Variational Bayesian Unlearning

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet

Presented by Ananth Mahadevan

February 10, 2021

Contents

Bayesian Basics

Learning

- Unknown model parameters $\boldsymbol{\theta}$
- Prior belief $p(\boldsymbol{\theta})$
- Set $\mathcal{D}$ of training data
- Learn approximate posterior belief $q(\boldsymbol{\theta}|\mathcal{D}) \approx p(\boldsymbol{\theta}|\mathcal{D})$

Unlearning

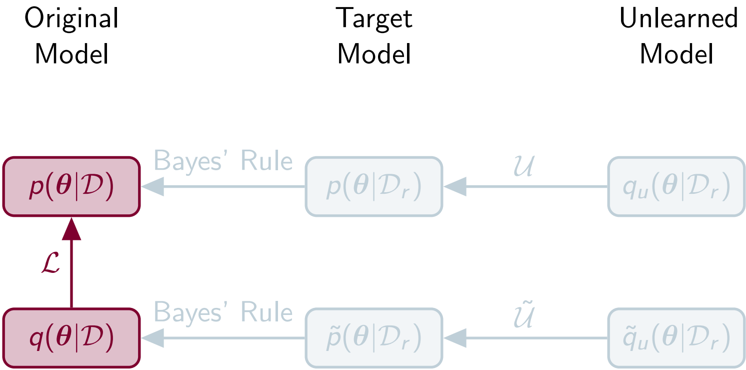
- $\mathcal{D}$ partitioned into $\mathcal{D}_e$ *erased data* and $\mathcal{D}_r$ of *remaining data*
- $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_e$ and $\mathcal{D}_r \cap \mathcal{D}_e = \emptyset$.
- Approximate $p(\boldsymbol{\theta}|\mathcal{D}_r)$

Overview

Original Model | Target Model | Unlearned Model

$$
\begin{tikzcd}
p(\boldsymbol{\theta}|\mathcal{D}) & p(\boldsymbol{\theta}|\mathcal{D}_r) \arrow[l, "\text{Bayes' Rule}"'] & q_u(\boldsymbol{\theta}|\mathcal{D}_r) \arrow[l, "\mathcal{U}"'] \\
q(\boldsymbol{\theta}|\mathcal{D}) \arrow[u, "\mathcal{L}"] & \tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r) \arrow[l, "\text{Bayes' Rule}"'] & \tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r) \arrow[l, "\tilde{\mathcal{U}}"']
\end{tikzcd}
$$

Overview


Original Model
Target Model
Unlearned Model
$p(\boldsymbol{\theta}|\mathcal{D})$
Bayes' Rule
$p(\boldsymbol{\theta}|\mathcal{D}_r)$
$\mathcal{U}$
$q_u(\boldsymbol{\theta}|\mathcal{D}_r)$
$\mathcal{L}$
$q(\boldsymbol{\theta}|\mathcal{D})$
Bayes' Rule
$\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r)$
$\tilde{\mathcal{U}}$
$\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r)$

Evidence Lower Bound (ELBO)

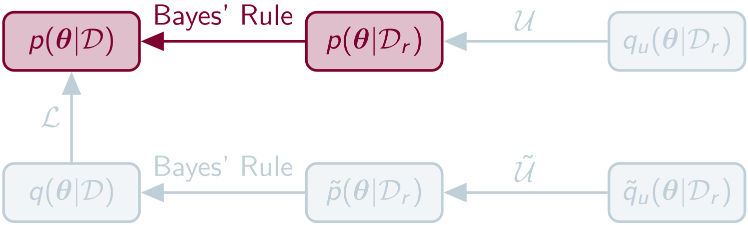
- Need to **minimize** the KL divergence $\mathrm{KL}[q(\boldsymbol{\theta}|\mathcal{D}) \parallel p(\boldsymbol{\theta}|\mathcal{D})] \triangleq \int q(\boldsymbol{\theta}|\mathcal{D}) \; \log(q(\boldsymbol{\theta}|\mathcal{D})/p(\boldsymbol{\theta}|\mathcal{D})) \; \mathrm{d}\boldsymbol{\theta}$
- Or, **maximize** the *evidence lower bound* (ELBO)

$$\mathcal{L} \triangleq \int \underbrace{q(\boldsymbol{\theta}|\mathcal{D}) \; \log p(\mathcal{D}|\boldsymbol{\theta}) \; \mathrm{d}\boldsymbol{\theta}}_{\text{increase likelihood}} - \underbrace{\mathrm{KL}[q(\boldsymbol{\theta}|\mathcal{D}) \parallel p(\boldsymbol{\theta})]}_{\text{remember prior}}$$

- Use **Black Box Variational Inference** (BBVI) if $\mathcal{L}$ cannot be evaluated in closed form
  - Stochastic gradient estimates for Gradient Ascent

Overview

Original Model | Target Model | Unlearned Model



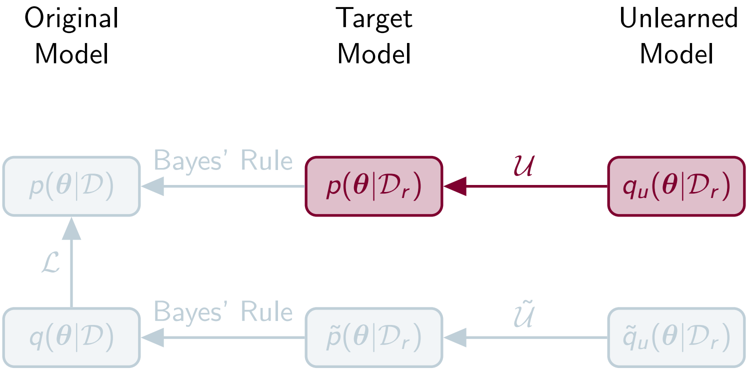
Exact Bayesian Unlearning

- Retraining on $\mathcal{D}_r$ we can get $p(\boldsymbol{\theta}|\mathcal{D}_r)$, but computationally costly,
- Alternatively, using Bayes' rule

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

- For discrete $\boldsymbol{\theta}$ and conjugate priors can be obtained directly
- Paper focuses on **non-conjugate priors**

Overview


Original Model
Target Model
Unlearned Model
$p(\boldsymbol{\theta}|\mathcal{D})$
Bayes' Rule
$p(\boldsymbol{\theta}|\mathcal{D}_r)$
$\mathcal{U}$
$q_u(\boldsymbol{\theta}|\mathcal{D}_r)$
$\mathcal{L}$
$q(\boldsymbol{\theta}|\mathcal{D})$
Bayes' Rule
$\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r)$
$\tilde{\mathcal{U}}$
$\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r)$

Approximate Bayesian Unlearning

- Find $q_u(\boldsymbol{\theta}|\mathcal{D}_r) \approx p(\boldsymbol{\theta}|\mathcal{D}_r)$ by unlearning from erased data $\mathcal{D}_e$
- Predictive distributions
  - $q_u(y|\mathcal{D}_r) \triangleq \int p(y|\boldsymbol{\theta})\ q_u(\boldsymbol{\theta}|\mathcal{D}_r)\ \mathrm{d}\boldsymbol{\theta}$
  - $p(y|\mathcal{D}_r) = \int p(y|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathcal{D}_r)\ \mathrm{d}\boldsymbol{\theta}$
- Loss to minimize: $\mathrm{KL}[q_u(y|\mathcal{D}_r) \parallel p(y|\mathcal{D}_r)]$
  - Closed form may not exist
  - Hard to estimate and optimize

Approximate Bayesian Unlearning

- Find $q_u(\boldsymbol{\theta}|\mathcal{D}_r) \approx p(\boldsymbol{\theta}|\mathcal{D}_r)$ by unlearning from erased data $\mathcal{D}_e$
- Predictive distributions
  - $q_u(y|\mathcal{D}_r) \triangleq \int p(y|\boldsymbol{\theta})\ q_u(\boldsymbol{\theta}|\mathcal{D}_r)\ \mathrm{d}\boldsymbol{\theta}$
  - $p(y|\mathcal{D}_r) = \int p(y|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathcal{D}_r)\ \mathrm{d}\boldsymbol{\theta}$
- Loss to minimize: $\mathrm{KL}[q_u(y|\mathcal{D}_r) \parallel p(y|\mathcal{D}_r)]$
  - Closed form may not exist
  - Hard to estimate and optimize

Propostion 1 Bound

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

- How to **minimize** $\mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel p(\boldsymbol{\theta}|\mathcal{D}_r)]$?

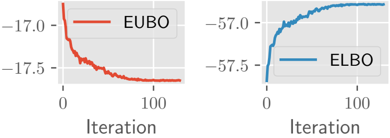
Evidence Upper Bound (EUBO)

- Similar to ELBO define an *evidence upper bound*

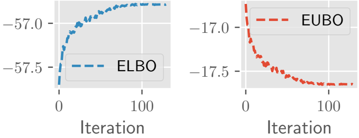
$$\mathcal{U} \triangleq \int \underbrace{q_u(\boldsymbol{\theta}|\mathcal{D}_r)\ \log p(\mathcal{D}_e|\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}}_{\text{Forget } \mathcal{D}_e} + \underbrace{\mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r)\ \|\ p(\boldsymbol{\theta}|\mathcal{D})]}_{\text{Remember } \mathcal{D}(\text{incudes } \mathcal{D}_r)}$$

- We can minimize $\mathrm{KL}[q_u(\boldsymbol{\theta}|\mathcal{D}_r)\ \|\ p(\boldsymbol{\theta}|\mathcal{D}_r)]$ in two ways
  1. **Maximize** ELBO while retraining using $\mathcal{D}_r$
  2. **Minimize** EUBO when unlearning with $\mathcal{D}_e$
- EUBO naturally has regularizing term to avoid *catastrophic forgetting*

EUBO and ELBO



(a) Unlearning from $\mathcal{D}_e$ by minimizing EUBO

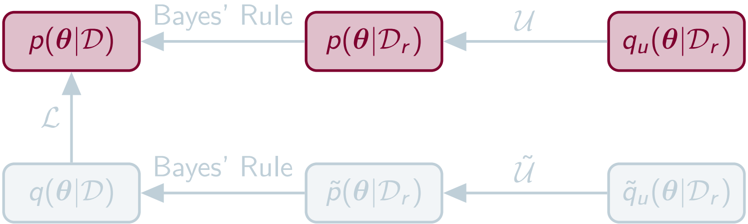


(b) Retraining with $\mathcal{D}_r$ by maximizing ELBO

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

Overview

Original Model | Target Model | Unlearned Model



Reality Check

- In reality we only obtain approximations
- VI training gives approximate posterior $q(\boldsymbol{\theta}|\mathcal{D})$
- We can estimate unknown $p(\boldsymbol{\theta}|\mathcal{D}_r)$ using

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

- $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r)$ from minimizing loss $\mathrm{KL}[\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r) \parallel \tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r)]$
- Define following EUBO

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

Overview

Original Model	Target Model	Unlearned Model

$$
\begin{tikzcd}
p(\boldsymbol{\theta}|\mathcal{D}) & p(\boldsymbol{\theta}|\mathcal{D}_r) \arrow[l, "\text{Bayes' Rule}"'] & q_u(\boldsymbol{\theta}|\mathcal{D}_r) \arrow[l, "\mathcal{U}"'] \\
q(\boldsymbol{\theta}|\mathcal{D}) \arrow[u, "\mathcal{L}"] & \tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r) \arrow[l, "\text{Bayes' Rule}"'] & \tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r) \arrow[l, "\tilde{\mathcal{U}}"']
\end{tikzcd}
$$

Issues

Two possible sources of inaccuracy in $q(\boldsymbol{\theta}|\mathcal{D})$

1. $q(\boldsymbol{\theta}|\mathcal{D})$ often underestimates the variance of $p(\boldsymbol{\theta}|\mathcal{D})$
2. Unlikely that the ELBO is maximized using samples of $\boldsymbol{\theta}$ with small $q(\boldsymbol{\theta}|\mathcal{D})$

Hence, curb unlearning at values of $\boldsymbol{\theta}$ with small $q(\boldsymbol{\theta}|\mathcal{D})$

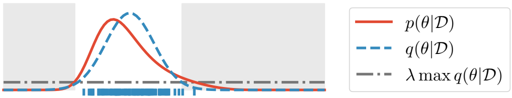
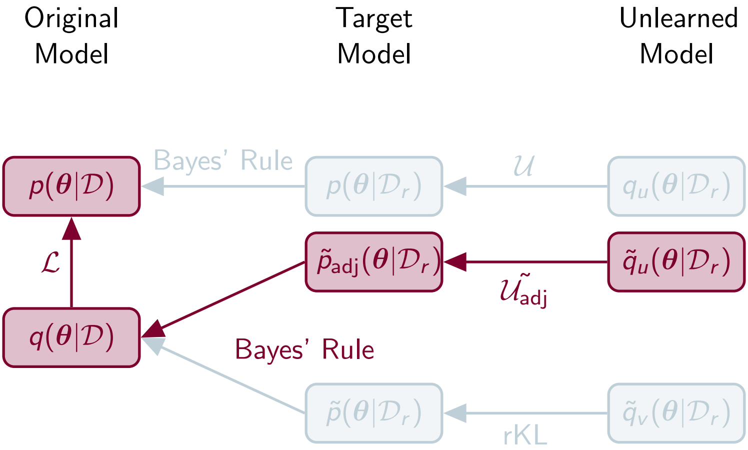


Figure: Plot of $q(\boldsymbol{\theta}|\mathcal{D})$ learned using VI. Gray shaded region corresponds to values of $\boldsymbol{\theta}$ where $q(\boldsymbol{\theta}|\mathcal{D}) \leq \lambda \max_{\boldsymbol{\theta}'} q(\boldsymbol{\theta}'|\mathcal{D})$. Vertical blue strips on horizontal axis show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

Overview


Original Model
Target Model
Unlearned Model
Bayes' Rule
$p(\boldsymbol{\theta}|\mathcal{D})$
$p(\boldsymbol{\theta}|\mathcal{D}_r)$
$\mathcal{U}$
$q_u(\boldsymbol{\theta}|\mathcal{D}_r)$
$\mathcal{L}$
$\tilde{p}_{\text{adj}}(\boldsymbol{\theta}|\mathcal{D}_r)$
$\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r)$
$\tilde{\mathcal{U}}_{\text{adj}}$
$q(\boldsymbol{\theta}|\mathcal{D})$
Bayes' Rule
$\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r)$
rKL
$\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r)$

Adjusted EUBO

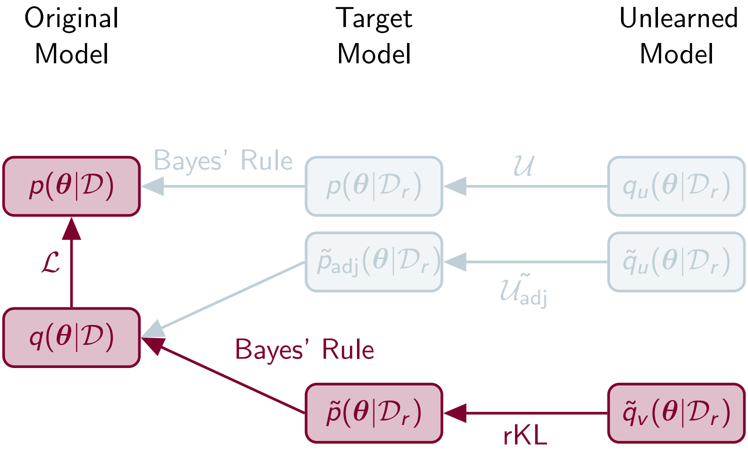
Introduce $\lambda \in [0, 1]$ to control focus of unlearning

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

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

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

Overview


Original Model
Target Model
Unlearned Model
Bayes' Rule
$p(\boldsymbol{\theta}|\mathcal{D})$
$p(\boldsymbol{\theta}|\mathcal{D}_r)$
$\mathcal{U}$
$q_u(\boldsymbol{\theta}|\mathcal{D}_r)$
$\mathcal{L}$
$\tilde{p}_{\text{adj}}(\boldsymbol{\theta}|\mathcal{D}_r)$
$\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r)$
$\tilde{\mathcal{U}}_{\text{adj}}$
$q(\boldsymbol{\theta}|\mathcal{D})$
Bayes' Rule
$\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r)$
rKL
$\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r)$

Reverse KL

- Minimize $\mathrm{KL}[\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r) \parallel \tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r)]$ instead
- Now, $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r)$ **overestimates** the variance of $\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r)$
- Initialize $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r)$ at $q(\boldsymbol{\theta}|\mathcal{D})$ for faster convergence
- Now, SGA naturally curbs unlearning at values of $\boldsymbol{\theta}$ with small $q(\boldsymbol{\theta}|\mathcal{D})$

Adjusted EUBO and rKL

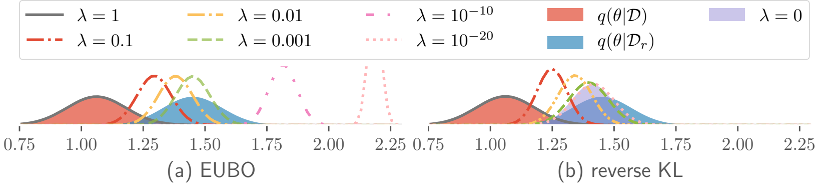
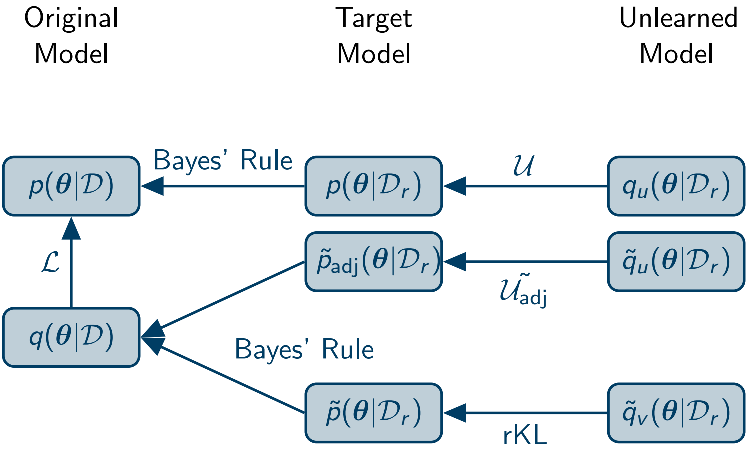


Figure: Plot of approximate posterior beliefs with varying λ obtained by minimizing (a) EUBO (i.e., $\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$) and (b) reverse KL (i.e., $\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r;\lambda)$); horizontal axis denotes $\theta = \alpha$. In (a), a huge 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.

Overview


Original Model
Target Model
Unlearned Model
Bayes' Rule
$p(\boldsymbol{\theta}|\mathcal{D})$
$p(\boldsymbol{\theta}|\mathcal{D}_r)$
$\mathcal{U}$
$q_u(\boldsymbol{\theta}|\mathcal{D}_r)$
$\mathcal{L}$
$\tilde{p}_{\text{adj}}(\boldsymbol{\theta}|\mathcal{D}_r)$
$\tilde{q}_u(\boldsymbol{\theta}|\mathcal{D}_r)$
$\tilde{\mathcal{U}}_{\text{adj}}$
$q(\boldsymbol{\theta}|\mathcal{D})$
Bayes' Rule
$\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_r)$
$\tilde{q}_v(\boldsymbol{\theta}|\mathcal{D}_r)$
rKL

Reults