Machine Unlearning with Variational Inference

Instructor: Prof. Piyush Rai

Ali Faraz (17807078) Jayant Ranwka (170323)
Shashank Kumar(170642) Vishweshwar Tyagi (191173)

Indian Institute of Technology Kanpur

May 20, 2021

Introduction

Problem overview

- Partition train-dataset $D = D_e \cup D_r$
- D_e - dataset to be *forgotten/deleted*
- D_r - remaining dataset
- *Task* - To unlearn model parameters with D_e without having to re-train them on D_r

Why is Machine Unlearning necessary?

- Parts of data can get corrupted
- User may exercise their *right to be forgotten*
- Retraining can be pretty cost-inefficient

Exact Bayesian Unlearning

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \mathcal{D}_r) &= \frac{p(\mathcal{D}_r \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathcal{D}_r)} \\
&= \frac{p(\mathcal{D}_r \mid \boldsymbol{\theta})\, p(\mathcal{D}_e \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathcal{D}_r)\, p(\mathcal{D}_e \mid \boldsymbol{\theta})} \\
&= \frac{p(\mathcal{D}, \boldsymbol{\theta})}{p(\mathcal{D}_r)\, p(\mathcal{D}_e \mid \boldsymbol{\theta})} \\
&= \frac{p(\boldsymbol{\theta} \mid \mathcal{D})\, p(\mathcal{D})}{p(\mathcal{D}_r)\, p(\mathcal{D}_e \mid \boldsymbol{\theta})} \\
&= \frac{p(\boldsymbol{\theta} \mid \mathcal{D})\, p(\mathcal{D}_e, \mathcal{D}_r)}{p(\mathcal{D}_r)\, p(\mathcal{D}_e \mid \boldsymbol{\theta})} \\
&= \frac{p(\boldsymbol{\theta} \mid \mathcal{D})\, p(\mathcal{D}_e \mid \mathcal{D}_r)}{p(\mathcal{D}_e \mid \boldsymbol{\theta})} \propto \frac{p(\boldsymbol{\theta} \mid \mathcal{D})}{p(\mathcal{D}_e \mid \boldsymbol{\theta})}
\end{aligned}
$$

Approximate Bayesian Unlearning with Exact Posterior Belief

- Need conjugate prior for $p\,(\boldsymbol{\theta} \mid \mathcal{D}_r)$, otherwise approximate it by $q_u(\boldsymbol{\theta} \mid \mathcal{D}_r)$
- Leads to *approximate Bayesian unlearning with exact posterior* $p(\boldsymbol{\theta} \mid \mathcal{D})$

- **Loss function**: $KL[q_u(y \mid \mathcal{D}_r)\;||p(y \mid \mathcal{D}_r)]$ **[hard]**
- Can show that

$$KL[q_u(y \mid \mathcal{D}_r)\;||p(y \mid \mathcal{D}_r)] \leq KL[q_u(\boldsymbol{\theta} \mid \mathcal{D}_r)\;||p(\boldsymbol{\theta} \mid \mathcal{D}_r)]$$

- Can minimize $KL[q_u(\boldsymbol{\theta} \mid \mathcal{D}_r)\;||p(\boldsymbol{\theta} \mid \mathcal{D}_r)]$ using *Evidence Upper Bound (EUBO)*

EUBO

Evidence Upper Bound (EUBO) $\mathcal{U}$

$$\mathcal{U} \triangleq \int q_u(\boldsymbol{\theta} \mid \mathcal{D}_r) \log p(\mathcal{D}_e \mid \boldsymbol{\theta})\, d\boldsymbol{\theta} + KL[q_u(\boldsymbol{\theta} \mid \mathcal{D}_r) \| p(\boldsymbol{\theta} \mid \mathcal{D})]$$

- The first term ensures that the likelihood of the erased data is low while the second term ensures that the new posterior is close to the original posterior to prevent catastrophic unlearning
- Minimising $KL[q_u(\boldsymbol{\theta} \mid \mathcal{D}_r) \| p(\boldsymbol{\theta} \mid \mathcal{D}_r)]$ using VI is equivalent to minimising EUBO since it can be simplified to $\mathcal{U} = \log p(\mathcal{D}_e \mid \mathcal{D}_r) + KL[q_u(\boldsymbol{\theta} \mid \mathcal{D}_r) \| p(\boldsymbol{\theta} \mid \mathcal{D}_r)]$.

Unlearning with Approximate Posterior Belief

- When exact $p(\boldsymbol{\theta} \mid \mathcal{D})$ is not known, we use its approximation $q(\boldsymbol{\theta} \mid \mathcal{D})$
- Then find $\tilde{q}_u(\boldsymbol{\theta} \mid \mathcal{D}_r)$ that minimizes $KL[\tilde{q}_u(\boldsymbol{\theta} \mid \mathcal{D}_r) \| \tilde{p}(\boldsymbol{\theta} \mid \mathcal{D}_r)]$
- Here,

$$\begin{aligned}\tilde{p}(\boldsymbol{\theta} \mid \mathcal{D}_r) &= \frac{q(\boldsymbol{\theta} \mid \mathcal{D})\, p(\mathcal{D}_e \mid \mathcal{D}_r)}{p(\mathcal{D}_e \mid \boldsymbol{\theta})} \\ &\propto \frac{q(\boldsymbol{\theta} \mid \mathcal{D})}{p(\mathcal{D}_e \mid \boldsymbol{\theta})}\end{aligned}$$

- Hence, the expression of EUBO changes accordingly.

$$\tilde{U} = \mathop{\mathbb{E}}_{\tilde{q}_u(\boldsymbol{\theta} \mid \mathcal{D}_r)}[\log p(\mathcal{D}_e \mid \boldsymbol{\theta})] + KL[\tilde{q}_u(\boldsymbol{\theta} \mid \mathcal{D}_r) \| q(\boldsymbol{\theta} \mid \mathcal{D})]$$

EUBO with adjusted likelihood

- $q(\boldsymbol{\theta} \mid \mathcal{D})$ can differ largely from $p(\boldsymbol{\theta} \mid \mathcal{D})$ where $q(\boldsymbol{\theta} \mid \mathcal{D}) \simeq 0$
- **Remedy** - Use *adjusted* likelihood of the erased data $\left(\lambda \in [0, 1]\right)$

$$p_{adj}\left(\mathcal{D}_e \mid \boldsymbol{\theta}; \lambda\right) \triangleq \begin{cases} p\left(\mathcal{D}_e \mid \boldsymbol{\theta}\right) & \text{if } q\left(\boldsymbol{\theta} \mid \mathcal{D}\right) > \lambda \max_{\boldsymbol{\theta}'} q\left(\boldsymbol{\theta}' \mid \mathcal{D}\right), \\ 1 & \text{otherwise} \end{cases}$$

- This helps curb unlearning in region where $q(\boldsymbol{\theta} \mid \mathcal{D}) \simeq 0$ since

$$\tilde{p}_{adj}\left(\boldsymbol{\theta} \mid \mathcal{D}_r; \lambda\right) \propto \begin{cases} \frac{q(\boldsymbol{\theta} \mid \mathcal{D})}{p(\mathcal{D}_e \mid \boldsymbol{\theta})} & \text{if } q\left(\boldsymbol{\theta} \mid \mathcal{D}\right) > \lambda \max_{\boldsymbol{\theta}'} q\left(\boldsymbol{\theta}' \mid \mathcal{D}\right), \\ q(\boldsymbol{\theta} \mid \mathcal{D}) & \text{otherwise} \end{cases}$$

Effect of hyperparameter λ

- $\lambda \simeq 0$ implies that we unlearn without $p_{adj}\left(\mathcal{D}_e \mid \boldsymbol{\theta}; \lambda\right)$, undesirable.
- As $\lambda \uparrow$ unlearning happens in region with sufficiently large $q(\boldsymbol{\theta} \mid \mathcal{D})$
- $\lambda \simeq 1$ implies no unlearning takes place, again undesirable

- Hence, choosing approprite $\lambda \in (0, 1)$ becomes essential.
- **Remedy** - Can tune λ on validation set or use *Reverse KL*

Reverse KL

- Here we minimize $KL\left[\tilde{p}\left(\boldsymbol{\theta} \mid \mathcal{D}_r\right) \| \tilde{q}_v\left(\boldsymbol{\theta} \mid \mathcal{D}_r\right)\right]$ instead of $KL\left[\tilde{q}_u\left(\boldsymbol{\theta} \mid \mathcal{D}_r\right) \| \tilde{p}\left(\boldsymbol{\theta} \mid \mathcal{D}_r\right)\right]$
- Now $\tilde{q}_v\left(\boldsymbol{\theta} \mid \mathcal{D}_r\right)$ overestimates the variance of $\tilde{p}\left(\boldsymbol{\theta} \mid \mathcal{D}_r\right)$ (earlier, it underestimated using EUBO)
- More desirable since sources of inaccuracies lie in region with $\tilde{q}_u\left(\boldsymbol{\theta} \mid \mathcal{D}_r\right) \simeq 0$

- Can further combine with *adjusted* likelihood trick to minimize $KL\left[\tilde{p}_{adj}\left(\boldsymbol{\theta} \mid \mathcal{D}_r\right) \| \tilde{q}_v\left(\boldsymbol{\theta} \mid \mathcal{D}_r\right)\right]$
- Or rather maximize $\mathbb{E}_{q(\boldsymbol{\theta}|\mathcal{D})}\left[\ln \tilde{q}_v\left(\boldsymbol{\theta} \mid \mathcal{D}_r\right) / \tilde{p}_{adj}\left(\mathcal{D}_e \mid \boldsymbol{\theta}\right)\right]$ since,

$$KL\left[\tilde{p}_{adj}\left(\boldsymbol{\theta} \mid \mathcal{D}_r\right) \| \tilde{q}_v\left(\boldsymbol{\theta} \mid \mathcal{D}_r\right)\right] = C_0 - C_1 \mathbb{E}_{q(\boldsymbol{\theta}|\mathcal{D})}\left[\frac{\ln \tilde{q}_v\left(\boldsymbol{\theta} \mid \mathcal{D}_r\right)}{\tilde{p}_{adj}\left(\mathcal{D}_e \mid \boldsymbol{\theta}\right)}\right]$$

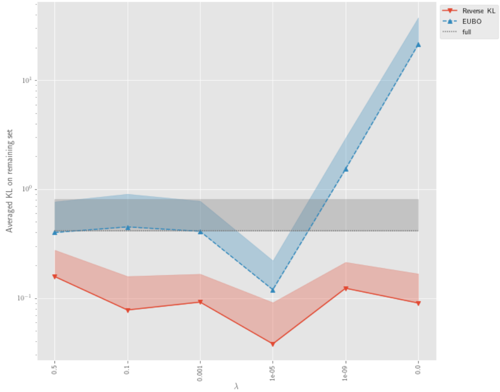
Experiment

- We implement *Sparse Gaussian Process classification* on *synthetic 2-D moon dataset* with binary response.
- Consists of 100 examples of which 20 are to be erased.

- Used *Adam* optimizer
- Batch size $= 100$ (Batch Stochastic-Gradient Descent)
- No. of training iterations $= 30,000$

- We compare the effect of λ on *averaged* KL using EUBO and Reverse KL methods.

Results



- EUBO results in catastrophic unlearning near small values of λ
- Reverse KL seems more robust against the effect of λ

Limitations

- Approach only applies to parametric models
- Implementation only on small scale models. Large models like BNNs have been avoided
- No theoretical guarantees on knowledge removal (essential for legal requirements)

Bayesian Inference Forgetting Framework

- ϵ-certified knowledge removal in Bayesian inference: quantifies performance of knowledge removal
- Energy Functions in Bayesian Inference
- Forgetting Algorithms for Bayesian inference designed based on energy functions
- Theoretical guarantee

ϵ-certified knowledge removal

For any subset $S' \subset S$ and $\epsilon > 0$, we say that algorithm $\mathcal{A}$ performs ϵ-certified knowledge removal if

$$KL\left(\hat{p}_S^{-S'}, \hat{p}_{S-S'}\right) \leq \epsilon$$

Here,

$$\hat{p}_S^{-S'} = \mathcal{A}\left(\hat{p}_S, S'\right)$$

and $\mathcal{A}$ is a forgetting algorithm designed to process the distribution $\hat{p}_{S-S'}$.

Energy Functions in Bayesian Inference

$F(\gamma, S) \rightarrow$ energy function of a probabilistic distribution $\chi(\gamma)$ parameterised by γ over the model parameter space

$$F(\gamma, S) = \sum_{i=1}^{n} h(\gamma, z_i) + f(\gamma)$$

$h(\gamma, z) \rightarrow$ characterises the influence from individual datums
$f(\gamma) \rightarrow$ characterises the influence from the prior

Energy Functions in Bayesian Inference

Similarly, the energy function for $S - S'$ is $F(\gamma, S - S')$. Clearly,

$$F(\gamma, S - S') = F(\gamma, S) - \sum_{z \in S'} h(\gamma, z)$$

Forgetting Algorithm for Bayesian inference

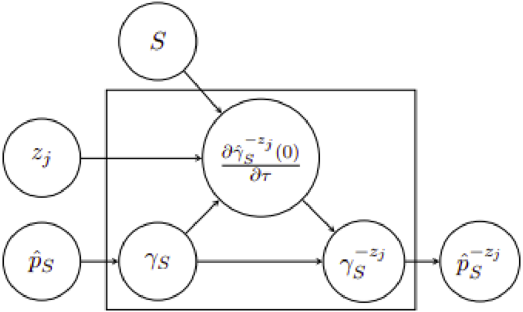


Figure: **Workflow of BIF framework**

Forgetting Algorithm for Bayesian inference

$$F\left(\gamma, S-\{z_j\}\right)=\sum_{i=1}^{n} h\left(\gamma, z_i\right)+f\left(\gamma\right)-h\left(\gamma, z_j\right)$$

$$\nabla_\gamma F\left(\gamma_S, S\right)=\sum_{i=1}^{n} \nabla_\gamma h\left(\gamma_S, z_i\right)+\nabla_\gamma f\left(\gamma_S\right)=0$$

$$\nabla_\gamma F\left(\gamma_{S-\{z_j\}}, S-\{z_j\}\right)=\nabla_\gamma F\left(\gamma_{S-\{z_j\}}, S\right)-\nabla_\gamma h\left(\gamma_S-\{z_j\}, z_j\right)=0$$

$$\nabla_\gamma F\left(\gamma, S\right)+\tau.\nabla_\gamma h\left(\gamma, z_j\right)=0$$

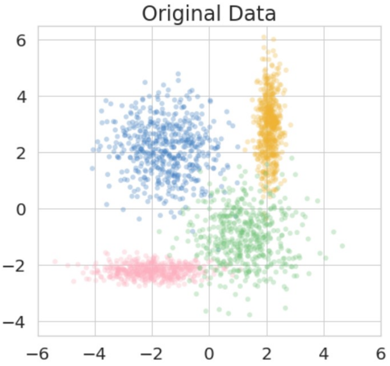
$$\gamma_S^{-z_j}:=\gamma_S-\frac{\partial\hat{\gamma}_S^{-z_j}\left(0\right)}{\partial\tau}$$

Theoretical Guarantee

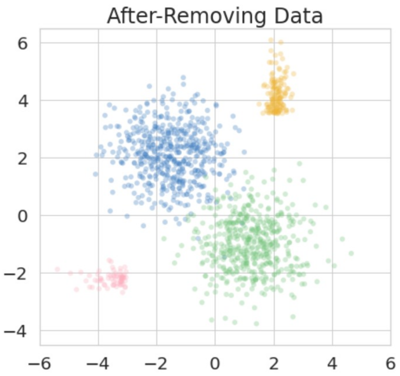
- The mapping $\hat{\gamma}_S^{-z_j}$ uniquely exists
- $\gamma_{S-\{z_j\}} = \hat{\gamma}_S(-1)$ is the global minimiser of $F(\gamma, S - \{z_j\})$
- Approximation error between $\gamma_S^{-z_j}$ and $\gamma_{S-\{z_j\}}$ is not larger than order $O(1/n^2)$, where n is the training set size

Experiment

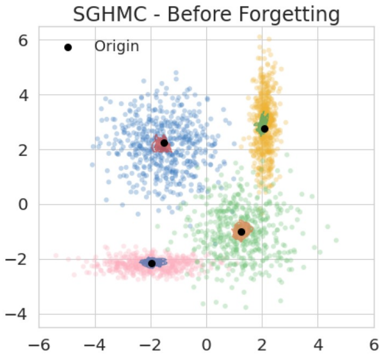
- SVI, SGLD and SGHMC applied to GMM on synthetic data
- Training phase, Forgetting phase



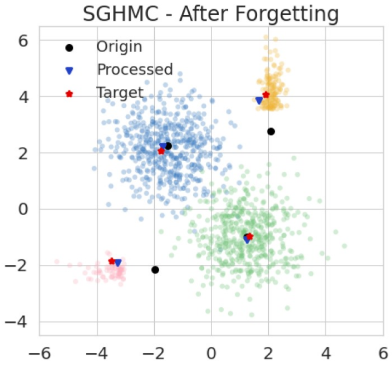
Experiment



Experiment



Experiment



SGHMC - After Forgetting
Origin
Processed
Target
6
4
2
0
−2
−4
−6
−4
−2
0
2
4
6

The End