We thank the reviewers for their careful evaluation of our manuscript.

R1 (correctness of prop.5,6,7): We believe that the results in the paper are correct as stated. First, Props 6 and 7 have no issue with disjoint support since $X_n \sim q$; X_n will not fall in a region where $p > 0$ and $q = 0$ almost surely. And although $\mathbb{E}\left[(1+R)^2\right]$ may be infinite, in this case the second bound in Prop 5 says that $d_H(p,q) \geq 0$, which is true (albeit vacuous). Likewise, in the first bound, any time R is large (or infinite) we have that the indicator function returns 0 and the argument in the expectation is $(R/(1+R))^2 \leq 1$.

R1 (uniqueness of $\hat{p}$): In the paragraph before Theorem 2, we note that $\mathcal{Q}$ is the closure of the mixtures, and hence the best possible approximation $\hat{p}$ exists and is unique since $\mathcal{Q}$ is convex and closed. But we realize now that earlier in the paper we denoted $\mathcal{Q}$ to be the finite mixtures (not the closure), which is certainly confusing. Thank you for pointing this out, we will make sure to clarify this and keep notation consistent in the final draft.

R1 (solving eq.4 with $(1-\delta)$ relative error):. A great point! We agree that this was not clear and will improve our explanation for the final draft. By "solving eq. 4 with $(1-\delta)$ relative error", we mean that the component we find has objective value at least $(1-\delta)$ times the maximum value of the optimum (i.e. it is not necessary to solve the problem exactly for our result to hold). Note that this is equivalent for both formulae in eq. 4, since $f - \langle f, \bar{g}_n \rangle \bar{g}_n$ is orthogonal to $\bar{g}_n$ and thus the two only differ by a multiplicative constant $\|f - \langle f, \bar{g}_n \rangle \bar{g}_n\|$. However, in practice we currently cannot assert that we have solved eq. 4 with a certain precision as it is a non-convex optimization problem. We are currently working towards a possible solution to this problem as a follow up to the present work.

R1 (references on l. 37):. Dehaene and Barthelme (2018) is indeed a relevant reference to add here; thanks for pointing this out to us. We will add this as well as a reference to the classical BvM theorem in the final draft.

R3 (disadvantages of Hellinger): Another excellent point. We will include a discussion of the drawbacks of the Hellinger distance in the final draft. While Hellinger has many statistical strengths (as outlined in the text), its weaknesses tend to be computational in nature. In particular, Monte Carlo gradient estimates for our objectives are both biased and tend to be higher variance than the corresponding KL gradient estimates (due to the $\log \mathbb{E}\left[e^{\cdots}\right]$ objective structure), making stochastic optimization a bit more difficult. We have further upcoming work that addresses both of these issues, which we hope to be out shortly.

R3 (nonexponential family components): It is certainly possible that other densities may yield a computationally tractable procedure; as long as the relevant inner products are easily evaluated / approximated, all should work out the same. Note that our theory also holds in the general (possibly misspecified) setting. Generally though, exponential families are a broad enough class to capture most settings of interest (full support, nonnegative, discrete, etc).

R3 (BBVI): Good eye – we will ensure that this initialism ("boosting black box variational inference") is defined and cited more directly (ref [27]) in the final draft.

R3 (complexity): The main cost of our algorithm lies in finding the next component, which is a nonconvex optimization problem; characterizing the complexity of such problems is generally nontrivial. The other expensive step is inverting Z in the weights optimization, where incremental methods using block matrix inversion reduce the cost at iteration n to $O(n^2)$ and overall cost to $O(N^3)$. Note that this was discussed in the paper just before Section 3.2 on p. 5.

R3 (scaling to high dimensions / large data): In order to make the method computationally tractable in high dimensions, we switch to diagonal covariances in the mixture components (as mentioned at the beginning of the experiments section). Note that our theoretical guarantees do not explicitly depend on dimension and hold equally well in the high-dimensional setting; the variable τ implicitly serves the role of "dimension" and captures the complexity of the densities to approximate (as mentioned just before eq. 2). Further, as mentioned in the appendix, our algorithm should also scale to the big data setting by using Monte Carlo gradient estimates similar to those in Rényi variational inference (Li & Turner 2016); but see the remark below about this setting. We will add remarks about dimension, diagonal vs full rank covariances, as well as the big data setting, in the final draft.

R2,3 (empirical results):. We certainly agree that testing the approach on high dimensional problems with large datasets is an important task to complete. But note that the primary purpose of the present work was to obtain a tuning-free variational method with statistical quality guarantees. To the best of our knowledge, this is the first method that accomplishes this goal, and we believe our experiments justify the claims in the paper. A comprehensive treatment of the high-dimensional / large-data regimes, with guarantees / computational techniques directed to these settings, is the focus of an ongoing follow-up project; based on the empirical results from recent work on Hellinger VI (Li & Turner 2016) and boosting VI (Locatello et al. 2018), we believe it will be possible to successfully apply our method to these settings.