Compressed Sensing with Approximate Priors via Conditional Resampling

Ajil Jalal [*]
ECE, UT Austin

Sushrut Karmalkar
CS, UT Austin

Alexandros G. Dimakis
ECE, UT Austin

Eric Price
CS, UT Austin

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 conditional resampling 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 conditional resampling 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 [55, 54], neuronal spike trains [37] and efficient sensing cameras [25]. 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 [17] 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 [19], graphs [78], manifolds [20, 77], or deep generative models [13, 5]. 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, [13] 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

[*] Correspondence to: ajiljalal@utexas.edu

NeurIPS 2020 Workshop on Deep Learning and Inverse Problems, virtual.

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 [48] and pixelRNN [60] 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 [2, 81] 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?

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

Definition 1.1. *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*

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

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

Definition 1.2. *Given an observation $y = Ax^* + \xi$ with $\xi \sim N(0, \frac{\sigma^2}{m} I)$, the* conditional resampling *recovery algorithm with respect to P outputs $\widehat{x}$ according to the conditional distribution $P(\cdot \mid y)$.*

Our main positive result is that conditional resampling 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: conditional resampling with respect to $P \neq R$ still works, as long as P and R are close in Wasserstein distance:

Theorem 1.3 (Upper bound)**.** *Let R, P be distributions with $\mathcal{W}_1(P, R) \leq \sigma$. Let $x^* \sim R$ and $\widehat{x}$ be conditionally resampled from y with respect to P. 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 $\widehat{x}$.

Our second main result shows that conditional resampling is a nearly optimal algorithm. 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 [65, 16, 41]; the closest result we are aware of is [2], which characterizes product distributions.

Theorem 1.4 (Lower bound)**.** *Let R be any distribution over the unit ball, and consider any method to achieve $\|\widehat{x} - x^*\| \leq \eta$ with 99% probability. This must have*

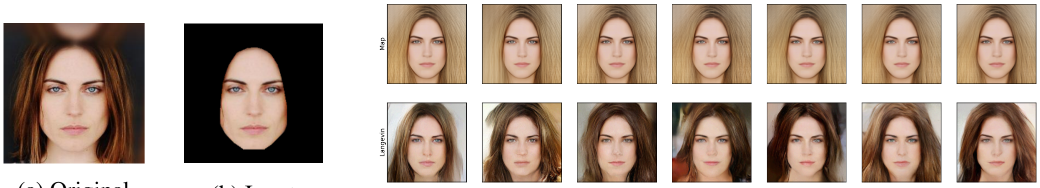
$$m \geq \frac{C'}{\log(1 + 1/\sigma^2)} \log \mathrm{Cov}_{C'\eta,0.04}(R).$$

for some constant $C' > 0$.

For more precisely stated and general versions of these results, including dependence on the failure probability δ, see Theorems 5.4 and 5.5.

1.1 Related Work

Generative priors have shown great promise in compressed sensing and other inverse problems, starting with [13], who generalized the theoretical framework of compressive sensing and restricted eigenvalue conditions [72, 24, 12, 15, 38, 10, 9, 26] for signals lying on the range of a deep generative



(a) Original

(b) Input

(c) Reconstructions by MAP and Langevin

Figure 1: We compare MAP and Langevin dynamics for the inverse problem of inpainting missing pixels. The hair and background of the original image in Figure 1a is removed to get the observations in Figure 1b. In Figure 1c we show reconstructions by MAP and Langevin given the observations in Figure 1b, and each column corresponds to a random initialization for MAP and Langevin. Langevin dynamics produces diverse images that satisfy the observations, while MAP consistently produces essentially the same, somewhat washed-out, image.

model [30, 47]. Similar ideas like projections on smooth manifolds and additional structure beyond sparsity in inverse problems have been studied in earlier signal processing work, e.g. [38, 10, 9, 26].

Results in [45, 52, 43] established that the sample complexities in [13] are order optimal. The approach in [13] has been generalized to tackle different inverse problems [32, 8, 6, 58, 7, 66, 8, 51, 42, 31, 3]. Alternate algorithms for reconstruction include [14, 22, 44, 28, 27, 70, 56, 22, 61, 34, 35]. The complexity of optimization algorithms using generative models have been analyzed in [29, 36, 49, 33]. Experimental results in [5, 74, 50] show that invertible models have superior performance in comparison to low dimensional models. See [59] 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 [1, 40, 57, 39].

Lower bounds for ℓ_2/ℓ_2 recovery of sparse vectors can be found in [68, 65, 2, 41, 16], and these are related to the lower bound in (1.4). The closest result is that of [2], 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 [76, 75]. The $\delta-$Minkowski dimension of a probability measure used in [76, 75, 62] can be derived from our $(\varepsilon, \delta)-$covering number by taking the limit $\varepsilon \to 0$. [67] 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 [79, 11].

2 Experiments

In this section we discuss our algorithm for conditional resampling, and briefly discuss why existing algorithms can fail.

We consider distributions P induced by generative models G, such that the distribution $P = G(z)$ for z a standard Gaussian. Conditionally resampling 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 $\widehat{x} \sim p(x|y)$.

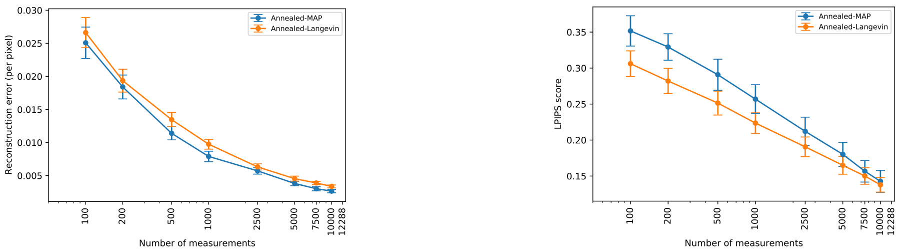
In order to sample from $p(z|y)$, we use *Langevin dynamics*, which samples from a given distribution by moving a random initial sample along a vector field given by the distribution. Langevin dynamics tells us that if we sample $z_0 \sim \mathcal{N}(0, 1)$, and run the following iterative procedure:

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

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 [71]. We instead use an annealed version of the algorithm to sample from the conditional distribution.

Given the measurements y, measurement matrix A, generative model G, and noise scale σ_i, we define $p_i(z|y)$ as

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



(a) MSE vs. m

(b) LPIPS scores vs. m

Figure 2: We compare the performance of our algorithm with baselines for compressed sensing using Gaussian matrices on the CelebA dataset with a RealNVP model in a low PSNR setting. Figure 2a shows per pixel reconstruction error while figure 2b shows LPIPS scores as m varies. We find no statistically significant difference in MSE of the reconstructions, but our reconstructions are perceptually closer to the ground truth. Error bars indicate 95% confidence intervals.

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_i that approach the true value of σ, we can anneal Langevin dynamics and sample from $p(z|y)$. Please refer to the algorithm proposed in [71] for details on how σ_i and the learning rates α_t are varied.

2.1 Experimental Results

We perform our experiments on the CelebA dataset [53, 46], using a RealNVP [23, 74] generative model whose output size is $64 \times 64 \times 3$ and a Glow model [48] whose output size is $256 \times 256 \times 3$. Details about our experiments are in Appendix C. In Figure 2, we show the performance of our proposed algorithm with respect to the MAP baseline for compressed sensing on CelebA with RealNVP in a low PSNR setting. MAP directly maximizes $\log p_i(z|y)$ defined in Eqn (2).

MAP estimation does not work on general distribution: 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. Conditional resampling prefers the opposite.

However, in our experiments on RealNVP and Glow we did not find any statistically significant differences in their reconstruction accuracy in compressed sensing. What we did observe is significant differences in the *perceptual quality* and *diversity* of the resulting images. The LPIPS score[80] is a measure of perceptual distance between images, and Figure 2b shows that our algorithm produces images that are perceptually closer to the ground truth. In order to highlight the difference in diversity, we evaluate MAP and conditional resampling on the inverse problem of inpainting missing pixels. As shown in Figure 1, 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 in Figure 1 corresponds to a run of MAP and conditional resampling starting from a random 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.

3 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 conditional resampling, even with respect to a distribution $P \neq R$ that is close in Wasserstein distance. Finally, this insight suggested a heuristic algorithm with improved performance in practice.

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.

Acknowledgments and Disclosure of Funding

Ajil Jalal and Alex Dimakis have been supported by NSF Grants CCF 1763702,1934932, AF 1901292, 2008710, 2019844 research gifts by NVIDIA, Western Digital, WNCG IAP, computing resources from TACC and the Archie Straiton Fellowship. Eric Price was supported in part by NSF Award CCF-1751040 (CAREER). Sushrut Karmalkar was supported in part by NSF CNS 1414082.

References

[1] Anders Aamand, Piotr Indyk, and Ali Vakilian. (learned) frequency estimation algorithms under zipfian distribution. *arXiv preprint arXiv:1908.05198*, 2019.

[2] Shuchin Aeron, Venkatesh Saligrama, and Manqi Zhao. Information theoretic bounds for compressed sensing. *IEEE Transactions on Information Theory*, 56(10):5111–5130, 2010.

[3] Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, and Peer-Timo Bremer. Mimicgan: Robust projection onto image manifolds with corruption mimicking. *International Journal of Computer Vision*, pages 1–19, 2020.

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[5] Muhammad Asim, Ali Ahmed, and Paul Hand. Invertible generative models for inverse problems: mitigating representation error and dataset bias. *arXiv preprint arXiv:1905.11672*, 2019.

[6] Muhammad Asim, Fahad Shamshad, and Ali Ahmed. Blind image deconvolution using deep generative priors. *arXiv preprint arXiv:1802.04073*, 2018.

[7] Muhammad Asim, Fahad Shamshad, and Ali Ahmed. Solving bilinear inverse problems using deep generative priors. *CoRR, abs/1802.04073*, 3(4):8, 2018.

[8] Benjamin Aubin, Bruno Loureiro, Antoine Baker, Florent Krzakala, and Lenka Zdeborová. Exact asymptotics for phase retrieval and compressed sensing with random generative priors. *arXiv preprint arXiv:1912.02008*, 2019.

[9] Richard G Baraniuk, Volkan Cevher, Marco F Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

[10] Richard G Baraniuk and Michael B Wakin. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

[11] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[12] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[13] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 537–546. JMLR. org, 2017.

[14] Ashish Bora, Eric Price, and Alexandros G Dimakis. Ambientgan: Generative models from lossy measurements. *ICLR*, 2:5, 2018.

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

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

[17] Emmanuel J Candes, Justin K Romberg, and Terence Tao. 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.

[18] Thierry Champion, Luigi De Pascale, and Petri Juutinen. The ∞-Wasserstein distance: Local solutions and existence of optimal transport maps. *SIAM Journal on Mathematical Analysis*, 40(1):1–20, 2008.

[19] Chen Chen and Junzhou Huang. Compressive sensing mri with wavelet tree sparsity. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1115–1123. Curran Associates, Inc., 2012.

[20] Minhua Chen, Jorge Silva, John Paisley, Chunping Wang, David Dunson, and Lawrence Carin. 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.

[21] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[22] Manik Dhar, Aditya Grover, and Stefano Ermon. Modeling sparse deviations for compressed sensing using generative models. *arXiv preprint arXiv:1807.01442*, 2018.

[23] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

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

[25] Marco F Duarte, Mark A Davenport, Dharmpal Takhar, Jason N Laska, Ting Sun, Kevin F Kelly, and Richard G Baraniuk. Single-pixel imaging via compressive sampling. *IEEE signal processing magazine*, 25(2):83–91, 2008.

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

[27] Alyson K Fletcher, Parthe Pandit, Sundeep Rangan, Subrata Sarkar, and Philip Schniter. Plug-in estimation in high-dimensional linear inverse problems: A rigorous analysis. In *Advances in Neural Information Processing Systems*, pages 7440–7449, 2018.

[28] Alyson K Fletcher, Sundeep Rangan, and Philip Schniter. Inference in deep networks in high dimensions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1884–1888. IEEE, 2018.

[29] Fabian Latorre Gómez, Armin Eftekhari, and Volkan Cevher. Fast and provable admm for learning with generative priors. *arXiv preprint arXiv:1907.03343*, 2019.

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[31] Paul Hand and Babhru Joshi. Global guarantees for blind demodulation with generative priors. In *Advances in Neural Information Processing Systems*, pages 11531–11541, 2019.

[32] Paul Hand, Oscar Leong, and Vlad Voroninski. Phase retrieval under a generative prior. In *Advances in Neural Information Processing Systems*, pages 9136–9146, 2018.

[33] Paul Hand and Vladislav Voroninski. Global guarantees for enforcing deep generative priors by empirical risk. *arXiv preprint arXiv:1705.07576*, 2017.

[34] Reinhard Heckel and Paul Hand. Deep decoder: Concise image representations from untrained non-convolutional networks. *arXiv preprint arXiv:1810.03982*, 2018.

[35] Reinhard Heckel and Mahdi Soltanolkotabi. Compressive sensing with un-trained neural networks: Gradient descent finds the smoothest approximation. *arXiv preprint arXiv:2005.03991*, 2020.

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

[37] Chinmay Hegde, Marco F Duarte, and Volkan Cevher. Compressive sensing recovery of spike trains using a structured sparsity model. In *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009.

[38] Chinmay Hegde, Michael Wakin, and Richard G Baraniuk. Random projections for manifold learning. In *Advances in neural information processing systems*, pages 641–648, 2008.

[39] Chen-Yu Hsu, Piotr Indyk, Dina Katabi, and Ali Vakilian. Learning-based frequency estimation algorithms. 2018.

[40] Piotr Indyk, Ali Vakilian, and Yang Yuan. Learning-based low-rank approximations. In *Advances in Neural Information Processing Systems*, pages 7400–7410, 2019.

[41] MA Iwen and AH Tewfik. Adaptive group testing strategies for target detection and localization in noisy environments. 2010.

[42] Gauri Jagatap and Chinmay Hegde. Phase retrieval using untrained neural network priors. 2019.

[43] Shirin Jalali and Xin Yuan. Solving linear inverse problems using generative models. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 512–516. IEEE, 2019.

[44] Maya Kabkab, Pouya Samangouei, and Rama Chellappa. Task-aware compressed sensing with generative adversarial networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[45] Akshay Kamath, Sushrut Karmalkar, and Eric Price. Lower bounds for compressed sensing with generative models. *arXiv preprint arXiv:1912.02938*, 2019.

[46] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[47] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[48] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.

[49] Qi Lei, Ajil Jalal, Inderjit S Dhillon, and Alexandros G Dimakis. Inverting deep generative models, one layer at a time. In *Advances in Neural Information Processing Systems*, pages 13910–13919, 2019.

[50] Erik M Lindgren, Jay Whang, and Alexandros G Dimakis. Conditional sampling from invertible generative models with applications to inverse problems. *arXiv preprint arXiv:2002.11743*, 2020.

[51] Zhaoqiang Liu, Selwyn Gomes, Avtansh Tiwari, and Jonathan Scarlett. Sample complexity bounds for 1-bit compressive sensing and binary stable embeddings with generative priors. *arXiv preprint arXiv:2002.01697*, 2020.

[52] Zhaoqiang Liu and Jonathan Scarlett. Information-theoretic lower bounds for compressive sensing with generative models. *arXiv preprint arXiv:1908.10744*, 2019.

[53] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15:2018, 2018.

[54] Michael Lustig, David Donoho, and John M Pauly. 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.

[55] Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly. Compressed sensing mri. *IEEE signal processing magazine*, 25(2):72–82, 2008.

[56] Morteza Mardani, Enhao Gong, Joseph Y Cheng, Shreyas S Vasanawala, Greg Zaharchuk, Lei Xing, and John M Pauly. Deep generative adversarial neural networks for compressive sensing mri. *IEEE transactions on medical imaging*, 38(1):167–179, 2018.

[57] Chris Metzler, Ali Mousavi, and Richard Baraniuk. Learned d-amp: Principled neural network based compressive image recovery. In *Advances in Neural Information Processing Systems*, pages 1772–1783, 2017.

[58] Lukas Mosser, Olivier Dubrule, and Martin J Blunt. Stochastic seismic waveform inversion using generative adversarial networks as a geological prior. *Mathematical Geosciences*, 52(1):53–79, 2020.

[59] Gregory Ongie, Ajil Jalal, Christopher A Metzler, Richard G Baraniuk, Alexandros G Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *arXiv preprint arXiv:2005.06001*, 2020.

[60] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[61] Parthe Pandit, Mojtaba Sahraee-Ardakan, Sundeep Rangan, Philip Schniter, and Alyson K Fletcher. Inference with deep generative priors in high dimensions. *arXiv preprint arXiv:1911.03409*, 2019.

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

[63] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

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

[65] Eric Price and David P Woodruff. (1+ eps)-approximate sparse recovery. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 295–304. IEEE, 2011.

[66] Shuang Qiu, Xiaohan Wei, and Zhuoran Yang. Robust one-bit recovery via relu generative networks: Improved statistical rates and global landscape analysis. *arXiv preprint arXiv:1908.05368*, 2019.

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

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

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

[70] Ganlin Song, Zhou Fan, and John Lafferty. Surfing: Iterative optimization over incrementally trained deep networks. In *Advances in Neural Information Processing Systems*, pages 15008–15017, 2019.

[71] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pages 11918–11930, 2019.

[72] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[73] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[74] Jay Whang, Qi Lei, and Alexandros G Dimakis. Compressed sensing with invertible generative models and dependent noise. *arXiv preprint arXiv:2003.08089*, 2020.

[75] Yihong Wu. *Shannon theory for compressed sensing*. Citeseer, 2011.

[76] Yihong Wu and Sergio Verdú. Optimal phase transitions in compressed sensing. *IEEE Transactions on Information Theory*, 58(10):6241–6263, 2012.

[77] Weiyu Xu and Babak Hassibi. Compressed sensing over the grassmann manifold: A unified analytical framework. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 562–567. IEEE, 2008.

[78] Weiyu Xu, Enrique Mallada, and Ao Tang. Compressive sensing over graphs. In *2011 Proceedings IEEE INFOCOM*, pages 2087–2095. IEEE, 2011.

[79] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[80] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[81] Zhou Zhou, Kaihui Liu, and Jun Fang. Bayesian compressive sensing using normal product priors. *IEEE Signal Processing Letters*, 22(5):583–587, 2014.

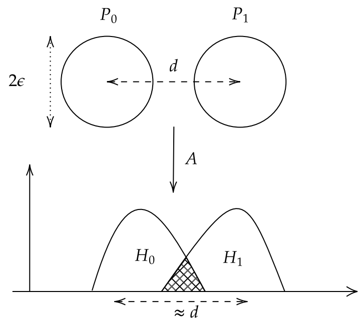


Figure 3: Illustrative example for the upper bound. The signal is x is drawn from a mixture of two well-separated balls. The observations $y = Ax$ are then drawn from a mixture of two distributions H_0, H_1 that may overlap. The probability that conditional resampling outputs something from the wrong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

4 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. We wish to handle cases where the distribution of reality R is not the same as the distribution P of our model. The metric we use to quantify the similarity between distributions is the Wassertein distance, which is defined for any $p \geq 1$ as

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

where μ, ν are probability distributions on some metric space Ω, and $\Pi(\mu, \nu)$ denotes the set of joint distributions whose marginals are μ, ν.

We will also need the limiting case of $p = \infty$. Following the definition in [18], the infinite Wasserstein distance is defined as

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

Informally, 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. Note that the definition uses the essential supremum, and hence whenever we use $\mathcal{W}_\infty$, all statements are made almost surely over γ.

See [73, 63] for more details, and [4] for applications of Wasserstein distances to generative modelling.

5 Theoretical Results

In this section we state the formal versions of Theorem 1.3, 1.4 and give proof sketches for them.

5.1 Upper Bound

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 ε radius balls can cover the whole space, so the parameters in Theorem (1.3) will be $\sigma = 0, \eta = \varepsilon$, and $\mathrm{Cov}_{\varepsilon,0}(P) = 2$. Applying Theorem (1.3) on P tells us that a constant number of measurements is enough to get $O(\varepsilon)$ close to the ground truth. We will now prove this claim.

Let $B_0, B_{\tilde{x}}$ denote $\varepsilon-$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 $d \gg \varepsilon$.

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 conditional 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.

Note that the ground truth and the reconstruction are far apart if and only if they are drawn from different components of P. It turns out that the probability of the event $\{x^* \in B_0, \widehat{x} \in B_{\tilde{x}}\}$ is bounded by how similar the distributions H_0, H_1 are:

Lemma 5.1. *For $c \in [0, 1]$, let $P := (1 - c)P_0 + cP_1$ be a mixture of two absolutely continuous distributions P_0, P_1 admitting densities p_0, p_1. Let y be a sample from the distribution P, such that $y|z^* \sim P_{z^*}$ where $z^* \sim Bernoulli(c)$.*

Define $\widehat{c}_y = \frac{cp_1(y)}{(1-c)p_0(y)+cp_1(y)}$, and let $\widehat{z}|y \sim Bernoulli(\widehat{c}_y)$ be the Bayes optimal estimate of z^* given y. Then we have

$$\Pr_{z^*, y, \widehat{z}}[z^* = 0, \widehat{z} = 1] \leq 1 - TV(P_0, P_1).$$

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 ε, a union bound of the above two probabilities gives:

$$\begin{aligned} \Pr\left[\|x^* - \widehat{x}\| > 2\varepsilon\right] &\leq \Pr\left[x^* \in B_0, \widehat{x} \in B_{\tilde{x}}\right] + \Pr\left[x^* \in B_{\tilde{x}}, \widehat{x} \in B_0\right], && (3) \\ &\leq 2\left(1 - TV\left(H_0, H_1\right)\right). && (4) \end{aligned}$$

If A is a Gaussian random matrix, the Johnson-Lindenstrauss (JL) Lemma tell us that it will with high probability preserve distances between vectors. 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:

Lemma 5.2. *Let $A \in \mathbb{R}^{m \times n}$ be a random Gaussian matrix with i.i.d. entries $A_{ij} \sim \mathcal{N}(0, 1/m)$, and let $\xi \sim \mathcal{N}(0, \frac{\sigma^2}{m} I_m)$. Consider observations of the form $y = Ax^* + \xi$. 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 $\tilde{x}$.

When $x^* \sim P_{\tilde{x}}$, and A is fixed, let $H_{\tilde{x}}$ denote the corresponding distribution of y. Similarly, let H_{out} denote the distribution of y when $x^* \sim P_{out}$ and A is fixed.

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

$$TV(H_0, H_1) \geq 1 - e^{-\Omega(m)}. \tag{5}$$

Substituting equation (5) in equation (4), we have

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

This shows that conditional sampling will produce a reconstruction which is close to the ground truth with overwhelmingly high probability for our current example.

This example provides intuition, but 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?

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 5.2 with a union bound. There are a few details (e.g., Lemma 5.2 shows you will not confuse any ball with faraway balls, but you might confuse it with nearby balls) but solving it is straightforward. This shows that, if $P = R$ and $\log \mathrm{Cov}_{\eta,0}(R)$ is bounded, then conditional resampling works well with $1 - e^{-\Omega(m)}$ probability.

Now, what if $P \neq R$? We would like to show that observing samples from R and resampling according to P gives good results. We first show that the results of this process are similar to those of sampling from R' and resampling according to P, where R' is any distribution $\mathcal{W}_\infty$-close to R:

Lemma 5.3. *Let R, R', denote arbitrary distributions over $\mathbb{R}^n$ such that $\mathcal{W}_\infty(R, R') \leq \varepsilon$.*

Let $A \in \mathbb{R}^{m\times n}$ be a random Gaussian matrix whose entries are i.i.d. and scaled so that $A_{ij} \sim \mathcal{N}(0, 1/m)$. For $x^* \sim R$, let $y = Ax^* + \xi$, where $\xi \sim \mathcal{N}(0, \frac{\sigma^2}{m} I_m)$. For $z^* \sim R'$, let $u = Az^* + \xi$ denote the corresponding measurements.

Let $\widehat{x}$ be a random variable whose distribution conditioned on y, A is $P(\cdot|y, A)$. Similarly, let $\widehat{z}$ be a random variable whose distribution conditioned on u, A is $P(\cdot|u, A)$. For any $d > 0$, we have

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

Now, if $\mathcal{W}_\infty(R, P) \ll \sigma$, we would be done: we could set $R' = P$, and find that the error probability of our problem is within $e^{o(m)}$ of that of sampling $x^* \sim P$ and resampling according to P, which (if P has a small cover) is $e^{-\Omega(m)}$.

Finally, we need to confront the third problem: we want to allow our distributions to have a small constant probability of behaving badly. R can have a δ mass outside the covering, and P is only close to R in $\mathcal{W}_1$ not $\mathcal{W}_\infty$. 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 compressed sensing would work with R' and P', it also works with R and P:

Theorem 5.4. *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 $A \in \mathbb{R}^{m\times n}$ be a random Gaussian matrix with i.i.d. entries $A_{ij} \sim \mathcal{N}(0, 1/m)$ for $m \geq O(\min(\log \mathrm{Cov}_{\eta,\delta}(R), \log \mathrm{Cov}_{\eta,\delta}(P)))$, and let $\xi \sim \mathcal{N}(0, \frac{\sigma^2}{m} I_m)$ for some $\sigma \gtrsim \varepsilon/\delta^{1/p}$.

For $x^* \sim R$, consider observations of the form $y = Ax^* + \xi$. Given y and the fixed matrix A, let $\widehat{x}$ be the reconstruction of x^*, sampled according to the conditional $P(\widehat{x}|y)$.

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.

5.2 Lower Bound

In the previous section, we showed, for any distribution R of signals, that $O(\log \mathrm{Cov}(R))$ measurements suffice for conditional resampling 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 5.5. *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*

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

This is proven using information theory, as a direct consequence of the following three lemmas. First, the measurement process reveals a limited amount of information:

Lemma 5.6. *Consider the setting of Theorem* (5.5)*. We have*

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

Second, a variant of the Data Processing Inequality is true for our measurement process, where A, x^*, are independent.

Lemma 5.7. *Consider the setting of Theorem* (5.5)*. We have*

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

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

Lemma 5.8 (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*

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

The proofs can be found in Appendix B.

A Upper Bound Proofs

A.1 Proof of Lemma 5.1

Lemma 5.1. *For* $c \in [0,1]$*, let* $P := (1-c)P_0 + cP_1$ *be a mixture of two absolutely continuous distributions* P_0, P_1 *admitting densities* p_0, p_1*. Let* y *be a sample from the distribution* P*, such that* $y|z^* \sim P_{z^*}$ *where* $z^* \sim Bernoulli(c)$*.*

Define $\widehat{c}_y = \frac{cp_1(y)}{(1-c)p_0(y)+cp_1(y)}$*, and let* $\widehat{z}|y \sim Bernoulli(\widehat{c}_y)$ *be the Bayes optimal estimate of* z^* *given* y*. Then we have*

$$\Pr_{z^*,y,\widehat{z}}[z^*=0, \widehat{z}=1] \leq 1 - TV(P_0, P_1).$$

Proof. We have

$$\begin{aligned}\Pr_{z^*,y,\widehat{z}}[z^*=0,\widehat{z}=1] &= \Pr[z^*=0] \mathop{\mathbb{E}}_{y\sim p_0, \widehat{z}|y}[1\{\widehat{z}=1\}], &(6)\\ &= (1-c)\int p_0(y)\Pr[\widehat{z}=1|y]dy. &(7)\end{aligned}$$

By definition, we have

$$\Pr[\widehat{z}=1|y] = \frac{cp_1(y)}{(1-c)p_0(y)+cp_1(y)}.$$

Substituting, we have

$$\begin{aligned}\Pr_{z^*,y,\widehat{z}}[z^*=0,\widehat{z}=1] &= \int \frac{(1-c)p_0(y)cp_1(y)}{(1-c)p_0(y)+cp_1(y)}dy\\ &\leq \int \frac{(1-c)p_0(y)\cdot cp_1(y)}{\max\{(1-c)p_0(y), cp_1(y)\}}dy\end{aligned}$$

$$\begin{aligned}
&= \int \min\{(1-c)p_0(y), cp_1(y)\}dy \\
&\leq \int \min\{p_0(y), p_1(y)\}dy \\
&= (1 - TV(P_0, P_1)).
\end{aligned}$$

□

A.2 Proof of Lemma 5.2

Lemma 5.2. *Let $A \in \mathbb{R}^{m\times n}$ be a random Gaussian matrix with i.i.d. entries $A_{ij} \sim \mathcal{N}(0, 1/m)$, and let $\xi \sim \mathcal{N}(0, \frac{\sigma^2}{m} I_m)$. Consider observations of the form $y = Ax^* + \xi$. 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 $\tilde{x}$.

When $x^* \sim P_{\tilde{x}}$, and A is fixed, let $H_{\tilde{x}}$ denote the corresponding distribution of y. Similarly, let H_{out} denote the distribution of y when $x^* \sim P_{out}$ and A is fixed.

Then we have,

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

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

$$\mathop{\mathbb{E}}_{A}[H_{out}(B)] \leq 2e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}, \tag{8}$$

$$\mathop{\mathbb{E}}_{A}[H_{\tilde{x}}(B)] \geq 1 - 2e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)}. \tag{9}$$

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 (8).

Consider the joint distribution of y, A. We have

$$\mathop{\mathbb{E}}_{A}\left[H_{out}(B)\right] = \mathop{\mathbb{E}}_{A}\left[\mathop{\mathbb{E}}_{x\sim P_{out}}\left[\mathcal{N}\left(Ax, \frac{\sigma^2}{m} I_m\right)(B)\right]\right], \tag{10}$$

$$= \mathop{\mathbb{E}}_{x\sim P_{out}}\left[\mathop{\mathbb{E}}_{A}\left[\mathcal{N}(Ax, \sigma^2/m)(B)\right]\right], \tag{11}$$

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

$$\Pr_{A}\left[\|Ax - A\tilde{x}\| < 2\sqrt{c}\,(\eta + \sigma)\right] \leq \Pr_{A}\left[\|Ax - A\tilde{x}\| \leq \frac{2}{\sqrt{c}}\|x - \tilde{x}\|\right],$$

$$
\begin{aligned}
&\leq \frac{2}{\sqrt{m\pi}} \left(\frac{2e}{\sqrt{c}} \right)^m, \\
&= \frac{2}{\sqrt{m\pi}} e^{-\frac{m}{2} \log\left(\frac{c}{4e^2}\right)}, \\
&\leq e^{-\frac{m}{2} \log\left(\frac{c}{4e^2}\right)} \quad \text{if } m > 1.
\end{aligned}
$$

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 (11), 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 (8).

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 (9).

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 $\min\{\log \mathrm{Cov}_{\eta,\delta}(P), \log \mathrm{Cov}_{\eta,\delta}(R)\} \leq k$, 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 $\log \mathrm{Cov}_{\eta,\delta}(P) \leq k$. 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

$$\begin{aligned} \sum_{u\in S} Q''(u) &= \sum_{u \in S} \int_{B(u,\eta)} f dP, \\ &= \int_{\cup_{u\in S} B(u,\eta)} dP = 1 - c_P \end{aligned}$$

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

$$c_P = \Pr_{(x_1,x_2)\sim\Pi}[\|x_1 - x_2\| \geq \eta] \leq \delta. \tag{12}$$

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

$$c_R = \Pr_{(x_1,x_2)\sim\Gamma}\left[\|x_1 - x_2\| \geq \frac{\rho}{\delta^{1/p}}\right] \leq \frac{\mathbb{E}\left[\|x_1 - x_2\|^p\right]}{\frac{\rho^p}{\delta}} \leq \delta. \tag{13}$$

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 (12). 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 $\operatorname{supp}(Q) < e^k$, and hence we have our result.

□

A.4 Proof of Lemma 5.3

Lemma 5.3. *Let R, R', denote arbitrary distributions over $\mathbb{R}^n$ such that $\mathcal{W}_\infty(R, R') \leq \varepsilon$.*

Let $A \in \mathbb{R}^{m \times n}$ be a random Gaussian matrix whose entries are i.i.d. and scaled so that $A_{ij} \sim \mathcal{N}(0, 1/m)$. For $x^* \sim R$, let $y = Ax^* + \xi$, where $\xi \sim \mathcal{N}(0, \frac{\sigma^2}{m} I_m)$. For $z^* \sim R'$, let $u = Az^* + \xi$ denote the corresponding measurements.

Let $\widehat{x}$ be a random variable whose distribution conditioned on y, A is $P(\cdot|y, A)$. Similarly, let $\widehat{z}$ be a random variable whose distribution conditioned on u, A is $P(\cdot|u, A)$. For any $d > 0$, we have

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

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, R'. 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 $W_\infty(R, R') \leq \varepsilon$, 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 $\|Ax^* - Az^*\| \le 2\varepsilon$, and another case where $\|Ax^* - Az^*\| > 2\varepsilon$. This gives

$$\begin{aligned}\Pr_{x^*,A,\xi,\widehat{x}\sim P(\cdot|A,y)}[B_1] &\le \mathop{\mathbb{E}}_{x^*,z^*}\mathop{\mathbb{E}}_{A}\left[1_{\|Ax^*-Az^*\|>2\varepsilon}\mathop{\mathbb{E}}_{y|A,x^*}\left[\mathop{\mathbb{E}}_{\widehat{x}\sim P(\cdot|y,A)}[1_{B_2}]\right]\right](*) \qquad (14)\\ &+ \mathop{\mathbb{E}}_{x^*,z^*}\mathop{\mathbb{E}}_{A}\left[1_{\|Ax^*-Az^*\|\le 2\varepsilon}\mathop{\mathbb{E}}_{y|A,x^*}\left[\mathop{\mathbb{E}}_{\widehat{x}\sim P(\cdot|y,A)}[1_{B_2}]\right]\right].(**) \qquad (15)\end{aligned}$$

Consider the term$(*)$ in line (14). We have

$$\begin{aligned}\mathop{\mathbb{E}}_{x^*,z^*}\mathop{\mathbb{E}}_{A}\left[1_{\|Ax^*-Az^*\|>2\varepsilon}\mathop{\mathbb{E}}_{y|A,x^*}\left[\mathop{\mathbb{E}}_{\widehat{x}\sim P(\cdot|y,A)}[1_{B_2}]\right]\right] &\le \mathop{\mathbb{E}}_{x^*,z^*}\mathop{\mathbb{E}}_{A}\left[1_{\|Ax^*-Az^*\|>2\varepsilon}\right], \qquad (16)\\ &\le \mathop{\mathbb{E}}_{x^*,z^*}\left[e^{-\Omega(m)}\right] \le e^{-\Omega(m)}, \qquad (17)\end{aligned}$$

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 (15). 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

   $$\begin{aligned}\|Ax^* - y\|^2 &\ge -4\varepsilon^2 - 8\sigma\varepsilon + \|y - Az^*\|^2, \qquad (18)\\ \Rightarrow \exp\left(-\frac{\|Ax^* - y\|^2}{2\sigma^2/m}\right) &\le \exp\left(\frac{4\varepsilon\left(\varepsilon + 2\sigma\right)m}{2\sigma^2}\right)\exp\left(-\frac{\|Az^* - y\|^2}{2\sigma^2/m}\right), \qquad (19)\\ & \qquad (20)\end{aligned}$$

   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

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

From the above inequality and eqn. (17), we have

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

□

A.5 Proof of Theorem 5.4

Theorem 5.4. *Let $\delta \in [0,1/4)$, $p \ge 1$, and $\varepsilon, \eta > 0$ be parameters. Let R, P be arbitrary distributions over $\mathbb{R}^n$ satisfying $\mathcal{W}_p(R,P) \le \varepsilon$.*

Let $A \in \mathbb{R}^{m\times n}$ be a random Gaussian matrix with i.i.d. entries $A_{ij} \sim \mathcal{N}(0,1/m)$ for $m \ge O(\min(\log \mathrm{Cov}_{\eta,\delta}(R), \log \mathrm{Cov}_{\eta,\delta}(P)))$, and let $\xi \sim \mathcal{N}(0, \frac{\sigma^2}{m}I_m)$ for some $\sigma \gtrsim \varepsilon/\delta^{1/p}$.

For $x^* \sim R$, consider observations of the form $y = Ax^* + \xi$. Given y and the fixed matrix A, let $\widehat{x}$ be the reconstruction of x^*, sampled according to the conditional $P(\widehat{x}|y)$.

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

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

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

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

$$\begin{aligned}&\Pr_{x^*\sim R,\widehat{x}\sim P(\cdot|y)}\left[\|x^*-\widehat{x}\| \ge (c+1)\eta + (c+1)\sigma\right] && (21)\\ &\le \Pr_{x^*\sim R,\widehat{x}\sim P(\cdot|y)}\left[\|x^*-\widehat{x}\| \ge (c+1)\eta + c\sigma + (\varepsilon/\delta^{1/p})\right] && (22)\\ &\le 2\delta + (1-2\delta)\Pr_{x^*\sim R',\widehat{x}\sim P(\cdot|y)}\left[\|x^*-\widehat{x}\| \ge (c+1)\eta + c\sigma + (\varepsilon/\delta^{1/p})\right] && (23)\end{aligned}$$

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 5.3, 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

$$\begin{aligned}&\Pr_{x^*\sim R',A,\xi,\widehat{x}\sim P(\cdot|A,y)}\left[\|x^*-\widehat{x}\| \ge (c+1)\,\eta + c\sigma + (\varepsilon/\delta^{1/p})\right] \le \\ &e^{-\Omega(m)} + e^{\left(\frac{2(\varepsilon/\delta^{1/p})\left((\varepsilon/\delta^{1/p})+2\sigma\right)m}{\sigma^2}\right)} \Pr_{z^*\sim P',A,\xi,\widehat{z}\sim P(\cdot|u,A)}\left[\|z^*-\widehat{z}\| \ge (c+1)\eta + c\sigma\right]. \qquad (24)\end{aligned}$$

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 $\|z^* - \tilde{z}^*\| \leq \eta$. 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,

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

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

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

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 5.2 we have

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

This implies

$$\begin{aligned}\Pr_{z^* \sim P', A, \xi, \widehat{z} \sim P(\cdot|u,A)}\left[\|z^* - \widehat{z}\| \geq (c+1)\eta + c\sigma\right] &\leq \frac{1}{1-2\delta}\sum_{\tilde{z}^* \in S}\mathop{\mathbb{E}}_{A}\left[(1 - TV(H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2}))\right],\\ &\leq \frac{1}{1-2\delta}4|S|e^{-\frac{m}{2}\log\left(\frac{c}{4e^2}\right)},\\ &\leq \frac{1}{1-2\delta}4e^{-\frac{m}{4}\log\left(\frac{c}{4e^2}\right)},\end{aligned}$$

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

Substituting in Eqn (24), 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 (21), 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*

$$\mathop{\mathbb{E}}_{A,\xi,z^*}\Pr\left[z^* \sim P_{\tilde{z}^*,0}, \widehat{z} \sim P_{\tilde{z}^*,2}(\cdot|u)\right] \leq \frac{1}{1-\delta_2}\mathop{\mathbb{E}}_{A}\left[1 - TV(H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2})\right], \tag{25}$$

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

$$\begin{aligned}\mathop{\mathbb{E}}_{A,\xi,z^*}\Pr\left[z^* \sim P_{\tilde{z}^*,0}, \widehat{z} \sim P_{\tilde{z}^*,2}(\cdot|u)\right] &= \mathop{\mathbb{E}}_{A}\int \frac{Q_{\tilde{z}^*}h_{\tilde{z}^*,0}(u)c_{\tilde{z}^*,2}h_{\tilde{z}^*,2}(u)}{(1-\delta_2)\,Q_{\tilde{z}^*,0}h_{\tilde{z}^*,0}(u) + c_{\tilde{z}^*,1}h_{\tilde{z}^*,1}(u) + c_{\tilde{z}^*,2}h_{\tilde{z}^*,2}(u)}du, &(26)\\ &\leq \mathop{\mathbb{E}}_{A}\int \frac{Q_{\tilde{z}^*}h_{\tilde{z}^*,0}(u)c_{\tilde{z}^*,2}h_{\tilde{z}^*,2}(u)}{(1-\delta_2)\,Q_{\tilde{z}^*,0}h_{\tilde{z}^*,0}(u) + c_{\tilde{z}^*,2}h_{\tilde{z}^*,2}(u)}du, &(27)\\ &\leq \mathop{\mathbb{E}}_{A}\int \frac{Q_{\tilde{z}^*}h_{\tilde{z}^*,0}(u)c_{\tilde{z}^*,2}h_{\tilde{z}^*,2}(u)}{(1-\delta_2)\,Q_{\tilde{z}^*,0}h_{\tilde{z}^*,0}(u) + (1-\delta_2)c_{\tilde{z}^*,2}h_{\tilde{z}^*,2}(u)}du, &(28)\\ &\leq \mathop{\mathbb{E}}_{A}\frac{1}{1-\delta_2}\int \frac{Q_{\tilde{z}^*}h_{\tilde{z}^*,0}(u)c_{\tilde{z}^*,2}h_{\tilde{z}^*,2}(u)}{Q_{\tilde{z}^*,0}h_{\tilde{z}^*,0}(u) + c_{\tilde{z}^*,2}h_{\tilde{z}^*,2}(u)}du, &(29)\end{aligned}$$

$$\leq \mathop{\mathbb{E}}_{A} \frac{1}{1-\delta_2} \int \frac{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u) c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)}{\max\{Q_{\tilde{z}^*,0} h_{\tilde{z}^*,0}(u)\,,\, c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)\}} du, \tag{30}$$

$$= \mathop{\mathbb{E}}_{A} \frac{1}{1-\delta_2} \int \min\{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u), c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)\} du, \tag{31}$$

$$\leq \mathop{\mathbb{E}}_{A} \frac{1}{1-\delta_2} \int \min\{h_{\tilde{z}^*,0}(u), h_{\tilde{z}^*,2}(u)\} du, \tag{32}$$

$$= \frac{1}{1-\delta_2} \mathop{\mathbb{E}}_{A} \left[1 - TV(H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2})\right]. \tag{33}$$

□

B Lower Bound Proofs

B.1 Proof of Lemma 5.6

Lemma 5.6. *Consider the setting of Theorem* (5.5)*. 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 [21, 64, 69], 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-addditivity of entropy implies,

$$I(y; z) = h(y) - h(y|z) = h(y) - h(y - z|z), \tag{34}$$

$$= h(y) - h(\xi|z) = h(y) - \sum h(\xi_i|z, \xi_1, \cdots, \xi_{i-1}), \tag{35}$$

$$= h(y) - \sum h(\xi_i), \tag{36}$$

$$\leq \sum h(y_i) - \sum h(\xi_i), \tag{37}$$

$$= \sum h(y_i) - \sum h(y_i|z_i), \tag{38}$$

$$= \sum I(y_i; z_i), \tag{39}$$

$$\leq \frac{m}{2} \log \left(1 + \frac{1}{\sigma^2}\right). \tag{40}$$

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

This gives

$$I(x^*; y|A) + I(z; y|x^*, A) = I(z; y|A), \tag{41}$$

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

Figure 4: DAG relating $x^*, A, z, y, \widehat{x}$. The conditional independencies we use are $x^* \perp\!\!\!\perp y|z, A$ and $A \perp\!\!\!\perp y|z$.

$$\Rightarrow I(x^*; y|A) \leq I(z; y|A). \tag{42}$$

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

$$\begin{aligned}
I(A, z; y) &= I(A; y) + I(z; y|A), & (43)\\
&= I(z; y) + I(A; y|z), & (44)\\
\Leftrightarrow I(A; y) + I(z; y|A) &= I(z; y) + I(A; y|z), & (45)\\
\Leftrightarrow I(A; y) + I(z; y|A) &= I(z; y), & (46)\\
\Rightarrow I(z; y|A) &\leq I(z; y), . & (47)
\end{aligned}$$

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

From Eqn (40), (42), (47), we have

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

□

B.2 Proof of Lemma 5.7

Lemma 5.7. *Consider the setting of Theorem* (5.5)*. We have*

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

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:

$$\begin{aligned}
I(x^*; y, A, \widehat{x}) &= I(x^*; y, A) + I(x^*; \widehat{x}|y, A), & (48)\\
&= I(x^*; \widehat{x}) + I(x^*; y, A|\widehat{x}). & (49)
\end{aligned}$$

As $\widehat{x}$ is a function of y, A, we have $I(x^*; \widehat{x}|y, A) = 0$. Also, $I(x^*; y, A|\widehat{x}) \geq 0$. Substituting in Eqn (48), (49), we have

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

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 5.8

We will build up Lemma 5.8 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 5.8).

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*

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

Then we have

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

Proof. Recall,

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

For any $v \in \operatorname{supp}(\widehat{x})$, 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 $G_1 \subseteq \operatorname{supp}(x, \widehat{x})$ with $\Pr[G_1] \ge 1 - \tau$, such that for all $(u, v) \in G_1$,

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

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 $G_2 \subseteq \operatorname{supp}(x, \widehat{x})$ satisfying $\Pr[G_2] \ge 1 - \tau - \delta$, such that for all $(u, v) \in G_2$, we have

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

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. □

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

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

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:

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

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

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

Applying Lemma B.1 on Q', we have

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

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

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

□

We now prove Lemma 5.8, which allows $(x, \widehat{x})$ to follow an arbitrary distribution.

Lemma 5.8 (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*

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

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

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

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 $\|x - \widehat{x}\| \le \varepsilon$ with probability $1 - \delta$, an application of the triangle inequality gives

$$\|\bar{x} - \bar{\widehat{x}}\| \le \varepsilon + 2\gamma\sqrt{n} \text{ with probability } \ge 1 - \delta. \tag{50}$$

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 $\bar{\widehat{x}}$ is a function of $\widehat{x}$ alone, by the Data Processing Inequality, we have

$$I(\bar{x}; \bar{\widehat{x}}) \le I(x; \widehat{x}). \tag{51}$$

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 [21], 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 (51) gives us,

$$I(x; \widehat{x}) \geq I(\bar{x}; \widehat{\bar{x}}) \geq (1 - c_1) I_Q(\bar{x}; \widehat{\bar{x}}) + c_1 I_D(\bar{x}; \widehat{\bar{x}}), \tag{52}$$

$$\geq (1 - c_1) I_Q(\bar{x}; \widehat{\bar{x}}), \tag{53}$$

$$\geq (1 - \beta) I_Q(\bar{x}; \widehat{\bar{x}}). \tag{54}$$

Now since the finite distribution Q has a TV distance of at most β to the countable distribution R, using Eqn (50), we have

$$\|\bar{x} - \widehat{\bar{x}}\| \leq \varepsilon + 2\gamma\sqrt{n} \text{ with probability } \geq 1 - \beta - \delta \text{ if } \bar{x} \sim Q. \tag{55}$$

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 (55), 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

$$\|\bar{x} - \widehat{\bar{x}}\| \leq \varepsilon + 2\gamma\sqrt{n} \text{ with probability } \geq 1 - \beta - \delta - \alpha N_\beta \text{ if } \bar{x} \sim Q'. \tag{56}$$

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 (54), we have

$$\frac{I(x; \widehat{x})}{1 - \beta} \geq I_Q(\bar{x}; \widehat{\bar{x}}) \geq (1 - c_2) I_{Q'}(\bar{x}; \widehat{\bar{x}}) + c_2 I_{Q''}(\bar{x}; \widehat{\bar{x}}), \tag{57}$$

$$\geq (1 - c_2) I_{Q'}(\bar{x}; \widehat{\bar{x}}), \tag{58}$$

$$\geq (1 - \alpha N_\beta) I_{Q'}(\bar{x}; \widehat{\bar{x}}). \tag{59}$$

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

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

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

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

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 5.5

Theorem 5.5. *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 5.7 and Lemma 5.6, 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 5.8 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

We borrowed the RealNVP model trained by [74], which was trained on the first 27,000 images in the CelebA-HQ dataset [46]. Please see Appendix C in [74] for detailed hyperparameters used in the training of the model. We evaluate our experiments on the first 30 images in the evaluation dataset of [5].

In our experiments, the Gaussian noise in the measurements (ξ) satisfies $\sqrt{\mathbb{E}\left[\|\xi\|^2\right]} = 20$.

For the inpainting experiment in Figure 1c we borrowed used the 256×256 GLOW model from the official repository.

C.2 Hyperparameter Selection

In order to pick the best hyperparameter, we use 500 fresh noisy measurements $y' = A'x^* + \xi'$, which were not used in optimization. The hyperparameters that give the lowest validation error on these measurements were picked.

For MAP, we tried an Adam and Gradient Descent optimizer. Langevin dynamics only uses Gradient Descent. Each algorithm was run with learning rates varying over $\left[0.1, 0.01, 0.001, 5\cdot10^{-4}, 10^{-4}, 5\cdot10^{-5}, 10^{-5}, 5\cdot10^{-6}, 10^{-6}\right]$. We also performed 2 random restarts for the initialization z_0. for all algorithms.

C.3 Computing Infrastructure

Experiments were run on an NVIDA Quadro P5000.