1. INTRODUCTION

Observations:
- Kernel methods can be used to embed a distribution to a Hilbert space and probability rules can be replaced by corresponding linear operators
- The kernel embedding of a conditional distribution has an optimization formulation
- The posterior distribution in Bayes’ rule has an optimization formulation

Does the kernel embedding of a posterior distribution have an optimization formulation?

Contributions:
- A theoretically justified affirmative answer to the question
- A simpler but faster regularization technique called thresholding regularization
- Posterior regularization for kernel Bayesian inference called kRegBayes, analogous to RegBayes

2. PRELIMINARIES

Kernel embedding:

\[ p_X \rightarrow \mathcal{E}_p[\phi(X)] =: \mu_X \in \mathcal{H}_X, \]

where \( \phi(X) := k(X, \cdot) \). (i) When \( p_X \) is a conditional distribution, \( \mu_X \) is called conditional embedding. (ii) When \( p_X \) is a posterior distribution in a Bayesian setting, \( \mu_X \) is called posterior embedding.

Optimization formulation of conditional embedding

\[ \mu^*_Y|X = \arg \inf_{\mu} \mathcal{E}_p[\mu] = \arg \inf_{\mu} \mathcal{E}_p(\psi(Y) - \mu(X))^2_{\mathcal{H}_Y} \]

Given i.i.d. samples \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \) from \( p(X, Y) \), the estimator is

\[ \hat{\mathcal{E}}_p[\mu] = \frac{1}{n} \| \psi(y_i) - \mu(x_i) \|^2_{\mathcal{H}_Y}, \]

Optimization formulation of posterior distribution

\[ p(Y \mid X = x) = \arg \min_{q(Y)} \left( KL(q(Y)) \| \pi(Y) \right) - \int \log p(X = x \mid Y)q(Y)dq(Y) \]

s.t. \( q(Y) \in \mathcal{P}_{\text{prob}} \)

Posterior regularization for Bayesian inference (RegBayes)

\[ \min_{\pi(Y) : \xi} \left\{ KL(q(Y)) \| \pi(Y) \right\} - \int \log p(X = x \mid Y)dq(Y) + U(\xi) \]

s.t. \( q(Y) \in \mathcal{P}_{\text{prob}}(\xi) \)

3. POSTERIOR EMBEDDING AS A REGRESSOR

Let \( \pi(Y) \) be the prior, \( p(X \mid Y) \) be the likelihood, \( p^*(X, Y) \) be the joint distribution and suppose we have samples to directly estimate \( \pi(Y) \) and \( p(x \mid Y) \). The posterior embedding \( \mu^*_Y|X \) is the same as conditional embedding

\[ \mu^*_Y|X = \arg \inf_{\mu} \mathcal{E}_p[\mu] = \arg \inf_{\mu} \mathcal{E}_p(\psi(Y) - \mu(X))^2_{\mathcal{H}_Y} \]

How to get a reasonable estimator of \( \hat{\mathcal{E}}_p[\mu] \) when we do not have i.i.d. samples from \( p^*(X, Y) \)?

Assuming \( f(x, y) = \| \psi(y) - \mu(x) \|^2_{\mathcal{H}_Y} \) in \( \mathcal{H}_X \), we have

\[ \mathcal{E}_p[\mu] = \mathbb{E}_{X, Y}[\| \psi(Y) - \mu(X) \|^2_{\mathcal{H}_Y}] = \langle f, \mu(X) \rangle_{\mathcal{H}_X} \]

We show in the paper that \( \mu(X, Y) \) can be estimated by

\[ \sum_{i=1}^n \beta_i \psi(Y_i) \odot \psi(X_i) \]

**Theorem 1** (Proof in Appendix). Under some conditions (details in paper), we have the following consistent estimator of \( \hat{\mathcal{E}}_p[\mu] \):

\[ \hat{\mathcal{E}}_p[\mu] = \sum_{i=1}^n \beta_i \| \psi(y_i) - \mu(x_i) \|^2_{\mathcal{H}_Y}, \]

where \( \beta \) is given by \( \beta = (\beta_1, \ldots, \beta_n)^T \).

What if some \( \beta_i \)'s are negative and \( \hat{\mathcal{E}}_p[\mu] \) has no minima?

Under some conditions, \( \hat{\mathcal{E}}_p^+[\mu] = \sum_{i=1}^n \beta_i^+ \| \psi(y_i) - \mu(x_i) \|^2_{\mathcal{H}_Y} \), where \( \beta_i^+ = \max(0, \beta_i) \) is also consistent. This is called thresholding regularization.

Finally, we can establish the consistency of \( \hat{\mu}_X|Y, n \) is the sample used for representing likelihood, \( \{ (x_i, t_i) \}_{i=1}^n \) is the sample used for nonparametric posterior regularization, \( \psi(t_i) \) is the kernel embedding of \( \delta = t_i \) and encourages \( p(Y \mid X = x) \) to be close to \( \delta = t_i \).

4. KREGBAYES

\[ \mathcal{L} := \sum_{i=1}^n \| \mu(x_i) - \psi(y_i) \|^2_{\mathcal{H}_Y} + \lambda \| \mu \|^2_{\mathcal{H}_Y} + \delta \sum_{i=1}^m \| \mu(x_i) - \psi(t_i) \|^2_{\mathcal{H}_Y} \]

where \( \{ (x_i, y_i) \}_{i=1}^m \) is the sample used for representing likelihood, \( \{ (x_i, t_i) \}_{i=1}^m \) is the sample used for nonparametric posterior regularization, \( \psi(t_i) \) is the kernel embedding of \( \delta = t_i \) and encourages \( p(Y \mid X = x) \) to be close to \( \delta = t_i \).