Bayesian Interpretations of Regularization

Charlie Frogner

9.520 Class 20

April 21, 2010

The Plan

Regularized least squares maps $\{(x_i, y_i)\}_{i=1}^n$ to a function that minimizes the regularized loss:

$$f_S = \underset{f \in \mathcal{H}}{\arg\min} \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2$$

Can we interpret RLS from a probabilistic point of view?

Some notation

- $S = \{(x_i, y_i)\}_{i=1}^n$ is the set of observed input/output pairs in $\mathbb{R}^d \times \mathbb{R}$ (the training set).
- X and Y denote the matrices $[x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times d}$ and $[y_1, \ldots, y_n]^T \in \mathbb{R}^n$, respectively.
- θ is a vector of parameters in $\mathbb{R}^p$.
- $p(Y|X, \theta)$ is the joint distribution over outputs Y given inputs X and the parameters.

Where do probabilities show up?

$$\frac{1}{2}\sum_{i=1}^{n} V(y_i, f(x_i)) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$$

becomes

$$p(Y|f,X) \cdot p(f)$$

- **Likelihood**, a.k.a. **noise model**: $p(Y|f,X)$.
  - Gaussian: $y_i \sim \mathcal{N}\left(f^*(x_i), \sigma_i^2\right)$
  - Poisson: $y_i \sim Pois\left(f^*(x_i)\right)$
- **Prior**: $p(f)$.

Where do probabilities show up?

$$\frac{1}{2}\sum_{i=1}^{n} V(y_i, f(x_i)) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$$

becomes

$$p(Y|f,X) \cdot p(f)$$

- **Likelihood**, a.k.a. **noise model**: $p(Y|f,X)$.
  - Gaussian: $y_i \sim \mathcal{N}\left(f^*(x_i), \sigma_i^2\right)$
  - Poisson: $y_i \sim Pois\left(f^*(x_i)\right)$
- **Prior**: $p(f)$.

Estimation

The estimation problem:

- Given data $\{(x_i, y_i)\}_{i=1}^N$ and model $p(Y|f, X), p(f)$.
- Find a good f to explain data.

The Plan

- Maximum likelihood estimation for ERM
- MAP estimation for linear RLS
- MAP estimation for kernel RLS
- Transductive model
- Infinite dimensions get more complicated

Maximum likelihood estimation

- Given data $\{(x_i, y_i)\}_{i=1}^N$ and model $p(Y|f, X), p(f)$.
- A good f is one that maximizes $p(Y|f, X)$.

Maximum likelihood and least squares

For least squares, noise model is:

$$y_i | f, x_i \sim \mathcal{N}\left(f(x_i), \sigma^2\right)$$

a.k.a.

$$Y | f, X \sim \mathcal{N}\left(f(X), \sigma^2 I\right)$$

So

$$p(Y|f,X) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\sum_{i=1}^{N} \frac{1}{\sigma^2}(y_i - f(x_i))^2\right\}$$

Maximum likelihood and least squares

For least squares, noise model is:

$$y_i|f, x_i \sim \mathcal{N}\left(f(x_i), \sigma^2\right)$$

a.k.a.

$$Y|f, X \sim \mathcal{N}\left(f(X), \sigma^2 I\right)$$

So

$$p(Y|f, X) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\sum_{i=1}^{N} \frac{1}{\sigma^2}(y_i - f(x_i))^2\right\}$$

Maximum likelihood and least squares

Maximum likelihood: maximize

$$p(Y|f,X) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{ -\sum_{i=1}^{N} \frac{1}{\sigma^2}(y_i - f(x_i)))^2 \right\}$$

Empirical risk minimization: minimize

$$\sum_{i=1}^{N}(y_i - f(x_i))^2$$

$$\sum_{i=1}^{N}(y_i - f(x_i))^2$$

$$e^{-\sum_{i=1}^{N} \frac{1}{\sigma^2}(y_i - f(x_i))^2}$$

What about regularization?

RLS:

$$\arg\min_f \frac{1}{2}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$$

Is there a model of Y and f that yields RLS?

Yes.

$$e^{-\frac{1}{2\sigma_\varepsilon^2}\left(\sum_{i=1}^{n}(y_i - f(x_i))^2\right) - \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2}$$

$$p(Y|f, X) \cdot p(f)$$

What about regularization?

RLS:

$$\arg\min_f \frac{1}{2}\sum_{i=1}^n (y_i - f(x_i))^2 + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$$

Is there a model of Y and f that yields RLS?

Yes.

$$e^{-\frac{1}{2\sigma_\varepsilon^2}\left(\sum_{i=1}^n (y_i - f(x_i))^2\right) - \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2}$$

$$p(Y|f, X) \cdot p(f)$$

What about regularization?

RLS:

$$\arg\min_f \frac{1}{2}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$$

Is there a model of Y and f that yields RLS?

Yes.

$$e^{-\frac{1}{2\sigma_\varepsilon^2}\left(\sum_{i=1}^{n}(y_i - f(x_i))^2\right)} \cdot e^{-\frac{\lambda}{2}\|f\|_{\mathcal{H}}^2}$$

$$p(Y|f,X) \cdot p(f)$$

What about regularization?

RLS:

$$\arg\min_f \frac{1}{2}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$$

Is there a model of Y and f that yields RLS?

Yes.

$$e^{-\frac{1}{2\sigma_\varepsilon^2}\left(\sum_{i=1}^{n}(y_i - f(x_i))^2\right)} \cdot e^{-\frac{\lambda}{2}\|f\|_{\mathcal{H}}^2}$$

$$p(Y|f, X) \cdot p(f)$$

Posterior function estimates

- Given data $\{(x_i, y_i)\}_{i=1}^N$ and model $p(Y|f, X), p(f)$.
- Find a good f to explain data.

(If we can get $p(f|Y, X)$)
Bayes least squares estimate:

$$\hat{f}_{BLS} = \mathbb{E}_{(f|X,Y)}[f]$$

i.e. the mean of the posterior.
MAP estimate:

$$\hat{f}_{MAP}(Y|X) = \arg\max_f p(f|X, Y)$$

i.e. a mode of the posterior.

Posterior function estimates

- Given data $\{(x_i, y_i)\}_{i=1}^N$ and model $p(Y|f, X), p(f)$.
- Find a good f to explain data.

(If we can get $p(f|Y, X)$)
Bayes least squares estimate:

$$\hat{f}_{BLS} = \mathbb{E}_{(f|X,Y)}[f]$$

i.e. the mean of the posterior.
MAP estimate:

$$\hat{f}_{MAP}(Y|X) = \arg\max_f p(f|X, Y)$$

i.e. a mode of the posterior.

Posterior function estimates

- Given data $\{(x_i, y_i)\}_{i=1}^N$ and model $p(Y|f, X), p(f)$.
- Find a good f to explain data.

(If we can get $p(f|Y, X)$)
Bayes least squares estimate:

$$\hat{f}_{BLS} = \mathbb{E}_{(f|X,Y)}[f]$$

i.e. the mean of the posterior.
MAP estimate:

$$\hat{f}_{MAP}(Y|X) = \arg\max_f p(f|X, Y)$$

i.e. a mode of the posterior.

A posterior on functions?

How to find $p(f|Y,X)$?
Bayes' rule:

$$p(f|X,Y) = \frac{p(Y|X,f)\cdot p(f)}{p(Y|X)} = \frac{p(Y|X,f)\cdot p(f)}{\int p(Y|X,f)df}$$

When is this well-defined?

A posterior on functions?

How to find $p(f|Y,X)$?
Bayes' rule:

$$p(f|X,Y) = \frac{p(Y|X,f)\cdot p(f)}{p(Y|X)} = \frac{p(Y|X,f)\cdot p(f)}{\int p(Y|X,f)df}$$

When is this well-defined?

Functions vs. parameters:

$$\mathcal{H} \cong \mathbb{R}^p$$

Represent functions in $\mathcal{H}$ by their coordinates w.r.t. a basis:

$$f \in \mathcal{H} \leftrightarrow \theta \in \mathbb{R}^p$$

Assume (for the moment): $p < \infty$

A posterior on functions?

Functions vs. parameters:

$$\mathcal{H} \cong \mathbb{R}^p$$

Represent functions in $\mathcal{H}$ by their coordinates w.r.t. a basis:

$$f \in \mathcal{H} \leftrightarrow \theta \in \mathbb{R}^p$$

Assume (for the moment): $p < \infty$

Posterior for linear RLS

Linear function:

$$f(x) = \langle x, \theta \rangle$$

Noise model:

$$Y|X, \theta \sim \mathcal{N}\left(X\theta, \sigma_\varepsilon^2 I\right)$$

Add a *prior*:

$$\theta \sim \mathcal{N}(0, \Lambda)$$

Posterior for linear RLS

Model:

$$Y|X,\theta \sim \mathcal{N}\left(X\theta, \sigma_\varepsilon^2 I\right), \qquad \theta \sim \mathcal{N}(0, \Lambda)$$

Joint over Y and θ:

$$\begin{bmatrix} Y \\ \theta \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} X\Lambda X^T + \sigma_\varepsilon^2 I & X\Lambda \\ \Lambda X^T & \Lambda \end{bmatrix}\right)$$

Condition on Y.

Posterior for linear RLS

Posterior:

$$\theta | X, Y \sim \mathcal{N}\left(\mu_{\theta|X,Y}, \Sigma_{\theta|X,Y}\right)$$

where

$$\mu_{\theta|X,Y} = \Lambda X^T (X \Lambda X^T + \sigma_\varepsilon^2 I)^{-1} Y$$
$$\Sigma_{\theta|X,Y} = \Lambda - \Lambda X^T (X \Lambda X^T + \sigma_\varepsilon^2 I)^{-1} X \Lambda$$

This is Gaussian, so

$$\hat{\theta}_{MAP}(Y|X) = \hat{\theta}_{BLS}(Y|X) = \Lambda X^T (X \Lambda X^T + \sigma_\varepsilon^2 I)^{-1} Y$$

Posterior for linear RLS

Posterior:

$$\theta|X, Y \sim \mathcal{N}\left(\mu_{\theta|X,Y}, \Sigma_{\theta|X,Y}\right)$$

where

$$\mu_{\theta|X,Y} = \Lambda X^T(X\Lambda X^T + \sigma_\varepsilon^2 I)^{-1} Y$$
$$\Sigma_{\theta|X,Y} = \Lambda - \Lambda X^T(X\Lambda X^T + \sigma_\varepsilon^2 I)^{-1} X\Lambda$$

This is Gaussian, so

$$\hat{\theta}_{MAP}(Y|X) = \hat{\theta}_{BLS}(Y|X) = \Lambda X^T(X\Lambda X^T + \sigma_\varepsilon^2 I)^{-1} Y$$

Linear RLS as a MAP estimator

Model:

$$Y|X,\theta \sim \mathcal{N}\left(X\theta, \sigma_\varepsilon^2 I\right), \qquad \theta \sim \mathcal{N}(0, \Lambda)$$

$$\hat{\theta}_{MAP}(Y|X) = \Lambda X^T (X\Lambda X^T + \sigma_\varepsilon^2 I)^{-1} Y$$

Recall the linear RLS solution:

$$\begin{aligned}\hat{\theta}_{RLS}(Y|X) &= \arg\min_\theta \frac{1}{2}\sum_{i=1}^N (y_i - \langle x_i, \theta\rangle)^2 + \frac{\lambda}{2}\|\theta\|^2 \\ &= \Lambda X^T (XX^T + \frac{\lambda}{2} I)^{-1} Y\end{aligned}$$

So what's Λ? λ?

Linear RLS as a MAP estimator

Model:

$$Y|X, \theta \sim \mathcal{N}\left(X\theta, \sigma_\varepsilon^2 I\right), \qquad \theta \sim \mathcal{N}(0, \Lambda)$$

$$\hat{\theta}_{MAP}(Y|X) = \Lambda X^T (X \Lambda X^T + \sigma_\varepsilon^2 I)^{-1} Y$$

Recall the linear RLS solution:

$$\hat{\theta}_{RLS}(Y|X) = \arg\min_\theta \frac{1}{2} \sum_{i=1}^{N} (y_i - \langle x_i, \theta \rangle)^2 + \frac{\lambda}{2} \|\theta\|^2$$

$$= \Lambda X^T (X X^T + \frac{\lambda}{2} I)^{-1} Y$$

So what's Λ? λ?

Represent functions in $\mathcal{H}$ by their coordinates w.r.t. a basis:

$$f \in \mathcal{H} \leftrightarrow \theta \in \mathbb{R}^p$$

Which basis?

Extending to kernel RLS

Mercer's theorem:

$$K(x_i, x_j) = \sum_k \nu_k \psi_k(x_i)\psi_k(x_j)$$

where $\nu_k \psi_k(\cdot) = \int K(\cdot, y)\psi_k(y)dy$ for all k. The functions $\{\sqrt{\nu_k}\psi_k(\cdot)\}$ form an *orthonormal basis* for $\mathcal{H}_K$.
Let $\phi(\cdot) = [\sqrt{\nu_1}\psi_1(\cdot), \ldots, \sqrt{\nu_p}\psi_p(\cdot)]$. Then:

$$\mathcal{H}_K = \{\phi(\cdot)\theta | \theta \in \mathbb{R}^p\}$$

Posterior for kernel RLS

Model for *linear* RLS:

$$Y|X,\theta \sim \mathcal{N}\left(X\theta, \sigma_\varepsilon^2 I\right), \qquad \theta \sim \mathcal{N}(0, I)$$

Model for kernel RLS?

$$Y|X,\theta \sim \mathcal{N}\left(\phi(X)\theta, \sigma_\varepsilon^2 I\right), \qquad \theta \sim \mathcal{N}(0, I)$$

Then:

$$\hat{\theta}_{MAP}(Y|X) = \phi(X)^T(\phi(X)\phi(X)^T + \sigma_\varepsilon^2 I)^{-1} Y$$

Potential problem?

Posterior for kernel RLS

Model for *linear* RLS:

$$Y|X,\theta \sim \mathcal{N}\left(X\theta, \sigma_\varepsilon^2 I\right), \qquad \theta \sim \mathcal{N}(0, I)$$

Model for kernel RLS?

$$Y|X,\theta \sim \mathcal{N}\left(\phi(X)\theta, \sigma_\varepsilon^2 I\right), \qquad \theta \sim \mathcal{N}(0, I)$$

Then:

$$\hat{\theta}_{MAP}(Y|X) = \phi(X)^T(\phi(X)\phi(X)^T + \sigma_\varepsilon^2 I)^{-1} Y$$

Potential problem?

Posterior for kernel RLS

Model for *linear* RLS:

$$Y|X, \theta \sim \mathcal{N}\left(X\theta, \sigma_\varepsilon^2 I\right), \qquad \theta \sim \mathcal{N}(0, I)$$

Model for kernel RLS?

$$Y|X, \theta \sim \mathcal{N}\left(\phi(X)\theta, \sigma_\varepsilon^2 I\right), \qquad \theta \sim \mathcal{N}(0, I)$$

Then:

$$\hat{\theta}_{MAP}(Y|X) = \phi(X)^T(K + \sigma_\varepsilon^2 I)^{-1} Y$$

Potential problem?

Prior on infinite-dimensional space

Problem: there's no such thing as

$$\theta \sim \mathcal{N}(0, I)$$

when $\theta \in \mathbb{R}^\infty$!

A quick recap

- **Empirical risk minimization** is ML.

$$p(Y|f, X) \propto e^{-\frac{1}{2}\sum_{i=1}^{N}(y_i - f(x_i))^2}$$

- **Linear RLS** is MAP.

$$p(Y, f|X) \propto e^{-\frac{1}{2}\sum_{i=1}^{N}(y_i - \langle x_i, \theta\rangle)^2} \cdot e^{-\frac{\lambda}{2}\theta^T\theta}$$

- **Kernel RLS** is also MAP.

$$p(Y, f|X) \propto e^{-\frac{1}{2}\sum_{i=1}^{N}(y_i - f(x_i))^2} \cdot e^{-\frac{\lambda}{2}\|f\|_{\mathcal{H}}^2}$$

But these aren't well-defined for infinite dimensional function spaces...

A quick recap

- **Empirical risk minimization** is ML.

$$p(Y|f,X) \propto e^{-\frac{1}{2}\sum_{i=1}^{N}(y_i - f(x_i))^2}$$

- **Linear RLS** is MAP.

$$p(Y,f|X) \propto e^{-\frac{1}{2}\sum_{i=1}^{N}(y_i - \langle x_i,\theta\rangle)^2} \cdot e^{-\frac{\lambda}{2}\theta^T\theta}$$

- **Kernel RLS** is also MAP.

$$p(Y,f|X) \propto e^{-\frac{1}{2}\sum_{i=1}^{N}(y_i - f(x_i))^2} \cdot e^{-\frac{\lambda}{2}\|f\|_{\mathcal{H}}^2}$$

But these aren't well-defined for infinite dimensional function spaces...

A quick recap

- **Empirical risk minimization** is ML.

$$p(Y|f,X) \propto e^{-\frac{1}{2}\sum_{i=1}^{N}(y_i - f(x_i))^2}$$

- **Linear RLS** is MAP.

$$p(Y,f|X) \propto e^{-\frac{1}{2}\sum_{i=1}^{N}(y_i - \langle x_i, \theta \rangle)^2} \cdot e^{-\frac{\lambda}{2}\theta^T\theta}$$

- **Kernel RLS** is also MAP.

$$p(Y,f|X) \propto e^{-\frac{1}{2}\sum_{i=1}^{N}(y_i - f(x_i))^2} \cdot e^{-\frac{\lambda}{2}\|f\|_{\mathcal{H}}^2}$$

But these aren't well-defined for infinite dimensional function spaces...

Transductive setting

We hinted at problems if $\dim \mathcal{H}_K = \infty$.

Idea: Forget about estimating θ (i.e. f).

Instead: Estimate *predicted outputs*

$$Y^* = [y_1^*, \ldots, y_M^*]^T$$

at *test inputs*

$$X^* = [x_1^*, \ldots, x_M^*]^T$$

Need the joint distribution over Y^* and Y.

Transductive setting

Say Y^* and Y are *jointly Gaussian*:

$$\begin{bmatrix} Y \\ Y^* \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mu_Y \\ \mu_{Y^*} \end{bmatrix}, \begin{bmatrix} \Lambda_Y & \Lambda_{YY^*} \\ \Lambda_{Y^*Y} & \Lambda_{Y^*} \end{bmatrix}\right)$$

Want: kernel RLS.

General form for the posterior:

$$Y^*|X, Y \sim \mathcal{N}\left(\mu_{Y^*|X,Y}, \Sigma_{Y^*|X,Y}\right)$$

where

$$\mu_{Y^*|X,Y} = \mu_{Y^*} + \Lambda_{YY^*}^T \Lambda_Y^{-1}(Y - \mu_Y)$$
$$\Sigma_{Y^*|X,Y} = \Lambda_{Y^*} - \Lambda_{YY^*}^T \Lambda_Y^{-1} \Lambda_{YY^*}$$

Transductive setting

Say Y^* and Y are *jointly Gaussian*:

$$\left[\begin{array}{c} Y \\ Y^* \end{array}\right] = \mathcal{N}\left(\left[\begin{array}{c} \mu_Y \\ \mu_{Y^*} \end{array}\right], \left[\begin{array}{cc} \Lambda_Y & \Lambda_{YY^*} \\ \Lambda_{Y^*Y} & \Lambda_{Y^*} \end{array}\right]\right)$$

Want: kernel RLS.

General form for the posterior:

$$Y^*|X, Y \sim \mathcal{N}\left(\mu_{Y*|X,Y}, \Sigma_{Y^*|X,Y}\right)$$

where

$$\mu_{Y^*|X,Y} = \mu_{Y^*} + \Lambda_{YY^*}^T \Lambda_Y^{-1}(Y - \mu_Y)$$
$$\Sigma_{Y^*|X,Y} = \Lambda_{Y^*} - \Lambda_{YY^*}^T \Lambda_Y^{-1} \Lambda_{YY^*}$$

Set $\Lambda_Y = K(X, X) + \sigma^2 I$, $\Lambda_{YY^*} = K(X, X^*)$, $\Lambda_{Y^*} = K(X^*, X^*)$.

Posterior:

$$Y^*|X, Y \sim \mathcal{N}\left(\mu_{Y*|X,Y}, \Sigma_{Y^*|X,Y}\right)$$

where

$$\mu_{Y^*|X,Y} = \mu_{Y^*} + K(X^*, X)(K(X, X + \sigma^2 I)^{-1}(Y - \mu_Y)$$
$$\Sigma_{Y*|X,Y} = K(X^*, X^*) - K(X^*, X)(K(X, X) + \sigma^2 I)^{-1} K(X, X^*)$$

So: $\hat{Y}^*_{MAP} = \hat{f}_{RLS}(X^*)$.

Transductive setting

Model:

$$\begin{bmatrix} Y \\ Y^* \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mu_Y \\ \mu_{Y^*} \end{bmatrix}, \begin{bmatrix} K(X,X) + \sigma_\varepsilon^2 I & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix} \right)$$

MAP estimate (posterior mean) $=$ RLS function *at every point* x^*, regardless of dim $\mathcal{H}_K$.

Are the prior and posterior (*on points*!) consistent with a distribution on $\mathcal{H}_K$?

Transductive setting

Strictly speaking, θ and f don't come into play here at all:

Have: $p(Y^*|X, Y)$
Do not have: $p(\theta|X, Y)$ or $p(f|X, Y)$

But, *if $\mathcal{H}_K$ is finite dimensional*, the joint over Y and Y^* is consistent with:

- $Y = f(X) + \varepsilon$,
- $Y^* = f(X)$, and
- $f \in \mathcal{H}_K$ is a random trajectory from a **Gaussian process** over the domain, with mean μ and covariance K.
- (Ergo, people call this "Gaussian process regression.") (Also "Kriging," because of a guy.)

Transductive setting

Strictly speaking, θ and f don't come into play here at all:

Have: $p(Y^*|X, Y)$
Do not have: $p(\theta|X, Y)$ or $p(f|X, Y)$

But, *if $\mathcal{H}_K$ is finite dimensional*, the joint over Y and Y^* is consistent with:

- $Y = f(X) + \varepsilon$,
- $Y^* = f(X)$, and
- $f \in \mathcal{H}_K$ is a random trajectory from a **Gaussian process** over the domain, with mean μ and covariance K.
- (Ergo, people call this "Gaussian process regression.") (Also "Kriging," because of a guy.)

Recap redux

- **Empirical risk minimization** is the maximum likelihood estimator when:
$$y = x^T\theta + \varepsilon$$
- **Linear RLS** is the MAP estimator when:
$$y = x^T\theta + \varepsilon, \qquad \theta \sim \mathcal{N}(0, I)$$
- **Kernel RLS** is the MAP estimator when:
$$y = \phi(x)^T\theta + \varepsilon, \qquad \theta \sim \mathcal{N}(0, I)$$
in finite dimensional $\mathcal{H}_K$.
- **Kernel RLS** is the MAP estimator *at points* when:
$$\begin{bmatrix} Y \\ Y^* \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mu_Y \\ \mu_{Y^*} \end{bmatrix}, \begin{bmatrix} K(X,X) + \sigma_\varepsilon^2 I & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix} \right)$$
in possibly infinite dimensional $\mathcal{H}_K$.

Is this useful in practice?

- Want confidence intervals + believe the posteriors are meaningful = yes
- Maybe other reasons?