

Copyright
by
Ajil Jalal
2022

The Dissertation Committee for Ajil Jalal
certifies that this is the approved version of the following dissertation:

**Compressed Sensing using Generative Models: Theory
and Applications**

Committee:

Alexandros G. Dimakis, Supervisor

Constantine Caramanis

Eric Price

Sanjay Shakkottai

Sujay Sanghavi

**Compressed Sensing using Generative Models: Theory
and Applications**

by

Ajil Jalal

DISSERTATION

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2022

This dissertation is dedicated to my brother. Thank you for the many laughs
and fights.

Acknowledgments

First and foremost, I sincerely thank my advisor, Alex Dimakis. Alex's ideas have laid the foundations of this dissertation (in fact, my first meeting with him was the basis for our ICML 2017 paper), and his infectious passion for research has played a significant role in my decision to pursue academia. He genuinely cares for the success and well-being of his students, and this dissertation would not have been possible without his unwavering support during the COVID-19 pandemic and the more difficult parts of my academic journey.

I have had the immense fortune of working with Eric Price throughout my Ph.D. Eric's mathematical ability is incredible, as is his ability to patiently explain himself even after I've asked him the same thing twenty times over. Looking back on the past six years, I am amazed at how much Alex and Eric have influenced my professional and personal growth. You have been phenomenal mentors and friends, and I wish I can be half as great as you to my own students.

The professors on my thesis committee have taught me several classes over the years, and they continue to provide invaluable advice. You are my reference for academic scholarship, and I have had a truly wonderful time learning from you. Thank you for everything!

I would like to thank my collaborators: interactions with you have a played a huge role in shaping my research progress and outlook. Chapter 2 is based on work with Ashish, who taught me the basics of Deep Learning. Chapter 3 is with Liu and Constantine, who helped me navigate the world of robust estimation. Chapter 4 and 5 are a result of long hours with Sushrut, who has been a good friend and roommate. Jessica was invaluable for the work in Chapter 5, and I have enjoyed our philosophical discussions. Chapter 6 is the result of a wonderful collaboration with Jon, Marius, and Giannis. My summer at IBM with Karthikeyan Shanmugam and Bhanukiran Vinzamuri introduced me to research in causality and led to a patent. My work with Giannis, Joseph, Qi, Akash, Eren, Dave, Mahdi, Andrew, Costis, Sriram, Jeff, and Inderjeet have not been included in this thesis. I have learned a lot from all of you, and look forward to future collaborations.

This thesis is one small part of a long and ongoing academic journey, and I would like to thank the teachers and mentors who have helped me along the way. Mrs. Jyotsna Shrivastava showed immense faith and belief in me during some of the rougher phases of my life. Professors Krishna Jagannathan, Umang Bhaskar, and Rahul Vaze were incredibly inspiring and crucial in developing my love for research, and they continue to be cherished sources of advice and perspective. I am in awe of the patience and kindness you all showed me.

My sincere gratitude to Karen, Apipol, Jaymie, Melody, and Melanie, whose administrative wizardry has helped ease my life as an international

student.

I have made incredible friends during my time in Austin. In no particular order: Murat, Rajat, Mridula, Soumya, Vampire who Baked, Shalmali, Eirini, Ahmad, Erik, Sara, Vatsal, Abhishek, Neha, Rajiv, Ethan, Isidoros, Nithin, Nihal, Liam, Karl, Manan, Advait, Parikshit, Sravan, Shubhankar, Anish, Giannis, Georgios, Sriram, Matt, Ashish, Arvind, Shivam, Sushrut, and Venkat have made my time in Austin truly special.

My dear friends Adit, Siddharth, and Amritha: thank you for the spontaneous calls and messages, I am blessed to have you in my life.

I am indebted to Feroze Gani and family, who have helped me right from my first move to Austin.

Nataša and Ricci have shaped me into a better person through their constant love, encouragement, and support. You made Austin feel like home, thank you for all the adventures.

Finally, my infinite love to Kukoo, Chechi, Ummichi, Vappichi, Umma, Vappa, Ummi, and Uppi. A paragraph in a dissertation is an inadequate way to thank you for all you have done for me. You make me want to be a better version of myself.

Compressed Sensing using Generative Models: Theory and Applications

Publication No. _____

Ajil Jalal, Ph.D.

The University of Texas at Austin, 2022

Supervisor: Alexandros G. Dimakis

Deep generative models, such as Generative Adversarial Networks, Variational Autoencoders, Flow-based models, and Score-based models, have shown excellent performance in modelling high-dimensional distributions. This dissertation proposes a novel framework that can leverage the power of these models for solving various problems in signal processing, such as compressed sensing, inpainting, and super-resolution, to name a few.

On the theoretical side, we propose algorithms that provably achieve optimal sample complexities and are also robust to different kinds of corruptions. On the practical side, we show that our algorithms can provide performance improvements over classical algorithms such as the LASSO. Furthermore, our most recent work shows that our approach achieves state-of-the art performance on real-world MRI reconstruction tasks.

Table of Contents

Acknowledgments	v
Abstract	viii
List of Tables	xv
List of Figures	xvi
Chapter 1. Introduction	1
1.1 Background	1
1.2 Summary of Contributions	2
1.3 Results	4
Chapter 2. Compressed Sensing using Generative Models	11
2.1 Abstract	11
2.2 Introduction	12
2.3 Our Algorithm	16
2.4 Related Work	17
2.5 Theoretical Results	19
2.6 Models	24
2.6.1 MNIST with VAE	24
2.6.2 CelebA with DCGAN	25
2.7 Experiments and Results	26
2.7.1 Reconstruction from Gaussian measurements	26
2.7.1.1 MNIST	27
2.7.1.2 celebA	27
2.7.2 Super-resolution	28
2.7.2.1 MNIST	29
2.7.2.2 celebA	29

2.7.3	Understanding sources of error	31
2.7.3.1	Sensing images from the range of the generator	32
2.7.3.2	Quantifying representation error	32
2.8	Conclusion	34
Chapter 3.	A Robust Median-of-Means Algorithm	35
3.1	Abstract	35
3.2	Introduction	36
3.2.1	Contributions	38
3.2.2	Related work	40
3.3	Notation	41
3.4	Problem formulation	42
3.5	Our Algorithm	43
3.5.1	MOM objective	44
3.6	Theoretical results	46
3.6.1	Set-Restricted Eigenvalue Condition for heavy-tailed distributions	46
3.6.2	Main results	47
3.7	Experiments	52
3.8	Conclusion	55
Chapter 4.	Instance-Optimality via Posterior Sampling	58
4.1	Abstract	58
4.2	Introduction	58
4.2.1	Contributions.	60
4.2.2	Related Work	65
4.3	Background and Notation	67
4.4	Upper Bound	68
4.4.1	Two-Ball Case	68
4.4.2	Going beyond two balls	71
4.5	Lower Bound	74
4.6	Experiments	76
4.6.1	Datasets and Models	76

4.6.2	Langevin Dynamics	78
4.6.3	MAP and Modified-MAP	79
4.6.4	Experimental Results	80
4.7	Conclusion	83
Chapter 5.	Fairness Aspects in the Presence of Uncertain Sensitive Attributes	85
5.1	Abstract	85
5.2	Introduction	86
5.2.1	Related Work	95
5.3	Fairness definitions for image generation	96
5.3.1	Representation Demographic Parity	96
5.3.2	Limitations of traditional group fairness definitions . . .	97
5.3.3	Conditional Proportional Representation	99
5.4	Posterior Sampling	101
5.4.1	Representation Cross-Entropy	102
5.5	Experiments	104
5.5.1	Langevin Dynamics	104
5.5.2	MNIST dataset	104
5.5.3	FlickrFaces dataset	105
5.5.4	AFHQ Cats and Dogs dataset	106
5.6	Limitations	108
5.7	Conclusion	109
Chapter 6.	Robust Compressed Sensing MRI	112
6.1	Abstract	112
6.2	Introduction	112
6.2.1	Contributions	115
6.2.2	Related Work	117
6.3	System Model and Algorithm	119
6.3.1	Multi-coil Magnetic Resonance Imaging	119
6.3.2	Posterior Sampling	122
6.4	Theoretical Results	123

6.5	Experimental Results	127
6.5.1	In-Distribution Performance	130
6.5.2	Out-of-Distribution Performance	131
6.5.3	Uncertainty Estimation	134
6.5.4	Radiologist Study	135
6.6	Limitations	135
6.7	Conclusions	136
6.8	Summary of Individual Contributions	137
Chapter 7.	Future Work	138
Appendices		142
Appendix A.	Appendix for Chapter 2	143
A.1	Proofs	144
A.1.1	Proof of Lemma 2.5.2	144
A.1.2	Proof of Lemma 2.5.3	151
A.1.3	Proof of Lemma 2.5.4	154
A.1.4	Lipschitzness of Neural Networks	155
A.2	Additional Experiments	157
A.2.1	Noise tolerance	157
A.2.2	Other models	158
A.2.2.1	End to end training on MNIST	158
A.2.3	More results	159
Appendix B.	Appendix for Chapter 3	164
B.1	Proof of Lemma 3.6.1	165
B.2	Proof of Lemma 3.6.2	172
B.3	Proof of Lemma 3.6.3	173
B.4	Proof of Lemma 3.6.4	181
B.5	Proof of Theorem 3.6.5	188
B.6	Experimental Setup	189
B.6.1	MNIST dataset	189

B.6.2 CelebA-HQ dataset	190
B.6.3 Hyperparameter selection	191
B.7 Background	191
Appendix C. Appendix for Chapter 4	193
C.1 Upper Bound Proofs	194
C.1.1 Proof of Lemma 4.4.1	194
C.1.2 Proof of Lemma 4.4.2	195
C.1.3 Proof of Lemma C.1.1	198
C.1.4 Proof of Lemma 4.4.3	201
C.1.5 Proof of Theorem 4.4.4	205
C.2 Lower Bound Proofs	211
C.2.1 Proof of Lemma 4.5.2	211
C.2.2 Proof of Lemma 4.5.3	214
C.2.3 Proof of Fano variant Lemma 4.5.4	216
C.2.4 Proof of Theorem 4.5.1	226
C.3 Experimental Setup	229
C.3.1 Datasets and Architecture	229
C.3.2 Hyperparameter Selection	229
C.3.3 Computing Infrastructure	231
Appendix D. Appendix for Chapter 5	232
D.1 FFHQ Experiments	233
D.2 AFHQ Experiments	237
D.2.1 50% cat generator	237
D.2.2 x^* drawn from generator	238
D.2.3 Varying training bias	238
D.3 Proofs	241
D.4 Langevin Dynamics	249
D.4.1 StyleGAN2	249
D.5 Code	252

Appendix E. Appendix for Chapter 6	253
E.1 Appendix: Additional Metrics	254
E.1.1 MVUE vs. RSS	254
E.2 Appendix: Theory	259
E.2.1 Proof of Theorem 6.4.3	261
E.2.2 Proof of Theorem 6.4.4	267
E.3 Appendix: fastMRI Brain	268
E.3.1 Examples of Sampling Masks	268
E.3.2 More Exemplar Reconstructions	269
E.4 Appendix: fastMRI Knee	277
E.5 Appendix: Abdomen	284
E.6 Appendix: Stanford Knee	284
E.7 Appendix: Implementation	285
E.7.1 Score-Based Generative Model	285
E.7.2 E2E-VarNet Baseline	288
E.7.3 MoDL Baseline	289
E.8 Appendix: Radiologist Study	290
Appendix F. Bibliography	294

List of Tables

5.1	Confusion matrix for super-resolution on the MNIST dataset	105
5.2	Confusion matrix for super-resolution on the FFHQ dataset	106
E.1	Results of blind radiologist study ranking our algorithm vs competing baselines on the fastMRI dataset	291
E.2	Statistical significances of the blind radiologist study	292
E.3	p-values and confidence intervals for differences in ranking between our method and baselines.	293

List of Figures

2.1	Quantitative PSNR plots of CSGM vs. sparsity on celebA faces and MNIST digits	23
2.2	Results on MNIST. Reconstruction with 100 measurements (left) and Super-resolution (right)	28
2.3	Reconstruction results on celebA with 500 Gaussian measurements	30
2.4	Super-resolution results on celebA	30
2.5	Representation error of DCGAN on celebA faces	30
2.6	Quantitative econstruction error for images in the range of MNIST and celebA generators.	31
2.7	Qualitative reconstruction error for MNIST digits in the range of the generator.	32
3.1	Comparison of CSGM versus Median-of-Means tournaments on heavy-tailed data	51
3.2	Running time of CSGM vs. Median-of-Means tournaments . .	52
3.3	Qualitative comparison of CSGM vs. MoM tournaments on adversarially corrupted data	56
3.4	Comparison of MOM and CSGM on a variety of additional settings	57
4.1	Qualitative comparison of MAP vs Langevin dynamics over varying number of measurements	61
4.2	Reconstruction results on inpainting faces using the GLOW generative model	62
4.3	Illustrative example of posterior sampling upper bound	69
4.4	Quantitative comparison of MAP vs. Langevin on celebA-HQ faces	77
4.5	Comparison of MAP vs Deep Decoder vs Langevin on FlickrFaces	81
5.1	Qualitative comparison of fairness of PULSE and Langevin dynamics on the FFHQ dataset	87

5.2	Incompatibilities of different fairness definitions	90
5.3	Qualitative comparison of PULSE and Langevin dynamics on the AnimalFaces dataset	93
5.4	Quantitative comparison of PULSE vs. Posterior Sampling on imbalanced animal faces datasets.	111
6.1	Qualitative comparison of Posterior Sampling versus baselines on the fastMRI dataset.	116
6.2	PSNR of Posterior Sampling vs. baselines on fastMRI brains, knees, and abdomens	119
6.3	Qualitative comparison of Posterior sampling vs. baselines on abdomen scans	128
6.4	Uncertainty estimation using Posterior Sampling	132
A.1	Noise tolerance of Lasso and CSGM	157
A.2	Quantitative results for end-to-end training of the measurement matrix for CSGM on the MNIST dataset	158
A.3	Qualitative results for end-to-end training of CSGM on the MNIST dataset	159
A.4	Reconstruction on MNIST. In each image, top row is ground truth, middle row is Lasso, bottom row is our algorithm. . . .	160
A.5	Reconstructions of Lasso and CSGM on celebA faces	161
A.6	Reconstruction results of Lasso and CSGM on celebA faces (part 2)	162
A.7	Reconstruction results of Lasso and CSGM on celebA faces (part 3)	163
C.1	DAG relating x^*, A, z, y, \hat{x} . The conditional independencies we use are $x^* \perp\!\!\!\perp y z, A$ and $A \perp\!\!\!\perp y z$	213
D.1	Qualitative comparison of fairness of PULSE and Langevin dynamics on the FFHQ dataset (part 2)	233
D.2	Qualitative comparison of fairness of PULSE and Langevin dynamics on the FFHQ dataset (part 3)	234
D.3	Qualitative comparison of fairness of PULSE and Langevin dynamics on the FFHQ dataset (part 4)	235
D.4	Qualitative comparison of fairness of PULSE and Langevin dynamics on the FFHQ dataset (part 5)	236

D.5	Quantitative comparison of PULSE vs. Posterior Sampling on balanced animal faces datasets	237
D.6	Quantitative comparison of PULSE vs. Posterior Sampling on imbalanced animal faces datasets (part 2)	238
D.7	Quantitative comparison of PULSE vs. Posterior Sampling on imbalanced animal faces datasets (part 3)	239
D.8	Verification of fairness over varying bias	240
E.1	SSIM on fastMRI brains, knees, and abdomens	255
E.2	Masked SSIM on fastMRI brains, knees, and abdomens	256
E.3	Masked PSNR on fastMRI brains, knees, and abdomens	257
E.4	Examples of sampling patterns in MRI	269
E.5	Qualitative comparison of Lasso, ConvDecoder, MoDL, E2E-Varnet, and Langevin dynamics on brains at R=3	271
E.6	Qualitative comparison of Lasso, ConvDecoder, MoDL, E2E-Varnet, and Langevin dynamics on brains at R=6	272
E.7	Qualitative comparison of Lasso, ConvDecoder, MoDL, E2E-Varnet, and Langevin dynamics on brains at R=12	273
E.8	Qualitative comparison of Lasso, ConvDecoder, MoDL, E2E-Varnet, and Langevin dynamics on brains at R=3 and a change in the k-space sampling pattern	274
E.9	Qualitative comparison of Lasso, ConvDecoder, MoDL, E2E-Varnet, and Langevin dynamics on brains at R=4 and under a different T1-contrast.	275
E.10	Qualitative comparison of Lasso, ConvDecoder, MoDL, E2E-Varnet, and Langevin dynamics on brains at R=4 and under a different FLAIR contrast.	276
E.11	Qualitative comparison of Lasso, ConvDecoder, MoDL, E2E-Varnet, and Langevin dynamics on knees at R=4	278
E.12	Qualitative comparison of Lasso, ConvDecoder, MoDL, E2E-Varnet, and Langevin dynamics on knees at R=8	279
E.13	Uncertainty of Langevin dynamics on MRI of a meniscus tear	280
E.14	Comparison of algorithms on MRI of a meniscus tear	280
E.15	Comparison of algorithms on MRI of a meniscus tear (example 2)	281
E.16	Quantitative results on fat-suppressed knees	281
E.17	Qualitative results on fat-suppressed knee scans	282
E.18	Qualitative results on fat-suppressed knee scans at R=8	283

E.19 Qualitative results on abdomen scans	284
E.20 Quantitative results on 3D knee scans	285
E.21 Qualitative results on 3D knee scans	286

Chapter 1

Introduction

1.1 Background

This dissertation proposes a novel framework for tackling inverse problems using the power of *deep generative models*.

Inverse problems seek to reconstruct an unknown image, signal, or multi-dimensional volume from observations of the data. The observations are obtained from the unknown data by a lossy *forward process*. For example, consider the problem of image super-resolution: you observe a low-resolution blur of a human face, and you would like to sharpen it to a higher resolution. Such problems are *ill-posed*, as finding a unique face that fits the low-resolution observations is difficult or impossible without imposing some prior knowledge on the original image.

Many imaging tasks fall under the umbrella of ill-posed inverse problems, including image super-resolution, deblurring, deconvolution, inpainting missing pixels, compressed sensing, and many more. Tomographic applications such as magnetic resonance imaging (MRI), X-ray computed tomography, and radar imaging are other examples of inverse problems.

Generative models, on the other hand, seek to learn structure present

in high-dimensional data. By viewing a dataset of images as empirical samples from a high-dimensional probability distribution, generative models are trained to produce samples from this target distribution. There are many popular frameworks for training generative models, the most popular of which are adversarial training [98], variational inference [158], likelihood estimation via normalizing flows [77], and score estimation [253]. All of these models can be trained on a dataset of representative images, and can produce *new* high-quality images. This leads to the fundamental question answered in this dissertation.

Question 1.1.1. *There is strong empirical evidence that generative models are able to approximate high-dimensional probability distributions. How can we leverage these models to solve inverse problems? Additionally, can we design algorithms with provable algorithmic and statistically complexities?*

1.2 Summary of Contributions

Our first paper, Compressed Sensing using Generative Models [41] (CSGM), showed that Maximum a posteriori (MAP) estimation using generative models can solve compressed sensing and other inverse problems: as generative models can *learn* the structure present in signals, they have a competitive edge over hand-crafted priors like sparsity. CSGM was an exciting result that inspired several extensions and follow-ups, but future work showed that it had several biases. The most striking was its majority bias: when a modification of CSGM was run on measurements of Barack Obama’s face,

the reconstruction was a white face [1]. CSGM was also fragile to changes in distribution of the ground-truth signal, for e.g., a generative model trained on human faces would fail on animated faces. This can be dangerous in applications like MRI, where rare pathologies must be accurately reconstructed.

In recent work, we showed that *posterior sampling* using full-dimensional generative models resolves many of these issues. Our theoretical analysis [133, 134] showed that posterior sampling achieves almost-optimal sample complexity and is robust to distribution mismatch between ground-truth and the generative model. Interestingly, posterior sampling is optimal with respect to *arbitrary metrics and measurement processes*: this is important in cases like MRI, where the probability of detecting a pathology is more important than PSNR values. Our experiments on the NYU fastMRI dataset showed that posterior sampling is more robust than deep learning heuristics to changes in the MRI scans, which was corroborated by a preliminary blind study with radiologists. With respect to the “White Obama” phenomenon, we showed that MAP estimation exacerbates the majority bias in the dataset: if the dataset contains an 80% majority class, then MAP can *always* produce reconstructions from this class, while posterior sampling preserves diversity in the dataset [135]. We gave the first codification of fairness for inverse problems and showed that posterior sampling provably satisfies certain definitions. Additionally, since attributes like race and gender are ambiguous to define, we showed that posterior sampling can be fair even if it is *oblivious* to the sensitive attribute.

1.3 Results

Compressed Sensing using Low-Dimensional Generative Models Our first paper, Compressed Sensing using Generative Models (CSGM) [41], showed that generative models, such as GANs [98] and VAEs [158], are excellent priors for compressed sensing. Specifically, in sparsity-based compressed sensing, one has to handcraft a good basis (or learn a dictionary with shallow features) such that a sparse combination of elements in this basis gives good approximations. Generative models, on the other hand, can *learn* low-dimensional structure implicitly present in large datasets, which gives them a natural competitive edge over other priors such as sparsity.

Our main theorem shows that if we have access to a good generative model, then the MAP estimate using this generative prior can recover signals using very few measurements. Specifically, if $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ is a ReLU neural network with d layers, then the number of measurements required to recover a signal living in \mathbb{R}^n is on the order of $kd\log n$. This bound is very good on generative models trained on established benchmark datasets – on the celebA [184] dataset, we trained a DCGAN [225] that had $n = 12288$, $k = 100$, $d = 5$, in which case $kd\log n \ll n$. We further generalized our results and showed that if G is an L -Lipschitz function, then the number of measurements scale as $k\log L$. Our proofs introduced the Set-Restricted Eigenvalue Condition (S-REC), which has proven to be useful in the analysis of optimization algorithms [244] and has been generalized to different inverse problems [126].

Empirically, we showed that our algorithm recovers signals using $5 - 10 \times$ fewer measurements than Lasso. An additional benefit of this framework is that it is highly modular – using a *single* trained generative model, one can solve different inverse problems such as super-resolution & inpainting. This also implies that using an improved generative model gives better results, as shown in a recent paper from our group [70].

This framework has proven to be quite general and has inspired many research directions. Our theorems were shown to be information-theoretically optimal [143, 182], and our framework can be used for channel estimation in MIMO networks [29, 19], seismic inversion [205], phase retrieval [105], and one-bit compressed sensing [223, 179], to name a few. Hand & Voroninski [106] showed that although the original CSGM algorithm considers a non-convex optimization problem, it converges in polynomial time if the weights of the generative model are random Gaussian, and these results were strengthened by [73]. In a different line of work, our group showed that matrix inversion is also a viable algorithm for recovery [171].

Robust Estimators via Median-of-Means Tournaments The above results assume that the measurement process follows a sub-Gaussian distribution. However, as our algorithm implements MAP via empirical risk minimization (ERM), it is fragile when the measurements contain adversarial corruptions and gives sub-optimal guarantees in the presence of heavy-tailed noise. To address this, we proposed a Median-of-Means tournament [136, 188] that

is robust to a constant-fraction of adversarial corruptions in the measurements and measurement matrix. Interestingly, the number of measurements matches [41] upto constants and we achieve sub-Gaussian guarantees even in the presence of heavy-tailed noise. Our estimator was the first feasible robust estimator for generative models, as the existing baselines [295] do not have feasible implementations. Additionally, we showed that the S-REC property is satisfied for matrices whose rows have constant Kurtosis, which weakens previous assumptions [41].

Full-Dimensional Generative Models and Posterior Sampling We noticed several interesting phenomena and limitations in the original CSGM algorithm. When reconstructing faces from the celebA dataset using a DC-GAN, we obtained extremely good reconstructions using only 500 measurements of a 12,288 dimensional image. Unfortunately, reconstruction quality plateaued and did not improve for > 500 measurements. Lasso did not exhibit this behaviour, and outperforms CSGM after 5,000 measurements.

This phenomenon occurs because our generative model was trained at $k = 100$. This introduces an inductive bias in our algorithm, as the data is assumed to lie strictly on a k -dimensional manifold. Lasso, on the other hand, can vary the dimensionality of the reconstruction as a free parameter, and can allow for richer reconstructions when the number of measurements is large. Recently, Asim et. al. [21] showed that a *modified* MAP estimate using an *invertible* generative model can mitigate the bias present in low-

dimensional generative models. However, [21] reported several regularizers in their experiments, as the true MAP estimate did not work well in practice. Additionally, as these models are bijective functions from \mathbb{R}^n to \mathbb{R}^n , the low-dimensional manifold assumption is no longer true and the theoretical results in literature are vacuous.

In recent work [134], we showed that *posterior sampling* is provably optimal for *any* distribution of the ground-truth, including those represented by full-dimensional models. We introduced the notion of *approximate covering numbers*, which characterizes the “compressibility” of full-dimensional models, and showed that the number of measurements required by posterior sampling is upper bounded by the logarithm of the approximate covering number. This requires new proof techniques, since the distributions we consider do not lie on low-dimensional manifolds. We also showed an almost-matching lower bound that applies to *any* distribution of the ground-truth (lower bounds in compressed sensing literature only apply to worst-case distributions, for e.g., uniform distributions over error correcting codes). By combining these results, we can infer that posterior sampling is a near-optimal algorithm.

Finally, using tools from Optimal Transport, we showed that posterior sampling is robust to distribution mismatch, i.e., our guarantees hold as long as the generative model’s distribution is close to the ground truth distribution in Wasserstein distance. We experimentally validated our results by implementing posterior sampling via annealed Langevin dynamics using the NCSNv2 [254] and GLOW [159] generative models, both of which are full-

dimensional models, and our results are competitive with *modified*-MAP and outperform the true MAP estimate.

Algorithmic Fairness In the aftermath of the “White Obama” [1] phenomenon, the prevailing public opinion was that this occurs purely due to the biased *dataset* used to train the generative model. Additionally, while this example is clearly “unfair”, there was no prior literature on what would be “fair” in the context of image generation and inverse problems.

We [135] showed that the above phenomenon is not purely a dataset problem, but there is a significant *algorithmic problem* as well: we ran a super-resolution experiment using a generative model trained on a dataset containing 80% cats and 20% dogs, and we found that the MAP estimate mistakenly reconstructs 1% of cats as dogs, while 98% of dogs are mistakenly reconstructed as cats. This shows that MAP estimation using a biased generative model can exacerbate the bias in the dataset and makes an overwhelming number of errors on the minority. On the other hand, posterior sampling using the same generative model was able to preserve diversity and makes an equal number of errors on both classes.

Defining fairness for inverse problems is challenging for two fundamental reasons. First, traditional algorithmic fairness requires some form of conditional independence between race and the outcome (for e.g., odds of receiving a bank loan). However, in the “White Obama” example, we want the sensitive attribute to be *explicitly preserved* in the output. Second, sensitive attributes

like race and gender are ambiguous to define, and defining partitions of protected groups based on these attributes is near-impossible. To address these challenges, we introduced and codified several definitions of fairness for inverse problems, and proved incompatibilities among these definitions. We showed that posterior sampling can be *obliviously fair*, meaning it can satisfy fairness guarantees without explicit knowledge of which groups need to be protected. If information of the protected groups is explicitly available, we proved that a reweighting of the generative model’s prior distribution can satisfy stronger fairness guarantees with respect to these specific groups. We experimentally validated our results using state-of-the-art StyleGAN2 [145] and NCSNv2 [254] generative models, and showed that posterior sampling (implemented via annealed Langevin dynamics) satisfies our theoretical guarantees in practice.

Applications to MRI Despite the surge of interest in the CSGM framework, practitioners were unconvinced by its empirical value. As generative priors produce low-resolution images, introduce artefacts, and are fragile out-of-distribution, CSGM could potentially be diagnostically dangerous in applications such as MRI. In our most recent work [133], we showed the first successful empirical application of the CSGM framework – the reconstruction quality of posterior sampling using score-based generative models [254] quantitatively and qualitatively matches state-of-the-art deep learning methods on the NYU fastMRI [297] and Stanford MRI [2] datasets. When applied to settings with changes in anatomy or scan parameters, we showed that posterior

sampling is more robust than baselines. We also performed a *preliminary* blind expert study, with radiologists comparing our reconstructions and baseline reconstructions with a “gold standard” fully-sampled scan. The experts scored our algorithm as better or equal to competing baselines, which suggests the possibility of clinical adoption of our techniques.

Chapter 2

Compressed Sensing using Generative Models

2.1 Abstract

The goal of compressed sensing is to estimate a vector from an underdetermined system of noisy linear measurements, by making use of prior knowledge on the structure of vectors in the relevant domain. For almost all results in this literature, the structure is represented by sparsity in a well-chosen basis. We show how to achieve guarantees similar to standard compressed sensing but without employing sparsity at all. Instead, we suppose that vectors lie near the range of a generative model $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$. Our main theorem is that, if G is L -Lipschitz, then roughly $O(k \log L)$ random Gaussian measurements suffice for an ℓ_2/ℓ_2 recovery guarantee. We demonstrate our results using generative models from published variational autoencoder and generative adversarial networks. Our method can use 5-10x fewer measurements than Lasso for the same accuracy.

These results were published at ICML 2017 [41].

2.2 Introduction

Compressive or compressed sensing is the problem of reconstructing an unknown vector $x^* \in \mathbb{R}^n$ after observing $m < n$ linear measurements of its entries, possibly with added noise:

$$y = Ax^* + \eta,$$

where $A \in \mathbb{R}^{m \times n}$ is called the measurement matrix and $\eta \in \mathbb{R}^m$ is noise. Even without noise, this is an underdetermined system of linear equations, so recovery is impossible unless we make an assumption on the structure of the unknown vector x^* . We need to assume that the unknown vector is “natural,” or “simple,” in some application-dependent way.

The most common structural assumption is that the vector x^* is k -sparse in some known basis (or approximately k -sparse). Finding the sparsest solution to an underdetermined system of linear equations is NP-hard, but still convex optimization can provably recover the true sparse vector x^* if the matrix A satisfies conditions such as the Restricted Isometry Property (RIP) or the related Restricted Eigenvalue Condition (REC) [263, 48, 82, 39]. The problem is also called high-dimensional sparse linear regression and there is vast literature on establishing conditions for different recovery algorithms, different assumptions on the design of A and generalizations of RIP and REC for other structures, see *e.g.* [39, 208, 9, 186, 26].

This significant interest is justified since a large number of applications can be expressed as recovering an unknown vector from noisy linear

measurements. For example, many tomography problems can be expressed in this framework: x^* is the unknown true tomographic image and the linear measurements are obtained by x-ray or other physical sensing system that produces sums or more general linear projections of the unknown pixels. Compressed sensing has been studied extensively for medical applications including computed tomography (CT) [55], rapid MRI [189] and neuronal spike train recovery [116]. Another impressive application is the “single pixel camera” [83], where digital micro-mirrors provide linear combinations to a single pixel sensor that then uses compressed sensing reconstruction algorithms to reconstruct an image. These results have been extended by combining sparsity with additional structural assumptions [32, 117], and by generalizations such as translating sparse vectors into low-rank matrices [208, 26, 90]. These results can improve performance when the structural assumptions fit the sensed signals. Other works perform “dictionary learning,” seeking overcomplete bases where the data is more sparse (see [56] and references therein).

In this paper instead of relying on sparsity, we use structure from a *generative model*. Recently, several neural network based generative models such as variational auto-encoders (VAEs) [158] and generative adversarial networks (GANs) [98] have found success at modeling data distributions. In these models, the generative part learns a mapping from a low dimensional representation space $z \in \mathbb{R}^k$ to the high dimensional sample space $G(z) \in \mathbb{R}^n$. While training, this mapping is encouraged to produce vectors that resemble the vectors in the training dataset. We can therefore use any pre-trained generator to

approximately capture the notion of a vector being “natural” in our domain: the generator defines a probability distribution over vectors in sample space and tries to assign higher probability to more likely vectors, for the dataset it has been trained on. We expect that vectors “natural” to our domain will be close to some point in the support of this distribution, *i.e.*, in the range of G .

Our Contributions: We present an algorithm that uses generative models for compressed sensing. Our algorithm simply uses gradient descent to optimize the representation $z \in \mathbb{R}^k$ such that the corresponding image $G(z)$ has small measurement error $\|AG(z) - y\|_2^2$. While this is a nonconvex objective to optimize, we empirically find that gradient descent works well, and the results can significantly outperform Lasso with relatively few measurements.

We obtain theoretical results showing that, as long as gradient descent finds a good approximate solution to our objective, our output $G(z)$ will be almost as close to the true x^* as the closest possible point in the range of G .

The proof is based on a generalization of the Restricted Eigenvalue Condition (*REC*) that we call the Set-Restricted Eigenvalue Condition (*S-REC*). Our main theorem is that if a measurement matrix satisfies the *S-REC* for the range of a given generator G , then the measurement error minimization optimum is close to the true x^* . Furthermore, we show that random Gaussian measurement matrices satisfy the *S-REC* condition with high probability for large classes of generators. Specifically, for d -layer neural networks such as VAEs and GANs, we show that $O(kd \log n)$ Gaussian measurements suffice to guarantee good reconstruction with high probability. One result, for

ReLU-based networks, is the following:

Theorem 2.2.1. *Let $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be a generative model from a d -layer neural network using ReLU activations. Let $A \in \mathbb{R}^{m \times n}$ be a random Gaussian matrix for $m = O(kd \log n)$, scaled so $A_{i,j} \sim N(0, 1/m)$. For any $x^* \in \mathbb{R}^n$ and any observation $y = Ax^* + \eta$, let \hat{z} minimize $\|y - AG(z)\|_2$ to within additive ε of the optimum. Then with $1 - e^{-\Omega(m)}$ probability,*

$$\|G(\hat{z}) - x^*\|_2 \leq 6 \min_{z^* \in \mathbb{R}^k} \|G(z^*) - x^*\|_2 + 3\|\eta\|_2 + 2\varepsilon.$$

Let us examine the terms in our error bound in more detail. The first two are the minimum possible error of any vector in the range of the generator and the norm of the noise; these are necessary for such a technique, and have direct analogs in standard compressed sensing guarantees. The third term ε comes from gradient descent not necessarily converging to the global optimum; empirically, ε does seem to converge to zero, and one can check post-observation that this is small by computing the upper bound $\|y - AG(\hat{z})\|_2$.

While the above is restricted to ReLU-based neural networks, we also show similar results for arbitrary L -Lipschitz generative models, for $m \approx O(k \log L)$. Typical neural networks have poly(n)-bounded weights in each layer, so $L \leq n^{O(d)}$, giving for all activation functions the same $O(kd \log n)$ sample complexity as for ReLU networks.

Theorem 2.2.2. *Let $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be an L -Lipschitz function. Let $A \in \mathbb{R}^{m \times n}$ be a random Gaussian matrix for $m = O(k \log \frac{Lr}{\delta})$, scaled so $A_{i,j} \sim N(0, 1/m)$.*

For any $x^* \in \mathbb{R}^n$ and any observation $y = Ax^* + \eta$, let \hat{z} minimize $\|y - AG(z)\|_2$ to within additive ε of the optimum over vectors with $\|\hat{z}\|_2 \leq r$. Then with $1 - e^{-\Omega(m)}$ probability,

$$\|G(\hat{z}) - x^*\|_2 \leq 6 \min_{\substack{z^* \in \mathbb{R}^k \\ \|z^*\|_2 \leq r}} \|G(z^*) - x^*\|_2 + 3\|\eta\|_2 + 2\varepsilon + 2\delta.$$

The downside is two minor technical conditions: we only optimize over representations z with $\|z\|$ bounded by r , and our error gains an additive δ term. Since the dependence on these parameters is $\log(rL/\delta)$, and L is something like $n^{O(d)}$, we may set $r = n^{O(d)}$ and $\delta = 1/n^{O(d)}$ while only losing constant factors, making these conditions very mild. In fact, generative models normally have the coordinates of z be independent uniform or Gaussian, so $\|z\| \approx \sqrt{k} \ll n^d$, and a constant signal-to-noise ratio would have $\|\eta\|_2 \approx \|x^*\| \approx \sqrt{n} \gg 1/n^d$.

We remark that, while these theorems are stated in terms of Gaussian matrices, the proofs only involve the distributional Johnson-Lindenstrauss property of such matrices. Hence the same results hold for matrices with subgaussian entries or fast-JL matrices [12].

2.3 Our Algorithm

All norms are 2-norms unless specified otherwise.

Let $x^* \in \mathbb{R}^n$ be the vector we wish to sense. Let $A \in \mathbb{R}^{m \times n}$ be the measurement matrix and $\eta \in \mathbb{R}^m$ be the noise vector. We observe the mea-

surements $y = Ax^* + \eta$. Given y and A , our task is to find a reconstruction \hat{x} close to x^* .

A generative model is given by a deterministic function $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$, and a distribution P_Z over $z \in \mathbb{R}^k$. To generate a sample from the generator, we can draw $z \sim P_Z$ and the sample then is $G(z)$. Typically, we have $k \ll n$, *i.e.* the generative model maps from a low dimensional representation space to a high dimensional sample space.

Our approach is to find a vector in representation space such that the corresponding vector in the sample space matches the observed measurements. We thus define the objective to be

$$\text{loss}(z) = \|AG(z) - y\|^2 \quad (2.1)$$

By using any optimization procedure, we can minimize $\text{loss}(z)$ with respect to z . In particular, if the generative model G is differentiable, we can evaluate the gradients of the loss with respect to z using backpropagation and use standard gradient based optimizers. If the optimization procedure terminates at \hat{z} , our reconstruction for x^* is $G(\hat{z})$. We define the measurement error to be $\|AG(\hat{z}) - y\|^2$ and the reconstruction error to be $\|G(\hat{z}) - x^*\|^2$.

2.4 Related Work

Several recent lines of work explore generative models for reconstruction. The first line of work attempts to project an image on to the representation space of the generator. These works assume full knowledge of the

image, and are special cases of the linear measurements framework where the measurement matrix A is identity. Excellent reconstruction results with SGD in the representation space to find an image in the generator range have been reported by [176] with stochastic clipping and [68] with logistic measurement loss. A different approach is introduced in [84] and [78]. In their method, a recognition network that maps from the sample space vector x to the representation space vector z is learned jointly with the generator in an adversarial setting.

A second line of work explores reconstruction with structured partial observations. The inpainting problem consists of predicting the values of missing pixels given a part of the image. This is a special case of linear measurements where each measurement corresponds to an observed pixel. The use of Generative models for this task has been studied in [294], where the objective is taken to be a combination of L_1 error in measurements and a perceptual loss term given by the discriminator. Super-resolution is a related task that attempts to increase the resolution of an image. We can view this problem as observing local spatial averages of the unknown higher resolution image and hence cast this as another special case of linear measurements. For prior work on super-resolution see *e.g.* [293, 81, 155] and references therein.

We also take note of the related work of [92] that connects model-based compressed sensing with the invertibility of Convolutional Neural Networks.

A related result appears in [33], which studies the measurement complexity of an RIP condition for smooth manifolds. This is analogous to our

S-REC for the range of G , but the range of G is neither smooth (because of ReLUs) nor a manifold (because of self-intersection). Their recovery result was extended in [115] to unions of two manifolds.

2.5 Theoretical Results

We begin with a brief review of the Restricted Eigenvalue Condition (REC) in standard compressed sensing. The REC is a sufficient condition on A for robust recovery to be possible. The REC essentially requires that all “approximately sparse” vectors are far from the nullspace of the matrix A . More specifically, A satisfies REC for a constant $\gamma > 0$ if for all approximately sparse vectors x ,

$$\|Ax\| \geq \gamma \|x\|. \quad (2.2)$$

It can be shown that this condition is sufficient for recovery of sparse vectors using Lasso. If one examines the structure of Lasso recovery proofs, a key property that is used is that the difference of any two sparse vectors is also approximately sparse (for sparsity up to $2k$). This is a coincidence that is particular to sparsity. By contrast, the difference of two vectors “natural” to our domain may not itself be natural. The condition we need is that the difference of any two natural vectors is far from the nullspace of A .

We propose a generalized version of the REC for a set $S \subseteq \mathbb{R}^n$ of vectors, the Set-Restricted Eigenvalue Condition (S-REC):

Definition 2.5.1. *Let $S \subseteq \mathbb{R}^n$. For some parameters $\gamma > 0$, $\delta \geq 0$, a matrix*

$A \in \mathbb{R}^{m \times n}$ is said to satisfy the S-REC(S, γ, δ) if $\forall x_1, x_2 \in S$,

$$\|A(x_1 - x_2)\| \geq \gamma \|x_1 - x_2\| - \delta.$$

There are two main differences between the S-REC and the standard REC in compressed sensing. First, the condition applies to differences of vectors in an *arbitrary* set S of “natural” vectors, rather than just the set of approximately k -sparse vectors in some basis. This will let us apply the definition to S being the range of a generative model.

Second, we allow an additive slack term δ . This is necessary for us to achieve the S-REC when S is the output of general Lipschitz functions. Without it, the S-REC depends on the behavior of S at arbitrarily small scales. Since there are arbitrarily many such local regions, one cannot guarantee the existence of an A that works for all these local regions. Fortunately, as we shall see, poor behavior at a small scale δ will only increase our error by $\mathcal{O}(\delta)$.

The S-REC definition requires that for any two vectors in S , if they are significantly different (so the right hand side is large), then the corresponding measurements should also be significantly different (left hand side). Hence we can hope to approximate the unknown vector from the measurements, if the measurement matrix satisfies the S-REC.

But how can we find such a matrix? To answer this, we present two lemmas showing that random Gaussian matrices of relatively few measurements m satisfy the S-REC for the outputs of large and practically useful classes of generative models $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$.

In the first lemma, we assume that the generative model $G(\cdot)$ is L -Lipschitz, *i.e.*, $\forall z_1, z_2 \in \mathbb{R}^k$, we have

$$\|G(z_1) - G(z_2)\| \leq L\|z_1 - z_2\|.$$

Note that state of the art neural network architectures with linear layers, (transposed) convolutions, max-pooling, residual connections, and all popular non-linearities satisfy this assumption. In Lemma A.1.6 in the Appendix we give a simple bound on L in terms of parameters of the network; for typical networks this is $n^{O(d)}$. We also require the input z to the generator to have bounded norm. Since generative models such as VAEs and GANs typically assume their input z is drawn with independent uniform or Gaussian inputs, this only prunes an exponentially unlikely fraction of the possible outputs.

Lemma 2.5.2. *Let $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be L -Lipschitz. Let*

$$B^k(r) = \{z \mid z \in \mathbb{R}^k, \|z\| \leq r\}$$

be an L_2 -norm ball in \mathbb{R}^k . For $\alpha < 1$, if

$$m = \Omega\left(\frac{k}{\alpha^2} \log \frac{Lr}{\delta}\right),$$

then a random matrix $A \in \mathbb{R}^{m \times n}$ with IID entries such that $A_{ij} \sim \mathcal{N}(0, \frac{1}{m})$ satisfies the S -REC($G(B^k(r))$, $1 - \alpha$, δ) with $1 - e^{-\Omega(\alpha^2 m)}$ probability.

All proofs, including this one, are deferred to Appendix A.

Note that even though we proved the lemma for an L_2 ball, the same technique works for any compact set.

For our second lemma, we assume that the generative model is a neural network with such that each layer is a composition of a linear transformation followed by a pointwise non-linearity. Many common generative models have such architectures. We also assume that all non-linearities are piecewise linear with at most two pieces. The popular ReLU or LeakyReLU non-linearities satisfy this assumption. We do not make any other assumption, and in particular, the magnitude of the weights in the network do not affect our guarantee.

Lemma 2.5.3. *Let $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be a d -layer neural network, where each layer is a linear transformation followed by a pointwise non-linearity. Suppose there are at most c nodes per layer, and the non-linearities are piecewise linear with at most two pieces, and let*

$$m = \Omega\left(\frac{1}{\alpha^2} kd \log c\right)$$

for some $\alpha < 1$. Then a random matrix $A \in \mathbb{R}^{m \times n}$ with IID entries $A_{ij} \sim \mathcal{N}(0, \frac{1}{m})$ satisfies the S -REC($G(\mathbb{R}^k), 1 - \alpha, 0$) with $1 - e^{-\Omega(\alpha^2 m)}$ probability.

To show Theorems 2.2.1 and 2.2.2, we just need to show that the S-REC implies good recovery. In order to make our error guarantee relative to ℓ_2 error in the image space \mathbb{R}^n , rather than in the measurement space \mathbb{R}^m , we also need that A preserves norms with high probability [64]. Fortunately, Gaussian matrices (or other distributional JL matrices) satisfy this property.

Lemma 2.5.4. *Let $A \in \mathbb{R}^{m \times n}$ be drawn from a distribution that (1) satisfies the S -REC(S, γ, δ) with probability $1 - p$ and (2) has for every fixed $x \in \mathbb{R}^n$, $\|Ax\| \leq 2\|x\|$ with probability $1 - p$.*

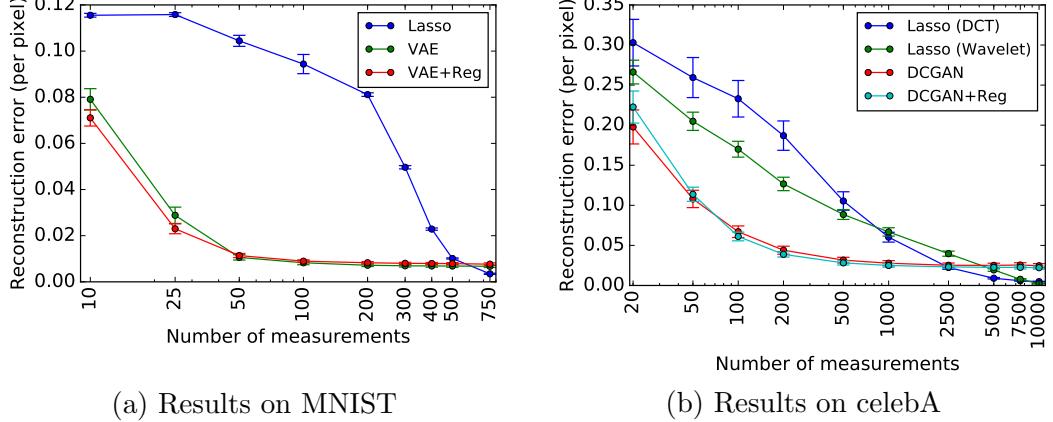


Figure 2.1: We compare the performance of our algorithm with baselines. We show a plot of per pixel reconstruction error as we vary the number of measurements. The vertical bars indicate 95% confidence intervals.

For any $x^* \in \mathbb{R}^n$ and noise η , let $y = Ax^* + \eta$. Let \hat{x} approximately minimize $\|y - Ax\|$ over $x \in S$, i.e.,

$$\|y - A\hat{x}\| \leq \min_{x \in S} \|y - Ax\| + \varepsilon.$$

Then,

$$\|\hat{x} - x^*\| \leq \left(\frac{4}{\gamma} + 1\right) \min_{x \in S} \|x^* - x\| + \frac{1}{\gamma} (2\|\eta\| + \varepsilon + \delta)$$

with probability $1 - 2p$.

Combining Lemma 2.5.2, Lemma 2.5.3, and Lemma 2.5.4 gives Theorems 2.2.1 and 2.2.2. In our setting, S is the range of the generator, and \hat{x} in the theorem above is the reconstruction $G(\hat{z})$ returned by our algorithm.

2.6 Models

In this section we describe the generative models used in our experiments. We used two image datasets and two different generative model types (a VAE and a GAN). This provides some evidence that our approach can work with many types of models and datasets.

In our experiments, we found that it was helpful to add a regularization term $L(z)$ to the objective to encourage the optimization to explore more in the regions that are preferred by the respective generative models (see comparison to unregularized versions in Fig. 2.1). Thus the objective function we use for minimization is

$$\|AG(z) - y\|^2 + L(z).$$

Both VAE and GAN typically imposes an isotropic Gaussian prior on z . Thus $\|z\|^2$ is proportional to the negative log-likelihood under this prior. Accordingly, we use the following regularizer:

$$L(z) = \lambda \|z\|^2, \tag{2.3}$$

where λ measures the relative importance of the prior as compared to the measurement error.

2.6.1 MNIST with VAE

The MNIST dataset consists of about 60,000 images of handwritten digits, where each image is of size 28×28 [168]. Each pixel value is either 0 (background) or 1 (foreground). No pre-processing was performed. We trained

VAE on this dataset. The input to the VAE is a vectorized binary image of input dimension 784. We set the size of the representation space $k = 20$. The recognition network is a fully connected $784 - 500 - 500 - 20$ network. The generator is also fully connected with the architecture $20 - 500 - 500 - 784$. We train the VAE using the Adam optimizer [157] with a mini-batch size 100 and a learning rate of 0.001.

We found that using $\lambda = 0.1$ in Eqn. (2.3) gave the best performance, and we use this value in our experiments.

The digit images are reasonably sparse in the pixel space. Thus, as a baseline, we use the pixel values directly for sparse recovery using Lasso. We set shrinkage parameter to be 0.1 for all the experiments.

2.6.2 CelebA with DCGAN

CelebA is a dataset of more than 200, 000 face images of celebrities [183]. The input images were cropped to a 64×64 RGB image, giving $64 \times 64 \times 3 = 12288$ inputs per image. Each pixel value was scaled so that all values are between $[-1, 1]$. We trained a DCGAN¹ [225, 156] on this dataset. We set the input dimension $k = 100$ and use a standard normal distribution. The architecture follows that of [225]. The model was trained by one update to the discriminator and two updates to the generator per cycle. Each update used the Adam optimizer [157] with minibatch size 64, learning rate 0.0002

¹Code reused from <https://github.com/carpedm20/DCGAN-tensorflow>

and $\beta_1 = 0.5$.

We found that using $\lambda = 0.001$ in Eqn. (2.3) gave the best results and thus, we use this value in our experiments.

For baselines, we perform sparse recovery using Lasso on the images in two domains: (a) 2D Discrete Cosine Transform (2D-DCT) and (b) 2D Daubechies-1 Wavelet Transform (2D-DB1). While we provide Gaussian measurements of the original pixel values, the L_1 penalty is on either the DCT coefficients or the DB1 coefficients of each color channel of an image. For all experiments, we set the shrinkage parameter to be 0.1 and 0.00001 respectively for 2D-DCT, and 2D-DB1.

2.7 Experiments and Results

2.7.1 Reconstruction from Gaussian measurements

We take A to be a random matrix with IID Gaussian entries with zero mean and standard deviation of $1/m$. Each entry of noise vector η is also an IID Gaussian random variable. We compare performance of different sensing algorithms qualitatively and quantitatively. For quantitative comparison, we use the reconstruction error $= \|\hat{x} - x^*\|^2$, where \hat{x} is an estimate of x^* returned by the algorithm. In all cases, we report the results on a held out test set, unseen by the generative model at training time.

2.7.1.1 MNIST

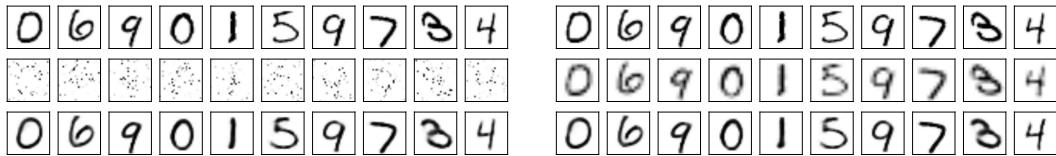
The standard deviation of the noise vector is set such that $\sqrt{\mathbb{E}[\|\eta\|^2]} = 0.1$. We use Adam optimizer [157], with a learning rate of 0.01. We do 10 random restarts with 1000 steps per restart and pick the reconstruction with best measurement error.

In Fig. 2.1a, we show the reconstruction error as we change the number of measurements both for Lasso and our algorithm. We observe that our algorithm is able to get low errors with far fewer measurements. For example, our algorithm’s performance with 25 measurements matches Lasso’s performance with 400 measurements. Fig. 2.2a shows sample reconstructions by Lasso and our algorithm.

However, our algorithm is limited since its output is constrained to be in the range of the generator. After 100 measurements, our algorithm’s performance saturates, and additional measurements give no additional performance. Since Lasso has no such limitation, it eventually surpasses our algorithm, but this takes more than 500 measurements of the 784-dimensional vector. We expect that a more powerful generative model with representation dimension $k > 20$ can make better use of additional measurements.

2.7.1.2 celebA

The standard deviation of entries in the noise vector is set such that $\sqrt{\mathbb{E}[\|\eta\|^2]} = 0.01$. We optimize use Adam optimizer [157], with a learning rate of 0.1. We do 2 random restarts with 500 update steps per restart and pick



(a) We show original images (top row) and reconstructions by Lasso (middle row) and our algorithm (bottom row).

(b) We show original images (top row), low resolution version of original images (middle row) and reconstructions (last row).

Figure 2.2: Results on MNIST. Reconstruction with 100 measurements (left) and Super-resolution (right)

the reconstruction with best measurement error.

In Fig. 2.1b, we show the reconstruction error as we change the number of measurements both for Lasso and our algorithm. In Fig. 2.3 we show sample reconstructions by Lasso and our algorithm. We observe that our algorithm is able to produce reasonable reconstructions with as few as 500 measurements, while the output of the baseline algorithms is quite blurry. Similar to the results on MNIST, if we continue to give more measurements, our algorithm saturates, and for more than 5000 measurements, Lasso gets a better reconstruction. We again expect that a more powerful generative model with $k > 100$ would perform better in the high-measurement regime.

2.7.2 Super-resolution

Super-resolution is the task of constructing a high resolution image from a low resolution version of the same image. This problem can be thought of as special case of our general framework of linear measurements, where the

measurements correspond to local spatial averages of the pixel values. Thus, we try to use our recovery algorithm to perform this task with measurement matrix A tailored to give only the relevant observations. We note that this measurement matrix may not satisfy the S-REC condition (with good constants γ and δ), and consequently, our theorems may not be applicable.

2.7.2.1 MNIST

We construct a low resolution image by spatial 2×2 pooling with a stride of 2 to produce a 14×14 image. These measurements are used to reconstruct the original 28×28 image. Fig. 2.2b shows reconstructions produced by our algorithm on images from a held out test set. We observe sharp reconstructions which closely match the fine structure in the ground truth.

2.7.2.2 celebA

We construct a low resolution image by spatial 4×4 pooling with a stride of 4 to produce a 16×16 image. These measurements are used to reconstruct the original 64×64 image. In Fig. 2.4 we show results on images from a held out test set. We see that our algorithm is able to fill in the details to match the original image.

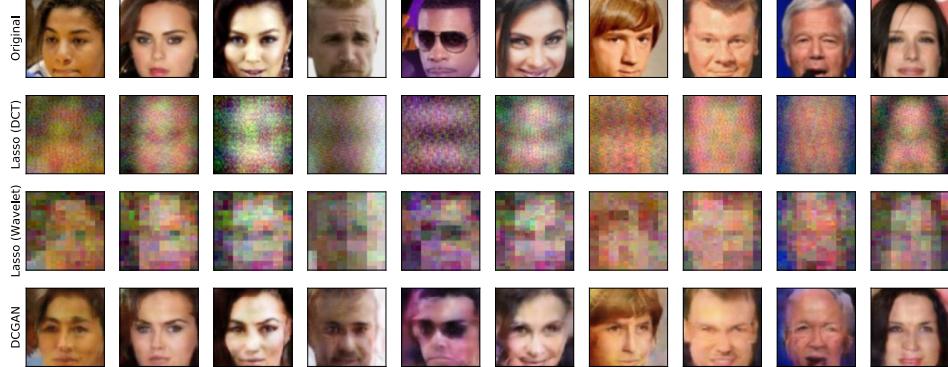


Figure 2.3: Reconstruction results on celebA with $m = 500$ measurements (of $n = 12288$ dimensional vector). We show original images (top row), and reconstructions by Lasso with DCT basis (second row), Lasso with wavelet basis (third row), and our algorithm (last row).

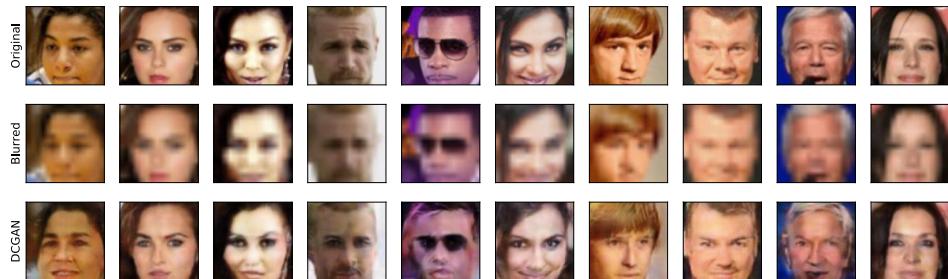


Figure 2.4: Super-resolution results on celebA. Top row has the original images. Second row shows the low resolution ($4x$ smaller) version of the original image. Last row shows the images produced by our algorithm.

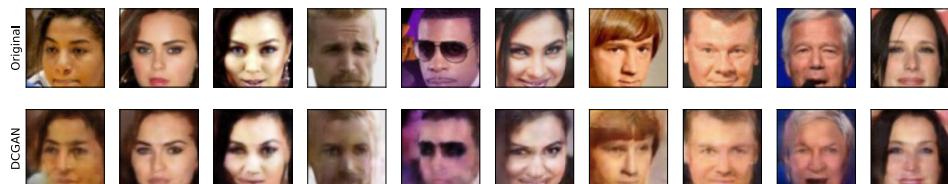


Figure 2.5: Results on the representation error experiments on celebA. Top row shows original images and the bottom row shows closest images found in the range of the generator.

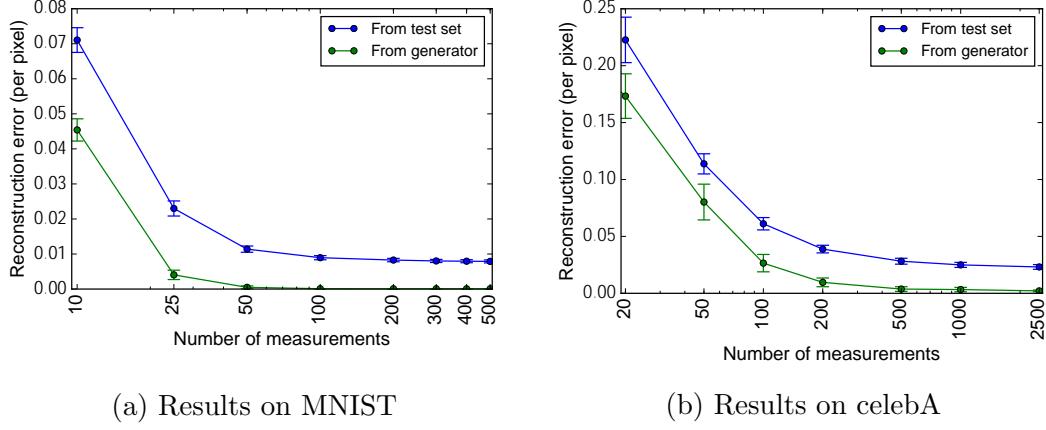


Figure 2.6: Reconstruction error for images in the range of the generator. The vertical bars indicate 95% confidence intervals.

2.7.3 Understanding sources of error

Although better than baselines, our reconstructions still admit some error. There are three sources of this error: (a) Representation error: the image being sensed is far from the range of the generator (b) Measurement error: The finite set of random measurements do not contain all the information about the unknown image (c) Optimization error: The optimization procedure did not find the best z .

In this section we present some experiments that suggest that the representation error is the dominant term. In our first experiment, we ensure that the representation error is zero, and try to minimize the sum of other two errors. In the second experiment, we ensure that the measurement error is zero, and try to minimize the sum of other two.

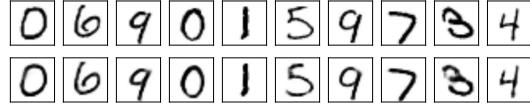


Figure 2.7: Results on the representation error experiments on MNIST. Top row shows original images and the bottom row shows closest images found in the range of the generator.

2.7.3.1 Sensing images from the range of the generator

Our first approach is to sense an image that *is* in the range of the generator. More concretely, we sample a z^* from P_Z . Then we pass it through the generator to get $x^* = G(z^*)$. Now, we pretend that this is a real image and try to sense that. This method eliminates the representation error and allows us to check if our gradient based optimization procedure is able to find z^* by minimizing the objective.

In Fig. 2.6a and Fig. 2.6b, we show the reconstruction error for images in the range of the generators trained on MNIST and celebA datasets respectively. We see that we get almost perfect reconstruction with very few measurements. This suggests that objective is being properly minimized and we indeed get \hat{z} close to z^* . *i.e.* the sum of optimization error and the measurement error is not very large, in the absence of the representation error.

2.7.3.2 Quantifying representation error

We saw that in absence of the representation error, the overall error is small. However from Fig. 2.1, we know that the overall error is still non-zero. So, in this experiment, we seek to quantify the representation error, *i.e.*, how far are the real images from the range of the generator?

From the previous experiment, we know that the \hat{z} recovered by our algorithm is close to z^* , the best possible value, if the image being sensed is in the range of the generator. Based on this, we make an assumption that this property is also true for real images. With this assumption, we get an estimate to the representation error as follows: We sample real images from the test set. Then we use the full image in our algorithm, *i.e.*, our measurement matrix A is identity. This eliminates the measurement error. Using these measurements, we get the reconstructed image $G(\hat{z})$ through our algorithm. The estimated representation error is then $\|G(\hat{z}) - x^*\|^2$. We repeat this procedure several times over randomly sampled images from our dataset and report average representation error values. The task of finding the closest image in the range of the generator has been studied in prior work [68, 84, 78].

On the MNIST dataset, we get average per pixel representation error of 0.005. The recovered images are shown in Fig. 2.7. In contrast with only 100 Gaussian measurements, we are able to get a per pixel reconstruction error of about 0.009.

On the celebA dataset, we get average per pixel representation error of 0.020. The recovered images are shown in Fig. 2.5. On the other hand, with only 500 Gaussian measurements, we get a per pixel reconstruction error of about 0.028.

These experiments suggest that the representation error is the major component of the total error. Thus, a more flexible generative model can help to decrease the overall error on both datasets.

2.8 Conclusion

We demonstrate how to perform compressed sensing using generative models from neural nets. These models can represent data distributions more concisely than standard sparsity models, while their differentiability allows for fast signal reconstruction. This will allow compressed sensing applications to make significantly fewer measurements.

Our theorems and experiments both suggest that, after relatively few measurements, the signal reconstruction gets close to the optimal within the range of the generator. To reach the full potential of this technique, one should use larger generative models as the number of measurements increase. Whether this can be expressed more concisely than by training multiple independent generative models of different sizes is an open question.

Generative models are an active area of research with ongoing rapid improvements. Because our framework applies to general generative models, this improvement will immediately yield better reconstructions with fewer measurements. We also believe that one could also use the performance of generative models for our task as one benchmark for the quality of different models.

Chapter 3

A Robust Median-of-Means Algorithm

3.1 Abstract

The goal of compressed sensing is to estimate a high dimensional vector from an underdetermined system of noisy linear equations. In analogy to classical compressed sensing, here we assume a generative model as a prior, that is, we assume the vector is represented by a deep generative model $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$. Classical recovery approaches such as empirical risk minimization (ERM) are guaranteed to succeed when the measurement matrix is sub-Gaussian. However, when the measurement matrix and measurements are heavy-tailed or have outliers, recovery may fail dramatically. In this paper we propose an algorithm inspired by the Median-of-Means (MOM). Our algorithm guarantees recovery for heavy-tailed data, even in the presence of outliers. Theoretically, our results show our novel MOM-based algorithm enjoys the same sample complexity guarantees as ERM under sub-Gaussian assumptions. Our experiments validate both aspects of our claims: other algorithms are indeed fragile and fail under heavy-tailed and/or corrupted data, while our approach exhibits the predicted robustness.

These results were published at NeurIPS 2020 [136].

3.2 Introduction

Compressive or compressed sensing is the problem of reconstructing an unknown vector $x^* \in \mathbb{R}^n$ after observing $m < n$ linear measurements of its entries, possibly with added noise: $y = Ax^* + \eta$, where $A \in \mathbb{R}^{m \times n}$ is called the measurement matrix and $\eta \in \mathbb{R}^m$ is noise. Even without noise, this is an underdetermined system of linear equations, so recovery is impossible without a structural assumption on the unknown vector x^* . The vast literature [263, 110, 208, 26, 46, 82, 10, 266, 31] on this subject typically assumes that the unknown vector is “natural,” or “simple,” in some application-dependent way.

Compressed sensing has been studied on a wide variety of structures such as sparse vectors [48], trees [53], graphs [291], manifolds [57, 290] or deep generative models [41]. In this paper, we concentrate on deep generative models, which were explored by [41] as priors for sample-efficient reconstruction. Theoretical results in [41] showed that if x^* lies close to the range of a generative model $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ with d -layers, a variant of ERM can recover x^* with $m = O(kd \log n)$ measurements. Empirically, [41] shows that generative models require $5 - 10 \times$ fewer measurements to obtain the same reconstruction accuracy as Lasso. This impressive empirical performance has motivated significant recent research to better understand the behaviour and theoretical limits of compressed sensing using generative priors [106, 142, 180].

A key technical condition for recovery is the Set Restricted Eigenvalue Condition (S-REC) [41], which is a generalization of the Restricted Eigenvalue Condition [39, 44] in sparse recovery. This condition is satisfied if A is a sub-

Gaussian matrix and the measurements satisfy $y = Ax^* + \eta$. This leads to the question: can the conditions on A be weakened, and can we allow for outliers in y and A ? This has significance in applications such as MRI and astronomical imaging, where data is often very noisy and requires significant pruning/cleansing.

As we show in this paper, the analysis and algorithm proposed by [41] are quite fragile in the presence of heavy-tailed noise or corruptions in the measurements. In the statistics literature, it is well known that algorithms such as empirical risk minimization (ERM) and its variants are not robust to even a *single* outlier. Since the algorithm in [41] is a variant of ERM, it is susceptible to the same failures in the presence of heavy-tails and outliers. Indeed, as we show empirically in Section 3.7, precisely this occurs.

Importantly, recovery failure in the setting of [41] (which is also the focus of this paper) can be pernicious, precisely because generative models (by design) output images in their range space, and for well-designed models, these have high perceptual quality. In contrast, when a classical algorithm like LASSO [263] fails, the typical failure mode is the output of a non-sparse vector. Thus in the context of generative models, resilience to outliers and heavy-tails is especially critical. This motivates the need for algorithms that do not require strong assumptions on the measurements.

In this paper, we propose an algorithm for compressed sensing using generative models, which is robust to heavy-tailed distributions and arbitrary outliers. We study its theoretical recovery guarantees as well as empirical per-

formance, and show that it succeeds in scenarios where other existing recovery procedures fail, without additional cost in sample complexity or computation.

3.2.1 Contributions

We propose a new reconstruction algorithm in place of ERM. Our algorithm uses a Median-of-Means (MOM) loss to provide robustness to heavy-tails and arbitrary corruptions. As S-REC may no longer hold, we necessarily use a different analytical approach. We prove recovery results and sample complexity guarantees for this setting even though previous assumptions such as the S-REC [41] condition do not hold. Specifically, our main contributions are as follows.

- (Algorithm) We consider robust compressed sensing for generative models where (i) a constant fraction of the measurements and measurement matrix are arbitrarily (perhaps maliciously) corrupted and (ii) the random ensemble only satisfies a weak moment assumption.

We propose a novel algorithm to replace ERM. Our algorithm uses a median-of-means (MOM) tournament [187, 165] i.e., a min-max optimization framework for robust reconstruction. Each iteration of our MOM-based algorithm comes at essentially no additional computational cost compared to an iteration of standard ERM. Moreover, as our code shows, it is straightforward to implement.

- (Analysis and Guarantees) We analyze the recovery guarantee and outlier-

robustness of our algorithm when the generative model is a d -layer neural network using ReLU activations. Specifically, in the presence of a constant fraction of outliers in y and A , we achieve $\|G(\hat{z}) - G(z^*)\|^2 \leq O(\sigma^2 + \tau)$ with sample size $m = O(kd \log n)$, where σ^2 is the variance of the heavy-tailed noise, and τ is the optimization accuracy. Using different analytical tools (necessarily, since we do not assume sub-Gaussianity), we show our algorithm, even under heavy-tails and corruptions, has the same sample complexity as the previous literature has achieved under much stronger sub-Gaussian assumptions. En route to our result, we also prove an interesting result for ERM: by avoiding the S-REC-based analysis, we show that the standard ERM algorithm does in fact succeed in the presence of a heavy-tailed measurement matrix, thereby strengthening the best-known recovery guarantees from [41]. This does not extend (as our empirical results demonstrate) to the setting of outliers, or of heavy-tailed measurement noise. For these settings, our new algorithm is required.

- (Empirical Support) We empirically validate the effectiveness of our robust recovery algorithm on MNIST and CelebA-HQ. Our results demonstrate that (as our theory predicts) our algorithm succeeds in the presence of heavy-tailed noise, heavy-tailed measurements, and also in the presence of arbitrary outliers. At the same time our experiments confirm that ERM can fail, and in fact fails dramatically: through an experiment on the CelebA-HQ data set, we demonstrate that the ERM recovery approach [41], as well as other natural approaches including ℓ_1 loss minimization and trimmed loss

minimization [246], can recover images that have little resemblance to the original.

3.2.2 Related work

Compressed sensing with outliers or heavy-tails has a long history. To deal with outliers only in y , classical techniques replace the ERM with a robust loss function such as ℓ_1 loss or Huber loss [172, 210, 185, 69], and obtain the optimal statistical rates. Much less is known for outliers in y and A for robust compressed sensing. Recent progress on robust sparse regression [60, 28, 61, 76, 220, 178, 177, 246] can handle outliers in y and A , but their techniques cannot be directly extended to arbitrary generative models G . Another line of research [121, 202, 187, 165] considers compressed sensing where the measurement matrix A and y have heavy-tailed distributions. Their techniques leverage variants of Median-of-Means (MOM) estimators on the loss function under weak moment assumptions instead of sub-Gaussianity, which generalize the classical MOM mean estimator in one dimension [209, 138, 14, 202].

[281] deals with compressed sensing of generative models when the measurements and the responses are non-Gaussian. However, the distribution model in [281] requires more stringent conditions compared to the weak moment assumption as will be specified in Definition 3.4.1, and their algorithm cannot tolerate arbitrary corruptions. [295] consider ℓ_1 -minimization for outlier detection using generative models, assuming the outliers in y are sparse.

Generative priors have shown great promise in compressed sensing and

other inverse problems, starting with [41], who generalized the theoretical framework of compressive sensing and restricted eigenvalue conditions [263, 82, 39, 44, 118, 33, 32, 87] for signals lying on the range of a deep generative model [98, 158]. Results in [142, 180, 137] established that the sample complexities in [41] are order optimal. The approach in [41] has been generalized to tackle different inverse problems [105, 24, 22, 205, 23, 223, 24, 179, 20, 132, 104, 16]. Alternate algorithms for reconstruction include [42, 75, 140, 89, 88, 252, 193, 75, 214, 111, 113]. The complexity of optimization algorithms using generative models have been analyzed in [95, 114, 171, 106]. See [211] 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 [7, 128, 200, 120].

3.3 Notation

For functions $f(n)$ and $g(n)$, we write $f(n) \lesssim g(n)$ to denote that there exists a universal constant $c_1 > 0$ such that $f(n) \leq c_1 g(n)$. Similarly, we write $f(n) \gtrsim g(n)$ to denote that there exists a universal constant $c_2 > 0$ such that $f(n) \geq c_2 g(n)$. We write $f(n) = O(g(n))$ to imply that there exists a positive constant c_3 and a natural number n_0 such that for all $n \geq n_0$, we have $|f(n)| \leq c_3 g(n)$. Similarly, we write $f(n) = \Omega(g(n))$ to imply that there exists a positive constant c_4 and a natural number n_1 such that for all $n \geq n_1$, we have $|f(n)| \geq c_4 g(n)$.

3.4 Problem formulation

Let $x^* = G(z^*) \in \mathbb{R}^n$ be the fixed vector of interest. The deep generative model $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ ($k \ll n$) maps from a low dimensional latent space to a higher dimensional space. In this paper, G is a feedforward neural network with ReLU activations and d layers.

Our definition of heavy-tailed samples assumes that the measurement matrix A only has bounded fourth moment. Our corruption model is Huber's ϵ -contamination model [123]. This model allows corruption in the measurement matrix A and measurements y . Precisely, these are:

Definition 3.4.1 (Heavy-tailed samples). *We say that a random vector a is heavy-tailed if for a universal constant $C > 0$, the 4th moment of a satisfies*

$$(\mathbb{E} [\langle a, u \rangle^4])^{\frac{1}{4}} \leq C (\mathbb{E} [\langle a, u \rangle^2])^{\frac{1}{2}}, \quad \forall u \in \mathbb{R}^n.$$

For all $\delta > 0$, the $(4 + \delta)^{th}$ moment of a need not exist, and we make no assumptions on them.

Definition 3.4.2 (ε -corrupted samples). *We say that a collection of samples $\{y_i, a_i\}$ is ε -corrupted if they are i.i.d. observations drawn from the mixture*

$$\{y_i, a_i\} \sim (1 - \varepsilon)P + \varepsilon Q,$$

where P is the uncorrupted distribution, Q is an arbitrary distribution.

Thus we assume that samples $\{y_i, a_i\}_{i=1}^m$ are generated from $(1 - \varepsilon)P + \varepsilon Q$, where Q is an adversary, and P satisfies the following:

Assumption 3.4.3. Samples $(y_i, a_i) \sim P$ satisfy $y_i = a_i^\top G(z^*) + \eta_i$, where the random vector a_i is isotropic and heavy-tailed as in Definition 3.4.2, and the noise term η_i is independent of a_i , i.i.d. with zero mean and bounded variance σ^2 .

3.5 Our Algorithm

$\|\cdot\|$ refers to ℓ_2 unless specified otherwise. The procedure proposed by [41] finds a reconstruction $\hat{x} = G(\hat{z})$, where \hat{z} solves:

$$\hat{z} := \arg \min_{z \in \mathbb{R}^k} \|AG(z) - y\|^2.$$

This is essentially an ERM-based approach. As is well known from the classical statistics literature, ERM's success relies on strong concentration properties, guaranteed, e.g., if the data are all sub-Gaussian. ERM may fail, however, in the presence of corruption or heavy-tails. Indeed, our experiments demonstrate that in the presence of outliers in y or A , or heavy-tailed noise in y , [41] fails to recover $G(z^*)$.

Remark Unlike typical problems in M -estimation and high dimensional statistics, the optimization problem that defines the recovery procedure here is non-convex, and thus in the worst case, computationally intractable. Interestingly, despite non-convexity, as demonstrated in [41], (some appropriate version of) gradient descent is empirically very successful. In this paper, we take this as a computational primitive, thus sidestepping the challenge of proving whether a gradient-descent based method can efficiently provide guar-

anted inversion of a generative model. Our theoretical guarantees are therefore statistical but our experiments show empirically excellent performance.

3.5.1 MOM objective

It is well known that the median of means estimator achieves nearly sub-Gaussian concentration for one dimensional mean estimation of variance bounded random variables [209, 138, 14]. Inspired by the median-of-means algorithm, we propose the following algorithm to handle heavy-tails and outliers in y and A . We partition the set $[m]$ into M disjoint batches $\{B_1, B_2, \dots, B_M\}$ such that each batch has cardinality $b = \frac{m}{M}$. Without loss of generality, we assume that M exactly divides m , so that b is an integer. For the j^{th} batch B_j , define the function

$$\ell_j(z) := \frac{1}{b} \|A_{B_j}G(z) - y_{B_j}\|^2, \quad (3.1)$$

where $A_{B_j} \in \mathbb{R}^{b \times n}$ denotes the submatrix of A corresponding to the rows in batch B_j . Similarly, $y_{B_j} \in \mathbb{R}^b$ denotes the entries of y corresponding to the batch B_j . Our workhorse is a novel variant of median-of-means (MOM) tournament procedure [187, 165] using the loss function eq. (3.1):

$$\hat{z} = \arg \min_{z \in \mathbb{R}^k} \max_{z' \in \mathbb{R}^k} \text{median}(\ell_j(z) - \ell_j(z')). \quad (3.2)$$

We do not assume that the minimizer is unique, since we only require a reconstruction $G(\hat{z})$ which is close to $G(z^*)$. Any value of z in the set of minimizers will suffice. The intuition behind this aggregation of batches is that if the inner player z' chooses a point close to z^* , then the outer player z must also choose

Algorithm 1 Robust compressed sensing of generative models

- 1: **Input:** Data samples $\{y_j, a_j\}_{j=1}^m$.
- 2: **Output:** $G(\hat{z})$.
- 3: **Parameters:** Number of batches M .
- 4: Initialize z and z' .
- 5: **for** $t = 0$ to $T - 1$, **do**
- 6: For each batch $j \in [M]$, calculate $\frac{1}{|B_j|}(\ell_j(z) - \ell_j(z'))$ by eq. (3.1).
- 7: Pick the batch with the median loss $\underset{1 \leq j \leq M}{\text{median}}(\ell_j(z) - \ell_j(z'))$, and evaluate the gradient for z and z' using backpropagation on that batch.
 - (i) perform gradient descent for z ;
 - (ii) perform gradient ascent for z' .
- 8: **end for**
- 9: Output the $G(\hat{z}) = G(z)$.

a point close to z^* in order to minimize the objective. Once this happens, there is no better option for z' . Hence a neighborhood around z^* is almost an equilibrium, and in fact there can be no neighborhood far from z^* with such an equilibrium.

Computational considerations. The objective function eq. (3.2) is not convex and we use Algorithm 1 as a heuristic to solve eq. (3.2). In Section 3.7, we empirically observe that gradient-based methods are able to minimize this objective and have good convergence properties. Our main theorem guarantees that a small value of the objective implies a good reconstruction and hence we can certify reconstruction quality using the obtained final value of the objective.

3.6 Theoretical results

We begin with a brief review of the Restricted Eigenvalue Condition in standard compressed sensing and show that S-REC is satisfied by heavy-tailed distributions.

3.6.1 Set-Restricted Eigenvalue Condition for heavy-tailed distributions

Most theoretical guarantees for compressed sensing rely on variants of the Restricted Eigenvalue Condition(REC) [39, 44] and the closest to our setting is the Set Restricted Eigenvalue Condition [41](S-REC). Formally, $A \in \mathbb{R}^{m \times n}$ satisfies $S\text{-REC}(S, \gamma, \delta)$ on a set $S \subseteq \mathbb{R}^n$ if for all $x_1, x_2 \in S$,

$$\|Ax_1 - Ax_2\| \geq \gamma \|x_1 - x_2\| - \delta.$$

While we can prove many powerful results using the REC condition, proving that a matrix satisfies REC typically involves sub-Gaussian entries in A . If we don't have sub-Gaussianity, proving REC requires a finer analysis. A recent technique called the *small-ball method* [197] requires significantly weaker assumptions on A , and can be used to show REC [197, 265] for A satisfying Assumption 3.4.3. While this technique can be used for sparse vectors, we do not have a general understanding of what structures it can handle, since existing proofs make heavy use of sparsity.

We now show that a random matrix whose rows satisfy Assumption 3.4.3 will satisfy S-REC over the range of a generator $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ with high probability. This generalizes Lemma 4.2 in [41]– the original lemma required i.i.d.

sub-Gaussian entries in the matrix A , whereas the following lemma only needs the rows to have bounded fourth moments.

Lemma 3.6.1. *Let $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be a d -layered neural network with ReLU activations. Let $A \in \mathbb{R}^{m \times n}$ be a matrix with i.i.d. rows satisfying Definition 3.4.1. For any $\gamma < 1$, if $m = \Omega\left(\frac{1}{1-\gamma^2}kd\log n\right)$, then with probability $1 - e^{-\Omega(m)}$, for all $z_1, z_2 \in \mathbb{R}^k$, we have*

$$\frac{1}{m} \|AG(z_1) - AG(z_2)\|^2 \geq \gamma^2 \|G(z_1) - G(z_2)\|^2.$$

This implies that the ERM approach of [41] still works when we only have a heavy-tailed measurement matrix A . However, as we show in our experiments, heavy-tailed noise in y and outliers in y, A will make ERM fail catastrophically. In order to solve this problem, we leverage the median-of-means tournament defined in eq. (3.2), and we will now show it is robust to heavy-tails and outliers in y, A .

3.6.2 Main results

We now present our main result. Theorem 3.6.5 provides recovery guarantees in terms of the error in reconstruction in the presence of heavy-tails and outliers, where \hat{z} is the (approximate) minimizer of eq. (3.2). First we show that the minimum value of the objective in eq. (3.2) is indeed small if there are no outliers.

Lemma 3.6.2. *Let M denote the number of batches. Assume that the measurements y and measurement matrix A are drawn from the uncorrupted dis-*

tribution satisfying Assumption 3.4.3. Then with probability $1 - e^{-\Omega(M)}$, the objective in Equation (3.2) satisfies

$$\min_{z \in \mathbb{R}^k} \max_{z' \in \mathbb{R}^k} \text{median}_{1 \leq j \leq M}(\ell_{B_j}(z) - \ell_{B_j}(z')) \leq 4\sigma^2. \quad (3.3)$$

We now introduce Lemma 3.6.3 and Lemma 3.6.4, which control two stochastic processes that appear in eq. (3.2). We show that minimizing the objective in eq. (3.2) implies that you are close to the unknown vector $G(z^*)$. Notice that since z^* is one feasible solution of the inner maximization step of z' , we can consider $z' = z^*$. Now consider the difference of square losses in eq. (3.2), which is given by:

$$\begin{aligned} \ell_j(\hat{z}) - \ell_j(z^*) &= \frac{1}{b} \|A_{B_j}G(\hat{z}) - y_{B_j}\|^2 - \frac{1}{b} \|A_{B_j}G(z^*) - y_{B_j}\|^2, \\ &= \frac{1}{b} \|A_{B_j}(G(\hat{z}) - G(z^*))\|^2 - \frac{2}{b} \eta_{B_j}^\top (A_{B_j}(G(\hat{z}) - G(z^*))), \end{aligned}$$

where the last line follows from an elementary arithmetic manipulation.

Assume we have the following bounds on a majority of batches:

$$\frac{1}{b} \|A_{B_j}(G(\hat{z}) - G(z^*))\|^2 \gtrsim \|G(\hat{z}) - G(z^*)\|^2, \quad (3.4)$$

$$-\frac{2}{b} \eta_{B_j}^\top (A_{B_j}(G(\hat{z}) - G(z^*))) \gtrsim -\|G(\hat{z}) - G(z^*)\|. \quad (3.5)$$

Since the objective is the median of the sum of the above terms, a small value of the objective implies that $\|G(\hat{z}) - G(z^*)\|$ is small. We formally show these bounds in Lemma 3.6.3, Lemma 3.6.4.

Lemma 3.6.3. *Let $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be a generative model from a d -layer neural network using ReLU activations. Let $A \in \mathbb{R}^{m \times n}$ be a matrix with i.i.d. uncorrupted rows satisfying Definition 3.4.1. Let the batch size $b = \Theta(C^4)$, let*

the number of batches satisfy $M = \Omega(kd \log n)$, and let γ be a constant which depends on the moment constant C . Then with probability at least $1 - e^{-\Omega(m)}$, for all $z_1, z_2 \in \mathbb{R}^k$ there exists a set $J \subseteq [M]$ of cardinality at least $0.9M$ such that

$$\frac{1}{b} \|A_{B_j}(G(z_1) - G(z_2))\|^2 \geq \gamma^2 \|G(z_1) - G(z_2)\|^2, \forall j \in J.$$

Lemma 3.6.4. Consider the setting of Lemma 3.6.3 with measurements satisfying $y = AG(z^*) + \eta$, where y, A, η satisfy Assumption 3.4.3 with noise variance σ^2 . For a constant batch size b and number of batches $M = \Omega(kd \log n)$, with probability at least $1 - e^{-\Omega(m)}$, for all $z \in \mathbb{R}^k$ there exists a set $J \subseteq [M]$ of cardinality at least $0.9M$ such that

$$\frac{1}{b} |\eta_{B_j}^T A_{B_j}(G(z) - G(z^*))| \leq \sigma \|G(z) - G(z^*)\|, \forall j \in J.$$

The above lemmas do not account for the ϵ -corrupted samples in Definition 3.4.2. However, since the batch size is constant in both the lemmas, there exists a value of ϵ such that sufficiently many batches have no corruptions. Hence we can apply Lemma 3.6.3, Lemma 3.6.4 to these uncorrupted batches. Using these lemmas with a constant batch size b , we obtain Theorem 3.6.5. We defer its proof to Section B.5.

Theorem 3.6.5. Let $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be a generative model from a d -layer neural network using ReLU activations. There exists a (sufficiently small) constant fraction ϵ which depends on the moment constant C in Definition 3.4.1 such that the following is true. We observe $m = O(kd \log n)$ ϵ -corrupted samples from Definition 3.4.2, under Assumption 3.4.3. For any $z^* \in \mathbb{R}^k$, let \hat{z}

minimize the objective function given by eq. (3.2) to within additive τ of the optimum. Then there exists a (sufficiently large) constant c , such that with probability at least $1 - e^{-\Omega(m)}$, the reconstruction $G(\hat{z})$ satisfies

$$\|G(\hat{z}) - G(z^*)\|^2 \leq c(\sigma^2 + \tau),$$

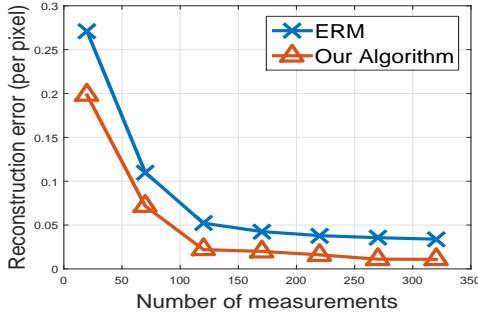
where σ^2 is the variance of noise under Assumption 3.4.3.

We briefly discuss the implications of Theorem 3.6.5, with regards to sample complexity and error in reconstruction.

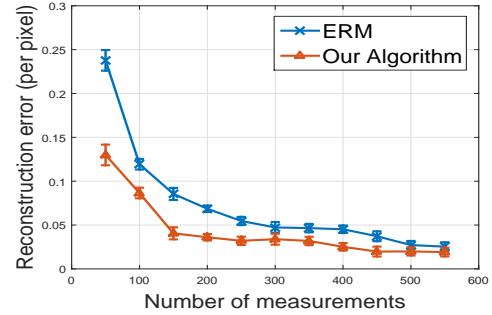
Sample Complexity. Our sample complexity matches that of [41] up to constant factors. This shows that the minimizer of eq. (3.2) in the presence of heavy-tails and outliers provides the same guarantees as in the case of ERM with sub-Gaussian measurements.

Statistical accuracy and robustness. Let us analyze the error terms in our theorem. The term τ is a consequence of the minimization algorithm not being perfect, since it only reaches within τ of the true minimum. Hence it cannot be avoided. The term σ^2 is due to the noise in measurements. In the main result of [41], the reconstruction $G(\hat{z})$ has error bounded by $\|G(\hat{z}) - G(z^*)\|^2 \lesssim \|\eta\|^2/m + \tau$.¹ This gives the following conditions:

¹In [41], the bound is stated as $\|\eta\|^2$, but our A has a different scaling, and hence the correct bound in our setting is $\|\eta\|^2/m$.



(a) Results on MNIST.



(b) Results on CelebA-HQ

Figure 3.1: We compare Algorithm 1 with the baseline ERM [41] under the heavy-tailed setting *without* arbitrary outliers. We fix $k = 100$ for the MNIST dataset and $k = 512$ for the CelebA-HQ dataset. We vary the number of measurements, and plot the reconstruction error per pixel averaged over multiple trials. With increasing number of measurements, we observe the reconstruction error decreases. For heavy-tailed y and A *without* arbitrary outliers, our method obtains significantly smaller reconstruction error in comparison to ERM.

- If η is sub-Gaussian with variance σ^2 , then $\|\eta\|^2/m \approx \sigma^2$ with high probability. Hence our bounds match up to constants.
- If higher order moments of η do not exist, an application of Chebyshev's inequality says that with probability $1 - \delta$, [41] has $\|G(z^*) - G(\tilde{z})\|^2 \approx \sigma^2/(m\delta)$, and this can be extremely large if we want $\delta = e^{-\Omega(m)}$.

Hence our method is clearly superior if η only has bounded variance, and if η is sub-Gaussian, then our bounds match up to constants. In the presence of corruptions, [41] has no provable guarantee.

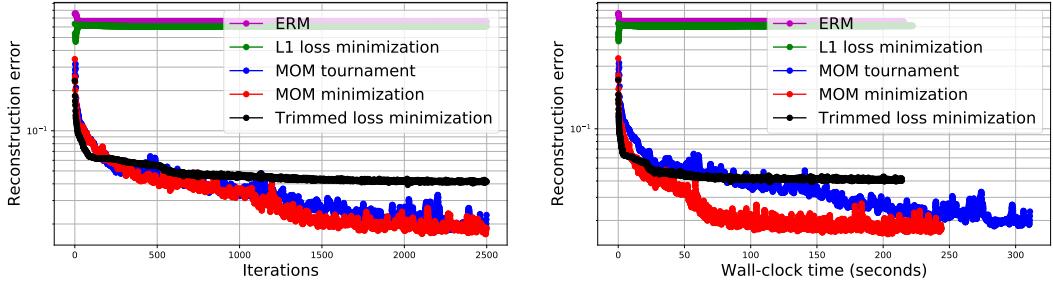


Figure 3.2: Plot of the reconstruction error versus the iteration number (left) and plot of the reconstruction error versus wall-clock time (right). ERM [41] and ℓ_1 minimization fail to converge. Our two proposed methods, MOM tournament(blue) and MOM minimization(red), have the smallest reconstruction error. We provide a theoretical analysis for the MOM tournament algorithm, and observe that direct minimization of the MOM objective also works in practice. The computation time of our algorithms is nearly the same as the baselines.

3.7 Experiments

In this section, we study the empirical performance of our algorithm on generative models trained on real image datasets. We show that we can reconstruct images under heavy-tailed samples and arbitrary outliers. For additional experiments and experimental setup details, see Section B.6.

Heavy-Tailed Samples In this experiment, we deal with the *uncorrupted* compressed sensing model P , which has heavy-tailed measurement matrix and stochastic noise: $y = AG(z^*) + \eta$. We use a Student's t -distribution (a typical example of heavy-tails) for A and η . We compare Algorithm 1 with the baseline ERM [41] for heavy-tailed data *without* arbitrary corruptions on MNIST [168] and CelebA-HQ [144, 184]. We trained a DCGAN [225] with $k = 100$ and $d = 5$ layers to produce 64×64 MNIST images. For CelebA-HQ, we used a

PG-GAN [144] with $k = 512$ to produce images of size $256 \times 256 \times 3 = 196,608$.

We vary the number of measurements m and obtain the reconstruction error $\|G(\hat{z}) - G(z^*)\|^2/n$ for Algorithm 1 and ERM, where $G(z^*)$ is the ground truth image. In Figure 3.1, Algorithm 1 and ERM both have decreasing reconstruction error per pixel with increasing number of measurements. To conclude, even for heavy-tailed noise *without* arbitrary outliers, Algorithm 1 obtains significantly smaller reconstruction error when compared to ERM.

Arbitrary corruptions. In this experiment, we use the same heavy-tailed samples as above, and we add $\varepsilon = 0.02$ -fraction of arbitrary corruption. We set the outliers of measurement matrix A as random sign matrix, and the outliers of y are fixed to be -1 . We note that we don't use any targeted attack to simulate the outliers. We perform our experiments on the CelebA-HQ dataset using a PG-GAN of latent dimension $k = 512$, and fix the number of measurements to $m = 1000$.

We compare our algorithm to a number of natural baselines. Our first baseline is ERM [41] which is not designed to deal with outliers. While its fragility is interesting to note, in this sense it is not unexpected. For outliers in y , classical robust methods replace the loss function by an ℓ_1 loss function or Huber loss function. This is done in order to avoid the squared loss, which makes recovery algorithms very sensitive to outliers. In this case, we have $\hat{z} := \arg \min \|y - AG(z)\|_1$.

We also investigate the performance of trimmed loss minimization,

which is a recent algorithm proposed by [246]. This algorithm picks the t -fraction of samples with smallest empirical loss for each update step, where t is a hyper-parameter.

We run Algorithm 1 and its variant MOM minimization. The MOM minimization directly minimizes

$$\hat{z} = \arg \min_{z \in \mathbb{R}^k} \text{median}_{1 \leq j \leq M}(\ell_j(z)), \quad (3.6)$$

and we use gradient-based methods similar to Algorithm 1 to solve it. Since Algorithm 1 optimizes z and z' in one iteration, the actual computation time of MOM tournament is twice that of MOM minimization. As shown in Figure 3.2, Figure 3.3, ERM [41] and ℓ_1 loss minimization fail to converge to the ground truth and in particular, they may recover a completely different person. Trimmed loss minimization [246] only succeeds on occasion, and when it fails, it obtains a visibly different person. The convergence of the MOM minimization per iteration is very similar to the MOM tournament, and they both achieve much smaller reconstruction error compared to trimmed loss minimization. The right panel of Figure 3.2 plots the reconstruction error versus the actual computation time, showing our algorithms match baselines. We plot the MSE vs. number of measurements in Figure 3.4b, where the fraction of corruptions is set to $\varepsilon = 0.02$.

Miscellaneous Experiments *Is ERM ever better than MOM?* So far we have analyzed cases where MOM performs better than ERM. Since ERM is

known to be optimal in linear regression when dealing with uncorrupted sub-Gaussian data, we expect it to be superior to MOM when our measurements are all sub-Gaussian. We evaluate this in Fig. 3.4a and observe that ERM obtains smaller MSE in this setting. Notice that as we reduce the number of batches in MOM, it approaches ERM.

How sensitive is MOM to the number of batches? In Figure 3.4c we study the MSE of MOM tournaments and MOM minimization as we vary the number of batches.

In order to select the optimal number of batches (M), we keep a set of validation measurements that we do not use in the optimization routines for estimating x . We can run MOM for different value of M to get multiple reconstructions, and then evaluate each reconstruction using the validation measurements to pick the best reconstruction. Note that one should use the median-of-means loss while evaluating the validation error as well.

3.8 Conclusion

The phenomenon observed in Figure 3.3 highlights the importance of our method. Our work raises several questions about why the objective we consider can be minimized, and suggests we need a new paradigm for analysis that accounts for similar instances that enjoy empirical success, even though they can be provably hard in the worst case.

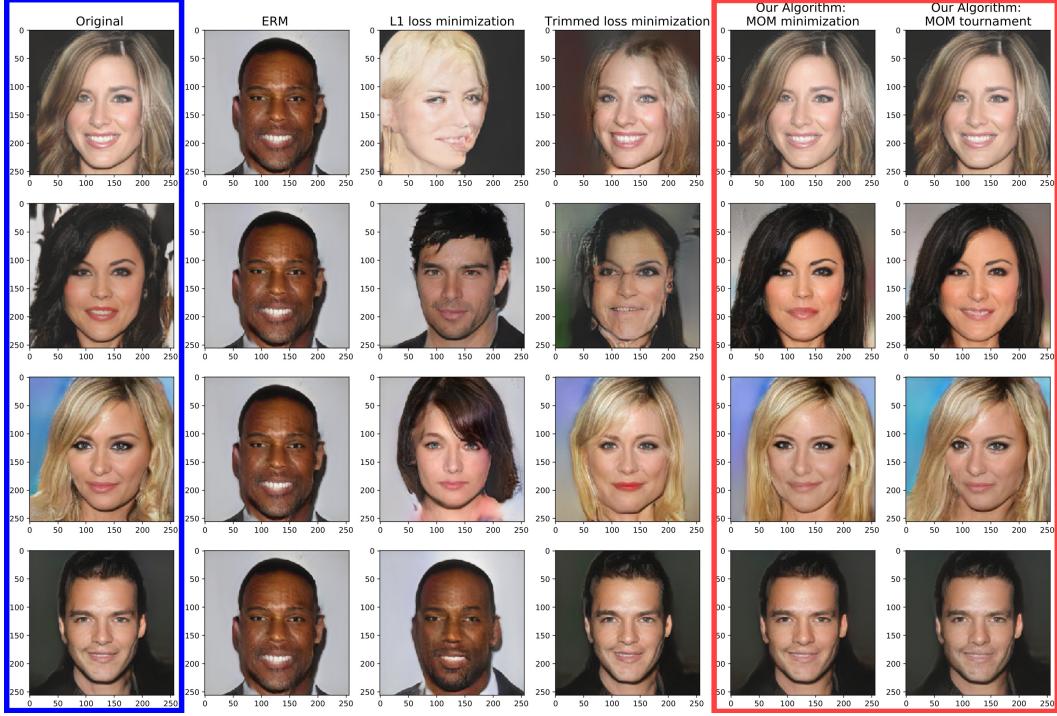


Figure 3.3: Reconstruction results on CelebA-HQ for $m = 1000$ measurements with 20 corrupted measurements. For each row, the first column is ground truth from a generative model. Subsequent columns show reconstructions by ERM [41], ℓ_1 -minimization, trimmed loss minimization [246]. In particular, vanilla ERM, ℓ_1 -minimization obtain completely different faces. Since we use the same outlier for different rows, vanilla ERM produces the same reconstruction irrespective of the ground truth. Trimmed loss minimization only succeeds on occasion (the last row), and when it fails, it obtains a similar but still different face. The last two columns show reconstructions by our proposed algorithms. The second to last one is directly minimizing the MOM objective eq. (3.6), and the last column minimizes the MOM tournaments algorithm, and observe that direct minimization of the MOM objective also works in practice. We provide a theoretical analysis for the MOM tournaments algorithm, and observe that direct minimization of the MOM objective also works in practice. We get a high quality reconstruction under heavy-tailed measurements and arbitrary outliers.

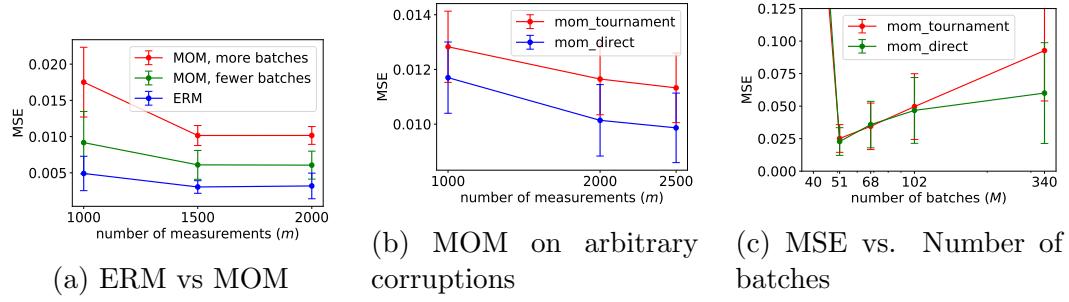


Figure 3.4: (a) We compare ERM and MOM by plotting MSE vs number of measurements when the measurements are *sub-Gaussian without corruptions*. (b) Aggregate statistics for MOM in the presence of corruptions. (c) MSE vs number of batches for MOM on 1000 heavy-tailed measurements and 20 corruptions. All error bars indicate 95% confidence intervals. Plots use a PGGAN on CelebA-HQ.

Chapter 4

Instance-Optimality via Posterior Sampling

4.1 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.

These results were published at ICML 2021 [134].

4.2 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 [190, 189], neuronal spike trains [116] and efficient sensing cameras [83]. 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 \hat{x} from (A, y) such that

$$\|x^* - \hat{x}\| \leq C\sigma \tag{4.1}$$

with high probability for some constant C . When x^* is k -sparse, this was shown by Candés, Romberg, and Tao [48] 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 [54], graphs [291], manifolds [57, 290], or deep generative models [41, 20]. 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, [41] uses the structure given by a deep generative model $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$; 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 [159] and pixelRNN [212] 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 [8, 302] 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?

4.2.1 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 4.2.1. *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 \hat{x} such that*

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



Figure 4.1: 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.

Then posterior sampling (see Definition 4.2.3) 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 \hat{x} satisfying

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

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

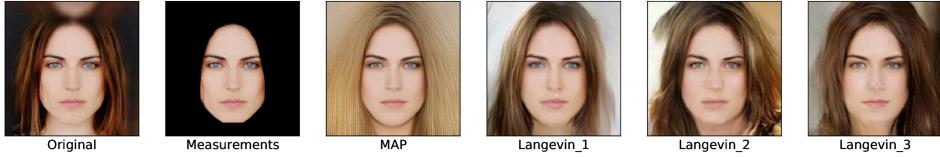


Figure 4.2: Reconstruction results for inpainting on CelebA-HQ using Glow. The first column shows the original image, second column shows the measurements by removing the hair and background, the third column shows reconstruction by MAP, and the last three columns show samples from posterior sampling via Langevin dynamics. MAP produces the same washed out image all the time, whereas posterior sampling produces images with diversity.

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 (4.1).

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 $\text{Cov}_{\eta,0.01}(R)$ the minimum number of η -radius balls needed to cover 99% of R . An elementary calculation shows

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

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 4.2.2. 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

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

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

When $\delta = 0$, this is $N_\eta(\text{supp } R)$, 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 [41], which depend on the covering numbers of low-dimensional generative models.

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

Definition 4.2.3. *Given an observation y , the posterior sampling recovery algorithm with respect to P outputs \hat{x} according to the posterior distribution $P(\cdot | y)$.*

Contribution 3: Sample complexity upper bound. Our main positive result is that posterior sampling achieves the guarantees of equation (4.1) for *general* distributions R , with $O(\log \text{Cov}_{\sigma,\delta}(R))$ 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 4.2.4 (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 $\hat{x} \sim P(\cdot | y)$. For any $\eta \geq \sigma$, with*

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

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

Contribution 4: Sample complexity lower bound. Our second main result lower bounds the sample complexity for *any* distribution. 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 [221, 45, 129]; the closest result we are aware of is [8], which characterizes product distributions.

Theorem 4.2.5 (Lower bound). *Let R be any distribution over an ℓ_2 ball of radius r , and consider any method to achieve $\|\hat{x} - x^*\| \leq \eta$ with 99% probability, using an arbitrary measurement matrix $A \in \mathbb{R}^{m \times n}$ with noise level σ . This must have*

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

for some constant $C' > 0$.

Note that Theorem 4.2.4 and 4.2.5 directly give Theorem ???. For more precisely stated and general versions of these results, including dependence on the failure probability δ , see Theorems 4.4.4 and 4.5.1.

4.2.2 Related Work

Generative priors have shown great promise in compressed sensing and other inverse problems, starting with [41], who generalized the theoretical framework of compressive sensing and restricted eigenvalue conditions [263, 82, 39, 44, 118, 33, 32, 87] for signals lying on the range of a deep generative model [98, 158].

Lower bounds in [142, 180, 137] established that the sample complexities in [41] are order optimal. The approach in [41] has been generalized

to tackle different inverse problems such as robust compressed sensing [136], phase retrieval [105, 24, 132], blind image deconvolution [22], seismic inversion [205], one-bit recovery [223, 179], and blind demodulation [104]. Alternate algorithms for reconstruction include sparse deviations from generative models [75], task-aware compressed sensing [140], PnP [214, 89, 88], iterative projections [193], OneNet [229] and Deep Decoder [111, 113]. The complexity of optimization algorithms using generative models have been analyzed for ADMM [95], PGD [114], layer-wise inversion [171], and gradient descent [106]. Experimental results in [20, 285, 175] show that invertible models have superior performance in comparison to low dimensional models. See [211] 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 [7, 128, 200, 120].

Lower bounds for ℓ_2/ℓ_2 recovery of sparse vectors can be found in [239, 221, 8, 129, 45], and these are related to the lower bound in (4.2.5). The closest result is that of [8], 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 [287, 286]. The δ -Minkowski dimension of a probability measure used in [287, 286, 217] can be derived from our (ε, δ) -covering number by taking the limit $\varepsilon \rightarrow 0$. [228] 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 [298, 34].

4.3 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 Wasserstein distance. For two probability distributions μ, ν supported on Ω , and for any $p \geq 1$, the Wasserstein- p [275, 18] and Wasserstein- ∞ [51] distances are defined as:

$$\begin{aligned}\mathcal{W}_p(\mu, \nu) &:= \inf_{\gamma \in \Pi(\mu, \nu)} \left(\mathbb{E}_{(u,v) \sim \gamma} [\|u - v\|^p] \right)^{1/p}, \\ \mathcal{W}_\infty(\mu, \nu) &:= \inf_{\gamma \in \Pi(\mu, \nu)} \left(\gamma\text{-ess sup}_{(u,v) \in \Omega^2} \|u - v\| \right),\end{aligned}$$

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 $A_{ij} \sim \mathcal{N}(0, 1/m)$.

4.4 Upper Bound

4.4.1 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 4.3). In this example, two η radius balls can cover the whole space, so the parameters in Theorem 4.2.4 will be $\sigma = 0$ and $\text{Cov}_{\eta,0}(P) = 2$. Applying Theorem 4.2.4 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 (\hat{x}) according to the posterior density

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

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 .

Since the balls $B_0 \& B_{\tilde{x}}$ are well separated, the ground truth and the

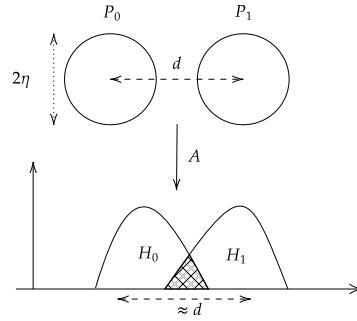


Figure 4.3: Illustrative example for the upper bound. The signal x^* is drawn from a mixture of two well-separated balls. The observations $y = Ax^*$ are then drawn from a mixture of two distributions H_0, H_1 that may overlap. The probability that posterior sampling outputs something from the wrong ball is proportional to the (shaded) overlap between these distributions, which is atmost $1 - TV(H_0, H_1)$.

reconstruction are far apart if and only if they lie in different balls, i.e., $\{x^* \in B_0, \hat{x} \in B_{\hat{x}}\}$, 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 4.4.1. *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 $z^* \sim \text{Bernoulli}(c)$.*

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

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

The proof of this, as well as all parts of the upper bound, can be found in Appendix C.1.

In our current example, this gives us

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

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

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

$$\begin{aligned} \Pr[\|x^* - \hat{x}\| > 2\eta] &\leq \Pr[x^* \in B_0, \hat{x} \in B_{\tilde{x}}] + \\ &\quad \Pr[x^* \in B_{\tilde{x}}, \hat{x} \in B_0], \\ &\leq 2(1 - TV(H_0, H_1)). \end{aligned} \tag{4.2}$$

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 4.4.2. *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 η -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:

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

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

$$TV(H_0, H_1) \geq 1 - e^{-\Omega(m)}. \quad (4.3)$$

Substituting equation (4.3) in equation (4.2), we have

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

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.

4.4.2 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 $\text{Cov}_{\eta,0}(R) \leq e^{o(m)}$, you can cover R with a small number of balls, and essentially apply Lemma 4.4.2 with a union bound. There are a few details (e.g., Lemma 4.4.2 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 $\log \text{Cov}_{\eta,0}(R)$ is bounded, then posterior sampling works well with $1 - e^{-\Omega(m)}$ probability.

Distribution mismatch in \mathcal{W}_∞ . 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}_∞ .

Lemma 4.4.3. *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 $\hat{x} \sim P(\cdot|y, A)$ and $\hat{z} \sim P(\cdot|u, A)$. For any $d > 0$, we have

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

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 4.4.3 says the situation is within $e^{o(m)}$ of the $R = P$ case, which we already know gives accurate recovery with $O(\log \text{Cov}_{\eta,0}(P))$ measurements.

Residual mass. There are just two main issues remaining: we want to depend on $\log \text{Cov}_{\eta,\delta}$ rather than $\log \text{Cov}_{\eta,0}$, 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}_∞ . 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 4.4.4. 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 $m \geq O(\min(\log \text{Cov}_{\eta,\delta}(R), \log \text{Cov}_{\eta,\delta}(P)))$ measurements. Given y and the fixed matrix A , let \hat{x} be the 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 ,

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

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

4.5 Lower Bound

In the previous section, we showed, for any distribution R of signals, that $O(\log \text{Cov}(R))$ 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 $\Omega(\log \text{Cov}(R))$ measurements.

Theorem 4.5.1. *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

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

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:

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

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 4.5.2. Consider the setting of Theorem (4.5.1). 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 $x^* \rightarrow y \rightarrow \hat{x}$ is a Markov chain, we can directly apply the Data Processing Inequality [67].

Lemma 4.5.3. Consider the setting of Theorem (4.5.1). If A is a deterministic matrix, we have

$$I(x^*; \hat{x}) \leq I(y; x^*).$$

If A is a random matrix, then

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

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

Lemma 4.5.4 (Fano variant). *Let (x, \hat{x}) be jointly distributed over $\mathbb{R}^n \times \mathbb{R}^n$, where $x \sim R$ and \hat{x} satisfies*

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

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

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

In order to complete the proof of Theorem 4.5.1, we need an additional counting argument to remove the extra τ term that appears in the left hand side of Lemma 4.5.4.

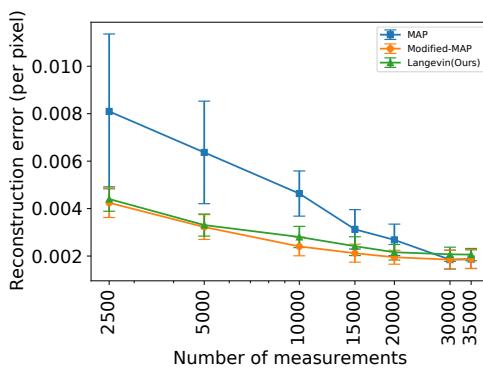
The proofs can be found in Appendix C.2.

4.6 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.

4.6.1 Datasets and Models

We perform our experiments on the CelebA-HQ [184, 144] and FlickrFaces-HQ [146] datasets. For the CelebA dataset, we run experiments using a Glow



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



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

Figure 4.4: 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.

generative model [159]. For the FlickrFaces-HQ dataset, we use the NCSNv2 model [254]. Both models have output size $256 \times 256 \times 3$. Details about our experiments are in Appendix C.3.

4.6.2 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 $G : \mathbb{R}^n \rightarrow \mathbb{R}^n$ 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 $\hat{z} \sim p(z|y)$ and setting $\hat{x} = G(\hat{z})$ is equivalent to sampling $\hat{x} \sim p(x|y)$.

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:

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

then $p(z|y)$ is the stationary distribution of z_t as $t \rightarrow \infty$ and $\alpha_t \rightarrow 0$. Unfortunately, this algorithm is slow to mix, as observed in [253]. We instead use an annealed version of the algorithm, where in step t we pretend that $p(z | y)$ has noise scale $\sigma_t \geq \sigma$ instead of σ . This gives

$$\log p_t(z|y) = \left(-\frac{\|y - AG(z)\|^2}{2\sigma_t^2/m} - \frac{\|z\|^2}{2} \right) + \log c(y), \quad (4.4)$$

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.3 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

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

Notice that we can also run MAP using this model. This can be achieved by simply following the gradient, and not adding noise:

$$x_{t+1} \leftarrow x_t + \frac{\alpha_t}{2} \nabla_x \log p(x_t|y).$$

This model also requires annealing, and we follow the schedule prescribed by [254]. Please see Appendix C.3 for more details.

4.6.3 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 [20].

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

$$\hat{z} := \arg \min_z \frac{\|y - AG(z)\|^2}{2\sigma^2/m} - \log q(z), \quad (4.5)$$

and set the estimate to be $\hat{x} = G(\hat{z})$. Typically, $q(z)$ is a standard Gaussian for many generative models. If one has access to $p(x)$, such as in NCSNv2 [252], 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, [20] introduce a parameter $\gamma > 0$ that weights the prior, and their estimate is

$$\hat{z}_{modified} := \arg \min_z \|y - AG(z)\|^2 - \gamma \log q(z), \quad (4.6)$$

Other examples of hyper-parameters include early stopping to avoid “overfitting” 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.

4.6.4 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

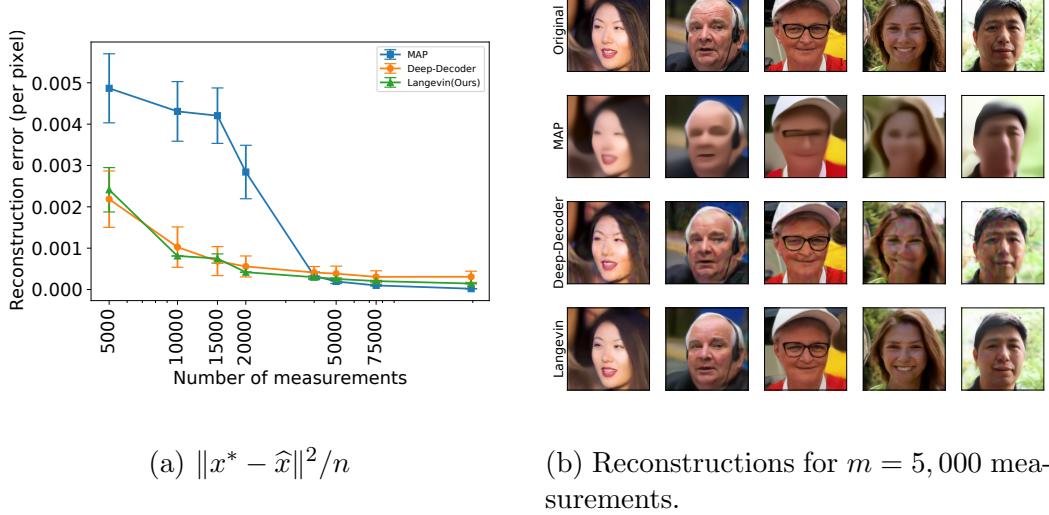


Figure 4.5: We compare our algorithm with the MAP and Deep-Decoder baselines on the FFHQ dataset, where the number of pixels is $n = 256 \times 256 \times 3 = 196,608$. Figure (a) plots per-pixel reconstruction error as we vary the number of measurements m . Figure (b) shows original images (top row), 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).

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 4.4, 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 (4.5) while Modified-MAP optimizes (4.6). The MAP baseline in Fig-

ure 4.4 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 [20], 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 [20].

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

FFHQ. In Figure 4.5, 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 [111] 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 match state-of-the-art [20].

We observe that for $m < 40,000$ measurements, Langevin dynamics beats MAP, and is competitive with Deep-Decoder. In Figure 4.1 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 4.5a, the images in Fig 4.1 produced by Langevin dynamics have better perceptual quality. Also see Fig 4.5b for more examples of reconstruc-

tions 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 4.2, 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 4.2 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.

4.7 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 $\log \text{Cov}_{\eta,\delta}(P)$ can be estimated or bounded when P is given by a neural network generative model.

Chapter 5

Fairness Aspects in the Presence of Uncertain Sensitive Attributes

5.1 Abstract

This work tackles the issue of fairness in the context of generative procedures, such as image super-resolution, which entail different definitions from the standard classification setting. Moreover, while traditional group fairness definitions are typically defined with respect to specified protected groups – camouflaging the fact that these groupings are artificial and carry historical and political motivations – we emphasize that there are no ground truth identities. For instance, should South and East Asians be viewed as a single group or separate groups? Should we consider one race as a whole or further split by gender? Choosing which groups are valid and who belongs in them is an impossible dilemma and being “fair” with respect to Asians may require being “unfair” with respect to South Asians. This motivates the introduction of definitions that allow algorithms to be *oblivious* to the relevant groupings.

We define several intuitive notions of group fairness and study their incompatibilities and trade-offs. We show that the natural extension of demographic parity is strongly dependent on the grouping, and *impossible* to

achieve obliviously. On the other hand, the conceptually new definition we introduce, Conditional Proportional Representation, can be achieved obliviously through Posterior Sampling. Our experiments validate our theoretical results and achieve fair image reconstruction using state-of-the-art generative models.

These results were published at ICML 2021 [135].

5.2 Introduction

Fairness, accountability, and transparency have taken a front-row seat in the machine learning community. Numerous recent controversies have erupted over how current machine learning systems already in use can be racist [249], sexist [149], homophobic [204], or all of the above [203]. In a recent controversy, a low-resolution image of Barack Obama was put into PULSE, a super-resolution generative model [199], but the resulting image was of a distinctly White man. While we generally have to be careful when identifying the race of a person that does not exist, such as the one represented by the generated image, multiple other reconstructions by PULSE strongly suggest that this algorithm contributes to the systemic bias against people of color.

Accuracy of representation as a fairness notion is a significant leap from the more traditional classification setting, in which we require some form of independence (or conditional independence) between the sensitive attributes and the algorithm prediction. In the context of image reconstruction, the output itself can be considered as having sensitive attributes, and we want

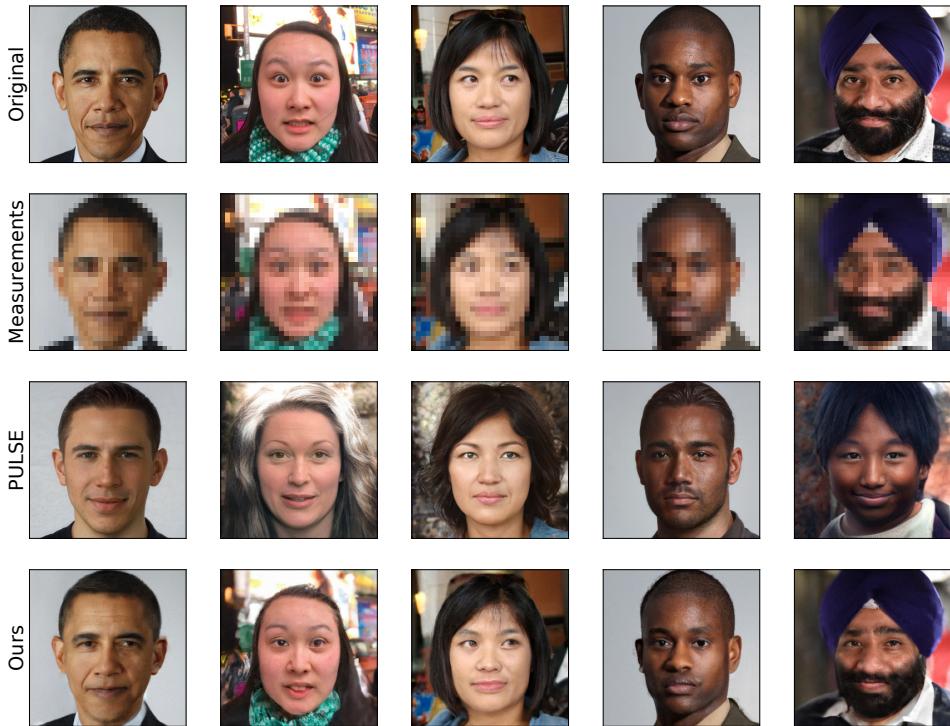


Figure 5.1: Super-resolution reconstructions on Barack Obama and four faces from the FFHQ dataset. The top row shows original images, the second row shows what the algorithms observe: blurry measurements after downsampling by $32\times$ in each dimension. The third row shows reconstructions by PULSE, and the last row shows reconstructions by Posterior Sampling via Langevin dynamics, the algorithm we are advocating for. These faces were chosen to compare performance on various ethnicities. Please see Appendix D.1 for images chosen at random from the dataset.

the sensitive attributes of the input to match the sensitive attributes of the output – which is fundamentally different from an independence condition. This leads us to introduce and discuss new fairness definitions, specific to the field of image generation, reconstruction, denoising and super-resolution.

In light of the “White Obama” controversy [199], it has been suggested

that reconstruction algorithms are biased because the datasets are not representative of the true population distribution. While it is true that the datasets are biased [43, 154], current algorithms also play their part in widening this gap [279, 261], such that majority classes get overrepresented, and minorities get further underrepresented. Indeed, when applying PULSE [199] to an unbalanced dataset with 80% dogs (majority class) and 20% cats, we observe that 80% of cats are mistakenly reconstructed as dogs, while only 2% of dogs are reconstructed as cats (see Figure 5.4b). When cats are the 80% majority, the situation reverses to 1% and 98% mistakes, respectively (see Figure 5.4d).

There is a simple intuitive reason why reconstruction algorithms designed to maximize accuracy will increase bias. Assume we observe a noisy version y of an image x^* that is either a dog or a cat. Assume cats are the minority, with the prior $\Pr(x^* \in \text{Dog}) = 0.8$. Further, assume that the measurements are always noisy and cannot definitively identify the species, so cat-like measurements are such that $p(y | x^* \in \text{Cat})/p(y | x^* \in \text{Dog}) \leq 2$.

Using Bayes, the posterior is

$$\begin{aligned} \Pr(x^* \in \text{Dog} | y) &= p(y | x^* \in \text{Dog}) \cdot \frac{\Pr(x^* \in \text{Dog})}{p(y)} \\ &\geq 1 \cdot \frac{0.8}{0.8 \cdot 1 + 0.2 \cdot 2} \\ &= 2/3. \end{aligned}$$

Therefore, regardless of the measurement, an algorithm that maximizes accuracy *will always produce images of dogs.*

This issue relates to a rich area of work on fairness in machine learning, including for classification or generation without measurements (see Section 5.2.1 for an overview). However, to the best of our knowledge, previous approaches always assume that the sensitive attributes are well-defined and unambiguous. While this assumption might hold for cats and dogs, as [38, 107] emphasize, race cannot be treated in the same way. First, it is unclear when to include subgroups within the larger group or when to treat them separately (for instance, when to consider South Asians as their own subgroup, or as Asians). This has major implications, as choosing which groups exist and what sensitive attributes are valid can already widen existing discrimination, as the long line of research on intersectionality shows. Second, even if we could decide on which groups are relevant, races are multidimensional and cannot be reduced to a simple categorical value: studies show that we can arrive at inconsistent conclusions about the same data depending on how race is measured (e.g. self-reported or observed) [119]. Our work therefore focuses on moving away from classifying people into partitions.

Problem Setting. Suppose that we have a distribution of users x^* ; each user x^* is observed through some lossy observation process to produce y (e.g., a low-resolution image); and our reconstruction algorithm produces \hat{x} from y . We are concerned about fairness with respect to a collection of protected groups $C = \{c_1, \dots, c_k\}$. Our setting therefore includes, but is not limited to,

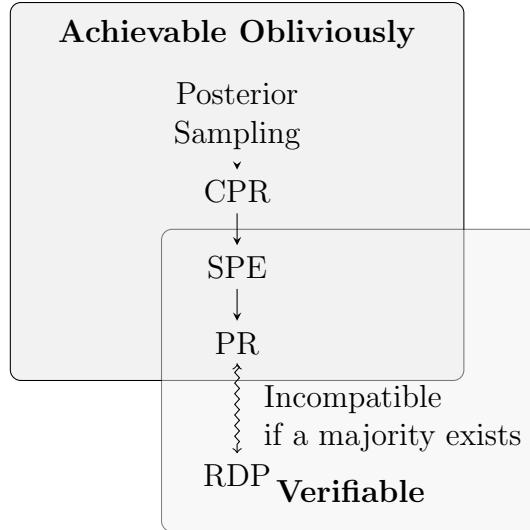


Figure 5.2: CPR, SPE and PR are achievable obliviously by Posterior Sampling. However, RDP cannot be achieved obliviously, and if a majority group exists it cannot be achieved simultaneously with PR.

the special case in which C is a partition¹.

The fairness concern we consider is that of *representation*: when users in each protected group use the algorithm, does the result adequately represent them and their group? When the observation process y is significantly lossy, there inevitably will be “representation errors” where a member of one group is reconstructed as being in a different group. How should we determine if the errors are equitable?

Our Contributions: Fairness Definitions. We introduce definitions for some natural notions of fairness in reconstruction. One is that the average

¹For simplicity of notation, each group c_i contains both *people* x^* and *images* \hat{x} .

representation rate should be independent of the group:

$$\Pr(\hat{x} \in c_i \mid x^* \in c_i) \quad (\text{RDP})$$

is the same value for all $i \in [k]$. We call this *Representation Demographic Parity* (RDP), by analogy to the binary classification setting, where Demographic Parity means that $\Pr(L = 1 \mid x^* \in c_i)$ is fixed. The difference here is that the “good” outcome ($\hat{x} \in c_i$) is different for each group, while typically in classification the “good” outcome (where, e.g., $L = 1$ means “offer a loan”) is the same across groups. RDP is simply requesting that the reconstructions have the same error rates across groups.

An alternative definition is that the demographics of the output should match those of the input:

$$\Pr(\hat{x} \in c_i) = \Pr(x^* \in c_i) \quad \forall i. \quad (\text{PR})$$

We call this *Proportional Representation* (PR). It simply says that the reconstruction process should not introduce bias in the distribution for or against any group.

Unfortunately, these two definitions are often *incompatible*. We show in Proposition 5.3.7 that, whenever a majority group exists and the measurements can confuse it with other groups, no algorithm can achieve both RDP and PR.

One weakness of both PR and RDP is that they only consider the *global* behavior of the reconstruction. But individual users want to be represented

well when they use the system, and may not be mollified by the knowledge that many other members of their group are being represented. On the other hand, some images are genuinely harder to reconstruct accurately, so expecting equal representation accuracy/RDP for every user would strongly limit overall accuracy. Our solution is to extend PR by incorporating the measurement process:

$$\Pr(\hat{x} \in c_i | y) = \Pr(x^* \in c_i | y) \quad \forall i, y. \quad (\text{CPR})$$

We call this *Conditional Proportional Representation* (CPR). The idea is that the population of users with each given y should have fair treatment (in the sense of PR). Of course, CPR implies PR by averaging over y .

Note that CPR implies that the reconstruction process must be *randomized*, not deterministic. This has other benefits: if the user is not satisfied with the result, they can rerun the algorithm until they get a result that represents them. Users can also get a collection of \hat{x}_i to observe the diversity of possible reconstructions.

Our Contributions: Algorithms. We show that CPR (and hence PR) are achievable with a simple-to-describe algorithm: posterior sampling, where we output $\hat{x} \sim p(x^* | y)$. This can be approximated well in practice using Langevin dynamics for state-of-the-art generative models representing $p(x^*)$, as we discuss in Section 5.5.1.

Posterior Sampling also satisfies one more fairness condition: the confusion matrix is symmetric, meaning that (for example) an equal number of

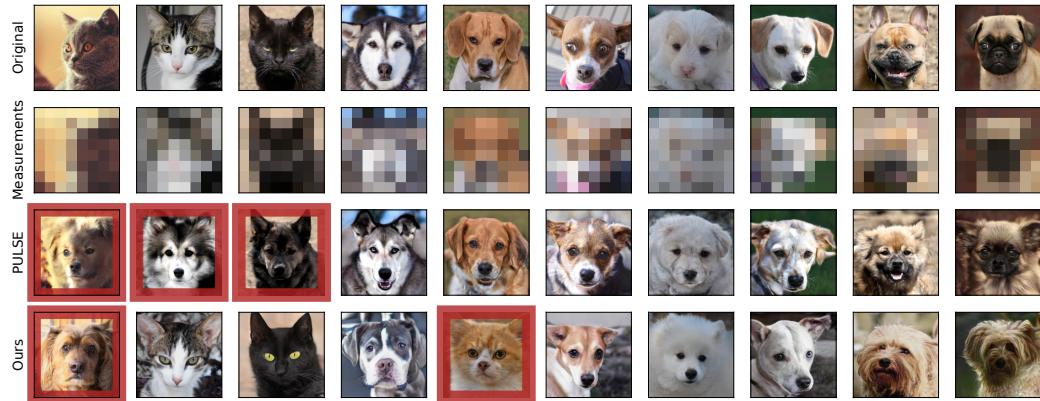


Figure 5.3: Super-resolution on the AFHQ cats & dogs dataset using StyleGAN2 trained on **20% cats** and **80% dogs**. The rows show, from top to bottom 1) original images 2) measurements after downsampling $64 \times$ in each dimension 3) reconstructions by PULSE 4) reconstructions by Posterior Sampling. The red bounding boxes denote the errors. PULSE converts almost all cats to dogs, and almost never does the reverse. Posterior Sampling makes roughly the same number of errors on cats and dogs.

Black users will be reconstructed as White as White users will be reconstructed as Black. We call this condition *Symmetric Pairwise Error* (SPE).

$$\Pr(\hat{x} \in c_i, x^* \in c_j) = \Pr(\hat{x} \in c_j, x^* \in c_i). \quad (\text{SPE})$$

for all $i, j \in [k]$. CPR implies SPE, and SPE implies PR (see Figure 5.2).

Since SPE implies PR, in general SPE is incompatible with RDP (per Proposition 5.3.7). But in the special case of two groups c_1, c_2 of equal size, then SPE actually implies RDP. This gives an algorithm to achieve RDP for the two-group setting: we reweight our input distribution such that each group has equal probability, then perform Posterior Sampling with respect to the reweighted distribution. With more than two groups, there still exists a reweighting of the groups such that Posterior Sampling on the reweighted

distribution satisfies RDP (see Theorem 5.4.5). This reweighted-resampling algorithm can be performed in practice by learning a GAN for the reweighted distribution and using Langevin dynamics on the reweighted GAN.

Our Contributions: Obliviousness. Posterior Sampling satisfies the CPR, PR, and SPE fairness criteria while retaining an invaluable property: the algorithm doesn't depend on the set of protected groups C . It satisfies the fairness properties for every set of protected groups, which is an algorithmically achievable way of addressing the issues raised in [107] about race being ambiguous and ill-defined. We say such an algorithm is *obliviously* fair. By contrast, our reweighted-resampling algorithm achieving RDP needs to know the protected groups, and would not satisfy RDP for a different collection of groups. Which fairness properties can be achieved obliviously, and under what circumstances?

Our main results here are twofold: first, Posterior Sampling is the *only* algorithm that satisfies CPR obliviously. Second, RDP *cannot* be satisfied obliviously. This impossibility applies even to obliviousness with respect to one of two plausible, socially meaningful partitions. Theorem 5.3.3 shows, for example, that you cannot satisfy RDP with respect to both {White, Asian} and {White, South Asian, East Asian} if your observations are lossy. This means that every algorithm can reasonably be viewed as unfair with respect to RDP.

Our Contributions: Experiments. We implement Posterior Sampling via Langevin dynamics, study its empirical performance and compare it to PULSE with respect to our defined metrics. We do this on the MNIST [167], FlickrFaces-HQ [146] and AFHQ cat & dog [63] datasets. We evaluate obliviousness and SPE of Posterior Sampling on the first two datasets. Using the AFHQ cat & dog dataset, we demonstrate empirically that Posterior Sampling satisfies SPE and PR over various imbalances between cats and dogs.

5.2.1 Related Work

Numerous works have attempted to tackle the issue of bias in the machine learning of images, either for data generation/reconstruction tasks or for downstream tasks such as face recognition and image quality assessment. One popular approach for dealing with bias consists of adversarially generating data or embeddings with a discriminator for different values of the sensitive attributes, yielding similar distributions for different values of the sensitive attribute [191, 288, 289, 96, 153, 237, 296]. Another approach focuses on learning explicitly the bias of the dataset, so as to remove it [154, 100, 62]. The special case of fair dimensionality reduction through principal component analysis is solved by [236]. Another research direction formulates the fairness constraints as an additional term in the loss [242]. Another approach focuses on minorities and learns their specific features [15, 97]. A related line of work improves the fairness of generative models without retraining [260], however we do not know how to use these for inverse problems.

Another relevant line of research studies fairness in the presence of uncertainty, either in the labels [160, 40, 277, 230] or in the sensitive attributes [25, 163, 50, 278]. In particular, one work studies overlapping groups [292].

Super resolution using deep learning has had remarkable success at producing accurate images. In [169], the authors provide an algorithm which performs photo-realistic super resolution using GANs. However, this model requires retraining of the GAN when the measurement operator changes. Subsequent work has overcome this hurdle. Some models independent of the forward operator include CSGM [41], OneNet [229], PULSE [199], Deep Image Prior [270] and Deep Decoder [111]. Another line of work has shown that Posterior Sampling using approximate deep generative priors is instance-optimal for compressed sensing [134].

5.3 Fairness definitions for image generation

5.3.1 Representation Demographic Parity

While multiple group fairness definitions (demographic parity, equalized odds or opportunity, calibration etc.[35, 108]) have been studied and widely accepted in the context of classification, their extension to the setting of image generation is not immediate. Here, we extend demographic parity.

Definition 5.3.1. *Let $x^* \in \mathbb{R}^n$ denote the ground truth, and P denote its distribution. Let $y \in \mathbb{R}^m$ be some measurements of x^* . For a collection $C = \{c_1, \dots, c_k\}$ of (potentially overlapping) sets, an algorithm which reconstructs*

x^* using y satisfies Representation Demographic Parity (RDP) if:

$$\forall i, j \in [k], \Pr(\hat{x} \in c_i | x^* \in c_i) = \Pr(\hat{x} \in c_j | x^* \in c_j).$$

Example 1. If $k = 2$, c_1 being all women, c_2 being all non-women, Representation Demographic Parity with respect to these two groups implies that women are as likely to be reconstructed as women as non-women are to be reconstructed as non-women.

5.3.2 Limitations of traditional group fairness definitions

Inspired by [107], we note several reasons for having fairness definitions that are more flexible with respect to the groups in the collection or partition.

Minorities are ill-defined: What constitutes a minority? Are South Asians their own subgroup, or are they assigned as Asians? The list of accepted minorities is not only inconsistent across location and purpose, but multiple levels of granularity could be equally valid. Similar concerns can be raised from the point of view of intersectionality: we might both be interested in the discrimination faced by all women, and all people of color, without wanting to erase the singular discrimination faced by women of color [43].

Races are multi-dimensional: As [234] argues, races are multi-dimensional, and these dimensions are all relevant, albeit in different settings. For instance, voting patterns are more accurately predicted based on self-identified race, while observed race is more informative when dealing with

discrimination. These differences are not minor: as [119] shows, measuring races in five different ways led to widely different interpretations of the same data.

Partitions reify the status quo: According to [107], widespread adoption of race categories participates in erasing their historical and social context [85, 250], as well as perpetuating the current system and creating new harm [148, 243].

Who chooses the partition: [30] raises concerns on who has the power to choose the partitions and what their intentions were. Historically, such partitions have done significant harm to the minorities they were supposed to protect [201, 107].

In response to these critiques, we study a novel property of fairness definitions.

Definition 5.3.2 (obliviously). *We say an algorithm satisfies a group fairness definition obliviously if the algorithm satisfies the fairness definition for any collection of sets and does not require knowledge of the collection of sets to perform reconstruction.*

Satisfying a fairness definition obliviously is one way of addressing the issues above, as it is now satisfied for all groups at the same time. This requirement may nevertheless be too strong, since most such groupings are not socially meaningful. This leads to more restricted versions of obliviousness,

ones that only hold for specific sets of collections. Unfortunately, RDP cannot be satisfied even with only two socially meaningful partitions.

Theorem 5.3.3. *Let A and B be disjoint groups (e.g., Asian and White people), and let $A_1, A_2 \subset A$ be disjoint groups that cannot be perfectly distinguished from measurements only (e.g., South Asians and East Asians). Then Representation Demographic Parity cannot be satisfied $\{\{A, B\}, \{A_1, A_2, B\}\}$ -obliviously.*

In the example stated in Theorem 5.3.3, it is impossible to be fair as defined by Representation Demographic Parity with respect to White people, South Asians, East Asians, and Asians as a whole. This holds even if we know exactly what the measurement process is, the demographics, and what the relevant groups are.

We can state this more generally:

Theorem 5.3.4 (Representation Demographic Parity cannot be satisfied obliviously). *The only way for an algorithm to satisfy Representation Demographic Parity obliviously is to achieve perfect reconstruction.*

5.3.3 Conditional Proportional Representation

An alternative fairness measure is that the distribution of the output of the algorithm should match the demographics of the input to the algorithm:

Definition 5.3.5 (Proportional Representation). *In the setting of Defini-*

tion 5.3.1, an algorithm satisfies Proportional Representation (PR) if:

$$\forall i \in [k], \Pr(\hat{x} \in c_i) = \Pr(x^* \in c_i).$$

One could also demand a much stricter fairness property, where the algorithm should satisfy PR among the population that maps to the same observation, for every possible observation:

Definition 5.3.6 (Conditional Proportional Representation). *In the setting of Definition 5.3.1, an algorithm satisfies Conditional Proportional Representation (CPR) if, almost surely over y :*

$$\forall i \in [k], \Pr(\hat{x} \in c_i | y) = \Pr(x^* \in c_i | y).$$

Intuitively, many images could yield the same lossy measurement. Because we have no way of knowing exactly from which image the measurement came, we reconstruct one at random based on how likely images in the same group are to have yielded this measurement in the first place. As such, it is “fair”: every image that could have led to the measurement gets a chance at being represented, not just the most likely. This also implies that the reconstruction cannot be deterministic. Unfortunately, while CPR can be achieved via Posterior Sampling (Theorem 5.4.1), the fact that the definition involves the posterior distribution makes it difficult to verify without full knowledge of the measurement process and the probability distribution.

It turns out that one cannot achieve RDP and PR simultaneously if you have a majority which has mass larger than $1/2$.

Proposition 5.3.7. *Whenever there exists a majority class that the measurements cannot 100% distinguish from the non-majority classes, PR and RDP are not simultaneously achievable.*

5.4 Posterior Sampling

The *Posterior Sampling* algorithm outputs a reconstruction \hat{x} drawn from the posterior $P(\cdot | y)$. It is known to be instance-optimal for compressed sensing [134] and to give fairly accurate results in practice when implemented via annealed Langevin dynamics [253, 254]. In this section, we show that it also has good fairness properties.

It is easy to see that if one has access to the distribution P over images and the likelihood function associated with the measurement process, then Posterior Sampling will satisfy the CPR. The following Theorem shows that this is the *only* algorithm that can satisfy CPR.

Theorem 5.4.1. *Posterior Sampling is the only algorithm that achieves oblivious Conditional Proportional Representation.*

Definition 5.4.2 (Symmetric Pairwise Error). *In the setting of Definition 5.3.1, an algorithm satisfies Symmetric Pairwise Error (SPE) if*

$$\Pr(\hat{x} \in c_i, x^* \in c_j) = \Pr(\hat{x} \in c_j, x^* \in c_i), \quad \forall i, j \in [k].$$

Using the fact that the ground truth and reconstruction are conditionally independent given the measurements, we can show that any algorithm that satisfies CPR will also satisfy SPE.

Theorem 5.4.3. *In the setting of Definition 5.3.1, Conditional Proportional Representation implies Symmetric Pairwise Error.*

Theorem 5.4.1 and Theorem 5.4.3 give the following Corollary.

Corollary 5.4.4. *Posterior Sampling achieves symmetric pairwise error for any pair of sets $U, V \subset \mathbb{R}^n$.*

Finally, for any partition C , there exists a reweighting of the underlying distribution such that Posterior Sampling achieves RDP with respect to the partition C .

Theorem 5.4.5. *Let $C = \{c_1, \dots, c_k\}$ be a partition. There exists a choice of weights $\lambda_i > 0$ with $\sum \lambda_i = 1$ such that Posterior Sampling with respect to the reweighted distribution*

$$p_\lambda(x) = \sum_i \lambda_i p(x \mid x \in c_i)$$

satisfies RDP with respect to C .

In the special case of 2 classes, the reweighting is very simple: $\lambda_1 = \lambda_2 = \frac{1}{2}$.

5.4.1 Representation Cross-Entropy

For the special case when the collection C is a partition, we can show that Posterior Sampling obviously minimizes a loss we call Representation

Cross-Entropy (RCE). Intuitively, one can think of this as the generative analogue of the cross-entropy loss popular in classification settings. Following the notation in Definition 5.3.1, we define RCE as:

Definition 5.4.6 (Representation Cross-Entropy). *Let $C = \{c_1, \dots, c_k\}$ form a disjoint partition of \mathbb{R}^n , and let U be a function such that $U(x)$ encodes where x lies in the partition. The Representation Cross-Entropy (RCE) of a reconstruction algorithm \mathcal{A} with respect to C is defined as*

$$RCE(\mathcal{A}) := - \mathbb{E}_{x^*, y} \log \Pr_{\hat{x}|y} [\hat{x} \in U(x^*)].$$

We show that if we want to minimize RCE over a partition, then we must have CPR on this partition:

Theorem 5.4.7. *Let $C = \{c_1, \dots, c_k\}$ form a disjoint partition of \mathbb{R}^n . An algorithm minimizes Representation Cross-Entropy on C iff the algorithm satisfies CPR on C .*

From Theorem 5.4.1, we know that Posterior Sampling is the only algorithm that can achieve CPR over all measurable sets. The same result holds if we restrict to measurable partitions, so Posterior Sampling is the only algorithm that minimizes RCE obliviously to the partition.

5.5 Experiments

So far we have discussed and analyzed properties of several different fairness metrics. In this section, we briefly describe how one can implement Posterior Sampling, and study the empirical performance of Posterior Sampling and PULSE with respect to our defined metrics, on the MNIST [167], FlickrFaces-HQ [146] and AFHQ cat&dog dataset [63].

5.5.1 Langevin Dynamics

We implement Posterior Sampling via Langevin dynamics, which states that if $x_0 \sim \mathcal{N}(0, cI_n)$, (for c appropriately small), then we can sample from $p(x|y)$ by running noisy gradient ascent:

$$x_{t+1} \leftarrow x_t + \gamma_t \nabla_{x_t} \log p(x_t|y) + \sqrt{2\gamma_t} \xi_t,$$

where $\xi_t \sim \mathcal{N}(0, I_n)$ is an i.i.d. standard Gaussian drawn at each iteration. It is well known [282, 253] that as $\gamma_t \rightarrow 0, t \rightarrow \infty$, we have $p(x_t|y) \rightarrow p(x|y)$. In practice, we need some form of annealing of the noise in order to mix efficiently. Since our algorithm is randomized, we always output the first obtained reconstruction. Please see Appendix D.4 for architecture-specific details.

5.5.2 MNIST dataset

We trained a VAE [158] on MNIST digits, and consider the groups $\{0, 1, 2, 3, 4, \geq 5\}$. As seen in the confusion matrix in Table 5.1, Posterior

	0	1	2	3	4	≥ 5
0	33	2	0	2	0	6
1	0	61	1	0	1	4
2	0	2	45	1	1	6
3	1	0	2	33	2	7
4	1	0	1	0	41	12
≥ 5	2	5	10	5	14	199

Table 5.1: Confusion matrix for super-resolution of MNIST digits after downampling by $4\times$ in each dimension. The rows denote the labels of original images, the columns denote the labels of reconstructed images. The symmetric nature of the matrix shows that Posterior Sampling achieves SPE obviously over multiple groups.

Sampling does satisfy SPE obviously over the groups (recall that SPE also implies PR).

5.5.3 FlickrFaces dataset

Dataset and generative models. We use a StyleGAN2 [147] model for PULSE, while Posterior Sampling uses the NCSNv2 generative model [253, 254]. We choose this model as it has been designed to produce images via Langevin dynamics, which is the practical implementation of Posterior Sampling.

Results In Figure 5.1, we show the results of super-resolution on Barack Obama and four faces from FFHQ, using PULSE and Posterior Sampling. As shown, Posterior Sampling preserves the image features better than PULSE. We use the CLIP classifier [224] to assign labels of {child with / without

	A	B	C	D
A	5	3	2	0
B	1	99	0	10
C	1	0	68	10
D	1	10	8	282

Table 5.2: Confusion matrix for super-resolution of FFHQ faces after $32\times$ down-sampling in each dimension. The categories are A: child with glasses, B: child without glasses, C: adult with glasses, D: adult without glasses. Rows denote labels of original images, columns denote labels of reconstructed images. The symmetric nature of the matrix shows that Posterior Sampling achieves SPE over multiple groups obviously.

glasses, adult with / without glasses}, and report the confusion matrix in Table 5.2. This shows that Posterior Sampling satisfies SPE over multiple groups obviously.

Please see Appendix D.1 for more representative samples. These correspond to images 69000-69020 in the FFHQ validation set (these were the first 20 images as we downloaded them in reverse-chronological order from the Google Drive folder).

5.5.4 AFHQ Cats and Dogs dataset

Dataset and models We trained StyleGAN2 [145] on the AFHQ cat & dog [63] training set. In order to study the effect of population bias on PULSE and Posterior Sampling, we trained three models on datasets with varying bias: (1) 20% cats and 80% dogs, (2) 80% cats and 20% dogs, and (3) 50% cats and 50% dogs.

In order to label the images generated by the GAN, we take a pre-

trained Resnet108 and retrain the last layer using labelled images from the AFHQ training set. We find that the classifier’s predictions does match the human perception of dogs and cats in general.

Posterior Sampling satisfies SPE and PR when the cats and dogs are unbalanced. For this experiment, we draw x^* from the AFHQ validation dataset, which contains 500 images of cats and 500 images of dogs. Since we want to study whether Posterior Sampling and PULSE satisfy SPE and PR, we construct the test set to match the training population of the generator. That is, for the 20% cat generator, we use 125 images of cats and all 500 images of dogs from the AFHQ dataset. Similarly, for the 80% cat generator, we use 500 images of cats and 125 images of dogs in the test set.

We then downscale the images, and vary the downscaling factor such that the observed measurements have resolution $1 \times 1, 2 \times 2, 4 \times 4, 8 \times 8$. We ran PULSE and Posterior Sampling to super-resolve the blurry measurements, and used a classifier to count how many cats and dogs were reconstructed in the wrong class. The results for the 20% cat generator are in Table 5.4a and Figure 5.4b, and the results for the 80% cat generator are in Table 5.4c and Figure 5.4d. In Figure 5.3 we show example reconstructions.

We find that PULSE consistently makes very few mistakes on the majority, and an overwhelming number of mistakes on the minority. Posterior Sampling, however, makes an approximately equal number of mistakes on each class (i.e., satisfies SPE). Equivalently for this 2-class setting, it generates cats

and dogs in proportion to their population (i.e., satisfies PR).

Posterior Sampling satisfies RDP, SPE, and PR when the cats and dogs are balanced. We use a generator trained on 50% cats and 50% dogs, and study whether Posterior Sampling and PULSE satisfy RDP, SPE, and PR in practice. In this case, we use all images of cats and dogs from the AFHQ validation set. These results are in Appendix D.2, Figure D.5. Please see Appendix D.2 for more results as we vary the training bias of the generator and test SPE for images drawn from the range of the generator.

5.6 Limitations

The fact that CPR can be satisfied obliviously is its main strength, as the subgroups one would like to protect are often not well defined or labeled in datasets. This is especially beneficial for overlooked groups that lack the power to convince an algorithm designer to cater to them. However, obliviousness can also be seen as a weakness, as it leads to symmetry in the *number* of errors in each group rather than the *fraction* of errors. For two groups, this means that the minority group will always have higher error rate than the majority.

Furthermore, the goal of CPR is to treat the members of each group equally. The philosophical stance behind this property implicitly views being “fair” as treating individuals equally, and hence representing groups in proportion to their size. However, alternative philosophical stances exist. In particular, it is at odds with the idea that historically oppressed minorities

should get particular attention [107]. One could adapt such an approach into our framework by reweighting the classes, analogous to Theorem 5.4.5, but doing so requires explicit group information.

Finally, all of the definitions we consider focus on representation but do not consider the quality of the reconstruction. If all reconstructions on minorities were of poor quality (for instance because the training set did not have enough images of this specific minority, and/or they were of poorer quality, as we know can happen [43]), the algorithm could still satisfy any of the definitions and be deemed “fair” according to it. Representation fairness is just one piece of the larger question of fairness in reconstruction.

5.7 Conclusion

In the image generation setting, fairness is related to the concept of *representation*: we assign a protected group to the output, which should match the protected group of the input. This is a stark contrast with the classification setting, in which we usually require some form of independence between the output and the sensitive attributes. We therefore introduce two notions of fairness, an extension of demographic parity called Representation Demographic Parity (RDP), and a conceptually new notion, Conditional Proportional Representation (CPR). We show that these notions are in general incompatible. Furthermore, we prove that RDP is strongly dependent on the choice of the protected groups. This is especially problematic for generating images of people, as races are usually ill-defined and/or ambiguous. CPR, however, does

not suffer from these downsides, and can even be satisfied obliviously (i.e., simultaneously for any choice of protected groups).

We prove that Posterior Sampling can achieve CPR, and is actually the only algorithm that can achieve CPR fully obliviously. We show how to experimentally implement our findings through Langevin dynamics, and our experiments exhibit the expected desirable properties.

We see our work as a first step towards better understanding ideas of fairness in the context of generating structured data – our paper deals with image generation, but the problem of generating structured data could be extended to other settings. What happens when the data is of a different type? For instance, one might want to predict pronouns in the context of text completion or generation, or provide the option to use a certain dialect.

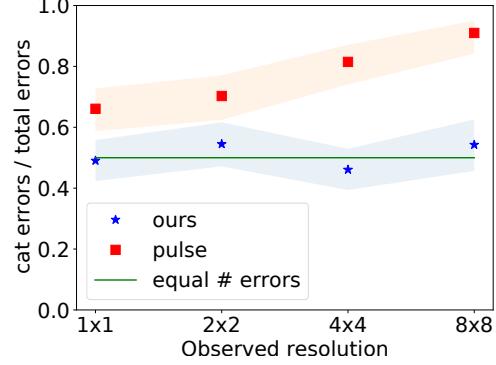
The definitions introduced in this paper are specific to generative procedures. However, the underlying issues – having definitions that do not strongly rely on the choice of the protected groups – can be found in the classification setting as well. It would be interesting to see if any analogs of CPR exist in this more traditional setting, and if there exist algorithms that can achieve it obliviously.

m	PULSE		Ours	
	Cats	Dogs	Cats	Dogs
1×1	113	58	102	106
2×2	104	44	97	81
4×4	110	25	94	110
8×8	101	10	70	59

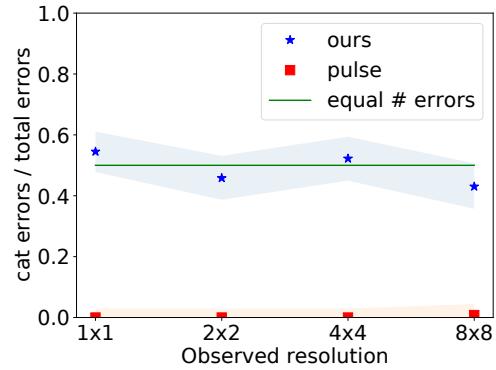
(a) Number of errors on 20% cat generator, for each resolution. Sampled test set has **125** cats and **500** dogs from the AFHQ validation set to mimick the generator's training distribution. PULSE makes errors on almost all the cats and a few dogs, while Posterior Sampling is relatively balanced.

m	PULSE		Ours	
	Cats	Dogs	Cats	Dogs
1×1	0	125	115	96
2×2	0	125	82	97
4×4	0	125	94	86
8×8	1	123	71	94

(c) Number of errors on 80% cat generator, for each resolution. Sampled test set has **500** cats and **125** dogs from the AFHQ validation set to mimick the generator's training distribution. PULSE makes errors on almost all the dogs and no cats, while Posterior Sampling is relatively balanced.



(b) Fraction of all errors on cats for 20% cat generator.



(d) Fraction of all errors on cats for 80% cat generator.

Figure 5.4: Figure (a): we use a StyleGAN2 model trained on 20% cats and report errors when reconstructing images from low-resolution measurements. The test set consists of 125 cats and 500 dogs from the AFHQ validation set to mimick the generator's training distribution (note that these correspond to all dogs in the AFHQ validation set). Figure (b) shows the proportion of all errors that are on cats, along with 95% confidence intervals from a binomial test. An algorithm that satisfies SPE would have this probability=0.5 (green line). Figure (c), (d), show analogous results when we use a StyleGAN2 generator trained on 80% cats. PULSE is clearly biased towards the majority, while Posterior Sampling via Langevin dynamics appears to satisfy SPE and PR. We remark that at 1×1 resolution, there is effectively no information and Posterior Sampling random guesses, while PULSE prefers the majority.

Chapter 6

Robust Compressed Sensing MRI

6.1 Abstract

The CSGM framework (Bora-Jalal-Price-Dimakis'17) has shown that deep generative priors can be powerful tools for solving inverse problems. However, to date this framework has been empirically successful only on certain datasets (for example, human faces and MNIST digits), and it is known to perform poorly on out-of-distribution samples. In this paper, we present the first successful application of the CSGM framework on clinical MRI data. We train a generative prior on brain scans from the fastMRI dataset, and show that posterior sampling via Langevin dynamics achieves high quality reconstructions. Furthermore, our experiments and theory show that posterior sampling is robust to changes in the ground-truth distribution and measurement process.

<https://github.com/utcsilab/csgm-mri-langevin>.

These results were published at NeurIPS 2021 [133].

6.2 Introduction

Compressed sensing [82, 44] has enabled reductions to the number of measurements needed for successful reconstruction in a variety of imaging in-

verse problems. In particular, it has led to shorter scan times for magnetic resonance imaging (MRI) [189, 273], and most MRI vendors have released products leveraging this framework to accelerate clinical workflows. Despite their successes, sparsity-based methods are limited by the achievable acceleration rates, as the sparsity assumptions are either hand-crafted or are limited to simple learned sparse codes [226, 227].

More recently, deep learning techniques have been used as powerful data-driven reconstruction methods for inverse problems [139, 211]. There are two broad families of deep learning inversion techniques [211]: end-to-end supervised and distribution-learning approaches. End-to-end supervised techniques use a training set of measured images and deploy convolutional neural networks (CNNs) and other architectures to learn the inverse mapping from measurements to image. Network architectures that include both CNN blocks and the imaging forward model have grown in popularity, as they combine deep learning with the compressed sensing optimization framework, see e.g. [102, 11, 193]. End-to-end methods are trained for specific imaging anatomy and measurement models and show excellent performance in these tasks. However, reconstruction quality is known to suffer when applied out of distribution, and recently has been shown to severely degrade [17, 71] under certain types of natural measurement and anatomy perturbations.

In this paper we study deep learning inversion techniques based on distribution learning. These models are trained without reference to measurements, and so easily adapt to changes in the measurement process. The most

common family of such techniques, known also as Compressed Sensing with Generative Models (CSGM) [41] uses pre-trained generative models as priors. Generative models are extremely powerful at representing image statistics and CSGM has been successfully applied to numerous inverse problems [41, 104] including non-linear phase retrieval [105], and improved with invertible models [20], sparsity based deviations [75], image adaptivity [124], and posterior sampling [253, 134]. These methods have only recently been applied to MRI and have not yet been shown to be competitive with supervised end-to-end methods. The very recent work [150] trains a StyleGAN for magnitude-only DICOM images but requires the presence of side-information and studies Gaussian, real-valued measurements for reconstruction. The deviation from the true MRI measurement model and the use of magnitude images are known to be problematic when evaluating performance [248]. Another work [151] trained an Invertible Neural Network on complex-valued single-coil MR images and showed very good performance in comparison to sparsity and GAN priors. Untrained and unamortized generators [111] have also been recently explored [71], showing promising results in some cases. Further, [66] studies the harder problem of learning a generative model for a class of images using only partial observations, as first proposed in AmbientGAN [42].

In this paper we train the first score-based generative model [254] for MR images. We show that we can faithfully represent MR images without any assumptions on the measurement system. As a consequence, we are able to reconstruct retrospectively under-sampled MRI data under a variety of realistic

sampling schemes. We show that our reconstruction algorithm is competitive with end-to-end supervised training when the test-data are matched to the training data and that it is robust to various out-of-distribution shifts, while in some cases end-to-end methods significantly degrade.

6.2.1 Contributions

- We successfully train a score-based deep generative model for complex-valued, T2-weighted brain MR images without any assumptions on the measurement scheme. When applied to multi-coil MRI reconstruction under the CSGM framework, we achieve competitive performance compared to end-to-end deep learning methods when the test-time data are sampled within distribution.
- We give evidence that posterior sampling should give high-quality reconstructions. First, we show that for any measurements (including the Fourier measurements in MRI) that posterior sampling with the correct prior is within constant factors of the optimal recovery method; second, even if the prior is wrong but gives α mass to the true distribution, we show that posterior sampling for Gaussian measurements is nearly optimal with just an additive $O(\log(1/\alpha))$ loss.
- We empirically show that our approach is robust to test-time distribution shifts including different sampling patterns and imaging anatomy. The former is unsurprising given that our model was trained without knowledge of the measurement scheme. As a consequence, our approach provides a degree

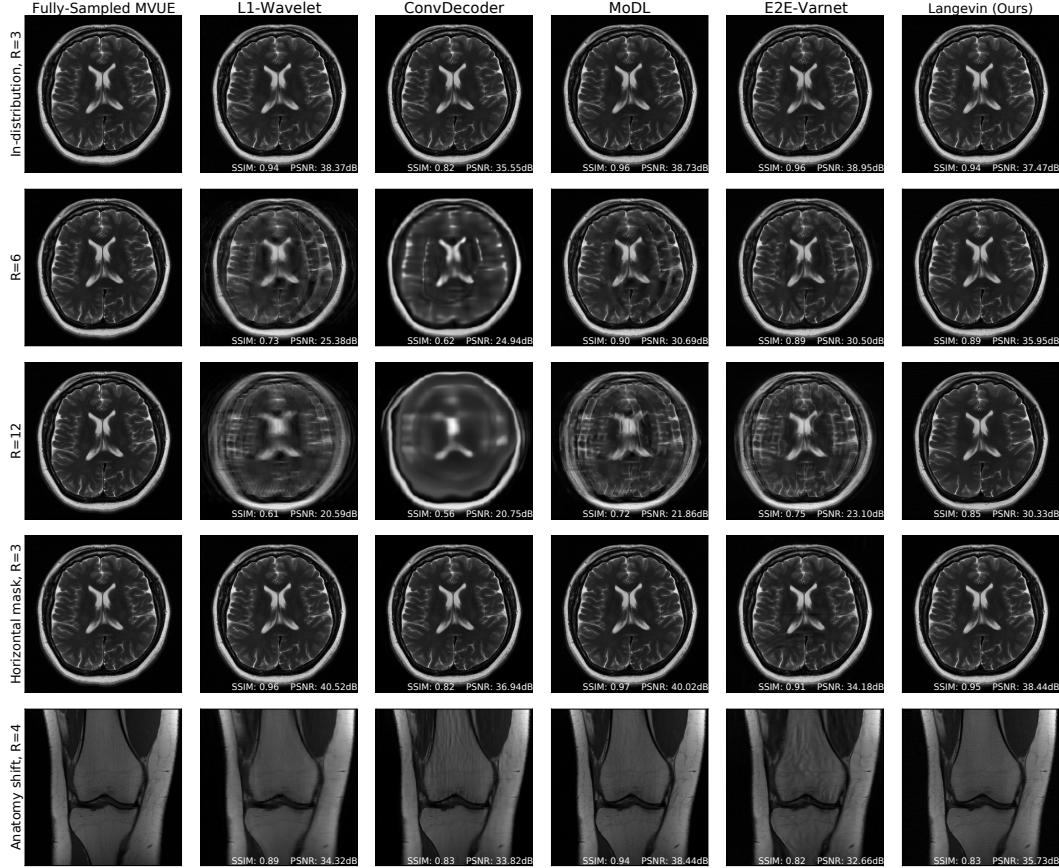


Figure 6.1: Comparison of reconstruction methods for in-distribution, sampling-shift, and anatomy-shift images. All methods and hyperparameters were optimized on T2-weighted *brain* scans with a vertical sampling mask, and tested at higher accelerations, horizontal masks, and on knee & abdomen scans. Our reconstructions are competitive with state-of-the-art methods, and introduce fewer artifacts out of distribution. All measurements are multicoil k-space from the NYU fastMRI dataset and the supervised baselines are trained from scratch on MVUE targets for a fair comparison.

of flexibility in choosing scan parameters – a common situation in routine clinical imaging. Perhaps surprisingly, the latter indicates that a specialized training set may offer sufficient regularization for a larger class of images. In

contrast, we empirically show that end-to-end methods do not always enjoy the same robustness guarantees, in some cases leading to severe degradation in reconstruction quality when applied out-of-distribution.

- Our method can be used to obtain multiple samples from the posterior by running Langevin dynamics with different random initializations. This allows us to get multiple reconstructions which can be used to obtain confidence intervals for each reconstructed voxel and visualize our reconstruction uncertainty on a voxel-by-voxel resolution. Uncertainty quantification can be incorporated into end-to-end methods, e.g., using variational auto-encoders [86], but this requires changes to the architecture. Our method does not require any modification and multiple reconstruction samplers can be run in parallel.

Our main results are succinctly summarized in Figure 6.1: we achieve equivalent reconstruction performance using a reduced training set when evaluated in-distribution and are robust when evaluated out-of-distribution.

6.2.2 Related Work

Generative priors have shown great utility to improving compressed sensing and other inverse problems, starting with [41], who generalized the theoretical framework of compressed sensing and restricted eigenvalue conditions [263, 82, 39, 44, 118, 33, 32, 87] for signals lying on the range of a deep generative model [98, 158, 256]. Lower bounds in [142, 180, 137] established

that the sample complexities in [41] are order optimal. The approach in [41] has been generalized to tackle different inverse problems [136, 105, 23, 223, 179, 192, 229, 29], and different reconstruction algorithms [75, 140, 214, 89, 88, 193, 111, 113, 70]. The complexity of optimization algorithms using generative models have been analyzed in [95, 114, 171, 106]. Our prior work shows that posterior sampling is instance-optimal for compressed sensing [134], and satisfies certain fairness guarantees without explicit information about protected sensitive groups [135].

Using compressed sensing for multi-coil MRI reconstruction has led to a rich body of work in the past two decades [189, 74, 268, 233]. See [80] and the recent special issue [131] for an overview of these methods. Classical approaches impose sparsity in a well-chosen basis, such as the wavelet domain [189], or apply shallow learning that leverages low-level redundancy in the images [226, 227, 283]. Recent research has demonstrated the superior performance of deep neural networks for MR image reconstruction [241, 102, 11, 257, 258]. A broad class of approaches is represented by end-to-end unrolled methods, which use deep networks as learned data priors in the image [11, 102, 257] or k-space domain [258]. Recent work has also investigated the performance of untrained methods [270, 113] for MR reconstruction and has shown competitive results. A much less explored line of research is MR image reconstruction with generative priors. The work in [207] proposes a CSGM-like algorithm that finetunes an entire pre-trained generator that requires a carefully tuned optimization algorithm during inference.

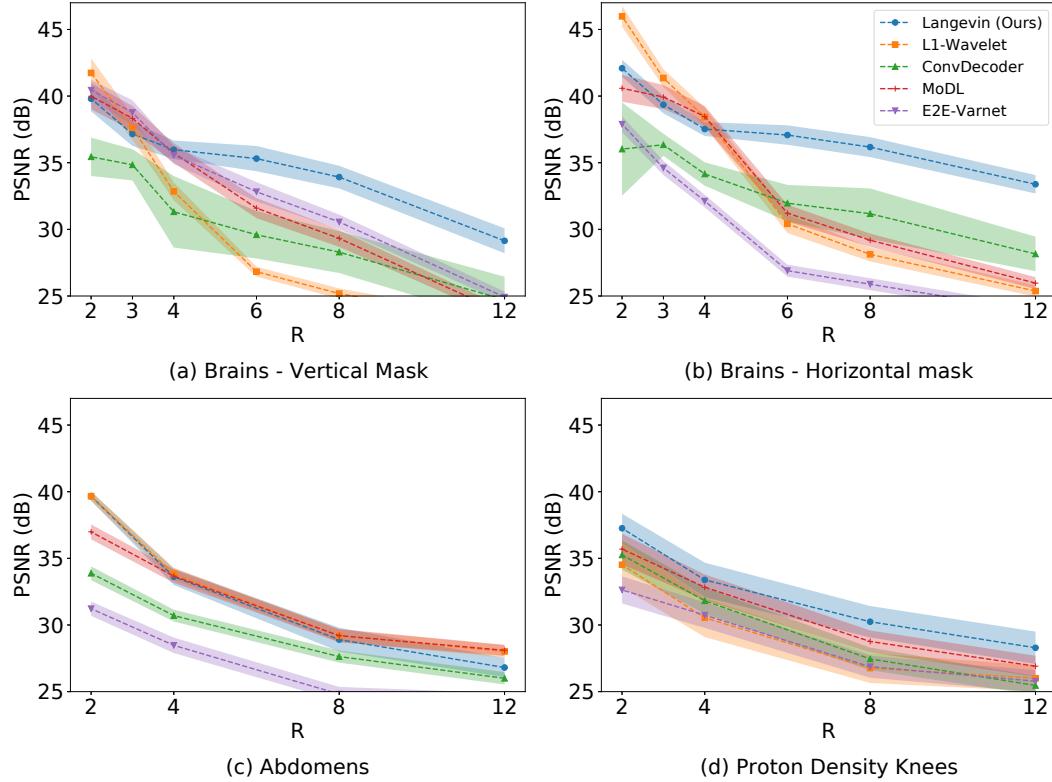


Figure 6.2: Average test PSNR in various scenarios, across a range of acceleration factors R . Higher R indicates a smaller number of acquired measurements. All methods and hyperparameters were optimized on brains with an equispaced vertical mask. Our approach mostly shows the best performance and lowest reconstruction variance both in- and out-of-distribution at test-time. Shaded regions indicate 95% confidence intervals. Note that we trained baselines on MVUE images and hence these numerical values should not be compared with those in literature trained on RSS images (see Appendix E.1.1 for a more detailed discussion).

6.3 System Model and Algorithm

6.3.1 Multi-coil Magnetic Resonance Imaging

MRI is a medical imaging modality that makes measurements using an array of radio-frequency coils placed around the body. Each coil is spatially

sensitive to a local region, and measurements are acquired directly in the spatial frequency, or *k-space*, domain. To decrease scan time, reduce operating costs, and improve patient comfort, a reduced number of k-space measurements are acquired in clinical use and reconstructed by incorporating explicit or implicit knowledge of the spatial sensitivity maps [251, 222, 99]. Formally, the vector of measurements $y_i \in \mathbb{C}^L$ acquired by the i^{th} coil can be characterized by the forward model [222]:

$$y_i = PFS_i x^* + w_i, \quad i = 1, \dots, N_c, \quad (6.1)$$

where $x^* \in \mathbb{C}^N$ is the image containing N pixels, S_i is an operator representing the point-wise multiplication of the i^{th} coil sensitivity map, F is the spatial Fourier transform operator, P represents the k-space sampling operator, and we assume $w_i \sim \mathcal{N}_c(0, \sigma^2 I)$ for simplicity. Importantly, note that the same under-sampling operator is applied to all N_c coils.

The acceleration factor R denotes the degree of under-sampling in the *k*-space domain, i.e., $R = N/L$. Due to the multiple coils, the measurements may not be compressive for small R . However, due to redundancy between the coils, the measurements are compressive for moderate values of R (even if $N_c \cdot L > N$) [122]. Also note that we use the *true acceleration factor* R , and this does not match the values in fastMRI [162]¹ on certain sampling patterns.

Given multi-coil measurements y , sensitivity maps represented by S and the sampling operator P , the goal of MR image reconstruction is to estimate

¹<https://github.com/facebookresearch/fastMRI/blob/main/fastmri/data/subsample.py>, line 247 has the fastMRI definition of equispaced acceleration factors.

the underlying image variable x^* . Prior work formulates this as a regularized optimization problem:

$$\arg \min_x \|y - Ax\|_2^2 + \lambda Q(x), \quad (6.2)$$

where we use the operator $A \in \mathbb{C}^{M \times N}$ (with $M = N_c \cdot L$) to subsume the discrete approximation to all linear effects, and Q is a suitably chosen functional prior for the image variable x . For example, to enforce a sparsity prior, one can penalize the ℓ_1 norm in the wavelet representation of x [189]. More recent approaches involve learned regularization terms parameterized by deep neural networks [241, 102, 11]. These models are typically trained *end-to-end* using a fixed training set and certain assumptions about the sampling operator. In the sequel, we present how score-based generative models can be combined with the posterior sampling [134] mechanism to reformulate (6.2) and achieve good quality reconstructions without any *a priori* assumptions about the sampling scheme.

When k-space is fully sampled at the Nyquist rate and no regularization is applied, the solution to (6.2) corresponds to the minimum-variance unbiased estimator (MVUE) of x^* , denoted by \hat{x}_{MVUE} [222]. Given fully sampled k-space data, this estimate can act as a reference image for evaluating reconstruction error as well as for end-to-end training. Alternatively, a reference image called the root-sum-of-squares (RSS) estimate can be formed by taking the inverse Fourier transform of each coil and subsequently applying the ℓ_2 norm for each pixel across the coil dimension, i.e. $\hat{x}_{\text{RSS}} = \sqrt{\sum_{i=1}^{N_c} |(F^H y_i)|^2}$, where F^H is the

Hermitian transpose of F (here the inverse DFT). Although the RSS estimate is a biased estimator, it is often used as it does not make any assumptions about the sensitivity maps, which are not explicitly measured by the MRI system. However, even if solving (6.2) results in perfect recovery of x^* , there will be a bias when comparing the result to \hat{x}_{RSS} and thus the RSS and MVUE cannot be directly compared numerically.

6.3.2 Posterior Sampling

The algorithm we consider is *posterior sampling* [134]. That is, given an observation of the form $y = Ax^* + w$, where $y \in \mathbb{C}^M$, $A \in \mathbb{C}^{M \times N}$, $w \sim \mathcal{N}_c(0, \sigma^2 I)$, and $x^* \sim \mu$, the posterior sampling recovery algorithm outputs \hat{x} according to the posterior distribution $\mu(\cdot|y)$.

In order to sample from the posterior, we use *Langevin Dynamics* [27]. Assuming we have access to $\nabla_x \log \mu(x|y)$, we can sample from $\mu(x|y)$ by running noisy gradient ascent:

$$x_{t+1} \leftarrow x_t + \eta_t \nabla_{x_t} \log \mu(x_t|y) + \sqrt{2\eta_t} \zeta_t, \quad \zeta_t \sim \mathcal{N}(0, 1). \quad (6.3)$$

Prior work [27] has shown that as $t \rightarrow \infty$ and $\eta_t \rightarrow 0$, Langevin dynamics will correctly sample from $\mu(x|y)$. In practice, vanilla Langevin Dynamics are slow to converge. Hence, the work in [253] proposes *annealed* Langevin Dynamics, where the marginal distribution of x at iteration t is modelled as $\mu_t = \mu * \mathcal{N}(0, \beta_t^2)$ and the generative model is trained to estimate the score function $f(x_t; \beta_t) := \nabla_{x_t} \log((\mu * \mathcal{N}(0, \beta_t^2))(x_t))$.

Since the distribution of $y|x^*$ is Gaussian in Eqn (6.2), we obtain $\nabla_{x_t} \log \mu(y|x_t) = \frac{A^H(y-Ax_t)}{\sigma^2}$. We find that it is also helpful to anneal this term, and we set it to $\frac{A^H(y-Ax_t)}{\sigma^2+\gamma_t^2}$, where $\gamma_t \rightarrow 0$ is a decreasing sequence. An application of Bayes' rule gives: $\nabla_{x_t} \log \mu(x_t|y) = f(x_t; \beta_t) + \frac{A^H(y-Ax_t)}{\sigma^2+\gamma_t^2}$.

Putting everything together, our final algorithm is: for $x_0 \sim \mathcal{N}_c(0, I)$ and for all $t = 0, \dots, T-1$,

$$x_{t+1} \leftarrow x_t + \eta_t \left(f(x_t; \beta_t) + \frac{A^H(y - Ax_t)}{\gamma_t^2 + \sigma^2} \right) + \sqrt{2\eta_t} \zeta_t, \quad \zeta_t \sim \mathcal{N}(0; I). \quad (6.4)$$

Note that the parameters $T, \{\beta_t\}_{t=0}^{T-1}$ were fixed during training of the generative model, and hence the only hyperparameters during inference are $\{\eta_t\}_{t=0}^{T-1}, \sigma$ and $\{\gamma_t\}_{t=0}^{T-1}$. Scripts in our codebase describe hyperparameter values used in our experiments.

6.4 Theoretical Results

Background and Notation. We first introduce background and notation required for our theoretical results. $\|\cdot\|$ refers to the ℓ_2 norm. In this section alone, for simplicity of exposition, we will assume that all matrices and vectors are real valued.

For two probability distributions μ, ν on some normed space Ω , and for any $q \geq 1$, the Wasserstein- q [275, 18] and Wasserstein- ∞ [51] distances are defined as:

$$\mathcal{W}_q(\mu, \nu) := \inf_{\gamma \in \Pi(\mu, \nu)} \left(\mathbb{E}_{(u,v) \sim \gamma} [\|u - v\|^q] \right)^{1/q}, \quad \mathcal{W}_\infty(\mu, \nu) := \inf_{\gamma \in \Pi(\mu, \nu)} \left(\gamma \text{-ess sup}_{(u,v) \in \Omega^2} \|u - v\| \right).$$

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.

The (ε, δ) -approximate covering number [134], is defined as the smallest number of ε -radius balls required to cover $1 - \delta$ mass under a distribution.

Definition 6.4.1 $((\varepsilon, \delta)$ -approximate covering number). *Let μ be a distribution on \mathbb{R}^N . For some parameters $\varepsilon > 0, \delta \in [0, 1]$, the (ε, δ) -approximate covering number of μ is defined as*

$$\text{Cov}_{\varepsilon, \delta}(\mu) := \min \left\{ k : \mu \left[\bigcup_{i=1}^k B(x_i, \varepsilon) \right] \geq 1 - \delta, x_i \in \mathbb{R}^N \right\},$$

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

Distributional robustness under Gaussian measurements. First, we consider mismatch between the ground-truth distribution, denoted by μ , and the generator distribution, denoted by ν . Prior work [134] has shown that if (i) $\mathcal{W}_q(\mu, \nu) \leq \varepsilon$ for some $q \geq 1$ and (ii) we are given $M \geq O(\log \text{Cov}_{\varepsilon, \delta}(\mu))$ Gaussian measurements, then posterior sampling with respect to ν will recover $x^* \sim \mu$ up to an error of $\varepsilon/\delta^{1/q}$ with probability $1 - \delta$. Closeness in Wasserstein distance is a reasonable assumption in certain examples, such as when μ is the distribution of celebrity faces and ν is the distribution of a generator trained on FlickrFaces [146]. However, this assumption is unsatisfactory when we

consider distributions of abdominal and brain MR scans, for example, since images of these anatomies look entirely different.

We define the following weaker notion of divergence between distributions. Informally, this new definition tells us that ν and μ are “close” if they can each be split into components which are close in \mathcal{W}_∞ distance, such that the close components contain a sufficiently large fraction under ν and μ . Formally, this is defined as:

Definition 6.4.2 $((\delta, \alpha)\text{-}\mathcal{W}_\infty$ divergence). *For two probability distributions ν and μ , and parameters $\delta, \alpha \in [0, 1]$, the (δ, α) - \mathcal{W}_∞ divergence is defined as*

$$(\delta, \alpha)\text{-}\mathcal{W}_\infty(\mu, \nu) := \inf \{\varepsilon \geq 0 : \quad$$

$$\exists \mu', \mu'', \nu', \nu'' \in \mathcal{M}(\mathbb{R}^N) \text{ s.t. } \mu = (1 - \delta)\mu' + \delta\mu'', \nu = (1 - \alpha)\nu' + \alpha\nu'', \mathcal{W}_\infty(\mu', \nu') = \varepsilon.\}$$

Lemma E.2.1 highlights that this is a strict generalization of Wasserstein distances, in the sense that closeness in Wasserstein distance implies closeness in this new divergence.

Since the $(\delta, \alpha)\text{-}\mathcal{W}_\infty$ divergence is a generalization of Wasserstein distances, it is not clear that the main Theorem in [134] holds for distributions that are close in this new divergence. The following result shows a rather surprising fact: if $(\delta, \alpha)\text{-}\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon$ then posterior sampling with $M = O\left(\log\left(\frac{1}{1-\alpha}\right) + \log \text{Cov}_{\varepsilon, \delta}(\mu)\right)$ measurements will still succeed with probability $\geq 1 - O(\delta)$.

Theorem 6.4.3. *Let $\delta, \alpha \in [0, 1]$, and $\varepsilon > 0$ be parameters. Let μ, ν be arbitrary distributions over \mathbb{R}^N satisfying $(\delta, \alpha)\text{-}\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon$. Let $x^* \sim \mu$*

and suppose $y = Ax^* + w$, where $A \in \mathbb{R}^{M \times N}$ and $w \in \mathbb{R}^M$ are i.i.d. Gaussian normalized such that $A_{ij} \sim \mathcal{N}(0, 1/M)$ and $w_i \sim \mathcal{N}(0, \sigma^2/M)$, with $\sigma \gtrsim \varepsilon$. Given y and the fixed matrix A , let \hat{x} be the output of posterior sampling with respect to ν .

Then for $M \geq O\left(\log\left(\frac{1}{1-\alpha}\right) + \min(\log \text{Cov}_{\sigma,\delta}(\mu), \log \text{Cov}_{\sigma,\delta}(\nu))\right)$, there exists a universal constant $c > 0$ such that with probability at least $1 - e^{-\Omega(M)}$ over A, w ,

$$\Pr_{x^* \sim \mu, \hat{x} \sim \nu(\cdot|y)} [\|x^* - \hat{x}\| \geq c(\varepsilon + \sigma)] \leq \delta + e^{-\Omega(M)}.$$

For our running example of ν being a generator trained on brain scans, and μ the distribution of abdominal scans, we can set ν' to be the distribution of our generator restricted to abdominal scans, and we can let μ' be the distribution restricted to “inliers” in μ . This shows that even if our generator places an *exponentially small* probability mass(i.e., $1 - \alpha \ll 1$) on the set of abdominal scans, we can still recover abdominal scans with a *polynomial additive* increase in the number of measurements (i.e., $\log(1/(1 - \alpha))$).

Near-optimality under arbitrary measurement processes. The previous result required Gaussian matrices to handle the distribution shift. Our next result shows that for an *arbitrary* measurement process, and assuming that there is no distribution shift between the generator and the ground truth distribution, posterior sampling is almost the best algorithm for this *fixed* measurement process. This result also shows that posterior sampling is good with respect to *any* metric.

Theorem 6.4.4. *Let $d(\cdot, \cdot)$ be an arbitrary metric over $\mathbb{R}^N \times \mathbb{R}^N$. Let $x^* \sim \mu$ and let $y = \mathcal{A}(x^*)$ be measurements generated from x^* for some arbitrary forward operator $\mathcal{A} : \mathbb{R}^N \rightarrow \mathbb{R}^M$. Then if there exists an algorithm that uses y as inputs and outputs x' such that*

$$d(x^*, x') \leq \varepsilon \text{ with probability } 1 - \delta,$$

then posterior sampling $\hat{x} \sim \mu(\cdot|y)$ will satisfy

$$d(x^*, \hat{x}) \leq 2\varepsilon \text{ with probability } \geq 1 - 2\delta.$$

Remark on combining these results. Our theoretical results above show that posterior sampling is (1) highly robust to distribution shift under Gaussian measurements, and (2) accurate with arbitrary measurements without distribution shift. A natural hope would be to combine these two results and show that it is robust to distribution shift under Fourier measurements. Unfortunately, this is *not* true for general distributions: for example, if μ and ν are both random distributions over Fourier-sparse signals, then Fourier measurements will usually give zero information about the signal, so cannot convince the sampler to sample near μ rather than ν .

6.5 Experimental Results

We perform retrospective under-sampling in all experiments, i.e., given fully-sampled k-space measurements from the NYU fastMRI [162, 297] and

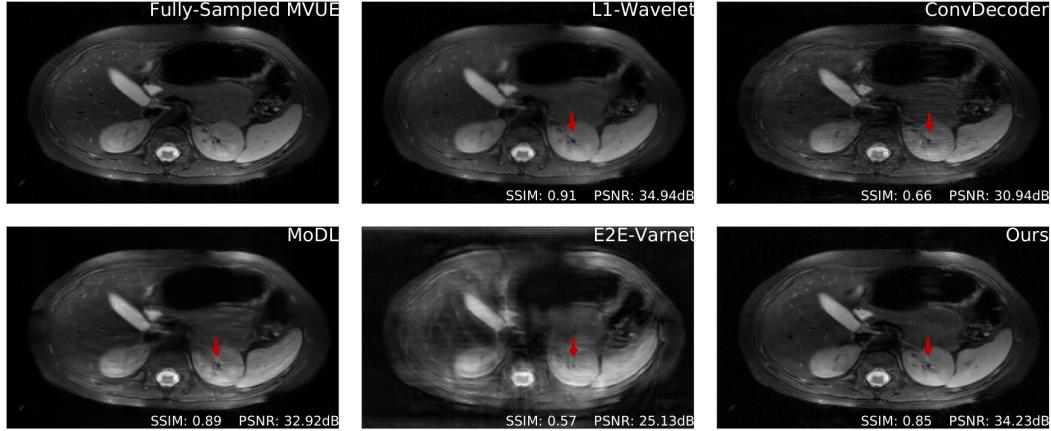


Figure 6.3: Comparative reconstructions of a 2D abdominal scan with uniform random under-sampling in the horizontal direction at $R = 4$. None of the methods were trained to reconstruct abdomen MRI. Our method uses a score-based generative model trained on brain images (as explained) and obtains good reconstructions. The red arrows indicate missing details or artifacts in the kidney structure.

Stanford MRI [3] datasets, we apply sampling masks and evaluate the performance of all considered algorithms on the reconstructed data. Depending on scan parameters (e.g., 3D scans for the Stanford knee data in Appendix E.6), we appropriately slice and sample the data in the proper dimension so as to not commit any inverse crime [101, 248].

We first highlight that an advantage of the proposed approach is the invariance to the sampling scheme during training. In contrast, this is a design choice that must be made for supervised end-to-end methods, which here were trained on equispaced, vertical sampling masks, following the fastMRI 2020 challenge guidelines [297, 206]. As our results show, this affords us a significant degree of robustness across a wide distribution of sampling masks during

inference.

We train a score-based model, NCSNv2 [254], on a small subset of scans from the NYU fastMRI brain dataset. Specifically, we train using T2-weighted images at a field strength of 3 Tesla for a total of 14,539 2D training slices. We calculate the MVUE from the fully sampled data and use the ESPIRiT algorithm [268, 130] applied to the fully-sampled central portion of k-space to estimate the sensitivity maps. The backbone network for our model is a RefineNet [174]. Since the generator’s output is expected to be complex-valued, we treat the real and imaginary parts as separate image channels. Details about the architectures are given in Appendix E.7.

We use an ℓ_1 -Wavelet regularized reconstruction algorithm [189] as a parallel imaging and compressed sensing baseline. This aims to solve the optimization problem given in (6.2) with $Q(x) = ||Wx||_1$, where W is a 2D Wavelet transform. We use the publicly available implementation from the BART toolbox [269, 267] and optimize the regularization hyper-parameter using the same subset of samples from the brain dataset that was used to train our method. We find that $\lambda = 0.01$ performs the best on the training data and use this value for all experiments. We consider three different deep learning baselines: MoDL [11], E2E-VarNet [257], and the ConvDecoder architecture [71].

We train the MoDL and E2E-VarNet baselines *from scratch* on the same training dataset as our method, at acceleration factors $R = \{3, 6\}$ and equispaced under-sampling, with a supervised SSIM loss on the magnitude

MVUE image, for 40 and 15 epochs, respectively, using a batch size of 1. For the ConvDecoder baseline, we use the architecture for brain data in [71] that outputs a complex image estimate and optimize the number of fitting iterations on a subset of samples from the training data. We find that 10000 iterations are sufficient to reach a stable average performance at $R = 3$. Put together, all of our baselines are tailored to estimate the complex image x , thus all comparisons are fair. We evaluate reconstruction performance using the complex MVUE of the fully sampled data as a reference image and measure the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [280] between the absolute values of the reconstruction and ground-truth MVUE images.

6.5.1 In-Distribution Performance

In this experiment, we test all models using the same forward model that matches the training conditions for the baselines: vertical, equispaced sampling patterns. Examples of various sampling patterns are shown in Appendix E.3.

Figure 6.1 (top three rows) shows qualitative results and Figures 6.2a & E.1a respectively show PSNR & SSIM values, for the case where there is no mismatch between the training and inference sampling patterns. As the baselines were trained to maximize SSIM at $R = 3$ & 6, we see that they achieve better SSIM scores than us at these accelerations, although there is clear aliasing in the baselines at $R = 6$. We achieve better PSNR values at

these accelerations, which supports the claim that our method does not overfit to a particular metric (Theorem 6.4.4). This also highlights the importance of qualitative evaluations in medical image reconstruction and the limitations of existing image quality metrics [194]. From the third row of Figure 6.1, and Figures 6.2a & E.1a, we notice that our method surpasses baselines at higher accelerations.

We find that ℓ_1 -Wavelet suffers both qualitatively and quantitatively at high acceleration factors, while the ConvDecoder is also a competitive architecture, but incurs a large computational cost. When benchmarked on an NVIDIA RTX 2080Ti GPU, our method takes 16 minutes and 0.95 GB of memory to reconstruct a high-resolution brain scan, whereas the ConvDecoder takes longer than 80 minutes and 6.6 GB of memory. While our method is limited by the inference time and is not in the range of end-to-end models (where reconstruction takes at most on the order of seconds and 3.5 GB of memory), multiple scans can be reconstructed in parallel due to the reduced memory footprint.

6.5.2 Out-of-Distribution Performance

Test-time sampling pattern shifts. Here we consider shifts in the forward sampling operator at test-time, while still evaluating on the same anatomy as the training conditions. We measure robustness by evaluating the average incurred performance loss when the sampling pattern changes. Recall that our proposed approach does not use any explicit information about the sampling

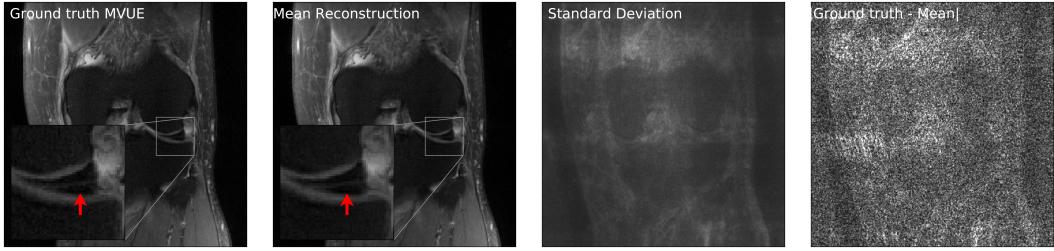


Figure 6.4: Our method successfully recovers fine details and can provide an estimate of the reconstruction error. The left column shows a knee from the fastMRI dataset, along with an annotated meniscus tear (indicated by red arrow in zoomed inset). Given measurements at an acceleration factor of $R = 4$, we obtain 48 independent reconstructions via posterior sampling. The second column shows the pixel-wise average of reconstructions, the third column shows the pixel-wise standard deviation, and the fourth column shows the magnitude of the error between the ground truth and the mean reconstruction. Note that our generative prior has never seen such pathology, as it was trained on T2-weighted brain scans.

pattern P during training, hence we anticipate the highest degree of robustness.

Figure 6.1 (fourth row) shows qualitative reconstructions when the measurements are obtained from an equispaced, horizontal sampling mask, with an acceleration factor $R = 3$. It can be observed that the reconstructions output by E2E-VarNet show aliasing artifacts. Based on the statistical results in Figure 6.2b & E.1b, our method retains its performance.

Furthermore, this experiment reveals that MoDL is more robust to this type of mask shift when compared to E2E-VarNet, even though it uses a smaller network. This is explained by the fact that E2E-VarNet does not use external sensitivity map estimates, but uses a deep neural network for end-to-end map estimation. While this improves performance on in-distribution

samples, the performance drop is strong evidence that accurate sensitivity map estimation is vital for robust generalization, and both our proposed approach and MoDL benefit from the external ESPIRiT algorithm, which is compatible with different sampling patterns.

We do note that retrospectively flipping the horizontal and vertical sampling direction is not necessarily representative of prospective sampling in the horizontal direction due to the discrete nature of the phase encoding direction in MRI, and this may contribute to the higher scores compared to the vertical mask experiments.

Test-time anatomy shifts. We now consider the more difficult problem of reconstructing different anatomies than the ones seen during. This was previously investigated in [71], which concluded that all methods suffer a drastic shift due to the various changes in scan parameters between body parts. In contrast to prior work, our main finding is that the proposed score-based model retains a significant degree of robustness under these shifts, and outputs excellent qualitative reconstructions. In some cases, some end-to-end methods retain robustness as well.

Figures 6.2c & E.1c show PSNR and SSIM scores obtained on reconstructed abdominal scans obtained from [3] at different acceleration factors. This represents both an anatomy and sampling pattern shift, and it can be seen that our method, MoDL, and the ℓ_1 -Wavelet algorithm retain their competitive advantage, while the ConvDecoder and E2E-VarNet suffer severe performance

losses. Figure 6.3 further shows a qualitative comparison of a reconstructed abdominal scan at $R = 4$, with highlighted artifacts. Appendix E.5 shows another abdomen scan.

Finally, Figures 6.2d & E.1d show PSNR and SSIM scores obtained on fastMRI knee reconstructions, while Figure 6.1 (bottom row) shows the accompanying qualitative plots. This anatomy is challenging especially because of the poor signal-to-noise ratio conditions, which can be seen even in the ground-truth image. It can be noticed that this is the most severe shift for all methods, but our approach still shows the best performance at $R = 2, 4$ and a significantly lower variance. Appendix E.4 shows more examples of knee reconstructions with and without fat suppression, and Figure E.16 shows metrics on fat suppressed knees.

6.5.3 Uncertainty Estimation

Our method can also provide uncertainty estimates for each reconstructed pixel by running multiple reconstruction samplers. For a given observation y , we can obtain independent samples $\hat{x}_1, \dots, \hat{x}_K \sim \mu(\cdot|y)$, for K sufficiently large. Now, using the conditional mean estimate $\bar{x} = \sum_{i=1}^K \hat{x}_i / K$, we can compute the pixel-wise standard deviation $\sqrt{\sum_{i=1}^K |\hat{x}_i - \bar{x}|^2 / K}$, and this gives an estimate of the error in each pixel. As shown in Fig 6.4, the pixel-wise standard deviation is a good estimate of the ground truth error $|x^* - \bar{x}|$. Additionally, notice that the reconstructions are able to recover

fine details such as the annotated meniscus tear² in Fig 6.4 and predict low uncertainty for these features.

Figure E.13 in Appendix E.4 shows another example of an annotated meniscus tear. Figures E.14 and E.15 show comparisons with baselines on the same examples.

6.5.4 Radiologist Study

We have conducted a preliminary blind assessment of overall image quality with two board-certified radiologists and one faculty member who uses neuroimaging for their research. These experts were *not* involved in our research. We have found that our algorithm was ranked best for knee scans, and tied with the baselines for abdominal and brain scans, supporting our robustness claims in the paper. For more details, please see Appendix E.8.

6.6 Limitations

We reported PSNR and SSIM values as they are correlated with radiologist evaluation upto an extent, and our preliminary radiologist study in Section 6.5.4 suggests the feasibility of clinical adoption. These metrics do not capture the needs of real-world radiologists, and a more detailed study is required before the proposed techniques can be clinically adopted.

Though promising, our initial results were still limited to fast spin-echo

²<https://discuss.fastmri.org/t/219>

imaging only and all data were retrospectively under-sampled. Further study is required to demonstrate prospective performance in a larger body of heterogeneous MRI data. Our method also currently requires a high compute cost at inference time, as well as the need for a pre-trained generative model. Clinical use requires fast reconstruction in addition to fast scanning. Future work should investigate whether score-based models can be trained without a fully-sampled training set as well as investigate approaches to reducing computation time.

Finally, there are potential issues related to discrimination. Specifically, it is possible that the quality of the reconstructed images varies across protected attributes, such as gender or race [164].

6.7 Conclusions

This paper reports the first successful application of the CSGM framework for robust multi-coil MR image reconstruction under realistic sampling conditions, and provides theoretical evidence for the robustness of posterior sampling. Our score-based model was trained on a small subset of brain MRI scans without any explicit information about the sampling scheme. This shows state-of-the-art performance under severe distributional shifts, making our model applicable in a wide range of clinical settings.

Our method shows a considerable degree of generalization to out-of-distribution samples such as abdomen and knee MRI, even when trained exclusively on brain MRI. Notably, these scans were acquired using different MRI

vendors with different pulse sequence parameters and at different institutions. We postulate that adding a small set of diverse training samples to our generative model could further improve robustness, and we hypothesize that these samples may not necessarily be restricted to MR images.

The results presented in this work represent an important step to applying deep learning models in the clinic, as there is a natural variation in sampling, image orientation, receive coils, scanner hardware, and anatomy in clinical practice.

6.8 Summary of Individual Contributions

Ajil Jalal designed the algorithmic approach, ran the experiments for the proposed algorithm, and proved the two theorems.

Marius Arvinte helped with the general formulation of the problem and was responsible for experimental parts of the paper – data pre-processing, implementing the baselines, and post-processing the results, as well as writing the paper.

Chapter 7

Future Work

The CSGM framework has inspired new research directions and several open problems at the intersection of theory and practice. Beyond developing new algorithms and proof techniques in this area, following are the topics that I am interested in:

Per-Instance Guarantees for Learning Structured Probability Distributions

Almost all our work has assumed the existence of a *good* generative model for our desired probability distribution. However, since these probability distributions live in extremely high dimensions, it is not clear why we can learn them with such few samples – the NCSNv2 model we trained on MRI brains needs only 14,000 samples to learn a probability distribution over $384 \times 384 \approx 148,000$ dimensional complex numbers! This suggests that real world distributions have structure in them that are not explained by worst-case analysis of learning probability distributions [271].

One encouraging result in this direction shows that certain generative models can be learned in polynomial time and samples [13]. The authors show that generative models possessing a property called “forward super-resolution” can be efficiently learned using Stochastic Gradient Descent-Ascent. There

are several open problems arising from this work: the authors do not take into account *convolutional* structure of real world image distributions, they assume that the probability distribution is strictly realized by a certain class of generative models, and they construct separate discriminators for learning different parameters in the generative model.

This line of thought also applies to score-based generative models [253, 254, 255], and we believe that convolutional structure explains their learnability. In [13], the authors use discriminators to learn moments of the probability distribution, and also exploit the fact that their target distribution is the push-forward of a Gaussian distribution. None of these assumptions hold true in score-based models and analyzing them is fundamentally challenging.

Closely related to this is the question of *certifying the quality of generative models*: how can we be sure that our generative model has actually learned a good approximation of the distribution? Our experimental results seem to suggest that generative models can generalize well to downstream tasks such as MRI, but a formal framework for verifying generative model quality is still an open problem.

Learning from Subsampled Data & Further MRI Applications End-to-end deep learning methods require “labelled” datasets, where the input is MRI data from \sim 1-2 minutes of scan time, while the label is the reconstruction that *would have been obtained* from \sim 5-10 minutes of scan time (note that a diagnostic exam would require \sim 5 scans, which makes the total time 30

minutes - 1 hour). This is done via retrospective undersampling – given data from a “fully-sampled” 10 minute scan, the data is subsampled to train models that can reconstruct subsampled data. This is clearly a limitation, as there are many cases [66] such as Dynamic Contrast Enhancement (DCE), real-time cardiac imaging, and 4D flow, where it is infeasible to acquire fully-sampled data and later subsample. The CSGM framework is a possible solution to this problem, as there have been cases where generative models can be learned from subsampled data [42, 66]. Unfortunately, these models do not yet give significant improvements over Lasso, and training score-based models from subsampled data is an exciting open problem.

Our work also raises a lot of questions that are directly related to MRI. In clinical settings, radiologists only care about certain small regions of the scan (such as tears in the knee meniscus) and our work raises the possibility of interactively refining the reconstruction based on a radiologist’s recommendation. As our algorithm produces randomized reconstructions from the posterior, it can also be used to provide confidence scores by quantifying its uncertainty. We are currently investigating how we can estimate these scores efficiently in theory and practice. In general, these questions all reflect theoretical questions for a broader class of inverse problems.

Adaptive Algorithms Lower bounds [143, 182] have shown that the theorems in [41] are optimal, and we cannot hope to reduce the number of measurements required for successful recovery. However, *adaptive* measurements are

extremely powerful in sparsity-based compressed sensing, sometimes allowing *exponential* reductions in the number of measurements [127]. Designing adaptive algorithms for generative priors is a fascinating open problem, and can have applications in MRI: in MRI, the scanner acquires Fourier coefficients of the target volume, and we believe the generative model can be used to decide *which* Fourier coefficients must be acquired in order to maximize efficiency. This is especially interesting in cases where the target is *dynamically changing*.

Algorithmic Guarantees Most of our work has concentrated on algorithm design and theoretical bounds on their *statistical* complexities. Follow-up work [106, 73, 244] has addressed the runtimes of our algorithms. However, these analyses require strong assumptions, such as the weights of the generative models being i.i.d. Gaussian or the existence of oracles that can project onto the range of a generative model in polynomial time. We believe this motivates the need for designing generative models that we understand better, which can lead to algorithmic guarantees under weaker assumptions on the generative model. One avenue for such research is via score-based generative models [254]. These models learn the score of probability distributions by training carefully designed denoisers. As there already exists a vast literature on solving inverse problems using denoisers [231], we believe establishing connections between these areas will provide algorithmic guarantees under weaker assumptions.

Appendices

Appendix A

Appendix for Chapter 2

A.1 Proofs

Lemma A.1.1. *Given $S \subseteq \mathbb{R}^n$, $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$, and $\gamma, \delta, \epsilon_1, \epsilon_2 > 0$, if matrix A satisfies the S -REC(S, γ, δ), then for any two $x_1, x_2 \in S$, such that $\|Ax_1 - y\| \leq \epsilon_1$ and $\|Ax_2 - y\| \leq \epsilon_2$, we have*

$$\|x_1 - x_2\| \leq \frac{\epsilon_1 + \epsilon_2 + \delta}{\gamma}.$$

Proof.

$$\begin{aligned} \|x_1 - x_2\| &\leq \frac{1}{\gamma} (\|Ax_1 - Ax_2\| + \delta), \\ &= \frac{1}{\gamma} (\|(Ax_1 - y) - (Ax_2 - y)\| + \delta), \\ &\leq \frac{1}{\gamma} (\|(Ax_1 - y)\| + \|(Ax_2 - y)\| + \delta), \\ &\leq \frac{\epsilon_1 + \epsilon_2 + \delta}{\gamma}. \end{aligned}$$

□

A.1.1 Proof of Lemma 2.5.2

Definition A.1.2. *A random variable X is said to be subgamma(σ, B) if $\forall \epsilon \geq 0$, we have*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq 2 \max \left(e^{-\epsilon^2/(2\sigma^2)}, e^{-B\epsilon/2} \right).$$

Lemma A.1.3. *Let $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be an L -Lipschitz function. Let $B^k(r)$ be the L_2 -ball in \mathbb{R}^k with radius r , $S = G(B^k(r))$, and M be a δ/L -net on $B^k(r)$ such that $|M| \leq k \log \left(\frac{4Lr}{\delta} \right)$. Let A be a $\mathbb{R}^{m \times n}$ random matrix with IID Gaussian*

entries with zero mean and variance $1/m$. If

$$m = \Omega\left(k \log \frac{Lr}{\delta}\right),$$

then for any $x \in S$, if $x' = \arg \min_{\hat{x} \in G(M)} \|x - \hat{x}\|$, we have $\|A(x - x')\| = \mathcal{O}(\delta)$ with probability $1 - e^{-\Omega(m)}$.

Note that for any given point x' in S , if we try to find its nearest neighbor of that point in an δ -net on S , then the difference between the two is at most the δ . In words, this lemma says that even if we consider measurements made on these points, *i.e.* a linear projection using a random matrix A , then as long as there are enough measurements, the difference between measurements is of the same order δ . If the point x' was in the net, then this can be easily achieved by Johnson-Lindenstrauss Lemma. But to argue that this is true for all x' in S , which can be an uncountably large set, we construct a chain of nets on S . We now present the formal proof.

Proof. Observe that $\frac{\|Ax\|^2}{\|x\|^2}$ is subgamma $\left(\frac{1}{\sqrt{m}}, \frac{1}{m}\right)$. Thus, for any $f > 0$,

$$\epsilon \geq 2 + \frac{4}{m} \log \frac{2}{f} \geq \max\left(\sqrt{\frac{2}{m} \log \frac{2}{f}}, \frac{2}{m} \log \frac{2}{f}\right)$$

is sufficient to ensure that

$$\mathbb{P}(\|Ax\| \geq (1 + \epsilon)\|x\|) \leq f.$$

Now, let $M = M_0 \subseteq M_1 \subseteq M_2, \dots \subseteq M_l$ be a chain of epsilon nets of $B^k(r)$ such that M_i is a δ_i/L -net and $\delta_i = \delta_0/2^i$, with $\delta_0 = \delta$. We know that

there exist nets such that

$$\log |M_i| \leq k \log \left(\frac{4Lr}{\delta_i} \right) \leq ik + k \log \left(\frac{4Lr}{\delta_0} \right).$$

Let $N_i = G(M_i)$. Then due to Lipschitzness of G , N_i 's form a chain of epsilon nets such that N_i is a δ_i -net of $S = G(B^k(r))$, with $|N_i| = |M_i|$.

For $i \in \{0, 1, 2 \dots, l-1\}$, let

$$T_i = \{x_{i+1} - x_i \mid x_{i+1} \in N_{i+1}, x_i \in N_i\}.$$

Thus,

$$\begin{aligned} |T_i| &\leq |N_{i+1}| |N_i|. \\ \implies \log |T_i| &\leq \log |N_{i+1}| + \log |N_i|, \\ &\leq (2i+1)k + 2k \log \left(\frac{4Lr}{\delta_0} \right), \\ &\leq 3ik + 2k \log \left(\frac{4Lr}{\delta_0} \right). \end{aligned}$$

Now assume $m = 3k \log \left(\frac{4Lr}{\delta_0} \right)$,

$$\log(f_i) = -(m + 4ik),$$

and

$$\begin{aligned} \epsilon_i &= 2 + \frac{4}{m} \log \frac{2}{f_i}, \\ &= 2 + \frac{4}{m} \log 2 + 4 + \frac{16ik}{m}, \\ &= O(1) + \frac{16ik}{m}. \end{aligned}$$

By choice of f_i and ϵ_i , we have $\forall i \in [l-1], \forall t \in T_i$,

$$\mathbb{P}(\|At\| > (1 + \epsilon_i)\|t\|) \leq f_i.$$

Thus by union bound, we have

$$\mathbb{P}(\|At\| \leq (1 + \epsilon_i)\|t\|, \forall i, \forall t \in T_i) \geq 1 - \sum_{i=0}^{l-1} |T_i|f_i.$$

Now,

$$\begin{aligned} \log(|T_i|f_i) &= \log(|T_i|) + \log(f_i), \\ &\leq -k \log\left(\frac{4Lr}{\delta_0}\right) - ik, \\ &= -m/3 - ik. \\ \implies \sum_{i=0}^{l-1} |T_i|f_i &\leq e^{-m/3} \sum_{i=0}^{l-1} e^{-ik}, \\ &\leq e^{-m/3} \left(\frac{1}{1 - e^{-1}}\right), \\ &\leq 2e^{-m/3}. \end{aligned}$$

Observe that for any $x \in S$, we can write

$$\begin{aligned} x &= x_0 + (x_1 - x_0) + (x_2 - x_1) \dots (x_l - x_{l-1}) + x^f. \\ x - x_0 &= \sum_{i=0}^{l-1} (x_{i+1} - x_i) + x^f. \end{aligned}$$

where $x_i \in N_i$ and $x_f = x - x_l$.

Since each $x_{i+1} - x_i \in T_i$, with probability at least $1 - 2e^{-m/3}$, we have

$$\begin{aligned} \sum_{i=0}^{l-1} \|A(x_{i+1} - x_i)\| &= \sum_{i=0}^{l-1} (1 + \epsilon_i) \|(x_{i+1} - x_i)\|, \\ &\leq \sum_{i=0}^{l-1} (1 + \epsilon_i) \delta_i, \\ &= \delta_0 \sum_{i=0}^{l-1} \frac{1}{2^i} \left(O(1) + \frac{16ik}{m} \right), \\ &= O(\delta_0) + \delta_0 \frac{16k}{m} \sum_{i=0}^{l-1} \left(\frac{i}{2^i} \right), \\ &= O(\delta_0). \end{aligned}$$

Now, $\|x^f\| = \|x - x_l\| \leq d_l = \frac{\delta_0}{2^l}$, and $\|x_{i+1} - x_i\| \leq \delta_i$ due to properties of epsilon-nets. We know that $\|A\| \leq 2 + \sqrt{n/m}$ with probability at least $1 - 2e^{-m/2}$ (Corollary 5.35 [274]). By setting $l = \log(n)$, we get that, $\|A\| \|x^f\| \leq \left(2 + \sqrt{\frac{n}{m}}\right) \frac{\delta_0}{2^l} = O(\delta_0)$ with probability $\geq 1 - 2e^{-m/2}$.

Combining these two results, and noting that it is possible to choose $x' = x_0$, we get that with probability $1 - e^{-\Omega(m)}$,

$$\begin{aligned} \|A(x - x')\| &= \|A(x - x_0)\|, \\ &\leq \sum_{i=0}^{l-1} \|A(x_{i+1} - x_i)\| + \|Ax^f\|, \\ &= \mathcal{O}(\delta_0) + \|A\| \|x^f\|, \\ &= \mathcal{O}(\delta). \end{aligned}$$

□

Lemma. Let $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be L -Lipschitz. Let

$$B^k(r) = \{z \mid z \in \mathbb{R}^k, \|z\| \leq r\}$$

be an L_2 -norm ball in \mathbb{R}^k . For $\alpha < 1$, if

$$m = \Omega\left(\frac{k}{\alpha^2} \log \frac{Lr}{\delta}\right),$$

then a random matrix $A \in \mathbb{R}^{m \times n}$ with IID entries such that $A_{ij} \sim \mathcal{N}(0, \frac{1}{m})$ satisfies the S-REC($G(B^k(r))$, $1 - \alpha$, δ) with $1 - e^{-\Omega(\alpha^2 m)}$ probability.

Proof. We construct a $\frac{\delta}{L}$ -net, N , on $B^k(r)$. There exists a net such that

$$\log |N| \leq k \log \left(\frac{4Lr}{\delta} \right).$$

Since N is a $\frac{\delta}{L}$ -cover of $B^k(r)$, due to the L -Lipschitz property of $G(\cdot)$, we get that $G(N)$ is a δ -cover of $G(B^k(r))$.

Let T denote the pairwise differences between the elements in $G(N)$, *i.e.*,

$$T = \{G(z_1) - G(z_2) \mid z_1, z_2 \in N\}.$$

Then,

$$\begin{aligned} |T| &\leq |N|^2, \\ \implies \log |T| &\leq 2 \log |N|, \\ &\leq 2k \log \left(\frac{4Lr}{\delta} \right). \end{aligned}$$

For any $z, z' \in B^k$, $\exists z_1, z_2 \in N$, such that $G(z_1), G(z_2)$ are δ -close to $G(z)$ and $G(z')$ respectively. Thus, by triangle inequality,

$$\begin{aligned} \|G(z) - G(z')\| &\leq \|G(z) - G(z_1)\| + \\ &\quad \|G(z_1) - G(z_2)\| + \\ &\quad \|G(z_2) - G(z')\|, \\ &\leq \|G(z_1) - G(z_2)\| + 2\delta. \end{aligned}$$

Again by triangle inequality,

$$\begin{aligned} \|AG(z_1) - AG(z_2)\| &\leq \|AG(z_1) - AG(z)\| + \\ &\quad \|AG(z) - AG(z')\| + \\ &\quad \|AG(z') - AG(z_2)\|. \end{aligned}$$

Now, by Lemma A.1.3, with probability $1 - e^{-\Omega(m)}$, $\|AG(z_1) - AG(z)\| = \mathcal{O}(\delta)$, and $\|AG(z') - AG(z_2)\| = \mathcal{O}(\delta)$. Thus,

$$\|AG(z_1) - AG(z_2)\| \leq \|AG(z) - AG(z')\| + \mathcal{O}(\delta).$$

By the Johnson-Lindenstrauss Lemma, for a fixed $x \in \mathbb{R}^n$, $\mathbb{P}[\|Ax\|^2 < (1 - \alpha)\|x\|^2] < \exp(-\alpha^2 m)$. Therefore, we can union bound over all vectors in T to get

$$\mathbb{P}(\|Ax\|^2 \geq (1 - \alpha)\|x\|^2, \forall x \in T) \geq 1 - e^{-\Omega(\alpha^2 m)}.$$

Since $\alpha < 1$, and $z_1, z_2 \in N$, $G(z_1) - G(z_2) \in T$, we have

$$\begin{aligned} (1 - \alpha)\|G(z_1) - G(z_2)\| &\leq \sqrt{1 - \alpha}\|G(z_1) - G(z_2)\|, \\ &\leq \|AG(z_1) - AG(z_2)\|. \end{aligned}$$

Combining the three results above we get that with probability $1 - e^{-\Omega(\alpha^2 m)}$,

$$\begin{aligned} (1 - \alpha)\|G(z) - G(z')\| &\leq (1 - \alpha)\|G(z_1) - G(z_2)\| + \mathcal{O}(\delta), \\ &\leq \|AG(z_1) - AG(z_2)\| + \mathcal{O}(\delta), \\ &\leq \|AG(z) - AG(z')\| + \mathcal{O}(\delta). \end{aligned}$$

Thus, A satisfies $S\text{-REC}(S, 1 - \alpha, \delta)$ with probability $1 - e^{-\Omega(\alpha^2 m)}$.

□

A.1.2 Proof of Lemma 2.5.3

Lemma A.1.4. *Consider c different $k - 1$ dimensional hyperplanes in \mathbb{R}^k . Consider the k -dimensional faces (hereafter called k -faces) generated by the hyperplanes, i.e. the elements in the partition of \mathbb{R}^k such that relative to each hyperplane, all points inside a partition are on the same side. Then, the number of k -faces is $\mathcal{O}(c^k)$.*

Proof. Proof is by induction, and follows [195].

Let $f(c, k)$ denote the number of k -faces generated in \mathbb{R}^k by c different $(k - 1)$ -dimensional hyperplanes. As a base case, let $k = 1$. Then $(k - 1)$ -dimensional hyperplanes are just points on a line. c points partition \mathbb{R} into $c + 1$ pieces. This gives $f(c, 1) = \mathcal{O}(c)$.

Now, assuming that $f(c, k - 1) = \mathcal{O}(c^{k-1})$ is true, we need to show $f(c, k) = \mathcal{O}(c^k)$. Assume we have $(c-1)$ different hyperplanes $H = \{h_1, h_2, \dots, h_{c-1}\} \subset$

\mathbb{R}^k , and a new hyperplane h_c is added. h_c intersects H at $(c - 1)$ different $(k - 2)$ -faces given by $F = \{f_j \mid f_j = h_j \cap h_c, 1 \leq j \leq (c - 1)\}$. The $(k - 2)$ -faces in F partition h_c into $f(c - 1, k - 1)$ different $(k - 1)$ -faces. Additionally, each $(k - 1)$ -face in h_c divides an existing k -face into two. Hence the number of new k -faces introduced by the addition of h_c is $f(c - 1, k - 1)$. This gives the recursion

$$\begin{aligned} f(c, k) &= f(c - 1, k) + f(c - 1, k - 1), \\ &= f(c - 1, k) + \mathcal{O}(c^{k-1}), \\ &= \mathcal{O}(c^k). \end{aligned}$$

□

Lemma. *Let $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be a d -layer neural network, where each layer is a linear transformation followed by a pointwise non-linearity. Suppose there are at most c nodes per layer, and the non-linearities are piecewise linear with at most two pieces, and let*

$$m = \Omega\left(\frac{1}{\alpha^2}kd\log c\right)$$

for some $\alpha < 1$. Then a random matrix $A \in \mathbb{R}^{m \times n}$ with IID entries $A_{ij} \sim \mathcal{N}(0, \frac{1}{m})$ satisfies the S-REC($G(\mathbb{R}^k)$, $1 - \alpha, 0$) with $1 - e^{-\Omega(\alpha^2 m)}$ probability.

Proof. Consider the first layer of G . Each node in this layer can be represented as a hyperplane in \mathbb{R}^k , where the points on the hyperplane are those where the input to the node switches from one linear piece to the other. Since there are

at most c nodes in this layer, by Lemma A.1.4, the input space is partitioned by at most c different hyperplanes, into $\mathcal{O}(c^k)$ k -faces. Applying this over the d layers of G , we get that the input space \mathbb{R}^k is partitioned into at most c^{kd} sets.

Recall that the non-linearities are piecewise linear, and the partition boundaries were made precisely at those points where the non-linearities change from one piece to another. This means that within each set of the input partition, the output is a linear function of the inputs. Thus $G(\mathbb{R}^k)$ is a union of c^{kd} different k -faces in \mathbb{R}^n .

We now use an oblivious subspace embedding to bound the number of measurements required to embed the range of $G(\cdot)$. For a single k -face $S \subseteq \mathbb{R}^n$, a random matrix $A \in \mathbb{R}^{m \times n}$ with IID entries such that $A_{ij} \sim \mathcal{N}(0, \frac{1}{m})$ satisfies $S\text{-REC}(S, 1 - \alpha, 0)$ with probability $1 - e^{-\Omega(\alpha^2 m)}$ if $m = \Omega(k/\alpha^2)$.

Since the range of $G(\cdot)$ is a union of c^{kd} different k -faces, we can union bound over all of them, such that A satisfies the $S\text{-REC}(G(\mathbb{R}^k), 1 - \alpha, 0)$ with probability $1 - c^{kd}e^{-\Omega(\alpha^2 m)}$. Thus, we get that A satisfies the $S\text{-REC}(G(\mathbb{R}^k), 1 - \alpha, 0)$ with probability $1 - e^{-\Omega(\alpha^2 m)}$ if

$$m = \Omega\left(\frac{kd \log c}{\alpha^2}\right).$$

□

A.1.3 Proof of Lemma 2.5.4

Lemma. Let $A \in \mathbb{R}^{m \times n}$ be drawn from a distribution that (1) satisfies the $S\text{-REC}(S, \gamma, \delta)$ with probability $1 - p$ and (2) has for every fixed $x \in \mathbb{R}^n$, $\|Ax\| \leq 2\|x\|$ with probability $1 - p$. For any $x^* \in \mathbb{R}^n$ and noise η , let $y = Ax^* + \eta$. Let \hat{x} approximately minimize $\|y - Ax\|$ over $x \in S$, i.e.,

$$\|y - A\hat{x}\| \leq \min_{x \in S} \|y - Ax\| + \varepsilon.$$

Then

$$\|\hat{x} - x^*\| \leq \left(\frac{4}{\gamma} + 1\right) \min_{x \in S} \|x^* - x\| + \frac{1}{\gamma} (2\|\eta\| + \varepsilon + \delta)$$

with probability $1 - 2p$.

Proof. Let $\bar{x} = \arg \min_{x \in S} \|x^* - x\|$. Then we have by Lemma A.1.1 and the hypothesis on \hat{x} that

$$\begin{aligned} \|\bar{x} - \hat{x}\| &\leq \frac{\|A\bar{x} - y\| + \|A\hat{x} - y\| + \delta}{\gamma}, \\ &\leq \frac{2\|A\bar{x} - y\| + \varepsilon + \delta}{\gamma}, \\ &\leq \frac{2\|A(\bar{x} - x^*)\| + 2\|\eta\| + \varepsilon + \delta}{\gamma}, \end{aligned}$$

as long as A satisfies the S-REC, as happens with probability $1 - p$. Now, since \bar{x} and x^* are independent of A , by assumption we also have $\|A(\bar{x} - x^*)\| \leq 2\|\bar{x} - x^*\|$ with probability $1 - p$. Therefore

$$\|x^* - \hat{x}\| \leq \|\bar{x} - x^*\| + \frac{4\|\bar{x} - x^*\| + 2\|\eta\| + \varepsilon + \delta}{\gamma}$$

as desired. \square

A.1.4 Lipschitzness of Neural Networks

Lemma A.1.5. *Consider any two functions f and g . If f is L_f -Lipschitz and g is L_g -Lipschitz, then their composition $f \circ g$ is $L_f L_g$ -Lipschitz.*

Proof. For any two x_1, x_2 ,

$$\begin{aligned} \|f(g(x_1)) - f(g(x_2))\| &\leq L_f \|g(x_1) - g(x_2)\|, \\ &\leq L_f L_g \|x_1 - x_2\|. \end{aligned}$$

□

Lemma A.1.6. *If G is a d -layer neural network with at most c nodes per layer, all weights $\leq w_{\max}$ in absolute value, and M -Lipschitz non-linearity after each layer, then $G(\cdot)$ is L -Lipschitz with $L = (Mcw_{\max})^d$.*

Proof. Consider any linear layer with input x , weight matrix W and bias vector b . Thus, $f(x) = Wx + b$. Now for any two x_1, x_2 ,

$$\begin{aligned} \|f(x_1) - f(x_2)\| &= \|Wx_1 + b - Wx_2 + b\|, \\ &= \|W(x_1 - x_2)\|, \\ &\leq \|W\| \|(x_1 - x_2)\|, \\ &\leq cw_{\max} \|(x_1 - x_2)\|. \end{aligned}$$

Let $f_i(\cdot), i \in [d]$ denote the function for the i -th layer in G . Since each layer is a composition of a linear function and a non-linearity, by Lemma A.1.5, have that f_i is Mcw_{\max} -Lipschitz.

Since $G = f_1 \circ f_2 \circ \dots \circ f_d$, by repeated application of Lemma A.1.5, we get that G is L -Lipschitz with $L = (Mcw_{\max})^d$. \square

A.2 Additional Experiments

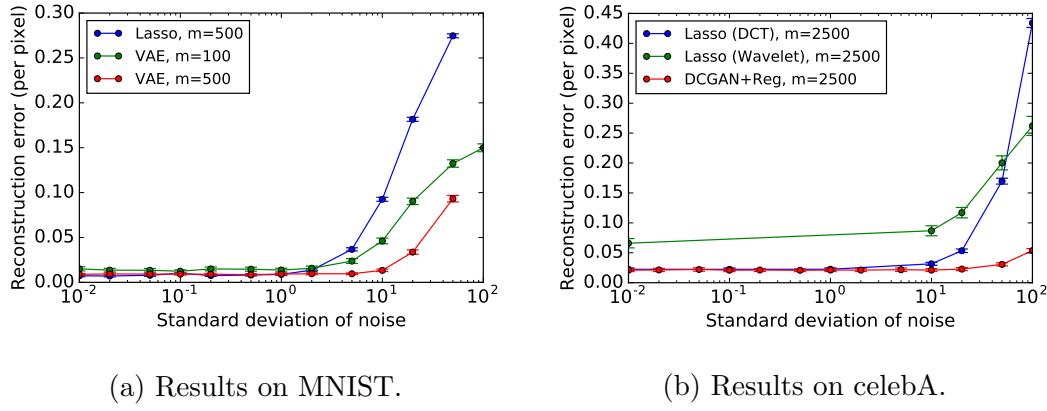


Figure A.1: Noise tolerance. We show a plot of per pixel reconstruction error as we vary the noise level ($\sqrt{\mathbb{E}[\|\eta\|^2]}$). The vertical bars indicate 95% confidence intervals.

A.2.1 Noise tolerance

To understand the noise tolerance of our algorithm, we do the following experiment: First we fix the number of measurements so that Lasso does as well as our algorithm. From Fig. 2.1a, and Fig. 2.1b we see that this point is at $m = 500$ for MNIST and $m = 2500$ for celebA. Now, we look at the performance as the noise level increases. Hyperparameters are kept fixed as we change the noise level for both Lasso and for our algorithm.

In Fig. A.1a, we show the results on the MNIST dataset. In Fig. A.1a, we show the results on celebA dataset. We observe that our algorithm has more noise tolerance than Lasso.

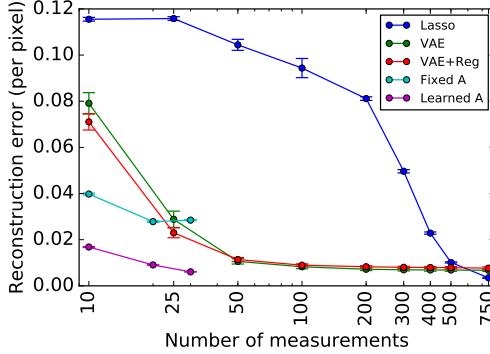


Figure A.2: Results on End to End model on MNIST. We show per pixel reconstruction error vs number of measurements. ‘Fixed A’ and ‘Learned A’ are two end to end models. The end to end models get noiseless measurements, while the other models get noisy ones. The vertical bars indicate 95% confidence intervals.

A.2.2 Other models

A.2.2.1 End to end training on MNIST

Instead of using a generative model to reconstruct the image, another approach is to learn from scratch a mapping that takes the measurements and outputs the original image. A major drawback of this approach is that it necessitates learning a new network if get a different set of measurements.

If we use a random matrix for every new image, the input to the network is essentially noise, and the network does not learn at all. Instead we are forced to use a fixed measurement matrix. We explore two approaches. First is to randomly sample and fix the measurement matrix and learn the rest of the mapping. In the second approach, we jointly optimize the measurement matrix as well.

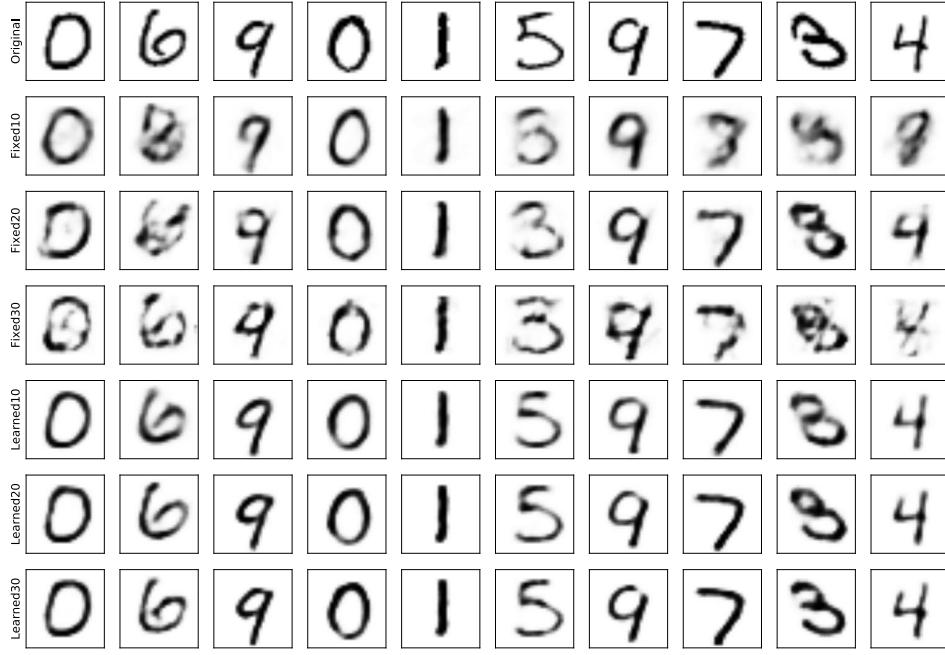
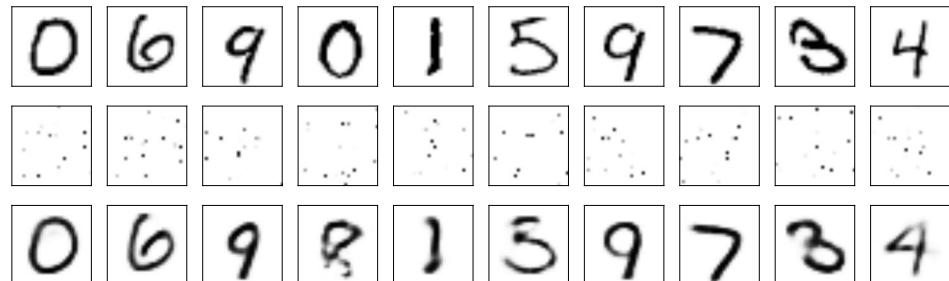


Figure A.3: MNIST End to end learned model. Top row are original images. The next three are recovered by model with fixed random A , with 10, 20 and 30 measurements. Bottom three rows are with learned A and 10, 20 and 30 measurements.

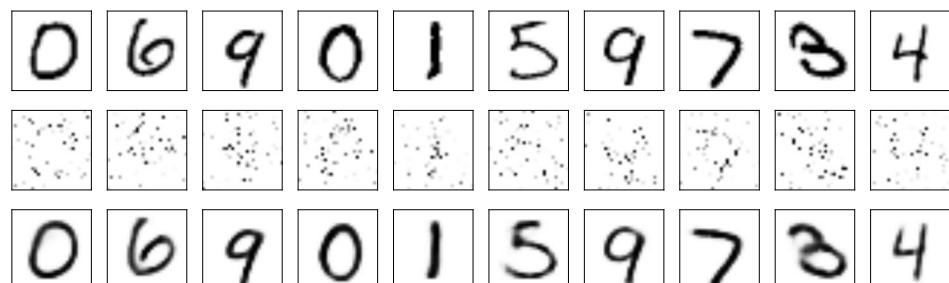
We do this for 10, 20 and 30 measurements for the MNIST dataset. We did not use additive noise. The reconstruction errors are shown in Fig. A.2. The reconstructions can be seen in Fig. A.3.

A.2.3 More results

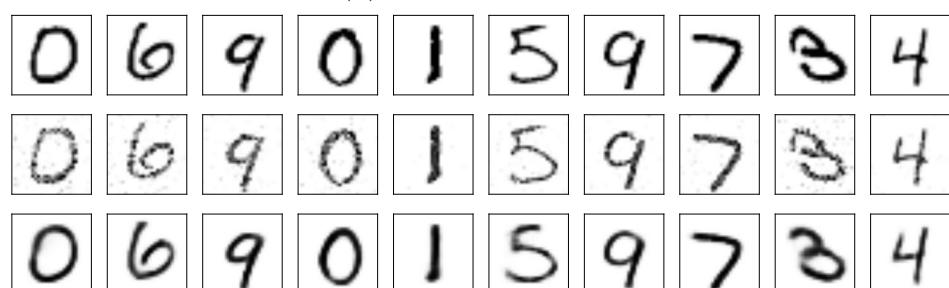
Here, we show more results on the reconstruction task, with varying number of measurements on both MNIST and celebA. Fig. A.4 shows reconstructions on MNIST with 25, 100 and 400 measurements. Fig. A.5, Fig. A.6 and Fig. A.7 show results on celebA dataset.



(a) 25 measurements

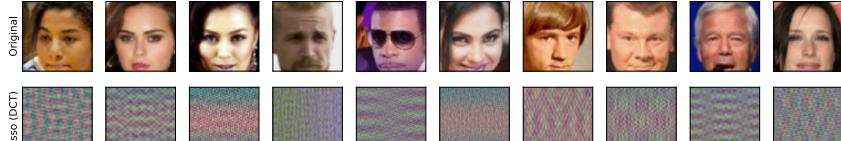


(b) 100 measurements

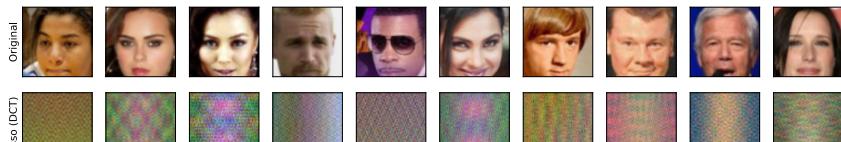


(c) 400 measurements

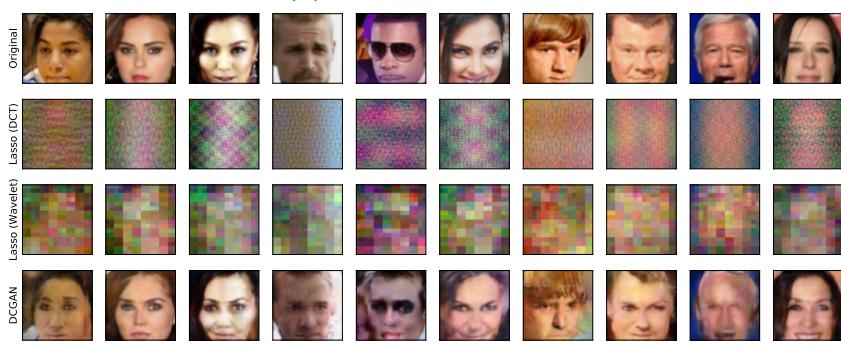
Figure A.4: Reconstruction on MNIST. In each image, top row is ground truth, middle row is Lasso, bottom row is our algorithm.



(a) 50 measurements



(b) 100 measurements

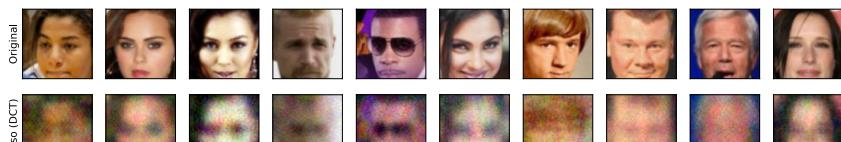


(c) 200 measurements

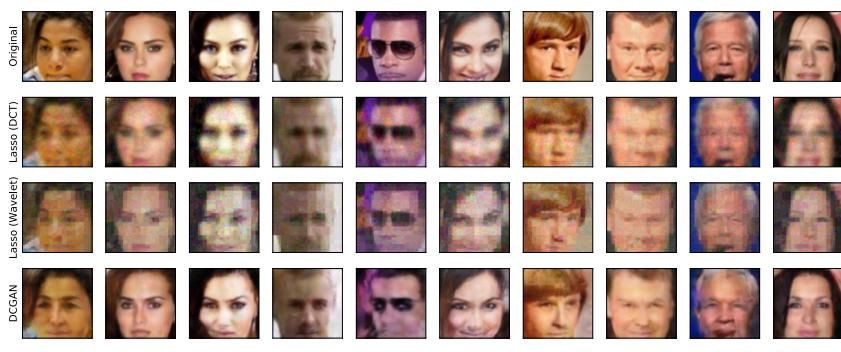
Figure A.5: Reconstruction on celebA. In each image, top row is ground truth, subsequent two rows show reconstructions by Lasso (DCT) and Lasso (Wavelet) respectively. The bottom row is the reconstruction by our algorithm.



(a) 500 measurements



(b) 1000 measurements



(c) 2500 measurements

Figure A.6: Reconstruction on celebA. In each image, top row is ground truth, subsequent two rows show reconstructions by Lasso (DCT) and Lasso (Wavelet) respectively. The bottom row is the reconstruction by our algorithm.



(a) 5000 measurements



(b) 7500 measurements



(c) 10000 measurements

Figure A.7: Reconstruction on celebA. In each image, top row is ground truth, subsequent two rows show reconstructions by Lasso (DCT) and Lasso (Wavelet) respectively. The bottom row is the reconstruction by our algorithm.

Appendix B

Appendix for Chapter 3

B.1 Proof of Lemma 3.6.1

Lemma B.1.1 (Lemma 3.6.1). *Let $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be a d -layered neural network with ReLU activations. Let $A \in \mathbb{R}^{m \times n}$ be a matrix with i.i.d rows satisfying Assumption 3.4.3. If $m = \Omega\left(\frac{1}{1-\gamma^2}kd\log n\right)$, then with probability $1 - e^{-\Omega(m)}$, A satisfies*

$$\frac{1}{m}\|AG(z_1) - AG(z_2)\|^2 \geq \gamma^2\|G(z_1) - G(z_2)\|^2$$

for all $z_1, z_2 \in \mathbb{R}^k$.

Proof. The proof is based on Proposition B.1.2 and Proposition B.1.3, which will be introduced as follows. Proposition B.1.2 shows that the set $S_G = \{G(z_1) - G(z_2) : z_1, z_2 \in \mathbb{R}^k\}$ lies in the range of $e^{O(kd\log n)}$ different $2k$ -dimensional subspaces.

Proposition B.1.3 guarantees the result for a single subspace with probability $1 - e^{-m}$. Since $m = \Omega(kd\log n)$, the proof follows from a union bound over the $e^{O(kd\log n)}$ subspaces in Proposition B.1.2. \square

Proposition B.1.2. *If $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ is a d -layered neural network with ReLU activations, then the set $S_G = \{G(z_1) - G(z_2) : z_1, z_2 \in \mathbb{R}^k\}$ lies in the union of $O(n^{2kd})$ different $2k$ -dimensional subspaces.*

Proof of Proposition (B.1.2). From Lemma 8.3 in [41], the set $\{G(z) : z \in \mathbb{R}^k\}$ lies in the union of $O(n^{kd})$ different k -dimensional subspaces.

This implies that the set

$$\{G(z_1) - G(z_2) : z_1, z_2 \in \mathbb{R}^k\}$$

lies in the union of $M = O(n^{2kd})$ different $2k$ -dimensional subspaces.

□

Proposition B.1.3. *Consider a single $2k$ -dimensional subspace given by $S_1 = \{Wz : W \in \mathbb{R}^{n \times 2k}, W^T W = I_{2k}, z \in \mathbb{R}^{2k}\}$. Let $A \in \mathbb{R}^{m \times n}$ be a matrix with i.i.d rows drawn from a distribution satisfying Assumption (3.4.3). If $m = O(\frac{C^2 k}{\frac{3}{4} - \gamma^2})$, with probability $1 - e^{-\Omega(m)}$, A satisfies*

$$\frac{1}{m} \|Av\|^2 \geq \gamma^2 \|v\|^2, \forall v \in S_1.$$

Proof. The proof follows Theorem 14.12 in [276], with non-trivial modifications for our setting.

We want to show that for all vectors $v \in S_1$,

$$\frac{1}{m} \|Av\|^2 \geq \gamma^2 \|v\|^2.$$

For $u, \tau \in \mathbb{R}$, define the truncated quadratic function

$$\phi_\tau(u) = \begin{cases} u^2 & \text{if } |u| \leq \tau, \\ \tau^2 & \text{otherwise.} \end{cases} \quad (\text{B.1})$$

By construction, $\phi_\tau(\langle a_i, v \rangle) \leq \langle a_i, v \rangle^2$.

This implies that

$$\frac{1}{m} \|Av\|^2 = \frac{1}{m} \sum_{i=1}^m \langle a_i, v \rangle^2 = \frac{\|v\|^2}{m} \sum_{i=1}^m \langle a_i, \frac{v}{\|v\|} \rangle^2 \quad (\text{B.2})$$

$$\geq \frac{\|v\|^2}{m} \sum_{i=1}^m \phi_\tau(\langle a_i, \frac{v}{\|v\|} \rangle) \quad (\text{B.3})$$

$$\geq \|v\|^2 \mathbb{E} \left[\frac{\sum_{i=1}^m \phi_\tau(\langle a_i, \frac{v}{\|v\|} \rangle)}{m} \right] - \|v\|^2 \left| \frac{\sum_{i=1}^m \phi_\tau(\langle a_i, \frac{v}{\|v\|} \rangle)}{m} - \mathbb{E} \left[\frac{\sum_{i=1}^m \phi_\tau(\langle a_i, \frac{v}{\|v\|} \rangle)}{m} \right] \right| \quad (\text{B.4})$$

$$= \|v\|^2 \mathbb{E} \left[\phi_\tau(\langle a, \frac{v}{\|v\|} \rangle) \right] - \|v\|^2 \left| \frac{\sum_{i=1}^m \phi_\tau(\langle a_i, \frac{v}{\|v\|} \rangle)}{m} - \mathbb{E} \left[\phi_\tau(\langle a, \frac{v}{\|v\|} \rangle) \right] \right| \quad (\text{B.5})$$

$$\geq \|v\|^2 \mathbb{E} \left[\phi_\tau(\langle a, \frac{v}{\|v\|} \rangle) \right] - \|v\|^2 \sup_{v \in S_1} \left| \frac{1}{m} \sum_{i=1}^m \phi_\tau(\langle a_i, \frac{v}{\|v\|} \rangle) - \mathbb{E} \left[\phi_\tau(\langle a, \frac{v}{\|v\|} \rangle) \right] \right| \quad (\text{B.6})$$

In Claim B.1.4 we will show that for $\tau^2 = \frac{C^4}{\frac{3}{4} - \gamma^2}$, we have

$$\mathbb{E} \left[\phi_\tau(\langle a, \frac{v}{\|v\|} \rangle) \right] \geq (\gamma^2 + \frac{1}{4}).$$

In Claim B.1.5 we will show that with overwhelming probability in m ,

$$\sup_{v: \|v\| \leq 1} \left| \frac{1}{m} \sum_{i=1}^m \phi_\tau(\langle a_i, \frac{v}{\|v\|} \rangle) - \mathbb{E} \left[\phi_\tau(\langle a, \frac{v}{\|v\|} \rangle) \right] \right| \leq \frac{1}{4}.$$

These two results together imply that

$$\frac{1}{m} \|Av\|^2 \geq \gamma^2 \|v\|^2.$$

with overwhelming probability in m . \square

Claim B.1.4. *Assume that the random vector a satisfies Assumption (3.4.3) with constant C . Let ϕ_τ be the thresholded quadratic function defined in*

Eqn (B.1). For all $v \in \mathbb{R}^n$, $\|v\| \leq 1$, we have

$$\mathbb{E} [\phi_\tau(\langle a, v \rangle)] \geq \left(1 - \frac{C^4}{\tau^2}\right) \|v\|^2.$$

Proof.

$$\|v\|^2 - \mathbb{E} [\phi_\tau(\langle a, v \rangle)] = \mathbb{E} [\langle a, v \rangle^2] - \mathbb{E} [\phi_\tau(\langle a, v \rangle)] \quad (\text{B.7})$$

$$= \mathbb{E} [(\langle a, v \rangle^2 - \tau^2) 1_{\{|\langle a, v \rangle| \geq \tau\}}] \quad (\text{B.8})$$

$$\leq \mathbb{E} [\langle a, v \rangle^2 1_{\{|\langle a, v \rangle| \geq \tau\}}] \quad (\text{B.9})$$

By the Cauchy-Schwartz inequality,

$$\mathbb{E} [\langle a, v \rangle^2 1_{\{|\langle a, v \rangle| \geq \tau\}}] \leq (\mathbb{E} [\langle a, v \rangle^4])^{\frac{1}{2}} (\Pr [|\langle a, v \rangle| \geq \tau])^{\frac{1}{2}} \quad (\text{B.10})$$

From Assumption (3.4.3), we have

$$(\mathbb{E} [\langle a, v \rangle^4])^{\frac{1}{2}} \leq C^2 \mathbb{E} [\langle a, v \rangle^2].$$

From Chebyshev's inequality and Assumption (3.4.3), we have

$$(\Pr [|\langle a, v \rangle| \geq \tau])^{\frac{1}{2}} \leq \left(\frac{\mathbb{E} [|\langle a, v \rangle|^4]}{\tau^4}\right)^{\frac{1}{2}} \leq \left(\frac{C^4 \mathbb{E} [|\langle a, v \rangle|^2]^2}{\tau^4}\right)^{\frac{1}{2}} = \frac{C^2 \mathbb{E} [|\langle a, v \rangle|^2]}{\tau^2}. \quad (\text{B.11})$$

Substituting the above two inequalities into eq. (B.10), we get

$$\mathbb{E} [\langle a, v \rangle^2 1_{\{|\langle a, v \rangle| \geq \tau\}}] \leq \frac{C^4 \mathbb{E} [\langle a, v \rangle^2]^2}{\tau^2} \quad (\text{B.12})$$

$$= \frac{C^4 \|v\|^4}{\tau^2} \leq \frac{C^4 \|v\|^2}{\tau^2}. \quad (\text{B.13})$$

Substituting into Eqn (B.7),

$$\|v\|^2 - \mathbb{E} [\phi_\tau(\langle a, v \rangle)] \leq \frac{C^4 \|v\|^2}{\tau^2}, \quad (\text{B.14})$$

which completes the proof. \square

Claim B.1.5. *For an orthonormal matrix $U \in \mathbb{R}^{n \times 2k}$, let $S := \{v : v = Uz, \|v\| = 1\}$. Let ϕ_τ be the function defined in Proposition B.1.3. For $m = \Omega(\tau^2 k)$, we have*

$$\sup_{v \in S} \left| \frac{1}{m} \sum_{i=1}^m \phi_\tau(\langle a_i, v \rangle) - \mathbb{E} [\phi_\tau(\langle a, v \rangle)] \right| \leq \frac{1}{4}.$$

with probability $1 - e^{-\Omega(m)}$.

Proof. Define

$$Z_m = \sup_{v \in S} \left| \frac{1}{m} \sum_{i=1}^m \phi_\tau(\langle a_i, v \rangle) - \mathbb{E} [\phi_\tau(\langle a, v \rangle)] \right|.$$

We will first show that

$$\mathbb{E}_A [Z_m] \leq \frac{1}{8}$$

for large enough m . Then we use Talagrand's inequality [259] to show that

$$\Pr \left[Z_m \geq \mathbb{E} [Z_m] + \frac{1}{8} \right] \leq e^{-\Omega(m)},$$

using which we can conclude that $Z_m \leq \frac{1}{4}$ with probability $1 - e^{-\Omega(m)}$.

By the symmetrization inequality, we have

$$\mathbb{E}_A [Z_m] \leq 2 \mathbb{E}_{\epsilon, A} \left[\sup_{v \in S} \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \phi_\tau(\langle a_i, v \rangle) \right| \right]$$

where $\{\epsilon_i\}_{i=1}^m$ are i.i.d Bernoulli ± 1 random variables.

Since ϕ_τ is a Lipschitz function with Lipschitz constant 2τ , we can apply the Ledoux-Talagrand contraction inequality [170] (refer to Section B.7 for the sake of completeness) to get

$$\begin{aligned} & 2 \mathbb{E}_{\epsilon, A} \left[\sup_{v \in S} \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \phi_\tau(\langle a_i, v \rangle) \right| \right] \\ & \leq 8\tau \mathbb{E}_{\epsilon, A} \left[\sup_{v \in S} \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i \langle a_i, v \rangle \right| \right] \end{aligned} \quad (\text{B.15})$$

$$= 8\tau \mathbb{E}_{\epsilon, A} \left[\sup_{v \in S} \left| \frac{1}{m} \epsilon^T A v \right| \right]. \quad (\text{B.16})$$

Since $S := \{v : v = Uz, \|v\| = 1\}$, we have

$$8\tau \mathbb{E}_{\epsilon, A} \left[\sup_{v \in S} \left| \frac{1}{m} \epsilon^T A v \right| \right] \quad (\text{B.17})$$

$$= 8\tau \mathbb{E}_{\epsilon, A} \left[\sup_{z: \|z\|=1} \left| \frac{8\tau}{m} \epsilon^T A U z \right| \right] \quad (\text{B.18})$$

$$\leq \frac{8\tau}{m} \mathbb{E}_{\epsilon, A} [\|\epsilon^T A U\|_2] \quad (\text{B.19})$$

$$\leq \frac{8\tau}{m} \sqrt{\mathbb{E}_{\epsilon, A} [\|\epsilon^T A U\|_2^2]} \quad (\text{B.20})$$

The third line follows from the Cauchy-Schwartz inequality, and the fourth line follows from Jensen's inequality.

Notice that

$$\mathbb{E}_\epsilon [\|\epsilon^T A U\|_2^2] = \text{trace}(A U U^T A^T) = \text{trace}(U^T A^T A U)$$

Since $U^T U = I_{2k}$, we have

$$\mathbb{E}_{\epsilon, A} [\|\epsilon^T A U\|_2^2] = \mathbb{E}_A [\text{trace}(U^T A^T A U)] \quad (\text{B.21})$$

$$= \sum_{i=1}^m \mathbb{E}_{a_i} \text{trace}(U^T a_i a_i^T U) \quad (\text{B.22})$$

$$= \sum_{i=1}^m \text{trace}(U^T I_n U) = m \text{ trace}(I_{2k}) = 2km. \quad (\text{B.23})$$

Putting this together, and choosing $m = \Omega(\tau^2 k)$, we have

$$\mathbb{E}_A [Z_m] \leq 8\tau \sqrt{\frac{2k}{m}} \leq \frac{1}{8}.$$

We now need to show that

$$\Pr \left[Z_m \geq \mathbb{E}_A [Z_m] + \frac{1}{8} \right] \leq e^{-\Omega(m)}.$$

By construction, $\phi_\tau(\langle a_i, v \rangle) \leq \tau^2$ for all $v \in S$.

In order to apply Talagrand's inequality, we need to bound

$$\sigma^2 = \sup_{v \in S} \mathbb{E} [(\phi_\tau(\langle a, v \rangle) - \mathbb{E} [\phi_\tau(\langle a, v \rangle)])^2].$$

We can bound this by

$$\text{var}(\phi_\tau(\langle a, v \rangle)) \leq \mathbb{E} [\phi_\tau^2(\langle a, v \rangle)] \quad (\text{B.24})$$

$$\leq \tau^2 \mathbb{E} [\phi_\tau(\langle a, v \rangle)] \leq \tau^2 \quad (\text{B.25})$$

Applying Talagrand's inequality, we have

$$\Pr [Z_m \geq \mathbb{E}_A [Z_m] + t] \leq C_1 \exp \left(-\frac{C_2 m t^2}{\tau^2 + \tau^2 t} \right).$$

Setting $t = \frac{1}{8}$, $m = \Omega(\tau^2 k)$ we get

$$\Pr[Z_m \geq \frac{1}{4}] \leq \Pr\left[Z_m \geq \mathbb{E}[Z_m] + \frac{1}{8}\right] \leq C_1 e^{-\frac{C_2 m}{\tau^2}} = e^{-\Omega(m)}.$$

This concludes the proof. \square

B.2 Proof of Lemma 3.6.2

Lemma B.2.1. *Let M denote the number of batches. Then with probability $1 - e^{-\Omega(M)}$, the objective in Equation (3.2) satisfies*

$$\min_{z \in \mathbb{R}^k} \max_{z' \in R^k} \operatorname{median}_{B_j}(z) - \ell_{B_j}(z') \leq 4\sigma^2. \quad (\text{B.26})$$

Proof. By setting $z \leftarrow z^*$, for all $z' \in \mathbb{R}^k$, for any $j \in [M]$, we have

$$\ell_{B_j}(z^*) - \ell_{B_j}(z') \leq \ell_{B_j}(z^*) = \frac{1}{b} \|\eta_{B_j}\|^2. \quad (\text{B.27})$$

Since the noise is i.i.d. and has variance σ^2 , we have $\mathbb{E}[\ell_{B_j}(z^*)] = \mathbb{E}\frac{1}{b} \|\eta_{B_j}\|^2 = \sigma^2$.

For batch $j \in [M]$, define the indicator random variable

$$Y_j = \mathbf{1}\{\ell_{B_j}(z^*) \geq 4\sigma^2\}.$$

By Markov's inequality, since $\mathbb{E}[\ell_{B_j}(z^*)] = \sigma^2$, we have

$$\Pr[Y_j = 1] \leq \frac{1}{4} \Rightarrow \mathbb{E}\left[\sum_{j=1}^M Y_j\right] \leq \frac{M}{4}. \quad (\text{B.28})$$

By the Chernoff bound,

$$\Pr \left[\sum_{j=1}^M Y_j \geq \frac{M}{2} \right] \leq \Pr \left[\sum_{j=1}^M Y_j \geq 2 \mathbb{E} \left[\sum_{j=1}^M Y_j \right] \right] \leq e^{-\Omega(M)}. \quad (\text{B.29})$$

The above inequality implies that with probability $1 - e^{-\Omega(M)}$, for all $z' \in \mathbb{R}^k$, at least $\frac{M}{2}$ batches satisfy

$$\ell_{B_j}(z^*) - \ell_{B_j}(z') \leq 4\sigma^2.$$

This gives

$$\min_{z \in \mathbb{R}^k} \max_{z' \in R^k} \text{median}_{1 \leq j \leq M} (\ell_{B_j}(z) - \ell_{B_j}(z')) \leq 4\sigma^2. \quad (\text{B.30})$$

□

B.3 Proof of Lemma 3.6.3

Lemma B.3.1 (Lemma 3.6.3). *Let $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be a generative model from a d -layer neural network using ReLU activations. Let $A \in \mathbb{R}^{m \times n}$ be a matrix with i.i.d rows satisfying Assumption 3.4.3. Let the batch size $b = \Theta(C^4)$, let the number of batches satisfy $M = \Omega(kd \log n)$, and let γ be a constant which depends on the moment constant C . Then with probability at least $1 - e^{-\Omega(m)}$, for all $z_1, z_2 \in \mathbb{R}^k$ there exists a set $J \subseteq [M]$ of cardinality at least $0.9M$ such that*

$$\frac{1}{b} \|A_{B_j}(G(z_1) - G(z_2))\|^2 \geq \gamma^2 \|G(z_1) - G(z_2)\|^2, \forall j \in J.$$

Proof. Proposition B.1.2 shows that the set $S_G = \{G(z_1) - G(z_2) : z_1, z_2 \in \mathbb{R}^k\}$ lies in the range of $e^{O(kd \log n)}$ different $2k$ -dimensional subspaces.

Proposition B.3.2 guarantees the result for a single subspace with probability $1 - e^{-\Omega(M)}$. Since $M = \Omega(kd \log n)$ and the batch size is constant which depends on the moment constant C , the lemma follows from a union bound over the $e^{O(kd \log n)}$ subspaces in Proposition B.1.2. \square

Proposition B.3.2. *Consider a single $2k$ -dimensional subspace given by $S = \{Wz : W \in \mathbb{R}^{n \times 2k}, W^T W = I_{2k}, z \in \mathbb{R}^{2k}\}$. Let $A \in \mathbb{R}^{m \times n}$ be a matrix with i.i.d rows drawn from a distribution satisfying Assumption (3.4.3) with constant C . If the batch size $b = O(C^4)$ and the number of batches satisfies $M = \Omega(k \log \frac{1}{\varepsilon})$, with probability $1 - e^{-\Omega(M)}$, for all $x \in S$, there exist a subset of batches $J_x \subseteq [M]$ with $|J_x| \geq 0.90M$ such that*

$$\frac{1}{b} \|A_{B_j} x\|^2 \geq \gamma^2 \|x\|^2 \quad \forall j \in J_x,$$

where $\gamma = \Theta(\frac{1}{C^2})$ is a constant that depends on the moment constant C .

Proof. Since the bound we want to prove is homogeneous, it suffices to show it for all vectors in S that have unit norm. Let $W \in \mathbb{R}^{n \times 2k}$ be the orthonormal matrix spanning S , and S_1 denote the set of unit norm vectors in its span. That is,

$$S_1 = \{Wz : z \in \mathbb{R}^{2k}, \|z\| = 1, W \in \mathbb{R}^{n \times 2k}, W^T W = I_{2k}\}.$$

For a fixed $x \in S_1$ and $0 < t < 1$, we have

$$\mathbb{E} [\langle a, x \rangle^2] = \mathbb{E} [\langle a, x \rangle^2 \mathbf{1}\{\langle a, x \rangle \leq t^2 \|x\|^2\}] \mathbb{E} [\langle a, x \rangle^2 \mathbf{1}\{\langle a, x \rangle > t^2 \|x\|^2\}] \quad (\text{B.31})$$

$$\leq t^2 \|x\|^2 + \mathbb{E} [\langle a, x \rangle^4]^{\frac{1}{2}} (\Pr [\langle a, x \rangle^2 \geq t^2 \|x\|^2])^{\frac{1}{2}} \quad (\text{B.32})$$

$$\leq t^2 \|x\|^2 + C^2 \|x\|^2 (\Pr [\langle a, x \rangle^2 \geq t^2 \|x\|^2])^{\frac{1}{2}} \quad (\text{B.33})$$

$$\Rightarrow \Pr [\langle a, x \rangle^2 \geq t^2 \|x\|^2] \geq \frac{(1-t^2)^2 \|x\|^4}{C^4 \|x\|^4} = \frac{(1-t^2)^2}{C^4} = C_1. \quad (\text{B.34})$$

This is essentially a modified version of the Paley-Zigmund inequality [213].

Consider a batch B_j , which has b samples. By the concentration of Bernoulli random variables, with probability $1 - 2e^{-\Omega(C_1 b)}$, we have

$$\sum_{i \in B_j} \mathbf{1} \{ \langle a_i, x \rangle^2 \geq t^2 \|x\|^2 \} \geq \frac{bC_1}{2}$$

This implies that if we set b such that $1 - 2e^{-\Omega(C_1 b)} = 0.975$, then with probability 0.975, B_j has $\frac{bC_1}{2}$ samples $\langle a_i, x \rangle$ whose magnitude is at least $t\|x\|$. This implies that the average square magnitude over the batch satisfies

$$\frac{1}{b} \|A_{B_j} x\|^2 = \frac{1}{b} \sum_{i \in B_j} \langle a_i, x \rangle^2 \geq t^2 \|x\|^2 \frac{bC_1}{2b} = \frac{C_1 t^2 \|x\|^2}{2}, \quad (\text{B.35})$$

with probability 0.975.

Consider the indicator random variable associated with the complement of the above event. That is,

$$Y_j(x) = \left\{ \frac{1}{b} \|A_{B_j} x\|^2 \leq \frac{C_1 t^2}{2} \|x\|^2 \right\}$$

From (B.35) we have that $\mathbb{E}[Y_j(x)] \leq 0.025$.

Consider the sum of indicator random variables over M batches. By standard concentrations of Bernoulli random variables, we have with probability $1 - e^{-\Omega(M)}$,

$$\sum_{j=1}^M Y_j(x) \leq 2 \mathbb{E} \left[\sum_{j=1}^M Y_j(x) \right] \leq 0.05.$$

This implies that there exist a subset of batches $J \subseteq [M]$ with $|J| \geq 0.95M$ such that

$$\frac{1}{b} \|A_{B_j} x\|^2 \geq \frac{C_1 t^2 \|x\|^2}{2} \quad \forall j \in J,$$

with probability $1 - e^{-\Omega(M)}$. This shows that we have the statement of the proposition for a fixed vector in S_1 .

We now show that this holds true for an ε -cover of S_1 . Let S_ε denote a minimial ε -covering of S_1 . That is, S_ε is a finite subset of S_1 such that for all $x \in S_1$, there exists $\tilde{x} \in S_\varepsilon$ such that $\|x - \tilde{x}\| \leq \varepsilon$. Since S_1 has dimension $2k$ and diameter 1, we can find a set S_ε whose cardinality is at most $(O(\frac{1}{\varepsilon}))^{2k}$.

By a union bound, with probability $1 - e^{-\Omega(M)} |S_\varepsilon|$, for all $\tilde{x} \in S_\varepsilon$ there exists a subset of batches $J_{\tilde{x}} \subset [M]$ with $|J_{\tilde{x}}| \geq 0.95M$ such that

$$\frac{1}{b} \|A_{B_j} \tilde{x}\|^2 \geq \frac{C_1 t^2}{2} \quad \forall j \in J_{\tilde{x}} \tag{B.36}$$

Since $|S|_\varepsilon \leq e^{O(k \log \frac{1}{\varepsilon})}$, if $M = \Omega(k \log \frac{1}{\varepsilon})$, the above statement holds with probability $1 - e^{-\Omega(M)}$.

We now show that the statement of the proposition is true for all vectors in S_1 . Since the proposition statement holds for an ε -cover of S_1 , we now only need to consider the effect of A at a scale of ε .

Now consider the set

$$S_2 = \{x - \tilde{x} : x \in S_1, \tilde{x} \in S_\varepsilon, \|x - \tilde{x}\| \leq \varepsilon\}.$$

Note that this a subset of all vectors in the span of W that have norm at most ε . That is, if

$$S_3 = \{Wz : z \in \mathbb{R}^{2k}, \|z\| \leq \varepsilon\},$$

we have $S_2 \subseteq S_3$.

For a vector $v \in \mathbb{R}^n$, consider the random variable

$$Z_i(v) = \mathbf{1} \left[\langle a_i, v \rangle \geq \frac{\sqrt{C_1}t}{2\sqrt{2}} \right].$$

Define the random process

$$\Psi(a_1, a_2, \dots, a_m) = \sup_{v \in S_2} \frac{1}{m} \sum_{i=1}^m \mathbf{1} \left[|\langle a_i, v \rangle| \geq \frac{\sqrt{C_1}t}{2\sqrt{2}} \right].$$

By the bounded difference inequality, with probability $1 - 2e^{-C_2\delta^2}$,

$$\Psi(a_1, a_2, \dots, a_m) \leq \mathbb{E} [\Psi(a_1, a_2, \dots, a_m)] + \frac{\delta}{\sqrt{m}}$$

Since $S_2 \subseteq S_3$, we can bound the expectation of Ψ by

$$\mathbb{E} [\Psi(a_1, \dots, a_m)] \leq \mathbb{E} \sup_{v \in S_3} \frac{1}{m} \sum_{i=1}^m \mathbf{1} \left[|\langle a_i, v \rangle| \geq \frac{\sqrt{C_1}t}{2\sqrt{2}} \right] \quad (\text{B.37})$$

$$\leq \mathbb{E} \sup_{v \in S_3} \sum_{i=1}^m \frac{|\langle a_i, v \rangle|}{mt\sqrt{C_1}/2\sqrt{2}} \quad (\text{B.38})$$

$$= \mathbb{E} \sup_{v \in S_3} \sum_{i=1}^m \frac{2\sqrt{2}|\langle a_i, v \rangle|}{mt\sqrt{C_1}} \quad (\text{B.39})$$

$$\leq \mathbb{E} \sup_{v \in S_3} \left| \sum_{i=1}^m 2\sqrt{2} \frac{|\langle a_i, v \rangle| - \mathbb{E}[|\langle a, v \rangle|]}{mt\sqrt{C_1}} \right| + \sup_{v \in S_3} \sum_{i=1}^m \frac{2\sqrt{2}\mathbb{E}[|\langle a, v \rangle|]}{mt\sqrt{C_1}} \quad (\text{B.40})$$

Since a is isotropic and v has norm at most ε , by Jensen's inequality, we can bound the second term in the RHS by

$$\mathbb{E} \sup_{v \in S_3} \sum_{i=1}^m \frac{2\sqrt{2}\mathbb{E}[|\langle a, v \rangle|]}{mt\sqrt{C_1}} \lesssim \frac{\varepsilon}{t\sqrt{C_1}}. \quad (\text{B.41})$$

To bound the first term in the RHS, we use the Gine-Zinn symmetrization inequality [93, 198, 170]

$$\mathbb{E} \sup_{v \in S_3} \left| \sum_{i=1}^m 2\sqrt{2} \frac{|\langle a_i, v \rangle| - \mathbb{E}[|\langle a, v \rangle|]}{mt\sqrt{C_1}} \right| \lesssim \mathbb{E} \sup_{v \in S_3} \left| \sum_{i=1}^m \frac{\xi_i \langle a_i, v \rangle}{mt\sqrt{C_1}} \right| \quad (\text{B.42})$$

where $\xi_i, i \in [m]$ are i.i.d ± 1 Bernoulli variables.

We can bound this by

$$\mathbb{E} \sup_{v \in S_3} \left| \sum_{i=1}^m \frac{\xi_i \langle a_i, v \rangle}{mt\sqrt{C_1}} \right| = \mathbb{E}_{\xi, A} \left[\sup_{v \in S_3} \left| \frac{\xi^T A v}{mt\sqrt{C_1}} \right| \right], \quad (\text{B.43})$$

$$= \mathbb{E}_{\xi, A} \left[\sup_{z: \|z\| \leq \varepsilon} \left| \frac{\xi^T A W z}{mt\sqrt{C_1}} \right| \right] \quad (\text{B.44})$$

$$\leq \mathbb{E}_{\xi, A} \left[\frac{\epsilon \|\xi^T A W\|}{mt\sqrt{C_1}} \right] \quad (\text{B.45})$$

$$\leq \frac{\epsilon \sqrt{\mathbb{E}_{\xi, A} \|\xi^T A W\|^2}}{mt\sqrt{C_1}} \quad (\text{B.46})$$

$$= \frac{\epsilon \sqrt{\mathbb{E}_A \text{trace}(A W W^T A^T)}}{mt\sqrt{C_1}} \quad (\text{B.47})$$

$$= \frac{\epsilon \sqrt{2km}}{mt\sqrt{C_1}} \lesssim \frac{\varepsilon}{t} \sqrt{\frac{k}{mC_1}} \quad (\text{B.48})$$

The third line follows from the Cauchy-Schwartz inequality, and the fourth line follows from Jensen's inequality.

Since $m = Mb$, from the above inequality and Eqn (B.41) we can now bound $\mathbb{E} \Psi$ as

$$\mathbb{E} [\Psi(a_1, \dots, a_m)] \lesssim \frac{\varepsilon}{t} \sqrt{\frac{k}{MbC_1}} + \frac{\varepsilon}{t\sqrt{C_1}} \quad (\text{B.49})$$

Substituting the above inequality into the bounded difference inequality, we have with probability at least $1 - e^{-\Omega(\delta^2)}$,

$$\Psi(a_1, a_2, \dots, a_m) \lesssim \frac{\varepsilon}{t} \sqrt{\frac{k}{MbC_1}} + \frac{\varepsilon}{t\sqrt{C_1}} + \frac{\delta}{\sqrt{Mb}} \quad (\text{B.50})$$

Setting $M = \Omega(k)$, $\delta = O\left(\sqrt{\frac{M}{b}}\right)$, $\varepsilon = O\left(\frac{t}{b}\sqrt{C_1}\right)$, we can reduce the

terms in the above inequality to

$$\frac{\varepsilon}{t} \sqrt{\frac{k}{MbC_1}} \leq O\left(\frac{1}{b^{\frac{3}{2}}}\right), \quad (\text{B.51})$$

$$\frac{\varepsilon}{t\sqrt{C_1}} \leq O\left(\frac{1}{b}\right), \quad (\text{B.52})$$

$$\frac{\delta}{\sqrt{Mb}} \leq O\left(\frac{1}{b}\right), \quad (\text{B.53})$$

Since $b > 1$, the sum of these three terms is dominated by $O\left(\frac{1}{b}\right)$. From this, we can conclude that for small enough ε, δ , with probability $1 - e^{-\Omega(\frac{M}{b})}$,

$$\Psi(a_1, a_2, \dots, a_m) \leq \frac{0.05}{b} \quad (\text{B.54})$$

$$\Rightarrow \sup_{v \in S_3} \sum_{i=1}^m \mathbf{1} \left[|\langle a_i, v \rangle| \geq \frac{t\sqrt{C_1}}{2\sqrt{2}} \right] \leq 0.05M. \quad (\text{B.55})$$

This allows us to control the effect of A at a scale of ε . It says that there at most $0.05M$ samples on which vectors with magnitude at most ε have a magnitude greater than $\frac{t\sqrt{C_1}}{2\sqrt{2}}$ after interacting with A . This implies that there at least $0.95M$ batches in which all samples are well behaved.

Since we have control over an ε -cover of S_1 as well as vectors at a scale of ε in S_1 , we can now prove our result for all vectors in S_1 .

For any $x \in S_1$, let $\tilde{x} \in S_\varepsilon$ be the point in the ε -cover which is closest to x . For a batch B_j , we can express $\|A_{B_j}x\|$ as

$$\frac{1}{\sqrt{b}} \|A_{B_j}x\| \geq \frac{1}{\sqrt{b}} \|A_{B_j}\tilde{x}\| - \frac{1}{\sqrt{b}} \|A_{B_j}(x - \tilde{x})\|. \quad (\text{B.56})$$

From (B.36), there exists a subset of batches $J_{\tilde{x}} \subseteq [M]$ with $|J_{\tilde{x}}| \geq 0.95M$ such that

$$\frac{1}{\sqrt{b}} \|A_{B_j}\tilde{x}\| \geq \frac{\sqrt{C_1}t}{\sqrt{2}} \quad \forall j \in J_{\tilde{x}}. \quad (\text{B.57})$$

From (B.55), there exists a subset of batches $J_{x-\tilde{x}} \subseteq [M]$ with $|J_{x-\tilde{x}}| \geq 0.95M$ such that for all $j \in J_{x-\tilde{x}}$,

$$|\langle a_i, x - \tilde{x} \rangle| \leq \frac{\sqrt{C_1}t}{2\sqrt{2}} \quad \forall i \in B_j \quad (\text{B.58})$$

$$\Rightarrow \frac{1}{\sqrt{b}} \|A_{B_j}(x - \tilde{x})\| \leq \frac{\sqrt{C_1}t}{2\sqrt{2}}, \quad (\text{B.59})$$

$$\Rightarrow -\frac{1}{\sqrt{b}} \|A_{B_j}(x - \tilde{x})\| \geq -\frac{\sqrt{C_1}t}{2\sqrt{2}}. \quad (\text{B.60})$$

From the bounds on $\|A_{B_j}\tilde{x}\|$ and the bound on $\|A_{B_j}(x - \tilde{x})\|$, we can conclude that for all $x \in S_1$ there exist a subset of batches $J_x = J_{\tilde{x}} \cap J_{x-\tilde{x}}$ with cardinality at least $0.9M$ such that

$$\frac{1}{\sqrt{b}} \|A_{B_j}x\| \geq \frac{\sqrt{C_1}t}{2\sqrt{2}}, \quad \forall j \in J_x. \quad (\text{B.61})$$

This completes the proof, with $\gamma = \frac{\sqrt{C_1}t}{2\sqrt{2}} = \frac{t(1-t^2)}{C^2 2\sqrt{2}}$. \square

B.4 Proof of Lemma 3.6.4

Lemma B.4.1 (Lemma 3.6.4). *Consider the setting of Lemma 3.6.3 with measurements satisfying $y = AG(z^*) + \eta$. For any $t > 0$ and noise variance σ^2 , let the batch size b and number of batches M satisfy $b = \Theta(\frac{\sigma^2}{t^2})$ and $M = \Omega(kd \log n)$. Then with probability at least $1 - e^{-\Omega(m)}$, for all $z \in \mathbb{R}^k$ there exists a set $J \subseteq [M]$ of cardinality at least $0.9M$ such that*

$$\frac{1}{b} |\eta_{B_j}^T A_{B_j}(G(z) - G(z^*))| \leq t \|G(z) - G(z^*)\|, \quad \forall j \in J.$$

Proof. Proposition B.1.2 shows that the set $S_G = \{G(z_1) - G(z_2) : z_1, z_2 \in \mathbb{R}^k\}$ lies in the range of $e^{O(kd \log n)}$ different $2k$ -dimensional subspaces. This trivially

implies that for a fixed $z^* \in \mathbb{R}^k$, the set $\{G(z) - G(z^*) : z \in \mathbb{R}^k\}$ also lies in the range of $e^{O(kd \log n)}$ different $2k$ -dimensional subspaces.

Proposition B.4.2 guarantees the result for a single subspace with probability $1 - e^{-\Omega(M)}$. Since $M = \Omega(kd \log n)$ and the batch size is constant which depends on the noise variance σ^2 and t^2 , the lemma follows from a union bound over the $e^{O(kd \log n)}$ subspaces. \square

Proposition B.4.2. *Consider a single $2k$ -dimensional subspace given by $S = \{Wz : W \in \mathbb{R}^{n \times 2k}, W^T W = I_{2k}, z \in \mathbb{R}^{2k}\}$. Let $A \in \mathbb{R}^{m \times n}$ be a matrix with i.i.d rows drawn from a distribution satisfying Assumption (3.4.3) with constant C . If the batch size $b = \Theta\left(\frac{\sigma^2}{t^2}\right)$ and the number of batches satisfies $M = \Omega\left(k \log \frac{1}{\varepsilon}\right)$, with probability $1 - e^{-\Omega(M)}$, for all $x \in S$, there exist a subset of batches $J_x \subseteq [M]$ with $|J_x| \geq 0.90M$ such that*

$$\frac{1}{b} |\eta_{B_j}^T A_{B_j} x| \leq t \|x\|, \forall j \in J.$$

Proof. Since the bound we want to prove is homogeneous, it suffices to show it for all vectors in S that have unit norm. Let $W \in \mathbb{R}^{n \times 2k}$ be the orthonormal matrix spanning S , and S_1 denote the set of unit norm vectors in its span. That is,

$$S_1 = \{Wz : z \in \mathbb{R}^{2k}, \|z\| = 1, W \in \mathbb{R}^{n \times 2k}, W^T W = I_{2k}\}.$$

Consider the set S_ε , which is a minimal ε -covering of S_1 . That is, for every $x \in S_1$, there exists $\tilde{x} \in S_\varepsilon$ such that $\|\tilde{x} - x\| \leq \varepsilon$.

For a fixed $\tilde{x} \in S_\varepsilon$, and $t > 0$, by Chebyshev's inequality,

$$\Pr\left[\frac{1}{b}|\eta^T A_{B_j} \tilde{x}| \geq \frac{t}{2}\right] \leq \frac{\sum_{i \in B_j} (\eta_i^2 \langle a_i, \tilde{x} \rangle^2)}{b^2 t^2 / 4} \quad (\text{B.62})$$

$$= \frac{b\sigma^2 \|\tilde{x}\|^2}{b^2 t^2 / 4} \quad (\text{B.63})$$

$$= \frac{\sigma^2 4}{bt^2} \leq \frac{1}{40}, \quad (\text{B.64})$$

if $b \geq \frac{160\sigma^2}{t^2}$.

Define the indicator random variable

$$Y_i(x) = \mathbf{1} \left\{ \frac{1}{b} |\eta^T A_{B_i} x| \geq \frac{t}{2} \right\}.$$

From Eqn (B.64) we have

$$\mathbb{E}[Y_i(\tilde{x})] \leq \frac{1}{40}.$$

By concentration of Bernoulli variables, with probability $1 - e^{-\Omega(M)}$,

$$\sum_{j=1}^M Y_i(\tilde{x}) \leq 2 \mathbb{E}[Y_i(\tilde{x})] \leq \frac{1}{20}.$$

This implies that for a fixed $\tilde{x} \in S_\varepsilon$, with probability $1 - e^{-\Omega(M)}$, there exist a subset of batches $J_{\tilde{x}} \subseteq [M]$ with cardinality $0.95M$ such that

$$\frac{1}{b} |\eta^T A_{B_j} \tilde{x}| \leq \frac{t}{2} \quad \forall j \in J_{\tilde{x}}. \quad (\text{B.65})$$

Since the size of S_ε is at most $(O(\frac{1}{\varepsilon}))^{2k}$, we can union bound over all \tilde{x} in S_ε . Hence, if $M = \Omega(k \log \frac{1}{\varepsilon})$, then with probability $1 - e^{-\Omega(M)}$, for all $\tilde{x} \in S_\varepsilon$, there exist a subset $J_{\tilde{x}} \subseteq [M]$ with cardinality $0.95M$ such that

$$\frac{1}{b} |\eta^T A_{B_j} \tilde{x}| \leq \frac{t}{2} \quad \forall j \in J_{\tilde{x}}. \quad (\text{B.66})$$

This shows that the multiplier component is well behaved on a large fraction of the batches for an ε -cover of S_1 . Now we need to extend the argument to all vectors in S_1 .

Now consider the set

$$S_2 = \{x - \tilde{x} : x \in S_1, \tilde{x} \in S_\varepsilon, \|x - \tilde{x}\| \leq \varepsilon\}.$$

Note that this is a subset of all vectors in the span of W that have norm at most ε . That is, if

$$S_3 = \{Wz : z \in \mathbb{R}^{2k}, \|z\| \leq \varepsilon\},$$

we have $S_2 \subseteq S_3$.

For any $v \in \mathbb{R}^n$, define the random variable

$$Z_j(v) = \mathbf{1} \left\{ |\eta_i a_i^T v| \geq \frac{t}{2} \right\}. \quad (\text{B.67})$$

Now define the random process

$$\Psi(a_1, \dots, a_m) = \sup_{v \in S_2} \frac{1}{m} \sum_{i=1}^m Z_i(v) \quad (\text{B.68})$$

Since $S_2 \subseteq S_3$, we can bound $\mathbb{E}[\Psi]$ via

$$\mathbb{E}[\Psi] \leq \mathbb{E} \left[\sup_{v \in S_3} \frac{1}{m} \sum_{i=1}^m Z_i(v) \right] \quad (\text{B.69})$$

$$\leq \mathbb{E} \left[\sup_{v \in S_3} \frac{1}{m} \sum_{i=1}^m \frac{|\eta_i a_i^T v|}{t/2} \right] \quad (\text{B.70})$$

$$\begin{aligned} &\leq \mathbb{E} \left[\sup_{v \in S_3} \left| \frac{1}{m} \sum_{i=1}^m \frac{|\eta_i a_i^T v| - \mathbb{E} |\eta_i a_i^T v|}{t/2} \right| \right] \\ &+ \mathbb{E} \left[\sup_{v \in S_3} \frac{1}{m} \sum_{i=1}^m \frac{\mathbb{E} |\eta_i a_i^T v|}{t/2} \right] \end{aligned} \quad (\text{B.71})$$

We can bound the term on the right by

$$\mathbb{E} \left[\sup_{v \in S_3} \frac{1}{m} \sum_{i=1}^m \frac{\mathbb{E} |\eta_i a_i^T v|}{t/2} \right] \leq \frac{\mathbb{E} \left[\sup_{v \in S_3} \|\eta_i\|_2 |\langle a_i, v \rangle| \right]}{t/2} \quad (\text{B.72})$$

$$\lesssim \frac{\sigma \varepsilon}{t}, \quad (\text{B.73})$$

where we have used the Cauchy Schwartz inequality, followed by the fact that η is independent noise and has variance σ^2 , a is isotropic, and $v \in S_3$ has norm at most ε .

To bound the term on the left, we use the Gine-Zinn symmetrization inequality [93, 198, 170]

$$\mathbb{E} \left[\sup_{v \in S_3} \left| \frac{1}{m} \sum_{i=1}^m \frac{|\eta_i a_i^T v| - \mathbb{E} |\eta_i a_i^T v|}{t/2} \right| \right] \lesssim \mathbb{E} \left[\sup_{v \in S_3} \left| \frac{1}{m} \sum_{i=1}^m \frac{\xi_i \eta_i a_i^T v}{t/2} \right| \right] \quad (\text{B.74})$$

where $\xi_i, i \in [m]$ are i.i.d \pm Bernoulli random variables.

Let $\xi\eta = (\xi_1\eta_1, \xi_2\eta_2, \dots, \xi_m\eta_m)$ denote the element wise product of the vectors $\xi = (\xi_1, \xi_2, \dots, \xi_m)$ and $\eta = (\eta_1, \eta_2, \dots, \eta_m)$. We can bound the

above inequality by

$$\mathbb{E} \sup_{v \in S_3} \left| \sum_{i=1}^m \frac{\xi_i \eta_i \langle a_i, v \rangle}{mt/2} \right| = \mathbb{E}_{\xi, \eta, A} \left[\sup_{v \in S_3} \left| \frac{(\xi \eta)^T A v}{mt/2} \right| \right], \quad (\text{B.75})$$

$$= \mathbb{E}_{\xi, \eta, A} \left[\sup_{z: \|z\| \leq \varepsilon} \left| \frac{(\xi \eta)^T A W z}{mt/2} \right| \right] \quad (\text{B.76})$$

$$\leq \mathbb{E}_{\xi, \eta, A} \left[\frac{\epsilon \|(\xi \eta)^T A W\|}{mt/2} \right] \quad (\text{B.77})$$

$$\leq \frac{\epsilon \sqrt{\mathbb{E}_{\xi, \eta, A} \|(\xi \eta)^T A W\|^2}}{mt/2} \quad (\text{B.78})$$

$$= \frac{\epsilon \sigma \sqrt{\mathbb{E}_A \text{trace}(A W W^T A^T)}}{mt/2} \quad (\text{B.79})$$

$$= \frac{\epsilon \sigma \sqrt{2km}}{mt/2} \lesssim \frac{\epsilon \sigma}{t} \sqrt{\frac{k}{m}} \quad (\text{B.80})$$

The third line follows from the Cauchy-Schwartz inequality, and the fourth line follows from Jensen's inequality, and the fifth line follows from the fact that $\xi \eta$ has i.i.d coordinates that are independent of A and have variance σ^2 .

From the above inequality and eq. (B.72), we get

$$\mathbb{E}[\Psi(a_1, a_2, \dots, a_m)] \lesssim \frac{\sigma \varepsilon}{t} \sqrt{\frac{k}{m}} + \frac{\sigma \varepsilon}{t} \lesssim \frac{\sigma \varepsilon}{t} \quad (\text{B.81})$$

If we choose $\varepsilon = c_1 \frac{t}{\sigma b}$ for a small enough constant c_1 , then we can bound the expectation as

$$\mathbb{E} [\Psi(a_1, \dots, a_m)] \leq \frac{0.025}{b} \quad (\text{B.82})$$

By the bounded differences inequality, with probability $1 - e^{-\Omega(\delta^2)}$,

$$\Psi(a_1, \dots, a_m) \leq \mathbb{E} [\Psi(a_1, \dots, a_m)] + \frac{\delta}{\sqrt{m}} \quad (\text{B.83})$$

Setting $\delta = 0.025\sqrt{\frac{M}{b}}$, we get $\frac{\delta}{\sqrt{m}} = \frac{0.025}{\sqrt{Mb}}\sqrt{\frac{M}{b}} = \frac{0.025}{b}$. This gives

$$\Psi(a_1, \dots, a_m) \leq \frac{0.025}{b} + \frac{0.025}{b} = \frac{0.05}{b}. \quad (\text{B.84})$$

From which we conclude that

$$\Rightarrow \sup_{v \in S_2} \sum_{i=1}^m \mathbf{1} \left\{ |\eta_i a_i^T v| \geq \frac{t}{2} \right\} \leq \frac{0.05m}{b} = 0.05M. \quad (\text{B.85})$$

Now consider any $x \in S_1$. There exists $\tilde{x} \in S_\varepsilon$ such that $\|\tilde{x} - x\| \leq \varepsilon$. From eq. (B.66) there exist a subset $J_{\tilde{x}} \subseteq [M]$ with cardinality $0.95M$ such that

$$\frac{1}{b} |\eta_{B_j}^T A_{B_j} \tilde{x}| \leq \frac{t}{2} \quad \forall j \in J_{\tilde{x}}. \quad (\text{B.86})$$

Similarly, from eq. (B.85), there exists a subset $J_{x-\tilde{x}} \subseteq [M]$ with cardinality $0.95M$ such that for all $j \in J_{x-\tilde{x}}$, we have

$$|\eta_i a_i^T (x - \tilde{x})| \leq \frac{t}{2} \quad \forall i \in B_j, \quad (\text{B.87})$$

$$\Rightarrow \frac{1}{b} |\eta_{B_j}^T A_{B_j} (x - \tilde{x})| \leq \frac{t}{2}. \quad (\text{B.88})$$

From the triangle inequality and a simple union bound, for all $x \in S_1$, there exists a subset $J_x = J_{\tilde{x}} \cap J_{x-\tilde{x}}$ with cardinality $0.9M$ such that

$$\frac{1}{b} |\eta_{B_j}^T A_{B_j} x| \leq \frac{1}{b} |\eta_{B_j}^T A_{B_j} (x - \tilde{x})| + \frac{1}{b} |\eta_{B_j}^T A_{B_j} \tilde{x}| \quad (\text{B.89})$$

$$\leq \frac{t}{2} + \frac{t}{2} = t \quad (\text{B.90})$$

This completes the proof. □

B.5 Proof of Theorem 3.6.5

Proof. In Theorem 3.6.5, we fix the batch size b to be a suitable constant, specified in Lemma 3.6.3, Lemma 3.6.4. Then for $\varepsilon \leq \frac{0.01}{b}$, the number of arbitrarily corrupted samples of A and y are at most $\frac{0.01}{b}bM = 0.01M$. This implies that there exist $0.99M$ batches with uncorrupted samples of A, y . For the rest of the proof, consider only these uncorrupted batches, and ignore the corrupted batches.

For a batch j , define the following

$$\mathbb{Q}_j(\hat{z}, z^*) := \frac{1}{b} \|A_{B_j}(G(\hat{z}) - G(z^*))\|^2, \quad (\text{B.91})$$

$$\mathbb{M}_j(\hat{z}) := \frac{2}{b} \eta_{B_j}^\top (A_{B_j}(G(\hat{z}) - G(z^*))). \quad (\text{B.92})$$

it is easy to verify that $\ell_j(\hat{z}) - \ell_j(z^*) = \mathbb{Q}_j(\hat{z}, z^*) - \mathbb{M}_j(\hat{z})$. The component $\mathbb{Q}_j(\hat{z}, z^*)$ is commonly called the quadratic component, and $\mathbb{M}_j(\hat{z})$ is called the multiplier component.

By Lemma 3.6.2, the minimum value of the MOM objective is at most $4\sigma^2$ with high probability. Since \hat{z} minimizes the objective eq. (3.2) to within additive τ of the optimum, it implies that the median batch satisfies

$$\mathbb{Q}_j(\hat{z}, z^*) - \mathbb{M}_j(\hat{z}) \leq 4\sigma^2 + \tau. \quad (\text{B.93})$$

Using Lemma 3.6.3, Lemma 3.6.4 on the $0.99M$ batches that do not have corruptions, if the batch size is a large enough constant, we see that there exist $0.78M$ batches on which both the following inequalities hold

$$\gamma^2 \|G(\hat{z}) - G(z^*)\|^2 \leq \mathbb{Q}_j(\hat{z}, z^*) \quad \text{and} \quad -\sigma \|G(\hat{z}) - G(z^*)\| \leq -\mathbb{M}_j(\hat{z}). \quad (\text{B.94})$$

Putting the above two inequalities together, the median batch satisfies

$$\gamma^2 \|G(\hat{z}) - G(z^*)\|^2 - \sigma \|G(\hat{z}) - G(z^*)\| \leq 4\sigma^2 + \tau.$$

Solving the quadratic inequality for $\|G(\hat{z}) - G(z^*)\|$, we have

$$\|G(\hat{z}) - G(z^*)\|^2 \lesssim \sigma^2 + \tau. \quad \square$$

B.6 Experimental Setup

B.6.1 MNIST dataset

We first compare Algorithm 1 with the baseline ERM [41] for heavy tailed dataset *without* arbitrary corruptions on MNIST dataset [168]. We trained a DCGAN [225] to produce 64×64 MNIST images.¹ We choose the dimension of the latent space as $k = 100$, and the model has 5 layers.

Based on this generative model, the uncorrupted compressed sensing model P has heavy tailed measurement matrix and stochastic noise: $y = AG(z^*) + \eta$. We consider a Student's t distribution (a typical example of heavy tails) – the measurement matrix A is generated from a Student's t distribution with degrees of freedom 4, and η with degrees of freedom 3 with bounded variance σ^2 . We vary the number of measurement m and obtain the reconstruction error $\|G(\hat{z}) - G(z^*)\|^2$ for Algorithm 1 and ERM, where $G(z^*)$ is the ground truth image. Each curve in Figure 3.1a demonstrates the

¹Code was cloned from the following repository <https://github.com/pytorch/examples/tree/master/dcgan>.

averaged reconstruction error for 50 trials. In Figure 3.1a, Algorithm 1 and ERM both have decreasing reconstruction error per pixel with increasing number of measurement. In particular, Algorithm 1 obtains significantly smaller reconstruction error comparing with the baseline ERM.

B.6.2 CelebA-HQ dataset

We continue the study of empirical performance of our algorithm on real image datasets with higher quality. We generate high quality RGB images with size 256×256 from CelebA-HQ². Hence the dimension of each image is $256 \times 256 \times 3 = 196608$. In all of our experiments, we fix the dimension of the latent space as $k = 512$, and train a DCGAN on this dataset to obtain a generative model G .

We first compare our algorithm with the baseline ERM [41] for heavy tailed dataset without arbitrary corruptions, and then deal with the situation of outliers.

Heavy tailed samples. In this experiment, we deal with the *uncorrupted* compressed sensing model P , which has heavy tailed measurement matrix and stochastic noise: $y = AG(z^*) + \eta$. We also use a Student's t distribution for A and η – the measurement matrix A is generated from a Student's t distribution with degrees of freedom 4, and stochastic noise η with degrees of freedom 3

²Code was cloned from the following repository: https://github.com/facebookresearch/pytorch_GAN_zoo.

with a bounded variance.

We obtain the reconstruction error $\|G(\hat{z}) - G(z^*)\|$ vs. the number of measurement m for our algorithm and ERM, where z^* is the ground truth. In Figure 3.1b, each curve is an average of 20 trials. For heavy tailed y and A without any corruption, both methods are consistent, and have decaying reconstruction error with increasing sample size. Our method obtains significantly smaller reconstruction error, and shows competitive results over the baseline ERM for heavy tailed data set, even without any arbitrary outliers.

B.6.3 Hyperparameter selection

When using the Adam [157] optimizer, we varied the learning rate over $[0.1, 0.05, 0.01, 0.005]$ for our algorithm and baselines. When using the Yellownfin [300] optimizer, we varied our learning rates over $[10^{-4}, 5 \cdot 10^{-5}, 10^{-5}, 5 \cdot 10^{-6}, 10^{-6}]$. We selected the best learning rate based on fresh measurements that were not used for optimization.

B.7 Background

Theorem B.7.1 (Ledoux-Talagrand Contraction Inequality). *For a compact set \mathcal{T} , let x_1, \dots, x_m be i.i.d vectors whose real valued components are indexed by \mathcal{T} , i.e., $x_i = (x_{i,s})_{s \in \mathcal{T}}$. Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be a 1-Lipschitz function such that $\phi(0) = 0$. Let $\epsilon_1, \dots, \epsilon_m$ be independent Rademacher random variables. Then*

$$\mathbb{E} \left[\sup_{s \in \mathcal{T}} \left| \sum_{i=1}^m \epsilon_i \phi(x_{i,s}) \right| \right] \leq 2 \mathbb{E} \left[\sup_{s \in \mathcal{T}} \left| \sum_{i=1}^m \epsilon_i x_{i,s} \right| \right].$$

Theorem B.7.2 (Talagrand's Inequality for Bounded Empirical Processes).

For a compact set \mathcal{T} , let x_1, \dots, x_m be i.i.d vectors whose real valued components are indexed by \mathcal{T} , i.e., $x_i = (x_{i,s})_{s \in \mathcal{T}}$. Assume that $\mathbb{E} x_{i,s} = 0$ and $|x_{i,s}| \leq b$ for all $s \in \mathcal{T}$. Let $Z = \sup_{s \in \mathcal{T}} \left| \frac{1}{m} \sum_{i=1}^m x_{i,s} \right|$. Let $\sigma^2 = \sup_{s \in \mathcal{T}} \mathbb{E} x_s^2$ and $\nu = 2b \mathbb{E} Z + \sigma^2$. Then

$$\Pr [Z \geq \mathbb{E} Z + t] \leq C_1 \exp \left(- \frac{C_2 m t^2}{\nu + bt} \right).$$

where C_1, C_2 are absolute constants.

Appendix C

Appendix for Chapter 4

C.1 Upper Bound Proofs

C.1.1 Proof of Lemma 4.4.1

Lemma 4.4.1. *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 $z^* \sim \text{Bernoulli}(c)$.*

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

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

Proof. We have

$$\Pr_{z^*, y, \hat{z}}[z^* = 0, \hat{z} = 1] = \Pr[z^* = 0] \mathbb{E}_{y \sim h_0, \hat{z}|y}[1\{\hat{z} = 1\}], \quad (\text{C.1})$$

$$= (1 - c) \int h_0(y) \Pr[\hat{z} = 1|y] dy. \quad (\text{C.2})$$

By definition, we have

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

Substituting, we have

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

□

C.1.2 Proof of Lemma 4.4.2

Lemma 4.4.2. *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 η -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:*

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

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

$$\mathbb{E}_A [H_{out}(B)] \leq 2e^{-\frac{m}{2} \log(\frac{c}{4e^2})}, \quad (\text{C.3})$$

$$\mathbb{E}_A [H_{\tilde{x}}(B)] \geq 1 - 2e^{-\frac{m}{2} \log(\frac{c}{4e^2})}. \quad (\text{C.4})$$

Using the above bounds, we can conclude that

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

First we prove Equation (C.3).

Consider the joint distribution of y, A . We have

$$\mathbb{E}_A[H_{out}(B)] = \mathbb{E}_A \left[\mathbb{E}_{x \sim P_{out}} \left[\mathcal{N} \left(Ax, \frac{\sigma^2}{m} I_m \right) (B) \right] \right], \quad (\text{C.5})$$

$$= \mathbb{E}_{x \sim P_{out}} \left[\mathbb{E}_A [\mathcal{N}(Ax, \sigma^2/m)(B)] \right], \quad (\text{C.6})$$

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

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

This implies

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

- 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

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

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 $2(\sqrt{c}-1)^2 > \frac{c}{2}$ if $c \geq 4$.

Substituting the inequalities from Case 1 and Case 2 in Eqn (C.6), we have

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

This proves Eqn (C.3).

A similar proof can be used to show that

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

This proves Eqn (C.4).

Putting the two above inequalities together, we have

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

This concludes the proof. \square

C.1.3 Proof of Lemma C.1.1

Lemma C.1.1. *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 $\min\{\log \text{Cov}_{\eta, \delta}(P), \log \text{Cov}_{\eta, \delta}(R)\} \leq k$, then there exist distributions R', R'', P', P'' , and a finite discrete distribution Q with $|\text{supp}(Q)| \leq e^k$ 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 $\log \text{Cov}_{\eta, \delta}(P) \leq k$. 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}_+$ as

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

By construction, f is a piecewise constant function that is inversely proportional to the number of η -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

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

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

$$c_P = \Pr_{(x_1, x_2) \sim \Pi} [\|x_1 - x_2\| \geq \eta] \leq \delta. \quad (\text{C.7})$$

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

$$c_R = \Pr_{(x_1, x_2) \sim \Gamma} \left[\|x_1 - x_2\| \geq \frac{\rho}{\delta^{1/p}} \right] \leq \frac{\mathbb{E} [\|x_1 - x_2\|^p]}{\frac{\rho^p}{\delta}} \leq \delta. \quad (\text{C.8})$$

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 $E := \{(x, y, z) \mid \|x - z\| \leq \rho/\delta^{1/p} \text{ and } \|x - y\| \leq \eta\}$. To estimate the probability of E , we define $E_1 := \{(x, y, z) \mid z \in \mathbb{R}^n \text{ and } \|x - y\| > \eta\}$ and $E_2 := \{(x, y, z) \mid \|x - z\| > \rho/\delta^{1/p} \text{ and } y \in \mathbb{R}^n\}$. 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 inequality follows from equation (C.7). 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 $P'(\cdot) := \Omega(\cdot, \mathbb{R}^n, \mathbb{R}^n \mid E)$, $Q(\cdot) :=$

$\Omega(\mathbb{R}^n, \cdot, \mathbb{R}^n | E)$ and $R'(\cdot) := \Omega(\mathbb{R}^n, \mathbb{R}^n, \cdot | E)$. P'' and R'' are defined similarly via conditioning on \bar{E} . Hence, $P(\cdot) = \Omega(\cdot, \mathbb{R}^n, \mathbb{R}^n) = \Omega(E)\Omega(\cdot, \mathbb{R}^n, \mathbb{R}^n | E) + \Omega(\bar{E})\Omega(\cdot, \mathbb{R}^n, \mathbb{R}^n | \bar{E}) = (1 - 2\delta)P'(\cdot) + (2\delta)P''(\cdot)$. 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 $\text{supp}(Q) < e^k$, and hence we have our result.

□

C.1.4 Proof of Lemma 4.4.3

Lemma 4.4.3. *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 $\hat{x} \sim P(\cdot|y, A)$ and $\hat{z} \sim P(\cdot|u, A)$. For any $d > 0$, we have

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

Proof. Let B_1 denote the event

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

Similarly, let B_2 denote the event

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

We have

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

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

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

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,

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

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

$$\Pr_{x^*, A, \xi, \hat{x} \sim P(\cdot | A, y)} [B_1] \leq \mathbb{E}_{x^*, z^*} \mathbb{E}_A \left[\mathbb{1}_{\|Ax^* - Az^*\| > 2\varepsilon} \mathbb{E}_{y|A, x^*} \left[\mathbb{E}_{\hat{x} \sim P(\cdot | y, A)} [1_{B_2}] \right] \right] (*) \quad (\text{C.9})$$

$$+ \mathbb{E}_{x^*, z^*} \mathbb{E}_A \left[\mathbb{1}_{\|Ax^* - Az^*\| \leq 2\varepsilon} \mathbb{E}_{y|A, x^*} \left[\mathbb{E}_{\hat{x} \sim P(\cdot | y, A)} [1_{B_2}] \right] \right]. (***) \quad (\text{C.10})$$

Consider the term(*) in line (C.9). We have

$$\mathbb{E}_{x^*, z^*} \mathbb{E}_A \left[\mathbb{1}_{\|Ax^* - Az^*\| > 2\varepsilon} \mathbb{E}_{y|A, x^*} \left[\mathbb{E}_{\hat{x} \sim P(\cdot | y, A)} [1_{B_2}] \right] \right] \leq \mathbb{E}_{x^*, z^*} \mathbb{E}_A \left[\mathbb{1}_{\|Ax^* - Az^*\| > 2\varepsilon} \right], \quad (\text{C.11})$$

$$\leq \mathbb{E}_{x^*, z^*} [e^{-\Omega(m)}] \leq e^{-\Omega(m)}, \quad (\text{C.12})$$

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 (C.10). Notice that since the noise in the measurements is Gaussian, we have

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

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 $\mathcal{N}\left(Ax^*, \frac{\sigma^2}{m} I_m\right)$, by standard Gaussian concentration, we have

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

2. Case 2: $\|y - Ax^*\| \leq 2\sigma$. 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

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

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

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

Substituting the above two inequalities, we have

$$\|Ax^* - y\|^2 \geq -4\varepsilon^2 - 8\sigma\varepsilon + \|y - Az^*\|^2, \quad (\text{C.13})$$

$$\Rightarrow \exp\left(-\frac{\|Ax^* - y\|^2}{2\sigma^2/m}\right) \leq \exp\left(\frac{4\varepsilon(\varepsilon + 2\sigma)m}{2\sigma^2}\right) \exp\left(-\frac{\|Az^* - y\|^2}{2\sigma^2/m}\right), \quad (\text{C.14})$$

$$(\text{C.15})$$

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 posterior sampling now uses u in place of y , we can replace \hat{x} in (**) with \hat{z} .

Combining Case 1 and 2 gives

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

From the above inequality and eqn. (C.12), we have

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

□

C.1.5 Proof of Theorem 4.4.4

Theorem 4.4.4. *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 $m \geq O(\min(\log \text{Cov}_{\eta, \delta}(R), \log \text{Cov}_{\eta, \delta}(P)))$ measurements. Given y and the fixed matrix A , let \hat{x} be the 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 ,

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

Proof. We know from Lemma C.1.1 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

$$\begin{aligned} & \Pr_{x^* \sim R, \hat{x} \sim P(\cdot|y)} [\|x^* - \hat{x}\| \geq (c+1)\eta + (c+1)\sigma] \\ & \leq \Pr_{x^* \sim R, \hat{x} \sim P(\cdot|y)} [\|x^* - \hat{x}\| \geq (c+1)\eta + c\sigma + (\varepsilon/\delta^{1/p})] \\ & \leq 2\delta + (1 - 2\delta) \Pr_{x^* \sim R', \hat{x} \sim P(\cdot|y)} [\|x^* - \hat{x}\| \geq (c+1)\eta + c\sigma + (\varepsilon/\delta^{1/p})], \end{aligned} \quad (\text{C.16})$$

where the first line follows since $\sigma \geq \varepsilon/\delta^{1/p}$, and the second line follows by decomposing $R = (1 - 2\delta)R' + 2\delta R''$.

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, ξ, \hat{x} . By Lemma 4.4.3, we can replace $x^* \sim R'$ with $z^* \sim P'$, replace $y = Ax^* + \xi$ with $u = Az^* + \xi$, and replace $\hat{x} \sim P(\cdot|A, y)$ with $\hat{z} \sim P(\cdot|A, u)$ to get the following bound

$$\begin{aligned} & \Pr_{x^* \sim R', A, \xi, \hat{x} \sim P(\cdot|A, y)} [\|x^* - \hat{x}\| \geq (c+1)\eta + c\sigma + (\varepsilon/\delta^{1/p})] \leq \\ & e^{-\Omega(m)} + e^{\left(\frac{2(\varepsilon/\delta^{1/p})(\varepsilon/\delta^{1/p} + 2\sigma)m}{\sigma^2}\right)} \Pr_{z^* \sim P', A, \xi, \hat{z} \sim P(\cdot|u, A)} [\|z^* - \hat{z}\| \geq (c+1)\eta + c\sigma]. \end{aligned} \quad (\text{C.17})$$

We now bound the second term in the right hand side of the above inequality. Let Γ denote an optimal \mathcal{W}_∞ -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}_∞ condition, each $\Gamma(\cdot|\tilde{z})$ is supported on a ball of radius η around \tilde{z} .

Let $E = \{z^*, \hat{z} \in \mathbb{R}^n : \|z^* - \hat{z}\| \geq (c+1)\eta + c\sigma\}$ denote the event that z^*, \hat{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

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

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

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 $B_{\tilde{z}^*}^c$ be its complement. The constants are defined via the following Lebesgue integrals, and the mixture components for any Borel measurable B

are defined as

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

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

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

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

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

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 \hat{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,

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

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 C.1.2.

Notice that if we sum over all $\tilde{z}^* \in S$, then the LHS of the above

inequality is an expectation over $z^* \sim P'$. This gives:

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

Notice that $P_{\tilde{z}^*, 0}$ is supported within an η -ball around \tilde{z}^* , and $P_{\tilde{z}^*, 2}$ is supported outside a $c(\eta + \sigma)$ -ball of \tilde{z}^* . By Lemma 4.4.2 we have

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

This implies

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

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

Substituting in Eqn (C.17), if $c > 4 \exp\left(2 + \frac{8(\varepsilon/\delta^{1/p})((\varepsilon/\delta^{1/p}) + 2\sigma)}{\sigma^2}\right)$, we have

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

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

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

Substituting in Eqn (C.16), we have

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

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 C.1.2. *Consider the setting of the previous theorem. We have*

$$\mathbb{E}_{A, \xi, z^*} \Pr [z^* \sim P_{\tilde{z}^*, 0}, \hat{z} \sim P_{\tilde{z}^*, 2}(\cdot|u)] \leq \frac{1}{1-\delta_2} \mathbb{E}_A [1 - TV(H_{\tilde{z}^*, 0}, H_{\tilde{z}^*, 2})].$$

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

$$\begin{aligned}
\mathbb{E}_{A,\xi,z^*} \Pr [z^* \sim P_{\tilde{z}^*,0}, \tilde{z} \sim P_{\tilde{z}^*,2}(\cdot|u)] &= \mathbb{E}_A \int \frac{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u) c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)}{(1 - \delta_2) Q_{\tilde{z}^*,0} h_{\tilde{z}^*,0}(u) + c_{\tilde{z}^*,1} h_{\tilde{z}^*,1}(u) + c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)} du, \\
&\leq \mathbb{E}_A \int \frac{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u) c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)}{(1 - \delta_2) Q_{\tilde{z}^*,0} h_{\tilde{z}^*,0}(u) + c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)} du, \\
&\leq \mathbb{E}_A \int \frac{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u) c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)}{(1 - \delta_2) Q_{\tilde{z}^*,0} h_{\tilde{z}^*,0}(u) + (1 - \delta_2) c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)} du, \\
&\leq \mathbb{E}_A \frac{1}{1 - \delta_2} \int \frac{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u) c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)}{Q_{\tilde{z}^*,0} h_{\tilde{z}^*,0}(u) + c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)} du, \\
&\leq \mathbb{E}_A \frac{1}{1 - \delta_2} \int \frac{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u) c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)}{\max\{Q_{\tilde{z}^*,0} h_{\tilde{z}^*,0}(u), c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)\}} du, \\
&= \mathbb{E}_A \frac{1}{1 - \delta_2} \int \min\{Q_{\tilde{z}^*} h_{\tilde{z}^*,0}(u), c_{\tilde{z}^*,2} h_{\tilde{z}^*,2}(u)\} du, \\
&\leq \mathbb{E}_A \frac{1}{1 - \delta_2} \int \min\{h_{\tilde{z}^*,0}(u), h_{\tilde{z}^*,2}(u)\} du, \\
&= \frac{1}{1 - \delta_2} \mathbb{E}_A [1 - TV(H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2})].
\end{aligned}$$

□

C.2 Lower Bound Proofs

C.2.1 Proof of Lemma 4.5.2

Lemma 4.5.2. *Consider the setting of Theorem (4.5.1). 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

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

Proof. First, we consider the case where A is a deterministic matrix.

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 sphere of radius r , we have $\mathbb{E}[z_i^2] = \mathbb{E}[\langle a_i, x^* \rangle^2] \leq \|a_i\|^2 r^2$. Since the Gaussian noise ξ has variance σ^2/m in each coordinate, every coordinate of y_i is a Gaussian channel with power constraint $\|a_i\|^2 r^2$ and noise variance σ^2/m . Using Shannon's AWGN theorem [67, 219, 245], the mutual information between y_i, z_i , is bounded by

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

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

$$\begin{aligned} I(y; z) &= h(y) - h(y|z) = h(y) - h(y - z|z), \\ &= h(y) - h(\xi|z) = h(y) - \sum h(\xi_i|z, \xi_1, \dots, \xi_{i-1}), \\ &= h(y) - \sum h(\xi_i), \\ &\leq \sum h(y_i) - \sum h(\xi_i), \\ &= \sum h(y_i) - \sum h(y_i|z_i), \\ &= \sum I(y_i; z_i), \\ &\leq \sum_{i=1}^m \frac{1}{2} \log \left(1 + \frac{\|a_i\|^2 r^2 m}{\sigma^2} \right), \\ &\leq \frac{m}{2} \log \left(1 + \frac{mr^2 \|A\|_\infty^2}{\sigma^2} \right). \end{aligned}$$

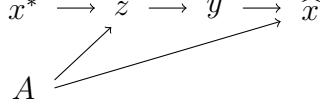


Figure C.1: DAG relating x^* , A , z , y , \hat{x} . The conditional independencies we use are $x^* \perp\!\!\!\perp y|z, A$ and $A \perp\!\!\!\perp y|z$.

Since $x^* \rightarrow z \rightarrow y$ is a Markov chain, we can conclude that

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

Now, if A is a Gaussian matrix with i.i.d. entries drawn from $\mathcal{N}(0, 1/m)$, then the power constraint is $\mathbb{E}[\langle a_i, x \rangle^2] \leq r^2/m$. This gives us

$$I(y; z) \leq \frac{m}{2} \log \left(1 + \frac{r^2}{\sigma^2} \right). \quad (\text{C.18})$$

Now since A is a random matrix, we cannot directly apply the Data Processing Inequality of x^*, y, z as before, and need to 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 C.1, note that x^*, y , are conditionally independent given z, A . This gives $I(x^*; y|z, A) = 0$.

This gives

$$I(x^*; y|A) + I(z; y|x^*, A) = I(z; y|A), \quad (\text{C.19})$$

$$\Rightarrow I(x^*; y|A) \leq I(z; y|A). \quad (\text{C.20})$$

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

$$I(A, z; y) = I(A; y) + I(z; y|A), \quad (\text{C.21})$$

$$= I(z; y) + I(A; y|z), \quad (\text{C.22})$$

$$\Leftrightarrow I(A; y) + I(z; y|A) = I(z; y) + I(A; y|z), \quad (\text{C.23})$$

$$\Leftrightarrow I(A; y) + I(z; y|A) = I(z; y), \quad (\text{C.24})$$

$$\Rightarrow I(z; y|A) \leq I(z; y), . \quad (\text{C.25})$$

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 (C.18), (C.20), (C.25), we have

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

□

C.2.2 Proof of Lemma 4.5.3

Lemma 4.5.3. *Consider the setting of Theorem (4.5.1). If A is a deterministic matrix, we have*

$$I(x^*; \hat{x}) \leq I(y; x^*).$$

If A is a random matrix, then

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

Proof. When A is a deterministic matrix, the proof follows directly from the Data Processing Inequality [67]. Since $x^* \rightarrow y \rightarrow \hat{x}$ is a Markov chain, we get

$$I(x^*; \hat{x}) \leq I(y; x^*).$$

Now when A is a random matrix, we need to show $I(x^*; \hat{x}) \leq I(y; x^*|A)$. Consider the mutual information $I(x^*; y, A, \hat{x})$. By the chain rule of mutual information, we can express it in two ways:

$$I(x^*; y, A, \hat{x}) = I(x^*; y, A) + I(x^*; \hat{x}|y, A), \quad (\text{C.26})$$

$$= I(x^*; \hat{x}) + I(x^*; y, A|\hat{x}). \quad (\text{C.27})$$

As \hat{x} is a function of y, A , we have $I(x^*; \hat{x}|y, A) = 0$. Also, $I(x^*; y, A|\hat{x}) \geq 0$. Substituting in Eqn (C.26), (C.27), we have

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

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

□

C.2.3 Proof of Fano variant Lemma 4.5.4

We will build up Lemma 4.5.4 in sequence. Before showing it in its full generality, we will show when x, \hat{x} , are discrete random variables and x is uniform (Lemma C.2.1). We then lift the uniformity restriction on x (Lemma C.2.3) before extending to continuous distributions (Lemma 4.5.4).

Lemma C.2.1. *Let Q be the uniform distribution over an arbitrary discrete finite set S . Let (x, \hat{x}) be jointly distributed, where $x \sim Q$ and \hat{x} is distributed over an arbitrary countable set, satisfying*

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

Then for all $\tau \in (0, 1)$, we have

$$\tau(1 - \delta) \log \text{Cov}_{2\varepsilon, \delta+\tau}(Q) \leq I(x; \hat{x}) + 2.$$

Proof. Let $E = \mathbf{1}\{\|x - \hat{x}\| \leq \varepsilon\}$ be the indicator random variable for x and \hat{x} being close.

Via claim C.2.2, we get

$$H(x|E = 1) \geq \log |S| - \frac{1}{1 - \delta}. \quad (\text{C.28})$$

Recall,

$$I(x; \hat{x}|E = 1) = H(x|E = 1) - H(x|\hat{x}, E = 1)$$

By the Law of total probability, we have:

$$I(x; \hat{x}|E = 1) = \sum_v \Pr[\hat{x} = v|E = 1] (H(x|E = 1) - H(x|\hat{x} = v, E = 1)).$$

We would like to apply a version of Markov's inequality to the above equation. However, the terms in the summation could be negative. However, from (C.28) we have that $H(x|E = 1) + \frac{1}{1-\delta} \geq \log |S|$. Furthermore, since x is supported on a discrete set of cardinality $|S|$, we have $H(x|\hat{x} = v, E = 1) \leq \log |S|$. Adding and subtracting $\frac{1}{1-\delta}$, in the above equation, we have

$$\begin{aligned} & I(x; \hat{x}|E = 1) \\ &= \sum_v \Pr[\hat{x} = v|E = 1] \left(H(x|E = 1) + \frac{1}{1-\delta} - H(x|\hat{x} = v, E = 1) - \frac{1}{1-\delta} \right), \\ &= \sum_v \Pr[\hat{x} = v|E = 1] \left(H(x|E = 1) + \frac{1}{1-\delta} - H(x|\hat{x} = v, E = 1) \right) - \frac{1}{1-\delta}, \\ &\Leftrightarrow I(x; \hat{x}|E = 1) + \frac{1}{1-\delta} = \\ & \quad \sum_v \Pr[\hat{x} = v|E = 1] \left(H(x|E = 1) + \frac{1}{1-\delta} - H(x|\hat{x} = v, E = 1) \right) \end{aligned}$$

Since the above summation has only non-negative terms that average to $I(x; \hat{x}|E = 1) + \frac{1}{1-\delta}$, for all $\tau \in (0, 1)$, there exists $G_1 \subseteq \text{supp}(\hat{x})$ with $\Pr[G_1|E = 1] \geq 1 - \tau$, such that for all $v \in G_1$, we have

$$H(x|E = 1) + \frac{1}{1-\delta} - H(x|\hat{x} = v, E = 1) \leq \frac{I(x; \hat{x}|E = 1) + \frac{1}{1-\delta}}{\tau}.$$

From (C.28), we have $H(x|E = 1) + \frac{1}{1-\delta} \geq \log |S|$. Hence for all $v \in G_1$,

we have

$$\begin{aligned}
\log |S| - H(x|\hat{x} = v, E = 1) &\leq \frac{I(x; \hat{x}|E = 1) + \frac{1}{1-\delta}}{\tau}, \\
\Leftrightarrow H(x|\hat{x} = v, E = 1) &\geq \log |S| - \frac{I(x; \hat{x}) + \frac{1}{1-\delta}}{\tau}, \\
\Rightarrow \log |\text{supp}(x|\hat{x} = v, E = 1)| &\geq \log |S| - \left(\frac{I(x; \hat{x}) + \frac{1}{1-\delta}}{\tau} \right), \\
\Rightarrow \log |S \cap B(v, \varepsilon)| &\geq \log |S| - \left(\frac{I(x; \hat{x}) + \frac{1}{1-\delta}}{\tau} \right),
\end{aligned} \tag{C.29}$$

where the last inequality follows as conditioned on $E = 1$, x must be supported on an ε -radius ball around \hat{x} .

Now consider the set $G_2 = (S \times G_1) \wedge E_1$. That is, $G_2 \subseteq \text{supp}(x, \hat{x})$, such that $(u, v) \in G_2$ if and only if $\|u - v\| \leq \varepsilon$ and $u \in S, v \in G_1$. Since $\Pr[E_1] \geq 1 - \delta$ by the statement of the lemma, and $\Pr[G_1|E_1] \geq 1 - \tau$ by construction, we have

$$\Pr[G_2] \geq (1 - \delta)(1 - \tau) \geq 1 - \delta - \tau.$$

Now for all $(u, v) \in G_2$, we have

$$\begin{aligned}
\|u - v\| &\leq \varepsilon, \\
\log |S \cap B(v, \varepsilon)| &\geq \log |S| - \left(\frac{I(x; \hat{x}|E = 1) + \frac{1}{1-\delta}}{\tau} \right).
\end{aligned} \tag{C.30}$$

Note that by the construction of G_2 , the set $\bigcup_{v \in G_2} B(v, \varepsilon)$ covers a $1 - \delta - \tau$ fraction of S . As each ball $B(v, \varepsilon)$ also has a large intersection with S , by the pigeon-hole principle, any 2ε -packing of this $1 - \delta - \tau$ fraction of S must have size at most $2^{(I(x; \hat{x}|E = 1) + \frac{1}{1-\delta})/\tau}$.

Hence, we can find a 2ε -cover of a $1 - \delta - \tau$ fraction of S that has size at most $2^{(I(x; \hat{x}|E=1) + \frac{1}{1-\delta})/\tau}$.

This gives

$$\log \text{Cov}_{2\varepsilon, \delta+\tau}(Q) \leq \frac{I(x; \hat{x}|E=1) + \frac{1}{1-\delta}}{\tau}. \quad (\text{C.31})$$

We are almost done, since we now only need to relate $I(x; \hat{x}|E=1)$ to $I(x; \hat{x})$.

By the chain rule of mutual information, we have

$$\begin{aligned} I(x; \hat{x}, E) &= I(x; \hat{x}) + I(x; E|\hat{x}) = I(x; E) + I(x; \hat{x}|E), \\ \Rightarrow I(x; \hat{x}|E) &\leq I(x; \hat{x}) + I(x; E|\hat{x}), \\ &\leq I(x; \hat{x}) + 1, \\ \Leftrightarrow I(x; \hat{x}|E=0) \Pr[E=0] + I(x; \hat{x}|E=1) \Pr[E=1] &\leq I(x; \hat{x}) + 1, \\ \Rightarrow I(x; \hat{x}|E=1) &\leq \frac{I(x; \hat{x}) + 1}{1 - \delta}. \end{aligned}$$

Substituting in Eqn (C.31), we have

$$\begin{aligned} \log \text{Cov}_{2\varepsilon, \delta+\tau}(Q) &\leq \frac{I(x; \hat{x}) + 2}{\tau(1 - \delta)}, \\ \Rightarrow \tau(1 - \delta) \log \text{Cov}_{2\varepsilon, \delta+\tau}(Q) &\leq I(x; \hat{x}) + 2. \end{aligned}$$

□

Claim C.2.2. *Let $x \sim Q$, where Q is the uniform distribution over an arbitrary discrete finite set S . Let E be a binary random variable such that $\Pr[E=1] \geq 1 - \delta$.*

Then we have

$$H(x|E = 1) \geq \log |S| - \frac{1}{1 - \delta}.$$

Proof. Let $p = \Pr[E = 1]$. By the definition of conditional entropy, we have

$$\begin{aligned} H(x|E) &= (1 - p)H(x|E = 0) + pH(x|E = 1), \\ \Leftrightarrow H(x|E = 1) &= \frac{1}{p}(H(x|E) - (1 - p)H(x|E = 0)), \\ &= \frac{1}{p}(H(x) - I(x; E) - (1 - p)H(x|E = 0)), \\ &= \frac{1}{p}(\log |S| - I(x; E) - (1 - p)H(x|E = 0)), \\ &\geq \frac{1}{p}(\log |S| - I(x; E) - (1 - p)\log |S|), \\ &= \log |S| - \frac{I(x; E)}{p}, \\ &\geq \log |S| - \frac{1}{1 - \delta}, \end{aligned}$$

where the fourth line follows from $H(x) = \log |S|$ since x is uniform, the fifth line follows from $H(x|E = 0) \leq \log |S|$ since x is supported on a discrete set of size $|S|$, and the last line follows from $p \geq 1 - \delta$ and $I(x; E) \leq H(E) \leq 1$. \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 C.2.3. *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, \hat{x}) be jointly distributed, where $x \sim Q$ and \hat{x} is distributed over a countable set, satisfying

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

Then we have

$$\tau(1 - \delta) \log \text{Cov}_{2\varepsilon, \tau+\delta}(Q) \leq I(x; \hat{x}) + 2\delta.$$

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 $x' = (x, j)$, $x \in \text{supp}(Q)$, $j \in [N_2]$ 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 C.2.1 on Q' , it suffices to show that $I(x; \hat{x}) = I(x'; \hat{x})$.

By the chain rule of mutual information, we have

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

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

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

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

$$\|x' - \hat{x}'\| \leq \varepsilon \text{ w.p. } 1 - \delta,$$

$$I(x'; \hat{x}') = I(x; \hat{x})$$

Applying Lemma C.2.1 on Q' , we have

$$\tau(1 - \delta) \log \text{Cov}_{2\varepsilon, \tau+\delta}(Q') \leq I(x; \hat{x}) + 2.$$

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

$$\tau(1 - \delta) \log \text{Cov}_{2\varepsilon, \tau+\delta}(Q) \leq I(x; \hat{x}) + 2.$$

□

We now prove Lemma 4.5.4, which allows (x, \hat{x}) to follow an arbitrary distribution.

Lemma 4.5.4 (Fano variant). *Let (x, \hat{x}) be jointly distributed over $\mathbb{R}^n \times \mathbb{R}^n$, where $x \sim R$ and \hat{x} satisfies*

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

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

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

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 $x, \hat{x} \in \mathbb{R}^n$, let $\bar{x}, \hat{\bar{x}}$ denote the rounding of x, \hat{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, \hat{x} in the following way. If $\|x - \hat{x}\| \leq \varepsilon$ with probability $1 - \delta$, an application of the triangle inequality gives

$$\|\bar{x} - \hat{\bar{x}}\| \leq \varepsilon + 2\gamma\sqrt{n} \text{ with probability } \geq 1 - \delta. \quad (\text{C.32})$$

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

$$I(\bar{x}; \hat{\bar{x}}) \leq I(x; \hat{x}). \quad (\text{C.33})$$

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 $\bar{R} = (1 - c_1)Q + c_1D$, with $c_1 \leq \beta$ and D is some other probability distribution. This is feasible because the probabilities of the infinite support of \bar{R} must sum to 1, and hence we can find a finite subset that sums to at least $1 - \beta$ 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 $\hat{\bar{x}}|\bar{x}$. Let $I(\bar{x}; \hat{\bar{x}})$, $I_Q(\bar{x}; \hat{\bar{x}})$, $I_D(\bar{x}; \hat{\bar{x}})$ denote the mutual information between $\bar{x}, \hat{\bar{x}}$ when the marginal of \bar{x} is \bar{R}, Q, D ,

respectively. From Theorem 2.7.4 in [67], mutual information is a concave function of the marginal distribution of \bar{x} for a fixed conditional distribution of $\widehat{x}|\bar{x}$. An application of Eqn (C.33) gives us,

$$I(x; \widehat{x}) \geq I(\bar{x}; \widehat{x}) \geq (1 - c_1)I_Q(\bar{x}; \widehat{x}) + c_1I_D(\bar{x}; \widehat{x}), \quad (\text{C.34})$$

$$\geq (1 - c_1)I_Q(\bar{x}; \widehat{x}), \quad (\text{C.35})$$

$$\geq (1 - \beta)I_Q(\bar{x}; \widehat{x}). \quad (\text{C.36})$$

Now since the finite distribution Q has a TV distance of at most β to the countable distribution R , using Eqn (C.32), we have

$$\|\bar{x} - \widehat{x}\| \leq \varepsilon + 2\gamma\sqrt{n} \text{ with probability } \geq 1 - \beta - \delta \text{ if } \bar{x} \sim Q. \quad (\text{C.37})$$

In order to apply Lemma C.2.3 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 cardinality of the support of Q . This gives

$$Q = (1 - c_2)Q' + c_2Q'', \quad c_2 \leq \alpha N_\beta.$$

From Eqn (C.37), we have $\|\bar{x} - \widehat{x}\| \leq \varepsilon + 2\gamma\sqrt{n}$ 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

$$\|\bar{x} - \widehat{x}\| \leq \varepsilon + 2\gamma\sqrt{n} \text{ with probability } \geq 1 - \beta - \delta - \alpha N_\beta \text{ if } \bar{x} \sim Q'. \quad (\text{C.38})$$

Let $I_Q(\bar{x}; \widehat{x}), I_{Q'}(\bar{x}; \widehat{x}), I_{Q''}(\bar{x}; \widehat{x})$ denote the mutual information between \bar{x}, \widehat{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 $\widehat{x}|\bar{x}$. Hence using Eqn (C.36), we have

$$\frac{I(x; \widehat{x})}{1 - \beta} \geq I_Q(\bar{x}; \widehat{x}) \geq (1 - c_2)I_{Q'}(\bar{x}; \widehat{x}) + c_2I_{Q''}(\bar{x}; \widehat{x}), \quad (\text{C.39})$$

$$\geq (1 - c_2)I_{Q'}(\bar{x}; \widehat{x}), \quad (\text{C.40})$$

$$\geq (1 - \alpha N_\beta)I_{Q'}(\bar{x}; \widehat{x}). \quad (\text{C.41})$$

Hence if $\bar{x} \sim Q'$, we have $I(\bar{x}; \widehat{x}) \leq \frac{I(x; \widehat{x})}{(1 - \alpha N_\beta)(1 - \beta)}$. Applying Lemma C.2.3 on the distribution Q' , for any $\tau > 0$, we have

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

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

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

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 $2\varepsilon + 4\gamma\sqrt{n}$ for \bar{R} by $2\varepsilon + 6\gamma\sqrt{n}$ for R . This gives

$$\tau(1 - \beta - \delta - \alpha N_\beta) \log \text{Cov}_{2\varepsilon+6\gamma\sqrt{n}, \tau+\beta+2\delta+2\alpha N_\beta}(R) \leq \frac{I(x; \hat{x})}{(1 - \alpha N_\beta)(1 - \beta)} + 2.$$

For $\gamma = \frac{\varepsilon}{6\sqrt{n}}$, $\beta = \min\left\{\frac{\delta}{3}, 1 - \sqrt{0.99}\right\}$, $\alpha N_\beta = \min\left\{\frac{\delta}{3}, 1 - \sqrt{0.99}\right\}$, we have

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

□

C.2.4 Proof of Theorem 4.5.1

Theorem 4.5.1. *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

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

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:

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

Proof. Throughout the proof, we use the notation $N(R, \delta)$ to denote a minimal set of 3η -radius balls that cover at least $1 - \delta$ mass under the distribution R .

Let B be the ball in $N(R, 10\delta)$ with smallest marginal probability. If we set $S \leftarrow N(R, 10\delta) \setminus B$, then S contains smaller than $1 - 10\delta$ of R .

Let $R = (1 - c)R' + cR''$, where the components R' and R'' are probability distributions restricted to S and its complement S^c respectively. By the construction of S , we have $c > 10\delta$. Note that since R'' contributes at least 10δ to R , any algorithm that succeeds with probability $\geq 1 - \delta$ over R must succeed with probability ≥ 0.9 over R'' .

Now consider $x \sim R''$. By Lemma 4.5.3 and Lemma 4.5.2, we have

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

Applying Lemma 4.5.4 on R'' with parameters $\tau = \delta = 0.1$, for the failure probability, we can conclude that

$$\begin{aligned} 0.99 \cdot 0.1 \cdot (1 - 0.2) \log |N(R'', 0.4)| &\leq I(x; \hat{x}) + 1.98 \leq \frac{m}{2} \log \left(1 + \frac{r^2}{\sigma^2} \right) + 1.98, \\ \Leftrightarrow m &\geq \frac{0.1584 \log |N(R'', 0.4)| - 3.96}{\log \left(1 + \frac{r^2}{\sigma^2} \right)}. \quad (\text{C.42}) \end{aligned}$$

We now need to express the covering number of R'' in terms of the covering number of R .

Note that as R'' contains at least 10δ mass under R , $N(R'', 0.4)$ contains at least 6δ mass under R . Similarly, since $N(R, 10\delta)$ contains at least $1 - 10\delta$

mass under R , $N(R'', 0.4) \cup N(R, 10\delta)$ will contain at least 4δ mass under R .

Hence, we get

$$|N(R'', 0.4)| + |N(R, 10\delta)| \geq |N(R, 4\delta)| \Leftrightarrow |N(R'', 0.4)| \geq |N(R, 4\delta)| - |N(R, 10\delta)|. \quad (\text{C.43})$$

Now we need to relate $N(R, 4\delta)$ with $N(R, 10\delta)$. This can be accomplished via a simple counting argument. Assume that the balls in $N(R, 4\delta)$ are ordered in decreasing order of their marginal probability, then the last $\frac{10\delta}{1-4\delta}$ -fraction of balls in $N(R, 4\delta)$ must contain at most 10δ mass. This implies that the first $\frac{1-10\delta}{1-4\delta}$ -fraction of $N(R, 4\delta)$ must contain at least $1 - 10\delta$ mass. This gives:

$$\frac{1 - 10\delta}{1 - 4\delta} N(R, 4\delta) \geq N(R, 10\delta). \quad (\text{C.44})$$

Combining Eqn (C.43), (C.44), we get

$$\begin{aligned} |N(R'', 0.4)| &\geq |N(R, 4\delta)| - \frac{1 - 10\delta}{1 - 4\delta} |N(R, 4\delta)|, \\ &= \frac{6\delta}{1 - 4\delta} |N(R, 4\delta)|, \\ &\geq 6\delta |N(R, 4\delta)|, \end{aligned}$$

$$\Leftrightarrow \log |N(R'', 0.4)| \geq \log |N(R, 4\delta)| + \log(6\delta).$$

Substituting in Eqn (C.42), we get

$$m \geq \frac{0.1584 (\log |N(R, 4\delta)| + \log(6\delta)) - 3.96}{\log \left(1 + \frac{r^2}{\sigma^2}\right)}.$$

Since $|N(R, 4\delta)| = \text{Cov}_{3\eta, 4\delta}(R)$ by definition, this completes the proof. \square

C.3 Experimental Setup

C.3.1 Datasets and Architecture

For the compressed sensing experiment in Fig 4.4a and the inpainting experiment in Figure 4.2 we used the 256×256 GLOW model [159] from the official repository. The test set for Fig 4.4a consists of the first 10 images used by [20] in their experiments.

For the compressed sensing experiment in Fig 4.1, 4.5a, 4.5b, we used the FFHQ NCSNv2 model [254] from the official repository. The test set for Fig 4.5a 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 4.4a and Fig 4.5a, the measurements have noise satisfying $\sqrt{\mathbb{E} \|\xi\|^2} = 16$ and $\sqrt{\mathbb{E} \|\xi\|^2} = 4$ respectively.

C.3.2 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 (4.6) 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 sched-

ule proposed by [253]. 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 [253].

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 [254], 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 [20], 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.3.3 Computing Infrastructure

Experiments were run on an NVIDIA Quadro P5000.

Appendix D

Appendix for Chapter 5

D.1 FFHQ Experiments



Figure D.1: Super-resolution reconstructions on faces 69000-69004 from the FFHQ dataset. The top row shows original images, the second row shows what the algorithms observe: blurry measurements after downsample by $32\times$ in each dimension. The third row shows reconstructions by PULSE, and the last row shows reconstructions by Posterior Sampling via Langevin dynamics, the algorithm we are advocating for.

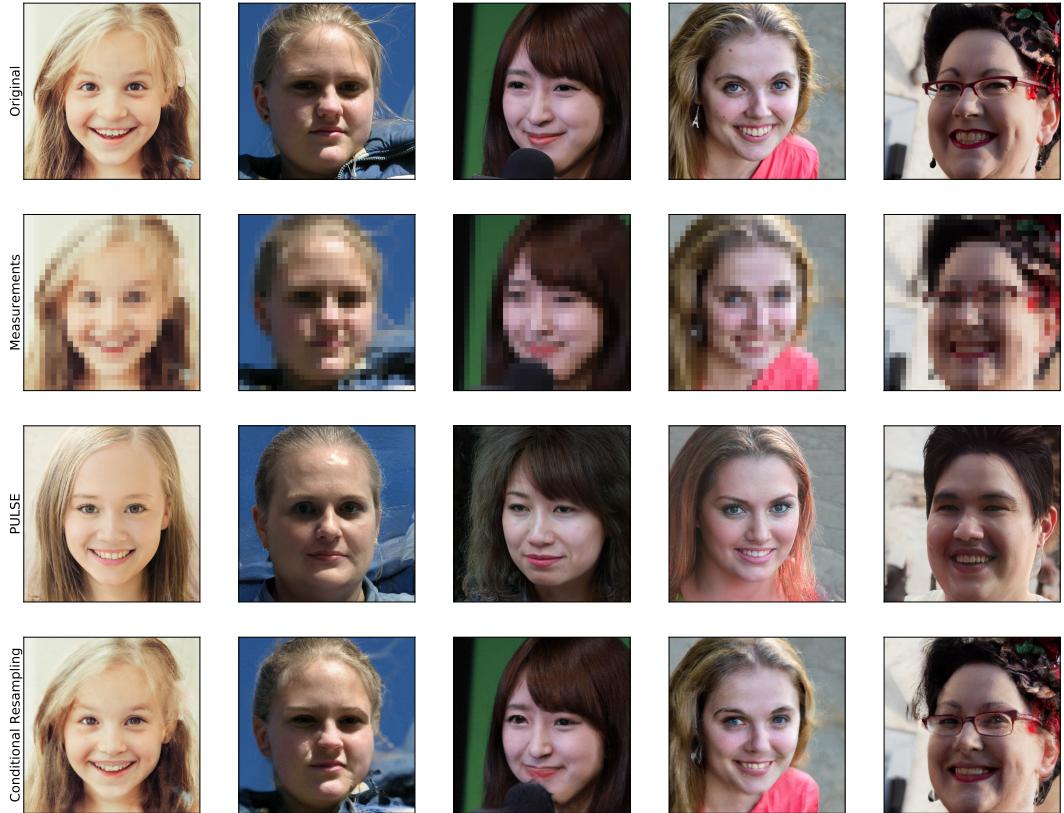


Figure D.2: Super-resolution reconstructions on faces 69005-69009 from the FFHQ dataset. The top row shows original images, the second row shows what the algorithms observe: blurry measurements after downsampling by $32\times$ in each dimension. The third row shows reconstructions by PULSE, and the last row shows reconstructions by Posterior Sampling via Langevin dynamics, the algorithm we are advocating for.

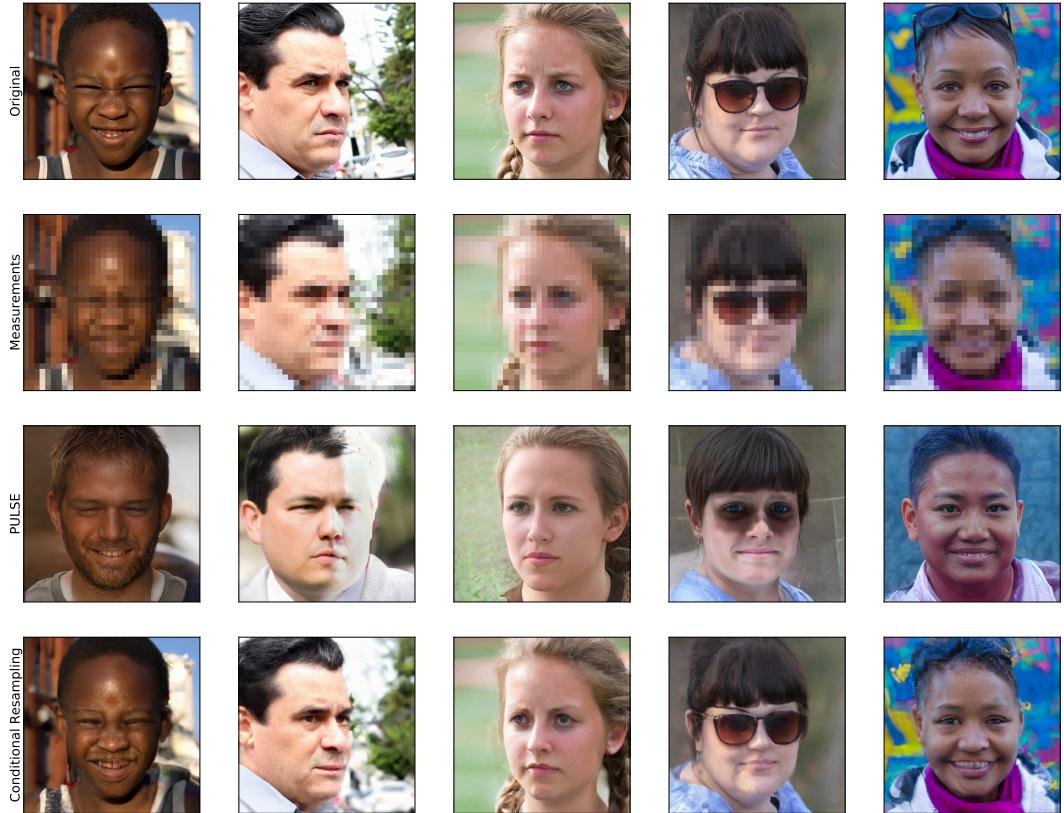


Figure D.3: Super-resolution reconstructions on faces 69010-69014 from the FFHQ dataset. The top row shows original images, the second row shows what the algorithms observe: blurry measurements after downsampling by $32\times$ in each dimension. The third row shows reconstructions by PULSE, and the last row shows reconstructions by Posterior Sampling via Langevin dynamics, the algorithm we are advocating for.

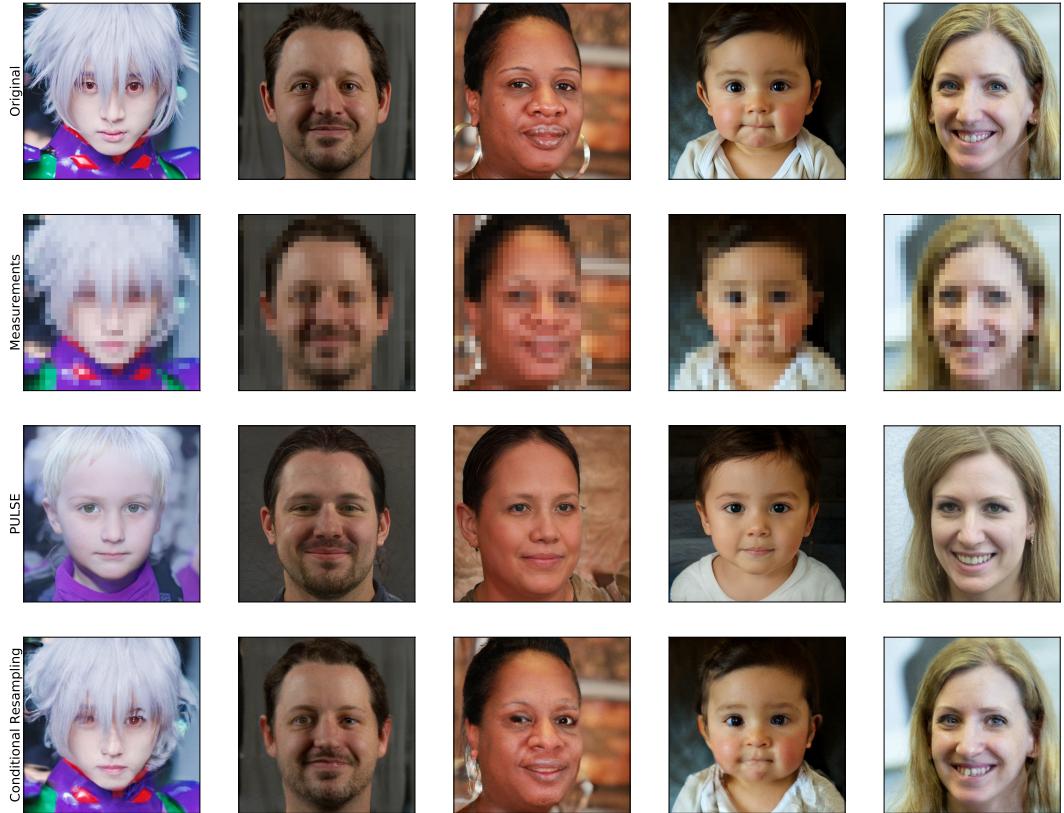
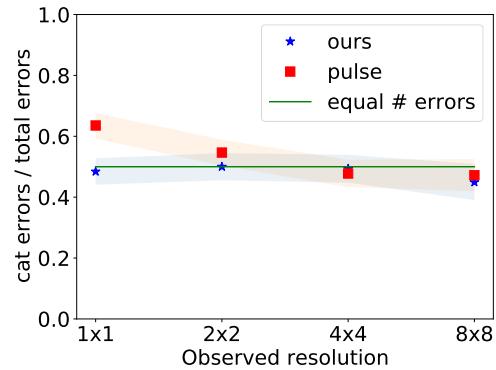


Figure D.4: Super-resolution reconstructions on faces 69015-69020 from the FFHQ dataset. The top row shows original images, the second row shows what the algorithms observe: blurry measurements after downsampling by $32\times$ in each dimension. The third row shows reconstructions by PULSE, and the last row shows reconstructions by Posterior Sampling via Langevin dynamics, the algorithm we are advocating for.

m	PULSE		Ours	
	Cats	Dogs	Cats	Dogs
1 × 1	319	183	245	261
2 × 2	282	234	239	239
4 × 4	225	246	223	229
8 × 8	160	179	119	146

(a) Number of errors. Test set has **500** cats and **500** dogs



(b) Fraction of all errors on cats for 50% cat generator.

Figure D.5: We use a StyleGAN2 model trained on 50% cats and report errors when reconstructing images from low-resolution measurements. The test set consists of 500 cats and 500 dogs from the AFHQ validation set to mimick the generator’s training distribution (note that these correspond to all cats and dogs in the AFHQ validation set). Figure (b) shows the proportion of all errors that are on cats, along with 95% confidence intervals from a binomial test. An algorithm that satisfies SPE would have this probability=0.5 (green line). In this case where the generator is balanced, Posterior Sampling via Langevin dynamics appears to achieve SPE, PR, and RDP. PULSE also appears to satisfy SPE, PR, and RDP, except when the resolution of measurements is 1×1 .

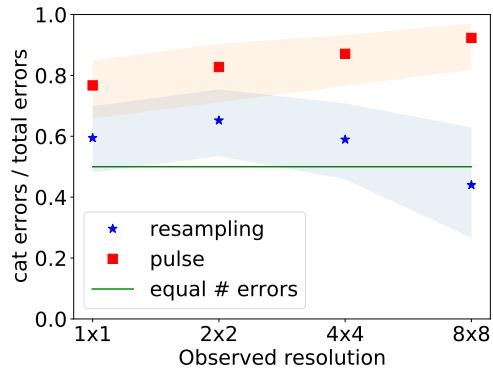
D.2 AFHQ Experiments

D.2.1 50% cat generator

For this experiment, we draw x^* from the validation set of the AFHQ dataset which contains 500 images of cats + 500 images of dogs. We use a generator trained on 50% cats and 50% dogs, and use it to study whether posterior sampling and PULSE satisfy RDP, SPE, and PR in practice. These results are in Figure D.5.

m	PULSE		Ours	
	Cats	Dogs	Cats	Dogs
1×1	56	17	44	30
2×2	48	10	45	24
4×4	54	8	33	23
8×8	48	4	11	14

(a) Number of errors on 20% cat generator, for each resolution. Sampled test set has **60** cats and **140** dogs. PULSE makes errors on almost all the cats and relatively few dogs, while Posterior Sampling is relatively balanced.



(b) Binomial hypothesis test for Symmetric Pairwise Error (SPE)

Figure D.6: We sample 200 images from a StyleGAN2 model trained on 20% cats, and report errors when reconstructing them from low-resolution measurements. Figure (b) shows the proportion of all errors that are on cats, along with 95% confidence intervals from a binomial test. An algorithm that satisfies SPE would have this probability=0.5 (green line). PULSE is clearly biased towards the majority, while Posterior Sampling via Langevin dynamics appears to satisfy SPE (except when $m = 2 \times 2$, but one failure is unsurprising as we are performing sequential hypothesis tests.)

D.2.2 x^* drawn from generator

In Figure D.6, we show results when 200 images drawn from the 20% cat generator are reconstructed.

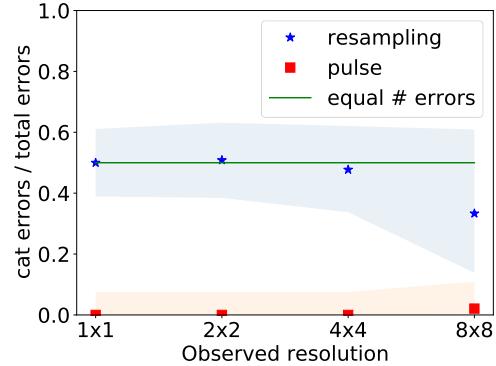
In Figure D.7, we show results when 200 images drawn from the 80% cat generator are reconstructed.

D.2.3 Varying training bias

We train StyleGAN2 models with 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% cats, and report the fraction of errors on cats when tested on

m	PULSE		Ours	
	Cats	Dogs	Cats	Dogs
1×1	0	47	37	37
2×2	0	47	30	29
4×4	0	47	21	23
8×8	1	47	4	8

(a) Number of errors on 80% cat generator, for each resolution. Sampled test set has **153** cats and **47** dogs. PULSE makes errors on almost all the cats and relatively few dogs, while posterior sampling is relatively balanced.



(b) Binomial hypothesis test for Symmetric Pairwise Error (SPE)

Figure D.7: We sample 200 images from a StyleGAN2 model trained on 80% cats, and report errors when reconstructing them from low-resolution measurements. Figure (b) shows the proportion of all errors that are on cats, along with 95% confidence intervals from a binomial test. An algorithm that satisfies SPE would have this probability=0.5 (green line). PULSE is clearly biased towards the majority, while posterior sampling via Langevin dynamics appears to satisfy SPE.

the AFHQ validation set. The results are in Figure D.8.

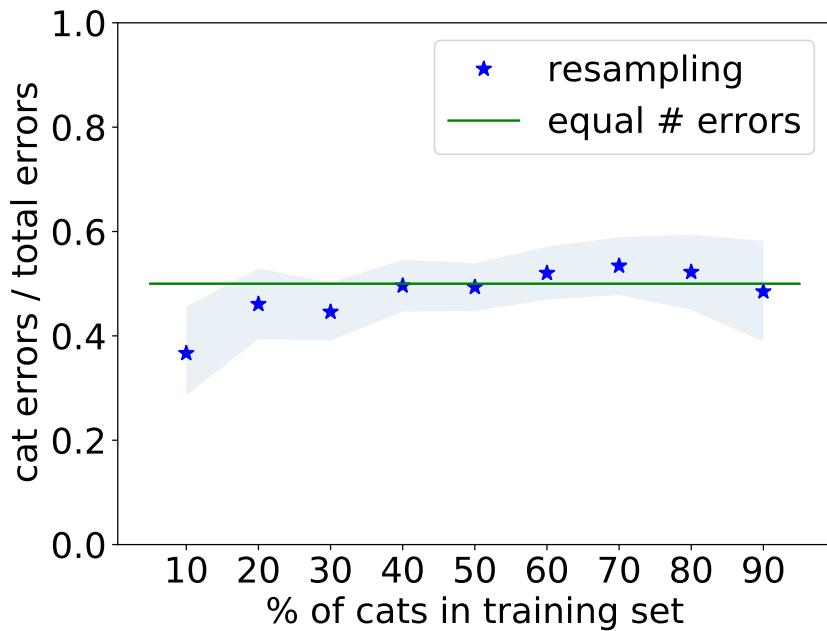


Figure D.8: We train StyleGAN2 generators of varying bias and test SPE. The ground truth images are from the validation set, the observed measurements have resolution 4×4 . Shaded areas denote 95% confidence intervals. We see that Posterior Sampling satisfies SPE. Note that the single failure in the 10% cat generator is not surprising as we are running sequential hypothesis tests on non-independent data.

D.3 Proofs

Theorem 5.3.3. *Let A and B be disjoint groups (e.g., Asian and White people), and let $A_1, A_2 \subset A$ be disjoint groups that cannot be perfectly distinguished from measurements only (e.g., South Asians and East Asians). Then Representation Demographic Parity cannot be satisfied $\{\{A, B\}, \{A_1, A_2, B\}\}$ -obliviously.*

Proof. Let $A = A_1 \cup A_2$. We write $p_a = \Pr(x^* \in a)$, $q_{ab} = \Pr(\hat{x} \in b | x^* \in a)$. Using Representation Demographic Parity, with respect first to $\{A, B\}$, then to $\{A_1, A_2, B\}$, we have:

$$q_{AA} = q_{BB}$$

$$q_{A_1 A_1} = q_{A_2 A_2} = q_{BB}$$

Since $A = A_1 \cup A_2$:

$$q_{AA} = \frac{p_{A_1}(q_{A_1 A_1} + q_{A_1 A_2}) + p_{A_2}(q_{A_2 A_1} + q_{A_2 A_2})}{p_{A_1} + p_{A_2}}$$

Writing $0 < \frac{p_{A_1}}{p_{A_1} + p_{A_2}} = \alpha < 1$, and replacing q_{AA} , $q_{A_1 A_1}$ and $q_{A_2 A_2}$ by q_{BB} , we have:

$$\begin{aligned} q_{AA} &= \alpha(q_{A_1 A_1} + q_{A_1 A_2}) + (1 - \alpha)(q_{A_2 A_1} + q_{A_2 A_2}) \\ q_{BB} &= \alpha(q_{BB} + q_{A_1 A_2}) + (1 - \alpha)(q_{BB} + q_{A_2 A_2}) \\ 0 &= \alpha q_{A_1 A_2} + (1 - \alpha) q_{A_2 A_2}. \end{aligned}$$

Therefore, an algorithm can satisfy Representation Demographic Parity $\{\{A_1 \cup A_2, B\}, \{A_1, A_2, B\}\}$ -obliviously if and only if there exists no confusion between A_1 and A_2 , i.e. $q_{A_1 A_2} = 0 = q_{A_2 A_1}$. \square

Theorem 5.3.4 (Representation Demographic Parity cannot be satisfied obliviously). *The only way for an algorithm to satisfy Representation Demographic Parity obviously is to achieve perfect reconstruction.*

Proof. Suppose there exists x such that $\Pr(x) > 0$, and $x_1 \neq x$ such that $\Pr(\hat{x} = x_1) > 0$. Let us split the space into two groups A and B , such that both x and x_1 belong in A . We now further split A into A_1 and A_2 , such that x_1 belongs in A_1 , and x belongs in A_2 . A_1 and A_2 now are not perfectly distinguishable, so using the claim above, Representation Demographic Parity is not satisfiable $\{\{A_1 \cup A_2, B\}, \{A_1, A_2, B\}\}$ -obviously, so it cannot be satisfiable obviously. \square

Proposition 5.3.7. *Whenever there exists a majority class that the measurements cannot 100% distinguish from the non-majority classes, PR and RDP are not simultaneously achievable.*

Proof. Suppose towards a contradiction that both PR and RDP hold, and the distribution is such that

$$\Pr(x^* \in c_1) > \frac{1}{2} > \sum_{i \neq 1} \Pr(x^* \in c_i).$$

Since PR holds, $\Pr(\hat{x} \in c_1) = \Pr(x^* \in c_1)$. However, since RDP holds and the algorithm does not reconstruct each class perfectly we have $\alpha = \Pr(\hat{x} \in c_i |$

$x^* \in c_i) < 1$ for all the i . We now observe the following contradiction.

$$\begin{aligned}\Pr(\hat{x} \in c_1) &\leq \sum_i \Pr(\hat{x} \in c_1 \mid x^* \in c_i) \Pr(x^* \in c_i) \\ &\leq (1 - \alpha) \sum_{i \neq 1} \Pr(x^* \in c_i) + \alpha \Pr(x^* \in c_1) \\ &< (1 - \alpha) \Pr(x^* \in c_1) + \alpha \Pr(x^* \in c_1) \\ &= \Pr(x^* \in c_1).\end{aligned}$$

□

Theorem 5.4.1. *Posterior Sampling is the only algorithm that achieves oblivious Conditional Proportional Representation.*

Proof. Let \mathcal{A} denote a reconstruction algorithm. Given measurements y , let $Q(U|y)$ denote the probability that the reconstruction from algorithm \mathcal{A} lies in the measurable set U .

If \mathcal{A} satisfies CPR, then for all measurable $U \subset \mathbb{R}^n$, and all $y \in \mathbb{R}^m$, we have

$$Q(U|y) = P(U|y).$$

By the definition of the total variation distance, we have

$$TV(Q(\cdot|y), P(\cdot|y)) = \sup_{U \in \mathcal{B}(\mathbb{R}^n)} |Q(U|y) - P(U|y)|.$$

Since we have $Q(U|y) = P(U|y)$ for all measurable $U \in \mathcal{B}(\mathbb{R}^n)$ and almost all measurements $y \in \mathbb{R}^m$, we have $TV(Q(\cdot|y), P(\cdot|y)) = 0$ for almost all $y \in \mathbb{R}^m$.

This shows that the output distribution of \mathcal{A} must exactly match the posterior distribution $P(\cdot|y)$, and hence posterior sampling is the only algorithm that can satisfy obliviousness and CPR. \square

Theorem 5.4.3. *In the setting of Definition 5.3.1, Conditional Proportional Representation implies Symmetric Pairwise Error.*

Proof. We want to show that if $\Pr(\hat{x} \in c_i|y) = \Pr(x^* \in c_i|y), \forall c_i \in C$, for almost all $y \in \mathbb{R}^m$, then we have $\Pr(\hat{x} \in c_i, x^* \in c_j) = \Pr(\hat{x} \in c_j, x^* \in c_i), \forall c_i, c_j \in C$.

Consider the term $\Pr(\hat{x} \in c_i, x^* \in c_j)$. We can write this as an average over y , to get:

$$\Pr(\hat{x} \in c_i, x^* \in c_j) = \mathbb{E}_y \Pr(\hat{x} \in c_i, x^* \in c_j|y).$$

Note that \hat{x} & x^* are conditionally independent given y . This is because \hat{x} is purely a function of y . This gives

$$\Pr(\hat{x} \in c_i, x^* \in c_j) = \mathbb{E}_y [\Pr(\hat{x} \in c_i|y) \Pr(x^* \in c_j|y)].$$

If we have CPR with respect to c_i and c_j , then we can rewrite the above equation as

$$\Pr(\hat{x} \in c_i, x^* \in c_j) = \mathbb{E}_y [\Pr(x^* \in c_i|y) \Pr(\hat{x} \in c_j|y)].$$

Using the conditional independence of x^* , \hat{x} given y , we now have

$$\begin{aligned} \Pr(\hat{x} \in c_i, x^* \in c_j) &= \mathbb{E}_y \Pr(x^* \in c_i, \hat{x} \in c_j|y), \\ &= \Pr(\hat{x} \in c_j, x^* \in c_i). \end{aligned}$$

This completes the proof. \square

Corollary 5.4.4. *Posterior Sampling achieves symmetric pairwise error for any pair of sets $U, V \subset \mathbb{R}^n$.*

Proof. The proof follows directly from Theorem 5.4.1 and Theorem 5.4.3 \square

Theorem 5.4.5. *Let $C = \{c_1, \dots, c_k\}$ be a partition. There exists a choice of weights $\lambda_i > 0$ with $\sum \lambda_i = 1$ such that Posterior Sampling with respect to the reweighted distribution*

$$p_\lambda(x) = \sum_i \lambda_i p(x \mid x \in c_i)$$

satisfies RDP with respect to C .

In the special case of 2 classes, the reweighting is very simple: $\lambda_1 = \lambda_2 = \frac{1}{2}$.

Proof. We will prove this theorem by contradiction. Before we start, observe that if we scale the mass of c_i by $\lambda_i \geq 0$, we have,

$$\begin{aligned} \alpha_i &:= \Pr(\hat{x} \in C_i \mid x^* \in C_i) \\ &= \mathbb{E}_{y|x^* \in C_i} \left[\frac{\lambda_i \Pr(x^* \in C_i \mid y)}{\sum_j \lambda_j \Pr(x^* \in C_j \mid y)} \right] \end{aligned}$$

WLOG, assume $\sum_i \lambda_i = 1$, this can be done by rescaling the λ_i 's by their sum. RDP is achieved if all the α_i are equal. Let the smallest α_i when all the λ_i 's are equal be ϵ . Consider the set $T := \{\vec{\lambda} \mid \sum_i \lambda_i = 1, \forall i \alpha_i \geq \epsilon\}$.

Towards a contradiction, suppose no assignment of $\vec{\lambda} \in T$ achieves $f(\vec{\lambda}) := \frac{\max_i \alpha_i}{\min_j \alpha_j} = 1$. Let $r := \min_{\lambda_1, \dots, \lambda_k} f(\vec{\lambda}) > 1$, and let $\vec{\lambda}^*$ be a point which achieves this. $\vec{\lambda}^*$ exists since $f(\vec{\lambda})$ is continuous over T , which is compact.

We will show that there exists $\vec{\lambda}' \in T$ such that $f(\lambda') < r$, which contradicts our hypothesis. Let $S := \{i \in [k] \mid \alpha_i \leq \sqrt{r} \min_i \alpha_i\}$ and,

$$\lambda'_i = \begin{cases} r^{1/4} \lambda_i^*, & \text{if } i \in S \\ \lambda_i^*, & \text{otherwise} \end{cases}$$

Let α'_i be $\Pr(\hat{x} \in C_i \mid x^* \in C_i)$ where the probability is with respect to the modified distribution. For $i \in S$,

$$\begin{aligned} \alpha'_i &= \mathbb{E}_{y|x^* \in C_i} \left[\frac{\lambda'_i \Pr(x^* \in C_i \mid y)}{\sum_j \lambda'_j \Pr(x^* \in C_j \mid y)} \right] \\ &= r^{1/4} \mathbb{E}_{y|x^* \in C_i} \left[\frac{\lambda_i \Pr(x^* \in C_i \mid y)}{\sum_j \lambda'_j \Pr(x^* \in C_j \mid y)} \right] \\ &\leq r^{1/4} \mathbb{E}_{y|x^* \in C_i} \left[\frac{\lambda_i \Pr(x^* \in C_i \mid y)}{\sum_j \lambda_j \Pr(x^* \in C_j \mid y)} \right] \end{aligned}$$

Where the last line follows from the fact that $\lambda'_i \geq \lambda_i$ for all i , since $r > 1$, which means the denominator only increases. A similar calculation shows that if $i \notin S$, then each α_i is multiplied by a factor between $r^{-1/4}$ and 1.

We notice that, by definition of S , all the j such that $\alpha_j = \min_i \alpha_i$ are in S , and all the k such that $\alpha_k = \max_i \alpha_i$ are not in S . This ensures that $\max_i \alpha'_i < \max_i \alpha_i$ and $\min_i \alpha'_i > \min_i \alpha_i \geq \epsilon$. Which, in turn, contradicts the hypothesis that r was the smallest achievable ratio with the original constraints, since we can always renormalize λ'_i without affecting the α_i .

$$\frac{\max_i \alpha'_i}{\min_i \alpha'_i} < \frac{\max_i \alpha_i}{\min\{r^{-1/4} \sqrt{r} \min_i \alpha_i, \min_i \alpha_i\}} \leq r.$$

□

Theorem 5.4.7. *Let $C = \{c_1, \dots, c_k\}$ form a disjoint partition of \mathbb{R}^n . An algorithm minimizes Representation Cross-Entropy on C iff the algorithm satisfies CPR on C .*

Proof. The proof follows from Lemma D.3.1. Note that $H(U|Y)$ is a function of $x^* \& y$ and hence has no dependence on the reconstruction algorithm. By the non-negativity of KL divergence, the representation cross-entropy is minimized when $Q(U_i|y) = P(U_i|y)$ for each $i \in [N]$, almost surely over y . □

Lemma D.3.1. *Let $U : \mathbb{R}^n \rightarrow \{c_1, c_2, \dots, c_k\}$ be a function that encodes which group contains an image, and assume that the groups $c_1, \dots, c_k \subset \mathbb{R}^n$ are disjoint and form a partition of \mathbb{R}^n .*

For a reconstruction algorithm \mathcal{A} , let $Q(c_i|Y)$ denote the probability that the reconstruction lies in the set c_i given measurements y . Let $P(c_i|y)$ denote the probability that x^ lies in U_i conditioned on y .*

Then we have

$$RCE(\mathcal{A}) = H_P(U|y) + \mathbb{E}_y [KL(P(U|y)\|Q(U|y))],$$

where

$$\begin{aligned} H_P(U|y) &:= -\mathbb{E}_y \left[\sum_{i \in [k]} P(c_i|y) \log P(c_i|y) \right], \\ KL(P(U|y)\|Q(U|y)) &:= \sum_{i \in [k]} P(c_i|y) \log \left(\frac{P(c_i|y)}{Q(c_i|y)} \right). \end{aligned}$$

Remark: There is a slight abuse of notation in the lemma. Since U is a function of x^* , when treating x^* as a random variable, we also treat U as a random variable.

Proof. By the definition of RCE and the tower property of expectations, we have

$$\begin{aligned} -RCE(\mathcal{A}) &= \mathbb{E}_{x^*, y} \log \Pr[\hat{x} \in U(x^*)|y] = \mathbb{E}_y \mathbb{E}_{x^*|y} [\log (\Pr[\hat{x} \in U(x^*)|y])], \\ &= \mathbb{E}_y \mathbb{E}_{x^*|y} \left[\sum_{i \in [N]} \mathbf{1}\{x^* \in U_i\} \log (\Pr[\hat{x} \in U(x^*)|y]) \right], \\ &= \mathbb{E}_y \mathbb{E}_{x^*|y} \left[\sum_{i \in [N]} \mathbf{1}\{x^* \in U_i\} \log (\Pr[\hat{x} \in U_i|y]) \right], \\ &= \mathbb{E}_y \mathbb{E}_{x^*|y} \left[\sum_{i \in [N]} \mathbf{1}\{x^* \in U_i\} \log (Q(U_i|y)) \right], \\ &= \mathbb{E}_y \left[\sum_{i \in [N]} P(U_i|y) \log (Q(U_i|y)) \right]. \quad .(*) \end{aligned}$$

where the second line follows because the U_i s form a partition, the third line follows since $\hat{x} \in U(x^*)$ is equivalent to $\hat{x} \in U_i$ if we know that $x^* \in U_i$, the fourth line follows from the definition of $Q(U_i|y)$ and the last line follows from linearity of expectation.

Now we can multiply and divide $P(U_i|y)$ within the log term above.

This gives

$$\begin{aligned}
(*) &= \mathbb{E}_y \left[\sum_{i \in [N]} P(U_i|y) \log \left(\frac{Q(U_i|y)P(U_i|y)}{P(U_i|y)} \right) \right], \\
&= \mathbb{E}_y \left[\sum_{i \in [N]} P(U_i|y) \log (P(U_i|y)) \right] \\
&\quad + \mathbb{E}_y \left[\sum_{i \in [N]} P(U_i|y) \log \left(\frac{Q(U_i|y)}{P(U_i|y)} \right) \right], \\
&= -H(U|y) - \mathbb{E}_y [KL(P(U|y)\|Q(U|y))].
\end{aligned}$$

This concludes the proof. □

D.4 Langevin Dynamics

D.4.1 StyleGAN2

We want to sample from the distribution $p(x|y)$ induced by a StyleGAN2. Note that sampling from the marginal distribution $p(x)$ of a StyleGAN2 is achieved by sampling a latent variable $z \in \mathbb{R}^{512}$, and 18 noise variables $n_i \in \mathbb{R}^{d_i}$ of varying sizes, and setting $x = G(z, n_1, \dots, n_{18})$. Hence, we

can sample from $p(x|y)$ by sampling $\hat{z}, \hat{n}_1, \dots, \hat{n}_{18}$, from $p(z, n_1, \dots, n_{18}|y)$, and setting $\hat{x} = G(\hat{z}, \hat{n}_1, \dots, \hat{n}_{18})$.

The prior of the latent and noise variables is a standard Gaussian distribution. Since we know the prior distribution of these variables, if we know the distribution of the measurement process, we can write out the posterior distribution.

For the measurement process we consider, we have $y = Ax^*$, where A is a blurring matrix of appropriate dimension. Note that in the absence of noise, posterior sampling must sample solutions that exactly satisfy the measurements. However, this is difficult to enforce in practice, and hence we assume that there is some small amount of Gaussian noise in the measurements. In this case, the posterior distribution becomes:

$$p(z, n_1, \dots, n_{18}|y) \propto p(y|z, n_1, \dots, n_{18})p(z, n_1, \dots, n_{18}), \quad (\text{D.1})$$

$$\Leftrightarrow \log p(z, n_1, \dots, n_{18}|y) = -\frac{\|y - AG(z, n_1, \dots, n_{18})\|^2}{2\sigma^2/m} - \|z\|^2/2 - \sum_{i=1}^1 8\|n_i\|^2/2 + c(y), \quad (\text{D.2})$$

where $c(y)$ is an additive constant which depends only on y .

Now, Langevin dynamics tells us that if we run gradient ascent on the above log-likelihood, and add noise at each step, then we will sample from the conditional distribution asymptotically. Please note that we sample z and *all noise variables* n_1, \dots, n_{18} .

In our experiments, we do 1500 gradient steps. In practice, we replace the σ in the equation above with σ_t , where t is the iteration number. When the

measurements have resolution 8×8 or 4×4 , we find that $\sigma_1 = 1.0, \sigma_{1500} = 0.1$ works best. When the resolution of the measurements is 2×2 or 1×1 , we find that $\sigma_1 = 1.0, \sigma_{1500} = 0.01$ works best. We change the value of σ_t after every 3 gradient steps, such that $\sigma_1, \sigma_4, \sigma_7, \dots, \sigma_{1497}$ form a geometrically decreasing sequence. The learning rate γ_{1500} is also tuned to be a decreasing geometric sequence, such that $\gamma_t = 5 \cdot 10^{-6}$. Please see [253] for the equations specifying the learning rate tuning, and the logic behind it.

We also find that adding a small amount of noise corresponding to σ_{1500} in the measurements helps Langevin mix better.

We note that our approach is different from prior work [147, 199], which optimizes a function of our variable z , and a subset of the noise variables.

NCSNv2 The NCSNv2 model [254] has been designed such that sampling from the marginal distribution requires Langevin dynamics. This model is given by a function $s(x; \sigma)$, which outputs $\nabla \log p_\sigma(x)$, where $p_\sigma(x)$ is the distribution obtained by convolving the distribution $p(x)$ with Gaussian noise of variance σ^2 . That is, $p_\sigma(x) = (p * \mathcal{N}(0, \sigma^2))(x)$.

It is easy to adapt the NCSNv2 model to sample from posterior distributions, see [254] for inpainting examples. In our super-resolution experiments, one can compute the gradient of $p(y|x)$, to get the following update rule for Langevin dynamics:

$$x_{t+1} \leftarrow x_t + \gamma_t(s(x_t; \sigma_t) - A^T(Ax_t - y)/\sigma_t^2) + \sqrt{2\gamma_t}\xi_t, \quad (\text{D.3})$$

where A is the blurring matrix, and $\xi_t \sim \mathcal{N}(0, I_n)$ is i.i.d. Gaussian noise sampled at each step. We use the default values of noise and learning rate specified in <https://github.com/ermongroup/ncsnv2/blob/master/configs/ffhq.yml>. That is, $\sigma_1 = 348$, $\sigma_{6933} = 0.01$, and $\gamma_{6933} = 9 \cdot 10^{-7}$. Note that the value of σ, γ changes every 3 iterations, and both γ_t and σ_t decay geometrically. See [254] for specific details on how these are tuned.

D.5 Code

All code and generative models, along with hyperparameters and README are available at <https://github.com/ajiljalal/code-cs-fairness>.

Appendix E

Appendix for Chapter 6

E.1 Appendix: Additional Metrics

Figure E.1 shows the test SSIM evaluated in the same conditions as Figure 6.2 in the main text. This highlights that our model is also robust in this metric.

We observe that our method has significant noise in the background. Hence, we also report the masked SSIM and PSNR values in Figures E.2 and E.3. The mask zeros out all coordinates whose absolute value is smaller than 0.05 times the maximum absolute value in the fully-sampled MVUE.

E.1.1 MVUE vs. RSS

The difference in numerical values between our results and the publicly available fastMRI leaderboard, as well as original results in the published baseline papers baselines comes from training and evaluating all methods on MVUE instead of RSS images. This is a design choice that we have made for all baselines, since our goal is to compare with a wide range of previous methods in a fair way.

Algorithms that output a complex-valued image (such as ours and L1-Wavelet) as a solution to the optimization in Eqn (6.2) will artificially perform worse (w.r.t. E2E methods) when compared to the RSS ground truth, even when the output is of similar or higher quality, due to the bias in the RSS. Since there is no way to obtain a good RSS score with these algorithms, this justifies our choice to train and evaluate all methods on MVUE.

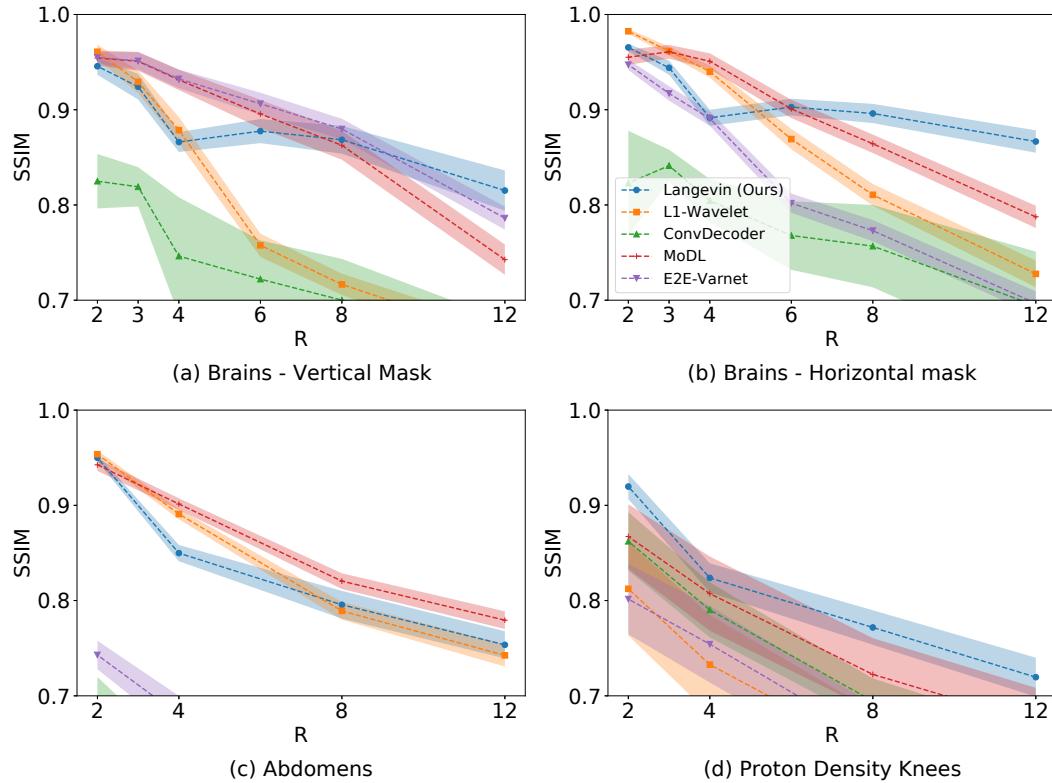


Figure E.1: Average test SSIM in various scenarios, across a range of acceleration factors R . Higher R indicates a smaller number of acquired measurements. Our approach mostly shows the best performance and lowest reconstruction variance both in- and out-of-distribution at test-time. Shaded regions indicate 95% confidence intervals. Note that we trained baselines on MVUE images and hence these numerical values should not be compared with those in literature trained on RSS images (see Appendix E.1.1 for a more detailed discussion).

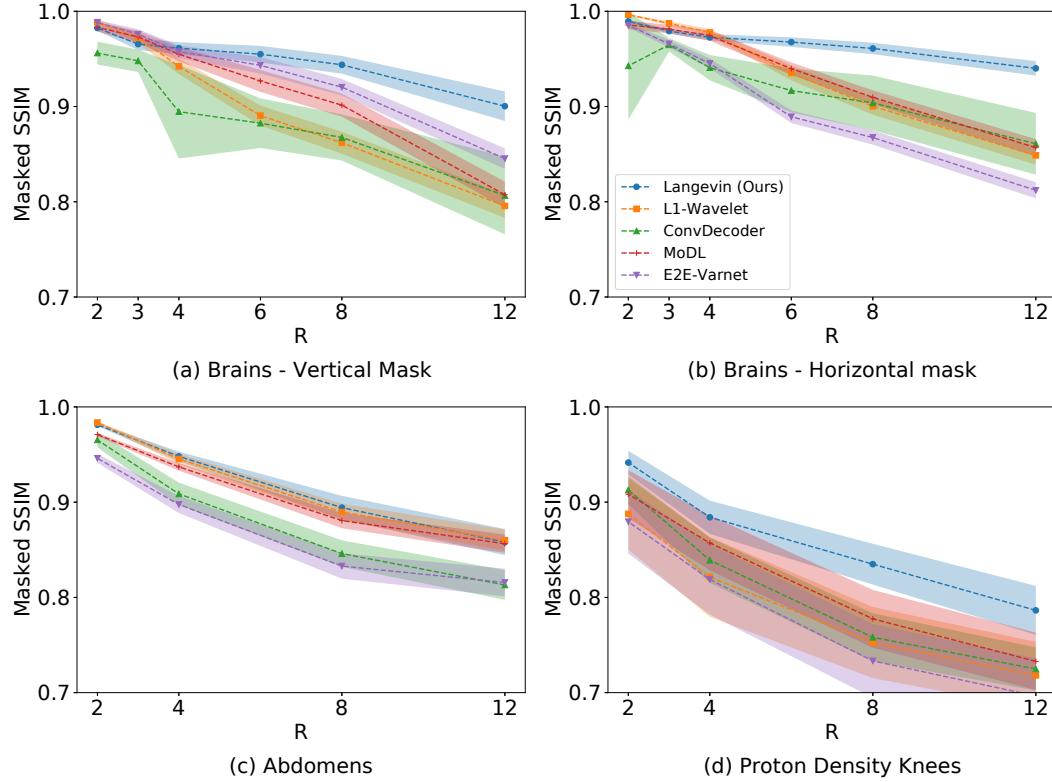


Figure E.2: Average test SSIM, with masking, in various scenarios across a range of acceleration factors R . The mask zeros out all coordinates whose absolute value is smaller than 0.05 times the maximum absolute value in the fully-sampled MVUE, and this reduces the effect of noise in the background. Higher R indicates a smaller number of acquired measurements. Our approach mostly shows the best performance and lowest reconstruction variance both in- and out-of-distribution at test-time. Shaded regions indicate 95% confidence intervals. Note that we trained baselines on MVUE images and hence these numerical values should not be compared with those in literature trained on RSS images (see Appendix E.1.1 for a more detailed discussion).

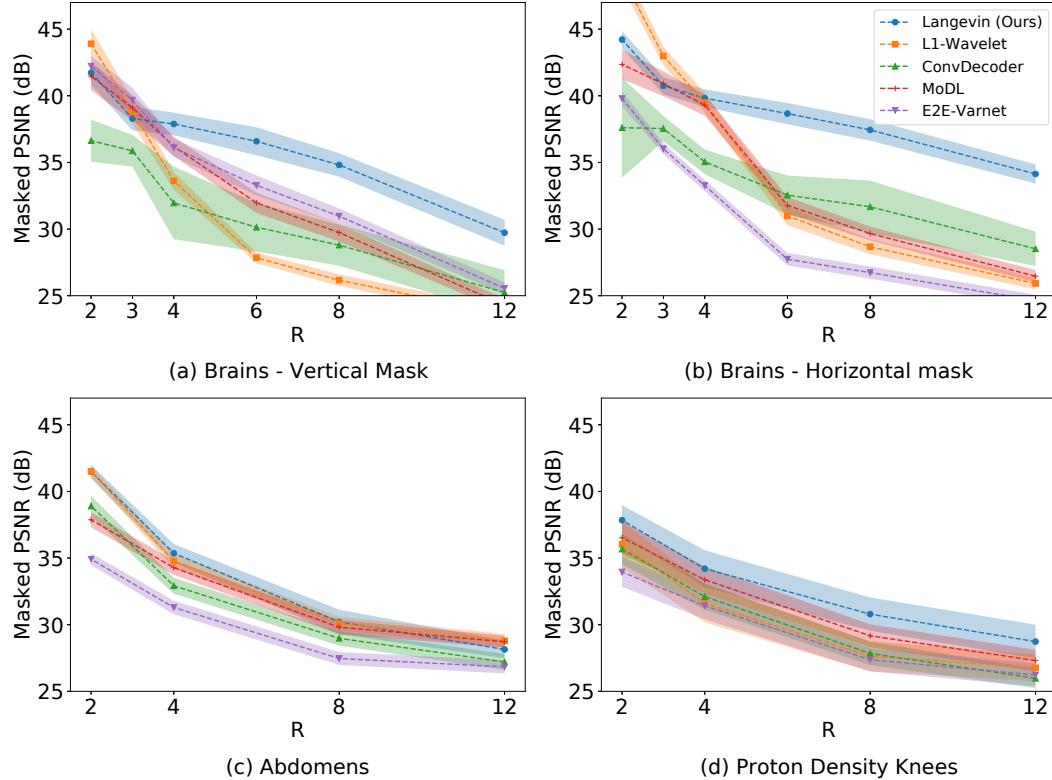


Figure E.3: Average test PSNR, with masking, in various scenarios across a range of acceleration factors R . The mask zeros out all coordinates whose absolute value is smaller than 0.05 times the maximum absolute value in the fully-sampled MVUE, and this reduces the effect of noise in the background. Higher R indicates a smaller number of acquired measurements. Our approach mostly shows the best performance and lowest reconstruction variance both in- and out-of-distribution at test-time. Shaded regions indicate 95% confidence intervals. Note that we trained baselines on MVUE images and hence these numerical values should not be compared with those in literature trained on RSS images (see Appendix E.1.1 for a more detailed discussion).

To the best of our knowledge, a rigorous, reproducible comparison between end-to-end models trained on RSS or MVUE images has not been made in prior work. The recent work of [103] has also discussed this point. To illustrate our claim of incompatibility between the two estimates, as well as the importance of qualitative inspection, we provide two simple, easy-to-verify examples.

1. We compare the fully sampled MVUE reconstruction (with ESPiRIT estimated maps) with the fully sampled RSS reconstruction, on T2 brain scans: we find that the SSIM is slightly larger than 0.8. This is a large penalty (as per Fig. 6.1), even though the two images are virtually indistinguishable and known to be clinically equivalent (see discussions of SENSE vs. GRAPPA in [103]). This would unfairly penalize the family of methods that explicitly solve the inverse problem. Since E2E methods can be trained to target the MVUE directly, this justifies our choice for using the MVUE as the reference image.
2. We point to the public knee fastMRI leaderboard at <https://fastmri.org/leaderboards>. Selecting "Multi-coil Knee" and "4x" acceleration, we inspect the two following submissions:
 - "zero-filling", which does zero-filling RSS reconstruction, has an SSIM of 0.804 and considerable artifacts.
 - "Baseline Classical Reconstruction Model", which applies compressed sensing with the ESPiRIT algorithm, has a much poorer SSIM score

of 0.6275, but produces qualitatively superior reconstructions.

E.2 Appendix: Theory

Lemma E.2.1 (\mathcal{W}_q implies (δ, α) - \mathcal{W}_∞). *If two distributions μ and ν satisfy $\mathcal{W}_q(\mu, \nu) \leq \varepsilon$ for some $q \geq 1$, then they satisfy (δ, δ) - $\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon/\delta^{1/q}$. Furthermore, there exist distributions that satisfy (δ, δ) - $\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon$, but $\mathcal{W}_q(\mu, \nu) = \infty$ for all $q \geq 1$.*

Proof. Let Γ be a coupling between μ, ν such that $\mathbb{E}_{(u,v) \sim \Gamma} [\|u - v\|^q] \leq \varepsilon^q$.

Then an application of Markov's inequality gives

$$\Pr[\|u - v\| \geq \varepsilon/\delta^{1/q}] \leq \delta. \quad (\text{E.1})$$

Now, we can split the distribution Γ into two unnormalized components Γ', Γ'' defined as

$$\begin{aligned}\Gamma'(u, v) &= \Gamma(u, v)\mathbf{1}\{\|u - v\| < \varepsilon/\delta^{1/q}\}, \\ \Gamma''(u, v) &= \Gamma(u, v)\mathbf{1}\{\|u - v\| \geq \varepsilon/\delta^{1/q}\}.\end{aligned}$$

Using Γ', Γ'' , we can define measures μ', μ'', ν', ν'' , via

$$\begin{aligned}\mu'(B) &:= \Gamma'(B, \Omega), \\ \mu''(B) &:= \Gamma''(B, \Omega), \\ \nu'(B) &:= \Gamma'(\Omega, B), \\ \nu''(B) &:= \Gamma''(\Omega, B),\end{aligned}$$

where B is any measurable set and Ω is the state-space.

Since Γ is a valid coupling between μ, ν , and Γ', Γ'' are disjoint distributions, for any measurable $B \subseteq \Omega$, we have:

$$\begin{aligned}\mu(B) &= \Gamma(B, \Omega), \\ &= \Gamma'(B, \Omega) + \Gamma''(B, \Omega), \\ &= \mu'(B) + \mu''(B), \\ &= \mu'(\Omega) \frac{\mu'(B)}{\mu'(\Omega)} + \mu''(\Omega) \frac{\mu''(B)}{\mu''(\Omega)}.\end{aligned}$$

Using Eqn (E.1), we can conclude that $\mu'(\Omega) \geq 1 - \delta, \mu''(\Omega) \leq \delta$. Setting $\mu' \leftarrow \mu'/\mu'(\Omega)$ and $\mu'' \leftarrow \mu''/\mu''(\Omega)$, we can now rewrite μ as $\mu = (1 - \delta)\mu' + \delta\mu''$. A similar argument for ν gives $\nu = (1 - \delta)\nu' + \delta\nu''$.

By construction, μ', ν' can be \mathcal{W}_∞ coupled via Γ' to within a distance of $\varepsilon/\delta^{1/q}$. This shows that $(\delta, \delta)\text{-}\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon/\delta^{1/q}$.

Now we need to show that two distributions can be close in $(\delta, \delta)\text{-}\mathcal{W}_\infty$, but $\mathcal{W}_q = \infty$ for all q . Consider two scalar distributions μ, ν defined as

$$\begin{aligned}\mu &= \begin{cases} 0 & \text{with probability } 1 - \delta, \\ r & \text{with probability } \delta, \end{cases}, \\ \nu &= \begin{cases} \varepsilon & \text{with probability } 1 - \delta, \\ -r & \text{with probability } \delta. \end{cases}\end{aligned}$$

Clearly, these distributions satisfy $(\delta, \delta)\text{-}\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon$, but $\mathcal{W}_q(\mu, \nu) \approx r$ for all q . As $r \rightarrow \infty$, we get $\mathcal{W}_q(\mu, \nu) \rightarrow \infty$ for all $q \geq 1$.

□

E.2.1 Proof of Theorem 6.4.3

In order to prove the Theorem, we make use of the following three Lemmas from [134].

Lemma E.2.2. [134] 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 $z^* \sim \text{Bernoulli}(c)$.

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

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

Lemma E.2.3. [134] 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 η -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:

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

Lemma E.2.4. [134] 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 $\hat{x} \sim P(\cdot|y, A)$ and $\hat{z} \sim P(\cdot|u, A)$. For any $d > 0$, we have

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

Theorem 6.4.3. Let $\delta, \alpha \in [0, 1]$, and $\varepsilon > 0$ be parameters. Let μ, ν be arbitrary distributions over \mathbb{R}^N satisfying $(\delta, \alpha)\text{-}\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon$. Let $x^* \sim \mu$ and suppose $y = Ax^* + w$, where $A \in \mathbb{R}^{M \times N}$ and $w \in \mathbb{R}^M$ are i.i.d. Gaussian normalized such that $A_{ij} \sim \mathcal{N}(0, 1/M)$ and $w_i \sim \mathcal{N}(0, \sigma^2/M)$, with $\sigma \gtrsim \varepsilon$. Given y and the fixed matrix A , let \hat{x} be the output of posterior sampling with respect to ν .

Then for $M \geq O\left(\log\left(\frac{1}{1-\alpha}\right) + \min(\log \text{Cov}_{\sigma, \delta}(\mu), \log \text{Cov}_{\sigma, \delta}(\nu))\right)$, there exists a universal constant $c > 0$ such that with probability at least $1 - e^{-\Omega(M)}$ over A, w ,

$$\Pr_{x^* \sim \mu, \hat{x} \sim \nu(\cdot|y)} [\|x^* - \hat{x}\| \geq c(\varepsilon + \sigma)] \leq \delta + e^{-\Omega(M)}.$$

Proof. We know from $(\delta, \alpha)\text{-}\mathcal{W}_\infty(\mu, \nu) \leq \varepsilon$ that there exist μ', ν', μ'', ν'' and a finite distribution Q supported on a set S such that

1. $\mathcal{W}_\infty(\mu', \nu') \leq \varepsilon$,
2. $\min\{\mathcal{W}_\infty(\nu', Q), \mathcal{W}_\infty(\mu', Q)\} \leq \sigma$,

3. $\mu = (1 - \delta)\mu' + \delta\mu''$ and $\nu = (1 - \alpha)\nu' + \alpha\nu''$.

Suppose $\mathcal{W}_\infty(\nu', Q) \leq \sigma$. If not, then $\mathcal{W}_\infty(\mu', Q) \leq \sigma$, and by (1), we see that $\mathcal{W}_\infty(\nu', Q) \leq \sigma + \varepsilon$, and we will use this in the proof instead. By decomposing $\mu = (1 - \delta)\mu' + \delta\mu''$, we have

$$\Pr_{x^* \sim \mu, \hat{x} \sim \nu(\cdot|y)} [\|x^* - \hat{x}\| \geq (2c + 1)\sigma + \varepsilon] \leq \delta + (1 - \delta) \Pr_{x^* \sim \mu', \hat{x} \sim \nu(\cdot|y)} [\|x^* - \hat{x}\| \geq (2c + 1)\sigma + \varepsilon]. \quad (\text{E.2})$$

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, w, \hat{x} . By Lemma E.2.4, we can replace $x^* \sim \mu'$ with $z^* \sim \nu'$, replace $y = Ax^* + w$ with $u = Az^* + w$, and replace $\hat{x} \sim \nu(\cdot|A, y)$ with $\hat{z} \sim \nu(\cdot|A, u)$ to get the following bound

$$\begin{aligned} & \Pr_{x^* \sim \mu', A, w, \hat{x} \sim \nu(\cdot|A, y)} [\|x^* - \hat{x}\| \geq (2c + 1)\sigma + \varepsilon] \\ & \leq e^{-\Omega(m)} + e^{\left(\frac{2\varepsilon(\varepsilon+2\sigma)m}{\sigma^2}\right)} \Pr_{z^* \sim \nu', A, w, \hat{z} \sim \nu(\cdot|u, A)} [\|z^* - \hat{z}\| \geq (2c + 1)\sigma]. \end{aligned} \quad (\text{E.3})$$

We now bound the second term in the right hand side of the above inequality. Let Γ denote an optimal \mathcal{W}_∞ -coupling between ν' 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}_∞ condition, each $\Gamma(\cdot|\tilde{z})$ is supported on a ball of radius σ around \tilde{z} .

Let $E = \{z^*, \tilde{z} \in \mathbb{R}^n : \|z^* - \tilde{z}\| \geq (2c + 1)\sigma\}$ denote the event that z^*, \tilde{z} are far apart. By the coupling, we can express ν' as

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

This gives

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

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

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

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

as

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

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

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

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

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

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

$$\begin{aligned} Q(\tilde{z}^*) \mathbb{E}_{z^* \sim \Gamma(\cdot | \tilde{z}^*), A, w, \hat{z} \sim \nu(\cdot | A, u)} [1_E] &\leq \mathbb{E}_{A, w, z^*} \Pr[z^* \sim \nu_{\tilde{z}^*,0}, \hat{z} \sim \nu_{\tilde{z}^*,2}(\cdot | u)] \\ &\leq \frac{1}{1-\alpha} \mathbb{E}_A [1 - TV(H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2})], \end{aligned}$$

where $H_{\tilde{z}^*,0}, H_{\tilde{z}^*,2}$ are the push-forwards of $\nu_{\tilde{z}^*,0}, \nu_{\tilde{z}^*,2}$ for A fixed and the last inequality follows from Lemma E.2.2.

Notice that if we sum over all $\tilde{z}^* \in S$, then the LHS of the above

inequality is an expectation over $z^* \sim \nu'$. This gives:

$$\Pr_{z^* \sim \nu', A, w, \tilde{z} \sim \nu(\cdot | u, A)} [E] \leq \frac{1}{1 - \alpha} \sum_{\tilde{z}^* \in S} \mathbb{E}_A [1 - TV(H_{\tilde{z}^*, 0}, H_{\tilde{z}^*, 2})].$$

Notice that $\nu_{\tilde{z}^*, 0}$ is supported within an σ -ball around \tilde{z}^* , and $\nu_{\tilde{z}^*, 2}$ is supported outside a $2c\sigma$ -ball of \tilde{z}^* . By Lemma E.2.3 we have

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

This implies

$$\begin{aligned} \Pr_{z^* \sim \nu', A, w, \tilde{z} \sim \nu(\cdot | u, A)} [\|z^* - \tilde{z}\| \geq (2c + 1)\sigma] &\leq \frac{1}{1 - \alpha} \sum_{\tilde{z}^* \in S} \mathbb{E}_A [(1 - TV(H_{\tilde{z}^*, 0}, H_{\tilde{z}^*, 2}))], \\ &\leq \frac{1}{1 - \alpha} 4|S| e^{-\frac{m}{2} \log(\frac{c}{4e^2})}, \\ &\leq 4e^{-\frac{m}{4} \log(\frac{c}{4e^2})}, \end{aligned}$$

where the last inequality is satisfied if $m \geq 4 \log(\frac{1}{1 - \alpha}) + 4 \log(|S|)$.

Substituting in Eqn (E.3), if $c > 4 \exp\left(2 + \frac{8\varepsilon(\varepsilon + 2\sigma)}{\sigma^2}\right)$, we have

$$\Pr_{x^* \sim \mu', A, w, \hat{x} \sim \nu(\cdot | A, y)} [\|x^* - \hat{x}\| \geq (2c + 1)\sigma + \varepsilon] \leq e^{-\Omega(m)}.$$

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

$$\Pr_{x^* \sim \mu', \hat{x} \sim \nu(\cdot | y)} [\|x^* - \hat{x}\| \geq (2c + 1)\sigma + \varepsilon] \leq e^{-\Omega(m)}.$$

Substituting in Eqn (E.2), we have

$$\Pr_{x^* \sim \mu, \hat{x} \sim \nu(\cdot | y)} [\|x^* - \hat{x}\| \geq (2c + 1)\sigma + \varepsilon] \leq \delta + e^{-\Omega(m)}.$$

Rescaling c gives us our result.

At the beginning of the proof, we had assumed that $\mathcal{W}_\infty(\nu', Q) \leq \sigma$. If instead $\mathcal{W}_\infty(\mu', Q) \leq \sigma$, then we need to replace σ in the above bound by $\sigma + \varepsilon$. Rescaling c in the above bound gives us the Theorem statement.

□

E.2.2 Proof of Theorem 6.4.4

Theorem 6.4.4. *Let $d(\cdot, \cdot)$ be an arbitrary metric over $\mathbb{R}^N \times \mathbb{R}^N$. Let $x^* \sim \mu$ and let $y = \mathcal{A}(x^*)$ be measurements generated from x^* for some arbitrary forward operator $\mathcal{A} : \mathbb{R}^N \rightarrow \mathbb{R}^M$. Then if there exists an algorithm that uses y as inputs and outputs x' such that*

$$d(x^*, x') \leq \varepsilon \text{ with probability } 1 - \delta,$$

then posterior sampling $\hat{x} \sim \mu(\cdot|y)$ will satisfy

$$d(x^*, \hat{x}) \leq 2\varepsilon \text{ with probability } \geq 1 - 2\delta.$$

Proof. By the statement of the Lemma, and conditioning on the measurements y , we have

$$1 - \delta = \Pr[d(x^*, x') \leq \varepsilon] = \mathbb{E}_y (\Pr[d(x^*, x') \leq \varepsilon | y]).$$

Using a similar conditioning for the event $d(x^*, \hat{x}) \leq 2\varepsilon$, we get

$$\begin{aligned}
\Pr[d(x^*, \hat{x}) \leq 2\varepsilon] &= \mathbb{E}_y (\Pr[d(x^*, \hat{x}) \leq 2\varepsilon | y]) , \\
&\geq \mathbb{E}_y (\Pr[d(x^*, x') \leq \varepsilon \wedge d(x', \hat{x}) \leq \varepsilon | y]) , \\
&= \mathbb{E}_y (\Pr[d(x^*, x') \leq \varepsilon | y] \cdot \Pr[d(x', \hat{x}) \leq \varepsilon | y]) , \\
&= \mathbb{E}_y (\Pr[d(x^*, x') \leq \varepsilon | y]^2) , \\
&\geq \left(\mathbb{E}_y (\Pr[d(x^*, x') \leq \varepsilon | y]) \right)^2 , \\
&= (1 - \delta)^2 \geq 1 - 2\delta ,
\end{aligned}$$

where the second line follows from a triangle inequality, the third line follows since x^*, \hat{x} are independent conditioned on y , the fourth line follows since $\hat{x}|y$ is distributed according to $x^*|y$, and the fifth line follows from Jensen's inequality.

□

E.3 Appendix: fastMRI Brain

E.3.1 Examples of Sampling Masks

Figure E.4 shows example of some of the masks used throughout the experiments in the paper and their corresponding reconstructions. Note that the type of mask used is coupled with the scan parameters (e.g., two-dimensional slices from a three-dimensional scan will use a 2D grid of points).

We also highlight that, in all cases, a central region of the k-space is kept fully sampled and is used to estimate the coil sensitivity maps for all methods.

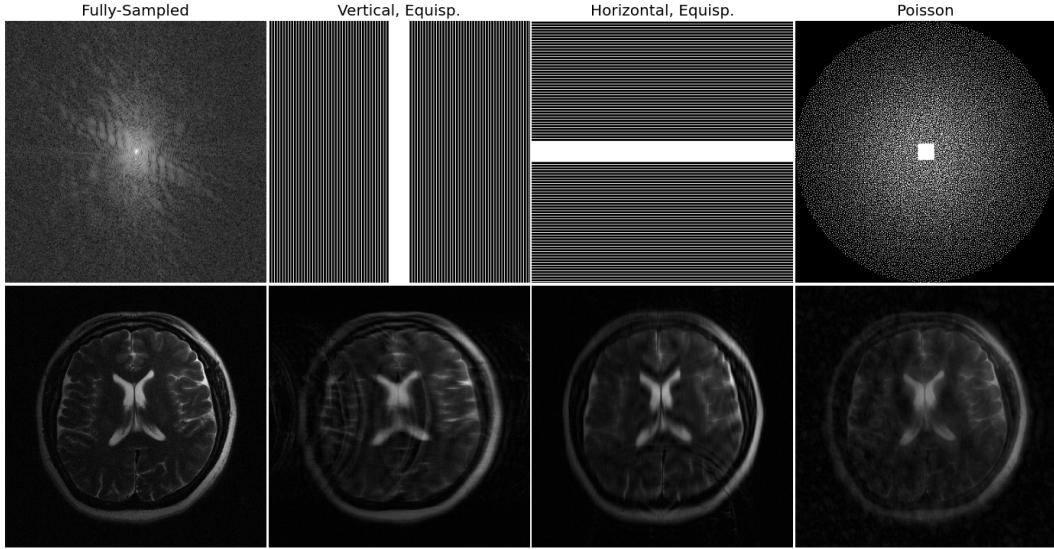


Figure E.4: Examples of sampling patterns used throughout the experiments (top) and naive reconstructions (bottom). Top: The leftmost image shows the log-magnitude of the fully sampled k-space measurements corresponding to a single coil. The remaining images show three possible sampling masks, all with acceleration factor $R = 4$ but drastically different patterns. Bottom: Each image shows the magnitude of the reconstruction obtained by a two-dimensional IFFT applied to the sampled k-space.

The bottom row of Figure E.4 shows naive reconstructions of a single coil image using the zero-filled k-space. This shows that different types of masks lead to different types of aliasing patterns in the image domain, motivating the need for robust image reconstruction algorithms.

E.3.2 More Exemplar Reconstructions

Figures E.5 throughout E.10 show detailed qualitative reconstructions on different brain scans from the fastMRI dataset. We highlight Figures E.9 and E.10, which represent a contrast shift from the in-distribution data (T1

and FLAIR vs. T2, respectively). Our method still produces excellent qualitative reconstructions.

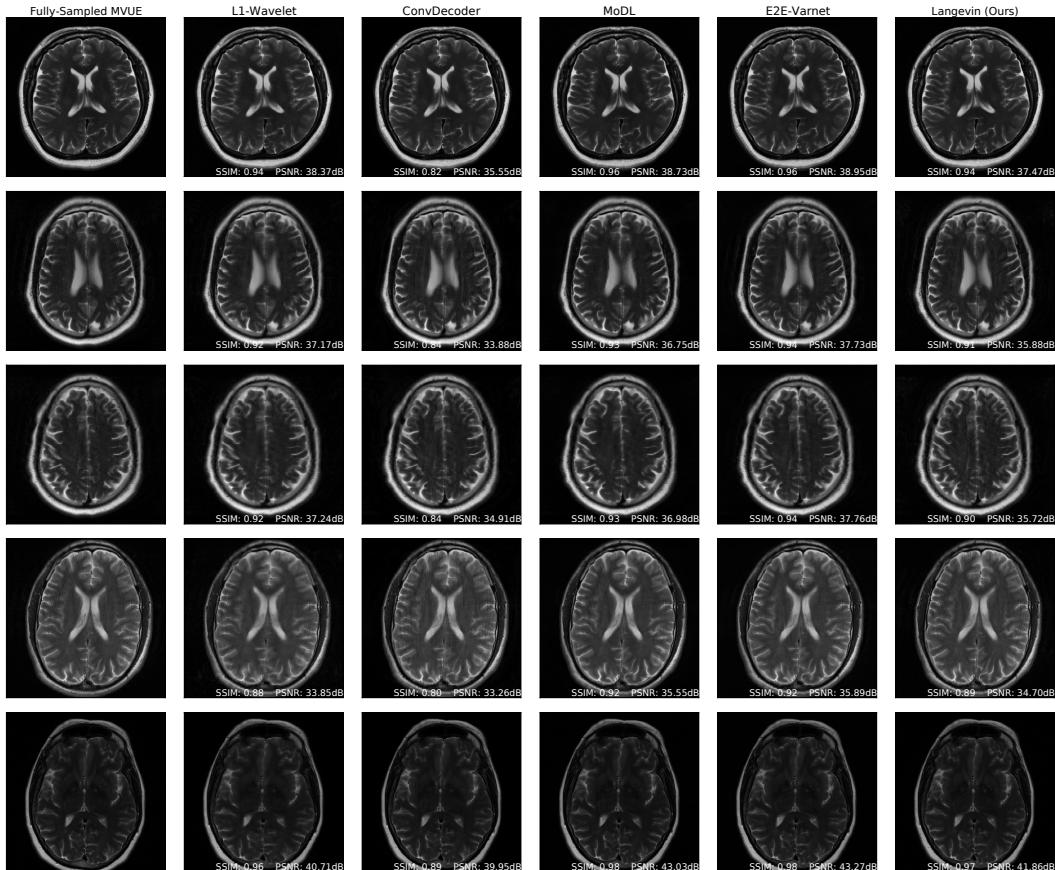


Figure E.5: In-distribution brain reconstructions, at an acceleration factor of $R = 3$ and an equispaced vertical mask in k-space. Our model was trained on T2-weighted brain images from the fastMRI dataset. These results show that our method is competitive with state-of-the-art methods such as E2E-VarNet.

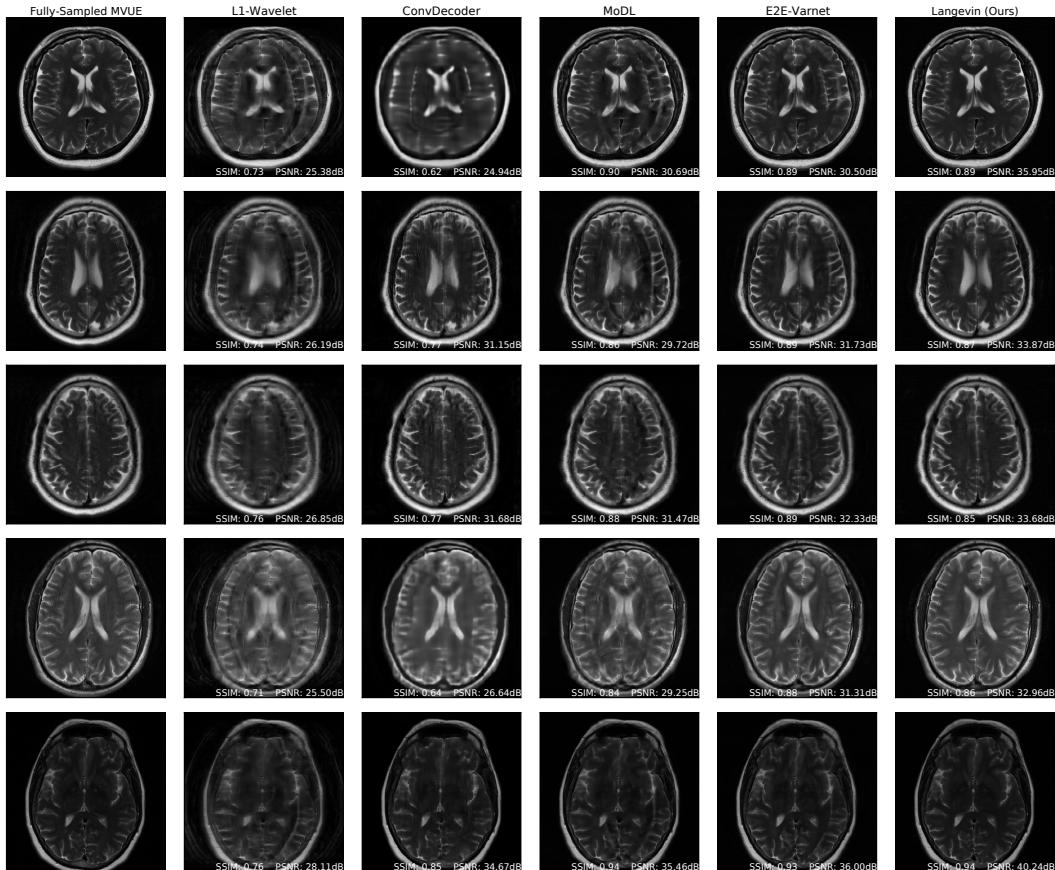


Figure E.6: In-distribution brain reconstructions, at an acceleration factor of $R = 6$ and an equispaced vertical mask in k-space. Our model was trained on T2-weighted brain images from the fastMRI dataset. These results show that our method retains its performance at higher acceleration factors.

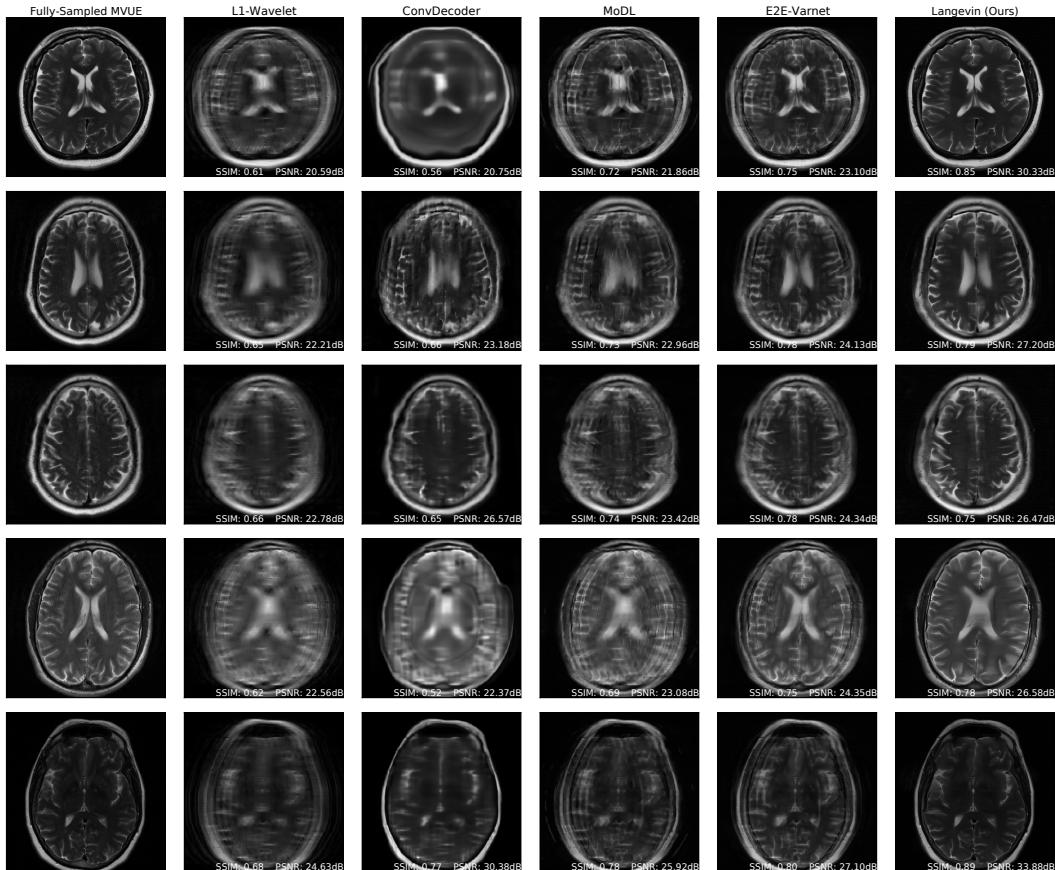


Figure E.7: Brain reconstructions, at an acceleration factor of $R = 12$ and an equispaced vertical mask in k-space. Our model was trained on T2-weighted brain images from the fastMRI dataset. These results show that our method has significantly fewer artifacts than baselines.

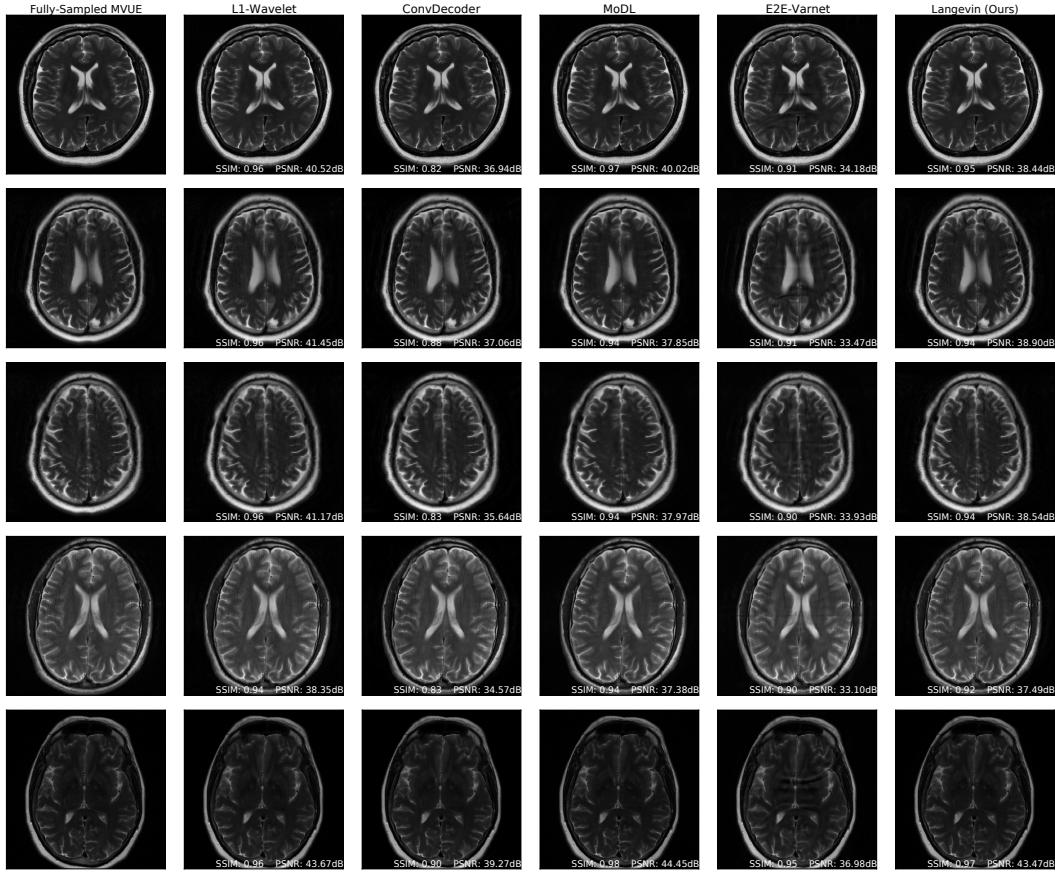


Figure E.8: Brain reconstructions under a mask shift, at an acceleration of $R = 3$. MoDL and E2E-VarNet were trained using an equispaced vertical mask, while these experiments were run using an equispaced *horizontal* mask. Our method is robust to the mask shift, as our generative prior was trained without any knowledge of the measurement process. ConvDecoder and L1-Wavelets are untrained methods, and hence are robust to the mask shift.

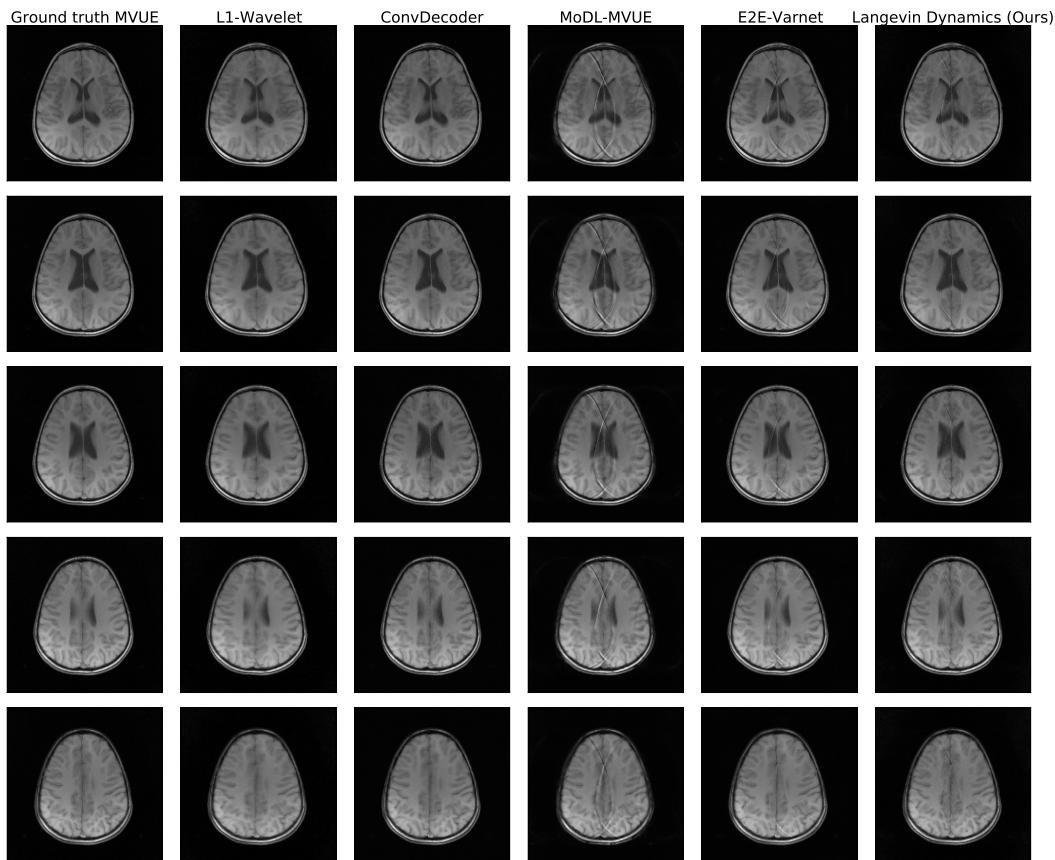


Figure E.9: Brain reconstructions under a contrast shift, at an acceleration of $R = 4$. Our method was trained on T2-weighted brains, while these are T1-weighted brains, and our method is clearly robust to this contrast shift.

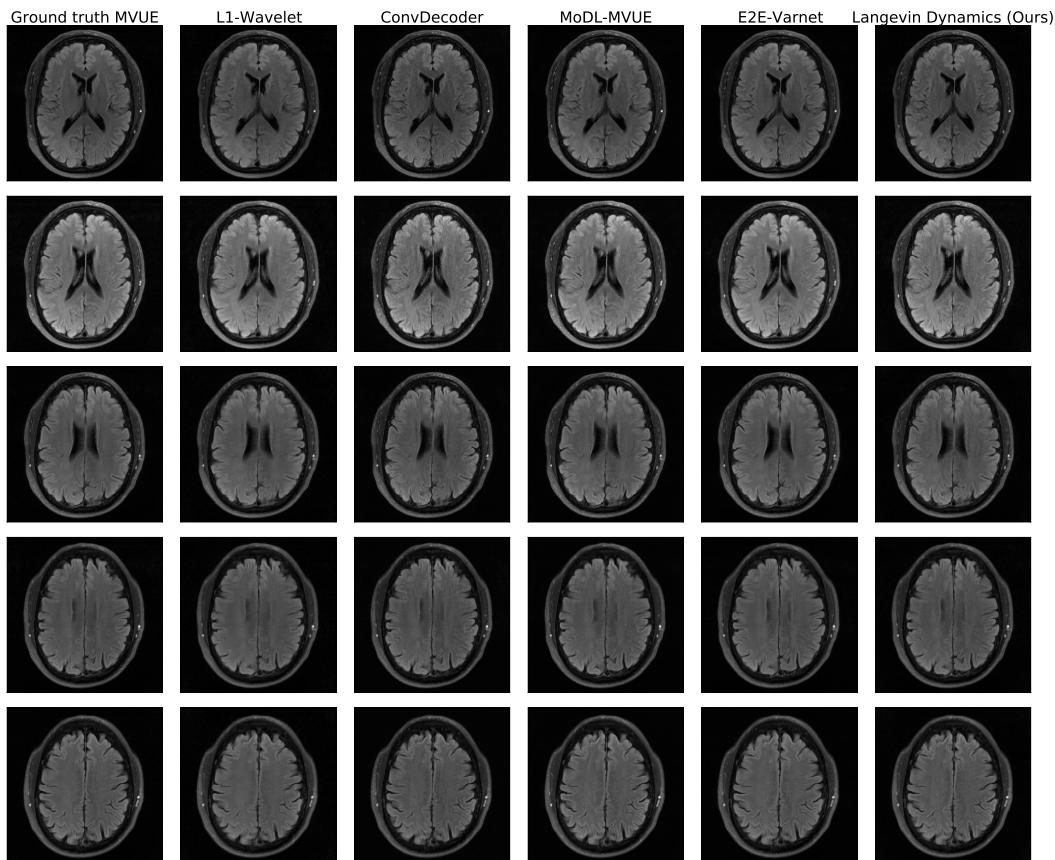


Figure E.10: Brain reconstructions under a contrast shift, at an acceleration of $R = 4$. Our method was trained on T2-weighted brains, while these are FLAIR brains, and our method is clearly robust to this contrast shift.

E.4 Appendix: fastMRI Knee

Figure E.11 and Figure E.12 show further examples of proton density knee reconstructions.

Figure E.14 and Figure E.15 show comparisons of our method and baselines on knees with meniscus tears. Figure E.13 shows uncertainty estimates from our algorithm on a knee with a meniscus tear.

Figure E.16 shows PSNR and SSIM on fat-suppressed(FS) knees. Our approach is not optimal numerically, likely due to a much lower signal-to-noise ratio in FS knees than the brain training data. However, Figures E.14, E.15, E.17, E.18 show that our qualitative reconstructions are competitive, and recovers fine details (like meniscus tears) better than the deep learning baselines.

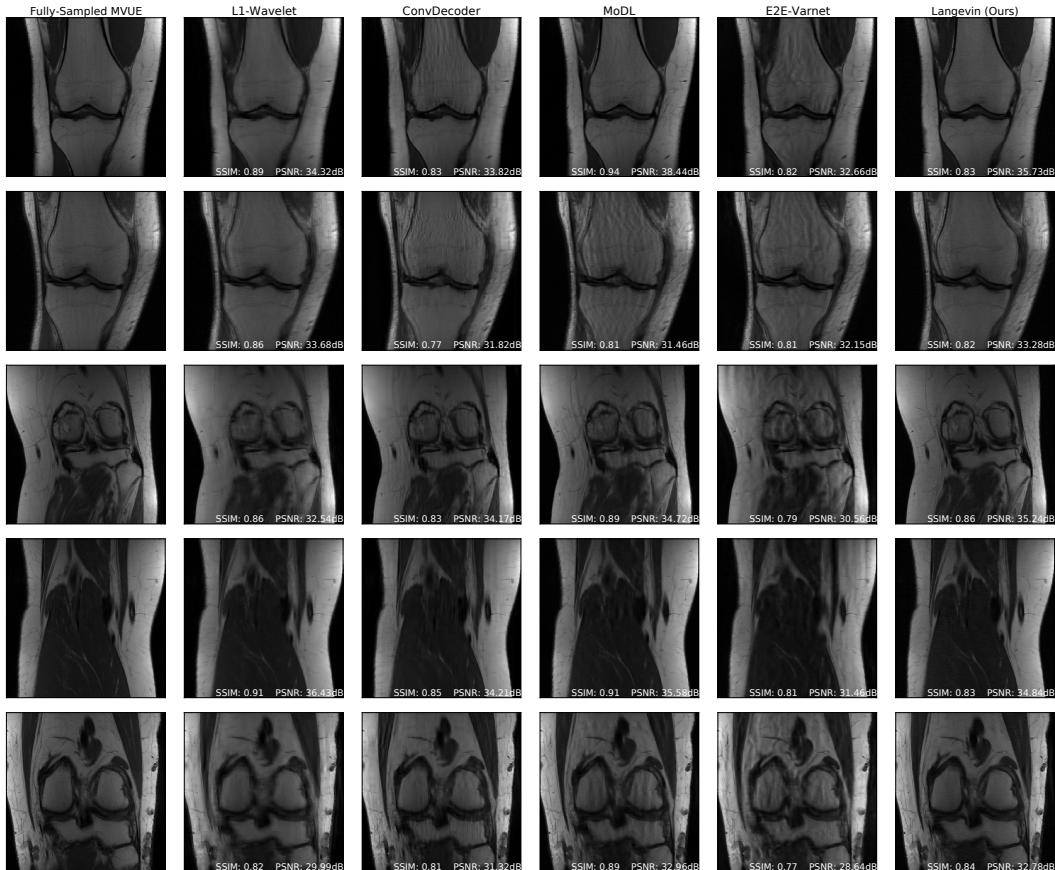


Figure E.11: fastMRI knee reconstructions at an acceleration factor of $R = 4$ and a random vertical mask in k-space. All methods were trained on fastMRI brains, and this shows that our method is more robust than other methods with respect to anatomy shift.

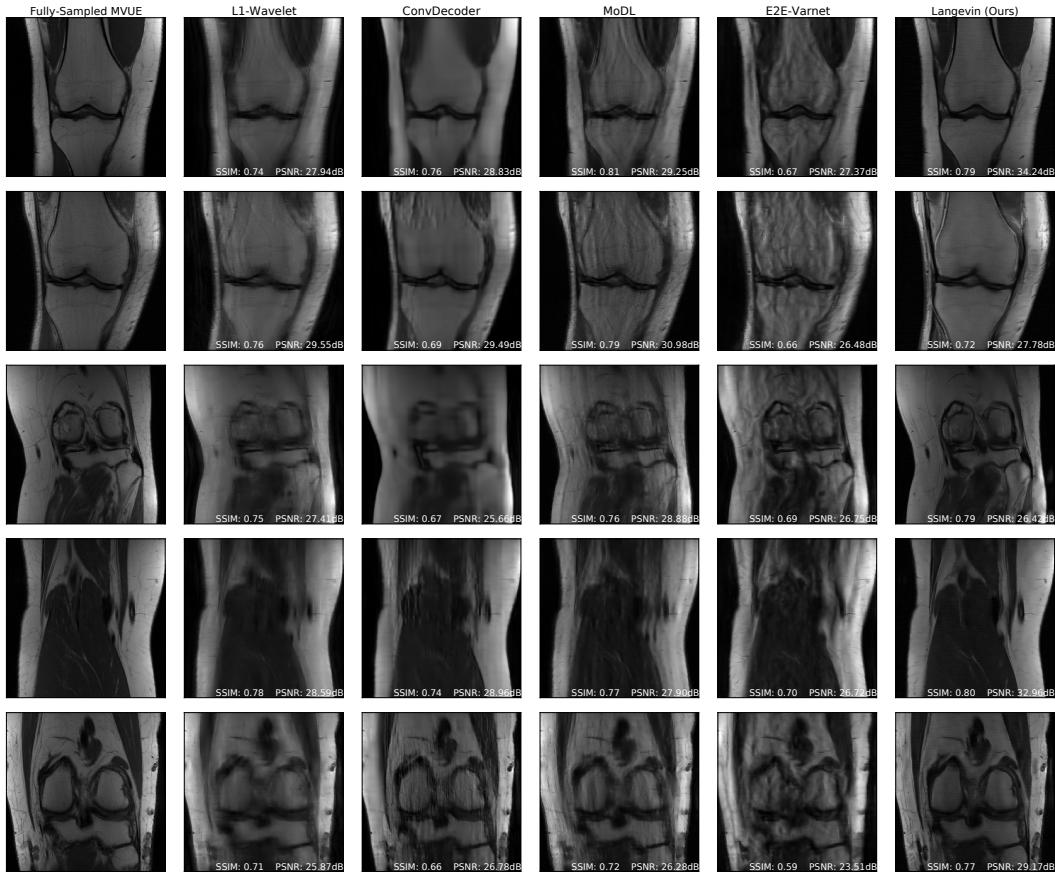


Figure E.12: fastMRI knee reconstructions at an acceleration factor of $R = 8$ and a random vertical mask in k-space. All methods were trained on fastMRI brains, and this shows that our method is more robust than other methods with respect to anatomy shift.

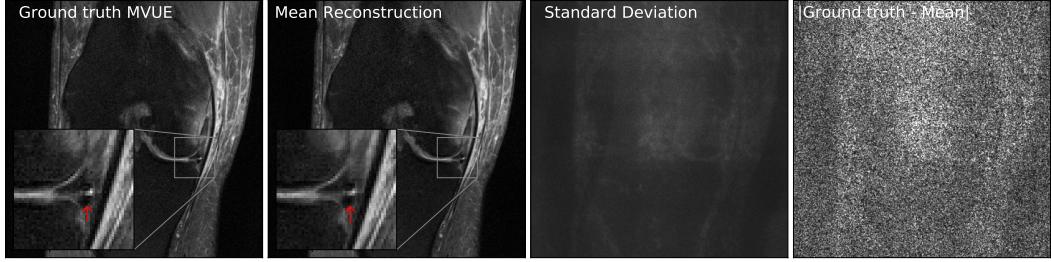


Figure E.13: Our method successfully recovers fine details and can provide an estimate of the reconstruction error. The left column shows a knee from the fastMRI dataset, along with an annotated meniscus tear (indicated by red arrow in zoomed inset). Given measurements at an acceleration factor of $R = 4$, we obtain 48 independent reconstructions via posterior sampling. The second column shows the pixel-wise average of reconstructions, the third column shows the pixel-wise standard deviation, and the fourth column shows the magnitude of the error between the ground truth and the mean reconstruction. Note that our generative prior has never seen such pathology, as it was trained on T2-weighted brain scans.

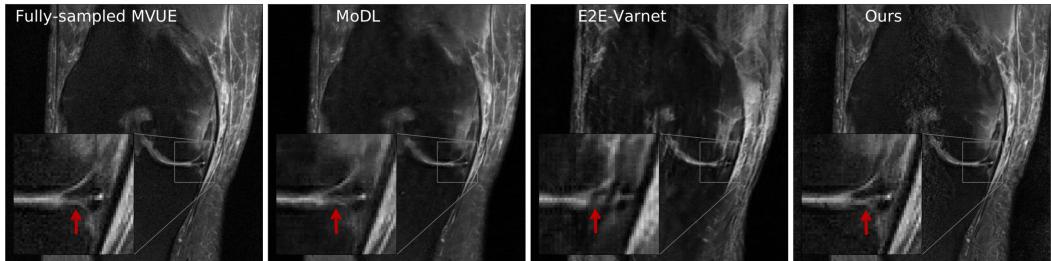


Figure E.14: The left column shows a knee from the fastMRI dataset, along with an annotated meniscus tear (indicated by red arrow in zoomed inset). Given measurements at an acceleration factor of $R = 4$, we observe that our method preserves fine details better than the baselines. None of the methods have seen such a pathology, as they were all trained on T2-weighted brain scans.

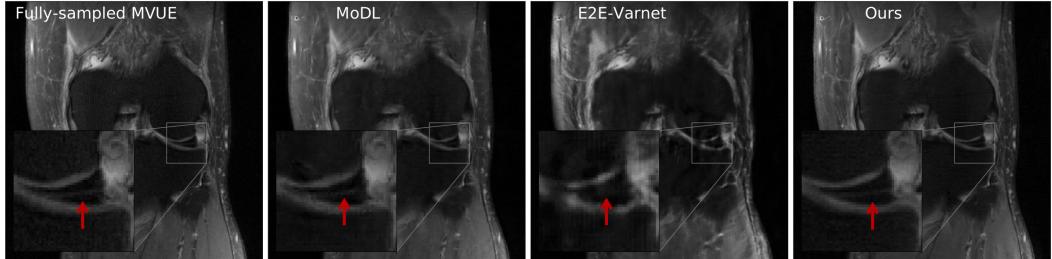


Figure E.15: The left column shows a knee from the fastMRI dataset, along with an annotated meniscus tear (indicated by red arrow in zoomed inset). Given measurements at an acceleration factor of $R = 4$, we observe that our method preserves fine details better than the baselines. None of the methods have seen such a pathology, as they were all trained on T2-weighted brain scans.

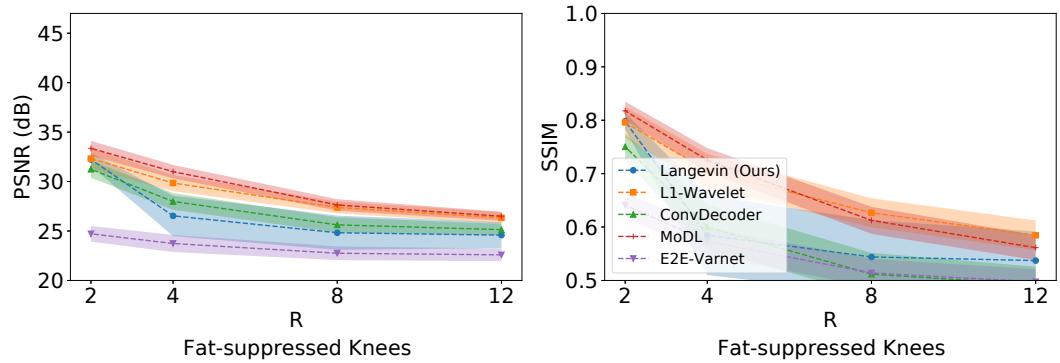


Figure E.16: Average test PSNR and SSIM on fat-suppressed (FS) knees, across a range of acceleration factors R and a random vertical mask in k-space. Higher R indicates a smaller number of acquired measurements. All methods were trained on fastMRI brains. Our approach is not optimal numerically, likely due to a much lower signal-to-noise ratio in FS knees than the brain training data. However, Figures E.14, E.15, E.17, E.18 show that our qualitative reconstructions are competitive, and recover fine details like meniscus tears better than the deep learning baselines. Shaded regions indicate 95% confidence intervals. Note that we trained baselines on MVUE images and hence these numerical values should not be compared with those in literature trained on RSS images (see Appendix E.1.1 for a more detailed discussion).

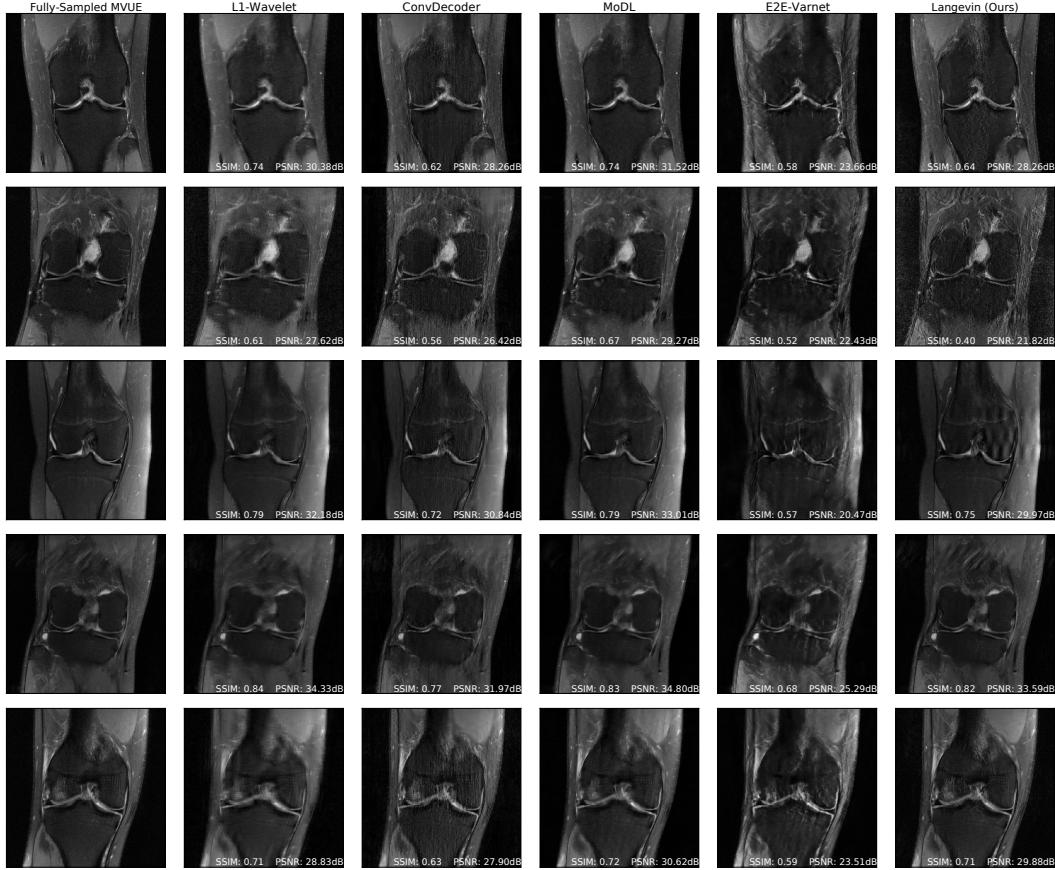


Figure E.17: fastMRI fat-suppressed(FS) knee reconstructions at an acceleration factor of $R = 4$ and a random vertical mask in k-space. All methods were trained on fastMRI brains. Our approach is not optimal numerically, likely due to a much lower signal-to-noise ratio in FS knees than the brain training data. However, the reconstructions in this figure and Figures E.14, E.15, E.18 show that our qualitative reconstructions are competitive, and recovers fine details like meniscus tears better than the deep learning baselines.

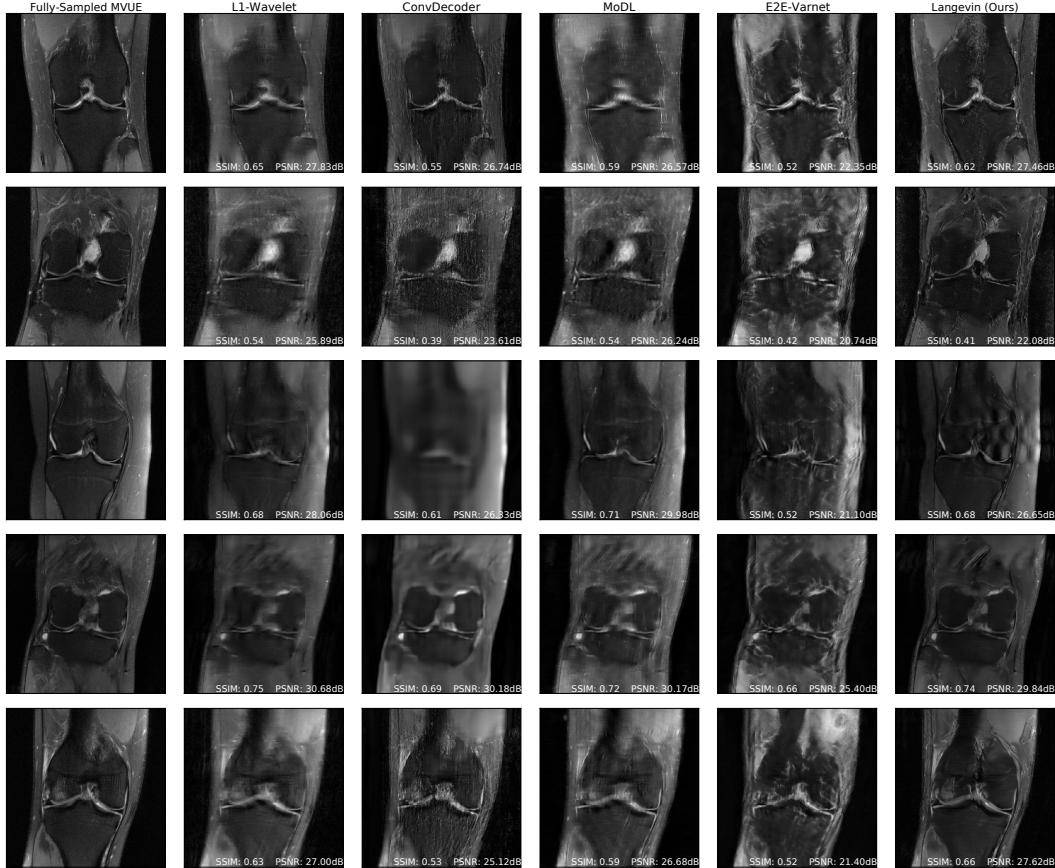


Figure E.18: fastMRI fat-suppressed knee reconstructions at an acceleration factor of $R = 8$ and a random vertical mask in k-space. All methods were trained on fastMRI brains. Our approach is not optimal numerically, likely due to a much lower signal-to-noise ratio in FS knees than the brain training data. However, the reconstructions in this figure and Figures E.14, E.15, E.17 show that our qualitative reconstructions are competitive, and recovers fine details like meniscus tears better than the deep learning baselines.

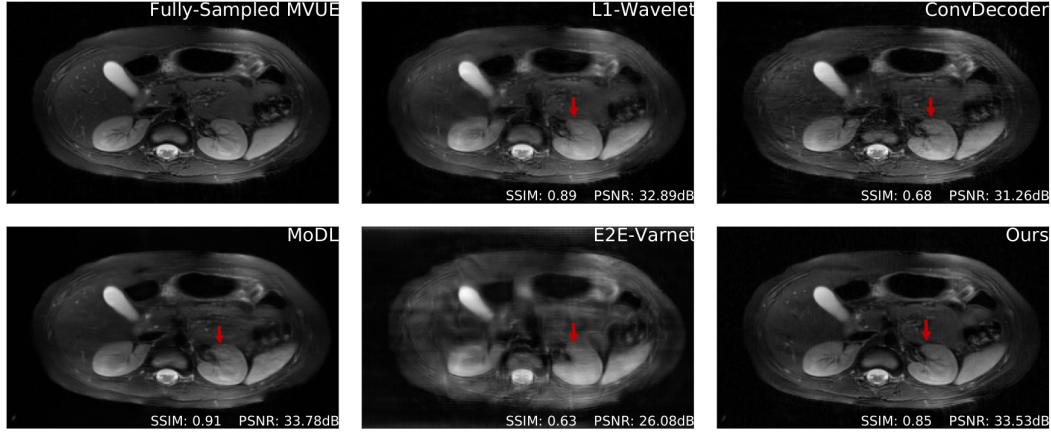


Figure E.19: Comparative reconstructions of a 2D abdominal scan with uniform random under-sampling in the horizontal direction at $R = 4$. None of the methods were trained to reconstruct abdomen MRI. Our method uses a score-based generative model trained on brain images (as explained) and obtains good reconstructions. The red arrows indicate missing details or artifacts in the kidney structure.

E.5 Appendix: Abdomen

Figure E.19 shows an additional example of a reconstructed abdominal scan. This is obtained from the same volume as the figure in the main text, and has a resolution of 158×320 voxels, but a much larger field of view, leading to a resolution shift for all models.

E.6 Appendix: Stanford Knee

Figures E.20 and E.21 show quantitative and qualitative reconstruction under an anatomy shift induced by testing axial knee scans. In this case, we first obtain a complete three-dimensional fast spin echo (3D-FSE) knee scan from the publicly available repository at mridata.org. To obtain two-

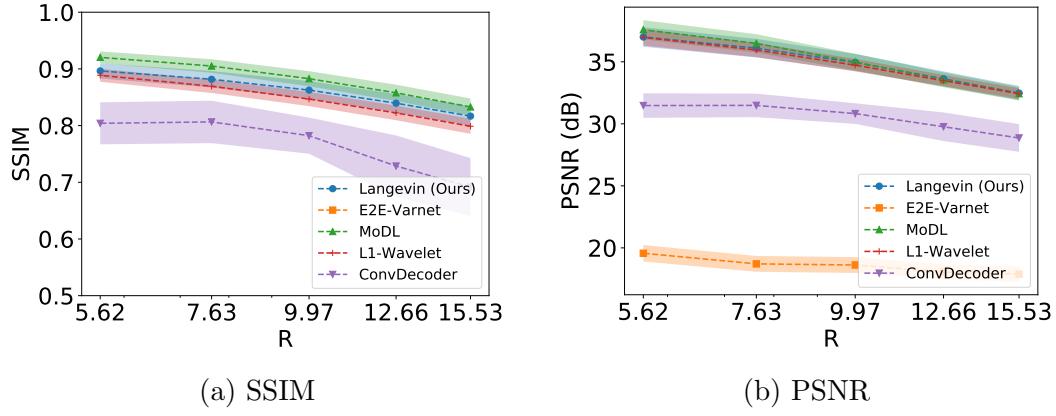


Figure E.20: Reconstruction SSIM and PSNR on Stanford Knees as a function of the acceleration R . This dataset is considerably different from the others, as they are 3D scans. We sample k-space measurements according to Poisson masks, which gives improved incoherence, and hence we find no statistical difference between L1-Wavelet, MoDL, and our method. Note that all hyper-parameter selection and model training was done on brains from the fastMRI dataset.

dimensional slices, we apply an IFFT operator on the readout axis and select 24 equally spaced slices for evaluation. Each slice has a resolution of 320×256 pixels.

E.7 Appendix: Implementation

E.7.1 Score-Based Generative Model

Training the model We use the implementation from <https://github.com/ermongroup/ncsnv2>. As raw MRI scans are complex valued, we changed the generator such that the output and input have two channels, one each for the real and imaginary components. We did not change the architecture otherwise.

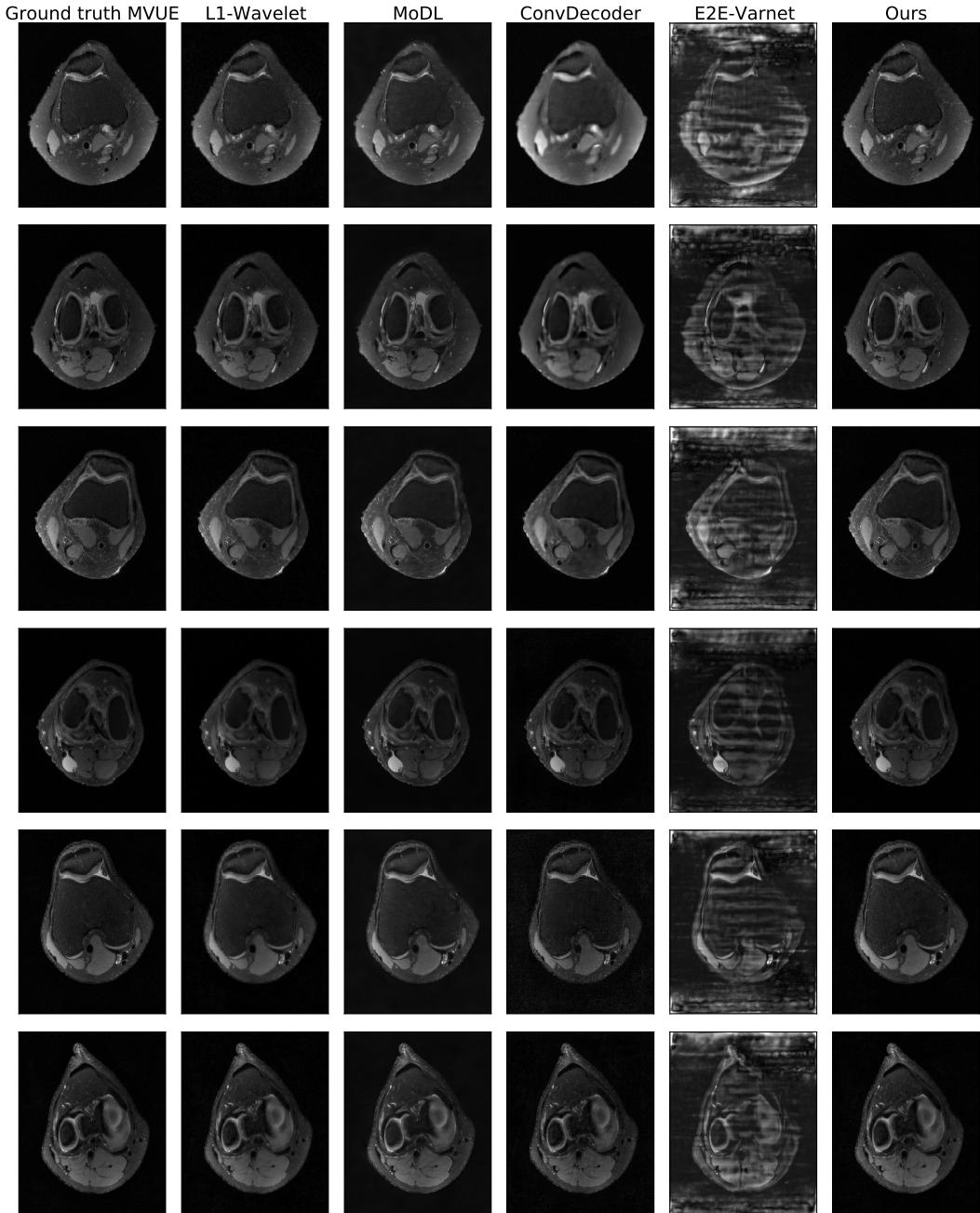


Figure E.21: Qualitative reconstructions obtained by all methods on the Stanford Knees dataset at an acceleration of $R = 5.62$. This dataset is considerably different from the others, as they are 3D scans. We sample k-space measurements according to Poisson masks, which gives improved incoherence, and hence we find no statistical difference between L1-Wavelet, MoDL, and our method. Note that all hyper-parameter selection and model training was done on brains from the fastMRI dataset.

We used the FlickrFaces (FFHQ) configs file from the NCSNv2 repo, except we set `sigma_begin` = 232, and `sigma_end` = 0.0066. This is because of the smaller number of channels in MRI when compared to FFHQ.

Dynamic range of the data. MRI data exhibits a lot of variation in the dynamic range. For example, the fastMRI dataset has max pixel value on the order of 10^{-4} , while the abdomen and Stanford knee data has max pixels on the order of 10^5 . In order to deal with this variation, during *training*, we normalize each image by the 99 percentile pixel value. During inference time, when we do not have access to the ground-truth image, we normalize the reconstruction using the 99 percentile pixel value of the *pseudo-inverse* complex image. We observe that this heuristic is sufficient to get good results.

Invariance to image shapes. Due to the convolutional nature of NCSNv2, although we trained on 384×384 images, we can still apply them to knees, T1-weighted & FLAIR brains, and abdomens, although all of these have different dimension shapes.

Hyperparameters We tuned our hyperparameters on two validation brain scans, at an acceleration of $R = 4$. We then reused these hyperparameters on *all anatomies, all accelerations*. Please see our GitHub link: <https://github.com/utcsilab/csgm-mri-langevin> for the hyperparameter values.

E.7.2 E2E-VarNet Baseline

We use the architecture publicly available in the fastMRI official repository. The backbone for the image reconstruction network is a U-Net with a depth of four stages, and 18 hidden channels in the first stage, for a total of 29 million learnable parameters. This model also include a smaller deep neural network that is used to estimate the sensitivity maps. This is also a U-Net, with four stages, but only eight hidden channels after the first stage, for an additional 0.7 million parameters. The model is trained for a number of 12 unrolls, and separate image networks are used at each unroll.

We train this model from scratch for a number of 40 epochs, using an Adam optimizer with default PyTorch parameters and a learning rate of 2e−4, decayed by 0.5 after 20 epochs, as well as gradient clipping to a maximum magnitude of 1. We use the fully-sampled MVUE reconstructions from the brain T2 contrast in fastMRI to train all methods. We use a batch size of 1 and a supervised SSIM loss between the absolute values of ground truth MVUE and the absolute value of the complex output of the network at acceleration factors $R = \{3, 6\}$ (chosen with equal probability), using a vertical, equispaced sampling pattern, same as all other baselines.

Finally, it is worth mentioning that the network used to estimate the sensitivity maps explicitly uses the fully-sampled, vertical ACS region, as shown in Figure E.4, both during training and inference. This makes testing with other mask patterns non-trivial for this baseline. To alleviate this, we always feed the image obtained from the *vertical* ACS region (for example,

in the case of horizontal masks, we intentionally zero out other sampled lines that would fall in this region), to not introduce incoherent aliasing in this image.

E.7.3 MoDL Baseline

We use the PyTorch MoDL implementation publicly available at <https://github.com/utcsilab/deep-j-sense> and train a MoDL model that uses a backbone residual network with a depth of six layers, three equispaced residual connections (that feed hidden signals from the first three layers to the last three layers) and 64 hidden channels, with a total of 220000 trainable parameters. Unlike E2E-VarNet, the same backbone network is used across all unrolls, and the data consistency term is given by a Conjugate Gradient (CG) operator, truncated to six steps.

We train MoDL for a number of six unrolls, leading to a total of 36 CG steps and six network applications in the unroll. We use the Adam optimizer with default PyTorch parameters and learning rate 2e-4, as well as gradient clipping to a maximum magnitude of 1. We train for 15 epochs and decay the learning rate by 0.5 after 8 epochs, using a batch size of 1 on exactly the same T2 brain scans as all methods and a supervised SSIM loss at $R = \{3, 6\}$ (chosen with equal probability) between the magnitude of the ground-truth MVUE image and the magnitude of the complex network output. We find that, although relatively small, the backbone network architecture is sufficient to achieve good in-distribution reconstruction, and serve as a strong baseline.

Since MoDL and all other methods (including ours) except E2E-VarNet, require external sensitivity map estimates to be provided to them, we use the ESPiRiT algorithm from the BART toolbox [267] without any eigenvalue cropping to estimate a single set of sensitivity maps, one for each coil.

E.8 Appendix: Radiologist Study

We performed a preliminary image quality assessment experiment with two board-certified radiologists and a faculty member that uses neuro-imaging in their research.

The three external experts were not involved with our research and have performed the image quality assessment blindly. Each of them was presented with ten scans from the following anatomies and scan parameters: abdominal scans, knee scans and brain scans with a horizontal readout direction, leading to a total of 30 quality assessment questions. Note that all anatomies represent test-time distributional shifts in at least one aspect.

In each question, the experts were shown four images:

- The fully-sampled reference image, explicitly marked as "Reference".
- The results of three reconstruction algorithms at acceleration factor R=3: MoDL, ConvDecoder and our method. The order of the reconstructions was shuffled for each question, and the reconstructions were labeled as "1", "2" and "3".

Anatomy	MoDL	ConvDec	Ours
Knee	1.87(0.34)	2.97(0.18)	1.17(0.45)
Abdomen	1.87(0.76)	2.17(0.93)	1.97(0.71)
Brain	2.00(0.82)	2.07(0.77)	1.93(0.85)

Table E.1: Ranking of algorithms by experts. A lower ranking is better: the best possible ranking is 1, and the worst 3. The values show the average and standard deviation (in parentheses) of the ranking for each anatomy, using a total of 30 data points (3 participants x 10 scans per anatomy).

We chose to compare with MoDL and ConvDecoder since these method had the best overall quantitative and qualitative (according to our own pre-assessment) robust performance. The participants were instructed to rank the three reconstructions from best to worst quality, while using the "Reference" image as a perceptual guideline. Table E.1 shows the average and standard deviation (in parentheses) of the ranking for each anatomy, obtained using a total of 30 data points (3 participants x 10 scans per anatomy).

In Table E.1, a lower ranking is better, the best possible ranking is 1, and the worst 3. We draw the following conclusions:

- Participants consistently ranked our method as best on the knee scans, which supports the distributional shift robustness claimed in the main paper, and detailed in Appendices E.4, E.5 and E.1.
- Participants did not perceive a significant difference between all methods when applied to abdominal or brain scans with a horizontal phase encode direction. In the brain case, this supports the qualitative results shown in Appendix E.3, Figure E.5.

Anatomy	Ours vs. MoDL	Ours vs. ConvDec
Knee	$1.53e - 10$	$2.77e - 6$
Abdomen	0.610	0.340
Brain	0.767	0.550

Table E.2: p-values from the Wilcoxon Rank Sum test to determine if the rankings of different algorithms are drawn from different populations. There is a significant difference in the case of knees, and no significant difference in the case of abdomens and brains.

- In the abdominal case, this partially correlates with Figure 6.2c, regarding the quantitative tie between our approach and MoDL.

To quantify the statistical significance of the above results, we perform a Wilcoxon Rank Sum test to determine if the rankings of different algorithms are drawn from different populations. We evaluate if our proposed method leads to different rankings than MoDL and the ConvDecoder, and show the p-values in Table E.2.

The results show a significant difference in the case of knees, while no significant difference is present for abdomen and brain. Finally, to evaluate inter-observer agreement between the three reviewers, we calculated the intra-class correlation (ICC) coefficient separately for each anatomy by aggregating the ten questions related to that anatomy and evaluating the ICC2 coefficient [4] in a pairwise manner at a 5% significance level.

The results are shown in Table E.3, where we also include the p-value and the 95% confidence interval for the ICC2 estimate. This indicates that there exists a very strong consensus regarding the ranking on the knee

Anatomy	ICC2	p-value	95% CI
Knee	0.980	0.0004	[0.81, 1]
Abdomen	-0.222	0.576	[-0.89, 0.92]
Brain	-0.818	0.907	[-0.98, 0.59]

Table E.3: p-values and confidence intervals for differences in ranking between our method and baselines.

anatomy, while for abdomen and brain this consensus is much weaker, which together with Table E.2 indicates that the images were considered equivalent.

This preliminary image quality assessment gives additional evidence (in addition to the quantitative metrics of SSIM and PSNR) that our method maintains robustness to distribution shifts at test time. As our quantitative results show, other methods maintain robustness in some but not all cases. Due to time limitations, we were not able to ask the reviewers to evaluate every algorithm and every distribution shift including different levels of acceleration. We stress that this preliminary study is not a substitute for a rigorous clinical evaluation which is necessary before considering using our proposed method in a clinical setting.

Appendix F

Bibliography

Bibliography

- [1] <https://twitter.com/chicken3gg/status/1274314622447820801?lang=en>.
- [2] <http://mridata.org/>.
- [3] <http://mridata.org/>.
- [4] https://pingouin-stats.org/generated/pingouin.intraclass_corr.html.
- [5] <https://github.com/facebookresearch/fastMRI>.
- [6] <https://discuss.fastmri.org/t/annotated-pathologies-in-the-fastmri-knee-data/219>.
- [7] Anders Aamand, Piotr Indyk, and Ali Vakilian. (learned) frequency estimation algorithms under zipfian distribution. *arXiv preprint arXiv:1908.05198*, 2019.
- [8] Shuchin Aeron, Venkatesh Saligrama, and Manqi Zhao. Information theoretic bounds for compressed sensing. *IEEE Transactions on Information Theory*, 56(10):5111–5130, 2010.

- [9] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.
- [10] Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.*, 40(5):2452–2482, 10 2012.
- [11] Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. Modl: Model-based deep learning architecture for inverse problems. *IEEE transactions on medical imaging*, 38(2):394–405, 2018.
- [12] Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.
- [13] Zeyuan Allen-Zhu and Yuanzhi Li. Forward super-resolution: How can gans learn hierarchical generative models for real-world distributions. *arXiv preprint arXiv:2106.02619*, 2021.
- [14] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58(1):137–147, 1999.
- [15] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias

through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019.

- [16] Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, and Timo Bremer. Mimicgan: Robust projection onto image manifolds with corruption mimicking. *arXiv preprint arXiv:1912.07748*, pages 1–19, 2019.
- [17] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen. On instabilities of deep learning in image reconstruction and the potential costs of ai. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020.
- [18] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.
- [19] Marius Arvinte and Jonathan I Tamir. Deep diffusion models for robust channel estimation. *arXiv preprint arXiv:2111.08177*, 2021.
- [20] 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.
- [21] Muhammad Asim, Max Daniels, Oscar Leong, Ali Ahmed, and Paul Hand. Invertible generative models for inverse problems: mitigating representation error and dataset bias. In *International Conference on Machine Learning*, pages 399–409. PMLR, 2020.

- [22] Muhammad Asim, Fahad Shamshad, and Ali Ahmed. Blind image deconvolution using deep generative priors. *arXiv preprint arXiv:1802.04073*, 2018.
- [23] Muhammad Asim, Fahad Shamshad, and Ali Ahmed. Solving bilinear inverse problems using deep generative priors. *CoRR, abs/1802.04073*, 3(4):8, 2018.
- [24] 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.
- [25] Pranjal Awasthi, Matthias Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In *International Conference on Artificial Intelligence and Statistics*, pages 1770–1780. PMLR, 2020.
- [26] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.
- [27] Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Seminaire de probabilités XIX 1983/84*, pages 177–206. Springer, 1985.
- [28] Sivaraman Balakrishnan, Simon S. Du, Jerry Li, and Aarti Singh. Com-

putationally efficient robust sparse estimation in high dimensions. In *Proceedings of the 2017 Conference on Learning Theory*, 2017.

- [29] Eren Balevi, Akash Doshi, Ajil Jalal, Alexandros Dimakis, and Jeffrey G Andrews. High dimensional channel estimation using deep generative networks. *IEEE Journal on Selected Areas in Communications*, 39(1):18–30, 2020.
- [30] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 167–176, 2020.
- [31] Richard G Baraniuk. Compressive sensing [lecture notes]. *IEEE signal processing magazine*, 24(4):118–121, 2007.
- [32] 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.
- [33] Richard G Baraniuk and Michael B Wakin. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.
- [34] 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.

- [35] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2, 2017.
- [36] Jonathan T Barron and Yun-Ta Tsai. Fast fourier color constancy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–894, 2017.
- [37] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.
- [38] Sebastian Bentham and Bruce D Haynes. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 289–298, 2019.
- [39] 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.
- [40] Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? *arXiv preprint arXiv:1912.01094*, 2019.
- [41] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th In-*

ternational Conference on Machine Learning- Volume 70, pages 537–546. JMLR.org, 2017.

- [42] Ashish Bora, Eric Price, and Alexandros G Dimakis. Ambientgan: Generative models from lossy measurements. *ICLR*, 2:5, 2018.
- [43] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.
- [44] Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9–10):589–592, 2008.
- [45] 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.
- [46] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

- [47] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52:489–509, 2004.
- [48] 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.
- [49] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.
- [50] L Elisa Celis, Lingxiao Huang, and Nisheeth K Vishnoi. Fair classification with noisy protected attributes. *arXiv preprint arXiv:2006.04778*, 2020.
- [51] 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.
- [52] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018.

- [53] Chen Chen and Junzhou Huang. Compressive sensing mri with wavelet tree sparsity. In *Advances in neural information processing systems*, pages 1115–1123, 2012.
- [54] 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.
- [55] Guang-Hong Chen, Jie Tang, and Shuai Leng. Prior image constrained compressed sensing (piccs): a method to accurately reconstruct dynamic ct images from highly undersampled projection data sets. *Medical physics*, 35(2):660–663, 2008.
- [56] Guangliang Chen and Deanna Needell. Compressed sensing and dictionary learning. *Proceedings of Symposia in Applied Mathematics*, 73, 2016.
- [57] 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.
- [58] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

- [59] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2492–2501, 2018.
- [60] Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*, pages 774–782, 2013.
- [61] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017.
- [62] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, pages 1887–1898. PMLR, 2020.
- [63] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.
- [64] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22(1):211–231, 2009.

- [65] EK Cole, JM Pauly, SS Vasanawala, and F Ong. Unsupervised mri reconstruction with generative adversarial networks. arxiv 2020. *arXiv preprint arXiv:2008.13065*.
- [66] Elizabeth K Cole, Frank Ong, Shreyas S Vasanawala, and John M Pauly. Fast unsupervised mri reconstruction without fully-sampled ground truth data using generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3988–3997, 2021.
- [67] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [68] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *arXiv preprint arXiv:1611.05644*, 2016.
- [69] Arnak Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1 -penalized huber’s m -estimator. In *Advances in Neural Information Processing Systems*, pages 13188–13198, 2019.
- [70] Giannis Daras, Joseph Dean, Ajil Jalal, and Alexandros G Dimakis. Intermediate layer optimization for inverse problems using deep generative models. *arXiv preprint arXiv:2102.07364*, 2021.
- [71] Mohammad Zalbagi Darestani, Akshay Chaudhari, and Reinhard

Heckel. Measuring robustness in deep learning based compressive sensing. *arXiv preprint arXiv:2102.06103*, 2021.

- [72] Mohammad Zalbagi Darestani and Reinhard Heckel. Can un-trained neural networks compete with trained neural networks at image reconstruction? *arXiv preprint arXiv:2007.02471*, 2020.
- [73] Constantinos Daskalakis, Dhruv Rohatgi, and Emmanouil Zampetakis. Constant-expansion suffices for compressed sensing with generative priors. *Advances in Neural Information Processing Systems*, 33:13917–13926, 2020.
- [74] Anagha Deshmane, Vikas Gulani, Mark A Griswold, and Nicole Seibertlich. Parallel mr imaging. *Journal of Magnetic Resonance Imaging*, 36(1):55–72, 2012.
- [75] Manik Dhar, Aditya Grover, and Stefano Ermon. Modeling sparse deviations for compressed sensing using generative models. *arXiv preprint arXiv:1807.01442*, 2018.
- [76] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018.
- [77] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

- [78] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.
- [79] Kate Donahue and Jon Kleinberg. Fairness and utilization in allocating resources with uncertain demand. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 658–668, 2020.
- [80] Mariya Doneva. Mathematical models for magnetic resonance imaging reconstruction: An overview of the approaches, problems, and future research areas. *IEEE Signal Processing Magazine*, 37(1):24–32, 2020.
- [81] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.
- [82] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.
- [83] 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.
- [84] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.
- [85] Troy Duster. Race and reification in science, 2005.

- [86] Vineet Edupuganti, Morteza Mardani, Shreyas Vasanawala, and John Pauly. Uncertainty quantification in deep mri reconstruction. *IEEE Transactions on Medical Imaging*, 40(1):239–250, 2021.
- [87] 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.
- [88] 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.
- [89] 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.
- [90] Rina Foygel and Lester Mackey. Corrupted sensing: Novel guarantees for separating structured signals. *IEEE Transactions on Information Theory*, 60(2):1223–1247, 2014.
- [91] Eric Frankel and Edward Vendrow. Fair generation through prior modification.
- [92] Anna C. Gilbert, Yi Zhang, Kibok Lee, Yuting Zhang, and Honglak Lee. Towards understanding the invertibility of convolutional neural networks. 2017.

- [93] Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, pages 929–989, 1984.
- [94] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.
- [95] Fabian Latorre Gómez, Armin Eftekhari, and Volkan Cevher. Fast and provable admm for learning with generative priors. *arXiv preprint arXiv:1907.03343*, 2019.
- [96] Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *European Conference on Computer Vision*, pages 330–347. Springer, 2020.
- [97] Sixue Gong, Xiaoming Liu, and Anil K Jain. Mitigating face recognition bias via group adaptive classifier. *arXiv preprint arXiv:2006.07576*, 2020.
- [98] 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.
- [99] Mark A. Griswold, Peter M. Jakob, Robin M. Heidemann, Mathias Nitka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase.

- Generalized autocalibrating partially parallel acquisitions (grappa). *Magnetic Resonance in Medicine*, 47(6):1202–1210, 2002.
- [100] Aditya Grover, Kristy Choi, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. 2019.
- [101] Matthieu Guerquin-Kern, Laurent Lejeune, Klaas Paul Pruessmann, and Michael Unser. Realistic analytical phantoms for parallel magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 31(3):626–636, 2011.
- [102] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.
- [103] Kerstin Hammernik, Jo Schlemper, Chen Qin, Jinming Duan, Ronald M. Summers, and Daniel Rueckert. Systematic evaluation of iterative deep neural networks for fast parallel mri reconstruction with sensitivity-weighted coil combination. *Magnetic Resonance in Medicine*, n/a(n/a), 2021.
- [104] Paul Hand and Babhru Joshi. Global guarantees for blind demodulation with generative priors. In *Advances in Neural Information Processing Systems*, pages 11531–11541, 2019.

- [105] Paul Hand, Oscar Leong, and Vlad Voroninski. Phase retrieval under a generative prior. In *Advances in Neural Information Processing Systems*, pages 9136–9146, 2018.
- [106] Paul Hand and Vladislav Voroninski. Global guarantees for enforcing deep generative priors by empirical risk. *arXiv preprint arXiv:1705.07576*, 2017.
- [107] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 501–512, 2020.
- [108] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS’16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.
- [109] Ralph VL Hartley. Transmission of information 1. *Bell System technical journal*, 7(3):535–563, 1928.
- [110] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.
- [111] Reinhard Heckel and Paul Hand. Deep decoder: Concise image repre-

- sentations from untrained non-convolutional networks. *arXiv preprint arXiv:1810.03982*, 2018.
- [112] Reinhard Heckel and Mahdi Soltanolkotabi. Denoising and regularization via exploiting the structural bias of convolutional generators. *arXiv preprint arXiv:1910.14634*, 2019.
- [113] Reinhard Heckel and Mahdi Soltanolkotabi. Compressive sensing with un-trained neural networks: Gradient descent finds the smoothest approximation. *arXiv preprint arXiv:2005.03991*, 2020.
- [114] 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.
- [115] Chinmay Hegde and Richard G Baraniuk. Signal recovery on incoherent manifolds. *IEEE Transactions on Information Theory*, 58(12):7204–7214, 2012.
- [116] 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.
- [117] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. A nearly-linear time framework for graph-structured sparsity. In *Proceedings of the 32nd*

International Conference on Machine Learning (ICML-15), pages 928–937, 2015.

- [118] Chinmay Hegde, Michael Wakin, and Richard G Baraniuk. Random projections for manifold learning. In *Advances in neural information processing systems*, pages 641–648, 2008.
- [119] Junia Howell and Michael O. Emerson. So what “should” we use? evaluating the impact of five racial measures on markers of social inequality. *Sociology of Race and Ethnicity*, 3(1):14–30, 2017.
- [120] Chen-Yu Hsu, Piotr Indyk, Dina Katabi, and Ali Vakilian. Learning-based frequency estimation algorithms. 2018.
- [121] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.
- [122] Feng Huang, Sathya Vijayakumar, Yu Li, Sarah Hertel, and George R Duensing. A software channel compression technique for faster reconstruction with many channels. *Magnetic resonance imaging*, 26(1):133–141, 2008.
- [123] Peter J Huber. Robust estimation of a location parameter. *The annals of mathematical statistics*, pages 73–101, 1964.

- [124] Shady Abu Hussein, Tom Tirer, and Raja Giryes. Image-adaptive gan based reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3121–3129, 2020.
- [125] Sunhee Hwang and Hyeran Byun. Unsupervised image-to-image translation via fair representation of gender bias. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1953–1957. IEEE, 2020.
- [126] Rakib Hyder, Viraj Shah, Chinmay Hegde, and M Salman Asif. Alternating phase projected gradient descent with generative priors for solving compressive phase retrieval. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7705–7709. IEEE, 2019.
- [127] Piotr Indyk, Eric Price, and David P Woodruff. On the power of adaptivity in sparse recovery. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 285–294. IEEE, 2011.
- [128] Piotr Indyk, Ali Vakilian, and Yang Yuan. Learning-based low-rank approximations. In *Advances in Neural Information Processing Systems*, pages 7400–7410, 2019.
- [129] MA Iwen and AH Tewfik. Adaptive group testing strategies for target detection and localization in noisy environments. 2010.

- [130] Siddharth Iyer, Frank Ong, Kawin Setsompop, Mariya Doneva, and Michael Lustig. Sure-based automatic parameter selection for espirit calibration. *Magnetic Resonance in Medicine*, 84(6):3423–3437, 2020.
- [131] Mathews Jacob, Jong Chul Ye, Leslie Ying, and Mariya Doneva. Computational mri: Compressive sensing and beyond [from the guest editors]. *IEEE Signal Processing Magazine*, 37(1):21–23, 2020.
- [132] Gauri Jagatap and Chinmay Hegde. Phase retrieval using untrained neural network priors. 2019.
- [133] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jonathan Tamir. Robust compressed sensing mri with deep generative priors. *Advances in Neural Information Processing Systems*, 34, 2021.
- [134] Ajil Jalal, Sushrut Karmalkar, Alexandros G Dimakis, and Eric Price. Instance-optimal compressed sensing via posterior sampling. *arXiv preprint arXiv:2106.11438*, 2021.
- [135] Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alexandros G Dimakis, and Eric Price. Fairness for image generation with uncertain sensitive attributes. *arXiv preprint arXiv:2106.12182*, 2021.
- [136] Ajil Jalal, Liu Liu, Alexandros G Dimakis, and Constantine Caramanis. Robust compressed sensing using generative models. *Advances in Neural Information Processing Systems*, 33, 2020.

- [137] 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.
- [138] Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.
- [139] Kyong Hwan Jin, Michael T. McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.
- [140] Maya Kabkab, Pouya Samangouei, and Rama Chellappa. Task-aware compressed sensing with generative adversarial networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [141] Mohammad Mahdi Kamani, Farzin Haddadpour, Rana Forsati, and Mehrdad Mahdavi. Efficient fair principal component analysis. *arXiv preprint arXiv:1911.04931*, 2019.
- [142] Akshay Kamath, Sushrut Karmalkar, and Eric Price. Lower bounds for compressed sensing with generative models. *arXiv preprint arXiv:1912.02938*, 2019.
- [143] Akshay Kamath, Eric Price, and Sushrut Karmalkar. On the power of compressed sensing with generative models. In *International Conference on Machine Learning*, pages 5101–5109. PMLR, 2020.

- [144] 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.
- [145] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020.
- [146] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
- [147] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.
- [148] Jay S Kaufman. How inconsistencies in racial classification demystify the race construct in public health statistics. *Epidemiology*, pages 101–103, 1999.
- [149] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828, 2015.

- [150] Varun A Kelkar and Mark A Anastasio. Prior image-constrained reconstruction using style-based generative models. *arXiv preprint arXiv:2102.12525*, 2021.
- [151] Varun A Kelkar, Sayantan Bhadra, and Mark A Anastasio. Compressible latent-space invertible networks for generative model-constrained image reconstruction. *IEEE Transactions on Computational Imaging*, 7:209–223, 2021.
- [152] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.
- [153] Moein Khajehnejad, Ahmad Asgharian Rezaei, Mahmoudreza Babaei, Jessica Hoffmann, Mahdi Jalili, and Adrian Weller. Adversarial graph embeddings for fair influence maximization over social networks.
- [154] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012.
- [155] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016.

- [156] Taehoon Kim. A tensorflow implementation of “deep convolutional generative adversarial networks”. <https://github.com/carpedm20/DCGAN-tensorflow>, 2017.
- [157] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [158] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [159] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.
- [160] Jon Kleinberg and Manish Raghavan. Selection problems in the presence of implicit bias. *arXiv preprint arXiv:1801.03533*, 2018.
- [161] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.
- [162] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, et al. fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020.

- [163] Alexandre Louis Lamy, Ziyuan Zhong, Aditya Krishna Menon, and Nakul Verma. Noise-tolerant fair classification. *arXiv preprint arXiv:1901.10837*, 2019.
- [164] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.
- [165] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *arXiv preprint arXiv:1711.10306*, 2017.
- [166] Guillaume Lecué and Shahar Mendelson. Sparse recovery under weak moment assumptions. *arXiv preprint arXiv:1401.2188*, 2014.
- [167] Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- [168] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [169] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of*

- the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.
- [170] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.
- [171] 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.
- [172] Xiaodong Li. Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37(1):73–99, 2013.
- [173] Orly Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiupei He, Jonathan T Barron, Dillon Sharlet, Ryan Geiss, Samuel W Hasinoff, and Marc Levoy. Handheld mobile photography in very low light. *ACM Transactions on Graphics (TOG)*, 38(6):1–16, 2019.
- [174] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.
- [175] 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.
- [176] Zachary C Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782*, 2017.
- [177] Liu Liu, Tianyang Li, and Constantine Caramanis. High dimensional robust m-estimation: Arbitrary corruption and heavy tails. *arXiv preprint arXiv:1901.08237*, 2019.
- [178] Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. *arXiv preprint arXiv:1805.11643*, 2018.
- [179] 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.
- [180] Zhaoqiang Liu and Jonathan Scarlett. Information-theoretic lower bounds for compressive sensing with generative models. *arXiv preprint arXiv:1908.10744*, 2019.
- [181] Zhaoqiang Liu and Jonathan Scarlett. Sample complexity lower bounds for compressive sensing with generative models. In *NeurIPS 2019 Workshop on Solving Inverse Problems with Deep Networks*, 2019.

- [182] Zhaoqiang Liu and Jonathan Scarlett. Information-theoretic lower bounds for compressive sensing with generative models. *IEEE Journal on Selected Areas in Information Theory*, 1(1):292–303, 2020.
- [183] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.
- [184] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15:2018, 2018.
- [185] Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust m -estimators. *The Annals of Statistics*, 45(2):866–896, 2017.
- [186] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.
- [187] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*, 2016.
- [188] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22(3):925–965, 2019.

- [189] 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.
- [190] Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly. Compressed sensing mri. *IEEE signal processing magazine*, 25(2):72–82, 2008.
- [191] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.
- [192] Morteza Mardani, Enhao Gong, Joseph Y Cheng, Shreyas Vasanawala, Greg Zaharchuk, Marcus Alley, Neil Thakur, Song Han, William Dally, John M Pauly, et al. Deep generative adversarial networks for compressed sensing automates mri. *arXiv preprint arXiv:1706.00051*, 2017.
- [193] 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.
- [194] Allister Mason, James Rioux, Sharon E Clarke, Andreu Costa, Matthias Schmidt, Valerie Keough, Thien Huynh, and Steven Beyea. Comparison of objective image quality metrics to expert radiologists’ scoring of

- diagnostic quality of mr images. *IEEE transactions on medical imaging*, 39(4):1064–1072, 2019.
- [195] Jiri Matousek. *Lectures on discrete geometry*, volume 212. Springer Science & Business Media, 2002.
- [196] Shahar Mendelson. Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, pages 29–43. Springer, 2002.
- [197] Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.
- [198] Shahar Mendelson. On aggregation for heavy-tailed classes. *Probability Theory and Related Fields*, 168(3-4):641–674, 2017.
- [199] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2437–2445, 2020.
- [200] 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.
- [201] Charles W Mills. *The racial contract*. Cornell University Press, 2014.

- [202] Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.
- [203] Madison Moore. Microsoft deletes ai chatbot after racist, homophobic tweets, according to report. *SD Times, March*, 2016.
- [204] Jack Morse. Google’s ai has some seriously messed up opinions about homosexuality. *Mashable, October*, 2017.
- [205] 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.
- [206] Matthew J. Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, Simon Arberet, Dominik Nickel, Zacccharie Ramzi, Philippe Ciuciu, Jean-Luc Starck, Jonas Teuwen, Dimitrios Karkalousos, Chaoping Zhang, Anuroop Sriram, Zhengnan Huang, Nafissa Yakubova, Yvonne W. Lui, and Florian Knoll. Results of the 2020 fastmri challenge for machine learning mr image reconstruction. *IEEE Transactions on Medical Imaging*, pages 1–1, 2021.
- [207] Dominik Narnhofer, Kerstin Hammernik, Florian Knoll, and Thomas Pock. Inverse gans for accelerated mri reconstruction. In *Wavelets and Sparsity XVIII*, volume 11138, page 111381A. International Society for Optics and Photonics, 2019.

- [208] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m -estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.
- [209] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.
- [210] Nam H Nguyen and Trac D Tran. Exact recoverability from dense corrupted observations via l1-minimization. *IEEE transactions on information theory*, 59(4):2017–2035, 2013.
- [211] 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.
- [212] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.
- [213] REAC Paley and Antoni Zygmund. A note on analytic functions in the unit circle. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pages 266–272. Cambridge University Press, 1932.
- [214] 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.

- [215] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019.
- [216] Andrew M Penner and Aliya Saperstein. Disentangling the effects of racial self-identification and classification by others: the case of arrest. *Demography*, 52(3):1017–1024, 2015.
- [217] Yakov B Pesin. *Dimension theory in dynamical systems: contemporary views and applications*. University of Chicago Press, 2008.
- [218] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
- [219] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and, 6(2012-2016):7*, 2014.
- [220] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.
- [221] 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.
- [222] Klaas P Pruessmann, Markus Weiger, Markus B Scheidegger, and Peter Boesiger. Sense: sensitivity encoding for fast mri. *Magnetic Resonance in Medicine*, 42(4):596–603, 1999.

Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 42(5):952–962, 1999.

- [223] 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.
- [224] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.
- [225] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [226] Saiprasad Ravishankar and Yoram Bresler. Mr image reconstruction from highly undersampled k-space data by dictionary learning. *IEEE Transactions on Medical Imaging*, 30(5):1028–1041, 2011.
- [227] Saiprasad Ravishankar and Jeffrey A. Fessler. Data-driven models and approaches for imaging. In *Imaging and Applied Optics 2017 (3D, AIO, COSI, IS, MATH, pcAOP)*, page MW2C.4. Optical Society of America, 2017.
- [228] Galen Reeves and Michael Gastpar. The sampling rate-distortion trade-

- off for sparsity pattern recovery in compressed sensing. *IEEE Transactions on Information Theory*, 58(5):3065–3092, 2012.
- [229] JH Rick Chang, Chun-Liang Li, Barnabas Poczos, BVK Vijaya Kumar, and Aswin C Sankaranarayanan. One network to solve them all—solving linear inverse problems using deep projection models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5888–5897, 2017.
- [230] Esther Rolf, Max Simchowitz, Sarah Dean, Lydia T Liu, Daniel Bjorkgren, Moritz Hardt, and Joshua Blumenstock. Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning. In *International Conference on Machine Learning*, pages 8158–8168. PMLR, 2020.
- [231] Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017.
- [232] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [233] Sebastian Rosenzweig, Hans Christian Martin Holme, Robin N Wilke, Dirk Voit, Jens Frahm, and Martin Uecker. Simultaneous multi-slice

- mri using cartesian and radial flash and regularized nonlinear inversion: Sms-nlinv. *Magnetic resonance in medicine*, 79(4):2057–2066, 2018.
- [234] Wendy D Roth. The multiple dimensions of race. *Ethnic and Racial Studies*, 39(8):1310–1338, 2016.
- [235] Joni Salminen, Soon-gyo Jung, Shammur Chowdhury, and Bernard J Jansen. Analyzing demographic bias in artificially generated facial pictures. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020.
- [236] Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair pca: One extra dimension. In *Advances in neural information processing systems*, pages 10976–10987, 2018.
- [237] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan. *arXiv preprint arXiv:1805.09910*, 2018.
- [238] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019.
- [239] 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.

- [240] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–35, 2020.
- [241] Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony N Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE transactions on Medical Imaging*, 37(2):491–503, 2017.
- [242] Ignacio Serna, Aythami Morales, Julian Fierrez, Manuel Cebrian, Nick Obradovich, and Iyad Rahwan. Sensitiveloss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *arXiv preprint arXiv:2004.11246*, 2020.
- [243] Abigail A Sewell. The racism-race reification process: A mesolevel political economic framework for understanding racial health disparities. *Sociology of Race and Ethnicity*, 2(4):402–432, 2016.
- [244] Viraj Shah and Chinmay Hegde. Solving linear inverse problems using gan priors: An algorithm with provable guarantees. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4609–4613. IEEE, 2018.
- [245] Claude E Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

- [246] Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. *arXiv preprint arXiv:1810.11874*, 2018.
- [247] Efrat Shimron, Jonathan Tamir, Ke Wang, and Michael Lustig. Subtle inverse crimes: Naively using publicly available images could make reconstruction results seem misleadingly better! *Proceedings of The ISMRM*, 2021.
- [248] Efrat Shimron, Jonathan I Tamir, Ke Wang, and Michael Lustig. Subtle inverse crimes: Naively training machine learning algorithms could lead to overly-optimistic results. *arXiv preprint arXiv:2109.08237*, 2021.
- [249] Tom Simonite. When it comes to gorillas, google photos remains blind. *Wired*, January, 11, 2018.
- [250] Andrew Smart, Richard Tutton, Paul Martin, George TH Ellison, and Richard Ashcroft. The standardization of race and ethnicity in biomedical science editorials and uk biobanks. *Social Studies of Science*, 38(3):407–423, 2008.
- [251] Daniel K Sodickson and Warren J Manning. Simultaneous acquisition of spatial harmonics (smash): fast imaging with radiofrequency coil arrays. *Magnetic resonance in medicine*, 38(4):591–603, 1997.
- [252] Ganlin Song, Zhou Fan, and John Lafferty. Surfing: Iterative optimization

- tion over incrementally trained deep networks. In *Advances in Neural Information Processing Systems*, pages 15008–15017, 2019.
- [253] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 11918–11930. Curran Associates, Inc., 2019.
- [254] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *arXiv preprint arXiv:2006.09011*, 2020.
- [255] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.
- [256] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.
- [257] Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated mri reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 64–73. Springer, 2020.

- [258] Anuroop Sriram, Jure Zbontar, Tullie Murrell, C Lawrence Zitnick, Aaron Defazio, and Daniel K Sodickson. Grappanet: Combining parallel imaging with deep learning for multi-coil mri reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14315–14322, June 2020.
- [259] Michel Talagrand. A new isoperimetric inequality and the concentration of measure phenomenon. In *Geometric Aspects of Functional Analysis*, pages 94–124. Springer, 1991.
- [260] Shuhan Tan, Yujun Shen, and Bolei Zhou. Improving the fairness of deep generative models without retraining. *arXiv preprint arXiv:2012.04842*, 2020.
- [261] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2020.
- [262] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognition Letters*, 140:332–338, 2020.
- [263] Robert Tibshirani. Regression shrinkage and selection via the lasso.

Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

- [264] Subarna Tripathi, Zachary C Lipton, and Truong Q Nguyen. Correction by projection: Denoising images with generative adversarial networks. *arXiv preprint arXiv:1803.04477*, 2018.
- [265] Joel A Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Springer, 2015.
- [266] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.
- [267] Martin Uecker, Christian Holme, Moritz Blumenthal, Xiaoqing Wang, Zhengguo Tan, Nick Scholand, Siddharth Iyer, Jon Tamir, and Michael Lustig. mrrecon/bart: version 0.7.00, March 2021.
- [268] Martin Uecker, Peng Lai, Mark J Murphy, Patrick Virtue, Michael Elad, John M Pauly, Shreyas S Vasanawala, and Michael Lustig. Espirit—an eigenvalue approach to autocalibrating parallel mri: where sense meets grappa. *Magnetic resonance in medicine*, 71(3):990–1001, 2014.
- [269] Martin Uecker, Frank Ong, Jonathan I Tamir, Dara Bahri, Patrick Virtue, Joseph Y Cheng, Tao Zhang, and Michael Lustig. Berkeley

- advanced reconstruction toolbox. In *Proc. Intl. Soc. Mag. Reson. Med*, volume 23, 2015.
- [270] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.
- [271] Ananya Uppal, Shashank Singh, and Barnabas Poczos. Nonparametric density estimation & convergence rates for gans under besov ipm losses. *Advances in Neural Information Processing Systems*, 32:9089–9100, 2019.
- [272] Dave Van Veen, Ajil Jalal, Mahdi Soltanolkotabi, Eric Price, Sriram Vishwanath, and Alexandros G Dimakis. Compressed sensing with deep image prior and learned regularization. *arXiv preprint arXiv:1806.06438*, 2018.
- [273] Shreyas S. Vasanawala, Marcus T. Alley, Brian A. Hargreaves, Richard A. Barth, John M. Pauly, and Michael Lustig. Improved pediatric mr imaging with compressed sensing. *Radiology*, 256(2):607–616, 2010. PMID: 20529991.
- [274] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
- [275] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

- [276] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- [277] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. *arXiv preprint arXiv:2011.00379*, 2020.
- [278] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael I Jordan. Robust optimization for fairness with noisy protected groups. *arXiv preprint arXiv:2002.09343*, 2020.
- [279] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019.
- [280] Zhou Wang and Alan C Bovik. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2(1):1–156, 2006.
- [281] Xiaohan Wei, Zhuoran Yang, and Zhaoran Wang. On the statistical rate of nonlinear recovery in generative models with heavy-tailed data. In *International Conference on Machine Learning*, pages 6697–6706, 2019.
- [282] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

- [283] Bihan Wen, Saiprasad Ravishankar, Luke Pfister, and Yoram Bresler. Transform learning for magnetic resonance image reconstruction: From model-based learning to building neural networks. *IEEE Signal Processing Magazine*, 37(1):41–53, 2020.
- [284] Catherine Westbrook. *Handbook of MRI technique*. John Wiley & Sons, 2014.
- [285] Jay Whang, Qi Lei, and Alexandros G Dimakis. Compressed sensing with invertible generative models and dependent noise. *arXiv preprint arXiv:2003.08089*, 2020.
- [286] Yihong Wu. *Shannon theory for compressed sensing*. Citeseer, 2011.
- [287] Yihong Wu and Sergio Verdú. Optimal phase transitions in compressed sensing. *IEEE Transactions on Information Theory*, 58(10):6241–6263, 2012.
- [288] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE, 2018.
- [289] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan+: Achieving fair data generation and classification through generative adversarial nets. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1401–1406. IEEE, 2019.

- [290] 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.
- [291] Weiyu Xu, Enrique Mallada, and Ao Tang. Compressive sensing over graphs. In *2011 Proceedings IEEE INFOCOM*, pages 2087–2095. IEEE, 2011.
- [292] Forest Yang, Moustapha Cisse, and Sanmi Koyejo. Fairness with overlapping groups. *arXiv preprint arXiv:2006.13485*, 2020.
- [293] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.
- [294] Raymond Yeh, Chen Chen, Teck Yian Lim, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*, 2016.
- [295] Jirong Yi, Anh Duc Le, Tianming Wang, Xiaodong Wu, and Weiyu Xu. Outlier detection using generative models with theoretical performance guarantees, 2018.
- [296] Ning Yu, Ke Li, Peng Zhou, Jitendra Malik, Larry Davis, and Mario Fritz. Inclusive gan: Improving data and minority coverage in generative

- models. In *European Conference on Computer Vision*, pages 377–393. Springer, 2020.
- [297] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastmri: An open dataset and benchmarks for accelerated mri. 2018.
- [298] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.
- [299] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.
- [300] Jian Zhang and Ioannis Mitliagkas. Yellowfin and the art of momentum tuning. *arXiv preprint arXiv:1706.03471*, 2017.
- [301] 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.

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