2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

# Finite-Sample Bounds for Adaptive Inverse Reinforcement Learning using Passive Langevin Dynamics

Luke Snow, Vikram Krishnamurthy

***Abstract*— Stochastic gradient Langevin dynamics (SGLD) are a useful methodology for sampling from probability distributions. This paper provides a finite sample analysis of a *passive* stochastic gradient Langevin dynamics algorithm (PSGLD) designed to achieve inverse reinforcement learning. 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). The PSGLD algorithm acts as a randomized sampler which recovers the cost function being optimized by this external process. Previous work has analyzed the asymptotic performance of this passive algorithm using stochastic approximation techniques; in this work we analyze the *non-asymptotic* performance. Specifically, we provide finite-time bounds on the 2-Wasserstein distance between the passive algorithm and its stationary measure, from which the reconstructed cost function is obtained.**

## I. Introduction

We derive non-asymptotic (finite-sample) bounds for a Langevin dynamics algorithm performing real-time inverse reinforcement learning (IRL). Traditional IRL [1], [2], [3] 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 performing stochastic gradient descent (e.g, policy gradient reinforcement learning) on a cost function $J$, and attempt to reconstruct $J$ in real-time.

To accomplish real-time IRL, we employ a *passive* stochastic gradient Langevin dynamics (PSGLD) algorithm. Given observations of an agent's sequential stochastic gradient descent (SGD) evaluations on $J$, the PSGLD algorithm acts as a Markov chain Monte Carlo (MCMC) sampler designed to reconstruct $J$. The algorithm relies on stochastic gradient Langevin dynamics [4], [5], which has emerged as a general MCMC technique for sampling from probability distributions. The algorithm is considered *passive* because the sequential stochastic gradients are not directly controlled, but are provided by the observed SGD process. Thus, this technique can be considered an inverse stochastic gradient algorithm. It can apply to IRL problems in a variety of contexts, such as adaptive Bayesian learning, constrained Markov Decision Processes, and logistic regression classification [6].

The PSGLD algorithm we consider was initially proposed in [6], in which stochastic approximation arguments were used to show that the algorithm asymptotically samples from the Gibbs measure encoding the cost function. Similar passive schemes and stochastic approximation analyses have been investigated in [7], [8], [9]. In this work we present a *non-asymptotic* analysis of this PSGLD algorithm; we provide finite-time 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 [10], [11], [12]. In our case the algorithm is *passive*; so our analysis generalizes and extends previous works to handle this complexity. To obtain our bound, we decompose the desired 2-Wasserstein distance into the sum of distances between the law of the PSGLD algorithm and a particular continuous time diffusion, and that between the diffusion and its stationary Gibbs measure. The former bound relies on a Girsanov-type change of measure technique and a weighted transportation cost inequality, as in [10]. To obtain the latter bound we 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 convergence in 2-Wasserstein distance.

The paper is organized as follows: Section II provides background on passive stochastic gradient Langevin dynamics. Section III discusses our main results, namely, a non-asymptotic 2-Wasserstein bound. In section IV we provide additional background for the proof of our bound, and in section V provides further proof details.

## II. Passive Langevin Dynamics

In this section we first present the PSGLD algorithm and its setting. We then discuss recent work providing asymptotic guarantees for this algorithm, and motivate the non-asymptotic analysis to follow.

Consider a forward learning agent running a stochastic gradient descent (SGD) to minimize a cost function $J : \mathbb{R}^N \to \mathbb{R}_+$. We aim to observe this process and

This research was supported in part by the National Science Foundation grants CCF-2112457 and CCF-2312198, Army Research Office grant W911NF-19-1-0365, and Air Force Office of Scientific Research grant FA9550-22-1-0016

Luke Snow `las474@cornell.edu`, and Vikram Krishnamurthy `vikramk@cornell.edu` are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA

reconstruct $J$. In order to observe sufficient richness of gradient samples from this cost function, we assume the stochastic gradient algorithm resets after some finite time. Thus we have, for $n \in \mathbb{N}$ representing each "run" of the SGD, and $\tau_n$ stopping times:

$$\theta_{k+1} = \theta_k - \eta \hat{\nabla} J(\theta_k), \quad k \in \{\tau_n, \dots, \tau_{n+1} - 1\} \tag{1}$$

where each $\theta_{\tau_n} \sim \pi_{0,\lambda}$ and $\eta > 0$ is some step-size. Here $\pi_{0,\lambda}$ is the "sampling distribution" with scale parameter $\lambda$, defined as

$$\pi_{0,\lambda}(x) = \frac{\pi_0(x/\lambda)}{\int_{\mathbb{R}^N} \pi_0(x/\lambda) dx} \tag{2}$$

for some base distribution $\pi_0$ on $\mathbb{R}^N$. $\hat{\nabla} J(\theta_k)$ is an unbiased estimate of the gradient $\nabla J(\theta_k)$, with bounded variance, see III-B. Algorithm 1 displays this randomly re-initializing stochastic gradient descent.

We consider an inverse learning agent who observes the SGD process, and attempts to reconstruct the cost function $J$ being optimized. We assume the observer knows the initialization density $\pi_{0,\lambda}$ and the step size $\eta$, and can observe evaluations $\theta_k, k \in \mathbb{N}$. The agent then performs the following *passive stochastic gradient Langevin dynamics* update:

$$\begin{aligned} \alpha_{k+1} = \alpha_k - \epsilon \Bigg[ & \frac{1}{\Delta^N} K\left(\frac{\theta_k - \alpha_k}{\Delta}\right) \frac{\beta}{2} \hat{\nabla} J(\theta_k) \\ & + \nabla \pi_{0,\lambda}(\alpha_k) \Bigg] \pi_{0,\lambda}(\alpha_k) + \sqrt{\epsilon} \pi_{0,\lambda}(\alpha_k) w_k \end{aligned} \tag{3}$$

initialized by $\alpha_0 \sim \pi_{0,\lambda}$. $\Delta$ is a constant step size parameter, $\{w_k, k \geq 0\}$ is an i.i.d. sequence of standard $N-$variate Gaussian random variables, $\beta$ is the inverse temperature parameter, and $K\left(\frac{\theta_k - \alpha_k}{\Delta}\right)$ is a kernel function weighting the relevance of the stochastic gradient $\hat{\nabla} J(\theta_k)$ to the current update $\alpha_{k+1}$. Algorithm 2 displays this passive stochastic gradient Langevin dynamics algorithm, which takes as input the sequential evaluations $\theta_k$ made in Algorithm 1.

The kernel function $K(\cdot)$ is a key element of this passive scheme. Since we do not know $J(\cdot)$ we cannot evaluate $\nabla J(\alpha_k)$, and so we instead employ the estimator $\hat{\nabla} J(\theta_k)$ obtained by observing the SGD (1). Thus, we want to weight our algorithm's dependence on this *biased* estimator by the proximity between evaluations $\theta_k$ and $\alpha_k$ through the kernel function. The kernel[1] can be chosen by the observer as any function $K : \mathbb{R}^N \to \mathbb{R}$ satisfying:

$$\begin{aligned} & K(u) \geq 0, \quad K(u) = K(-u), \quad \sup_u K(u) < \infty, \\ & \int_{\mathbb{R}^N} K(u) du = 1, \quad \int_{\mathbb{R}^N} |u|^2 K(u) < \infty \end{aligned} \tag{4}$$

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

The idea behind Algorithm 2, developed in [6], is that in the asymptotic limit the samples $\alpha_k$ are generated according to the Gibbs measure

$$\pi_\infty(\alpha) := \frac{\exp(-\beta J(\alpha))}{Z}, \quad \alpha \in \mathbb{R}^N \tag{5}$$

where $Z = \int_{\mathbb{R}^N} \exp(-\beta J(\alpha)) d\alpha$ is a normalizing constant. Thus, the true cost function $J(\cdot)$ driving the SGD (1) can be recovered by taking the log-density of asymptotic Markov chain Monte Carlo samples. This idea is presented as the following informal result, see [6] for more details.

**Proposition 1** (Weak Convergence Analysis [6])**.** *Let $\alpha^\epsilon(t) = \alpha_k$ for $t \in [\epsilon k, \epsilon(k+1)]$ be the continuous-time interpolation of Algorithm 2. Under assumptions (A1)-(A4) of [6], the process $\alpha^\epsilon(t)$ converges weakly to the solution of the stochastic differential equation*

$$\begin{aligned} & d\alpha(t) = \pi_{0,\lambda}(\alpha(t))\, dW(t) \\ & + \left[ \nabla \pi_{0,\lambda}(\alpha(t)) \pi_{0,\lambda}(\alpha(t)) - \frac{\beta}{2} \pi_{0,\lambda}^2(\alpha(t)) \nabla J(\alpha(t)) \right] dt \\ & \alpha(0) = \alpha_0 \sim \pi_{0,\lambda} \end{aligned} \tag{6}$$

*where $W(t)$ is standard $N$-dimensional Brownian motion. Furthermore, The stochastic differential equation* (6) *has* $\pi_\infty$ (5) *as its stationary distribution.*

**Motivation**: Proposition 1 shows that Algorithm 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, for any practical implementation it should be quantified how well this sampling algorithm approximates the Gibbs measure after a *finite run-time*. Our main result provides such non-asymptotic (finite-time) guarantees.

**Algorithm 1** Randomly Re-Initializing SGD Process

```
initialize τ_0 = 1, k = 1
while n ≥ 0 do
    generate τ_{n+1} > τ_n,   θ_{τ_n} ~ π_{0,λ}
    for k = τ_n : τ_{n+1} − 1 do
        θ_{k+1} ← θ_k − η ∇̂J(θ_k)
    end for
end while
```

**Algorithm 2** PSGLD

```
initialize α_1 ~ π_{0,λ}
while k ≥ 1 do
    obtain θ_k from Algorithm 1
    if k ≥ 2 then
        β ∇̂J(θ_k) = (1/ε)(θ_k − θ_{k−1}),   K̂ = (1/Δ^N) K((θ_k − α_k)/Δ)
        sample w_k ~ N(0, I_N)
        α_{k+1} ← α_k − ε [ K̂ (β/2) ∇̂J(θ_k) + ∇π_0(α_k) ] π_0(α_k)
        α_{k+1} ← α_{k+1} + √ε π_0(α_k) w_k
    end if
end while
```

## III. Main Result. Non-Asymptotic Analysis

In this section we construct finite-sample bounds on the 2-Wasserstein distance between the sample density produced by Algorithm 2 and the Gibbs measure (5) encoding the cost function $J$. We provide a brief overview of the 2-Wasserstein metric and the non-asymptotic analysis techniques, specify assumptions on the cost function $J$ and initial distribution $\pi_{0,\lambda}$, provide the main bound in the form of Theorem 1 , and briefly discuss the application to adaptive inverse reinforcement learning.

### A. 2-Wasserstein Distance

We provide a non-asymptotic bound on the convergence of (3) to the Gibbs measure $\pi_\infty$ (5), in terms of the *2-Wasserstein distance*:

$$\mathcal{W}_2(\mu,\nu) := \inf_{\gamma\in\Gamma(\mu,\nu)} \left(\mathbb{E}_{(x,y)\sim\gamma}\|x-y\|^2\right)^{1/2}$$

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.,

$$\gamma(A,\mathbb{R}^N)=\mu(A),\quad \gamma(\mathbb{R}^N,B)=\nu(B),\quad \forall A,B\in\mathcal{B}(\mathbb{R}^N)$$

where $\mathcal{B}(\mathbb{R}^N)$ is the Borel $\sigma$-algebra of $\mathbb{R}^N$.

Letting

$$\pi_k := \mathrm{Law}(\alpha_k),\quad \nu_{k\epsilon} := \mathrm{Law}(\alpha(k\epsilon))$$

be the distributions of the sampling density produced by iterates $\alpha_k$ (3) and the continuous time diffusion (6), respectively, we bound

$$\mathcal{W}_2(\pi_k,\pi_\infty) \le \mathcal{W}_2(\pi_k,\nu_{k\epsilon}) + \mathcal{W}_2(\nu_{k\epsilon},\pi_\infty)$$

i.e., by simple triangle inequality we can first control the distance between the law of the discretization (3) and that of the continuous-time diffusion (6), then control the convergence of the continuous time diffusion to its stationary measure $\pi_\infty$. Section V provides further details on the methods for obtaining these bounds.

Simple non-asymptotic convergence bounds for Markov diffusions have been been established in [13] in terms of total-variation norm. However, recent works [11], [10] study the convergence in 2-Wasserstein distance; this is a more suitable metric for assessing the quality of approximate sampling schemes since it gives direct guarantees on the accuracy of approximating higher order moments [11].

### B. Assumptions

Here we make several assumptions on the cost function $J$ and the base sampling distribution $\pi_0$. Recall that the sampling distribution $\pi_{0,\lambda}$ (2) is simply a scaled version of the base distribution $\pi_0$. Throughout the paper we will use $\|\cdot\|$ to denote the $l_2$ norm.

**A 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 $x,y\in\mathbb{R}^N$,*

$$\|J(x)-J(y)\| \le L_J\|x-y\|$$
$$\|\nabla J(x)-\nabla J(y)\| \le L_{\nabla J}\|x-y\|$$

**A 2** (Dissipativity)**.** *$J$ is $(m,b)$-dissipative:*

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

**A 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 $\zeta\ge 0$ such that*

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

**A 4** ($\pi_0$ exponential decay)**.** *There exists $M\in\mathbb{N},\tilde{C}>0$ such that for all $\|x\|>M$*

$$\pi_0(x)\le\exp(-\|x\|^2),\quad \|\nabla\pi_0(x)\|\le\frac{\tilde{C}}{\|x\|}$$

**A 5** ($\pi_0$ Lipschitz-continuity)**.**

$$\exists L_{\pi_0}>0 : \|\pi_0(x)-\pi_0(y)\|\ \le L_{\pi_0}\|x-y\|\ \forall x,y\in\mathbb{R}^N$$

**A 6** ($\pi_0$ unimodal)**.** *The sampling distribution $\pi_0$ is unimodal, and has $\sup_x\pi_0(x)=1$*

**A 7** (kernel structure)**.** *The kernel function $K(\cdot)$ satisfies (4).*

**A 8** (feasible parameter ranges)**.** *Here $\wedge$ denotes the min operator and $\vee$ the max operator. Assume*

*i)* $\eta\in(0,1\wedge\frac{m}{4L_{\nabla J}^2})$

*ii)* $\epsilon\in\left(0,1\wedge\sqrt{\frac{1}{249}}L_{\nabla J}^{-1}\right)$

*iii)* $\beta\ge\frac{1}{4L_{\nabla J}^2}\vee\frac{\sqrt{2\pi+4}}{m\sqrt{L_{\nabla J}}}$

Assumptions on $J$ (A1-A3) are standard and equivalent to those in [10]. Assumptions on the base sampling distribution $\pi_0$ hold for a wide class of probability density functions, including Gaussian densities. A7 admits a wide range of kernel functions, including Gaussian densities. Range specifications on $\epsilon,\beta$ in A8 can be satisfied once a feasible range for the Lipschitz constant $L_{\nabla J}$ is known to the inverse learner.

Notice that the feasible range for $\eta$ can always be satisfied; the SGD process (1) optimizing cost function $J$ with step $\hat{\eta}\ge(1\wedge\frac{m}{4L_{\nabla J}^2})$ is equivalent to another SGD with step $\eta<\frac{m}{4L_{\nabla J}^2}$ which optimizes $\frac{\eta}{\hat{\eta}}J$. So assuming $\eta$ satisfies A8 we can sample from $\pi_\infty\propto\exp(-\frac{\eta}{\hat{\eta}}\beta J)$, from which $J$ can be recovered since the scale $\frac{\eta}{\hat{\eta}}\beta$ disappears upon MCMC sample measure normalization.

### C. Main Result. Finite-Sample Bound

Recall that $\pi_k$ is the distribution of $\alpha_k$ in Algorithm 2, $\nu_{k\epsilon}$ is the distribution of $\alpha(t)$ at time $t=k\epsilon$ in diffusion (6), and $\pi_\infty\propto\exp(-\beta J(\alpha))$ is the Gibbs measure (5) encoding the cost function $J$ we aim to reconstruct.

We present our Wasserstein bound in a way that explicitly depends on a hyperparameter $\delta$: $\mathcal{W}_2(\pi_k,\pi_\infty)\le f(\delta)$ for some function $f$ which is monotonically increasing and has $\lim_{\delta\to 0}f(\delta)=0$. Our main result is that for any arbitrarily small $f(\delta)$, we can choose the step size $\epsilon$, algorithmic iterations $k$, kernel scale parameter $\Delta$

and sampling distribution scale parameter $\lambda$ *as follows* to achieve $\mathcal{W}_2(\pi_k, \pi_\infty) \leq f(\delta)$.

$$k\epsilon \geq \beta\, c_{LS} \log\left(\frac{1}{\delta}\right) \qquad \epsilon \leq \left(\frac{\delta}{\log\left(\frac{1}{\delta}\right)}\right)^2$$
$$\Delta \leq \inf_{x \in [\epsilon, \hat{K}_\epsilon]} \frac{K^{-1}(\frac{\hat{K}_1\sqrt{2\pi}}{2\epsilon} e^{x^2/2})}{K^{-2}(x\epsilon^{2N})} \qquad \lambda \in [\epsilon^2, \epsilon^{3/2}] \tag{7}$$

where $c_{LS}$ is the logarithmic-Sobolev constant of diffusion (6), explicitly bounded in (25). $K^{-1}$ denotes the inverse of $K$ and $K^{-2}$ denotes the inverse of $K^2$, both mapping to the non-negative orthant, and for general $\alpha \in \mathbb{R}_+$, $K_\alpha(\cdot) := \frac{1}{\alpha^N} K(\frac{\cdot}{\alpha})$, $\hat{K}_\alpha := \sup_{x \in \mathbb{R}^N} K_\alpha(x)$. So $\hat{K}_1 := \sup_x K(x)$.

**Theorem 1** (Finite-Sample 2-Wasserstein Bound)**.** *Consider the PSGLD Algorithm 2 with iterates $\alpha_k \in \mathbb{R}^N$. For any*

$$\delta \in \left[0, \exp\left(-\frac{1}{\beta c_{LS}}\right)\right] \tag{8}$$

*choose step size $\epsilon$, number of iterations $k$, kernel scale $\Delta$ and sampling distribution scale $\lambda$ according to* (7)*. Then, under assumptions (A1)-(A8), at iterate $k$ the 2-Wasserstein distance between the distribution $\pi_k$, generated by the PSGLD algorithm, and the Gibbs measure $\pi_\infty$* (5)*, satisfies:*

$$\mathcal{W}_2(\pi_k, \pi_\infty) \leq \delta\left[C_4 + \sqrt{2c_{LS}C_3}\right] + \delta\sqrt{10 c_{LS} N \log(1/\delta)} \tag{9}$$

*where $C_3, C_4$ are constants dependent on structural specifications of $J$ and the process* (3)*, and are provided explicitly in Appendix VII. $c_{LS}$ is the logarithmic-Sobolev constant bounded explicitly in Proposition 4.*

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

$$\lim_{\delta \to 0} \delta\sqrt{\alpha \log(1/\delta)} = 0$$

So, $\mathcal{W}_2(\pi_k, \pi_\infty)$ is monotonically increasing in $\delta$ for $\delta \in (0, 0.607)$ and

$$\lim_{\delta \to 0} \mathcal{W}_2(\pi_k, \pi_\infty) = 0$$

Thus, Theorem 1 asserts that, through hyperparameter $\delta$, we can take the number of iterations $k$ large enough, step size $\epsilon$, kernel parameter $\Delta$ and sampling distribution scale $\lambda$ small enough, in accordance with (7), such that the PSGLD algorithm (3) is within any arbitrarily small desired 2-Wasserstein distance (9) to the Gibbs measure (5). Here $\delta$ acts as a precision parameter; smaller $\delta$ yields a tighter approximation (9) at the expense of larger number of iterations $k$ and smaller step size $\epsilon$, kernel scale $\Delta$ and sampling distribution scale $\lambda$.

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 $\delta \to 0$.

### D. Example: Adaptive Inverse Reinforcement Learning

As an example, we now briefly discuss how adaptive IRL for an infinite horizon discounted cost Markov Decision Process (MDP) fits into our framework. Let $\{x_n\}$ denote a finite state Markov chain with controlled transition probabilities $P_{ij}(u) = \mathbb{P}[x_{n+1} = j | x_n = i, u_n = u]$ where action $u_n$ is chosen from policy $\boldsymbol{u}_\theta$ parametrized by $\theta$ as $u_n = \boldsymbol{u}_\theta(x_n)$. Solving a discounted average cost MDP requires computing the optimal parameter $\theta^* = \min\{\theta : J(\theta)\}$ where the cumulative cost is

$$J(\theta) = \lim_{N \to \infty} \mathbb{E}_\theta\left[\sum_{n=1}^{N} \gamma^n \rho(x_n, u_n) | x_0 = x\right]$$

Here $u_n = \boldsymbol{u}_\theta(x_n)$, $\gamma \in (0,1)$ is the discount factor, and $\rho(x_n, u_n)$ is the cost of taking action $u_n$ in state $x_n$.

One canonical scheme for achieving this minimum is the REINFORCE algorithm, which proceeds by evaluating sequential sample trajectories $\{s_0, a_0 \ldots, s_T, a_T;\, \theta_k\}$ under policy $\boldsymbol{u}_{\theta_k}(\cdot)$, and updating $\theta_k$ as

$$\begin{aligned}\theta_{k+1} &= \theta_k - \eta \sum_{t=0}^{T}\left[\gamma^t \nabla_\theta \log \pi(s_t, a_t; \theta) \sum_{k=t}^{T} \gamma^{k-t}\, \rho(s_k, a_k)\right] \\ &= \theta_k - \eta \hat{\nabla} J(\theta)\end{aligned}$$

where $\hat{\nabla} J(\theta)$ is an unbiased estimate of $\nabla J(\theta)$ by the Policy Gradient Theorem [14].

Suppose the forward learner runs such a policy gradient algorithm to obtain $\theta^* = \arg\min J(\theta)$. Given sequential observations of the estimates $\theta_k$, through e.g., state-action trajectories $\{s_0, a_0 \ldots, s_T, a_T;\, \theta_k\}$, our PSGLD algorithm can approximate, within 2-Wasserstein distance (9) depending on parameter specifications (7), the Gibbs measure $\exp(-\beta J(\theta))$ through Markov chain Monte Carlo sampling. So the cost function $J(\theta)$ can be recovered by taking the logarithm of the sample density.

Note that traditional IRL methods aim to reconstruct $\rho(x, a)$, rather than $J(\theta)$, given *optimal* policy demonstrations. In our case $\rho(x, a)$ 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 $\rho(x, a)$ with respect to the stationary measure induced by the policy $\boldsymbol{u}_\theta(\cdot)$ and the dynamics $P_{ij}(u)$. Furthermore, in contrast to traditional methods [15], [1], we operate in the transient regime where the observed agent is *in the process of learning* an optimal policy.

See [6] for more details on how a more broad class of RL frameworks fit into this work, and [16] for rigorous details on structural estimation and identification of MDPs.

## IV. Stochastic Differential Equation Preliminaries

Here we provide mathematical background on diffusion processes and functional inequalities which are indispensable to the proof of our bound (9).

*1) Infinitesimal Generator:* First we state some background on the infinitesimal generator of an Itô diffusion. Let $X_t$ be the $\mathbb{R}^n$-valued Itô diffusion solving the stochastic differential equation

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dW(t), \quad X_0 = x \in \mathbb{R}^n \tag{10}$$

where $b : \mathbb{R}^n \to \mathbb{R}^n$ is the drift function, $\sigma : \mathbb{R}^n \to \mathbb{R}$ is the diffusion 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 $f : \mathbb{R}^n \to \mathbb{R}$, in domain $\mathcal{D}(\mathcal{L})$, as

$$\begin{aligned}\mathcal{L}f(x) &= \lim_{t\downarrow 0} \frac{\mathbb{E}^x[f(X_t) - f(x)]}{t} \\ &= \sum_{i=1}^{n} b_i(x)\frac{\partial f}{\partial x_i}(x) + \frac{1}{2}\sum_{i,j}\sigma^2(x)\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\end{aligned} \tag{11}$$

where $b_i(x)$ is the $i$'th element of $b(x) \in \mathbb{R}^n$. Thus $\mathcal{L}$ is an operator acting on $f \in C^2(\mathbb{R}^n)$ 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 (6), which has:

$$b(x) = -\frac{\beta}{2}\pi_0^2(x)\nabla J(x) - \pi_0(x)\nabla\pi_0(x), \quad \sigma(x) = \pi_0(x)$$

Thus, the infinitesimal generator of our process is given as

$$\mathcal{L}f = \frac{1}{2}\pi_0^2\Delta f - \frac{\beta}{2}\pi_0^2\langle\nabla J, \nabla f\rangle - \pi_0\langle\nabla\pi_0, \nabla f\rangle \tag{12}$$

and note that by assumptions (A2),(A6) and by Theorem 2.5 of [17], we have that (6) admits a unique strong solution.

*2) Poincaré and logarithmic Sobolev inequalities:* From the generator $\mathcal{L}$ we can define the *Dirichlet form*

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

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* [18] with constant $c$ if

$$\chi^2(\mu||\pi) \le c\,\mathcal{E}\left(\sqrt{\frac{d\mu}{d\pi}}\right)$$

for all probability measures $\mu \ll \pi$ ($\mu$ absolutely continuous w.r.t $\pi$), where $\chi^2(\mu||\pi) := ||\frac{d\mu}{d\pi} - 1||^2_{L^2(\pi)}$ is the $\chi^2$ divergence between $\mu$ and $\pi$. We say that $\pi$ satisfies a *logarithmic Sobolev inequality* [18] with constant $c$ if

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

for all $\mu \ll \pi$, where

$$D(\mu||\pi) = \int d\mu \log\frac{d\mu}{d\pi}$$

is the Kullback-Leibler divergence between $\mu$ and $\pi$.

In this paper we will show that the diffusion (6) satisfied a log-Sobolev inequality, because several useful results then apply. Specifically, letting $\{X(t)\}_{t\ge 0}$ be a Markov process with stationary distribution $\pi$ and Dirichlet form $\mathcal{E}$, then we have:

**Lemma 1** (Exponential decay of entropy [18], Th. 5.2.1)**.** *Let $\mu_t := Law(X(t))$. If $\pi$ satisfies a logarithmic-Sobolev inequality with constant $c$, then*

$$D(\mu_t||\pi) \le D(\mu_0||\pi)e^{-2t/c} \tag{13}$$

**Lemma 2** (Otto-Villani theorem [18], Th. 9.6.1)**.** *If $\pi$ satisfies a logarithmic-Sobolev inequality with constant $c$, then, for any $\mu \ll \pi$*

$$\mathcal{W}_2(\mu, \pi) \le \sqrt{2cD(\mu||\pi)} \tag{14}$$

The following Proposition will be a crucial tool allowing us to show that our diffusion (6) satisfies a log-Sobolev inequality.

**Proposition 2** (Cattiaux et. al. (2010) [19])**.** *Let $\pi(dx) = \exp(-H(x))dx$ be a probability measure on $\mathbb{R}^N$ with $H \in C^2(\mathbb{R}^N)$ 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 $V : \mathbb{R}^d \to [1, \infty)$ such that*

$$\frac{\mathcal{L}V(w)}{V(w)} \le \kappa - \gamma\|w\|^2 \ \forall w \in \mathbb{R}^d \tag{15}$$

2) *$\pi$ satisfies a Poincaré inequality with constant $c_P$.*
3) *There exists some constant $K > 0$, such that $\nabla^2 H \succcurlyeq -KI_d$*

*Let $Z_1, Z_2$ be defined as*

$$Z_1 = \frac{2K}{\gamma} + \frac{2}{K}, \quad Z_2 = \frac{2K}{\gamma}\left(\kappa + \gamma\int_{\mathbb{R}^N}\|w\|^2\pi(dw)\right) \tag{16}$$

*Then $\pi$ satisfies a logarithmic Sobolev inequality with constant $c_{LS} = Z_1 + (Z_2 + 2)c_P$.*

Condition (2) of the above Proposition requires that the measure $\pi$ satisfy a Poincarê inequality. This can be shown by employing the following result.

**Proposition 3** (Bakry 2008 [20])**.** *Let $\pi(dx) = \exp(-H(x))dx$ be a probability measure on $\mathbb{R}^N$ with $H \in C^2(\mathbb{R}^N)$ and lower bounded. Let $\mathcal{L}$ be the infinitesimal generator of a Markov process with stationary measure $\pi$. Suppose there exist constants $\kappa_0, \zeta_0 > 0$, $r \ge 0$ and a $C^2$ function $V : \mathbb{R}^N \to [1, \infty)$ such that*

$$\frac{\mathcal{L}V(w)}{V(w)} \le -\zeta_0 + \kappa_0\mathbf{1}\{\|w\| \le r\} \tag{17}$$

*Then $\pi$ satisfies a Poincaré inequality with constant*

$$c_P \leq \frac{1}{\zeta_0}\left(1 + C\kappa_0 r^2 \exp(O_r(H))\right) \tag{18}$$

*where $C > 0$ is a universal constant and $O_r(H) := \max_{\|w\|\leq r} H(w) - \min_{\|w\|\leq r} H(w)$*

The following Corollary is unrelated to Poincarê and logarithmic-Sobolev inequalities, but will be useful in relating a KL-divergence bound to a 2-Wasserstein bound.

**Corollary 1** (Bolley and Villani 2005 [21] Cor. 2.3)**.** *For any two Borel probability measures $\mu, \nu$ on $\mathbb{R}^N$,*

$$\mathcal{W}_2(\mu,\nu) \leq C_\nu \left[\sqrt{D(\mu||\nu)} + \left(\frac{D(\mu||\nu)}{2}\right)^{1/4}\right]$$

$$C_\nu = 2\inf_{\lambda>0}\left(\frac{1}{\lambda}\left(\frac{3}{2} + \log\int_{\mathbb{R}^N} e^{\lambda\|w\|^2}\nu(dw)\right)\right)^{1/2}$$

Next we provide details on the structure of the proof of Theorem 1, utilizing the tools presented above.

## V. Main Result: Proof Outline

Here we provide the proof structure for our bound on $\mathcal{W}_2(\pi_k, \pi_\infty)$, provided as (9) in Theorem 1. The high level proof structure is as follows: We bound $\mathcal{W}_2(\pi_k,\pi_\infty) \leq \mathcal{W}_2(\pi_k,\nu_{k\epsilon}) + \mathcal{W}_2(\nu_{k\epsilon},\pi_\infty)$, i.e., we first control the discretization error between passive algorithm 2 and diffusion 6, 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 $O(k\epsilon\sqrt{\epsilon})$, we employ a Girsanov change of measure (controlling the KL-divergence), followed by Corollary 1 (to relate back to Wasserstein distance), as in [10]. This procedure relies crucially on the exponential integrability of the diffusion (6), which we prove as Lemma 5.

To bound the latter ($\mathcal{W}_2(\nu_{k\epsilon},\pi_\infty)$), we first show that $\pi_\infty$ satisfies a logarithmic-Sobolev inequality, by satisfying the conditions of Proposition 2 [19]. This result is given as Proposition 4. We then apply exponential decay of entropy [18], given as Lemma 1, and the Otto-Villani Theorem [20], given as 2. This procedure provides an exponentially decaying bound on $\mathcal{W}_2(\nu_{k\epsilon},\pi_\infty)$.

### A. Technical Results

Here we list several technical results which will be utilized in the proof methodology that follows. The proofs of all of these can be found in [22]. We denote $\bar{\pi}_0 := \sup_x \pi_0(x)$ and $\bar{\pi}_{0,\lambda} := \sup_x \pi_{0,\lambda}(x)$. $A = \|J(0)\|, B = \|\nabla J(0)\|$, and $I, I'$ are constants provided in Lemma 7.4 of [22].

**Lemma 3** ($\pi_{0,\lambda}$ exponential integrability)**.** *For all $\lambda \leq 1$, $\pi_{0,\lambda}$ has a bounded and strictly positive density with respect to the Lebesgue measure on $\mathbb{R}^N$, and*

$$\kappa_0^\lambda := \log\int_{\mathbb{R}^N} e^{\|x\|^2} d\pi_{0,\lambda}(x) < \infty \tag{19}$$

*and denote $\kappa_0 := \kappa_0^\lambda|_{\lambda=1}$.*

**Lemma 4** (relative entropy bound)**.**

$$\bar{D}_0^\lambda := D(\pi_{0,\lambda}||\pi_\infty) \leq \log\bar{\pi}_{0,\lambda} + \frac{N}{2}\log\frac{3\pi}{m\beta} + \frac{\beta b}{2}\log 3 + \beta\left(\frac{L_{\nabla J}}{3}\kappa_0^\lambda + B\sqrt{\kappa_0^\lambda} + A\right) \tag{20}$$

**Lemma 5** (exponential integrability of Langevin diffusion)**.**

$$\log\mathbb{E}[e^{\|\alpha(t)\|^2}] \leq \kappa_0^\lambda + ((\beta b + N)2\epsilon + 2I')t$$

*where $\kappa_0$ is given in* (19)*.*

**Lemma 6** ($L^2$ bound on Langevin diffusion)**.**

$$\mathbb{E}\|\alpha(t)\|^2 \leq \kappa_0^\lambda + \frac{(\beta b + N)\bar{\pi}_{0,\lambda} + 2I}{(m\beta)\bar{\pi}_{0,\lambda}}$$

### B. 2-Wasserstein Bound for Diffusion Approximation

The following Lemma provides a bound on $\mathcal{W}_2(\pi_k, \gamma_{k\epsilon})$.

**Lemma 7.** *Fixing the step size $\epsilon$ and time horizon $k\epsilon$, take the kernel scale parameter $\Delta$ and sampling distribution scale parameter $\lambda$ small enough to satisfy* (7)*. Then we have*

$$\mathcal{W}_2(\pi_k,\nu_{k\epsilon}) \leq k\epsilon\sqrt{\epsilon}\Bigg[6\sqrt{12C_0+3} + 3\sqrt{2} + 4\left(\frac{3}{2} + C_1\right)^{1/2}\left(4\sqrt{C_2} + 2\sqrt{2L_J^2 + 4C_0}\right)\Bigg] \tag{21}$$

where $C_0, C_1, C_2$ are constants provided in Appendix VII. $M_\theta$ is a bound on $\mathbb{E}\|\theta_k\|^2$, see the Appendix or Lemma 4.9 of [22].

*Proof Sketch*: The full proof is available in [22]. We aim to relate the measures $\pi_k$ and $\nu_{k\epsilon}$ through Girsanov's formula, as in [10], to obtain a desirable bound. However, due to an incompatibility between the algorithm 2 and continuous time diffusion (6) (specifically lack of absolute continuity between measures $\pi_k$ and $\nu_{k\epsilon}$), we cannot directly apply Girsanov's formula, see [22] for extended discussion of this phenomena. To solve this, we introduce the intermediate process

$$X(t) = \alpha_0 - \int_0^t \hat{g}_s(\theta_{\bar{s}}, X(s))ds + \int_0^t \pi_{0,\lambda}(X(s))dW(s) \tag{22}$$

where

$$\hat{g}_s(\theta_{\bar{s}}, X(s)) = \mathbb{E}\Bigg[\Bigg(K_\Delta(\theta_{\bar{s}},\bar{\alpha}(s))\frac{\beta}{2}\hat{\nabla}J(\theta_{\bar{s}}) + \nabla\pi_{0,\lambda}(\bar{\alpha}(s))\Bigg)\pi_{0,\lambda}(\bar{\alpha}(s))\,\Bigg|\,\bar{\alpha}(s) = X(s)\Bigg]$$

Notice that $X(t)$ is a stochastic differential equation with the same volatility term as the diffusion (6). Thus, Letting $\gamma_{k\epsilon}$ denote the law of $X(t)$ at time $t$, we can apply Girsanov's formula to relate $\gamma_{k\epsilon}$ and $\nu_{k\epsilon}$. We can bound $\mathcal{W}_2(\pi_k,\nu_{k\epsilon}) \leq \mathcal{W}_2(\pi_k,\gamma_{k\epsilon}) + \mathcal{W}_2(\gamma_{k\epsilon},\nu_{k\epsilon})$.

We bound $\mathcal{W}_2(\pi_k, \gamma_{k\epsilon})$ by $\mathbb{E}_{(x\sim\pi_k, y\sim\gamma_{k\epsilon})}\|x-y\|^2$, and obtain (see Lemma 5.1 of [22])

$$\mathcal{W}_2(\pi_k, \gamma_{k\epsilon}) \leq 6(k\epsilon)\epsilon\sqrt{12C_0+3} + 3\sqrt{2(k\epsilon)\epsilon} \tag{23}$$

where $C_0$ is a constant provided in Appendix VII.

Then, we use Girsanov's formula within the definition of the KL-divergence (see Lemma 5.2 of [22]) to obtain

$$D(\gamma_{k\epsilon}\|\nu_{k\epsilon}) \leq (k\epsilon)^3\,\epsilon^3\left[4\beta L_{\nabla J}^2\left(72C_0 + 6\sqrt{C_0} + 18 + \sqrt{2}\right)\right] + (k\epsilon)\,\epsilon\,(2L_J^2 + 4C_0)$$

Now applying Corollary 1 gives us a way to relate this KL-divergence bound to a 2-Wasserstein bound, provided that the measure $\nu_{k\epsilon}$ is exponentially integrable, i.e., $\mathbb{E}[\exp(\|\alpha(t)\|^2)] < \infty$. Lemma 5, in Appendix V-A, provides such a bound on $\mathbb{E}[\exp(\|\alpha(t)\|^2)]$, so we employ this within Corollary 1 to produce

$$\begin{aligned}&\mathcal{W}_2(\gamma_{k\epsilon}, \nu_{k\epsilon})\\ &\quad\leq 4\left(\frac{3}{2} + C_1 k\epsilon\right)^{1/2}\sqrt{k\epsilon}\sqrt{\epsilon}\left(4\sqrt{\beta L_{\nabla J}^2 C_2} + 2\sqrt{2L_J^2 + 4C_0}\right)\end{aligned} \tag{24}$$

with $C_0.C_1, C_2$ constants defined in Appendix VII. Combining (24) with (23) yields (21).

### C. 2-Wasserstein Distance for Diffusion Convergence

Here we describe the method to bound $\mathcal{W}_2(\nu_{k\epsilon}, \pi_\infty)$. The strategy is as follows:

i) Show that $\pi_\infty$ satisfies a logarithmic-Sobolev inequality.
ii) Apply exponential decay of entropy, given as Lemma 1, with the relative entropy bound in Lemma 4, to derive a bound on $D(\nu_{k\epsilon}\|\pi_\infty)$
iii) Apply the Otto-Villani Theorem, given as Lemma 2, to relate this to a bound on $\mathcal{W}_2(\nu_{k\epsilon}, \pi_\infty)$.

We accomplish (i) in the following proposition, establishing that the Gibbs measure $\pi_\infty$ satisfies a log-Sobolev inequality:

**Proposition 4.** *For $\beta$ satisfying Assumption 8, the Gibbs measure $\pi_\infty$ satisfies a logarithmic Sobolev inequality with constant $c_{LS}$:*

$$\begin{aligned}0 \leq c_{LS} &\leq \frac{2\beta L_{\nabla J}}{\gamma} + \frac{2}{\beta L_{\nabla J}}\\ &+ \frac{1}{\lambda}\left(\frac{2\beta L_{\nabla J}}{\gamma}\left(\kappa + \gamma\left(\kappa_0 + \frac{(\beta b + N)\bar{\pi}_{0,\lambda} + 2I}{(m\beta)\bar{\pi}_{0,\lambda}}\right)\right) + 2\right)\end{aligned} \tag{25}$$

where

$$\begin{aligned}\frac{1}{\lambda} &\leq \frac{1}{2\kappa}\left(1 + \frac{4C\kappa^2}{\gamma}\exp\left(\beta\left(\frac{(L_{\nabla J}+B)\kappa}{\gamma} + A + B\right)\right)\right)\\ \kappa &= \left(\frac{1}{2}\beta m N + \beta m I\right) + \frac{1}{2}\left[\beta^2 m b + (\beta m M)^2\right]\\ \gamma &= \frac{1}{2}\left((\beta m)^2 + \left(1 - \frac{1}{\bar{\pi}_0^2 + 1}\right)\right)\end{aligned} \tag{26}$$

*Proof Sketch*: The full proof is available in [22]. The key tool we use is the main Theorem in [19], reproduced as Proposition 2. To satisfy condition (1) of Proposition 2 we show that the Lyapunov function

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

and the infinitesimal generator (12) satisfy (15), with $\kappa$ and $\gamma$ given in (26). Then, Proposition 3 is used to show that condition (2) is satisfied. Condition (3) is satisfied with $K = \beta L_{\nabla J}$ by assumption 1.

Now since $D(\nu_0||\pi_\infty) = D(\pi_0||\pi_\infty) < \infty$ by Lemma 4, we can apply the exponential decay of entropy (Lemma 1) to obtain

$$D(\nu_t||\pi_\infty) \leq D(\pi_{0,\lambda}||\pi_\infty)e^{-2t/\beta c_{LS}} \tag{27}$$

Then by the Otto-Villani Theorem and Lemma 4, we have

$$\mathcal{W}_2(\nu_t, \pi_\infty) \leq \sqrt{2c_{LS}\bar{D}_0^\lambda}e^{-t/\beta c_{LS}} \tag{28}$$

where $\bar{D}_0^\lambda$ is the relative entropy bound given in (20) and $c_{LS}$ is bounded in (25).

### D. Controlling the 2-Wasserstein Distance

Combining the bounds (21) and (28) yields

$$\begin{aligned}&\mathcal{W}_2(\pi_k, \pi_\infty)\\ &\quad\leq k\epsilon\sqrt{\epsilon}\Big[6\sqrt{12C_0+3} + 3\sqrt{2}\\ &\qquad + 4\left(\frac{3}{2} + C_1\right)^{1/2}\left(4\sqrt{C_2} + 2\sqrt{2L_J^2 + 4C_0}\right)\Big]\\ &\qquad + \sqrt{2c_{LS}\bar{D}_0^\lambda}e^{-k\epsilon/\beta c_{LS}}\end{aligned} \tag{29}$$

The strategy to control (29) 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 $\bar{D}_0^\lambda$ may depend inconveniently on $\lambda$, and thus on $\epsilon$. In [22] we provide details on how the parameter specifications inherently control this relative entropy term. In short, using Lemma 4 and the choices of $k$, $\epsilon$, and $\lambda$ in

(7), we obtain:

$$\begin{aligned}
&\mathcal{W}_2(\pi_k, \pi_\infty) \\
&\quad \leq \delta \left[ 6\sqrt{12C_0 + 3} + 3\sqrt{2} + 4\left(\frac{3}{2} + C_1\right)^{1/2} \left(4\sqrt{C_2} \right.\right. \\
&\quad \left.\left. + 2\sqrt{2L_J^2 + 4C_0}\right)\right] \\
&\quad + \delta \sqrt{2c_{LS} N \log\left(\frac{1}{\lambda}\right)} + \delta \sqrt{2c_{LS} C_3}
\end{aligned} \tag{30}$$

where $C_3$ is listed in the Appendix. Then, since $\lambda \in [\epsilon^2, \epsilon^{3/2}]$ and $\epsilon \leq \left(\frac{\delta}{\log(1/\delta)}\right)^2$ we have

$$\log(\frac{1}{\lambda}) \leq 4 \log\left(\frac{\log(1/\delta)}{\delta}\right) \leq 5 \log\left(\frac{1}{\delta}\right)$$

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 (8).

Plugging this into (30) gives the 2-Wasserstein bound presented in Theorem 1.

## VI. 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 [6]. The passive Langevin algorithms analyzed in this paper use sequential evaluations of a stochastic gradient descent by an external agent (forward learner), and reconstruct the cost function. Thus *real-time* inverse reinforcement learning is achieved, in that we (the inverse learner) reconstruct the cost function while it is in the process of being optimized. Specifically, we have provided finite-time 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 [10] and uses techniques in the analysis of Markov Diffusion Operators [18] to achieve the bound.

## References


[1] A. Y. Ng, S. Russell *et al.*, "Algorithms for inverse reinforcement learning." in *Icml*, vol. 1, 2000, p. 2.

[2] 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.

[3] 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.

[4] 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.

[5] S. B. Gelfand and S. K. Mitter, "Simulated annealing type algorithms for multivariate optimization," *Algorithmica*, vol. 6, pp. 419–436, 1991.

[6] V. Krishnamurthy and G. Yin, "Langevin dynamics for adaptive inverse reinforcement learning of stochastic gradient algorithms." *J. Mach. Learn. Res.*, vol. 22, pp. 121–1, 2021.

[7] ——, "Multikernel passive stochastic gradient algorithms and transfer learning," *IEEE Transactions on Automatic Control*, vol. 67, no. 4, pp. 1792–1805, 2022.

[8] G. Yin and K. Yin, "Passive stochastic approximation with constant step size and window width," *IEEE transactions on automatic control*, vol. 41, no. 1, pp. 90–106, 1996.

[9] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Recursive Algorithms and Applications*, 2nd ed. Springer-Verlag, 2003.

[10] 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.

[11] A. S. Dalalyan and A. Karagulyan, "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient," *Stochastic Processes and their Applications*, vol. 129, no. 12, pp. 5278–5311, 2019.

[12] 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.

[13] C. Villani *et al.*, *Optimal transport: old and new*. Springer, 2009, vol. 338.

[14] 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*, vol. 12, 1999.

[15] 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.

[16] J. Rust, "Structural estimation of Markov decision processes," *Handbook of econometrics*, vol. 4, pp. 3081–3143, 1994.

[17] I. Karatzas and S. E. Shreve, *Brownian motion and stochastic calculus*. Springer Science & Business Media, 1991, vol. 113.

[18] D. Bakry, I. Gentil, M. Ledoux *et al.*, *Analysis and geometry of Markov diffusion operators*. Springer, 2014, vol. 103.

[19] P. Cattiaux, A. Guillin, and L. Wu, "A note on Talagrand's transportation inequality and logarithmic sobolev inequality," *arXiv preprint arXiv:0810.5435*, 2008.

[20] D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin, "A simple proof of the Poincaré inequality for a large class of probability measures," 2008.

[21] 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, no. 3, 2005, pp. 331–352.

[22] L. Snow and V. Krishnamurthy, "Finite-sample bounds for adaptive inverse reinforcement learning using passive Langevin dynamics," *ArXiv preprint*, 2023.


## VII. Appendix: Bound Constants

$$\begin{aligned}
C_0 &:= 3L_{\nabla J}^2 (M_\theta + 2B^2 M_\theta) + B^2 + \zeta \\
C_1 &:= \kappa_0^\lambda + (\beta b + N) 2\epsilon + 2I') \\
C_2 &:= \beta L_{\nabla J}^2 \left(72C_0 + 6\sqrt{C_0} + 18 + \sqrt{2}\right) \\
C_3 &:= \log(\bar{\pi}) + \frac{N}{2} \log \frac{3\pi}{m\beta} + \frac{\beta b}{2} \log 3 \\
&\quad + \beta \left( \frac{L_{\nabla J}}{3} \kappa_0 + B\sqrt{\kappa_0} + A \right) \\
C_4 &:= \left[ 6\sqrt{12C_0 + 3} + 3\sqrt{2} \right. \\
&\quad \left. + 4\left(\frac{3}{2} + C_1\right)^{1/2} \left(4\sqrt{C_2} + 2\sqrt{2L_J^2 + 4C_0}\right)\right] \\
M_\theta &= \kappa_0 + 2\left(1 \vee \frac{1}{m}\right)\left(b + 2B^2\right)
\end{aligned}$$