QUANTIZED COMPRESSED SENSING WITH SCORE-BASED GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the general problem of recovering a high-dimensional signal from noisy quantized measurements. Quantization, especially coarse quantization such as one-bit sign measurements, leads to severe information loss and thus a good prior knowledge of the unknown signal is helpful for accurate recovery. Motivated by the power of score-based generative models (SGM, also known as diffusion models) in capturing the rich structure of natural signals beyond simple sparsity, we propose an unsupervised data-driven approach called quantized compressed sensing with SGM (QCS-SGM), where the prior distribution is modeled by a pre-trained SGM. To perform posterior sampling, an annealed likelihood score called *noise perturbed likelihood score* is introduced and combined with the prior score of SGM. The proposed QCS-SGM applies to arbitrary number of quantization bits. Experiments on a variety of baseline datasets demonstrate that the proposed QCS-SGM significantly outperforms existing state-of-the-art algorithms by a large margin for both in-distribution and out-of-distribution samples. Moreover, as a posterior sampling method, QCS-SGM can be easily used to obtain confidence intervals or uncertainty estimates of the reconstructed results. *Our code will be open-sourced after acceptance.*

1 INTRODUCTION

Many problems in science and engineering such as signal processing, computer vision, machine learning, statistics can be cast as linear inverse problems:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ is a known linear mixing matrix, $\mathbf{n} \sim \mathcal{N}(\mathbf{n}; 0, \sigma^2\mathbf{I})$ is an i.i.d. additive Gaussian noise with known variance, and the goal is to recover an unknown signal $\mathbf{x} \in \mathbb{R}^{N \times 1}$ from the noisy linear measurements $\mathbf{y} \in \mathbb{R}^{M \times 1}$. Among various applications, compressed sensing (CS) provides a highly efficient paradigm which makes it possible to recover a high-dimensional signal from a far smaller number of $M \ll N$ measurements (Candès & Wakin, 2008). The underlying wisdom of CS is to leverage the intrinsic structure of the unknown signal $\mathbf{x}$ as some prior information to aid the recovery. One of the most widely used structure is *sparsity*, i.e., most elements of $\mathbf{x}$ are zero under certain transform domains, e.g., wavelet and Fourier transformation (Candès & Wakin, 2008). In other words, the standard CS exploits the fact that many natural signals in real-world applications are (approximately) sparse or compressible. This direction has spurred a huge active field of research during the last two decades, including efficient algorithm design (Tibshirani, 1996; Beck & Teboulle, 2009; Tropp & Wright, 2010; Kabashima, 2003; Donoho et al., 2009), theoretical analysis (Candès et al., 2006; Donoho, 2006; Bach et al., 2012), as well as all kinds of applications (Lustig et al., 2007; 2008), just to name a few.

Despite the significant success of the traditional CS, sparsity-based methods are still limited by the achievable rates since the sparsity assumptions, whether being naive sparsity or block sparsity (Duarte & Eldar, 2011), are still too simple to capture the complex and rich structure of natural signals of interest. For example, most natural signals are not strictly sparse even in the specified transform domains, and thus simply relying on sparsity alone for reconstruction could lead to inaccurate results. Indeed, researchers have proposed to combine sparsity with additional structure assumptions, such as low-rank assumption (Fazel et al., 2008; Foygel & Mackey, 2014), total variation (Candès et al.,

2006; Tang et al., 2009), to further improve the reconstruction performances. Nevertheless, these hand-crafted priors can only apply, if not exactly, to a very limited range of signals and are difficult to generalize to other cases. To address this problem, driven by the success of deep learning, in particular deep generative models (Goodfellow et al., 2014; Kingma & Welling, 2013; Rezende & Mohamed, 2015), there has been a surge of interests in developing CS methods with data-driven priors (Bora et al., 2017; Hand & Joshi, 2019; Asim et al., 2020; Pan et al., 2021). The basic idea is that, as first illustrated in Bora et al. (2017), instead of relying on hand-crafted sparsity, the prior structure of the unknown signal is learned through a generative model, such as VAE (Kingma & Welling, 2013) or GAN (Goodfellow et al., 2014). Notably, the past few years have witnessed a significant success of one new family of probabilistic generative models called diffusion models, in particular score matching with Langevin daynamics (SMLD) (Song & Ermon, 2019; 2020) and denoising diffusion probabilistic modeling (DDPM) (Ho et al., 2020; Nichol & Dhariwal, 2021), which have proven extremely effective and even outperform the state-of-the-art (SOTA) GAN (Goodfellow et al., 2014) and VAE (Kingma & Welling, 2013) in density estimation and generation of various natural sources such as images and audios. As both DDPM and SMLD estimate, implicitly or explicitly, the *score* (i.e., gradient of the log probability density with respect to data), they are also referred to together as score-based generative models (SGM) (Song et al., 2020). Several CS methods with SGM have been proposed very recently (Jalal et al., 2021a;b; Kawar et al., 2021; 2022; Chung et al., 2022) and perform quite well in recovering $\mathbf{x}$ with only a few linear measurements in (1).

Nevertheless, the linear model (1) ideally assumes that the measurements have infinite precision, which is not the case in realistic acquisition scenarios. In practice, the obtained measurements have to be quantized to a finite number of Q bits before transmission and/or storage (Zymnis et al., 2009; Dai & Milenkovic, 2011). Quantization leads to information loss which makes the recovery particularly challenging. For moderate and high quantization resolutions, i.e., Q is large, the quantization impact is usually modeled as a mere additive Gaussian noise whose variance is determined by quantization distortion (Dai & Milenkovic, 2011; Jacques et al., 2010). Subsequently, most CS algorithms originally designed for the linear model (1) can then be applied with some modifications. However, such an approach is apparently suboptimal since the information about the quantizer is not utilized to a full extent (Dai & Milenkovic, 2011). This is true especially in the case of coarse quantization, i.e., Q is small. An extreme and important case of coarse quantization is *one-bit* quantization, where $Q = 1$ and only the *sign* of the measurements are observed (Boufounos & Baraniuk, 2008). Apart from extremely low cost in storage, one-bit quantization is particularly appealing in hardware implementations and has also proven robust to both nonlinear distortions and dynamic range issues (Boufounos & Baraniuk, 2008). Consequently, there have been extensive studies on quantized CS, particularly 1-bit CS, in the past decades and a variety of algorithms have been proposed, e.g., (Zymnis et al., 2009; Dai & Milenkovic, 2011; Plan & Vershynin, 2012; 2013; Jacques et al., 2013; Xu & Kabashima, 2013; Xu et al., 2014; Awasthi et al., 2016; Meng et al., 2018; Jung et al., 2021; Liu et al., 2020; Liu & Liu, 2022). However, most existing methods are based on standard CS methods, which, therefore, inevitably inherit their inability to capture rich structures of natural signals beyond sparsity. While several recent works (Liu et al., 2020; Liu & Liu, 2022) studied one-bit CS using generative priors, their main focuses are VAE and/or GAN, rather than SGM.

1.1 Contributions

- We propose a novel framework, termed as Quantized Compressed Sensing with Score-based Generative Models (QCS-SGM in short), to recover unknown signals from noisy *quantized* measurements. QCS-SGM is applicable to arbitrary number of quantization bits, including the extremely coarse one-bit measurements. To the best of our knowledge, this is the first time that SGM has been utilized for quantized CS.
- We consider one popular SGM model called NCSNv2 (Song & Ermon, 2020) and train it on the MNIST, Cifar-10, and CelebA datasets, respectively. Using the pre-trained SGM as a generative prior, the proposed QCS-SGM significantly outperforms all existing SOTA algorithms by a large margin under quantized measurements. Perhaps surprisingly, even in the extreme case of one-bit measurements, QCS-SGM can still faithfully recover the original images with much fewer measurements than the original signal dimension.
- As one kind of posterior sampling method, QCS-SGM can be readily used to obtain multiple samples with different random initializations. As a result, with QCS-SGM one can easily obtain confidence intervals or uncertainty estimates of the reconstructed results.

- We propose an annealed likelihood score called *noise-perturbed likelihood score* as a principled way to incorporate information from the measurements, which is one key to the success of QCS-SGM. Interestingly, when degenerated to the linear case (1), if $\mathbf{A}\mathbf{A}^T$ is a diagonal matrix, QCS-SGM reduces to a form similar to Jalal et al. (2021a). While the annealing term (denoted as γ_t^2 in (4) of Jalal et al. (2021a) is added heuristically as one additional hyper-parameter, it appears naturally within our framework and admits an analytical solution. More importantly, for general matrices $\mathbf{A}$, our method significantly outperforms Jalal et al. (2021a). It is expected that the idea of noise-perturbed likelihood score can be generalized to other conditional generative models as a general rule.

1.2 Related works

The proposed QCS-SGM is one kind of data-driven methods for CS which generally include two lines of research: unsupervised and supervised. The supervised approach trains neural networks from end to end using pairs of original signals and observations (Jin et al., 2017; Aggarwal et al., 2018; Wu et al., 2019; Yao et al., 2019; Yang et al., 2020; Antun et al., 2020). Despite good performance on specific problems, the supervised approach is severely limited and requires re-training even with a slight change of the measurement model. The unsupervised approach, including the proposed QCS-SGM, only learns a prior and thus avoids such problem-specific training since the measurement model is only known and used for inference (Bora et al., 2017; Hand & Joshi, 2019; Asim et al., 2020; Pan et al., 2021). Among various others, the CSGM framwork proposed in Jin et al. (2017) is the most popular one which learns the prior with generative models.

Recent works Liu et al. (2020); Liu & Liu (2022) extended CSGM to non-linear observations including one-bit CS, which are the two most related studies to current work. However, the main focuses of Liu et al. (2020); Liu & Liu (2022) are limited to VAE and GAN (in particular DCGAN (Radford et al., 2015)). As is widely known, GAN suffers from unstable training and less diversity due to the adversarial training nature while VAE uses a surrogate loss. Another significant drawback of priors with VAE/GAN is that they can have large representation errors or biases due to architecture and training, which can be easily caused by inappropriate latent dimensionality and/or mode collapse (Asim et al., 2020). With the advent of SGM such as SMLD (Song & Ermon, 2019; 2020) and DDPM (Ho et al., 2020; Nichol & Dhariwal, 2021), there have been several recent studies (Jalal et al., 2021a;b; Kawar et al., 2021; 2022; Chung et al., 2022) for CS which learn the prior using SGM and outperform conventional methods. Nevertheless, existing works on SGM for CS all focus on linear measurements (1). Another related work (Qiu et al., 2020) studied one-bit CS where the sparsity is implicitly enforced via mapping a low dimensional representation through a known ReLU generative network. Wei et al. (2019) also considers a non-linear recovery using a generative model assuming that the original signal is generated from a low-dimensional latent vector.

2 Background

2.1 Inverse Problems from Quantized Measurements

The inverse problem from noisy quantized measurements is posed as follows (Zymnis et al., 2009)

$$\mathbf{y} = \mathsf{Q}(\mathbf{A}\mathbf{x} + \mathbf{n}), \tag{2}$$

where the goal is to recover the unknown signal $\mathbf{x} \in \mathbb{R}^{N\times 1}$ from quantized measurements $\mathbf{y} \in \mathbb{R}^{M\times 1}$, where $\mathbf{A} \in \mathbb{R}^{M\times N}$ is a known linear mixing matrix, $\mathbf{n} \sim \mathcal{N}(\mathbf{n}; 0, \sigma^2\mathbf{I})$ is an i.i.d. additive Gaussian noise with known variance, and $\mathsf{Q}(\cdot) : \mathbb{R}^{M\times 1} \to \mathcal{Q}^{M\times 1}$ is an *element-wise* [1] quantizer function which maps each element into a finite (or countable) set of codewords $\mathcal{Q}$, i.e., $y_m = \mathsf{Q}(z_m + n_m) \in \mathcal{Q}$, or equivalently $(z_m + n_m) \in \mathsf{Q}^{-1}(y_m), m = 1, 2, ..., M$, where z_m is the m-th element of $\mathbf{z} = \mathbf{A}\mathbf{x}$.

Usually the quantization codewords $\mathcal{Q}$ correspond to intervals (Zymnis et al., 2009), i.e., $\mathsf{Q}^{-1}(y_m) = [l_{y_m}, u_{y_m})$ where l_{y_m}, u_{y_m} denote prescribed lower and upper thresholds associated with the quantized measurement y_m, i.e., $l_{y_m} \leq (z_m + n_m) < u_{y_m}$. For example, for a uniform quantizer with Q

[1] While more sophisticated quantization schemes exist (Dirksen, 2019), here we focus on memoryless scalar quantization due to its popularity and simplicity where the quantizer acts element-wise.

quantization bits (resolution), the quantization codewords $\mathcal{Q} = \{q_r\}_{r=1}^{2^Q}$ consist of 2^Q elements, i.e.,

$$q_r = \left(2r - 2^Q - 1\right)\Delta/2, \quad r = 1, ..., 2^Q, \tag{3}$$

where $\Delta > 0$ is the quantization interval, and the lower and upper thresholds associated q_r are

$$l_{q_r} = \begin{cases} -\infty, r = 1; \\ \left(r - 2^{Q-1} - 1\right)\Delta, \ r = 2, ..., 2^Q. \end{cases} \quad u_{q_r} = \begin{cases} \left(r - 2^{Q-1}\right)\Delta, \ r = 1, ..., 2^Q; \\ +\infty, \ , r = 2^Q. \end{cases} \tag{4}$$

In the extreme one-bit case, i.e., $Q = 1$, only the sign values are observed, i.e.,

$$\mathbf{y} = \text{sign}(\mathbf{A}\mathbf{x} + \mathbf{n}), \tag{5}$$

where the quantization codewords $\mathcal{Q} = \{-1, +1\}$ and the associated thresholds are $l_{-1} = -\infty, u_{-1} = 0$ and $l_{+1} = 0, u_{+1} = +\infty$, respectively.

2.2 Score-based Generative Models

For any continuously differentiable probability density function $p(\mathbf{x})$, if we have access to its score function, i.e., $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, we can iteratively sample from it using Langevin dynamics (Turq et al., 1977; Bussi & Parrinello, 2007; Welling & Teh, 2011)

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \alpha_t \nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{x}_{t-1}) + \sqrt{2\alpha_t}\mathbf{z}_t, \ 1 \leq t \leq T, \tag{6}$$

where $\mathbf{z_t} \sim \mathcal{N}(\mathbf{z_t}; \mathbf{0}, \mathbf{I})$, $\alpha_t > 0$ is the step size, and T is the total number of iterations. It has been demonstrated that when α_t is sufficiently small and T is sufficiently large, the distribution of $\mathbf{x}_T$ will converge to $p(\mathbf{x})$ (Roberts & Tweedie, 1996; Welling & Teh, 2011). In practice, the score function $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ is unknown and can be estimated using a *score network* $\mathrm{s}_{\boldsymbol{\theta}}(\mathbf{x})$ via score matching (Hyvärinen, 2006; Vincent, 2011). However, the vanilla Langevin dynamics faces a variety of challenges such as slow convergence. To address this challenge, inspired from simulated annealing (Kirkpatrick et al., 1983; Neal, 2001), Song & Ermon (2019) proposed an annealed version of Langevin dynamics, which perturbs the data with Gaussian noise of different scales and jointly estimates the score functions of noise-perturbed data distributions. Accordingly, during the inference, an annealed Langevin dynamics (ALD) is performed to leverage the information from all noise scales.

Specifically, assume that $p_\beta(\tilde{\mathbf{x}} \mid \mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \beta^2\mathbf{I})$, and so we have $p_\beta(\tilde{\mathbf{x}}) = \int p_{\text{data}}(\mathbf{x}) p_\beta(\tilde{\mathbf{x}} \mid \mathbf{x}) d\mathbf{x}$, where $p_{\text{data}}(\mathbf{x})$ is the data distribution. Consider we have a sequence of noise scales $\{\beta_t\}_{t=1}^T$ satisfying $\beta_{\max} = \beta_1 > \beta_2 > \cdots > \beta_T = \beta_{\min} > 0$. The $\beta_{\min}$ is small enough so that $p_{\beta_{\min}}(\tilde{\mathbf{x}}) \approx p_{\text{data}}(\mathbf{x})$, and $\beta_{\max}$ is large enough so that $p_{\beta_{\max}}(\tilde{\mathbf{x}}) \approx \mathcal{N}(\mathbf{x}; 0, \beta_{\max}^2 I)$. The noise conditional score network (NCSN) $\mathrm{s}_{\boldsymbol{\theta}}(\mathbf{x}, \beta)$ proposed in Song & Ermon (2019) aims to estimate the score function of each $p_{\beta_t}(\tilde{\mathbf{x}})$ by optimizing the following weighted sum of score matching objective

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathbb{E}_{p_{\beta_t}(\tilde{\mathbf{x}}|\mathbf{x})} \left[\|\mathrm{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, \beta_t) - \nabla_{\tilde{\mathbf{x}}} \log p_{\beta_t}(\tilde{\mathbf{x}} \mid \mathbf{x})\|_2^2 \right]. \tag{7}$$

After training the NCSN, for each noise scale, we can run K steps of Langevin MCMC to obtain a sample for each $p_{\beta_t}(\tilde{\mathbf{x}})$ as

$$\mathbf{x}_t^k = \mathbf{x}_t^{k-1} + \alpha_t \mathrm{s}_{\boldsymbol{\theta}}(\mathbf{x}_t^{k-1}, \beta_t) + \sqrt{2\alpha_t}\mathbf{z}_t^k, \ k = 1, 2, ..., K. \tag{8}$$

The sampling process is repeated for $t = 1, 2, ..., T$ sequentially with $\mathbf{x}_1^0 \sim \mathcal{N}(\mathbf{x}; \mathbf{0}, \beta_{\max}^2\mathbf{I})$ and $\mathbf{x}_{t+1}^0 = \mathbf{x}_t^K$ when $t < T$. As shown in Song & Ermon (2019), when $K \to \infty$ and $\alpha_t \to 0$ for all t, the final sample $\mathbf{x}_T^K$ will become an exact sample from $p_{\beta_{\min}}(\tilde{\mathbf{x}}) \approx p_{\text{data}}(\mathbf{x})$ under some regularity conditions. Later, by a theoretical analysis of learning and sampling process of NCSN, an improved version of NCSN, termed as NCSNv2, was proposed in Song & Ermon (2020) which is more stable and can scale to various datasets with high resolutions.

3 Quantized CS with SGM

In the case of quantized measurements (2) with access to observations $\mathbf{y}$, the goal is to sample from the posterior distribution $p(\mathbf{x} \mid \mathbf{y})$ rather than $p(\mathbf{x})$. To this end, the sampling process in (6) using the Langevin dynamics can be modified as follows

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \alpha_t \nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{x}_{t-1} \mid \mathbf{y}) + \sqrt{2\alpha_t}\mathbf{z}_t, \ 1 \leq t \leq T, \tag{9}$$

where the conditional (*posterior*) score $\nabla_{\mathbf{x}} \log p(\mathbf{x} \mid \mathbf{y})$ is required. One direct solution is to specially train a score network work that matches the score $\nabla_{\mathbf{x}} \log p(\mathbf{x} \mid \mathbf{y})$. However, this method is rather inflexible and needs to re-train the network when confronted with different measurements. Instead, inspired from conditional diffusion models (Jalal et al., 2021a), we resort to compute the posterior score using the Bayesian rule from which $\nabla_{\mathbf{x}} \log p(\mathbf{x} \mid \mathbf{y})$ is decomposed into two terms

$$\nabla_{\mathbf{x}} \log p(\mathbf{x} \mid \mathbf{y}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x}), \tag{10}$$

which include the unconditional score $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ (we call it *prior score*), and the conditional score $\nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x})$ (we call it *likelihood score*), respectively. The prior score $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ can be easily obtained using a trained score network such as NCSN or NCSNv2, so the remaining goal is to compute the likelihood score $\nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x})$ from the quantized measurements.

3.1 Noise-perturbed likelihood score

In the case without noise perturbing in $\mathbf{x}$, one can calculate the likelihood score $\nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x})$ directly from the measurement process (2) and then substitute it into (10) to obtain the posterior score $\nabla_{\mathbf{x}} \log p(\mathbf{x} \mid \mathbf{y})$. However, this does not work well and might even diverge if the prior score $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ is estimated using noise perturbation such as the case in NCSN (Song & Ermon, 2019) or NCSNv2 (Song & Ermon, 2020), which is the most popular and efficient score-based method. The reason is that in NCSN/NCSNv2, as shown in (7), the estimated score is not the original prior score $\nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})$ but rather a noise-perturbed prior score $\nabla_{\tilde{\mathbf{x}}} \log p_{\beta_t}(\tilde{\mathbf{x}})$ for each noise scale β_t. As a result, the likelihood score directly computed from (2) does not match the noise-perturbed prior score. To address this problem, we propose to calculate the *noise-perturbed likelihood score* by expliciitly taking into consideration the noise perturbation in NCSN/NCSNv2, as shown in Theorem 1.

Theorem 1. *(noise-perturbed likelihood score) For each noise scale $\beta_t > 0$, if $p_{\beta_t}(\tilde{\mathbf{x}} \mid \mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \beta_t^2 \mathbf{I})$, suppose that $\mathbf{A}\mathbf{A}^T$ is a diagonal matrix, then the noise-perturbed likelihood score $\nabla_{\tilde{\mathbf{x}}} \log p_{\beta_t}(\mathbf{y} \mid \tilde{\mathbf{x}})$ for the quantized measurements $\mathbf{y}$ in (2) can be computed as*

$$\nabla_{\tilde{\mathbf{x}}} \log p_{\beta_t}(\mathbf{y} \mid \tilde{\mathbf{x}}) = \mathbf{A}^T \mathbf{G}(\beta_t, \mathbf{y}, \mathbf{A}, \tilde{\mathbf{x}}), \tag{11}$$

where $\mathbf{G}(\beta_t, \mathbf{y}, \mathbf{A}, \tilde{\mathbf{x}}) = [g_1, g_2, ..., g_M]^T \in \mathbb{R}^{M \times 1}$ with each element being

$$g_m = \frac{\exp\left(-\frac{\tilde{u}_{y_m}^2}{2}\right) - \exp\left(-\frac{\tilde{l}_{y_m}^2}{2}\right)}{\sqrt{\sigma^2 + \beta_t^2 \left\|\mathbf{a}_m\right\|_2^2} \int_{\tilde{l}_{y_m}}^{\tilde{u}_{y_m}} \exp\left(-\frac{t^2}{2}\right) dt}, \ m = 1, 2, ..., M, \tag{12}$$

where

$$\tilde{u}_{y_m} = \frac{\mathbf{a}_m^T \tilde{\mathbf{x}} - u_{y_m}}{\sqrt{\sigma^2 + \beta_t^2 \left\|\mathbf{a}_m\right\|_2^2}}, \ \tilde{l}_{y_m} = \frac{\mathbf{a}_m^T \tilde{\mathbf{x}} - l_{y_m}}{\sqrt{\sigma^2 + \beta_t^2 \left\|\mathbf{a}_m\right\|_2^2}} \tag{13}$$

Proof. The proof is shown in Appendix A. □

Remark. *Note that Theorem 1 only holds exactly under an assumption that $\mathbf{A}\mathbf{A}^T$ is a diagonal matrix since otherwise one cannot obtain a closed-form solution. Nevertheless, for a variety of typical applications in CS with handcrafted sensing matrices $\mathbf{A}$, this assumption will mostly hold, exactly or approximately, e.g., i.i.d. Gaussian matrices, discrete cosine transform (DCT) matrices. As demonstrated in Section 4, remarkable reconstruction performances can still be achieved using results in Theorem 1 even when it holds approximately. We leave the case for general measurement matrices $\mathbf{A}$ for quantized CS as an important future work.*

In the extreme one-bit case, a more concise result can be obtained, as shown in Corollary 1.1:

Corollary 1.1. *(noise-perturbed likelihood score in the one-bit case) For each noise scale $\beta_t > 0$, if $p_{\beta_t}(\tilde{\mathbf{x}} \mid \mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \beta_t^2 \mathbf{I})$, suppose that $\mathbf{A}\mathbf{A}^T$ is a diagonal matrix, then the noise-perturbed likelihood score $\nabla_{\tilde{\mathbf{x}}} \log p_{\beta_t}(\mathbf{y} \mid \tilde{\mathbf{x}})$ for the sign measurements $\mathbf{y} = \operatorname{sign}(\mathbf{A}\mathbf{x} + \mathbf{n})$ in (5) can be computed as*

$$\nabla_{\tilde{\mathbf{x}}} \log p_{\beta_t}(\mathbf{y} \mid \tilde{\mathbf{x}}) = \mathbf{A}^T \mathbf{G}(\beta_t, \mathbf{y}, \mathbf{A}, \tilde{\mathbf{x}}), \tag{14}$$

where $\mathbf{G}(\beta_t, \mathbf{y}, \mathbf{A}, \tilde{\mathbf{x}}) = [g_1, g_2, ..., g_M]^T \in \mathbb{R}^{M\times 1}$ *with each element being*

$$g_m = \left[\frac{1+y_m}{2\Phi(\tilde{z}_m)} - \frac{1-y_m}{2\left(1-\Phi(\tilde{z}_m)\right)}\right] \frac{\exp\left(-\frac{\tilde{z}_m^2}{2}\right)}{\sqrt{2\pi(\sigma^2+\beta_t^2\left\|\mathbf{a}_m\right\|_2^2)}},\ m=1,2,...,M, \tag{15}$$

where $\tilde{z}_m = \frac{\mathbf{a}_m^T\tilde{\mathbf{x}}}{\sqrt{\sigma^2+\beta_t^2\|\mathbf{a}_m\|_2^2}}$ *and* $\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{t^2}{2}}dt$ *is the cumulative distribution function of the standard normal distribution.*

Proof. It can be easily proved following the proof in Theorem 1. □

Interestingly, if no quantization is used, or equivalently $Q \to \infty$, the model in (2) reduces to the standard linear model (1). Accordingly, results in Theorem 1 can be modified as in Corollary 1.2:

Corollary 1.2. *(noise-perturbed likelihood score without quantization) For each noise scale* $\beta_t > 0$*, if* $p_{\beta_t}(\tilde{\mathbf{x}} \mid \mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \beta_t^2\mathbf{I})$*, then the noise-perturbed likelihood score* $\nabla_{\tilde{\mathbf{x}}}\log p_{\beta_t}(\mathbf{y} \mid \tilde{\mathbf{x}})$ *for linear measurements* $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$ *in (1) can be computed as*

$$\nabla_{\tilde{\mathbf{x}}}\log p_{\beta_t}(\mathbf{y} \mid \tilde{\mathbf{x}}) = \mathbf{A}^T\left(\sigma^2\mathbf{I} + \beta_t^2\mathbf{A}\mathbf{A}^T\right)^{-1}\left(\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}\right). \tag{16}$$

If $\mathbf{A}\mathbf{A}^T$ *is a diagonal matrix, it can be further simplified as*

$$\nabla_{\tilde{\mathbf{x}}}\log p_{\beta_t}(\mathbf{y} \mid \tilde{\mathbf{x}}) = \frac{\mathbf{A}^T\left(\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}\right)}{\sigma^2 + \beta_t^2\left\|\mathbf{a}_m\right\|_2^2}. \tag{17}$$

Proof. The proof is shown in Appendix B. □

Remark. *In contrast to the quantized case, here we obtain an exact closed-form solution (16) for any kind of matrices* $\mathbf{A}$*. Interestingly, when* $\mathbf{A}\mathbf{A}^T$ *becomes a diagonal matrix, it reduces to (17), which is similar to Jalal et al. (2021a) without quantization where the annealing term* $\beta_t^2\left\|\mathbf{a}_m\right\|_2^2$ *in (17) corresponds to* γ_t^2 *in Jalal et al. (2021a). However,* γ_t^2 *in Jalal et al. (2021a) is simply added heuristically as an additional hyper-parameter while we derive it in closed-form under the principled perturbed likelihood score. More importantly, for general matrices* $\mathbf{A}$*, we obtain an exact form (16) which significantly outperforms Jalal et al. (2021a), as demonstrated in Appendix C. Therefore, we not only explains the necessity of an annealing term in Jalal et al. (2021a) but also extends and improves Jalal et al. (2021a) in the general case.*

3.2 Posterior sampling via annealed Langevin dynamics

By combining the prior score from SGM and noise-perturbed likelihood score in Theorem 1, we can then perform posterior sampling via annealed Langevin dynamics (9). The resultant algorithm is shown in Algorithm 1 and we call it Quantized CS with SGM (QCS-SGM in short).

Algorithm 1: Quantized Compressed Sensing with SGM (QCS-SGM)

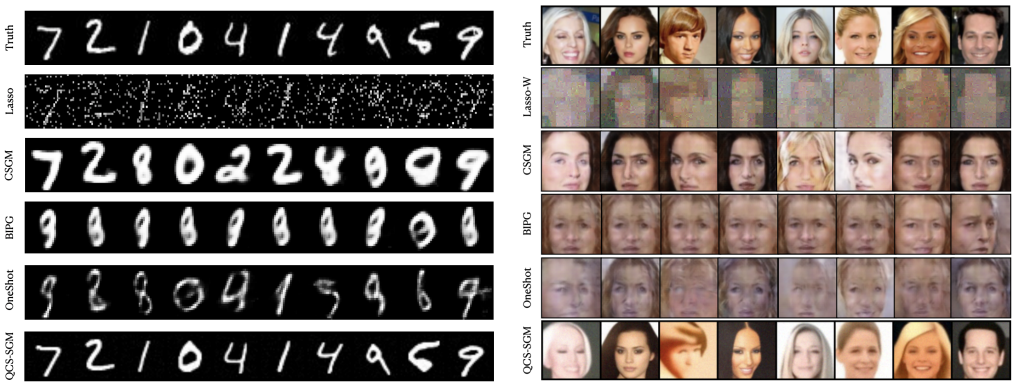
Input: $\{\beta_t\}_{t=1}^T, \epsilon, K, \mathbf{y}, \mathbf{A}, \sigma^2$, quantization codewords $\mathcal{Q}$ and thresholds $\{[l_q, u_q) | q \in \mathcal{Q}\}$
Initialization: $\mathbf{x}_1^0 \sim \mathcal{U}(0,1)$
for $t = 1$ ***to*** T **do**
  $\alpha_t \leftarrow \epsilon\beta_t^2/\beta_T^2$
  for $k = 1$ ***to*** K **do**
    Draw $\mathbf{z}_t^k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    Compute $\mathbf{G}(\beta_t, \mathbf{y}, \mathbf{A}, \mathbf{x}_t^{k-1})$ as (26) (or (15) for one-bit)
    $\mathbf{x}_t^k = \mathbf{x}_t^{k-1} + \alpha_t\left[\mathrm{s}_{\boldsymbol{\theta}}(\mathbf{x}_t^{k-1}, \beta_t) + \mathbf{A}^T\mathbf{G}(\beta_t, \mathbf{y}, \mathbf{A}, \mathbf{x}_t^{k-1})\right] + \sqrt{2\alpha_t}\mathbf{z}_t^k$
  $\mathbf{x}_{t+1}^0 \leftarrow \mathbf{x}_t^K$
Output: $\hat{\mathbf{x}} = \mathbf{x}_T^K$
.

4 EXPERIMENTS

In this section, we empirically demonstrate the efficacy of the proposed QCS-SGM in various scenarios. The widely-used i.i.d. Gaussian matrix $\mathbf{A}$, i.e., $A_{ij} \sim \mathcal{N}(0, 1/M)$ are considered. *Our code will be open-sourced after acceptance*.

Datasets: Three popular datasets are considered: MNIST (LeCun & Cortes, 2010) , Cifar-10 (Krizhevsky & Hinton, 2009), and CelebA (Liu et al., 2015). MNIST (LeCun & Cortes, 2010) is a standard database of handwritten digits which contains 60,000 training grayscale images and 10,000 testing grayscale images, each of size 28×28 pixels, so that the input dimension for MNIST is $N = 28 \times 28 = 784$ per image. Cifar-10 dataset (Krizhevsky & Hinton, 2009) consists of natural RGB images of size 32×32 pixels , resulting in $N = 32 \times 32 \times 3 = 3072$ inputs per image. Each image is classified into 1 of 10 classes, such as dog, cat, automobile, etc. The training set contains 50,000 images, while the test set contains 10,000 images. The CelebA dataset (Liu et al., 2015) consists of 202,599 face images of various celebrities. We cropped each face image to a 64×64 RGB image, resulting in $N = 64 \times 64 \times 3 = 12288$ inputs per image. To evaluate the out-of-distribution (OOD) performance, the Flickr Faces High Quality (FFHQ) Dataset (Karras et al., 2018) is also considered as an OOD dataset, whose preprocessing is the same CelebA.

QCS-SGM: Regarding the SGM models, we adopt the popular NCSNv2 (Song & Ermon, 2020) in all cases. Specifically, for MNIST, Cifar-10, and CelebA, we train NCSNv2 (Song & Ermon, 2020) models on the corresponding training datasets, respectively, using the same training setup as Song & Ermon (2020). As a result, the prior score can be estimated using these pre-trained NCSNv2 models. After observing the quantized measurements as (2), we can infer the original $\mathbf{x}$ by posterior sampling via QCS-SGM in Algorithm 1. It is important to note that, in evaluating the performances of QCS-SGM, **we select images $\mathbf{x}$ from the test sets** so that they are unseen by the pre-trained SGM models. For details of the experimental setting, please refer to Appendix D.



(a) MNIST, $M = 200, \sigma = 0.05$ (b) CelebA, $M = 4000, \sigma = 0.001$

Figure 1: Typical reconstructed images from one-bit measurements on MNIST and CelebA. QCS-SGM faithfully recovers original images from one-bit measurements even when $M \ll N$.

4.1 ONE-BIT QUANTIZATION

First, we perform experiments on the extreme case, i.e., one-bit measurements. Specifically, we consider images of the MNIST (LeCun & Cortes, 2010) and CelebA datasets (Liu et al., 2015) in the same setting as Liu & Liu (2022). Apart from QCS-SGM, we also show results of LASSO (Tibshirani, 1996) (for CelebA, Lasso in the wavelet domain using 2D Daubechies-1 Wavelet Transform, denoted as Lasso-W, is used), CSGM (Bora et al., 2017), BIPG (Liu et al., 2020), and OneShot (Liu & Liu, 2022). For Lasso (Lasso-W), CSGM, BIPG, OneShot, we follow the setting as the open-sourced code of Liu & Liu (2022).

The typical reconstructed images from one-bit measurements with fixed $M \ll N$ are shown in Figure 1 for both MNIST and CelebA. Perhaps surprisingly, QCS-SGM can faithfully recover the images from one-bit measurements with $M \ll N$ measurements while other methods fail or only recover

quite vague images. To quantitatively evaluate the effect of M, we compare different algorithms using three popular metrics, i.e., peak signal-to-noise ratio (PSNR), cosine similarity, and structural similarity index (SSIM) for different M, and the results are shown in Figure 2. It can be seen that, under all three metrics, the proposed QCS-SGM significantly outperforms all the other algorithms by a large margin. Please refer to Appendix F for more results.

Multiple samples & **Uncertainty Estimates**: It is worth pointing out that, as one kind of sampling method, QCS-SGM can yield multiple samples with different random initialization so that we can easily obtain confidence intervals or uncertainty estimates of the reconstructed results. Please refer to Appendix E for more typical samples of QCS-SGM. In contrast, OneShot (Liu & Liu, 2022) can easily get stuck in a local minimum with different random initializations (Liu & Liu, 2022).

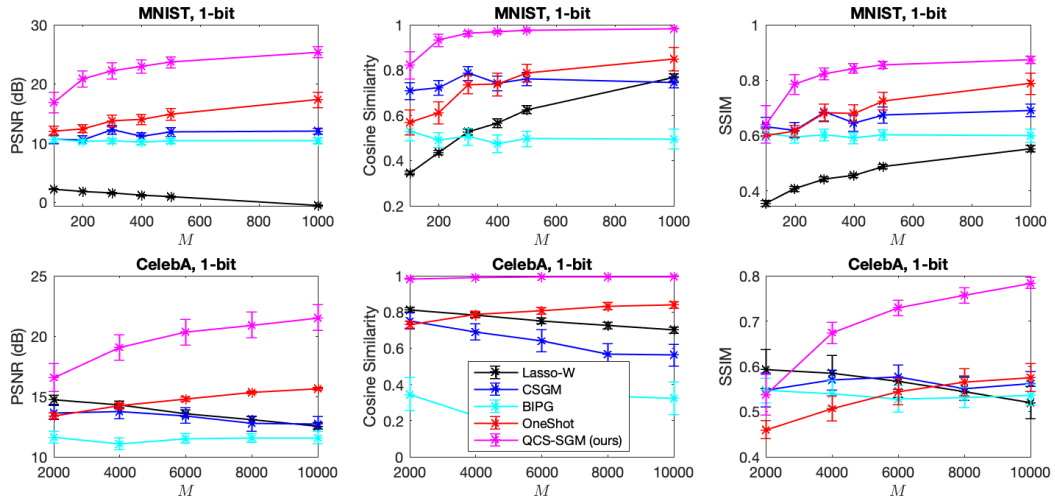
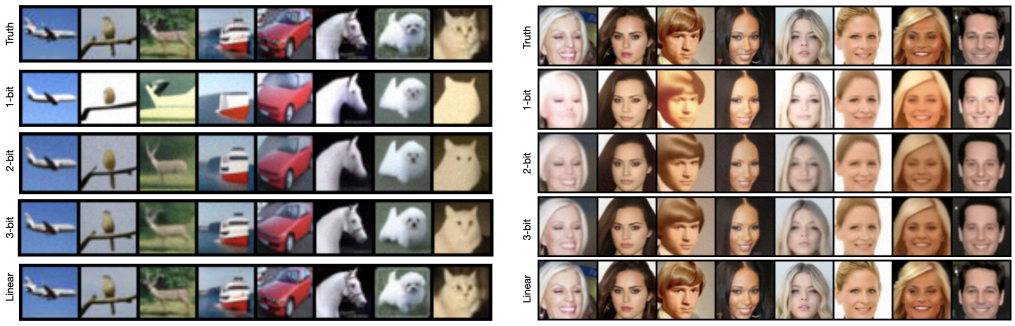


Figure 2: Quantitative comparisons based on different metrics for one-bit MNIST and CelebA. QCS-SGM remarkably outperforms all other methods under all metrics.

4.2 Multi-bit Quantization

Then, we evaluate the efficacy of QCS-SGM in the case of multi-bit quantization, e.g., 2-bit, 3-bit. The results on Cifar-10 and CelebA are shown in Figure 3 and Figure 4. The results of linear case without quantization are also shown for comparison using (17) in Corollary 1.2. As expected, as the increase of quantization resolution, the reconstruction performances get better. For more results, please refer to Appendix F.



(a) Cifar-10, $M = 2000$, $\sigma = 0.001$ (b) CelebA, $M = 4000$ $\sigma = 0.001$

Figure 3: Results of QCS-SGM for Cifar-10 and CelebA images under different quantization bits.

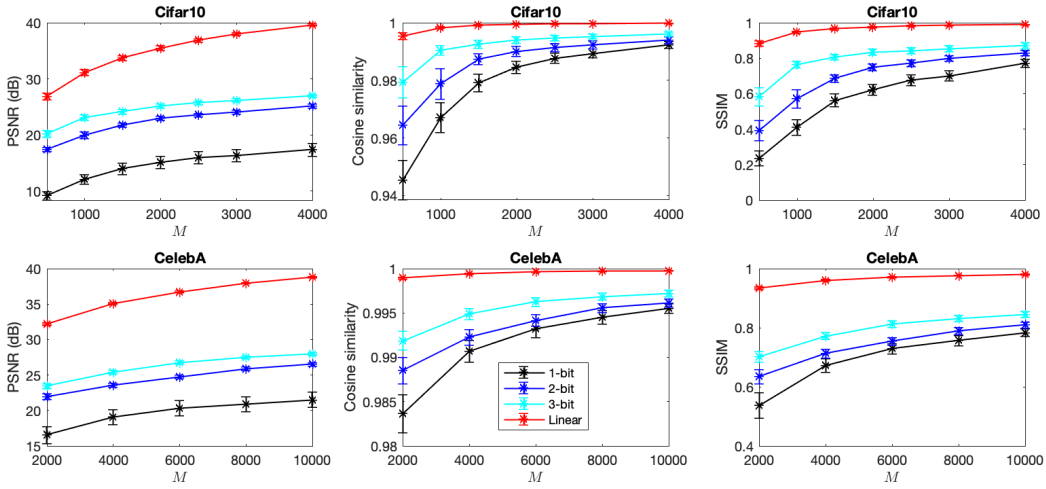


Figure 4: Quantitative results of QCS-SGM for different quantization bits, $\sigma = 0.001$.

4.3 Out-of-Distribution Performance

In practice, the signals of interests may be out-of-distribution (OOD) because the available training dataset has bias and is unrepresentative of the true underlying distribution. To assess the performance of QCS-SGM for OOD datasets, we choose the CelebA as the training set while evaluate results on OOD images randomly sampled from the FFHQ dataset, which have features of variation that are rare among CelebA images (e.g., skin tone, age, beards, and glasses) (Jalal et al., 2021b). As shown in Figure 5, even on OOD images, the proposed QCS-SGM can still obtain significantly competitive results compared with other existing algorithms.

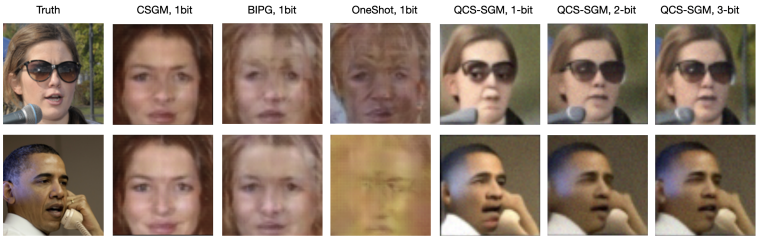


Figure 5: Results on out-of-distribution FFHQ dataset from one-bit (2-bit and 3-bit are also shown for QCS-SGM) measurements. $M = 10000, \sigma = 0.001$.

5 Conclusion

In this paper, a novel framework called Quantized Compressed Sensing with Score-based Generative Models (QCS-SGM) is proposed to reconstruct a high-dimensional signal from noisy quantized measurements. Thanks to the power of SGM in capturing rich structure of natural signals beyond simple sparsity, the proposed QCS-SGM significantly outperforms existing SOTA algorithms by a large margin under coarsely quantized measurements on a variety of experiments, for both in-distribution and out-of-distribution datasets. There are still some limitations in QCS-SGM. For example, it is limited to the case where measurement matrix $\mathbf{A}$ (approximately) satisfies that $\mathbf{A}\mathbf{A}^T$ is a diagonal matrix. Generalization to general matrices is an important future work. Generative models are an active area of research with ongoing rapid improvements and we believe that SGM could open up new opportunities in quantized CS in the future, and hope that our work as a first step could inspire more studies on this fascinating topic.

REFERENCES

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.

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.

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*, pp. 399–409. PMLR, 2020.

Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Conference on Learning Theory*, pp. 152–192. PMLR, 2016.

Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pp. 537–546. PMLR, 2017.

Petros T Boufounos and Richard G Baraniuk. 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pp. 16–21. IEEE, 2008.

Giovanni Bussi and Michele Parrinello. Accurate sampling using langevin dynamics. *Physical Review E*, 75(5):056707, 2007.

Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

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.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *arXiv preprint arXiv:2206.00941*, 2022.

Wei Dai and Olgica Milenkovic. Information theoretical and algorithmic approaches to quantized compressive sensing. *IEEE transactions on communications*, 59(7):1857–1866, 2011.

Sjoerd Dirksen. Quantized compressed sensing: a survey. In *Compressed Sensing and Its Applications*, pp. 67–95. Springer, 2019.

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

David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

Marco F Duarte and Yonina C Eldar. Structured compressed sensing: From theory to applications. *IEEE Transactions on signal processing*, 59(9):4053–4085, 2011.

Maryam Fazel, E Candes, Benjamin Recht, and P Parrilo. Compressed sensing and robust recovery of low rank matrices. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pp. 1043–1047. IEEE, 2008.

Rina Foygel and Lester Mackey. Corrupted sensing: Novel guarantees for separating structured signals. *IEEE Transactions on Information Theory*, 60(2):1223–1247, 2014.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Paul Hand and Babhru Joshi. Global guarantees for blind demodulation with generative priors. *Advances in Neural Information Processing Systems*, 32, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Aapo Hyvärinen. Consistency of pseudolikelihood estimation of fully visible boltzmann machines. *Neural Computation*, 18(10):2283–2292, 2006.

Laurent Jacques, David K Hammond, and Jalal M Fadili. Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine. *IEEE Transactions on Information Theory*, 57(1):559–571, 2010.

Laurent Jacques, Jason N Laska, Petros T Boufounos, and Richard G Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE transactions on information theory*, 59(4):2082–2102, 2013.

Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. Robust compressed sensing mri with deep generative priors. *Advances in Neural Information Processing Systems*, 34:14938–14954, 2021a.

Ajil Jalal, Sushrut Karmalkar, Alex Dimakis, and Eric Price. Instance-optimal compressed sensing via posterior sampling. In *International Conference on Machine Learning*, pp. 4709–4720. PMLR, 2021b.

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.

Hans Christian Jung, Johannes Maly, Lars Palzer, and Alexander Stollenwerk. Quantized compressed sensing by rectified linear units. *IEEE transactions on information theory*, 67(6):4125–4149, 2021.

Yoshiyuki Kabashima. A cdma multiuser detection algorithm on the basis of belief propagation. *Journal of Physics A: Mathematical and General*, 36(43):11111, 2003.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022.

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

Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Jiulong Liu and Zhaoqiang Liu. Non-iterative recovery from nonlinear observations using generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 233–243, 2022.

Zhaoqiang Liu, Selwyn Gomes, Avtansh Tiwari, and Jonathan Scarlett. Sample complexity bounds for 1-bit compressive sensing and binary stable embeddings with generative priors. In *International Conference on Machine Learning*, pp. 6216–6225. PMLR, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

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.

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

Xiangming Meng, Sheng Wu, and Jiang Zhu. A unified bayesian inference framework for generalized linear models. *IEEE Signal Processing Letters*, 25(3):398–402, 2018.

Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2012.

Yaniv Plan and Roman Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

Shuang Qiu, Xiaohan Wei, and Zhuoran Yang. Robust one-bit recovery via relu generative networks: Near-optimal statistical rate and global landscape analysis. In *International Conference on Machine Learning*, pp. 7857–7866. PMLR, 2020.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pp. 341–363, 1996.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

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.

Jie Tang, Brian E Nett, and Guang-Hong Chen. Performance comparison between total variation (tv)-based compressed sensing and statistical iterative reconstruction algorithms. *Physics in Medicine & Biology*, 54(19):5781, 2009.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Joel A Tropp and Stephen J Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, 2010.

Pierre Turq, Frédéric Lantelme, and Harold L Friedman. Brownian dynamics: Its application to ionic solutions. *The Journal of Chemical Physics*, 66(7):3039–3044, 1977.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

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*, pp. 6697–6706. PMLR, 2019.

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)*, pp. 681–688. Citeseer, 2011.

Yan Wu, Mihaela Rosca, and Timothy Lillicrap. Deep compressed sensing. In *International Conference on Machine Learning*, pp. 6850–6860. PMLR, 2019.

Yingying Xu and Yoshiyuki Kabashima. Statistical mechanics approach to 1-bit compressed sensing. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(02):P02041, feb 2013. doi: 10.1088/1742-5468/2013/02/p02041. URL https://doi.org/10.1088/1742-5468/2013/02/p02041.

Yingying Xu, Yoshiyuki Kabashima, and Lenka Zdeborová. Bayesian signal reconstruction for 1-bit compressed sensing. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(11):P11015, nov 2014. doi: 10.1088/1742-5468/2014/11/p11015. URL https://doi.org/10.1088/1742-5468/2014/11/p11015.

Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Admm-csnet: A deep learning approach for image compressive sensing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3): 521–538, 2020. doi: 10.1109/TPAMI.2018.2883941.

Hantao Yao, Feng Dai, Shiliang Zhang, Yongdong Zhang, Qi Tian, and Changsheng Xu. Dr2-net: Deep residual reconstruction network for image compressive sensing. *Neurocomputing*, 359: 483–493, 2019.

Argyrios Zymnis, Stephen Boyd, and Emmanuel Candes. Compressed sensing with quantized measurements. *IEEE Signal Processing Letters*, 17(2):149–152, 2009.

A PROOF OF THEOREM 1

Proof. Let us denote $\mathbf{z} = \mathbf{A}\mathbf{x}$. For each noise scale $\beta_t > 0$, if $p_{\beta_t}(\tilde{\mathbf{x}} \mid \mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \beta_t^2 \mathbf{I})$, then we can equivalently write

$$\tilde{\mathbf{x}} = \mathbf{x} + \beta_t \mathbf{w}, \tag{18}$$

where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. As a result, $\mathbf{z} = \mathbf{A}\mathbf{x} = \mathbf{A}(\tilde{\mathbf{x}} - \beta_t \mathbf{w}) = \mathbf{A}\tilde{\mathbf{x}} - \beta_{\mathbf{t}} \mathbf{A}\mathbf{w}$. Then, from (2), we obtain

$$\mathbf{y} = \mathsf{Q}\left(\mathbf{A}\tilde{\mathbf{x}} + \tilde{\mathbf{n}}\right), \tag{19}$$

where $\tilde{\mathbf{n}} = \mathbf{n} - \beta_t \mathbf{A}\mathbf{w}$. Since $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and are independent to each other, it can be concluded that $\tilde{\mathbf{n}}$ is also Gaussian with mean zero and covariance $\sigma^2 \mathbf{I} + \beta_t^2 \mathbf{A}\mathbf{A}^T$, i.e., $\tilde{\mathbf{n}} \sim \mathcal{N}(\tilde{\mathbf{n}}; \mathbf{0}, \sigma^2 \mathbf{I} + \beta_t^2 \mathbf{A}\mathbf{A}^T)$.

If $\mathbf{A}\mathbf{A}^T$ is a diagonal matrix, each element $\tilde{n}_m \sim \mathcal{N}(\tilde{n}_m; 0, \sigma^2 + \beta_t^2 \left\|\mathbf{a}_m\right\|_2^2)$ of $\tilde{\mathbf{n}}$ will be independent to each other and thus from (28) we can obtain a closed-form solution for the likelihood distribution $p(\mathbf{y}|\hat{\mathbf{z}} = \mathbf{A}\tilde{\mathbf{x}})$ as follows

$$p(\mathbf{y}|\hat{\mathbf{z}} = \mathbf{A}\tilde{\mathbf{x}}) = \prod_{m=1}^{M} p\left(y_m \mid \hat{z}_m = \mathbf{a}_m^T \tilde{\mathbf{x}}\right) \tag{20}$$

where, from the definition of quantizer Q,

$$p\left(y_m \mid \hat{z}_m = \mathbf{a}_m^T \tilde{\mathbf{x}}\right) = p\left(l_{y_m} \leq \hat{z}_m + \tilde{n}_m < u_{y_m}\right) \tag{21}$$

$$= \Phi\left(\frac{-\hat{z}_m + u_{y_m}}{\sqrt{\sigma^2 + \beta_t^2 \left\|\mathbf{a}_m\right\|_2^2}}\right) - \Phi\left(\frac{-\hat{z}_m + l_{y_m}}{\sqrt{\sigma^2 + \beta_t^2 \left\|\mathbf{a}_m\right\|_2^2}}\right) \tag{22}$$

$$= \Phi\left(-\tilde{u}_{y_m}\right) - \Phi\left(-\tilde{l}_{y_m}\right) \tag{23}$$

where $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$ is the cumulative distribution function of the standard normal distribution and

$$\tilde{u}_{y_m} = \frac{\mathbf{a}_m^T \tilde{\mathbf{x}} - u_{y_m}}{\sqrt{\sigma^2 + \beta_t^2 \left\|\mathbf{a}_m\right\|_2^2}}, \; \tilde{l}_{y_m} = \frac{\mathbf{a}_m^T \tilde{\mathbf{x}} - l_{y_m}}{\sqrt{\sigma^2 + \beta_t^2 \left\|\mathbf{a}_m\right\|_2^2}}. \tag{24}$$

As a result, it can be calculated that the noise-perturbed likelihood score $\nabla_{\tilde{\mathbf{x}}} \log p_{\beta_t}(\mathbf{y} \mid \tilde{\mathbf{x}})$ for the quantized measurements $\mathbf{y}$ in (2) can be computed as

$$\nabla_{\tilde{\mathbf{x}}} \log p_{\beta_t}(\mathbf{y} \mid \tilde{\mathbf{x}}) = \mathbf{A}^T \mathbf{G}(\beta_t, \mathbf{y}, \mathbf{A}, \tilde{\mathbf{x}}) \tag{25}$$

where $\mathbf{G}(\beta_t, \mathbf{y}, \mathbf{A}, \tilde{\mathbf{x}}) = [g_1, g_2, ..., g_M]^T \in \mathbb{R}^{M \times 1}$ with each element being

$$g_m = \frac{\exp\left(-\frac{\tilde{u}_{y_m}^2}{2}\right) - \exp\left(-\frac{\tilde{l}_{y_m}^2}{2}\right)}{\sqrt{\sigma^2 + \beta_t^2 \left\|\mathbf{a}_m\right\|_2^2} \int_{\tilde{l}_{y_m}}^{\tilde{u}_{y_m}} \exp\left(-\frac{t^2}{2}\right) dt}, \; m = 1, 2, ..., M, \tag{26}$$

□

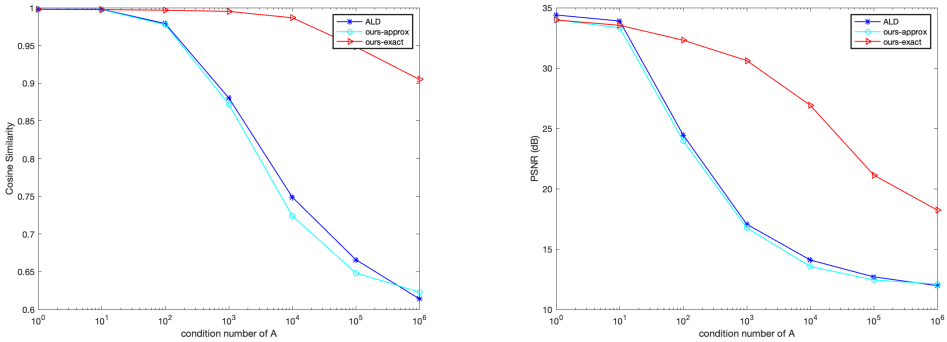
B PROOF OF COROLLARY 1.2

Proof. Similarly in the proof of Theorem 1, let us denote $\mathbf{z} = \mathbf{A}\mathbf{x}$. For each noise scale $\beta_t > 0$, if $p_{\beta_t}(\tilde{\mathbf{x}} \mid \mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \beta_t^2 \mathbf{I})$, then we can equivalently write

$$\tilde{\mathbf{x}} = \mathbf{x} + \beta_t \mathbf{w}, \tag{27}$$

where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. As a result, $\mathbf{z} = \mathbf{A}\mathbf{x} = \mathbf{A}(\tilde{\mathbf{x}} - \beta_t \mathbf{w}) = \mathbf{A}\tilde{\mathbf{x}} - \beta_{\mathbf{t}} \mathbf{A}\mathbf{w}$. Then, in the case of linear model, from (1), we obtain

$$\mathbf{y} = \mathbf{A}\tilde{\mathbf{x}} + \tilde{\mathbf{n}}, \tag{28}$$



(a) Cosine Similarity (b) PSNR

Figure 6: Averaged Cosine Similarity and PSNR of reconstructed MNIST images of ALD in Jalal et al. (2021a) and ours (with formula (17) and (16), respectively) when $M = 200, \sigma = 0.1$ for different condition number of matrix $\mathbf{A}$ is $\text{cond}(\mathbf{A}) = 1000$. It can be seen that our method with exact (16) significantly outperforms ALD in Jalal et al. (2021a) at high condition number while performs similarly at low condition number. Ours with diagonal approximation (17) is about the same as ALD as expected.

where $\tilde{\mathbf{n}} = \mathbf{n} - \beta_t \mathbf{A}\mathbf{w}$. Since $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and are independent to each other, it can be concluded that $\tilde{\mathbf{n}}$ is also Gaussian with mean zero and covariance $\sigma^2\mathbf{I} + \beta_t^2\mathbf{A}\mathbf{A}^T$, i.e., $\tilde{\mathbf{n}} \sim \mathcal{N}(\tilde{\mathbf{n}}; \mathbf{0}, \sigma^2\mathbf{I} + \beta_t^2\mathbf{A}\mathbf{A}^T)$. Therefore, a closed-form solution for the likelihood distribution $p(\mathbf{y}|\hat{\mathbf{z}} = \mathbf{A}\tilde{\mathbf{x}})$ can be obtained as follows

$$p(\mathbf{y}|\hat{\mathbf{z}} = \mathbf{A}\tilde{\mathbf{x}}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}})^T \left(\sigma^2\mathbf{I} + \beta_t^2\mathbf{A}\mathbf{A}^T\right)^{-1} (\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}})\right)}{\sqrt{(2\pi)^M \det\left(\sigma^2\mathbf{I} + \beta_t^2\mathbf{A}\mathbf{A}^T\right)}}. \tag{29}$$

As a result, we can readily obtain a closed-form solution for the noise-perturbed likelihood score $\nabla_{\tilde{\mathbf{x}}} \log p_{\beta_t}(\mathbf{y} \mid \tilde{\mathbf{x}})$ as follows

$$\nabla_{\tilde{\mathbf{x}}} \log p_{\beta_t}(\mathbf{y} \mid \tilde{\mathbf{x}}) = \mathbf{A}^T\left(\sigma^2\mathbf{I} + \beta_t^2\mathbf{A}\mathbf{A}^T\right)^{-1} (\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}}). \tag{30}$$

Furthermore, if $\mathbf{A}\mathbf{A}^T$ is a diagonal matrix, it can be simplified as

$$\nabla_{\tilde{\mathbf{x}}} \log p_{\beta_t}(\mathbf{y} \mid \tilde{\mathbf{x}}) = \frac{\mathbf{A}^T(\mathbf{y} - \mathbf{A}\tilde{\mathbf{x}})}{\sigma^2 + \beta_t^2 \|\mathbf{a}_m\|_2^2}, \tag{31}$$

which completes the proof. □

C Comparison with ALD in Jalal et al. (2021a) in the linear case

As shown in Corollary 1.2, in the special case without quantization, our results in Thoerem 1 can be reduced to a form similar to the ALD in Jalal et al. (2021a). However, there are several different important differences. First, our results are derived from a principled way as noise-perturbed likelihood score and admit closed-form solutions, while the results in Jalal et al. (2021a) are obtained heuristically by adding an additional hyper-parameter (and thus needs fine-tuning). Second, the results in Jalal et al. (2021a) are similar to an approximate version (17) of ours, which holds only when $\mathbf{A}\mathbf{A}^T$ is a diagonal matrix. For general matrices $\mathbf{A}$, we have a exact form (16). We compare our results with ALD (Jalal et al., 2021a) and the results are shown in Figure 6 and Figure 7. It can be seen that when at low condition number of $\mathbf{A}$ when $\mathbf{A}\mathbf{A}^T$ is approximately a diagonal matrix, our with diagonal approximation (17) performs similarly as exact one (16), both of which are similar to ALD (Jalal et al., 2021a), as expected. However, when the condition number of $\mathbf{A}$ is large so that $\mathbf{A}\mathbf{A}^T$ is far from a diagonal matrix, our method with the exact form (16) significantly outperforms ALD in Jalal et al. (2021a).

(a) Truth (b) ALD (Jalal et al., 2021a) (c) Ours (approx (17)) (d) Ours (exact (16))

Figure 7: Tyical recovered MNIST images of ALD in Jalal et al. (2021a) and ours (with formula (17) and (16), respectively) when $M = 200, \sigma = 0.05$ and the condition number of matrix $\mathbf{A}$ is $\text{cond}(\mathbf{A}) = 1000$. It can be seen that our method with exact (16) significantly outperforms ALD in Jalal et al. (2021a), which performs about the same as ours with diagonal approximation (17).

D Detailed Experimental Settings

In training NCSNv2 for MNIST, Cifar-10, and CelebA, we used exactly same training setup as Song & Ermon (2020), please refer to Song & Ermon (2020) and its open-sourced code for details of traning NCSNv2 models. When performing posterior sampling using the QCS-SGM in 1, for simplicity, we set a constant value $\epsilon = 0.0002$ for all quantized measurements (e.g., 1-bit, 2-bit, 3-bit) and $\epsilon = 0.00002$ for all linear measurements in all the experiments in this paper. It is believed that some improvement can be achieved with further fine-tuning of ϵ for different scenarios. For MNIST and Cifar-10, we set $\beta_1 = 50, \beta_T = 0.01, T = 232$, while for CelebA, we set $\beta_1 = 90, \beta_T = 0.01, T = 500$, which are the same as Song & Ermon (2020). The number of steps K in QCS-SGM for each noise scale is set to be $K = 5$ in all experiments.

E Multiple Samples and Uncertainty Estimates

As one kind of posterior sampling method, QCS-SGM can yield multiple samples with different random initialization so that we can easily obtain confidence intervals or uncertainty estimates of the reconstructed results. For example, typical samples as well as mean and std are shown in Figure 8 for MNIST and CelebA in the one-bit case.

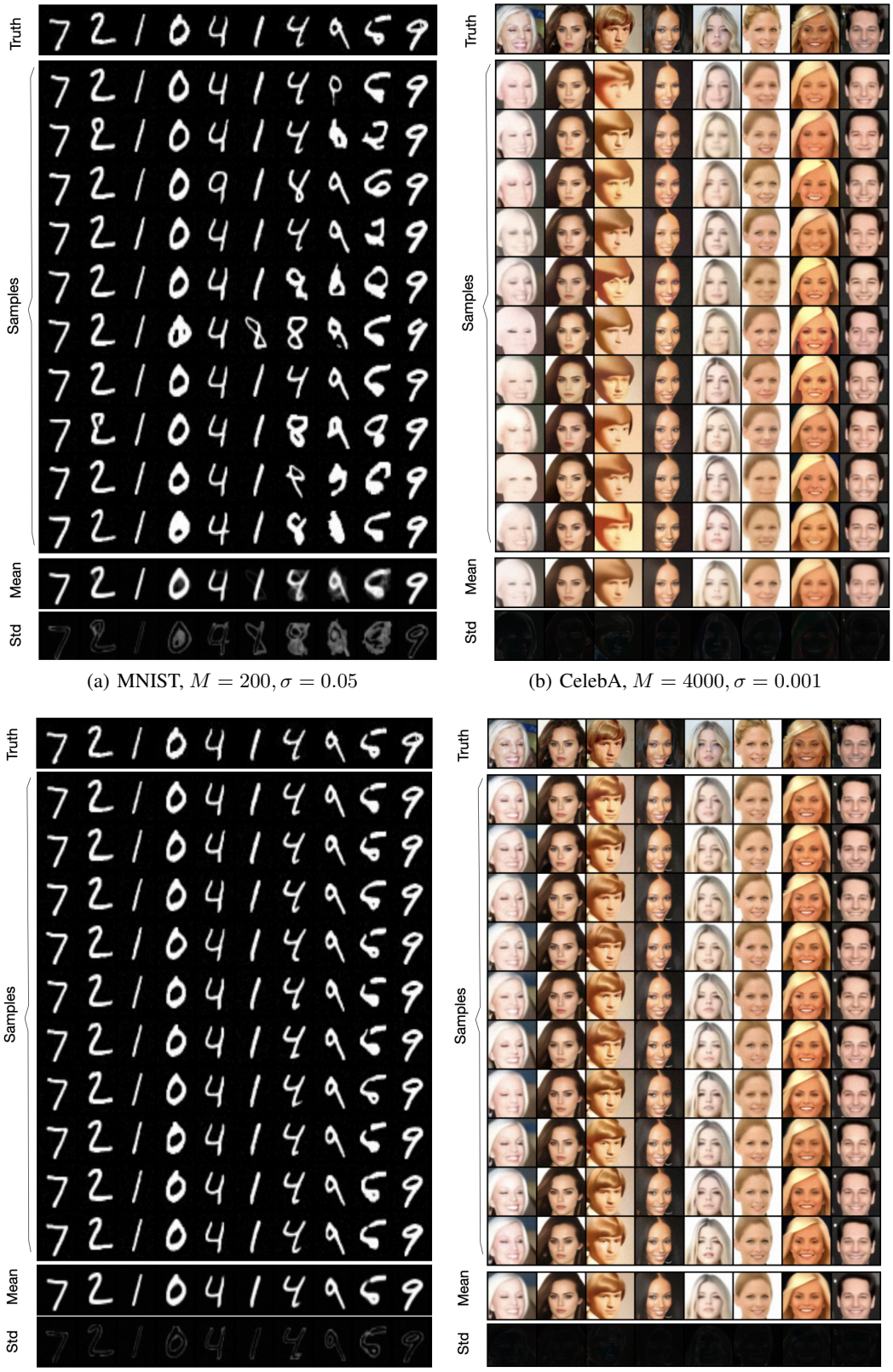
F Additional Results

Some additional results are shown in this section.

Figure 9 and Figure 10 show results with relatively large value of M in the same setting as Figure 1 and Figure 3, respectively.

F.1 Comparison in the exactly same setting as Liu & Liu (2022)

Note that there is a slight difference in the modeling of measurement matrix $\mathbf{A}$ and noise $\mathbf{n}$ between ours and that in Liu & Liu (2022). In Liu & Liu (2022), it is assumed that $\mathbf{A}$ is an i.i.d. Gaussian matrix where each element A_{ij} follows $A_{ij} \sim \mathcal{N}(0, 1)$ and that $\mathbf{n}$ is an i.i.d. Gaussian with variance σ^2, i.e., $n_i \sim \mathcal{N}(0, \sigma^2)$. However, in practice, the measurement matrix $\mathbf{A}$ is usually normalized so that the norm of each column equals 1. As a result, in our setting, we assume that each element of i.i.d. Gaussian matrix follows $A_{ij} \sim \mathcal{N}(0, 1/M)$. Mathematically, there is a one-to-one correspondence between the two settings, but the simulation setting is different due to the different measurement size M. As a result, for an exact comparison with results in Liu & Liu (2022), we also conducted experiments assuming the exactly the same setting as Liu & Liu (2022), i.e., $A_{ij} \sim \mathcal{N}(0, 1)$, and $n_i \sim \mathcal{N}(0, \sigma^2)$. The results are shown in Figures 11, 12, 13, 14.



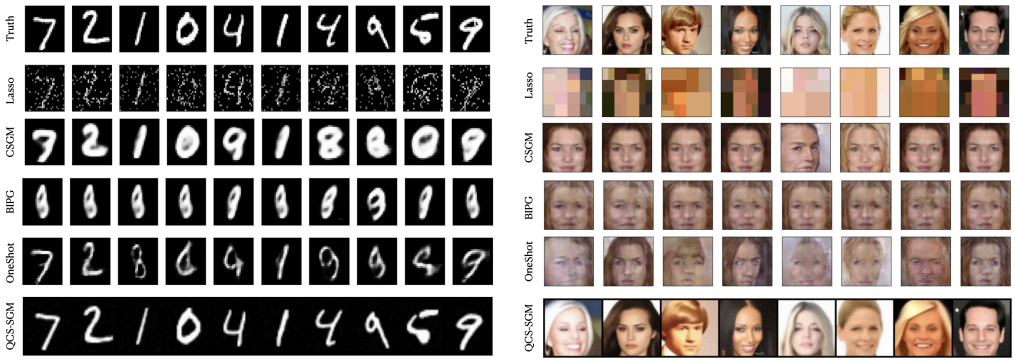
(a) MNIST, $M = 200, \sigma = 0.05$

(b) CelebA, $M = 4000, \sigma = 0.001$

(c) MNIST, $M = 400, \sigma = 0.05$

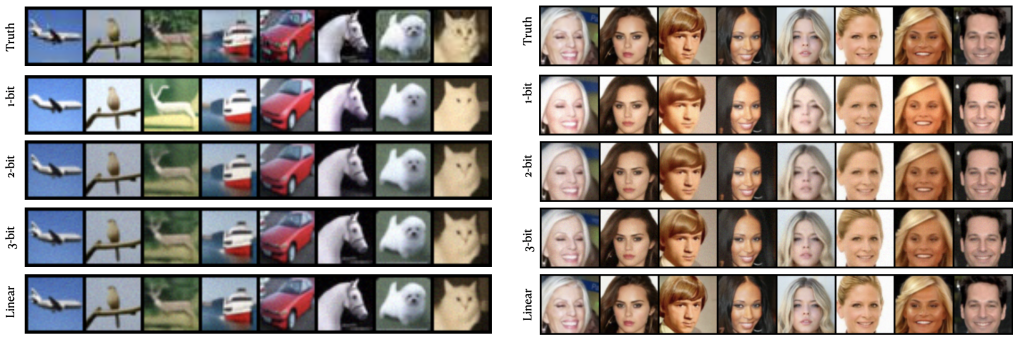
(d) CelebA, $M = 10000, \sigma = 0.001$

Figure 8: Multiple samples of QCS-SGM on MNIST ($M = 200, 400, \sigma = 0.05$) and CelebA datasets ($M = 4000, 1000, \sigma = 0.001$) from one-bit measurements. The mean and std of the samples are also shown.



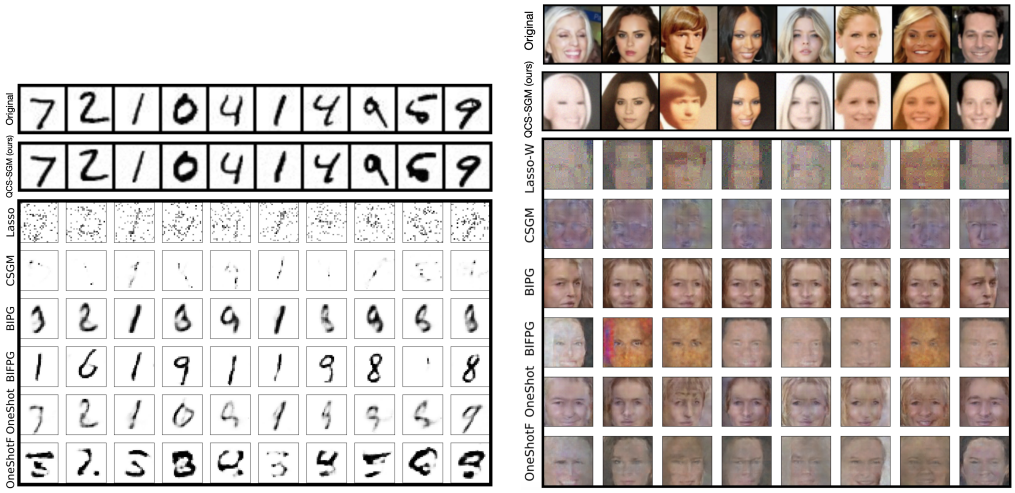
(a) MNIST, $M = 400, \sigma = 0.05$ (b) CelebA, $M = 10000, \sigma = 0.001$

Figure 9: Typical reconstructed images from one-bit measurements on MNIST ($M = 400$) and CelebA ($M = 10000$).



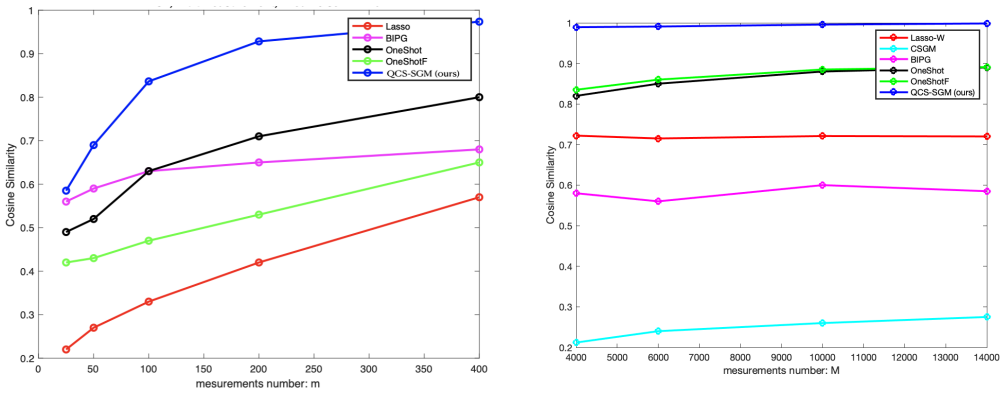
(a) Cifar-10, $M = 4000$, $\sigma = 0.001$ (b) CelebA, $M = 16000$ $\sigma = 0.001$

Figure 10: Results of QCS-SGM for Cifar-10 ($M = 4000$) and CelebA ($M = 16000$) images under different quantization bits.



(a) MNIST, $M = 200, \sigma = 1$ (b) CelebA, $M = 4000, \sigma = 0.01$

Figure 11: Results of reconstructed images from one-bit measurements on MNIST and CelebA images, following exactly same setting as Liu & Liu (2022), i.e., $A_{ij} \sim \mathcal{N}(0, 1)$, and $n_i \sim \mathcal{N}(0, \sigma^2)$.

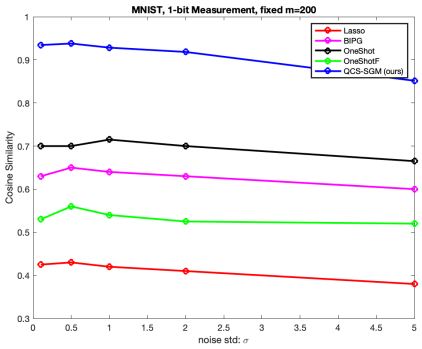


(a) MNIST, $\sigma = 1$ (b) CelebA, $\sigma = 0.01$

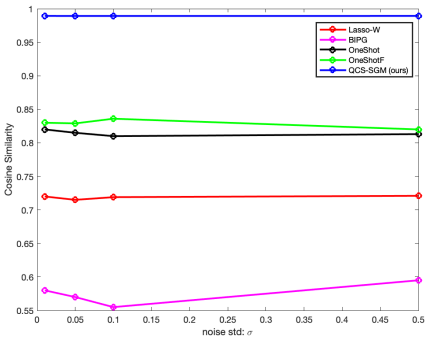
Figure 12: Quantitative comparisons based on cosine similarity for one-bit MNIST and CelebA images, following exactly same setting as Liu & Liu (2022), i.e., $A_{ij} \sim \mathcal{N}(0, 1)$, and $n_i \sim \mathcal{N}(0, \sigma^2)$.

(a) MNIST, $M = 400, \sigma = 1$ (b) CelebA, $M = 10000, \sigma = 0.01$

Figure 13: Results of reconstructed images from one-bit measurements on MNIST and CelebA images, following exactly same setting as Liu & Liu (2022), i.e., $A_{ij} \sim \mathcal{N}(0, 1)$, and $n_i \sim \mathcal{N}(0, \sigma^2)$.



(a) MNIST, $M = 200$



(b) CelebA, $M = 4000$

Figure 14: Results of reconstructed images from one-bit measurements on MNIST and CelebA images for different σ, following exactly same setting as Liu & Liu (2022), i.e., $A_{ij} \sim \mathcal{N}(0, 1)$, and $n_i \sim \mathcal{N}(0, \sigma^2)$.