Diffusion at Absolute Zero: Langevin Sampling Using Successive Moreau Envelopes[*]

Andreas Habring[†], Alexander Falk[†], Martin Zach[‡], and Thomas Pock[†]

Abstract. We propose a method for sampling from Gibbs distributions of the form $\pi(x) \propto \exp(-U(x))$ that leverages a family $(\pi^t)_t$ of approximations of the target density which is deliberately constructed such that π^t exhibits favorable properties for sampling when t is large, and such that π^t approaches π as t approaches 0. This sequence is obtained by replacing (parts of) the potential U with its Moreau envelope. Through the sequential sampling from π^t for decreasing values of t by a Langevin algorithm with appropriate step size, the samples are guided from a simple starting density to the more complex target quickly. We prove that $t \mapsto \pi^t$ is Lipschitz continuous in the total variation distance and Hölder continuous in the Wasserstein-p distance, that the sampling algorithm is ergodic, and that it converges to the target density without assuming convexity or differentiability of the potential U. In addition to the theoretical analysis, we show experimental results that support the superiority of the method in terms of convergence speed and mode-coverage of multi-modal densities over current algorithms. The experiments range from one-dimensional toy-problems to high-dimensional inverse imaging problems with learned potentials.

Key words. Langevin diffusion, Markov chain Monte Carlo, sampling, inverse imaging

MSC codes. 65C40, 65C05, 68U10, 65C60

1. Introduction.

This article is concerned with sampling from distributions of the form

$$\pi(x) = \frac{\exp\left(-U(x)\right)}{\int \exp\left(-U(y)\right) \mathrm{d}y}. \tag{1.1}$$

The potential $U = F + G$ is composed of the functions $F, G : \mathbb{R}^d \to [0, \infty)$ and is such that $\int \exp\left(-U(x)\right) \mathrm{d}x < \infty$. The consideration of the additive structure $U = F + G$ is motivated by frequent applications where U can be split into a well-behaved function F and a more difficult-to-handle function G (*cf.* [23, 32]). In particular, we cover settings where G and, consequently, U are nonconvex or nondifferentiable.

Sampling from distributions of the form (1.1) is a task that frequently arises in, *e.g.*, Bayesian inverse problems [27], mathematical imaging [13, 35], machine learning [47, 48], and uncertainty quantification [32]. As an example, training strategies for machine learning such as maximum-likelihood estimation [9, 33, 47], which are becoming increasingly popular, often rely on sampling algorithms as subroutines. For most interesting potentials, sampling from (1.1) necessitates Markov chain Monte Carlo (MCMC) methods where a Markov chain $(X_k)_{k\geq 1} \subset \mathbb{R}^d$

[*] Submitted to the editors on March 27, 2025. The concepts presented in this manuscript have already been introduced in a (very preliminary) workshop paper submitted to *2025 IEEE Statistical Signal Processing Workshop*. This version of the manuscript is available at [17].

[†] Institute of Visual Computing, Graz University of Technology (andreas.habring@tugraz.at, falk@tugraz.at, thomas.pock@tugraz.at).

[‡] Biomedical Imaging Group, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland and Center for Biomedical Imaging, 1015 Lausanne, Switzerland (martin.zach@epfl.ch)

is deliberately designed such that the law of X_k converges to the target distribution π as $k \to \infty$. To tackle high-dimensional problems often encountered in imaging applications, Markov chains derived from discretizations of the Langevin diffusion stochastic differential equation (SDE)

$$\mathrm{d}X_t = -\nabla U(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \tag{1.2}$$

where $(W_t)_t$ denotes Brownian motion, have become popular. Note, however, that (1.2) requires U to be differentiable. A straightforward way to obtain a Markov chain that approximates (1.2) is via the Euler-Maruyama (EM) discretization that leads to the unadjusted Langevin algorithm (ULA)

$$X_{k+1} = X_k - \tau\nabla U(X_k) + \sqrt{2\tau}Z_k, \tag{1.3}$$

where $\tau > 0$ is the step size of the discretization and $(Z_k)_k$ are independent and identically distributed (i.i.d.) standard Gaussian random vectors. Various convergence results of such schemes (under stronger assumptions than in this paper) can be found in [8,10–12,16,18,25,38]. While more sophisticated schemes have been proposed, the EM discretization remains the most popular due to the simple implementation and diminishing returns for higher-order schemes in practice.

Practitioners face two challenges when using ULA: Slow mixing and strong assumptions on the potential. The former ultimately led to the birth of diffusion models in [42], while the latter sparked research on Langevin algorithms that are applicable to nondifferentiable and nonconvex potentials [13,28]. In this work, we propose diffusion at absolute zero (DAZ) as a novel sampling method that combines ideas from diffusion models with ideas from nondifferentiable Langevin sampling. In particular, we propose the successive Langevin sampling of a sequence of distributions π^t for $t \to 0$ which is obtained by replacing G with its Moreau envelope with parameter t. The Moreau envelope convexifies the potential (we formalize this statement in Theorem 4.1), allows using larger step sizes, and, consequently, has favorable mode-coverage behavior when t is large and converges to the target as $t \to 0$. In addition and in contrast to diffusion models, the proposed sampling algorithm does not require a trained model for all $t > 0$ since the minimum mean-squared-error (MMSE) denoising step is replaced with a maximim a-posteriori (MAP) denoising problem that can be solved with efficient optimization algorithms for a large class of potentials. Consequently, the proposed method is applicable to a large class of models used for inverse problems.

Contributions. We provide the following contributions with regards to the proposed sampling algorithm that is summarized in Algorithm 1.1:

1. In the nondifferentiable and nonconvex setting, we prove the exponential ergodicity of the Langevin diffusion as well as the geometric ergodicity of the corresponding discretization for any fixed Moreau parameter t. Moreover, we show that the stationary distribution of the discretization converges to that of the continuous-time Langevin diffusion as the step size vanishes (Theorems 4.14 and 4.15).
2. We show that the map $t \mapsto \pi^t$ is Lipschitz continuous with respect to the total variation distance, as well as Hölder continuous with Hölder exponent $p \in [1, \infty)$ under appropriate integrability assumptions on the density π^t (Theorem 4.22). This

Algorithm 1.1 Diffusion at Absolute Zero (DAZ)

Require: Number of Moreau levels N with schedule $\{t_n\}_{n=1}^N$ and corresponding Langevin step sizes $\{\tau_n\}_{n=1}^N$, number of Langevin steps per Moreau level K, initial condition X_1^N

```
for $n = N, \dots, 1$ do
    for $k = 1, \dots K$ do
        $Z_k^n \sim \mathcal{N}(0, I)$
        $X_{k+1}^n = X_k^n - \tau_n \nabla F(X_k^n) - \frac{\tau_n}{t_n}\left(X_k^n - \mathrm{prox}_{t_n G}(X_k^n)\right) + \sqrt{2\tau_n} Z_k^n$
    $X_0^{n-1} = X_K^n$
return $X_K^1$
```

continuity ensures that a small change in t implies a small change in π^t, which is a crucial requirement for a meaningful annealing strategy.

3. We relate the proposed sampling algorithm to diffusion models by showing that the distributions $(\pi^t)_t$ can be understood as a *zero-temperature* limit of the corresponding diffusion distributions (Theorems 4.24 and 4.26).
4. We provide an extensive set of numerical experiments. We demonstrate the efficiency of the proposed sampling algorithm quantitatively on toy examples that allow for an efficient computation of reference distributions and qualitatively on several high-dimensional applications in imaging that leverage potentials that are learned from data.

2. Related Work. In recent years, data-driven approaches have significantly influenced applied mathematics and related fields. A central aspect of these data-driven approaches is their close relation to probabilistic modeling, which naturally led to an increased interest in sampling strategies. In the following, we provide an overview of relevant works that are closely related to the present article. To facilitate the comparison between our work and some of these works, we revisit some of them—after the presentation of our results—in more detail in subsection 4.1.1.

Langevin sampling. A large body of work investigates conditions for ergodicity of the Langevin diffusion (1.2) and ULA (1.3) and the relation of the respective stationary distributions to the target under various assumptions on the potential U. Under the assumption that U is differentiable, the authors of the early work [38] show exponential ergodicity of the diffusion with stationary measure π, as well as geometric ergodicity of ULA under growth conditions on ∇U. Later works focused on establishing nonasymptotic convergence results with explicit rates. For strongly convex (at least outside a ball) and differentiable potentials with Lipschitz continuous gradient, explicit rates are provided in [6,8]. Weaker conditions on the growth of U or ∇U are considered in [11]. In [10], ULA (and similar schemes) are interpreted as an iterative minimization of the Kullback-Leibler divergence to the target density (*cf.* [20]) and the authors provide nonasymptotic convergence results.

The popularity of nondifferentiable regularizers (see, *e.g.*, [3,40]) in Bayesian inverse problems in imaging sparked research on sampling algorithms that can handle nondifferentiable potentials. Methods that rely on the subgradient or the proximal map of (nondifferentiable parts of) U are proposed and analyzed in [10, 16, 18]. The resulting sampling algorithms closely resemble proximal-gradient-style methods from optimization. In [4, 32], a primal-dual sampling scheme inspired by the Chambolle-Pock method for optimization [5] is used to tackle nondifferentiable potentials. The primal-dual-inspired methods are ad-hoc algorithms that have been demonstrated to work well in practice (see, *e.g.*, [32]), but convergence has only been proven under strong differentiability assumptions [4]. An alternative approach for sampling from nondifferentiable densities is to replace the target density with a differentiable surrogate. This has been proposed, *e.g.*, in [13, 35] where the nondifferentiable part of the potential is substituted with its Moreau envelope; the authors named the resulting algorithm Moreau-Yosida regularized unadjusted Langevin algorithm (MYULA). In [14], the convergence of such methods is analyzed when the proximal map is inexact.

The Moreau envelope-based approach from [13, 35] is extended to nonconvex potentials in [28] where a result about convergence of the EM discretization to the continuous-time diffusion in expectation for finite time is presented [28, Theorem 1]. We extend this by an extensive *ergodic analysis* of the Langevin diffusion and its discretization and a proof of the convergence of the stationary distribution of the discrete chain to the target as the step size vanishes (in the total variation (TV) norm, among others). Moreover, in [28] like in [13], the Moreau envelope was used to obtain a single surrogate density for π. In contrast, we consider a sequence of densities $(\pi^t)_t$ with different Moreau parameters that approach π, similar to diffusion models. The use of different Moreau parameters necessitates distinguishing the case of small parameters that lead to differentiable envelopes and large parameters where differentiability is not guaranteed. In both cases, we prove the ergodicity of the discrete chain.

In [37] the authors propose to replace the simple EM discretization used in ULA by a Runge-Kutta stochastic integration scheme which extends the deterministic Chebyshev method [41] to SDEs. The proposed method is coined stochastic orthogonal Runge–Kutta–Chebyshev method (SK-ROCK). While theoretical results about the convergence speed of SK-ROCK are unfortunately not yet available [37, Section 3.1.1], the method provides a significant acceleration in practice (see also section 5).

Annealed Langevin sampling. A particular inspiration for the present article has been annealed Langevin sampling [42], which later lead to diffusion models [43]. Annealed Langevin sampling is related to DAZ in the sense that both approaches propose to approximate the target π by a family of distributions $(\pi^t)_t$ which is designed so that π^t exhibits favorable properties for large t and so that $\pi^t \to \pi$ as $t \to 0$ in an appropriate sense. The difference between the two approaches lies in the definition of the family π^t. In annealed Langevin sampling, π^t is defined as the convolution of π with a Gaussian of variance t, *i.e.*, $\pi^t = \pi * \mathcal{N}(0, \sqrt{t})$. In DAZ, π^t is defined by replacing G by its Moreau envelope. In order to draw from π, both DAZ and annealed Langevin sampling sample successively from their respective regularized distributions π^{t_k} for $k = n, n-1, \dots, 0$ using ULA where $t_n > t_{n-1} > \dots > t_0$ is a predefined parameter schedule. Initially, for t_k large, the distribution π^{t_k} admits a rather favorable structure and, thus, allows for efficient sampling. As t_k is decreased, the samples are guided to the complex target distribution.

A crucial difference between annealed Langevin sampling and the proposed method is their applicability to a given and generic (*i.e.*, without special structure and not trained in a method-specific manner) potential U. In particular, when $\pi^t = \pi * \mathcal{N}(0, \sqrt{t})$ is constructed for annealed

Langevin sampling, one evaluation of $\nabla \log \pi^t$ (which is required for the Langevin sampling) at any point requires the computation of the MMSE-estimate of the denoising of that point under the prior π. For a general potential, for instance a deep neural network, this task is as hard as the original problem of sampling from π. This challenge is typically circumvented by the off-line direct training of a family of such MMSE denoisers [42]. However, this renders the method impractical for the sampling of a given potential U. In contrast, when $\pi^t(x) \propto \exp\big(-U^t(x)\big)$ is constructed for DAZ, one evaluation of $\nabla \log \pi^t$ can be computed efficiently whenever the proximal map of G can be computed efficiently. This can be accomplished by the resolution of a MAP denoising problem under a prior with potential G (see Theorem 3.2 and the discussion thereafter) by efficient algorithms for a large class of functions that includes neural networks. Consequently, the proposed method can be readily applied to a large class of potentials. The relation between annealed Langevin sampling and DAZ is formalized more rigorously in Theorems 4.24 to 4.26 where we show that the sequence of distributions used in the present work can be viewed as a zero-temperature limit of the sequence of distributions used in diffusion models.

Successive regularization. DAZ can be understood as a sampling analog to a successive regularization approach that has recently been proposed in the context of optimization. For the minimization of a possibly nonconvex and nondifferentiable function H, the authors of [19] propose to alternate gradient descent steps on the Moreau envelope of H and update steps on the Moreau parameter t. Since the Moreau envelope leaves global minima unchanged and, under certain conditions, local minimizers of the Moreau envelope are global minimizers of H [19, Lemma A.6], this constitutes a valid approach to finding the global minimizers of H. However, while for our sampling approach we require that the Moreau parameter approaches 0, this need not be the case for optimization [19].

3. Preliminaries.

In this section, we introduce some notations and mathematical preliminaries that are used throughout the paper. We start with the definition of the (regular) subdifferential, which is a crucial part in the rigorous treatment of nondifferentiable functions.

Definition 3.1 (Regular subdifferential). *The* (regular) subdifferential *of a function $H : \mathbb{R}^d \to \mathbb{R}$ at the point $x \in \mathbb{R}^d$ is the set*

$$\partial H(x) = \left\{ v \in \mathbb{R}^d \;\middle|\; \liminf_{\substack{y \to x \\ y \neq x}} \frac{H(y) - H(x) - \langle v, y - x \rangle}{\|y - x\|} \geq 0 \right\}.$$

An element $v \in \partial H(x)$ is referred to as a (regular) subgradient *of H at x.*

Our strategy relies heavily on the Moreau envelope, which we now formally define.

Definition 3.2 (Moreau envelope). *For a function $H : \mathbb{R}^d \to \mathbb{R}$ the* Moreau envelope *with* Moreau parameter *$t > 0$ is defined as*

$$M_H^t(x) \coloneqq \inf_{y \in \mathbb{R}^d} H(y) + \tfrac{1}{2t}\|x - y\|^2. \tag{3.1}$$

The Moreau envelope has the property that $M_H^t(x) \leq H(x)$ for all $x \in \mathbb{R}^d$ and $t > 0$ and, if H is proper, lower semicontinuous (l.sc.) and there exists $t > 0$ such that $M_H^t(x) > -\infty$ for some

$x \in \mathbb{R}^d$, the map $(x,t) \mapsto M_H^t(x)$ is continuous on $\mathbb{R}^d \times (0,\infty)$. In addition $M_H^t(x) \to H(x)$ for all $x \in \mathbb{R}^d$ as $t \to 0$, which motivates its use it as an approximation of H [39, Theorem 1.25].

Definition 3.3 (Proximal map). *The* proximal map *of a function $H : \mathbb{R}^d \to \mathbb{R}$ with Moreau parameter $t > 0$ assigns to any $x \in \mathbb{R}^d$ the (possibly multi-valued or empty) set*

$$\operatorname{prox}_{tH}(x) = \operatorname*{argmin}_{y\in\mathbb{R}^d} H(y) + \tfrac{1}{2t}\|x-y\|^2.$$

A crucial relationship is that $\nabla M_H^t(x) \in \frac{1}{t}\big(x - \operatorname{prox}_{tH}(x)\big)$ at all points $x \in \mathbb{R}^d$ where M_H^t is differentiable [39, Example 10.32]. This yields a practical way of computing the gradient of M_H^t by solving an optimization problem.

Probability. Let $\mathcal{B}(\mathbb{R}^d)$ be the Borel σ-algebra on $\mathbb{R}^d$. We denote the space of all probability measures on $\mathcal{B}(\mathbb{R}^d)$ as $\mathcal{P}(\mathbb{R}^d)$ and the subspace of all probability measures with bounded p-th moment as $\mathcal{P}_p(\mathbb{R}^d) := \big\{\mu \in \mathcal{P}(\mathbb{R}^d) \mid \int \|x\|^p \,\mathrm{d}\mu(x) < \infty\big\}$. The Dirac measure concentrated at $x \in \mathbb{R}^d$ is denoted as $\delta_x \in \mathcal{P}(\mathbb{R}^d)$. For two probability measures $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$, we denote their Wasserstein-p distance as

$$\mathcal{W}_p(\mu,\nu) = \inf_{X\sim\mu,\, Y\sim\nu} \mathbb{E}[\|X-Y\|^p]^{\frac{1}{p}} = \inf_{\gamma\in\Pi(\mu,\nu)} \left(\int \|x-y\|^p d\gamma(x,y)\right)^{\frac{1}{p}},$$

where $\Pi(\mu,\nu)$ denotes the set of all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν [45, Definition 1.6], and their TV distance as

$$\|\mu-\nu\|_{\mathrm{TV}} = \sup_{A\in\mathcal{B}(\mathbb{R}^d)} |\mu(A) - \nu(A)|.$$

Moreover, we define a Markov kernel on $\mathbb{R}^d$ as a map $M : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$ such that for any $x \in \mathbb{R}^d$, $M(x,\cdot)$ is a probability measure and for any $A \in \mathcal{B}(\mathbb{R}^d)$, $M(\cdot,A)$ is measurable. For $\mu \in \mathcal{P}(\mathbb{R}^d)$, we define the probability measure $\mu M \in \mathcal{P}(\mathbb{R}^d)$ by

$$\mu M(A) = \int M(y,A)\,\mathrm{d}\mu(y)$$

for any $A \in \mathcal{B}(\mathbb{R}^d)$. We denote the k-fold application of a Markov kernel M to a measure μ as $\mu M^k = \mu M \dots M$ (k times). A Markov kernel can also be interpreted as a linear map that maps bounded and measurable functions to bounded and measurable functions via $f \mapsto Mf$, with $Mf(x) = \int f(y)\,\mathrm{d}M(x,y)$.

Let $\mu \in \mathcal{B}(\mathbb{R}^d)$. A Markov semi-group is a family $(P_t)_{t\geq 0}$ where P_t is a linear operator that maps bounded and measurable functions to bounded and measurable functions and is such that $P_t 1 = 1$ with 1 denoting the constant function with value one, $P_t f \geq 0$ if $f \geq 0$, for every $f \in L^2(\mathbb{R}^d,\mu)$ it holds that $P_t f \to f$ in L^2 as $t \to 0$, for every $1 \leq p < \infty$, P_t extends to a bounded (contraction) operator on $L^p(\mathbb{R}^d,\mu)$, and $P_t \circ P_s = P_{t+s}$, $s,t \geq 0$. The generator $\mathcal{A}$ of a Markov semi-group is an operator defined via

$$\mathcal{A}f = \lim_{t\to 0} \frac{1}{t}(P_t f - f)$$

whose domain consists of all functions $f \in L^2(\mathbb{R}^d,\mu)$ for which the limit exists. For details on Markov semi-groups and their generators, see [1]. In most cases the the Markov semi-group is given as a family of Markov kernels [1, Section 1.2.2].

Miscellaneous. We denote the Lipschitz constant of a Lipschitz continuous function $H : \mathbb{R}^d \to \mathbb{R}$ as L_H. We denote the ball centered at $m \in \mathbb{R}^d$ with radius $R > 0$ as $B_R(m) := \{x \in \mathbb{R}^d \mid \|x - m\| \leq R\}$, and its complement in $\mathbb{R}^d$ as $B_R^c(m) = \mathbb{R}^d \setminus B_R(m)$.

4. Diffusion at Absolute Zero.

We propose DAZ to sample from π as defined in (1.1): DAZ combines ideas from diffusion methods [43] and nondifferentiable sampling [13] by considering the sequence of perturbed densities $\pi^t(x) \propto \exp\big(-U^t(x)\big)$ where

$$U^t(x) := F(x) + M_G^t(x)$$

for $t > 0$. That is, we replace the function G with its Moreau envelope, whereas F is left as is. This splitting strategy is advantageous when the proximal map of G can be computed efficiently and the proximal map of $F + G$ is difficult to compute. However, we show in Theorem 4.17 that the choice $G = U$ and $F \equiv 0$ has favorable properties with respect to the selectable step sizes when the proximal map of the original potential can be computed efficiently. As a sampling strategy, we apply annealed Langevin sampling to $(\pi^t)_t$. The proposed sampling algorithm is summarized in Algorithm 1.1. There, $0 < t_0 < t_1 < \cdots < t_N$ denotes a sequence of Moreau parameters and $(\tau_n)_n$ a corresponding sequence of step sizes used within the inner loop. This inner loop consists simply of several updates of the conventional ULA with step size τ_n applied to the potential U^{t_n}, *i.e.*,

$$X_n^{k+1} = X_n^k - \tau_n \nabla U^{t_n}(X_n^k) + \sqrt{2\tau_n} Z_n^k,$$

starting from some X_n^0 that is informed from the last iterate of the previous Moreau parameter. The update step in line 4 in Algorithm 1.1 is obtained by plugging in $U^t = F + M_G^t$ as well as $\nabla M_G^t(x) = \frac{1}{t}(x - \operatorname{prox}_{tG}(x))$. In [13], the main motivation for using the Moreau envelope was to handle nondifferentiable points of G. In the present work, we also exploit the *convexifying* behavior of the Moreau envelope. This behavior is exemplified in Figure 1 where the Moreau envelope connects the separated local minima of the potential as t increases. The convexifying property is now formalized.

Lemma 4.1. *Define the* nonconvexity *of a function $H : \mathbb{R}^d \to \mathbb{R}$ as the (possibly infinite) number*

$$\operatorname{NC}(H) := \sup_{\substack{x,y\in\mathbb{R}^d \\ \lambda\in[0,1]}} H(\lambda x + (1-\lambda)y) - \lambda H(x) - (1-\lambda)H(y). \tag{4.1}$$

Assume $\operatorname{prox}_{tH}(x)$ is nonempty for all $x \in \mathbb{R}^d$. Then, $\operatorname{NC}(M_H^t) \leq \operatorname{NC}(H)$.

Proof. Let x, x' be arbitrary points in $\mathbb{R}^d$ and let $p \in \operatorname{prox}_{tH}(x)$, $p' \in \operatorname{prox}_{tH}(x')$, $x_\lambda = \lambda x + (1-\lambda)x'$, and $p_\lambda = \lambda p + (1-\lambda)p'$. It follows by definition of the Moreau envelope that

$$\begin{aligned} &M_H^t(x_\lambda) - \lambda M_H^t(x) - (1-\lambda)M_H^t(x') \\ &\qquad \leq H(p_\lambda) - \lambda H(p) - (1-\lambda)H(p') + \tfrac{1}{2t}\|x_\lambda - p_\lambda\|^2 - \tfrac{\lambda}{2t}\|x - p\|^2 - \tfrac{1-\lambda}{2t}\|x' - p'\|^2 \\ &\qquad \leq \operatorname{NC}(H). \end{aligned}$$

Since $\frac{1}{2t}\|x_\lambda - p_\lambda\|^2 - \frac{\lambda}{2t}\|x - p\|^2 - \frac{1-\lambda}{2t}\|x' - p'\|^2 \leq 0$ due to (strict) convexity. Taking the supremum over x, x', and λ on the left-hand side concludes the proof. ■

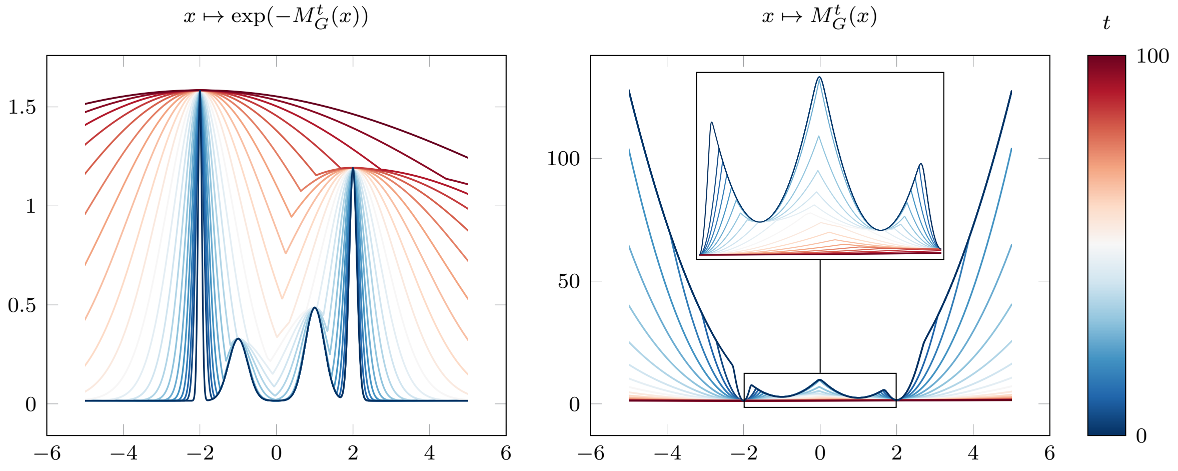

Figure 1. *Moreau envelopes of a Gaussian mixture for a sequence of Moreau parameters* $t \in [1\times 10^{-2}, 1\times 10^{2}]$*. Note how the Moreau envelope* convexifies *the potential with increasing* t*.*

Remark 4.2. $\mathrm{NC}(H) = 0$ if and only if H is convex: If $\mathrm{NC}(H) = 0$ it follows immediately that H is convex. If, on the other hand, H is convex, $\mathrm{NC}(H) \leq 0$ by definition and it can be lower bounded by 0 by choosing $x = y$.

The convexifying behavior of the Moreau enables DAZ to cover several modes of a multi-modal distribution more quickly.

We now give the assumptions on the potential $U = F + G$ that enable a thorough theoretical analysis of the proposed sampling algorithm. Remarks about the assumptions are discussed immediately below.

Assumption 4.3. We make the following assumptions on the potential $U = F + G$:

1. F is differentiable and its gradient is Lipschitz continuous.
2. G is locally bounded and weakly convex, *i.e.*, there exists a modulus of weak convexity $\rho_G > 0$ such that $G + \frac{\rho_G}{2}\|\cdot\|^2$ is convex. In addition, G is either Lipschitz continuous or such that $\int \exp\big(-G(x)\big)\,\mathrm{d}x < \infty$.
3. F and G are convex outside a ball, *i.e.*, there exists an $M > 0$ such that for any $x, y \in B_M^c(0)$ it holds that

$$G(\lambda x + (1-\lambda)y) \leq \lambda G(x) + (1-\lambda)G(y)$$

and identically for F.

4. Let $t_{\max} < \frac{1}{\rho_G}$ and define $\phi(s) := \sup\{\|p\| \mid p \in \partial G\big(\mathrm{prox}_{tG}(x)\big),\ \|x\| \leq s,\ t \leq t_{\max}\}$. It holds true that

$$\int \phi(\|x\|)^2 \exp\big(-U^{t_{\max}}(x)\big)\,\mathrm{d}x =: K_{\max} < \infty. \tag{4.2}$$

Remark 4.4. We provide a list of implications of Theorem 4.3 to facilitate the understanding of the types of potentials our theory covers. Moreover, some of the implications will be used within the proofs.

1. The weak convexity of G is relevant thrice. First, it implies that G admits a regular subgradient everywhere: The function $\widetilde{G} \coloneqq G + \frac{\rho_G}{2}\|\cdot\|^2$ admits a (convex) subgradient everywhere and by [39, 8.8 Exercise, 8.12 Proposition] it follows that $\partial G(x) = \partial\widetilde{G}(x) - \rho_G x$. Second, when $t < \frac{1}{\rho_G}$ the map $y \mapsto \frac{1}{2t}\|x-y\|^2 + G(y)$ is strongly convex and, consequently, for such t the proximal map of G is at most single-valued. Third, it implies local boundedness of the subgradients of G, that is, boundedness of the set
$$\bigcup_{x\in B_R(0)} \partial G(x)$$
for any $R > 0$: Since $\widetilde{G}$ is bounded from below and locally bounded from above, by [39, 9.14 Example], $\widetilde{G}$ is locally Lipschitz continuous and, consequently, $\partial\widetilde{G}$ is locally bounded by the respective Lipschitz constant. Local boundedness of ∂G follows from $\partial G(x) = \partial\widetilde{G}(x) - \rho_G x$. Moreover, local Lipschitz continuity of G also implies *global* continuity. Continuity of G will be used frequently, *e.g.*, in order to ensure continuity of the proximal map in Theorem 4.6. Local boundedness of ∂G is needed to ensure coercivity of the proximal map in Theorem 4.9.
2. Since G is bounded from below and continuous, $\operatorname{prox}_{tG}(x)$ is nonempty for all $x \in \mathbb{R}^d$.
3. The assumption that G is either Lipschitz continuous or satisfies the integrability condition $\int \exp\big(-G(x)\big)\,\mathrm{d}x < \infty$ ensures that $\int \exp\big(-U^t(x)\big)\,\mathrm{d}x < \infty$, *i.e.*, that the potential U^t defines a proper Gibbs distribution, by an adaptation of [13, Proposition 1] together with Theorem A.1. Without this assumption, the smoothing properties of the Moreau envelope might lead to a violation of the growth requirements on U^t that are needed for the integrability of $x \mapsto \exp\big(-U^t(x)\big)$.
4. The definition of *convexity outside a ball* implies also a *first-order* version, that is, for any $x, y \in B_M^c(0)$ and $v_x \in \partial G(x)$, $v_y \in \partial G(y)$ it holds true that
$$\langle v_x - v_y, x - y\rangle \geq 0.$$
The converse implication, however, is only true in a weaker sense. First order convexity outside the ball $B_M(0)$ implies $G(\lambda x + (1-\lambda)y) \leq \lambda G(x) + (1-\lambda)G(y)$ only whenever the entire line connecting x and y is outside the ball, *i.e.*, $[x, y] \subset B_M^c(0)$.
5. The last integrability assumption is technical and typically no issue in practice. When G is Lipschitz continuous (*e.g.*, G in the ℓ^1 norm or TV) it is equivalent to integrability of $x \mapsto \exp\big(-U^{t_{\max}}(x)\big)$. Otherwise, Theorem A.1 shows that there there exist $r, c > 0$ such that $U(x) \geq c\|x\|$ for all $x \in B_r^c(0)$. The same holds true for $U^{t_{\max}}$ as a consequence of Theorem 4.11. Therefore, for x large enough $\exp\big(-U^{t_{\max}}(x)\big) \leq \exp\left(-c\|x\|\right)$, which leads to (4.2) being satisfied as long as $\partial G(\operatorname{prox}_{tG}(x))$ does not grow exponentially fast in $\|x\|$. In particular, if $U^{t_{\max}}$ is superexponential (*cf.*, Theorem 4.12 below) it also follows that for $\|x\|$ sufficiently large $\|\operatorname{prox}_{tG}(x)\| \leq \|x\| + c'$ for some $c' > 0$ (Theorem A.2). Since G is locally Lipschitz, it is also almost everywhere differentiable

and the constraint reduces to integrability of

$$x \mapsto \sup_{\|y\|\leq\|x\|+c'} \|\nabla G(y)\|^2 \exp\left(-c\|x\|\right).$$

4.1. Ergodicity of Diffusion at Absolute Zero.

In this section we investigate ergodicity and convergence properties of the ULA subroutine in Algorithm 1.1 for a fixed Moreau parameter t_n. This is the Markov chain

$$\begin{cases} X_0 = x_0, \\ X_{k+1} = X_k - \tau\nabla F(X_k) - \frac{\tau}{t}\left(X_k - \operatorname{prox}_{tG}(X_k)\right) + \sqrt{2\tau}Z_k, \end{cases} \tag{4.3}$$

initialized at an arbitrary $x_0 \in \mathbb{R}^d$. We omit the subscript n since the Moreau parameter is fixed. We denote the Markov kernel that corresponds to one iteration of (4.3) as R_τ, that is, if $X_0 \sim \mu$, then $X_1 \sim \mu R_\tau$. This problem has been thoroughly analyzed in [13] for convex G and differentiable F with Lipschitz continuous gradient. We recall a shortened version of the main result from [13, Section 3.2].

Theorem 4.5. *Let F and G be lower bounded, F convex and differentiable with Lipschitz continuous gradient, and G proper, convex, and l.sc. In addition, assume that G is either Lipschitz continuous or such that $\int \exp\left(-G(x)\right) \mathrm{d}x < \infty$. Then for any $x_0 \in \mathbb{R}^d$ the Markov chain (4.3) is geometrically ergodic, i.e., there exists a probability measure π^t_τ, and constants $C > 0$, $\kappa \in (0, 1)$ such that*

$$\|\delta_{x_0}R^k_\tau - \pi^t_\tau\|_{\mathrm{TV}} \leq C\kappa^k.$$

Moreover, $\|\pi^t_\tau - \pi^t\|_{\mathrm{TV}} \to 0$ as $\tau \to 0$.

In addition to ergodicity, explicit and nonasymptotic convergence rates can be found in [13].

In the sequel, we analyze the ergodicity of (4.3) in the nonconvex case. We provide two main results: The ergodicity of the Markov chain independent of the Moreau parameter t (Theorem 4.14) and a stronger result that proves the convergence of the chain to the continuous-time diffusion when $t < \frac{1}{\rho_G}$ (Theorem 4.15). The theorems rely on several preliminary results concerning the Moreau envelope and the proximal map. Parts of these are covered in [39] but we include results that are tailored to the setting in this work in order to provide a self-contained article. The proofs of Theorems 4.6, 4.7, 4.9, 4.10, and 4.11 are given in Subsections A.1.1 to A.1.5.

Lemma 4.6. *The map $(x, t) \mapsto \operatorname{prox}_{tG}(x)$ is continuous on $\mathbb{R}^d \times \left(0, \frac{1}{\rho_G}\right)$ and, for $t \in \left(0, \frac{1}{\rho_G}\right)$, $x \mapsto \operatorname{prox}_{tG}(x)$ is Lipschitz continuous with Lipschitz constant $\frac{1}{1-\rho_G t}$.*

Lemma 4.7. *For $t \in \left(0, \frac{1}{\rho_G}\right)$, M^t_G is differentiable and the gradient is given by $\nabla M^t_G(x) = \frac{1}{t}\left(x - \operatorname{prox}_{tG}(x)\right)$.*

Corollary 4.8. *For $t \in \left(0, \frac{1}{\rho_G}\right)$, ∇M^t_G is Lipschitz continuous with Lipschitz constant $\frac{2-\rho_G t}{t(1-\rho_G t)}$.*

Proof. Since $\nabla M^t_G(x) = \frac{1}{t}\left(x - \operatorname{prox}_{tG}(x)\right)$ the result follows from Lipschitz continuity of the proximal map. ■

When $t \geq \frac{1}{\rho_G}$, regularity properties of the Moreau envelope do not follow as directly. However, we can obtain guarantees by using convexity of G outside the ball $B_M(0)$. To proceed, we require that $\operatorname{prox}_{tG}(x)$ is outside of $B_M(0)$ for sufficiently large $x \in \mathbb{R}^d$, which the following lemma asserts.

Lemma 4.9. *The proximal map is coercive in the sense that for all $t > 0$ it holds that*

$$\lim_{R\to\infty} \inf_{x\in B_R^c(0)} \inf_{p\in \operatorname{prox}_{tG}(x)} \|p\| = \infty. \tag{4.4}$$

Corollary 4.10. *For any $t > 0$, there exists $R > 0$ such that M_G^t is differentiable and prox_{tG} is single-valued and 1-Lipschitz on $B_R^c(0)$. That is,*

$$\|\operatorname{prox}_{tG}(x) - \operatorname{prox}_{tG}(y)\| \leq \|x - y\|, \quad x, y \in B_R^c(0).$$

In addition, $\|\nabla M_G^t(x) - \nabla M_G^t(y)\| \leq \frac{1}{t}\|x - y\|$ for $x, y \in B_R^c(0)$.

Corollary 4.11. *The Moreau envelope M_G^t is convex outside a ball. That is, there exists $R > 0$ such that for $x, y \notin B_R(0)$ it holds that*

$$M_G^t(\lambda x + (1-\lambda)y) \leq \lambda M_G^t(x) + (1-\lambda)M_G^t(y).$$

Within the proof of convergence of ULA to the target density, we will make use of the following growth property [11, Section 3].

Definition 4.12 (Superexponential). *A function $H : \mathbb{R}^d \to \mathbb{R}$ is* superexponential *if there exist a minimizer $x^* \in \mathbb{R}^d$ of H and $\rho, M_\rho > 0$ such that for any $x \in B_{M_\rho}^c(x^*)$ and $v \in \partial H(x)$ it holds that*

$$\langle v, x - x^*\rangle \geq \rho\|x - x^*\|^2. \tag{4.5}$$

We refer to M_ρ and ρ as the supex radius *and* supex modulus *of H, respectively.*

Lemma 4.13. *The Moreau envelope preserves the superexponential property. More specifically, if G is superexponential with radius M_ρ and modulus ρ, then there exists $M_\rho' > 0$ such that for all $x \in B_{M_\rho'}^c(x^*)$*

$$\langle \nabla M_G^t(x), x - x^*\rangle \geq \min\left(\frac{\rho}{4}, \frac{1}{2t}\right)\|x - x^*\|^2. \tag{4.6}$$

The proof is given in Subsection A.1.6. We are now in the position to prove the main results. The first results concerns ergodicity of the ULA chain, that is, existence of a unique stationary distribution and convergence of the chain to said distribution.

Theorem 4.14 (Ergodicity for arbitrary $t > 0$). *Let Theorem 4.3 be satisfied and assume that G or F is superexponential. Then, for a small enough $\tau > 0$, the chain (4.3) is geometrically ergodic. That is, there exists a stationary measure π_τ^t and $\lambda \in (0,1)$ such that for any $x_0 \in \mathbb{R}^d$ there exists $C > 0$ with*

$$\|\delta_{x_0} R_\tau^k - \pi_\tau^t\|_{\mathrm{TV}} \leq C\lambda^k$$

for $k \in \mathbb{N}$.

Proof. The result follows as in [16, Theorem 5.3] if we can prove that the potential U^t is superexponential. Therefore, we now show that F or G being superexponential implies that U^t is superexponential with appropriate radius and modulus. When G is superexponential, Theorem 4.13 implies that M_G^t is superexponential with appropriate radius and modulus which we denote as M_ρ and ρ. Using convexity of U^t outside a ball Theorem A.1 shows that U^t grows asymptotically at least linearly. Thus U^t is coercive, bounded from below, and continuous and, consequently, admits a minimizer. Let x_G^* be a minimizer of G (and, consequently, M_G^t) and x_U^* be a minimizer of U^t. Moreover, recall that M denotes the radius of convexity of F and G. Pick $M'_\rho > \max(M_\rho, M)$ such that $x \in B^c_{M'_\rho}(x_U^*)$ implies that $\|x\| > M$ and, consequently, differentiability and convexity of M_G^t in a neighborhood around x. Consequently, for $x \in B^c_{M'_\rho}(x_U^*)$ it follows that

$$\begin{aligned}
\left\langle \nabla U^t(x), x - x_U^* \right\rangle &= \underbrace{\left\langle \nabla F(x) - \nabla F\left(M'_\rho \frac{x}{\|x\|}\right), x - M'_\rho \frac{x}{\|x\|} \right\rangle}_{\geq 0 \text{ by convexity outside } B_M(0)} \\
&\quad + \left\langle \nabla F\left(M'_\rho \frac{x}{\|x\|}\right), x - M'_\rho \frac{x}{\|x\|} \right\rangle \\
&\quad + \left\langle \nabla F(x), M'_\rho \frac{x}{\|x\|} - x_U^* \right\rangle + \langle \nabla M_G^t(x), x - x_U^* \rangle \\
&\geq \langle \nabla M_G^t(x), x - x_U^* \rangle - c_1\|x\| - c_2 \\
&\geq \langle \nabla M_G^t(x), x - x_G^* \rangle + \langle \nabla M_G^t(x), x_G^* - x_U^* \rangle - c_1\|x\| - c_2 \\
&\geq \rho\|x - x_U^*\|^2 - \tilde{c}_1\|x\| - \tilde{c}_2
\end{aligned} \tag{4.7}$$

where we used Lipschitz continuity of ∇F and ∇M_G^t to obtain the linear bounds with appropriate constants $\tilde{c}_1, \tilde{c}_2 > 0$. Thus, U^t is superexponential. In the case that instead F is superexponential, we can use the convexity of M_G^t outside a ball and apply the same arguments. As in [16, Theorem 5.3] this yields the existence of a probability distribution π_τ^t, a constant $C > 0$, and $\lambda \in (0, 1)$ such that

$$\|\delta_{x_0} R_\tau^k - \pi_\tau^t\|_V \leq C\lambda^k$$

where

$$\|\nu\|_f := \sup_{g:|g|\leq f} \left| \int g(x) \, \mathrm{d}\nu(x) \right|. \tag{4.8}$$

and $V(x) = \exp\left(\|x - x_U^*\|\right)$. In particular, since for any measurable set A and $x \in \mathbb{R}^d$ it holds that $|1_A(x)| \leq V(x)$ we obtain that

$$\|\delta_{x_0} R_\tau^k - \pi_\tau^t\|_{\mathrm{TV}} = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} \left| \int 1_A(x) \, \mathrm{d}(\delta_{x_0} R_\tau^k - \pi_\tau^t)(x) \right| \leq \|\delta_{x_0} R_\tau^k - \pi_\tau^t\|_V,$$

which proves the desired convergence in TV. ■

When $t < \frac{1}{\rho_G}$ we obtain convergence of the stationary measure π^t_τ to the target π^t when the step size τ vanishes in addition to ergodicity.

Theorem 4.15 (Convergence for $t < \frac{1}{\rho_G}$). *Assume that $t < \frac{1}{\rho_G}$. Let Theorem 4.3 be satisfied and assume that G or F is superexponential. Then, the stationary measure π^t_τ of the discrete time Markov chain (4.3) approaches π^t as $\tau \to 0$ in TV. More precisely, for any $\epsilon > 0$ there exists $\bar\tau > 0$ and $N > 0$ such that for $\tau < \bar\tau$ and $\|\pi^t_\tau - \pi^t\|_{\mathrm{TV}} < \epsilon$ and if $k > N$, $\|\delta_{x_0} R^k_\tau - \pi^t\|_{\mathrm{TV}} < \epsilon$ and it holds $\bar\tau = \mathcal{O}(-\log^{-1}(\epsilon)\epsilon^2)$ and $N = \mathcal{O}(\log^2(\epsilon)\epsilon^{-2})$.*

Proof. The ergodicity of the discrete chain was already proved in Theorem 4.14. To prove that $\pi^t_\tau \to \pi^t$ as $\tau \to 0$, we follow the three-step strategy of [11]. In the first step, we establish exponential ergodicity of the corresponding continuous-time Langevin diffusion with stationary measure π^t. In the second step, we combine the exponential ergodicity with the approximation of the continuous time diffusion by the discrete chain to bound $\|\delta_{x_0} R^k_\tau - \pi^t\|_{\mathrm{TV}}$. This bound can be made small as $\tau \to 0$ if the second moments of the iterates of the discrete chain are bounded uniformly in τ. Thus, in the third step we establish such a uniform bound.

Step 1: Ergodicity of the continuous time process. The chain (4.3) is a discretization of the Langevin SDE on the potential U^t defined via

$$\begin{cases} X_0 = x_0 \\ \mathrm{d}X_s = -\nabla U^t(X_s)\mathrm{d}t + \sqrt{2}\mathrm{d}W_s. \end{cases} \tag{4.9}$$

The existence of a unique solution for all time is a standard result when ∇U^t is Lipschitz continuous, which is guaranteed for $t < \frac{1}{\rho_G}$. As in [38, Theorem 2.1] it follows that $(X_s)_{s>0}$ is nonexplosive, irreducible with respect to the Lebesgue measure, strong Feller and, as a consequence, all compact sets are small. The SDE defines a Markov semi-group via $P_s f(x) = \mathbb{E}[f(X_s)|X_0 = x]$. By Ito's lemma, the generator of the semi-group is given by

$$\mathcal{A}V(x) = \lim_{s\to+0} \frac{\mathbb{E}[V(X_s)|X_0 = x] - V(x)}{s} = -\langle \nabla U^t(x), \nabla V(x)\rangle + \Delta V(x)$$

for any twice continuously differentiable $V : \mathbb{R}^d \to \mathbb{R}$ (*cf.* [16, Lemma 4.5]). In particular, for $V(x) = \|x - x^*_U\|^2$ the generator reads

$$\mathcal{A}V(x) = -2\langle \nabla U^t(x), x - x^*_U\rangle + 2d. \tag{4.10}$$

For any $x \in B^c_{M_\rho}(x^*_U)$, the generator satisfies the drift condition

$$-2\langle \nabla U^t(x), x - x^*_U\rangle + 2d \le -2\rho\|x - x^*_U\|^2 + 2d \le -2\rho V(x) + 2d \tag{4.11}$$

since U^t is superexponential. By [30, Theorem 6.1], the combination of the drift condition with the fact that all compact sets are small implies the existence of $B > 0$, $\kappa \in (0,1)$, and a stationary measure (which has to be the target measure π^t as shown in [16, Theorem 3]) such that

$$\|\delta_{x_0} P_s - \pi^t\|_{V+1} \le (\|x_0 - x^*_U\|^2 + 1)B\kappa^s. \tag{4.12}$$

It follows for any μ with bounded second moment that

$$
\begin{aligned}
\|\mu P_s - \pi^t\|_{V+1} &= \sup_{g \le V+1} \left| \int g(x) \mathrm{d}(\mu P_s - \pi^t)(x) \right| \\
&= \sup_{g \le V+1} \left| \int \int g(x) \mathrm{d}P_s(y,\cdot)(x) \mathrm{d}\mu(y) - \int g(x) \mathrm{d}\pi^t(x) \right| \\
&= \sup_{g \le V+1} \left| \int \int g(x) \mathrm{d}(P_s(y,\cdot) - \pi^t)(x) \mathrm{d}\mu(y) \right| \\
&\le \sup_{g \le V+1} \left| \int (\|y - x_U^*\|^2 + 1) B\kappa^t \mathrm{d}\mu(y) \right| \le \left(\int \|y - x_U^*\|^2 \, \mathrm{d}\mu(y) + 1 \right) B\kappa^s.
\end{aligned} \tag{4.13}
$$

As previously, since $V+1$ dominates $|1_A|$ for arbitrary measurable sets A, the inequality also holds for the TV norm.

Step 2: Estimating the error. Let us define $C : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ as $C(\mu) = (\int \|x - x_U^*\|^2 \, \mathrm{d}\mu(x) + 1)B$, the *constant* in (4.13). Proposition 2 in [11] states that for any $k, n \in \mathbb{N}$, $k > n > 0$

$$
\begin{aligned}
\|\delta_{x_0} R_\tau^k - \pi^t\|_{\mathrm{TV}} &\le \|\delta_{x_0} R_\tau^k - \delta_{x_0} R_\tau^n P_{(k-n)\tau}\|_{\mathrm{TV}} + \|\delta_{x_0} R_\tau^n P_{(k-n)\tau} - \pi^t\|_{\mathrm{TV}} \\
&\le \frac{L_{\nabla U^t}}{\sqrt{2}} \sqrt{k-n} \left(\frac{\tau^3}{3} A(x_0,\tau) + d\tau^2 \right)^{1/2} + C(\delta_{x_0} R_\tau^n) \kappa^{(k-n)\tau}
\end{aligned} \tag{4.14}
$$

where $A(x_0,\tau) = \sup_{k \ge 0} \int \|\nabla U^t(y)\|^2 \, \mathrm{d}(\delta_{x_0} R_\tau^k)(y)$. The bound on the first term results from the discretization (4.3) of the SDE and the bound on the second term from exponential ergodicity of the continuous-time diffusion. Thus, it remains to bound the constants $A(x_0,\tau)$ and $C(\delta_{x_0} R_\tau^n)$ where it is crucial that the bounds are independent of τ. By their respective definitions and Lipschitz continuity of ∇U^t, it is sufficient to bound the second moments of the Markov chain $(X_k)_k$.

Step 3: Bounding the second moments of $(X_k)_k$. The proof is adapted from [16] and provided for the sake of completeness. Since U^t is superexponential, there exists $M_\rho, \rho > 0$ such that for all $x \in \mathbb{R}^d$ such that $\|x - x_U^*\| \ge M_\rho$ it holds that

$$
\begin{aligned}
\|x - \tau \nabla U^t(x) - x_U^*\|^2 &\le \|x - x_U^*\|^2 - 2\tau\rho \|x - x_U^*\|^2 + \tau^2 (L_{\nabla U^t} \|x - x_U^*\|)^2 \\
&\le \|x - x_U^*\|^2 \big(1 - \tau(2\rho - \tau L_{\nabla U^t}^2)\big)
\end{aligned} \tag{4.15}
$$

Thus, if $\tau \le \frac{\rho}{L_{\nabla U^t}^2}$, we find that $\|x - \tau \nabla U^t(x) - x_U^*\|^2 \le \|x - x_U^*\|^2 (1 - \tau\rho)$. Further, for $x \in \mathbb{R}^d$ with $\|x - x_U^*\| < M_\rho$, the Lipschitz continuity of ∇U^t implies that

$$
\begin{aligned}
\|x - \tau \nabla U^t(x) - x_U^*\| &\le \|x - x_U^*\| + L_{\nabla U^t} \tau \|x - x_U^*\| \\
&\le (1 + L_{\nabla U^t} \tau) M_\rho =: M_\tau.
\end{aligned} \tag{4.16}
$$

For the iterates of our chain, this yields (by induction)

$$\begin{aligned}\|X_k - x_U^*\|^2 \le 1_{\{\|X_n - x_U^*\| > R,\, n=0,\dots,k-1\}}(1-\rho\tau)^k\|X_0 - x_U^*\|^2 \\ + \sum_{n=0}^{k-1}\Bigg\{ M_\tau^2 1_{\|X_n - x_U^*\|\le R}\prod_{\ell=n+1}^{k-1} 1_{\|X_\ell - x_U^*\|>R} + (1-\rho\tau)^n\Big[\\ 2\tau\|W_{k-1-n}\|^2 + 2\langle X_{k-1-n}^\tau - \tau\partial U(X_{k-1-n}^\tau) - x_U^*, \sqrt{2\tau}W_{k-1-n}\rangle\Big]\Bigg\}.\end{aligned}$$

Taking the expectation and noting that W_{k-1-n} and X_{k-1-n} are independent, it follows that

$$\begin{aligned}\mathbb{E}\big[\|X_k - x_U^*\|^2\big] &\le (1-\rho\tau)^k\mathbb{E}\big[\|X_0 - x_U^*\|^2\big] + M_\tau^2\underbrace{\mathbb{E}\left[\sum_{n=0}^{k-1} 1_{\|X_n - x_U^*\|\le R}\prod_{\ell=n+1}^{k-1} 1_{\|X_\ell - x_U^*\|>R}\right]}_{\le 1} \\ &\quad + \sum_{n=0}^{k-1}(1-\rho\tau)^n 2\tau\mathbb{E}\big[\|W_{k-1-n}\|^2\big] \\ &\le (1-\rho\tau)^k\mathbb{E}\big[\|X_0 - x_U^*\|^2\big] + M_\tau^2 + \sum_{n=0}^{k-1}(1-\rho\tau)^n 2\tau \\ &\le (1-\rho\tau)^k\mathbb{E}\big[\|X_0 - x_U^*\|^2\big] + M_\tau^2 + \frac{2}{\rho},\end{aligned}$$

which is bounded as long as τ is bounded. Thus, the second moments of $(X_k)_k$ are bounded uniformly with respect to τ. As a consequence for any x_0, $\sup_\tau A(x_0,\tau) < \infty$ and $\sup_n C(\delta_{x_0}R_\tau^n) < \infty$. Now fix $k-n=m$ in (4.14) and let $k\to\infty$. It follows that $\|\pi_\tau^t - \pi^t\|_{\mathrm{TV}} \le C_1\sqrt{m\tau} + C_2\kappa^{m\tau}$. Now for arbitrary $\epsilon > 0$ define $T > 0$ such that $C_2\kappa^{T-1} < \frac{\epsilon}{2}$. Then for $\tau < \min\Big(\frac{\epsilon^2}{4c_1^2 T}\Big)$ it follows with $m = \lfloor\frac{T}{\tau}\rfloor$ that $\|\pi_\tau^t - \pi^t\|_{\mathrm{TV}} \le \epsilon$ which is the desired convergence and the complexity $\tau = \mathcal{O}(-\log^{-1}(\epsilon)\epsilon^2)$ and, since $k \ge m$, for the number of iterations $k = \mathcal{O}(\log^2(\epsilon)\epsilon^2)$. ■

Remark 4.16. We show later that the density π^t approaches π as $t\to 0$ in the TV distance and, consequently, DAZ constitutes a method for (approximate) sampling from nondifferentiable *and* non-log-concave densities.

Remark 4.17. The crucial inequality that guarantees ergodicity and enables one to bound the bias of the discretization in the previous proofs (see also [16, Proposition 5.3]) is

$$\tau < \tau_{\max} = \frac{2\rho}{L_{\nabla U^t}^2}$$

where ρ is the supex modulus of U^t (see, *e.g.*, (4.15)). Both, ρ and $L_{\nabla U^t}$ depend on the Moreau parameter t, which begs the question how the step-size requirements depend on t. Let us, therefore, distinguish the two different Moreau parameter regimes:

1. Behavior for large t: For $U^t = F + M_G^t$, it holds that $L_{\nabla U^t} \le L_{\nabla F} + \frac{1}{t}$ [13, Section 3.3] where the upper bound saturates at $L_{\nabla F}$ for $t\to\infty$. However, this can be circumvented

by the choice $F \equiv 0$ and $G = U$, in which case U^t inherits the superexponential property directly from G due to Theorem 4.13. Moreover, we find that $\rho \geq \frac{1}{2t}$ when t is large enough. Combining this with $L_{\nabla U^t} \leq \frac{1}{t}$ from Theorem 4.10 we find that

$$\tau_{\max} = \frac{2\rho}{L^2_{\nabla U^t}} \geq \frac{\frac{1}{t}}{\frac{1}{t^2}} = t. \tag{4.17}$$

Thus, the choice $F \equiv 0$ and $G = U$ enables the use of increasingly large step sizes for increasingly large t.

2. Behavior for small t: The above estimate on $\tau_{\max}$ tends to zero as $t \to 0$. However, if the original potential $U = G$ admits a Lipschitz gradient with Lipschitz constant L, we can improve the estimate. First, for $t \to 0$, we obtain that ρ saturates due to Theorem 4.13. Second, we can show that the Lipschitz constant of $\nabla U^t = \nabla M^t_G$ converges to that of $\nabla U = \nabla G$. To do so, we need to bound

$$\|\nabla M^t_G(x) - \nabla M^t_G(y)\| = \frac{1}{t}\|x - \operatorname{prox}_{tG}(x) - (y - \operatorname{prox}_{tG}(y))\|.$$

To this end, let us denote $q_x = x - \operatorname{prox}_{tG}(x)$ and $q_y = y - \operatorname{prox}_{tG}(y)$. It follows that $q_x = t\nabla G(x - q_x)$ and analogously for y. Therefore, $\|q_x - q_y\| \leq tL(\|x-y\| + \|q_x - q_y\|)$ and, consequently,

$$\|\nabla M^t_G(x) - \nabla M^t_G(y)\| \leq \frac{L}{1 - tL}\|x - y\|.$$

We find that as $t \to 0$, the Lipschitz constant *saturates* at L and the step size τ can be chosen as $\frac{\rho}{2L^2}$.

Consequently, when the proximal map of the composite potential can be computed efficiently, it is advantageous to choose $F \equiv 0$. When this computation is inefficient, it is typically advantageous to split the composite potential at the cost of stronger restrictions on the step size.

4.1.1. Comparison to related works.

In the following we put the presented ergodicity results into context with respect to the literature on Langevin based sampling. The strongest assumptions on the function G we have made in this paper are weak convexity, convexity outside a ball, and superexponential growth. Thus, the main contribution in comparison to the literature is that we provide theoretical results assuming neither convexity nor differentiability of the original potential U.

Early works on Langevin sampling such as [8,12] provide theoretical guarantees under the assumption that the potential is differentiable *and* convex. In [46], guarantees are provided for distributions that satisfy the log-Sobolev inequality, which is implied by strong log-concavity. Despite the fact that the proximal algorithm in [46] can formally be applied also in the non-differentiable case, the analysis still requires a potential that is thrice continuously differentiable and has a Lipschitz continuous gradient.

Theorem 9 in [11] is a result about ergodicity and convergence of ULA for potentials which are superexponential but potentially nonconvex. However, the analysis assumes that the potential is twice continuously differentiable and has a Lipschitz continuous gradient. In our approach, for the *original* potential U, we assume neither of these properties. For the *regularized* potential U^t, we only require that it has a Lipschitz continuous gradient—which is guaranteed through its construction—but do not assume that it is twice continuously differentiable. This relaxation is particularly relevant for DAZ as the Moreau envelope is frequently not twice continuously differentiable. A prominent example thereof is $G = \|\cdot\|_1$, whose Moreau envelope is a sum of Huber functions that are only once continuously differentiable. In [11, Theorem 12] the authors additionally prove a separate result about ergodicity and convergence without assuming that the potential is twice differentiable, and provide more explicit constants that depend on the dimension of the space by invoking a Poincaré inequality. Indeed, Theorem 12 in [11] is applicable for sampling from U^t in the inner loop of DAZ for fixed t. However, the result requires a non-constant sequence of step sizes $(\tau_n)_n$ which is square summable but not summable, which prohibits the direct control of the bias (see the discussion immediately after [11, Theorem 12]). The works [10, 13, 16, 18, 35] focus on results that heavily rely on the convexity of the potential with less smoothness assumptions. In [28] the authors also consider potentials which are weakly convex and nondifferentiable. Like in the proposed approach, they tackle the nondifferentiability of the potential by using Moreau envelopes. However, the work does not provide any results about the ergodicity of the continuous-time SDE solution or the discretization of it. Moreover, error bounds between continuous-time SDE and discrete Markov chain are provided only in expectation and for finite time.

4.2. Consistency of Diffusion at Absolute Zero. The proposed approach of successively sampling from π^t with decreasing values of t only makes sense if (i) the distributions π^t are similar for similar values of t and (ii) $\pi^t \to \pi$ as $t \to 0$. We verify these properties in this section. Recall that in the following $t_{\max} > 0$ is fixed and satisfies $t_{\max} < \frac{1}{\rho_G}$ (*cf.* Assumption 4.3.4). The main results will make use of the following preliminary results.

Lemma 4.18. *For any $x \in \mathbb{R}^d$, the function $t \mapsto M_G^t(x)$ is differentiable on $(0, t_{\max})$ with derivative $\partial_t M_G^t(x) = -\frac{1}{2t^2}\|x - \operatorname{prox}_{tG}(x)\|^2$.*

The proof can be found in Subsection A.2.1.

Remark 4.19. In contrast to the standard result [15, Theorem 5, Section 3.3.2], we proved that M_G^t satisfies the Hamilton-Jacobi equation

$$\partial_t M_G^t(x) + \frac{1}{2}\|\nabla M_G^t(x)\|^2 = 0$$

without assuming Lipschitz continuity of G.

Proposition 4.20. *For any $x \in \mathbb{R}^d$, the function $t \to M_G^t(x)$ is Lipschitz continuous on $[0, t_{\max})$. More precisely,*

$$|M_G^t(x) - M_G^s(x)| \leq \frac{|s-t|}{2}\phi(\|x\|)^2 \tag{4.18}$$

for $0 \leq s, t < t_{\max}$ and $x \in \mathbb{R}^d$.

The proof can be found in Subsection A.2.2.

Corollary 4.21. *The map $t \mapsto Z_t := \int \exp\big(-U^t(y)\big)\,\mathrm{d}y$ is Lipschitz continuous on $[0, t_{\max})$ with Lipschitz constant $\frac{K_{\max}}{2}$ where $K_{\max}$ is defined in Assumption 4.3.4.*

Proof. Let $s, t \in [0, t_{\max})$ and assume without loss of generality that $s < t$ so that $M_G^s(x) \geq M_G^t(x)$ and, consequently, $\left|1 - \exp\left(-M_G^s(x) - M_G^t(x)\right)\right| = 1 - \exp\left(-M_G^s(x) - M_G^t(x)\right) \leq M_G^s(x) - M_G^t(x)$. Using Theorem 4.20 we can compute that

$$\begin{aligned}
|Z_s - Z_t| &\leq \int \left|\exp\left(-U^s(x)\right) - \exp\left(-U^t(x)\right)\right| \, \mathrm{d}x \\
&= \int \left|1 - \exp\left(-(M_G^t(x) - M_G^s(x))\right)\right| \exp\left(-U^s(x)\right) \mathrm{d}x \\
&\leq \int \left|M_G^t(x) - M_G^s(x)\right| \exp\left(-U^s(x)\right) \mathrm{d}x \\
&\leq \int \frac{|s-t|}{2} \phi(\|x\|)^2 \exp\left(-U^{t_{\max}}(x)\right) \mathrm{d}x \\
&= \frac{K_{\max}}{2} |s - t|,
\end{aligned} \tag{4.19}$$

where we used Assumption 4.3.4 for the last equality. ■

Proposition 4.22. *The curve $t \mapsto \pi^t$ is Lipschitz continuous on $[0, t_{\max})$ with respect to the TV norm with Lipschitz constant $\frac{K_{\max}}{2Z_0}\left(1 + \frac{Z_{t_{\max}}}{Z_0}\right)$. If, in addition,*

$$\int \|x\|^p \phi(\|x\|)^2 \exp\left(-U^{t_{\max}}(x)\right) \mathrm{d}x < \infty$$

holds for any $p \in [1, \infty)$, then $t \mapsto \pi^t$ is also Hölder continuous with Hölder exponent $\frac{1}{p}$ for the Wasserstein-p distance.

Proof. Since the TV distance can be computed as $\|\pi^s - \pi^t\|_{\mathrm{TV}} = \int |\pi^s(x) - \pi^t(x)| \, \mathrm{d}x$ and the Wasserstein-p distance for $p \in [1, \infty)$ satisfies $\mathcal{W}_p^p(\pi^s, \pi^t) \leq \int 2^{p-1} \|x\|^p |\pi^s(x) - \pi^t(x)| \, \mathrm{d}x$ (see [45, Theorem 6.15]) we can prove both assertions by bounding $\int g(x) |\pi^s(x) - \pi^t(x)| \, \mathrm{d}x$ and afterwards considering respective choices of g. Let $s, t \in [0, t_{\max})$ and assume as before $s < t$. It follows that

$$\begin{aligned}
&\int_{\mathbb{R}^d} g(x) \left|\pi^s(x) - \pi^t(x)\right| \, \mathrm{d}x \\
&\quad = \frac{1}{Z_s} \int_{\mathbb{R}^d} g(x) \left|\exp\left(-U^s(x)\right) - \exp\left(-U^t(x)\right)\right| \, \mathrm{d}x \\
&\qquad + \int_{\mathbb{R}^d} g(x) \left|\tfrac{1}{Z_s} - \tfrac{1}{Z_t}\right| \exp\left(-U^t(x)\right) \mathrm{d}x \\
&\quad \leq \frac{1}{Z_0} \int_{\mathbb{R}^d} g(x) \left|1 - \exp\left(-(M_G^t(x) - M_G^s(x))\right)\right| \exp\left(-U^{t_{\max}}(x)\right) \, \mathrm{d}x \\
&\qquad + \left|\tfrac{1}{Z_s} - \tfrac{1}{Z_t}\right| Z_{t_{\max}} \int_{\mathbb{R}^d} g(x) \pi^{t_{\max}}(x) \, \mathrm{d}x \\
&\quad \leq \tfrac{1}{2Z_0} |s - t| \int_{\mathbb{R}^d} g(x) \phi(\|x\|)^2 \exp\left(-U^{t_{\max}}(x)\right) \, \mathrm{d}x \\
&\qquad + \left|\tfrac{Z_t - Z_s}{Z_0^2}\right| Z_{t_{\max}} \mathbb{E}_{X \sim \pi^{t_{\max}}}[g(X)] \\
&\quad \leq \tfrac{K_{\max}}{2Z_0} \left(1 + \frac{Z_{t_{\max}}}{Z_0} \mathbb{E}_{X \sim \pi^{t_{\max}}}[g(X)]\right) |s - t|
\end{aligned} \tag{4.20}$$

where we made again use of Assumption 4.3.4. The choice $g \equiv 1$ proves Lipschitz continuity with respect to the TV norm with the stated Lipschitz constant. The choice $g(x) = 2^{p-1}\|x\|^p$ proves Hölder continuity with exponent $\frac{1}{p}$ with respect to the Wasserstein-p distance under the additional assumption. ■

Remark 4.23. The estimation of the Lipschitz constant in Theorem 4.22 can inform the choice of the sequence of Moreau parameters $t_N, t_{N-1}, \dots, t_1$ since it can be used to bound the TV distance between the consecutive distributions $\pi^{t_{n+1}}$ and π^{t_n}. While the Lipschitz constant is difficult to estimate without further assumptions on U, one may obtain upper bounds using, *e.g.*, Theorem A.1 and, if available, growth bounds on the subgradient of G. As an illustrative example, consider the Gaussian $G(x) = \frac{\|x\|^2}{2}$. Then, $\pi^t(x) \propto \exp\left(-\frac{\|x\|^2}{2(1+t)}\right)$ and $Z_0 = (2\pi)^{d/2}$ and $Z_t = (2\pi(1+t)^2)^{d/2}$. Moreover, one can easily check that $K_{\max} = Z_{t_{\max}}$. Thus, the Lipschitz constant of $t \mapsto \pi^t$ is

$$\frac{K_{\max}}{2Z_0}\left(1 + \frac{Z_{t_{\max}}}{Z_0}\right) = \frac{1}{2}(1+t_{\max})^d(1+(1+t_{\max})^d) = \frac{1}{2}(1+t_{\max})^d + (1+t_{\max})^{2d}.$$

In particular, the Lipschitz constant scales exponentially with respect to the dimension and polynomially with respect to $t_{\max}$. We want to point out, however, that the proposed scheme will sample correctly for any sequence of Moreau parameters $(t_n)_n$ as long as the final Moreau parameter is sufficiently small since the inner loop in Algorithm 1.1 is ergodic with stationary distribution close to the target (for small t and appropriate τ). The annealing scheme for DAZ is solely an acceleration so that the exponential scaling of the Moreau levels may be ignored without the risk of losing the correct convergence of the scheme.

As a numerical confirmation of Theorem 4.22 we show in Figure 2 the estimated TV distances $\|\pi^t_\tau - \pi\|_{\mathrm{TV}}$ as well as $\|\pi^t - \pi\|_{\mathrm{TV}}$ in the case of a Laplace distribution[1] $U(x) = G(x) = |x|$ for $x \in \mathbb{R}$ for different values of the Moreau parameter $t \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0\}$. The distance $\|\pi^t - \pi\|_{\mathrm{TV}}$ is computed by numerical integration over $[-10, 10]$. The distance $\|\pi^t_\tau - \pi\|_{\mathrm{TV}}$ is estimated by sampling 10 000 independent chains for 2000 iterations using ULA with the potential M^t_G and step size $\tau = \frac{t}{2}$ in accordance with (4.17) and afterwards computing the TV distance between π and the empirical sample distribution, again, using numerical integration. The theoretical Lipschitz continuity of $t \mapsto \pi^t$ with respect to the TV norm is also reflected in the practical experiments. Moreover, the errors $\|\pi^t_\tau - \pi\|_{\mathrm{TV}}$ increase as the step size τ increases.

4.3. Relation to diffusion models.

The practical differences between the proposed approach and annealed Langevin sampling—namely, that the proposed approach is applicable to any given potential (so long as it satisfies our assumptions) without training—were already discussed in section 2. In this section, we establish a theoretical relation between the proposed method and annealed Langevin sampling by providing convergence results of the potential used in annealed Langevin sampling to that used in DAZ through a zero-temperature limit. In the following we consider for simplicity the setting $U = G$ and $F \equiv 0$.

[1] While $U(x) = |x|$ is not superexponential, ergodicity and convergence of ULA applied to the potential M^t_G in this case follow from [11, Section 3.2].

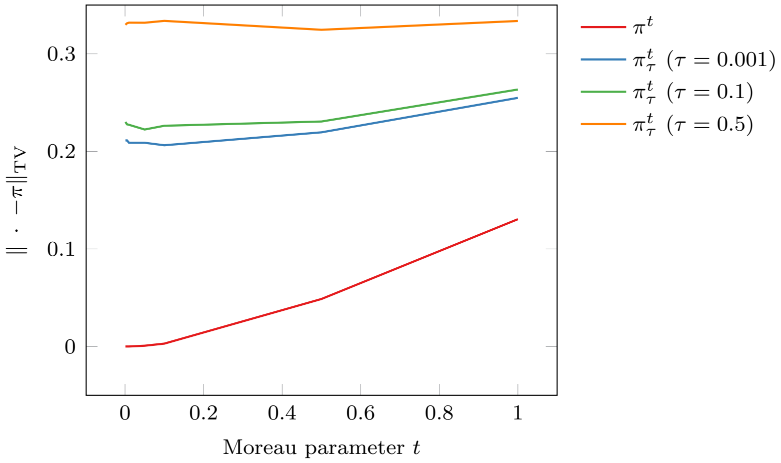

Figure 2. *TV distance between the target $\pi(x) \propto \exp(-G(x))$ with $G(x) = |x|$ and the Moreau envelope based potentials $\pi^t(x) \propto \exp(-M^t_G(x))$ as well as the corresponding stationary distributions of the ULA chain π^t_τ for step sizes $\tau \in \{0.001, 0.1, 0.5\}$. We clearly observe the proven Lipschitz continuity of $t \mapsto \pi^t$ with $\pi^t \to \pi$ as $t \to 0$.*

As elaborated in section 2, one variant of annealed Langevin sampling is to consider a family of distributions indexed by t and defined as $\pi * \mathcal{N}(0, \sqrt{t})$. The corresponding log-density reads as

$$\log(\pi * \mathcal{N}(0, \sqrt{t})) = \log\left(\frac{1}{(2\pi t)^{d/2}} \int_{\mathbb{R}^d} \exp\left(-G(y) - \frac{1}{2t}\|\cdot - y\|^2\right) \mathrm{d}y\right) + const. \tag{4.21}$$

where the (unknown) *const.* ensures normalization. A major inspiration for DAZ is the observation that the finite-sum version of this log-integral-exp expression is frequently encountered in machine learning and related fields and is also referred to as a *softmax*. One way to interpret the present work is that this softmax is replaced by a strict maximum, *i.e.*,

$$\log(\pi * \mathcal{N}(0, \sqrt{t})) \approx \frac{1}{(2\pi t)^{d/2}} \max_{y \in \mathbb{R}^d} -G(y) - \frac{1}{2t}\|\cdot - y\|^2 = -\frac{1}{(2\pi t)^{d/2}} M^t_G, \tag{4.22}$$

where we obtain precisely the Moreau envelope up to a multiplicative factor.

In this section, we formalize the relation between the softmax potential and the Moreau potential via the introduction of a (Boltzmann) temperature T. For the finite-sum case, it is well known[2] that for any $x \in \mathbb{R}^k$,

$$\max_{i \in \{1,\ldots,k\}} x_i = \lim_{T \to 0} T \log\left(\sum_{i=1}^k \exp\left(\frac{x_i}{T}\right)\right). \tag{4.23}$$

[2] This can be derived in various ways; a particularly cute one is to identify that the convex conjugate of $l(x) = \log(\sum_{i=1}^k \exp(x_i))$ is the Shannon entropy restricted to the simplex $\triangle^k$, that is $l^*(y) = \sum_{i=1}^k y_i \log(y_i) + \iota_{\triangle^k}(y)$, see [2, Example 3.25]. Since $l_T \coloneqq Tl(\,\cdot\,/T)$ is closed and convex, $l_T = l_T^{**}$, and standard scaling laws imply that $l_T^* = Tl^*$. Thus, $l_T = l_T^{**} = \sup_{y \in \triangle^k} \langle\,\cdot\,, y\rangle - Tl^*(y)$, which is the standard maximum when $T = 0$.

In analogy, we might expect that

$$-M_G^t(y) = \lim_{T\to 0} T \log\left(\int \exp\left(-\frac{1}{T}(G(x) + \frac{1}{2t}\|x-y\|^2)\right) \mathrm{d}x\right).$$

We will now establish this convergence in a rigorous manner, pointwise for the potential and its gradient as well as in total variation. To do so, motivated by the above, we define the following unnormalized diffusion-based potential with temperature $T > 0$

$$G_T^t(x) = -T \log\left(\frac{1}{(2\pi T t)^{d/2}} \int_{\mathbb{R}^d} \exp\left(-\frac{G(y)}{T} - \frac{1}{2Tt}\|x-y\|^2\right) \mathrm{d}y\right). \tag{4.24}$$

If the temperature T is set to 1, this is precisely the negative log-density of the distribution of variance exploding diffusion (4.21). By letting $T \to 0$, on the other hand, we recover the DAZ potential as will be shown below.

An alternative motivation for the consideration of the potential G_T^t arises through an analysis of the Hamilton-Jacobi equation: As shown above, the Moreau envelope satisfies

$$\begin{cases} \partial_t M_G^t(x) + \frac{1}{2}\|\nabla M_G^t(x)\|^2 = 0, & t > 0 \\ M_G^0 = G. \end{cases} \tag{4.25}$$

As an approximation, one might consider for small $T > 0$ the function G_T^t that solves

$$\begin{cases} \partial_t G_T^t(x) + \frac{1}{2}\|\nabla G_T^t(x)\|^2 = T\Delta G_T^t, & t > 0 \\ G_T^0 = G \end{cases} \tag{4.26}$$

whose solution converges to M_G^t uniformly as $T \to 0$ if G is bounded and Lipschitz[3] [7, Theorem 5.1]. The function G_T^t from (4.24) is, indeed, a solution to (4.26). Based on these observations, in [19, 34] the authors propose to estimate the Moreau envelope by computing (4.24) for small $T > 0$, which is done via Monte Carlo integration.

The following proposition formally states a result about the convergence of the two potentials.

Proposition 4.24. *The DAZ potential M_G^t is obtained as the zero-temperature limit of the diffusion potential G_T^t from (4.24). That is, $G_T^t(x) \to M_G^t(x)$, as $T \to 0^+$ for any $x \in \mathbb{R}^d$. In addition, if G is differentiable and ∇G is locally Lipschitz continuous, the convergence is uniform on compact sets.*

Proof. The result is an extension of the convergence $\|\cdot\|_p \to \|\cdot\|_\infty$ as $p \to \infty$. Using

[3] Neither of which is assumed in this manuscript.

Hölder's inequality, we find that

$$
\begin{aligned}
&\frac{1}{(2\pi Tt)^{d/2}} \int_{\mathbb{R}^d} \exp\left(\frac{-G(y) - \frac{1}{2t}\|x-y\|^2}{T}\right) \mathrm{d}y \\
&= \frac{1}{(2\pi Tt)^{d/2}} \int_{\mathbb{R}^d} \exp\left(-G(y) - \frac{1}{2t}\|x-y\|^2\right)^{\frac{1}{T}-1} \exp\left(-G(y) - \frac{1}{2t}\|x-y\|^2\right) \mathrm{d}y \\
&\le \exp\left(-M_G^t(x)\right)^{\frac{1}{T}-1} \frac{1}{(2\pi Tt)^{d/2}} \int_{\mathbb{R}^d} \exp\left(-G(y) - \frac{1}{2t}\|x-y\|^2\right) \mathrm{d}y \\
&\le \exp\left(-M_G^t(x)\right)^{\frac{1}{T}-1} \frac{1}{(2\pi Tt)^{d/2}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2t}\|x-y\|^2\right) \mathrm{d}y \\
&= \exp\left(-M_G^t(x)\right)^{\frac{1}{T}-1} \frac{1}{T^{d/2}},
\end{aligned}
\tag{4.27}
$$

where we used that $G \ge 0$. Taking the negative logarithm on both sides and multiplying by $T > 0$ yields

$$
G_T^t(x) \ge (1-T)M_G^t(x) + \tfrac{d}{2}T\log(T) \tag{4.28}
$$

Letting $T \to 0$ proves that $\liminf_{T\to 0} G_T^t(x) \ge M_G^t(x)$. For the opposite bound, let us consider $\Omega_\epsilon(x) := \left\{y \in \mathbb{R}^d \;\middle|\; G(y) + \frac{1}{2t}\|x-y\|^2 \le M_G^t(x) + \epsilon\right\}$ for some small $\epsilon > 0$. The continuity of G implies that this set has nonzero Lebesgue measure that we denote by $|\Omega_\epsilon(x)|$. Moreover,

$$
\frac{1}{(2\pi Tt)^{d/2}} \int_{\mathbb{R}^d} \exp\left(\frac{-G(y) - \frac{1}{2t}\|x-y\|^2}{T}\right) \mathrm{d}y \ge \frac{1}{(2\pi Tt)^{d/2}} |\Omega_\epsilon(x)| \exp\left(-\frac{M_G^t(x)+\epsilon}{T}\right)
\tag{4.29}
$$

which shows that $|\Omega_\epsilon(x)| < \infty$. Taking the negative logarithm and multiplying by $T > 0$ again, we find

$$
G_T^t(x) \le -T\log(|\Omega_\epsilon(x)|) + Td/2\log(2\pi Tt) + M_G^t(x) + \epsilon. \tag{4.30}
$$

Since ϵ was arbitrary, we conclude that $\limsup_{T\to 0} G_T^t(x) \le M_G^t(x)$ which shows the pointwise convergence.

To obtain uniform convergence on compact sets under the additional assumptions, we first note that if $x \in K$ with K compact, $G(x)$ and $M_G^t(x)$ are bounded and, consequently, T can be chosen uniformly for $x \in K$ in (4.28). For the upper bound we need to estimate $\log(|\Omega_\epsilon(x)|)$. By local Lipschitz continuity, there exists $L > 0$ such that ∇G is L-Lipschitz on the compact set K. The first-order approximation around $p \in \mathrm{prox}_{tG}(x)$ combined with the descent lemma shows that

$$
G(y) + \frac{1}{2t}\|y-x\|^2 \le M_G^t(x) + \left(\frac{L}{2} + \frac{1}{2t}\right)\|y-p\|^2. \tag{4.31}
$$

Consequently, $\left\{y \;\middle|\; (\frac{L}{2} + \frac{1}{2t})\|y-p\|^2 \le \epsilon\right\} \subset \Omega_\epsilon(x)$. Thus, we can lower bound $|\Omega_\epsilon(x)|$ uniformly for $x \in K$, which allows us also to choose T uniformly for $x \in K$ in (4.30) concluding the proof. ■

Proposition 4.25. *If ∇G is globally Lipschitz then, the distribution of annealed Langevin sampling converges to that of DAZ in TV. More precisely, we have for the probability densities $\pi_T^t(x) \propto \exp\left(-G_T^t(x)\right)$ and $\pi^t(x) \propto \exp\left(-M_G^t(x)\right)$ for any $t > 0$ that*

$$\lim_{T\to 0} \|\pi_T^t - \pi^t\|_{\mathrm{TV}} = 0.$$

Proof. If ∇G is globally Lipschitz, then by the same arguments as in Theorem 4.24 we find that $|\Omega_\epsilon| \geq c\epsilon^{d/2}$ for some dimension-dependent constant c.[4] It follows that

$$\begin{aligned}
&(1-T)M_G^t(x) + \tfrac{d}{2}T\log(T) \\
&\quad\leq G_T^t(x) \\
&\quad\leq -T\log(|\Omega_\epsilon(x)|) + Td/2\log(2\pi Tt) + M_G^t(x) + \epsilon \\
&\quad\leq -T\log(c\epsilon^{d/2}) + Td/2\log(2\pi Tt) + M_G^t(x) + \epsilon.
\end{aligned} \tag{4.32}$$

Thus, we can estimate

$$\begin{aligned}
|G_T^t(x) - M_G^t(x)| &\leq \max\left\{TM_G^t(x) + \tfrac{d}{2}T|\log(T)|, T|\log(c\epsilon^{d/2})| + Td/2|\log(2\pi Tt)| + \epsilon\right\} \\
&\leq TM_G^t(x) - cT(\log(T\epsilon) - 1) + \epsilon,
\end{aligned} \tag{4.33}$$

where $c > 0$ is an appropriate constant and assuming without loss of generality $T, \epsilon < 1$. We begin by showing the convergence

$$\int_{\mathbb{R}^d} |\exp\left(-G_T^t(x)\right) - \exp\left(-M_G^t(x)\right)|\mathrm{d}x \to 0 \tag{4.34}$$

as $T \to 0$ for which we use Lebesgue's dominated convergence theorem. The integrand converges to zero pointwise by Theorem 4.24. Thus, we are left to show that there exists a nonnegative integrable upper bound. We can estimate

$$\begin{aligned}
\left|\exp\left(-G_T^t(x)\right) - \exp\left(-M_G^t(x)\right)\right| &= |\exp\left(M_G^t(x) - G_T^t(x)\right) - 1|\exp\left(-M_G^t(x)\right) \\
&\leq \left(\exp\left(|M_G^t(x) - G_T^t(x)|\right) + 1\right)\exp\left(-M_G^t(x)\right).
\end{aligned} \tag{4.35}$$

Inserting (4.33) yields

$$\begin{aligned}
&\left|\exp\left(-G_T^t(x)\right) - \exp\left(-M_G^t(x)\right)\right| \\
&\quad\leq \left(\exp\left(TM_G^t(x) - cT(\log(T\epsilon) - 1) + \epsilon\right) + 1\right)\exp\left(-M_G^t(x)\right) \\
&\quad= \exp\left(-(1-T)M_G^t(x) - cT(\log(T\epsilon) - 1) + \epsilon\right) + \exp\left(-M_G^t(x)\right).
\end{aligned} \tag{4.36}$$

Since we are interested in the convergence for $T \to 0$, we can without loss of generality assume $T \in (0, 1/2]$ for which the above admits an integrable upper bound by Theorem A.1 together with the fact that $\int \exp\left(-M_G^t(x)\right)\mathrm{d}x < \infty$ as elaborated in Theorem 4.4, Item 3. Thus, we have proven (4.34). Let us now denote the respective partition functions as

$$Z(T) = \int_{\mathbb{R}^d} \exp\left(-G_T^t(x)\right)\mathrm{d}x, \quad Z = \int_{\mathbb{R}^d} \exp\left(-M_G^t(x)\right)\mathrm{d}x.$$

[4] Of course, c is explicit. The exact form is, however, not relevant for our purposes.

We find that

$$\begin{aligned}\|\pi_T^t - \pi^t\|_{\mathrm{TV}} &= \int_{\mathbb{R}^d} \left| \frac{1}{Z(T)} \exp\left(-G_T^t(x)\right) - \frac{1}{Z} \exp\left(-M_G^t(x)\right) \right| \mathrm{d}x \\ &= \frac{1}{Z} \int_{\mathbb{R}^d} \left|\exp\left(-G_T^t(x)\right) - \exp\left(-M_G^t(x)\right)\right| \mathrm{d}x \\ &\quad + \int_{\mathbb{R}^d} \left| \frac{1}{Z(T)} - \frac{1}{Z} \right| \exp\left(-G_T^t(x)\right) \mathrm{d}x \\ &= \frac{1}{Z} \int_{\mathbb{R}^d} \left|\exp\left(-G_T^t(x)\right) - \exp\left(-M_G^t(x)\right)\right| \mathrm{d}x + \frac{|Z(T) - Z|}{Z}\end{aligned} \tag{4.37}$$

which tends to zero as $T \to 0$ by (4.34) as shown in the first part of the proof. ■

Proposition 4.26. *The DAZ score is obtained as the zero-temperature limit of the diffusion score. That is, $\nabla G_T^t(x) \to \nabla M_G^t(x)$, as $T \to 0^+$ for any $x \in \mathbb{R}^d$. In addition, if G is differentiable and ∇G is locally Lipschitz continuous, the convergence is uniform on compact sets.*

Proof. Using Lebesgue's dominated convergence theorem to swap integration and differentiation we obtain that

$$\frac{\partial}{\partial x_i} \int_{\mathbb{R}^d} \exp\left(\frac{-G(y) - \frac{1}{2t}\|x-y\|^2}{T}\right) \mathrm{d}y = -\int_{\mathbb{R}^d} \frac{1}{Tt}(x_i - y_i) \exp\left(\frac{-G(y) - \frac{1}{2t}\|x-y\|^2}{T}\right) \mathrm{d}y.$$

By the chain rule it follows that

$$\begin{aligned}\nabla G_T^t(x) &= \int_{\mathbb{R}^d} \frac{1}{t}(x-y) \underbrace{\left(\frac{\exp\left(\frac{-G(y) - \frac{1}{2t}\|x-y\|^2}{T}\right)}{\int_{\mathbb{R}^d} \exp\left(\frac{-G(z) - \frac{1}{2t}\|x-z\|^2}{T}\right) dz} \right)}_{=:\rho_T(y|x)} \mathrm{d}y \\ &= \tfrac{1}{t}(x - \mathbb{E}_{Y \sim \rho_T(y|x)}[Y])\end{aligned} \tag{4.38}$$

Denoting $p = \mathrm{prox}_{tG}(x)$ it follows for the difference between the DAZ score and the diffusion score that for any $\nu > 0$

$$\begin{aligned}\|\nabla M_G^t(x) - \nabla G_T^t(x)\| &\leq \int_{\mathbb{R}^d} \tfrac{1}{t}\|y-p\|\rho_T(y|x)\,\mathrm{d}y \\ &= \int_{B_\nu(p)} \tfrac{1}{t}\|y-p\|\rho_T(y|x)\,\mathrm{d}y + \int_{B_\nu^c(p)} \tfrac{1}{t}\|y-p\|\rho_T(y|x)\,\mathrm{d}y \\ &\leq \frac{\nu}{t} + \int_{B_\nu^c(p)} \tfrac{1}{t}\|y-p\|\rho_T(y|x)\,\mathrm{d}y.\end{aligned} \tag{4.39}$$

To show that the above tends to zero we have to show that the density $\rho_T(y|x)$ approaches a Dirac measure concentrated at $\mathrm{prox}_{tG}(x)$ *quickly enough* as $T \to 0$. For a small enough $t > 0$, weak convexity of G implies that $y \mapsto G(y) + \frac{1}{2t}\|x-y\|^2$ is strongly convex and, consequently, that there exists a constant $C > 0$ such that $y \mapsto G(y) + \frac{1}{2t}\|x-y\|^2 - M_G^t(x) \geq C\|y-p\|^2$. As mentioned before, for any $\epsilon > 0$ the set $\Omega_\epsilon(x) := \left\{ y \in \mathbb{R}^d \;\middle|\; G(y) + \frac{1}{2t}\|x-y\|^2 \leq M_G^t(x) + \epsilon \right\}$ has positive Lebesgue measure and we can compute that

$$\begin{aligned} \frac{\int_{B_\nu^c(p)} \|y-p\| \exp\left(\frac{-G(y) - \frac{1}{2t}\|x-y\|^2}{T}\right) \mathrm{d}y}{\int_{\mathbb{R}^d} \exp\left(\frac{-G(y)-\frac{1}{2t}\|x-y\|}{T}\right) \mathrm{d}y} &\leq \frac{\int_{B_\nu^c(p)} \exp\left(\frac{-C\|y-p\|^2 - M_G^t(x)}{T}\right) \mathrm{d}y}{|\Omega_\epsilon(x)| \exp(-\frac{M_G^t(x)+\epsilon}{T})} \\ &= \frac{1}{|\Omega_\epsilon(x)|} \int_{B_\nu^c(p)} \|y-p\| \exp\left(\frac{-C\|y-p\|^2 + \epsilon}{T}\right) \mathrm{d}y \end{aligned} \tag{4.40}$$

Picking $\epsilon < C\nu^2$ and using Lebesgue's dominated convergence theorem we find that as $T \to 0$ the above integral tends to zero for any $\nu > 0$. Therefore by (4.39),

$$\limsup_{T\to 0} |\nabla M_G^t(x) - \nabla G_T^t(x)| \leq \frac{\nu}{t}.$$

Since $\nu > 0$ was arbitrary, this concludes the proof. Uniform convergence on compact sets in the case that ∇G is locally Lipschitz is obtained by the same arguments as above. ■

Within the above proof we have made use of the well-known Tweedie formula $\nabla G_T^t(x) = \frac{1}{t}(x - \mathbb{E}_{Y\sim\rho_T(y|x)}[Y])$, which relates the gradient of the potential with the MMSE for denoising. By viewing the proximal operator as a MAP denoiser, the result effectively states that the *MMSE-Tweedie* formula converges to the Moreau gradient formula $\nabla M_G^t(x) = \frac{1}{t}(x - \mathrm{prox}_{tG}(x))$, which constitutes a *MAP-version* of Tweedie.

5. Numerical Experiments.

In this section, we demonstrate the advantages of the proposed sampling algorithm over current algorithms on an extensive set of numerical experiments. We begin with a one-dimensional example that allows for the efficient computation of error metrics between the sample distribution and the reference distribution. Then, we consider the high-dimensional examples of TV prior sampling, TV-L2 denoising on a chain, TV-L2 denoising on images, sampling from the total deep variation (TDV) prior [23], and magnetic resonance imaging (MRI) reconstruction using the energy-based prior proposed in [47]. In all experiments, we compare several different sampling methods: ULA [38], MYULA [13], SK-ROCK [37], ULA with decreasing step sizes which we refer to as annealed Langevin dynamics (ALD)[5], the proposed sampling algorithm DAZ, and DAZ-SK-ROCK. In the latter, we use SK-ROCK instead of basic ULA for the inner loop in Algorithm 1.1. We use subgradient steps in the updates of ULA and ALD when the potential is nondifferentiable.

The Moreau parameters and step sizes will always be chosen as follows: The sequence of Moreau parameters $(t_n)_{N\geq n\geq 1}$ is determined by the specification of the endpoints t_N and t_1 and the loglinear computation of the intermediate values as

$$t_n = 10^{\frac{n-1}{N-1}\log_{10}\frac{t_N}{t_1} + \log_{10}(t_1)}, \quad n = 1, \ldots, N. \tag{5.1}$$

[5] This constitutes a small abuse of terminology since the potential remains unchanged as the step size changes, which is contrary to the use of this terminology in the early papers on diffusion models.

The endpoints are specified in the respective section of each experiment. The sequence of step sizes $(\tau_n)_n$ is derived from $(t_n)_{N\geq n\geq 1}$ such that it satisfies the step size requirements discussed in Theorem 4.17. For ULA, we use the final step size τ_1 across all iterations. For MYULA, we use the final Moreau parameter and step size, t_1 and τ_1, respectively, across all iterations. For SK-ROCK we fix the parameters $\eta = 0.05$ and $s = 5$ [37, Algorithm 3.1]. The Moreau parameter for SK-ROCK is chosen as the final (smallest) Moreau parameter t_1 with the step size for SK-ROCK as $0.9 \times \delta_s^{\max}$ with $\delta_s^{\max}$ according to [37, Algorithm 3.1]. For ALD, we use the same step sizes for each iteration as for DAZ. However, we want to emphasize that ALD has the same update rule as ULA directly applied to $\pi \propto \exp(-U(x))$ and the step size requirements for ergodicity of ULA are stricter than those of DAZ (*cf.* Theorem 4.17). Therefore, especially at the beginning of the iterations where the step sizes are largest it might be the case that ALD violates these step-size requirements. For DAZ-SK-ROCK we use the same Moreau parameter scheme as for DAZ and choose again always the largest feasible step size according to [37, Algorithm 3.1]. For methods involving SK-ROCK we count each update of the method as $s = 5$ iterations in order to provide a fair comparison.[6]

In all experiments, we denote the reference distribution with π and the sample distributions after k iterations of the various methods with π_k. More specifically, in each experiment we run many parallel Markov chains and at any iteration k, π_k is the sample distribution across all those parallel chains. For the experiments in Subsections 5.1, 5.3, and 5.4, we simulate 1000 parallel chains. For the TV prior experiment (Subsection 5.2), we simulate 100 000 parallel chains, and for the experiment that involves costly neural-network evaluations (Subsection 5.6), we simulate 500 parallel chains due to computational limitations. For the prior sampling experiment in Subsection 5.5 we only compute a single Markov chain since we do not compute any statistics. We define one iteration as one update step of Line 4 in Algorithm 1.1, *i.e.*, one evaluation of the gradient of the Moreau envelope. Hence, this does not account for any inner iterations when the proximal map of G is computed iteratively. When the experiments involve nonconvex potentials, which is the case in Subsections 5.5 and 5.6, the computations of the proximal maps might yield local minima.

To cover a range of design choices that the proposed sampling algorithm allows, we choose the number of iterations per Moreau level as $K = 1$ (which resembles diffusion models) in Subsection 5.2, $K = 20$ in Subsections 5.1, 5.3, and 5.4, and $K = 200$ (which resembles annealed Langevin sampling as in [42]) in Subsections 5.5 and 5.6.

5.1. One-dimensional Gaussian mixture. We consider a multi-modal (and, consequently, not log-concave) one-dimensional Gaussian mixture $\pi = \sum_{i=1}^{4} w_i \mathcal{N}(\mu_i, \sigma_i)$. The weights and parameters of the individual Gaussians are chosen as $w = (0.2, 0.2, 0.3, 0.3)$, $\mu = (-2, -1, 1, 2)$, and $\sigma = (0.05, 0.25, 0.25, 0.1)$. (These parameters were also used to produce Figure 1.) This potential satisfies Theorem 4.3 and is superexponential and, consequently, the presented theory applies. In Figure 3 we show the convergence of the sample distribution in the TV distance. We choose $N = 50$ Moreau parameters with endpoints $t_1 = 1 \times 10^{-4}$ and $t_N = 1 \times 10^{-2}$, each of which is used $K = 20$ times in the inner loop in Algorithm 1.1. The step size for DAZ is chosen as $\tau_n = \frac{t_n}{2}$ based on (4.17). To showcase the influence of the initialization, the

[6] The parameter s in SK-ROCK constitutes the order of the scheme so that one update within SK-ROCK is comparable to s (sub-)gradient or proximal updates.

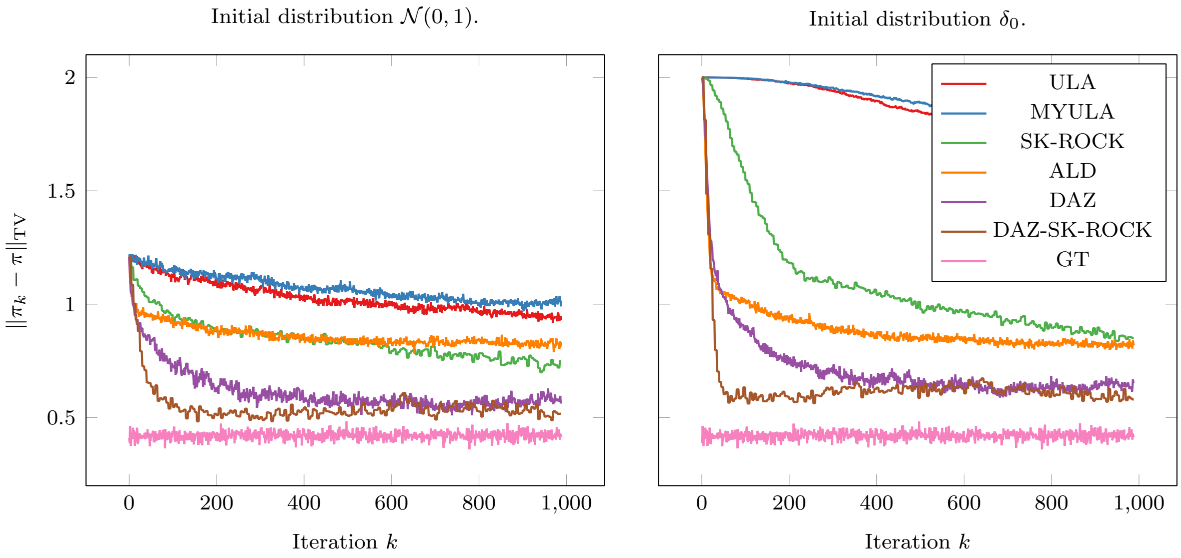

Figure 3. *TV distance between the sample distribution and the target Gaussian mixture. Left: Initializing the chains with a standard normal distribution. Right: Initializing with a Dirac distribution concentrated at zero. GT denotes the TV error of a random ground truth sample of the same size as the number of simulated chains. DAZ converges fastest with additional acceleration by combining it with SK-ROCK.*

experiment is done once with $\pi_0 = \mathcal{N}(0,1)$ and once with $\pi_0 = \delta_0$. The samples obtained from the Markov chains are then compared to the discretized target distribution in TV distance. To facilitate the interpretation of the results and to remove the influence of the finite sample and the discretization of the target distribution, we also plot the TV distance of a ground truth sample of the same finite size. (Note that the Gaussian mixture allows for easy direct sampling.)

The results in Figure 3 demonstrate that the DAZ sample converges to the reference sample significantly faster than the other methods, which is due to the the convexification of the Moreau envelope combined with the possibility of using larger steps. Indeed, for this experiment our method showed the greatest improvement over the others, which we hypothesize is due to the nonconvex potential. Combining DAZ with SK-ROCK leads to an additional speed increase at the beginning of the iterations. Regarding the slow convergence of ULA and MYULA we want to point out that for these methods which use only the smallest step size the trade-off between bias and convergence speed is especially crucial. While it might be possible to increase the convergence speed by using larger step sizes at the cost of an increased bias, we decided to stick to the chosen step-sizes as SK-ROCK—a comparison method that also uses only the smallest Moreau parameter—converges rather quickly. The remaining error of the reference sample (denoted as GT in Figure 3) of approximately 0.4 in TV to the reference distribution is a consequence of the finite number of chains and the computation of the TV distance rather than the sampling method.

5.2. TV prior sampling. In this experiment, we sample from a multi-dimensional ($d = 10$) distribution with potential $U = G\colon \mathbb{R}^d \to \mathbb{R}$ where $G(x) = \sum_{i=1}^{d-1} |x_{i+1} - x_i|$, which is the standard TV functional. The distribution of this potential is not integrable over $\mathbb{R}^d$, but it is integrable over the quotient space $\mathbb{R}^d/\{t \cdot 1 \in \mathbb{R}^d \mid t \in \mathbb{R}\}$. To account for this in our implementation, we project onto this linear subspace in each iteration by subtracting the mean of the iterate.[7]

We initialize all chains with samples from $\mathcal{N}(0, 0.1 \cdot I), I \in \mathbb{R}^{d\times d}$ and simulate them for $N = 1000$ Moreau levels with endpoints $t_1 = 2 \times 10^{-4}$ and $t_N = 1 \times 10^{-1}$, each of which is used $K = 1$ times. In accordance to (4.17) we set the step size to $\tau_n = \frac{t_n}{2}$. The proximal map of G is computed with the efficient method proposed in [24] that is based on dynamic programming.

The high dimensionality of this problem prohibits the computation of π and any distance to it. Moreover, since the density is only well-defined on a subspace of $\mathbb{R}^d$ also the computation of distances to marginals of the density is not straight forward (contrary to the subsequent TV experiments). However, by the transformation theorem for integrals, if $X \in \mathbb{R}^d/\{t \cdot 1 \in \mathbb{R}^d \mid t \in \mathbb{R}\}$ follows the distribution $\frac{1}{Z}\exp(-U(x))$, then the finite differences $X_{i+1} - X_i$ for $i = 1, \dots, d-1$ follow a Laplace distribution. Thus, we can compare the distribution of the finite differences of the samples of the various sampling algorithms at any iteration k, that we denote with $\Pi_{(i+1)\to i}(\pi_k)$[8] for $i = 1, \dots, d-1$ to the known reference distribution of these finite differences, that we denote with $\Pi(\pi) \propto \exp(-|\cdot|)$. (The reference distribution of the finite differences is independent of the index.) This evaluation gives us access to 9 marginal distributions that we could compare. For plotting we pick only three representative marginals, namely those with the smallest, median, and largest TV error at the end of the iterations[9]. The results are shown in Figure 4, where we find that DAZ and ALD converge significantly faster than ULA and MYULA. SK-ROCK gives mediocre performance while the combination of DAZ and SK-ROCK works very well. We hypothesize that ALD being on par with our method can be attributed to the convexity of U. The plots in Figure 5 visualize the negative-log histograms of the three representative marginal distribution obtained from the final steps of the simulated chains (*i.e.*, $-\log(\Pi_{(i+1)\to i}(\pi_{1000}))$ for the three representative i's) for the different algorithms. Again, we find that with DAZ, ALD, and DAZ-SK-ROCK the ground truth potential is approximated accurately.

5.3. TV-L2 denoising on a chain.

The next example we consider is TV denoising on a chain, where $F, G : \mathbb{R}^d \to \mathbb{R}$, $F(x) := \frac{1}{2\sigma^2}\|x - y\|^2$ and $G(x) = \lambda \sum_{i=1}^{d-1} |x_{i+1} - x_i|$ with $d = 100$,

[7] This experiment is not covered by the theory presented in the paper for two reasons. First, our theoretical results do not cover projections. However, we believe that this particular projection is unproblematic since the subgradient of TV and, consequently, its proximal map have zero mean when the input has zero mean. In addition, the Gaussian random vector added in each iteration has zero mean in expectation. Thus, the projection will be close to the identity in practice. Second, the potential is not superexponential. Both of these issues can be avoided by compensating for the non-trivial kernel of TV via considering a regularized version of the potential $U + \frac{\delta}{2}\|\cdot\|^2, 0 < \delta \ll 1$ as is done in the two subsequent experiments. We decided to stick to the projection instead in this experiment in order not to modify the potential and since the solution of adding a squared term is covered in the subsequent experiments anyway.

[8] We used `numpy.histogram` with options `bins="auto"` and `density=True` to generate histograms based on the samples (*cf.* https://numpy.org/doc/stable/reference/generated/numpy.histogram.html).

[9] As the empirical TV error oscillates significantly, determining the three representative marginals using only the TV error at the final iteration might lead to skewed results. Thus, we compute for each marginal the average TV error across the last 50 iterations. The obtained averages are used to find the three representative marginals.

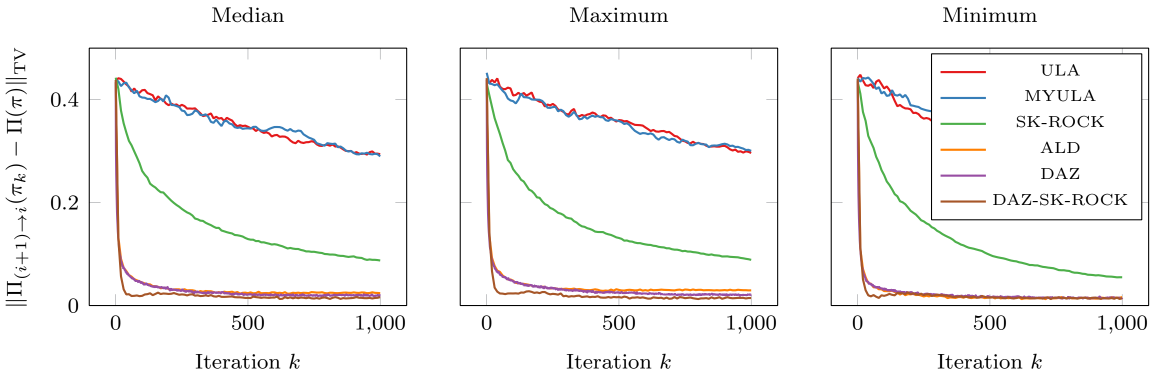

Figure 4. *TV prior sampling. TV distance between three finite difference marginals and the known ground truth. DAZ and ALD converge fastest, again, with additional acceleration by adding SK-ROCK in the inner loop of DAZ. The good performance of ALD might be due to convexity of the potential.*

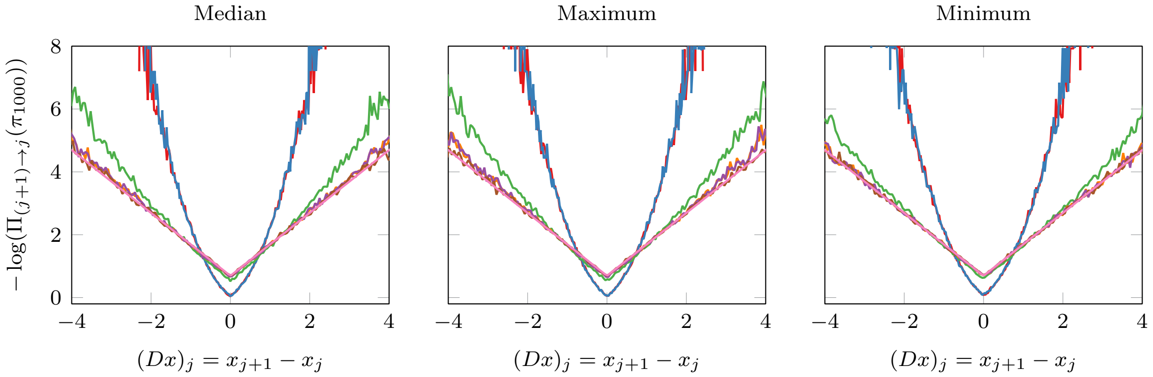

Figure 5. *TV difference marginals. The closer a method resembles the absolute value function, the better. We find that the samples obtained with DAZ, ALD, and DAZ-SK-ROCK approximate the target significantly more accurately.*

$\sigma = 0.1$, and $\lambda = 30$. To generate the data $y \in \mathbb{R}^d$ we first construct a piecewise constant vector $y^\dagger \in \mathbb{R}^d$ with values

$$y_i^\dagger = \begin{cases} -3 & \text{for } i = 1, \dots, 10 \\ -1 & \text{for } i = 11, \dots, 30 \\ 3 & \text{for } i = 31, \dots, 35 \\ 2 & \text{for } i = 36, \dots, 75 \\ 0 & \text{else.} \end{cases} \tag{5.2}$$

and compute $y = y^\dagger + \sigma z$ with $z \sim \mathcal{N}(0, I)$ and $I \in \mathbb{R}^{d\times d}$ the identity matrix. The potential satisfies Theorem 4.3 and is superexponential and, consequently, the theory presented in this paper applies. The endpoints of the Moreau parameters are $t_N = 1 \times 10^{-3}$ and $t_1 = 1 \times 10^{-4}$, each of which is used $K = 20$ times, and again $\tau_n = \frac{t_n}{2}$. We compute the proximal map of λG with the efficient method proposed in [24] which is based on dynamic programming.

Like in the previous section, the high dimensionality makes the computation of the reference distribution and any distances to it prohibitively expensive. However, estimates of the d reference marginal distributions—that we denote with $\Pi_1(\pi), \ldots, \Pi_d(\pi)$—can be computed efficiently using belief propagation (BP) algorithms [21, 32, 44]. In turn, this enables the verification of the convergence to the reference distribution through the computation of the TV distances between the (one-dimensional) marginals obtained from BP and those obtained from the various sampling algorithms. Similar to the previous section, we choose three representative marginals to visualize by estimating the averaging the TV distance over the last 20 iterations and then choose those marginals that correspond to the 5th, 50th, and 95th percentile. The results are shown in Figure 6. We find that SK-ROCK and DAZ-SK-ROCK perform best, closely followed by DAZ and afterwards ALD. As mentioned above, we believe that the advantages of DAZ are most prevalent in the nonconvex setting, so that in this experiment the acceleration by SK-ROCK is already at a similar level. The double dip in the convergence (especially visible for ULA and MYULA) might be a consequence of the fact that we measure the distance to π, whereas the chains target a biased version thereof.

In Figure 8 we plot for each method for one chain the value of the potential $(U(X_k))_k$. For convex potentials this value concentrates on a so-called *typical set* close to $\mathbb{E}[U(X)]$[10] [36, 37] so that we can use the convergence speed of the sequence $(U(X_k))_k$ as a proxy for the convergence speed of the Markov chain itself. We find in this experiment that DAZ and ALD converge fastest, closely followed by SK-ROCK. DAZ-SK-ROCK converges roughly at the same speed, however, including strong oscillations. ULA and MYULA converge significantly slower.

5.4. TV-L2 denoising for images.

For image denoising, $F, G : \mathbb{R}^{N\times M} \to \mathbb{R}$ with $F(x) := \frac{1}{2\sigma^2}\|x-y\|^2$ again and $G(x) = \lambda \sum_{i,j} |(Dx)_{i,j,1}| + |(Dx)_{i,j,2}|$ is the anisotropic TV, where $D : \mathbb{R}^{N\times M} \to \mathbb{R}^{N\times M\times 2}$ is a forward-finite-differences operator [5, Section 6.1]. We set $N = M = 200$, $\sigma = 0.05$, and $\lambda = 30$. The data y was again computed as $y = y^\dagger + \sigma z$ with $z \in \mathbb{R}^{N\times M}$, $(z_{i,j})_{i,j}$ i.i.d, $z_{i,j} \sim \mathcal{N}(0,1)$ and $y^\dagger$ a 200×200 crop of the ground truth watercastle image. For the computation of the proximal map of λG we use again [24]. The endpoints of the Moreau parameters are $t_N = 1 \times 10^{-3}$ and $t_1 = 1 \times 10^{-5}$, each of which is used $K = 20$ times, and again $\tau_n = \frac{t_n}{2}$.

As in Subsection 5.3, we obtain $M \times N$ reference marginal distributions using the BP algorithm and compute the one-dimensional TV distances to the sample distributions for each of them. Again, as in Subsection 5.3, we plot in Figure 7 the convergence of three representative marginals, and in Figure 8 on the right the convergence to the typical set. We observe best convergence from DAZ, ALD, and SK-ROCK. DAZ converges fastest on the medium marginal, on par with ALD on the slowest marginal, and SK-ROCK being fastest on the fastest converging marginal. In this experiment the combination DAZ-SK-ROCK induces strong oscillations which, however, do not disrupt the overall convergence. We speculate that the combination of SK-ROCK with annealing, *i.e.*, successive alteration of the target distribution, exhibits slightly instable behavior. The convergence to the typical set is in this experiment fastest for DAZ, ALD, and SK-ROCK[11]. The fact that for denoising on a chain the value of $U(x)$ increases during iterations, for images, however, it decreases is caused since

[10] up to the discrepancy induced by the difference between U and U^t

[11] Note that we focus here on the time until stationarity is reached, not reached the value of $U(x)$.

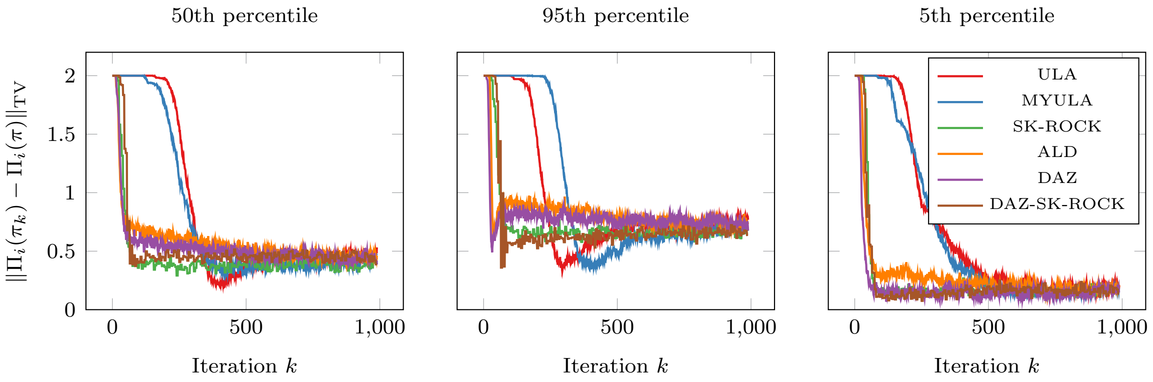

Figure 6. *TV denoising on a chain. TV distance between three different marginals of sample distribution and the target for TV-denoising. Fastest convergence by DAZ, SK-ROCK, and DAZ-SK-ROCK. The double dip behavior might be a consequence of the fact that we compare to the target distribution and not to the stationary measure of each chain.*

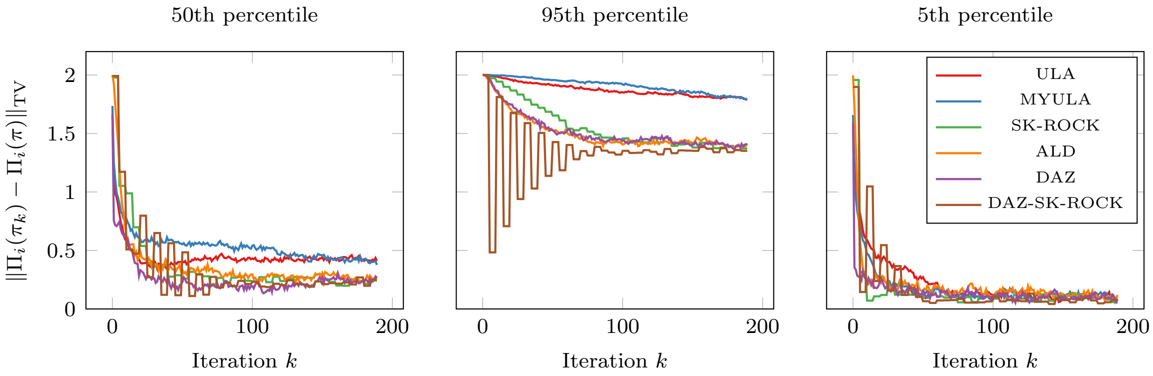

Figure 7. *TV denoising on an image. TV distance between three different marginals of sample distribution and the target for TV-denoising. Fastest convergence for DAZ, ALD, and SK-ROCK. DAZ-SK-ROCK exhibits strong oscillations which, however, do not bother the overall convergence. The piecewise constant shape of SK-ROCK is due to the fact that for SK-ROCK we count one iteration per prox evaluation which is plotted by piecewise constant upsampling.*

for chains we initialized at zero and for images at the given noisy observation.

5.5. Sampling from a deep prior. To showcase the applicability of the proposed sampling algorithm to nonconvex potentials in high-dimensional settings, we consider the TDV potential introduced in [23]. This potential is twice continuously differentiable and the second derivative is bounded and, consequently, the potential is Lipschitz continuous and weakly convex[12]. However, a naive usage of the potential does not meet the integrability requirements and instead we consider the potential $U(x) = \lambda \,\mathrm{TDV}(x) + \frac{1}{2\gamma_1}\big(\frac{1}{d}\sum_{i=1}^d x_i - \mu\big)^2 + \frac{1}{2\gamma_2}(\mathrm{var}[x] - \sigma^2)^2$ which quadratically penalizes the statistical deviations of a given patch to ensure integrability. In addition to the potential now being integrable we can also enforce the mean and variance of the samples staying close to μ and σ^2 which are chosen as the empirical mean and variance of

[12] Unfortunately, convexity outside a ball is difficult to verify for TDV and similar deep models.

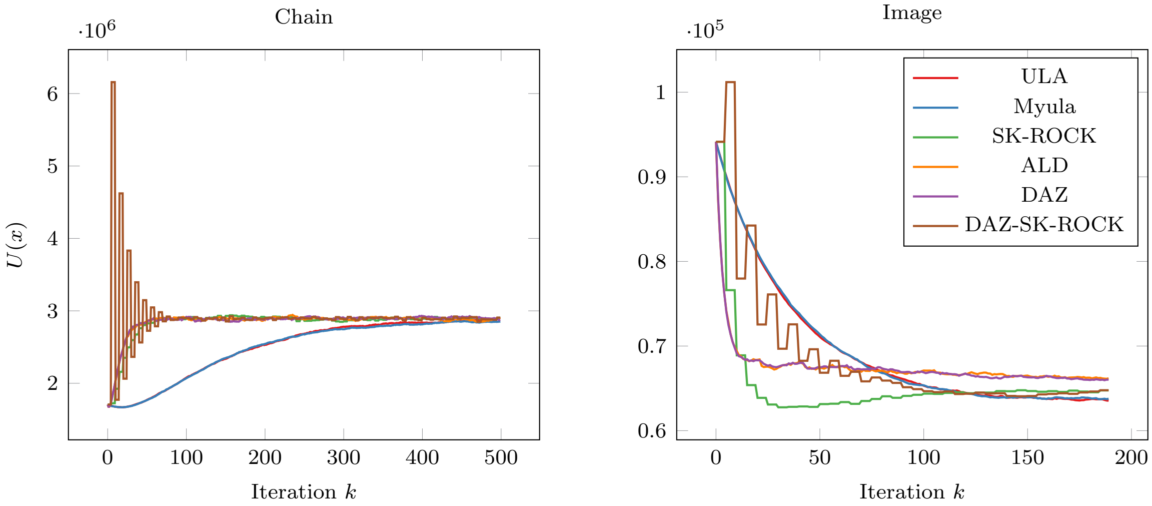

Figure 8. *TV denoising convergence to the typical set. Left: denoising on a chain, right: denoising on images. Fastest convergence for DAZ, ALD, and SK-ROCK. DAZ-SK-ROCK exhibits strong oscillations which, however, do not bother the overall convergence. The piecewise constant shape of SK-ROCK is due to the fact that for SK-ROCK we count one iteration per prox evaluation which is plotted by piecewise constant upsampling.*

the training dataset which were computed over 10 000 patches of size (96×96) extracted from the BSDS [29] dataset. For our experiment, we utilize the publicly available TDV weights[13] that were obtained from optimizing the denoising performance on perturbed BSDS500 [29] grayscale image patches of size 96×96.

Throughout this experiment, we use $\lambda = 8$, $\gamma_1 = \gamma_2 = 1 \times 10^{-8}$, and work on $d = 256 \times 256$-dimensional image patches. We simulate $N = 5$ Moreau parameters whose endpoints are are chosen as $t_1 = 1 \times 10^{-4}$ to $t_N = 5t_1$ and each of which is used for $K = 200$ Langevin steps, which results in a total of 1000 iterations. The Langevin step size in each level is set to to $\tau_n = t_n/2$ in accordance with Theorem 4.17. The proximal operator of G is computed with accelerated proximal gradient descent (APGD), which is given in Algorithm B.1, where we perform gradient steps on the potential and proximal steps on the proximity term. The proximity term could also be handled through gradient steps, but we observed slightly faster convergence when splitting the terms. All trajectories are initialized at the same $x_0 = \mu + \sigma z$, $z \sim \mathcal{N}(0, I), I \in \mathbb{R}^{d \times d}$.

The high dimensionality and the nonconvexity of the potential prohibits a quantitative analysis of the convergence. Consequently, we only provide a qualitative comparison of the methods in Figure 9. The results show that the chains of ALD and DAZ reach a reasonable sample after around 200 iterations, whereas MYULA and ULA require roughly 600 to 800 iterations to reach a result that is of similar quality.

In addition, we point out that we sometimes observed instabilities for ALD when the step sizes were too large. This underpins the relevance of the proposed sampling algorithm from a stability and convergence speed perspective and supports the findings regarding possible step sizes in Theorem 4.17.

[13] See: https://github.com/VLOGroup/tdv

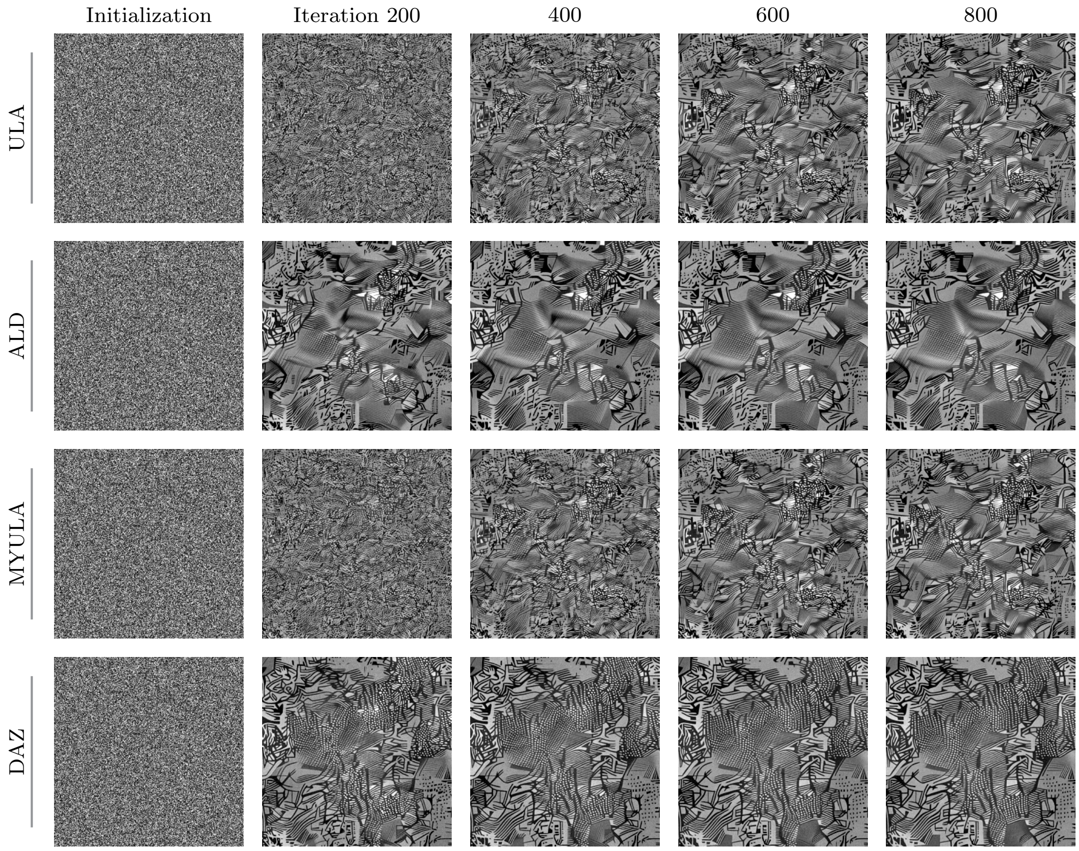

Figure 9. *Exemplary trajectories obtained by various sampling algorithms applied to the TDV potential. From top to bottom: ULA; ALD; MYULA; DAZ;. From left to right: Iterations 0; 200; 400; 600; 800. ULA and MYULA take significantly longer to reach a characteristic sample visually.*

5.6. Accelerated MRI reconstruction.

In this section we showcase the applicability of the proposed sampling algorithm to high-dimensional inverse problems with a nontrivial forward operator inspired by accelerated MRI. We consider the problem of recovering a real-valued image from undersampled and noisy Fourier data and thus $F, G : \mathbb{R}^{M\times N} \to \mathbb{R}$ with $F(x) \coloneqq \frac{\lambda}{2}\|M\mathcal{F}x - y\|_2^2$, where $\mathcal{F} : \mathbb{R}^{M\times N} \to \mathbb{C}^{M\times(\lfloor N/2\rfloor+1)}$ is the two-dimensional Fourier transform that accounts for the conjugate symmetry of the spectrum of a real signal and $M : \mathbb{C}^{M\times(\lfloor N/2\rfloor+1)} \to \mathbb{C}^n$ is a frequency selection operator that models the undersampling that is used to accelerate imaging speed. The function $G = \mathrm{EBM} + \frac{1}{2\gamma}\|\cdot\|^2$ incorporates the learned energy-based model EBM from [47]. This model is a cascade of convolution operators with stride two and point-wise leaky-rectified-linear activation functions, followed by a linear layer that maps to a scalar and finally the absolute value to ensure boundedness from below. To ensure weak convexity of G,[14] we replaced the leaky-rectified-linear activation functions $x \mapsto \max(\alpha x, x)$ with $\alpha = 0.05$ with the surrogate $x \mapsto \frac{1}{\beta}\log\bigl(\exp(\alpha\beta x) + \exp(\beta x)\bigr)$ with

[14] Again, convexity outside a ball is difficult to verify for deep models.

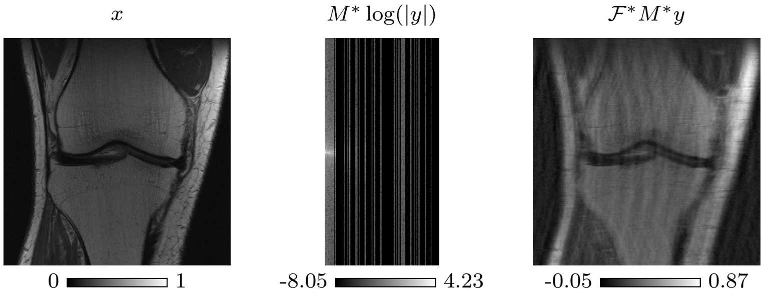

Figure 10. *From left to right: reference reconstruction; visualization of the data; least-squares reconstruction.*

$\beta = 1000$. This surrogate is infinitely often differentiable with bounded second derivative and, consequently, the potential is Lipschitz continuous and weakly convex. In addition, the small quadratic norm with $\gamma = 1 \times 10^9$ ensures the integrability of the Gibbs distribution. The data $y \in \mathbb{C}^n$ are constructed by $y = M\mathcal{F}x^* + \sigma z$ where $x^* \in \mathbb{R}^{320\times 320}$ is the reference root-sum-of-squares reconstruction of the 17th slice in the file `file1000005.h5` in the `multicoil_train` folder of the fastMRI knee dataset [22] and $z \sim \mathcal{N}(0, I)$. Like in the simulation studies in the original publication [47], the intensity values of the reference root-sum-of-squares reconstruction were affinely mapped such that their minimum is zero and their maximum is one, such that the intensity values consistent with those in the training. The standard deviation of the noise was chosen as $\sigma = 2 \times 10^{-2}$ and we set $\lambda = 1 \times 10^5$ based on manual search optimized on visually appealing reconstructions.

For the sampling algorithms, $\tau_1 = 7.5 \times 10^{-3}$ is determined by the step size used in the ULA algorithm used in training the model. We chose $N = 5$, $K = 200$ and the endpoints or the sequence of Moreau parameters as $t_1 = \frac{\tau_1}{\tau_1 - \tau\lambda}$ (to comply with step size restrictions) and $t_N = 5t_1$. As in the previous section, we computed the proximal map of G using APGD where we took gradient steps on G and proximal steps on the proximity term. We ran 500 parallel chains all initialized at the naive reconstruction $\mathcal{F}^* M^* y$ that is shown in Figure 10 along with the reference reconstruction and a visualization of the data.

Like in the previous section, the high dimensionality and nonconvexity makes this an extremely challenging problem. Due to the absence of any ground-truth samples we describe the results of the four sampling algorithms shown in Figure 11 again only qualitatively. The marginal standard deviations obtained by the gradient-based samplers ULA and ALD are significantly more blurry than those obtained by the proximal-based samplers MYULA and DAZ. In addition, ALD and DAZ converge significantly faster than MYULA and ULA. This is examplified by the backfolding artifact indicated by the gray arrow in Figure 11, that is clearly visible in the MMSE estimate even after 1000 iterations of MYULA and ULA. In contrast, it is almost fully removed after 600 iterations of DAZ.

6. Conclusion.

In this article, we proposed a method for the efficient sampling from Gibbs distributions π whose potential may be nondifferentiable and nonconvex. Inspired by diffusion models, we consider a sequence of distributions π^t that has favorable properties for sampling when t is large and approaches the target π as $t \to 0$. The sequence is obtained by replacing the parts of the potential with its Moreau envelope. Within our approach we then

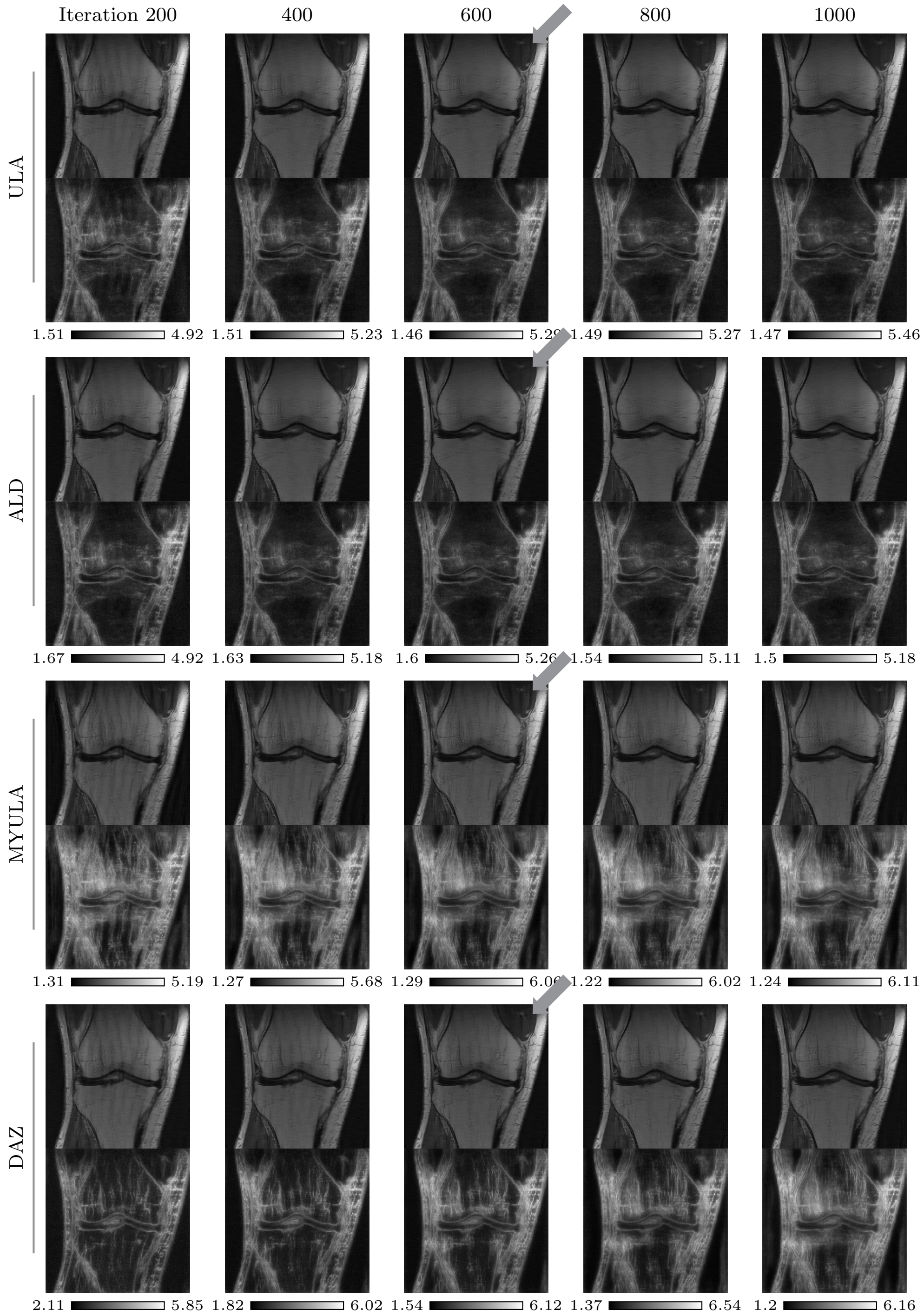

Figure 11. *MMSE and standard deviation estimates obtained by various sampling algorithms. From top to bottom: ULA; ALD; MYULA; DAZ. For each algorithm, the MMSE estimate is shown on top of the variance estimate. From left to right: Iterations 200; 400; 600; 800; 1000. The colormap values are multiplied by* 100 *and apply only to the standard deviation. The gray arrow indicates a prominent backfolding artifact. ULA and ALD lead to blurrier results. ALD and DAZ remove the artifact significantly faster.*

successively sample from π^t for decreasing values of t by constructing a Markov chain from the corresponding Langevin diffusion. We proved ergodicity of the chain for sampling from π^t for fixed t and convergence to the target density in TV when t is sufficiently small. Moreover, we showed that the map $t \mapsto \pi^t$ is Lipschitz continuous in the TV norm, which justifies the approach of successively sampling from π^t for decreasing t. In addition, we proved that all distributions π^t can be understood as a zero-temperature limit of the Gibbs distributions corresponding to a variance exploding diffusion model for π.

An extensive set of numerical experiments that contains one-dimensional toy problems and high-dimensional Bayesian inverse problems in imaging confirmed the efficacy of the method compared to ULA, MYULA, ALD, and SK-ROCK. In addition to the broad applicability of DAZ to potentials of rather general form, the conducted experiments confirm that the proposed method yields a significant speedup particularly for nonconvex or not strongly convex potentials (subsections 5.1 and 5.2). In the strongly convex case, DAZ yields faster convergence than ULA or MYULA due to the relaxed step size conditions and provides performance comparable to ALD and SK-ROCK.

Limitations and future work. Within the sampling procedure, we successively sample from the distributions π^t by initializing each Markov chain with the last iterate of the previous chain. Contrary, *e.g.*, in diffusion models, it is possible to discretize the backward SDE in order to obtain a theoretically well-founded time step from $\pi^{t+\Delta t} \to \pi^t$ where Δt denotes the size of the time step. Future work will investigate if, *e.g.*, the sequence $(\pi^t)_t$ admits a governing transport equation that would allow a similar discretization of the time evolution $t \mapsto \pi^t$.

An additional direction for future work is to extend the proposed scheme to Bregman-Moreau envelopes similar to [26] and the theoretical investigation of advanced sampling schemes like SK-ROCK as inner algorithms for DAZ.

REFERENCES

[1] D. Bakry, I. Gentil, and M. Ledoux, *Markov Semigroups*, Springer International Publishing, Cham, 2014, pp. 3–75, https://doi.org/10.1007/978-3-319-00227-9_1, https://doi.org/10.1007/978-3-319-00227-9_1.

[2] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[3] K. Bredies, K. Kunisch, and T. Pock, *Total generalized variation*, SIAM Journal on Imaging Sciences, 3 (2010), pp. 492–526, https://doi.org/10.1137/090769521.

[4] M. Burger, M. J. Ehrhardt, L. Kuger, and L. Weigand, *Coupling analysis of the asymptotic behaviour of a primal-dual Langevin algorithm*, arXiv preprint arXiv:2405.18098, (2024), https://doi.org/10.48550/arXiv.2405.18098.

[5] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of Mathematical Imaging and Vision, 40 (2011), pp. 120–145, https://doi.org/10.1007/s10851-010-0251-1.

[6] 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), https://doi.org/10.48550/arXiv.1805.01648.

[7] M. Crandal and P. Lions, *Two approximations of solutions of hamilton–jacobi equations*, Math. Comp, 43 (1984), pp. 1–19.

[8] 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 (2016), pp. 651–676, https://doi.org/10.1111/rssb.12183.

[9] Y. Du, S. Li, J. Tenenbaum, and I. Mordatch, *Improved contrastive divergence training of energy-based models*, in Proceedings of the 38th International Conference on Machine Learning, M. Meila and T. Zhang, eds., vol. 139 of Proceedings of Machine Learning Research, PMLR, 18–24 Jul 2021, pp. 2837–2848, https://proceedings.mlr.press/v139/du21b.html.

[10] A. Durmus, S. Majewski, and B. Miasojedow, *Analysis of Langevin Monte Carlo via convex optimization*, The Journal of Machine Learning Research, 20 (2019), pp. 2666–2711.

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

[12] A. Durmus and É. Moulines, *High-dimensional Bayesian inference via the unadjusted Langevin algorithm*, Bernoulli, 25 (2019), pp. 2854–2882.

[13] A. Durmus, E. Moulines, and M. Pereyra, *Efficient Bayesian computation by proximal Markov chain Monte Carlo: When Langevin meets Moreau*, SIAM Journal on Imaging Sciences, 11 (2018), pp. 473–506, https://doi.org/10.1137/16M1108340.

[14] M. J. Ehrhardt, L. Kuger, and C.-B. Schönlieb, *Proximal Langevin sampling with inexact proximal mapping*, SIAM Journal on Imaging Sciences, 17 (2024), pp. 1729–1760, https://doi.org/10.1137/23M1593565.

[15] L. C. Evans, *Partial differential equations*, vol. 19, American Mathematical Society, 2022.

[16] L. Fruehwirth and A. Habring, *Ergodicity of Langevin dynamics and its discretizations for non-smooth potentials*, arXiv preprint arXiv:2411.12051, (2024).

[17] A. Habring, A. Falk, and T. Pock, *Diffusion at absolute zero: Langevin sampling using successive moreau envelopes*, arXiv preprint arXiv:2502.01358, (2025).

[18] A. Habring, M. Holler, and T. Pock, *Subgradient Langevin methods for sampling from nonsmooth potentials*, SIAM Journal on Mathematics of Data Science, 6 (2024), pp. 897–925, https://doi.org/10.1137/23M1591451.

[19] H. Heaton, S. Wu Fung, and S. Osher, *Global solutions to nonconvex problems by evolution of hamilton-jacobi pdes*, Communications on Applied Mathematics and Computation, 6 (2024), pp. 790–810, https://doi.org/10.1007/s42967-022-00239-5, https://doi.org/10.1007/s42967-022-00239-5.

[20] R. Jordan, D. Kinderlehrer, and F. Otto, *The variational formulation of the Fokker-Planck equation*, SIAM Journal on Mathematical Analysis, 29 (1998), pp. 1–17, https://doi.org/10.1137/S0036141096303359.

[21] P. Knobelreiter, C. Sormann, A. Shekhovtsov, F. Fraundorfer, and T. Pock, *Belief propagation reloaded: Learning bp-layers for labeling problems*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7900–7909.

[22] F. Knoll, J. Zbontar, A. Sriram, M. J. Muckley, M. Bruno, A. Defazio, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana, Z. Zhang, M. Drozdzalv, A. Romero, M. Rabbat, P. Vincent, J. Pinkerton, D. Wang, N. Yakubova, E. Owens, C. L. Zitnick, M. P. Recht, D. K. Sodickson, and Y. W. Lui, *fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning*, Radiology: Artificial Intelligence, 2 (2020), p. e190007, https://doi.org/10.1148/ryai.2020190007.

[23] E. Kobler, A. Effland, K. Kunisch, and T. Pock, *Total deep variation for linear inverse problems*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[24] V. Kolmogorov, T. Pock, and M. Rolinek, *Total variation on a tree*, SIAM Journal on Imaging Sciences, 9 (2016), pp. 605–636, https://doi.org/10.1137/15M1010257.

[25] D. Lamberton and G. Pages, *Recursive computation of the invariant distribution of a diffusion: The case of a weakly mean reverting drift*, Stochastics and dynamics, 3 (2003), pp. 435–451, https://doi.org/10.1142/S0219493703000838.

[26] T. T.-K. Lau and H. Liu, *Bregman proximal langevin monte carlo via bregman-moreau envelopes*, in International Conference on Machine Learning, PMLR, 2022, pp. 12049–12077.

[27] G. Luo, M. Blumenthal, M. Heide, and M. Uecker, *Bayesian MRI reconstruction with joint uncertainty estimation using diffusion models*, Magnetic Resonance in Medicine, 90 (2023), pp. 295–311, https://doi.org/https://doi.org/10.1002/mrm.29624.

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

//doi.org/10.1007/s11009-020-09809-7.

[29] D. Martin, C. Fowlkes, D. Tal, and J. Malik, *A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics*, in Proceedings eighth IEEE international conference on computer vision. ICCV 2001, vol. 2, IEEE, 2001, pp. 416–423.

[30] S. P. Meyn and R. L. Tweedie, *Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes*, Advances in Applied Probability, 25 (1993), pp. 518–548, https://doi.org/10.2307/1427522.

[31] M. C. Mukkamala, P. Ochs, T. Pock, and S. Sabach, *Convex-concave backtracking for inertial Bregman proximal gradient algorithms in nonconvex optimization*, SIAM Journal on Mathematics of Data Science, 2 (2020), pp. 658–682, https://doi.org/10.1137/19M1298007.

[32] D. Narnhofer, A. Habring, M. Holler, and T. Pock, *Posterior-variance–based error quantification for inverse problems in imaging*, SIAM Journal on Imaging Sciences, 17 (2024), pp. 301–333, https://doi.org/10.1137/23M1546129.

[33] E. Nijkamp, M. Hill, T. Han, S.-C. Zhu, and Y. N. Wu, *On the anatomy of MCMC-based maximum likelihood learning of energy-based models*, Proceedings of the AAAI Conference on Artificial Intelligence, 34 (2020), pp. 5272–5280, https://doi.org/10.1609/aaai.v34i04.5973.

[34] S. Osher, H. Heaton, and S. W. Fung, *A hamilton–jacobi-based proximal operator*, Proceedings of the National Academy of Sciences, 120 (2023), p. e2220469120, https://doi.org/10.1073/pnas.2220469120, https://www.pnas.org/doi/abs/10.1073/pnas.2220469120, https://arxiv.org/abs/https://www.pnas.org/doi/pdf/10.1073/pnas.2220469120.

[35] M. Pereyra, *Proximal Markov chain Monte Carlo algorithms*, Statistics and Computing, 26 (2016), pp. 745–760, https://doi.org/10.1007/s11222-015-9567-4.

[36] M. Pereyra, *Maximum-a-posteriori estimation with bayesian confidence regions*, SIAM Journal on Imaging Sciences, 10 (2017), pp. 285–302, https://doi.org/10.1137/16M1071249, https://doi.org/10.1137/16M1071249, https://arxiv.org/abs/https://doi.org/10.1137/16M1071249.

[37] M. Pereyra, L. V. Mieles, and K. C. Zygalakis, *Accelerating proximal markov chain monte carlo by using an explicit stabilized method*, SIAM Journal on Imaging Sciences, 13 (2020), pp. 905–935, https://doi.org/10.1137/19M1283719, https://doi.org/10.1137/19M1283719, https://arxiv.org/abs/https://doi.org/10.1137/19M1283719.

[38] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363.

[39] T. Rockafellar and R. J. B. Wets, *Variational analysis*, Springer Berlin, Heidelberg, 2009, https://doi.org/https://doi.org/10.1007/978-3-642-02431-3.

[40] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena, 60 (1992), pp. 259–268, https://doi.org/10.1016/0167-2789(92)90242-F.

[41] L. M. Skvortsov, *Explicit stabilized runge-kutta methods*, Computational Mathematics and Mathematical Physics, 51 (2011), pp. 1153–1166, https://doi.org/10.1134/S0965542511070165, https://doi.org/10.1134/S0965542511070165.

[42] Y. Song and S. Ermon, *Generative modeling by estimating gradients of the data distribution*, in Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., vol. 32, Curran Associates, Inc., 2019.

[43] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, *Score-based generative modeling through stochastic differential equations*, in International Conference on Learning Representations, 2021.

[44] M. F. Tappen and F. W. T., *Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters*, in Proceedings Ninth IEEE International Conference on Computer Vision, 2003, pp. 900–906 vol.2.

[45] C. Villani, *Optimal transport*, Springer, 2009, https://doi.org/10.1007/978-3-540-71050-9.

[46] A. Wibisono, *Proximal langevin algorithm: Rapid convergence under isoperimetry*, arXiv preprint arXiv:1911.01469, (2019).

[47] M. Zach, F. Knoll, and T. Pock, *Stable deep MRI reconstruction using generative priors*, IEEE Transactions on Medical Imaging, 42 (2023), pp. 3817–3832, https://doi.org/10.1109/TMI.2023.3311345.

[48] M. Zach, E. Kobler, and T. Pock, *Computed tomography reconstruction using generative energy-based*

priors, in Proceedings of the OAGM Workshop 2021, M. Seidl, M. Zeppelzauer, and P. Roth, eds., Verlag der Technischen Universität Graz, Dec. 2021, pp. 52–58, https://doi.org/10.3217/978-3-85125-869-1-09.

Appendix A. Proofs.

A.1. Ergodicity of DAZ.

A.1.1. Proof of Theorem 4.6.

Proof. In the following, we will frequently denote an element $p \in \operatorname{prox}_{\tau G}(x)$ as $p = p_\tau(x)$ for simplicity. If it is clear from context, we will also omit either the argument x or the parameter τ and denote the prox as p_τ, or simply p. We begin with the first assertion by letting $(x_n, t_n) \to (x, t)$ in $\mathbb{R}^d \times (0, \frac{1}{\rho_G})$. By definition of the proximal map, we have that

$$\frac{1}{2t_n}\|x_n - p_{t_n}(x_n)\|^2 + G(p_{t_n}(x_n)) \leq \frac{1}{2t_n}\|x_n\|^2 + G(0).$$

The right-hand side is bounded by convergence of $x_n \to x$, $t_n \to t$ and, thus, by boundedness from below of G, it follows that $(p_{t_n}(x_n))_n$ is bounded. Therefore, there exists a convergent subsequence $p_{t_n}(x_n) \to \hat{p}$, which we shall not relabel for the sake of a simpler notation. It follows for arbitrary $y \in \mathbb{R}^d$ that

$$\begin{aligned} \frac{1}{2t}\|x - \hat{p}\|^2 + G(\hat{p}) &\underset{(i)}{=} \lim_{n\to\infty} \frac{1}{2t_n}\|x_n - p_{t_n}(x_n)\|^2 + G(p_{t_n}(x_n)) \\ &\underset{(ii)}{\leq} \lim_{n\to\infty} \frac{1}{2t_n}\|x_n - y\|^2 + G(y) \\ &= \frac{1}{2t}\|x - y\|^2 + G(y) \end{aligned} \tag{A.1}$$

where (i) is a consequence of continuity and (ii) follows from optimality of $p_{t_n}(x_n)$. Since y was arbitrary we can conclude $\hat{p} \in \operatorname{prox}_{tG}(x)$, which is single-valued for $t < \frac{1}{\rho_G}$. Convergence of the original sequence $(p_{t_n}(x_n))_n$ follows from a standard subsequence argument.

Next, we consider the Lipschitz continuity for fixed t. The optimality condition of the proximal map implied that for any $z \in \mathbb{R}^d$ there exists a $v_z \in \partial G(p(z))$ such that $0 = \frac{1}{t}(p(z) - z) + v_z$. It follows

$$\begin{aligned} \|p(x) - p(y)\|^2 &= \langle p(x) - p(y), x - tv_x - (y - tv_y)\rangle \\ &= \langle p(x) - p(y), x - y\rangle - t\langle p(x) - p(y), v_x - v_y\rangle. \end{aligned} \tag{A.2}$$

As mentioned in Theorem 4.4, the regular subgradient of the convex function $G + \frac{\rho_G}{2}\|\cdot\|^2$ satisfies $\partial(G + \frac{\rho_G}{2}\|\cdot\|^2)(x) = \partial G(x) + \rho_G x$ (see [39, 10.10 Exercise]). Therefore, it holds that

$$0 \leq \langle p(x) - p(y), (v_x + \rho_G p(x)) - (v_y + \rho_G p(y))\rangle = \langle p(x) - p(y), v_x - v_y\rangle + \rho_G\|p(x) - p(y)\|^2.$$

Plugging this into (A.2) yields that

$$\|p(x) - p(y)\|^2 \leq \langle p(x) - p(y), x - y\rangle + t\rho_G\|p(x) - p(y)\|^2$$

Thus, $\|p(x) - p(y)\|^2 \leq \frac{1}{1-\rho_G t}\langle p(x) - p(y), x - y\rangle$. Applying the Cauchy-Schwartz inequality yields the desired result. ■

A.1.2. Proof of Theorem 4.7.

Proof. Let $v \in \mathbb{R}^d$ be arbitrary. By the definition of the proximal map, we find that

$$\begin{aligned} M_G^t(x+v) - M_G^t(x) &\leq \frac{1}{2t}\|x+v-p_t(x)\|^2 + G(p_t(x)) - \frac{1}{2t}\|x-p_t(x)\|^2 - G(p_t(x)) \\ &= \frac{1}{t}\langle v, x-p_t(x)\rangle + \frac{1}{2t}\|v\|^2. \end{aligned} \tag{A.3}$$

Similarly, it holds true that $M_G^t(x+v) - M_G^t(x) \geq \frac{1}{t}\langle v, x-p_t(x+v)\rangle + \frac{1}{2t}\|v\|^2$. In summary, we have that

$$\begin{aligned} |M_G^t(x+v) - M_G^t(x) - \langle v, \tfrac{1}{t}(x-p_t(x))\rangle| &\leq \tfrac{1}{t}\|v\|\|p_t(x+v)-p_t(x)\| + \tfrac{1}{2t}\|v\|^2 \\ &\leq \tfrac{1}{t}\|v\|(\tfrac{1}{1-\rho_G t}\|v\| + \tfrac{1}{2}\|v\|) = o(\|v\|). \end{aligned} \tag{A.4}$$

■

A.1.3. Proof of Theorem 4.9.

Proof. By continuity, the set $\mathrm{prox}_{tG}(x)$ is closed and, consequently, the infimum is a minimum. Denote $\hat{p}_t(x) = \arg\min_{p\in\mathrm{prox}_{tG}(x)}\|p\|$. The optimality condition

$$0 \in \tfrac{1}{t}(\hat{p}_t(x) - x) + \partial G(\hat{p}_t(x))$$

implies the existence of a $v_t(x) \in \partial G(\hat{p}_t(x))$ such that $\|x\| \leq \|\hat{p}_t(x)\| + t\|v_t(x)\|$. Now assume to the contrary, there exist sequences $R_n \to \infty$, $\|x_n\| \geq R_n$ with $\|\hat{p}_t(x_n)\| \leq C < \infty$ for all $n \in \mathbb{N}$. But then, the local boundedness of ∂G (see Theorem 4.4) implies a uniform bound on $\|v_t(x_n)\|$ and thus on $\|x_n\|$, which is a contradiction. ■

A.1.4. Proof of Theorem 4.10.

Proof. By Theorem 4.9 there exists $R > 0$ such that for $x \in B_R^c(0)$, $\mathrm{prox}_{tG}(x) \subset B_M^c(0)$ where M is the radius outside of which G is convex, see Assumption 4.3.3. Thus, the map $p \mapsto \frac{1}{2t}\|x-p\|^2 + G(p)$ is strictly convex outside $B_M(0)$. Since all minimizers of this map are elements of $B_M^c(0)$, the minimizer has to be unique, that is, $\mathrm{prox}_{tG}(x)$ is single-valued. As in Theorem 4.6, we obtain Lipschitz continuity of the proximal map, but using now convexity of G outside the ball $B_M(0)$ instead of weak convexity. This also implies differentiability of M_G^t. Regarding the Lipschitz continuity of ∇M_G^t, let us denote $q_z = z - p(z)$ for $z = x, y$, so that $\nabla M_G^t(z) = \frac{1}{t}q_z$. Moreover, by the definition of the prox $q_z \in t\partial G(p(z)) = t\partial G(z - q_z)$. As a consequence, since $p(z) \notin B_M(0)$ so that convexity of G outside a ball can be used, $\langle p(x)-p(y), q_x - q_y\rangle = \langle x - q_x - (y-q_y), q_x - q_y\rangle \geq 0$. It follows

$$\|q_x - q_y\|^2 = \langle x-y, q_x-q_y\rangle - \langle x-q_x-(y-q_y), q_x-q_y\rangle \leq \|x-y\|\|q_x-q_y\|. \tag{A.5}$$

Therefore, ∇M_G^t is Lipschitz continuous with Lipschitz constant $\frac{1}{t}$. ■

A.1.5. Proof of Theorem 4.11.

Proof. Let R as in the proof of Theorem 4.10 be large enough such that $x \in B_R^c(0)$ implies that $\operatorname{prox}_{tG}(x) \in B_M^c(0)$. A simple computation then yields for $x, y \in B_R^c(0)$

$$\begin{aligned} M_G^t(\lambda x + (1-\lambda)y) &\leq \frac{1}{2t}\|\lambda p(x) + (1-\lambda)p(y) - (\lambda x + (1-\lambda)y)\|^2 + G(\lambda p(x) + (1-\lambda)p(y)) \\ &\leq \frac{\lambda}{2t}\|p(x) - x\|^2 + \lambda G(p(x)) + \frac{1-\lambda}{2t}\|p(y) - y\|^2 + (1-\lambda)G(p(y)) \\ &= \lambda M_G^t(x) + (1-\lambda)M_G^t(y) \end{aligned} \tag{A.6}$$

where the first inequality follows from definition of the Moreau envelope and the second inequality follows from from convexity of $\|\cdot\|$ together with convexity of G outside a ball. ■

A.1.6. Proof of Theorem 4.13.

Proof. First, we note that x^* is also a minimizer of M_G^t. Let $R > 0$ be such that $M_G^t(x)$ is differentiable at $x \in \mathbb{R}^d$ that fulfills $\|x\| > R$. Pick $M'_\rho > 0$ such that $M'_\rho \geq 2M_\rho$ and such that for all $x \in \mathbb{R}^d$, $\|x - x^*\| \geq M'_\rho$ implies that $\|x\| > R$. Let $p = \operatorname{prox}_{tG}(x)$ and where $x \in \mathbb{R}^d$ is such that $\|x - x^*\| \geq M'_\rho$. We make a case distinction: When $\|p - x^*\| \leq \frac{\|x-x^*\|}{2}$, we can compute that

$$\begin{aligned} \langle \nabla M_G^t(x), x - x^* \rangle &= \langle \tfrac{1}{t}(x - x^* + x^* - p), x - x^* \rangle \\ &\geq \tfrac{1}{t}(\|x - x^*\|^2 - \|p - x^*\|\|x - x^*\| \\ &\geq \tfrac{1}{2t}\|x - x^*\|^2. \end{aligned} \tag{A.7}$$

Conversely, assume now $\|p - x^*\| > \frac{\|x-x^*\|}{2} \geq M_\rho$ and let $v \in \partial G(p)$ such that $0 = \frac{1}{t}(p - x) + v$. It follows

$$\begin{aligned} \langle \nabla M_G^t(x), x - x^* \rangle &= \langle v, x - x^* \rangle \\ &= \langle v, p - x^* + tv \rangle \\ &\geq \langle v, p - x^* \rangle \\ &\underset{(*)}{\geq} \rho\|p - x^*\|^2 \geq \frac{\rho}{4}\|x - x^*\|^2 \end{aligned} \tag{A.8}$$

where $(*)$ follows from G being superexponential. In summary, we find that for $x \in \mathbb{R}^d$ such that $\|x - x^*\| \geq M'_\rho$, it holds that

$$\langle \nabla M_G^t(x), x - x^* \rangle \geq \min\left(\frac{\rho}{4}, \frac{1}{2t}\right)\|x - x^*\|^2. \tag{A.9}$$

■

A.2. Consistency of DAZ.

A.2.1. Proof of Theorem 4.18.

Proof. Let $0 < s < t < t_{\max}$. By definition of the Moreau envelope, we find that

$$\begin{aligned} &\frac{1}{2t}\|x - p_t\|^2 + G(p_t) - \frac{1}{2s}\|x - p_t\|^2 - G(p_t) \\ &\leq M_G^t(x) - M_G^s(x) \\ &\leq \frac{1}{2t}\|x - p_s\|^2 + G(p_s) - \frac{1}{2s}\|x - p_s\|^2 - G(p_s), \end{aligned} \tag{A.10}$$

and, consequently, that

$$-\frac{1}{2st}\|x-p_t\|^2 \leq \frac{M_G^t(x)-M_G^s(x)}{t-s} \leq -\frac{1}{2st}\|x-p_s\|^2.$$

By Theorem 4.6 the result follows and, in particular, we obtain $\partial_t M_G^t(x) = -\frac{1}{2t^2}\|x-p_t\|^2$. ■

A.2.2. Proof of Theorem 4.20.

Proof. Using Theorem 4.18 we find for $0 < s < t$

$$|M_G^t(x) - M_G^s(x)| \leq \int_s^t |\partial_\tau M_G^\tau(x)|\, d\tau = \frac{1}{2}\int_s^t \frac{1}{\tau^2}\|x-p_\tau\|^2\, d\tau. \tag{A.11}$$

By definition of the Moreau envelope $\frac{1}{\tau}(x-p_\tau) \in \partial G(p_\tau)$. Therefore, $\frac{1}{\tau^2}\|x-p_\tau\|^2 \leq \phi(\|x\|)^2$ and consequently $|M_G^t(x) - M_G^s(x)| \leq \frac{|t-s|}{2}\phi(\|x\|)^2$. When $s=0$, it holds that $|M_G^0(x) - M^t(x)| = \lim_{s\to 0}|M_G^s(x) - M_G^t(x)| \leq \lim_{s\to 0}\frac{|t-s|}{2}\phi(\|x\|)^2 = \frac{|0-t|}{2}\phi(\|x\|)^2$. ■

A.2.3. Miscellaneous results.

Lemma A.1. *Let $H : \mathbb{R}^d \to \mathbb{R}$ be locally bounded and convex outside the ball $B_M(0)$ and such that $\int \exp(-H(x))\mathrm{d}x < \infty$. Then, H grows asymptotically at least linearly, i.e., there exist a $C > 0$ such that $H(x) \geq C\|x\|$ for sufficiently large x.*

Proof. The proof is similar to [10, Lemma 4]. Let M be such that H is convex outside of $B_M(0)$. Define $m = \sup_{x\in B_{2M}(0)} H(x) < \infty$. We claim now that the set $\{H \leq m+1\}$ is bounded. Assume to the contrary, there exists y_n, $\|y_n\| \geq n$ such that $H(y_n) \leq m+1$. By convexity outside a ball it follows that $\mathrm{co}(B_{2M}(0) \cup \{y_n\}) \subset \{H \leq m+1\}$ where co denotes the convex hull. To see that, let $z \in \mathrm{co}(B_{2M}(0)\cup\{y_n\})$, *i.e.*, there exists $k \in \mathbb{N}$, $x_1,\dots,x_k \in B_{2M}(0)$, and $\lambda_1,\dots,\lambda_{k+1} \geq 0$ with $\sum_i \lambda_i = 1$ such that $z = \sum_{i=1}^k \lambda_i x_i + \lambda_{k+1}y_n$. We have to show, that $H(z) \leq m+1$. If $\|z\| \leq 2M$ the result follows by definition of m, thus, let us assume that $\|z\| > 2M$ which also implies $\lambda_{k+1} > 0$. Note that

$$z = \left(\sum_{i=1}^k \lambda_i\right)\underbrace{\frac{\sum_{i=1}^k \lambda_i x_i}{\sum_{i=1}^k \lambda_i}}_{\in B_{2M}(0)} + \left(1 - \sum_{i=1}^k \lambda_i\right) y_n.$$

That is, z is a convex combination of a single element in $B_{2M}(0)$ and y_n. It is easy to see that since $\|z\| > 2M$ we can also find $x \in \mathbb{R}^d$ with $\|x\| = 2M$ such that z is a convex combination of this specific x and y_n, that is, there exists $\mu \in [0,1]$ such that $z = \mu x + (1-\mu)y_n$. But then, convexity outside a ball implies $H(z) \leq m+1$, yielding that $\mathrm{co}(B_{2M}(0)\cup\{y_n\}) \subset \{H \leq m+1\}$.

Since the Lebesgue measure of $\mathrm{co}(B_{2M}(0)\cup\{y_n\})$ tends to infinity as $n\to\infty$, we obtain a contradiction to $\int \exp\bigl(-H(x)\bigr)\,\mathrm{d}x < \infty$ (*cf.* [13, Lemma 4]). As a consequence the set $\{H \leq m+1\}$ is bounded, *i.e.*, there exists $R > 2M$ such that for $\|x\| \geq R$ it follows $H(x) > m+1$.

Now take $x \in B_R^c(0)$. We can write $R\frac{x}{\|x\|} = (1-\lambda)2M\frac{x}{\|x\|} + \lambda x$ with $\lambda = \frac{R-2M}{\|x\|-2M}$. It follows by convexity outside a ball, that

$$H\bigl(R\tfrac{x}{\|x\|}\bigr) \leq \lambda H(x) + (1-\lambda)H\bigl(2M\tfrac{x}{\|x\|}\bigr) \tag{A.12}$$

and as a result

$$
\begin{aligned}
H(x) - H\big(2M\tfrac{x}{\|x\|}\big) &\geq \frac{1}{\lambda}\Big(H\big(R\tfrac{x}{\|x\|}\big) - H\big(2M\tfrac{x}{\|x\|}\big)\Big) \\
&\geq \frac{\|x\| - 2M}{R - 2M}\Big(\inf_{\|y\|=R} H(y) - \sup_{\|y\|=2M} H(y)\Big).
\end{aligned}
\tag{A.13}
$$

Since $H(y) \leq m$ in $B_{2M}(0)$ and $H(y) > m+1$ for $\|x\| \geq R$ the result follows. ■

Lemma A.2. *Let G be superexponential, then for a sufficiently large $x \in \mathbb{R}^d$, it follows that $\|\operatorname{prox}_{tG}(x)\| \leq \|x\| + c$ for some $c > 0$.*

Proof. Let $p \in \operatorname{prox}_{tG}(x)$. By definition of the proximal map there exists a $v \in \partial G(p)$ such that $0 = \frac{1}{t}(p - x) + v$. Since G is superexponential, there exist $x^* \in \mathbb{R}^d$, $\rho > 0$, and $M_\rho > 0$ such that $\|x - x^*\| \geq M_\rho$ implies $\langle v, x - x^*\rangle \geq \rho\|x - x^*\|^2$, $v \in \partial G(x)$. Now let $R > 0$ be such that $x \in B_R^c(0)$ implies $\|p - x^*\| \geq M_\rho$ (*cf.*Theorem 4.9). Then, we can conclude for such x that

$$
\begin{aligned}
\|x - x^*\|^2 &= \|p - x^*\|^2 + 2t\langle p - x^*, v\rangle + t^2\|v\|^2 \\
&\geq \|p - x^*\|^2.
\end{aligned}
\tag{A.14}
$$

■

Appendix B. Implementational Details.

Algorithm B.1 APGD algorithm with Lipschitz backtracking.

```
Require: Number of iterations $K$, initial condition $x^0$, initial $L_0$, number of backtracking
         iterations $J$, $\beta \in (0,1)$, $\gamma > 1$, relative tolerance $r$
$x^{-1} = x^0$
for $k = 0, 1, \dots, K-1$ do
    $\bar{x} = x^k + (x^k - x^{k-1})/\sqrt{2}$
    for $j = 0, 1, \dots, J-1$ do                              ▷ Lipschitz backtracking procedure [31]
        $x^{k+1} = \operatorname{prox}_{\frac{1}{L_k}g}(x - \nabla f(\bar{x}/L_k))$
        if $f(x^{k+1}) \leq f(\bar{x}) + \langle \nabla f(\bar{x}), x^{k+1} - \bar{x}\rangle + \frac{L_k}{2}\|\bar{x} - x^{k+1}\|^2$ then
            $L_k = \beta L_k$
            break
        $L_k = \gamma L_k$
    if $\|x^k - x^{k+1}\|/\|x^k\| \leq r$ then                      ▷ Stopping criterion
        break
return $x^k$
```