We thank the reviewers for their time, and are happy with the generally positive view of our work.

Reviewers 2+3 find that our proposed f-divergence estimator is “applicable to many machine learning problems” (R3) and gives “theoretical justification to several well-established methods” (R2). The “theoretical analysis [of our estimator] is complete” (R3) and provides “nice and comprehensive rates of convergence of both the estimator’s expectation, as well as its MC-estimator, for different f-functions” (R2). Both reviewers recognise the significance of our proposal and are in favour of acceptance.

It seems, however, that there is some disagreement between Reviewers 2+3 and Reviewer 1.

The disagreement seems to stem from Reviewer 1 having misunderstood a fundamental premise of the paper: namely, that we study f-divergence estimation under strong assumptions (lines 55-64) which are nonetheless realistic in many modern applications (lines 76-79, Section 4). Hence, although our setting “seems restrictive” (R1) it is in fact still very much applicable. Moreover, it is precisely because of our assumptions that we derive superior rates compared to the weak-assumption setting (lines 72-75). In light of this, Reviewer 1’s requests to “provide additional theoretical analysis for the estimator” and that “the estimator should be investigated without known density functions” for the conditionals $Q_{Z|X}$ do not make much sense, simply because our proposed estimator can not be computed in this setting.

We hope that this clears up the misunderstanding and that Reviewer 1 will consider raising their rating.

Comments specifically for Reviewer 3:

- Thank you for spotting the overloaded notation. We will fix this to avoid ambiguity.
- We included RAM-MC with both $N = 1$ and $N = 500$ to show that increasing N results in decreased bias and variance in order to validate Theorems 1, 2 & 4. We discuss this in lines 216-217, but will update the paragraph beginning line 192 to also explain why we do this. We did not include $N = 1$ for the other methods because the plots would have become too cluttered.
- Table with other rates: this is a great suggestion, and we will definitely include it. Below is a table with rates and assumptions beneath. We will update Table 1 in our paper accordingly.

Table 1: Rate of bias for estimators of $D_f(P, Q)$.

f-divergence	KL	TV	χ^2	H^2	JS	D_{f_β}		D_{f_α}
						$\frac{1}{2}<\beta<1$	$1<\beta<\infty$	$-1<\alpha<1$
Krishnamurthy et al. [22]	-	-	-	-	-	-	-	$N^{-\frac{1}{2}}+N^{\frac{-3s}{2s+d}}$
Nguyen et al. [28]	$N^{-\frac{1}{2}}$	-	-	-	-	-	-	-
Moon and Hero [26]	$N^{-\frac{1}{2}}$	-	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$

Assumptions: [22]: Both densities p and q must belong to the Hölder class of smoothness s, be supported on $[0, 1]^d$ and satisfy $0 < \eta_1 < p, q < \eta_2 < \infty$ on the support for known constants η_1, η_2. [28]: The density ratio p/q must satisfy $0 < \eta_1 < p/q < \eta_2 < \infty$ and belong to a function class G whose *bracketing entropy* (a measure of the complexity of a function class) is properly bounded. The condition on the bracketing entropy is quite strong and ensures that the density ratio is well behaved. [26]: This estimator makes **strong assumptions** to avoid non-parametric rates. Both p and q must have the same bounded support and satisfy $0 < \eta_1 < p, q < \eta_2 < \infty$ on the support. p and q must have *continuous bounded* derivatives of order d (which is stronger than assumptions of [22]). f has derivatives of order at least d.

References

[22] A. Krishnamurthy, A. Kandasamy, B. Póczos, and L. Wasserman. Nonparametric estimation of Rényi divergence and friends. In ICML, 2014.

[26] K. Moon and A. Hero. Ensemble estimation of multivariate f-divergence. In 2014 IEEE International 360 Symposium on Information Theory, pages 356–360, 2014.

[28] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and 363 the likelihood ratio by convex risk minimization. IEEE Trans. Information Theory, 56(11):5847–5861, 364 2010.