Diffusion Posterior Sampling is Computationally Intractable

Shivam Gupta
UT Austin
shivamgupta@utexas.edu

Ajil Jalal
UC Berkeley
ajiljalal@berkeley.edu

Aditya Parulekar
UT Austin
adityaup@cs.utexas.edu

Eric Price
UT Austin
ecprice@cs.utexas.edu

Zhiyang Xun
UT Austin
zxun@cs.utexas.edu

February 21, 2024

Abstract

Diffusion models are a remarkably effective way of learning and sampling from a distribution $p(x)$. In posterior sampling, one is also given a measurement model $p(y \mid x)$ and a measurement y, and would like to sample from $p(x \mid y)$. Posterior sampling is useful for tasks such as inpainting, super-resolution, and MRI reconstruction, so a number of recent works have given algorithms to heuristically approximate it; but none are known to converge to the correct distribution in polynomial time.

In this paper we show that posterior sampling is *computationally intractable*: under the most basic assumption in cryptography—that one-way functions exist—there are instances for which *every* algorithm takes superpolynomial time, even though *unconditional* sampling is provably fast. We also show that the exponential-time rejection sampling algorithm is essentially optimal under the stronger plausible assumption that there are one-way functions that take exponential time to invert.

Contents

1 Introduction

Over the past few years, diffusion models have emerged as a powerful way for representing distributions of images. Such models, such as Dall-E [RDN+22] and Stable Diffusion [RBL+21], are very effective at learning and sampling from distributions. These models can then be used as priors for a wide variety of downstream tasks, including inpainting, superresolution, and MRI reconstruction.

Diffusion models are based on representing the *smoothed scores* of the desired distribution. For a distribution $p(x)$, we define the smoothed distribution $p_\sigma(x)$ to be p convolved with $\mathcal{N}(0, \sigma^2 I)$. These have corresponding smoothed scores $s_\sigma(x) := \nabla \log p_\sigma(x)$. Given the smoothed scores, the distribution p can be sampled using an SDE [HJA20] or an ODE [SME21]. Moreover, the smoothed score is the minimizer of what is known as the *score-matching* objective, which can be estimated from samples.

Sampling via diffusion models is fairly well understood from a theoretical perspective. The sampling SDE and ODE are both fast (polynomial time) and robust (tolerating L_2 error in the estimation of the smoothed score). Moreover, with polynomial training samples of the distribution, the empirical risk minimizer (ERM) of the score matching objective will have bounded L_2 error, leading to accurate samples [BMR20, GPPX23]. So diffusion models give fast and robust unconditional samples.

But sampling from the original distribution is not the main utility of diffusion models: that comes from using the models to solve downstream tasks. A natural goal is to sample from the *posterior*: the distribution gives a prior $p(x)$ over images, so given a noisy measurement y of x with known measurement model $p(y \mid x)$, we can in principle use Bayes' rule to compute and sample from $p(x \mid y)$. Often (such as for inpainting, superresolution, MRI reconstruction) the measurement process is the noisy linear measurement model, with measurement $y = Ax + \eta$ for a known measurement matrix $A \in \mathbb{R}^{m \times d}$ with $m < d$, and Gaussian noise $\eta = \beta \mathcal{N}(0, I_m)$; we will focus on such linear measurements in this paper.

Posterior sampling has many appealing properties for image reconstruction tasks. For example, if you want to identify x precisely, posterior sampling is within a factor 2 of the minimum error possible for *every* measurement model and *every* error metric [JAD+21]. When ambiguities do arise, posterior sampling has appealing fairness guarantees with respect to sensitive attributes [JKH+21].

Given the appeal of posterior sampling, the natural question is: is efficient posterior sampling possible given approximate smoothed scores? A large number of recent papers [JAD+21, CKM+23, KVE21, TWN+23, SVMK23, KEES22, DS24] have studied algorithms for posterior sampling, with promising empirical results. But all these fail on some inputs; can we find a better posterior sampling algorithm that is fast and robust in all cases?

There are several reasons for optimism. First, there's the fact that *unconditional* sampling is possible from approximate smoothed scores; why not posterior sampling? Second, we know that *information-theoretically, it is possible*: rejection sampling of the unconditional samples (as produced with high fidelity by the diffusion process) is very accurate with fairly minimal assumptions. The only problem is that rejection sampling is slow: you need to sample until you get lucky enough to match on every measurement, which takes time exponential in m.

And third, we know that the *unsmoothed* score of the posterior $p(x \mid y)$ is computable efficiently from the unsmoothed score of $p(x)$ and the measurement model: $\nabla_x \log p(x \mid y) = \nabla \log p(x) + \nabla \log p(y \mid x)$. This is sufficient to run Langevin dynamics to sample from $p(x \mid y)$. Of course, this has the same issues that Langevin dynamics has for unconditional sampling: it can take exponential time to mix, and is not robust to errors in the score. Diffusion models fix this by using the *smoothed* score to get robust and fast (unconditional) sampling. It seems quite plausible that a sufficiently

clever algorithm could also get robust and fast posterior sampling.

Despite these reasons for optimism, in this paper we show that **no fast posterior sampling algorithm exists**, even given good approximations to the smoothed scores, under the most basic cryptographic assumption that one-way functions exist. In fact, under the further assumption that some one-way function is exponentially hard to invert, there exists a distribution—one for which the smoothed scores are well approximated by a neural network so that *unconditional* sampling is fast—that takes exponential in m time for posterior sampling. Rejection sampling takes time exponential in m, and so, one can no longer hope for much general improvement over rejection sampling.

Precise statements. To more formally state our results, we make a few definitions. We say a distribution is "well-modeled" if its smoothed scores can be represented by a polynomial size neural network to polynomial precision:

Definition 1.1 (C-Well-Modeled Distribution)**.** *For any constant $C > 0$, we say a distribution p over $\mathbb{R}^d$ with covariance Σ is "C-well-modeled" by score networks if $\|\Sigma\| \lesssim 1$ and there are approximate scores $\widehat{s}_\sigma$ that satisfy*

$$\mathop{\mathbb{E}}_{x\sim p_\sigma}[\|\widehat{s}_\sigma(x) - s_\sigma(x)\|^2] < \frac{1}{d^C\sigma^2}$$

and can be computed by a $\mathrm{poly}(d)$*-parameter neural network with* $\mathrm{poly}(d)$*-bounded weights for every $\frac{1}{d^C} \le \sigma \le d^C$.*

Throughout our paper we will be comparing similar distributions. We say distributions are (τ, δ) close if they are close up to some shift τ and failure probability δ:

Definition 1.2 ((τ,δ)-Close Distribution)**.** *We say the distribution of x and $\widehat{x}$ are (τ, δ) close if they can be coupled such that*

$$\mathbb{P}[\|x - \widehat{x}\| > \tau] < \delta.$$

An *unconditional* sampler is one that is (τ, δ) close to the true distribution.

Definition 1.3 ((τ,δ)-Unconditional Sampler)**.** *A (τ, δ) unconditional sampler of a distribution $\mathcal{D}$ is one where its samples $\widehat{x}$ are (τ, δ) close to the true $x \sim \mathcal{D}$.*

The theory of diffusion models [CCL+23] says that the diffusion process gives an unconditional sampler for well-modeled distributions that takes polynomial time (with the precise polynomial improved by subsequent work [BBDD24]).

Theorem 1.4 (Unconditional Sampling for Well-Modeled Distributions)**.** *For an $O(C)$-well-modeled distribution p, the discretized reverse diffusion process with approximate scores gives a $\left(\frac{1}{d^C}, \frac{1}{d^C}\right)$-unconditional sampler (as defined in Definition 1.3) for any constant $C > 0$ in* $\mathrm{poly}(d)$ *time.*

But what about *posterior* samplers? We want that, for most measurements y, the conditional distribution is (τ, δ) close to the truth:

Definition 1.5 ((τ,δ)-Posterior Sampler)**.** *Let $\mathcal{D}$ be a distribution over $X \times Y$ with density $p(x, y)$. Let C be an algorithm that takes in $y \in Y$ and outputs samples from some distribution $\widehat{p}_{|y}$ over X. We say C is a (τ, δ)-Posterior Sampler for $\mathcal{D}$ if, with $1 - \delta$ probability over $y \sim \mathcal{D}_Y$, $\widehat{p}_{|y}$ and $p(x \mid y)$ are (τ, δ) close.*

As described above, we consider the linear measurement model:

Definition 1.6 (Linear measurement model)**.** *In the linear measurement model with m measurements and noise parameter β, we have for $x \in \mathbb{R}^d$, the measurement $y = Ax + \eta$ for $A \in \mathbb{R}^{m \times d}$ normalized such that $\|A\| \leq 1$, and $\eta = \beta\mathcal{N}(0, I_m)$.*

One way to implement posterior sampling is by rejection sampling. As long as the measurement noise β is much bigger than the error $\tau = \frac{1}{\mathrm{poly}(d)}$ from the diffusion process, this is accurate. However, the running time is exponentially large in m:

Theorem 1.7 (Upper bound)**.** *Let $C > 1$ be a constant. Consider an $O(C)$-well-modeled distribution and a linear measurement model with $\beta > \frac{1}{d^C}$. When $\delta > \frac{1}{d^C}$, rejection sampling of the diffusion process gives a $(\frac{1}{d^C}, \delta)$-posterior sampler that takes $\mathrm{poly}(d)(\frac{O(1)}{\beta\sqrt{\delta}})^m$ time.*

Our main result is that this is nearly tight:

Theorem 1.8 (Lower bound)**.** *Suppose that one-way functions exist. Then for any $d^{0.01} < m < d/2$, there exists a 10-well-modeled distribution over $\mathbb{R}^d$, and linear measurement model with m measurements and noise parameter $\beta = \Theta(\frac{1}{\log^2 d})$, such that $(\frac{1}{10}, \frac{1}{10})$-posterior sampling requires superpolynomial time.*

To be a one-way function, inversion must take superpolynomial time (on average). For many one-way function candidates used in cryptography, the best known inversion algorithms actually take exponential time. Under the stronger assumption that some one-way function exists that requires exponential time to invert with non-negligible probability, we can show that posterior sampling takes $2^{\Omega(m)}$ time:

Theorem 1.9 (Lower bound: exponential hardness)**.** *For any $m \leq d/2$, suppose that there exists a one-way function $f : \{\pm 1\}^m \to \{\pm 1\}^m$ that requires $2^{\Omega(m)}$ time to invert. Then for any $C > 1$, there exists a C-well-modeled distribution over $\mathbb{R}^d$ and linear measurement model with m measurements and noise level $\beta = \frac{1}{C^2 \log^2 d}$, such that $(\frac{1}{10}, \frac{1}{10})$-posterior sampling takes at least $2^{\Omega(m)}$ time.*

Assuming such strong one-way functions exist, then for the lower bound instance, $2^{\Omega(m)}$ time is necessary and rejection sampling takes $2^{O(m \log\log d)}\,\mathrm{poly}(d)$ time. Up to the $\log\log d$ factor, this shows that rejection sampling is the best one can hope for in general.

Remark 1.10. *The lower bound produces a "well-modeled" distribution, meaning that the scores are representable by a polynomial-size neural network, but there is no requirement that the network be shallow. One could instead consider only shallow networks; the same theorem holds, except that f must also be computable by a shallow depth network. Many candidate one-way functions can be computed in NC^0 (i.e., by a constant-depth circuit) [AIK04], so the cryptographic assumption is still mild.*

2 Related Work

Diffusion models [SDWMG15, DN21, SE19] have emerged as the most popular approach to deep generative modeling of images, serving as the backbone for the recent impressive results in text-to-image generation [RDN+22, RBL+21], along with state-of-the-art results in video [BDK+23, HSG+22] and audio [KPH+21, CZZ+21] generation.

Noisy linear inverse problems capture a broad class of applications such as image inpainting, super-resolution, MRI reconstruction, deblurring, and denoising. The empirical success of diffusion models has motivated their use as a data prior for linear inverse problems, *without* task-specific

training. There have been several recent theoretical and empirical works [JAD+21, CKM+23, KVE21, TWN+23, SVMK23, KEES22, DS24] proposing algorithms to sample from the posterior of a noisy linear measurement. We highlight some of these approaches below.

Posterior Score Approximation. One class of approaches [CKM+23, KVE21, SVMK23] *approximates* the intractable posterior score $\nabla \log p_t(x_t|y) = \nabla \log p_t(x_t) + \nabla \log p(y|x_t)$ at time t of the reverse diffusion process, and uses this approximation to sample. Here, $y = Ax_0 + \eta$ is the noisy measurement of $x_0 \sim p_0$, where p_t is the density at time t. For instance, [CKM+23] proposes the approximation $\nabla \log p(y|x_t) \approx \nabla \log p(x|\,\mathbb{E}\,[x_0|x_t])$, thereby incurring error quantified by the so-called *Jensen gap*. [SVMK23] proposes an approximation based on the pseudoinverse of A, while [KVE21] proposes to use the score of the posterior wrt measurement y_t of x_t.

Replacement Method. Another approach, first introduced in the context of inpainting [LDR+22], replaces the observed coordinates of the sample with a noisy version of the observation during the reverse diffusion process. An extension was proposed for general noisy linear measurements [KEES22]. This approach essentially also attempts to sample from an approximation to the posterior.

Particle Filtering. A recent set of works [TWN+23, TYT+23, DS24] makes use of Sequential Monte Carlo (SMC) methods to sample from the posterior. These methods are guaranteed to sample from the correct distribution as the number of particles goes to ∞. Our paper implies a lower bound on the number of particles necessary for good convergence. Assuming one-way functions exist, polynomially many particles are insufficient in general, so that these algorithms takes superpolynomial time; assuming some one-way function requires exponential time to invert, particle filtering requires exponentially many particles for convergence.

To summarize, our lower bound implies that these approaches are either approximations that fail to sample from the posterior, and/or suffer from prohibitively large runtimes in general.

3 Proof Overview – Lower Bound

In this section, we give an overview of the proof of our main Theorem 1.8, which states that there is some well-modeled distribution for which posterior sampling is hard. The full proof can be found in the Appendix.

An initial attempt. Given a one-way function $f : \{-1,1\}^d \to \{-1,1\}^d$, consider the distribution that is uniform over $(s, f(s)) \in \{-1,1\}^{2d}$ for all $s \in \{-1,1\}^d$. This distribution is easy to sample from unconditionally: sample s uniformly, then compute $f(s)$. At the same time, posterior sampling is hard: if you observe the last d bits, i.e. $f(s)$, a posterior sample should be from $f^{-1}(f(s))$; and if f is a one-way function, finding any point in this support is computationally intractable on average.

However, it is not at all clear that this distribution is well-modeled as per Definition 1.1; we would need to be able to accurately represent the smoothed scores by a polynomial size neural network. The problem is that for smoothing levels $1 \ll \sigma \ll \sqrt{d}$, the smoothed score can have nontrivial contribution from many different $(s, f(s))$; so it's not clear one can compute the smoothed scores efficiently. Thus, while posterior sampling is intractable in this instance, it's possible the hardness lies in representing and computing the smoothed scores using a diffusion model, rather than in *using* the smoothed scores for posterior sampling.

However, for smoothing levels $\sigma \ll \frac{1}{\sqrt{\log d}}$, the smoothed scores *are* efficiently computable with high accuracy. The smoothed distribution is a mixture of Gaussians with very little overlap, so rounding to a nearby Gaussian and taking its score gives very high accuracy.

To design a better lower bound, we modify the distribution to encode $f(s)$ differently: into the phase of the discretization of a Gaussian. At large smoothing levels, a discretized Gaussian looks essentially like an undiscretized Gaussian, and the phase information disappears. Thus at large smoothing levels, the distribution is essentially like a product distribution, for which the scores are easy to compute. At the same time, conditioning on the observations still implies inverting f, so this is still hard to conditionally sample; and it's still the case that small smoothing levels are efficiently computable.

Based on the above, we define our lower bound instance formally in Section 3.1. Then, in Section 3.2 we sketch a proof of Lemma 3.4, which shows that it is impossible to perform accurate posterior sampling for our instance, under standard cryptographic assumptions. Section 3.3 shows that our lower bound distribution is well-modeled by a small ReLU network, which means that the hardness is not coming merely from inability to represent the scores, and that *unconditional* sampling is provably efficient. Finally, we put these observations together to show the theorem.

3.1 Lower Bound Instance

We define our lower bound instance here formally. Let $w_\sigma(x)$ denote the density of a Gaussian with mean zero and standard deviation σ, and let $\mathrm{comb}_\varepsilon$ denote the Dirac Comb distribution with period ε, given by

$$\mathrm{comb}_\varepsilon(x) = \sum_{k=-\infty}^{\infty} \delta(x - k\varepsilon)$$

For any $b \in \{-1, 1\}$, let ψ_b be the density of a standard Gaussian discretized to multiples of ε, with phase either 0 or $\frac{\varepsilon}{2}$ depending on b:

$$\psi_b(x) \propto w_1(x) \cdot \mathrm{comb}_\varepsilon\left(x - \varepsilon/2 \cdot \frac{1-b}{2}\right).$$

Definition 3.1 (Unscaled Lower Bound Distribution)**.** *Let $f : \{\pm 1\}^d \to \{\pm 1\}^{d'}$ be a given function. For $R > 0$ and for any $s \in \{\pm 1\}^d$, define the product distribution g_s over $x \in \mathbb{R}^{d+d'}$ such that*

$$\begin{aligned} x_i &\sim w_1(x_i - R \cdot s_i) && \text{for } i \le d \\ x_i &\sim \psi_{f(s)_{i-d}} && \text{for } i > d. \end{aligned}$$

The unconditional distribution g we consider is the uniform mixture of g_s over $s \in \{\pm 1\}^d$.

We will have $d' = O(d)$ throughout. Figure 1 gives a visualization of g_s; the final distribution is the mixture of g_s over uniformly random s.

For ease of exposition, we will also define a scaled version of our distribution g such that its covariance Σ has $\|\Sigma\| \lesssim 1$.

Definition 3.2 (Scaled Lower Bound Distribution)**.** *Let $\widetilde{g}(x) = R^{d+d'} g(R \cdot x)$ be the scaled version of the distribution with density g defined in Definition 3.1. Similarly, let $\widetilde{g}_s = R^{d+d'} g_s(R \cdot x)$.*

The measurement process then takes sample $x \sim \widetilde{g}$ and computes $Ax + \eta$, where $\eta = N(0, \beta^2 I_{d'})$ and $A = \begin{pmatrix} 0^{d' \times d} & I_{d'} \end{pmatrix}$. That is, we observe the last d' bits of x, with variance β^2 Gaussian noise added to each coordinate.

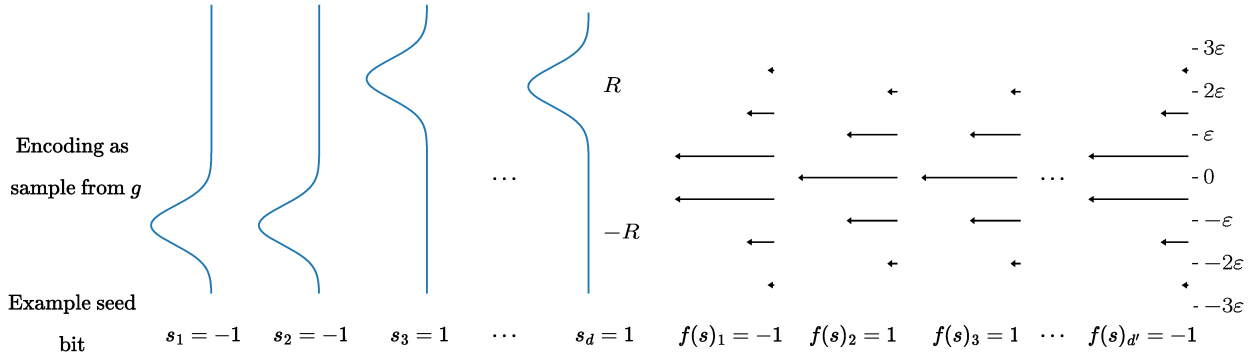


Figure 1: The distribution of each coordinate in g_s, has independent coordinates. For any seed $s \in \{\pm 1\}^d$, the first d bits are normal distributions whose *mean* is specified by s_i, and the last d' bits are a discretized standard normal where the *discretization* is specified by $f(s)_j$. The full distribution g is a mixture over all seeds s of g_s.

3.2 Posterior Sampling Implies Inversion

Below, we state the main result of this section, and give a sketch of the proof. We show that given any function $f : \mathbb{R}^d \to \mathbb{R}^m$, if we can conditionally sample the above measurement process, then we can invert f. For the sake of exposition, we assume here that f has unique inverses; a similar argument applies in general. The full proof of this Lemma is given in the Appendix.

Lemma 3.3. *For any function f, suppose C is an $(1/10, 1/10)$-posterior sampler in the linear measurement model with noise parameter β for distribution with density $\widetilde{g}$ as defined in Definition 3.2, with $\varepsilon \geq \beta\sqrt{32 \log d}$ and $R \geq 32\sqrt{\log d}$. If C takes time T to run, then there exists an algorithm A that runs in time $T + O(d)$ such that*

$$\mathbb{P}_{s,\mathsf{A}}[f(\mathsf{A}(f(s))) \neq f(s)] \leq \frac{3}{4}$$

Take some $z \in \{\pm 1\}^{d'}$. Our goal is to compute $f^{-1}(z)$, using the posterior sampler for $\widetilde{g}$. To do this, we take a sample $\overline{z}_i \sim \psi_{z_i} * N(0, \beta^2)$, for $i \in [d']$, and feed in $\overline{z}$ into our posterior sampler, to output $\widehat{x}$. We then take the first d bits of $\widehat{x}$, round each entry to the nearest ± 1, and output the result.

To see why this works, let's analyze what the resulting conditional distribution looks like. First, note that any sample $x \sim \widetilde{g}$ encodes some $(s, f(s))$ coordinate-wise so that the encoding of $f(s)$ is one of two discretizations of a normal distribution, with width ε, offset by $\varepsilon/2$ from each other (see Figure 1). Furthermore, since $\beta \ll \varepsilon$, these two encodings are distinguishable with high probability even after adding noise with variance β^2. Therefore, with high probability, our sample $\overline{z}$, which is a noised and discretized encoding of the input z we want to invert, will be such that each coordinate is within $\varepsilon/4$ of the correct discretization. Consequently, a posterior sample with this observation will correspond to an encoding of $(s, f(s))$ where $s = f^{-1}(z)$, with high probability. The first d bits of this encoding are just the bits of $f^{-1}(z)$ smoothed by a gaussian with variance $1/R^2$, and since $R \gg 1$, rounding these coordinates to the nearest ± 1 returns $f^{-1}(z)$, with high probability.

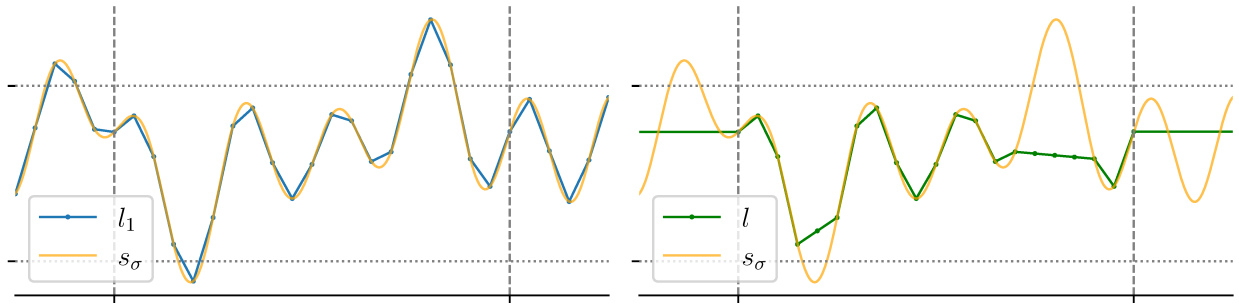
So, we showed how to invert an arbitrary f using a posterior sampler. The runtime of this procedure was just the runtime of the posterior sampler, along with some small overhead. In

particular, if f were a one-way function that takes superpolynomial time to invert, posterior sampling *must* take superpolynomial time. Formally, we show the following:

Lemma 3.4. *Suppose $d^{0.01} \le m \le d/2$ and one-way functions exist. Then, for $\widetilde{g}$ as defined in Definition 3.2 with $\varepsilon = \frac{1}{C\sqrt{\log d}}$ and $R = C \log d$, and linear measurement model with noise parameter $\beta = \frac{1}{C^2 \log^2 d}$ and measurement matrix $A \in \mathbb{R}^{m \times d}$, $(\frac{1}{10}, \frac{1}{10})$-posterior sampling takes superpolynomial time.*

One minor detail is that a one-way function is defined to map $\{0,1\}^n \to \{0,1\}^{n'}$ for an unconstrained n', while we want one that maps $\{0,1\}^{d-m} \to \{0,1\}^m$. Standard arguments imply that we can get such a function from the assumption; see Section G for details.

3.3 ReLU Approximation of Lower Bound Score



(a) l_1 is unbounded and has an unbounded number of pieces (b) l is bounded, has a small number of pieces.

Figure 2: Piecewise-Linear Approximations of Score s_σ

We have shown that our (scaled) lower bound distribution $\widetilde{g}$ (as defined in Definition 3.2) is computationally intractable to sample from. Now, we sketch our proof showing that $\widetilde{g}$ is well-modeled: the σ-smoothed scores are well approximated by a polynomially bounded ReLU network. The main result of this section is the following.

Corollary 3.5 (Lower Bound Distribution is Well-Modeled)**.** *Let C be a sufficiently large constant. Given a ReLU network $f : \{\pm 1\}^d \to \{\pm 1\}^{d'}$ with $\mathrm{poly}(d)$ parameters bounded by $\mathrm{poly}(d)$ in absolute value, the distribution $\widetilde{g}$ defined in Definition 3.2 for $R = C \log d$ and $\frac{1}{\mathrm{poly}(d)} < \varepsilon < \frac{1}{C\sqrt{\log d}}$, is $O(C)$-well-modeled.*

To show this, we will first show that the unscaled distribution g has a score approximation representable by a and polynomially bounded ReLU net. Rescaling by a factor of $R = C \log d$ then shows the above.

Notation. We will let h be the σ-smoothed version of g, and h_r be the σ-smoothed version of g_r.

Strategy. We will first show how to approximate the score of any σ-smoothed *product* distribution using a polynomial-size ReLU network with polynomially bounded weights in our dimension d, $\frac{1}{\sigma}$ and $\frac{1}{\gamma}$ for L^2 error γ^2.

Then, we will observe that when σ is *large*, so that $\mathrm{poly}(d) \ge \sigma \gg \varepsilon\sqrt{\log d}$, h becomes very close to a mixture of $(1+\sigma^2)I_{d+d'}$-covariance Gaussians placed at the vertices of a scaled hypercube (in

the first d coordinates). Since this is a product distribution, we can represent its score using our ReLU construction.

On the other hand, when σ is *small*, for $R \gg \log d$ and $\frac{1}{\text{poly}(d)} \leq \sigma \ll \frac{R}{\sqrt{\log d}}$, the score of h at any point x is well approximated by the distribution h_r, where $r \in \{\pm 1\}^d$ represents the orthant containing the first d coordinates of x. Since h_r is a product distribution, our ReLU construction applies.

Finally, we set $R \gg \log d$ so that for any $\frac{1}{\text{poly}(d)} \leq \sigma \leq \text{poly}(d)$, there is a polynomially bounded ReLU net that approximates the score of h. We now describe each of these steps in more detail.

3.3.1 ReLU Approximation for Score of Product Distribution

We will show first how to construct a ReLU network approximating the score of a *one-dimensional* distribution – the construction generalizes to product distributions in a straightforward way.

Consider any one-dimensional distribution p with σ-smoothed version p_σ, and corresponding score s_σ. Suppose p_σ has standard deviation m_2. We will first construct a *piecewise-linear* function l that approximates s_σ in L^2.

Since s_σ is σ-smoothed, its value does not change much in *most* σ-sized regions. More precisely, Lemma H.1 shows that

$$\mathop{\mathbb{E}}_{x \sim p_\sigma} \left[\sup_{|c| \leq \sigma} s'_\sigma(x+c)^2 \right] \lesssim \frac{1}{\sigma^4}$$

This immediately gives a piecewise linear-approximation l_1 with $O(\gamma\sigma^2)$-width pieces: By Taylor expansion, we can write any $s_\sigma(x) = s_\sigma(\alpha_x) + (x - \alpha_x)s'_\sigma(\xi)$ for some ξ between α_x and x. Then, if α_x is the largest discretization point smaller than x (so that $|x - \alpha_x| \lesssim \gamma\sigma^2$), this gives that

$$\begin{aligned}\mathbb{E}\left[(s_\sigma(x) - s_\sigma(\alpha_x))^2\right] &\lesssim \gamma^2\sigma^4 \,\mathbb{E}[\sup_c s'_\sigma(x+c)^2] \\ &\lesssim \gamma^2\end{aligned}$$

So, we can approximate every $s_\sigma(x)$ with $s_\sigma(\alpha_x)$, yielding a piecewise-*constant* approximation. Then, we can similarly obtain another piecewise-constant approximation by replacing $s_\sigma(x)$ with $s_\sigma(\beta_x)$ for β_x the smallest discretization point larger than x. By convexity, we can linearly interpolate between $s_\sigma(\alpha_x)$ and $s_\sigma(\beta_x)$ to obtain our piecewise-*linear* approximation l_1 (see Fig. 2).

Unfortunately, l_1 suffers from two issues: 1) It is potentially unbounded, and 2) It has an unbounded number of pieces.

For 1), since s_σ is σ-smoothed, it is bounded by with high probability, so that we can ensure that our approximation is also bounded without increasing its error much. For 2), since p_σ has standard deviation m_2, Chebyshev's inequality gives that the total probability outside a radius $\frac{m_2}{\gamma\sigma^2}$ region is small, so that we can use a constant approximation outside this region. This allows us to bound the number of pieces by poly $\left(\frac{m_2}{\gamma\sigma}\right)$, yielding our final approximation l.

As is well-known, such a piecewise linear function can be represented using a ReLU network with poly $\left(\frac{m_2}{\gamma\sigma}\right)$ parameters, and each parameter bounded by poly $\left(\frac{m_2}{\gamma\sigma}\right)$ in absolute value. For product distributions, we simply construct ReLU networks for each coordinate individually, and then append them, for bounds polynomial in d and $1/\sigma$, $1/\gamma$ and m_2. In the remaining proof, whenever this construction is used, all these parameters are set to polynomial in d, for final bounds poly(d).

3.3.2 ReLU Approximation for Large σ

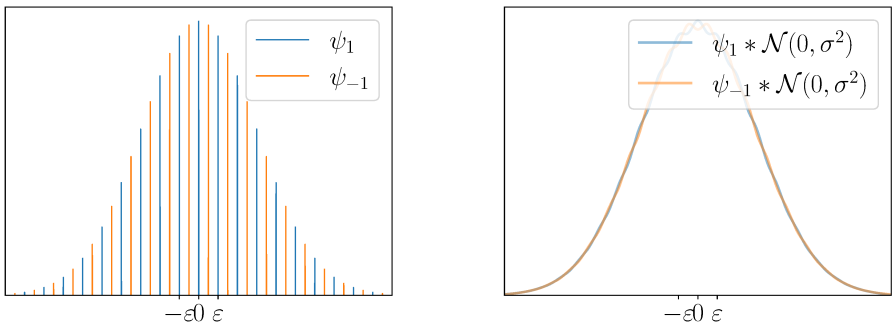


Figure 3: ψ_1 and ψ_{-1} are discretized Gaussians with discretization width ε and phase 0 and $\varepsilon/2$ respectively. If we convolve with $\mathcal{N}(0, \sigma^2)$, we get a distribution close to Gaussian when $\sigma \geq \varepsilon$ for each of ψ_1, ψ_{-1}.

Note that our lower bound distribution g is such that the first d coordinates are simply a mixture of Gaussians placed on the vertices of a (scaled) hypercube, while the last d' coordinates are discretized Gaussians ψ_1 or ψ_{-1}, with the choice of discretization depending on the first d coordinates.

The only reason g is not already a product distribution is that ψ_1 and ψ_{-1} are different. But for smoothing $\sigma \gg \varepsilon\sqrt{\log d}$, a Fourier argument shows that the smoothed versions of ψ_1 and ψ_{-1} are polynomially close to each other. See Figure 3 for an illustration.

3.3.3 ReLU Approximation for Small σ

When $\sigma \ll \frac{R}{\sqrt{\log d}}$ and $R \gg \log d$, consider the density $h(x)$ for $x_{1,\ldots,d}$ lying in the orthant identified by $r \in \{\pm 1\}^d$. Recall that

$$h(x) = \frac{1}{2^d} \sum_{s \in \{\pm 1\}^d} h_s(x)$$

where h_s is the product distribution that is Gaussian with mean $R \cdot s_i$ in the first d coordinates and is a smoothed discretized Gaussian with mean 0 in the remaining d' coordinates.

We first show that $h(x)$ is approximated by $\frac{h_r(x)}{2^d}$ up to small additive error. This is because every h_s has radius at most $\sqrt{1+\sigma^2} \lesssim \frac{R}{\sqrt{\log d}}$ and there are $\approx \binom{d}{k}$ points $s \neq r$ with the mean of h_s at least $\Omega(\sqrt{k}R)$ away from x. So, the total contribution of all the terms involving $h_s(x)$ to $h(x)$ for $s \neq r$ is at most $\approx \frac{1}{2^d} \cdot \frac{1}{\mathrm{poly}(d)}$. We can show that $\nabla h(x)$ is approximated by $\frac{\nabla h_r(x)}{2^d}$ in L^2 up to similar additive error in an analogous way.

We then show that the score of h_r serves as a good approximation to the score of h for all such points x such that $x_{1,\ldots,d}$ lies in the orthant identified by r. For x close to the mean of h_r (to within $R/10$, say), the above gives that $h(x)$ is approximated up to *multiplicative* error by $\frac{h_r(x)}{2^d}$, and $\nabla h(x)$ is approximated up to multiplicative error by $\frac{\nabla h_r(x)}{2^d}$. Together, this gives that the *score* of h at x, $\frac{\nabla h(x)}{h(x)}$ is approximated by the score of h_r at x up to $\frac{1}{\mathrm{poly}(d)}$ error. On the other hand, for x far from the mean of h_r, since the density itself is small, the total contribution of such points to the score error is negligible.

Since the score of h is well-approximated by the score of h_r, and h_r is a *product* distribution, we can essentially use our ReLU construction for product distributions to represent its score, after using a small gadget to identify the orthant that $x_{1,\dots,d}$ lies in.

3.4 Putting it all Together

Lemma 3.4 shows that it is computationally hard to sample from $\widetilde{g}$ from the posterior of a noisy linear measurement when f is a one-way funciton, while Corollary 3.5 shows that $\widetilde{g}$ has score that is well-modeled by a ReLU network when f can be represented by a polynomial-sized ReLU network. In Section G, we show that any one-way function can be represented using a polynomial-sized ReLU network. Thus, together, these imply our lower bound, Theorem 1.8.

Essentially the same argument holds under the stronger guarantee that there exists a one-way function that takes exponential time to invert, for a lower bound *exponential* in the number of measurements m.

4 Proof Overview - Upper Bound

Algorithm 1: Rejection Sampling Algorithm

Input: $y \in Y$

while True **do**
  Sample $x \sim \mathcal{D}_x$
  Compute $q := e^{\frac{-\|Ax-y\|^2}{2\beta^2}}$ (proportional to $p(y \mid x)$)
  Generate a random number $r \sim U(0,1)$
  if $r < q$ **then**
    return x
  end if
end while

In this section, we sketch the proof of Theorem 1.7 in Section E: the time complexity of posterior sampling by rejection sampling (Algorithm 1). For ease of discussion, we only consider the case when $\delta = \Theta(1)$. The proof overview below will repeatedly refer to events as occurring with "arbitrarily high probability"; this means the statement is true for every constant probability $p < 1$. (Usually there will be a setting of constants in big-O notation nearby that depends on p.)

Sampling Correctness With Ideal Sampler. To illustrate the idea of the proof, we first focus on the scenario where we can sample from the distribution of x perfectly. We aim to show that rejection sampling perfectly samples $x \mid y$. To prove the correctness of Algorithm 1, noting that each round is independent, it suffices to verify that each round outputs x with probability density proportional to $p(x \mid y)$. We have

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} \propto p(y \mid Ax)p(x) \propto e^{-\frac{\|Ax-y\|}{2\beta^2}} p(x).$$

Therefore, with a perfect unconditional sampler for $\mathcal{D}_x$ (sampling x according to density $p(x)$), rejection sampling perfectly samples $x \mid y$.

Running time. Now we show that for linear measurements $y = Ax + \beta\mathcal{N}(0, I_d)$, with arbitrarily high probability over $x \sim \mathcal{D}$, the acceptance probability per round is at least $\Theta(\beta)^m$; this implies the algorithm terminates in $(O(1)/\beta)^m$ rounds with arbitrarily high probability. For a given y, the acceptance probability per round is equal to

$$\mathop{\mathbb{E}}_{x}\left[e^{-\frac{\|Ax-y\|^2}{2\beta^2}}\right] \geq \mathbb{P}_x\left[\|Ax-y\| \leq O(\beta\sqrt{m})\right] \cdot e^{-O(m)}.$$

We first focus on the case when $m = 1$. We aim to show that with arbitrarily high probability over y,

$$\mathbb{P}_x\left[\|Ax-y\| \leq O(\beta)\right] \geq \beta.$$

For well-modeled distributions, the covariance matrix of x has constant singular values. Then with arbitrarily high probability, x is $O(1)$ in each direction. Since every singular value of A is at most 1, the projection Ax onto $\mathbb{R}$ will lie in $[-C, +C]$ for some constant C with arbitrarily high probability.

We divide $[-C, +C]$ into $N = \frac{2C}{\beta}$ segments of length β, forming set S. Now we only need to prove that with arbitrarily high probability over y, there exists a segment $\theta \in S$ satisfying for all $x \in \theta$, $|x - y| \leq O(\beta)$, and $\mathbb{P}_{x\sim\mathcal{D}_x}[Ax \in \theta] \gtrsim \beta$. For any constant $c > 0$, define

$$S' := \{\theta \in S \mid \mathbb{P}_{x\sim\mathcal{D}_x}[Ax \in \theta] > \frac{c}{N}\}.$$

Each segment in S' has probability at least $\Omega(1/N) \gtrsim \beta$ to be hit. Therefore, we only need to prove that, with arbitrarily high probability, $y = Ax + \eta$ satisfies these two independent events simultaneously: (1) Ax lands in some segment $\theta \in S'$; (2) $\eta \lesssim \beta$.

By a union bound, the probability that Ax lies in a segment in $S \setminus S'$ is at most $N \cdot \frac{c}{N} \leq c$. For sufficiently small c, combining with the fact that $Ax \in S$ with arbitrarily high probability, we have (1) with arbitrarily high probability. Since that $\eta \sim \mathcal{N}(0, \beta^2)$. By the concentration of Gaussian distribution, (2) is satisfied with arbitrarily high probability.

For the general case when $m > 1$, with arbitrarily high probability, Ax will lie in $\{x \in \mathbb{R}^m \mid \|x\| \leq C\sqrt{m}\}$ for some $C > 0$. Instead of segments, we use $N = (\frac{O(1)}{\beta})^m$ balls with radius β to cover $\{x \in \mathbb{R}^m \mid \|x\| \leq C\sqrt{m}\}$. Following a similar argument, we can prove that with arbitrarily high probability over y,

$$\mathbb{P}_x\left[\|Ax-y\| \leq O(\beta\sqrt{m})\right] \geq \Theta(\beta)^m.$$

Diffusion as unconditional sampler. In practice, we do not have a perfect sampler for $\mathcal{D}_x$. Theorem 1.4 states that for $O(C)$-well-modeled distributions, diffusion model gives an unconditional sampler that samples from approximation distribution $\widehat{\mathcal{D}}_x$ satisfying that there exists a coupling between $x \sim \mathcal{D}_x$ and $\widehat{x} \sim \widehat{\mathcal{D}}_x$ such that with arbitrarily high probability, $\|x - \widehat{x}\| \leq 1/d^{2C}$.

For $(x, \widehat{x})$ drawn from this coupling, we know from our previous analysis that rejection sampling based on x is correct. But the algorithm only knows $\widehat{x}$, which changes its behavior in two ways: (1) it chooses to accept based on $p(y \mid \widehat{x})$ rather than $p(y \mid x)$, and (2) it returns $\widehat{x}$ rather than x on acceptance. The perturbation from (2) is easily within our tolerance, since it is $\frac{1}{d^{2C}}$ close to x with arbitrarily high probability.

For (1), we can show when x and $\widehat{x}$ are close, these two probabilities are nearly the same. When $\|x - \widehat{x}\| \leq \frac{1}{d^{2C}} \leq o(\beta/\sqrt{m})$, we have

$$\left|\log\frac{p(y \mid \widehat{x})}{p(y \mid x)}\right| = \left|\frac{\|Ax-y\|^2}{2\beta^2} - \frac{\|A\widehat{x}-y\|^2}{2\beta^2}\right| \leq o(1).$$

This implies that $p(y \mid \widehat{x}) = (1 \pm o(1))p(y \mid x)$ and proves Theorem 1.7.

5 Conclusion and Future Work

We have shown that one cannot hope for a fast general algorithm for posterior sampling from diffusion models, in the way that diffusion gives general guarantees for unconditional sampling. Rejection sampling, slow as it may be, is about the fastest one can hope for on some distributions. However, people run algorithms that attempt to approximate the posterior sampling every day; they might not be perfectly accurate, but they seem to do a decent job. What might explain this?

Given our lower bound, a positive theory for posterior sampling of diffusion models must invoke distributional assumptions on the data. Our lower bound distribution is derived from a one-way function, and not very "nice". It would be interesting to identify distributional properties under which posterior sampling is possible, as well as new algorithms that work under plausible assumptions.

Acknowledgements

We thank Xinyu Mao for discussions about the cryptographic assumptions. SG, AP, EP and ZX are supported by NSF award CCF-1751040 (CAREER) and the NSF AI Institute for Foundations of Machine Learning (IFML). AJ is supported by ARO 051242-002.

6 References

[AIK04] B. Applebaum, Y. Ishai, and E. Kushilevitz. Cryptography in NC^0. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 166–175, 2004.

[BBDD24] Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization, 2024.

[BDK+23] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023.

[BMR20] Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling. *ArXiv*, abs/2002.00107, 2020.

[CCL+23] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023.

[CKM+23] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023.

[CZZ+21] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2021.

[DN21] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman

Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

[DS24] Zehao Dou and Yang Song. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

[GLPV22] Shivam Gupta, Jasper Lee, Eric Price, and Paul Valiant. Finite-sample maximum likelihood estimation of location. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30139–30149. Curran Associates, Inc., 2022.

[GPPX23] Shivam Gupta, Aditya Parulekar, Eric Price, and Zhiyang Xun. Sample-efficient training for diffusion, 2023.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[HSG+22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022.

[JAD+21] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. Robust compressed sensing mri with deep generative priors. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14938–14954. Curran Associates, Inc., 2021.

[JKH+21] Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alex Dimakis, and Eric Price. Fairness for image generation with uncertain sensitive attributes. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4721–4732. PMLR, 18–24 Jul 2021.

[KEES22] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[KPH+21] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

[KVE21] Bahjat Kawar, Gregory Vaksman, and Michael Elad. SNIPS: Solving noisy inverse problems stochastically. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[LDR+22] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11451–11461, 2022.

[LM00] B. Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28, 10 2000.

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. 2018.

[RBL+21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021.

[RDN+22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[SDWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

[SE19] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Neural Information Processing Systems*, 2019.

[SME21] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[SVMK23] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023.

[TWN+23] Brian L. Trippe, Luhuan Wu, Christian A. Naesseth, David Blei, and John Patrick Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.

[TYT+23] Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi S. Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *The Eleventh International Conference on Learning Representations*, 2023.

A Lower Bound instance

We first define our Lower Bound Distribution g (up to scaling). Let $w_\sigma(x)$ denote the density of a Gaussian with mean zero and standard deviation σ, and let $\mathrm{comb}_\varepsilon$ denote the Dirac Comb distribution with period ε, given by

$$\mathrm{comb}_\varepsilon(x) = \sum_{k=-\infty}^{\infty} \delta(x - k\varepsilon)$$

For any $b \in \{-1, 1\}$, let ψ_b be the density of a standard Gaussian discretized to multiples of ε, with phase either 0 or $\frac{\varepsilon}{2}$ depending on b:

$$\psi_b(x) \propto w_1(x) \cdot \mathrm{comb}_\varepsilon\left(x - \varepsilon/2 \cdot \frac{1-b}{2}\right).$$

Definition 3.1 (Unscaled Lower Bound Distribution)**.** *Let $f : \{\pm 1\}^d \to \{\pm 1\}^{d'}$ be a given function. For $R > 0$ and for any $s \in \{\pm 1\}^d$, define the product distribution g_s over $x \in \mathbb{R}^{d+d'}$ such that*

$$\begin{aligned} x_i &\sim w_1(x_i - R \cdot s_i) && \text{for } i \le d \\ x_i &\sim \psi_{f(s)_{i-d}} && \text{for } i > d. \end{aligned}$$

The unconditional distribution g we consider is the uniform mixture of g_s over $s \in \{\pm 1\}^d$.

We define our final Lower Bound distribution below, which is a scaled version of g.

Definition 3.2 (Scaled Lower Bound Distribution)**.** *Let $\widetilde{g}(x) = R^{d+d'} g(R \cdot x)$ be the scaled version of the distribution with density g defined in Definition 3.1. Similarly, let $\widetilde{g}_s = R^{d+d'} g_s(R \cdot x)$.*

B Lower Bound – Posterior Sampling implies Inversion of One-Way Function

B.1 Notation

Let $l \coloneqq [d] = \{1, 2, 3, \ldots, d\}$, and let $r \coloneqq \{d+1, d+2, \ldots, d+d'\}$, so that for any $x \in \mathbb{R}^{d+d'}$, $x_{[:d]} \in \mathbb{R}^d$ is a vector containing the first d entries of x, and $x_{[-d':]} \in \mathbb{R}^{d'}$ is a vector containing the last d' entries of x.

Let $\mathrm{Round}_R : \mathbb{R}^k \to \{\pm R\}^k$ be such that for all $i \in [k]$,

$$\mathrm{Round}_R(x)_i = \arg\min_{v \in \{\pm R\}} |x_i - v| .$$

Let $\mathrm{parity} : \mathbb{Z} \to \{-1, +1\}$ be such that $\mathrm{parity}(2i) = -1, \mathrm{parity}(2i+1) = 1$ for all $i \in Z$. Let $\mathrm{Bits}_\varepsilon : \mathbb{R}^k \to \{\pm 1\}^k$ be such that for all $i \in [k]$,

$$(\mathrm{Bits}_\varepsilon(y))_i = \mathrm{parity}\left(\arg\min_{i \in \mathbb{Z}} \left| i \cdot \frac{\varepsilon}{2} - y_i \right| \right)$$

This function takes a value y and returns a guess for whether y comes from a smoothed distribution discretized to even multiples of $\varepsilon/2$ or odd multiples of $\varepsilon/2$, based on which is closer.

Definition B.1 (Conditional Distribution)**.** *Let g be the distribution defined in 3.1, parameterized by a function f, and real values $R, \varepsilon > 0$. For some noise pdf h, we define $\mathcal{X}^h_{f,R,\varepsilon}$ to be the distribution over (x, y) where $x \sim g$ and $y \sim x_{[-d':]} + h$.*

We also explicitly define the two noise models we will be using for the lower bound: we take

$$\mathcal{X}^\beta_{f,R,\varepsilon} \coloneqq \mathcal{X}^{w_\beta}_{f,R,\varepsilon}, \qquad w_\beta = N(0, \beta^2). \tag{1}$$

Let $(\mathcal{X}^\beta_{f,R,\varepsilon})_y$ denote the marginal over y. Further, $\mathcal{X}^{\beta,\beta_{\max}}_{f,R,\varepsilon} \coloneqq \mathcal{X}^b_{f,R,\varepsilon}$ where b is a clipped normal distribution: $b \coloneqq \mathrm{clip}(\beta_{\max}, N(0, \beta^2))$.

B.2 Inverting f via Posterior Sampling

Lemma B.2. *Let $\beta_{max} \le \varepsilon/4$ and $\sqrt{32 \log \frac{d}{\delta}} \le R$. Then,*

$$\mathbb{P}_{x^b, y^b \sim \mathcal{X}_{f,R,\varepsilon}^{\beta, \beta_{max}}} \left[f(Round_R(x^b_{[:d]})) = Bits_\varepsilon(y^b) \right] \ge 1 - \delta$$

Proof. Let $x^b, y^b \sim \mathcal{X}_{f,R,\varepsilon}^{\beta,\beta_{\max}}$. By definition, we know that $y_b \sim x^b_{[-d':]} + \mathrm{clip}(\beta_{\max}, N(0, \beta^2)$. Further, for all indices i, $(x^b_{[-d':]})_i = j\varepsilon/2$ for some integer j. So, if $\beta_{\max} \le \varepsilon/4$, then

$$\mathrm{Bits}_\varepsilon(y^b) = \mathrm{Bits}_\varepsilon(x^b_{[-d':]}). \tag{2}$$

We know that x^b is drawn from a uniform mixture over $g_s(x)$, as defined in 3.1. So, fixing an $s \in \{\pm 1\}^d$. We have that

$$\mathrm{Bits}_\varepsilon(x^b_{[-d':]}) = s. \tag{3}$$

On the other hand, $x_{[:d]}$ is a product of gaussians centered at Rs_i in the ith coordinate. Therefore, for all $i < d$,

$$\mathbb{P}_{x^b} \left[\left| (x^b_{[:d]})_i - Rs_i \right| \le \sqrt{2 \log \frac{d}{\delta}} \right] \ge 1 - \frac{\delta}{d}$$

Since $\sqrt{2 \log \frac{d}{\delta}} \le R/4$, we get that

$$\mathbb{P}_{x^b} \left[\mathrm{Round}_R(x^b_{[:d]}) = s \right] \ge 1 - \delta. \tag{4}$$

Putting together eq. (2), eq. (3), and eq. (4) we get

$$\mathbb{P}_{x^b} \left[\mathrm{Round}_R(x^b_{[:d]}) = \mathrm{Bits}_\varepsilon(y^b) \right] \ge 1 - \delta$$

□

Lemma B.3. *Let C be a (τ, δ)-conditional sampling algorithm for $\mathcal{X}_{f,R,\varepsilon}^{\beta}$. If $\varepsilon \ge \beta \sqrt{32 \log \frac{d}{\delta}}$, $\tau \le R/4$, and $32 \log \frac{d}{\delta} \le R^2$, then for $y \sim (\mathcal{X}_{f,R,\varepsilon}^{\beta})_y$ and $\widehat{x} \sim \mathsf{C}(y)$,*

$$\mathbb{P}[f(Round_R(\widehat{x}_{[:d]})) \ne Bits_\varepsilon(y)] \le 5\delta.$$

Proof. Let $\mathcal{X}_{f,R,\varepsilon}^{\beta}$ have pdf p^β. Assume we have a (τ, δ)-posterior sampler over $\mathcal{X}_{f,R,\varepsilon}^{\beta}$ that outputs sample from distribution $\widehat{\mathcal{X}}$ with distribution $\widehat{p}$. This means that with probability $1 - \delta$ over y, there exists a coupling $\mathcal{P}$ over $(x, \widehat{x})$ such that $(x, \widehat{x})$ are (τ, δ)-close. Therefore, there exists a distribution $\mathcal{P}$ over $(x, \widehat{x}, y) \in \mathbb{R}^{d+d'} \times \mathbb{R}^{d+d'} \times \mathbb{R}^{d'}$ with density $p^{\mathcal{P}}$ such that $p^{\mathcal{P}}(x, y) = p^\beta(x, y)$, $p^{\mathcal{P}}(\widehat{x} \mid y) = \widehat{p}(\widehat{x} \mid y)$, and

$$\mathbb{P}_{x, \widehat{x} \sim \mathcal{P}} \left[\|x - \widehat{x}\|_2 \le \tau \right] \ge 1 - 2\delta.$$

Now, let $\mathcal{X}_{f,R,\varepsilon}^{\beta,\beta_{\max}}$ have pdf $p^{\beta,\beta_{\max}}$, with $\beta_{\max} = \beta \sqrt{2 \log \frac{1}{\delta}}$. We have

$$TV(\mathcal{X}_{f,R,\varepsilon}^{\beta}, \mathcal{X}_{f,R,\varepsilon}^{\beta,\beta_{\max}}) \lesssim e^{-\beta_{\max}^2 / 2\beta^2} \le \delta$$

Therefore, building on $\mathcal{P}$, we can construct a new distribution $\mathcal{P}'$ over $(x, \widehat{x}, x^b, y, y^b) \in \mathbb{R}^{d+d'} \times \mathbb{R}^{d+d'} \times \mathbb{R}^{d+d'} \times \mathbb{R}^{d'} \times \mathbb{R}^{d'}$ with density $p^{\mathcal{P}'}$ such that $p^{\mathcal{P}'}(x, y) = p^{\beta}(x, y)$, $p^{\mathcal{P}'}(\widehat{x} \mid y) = \widehat{p}(\widehat{x} \mid y)$, $p^{\mathcal{P}'}(x^b, y^b) = p^{\beta, \beta_{\max}}(x^b, y^b)$, $(x, y) = (x^b, y^b)$ with probability $1 - \delta$, and

$$\mathbb{P}_{x, \widehat{x} \sim \mathcal{P}'} \left[\|x - \widehat{x}\|_2 \le \tau \right] \ge 1 - 2\delta$$

Therefore, under this distribution,

$$\mathbb{P}_{\widehat{x}, x^b \sim \mathcal{P}'} \left[\|\widehat{x} - x^b\|_2 \le \tau \right] \ge 1 - 3\delta$$

In particular, we apply the fact that $\|\widehat{x}_{[:d]} - x^b_{[:d]}\|_\infty \le \|\widehat{x} - x^b\|_\infty \le \|\widehat{x} - x^b\|_2$ to get

$$\mathbb{P}_{\widehat{x}, x^b \sim \mathcal{P}'} \left[\|\widehat{x}_{[:d]} - x^b_{[:d]}\|_\infty \le \tau \right] \ge 1 - 3\delta. \tag{5}$$

Now, by the definition of $\mathcal{X}^{\beta, \beta_{\max}}_{f, R, \varepsilon}$, for all $i < d$, x^b_i is a mixture of variance 1 normal distributions centered at $\pm R$. So, for any $i < d$,

$$\mathbb{P}_{x^b} \left[\left| x^b_i - \mathrm{Round}_R(x^b_i) \right| \ge \sqrt{2 \log \frac{d}{\delta}} \right] \le \frac{\delta}{d}$$

Applying a union bound over $i \in [d]$ and putting this together with eq. (5),

$$\mathbb{P}_{\widehat{x}, x^b \sim P'} \left[\|\widehat{x}_{[:d]} - \mathrm{Round}_R(x^b_{[:d]})\|_\infty \le \sqrt{2 \log \frac{1}{\delta}} + \tau \right] \ge 1 - 4\delta$$

So, since $\sqrt{2 \log \frac{d}{\delta}} + \tau \le \frac{R}{4} + \frac{R}{4} = \frac{R}{2}$, and $\mathrm{Round}_R((x^b_{[:d]})_i) \in \pm R$, we have

$$\mathbb{P}_{\widehat{x}, x^b \sim P'} \left[\|\mathrm{Round}_R(\widehat{x}_{[:d]}) - \mathrm{Round}_R(x^b_{[:d]})\|_\infty \le \sqrt{2 \log \frac{d}{\delta}} + \tau \right] \le 1 - 3\delta$$

Again, the output of Round_R is always $\pm R$, so this means

$$\mathbb{P}_{\widehat{x}, x^b \sim P'} \left[\mathrm{Round}_R(\widehat{x}_{[:d]}) = \mathrm{Round}_R(x^b_{[:d]}) \right] \ge 1 - 3\delta$$

Now, by Lemma B.2, since $\beta_{\max} < \varepsilon/4$ and $R \ge \sqrt{32 \log \frac{d}{\delta}}$, we have

$$\mathbb{P}_{x^b, y^b \sim P'} \left[f(\mathrm{Round}_R(x^b_{[:d]})) = \mathrm{Bits}_\varepsilon(y^b) \right] \ge 1 - \delta$$

Therefore,

$$\mathbb{P}_{\widehat{x}, y^b \sim P'} \left[f(\mathrm{Round}_R(\widehat{x}_{[:d]})) = \mathrm{Bits}_\varepsilon(y^b) \right] \ge 1 - 4\delta$$

Finally, we know that $y = y^b$ with probability $1 - \delta$. Therefore, we get

$$\mathbb{P}_{\widehat{x}, y \sim P'} \left[f(\mathrm{Round}_R(\widehat{x}_{[:d]})) = \mathrm{Bits}_\varepsilon(y) \right] \ge 1 - 5\delta$$

□

Theorem B.4. *For any function f, let C be a $(R/4, \delta)$-posterior sampler (1.5) for $\mathcal{X}^{\beta}_{f,R,\varepsilon}$, as defined in (1), with $\varepsilon \geq \beta\sqrt{32 \log \frac{d}{\delta}}$, and $R \geq \sqrt{32 \log \frac{d}{\delta}}$, that takes time T to run. Then, there exists an algorithm A that runs in time $T + O(d)$ such that*

$$\mathbb{P}_{s,\mathsf{A}}[f(\mathsf{A}(f(s))) \neq f(s)] \leq 6\delta$$

Proof. Sample $y \sim h_{f(r)}$, where

$$h_s(y) = \begin{cases} (w_1(y) \cdot \mathrm{comb}_\varepsilon(y))) * N(0, \beta^2), & s_i = 1 \\ \left(w_1(y) \cdot \mathrm{comb}_\varepsilon\left(y - \frac{\varepsilon}{2}\right)\right) * N(0, \beta^2) & s_i = -1 \end{cases}$$

Now, since $\beta \leq \frac{\varepsilon}{\sqrt{32 \log \frac{d}{\delta}}}$, each coordinate of the noise, drawn from $N(0, \beta^2)$, is less than $\varepsilon/4$ with probability $1 - \delta/d$. Therefore,

$$\mathbb{P}\left[\mathrm{Bits}_\varepsilon(y) = f(r)\right] \geq 1 - \delta$$

By definition, h_s is the same as the density of $(\mathcal{X}^{\beta}_{f,R,\varepsilon})_y$. So, by Lemma B.3, since $R \geq \tau/4, R \geq \sqrt{32 \log \frac{d}{\delta}}$, and we take $\hat{x} \sim \mathsf{C}(y)$, we have

$$\mathbb{P}_{\widehat{x},y}\left[f(\mathrm{Round}_R(\widehat{x}_{[:d]})) \neq \mathrm{Bits}_\varepsilon(y)\right] \leq 5\delta$$

Therefore,

$$\mathbb{P}_{\widehat{x},y}\left[f(\mathrm{Round}_R(\widehat{x}_{[:d]})) \neq f(r)\right] \leq 6\delta$$

So, our algorithm A can output $\mathrm{Round}_R(\widehat{x}_l)$. All we had to do to run this algorithm was to sample d normal random variables, and then run our posterior sampler. This takes $T + O(d)$ time. □

Lemma 3.3. *For any function f, suppose C is an $(1/10, 1/10)$-posterior sampler in the linear measurement model with noise parameter β for distribution with density $\widetilde{g}$ as defined in Definition 3.2, with $\varepsilon \geq \beta\sqrt{32 \log d}$ and $R \geq 32\sqrt{\log d}$. If C takes time T to run, then there exists an algorithm A that runs in time $T + O(d)$ such that*

$$\mathbb{P}_{s,\mathsf{A}}[f(\mathsf{A}(f(s))) \neq f(s)] \leq \frac{3}{4}$$

Proof. This follows from Theorem B.4, using the fact that after rescaling down by R, $\mathcal{X}^{\beta}_{f,R,\varepsilon}$ as a distribution over (x, y) is the same distribution as $x \sim \widetilde{g}$, with $y = Ax + N(0, \beta^2)$. □

B.3 Inverting a One-Way function via Posterior Sampling

Lemma 3.4. *Suppose $d^{0.01} \leq m \leq d/2$ and one-way functions exist. Then, for $\widetilde{g}$ as defined in Definition 3.2 with $\varepsilon = \frac{1}{C\sqrt{\log d}}$ and $R = C \log d$, and linear measurement model with noise parameter $\beta = \frac{1}{C^2 \log^2 d}$ and measurement matrix $A \in \mathbb{R}^{m \times d}$, $(\frac{1}{10}, \frac{1}{10})$-posterior sampling takes superpolynomial time.*

Proof. Since $d^{0.01} < m < d/2$, d and m are only polynomially separated. So, by G.2, we can construct a one-way function $f : \{\pm 1\}^{d-m} \to \{\pm 1\}^m$. By definition, we can see that $\widetilde{g}$, with measurement noise β is the same distribution as $\mathcal{X}^{\beta R}_{f,R,\varepsilon}$, scaled down by R. Therefore, by Theorem B.4, since $R \geq 32\sqrt{\log \frac{d}{\delta}}$, $\varepsilon \geq \beta R\sqrt{\log \frac{d}{\delta}}$, if we can run a posterior sampler in time T, we can invert f with probability $1 - 6\delta$ in time $T + O(m)$. So, if f takes time superpolynomial in m to invert, then $T + O(m)$ is superpolynomial. Since $m > d^{0.01}$, this means that T itself is superpolynomial. □

Lemma B.5. *For any $m \le d/2$, suppose that there exists a one-way function $f : \{\pm 1\}^m \to \{\pm 1\}^m$ that requires $2^{\Omega(m)}$ time to invert. Then, for $\widetilde{g}$ as defined in Definition 3.2 with $\varepsilon = \frac{1}{C\sqrt{\log d}}$ and $R = C \log d$, and linear measurement model with noise parameter $\beta = \frac{1}{C^2 \log^2 d}$ and measurement matrix $A \in \mathbb{R}^{m \times d}$, $(\frac{1}{10}, \frac{1}{10})$-conditional sampling takes at least $2^{\Omega(m)}$ time.*

Proof. By definition, we can see that $\widetilde{g}$, with measurement noise β is the same distribution as $\mathcal{X}^{\beta R}_{f,R,\varepsilon}$, scaled down by R. Therefore, by Theorem B.4, since $R \ge 32\sqrt{\log d}$, $\varepsilon \ge \beta R \sqrt{\log d}$, if we can run a posterior sampler in time T, we can invert f with probability 0.4 in time $T + O(m)$. So, if f takes at least time $2^{\Omega(m)}$ to run, then we must have $T + O(m) \ge 2^{\Omega(m)}$, which means $T \ge 2^{\Omega(m)}$. □

C Lower Bound – ReLU Approximation of Score

C.1 Piecewise Linear Approximation of σ-smoothed score in One Dimension

In this section, we analyze the error of a piecewise linear approximation to a smoothed score. We first show that for one dimensional distributions, we can get good approximations, and later extend it to product distributions in higher dimensions.

First, we show that a piecewise linear approximation that discretizes the space into intervals of width γ has low error.

Lemma C.1. *Let p be a distribution over $\mathbb{R}$, and let $p_\sigma = p * N(0, \sigma^2)$ have score s_σ. Let $\gamma \le \sigma$, and let $S_i = [i\gamma, (i+1)\gamma)$ for all $i \in \mathbb{Z}$. Define a piecewise linear function $f : \mathbb{R} \to \mathbb{R}$ so that: for all x, if i is such that $S_i \ni x$, then*

$$f(x) = \frac{((i+1)\gamma - x) \cdot s(i\gamma) + (x - i\gamma) \cdot s((i+1)\gamma)}{\gamma}.$$

Then f is continuous and satisfies

$$\mathbb{E}\left[(s(x) - f(x))^2\right] \lesssim \frac{\gamma^2}{\sigma^4}$$

Proof. Define the left and right piecewise constant approximations $l(x) = s(i\gamma), r(x) = s((i+1)\gamma)$ for all $x \in S_i$.

We know that for any $y \in S_i$, there is some $y' \in [i\gamma, y]$ such that $s(y) = s(i\gamma) + (y - i\gamma)s'(y')$. So, we get

$$\forall y \in S_i,\ s(y) \le s(i\gamma) + \gamma \sup_{z \in S_i} s'(z) \le s(i\gamma) + \gamma \sup_{|c| \le \gamma} s'(y + c).$$

Therefore,

$$\mathop{\mathbb{E}}_{x \sim p}\left[(s_\sigma(x) - l(x))^2\right] \le \gamma^2 \mathop{\mathbb{E}}_{x \sim p}[\sup_{|c| \le \gamma} s'(y + c)^2] \lesssim \frac{\gamma^2}{\sigma^4}$$

By Lemma H.1. The same holds for $r(x)$. Now, recall that f satisfies

$$\forall i \in \mathbb{Z},\ \forall x \in S_i,\ f(x) = \frac{(i+1)\gamma - x}{\gamma} \cdot s(i\gamma) + \frac{x - i\gamma}{\gamma} \cdot s((i+1)\gamma).$$

The coefficients $\frac{(i+1)\gamma - x}{\gamma}$ and $\frac{x - i\gamma}{\gamma}$ sum to 1 and are within the interval $[0, 1]$. So, at each point, f is just a convex combination of the two approximations l and r. Therefore, by convexity, for any S_i, if $x \in S_i$,

$$\mathop{\mathbb{E}}_{x \in S_i}[(s_\sigma(x) - f(x))^2] \le \mathop{\mathbb{E}}_{x \in S_i}[(s_\sigma(x) - l(x))^2] + \mathop{\mathbb{E}}_{x \in S_i}[(s_\sigma(x) - r(x))^2] \tag{6}$$

This immediately gives us that

$$\mathbb{E}[(s_\sigma(x) - f(x))^2] \le \mathbb{E}[(s_\sigma(x) - l(x))^2] + \mathbb{E}[(s_\sigma(x) - r(x))^2] \lesssim \frac{\varepsilon^2}{\sigma^4}$$

Within each interval, the function is linear and so it is continous. We just need to check continuity at the endpoints. However, we can see that for any $i \in \mathbb{Z}$, $\lim_{x \to i\gamma^-} = \lim_{x \to i\gamma^+} = s(i\gamma)$, and so we also have continuity. □

Unfortunately, the above approximation has an infinite number of pieces. To handle this, we show that in regions far away from the mean, a zero-approximation is good enough, given that the distribution has bounded second moment m_2.

Lemma C.2. *Let p be some distribution over $\mathbb{R}$ with mean μ, and let $p_\sigma = p * N(0, \sigma^2)$ have score s_σ. Let $m_2^2 \coloneqq \mathbb{E}_{x \sim p}\left[(x-\mu)^2\right]$ be the second moment of p_σ. Further, let $|\varphi| \le \frac{1}{\sigma} \log \frac{1}{\delta}$ be some constant. Then,*

$$\mathbb{E}\left[(s_\sigma(x) - \varphi)^2 \cdot \mathbb{1}_{|x-\mu| > \frac{m_2}{\sqrt{\delta}}}\right] \lesssim \frac{\sqrt{\delta}}{\sigma^2}$$

Proof. We have

$$\mathbb{E}\left[(s_\sigma(x) - \varphi)^2 \cdot \mathbb{1}_{|x-\mu| > \frac{m_2}{\sqrt{\delta}}}\right] \lesssim \mathbb{E}\left[s_\sigma(x)^2 \cdot \mathbb{1}_{|x-\mu| > \frac{m_2}{\sqrt{\delta}}}\right] + \mathbb{E}\left[\varphi^2 \cdot \mathbb{1}_{|x-\mu| > \frac{m_2}{\sqrt{\delta}}}\right]$$

First, by Chebyshev's inequality, we know that

$$\mathbb{P}\left[|x - \mu| \ge \frac{m_2}{\delta}\right] \le \delta$$

Now, we use Cauchy Schwarz to bound the first term:

$$\begin{aligned}
\mathbb{E}\left[s_\sigma(x)^2 \cdot 1_{|x-\mu| > \frac{m_2}{\sqrt{\delta}}}\right] &\le \sqrt{\mathbb{E}\left[s_\sigma(x)^4\right] \mathbb{E}\left[1_{|x-\mu| > \frac{m_2}{\sqrt{\delta}}}\right]} \\
&= \sqrt{\mathbb{E}\left[s_\sigma(x)^4\right] \mathbb{P}\left[|x-\mu| \ge \frac{m_2}{\sqrt{\delta}}\right]} \\
&= \sqrt{\mathbb{E}\left[s_\sigma(x)^4\right] \cdot \delta} \lesssim \sqrt{\delta/\sigma^4} = \sqrt{\delta}/\sigma^2
\end{aligned}$$

where the last line is by Corollary H.8. Finally, for the second term, we know that

$$\begin{aligned}
\mathbb{E}\left[\varphi^2 \cdot \mathbb{1}_{|x-\mu| > \frac{m_2}{\sqrt{\delta}}}\right] &\le \mathbb{E}\left[\frac{1}{\sigma^2} \log^2 \frac{1}{\delta} \cdot \mathbb{1}_{|x-\mu| > \frac{m_2}{\sqrt{\delta}}}\right] \\
&= \frac{1}{\sigma^2} \log^2 \frac{1}{\delta} \mathbb{P}\left[|x-\mu| > \frac{m_2}{\sqrt{\delta}}\right] \\
&= \frac{\delta}{\sigma^2} \log^2 \frac{1}{\delta} \lesssim \frac{\sqrt{\delta}}{\sigma^2}
\end{aligned}$$

The last line here uses the fact that for all x, $x \log^2(1/x) \le 3\sqrt{x}$. Summing the two terms gives the desired result. □

Then, we show that neighborhoods where the magnitude of the score can be large are rare and can also be approximated by the zero function. This allows us to control the slope of the piecewise linear approximation in each piece.

Lemma C.3. *Let p be a distribution over $\mathbb{R}$. Let $p_\sigma = p * N(0, \sigma^2)$ have score s_σ. Let $\gamma \le \frac{\sigma}{2}$, and let $m(x) = \sup_{y \in [x-\gamma, x+\gamma]} s(x)$. Then,*

$$\mathbb{E}\left[s(x)^2 \cdot \mathbb{1}_{m(x) > \frac{\log \frac{1}{\delta}}{\sigma}}\right] \lesssim \frac{\sqrt{\delta}}{\sigma^2}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[m(x)^2 \cdot \mathbb{1}_{m(x) > \frac{\log \frac{1}{\delta}}{\sigma}}\right] &\le \sqrt{\mathbb{E}\left[m(x)^4\right] \cdot \mathbb{E}\left[\mathbb{1}_{m(x) > \frac{\log \frac{1}{\delta}}{\sigma}}\right]} && \text{by Cauchy-Schwarz} \\
&\le \sqrt{\mathbb{E}\left[\left(\sup_{y \in [x-\gamma, x+\gamma]} s(x)\right)^4\right] \cdot \mathbb{P}\left[m(x) > \frac{\log \frac{1}{\delta}}{\sigma}\right]} \\
&\lesssim \sqrt{\frac{1}{\sigma^4} \cdot \mathbb{P}\left[m(x) > \frac{\log \frac{1}{\delta}}{\sigma}\right]} && \text{by Lemma H.8} \\
&\le \frac{\sqrt{\delta}}{\sigma^2} && \text{by Lemma H.2}
\end{aligned}$$

□

We put these lemmas together to show that a piecewise linear function with a bounded number of pieces and bounded slope in each piece is a good approximation to the smoothed score.

Lemma C.4. *Let p be a distribution over $\mathbb{R}$ with mean μ, and let $p_\sigma = p * N(0, \sigma^2)$ have score s_σ and second moment m_2^2. Then, for any $\alpha \le 1/4$ there exists a function $l : \mathbb{R} \to \mathbb{R}$ that satisfies*

1. *l is piecewise linear with at most $\Theta(\frac{m_2}{\sigma \kappa^{3/2}})$ pieces,*
2. *if x is a transition point between two pieces, then $|x - \mu| \le \frac{m_2}{\kappa}$*
3. *the slope of each piece is bounded by $\Theta\left(\frac{\log \frac{1}{\kappa}}{\sigma^2 \sqrt{\kappa}}\right)$,*
4. *$|l| \lesssim \frac{1}{\sigma} \log \frac{1}{\kappa}$*
5. $$\mathop{\mathbb{E}}_{x \sim p}[(l(x) - s(x))^2] \lesssim \frac{\kappa}{\sigma^2}$$

Proof. First, we partition the real line into $S_i = [i\gamma, (i+1)\gamma)$ for all $i \in \mathbb{Z}$, where $\gamma < \sigma/2$. Define the function $l_1 : \mathbb{R} \to \mathbb{R}$ so that if $S_i \ni x$, then

$$l_1(x) = \frac{((i+1)\gamma - x)s(i\gamma) + (x - i\gamma)s((i+1)\gamma)}{\gamma}. \tag{7}$$

As in Lemma C.1, this is the linear interpolation between $s(i\gamma)$ and $s((i+1)\gamma)$ on the interval $[i\gamma, (i+1)\gamma)$. By Lemma C.1, when $\gamma < \sigma/2$, we have

$$\mathbb{E}\left[(s(x) - l_1(x))^2\right] \lesssim \frac{\gamma^2}{\sigma^4}$$

Now, we define $l_2 : \mathbb{R} \to \mathbb{R}$. This function uses the piecewise linear l_1 to create a linear approximation that has small slopes on all of the pieces. Define first a set of "good" sets

$$G = \left\{ S_i : \sup_{y \in S_i} s(x) \leq \frac{1}{\sigma} \log \frac{1}{\delta} \right\}.$$

These are the intervals on which the score is always bounded. Further, define two helper maps $L(x)$ and $U(x)$:

$$L(x) = \text{the largest } i \text{ such that } i\gamma < x, S_{i-1} \in G$$
$$R(x) = \text{the smallest } i \text{ such that } i\gamma \geq x, S_i \in G$$

These represent the nearest endpoint of a "good" interval to the left and right, respectively. We then interpolate linearly between $s(\gamma L(x))$ and $s(\gamma R(x))$ to evaluate $l_2(x)$. That is,

$$l_2(x) = \frac{(\gamma R(x) - x)s(\gamma L(x)) + (x - \gamma L(x))s(\gamma R(x))}{\gamma(R(x) - L(x))} \tag{8}$$

Note that by assumption, we have that $|s(\gamma R(x))|, |s(\gamma L(x))| \leq \frac{1}{\sigma} \log \frac{1}{\delta}$, and so $|l_2(x)| \leq \frac{1}{\sigma} \log \frac{1}{\delta}$. We now analyze the error of l_2 against s. First, we note that on the sets outside G, the error is bounded, using Lemma C.3:

$$\begin{aligned}
\sum_{S_i \notin G} \mathbb{E}\left[(s(x) - l_2(x))^2 \mathbb{1}_{x \in S_i}\right] &\leq 2 \sum_{S_i \notin G} \left(\mathbb{E}\left[s(x)^2 \mathbb{1}_{x \in S_i}\right] + \mathbb{E}\left[l_2(x)^2 \mathbb{1}_{x \in S_i}\right]\right) \\
&\lesssim \frac{\sqrt{\delta}}{\sigma^2} + \sum_{S_i \notin G} \mathbb{E}\left[\frac{1}{\sigma^2} \log^2 \frac{1}{\delta} \mathbb{1}_{x \in S_i}\right] && \text{by Lemma C.3} \\
&= \frac{\sqrt{\delta}}{\sigma^2} + \frac{1}{\sigma^2} \log^2 \frac{1}{\delta} \mathbb{P}\left[x \notin G\right] \\
&\leq \frac{\sqrt{\delta}}{\sigma^2} + \frac{1}{\sigma^2} \log^2 \frac{1}{\delta} \mathbb{P}\left[\sup_{y \in [x-\gamma, x+\gamma]} s(x) \geq \frac{1}{\sigma} \log \frac{1}{\delta}\right] \\
&\leq \frac{\sqrt{\delta}}{\sigma^2} + \frac{\delta}{\sigma^2} \log^2 \frac{1}{\delta} && \text{by Lemma H.2}
\end{aligned}$$

Further, if x is in a "good" interval, then $L(x), R(x)$ are simply the left and right endpoints of the interval that x is in. This means that $l_2(x) = l_1(x)$. So,

$$\begin{aligned}
\sum_{S_i \in G} \mathbb{E}\left[(s(x) - l_2(x))^2 \mathbb{1}_{x \in S_i}\right] &= \sum_{S_i \in G} \mathbb{E}\left[(s(x) - l_1(x))^2 \mathbb{1}_{x \in S_i}\right] \\
&\leq \sum_i \mathbb{E}\left[(s(x) - l_1(x))^2 \mathbb{1}_{x \in S_i}\right] \lesssim \frac{\sqrt{\delta}}{\sigma^2}
\end{aligned}$$

Putting these two together, we get that

$$\mathbb{E}\left[(l_2(x) - s(x))^2\right] \lesssim \frac{\sqrt{\delta}}{\sigma^2} + \frac{\gamma^2}{\sigma^4} + \frac{\delta}{\sigma^2} \log^2 \frac{1}{\delta}$$

Now, define $l_3 : \mathbb{R} \to \mathbb{R}$ as follows:

$$l_3(x) = \begin{cases} l_2(x) & |x - \mu| \leq \frac{m_2}{\sqrt{\delta}} \\ l_2\left(\mu - \frac{m_2}{\sqrt{\delta}}\right) & x < \mu - \frac{m_2}{\sqrt{\delta}} \\ l_2\left(\mu + \frac{m_2}{\sqrt{\delta}}\right) & x > \mu - \frac{m_2}{\sqrt{\delta}} \end{cases} \tag{9}$$

This takes our previous approximation l_2 and holds it constant on values of x far away from the mean.

Let B be the integers i such that $x \in S_i \implies |x-\mu| \geq m_2/\sqrt{\delta}$. In other words, the set B enumerates the intervals on which $l_2 \neq l_1$, and equivalently, $l_2 = 0$. Note that since $|l_3(x)| \leq \frac{1}{\sigma}\log\frac{1}{\delta}$, we have in particular that $\left|l_3\left(\mu \pm \frac{m_2}{\sqrt{\delta}}\right)\right| \leq \frac{1}{\sigma}\log\frac{1}{\delta}$. Therefore, for some $|\varphi| \leq \frac{1}{\sigma}\log\frac{1}{\delta}$, we have

$$\begin{aligned}
\mathbb{E}\left[(s(x)-l_3(x))^2\right] &= \sum_i \mathbb{E}\left[(s(x)-l_3(x))^2 \mathbb{1}_{x\in S_i}\right] \\
&= \sum_{i\in B} \mathbb{E}\left[(s(x)-l_3(x))^2 \mathbb{1}_{x\in S_i}\right] + \sum_{i\notin B} \mathbb{E}\left[(s(x)-l_3(x))^2 \mathbb{1}_{x\in S_i}\right] \\
&= \sum_{i\in B} \mathbb{E}\left[(s(x)-\varphi)^2 \mathbb{1}_{|x-\mu|\geq m_2/\sqrt{\delta}}\right] + \sum_i \mathbb{E}\left[(s(x)-l_2(x))^2 \mathbb{1}_{x\in S_i}\right] \\
&\lesssim \frac{\sqrt{\delta}}{\sigma^2} + \frac{\gamma^2}{\sigma^4}
\end{aligned}$$

where this last line uses Lemma C.2.

Finally, each piece of l_3 has slope at most $\Theta\left(\frac{\log\frac{1}{\delta}}{\gamma\sigma}\right)$ since the endpoints of each interval are bounded in magnitude by $\frac{1}{\sigma}\log\frac{1}{\delta}$ and each interval is at least γ in width. Also, we can see that l_3 has at most as many pieces as l_2, which has $\Theta\left(\frac{m_2}{\gamma\sqrt{\delta}}\right)$ pieces, with each endpoint being within $m_2/\sqrt{\delta}$ of the mean.

So, we take l to be l_3 with $\delta = \kappa^2$, and $\gamma = \sigma\sqrt{\kappa}$. Note that when $\kappa < 1/4$, we have $\gamma < \sigma/2$. Plugging these in, and using the fact that $x\log^2(1/x) \leq 3\sqrt{x}$, we get that the number of pieces is $\Theta\left(\frac{m_2}{\sigma\kappa^{3/2}}\right)$, the slope of each piece is bounded by $\Theta\left(\frac{\log\frac{1}{\kappa^2}}{\sigma^2\sqrt{\kappa}}\right)$, the function itself is always bounded by $\frac{1}{\sigma}\log\frac{1}{\kappa^2}$, and $\mathbb{E}_{x\sim p}[\|l(x)-s(x)\|_2^2] \leq \frac{\kappa}{\sigma^2}$. □

Finally, we show that if we have a product distribution over d dimensions, we can simply use the product of the one dimensional linear approximations along each coordinate to give a good approximation for the full score.

Lemma C.5. *Let p be a product distribution over $\mathbb{R}^d$, such that $p(x) = \prod_{i=1}^d p_i(x_i)$. Let $s : \mathbb{R}^d \to \mathbb{R}^d$ be the score of p and let $s_i : \mathbb{R} \to \mathbb{R}$ be the score of p_i. If $l_i : \mathbb{R} \to \mathbb{R}$ is an approximation to s_i such that*

$$\mathop{\mathbb{E}}_{x_i\sim p_i}\left[(l_i(x_i)-s_i(x_i))^2\right] \leq \varepsilon/d,$$

then the function $l : \mathbb{R}^d \to \mathbb{R}^d$ defined as $l(x) = (l_i(x_i))$ satisfies

$$\mathop{\mathbb{E}}_{x\sim p}\left[\|l(x)-s(x)\|_2^2\right] \leq \varepsilon$$

Proof. We have

$$s(x)_i = (\nabla \log p(x))_i = \frac{\partial}{\partial x_i}\log\prod_{i=1}^d p_i(x_i) = \frac{\partial}{\partial x_i}\sum_{i=1}^d \log p_i(x_i) = \frac{\partial}{\partial x_i}\log p_i(x_i) = s_i(x_i).$$

Therefore,

$$\mathop{\mathbb{E}}_{x\sim p}\left[\|l(x)-s(x)\|_2^2\right] = \mathop{\mathbb{E}}_{x\sim p}\left[\sum_{i=1}^d \|l_i(x_i)-s_i(x_i)\|_2^2\right]$$

$$= \sum_{i=1}^{d} \mathop{\mathbb{E}}_{x_i \sim p_i} \left[\|l_i(x_i) - s_i(x_i)\|_2^2 \right] \leq d \cdot \varepsilon/d = \varepsilon$$

□

C.2 Small noise level – Score of vertex distribution close to full score in vertex orthant

Lemma C.6 (Density $g_s(x)$ is close to $g(x)$ for $s \in \{\pm 1\}^d$ closest to x)**.** *Let $d' = O(d)$. Consider g_s and g as in Definition 3.1. We have that for $x \in \{\pm 1\}^d$ such that s is closest to $x_{1,\ldots,d}$ among points in $\{\pm 1\}^d$, for the σ-smoothed versions $h_s = g_s * \mathcal{N}(0, \sigma^2 I_{d+d'})$ of g_s and $h = g * \mathcal{N}(0, \sigma^2 I_{d+d'})$ of g, for $\frac{R^2}{1+\sigma^2} > C \log d$ for sufficiently large constant C,*

$$\left| \frac{1}{2^d} h_s(x) - h(x) \right| \lesssim \frac{1}{2^d} \cdot e^{-\frac{R^2}{4(1+\sigma^2)}}$$

Proof. We have that there are $\binom{d}{k}$ vectors $z \in \{\pm 1\}^d$ such that $\|R \cdot z - x_{1,\ldots,d}\|^2 \geq kR^2$. For such a z,

$$h_z(y) \lesssim e^{-\frac{kR^2}{2(1+\sigma^2)}}$$

So,

$$\left| \frac{1}{2^d} h_s(x) - h(x) \right| = \left| \frac{1}{2^d} \sum_{r \neq s} h_r(x) \right| \lesssim \left| \frac{1}{2^d} \sum_{k=1}^{d} d^k e^{\frac{-kR^2}{2(1+\sigma^2)}} \right| \lesssim \frac{1}{2^d} \cdot e^{-\frac{R^2}{4(1+\sigma^2)}}$$

since $\frac{R^2}{1+\sigma^2} > C \log d$. □

Lemma C.7 (Gradient of density $g_x(y)$ is close to $g(y)$ for $x \in \{\pm 1\}^d$ closest to y)**.** *Let $d' = O(d)$ and consider g_s and g as in Definition 3.1, and $x \in \mathbb{R}^{d+d'}$. We have that for $s \in \{\pm 1\}^d$ such that s is closest to $x_{1,\ldots,d}$ among points in $\{\pm 1\}^d$, for $\sigma \geq \tau$, $\tau = \frac{1}{d^C}$ and $\varepsilon > \frac{1}{\mathrm{poly}(d)}$, for the σ-smoothed versions $h_s = g_s * \mathcal{N}(0, \sigma^2 I_{d+d'})$ of g_s and $h = g * \mathcal{N}(0, \sigma^2 I_{d+d'})$ of g, for $\frac{R^2}{1+\sigma^2} > C \log d$ for sufficiently large constant C,*

$$\left\| \frac{1}{2^d} \nabla h_s(x) - \nabla h(x) \right\|^2 \lesssim \frac{1}{2^d} \cdot e^{-\frac{R^2}{16(1+\sigma^2)}}$$

Proof. We will let $\widetilde{h}_{s,i} = \widetilde{g}_{s,i} * \mathcal{N}(0, \sigma^2)$, where $\widetilde{g}_{s,i}$ is defined in Definition 3.1. So, $h_s(x) = \prod_{i=1}^{d+d'} \widetilde{h}_{s,i}(x_i)$. We have that there are $\binom{d}{k}$ vectors $z \in \{\pm 1\}^d$ such that $\|R \cdot z - x_{1,\ldots,d}\|^2 \geq kR^2$. So, for $i \in [d]$, for such a z,

$$|\left(\nabla h_z(x)\right)_i| \lesssim e^{-\frac{kR^2}{4(1+\sigma^2)}}$$

On the other hand, for $i > d$, by Lemma C.16, since $\sigma > \varepsilon^2$ and $\varepsilon > \frac{1}{\mathrm{poly}(d)}$,

$$\left| \widetilde{h}'_{z,i}(x_i) - w'_{\sqrt{\sigma^2+1}}(x_i) \right| \lesssim e^{-\frac{\sigma^2}{2\varepsilon^2(1+\sigma^2)}} + \sum_{j>0} e^{-\frac{j^2\sigma^2}{2\varepsilon^2(1+\sigma^2)} + \log \frac{j}{\varepsilon(1+\sigma^2)}}$$

$$\leq e^{-\frac{\tau^2}{2\varepsilon^2(1+\tau^2)}} + \sum_{j>0} e^{-\frac{j^2\tau^2}{2\varepsilon^2(1+\tau^2)} + \log\frac{j}{\varepsilon(1+\tau^2)}}$$
$$\lesssim \varepsilon\sqrt{1+\frac{1}{\tau^2}}$$

So, we have that for $z \in \{\pm 1\}^d$ such that $\|R\cdot z - x_{1,\dots,d}\|^2 > kR^2$, since $\varepsilon > \frac{1}{\mathrm{poly}(d)}$, $\tau = \frac{1}{\mathrm{poly}(d)}$ and $\frac{R^2}{1+\sigma^2} > C\log d$,

$$|(\nabla h_z(x))_i| \lesssim \varepsilon\sqrt{1+\frac{1}{\tau^2}}\cdot e^{-\frac{kR^2}{2(1+\sigma^2)}} \lesssim e^{-\frac{kR^2}{4(1+\sigma^2)}}$$

So, finally, for such z,

$$\|\nabla h_z(x)\|^2 \lesssim e^{-\frac{kR^2}{8(1+\sigma^2)}}$$

Thus,

$$\left\|\frac{1}{2^d}\nabla h_s(x) - \nabla h(x)\right\|^2 = \left\|\frac{1}{2^d}\sum_{r\neq s}\nabla h_r(x)\right\|^2 \lesssim \frac{1}{2^d}\sum_{k=1}^{d} d^k e^{-\frac{kR^2}{8(1+\sigma^2)}} \lesssim \frac{1}{2^d}\cdot e^{-\frac{R^2}{16(1+\sigma^2)}}$$

□

Lemma C.8 (Score of mixture close to score of closest (discretized) Gaussian)**.** *Let $d = O(d')$, and consider g_s, g as in Definition 3.1 for any $s \in \{\pm 1\}^d$, with $\frac{R^2}{1+\sigma^2} > C\log d$ for sufficiently large constant C. Let $S \subset \mathbb{R}^d$ be the orthant containing s. Let $\sigma \geq \tau$ for $\tau = \frac{1}{\mathrm{poly}(d)}$, and let $\varepsilon > \frac{1}{\mathrm{poly}(d)}$. We have that, for the σ-smoothed scores $s_{\sigma,s}$ of g_s and s_σ of g,*

$$\mathop{\mathbb{E}}_{x\sim h}\left[\|s_{\sigma,s}(x) - s_\sigma(x)\|^2 \middle| 1_{x_{1,\dots,d}\in S}\right] \lesssim e^{-\Omega\left(\frac{R^2}{1+\sigma^2}\right)}$$

where h is the σ-smoothed version of g, given by $h = g * \mathcal{N}(0,\sigma^2 I_{d+d'})$.

Proof. Let h_s be the σ-smoothed version of g_s, given by $h_s = g_s * \mathcal{N}(0,\sigma^2 I_{d+d'})$. Let $\widetilde{s} \in \mathbb{R}^{d+d'}$ be such that the first d coordinates are given by s, and the remaining d' coordinates are 0. We have

$$\mathop{\mathbb{E}}_{x\sim h}\left[\|s_{\sigma,s}(x) - s_\sigma(x)\|^2 \middle| 1_{x_{1,\dots,d}\in S}\right] = \mathop{\mathbb{E}}_{x\sim h}\left[\|s_{\sigma,s}(x) - s_\sigma(x)\|^2 \cdot 1_{\|x-\widetilde{s}\|\leq R/10} \middle| 1_{x_{1,\dots,d}\in S}\right]$$
$$+ \mathop{\mathbb{E}}_{x\sim h}\left[\|s_{\sigma,s}(x) - s_\sigma(x)\|^2 \cdot 1_{\|x-\widetilde{s}\|> R/10} \middle| 1_{x_{1,\dots,d}\in S}\right]$$

Note that when $\|x-\widetilde{s}\| \leq R/10$, by Lemma C.16, $h_s(x) \gtrsim e^{-\frac{R^2}{64(1+\sigma^2)}}$ since $\sigma \geq \tau$ for $\tau = \frac{1}{\mathrm{poly}(d)}$, $\varepsilon > \frac{1}{\mathrm{poly}(d)}$ and $\frac{R^2}{1+\sigma^2} > C\log d$. So, by Lemmas C.6 and C.7, $h(x) = \frac{1}{2^d}h_s(x)\left(1 + O\left(e^{-\frac{R^2}{8(1+\sigma^2)}}\right)\right)$, and $\|\nabla h(x) - \frac{1}{2^d}\nabla h_s(x)\|^2 \lesssim \frac{1}{2^d}e^{-\frac{R^2}{16(1+\sigma^2)}}$. Also note that by Lemma C.6, $h(x|x_{1,\dots,d}\in S) \leq h_s(x) + O(e^{-\frac{R^2}{8(1+\sigma^2)}}) \leq h_s(x)\cdot\left(1 + O\left(e^{-\frac{R^2}{32(1+\sigma^2)}}\right)\right)$. So, for the first term,

$$\mathop{\mathbb{E}}_{x\sim h}\left[\|s_{\sigma,s}(x) - s_\sigma(x)\|^2 \cdot 1_{\|x-\widetilde{s}\|\leq \frac{R}{10}} \middle| 1_{x_{1,\dots,d}\in S}\right]$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{x\sim h}\left[\left\|\frac{\frac{1}{2^d}\nabla h_s(x)}{\frac{1}{2^d}h_s(x)} - \frac{\nabla h(x)}{h(x)}\right\|^2 \cdot 1_{\|x-\widetilde{s}\|\le R/10}\middle|1_{x_{1,\dots,d}\in S}\right]\\
&\lesssim \mathop{\mathbb{E}}_{x\sim h}\left[\frac{\frac{1}{2^d}e^{-\frac{R^2}{16(1+\sigma^2)}} + e^{-\frac{R^2}{8(1+\sigma^2)}}\cdot\frac{1}{2^d}\cdot\|\nabla h_s(x)\|^2}{\frac{1}{2^d}h_s(x)^2}\cdot 1_{\|x-\widetilde{s}\|\le R/10}\middle|1_{x_{1,\dots,d}\in S}\right]\\
&\lesssim e^{-\frac{R^2}{32(1+\sigma^2)}} + e^{-\frac{R^2}{8(1+\sigma^2)}}\cdot\mathop{\mathbb{E}}_{x\sim h_s}\left[\frac{\|\nabla h_s(x)\|^2}{h_s(x)^2}\right]\\
&\lesssim e^{-\frac{R^2}{32(1+\sigma^2)}} + \frac{de^{-\frac{R^2}{8(1+\sigma^2)}}}{\sigma^2}\\
&\lesssim e^{-\frac{R^2}{64(1+\sigma^2)}}
\end{aligned}
$$

since $\sigma \ge \frac{1}{\mathrm{poly}(d)}$, $\varepsilon > \frac{1}{\mathrm{poly}(d)}$ and $\frac{R^2}{1+\sigma^2} > C\log d$.

For the second term, by Cauchy-Schwarz,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim h}\left[\|s_{\sigma,s}(x) - s_\sigma(x)\|^2\cdot 1_{\|x-\widetilde{s}\|>\frac{R}{10}}\middle|1_{x_{1,\dots,d}\in S}\right]\\
&\lesssim \sqrt{\left(\mathop{\mathbb{E}}_{x\sim h}\left[\|s_{\sigma,s}(x)\|^4 + \|s_\sigma(x)\|^4\middle|1_{x_{1,\dots,d}\in S}\right]\right)\cdot\mathbb{E}\left[1_{\cap\|x-\widetilde{s}\|>R/10}\middle|1_{x_{1,\dots,d}\in S}\right]}\\
&\lesssim \sqrt{\frac{R^4}{\sigma^4} + \frac{1}{\sigma^4}\mathbb{E}\left[\|x\|^4\middle|1_{x_{1,\dots,d}\in S}\right]}\cdot e^{-\Omega\left(\frac{R^2}{1+\sigma^2}\right)}\\
&= \frac{1}{\sigma^2}\sqrt{R^4 + \mathop{\mathbb{E}}_{s\sim\{\pm1\}^d}\left[\mathbb{E}\left[\|x\|^4\middle|1_{x_{1,\dots,d}\in S, x\sim g_s}\right]\right]}e^{-\Omega\left(\frac{R^2}{1+\sigma^2}\right)}\\
&\lesssim \frac{R^2}{\sigma^2}\cdot e^{-\Omega\left(\frac{R^2}{1+\sigma^2}\right)}\\
&\lesssim e^{\Omega\left(\frac{R^2}{1+\sigma^2}\right)}
\end{aligned}
$$

So, we have the claim. □

C.3 ReLU Network approximation of σ-smoothed Scores of Product Distributions

Once we have this, we also need to go from being close to mixture of Gaussians to being close to mixture of discretized Gaussians.

Lemma C.9. *Let $f:\mathbb{R}\to\mathbb{R}$ be a continuous piecewise linear function with D segments. Then, f can be represented by a ReLU network with $O(D)$ parameters. If each segment's slope, each transition point, and the values of the transition points are at most β in absolute value, each parameter of the network is bounded by $O(\beta)$ in absolute value.*

Proof. Since f is piecewise linear, we can define f as follows: there exists $-\infty = \gamma_0 < \gamma_1 < \gamma_2 < \cdots < \gamma_{D-1} = \gamma_D = +\infty$ such that

$$
f(x) = \begin{cases} a_1x + b_1, & x \le \gamma_1\\ a_2x + b_2, & \gamma_1 < x \le \gamma_2\\ \quad\vdots & \\ a_Dx + b_D, & \gamma_{D-1} < x, \end{cases}
$$

where $a_k\gamma_k + b_k = a_{k+1}\gamma_k + b_{k+1}$ for each $k \in [D-1]$. Now we will show that $f(x)$ equals $g(x)$ defined below:

$$g(x) := a_1 x + b_1 + \sum_{i=2}^{D} \mathrm{ReLU}((a_i - a_{i-1})(x - \gamma_{i-1})).$$

We observe that for $\gamma_{k-1} < x \le \gamma_k$,

$$g(x) = a_1 x + b_1 + \sum_{i=2}^{k} (a_i - a_{i-1})(x - \gamma_{i-1}) = a_k x - \sum_{i=2}^{k} (a_i - a_{i-1})\gamma_{i-1}.$$

Then, when $k > 1$, for $\gamma_{k-1} < x \le \gamma_k$, we have

$$\begin{aligned} g(x) &= a_k x - \sum_{i=2}^{k} (a_i - a_{i-1})\gamma_{i-1} \\ &= \left(a_{k-1}\gamma_{k-1} - \sum_{i=2}^{k-1} (a_i - a_{i-1})\gamma_{i-1} \right) + a_k x - a_{k-1}\gamma_{k-1} - (a_k - a_{k-1})\gamma_{k-1} \\ &= g(\gamma_{k-1}) + a_k x - a_k \gamma_{k-1}. \end{aligned}$$

Using these observations, we can inductively show that for each $k \in [D]$, $g(x) = f(x)$ holds for $\gamma_{k-1} < x \le \gamma_k$. For $x \le \gamma_1$,

$$g(x) = a_1 x + b_1 = f(x).$$

Assuming for $\gamma_{k-2} < x \le \gamma_{k-1}$, $g(x) = f(x)$. Then $g(\gamma_{k-1}) = f(\gamma_{k-1}) = a_k\gamma_{k-1} + b_k$. Therefore, for $\gamma_{k-1} < x \le \gamma_k$, we have

$$g(x) = g(\gamma_{k-1}) + a_k x - a_k \gamma_{k-1} = a_k x + b_k = f(x).$$

This proves that $g(x) = f(x)$ for $x \in \mathbb{R}$ and we only need to design neural network to represent g. By employing one neuron for $a_1 x + b_1$ and $D-1$ neurons for $\mathrm{ReLU}((a_i - a_{i-1})(x - \gamma_{i-1}))$, and aggregating their outputs, we obtain the function g. There are $O(D)$ parameters in total, and each parameter is bounded by $O(\beta)$ in absolute value. □

Lemma C.10. *Let $f_1, \ldots, f_k$ be functions mapping $\mathbb{R}$ to $\mathbb{R}$. Suppose each f_i can be represented by a neural network with p parameters bounded by β in absolute value. Then, function $g : \mathbb{R}^k \to \mathbb{R}^k$ defined by*

$$g(x_1, \ldots, x_k) := (f_1(x_1), \ldots, f_k(x_k))$$

can be represented by a neural network with $O(pk)$ parameters bounded by β in absolute value.

Proof. We just need to deal with each coordinate separately and use the neural network representation for each f_i. We just need to concatenate each result of f_i together as the final output. □

Lemma C.11 (ReLU network implementing the score of a one-dimensional σ-smoothed distribution)**.** *Let p be a distribution over $\mathbb{R}$ with mean μ, and let $p_\sigma = p * \mathcal{N}(0, \sigma^2)$ have variance m_2^2 and score s_σ. There exists a constant-depth ReLU network $f : \mathbb{R} \to \mathbb{R}$ with $O(\frac{m_2}{\gamma^3\sigma^4})$ parameters with absolute values bounded by $O(\frac{m_2}{\sigma^2\gamma^2} + \frac{\log\frac{1}{\gamma}}{\sigma^3\gamma} + |\mu|)$ such that*

$$\mathop{\mathbb{E}}_{x\sim p_\sigma} \left[\|s_\sigma(x) - f(x)\|^2 \right] \lesssim \gamma^2$$

and

$$|f(x)| \lesssim \frac{1}{\sigma} \log \frac{1}{\sigma\gamma}$$

Proof. By Lemma C.4, there exists a continuous piecewise approximation of p with $O(\frac{m_2}{\sigma^3\gamma^4})$ pieces with each segment's slope, each transition point, and function value all bounded in $O(\frac{m_2}{\sigma^2\gamma^2} + \frac{1}{\sigma^3\gamma}\log\frac{1}{\sigma\gamma} + \frac{1}{\sigma}\log\frac{1}{\sigma\gamma} + |\mu|)$. Taking this into C.9 and we have the bound. □

Lemma C.12. *Let p be a product distribution over $\mathbb{R}^d$ such that $p(x) = \prod_{i=1}^d p_i(x_i)$, and let $p_\sigma = p * \mathcal{N}(0, \sigma^2 I_d)$ have score s_σ. Assume p_σ has mean μ and variance $m_2^2 = \mathbb{E}_p[\|x-\mu\|_2^2]$. Then, there exists a constant-depth ReLU network $f : \mathbb{R}^d \to \mathbb{R}^d$ with $O(\frac{dm_2}{\gamma^3\sigma^4})$ parameters with absolute values bounded by $O(\frac{dm_2}{\sigma^2\gamma^2} + \frac{\sqrt{d}}{\sigma^3\gamma}\log\frac{d}{\sigma\gamma} + \|\mu\|_1)$ such that*

$$\mathop{\mathbb{E}}_{x\sim p_\sigma}\left[\|s_\sigma(x) - f(x)\|^2\right] \lesssim \gamma^2.$$

and

$$|f(x)_i| \lesssim \frac{1}{\sigma}\log\frac{1}{\sigma\gamma}$$

Proof. Consider distribution $p_i : \mathbb{R} \to \mathbb{R}$ and its σ-smoothed version $p_{i\sigma} = p_i * \mathcal{N}(0, \sigma^2)$. Let μ_i and m_{2i} be the mean and the variance of p_i respectively. Let $s_{\sigma i}$ be the i-th component of s_σ. Then, Lemma C.11 shows that for each $i \in [d]$, there exists a constant-depth ReLU network $f_i : \mathbb{R} \to \mathbb{R}$ with $O(\frac{m_{2i}}{\gamma^3\sigma^4})$ parameters with absolute values bounded by $O(\frac{dm_2}{\sigma^2\gamma^2} + \frac{\sqrt{d}}{\sigma^3\gamma}\log\frac{d}{\sigma\gamma} + |\mu_i|)$ such that

$$\mathop{\mathbb{E}}_{x\sim p_{\sigma i}}\left[\|s_{\sigma i}(x) - f_i(x)\|^2\right] \lesssim \frac{\gamma^2}{d}.$$

Then, we can use the product function $f = (f_1, \ldots, f_d)$ as the approximation for s_σ. By Lemma C.5,

$$\mathop{\mathbb{E}}_{x\sim p_\sigma}\left[\|s_\sigma(x) - f(x)\|^2\right] \lesssim \gamma^2.$$

Taking the fact that $\sum_{i\in[d]} |\mu_i| = \|\mu_1\|$ and $\sum_{i\in[d]} m_{2i} \le dm_2$ into Lemma C.10, and we prove the statement.

□

C.4 ReLU network for Score at Small smoothing level

Lemma C.13 (Vertex Identifier Network)**.** *For any $0 < \alpha < 1$, there exists a ReLU network $h : \mathbb{R}^d \to \mathbb{R}^d$ with $O(d/\alpha)$ parameters, constant depth, and weights bounded by $O(1/\alpha)$ such that*

- *If $|x_i| > \alpha$, for all $i \in [d]$, then $h(x)_i = \frac{x_i}{|x_i|}$ for all $i \in [d]$.*

Proof. Consider the one-dimensional function

$$g(y) = \begin{cases} -1, & y \le -\alpha \\ \frac{y}{\alpha}, & -\alpha < y < \alpha \\ 1, & y \ge \alpha \end{cases}$$

This is a piecewise linear function, where the derivative of each piece is bounded by $\frac{1}{\alpha}$, the value of the transition points are at most α in absolute value, and $|h|$ itself is bounded by 1. Thus, by Lemma C.9, we can represent the function $h(x) = (g(x_1), \ldots, g(x_d))$ using $O(d/\alpha)$ parameters, with each parameter's absolute value bounded by $O(1/\alpha)$. Moreover, clearly $h(x)_i = \frac{x_i}{|x_i|}$ for all $i \in [d]$ whenever $|x_i| \ge \frac{1}{C}$. □

Lemma C.14 (Switch Network)**.** *Consider any function switch* $: \mathbb{R}^{d+1} \to \mathbb{R}^d$ *such that for* $x \in \mathbb{R}^d$*,* $y \in \mathbb{R}$*, with* $|x_i| \le T$ *for all* $i \in [d]$*,*

$$switch(x, y) = \begin{cases} x & \text{if } y = 1 \\ 0 & \text{if } y = -1 \end{cases}$$

switch can be implemented using a constant depth ReLU network with $O(dT)$ *parameters, with each parameter's absolute value bounded by* $O(T)$*.*

Proof. Consider the ReLU network given by

$$\text{switch}(x, y)_i = \text{ReLU}((x_i - 2T) + 2T \cdot y) - \text{ReLU}((-x_i - 2T) + 2T \cdot y)$$

It computes our claimed function. Moreover, it is constant-depth, the number of parameters is $O(dT)$, and each parameter is bounded by $O(T)$ in absolute value, as claimed. □

Lemma C.15. *Let* $d' = O(d)$*. Given a constant-depth ReLU network representing a one-way function* $f : \{-1, 1\}^d \to \{-1, 1\}^{d'}$ *with* $\text{poly}(d)$ *parameters, there is a constant-depth ReLU network* $h : \mathbb{R}^{d+d'} \to \mathbb{R}^{d+d'}$ *with* $\text{poly}\left(\frac{d}{\sigma\gamma}\right)$ *parameters with each parameter bounded in absolute value by* $\text{poly}\left(\frac{d}{\sigma\gamma}\right)$ *such that for the unconditional distribution* g *defined in Definition 3.1 with* σ*-smoothed version* g_σ *and corresponding score* s_σ*, for* $\tau = \frac{1}{d^C}$ *and* $\tau \le \sigma < \frac{R}{C\sqrt{\log d}}$ *for sufficiently large constant* C*, and* $R > C \log d$*,* $\varepsilon > \frac{1}{\text{poly}(d)}$*,* $\gamma > \frac{1}{d^{C/100}}$

$$\mathop{\mathbb{E}}_{x \sim g_\sigma} \left[\|s_\sigma(x) - h(x)\|^2\right] \lesssim \gamma^2$$

Proof. We will let our ReLU network h be as follows. Let r be the ReLU network from Lemma C.13 that identifies the closest hypercube vertex with any constant parameter $\alpha < 1$.

For each $i \in [d]$, we will let $\widetilde{h}_i$ be the ReLU network that implements the approximation to the score of the one-dimensional distribution $w_1(x) * \mathcal{N}(0, \sigma^2)$ from Lemma C.11. By the lemma, it satisfies

$$\mathop{\mathbb{E}}_{x \sim w_{\sqrt{\sigma^2+1}}} \left[(\widetilde{h}_i(x) - \nabla \log w_{\sqrt{\sigma^2+1}}(x))^2\right] \lesssim \gamma^2 \tag{10}$$

For $i \in d + [d']$, we will let $\widetilde{h}_{i,1}$ be the ReLU network that implements the approximation to the score $\widetilde{s}_{\sigma,i,-1}$ of $\widetilde{g}_{\sigma,i,-1} = (\frac{w_1 \cdot \text{comb}_\varepsilon}{\int w_1(x) \cdot \text{comb}_\varepsilon(x) dx}) * \mathcal{N}(0, \sigma^2)$, and we will let $\widetilde{h}_{i,-1}$ implement the approximation to the score of $\left(\frac{w_1(x) \cdot \text{comb}_\varepsilon(x - \varepsilon/2)}{\int w_1(x) \cdot \text{comb}_\varepsilon(x - \varepsilon/2) dx}\right) * \mathcal{N}(0, \sigma^2)$, as given by Lemma C.11. By the Lemma, for every $i \in d + [d']$ and $j \in \{\pm 1\}$, we have

$$\mathop{\mathbb{E}}_{x \sim \widetilde{g}_{\sigma,i,j}} \left[(\widetilde{h}_{i,j}(x) - s_{\sigma,i,j}(x))^2\right] \lesssim \gamma^2 \tag{11}$$

Note that each $|\widetilde{h}_i| \le \frac{C}{\sigma} \log \frac{1}{\sigma\gamma}$ for $i \le d$, and $|\widetilde{h}_{i,\pm 1}| \le \frac{C}{\sigma} \log \frac{1}{\sigma\gamma}$ for $i > d$, for sufficiently large constant C.

Now let switch be the ReLU network described in Lemma C.14 for $T = \frac{C}{\sigma} \log \frac{1}{\sigma\gamma}$.

Consider the network $h : \mathbb{R}^{d+d'} \to \mathbb{R}^{d+d'}$ given by

$$h(x)_i = \begin{cases} \widetilde{h}_i(x_i - r(x)_i \cdot R) & \text{for } i \le d \\ \text{switch}(\widetilde{h}_{i,1}(x_i), f(r(x))_{i-d}) + \text{switch}(\widetilde{h}_{i,-1}(x_i), -f(r(x))_{i-d}) & \text{for } i > d \end{cases}$$

Note that h can be represented with poly $\left(\frac{d}{\sigma\gamma}\right)$ parameters with absolute value of each parameter bounded in poly $\left(\frac{d}{\sigma\gamma}\right)$. We will show that h approximates s_σ well in multiple steps.

For $r(x) \in \{\pm 1\}^d$, consider the score $s_{\sigma,r(x)}$ of $g_{\sigma,r(x)}$, the σ-smoothed version of the distribution $g_{r(x)}$ centered at $\widetilde{r(x)} \in \mathbb{R}^{d+d'}$, as described in Definition 3.1, where $\widetilde{r(x)}$ has the first d coordinates given by $r(x)$, and the remaining coordinates set to 0.

Whenever $r(x) = j \in \{\pm 1\}^d$, h approximates $s_{\sigma,j}$ well over $g_{\sigma,j}$. We will show that for fixed $j \in \{\pm 1\}^d$

$$\mathop{\mathbb{E}}_{x\sim g_{\sigma,j}}\left[\|s_{\sigma,j}(x) - h(x)\|^2 \cdot 1_{r(x)=j}\right] \lesssim d\gamma^2$$

First, note that for $i \le d$, by (10) and our definition of h,

$$\mathop{\mathbb{E}}_{x\sim w_{\sqrt{\sigma^2+1}}}\left[(h(x)_i - \nabla \log w_{\sqrt{\sigma^2+1}}(x - R\cdot j))^2 \cdot 1_{r(x)=j}\right] \lesssim \gamma^2$$

On the other hand, for $i > d$, by (11) and our definition of h,

$$\mathop{\mathbb{E}}_{x\sim \widetilde{g}_{\sigma,i,f(j)_{i-d}}}\left[(h(x)_i - s_{\sigma,i,f(j)_{i-d}})^2 \cdot 1_{r(x)=j}\right] \lesssim \gamma^2$$

Since by Definition 3.1, for $j \in \{\pm 1\}^d$, $g_{\sigma,j}(x) = \prod_{i=1}^d w_{\sqrt{\sigma^2+1}}(x) \cdot \prod_{i=d+1}^{d+d'} \widetilde{g}_{\sigma,i,f(j)_{i-d}}(x)$, we have by Lemma C.5,

$$\mathop{\mathbb{E}}_{x\sim g_{\sigma,j}}\left[\|h(x) - s_{\sigma,j}(x)\|^2 \cdot 1_{r(x)=j}\right] \lesssim d\gamma^2$$

h approximates $s_{\sigma,j}$ exponentially accurately over g_σ. By Lemma C.6, we have that for x such that $r(x) = j$,

$$\left|\frac{1}{2^d} g_{\sigma,j}(x) - g_\sigma(x)\right| \lesssim \frac{1}{2^d} \cdot e^{-\frac{R^2}{4(1+\sigma^2)}} \lesssim \frac{1}{2^d}$$

for our choice of R, σ.

So, we have

$$\mathop{\mathbb{E}}_{x\sim g_\sigma}\left[\|h(x) - s_{\sigma,j}(x)\|^2 \cdot 1_{r(x)=j}\right] \lesssim \frac{d\gamma^2}{2^d}$$

h approximates s_σ well over g_σ whenever $r(x) \in \{\pm 1\}^d$. Summing the above over $j \in \{\pm 1\}^d$ gives

$$\mathop{\mathbb{E}}_{x\sim g_\sigma}\left[\|h(x) - s_{\sigma,r(x)}(x)\|^2 \cdot 1_{r(x)\in\{\pm 1\}^d}\right] \lesssim d\gamma^2$$

Moreover, by Lemma C.8, for $\bar{r}(x) = y$ where $y \in \{\pm 1\}^d$ represents the orthant that $x \in \{\pm 1\}^d$ belongs to,

$$\mathop{\mathbb{E}}_{x\sim g_\sigma}\left[\|s_{\sigma,\bar{r}(x)} - s_\sigma(x)\|^2\right] \lesssim e^{-\Omega\left(\frac{R^2}{1+\sigma^2}\right)} \lesssim \frac{1}{d^{C^2/10}}$$

So, by the above, we have that

$$\mathop{\mathbb{E}}_{x\sim g_\sigma}\left[\|h(x) - s_\sigma(x)\|^2 \cdot 1_{r(x)\in\{\pm 1\}^d}\right] \lesssim d\gamma^2 + \frac{1}{d^{C^2/10}}$$

Contribution of x such that $r(x) \notin \{\pm 1\}^d$ is small. By the definition of r,

$$\mathbb{P}_{x\sim g_\sigma}\left[r(x) \notin \{\pm 1\}^d\right] \lesssim de^{-\frac{(R-\alpha)^2}{2}} \lesssim e^{-\frac{R^2}{4}}$$

So, by Cauchy-Schwarz,

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim g_\sigma}\left[\|s_\sigma(x)\|^2 \cdot 1_{r(x)\notin\{\pm1\}^d}\right] &\le \sqrt{\mathop{\mathbb{E}}_{x\sim g_\sigma}\left[\|s_\sigma(x)\|^4\right] \cdot \mathbb{P}_{x\sim g_\sigma}\left[r(x) \notin \{\pm 1\}^d\right]} \\ &\lesssim \frac{1}{\sigma^2} e^{-\frac{R^2}{4}} \\ &\lesssim e^{-\frac{R^2}{8}}\end{aligned}$$

Similarly, since $|h(x)_i| \le \frac{C}{\sigma}\log\frac{1}{\sigma\gamma}$, we have

$$\mathop{\mathbb{E}}_{x\sim g_\sigma}\left[\|h(x)\|^2 \cdot 1_{r(x)\notin\{\pm1\}^d}\right] \lesssim \frac{1}{\sigma^2}\log^2\frac{1}{\sigma\gamma}\cdot e^{-\frac{R^2}{4}} \lesssim e^{-\frac{R^2}{8}}$$

Thus, we have

$$\mathop{\mathbb{E}}_{x\sim g_\sigma}\left[\|h(x)-s_\sigma(x)\|^2 \cdot 1_{r(x)\notin\{\pm1\}^d}\right] \lesssim e^{-\frac{R^2}{8}} \lesssim \frac{1}{d^{C^2/40}}$$

Putting it together. By the above, we have,

$$\mathop{\mathbb{E}}_{x\sim g_\sigma}\left[\|h(x)-s_\sigma(x)\|^2\right] \lesssim d\gamma^2 + \frac{1}{d^{C^2/40}}$$

Reparameterizing γ and noting that $\gamma > \frac{1}{d^{C/100}}$ gives the claim. □

C.5 Smoothing a discretized Gaussian

Lemma C.16. *For any ϕ, let g be the univariate discrete Gaussian with pdf*

$$g(x) \propto w_1(x)\cdot comb_\varepsilon(x-\phi)$$

Consider the ρ-smoothed version of g, given by $g_\rho = g * w_\rho$. We have that

$$\left|g_\rho(x) - w_{\sqrt{\rho^2+1}}(x)\right| \lesssim e^{-\frac{\rho^2}{2\varepsilon^2(1+\rho^2)}}\cdot w_{\sqrt{\rho^2+1}}(x)$$

and

$$\left|g'_\rho(x) - w'_{\sqrt{\rho^2+1}}(x)\right| \lesssim e^{-\frac{\rho^2}{2\varepsilon^2(1+\rho^2)}}\cdot |w'_{\sqrt{\rho^2+1}}(x)| + \sum_{j>0} e^{-\frac{j^2\rho^2}{2\varepsilon^2(1+\rho)^2}}\cdot\frac{j}{\varepsilon(1+\rho^2)}\cdot w_{\sqrt{\rho^2+1}}(x)$$

Proof. The Fourier transform of the $\mathrm{comb}_\varepsilon$ distribution is given by $\frac{1}{\varepsilon}\mathrm{comb}_{1/\varepsilon}$. So, for the discrete Gaussian g, we have that its Fourier Transform is given by

$$\begin{aligned}\widehat{g}(\xi) &= \left(\widehat{w}_1 * \left(\frac{e^{-i\xi\phi}}{\varepsilon}\mathrm{comb}_{1/\varepsilon}\right)\right)(\xi) \\ &= \frac{1}{\varepsilon}\cdot\sum_{j\in\mathbb{Z}} e^{-i\frac{j}{\varepsilon}\phi}\cdot e^{-\frac{(\xi-\frac{j}{\varepsilon})^2}{2}}\end{aligned}$$

Then, for g_ρ, the ρ-smoothed version of g, we have that its Fourier Transform is

$$\begin{aligned}\widehat{g}_\rho(\xi) &= (\widehat{g}\cdot\widehat{w}_{1/\rho})(\xi)\\ &= \frac{1}{\varepsilon}\sum_{j=-k}^{k} e^{-\frac{ij}{\varepsilon}\phi} e^{-\frac{\xi^2\rho^2}{2}}\cdot e^{-\frac{\left(\xi-\frac{j}{\varepsilon}\right)^2}{2}}\end{aligned}$$

So, we have that, by the inverse Fourier transform,

$$\begin{aligned}g_\rho(x) &= w_{\sqrt{\rho^2+1}}(x) + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\xi x}\sum_{j\neq 0} e^{-\frac{ij}{\varepsilon}\phi} e^{-\frac{\xi^2\rho^2}{2}-\frac{(\xi-\frac{j}{\varepsilon})^2}{2}} d\xi\\ &= w_{\sqrt{\rho^2+1}}(x) + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\xi x}\cdot e^{-\frac{j^2}{2\varepsilon^2}+\frac{j^2}{2\varepsilon^2(\rho^2+1)}}\sum_{j\neq 0} e^{-\frac{ij}{\varepsilon}\phi}\cdot e^{-\frac{\rho^2+1}{2}\left(\xi-\frac{j}{\varepsilon(\rho^2+1)}\right)^2}\\ &= w_{\sqrt{\rho^2+1}}(x) + \sum_{j\neq 0} e^{-\frac{ij}{\varepsilon}\phi}\cdot e^{-\frac{j^2\rho^2}{2\varepsilon^2(\rho^2+1)}} e^{ix\frac{j}{\varepsilon(\rho^2+1)}}\cdot w_{\sqrt{\rho^2+1}}(x)\end{aligned}$$

so that

$$\left|g_\rho(x) - w_{\sqrt{\rho^2+1}}(x)\right| \leq \sum_{j\neq 0} e^{-\frac{j^2\rho^2}{2\varepsilon^2(\rho^2+1)}} w_{\sqrt{\rho^2+1}}(x)$$

giving the first claim. For the second claim, note that the above gives

$$g'_\rho(x) = w'_{\sqrt{\rho^2+1}}(x) + \sum_{j\neq 0} e^{-\frac{ij\phi}{\varepsilon}}\cdot e^{-\frac{j^2\rho^2}{2\varepsilon^2(\rho^2+1)}} e^{\frac{ixj}{\varepsilon(\rho^2+1)}}\cdot\left(\frac{ij}{\varepsilon(\rho^2+1)} w_{\sqrt{\rho^2+1}}(x) + w'_{\sqrt{\rho^2+1}}(x)\right)$$

So,

$$\left|g'_\rho(x) - w'_{\sqrt{\rho^2+1}}(x)\right| \lesssim e^{-\frac{\rho^2}{2\varepsilon^2(1+\rho^2)}}\cdot\left|w'_{\sqrt{\rho^2+1}}(x)\right| + \sum_{j>0} e^{-\frac{j^2\rho^2}{2\varepsilon^2(1+\rho^2)}}\cdot\frac{j}{\varepsilon(1+\rho^2)} w_{\sqrt{\rho^2+1}}(x)$$

□

C.6 Large Noise Level - Distribution is close to a mixture of Gaussians

Lemma C.17. *Let C be a sufficiently large constant. Let $g^j(x) \propto \prod_{i=1}^{d} w_1(x_i)\cdot\prod_{i=d+1}^{d'} w_1(x_i)\cdot comb_\varepsilon(x_i-\phi_{i,j})$ be the pdf of a distribution on $\mathbb{R}^{d+d'}$ with shifts $\phi_{i,j}$. Consider a mixture of discrete d-dimensional Gaussians, given by the pdf*

$$h(x) = \sum_{j=1}^{k}\beta_j g^j(x-\mu_j)$$

Let $h_\rho(x) = (h * w_\rho)(x)$ be the smoothed version of h. Then, for the mixture of standard $(d+d')$-dimensional Gaussians given by

$$f_\rho(x) = \sum_{j=1}^{k}\beta_j w_{\sqrt{\rho^2+1}}(x-\mu_j)$$

for $\frac{\rho^2}{\varepsilon^2(1+\rho^2)} > C\log d$, *we have that*

$$\mathop{\mathbb{E}}_{x\sim h_\rho}\left[\left\|\frac{\nabla h_\rho(x)}{h_\rho(x)} - \frac{\nabla f_\rho(x)}{f_\rho(x)}\right\|^2\right] \lesssim e^{-\frac{\rho^2}{2\varepsilon^2(1+\rho^2)}}\cdot\left(1+m_2^2+\sup_j\|\mu_j\|^2\right) + \sum_{j>0} e^{-\frac{j^2\rho^2}{2\varepsilon^2(1+\rho^2)}}\frac{j}{\varepsilon(1+\rho^2)}$$

where $m_2^2 = \mathbb{E}_{x\sim h_\rho}\left[\|x\|^2\right]$.

Proof. We have that

$$h_\rho(x) = \sum_{i=1}^k \beta_j g_\rho^j(x-\mu_j)$$

where $g_\rho^j(x) = g^j(x) * w_\rho(x)$. By Lemma C.16, we have that for every i, j,

$$\begin{aligned}&\left|\left(\nabla g_\rho^j(x)\right)_i - \left(\nabla w_{\sqrt{\rho^2+1}}(x)\right)_i\right| \\ &\lesssim d e^{-\frac{\rho^2}{2\varepsilon^2(1+\rho^2)}}\left|\left(\nabla w_{\sqrt{\rho^2+1}}(x)\right)_i\right| + d\cdot\sum_{j>0} e^{-\frac{j^2\rho^2}{2\varepsilon^2(1+\rho^2)}}\cdot\frac{j}{\varepsilon(1+\rho^2)}\cdot w_{\sqrt{\rho^2+1}}(x)\end{aligned}$$

So,

$$\begin{aligned}&\left|\left(\nabla h_\rho(x)\right)_i - \left(\nabla f_\rho(x)\right)_i\right| \\ &= \left|\sum_{j=1}^k \beta_j\cdot\left(\nabla g_\rho^j(x-\mu_j) - \nabla w_{\sqrt{\rho^2+1}}(x-\mu_j)\right)_i\right| \\ &\lesssim d\sum_{j=1}^k\beta_j\cdot\left(e^{-\frac{\rho^2}{2\varepsilon^2(1+\rho^2)}}\left|\nabla\left(w_{\sqrt{\rho^2+1}}(x-\mu_j)\right)_i\right| + \sum_{j>0} e^{-\frac{j^2\rho^2}{2\varepsilon^2(1+\rho^2)}}\frac{j}{\varepsilon(1+\rho^2)}\cdot w_{\sqrt{\rho^2+1}}(x)\right)\end{aligned}$$

Similarly, for the density, by Lemma C.16

$$\left|g_\rho^j(x) - w_{\sqrt{\rho^2+1}}(x)\right| \lesssim d e^{-\frac{\rho^2}{2\varepsilon^2(1+\rho^2)}} w_{\sqrt{\rho^2+1}}(x)$$

So,

$$\begin{aligned}|h_\rho(x) - f_\rho(x)| &= \left|\sum_{j=1}^k\beta_j\cdot\left(g_\rho^j(x-\mu_j) - w_{\sqrt{\rho^2+1}}(x-\mu_j)\right)\right| \\ &\lesssim d e^{-\frac{\rho^2}{2\varepsilon^2(1+\rho^2)}}\sum_{j=1}^k\beta_j\cdot w_{\sqrt{\rho^2+1}}(x-\mu_j) \\ &\lesssim d e^{-\frac{\rho^2}{2\varepsilon^2(1+\rho^2)}} f_\rho(x)\end{aligned}$$

Thus, we have

$$\mathop{\mathbb{E}}_{x\sim h_\rho}\left[\left(\frac{(\nabla h_\rho(x))_i}{h_\rho(x)} - \frac{(\nabla f_\rho(x))_i}{f_\rho(x)}\right)^2\right]$$

$$
\begin{aligned}
&\leq \mathop{\mathbb{E}}_{x\sim h_\rho}\left[\left(\frac{\nabla h_\rho(x))_i}{f_\rho(x)\cdot\left(1+O\left(de^{-\frac{\rho^2}{2\varepsilon^2(1+\rho^2)}}\right)\right)}-\frac{(\nabla f_\rho(x))_i}{f_\rho(x)}\right)^2\right]\\
&\lesssim de^{-\frac{\rho^2}{\varepsilon^2(1+\rho^2)}}\cdot\mathbb{E}\left[\left(\frac{(\nabla f_\rho(x)\ )_i}{f_\rho(x)}\right)^2\right]\\
&+d\cdot\mathbb{E}\left[\left(\frac{\sum_{j=1}^k\beta_j\cdot\left(e^{-\frac{\rho^2}{2\varepsilon^2(1+\rho^2)}}\left|\nabla w_{\sqrt{\rho^2+1}}(x-\mu_j)\right|_i+\sum_{j>0}e^{-\frac{j^2\rho^2}{2\varepsilon^2(1+\rho^2)}}\frac{j}{\varepsilon(1+\rho^2)}\cdot w_{\sqrt{\rho^2+1}}(x)\right)}{f_\rho(x)}\right)^2\right]\\
&\lesssim\frac{d}{\rho^2}e^{-\frac{\rho^2}{\varepsilon^2(1+\rho^2)}}+d\sum_{j>0}e^{-\frac{j^2\rho^2}{\varepsilon^2(\rho^2+1)}}\frac{j}{\varepsilon(1+\rho^2)}+de^{-\frac{\rho^2}{\varepsilon^2(1+\rho^2)}}\cdot\mathbb{E}\left[\left(\frac{\sum_{j=1}^k\beta_j\cdot|x-\mu_j|_i\cdot w_{\sqrt{\rho^2+1}}(x-\mu_j)}{f_\rho(x)}\right)^2\right]\\
&\lesssim de^{-\frac{\rho^2}{\varepsilon^2(1+\rho^2)}}+d\sum_{j>0}e^{-\frac{j^2\rho^2}{\varepsilon^2(1+\rho^2)}}\frac{j}{\varepsilon(1+\rho^2)}+de^{-\frac{\rho^2}{\varepsilon^2(1+\rho^2)}}\cdot\mathbb{E}\left[\sup_j|x-\mu_j|_i^2\right]\\
&\lesssim de^{-\frac{\rho^2}{\varepsilon^2(1+\rho^2)}}\left(1+\mathbb{E}\left[x_i^2\right]+\sup_j|\mu_j|_i^2\right)+d\sum_{j>0}e^{-\frac{j^2\rho^2}{\varepsilon^2(1+\rho^2)}}\frac{j}{\varepsilon(1+\rho^2)}
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\mathop{\mathbb{E}}_{x\sim h_\rho}\left[\left\|\frac{\nabla h_\rho(x)}{h_\rho(x)}-\frac{\nabla f_\rho(x)}{f_\rho(x)}\right\|^2\right]&\lesssim d^2e^{-\frac{\rho^2}{\varepsilon^2(1+\rho^2)}}\cdot\left(1+\mathbb{E}\left[\|x\|^2\right]+\sup_j\|\mu_j\|^2\right)+d^2\sum_{j>0}e^{-\frac{j^2\rho^2}{2\varepsilon^2(1+\rho^2)}}\frac{j}{\varepsilon(1+\rho^2)}\\
&\lesssim e^{-\frac{\rho^2}{2\varepsilon^2(1+\rho^2)}}\cdot\left(1+m_2^2+\sup_j\|\mu_j\|^2\right)+\sum_{j>0}e^{-\frac{j^2\rho^2}{2\varepsilon^2(1+\rho^2)}}\frac{j}{\varepsilon(1+\rho^2)}
\end{aligned}
$$

since $\frac{\rho^2}{\varepsilon^2(1+\rho^2)}>C\log d$ □

Corollary C.18. *Let $d'=O(d)$, and let g be as defined in Definition 3.1, and let $g_\rho=g*\mathcal{N}(0,\rho^2I_{d+d'})$ be the ρ-smoothed version of g. Let f_ρ be the mixture of $(d+d')$-dimensional standard Gaussians, given by*

$$
f_\rho(y)=\frac{1}{2^d}\sum_{x\in\{\pm1\}^d}w_{\sqrt{\rho^2+1}}(y-R\cdot\widetilde{x})
$$

where $\widetilde{x}\in\mathbb{R}^{d+d'}$ has the first d coordinates given by x, and the last d' coordinates 0. Then, for $\frac{\rho^2}{\varepsilon^2(1+\rho^2)}>C\log d$ for sufficiently large constant C, we have

$$
\mathop{\mathbb{E}}_{x\sim g_\rho}\left[\|\nabla\log g_\rho(x)-\nabla\log f_\rho(x)\|^2\right]\lesssim e^{-\frac{\rho^2}{2\varepsilon^2(1+\rho^2)}}\cdot\left(1+R^2+\rho^2\right)+\sum_{j>0}e^{-\frac{j^2\rho^2}{2\varepsilon^2(1+\rho^2)}}\frac{j}{\varepsilon(1+\rho^2)}
$$

Proof. Follows from the facts that $m_2^2\lesssim d\left(R^2+\rho^2\right)$ and $\mu_j^2\lesssim dR^2$ for all j. □

C.7 ReLU network for Score at Large smoothing Level

This section shows how to represent the score of the σ-smoothed unconditional distribution defined in Definition 3.1 for large σ using a ReLU network with a polynomial number of parameters bounded by a polynomial in the relevant quantities. We proceed in two stages – first, we show how to represent the score of a mixture of Gaussians placed on the vertices of a scaled hypercube. Then, we show that for large σ, this network is close to the score of the σ-smoothed unconditional distribution.

Lemma C.19 (ReLU network representing score of mixture of Gaussians on hypercube)**.** *For any $\sigma > 0$ and $R > 1$ consider the distribution on $\mathbb{R}^d$ with pdf*

$$f_\sigma(x) = \frac{1}{2^d} \sum_{\mu \in \{\pm 1\}^d} w_\sigma(x - R\mu)$$

where w_σ is the pdf of $\mathcal{N}(0, \sigma^2 I_d)$.

There is a constant depth ReLU network $h : \mathbb{R}^d \to \mathbb{R}^d$ with $O\left(\frac{dR}{\gamma^3\sigma^4}\right)$ parameters, with absolute values bounded by $O\left(\frac{dR}{\sigma^3\gamma^2}\right)$ such that

$$\mathop{\mathbb{E}}_{x \sim f_\sigma} \left[\|\nabla \log f_\sigma(x) - h(x)\|^2\right] \lesssim \gamma^2$$

Proof. Note that g_σ is a product distribution. So, the claim follows by Lemma C.12. □

Lemma C.20. *Let $d' = O(d)$, and let $R \le \mathrm{poly}(d)$. Let g be the pdf of the unconditional distribution on $\mathbb{R}^{d+d'}$, as defined in Definition 3.1, and let g_σ be its σ-smoothed version with score s_σ. For $\varepsilon < \frac{1}{C\sqrt{\log d}}$, and $\sigma > C\varepsilon\left(\sqrt{\log d} + \sqrt{\log \frac{1}{\varepsilon}}\right)$ for sufficiently large constant C, there is a constant depth ReLU network h with $O\left(\frac{dR}{\gamma^3\sigma^4}\right)$ parameters with absolute values bounded by $O\left(\frac{dR}{\sigma^3\gamma^2}\right)$ such that*

$$\mathop{\mathbb{E}}_{x \sim g_\sigma} \left[\|s_\sigma(x) - h(x)\|^2\right] \lesssim \gamma^2 + \frac{1}{d^{C^2/20}}$$

Proof. Let h be the ReLU network from Lemma C.19 for smoothing σ. It satisfies our bounds on the number of parameters and the absolute values.

Note that for our setting of ε and σ, we have that

$$\frac{\sigma^2}{\varepsilon^2(1+\sigma^2)} = \frac{1}{\varepsilon^2\left(1 + \frac{1}{\sigma^2}\right)} > \frac{1}{\varepsilon^2 + \frac{1}{C^2 \log d}} > \frac{1}{\frac{2}{C^2 \log d}} > \frac{C^2 \log d}{2}$$

and

$$\frac{\sigma^2}{\varepsilon^2(1+\sigma^2)} > \frac{C^2(\log d + \log \frac{1}{\varepsilon})}{2} > \log \frac{1}{\varepsilon(1+\sigma^2)}$$

So, by Lemma C.18, for the mixture of Gaussians f_σ as described in Lemma C.19, for $R \le \mathrm{poly}(d)$,

$$\mathop{\mathbb{E}}_{x \sim g_\sigma} \left[\|s_\sigma(x) - \nabla \log f_\sigma(x)\|^2\right] \lesssim e^{-\frac{\sigma^2}{10\varepsilon^2(1+\sigma^2)}} \lesssim \frac{1}{d^{C^2/20}}$$

Also, by Lemma C.16,

$$|g_\sigma(x) - f_\sigma(x)| \lesssim \frac{f_\sigma(x)}{d^{C^2/20}}$$

So, by Lemma C.19,

$$\mathop{\mathbb{E}}_{x\sim g_\sigma}\left[\|\nabla \log f_\sigma(x) - h(x)\|^2\right] \lesssim \mathop{\mathbb{E}}_{x\sim f_\sigma}\left[\|\nabla \log f_\sigma(x) - h(x)\|^2\right] \lesssim \gamma^2$$

So we have

$$\mathop{\mathbb{E}}_{x\sim g_\sigma}\left[\|s_\sigma(x) - h(x)\|^2\right] \lesssim \gamma^2 + \frac{1}{d^{C^2/20}}$$

□

C.8 ReLU Network Approximating score of Unconditional Distribution

Theorem C.21 (ReLU Score Approximation for Lower bound Distribution)**.** *Let C be a sufficiently large constant, and let $d' = O(d)$. Fix any $\sigma \geq \tau$ for $\tau = \frac{1}{d^C}$. Given a constant-depth ReLU network representing a function $f : \{-1,1\}^d \to \{-1,1\}^{d'}$ with $\mathrm{poly}(d)$ parameters, there is a constant-depth ReLU network $h : \mathbb{R}^{d+d'} \to \mathbb{R}^{d+d'}$ with $\mathrm{poly}\,(d)$ parameters with each parameter bounded in absolute value by $\mathrm{poly}\,(d)$ such that for the unconditional distribution g defined in Definition 3.1 with σ-smoothed version g_σ and corresponding score s_σ, for $R > C\log d$, $\frac{1}{\mathrm{poly}(d)} < \varepsilon < \frac{1}{C\sqrt{\log d}}$,*

$$\mathop{\mathbb{E}}_{x\sim g_\sigma}\left[\|s_\sigma(x) - h(x)\|^2\right] \lesssim \frac{1}{d^{C/200}}$$

Proof. Follows by Lemmas C.15 and C.20. □

Corollary 3.5 (Lower Bound Distribution is Well-Modeled)**.** *Let C be a sufficiently large constant. Given a ReLU network $f : \{\pm 1\}^d \to \{\pm 1\}^{d'}$ with $\mathrm{poly}(d)$ parameters bounded by $\mathrm{poly}(d)$ in absolute value, the distribution $\widetilde{g}$ defined in Definition 3.2 for $R = C\log d$ and $\frac{1}{\mathrm{poly}(d)} < \varepsilon < \frac{1}{C\sqrt{\log d}}$, is $O(C)$-well-modeled.*

Proof. Follows via reparameterization from the Theorem, and rescaling. □

D Lower Bound – Putting it all Together

Theorem 1.8 (Lower bound)**.** *Suppose that one-way functions exist. Then for any $d^{0.01} < m < d/2$, there exists a 10-well-modeled distribution over $\mathbb{R}^d$, and linear measurement model with m measurements and noise parameter $\beta = \Theta(\frac{1}{\log^2 d})$, such that $(\frac{1}{10}, \frac{1}{10})$-posterior sampling requires superpolynomial time.*

Proof. First, by Lemma G.4, there exists a ReLU network that represents a one-way function $f : \{\pm 1\}^m \to \{\pm 1\}^m$, with constant weights, polynomial size, and parameters bounded in magnitude by $\mathrm{poly}(d)$.

Therefore, by Corollary 3.5, the distribution $\widetilde{g}$ over $\mathbb{R}^d$ is a C-well-modeled distribution, if we take $R = C\log d$, $\varepsilon = \frac{1}{C\sqrt{\log d}}$. Further, if we take a linear measurement model with $\beta = \frac{1}{C^2\log^2(d)}$, then by Lemma 3.4, any $(1/10, 1/10)$-posterior sampler for this distribution takes at least $2^{\Omega(m)}$ time to run. □

Theorem 1.9 (Lower bound: exponential hardness)**.** *For any $m \le d/2$, suppose that there exists a one-way function $f : \{\pm 1\}^m \to \{\pm 1\}^m$ that requires $2^{\Omega(m)}$ time to invert. Then for any $C > 1$, there exists a C-well-modeled distribution over $\mathbb{R}^d$ and linear measurement model with m measurements and noise level $\beta = \frac{1}{C^2 \log^2 d}$, such that $(\frac{1}{10}, \frac{1}{10})$-posterior sampling takes at least $2^{\Omega(m)}$ time.*

Proof. First, by Lemma G.4, there exists a ReLU network that represents a one-way function $f : \{\pm 1\}^m \to \{\pm 1\}^m$, with constant weights, polynomial size, and parameters bounded in magnitude by $\mathrm{poly}(d)$.

Therefore, by Corollary 3.5, the distribution $\widetilde{g}$ over $\mathbb{R}^d$ is a C-well-modeled distribution, if we take $R = C \log d$, $\varepsilon = \frac{1}{C\sqrt{\log d}}$. Further, if we take a linear measurement model with $\beta = \frac{1}{C^2 \log^2(d)}$, then by Lemma B.5, any $(1/10, 1/10)$-posterior sampler for this distribution takes at least $2^{\Omega(m)}$ time to run. □

E Upper Bound

Lemma E.1. *Let q be a distribution over $\mathbb{R}^m$ such that $\mathbb{E}_{w\sim q}[\|w\|_2^2] = O(m)$. Let $w \sim q$ and $y = w + \beta\mathcal{N}(0, I_m)$. Then, there exists a constant $c > 0$ such that*

$$\mathbb{P}_y\left[\mathbb{P}_w\left[\|y - w\| \le 10\gamma\sqrt{m + \log(1/\delta)} \;\middle|\; y\right] \ge (c\gamma)^m \cdot \delta^{m/2+1}\right] \ge 1 - \delta.$$

Proof. Since $\mathbb{E}_{w\sim q}[\|w\|_2^2] \lesssim m$, there exists a constant C such that

$$\mathbb{P}_{w\sim q}\left[\|w\|_2^2 > \frac{Cm}{\delta}\right] < \frac{\delta}{3}.$$

Lemma H.10 shows that there exists a covering over $\{x \in \mathbb{R}^m \mid \|x\|_2 \le \sqrt{Cm/\delta}\}$ with $N = O(\frac{1}{\sqrt{\delta}\beta})^m$ balls of radius $\beta\sqrt{m + \log(1/\delta)}$. Let S be the set of all the covering balls. This means that

$$\mathbb{P}[\exists \theta \in S : w \in \theta] \ge 1 - \frac{\delta}{3}.$$

Define

$$S' := \{\theta \in S \mid \mathbb{P}_w[w \in \theta] > \frac{\delta}{3N}\}.$$

Then we have that with high probability, w will land in one of the cells in S':

$$\mathbb{P}_w\left[\forall \theta \in S' : w \notin \theta\right] \le \mathbb{P}[\forall \theta \in S : w \notin \theta] + \mathbb{P}[\bigvee_{\theta \in S \setminus S'} w \in \theta] \le \frac{\delta}{3} + N \cdot \frac{\delta}{3N} \le \frac{2\delta}{3}.$$

Moreover, we define

$$S^+ := \{y \in \mathbb{R}^m \mid \exists \theta \in S', \forall w \in \theta : \|w - y\| \le 10\beta\sqrt{m + \log\frac{1}{\delta}}\}.$$

By the sampling process of y, we have that

$$\begin{aligned}
\mathbb{P}_y\left[y \in S^+\right] &= \mathbb{P}_{w\sim q, z\sim\mathcal{N}(0,I_m)}\left[w + \beta z \in S^+\right] \\
&\ge \mathbb{P}_{w\sim q, z\sim\mathcal{N}(0,I_m)}\left[(\exists \theta \in S' : w \in \theta) \wedge (\|z\| \le 8\sqrt{m})\right] \\
&\ge 1 - \mathbb{P}_w\left[\forall \theta \in S' : w \notin \theta\right] - \mathbb{P}_{z\sim\mathcal{N}(0,I_m)}\left[\|z\|^2 > 64(m + \log\frac{1}{\delta})\right]
\end{aligned}$$

By Lemma H.9, we have

$$\mathbb{P}_{z\sim\mathcal{N}(0,I_m)}\left[\|z\|^2 > 64(m+\log\frac{1}{\delta})\right] < \frac{\delta}{3}.$$

Therefore,

$$\mathbb{P}_y\left[y \in S^+\right] \geq 1-\delta.$$

This implies that with $1-\delta$ probability over y, there exists a cell $\theta \in S$ such that $\left\|y-\tilde{t}\right\| \leq 10\beta$ and $\mathbb{P}_w[w\in\theta] \geq \frac{\delta}{3N} \geq \delta\cdot\Theta(\sqrt{\delta}\beta)^m$.

□

Lemma E.2. *Consider a well-modeled distribution and a linear measurement model. Suppose we have a (τ,δ)-unconditional sampler for the distribution, where $\tau < \frac{c\delta\beta}{\sqrt{m+\log(1/\delta)}}$ for a sufficiently small constant $c>0$. Then rejection sampling (Algorithm 1) gives a $(\tau,2\delta)$-posterior sampler using at most $\frac{\log(1/\delta)}{\delta^2}(\frac{O(1)}{\beta\sqrt{\delta}})^m$ samples .*

Proof. Let $\mathcal{P}$ be the distribution that couples true distribution $\mathcal{D}$ over (x,y) and the output distribution of the posterior sampler $\widehat{p}_{|y}$. Rigorously, we define $\mathcal{P}$ over $(x,\widehat{x},y) \in X\times X\times Y$ with density $p^{\mathcal{P}}$ such that $p^{\mathcal{P}}(x,y) = p^{\mathcal{D}}(x,y)$, $p^{\mathcal{P}}(\widehat{x}\mid y) = \widehat{p}_{|y}(\widehat{x})$. Similarly, we let $\widetilde{\mathcal{P}}$ over $(x,\widehat{x},y)\in\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^\alpha$ be the joint distribution between the unconditional sampler over $(x,\widehat{x})$ and the measurement process $\mathcal{D}$ over (x,y). Then by the definition of unconditional samplers, we have

$$\mathbb{P}_{x,\hat{x}\sim\widetilde{\mathcal{P}}}\left[\|x-\widehat{x}\|\geq\tau\right] \leq \delta.$$

Therefore, to prove the correctness of the algorithm, we only need to show that there exists a $\widehat{\mathcal{P}}$ over $(x,\widehat{x},y)$ such that $\widehat{\mathcal{P}}(\widehat{x}\mid y) = \widehat{p}_{|y}(\widehat{x})$ and $\mathsf{TV}(\widehat{\mathcal{P}},\widetilde{\mathcal{P}}) \leq \delta$. By Lemma H.9,

$$\mathbb{P}_{\widetilde{\mathcal{P}}'}\left[\|Ax-y\|^2 \geq 4\beta^2(m+\log\frac{1}{\delta})\right] \leq \frac{\delta}{4}.$$

Therefore, we define $\widetilde{\mathcal{P}}'$ as $\widetilde{\mathcal{P}}$ conditioned on $\|x-\widehat{x}\| < \tau$ and $\frac{\|Ax-y\|^2}{2\beta^2} \leq 2(m+\log\frac{1}{\delta})$. Then we have

$$\mathsf{TV}(\widetilde{\mathcal{P}},\widetilde{\mathcal{P}}') \leq \frac{3\delta}{2}.$$

Algorithm correctness. We have

$$\widehat{p}_{|y}(\widehat{x}) = \frac{p^{\widetilde{\mathcal{P}}'}(\widehat{x})\cdot e^{\frac{-\|A\widehat{x}-y\|^2}{2\beta^2}}}{\int p^{\widetilde{\mathcal{P}}'}(\widehat{x})\cdot e^{\frac{-\|A\widehat{x}-y\|^2}{2\beta^2}}\,\mathrm{d}\widehat{x}} = \frac{\int p^{\widetilde{\mathcal{P}}'}(x,\widehat{x})\cdot e^{\frac{-\|A\widehat{x}-y\|^2}{2\beta^2}}\,\mathrm{d}x}{\int p^{\widetilde{\mathcal{P}}'}(\widehat{x})\cdot e^{\frac{-\|A\widehat{x}-y\|^2}{2\beta^2}}\,\mathrm{d}\widehat{x}},$$

Then we define

$$r(\widehat{x}) := \frac{\int p^{\widetilde{\mathcal{P}}'}(x,\widehat{x})\cdot e^{\frac{-\|Ax-y\|^2}{2\beta^2}}\,\mathrm{d}x}{\int p^{\widetilde{\mathcal{P}}'}(x,\widehat{x})\cdot e^{\frac{-\|A\widehat{x}-y\|^2}{2\beta^2}}\,\mathrm{d}x}.$$

Conditioned on $\|x-\widehat{x}\|\leq\tau$ and $\frac{\|Ax-y\|^2}{2\beta^2} \leq 2(m+\log\frac{1}{\delta})$, we have

$$|\log r(\widehat{x})| \leq \sup_x \frac{|\|Ax-y\|^2 - \|A\widehat{x}-y\|^2|}{2\beta^2}$$

$$\leq \frac{\tau^2 \|A\|_2^2 + 2\tau \|A\|_2 \|Ax - y\|}{2\beta^2}$$
$$\lesssim \frac{\tau^2}{\beta^2} + \frac{\tau\sqrt{m + \log(1/\delta)}}{\beta}.$$

By our setting of τ, we have $1 - \delta/8 < r(\widehat{x}) < 1 + \delta/8$.

So we have

$$\int p^{\widetilde{\mathcal{P}}'}(\widehat{x}) \cdot e^{\frac{-\|A\widehat{x}-y\|^2}{2\beta^2}} \,\mathrm{d}\widehat{x} = \int p^{\widetilde{\mathcal{P}}'}(x, \widehat{x}) \cdot e^{\frac{-\|A\widehat{x}-y\|^2}{2\beta^2}} \,\mathrm{d}x \,\mathrm{d}\widehat{x} = \left(1 \pm \frac{\delta}{8}\right) \int p^{\widetilde{\mathcal{P}}'}(x, \widehat{x}) \cdot e^{\frac{-\|Ax-y\|^2}{2\beta^2}} \,\mathrm{d}x \,\mathrm{d}\widehat{x}.$$

Hence,

$$\begin{aligned}
\widehat{p}_{|y}(\widehat{x}) &= \frac{r(\widehat{x}) \cdot \int p^{\widetilde{\mathcal{P}}'}(x, \widehat{x}) e^{\frac{-\|Ax-y\|^2}{2\beta^2}} \,\mathrm{d}x}{(1 \pm \frac{\delta}{8}) \int p^{\widetilde{\mathcal{P}}'}(x, \widehat{x}) e^{\frac{-\|Ax-y\|^2}{2\beta^2}} \,\mathrm{d}x \,\mathrm{d}\widehat{x}} \\
&= \frac{r(\widehat{x}) \cdot \int p^{\widetilde{\mathcal{P}}'}(x, \widehat{x}) p^{\widetilde{\mathcal{P}}'}(y \mid x) \,\mathrm{d}x}{(1 \pm \frac{\delta}{8}) p^{\widetilde{\mathcal{P}}'}(y)} \\
&= \left(1 \pm \frac{\delta}{4}\right) r(\widehat{x}) \int p^{\widetilde{\mathcal{P}}'}(x, \widehat{x} \mid y) \,\mathrm{d}x \\
&= \left(1 \pm \frac{\delta}{2}\right) p^{\widetilde{\mathcal{P}}'}(\widehat{x} \mid y).
\end{aligned}$$

Finally, we have

$$\int \left| p^{\widetilde{\mathcal{P}}'}(\widehat{x} \mid y) - \widehat{p}_{|y}(\widehat{x}) \right| \mathrm{d}\widehat{x} \,\mathrm{d}p^{\widetilde{\mathcal{P}}'}(y) = \int \left| \left(1 \pm \frac{\delta}{2}\right) p^{\widetilde{\mathcal{P}}'}(\widehat{x} \mid y) - p^{\widetilde{\mathcal{P}}'}(\widehat{x} \mid y) \right| \mathrm{d}\widehat{x} \,\mathrm{d}p^{\widetilde{\mathcal{P}}'}(y) \leq \frac{\delta}{2}.$$

This implies that

$$\mathsf{TV}(\widehat{\mathcal{P}}_{\widehat{x}}, \widetilde{\mathcal{P}}_{\widehat{x}}) = \mathsf{TV}(\widehat{\mathcal{P}}_{\widehat{x}}, \widetilde{\mathcal{P}}'_{\widehat{x}}) + \mathsf{TV}(\widetilde{\mathcal{P}}'_{\widehat{x}}, \widetilde{\mathcal{P}}_{\widehat{x}}) \leq \frac{\delta}{4} + \frac{\delta}{2} \leq \frac{3\delta}{4}.$$

Hence,

$$\mathbb{P}_{x,\widehat{x} \sim \widehat{\mathcal{P}}} \left[\|x - \widehat{x}\| \geq \tau \right] \leq \frac{3\delta}{4} + \delta \leq \frac{7\delta}{4}.$$

Running time. Now we prove that for most y For $y \in Y$, for each round, the acceptance probability $q(y)$ each round is that

$$\begin{aligned}
q(y) &= \int p^{\widetilde{\mathcal{P}}'}(\widehat{x}) e^{-\frac{\|y-A\widehat{x}\|^2}{2\gamma^2}} \,\mathrm{d}\widehat{x} \\
&= (1 \pm \frac{\delta}{8}) \int p^{\widetilde{\mathcal{P}}'}(x, \widehat{x}) \cdot e^{\frac{-\|Ax-y\|^2}{2\beta^2}} \,\mathrm{d}x \,\mathrm{d}\widehat{x} \\
&\geq \frac{1}{2} \int p^{\mathcal{X}}(x) \cdot e^{\frac{-\|Ax-y\|^2}{2\beta^2}} \,\mathrm{d}x \\
&= \frac{1}{2} \mathop{\mathbb{E}}_{x \sim \mathcal{X}} \left[e^{-\frac{\|Ax-y\|^2}{2\beta^2}} \right] \\
&\geq \frac{1}{2} \mathbb{P}_{x \sim \mathcal{X}} \left[\|Ax - y\| \leq 10\sqrt{m + \log(1/\delta)}\beta \right] \cdot e^{-\frac{100(m+\log(1/\delta))\beta^2}{2\beta^2}}
\end{aligned}$$

$$= \frac{1}{2}\mathbb{P}_{x\sim\mathcal{X}}\left[\|Ax-y\| \le 10\sqrt{m+\log(1/\delta)}\beta\right]\cdot\delta e^{-50m}$$

By Lemma H.11, $\mathbb{E}_{x\sim\mathcal{X}}[\|Ax\|_2^2] = O(m)$. By Lemma E.1, we have that for $1-\delta/8$ probability over y, for some $c>0$,

$$\mathbb{P}_{x\sim\mathcal{X}}\left[\|Ax-y\| \le 10\sqrt{m+\log(1/\delta)}\beta\right] \ge (c\beta)^m\cdot\delta^{m/2+1}.$$

Therefore, for some $c>0$,

$$\mathbb{P}_{y\sim\mathcal{Y}}\left[q(y)\ge(c\beta)^m\cdot\delta^{m/2+2}\right]\ge 1-\frac{\delta}{8}.$$

Hence, for some $C>0$,

$$\mathbb{P}\left[\text{Rejection sampling terminates in } \frac{\log(1/\delta)}{\delta^2}\left(\frac{C}{\beta\sqrt{\delta}}\right)^m \text{ rounds}\right]\ge 1-\frac{\delta}{4}.$$

□

Theorem 1.7 (Upper bound)**.** *Let $C>1$ be a constant. Consider an $O(C)$-well-modeled distribution and a linear measurement model with $\beta > \frac{1}{d^C}$. When $\delta > \frac{1}{d^C}$, rejection sampling of the diffusion process gives a $(\frac{1}{d^C},\delta)$-posterior sampler that takes $\mathrm{poly}(d)(\frac{O(1)}{\beta\sqrt{\delta}})^m$ time.*

Proof. Theorem 1.4 suggests that for an $O(C)$-well-modeled distribution, a $\mathrm{poly}(d)$ time $(\frac{1}{d^{3C}},\frac{1}{2d^C})$-unconditional sampler exists. Since

$$\frac{1}{d^{3C}} < o\left(\frac{\frac{1}{2d^C}\cdot\frac{1}{d^C}}{\sqrt{d}}\right) < o\left(\frac{\delta\beta^2}{\sqrt{m+\log(1/\delta)}}\right).$$

By lemma E.2, a $(\frac{1}{d^{3C}},\frac{1}{d^C})$-posterior sampler exists using $\frac{\log(1/\delta)}{\delta^2}(\frac{O(1)}{\beta\sqrt{\delta}})^m \le \mathrm{poly}(d)(\frac{O(1)}{\beta\sqrt{\delta}})^m$ samples. Since generating each sample costs $\mathrm{poly}(d)$ time. The total time is $\mathrm{poly}(d)(\frac{O(1)}{\beta\sqrt{\delta}})^m$. □

F Well-Modeled Distributions Have Accurate Unconditional Samplers

Notation. For the purposes of this section, we let $\widetilde{s}_t = s_{\sigma^2}$ denote the score at time t.

Definition F.1 (Forward and Reverse SDE)**.** *For distribution q_0 over $\mathbb{R}^d$, consider the Variance Exploding (VE) Forward SDE, given by*

$$dx_t = dB_t, \quad x_0\sim q_0$$

where B_t is Brownian motion, so that $x_t \sim x_0+\mathcal{N}(0,tI_d)$. Let q_t be the distribution of x_t.

There is a VE Reverse SDE associated with the above Forward SDE given by

$$dx_{T-t} = \widetilde{s}_{T-t}(x_{T-t})+dB_t \tag{12}$$

for $x_T\sim q_T$.

Theorem F.2 (Unconditional Sampling Theorem, Implied by [BBDD24], adapted from [GPPX23])**.** *Let q be a distribution over $\mathbb{R}^d$ with second moment $m_2^2 = \mathbb{E}_{x\sim q}\left[\|x\|^2\right]$ between $\frac{1}{\mathrm{poly}(d)}$ and $\mathrm{poly}(d)$. Let $q_t = q * \mathcal{N}(0, tI_d)$ be the $\sqrt{t}$-smoothed version of q, with corresponding score $\widetilde{s}_t$. Suppose $T = d^C$. For any $\gamma > 0$, there exist $N = \widetilde{O}\left(\frac{d}{\varepsilon^2}\log^2\frac{1}{\gamma}\right)$ discretization times $0 = t_0 < \cdots < t_N \le T - \gamma$ such that, given score approximations h_{T-t_k} of $\widetilde{s}_{T-t_k}$ that satisfy*

$$\mathop{\mathbb{E}}_{x\sim q_{T-t_k}}\left[\|\widetilde{s}_{T-t_k} - h_{T-t_k}\|^2\right] \lesssim \frac{\varepsilon^2}{C\cdot(T-t_k)\cdot\log\frac{d}{\gamma}}$$

for sufficiently large constant C, then, the discretization of the VE Reverse SDE defined in (12) using the score approximations can sample from a distribution $\varepsilon + \frac{1}{d^{C/2}}$ close in TV to a distribution γm_2-close in 2-Wasserstein to q in N steps.

Theorem 1.4 (Unconditional Sampling for Well-Modeled Distributions)**.** *For an $O(C)$-well-modeled distribution p, the discretized reverse diffusion process with approximate scores gives a $\left(\frac{1}{d^C}, \frac{1}{d^C}\right)$-unconditional sampler (as defined in Definition 1.3) for any constant $C > 0$ in $\mathrm{poly}(d)$ time.*

Proof. The definition of a well-modeled distribution gives that, for every $\frac{1}{d^C} < \sigma < d^C$ there is an approximate score $\widehat{s}_\sigma$ such that

$$\mathop{\mathbb{E}}_{x\sim p_\sigma}\left[\|\widehat{s}_\sigma(x) - s_\sigma(x)\|^2\right] < \frac{1}{d^C\sigma^2}$$

and $\widehat{s}_\sigma$ can be computed by a $\mathrm{poly}(d)$-parameter neural network with $\mathrm{poly}(d)$ bounded weights. Here p_σ is the σ-smoothed version of p with score s_σ.

Then, by Theorem F.2, this means that the discretized reverse diffusion process can use the $\widehat{s}_\sigma$ to produce a sample $\widehat{x}$ from a distribution $\widehat{p}$ that is $\frac{1}{d^{C/3}}$ close in TV to a distribution $\frac{1}{d^{C/3}}$ close in 2-Wasserstein. This means there exists a coupling between $\widehat{x}\sim\widehat{p}$ and $x\sim p$ such that

$$\mathbb{P}\left[\|\widehat{x} - x\| > \frac{1}{d^{C/6}}\right] < \frac{1}{d^{C/6}}$$

The claim follows via reparameterization. □

G Cryptographic Hardness

A one-way function f is a function such that every polynomial-time algorithm fails to find a pre-image of a random output of f with high probability.

Definition G.1. *A function $f : \{\pm1\}^n \to \{\pm1\}^{m(n)}$ is one-way if, for any polynomial-time algorithm $\mathcal{A}$, any constant $c > 0$, and all sufficiently large n,*

$$\mathbb{P}_{x\sim\mathcal{U}_n}\left[f(\mathcal{A}(f(x))) = f(x)\right] < n^{-c}.$$

Lemma G.2. *If a one-way function $f : \{\pm1\}^n \to \{\pm1\}^{m(n)}$ exists, then for any $\frac{1}{\mathrm{poly}(n)} \le l(n) \le \mathrm{poly}(n)$, there exists a one-way function $g : \{\pm1\}^n \to \{\pm1\}^{l(n)}$.*

Proof. For $l(n) > m(n)$, we just need to pad $l(n) - m(n)$ 1's at the end of the output, i.e.,

$$g(x) := (f(x), 1^{l(n)-m(n)}).$$

For $\frac{1}{\text{poly}(n)} \le l(n) < m(n)$, for each n, there exists a constant $c < 1$ such that $l(n) = m(n^c)$. Then we can satisfies the requirement by defining

$$g(x) := f(\text{first } n^c \text{ bits of } x).$$

□

Lemma G.3. *Every circuit $f : \{\pm 1\}^n \to \{\pm 1\}^{m(n)}$ of $\text{poly}(n)$ size can be simulated by a ReLU network with $\text{poly}(n)$ parameters and constant weights.*

Proof. In the realm of $\{+1, -1\}$, -1 corresponds to True and $+1$ corresponds to False. We can use a layer of neurons to translate it to $\{0, 1\}$ first, where 1 corresponds to True and -1 corresponds to False. We will translate $\{0, 1\}$ back to $\{+1, -1\}$ when output.

Now we only need to show that the logic operation ($\neg$, $\wedge$, $\vee$) in each gate of the circuit can be simulated by a constant number of neurons with constant weights in ReLU network when the input is in $\{0, 1\}^n$:

- For each AND ($\wedge$) gate, we use $\text{ReLU}(\sum(y_i - 1) + 1)$ to calculate $\bigwedge y_i$.
- For each OR ($\vee$) gate, we use $\text{ReLU}(1 - \text{ReLU}(1 - \sum y_i))$ to calculate $\bigvee y_i$.
- For each NOT ($\neg$) gate , we use $\text{ReLU}(1 - y_i)$ to calculate $\neg y_i$.

It is easy to verify that for $\{0, 1\}$ input, the output of each neuron-simulated gate will remain in $\{0, 1\}^n$ and equal to the result of the logical operation. □

Then the next corollary directly follows.

Corollary G.4. *Every one-way function can be computed by a ReLU network with $\text{poly}(n)$ parameters, and constant weights.*

H Utility Results

Lemma H.1. *Let p_σ be some σ-smoothed distribution with score s_σ. For any $\varepsilon \le \sigma$,*

$$\mathop{\mathbb{E}}_{x\sim p_\sigma} \sup_{|c|\le\varepsilon} s'_\sigma(x + c)^2 \lesssim \frac{1}{\sigma^4}$$

Proof. Draw $x \sim p_\sigma$, and let $z \sim N(0, \sigma^2)$ be independent of x. By Lemma H.3,

$$s_\sigma(x) = \mathop{\mathbb{E}}_{z|x}\left[\frac{z}{\sigma^2}\right].$$

Moreover, by Corollary H.4,

$$s_\sigma(x + c) = \frac{\mathbb{E}_{z|x}\left[e^{\frac{2cz-c^2}{2\sigma^2}}\left(\frac{z-c}{\sigma^2}\right)\right]}{\mathbb{E}_{z|x}\left[e^{\frac{2cz-c^2}{2\sigma^2}}\right]} = \frac{\mathbb{E}_{z|x}\left[e^{cz/\sigma^2}\left(\frac{z-c}{\sigma^2}\right)\right]}{\mathbb{E}_{z|x}[e^{cz/\sigma^2}]}$$

Taking the derivative with respect to c, since $(a/b)' = (a'b - ab')/b^2$,

$$s'_\sigma(x + c) = \frac{\mathbb{E}_{z|x}\left[e^{cz/\sigma^2}\left(\frac{z^2-zc-\sigma^2}{\sigma^4}\right)\right]\mathbb{E}_{z|x}[e^{cz/\sigma^2}] - \mathbb{E}_{z|x}\left[e^{cz/\sigma^2}\left(\frac{z-c}{\sigma^2}\right)\right]\mathbb{E}_{z|x}\left[\frac{z}{\sigma^2}e^{cz/\sigma^2}\right]}{\mathbb{E}_{z|x}[e^{cz/\sigma^2}]^2}$$

$$= \frac{\mathbb{E}_{z|x}\left[e^{cz/\sigma^2}\left(\frac{z^2-\sigma^2}{\sigma^4}\right)\right]\mathbb{E}_{z|x}[e^{cz/\sigma^2}] - \mathbb{E}_{z|x}\left[e^{cz/\sigma^2}\frac{z}{\sigma^2}\right]^2}{\mathbb{E}_{z|x}[e^{cz/\sigma^2}]^2}$$

$$\leq \frac{\mathbb{E}_{z|x}[e^{\varepsilon z/\sigma^2}\frac{z^2}{\sigma^4}]}{\mathbb{E}_{z|x}[e^{\varepsilon z/\sigma^2}]} \tag{13}$$

Now we take the supremum over all $|c| \leq \varepsilon$, and take the expectation of this quantity over x to get the desired moment:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x}\left[\sup_{|c|\leq\varepsilon} s'_\sigma(x+c)^2\right] &\leq \mathop{\mathbb{E}}_{x}\left[\sup_{|c|\leq\varepsilon}\frac{\mathbb{E}_{z|x}[e^{cz/\sigma^2}\frac{z^2}{\sigma^4}]^2}{\mathbb{E}_{z|x}[e^{cz/\sigma^2}]^2}\right] \\
&\leq \mathop{\mathbb{E}}_{x}\left[\left(\sup_{|c|\leq\varepsilon}\mathop{\mathbb{E}}_{z|x}\left[e^{cz/\sigma^2}\frac{z^2}{\sigma^4}\right]^2\right)\left(\sup_{|c|\leq\varepsilon}\mathop{\mathbb{E}}_{z|x}\left[e^{cz/\sigma^2}\right]^{-2}\right)\right] \\
&\leq \sqrt{\mathop{\mathbb{E}}_{x}\left[\sup_{|c|\leq\varepsilon}\mathop{\mathbb{E}}_{z|x}\left[e^{cz/\sigma^2}\frac{z^2}{\sigma^4}\right]^4\right]\mathop{\mathbb{E}}_{x}\left[\sup_{|c|\leq\varepsilon}\mathop{\mathbb{E}}_{z|x}\left[e^{cz/\sigma^2}\right]^{-4}\right]}
\end{aligned} \tag{14}$$

The last inequality here follows from Cauchy-Schwarz. For the first term of equation 14, we have

$$\mathop{\mathbb{E}}_{x}\left[\sup_{|c|\leq\varepsilon}\mathop{\mathbb{E}}_{z|x}\left[e^{cz/\sigma^2}\frac{z^2}{\sigma^4}\right]^4\right] \leq \mathop{\mathbb{E}}_{x}\mathop{\mathbb{E}}_{z|x}\left[(e^{\varepsilon z/\sigma^2}+e^{-\varepsilon z/\sigma^2})\frac{z^2}{\sigma^4}\right] =: g(x)$$

We compute the 4th moment of this term directly:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x}[g(x)^4] &= \mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{z|x}\left[(e^{\varepsilon z/\sigma^2}+e^{\varepsilon z/\sigma^2})\frac{z^2}{\sigma^4}\right]^4\right] \\
&\leq \mathop{\mathbb{E}}_{z}\left[(e^{\varepsilon z/\sigma^2}+e^{-\varepsilon z/\sigma^2})^4\frac{z^8}{\sigma^{16}}\right] \\
&\leq \sqrt{\mathop{\mathbb{E}}_{z}[(e^{\varepsilon z/\sigma^2}+e^{-\varepsilon z/\sigma^2})^8]\mathop{\mathbb{E}}_{z}\left[\frac{z^{16}}{\sigma^{32}}\right]} \\
&\leq \sqrt{\mathop{\mathbb{E}}_{z}[2^8(e^{8\varepsilon z/\sigma^2}+e^{-8\varepsilon z/\sigma^2})]\mathop{\mathbb{E}}_{z}\left[\frac{z^{16}}{\sigma^{32}}\right]} \\
&\lesssim \sqrt{e^{32\varepsilon^2/\sigma^2}\cdot\frac{1}{\sigma^{16}}} = \frac{e^{16\varepsilon^2/\sigma^2}}{\sigma^8}
\end{aligned} \tag{15}$$

For the second term of equation 14,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x}\left[\sup_{|c|\leq\varepsilon}\mathop{\mathbb{E}}_{z|x}\left[e^{cz/\sigma^2}\right]^{-4}\right] &\leq \mathop{\mathbb{E}}_{x}\left[\sup_{|c|\leq\varepsilon}\mathop{\mathbb{E}}_{z|x}\left[e^{-4cz/\sigma^2}\right]\right] && \text{by Jensen's} \\
&\leq \mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{z|x}\left[e^{4\varepsilon|z|/\sigma^2}\right]\right] \\
&\leq \mathop{\mathbb{E}}_{z}\left[e^{4\varepsilon z/\sigma^2}+e^{-4\varepsilon z/\sigma^2}\right] \\
&= 2e^{\frac{1}{2}\sigma^2\cdot(4\varepsilon/\sigma^2)^2} = 2e^{8\varepsilon^2/\sigma^2}
\end{aligned} \tag{16}$$

So, putting equations 16 and 15 into equation 14, we get

$$\mathop{\mathbb{E}}_{x}\left[\sup_{|c|\le\varepsilon} s'_\sigma(x+c)^2\right] \le \sqrt{2e^{8\varepsilon^2/\sigma^2}\cdot\frac{e^{16\varepsilon^2/\sigma^2}}{\sigma^8}}$$

Now, by assumption, $\varepsilon \le \sigma$. So, we finally get that

$$\mathop{\mathbb{E}}_{x}\left[\sup_{|c|\le\varepsilon} s'_\sigma(x+c)^2\right] \lesssim \sqrt{\frac{1}{\sigma^8}} = \frac{1}{\sigma^4}$$

□

Lemma H.2. *Let p be a distribution over $\mathbb{R}$, and let $p_\sigma = p * N(0,\sigma^2)$ have score s_σ. If $\gamma \le \sigma/4$, then,*

$$\mathbb{P}\left[\sup_{y\in[x-\gamma,x+\gamma]} s(y) \ge t\right] \le e^{-\sigma t}$$

Proof. From Corollary H.8, we have

$$\mathbb{E}\left[\sup_{y\in[x-\gamma,x+\gamma]} s(y)^k\right] \le \frac{k^k 15^k}{\sigma^k}$$

So, we have

$$\mathbb{P}\left[\sup_{y\in[x-\gamma,x+\gamma]} s(x)^k \ge t^k\right] \le \frac{\mathbb{E}\left[\sup_{y\in[x-\gamma,x+\gamma]} s(x)^k\right]}{t^k} \le \left(\frac{15k}{t\sigma}\right)^k$$

Setting $k = \log\frac{1}{\delta}$, we get

$$\mathbb{P}\left[\sup_{y\in[x-\gamma,x+\gamma]} |s(x)| \ge \frac{15}{e\sigma}\log\frac{1}{\delta}\right] \le \delta$$

□

For the following Lemmas, If p is a distribution over $\mathbb{R}$ and has score s, define the Fisher information $\mathcal{I}$ as

$$\mathcal{I} := \mathop{\mathbb{E}}_{x\sim p}[s^2(x)]$$

Lemma H.3 (Lemma A.1 from [GLPV22])**.** *Let p be a distribution over $\mathbb{R}$, and let $p_\sigma = p * N(0,\sigma^2)$ have score s_σ. Let (x,y,z) be the joint distribution such that $y\sim p$, $z\sim N(0,\sigma^2)$ are independent, and $x = y+z$. For all $\varepsilon > 0$,*

$$\frac{p(x+\varepsilon)}{p(x)} = \mathop{\mathbb{E}}_{z|x}\left[e^{\frac{2\varepsilon z-\varepsilon^2}{2\sigma^2}}\right] \text{ and } s_\sigma(x) = \mathop{\mathbb{E}}_{z|x}\left[\frac{z}{\sigma^2}\right]$$

Corollary H.4. *Let p be a distribution over $\mathbb{R}$, and let $p_\sigma = p * N(0,\sigma^2)$ have score s_σ.*

$$s_\sigma(x+\varepsilon) = \frac{\mathbb{E}_{z|x}\left[e^{\varepsilon z/\sigma^2}\left(\frac{z-\varepsilon}{\sigma^2}\right)\right]}{\mathbb{E}_{z|x}[e^{\varepsilon z/\sigma^2}]}$$

Proof. This proof is given in Lemma A.2 of [GLPV22], and is reproduced here for convenience and completeness, since a statement in the middle of their proof is what we use.

By Lemma H.3, we have

$$\frac{p_\sigma(x+\varepsilon)}{p_\sigma(x)} = \mathop{\mathbb{E}}_{z|x}\left[e^{\frac{2\varepsilon z-\varepsilon^2}{2\sigma^2}}\right]$$

Taking the derivative with respect to ε, we have

$$\frac{p'_\sigma(x+\varepsilon)}{p_\sigma(x)} = \mathbb{E}_{z|x}\left[e^{\frac{2\varepsilon z-\varepsilon^2}{2\sigma^2}}\left(\frac{z-\varepsilon}{\sigma^2}\right)\right]$$

So,

$$\begin{aligned} s_\sigma(x+\varepsilon) = \frac{p'_\sigma(x+\varepsilon)}{p_\sigma(x+\varepsilon)} &= \frac{p'_\sigma(x+\varepsilon)}{p_\sigma(x)}\frac{p_\sigma(x)}{p_\sigma(x+\varepsilon)} \\ &= \frac{\mathbb{E}_{z|x}\left[e^{\frac{2\varepsilon z-\varepsilon^2}{2\sigma^2}}\left(\frac{z-\varepsilon}{\sigma^2}\right)\right]}{\mathbb{E}_{z|x}\left[e^{\frac{2\varepsilon z-\varepsilon^2}{2\sigma^2}}\right]} = \frac{\mathbb{E}_{z|x}\left[e^{\varepsilon z/\sigma^2}\left(\frac{z-\varepsilon}{\sigma^2}\right)\right]}{\mathbb{E}_{z|x}[e^{\varepsilon z/\sigma^2}]} \end{aligned}$$

□

Lemma H.5 (Lemma 3.1 from [GLPV22])**.** *Let p be a distribution over $\mathbb{R}$ and let $p_\sigma = p * N(0,\sigma^2)$ have Fisher information $\mathcal{I}_\sigma$. Then, $\mathcal{I}_\sigma \le \frac{1}{\sigma^2}$.*

Lemma H.6 (Lemma B.3 from [GLPV22])**.** *Let p be a distribution over $\mathbb{R}$ and let $p_\sigma = p * N(0,\sigma^2)$ have score s_σ and Fisher information $\mathcal{I}_\sigma$. If $|\gamma| \le \sigma/2$, then*

$$\mathbb{E}[s^2(x+\gamma)] \le \mathcal{I}_\sigma + O\left(\frac{\gamma}{\sigma}\mathcal{I}_\sigma\sqrt{\log\frac{1}{\sigma^2\mathcal{I}_\sigma}}\right)$$

Lemma H.7 (Lemma A.6 from [GLPV22])**.** *Let p be a distribution over $\mathbb{R}$ and let $p_\sigma = p * N(0,\sigma^2)$ have score s_σ and Fisher information $\mathcal{I}_\sigma$. Then, for $k \ge 3$ and $|\gamma| \le \sigma/2$,*

$$\mathbb{E}[|s_\sigma(x+\gamma)|^k] \le \frac{k!}{2}(15/\sigma)^{k-2}\max\left(\mathbb{E}[s_\sigma^2(x+\gamma)], \mathcal{I}_\sigma\right)$$

Corollary H.8. *Let p be a distribution over $\mathbb{R}$ and let $p_\sigma = p * N(0,\sigma^2)$ have score s_σ. Then, for $k \ge 3$ and $|\gamma| \le \sigma/2$,*

$$\mathbb{E}[|s_\sigma(x+\gamma)|^k] \le \left(\frac{15k}{\sigma}\right)^k$$

Proof. Consider the continuous function $f(x) = x\sqrt{\log\frac{1}{\sigma^2 x}}$. This function is only defined on $0 < x \le 1/\sigma^2$. We have

$$f'(x) = \frac{2\log\frac{1}{\sigma^2 x} - 1}{2\sqrt{\log\frac{1}{\sigma^2 x}}}.$$

Setting this equal to zero gives $x = \frac{1}{\sigma^2\sqrt{e}}$. $f(\frac{1}{\sigma^2\sqrt{e}}) = \frac{1}{\sigma^2\sqrt{2e}}$. Since $f(1/\sigma^2) = 0$ and $\lim_{x\to 0^+} f(x) = 0$, we have this is the maximum value of the function. Further, we know by Lemma H.5 that $I_\sigma \le 1/\sigma^2$. So, along with the fact that $|\gamma| \le \sigma/2$, we have

$$\frac{\gamma}{\sigma}I_\sigma\sqrt{\log\frac{1}{\sigma^2 I_\sigma}} \lesssim \frac{1}{\sigma^2}$$

Therefore, from Lemma H.6, and using Lemma H.5 again, we get

$$\mathbb{E}[s^2(x+\gamma)] \le \mathcal{I}_\sigma + O\left(\frac{\gamma}{\sigma} I_\sigma \sqrt{\log \frac{1}{\sigma^2 I_\sigma}}\right) \lesssim \frac{1}{\sigma^2}$$

Finally, we can plug this into Lemma H.7 to get

$$\begin{aligned}\mathbb{E}[|s_\sigma(x+\gamma)|^k] &\le \frac{k!}{2}(15/\sigma)^{k-2} \max\left(\mathbb{E}[s_\sigma^2(x+\gamma)], I_\sigma\right) \\ &\lesssim k^k \frac{15^{k-2}}{\sigma^{k-2}} \cdot \frac{1}{\sigma^2} \le k^k \frac{15^k}{\sigma^k}\end{aligned}$$

□

Lemma H.9 (Laurent-Massart Bounds[LM00])**.** *Let $v \sim \mathcal{N}(0, I_n)$. For any $t > 0$,*

$$\mathbb{P}[\|v\|^2 - n \ge 2\sqrt{nt} + 2t] \le e^{-t}.$$

Lemma H.10 (See [MRT18, Lemma 6.27])**.** *There exist $\Theta(R/\varepsilon)^d$ d-dimensional balls of radius ε that cover $\{x \in \mathbb{R}^d \mid \|x\|_2 \le R\}$.*

Lemma H.11. *Let p be a distribution over $\mathbb{R}^d$ with covariance Σ such that $\|\Sigma\| \lesssim 1$, and let $A \in \mathbb{R}^{m \times d}$ be a matrix with $\|A\| \le 1$. Then*

$$\mathop{\mathbb{E}}_{x \sim p}[\|Ax\|^2] \lesssim m.$$

Proof. Note the expectation of the squared norm $\|Ax\|^2$ can be expressed as:

$$\mathbb{E}_{x \sim p}[\|Ax\|^2] = \operatorname{trace}(A^T A \Sigma).$$

Given that $\|A\| \le 1$, the singular values of A are at most 1. Hence, the matrix $A^T A$, which represents the sum of squares of these singular values, will have its trace (sum of eigenvalues) bounded by m:

$$\operatorname{trace}(A^T A) \le m.$$

Hence, given that $\|\Sigma\| \lesssim 1$, we have :

$$\mathbb{E}_{x \sim p}[\|Ax\|^2] = \operatorname{trace}(A^T A \Sigma) \le \|\Sigma\| \cdot \operatorname{trace}(A^T A) \lesssim m.$$

□