# FINITE-SAMPLE BOUNDS FOR ADAPTIVE INVERSE REINFORCEMENT LEARNING USING PASSIVE LANGEVIN DYNAMICS* 

LUKE SNOW AND VIKRAM KRISHNAMURTHY $\dagger$


#### Abstract

This paper provides a finite-sample analysis of a passive stochastic gradient Langevin dynamics algorithm (PSGLD) designed to achieve adaptive inverse reinforcement learning (IRL). By passive, we mean that the noisy gradients available to the PSGLD algorithm (inverse learning process) are evaluated at randomly chosen points by an external stochastic gradient algorithm (forward learner) that aims to optimize a cost function. The PSGLD algorithm acts as a randomized sampler to achieve adaptive IRL by reconstructing this cost function nonparametrically from the stationary measure of a Langevin diffusion. Previous work has analyzed the asymptotic performance of this passive algorithm using weak convergence techniques. This paper analyzes the non-asymptotic (finite-sample) performance using a logarithmic-Sobolev inequality and the Otto-Villani Theorem. We obtain finite-sample bounds on the 2-Wasserstein distance between the estimates generated by the PSGLD algorithm and the cost function. Apart from achieving finite-sample guarantees for adaptive IRL, this work extends a line of research in analysis of passive stochastic gradient algorithms to the finite-sample regime for Langevin dynamics.


Key words. stochastic gradient Langevin dynamics, passive learning, inverse reinforcement learning, finite-sample analysis, logarithmic-Sobolev inequality, Wasserstein distance

1. Introduction. We derive non-asymptotic bounds for a Langevin dynamics algorithm performing real-time inverse reinforcement learning (IRL). Traditional IRL [21], [12], [4] reconstructs the cost function of a Markov Decision Process by observing decisions taken from an optimal policy, i.e., after an observed agent has completed learning the optimal policy. Here, we consider real-time (adaptive) IRL. We observe an agent (forward learner) performing stochastic gradient descent (e.g, policy gradient reinforcement learning) on a cost function $J$, and attempt to reconstruct $J$ in realtime. Thus, this technique can be regarded as an inverse stochastic gradient algorithm. It applies to IRL problems in several contexts such as adaptive Bayesian learning, constrained Markov Decision Processes, and logistic regression classification [16].

To accomplish real-time IRL, we employ a passive stochastic gradient Langevin dynamics (PSGLD) algorithm, initially proposed in [16]. Given observations of sequential stochastic gradient descent (SGD) evaluations on $J$, the PSGLD algorithm acts as a Markov chain Monte Carlo (MCMC) sampler designed to reconstruct $J$. Classical stochastic gradient Langevin dynamics [28], [10], evolves in the direction of a stochastic gradient evaluation (of a cost function) at the current iterate, plus an independent Gaussian perturbation. It asymptotically samples from the Gibbs measure encoding the cost function, and thus can be used to sample from probability distributions in the context of e.g., Bayesian learning [28] or empirical risk minimization [22]. Our PSGLD algorithm is considered passive because the stochastic gradient evaluations are not directly controlled, but are provided by the observed SGD process. Remarkably, [16] use stochastic approximation techniques to show that the PSGLD

[^0]algorithm asymptotically samples from the Gibbs measure encoding the cost function being optimized by the observed SGD process. So the PSGLD algorithm asymptotically achieves inverse reinforcement learning. Similar passive schemes and stochastic approximation analyses have been investigated in [17], [30], [19].

In any practical IRL implementation it is necessary to understand how well the PSGLD algorithm recovers this cost function after a finite run-time. In this work we present a non-asymptotic analysis of this PSGLD algorithm; we provide finite-sample bounds on the 2-Wasserstein distance between the law of the algorithm and that of the Gibbs measure encoding the cost function. Non-asymptotic analysis of stochastic gradient Langevin dynamics has been investigated in [22], [8], [6]. In our case the algorithm is passive; so our analysis generalizes and extends previous works to handle this complexity.

Main Result: Recall $J$ denotes the cost function being optimized by the forward learner, and which we aim to reconstruct. Denote $\pi_{k}$ the sampling measure of our PSGLD algorithm at iterate $k \in \mathbb{N}$, $\pi_{\infty}$ the Gibbs measure w.r.t. $J\left(\pi_{\infty} \propto \exp (-\beta J)\right.$, with $\beta$ a controllable PSGLD parameter), and $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right)$ the 2 -Wasserstein distance between these. Observe that sampling from $\pi_{\infty}$, achieved when $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right)=0$, suffices to reconstruct $J$ by taking the log-sample density. We obtain a bound on $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right)$ that scales as $\mathcal{O}(k \epsilon \sqrt{\epsilon}+\exp (-k \epsilon))$, where $\epsilon$ is the algorithm step size. So we can choose $k \epsilon$ fixed and large enough to diminish $\exp (-k \epsilon)$, then $\epsilon$ small enough (with $k$ simultaneously increasing to fix $k \epsilon$ ) to diminish $k \epsilon \sqrt{\epsilon}$ arbitrarily. Thus, for any $y>0$ we can choose the step size $\epsilon$ small enough and algorithmic iterate $k$ large enough such that $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right) \leq \mathcal{O}(y)$. The main result of this paper is a precise formulation of this statement.

Proof Technique: We bound $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right) \leq \mathcal{W}_{2}\left(\pi_{k}, \nu_{k \epsilon}\right)+\mathcal{W}_{2}\left(\nu_{k \epsilon}, \pi_{\infty}\right)$, where $\nu_{k \epsilon}$ is the measure, at time $k \epsilon$, of a particular continuous time diffusion with stationary measure $\pi_{\infty}$. We obtain $\mathcal{W}_{2}\left(\pi_{k}, \nu_{k \epsilon}\right) \leq \mathcal{O}(k \epsilon \sqrt{\epsilon})$ through a Girsanov change of measure technique and a weighted transportation cost inequality. We then show that the diffusion satisfies a logarithmic-Sobolev inequality, allowing us to employ exponential decay of entropy and the Otto-Villani Theorem to show exponential decay of $\mathcal{W}_{2}\left(\nu_{k \epsilon}, \pi_{\infty}\right)$. This proof structure is mirrored in [22]. However, our algorithm necessitates a non-trivial extension of the methods in [22]; we utilize both a generalized stochastic gradient Langevin dynamics form and a weighting kernel to control the external gradient evaluations. The generalized form disrupts absolute continuity of measure between the algorithm and continuous time diffusion, necessitating the introduction and control of an intermediate process in order to apply Girsanov's Theorem. It also necessitates control of the "sampling distribution" (from which initial SGD and PSGLD points are taken) to decrease discretization error. However, this control simultanesouly increases a relative entropy term appearing in the final 2-Wasserstein bound; this is handled by a careful specification of other algorithmic parameters. Finally, the continuous time diffusion is also distinct from that in [22], so we must prove logarithmic-Sobolev inequality satisfaction by a novel Lyapunov function. Furthermore, there are a host of supporting Lemmas, such as exponential integrability of the generalized diffusion, which have been necessarily derived for our analysis.

Another motivation for our work is the extension of a line of research in analysis of passive stochastic gradient algorithms. Such work [23] [13], [20], [31] has historically focused on passive stochastic approximation for e.g., sequential non-parametric estimation of regression functions. Recently, [16] extended this analysis to stochastic gradient Langevin dynamics in the passive regime. This paper serves as the first finite-sample result for stochastic gradient Langevin dynamics in the passive regime.
1.1. Organization. Section 2 provides background on passive stochastic gradient Langevin dynamics and the asymptotic result of [16]. Section 3 discusses our main result, a non-asymptotic 2-Wasserstein bound (Theorem 3.9), and provides an example of how this IRL result interfaces with a canonical reinforcement learning algorithm. In section 4 we provide additional mathematical background for the proof of our bound, and section 5 provides details on the structure of our proof of Theorem 3.9. The complete proof details appear in the supplementary material.
2. Background: Passive Langevin Dynamics. Motivating our finite-sample analysis, this section introduces Langevin dynamics, presents the adaptive IRL setting and discusses the weak convergence asymptotic analysis of the PSGLD algorithm in [16].
2.1. Langevin Dynamics. The classical Langevin dynamics algorithm is given, with step size $\epsilon_{k}$, objective function $J$, noise parameter $\beta$, and i.i.d. standard $N$ variate Gaussian noise $w_{k}$, as

$$
\begin{equation*}
\theta_{k+1}=\theta_{k}-\epsilon_{k} \nabla J\left(\theta_{k}\right)+\sqrt{2 \epsilon \beta^{-1}} w_{k}, \quad k \in \mathbb{N} \tag{2.1}
\end{equation*}
$$

Here $\theta_{k} \in \mathbb{R}^{N}$ is initialized by $\theta_{0} \sim \pi_{0}$ for some sampling distribution $\pi_{0}$ on $\mathbb{R}^{N}$. The algorithm (2.1) is used primarily for either non-convex optimization [29] or to sample from probability distributions via MCMC [28]. The former is accomplished by treating (2.1) as a simulated annealer, and letting the step size $\epsilon_{k}$ and 'temperature' $\beta^{-1}$ decrease to zero as $k \rightarrow \infty$. To accomplish the latter, the step size $\epsilon_{k}$ and temperature $\beta^{-1}$ are fixed for all $k$. This work considers the latter case of constant step-size Langevin dynamics, with $\epsilon_{k}=\epsilon \forall k \in \mathbb{N}$. It is well known that the Markov process (2.1), with constant step-size, asymptotically samples from the Gibbs measure

$$
\begin{equation*}
\pi_{\infty}(\theta)=\frac{1}{\Lambda} \exp (-\beta J(\theta)) \tag{2.2}
\end{equation*}
$$

Here $\Lambda=\int_{\mathbb{R}^{N}} \exp (-\beta J(\alpha)) d \alpha$ is a normalizing constant and $J$ is called the potential function. Indeed, (2.1) corresponds to a discretization of the continuous-time Langevin diffusion given by, with $\theta(t) \in \mathbb{R}^{N}$ and $W(t)$ standard Brownian motion in $\mathbb{R}^{N}$, the Itô stochastic differential equation (SDE)

$$
\begin{equation*}
d \theta(t)=-\nabla J(\theta(t)) d t+\sqrt{2 \beta^{-1}} d W(t), \quad t \geq 0 \tag{2.3}
\end{equation*}
$$

Under suitable conditions on $J$ and $\beta$, this SDE has the Gibbs measure (2.2) as its unique stationary measure [7]. Thus, the Langevin dynamics algorithm (2.1) can be used as a MCMC algorithm to asymptotically sample from any probability distribution which can be expressed as (2.2) with some potential function $J$.

In [25] more general reversible diffusions of the form, with $\sigma: \mathbb{R}^{N} \rightarrow \mathbb{R}$ differentiable,

$$
\begin{equation*}
d \theta(t)=\left[-\frac{\beta}{2} \nabla J(\theta) d t-\nabla \sigma(\theta) d t+d W(t)\right] \sigma(\theta) \tag{2.4}
\end{equation*}
$$

are studied, and it is shown that (2.4) has the same stationary measure (2.2) as the classical Langevin diffusion (2.3). The corresponding Euler-Maruyama time discretization results in the following discrete-time Markov process

$$
\begin{equation*}
\theta_{k+1}=\theta_{k}-\epsilon\left[\frac{\beta}{2} \nabla J\left(\theta_{k}\right)+\nabla \sigma\left(\theta_{k}\right)\right] \sigma\left(\theta_{k}\right)+\sqrt{\epsilon} \sigma\left(\theta_{k}\right) w_{k} \tag{2.5}
\end{equation*}
$$

which can thus equivalently be used as a MCMC sampler from (2.2). We will utilize the generalized process (2.5), as opposed to the classical Langevin dynamics (2.1), for our PSGLD algorithm ${ }^{1}$.

Active vs. Passive Gradient Evaluation: Notice that the above SGLD algorithms utilize, at each time step, the gradient $\nabla J\left(\theta_{k}\right)$ evaluated at the current iterate $\theta_{k}$. We term this active gradient evaluation, and distinguish this from passive gradient evaluation, where the gradient is evaluated at a different (uncontrolled) point. The following section introduces the adaptive IRL setting and motivates the need for passive gradient evaluation in our PSGLD algorithm.
2.2. Adaptive IRL: Forward and Inverse Learning. This section motivates the setting of adaptive IRL. An agent (forward learner) is in the process of optimizing a cost function. An inverse learner passively observes sequential algorithmic iterates of the forward learner, and attempts to reconstruct (learn) the cost function being optimized. This paper provides non-asymptotic guarantees on the inverse learner's cost function reconstruction. Here we first introduce the dynamics of the forward learner, then the inverse learning setting and PSGLD algorithm.
2.2.1. Forward Learner. The forward learner runs a stochastic gradient descent (SGD) on a non-negative cost function

$$
\begin{equation*}
J: \mathbb{R}^{N} \rightarrow \mathbb{R}_{+} \tag{2.6}
\end{equation*}
$$

where the initial point evaluation is sampled randomly from sampling distribution $\pi_{0, \gamma}$ on $\mathbb{R}^{N}$. Examples include policy gradient reinforcement learning [26], neural network optimization [24] and federated learning [14]. We define the sampling distribution $\pi_{0, \gamma}$ as follows

$$
\begin{equation*}
\pi_{0, \gamma}(x):=\frac{\pi_{0}\left(\frac{x}{\gamma}\right)}{\int_{\mathbb{R}^{N}} \pi_{0}\left(\frac{x}{\gamma}\right) d x} \tag{2.7}
\end{equation*}
$$

where $\pi_{0}$ is an arbitrary density function and $\gamma$ is a scale parameter ${ }^{2}$.
We assume the stochastic gradient algorithm (forward learner) resets after some finite time, so that the SGD process repeats indefinitely. This can be motivated by optimization of a non-convex function, in which re-initialization allows sufficient exploration, or multiple agents each learning to optimize the same cost function. Thus we have, for $n \in \mathbb{N}$ representing each "run" of the SGD , and $\tau_{n}$ stopping times:

$$
\begin{equation*}
\theta_{k+1}=\theta_{k}-\eta \hat{\nabla} J\left(\theta_{k}\right), \quad k \in\left\{\tau_{n}, \ldots, \tau_{n+1}-1\right\} \tag{2.8}
\end{equation*}
$$

where each $\theta_{\tau_{n}} \stackrel{i . i . d .}{\sim} \pi_{0, \gamma}$ and $\eta>0$ is a fixed step-size. Here $\hat{\nabla} J\left(\theta_{k}\right)$ is an unbiased estimate of the true gradient $\nabla J\left(\theta_{k}\right)$, with bounded variance, see 3.2. Algorithm 2.1 displays this randomly re-initializing stochastic gradient descent.

Remark: The forward learner is not necessarily restricted to implementing an SGD algorithm; any process which provides sequential stochastic gradients of the cost function $J$, including i.i.d. samples from measure $\pi_{0, \gamma}$ or sequential algorithmic iterates with more sophisticated dependencies than (2.8) will suffice, see [16].

[^1]```
Algorithm 2.1 Randomly Re-Initializing SGD Process
    initialize \(\tau_{0}=n=0\)
    while \(n \geq 0\) do
        generate \(\tau_{n+1}>\tau_{n}, \quad \theta_{\tau_{n}} \stackrel{i i d}{\sim} \pi_{0, \gamma}\)
        for \(k=\tau_{n}: \tau_{n+1}-1\) do
            \(\theta_{k+1} \leftarrow \theta_{k}-\eta \hat{\nabla} J\left(\theta_{k}\right)\)
        end for
    end while
```

2.2.2. Inverse Learning: PSGLD. In this paper we take the perspective of an inverse learner who observes the SGD process (2.8), and attempts to reconstruct the cost function $J$ being optimized. We assume this observer knows the sampling distribution $\pi_{0, \gamma}$ and can observe evaluations $\theta_{k}, k \in \mathbb{N}$. The agent recovers noisy gradient evaluations $\hat{\nabla} J\left(\theta_{k}\right)=\frac{\theta_{k+1}-\theta_{k}}{\eta}$.

Using only these sequential noisy gradient evaluations, how can the agent learn $J$ ? This is accomplished via MCMC sampling, using the following passive stochastic gradient Langevin dynamics (PSGLD) updates:

$$
\begin{align*}
\alpha_{k+1}= & \alpha_{k}  \tag{2.9}\\
& -\epsilon\left[K_{\Delta}\left(\theta_{k}-\alpha_{k}\right) \frac{\beta}{2} \hat{\nabla} J\left(\theta_{k}\right)+\nabla \pi_{0, \gamma}\left(\alpha_{k}\right)\right] \pi_{0, \gamma}\left(\alpha_{k}\right)+\sqrt{\epsilon} \pi_{0, \gamma}\left(\alpha_{k}\right) w_{k} \\
& \alpha_{0} \sim \pi_{0, \gamma}
\end{align*}
$$

Note that $\alpha_{0}$ is sampled randomly from the sampling distribution $\pi_{0, \gamma}$ of the SGD process (2.8). Here $\left\{w_{k}, k \geq 0\right\}$ is an i.i.d. sequence of standard $N$-variate Gaussian random variables,

$$
\begin{equation*}
K_{\Delta}\left(\theta_{k}-\alpha_{k}\right):=\frac{1}{\Delta^{N}} K\left(\frac{\theta_{k}-\alpha_{k}}{\Delta}\right) \tag{2.10}
\end{equation*}
$$

is the $\Delta$-parametrized kernel function, and $\beta$ is the inverse temperature parameter. The algorithm is passive since the stochastic gradients $\hat{\nabla} J\left(\theta_{k}\right)$ and evaluation points $\theta_{k}$ are passively observed from SGD process (2.8). The kernel ${ }^{3}$ function $K(\cdot)$ controls for bias in these passive gradient evaluations, and can be chosen by the observer as any function $K: \mathbb{R}^{N} \rightarrow \mathbb{R}$ satisfying:

$$
\begin{align*}
& K(u) \geq 0, \quad K(u)=K(-u), \quad \sup _{u} K(u)<\infty \\
& \int_{\mathbb{R}^{N}} K(u) d u=1, \quad \int_{\mathbb{R}^{N}}|u|^{2} K(u)<\infty \tag{2.11}
\end{align*}
$$

$K_{\Delta}$ weights the relevance of stochastic gradient $\hat{\nabla} J\left(\theta_{k}\right)$ to the current iterate $\alpha(t)$. We obtain $K_{\Delta}$ by modulating $K$ by the domain scaling parameter $\Delta$ as (2.10). So $\Delta$ modulates the degree to which samples $\theta_{k}$ at a fixed distance from current iterate $\alpha(t)$ impact the algorithm's evolution.

Algorithm 2.2 displays this passive stochastic gradient Langevin dynamics algorithm, which takes as input the sequential evaluations $\theta_{k}$ made in Algorithm 2.1.

We claim that Algorithm 2.2 achieves adaptive inverse reinforcement learning, reconstructing $J$ by taking only $\hat{\nabla} J\left(\theta_{k}\right)$ as input. Next a background result is presented which establishes this claim rigorously.

[^2]```
Algorithm 2.2 PSGLD
    parameters: step size \(\epsilon\), inverse temperature \(\beta\), kernel scale \(\Delta\), re-sampling distri-
    bution scale \(\gamma\)
    initialize \(\alpha_{0} \sim \pi_{0, \gamma}\)
    while \(k \geq 0\) do
        obtain \(\theta_{k}\) from a Algorithm 2.1
        if \(k \geq 1\) then
            \(\hat{\nabla} J\left(\theta_{k}\right)=\frac{1}{\eta}\left(\theta_{k}-\theta_{k-1}\right), \quad K_{k-1}=\frac{1}{\Delta^{N}} K\left(\frac{\theta_{k-1}-\alpha_{k-1}}{\Delta}\right)\)
            sample \(w_{k} \sim \mathcal{N}\left(0, I_{N}\right)\)
            \(\alpha_{k} \leftarrow \alpha_{k-1}-\epsilon\left[K_{k-1} \frac{\beta}{2} \hat{\nabla} J\left(\theta_{k}\right)+\nabla \pi_{0, \gamma}\left(\alpha_{k-1}\right)\right] \pi_{0, \gamma}\left(\alpha_{k-1}\right)+\sqrt{\epsilon} \pi_{0, \gamma}\left(\alpha_{k-1}\right) w_{k}\)
        end if
    end while
```

2.3. Passive SGLD: Asymptotic Convergence. [16] provides the following result:

Proposition 2.1 (Weak Convergence [16]). Let $\alpha^{\epsilon}(t)=\alpha_{k}$ for $t \in[\epsilon k, \epsilon(k+1)]$ be the continuous-time interpolation of PSGLD (2.9). Under assumptions (A1)-(A4) of [16], the process $\alpha^{\epsilon}(t)$ converges weakly to the solution of the stochastic differential equation

$$
\begin{align*}
& d \alpha(t)=-\left[\frac{\beta}{2} \pi_{0, \gamma}^{2}(\alpha(t)) \nabla J(\alpha(t))+\nabla \pi_{0, \gamma}(\alpha(t)) \pi_{0, \gamma}(\alpha(t))\right] d t+\pi_{0, \gamma}(\alpha(t)) d W(t)  \tag{2.12}\\
& \alpha(0)=\alpha_{0} \sim \pi_{0, \gamma}
\end{align*}
$$

where $W(t)$ is standard $N$-dimensional Brownian motion. Furthermore, the stochastic differential equation (2.12) has $\pi_{\infty}(2.2)$ as its stationary distribution.

Thus, we can use the algorithm (2.9) to generate asymptotic samples

$$
\alpha_{k} \sim \pi_{\infty}(\alpha) \propto \exp (-\beta J(\alpha))
$$

and reconstruct $J$ from the sample log-density.
Motivation: Proposition 2.1 shows that Algorithm 2.2 asymptotically produces samples $\alpha_{k} \sim \pi_{\infty}$, and so the cost function $J$ can be reconstructed from the logarithm of the asymptotic sample density. However, in this paper we are interested in quantifying how well this sampling algorithm approximates the Gibbs measure after a finite run-time. Our main result gives non-asymptotic (finite-sample) bounds on the 2-Wasserstein distance between the distribution of the sampling density produced by Algorithm 2.2 and the Gibbs measure $\pi_{\infty}$ (2.2).
3. Main Result. Non-Asymptotic Analysis of Passive Stochastic Gradient Langevin Dynamics. Recall from the Introduction that $\pi_{k}$ is the sampling measure of Algorithm 2.2 at iterate $k, \pi_{\infty}$ is the Gibbs measure proportional to $\exp (-\beta J)$, and $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right)$ is the 2 -Wasserstein distance between these. Our main result is as follows: for any $y>0$ we can choose the step size $\epsilon$ small enough and iteration number $k$ large enough such that $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right) \leq \mathcal{O}(y)$. In this section we formulate this result precisely. We provide a brief overview of the 2-Wasserstein metric, specify assumptions on the cost function $J$ and sampling distribution $\pi_{0, \gamma}$, provide the main bound in the form of Theorem 3.9, and discuss the application to adaptive inverse reinforcement learning in a Markov Decision Process framework.
3.1. 2-Wasserstein Distance. We provide a non-asymptotic bound on the convergence of (2.9) to the Gibbs measure $\pi_{\infty}(2.2)$, in terms of the 2-Wasserstein distance:

$$
\begin{equation*}
\mathcal{W}_{2}(\mu, \nu):=\inf _{\gamma \in \Gamma(\mu, \nu)}\left(\mathbb{E}_{(x, y) \sim \gamma}\|x-y\|^{2}\right)^{1 / 2} \tag{3.1}
\end{equation*}
$$

Here $\Gamma(\mu, \nu)$ is the set of all couplings of measures $\mu$ and $\nu$, where a coupling $\gamma$ is a joint probability measure on $\mathbb{R}^{N} \times \mathbb{R}^{N}$ with marginals $\mu$ and $\nu$, i.e., $\gamma\left(A, \mathbb{R}^{N}\right)=$ $\mu(A), \gamma\left(\mathbb{R}^{N}, B\right)=\nu(B), \forall A, B \in \mathcal{B}\left(\mathbb{R}^{N}\right)$, where $\mathcal{B}\left(\mathbb{R}^{N}\right)$ is the Borel $\sigma$-algebra of $\mathbb{R}^{N}$. Notice that the Wasserstein distance (3.1) indeed satisfies all axioms of a metric on the space of measures. The 2-Wasserstein distance is a more suitable metric for assessing the quality of approximate sampling schemes [8], [22], than others such as total-variation norm, since it gives direct guarantees on the accuracy of approximating higher order moments [8]. However, it also precludes us from using SDE discretization analysis presented in the seminal book [?], which utilizes total-variational norm.
3.2. Assumptions. Here we list several assumptions on the cost function $J$ (2.6) of the forward learner and the base sampling distribution $\pi_{0}$, required for the finite-sample analysis. Assumptions on $J$ are standard and equivalent to those taken in [22]. Assumptions on the base sampling distribution $\pi_{0}$ hold for a wide class of probability density functions, including Gaussian densities.

A 3.1 ( $J$ regularity). $J$ is $L_{J}$-Lipschitz continuous and $L_{\nabla J}$-smooth: $\exists L_{J}$, $L_{\nabla J}>0$ such that for all $x, y \in \mathbb{R}^{N}$,

$$
\|J(x)-J(y)\| \leq L_{J}\|x-y\|, \quad\|\nabla J(x)-\nabla J(y)\| \leq L_{\nabla J}\|x-y\|
$$

A 3.2 (Dissipativity). $J$ is $(m, b)$-dissipative:

$$
\exists m>0, b \geq 0:\langle x, \nabla J(x)\rangle \geq m\|x\|^{2}-b, \forall x \in \mathbb{R}^{N}
$$

A 3.3 (Gradient Noise Variance). The noisy SGD gradient evaluation is unbiased, i.e. $\mathbb{E}[\hat{\nabla} J(x)]=\nabla J(x) \forall x \in \mathbb{R}^{N}$. Furthermore, the noise is additive such that $\hat{\nabla} J(x)-\nabla J(x)$ is i.i.d. with variance bounded uniformly in $x$, i.e., there exists a constant $\zeta \geq 0$ such that

$$
\mathbb{E}\left[\|\hat{\nabla} J(x)-\nabla J(x)\|^{2}\right] \leq \zeta, \forall x \in \mathbb{R}^{N}
$$

A 3.4 ( $\pi_{0}$ Exponential Decay). The base sampling distribution $\pi_{0}$ has an exponential tail decay and differential decay $\mathcal{O}\left(\|x\|^{-1}\right)$, i.e.,

$$
\exists M \in \mathbb{N}, \tilde{C}>0: \pi_{0}(x) \leq \exp \left(-\|x\|^{2}\right), \quad\left\|\nabla \pi_{0}(x)\right\| \leq \frac{\tilde{C}}{\|x\|} \forall\|x\|>M
$$

A 3.5 ( $\pi_{0}$ Lipschitz-continuity).

$$
\exists L_{\pi_{0}}>0:\left\|\pi_{0}(x)-\pi_{0}(y)\right\| \leq L_{\pi_{0}}\|x-y\| \forall x, y \in \mathbb{R}^{N}
$$

A 3.6 ( $\pi_{0}$ Structure). $\pi_{0}$ is unimodal and has support on $\mathbb{R}^{N}$.
A 3.7 (Kernel Structure). The kernel function $K(\cdot)$ satisfies (2.11).
A 3.8 (Feasible Parameter Ranges). Here $\wedge$ denotes the min operator and $\vee$ the max operator. Assume
i) $\eta \in\left(0,1 \wedge \frac{m}{4 L_{\nabla J}^{2}}\right)$
ii) $\epsilon \in\left(0,1 \wedge \sqrt{\frac{1}{249}} L_{\nabla J}^{-1}\right)$
iii) $\beta \geq \frac{1}{4 L_{\nabla J}^{2}} \vee \frac{\sqrt{2 \pi+4}}{m \sqrt{L_{\nabla J}}}$


Figure 1: High level procedure for achieving inverse reinforcement learning. The forward learning process is represented by a stochastic gradient descent (SGD), and the inverse learner incorporates sequential SGD evaluations $\theta_{k}$ into its PSGLD algorithm to reconstruct $J$. The PSGLD algorithm reconstructs $J$ by approximately sampling from the Gibbs measure $\pi_{\infty}$ (then taking the log-sample density). We measure the proximity of the PSGLD algorithm to $\pi_{\infty}$ by $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right)$, the 2 -Wasserstein distance between the sample law of $\alpha_{k}$ and the measure $\pi_{\infty}$. We control this distance by bounding it by $\mathcal{W}_{2}\left(\pi_{k}, \nu_{k \epsilon}\right)+\mathcal{W}_{2}\left(\nu_{k \epsilon}, \pi_{\infty}\right)$, where $\nu_{k \epsilon}$ is the law of $\alpha(t)$ at time $t=k \epsilon$.
3.2.1. Discussion of assumptions. A3.1-A3.4 are equivalent to those taken for the objective function in [22]. A3.1 is widely used in the literature on non-convex optimization and sampling. A3.2 can be enforced through weight decay regularization [18], see section 4 of [22] for more details. A3.3 is a standard assumption for stochastic gradient evaluations. A3.4-A3.6 admit a wide range of probability density functions, including Gaussians. We note that for A3.8 to be satisfied in practice, the inverse learner must have some knowledge of feasible ranges for Hessian bound $L_{\nabla J}$ and dissipativity constant $m$; once these ranges are known then $\epsilon$ can be taken small enough and $\beta$ large enough so that (ii) and (iii) are satisfied. Notice that the feasible range for $\eta$ can always be satisfied; the SGD process (2.8) optimizing cost function $J$ with step $\hat{\eta} \geq\left(1 \wedge \frac{m}{4 L_{\nabla J}^{2}}\right)$ is equivalent to another SGD with step $\eta<\frac{m}{4 L_{\nabla J}^{2}}$ which optimizes $\frac{\eta}{\hat{\eta}} J$. So assuming $\eta$ which satisfies A3.8 we can sample from $\pi_{\infty} \propto \exp \left(-\frac{\eta}{\hat{\eta}} \beta J\right)$, from which $J$ can be recovered since the scale $\frac{\eta}{\hat{\eta}} \beta$ disappears upon MCMC sample measure normalization.

### 3.3. Main Result. Finite-Sample Bound. Letting

$$
\pi_{k}:=\operatorname{Law}\left(\alpha_{k}\right), \quad \nu_{k \epsilon}:=\operatorname{Law}(\alpha(k \epsilon))
$$

be the respective measures of the sampling density produced by iterates $\alpha_{k}$ (2.9) and the continuous time diffusion $\alpha(t)$ (2.12) at time $t=k \epsilon$, we may bound

$$
\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right) \leq \mathcal{W}_{2}\left(\pi_{k}, \nu_{k \epsilon}\right)+\mathcal{W}_{2}\left(\nu_{k \epsilon}, \pi_{\infty}\right)
$$

Figure 1 shows the high level procedure for achieving inverse reinforcement learning. The forward learner is represented by a stochastic gradient descent (SGD) process
which optimizes $J$. The PSGLD algorithm takes in sequential SGD evaluations $\theta_{k}$ and produces samples $\alpha_{k}$ which approximately sample from the Gibbs measure $\pi_{\infty}$, allowing for reconstruction of $J$ by taking the log-sample density. We measure this approximation by the distance $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right)$, which can be bounded by introducing the intermediate continuous-time diffusion (2.12), since PSGLD (2.9) is an approximate discretization of (2.12) and (2.12) has Gibbs measure (2.2) as its stationary measure.

We present our Wasserstein bound in a way that explicitly depends on a hyperparameter $\delta$, e.g., $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right) \leq f(\delta)$ for some function $f$ which is monotonically increasing and has $\lim _{\delta \rightarrow 0} f(\delta)=0$. Both the Wasserstein bound and certain algorithmic parameters have a functional dependence on $\delta$ : for decreasing $\delta$ (decreasing $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right)$ ), we require e.g., increasing the algorithmic iterations and decreasing the step size. Specifically, our main result states that for any arbitrarily small $f(\delta)$, we can choose the step size $\epsilon$ small enough, algorithmic iterations $k$ large enough, kernel scale parameter $\Delta$ small enough, and sampling distribution scale parameter $\gamma$ small enough, such that $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right) \leq f(\delta)$. Next these qualitative parameter specifications are shown explicitly, as functions of control hyperparameter $\delta$. Then our main bound on $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right)$ is presented in Theorem 3.9.
3.3.1. Algorithmic Parameter Specifications. Here we show the dependence of algorithmic parameters on the hyperparameter $\delta$, which controls the main Wasserstein bound presented in Theorem 3.9. $\delta$ acts as a one-dimensional "knob" that can be turned, which reveals the step size $\epsilon$, iteration number $k$, etc., required to achieve a 2 -Wasserstein bound proportional to $\delta$. The main idea is that Theorem 3.9 presents a (monotonically increasing) function $f(\delta)$, with $\lim _{\delta \rightarrow 0} f(\delta)=0$, such that for any $\delta>0$ we can take algorithmic parameters as follows to obtain $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right) \leq f(\delta)$.

Step Size:

$$
\begin{equation*}
\epsilon \leq\left(\frac{\delta}{\log \left(\frac{1}{\delta}\right)}\right)^{2} \wedge 1 \tag{3.2}
\end{equation*}
$$

Algorithmic Iterations:

$$
\begin{equation*}
k \epsilon \geq \beta c_{L S} \log \left(\frac{1}{\delta}\right) \tag{3.3}
\end{equation*}
$$

where $c_{L S}$ is the logarithmic-Sobolev constant of diffusion (2.12), explicitly bounded in (5.5).

Kernel Scale: Recalling, for general $\alpha \in \mathbb{R}_{+}, K_{\alpha}(\cdot)=\frac{1}{\alpha^{N}} K(\dot{\bar{\alpha}})$, define $\hat{K}_{\alpha}:=$ $\sup _{x \in \mathbb{R}^{N}} K_{\alpha}(x)$. Also let $K^{-1}$ denote the inverse of $K$ and $K^{-2}$ denote the inverse of $K^{2}$, both mapping to the non-negative orthant, i.e., for $x \in \mathbb{R}, K^{-1}(x):=\left\{y \in \mathbb{R}_{+}^{N}\right.$ : $K(y)=x\}, \quad K^{-2}(x):=\left\{y \in \mathbb{R}_{+}^{N}: K^{2}(y)=x\right\}$ where $\mathbb{R}_{+}^{N}$ is the set of $N$-dimensional vectors with all non-negative elements. This definition is without loss of generality, since $K$ is chosen to be symmetric by (2.11). Then take

$$
\begin{equation*}
\Delta \leq \inf _{x \in\left[\epsilon, \hat{K}_{\epsilon}\right]} \frac{K^{-1}\left(\frac{\hat{K}_{1} \sqrt{2 \pi}}{2 \epsilon} e^{x^{2} / 2}\right)}{K^{-2}\left(x \epsilon^{2 N}\right)} \tag{3.4}
\end{equation*}
$$

Sampling Distribution Scale: Choose the base sampling distribution $\pi_{0}$ such that $\bar{\pi}_{0}:=\sup _{x} \pi_{0}(x)=1$, and sampling distribution scale parameter $\gamma$ as

$$
\begin{gather*}
\gamma \in\left[\epsilon^{2}, \epsilon^{3 / 2}\right]  \tag{3.5}\\
9
\end{gather*}
$$

### 3.3.2. Main 2-Wasserstein Bound.

Theorem 3.9 (Finite-Sample 2-Wasserstein Bound). Consider the PSGLD Algorithm 2.2 with iterates $\alpha_{k} \in \mathbb{R}^{N}$. Recall $c_{L S}$ is the logarithmic-Sobolev constant for Itô diffusion (2.12), bounded in (5.5). For any

$$
\begin{equation*}
\delta \in\left[0, \exp \left(-\frac{1}{\beta c_{L S}}\right)\right] \tag{3.6}
\end{equation*}
$$

choose step size $\epsilon$ according to (3.2), number of iterations $k$ according to (3.3), kernel scale $\Delta$ according to (3.4), and sampling distribution $\pi_{0, \gamma}$ with $\gamma$ satisfying (3.5). Then, under assumptions (A3.1)-(A3.8), the 2-Wasserstein distance between the distribution $\pi_{k}$, generated by the PSGLD algorithm, and the Gibbs measure $\pi_{\infty}$ (2.2), satisfies:

$$
\begin{equation*}
\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right) \leq \delta\left[C_{4}+\sqrt{2 c_{L S} C_{3}}\right]+\delta \sqrt{10 c_{L S} N \log (1 / \delta)} \tag{3.7}
\end{equation*}
$$

$C_{3}, C_{4}$ are constants dependent on structural specifications of $J$ and the process (2.9), provided explicitly in supplementary material 7.1. $c_{L S}$ is the logarithmic-Sobolev constant bounded explicitly in Proposition 5.3.

Bound Discussion: For any $\alpha>0, \delta \sqrt{\alpha \log (1 / \delta)}$ is monotonically increasing in $\delta$ for $\delta \in(0,0.607)$ and

$$
\lim _{\delta \rightarrow 0} \delta \sqrt{\alpha \log (1 / \delta)}=0
$$

So, $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right)$ is monotonically increasing in $\delta$ for $\delta \in(0,0.607)$ and

$$
\lim _{\delta \rightarrow 0} \mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right)=0
$$

Thus, Theorem 3.9 asserts that, through hyperparameter $\delta$, we can control the number of iterations $k$ as (3.3), step size $\epsilon$ as (3.2), kernel scale $\Delta$ as (3.4) and sampling distribution scale $\gamma$ as (3.5), such that the PSGLD algorithm (2.9) is within any arbitrarily small desired 2-Wasserstein distance (3.7) to the Gibbs distribution (2.2). Here $\delta$ acts as a precision parameter; smaller $\delta$ yields a tighter approximation (3.7) at the expense of larger number of iterations $k$ and smaller step size $\epsilon$, kernel scale $\Delta$ and sampling distribution scale $\gamma$.

Recalling $\pi_{\infty}(\alpha) \propto \exp (-\beta J(\alpha))$, the cost function $J$ can be approximately reconstructed as the logarithm of sample density produced by $\alpha_{k}$. This reconstruction approaches the true cost function $J$ as $\delta \rightarrow 0$. This result generalizes the nonasymptotic bound obtained in [22] (equation 3.3) to our passive stochastic gradient Langevin dynamics algorithm.

Parameter Specifications Discussion: Observe the parameter specifications (3.2) (3.5) necessary for acheiving a given Wasserstein bound (3.7). Specifications (3.2) and (3.3) are intuitive; as we decrease the step size $\epsilon$ we should decrease the discretization error between algorithm (2.9) and continuous diffusion (2.12), and as we increase the iterations $k$ we will decrease the distance from the diffusion (2.12) to its stationary measure $\pi_{\infty}$.

However, the algorithm (2.9) is not an exact discretization of diffusion (2.12), as it has a gradient term governed by the external SGD process (2.8). The weighting kernel $K$ is introduced to control for biases in this SGD-evaluated gradient, but for any non-zero variance of $K$ there will still be biased gradient evaluations entering the algorithm (2.9) which prevent it from converging to $\pi_{\infty}$. To minimize these,
the kernel scale $\Delta$ can be reduced; however note that this should come at the cost of increasing the time needed to reach a specified Wasserstein bound, since "useful" gradient information will be integrated into the passive algorithm less often. We see this as an unavoidable tradeoff, one which necessitates taking $\Delta$ as (3.4), but which has not been fully quantified in this work.

Note that we also require decreasing the sampling distribution scale $\gamma$ to obtain a tighter Wasserstein bound. This has arisen as a quantitative necessity in obtaining bounds in Lemma 5.1 and Lemma 5.2 (which are key developments in the proof of Theorem 3.9). The intuition is as follows: this specification allows us to control $\mathbb{E}\left\|\pi_{0, \gamma}\left(\alpha_{k}\right)\right\|^{2}$ and $\mathbb{E}\left\|\nabla \pi_{0, \gamma}\left(\alpha_{k}\right)\right\|^{2}$, such that the influence of $\pi_{0, \gamma}\left(\alpha_{k}\right)$ in the algorithm (2.9) does not outweigh that of $K_{\Delta}\left(\theta_{k}, \alpha_{k}\right) \hat{\nabla} J\left(\theta_{k}\right)$; (as in the previous paragraph explanation) as $\Delta$ gets smaller, $K_{\Delta}\left(\theta_{k}, \alpha_{k}\right) \hat{\nabla} J\left(\theta_{k}\right)$ contributes less often (but is more accurate when it does), and the contribution of $\pi_{0, \gamma}\left(\alpha_{k}\right)$ should balance this. Notice that we also must choose the base distribution $\pi_{0}$ wide enough (such that $\bar{\pi}_{0}=1$ ), and scale parameter not too small (lower bounded by $\epsilon^{2}$ ), so that there is always some non-zero probability of sampling from any point in the domain, allowing for sufficient exploration.
3.4. Example. Adaptive MDP Inverse Reinforcement Learning. As discussed in [16], the PSGLD algorithm can be applied to a variety of settings including constrained MDPS, adaptive Bayesian inference, and logistic regression classification. Here we illustrate how the PSGLD algorithm (Alg 2.2), and finite-sample guarantees of Theorem 3.9, interface with a well-known practical reinforcement learning (RL) algorithm. We first outline the details of the forward RL procedure, then show how our PSGLD algorithm can be used to achieve inverse RL with nonasymptotic accuracy guarantees.
3.4.1. Forward Reinforcement Learning: Policy Gradient. Here we present one canonical and well-established algorithm for reinforcement learning (RL), the Reinforce algorithm. Consider the following notation

- discount factor $\gamma \in(0,1)$, discrete-time index $t \in \mathbb{N}$
- states $s_{t} \in \mathcal{S}$, actions $a_{t} \in \mathcal{A}$, cost $c_{t}=C\left(s_{t}, a_{t}\right), C: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}_{+}$
- state transition probabilities $P_{s, s^{\prime}}^{a}=P\left(s_{t+1}=s^{\prime} \mid s_{t}=s, a_{t}=a\right)$
- initial state probability distribution $\rho_{0}(s)=\mathbb{P}\left(s_{0}=s\right)$
- policy function $\pi(s, a ; \theta)=\mathbb{P}\left[a_{t}=a \mid s_{t}=s ; \theta\right]$ under policy parameter $\theta \in \mathbb{R}^{d}$ Now let

$$
J(\theta):=\mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} c_{t}\right]: \mathbb{R}^{d} \rightarrow \mathbb{R}_{+}
$$

be the expected cost objective, where the expectation is taken with respect to the probability distribution of trajectories $\left(s_{0}, a_{0}, \ldots, s_{t}, a_{t}, \ldots\right)$ induced by $\rho_{0}, P_{\left(s, s^{\prime}\right)}^{a}$, and $\pi(s, a ; \theta)$. The reinforcement learning goal is to find a policy parameter $\theta$ such that $J(\theta)$ is minimized. This minima can be achieved by sequentially updating $\theta$ via stochastic gradient descent on $J(\theta)$. This is the methodology of the canonical Reinforce algorithm, which has sequential policy parameter updates given by

$$
\begin{equation*}
\theta_{k+1}=\theta_{k}-\eta \sum_{t=0}^{T}\left[\gamma^{t} \nabla_{\theta} \log \pi\left(s_{t}, a_{t} ; \theta\right) \sum_{k=t}^{T} \gamma^{k-t} r_{k}\right]=\theta_{k}-\eta \hat{\nabla} J(\theta) \tag{3.8}
\end{equation*}
$$

where, crucially, $\hat{\nabla} J(\theta)$ is an unbiased estimate of $\nabla J(\theta)$ by the Policy Gradient Theorem [26].
3.4.2. Adaptive Inverse Reinforcement Learning: PSGLD. Observe that the RL algorithm 3.8 works by employing a stochastic gradient descent in the space of policies. This suggests that we can employ our PSGLD algorithm, taking as input the sequential policy evaluations and outputting the expected cost function.

Suppose we observe an agent performing policy gradient RL by enacting sequential sample paths $\left\{\left(s_{0}, a_{0}, \ldots, s_{T}, a_{T}\right)_{k} \sim \pi\left(\cdot, \cdot ; \theta_{k}\right)\right\}_{k=1}^{K}$, and updating $\theta_{k}$ according to Algorithm 3.8. Notice that by observing a sample path $\left(s_{0}, a_{0}, \ldots, s_{T}, a_{T}\right) \sim \pi(\cdot, \cdot ; \theta)$ we can obtain an unbiased estimate $\hat{\theta}$ of $\theta$ by simply taking the empirical distribution of observed state-action pairs. Then form $\tilde{\nabla} J\left(\theta_{k}\right)=\frac{\hat{\theta}_{k}-\hat{\theta}_{k-1}}{\eta}$, and notice that

$$
\mathbb{E}\left[\tilde{\nabla} J\left(\theta_{k}\right) \mid \theta_{k}, \theta_{k-1}\right]=\hat{\nabla} J\left(\theta_{k}\right)
$$

i.e., $\tilde{\nabla} J\left(\theta_{k}\right)$ is a second order stochastic gradient estimate; it is an unbiased estimate of the stochastic gradient $\hat{\nabla} J\left(\theta_{k}\right)$, and $\hat{\nabla} J\left(\theta_{k}\right)$ is fully determined by the (random) evaluations $\theta_{k}, \theta_{k+1}$. So we have

$$
\mathbb{E}\left[\tilde{\nabla} J\left(\theta_{k}\right)\right]=\mathbb{E}\left[\mathbb{E}\left[\tilde{\nabla} J\left(\theta_{k}\right) \mid \theta_{k}, \theta_{k+1}\right]\right]=\nabla J\left(\theta_{k}\right)
$$

and thus $\tilde{\nabla} J\left(\theta_{k}\right)$ can be utilized in our passive stochastic gradient Langevin dynamics algorithm (Alg 2.2) since it is an unbiased estimate of $\nabla J\left(\theta_{k}\right)$. Thus, Theorem 3.9 applies and we can control the non-asymptotic Wasserstein distance arbitrarily. Fixing any $\delta>0$, Algorithm 2.2, with parameters specified by (3.2) and with $\theta_{k}$ replaced by $\hat{\theta}_{k}$ above, produces iterates $\alpha_{l}, l \geq k$, that sample from a measure $\pi$ with

$$
\left.\mathcal{W}_{2}\left(\pi . \pi_{\infty}\right) \leq \mathcal{O}(\delta+\delta \sqrt{\log (1 / \delta})\right), \quad \pi_{\infty}(\theta) \propto \exp (-\beta J(\theta))
$$

Then $J$ can be approximately recovered by taking the logarithm of MCMC sample density.

Note that traditional IRL methods aim to reconstruct $C(s, a)$, rather than $J(\theta)$, given optimal policy demonstrations. In our case $C(s, a)$ can be recovered up to a constant multiplicative factor once $J(\theta)$ and the MDP transition dynamics are known, since $J(\theta)$ is the expectation of $C(s, a)$ with respect to the stationary measure induced by the policy $\pi(\cdot, \cdot ; \theta)$ and the dynamics $P_{s, s^{\prime}}^{a}$. Furthermore, in contrast to traditional methods [32], [21], we operate in the transient regime where the observed agent is in the process of learning an optimal policy.
4. Preliminaries for Bound Proof. Since finite-sample bounds for stochastic diffusions may not be widely known in the stochastic control community, this short section briefly summarizes the main tools required for the proof of Theorem 3.9.
4.1. Infinitesimal Generator. Let $X_{t}$ be an $\mathbb{R}^{N}$-valued diffusion defined by the stochastic differential equation

$$
\begin{equation*}
d X_{t}=b\left(X_{t}\right) d t+\sigma\left(X_{t}\right) d W(t), \quad X_{0}=x \in \mathbb{R}^{N} \tag{4.1}
\end{equation*}
$$

where $b: \mathbb{R}^{N} \rightarrow \mathbb{R}^{N}$ is the drift function, $\sigma: \mathbb{R}^{N} \rightarrow \mathbb{R}$ is the volatility function, and $W(t)$ is standard $N$-dimensional Brownian motion. Fixing a point $x \in \mathbb{R}^{N}$, let $P^{x}$ denote the law of $X_{t}$ given $X_{0}=x$, and $\mathbb{E}^{x}$ denote expectation with respect to $P^{x}$. Let $\mathcal{L}$ be the infinitesimal generator of $X_{t}$, defined by its action on compactly-supported $C^{2}$ functions $f: \mathbb{R}^{N} \rightarrow \mathbb{R}$ in domain $\mathcal{D}(\mathcal{L}) \subseteq C^{2}\left(\mathbb{R}^{N}\right)$, as

$$
\begin{equation*}
\mathcal{L} f(x)=\lim _{t \downarrow 0} \frac{\mathbb{E}^{x}\left[f\left(X_{t}\right)-f(x)\right]}{t}=\sum_{i=1}^{n} b_{i}(x) \frac{\partial f}{\partial x_{i}}(x)+\frac{1}{2} \sum_{i, j} \sigma^{2}(x) \frac{\partial^{2} f}{\partial x_{i} \partial x_{j}}(x) \tag{4.2}
\end{equation*}
$$

where $b_{i}(x)$ is the $i$ 'th element of $b(x) \in \mathbb{R}^{N}$. Thus $\mathcal{L}$ is an operator acting on $f \in C^{2}\left(\mathbb{R}^{N}\right)$ as

$$
\mathcal{L} f=\frac{1}{2} \sigma^{2} \Delta f+\langle b, \nabla f\rangle
$$

where $\Delta:=\nabla \cdot \nabla$ denotes the standard Laplacian operator. We say $\pi$ is an invariant probability measure w.r.t $\mathcal{L}$ if and only if $\int_{\mathbb{R}^{N}} \mathcal{L} g d \pi=0$ for all $g \in \mathcal{D}(\mathcal{L})$.

In this work we consider the diffusion which solves the stochastic differential equation (2.12), which has:

$$
b(x)=-\frac{\beta}{2} \pi_{0, \gamma}^{2}(x) \nabla J(x)-\pi_{0, \gamma}(x) \nabla \pi_{0, \gamma}(x), \quad \sigma(x)=\pi_{0, \gamma}(x)
$$

Thus, the infinitesimal generator of our diffusion process (2.12) is given as

$$
\begin{equation*}
\mathcal{L} f=\frac{1}{2} \pi_{0, \gamma}^{2} \Delta f-\frac{\beta}{2} \pi_{0, \gamma}^{2}\langle\nabla J, \nabla f\rangle-\pi_{0, \gamma}\left\langle\nabla \pi_{0, \gamma}, \nabla f\right\rangle \tag{4.3}
\end{equation*}
$$

and note that by assumptions (A2),(A6) and by Theorem 2.5 of [15], we have that (2.12) admits a unique strong solution.
4.2. Poincaré and logarithmic Sobolev inequalities. Considering a general infinitesimal generator $\mathcal{L}$, with stationary measure $\pi$, we can define the Dirichlet form

$$
\mathcal{E}(g):=-\int_{\mathbb{R}^{N}} g \mathcal{L} g d \pi
$$

and the spectral gap $\lambda$ as

$$
\begin{equation*}
\lambda:=\inf \left\{\frac{\int_{\mathbb{R}^{N}} \mathcal{E}(g) d \pi}{\int_{\mathbb{R}^{N}} g^{2} d \pi}: g \in C^{1}\left(\mathbb{R}^{N}\right) \cap L^{2}(\pi), g \neq 0, \int_{\mathbb{R}^{N}} g d \pi_{\infty}=0\right\} \tag{4.4}
\end{equation*}
$$

Let us consider a Markov process $X_{t}$ with unique invariant distribution $\pi$ and infinitesimal generator $\mathcal{L}$. We say that $\pi$ satisfies a Poincaré (spectral gap) inequality with constant $c$ if

$$
\begin{equation*}
\chi^{2}(\mu \| \pi) \leq c \mathcal{E}\left(\sqrt{\frac{d \mu}{d \pi}}\right) \tag{4.5}
\end{equation*}
$$

for all probability measures $\mu \ll \pi$ ( $\mu$ absolutely continuous w.r.t $\pi$ ), where $\chi^{2}(\mu \| \pi):=\left\|\frac{d \mu}{d \pi}-1\right\|_{L^{2}(\pi)}^{2}$ is the $\chi^{2}$ divergence between $\mu$ and $\pi$. If (4.5) is satisfied for some $c$, then we have $\frac{1}{c} \leq \lambda$ where $\lambda$ is the spectral gap given in (4.4). In particular, letting $c_{P}$ denote the Poincaré constant, given as the smallest $c$ such that (4.5) holds,

$$
c_{P}=\inf \left\{c: \chi^{2}(\mu \| \pi) \leq c \mathcal{E}\left(\sqrt{\frac{d \mu}{d \pi}}\right) \forall \mu \ll \pi\right\}
$$

where $\ll$ denotes absolute continuity, then we have $\frac{1}{c_{P}}=\lambda$, and the eigenspectrum of $-\mathcal{L}$ is contained in $\{0\} \cup\left[\frac{1}{c P}, \infty\right)$.

We say that $\pi$ satisfies a logarithmic Sobolev inequality with constant $c$ if

$$
D(\mu \| \pi) \leq 2 c \mathcal{E}\left(\sqrt{\frac{d \mu}{d \pi}}\right)
$$

for all $\mu \ll \pi$, where $D(\mu \| \pi)=\int d \mu \log \frac{d \mu}{d \pi}$ is the Kullback-Leibler divergence.
One of the main efforts of this work will be to show that the diffusion (2.12) satisfies a log-Sobolev inequality; this then allows us to utilize several useful properties in the non-asymptotic analysis. Specifically, letting $\{X(t)\}_{t \geq 0}$ be a Markov process with stationary distribution $\pi$ and Dirichlet form $\mathcal{E}$, then we have:

Lemma 4.1 (Exponential decay of entropy [2], Th. 5.2.1). Let $\mu_{t}:=\operatorname{Law}(X(t))$. If $\pi$ satisfies a logarithmic-Sobolev inequality with constant $c$, then

$$
\begin{equation*}
D\left(\mu_{t} \| \pi\right) \leq D\left(\mu_{0} \| \pi\right) e^{-2 t / c} \tag{4.6}
\end{equation*}
$$

Lemma 4.2 (Otto-Villani theorem [2], Th. 9.6.1). If $\pi$ satisfies a logarithmicSobolev inequality with constant $c$, then, for any $\mu \ll \pi$

$$
\begin{equation*}
\mathcal{W}_{2}(\mu, \pi) \leq \sqrt{2 c D(\mu \| \pi)} \tag{4.7}
\end{equation*}
$$

The following two results give sufficient conditions for a measure $\pi$ to satisfy Poincare and logarithmic-Sobolev inequalities, using Lyapunov function criteria.

Proposition 4.3 (Bakry 2008 [1]). Let $\pi(d x)=\exp (-H(x)) d x$ be a probability measure on $\mathbb{R}^{N}$ with $H \in C^{2}\left(\mathbb{R}^{N}\right)$ and lower bounded. Let $\mathcal{L}$ be the infinitesimal generator of a Markov process with stationary measure $\pi$. Suppose there exist constants $\kappa_{0}, \zeta_{0}>0, r \geq 0$ and a $C^{2}$ function $V: \mathbb{R}^{N} \rightarrow[1, \infty)$ such that

$$
\begin{equation*}
\frac{\mathcal{L} V(w)}{V(w)} \leq-\zeta_{0}+\kappa_{0} \mathbf{1}\{\|w\| \leq r\} \tag{4.8}
\end{equation*}
$$

Then $\pi$ satisfies a Poincaré inequality with constant

$$
\begin{equation*}
c_{P} \leq \frac{1}{\zeta_{0}}\left(1+C \kappa_{0} r^{2} \exp \left(O_{r}(H)\right)\right) \tag{4.9}
\end{equation*}
$$

where $C>0$ is a universal constant and $O_{r}(H):=\max _{\|w\| \leq r} H(w)-\min _{\|w\| \leq r} H(w)$
Proposition 4.4 (Cattiaux et. al. (2010) [5]). Let $\pi(d x)=\exp (-H(x)) d x$ be a probability measure on $\mathbb{R}^{N}$ with $H \in C^{2}\left(\mathbb{R}^{N}\right)$ and lower bounded. Let $\mathcal{L}$ be the infinitesimal generator of a Markov process with stationary measure $\pi$. Suppose the following conditions hold:

1. There exist constants $\kappa, \gamma>0$ and a $C^{2}$ function $V: \mathbb{R}^{d} \rightarrow[1, \infty)$ such that

$$
\begin{equation*}
\frac{\mathcal{L} V(w)}{V(w)} \leq \kappa-\gamma\|w\|^{2} \forall w \in \mathbb{R}^{d} \tag{4.10}
\end{equation*}
$$

2. $\pi_{\infty}$ satisfies a Poincaré inequality with constant $c_{P}$.
3. There exists some constant $K \geq 0$, such that $\nabla^{2} H \succcurlyeq-K I_{d}$

Let $Z_{1}, Z_{2}$ be defined, for some $\zeta>0$ as

$$
\begin{equation*}
Z_{1}=\frac{2}{\gamma}\left(\frac{1}{\zeta}+\frac{K}{2}\right)+\zeta, \quad Z_{2}=\frac{2}{\gamma}\left(\frac{1}{\zeta}+\frac{K}{2}\right)\left(\kappa+\gamma \int_{\mathbb{R}^{N}}\|w\|^{2} \pi(d w)\right) \tag{4.11}
\end{equation*}
$$

Then $\pi$ satisfies a logarithmic Sobolev inequality with constant $c_{L S}=Z_{1}+\left(Z_{2}+2\right) c_{P}$.
We will be interested in showing that the invariant (Gibbs) measure $\pi_{\infty}(2.2)$ of our particular diffusion (2.12) satisfies a log-Sobolev inequality, so that we can apply Lemmas 4.1 and 4.2 to obtain exponential convergence of $\mathcal{W}_{2}\left(\nu_{k \epsilon}, \pi_{\infty}\right)$. To show the
log-Sobolev inequality holds, we will show that the conditions of Proposition 4.4 hold, using Proposition 4.3 as an intermediate step. More details on this procedure will be outlined in Section 5 .

The following result is unrelated to Poincaré and log-Sobolev inequalities, but gives a way to bound a general Wasserstein distance once a KL-divergence is known. We will utilize this in the bound on $\mathcal{W}_{2}\left(\pi_{k}, \nu_{k \epsilon}\right)$.

Corollary 4.5 (Bolley and Villani 2005 [3] Cor. 2.3). For any two Borel probability measures $\mu, \nu$ on $\mathbb{R}^{N}$,
$\mathcal{W}_{2}(\mu, \nu) \leq 2 \inf _{\lambda>0}\left(\frac{1}{\lambda}\left(\frac{3}{2}+\log \int_{\mathbb{R}^{N}} e^{\lambda\|w\|^{2}} \nu(d w)\right)\right)^{1 / 2}\left[\sqrt{D(\mu \| \nu)}+\left(\frac{D(\mu \| \nu)}{2}\right)^{1 / 4}\right]$
5. Proof of Main Result (Theorem 3.9). Outline. Here we provide the proof structure for our bound on $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right)$, provided as (3.7) in Theorem 3.9 (complete proofs can be found in the supplementary material). The block diagram in Figure 2 displays the relations between our main supporting results in the proof of Theorem 3.9. The high level proof structure is as follows: We bound $\mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right) \leq \mathcal{W}_{2}\left(\pi_{k}, \nu_{k \epsilon}\right)+\mathcal{W}_{2}\left(\nu_{k \epsilon}, \pi_{\infty}\right)$, i.e., we first control the discretization error between passive algorithm 2.2 and diffusion 2.12 , then control the convergence rate of this diffusion to its stationary distribution $\pi_{\infty}$.

In order to achieve a useful bound on the former, scaling as $\mathcal{O}(k \epsilon \sqrt{\epsilon})$, we employ a Girsanov change of measure (controlling the KL-divergence), given as Lemma 5.2, followed by Corollary 4.5 (to relate back to 2-Wasserstein distance), as in [22]. This procedure relies crucially on the exponential integrability of the diffusion (2.12), which we prove as Lemma 7.3. To handle lack of measure absolute continuity, as discussed below, we must introduce an intermediate process (with law $\gamma_{k \epsilon}$ ), perform the above procedure on the error between $\gamma_{k \epsilon}$ and $\nu_{k \epsilon}$, then bound $\mathcal{W}_{2}\left(\pi_{k}, \nu_{k \epsilon}\right) \leq \mathcal{W}_{2}\left(\pi_{k}, \gamma_{k \epsilon}\right)+$ $\mathcal{W}_{2}\left(\gamma_{k \epsilon}, \nu_{k \epsilon}\right)$. The result providing a bound on $\mathcal{W}_{2}\left(\pi_{k}, \gamma_{k \epsilon}\right)$, completing this approach, is given as Lemma 5.1.

To bound $\mathcal{W}_{2}\left(\nu_{k \epsilon}, \pi_{\infty}\right)$, we first show that $\pi_{\infty}$ satisfies a logarithmic-Sobolev inequality, by satisfying the conditions of Proposition 4.4 [5]. This result is given as Proposition 5.3. We then apply exponential decay of entropy [2], given as Lemma 4.1, and the Otto-Villani Theorem [1], given as 4.2. This procedure provides an exponentially decaying bound on $\mathcal{W}_{2}\left(\nu_{k \epsilon}, \pi_{\infty}\right)$.
5.1. 2-Wasserstein Bound for Diffusion Approximation. Here we obtain a bound on $\mathcal{W}_{2}\left(\pi_{k}, \nu_{k \epsilon}\right)$. Consider the continuous-time interpolation of the process (2.9):

$$
\begin{align*}
\bar{\alpha}(t)=\alpha_{0} & -\int_{0}^{t}\left[K_{\Delta}\left(\theta_{\bar{s}}, \bar{\alpha}(\bar{s})\right) \frac{\beta}{2} \hat{\nabla} J\left(\theta_{\bar{s}}\right)+\nabla \pi_{0, \gamma}(\bar{\alpha}(\bar{s}))\right] \pi_{0, \gamma}(\bar{\alpha}(\bar{s})) d s  \tag{5.1}\\
& +\int_{0}^{t} \pi_{0, \gamma}(\bar{\alpha}(\bar{s})) d W(s)
\end{align*}
$$

where $\bar{s}=\lfloor s / \epsilon\rfloor \epsilon$, and $\theta_{\bar{s}}:=\theta_{k}$ for $k=\lfloor s / \epsilon\rfloor$. Note that, for each $k, \bar{\alpha}(k \epsilon)$ and $\alpha_{k}$ have the same probability law $\pi_{k}$. We aim to relate this process to the diffusion (2.12) through a Girsanov change of measure; but the process (5.1) is not Markovian and is therefore not an Itô diffusion. However, by the result of [11], the process $\bar{\alpha}(t)$ has the


Figure 2: Theorem 3.9 proof structure. First the 2-Wasserstein distance between discretetime algorithm (2.9) (with measure $\pi_{k}$ ) and continuous-time diffusion (2.12) (with measure $\nu_{k \epsilon}$ ) is bounded. We must introduce an intermediate process (with law $\gamma_{k \epsilon}$ ). Lemma 5.1 bounds the Wasserstein distance between $\pi_{k}$ and $\gamma_{k \epsilon}$. Lemma 5.2 bounds the KL-divergence between $\pi_{k}$ and $\gamma_{k \epsilon}$. Corollary 4.5 is then used, along with Lemma 7.3 to relate this KL bound to a 2 -Wasserstein bound. Proposition 5.3 is the key tool in bounding $\mathcal{W}_{2}\left(\nu_{k \epsilon}, \pi_{\infty}\right)$, establishing that $\pi_{\infty}$ satisfies a log-Sobolev inequality. We then employ exponential decay of entropy (Lemma 4.1) and the Otto-Villani Theorem (Lemma 4.2) to obtain exponential decay of $\mathcal{W}_{2}\left(\nu_{k \epsilon}, \pi_{\infty}\right)$.
same one-time marginals as the Itô process $Y(t)$, where

$$
\left.\begin{array}{rl}
(5.2) & Y(t) \tag{5.2}
\end{array}=\alpha_{0}-\int_{0}^{t} g_{s}\left(\theta_{\bar{s}}, Y(s)\right) d s+\int_{0}^{t} \mathbb{E}\left[\pi_{0, \gamma}(\bar{\alpha}(\bar{s})) \mid \bar{\alpha}(s)=Y(s)\right] d W(s)\right] .
$$

i.e., $\operatorname{Law}(Y(k \epsilon))=\pi_{k} \forall k \in \mathbb{N}$, and it is apparent that (5.2) is Markovian. However, we cannot apply Girsanov's formula to relate (5.2) and (2.12) because the volatility functions are different; so the measures $\pi_{k}$ and $\nu_{k \epsilon}$ are not absolutely continuous.

To solve this, we introduce the intermediate process

$$
\begin{equation*}
X(t)=\alpha_{0}-\int_{0}^{t} \hat{g}_{s}\left(\theta_{\bar{s}}, X(s)\right) d s+\int_{0}^{t} \pi_{0, \gamma}(X(s)) d W(s) \tag{5.3}
\end{equation*}
$$

where $\hat{g}_{s}\left(\theta_{\bar{s}}, X(s)\right)=\left(K_{\Delta}\left(\theta_{\bar{s}}, X(s)\right) \frac{\beta}{2} \hat{\nabla} J\left(\theta_{\bar{s}}\right)+\nabla \pi_{0, \gamma}(X(s))\right) \pi_{0, \gamma}(X(s))$. Let $\gamma_{k \epsilon}$ denote the law of (5.3) at time $t=k \epsilon$. This process (5.3) is similar enough to (5.2) to allow a tractable bound on $\mathcal{W}_{2}\left(\pi_{k}, \gamma_{k \epsilon}\right)$, and since (5.3) has the same volatility function as (2.12) we can relate these two via Girsanov's formula to obtain a desirable bound on $\mathcal{W}_{2}\left(\gamma_{k \epsilon}, \nu_{k \epsilon}\right)$. Then we simply bound $\mathcal{W}_{2}\left(\pi_{k}, \nu_{k \epsilon}\right) \leq \mathcal{W}_{2}\left(\pi_{k}, \gamma_{k \epsilon}\right)+\mathcal{W}_{2}\left(\gamma_{k \epsilon}, \nu_{k \epsilon}\right)$.

The following Lemma provides a bound on $\mathcal{W}_{2}\left(\pi_{k}, \gamma_{k \epsilon}\right)$.
Lemma 5.1. Fixing the step size $\epsilon$ and time horizon $k \epsilon$, take the kernel scale parameter $\Delta$ small enough to satisfy (3.4), and sampling distribution scale parameter $\gamma$ small enough to satisfy (3.5). Then we have

$$
\mathcal{W}_{2}\left(\pi_{k}, \gamma_{k \epsilon}\right) \leq 6(k \epsilon) \epsilon \sqrt{12 C_{0}+3}+3 \sqrt{2(k \epsilon) \epsilon}
$$

where $C_{0}$ is a constant provided in supplementary material 7.1, $M_{\theta}$ is a uniform bound on $\mathbb{E}\left\|\theta_{k}\right\|^{2}$, see Lemma $7.4, L_{\nabla J}$ is the Lipchitz constant for $\nabla J$, see Assumption 3.1, $B=\|\nabla J(0)\|$, and $\zeta$ is the uniform noise variance bound in Assumption 3.3. Now, the following Lemma provides a bound on $D\left(\gamma_{k \epsilon} \| \nu_{k \epsilon}\right)$, the Kullback-Leibler (KL) divergence between measures $\gamma_{k \epsilon}$ and $\nu_{k \epsilon}$, using a similar Girsanov change of measure as presented in [22].

Lemma 5.2. Fixing the step size $\epsilon$ and time horizon $k \epsilon$, taking the kernel scale parameter $\Delta$ small enough to satisfy (3.4), and sampling distribution scale parameter $\gamma$ small enough to satisfy (3.5). Then we have:

$$
D\left(\gamma_{k \epsilon} \| \nu_{k \epsilon}\right) \leq(k \epsilon)^{3} \epsilon^{3}\left[4 \beta L_{\nabla J}^{2}\left(72 C_{0}+6 \sqrt{C_{0}}+18+\sqrt{2}\right)\right]+(k \epsilon) \epsilon\left(2 L_{J}^{2}+4 C_{0}\right)
$$

Now we relate this KL divergence to a Wasserstein distance through Corollary 4.5, with $\mu=\gamma_{k \epsilon}, \nu=\nu_{k \epsilon}, \lambda=1$. By Lemma 7.3 we have exponential integrability of the diffusion 2.12:

$$
\log \int_{\mathbb{R}^{N}} e^{\lambda\|w\|^{2}} \nu_{k \epsilon}(d w) \leq \kappa_{0}^{\gamma}+\left((\beta b+N) 2 \epsilon+2 I^{\prime}\right) k \epsilon
$$

Now, since $k \epsilon \geq 1$ by (3.6) and (3.3), and $\kappa_{0}^{\gamma} \leq \kappa_{0}$, we can bound
$\mathcal{W}_{2}\left(\gamma_{k \epsilon}, \nu_{k \epsilon}\right) \leq 2 \sqrt{\frac{3}{2}+\left(\kappa_{0}+(\beta b+N) 2 \epsilon+2 I^{\prime}\right) k \epsilon}\left(\sqrt{D\left(\gamma_{k \epsilon} \| \nu_{k \epsilon}\right)}+\left(D\left(\gamma_{k \epsilon} \| \nu_{k \epsilon}\right)\right)^{1 / 4}\right)$
Applying Lemma 5.2, we have:

$$
\mathcal{W}_{2}\left(\gamma_{k \epsilon}, \nu_{k \epsilon}\right) \leq 4 \sqrt{\frac{3}{2}+C_{1} k \epsilon} \sqrt{k \epsilon} \sqrt{\epsilon}\left(4 \sqrt{\beta L_{\nabla J}^{2} C_{2}}+2 \sqrt{2 L_{J}^{2}+4 C_{0}}\right)
$$

with $C_{0}, C_{1}, C_{2}$ in supplementary material 7.1. So finally,

$$
\begin{equation*}
\mathcal{W}_{2}\left(\pi_{k}, \nu_{k \epsilon}\right) \leq k \epsilon \sqrt{\epsilon}\left[6 \sqrt{12 C_{0}+3}+3 \sqrt{2}+4 \sqrt{3 / 2+C_{1}}\left(4 \sqrt{C_{2}}+2 \sqrt{2 L_{J}^{2}+4 C_{0}}\right)\right] \tag{5.4}
\end{equation*}
$$

So we achieve a discretization error bound $\mathcal{W}_{2}\left(\pi_{k}, \nu_{k \epsilon}\right)$ which scales as $\mathcal{O}(k \epsilon \sqrt{\epsilon})$. In fact, this is tighter than the bound obtained in [22], which scales as $\mathcal{O}\left(k \epsilon \epsilon^{1 / 4}\right)$. We represent this bound in terms of distinct units $k \epsilon$ and $\sqrt{\epsilon}$ (rather than $k \epsilon^{3 / 2}$ ) since in our final analysis we will take $k \epsilon$ large enough (but fixed), then $\epsilon$ small enough, so that $\mathcal{W}_{2}\left(\pi_{k}, \nu_{k \epsilon}\right)$ decreases arbitrarily. We will need to first take $k \epsilon$ large enough to control the diffusion (2.12) convergence to the Gibbs measure. The following presents this convergence in terms of exponentially decaying distance $\mathcal{W}_{2}\left(\nu_{k \epsilon}, \pi_{\infty}\right)$.
5.2. 2-Wasserstein Distance for Diffusion Convergence. Here we describe the method to bound $\mathcal{W}_{2}\left(\nu_{k \epsilon}, \pi_{\infty}\right)$. The strategy is as follows:
i) Show that $\pi_{\infty}$ satisfies a logarithmic-Sobolev inequality.
ii) Apply exponential decay of entropy, given as Lemma 4.1, with the relative entropy bound in Lemma 7.2, to derive a bound on $D\left(\nu_{k \epsilon} \| \pi_{\infty}\right)$
iii) Apply the Otto-Villani Theorem, given as Lemma 4.2, to relate this to a bound on $\mathcal{W}_{2}\left(\nu_{k \epsilon}, \pi_{\infty}\right)$.
We accomplish (i) in the following proposition, establishing that the Gibbs measure $\pi_{\infty}$ satisfies a log-Sobolev inequality:

Proposition 5.3. For $\beta$ satisfying Assumption 3.8, the Gibbs measure $\pi_{\infty}$ satisfies a logarithmic Sobolev inequality with constant $c_{L S}$ :

$$
\begin{equation*}
0 \leq c_{L S} \leq \frac{2 \beta L_{\nabla J}}{\gamma}+\frac{2}{\beta L_{\nabla J}}+\frac{1}{\lambda}\left(\frac{2 \beta L_{\nabla J}}{\gamma}\left(\kappa+\gamma\left(\kappa_{0}+\frac{(\beta b+N) \bar{\pi}_{0, \gamma}+2 I}{(m \beta) \bar{\pi}_{0, \gamma}}\right)\right)+2\right) \tag{5.5}
\end{equation*}
$$

where

$$
\begin{align*}
& \frac{1}{\lambda} \leq \frac{1}{2 \kappa}\left(1+\frac{4 C \kappa^{2}}{\gamma} \exp \left(\beta\left(\frac{\left(L_{\nabla J}+B\right) \kappa}{\gamma}+A+B\right)\right)\right)  \tag{5.6}\\
& \kappa=\left(\frac{1}{2} \beta m N+\beta m I\right)+\frac{1}{2}\left[\beta^{2} m b+(\beta m M)^{2}\right], \quad \gamma=\frac{1}{2}\left((\beta m)^{2}+\left(1-\frac{1}{\bar{\pi}_{0}^{2}+1}\right)\right)
\end{align*}
$$

and $\bar{\pi}_{0, \gamma}=\sup _{x} \pi_{0, \gamma}(x), \bar{\pi}_{0}=\sup _{x} \pi_{0}(x)$.
Proof Sketch: The full proof is available in supplementary material 7.4. The key tool we use is the main Theorem in [5], reproduced as Proposition 4.4. To satisfy condition (1) of Proposition 4.4 we show that the Lyapunov function

$$
V(w)=\exp \left(\frac{\beta m\|w\|^{2}}{2\left(\bar{\pi}_{0, \gamma}^{2}+1\right)}\right)
$$

and the infinitesimal generator (4.3) satisfy (4.10), with $\kappa$ and $\gamma$ given in (5.6). Then, Proposition 4.3 is used to show that condition (2) is satisfied. Condition (3) is satisfied with $K=\beta L_{\nabla J}$ by assumption 3.1.

Now since $D\left(\nu_{0} \| \pi_{\infty}\right)=D\left(\pi_{0} \| \pi_{\infty}\right)<\infty$ by Lemma 7.2 , we can apply the exponential decay of entropy (Lemma 4.1) to obtain

$$
\begin{equation*}
D\left(\nu_{t} \| \pi_{\infty}\right) \leq D\left(\pi_{0, \gamma} \| \pi_{\infty}\right) e^{-2 t / \beta c_{L S}} \tag{5.7}
\end{equation*}
$$

Then by the Otto-Villani Theorem and Lemma 7.2, we have

$$
\begin{equation*}
\mathcal{W}_{2}\left(\nu_{t}, \pi_{\infty}\right) \leq \sqrt{2 c_{L S} \bar{D}_{0}^{\gamma}} e^{-t / \beta c_{L S}} \tag{5.8}
\end{equation*}
$$

where $\bar{D}_{0}^{\gamma}$ is the relative entropy bound given in (7.2) and $c_{L S}$ is bounded in (5.5).
5.3. Controlling the 2-Wasserstein Distance. Combining the bounds (5.4) and (5.8) yields

$$
\begin{align*}
& \mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right) \leq k \epsilon \sqrt{\epsilon}\left[6 \sqrt{12 C_{0}+3}+3 \sqrt{2}+4 \sqrt{\frac{3}{2}+C_{1}}\left(4 \sqrt{C_{2}}+2 \sqrt{2 L_{J}^{2}+4 C_{0}}\right)\right] \\
& \quad+\sqrt{2 c_{L S} \bar{D}_{0}^{\gamma}} e^{-k \epsilon / \beta c_{L S}} \tag{5.9}
\end{align*}
$$

The strategy to control (5.9) is to take $k \epsilon$ large enough so that the exponential term dies away, then (fixing $k \epsilon$ ) take $\epsilon$ small enough so that the first term decreases arbitrarily. However, we encounter a subtle problem: the term $\bar{D}_{0}^{\gamma}$ may depend inconveniently on $\gamma$, and thus on $\epsilon$, since we take $\gamma$ satisfying (3.5) in order to obtain Lemmas 5.1, 5.2.

Let us investigate this. Lemma 7.2 , with $\gamma \leq 1$ and $\bar{\pi}_{0, \gamma}$ expanded, gives

$$
\bar{D}_{0}^{\gamma} \leq \log \left(\bar{\pi}_{0}\right)+\log \frac{1}{\gamma^{N}}+\frac{N}{2} \log \frac{3 \pi}{m \beta}+\frac{\beta b}{2} \log 3+\beta\left(\frac{L_{\nabla J}}{3} \kappa_{0}+B \sqrt{\kappa_{0}}+A\right)
$$

so we see that $\bar{D}_{0}^{\gamma}$ depends on $\gamma$ as $N \log \left(\frac{1}{\gamma}\right)$. Observe that taking $k \epsilon$ as (3.3) and $\epsilon$ as (3.2) yields

$$
\begin{align*}
& \mathcal{W}_{2}\left(\pi_{k}, \pi_{\infty}\right) \leq \delta\left[6 \sqrt{12 C_{0}+3}+3 \sqrt{2}+4 \sqrt{\frac{3}{2}+C_{1}}\left(4 \sqrt{C_{2}}+2 \sqrt{2 L_{J}^{2}+4 C_{0}}\right)\right] \\
& +\delta \sqrt{2 c_{L S} N \log \left(\frac{1}{\gamma}\right)}+\delta \sqrt{2 c_{L S} C_{3}} \tag{5.10}
\end{align*}
$$

where $C_{3}:=\log (\bar{\pi})+\frac{N}{2} \log \frac{3 \pi}{m \beta}+\frac{\beta b}{2} \log 3+\beta\left(\frac{L_{\nabla J}}{3} \kappa_{0}+B \sqrt{\kappa_{0}}+A\right)$. Then, since $\gamma \in\left[\epsilon^{2}, \epsilon^{3 / 2}\right]$ and $\epsilon \leq\left(\frac{\delta}{\log (1 / \delta)}\right)^{2}$ we have

$$
\log \left(\frac{1}{\gamma}\right) \leq \log \left(\frac{1}{\epsilon^{2}}\right) \leq \log \left(\frac{\log (1 / \delta)}{\delta}\right)^{4} \leq \log \left(\frac{1}{\delta^{5}}\right)=5 \log (1 / \delta)
$$

where we use that $\left(\frac{\log (1 / \delta)}{\delta}\right)^{4} \leq \delta^{-5}$ for all $\delta \leq 1$, satisfied by the feasible $\delta$ range (3.6). Then we obtain the bound displayed in Theorem 3.9.
6. Conclusion. We derived non-asymptotic (finite-sample) bounds for a passive stochastic gradient Langevin dynamics algorithm. These results complement recent asymptotic weak convergence analysis of the passive Langevin algorithm in [16]. The passive Langevin algorithm analyzed in this paper uses sequential evaluations of a stochastic gradient descent by an external agent (forward learner), and reconstructs the cost function being optimized. Thus it achieves real-time (adaptive) inverse reinforcement learning, in that we (the inverse learner) reconstruct the cost function while it is in the process of being optimized. Specifically, we have provided finite-sample bounds on the 2-Wasserstein distance between the sample distribution induced by our algorithm and the Gibbs measure encoding the cost function to be reconstructed. Our paper builds on the seminal paper [22] and utilizes techniques in the analysis of Markov Diffusion Operators [2] to achieve the bound.

## REFERENCES

[1] D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin, A simple proof of the Poincaré inequality for a large class of probability measures, (2008).
[2] D. Bakry, I. Gentil, M. Ledoux, et al., Analysis and geometry of Markov diffusion operators, vol. 103, Springer, 2014.
[3] F. Bolley and C. Villani, Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities, in Annales de la Faculté des sciences de Toulouse: Mathématiques, vol. 14, 2005, pp. 331-352.
[4] D. Brown, W. Goo, P. Nagarajan, and S. Niekum, Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations, in International conference on machine learning, PMLR, 2019, pp. 783-792.
[5] P. Cattiaux, A. Guillin, and L. Wu, A note on Talagrand's transportation inequality and logarithmic sobolev inequality, arXiv preprint arXiv:0810.5435, (2008).
[6] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan, Underdamped Langevin MCMC: A non-asymptotic analysis, in Conference on learning theory, PMLR, 2018, pp. 300-323.
[7] T.-S. Chiang, C.-R. Hwang, and S. J. Sheu, Diffusion for global optimization in r ${ }^{\wedge} n$, SIAM Journal on Control and Optimization, 25 (1987), pp. 737-753.
[8] A. S. Dalalyan and A. Karagulyan, User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient, Stochastic Processes and their Applications, 129 (2019), pp. 52785311.
[9] H. Djellout, A. Guillin, and L. Wu, Transportation cost-information inequalities and applications to random dynamical systems and diffusions, The Annals of Probability, 32 (2004), pp. 2702-2732.
[10] S. B. Gelfand and S. K. Mitter, Simulated annealing type algorithms for multivariate optimization, Algorithmica, 6 (1991), pp. 419-436.
[11] I. GYÖngy, Mimicking the one-dimensional marginal distributions of processes having an Itô differential, Probability theory and related fields, 71 (1986), pp. 501-516.
[12] D. Hadfield-Menell, S. Russell, P. Abbeel, and A. Dragan, Cooperative inverse reinforcement learning, in Advances in neural information processing systems, IEEE, 2016, pp. 3909-3917.
[13] W. Hardle and R. Nixdorf, Nonparametric sequential estimation of zeros and extrema of regression functions, IEEE transactions on information theory, 33 (1987), pp. 367-372.
[14] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al., Advances and open problems in federated learning, Foundations and Trends® in Machine Learning, 14 (2021), pp. 1-210.
[15] I. Karatzas and S. E. Shreve, Brownian motion and stochastic calculus, vol. 113, Springer Science \& Business Media, 1991.
[16] V. Krishnamurthy and G. Yin, Langevin dynamics for adaptive inverse reinforcement learning of stochastic gradient algorithms., J. Mach. Learn. Res., 22 (2021), pp. 121-1.
[17] V. Krishnamurthy and G. Yin, Multikernel passive stochastic gradient algorithms and transfer learning, IEEE Transactions on Automatic Control, 67 (2022), pp. 1792-1805.
[18] A. Krogh and J. Hertz, A simple weight decay can improve generalization, Advances in neural information processing systems, 4 (1991).
[19] H. J. Kushner and G. Yin, Stochastic Approximation Algorithms and Recursive Algorithms and Applications, Springer-Verlag, 2nd ed., 2003.
[20] A. V. Nazin, B. T. Polyak, and A. B. Tsybakov, Passive stochastic approximation, Avtomatika i Telemekhanika, (1989), pp. 127-134.
[21] A. Y. Ng, S. Russell, et Al., Algorithms for inverse reinforcement learning., in Icml, vol. 1, 2000, p. 2.
[22] M. Raginsky, A. Rakhlin, and M. Telgarsky, Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis, in Conference on Learning Theory, PMLR, 2017, pp. 1674-1703.
[23] P. RÉvész, How to apply the method of stochastic approximation in the non-parametric estimation of a regression function, Statistics: A Journal of Theoretical and Applied Statistics, 8 (1977), pp. 119-126.
[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning representations by backpropagating errors, nature, 323 (1986), pp. 533-536.
[25] O. Stramer and R. Tweedie, Langevin-type models $i$ : Diffusions with given stationary distributions and their discretizations, Methodology and Computing in Applied Probability, 1 (1999), pp. 283-306.
[26] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, Policy gradient methods for reinforcement learning with function approximation, Advances in neural information processing systems, 12 (1999).
[27] C. Villani, Topics in optimal transportation, vol. 58, American Mathematical Soc., 2021.
[28] M. Welling and Y. W. Teh, Bayesian learning via stochastic gradient Langevin dynamics, in Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 681-688.
[29] P. Xu, J. Chen, D. Zou, and Q. Gu, Global convergence of Langevin dynamics based algorithms for nonconvex optimization, Advances in Neural Information Processing Systems, 31 (2018).
[30] G. Yin and K. Yin, Passive stochastic approximation with constant step size and window width, IEEE transactions on automatic control, 41 (1996), pp. 90-106.
[31] G. G. Yin and K. Yin, Passive stochastic approximation with constant step size and window width, IEEE transactions on automatic control, 41 (1996), pp. 90-106.
[32] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey, et al., Maximum entropy inverse reinforcement learning., in Aaai, vol. 8, Chicago, IL, USA, 2008, pp. 1433-1438.

## 7. Supplementary Material.

Glossary of Symbols.

| $\theta_{k}$ | SGD (2.8) iterate, $\theta_{k} \in \mathbb{R}^{N}$ |
| :--- | :--- |
| $\alpha_{k}$ | PSGLD (2.9) iterate, $\alpha_{k} \in \mathbb{R}^{N}$ |
| $\epsilon$ | PSGLD step size |
| $\pi_{0, \gamma}$ | sampling distribution $(2.7)$, maximum value: $\bar{\pi}_{0, \gamma}$ |
| $\beta$ | PSGLD inverse temperature parameter |
| $K$ | PSGLD kernel function |
| $\Delta$ | PSGLD kernel function scale parameter |
| J | cost function, $J: \mathbb{R}^{N} \rightarrow \mathbb{R}_{+}$ |
| $\pi_{\infty}$ | Gibbs measure $\left(\pi_{\infty} \propto \exp (-\beta J)\right)$ |
| $\alpha(t)$ | solution of Itô diffusion $(2.12)$ |
| $W(t)$ | standard Brownian motion |
| $c_{L S}$ | diffusion $(2.12)$ log-Sobolev constant |
| $c_{P}$ | diffusion $(2.12)$ Poincarê constant |
| $L \nabla_{\nabla J}$ | $\nabla J$ Lipschitz constant |
| $(m, b)$ | $J$ dissipativity constants |
| $\zeta$ | uniform stochastic gradient $\hat{\nabla} J(\cdot)$ variance bound |
| $L_{\pi_{0}}$ | $\pi_{0}$ Lipschitz constant |
| $M_{\theta}$ | uniform SGD bound, $M_{\theta}:=\sup p_{k \geq 0} \mathbb{E}\left\\|\theta_{k}\right\\|^{2}$ |
| $\kappa_{0}^{\gamma}$ | $\kappa_{0}^{\gamma}:=\log \mathbb{E}_{\pi_{0, \gamma}}\left[\exp \left(\\|x\\|^{2}\right)\right]<\infty\left(\kappa_{0}:=\left.\kappa_{0}^{\gamma}\right\|_{\gamma=1}\right)$ |
| $I$ | $(7.6)$ inner product bound 1 |
| $I^{\prime}$ | $(7.6)$ inner product bound 2 |
| $A$ | $\|J(0)\|$ |
| $B$ | $\\|\nabla J(0)\\|$ |
| $\\|\cdot\\|$ | $l_{2}$ norm |

### 7.1. Bound Constants.

$$
\begin{aligned}
& C_{0}:=3 L_{\nabla J}^{2}\left(M_{\theta}+2 B^{2} M_{\theta}\right)+B^{2}+\zeta \\
& \left.C_{1}:=\kappa_{0}+(\beta b+N) 2 \epsilon+2 I^{\prime}\right) \\
& C_{2}:=\beta L_{\nabla J}^{2}\left(72 C_{0}+6 \sqrt{C_{0}}+18+\sqrt{2}\right) \\
& C_{3}:=\log (\bar{\pi})+\frac{N}{2} \log \frac{3 \pi}{m \beta}+\frac{\beta b}{2} \log 3+\beta\left(\frac{L_{\nabla J}}{3} \kappa_{0}+B \sqrt{\kappa_{0}}+A\right) \\
& C_{4}:=\left[6 \sqrt{12 C_{0}+3}+3 \sqrt{2}+4\left(\frac{3}{2}+C_{1}\right)^{1 / 2}\left(4 \sqrt{C_{2}}+2 \sqrt{2 L_{J}^{2}+4 C_{0}}\right)\right] \\
& M_{\theta}=\kappa_{0}+2\left(1 \vee \frac{1}{m}\right)\left(b+2 B^{2}\right)
\end{aligned}
$$

7.2. Technical Results. Here we present several technical Lemmas which are necessary for the results derived in Section 5.

The proofs for all of these can be found in supplementary material 7.3. We denote $\bar{\pi}_{0}:=\sup _{x} \pi_{0}(x)$ and $\bar{\pi}_{0, \gamma}:=\sup _{x} \pi_{0, \gamma}(x) . A=\|J(0)\|, B=\|\nabla J(0)\|$, and $I, I^{\prime}$ are constants provided in Lemma 7.11.

LEMMA 7.1 ( $\pi_{0, \gamma}$ exponential integrability). For all $\gamma \leq 1, \pi_{0, \gamma}$ has a bounded
and strictly positive density with respect to the Lebesgue measure on $\mathbb{R}^{N}$, and

$$
\begin{equation*}
\kappa_{0}^{\gamma}:=\log \int_{\mathbb{R}^{N}} e^{\|x\|^{2}} d \pi_{0, \gamma}(x)<\infty \tag{7.1}
\end{equation*}
$$

and denote $\kappa_{0}:=\left.\kappa_{0}^{\gamma}\right|_{\gamma=1}$ so that $\kappa_{0}^{\gamma} \leq \kappa_{0} \forall \gamma \leq 1$.
LEmma 7.2 (relative entropy bound).

$$
\begin{array}{r}
\bar{D}_{0}^{\gamma}:=D\left(\pi_{0, \gamma} \| \pi_{\infty}\right) \leq \log \bar{\pi}_{0, \gamma}+\frac{N}{2} \log \frac{3 \pi}{m \beta}+\frac{\beta b}{2} \log 3 \\
+\beta\left(\frac{L_{\nabla J}}{3} \kappa_{0}^{\gamma}+B \sqrt{\kappa_{0}^{\gamma}}+A\right) \tag{7.2}
\end{array}
$$

Lemma 7.3 (exponential integrability of Langevin diffusion).

$$
\log \mathbb{E}\left[e^{\left.\|\alpha(t)\|^{2}\right]}\right] \leq \kappa_{0}^{\gamma}+\left((\beta b+N) 2 \epsilon+2 I^{\prime}\right) t
$$

where $\kappa_{0}$ is given in (7.1) and $I^{\prime}$ is given in (7.6).
Lemma 7.4 (uniform $L^{2}$ bound on SGD). For $\eta \in\left(0,1 \wedge \frac{m}{4 L_{\nabla J}^{2}}\right)$,

$$
\begin{equation*}
\sup _{k \geq 0} \mathbb{E}\left\|\theta_{k}\right\|^{2} \leq \kappa_{0}+2\left(1 \vee \frac{1}{m}\right)\left(b+2 B^{2}\right)=: M_{\theta} \tag{7.3}
\end{equation*}
$$

Lemma 7.5 ( $L^{2}$ bound on Langevin diffusion).

$$
\mathbb{E}\|\alpha(t)\|^{2} \leq \kappa_{0}^{\gamma}+\frac{(\beta b+N) \bar{\pi}_{0, \gamma}+2 I}{(m \beta) \bar{\pi}_{0, \gamma}}
$$

Lemma 7.6. Taking

$$
\begin{equation*}
\Delta \leq \inf _{x \in\left[\epsilon, \hat{K}_{\epsilon}\right]} \frac{K^{-1}\left(\frac{\hat{K}_{1} \sqrt{2 \pi}}{2 \epsilon} e^{x^{2} / 2}\right)}{K^{-2}\left(x \epsilon^{2 N}\right)} \tag{7.4}
\end{equation*}
$$

gives

$$
\mathbb{E}\left\|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right) \hat{\nabla} J\left(\theta_{k}\right)\right\|^{2} \leq 12 \epsilon\left(L_{\nabla J}^{2}\left(M_{\theta}+2 B^{2} M_{\theta}\right)+B^{2}+\zeta\right)
$$

Lemma 7.7. Taking $\pi_{0}$ such that $\bar{\pi}_{0}=1$, and

$$
\begin{equation*}
\gamma \leq \epsilon^{3 / 2} \tag{7.5}
\end{equation*}
$$

gives

$$
\mathbb{E}\left\|\pi_{0, \gamma}\left(\alpha_{k}\right)\right\|^{2} \leq \epsilon, \quad \mathbb{E}\left\|\nabla \pi_{0, \gamma}\left(\alpha_{k}\right)\right\|^{2} \leq \epsilon
$$

Lemma 7.8. Let $R$ be an $N$-dimensional random variable on the same probability space as $\bar{\alpha}(s)$. Then

$$
\mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(s)) R\right\|^{2} \leq 2 \mathbb{E}\left[\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2}\right] \mathbb{E}\left[\|R\|^{2}\right]
$$

Lemma $7.9\left(\pi_{0, \gamma}\right.$ Quadratic Decay). $\quad \pi_{0, \gamma}$ has tail value decay $\mathcal{O}\left(\|x\|^{-2}\right)$ and differential decay $\mathcal{O}\left(\|x\|^{-1}\right)$. Specifically,

$$
\exists M \in \mathbb{N}, \tilde{C}>0: \pi_{0, \gamma}(x) \leq \frac{2}{\left(\beta m^{*}\right)^{2}\|x\|^{2}},\left\|\nabla \pi_{0, \gamma}(x)\right\| \leq \frac{\tilde{C}}{\|x\|} \quad \forall\|x\|>M
$$

Lemma 7.10 (quadratic bounds on $J$ ). For all $w \in \mathbb{R}^{d}$,

$$
\|\nabla J(w)\| \leq L_{\nabla J}\|w\|+B
$$

and

$$
\frac{m}{3}\|w\|^{2}-\frac{b}{2} \leq J(w) \leq \frac{L_{\nabla J}}{2}\|w\|^{2}+B\|w\|+A
$$

Lemma 7.11 (Uniform Gradient Inner Product Bound). We have

$$
\begin{align*}
& \exists I>0:\left\|\left\langle x, \nabla \pi_{0, \gamma}(x)\right\rangle\right\| \leq I \forall x \in \mathbb{R}^{N}, \gamma \in \mathbb{R}_{+}  \tag{7.6}\\
& \exists I^{\prime}>0:\left\|\left\langle x, \pi_{0, \gamma}(x) \nabla \pi_{0, \gamma}(x)\right\rangle\right\| \leq I^{\prime} \forall x \in \mathbb{R}^{N}, \gamma \in \mathbb{R}_{+}
\end{align*}
$$

### 7.3. Proofs of Technical Lemmas.

Proof of Lemma 7.1.
Proof. This follows from Assumption 3.4.
Proof of Lemma 7.2.
Proof. Recall $\pi_{\infty}(w):=\frac{1}{\Lambda} \exp (-\beta J(w))$, where $\Lambda=\int_{\mathbb{R}^{N}} \exp (-\beta J(w)) d w$. Since $\pi_{\infty}>0$ everywhere, we can write

$$
\begin{aligned}
D\left(\pi_{0, \gamma} \| \pi_{\infty}\right) & =\int_{\mathbb{R}^{N}} \pi_{0, \gamma}(x) \log \left(\frac{\pi_{0, \gamma}(x)}{\pi_{\infty}(x)}\right) d x \\
& =\int_{\mathbb{R}^{N}} \pi_{0, \gamma}(x) \log \pi_{0, \gamma}(x) d x+\log \Lambda+\beta \int_{\mathbb{R}^{N}} \pi_{0, \gamma}(x) J(x) d x \\
& \leq \log \left\|\pi_{0, \gamma}\right\|_{\infty}+\log \Lambda+\beta \int_{\mathbb{R}^{N}} \pi_{0, \gamma}(x) J(x) d x \\
& \leq \log \bar{\pi}_{0, \gamma}+\log \Lambda+\beta \int_{\mathbb{R}^{N}} \pi_{0, \gamma}(x) J(x) d x
\end{aligned}
$$

First let us upper bound the normalization constant:

$$
\begin{aligned}
\Lambda & =\int_{\mathbb{R}^{N}} e^{-\beta J(x)} d x \\
& \leq e^{\frac{1}{2} \beta b \log 3} \int_{\mathbb{R}^{N}} e^{-\frac{m \beta\|x\|^{2}}{3}} d x \\
& =3^{\beta b / 2}\left(\frac{3 \pi}{m \beta}\right)^{N / 2}
\end{aligned}
$$

where the inequality follows from Lemma 7.10. Thus,

$$
\log \Lambda \leq \frac{N}{2} \log \frac{3 \pi}{m \beta}+\frac{\beta b}{2} \log 3
$$

By Lemma 7.10 we also have

$$
\begin{aligned}
\int_{\mathbb{R}^{N}} J(x) \pi_{0, \gamma}(x) d x & \leq \int_{\mathbb{R}^{N}} \pi_{0, \gamma} d x\left(\frac{L_{\nabla J}}{3}\|x\|^{2}+B\|x\|+A\right) \\
& \leq \frac{L_{\nabla J}}{3} \kappa_{0}^{\gamma}+B \sqrt{\kappa_{0}^{\gamma}}+A
\end{aligned}
$$

Thus

$$
D\left(\pi_{0, \gamma} \| \pi_{\infty}\right) \leq \log \bar{\pi}_{0, \gamma}+\frac{N}{2} \log \frac{3 \pi}{m \beta}+\frac{\beta b}{2} \log 3+\beta\left(\frac{L_{\nabla J}}{3} \kappa_{0}^{\gamma}+B \sqrt{\kappa_{0}^{\gamma}}+A\right)
$$

## Proof of Lemma 7.3.

Proof. Let us denote $\alpha_{t}:=\alpha(t)$ for notational convenience, and define $L(t):=$ $e^{\left\|\alpha_{t}\right\|^{2}}$. Similarly denote $L_{t}=L(t)$. By Itô's Lemma we have

$$
\left.d L_{t}=\left\{\left(\nabla_{\alpha_{t}} L_{t}\right)^{T} \boldsymbol{\mu}_{t}+\frac{1}{2} \operatorname{Tr}\left[\boldsymbol{G}_{t}^{T}\left(H_{\alpha_{t}} L_{t}\right) \boldsymbol{G}_{t}\right]\right\} d t+\left(\nabla_{\alpha_{t}} L_{t}\right)^{T} \boldsymbol{G}_{t} d W(t)\right)
$$

where from (2.12) we have

$$
\boldsymbol{\mu}_{t}=-\frac{\beta}{2} \pi_{0, \gamma}^{2}\left(\alpha_{t}\right) \nabla J\left(\alpha_{t}\right)-\pi_{0, \gamma}\left(\alpha_{t}\right) \nabla \pi_{0, \gamma}\left(\alpha_{t}\right), \boldsymbol{G}_{t}=\pi_{0, \gamma}\left(\alpha_{t}\right)
$$

Thus,

$$
\left(\nabla_{\alpha_{t}} L_{t}\right)^{T} \boldsymbol{\mu}_{t}=-\beta\left\langle\alpha_{t} L_{t}, \pi_{0, \gamma}^{2}\left(\alpha_{t}\right) \nabla J\left(\alpha_{t}\right)\right\rangle-2\left\langle\alpha_{t} L_{t}, \pi_{0, \gamma}\left(\alpha_{t}\right) \nabla \pi_{0, \gamma}\left(\alpha_{t}\right)\right\rangle
$$

and

$$
\frac{1}{2} \operatorname{Tr}\left[\boldsymbol{G}_{t}^{T}\left(H_{\alpha_{t}} L_{t}\right) \boldsymbol{G}_{t}\right]=\frac{1}{2} \operatorname{Tr}\left[\pi_{0, \gamma}^{2}\left(\alpha_{t}\right) H_{\alpha_{t}} L_{t}\right]=\pi_{0, \gamma}^{2}\left(\alpha_{t}\right)\left(\left\|\alpha_{t}\right\|^{2} L_{t}+N L_{t}\right)
$$

and

$$
\left(\nabla_{\alpha_{t}} L_{t}\right)^{T} \boldsymbol{G}_{t}=2 \alpha_{t}^{*} \pi_{0, \gamma}\left(\alpha_{t}\right) L_{t}
$$

Putting these together and integrating,

$$
\begin{aligned}
L_{t}= & L(0)-\beta \int_{0}^{t}\left\langle\alpha_{s} L_{s}, \pi_{0, \gamma}^{2}\left(\alpha_{s}\right) \nabla J\left(\alpha_{s}\right)\right\rangle d s-2 \int_{0}^{t}\left\langle\alpha_{s} L_{s}, \pi_{0, \gamma}\left(\alpha_{s}\right) \nabla \pi_{0, \gamma}\left(\alpha_{s}\right)\right\rangle d s \\
& +\int_{0}^{t} \pi_{0, \gamma}^{2}\left(\alpha_{s}\right)\left(\left\|\alpha_{s}\right\|^{2} L_{s}+N L_{s}\right) d s+\int_{0}^{t} 2 \alpha_{s}^{*} \pi_{0, \gamma}\left(\alpha_{s}\right) L_{s} d W(s) \\
= & L(0)+\int_{0}^{t}\left(\pi_{0, \gamma}^{2}\left(\alpha_{s}\right)\left\|\alpha_{s}\right\|^{2}-\beta\left\langle\alpha_{s}, \pi_{0, \gamma}^{2}\left(\alpha_{s}\right) \nabla J\left(\alpha_{s}\right)\right\rangle\right) L_{s} d s \\
& -2 \int_{0}^{t}\left\langle\alpha_{s} L_{s}, \pi_{0, \gamma}\left(\alpha_{s}\right) \nabla \pi_{0, \gamma}\left(\alpha_{s}\right)\right\rangle d s+\int_{0}^{t} N L_{s} \pi_{0, \gamma}^{2}\left(\alpha_{s}\right) d s \\
& +\int_{0}^{t} 2 \alpha_{s}^{*} \pi_{0, \gamma}\left(\alpha_{s}\right) L_{s} d W(s)
\end{aligned}
$$

Now, from the dissipativity condition 3.2 , we can obtain the following bound:

$$
\begin{aligned}
& \pi_{0, \gamma}^{2}\left(\alpha_{s}\right)\left(\left\|\alpha_{s}\right\|^{2}-\beta\left\langle\alpha_{s}, \nabla J\left(\alpha_{s}\right)\right\rangle\right) \leq \pi_{0, \gamma}^{2}\left(\alpha_{s}\right)\left(\left\|\alpha_{s}\right\|^{2}+\beta\left[-m\left\|\alpha_{s}\right\|^{2}+b\right]\right) \\
& =\left\|\alpha_{s}\right\|^{2}\left(\pi_{0, \gamma}^{2}\left(\alpha_{s}\right)-\beta m\right)+\pi_{0, \gamma}^{2}\left(\alpha_{s}\right) \beta b \leq \pi_{0, \gamma}^{2}\left(\alpha_{s}\right) \beta b
\end{aligned}
$$

Making this substitution, we now work with

$$
\begin{aligned}
L_{t} \leq & L(0)+(\beta b+N) \int_{0}^{t} \pi_{0, \gamma}^{2}(\alpha(s)) L_{s} d s-2 \int_{0}^{t}\left\langle\alpha_{s} L_{s}, \pi_{0, \gamma}\left(\alpha_{s}\right) \nabla \pi_{0, \gamma}\left(\alpha_{s}\right)\right\rangle d s \\
& +\int_{0}^{t} 2 \alpha_{s}^{*} \pi_{0, \gamma}\left(\alpha_{s}\right) L_{s} d W(s)
\end{aligned}
$$

It can be shown (e.g., proof of Corollary 4.1 in [9]) that $\int_{0}^{T} \mathbb{E}\left\|L_{t} \alpha(t)\right\|^{2} d t<\infty \forall T \geq 0$. Therefore the Itô integral $\int L_{s} \alpha_{s}^{*} d W(s)$ is a zero-mean martingale. Thus, taking expectations leaves us with

$$
\begin{aligned}
\mathbb{E}\left[L_{t}\right] \leq & \mathbb{E}[L(0)]+\mathbb{E}\left[(\beta b+N) \int_{0}^{t} \pi_{0, \gamma}^{2}(\alpha(s)) L_{s} d s\right] \\
& -\mathbb{E}\left[2 \int_{0}^{t}\left\langle\alpha_{s}, \pi_{0, \gamma}\left(\alpha_{s}\right) \nabla \pi_{0, \gamma}\left(\alpha_{s}\right)\right\rangle L_{s} d s\right] \\
= & \mathbb{E}[L(0)]+(\beta b+N) \int_{0}^{t} \mathbb{E}\left[\pi_{0, \gamma}^{2}(\alpha(s)) L_{s}\right] d s \\
& \left.-2 \int_{0}^{t} \mathbb{E}\left\langle\alpha_{s}, \pi_{0, \gamma}\left(\alpha_{s}\right) \nabla \pi_{0, \gamma}\left(\alpha_{s}\right)\right\rangle L_{s} d s\right] \\
\leq & \mathbb{E}[L(0)]+(\beta b+N) 2 \epsilon \int_{0}^{t} \mathbb{E}\left[L_{s}\right] d s+2 I^{\prime} \int_{0}^{t} \mathbb{E}\left[L_{s}\right] d s \\
= & \mathbb{E}[L(0)]+\left((\beta b+N) 2 \epsilon+2 I^{\prime}\right) \int_{0}^{t} \mathbb{E}\left[L_{s}\right] d s \\
= & e^{\kappa_{0}^{\gamma}}+\left((\beta b+N) 2 \epsilon+2 I^{\prime}\right) \int_{0}^{t} \mathbb{E}\left[L_{s}\right] d s
\end{aligned}
$$

By application of the Gronwall Inequality, we obtain

$$
\begin{aligned}
\mathbb{E}\left[L_{t}\right] & \leq \exp \left(\kappa_{0}^{\gamma}+\int_{0}^{t}\left((\beta b+N) 2 \epsilon+2 I^{\prime}\right) d s\right. \\
& =\exp \left(\kappa_{0}^{\gamma}+\left((\beta b+N) 2 \epsilon+2 I^{\prime}\right) t\right)
\end{aligned}
$$

Thus,

$$
\log \mathbb{E}\left[e^{\left.\|\alpha(t)\|^{2}\right]}\right] \leq \kappa_{0}^{\gamma}+\left((\beta b+N) 2 \epsilon+2 I^{\prime}\right) t
$$

## Proof of Lemma 7.4.

Proof. See Lemma 3 of [22], with $\beta=\infty$.
Proof of Lemma 7.5.
Proof. We consider the diffusion given by (2.12). Letting $Y(t)=\|\alpha(t)\|^{2}$, Itô's Lemma gives

$$
\begin{aligned}
d Y(t)= & {\left[-2\left\langle\alpha(t), \frac{\beta}{2} \pi_{0, \gamma}^{2}(\alpha(t)) \nabla J(\alpha(t))+\pi_{0, \gamma}(\alpha(t)) \nabla \pi_{0, \gamma}(\alpha(t))\right\rangle+N \pi_{0, \gamma}^{2}(\alpha(t))\right] d t } \\
& +\pi_{0, \gamma}(\alpha(t)) \alpha(t)^{*} d W(t)
\end{aligned}
$$

where $\alpha(t)^{*} d W(t):=\sum_{i=1}^{N} \alpha_{i}(t) d W_{i}(t)$. Letting $m:=\frac{m \beta}{2} \bar{\pi}_{0, \gamma}^{2}$, we then form

$$
\begin{aligned}
d\left(e^{2 m t} Y(t)\right)= & 2 m e^{2 m t} Y(t)+e^{2 m t} d Y(t) \\
= & {\left[-2 e^{2 m t}\left\langle\alpha(t), \frac{\beta}{2} \pi_{0, \gamma}^{2}(\alpha(t)) \nabla J(\alpha(t))+\pi_{0, \gamma}(\alpha(t)) \nabla \pi_{0, \gamma}(\alpha(t))\right\rangle\right.} \\
& \left.+N \pi_{0, \gamma}^{2}(\alpha(t)) e^{2 m t}+2 m e^{2 m t} Y(t)\right] d t \\
& +e^{2 m t} \pi_{0, \gamma}(\alpha(t)) \alpha(t)^{*} d W(t)
\end{aligned}
$$

Then integrating yields

$$
Y(t)=e^{-2 m t} Y(0)-2 \int_{0}^{t} e^{2 m(s-t)}\left\langle\alpha(s), \frac{\beta}{2} \pi_{0, \gamma}^{2}(\alpha(s)) \nabla J(\alpha(s))\right.
$$

$$
\begin{aligned}
& \left.+\pi_{0, \gamma}(\alpha(s)) \nabla \pi_{0, \gamma}(\alpha(s))\right\rangle d s \\
& +2 m \int_{0}^{t} e^{2 m(s-t)} Y(s) d s+\int_{0}^{t} N \pi_{0, \gamma}^{2}(\alpha(s)) e^{2 m(s-t)} d s \\
& +2 \int_{0}^{t} e^{2 m(s-t)} \pi_{0, \gamma}(\alpha(s)) \alpha(s)^{*} d W(s)
\end{aligned}
$$

Then using the dissipativity condition 3.2 we get

$$
\begin{aligned}
Y(t) \leq & e^{-2 m t} Y(0)+\beta \int_{0}^{t} \pi_{0, \gamma}^{2}(\alpha(s)) e^{2 m(s-t)}(b-m Y(s)) d s \\
& +2 \int_{0}^{t} \pi_{0, \gamma}(\alpha(s)) e^{2 m(s-t)} I d s \\
& +m \beta \int_{0}^{t} \bar{\pi}_{0, \gamma}^{2} e^{2 m(s-t)} Y(s) d s+\int_{0}^{t} N \bar{\pi}_{0, \gamma}^{2} e^{2 m(s-t)} d s \\
& +2 \int_{0}^{t} e^{2 m(s-t)} \pi_{0, \gamma}(\alpha(s)) \alpha(s)^{*} d W(s) \\
\leq & e^{-2 m t} Y(0)+\beta b \int_{0}^{t} \bar{\pi}_{0, \gamma}^{2} e^{2 m(s-t)} d s+2 I \int_{0}^{t} \bar{\pi}_{0, \gamma} e^{2 m(s-t)} d s \\
& +\int_{0}^{t} N \bar{\pi}_{0, \gamma}^{2} e^{2 m(s-t)} d s
\end{aligned}
$$

Then grouping terms and evaluating the integral yields

$$
\begin{aligned}
Y(t) \leq & e^{-2 m t} Y(0)+\frac{(\beta b+N) \bar{\pi}_{0, \gamma}^{2}+2 I \bar{\pi}_{0, \gamma}}{2 m}\left(1-e^{-2 m t}\right) \\
& +\bar{\pi}_{0, \gamma} \int_{0}^{t} e^{2 m(s-t)} \alpha(s)^{*} d W(s)
\end{aligned}
$$

Now taking expectations, and by the Martingale property of the Itô integral, we have

$$
\begin{aligned}
\mathbb{E}\left[\|\alpha(t)\|^{2}\right] & \leq e^{-2 m t} \mathbb{E}\|\alpha(0)\|^{2}+\frac{(\beta b+N) \bar{\pi}_{0, \gamma}^{2}+2 I \bar{\pi}_{0, \gamma}}{2 m}\left(1-e^{-2 m t}\right) \\
& \leq e^{-2 m t} \mathbb{E}\|\alpha(0)\|^{2}+\frac{(\beta b+N) \bar{\pi}_{0, \gamma}^{2}+2 I \bar{\pi}_{0, \gamma}}{2 m}
\end{aligned}
$$

and from (7.1), and using $m=\frac{m \beta}{2} \bar{\pi}_{0, \gamma}^{2}$, and taking the maximum over $t>0$ gives

$$
\mathbb{E}\|\alpha(t)\|^{2} \leq \kappa_{0}^{\gamma}+\frac{(\beta b+N) \bar{\pi}_{0, \gamma}+2 I}{(m \beta) \bar{\pi}_{0, \gamma}}
$$

## Proof of Lemma 7.6.

Proof.

$$
\begin{aligned}
& \mathbb{E}\left\|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right) \hat{\nabla} J\left(\theta_{k}\right)\right\|^{2}=\mathbb{E}\left|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right)\right|^{2}\left\|\hat{\nabla} J\left(\theta_{k}\right)\right\|^{2} \\
& \leq \mathbb{E}\left|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right)\right|^{2} \mathbb{E}\left\|\hat{\nabla} J\left(\theta_{k}\right)\right\|^{2}+\left[\operatorname{Cov}\left(\left|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right)\right|^{2},\left\|\hat{\nabla} J\left(\theta_{k}\right)\right\|^{2}\right)\right]
\end{aligned}
$$

By Cauchy-Schwarz,

$$
\begin{aligned}
& \operatorname{Cov}\left(\left|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right)\right|^{2},\left\|\hat{\nabla} J\left(\theta_{k}\right)\right\|^{2}\right) \leq \operatorname{Var}\left(\left|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right)\right|^{2}\right) \operatorname{Var}\left(\left\|\hat{\nabla} J\left(\theta_{k}\right)\right\|^{2}\right) \\
& \quad \leq \mathbb{E}\left[\left|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right)\right|^{2}\right] \mathbb{E}\left[\left\|\hat{\nabla} J\left(\theta_{k}\right)\right\|^{2}\right] \leq \mathbb{E}\left|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right)\right|^{2} \mathbb{E}\left\|\hat{\nabla} J\left(\theta_{k}\right)\right\|^{2}
\end{aligned}
$$

so

$$
\mathbb{E}\left\|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right) \hat{\nabla} J\left(\theta_{k}\right)\right\|^{2} \leq 2 \mathbb{E}\left|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right)\right|^{2} \mathbb{E}\left\|\hat{\nabla} J\left(\theta_{k}\right)\right\|^{2}
$$

Then we bound

$$
\begin{aligned}
& \mathbb{E}\left[\left\|\hat{\nabla} J\left(\theta_{k}\right)\right\|^{2}\right] \\
& \leq 3\left(\mathbb{E}\left\|L_{\nabla J}\left|\theta_{k}\right|+B\right\|^{2}+\zeta\right)=3\left(\mathbb{E}\left[L_{\nabla J}^{2}\left|\theta_{k}\right|^{2}\right]+2 \mathbb{E} \sup _{k \in \mathbb{N}}\left[L_{\nabla J}^{2} B^{2}\left|\theta_{k}\right|^{2}\right]+B^{2}+\zeta\right) \\
& \leq 3\left(L_{\nabla J}^{2}\left(M_{\theta}+2 B^{2} M_{\theta}\right)+B^{2}+\zeta\right)=: C_{0}
\end{aligned}
$$

and thus we have

$$
\begin{equation*}
\mathbb{E}\left\|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right) \hat{\nabla} J\left(\theta_{k}\right)\right\|^{2} \leq 6 \mathbb{E}\left\|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right)\right\|^{2}\left(L_{\nabla J}^{2}\left(M_{\theta}+2 B^{2} M_{\theta}\right)+B^{2}+\zeta\right) \tag{7.7}
\end{equation*}
$$

so we can control this quantity directly by controlling $\mathbb{E}\left\|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right)\right\|^{2}$, which is done as follows: Notice that

$$
\mathbb{P}\left(\frac{1}{\Delta^{2 N}} K^{2}\left(\frac{\left|\theta_{k}-\alpha_{k}\right|}{\Delta}\right)>x\right)=\mathbb{P}\left(K\left(\theta_{k}, \alpha_{k}\right)>K\left(\Delta K^{2^{-1}}\left(x \Delta^{2 N}\right)\right)\right)
$$

By Markov's Inequality we have

$$
\left.\mathbb{P}\left(K\left(\theta_{k}, \alpha_{k}\right)>K\left(\Delta K^{2^{-1}}\left(x \Delta^{2 N}\right)\right)\right)\right) \leq \frac{\mathbb{E}\left\|K\left(\theta_{k}, \alpha_{k}\right)\right\|}{\left.K\left(\Delta K^{2^{-1}}\left(x \Delta^{2 N}\right)\right)\right)} \leq \frac{\hat{K}}{\left.K\left(\Delta K^{2-1}\left(x \Delta^{2 N}\right)\right)\right)}
$$

Then, for all $x \in\left[\epsilon, \hat{K}_{\epsilon}\right]$, by choosing $\Delta$ as (7.4) we have

$$
\mathbb{P}\left(\frac{1}{\Delta^{2 N}} K^{2}\left(\frac{\left|\theta_{k}-\alpha_{k}\right|}{\Delta}\right)>x\right) \leq \frac{2 \epsilon}{\sqrt{2 \pi}} e^{-x^{2}}
$$

So now observe

$$
\begin{aligned}
& \mathbb{E}\left\|K_{\Delta}\left(\theta_{k}, \alpha_{k}\right)\right\|^{2}=\mathbb{E}\left[\frac{1}{\Delta^{2 N}} K^{2}\left(\frac{\left|\theta_{k}-\alpha_{k}\right|}{\Delta}\right)\right] \\
& =\int_{0}^{\infty} \mathbb{P}\left(\frac{1}{\Delta^{2 N}} K^{2}\left(\frac{\left|\theta_{k}-\alpha_{k}\right|}{\Delta}\right)>x\right) d x \\
& \quad=\int_{0}^{\epsilon} \mathbb{P}\left(\frac{1}{\Delta^{2 N}} K^{2}\left(\frac{\left|\theta_{k}-\alpha_{k}\right|}{\Delta}\right)>x\right) d x+\int_{\epsilon}^{\hat{K}_{\epsilon}} \mathbb{P}^{\hat{R}^{\epsilon}}\left(\frac{1}{\Delta^{2 N}} K^{2}\left(\frac{\left|\theta_{k}-\alpha_{k}\right|}{\Delta}\right)>x\right) d x \\
& \quad \leq \epsilon+\int_{\epsilon} \frac{2 \epsilon}{\sqrt{2 \pi}} e^{-x^{2}} d x \leq 2 \epsilon
\end{aligned}
$$

## Proof of Lemma 7.7.

Proof.

$$
\mathbb{E}\left[\left\|\pi_{0, \gamma}\left(\alpha_{k}\right)\right\|^{2}\right]=\int_{0}^{\infty} \mathbb{P}\left(\pi_{0, \gamma}^{2}\left(\alpha_{k}\right)>x\right) d x
$$

We define the level set

$$
\Delta_{x}^{\gamma}:=\left\{y \in \mathbb{R}^{N}: \pi_{0, \gamma}^{2}(y)>x\right\}
$$

so that

$$
\mathbb{P}\left(\pi_{0, \gamma}^{2}\left(\alpha_{k}\right)>x\right)=\mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma}\right)
$$

Now split this term as

$$
\begin{aligned}
\mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma}\right)= & \mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma} \mid \alpha_{k-1} \in \Delta_{x}^{\gamma}\right) \mathbb{P}\left(\alpha_{k-1} \in \Delta_{x}^{\gamma}\right) \\
& +\mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma} \mid \alpha_{k-1} \notin \Delta_{x}^{\gamma}\right) \mathbb{P}\left(\alpha_{k-1} \notin \Delta_{x}^{\gamma}\right) \\
\leq & \mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma} \mid \alpha_{k-1} \in \Delta_{x}^{\gamma}\right)+\mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma} \mid \alpha_{k-1} \notin \Delta_{x}^{\gamma}\right)
\end{aligned}
$$

Now, from (2.9), denote

$$
\nabla_{k}:=\epsilon\left[K\left(\frac{\theta_{k}-\alpha_{k}}{\Delta}\right) \frac{\beta}{2} \hat{\nabla} J\left(\theta_{k}\right)+\nabla \pi_{0, \gamma}\left(\alpha_{k}\right)\right] \pi_{0, \gamma}\left(\alpha_{k}\right), \quad \tilde{w}_{k}:=\sqrt{\epsilon} \pi_{0, \gamma}\left(\alpha_{k}\right) w_{k}
$$

so that $\tilde{w}_{k} \sim \mathcal{N}\left(0, \epsilon \pi_{0, \gamma}\left(\alpha_{k}\right)^{2}\right)$. Then observe that, given $\tilde{w}_{k}$ is symmetric with mean zero,

$$
\begin{aligned}
\mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma} \mid \alpha_{k-1} \in \Delta_{x}^{\gamma}\right) & =\mathbb{P}\left(\alpha_{k-1}-\nabla_{k-1}+\tilde{w}_{k} \in \Delta_{x}^{\gamma} \mid \alpha_{k-1} \in \Delta_{x}^{\gamma}\right) \\
& \leq \mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma} \mid \alpha_{k-1}-\nabla_{k-1} \in \Delta_{x}^{\gamma}\right)
\end{aligned}
$$

Similarly,

$$
\mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma} \mid \alpha_{k-1} \notin \Delta_{x}^{\gamma}\right) \leq \mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma} \mid \alpha_{k-1}-\nabla_{k} \in \Delta_{x}^{\gamma}\right)
$$

so that

$$
\mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma}\right) \leq 2 \mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma} \mid \alpha_{k-1}-\nabla_{k-1} \in \Delta_{x}^{\gamma}\right)
$$

Now, let $\zeta_{z}^{\epsilon, \gamma}:=\left(\epsilon \pi_{0, \gamma}^{2}(z)\right)^{-1}$, define

$$
\hat{\Delta}_{x}^{\gamma}(z):=\left\{y \in \mathbb{R}^{N}: y / z \in \Delta_{x}^{\gamma}\right\}
$$

and notice that

$$
\begin{aligned}
& \mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma} \mid \alpha_{k-1}-\nabla_{k-1} \in \Delta_{x}^{\gamma}\right) \\
& =\mathbb{P}\left(\zeta_{\alpha_{k-1}}^{\epsilon, \gamma} \alpha_{k} \in \hat{\Delta}_{x}^{\gamma}\left(\zeta_{\alpha_{k-1}}^{\epsilon, \gamma}\right) \mid \zeta_{\alpha_{k-1}}^{\epsilon, \gamma} \alpha_{k-1}-\zeta_{\alpha_{k-1}}^{\epsilon, \gamma} \nabla_{k-1} \in \hat{\Delta}_{x}^{\gamma}\left(\zeta_{\alpha_{k-1}}^{\epsilon, \gamma}\right)\right) \\
& =\mathbb{P}\left(\zeta_{\alpha_{k-1}}^{\epsilon, \gamma}\left(\alpha_{k-1}-\nabla_{k-1}+\sqrt{\epsilon} \pi_{0, \gamma}\left(\alpha_{k-1}\right) w_{k}\right) \in \hat{\Delta}_{x}^{\gamma}\left(\zeta_{\alpha_{k-1}}^{\epsilon, \gamma}\right) \mid\right. \\
& \left.\quad \zeta_{\alpha_{k-1}}^{\epsilon, \gamma}\left(\alpha_{k-1}-\nabla_{k-1}\right) \in \hat{\Delta}_{x}^{\gamma}\left(\zeta_{\alpha_{k-1}}^{\epsilon, \gamma}\right)\right)
\end{aligned}
$$

Then, since $\zeta_{\alpha_{k-1}}^{\epsilon, \gamma} \sqrt{\epsilon} \pi_{0, \gamma}\left(\alpha_{k-1}\right) w_{k} \sim \mathcal{N}\left(0,\left(\zeta_{\alpha_{k-1}}^{\epsilon, \gamma}\right)^{2} \epsilon \pi_{0, \gamma}^{2}\left(\alpha_{k-1}\right)\right)$ we have that,

$$
\begin{aligned}
& \mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma} \mid \alpha_{k-1}-\nabla_{k-1} \in \Delta_{x}^{\gamma}\right) \\
& \quad \leq \int_{\hat{\Delta}_{x}^{\gamma}\left(\zeta_{\alpha_{k-1}}^{\epsilon, \gamma}\right)} \mathcal{N}\left(\gamma ; \hat{c},\left(\zeta_{\alpha_{k-1}}^{\epsilon, \gamma}\right)^{2} \epsilon \pi_{0, \gamma}^{2}\left(\alpha_{k-1}\right)\right) d \gamma=\int_{\hat{\Delta}_{x}^{\gamma}\left(\zeta_{\alpha_{k-1}}^{\epsilon, \gamma}\right)} \mathcal{N}\left(\gamma ; \hat{c}, \zeta_{\alpha_{k-1}}^{\epsilon, \gamma}\right) d \gamma \\
&
\end{aligned}
$$

Now, crucially, observe that the volume of $\hat{\Delta}_{x}^{\gamma}\left(\zeta_{z}^{\epsilon, \gamma}\right)$ scales, w.r.t $z$, at the same rate as the variance of $\mathcal{N}\left(\cdot, \hat{c}, \zeta_{z}^{\epsilon, \gamma}\right)$. Thus we have

$$
\int_{\hat{\Delta}_{x}^{\gamma}\left(\zeta_{z_{1}^{\epsilon}}^{\epsilon}\right)} \mathcal{N}\left(\gamma ; \hat{c}, \zeta_{z_{1}^{\prime}}^{\epsilon, \gamma}\right) d \gamma=\int_{\hat{\Delta}_{x}^{\gamma}\left(\zeta_{z_{2}^{\epsilon}}^{\epsilon,}\right)} \mathcal{N}\left(\gamma ; \hat{c}, \zeta_{z_{2}}^{\epsilon, \gamma}\right) d \gamma \quad \forall z_{1}, z_{2}>0
$$

In particular, take $z$ such that $\zeta_{z}^{\epsilon, \gamma}=\left(\epsilon \pi_{0, \gamma}^{2}(z)\right)^{-1}=\epsilon^{2}$. Note that this necessitates $\bar{\pi}_{0, \gamma} \geq\left(\frac{1}{\epsilon}\right)^{3 / 2}$, which is given from the condition $\gamma \leq \bar{\pi}_{0} \epsilon^{3 / 2}$. Then we have

$$
\begin{aligned}
\mathbb{P}\left(\alpha_{k} \in \Delta_{x}^{\gamma} \mid \alpha_{k-1}-\nabla_{k-1} \in \Delta_{x}^{\gamma}\right) & \leq \int_{\hat{\Delta}_{x}^{\gamma}\left(\epsilon^{2}\right)} \mathcal{N}\left(\gamma ; \hat{c}, \epsilon^{2}\right) d \gamma \\
& \leq \int_{\hat{\Delta}_{x}^{\gamma}\left(\epsilon^{2}\right)} \frac{1}{\epsilon \sqrt{2 \pi}} d \gamma=\int_{\hat{\Delta}_{x}^{\gamma}(1)} \frac{\epsilon}{\sqrt{2 \pi}} d \gamma
\end{aligned}
$$

Then

$$
\begin{aligned}
\mathbb{E}\left[\left\|\pi_{0, \gamma}\left(\alpha_{k}\right)\right\|^{2}\right] & =\int_{0}^{\infty} \mathbb{P}\left(\pi_{0, \gamma}^{2}\left(\alpha_{k}\right)>x\right) d x \\
& \leq 2 \int_{0}^{\infty} \int_{\Delta_{x}^{\gamma}} \frac{\epsilon}{\sqrt{2 \pi}} d \gamma d x=\sqrt{\frac{2}{\pi}} \epsilon \int_{0}^{\infty} \int_{\mathbb{R}^{N}} \mathbb{1}\left\{\pi_{0, \gamma}(\gamma)>\sqrt{x}\right\} d \gamma d x
\end{aligned}
$$

but observe that, for $\gamma \leq 1$,

$$
\int_{0}^{\infty} \int_{\mathbb{R}^{N}} \mathbb{1}\left\{\pi_{0, \gamma}(\gamma)>\sqrt{x}\right\} d \gamma d x \leq \int_{0}^{\infty} \int_{\mathbb{R}^{N}} \mathbb{1}\left\{\pi_{0}(\gamma)>\sqrt{x}\right\} d \gamma d x
$$

So now

$$
\begin{aligned}
& \mathbb{E}\left[\left\|\pi_{0, \gamma}\left(\alpha_{k}\right)\right\|^{2}\right] \leq \sqrt{\frac{2}{\pi}} \epsilon \int_{0}^{\infty} \int_{\mathbb{R}^{N}} \mathbb{1}\left\{\pi_{0}(\gamma)>\sqrt{x}\right\} d \gamma d x \\
&= \sqrt{\frac{2}{\pi}} \epsilon\left[\int_{0}^{1} \int_{\mathbb{R}^{N}} \mathbb{1}\left\{\pi_{0}(\gamma)>\sqrt{x}\right\} d \gamma d x\right. \\
&\left.\quad+\int_{1}^{\bar{\pi}_{0, \gamma}^{2}} \int_{\mathbb{R}^{N}} \mathbb{1}\left\{\pi_{0}(\gamma)>\sqrt{x}\right\} d \gamma d x\right] \\
& \leq \sqrt{\frac{2}{\pi}} \epsilon\left[\int_{0}^{1} \int_{\mathbb{R}^{N}} \mathbb{1}\left\{\pi_{0}(\gamma)>x\right\} d \gamma d x+\bar{\pi}_{0, \gamma}^{2} V\left(\left\{\pi_{0}>1\right\}\right)\right] \\
& \leq \epsilon\left[1+V\left(\left\{\pi_{0}>1\right\}\right)\right]=: \epsilon V_{1}
\end{aligned}
$$

where $V\left(\left\{\pi_{0}>1\right\}\right)$ is shorthand for $\int_{\mathbb{R}^{N}} \mathbb{1}\left\{\pi_{0}(\gamma)>1\right\} d \gamma$.
Define, analagously,

$$
\Gamma_{x}^{\gamma}=\left\{y \in \mathbb{R}^{N}:\left\|\nabla \pi_{0, \gamma}(y)\right\|^{2}>x\right\}
$$

By the same procedure as above, we can obtain

$$
\mathbb{P}\left(\alpha_{k} \in \Gamma_{x}^{\gamma}\right) \leq 2 \int_{\Gamma_{x}^{\gamma}} \frac{\epsilon}{\sqrt{2 \pi}} d \gamma
$$

and so

$$
\begin{aligned}
& \mathbb{E}\left\|\nabla \pi_{0, \gamma}\left(\alpha_{k}\right)\right\|^{2}=\int_{0}^{\infty} \mathbb{P}\left(\left\|\nabla \pi_{0, \gamma}\left(\alpha_{k}\right)\right\|^{2}>x\right) d x \leq 2 \int_{0}^{\infty} \int_{\Gamma_{x}^{\gamma}} \sqrt{\frac{2}{\pi}} \epsilon d \gamma d x \\
& \leq \epsilon\left[1+\bar{\pi}_{0}^{\prime} V\left(\left\{\left\|\nabla \pi_{0}\right\|>1\right\}\right)\right]=: \epsilon V_{2} \\
& 29
\end{aligned}
$$

where $\bar{\pi}_{0}^{\prime}=\sup _{x \in \mathbb{R}^{N}} \nabla \pi_{0}(x), V\left(\left\{\left\|\nabla \pi_{0}\right\|>1\right\}\right)=\int_{\mathbb{R}^{N}} \mathbb{1}\left\{\left\|\nabla \pi_{0}(\gamma)\right\|>1\right\} d \gamma$ Taking $\pi_{0}$ such that $\bar{\pi}_{0}=1$ gives

$$
\mathbb{E}\left[\left\|\pi_{0, \gamma}\left(\alpha_{k}\right)\right\|^{2}\right] \leq \epsilon, \quad \mathbb{E}\left\|\nabla \pi_{0, \gamma}\left(\alpha_{k}\right)\right\|^{2} \leq \epsilon
$$

## Proof of Lemma 7.8.

Proof. First take

$$
\begin{aligned}
& \mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(s)) R\right\|^{2}=\mathbb{E}\left[\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2}\|R\|^{2}\right] \\
& =\mathbb{E}\left[\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2}\right] \mathbb{E}\left[\|R\|^{2}\right]+\operatorname{Cov}\left(\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2},\|R\|^{2}\right)
\end{aligned}
$$

By Cauchy-Schwarz,

$$
\begin{aligned}
& \operatorname{Cov}\left(\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2},\|R\|^{2}\right) \leq \operatorname{Var}\left(\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2}\right) \operatorname{Var}\left(\|R\|^{2}\right) \\
& \quad \leq \mathbb{E}\left[\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2}\right] \mathbb{E}\left[\|R\|^{2}\right]
\end{aligned}
$$

So we have

$$
\begin{equation*}
\mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(s)) R\right\|^{2} \leq 2 \mathbb{E}\left[\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2}\right] \mathbb{E}\left[\|R\|^{2}\right] \tag{7.8}
\end{equation*}
$$

## Proof of Lemma 7.9.

Proof. This follows from Assumption 3.4
Proof of Lemma 7.10.
Proof. Lemma 7.10 equivalent to Lemma 2 of [22], and follows from Assumptions A1, A2, and A3.

## Proof of Lemma 7.11.

Proof. First note that for all $x \in \mathbb{R}^{N}$ we have $\nabla \pi_{0, \gamma}(x)=\frac{\nabla \pi_{0}(x)}{\gamma}$, so

$$
\arg \max _{y \in \mathbb{R}^{N}}\left\|\left\langle y, \nabla \pi_{0, \gamma}(y)\right\rangle\right\|=\arg \max _{y \in \mathbb{R}^{N}}\left\|\left\langle\gamma y, \frac{\nabla \pi_{0}(y)}{\gamma}\right\rangle\right\|=\arg \max _{y \in \mathbb{R}^{N}}\left\|\left\langle y, \nabla \pi_{0}(y)\right\rangle\right\|
$$

and

$$
\begin{aligned}
& \arg \max _{y \in \mathbb{R}^{N}}\left\|\left\langle y, \pi_{0, \gamma}(y) \nabla \pi_{0, \gamma}(y)\right\rangle\right\|=\arg \max _{y \in \mathbb{R}^{N}}\left\|\left\langle\gamma y, \pi_{0, \gamma}(y) \frac{\nabla \pi_{0}(y)}{\gamma}\right\rangle\right\| \\
& \quad=\arg \max _{y \in \mathbb{R}^{N}}\left\|\left\langle y, \pi_{0}(y) \nabla \pi_{0}(y)\right\rangle\right\|
\end{aligned}
$$

Thus (7.6) is equivalent to:

$$
\begin{gather*}
\exists I>0:\left\|\left\langle x, \nabla \pi_{0}(x)\right\rangle\right\| \leq I \forall x \in \mathbb{R}^{N}  \tag{7.9}\\
\exists I^{\prime}>0:\left\|\left\langle x, \pi_{0}(x) \nabla \pi_{0}(x)\right\rangle\right\| \leq I^{\prime} \forall x \in \mathbb{R}^{N} \tag{7.10}
\end{gather*}
$$

Now we prove by reductio ad absurdum: Suppose (7.9), (7.10) do not hold. Then we have:

$$
\begin{align*}
& \forall y>0 \exists x \in \mathbb{R}^{N}:\left\|\left\langle x, \nabla \pi_{0}(x)\right\rangle\right\|>y  \tag{7.11}\\
& \forall y>0 \exists x \in \mathbb{R}^{N}:\left\|\left\langle x, \pi_{0}(x) \nabla \pi_{0}(x)\right\rangle\right\|>y
\end{align*}
$$

30

Recall $M$ as defined in Assumption 3.4. Denote

$$
D_{1}=\left\{x \in \mathbb{R}^{N}:\|x\| \leq M \vee 1\right\}, D_{2}=D_{1}^{C}=\mathbb{R}^{N} \backslash D_{1}
$$

Recall that we assume Lipshitz-continuity of $\pi_{0}(\cdot)$ in (A3.5), with Lipschitz constant $L_{\pi_{0}}$. Thus, $\left\|\nabla \pi_{0}(x)\right\|$ is bounded by $L_{\pi_{0}}$ for all $x \in \mathbb{R}^{N}$. In particular notice that since $\left\|\nabla \pi_{0}(x)\right\|,\left\|\pi_{0}(x)\right\|$, and $\|x\|$ are bounded for $x \in D_{1}$, there exists some $M^{*}, M^{* *}$ such that

$$
\begin{gathered}
\left\|\left\langle x, \nabla \pi_{0}(x)\right\rangle\right\| \leq M^{*} \forall x \in D_{1} \\
\left\|\left\langle x, \pi_{0}(x) \nabla \pi_{0}(x)\right\rangle\right\| \leq M^{* *} \forall x \in D_{1}
\end{gathered}
$$

Then (7.11) requires both:

$$
\begin{align*}
& \forall y>M^{*} \exists x \in D_{2}:\left\|\left\langle x, \nabla \pi_{0}(x)\right\rangle\right\|>y \\
& \forall y>M^{* *} \exists x \in D_{2}:\left\|\left\langle x, \nabla \pi_{0}(x) \nabla \pi_{0}(x)\right\rangle\right\|>y \tag{7.12}
\end{align*}
$$

But by assumption $7.9 \nabla \pi_{0}(x)$ decays as $\mathcal{O}\left(\|x\|^{-1}\right)$ for $x \in D_{2}$, and in Lemma 7.9 $\pi_{0}(x)$ decays as $\mathcal{O}\left(\|x\|^{-2}\right)$. Specifically, there exists $\tilde{C}$ such that

$$
\left\|\nabla \pi_{0}(x)\right\| \leq \frac{\tilde{C}}{\|x\|^{-1}}, \quad\left\|\pi_{0}(x)\right\| \leq \frac{2}{\left(\beta m^{*}\right)^{2}\|x\|^{2}}, \forall x \in D_{2}
$$

Thus we have that

$$
\begin{aligned}
\left\|\left\langle x, \nabla \pi_{0}(x)\right\rangle\right\| & \leq \tilde{C} \forall x \in D_{2} \\
\left\|\left\langle x, \pi_{0}(x) \nabla \pi_{0}(x)\right\rangle\right\| & \leq \frac{2 \tilde{C}}{\left(\beta m^{*}\right)^{2}} \forall x \in D_{2}
\end{aligned}
$$

which contradicts (7.12) and thus (7.11) is refuted. So (7.6) holds.

### 7.4. Proofs of Section 5 Results.

## Proof of Lemma 5.1.

Proof. We begin by bounding this Wasserstein distance by the mean-square error between the processes (5.2) and (5.3). Recall that $Y(k \epsilon)$ has probability law $\pi_{k}$.

$$
\begin{aligned}
\mathcal{W}_{2}\left(\pi_{k}, \gamma_{k \epsilon}\right)= & \inf _{\gamma \in \Gamma\left(\pi_{k}, \gamma_{k \epsilon}\right)}\left(\mathbb{E}_{(x, y) \sim \gamma}\|x-y\|^{2}\right)^{1 / 2} \\
& \leq \sqrt{\mathbb{E}_{x \sim \pi_{k}, y \sim \gamma_{k \epsilon}}\|x-y\|^{2}}
\end{aligned}
$$

Then we take $t=k \epsilon$, and bound

$$
\begin{align*}
& \mathbb{E}\|Y(t)-X(t)\|^{2} \\
& \leq 3 \mathbb{E}\left\|\int_{0}^{t} g_{s}\left(\theta_{\bar{s}}, Y(s)\right)-\hat{g}_{s}\left(\theta_{\bar{s}}, X(s)\right) d s\right\|^{2}  \tag{7.13}\\
& \quad+3 \mathbb{E}\left\|\int_{0}^{t} \mathbb{E}\left[\pi_{0, \gamma}(\bar{\alpha}(\bar{s})) \mid \bar{\alpha}(s)=Y(s)\right]-\pi_{0, \gamma}(X(s)) d W(s)\right\|^{2}
\end{align*}
$$

First we bound, using Jensen's inequality:

$$
\mathbb{E}\left\|\int_{0}^{t} g_{s}\left(\theta_{\bar{s}}, Y(s)\right)-\hat{g}_{s}\left(\theta_{\bar{s}}, X(s)\right) d s\right\|^{2}
$$

$$
\begin{aligned}
&= \mathbb{E} \| \int_{0}^{t} \mathbb{E}\left[\left.\left(K_{\Delta}\left(\theta_{\bar{s}}, \bar{\alpha}(\bar{s})\right) \frac{\beta}{2} \hat{\nabla} J\left(\theta_{\bar{s}}\right)+\nabla \pi_{0, \gamma}(\bar{\alpha}(\bar{s}))\right) \pi_{0, \gamma}(\bar{\alpha}(\bar{s})) \right\rvert\, \bar{\alpha}(s)=Y(s)\right] \\
&-\mathbb{E}\left[\left.\left(K_{\Delta}\left(\theta_{\bar{s}}, \bar{\alpha}(\bar{s})\right) \frac{\beta}{2} \hat{\nabla} J\left(\theta_{\bar{s}}\right)+\nabla \pi_{0, \gamma}(\bar{\alpha}(\bar{s}))\right) \pi_{0, \gamma}(\bar{\alpha}(\bar{s})) \right\rvert\, \bar{\alpha}(s)=X(s)\right] d s \|^{2} \\
& \leq \mathbb{E} \| \int_{0}^{t} \sqrt{\mathbb{E}\left[\left.\left\|K_{\Delta}\left(\theta_{\bar{s}}, \bar{\alpha}(\bar{s})\right) \frac{\beta}{2} \hat{\nabla} J\left(\theta_{\bar{s}}\right) \pi_{0, \gamma}(\bar{\alpha}(\bar{s}))\right\|^{2} \right\rvert\, \bar{\alpha}(s)=Y(s)\right]} \\
& \quad-\sqrt{\mathbb{E}\left[\left.\left\|K_{\Delta}\left(\theta_{\bar{s}}, \bar{\alpha}(\bar{s})\right) \frac{\beta}{2} \hat{\nabla} J\left(\theta_{\bar{s}}\right) \pi_{0, \gamma}(\bar{\alpha}(\bar{s}))\right\|^{2} \right\rvert\, \bar{\alpha}(s)=X(s)\right]} \\
& \quad+\sqrt{\mathbb{E}\left[\left\|\pi_{0, \gamma}(\bar{\alpha}(\bar{s})) \nabla \pi_{0, \gamma}(\bar{\alpha}(\bar{s}))\right\|^{2} \mid \bar{\alpha}(s)=Y(s)\right]} \\
& \quad-\sqrt{\mathbb{E}\left[\left\|\pi_{0, \gamma}(\bar{\alpha}(\bar{s})) \nabla \pi_{0, \gamma}(\bar{\alpha}(\bar{s}))\right\|^{2} \mid \bar{\alpha}(s)=X(s)\right] d s \|^{2}} \\
& \leq \mathbb{E}\left\|\int_{0}^{t} 4 \epsilon \sqrt{2 C_{0}}+2 \sqrt{2} \epsilon d s\right\|^{2} \leq(k \epsilon)^{2} \epsilon^{2}\left(96 C_{0}+24\right)
\end{aligned}
$$

We bound the second term, using the Itô Isometry:

$$
\begin{aligned}
& \mathbb{E}\left\|\int_{0}^{t} \mathbb{E}\left[\pi_{0, \gamma}(\bar{\alpha}(\bar{s})) \mid \bar{\alpha}(s)=Y(s)\right]-\pi_{0, \gamma}(X(s)) d W(s)\right\|^{2} \\
& =\mathbb{E}\left\|\int_{0}^{t} \mathbb{E}\left[\pi_{0, \gamma}(\bar{\alpha}(\bar{s}))-\pi_{0, \gamma}(X(s)) \mid \bar{\alpha}(s)=Y(s)\right] d W(s)\right\|^{2} \\
& =\mathbb{E}\left[\int_{0}^{t} \mathbb{E}\left[\left\|\pi_{0, \gamma}(\bar{\alpha}(\bar{s}))-\pi_{0, \gamma}(X(s))\right\|^{2} \mid \bar{\alpha}(s)=Y(s)\right] d s\right] \\
& \leq 6 \mathbb{E}\left[\int_{0}^{t} \epsilon d s\right]=6 k \epsilon^{2}
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\mathcal{W}_{2}\left(\pi_{k}, \gamma_{k \epsilon}\right) & \leq \sqrt{72(k \epsilon)^{2} \epsilon^{2}\left(4 C_{0}+1\right)+18(k \epsilon) \epsilon} \\
& \leq 6(k \epsilon) \epsilon \sqrt{12 C_{0}+3}+3 \sqrt{2(k \epsilon) \epsilon}
\end{aligned}
$$

## Proof of Lemma 5.2.

Proof. Let $\mathbb{P}_{X}^{t}:=\operatorname{Law}(X(s): 0 \leq s \leq t)$ for $X(s)$ in 5.3 and $\mathbb{P}_{A}^{t}:=\operatorname{Law}(\alpha(s)$ : $0 \leq s \leq t)$ for $\alpha(t)$ in 2.12. The Radon-Nikodym derivative of $\mathbb{P}_{A}^{t}$ with respect to $\mathbb{P}_{X}^{t}$ is given by the Girsanov formula:

$$
\begin{aligned}
& \frac{d \mathbb{P}_{A}^{t}}{d \mathbb{P}_{X}^{t}}(X)=\exp \left\{\int_{0}^{t}\left(G(X(s))-g_{s}\left(\theta_{\bar{s}}, X(s)\right)\right)^{*} \pi_{0, \gamma}(X(s))^{-2} d W(s)\right. \\
& \left.\left.-\frac{1}{2} \int_{0}^{t} \| G(X(s))-g_{s}\left(\theta_{\bar{s}}, X(s)\right)\right) \|^{2} \pi_{0, \gamma}(X(s))^{-2} d s\right\}
\end{aligned}
$$

where $G(X(s))=\frac{\beta}{2} \pi_{0, \gamma}^{2}(X(s)) \nabla J(X(s))+\pi_{0, \gamma}(X(s)) \nabla \pi_{0, \gamma}(X(s))$
Then by the martingale property of the Itô integral,

$$
D\left(\mathbb{P}_{X}^{t} \| \mathbb{P}_{A}^{t}\right)=-\int d \mathbb{P}_{X}^{t} \log \frac{d \mathbb{P}_{A}^{t}}{d \mathbb{P}_{X}^{t}}
$$

$$
\begin{aligned}
= & \left.\frac{1}{2} \int_{0}^{t} \mathbb{E}\left[\| G(X(s))-g_{s}\left(\theta_{\bar{s}}, X(s)\right)\right) \|^{2} \pi_{0, \gamma}(X(s))^{-2}\right] d s \\
= & \frac{1}{2} \int_{0}^{t} \mathbb{E}\left[\| \frac{\beta}{2} \pi_{0, \gamma}^{2}(X(s)) \nabla J(X(s))+\pi_{0, \gamma}(X(s)) \nabla \pi_{0, \gamma}(X(s))\right. \\
& \left.-g_{s}\left(\theta_{\bar{s}}, X(s)\right) \|^{2} \pi_{0, \gamma}(X(s))^{-2}\right] d s \\
= & \frac{1}{2} \int_{0}^{t} \mathbb{E}\left[\| \frac{\beta}{2} \pi_{0, \gamma}^{2}(\bar{\alpha}(s)) \nabla J(\bar{\alpha}(s))+\pi_{0, \gamma}(\bar{\alpha}(s)) \nabla \pi_{0, \gamma}(\bar{\alpha}(s))\right. \\
& \left.-g_{s}\left(\theta_{\bar{s}}, \bar{\alpha}(s)\right) \|^{2} \pi_{0, \gamma}(\bar{\alpha}(s))^{-2}\right] d s
\end{aligned}
$$

where the last line follows from the fact that $\operatorname{Law}(\bar{\alpha}(s))=\operatorname{Law}(X(s)) \forall s \leq t$.
Now let $t=k \epsilon$ for some $k \in \mathbb{N}$. Then, expanding $g_{s}\left(\theta_{\bar{s}}, Y(s)\right)$ and using Jensen's inequality, we get

$$
\begin{align*}
D\left(\mathbb{P}_{X}^{t} \| \mathbb{P}_{A}^{t}\right) \leq & \frac{1}{2} \int_{0}^{t} \mathbb{E}\left[\| \frac{\beta}{2} \pi_{0, \gamma}^{2}(\bar{\alpha}(s)) \nabla J(\bar{\alpha}(s))+\pi_{0, \gamma}(\bar{\alpha}(s)) \nabla \pi_{0, \gamma}(\bar{\alpha}(s))\right.  \tag{7.14}\\
& \left.-\left(K_{\Delta}\left(\theta_{\bar{s}}, \bar{\alpha}(s)\right) \frac{\beta}{2} \hat{\nabla} J\left(\theta_{\bar{s}}\right)+\nabla \pi_{0, \gamma}(\bar{\alpha}(s))\right) \pi_{0, \gamma}(\bar{\alpha}(s)) \|^{2} \pi_{0, \gamma}(\bar{\alpha}(s))^{-2}\right] d s \\
= & \frac{\beta}{4} \int_{0}^{t} \mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(s)) \nabla J(\bar{\alpha}(s))-K_{\Delta}\left(\theta_{\bar{s}}, \bar{\alpha}(s)\right) \hat{\nabla} J\left(\theta_{\bar{s}}\right)\right\|^{2} d s \\
= & \frac{\beta}{4} \sum_{j=0}^{k-1} \int_{j \epsilon}^{(j+1) \epsilon} \mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(s)) \nabla J(\bar{\alpha}(s))-K_{\Delta}\left(\theta_{\bar{s}}, \bar{\alpha}(s)\right) \hat{\nabla} J\left(\theta_{\bar{s}}\right)\right\|^{2} d s \\
\leq & \frac{\beta}{2} \sum_{j=0}^{k-1} \int_{j \epsilon}^{(j+1) \epsilon} \mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(s)) \nabla J(\bar{\alpha}(s))-\pi_{0, \gamma}(\bar{\alpha}(s)) \nabla J(\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon))\right\|^{2} d s \\
& +\frac{\beta}{2} \sum_{j=0}^{k-1} \int_{j \epsilon}^{(j+1) \epsilon} \mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(s)) \nabla J(\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon))-K_{\Delta}\left(\theta_{\bar{s}}, \bar{\alpha}(s)\right) \hat{\nabla} J\left(\theta_{\bar{s}}\right)\right\|^{2} d s \\
\leq & \beta \sum_{j=0}^{k-1} \int_{j \epsilon}^{(j+1) \epsilon} \mathbb{E}\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2} \mathbb{E}\|\nabla J(\bar{\alpha}(s))-\nabla J(\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon))\|^{2} d s \\
& +3 \beta \sum_{j=0}^{k-1} \int_{j \epsilon}^{(j+1) \epsilon} \mathbb{E}\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2} \mathbb{E}\|\nabla J(\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon))-\nabla J(\alpha(0))\|^{2} \\
& +3 \beta \sum_{j=0}^{k-1} \int_{j \epsilon}^{(j+1) \epsilon} \mathbb{E}\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2} \mathbb{E}\|\nabla(J(\alpha(0)))\|^{2} \\
& +3 \beta \sum_{j=0}^{k-1} \int_{j \epsilon}^{(j+1) \epsilon} \mathbb{E}\left\|K_{\Delta}\left(\theta_{\bar{s}}, \bar{\alpha}(s)\right) \hat{\nabla} J\left(\theta_{\bar{s}}\right)\right\|^{2} d s
\end{align*}
$$

where we use Lemma 7.8 in the last inequality.
First Term: By $L_{\nabla J \text {-smoothness }}$ (Assumption 3.1), we begin to control the first
term as:

$$
\mathbb{E}\|\nabla J(\bar{\alpha}(s))-\nabla J(\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon))\|^{2} \leq L_{\nabla J}^{2} \mathbb{E}\|(\bar{\alpha}(s)-\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon))\|^{2}
$$

Then for $s \in[j \epsilon,(j+1) \epsilon]$ :

$$
\begin{aligned}
& (\bar{\alpha}(s)-\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon)) \\
& =(s-j \epsilon) \frac{\beta}{2} K_{\Delta}\left(\theta_{j}, \bar{\alpha}(j \epsilon)\right) \hat{\nabla} J\left(\theta_{j}\right) \pi_{0, \gamma}(\bar{\alpha}(j \epsilon)) \\
& \quad+(s-j \epsilon) \pi_{0, \gamma}(\bar{\alpha}(j \epsilon)) \nabla \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))-\pi_{0, \gamma}(\bar{\alpha}(j \epsilon))(W(s)-W(j \epsilon))
\end{aligned}
$$

Then bound $\mathbb{E}\|(\bar{\alpha}(s)-\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon))\|^{2}$ as:

$$
\begin{align*}
& \mathbb{E}\|(\bar{\alpha}(s)-\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon))\|^{2} \\
& \leq 6 \epsilon^{2} \mathbb{E}\left\|K_{\Delta}\left(\theta_{j}, \bar{\alpha}(j \epsilon)\right) \frac{\beta}{2} \hat{\nabla} J\left(\theta_{j}\right)\right\|^{2} \mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2}  \tag{7.15}\\
&+3 \epsilon^{2} \mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(j \epsilon)) \nabla \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2}+3 \epsilon \mathbb{E} \| \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\left(W(s)-W(j \epsilon) \|^{2}\right.
\end{align*}
$$

But observe that

$$
\begin{aligned}
& \mathbb{E}\|(\bar{\alpha}(s)-\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon))\|^{2} \\
&= \mathbb{E}\left[(s-j \epsilon)^{2}\left\|K_{\Delta}\left(\theta_{j}, \bar{\alpha}(j \epsilon)\right) \hat{\nabla} J\left(\theta_{j}\right) \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2}\right] \\
&+\mathbb{E}\left[(s-j \epsilon)^{2}\left\|\pi_{0, \gamma}(\bar{\alpha}(j \epsilon)) \nabla \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2}\right] \\
&+2 \mathbb{E}\left[(s-j \epsilon)^{2}\left\langle K_{\Delta}\left(\theta_{j}, \bar{\alpha}(j \epsilon)\right) \hat{\nabla} J\left(\theta_{j}\right), \pi_{0, \gamma}^{2}(\bar{\alpha}(j \epsilon)) \nabla \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\rangle\right] \\
&\left.-\mathbb{E}\left[(s-j \epsilon)\left\langle K_{\Delta}\left(\theta_{j}\right), \bar{\alpha}(j \epsilon)\right) \hat{\nabla} J\left(\theta_{j}\right), \pi_{0, \gamma}^{2}(W(s)-W(j \epsilon))\right\rangle\right] \\
&-\mathbb{E}\left[(s-j \epsilon)\left\langle\nabla \pi_{0, \gamma}(\bar{\alpha}(j \epsilon)), \pi_{0, \gamma}^{2}(W(s)-W(j \epsilon)\rangle\right]\right. \\
&+\mathbb{E}\left[\pi_{0, \gamma}^{2} \| \bar{\alpha}(j \epsilon)\right)\left(W(s)-W(j \epsilon) \|^{2}\right]
\end{aligned}
$$

then combining (7.16) with (7.15), and using the Martingale property of Brownian motion, Jensen's Inequality, and Cauchy-Schwarz, gives:

$$
\begin{aligned}
\mathbb{E}[ & \left.\pi_{0, \gamma}^{2}(\bar{\alpha}(j \epsilon))\|W(s)-W(j \epsilon)\|^{2}\right] \\
\leq & 2 \epsilon^{2} \mathbb{E}\left\|K_{\Delta}\left(\theta_{j}, \bar{\alpha}(j \epsilon)\right) \hat{\nabla} J\left(\theta_{j}\right) \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2} \\
& +2 \epsilon^{2} \mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(j \epsilon)) \nabla \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2} \\
& +\epsilon^{2} \sqrt{\mathbb{E}\left[\left\|\left\langle K_{\Delta}\left(\theta_{j}, \bar{\alpha}(j \epsilon)\right) \hat{\nabla} J\left(\theta_{j}\right) \pi_{0, \gamma}^{2}(\bar{\alpha}(j \epsilon)), \pi_{0, \gamma}^{2}(\bar{\alpha}(j \epsilon)) \nabla \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\rangle\right\|^{2}\right]} \\
\leq & 2 \epsilon^{2} \mathbb{E}\left\|K_{\Delta}\left(\theta_{j}, \bar{\alpha}(j \epsilon)\right) \hat{\nabla} J\left(\theta_{j}\right) \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2} \\
& +2 \epsilon^{2} \mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(j \epsilon)) \nabla \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2} \\
& +\epsilon^{2} \sqrt{\mathbb{E}\left\|K_{\Delta}\left(\theta_{j}, \bar{\alpha}(j \epsilon)\right) \hat{\nabla} J\left(\theta_{j}\right) \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2}} \\
& +\epsilon^{2} \sqrt{\mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(j \epsilon)) \nabla \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2}}
\end{aligned}
$$

Thus we have:

$$
\mathbb{E}\|(\bar{\alpha}(s)-\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon))\|^{2}
$$

$$
\begin{aligned}
\leq & 18 \epsilon^{2} \mathbb{E}\left\|K_{\Delta}\left(\theta_{j}, \bar{\alpha}(j \epsilon)\right) \frac{\beta}{2} \hat{\nabla} J\left(\theta_{j}\right)\right\|^{2} \mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2} \\
& +9 \epsilon^{2} \mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(j \epsilon)) \nabla \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2} \\
& +3 \epsilon^{2} \sqrt{\mathbb{E}\left\|K_{\Delta}\left(\theta_{j}, \bar{\alpha}(j \epsilon)\right) \frac{\beta}{2} \hat{\nabla} J\left(\theta_{j}\right)\right\|^{2} \mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2}} \\
& +3 \epsilon^{2} \sqrt{\mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(j \epsilon)) \nabla \pi_{0, \gamma}(\bar{\alpha}(j \epsilon))\right\|^{2}} \\
\leq & \epsilon^{4}\left(72 C_{0}+18\right)+\epsilon^{3}\left(6 \sqrt{C_{0}}+\sqrt{2}\right)
\end{aligned}
$$

where recall $C_{0}:=3 L_{\nabla J}^{2}\left(M_{\theta}+2 B^{2} M_{\theta}\right)+B^{2}+\zeta$. Consequently,

$$
\begin{align*}
& \beta \sum_{j=0}^{k-1} \int_{j \epsilon}^{(j+1) \epsilon} \mathbb{E}\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2} \mathbb{E}\|\nabla J(\bar{\alpha}(s))-\nabla J(\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon))\|^{2} d s \\
& \leq \beta \sum_{j=0}^{k-1} \int_{j \epsilon}^{(j+1) \epsilon} L_{\nabla J}^{2} \mathbb{E}\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2} \mathbb{E}\|(\bar{\alpha}(s)-\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon))\|^{2} d s  \tag{7.17}\\
& \leq \beta L_{\nabla J}^{2} k \epsilon\left(\epsilon^{4}\left(72 C_{0}+18\right)+\epsilon^{3}\left(6 \sqrt{C_{0}}+\sqrt{2}\right)\right)
\end{align*}
$$

Second Term: We now bound the second term as

$$
\begin{aligned}
& 3 \beta \sum_{j=0}^{k-1} \int_{j \epsilon}^{(j+1) \epsilon} \mathbb{E}\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2} \mathbb{E}\|\nabla J(\bar{\alpha}(\lfloor s / \epsilon\rfloor \epsilon))-\nabla J(\bar{\alpha}(0))\|^{2} d s \\
& \quad=3 \beta \sum_{j=0}^{k-1} \epsilon \mathbb{E}\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2} \mathbb{E}\|\nabla J(\bar{\alpha}(j \epsilon))-\nabla J(\bar{\alpha}(0))\|^{2} d s \\
& \quad \leq 3 \beta \sum_{j=0}^{k-1} \epsilon \mathbb{E}\left|\pi_{0, \gamma}(\bar{\alpha}(s))\right|^{2} \mathbb{E}\left\|\sum_{i=0}^{j-1}|\nabla J(\bar{\alpha}((i+1) \epsilon))-\nabla J(\bar{\alpha}(i \epsilon))|\right\|^{2} \\
& \quad \leq 3 \beta \sum_{j=0}^{k-1} \epsilon^{2} L_{\nabla J}^{2} \mathbb{E}\left\|\sum_{i=0}^{j-1}|\bar{\alpha}((i+1) \epsilon)-\alpha(i \epsilon)|\right\|^{2} \\
& \quad=3 \beta \sum_{j=0}^{k-1} \epsilon^{2} L_{\nabla J}^{2} \mathbb{E}\left[\sum_{i=0}^{j-1} \sum_{l=0}^{j-1}|\bar{\alpha}((i+1) \epsilon)-\bar{\alpha}(i \epsilon) \| \bar{\alpha}((l+1) \epsilon)-\bar{\alpha}(l \epsilon)|\right] \\
& \quad \leq 3 \beta \sum_{j=0}^{k-1} \epsilon^{2} L_{\nabla J}^{2}\left[\sum_{i=0}^{j-1} \sum_{l=0}^{j-1} \mathbb{E}[|\bar{\alpha}((i+1) \epsilon)-\bar{\alpha}(i \epsilon) \| \bar{\alpha}((l+1) \epsilon)-\bar{\alpha}(l \epsilon)|]\right] \\
& \quad \leq 3 \beta \sum_{j=0}^{k-1} \epsilon^{2} L_{\nabla J}^{2}\left[\sum_{i=0}^{j-1} \sum_{l=0}^{j-1}\left(\epsilon^{4}\left(72 C_{0}+18\right)+\epsilon^{3}\left(6 \sqrt{C_{0}}+\sqrt{2}\right)\right)\right] \\
& \quad=3 \beta \sum_{j=0}^{k-1} \epsilon^{2} L_{\nabla J}^{2} j^{2}\left(\epsilon^{4}\left(72 C_{0}+18\right)+\epsilon^{3}\left(6 \sqrt{C_{0}}+\sqrt{2}\right)\right) \\
& \quad \leq 3 \beta k \epsilon^{2} L_{\nabla J}^{2} k^{2}\left(\epsilon^{4}\left(72 C_{0}+18\right)+\epsilon^{3}\left(6 \sqrt{C_{0}}+\sqrt{2}\right)\right) \\
& \quad=3 \beta(k \epsilon)^{3} L_{\nabla J}^{2}\left(\epsilon^{4}\left(72 C_{0}+18\right)+\epsilon^{3}\left(6 \sqrt{C_{0}}+\sqrt{2}\right)\right)
\end{aligned}
$$

Third Term: We bound $\mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(s)) \nabla(J(\alpha(0)))\right\|^{2}$ by controlling the sampling distribution $\pi_{0, \gamma}(\cdot)$. By Lemma 7.8 and Assumption 3.1 we have

$$
\begin{align*}
\mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(s)) \nabla(J(\alpha(0)))\right\|^{2} & \leq 2 \mathbb{E}\left[\left\|\pi_{0, \gamma}(\bar{\alpha}(s))\right\|^{2}\right] \mathbb{E}\left[\|\nabla(J(\alpha(0)))\|^{2}\right] \\
& \leq 2 L_{J}^{2} \mathbb{E}\left[\left\|\pi_{0, \gamma}(\bar{\alpha}(s))\right\|^{2}\right] \leq 2 L_{J}^{2} \epsilon \tag{7.19}
\end{align*}
$$

And so

$$
\begin{aligned}
& \sum_{j=0}^{k-1} \int_{j \epsilon}^{(j+1) \epsilon} \mathbb{E}\left\|\pi_{0, \gamma}(\bar{\alpha}(s)) \nabla(J(\alpha(0)))\right\|^{2} \\
& \quad \leq 2 k \epsilon L_{J}^{2} \epsilon
\end{aligned}
$$

Fourth Term: By Lemma 7.6, we have

$$
\mathbb{E}\left\|K_{\Delta}\left(\theta_{\bar{s}}, \bar{\alpha}(s)\right) \hat{\nabla} J\left(\theta_{\bar{s}}\right)\right\|^{2} \leq 12 \epsilon\left(L_{\nabla J}^{2}\left(M_{\theta}+2 B^{2} M_{\theta}\right)+B^{2}+\zeta\right)=4 \epsilon C_{0}
$$

and so

$$
\begin{equation*}
\sum_{j=0}^{k-1} \int_{j \epsilon}^{(j+1) \epsilon} \mathbb{E}\left\|K_{\Delta}\left(\theta_{\bar{s}}, \bar{\alpha}(s)\right) \hat{\nabla} J\left(\theta_{\bar{s}}\right)\right\|^{2} d s \leq k \epsilon\left(4 \epsilon C_{0}\right) \tag{7.20}
\end{equation*}
$$

Combining (7.17), (7.18), (7.19), (7.20) in (7.14), we obtain

$$
\begin{aligned}
D & \left(\mathbb{P}_{X}^{t} \| \mathbb{P}_{A}^{t}\right) \\
\leq & \beta L_{\nabla J}^{2} k \epsilon\left(\epsilon^{4}\left(72 C_{0}+18\right)+\epsilon^{3}\left(6 \sqrt{C_{0}}+\sqrt{2}\right)\right) \\
& +3 \beta(k \epsilon)^{3} L_{\nabla J}^{2}\left(\epsilon^{4}\left(72 C_{0}+18\right)+\epsilon^{3}\left(6 \sqrt{C_{0}}+\sqrt{2}\right)\right) \\
& +2 k \epsilon L_{J}^{2} \epsilon \\
& +k \epsilon\left(4 \epsilon C_{0}\right) \\
\leq & (k \epsilon)^{3} \epsilon^{3}\left[4 \beta L_{\nabla J}^{2}\left(72 C_{0}+6 \sqrt{C_{0}}+18+\sqrt{2}\right)\right]+(k \epsilon) \epsilon\left(2 L_{J}^{2}+4 C_{0}\right)
\end{aligned}
$$

Now since $\pi_{k}=\operatorname{Law}\left(\alpha_{k}\right)$ and $\nu_{k \epsilon}=\operatorname{Law}(\alpha(t))$, the KL divergence data-processing inequality yields

$$
D\left(\pi_{k} \| \nu_{k \epsilon}\right) \leq D\left(\mathbb{P}_{X}^{t} \| \mathbb{P}_{A}^{t}\right)
$$

### 7.5. Proof of Proposition 5.3.

Proof. Recall that the continuous time diffusion of interest (2.12) has infinitesimal generator $\mathcal{L}$ acting on $C^{2}$ function $f$ as

$$
\mathcal{L} f=\frac{1}{2} \pi_{0, \gamma}^{2} \Delta f-\frac{\beta}{2} \pi_{0, \gamma}\langle\nabla J, \nabla f\rangle-\pi_{0, \gamma}\left\langle\nabla \pi_{0, \gamma}, \nabla f\right\rangle
$$

We will show that the conditions of Proposition 4.4 hold:

1. Consider the Lyapunov function

$$
V(x)=\exp \left(\frac{\beta m\|x\|^{2}}{2\left(\bar{\pi}_{0, \gamma}^{2}+1\right)}\right)
$$

Then we have

$$
\begin{aligned}
& \mathcal{L} V(x)=-\frac{\beta}{2} \pi_{0, \gamma}^{2}(x)\langle\nabla J(x), \nabla V(x)\rangle-\pi_{0, \gamma}(x)\left\langle\nabla \pi_{0, \gamma}(x), \nabla V(x)\right\rangle \\
& +\frac{1}{2} \pi_{0, \gamma}^{2}(x) \Delta V(x) \\
& =\left\{-\frac{\beta}{2} \frac{m \beta}{\left(\bar{\pi}_{0, \gamma}^{2}+1\right)} \pi_{0, \gamma}^{2}(x)\langle\nabla J(x), x\rangle-\frac{m \beta}{\bar{\pi}_{0, \gamma}^{2}+1} \pi_{0, \gamma}(x)\left\langle\nabla \pi_{0, \gamma}(x), x\right\rangle\right. \\
& \left.\left.+\frac{1}{2} \pi_{0, \gamma}^{2}(x)\left(\frac{m \beta N}{\bar{\pi}_{0, \gamma}^{2}+1}+\left(\frac{m \beta}{\bar{\pi}_{0, \gamma}^{2}+1}\right)^{2}\|x\|^{2}\right)\right)\right\} V(x) \\
& \leq\left\{-\frac{\beta}{2} \frac{m \beta}{\bar{\pi}_{0, \gamma}^{2}+1} \pi_{0, \gamma}^{2}(x)\left(m\|x\|^{2}-b\right)+\frac{m \beta I}{\bar{\pi}_{0, \gamma}^{2}+1} \pi_{0, \gamma}(x)\right. \\
& \left.\left.+\frac{1}{2} \pi_{0, \gamma}^{2}(x)\left(\frac{m \beta N}{\bar{\pi}_{0, \gamma}^{2}+1}+\left(\frac{m \beta}{\bar{\pi}_{0, \gamma}^{2}+1}\right)^{2}\|x\|^{2}\right)\right)\right\} V(x) \\
& \leq\left\{\left(\frac{1}{2} \beta m N+m \beta I+\frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{\beta^{2} m b}{\bar{\pi}_{0, \gamma}^{2}+1}\right)\right. \\
& \left.-\left(\frac{\beta^{2} m}{2\left(\bar{\pi}_{0, \gamma}^{2}+1\right)} \pi_{0, \gamma}^{2}(x) m-\frac{1}{2} \pi_{0, \gamma}^{2}(x)\left(\frac{\beta m}{\left(\bar{\pi}_{0, \gamma}^{2}+1\right)}\right)^{2}\right)\|x\|^{2}\right\} V(x) \\
& \leq\left\{\left(\frac{1}{2} \beta m N+m \beta I\right)+\frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{\beta^{2} m b}{\bar{\pi}_{0, \gamma}^{2}+1}\right. \\
& \left.-\left(\frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1}-\frac{1}{2} \pi_{0, \gamma}^{2}(x)\left(\frac{\beta m}{\bar{\pi}_{0, \gamma}^{2}+1}\right)^{2}\right)\|x\|^{2}\right\} V(x) \\
& \leq\left\{\left(\frac{1}{2} \beta m N+m \beta I\right)+\frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{\beta^{2} m b}{\bar{\pi}_{0, \gamma}^{2}+1}\right. \\
& \left.-\left(\frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1}\left(1-\frac{1}{\bar{\pi}_{0, \gamma}^{2}+1}\right)\right)\|x\|^{2}\right\} V(x) \\
& \leq\left\{\left(\frac{1}{2} \beta m N+m \beta I\right)+\left[\frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{\beta^{2} m b}{\bar{\pi}_{0, \gamma}^{2}+1}\right.\right. \\
& +\frac{1}{2} \pi_{0, \gamma}^{2}(x)\left(\frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1} M^{2}\right)+1 \\
& \left.\left.-\left(\frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1}\left(1-\frac{1}{\bar{\pi}_{0, \gamma}^{2}+1}\right)\right)\|x\|^{2}\right]\right\} V(x)
\end{aligned}
$$

where in the last inequality statement we append

$$
\frac{1}{2} \pi_{0, \gamma}^{2}(x)\left(\frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1} M^{2}\right)+1
$$

for the following reason: By assumption 3.4 we have that $\exists M: \pi_{0, \gamma}(x)<$ $\frac{2}{(\beta m)^{2}\|x\|^{2}} \forall\|x\|>M, \gamma \leq 1$. We aim to show that the term inside brackets in the last inequality line of (1) is positive for all $x$. First take $\|x\|<M$ : we
have that:

$$
\left(\frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1}\left(1-\frac{1}{\bar{\pi}_{0, \gamma}^{2}+1}\right)\right)\|x\|^{2}<\frac{1}{2} \pi_{0, \gamma}^{2}(x)\left(\frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1} M^{2}\right)
$$

Now consider $\|x\| \geq M$. By Lemma 7.9 we have

$$
\begin{aligned}
& \frac{1}{2} \pi_{0, \gamma}(x)(\beta m)^{2}\|x\|^{2}<1 \\
& \Rightarrow \frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1}\|x\|^{2}<1 \\
& \Rightarrow \frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1}\left(1-\frac{1}{\bar{\pi}_{0, \gamma}^{2}+1}\right)\|x\|^{2}<1
\end{aligned}
$$

Thus we have:

$$
\begin{aligned}
& {\left[\frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{\beta^{2} m b}{\bar{\pi}_{0, \gamma}^{2}+1}+\frac{1}{2} \pi_{0, \gamma}^{2}(x)\left(\frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1} M^{2}\right)+1\right.} \\
& \left.\quad-\left(\frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1}\left(1-\frac{1}{\bar{\pi}_{0, \gamma}^{2}+1}\right)\right)\|x\|^{2}\right]>0 \quad \forall x \in \mathbb{R}^{N}
\end{aligned}
$$

and now observe that

$$
\begin{aligned}
{[ } & \frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{\beta^{2} m b}{\bar{\pi}_{0, \gamma}^{2}+1}+\frac{1}{2} \pi_{0, \gamma}^{2}(x)\left(\frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1} M^{2}\right)+1 \\
& \left.-\left(\frac{1}{2} \pi_{0, \gamma}^{2}(x) \frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1}\left(1-\frac{1}{\bar{\pi}_{0, \gamma}^{2}+1}\right)\right)\|x\|^{2}\right] \\
\leq & {\left[\frac{1}{2}\left(\pi_{0, \gamma}^{2}(x)+1\right) \frac{\beta^{2} m b}{\bar{\pi}_{0, \gamma}^{2}+1}+\frac{1}{2}\left(\pi_{0, \gamma}^{2}(x)+1\right)\left(\frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1} M^{2}\right)\right.} \\
& \left.-\left(\frac{1}{2}\left(\pi_{0, \gamma}^{2}(x)+1\right) \frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1}\left(1-\frac{1}{\bar{\pi}_{0, \gamma}^{2}+1}\right)\right)\|x\|^{2}\right] \\
\leq & \left(\pi_{0, \gamma}^{2}(x)+1\right)\left[\frac{1}{2} \frac{\beta^{2} m b}{\bar{\pi}_{0, \gamma}^{2}+1}+\frac{1}{2}\left(\frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1} M^{2}\right)\right. \\
& \left.-\left(\frac{1}{2} \frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1}\left(1-\frac{1}{\bar{\pi}_{0, \gamma}^{2}+1}\right)\right)\|x\|^{2}\right] \\
\leq & \left(\bar{\pi}_{0, \gamma}^{2}+1\right)\left[\frac{1}{2} \frac{\beta^{2} m b}{\bar{\pi}_{0, \gamma}^{2}+1}+\frac{1}{2}\left(\frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1} M^{2}\right)\right. \\
& \left.-\left(\frac{1}{2} \frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1}\left(1-\frac{1}{\bar{\pi}_{0, \gamma}^{2}+1}\right)\right)\|x\|^{2}\right]
\end{aligned}
$$

where we use that $1 \leq \frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}+1} M^{2}$, which is derived from Assumption (3.8). Thus we have:

$$
\frac{\mathcal{L} V(x)}{V(x)} \leq\left\{\left(\frac{1}{2} \beta m N+m \beta I\right)\right.
$$

$$
\begin{aligned}
& +\left(\bar{\pi}_{0, \gamma}^{2}+1\right)\left[\frac{1}{2} \frac{\beta^{2} m b}{\bar{\pi}_{0, \gamma}^{2}+1}+\frac{1}{2}\left(\frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1} M^{2}\right)\right] \\
& \left.-\left(\bar{\pi}_{0, \gamma}^{2}+1\right)\left(\frac{1}{2} \frac{(\beta m)^{2}}{\bar{\pi}_{0, \gamma}^{2}+1}\left(1-\frac{1}{\bar{\pi}_{0, \gamma}^{2}+1}\right)\right)\|x\|^{2}\right\} \\
& \leq\left(\frac{1}{2} \beta m N+\beta m I\right)+\frac{1}{2}\left[\beta^{2} m b+(\beta m M)^{2}\right] \\
& -\frac{1}{2}\left((\beta m)^{2}+\left(1-\frac{1}{\text { idistmax }^{2}+1}\right)\right)\|x\|^{2} \\
& :=\kappa-\gamma\|x\|^{2}
\end{aligned}
$$

and observe that $\kappa, \gamma>0$. Thus, condition (1) of Proposition 4.4 holds.
2. From (1), we have

$$
\frac{\mathcal{L} V(x)}{V(x)} \leq \kappa-\gamma\|x\|^{2}
$$

Observe that this implies

$$
\frac{\mathcal{L} V(x)}{V(x)} \leq-\kappa+2 \kappa \mathbf{1}_{\left(\|x\|^{2} \leq 2 \kappa / \gamma\right)}
$$

Moreover, by Lemma 7.10 and since $J(x) \geq 0$, we have

$$
O_{r}(\beta J) \leq \beta\left(\frac{L_{\nabla J}}{2}\|x\|^{2}+\|x\|+A\right) \leq \beta\left(\frac{\left(L_{\nabla J}+B\right) r^{2}}{2}+A+B\right)
$$

and so by Proposition 4.3, with $\kappa_{0}=2 \kappa, \zeta_{0}=\kappa, r^{2}=2 \kappa / \gamma$, we have that $\pi_{\infty}$ satisfies a Poincarê inequality with constant

$$
c_{P}=\frac{1}{\lambda} \leq \frac{1}{2 \kappa}\left(1+\frac{4 C \kappa^{2}}{\gamma} \exp \left(\beta\left(\frac{\left(L_{\nabla J}+B\right) \kappa}{\gamma}+A+B\right)\right)\right)
$$

where $\kappa$ and $\gamma$ are defined above and provided in (5.6).
3. By assumption 3.1, we have

$$
\nabla^{2} \beta J \succeq-\beta L_{\nabla J} I_{d} \succeq 0
$$

Thus the conditions of Proposition 4.4 are met, with $K=0$. So, letting $\zeta=1$ :

$$
Z_{1}=\frac{2}{\gamma}+1, \quad Z_{2}=\frac{2}{\gamma}\left(\kappa+\gamma \int_{\mathbb{R}^{N}}\|w\|^{2} \pi_{\infty}(d w)\right)
$$

Now we would like to make the bound on $c_{L S}$ (5.5) more explicit by providing a bound on $\int_{\mathbb{R}^{N}}\|w\|^{2} \pi_{\infty}(d w)$. From (5.8) we have $\mathcal{W}_{2}\left(\nu_{t}, \pi_{\infty}\right) \rightarrow 0$ as $t \rightarrow \infty$, and thus by Theorem 7.12 of [27] and Lemma 7.5 it follows that (with $\gamma \leq 1$ )

$$
\begin{align*}
& \int_{\mathbb{R}^{N}}\|w\|^{2} \pi_{\infty}(d w)=\lim _{t \rightarrow \infty} \int_{\mathbb{R}^{N}}\|w\|^{2} \nu_{t}(d w) \\
& \quad \leq \kappa_{0}^{\gamma}+\frac{(\beta b+N) \bar{\pi}_{0, \gamma}+2 I}{(m \beta) \bar{\pi}_{0, \gamma}} \leq \kappa_{0}+\frac{(\beta b+N) \bar{\pi}+2 I}{(m \beta) \bar{\pi}} \tag{7.21}
\end{align*}
$$

So now letting

$$
Z_{2}=\frac{2}{\gamma}\left(\kappa+\gamma\left(\kappa_{0}+\frac{(\beta b+N) \bar{\pi}+2 I}{(m \beta) \bar{\pi}}\right)\right)
$$

we have that $\pi_{\infty}$ satisfies a log-Sobolev inequality with constant

$$
c_{L S}=Z_{1}+\left(Z_{2}+2\right) c_{P}
$$


[^0]:    *The results in this paper have appeared in the Proceedings of the 2023 IEEE Conference on Decision and Control in reduced and abbreviated form. This manuscript substantially expands on the exposition of the Conference version, provides all proofs, and includes more detailed examples, discussion, and mathematical details.

    Funding: This research was supported in part by the National Science Foundation grants CCF2112457 and CCF-2312198, Army Research Office grant W911NF-19-1-0365, and Air Force Office of Scientific Research grant FA9550-22-1-0016
    ${ }^{\dagger}$ Department of Electrical \& Computer Engineering, Cornell University, Ithaca, NY

[^1]:    ${ }^{1}$ see [16] for motivation
    ${ }^{2}$ This construction is purely for notational convenience.

[^2]:    ${ }^{3}$ An example kernel function is the multivariate normal $\mathcal{N}\left(0, \sigma^{2} I_{N}\right)$ density with $\sigma=\Delta$, i.e., $\frac{1}{\Delta^{N}} K\left(\frac{\theta-\alpha}{\Delta}\right)=(2 \pi)^{-N / 2} \Delta^{-N} \exp \left(-\frac{\|\theta-\alpha\|^{2}}{2 \Delta^{2}}\right)$

