Reviewer #4: Thank you for a detailed review. Regarding your comments on the experiments section, please see the end of this page. We appreciate your comments on the presentation and we will consider incorporating your suggestions in the final version. In terms of content however, we believe that our contributions have been mischaracterized.
Q: "The contributions of the paper are very close from the one of [12]"
A: We respectfully strongly disagree. The authors in [12] propose an algorithm following the FTRL scheme, our algorithm is based on minimizing newly proposed surrogate cost functions. They study static regret, we study dynamic regret. They use a generic probabilistic bound at the core of their analysis, we use the specific dynamics of SGD/SGLD.
Q: "Novelty of the bound on dynamic regret compared to the one of [18]"; "The authors should discuss the links of assumption 3 and the different path-length assumptions in the literature"; "The links with (strong) convexity are not sufficiently discussed"
A: The regret bound derived in [18] assumes boundedness of the gradients of the cost functions which does not hold for our problem. Furthermore, our assumptions do no imply any of the path-length assumptions in [18]. Our bound and theirs are therefore not comparable. Besides, the work in [18] deals with the general online learning problem, whereas we are only concerned with the specific problem of adaptive importance sampling for SGD/SGLD. There is not much to be said about (strong) convexity apart from mentioning that our cost functions are convex but not strongly convex, which adds to the difficulty of our problem. The whole point of our work is to show that we are able to achieve a good dynamic regret bound even though most of the commonly used assumptions to study dynamic regret do not hold.
Q: "Why talking about the Langevin method (SGLD) ... it would be just as legitimate as comparing with all stochastic gradient descent methods (SAG, SAGA, SVRG,...)"
A: It is precisely because the only other method aside from SGD for which our algrorithm/analysis work is SGLD that we mention it. In particular, the variance of the SAG/SAGA/SVRG estimators has a completely different form, and designing adaptive importance sampling techniques for these estimators remains an open problem. Furthermore there has been a lot of work in applying variance reduction techniques initially designed for optimization to sampling (see e.g. [35]), and we view our work as a contribution to that literature as well.

Reviewer #6: Thank you for an in-depth read of our paper. We address your concern below.
Iteration complexity: At the theoretical level, we can show better convergence rates by replacing our dynamic regret bound directly in the standard convergence bound of averaged SGD as is done in ([11], Theorem 3 and Corollary 1). This does not cover the last-iterate case however. Since we are able to obtain per-step regret guarantees (Lemma 2 + Lemma 4), we can also obtain bounds on a weighted version of the dynamic regret, which allows us to obtain improved last-iterate guarantees for SGD compared to uniform sampling. We will consider including these results in a new section in the appendix if the reviewer thinks it would be a good addition. In practice, the magnitude of the improvement depends on how much variance reduction can be achieved through the use of the optimal probabilities vs uniform sampling. This is what we addressed in the experiments section, where we mentioned over-parametrized models and imbalanced datasets as examples of cases in which this improvement is significant. **Cost per iteration:** We would like to point out that while our algorithm is $\mathcal{O}(N)$ in both time and space, it has no dimension dependence. For large scale high-dimensional problems, the cost of the final optimization algorithm will still be largely dominated by the gradient evaluations, and the overhead from our sampler will be negligible. The fact that the memory cost is independent of the dimension of the problem is what makes our algorithm an attractive alternative to control-variate methods like SAG/SAGA/SVRG.

Reviewer #7: Thank you for your very positive comments and helpful suggestions! We will incorporate your suggestions for the experiments section in the final version, please see the paragraph below.

Reviewer #8: Thank you for taking the time to read our paper.
Concern 1: Their framework is slightly more general than ours, but when specialized to importance sampling for SGD it becomes exactly the same. Please see the last two paragraphs of the introduction in their paper.
Concerns 2, 3: See the discussion below. We disagree however on the significance of the theoretical improvement.
Concern 4: This was a mistake. Please see the supplementary version of the paper.

Comments on the experiments section: We would like to emphasize that our aim was not to provide a comprehensive empirical study of our algorithm, but only to illustrate its performance on a few datasets, and point to cases in which the improvement is substantial. This is inevitable given the space constraint. That being said, we agree with the reviewers that this section of our work can be improved, and we plan on making full use of the additional page in the final version to do so. In particular we plan on: (i) Giving more details on the setup of the experiments. (ii) Include an experiment that explores more the effect of the choice of step-size (decreasing vs constant) and batch size on the performance of the algorithm. This will include experiments with decreasing step-sizes and a batch-size of 1 which corresponds to the setting of our regret bound. (iii) Include results from other baselines in all of the experiments. (iv) Adding a paragraph analyzing the results and more explicitly explaining when and why the algorithm leads to significant improvements.