A unified PAC-Bayesian framework for machine unlearning via information risk minimization

Jose, Sharu; Simeone, Osvaldo

DOI:
10.1109/MLSP52302.2021.9596170

License:
Other (please provide link to licence statement

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Jose, S & Simeone, O 2021, A unified PAC-Bayesian framework for machine unlearning via information risk minimization. in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP).*, 9596170, Machine learning for signal processing , IEEE. https://doi.org/10.1109/MLSP52302.2021.9596170

Link to publication on Research at Birmingham portal

2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP) | 978-1-7281-6338-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/MLSP52302.2021.9596170

A UNIFIED PAC-BAYESIAN FRAMEWORK FOR MACHINE UNLEARNING VIA INFORMATION RISK MINIMIZATION

Sharu Theresa Jose, Osvaldo Simeone

King's Communications, Learning, and Information Processing (KCLIP) Lab
Department of Engineering
King's College London
London, WC2R 2LS
emails: sharu.jose@kcl.ac.uk, osvaldo.simeone@kcl.ac.uk

ABSTRACT

Machine unlearning refers to mechanisms that can remove the influence of a subset of training data upon request from a trained model without incurring the cost of re-training from scratch. This paper develops a unified PAC-Bayesian framework for machine unlearning that recovers the two recent design principles – variational unlearning [1] and forgetting Lagrangian [2]– as information risk minimization problems [3]. Accordingly, both criteria can be interpreted as PAC-Bayesian upper bounds on the test loss of the unlearned model that take the form of free energy metrics.

***Index Terms*—** Machine unlearning, PAC-Bayesian bounds, free energy minimization

1. INTRODUCTION

AI tools are increasingly widespread and subject to privacy attacks and data misuse. Recent regulations, such as the European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act, has enshrined in law the right for individuals to withdraw consent to the use of their personal data for training machine learning models. The mere deletion of the requested data from the training data set does not serve the purpose, as information about the deleted data can still be retrieved from already trained machine learning models [4]. Thus, data deletion necessitates the machine learning model to *unlearn* the contribution of the deleted data to the training process, such that the resulting model behaves as if it has never observed the data in the first place.

A straightforward approach to unlearn is to retrain the model from scratch by using only data remaining after deletion of the data to be unlearnt. However, this is computationally intensive and resource expensive. *Machine unlearning* refers to mechanisms that can remove the influence of a specific subset of the training data on a trained machine learning model, without incurring the cost of retraining from scratch [5], [6].

Several machine unlearning approaches have been studied since the introduction of the concept in [5], where the problem was studied in the context of statistical query learning. [7] proposes an unlearning approach that partitions the data set into *shards* that are used to train multiple models in *isolation* and finally *aggregrated*. This allows unlearning to be carried out by aggregating only the remaining shards, avoiding the need for retraining.

Our work is motivated by two recently proposed machine unlearning mechanisms. The first proposes a design criterion, termed Evidence Upper BOund (EUBO), for *variational unlearning* within a Bayesian setting [1], while the second optimizes over a "scrubbing function" by minimizing a *forgetting Lagrangian* criterion [2]. Although prima facie these two approaches seem different, we demonstrate that the two design principles can be interpreted in a unified manner in the context of PAC-Bayesian theory [8], [9]. PAC-Bayesian theory develops high-probability upper bounds on the population loss of a learning algorithm in terms of a free energy metric that includes the sum of a training loss and the Kullback-Leibler (KL) divergence between the learning algorithm and a data-independent *prior* distribution [9, 10].

The main contributions of the paper are summarized as follows. We develop a unified PAC-Bayesian framework for machine unlearning that explains the unlearning design principles in [1] and [2] through the principle of information risk minimization (IRM) [3]. The PAC-Bayesian formulation makes use of the recent result in [11] that accounts for data-dependent priors. We show that the design criteria – EUBO and forgetting Lagrangian – optimize PAC-Bayesian bounds with appropriate choices of training loss and data-dependent prior. Finally, the proposed framework motivates the design of *amortized* variants of variational unlearning and forgetting

The authors have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731)

978-1-7281-6338-3/21/$31.00 ©2021 Crown

Lagrangian-based mechanisms, which are also described.

2. LEARNING AND UNLEARNING ALGORITHMS

In this section, we start by defining the operation and performance criteria of learning and unlearning algorithms. These are described as stochastic mappings as in the standard PAC-Bayes framework.

2.1. Learning Algorithm

Let $D = (Z_1, \dots, Z_n)$ denote a training data set of n samples generated i.i.d. according to an unknown population distribution $P_Z \in \mathcal{P}(\mathcal{Z})$. A learning algorithm uses the data set D to infer a model parameter W belonging to a model class $\mathcal{W}$. We define the learning algorithm as a stochastic mapping, $P_{W|D} \in \mathcal{P}(\mathcal{W})$[1], from the input training set D to the model class, $\mathcal{W}$. The probabilistic mapping $P_{W|D}$ describes a distribution over all possible outcomes W in the model class $\mathcal{W}$.

Let $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_+$ denote a loss function. The goal of the learning algorithm is to find a model parameter $w \in \mathcal{W}$ that minimizes the *population loss*,

$$L(w) = \mathbb{E}_{P_Z}[\ell(w, Z)], \tag{1}$$

which is the average loss of the model parameter w incurred on a new test data point $Z \sim P_Z$. The population loss (1) is unknown to the learner, since the underlying population distribution P_Z is not available. Instead, the learner uses the *empirical training loss* on the data set D, i.e.,

$$\widehat{L}(w|D) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i) \tag{2}$$

as the training criterion. For a given training data set D, we define the *generalization error*, $\Delta\mathcal{L}(P_{W|D})$, of a learning mechanism $P_{W|D}$ as the average difference between the population loss (1) and the training loss (2), i.e.,

$$\Delta\mathcal{L}(P_{W|D}) = \mathbb{E}_{P_{W|D}}[L(W) - \widehat{L}(W|D)]. \tag{3}$$

The generalization error (3) quantifies the extent to which the training loss (2) can be reliably used as a proxy measure for the unknown population loss.

2.2. Machine Unlearning

Consider a model $W_l \sim P_{W|D}$ learned using the data set D. When a request is received to "delete" a subset $D_e \subset D$ of m samples, the learned model W_l must be updated so as to "unlearn" the information extracted from the data set D_e by the learning process. We refer to data set D_e as the *unlearning data set*. Ideally, this could be done by re-training from scratch by using the remaining data, $D_r = D \setminus D_e$, i.e., by applying the stochastic mapping $P_{W|D_r}$. Given the large computational cost of re-training, *machine unlearning* aims to remove the influence of the data D_e on the learned model W_l without incurring the full cost of re-training from scratch. Formally, we define an unlearning algorithm as follows [12].

Definition 2.1 (Unlearning Algorithm) *An unlearning algorithm $P_{W|W_l,T(D),D_e}$ is a stochastic mechanism that maps the learned model parameter $W_l \sim P_{W|D}$, a statistic $T(D)$ of data set D, and the unlearning data set D_e to the space of model parameters $\mathcal{W}$.*

We note that the rationale for making the unlearned model $W \sim P_{W|W_l,T(D),D_e}$ depend on a statistic $T(D)$ of D is to rule out training from scratch. In fact, if the statistic is $T(D) = D$, the unlearning algorithm can ignore W_l and retrain from scratch, while more restrictive choices of $T(D)$ make this impossible.

In order to ensure successful unlearning, one needs to impose that the distribution of the unlearned model W be close to that obtained by training from scratch. For fixed data sets D and D_e, the latter distribution is $P_{W|D_r}$, while the former is given by the average $\mathbb{E}_{P_{W_l|D}}[P_{W|W_l,T(D),D_e}]$ over the learning mechanism. Note that the expectation marginalizes over the learned models. This constraint can be formalized as follows.

Definition 2.2 (ϵ-certified unlearning) *An unlearning algorithm $P_{W|W_l,T(D),D_e}$ is said to satisfy ϵ-certified unlearning for $\epsilon > 0$ if*

$$D_{\mathrm{KL}}(\mathbb{E}_{W_l \sim P_{W|D}}[P_{W|W_l,T(D),D_e}]||P_{W|D_r}) \leq \epsilon, \tag{4}$$

where $D_{\mathrm{KL}}(P||Q)$ denotes the KL divergence between distributions P and Q.

By the biconvexity of the KL divergence, it is easy to see that the unlearning certificate in (4) is implied by the stronger condition that the inequality

$$D_{\mathrm{KL}}(P_{W|W_l,T(D),D_e}||P_{W|D_r}) \leq \epsilon \tag{5}$$

applies for all $W_l \in \mathcal{W}$ in the support of $P_{W|D}$.

3. PRELIMINARIES

In this section, we briefly review the classical PAC-Bayesian framework, which underlies the proposed unified approach to machine unlearning. PAC Bayesian theory [8, 13] provides upper bounds on the average population loss, $\mathbb{E}_{P_{W|D}}[L(W)]$, of a learning algorithm $P_{W|D}$ in terms of: (a) the average training loss, $\mathbb{E}_{P_{W|D}}[\widehat{L}(W|D)]$, and (b) the KL divergence between the distribution $P_{W|D}$ and an arbitrary data-independent "prior" Q_W. The PAC-Bayesian bounds hold with high probability over random draws of the training data set D. There has been extensive study on various refinements

[1] We use $\mathcal{P}(\cdot)$ to denote the space of all probability distributions on '$\cdot$'.

to the original PAC-Bayesian bound of [8] (see [14] for a review). More recently, PAC-Bayesian bounds have been extended to account for *data-dependent* priors [15], [11].

In this work, we make use of the general PAC-Bayesian bound derived in Theorem 2 of [11] that allows for data-dependent priors. This turns out to be important for unlearning, since the prior will be used to account for the learning algorithm. The next lemma restates Theorem 2 in [11] by using our notation and by adopting a conventional formulation in terms of uniform bounds over all posteriors $P_{W|D}$. A proof is provided for completeness in Section 6.1.

Lemma 3.1 *Let $Q_{W|D}$ denote a data-dependent prior. For any (measurable) function $A : \mathcal{Z}^n \times \mathcal{W} \to \mathbb{R}^2$ and convex function $F : \mathbb{R}^2 \to \mathbb{R}$, let $f : \mathcal{Z}^n \times \mathcal{W} \to \mathbb{R}$ be the composition of F and A, and let $\xi = \mathbb{E}_{P_Z^{\otimes n}} \mathbb{E}_{Q_{W|D}}[\exp(f(D, W))]$. Then, with probability at least $1 - \delta$, with $\delta \in (0, 1)$, over the random draw of data set $D \sim P_Z^{\otimes n}$, the following inequality holds uniformly over all stochastic mappings $P_{W|D}$*

$$\begin{aligned} &F(\mathbb{E}_{P_{W|D}}[A(D, W)]) \\ &\leq D_{\mathrm{KL}}(P_{W|D}||Q_{W|D}) + \log(\xi/\delta). \end{aligned} \tag{6}$$

In the rest of the paper, we will use Lemma 3.1 by selecting function $A(D, W)$ to output a tuple including the population loss $L(W)$ and a training loss metric to be specialized for different unlearning methods. Furthermore, the convex function F will be chosen to output the difference of its inputs, i.e., $F(a, b) = a - b$. With these choices, the PAC-Bayesian bound in (6) will allow us to relate the empirical training metrics and the unknown population loss.

For reference, in the standard analysis of learning algorithms, the function $A(D, W)$ is selected to be the two-dimensional vector $[\beta L(W), \beta \widehat{L}(W|D)]$. With this choice, the bound in (6) can be re-written as an upper bound on the population loss that holds for all $P_{W|D}$:

$$\mathbb{E}_{P_{W|D}}[L(W)] \leq \mathcal{F}_{\mathrm{IRM}} + \frac{1}{\beta}\log(\xi/\delta), \quad \text{where,} \tag{7}$$

$$\mathcal{F}_{\mathrm{IRM}} = \mathbb{E}_{P_{W|D}}[\widehat{L}(W|D)] + \frac{1}{\beta} D_{\mathrm{KL}}(P_{W|D}||Q_{W|D}).$$

Important to our framework is the observation that the PAC-Bayesian bound (6), and hence also (7), hold uniformly over all choices of the learning algorithm $P_{W|D}$. As such, one can optimize the right-hand side of (7) ove the learning algorithm $P_{W|D}$ by considering the problem $\min_{P_{W|D}} \mathcal{F}_{\mathrm{IRM}}$. By minimizing an upper bound on the population loss, the learning criterion (7) facilitates generalization. This approach is known as *Information Risk Minimization (IRM)* [3], and it amounts to the minimization of a *free energy criterion* [10]. A free energy criterion is given by the sum of a training loss and of an information-theoretic regularization.

The PAC-Bayesian bound in (6) contains a constant term ξ, bounding which ensures non-vacuous bounds on the generalization error. For data-independent priors, under suitable assumptions on the loss function, such as boundedness or sub-Gaussianity, the constant ξ can be easily upper bounded. An upper bound on ξ for a data-dependent prior has been recently obtained in [11]. Since we will use (6) to justify unlearning criteria via variants of the IRM problem, we will not be further concerned with bounding ξ.

4. VARIATIONAL UNLEARNING

In this section, we study the Bayesian unlearning framework introduced in the recent work [1]. As we first review, this paper presents a new unlearning criterion, termed *Evidence Upper BOund* (EUBO), that enables variational unlearning. To be consistent with Definition 2.1, we specifically describe here an amortized variational unlearning variant of the approach proposed in [1]. We then show that the resulting unlearning algorithm can be interpreted as IRM, which is obtained through a specific instantiation of the PAC-Bayesian bound (6).

4.1. Amortized Variational Unlearning

In order to meet the unlearning requirement (4) for some $\epsilon > 0$, the variational unlearning framework proposed in [1] finds a distribution in the model parameter space $\mathcal{W}$ that is closest, in terms of KL divergence, to the distribution $P_{W|D_r}$ resulting from re-training on the remaining data D_r. Optimization is restricted to a given family of distributions.

The approach requires the variational optimization to be carried out separately for any given selection of data sets D and D_e. Furthermore, it relies on access to the distribution $P_{W|D}$ and not solely on a trained model W_l. In contrast, an efficient unlearning mechanism conforming to Definition 2.1 must define a conditional probability distribution $P_{W|W_l,T(D),D_e}$ that can be instantiated for any choice of learned model W_l, statistic $T(D)$ of the data, and unlearning data set D_e. To this end, in this section, we develop an amortized variant of variational unlearning [1] that enables optimization over an unlearning mechanism $P_{W|W_l,T(D),D_e}$. We refer to this approach as *amortized variational unlearning* (AVU).

The proposed AVU framework constrains the unlearning mechanism $P_{W|W_l,T(D),D_e}$ to belong to a family $\mathcal{Q}^{\mathrm{AVU}}$ of (parameterized) conditional distributions on $\mathcal{W}$. AVU seeks to find the unlearning mechanism $P_{W|W_l,T(D),D_e}$ that solves the following problem

$$\min_{\substack{P_{W|W_l,T(D),D_e} \\ \in \mathcal{Q}^{\mathrm{AVU}}}} \mathbb{E}_{P_{D,D_e} P_{W_l|D}} \Big[D_{\mathrm{KL}}(P_{W|W_l,T(D),D_e} || P_{W|D_r}) \Big], \tag{8}$$

where $P_{W_l|D}$ denote the distribution of the learned model $W_l \sim P_{W|D}$, and P_{D,D_e} denote the probability distribution of the training data $D \sim P_Z^{\otimes n}$ and of the unlearning data set

$D_e \sim P_{D_e|D}$. The conditional distribution $P_{D_e|D}$ describes a uniformly distributed stochastic selection of a subset D_e of m samples from D. Problem (8) aims at ensuring that the unlearning condition (5) be satisfied on average over all training data set D and unlearning data set D_e for small value of $\epsilon > 0$.

Following [1], the optimization problem in (8) can be equivalently formulated as

$$\min_{\substack{P_{W|W_l,T(D),D_e} \\ \in \mathcal{Q}^{\text{AVU}}}} \mathbb{E}_{P_{D,D_e}P_{W_l|D}} \Big[\text{EUBO}(P_{W|W_l,T(D),D_e}, P_{W|D})\Big], \tag{9}$$

where the Evidence Upper BOund (EUBO) is defined as

$$\begin{aligned} &\text{EUBO}(P_{W|W_l,T(D),D_e}, P_{W|D}) \\ =&\mathbb{E}_{P_{W|W_l,T(D),D_e}}[\log P_{D_e|W}] \\ &+ D_{\text{KL}}(P_{W|W_l,T(D),D_e}||P_{W|D}). \end{aligned} \tag{10}$$

The EUBO (10) comprises of two terms: (i) the average positive log-likelihood of the unlearning data set D_e obtained after unlearning; and (ii) the deviation of the unlearning mechanism from the learning algorithm $P_{W|D}$. Intuitively, the first term should be small for effective unlearning, while the second is a regularization penalty that accounts for the residual epistemic uncertainty associated with the training algorithm.

4.2. A PAC-Bayesian View of Variational Unlearning

We now demonstrate that the optimization (10) can be justified as an IRM obtained from the PAC-Bayesian bound in (6). To instantiate the PAC-Bayesian bound in (6) for unlearning, we note that the unlearning mechanism in Definition 2.1 is a cascade of two operations: (a) sample model parameter $W_l \sim P_{W|D}$ according to the learning mechanism; and then (b) apply the unlearning mechanism $P_{W|W_l,T(D),D_e}$ on the learned model W_l. This process is subject to the random draw of data $D \sim P_Z^{\otimes n}$ and to the random selection of subset of data to be removed, $D_e \sim P_{D_e|D}$. In line with this observation, we have the following PAC-Bayesian bound for the unlearning mechanism.

Corollary 4.1 *Let the data dependent prior be fixed as the learning mechanism $P_{W|D}$. With probability at least $1-\delta$, with $\delta \in (0,1)$, over the random draw of the data set $D \sim P_Z^{\otimes n}$ and the subset $D_e \subset D$ to be removed, the following inequality holds uniformly for all unlearning algorithms $P_{W|W_l,T(D),D_e}$:*

$$\begin{aligned} &\mathbb{E}_{P_{W_l|D}}\mathbb{E}_{P_{W|W_l,T(D),D_e}}[-\mathbb{E}_{P_Z}[\log P_{Z|W}]] \\ \leq&\mathbb{E}_{P_{W_l|D}}\left[\frac{1}{m}\text{EUBO}(P_{W|W_l,T(D),D_e}, P_{W|D})\right] \\ &+ \frac{1}{m}\log\frac{\bar{\xi}_{\text{AVU}}}{\delta}, \end{aligned} \tag{11}$$

where $\bar{\xi}_{\text{AVU}} = \mathbb{E}_{P_{D,D_e}P_{W|D}}[\exp(m(-\mathbb{E}_{P_Z}[\log P_{Z|W}] - (1/m)\log P_{D_e|W})]$.

Proof: This result is obtained from Lemma 3.1 by selecting $A(D,W)$ as the two-dimensional vector $[-m\mathbb{E}_{P_Z}[\log P_{Z|W}], \log P_{D_e|W}]$ and $F(a,b) = a-b$. Details can be found in Section 6.2. ■

The left-hand side in (11) is the average test log-loss obtained by the unlearnt model. Therefore, by (11), the variational unlearning mechanism introduced in [1] can be interpreted as minimizing an upper bound on the test log-loss over the unlearning mechanism $P_{W|W_l,T(D),D_e}$ (assuming knowledge of $P_{W|D}$). By (10), this minimization is of the form (7) assumed by IRM problems [3]. As $\delta \to 0$, the inequality in (11) holds almost surely, which justifies taking the average in (10) over the draws of D and D_e. It follows that the proposed AVU (10) can be similarly interpreted in terms of the minimization of a PAC-Bayes upper bound on the average test log-loss, and hence in terms of an IRM problem.

5. FORGETTING LAGRANGIAN-BASED UNLEARNING

In this section, we first review the unlearning framework introduced in [2], the *Forgetting Lagrangian*, and show that this can also be intrepreted as an IRM obtained as a specific instantiation of (6).

5.1. Forgetting Lagrangian

Reference [2] considers a stochastic learning mechanism $P_{W|D}$ that trains the model parameter vector W of a deep neural network (DNN) using data set D. The unlearning mechanism $P_{W|W_l,T(D),D_e}$ ignores the statistic $T(D)$ and yields a *stochastic scrubbing function* $P_{W|W_l,D_e}$ that "scrubs off" the influence of the unlearning data set D_e on the learned model $W_l \sim P_{W|D}$.

The scrubbing function $P_{W|W_l,D_e}$ is designed so as to optimize the *Forgetting Lagrangian*,

$$\begin{aligned} &\mathcal{FL}(P_{W|W_l,D_e}, \lambda) = \mathbb{E}_{P_{W|W_l,D_e}}[\widehat{L}(W|D_r)] \\ &\quad + \lambda D_{\text{KL}}(\mathbb{E}_{P_{W_l|D}}[P_{W|W_l,D_e}]||\mathbb{E}_{P_{W_l|D_r}}[\tilde{P}_{W|W_l}]) \end{aligned} \tag{12}$$

where $\lambda > 0$ denotes a Lagrangian multiplier, and $\tilde{P}_{W|W_l}$ is a an arbitrary 'reference' distribution that maps the model $W_l \sim P_{W|D_r}$, obtained by retraining on the data set D_r, to a "noisy" version $W \in \mathcal{W}$. The forgetting Lagrangian in (12) thus aims at finding an unlearning mechanism that (a) minimizes the average training loss $\widehat{L}(w|D_r)$ on the remaining data D_r; while (b) ensuring that the unlearning mechanism $P_{W|W_l,D_e}$ applied on the learned model $W_l \sim P_{W|D}$ is close, in terms of KL divergence, to the reference distribution

$\tilde{P}_{W|W_l}$ applied on the model $W_l \sim P_{W|D_r}$ obtained after retraining from scratch. Thus, the KL divergence term in (12) ensures a "certificate of unlearning" with respect to the reference $\tilde{P}_{W|W_l}$ in the sense of Definition 2.2. Moreover, the KL divergence term can be interpreted as an upper bound on the information about the unlearning data set D_e that can be read out from observing the unlearned model $W \sim P_{W|W_l,D_e}$ [2].

As discussed in Section 4.1, designing the unlearning mechanism via the forgetting Lagrangian in (12) requires the optimization to be performed for each selection of the learned model W_l and the data sets D and D_e. Furthermore, it depends directly on the distribution $P_{W|D}$. Following the discussion in Section 4.1, we could address this problem by considering an *amortized forgetting Lagrangian* approach so as to optimize a conditional distribution $P_{W|W_l,D_e}$ that can be instantiated for any choice of learned model W_l, and unlearning data D_e. We do not pursue this here, since reference [2] shows that an approximate solution $P_{W|W_l,D_e}$ to problem (12) can be found that does not require a separate optimization for all D and D_e.

5.2. A PAC-Bayesian view of forgetting Lagrangian

We now show that the forgetting Lagrangian (12) follows from a specific instantiation of the PAC-Bayesian bound (6) for unlearning mechanisms.

Corollary 5.1 *Let the data dependent prior be fixed as* $\tilde{P}_{W|D,D_e} = \mathbb{E}_{P_{W_l|D_r}}[\tilde{P}_{W|W_l}]$. *Then, for all* $\beta > 0$, *with probability at least* $1-\delta$, *with* $\delta \in (0,1)$, *over the random draw of the data set* $D \sim P_Z^{\otimes n}$ *and the subset* $D_e \subset D$ *to be removed, the following inequality holds uniformly for all* $P_{W|W_l,D_e}$,

$$\mathbb{E}_{P_{W_l|D}P_{W|W_l,D_e}}[L(W)] \\ \leq \mathbb{E}_{P_{W_l|D}}[\mathcal{FL}(P_{W|W_l,D_e},\beta^{-1})] + \frac{1}{\beta}\log\frac{\xi_{\mathcal{FL}}}{\delta_2}, \qquad (13)$$

where $\xi_{\mathcal{FL}} = \mathbb{E}_{P_{D,D_e}\tilde{P}_{W|D,D_e}}[\exp(\beta(L(W) - \widehat{L}(W|D_r))]$.

Proof: The proof follows in the same steps as the proof of Corollary 4.1 with $A(D,W) = [\beta L(W), \beta\widehat{L}(W|D_r)]$. Details in Section 6.3. ■

The left-hand side of (13) is the average test loss, and hence the forgetting Lagrangian framework introduced in [2] can be again interpreted as minimizing an upper bound on the average test loss.

6. PROOFS OF MAIN RESULTS

6.1. Proof of Lemma 3.1

The PAC-Bayesian bound in (6) is obtained by first using a Markov inequality, and then applying change of measure as detailed next. The Markov inequality for a non-negative random variable Y states that with probability at least $1-\delta$, with $\delta \in (0,1)$, we have $Y \leq \mathbb{E}[Y]/\delta$. Precisely, the following inequality holds,

$$\Pr(Y \leq \mathbb{E}[Y]/\delta) \geq 1-\delta.$$

We specialize the above Markov inequality to our setting by taking $Y = \mathbb{E}_{Q_{W|D}}[\exp(f(D,W))]$. Note that Y is a function of the random variable D, and that $\mathbb{E}_{P_Z^{\otimes n}}[Y] = \xi$. Markov's inequality then gives that

$$\Pr_D\left(\mathbb{E}_{Q_{W|D}}[\exp(f(D,W)] \leq \frac{\xi}{\delta}\right) \geq 1-\delta. \qquad (14)$$

Applying change of measure then results in the following inequality

$$\Pr_D\left(\forall P_{W|D}, \mathbb{E}_{P_{W|D}}\left[\exp\left(f(D,W) - \log\frac{P_{W|D}(W|D)}{Q_{W|D}(W|D)}\right)\right] \\ \leq \frac{\xi}{\delta}\right) \geq 1-\delta. \qquad (15)$$

Using Jensen's inequality to take expectation inside the exponential term, and subsequently applying log on both sides of the inequality then results in

$$\Pr_D\left(\forall P_{W|D},\ \mathbb{E}_{P_{W|D}}[f(D,W)] - D_{\mathrm{KL}}(P_{W|D}||Q_{W|D}) \\ \leq \log\frac{\xi}{\delta}\right) \geq 1-\delta. \qquad (16)$$

Finally, noting that $f(D,W) = F(A(D,W))$ where F is convex, and applying Jensen's inequality again results in the PAC-Bayesian bound in (6).

6.2. Proof of Corollary 4.1

The required bound follows by instantiating the general PAC-Bayesian bound in Lemma 3.1 for unlearning. As such, the unlearning PAC-Bayesian bound depends on the cascade operation of learning a model $W_l \sim P_{W|D}$, and subsequent unlearning using $P_{W|W_l,T(D),D_e}$. This process is subject to the random draw of $D \sim P_Z^{\otimes n}$, and to the random selection of the subset $D_e \subset D$. Consequently, we consider the prior in Lemma 3.1 as $Q_{W|D,D_e}$, depending on both data sets D and D_e.

Lemma 3.1 then gives that with probability at least $1-\delta$ over the random draw of data set D, and that of the unlearning data set D_e, the following inequality holds uniformly over all distributions $P_{W|D,D_e}$,

$$F(\mathbb{E}_{P_{W|D,D_e}}[A(D,W)]) - D_{\mathrm{KL}}(P_{W|D,D_e}||Q_{W|D,D_e}) \\ \leq \log\frac{\xi}{\delta}. \qquad (17)$$

In particular, (17) holds for all learning mechanisms $P_{W|D}$ and unlearning mechanisms $P_{W|W_l,T(D),D_e}$ such that $P_{W|D,D_e} = \mathbb{E}_{P_{W_l|D}}[P_{W|W_l,T(D),D_e}]$ is the marginal of the joint distribution $P_{W_l|D} \otimes P_{W|W_l,T(D),D_e}$.

To get to (11), we consider the PAC-Bayesian bound (17) for a fixed learning algorithm $P_{W|D}$. Further, we take $Q_{W|D,D_e} = P_{W|D}$, $A(D,W) = [-m\mathbb{E}_{P_Z}[\log P_{Z|W}], \log P_{D_e|W}]$ and $F(a,b) = a - b$. Noting that $P_{W|D,D_e} = \mathbb{E}_{P_{W_l|D}}[P_{W|W_l,T(D),D_e}]$, we use the biconvexity of KL divergence to upper bound

$$\begin{aligned} &D_{\mathrm{KL}}(P_{W|D,D_e}||P_{W|D}) \\ &\leq \mathbb{E}_{P_{W_l|D}}[D_{\mathrm{KL}}[P_{W|W_l,T(D),D_e}||P_{W|D}]. \end{aligned}$$

Using all these in (17) yields the required bound in (11).

6.3. Proof of Corollary 5.1

The proof follows the same line as the proof of Corollary 4.1 in Section 6.2. To get to (13), we use (17) with $P_{W|D,D_e} = \mathbb{E}_{P_{W_l|D}}[P_{W|W_l,T(D),D_e}]$, $Q_{W|D,D_e} = \mathbb{E}_{P_{W_l|D_r}}[\tilde{P}_{W|W_l}]$ and $A(D,W) = [\beta L(W), \beta\widehat{L}(W|D_r)]$.

7. CONCLUSION

The paper presents a unified PAC-Bayesian framework for the design of machine unlearning algorithms. We show that two unlearning design criteria studied in literature – EUBO for variational unlearning [1] and Forgetting Lagrangian [2] can be interpreted as IRM obtained via specific instantiation of the proposed PAC-Bayesian framework.

8. REFERENCES

[1] Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet, "Variational bayesian unlearning," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[2] Aditya Golatkar, Alessandro Achille, and Stefano Soatto, "Eternal sunshine of the spotless net: Selective forgetting in deep networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9304–9312.

[3] Tong Zhang, "Information-Theoretic Upper and Lower Bounds for Statistical Estimation," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1307–1321, 2006.

[4] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019, pp. 267–284.

[5] Yinzhi Cao and Junfeng Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE Symposium on Security and Privacy*. IEEE, 2015, pp. 463–480.

[6] Antonio Ginart, Melody Y Guan, Gregory Valiant, and James Zou, "Making ai forget you: Data deletion in machine learning," *arXiv preprint arXiv:1907.05012*, 2019.

[7] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot, "Machine unlearning," *arXiv preprint arXiv:1912.03817*, 2019.

[8] David A McAllester, "PAC-Bayesian Model Averaging," in *Proc. of Annual Conf. Computational Learning Theory (COLT)*, July 1999, pp. 164–170.

[9] Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand, "Pac-bayesian learning of linear classifiers," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 353–360.

[10] Sharu Theresa Jose and Osvaldo Simeone, "Free energy minimization: A unified framework for modeling, inference, learning, and optimization [lecture notes]," *IEEE Signal Processing Magazine*, vol. 38, no. 2, pp. 120–125, 2021.

[11] Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor, "Pac-bayes analysis beyond the usual bounds," *arXiv preprint arXiv:2006.13057*, 2020.

[12] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh, "Remember what you want to forget: Algorithms for machine unlearning," *arXiv preprint arXiv:2103.03279*, 2021.

[13] David A McAllester, "PAC-Bayesian stochastic model selection," *Machine Learning*, vol. 51, no. 1, pp. 5–21, 2003.

[14] Benjamin Guedj, "A primer on PAC-Bayesian learning," *arXiv preprint arXiv:1901.05353*, 2019.

[15] Gintare Karolina Dziugaite and Daniel M Roy, "Data-dependent pac-bayes priors via differential privacy," *arXiv preprint arXiv:1802.09583*, 2018.