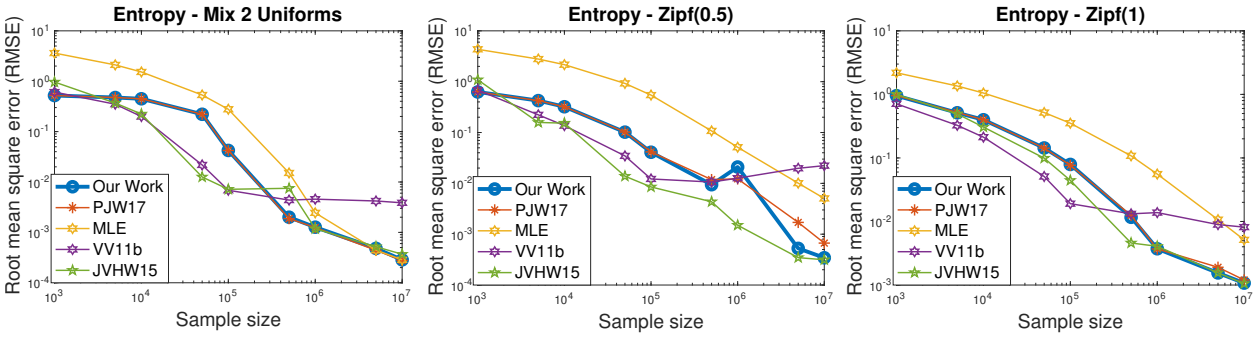
We thank the reviewers for the helpful comments, great suggestions, and positive feedback. In the final version, we will be sure to add further intuition about the proofs and insight into parameter choices to clarify the presentation. Further, we ran multiple experiments (detailed below) corroborating our theoretical claims. Finally, we thank reviewer 3 for pointing out that the result of [HJW18] holds only when $\epsilon > n^{-0.33}$ and therefore our result outperforms all previous universal estimators when applied to estimating entropy and distance to uniformity for $\epsilon < n^{-0.33}$. We hope the collection of proposed changes and new experiments elevate your view of the paper.

Experiments: We performed two sets of experiments. First, we compared the error of multiple estimators for computing entropy (in the full version we will repeat these experiments for distance to uniformity and more distributions).



Each plot depicts the performance of various algorithms for estimating entropy of different distributions with domain size $N = 10^5$. "Mix 2 Uniforms" is a mixture of two uniform distributions, with half the probability mass on the first $N/10$ symbols, and $\mathrm{Zipf}(\alpha) \sim 1/i^{\alpha}$ with $i \in [N]$. MLE is the naive approach of using the empirical distribution with correction bias, [PJW17] is arXiv:1712.07177; all the remaining are cited in our paper. Each data point represents 50 random trials. In our algorithm we pick $threshold = 18$ (same as [WY16]) and our set $\mathrm{F} = [0, 18]$ (input of algorithm 1), meaning, we run PML on frequencies ≤ 18 and empirical on the rest. We use the heuristic algorithm in [PJW17] to compute an approximate PML distribution. Our results are competitive with other state of the art entropy estimators.

Second, to demonstrate the practicality of our approach, we compare the running time of our algorithm using [PJW17] as a subroutine to the raw [PJW17] algorithm on the $\mathrm{Zipf}(1)$ distribution. The second row is the fraction of samples on which our algorithm performs empirical estimate (plus correction bias). The third row is the ratio of running time of [PJW17] to our algorithm. For larger sizes we get $\geq$ 10x speedup.

Samples size	10^3	$5*10^3$	10^4	$5*10^4$	10^5	$5*10^5$	10^6	$5*10^6$	10^7
EmpFrac	0.18382	0.31654	0.37150	0.50457	0.56239	0.69533	0.75245	0.88554	0.94282
Speedup	0.824	1.205	1.669	3.561	4.852	9.552	13.337	12.196	10.204

Reviewer 1: Thank you! As discussed above we will be sure to add more intuition in the final version.

Reviewer 2. Thank you! Comparing the relative strengths of pseudoPML versus PML is an interesting direction for future research. One advantage of pseudoPML is that it is a simple approach that facilitates provable guarantees for a broader range of parameter ϵ. For simulations see above and further we will address all the writing related suggestions.

Reviewer 3: Thank you! See below for response to detailed comments and beginning for response to improvements.

Point 1, 5. Great points; We agree and will address all in the final verison.

Point 2. Beyond the experiments discussed above, we note that the key idea of pseudoPML is to use an algorithm for PML as a subroutine in a black box way. For example, for entropy if we invoke the [CSS19] efficient approximate PML algorithm on frequencies up to $O(\log N)$, then our pseudo-PML algorithm gives a nearly linear time sample-optimal estimator whenever $\epsilon > N^{-0.33}$, improving upon [CSS19] which applied only when $\epsilon > N^{-0.199}$. Similarly for distance to uniformity if we invoke the [CSS19] efficient approximate PML algorithm on $O(\sqrt{n \log n/N})$ distinct frequencies, then our pseudo-PML algorithm gives a nearly linear time sample-optimal estimator whenever $\epsilon > N^{-0.33}$, improving upon [CSS19] which applied only when $\epsilon > N^{-0.249}$.

Point 3. In the proof of Theorem 3.1, the inequality above line 357 should be $\Pr(S(p) = \mathrm{Distinct}(\phi)) >= 1 - k\exp(-n/k)$. Since $\epsilon > 1/k^{0.249}$ is handled in [ADOS16], we consider $\epsilon \leq 1/k^{0.249}$. Using sample size $n = \Theta(k \log^2(1/\epsilon)/\log k)$, we have $n \geq ck \log k$ for constant $c \geq 2$. Combining this with $\Pr(S(p_\phi) = \mathrm{Distinct}(\phi)) = 1$ (where p_ϕ is the PML) we have that our success probability, $\Pr(S(p_\phi) = S(p)) \geq 1 - k\exp(-n/k) \geq 1 - \exp(-\log k)$.

Point 4. We agree that it is more accurate to say our method "weakly depends" on the property and we will update.