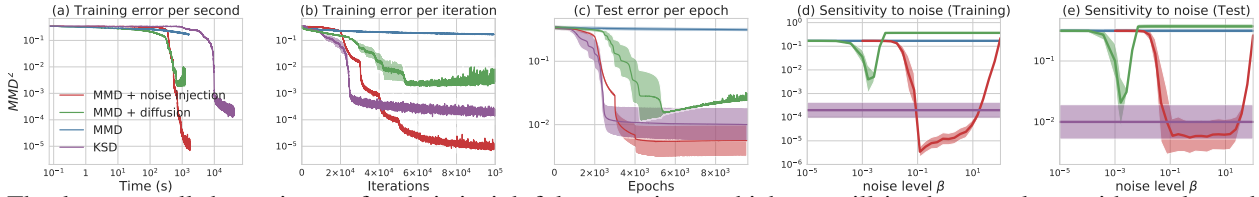

Thank you to all the reviewers for their insightful suggestions, which we will implement along with results and clarifications below. **When do prop. 7 conditions hold:** we relax the condition by restricting the set of test functions in Eq 15 to functions in the RKHS and exactly obtain the KSD [31], allowing us to compute and verify it in simple cases. E.g., when both μ and ν_0 are gaussians and the kernel is $k(x, y) = xy + x^2y^2$, then ν_t is also gaussian with mean a_t and variance σ_t satisfying a coupled differential equation. The KSD can be computed in closed form in terms of a_t and σ_t. The condition is then verified using standard dynamical systems techniques. A simpler example: linear kernel; both μ and ν_0 are gaussians with the same variance. The dynamical system can be solved in closed form and the condition in Prop. 7 is satisfied. We will add both examples with detailed proofs in the appendix. In the more general case, weighted negative sobolev distance (WNSD) can be controlled if ν_t has a density bounded below [35]. Getting milder conditions remains an open question. **Conditions that depend on the whole trajectory** have been considered previously: [A, Theorem 6.1], assumes that the first moment remains constant to get a stronger convergence result. We agree with R.1 that Prop. 7 can be interpreted as a locality condition. Unlike in the finite dimensional case, however, it is unclear a priori what locality condition is best suited to characterize convergence of the MMD flow. Our contribution is to use the WNSD to define the notion of locality, which may be useful for further analysis of MMD flow convergence.

Reviewer 1. We've added experiments (the figure shows results averaged over 10 runs. Same setting as in Sec G.1) and now compare with an entropy-regularized flow [30] and KSD flow [31]. (d,e) our noise injection method (red) is robust to the amount of noise and achieves best performance over a wide region. MMD + diffusion (green) has worse performance over a much narrower optimal region. KSD (purple) behaves better than MMD (blue) (b and c), however we (red) still outperform it. Moreover, the computational cost is much higher for KSD (a). We didn't include SVGD in this comparison since it requires the closed form unnormalised density of the target, contrary to our paper setting. **Validation error plateaus** even for $\beta_n > 0$, since the kernel is estimated using finitely many random features in training (empirical version, Eq 42 in Sec B.1). Validation error is estimated using new RFs. **MMD flow vs training NNs** we'll add this clarification: "The MMD flow (without noise) with a finite number of particles is equivalent to standard GD with a quadratic loss, as shown in Sec B.1 and G.1 and in [37]. Therefore, our results could be used to analyze convergence of GD in NNs even when a non-linearity is applied after the final activation. This is by contrast to [37,11] where the final layer must be linear to get convergence results." In the figure, MMD flow (blue) is in fact equivalent to SGD. Stochasticity is due to the estimation of the kernel using RFs (Eq 42+43 in Sec B.1). Our algorithm based on noise injection (red) can therefore be applied to train NNs as shown in Alg. 2 of App. G.1. We'll clarify this in the main paper, along with extensions to more general cost functions.

Reviewer 2. We'll gladly implement all suggested clarifications and provide more intuition for displacement convexity.

Reviewer 3. As suggested, we will provide more discussion of prior works, and how they differ from ours, as follows. [Carillo+, Thm. 6.1] provides a convergence results when both potentials satisfy convexity assumptions. This can't be applied for MMD flow as it requires convexity of either the potential or interaction term. Both terms involve the same kernel but with opposite signs, so even under convexity of the kernel, a concave term appears and cancels the effect of the convex term, and [Carillo+, Remark 6.4] fails to hold. Moreover, the requirement that the kernel be positive semi-definite makes it hard to construct interesting convex kernels. In [30], an entropic regularization is used and allows to prove convergence to the global optimum. However, the latter is in general different from the global optimum of the un-regularized loss. To get an accurate solution, small levels of noise are required which can be of limited interest in practice: green traces, Figure. Our proposed algorithm in (Eq. 21) is different from entropic regularization, and the global optimum of the MMD remains a fixed point of the algorithm. Qualitatively, the behavior is also very different: Figure (red). In [37,11], the loss function has a particular structure: '1-homogeneity'.This is well suited for NNs with a linear final layer and leads to an elegant proof for global optimality. In our case, this corresponds to a kernel k of the form $k(x, x') = cc'\kappa(\theta, \theta')$ where x and x' are of the form $x = (c, \theta)$ and $x' = (c', \theta')$. However, when a characteristic kernel is required (to ensure the MMD is a metric), such a structure can't be exploited. KSD flow [31] is also shown to minimize the MMD. However, those are two different functionals and behave differently: Figure (purple vs blue/red). See also our App. E.1 discussion on the global optimality condition in [31]. SVGD [28], was introduced as a gradient flow of the KL w.r.t. a metric [28, Eq. 20] that is not the Wasserstein metric, and requires a closed form target density. Finally, we emphasize our new noise injected flow in Sec. 4., which improves over the Sec. 3 "vanilla" MMD flow and has a global convergence result (Prop. 8) that supplants the Prop. 7 conditions. We provide empirical evidence of its benefits compared to other methods in the figure. The algorithm can be best understood as a generalization of *continuation methods* to interacting potentials. To our knowledge, our work is the first to propose this approach.