Variational Bayesian Unlearning

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet[†]
Dept. of Computer Science, National University of Singapore, Republic of Singapore
Dept. of Electrical Engineering and Computer Science, MIT, USA[†]
{qphong,lowkh}@comp.nus.edu.sg, jaillet@mit.edu[†]

Abstract

This paper studies the problem of approximately unlearning a Bayesian model from a small subset of the training data to be erased. We frame this problem as one of minimizing the Kullback-Leibler divergence between the approximate posterior belief of model parameters after directly unlearning from erased data vs. the exact posterior belief from retraining with remaining data. Using the *variational inference* (VI) framework, we show that it is equivalent to minimizing an evidence upper bound which trades off between fully unlearning from erased data vs. not entirely forgetting the posterior belief given the full data (i.e., including the remaining data); the latter prevents catastrophic unlearning that can render the model useless. In model training with VI, only an approximate (instead of exact) posterior belief given the full data can be obtained, which makes unlearning even more challenging. We propose two novel tricks to tackle this challenge. We empirically demonstrate our unlearning methods on Bayesian models such as sparse Gaussian process and logistic regression using synthetic and real-world datasets.

1 Introduction

Our interactions with *machine learning* (ML) applications have surged in recent years such that large quantities of users' data are now deeply ingrained into the ML models being trained for these applications. This greatly complicates the regulation of access to each user's data or implementation of *personal data ownership*, which are enforced by the General Data Protection Regulation in the European Union [24]. In particular, if a user would like to exercise her *right to be forgotten* [24] (e.g., when quitting an ML application), then it would be desirable to have the trained ML model "unlearn" from her data. Such a problem of *machine unlearning* [4] extends to the practical scenario where a small subset of data previously used for training is later identified as malicious (e.g., anomalies) [4, 9] and the trained ML model can perform well once again if it can unlearn from the malicious data.

A naive alternative to machine unlearning is to simply retrain an ML model from scratch with the data *remaining* after *erasing* that to be unlearned from. In practice, this is prohibitively expensive in terms of time and space costs since the remaining data is often large such as in the above scenarios. How then can a trained ML model *directly* and efficiently unlearn from a small subset of data to be erased to become (a) exactly and if not, (b) approximately close to that from retraining with the large remaining data? Unfortunately, (a) exact unlearning is only possible for selected ML models (e.g., naive Bayes classifier, linear regression, k-means clustering, and item-item collaborative filtering [4, 12, 30]). This motivates the need to consider (b) approximate unlearning as it is applicable to a broader family of ML models like neural networks [9, 13] but, depending on its choice of loss function, may suffer from *catastrophic unlearning*[1] that can render the model useless. For example, to mitigate this issue, the works of [9, 13] have to "patch up" their loss functions by additionally bounding the loss incurred

[1] A trained ML model is said to experience *catastrophic unlearning* from the erased data when its resulting performance is considerably worse than that from retraining with the remaining data.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

by erased data with a rectified linear unit and injecting a regularization term to retain information of the remaining data, respectively. This begs the question whether there exists a loss function that can *directly* quantify the approximation gap and *naturally* prevent catastrophic unlearning.

Our work here addresses the above question by focusing on the family of Bayesian models. Specifically, our proposed loss function measures the *Kullback-Leibler* (KL) divergence between the approximate posterior belief of model parameters by directly unlearning from erased data vs. the exact posterior belief from retraining with remaining data. Using the *variational inference* (VI) framework, we show that minimizing this KL divergence is equivalent to *minimizing* (instead of maximizing) a counterpart of the evidence lower bound called the *evidence upper bound* (EUBO) (Sec. 3.2). Interestingly, the EUBO lends itself to a natural interpretation of a trade-off between fully unlearning from erased data vs. not entirely forgetting the posterior belief given the *full* data (i.e., including the remaining data); the latter prevents catastrophic unlearning induced by the former.

Often, in model training, only an approximate (instead of exact) posterior belief of model parameters given the full data can be learned, say, also using VI. This makes unlearning even more challenging. To tackle this challenge, we analyse two sources of inaccuracy in the approximate posterior belief learned using VI, which lay the groundwork for proposing our first trick of an *adjusted likelihood* of erased data (Sec. 3.3.1): Our key idea is to curb unlearning in the region of model parameters with low approximate posterior belief where both sources of inaccuracy primarily occur. Additionally, to avoid the risk of incorrectly tuning the adjusted likelihood, we propose another trick of *reverse KL* (Sec. 3.3.2) which is naturally more protected from such inaccuracy without needing the adjusted likelihood. Nonetheless, our adjusted likelihood is general enough to be applied to reverse KL.

VI is a popular approximate Bayesian inference framework due to its scalability to massive datasets [15, 18] and its ability to model complex posterior beliefs using generative adversarial networks [33] and normalizing flows [21, 29]. Our work in this paper exploits VI to broaden the family of ML models that can be unlearned, which we empirically demonstrate using synthetic and real-world datasets on several Bayesian models such as sparse Gaussian process and logistic regression with the approximate posterior belief modeled by a normalizing flow (Sec. 4).

2 Variational Inference (VI)

In this section, we revisit the VI framework [2] for learning an approximate posterior belief of the parameters $\boldsymbol{\theta}$ of a Bayesian model. Given a prior belief $p(\boldsymbol{\theta})$ of the unknown model parameters $\boldsymbol{\theta}$ and a set $\mathcal{D}$ of training data, an approximate posterior belief $q(\boldsymbol{\theta}|\mathcal{D}) \approx p(\boldsymbol{\theta}|\mathcal{D})$ is being optimized by minimizing the KL divergence $\mathrm{KL}[q(\boldsymbol{\theta}|\mathcal{D}) \parallel p(\boldsymbol{\theta}|\mathcal{D})] \triangleq \int q(\boldsymbol{\theta}|\mathcal{D}) \ \log(q(\boldsymbol{\theta}|\mathcal{D})/p(\boldsymbol{\theta}|\mathcal{D}))\ \mathrm{d}\boldsymbol{\theta}$ or, equivalently, maximizing the *evidence lower bound* (ELBO) $\mathcal{L}$ [2]:

$$\mathcal{L} \triangleq \int q(\boldsymbol{\theta}|\mathcal{D}) \ \log p(\mathcal{D}|\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta} - \mathrm{KL}[q(\boldsymbol{\theta}|\mathcal{D}) \parallel p(\boldsymbol{\theta})]\ . \tag{1}$$

Such an equivalence follows directly from $\mathcal{L} = \log p(\mathcal{D}) - \mathrm{KL}[q(\boldsymbol{\theta}|\mathcal{D}) \parallel p(\boldsymbol{\theta}|\mathcal{D})]$ where the log-marginal likelihood $\log p(\mathcal{D})$ is independent of $q(\boldsymbol{\theta}|\mathcal{D})$. Since $\mathrm{KL}[q(\boldsymbol{\theta}|\mathcal{D}) \parallel p(\boldsymbol{\theta}|\mathcal{D})] \geq 0$, the ELBO $\mathcal{L}$ is a lower bound of $\log p(\mathcal{D})$. The ELBO $\mathcal{L}$ in (1) can be interpreted as a trade-off between attaining a higher likelihood of $\mathcal{D}$ (first term) vs. not entirely forgetting the prior belief $p(\boldsymbol{\theta})$ (second term).

When the ELBO $\mathcal{L}$ (1) cannot be evaluated in closed form, it can be maximized using *stochastic gradient ascent* (SGA) by approximating the expectation in

$$\mathcal{L} = \mathbb{E}_{q(\boldsymbol{\theta}|\mathcal{D})}[\log p(\mathcal{D}|\boldsymbol{\theta}) + \log(p(\boldsymbol{\theta})/q(\boldsymbol{\theta}|\mathcal{D}))] = \int q(\boldsymbol{\theta}|\mathcal{D})\ (\log p(\mathcal{D}|\boldsymbol{\theta}) + \log(p(\boldsymbol{\theta})/q(\boldsymbol{\theta}|\mathcal{D})))\ \mathrm{d}\boldsymbol{\theta}$$

with stochastic sampling in each iteration of SGA. The approximate posterior belief $q(\boldsymbol{\theta}|\mathcal{D})$ can be represented by a simple distribution (e.g., in the exponential family) for computational ease or a complex distribution (e.g., using generative neural networks) for expressive power. Note that when the distribution of $q(\boldsymbol{\theta}|\mathcal{D})$ is modeled by a generative neural network whose density cannot be evaluated, the ELBO can be maximized with adversarial training by alternating between estimating the log-density ratio $\log(p(\boldsymbol{\theta})/q(\boldsymbol{\theta}|\mathcal{D}))$ and maximizing the ELBO [33]. On the other hand, when the distribution of $q(\boldsymbol{\theta}|\mathcal{D})$ is modeled by a normalizing flow (e.g., *inverse autoregressive flow* (IAF) [21]) whose density can be computed, the ELBO can be maximized with SGA.

3 Bayesian Unlearning

3.1 Exact Bayesian Unlearning

Let the (*full*) training data $\mathcal{D}$ be partitioned into a small subset $\mathcal{D}_e$ of data to be *erased* and a (large) set $\mathcal{D}_r$ of *remaining* data, i.e., $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_e$ and $\mathcal{D}_r \cap \mathcal{D}_e = \emptyset$. The problem of *exact Bayesian unlearning* involves recovering the exact posterior belief $p(\boldsymbol{\theta}|\mathcal{D}_r)$ of model parameters $\boldsymbol{\theta}$ given remaining data $\mathcal{D}_r$ from that given full data $\mathcal{D}$ (i.e., $p(\boldsymbol{\theta}|\mathcal{D})$ assumed to be available) by directly unlearning from erased data $\mathcal{D}_e$. Note that $p(\boldsymbol{\theta}|\mathcal{D}_r)$ can also be obtained from retraining with remaining data $\mathcal{D}_r$, which is computationally costly, as discussed in Sec. 1. By using Bayes' rule and assuming conditional independence between $\mathcal{D}_r$ and $\mathcal{D}_e$ given $\boldsymbol{\theta}$,

$$p(\boldsymbol{\theta}|\mathcal{D}_r) \;=\; p(\boldsymbol{\theta}|\mathcal{D})\, p(\mathcal{D}_e|\mathcal{D}_r)/p(\mathcal{D}_e|\boldsymbol{\theta}) \;\propto\; p(\boldsymbol{\theta}|\mathcal{D})/p(\mathcal{D}_e|\boldsymbol{\theta})\;. \tag{2}$$

When the model parameters $\boldsymbol{\theta}$ are discrete-valued, $p(\boldsymbol{\theta}|\mathcal{D}_r)$ can be obtained from (2) directly. The use of a conjugate prior also makes unlearning relatively simple. We will investigate the more interesting case of a non-conjugate prior in the rest of Sec. 3.

3.2 Approximate Bayesian Unlearning with Exact Posterior Belief $p(\boldsymbol{\theta}|\mathcal{D})$

The problem of *approximate Bayesian unlearning* differs from that of exact Bayesian unlearning (Sec. 3.1) in that only the approximate posterior belief $q_u(\boldsymbol{\theta}|\mathcal{D}_r)$ (instead of the exact one $p(\boldsymbol{\theta}|\mathcal{D}_r)$) can be recovered by directly unlearning from erased data $\mathcal{D}_e$. Since existing unlearning methods often use their model predictions to construct their loss functions [3, 4, 12, 14], we have initially considered doing likewise (albeit in the Bayesian context) by defining the loss function as the KL divergence between the approximate predictive distribution $q_u(y|\mathcal{D}_r) \triangleq \int p(y|\boldsymbol{\theta})\; q_u(\boldsymbol{\theta}|\mathcal{D}_r)\; \mathrm{d}\boldsymbol{\theta}$ vs. the exact predictive distribution $p(y|\mathcal{D}_r) = \int p(y|\boldsymbol{\theta})\; p(\boldsymbol{\theta}|\mathcal{D}_r)\; \mathrm{d}\boldsymbol{\theta}$ where the observation y (i.e., drawn from a model with parameters $\boldsymbol{\theta}$) is conditionally independent of $\mathcal{D}_r$ given $\boldsymbol{\theta}$. However, it may not be possible to evaluate these predictive distributions in closed form, hence making the optimization of this loss function computationally difficult. Fortunately, such a loss function can be bounded from above by the KL divergence between posterior beliefs $q_u(\boldsymbol{\theta}|\mathcal{D}_r)$ vs. $p(\boldsymbol{\theta}|\mathcal{D}_r)$, as proven in Appendix A:

Proposition 1. $\mathrm{KL}[q_u(y|\mathcal{D}_r) \parallel p(y|\mathcal{D}_r)] \leq \mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D}_r)]$.[2]

Proposition 1 reveals that reducing $\mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D}_r)]$ decreases $\mathrm{KL}[q_u(y|\mathcal{D}_r) \parallel p(y|\mathcal{D}_r)]$, thus motivating its use as the loss function instead. In particular, it follows immediately from our result below (i.e., proven in Appendix B) that minimizing $\mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D}_r)]$ is equivalent to *minimizing* a counterpart of the ELBO called the *evidence upper bound* (EUBO) $\mathcal{U}$:

Proposition 2. Define the EUBO $\mathcal{U}$ as

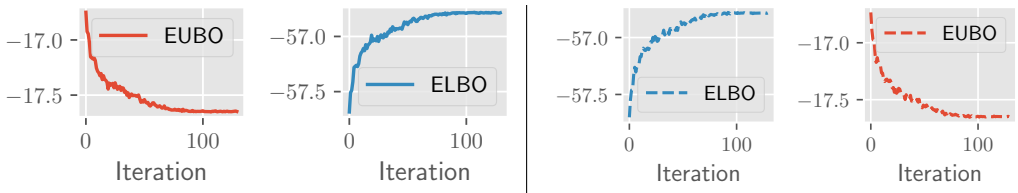
$$\mathcal{U} \triangleq \int q_u(\boldsymbol{\theta}|\mathcal{D}_r)\; \log p(\mathcal{D}_e|\boldsymbol{\theta})\; \mathrm{d}\boldsymbol{\theta} + \mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D})]\;. \tag{3}$$

Then, $\mathcal{U} = \log p(\mathcal{D}_e|\mathcal{D}_r) + \mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D}_r)] \geq \log p(\mathcal{D}_e|\mathcal{D}_r)$ such that $p(\mathcal{D}_e|\mathcal{D}_r)$ is independent of $q_u(\boldsymbol{\theta}|\mathcal{D}_r)$.

From Proposition 2, minimizing EUBO (3) is equivalent to minimizing $\mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D}_r)]$ which is precisely achieved using VI (i.e., by maximizing ELBO (1)) from retraining with remaining data $\mathcal{D}_r$. This is illustrated in Fig. 1a where unlearning from $\mathcal{D}_e$ by minimizing EUBO maximizes ELBO w.r.t. $\mathcal{D}_r$; in Fig. 1b, retraining with $\mathcal{D}_r$ by maximizing ELBO minimizes EUBO w.r.t. $\mathcal{D}_e$.

The EUBO $\mathcal{U}$ (3) can be interpreted as a trade-off between fully unlearning from erased data $\mathcal{D}_e$ (first term) vs. not entirely forgetting the exact posterior belief $p(\boldsymbol{\theta}|\mathcal{D})$ given the full data $\mathcal{D}$ (i.e., including the remaining data $\mathcal{D}_r$) (second term). The latter can be viewed as a regularization term to prevent catastrophic unlearning[1] (i.e., potentially induced by the former) that *naturally* results from minimizing our loss function $\mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D}_r)]$, which differs from the works of [9, 13] needing to "patch up" their loss functions (Sec. 1). Generative models can be used to model the approximate posterior belief $q_u(\boldsymbol{\theta}|\mathcal{D}_r)$ in the EUBO $\mathcal{U}$ (3) in the same way as that in the ELBO $\mathcal{L}$ (1).

[2]Similarly, $\mathrm{KL}[p(y|\mathcal{D}_r) \parallel q_u(y|\mathcal{D}_r)] \leq \mathrm{KL}[p(\boldsymbol{\theta}|\mathcal{D}_r) \parallel q_u(\boldsymbol{\theta}|\mathcal{D}_r)]$ holds.



(a) Unlearning from $\mathcal{D}_e$ by minimizing EUBO (b) Retraining with $\mathcal{D}_r$ by maximizing ELBO

Figure 1: Plots of EUBO and ELBO when (a) unlearning from $\mathcal{D}_e$ and (b) retraining with $\mathcal{D}_r$.

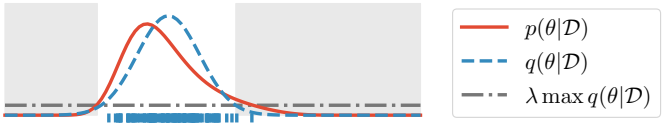


Figure 2: Plot of $q(\boldsymbol{\theta}|\mathcal{D})$ learned using VI. Gray shaded region corresponds to values of $\boldsymbol{\theta}$ where $q(\boldsymbol{\theta}|\mathcal{D}) \leq \lambda \max_{\boldsymbol{\theta}'} q(\boldsymbol{\theta}'|\mathcal{D})$. Vertical blue strips on horizontal axis show 100 samples of $\boldsymbol{\theta} \sim q(\boldsymbol{\theta}|\mathcal{D})$.

3.3 Approximate Bayesian Unlearning with Approximate Posterior Belief $q(\boldsymbol{\theta}|\mathcal{D})$

Often, in model training, only an approximate posterior belief[3] $q(\boldsymbol{\theta}|\mathcal{D})$ (instead of the exact $p(\boldsymbol{\theta}|\mathcal{D})$ in Sec. 3.2) of model parameters $\boldsymbol{\theta}$ given full data $\mathcal{D}$ can be learned, say, using VI by maximizing the ELBO (Sec. 2). Our proposed unlearning methods are parsimonious in requiring only $q(\boldsymbol{\theta}|\mathcal{D})$ and erased data $\mathcal{D}_e$ to be available, which makes unlearning even more challenging since there is no further information about $p(\boldsymbol{\theta}|\mathcal{D})$ nor the difference between $p(\boldsymbol{\theta}|\mathcal{D})$ vs. $q(\boldsymbol{\theta}|\mathcal{D})$. So, we estimate the unknown $p(\boldsymbol{\theta}|\mathcal{D}_r)$ (2) with

$$\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r) \;\propto\; q(\boldsymbol{\theta}|\mathcal{D})/p(\mathcal{D}_e|\boldsymbol{\theta}) \tag{4}$$

and minimize the KL divergence between the approximate posterior belief recovered by directly unlearning from erased data $\mathcal{D}_e$ vs. $\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r)$ (4) instead. We will discuss two novel tricks below to alleviate the undesirable consequence of using $\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r)$ instead of the unknown $p(\boldsymbol{\theta}|\mathcal{D}_r)$ (2).

3.3.1 EUBO with Adjusted Likelihood

Let the loss function $\mathrm{KL}[\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel \tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r)]$ be minimized w.r.t. the approximate posterior belief $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r)$ that is recovered by directly unlearning from erased data $\mathcal{D}_e$. Similar to Proposition 2, $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r)$ can be optimized by minimizing the following EUBO:

$$\widetilde{\mathcal{U}} \triangleq \int \tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r)\; \log p(\mathcal{D}_e|\boldsymbol{\theta})\; \mathrm{d}\boldsymbol{\theta} + \mathrm{KL}[\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel q(\boldsymbol{\theta}|\mathcal{D})] \tag{5}$$

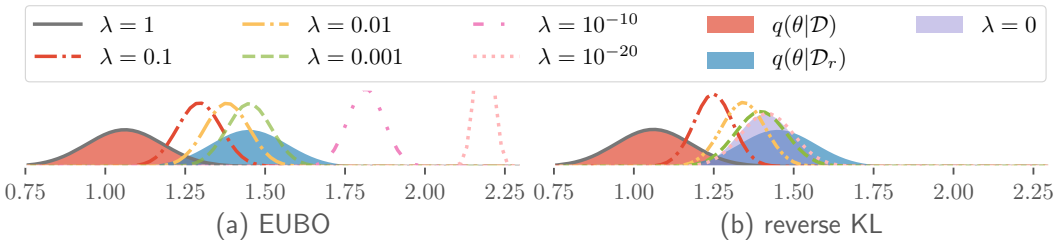
which follows from simply replacing the unknown $p(\boldsymbol{\theta}|\mathcal{D})$ in $\mathcal{U}$ (3) with $q(\boldsymbol{\theta}|\mathcal{D})$. We discuss the difference between $p(\boldsymbol{\theta}|\mathcal{D})$ vs. $q(\boldsymbol{\theta}|\mathcal{D})$ in the remark below:

Remark 1. We analyze two possible sources of inaccuracy in $q(\boldsymbol{\theta}|\mathcal{D})$ that is learned using VI by minimizing the loss function $\mathrm{KL}[q(\boldsymbol{\theta}|\mathcal{D}) \parallel p(\boldsymbol{\theta}|\mathcal{D})]$ (Sec. 2). Firstly, $q(\boldsymbol{\theta}|\mathcal{D})$ often underestimates the variance of $p(\boldsymbol{\theta}|\mathcal{D})$: Though $q(\boldsymbol{\theta}|\mathcal{D})$ tends to be close to 0 at values of $\boldsymbol{\theta}$ where $p(\boldsymbol{\theta}|\mathcal{D})$ is close to 0, the reverse is not enforced [1] (see, for example, Fig. 2). So, $q(\boldsymbol{\theta}|\mathcal{D})$ can differ from $p(\boldsymbol{\theta}|\mathcal{D})$ at values of $\boldsymbol{\theta}$ where $q(\boldsymbol{\theta}|\mathcal{D})$ is close to 0. Secondly, if $q(\boldsymbol{\theta}|\mathcal{D})$ is learned through stochastic optimization of the ELBO (i.e., with stochastic samples of $\boldsymbol{\theta} \sim q(\boldsymbol{\theta}|\mathcal{D})$ in each iteration of SGA), then it is unlikely that the ELBO is maximized using samples of $\boldsymbol{\theta}$ with small $q(\boldsymbol{\theta}|\mathcal{D})$ (Fig. 2). Thus, both sources of inaccuracy primarily occur at values of $\boldsymbol{\theta}$ with small $q(\boldsymbol{\theta}|\mathcal{D})$. Though it can also be inaccurate at values of $\boldsymbol{\theta}$ with large $q(\boldsymbol{\theta}|\mathcal{D})$, such an inaccuracy can be reduced by representing $q(\boldsymbol{\theta}|\mathcal{D})$ with a complex distribution (Sec. 2).

Remark 1 motivates us to curb unlearning at values of $\boldsymbol{\theta}$ with small $q(\boldsymbol{\theta}|\mathcal{D})$ by proposing our first novel trick of an adjusted likelihood of the erased data:

$$p_{\text{adj}}(\mathcal{D}_e|\boldsymbol{\theta};\lambda) \triangleq \begin{cases} p(\mathcal{D}_e|\boldsymbol{\theta}) & \text{if } q(\boldsymbol{\theta}|\mathcal{D}) > \lambda \max_{\boldsymbol{\theta}'} q(\boldsymbol{\theta}'|\mathcal{D})\;, \\ 1 & \text{otherwise (i.e., shaded area in Fig. 2)}\;; \end{cases} \tag{6}$$

[3]With a slight abuse of notation, we let $q(\boldsymbol{\theta}|\mathcal{D})$ be the approximate posterior belief that maximizes the ELBO $\mathcal{L}$ (1) (Sec. 2) from Sec. 3.3 onwards.



(a) EUBO (b) reverse KL

Figure 3: Plot of approximate posterior beliefs with varying λ obtained by minimizing (a) EUBO (i.e., $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$) and (b) reverse KL (i.e., $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$); horizontal axis denotes $\theta = \alpha$. In (a), a huge probability mass of $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r, \lambda = 0)$ is at large values of α beyond the plotting area and the top of the plot of $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r, \lambda = 10^{-20})$ is cut off due to lack of space.

for any $\boldsymbol{\theta}$ where $\lambda \in [0, 1]$ controls the magnitude of a threshold under which $q(\boldsymbol{\theta}|\mathcal{D})$ is considered small. To understand the effect of λ, let $\tilde{p}_{\text{adj}}(\boldsymbol{\theta}|\mathcal{D}_r;\lambda) \propto q(\boldsymbol{\theta}|\mathcal{D})/p_{\text{adj}}(\mathcal{D}_e|\boldsymbol{\theta};\lambda)$, i.e., by replacing $p(\mathcal{D}_e|\boldsymbol{\theta})$ in (4) with $p_{\text{adj}}(\mathcal{D}_e|\boldsymbol{\theta};\lambda)$. Then, using (6),

$$\tilde{p}_{\text{adj}}(\boldsymbol{\theta}|\mathcal{D}_r;\lambda) \propto \begin{cases} q(\boldsymbol{\theta}|\mathcal{D})/p(\mathcal{D}_e|\boldsymbol{\theta}) & \text{if } q(\boldsymbol{\theta}|\mathcal{D}) > \lambda \max_{\boldsymbol{\theta}'} q(\boldsymbol{\theta}'|\mathcal{D}) \text{ ,} \\ q(\boldsymbol{\theta}|\mathcal{D}) & \text{otherwise (i.e., shaded area in Fig. 2) .} \end{cases} \tag{7}$$

According to (7), unlearning is therefore focused at values of $\boldsymbol{\theta}$ with sufficiently large $q(\boldsymbol{\theta}|\mathcal{D})$ (i.e., $q(\boldsymbol{\theta}|\mathcal{D}) > \lambda \max_{\boldsymbol{\theta}'} q(\boldsymbol{\theta}'|\mathcal{D})$). Let the loss function $\text{KL}[\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda) \parallel \tilde{p}_{\text{adj}}(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)]$ be minimized w.r.t. the approximate posterior belief $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$ that is recovered by directly unlearning from erased data $\mathcal{D}_e$. Similar to (5), $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$ can be optimized by minimizing the following EUBO:

$$\widetilde{\mathcal{U}}_{\text{adj}}(\lambda) \triangleq \int \tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda) \log p_{\text{adj}}(\mathcal{D}_e|\boldsymbol{\theta};\lambda) \, \text{d}\boldsymbol{\theta} + \text{KL}[\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda) \parallel q(\boldsymbol{\theta}|\mathcal{D})] \tag{8}$$

which follows from replacing $p(\mathcal{D}_e|\boldsymbol{\theta})$ in (5) with $p_{\text{adj}}(\mathcal{D}_e|\boldsymbol{\theta};\lambda)$. Note that $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$ can be represented by a simple distribution (e.g., Gaussian) or a complex one (e.g., generative neural network, IAF). We initialize $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$ at $q(\boldsymbol{\theta}|\mathcal{D})$ for achieving empirically faster convergence. When $\lambda = 0$, $\widetilde{\mathcal{U}}_{\text{adj}}(\lambda = 0)$ reduces to $\widetilde{\mathcal{U}}$ (5), i.e., EUBO is minimized without the adjusted likelihood. As a result, $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda = 0) = \tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r)$. As λ increases, unlearning is focused on a smaller and smaller region of $\boldsymbol{\theta}$ with sufficiently large $q(\boldsymbol{\theta}|\mathcal{D})$. When λ reaches 1, no unlearning is performed since $\tilde{p}_{\text{adj}}(\boldsymbol{\theta}|\mathcal{D}_r;\lambda = 1) = q(\boldsymbol{\theta}|\mathcal{D})$, which results in $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda = 1) = q(\boldsymbol{\theta}|\mathcal{D})$ minimizing the loss function $\text{KL}[\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda = 1) \parallel \tilde{p}_{\text{adj}}(\boldsymbol{\theta}|\mathcal{D}_r;\lambda = 1)]$.

Example 1. To visualize the effect of varying λ on $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$, we consider learning the shape α of a Gamma distribution with a known rate (i.e., $\boldsymbol{\theta} = \alpha$): $\mathcal{D}$ are 20 samples of the Gamma distribution, $\mathcal{D}_e$ are the smallest 5 samples in $\mathcal{D}$, and the (non-conjugate) prior belief and approximate posterior beliefs of α are all Gaussians. Fig. 3a shows the approximate posterior beliefs $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$ with varying λ as well as $q(\boldsymbol{\theta}|\mathcal{D})$ and $q(\boldsymbol{\theta}|\mathcal{D}_r)$ learned using VI. As explained above, $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r, \lambda = 1) = q(\boldsymbol{\theta}|\mathcal{D})$. When $\lambda = 0.001$, $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r, \lambda = 0.001)$ is close to $q(\boldsymbol{\theta}|\mathcal{D}_r)$. However, as λ decreases to 0, $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r, \lambda)$ moves away from $q(\boldsymbol{\theta}|\mathcal{D}_r)$.

The optimized $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$ suffers from the same issue of underestimating the variance as $q(\boldsymbol{\theta}|\mathcal{D})$ learned using VI (see Remark 1), especially when λ tends to 0 (e.g., see $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda = 10^{-20})$ in Fig. 3a). Though this issue can be mitigated by tuning λ in the adjusted likelihood (6), we may not want to risk facing the consequence of picking a value of λ that is too small. So, in Sec. 3.3.2, we will propose another novel trick that is not inconvenienced by this issue.

3.3.2 Reverse KL

Let the loss function be the reverse KL divergence $\text{KL}[\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r) \parallel \tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r)]$ that is minimized w.r.t. the approximate posterior belief $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r)$ recovered by directly unlearning from erased data $\mathcal{D}_e$. In contrast to the optimized $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$ from minimizing EUBO (8), the optimized $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r)$ from minimizing the reverse KL divergence overestimates (instead of underestimates) the variance of $\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r)$ [1]: If $\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r)$ is close to 0, then $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r)$ is not necessarily close to 0. From (4), the reverse KL divergence can be expressed as follows:

$$\text{KL}[\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r) \parallel \tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r)] = C_0 - C_1 \, \mathbb{E}_{q(\boldsymbol{\theta}|\mathcal{D})} \left[(\log \tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r))/p(\mathcal{D}_e|\boldsymbol{\theta})\right] \tag{9}$$

where C_0 and C_1 are constants independent of $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r)$. So, the reverse KL divergence (9) can be minimized by maximizing $\mathbb{E}_{q(\boldsymbol{\theta}|\mathcal{D})}[(\log \tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r))/p(\mathcal{D}_e|\boldsymbol{\theta})]$ with *stochastic gradient ascent* (SGA). We also initialize $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r)$ at $q(\boldsymbol{\theta}|\mathcal{D})$ for achieving empirically faster convergence. Since stochastic optimization is performed with samples of $\boldsymbol{\theta} \sim q(\boldsymbol{\theta}|\mathcal{D})$ in each iteration of SGA, it is unlikely that the reverse KL divergence (9) is minimized using samples of $\boldsymbol{\theta}$ with small $q(\boldsymbol{\theta}|\mathcal{D})$. This naturally curbs unlearning at values of $\boldsymbol{\theta}$ with small $q(\boldsymbol{\theta}|\mathcal{D})$, as motivated by Remark 1. On the other hand, it is still possible to employ the same trick of adjusted likelihood (Sec. 3.3.1) by minimizing the reverse KL divergence $\mathrm{KL}[\tilde{p}_{\text{adj}}(\boldsymbol{\theta}|\mathcal{D}_r;\lambda) \parallel \tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)]$ as the loss function or, equivalently, maximizing $\mathbb{E}_{q(\boldsymbol{\theta}|\mathcal{D})}[(\log \tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r;\lambda))/p_{\text{adj}}(\mathcal{D}_e|\boldsymbol{\theta};\lambda)]$ where $p_{\text{adj}}(\mathcal{D}_e|\boldsymbol{\theta};\lambda)$ and $\tilde{p}_{\text{adj}}(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$ are previously defined in (6) and (7), respectively. The role of λ here is the same as that in (8).

To illustrate the difference between minimizing the reverse KL divergence (9) and EUBO (8), Fig. 3b shows the approximate posterior beliefs $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$ with varying λ. It can be observed that $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r;\lambda=1) = q(\boldsymbol{\theta}|\mathcal{D})$ (i.e., no unlearning). In contrast to minimizing EUBO (Fig. 3a), as λ decreases to 0, $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$ does not deviate that much from $q(\boldsymbol{\theta}|\mathcal{D}_r)$, even when $\lambda = 0$ (i.e., the reverse KL divergence is minimized without the adjusted likelihood). This is because the optimized $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$ is naturally more protected from both sources of inaccuracy (Remark 1), as explained above. Hence, we do not have to be as concerned about picking a small value of λ, which is also consistently observed in our experiments (Sec. 4).

4 Experiments and Discussion

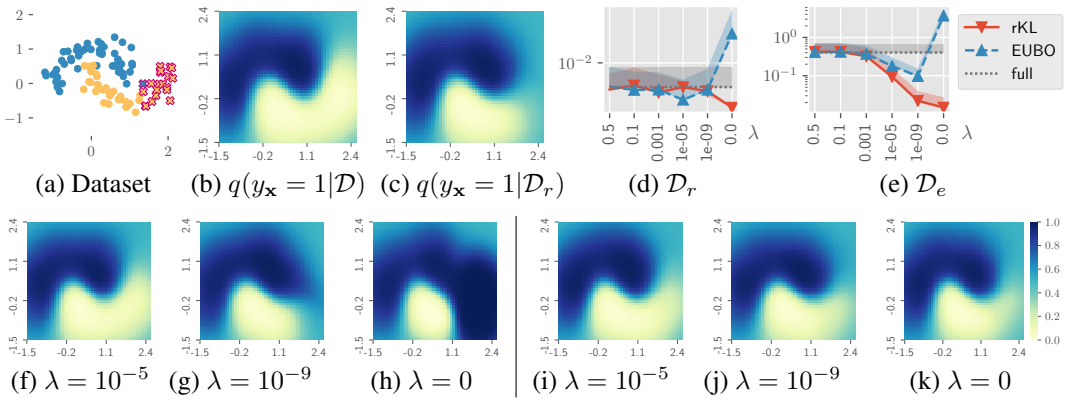
This section empirically demonstrates our unlearning methods on Bayesian models such as sparse Gaussian process and logistic regression using synthetic and real-world datasets. Further experimental results on Bayesian linear regression and with a bimodal posterior belief are reported in Appendices C and D, respectively. In our experiments, each dataset comprises pairs of input $\mathbf{x}$ and its corresponding output/observation $y_{\mathbf{x}}$. We use RMSProp as the SGA algorithm with a learning rate of 10^{-4}. To assess the performance of our unlearning methods (i.e., by directly unlearning from erased data $\mathcal{D}_e$), we consider the difference between their induced predictive distributions vs. that obtained using VI from retraining with remaining data $\mathcal{D}_r$, as motivated from Sec. 3.2. To do this, we use a **performance metric** that measures the KL divergence between the approximate predictive distributions

$$\tilde{q}_u(y_{\mathbf{x}}|\mathcal{D}_r) \triangleq \int p(y_{\mathbf{x}}|\boldsymbol{\theta})\ \tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)\ \mathrm{d}\boldsymbol{\theta} \quad \text{or} \quad \tilde{q}_v(y_{\mathbf{x}}|\mathcal{D}_r) \triangleq \int p(y_{\mathbf{x}}|\boldsymbol{\theta})\ \tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)\ \mathrm{d}\boldsymbol{\theta}$$

vs. $q(y_{\mathbf{x}}|\mathcal{D}_r) \triangleq \int p(y_{\mathbf{x}}|\boldsymbol{\theta})\ q(\boldsymbol{\theta}|\mathcal{D}_r)\ \mathrm{d}\boldsymbol{\theta}$ where $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$ and $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$ are optimized by, respectively, minimizing EUBO (8) and *reverse KL* (rKL) divergence (9) while requiring only $q(\boldsymbol{\theta}|\mathcal{D})$ and erased data $\mathcal{D}_e$ (Sec. 3.3), and $q(\boldsymbol{\theta}|\mathcal{D}_r)$ is learned using VI (Sec. 2). The above predictive distributions are computed via sampling with 100 samples of $\boldsymbol{\theta}$. For tractability reasons, we evaluate the above performance metric over finite input domains, specifically, over that in $\mathcal{D}_e$ and $\mathcal{D}_r$, which allows us to assess the performance of our unlearning methods on both the erased and remaining data, i.e., whether they can fully unlearn from $\mathcal{D}_e$ yet not forget nor catastrophically unlearn from $\mathcal{D}_r$, respectively. For example, the performance of our EUBO-based unlearning method over $\mathcal{D}_e$ is shown as an average (with standard deviation) of the KL divergences between $\tilde{q}_u(y_{\mathbf{x}}|\mathcal{D}_r)$ vs. $q(y_{\mathbf{x}}|\mathcal{D}_r)$ over all $(\mathbf{x}, y_{\mathbf{x}}) \in \mathcal{D}_e$. We also plot an average (with standard deviation) of the KL divergences between $q(y_{\mathbf{x}}|\mathcal{D})$ vs. $q(y_{\mathbf{x}}|\mathcal{D}_r)$ over $\mathcal{D}_r$ and $\mathcal{D}_e$ as *baselines* (i.e., representing no unlearning), which is expected to be larger than that of our unlearning methods (i.e., if performing well) and labeled as *full* in the plots below.

4.1 Sparse Gaussian Process (GP) Classification with Synthetic Moon Dataset

This experiment is about unlearning a binary classifier that is previously trained with the synthetic moon dataset (Fig. 4a). The probability of input $\mathbf{x} \in \mathbb{R}^2$ being in the 'blue' class (i.e., $y_{\mathbf{x}} = 1$ and denoted by blue dots in Fig. 4a) is defined as $1/(1+\exp(f_{\mathbf{x}}))$ where $f_{\mathbf{x}}$ is a latent function modeled by a sparse GP [27], which is elaborated in Appendix E. The parameters $\boldsymbol{\theta}$ of the sparse GP consist of 20 inducing variables; the approximate posterior beliefs of $\boldsymbol{\theta}$ are thus multivariate Gaussians (with full covariance matrices), as shown in Appendix E. By comparing Figs. 4b and 4c, it can be observed that after erasing $\mathcal{D}_e$ (i.e., mainly in 'yellow' class), $q(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ increases at $\mathbf{x} \in \mathcal{D}_e$. Figs. 4d and 4e show results of the performance of our EUBO- and rKL-based unlearning methods



(a) Dataset (b) $q(y_{\mathbf{x}} = 1|\mathcal{D})$ (c) $q(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ (d) $\mathcal{D}_r$ (e) $\mathcal{D}_e$

(f) $\lambda = 10^{-5}$ (g) $\lambda = 10^{-9}$ (h) $\lambda = 0$ (i) $\lambda = 10^{-5}$ (j) $\lambda = 10^{-9}$ (k) $\lambda = 0$

Figure 4: Plots of (a) synthetic moon dataset with erased data $\mathcal{D}_e$ (crosses) and remaining data $\mathcal{D}_r$ (dots), and of predictive distributions obtained using VI from (b) training with full data $\mathcal{D}$ and (c) retraining with $\mathcal{D}_r$. Graphs of averaged KL divergence vs. λ achieved by EUBO, rKL, and $q(\boldsymbol{\theta}|\mathcal{D})$ (i.e., baseline labeled as *full*) over (d) $\mathcal{D}_r$ and (e) $\mathcal{D}_e$. Plots of predictive distributions (f-h) $\tilde{q}_u(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ and (i-k) $\tilde{q}_v(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ induced, respectively, by EUBO and rKL for varying λ.

over $\mathcal{D}_r$ and $\mathcal{D}_e$ with varying λ, respectively.[4] When $\lambda = 10^{-9}$, EUBO performs reasonably well (compare Figs. 4g vs. 4c) as its averaged KL divergence is smaller than that of $q(\boldsymbol{\theta}|\mathcal{D})$ (i.e., baseline labeled as *full*). When $\lambda = 0$, EUBO performs poorly (compare Figs. 4h vs. 4c) as its averaged KL divergence is much larger than that of $q(\boldsymbol{\theta}|\mathcal{D})$, as shown in Figs. 4d and 4e. This agrees with our discussion of the issue with picking too small a value of λ for EUBO at the end of Sec. 3.3.1. In particular, catastrophic unlearning is observed as the input region containing $\mathcal{D}_e$ (i.e., mainly in 'yellow' class) has a high probability in 'blue' class after unlearning in Fig. 4h. On the other hand, when $\lambda = 0$, rKL performs well (compare Figs. 4k vs. 4c) as its KL divergence is much smaller than that of $q(\boldsymbol{\theta}|\mathcal{D})$, as seen in Figs. 4d and 4e. This agrees with our discussion at the end of Sec. 3.3.2 that rKL can work well without needing the adjusted likelihood.

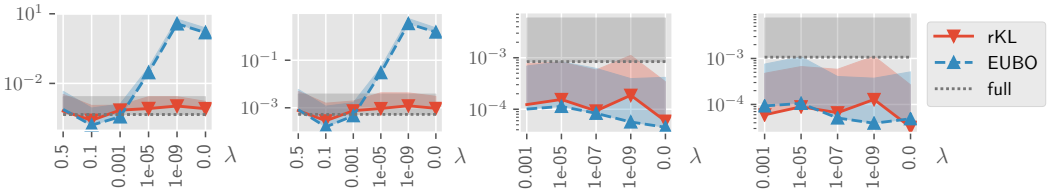
One may question how the performance of our unlearning methods would vary when erasing a large quantity of data or with different distributions of erased data (e.g., erasing the data randomly vs. deliberately erasing all data in a given class). To address this question, we have discovered that a key factor influencing their unlearning performance in these scenarios is the difference between the posterior beliefs of model parameters $\boldsymbol{\theta}$ given remaining data $\mathcal{D}_r$ vs. that given full data $\mathcal{D}$, especially at values of $\boldsymbol{\theta}$ with small $q(\boldsymbol{\theta}|\mathcal{D})$ since unlearning in such a region is curbed by the adjusted likelihood and reverse KL. In practice, we expect such a difference not to be large due to the small quantity of erased data and redundancy in real-world datasets. We will present the details of this study in Appendix F due to lack of space by considering how much $\mathcal{D}_e$ reduces the entropy of $\boldsymbol{\theta}$ given $\mathcal{D}_r$.

4.2 Logistic Regression with Banknote Authentication Dataset

The banknote authentication dataset [10] of size $|\mathcal{D}| = 1372$ is partitioned into erased data of size $|\mathcal{D}_e| = 412$ and remaining data of size $|\mathcal{D}_r| = 960$. Each input $\mathbf{x}$ comprises 4 features extracted from an image of a banknote and its corresponding binary label $y_{\mathbf{x}}$ indicates whether the banknote is genuine or forged. We use a logistic regression model with 5 parameters that is trained with this dataset. The prior beliefs of the model parameters are independent Gaussians $\mathcal{N}(0, 100)$.

Unlike the previous experiment, the erased data $\mathcal{D}_e$ here is randomly selected and hence does not reduce the entropy of model parameters $\boldsymbol{\theta}$ given $\mathcal{D}_r$ much, as explained in Appendix F; a discussion on erasing informative data (such as that in Sec. 4.1) is in Appendix F. As a result, Figs. 5a and 5b show a very small averaged KL divergence of about 10^{-3} between $q(y_{\mathbf{x}}|\mathcal{D})$ vs. $q(y_{\mathbf{x}}|\mathcal{D}_r)$ (i.e., baselines) over $\mathcal{D}_r$ and $\mathcal{D}_e$.[4] Figs. 5a and 5b also show that our unlearning methods do not perform well when using multivariate Gaussians to model the approximate posterior beliefs of $\boldsymbol{\theta}$: While rKL still gives a useful $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r; \lambda)$ achieving an averaged KL divergence close to that of $q(\boldsymbol{\theta}|\mathcal{D})$, EUBO gives a useless $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r; \lambda)$ incurring a large averaged KL divergence when λ is small. On the other hand,

[4]Note that the log plots can only properly display the upper confidence interval of 1 standard deviation (shaded area) and hence do not show the lower confidence interval.



(a) $\mathcal{D}_r$ (b) $\mathcal{D}_e$ (c) $\mathcal{D}_r$ (d) $\mathcal{D}_e$

Figure 5: Graphs of averaged KL divergence vs. λ achieved by EUBO, rKL, and $q(\boldsymbol{\theta}|\mathcal{D})$ (i.e., baseline labeled as *full*) over $\mathcal{D}_r$ and $\mathcal{D}_e$ for the banknote authentication dataset with the approximate posterior beliefs of model parameters represented by (a-b) multivariate Gaussians and (c-d) normalizing flows.

when more complex models like normalizing flows with the MADE architecture [26] are used to represent the approximate posterior beliefs, EUBO and rKL can unlearn well (Figs. 5c and 5d).

4.3 Logistic Regression with Fashion MNIST Dataset

The fashion MNIST dataset of size $|\mathcal{D}| = 60000$ (28×28 images of fashion items in 10 classes) is partitioned into erased data of size $|\mathcal{D}_e| = 10000$ and remaining data of size $|\mathcal{D}_r| = 50000$. The classification model is a neural network with 3 fully-connected hidden layers of 128, 128, 64 hidden neurons and a softmax layer to output the 10-class probabilities. The model can be interpreted as one of logistic regression on 64 features generated from the hidden layer of 64 neurons. Since modeling all weights of the neural network as random variables can be costly, we model only 650 weights in the transformation of the 64 features to the inputs of the softmax layer. The other weights remain constant during unlearning and retraining. The prior beliefs of the network weights are $\mathcal{N}(0, 10)$. The approximate posterior beliefs are modeled with independent Gaussians. Though a large part of the network is fixed and we use simple models to represent the approximate posterior beliefs, we show that unlearning is still fairly effective.

As discussed in Sec. 4.1, 4.2, and Appendix F, the random selection of erased data $\mathcal{D}_e$ and redundancy in $\mathcal{D}$ lead to a small averaged KL divergence of about 0.1 between $q(y_{\mathbf{x}}|\mathcal{D})$ vs. $q(y_{\mathbf{x}}|\mathcal{D}_r)$ (i.e., baselines) over $\mathcal{D}_r$ and $\mathcal{D}_e$ (Figs. 6a and 6b) despite choosing a relatively large $|\mathcal{D}_e|$. Figs. 6a and 6b show that when $\lambda \geq 10^{-9}$, EUBO and rKL achieve averaged KL divergences comparable to that of $q(\boldsymbol{\theta}|\mathcal{D})$ (i.e., baseline labeled as *full*), hence making their unlearning insignificant.[4] However, at $\lambda = 0$, the unlearning performance of rKL improves by achieving a smaller averaged KL divergence than that of $q(\boldsymbol{\theta}|\mathcal{D})$, while EUBO's performance deteriorates. Their performance can be further improved by using more complex models to represent their approximate posterior beliefs like that in Sec. 4.2, albeit high-dimensional. Figs. 6c and 6d show the class probabilities for two images evaluated at the mean of the approximate posterior beliefs with $\lambda = 0$. We observe that rKL induces the highest class probability for the same class as that of $q(\boldsymbol{\theta}|\mathcal{D}_r)$. The class probabilities for other images are shown in Appendix G. The two images are taken from a separate set of 10000 test images (i.e., different from $\mathcal{D}$) where rKL with $\lambda = 0$ yields the same predictions as $q(\boldsymbol{\theta}|\mathcal{D}_r)$ and $q(\boldsymbol{\theta}|\mathcal{D})$ in, respectively, 99.34% and 99.22% of the test images, the latter of which are contained in the former.

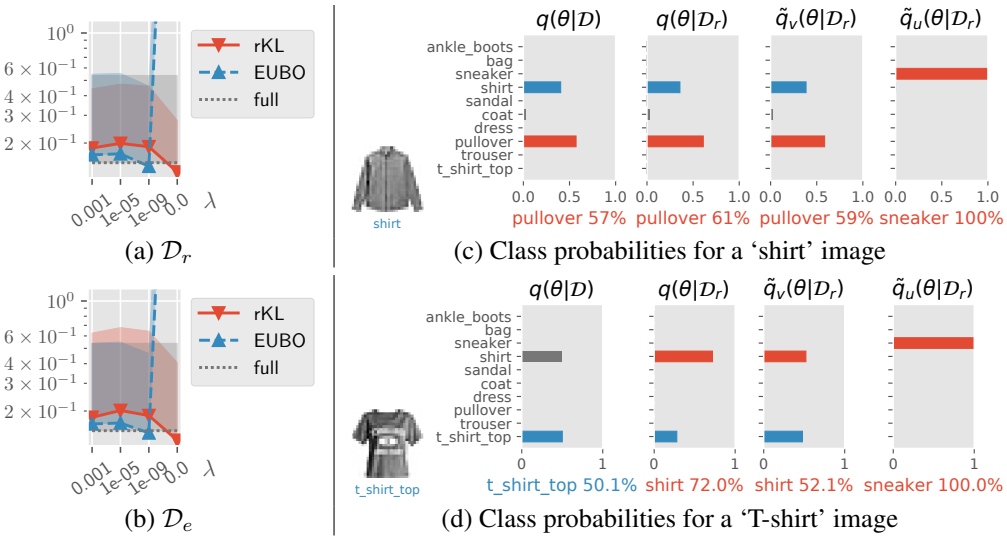
4.4 Sparse Gaussian Process (GP) Regression with Airline Dataset

This section illustrates the scalability of unlearning to the massive airline dataset of $\sim$ 2 million flights [15]. Training a sparse GP model with this massive dataset is made possible through stochastic VI [15]. Let $\mathcal{X}_{\mathbf{u}}$ denote the set of 50 inducing inputs in the sparse GP model and $\mathbf{f}_{\mathcal{X}_{\mathbf{u}}}$ be a vector of corresponding latent function values (i.e., inducing variables). The posterior belief $p(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{X}_{\mathbf{u}}}|\mathcal{D})$ is approximated as $q(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{X}_{\mathbf{u}}}|\mathcal{D}) \triangleq q(\mathbf{f}_{\mathcal{X}_{\mathbf{u}}}|\mathcal{D})\, p(\mathbf{f}_{\mathcal{D}}|\mathbf{f}_{\mathcal{X}_{\mathbf{u}}})$ where $\mathbf{f}_{\mathcal{D}} \triangleq (f_{\mathbf{x}})_{\mathbf{x}\in\mathcal{D}}$. Let the sets $\mathcal{X}_{\mathcal{D}}$ and $\mathcal{X}_{\mathcal{D}_e}$ denote inputs in the full and erased data, respectively. Then, the ELBO can be decomposed to

$$\mathcal{L} = \sum_{\mathbf{x}\in\mathcal{X}_{\mathcal{D}}} \int q(\mathbf{f}_{\mathcal{X}_{\mathbf{u}}}|\mathcal{D})\, p(f_{\mathbf{x}}|\mathbf{f}_{\mathcal{X}_{\mathbf{u}}}) \log p(y_{\mathbf{x}}|f_{\mathbf{x}})\, \mathrm{d}f_{\mathbf{x}}\, \mathrm{d}\mathbf{f}_{\mathcal{X}_{\mathbf{u}}} - \mathrm{KL}[q(\mathbf{f}_{\mathcal{X}_{\mathbf{u}}}|\mathcal{D}) \parallel p(\mathbf{f}_{\mathcal{X}_{\mathbf{u}}})] \tag{10}$$

where $\int p(f_{\mathbf{x}}|\mathbf{f}_{\mathcal{X}_{\mathbf{u}}}) \log p(y_{\mathbf{x}}|f_{\mathbf{x}})\, \mathrm{d}f_{\mathbf{x}}$ can be evaluated in closed form [11]. To unlearn such a trained model from $\mathcal{D}_e$ ($|\mathcal{D}_e| = 100$K here), the EUBO (8) can be expressed in a similar way as the ELBO:

$$\widetilde{\mathcal{U}}_{\text{adj}}(\lambda) = \sum_{\mathbf{x}\in\mathcal{X}_{\mathcal{D}_e}} \int \tilde{q}_u(\mathbf{f}_{\mathcal{X}_{\mathbf{u}}}|\mathcal{D}_r;\lambda) p(f_{\mathbf{x}}|\mathbf{f}_{\mathcal{X}_{\mathbf{u}}}) \log p_{\text{adj}}(y_{\mathbf{x}}|f_{\mathbf{x}};\lambda)\, \mathrm{d}f_{\mathbf{x}}\, \mathrm{d}\mathbf{f}_{\mathcal{X}_{\mathbf{u}}} + \mathrm{KL}[\tilde{q}_u(\mathbf{f}_{\mathcal{X}_{\mathbf{u}}}|\mathcal{D}_r;\lambda) \| q(\mathbf{f}_{\mathcal{X}_{\mathbf{u}}}|\mathcal{D})]$$



(a) $\mathcal{D}_r$ (c) Class probabilities for a 'shirt' image

(b) $\mathcal{D}_e$ (d) Class probabilities for a 'T-shirt' image

Figure 6: Graphs of averaged KL divergence vs. λ achieved by EUBO, rKL, and $q(\boldsymbol{\theta}|\mathcal{D})$ (i.e., baseline labeled as *full*) over (a) $\mathcal{D}_r$ and (b) $\mathcal{D}_e$. (c-d) Plots of class probabilities for two images in the fashion MNIST dataset obtained using $q(\boldsymbol{\theta}|\mathcal{D})$, $q(\boldsymbol{\theta}|\mathcal{D}_r)$, optimized $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r; \lambda = 0)$ and $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r; \lambda = 0)$.

Table 1: KL divergence achieved by EUBO (top row) and rKL (bottom row) with varying λ for airline dataset.

λ	10^{-11}	10^{-13}	10^{-20}	0
$\mathrm{KL}[\tilde{q}_u(\mathbf{f}_{\mathcal{X}_\mathbf{u}}\|\mathcal{D}_r;\lambda) \parallel q(\mathbf{f}_{\mathcal{X}_\mathbf{u}}\|\mathcal{D}_r)]$	2194.49	1943.00	1384.96	2629.71
$\mathrm{KL}[\tilde{q}_v(\mathbf{f}_{\mathcal{X}_\mathbf{u}}\|\mathcal{D}_r;\lambda) \parallel q(\mathbf{f}_{\mathcal{X}_\mathbf{u}}\|\mathcal{D}_r)]$	418.42	367.12	543.45	455.11

where $p_{\text{adj}}(y_\mathbf{x}|f_\mathbf{x};\lambda) = p(y_\mathbf{x}|f_\mathbf{x})$ if $q(f_\mathbf{x},\mathbf{f}_{\mathcal{X}_\mathbf{u}}|\mathcal{D}) > \lambda \max_{\mathbf{f}_{\mathcal{X}_\mathbf{u}}} q(f_\mathbf{x},\mathbf{f}_{\mathcal{X}_\mathbf{u}}|\mathcal{D})$, and $p_{\text{adj}}(y_\mathbf{x}|f_\mathbf{x};\lambda) = 1$ otherwise. EUBO can be minimized using stochastic gradient descent with random subsets (i.e., mini-batches of size 10K) of $\mathcal{D}_e$ in each iteration. For rKL, we use the entire $\mathcal{D}_e$ in each iteration. Since $\tilde{q}_u(\mathbf{f}_{\mathcal{X}_\mathbf{u}}|\mathcal{D}_r;\lambda)$, $\tilde{q}_v(\mathbf{f}_{\mathcal{X}_\mathbf{u}}|\mathcal{D}_r;\lambda)$, and $q(\mathbf{f}_{\mathcal{X}_\mathbf{u}}|\mathcal{D}_r)$ in (10) [11] are all multivariate Gaussians, we can directly evaluate the performance of EUBO and rKL with varying λ through their respective $\mathrm{KL}[\tilde{q}_u(\mathbf{f}_{\mathcal{X}_\mathbf{u}}|\mathcal{D}_r;\lambda) \parallel q(\mathbf{f}_{\mathcal{X}_\mathbf{u}}|\mathcal{D}_r)]$ and $\mathrm{KL}[\tilde{q}_v(\mathbf{f}_{\mathcal{X}_\mathbf{u}}|\mathcal{D}_r;\lambda) \parallel q(\mathbf{f}_{\mathcal{X}_\mathbf{u}}|\mathcal{D}_r)]$ which, according to Table 1, are smaller than $\mathrm{KL}[q(\mathbf{f}_{\mathcal{X}_\mathbf{u}}|\mathcal{D}) \parallel q(\mathbf{f}_{\mathcal{X}_\mathbf{u}}|\mathcal{D}_r)]$ of value 4344.09 (i.e., baseline representing no unlearning), hence demonstrating reasonable unlearning performance.

5 Conclusion

This paper describes novel unlearning methods for approximately unlearning a Bayesian model from a small subset of training data to be erased. Our unlearning methods are parsimonious in requiring only the approximate posterior belief of model parameters given the full data (i.e., obtained in model training with VI) and erased data to be available. This makes unlearning even more challenging due to two sources of inaccuracy in the approximate posterior belief. We introduce novel tricks of adjusted likelihood and reverse KL to curb unlearning in the region of model parameters with low approximate posterior belief where both sources of inaccuracy primarily occur. Empirical evaluations on synthetic and real-world datasets show that our proposed methods (especially reverse KL without adjusted likelihood) can effectively unlearn Bayesian models such as sparse GP and logistic regression from erased data. In practice, for the approximate posterior beliefs recovered by unlearning from erased data using our proposed methods, they can be immediately used in ML applications and continue to be improved at the same time by retraining with the remaining data at the expense of parsimony. In our future work, we will apply our our proposed methods to unlearning more sophisticated Bayesian models like the entire family of sparse GP models [5, 6, 7, 8, 16, 17, 18, 19, 20, 22, 23, 25, 31, 32, 34]) and deep GP models [33].

Broader Impact

As discussed in our introduction (Sec. 1), a direct contribution of our work to the society in this information age is to the implementation of *personal data ownership* (i.e., enforced by the General Data Protection Regulation in the European Union [24]) by studying the problem of machine unlearning for Bayesian models. Such an implementation can boost the confidence of users about sharing their data with an application/organization when they know that the trace of their data can be reduced/erased, as requested. As a result, organizations/applications can gather more useful data from users to enhance their service back to the users and hence to the society.

Our unlearning work can also contribute to the defense against data poisoning attacks (i.e., injecting malicious training data). Instead of retraining the tampered machine learning model from scratch to recover the quality of a service, unlearning the model from the detected malicious data may incur much less time, which improves the user experience and reduces the cost due to the service disruption.

In contrast, the ability to unlearn machine learning models may also open the door to new adversarial activities. For example, in the context of data sharing, multiple parties share their data to train a common machine learning model. An unethical party can deliberately share a low-quality dataset instead of its high-quality one. After obtaining the model trained on datasets from all parties (including the low-quality dataset), the unethical party can unlearn the low-quality dataset and continue to train the model with its high-quality dataset. By doing this, the unethical party achieves a better model than other parties in the collaboration. Therefore, the possibility of machine unlearning should be considered in the design of different data sharing frameworks.

Acknowledgments and Disclosure of Funding

This research/project is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *J. American Statistical Association*, 112(518):859–877, 2017.

[3] L. Bourtoule, V. Chandrasekaran, C. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot. Machine unlearning. arXiv:1912.03817, 2019.

[4] Y. Cao and J. Yang. Towards making systems forget with machine unlearning. In *Proc. IEEE S&P*, pages 463–480, 2015.

[5] J. Chen, N. Cao, B. K. H. Low, R. Ouyang, C. K.-Y. Tan, and P. Jaillet. Parallel Gaussian process regression with low-rank covariance matrix approximations. In *Proc. UAI*, pages 152–161, 2013.

[6] J. Chen, B. K. H. Low, P. Jaillet, and Y. Yao. Gaussian process decentralized data fusion and active sensing for spatiotemporal traffic modeling and prediction in mobility-on-demand systems. *IEEE Trans. Autom. Sci. Eng.*, 12:901–921, 2015.

[7] J. Chen, B. K. H. Low, and C. K.-Y. Tan. Gaussian process-based decentralized data fusion and active sensing for mobility-on-demand system. In *Proc. RSS*, 2013.

[8] J. Chen, B. K. H. Low, C. K.-Y. Tan, A. Oran, P. Jaillet, J. M. Dolan, and G. S. Sukhatme. Decentralized data fusion and active sensing with mobile sensors for modeling and predicting spatiotemporal traffic phenomena. In *Proc. UAI*, pages 163–173, 2012.

[9] M. Du, Z. Chen, C. Liu, R. Oak, and D. Song. Lifelong anomaly detection through unlearning. In *Proc. CCS*, pages 1283–1297, 2019.

[10] D. Dua and C. Graff. UCI machine learning repository, 2017.

[11] Y. Gal and M. van der Wilk. Variational inference in sparse Gaussian process regression and latent variable models–a gentle tutorial. *arXiv preprint arXiv*, 1402, 2014.

[12] A. Ginart, M. Guan, G. Valiant, and J. Y. Zou. Making AI forget you: Data deletion in machine learning. In *Proc. NeurIPS*, pages 3513–3526, 2019.

[13] A. Golatkar, A. Achille, and S. Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep neural networks. In *Proc. CVPR*, 2020.

[14] C. Guo, T. Goldstein, A. Hannun, and L. van der Maaten. Certified data removal from machine learning models. arXiv:1911.03030, 2019.

[15] J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Proc. UAI*, pages 282–290, 2013.

[16] Q. M. Hoang, T. N. Hoang, and B. K. H. Low. A generalized stochastic variational Bayesian hyperparameter learning framework for sparse spectrum Gaussian process regression. In *Proc. AAAI*, pages 2007–2014, 2017.

[17] Q. M. Hoang, T. N. Hoang, B. K. H. Low, and C. Kingsford. Collective model fusion for multiple black-box experts. In *Proc. ICML*, pages 2742–2750, 2019.

[18] T. N. Hoang, Q. M. Hoang, and B. K. H. Low. A unifying framework of anytime sparse Gaussian process regression models with stochastic variational inference for big data. In *Proc. ICML*, pages 569–578, 2015.

[19] T. N. Hoang, Q. M. Hoang, and B. K. H. Low. A distributed variational inference framework for unifying parallel sparse Gaussian process regression models. In *Proc. ICML*, pages 382–391, 2016.

[20] T. N. Hoang, Q. M. Hoang, B. K. H. Low, and J. P. How. Collective online learning of Gaussian processes in massive multi-agent systems. In *Proc. AAAI*, pages 7850–7857, 2019.

[21] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In *Proc. NeurIPS*, pages 4743–4751, 2016.

[22] B. K. H. Low, N. Xu, J. Chen, K. K. Lim, and E. B. Özgül. Generalized online sparse Gaussian processes with application to persistent mobile robot localization. In *Proc. ECML/PKDD Nectar Track*, pages 499–503, 2014.

[23] B. K. H. Low, J. Yu, J. Chen, and P. Jaillet. Parallel Gaussian process regression for big data: Low-rank representation meets Markov approximation. In *Proc. AAAI*, pages 2821–2827, 2015.

[24] A. Mantelero. The EU proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review*, 29(3):229–235, 2013.

[25] R. Ouyang and B. K. H. Low. Gaussian process decentralized data fusion meets transfer learning in large-scale distributed cooperative perception. In *Proc. AAAI*, pages 3876–3883, 2018.

[26] G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. In *Proc. NeurIPS*, pages 2338–2347, 2017.

[27] J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *JMLR*, 6:1939–1959, 2005.

[28] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[29] D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. In *Proc. ICML*, pages 1530–1538, 2015.

[30] S. Schelter. "Amnesia" – Towards machine learning models that can forget user data very fast. In *Proc. International Workshop on Applied AI for Database Systems and Applications*, 2019.

[31] T. Teng, J. Chen, Y. Zhang, and B. K. H. Low. Scalable variational Bayesian kernel selection for sparse Gaussian process regression. In *Proc. AAAI*, pages 5997–6004, 2020.

[32] N. Xu, B. K. H. Low, J. Chen, K. K. Lim, and E. B. Özgül. GP-Localize: Persistent mobile robot localization using online sparse Gaussian process observation model. In *Proc. AAAI*, pages 2585–2592, 2014.

[33] H. Yu, Y. Chen, Z. Dai, K. H. Low, and P. Jaillet. Implicit posterior variational inference for deep Gaussian processes. In *Proc. NeurIPS*, pages 14475–14486, 2019.

[34] H. Yu, T. N. Hoang, B. K. H. Low, and P. Jaillet. Stochastic variational inference for Bayesian sparse Gaussian process regression. In *Proc. IJCNN*, 2019.

A Proof of Proposition 1

We first follow the proof of the log-sum inequality to prove the following inequality:

$$q_u(y|\mathcal{D}_r)\ \log \frac{q_u(y|\mathcal{D}_r)}{p(y|\mathcal{D}_r)} \leq \int q_u(\boldsymbol{\theta}|\mathcal{D}_r)\ p(y|\boldsymbol{\theta})\ \log \frac{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}{p(\boldsymbol{\theta}|\mathcal{D}_r)}\ \mathrm{d}\boldsymbol{\theta} \tag{11}$$

where $q_u(y|\mathcal{D}_r) \triangleq \mathbb{E}_{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})] = \int q_u(\boldsymbol{\theta}|\mathcal{D}_r)\ p(y|\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}$ and $p(y|\mathcal{D}_r) \triangleq \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})] = \int p(\boldsymbol{\theta}|\mathcal{D}_r)\ p(y|\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}$.

Proof. Define the function $f(t) \triangleq t \log t$ which is convex. Then,

$$\begin{aligned}
&\int q_u(\boldsymbol{\theta}|\mathcal{D}_r)\ p(y|\boldsymbol{\theta})\ \log \frac{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}{p(\boldsymbol{\theta}|\mathcal{D}_r)}\ \mathrm{d}\boldsymbol{\theta} \\
&= \int p(\boldsymbol{\theta}|\mathcal{D}_r)\ p(y|\boldsymbol{\theta})\ f\left(\frac{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}{p(\boldsymbol{\theta}|\mathcal{D}_r)}\right)\ \mathrm{d}\boldsymbol{\theta} \\
&= \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})] \int \frac{p(\boldsymbol{\theta}|\mathcal{D}_r)\ p(y|\boldsymbol{\theta})}{\mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})]}\ f\left(\frac{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}{p(\boldsymbol{\theta}|\mathcal{D}_r)}\right)\ \mathrm{d}\boldsymbol{\theta} \\
&\geq \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})]\ f\left(\int \frac{p(\boldsymbol{\theta}|\mathcal{D}_r)\ p(y|\boldsymbol{\theta})}{\mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})]} \frac{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}{p(\boldsymbol{\theta}|\mathcal{D}_r)}\ \mathrm{d}\boldsymbol{\theta}\right) \\
&= \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})]\ f\left(\int \frac{p(y|\boldsymbol{\theta})\ q_u(\boldsymbol{\theta}|\mathcal{D}_r)}{\mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})]}\ \mathrm{d}\boldsymbol{\theta}\right) \\
&= \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})]\ f\left(\frac{\mathbb{E}_{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})]}{\mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})]}\right) \\
&= \mathbb{E}_{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})]\ \log \frac{\mathbb{E}_{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})]}{\mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_r)}[p(y|\boldsymbol{\theta})]} \\
&= q_u(y|\mathcal{D}_r)\ \log \frac{q_u(y|\mathcal{D}_r)}{p(y|\mathcal{D}_r)}
\end{aligned}$$

where the inequality is due to Jensen's inequality. □

Then, integrating both sides of (11) w.r.t. y,

$$\begin{aligned}
\int q_u(y|\mathcal{D}_r)\ \log \frac{q_u(y|\mathcal{D}_r)}{p(y|\mathcal{D}_r)}\ \mathrm{d}y &\leq \int\int q_u(\boldsymbol{\theta}|\mathcal{D}_r)\ p(y|\boldsymbol{\theta})\ \log \frac{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}{p(\boldsymbol{\theta}|\mathcal{D}_r)}\ \mathrm{d}\boldsymbol{\theta}\ \mathrm{d}y \\
\int q_u(y|\mathcal{D}_r)\ \log \frac{q_u(y|\mathcal{D}_r)}{p(y|\mathcal{D}_r)}\ \mathrm{d}y &\leq \int q_u(\boldsymbol{\theta}|\mathcal{D}_r) \left(\int p(y|\boldsymbol{\theta})\ \mathrm{d}y\right) \log \frac{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}{p(\boldsymbol{\theta}|\mathcal{D}_r)}\ \mathrm{d}\boldsymbol{\theta} \\
\int q_u(y|\mathcal{D}_r)\ \log \frac{q_u(y|\mathcal{D}_r)}{p(y|\mathcal{D}_r)}\ \mathrm{d}y &\leq \int q_u(\boldsymbol{\theta}|\mathcal{D}_r) \log \frac{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}{p(\boldsymbol{\theta}|\mathcal{D}_r)}\ \mathrm{d}\boldsymbol{\theta} \\
\mathrm{KL}[q_u(y|\mathcal{D}_r) \parallel p(y|\mathcal{D}_r)] &\leq \mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D}_r)]\ .
\end{aligned}$$

B Proof of Proposition 2

From (2),

$$\begin{aligned}
\log p(\mathcal{D}_e|\mathcal{D}_r) &= \log \frac{p(\mathcal{D}_e|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathcal{D}_r)}{p(\boldsymbol{\theta}|\mathcal{D})} \\
&= \log \frac{q_u(\boldsymbol{\theta}|\mathcal{D}_r)\ p(\mathcal{D}_e|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathcal{D}_r)}{q_u(\boldsymbol{\theta}|\mathcal{D}_r)\ p(\boldsymbol{\theta}|\mathcal{D})} \\
&= \log p(\mathcal{D}_e|\boldsymbol{\theta}) + \log \frac{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}{p(\boldsymbol{\theta}|\mathcal{D})} - \log \frac{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}{p(\boldsymbol{\theta}|\mathcal{D}_r)}\ .
\end{aligned}$$

Then, taking an expectation of both sides w.r.t. $q_u(\boldsymbol{\theta}|\mathcal{D}_r)$,

$$\begin{aligned}\log p(\mathcal{D}_e|\mathcal{D}_r) &= \int q_u(\boldsymbol{\theta}|\mathcal{D}_r) \log p(\mathcal{D}_e|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} + \int q_u(\boldsymbol{\theta}|\mathcal{D}_r) \log \frac{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}{p(\boldsymbol{\theta}|\mathcal{D})}\, \mathrm{d}\boldsymbol{\theta} - \int q_u(\boldsymbol{\theta}|\mathcal{D}_r) \log \frac{q_u(\boldsymbol{\theta}|\mathcal{D}_r)}{p(\boldsymbol{\theta}|\mathcal{D}_r)}\, \mathrm{d}\boldsymbol{\theta} \\ &= \int q_u(\boldsymbol{\theta}|\mathcal{D}_r)\ \log p(\mathcal{D}_e|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} + \mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D})] - \mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D}_r)] \\ &= \mathcal{U} - \mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D}_r)]\ .\end{aligned}$$

Therefore,

$$\mathcal{U} = \log p(\mathcal{D}_e|\mathcal{D}_r) + \mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D}_r)] \geq \log p(\mathcal{D}_e|\mathcal{D}_r)$$

since $\mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D}_r)] \geq 0$. So, $\mathcal{U}$ is an upper bound of $\log p(\mathcal{D}_e|\mathcal{D}_r)$.

C Bayesian Linear Regression

We perform unlearning of a simple Bayesian linear regression model: $y_x = ax^3 + bx^2 + cx + d + \epsilon$ where $a = 2$, $b = -3$, $c = 1$, and $d = 0$ are the model parameters $\boldsymbol{\theta}$, and the noise is $\epsilon \sim \mathcal{N}(0, 0.05^2)$. Though the exact posterior belief of $\boldsymbol{\theta}$ is known to be a multivariate Gaussian, we choose to use a low-rank approximation (i.e., multivariate Gaussian with a diagonal covariance matrice) and represent the approximate posterior beliefs of the model parameters with independent Gaussians so that the approximation is not exact.

Fig. 7a shows the remaining data $\mathcal{D}_r$ and erased data $\mathcal{D}_e$. Note that the erased data $\mathcal{D}_e$ is informative to the approximate posterior beliefs of the model parameters $\boldsymbol{\theta}$ as $\mathcal{D}_e$ are clustered. So, the difference between the samples drawn from predictive distributions $q(y_x|\mathcal{D})$ (Fig. 7b) vs. $q(y_x|\mathcal{D}_r)$ (Fig. 7c) is large.

Table 2: KL divergences achieved by EUBO (left column) and rKL (right column) with varying λ for synthetic linear regression dataset.

λ	$\mathrm{KL}[\tilde{q}_u(\boldsymbol{\theta}\|\mathcal{D}_r;\lambda) \parallel q(\boldsymbol{\theta}\|\mathcal{D}_r)]$	$\mathrm{KL}[\tilde{q}_v(\boldsymbol{\theta}\|\mathcal{D}_r;\lambda) \parallel q(\boldsymbol{\theta}\|\mathcal{D}_r)]$
0.5	0.1143	0.1012
0.1	0.0899	0.0600
0.0	266.68	0.0158

From Table 2, the KL divergences achieved by EUBO and rKL with $\lambda = 0.1, 0.5$ are smaller than $\mathrm{KL}[q(\boldsymbol{\theta}|\mathcal{D}) \parallel q(\boldsymbol{\theta}|\mathcal{D}_r)]$ of value 0.1170 (i.e., baseline representing no unlearning), hence demonstrating reasonable unlearning performance. When $\lambda = 0$, EUBO suffers from catastrophic unlearning, but rKL does not. The KL divergences in Table 2 also agree with the plots of samples drawn from the predictive distributions induced by EUBO and rKL in Fig. 7 by comparing with the samples drawn from the predictive distribution obtained using VI from retraining with $\mathcal{D}_r$ in Fig. 7c.

D Bimodal Posterior Belief

Let the posterior belief of model parameter θ given full data $\mathcal{D}$ be a Gaussian mixture (i.e., a bimodal distribution):

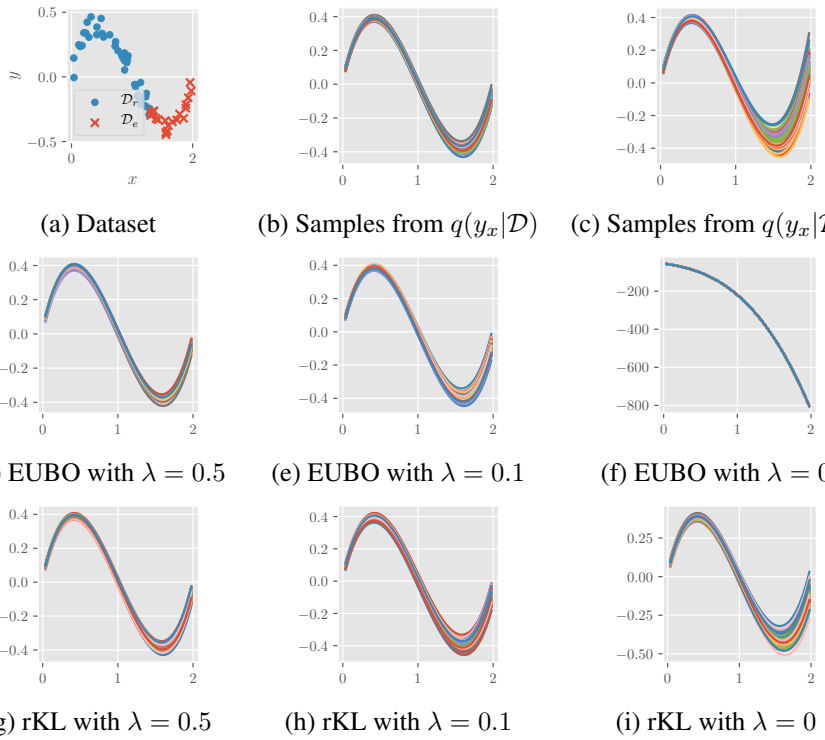
$$p(\theta|\mathcal{D}) \triangleq 0.5\ \phi(\theta; 0, 1) + 0.5\ \phi(\theta; 2, 1) \tag{12}$$

where $\phi(\theta; \mu, \sigma^2)$ is a Gaussian p.d.f. with mean μ and variance σ^2. We deliberately choose the likelihood of the erased data $\mathcal{D}_e$ to be

$$p(\mathcal{D}_e|\theta) \triangleq 1 + \frac{\phi(\theta; 2, 1)}{\phi(\theta; 0, 1)} \tag{13}$$

so that the posterior belief of θ given the remaining data $\mathcal{D}_r$ is a Gaussian:

$$p(\theta|\mathcal{D}_r) \propto \frac{p(\theta|\mathcal{D})}{p(\mathcal{D}_e|\theta)} = \phi(\theta; 0, 1) \tag{14}$$



(a) Dataset (b) Samples from $q(y_x|\mathcal{D})$ (c) Samples from $q(y_x|\mathcal{D}_r)$

(d) EUBO with $\lambda = 0.5$ (e) EUBO with $\lambda = 0.1$ (f) EUBO with $\lambda = 0$

(g) rKL with $\lambda = 0.5$ (h) rKL with $\lambda = 0.1$ (i) rKL with $\lambda = 0$

Figure 7: Plots of (a) synthetic linear regression dataset with erased data $\mathcal{D}_e$ (crosses) and remaining data $\mathcal{D}_r$ (dots), and samples from predictive distributions obtained using VI from (b) training with full data $\mathcal{D}$ and (c) retraining with $\mathcal{D}_r$. Plots of samples from predictive distributions (d-f) $\tilde{q}_u(y_x|\mathcal{D}_r)$ and (g-i) $\tilde{q}_v(y_x|\mathcal{D}_r)$ induced, respectively, by EUBO and rKL with varying λ.

where the proportionality is due to (2).

We assume to only have access to the likelihood of the erased data in (13); the exact posterior beliefs of θ given the full data (12) and that given the remaining data (14) are not available. Instead, we have access to an approximate posterior belief $q(\theta|\mathcal{D})$ given the full data obtained using VI by minimizing $\mathrm{KL}[q(\theta|\mathcal{D}) \,\|\, p(\theta|\mathcal{D})]$ or, equivalently, maximizing the ELBO (Section 2):

$$q(\theta|\mathcal{D}) = \phi(\theta; 1.004, 1.390^2) \,. \tag{15}$$

Given the likelihood $p(\mathcal{D}_e|\theta)$ of the erased data in (13) and the approximate posterior belief $q(\theta|\mathcal{D})$ given the full data (15), unlearning from $\mathcal{D}_e$ is performed using EUBO and rKL to obtain

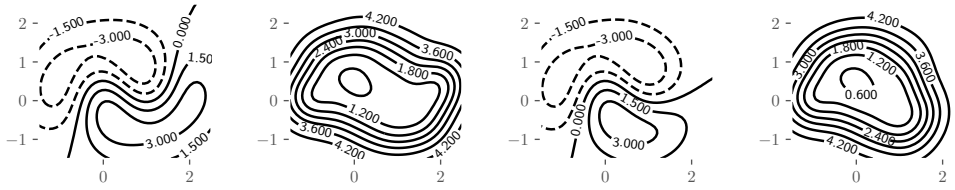
$$\tilde{q}_u(\theta|\mathcal{D}_r; \lambda = 0) = \phi(\theta; 0.060, 1.000^2) \quad \text{and} \quad \tilde{q}_v(\theta|\mathcal{D}_r; \lambda = 0) = \phi(\theta; 0.062, 1.018^2) \,,$$

respectively. Hence, both EUBO and rKL perform reasonably well since their respective $\tilde{q}_u(\theta|\mathcal{D}_r; \lambda = 0)$ and $\tilde{q}_v(\theta|\mathcal{D}_r; \lambda = 0)$ are close to $p(\theta|\mathcal{D}_r) = \phi(\theta; 0, 1)$ (14) when $p(\theta|\mathcal{D})$ is a bimodal distribution.

E Gaussian Process (GP) Classification with Synthetic Moon Dataset: Additional Details and Experimental Results

This section discusses the sparse GP model that is used in the classification of the synthetic moon dataset in Sec. 4.1. Let $y_{\mathbf{x}} \in \{0, 1\}$ be the class label of $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^2$; $y_{\mathbf{x}} = 1$ denotes the 'blue' class plotted as blue dots in Fig. 4a. The probability of $y_{\mathbf{x}}$ is defined as follows:

$$\begin{aligned} p(y_{\mathbf{x}} = 1|f_{\mathbf{x}}) &\triangleq \frac{1}{1 + \exp(f_{\mathbf{x}})} \\ p(y_{\mathbf{x}} = 0|f_{\mathbf{x}}) &\triangleq \frac{\exp(f_{\mathbf{x}})}{1 + \exp(f_{\mathbf{x}})} \end{aligned} \tag{16}$$



(a) $\mu_{\mathbf{x}|\mathcal{D}}$ (b) $\sigma^2_{\mathbf{x}|\mathcal{D}}$ (c) $\mu_{\mathbf{x}|\mathcal{D}_r}$ (d) $\sigma^2_{\mathbf{x}|\mathcal{D}_r}$

Figure 8: Plots of approximate posterior beliefs (a-b) $q(f_{\mathbf{x}}|\mathcal{D})$ and (c-d) $q(f_{\mathbf{x}}|\mathcal{D}_r)$.

where $f_{\mathbf{x}}$ is modeled using a GP [28], that is, every finite subset of $\{f_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{X}}$ follows a multivariate Gaussian distribution. A GP is fully specified by its *prior* mean (i.e., assumed to be 0 w.l.o.g.) and covariance $k_{\mathbf{x}\mathbf{x}'} \triangleq \mathrm{cov}(\mathbf{x},\mathbf{x}')$, the latter of which can be defined by the widely-used squared exponential covariance function $k_{\mathbf{x}\mathbf{x}'} \triangleq \sigma_f^2 \exp(-0.5\|\Lambda(\mathbf{x}-\mathbf{x}')\|_2^2)$ where $\Lambda = \mathrm{diag}[\lambda_1,\lambda_2]$ and σ_f^2 are the length-scale and signal variance hyperparameters, respectively. In this experiment, we set $\lambda_1 = 1.56$, $\lambda_2 = 1.35$, and $\sigma_f^2 = 4.74$.

We employ a sparse GP model, namely, the *deterministic training conditional* (DTC) [27] approximation of the GP model with a set $\mathcal{X}_u$ of 20 *inducing inputs*. These inducing inputs are randomly selected from $\mathcal{X}$ and remain the same (and fixed) for both model training and unlearning. Given the latent function values (i.e., also known as *inducing variables*) $\mathbf{f}_{\mathcal{X}_u} \triangleq (f_{\mathbf{x}})^\top_{\mathbf{x}\in\mathcal{X}_u}$ at these inducing inputs, the posterior belief of the latent function value $f_{\mathbf{x}}$ at a new input $\mathbf{x}$ is a Gaussian $p(f_{\mathbf{x}}|\mathbf{f}_{\mathcal{X}_u}) = \mathcal{N}(\mathbf{k}_{\mathbf{x}\mathcal{X}_u}\mathbf{K}^{-1}_{\mathcal{X}_u\mathcal{X}_u}\mathbf{f}_{\mathcal{X}_u}, k_{\mathbf{x}\mathbf{x}} - \mathbf{k}_{\mathbf{x}\mathcal{X}_u}\mathbf{K}^{-1}_{\mathcal{X}_u\mathcal{X}_u}\mathbf{k}_{\mathcal{X}_u\mathbf{x}})$ where $\mathbf{k}_{\mathbf{x}\mathcal{X}_u} \triangleq (k_{\mathbf{x}\mathbf{x}'})_{\mathbf{x}'\in\mathcal{X}_u}$, $\mathbf{k}_{\mathcal{X}_u\mathbf{x}} = \mathbf{k}^\top_{\mathbf{x}\mathcal{X}_u}$, and $\mathbf{K}_{\mathcal{X}_u\mathcal{X}_u} = (k_{\mathbf{x}\mathbf{x}'})_{\mathbf{x},\mathbf{x}'\in\mathcal{X}_u}$.

Using $p(f_{\mathbf{x}}|\mathbf{f}_{\mathcal{X}_u})$ and $q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D}) \triangleq \mathcal{N}(\boldsymbol{\mu}_{\mathcal{X}_u}, \boldsymbol{\Sigma}_{\mathcal{X}_u})$, it can be derived that the approximate posterior belief $q(f_{\mathbf{x}}|\mathcal{D})$ of $f_{\mathbf{x}}$ given full data $\mathcal{D}$ is also a Gaussian with the following respective *posterior* mean and variance:

$$\mu_{\mathbf{x}|\mathcal{D}} \triangleq \mathbf{k}_{\mathbf{x}\mathcal{X}_u}\mathbf{K}^{-1}_{\mathcal{X}_u\mathcal{X}_u}\boldsymbol{\mu}_{\mathcal{X}_u}\ , \tag{17}$$

$$\sigma^2_{\mathbf{x}|\mathcal{D}} \triangleq k_{\mathbf{x}\mathbf{x}} - \mathbf{k}_{\mathbf{x}\mathcal{X}_u}\mathbf{K}^{-1}_{\mathcal{X}_u\mathcal{X}_u}\mathbf{k}_{\mathcal{X}_u\mathbf{x}} + \mathbf{k}_{\mathbf{x}\mathcal{X}_u}\mathbf{K}^{-1}_{\mathcal{X}_u\mathcal{X}_u}\boldsymbol{\Sigma}_{\mathcal{X}_u}\mathbf{K}^{-1}_{\mathcal{X}_u\mathcal{X}_u}\mathbf{k}_{\mathcal{X}_u\mathbf{x}}\ . \tag{18}$$

The approximate posterior belief $q(f_{\mathbf{x}}|\mathcal{D}_r)$ of $f_{\mathbf{x}}$ from retraining with remaining data $\mathcal{D}_r$ using VI (specifically, using $q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D}_r)$) can be derived in the same way as that of $q(f_{\mathbf{x}}|\mathcal{D})$.

The parameters $\boldsymbol{\mu}_{\mathcal{X}_u}$, $\boldsymbol{\Sigma}_{\mathcal{X}_u}$ of the approximate posterior belief $q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D})$ is optimized by maximizing the ELBO with stochastic gradient ascent (let $\boldsymbol{\theta} = \mathbf{f}_{\mathcal{X}_u}$ in (1) in Sec. 2):

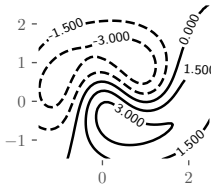
$$\mathbb{E}_{\mathbf{f}_{\mathcal{X}_u}\sim q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D})}\left[\log p(\mathcal{D}|\mathbf{f}_{\mathcal{X}_u}) - \log q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D}) + \log p(\mathbf{f}_{\mathcal{X}_u})\right]$$

where $p(\mathcal{D}|\mathbf{f}_{\mathcal{X}_u})$ is computed using (16), (17) and (18).

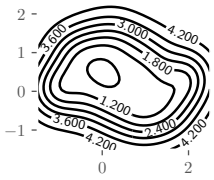
Fig. 8 visualizes $q(f_{\mathbf{x}}|\mathcal{D})$ (Figs. 8a and 8b) and $q(f_{\mathbf{x}}|\mathcal{D}_r)$ (Figs. 8c and 8d) whose corresponding predictive distributions $q(y_{\mathbf{x}} = 1|\mathcal{D})$ and $q(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ are shown in Figs. 4b and 4c, respectively. On the other hand, Figs. 9 and 10 visualize the approximate posterior beliefs $\tilde{q}_u(f_{\mathbf{x}}|\mathcal{D}_r;\lambda)$ and $\tilde{q}_v(f_{\mathbf{x}}|\mathcal{D}_r;\lambda)$ induced, respectively, by EUBO and rKL whose corresponding predictive distributions $\tilde{q}_u(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ and $\tilde{q}_v(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ are shown in Figs. 4f-k. Similar to the comparison between predictive distributions $\tilde{q}_u(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ vs. $q(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ in Sec. 4.1, it can be observed that the approximate posterior belief $\tilde{q}_u(f_{\mathbf{x}}|\mathcal{D}_r;\lambda = 10^{-9})$ induced by EUBO is similar to $q(f_{\mathbf{x}}|\mathcal{D}_r)$ obtained using VI from retraining with $\mathcal{D}_r$ (compare Figs. 9c vs. 8c and Figs. 9d vs. 8d). However, $\tilde{q}_u(f_{\mathbf{x}}|\mathcal{D}_r;\lambda = 0)$ induced by EUBO differs from $q(f_{\mathbf{x}}|\mathcal{D}_r)$ obtained using VI from retraining with $\mathcal{D}_r$ (compare Figs. 9e vs. 8c and Figs. 9f vs. 8d). On the other hand, both the approximate posterior beliefs $\tilde{q}_v(f_{\mathbf{x}}|\mathcal{D}_r;\lambda = 10^{-9})$ and $\tilde{q}_v(f_{\mathbf{x}}|\mathcal{D}_r;\lambda = 0)$ induced by rKL are similar to $q(f_{\mathbf{x}}|\mathcal{D}_r)$ obtained using VI from retraining with $\mathcal{D}_r$ (compare Fig. 10 vs. Figs. 8c-d).

F A Note on Erasing Informative Data

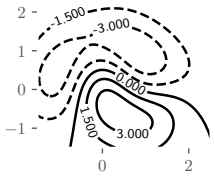
In this section, we study the performance of our unlearning methods when erasing a large quantity of data or with different distributions of erased data (i.e., erasing the data randomly vs. deliberately



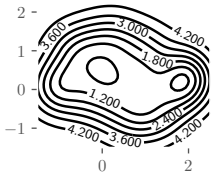
(a) Mean of $\tilde{q}_u(f_{\mathbf{x}}|\mathcal{D}_r;\lambda=10^{-5})$



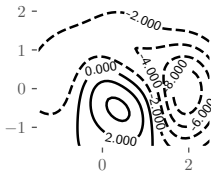
(b) Variance of $\tilde{q}_u(f_{\mathbf{x}}|\mathcal{D}_r;\lambda=10^{-5})$



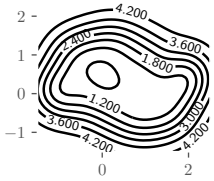
(c) Mean of $\tilde{q}_u(f_{\mathbf{x}}|\mathcal{D}_r;\lambda=10^{-9})$



(d) Variance of $\tilde{q}_u(f_{\mathbf{x}}|\mathcal{D}_r;\lambda=10^{-9})$

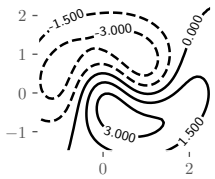


(e) Mean of $\tilde{q}_u(f_{\mathbf{x}}|\mathcal{D}_r;\lambda=0)$

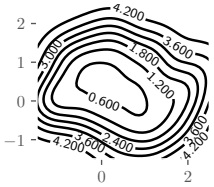


(f) Variance of $\tilde{q}_u(f_{\mathbf{x}}|\mathcal{D}_r;\lambda=0)$

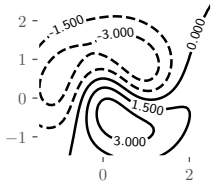
Figure 9: Plots of approximate posterior belief $\tilde{q}_u(f_{\mathbf{x}}|\mathcal{D}_r;\lambda)$ induced by EUBO for varying λ.



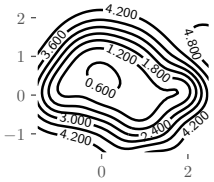
(a) Mean of $\tilde{q}_v(f_{\mathbf{x}}|\mathcal{D}_r;\lambda=10^{-5})$



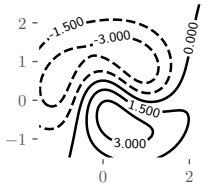
(b) Variance of $\tilde{q}_v(f_{\mathbf{x}}|\mathcal{D}_r;\lambda=10^{-5})$



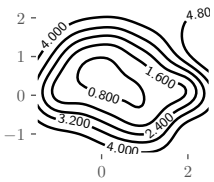
(c) Mean of $\tilde{q}_v(f_{\mathbf{x}}|\mathcal{D}_r;\lambda=10^{-9})$



(d) Variance of $\tilde{q}_v(f_{\mathbf{x}}|\mathcal{D}_r;\lambda=10^{-9})$



(e) Mean of $\tilde{q}_v(f_{\mathbf{x}}|\mathcal{D}_r;\lambda=0)$



(f) Variance of $\tilde{q}_v(f_{\mathbf{x}}|\mathcal{D}_r;\lambda=0)$

Figure 10: Plots of approximate posterior belief $\tilde{q}_v(f_{\mathbf{x}}|\mathcal{D}_r;\lambda)$ induced by rKL for varying λ.

erasing all data in a given class). Let us consider the experiment in Sec. 4.1 on the sparse GP model (i.e., the model parameters $\boldsymbol{\theta}$ in (1) in Sec. 2 are inducing variables $\mathbf{f}_{\mathcal{X}_u}$) in the classification of the synthetic moon dataset as it allows us to easily visualize both the approximate posterior beliefs of the latent function $f_{\mathbf{x}}$ and the predictive distributions of the output/observation $y_{\mathbf{x}}$. A key factor influencing the performance of our unlearning methods in the above-mentioned scenarios is the difference between the approximate posterior belief of model parameters $\mathbf{f}_{\mathcal{X}_u}$ given remaining data $\mathcal{D}_r$ vs. that given full data $\mathcal{D}$. We quantify such a difference by how much the erased data $\mathcal{D}_e$ reduces the entropy of model parameters/inducing variables $\mathbf{f}_{\mathcal{X}_u}$ given remaining data $\mathcal{D}_r$:

$$\mathcal{I} \triangleq H(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D}_r) - H(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D}) = -\int q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D}_r)\log q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D}_r)\,\mathrm{d}\mathbf{f}_{\mathcal{X}_u} + \int q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D})\log q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D})\,\mathrm{d}\mathbf{f}_{\mathcal{X}_u}\,. \tag{19}$$

Note that $\mathcal{I}$ (19) is not the same as the mutual information (i.e., information gain) between $\mathbf{f}_{\mathcal{X}_u}$ and $\mathbf{y}_{\mathcal{D}_e} \triangleq (y_{\mathbf{x}})^\top_{(\mathbf{x},y_{\mathbf{x}})\in\mathcal{D}_e}$ given $\mathcal{D}_r$, which is equal to $H(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D}_r) - \mathbb{E}_{p(\mathbf{y}_{\mathcal{D}_e}|\mathcal{D}_r)}\left[H(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D}_r,\mathbf{y}_{\mathcal{D}_e})\right]$ with an expensive-to-evaluate expectation term. Furthermore, the outputs/observations $\mathbf{y}_{\mathcal{D}_e}$ are known from $\mathcal{D}_e$. These therefore prompt us to choose $\mathcal{I}$ (19) as the measure of how much the erased data $\mathcal{D}_e$ reduces the entropy of model parameters/inducing variables $\mathbf{f}_{\mathcal{X}_u}$ given remaining data $\mathcal{D}_r$.

We investigate 4 different scenarios in the order of increasing $\mathcal{I}$:

1. Randomly selected $\mathcal{D}_e$ ($\mathcal{I} = 0.27$): The erased data of size $|\mathcal{D}_e| = 20$ are randomly selected from $\mathcal{D}$. Hence, they are not necessarily near the decision boundary, i.e., $\mathcal{D}_e$ does not reduce the entropy of model parameters/inducing variables $\mathbf{f}_{\mathcal{X}_u}$ given $\mathcal{D}_r$ much;
2. Partially 'yellow' $\mathcal{D}_e$ ($\mathcal{I} = 1.59$): The erased data of size $|\mathcal{D}_e| = 30$ are labeled with the 'yellow' class and comprise inputs $\mathbf{x}$ with the largest possible first component x_0. Such a choice ensures that the erased data group together to cover a part of the decision boundary, as shown in Fig. 11d;
3. Largely 'yellow' $\mathcal{D}_e$ ($\mathcal{I} = 2.06$): The erased data of size $|\mathcal{D}_e| = 40$ are labeled with the yellow class and comprise inputs $\mathbf{x}$ with the largest possible first component x_0. As the quantity of the erased data $\mathcal{D}_e$ increases from 30 (i.e., partially 'yellow' $\mathcal{D}_e$) to 40, $\mathcal{D}_e$ covers a larger part of the decision boundary (compare Figs. 11g vs. 11d); and
4. Fully 'yellow' $\mathcal{D}_e$ ($\mathcal{I} = 3.86$): The erased data of size $|\mathcal{D}_e| = 50$ comprise all data in the yellow class. In this case, $\mathcal{D}_e$ reduces the entropy of the model parameters/inducing variables $\mathbf{f}_{\mathcal{X}_u}$ given $\mathcal{D}_r$ the most when compared to the above 3 scenarios.

As $\mathcal{I}$ increases, the difference between the approximate posterior belief of $\mathbf{f}_{\mathcal{X}_u}$ given remaining data $\mathcal{D}_r$ vs. that given full data $\mathcal{D}$ increases. Though it is difficult to visualize such a difference directly, Proposition 1 tells us that this difference can be alternatively understood by comparing the predictive distributions $q(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ in Table 3 vs. $q(y_{\mathbf{x}} = 1|\mathcal{D})$ in Fig. 4b.

Fig. 11 shows results of averaged KL divergences (i.e., performance metric described in Sec. 4) achieved by EUBO, rKL, and $q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D})$ over $\mathcal{D}_r$ and $\mathcal{D}_e$ for the 4 scenarios above. Table 3 also analyzes the performance of our unlearning methods qualitatively by plotting the means of the approximate posterior beliefs $\tilde{q}_u(f_{\mathbf{x}}|\mathcal{D}_r;\lambda)$ and $\tilde{q}_v(f_{\mathbf{x}}|\mathcal{D}_r;\lambda)$ induced, respectively, by EUBO and rKL with the corresponding predictive distributions $\tilde{q}_u(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ and $\tilde{q}_v(y_{\mathbf{x}} = 1|\mathcal{D}_r)$, together with the mean of the approximate posterior belief $q(f_{\mathbf{x}}|\mathcal{D}_r)$ with the corresponding predictive distribution $q(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ obtained using VI from retraining with remaining data $\mathcal{D}_r$. The following observations result:

- Fig. 11 shows that as $\mathcal{I}$ increases across the 4 scenarios, the averaged KL divergence between $q(y_{\mathbf{x}}|\mathcal{D})$ vs. $q(y_{\mathbf{x}}|\mathcal{D}_r)$ over $\mathcal{D}_r$ and $\mathcal{D}_e$ (i.e., baseline labeled as *full*) generally increases.
- In the scenario of randomly selected $\mathcal{D}_e$ (i.e., $\mathcal{I}$ is small), we expect the difference between the predictive distributions $q(y_{\mathbf{x}}|\mathcal{D})$ vs. $q(y_{\mathbf{x}}|\mathcal{D}_r)$ over $\mathcal{D}_r$ and $\mathcal{D}_e$ to be small, which is reflected in the very small averaged KL divergences of about 0.002 and 0.004 achieved by $q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D})$ (i.e., baseline labeled as *full*) in Figs. 11b and 11c, respectively. It can also be observed that though EUBO and rKL with $\lambda \in \{10^{-5}, 10^{-9}\}$ achieve smaller averaged KL divergences than that of $q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D})$ (i.e., baseline), EUBO's averaged KL divergence increases beyond than that of the baseline when $\lambda = 0$, but remains very small. As a result, the first row in Table 3 shows that when $\lambda = 10^{-9}$ or $\lambda = 0$, the predictive distributions

$\tilde{q}_u(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ and $\tilde{q}_v(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ induced, respectively, by EUBO and rKL are similar to $q(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ obtained using VI from retraining with $\mathcal{D}_r$. Hence, we can conclude that both EUBO and rKL perform reasonably well in this scenario, even when $\lambda = 0$.

- In the scenarios of partially and largely 'yellow' $\mathcal{D}_e$, $\mathcal{I}$ is much larger than that in the scenario of randomly selected $\mathcal{D}_e$. So, we expect an increase in the difference between the predictive distributions $q(y_{\mathbf{x}}|\mathcal{D})$ vs. $q(y_{\mathbf{x}}|\mathcal{D}_r)$ over $\mathcal{D}_r$ and $\mathcal{D}_e$. It can be observed from Figs. 11e-f and 11h-i that when $\lambda = 0$, EUBO performs poorly as its averaged KL divergence is larger than that of $q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D})$ (i.e., baseline labeled as *full*), while rKL performs well as its averaged KL divergence is much smaller than that of the baseline. On the other hand, when $\lambda = 10^{-9}$, both EUBO and rKL perform well, which can also be observed from the second and third rows of Table 3. These plots also show that while the predictive distributions $\tilde{q}_v(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ induced by rKL with $\lambda = 10^{-9}$ are not as similar to $q(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ as $\tilde{q}_u(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ induced by EUBO with $\lambda = 10^{-9}$, the performance of rKL with $\lambda = 0$ is more robust.
- In the scenario of fully 'yellow' $\mathcal{D}_e$ (i.e., $\mathcal{I}$ is largest), the difference between the predictive distributions $q(y_{\mathbf{x}}|\mathcal{D})$ vs. $q(y_{\mathbf{x}}|\mathcal{D}_r)$ over $\mathcal{D}_r$ and $\mathcal{D}_e$ is larger than that in the above 3 scenarios. Except for EUBO with $\lambda = 0$, the predictive distributions $\tilde{q}_u(y_{\mathbf{x}}|\mathcal{D}_r)$ and $\tilde{q}_v(y_{\mathbf{x}}|\mathcal{D}_r)$ induced, respectively, by EUBO and rKL are closer to $q(y_{\mathbf{x}}|\mathcal{D}_r)$ than $q(y_{\mathbf{x}}|\mathcal{D})$ as they achieve smaller averaged KL divergences than that of $q(\mathbf{f}_{\mathcal{X}_u}|\mathcal{D})$, as shown in Figs. 11k-l. However, the fourth row of Table 3 shows that both EUBO and rKL do not perform that well. Nevertheless, it can be observed that when $\lambda = 0$, the predictive distribution $\tilde{q}_v(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ induced by rKL is still usable while $\tilde{q}_u(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ induced by EUBO is useless.

To summarize, when only an approximate posterior belief $q(\boldsymbol{\theta}|\mathcal{D})$ of model parameters $\boldsymbol{\theta} = \mathbf{f}_{\mathcal{X}_u}$ given full data $\mathcal{D}$ (i.e., obtained in model training with VI) is available, both EUBO and rKL can perform well if the difference between the approximate posterior belief of model parameters given remaining data $\mathcal{D}_r$ vs. that given full data $\mathcal{D}$ is sufficiently small. In practice, this is expected due to the small quantity of erased data and redundancy in real-world datasets. In the case where the erased data is highly informative, the approximate posterior belief $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r; \lambda = 0)$ induced by rKL remains usable by being close to $q(\boldsymbol{\theta}|\mathcal{D})$ and hence sacrificing its unlearning performance. On the other hand, EUBO may suffer from poor unlearning performance when λ is too small.

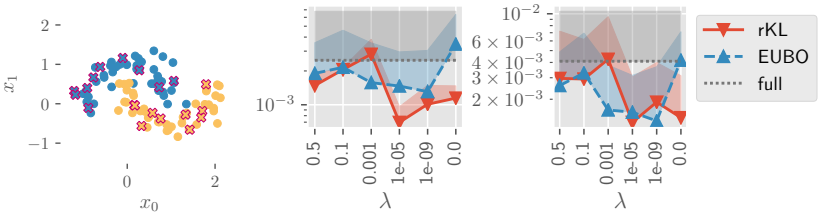
The above remark highlights the limitation of our unlearning methods when the erased data $\mathcal{D}_e$ is informative and only the approximate posterior belief $q(\boldsymbol{\theta}|\mathcal{D})$ is available. Such a limitation is due to the lack of information about the difference between the exact posterior belief $p(\boldsymbol{\theta}|\mathcal{D})$ vs. the approximate one $q(\boldsymbol{\theta}|\mathcal{D})$ (Sec. 3.3), which motivates future investigation into maintaining additional information about this difference during the model training with VI to improve the unlearning performance. In practice, an ML application may require an unlearning method to be time-efficient in order to satisfy the constraint on the response time to a user's request for her data to be erased while not rendering the model useless (e.g., due to catastrophic unlearning). After processing the user's request, the ML application can continue to improve the approximate posterior belief recovered by unlearning from erased data (i.e., using our proposed EUBO or rKL) by retraining with the remaining data at the expense of parsimony (i.e., in terms of time and space costs).

One may wonder how our unlearning methods can handle multiple users' request arriving sequentially over time. To avoid approximation errors from accumulating, we can adopt the approach of *lazy* unlearning by aggregating all the (past and new) users' erased data into $\mathcal{D}_e$ and performing unlearning (i.e., using only $q(\boldsymbol{\theta}|\mathcal{D})$ and $\mathcal{D}_e$) as and when necessary. As expected, our unlearning methods can perform well, provided that the aggregated erased data $\mathcal{D}_e$ remains sufficiently small or contains enough redundancy.

G Logistic Regression with Fashion MNIST Dataset: Additional Experimental Results

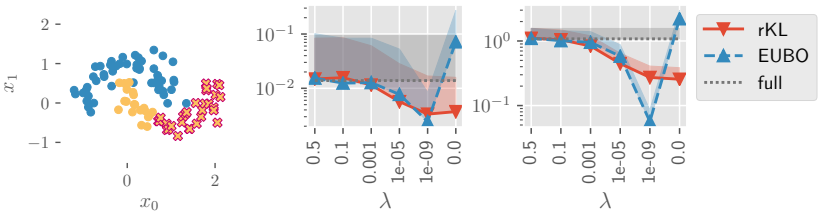
In this section, we will present the following:

- Additional visualizations of the class probabilities for images in $\mathcal{D}_r$ evaluated at the mean of the approximate posterior beliefs obtained using EUBO and rKL with $\lambda = 0$ in Fig. 13, and



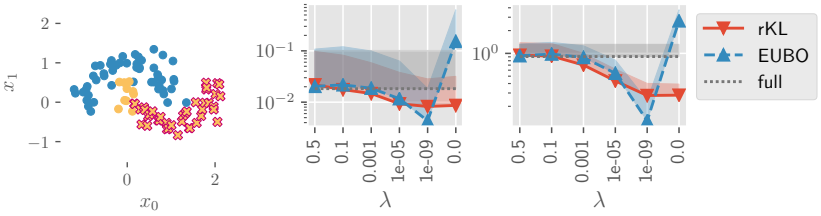
$\mathcal{I} = 0.27$

(a) Randomly selected $\mathcal{D}_e$ (b) $\mathcal{D}_r$ (c) $\mathcal{D}_e$



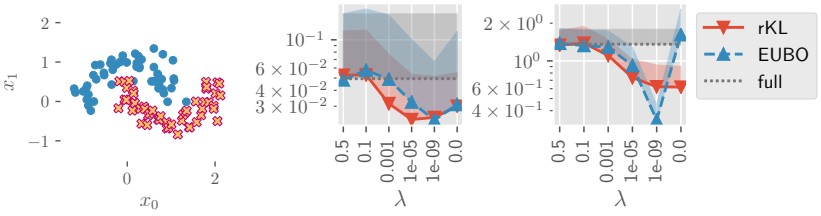
$\mathcal{I} = 1.59$

(d) Partially 'yellow' $\mathcal{D}_e$ (e) $\mathcal{D}_r$ (f) $\mathcal{D}_e$



$\mathcal{I} = 2.06$

(g) Largely 'yellow' $\mathcal{D}_e$ (h) $\mathcal{D}_r$ (i) $\mathcal{D}_e$



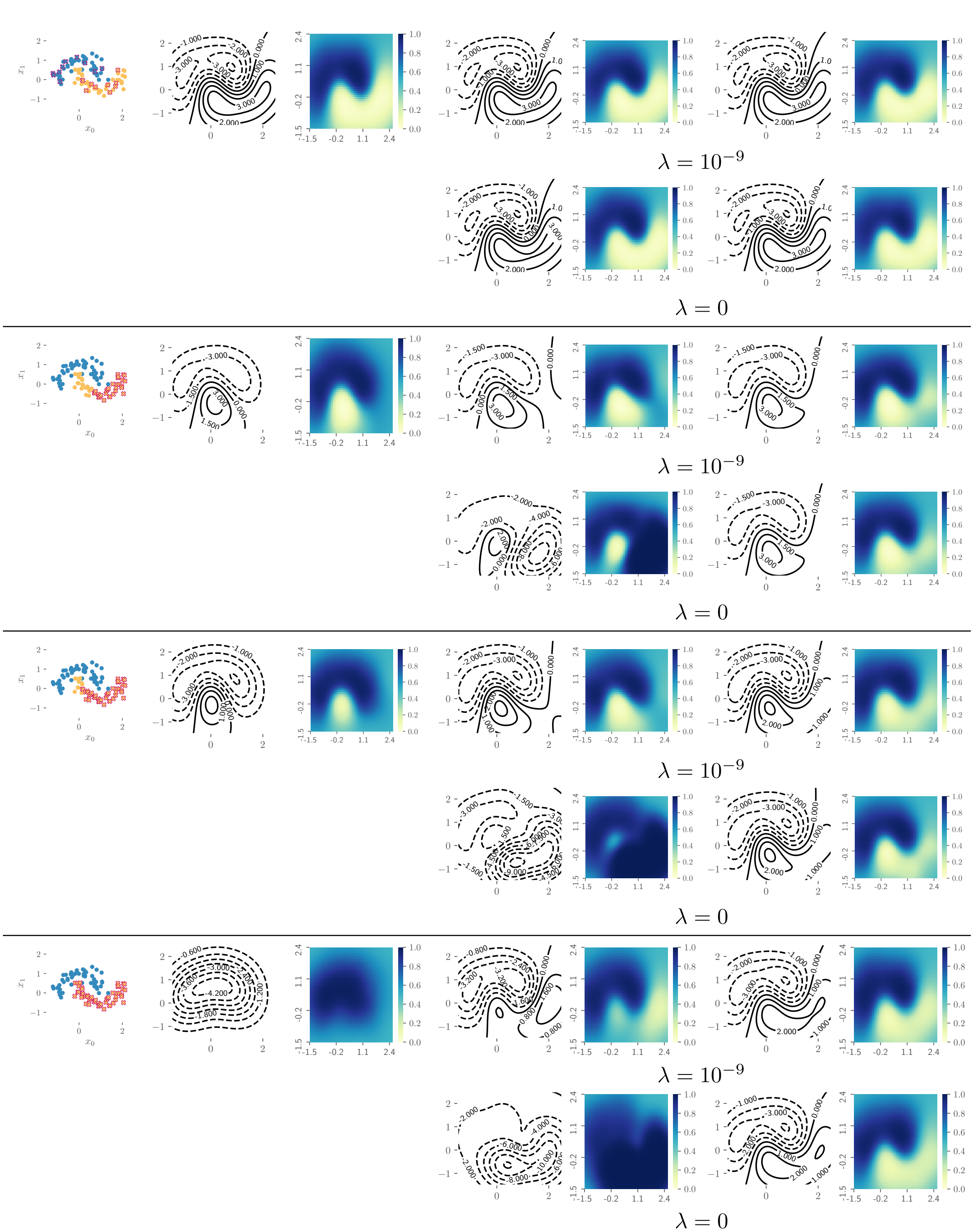
$\mathcal{I} = 3.86$

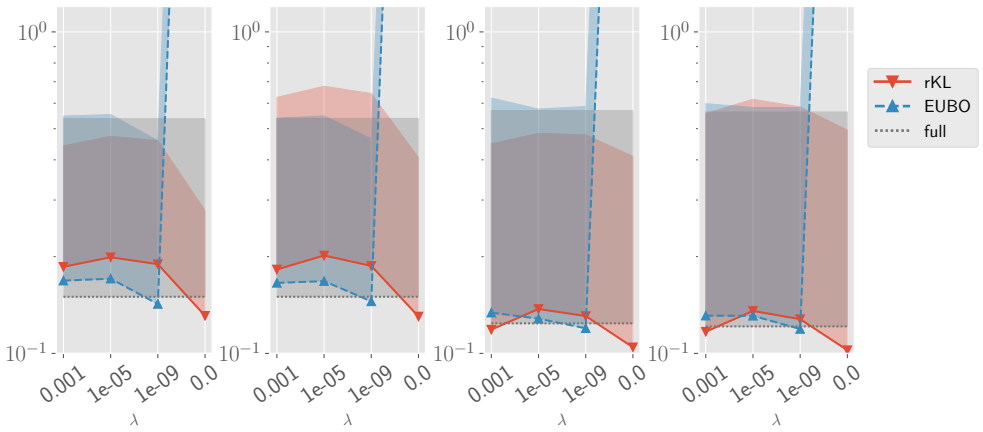
(j) Fully 'yellow' $\mathcal{D}_e$ (k) $\mathcal{D}_r$ (l) $\mathcal{D}_e$

Figure 11: Plots of (a,d,g,j) synthetic moon dataset with erased data $\mathcal{D}_e$ (crosses) and remaining data $\mathcal{D}_r$ (dots) in 4 different scenarios. Graphs of averaged KL divergence vs. λ achieved by EUBO, *reverse KL* (rKL), and $q(\boldsymbol{\theta}|\mathcal{D})$ (i.e., baseline labeled as *full*) over $\mathcal{D}_r$ and $\mathcal{D}_e$ in the following 4 scenarios: (b-c) randomly selected $\mathcal{D}_e$, (e-f) partially 'yellow' $\mathcal{D}_e$, (h-i) largely 'yellow' $\mathcal{D}_e$, and (k-l) fully 'yellow' $\mathcal{D}_e$.

Table 3: Plots of the mean of approximate posterior belief $q(f_{\mathbf{x}}|\mathcal{D}_r)$ with the corresponding predictive distribution $q(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ obtained using VI from retraining with remaining data $\mathcal{D}_r$, and also the means of approximate posterior beliefs $\tilde{q}_u(f_{\mathbf{x}}|\mathcal{D}_r; \lambda)$ and $\tilde{q}_v(f_{\mathbf{x}}|\mathcal{D}_r; \lambda)$ induced, respectively, by EUBO and rKL with the corresponding predictive distributions $\tilde{q}_u(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ and $\tilde{q}_v(y_{\mathbf{x}} = 1|\mathcal{D}_r)$ for $\lambda \in [10^{-9}, 0]$. The 1-st, 2-nd, 3-rd, and 4-th rows correspond to the following 4 respective scenarios: randomly selected $\mathcal{D}_e$, partially 'yellow' $\mathcal{D}_e$, largely 'yellow' $\mathcal{D}_e$, and fully 'yellow' $\mathcal{D}_e$.

Dataset	Retrained		EUBO		rKL	
	Mean $\mu_{\mathbf{x}|\mathcal{D}_r}$	$q(y_{\mathbf{x}} = 1|\mathcal{D}_r)$	Mean	$\tilde{q}_u(y_{\mathbf{x}} = 1|\mathcal{D}_r)$	Mean	$\tilde{q}_v(y_{\mathbf{x}} = 1|\mathcal{D}_r)$





(a) $\mathcal{D}_r$ (b) $\mathcal{D}_e$ (c) $\mathcal{D}_r$ (d) $\mathcal{D}_e$

Figure 12: Graphs of averaged KL divergence vs. λ achieved by EUBO, rKL, and $q(\boldsymbol{\theta}|\mathcal{D})$ (i.e., baseline labeled as *full*) over $\mathcal{D}_r$ and $\mathcal{D}_e$ for the fashion MNIST dataset. The approximate posterior beliefs of the model parameters/weights are represented by (a-b) independent Gaussians (i.e., diagonal covariance matrices) and (c-d) multivariate Gaussians (i.e., full covariance matrices).

- Comparison of the unlearning performance obtained using approximate posterior beliefs modeled with independent Gaussians (i.e., diagonal covariance matrices) vs. that modeled with multivariate Gaussians (i.e., full covariance matrices).

Fig. 13 shows the class probabilities for the images in $\mathcal{D}_r$ evaluated at the mean of the approximate posterior beliefs with $\lambda = 0$. Figs. 13a-d and 13g show that rKL induces the highest class probability for the same class as that of $q(\boldsymbol{\theta}|\mathcal{D}_r)$. In Figs. 13e-f and 13h, the class probabilities obtained using optimized $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r; \lambda = 0)$ resemble that obtained using $q(\boldsymbol{\theta}|\mathcal{D})$, though the probability of the correct class is reduced due to unlearning.

Fig. 12 shows the averaged KL divergences of EUBO, rKL, and $q(\boldsymbol{\theta}|\mathcal{D})$ where the approximate posterior beliefs are modeled with independent Gaussians (i.e., diagonal covariance matrices) in Figs. 12a-b and multivariate Gaussians (i.e., full covariance matrices) in Figs. 12c-d. It can be observed that the averaged KL divergences between $q(y_{\mathbf{x}}|\mathcal{D})$ vs. $q(y_{\mathbf{x}}|\mathcal{D}_r)$ over $\mathcal{D}_r$ and $\mathcal{D}_e$ (i.e., baselines labeled as *full*) decrease when multivariate Gaussians with full covariance matrices are used to model the approximate posterior beliefs instead (compare the baselines labeled as *full* in Figs. 12c-d vs. that in Figs. 12a-b). Furthermore, in such a case, the unlearning performance of both EUBO and rKL improve as their averaged KL divergences are not as large (relative to the baselines) as that using independent Gaussians.

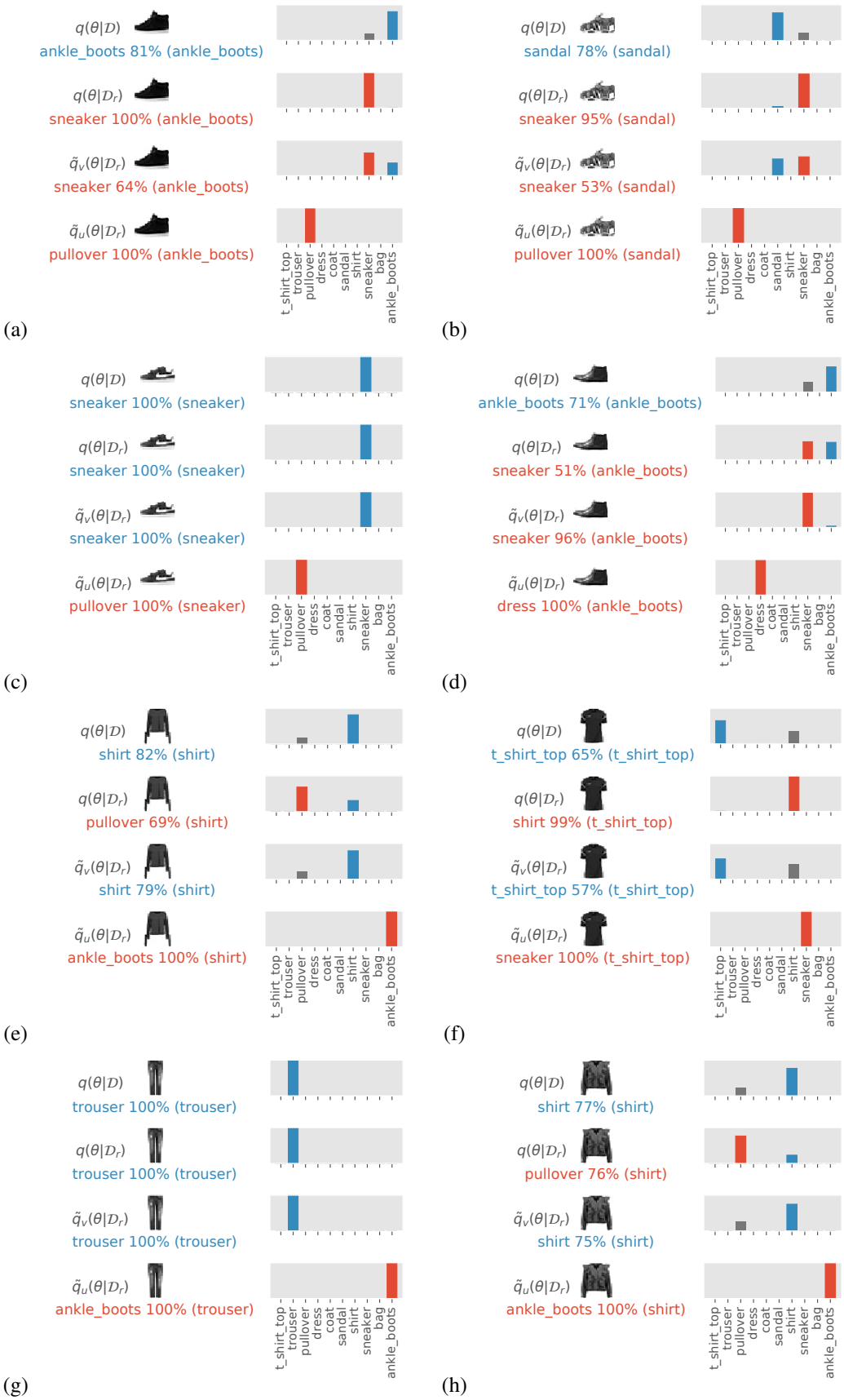


Figure 13: Plots of class probabilities for images in $\mathcal{D}_r$ obtained using $q(\boldsymbol{\theta}|\mathcal{D})$, $q(\boldsymbol{\theta}|\mathcal{D}_r)$, optimized $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r; \lambda = 0)$ and $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r; \lambda = 0)$.