Bayesian Interpretations of Regularization

Charlie Frogner

9.520 Class 10

March 12, 2012

The Plan

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

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

Can we interpret RLS from a probabilistic point of view?

Some notation

- Training set: $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- Inputs: $\mathbf{X} = \{x_1, \ldots, x_n\}$.
- Labels: $\mathbf{Y} = \{y_1, \ldots, y_n\}$.
- Parameters: $\theta \in \mathbb{R}^p$.
- $p(\mathbf{Y}|\mathbf{X}, \theta)$ is the joint distribution over labels **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

$$p(\mathbf{Y}|f, \mathbf{X}) \cdot p(f)$$

- **Likelihood**, a.k.a. **noise model**: $p(\mathbf{Y}|f, \mathbf{X})$.
  - 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

$$p(\mathbf{Y}|f,\mathbf{X}) \cdot p(f)$$

- **Likelihood**, a.k.a. **noise model**: $p(\mathbf{Y}|f,\mathbf{X})$.
  - 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 $p(\mathbf{Y}|f, \mathbf{X}), p(f)$.
- 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 $p(\mathbf{Y}|f, \mathbf{X}), p(f)$.
- A good f is one that maximizes $p(\mathbf{Y}|f, \mathbf{X})$.

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.

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

So

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

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.

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

So

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

Maximum likelihood and least squares

Maximum likelihood: maximize

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

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.

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

$$p(\mathbf{Y}|f, \mathbf{X}) \cdot p(f)$$

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.

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

$$p(\mathbf{Y}|f,\mathbf{X}) \cdot p(f)$$

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.

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

$$p(\mathbf{Y}|f, \mathbf{X}) \cdot p(f)$$

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 $\mathbf{Y}$ and f that yields RLS?

Yes.

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

$$p(\mathbf{Y}|f,\mathbf{X}) \cdot p(f)$$

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

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

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

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

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

i.e. a mode of the posterior.

Posterior function estimates

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

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

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

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

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

i.e. a mode of the posterior.

Posterior function estimates

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

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

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

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

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

i.e. a mode of the posterior.

A posterior on functions?

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

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

When is this well-defined?

A posterior on functions?

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

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

When is this well-defined?

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): $p < \infty$

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): $p < \infty$

A posterior on functions?

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:

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

Prior on infinite-dimensional space

Problem: there's no such thing as

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

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

Posterior for linear RLS

Linear function:

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

Noise model:

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

Add a *prior*:

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

Posterior for linear RLS

Model:

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

Joint over **Y** and θ:

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

Condition on **Y**.

Posterior for linear RLS

Posterior:

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

where

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

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

This is Gaussian, so

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

Posterior for linear RLS

Posterior:

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

where

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

This is Gaussian, so

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

Linear RLS as a MAP estimator

Model:

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

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

Recall the linear RLS solution:

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

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

So what's λ?

Linear RLS as a MAP estimator

Model:

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

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

Recall the linear RLS solution:

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

So what's λ*?*

Posterior for kernel RLS

Model for *linear* RLS:

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

Model for kernel RLS?

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

Then:

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

Posterior for kernel RLS

Model for *linear* RLS:

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

Model for kernel RLS?

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

Then:

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

Posterior for kernel RLS

Model for *linear* RLS:

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

Model for kernel RLS?

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

Then:

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

A quick recap

- **Empirical risk minimization** is ML.

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

- **Linear RLS** is MAP.

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

- **Kernel RLS** is also MAP.

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

A quick recap

- **Empirical risk minimization** is ML.

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

- **Linear RLS** is MAP.

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

- **Kernel RLS** is also MAP.

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

A quick recap

- **Empirical risk minimization** is ML.

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

- **Linear RLS** is MAP.

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

- **Kernel RLS** is also MAP.

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

Transductive setting

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

Instead: Estimate *predicted outputs*

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

at *test inputs*

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

Need the joint distribution over $\mathbf{Y}^*$ and $\mathbf{Y}$.

Transductive setting

Say $\mathbf{Y}^*$ and $\mathbf{Y}$ are *jointly Gaussian*:

$$\begin{bmatrix} \mathbf{Y} \\ Y^* \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Lambda_{\mathbf{Y}} & \Lambda_{\mathbf{Y}Y^*} \\ \Lambda_{Y^*\mathbf{Y}} & \Lambda_{Y^*} \end{bmatrix} \right)$$

Want: kernel RLS.

General form for the posterior:

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

where

$$\mu_{Y^*|\mathbf{X},\mathbf{Y}} = \Lambda_{\mathbf{Y}Y^*}^T \Lambda_{\mathbf{Y}}^{-1} \mathbf{Y}$$
$$\Sigma_{Y^*|\mathbf{X},\mathbf{Y}} = \Lambda_{\mathbf{Y}^*} - \Lambda_{\mathbf{Y}Y^*}^T \Lambda_{\mathbf{Y}}^{-1} \Lambda_{\mathbf{Y}Y^*}$$

Transductive setting

Say $\mathbf{Y}^*$ and $\mathbf{Y}$ are *jointly Gaussian*:

$$\begin{bmatrix} \mathbf{Y} \\ Y^* \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Lambda_{\mathbf{Y}} & \Lambda_{\mathbf{Y}Y^*} \\ \Lambda_{Y^*\mathbf{Y}} & \Lambda_{Y^*} \end{bmatrix} \right)$$

Want: kernel RLS.

General form for the posterior:

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

where

$$\mu_{Y^*|\mathbf{X},\mathbf{Y}} = \Lambda_{\mathbf{Y}Y^*}^T \Lambda_{\mathbf{Y}}^{-1} \mathbf{Y}$$
$$\Sigma_{Y^*|\mathbf{X},\mathbf{Y}} = \Lambda_{\mathbf{Y}^*} - \Lambda_{\mathbf{Y}Y^*}^T \Lambda_{\mathbf{Y}}^{-1} \Lambda_{\mathbf{Y}Y^*}$$

Transductive setting

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

Posterior:

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

where

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

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

Transductive setting

Model:

$$\begin{bmatrix} \mathbf{Y} \\ Y^* \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_\varepsilon^2 I & K(\mathbf{X}, X^*) \\ K(X^*, \mathbf{X}) & K(X^*, X^*) \end{bmatrix} \right)$$

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

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

- $\mathbf{Y} = f(\mathbf{X}) + \varepsilon$,
- $Y^* = f(\mathbf{X})$, 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: $p(Y^*|\mathbf{X}, \mathbf{Y})$
Do not have: $p(\theta|\mathbf{X}, \mathbf{Y})$ or $p(f|\mathbf{X}, \mathbf{Y})$

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

- $\mathbf{Y} = f(\mathbf{X}) + \varepsilon$,
- $Y^* = f(\mathbf{X})$, 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:
$$\begin{bmatrix} \mathbf{Y} \\ Y^* \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mu_{\mathbf{Y}} \\ \mu_{Y^*} \end{bmatrix}, \begin{bmatrix} K(\mathbf{X},\mathbf{X}) + \sigma_\varepsilon^2 I & K(\mathbf{X}, X^*) \\ K(X^*, \mathbf{X}) & K(X^*, X^*) \end{bmatrix}\right)$$
in possibly infinite dimensional $\mathcal{H}_K$.

Is this useful in practice?

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