Machine Unlearning for Gibbs Supervised Learning Algorithms

Yaiza Bermudez, Samir M. Perlaza, Iñaki Esnaola

Machine Unlearning for Gibbs Supervised Learning Algorithms

Yaiza Bermudez*, Samir M. Perlaza*†§, and Iñaki Esnaola‡§
Emails: name.lastname@inria.fr and esnaola@sheffield.ac.uk
*Centre Inria d'Université Côte d'Azur, INRIA, Sophia Antipolis, France.
†Laboratoire GAATI, Université de la Polynésie française, Fa'a'ā, French Polynesia.
‡School of Electrical and Electronic Engineering, University of Sheffield, Sheffield, United Kingdom.
§ECE Dept. Princeton University, Princeton, 08544 NJ, USA.

***Abstract*—In this paper, a method for achieving exact unlearning for Gibbs supervised learning algorithms is proposed using a variational formulation inspired by empirical risk minimization subject to relative entropy regularization (ERM-RER). Such a method consists of maximizing the expected empirical risk over the dataset to be unlearned subject to a regularization by relative entropy with respect to the original algorithm. The optimization variable is a probability measure on the models; and the solution is another Gibbs probability measure that represents a new Gibbs supervised learning algorithm. The method guarantees exact unlearning in the sense that the new Gibbs algorithm coincides in distribution with the algorithm that would have been obtained by retraining from scratch on the dataset to be retained. As a byproduct, a framework for reweighting data points in ERM-RER by strategically choosing both the reference measure and the regularization factor is obtained. In this framework, exact unlearning is the special case in which zero-weight is assigned to the contribution of the data points to be unlearned. More generally, depending on the choice of certain parameters, data points can be up-weighted or down-weighted in ERM-RER problems for particular purposes, e.g., controlling the generalization error of Gibbs algorithms. This paves the way for new constructive or adversarial views on classical reweighting data points in ERM-RER.**

I. Introduction

Modern learning systems are routinely trained on data that may later become unavailable or impermissible to use, e.g., due to user requests, contractual constraints, or regulatory requirements such as the right to erasure in Article 17 of the GDPR [1], also known as the "Right to be Forgotten". In this context, *machine unlearning* refers to the task of removing the influence of a part of the training dataset from a learning algorithm while avoiding the computational cost of retraining from scratch [2]–[4]. Early formulations of data deletion in learning formalize this task as producing an updated algorithm that is statistically consistent with training on the retained data only [5], [6]. A particularly stringent requirement is *exact unlearning*, in which the distribution of the post-unlearning algorithm is required to be identical to the algorithm obtained by retraining on the dataset to be retained [6]. These guarantees are difficult to obtain and practical approaches often trade computational efficiency for approximate removal [7]–[10], often referred to as *approximate unlearning* [11].

In this work, exact unlearning is formulated as a relative-entropy-regularized variational problem, which allows using existing tools previously developed in the framework of empirical risk minimization with f-divergence regularization [12]–[16]. The main contribution of this work is an exact unlearning method for Gibbs algorithms. More specifically, the method consists in implementing a Gibbs algorithm via the maximization of the empirical risk with respect to the dataset to be unlearned using as reference measure the Gibbs algorithm from which data must be unlearned. Interestingly, this method guarantees exact unlearning in the sense that the post-unlearning algorithm coincides with the solution to an ERM-RER problem with respect to the retained training dataset. The key observation is that exact unlearning is achieved thanks to the strategic choice of both the reference measure, as in [17]; and the regularization factor, as in [18]–[20]. This choice results in an effect that can be assimilated to a reweighting of the contribution of the data points to be unlearned. In particular, unlearning consists in assigning zero-weight to those data points. From this perspective, the results presented here can be placed in a more general scenario than unlearning. In particular, it can be placed in the scenario of reweighting data points in ERM-RER. More generally, this reweighting might be oriented to decrease the generalization error [21] of Gibbs algorithms by controlling the contribution of particular data points. On the other hand, reweighting data points in ERM-RER might be performed adversarially. In this case, the objective might be to increase the generalization error of Gibbs algorithms by up-weighting the outliers within the training dataset.

The paper is organized as follows. Section II introduces the notation and formalizes the supervised learning setting. Section III defines Gibbs conditional probability measures and their characterizations via ERM-RER optimization problems. Section IV introduces the main result, which consists in an exact unlearning method for Gibbs algorithms. Finally, Section V concludes the paper, and discusses open research paths.

This work is supported in part by the European Commission through the H2020-MSCA-RISE-2019 project 872172; the French National Agency for Research (ANR) through the Project ANR-21-CE25-0013 and the project ANR-22-PEFT-0010 of the France 2030 program PEPR Réseaux du Futur; and in part by the Agence de l'innovation de défense (AID) through the project UK-FR 2024352.

II. Supervised Machine Learning

Let $\mathcal{M}$, $\mathcal{X}$ and $\mathcal{Y}$, with $\mathcal{M} \subseteq \mathbb{R}^d$ and $d \in \mathbb{N}$, be sets of *models*, *patterns*, and *labels*, respectively. The training data available is assumed to be partitioned into K smaller datasets. For all $k \in \{1, 2, \ldots, K\}$, the k-th dataset consists of n_k data points $(x_{k,1}, y_{k,1}), \ldots, (x_{k,n_k}, y_{k,n_k})$, which are elements of the set $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$. Such data points form the dataset $z_k \in \mathcal{Z}^{n_k}$, which can be explicitly written as

$$z_k \triangleq \left((x_{k,1}, y_{k,1}), \ldots, (x_{k,n_k}, y_{k,n_k})\right) \in \mathcal{Z}^{n_k}. \tag{1}$$

The aggregated training dataset is then defined explicitly as,

$$z_0 \triangleq (z_1, \ldots, z_K) \in \mathcal{Z}^{n_1} \times \mathcal{Z}^{n_2} \times \cdots \times \mathcal{Z}^{n_K}, \tag{2}$$

with size $n_0 \triangleq \sum_{k=1}^{K} n_k$ data points. Given a model $\boldsymbol{\theta} \in \mathcal{M}$, the loss induced by such a model with respect to a data point $(x, y) \in \mathcal{Z}$ is $\ell(x, y, \boldsymbol{\theta})$, where the function

$$\ell : \mathcal{Z} \times \mathcal{M} \to [0, +\infty), \tag{3}$$

is referred to as the *loss function* and is assumed to be Borel measurable. Using such a loss function, the *empirical risk* induced by a model $\boldsymbol{\theta} \in \mathcal{M}$, with respect to a dataset $z_k \in \mathcal{Z}^{n_k}$, with $k \in \{0, 1, \ldots, K\}$, is determined by the function

$$\mathsf{L}_k : \begin{cases} \mathcal{Z}^{n_k} \times \mathcal{M} \longrightarrow [0, +\infty) \\ \quad (z_k, \boldsymbol{\theta}) \longmapsto \frac{1}{n_k} \sum_{i=1}^{n_k} \ell\left(x_{k,i}, y_{k,i}, \boldsymbol{\theta}\right). \end{cases} \tag{4}$$

The functions $\mathsf{L}_0, \mathsf{L}_1, \ldots, \mathsf{L}_K$ exhibit the following property.

Lemma 1. *Consider the training dataset z_0 in* (2)*, then the empirical risk functions $\mathsf{L}_0, \mathsf{L}_1, \ldots, \mathsf{L}_K$ in* (4) *satisfy for all $\boldsymbol{\theta} \in \mathcal{M}$,*

$$\mathsf{L}_0(z_0, \boldsymbol{\theta}) = \sum_{k=1}^{K} \frac{n_k}{n_0} \mathsf{L}_k(z_k, \boldsymbol{\theta}), \tag{5}$$

Proof: The proof is similar to the proof of [15, Lemma 4]. ■

The set of all probability measures on the measurable space $\left(\mathcal{M}_k, \mathscr{F}_{\mathcal{M}_k}\right)$ is denoted by $\triangle(\mathcal{M}_k)$. Using this notation, the relative entropy or KL-divergence is defined hereunder.

Definition 1 (Relative Entropy)**.** *Given a probability measure P and a σ-finite measure Q, both on the same measurable space, with P absolutely continuous with respect to Q. The relative entropy of P with respect to Q is*

$$D(P \parallel Q) \triangleq \int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}). \tag{6}$$

The set of all probability measures in $\mathcal{M}_k$ conditioned on an element of $\mathcal{Z}_k^{n_k}$ is denoted by $\triangle\left(\mathcal{M}_k \mid \mathcal{Z}_k^{n_k}\right)$. Moreover, the set of probability measures in $\triangle(\mathcal{M}_k)$ that are absolutely continuous with respect to Q_k is denoted by $\triangle_{Q_k}(\mathcal{M}_k)$. Using this notation, a supervised machine learning algorithm is represented by a conditional probability measure, as defined hereunder.

Definition 2 (Algorithm)**.** *A conditional probability measure $P_{\boldsymbol{\Theta}|\mathbf{Z}_k} \in \triangle\left(\mathcal{M} \mid (\mathcal{X} \times \mathcal{Y})^{n_k}\right)$ is said to represent a supervised machine learning algorithm. The instance of such an algorithm trained upon the dataset z_k in* (1) *is denoted by $P_{\boldsymbol{\Theta}|\mathbf{Z}_k = z_k} \in \triangle(\mathcal{M})$.*

A class of algorithms that are central in this work are known as Gibbs algorithms. These algorithms are introduced in the following section.

III. Gibbs Algorithms

Gibbs algorithms are represented by Gibbs conditional probability measures. Such conditional probability measures are parametrized by the empirical risk function L_k in (4); a σ-finite measure $Q \in \triangle(\mathcal{M})$; and a dataset $z_k \in \mathcal{Z}^{n_k}$, with $k \in \{0, 1, \ldots, K\}$. Using this notation, the definition of Gibbs conditional probability measures on the models is presented hereunder.

Definition 3. *Given the function L_k in* (4)*, with $k \in \{0, 1, \ldots, K\}$; a σ-finite measure $Q \in \triangle(\mathcal{M})$; and a $\lambda_k \in \mathbb{R} \setminus \{0\}$, the probability measure $P_{\boldsymbol{\Theta}|\mathbf{Z}_k}^{(Q,\lambda_k)} \in \triangle\left(\mathcal{M} \mid \mathcal{Z}^{n_k}\right)$ is said to be an $(\mathsf{L}_k, Q, \lambda_k)$-Gibbs conditional probability measure if*

$$\forall z_k \in \mathcal{Z}^{n_k}, \ \int \exp\left(-\frac{1}{\lambda_k} \mathsf{L}_k(z_k, \boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) < +\infty; \tag{7}$$

and for all $(z_k, \boldsymbol{\theta}) \in \mathcal{Z}^{n_k} \times \operatorname{supp} Q$,

$$\frac{\mathrm{d}P_{\boldsymbol{\Theta}|\mathbf{Z}_k = z_k}^{(Q,\lambda_k)}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{1}{\lambda_k}\mathsf{L}_k(z_k, \boldsymbol{\theta})\right)}{\int \exp\left(-\frac{1}{\lambda_k}\mathsf{L}_k(z_k, \boldsymbol{\nu})\right) \mathrm{d}Q(\boldsymbol{\nu})}. \tag{8}$$

Note that if $\lambda_k < 0$, the condition in (7) may fail. Therefore, throughout this work, the parameter λ_k is chosen so that (7) holds. Note that, while $P_{\boldsymbol{\Theta}|\mathbf{Z}_k}^{(Q,\lambda_k)}$ in (8) is referred to as a Gibbs conditional probability measure, the measure $P_{\boldsymbol{\Theta}|\mathbf{Z}_k = z_k}^{(Q,\lambda_k)}$, obtained by conditioning upon a given dataset $z_k \in \mathcal{Z}^{n_k}$, is referred to as a Gibbs probability measure. Consider the functional R_{z_k} defined as follows,

$$\mathsf{R}_{z_k} : \begin{cases} \triangle(\mathcal{M}) \longrightarrow [0, +\infty) \\ \quad P \longmapsto \int \mathsf{L}_k(z_k, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}), \end{cases} \tag{9}$$

where the function L_k is defined in (4). The Gibbs probability measure $P_{\boldsymbol{\Theta}|\mathbf{Z}_k = z_k}^{(Q,\lambda_k)}$ in (8) is related to the following optimization problem, under the assumption that $\lambda_k > 0$:

$$\min_{P \in \triangle_Q(\mathcal{M})} \mathsf{R}_{z_k}(P) + \lambda_k D(P \parallel Q). \tag{10}$$

Alternatively, when $\lambda_k < 0$, such a Gibbs probability measure is related to the following optimization problem:

$$\max_{P \in \triangle_Q(\mathcal{M})} \mathsf{R}_{z_k}(P) + \lambda_k D(P \parallel Q), \tag{11}$$

where the functional R_{z_k} is defined in (9). The importance of these optimization problems stems from the fact that the unlearning problem can be posed as the optimization problem in (11). The following lemma shows the connections between the Gibbs probability measure $P_{\boldsymbol{\Theta}|\mathbf{Z}_k = z_k}^{(Q,\lambda_k)}$ in (8) and the optimization problems in (10) and (11).

Lemma 2. *The $(\mathsf{L}_k, Q, \lambda_k)$-Gibbs probability measure $P_{\boldsymbol{\Theta}|\mathbf{Z}_k = z_k}^{(Q,\lambda_k)}$ in* (8) *is the unique solution to* (10) *(respectively, to* (11)*), if $\lambda_k > 0$ (respectively, if $\lambda_k < 0$).*

Proof: The proof is immediate from [22, Lemma 1]. ∎

Two optimization problems that are also closely related to the probability measure $P^{(Q,\lambda_k)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ in (8) are the following. When $\lambda_k > 0$, the optimization problem is:

$$\min_{P\in\triangle_Q(\mathcal{M})} \quad \mathsf{R}_{\boldsymbol{z}_k}(P) \tag{12a}$$

$$\text{s.t.} \quad D(P \parallel Q) \leqslant \gamma_k, \tag{12b}$$

for some $\gamma_k > 0$. Alternatively, when $\lambda_k < 0$, the optimization problem is:

$$\max_{P\in\triangle_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}_k}(P) \tag{13a}$$

$$\text{s.t. } D(P \parallel Q) \leqslant \gamma_k. \tag{13b}$$

The following lemma formalizes these observations.

Lemma 3. *Consider the* $(\mathsf{L}_k, Q, \lambda_k)$*-Gibbs probability measure* $P^{(Q,\lambda_k)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ *in* (8) *and assume that* $\lambda_k > 0$ *(respectively,* $\lambda_k < 0$*) is such that*

$$D\left(P^{(Q,\lambda_k)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_k=\boldsymbol{z}_k} \,\middle\|\, Q\right) = \gamma_k, \tag{14}$$

with γ_k *in* (12) *(respectively, in* (13)*). Then, the measure* $P^{(Q,\lambda_k)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ *is the unique solution to* (12) *(respectively, to* (13)*).*

Proof: The proof follows from [22, Lemma 4]. ∎

Lemma 3 implies that the probability measure $P^{(Q,\lambda_k)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ in (8) is the one that minimizes (respectively, maximizes) the training empirical risk over all probability measures in the following neighborhood of Q,

$$\left\{P \in \triangle_Q(\mathcal{M}) : D(P \parallel Q) \leqslant D\left(P^{(Q,\lambda_k)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_k=\boldsymbol{z}_k} \,\middle\|\, Q\right)\right\}. \tag{15}$$

The relevance of Gibbs algorithms, in particular when $\lambda_k > 0$, stems from the observation that the probability measure $P^{(Q,\lambda_k)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ in (8) is the long-run distribution of a stochastic gradient descent algorithm, as highlighted in [23].

IV. Main result

The main result of this work is presented using the following $(\mathsf{L}_0, Q, \lambda_0)$-Gibbs probability measure

$$P^{(Q,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0} \in \triangle(\mathcal{M}), \tag{16a}$$

with $\lambda_0 > 0$, which represents a Gibbs algorithm trained upon the aggregated dataset $\boldsymbol{z}_0$ in (2). That is, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{1}{\lambda_0}\mathsf{L}_0(\boldsymbol{z}_0,\boldsymbol{\theta})\right)}{\int \exp\left(-\frac{1}{\lambda_0}\mathsf{L}_0(\boldsymbol{z}_0,\boldsymbol{\nu})\right)\mathrm{d}Q(\boldsymbol{\nu})}. \tag{16b}$$

From Lemma 3, it follows that such an algorithm is optimal in the sense that the corresponding probability measure $P^{(Q,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}$ in (16) minimizes the expected training empirical risk with respect to $\boldsymbol{z}_0$, over all probability measures in the following neighborhood of Q,

$$\mathcal{G}_0 \triangleq \left\{P \in \triangle_Q(\mathcal{M}) : D(P \parallel Q) \leqslant D\left(P^{(Q,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0} \,\middle\|\, Q\right)\right\}. \tag{17}$$

Consider that the dataset $\boldsymbol{z}_0$ in (2) is partitioned into two smaller datasets $\boldsymbol{z}_1 \in \mathcal{Z}^{n_1}$ and $\boldsymbol{z}_2 \in \mathcal{Z}^{n_2}$, *i.e.* $K = 2$. The objective is to transform the $(\mathsf{L}_0, Q, \lambda_0)$-Gibbs probability measure $P^{(Q,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}$ in (16) into an $(\mathsf{L}_2, Q, \lambda_2)$-Gibbs probability measure

$$P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_2=\boldsymbol{z}_2} \in \triangle(\mathcal{M}), \tag{18a}$$

for some $\lambda_2 > 0$, which represents a Gibbs algorithm exclusively trained upon dataset $\boldsymbol{z}_2$. That is, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_2=\boldsymbol{z}_2}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{1}{\lambda_2}\mathsf{L}_2(\boldsymbol{z}_2,\boldsymbol{\theta})\right)}{\int \exp\left(-\frac{1}{\lambda_2}\mathsf{L}_2(\boldsymbol{z}_2,\boldsymbol{\nu})\right)\mathrm{d}Q(\boldsymbol{\nu})}. \tag{18b}$$

From Lemma 3, it follows that the Gibbs algorithm represented by the measure $P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_2=\boldsymbol{z}_2}$ in (18) minimizes the expected empirical risk with respect to $\boldsymbol{z}_2$ over all probability measures in the following neighborhood of Q,

$$\mathcal{G}_1 \triangleq \left\{P \in \triangle_Q(\mathcal{M}) : D(P \parallel Q) \leqslant D\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_2=\boldsymbol{z}_2} \,\middle\|\, Q\right)\right\}. \tag{19}$$

In a nutshell, the objective is to unlearn the dataset $\boldsymbol{z}_1$ from the Gibbs algorithm represented by the probability measure $P^{(Q,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}$ in (16), while retaining the dataset $\boldsymbol{z}_2$. This operation is often referred to as exact unlearning and is defined for this particular case as follows:

Definition 4 (Exact unlearning)**.** *Consider the algorithm represented by the* $(\mathsf{L}_0, Q, \lambda_0)$*-Gibbs probability measure* $P^{(Q,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}$ *in* (16) *and the datasets* $\boldsymbol{z}_1$ *and* $\boldsymbol{z}_2$ *to be unlearned and retained, respectively, with* $\boldsymbol{z}_0$*,* $\boldsymbol{z}_1$ *and* $\boldsymbol{z}_2$ *in* (2)*. An algorithm* $P \in \triangle_Q(\mathcal{M})$ *is said to achieve exact unlearning of* $\boldsymbol{z}_1$ *from* $P^{(Q,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}$*, if for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*,*

$$\frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_2=\boldsymbol{z}_2}}(\boldsymbol{\theta}) = 1, \tag{20}$$

where the probability measure $P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_2=\boldsymbol{z}_2}$ *is defined in* (18)*, for some* $\lambda_2 > 0$*.*

The equality in implies that the measures P and $P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_2=\boldsymbol{z}_2}$ are identical. See [24, Theorem 3].

The main result of this work consists in showing that exact unlearning of $\boldsymbol{z}_1$ from $P^{(Q,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}$ in (16) is achieved by solving the optimization problem

$$\max_{P\in\triangle_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}_1}(P) + \lambda_1 D\left(P \,\middle\|\, P^{(Q,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}\right), \tag{21}$$

for some $\lambda_1 < 0$, where the functional $\mathsf{R}_{\boldsymbol{z}_1}$ is defined in (9). From Lemma 2, it follows that the unique solution to (21) is an $(\mathsf{L}_1, P^{(Q,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}, \lambda_1)$-Gibbs probability measure

$$P^{\left(P^{(Q,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0},\lambda_1\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_1=\boldsymbol{z}_1} \in \triangle(\mathcal{M}). \tag{22}$$

The following theorem, which is the main result of this work, shows that a specific choice of λ_1, and λ_2 makes the $(\mathsf{L}_1, P^{(Q,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}, \lambda_1)$-Gibbs probability measure in (22) identical to the probability measure $P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}_2=\boldsymbol{z}_2}$ in (18), which implies exact unlearning (Definition 4).

Theorem 4. *Assume that*

$$\lambda_1 = -\frac{n_0}{n_1}\lambda_0 \text{ and } \lambda_2 = \frac{n_0}{n_2}\lambda_0, \tag{23}$$

for some $\lambda_0 > 0$. *Then, the* $(\mathsf{L}_2, Q, \lambda_2)$*-Gibbs probability measure* $P^{(Q,\lambda_2)}_{\mathbf{\Theta}|\mathbf{Z}_2=\mathbf{z}_2}$ *in* (18) *and the* $(\mathsf{L}_1, P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0}, \lambda_1)$*-Gibbs probability measure* $P^{\left(P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0},\lambda_1\right)}_{\mathbf{\Theta}|\mathbf{Z}_1=\mathbf{z}_1}$ *in* (22) *satisfy for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{\left(P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0},\lambda_1\right)}_{\mathbf{\Theta}|\mathbf{Z}_1=\mathbf{z}_1}}{\mathrm{d}P^{(Q,\lambda_2)}_{\mathbf{\Theta}|\mathbf{Z}_2=\mathbf{z}_2}}(\boldsymbol{\theta}) = 1. \tag{24}$$

Proof: The proof is presented in Section IV-A. ∎

Theorem 4 shows that the probability measure $P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0}$ in (16) can be transformed into $P^{(Q,\lambda_2)}_{\mathbf{\Theta}|\mathbf{Z}_2=\mathbf{z}_2}$ in (18) via the optimization problem in (21), achieving exact unlearning in the sense of Definition 4. An important observation is that only the dataset to be unlearned, i.e., dataset $\mathbf{z}_1$, is used in the optimization problem in (21). The dependence of such an optimization problem on the dataset to be retained, i.e., $\mathbf{z}_2$, is via the reference measure $P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0}$, which represents the Gibbs algorithm from which the dataset $\mathbf{z}_1$ must be unlearned. Hence, the optimization problem in (21) implies that unlearning the dataset $\mathbf{z}_1$ from the Gibbs algorithm $P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0}$ is achieved through training a new Gibbs algorithm exclusively upon the dataset to be unlearned, i.e., dataset $\mathbf{z}_1$, while using as reference measure the Gibbs algorithm $P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0}$ itself. Interestingly, this establishes a proof of exact unlearning as, under the scaling in (23), the $(\mathsf{L}_1, P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0}, \lambda_1)$-Gibbs probability measure $P^{\left(P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0},\lambda_1\right)}_{\mathbf{\Theta}|\mathbf{Z}_1=\mathbf{z}_1}$ in (22) coincides with the Gibbs measure that would have been obtained by training only on the retained dataset $\mathbf{z}_2$, i.e., the measure $P^{(Q,\lambda_2)}_{\mathbf{\Theta}|\mathbf{Z}_2=\mathbf{z}_2}$ in (18).

Another consequence of Theorem 4, and the uniqueness of the solutions to (10) and (21) implied by Lemma 2, is that

$$\begin{aligned}&\operatorname*{arg\,max}_{P\in\triangle_Q(\mathcal{M})} \mathsf{R}_{\mathbf{z}_1}(P) - \frac{n_0}{n_1}\lambda_0 D\left(P \,\middle\|\, P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0}\right)\\ &= \operatorname*{arg\,min}_{P\in\triangle_Q(\mathcal{M})} \mathsf{R}_{\mathbf{z}_2}(P) + \frac{n_0}{n_2}\lambda_0 D(P \,\|\, Q).\end{aligned} \tag{25}$$

From the perspective of Lemma 3, the equality in (25) implies that the probability measure in $\triangle_Q(\mathcal{M})$ that minimizes the expected empirical risk with respect to $\mathbf{z}_2$ (dataset to be retained) over all measures in the set $\mathcal{G}_1$ in (19) is identical to the measure that maximizes the expected empirical risk with respect to $\mathbf{z}_1$ (dataset to be unlearned) over all measures in the set

$$\mathcal{G}_2 \triangleq \left\{ P \in \triangle_Q(\mathcal{M}) : D\left(P \,\middle\|\, P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0}\right) \leqslant D\left(P^{\left(P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0},-\frac{n_0}{n_1}\lambda_0\right)}_{\mathbf{\Theta}|\mathbf{Z}_1=\mathbf{z}_1} \,\middle\|\, P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0}\right)\right\}. \tag{26}$$

Figure 1 provides a representation of (25) under the assumption in (23): it depicts how the common optimizer in (25)

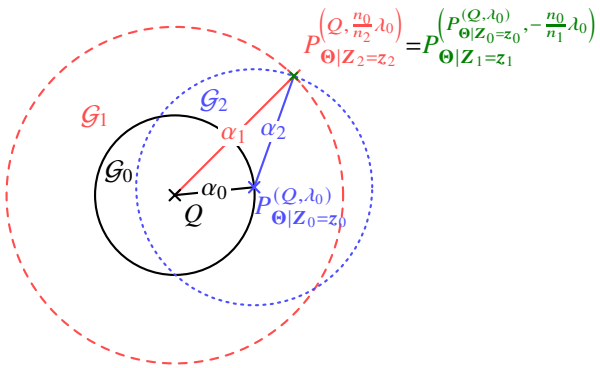

Figure 1. Representation, under the assumption in (23), of the set $\mathcal{G}_0$ in (17) as a solid black circle; the set $\mathcal{G}_1$ in (19) as a dashed red circle; and the set $\mathcal{G}_2$ in (26) as a dotted blue circle. The involved probability measures are defined in (16), (18), and (22). The definitions of α_0, α_1, and α_2 are in (27), (28) and (29), respectively.

can be equivalently characterized either as an empirical-risk minimizer over $\mathcal{G}_1$ in (19) (relative to $\mathbf{z}_2$) or as an empirical-risk maximizer over $\mathcal{G}_2$ in (26) (relative to $\mathbf{z}_1$), while also highlighting the relation with $\mathcal{G}_0$ in (17). In particular, the figure shows the sets $\mathcal{G}_0$, $\mathcal{G}_1$, and $\mathcal{G}_2$ together with the probability measures $P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0}$ in (16), $P^{(Q,\lambda_2)}_{\mathbf{\Theta}|\mathbf{Z}_2=\mathbf{z}_2}$ in (18), and $P^{\left(P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0},\lambda_1\right)}_{\mathbf{\Theta}|\mathbf{Z}_1=\mathbf{z}_1}$ in (22). In such a figure, the following notation is used:

$$\alpha_0 \triangleq D\left(P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0} \,\middle\|\, Q\right); \tag{27}$$

$$\alpha_1 \triangleq D\left(P^{\left(Q,\frac{n_0}{n_2}\lambda_0\right)}_{\mathbf{\Theta}|\mathbf{Z}_2=\mathbf{z}_2} \,\middle\|\, Q\right); \text{ and} \tag{28}$$

$$\alpha_2 \triangleq D\left(P^{\left(P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0},-\frac{n_0}{n_1}\lambda_0\right)}_{\mathbf{\Theta}|\mathbf{Z}_1=\mathbf{z}_1} \,\middle\|\, P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0}\right), \tag{29}$$

under the assumption that $\alpha_0 < \alpha_1$.

Finally, note that the parameter $\lambda_0 > 0$ remains a free parameter that can be further optimized, e.g., to minimize the generalization error of the resulting Gibbs algorithm $P^{\left(P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0},-\frac{n_0}{n_1}\lambda_0\right)}_{\mathbf{\Theta}|\mathbf{Z}_1=\mathbf{z}_1}$. See for instance [21], [25]–[28] and references therein for this purpose.

A. Proof of Theorem 4

The proof of Theorem 4 relies on the following result.

Theorem 5. *The* $(\mathsf{L}_1, P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0}, \lambda_1)$*-Gibbs probability measure* $P^{\left(P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0},\lambda_1\right)}_{\mathbf{\Theta}|\mathbf{Z}_1=\mathbf{z}_1}$ *in* (22), *with* $\lambda_0 > 0$ *and* $\lambda_1 \in \mathbb{R}\setminus\{0\}$, *satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\begin{aligned}&\frac{\mathrm{d}P^{\left(P^{(Q,\lambda_0)}_{\mathbf{\Theta}|\mathbf{Z}_0=\mathbf{z}_0},\lambda_1\right)}_{\mathbf{\Theta}|\mathbf{Z}_1=\mathbf{z}_1}}{\mathrm{d}Q}(\boldsymbol{\theta})\\ &= \frac{\exp\left(\frac{-1}{\lambda_0}\left(\left(\frac{\lambda_0}{\lambda_1}+\frac{n_1}{n_0}\right)\mathsf{L}_1(\mathbf{z}_1,\boldsymbol{\theta})+\frac{n_2}{n_0}\mathsf{L}_2(\mathbf{z}_2,\boldsymbol{\theta})\right)\right)}{\int \exp\left(\frac{-1}{\lambda_0}\left(\left(\frac{\lambda_0}{\lambda_1}+\frac{n_1}{n_0}\right)\mathsf{L}_1(\mathbf{z}_1,\boldsymbol{\nu})+\frac{n_2}{n_0}\mathsf{L}_2(\mathbf{z}_2,\boldsymbol{\nu})\right)\right)\mathrm{d}Q(\boldsymbol{\nu})},\end{aligned} \tag{30}$$

where the function L_2 *is defined in* (4).

Proof: The proof is presented in [29]. ∎

Theorem 5 makes explicit the role of λ_1 as a reweighting parameter of the contribution of the dataset z_1. The exact unlearning identity in Theorem 4 follows by choosing λ_1 to completely eliminate the contribution of dataset z_1. That is, λ_1 is chosen such that

$$\frac{\lambda_0}{\lambda_1} + \frac{n_1}{n_0} = 0. \tag{31}$$

More specifically, using (18b) and (31) in (30) yields,

$$\frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}^{\left(Q,\lambda_0\right),-\frac{n_0}{n_1}\lambda_0\right)}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}_2=\boldsymbol{z}_2}^{\left(Q,\frac{n_0}{n_2}\lambda_0\right)}}{\mathrm{d}Q}(\boldsymbol{\theta}). \tag{32}$$

From [12, Lemma 3] and (32), it follows that $Q \ll P_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}^{(Q,\lambda_0)} \ll P_{\boldsymbol{\Theta}|\boldsymbol{Z}_1=\boldsymbol{z}_1}^{\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}^{(Q,\lambda_0)},-\frac{n_0}{n_1}\lambda_0\right)} \ll P_{\boldsymbol{\Theta}|\boldsymbol{Z}_2=\boldsymbol{z}_2}^{\left(Q,\frac{n_0}{n_2}\lambda_0\right)} \ll Q$. Hence, from (32), it follows that

$$1 = \frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}_1=\boldsymbol{z}_1}^{\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}^{(Q,\lambda_0)},-\frac{n_0}{n_1}\lambda_0\right)}}{\mathrm{d}Q}(\boldsymbol{\theta}) \frac{\mathrm{d}Q}{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}_2=\boldsymbol{z}_2}^{\left(Q,\frac{n_0}{n_2}\lambda_0\right)}}(\boldsymbol{\theta}) \tag{33}$$

$$= \frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}_1=\boldsymbol{z}_1}^{\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}^{(Q,\lambda_0)},-\frac{n_0}{n_1}\lambda_0\right)}}{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}_2=\boldsymbol{z}_2}^{\left(Q,\frac{n_0}{n_2}\lambda_0\right)}}(\boldsymbol{\theta}), \tag{34}$$

where the equality in (33) follows from [24, Theorem 5]; and the equality in (34) follows from [24, Theorem 4]. This completes the proof of Theorem 4.

B. Reweighting Data Points in ERM-RER

Beyond the specific choices of λ_1 and λ_2 in (23), in general if $\lambda_0 > 0$ and $\lambda_1 \in \mathbb{R} \setminus \{0\}$, Theorem 5 unveils the following observation:

$$\begin{aligned}
&\operatorname*{arg\,min}_{P\in\triangle_Q(\mathcal{M})} \left(\frac{\lambda_0}{\lambda_1} + \frac{n_1}{n_0}\right) \mathsf{R}_{\boldsymbol{z}_1}(P) + \frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}(P) + \lambda_0 D\left(P \parallel Q\right) \\
&= \operatorname*{arg\,min}_{P\in\triangle_Q(\mathcal{M})} \frac{\lambda_0}{\lambda_1}\mathsf{R}_{\boldsymbol{z}_1}(P) + \mathsf{R}_{\boldsymbol{z}_0}(P) + \lambda_0 D\left(P \parallel Q\right) && (35) \\
&= \begin{cases} \operatorname*{arg\,min}_{P\in\triangle_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}_1}(P) + \lambda_1 D\left(P \,\middle\|\, P_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}^{(Q,\lambda_0)}\right) & \text{if } \lambda_1 > 0; \\ \operatorname*{arg\,max}_{P\in\triangle_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}_1}(P) + \lambda_1 D\left(P \,\middle\|\, P_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}^{(Q,\lambda_0)}\right) & \text{if } \lambda_1 < 0 \end{cases} && (36) \\
&= \left\{P_{\boldsymbol{\Theta}|\boldsymbol{Z}_1=\boldsymbol{z}_1}^{\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}^{(Q,\lambda_0)},\lambda_1\right)}\right\}, && (37)
\end{aligned}$$

where the equality in (35) follows from the fact that for all $P \in \triangle_Q(\mathcal{M})$,

$$\mathsf{R}_{\boldsymbol{z}_0}(P) = \frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}(P) + \frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}(P); \tag{38}$$

the equality in (36) follows from Theorem 5 and Lemma 2; and the equality in (37) follows from Lemma 2. The equality in (35) highlights that both λ_0 and λ_1 reweight the dataset z_1 in the optimization problem whose solution is the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}_1=\boldsymbol{z}_1}^{\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}^{(Q,\lambda_0)},\lambda_1\right)}$ in (22), relative to the optimization problem whose solution is $P_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}^{(Q,\lambda_0)}$ in (16). More specifically, from Lemma 2, it holds that

$$\begin{aligned}
\left\{P_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}^{(Q,\lambda_0)}\right\} = &\operatorname*{arg\,min}_{P\in\triangle_Q(\mathcal{M})} \left(\frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}(P) + \frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}(P)\right) \\
&+\lambda_0 D\left(P \parallel Q\right),
\end{aligned} \tag{39}$$

and from (37), it holds that

$$\begin{aligned}
\left\{P_{\boldsymbol{\Theta}|\boldsymbol{Z}_1=\boldsymbol{z}_1}^{\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}_0=\boldsymbol{z}_0}^{(Q,\lambda_0)},\lambda_1\right)}\right\} = &\operatorname*{arg\,min}_{P\in\triangle_Q(\mathcal{M})} \left(\frac{\lambda_0}{\lambda_1} + \frac{n_1}{n_0}\right)\mathsf{R}_{\boldsymbol{z}_1}(P) + \frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}(P) \\
&+\lambda_0 D\left(P \parallel Q\right).
\end{aligned} \tag{40}$$

The contribution of the dataset z_1 in (39) is proportional to $\frac{n_1}{n_0}$, while in (40), it is proportional to $\frac{\lambda_0}{\lambda_1} + \frac{n_1}{n_0}$. Interestingly, when $\lambda_1 > 0$, the contribution of the dataset z_1 is up-weighted in (40). That is, $\frac{\lambda_0}{\lambda_1} + \frac{n_1}{n_0} > \frac{n_1}{n_0}$. The converse holds when $\lambda_1 < 0$. It is important to highlight that the contribution of the dataset z_2 remains invariant in both (39) and (40).

V. Conclusions and Final Remarks

In this work, an exact unlearning method for Gibbs algorithms has been proposed in Theorem 4. The method consists in implementing a new Gibbs algorithm via the maximization of the empirical risk with respect to the dataset to be unlearned using as reference measure the algorithm from which data must be unlearned. Under the corresponding scaling, this construction guarantees exact unlearning in the sense that the new Gibbs algorithm coincides with the Gibbs algorithm that would have been obtained by retraining from scratch exclusively on the dataset to be retained. An important feature is that the corresponding optimization problem depends exclusively on the dataset to be unlearned, while the dependence on the dataset to be retained is captured solely through the reference measure. The proof of Theorem 4 relies on Theorem 5, which can be placed in a more general scenario than the problem of unlearning. In particular, it can be placed in the scenario of reweighting data points in ERM-RER [12]. From this perspective, unlearning appears as a special case in which zero weight is assigned to the contribution of the data points to be unlearned. More broadly, this reweighting offers a principled mechanism to modulate the contribution of particular data points, which can be leveraged to improve the generalization error by attenuating the effect of atypical or overly influential data points. Conversely, the same mechanism can be used adversarially: by up-weighting outliers, an attacker may deliberately increase the generalization error of Gibbs algorithms, yielding a model-poisoning strategy within the ERM-RER learning framework. The strategic choice of the parameters in Theorem 5, beyond unlearning, is left out of the scope of this paper. Nonetheless, the theorem provides a unifying theoretical foundation for exact unlearning and data points reweighting in ERM-RER. This opens the door to developing both robustness-oriented reweighting and its adversarial counterpart in statistical machine learning.

REFERENCES

[1] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 (General Data Protection Regulation)," Official Journal of the European Union, OJ L 119, 4.5.2016, pp. 1–88, 2016, art. 17.

[2] T. T. Nguyen, T. T. Huynh, Z. Ren, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen, "A survey of machine unlearning," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 5, pp. 1–46, Oct. 2025.

[3] S. Sai, U. Mittal, V. Chamola, K. Huang, I. Spinelli, S. Scardapane, Z. Tan, and A. Hussain, "Machine un-learning: An overview of techniques, applications, and future directions," *Cognitive Computation*, vol. 16, no. 2, pp. 482–506, Mar. 2024.

[4] J. Xu, Z. Wu, C. Wang, and X. Jia, "Machine unlearning: Solutions and challenges," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 3, pp. 2150–2168, Jun. 2024.

[5] A. Ginart, M. Guan, G. Valiant, and J. Y. Zou, "Making AI forget you: Data deletion in machine learning," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 32. Vancouver, Canada: Curran Associates, Inc., 2019.

[6] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *Proceedings of the IEEE Symposium on Security and Privacy*, San Jose, CA, USA, May 2015, pp. 463–480.

[7] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, May 2021, pp. 141–159.

[8] M. Egger, R. Bitar, and R. L. Urbanke, "Efficient machine unlearning by model splitting and core sample selection," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, Sydney, NSW, Australia, Sep. 2025, pp. 1–6.

[9] Y. Jiang, C.-W. Tan, and K.-Y. Lam, "Feduhb: Accelerating federated unlearning via Polyak heavy ball method," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, Shenzhen, China, Nov. 2024, pp. 235–240.

[10] Y. Jiang, X. Tong, Z. Liu, X. Zhang, K.-Y. Lam, and C.-W. Tan, "Certifying the right to be forgotten: Primal–dual optimization for sample and label unlearning in vertical federated learning," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 13 143–13 158, Nov. 2025.

[11] C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten, "Certified data removal from machine learning models," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, vol. 119, Virtual Event, Jul. 2020, pp. 3832–3842.

[12] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with relative entropy regularization," *IEEE Transactions on Information Theory*, vol. 70, no. 7, pp. 5122 – 5161, Jul. 2024.

[13] X. Zou, S. M. Perlaza, I. Esnaola, and E. Altman, "Generalization analysis of machine learning algorithms via the worst-case data-generating probability measure," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 15, Vancouver, Canada, Feb. 2024, pp. 17 271–17 279.

[14] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Asymmetry of the relative entropy in the regularization of empirical risk minimization," *IEEE Transactions on Information Theory*, vol. 71, no. 8, pp. 6198–6226, Aug. 2025.

[15] S. M. Perlaza, I. Esnaola, G. Bisson, and H. V. Poor, "On the validation of Gibbs algorithms: Training datasets, test datasets and their aggregation," in *Proceedings of the International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, Jun. 2023, pp. 328–333.

[16] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Equivalence of empirical risk minimization to regularization on the family of f-divergences," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Athens, Greece, Jul. 2024, pp. 759–764.

[17] Y. Bermudez, S. M. Perlaza, and I. Esnaola, "Decentralized machine learning with centralized performance guarantees via Gibbs algorithms," in *Proceedings of the International Symposium on Information Theory (ISIT)*, Guangzhou, China, Jun. 2026.

[18] Y. Bu, "Towards optimal inverse temperature in the Gibbs algorithm," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Athens, Greece, Jul. 2024, pp. 2257–2262.

[19] M. A. Medina, J. L. M. Olea, C. Rush, and A. Velez, "On the robustness to misspecification of α-posteriors and their variational approximations," *Journal of Machine Learning Research*, vol. 23, no. 147, pp. 1–51, 2022.

[20] R. Ray, M. A. Medina, and C. Rush, "Asymptotics for power posterior mean estimation," in *Proceedings of the 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Monticello, IL, USA, Sep. 2023.

[21] S. M. Perlaza and X. Zou, "The generalization error of supervised machine learning algorithms," *preprint arXiv:2411.12030*, 2024.

[22] S. M. Perlaza and G. Bisson, "Variations on the expectation due to changes in the probability measure," *Entropy*, vol. 27, no. 8:865, pp. 1–20, Aug. 2025.

[23] W. Azizian, F. Lutzeler, J. Malick, and P. Mertikopoulos, "What is the long-run distribution of stochastic gradient descent? A large deviations analysis," in *Proceedings of the International Conference on Machine Learning (ICML)*, Vienna, Austria, Jul. 2024, pp. 2168 – 2229.

[24] Y. Bermudez, G. Bisson, I. Esnaola, and S. M. Perlaza, "Proofs for folklore theorems on the Radon-Nikodym derivative," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9591, Jul. 2025.

[25] G. Aminian, Y. Bu, L. Toni, M. R. D. Rodrigues, and G. W. Wornell, "Information-theoretic characterizations of generalization error for the Gibbs algorithm," *IEEE Transactions on Information Theory*, vol. 70, no. 1, pp. 632–655, 2024.

[26] X. Zou, S. M. Perlaza, I. Esnaola, E. Altman, and H. V. Poor, "The worst-case data-generating probability measure in statistical learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 5, p. 175 – 189, Apr. 2024.

[27] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, "An exact characterization of the generalization error for the Gibbs algorithm," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 34, Virtual Event, Dec. 2021, pp. 8106–8118.

[28] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 2521–2530.

[29] Y. Bermudez, S. M. Perlaza, and I. Esnaola, "Machine unlearning for Gibbs supervised learning algorithms," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9610, Jan. 2026.