$$

Provided that $\epsilon' \le 1/2$ (which also guarantees that $\alpha < 1$ and our definition of s^2 is sensible), we can use the inequality $(1-x)^{-1} \le 1 + 2x$ for $x \in [0, 1/2]$ to show that

$$s^2 \le \sigma^2 \frac{n}{(\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*)^2} \frac{1}{1-\epsilon'} \le (1+2\epsilon')\,\sigma^2 \frac{n}{(\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*)^2}$$

and thus by (66)

$$\phi^2 \le s^2 \|Pv^*\|^2 \le (1+\epsilon_2)(1+2\epsilon')\,\sigma^2 \frac{n}{(\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*)^2} \le (1+\epsilon)\,\sigma^2 \frac{n}{(\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*)^2}$$

with $\epsilon = 2\epsilon' + 2\epsilon_2$. □

Finally, we are ready to prove our general norm bound.

Theorem 4 (General norm bound)**.** *There exists an absolute constant $C_2 \le 64$ such that the following is true. Under the model assumptions in (1) with any covariance split $\Sigma = \Sigma_1 \oplus \Sigma_2$, let $\|\cdot\|$ be an arbitrary norm, and fix $\delta \le 1/4$. Denote the ℓ_2 orthogonal projection matrix onto the space spanned by Σ_2 as P. Let $H \sim N(0, I_d)$, and let $v^* = \arg\min_{v \in \partial \|\Sigma_2^{1/2} H\|_*} \|v\|_{\Sigma_2}$. Suppose that there exist $\epsilon_1, \epsilon_2 \ge 0$ such that with probability at least $1 - \delta/4$*

$$\|v^*\|_{\Sigma_2} \le (1+\epsilon_1)\,\mathbb{E}\,\|v^*\|_{\Sigma_2} \qquad \textit{and} \qquad \|Pv^*\|^2 \le 1 + \epsilon_2; \tag{10}$$

let $\epsilon = C_2 \left(\sqrt{\frac{\log(1/\delta)}{r_{\|\cdot\|}(\Sigma_2)}} + \sqrt{\frac{\log(1/\delta)}{n}} + (1+\epsilon_1)^2 \frac{n}{R_{\|\cdot\|}(\Sigma_2)} + \epsilon_2 \right)$. Then if n and the effective ranks are large enough that $\epsilon \le 1$, with probability at least $1 - \delta$, it holds that

$$\|\hat{w}\| \le \|w^*\| + (1+\epsilon)^{1/2}\,\sigma \frac{\sqrt{n}}{\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*}. \tag{11}$$

Proof. By Lemmas 6 and 7, we show that for any t

$$\Pr(\|\hat{w}\| > t) \le \Pr(\Phi > t - \|w^*\|) \le 2\Pr(\phi \ge t - \|w^*\|).$$

By Lemma 8, the above is upper bounded by δ if we set $t - \|w^*\|$ according to (67) with δ replaced by $\delta/2$. Moving $\|w^*\|$ to the other side concludes the proof. □

C.2 Special Case: Euclidean Norm

Lemma 9. *For any covariance matrix Σ, it holds that*

$$\left(\mathbb{E}\,\|\Sigma^{1/2} H\|_2\right)^2 \ge \left(1 - \frac{1}{r(\Sigma)}\right) \mathrm{Tr}(\Sigma) \tag{72}$$

and

$$\frac{1}{\mathrm{Tr}(\Sigma)} \ge \left(1 - \sqrt{\frac{8}{r(\Sigma)}}\right) \mathbb{E}\left[\frac{1}{H^T \Sigma H}\right]. \tag{73}$$

As a result, it holds that

$$r(\Sigma) - 1 \le r_{\|\cdot\|_2}(\Sigma) \le r(\Sigma) \tag{74}$$

and

$$1 - \frac{4}{\sqrt{r(\Sigma)}} \le \frac{R_{\|\cdot\|_2}(\Sigma)}{R(\Sigma)} \le \left(1 - \sqrt{\frac{8}{r(\Sigma^2)}}\right)^{-1}. \tag{75}$$

Proof. Observe that if $f(H) = \|\Sigma^{1/2} H\|_2$, then it can easily be checked that

$$\|\nabla f\|_2^2 = \frac{\|\Sigma H\|_2^2}{\|\Sigma^{1/2} H\|_2^2} \le \|\Sigma\|_{op}$$

and so by the Gaussian Poincaré inequality (van Handel 2014, Corollary 2.27), we have

$$\begin{aligned}\mathrm{Tr}(\Sigma) = \mathbb{E}\,\|\Sigma^{1/2}H\|_2^2 &= (\mathbb{E}\,\|\Sigma^{1/2}H\|_2)^2 + \mathrm{Var}\,\|\Sigma^{1/2}H\|_2 \\ &\le (\mathbb{E}\,\|\Sigma^{1/2}H\|_2)^2 + \|\Sigma\|_{op} \\ &= (\mathbb{E}\,\|\Sigma^{1/2}H\|_2)^2 + \frac{\mathrm{Tr}(\Sigma)}{r(\Sigma)}.\end{aligned}$$

Rearranging the terms proves (72). To prove (73), without loss of generality assume that Σ is diagonal, with diagonal entries $\lambda_1 \ge \lambda_2 \ge ... \ge \lambda_d$. Observe that for any integer $\nu > 2$, we can pad Σ with 0's such that ν divides d, and we have

$$H^T\Sigma H = \sum_{i=1}^{d}\lambda_i H_i^2 \ge \sum_{i=1}^{d/\nu}\lambda_{\nu i}(H^2_{\nu(i-1)+1} + ... + H^2_{\nu i}).$$

By Jensen's inequality, $1/\,\mathbb{E}[X] \le \mathbb{E}[1/X]$; it follows that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{H^T\Sigma H}\right] &\le \mathbb{E}\left[\frac{1}{\sum_{i=1}^{d/\nu}\lambda_{\nu i}(H^2_{\nu(i-1)+1} + ... + H^2_{\nu i})}\right] \\ &= \frac{1}{\sum_{j=1}^{d/\nu}\lambda_{\nu j}}\,\mathbb{E}\left[\frac{1}{\sum_{i=1}^{d/\nu}\frac{\lambda_{\nu i}}{\sum_{j=1}^{d/\nu}\lambda_{\nu j}}(H^2_{\nu(i-1)+1} + ... + H^2_{\nu i})}\right] \\ &\le \frac{1}{\sum_{j=1}^{d/\nu}\lambda_{\nu j}}\,\mathbb{E}\left[\sum_{i=1}^{d/\nu}\frac{\lambda_{\nu i}}{\sum_{j=1}^{d/\nu}\lambda_{\nu j}}\frac{1}{H^2_{\nu(i-1)+1} + ... + H^2_{\nu i}}\right] \\ &= \frac{1}{\sum_{j=1}^{d/\nu}\lambda_{\nu j}}\frac{1}{\nu-2}.\end{aligned}$$

In the last equality, we use the fact that for each i the random variable $\left(H^2_{\nu(i-1)+1} + ... + H^2_{\nu i}\right)^{-1}$ follows an inverse Chi-square distribution with ν degrees of freedom; its expectation is $(\nu-2)^{-1}$. In addition, notice that

$$\nu\|\Sigma\|_{op} + \nu\sum_{i=1}^{d/\nu}\lambda_{\nu i} \ge (\lambda_1 + ... + \lambda_\nu) + \sum_{i=1}^{d/\nu-1}(\lambda_{\nu i+1} + ... + \lambda_{\nu(i+1)}) = \mathrm{Tr}(\Sigma).$$

Plugging the above estimate into our upper bound shows for any integer $\nu > 2$, it holds that

$$\mathbb{E}\left[\frac{1}{H^T\Sigma H}\right] \le \frac{1}{\mathrm{Tr}(\Sigma) - \nu\|\Sigma\|_{op}}\frac{\nu}{\nu-2} = \frac{1}{\mathrm{Tr}(\Sigma)}\left(1 - \frac{\nu}{r(\Sigma)} - \frac{2}{\nu} + \frac{2}{r(\Sigma)}\right)^{-1}.$$

We can show (73) by choosing $\nu = \lceil (2r(\Sigma))^{1/2}\rceil$:

$$\mathbb{E}\left[\frac{1}{H^T\Sigma H}\right] \le \frac{1}{\mathrm{Tr}(\Sigma)}\left(1 - \sqrt{\frac{8}{r(\Sigma)}}\right)^{-1}.$$

It remains to verify (74) and (75). By (72), we can check

$$r_{\|\cdot\|_2}(\Sigma) = \frac{(\mathbb{E}\,\|\Sigma^{1/2}H\|_2)^2}{\|\Sigma\|_{op}} \ge \left(1 - \frac{1}{r(\Sigma)}\right)\frac{\mathrm{Tr}(\Sigma)}{\|\Sigma\|_{op}} = r(\Sigma) - 1.$$

The other direction $r(\Sigma) \ge r_{\|\cdot\|_2}(\Sigma)$ follows directly from an application of the Cauchy-Schwarz inequality. By Jensen's inequality $1/\,\mathbb{E}[X] \le \mathbb{E}[1/X]$ and the Cauchy-Schwarz inequality, we show

$$\frac{1}{\mathrm{Tr}(\Sigma)}\left(\mathbb{E}\,\frac{1}{\|\Sigma H\|_2^2}\right)^{-1} \le \left(\mathbb{E}\,\frac{\|\Sigma^{1/2}H\|_2}{\|\Sigma H\|_2}\right)^{-2} \le \left(\mathbb{E}\,\frac{\|\Sigma H\|_2}{\|\Sigma^{1/2}H\|_2}\right)^{2} \le \mathrm{Tr}(\Sigma^2)\,\mathbb{E}\,\frac{1}{\|\Sigma^{1/2}H\|_2^2}.$$

Recall that $R_{\|\cdot\|_2}(\Sigma) = (\mathbb{E}\,\|\Sigma^{1/2}H\|_2)^2 \left(\mathbb{E}\,\frac{\|\Sigma H\|_2}{\|\Sigma^{1/2}H\|_2}\right)^{-2}$. By Cauchy-Schwarz inequality and (73), it follows that

$$R_{\|\cdot\|_2}(\Sigma) \le \mathrm{Tr}(\Sigma)^2 \left(\mathbb{E}\,\frac{1}{\|\Sigma H\|_2^2}\right) \le \left(1 - \sqrt{\frac{8}{r(\Sigma^2)}}\right)^{-1} R(\Sigma)$$

and also by (72)

$$\begin{aligned} R_{\|\cdot\|_2}(\Sigma) &\ge \left(1 - \frac{1}{r(\Sigma)}\right) \frac{\mathrm{Tr}(\Sigma)}{\mathrm{Tr}(\Sigma^2)} \left(\mathbb{E}\,\frac{1}{\|\Sigma^{1/2}H\|_2^2}\right)^{-1} \\ &\ge \left(1 - \frac{1}{r(\Sigma)}\right) \left(1 - \sqrt{\frac{8}{r(\Sigma)}}\right) R(\Sigma) \ge \left(1 - \frac{4}{\sqrt{r(\Sigma)}}\right) R(\Sigma). \end{aligned}$$

□

Lemma 10. *For any covariance matrix Σ, it holds that with probability at least $1-\delta$,*

$$1 - \frac{\|\Sigma^{1/2}H\|_2^2}{\mathrm{Tr}(\Sigma)} \lesssim \frac{\log(4/\delta)}{\sqrt{R(\Sigma)}} \tag{76}$$

and

$$\|\Sigma H\|_2^2 \lesssim \log(4/\delta)\,\mathrm{Tr}(\Sigma^2). \tag{77}$$

Therefore, provided that $R(\Sigma) \gtrsim \log(4/\delta)^2$, it holds that

$$\left(\frac{\|\Sigma H\|_2}{\|\Sigma^{1/2}H\|_2}\right)^2 \lesssim \log(4/\delta)\frac{\mathrm{Tr}(\Sigma^2)}{\mathrm{Tr}(\Sigma)}. \tag{78}$$

Proof. Because we are considering ℓ_2 norm and H is standard Gaussian, without loss of generality we can assume that Σ is diagonal and we denote the diagonals of Σ as $\lambda_1, ..., \lambda_d$. By the sub-exponential Bernstein inequality (Vershynin 2018, Corollary 2.8.3), we have with probability at least $1-\delta/2$

$$\left|\frac{\|\Sigma^{1/2}H\|_2^2}{\mathrm{Tr}(\Sigma)} - 1\right| = \left|\sum_{i=1}^p \frac{\lambda_i}{\sum_j \lambda_j}(H_i^2 - 1)\right| \lesssim \sqrt{\frac{\log(4/\delta)}{R(\Sigma)}} \vee \frac{\log(4/\delta)}{r(\Sigma)} \le \frac{\log(4/\delta)}{\sqrt{R(\Sigma)}}$$

where the last inequality uses that $R(\Sigma) \le r(\Sigma)^2$, shown in Lemma 5 of Bartlett et al. (2020). Using the sub-exponential Bernstein inequality again, we show with probability at least $1-\delta/2$

$$\left|\frac{\|\Sigma H\|_2^2}{\mathrm{Tr}(\Sigma^2)} - 1\right| \lesssim \sqrt{\frac{\log(4/\delta)}{R(\Sigma^2)}} \vee \frac{\log(4/\delta)}{r(\Sigma^2)}$$

From Lemma 5 of Bartlett et al. (2020), we know that the effective ranks are at least 1. This implies

$$\|\Sigma H\|_2^2 \lesssim \log(4/\delta)\,\mathrm{Tr}(\Sigma^2).$$

Provided that $R(\Sigma) \gtrsim \log(4/\delta)^2$, we have

$$\|\Sigma^{1/2}H\|_2^2 \ge \frac{1}{2}\,\mathrm{Tr}(\Sigma)$$

in which case it holds that

$$\frac{\|\Sigma H\|_2^2}{\|\Sigma^{1/2}H\|_2^2} \lesssim \log(4/\delta)\frac{\mathrm{Tr}(\Sigma^2)}{\mathrm{Tr}(\Sigma)}.$$

□

Theorem 2 (Euclidean norm bound; special case of Theorem 4)**.** *Fix any $\delta \le 1/4$. Under the model assumptions in (1) with any choice of covariance splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$, there exists some $\epsilon \lesssim \sqrt{\frac{\log(1/\delta)}{r(\Sigma_2)}} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{n\log(1/\delta)}{R(\Sigma_2)}$ such that the following is true. If n and the effective ranks are such that $\epsilon \le 1$ and $R(\Sigma_2) \gtrsim \log(1/\delta)^2$, then with probability at least $1-\delta$, it holds that*

$$\|\hat{w}\|_2 \le \|w^*\|_2 + (1+\epsilon)^{1/2}\,\sigma\sqrt{\frac{n}{\mathrm{Tr}(\Sigma_2)}}. \tag{6}$$

Proof. To apply Theorem 4, it is clear that $v^* = \frac{\Sigma_2^{1/2} H}{\|\Sigma_2^{1/2} H\|_2}$ and so $\|v^*\|_{\Sigma_2} = \frac{\|\Sigma_2 H\|_2}{\|\Sigma_2^{1/2} H\|_2}$. By (78), it suffices to pick ϵ_1 such that for some constant $c > 0$

$$(1+\epsilon_1)\,\mathbb{E}\,\|v^*\|_{\Sigma_2} = c\sqrt{\log(16/\delta)\frac{\mathrm{Tr}(\Sigma_2^2)}{\mathrm{Tr}(\Sigma_2)}}.$$

By (72) of Lemma 9, for sufficiently large effective rank, it holds that $\left(\mathbb{E}\,\|\Sigma_2^{1/2} H\|_2\right)^2 \gtrsim \mathrm{Tr}(\Sigma_2)$ and so

$$(1+\epsilon_1)^2 \frac{n}{R_{\|\cdot\|_2}(\Sigma_2)} = n\frac{(1+\epsilon_1)^2(\mathbb{E}\,\|v^*\|_{\Sigma_2})^2}{\left(\mathbb{E}\,\|\Sigma_2^{1/2} H\|_2\right)^2} \lesssim n\log(16/\delta)\frac{\mathrm{Tr}(\Sigma_2^2)}{\mathrm{Tr}(\Sigma_2)^2} = \frac{n\log(16/\delta)}{R(\Sigma_2)}.$$

Furthermore, it suffices to let $\epsilon_2 = 0$ because P is an ℓ_2 projection matrix. Combined with (74) of Lemma 9, we show

$$\epsilon \lesssim \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\log(1/\delta)}{r(\Sigma_2)}} + \frac{n\log(1/\delta)}{R(\Sigma_2)}.$$

Finally, using the inequality $(1-x)^{-1} \le 1 + 2x$ for $x \in [0, 1/2]$ and (72) of Lemma 9 again, we can conclude

$$\begin{aligned}(1+\epsilon)^{1/2}\,\sigma\frac{\sqrt{n}}{\mathbb{E}\,\|\Sigma_2^{1/2} H\|_2} &\le (1+\epsilon)^{1/2}\left(1 - \frac{1}{r(\Sigma_2)}\right)^{-1/2}\sigma\sqrt{\frac{n}{\mathrm{Tr}(\Sigma_2)}} \\ &\le \left(1 + 2\epsilon + \frac{2}{r(\Sigma_2)}\right)^{1/2}\sigma\sqrt{\frac{n}{\mathrm{Tr}(\Sigma_2)}}\end{aligned}$$

and we can replace ϵ with

$$\epsilon' = 2\epsilon + \frac{2}{r(\Sigma_2)} \lesssim \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\log(1/\delta)}{r(\Sigma_2)}} + \frac{n\log(1/\delta)}{R(\Sigma_2)}.$$

□

D Benign Overfitting

In this section, we will combine results from the previous two sections to study when interpolators are consistent.

D.1 General Norm

Theorem 5 (Benign overfitting with general norm)**.** *Fix any $\delta \le 1/2$. Under the model assumptions in (1), let $\|\cdot\|$ be an arbitrary norm and pick a covariance split $\Sigma = \Sigma_1 \oplus \Sigma_2$. Suppose that n and the effective ranks are sufficiently large such that $\gamma, \epsilon \le 1$ with the same choice of γ and ϵ as in Corollary 3 and Theorem 4. Then, with probability at least $1-\delta$,*

$$L(\hat{w}) \le (1+\gamma)(1+\epsilon)\left(\sigma + \|w^*\|\frac{\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*}{\sqrt{n}}\right)^2. \tag{12}$$

Proof. By Theorem 4, if we choose

$$B = \|w^*\| + (1+\epsilon)^{1/2}\,\sigma\frac{\sqrt{n}}{\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*}$$

then with large probability, $\{w : \|w\| \le B\}$ has non-empty intersection with $\{w : Xw = Y\}$, which contains the minimal norm interpolator $\hat{w}$. Also, it is clear that $B > \|w^*\|$ and so by Corollary 3, it holds that

$$
\begin{aligned}
L(\hat{w}) &\leq \sup_{\|w\| \leq B, \hat{L}(w)=0} L(w) \\
&\leq (1+\gamma)\left(\|w^*\| + (1+\epsilon)^{1/2}\,\sigma \frac{\sqrt{n}}{\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*}\right)^2 \frac{\left(\mathbb{E}\|\Sigma_2^{1/2} H\|_*\right)^2}{n} \\
&= (1+\gamma)\left(\|w^*\|\,\frac{\mathbb{E}\|\Sigma_2^{1/2} H\|_*}{\sqrt{n}} + (1+\epsilon)^{1/2}\,\sigma\right)^2 \\
&\leq (1+\gamma)(1+\epsilon)\left(\sigma + \|w^*\|\frac{\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*}{\sqrt{n}}\right)^2 .
\end{aligned}
$$

□

Theorem 11 (Sufficient conditions)**.** *Under the model assumptions in (1), let $\|\cdot\|$ be an arbitrary norm. Suppose that as n goes to ∞, there exists a sequence of covariance splits $\Sigma = \Sigma_1 \oplus \Sigma_2$ such that the following properties hold:*

1. *(Small large-variance dimension.)*
$$\lim_{n\to\infty} \frac{\mathrm{rank}(\Sigma_1)}{n} = 0. \tag{79}$$
2. *(Large effective dimension.)*
$$\lim_{n\to\infty} \frac{1}{r_{\|\cdot\|}(\Sigma_2)} = 0 \quad \textit{and} \quad \lim_{n\to\infty} \frac{n}{R_{\|\cdot\|}(\Sigma_2)} = 0. \tag{80}$$
3. *(No aliasing condition.)*
$$\lim_{n\to\infty} \frac{\|w^*\|\,\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*}{\sqrt{n}} = 0. \tag{81}$$
4. *(Contracting ℓ_2 projection condition.) With the same definition of P and v^* as in Theorem 4, it holds that for any $\eta > 0$,*
$$\lim_{n\to\infty} \Pr(\|P v^*\|^2 > 1 + \eta) = 0. \tag{82}$$

Then $L(\hat{w})$ converges to σ^2 in probability. In other words, minimum norm interpolation is consistent.

Proof. Fix any $\eta > 0$, for sufficiently small γ, ϵ and $\|w^*\|\frac{\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*}{\sqrt{n}}$, it is clear that

$$(1+\gamma)(1+\epsilon)\left(\sigma + \|w^*\|\frac{\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*}{\sqrt{n}}\right)^2 - \sigma^2 \leq \eta. \tag{83}$$

For any $\delta > 0$, by the definition of γ in Corollary 3 and our assumptions, the terms γ and $\|w^*\|\frac{\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*}{\sqrt{n}}$ can be made arbitrarily small for large enough n. Also by our assumption, ϵ_2 in the definition of ϵ in Theorem 4 can be arbitrarily small. Note that

$$\sqrt{\frac{n}{R_{\|\cdot\|}(\Sigma_2)}} = \mathbb{E}\left[\frac{\|v^*\|_{\Sigma_2}}{\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*/\sqrt{n}}\right]$$

converges to 0 by assumption. Then by Markov's inequality, for any $\eta' > 0$, it holds that for all sufficiently large n

$$\Pr\left(\frac{\|v^*\|_{\Sigma_2}}{\mathbb{E}\,\|\Sigma_2^{1/2} H\|_*/\sqrt{n}} > \sqrt{\eta'}\right) < \delta$$

and we can pick

$$(1+\epsilon_1)\,\mathbb{E}\,\|v^*\|_{\Sigma_2} = \sqrt{\eta'}\frac{\mathbb{E}\,\|\Sigma_2^{1/2}H\|_*}{\sqrt{n}}.$$

This implies that

$$(1+\epsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)} = \frac{n}{\left(\mathbb{E}\,\|\Sigma_2^{1/2}H\|_*\right)^2}\left((1+\epsilon_1)\,\mathbb{E}\,\|v^*\|_{\Sigma_2}\right)^2 = \eta'.$$

By Theorem 5, we have shown that for sufficiently large n such that γ, ϵ and $\|w^*\|\frac{\mathbb{E}\,\|\Sigma_2^{1/2}H\|_*}{\sqrt{n}}$ are small enough for (83) to hold, it holds that

$$\Pr(\,|L(\hat{w}) - \sigma^2| > \eta\,) \le \delta.$$

As a result, we show $\lim_{n\to\infty}\Pr(\,|L(\hat{w}) - \sigma^2| > \eta\,) \le \delta$ for any $\delta > 0$. To summarize, for any fixed $\eta > 0$, we have

$$\lim_{n\to\infty}\Pr(\,|L(\hat{w}) - \sigma^2| > \eta\,) = 0$$

and so $L(\hat{w})$ converges to σ^2 in probability. □

D.2 Euclidean Norm

Theorem 3 (Benign overfitting)**.** *Fix any $\delta \le 1/2$. Under the model assumptions in (1) with any covariance splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$, let γ and ϵ be as defined in Corollary 2 and Theorem 2. Suppose that n and the effective ranks are such that $R(\Sigma_2) \gtrsim \log(1/\delta)^2$ and $\gamma, \epsilon \le 1$. Then, with probability at least $1-\delta$,*

$$L(\hat{w}) \le (1+\gamma)(1+\epsilon)\left(\sigma + \|w^*\|_2\sqrt{\frac{\mathrm{Tr}(\Sigma_2)}{n}}\right)^2. \tag{7}$$

Proof. The proof follows the same strategy as Theorem 5. By Theorem 2, if we choose

$$B = \|w^*\|_2 + (1+\epsilon)^{1/2}\,\sigma\sqrt{\frac{n}{\mathrm{Tr}(\Sigma_2)}},$$

then with large probability, $\{w : \|w\|_2 \le B\}$ has non-empty intersection with $\{w : Xw = Y\}$. This intersection necessarily contains the minimal norm interpolator $\hat{w}$.

Also, it is clear that $B > \|w^*\|$ and so by Corollary 2, it holds that

$$\begin{aligned}
L(\hat{w}) &\le \sup_{\|w\|_2 \le B, \hat{L}(w)=0} L(w) \\
&\le (1+\gamma)\left(\|w^*\|_2 + (1+\epsilon)^{1/2}\,\sigma\sqrt{\frac{n}{\mathrm{Tr}(\Sigma_2)}}\right)^2\frac{\mathrm{Tr}(\Sigma_2)}{n} \\
&= (1+\gamma)\left(\|w^*\|_2\sqrt{\frac{\mathrm{Tr}(\Sigma_2)}{n}} + (1+\epsilon)^{1/2}\,\sigma\right)^2 \\
&\le (1+\gamma)(1+\epsilon)\left(\sigma + \|w^*\|_2\sqrt{\frac{\mathrm{Tr}(\Sigma_2)}{n}}\right)^2.
\end{aligned}$$

□

Theorem 12 (Sufficient conditions)**.** *Under the model assumptions in (1), let $\hat{w}$ be the minimal ℓ_2 norm interpolator. Suppose that as n goes to ∞, there exists a sequence of covariance splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$ such that the following conditions hold:*

1. *(Small large-variance dimension.)*

$$\lim_{n\to\infty}\frac{\mathrm{rank}(\Sigma_1)}{n} = 0. \tag{84}$$

2. *(Large effective dimension.)*

$$\lim_{n\to\infty}\frac{n}{R(\Sigma_2)} = 0. \tag{85}$$

3. (No aliasing condition.)

$$\lim_{n\to\infty} \frac{\|w^*\|_2 \, \mathbb{E}\, \|\Sigma_2^{1/2} H\|_2}{\sqrt{n}} = 0. \tag{86}$$

Then $L(\hat{w})$ converges to σ^2 in probability. In other words, minimum ℓ_2 norm interpolation is consistent.

Proof. Fix any $\eta > 0$, for sufficiently small γ, ϵ and $\|w^*\|_2 \sqrt{\frac{\mathrm{Tr}(\Sigma_2)}{n}}$, it is clear that

$$(1+\gamma)(1+\epsilon)\left(\sigma + \|w^*\|_2 \sqrt{\frac{\mathrm{Tr}(\Sigma_2)}{n}}\right)^2 - \sigma^2 \leq \eta. \tag{87}$$

From Lemma 5 of Bartlett et al. (2020), it holds that $R(\Sigma_2) \leq r(\Sigma_2)^2$, and so the condition $R(\Sigma_2) = \omega(n)$ implies that $r(\Sigma_2) = \omega(\sqrt{n}) = \omega(1)$. For any $\delta > 0$, by the definition of γ, ϵ in Corollary 2 and Theorem 2 and our assumptions, the terms γ, ϵ and $\|w^*\|_2 \sqrt{\frac{\mathrm{Tr}(\Sigma_2)}{n}}$ can be made small enough for Equation (87) to hold with a sufficiently large n. By Theorem 3, we show that

$$\lim_{n\to\infty} \Pr(\, |L(\hat{w}) - \sigma^2| > \eta \,) \leq \delta$$

Since the choice of $\delta > 0$ is arbitrary, we have shown that $L(\hat{w})$ converges to σ^2 in probability. □

D.2.1 Equivalence of Consistency Conditions

If we assume that $\|w^*\| = \Theta(1)$, our consistency condition (Theorem 12) for minimum ℓ_2 norm interpolation is the existence of a covariance splitting such that

$$\mathrm{rank}(\Sigma_1) = o(n), \qquad \mathrm{Tr}\,\Sigma_2 = o(n), \qquad \frac{(\mathrm{Tr}\,\Sigma_2)^2}{\mathrm{Tr}[(\Sigma_2)^2]} = \omega(n). \tag{88}$$

We compare the above conditions to the following conditions:

$$\mathrm{rank}(\Sigma_1) = o(n), \quad \mathrm{Tr}\,\Sigma_2 = o(n), \quad \frac{\mathrm{Tr}\,\Sigma_2}{\|\Sigma_2\|_{op}} = \omega(n), \quad \frac{(\mathrm{Tr}\,\Sigma_2)^2}{\mathrm{Tr}[(\Sigma_2)^2]} = \omega(n). \tag{89}$$

Obviously, the conditions in (89) imply (88), but we show in Theorem 13 that the existence of a splitting that satisfies (88) also implies the existence of a (potentially different) splitting that satisfies (89). This is one way to see that the particular choice of k^* from Bartlett et al. (2020) can be made without loss of generality, at least if we only consider the consistency conditions.

Theorem 13. *Suppose that there exists $\Sigma = \Sigma_1 \oplus \Sigma_2$ that satisfies the conditions in (88). Then there exists a $\Sigma = \Sigma_1' \oplus \Sigma_2'$ that satisfies the conditions in (89).*

Proof. Denote v as the vector of eigenvalues of Σ, and v_k as the vector obtained by setting the k coordinates of v corresponding to Σ_1 to be 0. By our assumptions in (88), there exists $k = o(n)$ such that

$$\|v_k\|_1 = o(n), \qquad \frac{\|v_k\|_1^2}{\|v_k\|_2^2} = \omega(n).$$

For any $\tau \geq 0$, we let $S_\tau = \{i \in [d] : |v_{k,i}| \geq \tau \|v_k\|_\infty\}$ and define $v_{k,\tau}$ by setting the coordinates of v_k in S_τ to be 0. For simplicity of notation, define $a = \|v_k\|_1^2 / \|v_k\|_2^2$ and $b = \|v_k\|_1 / \|v_k\|_\infty$. Observe that

$$\sum_{i \in S_\tau} |v_{k,i}| \leq \frac{1}{\tau \|v_k\|_\infty} \sum_{i \in S_\tau} v_{k,i}^2 \leq \frac{\|v_k\|_2^2}{\tau \|v_k\|_\infty} = \frac{\|v_k\|_1}{\tau} \frac{b}{a}.$$

This shows that

$$\|v_{k,\tau}\|_1 \geq \left(1 - \frac{b}{\tau a}\right) \|v_k\|_1$$

and

$$\|v_{k,\tau}\|_\infty \leq \tau \|v_k\|_\infty = \frac{\tau}{b} \|v_k\|_1.$$

In addition, observe that

$$\tau\|v_k\|_\infty \cdot |S_\tau| \le \sum_{i\in S_\tau} |v_{k,i}| \le \frac{\|v_k\|_1}{\tau}\frac{b}{a}.$$

The above inequalities imply that

$$\frac{\|v_{k,\tau}\|_1}{\|v_{k,\tau}\|_\infty} \ge \frac{b}{\tau}\left(1-\frac{b}{\tau a}\right)$$

and

$$|S_\tau| \le a\cdot\left(\frac{b}{\tau a}\right)^2$$

Finally, we pick τ by setting $b/(\tau a) = (n/a)^{3/4}$. By our assumption that $a = \omega(n)$, we can check

$$\frac{b}{\tau}\left(1-\frac{b}{\tau a}\right) = n^{3/4}a^{1/4}(1-(n/a)^{3/4}) = \omega(n)(1-o(1)) = \omega(n)$$

and

$$a\cdot\left(\frac{b}{\tau a}\right)^2 = a(n/a)^{3/2} = n(n/a)^{1/2} = o(n).$$

By Holder's inequality, we also have

$$\frac{\|v_{k,\tau}\|_1^2}{\|v_{k,\tau}\|_2^2} \ge \frac{\|v_{k,\tau}\|_1}{\|v_{k,\tau}\|_\infty} = \omega(n).$$

It is clear that $\|v_{k,\tau}\|_1 \le \|v_k\|_1 = o(n)$ and $k + |S_\tau| = o(n)$, so picking the covariance splitting that corresponds to $v_{k,\tau}$ concludes the proof. □

E Basis Pursuit (Minimum ℓ_1-Norm Interpolation)

In this section, we illustrate the consequences of our general theory for basis pursuit. The following generalization bound for basis pursuit follows immediately from Corollary 3:

Corollary 5 (Generalization bound for ℓ_1 norm balls)**.** *There exists an absolute constant $C_1 \le 66$ such that the following is true. Under the model assumptions in (1) with $\Sigma = \Sigma_1 \oplus \Sigma_2$, fix $\delta \le 1/4$ and let $\gamma = C_1\left(\sqrt{\frac{\log(1/\delta)}{r_1(\Sigma_2)}} + \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\mathrm{rank}(\Sigma_1)}{n}}\right)$. If $B \ge \|w^*\|_1$ and n is large enough that $\gamma \le 1$, then the following holds with probability at least $1-\delta$:*

$$\sup_{\|w\|_1\le B, \hat{L}(w)=0} L(w) \le (1+\gamma)\frac{\left(B\cdot\mathbb{E}\|\Sigma_2^{1/2}H\|_\infty\right)^2}{n}. \tag{90}$$

Proof. Recall that the dual of the ℓ_1 norm is the ℓ_∞ norm. By convexity

$$\max_{\|w\|_1\le 1}\|w\|_\Sigma = \sqrt{\max_i \langle e_i, \Sigma e_i\rangle} = \sqrt{\max_i \Sigma_{ii}}$$

and so we can use $r_1(\Sigma) = \frac{\left(\mathbb{E}\|\Sigma^{1/2}H\|_\infty\right)^2}{\max_i(\Sigma)_{ii}} = r_{\|\cdot\|_1}(\Sigma)$. □

The following norm bound for basis pursuit follows from Theorem 4:

Corollary 6 (ℓ_1 norm bound)**.** *There exists an absolute constant $C_2 \le 64$ such that the following is true. Under the model assumptions in (1), let $\Sigma = \Sigma_1 \oplus \Sigma_2$ such that Σ_2 is diagonal. Fix $\delta \le 1/4$ and let $\epsilon = C_2\left(\sqrt{\frac{\log(1/\delta)}{r_1(\Sigma_2)}} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{n}{r_1(\Sigma_2)}\right)$. Then if n and the effective rank $r_1(\Sigma_2)$ are large enough that $\epsilon \le 1$, with probability at least $1-\delta$, it holds that*

$$\|\hat{w}\|_1 \le \|w^*\|_1 + (1+\epsilon)^{1/2}\,\sigma\frac{\sqrt{n}}{\mathbb{E}\|\Sigma_2^{1/2}H\|_\infty}. \tag{91}$$

Proof. Recall that $\partial\|u\|_* = \text{conv}\{\text{sign}(u_i)\, e_i : i \in \arg\max |u_i|\}$, where $\text{conv}(S)$ denotes the convex hull of S. By definition, it holds almost surely that

$$\|v^*\|_{\Sigma_2} \le \max_{i\in[d]} \|e_i\|_\Sigma = \sqrt{\max_i \Sigma_{ii}}, \tag{92}$$

and so we can pick ϵ_1 such that

$$(1+\epsilon_1)\,\mathbb{E}\,\|v^*\|_{\Sigma_2} = \sqrt{\max_i \Sigma_{ii}}$$

and

$$(1+\epsilon_1)^2 \frac{n}{R_{\|\cdot\|_1}(\Sigma_2)} = n\frac{(1+\epsilon_1)^2(\mathbb{E}\,\|v^*\|_{\Sigma_2})^2}{\left(\mathbb{E}\,\|\Sigma_2^{1/2}H\|_\infty\right)^2} = \frac{n}{r_1(\Sigma_2)}.$$

In addition, since Σ_2 is diagonal, the coordinates of $\Sigma_2^{1/2}H$ that correspond to the zero diagonals of Σ_2 are 0. Therefore, v^* must also have zero entry in those coordinates. In other words, v^* lies in the span of Σ_2. As P is the orthogonal projection onto the space spanned by Σ_2, this implies $Pv^* = v^*$, and so $\|Pv^*\|_1 = \|v^*\|_1 = 1$, so that we can take $\epsilon_2 = 0$. Plugging ϵ_1, ϵ_2 into Theorem 4 concludes the proof. □

Theorem 14 (Benign overfitting)**.** *Fix any $\delta \le 1/2$. Under the model assumptions in (1), let $\Sigma = \Sigma_1 \oplus \Sigma_2$ such that Σ_2 is diagonal. Suppose that n and the effective rank $r_1(\Sigma_2)$ are sufficiently large such that $\gamma, \epsilon \le 1$ with the same choice of γ and ϵ as in Corollaries 5 and 6. Then, with probability at least $1-\delta$:*

$$L(\hat{w}) \le (1+\gamma)(1+\epsilon)\left(\sigma + \|w^*\|_1 \frac{\mathbb{E}\,\|\Sigma_2^{1/2}H\|_\infty}{\sqrt{n}}\right)^2. \tag{93}$$

The proof of Theorem 14 uses Corollaries 5 and 6, and follows the same lines as in Theorem 5. The details are repetitive, so we omit writing them out in full here. As before, we can use the finite sample bound to deduce sufficient conditions for consistency.

Theorem 15 (Sufficient conditions)**.** *Under the model assumptions in (1), let $\hat{w}$ be the minimal ℓ_1 norm interpolator. Suppose that as n goes to ∞, there exists a sequence of covariance splits $\Sigma = \Sigma_1 \oplus \Sigma_2$ such that Σ_2 is diagonal and the following conditions hold:*

1. *(Small large-variance dimension.)*
$$\lim_{n\to\infty} \frac{\text{rank}(\Sigma_1)}{n} = 0. \tag{94}$$
2. *(Large effective dimension.)*
$$\lim_{n\to\infty} \frac{n}{r_1(\Sigma_2)} = 0. \tag{95}$$
3. *(No aliasing condition.)*
$$\lim_{n\to\infty} \frac{\|w^*\|_1\,\mathbb{E}\,\|\Sigma_2^{1/2}H\|_\infty}{\sqrt{n}} = 0. \tag{96}$$

Then $L(\hat{w})$ converges to σ^2 in probability. In other words, minimum ℓ_1 norm interpolation is consistent.

Again, the proof of Theorem 15 is exactly analogous to Theorem 12, so we omit the full proof here.

E.1 Isotropic features

Theorem 16. *There exists an absolute constant $C_3 \le 140$ such that the following is true. Under the model assumptions in (1) with $\Sigma = I_d$, denote S as the support of w^*. Fix $\delta \le 1/4$ and let $\epsilon = C_3\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\log(1/\delta)}{\log(d-|S|)}} + \frac{n}{\log(d-|S|)}\right)$. Then if n and d are large enough that $\epsilon \le 1$, the following holds with probability $1-\delta$ where $H' \sim N(0, I_{d-|S|})$:*

$$\|\hat{w}\|_1 \le (1+\epsilon)^{1/2}\,(\sigma^2 + \|w^*\|_2^2)^{1/2}\,\frac{\sqrt{n}}{\mathbb{E}\,\|H'\|_\infty}. \tag{97}$$

Proof. Write $X = [X_S, X_{S^c}]$, where X_S is formed by selecting the columns of X in S. Also let $\xi' = X_S w_S^* + \xi$; then the entries of ξ' are i.i.d. $N(0, \sigma^2 + \|w^*\|_2^2)$ and independent of X_{S^c}. Observe that $Y = X_{S^c} 0 + \xi'$. By choosing $\Sigma_1 = 0$ in Corollary 6, we show with large probability

$$\min_{X_{S^c} w = Y} \|w\|_1 \le (1+\epsilon)^{1/2} (\sigma^2 + \|w^*\|_2^2)^{1/2} \frac{\sqrt{n}}{\mathbb{E}\,\|H'\|_\infty}$$

for some $\epsilon \le 64 \left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\log(1/\delta)}{r_1(I_{d-|S|})}} + \frac{n}{r_1(I_{d-|S|})} \right)$. By the bound of Kamath (2013), it holds that

$$r_1(I_{d-|S|}) = (\mathbb{E}\,\|H'\|_\infty)^2 \ge \frac{\log(d-|S|)}{\pi \log 2}$$

and so we can choose $C_3 \le 64\pi \log 2 < 140$. Observe that if $X_{S^c} w = Y$, then $X(0,w)^T = Y$ and $\|(0,w)\|_1 = \|w\|_1$. It follows that

$$\|\hat{w}\|_1 = \min_{Xw=Y} \|w\|_1 \le \min_{X_{S^c} w = Y} \|w\|_1. \qquad \square$$

Theorem 17. *Under the model assumptions in (1) with $\Sigma = I_d$, fix any $\delta \le 1/2$ and let $\eta = 368 \left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\log(1/\delta)+\log|S|}{\log(d-|S|)}} + \frac{n}{\log(d-|S|)} \right)$. Suppose that n and d are large enough that $\eta \le 1$. Then, with probability at least $1-\delta$,*

$$L(\hat{w}) \le (1+\eta)(\sigma^2 + \|w^*\|_2^2). \tag{98}$$

Proof. By Theorem 16, if we choose

$$B = (1+\epsilon)^{1/2} (\sigma^2 + \|w^*\|_2^2)^{1/2} \frac{\sqrt{n}}{\mathbb{E}\,\|H'\|_\infty}$$

then with large probability, $\mathcal{K} = \{w : \|w\|_1 \le B\}$ has non-empty intersection with $\{w : Xw = Y\}$, which contains the minimal ℓ_1 norm interpolator $\hat{w}$. It can be easily seen that

$$W(\mathcal{K}) = B\,\mathbb{E}\,\|H\|_\infty \quad \text{and} \quad R(\mathcal{K}) = B$$

and so by Theorem 1, with large probability

$$\begin{aligned}
L(\hat{w}) &\le \sup_{\|w\|_2 \le B, \hat{L}(w)=0} L(w) \\
&\le \frac{1+\beta}{n} \left(B\,\mathbb{E}\,\|H\|_\infty + B\sqrt{2\log\left(\frac{64}{\delta}\right)} + \|w^*\|_2 \sqrt{2\log\left(\frac{64}{\delta}\right)} \right)^2 \\
&= \frac{1+\beta}{n} B^2 (\mathbb{E}\,\|H\|_\infty)^2 (1+\gamma)^2 \\
&= (1+\beta)(1+\epsilon)(1+\gamma)^2 \left(\frac{\mathbb{E}\,\|H\|_\infty}{\mathbb{E}\,\|H'\|_\infty} \right)^2 (\sigma^2 + \|w^*\|_2^2)
\end{aligned}$$

where $\beta = 66\sqrt{\log(1/\delta)/n}$ and $\gamma = \frac{\sqrt{2\log\left(\frac{64}{\delta}\right)}}{\mathbb{E}\,\|H\|_\infty} + \frac{\|w^*\|_2 \sqrt{2\log\left(\frac{64}{\delta}\right)}}{B\,\mathbb{E}\,\|H\|_\infty}$. Observe that

$$B \ge \|w^*\|_2 \frac{\sqrt{n}}{\mathbb{E}\,\|H'\|_\infty} \quad \text{and} \quad \mathbb{E}\,\|H\|_\infty \ge \mathbb{E}\,\|H'\|_\infty.$$

Combined with the lower bound of Kamath (2013), we show

$$\gamma \le \sqrt{\frac{2\pi \log(128/\delta)}{\log d}} + \sqrt{\frac{2\log(64/\delta)}{n}} \le 8 \left(\sqrt{\frac{\log(1/\delta)}{\log d}} + \sqrt{\frac{\log(1/\delta)}{n}} \right).$$

In addition, we have

$$\frac{\mathbb{E}\,\|H\|_\infty}{\mathbb{E}\,\|H'\|_\infty} = 1 + \frac{\mathbb{E}\,\|H\|_\infty - \mathbb{E}\,\|H'\|_\infty}{\mathbb{E}\,\|H'\|_\infty} \le 1 + \sqrt{\frac{2\pi \log(2) \log|S|}{\log(d-|S|)}}.$$

Finally, it is a routine calculation to show

$$\begin{aligned}
&(1+\beta)(1+\epsilon)(1+\gamma)^2 \left(\frac{\mathbb{E}\,\|H\|_\infty}{\mathbb{E}\,\|H'\|_\infty}\right)^2 \\
&\leq 1 + 368\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\log(1/\delta)+\log|S|}{\log(d-|S|)}} + \frac{n}{\log(d-|S|)}\right) = 1+\eta
\end{aligned}$$

using the inequality $(1+x)(1+y) \leq 1+x+2y$ for $x \leq 1$. □