INVERSE IMAGING WITH GENERATIVE PRIORS VIA LANGEVIN DYNAMICS

Thanh V. Nguyen, Gauri Jagatap, Chinmay Hegde

ABSTRACT

Deep generative models have emerged as a powerful class of priors for signals in various inverse problems such as compressed sensing, phase retrieval and super-resolution. Here, we assume an unknown signal to lie in the range of some pre-trained generative model. A popular approach for signal recovery is via gradient descent in the low-dimensional latent space. While gradient descent has achieved good empirical performance, its theoretical behavior is not well understood. In this paper, we introduce the use of stochastic gradient Langevin dynamics (SGLD) for compressed sensing with a generative prior. Under mild assumptions on the generative model, we prove the convergence of SGLD to the true signal. We also demonstrate competitive empirical performance to standard gradient descent.

1. INTRODUCTION

We consider the familiar setting of inverse problems where the goal is to recover an n-dimensional signal (or image) x^* that is observed via a linear measurement operation $y = Ax^*$. The measurement vector can be noisy, and its dimension m may be less than n. Several imaging applications fit this setting, including super-resolution [1], in-painting, denoising [2], and compressed sensing [3].

Since such an inverse problem is ill-posed in general, the recovery of x^* from y often requires assuming a low-dimensional structure or *prior* on x^*. Choices of good priors have been extensively explored in the past three decades, including sparsity [4], structured sparsity [5], end-to-end training via convolutional neural networks [3], pre-trained generative priors [6], as well as untrained deep image priors [7, 8].

In this paper, we focus on a powerful class of priors based on deep generative models. The setup is the following: the unknown signal x^* is assumed to lie in the range of some pre-trained generator network, obtained from (say) a generative adversarial network (GAN) or a variational autoencoder (VAE). That is, $x^* = G(z^*)$ for some z^* in the latent space. The task is again to recover x^* from (noisy) linear measurements.

Such generative priors have been shown to achieve high empirical success [3, 6]. However, progress on the theoretical side for inverse problems with generative priors has been much more modest. On the one hand, the seminal work of [9] established the first *statistical* upper bounds (in terms of measurement complexity) for compressed sensing for fairly general generative priors, which was later shown in [10] to be nearly optimal. On the other hand, provable *algorithmic guarantees* for recovery using generative priors are only available in very restrictive cases. The paper [**?**] proves the convergence of (a variant of) gradient descent for shallow generative priors whose weights obey a distributional assumption. The paper [11] proves the convergence of projected gradient descent (PGD) under the assumption that the range of the (possibly deep) generative model G admits a polynomial-time oracle projection. To our knowledge, the most general algorithmic result in this line of work is by [12]. Here, the authors show that under rather mild and intuitive assumptions on G, a linearized alternating direction method of multipliers (ADMM) applied to a regularized mean-squared error loss converges to a (potentially large) neighborhood of x^*.

A barrier for obtaining guarantees for recovery algorithms based on gradient descent is the *non-convexity* of the recovery problem induced by the generator network. Therefore, in this paper we sidestep traditional gradient descent-style optimization methods, and instead show that a good estimate of x^* can be obtained by performing stochastic gradient Langevin Dynamics (SGLD) [**?**, 13–15]. We show that this dynamics amounts to *sampling* from a Gibbs distribution whose energy function is precisely the reconstruction loss [1].

As a stochastic version of gradient descent, SGLD is simple to implement. However, care must be taken in constructing the additive stochastic perturbation to each gradient update step. Nevertheless, the sampling viewpoint enables us to achieve finite-time convergence guarantees for compressed sensing recovery. To the best of our knowledge, this is the first theoretical result for solving inverse problems with generative neural priors using Langevin gradients. Moreover, our analysis succeeds under (slightly) weaker assumptions on the generator network than those made in [12].

Our specific contributions are as follows:

1. We propose a provable image recovery algorithm for generative priors based on stochastic gradient Langevin dynamics (SGLD).
2. We prove polynomial-time convergence of our proposed recovery algorithm to the true underlying solution, under assumptions of smoothness and near-isometry of G. These are technically weaker than the mild assumptions made in [12]. We emphasize that these conditions are valid for a wide range of generator networks. Section 3 describes them in greater details.
3. We provide several empirical results and demonstrate that our approach is competitive with existing (heuristic) methods based on gradient descent.

2. PRIOR WORK

We briefly review the literature on solving inverse imaging problems with deep generative models. For a more thorough survey on deep learning for inverse problems, see [17].

In [6], the authors provide sufficient conditions under which the solution of the inverse problem is a minimizer of the (possibly non-convex) program:

$$\min_{x=G(z)} \|Ax - y\|_2^2 \,. \tag{2.1}$$

Email: thanhng.cs@gmail.com, gbj221@nyu.edu, chinmay.h@nyu.edu. This work was partially done while the authors were with the Electrical and Computer Engineering Department at Iowa State University. CH is currently at the Tandon School of Engineering at New York University. This work was supported in part by NSF grants CCF-2005804 and CCF-1815101.

[1] While preparing this manuscript, we became aware of concurrent work by [16] which also pursues a similar Langevin-style approach for solving compressed sensing problems; however, they do not theoretically analyze its dynamics.

Specifically, they show that if A satisfies the so-called set-Restricted Eigenvalue Condition (REC), then the solution to (2.1) equals the unknown vector x^*. They also show that if the generator G has a latent dimension k and is L-Lipschitz, then a matrix $A \in \mathbb{R}^{m\times n}$ populated with i.i.d. Gaussian entries satisfies the REC, provided $m = O(k \log L)$. However, they propose gradient descent as a heuristic to solve (2.1), but do not analyze its convergence. In [11], the authors show that projected gradient descent (PGD) for (2.1) converges at a linear rate under the REC, but only if there exists a tractable *projection* oracle that can compute $\arg\min_z \|x - G(z)\|$ for any x. The recent work [18] provides sufficient conditions under which such a projection can be approximately computed. In [12], a provable recovery scheme based on ADMM is established, but guarantees recovery only up to a neighborhood around x^*.

Note that all the above works assume mild conditions on the weights of the generator, use variations of gradient descent to update the estimate for x, and require the forward matrix A to satisfy the REC over the range of G. [**?**] showed *global* convergence for gradient descent, but under the (strong) assumption that the weights of the trained generator are Gaussian distributed.

Generator networks trained with GANs are most commonly studied. However, more recently, [19] have advocated using *invertible* generative models, which use real-valued non-volume preserving (NVP) transformations [20]. An alternate strategy for sampling images consistent with linear forward models was proposed in [21] where the authors assume an invertible generative mapping and sample the latent vector z from a second generative invertible prior.

Our proposed approach also traces its roots to Bayesian sparse modeling [22], where instead of modeling the problem as estimating a (deterministic) sparse vector, one models the signal x to be sampled from a sparsity promoting distribution, such as a Laplace prior. One can then derive the maximum *a posteriori* (MAP) estimate of x under the constraint that the measurements $y = Ax$ are consistent. Our motivation is similar, except that we model the distribution of x as being supported on the range of a generative prior.

3. RECOVERY VIA LANGEVIN DYNAMICS

In the rest of the paper, $x \wedge y$ denotes $\min\{x, y\}$ and $x \vee y$ for $\max\{x, y\}$. Given a distribution μ and set $\mathcal{A}$, we denote $\mu(\mathcal{A})$ the probability measure of $\mathcal{A}$ with respect to μ. $\|\mu - \nu\|_{TV}$ is the total variation distance between two distributions μ and ν. Finally, we use standard big-O notation in our analysis.

3.1. Preliminaries

We focus on the problem of recovering a signal $x^* \in \mathbb{R}^n$ from a set of linear measurements $y \in \mathbb{R}^m$ where

$$y = Ax^* + \varepsilon.$$

To keep our analysis and results simple, we consider zero measurement noise, i.e., $\varepsilon = 0$[2]. Here, $A \in \mathbb{R}^{m\times n}$ is a matrix populated with i.i.d. Gaussian entries with mean 0 and variance $1/m$. We assume that x^* belongs to the range of a known generative model $G : \mathcal{D} \subset \mathbb{R}^d \to \mathbb{R}^n$; that is,

$$x^* = G(z^*) \text{ for some } z^* \in \mathcal{D}.$$

Following [9], we restrict z to belong to a d-dimensional Euclidean ball, i.e., $\mathcal{D} = \mathcal{B}(0, R)$. Then, given the measurements y, our goal is to recover x^*. Again following [9], we do so by solving the usual optimization problem:

$$\min_{z\in\mathcal{D}} F(z) \triangleq \|y - AG(z)\|^2. \tag{3.1}$$

Hereon and otherwise stated, $\|\cdot\|$ denotes the ℓ_2-norm. The most popular approach to solving (3.1) is to use gradient descent [9]. For generative models $G(z)$ defined by deep neural networks, the function $F(z)$ is highly non-convex, and as such, it is impossible to guarantee global signal recovery using regular (projected) gradient descent.

We adopt a slightly more nuanced approach. Starting from an initial point $z_0 \sim \mu_0$, our algorithm computes stochastic gradient updates of the form:

$$z_{k+1} = z_k - \eta\nabla_z F(z) + \sqrt{2\eta\beta^{-1}}\xi_k, \quad k = 0, 1, 2, \dots \tag{3.2}$$

where ξ_k is a unit Gaussian random vector in $\mathbb{R}^d$, η is the step size and β is an inverse temperature parameter. This update rule is known as *stochastic gradient Langevin dynamics* (SGLD) [13] and has been widely studied both in theory and practice [14, 15]. Intuitively, (3.2) is an Euler discretization of the continuous-time diffusion equation:

$$\mathrm{d}Z(t) = -\nabla_z F(Z(t))\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B(t), \quad t \geq 0, \tag{3.3}$$

where $Z(0) \sim \mu_0$. Under standard regularity conditions on $F(z)$, one can show that the above diffusion has a unique invariant Gibbs measure.

We refine the standard SGLD to account for the boundedness of z. Specifically, we require an additional Metropolis-like accept/reject step to ensure that z_{k+1} always belongs to the support $\mathcal{D}$, and also is not too far from z_k of the previous iteration. We study this variant for theoretical analysis; in practice we have found that this is not necessary. Algorithm 1 (CS-SGLD) shows the detailed algorithm. Note that we can use stochastic (mini-batch) gradient instead of the full gradient $\nabla_z F(z)$.

Let us derive sufficient conditions on the convergence (in distribution) of the random process in Algorithm 1 to the target distribution π, denoted by:

$$\pi(\mathrm{d}z) \propto \exp(-\beta F(z))\mathbf{1}(z \in \mathcal{D}), \tag{3.4}$$

and study its consequence in recovering the true signal x^*. This leads to the first guarantees of a stochastic gradient-like method for compressed sensing with generative priors. In order to do so, we make the following three assumptions on the generator network $G(z)$.

(A.1) Boundedness. For all $z \in \mathcal{D}$, we have that $\|G(z)\| \leq B$ for some $B > 0$.

Algorithm 1 CS-SGLD

Input: step size η; inverse temperature parameter β, radius r and Lipschitz constant L of $F(z)$.
Draw z_0 from $\mu_0 = \mathcal{N}(0, \frac{1}{2L\beta}I)$ truncated on $\mathcal{D}$.
for $k = 0, 1, \dots,$ **do**
  Randomly sample $\xi_k \sim \mathcal{N}(0, I)$.
  $z_{k+1} = z_k - \eta\nabla_z F(z_k) + \sqrt{2\eta/\beta}\xi_k$
  if $z_{k+1} \notin \mathcal{B}(z_k, r) \cap \mathcal{D}$ **then**
    $z_{k+1} = z_k$
  end if
end for
Output: $\widehat{z} = \{z_k\}$.

[2] We not in passing that our analysis techniques succeed for any vector ε with bounded ℓ_2 norm.

(A.2) Near-isometry. $G(z)$ is a near-isometric mapping if there exist $0 < \iota_G \leq \kappa_G$ such that the following holds for any $z, z' \in \mathcal{D}$:

$$\iota_G \|z - z'\| \leq \|G(z) - G(z')\| \leq \kappa_G \|z - z'\|.$$

(A.3) Lipschitz gradients. The Jacobian of $G(z)$ is M-Lipschitz, i.e., for any $z, z' \in \mathcal{D}$, we have

$$\|\nabla_z G(z) - \nabla_z G(z')\| \leq M\|z - z'\|,$$

where $\nabla_z G(z) = \frac{\partial G(z)}{\partial z}$ is the Jacobian of the mapping $G(\cdot)$ with respect to z.

All three assumptions are justifiable. Assumption **(A.1)** is reasonable due to the bounded domain K and for well-trained generative models $G(z)$ whose target data distribution is normalized. Assumption **(A.2)** is reminiscent of the ubiquitous restricted isometry property (RIP) used for compressed sensing analysis [23] and is recently adopted in [12]. Finally, Assumption **(A.3)** is needed so that the loss function $F(z)$ is smooth, following typical analyses of Markov processes.

Next, we introduce a new concept of smoothness for generative networks. This concept is a weaker version of a condition on $G(\cdot)$ introduced in [12].

Definition 3.1 (Strong smoothness)**.** *The generator network $G(z)$ is (α, γ)-strongly smooth if there exist $\alpha > 0$ and $\gamma \geq 0$ such that for any $z, z' \in \mathcal{D}$, we have*

$$\langle G(z) - G(z'), \nabla_z G(z)(z - z')\rangle \geq \alpha\|z - z'\|^2 - \gamma. \quad (3.5)$$

Following [12] (Assumption 2), we call this property "strong smoothness". However, our definition of strong smoothness requires two parameters instead of one, and is weaker since we allow for an additive slack parameter $\gamma \geq 0$.

Definition 3.1 can be closely linked to the following property of the loss function $F(z)$ that turns out to be crucial in establishing convergence results for CS-SGLD.

Definition 3.2 (Dissipativity [24])**.** *A differentiable function $F(z)$ on $\mathcal{D}$ is (α, γ)-dissipative around z^* if for constants $\alpha > 0$ and $\gamma \geq 0$, we have*

$$\langle z - z^*, \nabla_z F(z)\rangle \geq \alpha\|z - z^*\|^2 - \gamma. \quad (3.6)$$

It is straightforward to see that (3.6) essentially recovers the strong smoothness condition (3.5) if the measurement matrix A is assumed to be the identity matrix. In compressed sensing, it is often the case that A is a (sub)Gaussian matrix and that given a sufficient number of measurements as well as Assumptions **(A.1)**, **(A.2)** and **(A.3)**, the dissipativity of $F(z)$ for such an A can still be established.

Once F is shown to be dissipative, the machinery of [**?**,14,15] can be adapted to show that the convergence of CS-SGLD. The majority of the remainder of the paper is devoted to proving this series of technical claims.

3.2. Main results

We first show that a very broad class of generator networks satisfies the assumptions made above. The following proposition is an extension of a result in [12].

Proposition 3.1. *Suppose $G(z) : \mathcal{D} \subset \mathbb{R}^d \to \mathbb{R}^n$ is a feed-forward neural network with layers of non-decreasing sizes and compact input domain $\mathcal{D}$. Assume that the non-linear activation is a continuously differentiable, strictly increasing function. Then, $G(z)$ satisfies Assumptions **(A.2)** & **(A.3)** with constants ι_G, κ_G, M, and if $2\iota_G^2 > M\kappa_G$, the strong smoothness in Definition 3.1 also holds almost surely with respect to the Lebesgue measure.*

This proposition merits a thorough discussion. First, architectures with increasing layer sizes are common; many generative models (such as GANs) assume architectures of this sort. Observe that the non-decreasing layer size condition is much milder than the expansivity ratios of successive layers assumed in related work [**?**, 19].

Second, the compactness assumption of the domain of G is mild, and traces its provenance to earlier related works [9, 12]. Moreover, common empirical techniques for training generative models (such as GANs) indeed assume that the latent vectors z lie on the surface of a sphere [25].

Third, common activation functions such as the sigmoid, or the Exponential Linear Unit (ELU) are continuously differentiable and monotonic. Note that the standard Rectified Linear Unit (ReLU) activation does *not* satisfy these conditions, and establishing similar results for ReLU networks is deferred to future work.

The key for our theoretical analysis, as discussed above, is Definition 3.1, and establishing this requires Proposition 3.1. Interestingly however, in Section 4 below we provide *empirical* evidence that strong smoothness holds for generative adversarial networks with ReLU activation trained on the MNIST and CIFAR-10 image datasets.

We now obtain a measurement complexity result by deriving a bound on the number of measurements required for F to be dissipative.

Lemma 3.1. *Let $G(z) : \mathcal{D} \subset \mathbb{R}^d \to \mathbb{R}^n$ be a feed-forward neural network that satisfies the conditions in Proposition 3.1. Let κ_G be its Lipschitz constant. Suppose the number of measurements m satisfies:*

$$m = \Omega\left(\frac{d}{\delta^2}\log(\kappa_G/\gamma)\right),$$

for some small constant $\delta > 0$. If the elements of A are drawn according to $\mathcal{N}(0, \frac{1}{m})$, then the loss function $F(z)$ is $(1 - \delta, \gamma)$-dissipative with probability at least $1 - \exp(-\Omega(m\delta^2))$.

The above result can be derived using covering number arguments, similar to the treatment in [9]. Observe that the number of measurements scales linearly with the dimension of the *latent* vector z instead of the *ambient* dimension, keeping in line with the flavor of results in standard compressed sensing. Recent lower bounds reported in [10] also have shown that the scaling of m with respect to d and $\log L$ might be *tight* for compressed sensing recovery in several natural parameter regimes.

We need two more quantities to state our convergence guarantee. Both definitions are widely used in the convergence analysis of MCMC methods. The first quantity defines the goodness of an initial distribution μ_0 with respect to the target distribution π.

Definition 3.3 (λ-warm start, [**?**])**.** *Let ν be a distribution on $\mathcal{D}$. An initial distribution μ_0 is λ-warm start with respect to ν if*

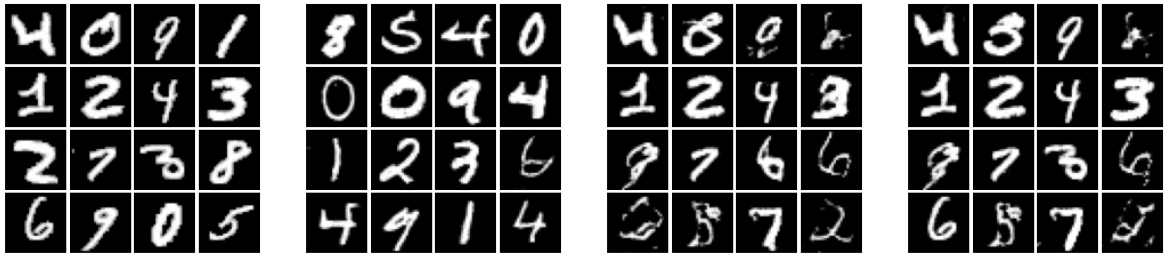
$$\sup_{\mathcal{A}:\mathcal{A}\subseteq\mathcal{D}} \frac{\mu_0(\mathcal{A})}{\nu(\mathcal{A})} \leq \lambda.$$

The next quantity is the Cheeger constant that connects the geometry of the objective function and the hitting time of SGLD to a particular set in the domain [15].

Definition 3.4 (Cheeger constant)**.** *Let μ be a probability measure on $\mathcal{D}$. We say μ satisfies the isoperimetric inequality with Cheeger constant ρ if for any $\mathcal{A} \subset \mathcal{D}$,*

$$\liminf_{h\to 0^+} \frac{\mu(\mathcal{A}_h) - \mu(\mathcal{A})}{h} \geq \rho \min\big\{\mu(\mathcal{A}), 1 - \mu(\mathcal{A})\big\},$$

where $\mathcal{A}_h = \{u \in K : \exists v \in \mathcal{A}, \|u - v\|_2 \leq h\}$.



(a) Ground truth (b) Initial (c) GD MSE = 0.0447 (d) SGLD MSE = 0.0275

Fig. 1. [MNIST] Comparing the recovery performance of SGLD and GD at $m = 0.2n$ measurements.

Our main theoretical result describing the convergence of Algorithm 1 (CS-SGLD) for compressed sensing recovery is given as follows. All proofs are in an expanded version of this paper [26].

Theorem 1 (Convergence of CS-SGLD)**.** *Assume that the generative network G satisfies Assumptions* **(A.1)** – **(A.3)** *as well as the strong smoothness condition. Consider a signal $x^* = G(z^*)$, and assume that it is measured with m (sub)Gaussian measurements such that $m = \Omega(d \log \kappa_G / \gamma)$. Choose an inverse temperature β and precision parameter $\epsilon > 0$. Then, after k iterations of SGLD in Algorithm 1, we obtain a latent vector z_k such that*

$$\mathbb{E}\left[F(z_k)\right] \leq \epsilon + O\left(\frac{d}{\beta}\log\left(\frac{\beta}{d}\right)\right), \tag{3.7}$$

provided the step size η and the number of iterations k are chosen such that:

$$\eta = \widetilde{O}\left(\frac{\rho^2\epsilon^2}{d^2\beta}\right), \text{ and } \quad k = \widetilde{O}\left(\frac{d^3\beta^2}{\rho^4\epsilon^2}\right).$$

In words, if we choose a high enough inverse temperature and appropriate step size, CS-SGLD converges (in expectation) to a signal estimate with very low loss within a polynomial number of iterations.

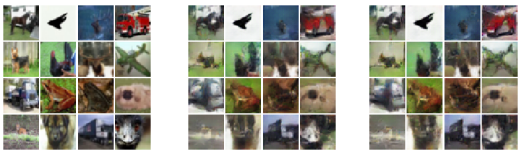
Let us parse the above result further. First, observe that the right hand side of (3.7) consists of two terms. The first term can be made arbitrarily small (at the cost of greater computational cost since η decreases). The second term represents the irreducible expected error of the exact sampling algorithm on the Gibbs measure $\pi(\mathrm{d}z)$, which is worse than the optimal loss obtained at $z = z^*$.

Second, suppose the right hand side of (3.7) is upper bounded by ϵ'. Once SGLD finds an ϵ'-approximate minimizer of the loss, in the regime of sufficient compressed sensing measurements (as specified by Lemma 3.1), we can invoke Theorem 1.1 in [9] along with Jensen's inequality to immediately obtain a recovery guarantee, i.e.,

$$\mathbb{E}\left[\|x^* - G(z_k)\|\right] \leq \sqrt{\epsilon'}.$$

Third, the convergence rate of CS-SGLD can be slow. In particular, SGLD may require a polynomial number of iterations to recover the true signal, while linearized ADMM [12] converges within a logarithmic number of iterations up to a *neighborhood* of the true signal. Obtaining an improved characterization of CS-SGLD convergence (or perhaps devising a new linearly convergent algorithm) is an important direction for future work.

Fourth, the above result is for noiseless measurements. A rather similar result can be derived with noisy measurements of bounded noise (says, $\|\varepsilon\| \leq \sigma$). This quantity (times a constant depending on A) will affect (3.7) up to an additive term that scales with σ. This is precisely in line with most compressed sensing recovery results and for simplicity we omit such a derivation.



(a) Ground truth (b) GD MSE = 0.0248 (c) SGLD MSE = 0.0246

Fig. 2. [CIFAR10] Comparing the recovery performance of SGLD and GD at $m = 0.3n$ measurements.

4. EXPERIMENTAL RESULTS

While we emphasize that the primary focus of our paper is theoretical, we corroborate our theory with representative experimental results on MNIST and CIFAR-10.

We test the SGLD reconstruction by using the update rule in (3.2) and compare against the optimizing the updates of z using standard gradient descent as in [9]. For all experiments, we use a pre-trained DCGAN generator, with network configuration described as follows: the generator consists of four different layers consisting of transposed convolutions, batch normalization and RELU activation; this is followed by a final layer with a transposed convolution and $\tanh$ activation [27].

We display the reconstructions on MNIST in Figure 1. Note that the implementation in [9] requires 10 random restarts for CS reconstruction and they report the results corresponding to the best reconstruction. This likely suggests that the standard implementation is likely to get stuck in bad local minima or saddle points. In Figure 1 we show reconstructions for the 16 different examples, which were all reconstructed at once using same $k = 2000$ steps, learning rate of $\eta = 0.02$ and the inverse temperature $\beta = 1$ for both approaches. The only difference is the additional noise term in SGLD (Figure 1 part (d)), which helps achieve better reconstruction performance compared to simple gradient descent.

Example reconstructions on CIFAR-10 images can be found in Fig. 2. More thorough empirical comparisons with PGD-based approaches [11, 28] are deferred to future work.

5. REFERENCES

[1] C. Dong, C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[2] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[3] J. Chang, C. Li, B. Póczos, B. Kumar, and A. Sankaranarayanan, "One network to solve them all—solving linear inverse problems using deep projection models," in *International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 5889–5898.

[4] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[5] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, pp. 1982–2001, 2010.

[6] A. Bora, A. Jalal, E. Price, and A. Dimakis, "Compressed sensing using generative models," in *International Conference on Machine Learning (ICML)*, 2017, pp. 537–546.

[7] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9446–9454.

[8] G. Jagatap and C. Hegde, "Algorithmic guarantees for inverse imaging with untrained network priors," in *Neural Information Processing Systems (NeurIPS)*, 2019.

[9] A. Bora, A. Jalal, E. Price, and A. Dimakis, "Compressed sensing using generative models," in *International Conference on Machine Learning (ICML)*, 2017, pp. 537–546.

[10] Z. Liu and J. Scarlett, "Information-theoretic lower bounds for compressive sensing with generative models," *IEEE Journal on Selected Areas in Information Theory*, 2020.

[11] V. Shah and C. 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)*. IEEE, 2018, pp. 4609–4613.

[12] F. Latorre, A. Eftekhari, and V. Cevher, "Fast and provable admm for learning with generative priors," in *Advances in Neural Information Processing Systems*, 2019, pp. 12 004–12 016.

[13] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *International Conference on Machine Learning (ICML)*, 2011.

[14] M. Raginsky, A. Rakhlin, and M. Telgarsky, "Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis," *arXiv preprint arXiv:1702.03849*, 2017.

[15] Y. Zhang, P. Liang, and M. Charikar, "A hitting time analysis of stochastic gradient langevin dynamics," in *Conference on Learning Theory (COLT)*, 2017, pp. 1980–2022.

[16] A. Jalal, S. Karmalkar, A. G. Dimakis, and E. Price, "Compressed sensing with approximate priors via conditional resampling," 2020, preprint.

[17] G. Ongie, A. Jalal, C. Metzler, R. Baraniuk, A. Dimakis, and R. Willett, "Deep learning techniques for inverse problems in imaging," *IEEE Journal on Selected Areas in Information Theory*, 2020.

[18] Q. Lei, A. Jalal, I. S. Dhillon, and A. G. Dimakis, "Inverting deep generative models, one layer at a time," in *Advances in Neural Information Processing Systems*, 2019, pp. 13 910–13 919.

[19] M. Asim, A. Ahmed, and P. Hand, "Invertible generative models for inverse problems: mitigating representation error and dataset bias," in *International Conference on Machine Learning (ICML)*, 2020.

[20] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," *International Conference on Learning Representations*, 2017.

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

[22] S. Ji and L. Carin, "Bayesian compressive sensing and projection optimization," in *International Conference on Machine Learning (ICML)*, 2007, pp. 377–384.

[23] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[24] J. Hale, "Asymptotic behavior of dissipative systems," *Bull. Am. Math. Soc*, vol. 22, pp. 175–183, 1990.

[25] T. White, "Sampling generative networks," *arXiv preprint arXiv:1609.04468*, 2016.

[26] T. Nguyen, G. Jagatap, and C. Hegde, "Provable compressed sensing with generative priors via langevin dynamics," *ArXiv preprint arXiv:2102.12643*, 2021.

[27] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[28] A. Raj, Y. Li, and Y. Bresler, "Gan-based projector for faster recovery with convergence guarantees in linear inverse problems," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 5602–5611.

[29] L. Lovász *et al.*, "Random walks on graphs: A survey," *Combinatorics*, vol. 2, no. 1, pp. 1–46, 1993.

[30] L. Lovász and S. Vempala, "The geometry of logconcave functions and sampling algorithms," *Random Structures & Algorithms*, vol. 30, no. 3, pp. 307–358, 2007.

[31] D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin, "A simple proof of the Poincaré inequality for a large class of probability measures," *Electronic Communications in Probability*, vol. 13, pp. 60–66, 2008.

[32] D. Zou, P. Xu, and Q. Gu, "Faster convergence of stochastic gradient langevin dynamics for non-log-concave sampling," in *Uncertainty in Artificial Intelligence (UAI)*, 2021, pp. 1152–1162.

[33] Y. T. Lee and S. S. Vempala, "Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation," in *Symposium on Theory of Computing (STOC)*, 2018, pp. 1115–1121.

A. PROOF OUTLINE

In this section, we provide a brief proof sketch of Theorem 1, while relegating details to the appendix.

At a high level, our analysis is an adaptation of the framework of [**?**, 15] specialized to the problem of compressed sensing recovery using generative priors. The basic ingredient in the proof is the use of conductance analysis to show the convergence of CS-SGLD to the target distribution in total variation distance.

Let μ_k denote the probability measure of Z_k generated by Algorithm 1 and π denote the target distribution in 3.4. The proof of Theorem 1 consists of three main steps:

1. First, we construct an auxiliary Metropolis-Hasting Markov process to show that μ_k converges to π in total variation for a sufficiently large k and a "good" initial distribution μ_0.
2. Then, we show that there exists an initial distribution μ_0 that serves as a λ-warm start with respect to π.
3. Finally, we show that a random draw from π is a near-minimizer of $F(z)$, proving that CS-SGLD recovers the signal to high fidelity.

We proceed with a characterization of the evolution of the distribution of z_k in Algorithm 1, which basically follows [**?**].

A.1. Construction of Metropolis-Hasting SGLD

Let $g(z) = \nabla_z F(z)$, and u and w be the points before and after one iteration of Algorithm 1; the Markov chain is written as $u \to v \to w$, where $v \sim \mathcal{N}(u - \eta g(u), \frac{2\eta}{\beta} I)$ with the following density:

$$P(v|u) = \left[\frac{1}{(4\pi\eta/\beta)^{d/2}} \exp\left(- \frac{\|v - u + \eta g(u)\|_2^2}{4\eta/\beta} \right) \middle| u \right]. \tag{A.1}$$

Without the correction step, $P(v|u)$ is exactly the transition probability of the standard Langevin dynamics. Note also that one can construct a similar density with a stochastic (mini-batch) gradient. The process of $v \to w$ is

$$w = \begin{cases} v & v \in \mathcal{B}(u, r) \cap \mathcal{D}; \\ u & \text{otherwise.} \end{cases} \tag{A.2}$$

Let $p(u) = \mathbb{P}_{v \sim P(\cdot|u)}[v \in \mathcal{B}(u, r) \cap \mathcal{D}]$ be the probability of accepting v. The conditional density $Q(w|u)$ is

$$Q(w|u) = (1 - p(u))\delta_u(w) + P(w|u) \cdot \mathbf{1}\big[w \in \mathcal{B}(u, r) \cap \mathcal{D}\big],$$

where $\delta_u(\cdot)$ is the Dirac-delta function at u. Similar to [**?**, 15], we consider the $1/2$-lazy version of the above Markov process, with the transition distribution

$$\mathcal{T}_u(w) = \frac{1}{2}\delta_u(w) + \frac{1}{2}Q(w|u), \tag{A.3}$$

and construct an auxiliary Markov process by adding an extra Metropolis accept/reject step. While proving the ergodicity of the Markov process with transition distribution $\mathcal{T}_u(w)$ is difficult, the auxiliary chain does indeed converge to a unique stationary distribution $\pi \propto e^{-\beta F(z)} \cdot \mathbf{1}(z \in \mathcal{D})$ due to the Metropolis-Hastings correction step.

The auxiliary Markov chain is given as follows: starting from u, let w be the state generated from $\mathcal{T}_u(\cdot)$. The Metropolis-Hasting SGLD accepts w with probability,

$$\alpha_u(w) = \min\left\{1, \frac{\mathcal{T}_w(u)}{\mathcal{T}_u(w)} \cdot \exp\big[-\beta\big(F(w) - F(u)\big)\big]\right\}.$$

Let $\mathcal{T}_u^\star(\cdot)$ denote the transition distribution of the auxiliary Markov process, such that

$$\mathcal{T}_u^\star(w) = (1 - \alpha_u(w))\delta(u) + \alpha_u(w)\mathcal{T}_u(w).$$

Below, we establish the connection between $\mathcal{T}_u(\cdot)$ and $\mathcal{T}_u^\star(\cdot)$, as well as the convergence of the original chain in Algorithm 1 through a conductance analysis on $\mathcal{T}_u^\star(\cdot)$.

Lemma A.1. *Under Assumptions, $F(z)$ is L-smooth and satisfies $\|\nabla_z F(z)\| \le D$ for $z \in \mathcal{D}$. For $r = \sqrt{10\eta d/\beta}$, the transition distribution of the chain in Algorithm 1 is δ-close the auxiliary chain, i.e., for any set $\mathcal{A} \subseteq \mathcal{D}$*

$$(1 - \delta)\mathcal{T}_u^\star(\mathcal{A}) \le \mathcal{T}_u(\mathcal{A}) \le (1 + \delta)\mathcal{T}_u^\star(\mathcal{A}).$$

where $\delta = 10Ld\eta + 10LDd^{1/2}\beta^{1/2}\eta^{3/2}$.

In Appendix D, we show that $F(z)$ is L-smooth with $L = (MB + \kappa_G^2)$ and its gradient is bounded by $D = \kappa_G^2\|A^\top A\|$.

One can verify that $\mathcal{T}_u^\star(\cdot)$ is time-reversible [15]. Moreover, following [29,30], the convergence of a time-reversible Markov chain to its stationary distribution depends on its conductance, which is defined as follows:

Definition A.1 (Restricted conductance)**.** *The conductance of a time-reversible Markov chain with transition distribution $\mathcal{T}_u^\star(\cdot)$ and stationary distribution π is defined by,*

$$\phi \triangleq \inf_{\mathcal{A}:\mathcal{A}\subseteq\mathcal{D},\pi(\mathcal{A})\in(0,1)} \frac{\int_{\mathcal{A}} \mathcal{T}_u(\mathcal{D}\backslash\mathcal{A})\pi(\mathrm{d}u)}{\min\{\pi(\mathcal{A}), \pi(\mathcal{D}\backslash\mathcal{A})\}}.$$

Using the conductance parameter ϕ and the closeness δ between $\mathcal{T}_u(\cdot)$ and $\mathcal{T}_u^\star(\cdot)$, we can derive the convergence of $\mathcal{T}_u(\cdot)$ in total variation distance.

Lemma A.2 ([**?**])**.** *Assume the conditions of Lemma A.1 hold. If $\mathcal{T}_u(\cdot)$ is δ-close to $\mathcal{T}_u^\star(\cdot)$ with $\delta \le \min\{1 - \sqrt{2}/2, \phi/16\}$, and the initial distribution μ_0 serves as a λ-warm start with respect to π, then*

$$\|\mu_k - \pi\|_{TV} \le \lambda\big(1 - \phi^2/8\big)^k + 16\delta/\phi.$$

We will further give a lower bound on δ in order to establish an explicit convergence rate.

Lemma A.3 ([**?**])**.** *Under the same conditions of Lemma A.1 and the step size $\eta \le \frac{1}{30Ld} \wedge \frac{d}{25\beta D^2}$, there exists a constant c_0 such that*

$$\phi \ge c_0\rho\sqrt{\eta/\beta}.$$

A.2. Convergence to the target distribution

Armed with these tools, we formally establish the first step of the proof.

Theorem 2. *Suppose that the generative network G satisfies Assumptions* **(A.1) – (A.3)** *as well as the strong smoothness condition. Set $\eta = O\big(d^{-1} \wedge \rho^2\beta^{-1}d^{-2}\big)$ and $r = \sqrt{10\eta d/\beta}$, then for any λ-warm start with respect to π, the output of Algorithm 1 satisfies*

$$\|\mu_k - \pi\|_{TV} \le \lambda(1 - C_0\eta)^k + C_1\eta^{1/2},$$

where ρ is the Cheeger constant of π, $C_0 = \widetilde{O}\big(\rho^2\beta^{-1}\big)$, and $C_2 = \widetilde{O}\big(d\beta^{1/2}\rho^{-1}\big)$. In particular, if the step size and the number of iterations satisfy:

$$\eta = \widetilde{O}\left(\frac{\rho^2\epsilon^2}{d^2\beta}\right), \text{ and } \quad k = \widetilde{O}\left(\frac{d^2\beta^2\log(\lambda)}{\rho^4\epsilon^2}\right),$$

then $\|\mu_k - \pi\|_{TV} \le \epsilon$ for $\epsilon > 0$.

The convergence rate is polynomial in the Cheeger constant ρ whose lower bound is difficult to obtain generally. A rough bound $\rho = e^{-\tilde{O}(d)}$ can be derived using the Poincaré constant of the distribution π, under the smoothness assumption. See [31] for details.

Proof outline of Theorem 2. To prove the result, we find a sufficient condition for η that fulfills the requirements of Lemmas A.1, A.2 and A.3 hold. For $\eta \leq \frac{d}{25\beta D^2}$, we have

$$\delta = 10Ld\eta + 10LDd^{1/2}\beta^{1/2}\eta^{3/2} \leq 12Ld\eta.$$

Moreover, Lemma A.2 requires $\delta \leq \min\{1-\sqrt{2}/2, \phi/16\}$, while $\phi \geq c_0\rho\sqrt{\eta/\beta}$ by Lemma A.3, so we can set

$$\eta = \min\left\{\frac{1}{30Ld}, \frac{d}{25\beta D^2}, \frac{c_0^2\rho^2}{(156Ld)^2\beta}\right\}$$

for these conditions to hold. Putting all together, we obtain

$$\begin{aligned}\|\mu_k - \pi\|_{TV} &\leq \lambda\big(1-\phi^2/8\big)^k + \frac{16\delta}{\phi} \\ &\leq \lambda(1-C_0\eta)^k + C_1\eta^{1/2},\end{aligned}$$

where $C_0 = c_0^2\rho^2/8\beta$, $C_1 = 156Ld\beta^{1/2}\rho^{-1}/c_0$. Therefore, we have proved the first part.

For the second part, to achieve ϵ-sampling error, it suffices to choose η and k such that

$$\lambda(1-C_0\eta)^k \leq \frac{\epsilon}{2}, \text{ and } \quad C_1\eta^{1/2} \leq \frac{\epsilon}{2}.$$

Plugging in C_0, C_1 above, we can choose

$$\eta = O\left(\frac{\rho^2\epsilon^2}{d^2\beta}\right) \text{ and } k = O\left(\frac{\log(\lambda/\epsilon)}{C_0\eta}\right) = \widetilde{O}\left(\frac{d^2\beta^2\log(\lambda)}{\rho^4\epsilon^2}\right)$$

such that $\|\mu_k - \pi\|_{TV} \leq \epsilon$, which completes the proof. □

A.3. Existence of warm start initial distribution

Apart from the step size and the number of iterations, the convergence depends on λ, the goodness of the initial distribution μ_0. In this part, we specify a particular choice of μ_0 in establish this.

Definition A.2 (Set-Restricted Eigenvalue Condition, [9])**.** *For some parameters $\tau > 0$ and $o \geq 0$, $A \in \mathbb{R}^{m\times n}$ is called S-REC(τ, o) if for all $z, z' \in \mathcal{D}$,*

$$\|A(G(z) - G(z'))\| \geq \tau\|G(z) - G(z')\| - o.$$

Lemma A.4. *Suppose that $G(z)$ satisfies the near-isometry property in Assumption **A.2**, and $F(z)$ is L-smooth. If A is S-REC$(\tau, 0)$, then the Gaussian distribution $\mathcal{N}(0, \frac{1}{2\beta L}I)$ supported on $\mathcal{D}$ is a λ-warm start with respect to π with $\lambda = e^{O(d)}$.*

Proof. Let μ_0 denote the truncated Gaussian distribution $\mathcal{N}(0, \frac{1}{2\beta L}I)$ on $\mathcal{D}$ whose measure is

$$\mu_0(\mathrm{d}z) = e^{-\beta L\|z\|_2^2}\mathbf{1}(z \in \mathcal{D})\mathrm{d}z/\Gamma$$

where $\Gamma = \int_{\mathcal{D}} e^{-\beta L\|z\|_2^2}\mathrm{d}z$ is the normalization constant. Along with the target measure π, we can easily verify that

$$\frac{\mu_0(\mathrm{d}z)}{\pi(\mathrm{d}z)} \leq \frac{\int_{\mathcal{D}} e^{-\beta F(z)}\mathrm{d}z}{\Gamma} \cdot e^{-\beta L\|z\|_2^2 + \beta F(z)}.$$

Our goal is to bound the right hand side. Using the smoothness and the simple fact $F(z^*) = 0$, we have

$$F(z) \leq \frac{L}{2}\|z - z^*\|_2^2 \leq L\|z^*\|_2^2 + L\|z\|_2^2,$$

which implies that $e^{-\beta L\|z\|_2^2 + \beta F(z)} \leq e^{\beta L\|z^*\|_2^2}$. To bound $\int_{\mathcal{D}} e^{-\beta F(z)}\mathrm{d}z$, we use the S-REC property of A as well as the near-isometry of $G(z)$. Recall the objective function:

$$\begin{aligned}F(z) &= \|y - AG(z)\|^2 = \|A(G(z) - G(z^*)\|^2 \\ &\geq \tau^2\|G(z) - G(z^*)\|^2 - o \geq \tau^2\iota_G^2\|z - z^*\|^2\end{aligned}$$

where we have dropped o for simplicity. Therefore,

$$\int_{\mathcal{D}} e^{-\beta F(z)}\mathrm{d}z \leq \int_{\mathcal{D}} e^{-\beta\tau^2\iota_G^2\|z-z^*\|^2}\mathrm{d}z \leq \left(\frac{\pi}{\beta\tau^2\iota_G^2}\right)^{d/2}.$$

Putting the above results together, we can get

$$\lambda \leq \max_{z\in K}\frac{\mu_0(\mathrm{d}z)}{\pi(\mathrm{d}z)} \leq \left(\frac{\pi}{\beta\tau^2\iota_G^2}\right)^{d/2}\frac{e^{\beta L\|z^*\|_2^2}}{\Gamma} = e^{O(d)},$$

and conclude the proof. □

A.4. Completing the proof

Proof of Theorem 1. Consider a random draw $\widehat{Z}$ from μ_k and another $\widehat{Z}^*$ from π. We have

$$\mathbb{E}[F(\widehat{Z})] = \Big(\mathbb{E}[F(\widehat{Z})] - \mathbb{E}[F(\widehat{Z}^*)]\Big) + \mathbb{E}[F(\widehat{Z}^*)]$$

We will first give a crude bound for the second term $\mathbb{E}[F(\widehat{Z}^*)]$ following the idea from [14]:

$$\mathbb{E}[F(\widehat{Z}^*)] = \int_{\mathcal{D}} F(z)\pi(\mathrm{d}z) \leq \mathcal{O}\left(\frac{d}{\beta}\log\frac{\beta}{d}\right).$$

The detailed proof is given in Appendix F.

The first term is related to the convergence of μ_k to π in total variation shown in Theorem 2. Notice that $F(z) \leq 2R\|A\|\kappa_G$ for all $z \in \mathcal{D}$ due the Lipschitz property of the generative network G. Moreover, by Theorem 2, we have $\|\mu_k - \pi\|_{TV} \leq \epsilon'$ for any $\epsilon' > 0$ and a sufficiently large k. Hence, the first term is upper bounded by

$$\begin{aligned}&\left|\int_{\mathcal{D}} F(z)\mu_k(\mathrm{d}z) - \int_{\mathcal{D}} F(z)\pi(\mathrm{d}z)\right| \\ &\leq 2R\|A\|\kappa_G\left|\int_{\mathcal{D}}\mu_k(\mathrm{d}z) - \int_{\mathcal{D}}\pi(\mathrm{d}z)\right| \leq 2R\|A\|\kappa_G\epsilon'.\end{aligned}$$

Given the target error ϵ, choose $\epsilon' = \epsilon/(2R\|A\|\kappa_G)$. By Lemma A.4, we have $\lambda = e^{O(d)}$. Then, for

$$\eta = \widetilde{O}\left(\frac{\rho^2\epsilon^2}{d^2\beta}\right), \text{ and } \quad k = \widetilde{O}\left(\frac{d^3\beta^2}{\rho^4\epsilon^2}\right), \text{ we have}$$

$$\mathbb{E}[F(\widehat{Z})] \leq \epsilon + \mathcal{O}\left(\frac{d}{\beta}\log\left(\frac{d+\gamma\beta}{\alpha\beta^2}\right)\right).$$

Therefore, we complete the proof of our main result. □

B. ADDITIONAL EXPERIMENTS

B.1. Validation of strong smoothness

We wish to verify whether the following condition holds for some $\alpha > 0$ and $\gamma \geq 0$:

$$\langle \nabla_z G(z)^\top (G(z) - G(z')), z - z' \rangle \geq \alpha \|z - z'\|^2 - \gamma \quad \text{(B.1)}$$

where z and z' are all possible pairs of latent vectors. To estimate these constants, we generate samples z and z' from $\mathcal{N}(0, \mathbb{I})$. To establish α and γ, we perform experiments on two different datasets (i) MNIST (Net1) and (ii) CIFAR10 (Net2). For both datasets, we compute the terms $u(z, z') = \nabla_z G(z)^\top (G(z) - G(z')), z - z' \rangle$ and $v(z, z') = \|z - z'\|^2$ for 500 different instantiations of z and z'. We then plot these pairs of $(\alpha v - \gamma, u)$ samples for different z's and z''s and tune the values of α and γ such that $u \geq \alpha v - \gamma$. We do this experiment for a DCGAN (Net1) generator trained on MNIST (Figure 3 (a)) as well as DCGAN (Net2) generator trained on CIFAR10 (Figure 3 (c)).

Similarly, we also derive values α_A and γ_A, where a compressive matrix A acts on the output of the generator G. Here we have picked $m = 0.1n$. This is encapsulated in the following equation:

$$\langle \nabla_z (AG(z))^\top (AG(z) - AG(z')), z - z' \rangle \geq \alpha_A \|z - z'\|^2 - \gamma_A \quad \text{(B.2)}$$

for all possible Gaussian matrices A and different instantiations of z and z'. Here, we capture the left side of the inequality in $u(z, z') = \langle \nabla_z (AG(z))^\top (AG(z) - AG(z')), z - z' \rangle$. We similarly plot points $(\alpha_A v - \gamma_A, u)$. The scatter plot generated for 50 different instantiations of z and z' and 5 different instantiations of A. We do this experiment for a DCGAN (Net1) generator trained on MNIST (Figure 3 (b)) as well as DCGAN (Net2) generator trained on CIFAR10 (Figure 3 (d)).

B.2. Reconstructions for CIFAR10

We display the reconstructions on CIFAR10 in Figure 4. As with the implementation for MNIST, for the sake of fair comparison, we fix the same random initialization of latent vector z for both GD and SGLD with no restarts. We select $m = 0.3n$. In Figure 4 we show reconstructions for the 16 different examples from MNIST, which were all reconstructed at once using same $k = 2000$ steps, learning rate of $\eta = 0.05$ and the inverse temperature $\beta = 1$ for both approaches. The only difference is the additional noise term in SGLD (Figure 1 part (d)). Similar to our experiments on MNIST we notice that this additional noise component helps achieve better reconstruction performance overall as compared to simple gradient descent.

Next, we plot phase transition diagrams by scanning the compression ratio $f = m/n = [0.2, 0.4, 0.6, 0.8, 1.0]$ for the MNIST dataset in Figure 5. For this experiment, we have chosen 5 different instantiations of the sampling matrix A for each compression ratio f. In Figure 5 we report the average Mean Square Error (MSE) of reconstruction $\|\hat{x} - x\|^2$ over 5 different instances of A. We conclude that SGLD gives improved reconstruction quality as compared to GD.

C. CONDITIONS ON THE GENERATOR NETWORK

Proposition C.1. *Suppose $G(z) : \mathcal{D} \subset \mathbb{R}^d \to \mathbb{R}^n$ is a feed-forward neural network with layers of non-increasing sizes and compact input domain $\mathcal{D}$. Assume that the non-linear activation is a continuously differentiable, strictly increasing function. Then, $G(z)$ satisfies Assumptions **(A.2)** & **(A.3)** with constants ι_G, κ_G, M, and if $2\iota_G^2 > M\kappa_G$, the strong smoothness in Definition 3.1 also holds almost surely with respect to the Lebesgue measure.*

Proof. The proof proceeds similar to [12], Appendix B. Since $G(z)$ is a composition of linear maps followed by C^1 activation functions, $G(z)$ is continuously differentiable. As a result, the Jacobian $\nabla_z G$ is a continuous matrix-valued function and its restriction to the compact domain $\mathcal{D} \subseteq \mathbb{R}^d$ is Lipschitz-continuous. Therefore, there exists $M \geq 0$ such that

$$\|\nabla_z G(z) - \nabla_z G(z')\| \leq M \|z - z'\|, \qquad \forall z, z' \in \mathcal{D}. \quad \text{(C.1)}$$

Thus, Assumption **(A.3)** holds. Assumption **(A.2)** is also satisfied according to [12], Lemma 5. To show the strong smoothness, we use the fundamental theorem of calculus with the Lipchitzness of $G(z)$ obtained by Assumption **(A.2)**. For every $z, z' \in \mathcal{D}$, and $u(t) = tz + (1-t)z'$:

$$\begin{aligned}
&\langle G(z) - G(z'), \nabla_z G(z)(z - z') \rangle \\
&\qquad = \|G(z) - G(z')\|^2 - \langle G(z) - G(z'), G(z) - G(z') - \nabla_z G(z \\
&\qquad = \|G(z) - G(z')\|^2 - \int_0^1 \langle G(z) - G(z'), \big(\nabla_z G(u(t)) - \nabla_z G \\
&\qquad \geq \iota_G^2 \|z - z'\|^2 - \kappa_G M \|z - z'\|^2 \int_0^1 (1-t) \mathrm{d}t \\
&\qquad = (\iota_G^2 - \frac{\kappa_G M}{2}) \|z - z'\|^2,
\end{aligned}$$

where in the last step we use the near-isometry and the Lipschitzness of $\nabla_z G(z)$ we have obtained. Consequently, $G(z)$ is $(\iota_G^2 - \frac{\kappa_G M}{2}, 0)$-strongly smooth, if $\iota_G^2 > \frac{\kappa_G M}{2}$. □

Lemma C.1 (Measurement complexity)**.** *Let $G(z) : \mathcal{D} \subset \mathbb{R}^d \to \mathbb{R}^n$ be a feed-forward neural network that satisfies the conditions in Proposition 3.1. Let L be its Lipschitz constant. If the number of measurements m satisfies:*

$$m = \Omega\left(\frac{d}{\delta^2} \log(\kappa_G / \gamma)\right),$$

for some small constant $\delta > 0$. If the elements of A are drawn according to $\mathcal{N}(0, \frac{1}{m})$, then the loss function $F(z)$ is $(\alpha - \delta\kappa_G^2, \gamma)$-dissipative with probability at least $1 - \exp(-\Omega(m\delta^2))$.

Proof. Using Proposition C.1, it follows that there exist $\alpha > 0$ and $\gamma \geq 0$ such that $G(z)$ is strongly smooth. Now, note that the left hand side of (3.6) is simplified as

$$\langle z - z^*, \nabla_z F(z) \rangle = \langle A(G(z) - G(z^*)), A \nabla_z G(z)(z - z^*) \rangle, \quad \text{(C.2)}$$

Denote $u = G(z) - G(z^*)$ and $v = \nabla_z G(z)(z - z^*)$, then

$$\langle z - z^*, \nabla_z F(z) \rangle = \langle Au, Av \rangle = \langle u, v \rangle - \langle (I - A^\top A)u, v \rangle.$$

Using standard result in random matrix theory, we can get $P(\|I - A^\top A\| \geq \delta) \leq \exp(-m\delta^2)$. Also, $\|u\|, \|v\| \leq \kappa_G \|z - z'\|$. Therefore,

$$\langle z - z^*, \nabla_z F(z) \rangle \geq \langle u, v \rangle - \delta \|z - z'\|^2.$$

For $m = \Omega\left(\frac{d}{\delta^2} \log(\kappa_G / \gamma)\right)$, then

$$\langle z - z^*, \nabla_z F(z) \rangle \geq (\alpha - \delta) \|z - z'\| - \gamma,$$

with probability at least $1 - \exp(-\Omega(m\delta^2))$. Therefore, the loss function $F(z)$ is $(\alpha - \delta\kappa_G^2, \gamma)$-dissipative with probability at least $1 - \exp(-\Omega(m\delta^2))$. □

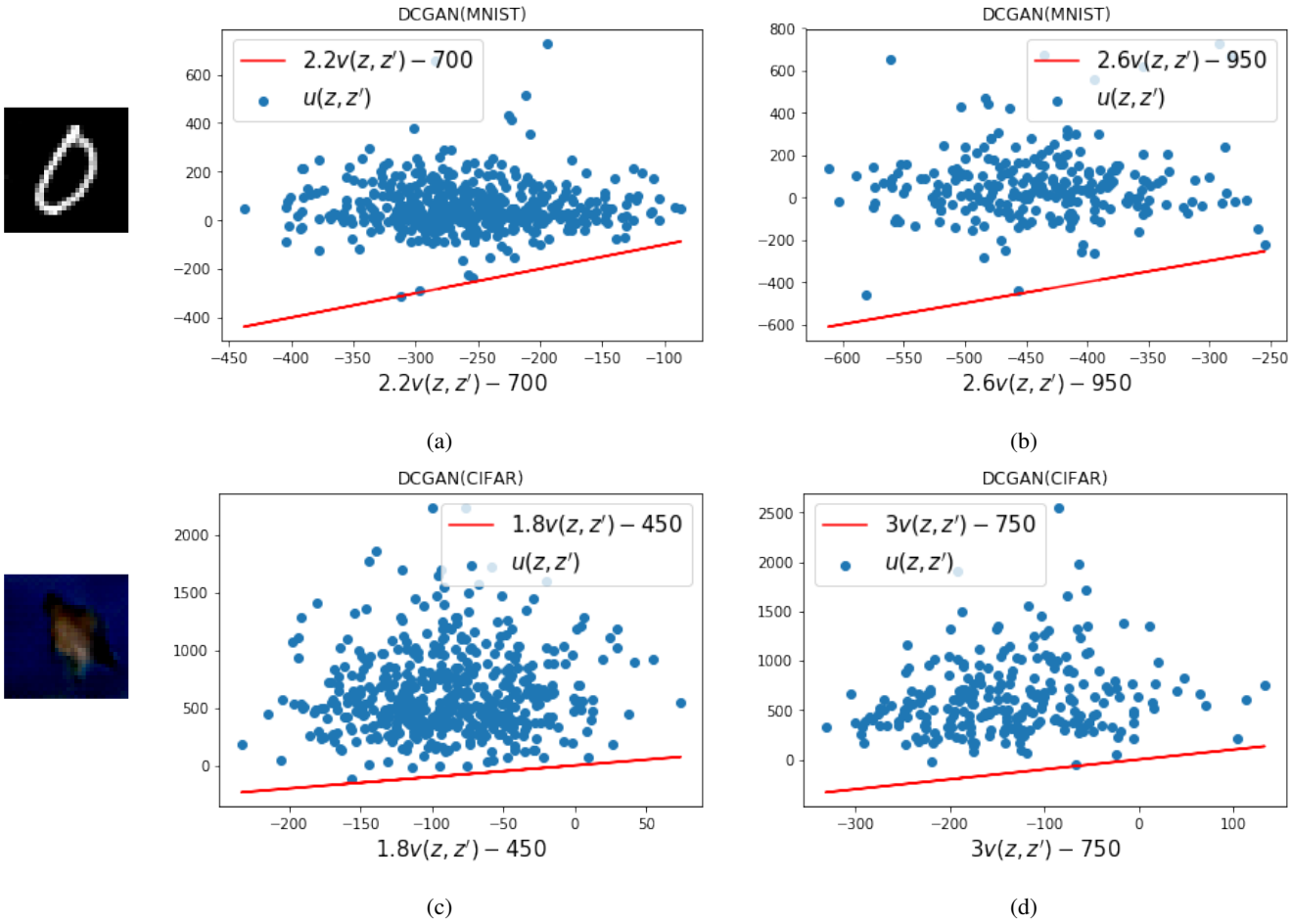


Fig. 3. [MNIST] selected base digit $G(z^*)$, evaluating (a) (B.1) (b) (B.2), [CIFAR] selected base image $G(z^*)$, evaluating (c) (B.1) (d) (B.2) .

(a) Ground truth (b) Initial (c) GD MSE = 0.0248 (d) SGLD MSE = 0.0246

Fig. 4. [CIFAR10] Comparing the recovery performance of SGLD and GD at $m = 0.3n$ measurements.

D. PROPERTIES OF THE LOSS FUNCTION

In this part, we establish some key properties of the loss function $F(z)$. We use Assumptions **(A.1)** – **(A.3)** on the boundedness, Lipschitz gradient and near-isometry to obtain an upper bound of $\|\nabla_z F(z)\|$ and the smoothness of $F(z)$.

Lemma D.1 (Lipschitzness of $F(z)$)**.** *We have* $\|\nabla_z F(z)\| \leq \kappa_G^2 \|A^\top A\| \|z - z^*\|$ *for any* $z \in \mathcal{D} \subset \mathbb{R}^d$.

Proof. Recall the gradient of $F(z)$:

$$\nabla_z F(z) = -(\nabla_z G(z))^\top A^\top (y - AG(z)) = -(\nabla_z G(z))^\top A^\top A(G(z^*) - G(z)$$

It follows from the Lipschitz assumption **(A.2)** that $\|G(z^*) - G(z)\| \leq \kappa_G \|z - z^*\|$, and hence $\|\nabla_z G(z)\| \leq \kappa_G$. Therefore,

$$\|\nabla_z F(z)\| \leq \kappa_G^2 \|A^\top A\| \|z - z^*\|.$$

□

Lemma D.2 (Smoothness of $F(z)$)**.** *For any* $z, z' \in \mathcal{D} \subset \mathbb{R}^d$, *we have*

$$\|\nabla_z F(z) - \nabla_z F(z')\| \leq (MB + \kappa_G^2) \|A^\top A\| \|z - z'\|.$$

Proof. We use the assumptions on $G(z)$ to derive the bound: $\|G(z^*)\| \leq B$.

$$\begin{aligned}\|\nabla_z F(z) - \nabla_z F(z')\| \leq\ & \|(\nabla_z G(z') - \nabla_z G(z))^\top A^\top A G(z^*)\| \\ & + \|(\nabla_z G(z))^\top A^\top A(G(z) - G(z'))\| \\ & + \|(\nabla_z G(z) - \nabla_z G(z'))^\top A^\top A G(z')\|\end{aligned}$$

Then, using the boundedness, Lipschitzness and smoothness, we arrive at:

$$\|\nabla_z F(z) - \nabla_z F(z')\| \leq (MB + \kappa_G^2) \|A^\top A\|.$$

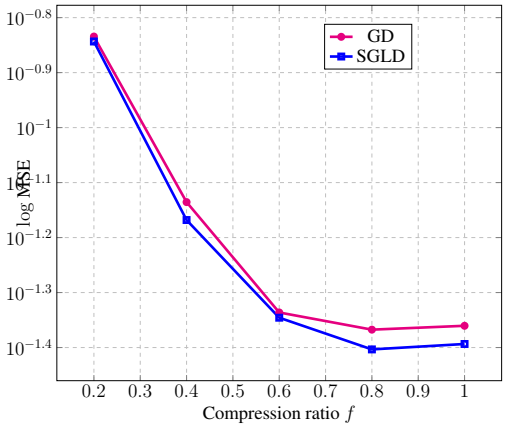


Fig. 5. Phase transition plots representing average MSE of reconstructed image using gradient descent and stochastic gradient Langevin dynamics.

Therefore, $F(z)$ is L-smooth, with $L = (MB + \kappa_G^2)\|A^\top A\|$. □

E. CONDUCTANCE ANALYSIS

In this section, we provide the proofs of Lemma A.1 and A.3 based on the conductance analysis laid out in [15] and similarly in [32]. The proof of A.2 directly follows from Lemma 6.3 of [32].

Proof of Lemma A.1. We use the same idea in Lemma 3 from [15] (and similarly in Lemma 6.1 from [32].) The main difference of our proof is that we use full gradient $\nabla_z F(z)$ in Algorithm 1, instead of stochastic mini-batch gradient, which simplifies the proof of this lemma a little.

We consider two cases for each u: $u \notin \mathcal{A}$ and $u \in \mathcal{A}$. As long as we can prove the first case, the second case easily follows, by splitting $\mathcal{A}$ into $\{u\}$ and $\mathcal{A}\backslash\{u\}$ and using the result of the first case. For a detailed treatment of the latter case, we refer the reader to the proof of Lemma 6.1 in [32].

Now that $u \notin \mathcal{A}$, we have

$$\mathcal{T}_u^\star(\mathcal{A}) = \int_{\mathcal{A}\cap\mathcal{B}(u,r)} \mathcal{T}_u^\star(w)\mathrm{d}w = \int_{\mathcal{A}\cap\mathcal{B}(u,r)} \alpha_u(w)\mathcal{T}_u(w)\mathrm{d}w. \tag{E.1}$$

where $\alpha_u(w)$ is the acceptance ratio of the Metropolis-Hasting. If suffices to show that $\alpha_u(w) \geq 1 - \delta/2$ for all $w \in K \cap \mathcal{B}(u, r)$, which implies

$$(1 - \delta/2)\mathcal{T}_u(\mathcal{A}) \leq \mathcal{T}_u^\star(\mathcal{A}) \leq \mathcal{T}_u(\mathcal{A}).$$

The right hand side is obvious by the definition of $\alpha_u(w)$ while we can ensure $\delta \leq 1/2$ with a sufficiently small η. What remains is to show that

$$\frac{\mathcal{T}_w(u)}{\mathcal{T}_u(w)} \cdot \exp(-\beta(F(w) - F(u))) \geq 1 - \delta/2. \tag{E.2}$$

The left hand side is simplified by definition of $\mathcal{T}_u(w)$ as

$$\exp\left(\frac{\|w - u + \eta g(u)\|_2^2}{4\eta/\beta} - \frac{\|u - w + \eta g(w)\|_2^2}{4\eta/\beta}\right)\exp(-\beta(F(w) - F(u))) \geq 1 - \delta/2$$

Note that $g(z) = \nabla_z F(z)$. Simplify the first exponent and combine with the second one gives the following form:

$$-\beta\left(F(w) - F(u) - \frac{1}{2}\langle w - u, \nabla_z F(w) + \nabla_z F(u)\rangle\right) + \frac{\eta\beta}{4}(\|\nabla_z F(u)\|^2 - \| \tag{E.3}$$

To lower bound the left hand side, we appeal to the smoothness of $F(z)$. Specifically, by Lemmas D.1 and D.2, we have F is L-smooth and $\|\nabla_z F(z)\| \leq D$ with $L = (MB + \kappa_G^2)$ and $D = \kappa_G^2\|A^\top A\|$. Then,

$$F(w) \leq F(u) + \langle w - u, \nabla F(u)\rangle + \frac{L\|w - u\|_2^2}{2},$$
$$F(u) \geq F(w) + \langle u - w, \nabla F(w)\rangle - \frac{L\|w - u\|_2^2}{2}.$$

This directly implies that

$$\left|F(w) - F(u) - \langle w - u, \frac{1}{2}\nabla F(w) + F(u)\rangle\right| \leq \frac{L\|w - u\|_2^2}{2}. \tag{E.4}$$

Moreover,

$$\begin{aligned}\left|\|\nabla_z F(u)\|_2^2 - \|\nabla_z F(w)\|_2^2\right| &\leq \|\nabla F(u) - \nabla F(w)\|_2 \cdot \|\nabla F(u) + \nabla F(w)\|_2 \\ &\leq 2LD\|w - u\|_2. \end{aligned} \tag{E.5}$$

Combining (E.4) and (E.5) in (E.3), and together with $w \in \mathcal{B}(u, r)$ with $r = \sqrt{10\eta d/\beta}$,

$$\begin{aligned}\text{LHS of (E.3)} &\geq -\frac{L\beta\|w - u\|^2}{2} - \frac{\eta\beta LD\|w - u\|}{2} \\ &\geq -5Ld\eta - 5LGd^{1/2}\beta^{-1/2}\eta^{3/2}.\end{aligned}$$

Pick $\delta/2 = 5Ld\eta + 5LDd^{1/2}\beta^{-1/2}\eta^{3/2}$, and use the fact $e^{-x} \geq 1 - x$ for $x \geq 0$, then we have proved the result. □

Next, we lower bound the conductance ϕ of $\mathcal{T}_u^\star(\cdot)$ using the idea in [32, 33], by first restating the following lemma:

Lemma E.1 (Lemma 13 in [33])**.** *Let $\mathcal{T}_u^\star(\cdot)$ be a time-reversible Markov chain on $\mathcal{D}$ with stationary distribution π. Suppose for any*

$u, v \in \mathcal{D}$ and a fixed $\Delta > 0$ such that $\|u - v\|_2 \leq \Delta$, we have $\|\mathcal{T}_u^\star(\cdot) - \mathcal{T}_v^\star(\cdot)\|_{TV} \leq 0.99$, then the conductance of $\mathcal{T}_u^\star(\cdot)$ satisfies $\phi \geq C\rho\Delta$ for some constant $C > 0$ and ρ is the Cheeger constant of π.

Proof of Lemma A.3. To apply Lemma E.1, we follow the same idea of [32] and reuse some of their results without proof. To this end, we prove that for some Δ, any pair of $u, v \in \mathcal{D}$ such that $\|u - v\|_2 \leq \Delta$, we have $\|\mathcal{T}_u^\star(\cdot) - \mathcal{T}_v^\star(\cdot)\|_{TV} \leq 0.99$. Recall the distribution of the iterate z after one-step standard SGLD without the accept/reject step in (A.1) is

$$P(z|u) = \frac{1}{(4\pi\eta/\beta)^{d/2}} \exp\left(- \frac{\|z - u + \eta g(u)\|_2^2}{4\eta/\beta} \right)$$

Since Algorithm 1 accepts the candidate only if it falls in the region $\mathcal{D} \cap \mathcal{B}(u, r)$, the acceptance probability is

$$p(u) = \mathbb{P}_{z \sim P(\cdot|u)}\big[z \in \mathcal{D} \cap \mathcal{B}(u, r)\big].$$

Therefore, the transition probability $\mathcal{T}_u^\star(z)$ for $z \in \mathcal{D} \cap \mathcal{B}(u, r)$ is given by

$$\mathcal{T}_u^\star(z) = \frac{2 - p(u) + p(u)(1 - \alpha_u(z))}{2} \delta_u(z) + \frac{\alpha_u(z)}{2} P(z|u) \cdot \mathbf{1}[z \in \mathcal{D} \cap \mathcal{B}(u, r)].$$

Take $u, v \in \mathcal{D}$ and let $\mathcal{S}_u = \mathcal{D} \cap \mathcal{B}(u, r)$ and $\mathcal{S}_v = \mathcal{D} \cap \mathcal{B}(v, r)$. By the definition of the total variation, there exists a set $\mathcal{A} \in \mathcal{D}$ such that

$$\begin{aligned}
\|\mathcal{T}_u^\star(\cdot) - \mathcal{T}_v^\star(\cdot)\|_{TV} &= |\mathcal{T}_u^\star(\mathcal{A}) - \mathcal{T}_v^\star(\mathcal{A})| \\
&\leq \underbrace{\max_{u,z} \left[\frac{2 - p(u) + p(u)(1 - \alpha_u(z))}{2} \right]}_{I_1} \\
&\quad + \frac{1}{2} \underbrace{\left| \int_{z \in \mathcal{A}} \alpha_u(z) P(z|u) \mathbf{1}(z \in \mathcal{S}_u) - \alpha_v(z) P(z|v) \mathbf{1}(z \in \mathcal{S}_v) \mathrm{d}z \right|}_{I_2}.
\end{aligned}$$

Using the mini-batch size that is exactly the same as the number of samples, we can reuse the bounds of I_1 and I_2 in Lemmas C.4 and C.5 of [32]. Consequently,

$$\|\mathcal{T}_u^\star(\cdot) - \mathcal{T}_v^\star(\cdot)\|_{TV} \leq I_1 + I_2/2 \leq 0.85 + 0.1\delta + \frac{\sqrt{\beta}\|u - v\|_2}{\sqrt{2\eta}}.$$

By Lemma A.1, we have $\delta = 10Ld\eta + 10LDd^{1/2}\beta^{1/2}\eta^{3/2} \leq 12Ld\eta$ if $\eta \leq \frac{d}{25\beta D^2}$. Thus if

$$\eta \leq \frac{1}{25\beta D^2} \wedge \frac{1}{30Ld\eta} \quad \text{and} \quad \|u - v\|_2 \leq \frac{\sqrt{2\eta}}{10\sqrt{\beta}} \leq 0.1r,$$

we have $\|\mathcal{T}_u^\star(\cdot) - \mathcal{T}_v^\star(\cdot)\|_{TV} \leq 0.99$. As the result of Lemma E.1, we prove a lower bound on the conductance ϕ of $\mathcal{T}_u^\star(\cdot)$

$$\phi \geq c_0 \rho \sqrt{\eta/\beta},$$

and finish the proof.

□

F. PROPERTY OF THE GIBBS ALGORITHM

Proposition F.1. *For $\mathcal{D} = \mathcal{B}(0, R)$, we have*

$$\int_{\mathcal{D}} F(z)\pi(\mathrm{d}z) \leq \mathcal{O}\left(\frac{d}{\beta} \log \frac{\beta L}{d} \right).$$

Proof. Let $p(z) = e^{-\beta F(z)}/\Lambda$ denote the density of π. $\Lambda \triangleq \int_{\mathcal{D}} e^{-\beta F(z)} \mathrm{d}z$ is the partition function. We start by writing

$$\int_{\mathcal{D}} F(z)\pi(\mathrm{d}z) = \frac{1}{\beta} \left(h(p) - \log \Lambda \right), \tag{F.1}$$

where

$$h(p) = -\int_{\mathcal{D}} p(z) \log p(z) \mathrm{d}z = -\int_K \frac{e^{-\beta F(z)}}{\Lambda} \log \frac{e^{-\beta F(z)}}{\Lambda} \mathrm{d}z$$

is the differential entropy of p. To upper-bound $h(p)$, we use the fact that the differential entropy of a probability density with a finite second moment is upper-bounded by that of a Gaussian density with the same second moment. Moreover, since p has the support in the Euclidean ball with radius R, its second moment is simply bounded by R^2. Therefore, we have

$$h(p) \leq h(\mathcal{N}(0, R^2 I)) = \frac{d}{2} \log \frac{2\pi R^2}{d}. \tag{F.2}$$

Next, we give a lower bound on the second term, $\log \Lambda$. We use the smoothness of $F(z)$ and the fact that z^* is the minimizer of F. We have $F(z) \leq \frac{L}{2}\|z - z^*\|^2$ for $z \in \mathcal{D}$. As such,

$$\log \Lambda = \log \int_{\mathcal{D}} e^{-\beta F(z)} \mathrm{d}z \geq \log \int_{\mathcal{D}} e^{-\beta L \|z - z^*\|^2/2} \mathrm{d}z \asymp O\left(\frac{d}{2} \log \frac{2\pi}{\beta L} \right). \tag{F.3}$$

Using (F.2) and (F.3) in (F.1) and simplifying, we prove the result. □