Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Approximating Langevin Monte Carlo with ResNet-like neural network architectures

Martin Eigel, Charles Miranda, Janina Enrica Schütte, David Sommer

submitted: December 21, 2023

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: martin.eigel@wias-berlin.de
janina.schuette@wias-berlin.de
charles.miranda@wias-berlin.de
david.sommer@wias-berlin.de

No. 3077
Berlin 2023

2020 *Mathematics Subject Classification.* 62F15, 65N75, 65C30, 60H35, 62H12, 65C05, 60H35, 68T07,

Key words and phrases. Langevin Monte Carlo, Approximate sampling, Rates of convergence, Neural Network approximation, Wasserstein distance.

We acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the the priority programme SPP 2298 "Theoretical Foundations of Deep Learning". ME, CM & DS acknowledge support by the ANR-DFG project *COFNET: Compositional functions networks - adaptive learning for high-dimensional approximation and uncertainty quantification*. This study does not have any conflicts to disclose.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Approximating Langevin Monte Carlo with ResNet-like neural network architectures

Martin Eigel, Charles Miranda, Janina Enrica Schütte, David Sommer

Abstract

We sample from a given target distribution by constructing a neural network which maps samples from a simple reference, e.g. the standard normal distribution, to samples from the target. To that end, we propose using a neural network architecture inspired by the Langevin Monte Carlo (LMC) algorithm. Based on LMC perturbation results, we show approximation rates of the proposed architecture for smooth, log-concave target distributions measured in the Wasserstein-2 distance. The analysis heavily relies on the notion of sub-Gaussianity of the intermediate measures of the perturbed LMC process. In particular, we derive bounds on the growth of the intermediate variance proxies under different assumptions on the perturbations. Moreover, we propose an architecture similar to deep residual neural networks and derive expressivity results for approximating the sample to target distribution map.

1 Introduction and main result

In this work we consider the task of measure transport in terms of a stochastic dynamical system described by a Langevin SDE. The main novelty is a complete analysis of the convergence and complexity when discretized as an Euler-Maruyama scheme in an appropriate deep neural network architecture. Several research areas are related to our approach for which we give an overview in the following.

Deep Neural Networks Sampling from probability densities is a common problem in fields such as Bayesian inference [61, 34] and generative modelling (GM) [60]. In GM in particular, the training of deep neural networks (DNNs) to sample from the target distribution has become widespread [56, 6]. Popular approaches in deep generative modelling (DGM) include normalizing flows, variational autoencoders [36, 37, 50], and generative adversarial networks (GANs) [28]. Beyond DGM, there is also growing interest in GM via score-based diffusion models [62, 60], where DNNs are trained to approximate the score function, i.e. the gradient log-density of the forward diffusion process. While most research has been focused on image, video and text generation, generative models can also be used in the context of differential equations related to engineering and the natural sciences [18].

From a mathematical perspective, there has been a lot of work on the expressivity analysis of fully connected neural networks (FCNNs), providing qualitative results regarding the complexity of representations. For this, classical and new approximation classes and function spaces are considered such as in [31]. Important approximation results include [2, 3, 4, 5, 23, 46, 47, 63, 64, 65]. An analysis for the FCNN approximation of (parametric) PDEs was for instance presented in [43, 38, 8, 39]. In a recent work [33], FCNNs have been shown to beat the curse of dimensionality in approximation of Kolmogorov backward equations (KBE) provided that the drift term can be approximated without curse of dimensionality. The central idea is the use of the Feynman-Kac formula, linking the KBE solution

DOI 10.20347/WIAS.PREPRINT.3077 Berlin 2023

to the expectation of an observable subject to an underlying Itô diffusion process. Under suitable assumptions on the observable (which coincides with the initial condition of the KBE) and the drift term, this diffusion process can be approximated by adding and composing FCNN layers in an imitation of the Euler-Maruyama discretization.

While standard feed-forward FCNNs offer conceptual simplicity, there are architectures arguably better suited to the approximation of differential equations. One such architecture is given by deep residual networks, also called *ResNets*, [32], which instead of the full mapping from input to output learn a residual component (respective to the layer input) in each layer. Due to this residual structure, ResNets have interesting theoretical connections to time discretizations of differential equations. In particular, the forward propagation of the inputs through the residual layers can be interpreted as time-discretization of an underlying (stochastic) differential equation. The continuous-time equivalent of a ResNet is called a neural ODE [10, 57].

Sampling and Langevin Monte Carlo There exists a vast literature on sampling strategies from unknown probability measures including Markov Chain Monte Carlo (MCMC) [52, 51, 7], Sequential Monte Carlo (SMC) [19] and Langevin dynamics [53, 54]. The Langevin method has strong historical connections to statistical physics [55] and can be seen as a stochastic analogue to gradient descent. Extensions of these methods have been proposed as Metropolis adjusted Langevin and Hamiltonian Monte Carlo (HMC) sampling methods defined on Riemannian manifolds [26] and also to ensemble methods [24, 25] ensuring affine invariance [29]. Under smoothness and growth conditions on the log-density, one can define a simple first order overdamped Langevin process, which admits the target density as invariant measure and which contracts exponentially (in relative entropy) to that invariant measure [44]. Methods obtained by discretization of this process are called Langevin Monte Carlo (LMC) methods. Errors bounds of LMC in case of an M-Lipchitz, m-strongly convex potential have been extensively studied in Wasserstein-2 distance, relative entropy and total variation distance, cf. [15, 16, 20, 17, 11, 21, 22]. There are also works aiming to extend the convergence analysis beyond the restricted log-concave setting [41, 12, 42, 49, 9, 67]. A good overview of the different approaches can be found in [66]. Another interesting work is [1], where a bound for the variance proxy of the sub-Gaussian invariant distribution of LMC is derived.

Goal and methodology of this work This work is concerned with sampling from measures of the form $\mathrm{d}\mu_\infty(x) = Z^{-1}e^{-V(x)}\mathrm{d}x$, where V is smooth and log-concave as per Assumption 1.1 and Z is an (unknown) normalization constant.

Assumption 1.1. *We make the following assumption on the potential V of μ_∞:*

- *V has a M-Lipschitz gradient:* $\forall x, y \in \mathbb{R}^d,\ \|\nabla V(x) - \nabla V(y)\|_{\ell^2} \leq M\|x - y\|_{\ell^2}$.
- *V is strongly convex with parameter m:* $\forall x, y \in \mathbb{R}^d,\ V(x) - V(y) - \langle \nabla V(y), x - y\rangle \geq \frac{m}{2}\|x - y\|^2_{\ell^2}$.

The goal is to derive complexity bounds for a neural network architecture, which takes inputs $Y_0 \sim \mu_0$ distributed according to a sub-Gaussian reference distribution μ_0 and outputs samples from μ_∞ up to an epsilon error in the Wasserstein-2 distance. To achieve this, a ResNet-like architecture inspired by the LMC algorithm is proposed. A sketch of the architecture is pictured in Figure 1.1. The introduced network has two important properties. By imitating the Langevin algorithm, the number of parameters in the expressivity results mainly depends on how well the drift can be approximated as well as on the

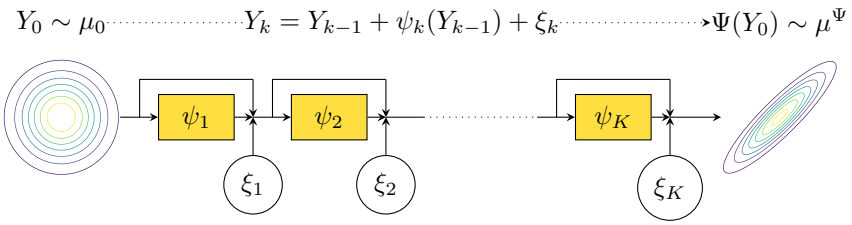

Figure 1.1: Sketch of the ResNet-like architecture used in this work.

number of realized time discretization steps. Furthermore, the architecture allows to train small parts of the network separately. The analysis combines known results for Wasserstein-2 convergence of LMC with perturbation arguments for the FCNN-approximated drift terms. Results are derived under the assumption of the availability of an FCNN able to approximate the drift on a ball as in Assumption 1.2 (ii).

Results are derived for two different assumptions on the availability of FCNN approximations, which are listed in Assumption 1.2.

Assumption 1.2. *We assume the following on the availibility of FCNN approximations of the drift.*

(i) For any $\varepsilon > 0$ there exists an FCNN ϕ_ε with realization $\mathcal{R}\phi_\varepsilon$ and $N(d, \varepsilon, m, M)$ parameters such that

$$\| -\nabla V(x) - \mathcal{R}\phi_\varepsilon(x)\|_{\ell^2} \leq \varepsilon(1 + \|x\|_{\ell^2}) \tag{1.1}$$

for all $x \in \mathbb{R}^d$.

(ii) For any $\varepsilon > 0$ and $r > 0$ there exists an FCNN $\phi_{\varepsilon,r}$ with realization $\mathcal{R}\phi_{\varepsilon,r}$, number of parameters $N(d, r, \varepsilon, m, M)$ and depth $L(d, r, \varepsilon, m, M)$ such that

$$\| -\nabla V - \mathcal{R}\phi_{\varepsilon,r}\|_{L^\infty(B_r(0);\mathbb{R}^d)} \leq \varepsilon/\sqrt{2},$$

where $B_r(0)$ is the closed ℓ^2-ball with radius r and center in 0 in $\mathbb{R}^d$.

In the analysis a perturbed LMC process is considered, where $-\nabla V$ is replaced by neural network approximations. It is then shown that under either of the assumptions in Assumption 1.2 the global L^2-error of the approximation can be bounded with respect to the current measure at any time in the process. The analysis differs depending on which item of Assumption 1.2 is assumed. In Assumption 1.2 (ii), only approximation on a ball of arbitrary radius is presupposed. In particular, it is not assumed that $-\nabla V$ can be globally approximated. In this case, concentration inequalities are used to bound the error outside of balls of suitable radius. The measures appearing in the concentration inequalities are ensured to be sub-Gaussian by a cutoff of the FCNNs via two additional ReLU layers. Assumption 1.2 (i), on the other hand, presupposes global approximation, with the error growing at most linearly and with sufficiently small slope G. While this is a strong assumption, note that the set of functions satisfying it is non-empty, since it includes quadratic potentials V. Why is the linear growth bound on the error desirable? The linear growth allows us to uniformly bound the variance proxies of the sub-Gaussian distributions induced by the approximate LMC process (where the drift $-\nabla V$ is replaced by FCNN approximations) by the sum of the variance proxies of the starting and target distributions. As we will show later, this greatly simplifies the analyis. Somewhat mitigating the strong assumptin of linear error growth, we show that Assumption 1.2 (ii) plus a constraint on the Lipschitz constant M of the potential gradient implies a similar uniform bound on the variance proxies.

Results This work gives complexity bounds for a ResNet-like neural network architecture approximating the Langevin Monte Carlo process. In doing so, we are able to show an interesting property of the underlying LMC process, which is that the variance proxies of the sub-Gaussian intermediate measures are uniformly bounded (Proposition 5.1). To the best of our knowledge, this result is not yet known to the community, with bounds on the variance proxy of the invariant measure having been shown as recent as [1]. The following is an informal version of the three main complexity results of this work, namely Theorem 5.4, Theorem 6.2 and Theorem 7.3.

Theorem 1.3 (Main result)**.** *Assume that $V : \mathbb{R}^d \to \mathbb{R}$ is an M-Lipschitz, m-strongly convex potential as in Assumption 1.1. Let μ_0 be sub-Gaussian with variance proxy $\sigma_0^2 > 0$ and $Y_0 \sim \mu_0$. Then, for any $\varepsilon > 0$, any $h \in (0, \frac{2}{m+M})$ and any $K \in \mathbb{N}$ there exists a ResNet-like network Ψ such the measure μ^Ψ of $\Psi(Y_0)$ satisfies*

$$\mathcal{W}_2(\mu_\infty, \mu^\Psi) \leq (1-mh)^K \mathcal{W}_2(\mu_\infty, \mu_0) + \frac{7\sqrt{2}}{6}\frac{M}{m}\sqrt{hd} + \frac{1-(1-mh)^K}{m}\varepsilon. \tag{1.2}$$

Furthermore, the complexity of Ψ can be bounded as follows.

(i) Under Assumption 1.2 (i), there exists a constant $c(d) \in \mathcal{O}(d^{-1})$ and a ResNet-like network Ψ satisfying (1.2) with number of parameters bounded by $KN(d, c(d)\varepsilon, m, M)$.

(ii) Under Assumption 1.2 (ii) and the additional assumption that $M < \sqrt{2}m$, there exists a ResNet-like network Ψ satisfying (1.2) with number of parameters in

$$K \cdot \mathcal{O}\big(d\log(2d\max\{1, r\sqrt{M^2-m^2}\}/\varepsilon) + N(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) + dL(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) + 2d^2\big),$$

where $r \in \mathcal{O}(d(1+\ln(d^2\varepsilon^{-4})^{-\frac{1}{2}}))$.

(iii) Under Assumption 1.2 (ii), there exists a ResNet-like network Ψ satisfying (1.2) with number of parameters bounded by $K(N(d, r, \varepsilon/\sqrt{d}, m, M) + 2d^2 + 2)$ where $r \in \mathcal{O}\left(d^{7/4}\varepsilon^{-1}(d^{9/4}\varepsilon^{-1})^{3(1.5^K-1)}\right)$.

We provide some remarks and intuition on this result. Consider first the perturbation in Wasserstein distance, (1.2). The first two terms result from the LMC process, which the *ResNet-like* network Ψ imitates. This process has step-size h and total number of steps K. The first term, $(1-mh)^K\mathcal{W}_2(\mu_\infty, \mu_0)$, results from the contraction property of the continuous Langevin dynamics. Note in particular that it decreases when increasing either the step-size or the number of steps, both of which correspond to letting the Langevin process run for a longer period Kh. If we fix a terminal time T such that $h = T/K$ and consider the continuous limit $K \to \infty$, this term recovers the exponential convergence of the continuous system, since $(1-mT/K)^K \to \exp(-mT)$. It can be clearly seen that the strong convexity constant m defines the speed of the contraction in the continuous (and low step-size) setting. The second term $\frac{7\sqrt{2}}{6}\frac{M}{m}\sqrt{hd}$ is a discretization error of the LMC algorithm, decreasing with the step size h. The third term results from a neural network approximation of the drift in every step. This error is bounded by ε/m and it inherits the contraction property of the LMC process: for $h = T/K$ and in the continuous limit $K \to \infty$, the term becomes $(1-\exp(-mT))\varepsilon/m$, which goes to 0 for $T \to 0$ with speed of contraction given by the strong convexity parameter m. Of course, the limit $K \to \infty$ alone implies an infinitely deep neural network, however this limit serves as a consistency check to ensure that the network does not introduce any additional bias.

We receive different upper bounds for the complexity of the network, depending on the assumption. In Theorem 1.3, the results are listed from strong to weak assumptions. Theorem 1.3 (i) presupposes

global approximation with linear error growth. Here, the network Ψ is constructed by "compositioning" K times the network $\phi_{c(d)\varepsilon}$ from Assumption 1.2 (i), i.e. $\psi_1 = \psi_2 = \ldots = \psi_K = h\phi_{c(d)\varepsilon}$ in the sense of Figure 1.1. In this case, the variance proxies of the intermediate measures (of the random variables Y_k in Figure 1.1) can be uniformly bounded and the error incurred by the network in every step can be bounded by upper bounds on second moments derived from the variance proxies. The resulting complexity is simply K times the complexity of the drift approximation $\phi_{c(d)\varepsilon}$. The analysis for this case is done in Section 5. Theorem 1.3 (ii) and Theorem 1.3 (iii) presuppose only local approximations of the gradient potential. Assuming no additional structure of the potential, we cannot derive uniform bounds on the intermediate variance proxies and hence the result in Theorem 1.3 (iii) yields a complexity growing exponentially in the number of steps K. The analysis involves a "worst case" upper bound on the variance proxies, using globally bounded neural networks, with the upper bound growing from step to step. This is the subject of Section 6. In Theorem 1.3 (ii), we require additional structure in order to receive a uniform bound on the variance proxies. Namely, V must not be "too far away" from a quadratic function, which is encoded in the additional condition that $M < \sqrt{2}m$. An alternative way to view this condition is that ∇V can be approximated globally with a linear function, incurring "small enough" error. Under this additional assumption, we can again show the existence of a network ϕ, such that ResNet-like network Ψ given by blocks $\psi_1 = \psi_2 = \ldots = \psi_K = h\phi$ (in the sense of Figure 1.1) satisfies (1.2). The construction of the network ϕ is quite technical, involving cut-offs and multiplication with approximate indicator functions of the networks provided by Assumption 1.2 (ii). Hence, the complexity is higher in this case than under the assumption of global linear error growth. Note however that the complexity again only grows linearly in the number of steps K. This analysis is performed in Section 7.

Structure of the paper We begin with the definition of the Langevin process, Wasserstein spaces and feed-forward neural networks and introduce our notation in Section 2. We continue by introducing the *ResNet-like* architecture in Section 3. A convergence result for the perturbed Langevin process with approximate drifts in every step is derived in Section 4. Our main results on the expressivity of the ResNet-like architecture to sample from the given distribution can be found in Section 5, Section 6 and Section 7. Herein, Section 5 is based on Assumption 1.2 (i), Section 6 and Section 7 are based on Assumption 1.2 (ii). The proofs of all results can be found in the appendix. In Section 8, numerical experiments for Gaussian and Gaussian mixture posterior distributions can be found. We summarize and discuss our results in Section 9.

2 Definitions and notation

We briefly recall some definitions and results that are used throughout the paper. An analysis of the Langevin Monte Carlo approximation is carried out involving the notion of Wasserstein spaces. Moreover, the multi-dimensional sub-Gaussian random vector is defined and a notion of Lyapunov functions is introduced as well as a formal notation for neural networks.

2.1 Langevin Monte Carlo and Wasserstein space

Throughout this manuscript, let $d \in \mathbb{N}$ be the dimension of the space, on which the target distribution lives. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $W : [0, \infty) \times \Omega \to \mathbb{R}^d$ be standard Brownian motion. Let $V \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ satisfy Assumption 1.1 with Lipschitz constant $M \in (0, \infty)$ and convexity

constant $m \in (0,\infty)$. Let $h \in (0,\infty)$. Let $\chi_h : [0,\infty) \to [0,\infty)$, $\chi_h(s) = \max\{kh : k \in \mathbb{N}_0, kh \leq s\}$ and $\xi_k = \sqrt{2}(W_{kh} - W_{(k-1)h}) \sim \mathcal{N}_d(0, 2hI_d)$ for $k \in \mathbb{N}$. Let μ_0 be a distribution on $\mathbb{R}^d$ with finite moments $\mathbb{E}_{x\sim\mu_0}[|x|^p]^{\frac{1}{p}} < \infty$ for all $p \in \mathbb{N}$ and $X_0, \tilde{X}_0 \colon \Omega \to \mathbb{R}^d$ be random variables independent of the Brownian motion W such that $X_0, \tilde{X}_0 \sim \mu_0$. Let $X, \tilde{X} \colon [0,\infty) \times \Omega \to \mathbb{R}^d$ be stochastic processes satisfying

$$X_t = X_0 - \int_0^t \nabla V(X_s)\mathrm{d}s + \sqrt{2}W_t, \tag{2.1a}$$

$$\tilde{X}_t = \tilde{X}_0 - \int_0^t \nabla V(\tilde{X}_{\chi_h(s)})\mathrm{d}s + \sqrt{2}W_t, \tag{2.1b}$$

for all $t \in [0,\infty)$ and denote the distributions of X_t and $\tilde{X}_t$ by μ_t^X, and $\mu_t^{\tilde{X}}$, respectively. We call $\tilde{X}$ the *Langevin-Monte-Carlo* process (LMC). Let μ_∞ be the probability distribution on $\mathbb{R}^d$ defined by $\mathrm{d}\mu_\infty(x) = \frac{1}{Z}e^{-V(x)}\mathrm{d}x$, where $Z \in (0,\infty)$ is a normalization constant such that $\int_{\mathbb{R}^d} \mathrm{d}\mu_\infty(x) = 1$.

Since they will be used frequently, we review the standard definition of L^p and ℓ^p spaces.

Definition 2.1 (L^p and ℓ^p spaces)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a probability space. For $p \in [1,\infty]$, we define the functional space*

$$\mathcal{L}^p(\Omega,\mathcal{F},\mu) \coloneqq \left\{ f \colon \Omega \to \mathbb{R} \text{ measurable} \mid \|f\|_{L^p_\mu(\Omega)} < \infty \right\},$$

where for all measurable functions f

$$\|f\|_{L^p_\mu(\Omega)} \coloneqq \begin{cases} \left(\int_\Omega |f|^p \mathrm{d}\mu\right)^{1/p} & 1 \leq p < \infty \\ \inf\{C \geq 0 : |f| \leq C, \mu - \text{a.s.}\} & p = \infty \end{cases}.$$

We also define the quotient space $L^p(\Omega,\mathcal{F},\mu) \coloneqq \mathcal{L}^p(\Omega,\mathcal{F},\mu)/\sim$ with the compact form $L^p_\mu(\Omega) \coloneqq L^p(\Omega,\mathcal{F},\mu)$, where $\sim$ denotes the usual equivalence of $\mu-$a.s. equal functions. In this space, $\|\cdot\|_{L^p_\mu(\Omega)}$ is a norm.

For $x \in \mathbb{R}^d$ we define $\|x\|_{\ell^p} \coloneqq (\sum_{i=1}^d x_i^p)^{1/p}$ for $p \in (0,\infty)$ and $\|x\|_\infty = \max_{i=1,\dots,d}|x_i|$. In case of a real-vector-valued function $f : \Omega \to \mathbb{R}^d$ define the Bochner space $L^p(\Omega,\mathcal{F},\mu;\mathbb{R}^d) = L^p_\mu(\Omega;\mathbb{R}^d)$ as the space of all measurable functions such that $\|f\|_{\ell^p} \in L^p(\Omega,\mathcal{F},\mu)$. The norm of this space is defined via

$$\|f\|_{L^p_\mu(\Omega;\mathbb{R}^d)} \coloneqq \|\|f\|_{\ell^p}\|_{L^p_\mu(\Omega)}.$$

Furthermore, the notion of balls and spheres will be needed.

Definition 2.2 (Balls and spheres)**.** *Let $(U, \|\cdot\|_U)$ be some normed space. We define the following:*

- ***Closed ball**: the* closed ball *of radius $r > 0$ centered at a point $x \in U$ is defined by*

$$B_r(x) \coloneqq \{y \in U : \|x - y\|_U \leq r\}.$$

- ***Sphere:** the* sphere *of radius $r > 0$ centered at a point $x \in U$ is defined by*

$$\mathbb{S}_r(x) \coloneqq \{y \in U : \|x - y\|_U = r\}.$$

We henceforth use $U = \mathbb{R}^d$ *and* $\|\cdot\|_U = \|\cdot\|_{\ell^p}$*. In this case, we will write* $B_r^p(x)$ *and* $\mathbb{S}_r^p(x)$ *to denote the norm dependence. In the special case of* $p = 2$ *we will often simply write* $B_r^2(x) = B_r(x)$ *and* $\mathbb{S}_r^2(x) = \mathbb{S}_r(x)$*.*

We henceforth use the *Kantorovich–Rubinstein metric* (or *Wasserstein-*p *distance*) of measures for $p = 2$.

Definition 2.3 (Wasserstein space)**.** *For* $p \geq 1$*, denote by* $\mathcal{D}_p(\mathbb{R}^d)$ *the set of probability measures on* $\mathbb{R}^d$ *endowed with the Wasserstein-*p *distance*

$$\mathcal{W}_p(\mu,\nu)^p := \min_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d} \|x-y\|_{\ell^p}^p \mathrm{d}\gamma(x,y),$$

where $\Pi(\mu,\nu) := \left\{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : (\pi_0)_\sharp\gamma = \mu, (\pi_1)_\sharp\gamma = \nu\right\}$ *is the set of transport plans with marginals* μ *and* ν*.*

The next theorem recalls Wasserstein-2 convergence results for the LMC scheme defined by Equation (2.1b).

Theorem 2.4 (Guarantees for the constant-step LMC [17, Theorem 1])**.** *Assume that* $h \in (0, \frac{2}{M})$ *and that* V *satisfies Assumption 1.1. For any* $K \in \mathbb{N}_0$*, the following claims hold:*

- *If* $h \leq \frac{2}{m+M}$ *then* $\mathcal{W}_2(\mu_\infty, \mu_{Kh}^{\tilde{X}}) \leq (1-mh)^K \mathcal{W}_2(\mu_\infty,\mu_0) + \frac{7\sqrt{2}}{6}\frac{M}{m}\sqrt{hd}$*.*
- *If* $h \geq \frac{2}{m+M}$ *then* $\mathcal{W}_2(\mu_\infty, \mu_{Kh}^{\tilde{X}}) \leq (Mh-1)^K \mathcal{W}_2(\mu_\infty,\mu_0) + \frac{7\sqrt{2}}{6}\frac{Mh}{2-Mh}\sqrt{hd}$*.*

2.2 Sub-Gaussianity

In the later analysis, we make use of sub-Gaussian random variables the distributions of which exhibit a strong tail decay dominated by the tails of a Gaussian.

Definition 2.5 (Sub-Gaussian random variable)**.** *Let* Z *be a random variable on* $\mathbb{R}$*.* Z *is said to be* sub-Gaussian *with variance proxy* $\sigma^2 > 0$ *if it satisfies one of these equivalent conditions:*

(i) For any $s \in \mathbb{R}$*,* $\mathbb{E}[\exp(sZ)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right)$*.*
(ii) For any $r > 0$*,* $\mathbb{P}(Z \geq r) \leq \exp\left(-\frac{r^2}{2\sigma^2}\right)$ *and* $\mathbb{P}(Z \leq -r) \leq \exp\left(-\frac{r^2}{2\sigma^2}\right)$*.*
(iii) For any $q \in \mathbb{N}$*,* $\mathbb{E}[|Z|^q] \leq (\sqrt{2}\sigma)^q q\Gamma(q/2)$*.*

We then write $Z \sim \mathrm{subG}(\sigma^2)$*.*

Definition 2.6 (Sub-Gaussian random vector)**.** *A random vector* $Z \in \mathbb{R}^d$ *is said to be* sub-Gaussian *with variance proxy* $\sigma^2 > 0$ *if for any* $u \in \mathbb{S}_1(0)$ *the real random variable* $\langle u, Z\rangle$ *is sub-Gaussian with variance proxy* σ^2*. We then write* $Z \sim \mathrm{subG}(\sigma^2)$*.*

Proposition 2.7 (ℓ^p-norm of a sub-Gaussian random vector is sub-Gaussian)**.** *Let* $Z \in \mathbb{R}^d$ *be a sub-Gaussian random vector with variance proxy* $\sigma^2 > 0$*. Then,* $\|Z\|_{\ell^p}$ *is a sub-Gaussian random variable for any* $p \geq 1$ *with variance proxy bounded by* $d^2\sigma^2$*.*

2.3 Lyapunov functions

There is an interesting connection between sub-Gaussianity and Lyapunov functions, which we exploit in our analysis. This connection presented in Lemma 2.9 was for instance used in [1] to bound the variance proxy of the invariant measure of the LMC algorithm in the smooth log-concave setting.

Definition 2.8 (Lyapunov function [1, Definition 3.1])**.** *For any weight $\lambda > 0$ the Lyapunov function $\mathcal{L}_\lambda \colon \mathbb{R}^d \to \mathbb{R}$ is defined by*

$$\mathcal{L}_\lambda(x) = \mathbb{E}_{v\sim\mathbb{S}_1(0)}\left[e^{\lambda\langle v,x\rangle}\right],$$

where $v \sim \mathbb{S}_1(0)$ denotes uniform sampling from the ℓ^2-unit sphere in $\mathbb{R}^d$. Furthermore, let $\ell \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be defined by

$$\ell(z) = \mathbb{E}_{v\sim\mathbb{S}_1(0)}\left[e^{z\langle v,e_1\rangle}\right]$$

such that it holds $\mathcal{L}_\lambda(x) = \ell(\lambda\|x\|)$ due to the rotational invariance of the Lypunov function.

The following lemma establishes the fact that sub-Gaussianity of a random variable is equivalent to an appropriate exponential upper bound on the expectation of the Lyapunov function.

Lemma 2.9 (Connection between sub-Gaussianity and Lyapunov functions)**.** *X is a sub-Gaussian random vector with variance proxy σ^2 if and only if it holds for all $\lambda > 0$ that*

$$\mathbb{E}_X \mathcal{L}_\lambda(X) \leq e^{\frac{\sigma^2\lambda^2}{2}}.$$

2.4 Neural networks

We recall the formal mathematical notation of neural networks as introduced in [48]. The neural network is defined as a set of weights and biases and the corresponding function is defined as the realization.

Definition 2.10 (Neural network architectures)**.** *Let*

$$\mathcal{N} = \bigcup_{L=2}^{\infty} \bigcup_{(W_0,W_1,\dots,W_{L+1})\in\mathbb{N}^{L+1}} \left(\bigtimes_{\ell=0}^{L} (\mathbb{R}^{W_{\ell+1}\times W_\ell} \times \mathbb{R}^{W_{\ell+1}})\right) \tag{2.2}$$

be the set of all fully connected neural networks *(FCNNs). We call $\sigma \in \mathcal{C}(\mathbb{R},\mathbb{R})$ the* activation function *and for every $n \in \mathbb{N}$ let $\sigma_n \in \mathcal{C}(\mathbb{R}^n,\mathbb{R}^n)$ be the function satisfying $\sigma_n(x) = (\sigma(x_1),\dots,\sigma(x_n))^\intercal$ for every $x \in \mathbb{R}^n$. Let $\mathcal{P}\colon \mathcal{N} \to \mathbb{N}$, $\mathbb{L} : \mathcal{N} \to \mathbb{N}$, $\mathcal{R}\colon \mathcal{N} \to \mathcal{C}(\mathbb{R}^{W_0},\mathbb{R}^{W_{L+1}})$ be the* number of nonzero parameters*,* number of layers *and the* realization*, respectively, which satisfy for all $L, W_0,\dots,W_{L+1} \in \mathbb{N}$, $\phi = ((A_0,b_0),\dots,(A_L,b_L)) \in \times_{\ell=0}^{L}(\mathbb{R}^{W_{\ell+1}\times W_\ell} \times \mathbb{R}^{W_{\ell+1}})$ and for all $x \in \mathbb{R}^{W_0}$*

$$x_0 = x, \tag{2.3a}$$
$$x_{\ell+1} = \sigma_{W_{\ell+1}}(A_\ell x_\ell + b_\ell), \qquad \ell = 0,\dots,L-1, \tag{2.3b}$$
$$\mathcal{R}\phi(x) = A_L x_L + b_L. \tag{2.3c}$$

If not specified otherwise, we let $\sigma\colon \mathbb{R} \to \mathbb{R}, x \mapsto \max\{0,x\}$ be the ReLU activation function.

For a simpler notation, we also make use of the set of FCNNs of a fixed number of layers and fixed maximal numbers of parameters per layer.

Definition 2.11 (Fully connected networks with fixed width and depth)**.** *Let* $n_0, n_1, W, L \in \mathbb{N}$ *and*

$$\begin{aligned}\mathcal{N}_{n_0,n_1}(W,L) \coloneqq \big\{((A_0,b_0),\dots,(A_L,b_L)) \in (\bigtimes_{\ell=0}^{L}(\mathbb{R}^{W_{\ell+1}\times W_\ell}\times\mathbb{R}^{W_{\ell+1}}))\\ : W_0 = n_0, W_{L+1} = n_1, W_\ell \le W \ \forall\ \ell\in\{1,\dots,L\}\big\}\end{aligned} \tag{2.4}$$

be the set of fully connected neural networks with fixed width and depth *with* n_0 *inputs,* n_1 *outputs, and a maximum of* W *neurons in each layer* $\ell \in \{1,\dots,L\}$*.*

3 ResNet-like architectures

We define neural network realizations that resembles residual neural networks as introduced in [32] including multiple skip connections from the input to intermediate results, see Figure 1.1 for an illustration. This architecture allows to efficiently approximate LMC and performs well in our numerical experiments.

Definition 3.1 (ResNet-like realization)**.** *Let* $K, n \in \mathbb{N}$ *and let* $\Phi \coloneqq \{\phi_i\}_{i=1}^K \subset \mathcal{N}$ *with* $W_0 = W_{L+1} = n$ *fixed for every network. A* ResNet-like realization $\tilde{\mathcal{R}}\Phi \in \mathcal{C}(\mathbb{R}^n \times \bigtimes_{i=1}^N \mathbb{R}^n, \mathbb{R}^n)$ *is defined for* $x \in \mathbb{R}^n$ *and* $y = (y_1,\dots,y_K) \in \bigtimes_{i=1}^K \mathbb{R}^n$ *by*

$$x_0 \coloneqq x, \tag{3.1a}$$

$$x_i \coloneqq x_{i-1} + \mathcal{R}\phi_i(x_{i-1}) + y_i \quad \text{for } i = 1,\dots,K, \tag{3.1b}$$

$$\widetilde{\mathcal{R}}\Phi(x,y) \coloneqq x_K. \tag{3.1c}$$

The ResNet-like architecture is defined in a way such that it emulates a perturbed version of a Langevin Monte Carlo process, where the drift in the k-th step is replaced by the realization of a FCNN $\mathcal{R}\phi_k$. The following definition makes this perturbed process concrete.

Definition 3.2 (Stochastic process driven by Φ)**.** *Let* $K \in \mathbb{N}$*,* $\phi_1,\dots,\phi_K \colon \mathbb{R}^d \to \mathbb{R}^d$ *be Lipschitz-continuous and* $\Phi = \{\phi_i\}_{i=1}^K$*. Let* $Y^\Phi \colon [0, Kh] \times \Omega \to \mathbb{R}^d$ *be the stochastic process defined by*

$$Y_t^\Phi = Y_0 + \int_0^t \phi_{1+\frac{1}{h}\chi_h(s)}\left(Y^\Phi_{\chi_h(s)}\right)\mathrm{d}s + \sqrt{2}W_t. \tag{3.2}$$

In this case Y^Φ *is called a* stochastic process driven by Φ*. We denote the law of the process* Y^Φ *by* μ_t^Φ*. When considering a set of neural networks* $\{\phi_1,\dots,\phi_K\}$*, we denote the stochastic process driven by the realizations* $\{\mathcal{R}\phi_k,\dots,\mathcal{R}\phi_K\}$ *by* Y^Φ*, suppressing the realizations for the sake of brevity.*

To prepare for approximations of LMC with neural networks, we first show that the previously defined ResNet-like realizations in Definition 3.1 are the appropriate architecture to represent stochastic processes driven by neural networks in the sense of Definition 3.2.

Proposition 3.3. *Let* $K \in \mathbb{N}$ *and* $\Phi = \{\phi_i\}_{i=1}^K \subset \mathcal{N}$*. Let* $\xi = (\xi_1,\dots,\xi_K)$*, where* ξ_i *are the Brownian increments as defined in Section 2. Let* $Y_0 \sim \mu_0$ *and let* Y^Φ *be a stochastic process driven by* Φ *in the sense of Definition 3.2. Then, there exists* K *neural networks* $\psi_1,\dots,\psi_K$ *such that* ψ_i *has the same width and number of layers as* ϕ_i *for* $i = 1,\dots,K$*, and* $\Psi = \{\psi\}_{i=1}^K$ *such that*

$$Y_{Kh}^\Phi = \tilde{\mathcal{R}}\Psi(Y_0,\xi). \tag{3.3}$$

4 Perturbed Langevin Monte Carlo

In this section, we derive a perturbation result in Wasserstein-2 distance for the standard Langevin process, when the drift is replaced by approximations with small global L^2-error w.r.t. the current measure. This is more or less a direct consequence of Theorem 2.4 and the proof follows similar arguments. Upper bounds for the convergence of the unperturbed Langevin process $\tilde{X}$ can be found in [16].

Theorem 4.1 (Perturbed LMC). *Assume $\varepsilon > 0$, $h \in (0, \frac{2}{M})$ and V satisfies Assumption 1.1. Let $Y_0 \sim \mu_0$, $K \in \mathbb{N}$ and $\Phi = \{\phi_i : \mathbb{R}^d \to \mathbb{R}^d\}_{i=1}^K$. Let Y^Φ be the stochastic process driven by Φ and assume that for $i = 0, \dots, K-1$,*

$$\| -\nabla V - \phi_{i+1} \|_{L^2_{\mu^\Phi_{ih}}(\mathbb{R}^d;\mathbb{R}^d)} < \varepsilon \tag{4.1}$$

is satisfied. Then the law μ_t^Φ of the process Y^Φ satisfies

- *If $h \leq \frac{2}{m+M}$ then $\mathcal{W}_2(\mu_\infty, \mu^\Phi_{Kh}) \leq (1-mh)^K \mathcal{W}_2(\mu_\infty, \mu_0) + \frac{7\sqrt{2}}{6}\frac{M}{m}\sqrt{hd} + \frac{1-(1-mh)^K}{m}\varepsilon$.*
- *If $h \geq \frac{2}{m+M}$ then $W_2(\mu_\infty, \mu^\Phi_{Kh}) \leq (Mh-1)^K \mathcal{W}_2(\mu_\infty, \mu_0) + \frac{7\sqrt{2}}{6}\frac{Mh}{2-Mh}\sqrt{hd} + \frac{1-(Mh-1)^K}{2-Mh}h\varepsilon$.*

5 Approximation of Langevin Monte Carlo under linear error growth constraints

The analysis in this section is inspired by [1], where bounds on the sub-Gaussian variance proxy of the invariant measure of the LMC (2.1b) are derived. We use similar arguments based on the connection between sub-Gaussianity and Lyapunov functions established in Lemma 2.9.

5.1 Sub-Gaussianity of perturbed Langevin Monte Carlo

First, we treat the standard LMC process (2.1b), without any approximation of the drift and without requiring any additional assumption. Instead of only bounding the variance proxy of the invariant measure, we derive a bound for all intermediate measures $\mu^{\tilde{X}}_{kh}$. In the limit of steps $k \to \infty$, our result coincides with the one obtained in [1] for the invariant measure.

Proposition 5.1 (Sub-Gaussianity of LMC). *Let $h \in (0, \frac{2}{M})$ and $\tilde{X}_0 \sim \mu_0$ be sub-Gaussian with variance proxy σ_0^2. Then, for $k \in \mathbb{N}$, $\tilde{X}_{kh}$ is sub-Gaussian with variance proxy*

$$\sigma_k^2 = 2h\frac{1-c^k}{1-c} + \sigma_0^2 c^k, \tag{5.1}$$

where $c = \max_{\rho \in \{m,M\}} |1-\rho h|$.

Note that in the limit $k \to \infty$, this estimate leads exactly to the bound in [1] for the invarient measure, i.e.,

$$\sigma_k^2 \longrightarrow \frac{2h}{1-c}, \tag{5.2}$$

since $c < 1$.

Remark 5.2. *Recall that the process $\tilde{X}$ for all $t \in [0, Kh] \setminus \mathbb{N}h$ can be written as*

$$\tilde{X}_t = \tilde{X}_{\chi_h(t)} - h\nabla V(\tilde{X}_{\chi_h(t)}) + \sqrt{2}(W_t - W_{\chi_h(t)}).$$

Therefore, since $\tilde{X}_{kh}$ is sub-Gaussian for all $k \in \mathbb{N}$ by Proposition 5.1, by linear interpolation $\tilde{X}_t$ is sub-Gaussian for all $t \in [0, Kh]$. Indeed, applying Lemma A.4 with $\sigma^2 = 2(t - \chi_h(t))$ leads to the variance proxy

$$\sigma_t^2 = 2(t - \chi_h(t)) + 2h\frac{1 - c^{\chi_h(t)/h}}{1 - c} + \sigma_0^2 c^{\chi_h(t)/h}$$

for $\tilde{X}_t$.

A main ingredient of the proof of Proposition 5.1 is the fact that the contractivity constant c can be factored out of the expectation of the Lyapunov function via Jensen's inequality. The basic idea of our analysis for the DNN driven LMC is to ensure that similar arguments can be applied for the perturbed process. The following theorem provides bounds on the variance proxies of the perturbed process under the condition that the global error of the neural network approximations grows at most linearly. The proof follows similar arguments as the one of Proposition 5.1. In particular, we again make frequent use of Jensen's inequality for concave functions. The additional assumption on the networks can be seen as a way to ensure that Jensen's inequality can be applied.

Proposition 5.3 (Sub-Gaussianity of DNN driven LMC)**.** *For $k \in \mathbb{N}$, let $\phi_k \colon \mathbb{R}^d \to \mathbb{R}^d$ and $\Phi = \{\phi_k\}_{k\in\mathbb{N}}$. Assume that there exist $\delta > 0$, $G < m$ such that*

$$\| - \nabla V(x) - \phi_k(x)\|_{\ell^2} \leq \delta + G\|x\|_{\ell^2}, \qquad \forall x \in \mathbb{R}, k \in \mathbb{N}. \tag{5.3}$$

Let $h \in (0, \frac{2}{m+M})$. Then, the stochastic process Y^Φ driven by Φ and given by Equation (3.2) with $Y_0^\Phi \sim \mu_0$ is sub-Gaussian for all t. In particular, at time kh, $k \in \mathbb{N}$, Y_{kh}^Φ is sub-Gaussian with variance proxy

$$\sigma_k^2 = \left(\frac{2h}{1 - (c + hG)} + h^2\delta^2\right)[1 - (c + hG)^k] + \sigma_0^2[c + hG]^k \leq \frac{2h}{1 - (c + hG)} + h^2\delta^2 + \sigma_0^2, \tag{5.4}$$

where $c = \max_{\rho\in\{m,M\}} |1 - \rho h|$.

5.2 Neural network driven LMC with approximate drift with global linear error growth

We can now prove a first theorem on approximations of the LMC process using ResNet-like ReLU networks. The goal is to approximate the gradient $-\nabla V$ in each step of the LMC process with an FCNN on the whole domain and apply Theorem 4.1. We use assumption Assumption 1.2 (i) on the existence of neural networks approximating the drift with a controlled linear growth. When considering the stochastic processes driven by these neural networks, the linear growth allows us to bound the variance proxy in every step according to Proposition 5.3. Due to the bounded variance proxies and the arbitrarily small error growth, the sub-Gaussianity of the intermediate measures ensures the desired global errors (4.1) of the drift approximations.

Theorem 5.4 (ResNet-like realization approximated LMC (I))**.** *We presuppose the conditions in Assumption 1.1 for the potential V and Assumption 1.2 (i) for the existence of FCNN approximations of $-\nabla V$ with parameters bounded by $N(d, \varepsilon, m, M)$. Let $h \in (0, \frac{2}{m+M})$ and μ_0 be sub-Gaussian*

with variance proxy $\sigma_0^2 > 0$*. Then, there exists for any* $K \in \mathbb{N}$ *and any* $\varepsilon > 0$ *a neural network* ψ *with number of parameters bounded by* $N(d, c(d)\varepsilon, m, M)$*, where* $c(d) \in \mathcal{O}(d^{-1})$ *such that the measure* μ^Ψ *of the ResNet-like realisation of* $\Psi := \{\psi_k := \psi\}_{k=1}^K$ *with input* (Y_0, ξ)*, i.e. the law of* $\tilde{\mathcal{R}}(Y_0, \xi)$ *satisfies*

$$\mathcal{W}_2(\mu_\infty, \mu^\Psi) \leq (1-mh)^K \mathcal{W}_2(\mu_\infty, \mu_0) + \frac{7\sqrt{2}}{6}\frac{M}{m}\sqrt{hd} + \frac{1-(1-mh)^K}{m}\varepsilon. \tag{5.5}$$

6 Approximation of Langevin Monte Carlo under local error constraints

In this section, we aim to derive an approximation result similar to Theorem 5.4 but replacing the restrictive assumption of linear error growth on the whole space with the assumption of arbitrarily good approximation on open balls, Assumption 1.2 (ii). In particular, we analyze the growth of the approximation domain with increasing number of steps K. Previously, we did not have to consider approximation domains due to the global nature of the approximation in Assumption 1.2 (i) and the uniformly bounded variance proxies of the resulting sub-Gaussian intermediate measures. However, we can not guarantee uniform bounds with local approximations. In fact, the result we derive in this section has the required approximation domain (and hence, the number of required parameters of the network) growing with the number of steps K. Finally, in Section 6.2, we give examples on complexity bounds for neural network approximation on balls, to arrive at a complete upper bound on the number of parameters needed for a neural network with ResNet-like realization to sample from the given target distribution.

6.1 Neural network driven LMC with approximate drift on bounded domain

We start our considerations by showing that we can extend FCNNs from a bounded domain to the whole space with arbitrary accuracy with respect to a sub-Gaussian measure with a fixed number of added parameters, if the bounded domain is large enough.

Proposition 6.1 (Approximation on $\mathbb{R}^d$)**.** *Let* $\varepsilon > 0$*,* V *fulfill Assumption 1.1 and let* μ *be a* sub-Gaussian measure *with variance proxy* σ^2*. Let* $r > 0$ *be given by*

$$r = \sqrt{2}d\sigma \ln\left(\frac{16(\|\nabla V(0)\|_{\ell^2}^2(1+d) + M^2d^2\sigma^2(8d+10))}{\varepsilon^4}\right)^{1/2}$$

and let ϕ_{L-2} *be a neural network such that for* $\Omega := B_r(0)$

$$\| -\nabla V - \mathcal{R}\phi_{L-2}\|_{L^2_\mu(\Omega;\mathbb{R}^d)} \leq \frac{\varepsilon}{\sqrt{2}}.$$

There exists a neural network ϕ_L *with two more layers and* $2d^2+2$ *additional weights such that*

$$\| -\nabla V - \mathcal{R}\phi_L\|_{L^2_\mu(\mathbb{R}^d;\mathbb{R}^d)} \leq \varepsilon$$

and $\|\mathcal{R}\phi_L\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^d)} \leq \|\nabla V\|_{L^\infty(\Omega;\mathbb{R}^d)}$.

We combine the result on the perturbed Langevin process with the extension of FCNNs from bounded domains to $\mathbb{R}^d$ in the following statement to arrive at error bounds for a FCNN driven process. The central ingredients for the construction of the ResNet-like neural network are summarized in the iterative procedure in Figure 6.1. As before, the architecture emulates an LMC process with sub-Gaussian laws μ_k in the k-th step. In Lemma A.11 we derive a radius r_k depending on the variance proxy σ_k of μ_k which guarantees that there is "sufficiently little" mass of μ_k outside the ball $B_{r_k}(0)$. This radius is used in Proposition 6.1 to construct a that approximates the drift with global ε-error and has bounded L^∞-norm. Finally, Theorem A.14 guarantees sub-Gaussianity of the next measure μ_{k+1} for which we can again a network with global ε-error. Application of Theorem 4.1 then yields the following result.

Theorem 6.2 (ResNet-like realization approximated LMC (II)). *We presuppose the conditions in Assumption 1.1 for the potential V and Assumption 1.2 (ii) as a feed-forward neural network approximation with parameters bounded by $N(d, r, \varepsilon/\sqrt{d}, M)$. Let $h \in (0, \frac{2}{M})$ and μ_0 be sub-Gaussian with variance proxy $\sigma_0^2 > 0$. Then for any $K \in \mathbb{N}$ there exist neural networks $\{\psi_k\}_{k=1}^K = \Psi$ with a summed number of parameters bounded by $K(N(d, r, \varepsilon, M) + 2d^2 + 2)$ where*

$$r \in \mathcal{O}\left(d^{7/4}\varepsilon^{-1}(d^{9/4}\varepsilon^{-1})^{3(1.5^K-1)}\right)$$

such that the law μ^Ψ of the ResNet-like realisation of Ψ with input (Y_0, ξ) (i.e. the random variable $\widetilde{\mathcal{R}}\Psi(Y_0, \xi)$) satisfies

- *If $h \le \frac{2}{m+M}$ then $\mathcal{W}_2(\mu_\infty, \mu^\Psi) \le (1-mh)^K \mathcal{W}_2(\mu_\infty, \mu_0) + \frac{7\sqrt{2}}{6}\frac{M}{m}\sqrt{hd} + \frac{1-(1-mh)^K}{m}\varepsilon$.*
- *If $h \ge \frac{2}{m+M}$ then $W_2(\mu_\infty, \mu^\Psi) \le (Mh-1)^K \mathcal{W}_2(\mu_\infty, \mu_0) + \frac{7\sqrt{2}}{6}\frac{Mh}{2-Mh}\sqrt{hd} + \frac{1-(Mh-1)^K}{2-Mh}h\varepsilon$.*

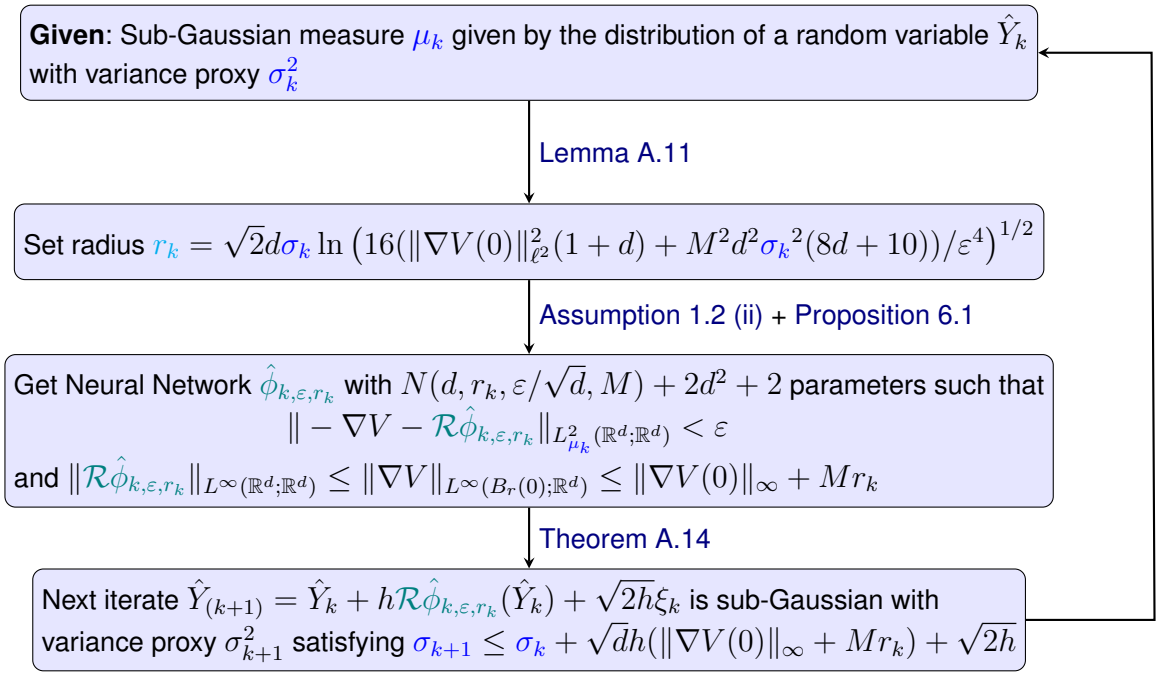

Figure 6.1: *Iterative procedure to build the sequences $(r_k)_{k=0}^{K-1} \subseteq [0, \infty)$, $(\hat{\phi}_{k,\varepsilon,r_k})_{k=0}^{K-1} \subseteq \mathcal{N}$ and $(\sigma_k)_{k=0}^{K-1} \subseteq (0, \infty)$ in the proof of Theorem 6.2.*

Note that the architecture of the neural networks in Assumption 1.2 (ii) as used in the last theorem is not specified.

Remark 6.3 (Activation function)**.** *We do not require the choice of a specific activation of the neural networks in Assumption 1.2. In the construction of the ResNet-like networks the last two layers use a ReLU activation due to the extension of the networks in local domains to the whole space in Proposition 6.1.*

6.2 FCNN approximation of the drift

In this section we derive estimates of $N(d, r, \varepsilon, M)$ such that $\phi_{\varepsilon,r}$ satisfies Assumption 1.2 (ii). The following is a result that was first proven by Yarotsky in [64] on the unit cube and extended later to cubes $[0, r]^d$, $r > 0$ [45]. We use the version for $[-r, r]^d$, $r > 0$, presented in [27].

Proposition 6.4 ([27, Theorem 1])**.** *Let $r > 0$ and $f\colon [-r, r]^d \to \mathbb{R}$ have modulus of continuity $\omega_f\colon [0, \infty) \to [0, \infty)$, i.e.*

$$|f(x) - f(y)| \leq \omega_f(\|x - y\|_\infty) \tag{6.1}$$

for all $x, y \in [-r, r]^d$. Then there exists a fully-connected ReLU network ϕ of width $2d+10$ and depth $L = \mathcal{O}(N)$, where N is the size of the network (i.e., the total number of neural network parameters), such that

$$\|f - \mathcal{R}\phi\|_{L^\infty([-r,r]^d)} \leq c_1(d)\omega_f(c_2(d)2rN^{-2/d}), \tag{6.2}$$

where $c_1(d)$ and $c_2(d)$ are positive constants depending on the input dimension d but independent of f and N.

Corollary 6.5 (Approximation on a ball)**.** *Assume that the potential V fulfills Assumption 1.1. Let $r > 0$, $x_0 \in \mathbb{R}^d$ and $\Omega := B_r^2(x_0)$. Then there exists a fully connected ReLU neural network $\tilde{\phi}$ of constant width $d(2d + 10)$ and depth $L = \mathcal{O}(N)$ where N is the size of the network (i.e., the total number of neural network parameters), such that*

$$\| - \nabla V - \mathcal{R}\tilde{\phi}\|_{L^\infty(\Omega;\mathbb{R}^d)} < c(d)MrN^{-2/d}, \tag{6.3}$$

where $c(d)$ is a positive constant depending on the input dimension d but independent of f and N. Hence, for any $\varepsilon > 0$ there exists a fully connected ReLU network with $N = (c(d)rM)^{d/2}2^{d/4}\varepsilon^{-d/2}$ parameters such that

$$\| - \nabla V - \mathcal{R}\tilde{\phi}\|_{L^\infty(\Omega,\mathbb{R}^d)} < \frac{\varepsilon}{\sqrt{2}}. \tag{6.4}$$

There are numerous other results on neural network approximation offering L^∞-approximation on compact domains, some of which do not (exclusively) use ReLU activation function. In [58] pointwise error bounds on the unit cube $[0, 1]^d$ are derived for Floor-ReLU networks. [59] derives similar bounds for FLES-nets with floor, exponential and step functions as activation functions. An overview of different expressivity results for neural networks can be found in [31].

7 Approximation of Langevin Monte Carlo under local error- and Lipschitz constraints

Under strong Lipschitz constraints on the potential gradient ∇V it is possible to derive uniform bounds on the sub-Gaussian variance proxies even when assuming only local approximation of the drift. In this section, we construct networks that fulfill both Assumption 1.2 (ii) and (5.3), under the additional

constraint that $M < \sqrt{2}m$, meaning that V can not be "too far away" from a quadratic function. In this section we will assume without loss of generality that the unique minimizer of V is given by $x^* = 0$ to simplify notation. First, we show that the drift can be approximated by a linear function with the required linear error growth.

Proposition 7.1 (Linear approximations of potential gradient with bounded error)**.** *Assume Assumption 1.1 with $m \leq M < \sqrt{2}m$ and assume that the unique minimizer of V is given by $x^* = 0$. Then for all $x \in \mathbb{R}^d$*

$$\| - \nabla V(x) - mx\|_{\ell^2} < \sqrt{M^2 - m^2}\|x\|_{\ell^2} < m\|x\|_{\ell^2}. \tag{7.1}$$

With this linear affine approximation we can prove our final theorem on approximations of the LMC process using ResNet-like ReLU networks. The goal is to approximate the gradient $-\nabla V$ in each step of the LMC process with an FCNN on the whole domain and apply Theorem 4.1. First, we use Proposition 7.1 and assumption Assumption 1.2 (ii) to construct appropriate neural networks approximating the drift arbitrarily well on a ball and with a controlled linear growth of the error outside of the ball. The key ideas are the following. First, by the bound on the Lipschitz constant M, the potential gradient $-\nabla V$ can be globally approximated by the linear function $x \mapsto mx$ with pointwise error growing at most linearly and with with slope m. According to our assumptions and results on the addition of neural networks, $-\nabla V - m\cdot$ can be approximated to arbitrary accuracy on any ball. With two additional ReLU layers the resulting network can be "cut off" so that it stays bounded even outside of the considered ball. The output of the network is thus confined to a hypercube, say $[-c, c]^d$ for some $c > 0$. The next step is to construct a network with the same output on the ball but with zero output outside of it. This can be done by approximating the multiplication of the cut-off network with an (approximate) indicator function constructed with ReLUs. Note that the cut-off is required for this approximation, because it ensures that all the inputs of the multiplication $(x, y) \mapsto xy$, i.e. the indicator function and the output of the cut-off network, live on compact sets. Adding to the resulting network again the neural network representation of the function $x \mapsto mx$ provides us with networks which satisfies (5.3) in addition to Assumption 1.2 (ii). In particular the error is controlled both on the ball ($\varepsilon/2$-accuracy) and outside of it (linear error growth). The ingredients for the construction of this neural network are visualized in Figure 7.1 for the one-dimensional case. The existence of such a network and its formal derivation are the subjects of the following proposition.

Proposition 7.2. *Presuppose Assumption 1.1, let $\varepsilon > 0$ and $r > 0$. Assume $m \leq M < \sqrt{2}m$. Furthermore, assume the existence of FCNN approximations as in Assumption 1.2 (ii) with ReLU activation function. Let where $\delta = 9\varepsilon/\sqrt{2d}$ and $G = \sqrt{M^2 - m^2}$. Then, there exists a FCNN ϕ with ReLU activation function and*

$$\mathcal{P}(\phi) = \mathcal{O}\left(d\log(2d\max\{1, rG\}/\varepsilon) + N(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) + dL(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) + 2d^2\right), \tag{7.2}$$

such that

$$\begin{aligned} &\| - \nabla V - \mathcal{R}\phi\|_{L^\infty(B_r^2(0))} \leq \varepsilon/\sqrt{2d}, \\ &\| - \nabla V(x) - \mathcal{R}\phi(x)\|_{\ell^2} \leq \delta + G\|x\|_{\ell^2}, \quad \forall x \in \mathbb{R}^d. \end{aligned}$$

In a nutshell, Proposition 7.2 states that Assumption 1.2 (ii) is sufficient to get global linear error bounds on the network approximations, provided that the Lipschitz constant M is not too far away from the convexity parameter m. This makes sense intuitively: The closeness of M and m implies that V is close to a quadratic function, and hence ∇V can be well approximated by a linear function, as Proposition 7.1 states. The important ingredient for proving the above FCNN result is the fact that this linear function can be represented by a ReLU network. More precisely, in the proof we construct

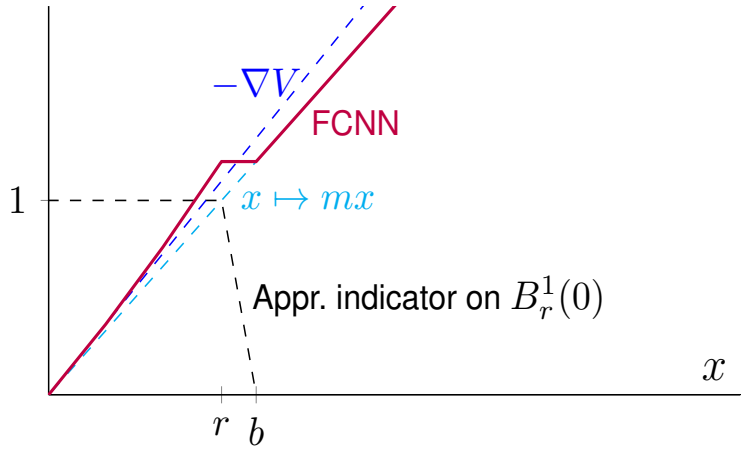

Figure 7.1: *Sketch of the construction of the network from Proposition 7.2. On the ℓ^1-ball of radius r, $B_r^1(0)$, the network approximates $-\nabla V$. On $B_b^1(0) \setminus B_r^1(0)$, where $b > r$, the network (approximately) interpolates linearly between $-\nabla V$ and $x \mapsto mx$. On $\mathbb{R}^d \setminus B_b^1(0)$, the network is identical to $x \mapsto mx$. In this way, global approximation with linearly growing error is achieved. For the precise construction and the choice of b, we refer to the proof of Proposition 7.2.*

a network that combines the local approximation on $B_r^2(0)$ granted by Assumption 1.2 (ii) with the global approximation property of the function $x \mapsto -mx$ on the complement of $B_r^2(0)$.

When considering the stochastic processes driven by these neural networks, the linear growth allows us to bound the variance proxy in every step, according to Proposition 5.3. Due the bounded variance proxies and the arbitrarily small errors on any ball, properties of the sub-Gaussian measures can be applied to obtain the desired global errors (4.1) of the drift approximations.

Theorem 7.3 (ResNet-like realization approximated LMC (III))**.** *We presuppose the conditions in Assumption 1.1 for the potential V and Assumption 1.2 (i) for the existence of FCNN approximations of $-\nabla V$ with parameters bounded by $N(d, r, \varepsilon, m, M)$. Furthermore, we assume $M < \sqrt{2}m$. Let $h \in (0, \frac{2}{m+M})$ and μ_0 be sub-Gaussian with variance proxy $\sigma_0^2 > 0$. Then, there exists for any $K \in \mathbb{N}$ and any $\varepsilon > 0$ a neural network ψ with number of parameters bounded by (7.2) with $r = \mathcal{O}(d(1+\ln(d^2\varepsilon^{-4})^{\frac{1}{2}}))$, such that the measure μ^Ψ of the ResNet-like realization of $\Psi := \{\psi_k := \psi\}_{k=1}^K$ with input (Y_0, ξ), i.e. the law of $\tilde{\mathcal{R}}(Y_0, \xi)$, satisfies*

$$\mathcal{W}_2(\mu_\infty, \mu^\Psi) \leq (1-mh)^K \mathcal{W}_2(\mu_\infty, \mu_0) + \frac{7\sqrt{2}}{6}\frac{M}{m}\sqrt{hd} + \frac{1-(1-mh)^K}{m}\varepsilon. \tag{7.3}$$

8 Experiments

With the numerical experiments in this section, our intention is to validate the error decay for a simple Gaussian distribution. Furthermore, we also consider a setting with a Gaussian mixture to illustrate that the method can assimilate more complex densities. The theoretical results show that if we can approximate the Langevin dynamics correctly, exponential converge to the posterior distribution can be expected Theorem 4.1. The experiments are designed to verify our results in practice.

Since in actual computations the evaluation of $\mathcal{W}_2(\mu_\infty, \mu_{kh}^{\tilde{X}})$ is intractable, an approximation is inevitable. The most common method is the *Sinkhorn algorithm* [14, 13].

Definition 8.1 (Entropy regularized optimal transport cost [13]). *Let $\mu, \nu \in \mathcal{D}_2(\mathbb{R}^d)$ be two probability measures on $\mathbb{R}^d$. Then, the* entropy regularized optimal transport cost *is defined as*

$$T_\lambda(\mu,\nu) := \min_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|_{\ell^2}^2 \mathrm{d}\gamma(x,y) + 2\lambda H(\gamma, \mu\otimes\nu),$$

where $\Pi(\mu,\nu)$ is the set of transport plans between μ and ν, $\lambda \geq 0$ is the regularization parameter, and $H(\gamma, \mu\otimes\nu)$ is the entropy of γ with respect to $\mu\otimes\nu$.

Note that $T_0(\mu,\nu) = \mathcal{W}_2^2(\mu,\nu)$ but the choice $T_\lambda(\hat{\mu}_n, \hat{\nu}_n)$ is not optimal since it introduces a large bias. The proposed estimator is the *Sinkhorn divergence* defined by

$$S_\lambda(\mu,\nu) := T_\lambda(\mu,\nu) - \frac{1}{2}(T_\lambda(\mu,\mu) + T_\lambda(\nu,\nu)).$$

For more information about the *Sinkhorn divergence* and its computation, we refer to [13].

Setting In the experiments, the model is composed of 200 deep neural networks with 2 hidden layers and 32 neurons each. The model is trained on a dataset of 10,000 samples during 50 epochs with a batch size of 64, and the time horizon is $T = 4$. For the optimization, the Adam optimizer [35] with a learning rate 5×10^{-4} is used.

8.1 Gaussian distribution

This experiment is meant to show that the model is able to approximate the Langevin dynamics for a Normal distribution defined by

$$\mu_\infty = \mathcal{N}_d(m, \Sigma),$$

where $m = 2((-1)^{i-1}(i-1))_{i=1}^d$, $\Sigma = I_d$ and $d = 10$.

As mentioned in the introduction of this section, in order to check if our model approximates the Langevin dynamics correctly, we have to (approximately) compute the Wasserstein distance $\mathcal{W}_2(\mu_\infty, \mu_{kh}^{\bar{X}})$.

The plot Figure 8.1 shows that the model is able to perform as well as the Euler-Maruyama scheme and that the costs decrease exponentially as expected according to Theorem 4.1.

8.2 Gaussian mixture

Although we have no theoretical convergence results for the approximation of the Langevin dynamics with Gaussian mixtures, intuitively it can be assumed that our approach also works in this case, which is illustrated with the next experiment. Here, the standard normal distribution is still the prior distribution and the posterior is chosen such that

$$\mathrm{d}\pi(x) = \left(\sum_{k=1}^C w_k p_k(x)\right)\mathrm{d}x,$$

where $(w_k)_{k=1}^C \in \mathbb{R}^+$ are the weights satisfying $|w|_1 = 1$ and $(p_k)_{k=1}^C$ are the densities of each normal component with mean μ_k and covariance matrix Σ_k. We confine the experiment to dimension

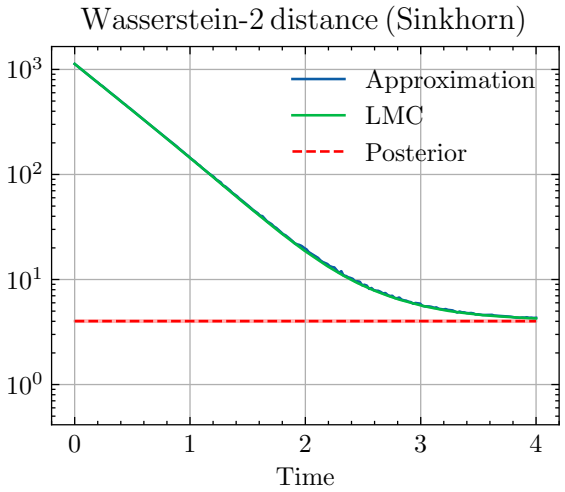

Figure 8.1: $\mathcal{W}_2^2(\mu_\infty, \mu_{kh}^{\tilde{X}})$ for the model and the Euler-Maruyama scheme. The standard deviation is also shown for each line. The regularization parameter is $\lambda = 10^{-2}$. Distances are averaged over 50 experiments.

2 with 2 components. The model is exactly the same as in the previous section. Again, the Wasserstein distance $\mathcal{W}_2(\mu_\infty, \mu_{kh}^{\tilde{X}})$ is computed. Figure 8.2 shows that the distance decreases exponentially and that we reach a plateau at time $t = 2$. This suggests that our results also work for non-strongly-convex potentials.

9 Conclusion

Expressivity results for ResNet-like neural networks mapping samples from a prior distribution to a smooth log-concave posterior distribution with arbitrary accuracy were derived. To that end, an upper bound for the decay of the Wasserstein-2 distance for the perturbed Langevin Monte Carlo process with approximate gradient steps was derived. Neural networks are used as an approximation for the drift in every step under different approximation assumptions.

In the first approach, global approximation of the drift with linear error growth is assumed. In this case, the variance proxies of the sub-Gaussian intermediate measures of the perturbed LMC process can be bounded uniformly (Proposition 5.3), and hence, the size of the ResNet-like neural network only depends linearly on the number of steps taken, see Theorem 5.4.

If the assumption on linear error growth is dropped and replaced with the assumption that the drift can be approximated arbitrarily well on a ball by an FCNN, the results are weaker. In particular, the bounds on the growth of the needed approximation domain in every step of the perturbed LMC depends exponentially on the number of steps taken. This can lead to an exponential growth of the number of parameters needed by a ResNet-like neural network in case the complexity on the FCNN depends on the size of the approximation domain, see Corollary 6.5 and Theorem 6.2.

Under an additional assumptions on the growth of the potential, i.e. that the Lipschitz constant of the gradient is smaller than $\sqrt{2}$ times the strong convexity constant, local approximation on a ball is sufficient to ensure a uniform bound for the variance proxies of the intermediate distributions of the FCNN process. Hence, robust error bounds, depending only linearly on the number of steps taken can be derived in this case. This is the result of Theorem 7.3.

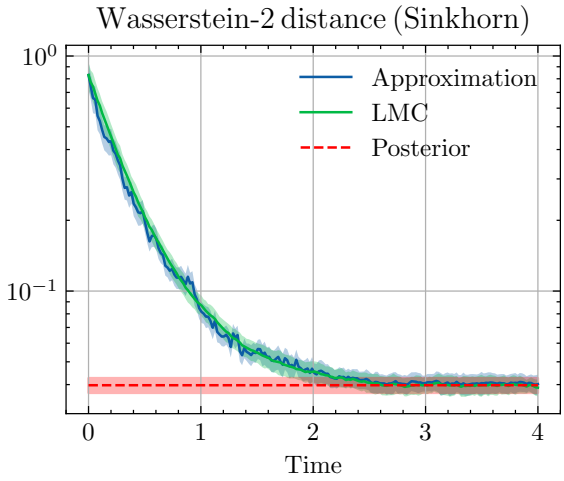

Figure 8.2: $\mathcal{W}_2^2(\mu_\infty, \mu_{kh}^{\tilde{X}})$ for the model and the Euler-Maruyama scheme. The standard deviation is also shown for each line. The regularization parameter is $\lambda = 10^{-2}$. The distances are averaged over 50 experiments.

The proposed architecture is tested on a Gaussian mixture and a Gaussian posterior distribution. We observe that the ResNet-like architecture with intermediate feed-forward networks of the same size show the same convergence as the Langevin process with the true drift, see Figure 8.1 and Figure 8.2. The architecture of the networks allows to train small networks in every step, which allows for short training processes.

Approximating Langevin dynamics with a ResNet-like neural network allows for an upper bound on the required steps for an effective posterior distribution approximation. Potentially, fewer network steps could suffice.

Acknowledgements

ME & JS acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the the priority programme SPP 2298 "Theoretical Foundations of Deep Learning". ME & CS & DS acknowledge support by the ANR-DFG project *COFNET: Compositional functions networks - adaptive learning for high-dimensional approximation and uncertainty quantification*. This study does not have any conflicts to disclose.

A Proofs of main results

A.2 Preliminaries: Sub-Gaussianity and Lyapunov functions

Proof of Proposition 2.7. By definition of a sub-Gaussian random vector, for any $u \in \mathbb{S}_1(0)$ and any $p \in \mathbb{N}$, we have

$$\mathbb{E}[|\langle u, Z\rangle|^p] \leq (\sqrt{2}\sigma)^p p\Gamma\left(p/2\right).$$

Now, let $q \geq 1$ be arbitrary and note that by Jensen's inequality we have

$$\|Z\|_{\ell^p} \leq \max(d^{1/p-1/q}, 1)\|Z\|_{\ell^q}.$$

Furthermore, for any $i \in \{1, \dots, d\}$ and $u = e_i$ we have $\mathbb{E}[|Z_i|^q] = \mathbb{E}[|\langle u, Z\rangle|^q] \leq (\sqrt{2}\sigma)^q q\Gamma\left(q/2\right)$. Taking the expectation leads to

$$\begin{aligned}\mathbb{E}[\|Z\|_{\ell^p}^q] &\leq \max(d^{q/p-1}, 1)\,\mathbb{E}[\|Z\|_{\ell^q}^q] \\ &\leq \max(d^{q/p}, d)(\sqrt{2}\sigma)^q q\Gamma\left(q/2\right) \\ &\leq (\sqrt{2}d\sigma)^q q\Gamma\left(q/2\right),\end{aligned}$$

Hence, $\|Z\|_{\ell^p}$ is sub-Gaussian. □

Proof of Lemma 2.9. If X is a sub-Gaussian random vector with variance proxy σ^2 then by Definition 2.6 and Definition 2.5 it holds that $\mathbb{E}\left[e^{s\langle v, X\rangle}\right] \leq e^{\frac{\sigma^2 s^2}{2}}$ for all $s \in \mathbb{R}$ and $v \sim \mathbb{S}_1(0)$. In particular it holds for any $\lambda > 0$ that

$$\mathbb{E}_X \mathcal{L}_\lambda(X) \leq e^{\frac{\sigma^2\lambda^2}{2}}.$$

The reverse is not obvious but does in fact also hold. Let σ^2 be such that $\mathbb{E}_X \mathcal{L}_\lambda(X) \leq e^{\frac{\sigma^2\lambda^2}{2}}$. Let $u \in \mathbb{S}_1$ be arbitrary and $\mathcal{R}$ denote the Haar measure over rotations in $\mathbb{R}^d$. Then it holds for $\lambda > 0$ that

$$\mathbb{E}_X \mathcal{L}_\lambda(X) = \mathbb{E}_X \mathbb{E}_{v\sim\mathbb{S}_1(0)}\left[e^{\lambda\langle v, X\rangle}\right] = \mathbb{E}_X \mathbb{E}_{R\sim\mathcal{R}}\left[e^{\lambda\langle Ru, X\rangle}\right] = \mathbb{E}_X \mathbb{E}_{R\sim\mathcal{R}}\left[e^{\lambda\langle u, RX\rangle}\right].$$

Hence,

$$\mathbb{E}_X \mathbb{E}_{R\sim\mathcal{R}}\left[e^{\lambda\langle u, RX\rangle}\right] \leq e^{\frac{\sigma^2\lambda^2}{2}} \qquad \text{for all } u \in \mathbb{S}_1,\ \lambda \in \mathbb{R},$$

where the inequality for negative λ follows since the negative sign can be absorbed into the rotation. By definition this implies that the random variable RX is sub-Gaussian with variance proxy σ^2. Let R^* denote the adjoint of the rotation R. It follows for any $r > 0$ and $u \in \mathbb{S}_1$ that

$$\mathbb{P}(|\langle u, X\rangle| \geq r) = \mathbb{P}(|\langle R^*(R^*)^{-1}u, X\rangle| \geq r) = \mathbb{P}(|\langle (R^*)^{-1}u, RX\rangle| \geq r) \leq \exp\left(-\frac{r^2}{2\sigma^2}\right),$$

since $(R^*)^{-1}u$ is again an element of $\mathbb{S}_1$. By definition this yields the sub-Gaussianity of X with variance proxy σ^2. □

A.3 ResNet-like architectures

Proof of Proposition 3.3. For $i = 1, \dots, K$, let $L_i, w_0, \dots, w_{L_i+1} \in \mathbb{N}$, and

$$((A_0^i, b_0^i), \dots, (A_{L_i}^i, b_{L_i}^i)) \in \bigtimes_{\ell=0}^{L_i} (\mathbb{R}^{w_{\ell+1} \times w_\ell} \times \mathbb{R}^{w_{n+1}})$$

be such that $\phi_i = ((A_0^i, b_0^i), \dots, (A_{L_i}^i, b_{L_i}^i))$ and let

$$\psi_i = ((A_0^i, b_0^i), \dots, (hA_{L_i}^i, hb_{L_i}^i)). \tag{A.1}$$

Then the ResNet-like realization of $\Psi = \{\psi_i\}_{i=1}^K$ satisfies for every $\omega \in \Omega$ with $x \coloneqq Y_0(\omega)$ that

$$\begin{aligned}
\widetilde{\mathcal{R}}\Psi(x, \xi(\omega)) &= x_{k-1} + \mathcal{R}\psi_k(x_{K-1}) + \xi_K(\omega) \\
&= x + \sum_{i=1}^K \mathcal{R}\psi_i(x_{i-1}) + \xi_i(\omega) \\
&= x + \sum_{i=1}^K h\mathcal{R}\phi_i(x_{i-1}) + \xi_i(\omega) \\
&= x + \int_0^{Kh} \mathcal{R}\phi_{1+\frac{1}{h}\chi_h(s)}(Y^\Phi_{\chi_h(s)}(\omega))\mathrm{d}s + \sqrt{2}W_{Kh}(\omega) = Y^\Phi_{Kh}(\omega).
\end{aligned} \tag{A.2}$$

□

A.4 Perturbed Langevin Monte Carlo

Proof of Theorem 4.1. Let $\rho\colon \left(0, \frac{2}{M}\right) \to [0, 1)$ be defined by

$$\rho(h) := \begin{cases} 1 - mh & \text{if } 0 < h < \frac{2}{m+M} \\ Mh - 1 & \text{if } \frac{2}{m+M} \le h < \frac{2}{M} \end{cases}.$$

Consider the triangle inequality

$$\mathcal{W}_2(\mu^\infty, \mu^\Phi_{Kh}) \le \mathcal{W}_2(\mu_\infty, \mu^{\tilde{X}}_{Kh}) + \mathcal{W}_2(\mu^{\tilde{X}}_{Kh}, \mu^\Phi_{Kh}).$$

By Theorem 2.4, the first term can be bounded by

$$\mathcal{W}_2(\mu_\infty, \mu^{\tilde{X}}_{Kh}) \le \rho(h)^K \mathcal{W}_2(\mu_\infty, \mu_0) + \begin{cases} \frac{7\sqrt{2}}{6}\frac{M}{m}\sqrt{hd}, & h \in \left(0, \frac{2}{m+M}\right] \\ \frac{7\sqrt{2}}{6}\frac{Mh}{2-Mh}\sqrt{hd}, & h \in \left[\frac{2}{m+M}, \frac{2}{M}\right) \end{cases}. \tag{A.3}$$

For $i = 0, \dots, K$, define $\Delta_i = \tilde{X}_{ih} - Y_{ih}$ and note that for $i = 0, \dots, K-1$ it holds that

$$\begin{aligned}
\Delta_{i+1} &= \tilde{X}_{ih} - Y_{ih} + h(-\nabla V(\tilde{X}_{ih}) - \phi_{i+1}(Y_{ih})) \\
&= \tilde{X}_{ih} - Y_{ih} - h\underbrace{(\nabla V(\tilde{X}_{ih}) - \nabla V(Y_{ih}))}_{:=u_i} + h\underbrace{(-\phi_{i+1}(Y_{ih}) - \nabla V(Y_{ih}))}_{:=v_i}
\end{aligned}$$

and

$$\mathbb{E}[\|\Delta_{i+1}\|_{\ell^2}^2]^{1/2} \le \mathbb{E}[\|\Delta_i - hu_i\|_{\ell^2}^2]^{1/2} + \mathbb{E}[\|hv_i\|_{\ell^2}^2]^{1/2}.$$

By the assumption that $\| -\nabla V - \phi_{i+1}\|_{L^2_{\mu^\Phi_{ih}}(\mathbb{R}^d;\mathbb{R}^d)} < \varepsilon$ for all $i = 0, \dots, K-1$, we have

$$\mathbb{E}[\|hv_i\|_{\ell^2}^2]^{\frac{1}{2}} \leq h\varepsilon.$$

Furthermore, we get for all $i = 0, \dots, K$ that

$$\begin{aligned}
\mathbb{E}[\|\Delta_i - hu_i\|_{\ell^2}^2]^{1/2} &\leq (1-mh)\mathbb{E}[\|\Delta_i\|_{\ell^2}^2]^{1/2}, \qquad h \in (0, 2/(m+M)]\,, \\
\mathbb{E}[\|\Delta_i - hu_i\|_{\ell^2}^2]^{1/2} &\leq (Mh-1)\mathbb{E}[\|\Delta_i\|_{\ell^2}^2]^{1/2}, \qquad h \in [2/(m+M), 2/M)\,,
\end{aligned}$$

by Lemma A.1. Combining these estimates, we arrive at

$$\begin{aligned}
\mathcal{W}_2(\mu_{Kh}^{\tilde{X}}, \mu_{Kh}^{\Phi}) &\leq \mathbb{E}[\|\Delta_K\|_{\ell^2}^2]^{1/2} \leq h\varepsilon + \rho(h)\mathbb{E}[\|\Delta_{K-1}\|_{\ell^2}^2]^{1/2} \\
&\leq \sum_{\ell=0}^{K-1} \rho(h)^\ell h\varepsilon + \rho(h)^K \mathbb{E}[\|\Delta_0\|_{\ell^2}^2]^{1/2} \\
&= \frac{1-\rho(h)^K}{1-\rho(h)} h\varepsilon + \rho(h)^K \mathbb{E}[\|\Delta_0\|_{\ell^2}^2]^{1/2}.
\end{aligned} \tag{A.4}$$

Now, since $\mu_0^{\tilde{X}} = \mu_0^{\Phi} = \mu_0$, the term $\mathbb{E}[\|\Delta_0\|_{\ell^2}^2]^{1/2}$ vanishes. Combining Equation (A.3) and Equation (A.4), we arrive at

$$\mathcal{W}_2(\mu_\infty, \mu_{Kh}^\Phi) \leq \rho(h)^K \mathcal{W}_2(\mu_\infty, \mu_0) + \begin{cases} \frac{7\sqrt{2}}{6}\frac{M}{m}\sqrt{hd} + \frac{1-(1-mh)^K}{m}\varepsilon, & h \in \left(0, \frac{2}{m+M}\right] \\ \frac{7\sqrt{2}}{6}\frac{Mh}{2-Mh}\sqrt{hd} + \frac{1-(Mh-1)^K}{2-Mh}h\varepsilon, & h \in \left[\frac{2}{m+M}, \frac{2}{M}\right) \end{cases}.$$

□

A.4.1 Auxiliary results

Lemma A.1 ([16, Lemma 1])**.** *Let* $\Delta_i := \tilde{X}_{ih} - Y_{ih}$ *and* $u_i := \nabla V(\tilde{X}_{ih}) - \nabla V(Y_{ih})$*. Let*

$$\gamma := \begin{cases} 1-mh & \text{if } h \leq 2/(m+M) \\ Mh-1 & \text{if } h \geq 2/(m+M) \end{cases}$$

which is in $(0,1)$ *since* $h \leq 2/M$ *by assumption. It holds that*

$$\mathbb{E}[\|\Delta_i - hu_i\|_{\ell^2}^2]^{1/2} \leq \gamma\, \mathbb{E}[\|\Delta_i\|_{\ell^2}^2]^{1/2}$$

A.5 Approximation of Langevin Monte Carlo under linear error growth constraints

A.5.1 Sub-Gaussianity of perturbed Langevin Monte Carlo

Proof of Proposition 5.1. We bound the Lyapunov-function of the LMC algorithm for a generic step. To simplify notation, we write X_k instead of $\tilde{X}_{kh}$ in this proof. Consider the process

$$X_k = X_{k-1} - h\nabla V(X_{k-1}) + \xi_k, \quad X_0 \sim \mu_0.$$

Let $c = \max_{\rho\in\{m,M\}} |1-\rho h|$ and note that $c < 1$. Hence, $x \mapsto x^c$ is a concave function and Jensen's inequality yields $\mathbb{E}[Z^c] \leq (\mathbb{E}[Z])^c$ for any random variable Z. We further assume without loss of generality that the unique minimizer of V is given by $x^* = 0$. Then, the Lyapunov function satisfies

$$\begin{aligned}\mathbb{E}_{\xi_k}[\mathcal{L}_\lambda(X_k)] &= e^{h\lambda^2}\mathcal{L}_\lambda(X_{k-1} - h\nabla V(X_{k-1})) \\ &\leq e^{h\lambda^2}\ell(c\lambda\|X_{k-1}\|_{\ell^2}) \leq e^{h\lambda^2}\ell(\lambda\|X_{k-1}\|_{\ell^2})^c \\ &= e^{h\lambda^2}\left(\mathcal{L}_\lambda(X_{k-1})\right)^c,\end{aligned}$$

where we have used the behaviour of the Lyapunov function under Gaussian convolution (Lemma A.4), the contractivity of the gradient descent step (Lemma A.7) and Jensen's inequality. Furthermore,

$$\begin{aligned}\mathbb{E}_{\xi_k,\xi_{k-1}}[\mathcal{L}_\lambda(X_k)] &\leq e^{h\lambda^2}\mathbb{E}_{\xi_{k-1}}\left[(\mathcal{L}_\lambda(X_{k-1}))^c\right] \\ &\leq e^{h\lambda^2}\left(\mathbb{E}_{\xi_{k-1}}\left[\mathcal{L}_\lambda(X_{k-1})\right]\right)^c\end{aligned}$$

by Jensen's inequality. Iteratively, this leads to

$$\mathbb{E}_{\xi_k,\xi_{k-1},\dots,\xi_1}[\mathcal{L}_\lambda(X_k)] \leq e^{h\lambda^2\sum_{i=0}^{k-1}c^i}\left(\mathcal{L}_\lambda(X_0)\right)^{c^k} = e^{h\lambda^2\frac{1-c^k}{1-c}}\left(\mathcal{L}_\lambda(X_0)\right)^{c^k}.$$

Finally, taking the expectation with respect to X_0 on both sides we find (again using Jensen's inequality) that

$$\mathbb{E}_{X_k}[\mathcal{L}_\lambda(X_k)] \leq e^{h\lambda^2\frac{1-c^k}{1-c}}\left(\mathbb{E}_{X_0}\left[\mathcal{L}_\lambda(X_0)\right]\right)^{c^k}.$$

Since $X_0 \sim \mu_0$ is sub-Gaussian with variance proxy σ_0^2, this yields

$$\mathbb{E}_{X_k}[\mathcal{L}_\lambda(X_k)] \leq \exp\left(h\lambda^2\frac{1-c^k}{1-c}\right)\exp\left(\frac{\sigma_0^2\lambda^2c^k}{2}\right) = \exp\left(h\lambda^2\frac{1-c^k}{1-c} + \frac{\sigma_0^2\lambda^2c^k}{2}\right).$$

By Lemma 2.9 it follows that X_k is a sub-Gaussian RV with variance proxy

$$\sigma_k^2 = 2h\frac{1-c^k}{1-c} + \sigma_0^2c^k.$$

□

Proof of Proposition 5.3. We simplify notation and consider the process $Y_k = Y_{k-1} + h\phi_k(Y_{k-1}) + \xi_k$. A similar analysis as in the proof of Proposition 5.1 leads to

$$\begin{aligned}\mathbb{E}_{\xi_k}\left[\mathcal{L}_\lambda(Y_k)\right] &= e^{h\lambda^2}\mathcal{L}_\lambda(Y_{k-1} + h\phi_k(Y_{k-1})) \\ &= e^{h\lambda^2}\mathcal{L}_\lambda(Y_{k-1} - h\nabla V(Y_{k-1}) + h\nabla V(Y_{k-1}) + h\phi_k(Y_{k-1})) \\ &= e^{h\lambda^2}\ell(\lambda\|Y_{k-1} - h\nabla V(Y_{k-1}) + h\nabla V(Y_{k-1}) + h\phi_k(Y_{k-1})\|_{\ell^2}) \\ &\leq e^{h\lambda^2}\ell(\lambda\|Y_{k-1} - h\nabla V(Y_{k-1})\|_{\ell^2} + \lambda\|h\nabla V(Y_{k-1}) + h\phi_k(Y_{k-1})\|_{\ell^2}) \\ &\leq e^{h\lambda^2}\ell(\lambda(c+hG)\|Y_{k-1}\|_{\ell^2} + \lambda h\delta),\end{aligned}$$

where we have used the behaviour of the Lyapunov function under Gaussian convolution (Lemma A.4), the fact that ℓ is monotonically increasing (Lemma A.2), assumption Equation (5.3) and the contractivity of the gradient descent step (Lemma A.7). Now, note that $(c+hG) < 1$ since for $h < \frac{2}{m+M}$,

$$c + hG = 1 - mh + hG < 1.$$

The fact that ℓ is a log-convex function (Lemma A.3) together with the upper bound $\ell(z) \leq \cosh(z)$ (Lemma A.6) and the fact that $\cosh(z) \leq e^{\frac{z^2}{2}}$ for all $z \in \mathbb{R}$ yields

$$\begin{aligned}\mathbb{E}_{\xi_k}\left[\mathcal{L}_\lambda(Y_k)\right] &\leq e^{h\lambda^2}\ell\left(\lambda\|Y_{k-1}\|\right)^{(c+hG)}\ell\left(\frac{\lambda h\delta}{1-(c+hG)}\right)^{1-(c+hG)} \\ &\leq \exp\left(h\lambda^2 + \frac{\lambda^2h^2\delta^2}{2(1-(c+hG))}\right)\mathcal{L}_\lambda\left(Y_{k-1}\right)^{(c+hG)}.\end{aligned}$$

Jensen's inequality further yields

$$\mathbb{E}_{\xi_k}\mathbb{E}_{\xi_{k-1}}\left[\mathcal{L}_\lambda(Y_k)\right] \leq \exp\left(h\lambda^2 + \frac{\lambda^2h^2\delta^2}{2(1-(c+hG))}\right)\left(\mathbb{E}_{\xi_{k-1}}[\mathcal{L}_\lambda\left(Y_{k-1}\right)]\right)^{(c+hG)}.$$

Iteratively it holds that

$$\begin{aligned}&\mathbb{E}_{\xi_k,\xi_{k-1},\dots,\xi_1}\left[\mathcal{L}_\lambda(Y_k)\right] \\ &\leq \exp\left(h\lambda^2 + \frac{\lambda^2h^2\delta^2}{2(1-(c+hG))}\right)^{\sum_{i=0}^{k-1}(c+hG)^i}\left(\mathcal{L}_\lambda\left(Y_0\right)\right)^{(c+hG)^k} \\ &= \exp\left(h\lambda^2 + \frac{\lambda^2h^2\delta^2}{2(1-(c+hG))}\right)^{\frac{1-(c+hG)^k}{1-(c+hG)}}\left(\mathcal{L}_\lambda\left(Y_0\right)\right)^{(c+hG)^k}.\end{aligned}$$

Finally, taking the expectation with respect to $Y_0 \sim \mu_0$, which is sub-Gaussian with variance proxy σ_0^2 and Jensen's inequality lead to

$$\begin{aligned}\mathbb{E}_{Y_k}\left[\mathcal{L}_\lambda(Y_k)\right] &\leq \exp\left(h\lambda^2 + \frac{\lambda^2h^2\delta^2}{2(1-(c+hG))}\right)^{\frac{1-(c+hG)^k}{1-(c+hG)}}\left(\mathbb{E}_{Y_0}[\mathcal{L}_\lambda\left(Y_0\right)]\right)^{(c+hG)^k} \\ &\leq \exp\left(\frac{1-(c+hG)^k}{1-(c+hG)}\left(h\lambda^2 + \frac{\lambda^2h^2\delta^2}{2(1-(c+hG))}\right) + \frac{\sigma_0^2\lambda^2[c+hG]^k}{2}\right)\end{aligned}$$

By Lemma 2.9, Y_k is sub-Gaussian with variance proxy

$$\sigma_k^2 = \left(\frac{2h}{1-(c+hG)} + h^2\delta^2\right)[1-(c+hG)^k] + \sigma_0^2[c+hG]^k. \tag{A.5}$$

As a consistency check, note that for $k \to \infty$ we have

$$\sigma_k^2 \longrightarrow \frac{2h}{1-c-hG} + h^2\delta^2,$$

which in the case of "perfect aproximation" with $\delta = 0$ and $G = 0$ leads to the known formula of $\frac{2h}{1-c}$ for the variance proxy of the invariant measure of LMC.

Since $c + hG < 1$, the sequence of sub-Gaussian proxies is bounded by

$$\sigma_k^2 \leq \underbrace{\left(\frac{2h}{1-(c+hG)} + h^2\delta^2\right)}_{\text{proxy of target dist.}} + \underbrace{\sigma_0^2}_{\text{proxy of initial dist.}}. \tag{A.6}$$

Sub-Gaussianity for all t can now be shown in exactly the same way as in Remark 5.2 for the standard LMC process. Recall that the process Y can be written for all $t \in [0, Kh] \setminus \mathbb{N}h$ as

$$Y_t = Y_{\chi_h(t)} + h\phi_{\chi_h(t)+1}(Y_{\chi_h(t)}) + \sqrt{2}(W_t - W_{\chi_h(t)}).$$

Therefore, since Y_{kh} is sub-Gaussian for all $k \in \mathbb{N}$, then by linear interpolation, Y_t is sub-Gaussian for all $t \in [0, Kh]$. Indeed, applying Lemma A.4 with $\sigma^2 = 2(t - \chi_h(t))$ leads to the variance proxy

$$\sigma_t^2 = 2(t - \chi_h(t)) + \left(\frac{2h}{1-(c+hG)} + h^2\delta^2\right)[1 - (c+hG)^{\chi_h(t)/h}] + \sigma_0^2[c+hG]^{\chi_h(t)/h}$$

for Y_t. □

A.5.2 Neural network driven LMC with approximate drift with global linear error growth

Proof of Theorem 5.4. Assumption 1.2 (i) guarantees for any $\delta_0 > 0$ the existence of a neural network ϕ_{δ_0} with $N(d, \delta_0, m, M)$ parameters such that

$$\| - \nabla - \phi_{\delta_0}\|_{\ell^2} \leq \delta_0(1 + \|x\|_{\ell^2})$$

for all $x \in \mathbb{R}^d$. Let $\phi \coloneqq \phi_\delta$, where

$$\delta = \varepsilon \left(1 + \left(2\pi d\sqrt{\frac{2}{m}} + \frac{4d^2}{m}\right)\left[4 + \frac{64}{m(m+M)^2} + m\sigma_0^2\right]\right)^{-1/2}.$$

Let $\Phi \coloneqq \{\phi\}_{k=0}^{K-1}$ and $Y^\Phi \colon \Omega \times [0, Kh] \to \mathbb{R}^d$ be the stochastic process driven by Φ, i.e.,

$$Y_t^\Phi = Y_0 + \int_0^t \mathcal{R}\phi_{\frac{1}{h}\chi_h(s)}\left(Y_{\chi_h(s)}^\Phi\right) \mathrm{d}s + \sqrt{2}W_t.$$

By Proposition 5.3, $Y_{kh}^\Phi \sim \mu_{kh}^\Phi$ is sub-Gaussian for all $k = 0, \dots, K$ with variance proxy σ_k^2 bounded by

$$\sigma_k^2 \leq \frac{2h}{1-(c+hG)} + h^2\delta^2 + \sigma_0^2.$$

We have

$$\begin{aligned} \| - \nabla V - \phi_{k+1}\|^2_{L^2_{\mu_{kh}^\Phi}(\mathbb{R}^d;\mathbb{R}^d)} &= \int_{\mathbb{R}^d} \| - \nabla V(x) - \phi_{k+1}(x)\|_{\ell^2}^2 \mathrm{d}\mu_{kh}^\Phi(x) \\ &\leq \int_{\mathbb{R}^d} \delta^2 + 2\delta^2\|x\|_{\ell^2} + \delta^2\|x\|_{\ell^2}^2 \, \mathrm{d}\mu_{kh}^\Phi(x) \\ &= \delta^2 + 2\delta^2\mathbb{E}_{Y_{kh}^\Phi}[\|Y_{kh}^\Phi\|_{\ell^2}] + \delta^2\mathbb{E}_{Y_{kh}^\Phi}[\|Y_{kh}^\Phi\|_{\ell^2}^2]. \end{aligned} \tag{A.7}$$

Furthermore, by Proposition 2.7, it holds for all $q \in \mathbb{N}$ that $\mathbb{E}_{Y_{kh}^\Phi}[\|Y_{kh}^\Phi\|_{\ell^2}^q] \leq (\sqrt{2}d\sigma_k)^q q\Gamma(q/2)$, leading to

$$\| - \nabla V - \phi_{k+1}\|^2_{L^2_{\mu_{kh}^\Phi}(\mathbb{R}^d;\mathbb{R}^d)} \leq \delta^2 + 2\delta^2\sqrt{2}d\sigma_k\Gamma(1/2) + 4\delta^2d^2\sigma_k^2\Gamma(1).$$

A simple calculation shows that this term is bounded from above by ε^2 for all k with the chosen δ (see Lemma A.8). Applying Theorem 4.1 for $h < \frac{2}{m+M}$, we get

$$\mathcal{W}_2(\mu_\infty, \mu_{Kh}^\Phi) \leq (1-mh)^K\mathcal{W}_2(\mu_\infty, \mu_0) + \frac{7\sqrt{2}}{6}\frac{M}{m}\sqrt{hd} + \frac{1-(1-mh)^K}{m}\varepsilon.$$

Finally, Proposition 3.3 guarantees the existence of a network ψ with number of parameters equal to the number of parameters of ϕ such that for $\Psi \coloneqq \{\psi\}_{k=0}^{K-1}$ it holds that $\mu_{Kh}^\Phi = \mu^\Psi$, where $\xi = (\xi_1, \dots, \xi_K)$ and $\mathcal{R}\Psi(Y_0, \xi) \sim \mu^\Psi$. This yields the claim. □

A.5.3 Auxiliray results: Lyapunov functions and contractivity

Lemma A.2 (Monotonicity of ℓ, [1, Lemma 3.4])**.** *For any $d \in \mathbb{N}$ the function ℓ in Definition 2.8 is monotonically increasing on $\mathbb{R}_{\geq 0}$.*

Lemma A.3 (Log-convexity of ℓ)**.** *For any $d \in \mathbb{N}$ the function ℓ is log-convex.*

Proof. By Hölder's inequality it holds that

$$\mathbb{E}(UV) \leq (\mathbb{E}|U|^p)^{1/p}(\mathbb{E}|V|^q)^{1/q}$$

for any $1 < p, q < \infty$ with $1/p + 1/q = 1$. Now, for $\theta \in (0,1)$ let $U = \exp(z_1\theta\langle v, e_1\rangle)$, $V = \exp(z_2(1-\theta)\langle v, e_1\rangle)$, $p = 1/\theta$, $q = 1/(1-\theta)$. Then

$$\begin{aligned}
\ell(\theta z_0 + (1-\theta)z_1) &= \mathbb{E}_{v\sim\mathbb{S}_1^2(0)}\left[e^{(\theta z_0 + (1-\theta)z_1)\langle v, e_1\rangle}\right] && \text{(A.8)}\\
&\leq \left(\mathbb{E}_{v\sim\mathbb{S}_1^2(0)}\left[e^{z_0\langle v, e_1\rangle}\right]\right)^{\theta}\left(\mathbb{E}_{v\sim\mathbb{S}_1^2(0)}\left[e^{z_1\langle v, e_1\rangle}\right]\right)^{1-\theta} && \text{(A.9)}\\
&= \ell(z_0)^{\theta}\ell(z_1)^{1-\theta}. && \text{(A.10)}
\end{aligned}$$

Taking the logarithm on both sides yields

$$\log \ell(\theta z_0 + (1-\theta)z_1) \leq \theta \log(\ell(z_0)) + (1-\theta)\log(\ell(z_1)), \tag{A.11}$$

showing log-convexity. □

Lemma A.4 (Behavior of the Lyapunov function under Gaussian convolution, see [1, Lemma 3.3])**.** *For any dimension $d \in \mathbb{N}$, point $x \in \mathbb{R}^d$, weight $\lambda > 0$ and noise variance σ^2 it holds that*

$$\mathbb{E}_{Z\sim\mathcal{N}(0,\sigma^2 I)}\left[\mathcal{L}_\lambda(x+Z)\right] = e^{\frac{\lambda^2\sigma^2}{2}}\mathcal{L}_\lambda(x). \tag{A.12}$$

In order to derive upper and lower bounds on the Lyapunov function, an explicit formula is useful.

Lemma A.5 (Explicit formula for the Lyapunov function, see [1, Lemma 3.2])**.** *For any dimensions $d \geq 2$ and argument $z > 0$ it holds that*

$$\ell(z) = \Gamma(\alpha+1)\cdot\left(\frac{2}{z}\right)^{\alpha}\cdot I_\alpha(z), \tag{A.13}$$

where $\alpha = (d-2)/d$, Γ is the Gamma function and I_n is the modified Bessel function of the first kind. For $d = 1$, it holds that $\ell(z) = \frac{1}{2}(e^{-z} + e^z) = \cosh(z)$.

The following bounds are shown in [40].

Lemma A.6 (Lower and upper bound of the Lyapunov function [40])**.** *For $z > 0$ and $\alpha > -\frac{1}{2}$ it holds that*

$$1 < \Gamma(\alpha+1)\cdot\left(\frac{2}{z}\right)^{\alpha}\cdot I_\alpha(z) < \cosh(z). \tag{A.14}$$

In particular it holds for all d and $z > 0$ that

$$1 \leq \ell(z) \leq \cosh(z). \tag{A.15}$$

Lemma A.7 (Contractivity of gradient descent step, [1, Lemma 4.2]). *Suppose V is m-strongly convex and has M-Lipschitz gradient and let $h \in (0, \frac{2}{M})$. Then it holds for all $x \in \mathbb{R}^d$ that*

$$\|x - h\nabla V(x) - x^*\|_{\ell^2} \leq c\|x - x^*\|_{\ell^2}, \tag{A.16}$$

where x^* is any minimizer of V and $c \coloneqq \max_{\rho\in\{m,M\}} |1 - \rho h| < 1$.

Lemma A.8 (Achieving ε-error with linear error growth assumption). *Let $\varepsilon > 0$. Let $\delta \leq \frac{m}{2}$ satisfy*

$$\delta \leq \varepsilon \left(1 + \left(2\pi d\sqrt{\frac{2}{m}} + \frac{4d^2}{m}\right)\left[4 + \frac{64}{m(m+M)^2} + m\sigma_0^2\right]\right)^{-\frac{1}{2}}$$

Let $\phi\colon \mathbb{R}^d \to \mathbb{R}^d$ satisfy

$$\|-\nabla V(x) - \phi(x)\|_{\ell^2} \leq \delta(1 + \|x\|_{\ell^2}) \tag{A.17}$$

for all $x \in \mathbb{R}^d$. Let $h \in (0, \frac{2}{m+M})$ and $\Phi = \{\phi_k = \phi\}_{k\in\mathbb{N}}$. Then, the stochastic process Y^Φ driven by Φ, given by Eq. (3.2), with $Y_0^\Phi \sim \mu_0$ has intermediate measures $Y_{ih}^\Phi \sim \mu_{ih}^\Phi$ satisfying for all $i \in \mathbb{N}$ that

$$\|-\nabla V - \phi_{i+1}\|_{L^2_{\mu_{ih}^\Phi}(\mathbb{R}^d;\mathbb{R}^d)} \leq \varepsilon. \tag{A.18}$$

Proof. We have

$$\begin{aligned}
\|-\nabla V - \phi_{i+1}\|^2_{L^2_{\mu_{ih}^\Phi}(\mathbb{R}^d;\mathbb{R}^d)} &= \int_{\mathbb{R}^d} \|-\nabla V(x) - \phi_{i+1}(x)\|^2_{\ell^2} \mathrm{d}\mu_{ih}^\Phi(x) \\
&\leq \int_{\mathbb{R}^d} \delta^2 + 2\delta^2\|x\|_{\ell^2} + \delta^2\|x\|^2_{\ell^2}\, \mathrm{d}\mu_{ih}^\Phi(x) \\
&= \delta^2 + 2\delta^2\mathbb{E}_{X\sim\mu_{ih}^\Phi}[\|X\|_{\ell^2}] + \delta^2\mathbb{E}_{X\sim\mu_{ih}^\Phi}[\|X\|^2_{\ell^2}].
\end{aligned} \tag{A.19}$$

Now, by Proposition 5.3 it holds that μ_{ih}^Φ is sub-Gaussian with variance proxy

$$\sigma_i \leq \frac{2h}{1 - (c + h\delta)} + h^2\delta^2 + \sigma_0^2. \tag{A.20}$$

Furthermore, by Proposition 2.7, it holds for all $q \in \mathbb{N}$ that $\mathbb{E}_{X\sim\mu_{ih}^\Phi}[\|X\|^q_{\ell^2}] \leq (\sqrt{2}d\sigma_i)^q q\Gamma(q/2)$. Hence,

$$\begin{aligned}
&\|-\nabla V - \phi_{i+1}\|^2_{L^2_{\mu_{ih}^\Phi}(\mathbb{R}^d;\mathbb{R}^d)} \leq \delta^2 + 2\delta^2\sqrt{2}d\sigma_i\Gamma(1/2) + \delta^2 4d^2\sigma_i^2\Gamma(1) \\
&\leq \delta^2 + 2\delta^2\sqrt{2}d\left[\frac{2h}{1 - (c + h\delta)} + h^2\delta^2 + \sigma_0^2\right]^{\frac{1}{2}}\Gamma(1/2) \\
&\quad + \delta^2 4d^2\left[\frac{2h}{1 - (c + h\delta)} + h^2\delta^2 + \sigma_0^2\right]\Gamma(1) \\
&= \delta^2 + 2\delta^2\sqrt{2}d\left[\frac{2h}{1 - (c + h\delta)} + h^2\delta^2 + \sigma_0^2\right]^{\frac{1}{2}}\sqrt{\pi} + \delta^2 4d^2\left[\left(\frac{2h}{1 - (c + h\delta)} + h^2\delta^2\right) + \sigma_0^2\right].
\end{aligned} \tag{A.21}$$

Note that $h < \frac{2}{m+M}$ implies $c = 1 - mh$. Since $\delta < \frac{m}{2}$, it holds that

$$\frac{2h}{1 - (c + h\delta)} = \frac{2}{m - \delta} < \frac{2}{m - \frac{m}{2}} = \frac{4}{m}.$$

Hence

$$
\begin{aligned}
&\| -\nabla V - \phi_{i+1}\|^2_{L^2_{\mu^\Phi_{ih}}(\mathbb{R}^d;\mathbb{R}^d)} \\
&\quad \leq \delta^2 + 2\delta^2\sqrt{\frac{2\pi}{m}}d\left[4 + mh^2\delta^2 + m\sigma_0^2\right]^{\frac{1}{2}}\sqrt{\pi} + \delta^2\frac{4}{m}d^2\left[4 + mh^2\delta^2 + m\sigma_0^2\right] \\
&\quad \leq \delta^2 + 2\delta^2 d\sqrt{\frac{2\pi}{m}}\left[4 + 16h^2/m + m\sigma_0^2\right]^{\frac{1}{2}}\sqrt{\pi} + \delta^2\frac{4}{m}d^2\left[4 + 16h^2/m + m\sigma_0^2\right] \\
&\quad = \delta^2\left(1 + 2\pi d\sqrt{\frac{2}{m}}\left[4 + 16h^2/m + m\sigma_0^2\right]^{\frac{1}{2}} + \frac{4d^2}{m}\left[4 + 16h^2/m + m\sigma_0^2\right]\right) \\
&\quad \leq \delta^2\left(1 + \left(2\pi d\sqrt{\frac{2}{m}} + \frac{4d^2}{m}\right)\left[4 + 16h^2/m + m\sigma_0^2\right]\right) \\
&\quad \leq \delta^2\left(1 + \left(2\pi d\sqrt{\frac{2}{m}} + \frac{4d^2}{m}\right)\left[4 + \frac{64}{m(m+M)^2} + m\sigma_0^2\right]\right) \\
&\quad \leq \varepsilon^2,
\end{aligned}
\tag{A.22}
$$

with the choice of δ in the lemma. □

A.6 Approximation of Langevin Monte Carlo under local error constraints

Proof of Proposition 6.1. We start by constructing a bounded neural network ϕ_L as in Lemma A.9, which does not change the approximation accuracy on the domain Ω. The neural network is bounded componentwise on $\mathbb{R}^d$ by $c \in \mathbb{R}^d$ with $c_i := \|\nabla V_i\|_{L^\infty(B_r(0))}$. It then holds that

$$
\| -\nabla V - \mathcal{R}\phi_L\|^2_{L^2_\mu(\mathbb{R}^d;\mathbb{R}^d)} \leq \| -\nabla V - \mathcal{R}\phi_{L-2}\|^2_{L^2_\mu(\Omega;\mathbb{R}^d)} + 2(\|\nabla V\|^2_{L^2_\mu(\mathbb{R}^d\setminus\Omega;\mathbb{R}^d)} + \|c\|^2_{L^2_\mu(\mathbb{R}^d\setminus\Omega;\mathbb{R}^d)}).
$$

The first part is bounded by assumption. Using the sub-Gaussian property and the Lipschitz continuity of ∇V, the second term can be bounded (see Lemma A.10) by

$$
2\int_{\mathbb{R}^d\setminus\Omega} \|\nabla V\|^2_{\ell^2} + \|c\|^2_{\ell^2}\mathrm{d}\mu \leq \left[4M^2(2d^2\sigma^2 + r^2) + 2(2\|\nabla V(x_0)\|^2_{\ell^2} + \|c\|^2_{\ell^2})\right]\exp\left(-\frac{r^2}{2d^2\sigma^2}\right).
$$

Finally, Lemma A.11 yields

$$
2(\|\nabla V\|^2_{L^2_\mu(\mathbb{R}^d\setminus\Omega;\mathbb{R}^d)} + \|c\|^2_{L^2_\mu(\mathbb{R}^d\setminus\Omega;\mathbb{R}^d)}) \leq \varepsilon^2/2
$$

for the chosen size of the domain $r > 0$. □

Proof of Theorem 6.2. We construct the network iteratively over the steps of the LMC algorithm. Starting with $k = 0$, we have a measure μ_0 with variance proxy σ_0^2. Now, let

$$
r_0 = \sqrt{2}d\sigma_0\ln\left(\frac{16(\|\nabla V(0)\|^2_{\ell^2}(1+d) + M^2d^2\sigma_0^2(8d+10))}{\varepsilon^4}\right)^{1/2}.
$$

Assumption 1.2 (ii) guarantees the existence of a DNN $\phi_{0,\varepsilon,t}$ with $N(d, r_0, \varepsilon/\sqrt{d}, M)$ parameters such that

$$
\| -\nabla V - \mathcal{R}\phi_{0,\varepsilon,t_0}\|_{L^2_{\mu_0}(B_r(0);\mathbb{R}^d)} \leq \sqrt{d}\| -\nabla V - \mathcal{R}\phi_{0,\varepsilon,t_0}\|_{L^\infty(B_r(0);\mathbb{R}^d)} \leq \varepsilon/\sqrt{2}. \tag{A.23}
$$

With this, Proposition 6.1 guarantees existence of a network $\hat{\phi}_{0,\varepsilon,t_0}$ with $N(d, r_0, \varepsilon/\sqrt{d}, M)+2d^2+2$ parameters such that

$$\| -\nabla V - \mathcal{R}\hat{\phi}_{0,\varepsilon,r_0}\|_{L^2_{\mu_0}(\mathbb{R}^d;\mathbb{R}^d)} \leq \varepsilon.$$

Now, let $\hat{Y}_0 \sim \mu_0$ and $\hat{Y}_h := \hat{Y}_0 + h\mathcal{R}\hat{\phi}_{0,\varepsilon,r_0}(\hat{Y}_0) + \sqrt{2h}\xi_1$. By Theorem A.14, $\hat{Y}_h$ is again a sub-Gaussian random variable with measure denoted μ_h and variance proxy σ_1^2 satisfying

$$\sigma_1 \leq \sigma_0 + \sqrt{d}h(\|\nabla V(0)\|_\infty + Mr_0) + \sqrt{2h}.$$

All of the above steps can be repeated for this measure, yielding r_1 according to Proposition 6.1 and a network $\hat{\phi}_{1,\varepsilon,r_1}$ with $N(d, r_1, \varepsilon, M) + 2d^2 + 2$ parameters such that

$$\| -\nabla V - \mathcal{R}\hat{\phi}_{1,\varepsilon,r_1}\|_{L^2_{\mu_1}(\mathbb{R}^d;\mathbb{R}^d)} \leq \varepsilon.$$

Iteratively, this procedure (summarized in Figure 6.1) yields sequences $(r_k)_{k=0}^{K-1} \subseteq [0,\infty)$, $(\hat{\phi}_{k,\varepsilon,r_k})_{k=0}^{K-1} \subseteq \mathcal{N}$ and $(\sigma_k)_{k=0}^{K-1} \subseteq (0,\infty)$ such that for $k = 0, \dots, K-1$

- $\hat{Y}_{k+1} := \hat{Y}_k + h\mathcal{R}\hat{\phi}_{k,\varepsilon,r_k}(\hat{Y}_k) + \sqrt{2h}\xi_{k+1}$ is sub-Gaussian with variance proxy σ_{k+1} and measure denoted by μ_{k+1}.
- $\hat{\phi}_{k,\varepsilon,r_k}$ has $N(d, r_k, \varepsilon, M) + 2d^2 + 2$ parameters and satisfies

$$\| -\nabla V - \mathcal{R}\hat{\phi}_{k,\varepsilon,r_k}\|_{L^2_{\mu_k}(\mathbb{R}^d;\mathbb{R}^d)} \leq \varepsilon.$$

From now on, let $\phi_k := \hat{\phi}_{k,\varepsilon,r_k}$ and $\Phi := \{\phi_k\}_{k=0}^{K-1}$. Let $Y^\Phi \colon \Omega \times [0, Kh] \to \mathbb{R}^d$ be the stochastic process driven by Φ, i.e.

$$Y_t^\Phi = Y_0 + \int_0^t \mathcal{R}\phi_{\frac{1}{h}\chi_h(s)}\left(Y^\Phi_{\chi_h(s)}\right) \mathrm{d}s + \sqrt{2}W_t.$$

By Theorem 4.1, the law μ_t^Φ of this process satisfies

- If $h \leq \frac{2}{m+M}$ then $\mathcal{W}_2(\mu_\infty, \mu^\Phi_{Kh}) \leq (1-mh)^K \mathcal{W}_2(\mu_\infty, \mu_0) + \frac{7\sqrt{2}}{6}\frac{M}{m}\sqrt{hd} + \frac{1-(1-mh)^K}{m}\varepsilon.$
- If $h \geq \frac{2}{m+M}$ then $W_2(\mu_\infty, \mu^\Phi_{Kh}) \leq (Mh-1)^K \mathcal{W}_2(\mu_\infty, \mu_0) + \frac{7\sqrt{2}}{6}\frac{Mh}{2-Mh}\sqrt{hd} + \frac{1-(Mh-1)^K}{2-Mh}h\varepsilon.$

Furthermore, Proposition 3.3 guarantees the existence of networks $\psi_0, \dots, \psi_{K-1}$ with the number of parameters of ψ_i equal to the number of parameters of ϕ_i for all $k = 0, \dots, K-1$ such that for $\Psi := \{\psi_k\}_{k=0}^{K-1}$ it holds that $\mu^\Phi_{Kh} = \mu^\Psi$, where $\xi = (\xi_1, \dots, \xi_K)$ and $\widetilde{\mathcal{R}}\Psi(Y_0, \xi) \sim \mu^\Psi$. To prove the theorem, it remains to show the bound for the complexity of Ψ. First, note that the complexity of Ψ is given by $\sum_{k=0}^{K-1}[N(d, r_k, \varepsilon/\sqrt{d}, M) + 2d^2 + 2]$, which is upper bounded by $K[N(d, r_{K-1}, \varepsilon/\sqrt{d}, M) + 2d^2 + 2]$, since $(r_k)_{k=0}^{K-1}$ is a monotonically increasing sequence. The proof is then completed by controlling the growth of r_{K-1} with respect to K, for which we refer to Proposition A.17. □

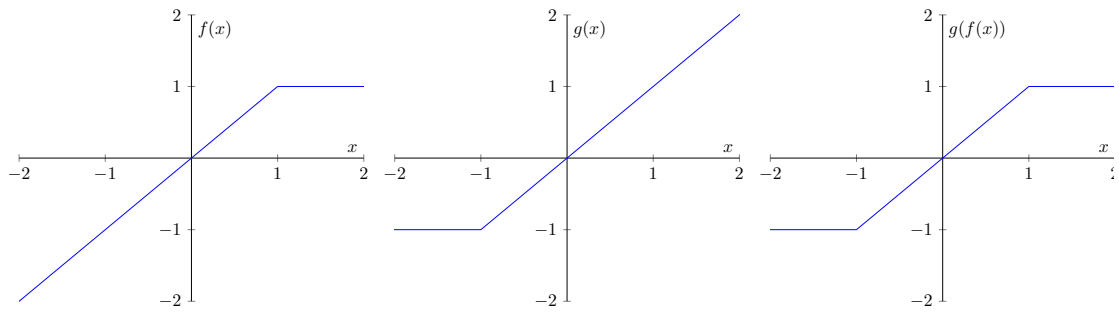

Figure A.1: *Visualization of the bounding of the neural network used in Lemma A.9 in one dimension. Here,* $f(x) = -\sigma(-x+1)+1$ *defines a bound from above by* 1 *and* $g(x) = \sigma(x+1)-1$ *defines a bound from below by 1. Note that* $g \circ f$ *is the identity on* $[-1,1]$ *and bounded by* ± 1 *on* $\mathbb{R}$*. The bounding of the network in Lemma A.9 corresponds to an application of similar functions in every dimension.*

A.6.1 Auxiliary results: Network cut-off, domain growth, asymptotics for r and σ

Lemma A.9. *Let* $p \in [1,\infty]$*,* μ *be a measure on* $\mathbb{R}^d$*,* $\Omega \subset \mathbb{R}^d$ *and let* ∇V *be bounded on* Ω*. Let* ϕ_{L-2} *be a neural network. There exists a neural network* ϕ_L *entrywise bounded by* $c \in \mathbb{R}^d$ *with* $c_i \coloneqq \|\nabla V_i\|_{L^\infty(\Omega;\mathbb{R}^d)}$ *with two more layers than* ϕ_{L-2} *and* $2d^2+2$ *additional weights such that*

$$\|-\nabla V - \mathcal{R}\phi_L\|^p_{L^p_\mu(\Omega;\mathbb{R}^d)} \leq \|-\nabla V - \mathcal{R}\phi_{L-2}\|^p_{L^p_\mu(\Omega,\mathbb{R}^d)} \quad \text{and}$$
$$\|-\nabla V - \mathcal{R}\phi_L\|^p_{L^p_\mu(\mathbb{R}^d;\mathbb{R}^d)} \leq \|-\nabla V - \mathcal{R}\phi_{L-2}\|^p_{L^p_\mu(\Omega,\mathbb{R}^d)} + 2^{p-1}(\|\nabla V\|^p_{L^p_\mu(\mathbb{R}^d\setminus\Omega;\mathbb{R}^d)} + \|c\|^p_{L^p_\mu(\mathbb{R}^d\setminus\Omega;\mathbb{R}^d)}).$$

Proof. We construct a bounded neural network ϕ_L, which does not change the approximation on Ω. In the proof a network ϕ and its realization $\mathcal{R}\phi$ will both be denoted by ϕ to avoid overloading notation.

Let $c_i \coloneqq \|\nabla V_i\|_{L^\infty(\Omega)} \geq 0$. Define the neural networks $\phi_{L-1}(x) = -\sigma(-\phi_{L-2}(x)+c)+c$ and $\phi_L(x) = \sigma(\phi_{L-1}(x)+c)-c$.

We show that $\|\phi_L(x)_i\|_{L^\infty(\Omega)} \leq c_i$ for all $x \in \mathbb{R}^d$ by contradiction.

- First assume that there exists $i \in [d]$ and $x \in \mathbb{R}^d$ such that $\phi_L(x)_i > c_i$. Then $\sigma(\phi_{L-1}(x)_i + c_i) > 2c_i \geq 0$ implying $\phi_{L-1}(x)_i + c_i = \sigma(\phi_{L-1}(x)_i + c_i) > 2c_i$ and therefore $\phi_{L-1}(x)_i > c_i$ and therefore $-\sigma(-\phi_{L-2}(x)_i + c_i) > 0$. This is a contradiction to σ being non-negative. Therefore, $\phi_L(x)_i \leq c_i$ holds.

- On the other hand, assume that there exist $i \in [d]$ and $x \in \mathbb{R}^d$ such that $\phi_L(x) < -c_i$. Then $\sigma(\phi_{L-1}(x)+c_i) < 0$. This again is a contradiction to σ being non-negative.

Therefore, we have that $-c_i \leq \phi_L(x)_i \leq c_i$ for all $i \in [d]$ and $x \in \mathbb{R}^d$ implying $\|\phi_L(x)_i\|_{L^\infty(\mathbb{R}^d)} \leq c_i$.

Furthermore, we show that the approximation on the domain Ω does not get worse. We consider the following cases.

- Let $x \in \Omega$ and $i \in [d]$ be such that $\phi_{L-2}(x)_i > c_i$. Then $\phi_{L-1}(x)_i = c_i$ and $\phi_L(x)_i = c_i$. We observe that $-\nabla V(x)_i \leq c_i = \phi_L(x)_i < \phi_{L-2}(x)_i$ implies $|-\nabla V(x)_i - \phi_L(x)_i| = \phi_L(x)_i + \nabla V(x)_i < \phi_{L-2}(x)_i + \nabla V(x)_i = |-\nabla V(x)_i - \phi_{L-2}(x)_i|$

- Let $x \in \Omega, i \in [d]$ such that $\phi_{L-2}(x)_i < -c_i$. Then $\phi_{L-1}(x)_i < -c_i$ and $\phi_L(x)_i = -c_i$. With $-\nabla V(x)_i \geq -c_i = \phi_L(x)_i > \phi_{L-2}(x)_i$, we observe that $|-\nabla V(x)_i - \phi_L(x)_i| = -\nabla V(x)_i - \phi_L(x)_i \leq -\nabla V(x)_i - \phi_{L-2}(x)_i = |-\nabla V(x)_i - \phi_{L-2}(x)_i|$

- Let $x \in \Omega, i \in [d]$ such that $-c_i \leq \phi_{L-2}(x)_i \leq c_i$. Then $\phi_{L-1}(x)_i = \phi_{L-2}(x)_i$ and $\phi_L(x)_i = \phi_{L-2}(x)_i$. Therefore by assumption $|-\nabla V(x)_i - \phi_L(x)_i| = |-\nabla V(x)_i - \phi_{L-2}(x)_i|$.

Therefore, for all $x \in \Omega$ we get $|-\nabla V(x)_i - \phi_L(x)_i| \leq |-\nabla V(x)_i - \phi_{L-2}(x)_i|$. We have

$$\begin{aligned}
\|-\nabla V - \phi_L\|^p_{L^p_\mu(\mathbb{R}^d,\mathbb{R}^d)} &= \|-\nabla V - \phi_L\|^p_{L^p_\mu(\Omega;\mathbb{R}^d)} + \|-\nabla V - \phi_L\|^p_{L^p_\mu(\mathbb{R}^d\setminus\Omega;\mathbb{R}^d)} \\
&= \|-\nabla V - \phi_L\|^p_{L^p_\mu(\Omega;\mathbb{R}^d)} + \int_{\mathbb{R}^d\setminus\Omega} \|-\nabla V - \phi_L\|^p_{\ell^p} d\mu \\
&= \|-\nabla V - \phi_L\|^p_{L^p_\mu(\Omega;\mathbb{R}^d)} + 2^{p-1}\int_{\mathbb{R}^d\setminus\Omega} \|-\nabla V\|^p_{\ell^p} + \|\phi_L\|^p_{\ell^p} d\mu \\
&\leq \|-\nabla V - \phi_{L-2}\|^p_{L^p_\mu(\Omega;\mathbb{R}^d)} + 2^{p-1}(\|\nabla V\|^p_{L^p_\mu(\mathbb{R}^d\setminus\Omega;\mathbb{R}^d)} + \|c\|^p_{L^p_\mu(\mathbb{R}^d\setminus\Omega;\mathbb{R}^d)})
\end{aligned}$$

□

Lemma A.10. *Assume that ∇V is M-Lipschitz continuous such that for any $x, y \in \mathbb{R}^d$*

$$\|\nabla V(x)\|_{\ell^2} \leq \|\nabla V(y)\|_{\ell^2} + M\|x-y\|_{\ell^2}$$

and let μ be a sub-Gaussian measure with variance proxy σ^2. Let $x_0 \in \mathbb{R}^d$, $r > 0$ and $\Omega := B^2_r(0)$. Let $c \in \mathbb{R}^d$. Then

$$2\int_{\mathbb{R}^d\setminus\Omega} \|\nabla V\|^2_{\ell^2} + \|c\|^2_{\ell^2} \mathrm{d}\mu \leq \left[4M^2(2d^2\sigma^2 + r^2) + 2(2\|\nabla V(x_0)\|^2_{\ell^2} + \|c\|^2_{\ell^2})\right] \exp\left(-\frac{r^2}{2d^2\sigma^2}\right) \tag{A.24}$$

Proof. The layer cake representation asserts that

$$\int_\Omega f(x)\mathrm{d}\mu(x) = \int_0^\infty \mu(\{x \in \Omega | f(x) > s\})\mathrm{d}s. \tag{A.25}$$

Plugging in $\Omega = \{\|x\|_{\ell^2} > r\}$ and $f(x) = \|x\|^2_{\ell^2}$, it holds that

$$\begin{aligned}
\int_{\|x\|_{\ell^2}>r} \|x\|^2_{\ell^2}\mathrm{d}\mu(x) &= \int_0^\infty \mu(\{\|x\|_{\ell^2} > r, \|x\|^2_{\ell^2} > s\})\mathrm{d}s \\
&= \int_{r^2}^\infty \mu(\{\|x\|^2_{\ell^2} > s\})\mathrm{d}s + \int_0^{r^2} \mu(\{\|x\|_{\ell^2} > r\})\mathrm{d}s \\
&\leq \int_{r^2}^\infty \exp\left(-\frac{s}{2d^2\sigma^2}\right)\mathrm{d}s + r^2\exp\left(-\frac{r^2}{2d^2\sigma^2}\right) \\
&= 2d^2\sigma^2\exp\left(-\frac{r^2}{2d^2\sigma^2}\right) + r^2\exp\left(-\frac{r^2}{2d^2\sigma^2}\right) = (2d^2\sigma^2 + r^2)\exp\left(-\frac{r^2}{2d^2\sigma^2}\right).
\end{aligned} \tag{A.26}$$

Now, note that $\mu(\mathbb{R}^d \setminus \Omega) = \mathbb{P}(\|X\|_2 \geq r)$ and apply Proposition 2.7 to get $\mu(\mathbb{R}^d \setminus \Omega) \leq$

$\exp\left(-\frac{r^2}{2d^2\sigma^2}\right)$. Building on this, we get

$$2\int_{\mathbb{R}^d\setminus\Omega} \|\nabla V\|_{\ell^2}^2 + \|c\|_{\ell^2}^2 \mathrm{d}\mu \tag{A.27}$$

$$= 2\int_{\mathbb{R}^d\setminus B_r^2(0)} \|\nabla V(\cdot + x_0)\|_{\ell^2}^2 + \|c\|_{\ell^2}^2 d\mu \tag{A.28}$$

$$\leq 2\mu_Y(\mathbb{R}^d\setminus B_r^2(0))(2\|\nabla V(x_0)\|_{\ell^2}^2 + \|c\|_{\ell^2}^2) + 4M^2\int_{\mathbb{R}^d\setminus B_r^2(0)} \|x\|_{\ell^2}^2 d\mu \tag{A.29}$$

$$\leq 2\exp\left(-\frac{r^2}{2d^2\sigma^2}\right)(2\|\nabla V(x_0)\|_{\ell^2}^2 + \|c\|_{\ell^2}^2) + 4M^2(2d^2\sigma^2 + r^2)\exp\left(-\frac{r^2}{2d^2\sigma^2}\right) \tag{A.30}$$

$$\leq \left[4M^2(2d^2\sigma^2 + r^2) + 2(2\|\nabla V(x_0)\|_{\ell^2}^2 + \|c\|_{\ell^2}^2)\right]\exp\left(-\frac{r^2}{2d^2\sigma^2}\right) \tag{A.31}$$

□

Lemma A.11. *Let $\varepsilon > 0$. Assume that ∇V is Lipschitz continuous such that for any $x, y \in \mathbb{R}^d$*

$$\|\nabla V(x)\|_{\ell^2} \leq \|\nabla V(y)\|_{\ell^2} + M\|x - y\|_{\ell^2}$$

and let μ be a sub-Gaussian measure with variance proxy σ^2. Let $x_0 \in \mathbb{R}^d$,

$$r = \sqrt{2}d\sigma\ln\left(\frac{16(\|\nabla V(x_0)\|_{\ell^2}^2(1+d) + M^2d^2\sigma^2(8d+10))}{\varepsilon^4}\right)^{1/2}.$$

and let $\Omega \coloneqq B_r^2(x_0)$. Let $c \in \mathbb{R}^d$ with $c_i \coloneqq \|\nabla V_i\|_{L^\infty(\Omega)}$. Then

$$2(\|\nabla V\|_{L^2_\mu(\mathbb{R}^d\setminus\Omega;\mathbb{R}^d)}^2 + \|c\|_{L^2_\mu(\mathbb{R}^d\setminus\Omega;\mathbb{R}^d)}^2) \leq \varepsilon^2/2$$

Proof. Let j be the index of the maximum of c and note that

$$\begin{aligned}
\|c\|_{\ell^2} &\leq \sqrt{d}c_j = \sqrt{d}\max_{x\in\Omega}|\nabla V(x)_j| \\
&\leq \sqrt{d}\max_{x\in\Omega}\|\nabla V(x)\|_{\ell^2} \\
&\leq \sqrt{d}\left(\max_{x\in\Omega}\|\nabla V(x_0)\|_{\ell^2} + M\|x - x_0\|_{\ell^2}\right) \\
&= \sqrt{d}(\|\nabla V(x_0)\|_{\ell^2} + Mr).
\end{aligned}$$

Then, using Lemma A.10, it holds that

$$2\int_{\mathbb{R}^d\setminus\Omega} \|\nabla V\|_{\ell^2}^2 + \|c\|_{\ell^2}^2 \mathrm{d}\mu \tag{A.32}$$

$$\leq \left[4M^2(2d^2\sigma^2 + r^2) + 2(2\|\nabla V(x_0)\|_{\ell^2}^2 + 2d(\|\nabla V(x_0)\|_{\ell^2}^2 + M^2r^2))\right]\exp\left(-\frac{r^2}{2d^2\sigma^2}\right) \tag{A.33}$$

$$\leq \left[(4+4d)M^2r^2 + 2\|\nabla V(x_0)\|_{\ell^2}^2(2+2d) + 2d^2\sigma^2 4M^2\right]\exp\left(-\frac{r^2}{2d^2\sigma^2}\right) \tag{A.34}$$

To simplify notation we define $a = 2d^2\sigma^2$, $b = (4{+}4d)M^2$, $c = 2\|\nabla V(x_0)\|_{\ell^2}^2(2{+}2d){+}2d^2\sigma^2 4M^2$. With this, it holds that

$$2\int_{\mathbb{R}^d\setminus\Omega} \|\nabla V\|_{\ell^2}^2 + \|c\|_{\ell^2}^2 \mathrm{d}\mu \leq \left[br^2 + c\right] e^{-r^2/a}. \tag{A.35}$$

Now let $x > 0$ be such that $r = \sqrt{a\ln(x)}$. Then, $2\int_{\mathbb{R}^d\setminus\Omega} \|\nabla V\|_{\ell^2}^2 + \|c\|_{\ell^2}^2 \mathrm{d}\mu < \frac{\varepsilon^2}{2}$ holds if

$$[ba\ln(x) + c]\frac{1}{x} < \frac{\varepsilon^2}{2}. \tag{A.36}$$

This inequality holds if

$$\frac{c}{x} < \frac{\varepsilon^2}{4}, \tag{A.37}$$

$$ba\frac{\ln(x)}{x} < \frac{\varepsilon^2}{4}. \tag{A.38}$$

The first of these inequalities leads to the condition

$$x > \frac{4c}{\varepsilon^2}. \tag{A.39}$$

Regarding the second inequality, note that $\ln(x)/x < 1/\sqrt{x}$ for all $x > 0$. Hence, the second inequality is satisfied if $\frac{ba}{\sqrt{x}} < \frac{\varepsilon^2}{4}$ which is equivalent to $x > \frac{16ba}{\varepsilon^4}$. In total,

$$x > \max\left\{\frac{4c}{\varepsilon^2}, \frac{16ba}{\varepsilon^4}\right\} \tag{A.40}$$

leads to the desired inequalities. For $\varepsilon < 1$, this is satisfied for

$$x = \frac{4c + 16ba}{\varepsilon^4}. \tag{A.41}$$

Plugging this x back into the expression for r leads to

$$r = \sqrt{a\ln\left(\frac{4c + 16ba}{\varepsilon^4}\right)} \tag{A.42}$$

and inserting all the correct expression again yields

$$r = \sqrt{2}d\sigma\ln\left(\frac{4(2\|\nabla V(x_0)\|_{\ell^2}^2(2+2d) + 2d^2\sigma^2 4M^2) + 16((4+4d)M)2d^2\sigma^2}{\varepsilon^4}\right)^{1/2}, \tag{A.43}$$

which simplifies to the expression given in the theorem.

□

Remark A.12. *The proof of Lemma A.11 leads to the following assertion. Let $a, b, c > 0$. Then*

$$\left[br^2 + c\right] e^{-r^2/a} < \frac{\varepsilon^2}{2}$$

is fulfilled for

$$r = \sqrt{a\ln\left(\frac{4c + 16ba}{\varepsilon^4}\right)}.$$

Remark A.13. *We can have another expression of r in the asymptotic case $\varepsilon \to 0$. Let $\varepsilon > 0$. We are trying to solve for r the following equation*

$$[br^2 + c]\exp(-r^2/a) \leq \varepsilon$$

with $a = 2d^2\sigma^2$, $b = (4+4d)M^2$, $c = 2\|\nabla V(x_0)\|_{\ell^2}^2(2+2d) + 2d^2\sigma^2 4M^2$.

If $-\frac{\exp(-\frac{c}{ab})\varepsilon}{ab} \in (-1/e, 0)$ (it should be satisfied for a small ε i.e. $\varepsilon \in (0, ab\exp\left(\frac{c}{ab} - 1\right))$), then

$$r = \sqrt{\frac{-abW_{-1}\left(-\frac{\exp(-\frac{c}{ab})\varepsilon}{ab}\right) - c}{b}}$$

where W_{-1} is the Lambert W function defined on the branch $[-1/e, 0)$.

We can show that $W_{-1}(x) = \ln(-x) - \ln(-\ln(-x)) + \mathcal{O}\left(\frac{\ln(-\ln(-x))}{\ln(-x)}\right)$ when $x \to 0^-$. Therefore, $W_{-1}(x) \underset{x\to 0^-}{\sim} \ln(-x)$.

By the properties of the asymptotic equivalence *$\sim$, we have*

$$\begin{aligned} -abW_{-1}\left(-\frac{\exp(-\frac{c}{ab})\varepsilon}{ab}\right) &\sim -ab\ln\left(\frac{\exp(-\frac{c}{ab})\varepsilon}{ab}\right) \\ &= -ab\left(-\frac{c}{ab} + \ln(\varepsilon) - \ln(ab)\right) \\ &= c - ab\ln(\varepsilon) + ab\ln(ab) \end{aligned}$$

and since for ε sufficiently small the expression above is positive, we have

$$r \underset{\varepsilon\to 0}{\sim} \sqrt{a}\sqrt{\ln(ab) - \ln(\varepsilon)}$$

and by the definitions of a, b and c,

$$r \underset{\varepsilon\to 0}{\sim} \sqrt{2}d\sigma\sqrt{\ln(8d^2\sigma^2G^2) - \ln(\varepsilon)}$$

Theorem A.14 (Growth of the sub-Gaussian variance proxy)**.** *Suppose $(\phi_k)_{k\geq 0}$ is sequence of bounded functions. Then, the stochastic process Y^Φ driven by Φ, as defined in Definition 3.2, is a sub-Gaussian random vector for all $t = kh$, and the variance proxy $\sigma_k^2 > 0$ is given by*

$$\sigma_k \leq 1 + \sqrt{2h}k + \sqrt{d}h\sum_{j=0}^{k-1}\|\mathcal{R}\phi_k\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^d)}$$

Proof. We prove this by induction. First $Y_0 \sim \mathcal{N}_d(0, I)$ so it is sub-Gaussian with variance proxy $\sigma_{\mathcal{N}}^2 = 1$. Now, we assume that Y_{kh} is sub-Gaussian with variance proxy $\sigma_k^2 > 0$, i.e.

$$\mathbb{E}[|\langle u, Y_{kh}\rangle|^p] \leq (\sqrt{2}\sigma_k)^p p\Gamma(p/2)$$

for all $p \geq 1$ and $u \in \mathbb{S}_1(0)$. We have $Y_{(k+1)h} = Y_{kh} + h\mathcal{R}\phi_k(Y_{kh}) + \sqrt{2h}\xi$, with $\xi \sim \mathcal{N}_d(0, I)$. Therefore, by the Minkowski inequality,

$$\begin{aligned} \mathbb{E}[|\langle u, Y_{(k+1)h}\rangle|^p]^{1/p} &\leq \mathbb{E}[|\langle u, Y_{kh}\rangle|^p]^{1/p} + h\,\mathbb{E}[|\langle u, \mathcal{R}\phi_k(Y_{kh})\rangle|^p]^{1/p} + \sqrt{2h}\,\mathbb{E}[|\langle u, \xi\rangle|^p]^{1/p} \\ &\leq \sqrt{2}\sigma_k p^{1/p}\Gamma(p/2)^{1/p} + h\|u\|_{\ell^1}\|\mathcal{R}\phi_k\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^d)} + 2\sqrt{h}p^{1/p}\Gamma(p/2)^{1/p} \\ &\leq \sqrt{2}\sigma_k p^{1/p}\Gamma(p/2)^{1/p} + \sqrt{2d}hp^{1/p}\Gamma(p/2)^{1/p}\|\mathcal{R}\phi_k\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^d)} + 2\sqrt{h}p^{1/p}\Gamma(p/2)^{1/p} \\ &\leq \sqrt{2}p^{1/p}\Gamma(p/2)^{1/p}(\sigma_k + \sqrt{d}h\|\mathcal{R}\phi_k\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^d)} + \sqrt{2h}) \end{aligned}$$

Finally, we have that $Y_{(k+1)h}$ is a *sub-Gaussian random vector* according to Definition 2.6 and Definition 2.5 (iii), and the variance proxy σ_{k+1}^2 satisfies

$$\sigma_{k+1} \leq \sigma_k + \sqrt{d}h\|\mathcal{R}\phi_k\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^d)} + \sqrt{2h}.$$

By telescoping we have,

$$\sigma_{k+1} \leq 1 + \sqrt{2h}k + \sqrt{d}h\sum_{j=1}^{k}\|\mathcal{R}\phi_j\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^d)}$$

□

Proposition A.15. *Asymptotic behavior of σ_k Let $d \in \mathbb{N}^*, k \in \mathbb{N}$. Then,*

$$\sigma_k = \underset{\varepsilon\to 0}{\mathcal{O}}\left([d^{\frac{9}{4}}\varepsilon^{-1}]^{2((3/2)^k-1)}\right)$$

where the constant in the big-O depends only on the regularity of ∇V and the step size h.

Proof. By Theorem A.14, we have

$$\sigma_{k+1} \leq \sigma_k + \sqrt{d}h\|\mathcal{R}\phi_k\|_\infty + \sqrt{2h}$$

Note that for $\Omega = B_{r_k}^2(x_k)$ we have

$$\begin{aligned}
\|\mathcal{R}\phi_k\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^d)} &\leq \|\nabla V\|_{L^\infty(\Omega,\mathbb{R}^d)} = \max_{x\in\Omega}\|\nabla V(x)\|_{\ell^\infty} \\
&\leq \|\nabla V(x_k)\|_{\ell^\infty} + \max_{x\in\Omega}\|\nabla V(x) - \nabla V(x_k)\|_{\ell^\infty} \\
&\leq \|\nabla V(x_k)\|_{\ell^\infty} + \max_{x\in\Omega}\|\nabla V(x) - \nabla V(x_k)\|_{\ell^2} \\
&\leq \|\nabla V(x_k)\|_{\ell^\infty} + \max_{x\in\Omega} M\|x - x_k\|_{\ell^2} \\
&\leq \|\nabla V(x_k)\|_{\ell^\infty} + Mr_k.
\end{aligned}$$

With Lemma A.11,

$$r_k = \underset{\varepsilon\to 0}{\mathcal{O}}\left(d\sigma_k \ln\left(\frac{d^3\sigma_k^2}{\varepsilon^4}\right)^{1/2}\right),$$

we have

$$\begin{aligned}
\sigma_{k+1} &\leq \sigma_k + \sqrt{2h} + \sqrt{d}h(\|\nabla V(x_k)\|_\infty + Mr_k) \\
&= \sigma_k + \sqrt{2h} + \mathcal{O}\left(d^{3/2}\sigma_k \ln\left(\frac{d^3\sigma_k^2}{\varepsilon^4}\right)^{1/2}\right) \\
&= \mathcal{O}\left(d^{3/2}\sigma_k \ln\left(\frac{d^3\sigma_k^2}{\varepsilon^4}\right)^{1/2}\right)
\end{aligned}$$

Recall that $\forall\gamma,\alpha > 0, |\ln x|^\gamma = \underset{x\to\infty}{\mathcal{O}}(x^\alpha)$. Then, choosing $\gamma = \frac{1}{2}$ and $\alpha = \frac{1}{4}$ leads to

$$\begin{aligned}
\sigma_{k+1} &= \mathcal{O}\left(d^{3/2}\sigma_k \ln\left(\frac{d^3\sigma_k^2}{\varepsilon^4}\right)^{1/2}\right) \\
&= \mathcal{O}\left(d^{9/4}\sigma_k^{3/2}\varepsilon^{-1}\right).
\end{aligned}$$

And by recurrence,

$$\sigma_k = \underset{\varepsilon\to 0}{\mathcal{O}}\left([d^{\frac{9}{4}}\varepsilon^{-1}]^{2((3/2)^k-1)}\right).$$

□

Remark A.16. *Using Remark A.13, we have*

$$\begin{aligned}\sigma_{k+1} &= \mathcal{O}\left(d^{3/2}\sigma_k \ln\left(\frac{d^3\sigma_k^2}{\varepsilon}\right)^{1/2}\right)\\ &= \mathcal{O}\left(d^{9/4}\sigma_k^{3/2}\varepsilon^{-1/4}\right)\end{aligned}$$

and therefore,

$$\sigma_k = \underset{\varepsilon\to 0}{\mathcal{O}}\left([d^{\frac{9}{4}}\varepsilon^{-1/4}]^{2((3/2)^k-1)}\right)$$

Proposition A.17. *Asymptotic behavior of r_k Let $d \in \mathbb{N}^*, k \in \mathbb{N}$. Then,*

$$r_k = \underset{\varepsilon\to 0}{\mathcal{O}}\left(d^{\frac{7}{4}}\varepsilon^{-1}[d^{\frac{9}{4}}\varepsilon^{-1}]^{3((3/2)^k-1)}\right)$$

where the constant in the big-O depends only on the regularity of ∇V and the step size h.

Proof. Recall that by Lemma A.11

$$r_k = \underset{\varepsilon\to 0}{\mathcal{O}}\left(d\sigma_k \ln\left(\frac{d^3\sigma_k^2}{\varepsilon^4}\right)^{1/2}\right)$$

and that $\forall\gamma,\alpha>0, |\ln x|^\gamma = \underset{x\to\infty}{\mathcal{O}}(x^\alpha)$. Then, choosing $\gamma = \frac{1}{2}$ and $\alpha = \frac{1}{4}$ leads to

$$r_k = \underset{\varepsilon\to 0}{\mathcal{O}}\left(d^{\frac{7}{4}}\sigma_k^{\frac{3}{2}}\varepsilon^{-1}\right)$$

Then, plugging in the result from Proposition A.15 leads to

$$r_k = \underset{\varepsilon\to 0}{\mathcal{O}}\left(d^{\frac{7}{4}}\varepsilon^{-1}[d^{\frac{9}{4}}\varepsilon^{-1}]^{3((3/2)^k-1)}\right).$$

□

Remark A.18. *Using Remarks A.13 and A.16, we have*

$$r_k = \underset{\varepsilon\to 0}{\mathcal{O}}\left(d^{\frac{7}{4}}\varepsilon^{-1/4}[d^{\frac{9}{4}}\varepsilon^{-1/4}]^{3((3/2)^k-1)}\right)$$

A.7 Approximation of Langevin Monte Carlo under local error- and Lipschitz constraints

Proof of Proposition 7.1. Since the minimizer of V is given by 0 we have $\nabla V(0) = 0$. The strong convexity then implies for all $x, z \in \mathbb{R}^d$ that

$$V(x) \geq m/2\|x\|_{\ell^2}^2 + V(0)$$

and

$$-\langle\nabla V(x), z-x\rangle \geq m/2\|z-x\|_{\ell^2}^2 - V(z) + V(x).$$

We get for $z = 0$ that

$$\langle\nabla V(x), x\rangle \geq m/2\|x\|_{\ell^2}^2 - V(0) + V(x).$$

and hence

$$\langle \nabla V(x), x\rangle \geq m\|x\|_{\ell^2}^2.$$

for all $x \in \mathbb{R}^d$. With the Lipschitz continuity, we get for any $a > 0$ that

$$\begin{aligned}\|\nabla V(x) - ax\|_{\ell^2}^2 &= \|\nabla V(x)\|_{\ell^2}^2 - 2a\langle \nabla V(x), x\rangle + a^2\|x\|_{\ell^2}^2\\ &\leq (M^2 + a^2)\|x\|^2 - 2am\|x\|^2\\ &= (M^2 + a^2 - 2am)\|x\|^2\end{aligned}$$

The inequality $g := \sqrt{M^2 + a^2 - 2am} < m$ is fulfilled for $a = m$, if $0 < M < \sqrt{2}m$. In this case, $g = \sqrt{M^2 - m^2}$. □

Proof of Proposition 7.2. We assume without loss of generality that the unique minimizer of V is given by $x^* = 0$ and hence $\nabla V(0) = 0$. Throughout this proof we overload notation by writing ϕ instead of $\mathcal{R}\phi$ to denote the realization of a network ϕ. With Assumption 1.2 (ii) there exists for any $\varepsilon, r > 0$ a neural network $\phi_r^{(3)}$ with number of parameters bounded by $N(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M)$ such that

$$\| - \nabla V - \mathcal{R}\phi_r^{(3)}\|_{L^\infty(B_r^1(0))} \leq \| - \nabla V - \mathcal{R}\phi_r^{(3)}\|_{L^\infty(B_r^2(0))} \leq \varepsilon/\sqrt{d}.$$

By Proposition B.1, we can represent $x \in \mathbb{R}^d \mapsto x$ by a neural network with exactly $4d$ parameters and depth 2. For summing neural networks, we use the result Proposition B.2. We construct a network $\phi_r^{(2)}$ by adding to $\phi_r^{(3)}$ a neural network representing the d-dimensional identity, whose last layer is multiplied with $-m$. Then, $\mathcal{R}\phi_r^{(2)}(x) = \mathcal{R}\phi_r^{(3)}(x) - mx$ holds and

$$\| - \nabla V - m \cdot -\mathcal{R}\phi_r^{(2)}\|_{L^\infty(B_r^1(0))} = \| - \nabla V - \mathcal{R}\phi_r^{(3)}\|_{L^\infty(B_r^1(0))} \leq \varepsilon/\sqrt{d}.$$

To keep track of complexity, note that $\phi_r^{(2)}$ is the sum of a network of $4d$ parameters and depth 2 and a network of $N(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M)$ parameters and depth $L(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M)$. Hence, by Proposition B.2, the number of parameters and the depth are bounded by

$$\begin{aligned}\mathcal{P}(\phi_r^{(2)}) &\leq d(L(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) - 1) + N(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) + 4d,\\ \mathbb{L}(\phi_r^{(2)}) &\leq L(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M).\end{aligned}$$

Let $\phi_r^{(1)}$ be the cut-off NN of $\phi_r^{(2)}$ as in Lemma A.9 with the same accuracy on the ℓ^1-ball and

$$\begin{aligned}\max_{x\in\mathbb{R}^d} \|\mathcal{R}\phi_r^{(1)}(x)\|_{\ell^\infty} &\leq \max_{x\in B_r^1(0)} \| - \nabla V - m \cdot \|_{\ell^\infty} \leq \max_{x\in B_r^1(0)} \| - \nabla V - m \cdot \|_{\ell^2} \leq rG,\\ \max_{x\in\mathbb{R}^d} \|\mathcal{R}\phi_1^{(1)}(x)\|_{\ell^2} &\leq \max_{x\in\mathbb{R}^d} \sqrt{d}\|\mathcal{R}\phi_r^{(1)}(x)\|_{\ell^\infty} \leq \sqrt{d}rG,\end{aligned}$$

where we used Proposition 7.1 with $G := \sqrt{M^2 - m^2}$. Again, keeping track of complexity, the cutoff from Lemma A.9 introduces two more layers and $2d^2 + 2$ more weights to the network. Hence

$$\begin{aligned}\mathcal{P}(\phi_r^{(1)}) &\leq d(L(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) - 1) + N(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) + 4d + 2d^2 + 2,\\ \mathbb{L}(\phi_r^{(1)}) &\leq L(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) + 2.\end{aligned}$$

Now, define for $b > r$ the following approximation to the indicator function on $B_r^1(0)$,

$$f(x) = 1 - \frac{\mathrm{ReLU}(\|x\|_{\ell^1} - r) - \mathrm{ReLU}(\|x\|_{\ell^1} - b)}{b - r}, \tag{A.44}$$

with complexity $\mathcal{P}(f) = 4d + 7$ and $\mathbb{L}(f) = 3$ (for a proof of the complexity, see Proposition B.4), and note that

$$f(x)\phi_r^{(1)}(x) = \begin{cases} \phi_r^{(1)}(x), & \|x\|_{\ell^1} \leq r \\ (1 - \frac{\|x\|_{\ell^1} - r}{b-r})\phi_r^{(1)}(x), & r \leq \|x\|_{\ell^1} \leq b \\ 0, & \|x\|_{\ell^1} \geq b \end{cases}. \tag{A.45}$$

Now, noting that $f(x) \in [0,1]$ and $\phi_r^{(1)}(x) \in [-rG, rG]^d$ for all $x \in \mathbb{R}^d$, this product can be approximated by composing the parallelization of f and $\phi_r^{(1)}$ with an approximation of the multiplication $(x,y) \mapsto xy$ on $\Omega := [0,1] \times [-rG, rG]^d$, using Proposition B.6. In particular, we want to approximate the product up to $\varepsilon/\sqrt{d}$ error in $L^\infty(\Omega)$. We call the network that accomplishes this $\tilde{\phi}_r^{(0)}$. The complexity of $\tilde{\phi}_r^{(0)}$ is given by

$$\begin{aligned} \mathcal{P}(\tilde{\phi}_r^{(0)}) &= \mathcal{O}\left(d\log(2d\max\{1, rG\}/\varepsilon) + \mathcal{P}(\phi_r^{(1)}) + \mathcal{P}(f)\right) \\ &= \mathcal{O}\Big(d\log(2d\max\{1, rG\}/\varepsilon) + d(L(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) - 1) \\ &\qquad + N(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) + 8d + 2d^2 + 9\Big) \\ \mathbb{L}(\tilde{\phi}_r^{(0)}) &= \mathcal{O}\left(\log(2d\max\{1, rG\}/\varepsilon) + \mathbb{L}(f) + \mathbb{L}(\phi_r^{(1)})\right) \\ &= \mathcal{O}\left(\log(2d\max\{1, rG\}/\varepsilon) + L(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) + 5\right) \end{aligned}$$

Finally, we define $\phi_r^{(0)}$ to be the network representing $\phi := \tilde{\phi}_r^{(0)} + m\cdot$. This is an addition of $\tilde{\phi}_r^{(0)}$ with a network of $4d$ parameters and depth 2. By the properties of neural network summation (Proposition B.2) and the properties of the "big-O" notation, the complexity of $\phi_r^{(0)}$ is given by

$$\begin{aligned} \mathcal{P}(\phi_r^{(0)}) &= d\left[\mathcal{O}\left(\log(2d\max\{1, rG\}/\varepsilon) + L(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) + 5\right) - 1\right] \\ &\quad + \mathcal{O}\Big(d\log(2d\max\{1, rG\}/\varepsilon) + d(L(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) - 1) \\ &\qquad + N(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) + 12d + 2d^2 + 9\Big) \\ &= \mathcal{O}\left(d\log(2d\max\{1, rG\}/\varepsilon) + N(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) + dL(d, r, \sqrt{2}\varepsilon/\sqrt{d}, m, M) + 2d^2\right). \end{aligned}$$

The goal now is to show that ϕ fulfills the condition

$$\| - \nabla V(x) - \phi(x)\|_{\ell^2} \leq c\varepsilon + G\|x\|_{\ell^2}, \quad x \in \mathbb{R}^d,$$

for some constant c independent of r. We treat three cases separately. First, consider points x inside the ℓ^1-ball of radius r, i.e. $\|x\|_{\ell^1} \leq r$. In this case, we have

$$\begin{aligned} \| - \nabla V(x) - \phi(x)\|_{\ell^2} &= \| - \nabla V(x) - \tilde{\phi}_r^{(0)}(x) - mx\|_{\ell^2} \\ &\leq \| - \nabla V(x) - f(x)\phi_r^{(1)}(x) - mx\|_{\ell^2} + \|f(x)\phi_r^{(1)}(x) - \tilde{\phi}_r^{(0)}(x)\|_{\ell^2} \\ &\leq \| - \nabla V(x) - f(x)\phi_r^{(1)}(x) - mx\|_{\ell^2} + \sqrt{d}\|f(x)\phi_r^{(1)}(x) - \tilde{\phi}_r^{(0)}(x)\|_{\ell^\infty} \\ &\leq \| - \nabla V(x) - f(x)\phi_r^{(1)}(x) - mx\|_{\ell^2} + \varepsilon \\ &= \| - \nabla V(x) - \phi_r^{(1)}(x) - mx\|_{\ell^2} + \varepsilon \\ &\leq \| - \nabla V(x) - \phi_r^{(2)}(x) - mx\|_{\ell^2} + \varepsilon \\ &= \| - \nabla V(x) - \phi_r^{(3)}(x)\|_{\ell^2} + \varepsilon \\ &\leq \varepsilon + \varepsilon \\ &\leq 2\varepsilon + G\|x\|_{\ell^2}. \end{aligned}$$

Next, let $\|x\|_{\ell^1} \geq b$. In this case, we have

$$\begin{aligned}
\| - \nabla V(x) - \phi(x)\|_{\ell^2} &= \| - \nabla V(x) - \tilde{\phi}_r^{(0)}(x) - mx\|_{\ell^2} \\
&\leq \| - \nabla V(x) - f(x)\phi_r^{(1)}(x) - mx\|_{\ell^2} + \|f(x)\phi_r^{(1)}(x) - \tilde{\phi}_r^{(0)}(x)\|_{\ell^2} \\
&\leq \| - \nabla V(x) - f(x)\phi_r^{(1)}(x) - mx\|_{\ell^2} + \sqrt{d}\|f(x)\phi_r^{(1)}(x) - \tilde{\phi}_r^{(0)}(x)\|_{\ell^\infty} \\
&\leq \| - \nabla V(x) - f(x)\phi_r^{(1)}(x) - mx\|_{\ell^2} + \varepsilon \\
&\leq \| - \nabla V(x) - 0 - mx\|_{\ell^2} + \varepsilon \\
&= \| - \nabla V(x) - mx\|_{\ell^2} + \varepsilon \\
&\leq G\|x\|_{\ell^2} + \varepsilon.
\end{aligned}$$

Finally, let $r \leq \|x\|_{\ell^1} \leq b$. We have not made a choice for b yet. Now (and in every line of the proof before), we let

$$b = r + \frac{\varepsilon}{\max\{L_{\phi_r^{(1)}}, M\}}, \tag{A.46}$$

$$\hat{x} = \frac{rx}{\|x\|_{\ell^1}}. \tag{A.47}$$

Note that $\|x - \hat{x}\|_{\ell^2} \leq b - r$. Then, we have

$$\begin{aligned}
&\| - \nabla V(x) - \phi(x)\|_{\ell^2} \\
&\leq \| - \nabla V(x) + \nabla V(\hat{x})\|_{\ell^2} + \| - \nabla V(\hat{x}) - \phi(\hat{x})\|_{\ell^2} + \|\phi(\hat{x}) - \phi(x)\|_{\ell^2}.
\end{aligned}$$

We treat the terms of the triangle inequality separately. First, note that due to the Lipschitz continuity of ∇V and the definition of b we have

$$\| - \nabla V(x) + \nabla V(\hat{x})\|_{\ell^2} \leq M\|x - \hat{x}\|_{\ell^2} \leq \varepsilon.$$

Consider next the second term. Since $\hat{x}$ is an element of $B_r^1(0)$ by construction, the approximation properties of the network guarantee that

$$\| - \nabla V(\hat{x}) - \phi(\hat{x})\|_{\ell^2} \leq 2\varepsilon,$$

which follows from case 1. The remaining term captures the growth of the network on the "slope domain" between $B_r^1(0)$ and $B_b^1(0)$, and can be bounded by using the Lipschitz continuity of $\phi_r^{(1)}$ and the definition of b. We have

$$\begin{aligned}
\|\phi(\hat{x}) - \phi(x)\|_{\ell^2} &= \|\tilde{\phi}_r^{(0)}(\hat{x}) + m\hat{x} - \tilde{\phi}_r^{(0)}(x) - mx\|_{\ell^2} \\
&\leq \left\|\phi_r^{(1)}(\hat{x}) + m\hat{x} - \left(1 - \frac{\|x\|_{\ell^1} - r}{b - r}\right)\phi_r^{(1)}(x) - mx\right\|_{\ell^2} \\
&\quad + \|\tilde{\phi}_r^{(0)}(\hat{x}) - f(\hat{x})\phi_r^{(1)}(\hat{x})\|_{\ell^2} + \|\tilde{\phi}_r^{(0)}(x) - f(x)\phi_r^{(1)}(x)\|_{\ell^2} \\
&\leq \left\|\phi_r^{(1)}(\hat{x}) + m\hat{x} - \left(1 - \frac{\|x\|_{\ell^1} - r}{b - r}\right)\phi_r^{(1)}(x) - mx\right\|_{\ell^2} \\
&\quad + \sqrt{d}\|\tilde{\phi}_r^{(0)}(\hat{x}) - f(\hat{x})\phi_r^{(1)}(\hat{x})\|_{\ell^\infty} + \sqrt{d}\|\tilde{\phi}_r^{(0)}(x) - f(x)\phi_r^{(1)}(x)\|_{\ell^\infty} \\
&\leq \left\|\phi_r^{(1)}(\hat{x}) + m\hat{x} - \left(1 - \frac{\|x\|_{\ell^1} - r}{b - r}\right)\phi_r^{(1)}(x) - mx\right\|_{\ell^2} + 2\varepsilon \\
&\leq \left\|\phi_r^{(1)}(\hat{x}) - \left(1 - \frac{\|x\|_{\ell^1} - r}{b - r}\right)\phi_r^{(1)}(x)\right\|_{\ell^2} + m\|x - \hat{x}\|_{\ell^2} + 2\varepsilon \\
&\leq \left\|\phi_r^{(1)}(\hat{x}) - \left(1 - \frac{\|x\|_{\ell^1} - r}{b - r}\right)\phi_r^{(1)}(x)\right\|_{\ell^2} + 3\varepsilon.
\end{aligned}$$

Thus, we have traced back the error in ϕ back to an error in the network $\phi_r^{(1)}$. This, we can bound by

$$\begin{aligned}
&\left\| \phi_r^{(1)}(\hat{x}) - \left(1 - \frac{\|x\|_{\ell^1} - r}{b - r}\right) \phi_r^{(1)}(x) \right\|_{\ell^2} \\
&\leq \left\| \phi_r^{(1)}(\hat{x}) - \phi_r^{(1)}(x) \right\|_{\ell^2} + \left(\frac{\|x\|_{\ell^1} - r}{b - r} \right) \left\| \phi_r^{(1)}(x) \right\|_{\ell^2} \\
&\leq L_{\phi_r^{(1)}} \|x - \hat{x}\| + \left\| \phi_r^{(1)}(x) \right\|_{\ell^2} \\
&\leq \varepsilon + \left(\left\| \phi_r^{(1)}(\hat{x}) \right\|_{\ell^2} + L_{\phi_r^{(1)}} \|x - \hat{x}\|_{\ell^2} \right) \\
&\leq 2\varepsilon + \| - \nabla V(\hat{x}) - m\hat{x} - \phi_r^{(1)}(\hat{x}) \|_{\ell^2} + \| -\nabla V(\hat{x}) - m\hat{x} \|_{\ell^2} \\
&\leq 3\varepsilon + G\|x\|_{\ell^2}.
\end{aligned}$$

Putting all of these inequalities together, we finally obtain

$$\| - \nabla V(x) - \phi(x) \|_{\ell^2} \leq 9\varepsilon + G\|x\|_{\ell^2},$$

yielding the claim. □

Proof of Theorem 7.3. We assume without loss of generality that the unique minimizer of V is given by $x^* = 0$ and hence $\nabla V(0) = 0$. For any $r > 0$ there exists by Proposition 7.2 a ReLU FCNN ϕ_r with N parameters such that

$$\begin{aligned}
&\| - \nabla V - \mathcal{R}\phi_r \|_{L^\infty(B_r^2(0))} \leq \varepsilon / \sqrt{2d}, \\
&\| - \nabla V(x) - \mathcal{R}\phi_r(x) \|_{\ell^2} \leq 9\varepsilon / \sqrt{2d} + \sqrt{M^2 - m^2} \|x\|_{\ell^2}, \quad \forall x \in \mathbb{R}.
\end{aligned}$$

Let $\Phi \coloneqq \{\phi_r\}_{k=0}^{K-1}$. Let $Y^\Phi \colon \Omega \times [0, Kh] \to \mathbb{R}^d$ be the stochastic process driven by Φ, i.e.

$$Y_t^\Phi = Y_0 + \int_0^t \mathcal{R}\phi_{\frac{1}{h}\chi_h(s)} \left(Y_{\chi_h(s)}^\Phi \right) \mathrm{d}s + \sqrt{2} W_t.$$

Let $G \coloneqq \sqrt{M^2 - m^2}$. By Proposition 5.3, $Y_{kh}^\Phi \sim \mu_{kh}^\Phi$ is sub-Gaussian for all $k = 0, \dots, K$ with variance proxy σ_k^2 bounded by

$$\sigma_k^2 \leq \frac{2h}{1 - (c + hG)} + \frac{81h^2\varepsilon^2}{2d} + \sigma_0^2 = \frac{2}{m - \sqrt{M^2 - m^2}} + \frac{81h^2\varepsilon^2}{2d} + \sigma_0^2. \tag{A.48}$$

Note that the right hand side is in $\mathcal{O}(1)$ as $\varepsilon \to 0$, $d \to \infty$. In particular, we find for $\varepsilon \in (0, 1)$ and $d \geq 1$ that

$$\sigma_k \leq \frac{2}{m - \sqrt{M^2 - m^2}} + \frac{81h^2}{2} + \sigma_0^2 =: \sigma.$$

Now, note that $\mu_{kh}^\Phi(\mathbb{R}^d \backslash B_r(0)) = \mathbb{P}(\|X\|_2 \geq r)$ and apply Proposition 2.7 to get $\mu_{kh}^\Phi(\mathbb{R}^d \backslash B_r^2(0)) \leq \exp\left(-\frac{r^2}{2d^2\sigma_k^2} \right)$. Hence, using the fact that for all $a, b \in \mathbb{R}$ it holds that $(a + b)^2 \leq 2(a^2 + b^2)$, we

have

$$\begin{aligned}
&\| -\nabla V - \mathcal{R}\phi_r\|^2_{L^2_{\mu^\Phi_{kh}}(\mathbb{R}^d;\mathbb{R}^d)} \\
&\quad= \int_{B^2_r(0)} \| -\nabla V(x) - \mathcal{R}\phi_r\|^2_{\ell^2} \mathrm{d}\mu^\Phi_{kh}(x) + \int_{\mathbb{R}^d\setminus B^2_r(0)} \| -\nabla V(x) - \mathcal{R}\phi_r\|^2_{\ell^2} \mathrm{d}\mu^\Phi_{kh}(x) \\
&\quad= \int_{B^2_r(0)} \left(\sqrt{d}\frac{\varepsilon}{\sqrt{2d}}\right)^2 \mathrm{d}\mu^\Phi_{kh}(x) + \int_{\mathbb{R}^d\setminus B^2_r(0)} (9\varepsilon/\sqrt{2d} + \sqrt{M^2-m^2}\|x\|_{\ell^2})^2 \mathrm{d}\mu^\Phi_{kh}(x) \\
&\quad\le \frac{\varepsilon^2}{2} + \frac{81\varepsilon^2}{d}\mu^\Phi_{kh}(\mathbb{R}^d\setminus B^2_r(0)) + 2\int_{\mathbb{R}^d\setminus B^2_r(0)} (M^2-m^2)\|x\|^2_{\ell^2}\mathrm{d}\mu^\Phi_{kh}(x) \\
&\quad\le \frac{\varepsilon^2}{2} + \left(\frac{81\varepsilon^2}{d} + 2(M^2-m^2)(2d^2\sigma^2 + r^2)\right)\exp\left(-\frac{r^2}{2d^2\sigma^2}\right),
\end{aligned}$$

where the layer cake representation (see Equation (A.26)) was used in the last inequality. Now, we use Remark A.12 with $a = 2d^2\sigma^2$, $b = 2(M^2-m^2)$, $c = 81\varepsilon^2/d + 4(M^2-m^2)d^2\sigma^2$, to see that

$$\left(\frac{81\varepsilon^2}{d} + 2(M^2-m^2)(2d^2\sigma^2 + r^2)\right)\exp\left(-\frac{r^2}{2d^2\sigma^2}\right) < \frac{\varepsilon^2}{2}$$

is satisfied if

$$r = \left[2d^2\sigma^2 \ln\left(\frac{4(81\varepsilon^2/d + 4(M^2-m^2)d^2\sigma^2) + 16\cdot 4(M^2-m^2)d^2\sigma^2}{\varepsilon^4}\right)\right]^{\frac{1}{2}}. \tag{A.49}$$

Hence, it holds for r as in (A.49) that

$$\| -\nabla V - \mathcal{R}\phi_r\|_{L^2_{\mu^\Phi_{kh}}(\mathbb{R}^d;\mathbb{R}^d)} < \varepsilon$$

Applying Theorem 4.1 for $h < \frac{2}{m+M}$ and $\Phi = \{\phi_r\}_{k=1}^K$, we get

$$\mathcal{W}_2(\mu_\infty, \mu^\Phi_{Kh}) \le (1-mh)^K \mathcal{W}_2(\mu_\infty, \mu_0) + \frac{7\sqrt{2}}{6}\frac{M}{m}\sqrt{hd} + \frac{1-(1-mh)^K}{m}\varepsilon.$$

Finally, Proposition 3.3 guarantees the existence of a network ψ with number of parameters equal to the number of parameters of ϕ_r such that for $\Psi := \{\psi\}_{k=1}^K$ it holds that $\mu^\Phi_{Kh} = \mu^\Psi$, where $\xi = (\xi_1, \dots, \xi_K)$ and $\tilde{\mathcal{R}}\Psi(Y_0, \xi) \sim \mu^\Psi$. Towards the asymptotic complexity of r for $\varepsilon \to 0$ and $d \to \infty$, note that

$$\begin{aligned}
&\left[2d^2\sigma^2 \ln\left(\frac{4(81\varepsilon^2/d + 4(M^2-m^2)d^2\sigma^2) + 16\cdot 4(M^2-m^2)d^2\sigma^2}{\varepsilon^4}\right)\right]^{\frac{1}{2}} \\
&= \left[2d^2\sigma^2 \ln\left(\frac{324\varepsilon^2/d + 80(M^2-m^2)d^2\sigma^2}{\varepsilon^4}\right)\right]^{\frac{1}{2}} \\
&\le \left[2d^2\sigma^2 \ln\left((324 + 80(M^2-m^2)\sigma^2)\varepsilon^{-4}d^2\right)\right]^{\frac{1}{2}} \\
&= \sqrt{2}d\sigma\left[\ln\left(324 + 80(M^2-m^2)\sigma^2\right) + \ln\left(\varepsilon^{-4}d^2\right)\right]^{\frac{1}{2}} \\
&\le \sqrt{2}d\sigma\left[\ln\left(324 + 80(M^2-m^2)\sigma^2\right)\right]^{\frac{1}{2}} + \sqrt{2}d\sigma\left[\ln\left(\varepsilon^{-4}d^2\right)\right]^{\frac{1}{2}} \\
&= \mathcal{O}(d) + \mathcal{O}\left(d\ln\left(\varepsilon^{-4}d^2\right)^{\frac{1}{2}}\right) \\
&= \mathcal{O}\left(d\left(1 + \ln\left(\varepsilon^{-4}d^2\right)^{\frac{1}{2}}\right)\right).
\end{aligned} \tag{A.50}$$

This yields the claim. □

B Standard Results for FCNNs

Proposition B.1 (Representation of the identity by ReLU neural networks)**.** *Let σ be the ReLU activation function and $d \in \mathbb{N}$. Then, there exists a fully connected neural network ϕ with $\mathcal{P}(\phi) = 4d$ and $\mathbb{L}(\phi) = 1$ such that*

$$\mathcal{R}\phi = \mathrm{Id}_{\mathbb{R}^d} .$$

Proof. Let

$$\begin{aligned} A_0 &:= \begin{pmatrix} I_d \\ \hline -I_d \end{pmatrix} \in \mathbb{R}^{2d\times d}, \\ b_0 &:= 0_{\mathbb{R}^{2d}}, \\ A_1 &:= \left(\begin{array}{c|c} I_d & -I_d \end{array}\right) \in \mathbb{R}^{d\times 2d}, \\ b_1 &:= 0_{\mathbb{R}^d}. \end{aligned}$$

Then, define $\phi := ((A_0, b_0), (A_1, b_1))$. We have

$$\begin{aligned} \mathcal{R}\phi(x) &= A_1\sigma(A_0x + b_0) + b_1 \\ &= \left(\begin{array}{c|c} I_d & -I_d \end{array}\right) \sigma\begin{pmatrix} x \\ \hline -x \end{pmatrix} \\ &= \sigma(x) - \sigma(-x) \\ &= x. \end{aligned}$$

□

Proposition B.2 (Sum of neural networks [30, Lemma 2.17])**.** *Let $\phi_1, \dots, \phi_n$ be n fully connected neural networks with d inputs and k outputs. Then, there exists a neural network ψ such that*

$$\begin{aligned} \mathcal{R}\psi &= \sum_{i=1}^{n} \mathcal{R}\phi_i, \\ \mathcal{P}(\psi) &\leq \delta + \sum_{i=1}^{n} \mathcal{P}(\phi_i), \\ \mathbb{L}(\psi) &= \max_{i=1\dots n} \mathbb{L}(\phi_i), \end{aligned}$$

where $\delta := \min(d, k)(\max_i \mathbb{L}(\phi_i) - \min_i \mathbb{L}(\phi_i))$.

Proposition B.3 (Representation of the Euclidean 1-norm)**.** *There exists a fully connected neural network ϕ with d inputs and 1 output such that*

$$\mathcal{R}\phi(x) = \|x\|_{\ell^1}$$

and

$$\begin{aligned} \mathcal{P}(\phi) &= 4d \\ \mathbb{L}(\phi) &= 1 \end{aligned}$$

Proof. Let

$$\begin{aligned} A_0 &:= \begin{pmatrix} I_d \\ -I_d \end{pmatrix} \in \mathbb{R}^{2d\times d}, \\ b_0 &:= 0_{\mathbb{R}^{2d}}, \\ A_1 &:= (1,\dots,1) \in \mathbb{R}^{1\times 2d}, \\ b_1 &:= 0. \end{aligned}$$

Then, define $\phi := ((A_0,b_0),(A_1,b_1))$. We have

$$\begin{aligned} \mathcal{R}\phi(x) &= A_1\sigma(A_0x+b_0)+b_1 \\ &= (1,\dots,1)\cdot\sigma\begin{pmatrix} x \\ -x \end{pmatrix} \\ &= \sum_{k=1}^{d}(\sigma(x_k)+\sigma(-x_k)) \\ &= \sum_{k=1}^{d}|x_k| \\ &= \|x\|_{\ell^1} \end{aligned}$$

□

Proposition B.4 (Approximation of the indicator function on $B_r(0)$)**.** *Let $\delta > 0$. Then, there exists a fully connected neural network ϕ with d inputs and 1 output such that*

$$\mathcal{R}(\phi)(x) = \begin{cases} 1 & \text{if } x \in B_r(0) \\ \frac{r+\delta-\|x\|_{\ell^1}}{\delta} & \text{if } x \in B_r(0)\cap B_{r+\delta}(0) \\ 0 & \text{otherwise} \end{cases}$$

and

$$\begin{aligned} \mathcal{P}(\phi) &= 4d+7 \\ \mathbb{L}(\phi) &= 3 \end{aligned}$$

Proof. By Proposition B.3, if

$$\begin{aligned} A_0 &:= \begin{pmatrix} I_d \\ -I_d \end{pmatrix} \in \mathbb{R}^{2d\times d}, \\ b_0 &:= 0_{\mathbb{R}^{2d}}, \\ A_1 &:= (1,\dots,1) \in \mathbb{R}^{1\times 2d}, \\ b_1 &:= 0. \end{aligned}$$

then the fully connected neural network $\tilde{\phi} := ((A_0,b_0),(A_1,b_1))$ satisfies $\mathcal{R}\tilde{\phi} = \|\cdot\|_{\ell^1}$. Let

$$\begin{aligned} A_2 &:= \begin{pmatrix} 1 \\ 1 \end{pmatrix} \in \mathbb{R}^{2\times 1}, \\ b_2 &:= \begin{pmatrix} -r \\ -(r+\delta) \end{pmatrix} \in \mathbb{R}^2, \\ A_3 &:= -\frac{1}{\delta}(1,-1) \in \mathbb{R}^{1\times 2}, \\ b_3 &:= 1. \end{aligned}$$

Then, define $\phi := ((A_0, b_0), (A_1, b_1), (A_2, b_2), (A_3, b_3))$. We have

$$\begin{aligned}\mathcal{R}\phi(x) &= A_3\sigma(A_2\|x\|_{\ell^1} + b_2) + b_3 \\ &= A_3\sigma(A_2\|x\|_{\ell^1} + b_2) + b_3 \\ &= -\frac{1}{\delta}(1,-1)\cdot\sigma\left(\frac{\|x\|_{\ell^1} - r}{\|x\|_{\ell^1} - (r+\delta)}\right) + 1 \\ &= -\frac{\sigma(\|x\|_{\ell^1} - r) - \sigma(\|x\|_{\ell^1} - (r+\delta))}{\delta} + 1\end{aligned}$$

□

Lemma B.5 ([63, Proposition 3]). *For $M > 0$ and $\varepsilon \in (0,1)$ there is a ReLU network ϕ^{mult} with $\mathcal{R}\phi : \mathbb{R}^2 \to \mathbb{R}$ such that*

1. $|\mathcal{R}\phi^{mult}(x,y) - xy| \leq \varepsilon$ *for all* $x, y \in [-M, M]$
2. $\mathcal{R}\phi^{mult}(x,y) = 0$, *if* $x = 0$ *or* $y = 0$
3. $\mathbb{L}(\phi^{mult}), \mathcal{P}(\phi^{mult}) \in \mathcal{O}(\log(1/\varepsilon) + \log(M))$

Proposition B.6 (Element-wise multiplication of neural networks). *Let $\varepsilon > 0$. Then there exists a fully connected neural network ϕ such that*

$$\|xy - \mathcal{R}\phi((x,y))\|_{\ell^1} \leq \varepsilon$$

where $x \in [A_1, B_1]$ and $y \in [A_2, B_2]^d$, and ϕ satisfies

$$\begin{aligned}\mathcal{P}(\phi) &= \mathcal{O}(d\log(dr/\varepsilon)) \\ \mathbb{L}(\phi) &= \mathcal{O}(\log(dr/\varepsilon))\end{aligned}$$

where $t = \min(A_1, A_2)$ and $r = \max((B_1 - t), (B_2 - t))$.

Proof. Let $\widetilde{\times} = ((A_0, b_0), \ldots, (A_L, b_L))$ Lemma B.5 be a fully connected neural network which satisfies, for all $\alpha \in [A_1, B_1]$ and $\beta \in [A_2, B_2]$,

$$|\mathcal{R}\widetilde{\times}(\alpha, \beta) - \alpha\beta| \leq \frac{\varepsilon}{d}$$

Construction of a neural network that can extract (x, y_j) Let $j \in \{1, \ldots, d\}$ and $z := (x, y) \in \mathbb{R}^{d+1}$. Let

$$\begin{aligned}\Gamma_j &:= E_{1,1} + E_{2,j+1} \in \mathbb{R}^{2\times(d+1)} \\ \widetilde{A}_0^j &:= \begin{pmatrix}\Gamma_j \\ \hline -\Gamma_j\end{pmatrix} \in \mathbb{R}^{4\times(d+1)}, \\ \widetilde{b}_0^j &:= 0_{\mathbb{R}^4}, \\ \widetilde{A}_1^j &:= \left(\begin{array}{cc|cc} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{array}\right) \in \mathbb{R}^{2\times 4}, \\ \widetilde{b}_1^j &:= 0_{\mathbb{R}^2}.\end{aligned}$$

where $E_{m,n} = e_m^T e_n$ is zero everywhere expect in (m,n) where it is equal to 1. Let $\widetilde{\phi}_j := ((\widetilde{A}_0^j, \widetilde{b}_0^j), (\widetilde{A}_1^j, \widetilde{b}_1^j))$. We have $\mathcal{P}(\widetilde{\phi}_j) = 8$ and $\mathbb{L}(\widetilde{\phi}_j) = 1$. Then,

$$\begin{aligned}
\mathcal{R}\widetilde{\phi}_j(z) &= \widetilde{A}_1^j \sigma(\widetilde{A}_0^j z + \widetilde{b}_0^j) + \widetilde{b}_1^j \\
&= \left(\begin{array}{cc|cc} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{array}\right) \cdot \begin{pmatrix} \sigma(x) \\ \sigma(y_j) \\ \sigma(-x) \\ \sigma(-y_j) \end{pmatrix} \\
&= \begin{pmatrix} \sigma(x) - \sigma(-x) \\ \sigma(y_j) - \sigma(-y_j) \end{pmatrix} \\
&= \begin{pmatrix} x \\ y_j \end{pmatrix}
\end{aligned}$$

Construction of a neural network that can approximate xy_j By concatenating $\widetilde{\phi}_j$ and $\widetilde{\times}$ [47, Definition 2.2], there exists a neural network $\phi_j := ((\widetilde{A}_0^j, \widetilde{b}_0^j), (A_0\widetilde{A}_1^j, A_0\widetilde{b}_1^j + b_0), (A_1, b_1), \ldots, (A_L, b_L))$ which satisfies $\mathcal{R}\phi_j = \mathcal{R}\widetilde{\times} \circ \mathcal{R}\widetilde{\phi}_j$, $\mathcal{P}(\phi_j) \leq 2(8 + \mathcal{P}(\widetilde{\times}))$ and $\mathbb{L}(\phi_j) = \mathbb{L}(\widetilde{\times}) + 1$. And so, for $z = (x, y_j) \in [A_1, B_1] \times [A_2, B_2]$,

$$\begin{aligned}
|\mathcal{R}\phi_j(x,y) - xy_j| &= |\mathcal{R}\widetilde{\times}(x, y_j) - xy_j| \\
&\leq \frac{\varepsilon}{d}
\end{aligned}$$

Parallelization of ϕ_j By parallelizing the $(\phi_j)_{j=1}^d$ [47, Defintion 2.7], there exists a neural network ϕ which satisfies

$$\begin{aligned}
\mathcal{R}\phi &= (\mathcal{R}\phi_1, \ldots, \mathcal{R}\phi_d) \\
\mathcal{P}(\phi) &= \sum_{j=1}^d \mathcal{P}(\phi_j) \\
&= 2d(8 + \mathcal{P}(\widetilde{\times})) \\
\mathbb{L}(\phi) &= \mathbb{L}(\widetilde{\times}) + 1
\end{aligned}$$

Finally, for $x \in [A_1, B_1]$ and $y \in [A_2, B_2]^d$,

$$\begin{aligned}
\|xy - \mathcal{R}\phi((x,y))\|_{\ell^1} &= \sum_{j=1}^d |\mathcal{R}\phi_j((x, y_j)) - xy_j| \\
&\leq \varepsilon.
\end{aligned}$$

□

References

[1] J. M. Altschuler and K. Talwar. Concentration of the langevin algorithm's stationary distribution, 2022.

[2] F. Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[3] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

[4] A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14:115–133, 1994.

[5] H. Bolcskei, P. Grohs, G. Kutyniok, and P. Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1):8–45, 2019.

[6] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[7] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

[8] S. Cai, Z. Mao, Z. Wang, M. Yin, and G. E. Karniadakis. Physics-informed neural networks (pinns) for fluid mechanics: A review, 2021.

[9] N. H. Chau, É. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang. On stochastic gradient langevin dynamics with dependent data streams: The fully nonconvex case. *SIAM Journal on Mathematics of Data Science*, 3(3):959–986, 2021.

[10] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[11] X. Cheng and P. Bartlett. Convergence of langevin mcmc in kl-divergence. In F. Janoos, M. Mohri, and K. Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 186–211. PMLR, 07–09 Apr 2018.

[12] X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan. Sharp convergence rates for langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

[13] L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[14] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[15] A. S. Dalalyan. Theoretical Guarantees for Approximate Sampling from Smooth and Log-Concave Densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 04 2016.

[16] A. S. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent, 2017.

[17] A. S. Dalalyan and A. G. Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *ArXiv*, abs/1710.00095, 2017.

[18] R. Dandekar, K. Chung, V. Dixit, M. Tarek, A. Garcia-Valadez, K. V. Vemula, and C. Rackauckas. Bayesian neural ordinary differential equations, 2022.

[19] P. Del Moral, A. Doucet, and A. Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 2006.

[20] A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551 – 1587, 2017.

[21] A. Durmus and E. Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. 2019.

[22] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 793–797. PMLR, 06–09 Jul 2018.

[23] D. Elbrächter, P. Grohs, A. Jentzen, and C. Schwab. Dnn expression rate analysis of high-dimensional pdes: Application to option pricing. *Constructive Approximation*, 55(1):3–71, 2022.

[24] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler. *SIAM Journal on Applied Dynamical Systems*, 19(1):412–441, 2020.

[25] A. Garbuno-Inigo, N. Nusken, and S. Reich. Affine invariant interacting Langevin dynamics for Bayesian inference. *SIAM Journal on Applied Dynamical Systems*, 19(3):1633–1658, 2020.

[26] M. Girolami and B. Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2):123–214, 2011.

[27] Q. Gong, W. Kang, and F. Fahroo. Approximation of compositional functions with relu neural networks. *Systems & Control Letters*, 175:105508, 2023.

[28] I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[29] J. Goodman and J. Weare. Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science*, 5(1):65–80, 2010.

[30] R. Gribonval, G. Kutyniok, M. Nielsen, and F. Voigtlaender. Approximation spaces of deep neural networks, 2020.

[31] I. Gühring, M. Raslan, and G. Kutyniok. *Expressivity of Deep Neural Networks*, page 149–199. Cambridge University Press, 2022.

[32] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] A. Jentzen, D. Salimova, and T. Welti. A proof that deep artificial neural networks overcome the curse of dimensionality in the numerical approximation of kolmogorov partial differential equations with constant diffusion and nonlinear drift coefficients. *Communications in Mathematical Sciences*, 19(5):1167–1205, 2021.

[34] J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media, 2006.

[35] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

[36] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[37] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[38] S. Lanthaler, S. Mishra, and G. E. Karniadakis. Error estimates for DeepONets: a deep learning framework in infinite dimensions. *Transactions of Mathematics and Its Applications*, 6(1):tnac001, 03 2022.

[39] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations, 2021.

[40] Y. L. Luke. Inequalities for generalized hypergeometric functions. *Journal of Approximation Theory*, 5(1):41–65, 1972.

[41] T. D. Luu, J. Fadili, and C. Chesneau. Sampling from non-smooth distributions through langevin diffusion. *Methodology and Computing in Applied Probability*, 23:1173–1201, 2021.

[42] M. B. Majka, A. Mijatović, and Ł. Szpruch. Nonasymptotic bounds for sampling algorithms without log-concavity. 2020.

[43] C. Marcati and C. Schwab. Exponential convergence of deep operator networks for elliptic partial differential equations, 2022.

[44] P. A. Markowich and C. Villani. On the trend to equilibrium for the fokker-planck equation: an interplay between physics and functional analysis. *Mat. Contemp*, 19:1–29, 2000.

[45] H. Montanelli and H. Yang. Error bounds for deep relu networks using the kolmogorov–arnold superposition theorem, 2020.

[46] D. Perekrestenko, P. Grohs, D. Elbrächter, and H. Bölcskei. The universal approximation power of finite-width deep relu networks. *arXiv preprint arXiv:1806.01528*, 2018.

[47] P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

[48] P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, Dec. 2018.

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

[50] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

[51] C. Robert and G. Casella. A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statistical Science*, 26(1):102 – 115, 2011.

[52] G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability surveys*, 1:20–71, 2004.

[53] G. O. Roberts and O. Stramer. Langevin diffusions and metropolis-hastings algorithms. *Methodology and computing in applied probability*, 4:337–357, 2002.

[54] G. O. Roberts and R. L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[55] P. J. Rossky, J. D. Doll, and H. L. Friedman. Brownian dynamics as smart monte carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, 1978.

[56] L. Ruthotto and E. Haber. An introduction to deep generative modeling. *GAMM-Mitteilungen*, 44(2):e202100008, 2021.

[57] M. Sander, P. Ablin, and G. Peyré. Do residual neural networks discretize neural ordinary differential equations? *Advances in Neural Information Processing Systems*, 35:36520–36532, 2022.

[58] Z. Shen, H. Yang, and S. Zhang. Deep Network With Approximation Error Being Reciprocal of Width to Power of Square Root of Depth. *Neural Computation*, 33(4):1005–1036, 03 2021.

[59] Z. Shen, H. Yang, and S. Zhang. Neural network approximation: Three hidden layers are enough. *Neural Networks*, 141:160–173, 2021.

[60] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[61] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta numerica*, 19:451–559, 2010.

[62] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2022.

[63] D. Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

[64] D. Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Conference on learning theory*, pages 639–649. PMLR, 2018.

[65] D. Yarotsky and A. Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. *Advances in neural information processing systems*, 33:13005–13015, 2020.

[66] K. S. Zhang, G. Peyré, J. Fadili, and M. Pereyra. Wasserstein control of mirror langevin monte carlo. In *Conference on Learning Theory*, pages 3814–3841. PMLR, 2020.

[67] Y. Zhang, Ö. D. Akyildiz, T. Damoulas, and S. Sabanis. Nonasymptotic estimates for stochastic gradient langevin dynamics under local conditions in nonconvex optimization. *Applied Mathematics & Optimization*, 87(2):25, 2023.