We thank the reviewers for their insightful and constructive comments.

Reviewer #1 and shared comments.

- ***(Shared by R2, R3) Stability of importance sampling, discussion & analysis on choice of proposal*** $q(x)$***.*** This is a challenge shared by almost all non-parametric density estimation models (e.g., NCE, DDE). Experimentally, our FML outperforms its counterparts which also involve importance sampling estimates. To keep the variance in check, a general guiding principle for choosing a good $q(x)$ is to make it as close to target $p(x)$ as possible, which is expected to yield small var[p/q]. To this end, for explicit FML we have proposed to use a pre-trained tractable sampler $q(x)$ modeled with generative flows (SM L.6, likelihood maximized wrt empirical data; other models like GMM are also applicable). For latent FML we maximize the mutual information (Sec 3.3). We have revised the paper to expand the theoretical discussions and elaborate implementation details on the choice of $q(x)$. Empirical comparisons with different proposals are also added for sensitivity analysis and show the gains with a good proposal, with both simple and complex datasets.
- ***MC samples & convergence.*** We remind the reviewer that one of the key features of our FML framework is that we replace the direct estimation of normalizing constant (typically requires multiple MC samples) with an optimization procedure, such that under the SGD setup 1 MC sample suffices. For convergence guarantees, we have proved with FML model parameters converge to the correct answer under both convex setting (Col 2.3) and more general non-convex setting (SM Thm G.2). Other competing MC-based solutions generally cannot guarantee this under finite sample.

Reviewer #2. We thank the reviewer for this very comprehensive review, which we really appreciate.

- ***Highlighting the contribution of unbiased estimation of likelihood.*** We agree this point needs to be reinforced. It is the key motivation of this study and we have rewritten relevant sections in the paper to reflect the reviewer's inputs. We have also added the discussion of the biased-estimation issue to the main part and updated the figs as suggested.
- ***Finite sample estimate of the partition.*** Our FML treats the partition function as a learnable parameter that is updated with *finite sample evaluations* of the inverse likelihood, so that the objective does not involve a $\log$ transform. Technically it is not a (direct) finite sample estimator. This differs from a direct $\log$ (finite sample estimate) adopted by competing solutions, which lead to biased estimation/gradient of likelihood, a key challenge that FML addressed. We agree FML itself cannot sidestep the challenge of choosing efficient sampling schemes for the evaluation of the inverse likelihood integral (i.e., choice of proposal $q(x)$ used in SGD), which is discussed in detail in our reply to R1 above.
- ***What's gained by the minimax game over plain MLE.*** In our FML the log-partition is modeled as a learnable parameter, and our theory guarantees convergence to the correct answer as long as the log-partition is estimated with bounded error. The major gain of minimax FML is unbiased estimation for *unnormalized* statistical models and latent variable models, where the exact likelihood is intractable and existing solutions typically settle for bounds.
- ***L127 Is*** b_θ ***fixed or learned.*** This is a misunderstanding that will be made more clear in our revision. $\log$-partition estimate b_θ is a free-parameter to be learned, and b_θ minimizes the objective iff it equals to the true log-partition.
- ***Why called a minimax formulation.*** We agree the explicit FML (Eq 3) can be understood as a min-min game, but since the latent FML formulation (Eq 7) is a strict min-max game, calling it a minimax game is more consistent.
- ***Response to improvement suggestions.*** We have rewritten relevant sections to highlight that our FML provides an unbiased estimate of the log-likelihood using the Fenchel mini-max setup as a key contribution, addressing a long-standing challenge in statistical estimation. We will remove, rephrase or clarify the claims the reviewer found inaccurate/confusing/unjustified. While current submission already includes experiments on high-dimensional complex data (e.g., image, language) with the latent variable FML, we will report more results with explicit FML in our revision.
- ***Misc issues.*** We thank the reviewer for mentioning additional relevant literature ([a-f]), which have been updated to our draft with the suggested discussions. The manuscript has been revised to clarify CD and correct tech conditions used in our theory. Edits are also made to incorporated all other suggestions, which further improved our presentation.

Reviewer #3.

- ***Derivation of Eq 2.*** Our math as presented is correct; the reviewer must have missed a minus sign somewhere. To derive Eq 2, let $t = \frac{1}{p(x)}$ and we have $-\log p(x) = -(-\log t) = -(\max_u\{-u - \exp(-u)t + 1\}) = \min_u\{u + \exp(-u)t - 1\}$. Our SM includes more on this equation, see our code there for implementation details.
- ***What's the benefit of using inverse likelihood MC evaluation over direct likelihood estimate.*** To understand the benefits we need to clarify how we estimate the likelihood (Eq 3) with how we compute the gradient for model updates (Eqs 4-6). FML likelihood is estimated through optimization (Eq 3, min step), and later used to adjust for the scaling of model parameter gradient (Eq 6) computed from MC inverse likelihood evaluations (Eq 4, unbiased). As a result, FML guarantees the model parameters θ will converge to the right answer even with less accurate likelihood estimate (bounded error, Col 2.3, SM Thm G.2). On the other hand, computing the gradient directly with a direct MC likelihood estimate introduces bias when updating model parameters (SM Sec C), and there is no guarantee of convergence with finite samples.
- ***Can it be generalized.*** The Fenchel conjugacy technique is applicable for other convex functions, with which more general likelihood evidence scores can be defined (ref [54]). However, (a) such criteria are less popular in practice; (b) Fenchel conjugacy does not necessarily have a closed form for an arbitrary convex function; and (c) unlike $\log(t)$ the unbiased estimation cannot be guaranteed in general. Further investigation is warranted for future study.