We thank reviewers for careful reading, and appreciate the honesty of Reviewer 2. The weakness pointed out here is the lack of numerical evidence, which we explain below (No. 2). As both other reviewers pointed out, the significance of the paper is two-fold: 1. We show, by giving a counter-example, that the direct application of RCD (random coordinate descent) on LMC (Langevin Monte Carlo) does not improve the numerical performance; 2. With variance reduction techniques incorporated, the numerical cost is significantly reduced. The reduction rate depends on the dimension of the problem: the algorithm saves more in high dimensional problems. Moreover, in the under-damped case, the method converges as fast as the vanilla LMC while requiring only one partial derivative instead of the full gradient per iteration. This is the optimal numerical cost one can possibly get. Below we address the weakness pointed out by the reviewers.

1. WHAT IF THE FORWARD COST (FUNCTION EVALUATION) AND THE BACKWARD COST (GRADIENT EVALUATION) HAVE THE SAME ORDER OF COMPUTATION? We agree that there are cases, as pointed out by Reviewer 3 when the two costs are similar, but in the most general setting, a problem does require a much higher cost for the gradient to be computed. In fact, most problems in atmospheric science and remote sensing cannot even have one gradient computed due to the high dimensionality (see Refs. 21, 36). This is exactly why the ensemble type sampling methods became popular that target at achieving "gradient-free" property. We would like to put ourselves in the most general footing. It is our principle, and we believe it is shared by most researchers, that investigation into special cases should come after a clear picture of general setups. The same question could have been asked to challenge the validity of RCD, but nevertheless RCD is a tremendously popular method in optimization. We agree with Reviewer 4 that we could have made some comments on the cases when RCD already performs well. We believe playing with directional Lipschitz constants would be the key but this is beyond the scope of the current paper.

2. WHY ARE THERE NO NUMERICAL RESULTS? We have not seen a single result in the literature, including the fundamental papers in the area (see Refs. 8, 10, 12), that truly demonstrates the convergence in Wasserstein-2 distance numerically. This is simply because there is no numerical method available yet that is even able to evaluate the criterion. The W_2 distance between two probability measures is hard to compute in high dimensions, especially when one probability is represented by one data point. In the plot below we show the decay of MSE (mean square error, a much weaker and loosened criterion). We could have chosen to demonstrate these in the original paper, but we preferred reserving the space for richer theoretical guarantees other than providing numerical results with mismatching norms.

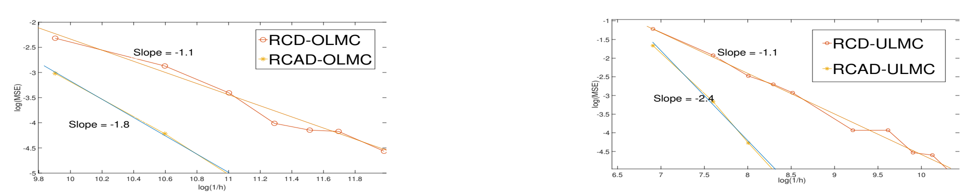

Figure 1: Decay of MSE of LMC in overdamped (left) and underdamped (right) settings. Test function: $\phi(x) = x_1^2$.

3. CAN WE ELIMINATE THE HESSIAN-LIPSCHITZ CONTINUITY ASSUMPTION? Yes we can. We did comment on it in the original paper (line 245-247). Since neither the result nor the proof is significant, we did not include the full statement in the paper. **Theorem:** *Suppose f satisfies Assumption 3.1, $h < O\left(\frac{1}{K}\right)$ and $\eta < h$, then:* $W_2(q_m^O, p) \lesssim \exp(-Mhm/4)W_2(q_0^O, p) + hK^{3/2} + h^{1/2}K$. Here $\lesssim$ means $\leq$ up to a constant independent of h, K.

4. CAN WE CHANGE OUR NORM? We can comment on other norms but we do not believe any theorems on other norms should be included. There is no single paper (either journal or conference) in the literature that studies convergence in more than one norm in one paper, exactly because different criteria are evaluated with different mathematical techniques, and the entire roadmap has to change. We do have a very simple corollary on MSE convergence. It is a standard derivation from W_2 convergence. **Corollary:** *Under conditions of Theorem 5.1, MSE decays with the rate* $|\mathbb{E}_{q_m^O}(\phi) - \mathbb{E}_p(\phi)| \lesssim \exp(-Mhm/4)W_2(q_0^O, p) + h(K^{3/2} + K)$, *for all Lipschitz test function* ϕ.

5. (FROM REVIEWER 4) WHY DON'T WE DO OPTIMIZATION FIRST? We very much appreciate Reviewer 4 raising this question. The fact is, we did. It was a surprising result for us that in optimization, this formulation does **NOT** help in saving numerical cost. We were left wondering if this is a known result in the community that we missed out on, or is this also new? We opt to investigate it in the optimization setup a bit more before claiming publicly a negative result.

Finally, while we agree a different layout of the paper, and some changes in phrases may help delivering stronger messages to some certain audience, and small typos on constants should also have been avoided, we are genuinely surprised that some tasks that have never been done in the literature are used to discount the significance of the current paper. We will be happily corrected by the reviewers if we miss any older results, and we will continue monitoring the area and NeurIPS selected publications for most recent progresses. We do not think the paper has ethical impact.