Score-Guided Intermediate Layer Optimization: Fast Langevin Mixing for Inverse Problems

Giannis Daras [*1] **Yuval Dagan** [*2] **Alexandros G. Dimakis** [3] **Constantinos Daskalakis** [2]

Abstract

We prove fast mixing and characterize the stationary distribution of the Langevin Algorithm for inverting random weighted DNN generators. This result extends the work of Hand and Voroninski from efficient inversion to efficient posterior sampling. In practice, to allow for increased expressivity, we propose to do posterior sampling in the latent space of a pre-trained generative model. To achieve that, we train a score-based model in the latent space of a StyleGAN-2 and we use it to solve inverse problems. Our framework, *Score-Guided Intermediate Layer Optimization (SGILO)*, extends prior work by replacing the sparsity regularization with a generative prior in the intermediate layer. Experimentally, we obtain significant improvements over the previous state-of-the-art, especially in the low measurement regime.

1. Introduction

We are interested in solving inverse problems with generative priors, a family of unsupervised imaging algorithms initiated by Compressed Sensing with Generative Models (CSGM) (Bora et al., 2017). This framework has been successfully applied to numerous inverse problems including non-linear phase retrieval (Hand et al., 2018), improved MR imaging (Kelkar & Anastasio, 2021; Darestani et al., 2021) and 3-D geometry reconstruction from a single image (Chan et al., 2021; Lin et al., 2022; Daras et al., 2021a), etc. CSGM methods can leverage any generative model including GANs and VAEs as originally proposed (Bora et al., 2017), but also invertible flows (Asim et al., 2019) or even untrained generators (Heckel & Hand, 2018).

One limitation of GAN priors when used for solving inverse problems is that the low-dimensionality of their latent space impedes the reconstruction of signals that lie outside their generation manifold. To mitigate this issue, sparse deviations were initially proposed in the pixel space (Dhar et al., 2018) and subsequently generalized to intermediate layers with Intermediate Layer Optimization (ILO) (Daras et al., 2021b). ILO extends the set of signals that can be reconstructed by allowing sparse deviations from the range of an intermediate layer of the generator. Regularizing intermediate layers is crucial when solving inverse problems to avoid overfitting to the measurements. In this work, we show that the sparsity prior is insufficient to prevent artifacts in challenging settings (e.g. inpainting with very few measurements, see Figure 1).

Recently, two new classes of probabilistic generative models, *Score-Based* networks (Song & Ermon, 2019) and *Denoising Diffussion Probabilistic Models (DDPM)* (Ho et al., 2020) have also been successfully used to solve inverse problems (Nichol et al., 2021; Jalal et al., 2021a; Song et al., 2021a; Meng et al., 2021; Whang et al., 2021). Score-Based networks and DDPMs both gradually corrupt training data with noise and then learn to reverse that process, i.e. they learn to create data from noise. A unified framework has been proposed in the recent Song et al. (2021b) paper and the broader family of such models is widely known as *Diffusion Models*. Diffusion models have shown excellent performance for conditional and unconditional image generation (Ho et al., 2020; Dhariwal & Nichol, 2021; Song et al., 2021b; Karras et al., 2022; Ramesh et al., 2022; Saharia et al., 2022), many times outpeforming GANs in image synthesis (Karras et al., 2019; 2020; Brock et al., 2019; Daras et al., 2020).

Unlike MAP methods, such as CSGM and ILO, solving inverse problems with Score-Based networks and DDPMs corresponds (assuming mixing) to sampling from the posterior. Recent work showed that posterior sampling has several advantages including diversity, optimal measurement scaling (Jalal et al., 2020; Nguyen et al., 2021) and reducing bias (Jalal et al., 2021c). The main weakness of this approach is that, in principle, mixing to the posterior

[*]Equal contribution [1]Department of Computer Science University of Texas at Austin [2]MIT CSAIL [3]Department of Electrical and Computer Engineering University of Texas at Austin. Correspondence to: Giannis Daras <giannisdaras@utexas.edu>, Yuval Dagan <dagan@mit.edu>, Alexandros G. Dimakis <dimakis@austin.utexas.edu>, Constantinos Daskalakis <costis@csail.mit.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

distribution can take exponentially many steps in the dimension n. In practice, Score-Based models usually require thousands of steps for a single reconstruction (Jolicoeur-Martineau et al., 2021; Xiao et al., 2021; Watson et al., 2021).

We show that (under the random weights assumption), CSGM with Stochastic Gradient Langevin Dynamics has polynomial (in the dimension) mixing to the stationary distribution. This result extends the seminal work of Hand & Voroninski (2018b); Huang et al. (2018) from MAP to posterior sampling. Specifically, Hand & Voroninski (2018b); Huang et al. (2018) established polynomial-time point convergence of Gradient Descent (with sign flips) for CSGM optimization for random weight ReLU Generators. We prove that, even without the sign flips, Langevin Dynamics will mix fast. Our result is important since prior work assumed mixing of the Markov Chain sampler to establish theoretical guarantees (e.g. see Jalal et al. (2020)).

Finally, we show how to solve inverse problems with posterior sampling in the latent space of a pretrained generator. Effectively, we combine ILO and Score-Based models into a single framework for inverse problems. We call our new method Score-Guided Intermediate Layer Optimization (SGILO). *The central idea is to create generative models that come endowed with a score-based model as a prior for one internal intermediate layer in their architecture*. This replaces the sparsity prior used by ILO with a learned intermediate layer regularizer.

We start with a StyleGAN2 (Karras et al., 2019; 2020) and train a score-based model to learn the distribution of the outputs of an intermediate layer. To solve an inverse problem, we optimize over an intermediate layer as in ILO (Daras et al., 2021b), but instead of constraining the solutions to sparse deviations near the range, we use the learned score as a regularization. Specifically, we are using Stochastic Gradient Langevin Dynamics (SGLD) to sample from the posterior distribution of the latents where the gradient of the log-density is provided by our score-based model.

Our Contributions:

1. We propose a novel framework, Score-Guided Intermediate Layer Optimization (SGILO), for solving general inverse problems. Our method replaces the sparsity prior of ILO (Daras et al., 2021b) with a learned score-based prior.

2. To learn this prior we train a score-based model on an intermediate latent space of StyleGAN using inversions of real images from FFHQ (Karras et al., 2019) obtained with ILO (Daras et al., 2021b). Our score-based models use a Vision Transformer (ViT) (Dosovitskiy et al., 2020) variant as the backbone architecture, demonstrating design flexibility when training score models for intermediate representations.

3. Given some measurements (e.g. inpainted image), we use the learned prior and the Langevin algorithm to do posterior sampling. Experimentally we show that our approach yields significant improvements over ILO (Daras et al., 2021b) and other prior work. Further, we show that our Langevin algorithm is much faster to train and to sample from, compared to standard score-based generators, since we work in the much lower dimension of the intermediate layer.

4. Theoretically we prove that the Langevin algorithm converges to stationarity in polynomial time. Our result extends prior work (Hand & Voroninski, 2018b; Huang et al., 2018) which analyzed MAP optimization to Langevin dynamics. Like prior work, our theory requires that the generator has random independent weights and an expansive architecture.

5. We open-source all our code and pre-trained models to facilitate further research on this area.

2. Score Guided Intermediate Layer Optimization

Setting Let $x^* \in \mathbb{R}^n$ be an unknown vector that is assumed to lie in the range of a pre-trained generator $G(z) : \mathbb{R}^k \to \mathbb{R}^n$, i.e. we assume that there is a $z^* \in \mathbb{R}^k$ such that: $x^* = G(z^*)$. We observe some noisy measurements of x^*, i.e. the vector $y = \mathcal{A}(x^*) + \xi \in \mathbb{R}^m$, where $A : \mathbb{R}^n \to \mathbb{R}^m$ is a known, differentiable forward operator and $\xi \sim \mathcal{N}(0, \sigma^2 I)$.

Posterior Sampling in the Latent Space We first want to characterize the posterior density $p(z|y)$. Applying Bayes rule, we get that: $p(z|y) = \frac{p(z,y)}{p(y)} \propto p(y|z)p(z)$. The noise is assumed Gaussian, so $p(y|z) = \mathcal{N}(y; \mu = \mathcal{A}(G(z)), \Sigma = \sigma^2 I)$. Hence,

$$\log p(z|y) \propto \underbrace{\frac{1}{2\sigma^2}||\mathcal{A}(G(z)) - y||_2^2 - \log p(z)}_{L(z)} \quad . \tag{1}$$

To derive this posterior, we assumed that x^* is in the range of the generator G. This assumption is somewhat unrealistic for the Gaussian latent space of state-of-the-art GANs, such as StyleGAN (Karras et al., 2019; 2020) which motivates optimization over an intermediate space, as done in ILO (Daras et al., 2021b).

ILO has two weaknesses: i) it is a MAP method while there is increasing research showing the benefits of posterior

Algorithm	Expressive	Sampling	Fast	Provable Convergence
Gradient Descent in $\mathbb{R}^k$ (CSGM (Bora et al., 2017))	✗	✗	✓	✓
Projected Gradient Descent in $\mathbb{R}^p$ (ILO (Daras et al., 2021b))	✓	✗	✓	✓
Langevin Dynamics in $\mathbb{R}^n$ (Jalal et al., 2021b)	✓	✓	✗	✗
Langevin Dynamics in $\mathbb{R}^p$ (**SGILO**)	✓	✓	✓	✓(under assumptions)

Table 1: Summary of different reconstruction algorithms for solving inverse problems with deep generative priors. For the GAN based methods (Rows 1, 2), we think of a generator as a composition over two transformations $G_1 : \mathbb{R}^k \to \mathbb{R}^p$ and $G_2 : \mathbb{R}^p \to \mathbb{R}^n$, where $k < p < n$. Gradient Descent in the intermediate space, as in the ILO paper, can be expressive (increased expressivity due to ILO) and fast (GAN-based methods) but does not offer diverse sampling. On the other hand, Stochastic Gradient Langevin Dynamics in the pixel space is slow as it is usually done with high-dimensional score-based models. SGILO (Row 4) combines the best of the two worlds.

Reference | Input | SGILO (Ours) | ILO | CSGM

Figure 1: Results on randomized inpainting in the very challenging regime of only $\mathbf{0.75}\%$ observed pixels (with random sampling). The input seems completely black unless zoomed in. The proposed SGILO benefits from the intermediate layer score-based model to remove artifacts and unnatural colors that appear in ILO (Daras et al., 2021b). CSGM (Bora et al., 2017) is constrained to be on the range of StyleGAN2 and hence produces high quality images that, however, do not resemble much the (unobserved) reference. We emphasize that these are real reference images that have not been used in training, for any of the models.

sampling (Jalal et al., 2020; 2021c; Nguyen et al., 2021), ii) it is assuming a handcrafted prior which is uniform in an l_1 dilation of the range of the previous layers and 0 elsewhere.

Instead, we propose a new framework, Score-Guided Intermediate Layer Optimization (SGILO), that trains a score-based model in the latent space of some intermediate layer and then uses it with Stochastic Gradient Langevin Dynamics to sample from $e^{-bL(z)}$ for some temperature parameter.

Figure 2 illustrates the central idea of SGILO. As shown, ILO optimizes in the intermediate layer $\mathbb{R}^p$ assuming a uniform prior over the expanded manifold (that is colored green). In this paper, we learn a distribution in the intermediate layer using a score based model. This learned distribution is shown by orange geodesics and can expand outside the ℓ_1-ball dilated manifold.

Table 1 summarizes the strengths and weaknesses of the following reconstruction algorithms: i) Gradient Descent (GD) in the latent space of the first layer of a pre-trained generator as in the CSGM (Bora et al., 2017) framework, ii) (Projected) GD in the latent space of an intermediate layer, as in ILO (Daras et al., 2021b), iii) Stochastic Gradient Langevin Dynamics (SGLD) in the pixel space, as done by Jalal et al. (2020); Song et al. (2021a) and others with Score-Based Networks and iv) SGLD in the intermediate space

of some generator, as we propose in this work. Notation wise, for the GAN based methods (Rows 1, 2), we think of a generator as a composition over two transformations $G_1 : \mathbb{R}^k \to \mathbb{R}^p$ and $G_2 : \mathbb{R}^p \to \mathbb{R}^n$, where $k < p < n$.

Gradient Descent in the intermediate space, as in the ILO paper, can be expressive (increased expressivity due to ILO) and fast (GAN-based methods) but does not offer diverse sampling. On the other hand, Stochastic Gradient Langevin Dynamics in the pixel space is slow as it is usually done with high-dimensional score-based models. SGILO (Row 4) combines the best of the worlds of GAN-based inversion and posterior sampling with Score-Based Networks. Specifically, it is expressive (optimization over an intermediate layer), it offers diverse sampling (posterior sampling method) and it is fast (dimensionality $p < n$). Experimental evidence that supports these claims is given in Section 4.

The last column of Table 1 characterizes the different algorithms with respect to what we know about their convergence. A recent line of work (Hand & Voroninski, 2018b; Huang et al., 2018; Daskalakis et al., 2020a) has been able to prove that despite the non-convexity, for neural networks with random Gaussian weights, a signal in the range of the network can be approximately recovered using Gradient Descent (with sign flips) in polynomial time under an expansion assumption in the dimension of the layers of the generator. This motivates the question of whether we can prove under the same setting, that a Langevin Sampling algorithm would converge fast to a stationary measure. The next section, answers affirmatively this question while even removing the need for sign flipping. The theoretical results hold for the CSGM setting, but can apply to the optimization in an intermediate layer with a uniform prior over the latents. Unfortunately, assuming uniformity in the intermediate layer is not a realistic assumption. Proving distributional convergence of SGILO under more realistic assumptions is left for future work.

3. Theoretical Results

We are now ready to state the main Theorem of our paper.

Theorem 3.1 (Informal)**.** *Consider the Markov Chain defined by the following Langevin Dynamics:*

$$z_{t+1} = z_t - \eta \nabla f(z_t) + \sqrt{2\eta\beta^{-1}}u \tag{2}$$

where u is a zero-mean, unit variance Gaussian vector, i.e. $u_{ij} \sim \mathcal{N}(0, \sigma^2 = 1)$, $G(z)$ is a fully-connected d-layer ReLU neural network,

$$G(z) = \text{ReLU}\left(W^{(d)}\left(\cdots \text{ReLU}\left(W^{(1)}z\right)\cdots\right)\right)$$

and $f(z)$ is the loss function:

$$f(z) = \beta\|AG(z) - y\|_2^2$$

where $A \in \mathbb{R}^{m\times k}$, and $y = AG(z^*)$, for some unknown vector $z^* \in \mathbb{R}^n$.

Define $\mu(z) \propto e^{-f(z)}$ and $z_t \sim Z_t$, then for any $\epsilon > 0$ and for $t \geq \Omega(\log(1/\epsilon)/\epsilon^2)$,

$$\mathcal{W}(Z_t, \mu) := \inf_{Q \in \{\text{couplings of } Z_t, \mu\}} \mathbb{E}_{(z_t,z)\sim Q}\|z_t - z\| \leq (\epsilon + e^{-\Omega(n)})\|z^*\|,$$

provided that $\eta = \Theta(\epsilon^2)$, that $\beta = Cn$ (for some sufficiently large constant C), that $\|z_0\| \leq O(\|z^*\|)$, that $W^{(i)}$ and A satisfy conditions WDC and RRIC (Hand & Voroninski, 2018b) and $d \geq 2$ can be any constant.

We note that $\beta = \Theta(n)$ is the right choice of parameters since a smaller β produces approximately random noise and a larger β produces a nearly deterministic output.

Sketch of the proof: We analyze the landscape of the loss function f. It was already noted by Hand & Voroninski (2018a) that it has three points where the gradient vanishes: at the optimum $\vec{x}^*$, at a point $-\rho\vec{x}^*$ for some $\rho \in (0, 1)$ and at 0, a local maxima. In order to escape the stationary point $-\rho\vec{x}^*$, Hand & Voroninski (2018a) proposed to flip the sign of $\vec{x}$ whenever such flipping reduces the loss. We write f in a more compact fashion, obtaining that $-\rho\vec{x}^*$ is a saddle point. We show that the noise added by the Langevin dynamics can help escaping this point, and converging to some ball around $\vec{x}^*$. This is proven via a potential function argument: we construct a potential V and show that it decreases in expectation after each iteration, as long as the current iteration is far from $\vec{x}^*$. We note that the expected change in V is measured in the continuous dynamics by a Laplace operator $\mathcal{L}V$. In this paper, we use this to show that the potential decreases in the continuous dynamics, and compare the continuous to the discrete dynamics.

Finally, our goal is to couple the discrete dynamics to the continuous dynamics that sample from μ. Once we establish that the continuous and discrete dynamics arrive close to $\vec{x}^*$, we use the fact that f is strongly convex in this region to couple them in such a way that they get closer in each iteration, until they are ϵ-close, and this concludes the proof. The full proof and the detailed formal statement of the theorem can be found in the Appendix.

4. Experimental Results

We use StyleGAN-2 (Karras et al., 2019; 2020) as our pretrained GAN generator. Score-based models are trained as priors for internal StyleGAN-2 layers. We use a variant of the Vision Transformer (Dosovitskiy et al., 2020) as the backbone architecture for our score-based models. To incorporate time information, we add Gaussian random features (Tancik et al., 2020) to the input sequence, as done

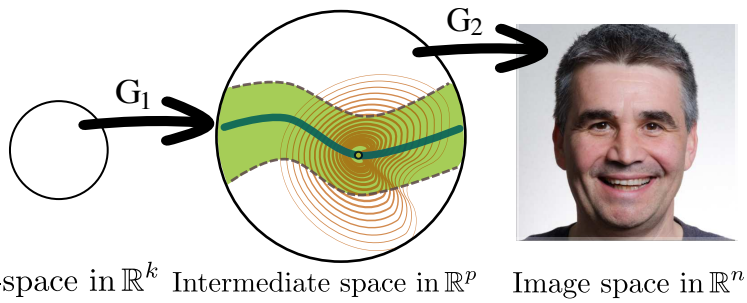

Figure 2: An illustration of SGILO. In previous work (ILO (Daras et al., 2021b)) a generator is considered as the composition of two transformations G_1 from the latent space to an intermediate space $\mathbb{R}^p$ and a second transformation G_2 from the intermediate layer to the image space. The range of the generator G_1 is a k dimensional manifold in $\mathbb{R}^p$ shown with a blue line in the figure. ILO expands this by taking the Minkowski sum of the manifold with the ℓ_1 ball. ILO optimizes in the intermediate layer $\mathbb{R}^p$ assuming a uniform prior over the expanded manifold, shown as the green set. In this paper we learn a distribution in the intermediate layer using a score based model. This learned distribution is shown by orange geodesics and can expand outside the ℓ_1-ball dilated manifold.

in Song & Ermon (2019) for the U-net (Ronneberger et al., 2015) architecture. The score-based models are trained with the Variance Preserving (VP) SDE, defined in Song et al. (2021b).

Transformers are not typically used for score-based modeling. This is probably due to the quadratic complexity of transformers with respect to the length of the input sequence, e.g. for training a 1024x1024x3 score-based model, the Transformer would require memory proportional to $1024^2 \times 1024^2 \times 3^2$. Since our score-based models learn the distribution of intermediate StyleGAN-2 layers, we work with much lower dimensional objects. For the score-based model, we use a ViT Transformer with 8 layers, 1 attention head and dimension 1024. For the VP-SDE we use the parameters in Song et al. (2021b). For more information on implementation and hyperparameters, please refer to our open-sourced code.

Dataset and training. The score-based model is trained by creating a dataset of intermediate StyleGAN inputs (latents and intermediate outputs). We inverted all images in FFHQ (Karras et al., 2019) with ILO, and used the intermediate outputs as training data for our score-based model.

We train score-based models to learn the distribution of: i) the latents that correspond to the inverted FFHQ and ii) the intermediate distributions of layers $\{1, 2, 3, 4\}$ of the generator. Consistently, we observed that the score-based models for the deeper priors were more powerful in terms of solving inverse problems. This is expected but comes with the cost of more expensive training, which is why we stopped at layer 4, which is already powerful enough to give us excellent reconstructions.

Unconditional Image Generation. The first experiment we run aims to demonstrate that the score-based models that we trained on the intermediate distributions are indeed capable of modeling accurately the gradient of the log-density. To this end, we use Annealed Langevin Dynamics, as in Song & Ermon (2019), to sample from the intermediate distribution of the fourth layer and the distribution of the inverted latents. The results are summarized in Figure 4. In the first two rows, we show results when sampling from the intermediate distribution (keeping the noises and the latent vectors fixed). In the last row, we show results when sampling from the distribution of the inverted latents (keeping the noises fixed). As shown, the combination of the score-models and the powerful StyleGAN generators leads to diverse, high-quality generations.

Quantitative Results on Inverse Problems. We want to evaluate if our method qualitatively improves upon ILO (Daras et al., 2021b) which is the previous state-of-the-art method for solving inverse problems with pre-trained generators. We also compare with vanilla CSGM (Bora et al., 2017) which performs much worse. For a fair comparison, we choose 8 real images from FFHQ to tune the hyperparameters for each method at each measurement level, and then measure performance with respect to the ground truth on 30 FFHQ test set images (never seen by the score-based model). For ILO, we also tried the default parameters (300, 300, 300, 100 steps) reported in Daras et al. (2021b). Finally, to make sure that the benefit of our method comes indeed from the prior and not from optimizing without the ILO sparsity constraints, we also test ILO without any constraint on the optimization space. In the Figures, for the ILO we report the minimum of ILO with tuned parameters, ILO with default parameters (from the paper) and ILO without any regularization. For the denoising experiments, we tried ILO with and without dynamic addition of noise (Stochastic Noise Addition) and we plotted the best score.

Figure 3 shows MSE and Perceptual distance between the

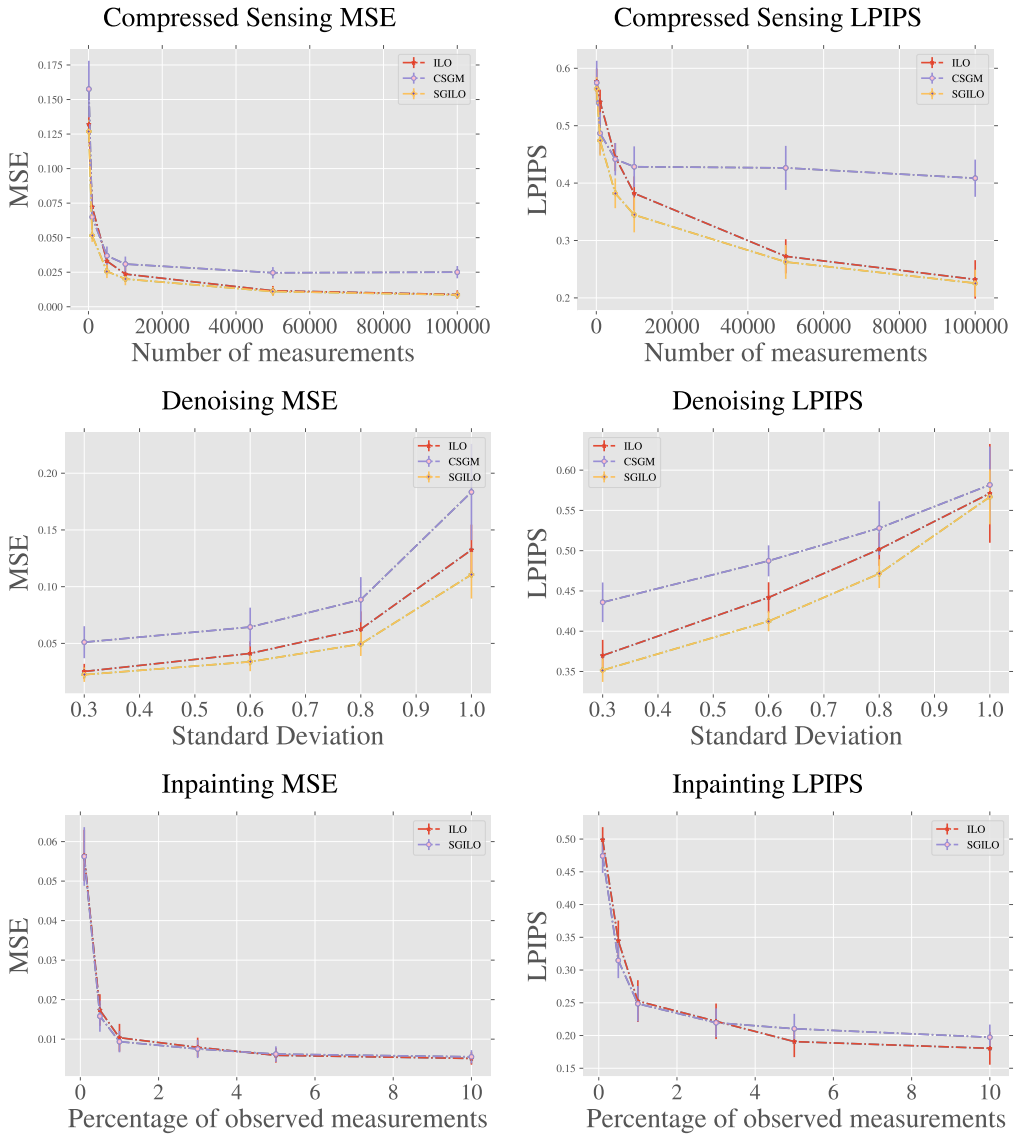

Figure 3: Quantitative results on the task of compressed sensing, denoising and inpainting. Our method, SGILO, significantly outperforms the state-of-the-art unsupervised method ILO when the measurements are scarce or the level of the noise is high. In this challenging regime, the prior from the score-based model is a much better regularizer than the sparse deviations constraint of ILO, yielding significant performance boosts.

ground truth image and the reconstructions of ILO, CSGM and SGILO (ours) as we vary the difficulty of the task (horizontal axis). The plots show results for denoising, compressed sensing and inpainting. As shown, in more challenging regimes, our method outperforms ILO and CSGM. When the task is not very challenging, e.g. denoising when the standard deviation of the noise is $\sigma = 0.1$, the prior becomes less important and SGILO performs on par with ILO. As expected, the contribution of the prior is significant when less information is available.

Out of distribution projections. This experiment demonstrates that following the learned vector field of the log-likelihood in the latent space leads to more natural images. Specifically, we start with an out-of-distribution image and we invert it with ILO. For the purposes of this experiment, we intentionally stop ILO early, to arrive at a solution that has visual artifacts. We now use solely the score-based prior to see if we can correct these artifacts. We obtain the early stopped ILO solution $z_0^* \in \mathbb{R}^p$ (where p is the dimension of the intermediate layer, optimizing over the third layer of StyleGAN-2), and use the Forward SDE to sample from

Figure 4: Images generated by a pre-trained StyleGAN-2 (Karras et al., 2020) with inputs to intermediate layers sampled with our trained-score based models and the Annealed Langevin Dynamics algorithm (Song & Ermon, 2019).

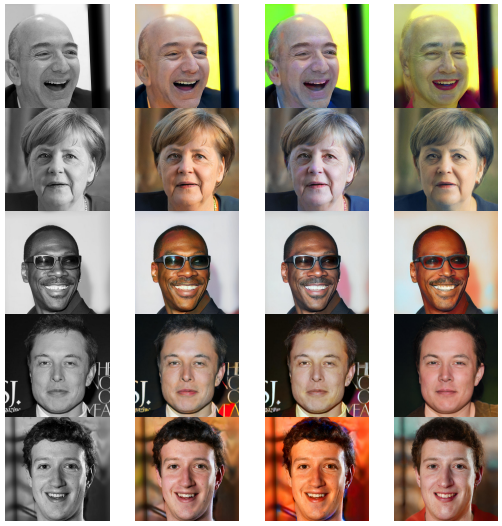

Figure 5: Results on colorization. ILO introduces artifacts, e.g. Row 1, column 3. Those artifacts are mostly corrected by SGILO, that displays more natural colors than prior work.

$p(z_t|z_0^*)$. This corresponds to sampling from a Gaussian centered in z_0^* with variance that grows with t. Then, we use Annealed Langevin Dynamics with the learned Score-Based network to sample from $p(z_0|z_t)$. The choice of t is affecting how close will be the final sample to z_0^*. Since we started with an unnatural latent, we expect that t is controlling the trade-off between matching the measurements and increasing the naturalness of the image. Results are shown in Figure 6.

Other Inverse Problems. SGILO is a general framework that can be applied to solve any inverse problem, as long as the forward operator $\mathcal{A}$ is known and differentiable. Figure 1 shows results for randomized inpainting in the extreme regime of 0.75% observed measurements.

Figure 5 shows results for the task of colorization. As shown, both ILO and CSGM introduce artifacts, e.g. see columns 3 and 4 of Row 1. Those artifacts are mostly corrected by our framework, SGILO, that displays more natural colors than prior work.

A final experiment we performed is generating samples using a pre-trained classifier to deviate from the learned distribution. We use a classifier to bias our Langevin algorithm to produce samples that look like ImageNet classes. We use gradients from robust classifiers (Santurkar et al., 2019) to get samples from the class 'Bullfrog'. As shown in Figure 7, SGILO is flexible to produce samples outside its learned distribution and retains interesting visual features.

Posterior Sampling Ablation As we argued in Sections 1, 2, SGILO is a posterior sampling method. Among others, posterior sampling offers i) diverse reconstructions, ii) reduced bias. We perform two experiments to examine how well SGILO performs with respect to i) and ii). Figure 8 shows different reconstructions we get for the measurements given in the first column. As shown, the generated images have variability with respect to age, ethnicity, eye color, etc. We also perform a preliminary experiment to examine whether SGILO has the potential to reduce dataset biases. To that end, we downsample 64 images of men and women (each) by $\times 256$ and then reconstruct them using ILO and SGILO. For each of the reconstructed images, we

Figure 6: Out of distribution projections. The initial painting image is our of the StyleGAN2 distribution. Using CSGM frequently fails especially when features are not perfectly aligned with the learned manifold. ILO produces a better image that is still not very realistic. Our method (SGILO) uses the Score-prior to improve on ILO and produce a more realistic image. Note that the goal is not to reconstruct the input image, but to demonstrate that by exclusively following the learned score in the latent space makes the generated image more natural.

Figure 7: Samples of the posterior using a Bullfrog classifier as a differentiable forward operator. SGILO is flexible and able to extend outside its learned distribution as it produces interesting blendings of human and frog characteristics.

Figure 8: Diverse reconstructions with posterior sampling.

use CLIP (Radford et al., 2021) to predict the gender. ILO predicts correctly the gender in $78/128$, while SGILO succeeds in $\mathbf{89/128}$. This experiment aligns with the findings of Jalal et al. (2021c) that shows that reconstruction methods based on posterior sampling usually lead to increased fairness.

Speed Ablation One advantage of SGILO over other sampling frameworks with conventional Score-Based networks is the speed. The reasons SGILO is fast are twofold: i) the model is working on a low-dimensional space and ii) one might not need to reverse the whole diffusion, since any step of SGLD can serve as a hot-start for the reverse diffusion. For most of our experiments, instead of using directly the gradient of $\log p(z_0)$, we sample for some small t one $p(z_t|z_0)$ according to the SDE and then we run the reverse SDE for the interval $[t, 0]$. This can give more flexibility to the score-based model to guide the solutions of the Intermediate Layer Optimization and is still pretty fast as long as t is small. This is similar in spirit to the SDEdit (Meng et al., 2021) paper. The only time we revert the whole diffusion is when we treat the score-based model as a generator (instead of regularizer for ILO), as we do for ablation purposes in Figure 4. Specifically, each inverse problem takes $1-2$ minutes to get solved on a single V100 GPU. Figure 9 of the Appendix shows how MSE changes as time goes. SGILO typically requires 300 function evaluations which corresponds to $1-2$ minutes. Most score-based models, like NCSNv3, require thousands of steps. For image generation, SGILO needs ~ 40 seconds for a single sample on a single GPU, which is $10\times$ faster than score models in the pixel space. We note that recently many other methods for accelerating diffusion models have been proposed (Karras et al., 2022; Salimans & Ho, 2022; Nichol & Dhariwal, 2021; Song et al., 2020; Jolicoeur-Martineau et al., 2021) that are orthogonal to (and hence, can be combined with) our approach.

5. Related Work

The CSGM paper (Bora et al., 2017) introduced the unsupervised framework for inverse problems and this has been shown to be optimal under natural assumptions (Liu & Scarlett, 2020; Kamath et al., 2019). Recent works have investigated methods of expanding the range of the generator. Optimizing an intermediate layer was first proposed in the context of inversion as a way to identify data distribution modes dropped by GANs (Bau et al., 2019). The same technique has been rediscovered in the GAN surgery paper (Park et al., 2020), in which the authors demonstrated (among other things) that the expansion of the range is useful for out-of-distribution generation. Intermediate Layer Optimization (Daras et al., 2021b) improved prior work by i) using the powerful StyleGAN generator (as was first pioneered in PULSE (Menon et al., 2020) for the special case of super-resolution), ii) gradually transitioning to higher layers with sequential optimization, iii) regularizing the solutions of CSGM (Bora et al., 2017) by only allowing sparse deviations from the range of some intermediate layer. Dhar et al. (2018) previously proposed extending the range by sparse deviations from the output space, but ILO generalized this by allowing deviations from the range of *any* layer.

Score-based modeling was proposed by Song & Ermon (2019) using Score Matching (Hyvärinen, 2005) and further work (Song et al., 2021b; Dhariwal & Nichol, 2021; Song & Ermon, 2020) significantly improved score-based performance. Our work is not the first one to train score-based model in the latent space. Vahdat et al. (2021) also trains score-based models in the latent space of a VAE to improve generation quality and sampling time. Our work is related but we are training score-based networks on already pre-trained generators and we are focusing on solving inverse problems (instead of generation) by formulating the SGILO algorithm. Algorithms for solving inverse problems with score-based models in pixel-space have been developed in the interesting works of Kawar et al. (2021; 2022); Jalal et al. (2021a). We do not compare directly with these methods since they use different generators as priors for solving inverse problems.

On the theoretical side, our work extends the seminal work of Hand & Voroninski (2018b); Huang et al. (2018). Prior work showed that a variant of Gradient Descent converges polynomially for MAP estimation using random weight ReLU Generators. Our result is that Langevin Dynamics gives polynomial convergence to the posterior distribution under the same setting. Prior work has also analyzed convergence of Langevin Dynamics for non-convex optimization under different set of assumptions, e.g. see Raginsky et al. (2017); Block et al. (2020); Xu et al. (2017). For theoretical guarantees for sampling with generative models with latent diffusions, we also refer the interested reader to the relevant work of Tzen & Raginsky (2019).

Finally, it is useful to underline that in the presence of enough training data, end-to-end supervised methods usually outperform unsupervised methods, e.g. see Tian et al. (2020); Sun et al. (2020); Tripathi et al. (2018) for denoising, Sun & Chen (2020); Yang et al. (2019) for super-resolution and Yu et al. (2019); Liu et al. (2019) for inpainting. The main disadvantages of solving inverse problems with end-to-end supervised methods are that: i) separate training is required for each problem, ii) there is significant fragility to forward operator changes (robustness issues) (Darestani et al., 2021; Ongie et al., 2020).

6. Conclusions

This paper introduced Score-Guided Intermediate Layer Optimization (SGILO), a framework for posterior sampling in the latent space of a pre-trained generator. Our work extends the sparsity prior that appeared in prior work, with a powerful generative prior that is used to solve inverse problems. On the theoretical side, we proved fast convergence of the Langevin Algorithm for random weights generators, for the simplified case of uniform prior over the latents.

7. Acknowledgments

This research has been supported by NSF Grants CCF 1763702, 1934932, AF 1901281, 2008710, 2019844 the NSF IFML 2019844 award as well as research gifts by Western Digital, WNCG and MLL, computing resources from TACC and the Archie Straiton Fellowship. This work is also supported by NSF Awards CCF-1901292, DMS-2022448 and DMS2134108, a Simons Investigator Award, the Simons Collaboration on the Theory of Algorithmic Fairness, a DSTA grant, the DOE PhILMs project (DE-AC05-76RL01830).

References

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

Bau, D., Zhu, J.-Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., and Torralba, A. Seeing what a gan cannot generate, 2019.

Block, A., Mroueh, Y., Rakhlin, A., and Ross, J. Fast mixing of multi-scale langevin dynamics under the manifold hypothesis, 2020.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. In *International*

Conference on Machine Learning, pp. 537–546. PMLR, 2017.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis, 2019.

Chan, E. R., Monteiro, M., Kellnhofer, P., Wu, J., and Wetzstein, G. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021. doi: 10.1109/cvpr46437.2021.00574. URL http://dx.doi.org/10.1109/CVPR46437.2021.00574.

Daras, G., Odena, A., Zhang, H., and Dimakis, A. G. Your local gan: Designing two dimensional local attention mechanisms for generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Daras, G., Chu, W.-S., Kumar, A., Lagun, D., and Dimakis, A. G. Solving inverse problems with nerfgans. *arXiv preprint arXiv:2112.09061*, 2021a.

Daras, G., Dean, J., Jalal, A., and Dimakis, A. G. Intermediate layer optimization for inverse problems using deep generative models. In *ICML 2021*, 2021b.

Darestani, M. Z., Chaudhari, A., and Heckel, R. Measuring robustness in deep learning based compressive sensing. *arXiv preprint arXiv:2102.06103*, 2021.

Daskalakis, C., Rohatgi, D., and Zampetakis, E. Constant-expansion suffices for compressed sensing with generative priors. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13917–13926. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/file/9fa83fec3cf3810e5680ed45f7124dce-Paper.pdf.

Daskalakis, C., Rohatgi, D., and Zampetakis, M. Constant-expansion suffices for compressed sensing with generative priors. in the. In *34th Annual Conference on Neural Information Processing Systems (NeurIPS), NeurIPS 2020*, 2020b.

Dhar, M., Grover, A., and Ermon, S. Modeling sparse deviations for compressed sensing using generative models. In *International Conference on Machine Learning*, pp. 1214–1223. PMLR, 2018.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Hand, P. and Voroninski, V. Global guarantees for enforcing deep generative priors by empirical risk. In *Conference On Learning Theory*, pp. 970–978. PMLR, 2018a.

Hand, P. and Voroninski, V. Global guarantees for enforcing deep generative priors by empirical risk. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 970–978. PMLR, 06–09 Jul 2018b. URL http://proceedings.mlr.press/v75/hand18a.html.

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

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

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

Huang, W., Hand, P., Heckel, R., and Voroninski, V. A provably convergent scheme for compressive sensing under random generative priors, 12 2018.

Huang, W., Hand, P., Heckel, R., and Voroninski, V. A provably convergent scheme for compressive sensing under random generative priors. *Journal of Fourier Analysis and Applications*, 27(2):1–34, 2021.

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL http://jmlr.org/papers/v6/hyvarinen05a.html.

Jalal, A., Karmalkar, S., Dimakis, A., and Price, E. Compressed sensing with approximate priors via conditional resampling. In *NeurIPS 2020 Workshop on Deep Learning and Inverse Problems*, 2020. URL https://openreview.net/forum?id=8ozSD4Oymw.

Jalal, A., Arvinte, M., Daras, G., Price, E., Dimakis, A. G., and Tamir, J. I. Robust compressed sensing mri with deep generative priors, 2021a.

Jalal, A., Karmalkar, S., Dimakis, A. G., and Price, E. Instance-optimal compressed sensing via posterior sampling, 2021b.

Jalal, A., Karmalkar, S., Hoffmann, J., Dimakis, A., and Price, E. Fairness for image generation with uncertain sensitive attributes. In *International Conference on Machine Learning*, pp. 4721–4732. PMLR, 2021c.

Jolicoeur-Martineau, A., Li, K., Piché-Taillefer, R., Kachman, T., and Mitliagkas, I. Gotta go fast when generating data with score-based models, 2021.

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

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. doi: 10.1109/cvpr.2019.00453. URL http://dx.doi.org/10.1109/CVPR.2019.00453.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. doi: 10.1109/cvpr42600.2020.00813. URL http://dx.doi.org/10.1109/cvpr42600.2020.00813.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models, 2022.

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

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models, 2022.

Kelkar, V. A. and Anastasio, M. A. Prior image-constrained reconstruction using style-based generative models. *arXiv preprint arXiv:2102.12525*, 2021.

Lin, C. Z., Lindell, D. B., Chan, E. R., and Wetzstein, G. 3d gan inversion for controllable portrait image animation, 2022.

Liu, H., Jiang, B., Xiao, Y., and Yang, C. Coherent semantic attention for image inpainting. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. doi: 10.1109/iccv.2019.00427. URL http://dx.doi.org/10.1109/ICCV.2019.00427.

Liu, Z. and Scarlett, J. Information-theoretic lower bounds for compressive sensing with generative models. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 292–303, May 2020. ISSN 2641-8770. doi: 10.1109/jsait.2020.2980676. URL http://dx.doi.org/10.1109/JSAIT.2020.2980676.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2021.

Menon, S., Damian, A., Hu, S., Ravi, N., and Rudin, C. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. doi: 10.1109/cvpr42600.2020.00251. URL http://dx.doi.org/10.1109/cvpr42600.2020.00251.

Nguyen, T. V., Jagatap, G., and Hegde, C. Provable compressed sensing with generative priors via langevin dynamics, 2021.

Nichol, A. and Dhariwal, P. Improved denoising diffusion probabilistic models, 2021.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2021.

Ongie, G., Jalal, A., Metzler, C. A., Baraniuk, R. G., Dimakis, A. G., and Willett, R. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020.

Park, J. Y., Smedemark-Margulies, N., Daniels, M., Yu, R., van de Meent, J.-W., and HAnd, P. Generator surgery for compressed sensing. In *NeurIPS 2020 Workshop on Deep Learning and Inverse Problems*, 2020. URL https://openreview.net/forum?id=s2EucjZ6d2s.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pp. 1674–1703. PMLR, 2017.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F. (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

Salimans, T. and Ho, J. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Santurkar, S., Tsipras, D., Tran, B., Ilyas, A., Engstrom, L., and Madry, A. Image synthesis with a single (robust) classifier, 2019.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models, 2020.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models, 2020.

Song, Y., Shen, L., Xing, L., and Ermon, S. Solving inverse problems in medical imaging with score-based generative models, 2021a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Sun, W. and Chen, Z. Learned image downscaling for upscaling using content adaptive resampler. *IEEE Transactions on Image Processing*, 29:4027–4040, 2020. ISSN 1941-0042. doi: 10.1109/tip.2020.2970248. URL http://dx.doi.org/10.1109/TIP.2020.2970248.

Sun, Y., Liu, J., and Kamilov, U. S. Block coordinate regularization by denoising. *IEEE Transactions on Computational Imaging*, 6:908–921, 2020. ISSN 2573-0436. doi: 10.1109/tci.2020.2996385. URL http://dx.doi.org/10.1109/TCI.2020.2996385.

Tancik, M., Srinivasan, P. P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J. T., and Ng, R. Fourier features let networks learn high frequency functions in low dimensional domains, 2020.

Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., and Lin, C.-W. Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275, Nov 2020. ISSN 0893-6080. doi: 10.1016/j.neunet.2020.07.025. URL http://dx.doi.org/10.1016/j.neunet.2020.07.025.

Tripathi, S., Lipton, Z. C., and Nguyen, T. Q. Correction by projection: Denoising images with generative adversarial networks. *arXiv preprint arXiv:1803.04477*, 2018.

Tzen, B. and Raginsky, M. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Conference on Learning Theory*, pp. 3084–3114. PMLR, 2019.

Vahdat, A., Kreis, K., and Kautz, J. Score-based generative modeling in latent space. *arXiv preprint arXiv:2106.05931*, 2021.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Watson, D., Ho, J., Norouzi, M., and Chan, W. Learning to efficiently sample from diffusion probabilistic models, 2021.

Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A. G., and Milanfar, P. Deblurring via stochastic refinement, 2021.

Xiao, Z., Kreis, K., and Vahdat, A. Tackling the generative learning trilemma with denoising diffusion gans, 2021.

Xu, P., Chen, J., Zou, D., and Gu, Q. Global convergence of langevin dynamics based algorithms for nonconvex optimization, 2017.

Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.-H., and Liao, Q. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21 (12):3106–3121, 2019.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. Free-form image inpainting with gated convolution. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. doi: 10.1109/iccv.2019.00457. URL http://dx.doi.org/10.1109/ICCV.2019.00457.

Notation

$\mathbb{P}_X^{x,t}$ Probability Density of a continuous Markov Chain that started from point x on the space of all continuous paths on $[0,t]$.

$g^{(d)}$ Synthesis of scalar function g with itself, d times.

$v\mathbb{P}_X^t$ Probability Density of a Markov Chain at time t.

$||x||$ l_2 norm of vector x.

A. Formal statement

In the theorem, we make use of a d-layer ReLU network G,

$$G(z) = \text{ReLU}\left(W^{(d)}\left(\cdots \text{ReLU}\left(W^{(1)}z\right)\cdots\right)\right)$$

where each $w_i \in \mathbb{R}^{n_i \times n_{i-1}}$. Further, we make use of a matrix A of dimension $n_d \times k$. We use the same assumptions on G and A as in Hand & Voroninski (2018b). Our setting has a minor difference compared to Hand & Voroninski (2018b): to remove unnecessary scalings, we scale the distribution of the weights by a factor of 2 at every layer.

Definition A.1. We say that the matrix $W \in \mathbb{R}^{n\times k}$ satisfies the *Weight Distribution Condition* with constant ϵ if for all nonzero $x, y \in \mathbb{R}^k$,

$$\left\|\sum_{i=1}^{n} 1_{w_i\cdot x>0}1_{w_i\cdot y>0}\cdot w_i w_i^t - Q_{x,y}\right\| \le \epsilon, \text{ with } Q_{x,y} = \frac{\pi-\theta_0}{2\pi}I_k + \frac{\sin\theta_0}{2\pi}M_{\hat{x}\leftrightarrow\hat{y}}, \tag{3}$$

where $w_i \in \mathbb{R}^k$ is the ith row of W; $M_{\hat{x}\leftrightarrow\hat{y}} \in \mathbb{R}^{k\times k}$ is the matrix[1] such that $\hat{x} \mapsto \hat{y}$, $\hat{y} \mapsto \hat{x}$, and $z \mapsto 0$ for all $z \in \text{span}(\{x,y\})^\perp$; $\hat{x} = x/\|x\|_2$ and $\hat{y} = y/\|y\|_2$; $\theta_0 = \angle(x,y)$; and 1_S is the indicator function on S.

Definition A.2. We say that the compression matrix $A \in \mathbb{R}^{m\times n}$ satisfies the *Range Restricted Isometry Condition (RRIC)* with respect to G with constant ϵ if for all $x_1, x_2, x_3, x_4 \in \mathbb{R}^k$,

$$\begin{aligned}\left|\left\langle A\big(G(x_1)-G(x_2)\big), A\big(G(x_3)-G(x_4)\big)\right\rangle - \left\langle G(x_1)-G(x_2), G(x_3)-G(x_4)\right\rangle\right| \\ \le \epsilon\|G(x_1)-G(x_2)\|_2\|G(x_3)-G(x_4)\|_2.\end{aligned} \tag{4}$$

We assume that each matrix $W^{(i)}$ in the network G satisfies WDC and that the matrix A satisfies RRIC. Such assumptions hold for random matrices:

Theorem A.3. *Let $\epsilon > 0$. Suppose that each entry of each weight matrix $W^{(i)} \in \mathbb{R}^{n_i\times n_{i-1}}$ is drawn i.i.d. $N(0,1)$, and suppose that each entry of $A \in \mathbb{R}^{m\times k}$ is drawn i.i.d. Further, suppose that for all $i = 1,\ldots,d$, $n_i/n_{i-1} \ge C\epsilon^{-2}\log(1/\epsilon)$ and suppose that $m \ge C\epsilon^{-1}\log(1/\epsilon)dn\log(\prod_{i=1}^d n_i)$, where $n = n_0$ and $C > 0$ is a universal constant. Then, with probability $1 - e^{-cn}$, WDC is satisfied for all matrices $W^{(i)}$ and RRIC is satisfied for A.*

We prove the following theorem, which assumes that WDC and RRIC are satisfied:

Theorem A.4. *Consider the Markov Chain defined by the following Langevin Dynamics:*

$$z_{t+1} = z_t - \eta\nabla f(z_t) + \sqrt{2\eta\beta^{-1}}u \tag{5}$$

where $u \sim N(0, I_n)$ is a zero-mean, unit variance Gaussian vector, $G(z)$ is a fully-connected d-layer ReLU neural network,

$$G(z) = \text{ReLU}\left(W^{(d)}\left(\cdots \text{ReLU}\left(W^{(1)}z\right)\cdots\right)\right)$$

[1] A formula for $M_{\hat{x}\leftrightarrow\hat{y}}$ is as follows. If $\theta_0 = \angle(\hat{x},\hat{y}) \in (0,\pi)$ and R is a rotation matrix such that $\hat{x}$ and $\hat{y}$ map to e_1 and $\cos\theta_0\cdot e_1 + \sin\theta_0\cdot e_2$ respectively, then $M_{\hat{x}\leftrightarrow\hat{y}} = R^t\begin{pmatrix}\cos\theta_0 & \sin\theta_0 & 0\\ \sin\theta_0 & -\cos\theta_0 & 0\\ 0 & 0 & 0_{k-2}\end{pmatrix}R$, where 0_{k-2} is a $k-2\times k-2$ matrix of zeros. If $\theta_0 = 0$ or π, then $M_{\hat{x}\leftrightarrow\hat{y}} = \hat{x}\hat{x}^t$ or $-\hat{x}\hat{x}^t$, respectively.

and $f(z)$ is the loss function:

$$f(z) = \beta \|AG(z) - y\|_2^2$$

where $A \in \mathbb{R}^{n_d \times k}$, and $y = AG(z^*)$, for some unknown vector $z^* \in \mathbb{R}^n$.

Define $\mu(z) \propto e^{-f(z)}$ and denote by Z_t the distribution over z_t, i.e. the t'th step of the dynamics. Then, there exist constants C_1, C_2, C_3, C_4 that depend only on d such that the following holds: For any $\epsilon > 0$ and for $t \geq C_1 \log(1/\epsilon)/\epsilon^2$,

$$\begin{aligned}\mathcal{W}(Z_t, \mu) := \inf_{Q \in \{\textit{couplings of } Z_t, \mu\}} \mathbb{E}_{(z_t, z) \sim Q} \|z_t - z\| \\ \leq (\epsilon + e^{-C_2 n}) \|z^*\|,\end{aligned}$$

provided that $C_3\epsilon^2 \leq \eta \leq 1000 C_3 \epsilon^2$, that $\beta = C_4 n$, that $\|z_0\| \leq 1000\|z^*\|$, that $W^{(i)}$ and A satisfy conditions WDC and RRIC with constant $\mathrm{poly}(\epsilon)$ and that $d \geq 2$. (above, 1000 can be replaced by any other constant)

B. Proof

B.1. Overview

We start by some preliminaries and definitions in Sections B.2, B.3 and B.4. Then, we analyze the loss function in Section B.5. Next, we show that with high probability, the norm of the iterates will neither be very small nor very large, in Section B.6 and Section B.7. Afterwards, we define a modified loss function, that is changed around the origin, in Section B.8. This change is necessary because the original loss function is not smooth next to the origin, and it is significantly simpler to analyze smooth losses. Since the norm of the iterate is not small with high probability, this change will not be apparent in the dynamics. Later, we would like to argue that the iterates will converge to some region around the optimum. In Section B.9 we argue that in order to show that the iterates converge to some region, it is sufficient to construct a potential function whose Laplacian is negative outside this region. Then, in Section B.10 we define such potential, concluding that the iterates converge to some region around the optimum. In that region, the function is convex. In Section B.11 we show that if the function is convex, then we can couple the continuous and discrete iterations such that they get closer and closer. In order to apply this argument, we have to guarantee that the iterates do not leave the convexity region. Indeed, in Section B.12 we show that if the iterates are already in a convexity region, they will stay there, enabling them to get closer, until they are ϵ-apart.

B.2. Preliminaries on polar coordinates

We start with some preliminaries. Denote by $\vec{\cdot}$ any vector from $\mathbb{R}^n$. Denote $e_1 = (1, 0, \ldots, 0)$. Given any $\vec{x}$, denote by $r = r(\vec{x}) = \|\vec{x}\|_2$, and denote by

$$\theta = \theta(\vec{x}) = \arccos\left(\frac{\langle x, e_1\rangle}{\|\vec{x}\|\|e_1\|}\right) = \arccos\left(\frac{\langle \vec{x}, e_1\rangle}{\|x\|}\right)$$

the angle $\theta \in [0, \pi]$ between $\vec{x}$ and e_1. Denote

$$\vec{r} = \vec{r}(\vec{x}) = \nabla_{\vec{x}} r(\vec{x}) = \frac{\vec{x}}{\|\vec{x}\|}$$

and

$$\vec{\theta} = \vec{\theta}(\vec{x}) = \frac{\nabla_{\vec{x}}\theta(\vec{x})}{\|\nabla_{\vec{x}}\theta(\vec{x})\|} = \frac{\cos(\theta(\vec{x}))}{\sin(\theta(\vec{x}))}\frac{\vec{x}}{\|\vec{x}\|} - \frac{e_1}{\sin(\theta(\vec{x}))} \ .$$

We will use $\theta, r, \vec{\theta}, \vec{r}$ without writing $\vec{x}$ when $\vec{x}$ is clear from context. We have the following properties:

Lemma B.1. *Let $\vec{x} \in \mathbb{R}^n$, denote by $V^\perp$ the vector space that is the orthogonal complement to $\mathrm{span}(\vec{\theta}(\vec{x}), \vec{r}(\vec{x}))$ as a subspace of $\mathbb{R}^n$ and let $\vec{\psi}_1, \ldots, \vec{\psi}_{n-2}$ denote an orthonormal basis of $V^\perp$. Then:*

1. *The set $\{\vec{r}(\vec{x}), \vec{\theta}(\vec{x}), \vec{\psi}_1, \ldots, \vec{\psi}_{n-2}\}$ forms an orthonormal basis to $\mathbb{R}^n$. In particular, $\|\vec{r}(\vec{x})\| = \|\vec{\theta}(\vec{x})\| = 1$ and $\langle \vec{r}(\vec{x}), \vec{\theta}(\vec{x})\rangle = 0$.*

2. *Let $f\colon [0,\infty)\times[0,\pi]\to\mathbb{R}$ be a C^2 function. Denote*

$$f_r=\frac{\partial f}{\partial r},\quad f_\theta=\frac{\partial f}{\partial\theta},\quad f_{rr}=\frac{\partial^2 f}{\partial r^2},\quad f_{\theta\theta}=\frac{\partial^2 f}{\partial\theta^2},\quad f_{r\theta}=f_{\theta r}=\frac{\partial^2 f}{\partial r\partial\theta}.$$

Denote $r=r(\vec{x})$, $\theta=\theta(\vec{x})$, $\vec{r}=\vec{r}(\vec{x})$, $\vec{\theta}=\vec{\theta}(\vec{x})$, $f_r=f_r(r,\theta)$, $f_\theta(r,\theta)$ etc. Then,

$$\nabla_{\vec{x}} f(r(\vec{x}),\theta(\vec{x}))=f_r\vec{r}+\frac{f_\theta}{r}\vec{\theta}.$$

Further, the Hessian of $f(r(\vec{x}),\theta(\vec{x}))$ with respect to $\vec{x}$ equals

$$\nabla^2_{\vec{x}} f(r(\vec{x}),\theta(\vec{x}))=f_{rr}\vec{r}\vec{r}^\top+\left(\frac{f_r}{r}+\frac{f_{\theta\theta}}{r^2}\right)\vec{\theta}\vec{\theta}^\top+\left(\frac{f_{r\theta}}{r}-\frac{f_\theta}{r^2}\right)\left(\vec{r}\vec{\theta}^\top+\vec{\theta}\vec{r}^\top\right)+\left(\frac{f_r}{r}+\frac{f_\theta}{r^2\tan(\theta)}\right)\left(\sum_{i=1}^{n-2}\vec{\psi}_i\vec{\psi}_i^\top\right)\ .$$

3. *It holds that*

$$\triangle f=\sum_{i=1}^{n}\frac{d^2 f}{dx_i^2}=f_{rr}+\left(\frac{f_r}{r}+\frac{f_{\theta\theta}}{r^2}\right)+(n-2)\left(\frac{f_r}{r}+\frac{f_\theta}{r^2\tan(\theta)}\right)=f_{rr}+(n-1)\frac{f_r}{r}+\frac{f_{\theta\theta}}{r^2}+(n-2)\frac{f_\theta}{r^2\tan\theta}\ .$$

4. *Assume that for all $\vec{x}$,*

$$\max\left(|f_{rr}|,\left|\frac{f_r}{r}+\frac{f_{\theta\theta}}{r^2}\right|,\left|\frac{f_{r\theta}}{r}-\frac{f_\theta}{r^2}\right|,\left|\frac{f_r}{r}+\frac{f_\theta}{r^2\tan(\theta)}\right|\right)\le s/2\ .$$

Then, f is s-smooth.

Proof. The first two items are folklore, and follow from a simple application of the chain rule. The third item follows from the fact that for any orthonormal basis $B=\{\vec{u}_1,\dots\vec{u}_n\}$ of $\mathbb{R}^n$, $\triangle f$ equals the trace of the Hessian of f, computed with respect to the basis B. In particular, the entries of the Hessian with respect to the basis $\{\vec{r},\vec{\theta},\vec{\psi}_1,\dots,\vec{\psi}_{n-2}\}$ are computed in item 2 and the trace equals the formula in item 3, as required. For the forth item, it holds that f is s-smooth if the spectral norm of $\nabla^2 f$ is bounded by s, while the Hessian $\nabla^2 f$ can be computed with respect to any orthonormal basis. We will write the Hessian with respect to the basis defined in item 1, and the Hessian's coefficients are computed in item 2. Since the Hessian is symmetric, its spectral norm is bounded by the ∞-norm, which is the maximum over the rows of the sum of absolute values, namely, $\|A\|_\infty=\max_i\sum_j|A_{ij}|$. The infinity norm of the Hessian is bounded by s, using the formula computed in item 2 and using the assumption of item 4. □

B.3. Definitions of Langevin dynamics

Assuming some potential function $H\colon\mathbb{R}^n\to\mathbb{R}$ and parameters η,β. The Langevin dynamics can be defined by

$$\vec{x}_t=\vec{x}_{t-1}-\eta\nabla H(\vec{x})+\vec{z}_t$$

where $\vec{z}_t\sim N(\vec{0},\sigma^2 I_n)$, $\sigma^2:=2\eta/\beta$. It is known that in the limit where $\eta\to 0$ (and under some regularity assumptions) the distribution of $\vec{x}_t$ as $t\to\infty$ converges to

$$\mu_{\beta H}(\vec{x}):=\frac{e^{-\beta H(\vec{x})}}{\int_{\mathbb{R}^n}e^{-\beta H(\vec{y})}d\vec{y}}\ .$$

Denote the distribution of $\vec{x}_t$ by $\mu_{\beta H,t}$.

B.4. Setting

We will use the same network as suggested by Huang et al. (2021) (multiplying the weights by 2 for convenience): given an input $\vec{x}$, the network is given by

$$G(\vec{x})=\mathrm{ReLU}(2W^d(\cdots(\mathrm{ReLU}(2W^1(\vec{x})))\cdots)).$$

A compressive map A is applied on the outcome, for obtaining an output of $AG(\vec{x})$. The goal is to recover some unknown $G(\vec{x}^*)$, given the measurement $AG(\vec{x}^*)$. We assume that each W^i satisfies the Weight Distribution Condition and that A satisfies Range Restricted Isometry Condition (Huang et al., 2021), both with parameter δ.[2] As shown by (Huang et al., 2021), this implies that

$$\forall \vec{x}, \vec{y}\colon |\angle(\mathrm{ReLU}(2W^i(\vec{x})), \mathrm{ReLU}(2W^i(\vec{y}))) - g(\angle(\vec{x}, \vec{y}))| \leq f(\delta),$$

where $f(\delta) \to 0$ as $\delta \to 0$, $\angle(\vec{x}, \vec{y}) \in [0, \pi]$ is the angle between $\vec{x}$ and $\vec{y}$ and

$$g(\theta) = \arccos\left(\frac{(\pi - \theta)\cos\theta + \sin\theta}{\pi}\right) .$$

Further,

$$|\|\mathrm{ReLU}(2W^i(\vec{x}))\| - \|\vec{x}\|| \leq f(\delta)\|\vec{x}\|$$

and

$$\forall \vec{x}\colon |\|A\vec{x}\| - \|\vec{x}\|| \leq f(\delta)\|\vec{x}\|.$$

The loss function is

$$\tilde{L}(\vec{x}) = \|AG(\vec{x}) - AG(\vec{x}^*)\|^2/2.$$

Let us compute the loss function assuming that $\delta = 0$. There, $f(\delta) = 0$ and further

$$\|AG(\vec{x}) - AG(\vec{x}^*)\|^2/2 = \|G(\vec{x}) - G(\vec{x}^*)\|^2/2.$$

Additionally, $\delta = 0$ implies that $\|G\vec{x}^*\| = \|\vec{x}^*\|$, $\|G\vec{x}\| = \|\vec{x}\|$ for all $\vec{x}$ and $\theta_d(\vec{x}) := \angle(G(\vec{x}), G(\vec{x}^*)) = g^{\circ d}(\angle(\vec{x}, \vec{x}^*))$ where $g^{\circ d}$ is a composition of g for d iterations. Assuming that $\vec{x}^* = (1, 0, \ldots, 0)$, and denoting by $\theta(\vec{x}) = \angle(\vec{x}, \vec{x}^*)$ we have that

$$\|G(\vec{x}) - G(\vec{x}^*)\|^2 = (\|\vec{x}\|\cos\theta_d(\vec{x}) - 1)^2 + (\|\vec{x}\|\sin\theta_d(\vec{x}))^2 = \|\vec{x}\|^2 - 2\|x\|\cos\theta_d(\vec{x}) + 1.$$

Denote by $L(\vec{x})$ the value of the loss function $\tilde{L}(\vec{x})$ when $\delta = 0$. As computed above,

$$L(\vec{x}) = \|\vec{x}\|^2/2 - \|x\|\cos\theta_d(\vec{x}) + 1/2.$$

Huang et al. (2021) have shown that the gradients of L are close to the gradients of $\tilde{L}$, under the above assumptions on the weights, in the following sense:

$$\forall \vec{x}\colon \|\nabla L(\vec{x}) - \nabla\tilde{L}(\vec{x})\| \leq (\|\vec{x}\| + 1)f(\delta, d), \tag{6}$$

for some $f(\delta, d)$ that decays to zero as $\delta \to 0$ while keeping d fixed.

It is sufficient to assume that $\vec{x}^* = e_1$, since Langevin dynamics is indifferent to scaling and rotations. Yet, once we consider $\vec{x}^*$ such that $\|\vec{x}^*\| \neq 1$, the error has to be multiplied by $\|\vec{x}^*\|$.

Notation. When using $O(\cdot)$-notation, we will ignore constants that depend on the depth d of the network. Given some parameter, e.g. $l > 0$, we denote by $C(l)$ a constant that may depend only on l (and perhaps also on the depth d), but not on the other parameters, in particular, not on n. We will use C, c to denote positive constants that depend only on d (and perhaps on other parameters that are explicitly defined as constants).

B.5. Properties of the loss function

Define by $\theta'_d = \frac{dg^{\circ d}(\theta)}{d\theta}$. And similarly define θ''_d as the second derivative. The loss function $L(\vec{x})$ can be computed as a function of $r = r(\vec{x})$ and $\theta = \theta(\vec{x})$, by the formula

$$L(r, \theta) = r^2/2 - r\cos\left(g^{\circ d}(\theta)\right) + 1/2 = r^2/2 - r\cos\theta_d + 1/2,$$

and the corresponding derivatives of L as a function of r and θ equal:

[2] The weight Distribution Condition holds with high probability for isotropic Gaussian matrices with constant expansion: namely, when the output dimension of each layer is at least a constant times larger than the input dimension. The required expansion constant depends on δ. Further, the Range Restricted Isometry Condition holds with high probability for matrices A with a constant output dimension, where the constant depends on δ (Huang et al., 2021; Daskalakis et al., 2020b).

- $L_r = r - \cos\theta_d$
- $L_\theta = r\sin\theta_d\theta'_d$
- $L_{rr} = 1$
- $L_{\theta\theta} = r\cos\theta_d(\theta'_d)^2 + r\sin\theta_d\theta''_d$
- $L_{r\theta} = \sin\theta_d\theta'_d$

Consequently, we have by Lemma B.1:

$$\nabla L(\vec{x}) = (r - \cos\theta_d)\vec{r} + \sin\theta_d\theta'_d\vec{\theta},$$

and

$$\nabla^2 L = \vec{r}\vec{r}^\top + \frac{r - \cos\theta_d + \cos\theta_d(\theta'_d)^2 + \sin\theta_d\theta''_d}{r}\vec{\theta}\vec{\theta}^\top + \sum_{i=1}^{n-2}\frac{(r-\cos\theta_d)\sin\theta + \sin\theta_d\theta'_d\cos\theta}{r\sin\theta}\vec{\psi}_i\vec{\psi}_i^\top \quad . \tag{7}$$

Before proceeding, let us analyze the function g and consequently, θ'_d and θ''_d:

Lemma B.2. *Let g' and g'' denote the first and second derivatives of g. Then, g is decreasing, $g'(\theta) \in [0,1]$ and $g''(\theta) \le 0$. Consequently, $\theta'_d \in [0,1]$ while $\theta''_d \le 0$.*

Proof. The properties of g follow directly by computing the derivatives of g and they were analyzed by Huang et al. (2021). The derivative of θ_d can be computed using the composition rule:

$$\theta'_d = \frac{dg^{\circ d}(\theta)}{d\theta} = \prod_{i=0}^{d-1} g'(g^{\circ i}(\theta)) \in [0,1]. \tag{8}$$

Notice that θ''_d is the derivative of (8) and it is non-positive as $g''(\theta) \le 0$. □

We have the following:

Lemma B.3. *Let $0 < r < R$. Then the loss L is $C(r,R)$-Lipschitz in $K = \{\|x\| : r \le \|\vec{x}\| \le R\}$ and $C(r)$-smooth in $K' = \{\|x\| : r \le \|\vec{x}\|\}$.*

Proof. First of all, we prove that the function is Lipschitz and smooth in K, and then we extend the smoothness to K'. It is sufficient to use Lemma B.1, and argue that the coefficients in the expansion of ∇L and $\nabla^2 L$ are bounded in absolute value in K. First, it is easy to verify that the two derivatives of g are finite, which implies that θ'_d, θ''_d are bounded. Further, r is bounded from below by assumption, hence, the only coefficient that could possibly go to infinity is

$$\frac{(r-\cos\theta_d)\sin\theta + \sin\theta_d\theta'_d\cos\theta}{r\sin\theta} = \frac{(r-\cos\theta_d)}{r} + \frac{\sin\theta_d\theta'_d\cos\theta}{r\sin\theta}.$$

In fact, the quantity that can possibly go to infinity in K is

$$\frac{\sin\theta_d\theta'_d\cos\theta}{r\sin\theta}. \tag{9}$$

since it has $\sin\theta$ in its denominator and $\sin\theta$ can be zero. Yet we would like to argue that when the denominator goes to zero, the numerator goes to 0 as well and the ratio does not go to infinity. Notice that the numerator contains the term $\theta_d = g^{\circ d}(\theta)$. Since $g(\theta) \le \theta$ (Huang et al., 2021) we derive that $\theta_d = g^{\circ d}(\theta) \le \theta$. If $\theta \le \pi/2$ then we have $0 \le \sin\theta_d \le \sin\theta$ which implies that the ratio in (9) is bounded. Otherwise, $\theta \ge \pi/2$ and the denominator can go to 0 only when $\theta \to \pi$. We would like to argue that the numerator also goes to 0 in this case. Indeed,

$$\theta'_d = \frac{dg^{\circ d}(\theta)}{d\theta} = \prod_{i=0}^{d-1} g'(g^{\circ i}(\theta)) \le g'(\theta),$$

since $0 \leq g' \leq 1$ (Huang et al., 2021), where $g'(\theta) = dg(\theta)/d\theta$. Further, $g'(\theta) \to 0$ as $\theta \to \pi$. Hence, by L'Hopital's rule,

$$\limsup_{\theta \to \pi} \frac{\theta'_d}{\sin\theta} = \limsup_{\theta \to \pi} \frac{\theta''_d}{\cos\theta} < \infty,$$

using the fact that θ''_d as argued above. By continuity, $\theta'_d / \sin\theta$ is uniformly bounded in $[0, \pi]$, which implies that (9) is uniformly bounded, as required.

Notice that the smoothness holds also over all of K', since, from the form of the second derivative and the arguments above, it is clear that these do not go to ∞ as $r \to \infty$. □

Further, we use the following lemma from Huang et al. (2021):

Lemma B.4. *There exist only three points where the gradient of L possibly equals 0: at the optimum $\vec{x}^*$, at $\vec{x} = -\vec{x}^* \cos g^{\circ d}(\pi)$, and at 0.*

We note the at 0 there is a local max, at $\vec{x}^*$ a local min and at $-\vec{x}^* \cos g^{\circ d}(\pi)$ a saddle point, that is a minimum with respect to r and a maximum with respect to θ.

We add the following lemma:

Lemma B.5. *Let $l > 0$. There exists a constant $C(l) > 0$ (independent of n) such that if $\vec{x}$ is at least l-far apart from any point where the gradient vanishes (in l_2-norm), then $\|\nabla L(\vec{x})\| \geq C(l)$.*

Proof. Notice first that as $\|\vec{x}\| \to \infty$ then the gradient-norm goes to infinity. Hence, it is sufficient to assume that $\|\vec{x}\| \leq M$ for some sufficiently large M. Secondly, notice that both the distance of $\vec{x}$ from any stationary point, and the norm of its gradient, are only functions of $r(\vec{x}) = \|\vec{x}\|$ and $\theta(\vec{x})$. Let K be the set of pairs (r, θ) such that (1) $r \leq M$ and (2) (r, θ) signify a point of distance at least l from any stationary point. This set is compact, hence the continuous function $(r, \theta) \to \|\nabla L(r, \theta)\|$ has a minimum in K, which is non-zero since we assumed that K does not contain any stationary point. □

Further, we have the following lemma:

Lemma B.6. *There exists some constant $l > 0$ such the function is 0.9-strongly convex in a ball of radius ℓ around $\vec{x}^*$.*

Proof. First of all, we will prove that the Hessian is PSD at $\vec{x}^*$. For this purpose, it is sufficient to prove that all the coefficients in (7) are positive at $\vec{x}^*$, since the basis $\vec{r}, \vec{\theta}, \vec{\psi}_1, \cdots, \vec{\psi}_{n-1}$ is orthonormal, as stated in Lemma B.1. The coefficient that multiplies $\vec{r}\vec{r}^\top$ is $1 > 0$. The second coefficient is

$$\frac{r - \cos\theta_d + \cos\theta_d(\theta'_d)^2 + \sin\theta_d\theta''_d}{r}.$$

We have that $r(\vec{x}^*) = 1$, $\theta(\vec{x}^*) = 0$, and $\theta_d(\vec{x}^*) = g^{\circ d}(\theta(\vec{x}^*)) = g^{\circ d}(0) = 0$ since $g(0) = 0$. Further, $\theta'_d(\vec{x}^*) = 1$ as $\frac{dg(\theta)}{d\theta}\big|_{\theta=0} = 1$. Hence,

$$\frac{r - \cos\theta_d + \cos\theta_d(\theta'_d)^2 + \sin\theta_d\theta''_d}{r} = \frac{1 - 1 + 1 + 0}{1} = 1.$$

For the last coefficient in (7), we have

$$\frac{(r - \cos\theta_d)\sin\theta + \sin\theta_d\theta'_d\cos\theta}{r\sin\theta} = \frac{(r - \cos\theta_d)}{r} + \frac{\sin\theta_d\theta'_d\cos\theta}{r\sin\theta}. \tag{10}$$

The first term is 0, while the second term is undefined, yet, can be computed using the limit $\theta \to 0$. In particular, using the calculations above and L-Hopital's rule,

$$\lim_{\theta \to 0} \frac{\sin\theta_d}{\sin\theta} = \lim_{\theta \to 0} \frac{\cos\theta_d\theta'_d}{\cos\theta} = 1.$$

In particular, the second term in (10) equals 1, hence $\nabla^2 L(\vec{x}^*) = I_n$. The minimal eigenvalue of the Hessian at $\vec{x}$, which is the minimal of the three coefficients in (7), is a function only of $r(\vec{x})$ and $\theta(\vec{x})$. By continuity, there exists a neighborhood $U \subseteq [0, \infty) \times [0, 2\pi]$ or pairs (r, θ), that contains the point $(r, \theta) = (1, 0)$, such that $\nabla^2(\vec{x}) \succeq 0.9 I_n$, for any $\vec{x}$ such that $(r(\vec{x}), \theta(\vec{x})) \in U$. This proves the result. □

Lastly, let us analyze $\triangle L(\vec{x})$. This will be useful later in the proof.

Lemma B.7.

$$\triangle L \leq \begin{cases} 2 + (n-2)(r - \cos\theta_d)/r & \theta \geq \pi/2 \\ n & \theta \leq \pi/2 \end{cases}$$

Proof. Using Lemma B.1, we have that,

$$\triangle L = 1 + \frac{r - \cos\theta_d + \cos\theta_d(\theta_d')^2 + \sin\theta_d\theta_d''}{r} + (n-2)\frac{(r-\cos\theta_d)\sin\theta + \sin\theta_d\theta_d'\cos\theta}{r\sin\theta} \quad . \tag{11}$$

First,

$$\frac{r - \cos\theta_d + \cos\theta_d(\theta_d')^2 + \sin\theta_d\theta_d''}{r} \leq 1, \tag{12}$$

since $-\cos\theta_d + \cos\theta_d(\theta_d')^2 \leq 0$, and $\sin\theta_d\theta_d'' \leq 0$ (as follows from the fact that $g'(\theta) \in [0,1]$ hence $\theta_d' = \frac{dg^{\circ d}(\theta)}{d\theta} \in [0,1]$, and $g''(\theta) \leq 0$ hence $\theta_d'' \leq 0$). To bound the last term in (11), first assume that $\theta \geq \pi/2$. Then,

$$\frac{(r-\cos\theta_d)\sin\theta + \sin\theta_d\theta_d'\cos\theta}{r\sin\theta} \leq \frac{r-\cos\theta_d}{r}$$

since $\sin\theta_d \geq 0$, $\theta_d' \geq 0$ and $\cos\theta \leq 0$. We conclude that

$$\Delta L \leq 2 + (n-2)\frac{r-\cos\theta_d}{r}.$$

Next, for $\theta \leq \pi/2$, we have that

$$\frac{(r-\cos\theta_d)\sin\theta + \sin\theta_d\theta_d'\cos\theta}{r\sin\theta} \leq \frac{r-\cos\theta_d+\cos\theta}{r} \leq 1,$$

using $0 \leq \sin\theta_d \leq \sin\theta$, $\theta_d' \in [0,1]$ and $0 \leq \cos\theta \leq \cos\theta_d$. This concludes the proof, in combination with (11) and (12). □

B.6. Escaping from the origin

Since the loss function is not well behaved around the origin, we want to show that the dynamics do not approach the origin with high probability, as stated below:

Lemma B.8. *Fix $t \geq 3/\eta$, define $A = \cos g^{\circ d}(\pi)$, and assume that we run the Langevin dynamics according to $\tilde{L}$. Define $\beta = 2\eta/\sigma^2$, as in Section B.3. Then, for any $a > 0$,*

$$\Pr\left[\|\vec{x}_t\| < 0.9A - a\right] \leq e^{-\beta a^2/4}.$$

The remainder of this subsection is devoted for the proof of this Lemma. Let us write the update rule:

$$\begin{aligned}\vec{x}_{t+1} &= \vec{x}_t - \eta\nabla\tilde{L}(\vec{x}) + \vec{z}_{t+1} = \vec{x}_t - \eta\nabla L(\vec{x}) - \eta(\nabla\tilde{L}(\vec{x}) - \nabla L(\vec{x})) + \vec{z}_{t+1} \\ &= \vec{x}_t - \eta(r-\cos\theta_d)\vec{r} - \eta\sin\theta_d\theta_d'\vec{\theta} + \vec{z}_{t+1} - \eta(\nabla\tilde{L}(\vec{x}) - \nabla L(\vec{x}))\end{aligned}$$

where $r, \vec{r}, \vec{\theta}$ etc. refer to $\vec{x}_t$ (as defined in Section B.2). We will define the following intermediate random variables, that help us transferring from $\vec{x}_t$ to $\vec{x}_{t+1}$:

$$\vec{x}_t' = \vec{x}_t - \eta(r-\cos\theta_d)\vec{r} = r\vec{r} - \eta(r-\cos\theta_d)\vec{r} = (r - \eta r + \eta\cos\theta_d)\vec{r},$$

$$\vec{x}_t'' = \vec{x}_t' - \eta\sin\theta_d\theta_d'\vec{\theta},$$

$$\vec{x}_t''' = \vec{x}_t'' + \langle\vec{z}_{t+1}, \vec{r}(\vec{x}_t'')\rangle\vec{r}(\vec{x}_t''),$$

$$\vec{x}_t^{(4)} = \vec{x}_t''' + \vec{z}_{t+1} - \langle \vec{z}_{t+1}, \vec{r}(\vec{x}_t'')\rangle \vec{r}(\vec{x}_t'')$$

and notice that

$$\vec{x}_{t+1} = \vec{x}_t^{(4)} - \eta(\nabla \tilde{L}(\vec{x}) - \nabla L(\vec{x})).$$

We will lower bound $\|\vec{x}_{t+1}\|$ as a function of x_t and of $\vec{z}_{t+1}$. First, we will lower bound the norms of these intermediate variables. Notice that

$$\|\vec{x}_t'\| = |r - \eta r + \eta \cos\theta_d| = (1-\eta)\|\vec{x}_t\| + \eta\cos\theta_d \geq (1-\eta)\|\vec{x}_t\| + \eta A,$$

where we use the fact by monotonicity of $g(\theta)$ (Lemma B.2),

$$\theta_d = g^{\circ d}(\theta) \leq g^{\circ d}(\pi) \Rightarrow \cos\theta_d \geq \cos g^{\circ d}(\pi) := A.$$

Further, notice that $\vec{x}_t'$ is a multiple of $\vec{r}$, and that $\vec{r}$ and $\vec{\theta}$ are orthogonal unit vectors by Lemma B.1, hence,

$$\|\vec{x}_t''\| = \sqrt{\|\vec{x}_t'\|^2 + \|\eta \sin\theta_d \theta_d'\|^2} \geq \|\vec{x}_t'\|.$$

Define $z_{t+1} = \langle \vec{z}_{t+1}, \vec{r}(\vec{x}_t'')\rangle$. Notice that

$$\|\vec{x}_t'''\| = \|\vec{x}_t'' + z_{t+1}\vec{r}(\vec{x}_t'')\| = \|\vec{r}(\vec{x}_t'')(\|\vec{x}_t''\| + z_{t+1})\| = |\|\vec{x}_t''\| + z_{t+1}| \geq \|\vec{x}_t''\| + z_{t+1} \ ,$$

using the fact that by Lemma B.1, $\|\vec{r}(\vec{x}_t'')\| = 1$. Recall that

$$\vec{x}_t^{(4)} = \vec{x}_t''' + \vec{z}_{t+1} - \langle \vec{z}_{t+1}, \vec{r}(\vec{x}_t'')\rangle \vec{r}(\vec{x}_t''),$$

and notice that $\vec{x}_t'''$ is a multiple of $\vec{r}(\vec{x}_t'')$ while $\vec{z}_{t+1} - \langle \vec{z}_{t+1}, \vec{r}(\vec{x}_t'')\rangle \vec{r}(\vec{x}_t'')$ is perpendicular to $\vec{r}(\vec{x}_t'')$ (namely, their inner product is 0). Hence,

$$\|\vec{x}_t^{(4)}\| = \sqrt{\|\vec{x}_t'''\|^2 + \|\vec{z}_{t+1} - \langle \vec{z}_{t+1}, \vec{r}(\vec{x}_t'')\rangle \vec{r}(\vec{x}_t'')\|^2} \geq \|\vec{x}_t'''\|.$$

Lastly, by the triangle inequality,

$$\|\vec{x}_{t+1}\| = \|\vec{x}_t^{(4)} - \eta(\nabla\tilde{L}(\vec{x}) - \nabla L(\vec{x}))\| \geq \|\vec{x}_t^{(4)}\| - \eta\|\nabla\tilde{L}(\vec{x}) - \nabla L(\vec{x})\| \geq \|\vec{x}_t^{(4)}\| - \eta(\|\vec{x}_t\| + 1) f(\delta, d),$$

where the last inequality follows from (6) and $f(\delta, d) \to 0$ as $\delta \to 0$. In particular,

$$\|\vec{x}_{t+1}\| \geq \|\vec{x}_t^{(4)}\| - \eta c(\|\vec{x}_t\| + 1),$$

where $c > 0$ can be chosen arbitrarily small, since δ can be chosen arbitrarily small (as assumed in this lemma). Combining all the above, we have

$$\begin{aligned}
\|\vec{x}_{t+1}\| &\geq \|\vec{x}_t^{(4)}\| - \eta c(\|\vec{x}_t\| + 1) \geq \|\vec{x}_t'''\| - \eta c(\|\vec{x}_t\| + 1) \geq \|\vec{x}_t''\| + z_{t+1} - \eta c(\|\vec{x}_t\| + 1) \\
&\geq \|\vec{x}_t'\| + z_{t+1} - \eta c(\|\vec{x}_t\| + 1) \geq (1-\eta)\|\vec{x}_t\| + \eta A + z_{t+1} - \eta c(\|\vec{x}_t\| + 1) \\
&= (1 - \eta - \eta c)\|\vec{x}_t\| + \eta(A - c) + z_{t+1}, \qquad (13)
\end{aligned}$$

where $z_{t+1} = \langle \vec{z}_{t+1}, \vec{r}(\vec{x}_t'')\rangle$. Since $\vec{z}_{t+1} \sim N(\vec{0}, \sigma^2 I)$ and since $\|\vec{r}(\vec{x}_t'')\| = 1$ (see Section B.2), it holds that $z_t \sim N(0, \sigma^2)$. Further, since $\vec{z}_1, \vec{z}_2, \ldots$ are i.i.d., then $z_1, z_2, \ldots$ are i.i.d. By expanding the recursive inequality in Section 13, we derive that

$$\|\vec{x}_t\| \geq (1 - \eta - \eta c)^t \|\vec{x}_0\| + \sum_{i=0}^{t-1} (1 - \eta - \eta c)^i \eta(A - c) + \sum_{i=0}^{t-1} (1 - \eta - \eta c)^i z_i. \qquad (14)$$

Let us lower bound the three terms above. The first term will be bounded by 0. The second term equals

$$\begin{aligned}
&\sum_{i=0}^{t-1} (1 - \eta - \eta c)^i \eta(A - c) = \frac{1 - (1 - \eta - \eta c)^t}{1 - (1 - \eta - \eta c)} \eta(A - c) \\
&\geq (1 - (1 - \eta - \eta c)^t)(A - c) \geq (1 - (1 - \eta)^t)(A - c) \geq (1 - e^{-\eta t})(A - c) \geq 0.95(A - c) \geq 0.9A.
\end{aligned}$$

Here, we used that $(1-\eta)^t \le e^{-\eta t} \le e^{-3} \le 0.05$, since $1-x \le e^{-x}$ for all $x \in \mathbb{R}$ and due to the assumption that $t \ge 3/\eta$, and further, we used the fact that $c > 0$ can be chosen arbitrarily small to bound $0.95(A-c) \ge 0.9A$. It remains to bound the third term in the expansion of (14), which is a Gaussian random variable, with zero mean and its variance can be computed as:

$$\mathrm{Var}\left(\sum_{i=0}^{t-1}(1-\eta-\eta c)^i z_t\right) = \sum_{i=0}^{t-1}\mathrm{Var}((1-\eta-\eta c)^i z_t) = \sum_{i=0}^{t-1}(1-\eta-\eta c)^{2i}\sigma^2$$
$$\le \sum_{i=0}^{t-1}(1-\eta)^{2i}\sigma^2 \le \sum_{i=0}^{\infty}(1-\eta)^{2i}\sigma^2 = \frac{\sigma^2}{1-(1-\eta)^2} = \frac{\sigma^2}{2\eta-\eta^2} \le \frac{\sigma^2}{\eta} = \frac{2}{\beta},$$

recalling that we defined $\beta = 2\eta/\sigma^2$. Denote by z the random variable corresponding to the third term of (14), then we have just shown that $\mathrm{Var}(z) \le 2/\beta$ and that $\|\vec{x}_t\| \ge 0.9A - z$. In order to conclude the proof, it is sufficient to bound $\Pr[z > a]$ for any $a > 0$. From standard concentration inequalities for Gaussians, we know that for any a,

$$\Pr[z > a] \le e^{-a^2/2\mathrm{Var}(z)} \le e^{-a^2\beta/4},$$

as required.

B.7. The norm is bounded from above

Here we prove the following proposition:

Proposition B.9. *For any $t \ge 0$,*

$$\Pr[\|\vec{x}_t\| \ge (1-\eta/2)^t\|\vec{x}_0\| + C + C\sqrt{n/\beta}] \le e^{-n/C}$$

for some universal $C > 0$.

We write the gradient of the loss as

$$\nabla\tilde{L}(\vec{x}) = \nabla L(\vec{x}) + (\nabla\tilde{L}(x) - \nabla L(\vec{x})) = (\|\vec{x}\| - \cos\theta_d(\vec{x}))\vec{r}(\vec{x}) + \sin\theta_d(\vec{x})\theta_d'(\vec{x})\vec{\theta}(\vec{x}) + (\nabla\tilde{L}(x) - \nabla L(\vec{x}))$$

and the Langevin step is

$$\vec{x}_t = \vec{x}_{t-1} - \eta\nabla\tilde{L}(\vec{x}) + \vec{z}_t$$

where $\vec{z}_t \sim N(\vec{0}, \sigma^2 I_n)$. We have

$$\vec{x}_t = (1-\eta)\vec{x}_{t-1} + \eta\vec{A}_t + \eta\vec{B}_t + \vec{z}_t,$$

where

$$\vec{A}_t := \cos\theta_d\vec{r} + \sin\theta_d\theta_d'\vec{\theta}, \quad \vec{B}_t = \nabla\tilde{L}(x) - \nabla L(\vec{x}).$$

Notice that

$$\|\vec{A}_t\|^2 \le \cos^2\theta_d + \sin^2\theta_d(\theta_d')^2 \le \cos^2\theta_d + \sin^2\theta_d \le 1$$

where we used that $\theta_d' \le 1$ (this follows from Lemma B.2). Further, by (6),

$$\|\vec{B}_t\| \le (\|\vec{x}_{t-1}\| + 1)f(\delta, d),$$

where $f(\delta, d) \to 0$ as $\delta \to 0$. In particular, since we assume that δ can be chosen arbitrarily small, we can assume that $f(\delta, d) \le c$ for some arbitrarily small constant $c > 0$.

Expanding on the definition of $\vec{x}_t$, we have

$$\vec{x}_t = (1-\eta)^t\vec{x}_0 + \sum_{i=1}^{t}(1-\eta)^{t-i}\vec{z}_i + \eta\sum_{i=1}^{t}(\vec{A}_i + \vec{B}_i)(1-\eta)^{t-i}.$$

Decompose $\vec{x}_t = \vec{y}_t + \vec{w}_t$ as follows:

$$\vec{y}_t = \sum_{i=1}^{t}(1-\eta)^{t-i}\vec{z}_i, \qquad \vec{w}_t = (1-\eta)^t\vec{x}_0 + \eta\sum_{i=1}^{t}(\vec{A}_i + \vec{B}_i)(1-\eta)^{t-i}.$$

Notice that

$$\vec{w}_t = (1-\eta)\vec{w}_{t-1} + \eta(\vec{A}_t + \vec{B}_t).$$

We have that

$$\|\vec{w}_t\| \leq (1-\eta)\|\vec{w}_{t-1}\| + \eta(\|\vec{A}_t\| + \|\vec{B}_t\|) \leq (1-\eta)\|\vec{w}_{t-1}\| + \eta + \eta c(\|\vec{x}_{t-1}\| + 1)$$
$$\leq (1-\eta)\|\vec{w}_{t-1}\| + \eta + \eta c(\|\vec{y}_{t-1}\| + \|\vec{w}_{t-1}\| + 1) \leq (1-\eta+\eta c)\|\vec{w}_{t-1}\| + \eta c\|\vec{y}_{t-1}\| + \eta(1+c),$$

and $\|\vec{w}_0\| = \|\vec{x}_0\|$. By expanding on this, we have that

$$\|\vec{w}_t\| \leq (1-\eta+\eta c)^t\|\vec{x}_0\| + \eta c\sum_{i=1}^{t-1}(1-\eta+\eta c)^{t-1-i}\|\vec{y}_i\| + \eta(1+c)\sum_{i=1}^{t}(1-\eta+\eta c)^{t-i}.$$

Hence,

$$\|\vec{x}_t\| \leq \|\vec{y}_t\| + \|\vec{w}_t\| \leq (1-\eta+\eta c)^t\|\vec{x}_0\| + \eta c\sum_{i=1}^{t}(1-\eta+\eta c)^{t-i}\|\vec{y}_i\| + \|\vec{y}_t\| + \eta(1+c)\sum_{i=1}^{t}(1-\eta+\eta c)^{t-i}.$$

Assuming that $c \leq 1/2$, we have

$$\eta(1+c)\sum_{i=1}^{t}(1-\eta+\eta c)^{t-i} \leq 1.5\eta\sum_{i=0}^{\infty}(1-\eta/2)^i = \frac{1.5\eta}{1-(1-\eta/2)} = 3.$$

Assuming again that $c \leq 1/2$, we have that

$$\|\vec{x}_t\| \leq (1-\eta/2)^t\|\vec{x}_0\| + \sum_{i=1}^{t}(1-\eta/2)^{t-i}\|\vec{y}_i\| + \|\vec{y}_t\| + 3.$$

Let us bound the term that corresponds to the $\vec{y}_i$, and notice that these are isotropic Gaussians with variance bounded as follows:

$$\mathrm{Var}(\vec{y}_t) = \mathrm{Var}\left(\sum_{i=1}^{t}(1-\eta)^{t-i}\vec{z}_i\right) = \sum_{i=1}^{t}\mathrm{Var}\left((1-\eta)^{t-i}\vec{z}_i\right) = \sum_{i=1}^{t}(1-\eta)^{2t-2i}\sigma^2 I \preceq \sigma^2\sum_{i=0}^{\infty}(1-\eta)^{2i}I$$
$$= \frac{\sigma^2}{1-(1-\eta)^2}I = \frac{\sigma^2}{2\eta-\eta^2}I \preceq \frac{\sigma^2 I}{\eta} = 2\beta I,$$

using the fact that $\eta \leq 1$ and recalling the definition of β from Section B.3. We will use the following definition of a sub-Gaussian random variable:

Definition B.10. A random variable X is L-subGaussian if $\mathbb{E}[\exp((X-\mathbb{E}x)/(2L))] \leq 2$.

We have the following properties of a subGaussian random variable (Vershynin, 2018):

Lemma B.11. *Let $\vec{X} \sim N(\vec{0}, \sigma^2 I)$. Then, $\|\vec{X}\|$ is an L-subGaussian for some universal constant $L > 0$.*

Lemma B.12. *If X is an L-subGaussian random variable than for any $t > 0$,*

$$\Pr[X \geq t] \leq \exp(-t^2/2CL)$$

for some universal constant $C > 0$.

Lemma B.13. *If $X_1, \ldots, X_n$ are L-subGaussian random variables, then $\sum_i \lambda_i X_i$ is $L\sqrt{\sum_i \lambda_i^2}$-subGaussian, hence it is $L\sum_i |\lambda_i|$-subGaussian.*

Since $\vec{y}_t$ is an isotropic random variable with variance bounded by 2β, we derive that $\|\vec{y}_t\|$ is $C\beta$ subGaussian for some $C > 0$. Further, we derive that

$$\eta\sum_{i=1}^{t}(1-\eta/2)^{t-i}\|\vec{y}_i\| + \|\vec{y}_t\|$$

is a subGaussian with parameter bounded by

$$C\beta\sum_{i=1}^{t}(1-\eta/2)^{t-i}+C\beta\leq C\beta\eta\sum_{i=1}^{\infty}(1-\eta/2)^{i}+C\beta=\frac{C\beta\eta}{\eta/2}+C\beta=3C\beta.$$

Further, let us compute:

$$\mathbb{E}\left[\eta\sum_{i=1}^{t}(1-\eta/2)^{t-i}\|\vec{y}_i\|+\|\vec{y}_t\|\right]\leq\eta\sum_{i=1}^{t}(1-\eta/2)^{t-i}\sqrt{\mathbb{E}[\|\vec{y}_i\|^2]}+\sqrt{\mathbb{E}[\|\vec{y}_t\|^2]}\leq\eta\sum_{i=1}^{t}(1-\eta/2)^{t-i}\sqrt{2\beta n}+\sqrt{2\beta n}$$
$$\leq\frac{\eta}{\eta/2}\sqrt{2\beta n}+\sqrt{2\beta n}\leq 3\sqrt{2\beta n}.$$

From Lemma B.11 we derive that for any $h\geq 0$

$$\Pr\left[\sum_{i=1}^{t}(1-\eta/2)^{t-i}\|\vec{y}_i\|+\|\vec{y}_t\|\geq 3\sqrt{2\beta n}+h\right]\leq\exp(-h^2/C'\beta),$$

for some universal constant $C'>0$. This implies that

$$\Pr\left[\|\vec{\vec{x}}_t\|\geq 3+(1-\eta/2)^t\|\vec{x}_0\|+3\sqrt{2\beta n}+h\right]\leq\exp(-h^2/(C'\beta)).$$

In particular, if we substitute $h=\sqrt{n/\beta}$, we get that

$$\Pr\left[\|\vec{\vec{x}}_t\|\geq 3+(1-\eta/2)^t\|\vec{x}_0\|+3\sqrt{2\beta n}\right]\leq\exp(-n/C'')$$

for some universal constant $C''>0$.

B.8. Defining a smooth loss function

One problem that arises with L is that it is not smooth around the origin. As we have shown, $\vec{x}$ does not approach the origin with high probability. Hence, it is sufficient to assume that the loss function is different around the origin. In particular, the dynamics will not reach a ball of radius $r_0:=\cos(g^{\circ d}(\pi))/2$ around the origin, w.h.p. We define a modified loss, $\hat{L}$, that is different in this ball. First, we define an auxiliary function, that is parameterized by $0\leq a<b$ and is a function of $r\geq 0$:

$$h^{a,b}(r)=\begin{cases}0 & r\leq a\\ 2(r-a)^2/(b-a)^2 & a\leq r\leq(a+b)/2\\ 1-2(b-r)^2/(b-a)^2 & (a+b)/2\leq r\leq b\\ 1 & r\geq b\end{cases}.$$

Notice that this function transitions smoothly from 0 to 1 in the interval $[a,b]$, it has a continuous first derivative and has a second derivative almost everywhere, with

$$\frac{dh^{ab}(r)}{dr}=h^{a,b}_r(r)=\begin{cases}0 & r\leq a\\ 4(r-a)/(b-a)^2 & a\leq r\leq(a+b)/2\\ 4(b-r)/(b-a)^2 & (a+b)/2\leq r\leq b\\ 0 & r\geq b\end{cases}$$

and

$$\frac{d^2h^{ab}(r)}{dr^2}=h^{a,b}_{rr}(r)=\begin{cases}0 & r<a\\ 4/(b-a)^2 & a<r<(a+b)/2\\ -4/(b-a)^2 & (a+b)/2<r<b\\ 0 & r>b\end{cases}.$$

We will define the following smoothed loss function function:

$$\hat{L}(r,\theta) = L(r,\theta)h^{r_0/3,2r_0/3}(r) + \xi(1 - h^{0,r_0}(r)),$$

for some parameter $\xi > 0$ to be determined. Denote $h^{r_0/3,2r_0/3} = h^1$ and $h^{0,r_0} = h^2$ for convenience. The derivatives of $\hat{L}$ can be computed as follows:

- $\hat{L}_r = L_r h^1 + L h^1_r - \xi h^2_r$.
- $\hat{L}_\theta = L_\theta h^1$
- $\hat{L}_{rr} = L_{rr} h^1 + 2L_r h^1_r + h^1_{rr} - \xi h^2_{rr}$
- $\hat{L}_{\theta\theta} = L_{\theta\theta} h^1$
- $\hat{L}_{r\theta} = L_{r\theta} h^1 + L_\theta h^1_r$

We conclude the following properties that the modified loss function satisfies everywhere, based on the properties computed above and the properties of L:

Lemma B.14. *Assume that ξ is a large enough universal constant. Then, the modified loss satisfies:*

- *$\hat{L}$ is $O(1)$ smooth everywhere. Further, $\triangle\hat{L} \le O(n)$.*
- *The critical points of $\hat{L}$ (those with zero derivative) are $0, \vec{x}^*$ and $-\cos g^{\circ d}(\pi)\vec{x}^*$. For any $l > 0$ there exists a constant $c(l) > 0$ such that any point whose distance from the critical points is at least l satisfies $\|\nabla\hat{L}(\vec{x})\| \ge c(l)$.*
- *At $\vec{x} \in \{\vec{x}\colon \|\vec{x}\| \le r_0/3\}$, $\triangle\hat{L}(\vec{x}) \le -\Omega(n)$.*

Proof. The smoothness in the ball $\{\vec{x}\colon \|\vec{x}\| \ge 2r_0/3\}$ follows from Lemma B.3, which argues that L is smooth in this region, due to the fact that $\hat{L} = L$ in this region. In the region $\{\vec{x}\colon \|\vec{x}\| \in \{r_0/3, 2r_0/3\}\}$ smoothness of $\hat{L}$ follows from the expression for the second derivative of $\hat{L}$, from Lemma B.1 and from the fact L, h^1, h^2 are smooth with bounded derivatives in this region. For $\{\vec{x}\colon \|\vec{x}\| \in \{0, r_0/3\}\}$ smoothness of $\hat{L}$ follows from the smoothness of h^2. Further, $|\triangle L| \le O(n)$ since any function f on n variables that is s-smooth satisfies $\triangle f \le sn$.

Next, we argue about the critical points of L. First, look at the region defined by $\|\vec{x}\| \ge r_0$, where, $\hat{L} = L$. In this region, the critical points of L are $\vec{x}^*$ and $-\cos(g^{\circ d}(\pi))\vec{x}^*$ and these are also the critical points of $\hat{L}$ in this region. Next, we study the region $\|\vec{x}\| \in [r_0/3, r_0]$. In this region, recall that

$$\hat{L}_r = L_r h^1 + L h^1_r - \xi h^2_r.$$

Now, the first two terms are bounded by a constant, using the calculations of the derivatives of L and of h^1. And the last term (which is being subtracted from the first two terms) is lower bounded by a constant times ξ. We can make the whole derivative negative by taking ξ to be a sufficiently large constant. In particular, in this region, the derivative with respect to r is nonzero, hence, by Lemma B.1, the gradient of $\hat{L}$ is nonzero. Lastly, for the region $r \in [0, r_0/3]$: Here, $\hat{L} = \xi(1 - h^2)$. By the derivative computation above, the only critical point is $\vec{0}$. In particular, this concludes that the critical points of $\hat{L}$ are $\vec{x}^*, -\cos(\theta^{\circ d}(\pi))\vec{x}^*$ and $\vec{0}$. Now, from continuity, for any $l > 0$ there exists $c(l, n) > 0$ such that any point $\vec{x}$ whose distance from any critical point is at least l, satisfies that its gradient norm is at least $c(l, n)$. Yet, notice that this constant can be taken independent of n. This is due to the fact that $\hat{L}(\vec{x})$ is only a function of $r(\vec{x})$ and $\theta(\vec{x})$, hence $\|\nabla\hat{L}(\vec{x})\|$ is as well, and there is no dimension dependence.

For the last item, notice that in the region $\|\vec{x}\| \le r_0/3$, $\hat{L}(\vec{x}) = \xi(1 - h^2(\vec{x}))$. By the computation of the second derivative, and by Lemma B.1, it follows that $\triangle\hat{L}(\vec{x}) \le -\Omega(\xi n) \le -\Omega(n)$ as required. □

B.9. Convergence assuming a potential function

In this section, we want to argue that certain potential functions decrease as a consequence of applying a Langevin step. We will use this in the future to prove that the iterations converge to a certain region where this potential is small.

Assume that there is a potential function $V\colon \mathbb{R}^d \to \mathbb{R}$. Further assume the Laplacian is defined as

$$\mathcal{L}V(\vec{x}) = \triangle V(\vec{x}) - \beta\langle \nabla H(\vec{x}), \nabla V(\vec{x})\rangle,$$

where $H(\vec{x})$ is the function that defines the Langevin dynamics as in Section B.3. We would like to show that if $\mathcal{L}V(\vec{x})$ is negative around $\vec{x}_{t-1}$ then $\mathbb{E}[V(\vec{x}_t) \mid \vec{x}_{t-1}] < V(\vec{x}_{t-1})$.

Lemma B.15. *Assume that the functions H, V are $O(1)$ smooth in $\mathbb{R}^n$, and that $\sigma^2 = 2\eta/\beta \le O(1/n)$. Assume that $\|\nabla H(\vec{x}_{t-1})\|, \|\nabla V(\vec{x}_{t-1})\| \le O(1)$. Let $-\kappa$ denote the maximum of $\mathcal{L}V$ in the ball of radius $r := 2\sqrt{n}\sigma = 2\sqrt{2n\eta/\beta}$ around $\vec{x}_{t-1}$, and assume that $\mathcal{L}(\vec{x})$ is bounded by $-\kappa + M$ in $\mathbb{R}^n$. Then,*

$$\mathbb{E}[V(\vec{x}_t) - V(\vec{x}_{t-1}) \mid \vec{x}_{t-1}] \le -\mu\kappa/\beta + e^{-cn}M/\beta + O(\eta\sqrt{\eta n/\beta}),$$

where $c > 0$ is a universal constant.

To prove the above, we use the following stochastic process:

$$\vec{y}_0 = \vec{x}_{t-1}; \quad d\vec{y}_s = -\eta\nabla H(\vec{y}_0)ds + \sqrt{2\eta/\beta}d\vec{B}_s \ .$$

Notice that $\vec{y}_1 \sim \vec{x}_t$ conditioned on $\vec{x}_{t-1}$. Let us assume that $\vec{x}_{t-1}$ is fixed for the calculations ahead. We would like to compute $\mathbb{E}V(\vec{y}_1)$. To do this, we can use It's formula, to derive that

$$\mathbb{E}[V(\vec{y}_1) - V(\vec{y}_0)] = \int_0^1 \mathbb{E}[-\eta\langle \nabla H(\vec{y}_0), \nabla V(\vec{y}_s)\rangle + \triangle V(\vec{y}_s)\eta/\beta]ds.$$

For a fixed s, the term under expectation equals

$$\frac{\eta}{\beta}\mathcal{L}V(\vec{y}_s) + \eta\langle \nabla H(\vec{y}_0) - \nabla H(\vec{y}_s), \nabla V(\vec{y}_s)\rangle.$$

Let us bound both terms in expectation. For the first term, we use the fact that since $\vec{y}_s \sim N(\vec{y}_0, \sigma^2 s I_n)$, $\Pr[\|\vec{y}_s - \vec{y}_0\| \ge 2\sqrt{\sigma s n}] \le e^{-cn}$. If the above does not hold, we Laplacian of $\vec{y}_s$ is at most $-\kappa$, and otherwise it is at most $-\kappa + M$, as assumed above. Hence,

$$\mathbb{E}[\mathcal{L}V(\vec{y}_s)] \le -\kappa(1 - e^{-cn}) + e^{-cn}(-\kappa + M) \le -\kappa + Me^{-cn}.$$

For the second term, we have, for some constant C,

$$\begin{aligned}
&\mathbb{E}\langle \nabla H(\vec{y}_0) - \nabla H(\vec{y}_s), \nabla V(\vec{y}_s)\rangle \le \mathbb{E}\|\nabla H(\vec{y}_0) - \nabla H(\vec{y}_s)\|\|\nabla V(\vec{y}_s)\| \\
&\le \mathbb{E}[\|\nabla H(\vec{y}_0) - \nabla H(\vec{y}_s)\|(\|\nabla V(\vec{y}_s) - \nabla V(\vec{y}_0)\| + \|\nabla V(\vec{y}_0)\|)] \le O(1)\mathbb{E}[\|\vec{y}_0 - \vec{y}_s\|(\|\vec{y}_s - \vec{y}_0\| + O(1))] \\
&\le O(1)\mathbb{E}\|\vec{y}_0 - \vec{y}_s\|^2 + O(1)\mathbb{E}\|\vec{y}_0 - \vec{y}_s\| \le O(\sqrt{\sigma^2 sn}) = O(\sqrt{\eta n/\beta}).
\end{aligned}$$

This completes the proof of the lemma above. As a consequence, we bound the number of times that it takes for the function to get to a region with positive value of $\mathcal{L}V$:

Lemma B.16. *Let $V \ge 0$ be an $O(1)$-smooth potential function, assume that H is $O(1)$ smooth, let $\kappa > 0$, define*

$$K = \{\vec{x}\colon \exists \vec{y},\ \|\vec{y} - \vec{x}\| \le 2\sqrt{n}\sigma,\ \mathcal{L}V(\vec{y}) > -\kappa\}.$$

Let $C_1 > 0$ and define

$$B = \{\vec{x}\colon \max(\|\nabla H(\vec{x})\|, \|\nabla V(\vec{x})\|) > C_1\}.$$

Assume that $\vec{x}_t$ is according to the Langevin dynamics with potential function H (see Section B.3). Let $\tau > 0$ be the first t such that $\vec{x}_t \in K \cup B$. Let M denote the maximum of $\mathcal{L}V$ over all $\mathbb{R}^n$. If $e^{-cn}M/\beta + O(\eta\sqrt{\eta n/\beta}) < \mu\kappa/2\beta$, then,

$$\mathbb{E}[\tau \mid \vec{x}_0] \le \frac{V(\vec{x}_0)}{\mu\kappa/\beta - (e^{-cn}M/\beta + O(\eta\sqrt{\eta n/\beta})} \le \frac{2V(\vec{x}_0)}{\mu\kappa/\beta}.$$

Proof. Denote $\Delta = \mu\kappa/\beta - (e^{-cn}M/\beta + O(\eta\sqrt{\eta n/\beta})$ If $\vec{x}_t \notin K \cup B$, we can apply Lemma B.15 to argue that $\mathbb{E}[V(\vec{x}_{t+1}) \mid \vec{x}_t] \le V(\vec{x}_t) - \Delta$. Since V cannot decrease below 0, the expected number of iterations that this happens is at most $V(\vec{x}_0)/\Delta$ as required. □

B.10. Defining a potential function

We would like to apply Lemma B.16 for the dynamics defined by the loss function $\hat{L}$. Notice that this function identifies with L except for some ball around 0. Define the potential function

$$V(\vec{x}) = \hat{L}(\vec{x}) - \lambda \cos(\theta) h^{r_0, 3r_0/2}(r) \mathbb{1}(\theta \geq \pi/2).$$

We prove the following:

Lemma B.17. *Assume that n is at least a sufficiently large constant. There exists some $\lambda = \Theta(1)$ such that the following holds. Let $l > 0$ be a constant. Then, there exist a constants $C, c > 0$ (depending possibly on l) such that for any $\beta \geq Cn$ and any $\vec{x}$ that satisfies $\|\vec{x} - \vec{x}^*\| \geq l$, we have that $\mathcal{L}V(\vec{x}) \leq -cn$. Further, $\mathcal{L}V \leq O(n)$ everywhere.*

Proof. For convenience, denote $h = h^{r_0, 3r_0/2}$. First of all, we explain how to set λ. For that purpose, recall that the Laplacian involves an inner product between the gradient of $\hat{L}$ and that of V, and we would like to make sure that this inner product is always non-positive (as it appears with a negative sign). Notice that

$$\langle \nabla \hat{L}(\vec{x}), \nabla V(\vec{x}) \rangle = \|\nabla \hat{L}(\vec{x})\|^2 + \lambda \langle \nabla \hat{L}(\vec{x}), \nabla - \cos\theta(\vec{x}) h^{r_0, 3r_0/2}(r(\vec{x})) \mathbb{1}(\theta(\vec{x}) \geq \pi/2) \rangle.$$

While the first term is always non-negative, we would like to make sure that the second term is not very negative. For that purpose, let us compute the gradient of the second term of the loss function, and notice that it is nonzero only if $\theta \geq \pi/2$ and $r \geq 3r_0/2$, and assume that we are in this region for convenience, and in particular, the indicator function $\mathbb{1}(\theta \geq \pi/2)$ can be replaced with 1. In order to compute the gradient, it is sufficient to compute the derivatives with respect to r and θ, as follows from Lemma B.1. We have the the derivative with respect to r equals

$$(-\cos(\theta) h(r))_r = -\cos(\theta) h_r(r),$$

and

$$(-\cos(\theta) h(r))_\theta = \sin\theta h(r).$$

Hence, the gradient equals

$$\nabla(-\cos(\theta) h(r)) = -\cos(\theta) h_r(r) \vec{r} + \frac{\sin\theta h(r)}{r} \vec{\theta}.$$

The inner product with the gradient of $\hat{L}$ equals

$$\langle (r - \cos\theta_d)\vec{r} + \sin\theta_d \theta_d' \vec{\theta}, -\cos(\theta) h_r(r) \vec{r} + \frac{\sin\theta h(r)}{r} \vec{\theta} \rangle = (r - \cos\theta_d)(-\cos(\theta) h_r(r)) + (\sin\theta_d \theta_d') \frac{\sin\theta h(r)}{r},$$

which follows since $\vec{r}, \vec{\theta}$ are orthonormal vectors, from Lemma B.1. The second term in the right hand side is nonnegative, since all the involved terms are positive, including θ_d' which is the derivative of $g^{\circ d}(\theta)$ that is nonnegative since g is increasing. Hence, we only have to take care of the first part. Notice that $h_r(r) = 0$ for any $r \geq 3r_0/2$, hence, we have to care only for $r \in [r_0, 3r_0/2]$. In this region, the norm of the gradient of $\hat{L}$ is at least some constant, since $\hat{L} = L$ in this region and due to Lemma B.5. Further, h_r is always bounded from above in an absolute value, as is $\cos\theta$ and r is bounded in this region, hence, we can set λ to be a constant such that for all $r \in [r_0, 3r_0/2]$,

$$\lambda \left\langle (r - \cos\theta_d)\vec{r} + \sin\theta_d \theta_d' \vec{\theta}, -\cos(\theta) h_r(r) \vec{r} + \frac{\sin\theta h(r)}{r} \vec{\theta} \right\rangle \geq (r - \cos\theta)(-\cos\theta h_r(r)) \geq -\|\nabla \hat{L}\|^2/2.$$

This concludes that

$$\langle \nabla V(\vec{x}), \nabla \hat{L}(\vec{x}) \rangle \geq \|\nabla \hat{L}\|^2/2, \tag{15}$$

for all $\vec{x}$, for a sufficiently small (but constant) λ.

Now, it is sufficient to prove that $\triangle V \leq -\Omega(n)$ in some balls of constant radius l' around $\vec{0}$ and the saddle point $-\vec{x}^* \cos g^{\circ d}(\pi)$. Indeed, assume that this is the case and let us conclude the proof. First of all, in the above two neighborhoods, we have that

$$\mathcal{L}V \leq \triangle V - \frac{\beta}{2} \|\nabla \hat{L}\| \leq \triangle V \leq -\Omega(n)$$

using (15). Next, we would analyze $\mathcal{L}V$ outside these regions and outside a ball of radius l around $\vec{x}^*$. Using Lemma B.14, $\|\nabla \hat{L}\| \geq \Omega(1)$ in these regions and $\triangle V = \triangle \hat{L} - \triangle \cos\theta h(r) \leq O(n)$, using Lemma B.14 to bound $\triangle \hat{L}$ and using the fact that $\cos\theta$ and $h(r)$ has bounded first and second derivatives to bound $\triangle \cos\theta h(r)$. In particular, using (15) we derive that

$$\mathcal{L}V = \triangle V - \frac{\beta}{2}\|\nabla \hat{L}\|^2 \leq O(n) - \Omega(\beta) \leq -\Omega(\beta) \leq -\Omega(n),$$

if $\beta \geq \Omega(n)$ for a sufficiently large constant. It remains to show that $\triangle V \leq -\Omega(n)$ in two balls of radius $l' > 0$ around 0 and around the saddle point. First of all, around 0 we have that $V = \hat{L}$, and using Lemma B.14 we have that $\triangle V \leq -\Omega(n)$. Secondly, let us look at the region around $-\vec{x}^* \cos g^{\circ d}(\pi)$. In that region $h = 1$, hence, using Lemma B.1 we can compute that

$$\triangle(-h(r)\cos(\theta)) = \triangle(-\cos\theta) = \frac{(n-1)\cos\theta}{r^2} \leq -\Omega(n),$$

around $-\vec{x}^* \cos g^{\circ d}(\pi)$, since $\theta(-\vec{x}^* \cos g^{\circ d}(\pi)) = \theta(-\vec{x}^*) = \pi$ and $\cos\pi = -1$. Next, from Lemma B.7, we have that around this point,

$$\triangle \hat{L} = \triangle L \leq 2 + \frac{(n-2)(r - \cos\theta_d)}{r}.$$

At the saddle point this equals 0. Yet, this can be positive if $r \geq \cos\theta$, however, it is bounded by $c(l') \cdot n$ in a ball of radius $l' > 0$ around the saddle point. By continuity, we can take $c(l')$ to zero as $l' \to 0$. Hence, if l' is taken as a sufficiently small constant, then $|\triangle \hat{L}| \leq |\triangle \lambda \cos\theta|/2$. In particular, we have that in a ball of radius l' around the saddle point,

$$\triangle V = \triangle L + \triangle \cos\theta \leq \frac{1}{2}\triangle \cos\theta \leq -\Omega(n),$$

as required.

Lastly, inside a ball of radius l around $\vec{x}^*$, we have that

$$\mathcal{L}V \leq \triangle V = \triangle \hat{L} = \triangle L \leq O(n),$$

where we used the computed bound on $\triangle L$ and the fact that V identifies with L around $\vec{x}^*$. □

For conclusion, let us bound the time that it takes the algorithm to get into the convexity region:

Lemma B.18. *Let $l > 0$ and assume that n is a sufficiently large constant. Assume that we run the dynamics according to the loss function L and let τ denote the first iteration such that $\|\vec{x}_t - \vec{x}^*\| \leq l$. Then, $\mathbb{E}\tau \leq O(1/\eta)$. Further, $\Pr[\tau \geq \Omega(\log(1/\epsilon)/\epsilon)] \leq \epsilon + e^{-cn}$.*

Proof. First, we will argue that if we run the dynamics according to L, the hitting hitting time is bounded by $O(1/\eta)$ in expectation. In order to prove that, we use the potential function V in combination with Lemma B.16. Recall that the set K defined in Lemma B.16 corresponds to the set of all points $\vec{x}$ where $\mathcal{L}V$ is smaller than $-\kappa$, in a neighborhood of radius $2\sqrt{n}\sigma$ around $\vec{x}$. First of all, notice that $2\sqrt{n}\sigma = 2\sqrt{2n\eta/\beta} \leq 2\sqrt{2n/\beta}$ and we will choose β sufficiently large such that this is smaller than $l/2$. Further, from Lemma B.17, we know that we can choose $\kappa = cn$ such that

$$\{\vec{x}\colon \mathcal{L}V(\vec{x}) \geq -cn\} \subset \{\vec{x}\colon \|\vec{x} - \vec{x}^*\| \leq l/2\}$$

($c > 0$ is a universal constant). This implies that the set K from Lemma B.16 is contained in a ball of radius l around $\vec{x}^*$. By the same lemma, the hitting time to $K \cup B$, is bounded by

$$\frac{V(\vec{x}_0)}{\kappa\eta/\beta - (e^{-cn}M/\beta + O(\eta\sqrt{\eta n/\beta})}, \tag{16}$$

where M is a total bound on $\mathcal{L}V$ and it is $O(n)$ using Lemma B.17. Since $\beta = \Theta(n) = \Theta(\kappa)$, we have that $\kappa\eta/\beta = \Theta(\eta)$. Further, $e^{-cn}M/\beta = O(e^{-cn}n/\beta) = O(e^{-cn})$, since $n = \Theta(\beta)$. Lastly, $\eta\sqrt{\eta n/\beta} = \Theta(\eta^{3/2})$ and this can be smaller than $\kappa\eta/\beta = \Theta(\eta)$ if η is sufficiently small. Hence, the denominator in (16) is $\Omega(\kappa\eta/\beta) \geq \Omega(1)$. Finally, we want to bound the numerator. We have that if $\vec{x}_0$ is bounded, then $\|V(\vec{x}_0)\| \leq O(1)$. This derives that the number of iterations required to hit either K or B is bounded by $O(1/\eta)$. Recall the definition of B from Lemma B.16, and notice that it contains only points where wither $\nabla \hat{L}$ or ∇V are larger than some constant $C_1 > 0$ that we can select. Hence, B only contains points of large

norm. In particular, the probability to hit B is very small, from Proposition B.9, hence, with high probability we first hit K. In particular, the hitting time is $O(1/\eta)$ with high probability.

Recall that this assumed that we run the dynamics according to L', yet the lemma is about running it according to L. Yet using Lemma B.8 we know that if we run the dynamics according to L, then with high probability, at iterations $t = 3/\eta, \cdots, 3/\eta + O(1/\eta)$, the norm is at least $r_0 = \cos g^{\circ d}(\pi)/2$. In this region, L and L' are the same and running the dynamics according to L is the same as running according to L'. In particular, the expected time to hit K after iteration $3/\eta$ is $O(1/\eta)$.

Notice that the above argument can fail with some probability, if at some point the norm of $\vec{x}_t$ is either very small or very large. Yet, if this holds, then by Lemma B.8 and Proposition B.9, after a small number of iterations the norm will be of the right order and again, one have a large chance of hitting K. Overall, a simple calculation shows that the expected number of iterations to hit K is $O(1/\eta)$ as required.

Lastly, we prove the high probability bound on τ. By iterating: fix $C > 0$ some appropriate constant, then for any t,

$$\Pr[\tau > C\log(1/\epsilon)/\eta] = \prod_{i=1}^{\log(1/\epsilon)} \Pr[\tau > Ci/\eta \mid \tau > C(i-1)/\eta] \leq \prod_{i=1}^{\log(1/\epsilon)} 0.1 \leq \epsilon,$$

where we use Markov's inequality to bound $\Pr[\tau > Ci/\eta \mid \tau > C(i-1)/\eta]$.[3] □

B.11. Continuous gets closer to discrete

Assume that a function H is M-smooth and μ-strongly convex. We want to compare the langevin iteration

$$\vec{x}_t = \vec{x}_{t-1} - \eta\nabla H(\vec{x}_{t-1}) + \vec{z}_t,$$

where $\vec{z} \sim N(\vec{0}, \sigma^2 I)$, to the continuous iteration defined by

$$d\vec{y}_t = -\eta\nabla H(\vec{y}_t)dt + \sigma dB_t.$$

Note that $\vec{x}_t$ runs in discrete times $t = 0, 1, 2, \ldots$ while the continuous runs continuous time $t \geq 0$. We want to show the following:

Lemma B.19. *Assume that we run the discrete and continuous time dynamics, $\vec{x}_t$ and $\vec{y}_t$, with respect to some function H, that is $\Omega(1)$ smooth and $O(1)$ strongly convex, assume that for all t, $\mathbb{E}\|\nabla H(\vec{x}_t)\| \leq O(1)$ and that $\|\vec{x}_0 - \vec{y}_0\| \leq O(1)$. Then, for $T \geq \Omega(\log(1/\eta)/\eta)$, one has a coupling between $\vec{x}_t$ and $\vec{y}_t$ such that*

$$\mathbb{E}\|\vec{x}_t - \vec{y}_t\| \leq O(\sqrt{\eta}).$$

Proof. To prove the lemma, let us first present $\vec{x}_t$ as continuous dynamics over $t \geq 0$:

$$d\vec{x}_t = -\eta\nabla H(\vec{x}_{\lfloor t \rfloor})dt + \sigma dB_t,$$

and note that the difference between $\vec{x}_t$ and $\vec{y}_t$ is that the gradient with respect to $\vec{x}_t$ is taken according to $\vec{x}_{\lfloor t \rfloor}$ and not to $\vec{x}_t$. This produces exactly the same distribution over $\vec{x}_0, \vec{x}_1, \ldots$ as the discrete dynamics. Let us couple $\vec{x}_t$ with $\vec{y}_t$, while using the same Gaussian noise dB_t. Then, if we take $\vec{x}_t - \vec{y}_t$ the noise cancels, and we have

$$d(\vec{x}_t - \vec{y}_t) = -\eta(\nabla H(\vec{x}_{\lfloor t \rfloor}) - \nabla H(\vec{y}_t))dt \ .$$

Applying Ito's lemma, and assuming that the function is c-strongly convex and C-smooth, one has

$$\begin{aligned}
&d\|\vec{x}_t - \vec{y}_t\|^2/2 = \langle \vec{x}_t - \vec{y}_t, d(\vec{x}_t - \vec{y}_t)\rangle = -\eta\left\langle \vec{x}_t - \vec{y}_t, \nabla H(\vec{x}_{\lfloor t \rfloor}) - \nabla H(\vec{y}_t)\right\rangle dt \\
&= -\eta\left\langle \vec{x}_t - \vec{y}_t, \nabla H(\vec{x}_t) - \nabla H(\vec{y}_t)\right\rangle dt + \eta\left\langle \vec{x}_t - \vec{y}_t, \nabla H(\vec{x}_t) - \nabla H(\vec{x}_{\lfloor t \rfloor})\right\rangle dt \\
&\leq -\eta c(\|\vec{x}_t - \vec{y}_t\|^2 + \eta\|\vec{x}_t - \vec{y}_t\|\|\nabla H(\vec{x}_t) - \nabla H(\vec{x}_{\lfloor t \rfloor})\|)dt \\
&\leq -\eta c(\|\vec{x}_t - \vec{y}_t\|^2 + \eta C\|\vec{x}_t - \vec{y}_t\|\|\vec{x}_t - \vec{x}_{\lfloor t \rfloor}\|)dt \ .
\end{aligned}$$

[3] There is one detail that should be taken care of: conditioned on $\tau > C(i-1)/\eta$, $\|\vec{x}_t\|$ may be large. Yet, since $\|\vec{x}_t\| \leq O(1)$ w.pr. e^{-cn} and since we can assume that $\epsilon \geq e^{-cn}$ (as an error of e^{-cn} is already present in the theorem statement), we will not encounter a large $\|\vec{x}_t\|$, even conditioned on a large τ.

Using Ito's formula again, one has

$$d\|\vec{x}_t - \vec{y}_t\| = d\sqrt{\|\vec{x}_t - \vec{y}_t\|^2} = \frac{d\|\vec{x}_t - \vec{y}_t\|^2}{2\|\vec{x}_t - \vec{y}_t\|} \leq (\eta c\|\vec{x}_t - \vec{y}_t\| + \eta C\|\vec{x}_t - \vec{x}_{\lfloor t \rfloor}\|)dt.$$

Integrating, one has

$$\|\vec{x}_T - \vec{y}_T\| = e^{-\eta cT}\|\vec{x}_0 - \vec{y}_0\| + \int_0^T \eta C\|\vec{x}_t - \vec{x}_{\lfloor t \rfloor}\|e^{-\eta c(T-t)}dt \ .$$

We would like to take an expectation, for that purpose, let us estimate $\|\vec{x}_t - \vec{x}_{\lfloor t \rfloor}\|$. Denote $s = t - \lfloor t \rfloor$. Then, $\vec{x}_t - \vec{x}_{\lfloor t \rfloor} \sim N(\nabla H(\vec{x}_{\lfloor t \rfloor})s\eta, s\sigma^2 I_n)$, in particular,

$$\mathbb{E}[\|\vec{x}_t - \vec{x}_{\lfloor t \rfloor}\|^2 \mid \vec{x}_{\lfloor t \rfloor}] = \|\nabla H(\vec{x}_{\lfloor t \rfloor})s\eta\|^2 + sn\sigma^2$$

which implies, by Jensen, that

$$\mathbb{E}\left[\|\vec{x}_t - \vec{x}_{\lfloor t \rfloor}\| \mid \vec{x}_{\lfloor t \rfloor}\right] \leq \|\nabla H(\vec{x}_{\lfloor t \rfloor})s\eta\| + \sqrt{ns}\sigma \leq \eta\|\nabla H(\vec{x}_{\lfloor t \rfloor})\| + \sigma\sqrt{n}.$$

Taking an outer expectation and using the bound on the gradient and that $\beta = \Theta(n)$, one has that

$$\mathbb{E}\left[\|\vec{x}_t - \vec{x}_{\lfloor t \rfloor}\|\right] \leq O(\eta + \sqrt{\eta n/\beta}) \leq O(\sqrt{\eta}).$$

Substituting this above, one has

$$\mathbb{E}[\|\vec{x}_T - \vec{y}_T\|] \leq e^{-\eta cT}\|\vec{x}_0 - \vec{y}_0\| + O(1)\int_0^T \eta^{3/2}e^{-\eta c(T-t)}dt \leq e^{-\eta cT}\|\vec{x}_0 - \vec{y}_0\| + O(\sqrt{\eta}).$$

The result follows by substituting $T \geq \Omega(\log(1/\eta)/\eta)$. □

B.12. Staying in the convexity region

We use the following known property for gradient descent:

Lemma B.20. *Let $f\colon K \to \mathbb{R}$, where $K \subset \mathbb{R}^n$ is a convex set. Assume that f is s-smooth and μ-strongly convex, let $\eta \leq 2/(s+\mu)$. Let $\vec{x}, \vec{y} \in K$ then,*

$$\|\vec{x} - \eta\nabla f(\vec{x}) - (\vec{y} - \eta\nabla f(\vec{y}))\| \leq \left(1 - \frac{\eta s\mu}{s+\mu}\right)\|\vec{x} - \vec{y}\|.$$

In particular, if K is a ball around the minima $\vec{x}^*$ of f, then

$$\|\vec{x} - \eta\nabla f(\vec{x}) - \vec{x}^*\| \leq \left(1 - \frac{\eta s\mu}{s+\mu}\right)\|\vec{x} - \vec{x}^*\|.$$

We would like to show that the Langevin rarely escapes some ball around $\vec{x}^*$. We have the following proposition:

Lemma B.21. *Let f be function that has local minima at $\vec{x}^*$ and which is μ-strongly convex and s-smooth in a ball of radius R around $\vec{x}^*$. Assume that we run Langevin dynamics, starting at $\vec{x}_0$, following*

$$\vec{x}_t = \vec{x}_{t-1} - \eta\nabla f(x) + \vec{z}_t,$$

where $\vec{z}_t \sim N(0, \sigma^2)$, and

$$\eta \leq \frac{2}{s+\mu}.$$

Further, assume the $\sigma\sqrt{n} \leq R/4$ and that $\|\vec{x}_0 - \vec{x}^*\| \leq R/2$ Then, for any $T > 0$,

$$\Pr[\forall i = 1, \ldots, T\colon \|\vec{x}_i - \vec{x}^*\| \leq R] \geq 1 - e^{-cn},$$

where $c > 0$ is a small universal constant.

To prove this lemma, we would like to couple $\vec{x}_0, \ldots, \vec{x}_T$ with auxiliary variables $\vec{y}_0, \ldots, \vec{y}_T$ such that for all $t \leq T$, if $\|\vec{y}_0\|, \ldots, \|\vec{y}_{t-1}\| \leq R$ then $\|\vec{x}_t\| \leq \|\vec{y}_t\|$. In particular, if $\|\vec{y}_0\|, \ldots, \|\vec{y}_T\| \leq R$ then $\|\vec{x}_0\|, \ldots, \|\vec{x}_T\| \leq R$. It will be convenient to bound the $\vec{y}_t$ variables.

Define $p = 1 - \frac{\eta s \mu}{s+\mu}$, then we define

$$\vec{y}_0 = \vec{x}_0; \quad \vec{y}_t = p\vec{y}_{t-1} + \vec{w}_t,$$

where $\vec{w}_t \sim N(0, \sigma^2)$. Let us show how to couple $\vec{x}_t$ and $\vec{y}_t$ and show by induction the required property. For $t = 0$ this holds by definition. Assume that this holds for all $i < t$ and prove for t. Let us assume that $\|\vec{y}_0\|, \ldots, \|\vec{y}_{t-1}\| \leq R$ otherwise the proof follows. By assumption we have that $\|\vec{x}_{t-1}\| \leq \|\vec{y}_t\| \leq R$. Denote by $\vec{x}'_{t-1} = \vec{x}_{t-1} - \eta \nabla f(\vec{x}_{t-1})$. By Lemma B.20, we have that

$$\|\vec{x}'_{t-1}\| \leq p\|\vec{x}_{t-1}\| \leq p\|\vec{y}_{t-1}\|.$$

Let us now couple the noise $\vec{z}_t$ added to $\vec{x}_{t-1}$ in the recursive formula, with the noise $\vec{w}_t$ added to $\vec{y}_{t-1}$. Denote $\tilde{\vec{x}}'_{t-1} = \frac{\vec{x}'_{t-1}}{\|\vec{x}'_{t-1}\|}$ and $\tilde{\vec{y}}_{t-1} = \frac{\vec{y}_{t-1}}{\|\vec{y}_{t-1}\|}$, $z_t = \langle \vec{z}_t, \tilde{\vec{x}}_{t-1} \rangle$, $w_t = \langle \vec{w}_t, \tilde{\vec{y}}_{t-1} \rangle$. Notice that $z_t, w_t \sim N(0, \sigma^2)$ and we have that:

$$\|\vec{x}_t\|^2 = \|\vec{x}'_{t-1} + \vec{z}_t\|^2 = \|\vec{x}'_{t+1} + z_t\tilde{\vec{x}}'_{t-1} + (\vec{z}_t - z_t\tilde{\vec{x}}'_{t-1})\|^2 = \|\vec{x}'_{t+1} + z_t\tilde{\vec{x}}'_{t-1}\|^2 + \|\vec{z}_t - z_t\tilde{\vec{x}}'_{t-1})\|^2, \tag{17}$$

where the last equality is due to the fact that $\vec{x}'_{t+1} + z_t\tilde{\vec{x}}_{t-1}$ is a multiple of $\|\vec{x}'_{t-1}\|$ while the second term is perpendicular to this vector. Similarly, we have that

$$\|\vec{y}_t\|^2 = \|p\vec{y}_{t-1} + \vec{w}_t\|^2 = \|p\vec{y}_{t-1} + w_t\tilde{\vec{y}}_{t-1} + (\vec{w}_t - w_t\tilde{\vec{y}}_{t-1})\|^2 = \|p\vec{y}_{t-1} + w_t\tilde{\vec{y}}_{t-1}\|^2 + \|\vec{w}_t - w_t\tilde{\vec{y}}_{t-1}\|^2. \tag{18}$$

We would couple $\vec{z}_t$ and $\vec{w}_t$ such that the first term in (17) is bounded by the first term in (18) and similarly for the second term. For the second term, notice that both $\vec{z}_t - z_t\tilde{\vec{x}}'_{t-1}$ and $\vec{w}_t - w_t\tilde{\vec{y}}_{t-1}$ are Gaussian variables with Isotropic covariance $\sigma^2 I_{n-1}$ over a subspace of dimension $n-1$, so we can couple them such that their absolute value is identical. Now, we argue for the first terms. Analyzing the first term in (17), we have that,

$$\vec{x}'_{t-1} + z_t\tilde{\vec{x}}'_{t-1} = (\|\vec{x}'_{t-1}\| + z_t)\tilde{\vec{x}}'_{t-1},$$

hence, the corresponding first term equals

$$\|\vec{x}'_{t-1} + z_t\tilde{\vec{x}}'_{t-1}\|^2 = (\|\vec{x}'_{t-1}\| + z_t)^2\|\tilde{\vec{x}}'_{t-1}\|^2 = (\|\vec{x}'_{t-1}\| + z_t)^2.$$

For the term corresponding (18), we have that

$$p\vec{y}_{t-1} + w_t\tilde{\vec{y}}_{t-1} = (\|p\vec{y}_{t-1}\| + w_t)\tilde{\vec{y}}_{y-1},$$

hence, the corresponding first term equals

$$\|p\vec{y}_{t-1} + w_t\tilde{\vec{y}}_{t-1}\|^2 = (\|p\vec{y}_{t-1}\| + w_t)^2\|\tilde{\vec{y}}_{y-1}\|^2 = (\|p\vec{y}_{t-1}\| + w_t)^2.$$

Hence, our goal is to couple z_t with w_t such that

$$(\|\vec{x}'_{t-1}\| + z_t)^2 \leq (\|p\vec{y}_{t-1}\| + w_t)^2,$$

or, equivalently,

$$|\|\vec{x}'_{t-1}\| + z_t| \leq |\|p\vec{y}_{t-1}\| + w_t|.$$

We have already argued that $\|\vec{x}'_{t-1}\| \leq \|p\vec{y}_{t-1}\|$. Further, notice that $z_t, w_t \sim N(0, \sigma^2)$. So, it is sufficiently to use the following lemma:

Lemma B.22. *Let $a \geq b \geq 0$. Then, we can couple two random variables, $z, w \sim N(0, \sigma^2)$ such that $|a + z| \geq |b + w|$.*

Proof. First of all, notice that two real-valued random variables, X, Y, can be coupled such that $X \geq Y$ whenever $\Pr[X \leq t] \leq \Pr[Y \leq t]$ for any $t \in \mathbb{R}$. This is a standard argument, and the proof is by first $b \in [0, 1]$ and then setting $X = x, Y = y$ for the values x, y such that $\Pr[X \leq x] = b$ and $\Pr[Y \leq y] = b$. By the assumption that $\Pr[X \leq t] \leq \Pr[Y \leq t]$ it holds that $x \geq y$ as required.

So, it suffices to show that $\Pr[|a+z| \leq t] \leq \Pr[|b+w| \leq t]$, for any $t \geq 0$. Indeed, if ϕ is the density of a random variable $N(0, \sigma^2)$, we have that

$$\begin{aligned}
&\Pr[|a+z| \leq t] - \Pr[|b+w| \leq t] \\
&= 1 - \Pr[|a+z| > t] - (1 - \Pr[|b+w| > t]) \\
&= \Pr[|b+w| > t] - \Pr[|a+z| > t] \\
&= \Pr[b+w > t] + \Pr[b+w < -t] - \Pr[a+z > t] - \Pr[a+z < -t] \\
&= \Pr[w > t-b] + \Pr[w < -t-b] - \Pr[z > t-a] - \Pr[z < -t-a] \\
&= \Pr[w > t-b] + \Pr[w > t+b] - \Pr[z > t-a] - \Pr[z > t+a] \\
&= \int_{t-b}^{\infty} \phi(u)du + \int_{t+b}^{\infty} \phi(u)du - \int_{t-a}^{\infty} \phi(u)du - \int_{t+a}^{\infty} \phi(u)du \\
&= \int_{t+b}^{t+a} \phi(u)du - \int_{t-a}^{t-b} \phi(u)du \\
&= \int_{b+t}^{a+t} \phi(u)du - \int_{b-t}^{a-t} \phi(u)du \\
&= \int_{b}^{a} (\phi(u+t) - \phi(u-t))du \\
&= \int_{b}^{a} (\phi(|u+t|) - \phi(|u-t|))du
\end{aligned}$$

using the fact that ϕ is symmetric around the origin. Recall that $a, b, t \geq 0$, hence $|u+t| \geq |u-t|$ which implies that $\phi(|u+t|) \leq \phi(|u-t|)$ as ϕ is decreasing. We derive that the desired quantity is negative and this is what we wanted to prove. □

This concludes the inductive proof that if $\|\vec{y}_1\|, \ldots, \|\vec{y}_T\| \leq R$ then $\|\vec{x}_1\|, \ldots, \|\vec{x}_T\| \leq R$. It suffices to bound the probability that $\|\vec{y}_1\|, \ldots, \|\vec{y}_T\| \leq R$. Notice that

$$\vec{y}_T = p\vec{y}_{T-1} + \vec{w}_T = p^T \vec{y}_0 + \sum_{t=1}^{T} p^{T-t} \vec{w}_t.$$

Notice that its mean is $p^T \vec{y}_0$ and its covariance is

$$\sum_{t=1}^{T} Cov(p^{T-t}\vec{w}_t) = \sum_{t=1}^{T} p^{2(T-t)} \sigma^2 I_n = \sigma^2 I_n \sum_{i=0}^{T-1} p^{2i} \leq \sigma^2 I_n \sum_{i=0}^{\infty} p^{2i} = \frac{\sigma^2}{1-p^2} I_n,$$

where the inequality corresponds to the constant that multiplies the identity matrix. Using this inequality, we can derive that

$$\begin{aligned}
\Pr[\forall i, \|\vec{y}_i\| \leq R] &\leq \sum_{i=1}^{T} \Pr[\|\vec{y}_i\| \leq R] \leq \sum_{i=1}^{T} \Pr[\|\vec{y}_i - \mathbb{E}\vec{y}_i\| \leq R - \|\mathbb{E}\vec{y}_i\|] \\
&\leq \sum_{i=1}^{T} \Pr[\|\vec{y}_i - \mathbb{E}\vec{y}_i\| \leq R - p^t\|\vec{y}_0\|] \leq \sum_{i=1}^{T} \Pr[\|\vec{y}_i - \mathbb{E}\vec{y}_i\| \leq R - \|\vec{y}_0\|] \\
&\leq \sum_{i=1}^{T} \Pr[\|\vec{y}_i - \mathbb{E}\vec{y}_i\| \leq R/2].
\end{aligned}$$

We can use the fact that for a random variable $\vec{X} \sim N(\vec{0}, \sigma^2 I_n)$, it holds that $\Pr[\|\vec{X}\| > 2\sigma\sqrt{n}] \leq e^{-cn}$ for some universal constant $c > 0$. In particular, applying $\vec{X} = \vec{y}_i - \mathbb{E}\vec{y}_i$, we derive that

$$\Pr[\forall i, \|\vec{y}_i\| \leq R] \leq Te^{-cn},$$

using that $R/2 \geq 2\sigma\sqrt{n}$. This concludes the proof.

B.13. Culminating the Proof

We start by arguing about the dynamics according to L and then we argue for $\tilde{L}$. First, assume that $t = \Theta(\log(1/\epsilon)/\epsilon^2)$. This assumption will be removed later. Let τ be the minimal t such that $\|\vec{x}_t - \vec{x}^*\| \leq l$ for some appropriately chosen constant $l > 0$. First, notice that by Lemma B.18, with probability $1 - \epsilon$ we have $\tau \leq O(\log(1/\epsilon)/\epsilon)$. By Lemma B.6 there is some radius around $\vec{x}^*$ where the function is $\Omega(1)$-strongly convex, and assume that this radius is $2l$. By Lemma B.21, with high probability, the dynamics stay within the ball of radius $2l$ for additional $O(\log(1/\epsilon)/\epsilon^2)$ iterations.

In order to bound the Wasserstein distance between $\vec{x}_T$ and μ, we would like to couple the discrete dynamics $\vec{x}_t$ to the continuous dynamics, defined by

$$d\vec{y}_t = -\eta \nabla L(\vec{y}_t)dt + \sqrt{2\eta/\beta}dB_t; \vec{y}_0 \sim \mu, \mu(\vec{y}) = \frac{e^{-\beta L(\vec{y})}}{\int e^{-\beta L(\vec{z})}d\vec{z}};$$

Notice that $\vec{y}_t \sim \mu$ for all t. The coupling is done as follows: the chains are run independently until time τ. Since $\vec{y}_\tau \sim \mu$ independently of $\vec{x}_\tau$, and since, assuming that $\beta \geq \Omega(n)$, μ has mass $1 - e^{-cn}$ in the ball of radius ℓ around $\vec{x}^*$ (as can be computed using a simple integral), one has that with probability $1 - e^{-cn}$, $\vec{x}_t$ and $\vec{y}_t$ are in this ball. From that point onward, using Lemma B.19 we can couple $\vec{x}_t$ and $\vec{y}_t$ such that after additional $O(\log(1/\epsilon)/\eta)$ iterations, $\mathbb{E}\|\vec{x}_t - \vec{y}_t\| \leq O(\epsilon)$. By taking into account the failure probability to stay and remain in the convexity region, we derive that after $T = O(\log(1/\epsilon)/\eta) = O(\log(1/\epsilon)/\epsilon^2)$ iterations, $\mathbb{E}\|\vec{x}_t - \vec{y}_t\| \leq O(\epsilon) + T2^{-cn}$. We can assume that $\epsilon \geq e^{-cn/10}$ without hurting the guarantee, hence $T \leq e^{cn/2}$, and the error is bounded by $e^{-cn/2} + O(\epsilon)$.

Next, we want to argue about the dynamics according to $\tilde{L}$ rather than L. Due to the added noise in each step, if the parameter δ in (6) is sufficiently small, then the KL divergence between the execution with L vs. $\tilde{L}$ is small.

Lastly, we argue about what happens when $T > \Omega(\log(1/\epsilon)/\epsilon^2)$: in this case, we disregard the initial iterations, and restart the above argument, replacing $t = 0$ with $t = T - \Theta(\log(1/\epsilon)/\epsilon^2)$ and replacing T with $\Theta(\log(1/\epsilon)/\epsilon^2)$.

C. Additional Experiments

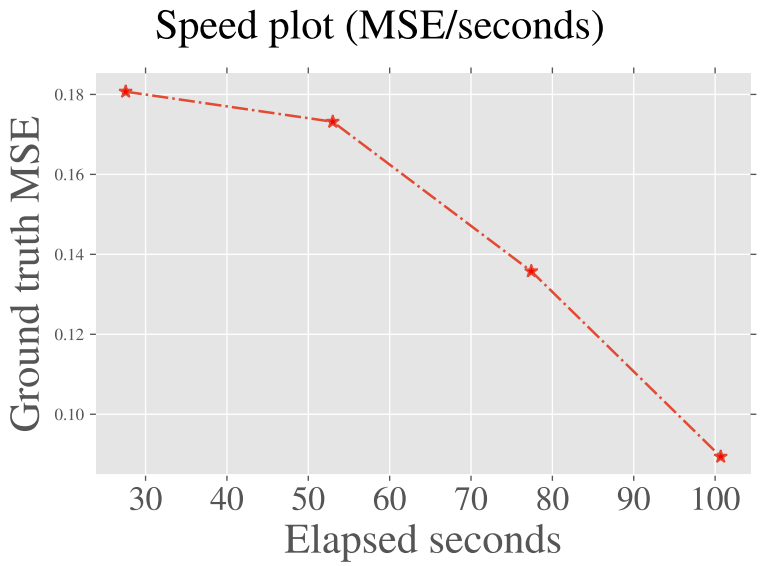

Figure 9: Speed plot that demonstrates how the loss is changing over time. Each inversion takes about $1 - 2$ minutes on single V100 GPU.