Instance-Optimal Compressed Sensing via Posterior Sampling

Ajil Jalal [1] **Sushrut Karmalkar** [2] **Alexandros G. Dimakis** [1] **Eric Price** [2]

Abstract

We characterize the measurement complexity of compressed sensing of signals drawn from a known prior distribution, even when the support of the prior is the entire space (rather than, say, sparse vectors). We show for Gaussian measurements and *any* prior distribution on the signal, that the posterior sampling estimator achieves near-optimal recovery guarantees. Moreover, this result is robust to model mismatch, as long as the distribution estimate (e.g., from an invertible generative model) is close to the true distribution in Wasserstein distance. We implement the posterior sampling estimator for deep generative priors using Langevin dynamics, and empirically find that it produces accurate estimates with more diversity than MAP.

1. Introduction

The goal of compressed sensing is to recover a structured signal from a relatively small number of linear measurements. The setting of such linear inverse problems has numerous and diverse applications ranging from Magnetic Resonance Imaging (Lustig et al., 2008; 2007), neuronal spike trains (Hegde et al., 2009) and efficient sensing cameras (Duarte et al., 2008). Estimating a signal in $\mathbb{R}^n$ would in general require n linear measurements, but because real-world signals are structured—i.e., compressible—one is often able to estimate them with $m \ll n$ measurements.

Formally, we would like to estimate a "signal" $x^* \in \mathbb{R}^n$ from noisy linear measurements,

$$y = Ax^* + \xi$$

for a measurement matrix $A \in \mathbb{R}^{m\times n}$ and noise vector $\xi \in \mathbb{R}^m$. We will focus on the i.i.d. Gaussian setting, where $A_{ij} \sim \mathcal{N}(0, \frac{1}{m})$ and $\xi_i \sim \mathcal{N}(0, \frac{\sigma^2}{m})$, and one would like to recover $\widehat{x}$ from (A, y) such that

$$\|x^* - \widehat{x}\| \leq C\sigma \tag{1}$$

with high probability for some constant C. When x^* is k-sparse, this was shown by Candés, Romberg, and Tao (Candes et al., 2006) to be possible for m at least $O(k \log \frac{n}{k})$.

Over the past 15 years, compressed sensing has been extended in a wide variety of remarkable ways, including by generalizing from sparsity to other signal structures, such as those given by trees (Chen & Huang, 2012), graphs (Xu et al., 2011), manifolds (Chen et al., 2010; Xu & Hassibi, 2008), or deep generative models (Bora et al., 2017; Asim et al., 2019). These are all essentially frequentist approaches to the problem: they define a small *set* of "structured" signals x, and ask for recovery of every such signal.

Such set-based approaches have limitations. For example, (Bora et al., 2017) uses the structure given by a deep generative model $G : \mathbb{R}^k \to \mathbb{R}^n$; with $O(kd \log n)$ measurements for d-layer networks, accurate recovery is guaranteed for every signal x^* near the range of G. But this completely ignores the *distribution* over the range. Generative models like Glow (Kingma & Dhariwal, 2018) and pixelRNN (Oord et al., 2016) have seed length $k = n$ and range equal to the entire $\mathbb{R}^n$. Yet because these models are designed to approximate reality, and real images can be compressed, we know that compressed sensing is possible in principle.

This leads to the question: Given signals drawn from some *distribution* R, can we characterize the number of linear measurements necessary for recovery, with both upper and lower bounds? Such a Bayesian approach has previously been considered for sparsity-inducing product distributions (Aeron et al., 2010; Zhou et al., 2014) but not general distributions.

Second, suppose that we don't know the real distribution R, but instead have an approximation P of R (e.g., from a GAN or invertible generative model). In what sense should P approximate R for compressed sensing with good guarantees to be possible?

Code and models available at: `https://github.com/ajiljalal/code-cs-fairness`. [1]University of Texas at Austin, Department of Electrical and Computer Engineering [2]University of Texas at Austin, Department of Computer Science. Correspondence to: Ajil Jalal <ajiljalal@utexas.edu>, Sushrut Karmalkar <sushrutk@cs.utexas.edu>, Alexandros G. Dimakis <dimakis@austin.utexas.edu>, Eric Price <ecprice@cs.utexas.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Figure 1: Reconstruction results on FFHQ for Gaussian measurements (here $n = 256 \times 256 \times 3 = 196,608$ pixels), using an NCSNv2 model. Each column shows the reconstruction obtained as the number of measurements m varies. The top row shows reconstructions by MAP, the middle row shows reconstruction by Deep-Decoder, and the bottom row shows reconstructions by Langevin dynamics, which is the practical implementation of our proposed posterior sampling estimator.

1.1. Contributions.

Our main theorem is that posterior sampling is a near optimal recovery algorithm for *any* distribution. Moreover, it is sufficient to learn the distribution in Wasserstein distance.

Theorem 1.1. *Let R be an arbitrary distribution over an ℓ_2 ball of radius r. Suppose that there exists an algorithm that uses an* arbitrary measurement matrix $A \in \mathbb{R}^{m \times n}$ *with noise level σ and finds a reconstruction $\widehat{x}$ such that*

$$\|x^* - \widehat{x}\| \lesssim \sigma \text{ with probability } \geq 1 - \delta.$$

Then posterior sampling (see Definition 1.3) with respect to R using $m' \geq O\left(m \log\left(1 + \frac{mr^2\|A\|_\infty^2}{\sigma^2}\right) + \log \frac{1}{\delta}\right)$ Gaussian measurements of noise level σ will output $\widehat{x}$ satisfying

$$\|x^* - \widehat{x}\| \lesssim \sigma \text{ with probability } \geq 1 - O(\delta).$$

Moreover, the same holds for posterior sampling with respect to any distribution P satisfying $\mathcal{W}_p(R, P) \lesssim \sigma\delta^{1/p}$ for some $p \geq 1$.

This theorem comprises three main contributions: the introduction of posterior sampling as *a new algorithm* for recovery with a generative prior; an *upper bound* on the sample complexity of the algorithm in terms of an approximate covering number that we introduce; and an *instance-optimal lower bound* in terms of the same approximate covering number that (unlike previous lower bounds in compressed sensing) applies to *any* distribution of input signals.

Contribution 1: Approximate covering numbers. The covering number of a set is the smallest number of balls that can cover the entire set. Standard compressed sensing is closely tied to the covering number $N_\eta(S)$ of the set S of possible signals x; for example, the set of unit-norm k-sparse vectors has $\log N_\eta = \Theta(k \log \frac{n}{k})$, which is precisely why Candés, Romberg, and Tao use this many linear measurements to achieve (1).

For distributions, we need a different concept of covering number. As a motivating example, consider a distribution R induced by a trivial *linear* generative model, $x = \Sigma z$ where $z \sim \mathcal{N}(0, I_n)$ and Σ is a fixed $n \times n$ matrix. Further suppose the singular values σ_i of Σ are Zipfian, so $\sigma_i = 1/i$. In this case, R's support is $\mathbb{R}^n$, so covering the entire support of R is infeasible. Instead we could denote by $\mathrm{Cov}_{\eta,0.01}(R)$ the minimum number of $\eta-$radius balls needed to cover 99%

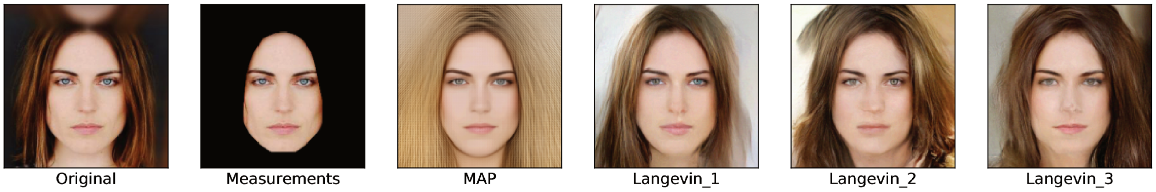


Figure 2: Reconstruction results for inpainting on CelebA-HQ using Glow. The first column shows the original image, second column shows the measurements by removing the hair and background, the third column shows reconstruction by MAP, and the last three columns show samples from posterior sampling via Langevin dynamics. MAP produces the same washed out image all the time, whereas posterior sampling produces images with diversity.

of R. An elementary calculation shows

$$\log \mathrm{Cov}_{\eta,0.01}(R) = \Theta(1/\eta^2),$$

which is (up to constants) precisely the number of linear measurements you need to estimate x to within η.

We show that an *approximate covering number* characterizes the measurement complexity of compressed sensing a general distribution R, and that recovery by *posterior sampling* achieves this bound.

Definition 1.2. *Let R be a distribution on $\mathbb{R}^n$. For some parameters $\eta > 0, \delta \in [0,1]$, we define the (η,δ)-*approximate covering number *of R as*

$$\mathrm{Cov}_{\eta,\delta}(R) := \min\left\{k : R\left[\cup_{i=1}^{k}\mathcal{B}(x_i,\eta)\right] \geq 1-\delta, x_i \in \mathbb{R}^n\right\},$$

where $\mathcal{B}(x,\eta)$ is the ℓ_2 ball of radius η centered at x.

When $\delta = 0$, this is $N_\eta(\operatorname{supp} R)$, the standard covering number of the support of R. Having $\delta > 0$ allows meaningful results for full-support distributions that are concentrated on smaller sets. This also generalizes our previous results in (Bora et al., 2017), which depend on the covering numbers of low-dimensional generative models.

Contribution 2: Recovery algorithm. The recovery algorithm we consider is posterior sampling:

Definition 1.3. *Given an observation y, the* posterior sampling *recovery algorithm with respect to P outputs $\widehat{x}$ according to the posterior distribution $P(\cdot \mid y)$.*

Contribution 3: Sample complexity upper bound. Our main positive result is that posterior sampling achieves the guarantees of equation (1) for *general* distributions R, with $O(\log \mathrm{Cov}_{\sigma,\delta}(R))$ measurements. Not only this, but the algorithm is robust to model mismatch: posterior sampling with respect to $P \neq R$ still works, as long as P and R are close in Wasserstein distance:

Theorem 1.4 (Upper bound)**.** *Let P, R be distributions with $\mathcal{W}_1(P,R) \leq \sigma$. Let $x^* \sim R$, let y be Gaussian measurements with noise level σ, and let $\widehat{x} \sim P(\cdot|y)$. For any $\eta \geq \sigma$, with*

$$m \geq O(\log \mathrm{Cov}_{\eta,0.01}(R))$$

measurements, the guarantee $\|\widehat{x} - x^*\| \leq C\eta$ is satisfied for some universal constant C with 97% probability over the signal x, measurement matrix A, noise ξ, and recovery algorithm $\widehat{x}$.

Contribution 4: Sample complexity lower bound. Our second main result lower bounds the sample complexity for *any* distribjution. This is, to our knowledge, the first lower bound for compressed sensing that applies to arbitrary distributions R. Most lower bounds in the area are minimax, and only apply to specific "hard" distributions R (Price & Woodruff, 2011; Candes & Davenport, 2013; Iwen & Tewfik, 2010); the closest result we are aware of is (Aeron et al., 2010), which characterizes product distributions.

Theorem 1.5 (Lower bound)**.** *Let R be any distribution over an ℓ_2 ball of radius r, and consider any method to achieve $\|\widehat{x} - x^*\| \leq \eta$ with 99% probability, using an* arbitrary measurement matrix *$A \in \mathbb{R}^{m\times n}$ with noise level σ. This must have*

$$m \geq \frac{C'}{\log(1 + \frac{mr^2\|A\|_\infty^2}{\sigma^2})} \log \mathrm{Cov}_{C'\eta,0.04}(R).$$

for some constant $C' > 0$.

Note that Theorem 1.4 and 1.5 directly give Theorem 1.1. For more precisely stated and general versions of these results, including dependence on the failure probability δ, see Theorems 3.4 and 4.1.

1.2. Related Work

Generative priors have shown great promise in compressed sensing and other inverse problems, starting with (Bora et al., 2017), who generalized the theoretical framework of compressive sensing and restricted eigenvalue conditions (Tibshirani, 1996; Donoho, 2006; Bickel et al., 2009; Candes, 2008; Hegde et al., 2008; Baraniuk & Wakin, 2009; Baraniuk et al., 2010; Eldar & Mishali, 2009) for signals lying on the range of a deep generative model (Goodfellow et al., 2014; Kingma & Welling, 2013).

Lower bounds in (Kamath et al., 2019; Liu & Scarlett, 2019; Jalali & Yuan, 2019) established that the sample complexities in (Bora et al., 2017) are order optimal. The approach in (Bora et al., 2017) has been generalized to tackle different inverse problems such as robust compressed sensing (Jalal et al., 2020), phase retrieval (Hand et al., 2018; Aubin et al., 2019; Jagatap & Hegde, 2019), blind image deconvolution (Asim et al., 2018), seismic inversion (Mosser et al., 2020), one-bit recovery (Qiu et al., 2019; Liu et al., 2020), and blind demodulation (Hand & Joshi, 2019). Alternate algorithms for reconstruction include sparse deviations from generative models (Dhar et al., 2018), task-aware compressed sensing (Kabkab et al., 2018), PnP (Pandit et al., 2019; Fletcher et al., 2018b;a), iterative projections (Mardani et al., 2018), OneNet (Rick Chang et al., 2017) and Deep Decoder (Heckel & Hand, 2018; Heckel & Soltanolkotabi, 2020). The complexity of optimization algorithms using generative models have been analyzed for ADMM (Gómez et al., 2019), PGD (Hegde, 2018), layer-wise inversion (Lei et al., 2019), and gradient descent (Hand & Voroninski, 2017). Experimental results in (Asim et al., 2019; Whang et al., 2020; Lindgren et al., 2020) show that invertible models have superior performance in comparison to low dimensional models. See (Ongie et al., 2020) for a more detailed survey on deep learning techniques for compressed sensing. A related line of work has explored learning-based approaches to tackle classical problems in algorithms and signal processing (Aamand et al., 2019; Indyk et al., 2019; Metzler et al., 2017; Hsu et al., 2018).

Lower bounds for ℓ_2/ℓ_2 recovery of sparse vectors can be found in (Scarlett & Cevher, 2016; Price & Woodruff, 2011; Aeron et al., 2010; Iwen & Tewfik, 2010; Candes & Davenport, 2013), and these are related to the lower bound in (1.5). The closest result is that of (Aeron et al., 2010), which characterizes the probability of error and ℓ_2 error of the reconstruction via covering numbers of the probability distribution. Their approach uses the rate distortion function of a scalar random variable $\mathbf{x}$, and provides guarantees for the product measure generated via an i.i.d. sequence of $\mathbf{x}$. A Shannon theory for compressed sensing was pioneered by (Wu & Verdú, 2012; Wu, 2011). The $\delta-$Minkowski dimension of a probability measure used in (Wu & Verdú, 2012; Wu, 2011; Pesin, 2008) can be derived from our $(\varepsilon, \delta)-$covering number by taking the limit $\varepsilon \to 0$. (Reeves & Gastpar, 2012) contains a related theory of rate distortion for compressed sensing. There is also related work in the statistical physics community under different assumptions on the signal structure (Zdeborová & Krzakala, 2016; Barbier et al., 2019).

2. Background and Notation

In this section, we introduce a few concepts that we will use throughout the paper. $\|\cdot\|$ refers to the ℓ_2 norm unless specified otherwise. The metric we use to quantify the similarity between distributions is the Wassertein distance. For two probability distributions μ, ν supported on Ω, and for any $p \geq 1$, the Wasserstein-p (Villani, 2008; Arjovsky et al., 2017) and Wasserstein-∞ (Champion et al., 2008) distances are defined as:

$$\mathcal{W}_p(\mu,\nu) := \inf_{\gamma\in\Pi(\mu,\nu)} \left(\mathop{\mathbb{E}}_{(u,v)\sim\gamma} [\|u-v\|^p] \right)^{1/p},$$

$$\mathcal{W}_\infty(\mu,\nu) := \inf_{\gamma\in\Pi(\mu,\nu)} \left(\gamma\text{-}\operatorname*{ess\,sup}_{(u,v)\in\Omega^2} \|u-v\| \right),$$

where $\Pi(\mu,\nu)$ denotes the set of joint distributions whose marginals are μ,ν. The above definition says that if $\mathcal{W}_\infty(\mu,\nu) \leq \varepsilon$, and $(u,v)\sim\gamma$, then $\|u-v\| \leq \varepsilon$ almost surely.

We say that y is generated from x^* by a Gaussian measurement process with m measurements and noise level σ, if $y = Ax^* + \xi$ where $\xi \sim \mathcal{N}(0, \frac{\sigma^2}{m} I_m)$ and $A \in \mathbb{R}^{m\times n}$ with $A_{ij} \sim \mathcal{N}(0, 1/m)$.

3. Upper Bound

3.1. Two-Ball Case

For simplicity, we will first demonstrate our proof techniques in the simple setting where $R = P$, the measurements are noiseless, and the ground truth distribution P is supported on two disjoint balls (illustrated in Figure 3). In this example, two η radius balls can cover the whole space, so the parameters in Theorem 1.4 will be $\sigma = 0$ and $\mathrm{Cov}_{\eta,0}(P) = 2$. Applying Theorem 1.4 on P tells us that a constant number of measurements is sufficient for posterior sampling to get $O(\eta)$-close to the ground truth, i.e., to return an element of the correct ball. We will now prove this claim.

Let $B_0, B_{\tilde{x}}$ denote η-radius balls centered at $0, \tilde{x} \in \mathbb{R}^n$ respectively. Suppose $P = 0.5P_0 + 0.5P_1$, where P_0, P_1, are uniform distributions on $B_0, B_{\tilde{x}}$. The centers of the balls are separated by a distance $d \gg \eta$.

The ground truth x^* will be sampled from P. For a fixed matrix $A \in \mathbb{R}^{m\times n}$ with $m \ll n$, let the noiseless measurements be $y = Ax^*$ and let H_0, H_1, denote the distributions over $\mathbb{R}^m$ induced by the projection of P_0, P_1, by A.

Given A, y, we sample the reconstruction $(\widehat{x})$ according to the posterior density

$$p(\widehat{x}|y) = c_y p_0(\widehat{x}|y) + (1-c_y) p_{\tilde{x}}(\widehat{x}|y),$$

where c_y is the posterior probability that y is a projection of x^* drawn from the P_0 component of P. Note that c_y depends on y.

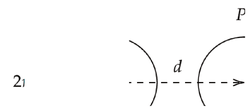


Figure 3: Illustrative example for the upper bound. The signal x^* is drawn from a mixture of two well-separated balls. The observations $y = Ax^*$ are then drawn from a mixture of two distributions H_0, H_1 that may overlap. The probability that posterior sampling outputs something from the wrong ball is proportional to the (shaded) overlap between these distributions, which is atmost $1 - TV(H_0, H_1)$.

Since the balls $B_0 \& B_{\tilde{x}}$ are well separated, the ground truth and the reconstruction are far apart if and only if they lie in different balls, i.e., $\{x^* \in B_0, \widehat{x} \in B_{\tilde{x}}\}$, or vice versa. It turns out quite generally that the probability of this event is bounded by how similar the distributions H_0, H_1 are:

Lemma 3.1. *For $c \in [0, 1]$, let $H := (1-c)H_0 + cH_1$ be a mixture of two absolutely continuous distributions H_0, H_1 admitting densities h_0, h_1. Let y be a sample from the distribution H, such that $y|z^* \sim H_{z^*}$ where $z^* \sim Bernoulli(c)$.*

Define $\widehat{c}_y = \frac{ch_1(y)}{(1-c)h_0(y)+ch_1(y)}$, and let $\widehat{z}|y \sim Bernoulli(\widehat{c}_y)$ be the posterior sampling of z^* given y. Then we have

$$\Pr_{z^*, y, \widehat{z}}[z^* = 0, \widehat{z} = 1] \leq 1 - TV(H_0, H_1).$$

The proof of this, as well as all parts of the upper bound, can be found in Appendix A.

In our current example, this gives us

$$\Pr[x^* \in B_0, \widehat{x} \in B_{\tilde{x}}] \leq 1 - TV(H_0, H_1) \text{ and}$$
$$\Pr[x^* \in B_{\tilde{x}}, \widehat{x} \in B_0] \leq 1 - TV(H_0, H_1).$$

Since B_0 and $B_{\tilde{x}}$ are balls of radius η, a union bound of the above two probabilities gives:

$$\begin{aligned} \Pr\left[\|x^* - \widehat{x}\| > 2\eta\right] \leq & \Pr\left[x^* \in B_0, \widehat{x} \in B_{\tilde{x}}\right] + \\ & \Pr\left[x^* \in B_{\tilde{x}}, \widehat{x} \in B_0\right], \\ \leq & 2\left(1 - TV\left(H_0, H_1\right)\right). \end{aligned} \tag{2}$$

If A is a Gaussian random matrix, the Johnson-Lindenstrauss (JL) Lemma tells us that it will preserve distances between vectors with high probability . This does not necessarily mean that every point in the distribution P will be preserved in norm. Still, we show that, since P_0 and P_1 have well-separated supports, their projected distributions H_0 & H_1 have very high TV distance. This also holds more generally, between any distribution on a ball and any distribution far from the ball and in the presence of noise.

Lemma 3.2. *Let y be generated from x^* by a Gaussian measurement process with noise level σ. For a fixed $\tilde{x} \in \mathbb{R}^n$, and parameters $\eta > 0, c \geq 4e^2$, let P_{out} be a distribution supported on the set*

$$S_{\tilde{x},out} := \{x \in \mathbb{R}^n : \|x - \tilde{x}\| \geq c(\eta + \sigma)\}.$$

Let $P_{\tilde{x}}$ be a distribution which is supported within an $\eta-$radius ball centered at $\tilde{x}$.

For a fixed A, let $H_{\tilde{x}}$ denote the distribution of y when $x^* \sim P_{\tilde{x}}$. Let H_{out} denote the corresponding distribution of y when $x^* \sim P_{out}$. Then we have:

$$\mathop{\mathbb{E}}_{A}\left[TV(H_{\tilde{x}}, H_{out})\right] \geq 1 - 4e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}.$$

By Markov's inequality, the expectation bound also gives a high probability bound over A.

For our current example, the above result implies that with probability $1 - e^{-\Omega(m)}$ over A, we have

$$TV(H_0, H_1) \geq 1 - e^{-\Omega(m)}. \tag{3}$$

Substituting equation (3) in equation (2), we have

$$\Pr\left[\|x^* - \widehat{x}\| > 2\eta\right] \leq 2e^{-\Omega(m)}.$$

This shows that posterior sampling will produce a reconstruction which is close to the ground truth with overwhelmingly high probability for the two-ball example.

3.2. Going beyond two balls

The two-ball example leaves three main questions unanswered:

1. How do we handle distributions over larger collections of balls?
2. How do we handle mismatch between the distribution of reality (R) and the model (P)?
3. How do we handle having a δ probability of lying outside any ball?

Unions of many balls. The first question is relatively easy to answer: if $\mathrm{Cov}_{\eta,0}(R) \leq e^{o(m)}$, you can cover R with a small number of balls, and essentially apply Lemma 3.2 with a union bound. There are a few details (e.g., Lemma 3.2

shows you will not confuse any ball with faraway balls, but you might confuse it with nearby balls) but solving them is straightforward. This shows that, if $P = R$ and $\log \mathrm{Cov}_{\eta,0}(R)$ is bounded, then posterior sampling works well with $1 - e^{-\Omega(m)}$ probability.

Distribution mismatch in $\mathcal{W}_\infty$. The above assumes we resample with respect to the true distribution R. But we only have a learned estimate P of R. We would like to show that observing samples from R and resampling according to P gives good results. We first show that resampling signals drawn from R with respect to P is not much worse than resampling signals drawn from P with respect to P, if P and R are close in $\mathcal{W}_\infty$.

Lemma 3.3. *Let R, P, denote arbitrary distributions over $\mathbb{R}^n$ such that $\mathcal{W}_\infty(R, P) \leq \varepsilon$.*

Let $x^* \sim R$ and $z^* \sim P$ and let y and u be generated from x^* and z^* via a Gaussian measurement process with m measurements and noise level σ. Let $\widehat{x} \sim P(\cdot|y, A)$ and $\widehat{z} \sim P(\cdot|u, A)$. For any $d > 0$, we have

$$\Pr_{x^*,A,\xi,\widehat{x}}[\|x^* - \widehat{x}\| \geq d + \varepsilon] \leq$$
$$e^{-\Omega(m)} + e^{\left(\frac{4\varepsilon(\varepsilon+2\sigma)m}{2\sigma^2}\right)} \Pr_{z^*,A,\xi,\widehat{z}}[\|z^* - \widehat{z}\| \geq d] .$$

The idea is that with σ Gaussian noise, measurements of a signal from R aren't too different in distribution from measurements of the corresponding nearby signal from P.

Now, if $\mathcal{W}_\infty(R, P) \ll \sigma$, we would be nearly done: Lemma 3.3 says the situation is within $e^{o(m)}$ of the $R = P$ case, which we already know gives accurate recovery with $O(\log \mathrm{Cov}_{\eta,0}(P))$ measurements.

Residual mass. There are just two main issues remaining: we want to depend on $\log \mathrm{Cov}_{\eta,\delta}$ rather than $\log \mathrm{Cov}_{\eta,0}$, and we only want to require a bound on $\mathcal{W}_1(R, P)$ not $\mathcal{W}_\infty(R, P)$. By Markov's inequality, these issues are very similar: we want to allow both R and P to have a small constant probability of behaving badly. To address this, we note the existence of two distributions R' and P', which are only δ-far in TV from R and P respectively, such that R' and P' do have a small cover & are close in $\mathcal{W}_\infty$. We show that, because posterior sampling would work with R' and P', it also works with R and P. This leads to our full upper bound:

Theorem 3.4. *Let $\delta \in [0, 1/4)$, $p \geq 1$, and $\varepsilon, \eta > 0$ be parameters. Let R, P be arbitrary distributions over $\mathbb{R}^n$ satisfying $\mathcal{W}_p(R, P) \leq \varepsilon$.*

Let $x^* \sim R$ and suppose y is generated by a Gaussian measurement process from x^* with noise level $\sigma \gtrsim \varepsilon/\delta^{1/p}$ and $m \geq O(\min(\log \mathrm{Cov}_{\eta,\delta}(R), \log \mathrm{Cov}_{\eta,\delta}(P)))$ measurements. Given y and the fixed matrix A, let $\widehat{x}$ output of posterior sampling with respect to P.

Then there exists a universal constant $c > 0$ such that with probability at least $1 - e^{-\Omega(m)}$ over A, ξ,

$$\Pr_{x^* \sim R, \widehat{x} \sim P(\cdot|y)}[\|x^* - \widehat{x}\| \geq c\eta + c\sigma] \leq 2\delta + 2e^{-\Omega(m)}.$$

Note that we can get a high-probability result by setting $p = \infty$: if $m \geq O(\log \mathrm{Cov}_{\eta,0}(R))$ and $\mathcal{W}_\infty(R, P) \leq \sigma$, the error is $O(\sigma + \eta)$ with $1 - e^{-\Omega(m)}$ probability.

4. Lower Bound

In the previous section, we showed, for any distribution R of signals, that $O(\log \mathrm{Cov}(R))$ measurements suffice for posterior sampling to recover most signals well. Now we show the converse: for any distribution of signals R, any algorithm for recovery must use $\Omega(\log \mathrm{Cov}(R))$ measurements.

Theorem 4.1. *Let R be a distribution supported on a ball of radius r in $\mathbb{R}^n$, and $x^* \sim R$. Let $y = Ax^* + \xi$, where A is any matrix, and $\xi \sim \mathcal{N}(0, \frac{\sigma^2}{m} I_m)$. Assuming $\delta < 0.1$, if there exists a recovery scheme that uses y and A as inputs and guarantees*

$$\|\hat{x} - x^*\| \leq O(\eta),$$

with probability $\geq 1 - \delta$, then we have

$$m \geq \frac{0.15}{\log\left(1 + \frac{mr^2\|A\|_\infty^2}{\sigma^2}\right)} \left(\log \mathrm{Cov}_{3\eta,4\delta}(R) + \log 6\delta - O(1)\right) .$$

If A is an i.i.d. Gaussian matrix where each element is drawn from $\mathcal{N}(0, 1/m)$, then the above bound can be improved to:

$$m \geq \frac{0.15}{\log\left(1 + \frac{r^2}{\sigma^2}\right)} \left(\log \mathrm{Cov}_{3\eta,4\delta}(R) + \log 6\delta - O(1)\right) .$$

This Theorem is proven using information theory, as an almost direct consequence of the following three Lemmas.

First, the measurement process reveals a limited amount of information:

Lemma 4.2. *Consider the setting of Theorem (4.1). If A is a deterministic matrix, we have*

$$I(y; x^*) \leq \frac{m}{2} \log\left(1 + \frac{mr^2\|A\|_\infty^2}{\sigma^2}\right).$$

If A is a Gaussian matrix, then $I(y; x^*|A) \leq \frac{m}{2} \log\left(1 + \frac{r^2}{\sigma^2}\right)$.

Second, since $x^* \to y \to \widehat{x}$ is a Markov chain, we can directly apply the Data Processing Inequality (Cover & Thomas, 2012).

Lemma 4.3. *Consider the setting of Theorem (4.1). If A is a deterministic matrix, we have $I(x^*; \widehat{x}) \leq I(y; x^*)$.*

If A is a random matrix, then $I(x^*; \widehat{x}) \leq I(y; x^* | A)$.

Finally, successful recovery must yield a large amount of information:

Lemma 4.4 (Fano variant)**.** *Let $(x, \widehat{x})$ be jointly distributed over $\mathbb{R}^n \times \mathbb{R}^n$, where $x \sim R$ and $\widehat{x}$ satisfies*

$$\Pr[\|x - \widehat{x}\| \leq \eta] \geq 1 - \delta.$$

Then for any $\tau \leq 1 - 3\delta, \delta < 1/3$, we have

$$0.99\tau(1 - 2\delta) \log \mathrm{Cov}_{3\eta, \tau+3\delta}(R) \leq I(x; \widehat{x}) + 1.98.$$

In order to complete the proof of Theorem 4.1, we need an additional counting argument to remove the extra τ term that appears in the left hand side of Lemma 4.4.

The proofs can be found in Appendix B.

5. Experiments

In this section we discuss our algorithm for posterior sampling, discuss why existing algorithms can fail, and show our empirical evaluation of posterior sampling versus baselines.

5.1. Datasets and Models

We perform our experiments on the CelebA-HQ (Liu et al., 2018; Karras et al., 2017) and FlickrFaces-HQ (Karras et al., 2019) datasets. For the CelebA dataset, we run experiments using a Glow generative model (Kingma & Dhariwal, 2018). For the FlickrFaces-HQ dataset, we use the NCSNv2 model (Song & Ermon, 2020). Both models have output size $256 \times 256 \times 3$. Details about our experiments are in Appendix C.

5.2. Langevin Dynamics

Glow trained on CelebA-HQ We first consider the Glow generative model, whose distribution P is induced by the random variable $G(z)$, where $G : \mathbb{R}^n \to \mathbb{R}^n$ is a fixed deterministic generative model, and $z \sim \mathcal{N}(0, I_n)$. Sampling from $p(z|y)$ is easier than sampling from $p(x|y)$, since it is easier to compute and we observe that sampling mixes quicker. Note that sampling $\widehat{z} \sim p(z|y)$ and setting $\widehat{x} = G(\widehat{z})$ is equivalent to sampling $\widehat{x} \sim p(x|y)$.

In order to sample from $p(z|y)$, we use *Langevin dynamics*, which samples from a given distribution by moving a random initial sample along a vector field given by the distribution. Langevin dynamics tells us that if we sample $z_0 \sim \mathcal{N}(0, 1)$, and run the following iterative procedure:

$$z_{t+1} \leftarrow z_t + \frac{\alpha_t}{2} \nabla_z \log p\left(z_t | y\right) + \sqrt{\alpha_t} \zeta_t, \quad \zeta_t \sim \mathcal{N}(0, I),$$

then $p(z|y)$ is the stationary distribution of z_t as $t \to \infty$ and $\alpha_t \to 0$. Unfortunately, this algorithm is slow to mix, as observed in (Song & Ermon, 2019). We instead use an annealed version of the algorithm, where in step t we pretend that $p(z \mid y)$ has noise scale $\sigma_t \geq \sigma$ instead of σ. This gives

$$\log p_t(z|y) = \left(-\frac{\|y - AG(z)\|^2}{2\sigma_t^2/m} - \frac{\|z\|^2}{2} \right) + \log c(y), \tag{4}$$

where $c(y)$ is a constant that depends only on y. Since we only care about the gradient of $\log p(z|y)$, we can ignore this constant $c(y)$. By taking a decreasing sequence of σ_t that approach the true value of σ, we can anneal Langevin dynamics and sample from $p(z|y)$. Please refer to Appendix C for more details about how σ_t varies.

NCSNv2 trained on FFHQ We also consider the NCSNv2 model, which takes as input the image x, and outputs $\nabla_x \log p(x)$. This model is designed such that sampling from its marginal involves running Langevin dynamics. Since we have access to $\nabla_x \log p(x)$, and if we know the functional form of $p(y|x)$, we can easily compute $\nabla_x \log p(x|y)$, and run Langevin dynamics via

$$x_{t+1} \leftarrow x_t + \frac{\alpha_t}{2} \nabla_x \log p\left(x_t | y\right) + \sqrt{\alpha_t} \zeta_t, \ \zeta_t \sim \mathcal{N}(0, I).$$

Notice that we can also run MAP using this model. This can be achieved by simply following the gradient, and not adding noise: $x_{t+1} \leftarrow x_t + \frac{\alpha_t}{2} \nabla_x \log p\left(x_t | y\right)$.

This model also requires annealing, and we follow the schedule prescribed by (Song & Ermon, 2020). Please see Appendix C for more details.

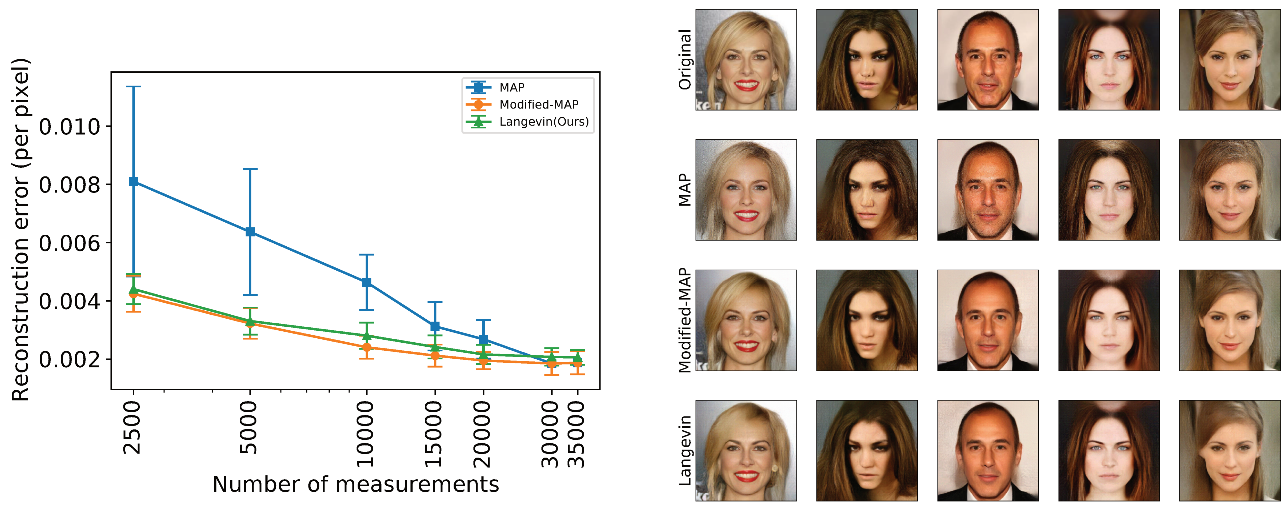
5.3. MAP and Modified-MAP

The most relevant baseline for our algorithm is MAP, which was shown to be state-of-the-art for compressed sensing using generative priors (Asim et al., 2019).

Given access to a generative model G such that the image $x = G(z)$, and $q(z)$ is the prior of z, the MAP estimate is

$$\widehat{z} := \arg\min_z \frac{\|y - AG(z)\|^2}{2\sigma^2/m} - \log q(z), \tag{5}$$

and set the estimate to be $\widehat{x} = G(\widehat{z})$. Typically, $q(z)$ is a standard Gaussian for many generative models. If one has



(a) $\|x^* - \widehat{x}\|^2/n$

(b) Reconstructions for $m = 20,000$ measurements.

Figure 4: We compare our algorithm with the MAP baseline on the CelebA-HQ dataset, where the number of pixels is $n = 256 \times 256 \times 3 = 196,608$. In Figure (a) we show a plot of the per-pixel reconstruction error as we vary the number of measurements m. In Figure (b) we show reconstructions obtained by each algorithm for $m = 20,000$ measurements. We show original images (top row), reconstructions by MAP (second row), Modified-MAP (third row), and Langevin dynamics (ours, bottom row). Note that MAP produces several artefacts that are not seen in Modified-MAP or Langevin dynamics. In these experiments, modified-MAP picks hyperparameters based on the reconstruction error evaluated on some validation images, while MAP and Langevin dynamics pick hyperparameters that maximize the posterior likelihood. Here MAP, modified-MAP, and Langevin dynamics all use the same Glow model.

access to $p(x)$, such as in NCSNv2 (Song et al., 2019), it is possible to also do MAP in x-space.

One may modify this algorithm and introduce hyperparameters for better reconstructions. We call such algorithms *modified-MAP*. For example, (Asim et al., 2019) introduce a parameter $\gamma > 0$ that weights the prior, and their estimate is

$$\widehat{z}_{modified} := \arg\min_{z} \|y - AG(z)\|^2 - \gamma \log q(z), \quad (6)$$

Other examples of hyper-parameters include early stopping to avoid "over-fitting" to the measurements, and choosing optimization parameters such that the reconstruction error is minimized on a validation set of images. Then these hyper-parameters are used for evaluating reconstruction error on a different test.

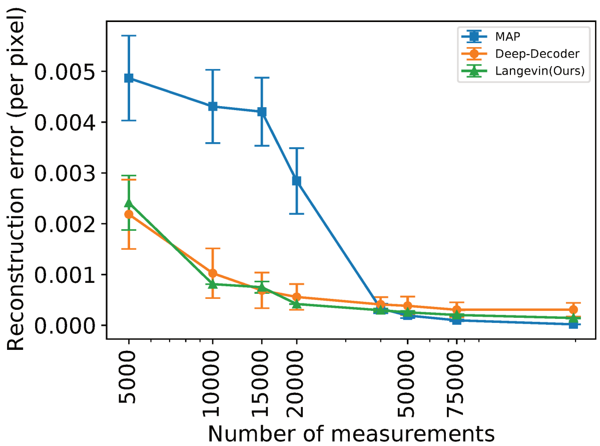
5.4. Experimental Results

MAP estimation does not work on general distributions: as an extreme example, if R is a mixture of some continuous distribution 99% of the time, and the all-zero image 1% of the time, it will always output the all-zero image, which is wrong 99% of the time. More generally, looking for high-likelihood *points* rather than *regions* means it prefers sharp but very narrow maxima to wide, but slightly shorter, maxima. Posterior sampling prefers the opposite. We now study this empirically.

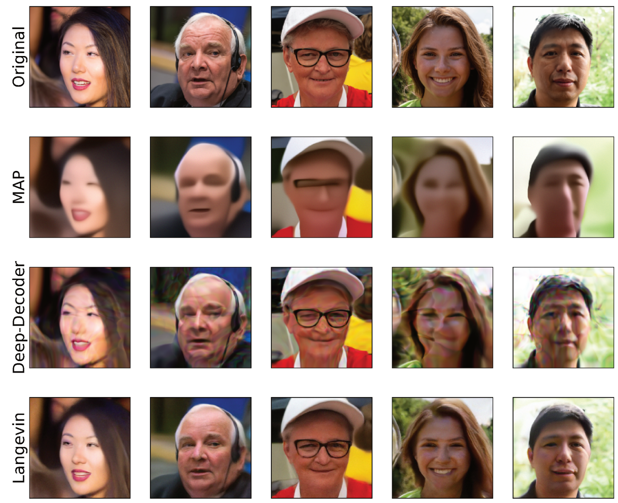
CelebA. In Figure 4, we show the performance of our proposed algorithm for compressed sensing on CelebA-HQ with Glow. The baselines we consider are MAP, and modified-MAP. MAP directly optimizes the objective defined in Eqn (5) while Modified-MAP optimizes (6). The MAP baseline in Figure 4 tries to maximize the posterior likelihood, and hence hyperparameters are selected so that the posterior is optimized. In contrast, what we term the modified-MAP algorithm was proposed by (Asim et al., 2019), and this algorithm picks hyperparameters that minimize reconstruction error on a holdout set of images. These hyperparameters are significantly worse at optimizing the MAP objective, but lead to more accurate recovered images, presumably due to some sort of implicit regularization. This modified-MAP method has shown to be state-of-the-art for compressed sensing on CelebA (Asim et al., 2019).

We find that our algorithm is competitive with respect to modified-MAP, and beats MAP when the measurements are $< 35,000$.

FFHQ. In Figure 5, we show the performance of our proposed algorithm for compressed sensing on FlickrFaces-HQ with the NCSNv2 generative model. We consider MAP and Deep-Decoder (Heckel & Hand, 2018) as the baselines. Note that the NCSNv2 model was designed for Langevin dynamics, and we adapt it to MAP. Hence, we choose the Deep-Decoder as a second baseline, as it has been shown to



(a) $\|x^* - \widehat{x}\|^2/n$



(b) Reconstructions for $m = 5,000$ measurements.

Figure 5: We compare our algorithm with the MAP and Deep-Decoder baselines on the FFHQ dataset, where the number of pixels is $n = 256 \times 256 \times 3 = 196,608$. Figure (a) plots per-pixel reconstruction error as we vary the number of measurements m. Figure (b) shows original images (top row), reconstructions by MAP (second row), Deep-Decoder (third row), and Langevin dynamics (bottom row). Langevin dynamics is the practical implementation of our proposed posterior sampling estimator. Note that although Deep Decoder and Langevin achieve similar value of reconstruction errors, Langevin produces images with higher perceptual quality, as can be seen in Figure (b).

match state-of-the-art (Asim et al., 2019).

We observe that for $m < 40,000$ measurements, Langevin dynamics beats MAP, and is competitive with Deep-Decoder. In Figure 1 we visually compare the reconstruction quality as the number of measurements increases. Note that although Langevin and Deep-Decoder have similar reconstruction errors in Fig 5a, the images in Fig 1 produced by Langevin dynamics have better perceptual quality. Also see Fig 5b for more examples of reconstructions at $m = 5,000$ measurements.

Inpainting. In order to highlight the difference in diversity between images produced by MAP and Langevin dynamics, we evaluate them on the inverse problem of inpainting missing pixels. As shown in Figure 2, when the hair and background of a ground truth image is removed, MAP produces a single "most likely" reconstruction, while Langevin produces diverse images that satisfy the measurements. Each column for Langevin dynamics in Figure 2 corresponds to a run starting from a random initial point. We do not observe any change in MAP reconstructions as we vary the initial point.

We believe that the MAP reconstruction, while in some sense a highly likely reconstruction, is abnormally "washed out" and indistinct; analogous to how zero is the most likely sample from $N(0, I_d)$, yet is extremely atypical of the distribution. We see this quantitatively in that the corresponding $\|z\|^2/n$ for MAP is 0.007, even though samples from R almost surely have $\|z\|^2/n \approx 1$, as do those of Langevin.

6. Conclusion

This paper studies the problem of compressed sensing a signal from a distribution R. We have shown that the measurement complexity is closely characterized by the log approximate covering number of R. Moreover, this recovery guarantee can be achieved by posterior sampling, even with respect to a distribution $P \neq R$ that is close in Wasserstein distance. Our experiments using Langevin dynamics to approximate posterior sampling match state-of-the-art recovery with a theoretically grounded algorithm.

This measurement complexity is inherent to the true distribution of images in the domain, and can't be improved. But perhaps it can be estimated: one open question is whether $\log \mathrm{Cov}_{\eta,\delta}(P)$ can be estimated or bounded when P is given by a neural network generative model.

7. Acknowledgements

Ajil Jalal and Alex Dimakis were supported by NSF Grants CCF 1934932, AF 1901292, 2008710, 2019844 the NSF IFML 2019844 award and research gifts by Western Digital, WNCG and MLL, computing resources from TACC and the Archie Straiton Fellowship. Sushrut Karmalkar was supported by a University Graduate Fellowship from UT Austin. Eric Price was supported by NSF Award CCF-1751040 (CAREER) and NSF IFML 2019844.

References

Aamand, A., Indyk, P., and Vakilian, A. (learned) frequency estimation algorithms under zipfian distribution. *arXiv preprint arXiv:1908.05198*, 2019.

Aeron, S., Saligrama, V., and Zhao, M. Information theoretic bounds for compressed sensing. *IEEE Transactions on Information Theory*, 56(10):5111–5130, 2010.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Asim, M., Shamshad, F., and Ahmed, A. Blind image deconvolution using deep generative priors. *arXiv preprint arXiv:1802.04073*, 2018.

Asim, M., Ahmed, A., and Hand, P. Invertible generative models for inverse problems: mitigating representation error and dataset bias. *arXiv preprint arXiv:1905.11672*, 2019.

Aubin, B., Loureiro, B., Baker, A., Krzakala, F., and Zdeborová, L. Exact asymptotics for phase retrieval and compressed sensing with random generative priors. *arXiv preprint arXiv:1912.02008*, 2019.

Baraniuk, R. G. and Wakin, M. B. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

Baraniuk, R. G., Cevher, V., Duarte, M. F., and Hegde, C. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 537–546. JMLR. org, 2017.

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

Candes, E. J. and Davenport, M. A. How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

Candes, E. J., Romberg, J. K., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

Champion, T., De Pascale, L., and Juutinen, P. The ∞-Wasserstein distance: Local solutions and existence of optimal transport maps. *SIAM Journal on Mathematical Analysis*, 40(1):1–20, 2008.

Chen, C. and Huang, J. Compressive sensing mri with wavelet tree sparsity. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1115–1123. Curran Associates, Inc., 2012.

Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D., and Carin, L. Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Transactions on Signal Processing*, 58(12):6140–6155, 2010.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.

Dhar, M., Grover, A., and Ermon, S. Modeling sparse deviations for compressed sensing using generative models. *arXiv preprint arXiv:1807.01442*, 2018.

Donoho, D. L. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Duarte, M. F., Davenport, M. A., Takhar, D., Laska, J. N., Sun, T., Kelly, K. F., and Baraniuk, R. G. Single-pixel imaging via compressive sampling. *IEEE signal processing magazine*, 25(2):83–91, 2008.

Eldar, Y. C. and Mishali, M. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009.

Fletcher, A. K., Pandit, P., Rangan, S., Sarkar, S., and Schniter, P. Plug-in estimation in high-dimensional linear inverse problems: A rigorous analysis. In *Advances in Neural Information Processing Systems*, pp. 7440–7449, 2018a.

Fletcher, A. K., Rangan, S., and Schniter, P. Inference in deep networks in high dimensions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1884–1888. IEEE, 2018b.

Gómez, F. L., Eftekhari, A., and Cevher, V. Fast and provable admm for learning with generative priors. *arXiv preprint arXiv:1907.03343*, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Hand, P. and Joshi, B. Global guarantees for blind demodulation with generative priors. In *Advances in Neural Information Processing Systems*, pp. 11531–11541, 2019.

Hand, P. and Voroninski, V. Global guarantees for enforcing deep generative priors by empirical risk. *arXiv preprint arXiv:1705.07576*, 2017.

Hand, P., Leong, O., and Voroninski, V. Phase retrieval under a generative prior. In *Advances in Neural Information Processing Systems*, pp. 9136–9146, 2018.

Heckel, R. and Hand, P. Deep decoder: Concise image representations from untrained non-convolutional networks. *arXiv preprint arXiv:1810.03982*, 2018.

Heckel, R. and Soltanolkotabi, M. Compressive sensing with un-trained neural networks: Gradient descent finds the smoothest approximation. *arXiv preprint arXiv:2005.03991*, 2020.

Hegde, C. Algorithmic aspects of inverse problems using generative models. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 166–172. IEEE, 2018.

Hegde, C., Wakin, M., and Baraniuk, R. G. Random projections for manifold learning. In *Advances in neural information processing systems*, pp. 641–648, 2008.

Hegde, C., Duarte, M. F., and Cevher, V. Compressive sensing recovery of spike trains using a structured sparsity model. In *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009.

Hsu, C.-Y., Indyk, P., Katabi, D., and Vakilian, A. Learning-based frequency estimation algorithms. 2018.

Indyk, P., Vakilian, A., and Yuan, Y. Learning-based low-rank approximations. In *Advances in Neural Information Processing Systems*, pp. 7400–7410, 2019.

Iwen, M. and Tewfik, A. Adaptive group testing strategies for target detection and localization in noisy environments. 2010.

Jagatap, G. and Hegde, C. Phase retrieval using untrained neural network priors. 2019.

Jalal, A., Liu, L., Dimakis, A. G., and Caramanis, C. Robust compressed sensing using generative models. *Advances in Neural Information Processing Systems*, 33, 2020.

Jalali, S. and Yuan, X. Solving linear inverse problems using generative models. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 512–516. IEEE, 2019.

Kabkab, M., Samangouei, P., and Chellappa, R. Task-aware compressed sensing with generative adversarial networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Kamath, A., Karmalkar, S., and Price, E. Lower bounds for compressed sensing with generative models. *arXiv preprint arXiv:1912.02938*, 2019.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pp. 10215–10224, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lei, Q., Jalal, A., Dhillon, I. S., and Dimakis, A. G. Inverting deep generative models, one layer at a time. In *Advances in Neural Information Processing Systems*, pp. 13910–13919, 2019.

Lindgren, E. M., Whang, J., and Dimakis, A. G. Conditional sampling from invertible generative models with applications to inverse problems. *arXiv preprint arXiv:2002.11743*, 2020.

Liu, Z. and Scarlett, J. Information-theoretic lower bounds for compressive sensing with generative models. *arXiv preprint arXiv:1908.10744*, 2019.

Liu, Z., Luo, P., Wang, X., and Tang, X. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15: 2018, 2018.

Liu, Z., Gomes, S., Tiwari, A., and Scarlett, J. Sample complexity bounds for 1-bit compressive sensing and binary stable embeddings with generative priors. *arXiv preprint arXiv:2002.01697*, 2020.

Lustig, M., Donoho, D., and Pauly, J. M. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

Lustig, M., Donoho, D. L., Santos, J. M., and Pauly, J. M. Compressed sensing mri. *IEEE signal processing magazine*, 25(2):72–82, 2008.

Mardani, M., Gong, E., Cheng, J. Y., Vasanawala, S. S., Zaharchuk, G., Xing, L., and Pauly, J. M. Deep generative adversarial neural networks for compressive sensing mri. *IEEE transactions on medical imaging*, 38(1):167–179, 2018.

Metzler, C., Mousavi, A., and Baraniuk, R. Learned d-amp: Principled neural network based compressive image recovery. In *Advances in Neural Information Processing Systems*, pp. 1772–1783, 2017.

Mosser, L., Dubrule, O., and Blunt, M. J. Stochastic seismic waveform inversion using generative adversarial networks as a geological prior. *Mathematical Geosciences*, 52(1): 53–79, 2020.

Ongie, G., Jalal, A., Metzler, C. A., Baraniuk, R. G., Dimakis, A. G., and Willett, R. Deep learning techniques for inverse problems in imaging. *arXiv preprint arXiv:2005.06001*, 2020.

Oord, A. v. d., Kalchbrenner, N., and Kavukcuoglu, K. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

Pandit, P., Sahraee-Ardakan, M., Rangan, S., Schniter, P., and Fletcher, A. K. Inference with deep generative priors in high dimensions. *arXiv preprint arXiv:1911.03409*, 2019.

Pesin, Y. B. *Dimension theory in dynamical systems: contemporary views and applications*. University of Chicago Press, 2008.

Polyanskiy, Y. and Wu, Y. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

Price, E. and Woodruff, D. P. (1+ eps)-approximate sparse recovery. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pp. 295–304. IEEE, 2011.

Qiu, S., Wei, X., and Yang, Z. Robust one-bit recovery via relu generative networks: Improved statistical rates and global landscape analysis. *arXiv preprint arXiv:1908.05368*, 2019.

Reeves, G. and Gastpar, M. The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing. *IEEE Transactions on Information Theory*, 58(5): 3065–3092, 2012.

Rick Chang, J., Li, C.-L., Poczos, B., Vijaya Kumar, B., and Sankaranarayanan, A. C. One network to solve them all–solving linear inverse problems using deep projection models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5888–5897, 2017.

Scarlett, J. and Cevher, V. Limits on support recovery with probabilistic models: An information-theoretic framework. *IEEE Transactions on Information Theory*, 63(1): 593–620, 2016.

Shannon, C. E. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

Song, G., Fan, Z., and Lafferty, J. Surfing: Iterative optimization over incrementally trained deep networks. In *Advances in Neural Information Processing Systems*, pp. 15008–15017, 2019.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pp. 11918–11930, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *arXiv preprint arXiv:2006.09011*, 2020.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Whang, J., Lei, Q., and Dimakis, A. G. Compressed sensing with invertible generative models and dependent noise. *arXiv preprint arXiv:2003.08089*, 2020.

Wu, Y. *Shannon theory for compressed sensing*. Citeseer, 2011.

Wu, Y. and Verdú, S. Optimal phase transitions in compressed sensing. *IEEE Transactions on Information Theory*, 58(10):6241–6263, 2012.

Xu, W. and Hassibi, B. Compressed sensing over the grassmann manifold: A unified analytical framework. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 562–567. IEEE, 2008.

Xu, W., Mallada, E., and Tang, A. Compressive sensing over graphs. In *2011 Proceedings IEEE INFOCOM*, pp. 2087–2095. IEEE, 2011.

Zdeborová, L. and Krzakala, F. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Zhou, Z., Liu, K., and Fang, J. Bayesian compressive sensing using normal product priors. *IEEE Signal Processing Letters*, 22(5):583–587, 2014.

A. Upper Bound Proofs

A.1. Proof of Lemma 3.1

Lemma 3.1. *For $c \in [0,1]$, let $H := (1-c)H_0 + cH_1$ be a mixture of two absolutely continuous distributions H_0, H_1 admitting densities h_0, h_1. Let y be a sample from the distribution H, such that $y|z^* \sim H_{z^*}$ where $z^* \sim Bernoulli(c)$.*

Define $\widehat{c}_y = \frac{ch_1(y)}{(1-c)h_0(y)+ch_1(y)}$, and let $\widehat{z}|y \sim Bernoulli(\widehat{c}_y)$ be the conditional resampling of z^* given y. Then we have

$$\Pr_{z^*,y,\widehat{z}}[z^* = 0, \widehat{z} = 1] \leq 1 - TV(H_0, H_1).$$

Proof. We have

$$\begin{aligned}\Pr_{z^*,y,\widehat{z}}[z^* = 0, \widehat{z} = 1] &= \Pr[z^* = 0] \mathop{\mathbb{E}}_{y\sim h_0, \widehat{z}|y}[1\{\widehat{z} = 1\}], &(8)\\ &= (1-c)\int h_0(y) \Pr[\widehat{z} = 1|y]dy. &(9)\end{aligned}$$

By definition, we have

$$\Pr[\widehat{z} = 1|y] = \frac{ch_1(y)}{(1-c)h_0(y) + ch_1(y)}.$$

Substituting, we have

$$\begin{aligned}\Pr_{z^*,y,\widehat{z}}[z^* = 0, \widehat{z} = 1] &= \int \frac{(1-c)h_0(y)ch_1(y)}{(1-c)h_0(y) + ch_1(y)}dy\\ &\leq \int \frac{(1-c)h_0(y)\cdot ch_1(y)}{\max\{(1-c)h_0(y), ch_1(y)\}}dy\\ &= \int \min\{(1-c)h_0(y), ch_1(y)\}dy\\ &\leq \int \min\{h_0(y), h_1(y)\}dy\\ &= (1 - TV(H_0, H_1)).\end{aligned}$$

□

A.2. Proof of Lemma 3.2

Lemma 3.2. *Let y be generated from x^* by a Gaussian measurement process with noise rate σ. For a fixed $\tilde{x} \in \mathbb{R}^n$, and parameters $\eta > 0, c \geq 4e^2$, let P_{out} be a distribution supported on the set*

$$S_{\tilde{x},out} := \{x \in \mathbb{R}^n : \|x - \tilde{x}\| \geq c(\eta + \sigma)\}.$$

Let $P_{\tilde{x}}$ be a distribution which is supported within an $\eta-$radius ball centered at $\tilde{x}$.

For a fixed A, let $H_{\tilde{x}}$ denote the distribution of y when $x^* \sim P_{\tilde{x}}$. Let H_{out} denote the corresponding distribution of y when $x^* \sim P_{out}$. Then we have:

$$\mathop{\mathbb{E}}_{A}[TV(H_{\tilde{x}}, H_{out})] \geq 1 - 4e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}.$$

Proof. In order to prove the lemma, it suffices to show that on the set

$$B := \{y \in \mathbb{R}^m : \|y - A\tilde{x}\| \leq \sqrt{c}\,(\eta + \sigma)\},$$

we have

$$\mathop{\mathbb{E}}_{A}[H_{out}(B)] \leq 2e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}, \tag{10}$$

$$\mathop{\mathbb{E}}_{A}[H_{\tilde{x}}(B)] \geq 1 - 2e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}. \tag{11}$$

Using the above bounds, we can conclude that

$$\mathop{\mathbb{E}}_{A}\left[TV(H_{out}, H_{\tilde{x}})\right] \geq \mathop{\mathbb{E}}_{A}[H_{\tilde{x}}(B)] - \mathop{\mathbb{E}}_{A}[H_{out}(B)] \geq 1 - 4e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}.$$

First we prove Equation (10).

Consider the joint distribution of y, A. We have

$$\begin{aligned} \mathop{\mathbb{E}}_{A}\left[H_{out}(B)\right] &= \mathop{\mathbb{E}}_{A}\left[\mathop{\mathbb{E}}_{x\sim P_{out}}\left[\mathcal{N}\left(Ax, \frac{\sigma^2}{m}I_m\right)(B)\right]\right], && (12)\\ &= \mathop{\mathbb{E}}_{x\sim P_{out}}\left[\mathop{\mathbb{E}}_{A}\left[\mathcal{N}(Ax, \sigma^2/m)(B)\right]\right], && (13) \end{aligned}$$

where the first line follows from the definition of H_{out} and the fact that x, A are independent. The last line follows by switching the order of integrating A, x. Here $\mathcal{N}(Ax, \sigma^2/m)(B)$ refers to the mass $\mathcal{N}(Ax, \sigma^2/m)$ places on B.

Consider a fixed $x \in S_{\tilde{x},out}$, that is, x lies in the support of P_{out} and satisfies $\|x - \tilde{x}\| \geq c(\eta + \sigma\sqrt{m})$. We split the above expectation into two conditions over the matrix A.

- Case 1: $\|Ax - A\tilde{x}\| \leq 2\sqrt{c}\,(\eta+\sigma)$. Since A is i.i.d. Gaussian, $A\,(x-\tilde{x})$ is distributed as $\mathcal{N}\left(0, \frac{\|x-\tilde{x}\|^2}{m}I_m\right)$. This gives

$$\begin{aligned} \Pr_{A}\left[\|Ax - A\tilde{x}\| < 2\sqrt{c}\,(\eta+\sigma)\right] &\leq \Pr_{A}\left[\|Ax - A\tilde{x}\| \leq \frac{2}{\sqrt{c}}\|x-\tilde{x}\|\right],\\ &\leq \frac{2}{\sqrt{m\pi}}\left(\frac{2e}{\sqrt{c}}\right)^m,\\ &= \frac{2}{\sqrt{m\pi}}e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)},\\ &\leq e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)} \quad \text{if } m > 1. \end{aligned}$$

  This implies

$$\begin{aligned} \mathop{\mathbb{E}}_{x\sim P_{out}}\left[\mathop{\mathbb{E}}_{A}\left[\mathcal{N}(Ax,\sigma^2/m)(B)1_{\|Ax-A\tilde{x}\|<2\sqrt{c}(\eta+\sigma)}\right]\right] &\leq \mathop{\mathbb{E}}_{x\sim P_{out}}\left[\mathop{\mathbb{E}}_{A}\left[1_{\|Ax-A\tilde{x}\|<2\sqrt{c}(\eta+\sigma)}\right]\right],\\ &= \mathop{\mathbb{E}}_{x\sim P_{out}}\left[\Pr_{A}\left[\|Ax - A\tilde{x}\| \leq 2\sqrt{c}\,(\eta+\sigma)\right]\right],\\ &\leq e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}. \end{aligned}$$

- Case 2: $\|Ax - A\tilde{x}\| > 2\sqrt{c}\,(\eta+\sigma)$.

  Recall the definition of $B := \{y \in \mathbb{R}^m : \|y - A\tilde{x}\| \leq \sqrt{c}\,(\eta+\sigma)\}$. For any $y \in B$, x in the support of P_{out} and for A such that $\|Ax - A\tilde{x}\| > 2\sqrt{c}\,(\eta+\sigma)$, we have

$$\|y - Ax\| \geq \|Ax - A\tilde{x}\| - \|y - A\tilde{x}\| \geq 2\sqrt{c}\,(\eta+\sigma) - \sqrt{c}\,(\eta+\sigma) = \sqrt{c}\,(\eta+\sigma)\,.$$

  For each x in the support of P_{out}, define the set $B_x := \{y \in \mathbb{R}^m : \|y - Ax\| \geq \sqrt{c}\,(\eta+\sigma)\}$. The above inequality gives $B \subseteq B_x$ for each x in the support of P_{out}. This gives

$$\mathcal{N}(Ax, \sigma^2)(B) \leq \mathcal{N}(Ax,\sigma^2)(B_x) \leq e^{-2\left(\sqrt{c}-1\right)^2 m} \leq e^{-\frac{mc}{2}}.$$

  where the last inequality follows by the definition of B_x and Gaussian concentration of $\mathcal{N}(Ax, \sigma^2)$ on the set B_x, and since $2\left(\sqrt{c}-1\right)^2 > \frac{c}{2}$ if $c \geq 4$.

Substituting the inequalities from Case 1 and Case 2 in Eqn (13), we have

$$\begin{aligned}\mathop{\mathbb{E}}_{A}\left[H_{out}(B)\right] &= \mathop{\mathbb{E}}_{x\sim P_{out}}\left[\mathop{\mathbb{E}}_{A}\left[\mathcal{N}(Ax,\sigma^2/m)(B)\right]\right],\\ &\leq e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}+e^{-\frac{cm}{2}},\\ &\leq 2e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)} \quad \text{if } c\geq 4e^2.\end{aligned}$$

This proves Eqn (10).

A similar proof can be used to show that

$$\mathop{\mathbb{E}}_{A}\left[H_{\tilde{x}}(B^c)\right]\leq 2e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}.$$

This proves Eqn (11).

Putting the two above inequalities together, we have

$$\mathop{\mathbb{E}}_{A} TV(H_{out},H_{\tilde{x}})\geq \mathop{\mathbb{E}}_{A}[H_{\tilde{x}}(B)]-\mathop{\mathbb{E}}_{A}[H_{out(B)}]\geq 1-4e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}.$$

This concludes the proof. □

A.3. Proof of Lemma A.1

Lemma A.1. *Let R, P be arbitrary distributions on $\mathbb{R}^n$. Let $p\geq 1$ and $\eta,\rho,\delta>0$, be parameters.*

If $\mathcal{W}_p(R,P)\leq\rho$ and $\min\{\log\operatorname{Cov}_{\eta,\delta}(P),\log\operatorname{Cov}_{\eta,\delta}(R)\}\leq k$, then there exist distributions R', R'', P', P'', and a finite discrete distirbution Q with $|\operatorname{supp}(Q)|\leq e^k$ satisfying:

1. $\min\{\mathcal{W}_\infty(P',Q),\mathcal{W}_\infty(R',Q)\}\leq\eta$,
2. $\mathcal{W}_\infty(R',P')\leq\frac{\rho}{\delta^{1/p}}$,
3. $P=(1-2\delta)P'+(2\delta)P''$ *and* $R=(1-2\delta)R'+(2\delta)R''$

Proof. Since the statement of the lemma is symmetric with respect to P and R, WLOG let $\log\operatorname{Cov}_{\eta,\delta}(P)\leq k$. Then there is an $S\subset\mathbb{R}^n$ such that $|S|\leq e^k$ and

$$\Pr_{x\sim P}[x\in\cup_{u\in S}B(u,\eta)]=1-c_P\geq 1-\delta,$$

We define the function $f:\mathbb{R}^n\to\mathbb{R}_+$ as

$$f(x)=\begin{cases}\frac{1}{|\{u\in S|x\in B(u,\eta)\}|} & \text{if } \exists u\in S\ s.t.\ x\in B(u,\eta),\\ 0 & \text{otherwise.}\end{cases}$$

By construction, f is a piecewise constant function that is inversely proportional to the number of $\eta-$radius balls centered around points in S cover a point x.

For each $u\in S$, we define the measure Q'' as

$$Q''(u):=\int_{B(u,\eta)} f\,dP.$$

Observe that

$$\sum_{u\in S}Q''(u)=\sum_{u\in S}\int_{B(u,\eta)}f dP,$$

$$= \int_{\cup_{u \in S} B(u,\eta)} dP = 1 - c_P$$

Notice that Q'' is not a probability distribution, since it only has mass $1 - c_P$. However we can create a distribution Q' from Q'' by putting an additional c_P mass on some arbitrary point in $\mathbb{R}^n$ (say, 0). By construction, there exists a coupling Π of P and Q' where the coupling distributes the mass at each point in $\mathbb{R}^n$ to points η close to it in S, such that

$$c_P = \Pr_{(x_1,x_2)\sim\Pi} [\|x_1 - x_2\| \geq \eta] \leq \delta. \tag{14}$$

Additionally, since $W_p(R, P) \leq \rho$, there exists a coupling Γ such that.

$$c_R = \Pr_{(x_1,x_2)\sim\Gamma} \left[\|x_1 - x_2\| \geq \frac{\rho}{\delta^{1/p}}\right] \leq \frac{\mathbb{E}\left[\|x_1 - x_2\|^p\right]}{\frac{\rho^p}{\delta}} \leq \delta. \tag{15}$$

where c_P is defined by the first equality. We can hence define a couple between P, Q', R whose distribution is given by the following – for any borel measurable sets B_1, B_2, B_3 we have $\Omega(B_1, B_2, B_3) = P(B_1)\Pi(B_2 \mid B_1)\Gamma(B_3 \mid B_1)$. To verify that this is indeed a coupling of the kind we want, we observe that the marginals of Ω are P, Q and R respectively.

1. $\Omega(B_1, \mathbb{R}^n, \mathbb{R}^n) = P(B_1)\Pi(\mathbb{R}^n \mid B_1)\Gamma(\mathbb{R}^n \mid B_1) = P(B_1)$.
2. $\Omega(\mathbb{R}^n, B_2, \mathbb{R}^n) = P(\mathbb{R}^n)\Pi(B_2 \mid \mathbb{R}^n)\Gamma(\mathbb{R}^n \mid \mathbb{R}^n) = 1 \cdot \frac{\Pi(B_2,\mathbb{R}^n)}{P(\mathbb{R}^n)} \cdot 1 = Q'(B_2)$.
3. $\Omega(\mathbb{R}^n, \mathbb{R}^n, B_3) = P(\mathbb{R}^n)\Pi(\mathbb{R}^n \mid \mathbb{R}^n)\Gamma(B_3 \mid \mathbb{R}^n) = R(B_3)$.

To define P', Q, R', we look at Ω conditioned on the event $E := \{(x, y, z) \mid \|x - z\| \leq \rho/\delta^{1/p}$ and $\|x - y\| \leq \eta\}$. To estimate the probability of E, we define $E_1 := \{(x, y, z) \mid z \in \mathbb{R}^n$ and $\|x - y\| > \eta\}$ and $E_2 := \{(x, y, z) \mid \|x - z\| > \rho/\delta^{1/p}$ and $y \in \mathbb{R}^n\}$. Then, $\overline{E} = E_1 \vee E_2$.

We now show that $\Omega(E_1) \leq \delta$. Let $(E_1)_I$ denote E_1 restricted to the coordinates in I.

$$\Omega(E_1) := P((E_1)_1)\Pi((E_1)_{1,2} \mid (E_1)_1)\Gamma((E_1)_{1,3} \mid (E_1)_1) \leq \Pi((E_1)_{1,2}) \leq \delta,$$

where the first inequality is because $\Gamma((E_1)_{1,3} \mid (E_1)_1) \leq 1$ and $\Pi((E_1)_{1,2} \mid (E_1)_1) = \Pi((E_1)_{1,2})/P((E_1)_1)$ and the final inequaity follows from equation (14). The bound for E_2 follows similarly. A union bound shows that $\Omega(E) \geq 1 - 2\delta$. We can restrict the event E further to have mass $1 - 2\delta$.

We look at the marginals of the conditional couple $\Omega(\cdot \mid E)$ to get distributions P', Q, R' as follows. We define $P'(\cdot) := \Omega(\cdot, \mathbb{R}^n, \mathbb{R}^n \mid E)$, $Q(\cdot) := \Omega(\mathbb{R}^n, \cdot, \mathbb{R}^n \mid E)$ and $R'(\cdot) := \Omega(\mathbb{R}^n, \mathbb{R}^n, \cdot \mid E)$. P'' and R'' are defined similarly via conditioning on $\overline{E}$. Hence, $P(\cdot) = \Omega(\cdot, \mathbb{R}^n, \mathbb{R}^n) = \Omega(E)\Omega(\cdot, \mathbb{R}^n, \mathbb{R}^n \mid E) + \Omega(\overline{E})\Omega(\cdot, \mathbb{R}^n, \mathbb{R}^n \mid \overline{E}) = (1 - 2\delta)P'(\cdot) + (2\delta)P''(\cdot)$. The statement for R follows similarly.

This finally gives distributions P', R', Q, such that:

1. $\mathcal{W}_\infty(P', Q) \leq \eta$
2. $\mathcal{W}_\infty(R', P') \leq \rho/\delta^{1/p}$
3. $P = (1 - 2\delta)P' + (2\delta)P''$ and $R = (1 - 2\delta)R' + (2\delta)R''$.

The first two statements follow because of the event we condition over.

Note that this restriction does not change the fact that $\mathrm{supp}(Q) < e^k$, and hence we have our result.

□

A.4. Proof of Lemma 3.3

Lemma 3.3. *Let R, P, denote arbitrary distributions over $\mathbb{R}^n$ such that $\mathcal{W}_\infty(R, P) \leq \varepsilon$.*

Let $x^* \sim R$ and $z^* \sim P$ and let y and u be generated from x^* and z^* via a Gaussian measurement process with m measurements and noise rate σ. Let $\widehat{x} \sim P(\cdot|y, A)$ and $\widehat{z} \sim P(\cdot|u, A)$. For any $d > 0$, we have

$$\begin{aligned}
&\Pr_{x^*, A, \xi, \widehat{x}} [\|x^* - \widehat{x}\| \geq d + \varepsilon] \\
&\leq e^{-\Omega(m)} + e^{\left(\frac{4\varepsilon(\varepsilon+2\sigma)m}{2\sigma^2}\right)} \Pr_{z^*, A, \xi, \widehat{z}} [\|z^* - \widehat{z}\| \geq d] .
\end{aligned}$$

Proof. Let B_1 denote the event

$$B_1 = \{\|x^* - \widehat{x}\| \geq d + \varepsilon\} .$$

Similarly, let B_2 denote the event

$$B_2 = \{\|z^* - \widehat{x}\| \geq d\} .$$

We have

$$\Pr_{x^* \sim R, A, \xi, \widehat{x} \sim P(\cdot|A,y)} [B_1] = \mathop{\mathbb{E}}_{x^* \sim R} \mathop{\mathbb{E}}_{A} \left[\mathop{\mathbb{E}}_{y|A,x^*} \left[\mathop{\mathbb{E}}_{\widehat{x} \sim P(\cdot|y,A)} [1_{B_1}] \right]\right] .$$

We can write the integral over R as an integral over the coupling Π between R, P. This gives

$$\Pr_{x^*, A, \xi, \widehat{x} \sim P(\cdot|A,y)} [B_1] = \mathop{\mathbb{E}}_{x^*, z^*} \mathop{\mathbb{E}}_{A} \left[\mathop{\mathbb{E}}_{y|A,x^*} \left[\mathop{\mathbb{E}}_{\widehat{x} \sim P(\cdot|y,A)} [1_{B_1}] \right]\right] .$$

Since x^*, z^* are coupled and $W_\infty(R, P) \leq \varepsilon$, we have $\|x^* - z^*\| \leq \varepsilon$ almost surely. This gives $B_1 \subseteq B_2$ if x^*, z^* are distributed according to Π. Hence,

$$\Pr_{x^*, A, \xi, \widehat{x} \sim P(\cdot|A,y)} [B_1] \leq \mathop{\mathbb{E}}_{x^*, z^*} \mathop{\mathbb{E}}_{A} \left[\mathop{\mathbb{E}}_{y|A,x^*} \left[\mathop{\mathbb{E}}_{\widehat{x} \sim P(\cdot|y,A)} [1_{B_2}] \right]\right] .$$

We can split the above integral into two parts: one where the matrix A satsifies $\|Ax^* - Az^*\| \leq 2\varepsilon$, and another case where $\|Ax^* - Az^*\| > 2\varepsilon$. This gives

$$\begin{aligned}
\Pr_{x^*, A, \xi, \widehat{x} \sim P(\cdot|A,y)} [B_1] &\leq \mathop{\mathbb{E}}_{x^*, z^*} \mathop{\mathbb{E}}_{A} \left[1_{\|Ax^* - Az^*\| > 2\varepsilon} \mathop{\mathbb{E}}_{y|A,x^*} \left[\mathop{\mathbb{E}}_{\widehat{x} \sim P(\cdot|y,A)} [1_{B_2}] \right]\right] (*) && (16) \\
&+ \mathop{\mathbb{E}}_{x^*, z^*} \mathop{\mathbb{E}}_{A} \left[1_{\|Ax^* - Az^*\| \leq 2\varepsilon} \mathop{\mathbb{E}}_{y|A,x^*} \left[\mathop{\mathbb{E}}_{\widehat{x} \sim P(\cdot|y,A)} [1_{B_2}] \right]\right] .(**) && (17)
\end{aligned}$$

Consider the term$(*)$ in line (16). We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{x^*, z^*} \mathop{\mathbb{E}}_{A} \left[1_{\|Ax^* - Az^*\| > 2\varepsilon} \mathop{\mathbb{E}}_{y|A,x^*} \left[\mathop{\mathbb{E}}_{\widehat{x} \sim P(\cdot|y,A)} [1_{B_2}] \right]\right] &\leq \mathop{\mathbb{E}}_{x^*, z^*} \mathop{\mathbb{E}}_{A} \left[1_{\|Ax^* - Az^*\| > 2\varepsilon} \right] , && (18) \\
&\leq \mathop{\mathbb{E}}_{x^*, z^*} \left[e^{-\Omega(m)} \right] \leq e^{-\Omega(m)}, && (19)
\end{aligned}$$

where the last inequality follows from the Johnson-Lindenstrauss lemma for a fixed x^*, z^*, and hence is true on average over x^*, z^* drawn independent of A.

Now consider the term $(**)$ in line (17). Notice that since the noise in the measurements is Gaussian, we have

$$y|x^*, A \sim \mathcal{N}(0, \sigma^2/m).$$

We break the integral over y in $(**)$ into two cases:

1. Case 1: $\|y - Ax^*\| > 2\sigma$. Since $p(y|A, x^*)$ is distributed as $\mathcal{N}\left(0, \frac{\sigma^2}{m} I_m\right)$, by standard Gaussian concentration, we have

$$\int_{y:\|y-Ax^*\|>2\sigma} p(y|A, x^*)dy \le e^{-\Omega(m)}.$$

2. Case 2: $\|y - Ax^*\| \le 2\sigma$. This gives

$$\begin{aligned}
\|Ax^* - y\|^2 &= \|Ax^* - y\|^2 - \|y - Az^*\|^2 + \|y - Az^*\|^2, \\
&= \|Ax^* - y\|^2 - \|y - Ax^* + Ax^* - Az^*\|^2 + \|y - Az^*\|^2, \\
&= -\|Ax^* - Az^*\|^2 - 2\langle y - Ax^*, Ax^* - Az^*\rangle + \|y - Az^*\|^2.
\end{aligned}$$

Observe that in $(**)$, we have

$$\|Ax^* - Az^*\| \le 2\varepsilon \Rightarrow \|Ax^* - Az^*\|^2 \le 4\varepsilon^2.$$

By the Cauchy-Schwartz inequality and the assumption that $\|y - Ax^*\| \le 2\sigma$, we have

$$2\langle y - Ax^*, Ax^* - Az^*\rangle \le 8\sigma\varepsilon.$$

Substituting the above two inequalities, we have

$$\|Ax^* - y\|^2 \ge -4\varepsilon^2 - 8\sigma\varepsilon + \|y - Az^*\|^2, \tag{20}$$

$$\Rightarrow \exp\left(-\frac{\|Ax^* - y\|^2}{2\sigma^2/m}\right) \le \exp\left(\frac{4\varepsilon\left(\varepsilon + 2\sigma\right)m}{2\sigma^2}\right) \exp\left(-\frac{\|Az^* - y\|^2}{2\sigma^2/m}\right), \tag{21}$$

(22)

Observe that the LHS has the density of measurements from x^*, while the RHS has the density of measurements from z^* with an exponential scaling. From the above inequality, we can replace the expectation over $y|A, x^*$ in $(**)$ with $u|A, z^*$ with an exponential factor.

Similarly, since the conditional sampling now uses u in place of y, we can replace $\widehat{x}$ in $(**)$ with $\widehat{z}$.

Combining Case 1 and 2 gives

$$\begin{aligned}
(**) &\le e^{-\Omega(m)} + e^{\left(\frac{4\varepsilon(\varepsilon+2\sigma)m}{2\sigma^2}\right)} \mathop{\mathbb{E}}_{x^*, z^*} \mathop{\mathbb{E}}_{A}\left[\mathop{\mathbb{E}}_{u|A,z^*}\left[\mathop{\mathbb{E}}_{\widehat{z}\sim P(\cdot|u,A)}[1_{B_2}]\right]\right], \\
&= e^{-\Omega(m)} + e^{\left(\frac{4\varepsilon(\varepsilon+2\sigma)m}{2\sigma^2}\right)} \mathop{\mathbb{E}}_{z^*\sim P} \mathop{\mathbb{E}}_{A}\left[\mathop{\mathbb{E}}_{u|A,z^*}\left[\mathop{\mathbb{E}}_{\widehat{z}\sim P(\cdot|u,A)}[1_{B_2}]\right]\right].
\end{aligned}$$

From the above inequality and eqn. (19), we have

$$\Pr_{x^*\sim R, \xi, A, \widehat{x}\sim P(\cdot|A,y)}\left[\|x^* - \widehat{x}\| \ge d + \varepsilon\right] \le e^{-\Omega(m)} + e^{\left(\frac{4\varepsilon(\varepsilon+2\sigma)m}{2\sigma^2}\right)} \Pr_{z^*\sim P, \xi, A, \widehat{z}\sim P(\cdot|u,A)}\left[\|z^* - \widehat{z}\| \ge d\right].$$

$\square$

A.5. Proof of Theorem 3.4

Theorem 3.4. *Let $\delta \in [0, 1/4)$, $p \geq 1$, and $\varepsilon, \eta > 0$ be parameters. Let R, P be arbitrary distributions over $\mathbb{R}^n$ satisfying $\mathcal{W}_p(R, P) \leq \varepsilon$.*

Let $x^* \sim R$ and suppose y is generated by a Gaussian measurement process from x^* with noise rate $\sigma \gtrsim \varepsilon/\delta^{1/p}$ and $m \geq O(\min(\log \mathrm{Cov}_{\eta,\delta}(R), \log \mathrm{Cov}_{\eta,\delta}(P)))$ measurements. Given y and the fixed matrix A, let $\widehat{x}$ output of conditional resampling with respect to P.

Then there exists a universal constant $c > 0$ such that with probability at least $1 - e^{-\Omega(m)}$ over A, ξ,

$$\Pr_{x^* \sim R, \widehat{x} \sim P(\cdot|y)} [\|x^* - \widehat{x}\| \geq c\eta + c\sigma] \leq 2\delta + 2e^{-\Omega(m)}.$$

Proof. We know from Lemma A.1 that there exist R', P', R'', P'' and a finite distribution Q supported on the set S such that

1. $\mathcal{W}_\infty(R', P') \leq \frac{\varepsilon}{\delta^{1/p}}$,
2. $\min\{\mathcal{W}_\infty(P', Q), \mathcal{W}_\infty(R', Q)\} \leq \eta$,
3. $R = (1 - 2\delta)R' + 2\delta R''$ and $P = (1 - 2\delta)P' + 2\delta P''$,
4. $|S| \leq e^k$.

Suppose $\mathcal{W}_\infty(P', Q) \leq \eta$. If not, then $\mathcal{W}_\infty(R', Q) \leq \eta$, and by (1), we see that $\mathcal{W}_\infty(P', Q) \leq \eta + \frac{\varepsilon}{\delta^{1/p}}$, and we will use this in the proof instead. This gives us

$$\Pr_{x^* \sim R, \widehat{x} \sim P(\cdot|y)} [\|x^* - \widehat{x}\| \geq (c+1)\eta + (c+1)\sigma] \tag{23}$$

$$\leq \Pr_{x^* \sim R, \widehat{x} \sim P(\cdot|y)} \left[\|x^* - \widehat{x}\| \geq (c+1)\eta + c\sigma + (\varepsilon/\delta^{1/p})\right] \tag{24}$$

$$\leq 2\delta + (1 - 2\delta) \Pr_{x^* \sim R', \widehat{x} \sim P(\cdot|y)} \left[\|x^* - \widehat{x}\| \geq (c+1)\eta + c\sigma + (\varepsilon/\delta^{1/p})\right] \tag{25}$$

We now bound the second term on the right hand side of the above equation. For this term, consider the joint distribution over $x^*, A, \xi, \widehat{x}$. By Lemma 3.3, we can replace $x^* \sim R'$ with $z^* \sim P'$, replace $y = Ax^* + \xi$ with $u = Az^* + \xi$, and replace $\widehat{x} \sim P(\cdot|A, y)$ with $\widehat{z} \sim P(\cdot|A, u)$ to get the following bound

$$\begin{aligned} &\Pr_{x^* \sim R', A, \xi, \widehat{x} \sim P(\cdot|A, y)} \left[\|x^* - \widehat{x}\| \geq (c+1)\,\eta + c\sigma + (\varepsilon/\delta^{1/p})\right] \leq \\ &e^{-\Omega(m)} + e^{\left(\frac{2(\varepsilon/\delta^{1/p})\left((\varepsilon/\delta^{1/p}) + 2\sigma\right)m}{\sigma^2}\right)} \Pr_{z^* \sim P', A, \xi, \widehat{z} \sim P(\cdot|u, A)} [\|z^* - \widehat{z}\| \geq (c+1)\eta + c\sigma] \,. \end{aligned} \tag{26}$$

We now bound the second term in the right hand side of the above inequality. Let Γ denote an optimal $\mathcal{W}_\infty-$coupling between P' and Q.

For each $\tilde{z} \in S$, the conditional coupling can be defined as

$$\Gamma(\cdot|\tilde{z}) = \frac{\Gamma(\cdot, \tilde{z})}{Q(\tilde{z})}.$$

By the $\mathcal{W}_\infty$ condition, each $\Gamma(\cdot|\tilde{z})$ is supported on a ball of radius η around $\tilde{z}$.

Let $E = \{z^*, \widehat{z} \in \mathbb{R}^n : \|z^* - \widehat{z}\| \geq (c+1)\,\eta + c\sigma\}$ denote the event that $z^*, \widehat{z}$ are far apart. By the coupling, we can express P' as

$$P' = \sum_{\tilde{z} \in S} Q(\tilde{z}) \Gamma(\cdot|\tilde{z}).$$

This gives

$$\Pr_{z^*\sim P',A,\xi,\widehat{z}\sim P(\cdot|A,u)}[E] = \sum_{\tilde{z}^*\in S} Q(\tilde{z}^*) \mathop{\mathbb{E}}_{z^*\sim\Gamma(\cdot|\tilde{z}^*),A,\xi,\widehat{z}\sim P(\cdot|A,u)}[1_E] \,.$$

For each $\tilde{z}^* \in S$, we now bound $Q(\tilde{z}^*)\,\mathbb{E}_{z^*\sim\Gamma(\cdot|\tilde{z}^*),A,\xi,\widehat{z}\sim P(\cdot|A,u)}[1_E]$.

For each $\tilde{z}^* \in S$, we can write P as $P = (1-2\delta)\,Q_{\tilde{z}^*}P_{\tilde{z}^*,0} + c_{\tilde{z}^*,1}P_{\tilde{z}^*,1} + c_{\tilde{z}^*,2}P_{\tilde{z}^*,2}$, where the components of the mixture are defined in the following way. The first component $P_{\tilde{z}^*,0}$ is $\Gamma(\cdot|\tilde{z}^*)$, the second component is supported within a $c(\eta+\sigma)$ radius of $\tilde{z}^*$, and the third component is supported outside a $c\,(\eta+\sigma)$ radius of $\tilde{z}^*$.

Formally, let $B_{\tilde{z}^*}$ denote the ball of radius $c(\eta+\sigma)$ centered at $\tilde{z}^*$, and let $B^c_{\tilde{z}^*}$ be its complement. The constants are defined via the following Lebsque integrals, and the mixture components for any Borel measurable B are defined as

$$c_{\tilde{z}^*,1} := \int_{B_{\tilde{z}^*}} dP - (1-2\delta)\,Q_{\tilde{z}^*}\int_{B_{\tilde{z}^*}} d\Gamma(\cdot|\tilde{z}^*),$$

$$c_{\tilde{z}^*,2} := \int_{B^c_{\tilde{z}^*}} dP - (1-2\delta)\,Q_{\tilde{z}^*}\int_{B^c_{\tilde{z}^*}} d\Gamma(\cdot|\tilde{z}^*),$$

$$P_{\tilde{z}^*,0}(B) := \Gamma(B\cap B_{\tilde{z}^*}|\tilde{z}^*) = \Gamma(B|\tilde{z}^*) \text{ since } \operatorname{supp}(\Gamma(\cdot|\tilde{z}^*)) \subset B_{\tilde{z}^*},$$

$$P_{\tilde{z}^*,1}(B) := \begin{cases} \frac{1}{c_{\tilde{z}^*,1}}P(B\cap B_{\tilde{z}^*}) - \frac{1-2\delta}{c_{\tilde{z}^*,1}}Q_{\tilde{z}^*}\Gamma(B\cap B_{\tilde{z}^*}|\tilde{z}^*) & \text{if } c_{\tilde{z}^*,1}>0, \\ \text{do not care} & \text{otherwise.} \end{cases},$$

$$P_{\tilde{z}^*,2}(B) := \begin{cases} \frac{1}{c_{\tilde{z}^*,2}}P(B\cap B^c_{\tilde{z}^*}) - \frac{1-2\delta}{c_{\tilde{z}^*,2}}Q_{\tilde{z}^*}\Gamma(B\cap B^c_{\tilde{z}^*}|\tilde{z}^*) & \text{if } c_{\tilde{z}^*,2}>0, dx^* \\ \text{do not care} & \text{otherwise.} \end{cases}.$$

Notice that if z^* is sampled from $\Gamma(\cdot|\tilde{z}^*)$, then by the W_∞ condition, we have $\|z^*-\tilde{z}^*\| \le \eta$. Furthermore, if $\widehat{z}$ is $(c+1)\,\eta + c\sigma$ far from z^*, an application of the triangle inequality implies that it must be distributed according to $P_{\tilde{z}^*,2}$. That is,

$$\begin{aligned} Q(\tilde{z}^*) \mathop{\mathbb{E}}_{z^*\sim\Gamma(\cdot|\tilde{z}^*),A,\xi,\widehat{z}\sim P(\cdot|A,u)}[1_E] &\le \mathop{\mathbb{E}}_{A,\xi,z^*}\Pr\left[z^*\sim P_{\tilde{z}^*,0}, \widehat{z}\sim P_{\tilde{z}^*,2}(\cdot|u)\right] \\ &\le \frac{1}{1-2\delta}\mathop{\mathbb{E}}_{A}\left[1-TV(H_{\tilde{z}^*,0},H_{\tilde{z}^*,2})\right], \end{aligned}$$

where $H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2}$ are the push-forwards of $P_{\tilde{z}^*,0}, P_{\tilde{z}^*,2}$ for A fixed and the last inequality follows from Claim A.2.

This gives

$$\Pr_{z^*\sim P',A,\xi,\widehat{z}\sim P(\cdot|u,A)}[E] \le \frac{1}{1-2\delta}\sum_{\tilde{z}^*\in S}\mathop{\mathbb{E}}_{A}\left[1-TV(H_{\tilde{z}^*,0},H_{\tilde{z}^*,2})\right].$$

Notice that $P_{\tilde{z}^*,0}$ is supported within an $\eta-$ball around $\tilde{z}^*$, and $P_{\tilde{z}^*,2}$ is supported outside a $c(\eta+\sigma)-$ball of $\tilde{z}^*$. By Lemma 3.2 we have

$$\mathop{\mathbb{E}}_{A}[TV(H_{\tilde{z}^*,0},H_{\tilde{z}^*,2})] \ge 1 - 4e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}.$$

This implies

$$\Pr_{z^*\sim P',A,\xi,\widehat{z}\sim P(\cdot|u,A)}\left[\|z^*-\widehat{z}\| \ge (c+1)\eta + c\sigma\right] \le \frac{1}{1-2\delta}\sum_{\tilde{z}^*\in S}\mathop{\mathbb{E}}_{A}\left[(1-TV(H_{\tilde{z}^*,0},H_{\tilde{z}^*,2}))\right],$$

$$\leq \frac{1}{1-2\delta} 4|S| e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)},$$
$$\leq \frac{1}{1-2\delta} 4 e^{-\frac{m}{4}\log\left(\frac{c}{4e^2}\right)},$$

where the last inequality is satisfied if $m \geq 4\log(|S|)$.

Substituting in Eqn (26), if $c > 4\exp\left(2 + \frac{8(\varepsilon/\delta^{1/p})\left((\varepsilon/\delta^{1/p})+2\sigma\right)}{\sigma^2}\right)$, we have

$$\Pr_{x^* \sim R', A, \xi, \widehat{x} \sim P(\cdot|A,y)}\left[\|x^* - \widehat{x}\| \geq (c+1)\,\eta + c\sigma + (\varepsilon/\delta^{1/p})\right] \leq e^{-\Omega(m)} + \frac{1}{1-2\delta} e^{-\Omega(m\log c)}.$$

This implies that there exists a set $S_{A,\xi}$ over A,ξ satisfying $\Pr_{A,\xi}[S_{A,\xi}] \geq 1 - e^{-\Omega(m)}$, such that for all $A,\xi \in S_{A,\xi}$, we have

$$\Pr_{x^* \sim R', \widehat{x} \sim P(\cdot|y)}\left[\|x^* - \widehat{x}\| \geq (c+1)\eta + c\sigma + (\varepsilon/\delta^{1/p})\right] \leq \frac{1}{1-2\delta} e^{-\Omega(m)}.$$

Substituting in Eqn (23), we have

$$\Pr_{x^* \sim R, \widehat{x} \sim P(\cdot|y)}\left[\|x^* - \widehat{x}\| \geq (c+1)\eta + c\sigma + (\varepsilon/\delta^{1/p})\right] \leq 2\delta + \frac{1}{1-2\delta} e^{-\Omega(m)} \leq 2\delta + 2e^{-\Omega(m)}.$$

Rescaling c gives us our result.

At the beginning of the proof, we had assumed that $\mathcal{W}_\infty(P', Q) \leq \eta$. If instead $\mathcal{W}_\infty(R', Q) \leq \eta$, then we need to replace η in the above bound by $\eta + \frac{\varepsilon}{\delta^{1/p}}$. Rescaling c in the above bound gives us the Theorem statement.

□

Claim A.2. *Consider the setting of the previous theorem. We have*

$$\mathop{\mathbb{E}}_{A,\xi,z^*} \Pr\left[z^* \sim P_{\tilde{z}^*,0}, \widehat{z} \sim P_{\tilde{z}^*,2}(\cdot|u)\right] \leq \frac{1}{1-\delta_2} \mathop{\mathbb{E}}_{A}\left[1 - TV(H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2})\right], \tag{27}$$

Proof. For a fixed A, let h_0, h_2 denote the corresponding densities of the push forward of $P_{\tilde{z}^*,0}, P_{\tilde{z}^*,2}$. Then we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{A,\xi,z^*} \Pr\left[z^* \sim P_{\tilde{z}^*,0}, \widehat{z} \sim P_{\tilde{z}^*,2}(\cdot|u)\right] &= \mathop{\mathbb{E}}_{A} \int \frac{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u) c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)}{(1-\delta_2)\, Q_{\tilde{z}^*,0} h_{\tilde{z}^*,0}(u) + c_{\tilde{z}^*,1} h_{\tilde{z}^*,1}(u) + c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)} du, && (28)\\
&\leq \mathop{\mathbb{E}}_{A} \int \frac{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u) c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)}{(1-\delta_2)\, Q_{\tilde{z}^*,0} h_{\tilde{z}^*,0}(u) + c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)} du, && (29)\\
&\leq \mathop{\mathbb{E}}_{A} \int \frac{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u) c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)}{(1-\delta_2)\, Q_{\tilde{z}^*,0} h_{\tilde{z}^*,0}(u) + (1-\delta_2) c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)} du, && (30)\\
&\leq \mathop{\mathbb{E}}_{A} \frac{1}{1-\delta_2} \int \frac{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u) c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)}{Q_{\tilde{z}^*,0} h_{\tilde{z}^*,0}(u) + c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)} du, && (31)\\
&\leq \mathop{\mathbb{E}}_{A} \frac{1}{1-\delta_2} \int \frac{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u) c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)}{\max\{Q_{\tilde{z}^*,0} h_{\tilde{z}^*,0}(u)\,,\, c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)\}} du, && (32)\\
&= \mathop{\mathbb{E}}_{A} \frac{1}{1-\delta_2} \int \min\{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u), c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)\} du, && (33)\\
&\leq \mathop{\mathbb{E}}_{A} \frac{1}{1-\delta_2} \int \min\{h_{\tilde{z}^*,0}(u), h_{\tilde{z}^*,2}(u)\} du, && (34)\\
&= \frac{1}{1-\delta_2} \mathop{\mathbb{E}}_{A}\left[1 - TV(H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2})\right]. && (35)
\end{aligned}$$

□

$$\begin{tikzcd} x^* \arrow[r] & z \arrow[r] & y \arrow[r] & \widehat{x} \\ A \arrow[u, shift left] \arrow[ur] \arrow[urrr] & & & \end{tikzcd}$$

Figure 7: DAG relating $x^*, A, z, y, \widehat{x}$. The conditional independencies we use are $x^* \perp\!\!\!\perp y|z, A$ and $A \perp\!\!\!\perp y|z$.

B. Lower Bound Proofs

B.1. Proof of Lemma 4.2

Lemma 4.2. *Consider the setting of Theorem* (4.1)*. We have*

$$I(y; x^*|A) \leq \frac{m}{2} \log \left(1 + \frac{1}{\sigma^2}\right).$$

Proof. We have $y = Ax^* + \xi$. Let $z = Ax^*$, which gives $y = z + \xi$.

We have $z_i = a_i^T x^*$ where a_i is the i^{th} row of A, and $y_i = z_i + \xi_i$. Since x^* is supported within the unit sphere and the elements of A are drawn from $\mathcal{N}(0, \frac{1}{m})$, we have $\mathbb{E}[y_i^2] = \mathbb{E}[\langle a_i, x \rangle^2] \leq \frac{1}{m}$. Since the Gaussian noise ξ has variance σ^2/m in each coordinate, every coordinate of y_i is a Gaussian channel with power constaint $1/m$ and noise variance σ^2/m. Using Shannon's AWGN theorem (Cover & Thomas, 2012; Polyanskiy & Wu, 2014; Shannon, 1948), the mutual information between y_i, z_i, is bounded by

$$I(y_i; z_i) \leq \frac{1}{2} \log \left(1 + \frac{1}{\sigma^2}\right).$$

The chain rule of entropy and sub-adddітivity of entropy implies,

$$\begin{aligned} I(y; z) &= h(y) - h(y|z) = h(y) - h(y - z|z), && (36)\\ &= h(y) - h(\xi|z) = h(y) - \sum h(\xi_i|z, \xi_1, \cdots, \xi_{i-1}), && (37)\\ &= h(y) - \sum h(\xi_i), && (38)\\ &\leq \sum h(y_i) - \sum h(\xi_i), && (39)\\ &= \sum h(y_i) - \sum h(y_i|z_i), && (40)\\ &= \sum I(y_i; z_i), && (41)\\ &\leq \frac{m}{2} \log \left(1 + \frac{1}{\sigma^2}\right). && (42) \end{aligned}$$

Now we prove that $I(x^*; y|A) \leq I(y; z)$.

Consider the mutual information $I(x^*, A, z; y)$. By the chain rule of mutual information, we have

$$\begin{aligned} I(x^*, A, z; y) &= I(A; y) + I(x^*; y|A) + I(z; y|x^*, A), \\ &= I(A; y) + I(z; y|A) + I(x^*; y|z, A), \\ \Leftrightarrow I(x^*; y|A) + I(z; y|x^*, A) &= I(z; y|A) + I(x^*; y|z, A). \end{aligned}$$

From Figure 7, note that x^*, y, are conditionally independent given z, A. This gives $I(x^*; y|z, A) = 0$.

This gives

$$I(x^*; y|A) + I(z; y|x^*, A) = I(z; y|A), \tag{43}$$

$$\Rightarrow I(x^*; y|A) \le I(z; y|A). \tag{44}$$

We can bound $I(z; y|A)$ in the following way.

$$I(A, z; y) = I(A; y) + I(z; y|A), \tag{45}$$
$$= I(z; y) + I(A; y|z), \tag{46}$$
$$\Leftrightarrow I(A; y) + I(z; y|A) = I(z; y) + I(A; y|z), \tag{47}$$
$$\Leftrightarrow I(A; y) + I(z; y|A) = I(z; y), \tag{48}$$
$$\Rightarrow I(z; y|A) \le I(z; y),. \tag{49}$$

where the second last line follows from $I(A; y|z) = 0$, and the last line follows from $I(A; y) \ge 0$.

From Eqn (42), (44), (49), we have

$$I(x^*; y|A) \le \frac{m}{2}\left(1 + \frac{1}{\sigma^2}\right).$$

□

B.2. Proof of Lemma 4.3

Lemma 4.3. *Consider the setting of Theorem (4.1). We have*

$$I(x^*; \widehat{x}) \le I(y; x^*|A).$$

Proof. Consider the mutual information $I(x^*; y, A, \widehat{x})$. By the chain rule of mutual information, we can express it in two ways:

$$I(x^*; y, A, \widehat{x}) = I(x^*; y, A) + I(x^*; \widehat{x}|y, A), \tag{50}$$
$$= I(x^*; \widehat{x}) + I(x^*; y, A|\widehat{x}). \tag{51}$$

As $\widehat{x}$ is a function of y, A, we have $I(x^*; \widehat{x}|y, A) = 0$. Also, $I(x^*; y, A|\widehat{x}) \ge 0$. Substituting in Eqn (50), (51), we have

$$\begin{aligned} I(x^*; \widehat{x}) &\le I(x^*; y, A), \\ &= I(x^*; A) + I(x^*; y|A), \\ &= I(x^*; y|A), \end{aligned}$$

where the second line follows from the chain rule of mutual information, and the last line follows because x^*, A, are independent.

□

B.3. Proof of Fano variant Lemma 4.4

We will build up Lemma 4.4 in sequence. Before showing it in its full generality, we will show when $x, \widehat{x}$, are discrete random variables and x is uniform (Lemma B.1. We then lift the uniformity restriction on x (Lemma B.2) before extending to continuous distributions (Lemma 4.4).

Lemma B.1. *Let Q be the uniform distribution over an arbitrary discrete finite set S. Let $(x, \widehat{x})$ be jointly distributed, where $x \sim Q$ and $\widehat{x}$ is distributed over an arbitrary countable set, satisfying*

$$\Pr\left[\|x - \widehat{x}\| \le \varepsilon\right] \ge 1 - \delta.$$

Then we have

$$I(x; \widehat{x}) \ge \tau \log \mathrm{Cov}_{4\varepsilon, \tau+\delta}(Q)$$

Proof. Recall,

$$H(x) = \log(|S|)$$
$$H(x \mid \widehat{x}) = H(x) - I(x, \widehat{x})$$

For any $v \in \mathrm{supp}(\widehat{x})$, since x is supported on a finite set of cardinality $|S|$, we have

$$H(x \mid \widehat{x} = v) \le \log(|S|),$$
$$\Rightarrow \log(|S|) - H(x \mid \widehat{x} = v) \ge 0.$$

By the law of total probability, we have

$$I(x; \widehat{x}) = \sum_v P(\widehat{x} = v) \left(\log(|S|) - H(x|\widehat{x} = v)\right).$$

Since the above summation has only non-negative terms that average to $I(x; \widehat{x})$, there exists $G_1 \subseteq \mathrm{supp}(x, \widehat{x})$ with $\Pr[G_1] \ge 1 - \tau$, such that for all $(u, v) \in G_1$,

$$\log(|S|) - H(x|\widehat{x} = v) \le \frac{I(x; \widehat{x})}{\tau},$$
$$\Rightarrow H(x|\widehat{x} = v) \ge \log(|S|) - \frac{I(x; \widehat{x})}{\tau},$$
$$\Rightarrow |\mathrm{supp}(x|\widehat{x} = v)| \ge \frac{|S|}{2^{I(x;\widehat{x})/\tau}}.$$

Let $B(v, \varepsilon)$ denote the ε-radius ball around v. By the hypothesis of the Lemma, we have $\|x - \widehat{x}\| \le \varepsilon$ with probability $\ge 1 - \delta$. By a union bound of the above two inequalities, there exists a set $G_2 \subseteq \mathrm{supp}(x, \widehat{x})$ satisfying $\Pr[G_2] \ge 1 - \tau - \delta$, such that for all $(u, v) \in G_2$, we have

$$|\mathrm{supp}(x|\widehat{x} = v)| \ge \frac{|S|}{2^{I(x;\widehat{x})/\tau}},$$
$$\mathrm{supp}(x|\widehat{x} = v) \subseteq B(v, \varepsilon).$$

The above two inequalities imply that for all $(u, v) \in G_2$, we have

$$|S \cap B(v, \varepsilon)| \ge \frac{|S|}{2^{I(x;\widehat{x})/\tau}}.$$

By the definition of G_2, $(u, v) \in G_2$ satisfy $\|u - v\| \le \varepsilon$. This gives

$$|S \cap B(u, 2\varepsilon)| \ge \frac{|S|}{2^{I(x;\widehat{x})/\tau}}.$$

Therefore any 4ε-packing of this $1 - (\delta + \tau)$ fraction of x must have size at most $2^{I(x;\widehat{x})/\tau}$ by the pigeon-hole principle. Hence there exists a size $2^{I(x;\widehat{x})/\tau}$ cover of radius 4ε containing $1 - (\delta + \tau)$ of Q. $\square$

The previous lemma handled the uniform distribution on x. Now we show that a similar result applies if x's distribution has quantized probability values.

Lemma B.2. *Let Q be a finite discrete distribution over $N \in \mathbb{N}$ points such that for each u in its support, $Q(u) = j\alpha$, where $j \in \mathbb{N}$ and $\alpha := \frac{1}{N_2}$ is a discretization level for $N_2 \in \mathbb{N}$ large enough.*

Let $(x, \widehat{x})$ be jointly distributed, where $x \sim Q$ and $\widehat{x}$ is distributed over a countable set, satisfying

$$\Pr\left[\|x - \widehat{x}\| \le \varepsilon\right] \ge 1 - \delta.$$

Then we have

$$I(x; \widehat{x}) \ge \tau \log \mathrm{Cov}_{4\varepsilon, \tau+\delta}(Q)$$

Proof. For each x in the support of Q, we know that its probability is an integral multiple of $\frac{1}{N_2}$. Hence we can define a new random variable $x' = (x, j), x \in \operatorname{supp}(Q), j \in [N_2]$ and a distribution Q' over x' in the following way:

$$Q'((x, j)) = \begin{cases} \alpha & \text{if } j\alpha \leq Q(x), \\ 0 & \text{otherwise .} \end{cases}$$

By definition, Q' is a uniform distribution, and its support is a discrete subset of $\mathbb{R}^n \times \mathbb{N}$.

Define the following norm for x'. For $x'_1 = (x_1, j_1), x'_2 = (x_2, j_2)$, define

$$\|(x_1, j_1) - (x_2, j_2)\| := \|x_1 - x_2\|.$$

In order to apply Lemma B.1 on Q', it suffices to show that $I(x; \widehat{x}) = I(x'; \widehat{x})$.

By the chain rule of mutual information, we have

$$\begin{aligned} I(x'; \widehat{x}) &= I((x, j); \widehat{x}) \\ &= I(x; \widehat{x}) + I(j; \widehat{x}|x). \end{aligned}$$

Since $\widehat{x}$ is purely a function of x, we have $I(j; \widehat{x}|x) = 0$. This gives

$$I(x'; \widehat{x}) = I(x; \widehat{x}).$$

Similarly construct a version $\widehat{x}' = (\widehat{x}, 0)$ of $\widehat{x}$, whose second coordinate is identically zero. Hence for $x' = (x, j) \sim Q'$, we have

$$\begin{aligned} \|x' - \widehat{x}'\| &\leq \varepsilon \text{ w.p. } 1 - \delta, \\ I(x'; \widehat{x}') &= I(x; \widehat{x}) \end{aligned}$$

Applying Lemma B.1 on Q', we have

$$\tau \log \operatorname{Cov}_{4\varepsilon, \tau+\delta}(Q') \leq I(x; \widehat{x}).$$

Since the support of the first coordinate of Q' is the same as the support of Q, we have

$$\tau \log \operatorname{Cov}_{4\varepsilon, \tau+\delta}(Q) \leq I(x; \widehat{x}).$$

□

We now prove Lemma 4.4, which allows $(x, \widehat{x})$ to follow an arbitrary distribution.

Lemma 4.4 (Fano variant)**.** *Let $(x, \widehat{x})$ be jointly distributed over $\mathbb{R}^n \times \mathbb{R}^n$, where $x \sim R$ and $\widehat{x}$ satisfies*

$$\Pr[\|x - \widehat{x}\| \leq \eta] \geq 1 - \delta.$$

Then for any $\tau > 0$, we have

$$I(x; \widehat{x}) \geq 0.99\tau \log \operatorname{Cov}_{5\eta, \tau+3\delta}(R).$$

Proof. Let $\varepsilon = \eta$, which is the error in the statement of the lemma. Let $\gamma > 0$ be a small enough discretization level to be specified later. For every $x, \widehat{x} \in \mathbb{R}^n$, let $\bar{x}, \bar{\widehat{x}}$ denoted the rounding of $x, \widehat{x}$ to the nearest multiple of γ in each coordinate.

Let $\bar{R}$ be the discrete distribution induced by this discretization of x. We can create such a distribution by assigning the probability of each cell in the grid to its corresponding coordinate-wise floor. This discretization of the support changes the error between $x, \widehat{x}$ in the following way. If $\|x - \widehat{x}\| \leq \varepsilon$ with probability $1 - \delta$, an application of the triangle inequality gives

$$\|\bar{x} - \bar{\widehat{x}}\| \leq \varepsilon + 2\gamma\sqrt{n} \text{ with probability} \geq 1 - \delta. \tag{52}$$

We also need to take into account the effect discretizing $x, \widehat{x}$ has on their mutual information. Note that since $\bar{x}$ is a function of x alone, and $\widehat{\bar{x}}$ is a function of $\widehat{x}$ alone, by the Data Processing Inequality, we have

$$I(\bar{x}; \widehat{\bar{x}}) \leq I(x; \widehat{x}). \tag{53}$$

Note that $\bar{R}$ is a distribution on a discrete but infinite set. However, for any $\beta \in (0, 1]$, we can find a discrete and finite distribution Q such that $\bar{R} = (1 - c_1)Q + c_1 D$, with $c_1 \leq \beta$ and D is some other probability distribution. This is feasible because the probabilites of the infinite support of $\bar{R}$ must sum to 1, and hence we can find a finite subset that sums to atleast $1 - \beta$ for any $\beta \in (0, 1]$. Note that in this process, we only change the marginal of $\bar{x}$ without changing the conditional distribution of $\widehat{\bar{x}}|\bar{x}$. Let $I(\bar{x}; \widehat{\bar{x}}), I_Q(\bar{x}; \widehat{\bar{x}}), I_D(\bar{x}; \widehat{\bar{x}})$ denote the mutual information between $\bar{x}, \widehat{\bar{x}}$ when the marginal of $\bar{x}$ is $\bar{R}, Q, D$, respectively. From Theorem 2.7.4 in (Cover & Thomas, 2012), mutual information is a concave function of the marginal distribution of $\bar{x}$ for a fixed conditional distribution of $\widehat{\bar{x}}|\bar{x}$. An application of Eqn (53) gives us,

$$\begin{aligned} I(x; \widehat{x}) \geq I(\bar{x}; \widehat{\bar{x}}) &\geq (1 - c_1) I_Q(\bar{x}; \widehat{\bar{x}}) + c_1 I_D(\bar{x}; \widehat{\bar{x}}), && (54)\\ &\geq (1 - c_1) I_Q(\bar{x}; \widehat{\bar{x}}), && (55)\\ &\geq (1 - \beta) I_Q(\bar{x}; \widehat{\bar{x}}). && (56) \end{aligned}$$

Now since the finite distribution Q has a TV distance of at most β to the countable distribution R, using Eqn (52), we have

$$\|\bar{x} - \widehat{\bar{x}}\| \leq \varepsilon + 2\gamma\sqrt{n} \text{ with probability } \geq 1 - \beta - \delta \text{ if } \bar{x} \sim Q. \tag{57}$$

In order to apply Lemma B.2 on the distribution Q, we need its probability values to be multiples of some discretization level α. Let α be a small enough quantization level for the probability values. We will specify the value of α later. We can now express the distribution Q as a mixture of two distributions Q', Q''. The distribution Q' is obtained by flooring the probability values under Q and renormalizing to make them sum to 1. The distribution Q'' is the mass not contained in Q', normalized to sum to 1. Since each element in the support of Q loses at most α mass, the total mass in Q'' prior to normalization is at most αN_β, where N_β is the cardinaltiy of the support of Q. This gives

$$Q = (1 - c_2)Q' + c_2 Q'', \; c_2 \leq \alpha N_\beta.$$

From Eqn (57), we have $\|\bar{x} - \widehat{\bar{x}}\| \leq \varepsilon + 2\gamma\sqrt{n}$ with probability $\geq 1 - \beta - \delta$ when $\bar{x} \sim Q$. Since Q' has a TV distance of at most αN_β to Q, if $\bar{x} \sim Q'$, we have

$$\|\bar{x} - \widehat{\bar{x}}\| \leq \varepsilon + 2\gamma\sqrt{n} \text{ with probability } \geq 1 - \beta - \delta - \alpha N_\beta \text{ if } \bar{x} \sim Q'. \tag{58}$$

Let $I_Q(\bar{x}; \widehat{\bar{x}}), I_{Q'}(\bar{x}; \widehat{\bar{x}}), I_{Q''}(\bar{x}; \widehat{\bar{x}})$ denote the mutual information between $\bar{x}, \widehat{\bar{x}}$ when the marginal of $\bar{x}$ is Q, Q', Q'' respectively. Mutual information is a concave function of the marginal distribution of $\bar{x}$ for a fixed conditional distribution of $\widehat{\bar{x}}|\bar{x}$. Hence using Eqn (56), we have

$$\begin{aligned} \frac{I(x; \widehat{x})}{1 - \beta} \geq I_Q(\bar{x}; \widehat{\bar{x}}) &\geq (1 - c_2) I_{Q'}(\bar{x}; \widehat{\bar{x}}) + c_2 I_{Q''}(\bar{x}; \widehat{\bar{x}}), && (59)\\ &\geq (1 - c_2) I_{Q'}(\bar{x}; \widehat{\bar{x}}), && (60)\\ &\geq (1 - \alpha N_\beta) I_{Q'}(\bar{x}; \widehat{\bar{x}}). && (61) \end{aligned}$$

Hence if $\bar{x} \sim Q'$, we have $I(\bar{x}; \widehat{\bar{x}}) \leq \frac{I(x; \widehat{x})}{(1 - \alpha N_\beta)(1 - \beta)}$. Applying Lemma B.2 on the distribution Q', for any $\tau > 0$, we have

$$\tau \log \mathrm{Cov}_{4\varepsilon + 8\gamma\sqrt{n}, \tau + \beta + \delta + \alpha N_\beta}(Q') \leq \frac{I(x; \widehat{x})}{(1 - \alpha N_\beta)(1 - \beta)}.$$

Now since Q' has at least $1 - \alpha N_\beta$ of the mass under Q and Q has at least $1 - \delta$ of the mass under $\bar{R}$, the mass $\tau + \beta + \delta + \alpha N_\beta$ not covered under Q' can be replaced with $\tau + \beta + 2\delta + 2\alpha N_\beta$ under $\bar{R}$. This gives

$$\tau \log \mathrm{Cov}_{4\varepsilon + 8\gamma\sqrt{n}, \tau + \beta + 2\delta + 2\alpha N_\beta}(\bar{R}) \leq \frac{I(x; \widehat{x})}{(1 - \alpha N_\beta)(1 - \beta)}.$$

Now since we can cover the whole distribution of R by extending each element in the support of $\bar{R}$ by γ in each coordinate, we can replace the radius $4\varepsilon + 8\gamma\sqrt{n}$ for $\bar{R}$ by $4\varepsilon + 10\gamma\sqrt{n}$ for R. This gives

$$\tau \log \mathrm{Cov}_{4\varepsilon+10\gamma\sqrt{n},\tau+\beta+2\delta+2\alpha N_\beta}(R) \leq \frac{I(x;\widehat{x})}{(1-\alpha N_\beta)(1-\beta)}.$$

For $\gamma = \frac{\varepsilon}{10\sqrt{n}}$, $\beta = \min\left\{\frac{\delta}{3}, 1-\sqrt{0.99}\right\}$, $\alpha N_\beta = \min\left\{\frac{\delta}{3}, 1-\sqrt{0.99}\right\}$, this becomes

$$0.99\tau \log \mathrm{Cov}_{5\varepsilon,\tau+3\delta}(R) \leq I(x;\widehat{x}).$$

□

B.4. Proof of Theorem 4.1

Theorem 4.1. *Let R be a distribution supported on the unit ball in $\mathbb{R}^n$, and $x^* \sim R$. Let $y = Ax^* + \xi$, where each element of A is drawn iid from $\mathcal{N}(0, \frac{1}{m})$, and $\xi \sim \mathcal{N}(0, \frac{\sigma^2}{m} I_m)$. If there exists a recovery scheme that uses y and A as inputs and guarantees*

$$\|\hat{x} - x^*\| \leq O(\eta),$$

with probability $\geq 1-\delta$, then for any $\eta > 0$, we have

$$m \geq \frac{1.98\delta}{\log\left(1+\frac{1}{\sigma^2}\right)} \log \mathrm{Cov}_{5\eta,4\delta}(R).$$

Proof. By Lemma 4.3 and Lemma 4.2, we have

$$\begin{aligned} I(x^*;\widehat{x}) &\leq I(x^*;y|A), \\ &\leq \frac{m}{2}\log\left(1+\frac{1}{\sigma^2}\right). \end{aligned}$$

Applying Lemma 4.4 on the pair $(x^*, \widehat{x})$, with $\tau = \delta$, we have

$$0.99\delta \log \mathrm{Cov}_{5\eta,4\delta}(R) \leq I(x;\widehat{x}) \leq \frac{m}{2}\log\left(1+\frac{1}{\sigma^2}\right).$$

or

$$m \geq \frac{1.98\delta}{\log\left(1+\frac{1}{\sigma^2}\right)} \log \mathrm{Cov}_{5\eta,4\delta}(R)$$

as desired. □

C. Experimental Setup

C.1. Datasets and Architecture

For the compressed sensing experiment in Fig 4 and the inpainting experiment in Figure 2 we used the 256×256 GLOW model (Kingma & Dhariwal, 2018) from the official repository. The test set for Fig 4 consists of the first 10 images used by (Asim et al., 2019) in their experiments.

For the compressed sensing experiment in Fig 1, 5, 6, we used the FFHQ NCSNv2 model (Song & Ermon, 2020) from the official repository. The test set for Fig 5 consists of the images 69000-69017 from the FFHQ dataset (this corresponds to the first 18 images in the last batch of FFHQ images).

In Fig 4 and Fig 5, the measurements have noise satisfying $\sqrt{\mathbb{E}\|\xi\|^2} = 4$.

C.2. Hyperparameter Selection

CelebA experiments For MAP, we used an Adam and Gradient Descent optimizer. Langevin dynamics only uses Gradient Descent. Each algorithm was run with learning rates varying over $[0.1, 0.01, 0.001, 5 \cdot 10^{-4}, 10^{-4}, 5 \cdot 10^{-5}, 10^{-5}, 5 \cdot 10^{-6}, 10^{-6}]$. For MAP and Modified-MAP, we also performed 2 random restarts for the initialization z_0.

The value of γ in Eqn (7) was varied over $[0, 0.1, 0.01, 0.001]$ for Modified-MAP. MAP uses the theoretically defined value of $\frac{\sigma^2}{m}$.

For Langevin dynamics, we vary the value of σ_i according to the schedule proposed by (Song & Ermon, 2019). We start with $\sigma_1 = 16.0$, and finish with $\sigma_{10} = 4.0$, such that σ_i decreases geometrically for $i \in [10]$. For each value of i, we do 200 steps of noisy gradient descent, with the learning rate schedule proposed by (Song & Ermon, 2019).

In order to select the optimal hyperparameters for each m, we chose the hyperparams that give maximum likelihood for Langevin and MAP. For Modified-MAP, we selected the hyperparameters based on reconstruction error on a holdout set of 5 images.

FFHQ experiments The NCSNv2 model is designed for Langevin dynamics. It can be adapted to MAP by simply not adding noise at each gradient step. We tune the initial and final values of σ used in (Song & Ermon, 2020), along with the initial learning rate.

Unfortunately, it is computationally difficult to obtain the likelihood associated with each reconstruction, since the NCSNv2 model only provides $\nabla \log p(x)$. Although one could, in theory, do numerical integration to find $p(x)$, we selected the optimal hyperparameters for each m based on reconstruction error on a holdout set of 5 images.

For the Deep-Decoder, we used the over-parameterized network described in (Asim et al., 2019), and tuned the learning rate over $[0.4, 0.004, 0.0004]$, and selected the hyper-parameters that optimized the reconstruction error on a holdout set of 5 images.

C.3. Computing Infrastructure

Experiments were run on an NVIDA Quadro P5000.

C.4. Additional Figures

C.5. Code

Our code can be found at `https://github.com/anonymous-panda-icml2021/cs-fairness.git` We will publicly release the code, generative model checkpoints, and optimal hypreparameters shortly.

Figure 8: Reconstructions for $m = 20,000$ for CelebA using GLOW. Note the artefacts in the reconstructions by MAP.

Figure 9: Reconstructions for $m = 40,000$ for FFHQ using NCSNv2. These are the measurements at which MAP competes with DeepDecoder and Langevin dynamics.