We thank all the reviewers for their time and valuable comments. For space limitation, we focus on addressing the main comments. **Reviewer #1** wants to see an algorithm that works when b^* has negative values. We show below that our algorithm can still be used in that case to recover part of the parameters with small number of samples. Both **Reviewer #2** and **Reviewer #3** ask about generalization to other settings. We discuss below one possible approach to learn a two-layer generative model. Extending our results to more general settings is definitely an interesting direction and we hope that our current work can encourage more people to work on this important problem.

Reviewer #1

"Provide an algorithm to output a distribution that's close to the target, even if b has negative components."

When b^* has negative components, running our algorithm can still recover part of the parameters. Specifically, let $\Omega := \{i \in [d] : b^*(i) \geq 0\}$ be the set of coordinates that b^* is non-negative, then the output of our algorithm $\widehat{b}$ and $\widehat{\Sigma}$ satisfies: 1) the sub-vector $\widehat{b}_\Omega$ is close to b^*_Ω; 2) the sub-matrix $\widehat{\Sigma}_{\Omega\times\Omega}$ is close to $W^*_\Omega W^{*T}_\Omega$. This is because our algorithm only uses the i-th and j-th coordinates of the samples to estimate $\langle W^*(i,:), W^*(j,:)\rangle$ and $b^*(i)$, $b^*(j)$. As a result, our guarantee (Theorem 1 in our paper) still holds for this part of the parameters. We will mention this in the paper.

For the rest part of the parameters, if the negative components of b^* are small (in absolute value), then the error of our algorithm will be also small. Specifically, let Ω^c be the complement of Ω. Suppose that $b^*(i) \geq -\eta\|W^*(i,:)\|_2$ for all $i \in \Omega^c$ and for some $\eta \geq 0$, then given $\widetilde{O}(\ln^2(d)/\epsilon^2)$ samples, the output of our algorithm satisfies $|\widehat{b}(i) - b^*(i)| \leq \max(\eta, \epsilon)\|W^*(i,:)\|_2$, for all $i \in \Omega^c$. One can show a similar result for $\langle W^*(i,:), W^*(j,:)\rangle$, for all $i \in \Omega^c$. We see that the error from negative bias is small if $\eta = O(\epsilon)$. If η is large, i.e., if b^* have large negative components, then estimating those parameters becomes difficult (as indicated by Claim 2 in our paper). In that case, maybe one should directly estimate the distribution (as suggested by the reviewer). This is an interesting direction for future research.

Reviewer #2 and #3

"What happens when we increase the number of layers?"

Besides the single-layer ReLU generative model considered in our paper, we also thought about extending our results to learning a two-layer generative model. Let $\mathcal{D}(A, W, b)$ be the distribution of a random variable $x \in \mathbb{R}^d$ defined by

$$x = A\ \mathrm{ReLU}(Wz + b),\ \text{where } z \sim \mathcal{N}(0, I_k), A \in \mathbb{R}^{d\times p}, W \in \mathbb{R}^{p\times k}, b \in \mathbb{R}^p.$$

Given i.i.d. samples $x \sim \mathcal{D}(A, W, b)$, can we recover the parameters A, W, b (up to permutation and scaling of the column vectors in A)? While this problem seems hard in general, we find an interesting connection between this problem and non-negative matrix factorization (NMF).

In MNF, we are given a non-negative matrix $X \in \mathbb{R}^{d\times n}$ and an integer $p > 0$, the goal is to find two non-negative matrices $A \in \mathbb{R}^{d\times p}, M \in \mathbb{R}^{p\times n}$ such that $X = AM$. This problem is NP-hard and [AGKM12] give the first polynomial-time algorithm under the "separability" assumption (Definition 5.1 in [AGKM12]).

In our problem, we are given n samples $\{x_i\}_{i=1}^n$ from $\mathcal{D}(A, W, b)$. Stacking the samples gives a matrix $X \in \mathbb{R}^{d\times n}$:

$$X = AM,\ \text{where } M(:,i) = \mathrm{ReLU}(Wz_i + b), i \in [n].$$

Note that $M \in \mathbb{R}^{p\times n}$ is non-negative while the entries of A can have *arbitrary* sign. If M satisfies the "separability" condition [AGKM12], and A has full column rank (i.e., the columns of A are linearly independent), then we can still use the same idea of [AGKM12] to *exactly* recover A and M (up to permutation and scaling of the column vectors in A). Once $M \in \mathbb{R}^{p\times n}$ is recovered, estimating W and b is the same problem as learning one-layer ReLU generative model, which can be done by our algorithm. One problem with the above approach is that it requires the $M \in \mathbb{R}^{p\times n}$ matrix to satisfy the "separability" condition. This is true when, e.g., W has full row rank, and the number of samples is $\Omega(2^k)$. Developing sample-efficient algorithms for more general cases is definitely an important research direction.

Reviewer #2

"Does similar results extend to more general input distributions?"

This is an interesting research direction. In our paper we focus on the standard Gaussian distribution for two reasons: 1) It has already been used in VAEs, GANs, and reversible generative models as the input distribution; 2) Even for this simple input distribution, we already encountered some technical difficulties such as negative bias vector (see our response to Reviewer #1). It is not easy to directly extend our algorithm to other input distributions, but our high-level idea, i.e., first estimate the norm and then estimate the pairwise angle, may still be useful.

References

[AGKM12] Arora, Sanjeev and Ge, Rong and Kannan, Ravindran and Moitra, Ankur. Computing a nonnegative matrix factorization–provably. *Forty-fourth Annual ACM Symposium on Theory of Computing (STOC)*, 2012.