We thank **R1, R2, R3, R4**, who provided overall scores (7-6-7-7) respectively, for their careful reading of the paper, their positive comments on its clarity and interest, and their relevant questions. We recall that our paper provides rates of convergence for the iterates $(\mu_n)_{n\geq 0}$ of SVGD in the infinite particle regime (Eq 14) (i.e., of the time-discretized version of the SVGD gradient flow $\mu_t$ (Eq 11)) to a target distribution $\pi \propto \exp(-V)$. At time $n$, the distribution $\mu_n$ is associated to a Mac-Kean Vlasov process $x_n$ whose dynamics depends on $\mu_n$ itself. Therefore, the practical implementation of SVGD relies on approximating $x_n$ with $N$ interacting particles $(\hat{x}_n^i)_{i=1}^N$ (Eq 38), where the empirical distribution $\hat{\mu}_n$ of the particles approximates $\mu_n$. In particular, we provide a $\mathcal{O}(1/n)$ convergence rate for the arithmetic mean of the Kernel Stein Discrepancy (KSD) (which metricizes weak convergence in many cases, see Sec. 3.3) between the iterates $\mu_n$ and $\pi$, under Assumptions A1–A3. It **does not rely on Stein LSI nor on convexity of** $V$, unlike most of the results on Langevin Monte Carlo (LMC) which assume either (standard) LSI or convexity of $V$.

**R1, R3. Finite number of particles.** We would like to clarify our result (Prop 12). It is non-asymptotic one in the sense that it provides an explicit bound. However, *it is not a bound that helps quantify the rate of minimization of the objective function*. Rather, it is a bound between the population distribution $\mu_n$ and its particle approximation $\hat{\mu}_n$. It states that for a fixed time horizon $T > 0$, $\mathbb{E}[W_2^2(\mu_n, \hat{\mu}_n)] \leq C\frac{1}{\sqrt{N}}$, where $N$ is the number of particles. Such results are referred to *propagation of chaos* in the PDE literature, where having $C$ depending on $T$ is common. Getting a similar bound with $C$ *not* depending on $T$ would be a much stronger result referred to as *uniform in time propagation of chaos* (see l.522-525). Such results, which are subject to active research in PDE, are hard to obtain. Among the recent exceptions is (Durmus, 2018a) who consider the process $dx_t = -\nabla U(x_t) - \nabla W * \mu_t(x_t)dt$ and manage to prove such results when $U$ is strictly convex outside of a ball. However in SVGD (see Eq 8), the attractive force $\nabla \log \pi(x)k(x,.)$ cannot be written as the gradient of a confinement potential $U$ in general. Hence these results do not apply. **R3 (2).** In our answer to R2 below, we discuss a contradiction in the Th 7 assumptions, discovered by R2. We will therefore acknowledge that the convergence rate for SVGD using $\hat{\mu}_n$ remains an open problem.

**R2, R4. Assumptions A1-A3.** A1 and A2 are mild smoothness assumptions on $(k, V)$. In particular, A2 is standard in the LMC algorithm literature. A1 is needed to obtain our core descent result, Prop 5, because KL is not smooth, see Remark 3 l.320-323. A3 can be checked in each specific context. For instance, in Lemma 17, we provide conditions under which A3 holds. The validity of this hypothesis is also confirmed in our experiments.

**R2.** Thank you for raising the question of whether there exists a distribution $\pi$ and kernel $k$ that simultaneously satisfy the Stein LSI and assumption A1. Having thought about this, *we believe that no such* $\pi$ *and* $k$ *exist* (at least for $\mathcal{X} = \mathbb{R}^d$). Given that both the kernel and its derivative are bounded, equation $\int \sum_{i=1}^d [(\partial_i V(x))^2 k(x,x) - \partial_i V(x)(\partial_i^1 k(x,x) + \partial_i^2 k(x,x)) + \partial_i^1 \partial_i^2 k(x,x)] d\pi(x) < \infty$ reduces to a property on $V$ which, as far as we can tell, always holds; and this implies that Stein LSI does not hold (see l.163-165). For instance, even when $V = -\log(\text{cauchy})$ or $V = -\log(\text{student})$, we quickly find that the resulting expectations are bounded. We will therefore remove Th 7 from our results, and replace it with a discussion on the difficulty in simultaneously ensuring conditions for a descent lemma and for Stein LSI. In particular, we recall that in the classical LMC setting, we would require only smoothness assumptions on $V$ (A2) and the classic LSI inequality to obtain exponential convergence in the objective (here KL), see (Vempala, 2019). This is also the approach in nonconvex optimization (where LSI is called Polyak-Lojasiewicz (PL) inequality, see Rk 3). Unfortunately, as KL is not smooth (see l.320-323), we had to further assume A1 i.e. boundedness of the kernel in Prop. 5, resulting in the contradiction in Thm. 7. We emphasize that Corollary 6 establishes convergence under very general conditions, and remains valid, however we cannot now show fast rates. **R2, R3. Validity of the Stein LSI itself.** (Duncan et al., 2019) discusses conditions for which the Stein LSI on its own is satisfied, among which are the 1-d examples provided Sec 3.3 (which, unfortunately, do not satisfy A.1). Construction of examples satisfying Stein LSI should begin by ensuring that l.164 does not hold, i.e. the integral is infinite (see above): eg, $k$ polynomial of order 3 or greater and $\pi$ with exploding $\beta$-moments, for $\beta \geq 3$ (e.g., a student distribution in $\mathcal{P}_2(\mathcal{X})$).

**R1.** The Wasserstein Hessian of $\mathrm{KL}(.|\pi)$ is briefly mentioned in (Villani, 2003, Sec 8.2) and (Wibisono, 2018, Sec 3.1.1) but was not particularly highlighted in the ML literature. We derive all the formulas in the proof of Prop 5.

**R3. (3)** We will correct this with a precise discussion about the dependence on the dimension $d$. In Cor. 6 and Th. 7, the constants $M, B, C, \mathrm{KL}(\mu_0|\pi)$ are parameters of the problem and depend indeed implicitly on $d$. To explicit the dependence of $\mathrm{KL}(\mu_0|\pi)$ on $d$, we can apply (Vempala, 2019, Lem. 1): under A2, we have $\mathrm{KL}(\mu_0|\pi) \leq V(x_\star) + \frac{d}{2}\log\left(\frac{M}{2\pi}\right)$, where $x_\star$ is a stationary point of $V$ and assuming that $\mu_0 \sim \mathcal{N}(x_\star, \frac{1}{M})$. We will explicit the dependence e.g. for $M, B$ for a gaussian kernel $k$ and quadratic potential $V$ for illustration. **(4)** Assuming $0 < \gamma < \min\left(\frac{1}{2\lambda}, \frac{1}{B^2(\alpha^2+M)}\right)$ is sufficient to obtain $0 < 2c_\gamma\lambda < 1$. Indeed, $\gamma < \frac{1}{B^2(\alpha^2+M)}$ implies $\frac{\gamma}{2} < c_\gamma$. Since $c_\gamma < \gamma$, we have $0 < \lambda\gamma < 2c_\gamma\lambda < 2\gamma\lambda < 1$, using the assumption on $\gamma$.

**R4.** SVGD can be seen as a gradient descent (GD) algorithm to minimize $\mathrm{KL}(\cdot|\pi)$. In this context, the KSD ($I_{stein}(\mu_n|\pi)$) plays the role of the squared norm of the gradient at time $n \geq 0$. Assumption A3 is analogous to assuming $\sup_n \|\nabla F(x_n)\|^2 < \infty$ when applying GD to minimize some function $F$ over $\mathbb{R}^d$, which is a standard bounded gradient assumption in optimization. It holds for instance if the objective function $F$ is Lipschitz.