We appreciate the constructive feedbacks from all reviewers, which will be taken into account when revising our paper.

**Reviewer #1**: Our work focuses on the setting of small erased data $\mathcal{D}_e$ (line 28). Since the effect of erasing $\mathcal{D}_e$ from small training data $\mathcal{D}$ is more noticeable when evaluating our methods, we use small- to moderate-sized $\mathcal{D}$ in our experiments. To scale to massive datasets, recall that our methods are parsimonious in requiring only $q(\boldsymbol{\theta}|\mathcal{D})$ and $\mathcal{D}_e$ (line 125), i.e., independent of $|\mathcal{D}|$. For example, we use larger-scale sparse GP models for regression on the complex ***Airline dataset*** ($|\mathcal{D}| = 2\text{M}$, $|\mathcal{D}_e| = 100\text{K}$) (Hensman et al., 2013): With $\lambda = 0$, EUBO and reverse KL are capable of unlearning by, respectively, achieving $\text{KL}[\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel q(\boldsymbol{\theta}|\mathcal{D}_r)] = 1697.25$ and $\text{KL}[\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r) \parallel q(\boldsymbol{\theta}|\mathcal{D}_r)] = 455.65$ which are smaller than $\text{KL}[q(\boldsymbol{\theta}|\mathcal{D}) \parallel q(\boldsymbol{\theta}|\mathcal{D}_r)] = 4344.09$.

We do not know any unlearning work for approximate Bayesian models; existing works consider MAP and MLE (e.g., ridge linear regression, logistic regression). So, there is no suitable existing work for comparison in our experiments.

The disadvantage of overestimating variance in reverse KL can be understood in our study in App. D (referred to in lines 239-247). In Appendix D, we also highlight and discuss practical implications of the main limitation of our approach when the erased data is informative and only an approximate posterior belief is available (lines 491-501).

To unlearn MCMC, we can re-weight (like importance sampling) MCMC samples by $1/p(\mathcal{D}_e|\boldsymbol{\theta})$ from Eq. (2). We will consider Laplace approximation for future work. We hope the above results would improve your opinion of our work.

**Reviewer #2**: As you have noticed, our unlearning performance improves when the approximation of the full-data posterior belief improves due to the challenging constraint of unknown exact full-data posterior belief (lines 125-127). The Airline experiment described in first paragraph for Reviewer #1 shows the scalability of our methods to a massive dataset (hence, more expensive model), which will be included in our revised paper. The limitation of our approach is the dependence on the approximation quality of the posterior belief, which is discussed in Appendix D.

We are not aware of any unlearning work for approximate Bayesian models (i.e., approximate posteriors instead of MAP or MLE). Therefore, there is no suitable existing work for empirical comparison.

Line 282 is not validated by a flow-based approach but with a multivariate Gaussian approximation (full covariance matrix) in Appendix E. We will clarify this and address your other comments (e.g., experimental details) in our revision.

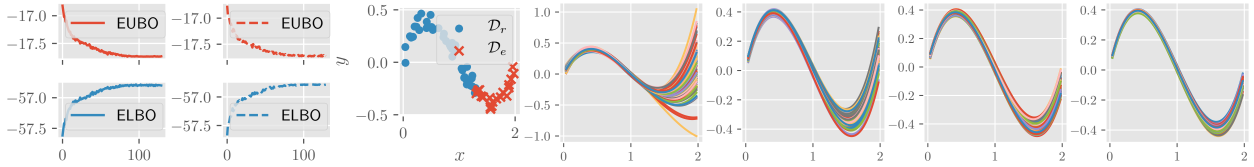
**Reviewer #3**: As you have noticed, both EUBO and ELBO minimize the same KL term $\text{KL}[q(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D}_r)]$, which guarantees their optimal solutions to be the same. We will show an empirical analysis here using the example of Bayesian linear regression: Fig. (a) below shows both EUBO and ELBO values when minimizing EUBO, while Fig. (b) shows their values when maximizing ELBO. We can observe that by minimizing EUBO, we maximize ELBO stably, and vice versa. However, EUBO and ELBO are bounding different quantities, i.e., $\log p(\mathcal{D}_e|\mathcal{D}_r) \neq \log p(\mathcal{D}_r)$. Therefore, the gap between them is not meaningful. We will discuss the above empirical analysis in our revised paper.

Both EUBO with adjusted likelihood and reverse KL can perform well as they are designed to resolve the issue in Remark 1 (lines 172-76, 180-81, 186-89). But, EUBO requires a more careful fine-tuning of $\lambda$ to perform well.

We will include more experimental details (which can be extracted from submitted code) and address your other comments in the revised paper. We hope that the above clarifications would improve your opinion of our work.

**Reviewer #4**: We perform a simple Bayesian regression $y_x = ax^3 + bx^2 + cx + d + \epsilon$ where $a = 2$, $b = -3$, $c = 1$, $d = 0$, and $\epsilon \sim \mathcal{N}(0, 0.05^2)$. Fig. (c) shows the data. The low-rank approximation of the posterior beliefs are diagonal Gaussians. Fig. (d) shows samples of $p(y_x|\mathcal{D}_r)$ (exact). Though reverse KL in Fig. (f) and EUBO in Fig. (g) generate different distributions from the exact $p(y_x|\mathcal{D}_r)$, they resemble $q(y_x|\mathcal{D}_r)$.

Following your suggestion, let $p(\theta|\mathcal{D}) = 0.5\phi(\theta; 0, 1) + 0.5\phi(\theta; 2, 1)$ be a Gaussian mixture (bi-modal) where $\phi(\theta; \mu, \sigma^2)$ is a Gaussian p.d.f. To easily compare the distributions, let the likelihood of erased data be $p(\mathcal{D}_e|\theta) = 1 + \phi(\theta; 2, 1)/\phi(\theta; 0, 1)$. So, $p(\theta|\mathcal{D}_r) = \phi(\theta; 0, 1)$ is a Gaussian by Eq. (2). Supposing the approximate posterior beliefs are Gaussians, minimizing the KL to the Gaussian mixture $p(\theta|\mathcal{D})$ (or, equivalently, maximizing ELBO) gives $q(\theta|\mathcal{D}) = \phi(\theta; 1.004, 1.390^2)$. Then, given only $q(\theta|\mathcal{D})$ and $p(\mathcal{D}_e|\theta)$, we can compute $\tilde{q}_u(\theta|\mathcal{D}_r; \lambda = 0) = \phi(\theta; 0.060, 1.000^2)$ (minimizing EUBO) and $\tilde{q}_v(\theta|\mathcal{D}_r; \lambda = 0) = \phi(\theta; 0.0618, 1.0184^2)$ (minimizing reverse KL). Hence, EUBO and reverse KL perform reasonably well (by being close to $p(\theta|\mathcal{D}_r) = \phi(\theta; 0, 1)$) even when $p(\theta|\mathcal{D})$ is bi-modal. We will include the above results in our revised paper and hope that they would improve your opinion of our work.



(a) Unlearn (b) Retrain (c) Data (d) $p(y_x|\mathcal{D}_r)$ (e) $q(y_x|\mathcal{D}_r)$ (f) $\tilde{q}_v(y_x|\mathcal{D}_r)$ (g) $\tilde{q}_u(y_x|\mathcal{D}_r)$