We are very grateful to the reviewers for their helpful feedback and suggestions, and are pleased to have received a generally positive response. Our responses to the main concerns are given as follows. All citations refer to the reference list in the main document.

[Responses to multiple reviewers] **(Experimental evaluations)** We appreciate that experiments are important for many papers, but do not believe it to be the case for this theory paper; please note that prominent related works such as [8,23,24,37] also do not include experiments. At least in the special case of linear measurements, extensive numerical results for (an approximation of) the $\mathcal{K}$-Lasso have already been presented in [2].

(Novelty and Insight) While our paper shares similar high-level insights to earlier works, in particular showing that non-linear observations may be treated as noisy linear observations [23,28], we believe that the *direct* study of generative priors adds significant value to existing approaches based on the Gaussian mean width (GMW) and related notions.

Our analysis builds on works such as [2,15,23,24], but we believe that these techniques are combined and extended in a novel manner, with distinct proofs. For instance: (i) Compared to [23,24], we attain $m^{-\frac{1}{2}}$ scaling instead of $m^{-\frac{1}{4}}$, and make no use of GMW throughout our main analysis; (ii) We require the careful control of several terms in (53), which in turn requires proving Lemmas 7 and 8, along with a more involved chaining argument compared to [2,15]; (iii) We address the uniform recovery case via the LEP, and are the first to do so to our knowledge.

[Responses to R1]: **(Assumptions)** The unit-norm assumption is standard in this line of works, and it (or similar) is indeed essential in the 1-bit model. More generally, the generalized Lasso does not depend on $f(\cdot)$, so if $\mathbf{x}_0$ and $c\mathbf{x}_0$ are both feasible under the prior, one cannot distinguish $f(\langle \mathbf{a}, \mathbf{x}_0 \rangle)$ from $\tilde{f}(\langle \mathbf{a}, c\mathbf{x}_0 \rangle)$, where $\tilde{f}(z) = f(z/c)$. See also Section 4.4 for a related discussion and generalizations to non-unit norms. Indeed, certain models such as phase retrieval do not satisfy our sub-Gaussianity assumption, and we will better highlight this in the revision. This assumption is still much more general than the 1-bit and linear models, and is adopted in many prior works including [9, 24, 27].

(GMW) We will make explicit that the GMW calculation assumes $G(B_2^k(r))$ is contained in the unit ball, and highlight the limitation mentioned for a large radius r. We believe that the unit-ball setting remains of significant interest in itself.

(Practicality in the 1-bit case) In [15] it is assumed that the feasible set lies in the unit sphere, so it is fair to assume the same for comparison. In more detail, [15, Corollary 3] gives a guarantee on any $\hat{\mathbf{x}}$ such that $\mathbf{A}\hat{\mathbf{x}}$ has small Hamming distance to $\tilde{\mathbf{y}}$, but does not specify an optimization problem for finding such $\hat{\mathbf{x}}$. The problem $\min_{\mathbf{x} \in \mathrm{Range}(G)} \mathrm{d_H}(\mathbf{A}\mathbf{x}, \tilde{\mathbf{y}})$ appears to be very hard to solve (e.g., being highly non-differentiable and combinatorial), and the heuristic in [15, Section V] can be viewed as first approximating $\mathrm{d_H}$ by a convex function, and then *further* approximating the minimizer of that function. In contrast, the generalized Lasso solution can be approximated *directly* using gradient methods. However, we acknowledge that both approaches still require some level of approximation, and will accordingly significantly tone down and clarify the claim of practicality.

(Corollary 1) The assumption $\frac{\bar{\mu} G(\mathbf{z}^*)}{\|G(\mathbf{z}^*)\|_2} \in \mathcal{K}$ will be satisfied, for instance, when the generative model is a ReLU network with no offsets (see [37, Remark 2.1]), due to $\mathcal{K}$ being cone-shaped. The sub-Gaussianity constant is indeed dependent on $\mathbf{z}^*$, but it can be upper bounded independently of $\mathbf{z}^*$ in special cases of interest, including any model in which the measurements are uniformly bounded (e.g., including not only 1-bit, but also more general multi-bit quantized models). We will point out these examples, but also highlight that these assumptions may pose some limitations.

Despite the slight limitations of the GMW-based and NN-based results, we note that these are relatively minor corollaries, and hope that the final decision is primarily based on our main theorems.

(Flow of main body) We would be happy to move some of the less central corollaries (e.g., Sections 4.2 and 4.5) to the appendix for improved flow in the main body. We could use the extra space for additional intuition and/or brief outlines for the main proofs, as suggested by R4.

[Responses to R2]: **(LEP in Assumption 1)** The intuition behind the LEP in Definition 2 is simply that if $\mathbf{x}_1$ is close to $\mathbf{x}_2$, then $\tilde{f}(\tilde{\mathbf{A}}\mathbf{x}_1)$ is close to $\tilde{f}(\tilde{\mathbf{A}}\mathbf{x}_2)$. The statement of Assumption 1 is somewhat more cumbersome for technical reasons (e.g., to make (114) rigorous), but we will aim to further highlight this intuition. Our paper includes formal verifications of the LEP for the linear and 1-bit models, and we expect that further examples could be established, but we defer this to future work. (Please see "Responses to multiple reviewers" above regarding the other comments.)

[Response to R3]: **(GMW)** Indeed, if the GMW is defined for $\mathcal{K}$ instead of $\mathcal{K} - \mathcal{K}$, the factor 2 can be omitted; however, $\mathcal{K} - \mathcal{K}$ is more commonly used. In the revision, we will further highlight that $Lr = n^{\Omega(1)}$ is typical for neural networks [2] after stating Theorem 1. We will also correct the typo in Line 202.

[Response to R4]: **(Proofs)** In our experience, having all proofs in the appendix is not uncommon for theory papers at NeurIPS. However, we would be happy to include some additional proof intuition/outlines in the revision, using the 9th page and/or some space freed up by moving Sections 4.2 and 4.5 to the appendix.