Tight bounds for minimum ℓ_1-norm interpolation of noisy data

Guillaume Wang* **Konstantin Donhauser*** **Fanny Yang**
ETH Zurich

Abstract

We provide matching upper and lower bounds of order $\sigma^2/\log(d/n)$ for the prediction error of the minimum ℓ_1-norm interpolator, a.k.a. basis pursuit. Our result is tight up to negligible terms when $d \gg n$, and is the first to imply asymptotic consistency of noisy minimum-norm interpolation for isotropic features and sparse ground truths. Our work complements the literature on "benign overfitting" for minimum ℓ_2-norm interpolation, where asymptotic consistency can be achieved only when the features are effectively low-dimensional.

1 INTRODUCTION

Recent experimental studies (Belkin et al., 2019; Zhang et al., 2021) reveal that in the modern high-dimensional regime, models that perfectly fit noisy training data can still generalize well. The phenomenon stands in contrast to the classical wisdom that interpolating the data results in poor statistical performance due to overfitting. Many theoretical papers have explored why, when, and to what extent interpolation can be harmless for generalization, suggesting a coherent storyline: High dimensionality itself can have a regularizing effect, in the sense that it lowers the model's sensitivity to noise. This intuition emerges from the fast-growing literature studying min-ℓ_2-norm interpolation in the regression setting with input dimension d substantially exceeding sample size n (see Bartlett et al. (2020); Dobriban and Wager (2018) and references therein). Results and intuition for this setting also extend to kernel methods (Ghorbani et al., 2021; Mei and Montanari, 2019).

However, a closer look at this literature reveals that while high dimensionality decreases the sensitivity to noise (error due to variance), the prediction error generally does not vanish as $d, n \to \infty$. Indeed, the bottleneck for asymptotic consistency is a non-vanishing bias term which can only be avoided when the features have low effective dimension $d_{\text{eff}} = \operatorname{Tr}\Sigma / |||\Sigma||| \ll n$, where Σ is the covariance matrix (Tsigler and Bartlett, 2020). Therefore, current theory does not yet provide a convincing explanation for why interpolating models generalize well for inherently high-dimensional input data. This work takes a step towards addressing this gap.

When the input data is effectively high-dimensional (e.g. isotropic and $d \gg n$), we generally cannot expect any data-driven estimator to generalize well unless there is underlying structure that can be exploited. In this paper, we hence focus on linear regression on isotropic Gaussian features with the simplest structural assumption: sparsity of the ground truth in the standard basis. For this setting, the ℓ_1-penalized regressor (LASSO, Tibshirani (1996)) achieves minimax optimal rates in the presence of noise (Van de Geer, 2008), while basis pursuit (BP, Chen et al. (1998)) – that is min-ℓ_1-norm interpolation – generalizes well in the noiseless case but is known to be very sensitive to noise (Candes, 2008; Donoho and Elad, 2006).

Given recent results on high dimensionality decreasing sensitivity of interpolators to noise, and classical results on the low bias of BP for learning sparse signals, the following question naturally arises:

Can we consistently learn sparse ground truth functions with minimum-norm interpolators on inherently high-dimensional features?

So far, upper bounds on the prediction error of the BP estimator of the order of the noise level $O(\sigma^2)$ have been derived for isotropic Gaussian (Koehler et al., 2021; Ju et al., 2020; Wojtaszczyk, 2010), sub-exponential (Foucart, 2014), or heavy-tailed (Chinot et al., 2021; Krahmer et al., 2018) features. In the case of isotropic Gaussian features, even though Chinot et al. (2021) show a tight matching lower bound for

---

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

adversarial noise, for i.i.d. noise the best known results are not tight: there is a gap between the non-vanishing upper bound $O(\sigma^2)$ (Wojtaszczyk, 2010) and the lower bound $\Omega\left(\frac{\sigma^2}{\log(d/n)}\right)$ (Chatterji and Long, 2021; Muthukumar et al., 2020). For i.i.d. noise, Chinot et al. (2021) conjecture that BP does not achieve consistency (see also Koehler et al. (2021)).

Contribution. We are the first to answer the above question in the affirmative. Specifically, we show that for isotropic Gaussian features, BP does in fact achieve asymptotic consistency when d grows superlinearly and subexponentially in n, disproving the recent conjecture by Chinot et al. (2021). Our result closes the aforementioned gap in the literature on BP: We give matching upper and lower bounds of order $\frac{\sigma^2}{\log(d/n)}$ on the prediction error of the BP estimator, exact up to terms that are negligible when $d \gg n$. Further, our proof technique is novel and may be of independent interest.

Structure of the paper. The rest of the article is structured as follows. In Section 2, we give our main result and discuss its implications. In Section 3, we present a proof sketch and provide insights on why our approach leads to tighter bounds than previous works. We discuss the scope of our assumptions and motivate future work in Section 4, and conclude the paper in Section 5.

2 MAIN RESULT

In this section we state our main result, followed by a discussion of its implications. We consider a linear regression model with input vectors $x \in \mathbb{R}^d$ drawn from an isotropic Gaussian distribution $x \sim \mathcal{N}(0, I_d)$, and response variable $y = \langle w^*, x\rangle + \xi$, where w^* is the ground truth to be estimated and $\xi \sim \mathcal{N}(0, \sigma^2)$ is a noise term independent of x. Given n i.i.d. random samples $(x_i, y_i)_{i=1}^n$, the goal is to estimate w^* and obtain a small prediction error (or risk) for the estimate $\hat{w}$

$$\mathbb{E}_{x,y}(\langle \hat{w}, x\rangle - y)^2 - \sigma^2 = \|\hat{w} - w^*\|_2^2$$

where we subtract the irreducible error σ^2. Note that this is also exactly the ℓ_2-error of the estimator. We study the min-ℓ_1-norm interpolator (or BP solution) defined by

$$\hat{w} = \arg\min_w \|w\|_1 \quad \text{such that} \quad \forall i,\ \langle x_i, w\rangle = y_i.$$

Our main result, Theorem 1, provides non-asymptotic matching upper and lower bounds for the prediction error of this estimator:

Theorem 1. *Suppose $\|w^*\|_0 \leq \kappa_1 \frac{n}{\log(d/n)^5}$ for some universal constant $\kappa_1 > 0$. There exist universal constants $\kappa_2, \kappa_3, \kappa_4, c_1, c_2, c_3 > 0$ such that, for any n, d with $n \geq \kappa_2$ and $\kappa_3 n \log(n)^2 \leq d \leq \exp(\kappa_4 n^{1/5})$, the prediction error satisfies*

$$\left| \|\hat{w} - w^*\|_2^2 - \frac{\sigma^2}{\log(d/n)} \right| \leq c_1 \frac{\sigma^2}{\log(d/n)^{3/2}} \tag{1}$$

with probability at least $1 - c_2 \exp\left(-\frac{n}{\log(d/n)^5}\right)$ $-d\exp(-c_3 n)$ over the draws of the dataset.

A proof sketch is presented in Section 3 and the full proof is given in Appendix A. We refer to Section 4 for a discussion on limitations of the assumptions.

This theorem proves an exact statistical rate with respect to the leading factor of order $\frac{\sigma^2}{\log(d/n)}$ for the prediction error of the BP solution. Previous lower bounds of order $\Omega\left(\frac{\sigma^2}{\log(d/n)}\right)$ for the same distributional setting (isotropic Gaussian features, i.i.d. noise) only apply under more restrictive assumptions, such as the zero-signal case $w^* = 0$ (Muthukumar et al., 2020), or assuming $d > n^4$ (Ju et al., 2020). On another note the best known upper bounds are of constant order $O(\sigma^2)$ (Chinot et al., 2021; Wojtaszczyk, 2010). Our result both proves the lower bound in more generality and significantly improves the upper bound that matches the lower bound, showing that the lower bound is in fact tight. An important implication of the upper bound is that BP achieves high-dimensional asymptotic consistency when $d = \omega(n)$, thus disproving to a recent conjecture by Chinot et al. (2021).

Dependency on w^*. We note that the bound for the risk in Theorem 1 is independent of the choice of w^* assuming that it is sparse (i.e, has bounded ℓ_0-norm). Essentially, this arises from the well known fact that in the noiseless case ($\sigma = 0$) we can achieve exact recovery Candes (2008) of sparse ground truths. More generally, existing upper bounds for the prediction error of the BP estimator for general ground truths w^* are of the form[1]

$$\|\hat{w} - w^*\|_2^2 \lesssim \frac{\|\xi\|_2^2}{n} + \|w^*\|_1^2 \frac{\log(d/n)}{n} \tag{2}$$

(see e.g. (Chinot et al., 2021, Theorem 3.1)). That is, they contain a first term reflecting the error due to overfitting of the noise ξ which is independent of w^* and a second term which can be understood as the noiseless error only depending on w^* but not on the noise ξ. In fact, the authors of both papers

[1] The notation $a \lesssim b$ means that there exists a universal constant $c_1 > 0$ such that $a \leq c_1 b$, and we write $a \asymp b$ for $a \lesssim b$ and $b \lesssim a$.

show that assuming the ground truth is hard-sparse (bounded ℓ_0-norm), the second term on the RHS in Equation (2) can be avoided, resulting in the bound $\|\hat{w} - w^*\|_2^2 \lesssim \frac{\|\xi\|_2^2}{n}$. Therefore, it is also not surprising that our tighter bound in Equation (1) does not explicitly depend on w^*.

2.1 Numerical simulations

We now present numerical simulations illustrating Theorem 1. Figure 1a shows the prediction error of BP plotted as a function of $\log(d/n)$ with varying d and $n = 400$ fixed, for isotropic inputs generated from the zero-mean and unit-variance *Normal*, *Log Normal* and *Rademacher* distributions. For all three distributions, the prediction error closely follows the trend line $\frac{\sigma^2}{\log(d/n)}$ (dashed curve). While Theorem 1 only applies for Gaussian features, the figure suggests that this statistical rate of BP holds more generally (see discussion in Section 4).

Figure 1b shows the prediction error of the min-ℓ_1-norm (BP) and min-ℓ_2-norm interpolators as a function of the noise σ^2, for fixed d and n. The prediction error of the former again aligns with the theoretical rate $\frac{\sigma^2}{\log(d/n)}$. Furthermore, we observe that the min-ℓ_1-norm interpolator is sensitive to the noise level σ^2, while the min-ℓ_2-norm interpolator has a similar (non-vanishing) prediction error across all values of σ^2.

For both plots we use $n = 400$ and average the prediction error over 20 runs; in Figure 1b we additionally show the standard deviation (shaded regions). The ground truth is $w^* = (1, 0, \cdots, 0)$. Finally, we choose $\sigma^2 = 1$ in Figure 1a, and $d = 20000$ in Figure 1b.

2.2 Implications and insights

We now discuss further high-level implications and insights that follow from Theorem 1.

High-dimensional asymptotic consistency. Our result proves consistency of BP for any asymptotic regime $d \asymp n^\beta$ with $\beta > 1$. In fact, we argue that those are the only regimes of interest. For d growing exponentially with n, known minimax lower bounds for sparse problems of order $\frac{\sigma^2 s \log(d/s)}{n}$ (with $s \leq n$ the ℓ_0-norm of the BP estimator), preclude consistency (Verzelen, 2012). On the other hand, for linear growth $d \asymp n$, i.e., $\beta = 1$ – studied in detail by Li and Wei (2021) –, the uniform prediction error lower bound $\frac{\sigma^2 n}{d-n}$ holding for all interpolators (Zhou et al., 2020; Muthukumar et al., 2020) also forbids vanishing prediction error. Note that for $d \asymp n^\beta$ $(\beta > 1)$, asymptotic consistency can also be achieved by a carefully designed "hybrid" interpolating estimator (Muthukumar et al., 2020, Section 5.2); contrary to BP, this estimator is not a minimum-norm interpolator, and is not structured (not n-sparse).

Trade-off between structural bias and sensitivity to noise. As mentioned in the introduction, our upper bound on the prediction error shows that, contrary to min-ℓ_2-norm interpolation, BP is able to learn sparse signals in high dimensions thanks to its structural bias towards sparsity. However, our lower bound can be seen as a tempering negative result: The prediction error decays only at a slow rate of $\frac{\sigma^2}{\log(d/n)}$.

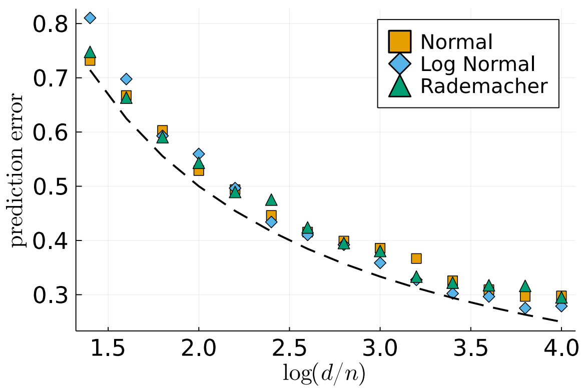
Compared to min-ℓ_2-norm interpolation, BP (min-ℓ_1-norm interpolation) suffers from a higher sensitivity to noise, but possesses a more advantageous structural bias. To compare the two methods' sensitivity to noise, consider the case $w^* = 0$, where the prediction error purely reflects the effect of noise. In this case, although both methods achieve vanishing error, the statistical rate for BP, $\frac{\sigma^2}{\log(d/n)}$, is much slower than that of min-ℓ_2-norm interpolation, $\sigma^2 \max(\frac{1}{\sqrt{n}}, \frac{n}{d})$ (Koehler et al., 2021, Theorem 3). Contrariwise, to compare the effect of structural bias, consider the noiseless case with a non-zero ground truth. It is well known that BP successfully learns sparse signals (Candes, 2008), while min-ℓ_2-norm interpolation always fails to learn the ground truth due to the lack of any corresponding structural bias.

Thus, there appears to be a trade-off between structural bias and sensitivity to noise: BP benefits from a strong structural bias, allowing it to have good performance for noiseless recovery of sparse signals, but in return displays a poor rate in the presence of noise – while min-ℓ_2-norm interpolation has no structural bias (except towards zero), causing it to fail to recover any non-zero signal even in the absence of noise, but in return does not suffer from overfitting of the noise. This behavior is also illustrated in Figure 1b.

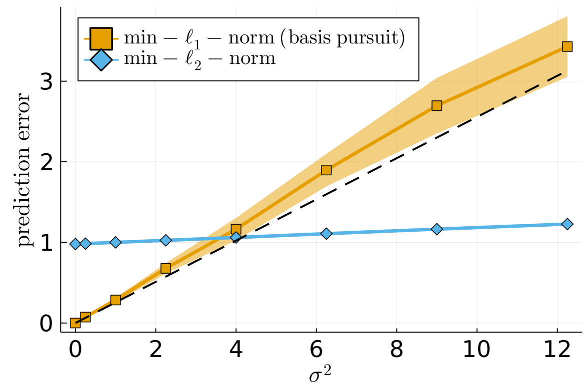
3 PROOF SKETCH

In this section we show the main ingredients that are key to prove our risk upper bound of $\frac{\sigma^2}{\log(d/n)} + O\left(\frac{\sigma^2}{\log(d/n)^{3/2}}\right)$. The proof sketch is interleaved with remarks providing insights on how our technique allows to improve upon previous works. For the sake of clarity, we omit the discussion of the matching lower bound, as its proof follows exactly the same ideas. The full proof is given in Appendix A.

The proof follows a standard localization/uniform convergence argument, where we first upper-bound the ℓ_1-error $\|\hat{w} - w^\star\|_1$ (localization) and then uniformly



(a) Prediction error of BP vs. $\log(d/n)$



(b) Prediction error vs. σ^2

Figure 1: Prediction error as a function of (a) $\log(d/n)$ with varying d and $n = 400$ fixed, and (b) $\sigma^2 = 1$ with $d = 20000, n = 400$. The features are generated by drawing from the isotropic zero-mean and unit-variance (b) *Normal* and (a) *Normal*, *Log Normal* and *Rademacher* distributions. For BP on Gaussian-distributed features (orange squares), the plots correctly reflect the theoretical rate $\frac{\sigma^2}{\log(d/n)}$ (dashed curve). See Section 2.1 for further details.

upper-bound the risk (i.e. ℓ_2-error) over all interpolators with bounded ℓ_1-error (uniform convergence).

3.1 Localization

We derive a high-probability upper bound on the ℓ_1-error: $\|\hat{w} - w^\star\|_1 \leq B(n,d)$, implying that the estimator of interest $\hat{w}$ is an interpolator located in the ℓ_1-ball of radius $B(n,d)$ centered at $w^\star$.

To upper-bound $\|\hat{w} - w^\star\|_1$ we first observe that, by definition of the BP estimator $\hat{w}$ and via a simple triangle inequality (Chinot et al., 2021), it holds that

$$\|\hat{w} - w^\star\|_1 \leq 2\sqrt{\|w^*\|_0}\,\|\hat{w} - w^*\|_2 + \min_{Xw=\xi} \|w\|_1$$

which is proven in Lemma 1. To control the first term, we make use of the loose high-probability upper bound $\|\hat{w} - w^*\|_2 \lesssim \sigma$ from previous works (Chinot et al., 2021; Wojtaszczyk, 2010). Thus we have $2\sqrt{\|w^*\|_0}\,\|\hat{w} - w^*\|_2 \lesssim \sigma\sqrt{\|w^*\|_0}$. The second term, $\Phi_N := \min_{Xw=\xi} \|w\|_1$, reflects how enforcing the interpolation of noise affects the ℓ_1-norm of the estimator. To control it with high probability directly is challenging, due to the randomness of both data X and noise ξ. Instead, we bound it using the Convex Gaussian Minimax Theorem (CGMT) (Thrampoulidis et al., 2015); we postpone the sketch of this derivation to Section 3.4 where we show how we derive a high probability bound $\Phi_N \leq M(n,d)$ (the precise expression can be found in Proposition 2). Having controlled the two terms separately, we get the high-probability bound

$$\|\hat{w} - w^*\|_1 \leq c\sigma\sqrt{\|w^*\|_0} + M(n,d) =: B(n,d)$$

for some universal constant $c > 0$. Note that by assumption on $\|w^*\|_0$, the first term is of order at most $\sqrt{\frac{\sigma^2 n}{\log(d/n)^5}}$, so negligible compared to $M(n,d)$ which is of order $\sqrt{\frac{\sigma^2 n}{\log(d/n)}}$.

How tightness of $M(n,d)$ affects the upper bound. Intriguingly, our analysis requires a very precise expression for the deterministic upper bound on $\Phi_N = \min_{Xw=\xi} \|w\|_1$. Let $M(n,d)$ be the one used in our analysis, given in Proposition 2. If we use a different expression $\tilde{M}(n,d)$ as the upper bound of Φ_N instead of $M(n,d)$ in the rest of our analysis,

1. for $\tilde{M}(n,d) = M(n,d)\left(1 + \Theta\left(\frac{1}{\log(d/n)^2}\right)\right)$: We would still get exactly the precise upper bound of Theorem 1.
2. for $\tilde{M}(n,d) = M(n,d)\left(1 + \Theta\left(\frac{1}{\log(d/n)}\right)\right)$: We would obtain an upper bound of the correct order up to a universal constant factor $\|\hat{w} - w^*\|_2^2 \lesssim \frac{\sigma^2}{\log(d/n)}$.
3. for $\tilde{M}(n,d) = M(n,d)\,(1 + \Theta(1))$: In this case our analysis would only yield an upper bound of constant order $\|\hat{w} - w^*\|_2^2 \lesssim \sigma^2$, same as obtained by Chinot et al. (2021); Wojtaszczyk (2010).

Yet, it is not clear whether the tightness of $M(n,d)$ is only needed due to our analysis, or whether any uniform convergence bound with the corresponding $B(n,d)$ would fail to yield tight bounds. We leave this question as an interesting direction for future work, described further in Section 4.

Comparison to Chinot et al. (2021); Wojtaszczyk (2010). As we discussed above, the tight expression for $M(n,d)$ is crucial for our analysis to yield the tight bounds in Theorem 1. In comparison, in the paper Chinot et al. (2021) the authors derive an upper bound for $M(n,d)$ of order $\sim \sqrt{2\frac{\sigma^2 n}{\log(d/n)}}$ and hence existing bounds are not sufficiently tight. However, we further note that the bound on $M(n,d)$ is not the only contribution of this paper as inserting our bound in the analysis conveyed in Chinot et al. (2021); Wojtaszczyk (2010); Koehler et al. (2021) would still result in a constant upper bound of order σ^2 (see the next subsection for a detailed discussion on Koehler et al. (2021)). Finally, we note that Chinot et al. (2021); Wojtaszczyk (2010) both study the case where the noise can also be adversarial, for which the rate of order σ^2 for the risk is optimal.

3.2 Uniform risk bound and reduction to auxiliary problem by GMT

Given that $\hat{w}$ belongs with high probability to the set of interpolators located in the ℓ_1-ball of radius $B(n,d)$ centered at $w^\star$, we proceed to upper-bound the risk of all such interpolators.

Concretely, we find a high-probability upper bound on

$$\begin{aligned}\Phi_+ &:= \max_w \|w - w^\star\|_2^2 \quad \text{s.t} \quad \|w - w^*\|_1 \leq B(n,d) \\ &\qquad\qquad \text{and } X(w - w^*) = \xi \\ &= \max_w \|w\|_2^2 \quad \text{s.t} \quad \|w\|_1 \leq B(n,d) \text{ and } Xw = \xi.\end{aligned}$$

While directly bounding Φ_+ is challenging due to the randomness of both X and ξ, we can instead make use of the Gaussian Minimax Theorem (GMT) which allows us to equivalently upper-bound the value of the so-called auxiliary problem

$$\begin{aligned}\phi_+ := \max_w \|w\|_2^2 \quad \text{s.t} \quad \|w\|_1 \leq B(n,d) \\ \text{and } \langle w, h\rangle^2 \geq (1-\rho)n(\sigma^2 + \|w\|_2^2)\end{aligned} \tag{3}$$

with h an i.i.d. Gaussian random vector and ρ a vanishing parameter. Indeed, the GMT ensures that $\mathbb{P}_{X,\xi}(\Phi_+ > t) \leq 2\mathbb{P}_h(\phi_+ \geq t) + \epsilon_\rho$ (see Proposition 1 for the expression of ϵ_ρ); in words, a high-probability upper bound on ϕ_+ gives a high-probability upper bound on Φ_+.

The crux of the analysis is thus to obtain a good upper bound for ϕ_+. Since ϕ_+ is defined as the optimal value of a maximization problem, a possible approach is to consider relaxations of it; as detailed in the next remark, this leads to loose bounds.

Comparison to Koehler et al. (2021). The optimization problem defining ϕ_+ in Equation (3) is a maximization problem of a convex function over nonconvex constraints. We now show how to obtain a first loose upper bound following Koehler et al. (2021) and briefly discuss why this methodology fails to give tight bounds (see also the paragraph "Application: Isotropic features" in that paper).

Using Hölder's inequality $\langle w, h\rangle \leq \|w\|_1 \|h\|_\infty$, we obtain a proper relaxation of the problem (3) if we replace the constraints by

$$B(n,d)^2 \|h\|_\infty^2 \geq (1-\rho)n(\sigma^2 + \|w\|_2^2).$$

This immediately implies the upper bound

$$\phi_+ \leq \frac{B(n,d)^2 \|h\|_\infty^2}{n(1-\rho)} - \sigma^2.$$

However this bound is loose, even when we plug in our tight localization bound for $B(n,d)^2 \approx \frac{\sigma^2 n}{2\log(d/n) - \log\log(d/n)}$. Indeed, with this estimate and by Gaussian concentration results, the above bound reads for $d \gg n$

$$\begin{aligned}\phi_+ &\leq \frac{B(n,d)^2 \|h\|_\infty^2}{n(1-\rho)} - \sigma^2 \\ &\approx \frac{\sigma^2 \; 2\log(d)}{2\log(d/n)} - \sigma^2 = \sigma^2 \frac{\log(n)}{\log(d/n)}.\end{aligned}$$

Note that this bound is constant in any polynomial growth regime $d \asymp n^\beta$, while we prove an upper bound in Theorem 1 which vanishes in these regimes as $d, n \to \infty$.

This looseness points to the fact that using Hölders inequality is too imprecise. In the next subsection, we describe our refined analysis which better takes into account the relationship between $\langle w, h\rangle$, $\|w\|_2$ and $\|w\|_1$.

3.3 Path approach: reparametrizing the auxiliary problem as a one-dimensional problem

The key observation that allows us to derive a tight bound for ϕ_+, is that we can cast the d-dimensional problem (3) into a one-dimensional problem, which we can study explicitly. Namely, we identify a path $\gamma : \mathbb{R} \to \mathbb{R}^d$ for which we show that the optimum in (3) is necessarily attained at $B(n,d)\gamma(\alpha)/\|\gamma(\alpha)\|_1$ for some $\alpha \in \mathbb{R}$. Note that our reduction is exact, not a relaxation.

More precisely, we define the path $\gamma : [1, \alpha_{\max}] \to \mathbb{R}^d$

by

$$\gamma(\alpha) = \arg\min_w \|w\|_2^2 \quad \text{s.t} \quad \begin{cases} \langle w, h \rangle = \|h\|_\infty \\ \forall i, h_i w_i \geq 0 \\ \|w\|_1 = \alpha \end{cases}$$

(in particular $\|\gamma(\alpha)\|_1 = \alpha$) and we show that

$$\phi_+ = B(n,d)^2 \max_{1 \leq \alpha \leq \alpha_{\max}} \left(\frac{\|\gamma(\alpha)\|_2}{\alpha} \right)^2 \qquad (4)$$
$$\text{s.t.} \quad \frac{\|h\|_\infty^2}{(1-\rho)n} \geq \frac{\alpha^2 \sigma^2}{B(n,d)^2} + \|\gamma(\alpha)\|_2^2$$

(see Appendix A.2). Because $\gamma(\alpha)$ is the argmin of a convex optimization problem, it is relatively easy to study, and we can even derive an exact expression for it (Lemma 3). We now discuss the two key steps to study the optimization problem in Equation (4).

a) Monotonicity of the objective. We observe that $\frac{\|\gamma(\alpha)\|_2}{\alpha}$ is monotonically decreasing and that $\|\gamma(\alpha)\|_2^2$ is a convex function (Lemma 5). This has two important consequences. Firstly, the set of α's which satisfy the constraints in (4) is an interval, denoted $[\underline{\alpha}_I, \overline{\alpha}_I]$. Secondly, denoting α^* ($\in [\underline{\alpha}_I, \overline{\alpha}_I]$) an argmax of (4), we have for any $\alpha < \underline{\alpha}_I$

$$B(n,d)^2 \left(\frac{\|\gamma(\alpha)\|_2}{\alpha} \right)^2 \geq B(n,d)^2 \left(\frac{\|\gamma(\alpha^*)\|_2}{\alpha^*} \right)^2 = \phi_+.$$

So to obtain an upper bound on ϕ_+, all we need is to find an α on the left of the feasible interval $[\underline{\alpha}_I, \overline{\alpha}_I]$. We do this by finding both an α_n such that $\underline{\alpha}_I \leq \alpha_n \leq \overline{\alpha}_I$ (i.e. α_n is feasible), and an α_s such that $\alpha_s < \alpha_n$ and $\alpha_s \notin [\underline{\alpha}_I, \overline{\alpha}_I]$ (i.e. α_s is not feasible).

b) Discretization of the path. We observe that there exist "breakpoints" $1 = \alpha_2 < ... < \alpha_{d+1} = \alpha_{\max}$ for which $\gamma(\alpha_s)$ has a special structure (in particular it is $(s-1)$-sparse). Further, applying Gaussian concentration results to h leads to high-probability estimates for α_s and $\|\gamma(\alpha_s)\|_2$ (Proposition 4). Thanks to those estimates, we show that α_n is feasible for (4) and we find a choice of $s < n$ such that α_s is not feasible, with high probability. Thus, with high probability ϕ_+ is upper-bounded by $B(n,d)^2 \left(\frac{\|\gamma(\alpha_s)\|}{\alpha_s} \right)^2$ – for which we have high-probability estimates.

Intuition for the definition of $\gamma(\alpha)$. As discussed in the previous subsection, the relaxation of (A_+) based on Hölder's inequality $\langle w, h \rangle \leq \|w\|_1 \|h\|_\infty$, used by Koehler et al. (2021), is too loose. Informally, it effectively amounts to forgetting the direction of h and only optimizing over the ℓ_1 and ℓ_2-norms of vectors. The main idea of our refined analysis is to introduce a path, $\{\gamma(\alpha)/\alpha\}_\alpha$, allowing us to better take into account the relationship between $\langle w, h \rangle$, $\|w\|_2$ and $\|w\|_1$. To intuitively understand how the path achieves this goal, it may be easier to use its following form:

$$\overline{\gamma}(\beta) = \arg\max_w \langle w, h \rangle \quad \text{s.t} \quad \begin{cases} \|w\|_2^2 \leq \beta \\ \forall i, h_i w_i \geq 0 \\ \|w\|_1 = 1 \end{cases},$$

which is equivalent to $\gamma(\alpha)$ up to linear reparametrization and rescaling (see Equation (10) in Appendix B.3). Note that $\langle \overline{\gamma}(1), h \rangle = \|h\|_\infty \|\overline{\gamma}(1)\|_1$, which exactly recovers the equality case of Hölder's inequality (i.e. $\overline{\gamma}(1)$ is a subgradient of the ℓ_1-norm at h). More generally, $\langle w, h \rangle \leq \langle \overline{\gamma}(\beta), h \rangle \|w\|_1$ for any w such that $\|w\|_2^2 \leq \beta \|w\|_1^2$, which can be understood as a refined Hölder's inequality for limited ℓ_2-norms.

3.4 Obtaining a good estimate for Φ_N

Finally, we unveil how we derive the high-probability upper bound $\Phi_N \leq M(n,d)$ in the localization step (1.): The derivation actually uses the same tools as for the upper bound of Φ_+. Using the Convex Gaussian Minimax Theorem (CGMT) (Thrampoulidis et al., 2015), which is a variant of the GMT for convex-concave functions, we can again introduce an auxiliary problem

$$\phi_N := \min_w \|w\|_1 \quad \text{s.t} \quad \langle w, h \rangle^2 \geq (1+\rho)n(\sigma^2 + \|w\|_2^2)$$

with h an i.i.d. Gaussian vector and ρ a vanishing parameter, with the property that high-probability upper bounds on ϕ_N give high-probability upper bounds on Φ_N. Further, we can again reduce this d-dimensional optimization problem to one over the same path $\{\gamma(\alpha)\}_\alpha$:

$$\phi_N = \min_{1 \leq \alpha \leq \alpha_{\max}} f(\alpha) \quad \text{s.t} \quad \|h\|_\infty^2 \geq (1+\rho)n \|\gamma(\alpha)\|_2^2 \qquad (5)$$

(see Appendix A.1.1 for the expression of $f(\alpha)$).

Since we want to upper-bound this minimum, it is sufficient to find some α which satisfies the constraints in Equation (5). In particular, we again focus on the breakpoints $\{\alpha_s\}_{s \in \{2,...,d+1\}}$, and show that with high probability $\alpha = \alpha_n$ is a valid choice which is approximately tight (see Remark 1).

Sparsity of $\gamma(\alpha_n)$. In summary, the proof is essentially based on the localization around a rescaled version of $\gamma(\alpha_n)$, which is a $(n-1)$-sparse vector. This choice can also be motivated by a different argument: It is well known that the minimizer of the optimization problem $\min_{Xw=\xi} \|w\|_1$ defining Φ_N is n-sparse. Hence, due to the strong connection between the optimization problems defining ϕ_N and Φ_N, we also expect the minimizer of Equation (5) to be approximately n-sparse.

4 FUTURE WORK

Our main result gives tight bounds for BP on isotropic Gaussian features. It would be interesting to extend the study to other connected settings, which we now motivate and for which we summarize key challenges. Furthermore, we pose a research question which aims to give a better intuition for the proof.

Necessity of tightness at the localization step. As discussed in Section 3.1, in order to obtain the right rate in Theorem 1, the localization step of our analysis needed to be very tight. The expression we derive for a high-probability upper bound $M(n, d)$ (from Proposition 2) on $\min_{Xw=\xi} \|w\|_1$ needs to be precise up to relative error of no more than $\Theta\left(\frac{1}{\log(d/n)^2}\right)$. This strikes us as an unusual feature of our derivation. Yet, it is unclear whether this is an artifact of our analysis via the application of the GMT, or whether this is due to the nature of the statistical problem itself. More specifically, we motivate future research to answer the question whether it is true that

1. for any $\tilde{M}(n, d) = cM(n, d)$ with $c > 1$, we have that with high probability,
$$\max_{\substack{\|w\|_1 \leq \tilde{M}(n,d) \\ Xw=\xi}} \|w\|_2^2 \asymp \sigma^2.$$
2. for any $\tilde{M}(n, d) = M(n, d)(1 + \omega(\frac{1}{\log(d/n)}))$, we have that with high probability,
$$\max_{\substack{\|w\|_1 \leq \tilde{M}(n,d) \\ Xw=\xi}} \|w\|_2^2 = \omega\left(\frac{\sigma^2}{\log(d/n)}\right).$$

Resolving this question is challenging due to the non-concavity of the maximization objective. While we can still use the GMT to upper-bound this quantity (see Proposition 1), we cannot use the CGMT to lower-bound it, and thus the methodologies used in this paper fall short. As a possible direction, we note that this hypothesis is related to the question of finding tight lower bounds for the diameter of the intersection of the kernel of X and the ℓ_1-ball (see Theorem 3.5 in Vershynin (2011)).

Non-isotropic features. Theorem 1 assumes isotropic features as we are interested in showing consistency of BP for inherently high-dimensional input data. By contrast, recently there has been an increased interest in studying spiked covariance data models (see e.g. Bartlett et al. (2020); Muthukumar et al. (2021); Chatterji and Long (2021)). In such settings even min-ℓ_2-norm interpolators can achieve consistency. The main obstacle to extending our methodology to non i.i.d. features lies in adapting the definition of the path $\{\gamma(\alpha)\}_\alpha$. Assuming a diagonal covariance matrix, such an extension should be relatively straightforward. We leave this task and the challenging non-diagonal case for future work.

Non-Gaussian features. The proof of Theorem 1 crucially relies on the (Convex) Gaussian Minimax Theorem (Thrampoulidis et al., 2015; Gordon, 1988), and hence on the assumption that the input features are drawn from a Gaussian distribution. In Figure 1a, we include plots of the prediction error $\|\hat{w} - w^*\|_2^2$ not only for Gaussian but also for Log Normal and Rademacher distributed features. We observe that in all three cases, the prediction error closely follows the trend line $\frac{\sigma^2}{\log(d/n)}$ (dashed curve). This leads us to conjecture that Theorem 1 can be extended to a more general class of distributions.

Generalizing our results in this direction appears to be a challenging task since the tools used in this paper are not directly applicable anymore. Instead, for heavy-tailed distributions, a popular theoretical framework is the small-ball method (Mendelson, 2014; Koltchinskii and Mendelson, 2015), which covers the Log Normal and Rademacher distributions. Chinot et al. (2021) apply this approach to min-ℓ_1-norm interpolation, and obtain the constant upper bound $O(\sigma^2)$, under more general assumptions than our setting (in particular their analysis handles adversarial noise with magnitude controlled by σ^2). Yet, it is unclear whether the looseness of their upper bound is an artifact of their proof, or whether the small-ball method itself is too general to capture the rates observed in Figure 1a.

Finally, we also leave it as future work to adapt our proof technique for minimum-norm interpolators with general norms, and for classification tasks.

5 CONCLUSION

By introducing a novel proof technique, we derive matching upper and lower bounds of order $\frac{\sigma^2}{\log(d/n)}$ on the prediction error of basis pursuit (BP, or min-ℓ_1-norm interpolation) in noisy sparse linear regression. Our result closes a gap in the minimum-norm interpolation literature, disproves a conjecture from Chinot et al. (2021), and is the first to imply asymptotic consistency of a minimum-norm interpolator for isotropic features. Furthermore, the prediction error decays with the amount of overparametrization d/n, confirming that BP also benefits from the regularization effect of high dimensionality, as suggested by the modern storyline on interpolating models.

Acknowledgements

K.D. is supported by the ETH AI Center and the ETH Foundations of Data Science. We would like to thank Geoffrey Chinot and Matthias Löffler for insightful discussions. Finally, we are very grateful to Sara van de Geer for her valuable advice.

References

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Boucheron, S. and Thomas, M. (2012). Concentration inequalities for order statistics. *Electronic Communications in Probability*, 17:1–12.

Candes, E. J. (2008). The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592.

Chatterji, N. S. and Long, P. M. (2021). Foolish crowds support benign overfitting. *arXiv preprint arXiv:2110.02914*.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61.

Chinot, G., Löffler, M., and van de Geer, S. (2021). On the robustness of minimum-norm interpolators. *arXiv preprint arXiv:2012.00807*.

Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279.

Donoho, D. L. and Elad, M. (2006). On the stability of the basis pursuit in the presence of noise. *Signal Processing*, 86(3):511–532.

Foucart, S. (2014). Stability and robustness of l1-minimizations with Weibull matrices and redundant dictionaries. *Linear Algebra and its Applications*, 441:4–21.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. (2021). Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054.

Gordon, Y. (1988). On Milman's inequality and random subspaces which escape through a mesh in rn. In *Geometric aspects of functional analysis*, pages 84–106. Springer.

Ju, P., Lin, X., and Liu, J. (2020). Overfitting Can Be Harmless for Basis Pursuit, But Only to a Degree. *Advances in Neural Information Processing Systems*, 33.

Koehler, F., Zhou, L., Sutherland, D. J., and Srebro, N. (2021). Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds, and Benign Overfitting. *arXiv preprint arXiv:2106.09276*.

Koltchinskii, V. and Mendelson, S. (2015). Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008.

Krahmer, F., Kümmerle, C., and Rauhut, H. (2018). A quotient property for matrices with heavy-tailed entries and its application to noise-blind compressed sensing. *arXiv preprint arXiv:1806.04261*.

Li, G., Xu, X., and Gu, Y. (2020). Lower Bound for RIP Constants and Concentration of Sum of Top Order Statistics. *IEEE Transactions on Signal Processing*, 68:3169–3178.

Li, Y. and Wei, Y. (2021). Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*.

Mei, S. and Montanari, A. (2019). The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve. *Communications on Pure and Applied Mathematics*.

Mendelson, S. (2014). Learning without concentration. *Conference on Learning Theory*, pages 25–39.

Muthukumar, V., Narang, A., Subramanian, V., Belkin, M., Hsu, D., and Sahai, A. (2021). Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69.

Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. (2020). Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83.

Thrampoulidis, C., Oymak, S., and Hassibi, B. (2015). Regularized linear regression: A precise analysis of the estimation error. *Conference on Learning Theory*, pages 1683–1709.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

Tsigler, A. and Bartlett, P. L. (2020). Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645.

Vershynin, R. (2011). Lectures in geometric functional analysis. *Unpublished manuscript. Available at http://www-personal. umich. edu/romanv/papers/GFA-book/GFA-book. pdf*, 3(3):3–3.

Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90.

Wojtaszczyk, P. (2010). Stability and instance optimality for Gaussian measurements in compressed sensing. *Foundations of Computational Mathematics*, 10(1):1–13.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Zhou, L., Sutherland, D. J., and Srebro, N. (2020). On uniform convergence and low-norm interpolation learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 6867–6877.

Supplementary Material: Tight bounds for minimum ℓ_1-norm interpolation of noisy data

A PROOF OF MAIN RESULT

In this section, we present the proof of our main result, Theorem 1. In Appendix A.1, we describe the main steps of the proof rigorously, in the form of three propositions which we then prove in Appendix A.2. Full proofs for the intermediary Lemmas and Propositions are given in Appendix B.

Notation. On the finite-dimensional space $\mathbb{R}^d$, we write $\|\cdot\|_2$ for the Euclidean norm and $\langle\cdot,\cdot\rangle$ for the Euclidean inner product. The ℓ_1 and ℓ_∞-norms are denoted by $\|\cdot\|_1$ and $\|\cdot\|_\infty$, respectively. The vectors of the standard basis are denoted by $e_1, ..., e_d$, and $\mathbf{1} \in \mathbb{R}^d$ is the vector with all components equal to 1. For $s \leq d$ and $H \in \mathbb{R}^d$, $H_{[s]}$ is the vector such that $(H_{[s]})_i = H_i$ if $i \leq s$ and 0 otherwise. $\mathcal{N}(\mu, \Sigma)$ is the normal distribution with mean μ and covariance Σ, $\Phi(x)$ is the cumulative distribution function of the scalar standard normal distribution, $\Phi^{\complement}(x) = 1 - \Phi(x)$, and log denotes the natural logarithm. For all $s \leq d$, we denote by $t_s \in \mathbb{R}$ the quantile of the standard normal distribution defined by $2\Phi^{\complement}(t_s) = s/d$. The n samples $x_i \in \mathbb{R}^d$ form the rows of the data matrix $X = [x_1 \dots x_n]^\top$, with $X_{ij} \sim \mathcal{N}(0, 1)$ for each i, j. The scalars y_i, ξ_i are also aggregated into vectors $y, \xi \in \mathbb{R}^n$ with $\xi \sim \mathcal{N}(0, \sigma^2 I_n)$ and $y = Xw^* + \xi$. With this notation, $\hat{w}$ interpolates the data $X\hat{w} = y$ which is equivalent to $X(\hat{w} - w^*) = \xi$. To easily keep track of the dependency on dimension and sample size, we reserve the $O(\cdot)$ notation to contain only universal constants, without any hidden dependency on d, n, σ^2 or $\|w^*\|_0$. We will also use $c_1, c_2, \dots$ and $\kappa_1, \kappa_2, \dots$ to denote positive universal constants reintroduced each time in the proposition and lemma statements, except for c and c_0 which should be considered as fixed throughout the whole proof.

A.1 Proof of Theorem 1

We proceed by a localized uniform convergence approach, similar to Chinot et al. (2021); Koehler et al. (2021); Ju et al. (2020); Muthukumar et al. (2020), and common in the literature, e.g., on structural risk minimization. That is, the proof consists of two steps:

1. *Localization.* We prove that, with high probability, the min-ℓ_1-norm interpolator $\hat{w}$ satisfies $\|\hat{w} - w^*\|_1 \leq c\sigma\sqrt{\|w^*\|_0} + \min_{Xw=\xi} \|w\|_1$ for some universal constant $c > 0$. We then derive a (finer than previously known) high-probability upper bound on the second term,

$$\min_{Xw=\xi} \|w\|_1 \;=: \Phi_N \;\leq M(n, d). \tag{P_N}$$

   Consequently, with high probability $\hat{w}$ satisfies

$$\|\hat{w} - w^*\|_1 \leq c\sigma\sqrt{\|w^*\|_0} + M(n, d) =: B(n, d).$$

2. *Uniform convergence.* We derive high-probability uniform upper and lower bounds on the prediction error for all interpolators located no farther than $B(n, d)$ from w^* in ℓ_1 norm. In symbols, we find a high-probability upper bound for

$$\max_{\substack{\|w-w^*\|_1 \leq B(n,d) \\ X(w-w^*)=\xi}} \|w - w^*\|_2^2 = \max_{\substack{\|w\|_1 \leq B(n,d) \\ Xw=\xi}} \|w\|_2^2 \;=: \Phi_+ \tag{P_+}$$

   and a high-probability lower bound for

$$\min_{\substack{\|w-w^*\|_1 \leq B(n,d) \\ X(w-w^*)=\xi}} \|w - w^*\|_2^2 = \min_{\substack{\|w\|_1 \leq B(n,d) \\ Xw=\xi}} \|w\|_2^2 \;=: \Phi_-. \tag{P_-}$$

By definition of $B(n,d)$ in (P_N), with high probability the min-ℓ_1-norm interpolator $\hat{w}$ belongs to the set of feasible solutions in (P_+) and (P_-), and hence the second step yields high-probability upper and lower bounds on its prediction error $\|\hat{w} - w^*\|_2^2$.

The key is thus to derive tight high-probability bounds for the quantities Φ_N, Φ_+, Φ_-. Our derivation proceeds in two parts, described below. The first part uses the CGMT to convert the original optimization problem to an auxiliary problem, similar to Koehler et al. (2021). The second part, which contains the crucial elements for our proof of the vanishing upper bound and is the key technical contribution of this paper, consists in reducing the d-dimensional auxiliary problem to a scalar one using a path reparametrization.

Preliminary: Localization around w^*. The following fact shows how, as announced, Φ_N can be used to derive a localization bound for $\hat{w}$.

Lemma 1. *Suppose $\|w^*\|_0 \le \kappa_1 \frac{n}{\log(d/n)^5}$ for some universal constant $\kappa_1 > 0$. There exist universal constants $\kappa_2, \kappa_3, \kappa_4, c, c_3 > 0$ such that, if $n \ge \kappa_2$ and $\kappa_3 n \log(n)^2 \le d \le \exp(\kappa_4 n^{1/5})$, then the min-$\ell_1$-norm interpolator $\hat{w}$ satisfies*

$$\|\hat{w} - w^*\|_1 \le c\sigma\sqrt{\|w^*\|_0} + \min_{Xw=\xi} \|w\|_1$$

with probability at least $1 - d\exp(-c_3 n)$.

The proof of Lemma 1 is given in Appendix B.1. Interestingly, it makes use of the loose upper bound $\|\hat{w} - w^*\|_2^2 \lesssim \sigma^2$, shown previously by Wojtaszczyk (2010) and Chinot et al. (2021), as an intermediate result.

A.1.1 (Convex) Gaussian Minimax Theorem

Since each of the quantities Φ_N, Φ_+, Φ_- is defined as the optimal value of a stochastic program with Gaussian parameters, we may apply the (Convex) Gaussian Minimax Theorem ((C)GMT) (Gordon, 1988; Thrampoulidis et al., 2015). On a high level, given a "primary" optimization program with Gaussian parameters, the (C)GMT relates it to an "auxiliary" optimization program, so that high-probability bounds on the latter imply high-probability bounds on the former. The following proposition applies the CGMT on Φ_N and the GMT on Φ_+, Φ_-.

Proposition 1. *For $h \sim \mathcal{N}(0, I_d)$, define the stochastic auxiliary optimization problems:*

$$\phi_N(\rho) = \min_w \|w\|_1 \quad s.t \quad \langle w, h\rangle^2 \ge (1+\rho)n(\sigma^2 + \|w\|_2^2) \tag{A_N}$$

$$\phi_+(\rho) = \max_w \|w\|_2^2 \quad s.t \quad \begin{cases} \|w\|_1 \le B(n,d) \\ \langle w, h\rangle^2 \ge (1-\rho)n(\sigma^2 + \|w\|_2^2) \end{cases} \tag{A_+}$$

$$\phi_-(\rho) = \min_w \|w\|_2^2 \quad s.t \quad \begin{cases} \|w\|_1 \le B(n,d) \\ \langle w, h\rangle^2 \ge (1-\rho)n(\sigma^2 + \|w\|_2^2) \end{cases} \tag{A_-}$$

where $0 < \rho < 1/2$ can be any small enough quantity. For any $t \in \mathbb{R}$, it holds that

$$\begin{aligned} &\mathbb{P}(\Phi_N > t) \le 2\mathbb{P}(\phi_N(\rho) \ge t) + 6\exp\left(-\frac{n\rho^2}{100}\right) \\ \text{and} \quad &\mathbb{P}(\Phi_+ > t) \le 2\mathbb{P}(\phi_+(\rho) \ge t) + 6\exp\left(-\frac{n\rho^2}{100}\right) \\ \text{and} \quad &\mathbb{P}(\Phi_- < t) \le 2\mathbb{P}(\phi_-(\rho) \le t) + 6\exp\left(-\frac{n\rho^2}{100}\right), \end{aligned}$$

where on the left-hand side $\mathbb{P}$ denotes the probability distribution over X and ξ, and on the right-hand side the distribution over h.

For the remainder of this proof, we choose[2]

$$\rho = \frac{10}{\log(d/n)^{5/2}}.$$

[2]This choice of ρ is justified by the proof of Proposition 2. Indeed, for an arbitrary choice of $\rho < 1/2$, one could still show the same bound with just an extra factor: $(\phi_N)^2 \le (1+\rho)M(n,d)$, holding with still the same probability.

As such, from now on, we simply write ϕ_N, ϕ_+, ϕ_-. The proof of Proposition 1, given in Appendix B.2, closely follows Lemmas 3-7 in the paper Koehler et al. (2021). For clarity, note that the three pairs of stochastic programs (P_N/A_N), (P_+/A_+), (P_-/A_-) are not coupled: Proposition 1 should be understood as consisting of three separate statements, each using a different independent copy of h.

As a result of the proposition, the goal of finding high-probability bounds on Φ_N, Φ_+, Φ_- now reduces to finding high-probability bounds on ϕ_N, ϕ_+, ϕ_-, respectively.

A.1.2 Bounds on ϕ_N, ϕ_+, ϕ_-

To obtain tight bounds on the auxiliary quantities ϕ_N, ϕ_+, ϕ_-, we adopt a significantly different approach from previous works. The main idea is to reduce the optimization problems (A_N), (A_+) and (A_-) to optimization problems over a parametric path $\{\gamma(\alpha)\}_\alpha \subset \mathbb{R}^d$. Here we only state the results and refer to Appendix A.2 for their proofs and further intuition. For the remainder of this proof, we denote by $t_n \in \mathbb{R}$ the quantile of the standard normal distribution defined by $2\Phi^{\complement}(t_n) = n/d$.

Proposition 2. *There exist universal constants $\kappa_2, \kappa_3, \kappa_4, c_0 > 0$ such that, if $n \geq \kappa_2$ and $\kappa_3 n \leq d \leq \exp(\kappa_4 n^{1/5})$, then*

$$\phi_N^2 \leq \frac{\sigma^2 n}{t_n^2}\left(1 - \frac{2}{t_n^2} + \frac{c_0}{t_n^4}\right)$$

with probability at least $1 - 6\exp\left(-2\frac{n}{\log(d/n)^5}\right)$ over the draws of h.

Consequently, (P_N) holds with

$$M(n,d) := \sqrt{\frac{\sigma^2 n}{t_n^2}\left(1 - \frac{2}{t_n^2} + \frac{c_0}{t_n^4}\right)}$$

with probability at least $1 - 18\exp\left(-\frac{n}{\log(d/n)^5}\right)$ over the draws of X and ξ.

Hence by Lemma 1, the min-ℓ_1-norm interpolator is located close to the true vector w^*, namely the ℓ_1 distance is bounded by the deterministic quantity

$$\|\hat{w} - w^*\|_1 \leq c\sigma\sqrt{\|w^*\|_0} + M(n,d) =: B(n,d)$$

with probability at least $1 - 18\exp\left(-\frac{n}{\log(d/n)^5}\right) - d\exp(-c_3 n)$ and with $c, c_3 > 0$ some universal constants. We now establish high-probability upper resp. lower bounds for ϕ_+ resp. ϕ_-.

Proposition 3. *Suppose $\|w^*\|_0 \leq \kappa_1 \frac{n}{\log(d/n)^5}$ for some universal constant $\kappa_1 > 0$. There exist universal constants $\kappa_2, \kappa_3, \kappa_4, c_1, c_3 > 0$ such that, if $n \geq \kappa_2$ and $\kappa_3 n \leq d \leq \exp(\kappa_4 n^{1/5})$, then each of the two events*

$$\phi_+ \leq \frac{\sigma^2}{\log(d/n)}\left(1 + \frac{c_1}{\sqrt{\log(d/n)}}\right) \qquad \text{and} \qquad \phi_- \geq \frac{\sigma^2}{\log(d/n)}\left(1 - \frac{c_1}{\sqrt{\log(d/n)}}\right)$$

happens with probability at least $1 - 18\exp\left(-\frac{n}{\log(d/n)^5}\right)$ over the draws of h.

Theorem 1 follows straightforwardly from Lemma 1 and Propositions 1, 2 and 3.

A.2 Proof of Propositions 2 and 3

In this section we detail our analysis of the auxiliary optimization problems (A_N), (A_+) and (A_-). We start by a remark that considerably simplifies notation: The definitions of ϕ_N, ϕ_+, ϕ_- are unchanged if, in (A_N),

This would translate to a bound on Φ_N holding with probability $1 - 12\exp\left(-2\frac{n}{\log(d/n)^5}\right) - 6\exp\left(-\frac{n\rho^2}{100}\right)$. So the choice $\rho = \frac{10}{\log(d/n)^{5/2}}$ "comes at no cost" in terms of the probability with which the bound holds, while being sufficiently small to allow for a satisfactory bound (it only affects the constant c_0 appearing in $M(n,d)$).

(A_+), (A_-), h is replaced by the reordered vector of its absolute order statistics, i.e., by H such that H_i is the i-th largest absolute value of h. Throughout this proof, we condition on the event where H has distinct and positive components: $H_1 > ... > H_d > 0$, which holds with probability one. Henceforth, unless specified otherwise, references to the optimization problems (A_N), (A_+) and (A_-) refer to the equivalent problems where h is replaced by H. Also recall that we choose $\rho = \frac{10}{\log(d/n)^{5/2}}$. The key steps in the proof of Propositions 2 and 3 are as follows.

- For each of the three optimization problems (A_N), (A_+) and (A_-), we show that the argmax (or argmin) is of the form $b\gamma(\alpha)$ for some $b > 0$ and a parametric path $\Gamma = \{\frac{\gamma(\alpha)}{\alpha}\}_\alpha$ (which depends on H). Hence we can restate (A_N), (A_+) and (A_-) as optimization problems over a scalar variable α and a scale variable $b > 0$. (Appendix A.2.1)
- Still conditioning on H, we explicitly characterize the parametric path Γ. In particular, we show that it is piecewise linear with breakpoints $\gamma(\alpha_s)$ having closed-form expressions. (Appendix A.2.2)
- Thanks to the concentration properties of H (Appendix A.2.3), evaluating at one of the breakpoints yields the desired high-probability upper bound on ϕ_N (Appendix A.2.4).
- A fine-grained study of the intersection of $\mathbb{R}_+\Gamma := \{b\frac{\gamma(\alpha)}{\alpha}\}_{b\in\mathbb{R}_+,\alpha}$ with the constraint set of (A_+) and (A_-), as well as the concentration properties of H, yield the desired high-probability bounds on ϕ_+ and ϕ_-. (Appendix A.2.5)

A.2.1 Parametrizing the argmax/argmin

Note that in the optimization problems (A_N), (A_+) and (A_-), the variable w only appears through $\|w\|_2$, $\|w\|_1$ and $\langle w, H\rangle$. Thus, we can add the constraint that $\forall i,\ w_i \geq 0$ without affecting the optimal solution. We will show that the path $\Gamma = \left\{\frac{\gamma(\alpha)}{\alpha}\right\}_\alpha$ can be used to parametrize the solutions of the optimization problems, where $\gamma : [1, \alpha_{\max}] \to \mathbb{R}^d$ is defined by

$$\gamma(\alpha) = \arg\min_w \|w\|_2^2 \quad \text{s.t} \quad \begin{cases} \langle w, H\rangle \geq \|H\|_\infty \\ \forall i, w_i \geq 0 \\ \mathbf{1}^\top w = \|w\|_1 = \alpha \end{cases}$$

and $\alpha_{\max} = d\frac{\|H\|_\infty}{\|H\|_1}$. Specifically, the following key lemma states that (at least one element of) the argmax/argmin of (A_N), (A_+) and (A_-) is of the form $b\frac{\gamma(\alpha)}{\alpha}$ for some $b > 0$ and $\alpha \in [1, \alpha_{\max}]$. This allows to reduce the optimization problems to a single scalar variable and a scale variable.

Lemma 2. *Denoting for concision $B = B(n, d)$, we have that:*

1. *The variable w in (A_N) can equivalently be constrained to belong to the set $\mathbb{R}_+\Gamma$, i.e.,*
$$\phi_N = \min_{b>0, 1\leq\alpha\leq\alpha_{\max}} b \quad s.t \quad b^2 \|H\|_\infty^2 \geq (1+\rho)n(\sigma^2 \|\gamma(\alpha)\|_1^2 + b^2 \|\gamma(\alpha)\|_2^2). \tag{A'_N}$$
2. *The variable w in (A_+) can equivalently be constrained to belong to the set $B\Gamma$, i.e.,*
$$\phi_+ = \max_{1\leq\alpha\leq\alpha_{\max}} B^2 \frac{\|\gamma(\alpha)\|_2^2}{\|\gamma(\alpha)\|_1^2} \quad s.t \quad B^2 \|H\|_\infty^2 \geq (1-\rho)n(\sigma^2 \|\gamma(\alpha)\|_1^2 + B^2 \|\gamma(\alpha)\|_2^2). \tag{A'_+}$$
3. *The variable w in (A_-) can equivalently be constrained to belong to the set $(0, B]\Gamma$, i.e.,*
$$\phi_- = \min_{\substack{0<b\leq B \\ 1\leq\alpha\leq\alpha_{\max}}} b^2 \frac{\|\gamma(\alpha)\|_2^2}{\|\gamma(\alpha)\|_1^2} \quad s.t \quad b^2 \|H\|_\infty^2 \geq (1-\rho)n(\sigma^2 \|\gamma(\alpha)\|_1^2 + b^2 \|\gamma(\alpha)\|_2^2). \tag{A'_-}$$

The proof of the lemma is given in Appendix B.3. To give an intuitive explanation for the equivalence between (A_N) and (A'_N), consider a penalized version of (A_N): $\min_w \|w\|_1 - \lambda\left(\langle w, h\rangle^2 - (1+\rho)n(\sigma^2 + \|w\|_2^2)\right)$ with $\lambda > 0$. For fixed values of $\|w\|_1$ and $\langle w, h\rangle$, minimizing this penalized objective is equivalent to minimizing $\|w\|_2^2$. Hence, we can expect the argmin to be attained at $b\frac{\gamma(\alpha)}{\alpha}$ for some $b > 0, \alpha$.

A.2.2 Characterizing the parametric path

As $\gamma(\alpha)$ is defined as the optimal solution of a convex optimization problem, we are able to obtain a closed-form expression, by a straightforward application of Lagrangian duality. The only other non-trivial ingredient is to notice that, at optimality, the inequality constraint $\langle w, H\rangle \geq \|H\|_\infty$ necessarily holds with equality. Denote $H_{[s]}$ the vector equal to H on the first s components and 0 on the last $(d-s)$, and similarly for $\mathbf{1}_{[s]}$. Define, for any integer $2 \leq s \leq d$,

$$\alpha_s = \frac{\left(\left\|H_{[s]}\right\|_1 - sH_s\right)\|H\|_\infty}{\left\|H_{[s]}\right\|_2^2 - \left\|H_{[s]}\right\|_1 H_s}.$$

Note that $\alpha_2 = 1$. Let $\alpha_{d+1} = \alpha_{\max}$.

Lemma 3. *For all $1 < \alpha \leq \alpha_{\max}$, denote s the unique integer in $\{2, ..., d\}$ such that $\alpha_s < \alpha \leq \alpha_{s+1}$. Then $\gamma(\alpha) = \lambda H_{[s]} - \mu \mathbf{1}_{[s]}$ (in particular it is s-sparse) where the dual variables λ and μ are given by*

$$\lambda = \frac{1}{s\left\|H_{[s]}\right\|_2^2 - \left\|H_{[s]}\right\|_1^2}\left(s\|H\|_\infty - \alpha\left\|H_{[s]}\right\|_1\right)$$
$$\textit{and} \quad \mu = \frac{1}{s\left\|H_{[s]}\right\|_2^2 - \left\|H_{[s]}\right\|_1^2}\left(\left\|H_{[s]}\right\|_1\|H\|_\infty - \alpha\left\|H_{[s]}\right\|_2^2\right).$$

The proof of the lemma is given in Appendix B.4.

A.2.3 Concentration of norms of $\gamma(\alpha_s)$

Given the explicit characterization of the parametric path, we now study its breakpoints $\gamma(\alpha_s)$ ($s \in \{2, ..., d\}$), and more precisely we estimate $\|\gamma(\alpha_s)\|_1 = \alpha_s$ and $\|\gamma(\alpha_s)\|_2$ as a function of s (we have by definition $\langle\gamma(\alpha_s), H\rangle = \|H\|_\infty$). Namely, we prove the following concentration result, where, analogously to t_n, we let $t_s \in \mathbb{R}$ denote the quantity such that $2\Phi^{\complement}(t_s) = s/d$.

Proposition 4. *There exist universal constants $\kappa_2, \kappa_3, \kappa_4, c_1 > 0$ such that for any s, d with $s \geq \kappa_2$ and $\kappa_3 s \leq d \leq \exp(\kappa_4 s^{1/5})$,*

$$\left|\frac{\|\gamma(\alpha_s)\|_1}{\|H\|_\infty} - \left(\frac{1}{t_s} - \frac{2}{t_s^3}\right)\right| \leq \frac{c_1}{t_s^5} \qquad \textit{and} \qquad \left|\frac{\|\gamma(\alpha_s)\|_2^2}{\|H\|_\infty^2} - \frac{2}{st_s^2}\right| \leq \frac{c_1}{st_s^4}, \tag{6}$$

with probability at least $1 - 6\exp\left(-2\frac{s}{\log(d/s)^5}\right)$ over the draws of h.

This proposition relies on and extends the literature studying concentration of order statistics Boucheron and Thomas (2012); Li et al. (2020). An important ingredient for the proof of the proposition is the following lemma, which gives a tight approximation for t_s.

Lemma 4. *There exist universal constants $\kappa_3, c_1 > 0$ such that, for all $s \leq d/\kappa_3$, t_s satisfies*

$$\bar{t}_s^2 - c_1 \leq t_s^2 \leq \bar{t}_s^2$$

where

$$\bar{t}_s = \sqrt{2\log(d/s) - \log\log(d/s) - \log(\frac{\pi}{2})}.$$

Furthermore, κ_3 and c_1 can be chosen (e.g. $\kappa_3 = 11$ and $c_1 = 1$) such that $\log(d/s) \leq t_s^2 \leq 2\log(d/s)$.

The proofs of Proposition 4 and of Lemma 4 are given in Appendix B.5.

A.2.4 Localization: Proof of Proposition 2 (upper bound for ϕ_N)

We now use the concentration bounds of Proposition 4 to obtain a high-probability upper bound for ϕ_N. Recall from Lemma 2 that it is given by (A_N'):

$$(\phi_N)^2 = \min_{b>0, 1\leq\alpha\leq\alpha_{\max}} b^2 \quad \text{s.t} \quad b^2\|H\|_\infty^2 \geq (1+\rho)n(\sigma^2\|\gamma(\alpha)\|_1^2 + b^2\|\gamma(\alpha)\|_2^2).$$

We may rewrite the constraint as

$$b^2 \|H\|_\infty^2 \left(1 - (1+\rho)n\frac{\|\gamma(\alpha)\|_2^2)}{\|H\|_\infty^2}\right) \geq (1+\rho)n\sigma^2 \|\gamma(\alpha)\|_1^2$$
$$\iff b^2 \geq \underbrace{\frac{\|\gamma(\alpha)\|_1^2}{\|H\|_\infty^2} \frac{\sigma^2 n(1+\rho)}{1-(1+\rho)n\frac{\|\gamma(\alpha)\|_2^2}{\|H\|_\infty^2}}}_{=:f(\alpha)^2} \quad \text{and} \quad (1+\rho)n\frac{\|\gamma(\alpha)\|_2^2}{\|H\|_\infty^2} < 1.$$

Thus minimizing over b shows that $(\phi_N)^2 = \min_{1\leq\alpha\leq\alpha_{\max}} f(\alpha)^2$ s.t $(1+\rho)n\frac{\|\gamma(\alpha)\|_2^2}{\|H\|_\infty^2} < 1$. Since we want to upper-bound this minimum, it is sufficient to further restrict the optimization problem by the constraint $\alpha \in \{\alpha_s | s \in \{2, ..., d\}\}$, yielding

$$(\phi_N)^2 \leq \min_{2\leq s\leq d+1} f(\alpha_s)^2 \ \text{ s.t } \ (1+\rho)n\frac{\|\gamma(\alpha_s)\|_2^2}{\|H\|_\infty^2} < 1.$$

We now show that for the choice $s = n$, the constraint is satisfied with high probability, and we give a high-probability estimate for the resulting upper bound $f(\alpha_n)^2$. See Remark 1 below for a justification of this choice. For the remainder of the proof of Proposition 2, we condition on the event where the inequalities in Equation (6) hold for $s = n$. By the concentration bound for $\|\gamma(\alpha_n)\|_2^2$, a sufficient condition for the choice $s = n$ to be feasible is

$$(1+\rho)\frac{2}{t_n^2}\left(1 + \frac{c_1}{t_n^2}\right) < 1$$

with $c_1 > 0$ some universal constant. Now $t_n^2 \geq \log(d/n)$ by Lemma 4, and recall that $\rho = \frac{10}{\log(d/n)^{5/2}}$. For κ_3 sufficiently large, the above inequality holds for any n, d with $\kappa_3 n \leq d$. Moreover, by the concentration bounds for $\|\gamma(\alpha_n)\|_2^2$ and $\|\gamma(\alpha_n)\|_1$, $f(\alpha_n)^2$ is upper-bounded by

$$f(\alpha_n)^2 \leq \frac{(1 - \frac{4}{t_n^2} + O(\frac{1}{t_n^4}))\sigma^2 n(1+\rho)}{t_n^2 - 2(1+\rho)(1+O(\frac{1}{t_n^2}))} \leq \frac{\sigma^2 n}{t_n^2}(1+\rho)\left(1 - \frac{2}{t_n^2} + O(\frac{1}{t_n^4})\right).$$

Furthermore, $\rho = O\left(\frac{1}{t_n^5}\right)$ by Lemma 4, so $f(\alpha_n)^2 \leq \frac{\sigma^2 n}{t_n^2}\left(1 - \frac{2}{t_n^2} + \frac{c_0}{t_n^4}\right) =: M(n,d)^2$ for a universal constant $c_0 > 0$. This concludes the proof of Proposition 2.

Remark 1. *Let us informally justify why we can expect the choice $s = n$ to approximately minimize $f(\alpha_s)$. A first justification is that the min-ℓ_1-norm interpolator $\hat{w}$, which is the solution of the optimization problem (P_N), is well-known to be n-sparse. Since the optimization problems (P_N) and (A_N) are intimately connected via the CGMT (Proposition 1), we can expect the optimal solution of (A_N) to have similar properties to $\hat{w}$ – in particular, to have the same sparsity $s = n$. A second, more technical, justification is as follows. Note that if we replace $\|\gamma(\alpha_s)\|_2^2$ and $\|\gamma(\alpha_s)\|_1$ by their estimates from Proposition 4 and ignore the higher-order terms, we have*

$$f(\alpha_s)^2 \approx \frac{1}{t_s^2}\frac{\sigma^2 n(1+\rho)}{1-(1+\rho)n\frac{2}{st_s^2}} = \frac{\sigma^2(1+\rho)}{\frac{t_s^2}{n} - (1+\rho)\frac{2}{s}}.$$

Thus a good choice for s is given by maximizing the denominator. By using the estimate $t_s^2 \approx 2\log(d/s)$ from Lemma 4, we can approximate it by

$$\frac{t_s^2}{n} - (1+\rho)\frac{2}{s} \approx \frac{2\log(d/s)}{n} - (1+\rho)\frac{2}{s} =: g(s).$$

Interpreting s as a continuous variable and setting $\frac{d}{ds}g(s) = 0$ yields the choice $s = (1+\rho)n \approx n$.

A.2.5 Uniform convergence: Proof of Proposition 3 (bounds for ϕ_+ and ϕ_-)

To obtain an upper bound for the maximization problem defining ϕ_+ (resp. lower bound for the minimization for ϕ_-), a typical approach would be to find a tractable relaxation of the problem. However, the more obvious relaxations already explored in the paper Koehler et al. (2021) turn out to be unsatisfactorily loose, as discussed in Section 3. Here, thanks to the one-dimensional structure of our reformulations (A'_+) and (A'_-), we take a different approach and study the monotonicity of the objectives.

We can decompose our proof in three steps. Firstly, we describe our overall monotonicity-based approach. Secondly, we find values $\underline{\alpha}, \overline{\alpha}$ that allow us to unroll our approach. Finally, we evaluate the bound that the first two steps give us, thus proving the proposition.

Step 1: Studying the feasible set of (A'_+) and (A'_-). Recall that ϕ_+, ϕ_- are respectively given by Equations (A'_+), (A'_-). We can also write them in the following form, using the fact that $\|\gamma(\alpha)\|_1 = \alpha$:

$$\phi_+ = \max_{1\leq\alpha\leq\alpha_{\max}} B^2\frac{\|\gamma(\alpha)\|_2^2}{\alpha^2} \text{ s.t } B^2\|H\|_\infty^2 \geq (1-\rho)n(\sigma^2\alpha^2 + B^2\|\gamma(\alpha)\|_2^2)$$

$$\phi_- = \min_{\substack{0<b\leq B\\ 1\leq\alpha\leq\alpha_{\max}}} b^2\frac{\|\gamma(\alpha)\|_2^2}{\alpha^2} \text{ s.t } b^2\|H\|_\infty^2 \geq (1-\rho)n(\sigma^2\alpha^2 + b^2\|\gamma(\alpha)\|_2^2).$$

We first study the sets of feasible solutions of (A'_+) and (A'_-). Denote by I the former set, i.e.,

$$I := \left\{\alpha\in[1,\alpha_{\max}] \;\middle|\; B^2\|H\|_\infty^2 \geq (1-\rho)n\left(\sigma^2\alpha^2 + B^2\|\gamma(\alpha)\|_2^2\right)\right\}.$$

Also let $\alpha_{d+1/2} = \frac{\|H\|_1\|H\|_\infty}{\|H\|_2^2}$; this choice of notation is purely symbolic, and is justified by the fact that $\alpha_d < \alpha_{d+1/2} < \alpha_{d+1}$.

Lemma 5. *The following statements hold:*

1. *The mapping $\alpha\mapsto\|\gamma(\alpha)\|_2^2$ is decreasing over $[1,\alpha_{d+1/2}]$ and increasing over $[\alpha_{d+1/2},\alpha_{\max}]$.*
2. *The mapping $\alpha\mapsto\|\gamma(\alpha)\|_2^2$ is convex over $[1,\alpha_{\max}]$, and I is an interval.*
3. *The mapping $\alpha\mapsto\frac{\|\gamma(\alpha)\|_2^2}{\alpha^2}$ is monotonically decreasing.*

These monotonicity properties lead us to a proof strategy that can be summarized as follows.

Lemma 6. *Denote $I = [\underline{\alpha}_I, \overline{\alpha}_I]$ the endpoints of I. For any $\underline{\alpha}\leq\underline{\alpha}_I$,*

$$\phi_+ \leq B^2\frac{\|\gamma(\underline{\alpha})\|_2^2}{\underline{\alpha}^2}.$$

If $\overline{\alpha}_I < \alpha_{d+1/2}$, then for any $\overline{\alpha}_I\leq\overline{\alpha}\leq\alpha_{d+1/2}$,

$$\phi_- \geq \frac{\sigma^2 n(1-\rho)}{\|H\|_\infty^2 - (1-\rho)n\|\gamma(\overline{\alpha})\|_2^2}\|\gamma(\overline{\alpha})\|_2^2.$$

The proofs of Lemma 5 and Lemma 6 are given in Appendix B.6.

Step 2: A tight admissible choice for $\underline{\alpha}$ and $\overline{\alpha}$. To apply Lemma 6 and obtain bounds on ϕ_+, ϕ_-, we need to find $\underline{\alpha}$ and $\overline{\alpha}$ lying on the left, respectively on the right of the interval I, and such that $\overline{\alpha}\leq\alpha_{d+1/2}$. By having a closer look at the way we derived the expression of $M(n,d)$, we have by construction that with high probability, $\alpha_n\in I$. In fact, we show that there exist integers $\underline{s}$ and $\overline{s}$ very close to n such that $\alpha_{\underline{s}}$ already falls to the left of I, and $\alpha_{\overline{s}}$ to the right of I, with high probability.

Lemma 7. *Suppose $\|w^*\|_1\leq\kappa_1\sqrt{\frac{\sigma^2 n}{\log(d/n)^5}}$ for some universal constant $\kappa_1>0$. There exist universal constants $\kappa_2,\kappa_3,\kappa_4,\lambda>0$ such that, for any d,n with $n\geq\kappa_2$ and $\kappa_3 n\leq d\leq\exp(\kappa_4 n^{1/5})$, we can find integers $\underline{s},\overline{s}\in\mathbb{N}_+$ satisfying*

$$\underline{s} = n\exp\left(-\frac{\lambda}{2t_n}\right)\left(1+O\left(\frac{1}{t_n^2}\right)\right) \quad \textit{and} \quad \overline{s} = n\exp\left(\frac{\lambda}{2t_n}\right)\left(1+O\left(\frac{1}{t_n^2}\right)\right) \tag{7}$$

and

$$\alpha_{\underline{s}} < \underline{\alpha}_I \leq \alpha_n \leq \overline{\alpha}_I \leq \alpha_{\overline{s}} \leq \alpha_{d+1/2},$$

with probability at least $1 - 18\exp\left(-\frac{n}{\log(d/n)^5}\right)$ *over the draws of* h*. Moreover,* $t_{\underline{s}}^2 = t_n^2 + O(1)$ *and* $t_{\overline{s}}^2 = t_n^2 + O(1)$*.*

The proof of the lemma is given in Appendix B.7. It relies in particular on the assumption that $\|w^*\|_0 \leq \kappa_1 \frac{n}{\log(d/n)^5}$ for some universal constant κ_1, which implies that $M(n,d)$ from Proposition 2 is the dominating term in B, and hence $B^2 = (M(n,d) + c\sigma\sqrt{\|w^*\|_0})^2 = \frac{\sigma^2 n}{t_n^2}\left(1 - \frac{2}{t_n^2} + O\left(\frac{1}{t_n^4}\right)\right)$. Furthermore, the equations in the lemma hold true conditionally on the event where the inequalities in Equation (6) hold simultaneously for $s = n$, $s = \underline{s}$, and $s = \overline{s}$ – which indeed occurs with the announced probability. These two elements of the proof will be reused in the following step.

Step 3: Applying Lemma 6. Lemma 7 provides us with a choice of $\underline{\alpha} = \alpha_{\underline{s}}$ and $\overline{\alpha} = \alpha_{\overline{s}}$ that satisfy the conditions of Lemma 6 with high probability. To conclude the proof of Proposition 3, all that remains to be done is to compute the bounds given by Lemma 6, i.e.,

$$\phi_+ \leq B^2 \frac{\|\gamma(\alpha_{\underline{s}})\|_2^2}{\|\gamma(\alpha_{\underline{s}})\|_1^2} \quad \text{and} \quad \phi_- \geq \frac{\sigma^2 n(1-\rho)}{\|H\|_\infty^2 - (1-\rho)n\|\gamma(\alpha_{\overline{s}})\|_2^2} \|\gamma(\alpha_{\overline{s}})^2\|_2^2.$$

For the remainder of the proof of Proposition 3, we condition on the event where the inequalities in Equation (6) hold simultaneously for $s = n$, $s = \underline{s}$, and $s = \overline{s}$. In particular, the conclusions of Lemma 7 hold, as discussed just above. We also recall that, because of the assumption on the growth of $\|w^*\|_0$, we have $B^2 = \frac{\sigma^2 n}{t_n^2}\left(1 - \frac{2}{t_n^2} + O\left(\frac{1}{t_n^4}\right)\right)$. By applying the concentration inequalities from Equation (6), and using the above estimate for B^2, we obtain

$$\phi_+ \leq \frac{\sigma^2 n}{t_n^2}\left(1 + O\left(\frac{1}{t_n^2}\right)\right)\frac{2}{\underline{s}t_{\underline{s}}^2} t_{\underline{s}}^2\left(1 + O\left(\frac{1}{t_{\underline{s}}^2}\right)\right) \quad \text{and} \quad \phi_- \geq \frac{2\sigma^2 n}{\overline{s}t_{\overline{s}}^2}\left(1 + O\left(\frac{1}{t_n^2}\right)\right).$$

By plugging in the approximate expressions of $\underline{s}$ and $\overline{s}$ from Equation (7), as well as the estimates $t_{\underline{s}}^2 = t_n^2 + O(1)$ and $t_{\overline{s}}^2 = t_n^2 + O(1)$ from Lemma 7, we further obtain

$$\phi_+ \leq \frac{2\sigma^2}{t_n^2}\exp\left(\frac{\lambda}{2t_n}\right)\left(1 + O\left(\frac{1}{t_n^2}\right)\right) \quad \text{and} \quad \phi_- \geq \frac{2\sigma^2}{t_n^2}\exp\left(-\frac{\lambda}{2t_n}\right)\left(1 + O\left(\frac{1}{t_n^2}\right)\right).$$

Finally, by the expansion $t_n^2 = 2\log(d/n) + O(\log\log(d/n))$ from Lemma 4 and by the Taylor series approximation $\exp(x) = 1 + x + O(x^2)$ (for bounded x), we obtain the desired bounds

$$\phi_+ \leq \frac{\sigma^2}{\log(d/n)}\left(1 + O\left(\frac{1}{\sqrt{\log(d/n)}}\right)\right) \quad \text{and} \quad \phi_- \geq \frac{\sigma^2}{\log(d/n)}\left(1 + O\left(\frac{1}{\sqrt{\log(d/n)}}\right)\right).$$

This concludes the proof of Proposition 3.

B PROOF DETAILS

In this appendix, we provide details of the proof of our main result, Theorem 1, omitted in Section A. We refer to that section for notation.

B.1 Proof of Lemma 1: Preliminary

Let $\mathcal{S} := \{j : w_j^* \neq 0\}$ (in particular, $\|w^*\|_0 = |\mathcal{S}|$). Denote by $\hat{w}_{\mathcal{S}} \in \mathbb{R}^d$ the vector with entries $(\hat{w}_{\mathcal{S}})_j = \hat{w}_j$ if $j \in \mathcal{S}$ and 0 otherwise, and let $\hat{w}_{-\mathcal{S}} = \hat{w} - \hat{w}_{\mathcal{S}}$. Now recall the definition of $\hat{w}$, and define $\hat{v}$ by

$$\hat{w} = \arg\min_w \|w\|_1 \quad \text{s.t} \quad X(w - w^*) = \xi$$
$$\hat{v} = \arg\min_v \|v\|_1 \quad \text{s.t} \quad Xv = \xi.$$

Clearly by definition of $\hat{v}$ and $\hat{w}$, we have that $\|\hat{w}\|_1 \leq \|\hat{v}\|_1 + \|w^*\|_1$. Therefore

$$\begin{aligned}0 &\leq \|w^*\|_1 - \|\hat{w}\|_1 + \|\hat{v}\|_1 \\ &= \|w^*\|_1 - \|\hat{w}_{\mathcal{S}}\|_1 - \|\hat{w}_{-\mathcal{S}}\| + \|\hat{v}\|_1 \\ &\leq \|w^* - \hat{w}_{\mathcal{S}}\|_1 - \|\hat{w}_{-\mathcal{S}}\|_1 + \|\hat{v}\|_1 \\ &= 2\|w^* - \hat{w}_{\mathcal{S}}\|_1 - \|\hat{w} - w^*\|_1 + \|\hat{v}\|_1 \\ &\leq 2\sqrt{\|w^*\|_0}\,\|w^* - \hat{w}_{\mathcal{S}}\|_2 - \|\hat{w} - w^*\|_1 + \|\hat{v}\|_1 .\end{aligned}$$

Hence,

$$\begin{aligned}\|\hat{w} - w^*\|_1 &\leq 2\sqrt{\|w^*\|_0}\,\|w^* - \hat{w}_{\mathcal{S}}\|_2 + \|\hat{v}\|_1 \\ &\leq 2\sqrt{\|w^*\|_0}\,\|w^* - \hat{w}\|_2 + \|\hat{v}\|_1 .\end{aligned}$$

Finally, we bound $\|w^* - \hat{w}\|_2$ by applying Theorem 3.1 of Chinot et al. (2021), noting that its assumptions are subsumed by the assumptions of our Theorem 1.

B.2 Proof of Proposition 1: Application of the (C)GMT

Proposition 1 reduces the estimation of the quantities Φ_N, Φ_+, Φ_- in Equations (P_N), (P_+), (P_-) to the estimation of auxiliary quantities ϕ_N, ϕ_+, ϕ_-, using the (C)GMT.

As a first step, we apply the CGMT to Φ_N and the GMT to Φ_+ analogously to (Koehler et al., 2021, Lemmas 4&7). We only restate the results here and refer the reader to that paper for details and proofs. Note that the (C)GMT is applied on X conditionally on ξ, so that the Gaussianity of the noise is not crucial.

Lemma 8 ((Koehler et al., 2021, Lemma 7), Application of CGMT). *Define*

$$\tilde{\phi}_N = \min_w \|w\|_1 \quad s.t \quad \|\xi - g\|w\|_2\|_2 \leq \langle w, h\rangle ,$$

where $g \sim \mathcal{N}(0, I_n)$ and $h \sim \mathcal{N}(0, I_d)$ are independent random variables. Then, for all $t \in \mathbb{R}$,

$$\mathbb{P}(\Phi_N > t) \leq 2\mathbb{P}(\tilde{\phi}_N > t),$$

where the probabilities on the left and on the right are over the draws of X, ξ and of g, h, ξ, respectively.

Lemma 9 ((Koehler et al., 2021, Lemma 4), Application of GMT). *Define*

$$\tilde{\phi}_+ = \max_w \|w\|_2^2 \quad s.t \quad \begin{cases}\|w\|_1 \leq B(n,d) \\ \|\xi - g\|w\|_2\|_2 \leq \langle w, h\rangle\end{cases} ,$$

where $g \sim \mathcal{N}(0, I_n)$ and $h \sim \mathcal{N}(0, I_d)$ are independent random variables. Then, for all $t \in \mathbb{R}$,

$$\mathbb{P}(\Phi_+ > t) \leq 2\mathbb{P}(\tilde{\phi}_+ > t),$$

where the probabilities on the left and on the right are over the draws of X, ξ and of g, h, ξ, respectively.

Following the same argument as in (Koehler et al., 2021, Lemma 4), we can show a corresponding lemma for Φ_- which we state without proof:

Lemma 10 (Application of GMT). *Define*

$$\tilde{\phi}_- = \min_w \|w\|_2^2 \quad s.t \quad \begin{cases}\|w\|_1 \leq B(n,d) \\ \|\xi - g\|w\|_2\|_2 \leq \langle w, h\rangle\end{cases} ,$$

where $g \sim \mathcal{N}(0, I_n)$ and $h \sim \mathcal{N}(0, I_d)$ are independent random variables. Then, for all $t \in \mathbb{R}$,

$$\mathbb{P}(\Phi_- < t) \leq 2\mathbb{P}(\tilde{\phi}_- < t),$$

where the probabilities on the left and on the right are over the draws of X, ξ and of g, h, ξ, respectively.

Next, by using Gaussian concentration results, we can formulate simpler versions of the above optimization problems defining the quantities $\tilde{\phi}$'s.

Simplify $\tilde{\phi}_N$. Following the same argument as in the first part of the proof of (Koehler et al., 2021, Lemma 8) (Equations (68)-(70)), we can show that for any $0 < \rho < 1/2$, with probability at least $1 - 6\exp(-n\rho^2/100)$, uniformly over w,

$$\|\xi - g\|w\|_2\|_2^2 \leq (1+\rho)n(\sigma^2 + \|w\|_2^2). \tag{8}$$

So on the event where Equation (8) holds, we have that

$$\tilde{\phi}_N \leq \min_w \|w\|_1 \quad \text{s.t} \quad \langle w, h\rangle^2 \geq (1+\rho)n(\sigma^2 + \|w\|_2^2) = \phi_N$$

which proves the first inequality in Proposition 1.

Simplify $\tilde{\phi}_+, \tilde{\phi}_-$. By the same argument as for $\tilde{\phi}_N$, we can show that for any $0 < \rho < 1/2$, with probability at least $1 - 6\exp(-n\rho^2/100)$, uniformly over w,

$$\|\xi - g\|w\|_2\|_2^2 \geq (1-\rho)n(\sigma^2 + \|w\|_2^2). \tag{9}$$

So on the event where Equation (9) holds, we have that

$$\tilde{\phi}_+ \leq \max_w \|w\|_2^2 \quad \text{s.t} \quad \begin{cases} \|w\|_1 \leq M + 2\|w^*\|_1 \\ \langle w, h\rangle^2 \geq (1-\rho)n(\sigma^2 + \|w\|_2^2) \end{cases} = \phi_+,$$

and similarly $\tilde{\phi}_- \geq \phi_-$. This proves the second and third inequalities in Proposition 1 and thus completes the proof.

B.3 Proof of Lemma 2: Parametrizing the argmax/argmin

We now prove our first key lemma: we show that, up to scaling, the argmax/argmin in (A_N), (A_+), and (A_-) belong to a certain parametric path $\Gamma = \{\frac{\gamma(\alpha)}{\alpha}\}_\alpha$. Throughout this section and the next, we consider H as a fixed vector such that $H_1 > ... > H_d > 0$. In other words, all of our statements should be understood as holding conditionally on h, and with h in general position.

For all $\beta \in [\frac{1}{d}, 1]$, define

$$\overline{\gamma}(\beta) = \arg\max_w \langle w, H\rangle \quad \text{s.t} \quad \begin{cases} \|w\|_2^2 \leq \beta \\ \forall i, w_i \geq 0 \\ \mathbf{1}^\top w = \|w\|_1 = 1 \end{cases}. \tag{10}$$

Importantly, note that the constraint $\|w\|_2^2 \leq \beta$ in the definition of $\overline{\gamma}(\beta)$ necessarily holds with equality at optimality. Indeed, suppose by contradiction $\|\overline{\gamma}(\beta)\|_2^2 < \beta \leq 1 = \|\overline{\gamma}(\beta)\|_1^2$. This implies that $\overline{\gamma}(\beta)$ has at least two nonzero components; denote $i \neq 1$ such that $\overline{\gamma}(\beta)_i > 0$. Then there exists some $\varepsilon > 0$ such that $\overline{\gamma}(\beta) + \varepsilon e_1 - \varepsilon e_i$ satisfies the constraints and achieves a higher objective value than $\overline{\gamma}(\beta)$, contradicting its optimality.

The first step of the proof is to show that (at least one element) of the argmax/argmin belongs to the set $\mathbb{R}_+\overline{\Gamma}$, where $\overline{\Gamma} = \left\{\overline{\gamma}(\beta); \frac{1}{d} \leq \beta \leq 1\right\}$.

Claim 1. *For each of the optimization problems (A_N), (A_+), and (A_-), there exist $b > 0$ and $\beta \in [\frac{1}{d}, 1]$ such that $b\overline{\gamma}(\beta)$ is an optimal solution.*

Proof. Let v be an optimal solution of (A_N). It is straightforward to check that we may assume w.l.o.g. that $\forall i, v_i \geq 0$. Choose $b = \|v\|_1$ and $\beta = \frac{\|v\|_2^2}{\|v\|_1^2}$; note that $\beta \in [\frac{1}{d}, 1]$. By definition, $\|b\overline{\gamma}(\beta)\|_2 = \|v\|_2$ and $\|b\overline{\gamma}(\beta)\|_1 = \|v\|_1$, and v/b is feasible for (10) so $\langle b\overline{\gamma}(\beta), H\rangle \geq \langle v, H\rangle$. Therefore, $b\overline{\gamma}(\beta)$ satisfies the constraint of (A_N) and achieves the optimal objective value, so is also an optimal solution of (A_N).

The statements for (A_+) and (A_-) follow by the exact same argument. □

Next, we show that $\{\overline{\gamma}(\beta)\}_\beta$ and $\{\frac{\gamma(\alpha)}{\alpha}\}_\alpha$ are two parametrizations of the same path.

Claim 2. *We have the equality*

$$\overline{\Gamma} := \left\{\overline{\gamma}(\beta); \beta \in \left[\frac{1}{d}, 1\right]\right\} \quad = \quad \left\{\frac{\gamma(\alpha)}{\alpha}; \alpha \in [1, \alpha_{\max}]\right\} =: \Gamma$$

where $\alpha_{\max} = d\frac{\|H\|_\infty}{\|H\|_1}$.

Proof. First note that we can characterize $\frac{\gamma(\alpha)}{\alpha}$ as the optimal solution of

$$\frac{\gamma(\alpha)}{\alpha} = \arg\min_w \|w\|_2^2 \quad \text{s.t} \quad \begin{cases} \langle w, H\rangle \geq \frac{\|H\|_\infty}{\alpha} \\ \forall i, w_i \geq 0 \\ \mathbf{1}^\top w = \|w\|_1 = 1 \end{cases}. \tag{11}$$

The optimization problems (10) and (11) are both convex and both satisfy the Linear Independence Constraint Qualification conditions. So, denoting $\Delta_d = \left\{w \in [0,1]^d; \mathbf{1}^\top w = 1\right\}$ the standard simplex, by the Lagrangian duality theorem (a.k.a. Karush-Kuhn-Tucker theorem) we have that for all $w \in \mathbb{R}^d$,

$$\begin{aligned} \exists \beta > 0; w = \overline{\gamma}(\beta) \quad &\iff \quad \exists \lambda > 0; w = \arg\max_{w \in \Delta_d} \langle w, H\rangle - \lambda \|w\|_2^2 \\ &\iff \quad \exists \mu > 0; w = \arg\min_{w\in\Delta_d} \|w\|_2^2 - \mu \langle w, H\rangle \quad \iff \quad \exists \alpha > 0; w = \frac{\gamma(\alpha)}{\alpha}. \end{aligned}$$

Thus, $\{\overline{\gamma}(\beta); \beta > 0\} = \left\{\frac{\gamma(\alpha)}{\alpha}; \alpha > 0\right\}$. However it is straightforward to check that $\{\overline{\gamma}(\beta); \beta > 0\} = \overline{\Gamma}$ and that $\left\{\frac{\gamma(\alpha)}{\alpha}; \alpha > 0\right\} = \Gamma$, which concludes the proof. □

Just as the first constraint in (10) holds with equality at optimality, so does the first constraint in (11); that is, $\langle \gamma(\alpha), H\rangle = \|H\|_\infty$ for all $\alpha \in [1, \alpha_{\max}]$. This would follow from a careful study of the equivalence between the two problems, but here we give a more direct proof.

Claim 3. *The inequality constraint $\langle w, H\rangle \geq \|H\|_\infty$ in the problem defining $\gamma(\alpha)$ holds with equality, at optimality.*

Proof. Denote $w = \gamma(\alpha)$. Suppose by contradiction $\langle w, H\rangle > \|H\|_\infty$. Let i resp. j the index of the largest resp. smallest component of w. First note that if $w_i = w_j$, then $w \propto \mathbf{1}$, i.e., $w = \frac{\alpha}{d}\mathbf{1}$ and so $\langle w, H\rangle = \frac{\alpha}{d}\|H\|_1 > \|H\|_\infty$, which would contradict $\alpha \leq \alpha_{\max}$; so we have the strict inequality $w_i > w_j$. Now for some $\varepsilon > 0$ to be chosen, let $w' = w - \varepsilon e_i + \varepsilon e_j$. Clearly $\varepsilon > 0$ can be chosen small enough so that w' satisfies all three constraints in the optimization problem defining $\gamma(\alpha)$. Furthermore, for small enough ε, $\|w\|_2^2 - \|w'\|_2^2 = w_i^2 - (w_i - \varepsilon)^2 + w_j^2 - (w_j + \varepsilon)^2 = 2\varepsilon(w_i - w_j - \varepsilon)$ is positive, i.e $\|w'\|_2^2 < \|w\|_2^2$, which contradicts optimality of $w = \gamma(\alpha)$. □

We now have all the necessary ingredients to prove Lemma 2. The equivalence between (A_N) and (A'_N) follows immediately from constraining the variable w (in the former) to belong to the set $\mathbb{R}_+\Gamma$. The equivalence between (A_-) and (A'_-) also follows immediately, noting that $\left\|b\frac{\gamma(\alpha)}{\alpha}\right\|_1 = b$ by definition. Finally, the equivalence between (A_+) and (A'_+) follows by noticing that the inequality constraint $\|w\|_1 \leq B$ (in the former) is necessarily saturated at optimality.

B.4 Proof of Lemma 3: Characterizing the parametric path

We now give a precise characterization of the parametric path Γ, by studying the optimization problem defining $\gamma(\alpha)$. Throughout this section (just as in the previous one), we consider H as a fixed vector such that $H_1 > ... > H_d > 0$. In other words, all of our statements should be understood as holding conditionally on h, and with h in general position.

Throughout the proof, consider a fixed $1 < \alpha \leq \alpha_{\max}$. The goal is to derive a closed-form expression of $\gamma(\alpha)$. We proceed by a Lagrangian duality approach, and first identify the dual variables λ, ν, μ (a.k.a. Lagrangian multipliers, a.k.a. KKT vectors) of the optimization problem defining $\gamma(\alpha)$. This first analysis yields an expression for $\gamma(\alpha)$ involving an unknown "sparsity" integer s, which depends on λ, ν, μ and hence indirectly on α. We finish by showing how to determine s explicitly from α.

Karush-Kuhn-Tucker (KKT) conditions. Recall that $\gamma(\alpha)$ is defined by the following optimization problem (note that the additional factor $\frac{1}{2}$ in the objective does not change the arg min):

$$\gamma(\alpha) = \arg\min_w \frac{1}{2} \|w\|_2^2 \quad \text{s.t} \quad \begin{cases} \langle w, H \rangle \geq \|H\|_\infty \\ \forall i, w_i \geq 0 \\ \mathbf{1}^\top w = \|w\|_1 = \alpha \end{cases} .$$

This is a convex optimization problem with Lagrangian

$$L(w; \lambda, \mu, \nu) = \frac{1}{2} \|w\|_2^2 - \lambda(\langle w, H \rangle - \|H\|_\infty) + \mu(\mathbf{1}^\top w - \alpha) - \nu^\top w.$$

The objective is convex and all the constraints are affine. So by Lagrangian duality, $w = \gamma(\alpha)$ if and only if there exist $\lambda, \mu \in \mathbb{R}$ and $\nu \in \mathbb{R}^d$ satisfying the KKT conditions:

- (Stationarity) $w - \lambda H + \mu \mathbf{1} - \nu = 0$ i.e. $w = \lambda H - \mu \mathbf{1} + \nu$
- (Primal feasibility) $\begin{cases} \langle w, H \rangle \geq \|H\|_\infty \\ \forall i, \; w_i \geq 0 \\ \mathbf{1}^\top w = \alpha \end{cases}$
- (Dual feasibility) $\lambda \geq 0$ and $\forall i, \; \nu_i \geq 0$
- (Complementary slackness) $\lambda(\langle w, H \rangle - \|H\|_\infty) = 0$, and $\forall i, \; \nu_i w_i = 0$.

In the rest of this proof, denote $w = \gamma(\alpha)$, and let λ, μ, ν as above.

Sparsity structure of w. Let s denote the largest $s' \in \{1, ..., d\}$ such that $\lambda H_{s'} > \mu$. Since $\lambda \geq 0$ and H is ordered, we have

$$\lambda H_1 \geq ... \geq \lambda H_s > \mu \geq \lambda H_{s+1} \geq ...$$

Consider the complementary slackness condition $\forall i, \; \nu_i w_i = 0$.

- If $w_i > 0$, then $\nu_i = 0$ so $w_i = \lambda H_i - \mu > 0$, and so $i \leq s$.
- If $\nu_i > 0$, then $w_i = \lambda H_i - \mu + \nu_i = 0$ so $\lambda H_i - \mu < 0$, and so $i > s$.
  So by contraposition, for all $i \leq s$, $\nu_i = 0$ and $w_i = \lambda H_i - \mu$.

Thus, $\text{supp}(w) \subset \{1, ..., s\}$ and $w = \lambda H_{[s]} - \mu \mathbf{1}_{[s]}$, where $H_{[s]}$ is the vector equal to H on the first s components and 0 on the last $(d - s)$, and similarly for $\mathbf{1}_{[s]}$.

Furthermore, note that the case $s = 1$ occurs only if $w \propto e_1$, and one can check that it implies $\alpha = 1$, which we excluded.

Closed-form expression of the dual variables λ, μ. We can compute λ and μ by substituting $w = \lambda H_{[s]} - \mu \mathbf{1}_{[s]}$ into the primal feasibility conditions.

- Since we know from Claim 3 (in Section B.3) that the first constraint in the problem defining $\gamma(\alpha)$ holds with equality at optimality, this means that the first primal feasibility condition holds with equality, i.e.,

$$\langle w, H \rangle = \lambda \left\|H_{[s]}\right\|_2^2 - \mu \left\|H_{[s]}\right\| = \|H\|_\infty .$$

- By the last primal feasibility condition, $\mathbf{1}^\top w = \lambda \left\|H_{[s]}\right\|_1 - \mu s = \alpha$.

So λ and μ are given by

$$\begin{cases} \lambda \left\|H_{[s]}\right\|_2^2 - \mu \left\|H_{[s]}\right\| & = \|H\|_\infty \\ \lambda \left\|H_{[s]}\right\|_1 - \mu s & = \alpha \end{cases} \iff \begin{cases} \lambda & = \frac{1}{s\left\|H_{[s]}\right\|_2^2 - \left\|H_{[s]}\right\|_1^2} \left(s \|H\|_\infty - \alpha \left\|H_{[s]}\right\|_1\right) \\ \mu & = \frac{1}{s\left\|H_{[s]}\right\|_2^2 - \left\|H_{[s]}\right\|_1^2} \left(\left\|H_{[s]}\right\|_1 \|H\|_\infty - \alpha \left\|H_{[s]}\right\|_2^2\right) \end{cases}.$$

Note that the denominator is positive, since $H_{[s]}$ has distinct components.

Closed-form characterization of s. We now show that there exists an increasing sequence $\alpha_2 = 1 < \cdots < \alpha_d < \alpha_{d+1} = \alpha_{\max}$ such that for all α, the sparsity s of w is exactly the index which satisfies $\alpha \in (\alpha_s, \alpha_{s+1}]$. By plugging the expressions of λ and μ into the condition defining s: $\lambda H_s > \mu \geq \lambda H_{s+1}$, we obtain

$$\left(s \|H\|_\infty - \alpha \left\|H_{[s]}\right\|_1\right) H_s > \left\|H_{[s]}\right\|_1 \|H\|_\infty - \alpha \left\|H_{[s]}\right\|_2^2 \geq \left(s \|H\|_\infty - \alpha \left\|H_{[s]}\right\|_1\right) H_{s+1}.$$

Rearranging, this is equivalent to

$$\alpha \left(\left\|H_{[s]}\right\|_2^2 - \left\|H_{[s]}\right\|_1 H_s\right) > \left(\left\|H_{[s]}\right\|_1 - sH_s\right) \|H\|_\infty$$
$$\text{and} \quad \alpha \left(\left\|H_{[s]}\right\|_2^2 - \left\|H_{[s]}\right\|_1 H_{s+1}\right) \leq \left(\left\|H_{[s]}\right\|_1 - sH_{s+1}\right) \|H\|_\infty.$$

One can check that $\left\|H_{[s]}\right\|_2^2 - \left\|H_{[s]}\right\|_1 H_{s+1} > \left\|H_{[s]}\right\|_2^2 - \left\|H_{[s]}\right\|_1 H_s > 0$. So the above is equivalent to

$$\alpha_s := \frac{\left(\left\|H_{[s]}\right\|_1 - sH_s\right) \|H\|_\infty}{\left\|H_{[s]}\right\|_2^2 - \left\|H_{[s]}\right\|_1 H_s} < \alpha \leq \frac{\left(\left\|H_{[s]}\right\|_1 - sH_{s+1}\right) \|H\|_\infty}{\left\|H_{[s]}\right\|_2^2 - \left\|H_{[s]}\right\|_1 H_{s+1}} =: \overline{\alpha}(s).$$

A straightforward calculation shows that $\overline{\alpha}(s) = \alpha_{s+1}$. Thus, using the convention $\alpha_{d+1} = \alpha_{\max}$, s is uniquely characterized by $\alpha_s < \alpha \leq \alpha_{s+1}$. This concludes the proof of Lemma 3.

B.5 Proof of Proposition 4: Concentration of norms of $\gamma(\alpha_s)$

In this section we prove Proposition 4 and Lemma 4 which establish concentration inequalities for $\gamma(\alpha)$ at the breakpoints α_s for $2 \leq s \leq d$. More precisely, we give high-probability estimates (with respect to the draws of h) of their ℓ_1 and ℓ_2 norms, since those are the quantities that appear in the stochastic optimization problems (A'_N), (A'_+) and (A'_-).

Plugging in $\alpha = \alpha_s$ into the closed-form expressions of λ and μ in Lemma 3, we obtain $\gamma(\alpha_s) = \frac{\|H\|_\infty}{\langle v_s, H\rangle} v_s$ where $v_s := H_{[s]} - H_s \mathbf{1}_{[s]}$. Thus, to estimate the norms of $\gamma(\alpha_s)$ it suffices to estimate the quantities

$$\|v_s\|_2^2 = \left\|H_{[s]}\right\|_2^2 - 2\left\|H_{[s]}\right\|_1 H_s + sH_s^2$$
$$\|v_s\|_1 = \left\|H_{[s]}\right\|_1 - sH_s$$
$$\langle v_s, H\rangle = \left\|H_{[s]}\right\|_2^2 - \left\|H_{[s]}\right\|_1 H_s.$$

Throughout the proofs in this section, we will use $c > 0$ to denote a universal constant (in particular, independent of d and s) which may change from display to display. Furthermore, in this section we let Z denote a standard normal distributed random variable, and recall that $\Phi^{\mathsf{C}}(x) = \mathbb{P}(Z > x) = \frac{1}{2}\mathbb{P}(|Z| > x)$ for $x > 0$ denotes its complementary cumulative distribution function.

B.5.1 Preliminary facts

We start by stating some auxiliary facts about Φ^{C}.

Fact 1. *Denote $h(x)$ the function such that $\forall x > 0$, $\Phi^{\mathsf{C}}(x) = \frac{\exp(-x^2/2)}{x\sqrt{2\pi}} h(x)$. We have the first-order and higher-order upper and lower bounds*

$$1 - \frac{1}{1+x^2} \leq h(x) \leq 1 \quad \text{and} \quad \left| h(x) - \left(1 - \frac{1}{x^2} + \frac{3}{x^4} - \frac{15}{x^6}\right)\right| \leq \frac{c}{x^8}$$

for all $x > 0$.

Proof. The first-order estimate follows from straightforward analysis. The higher-order estimate follows from the exact asymptotic expansion of the *complementary error function* erfc,since $2\Phi^{\complement}(x) = \text{erfc}(x/\sqrt{2})$. □

Fact 2. *By straightforward calculations, we have*

$$\forall x > 0,\ \mathbb{E}[Z|Z \geq x] = \frac{1}{\Phi^{\complement}(x)} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} = \frac{x}{h(x)}$$
$$\text{and} \quad \mathbb{E}[Z^2|Z \geq x] = \frac{1}{\Phi^{\complement}(x)} \left(\frac{x}{\sqrt{2\pi}} \exp(-x^2/2) + \Phi^{\complement}(x) \right) = 1 + \frac{x^2}{h(x)}.$$

We will also make repeated use of Lemma 4 (Section A.2.3), whose proof is deferred to Section B.5.4.

B.5.2 Proof of Proposition 4

We will show the following lemmas successively, in which $t \in \mathbb{R}$ denotes the quantity such that $2\Phi^{\complement}(t) = s/d$ (we drop the explicit dependency on s for concision in this section).

Lemma 11 (Concentration of H_s)**.** *Assume that $s < d/2$. With probability at least $1 - 2\delta$, we have*

$$|H_s - t| \leq c \left(\frac{1}{\sqrt{s}} + \sqrt{\frac{\log(1/\delta)}{s}} + \frac{\log(1/\delta)}{s} \right).$$

Lemma 12 (Concentration of $\left\|H_{[s]}\right\|_2^2$)**.** *Assume $s < d/5$. With probability at least $1 - 2\delta$, we have*

$$\left| \left\|H_{[s]}\right\|_2^2 - s\mathbb{E}[Z^2|Z \geq t] \right| \leq c\sqrt{s}(1 + \sqrt{\log(1/\delta)}) \left(\frac{1}{\sqrt{s}}(1 + \sqrt{\log(1/\delta)}) + t \right).$$

Lemma 13 (Concentration of $\left\|H_{[s]}\right\|_1$)**.** *With probability at least $1 - 2\delta$, we have*

$$\left| \left\|H_{[s]}\right\|_1 - s\mathbb{E}\left[Z|Z \geq t\right] \right| \leq c \left(\sqrt{s} + \sqrt{\log(1/\delta)s} \right).$$

Lemma 14 (Concentration of v_s)**.** *Assume $s < d/5$. For $\delta \geq \exp(-s)$, with probability at least $1 - 6\delta$, we have*

$$\left| \|v_s\|_2^2 - s\left(\frac{2}{t^2} - \frac{10}{t^4} \right) \right| \leq s\left(\frac{c}{t^6} + C_{s,\delta} \right) \text{ and}$$
$$\left| \|v_s\|_1 - s\left(\frac{1}{t} - \frac{2}{t^3} \right) \right| \leq s\left(\frac{c}{t^5} + \frac{C_{s,\delta}}{t} \right) \text{ and}$$
$$|\langle v_s, H\rangle - s| \leq sC_{s,\delta}$$

with $C_{s,\delta} = c\frac{t + t\sqrt{\log(1/\delta)}}{\sqrt{s}}$.

The proposition follows as a consequence of this last lemma:

Proof of Proposition 4. Let $c_1, c_2, \bar{t}$ as in Lemma 4, and assume $s \leq d/c_1$. In particular, $\log(d/s) \leq t^2 \leq 2\log(d/s)$.

We apply Lemma 14 with $\delta = \exp\left(-2\frac{s}{\log(d/s)^5}\right)$. Since $t^2 \leq 2\log(d/s)$, this choice ensures that $\frac{t\sqrt{\log(1/\delta)}}{\sqrt{s}} \leq 8/t^4$. Moreover, the assumption that $d \leq \exp\left(\kappa_4 s^{1/5}\right)$ ensures that $\frac{t}{\sqrt{s}} \leq c/t^4$. So we have $C_{s,\delta} \leq c/t^4$.

The proposition follows by substituting the estimates of $\|v_s\|_2^2$, $\|v_s\|_1$, $\langle v_s, H\rangle$ into $\gamma(\alpha_s) = \frac{\|H\|_\infty}{\langle v_s, H\rangle} v_s$, and making the appropriate simplifications. □

B.5.3 Proofs of the concentration lemmas

Proof of Lemma 11: Concentration of H_s. By observing that the random variable $\max\{s; H_s > t\}$ is binomially distributed with parameters d and $p = \mathbb{P}(|Z| > t)$, Li et al. (2020) show the following upper and lower tail bounds for H_s.

Claim 4. *Assume that $s < d/2$ and let t be such that $2\Phi^{\complement}(t) = s/d$. Then for all $\varepsilon > 0$, we have the lower resp. upper tail bounds*

$$\begin{aligned} &\mathbb{P}(H_s \leq t - \varepsilon) \leq \exp\left(-cs\varepsilon^2 \log(d/s)\right) \\ \text{and} \quad &\mathbb{P}(H_s \geq t + \varepsilon) \leq \exp\left(-cs\varepsilon^2 \log(d/s) \exp\left(-2\varepsilon\sqrt{2\log(d/s) - \log\log(d/s) - \log(\frac{\pi}{2}) - \varepsilon^2}\right)\right). \end{aligned}$$

Proof. This follows straightforwardly from Lemma 2 of Li et al. (2020) and from the estimate of t in Lemma 4. ☐

The lower tail bound is already sufficiently tight to show our high-probability lower bound on $H_s - t$. However we remark that the upper tail bound is too loose; indeed it is only reasonable when ε is sufficiently small. So to prove our high-probability upper bound, we instead start from the following one-sided concentration inequality from Boucheron and Thomas (2012).

Claim 5. *Assume $d \geq 3$ and $s < d/2$, then for all $z > 0$,*

$$\mathbb{P}\left(H_s - \mathbb{E}H_s \geq c(\sqrt{z/s} + z/s)\right) \leq \exp(-z).$$

Proof. The proof follows from the same argument as in Proposition 4.6 of Boucheron and Thomas (2012). ☐

It only remains to bound the distance between t and $\mathbb{E}H_s$.

Claim 6. *Assume that $s < d/2$ and let t be such that $2\Phi^{\complement}(t) = s/d$. Then*

$$|\mathbb{E}H_s - t| \leq c\frac{1}{\sqrt{s}}.$$

Proof. According to Proposition 4.2 of Boucheron and Thomas (2012),

$$\mathrm{Var}(H_s) \leq \frac{1}{s\log 2}\frac{8}{\log\frac{2d}{s} - \log(1 + \frac{4}{s}\log\log\frac{2d}{s})}$$

so by Chebyshev's inequality,

$$\mathbb{P}\left(|H_s - \mathbb{E}H_s| > \varepsilon'\right) \leq \frac{c'}{s}\frac{1}{(\varepsilon')^2}.$$

On the other hand, recall from Claim 4 that

$$\mathbb{P}\left(|H_s - t| > \varepsilon\right) \leq 2\exp\left(-cs\varepsilon^2 \log(d/s) \exp\left(-2\varepsilon\sqrt{2\log(d/s) - \log\log(d/s) - \log(\frac{\pi}{2}) - \varepsilon^2}\right)\right)$$

One can check that there exist universal constants c_1, c_2 such that, by picking $\varepsilon = c_1/\sqrt{s\log(d/s)}$ and $\varepsilon' = c_2/\sqrt{s}$, the sum of the right-hand sides is less than 1.

Thus, with positive probability we have

$$|\mathbb{E}H_s - t| \leq |H_s - \mathbb{E}H_s| + |H_s - t| \leq \frac{c_1/\sqrt{\log(2)} + c_2}{\sqrt{s}}.$$

☐

Proof of Lemma 12: Concentration of $\left\|H_{[s]}\right\|_2^2$. Let us first restate Proposition 2 of Li et al. (2020) in our notation. We remark that their statement contained an additional $\log(d/s)$ factor due to a mistake in the proof. Correcting this mistake, we have that with probability at least $1-2\delta$,

$$\left|\frac{1}{\sqrt{s}}\left\|H_{[s]}\right\|_2 - \sqrt{\mathbb{E}[Z^2|Z\geq t]}\right| \leq c\frac{1}{\sqrt{s}}(1+\sqrt{\log(1/\delta)}).$$

Since for all $a,b,\varepsilon>0$, $|a-b|\leq\varepsilon \implies \left|a^2-b^2\right|\leq\varepsilon(\varepsilon+2b)$, this implies

$$\left|\frac{1}{s}\left\|H_{[s]}\right\|_2^2 - \mathbb{E}[Z^2|Z\geq t]\right| \leq c\frac{1}{\sqrt{s}}(1+\sqrt{\log(1/\delta)})\left(\frac{1}{\sqrt{s}}(1+\sqrt{\log(1/\delta)})+\sqrt{\mathbb{E}[Z^2|Z\geq t]}\right).$$

Now $\mathbb{E}[Z^2|Z\geq t] = 1+\frac{t^2}{h(t)} \leq ct^2$ whenever $t\geq 1$, which is ensured by our assumption that $s/d=\Phi^{\complement}(t)\leq 0.2$. So

$$\left|\frac{1}{s}\left\|H_{[s]}\right\|_2^2 - \mathbb{E}[Z^2|Z\geq t]\right| \leq c\frac{1}{\sqrt{s}}(1+\sqrt{\log(1/\delta)})\left(\frac{1}{\sqrt{s}}(1+\sqrt{\log(1/\delta)})+t\right).$$

Proof of Lemma 13: Concentration of $\left\|H_{[s]}\right\|_1$. We use exactly the same argument as in the proof of Proposition 2 of Li et al. (2020). Namely, start by decomposing

$$\left|\frac{1}{s}\left\|H_{[s]}\right\|_1 - \mathbb{E}[Z|Z\geq t]\right| \leq \left|\frac{1}{s}\left\|H_{[s]}\right\|_1 - \frac{1}{s}\mathbb{E}\left\|H_{[s]}\right\|_1\right| + \left|\frac{1}{s}\mathbb{E}\left\|H_{[s]}\right\|_1 - \mathbb{E}[Z|Z\geq t]\right|.$$

For the first term, note that by rearrangement inequality, $Z\mapsto\frac{1}{s}\sum_{i=1}^s\left|Z_{(i)}\right|$ is $\frac{1}{\sqrt{s}}$-Lipschitz for the $\|\cdot\|_2$ norm, where $(Z_{(1)},...,Z_{(d)})$ is the nondecreasing reordering of the absolute values of Z.[3] So by concentration of Lipschitz-continuous functions of Gaussians,

$$\mathbb{P}\left(\left|\frac{1}{s}\sum_{i=1}^s H_i - \mathbb{E}\frac{1}{s}\sum_{i=1}^s H_i\right|\geq\varepsilon\right)\leq 2\exp(-s\varepsilon^2/2).$$

For the second term, we can apply exactly the same arguments as in the proof of Proposition 2 of Li et al. (2020), adapting equations (42) to (46), to obtain the bound

$$\left|\mathbb{E}\frac{1}{s}\sum_{i=1}^s H_i - \mathbb{E}[Z|Z\geq t]\right| \leq c\mathbb{E}\left|H_{s+1}-t\right| \leq c\frac{1}{\sqrt{s}}.$$

In particular, we use the fact that $x\mapsto\mathbb{E}[Z|Z\geq x]$ is a smooth function, which follows from its explicit expression given in Fact 2.

Proof of Lemma 14: Concentration of v_s. For brevity of notation, let $C_{s,\delta}=c\frac{t+t\sqrt{\log(1/\delta)}}{\sqrt{s}}$. Assume $\delta\geq e^{-s}$; in particular, $C_{s,\delta}\leq ct$. Collecting and simplifying the above results, so far we showed that

$$t\left|H_s-t\right|\leq C_{s,\delta} \text{ and}$$
$$\left|\frac{1}{s}\left\|H_{[s]}\right\|_2^2-(1+\frac{t^2}{h(t)})\right|\leq C_{s,\delta} \text{ and}$$
$$t\left|\frac{1}{s}\left\|H_{[s]}\right\|_1-\frac{t}{h(t)}\right|\leq C_{s,\delta}.$$

[3]Proof: Denote $\widetilde{Z}$ the reordering of Z such that $\left|\widetilde{Z}_1\right|\geq...\geq\left|\widetilde{Z}_d\right|$ (but still with $\widetilde{Z}$ signed). Then

$$\left|\sum_{i=1}^s\left|Z_{(i)}\right|-\left|Y_{(i)}\right|\right| = \left|\left\|\widetilde{Z}_{[s]}\right\|_1-\left\|\widetilde{Y}_{[s]}\right\|_1\right| \leq \left\|\widetilde{Z}_{[s]}-\widetilde{Y}_{[s]}\right\|_1 \leq \sqrt{s}\left\|\widetilde{Z}_{[s]}-\widetilde{Y}_{[s]}\right\|_2 \leq \sqrt{s}\left\|Z-Y\right\|_2$$

where the last inequality follows from the rearrangement inequality after taking squares and expanding.

- Substituting the deterministic estimates in the expression of $\|v_s\|_2^2$ and carrying over the above concentration bounds, we obtain
$$\left| \|v_s\|_2^2 - \left(s(1+\frac{t^2}{h(t)}) - 2\frac{st}{h(t)}t + st^2 \right) \right| \leq sC_{s,\delta}(1+C_{s,\delta})$$
and the deterministic estimate can be simplified to
$$s\frac{(1+t^2)h(t)-t^2}{h(t)} = s\frac{\frac{2}{t^2} - \frac{12}{t^4} + O\left(\frac{1}{t^6}\right)}{1-\frac{1}{t^2}+O\left(\frac{1}{t^4}\right)} = s\left(\frac{2}{t^2} - \frac{10}{t^4} + O\left(\frac{1}{t^6}\right)\right)$$
(where the $O(\cdot)$ hides a universal constant).

- Likewise for $\|v_s\|_1$ we get
$$\left| \|v_s\|_1 - \left(\frac{st}{h(t)} - st \right) \right| \leq \frac{s}{t}C_{s,\delta}$$
and the deterministic estimate can be simplified to
$$st\left(\frac{1}{h(t)} - 1\right) = s\left(\frac{1}{t^2} - \frac{2}{t^3} + O\left(\frac{1}{t^5}\right)\right).$$

- Likewise for $\langle v_s, H\rangle$ we get
$$\left| \langle v_s, H\rangle - \left(s\left(1+\frac{t^2}{h(t)}\right) - \frac{st}{h(t)}t \right) \right| \leq 3sC_{s,\delta}$$
and the deterministic estimate simplifies to s.

B.5.4 Proof of Lemma 4

Using the upper bound $\Phi^{\mathsf{C}}(x) \leq \frac{\exp(-x^2/2)}{x\sqrt{2\pi}}$ from the first part of Fact 1, it is straightforward to check that $2\Phi^{\mathsf{C}}(\bar{t}) \leq s/d = 2\Phi^{\mathsf{C}}(t)$, and so $\bar{t} \geq t$.

Let $\underline{t}^2 = \bar{t}^2 - c_1$ for some constant $c_1 > 0$ to be chosen. Using the lower bound $\Phi^{\mathsf{C}}(x) \geq \frac{\exp(-x^2/2)}{x\sqrt{2\pi}}\frac{x^2}{1+x^2}$, one can check that c_1 can be chosen such that $2\Phi^{\mathsf{C}}(\underline{t}) \geq s/d = 2\Phi^{\mathsf{C}}(t)$, and so $\underline{t} \leq t$.

Going through the calculations reveals that $\kappa_3 \geq e^{2/\pi} \approx 2$ ensures $\bar{t}^2 \leq 2\log(d/s)$, that $2\Phi^{\mathsf{C}}(\bar{t}) \leq s/d$ is always true, that $c_1 = 1 - \log(\frac{\pi}{2}) \approx 0.5$ ensures $\underline{t}^2 \geq \log(d/s)$ for all s, d, and that $\kappa_3 \geq e^{2.3415...} \approx 10.4$ ensures $2\Phi^{\mathsf{C}}(\underline{t}) \geq s/d$. This concludes the proof of the lemma.

Remark 2. *Tighter bounds for t can be derived while still only using the first-order estimate of $h(x)$ (the first part of Fact 1). Namely, by similar straightforward calculations as above, one can check that there exist universal constants $\kappa, \alpha_1, \alpha_2 > 0$ such that, for all $s \leq d/\kappa$, t is bounded as $\underline{t} \leq t \leq \bar{t}$ where*
$$\underline{t}^2 = 2\log(d/s) - \log\log(d/s) - \log(\pi) + \frac{\log\log(d/s)}{2\log(d/s)} - \frac{\alpha_1}{\log(d/s)}$$
$$\textit{and}\quad \bar{t}^2 = 2\log(d/s) - \log\log(d/s) - \log(\pi) + \frac{\log\log(d/s)}{2\log(d/s)} + \frac{\alpha_2}{\log(d/s)}.$$

B.6 Proofs of Lemmas 5 and 6: Studying the feasible set of (A'_+) and (A'_-)

B.6.1 Proof of Lemma 5

We give separate proofs for the statements 1-3 in the Lemma:

First statement: The mapping $\alpha \mapsto \|\gamma(\alpha)\|_2^2$ is decreasing over $[1, \alpha_{d+1/2}]$ and increasing over $[\alpha_{d+1/2}, \alpha_{\max}]$.

Using the notation of Section B.4, the optimization problem defining $\gamma(\alpha)$ has Lagrangian
$$L(w;\lambda,\mu,\nu) = \frac{1}{2}\|w\|_2^2 - \lambda(\langle w, H\rangle - \|H\|_\infty) + \mu(\mathbf{1}^\top w - \alpha) - \nu^\top w,$$

(up to the constant factor $\frac{1}{2}$ in the first term). By the envelope theorem, the marginal effect on the optimal value of increasing α, is equal to the associated Lagrangian multiplier at optimum: $\frac{d\|\gamma(\alpha)\|_2^2}{d\alpha} = -\mu$. A straightforward computation using the expression of μ from Lemma 3 reveals that $\mu > 0$ for $1 < \alpha < \alpha_{d+1/2}$ and $\mu < 0$ for $\alpha_{d+1/2} < \alpha \leq \alpha_{\max}$; hence the monotonicity of $\alpha \mapsto \|\gamma(\alpha)\|_2^2$.

Second statement: The mapping $\alpha \mapsto \|\gamma(\alpha)\|_2^2$ is convex over $[1, \alpha_{\max}]$, and I is an interval.

Recall that $\|\gamma(\alpha)\|_2^2$ is given by

$$\|\gamma(\alpha)\|_2^2 = \min_w \|w\|_2^2 \quad \text{s.t} \quad \begin{cases} \langle w, H\rangle \geq \|H\|_\infty \\ \forall i, w_i \geq 0 \\ \mathbf{1}^\top w = \|w\|_1 = \alpha \end{cases}.$$

Since α appears on the right-hand side of a linear constraint, it is straightforward to check directly that $\alpha \mapsto \|\gamma(\alpha)\|_2^2$ is convex. In detail: Let any α_0 and α_1, let $w_i = \gamma(\alpha_i)$ for $i \in \{1, 2\}$, and $\alpha_t = (1-t)\alpha_0 + t\alpha_1$ and $w_t = (1-t)w_0 + tw_1$ for $t \in [0, 1]$; then w_t is feasible for the optimization problem defining $\gamma(\alpha_t)$, so $\|\gamma(\alpha_t)\|_2^2 \leq \|w_t\|_2^2 \leq t\|w_0\|_2^2 + (1-t)\|w_1\|_2^2$ by convexity.

I is the $(B^2\|H\|_\infty^2)$-sublevel set of the function $\alpha \mapsto (1-\rho)n\left(\sigma^2\alpha^2 + B^2\|\gamma(\alpha)\|_2^2\right)$, which is convex, so I is an interval.

Third statement: The mapping $\alpha \mapsto \frac{\|\gamma(\alpha)\|_2^2}{\alpha^2}$ is monotonically decreasing.

Note that for each $\alpha \in [1, \alpha_{\max}]$, $\frac{\gamma(\alpha)}{\alpha}$ is the optimal solution of the optimization problem

$$\frac{\gamma(\alpha)}{\alpha} = \arg\min_w \|w\|_2^2 \quad \text{s.t} \quad \begin{cases} \langle w, H\rangle \geq \frac{\|H\|_\infty}{\alpha} \\ \forall i, w_i \geq 0 \\ \mathbf{1}^\top w = \|w\|_1 = 1 \end{cases}.$$

In particular, the constraint set is only increasing with α, implying that $\alpha \mapsto \left\|\frac{\gamma(\alpha)}{\alpha}\right\|_2^2$ is monotonically decreasing with α.

B.6.2 Proof of Lemma 6

The upper bound for ϕ_+ immediately follows from Equation (A'_+) and from the monotonicity of $\alpha \mapsto \left\|\frac{\gamma(\alpha)}{\alpha}\right\|_2^2$, which is the last statement of Lemma 5.

For ϕ_-, there is an extra scale variable $0 < b \leq B$ which we first minimize out, similarly to the proof of Proposition 2 in Section A.2.4. Starting from Equation (A'_-), first rewrite the constraint as

$$\begin{aligned} b^2\|H\|_\infty^2 &\geq (1-\rho)n\left(\sigma^2\alpha^2 + b^2\|\gamma(\alpha)\|_2^2\right) \\ \iff b^2 &\geq \underbrace{\frac{(1-\rho)n\sigma^2\alpha^2}{\|H\|_\infty^2 - (1-\rho)n\|\gamma(\alpha)\|_2^2}}_{=:\tilde{f}(\alpha)^2} \quad \text{and} \quad (1-\rho)n\|\gamma(\alpha)\|_2^2 < \|H\|_\infty^2. \end{aligned}$$

Then,

$$\begin{aligned}
\phi_- &= \min_{1\leq\alpha\leq\alpha_{\max}} \min_b \; b^2 \frac{\|\gamma(\alpha)\|_2^2}{\alpha^2} \quad \text{s.t} \quad \tilde{f}(\alpha) \leq b \leq B \text{ and } (1-\rho)n\|\gamma(\alpha)\|_2^2 < \|H\|_\infty^2 \\
&= \min_{1\leq\alpha\leq\alpha_{\max}} \tilde{f}(\alpha)^2 \frac{\|\gamma(\alpha)\|_2^2}{\alpha^2} \quad \text{s.t} \quad \tilde{f}(\alpha) \leq B \text{ and } (1-\rho)n\|\gamma(\alpha)\|_2^2 < \|H\|_\infty^2 \\
&= \min_{1\leq\alpha\leq\alpha_{\max}} \tilde{f}(\alpha)^2 \frac{\|\gamma(\alpha)\|_2^2}{\alpha^2} \quad \text{s.t} \quad \alpha \in I \\
&= \min_\alpha \frac{(1-\rho)n\sigma^2\|\gamma(\alpha)\|_2^2}{\|H\|_\infty^2 - (1-\rho)n\|\gamma(\alpha)\|_2^2} \quad \text{s.t} \quad \underline{\alpha}_I \leq \alpha \leq \overline{\alpha}_I.
\end{aligned}$$

This objective is increasing in $\|\gamma(\alpha)\|_2^2$. Now as shown in Lemma 5, $\alpha \mapsto \|\gamma(\alpha)\|_2^2$ is decreasing over $[1, \alpha_{d+1/2}]$. Therefore, this objective is decreasing in α over $[\underline{\alpha}_I, \alpha_{d+1/2}]$, and ϕ_- is lower-bounded by its value at any $\overline{\alpha}$ such that $\overline{\alpha}_I \leq \overline{\alpha} \leq \alpha_{d+1/2}$.

B.7 Proof of Lemma 7: A tight admissible choice for $\underline{\alpha}$ and $\overline{\alpha}$

Recall that we defined $B(n,d) = c\sigma\sqrt{\|w^*\|_0} + M(n,d)$ where $M(n,d)$ is given by Proposition 2 and $c>0$ some universal constant. For brevity of notation, abbreviate $B(n,d) = B$. We seek integers $\underline{s}, \overline{s}$ such that $\alpha_{\underline{s}}$ and $\alpha_{\overline{s}}$ lie on the left, respectively on the right of the interval I, and such that $\alpha_{\overline{s}} \leq \alpha_{d+1/2}$, with high probability over the draws of h.

B.7.1 Preliminaries

We recall the notations $B = M(n,d) + 2\|w^*\|_1$ and

$$I = \left\{\alpha \in [1,\alpha_{\max}] \middle| B^2\|H\|_\infty^2 \geq (1-\rho)n\left(\sigma^2\alpha^2 + B^2\|\gamma(\alpha)\|_2^2\right)\right\}$$
$$\text{and} \qquad \forall \alpha \in [1,\alpha_{\max}],\; \tilde{f}(\alpha)^2 = \frac{(1-\rho)n\sigma^2\alpha^2}{\|H\|_\infty^2 - (1-\rho)n\|\gamma(\alpha)\|_2^2}.$$

Note that we have the equivalence

$$\alpha \in I \iff (1-\rho)n\|\gamma(\alpha)\|_2^2 < \|H\|_\infty^2 \quad \text{and} \quad \tilde{f}(\alpha) \leq B. \tag{12}$$

Reference point: $\alpha_n \in I$. We have by construction that, conditionally on the event where the inequalities in Equation 6 hold for $s = n$, $\alpha_n \in I$ – in particular this holds with probability at least $1 - 6\exp\left(-2\frac{n}{\log(d/n)^5}\right)$. Indeed, let us take a closer look at the way we chose $M(n,d)$, from the proof of Proposition 2 (Section A.2.4). We showed that, conditionally on that event, α_n satisfies

$$(1+\rho)n\|\gamma(\alpha_n)\|_2^2 < \|H\|_\infty^2 \quad \text{and} \quad f(\alpha_n) \leq M(n,d)$$

where $f(\alpha)^2 = \frac{(1+\rho)n\sigma^2\alpha^2}{\|H\|_\infty^2 - (1+\rho)n\|\gamma(\alpha)\|_2^2}$. Since $\tilde{f}(\alpha_n) \leq f(\alpha_n)$ and $M(n,d) \leq B$, clearly α_n satisfies the condition (12), i.e., $\alpha_n \in I$.

Summary of (in)equalities to be used in the proof. For ease of presentation, let us recall some assumptions or definitions that we will use throughout this proof. For each integer s, $t_s \in \mathbb{R}$ denotes the quantity such that $2\Phi^{\complement}(t_s) = s/d$, and $t_s^2 \asymp \log(d/s)$ by Lemma 4. By assumption, $\|w^*\|_0 \leq \kappa_1 \frac{\sigma^2 n}{\log(d/n)^5}$ for some universal constant $\kappa_1 > 0$. By definition, $M(n,d)^2 = \frac{\sigma^2 n}{t_n^2}\left(1 - \frac{2}{t_n^2} + \frac{c_0}{t_n^4}\right)$ for some universal constant $c_0 > 0$. In particular, this implies that

$$B^2 = \left(M(n,d) + c\sigma\sqrt{\|w^*\|_0}\right)^2 = \frac{\sigma^2 n}{t_n^2}\left(1 - \frac{2}{t_n^2} + O\left(\frac{1}{t_n^4}\right)\right). \tag{13}$$

B.7.2 Finding $\underline{s}$ such that $\alpha_{\underline{s}} \leq \underline{\alpha}_I$

We want to find an $\underline{s}$ such that $\alpha_{\underline{s}}$ is on the left of the interval I, i.e., $\alpha_{\underline{s}} \leq \underline{\alpha}_I$. Conditioning on the event where $\alpha_n \in I$, it suffices to have $\alpha_{\underline{s}} < \alpha_n$ i.e. $\underline{s} < n$, and $\alpha_{\underline{s}} \notin I$ i.e.

$$\begin{aligned} &\frac{1}{B^2\|H\|_\infty^2}(1-\rho)n\left(\sigma^2 \left\|\gamma(\alpha_{\underline{s}})\right\|_1^2 + B^2 \left\|\gamma(\alpha_{\underline{s}})\right\|_2^2\right) \\ &= (1-\rho)n\frac{\sigma^2}{B^2}\frac{\left\|\gamma(\alpha_{\underline{s}})\right\|_1^2}{\|H\|_\infty^2} + (1-\rho)n\frac{\left\|\gamma(\alpha_{\underline{s}})\right\|_2^2}{\|H\|_\infty^2} > 1. \end{aligned}$$

Instead of working directly with $\underline{s}$, it is more convenient to define $\underline{s}$ implicitly through a condition on $t_{\underline{s}}$. Namely, we choose $\underline{s}$ such that $t_{\underline{s}}^2 \approx t_n^2 + \frac{\lambda}{t_n}$ for some constant $\lambda > 0$. We now make this choice formal and show that $\underline{s}$ is very close to n; the following step will be to show that this choice guarantees $\alpha_s \notin I$.

Claim 7. *Assume $\kappa_3 n \leq d$. Let any fixed constant $0 < \lambda \leq \sqrt{\log(\kappa_3)}$, and let $\underline{s}$ be the largest integer s such that $t_s^2 \geq t_n^2 + \frac{\lambda}{t_n}$. Then,*

$$\begin{aligned} &\underline{s} = n\exp\left(-\frac{\lambda}{2t_n}\right)\left(1 + O\left(\frac{1}{t_n^2}\right)\right) \\ \textit{and} \qquad &\left|t_{\underline{s}}^2 - \left(t_n^2 + \frac{\lambda}{t_n}\right)\right| \leq O\left(\frac{1}{n}\right). \end{aligned}$$

The first equation quantifies the fact that this choice of s is very close to n; the second equation controls the error due to rounding (due to the fact that there is no integer s such that $t_s^2 = t_n^2 + \frac{\lambda}{t_n}$ exactly).

Proof. For concision, in this proof we write s instead of $\underline{s}$. Denote $\bar{t}_n^2 = t_n^2 + \frac{\lambda}{t_n}$. By definition, $t_s^2 \geq \bar{t}_n^2 > t_{s+1}^2$.

For the first part of the claim, we apply Fact 1 several times. Firstly,

$$2\Phi^{\mathsf{C}}(t_s) = \frac{s}{d} \leq \frac{2}{\sqrt{2\pi}}\frac{1}{t_s}e^{-t_s^2/2} \leq \frac{2}{\sqrt{2\pi}}\frac{1}{t_n}e^{-\bar{t}_n^2/2} = \frac{2}{\sqrt{2\pi}}\frac{1}{t_n}e^{-t_n^2/2}\exp\left(-\frac{\lambda}{2t_n}\right).$$

Secondly,

$$\begin{aligned} &\frac{t_n^2}{1+t_n^2}\cdot\frac{2}{\sqrt{2\pi}}\frac{1}{t_n}e^{-t_n^2/2} \leq 2\Phi^{\mathsf{C}}(t_n) \\ \text{and hence} \qquad &\frac{2}{\sqrt{2\pi}}\frac{1}{t_n}e^{-t_n^2/2} \leq \frac{n}{d}\left(1 + \frac{1}{t_n^2}\right). \end{aligned}$$

This proves the upper bound on s. The lower bound can be proved in a similar fashion, by applying Fact 1 to lower-bound $\frac{s+1}{d} = 2\Phi^{\mathsf{C}}(t_{s+1})$, and again to upper-bound $\frac{1}{t_n^2}e^{-t_n^2/2}$. In this way we obtain

$$\frac{s+1}{d} \geq \frac{n}{d}\exp\left(-\frac{\lambda}{2t_n}\right)\left(1 - \frac{1}{1+t_{s+1}^2}\right),$$

and the bound $t_s^2 - t_{s+1}^2 \leq \frac{2}{s} \leq 2$ shown below implies that $\frac{1}{1+t_{s+1}^2} = O\left(\frac{1}{t_n^2}\right)$.

We now turn to the second part of the claim. By mean value theorem applied on Φ^{C}, there exists $\xi \in [t_{s+1}, t_s]$ such that

$$\begin{aligned} &\frac{\Phi^{\mathsf{C}}(t_{s+1}) - \Phi^{\mathsf{C}}(t_s)}{t_{s+1} - t_s} = \frac{1}{2d}\frac{1}{t_{s+1} - t_s} = (\Phi^{\mathsf{C}})'(\xi) = -\frac{1}{\sqrt{2\pi}}e^{-\xi^2/2} \\ \text{and} \qquad &0 < t_s - t_{s+1} = \frac{\sqrt{2\pi}}{2d}e^{\xi^2/2} \leq \frac{\sqrt{2\pi}}{2d}e^{t_s^2/2}. \end{aligned}$$

Now, by Fact 1, this can be further upper-bounded as

$$\frac{\sqrt{2\pi}}{2d}e^{t_s^2/2} \leq \frac{1}{2d}\frac{1}{\Phi^{\mathsf{C}}(t_s)}\frac{1}{t_s} = \frac{1}{st_s}.$$

So we have the bound:

$$t_s^2 - \bar{t}_n^2 \leq t_s^2 - t_{s+1}^2 \leq \frac{t_s + t_{s+1}}{s t_s} \leq \frac{2}{s}.$$

(This completes the proof of the first part of the claim, for the lower bound.) We can conclude by substituting s by its estimate from the first part of the claim, noting that $\frac{\lambda}{t_n}$ is uniformly bounded by assumption since $\lambda \leq \sqrt{\log(\kappa_3)} \leq \sqrt{\log(d/n)} \leq t_n$ by Lemma 4 (for an appropriate choice of κ_3). ☐

We now show that we can choose the constant $\lambda > 0$ such that $\alpha_{\underline{s}} \notin I$ with high probability.

Claim 8. *The constants $\kappa_2, \kappa_3, \kappa_4, \lambda > 0$ can be chosen such that for any n, d with $n \geq \kappa_2$ and $\kappa_3 n \leq d \leq \exp(\kappa_4 n)$,*

$$(1-\rho)n \frac{\sigma^2}{B^2} \frac{\left\|\gamma(\alpha_{\underline{s}})\right\|_1^2}{\|H\|_\infty^2} + (1-\rho)n \frac{\left\|\gamma(\alpha_{\underline{s}})\right\|_2^2}{\|H\|_\infty^2} > 1$$

with probability at least $1 - 12\exp\left(-\frac{n}{\log(d/n)^5}\right)$ over the draws of h, where $\underline{s}$ is defined as in Claim 7.

Proof. We will repeatedly use the following inequalities, which follow from Lemma 4 and Claim 7:

$$t_n^2 = \log(d/n) + O(\log\log(d/n))$$
$$t_{\underline{s}}^2 = \log(d/n) + O(\log\log(d/n))$$
$$\frac{t_n^2}{t_{\underline{s}}^2} = \frac{1}{1 + \frac{\lambda}{t_n^3} + O\left(\frac{1}{t_n^2 n}\right)} = 1 - \frac{\lambda}{t_n^3} + O\left(\frac{1}{t_n^2 n}\right) + O\left(\frac{\lambda^2}{t_n^6}\right).$$

Note that for appropriate choices of $\kappa_2, \kappa_3, \kappa_4$, Equation (6) in Proposition 4 holds simultaneously for $s = n$ and for $s = \underline{s}$ with probability at least $1 - 6\exp(-2\frac{n}{\log(d/n)^5}) - 6\exp(-2\frac{\underline{s}}{\log(d/\underline{s})^5}) \geq 1 - 12\exp(-\frac{n}{\log(d/n)^5})$. We condition on this event throughout the remainder of the proof.

We begin with the first term, where we use the upper estimate of B^2 from Equation (13):

$$\begin{aligned}
(1-\rho)n \frac{\sigma^2}{B^2} \frac{\left\|\gamma(\alpha_{\underline{s}})\right\|_1^2}{\|H\|_\infty^2} &\geq \frac{t_n^2}{1 - \frac{2}{t_n^2} + O\left(\frac{1}{t_n^4}\right)} \frac{1}{t_{\underline{s}}^2} \left(1 - \frac{4}{t_{\underline{s}}^2} + O\left(\frac{1}{t_{\underline{s}}^4}\right)\right) \\
&\geq \left(1 - \frac{2}{t_n^2} + O\left(\frac{1}{t_n^4}\right)\right) \left(1 - \frac{\lambda}{t_n^3} + O\left(\frac{1}{t_n^2 n}\right) + O\left(\frac{\lambda^2}{t_n^6}\right)\right) \\
&= 1 - \frac{2}{t_n^2} - \frac{\lambda}{t_n^3} + O\left(\frac{1}{t_n^4}\right) + O\left(\frac{\lambda^2}{t_n^6}\right) + O\left(\frac{\lambda}{t_n^2 n}\right). \qquad (14)
\end{aligned}$$

Next the second term:

$$\begin{aligned}
(1-\rho)n \frac{\left\|\gamma(\alpha_{\underline{s}})\right\|_2^2}{\|H\|_\infty^2} &\geq n \frac{2}{\underline{s} t_{\underline{s}}^2} \left(1 + O\left(\frac{1}{t_{\underline{s}}^2}\right)\right) \\
&\geq \frac{2}{t_{\underline{s}}^2} \exp\left(\frac{\lambda}{2t_n}\right) \left(1 + O\left(\frac{1}{t_n^2}\right)\right) \\
&\geq \frac{2}{t_n^2} \left(1 + \frac{\lambda}{2t_n} + \frac{\lambda^2}{4t_n^2} + O\left(\frac{\lambda^3}{t_n^3}\right)\right) \left(1 + O\left(\frac{1}{t_n^2}\right)\right). \qquad (15)
\end{aligned}$$

Summing up (14) and (15), we get:

$$(1-\rho)n \frac{\sigma^2}{B^2} \frac{\left\|\gamma(\alpha_{\underline{s}})\right\|_1^2}{\|H\|_\infty^2} + (1-\rho)n \frac{\left\|\gamma(\alpha_{\underline{s}})\right\|_2^2}{\|H\|_\infty^2} \geq 1 + \frac{\lambda^2}{2t_n^4} + O\left(\frac{1}{t_n^4}\right) + O\left(\frac{\lambda^2}{t_n^6}\right) + O\left(\frac{\lambda}{t_n^2 n}\right).$$

Clearly, we can choose the constants $\lambda, \kappa_2, \kappa_3, \kappa_4 > 0$ such that the right-hand side of the above equation is strictly greater than 1 for any n, d with $\kappa_3 n \leq d$, since $t_n^2 \asymp \log(d/n)$. ☐

B.7.3 Finding $\overline{s}$ such that $\overline{\alpha}_I \leq \alpha_{\overline{s}} \leq \alpha_{d+1/2}$

We take the exact same approach to find $\overline{s} \geq n$ such that $\alpha_{\overline{s}}$ is on the right of the interval I, i.e., $\overline{\alpha}_I \leq \alpha_{\overline{s}}$. The derivations can be straightforwardly adapted, and we get the analogous results:

Claim 9. *Assume $\kappa_3 n \leq d$. Let any fixed constant $0 < \lambda \leq \sqrt{\log(\kappa_3)}$, and let $\overline{s}$ be the smallest integer s such that $t_s^2 \leq t_n^2 - \frac{\lambda}{t_n}$. Then,*

$$\overline{s} = n \exp\left(\frac{\lambda}{2t_n}\right)\left(1 + O\left(\frac{1}{t_n^2}\right)\right) \qquad \text{and} \qquad \left|t_{\overline{s}}^2 - \left(t_n^2 - \frac{\lambda}{t_n}\right)\right| \leq O\left(\frac{1}{n}\right).$$

Claim 10. *The constants $\kappa_2, \kappa_3, \kappa_4, \lambda > 0$ can be chosen such that for any n, d with $n \geq \kappa_2$ and $\kappa_3 n \leq d \leq \exp(\kappa_4 n)$,*

$$(1-\rho)n\frac{\sigma^2}{B^2}\frac{\|\gamma(\alpha_{\overline{s}})\|_1^2}{\|H\|_\infty^2} + (1-\rho)n\frac{\|\gamma(\alpha_{\overline{s}})\|_2^2}{\|H\|_\infty^2} > 1$$

with probability at least $1 - 12\exp\left(-\frac{n}{\log(d/n)^5}\right)$ over the draws of h, where $\overline{s}$ is defined as in Claim 9.

It remains to check that this choice of $\overline{s}$ satisfies $\alpha_{\overline{s}} < \alpha_{d+1/2}$. But this is clearly the case, because $\overline{s} < d$ (by the first part of Claim 9, for appropriate choices of κ_3) and $\alpha_d < \alpha_{d+1/2}$ by definition. This concludes the proof of Lemma 7.