Bayesian Interpretations of Regularization

Charlie Frogner

9.520 Class 15

April 1, 2009

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 justify Tikhonov regularization from a probabilistic point of view?

The Plan

- Bayesian estimation basics
- Bayesian interpretation of ERM
- Bayesian interpretation of linear RLS
- Bayesian interpretation of kernel RLS
- Transductive model
- Infinite dimensions = weird

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.

The setup:

- A **model:** relates *observed* quantities (α) to an *unobserved* quantity (say β).
- Want: an **estimator** – maps observed data α back to an estimate of unobserved β.
- Nothing new yet...

Estimator

$\beta \in B$ is unobserved, $\alpha \in A$ is observed. An estimator for β is a function

$$\hat{\beta} : A \to B$$

such that $\hat{\beta}(\alpha)$ is an estimate of β given an observation α.

Estimation

Tikhonov fits in the estimation framework.

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

Regression model:

$$y_i = f(x_i) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\left(0, \sigma_\varepsilon^2 I\right)$$

Bayesian Estimation

Difference: *Bayesian* model specifies $\boxed{p(\beta, \alpha)}$, usually by a **measurement model**, $p(\alpha|\beta)$ and a **prior** $p(\beta)$.

Bayesian model

β is unobserved, α is observed.

$$p(\beta, \alpha) = p(\alpha|\beta) \cdot p(\beta)$$

ERM as a Maximum Likelihood Estimator

(Linear) Expected risk minimization:

$$f_S(x) = x^T \hat{\theta}_{ERM}(S), \qquad \hat{\theta}_{ERM}(S) = \arg\min_{\theta} \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2$$

Measurement model:

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

X fixed/non-random, θ is unknown.

ERM as a Maximum Likelihood Estimator

(Linear) Expected risk minimization:

$$f_S(x) = x^T \hat{\theta}_{ERM}(S), \qquad \hat{\theta}_{ERM}(S) = \arg\min_{\theta} \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2$$

Measurement model:

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

X fixed/non-random, θ is unknown.

ERM as a Maximum Likelihood Estimator

Measurement model:

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

Want to estimate θ.

- Can do this *without defining a prior on* θ.
- Maximize the **likelihood**, i.e. the probability of the observations.

Likelihood

The **likelihood** of any fixed parameter vector θ is:

$$L(\theta|X) = p(Y|X,\theta)$$

Note: we always condition on X.

ERM as a Maximum Likelihood Estimator

Measurement model:

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

Likelihood:

$$\begin{aligned} L(\theta|X) &= \mathcal{N}\left(Y; X\theta, \sigma_\varepsilon^2 I\right) \\ &\propto \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\|Y - X\theta\|^2\right) \end{aligned}$$

Maximum likelihood estimator is ERM:

$$\arg\min_\theta \frac{1}{2}\|Y - X\theta\|^2$$

ERM as a Maximum Likelihood Estimator

Measurement model:

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

Likelihood:

$$L(\theta|X) = \mathcal{N}\left(Y; X\theta, \sigma_\varepsilon^2 I\right) \propto \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\|Y - X\theta\|^2\right)$$

Maximum likelihood estimator is ERM:

$$\arg\min_\theta \frac{1}{2}\|Y - X\theta\|^2$$

Really?

$$\frac{1}{2}\|Y - X\theta\|^2$$

Really?

$$e^{-\frac{1}{2\sigma_\varepsilon^2}\|Y - X\theta\|^2}$$

What about regularization?

Linear regularized least squares:

$$f_S(x) = x^T\theta, \quad \hat{\theta}_{RLS}(S) = \arg\min_{\theta} \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \frac{\lambda}{2}\|\theta\|^2$$

Is there a model of Y and θ that yields linear RLS?

Yes.

$$e^{-\frac{1}{2\sigma_\varepsilon^2}\left(\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \frac{\lambda}{2}\|\theta\|^2\right)}$$

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

What about regularization?

Linear regularized least squares:

$$f_S(x) = x^T\theta, \quad \hat{\theta}_{RLS}(S) = \arg\min_{\theta} \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \frac{\lambda}{2}\|\theta\|^2$$

Is there a model of Y and θ that yields linear RLS?

Yes.

$$e^{-\frac{1}{2\sigma_\varepsilon^2}\left(\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \frac{\lambda}{2}\|\theta\|^2\right)}$$

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

What about regularization?

Linear regularized least squares:

$$f_S(x) = x^T\theta, \quad \hat{\theta}_{RLS}(S) = \arg\min_{\theta} \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \frac{\lambda}{2}\|\theta\|^2$$

Is there a model of Y and θ that yields linear RLS?

Yes.

$$e^{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2} \cdot e^{-\frac{\lambda}{2}\|\theta\|^2}$$

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

What about regularization?

Linear regularized least squares:

$$f_S(x) = x^T\theta, \quad \hat{\theta}_{RLS}(S) = \arg\min_{\theta} \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \frac{\lambda}{2}\|\theta\|^2$$

Is there a model of Y and θ that yields linear RLS?

Yes.

$$e^{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2} \cdot e^{-\frac{\lambda}{2}\|\theta\|^2}$$

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

Linear RLS as a MAP estimator

Measurement model:

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

Add a *prior*:

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

So $\sigma_\varepsilon^2 = \lambda$. How to estimate θ?

The Bayesian method

- Take $p(Y|X,\theta)$ and $p(\theta)$.
- Apply *Bayes' rule* to get **posterior**:

$$p(\theta|X,Y) = \frac{p(Y|X,\theta)\cdot p(\theta)}{p(Y|X)} = \frac{p(Y|X,\theta)\cdot p(\theta)}{\int p(Y|X,\theta)d\theta}$$

- Use the posterior to estimate θ.

Estimators that use the posterior

Bayes least squares estimator

The *Bayes least squares estimator* for θ given the observed Y is:

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

i.e. the mean of the posterior.

Maximum a posteriori estimator

The *MAP estimator* for θ given the observed Y is:

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

i.e. a mode of the posterior.

Estimators that use the posterior

Bayes least squares estimator

The *Bayes least squares estimator* for θ given the observed Y is:

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

i.e. the mean of the posterior.

Maximum a posteriori estimator

The *MAP estimator* for θ given the observed Y is:

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

i.e. a mode of the posterior.

Linear RLS as a MAP estimator

Model:

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

Joint over Y and θ:

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

Condition on $Y|X$.

Linear RLS as a MAP estimator

Model:

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

Posterior:

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

where

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

This is Gaussian, so

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

Linear RLS as a MAP estimator

Model:

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

Posterior:

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

where

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

This is Gaussian, so

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

Linear RLS as a MAP estimator

Model:

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

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

Recall the linear RLS solution:

$$\hat{\theta}_{RLS}(Y|X) = \arg\min_\theta \frac{1}{2}\sum_{i=1}^n (y_i - x_i^T\theta)^2 + \frac{\lambda}{2}\|\theta\|^2$$
$$= X^T(XX^T + \frac{\lambda}{2} I)^{-1} Y$$

How do we write the estimated function?

Linear RLS as a MAP estimator

Model:

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

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

Recall the linear RLS solution:

$$\hat{\theta}_{RLS}(Y|X) = \arg\min_\theta \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \frac{\lambda}{2}\|\theta\|^2$$
$$= X^T(XX^T + \frac{\lambda}{2}I)^{-1}Y$$

How do we write the estimated function?

What about Kernel RLS?

We can use basically the same trick to derive kernel RLS;

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

How?

Feature space: $f(x) = \langle \theta, \phi(x) \rangle_{\mathcal{F}}$

$$e^{-\frac{1}{2}(\sum_{i=1}^{n}(y_i - \phi(x_i)^T\theta)^2 + \frac{\lambda}{2}\theta^T\theta)}$$

Feature space must be finite-dimensional.

We can use basically the same trick to derive kernel RLS;

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

How?

Feature space: $f(x) = \langle \theta, \phi(x) \rangle_{\mathcal{F}}$

$$e^{-\frac{1}{2}(\sum_{i=1}^{n}(y_i - \phi(x_i)^T\theta)^2 + \frac{\lambda}{2}\theta^T\theta)}$$

Feature space must be finite-dimensional.

What about Kernel RLS?

We can use basically the same trick to derive kernel RLS;

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

How?

Feature space: $f(x) = \langle \theta, \phi(x) \rangle_{\mathcal{F}}$

$$e^{-\frac{1}{2} \sum_{i=1}^{n} (y_i - \phi(x_i)^T \theta)^2} \cdot e^{-\frac{\lambda}{2} \theta^T \theta}$$

Feature space must be finite-dimensional.

What about Kernel RLS?

We can use basically the same trick to derive kernel RLS;

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

How?

Feature space: $f(x) = \langle \theta, \phi(x) \rangle_{\mathcal{F}}$

$$e^{-\frac{1}{2} \sum_{i=1}^{n} (y_i - \phi(x_i)^T \theta)^2} \cdot e^{-\frac{\lambda}{2} \theta^T \theta}$$

Feature space must be finite-dimensional.

More notation

- $\phi(X) = [\phi(x_1), \ldots, \phi(x_n)]^T$
- $K(X, X)$ is the kernel matrix: $[K(X, X)]_{ij} = K(x_i, x_j)$
- $K(x, X) = [K(x, x_1), \ldots, K(x, x_n)]$
- $f(X) = [f(x_1), \ldots, f(x_n)]^T$

What about Kernel RLS?

Model:

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

Then:

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

What is $\phi(X)\phi(X)^T$?

It's $K(X, X)$.

What about Kernel RLS?

Model:

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

Then:

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

What is $\phi(X)\phi(X)^T$?

It's $K(X,X)$.

What about Kernel RLS?

Model:

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

Then:

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

Estimated function?

$$\begin{aligned}\hat{f}_{MAP}(x) &= \phi(x)\hat{\theta}_{MAP}(Y|X)\\ &= \phi(x)\phi(X)^T(K(X,X) + \sigma_\varepsilon^2 I)^{-1}Y\\ &= K(x,X)(K(X,X) + \frac{\lambda}{2}I)^{-1}Y\\ &= \hat{f}_{RLS}(x)\end{aligned}$$

What about Kernel RLS?

Model:

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

Then:

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

Estimated function?

$$\begin{aligned}\hat{f}_{MAP}(x) &= \phi(x)\hat{\theta}_{MAP}(Y|X) \\ &= \phi(x)\phi(X)^T(K(X,X) + \sigma_\varepsilon^2 I)^{-1}Y \\ &= K(x,X)(K(X,X) + \frac{\lambda}{2}I)^{-1}Y \\ &= \hat{f}_{RLS}(x)\end{aligned}$$

A prior over functions

Model:

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

Can we write this as a prior on $\mathcal{H}_K$?

A prior over functions

Remember *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:

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

A prior over functions

We showed: when $\theta \sim \mathcal{N}(0, I)$,

$$\hat{f}_{MAP}(\cdot) = \hat{\theta}_{MAP}(Y|X)^T \phi(\cdot) = \hat{f}_{RLS}(\cdot)$$

Taking $\phi(\cdot) = [\sqrt{\nu_1}\psi_1, \ldots, \sqrt{\nu_p}\psi_p]$, this prior is equivalently:

$$f(\cdot) = \theta^T \phi(\cdot) \sim \mathcal{N}(0, I)$$

i.e. the functions in $\mathcal{H}_K$ are Gaussian distributed:

$$p(f) \propto \exp\left(-\frac{1}{2}\|f\|^2_{\mathcal{H}_K}\right) = \exp\left(-\frac{1}{2}\theta^T\theta\right)$$

Note: again we need $\mathcal{H}_K$ to be finite-dimensional.

A prior over functions

So:

$$p(f) \propto \exp\left(-\frac{1}{2}\|f\|^2_{\mathcal{H}_K}\right) \Leftrightarrow \theta \sim \mathcal{N}(0, I) \Rightarrow \hat{f}_{MAP} = \hat{f}_{RLS}$$

Assuming $p(f) \propto \exp(-\frac{1}{2}\|f\|^2_{\mathcal{H}_K})$,

$$\begin{aligned} p(f|X,Y) &= \frac{p(Y|X,f)\cdot p(f)}{p(Y|X)} \\ &\propto \exp\left(-\frac{1}{2}\|Y - f(X)\|^2\right)\exp\left(-\frac{1}{2}\|f\|^2\right) \\ &= \exp\left(-\frac{1}{2}\|Y - f(X)\|^2 - \frac{1}{2}\|f\|^2\right) \end{aligned}$$

A quick recap

We wanted to know if RLS has a probabilistic interpretation.

- **Empirical risk minimization** is ML.

$$p(Y|X,\theta) \propto e^{-\frac{1}{2}\|Y - X\theta\|^2}$$

- **Linear RLS** is MAP.

$$p(Y,\theta|X) \propto e^{-\frac{1}{2}\|Y - X\theta\|^2} \cdot e^{-\frac{\lambda}{2}\theta^T\theta}$$

- **Kernel RLS** is also MAP.

$$p(Y,\theta|X) \propto e^{-\frac{1}{2}\|Y - \phi(X)\theta\|^2} \cdot e^{-\frac{\lambda}{2}\theta^T\theta}$$

- Equivalent to a Gaussian prior on $\mathcal{H}_K$:

$$p(Y,\theta|X) \propto e^{-\frac{1}{2}\|Y - f(X)\|^2} \cdot e^{-\frac{\lambda}{2}\|f\|^2_{\mathcal{H}_K}}$$

But these don't work for infinite dimensional function spaces...

A quick recap

We wanted to know if RLS has a probabilistic interpretation.

- **Empirical risk minimization** is ML.

$$p(Y|X,\theta) \propto e^{-\frac{1}{2}\|Y-X\theta\|^2}$$

- **Linear RLS** is MAP.

$$p(Y,\theta|X) \propto e^{-\frac{1}{2}\|Y-X\theta\|^2} \cdot e^{-\frac{\lambda}{2}\theta^T\theta}$$

- **Kernel RLS** is also MAP.

$$p(Y,\theta|X) \propto e^{-\frac{1}{2}\|Y-\phi(X)\theta\|^2} \cdot e^{-\frac{\lambda}{2}\theta^T\theta}$$

- Equivalent to a Gaussian prior on $\mathcal{H}_K$:

$$p(Y,\theta|X) \propto e^{-\frac{1}{2}\|Y-f(X)\|^2} \cdot e^{-\frac{\lambda}{2}\|f\|^2_{\mathcal{H}_K}}$$

But these don't work for infinite dimensional function spaces...

A quick recap

We wanted to know if RLS has a probabilistic interpretation.

- **Empirical risk minimization** is ML.

$$p(Y|X,\theta) \propto e^{-\frac{1}{2}\|Y-X\theta\|^2}$$

- **Linear RLS** is MAP.

$$p(Y,\theta|X) \propto e^{-\frac{1}{2}\|Y-X\theta\|^2} \cdot e^{-\frac{\lambda}{2}\theta^T\theta}$$

- **Kernel RLS** is also MAP.

$$p(Y,\theta|X) \propto e^{-\frac{1}{2}\|Y-\phi(X)\theta\|^2} \cdot e^{-\frac{\lambda}{2}\theta^T\theta}$$

- Equivalent to a Gaussian prior on $\mathcal{H}_K$:

$$p(Y,\theta|X) \propto e^{-\frac{1}{2}\|Y-f(X)\|^2} \cdot e^{-\frac{\lambda}{2}\|f\|^2_{\mathcal{H}_K}}$$

But these don't work for infinite dimensional function spaces...

A quick recap

We wanted to know if RLS has a probabilistic interpretation.

- **Empirical risk minimization** is ML.

$$p(Y|X,\theta) \propto e^{-\frac{1}{2}\|Y-X\theta\|^2}$$

- **Linear RLS** is MAP.

$$p(Y,\theta|X) \propto e^{-\frac{1}{2}\|Y-X\theta\|^2} \cdot e^{-\frac{\lambda}{2}\theta^T\theta}$$

- **Kernel RLS** is also MAP.

$$p(Y,\theta|X) \propto e^{-\frac{1}{2}\|Y-\phi(X)\theta\|^2} \cdot e^{-\frac{\lambda}{2}\theta^T\theta}$$

- Equivalent to a Gaussian prior on $\mathcal{H}_K$:

$$p(Y,\theta|X) \propto e^{-\frac{1}{2}\|Y-f(X)\|^2} \cdot e^{-\frac{\lambda}{2}\|f\|^2_{\mathcal{H}_K}}$$

But these don't work 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*:

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

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

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

Transductive setting

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 $\mathcal{H}_K$?

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:

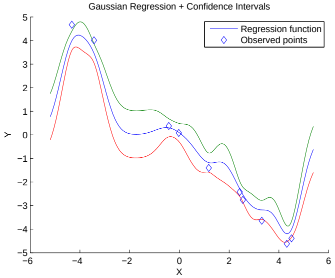
- $Y = f(X) + \varepsilon$,
- $Y^* = f(X)$, and
- $f \in \mathcal{H}_K$ is a random trajectory from a **Gaussian process** over the domain, with mean μ and covariance K.
- (Ergo, people call this "Gaussian process regression.") (Also "Kriging," because of a guy.)

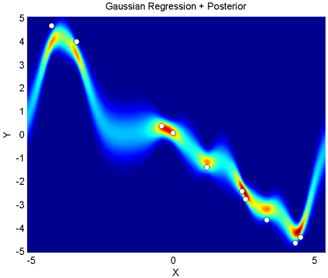
Recap redux

- **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?





What is going on with infinite-dimensional $\mathcal{H}_K$?

Wrote down statements like: $\theta \sim \mathcal{N}(0, I)$ and $f \sim \mathcal{N}(0, I)$. The space $\mathcal{H}_K$ can be written

$$\mathcal{H}_K = \{f : \|f\|^2_{\mathcal{H}_K} < \infty\} = \{\theta^T \phi(\cdot) : \sum_{i=1}^{\infty} \theta_i^2 < \infty\}$$

Difference between finite and infinite: not every θ yields a function $\theta^T \phi(\cdot)$ in $\mathcal{H}_K$.

A hint: $\theta \sim \mathcal{N}(0, I) \Rightarrow \mathbb{E}\|\theta^T \phi(\cdot)\|^2_{\mathcal{H}_K} = \infty$.
In fact: $\theta \sim \mathcal{N}(0, I) \Rightarrow \theta^T \phi(\cdot) \in \mathcal{H}_K$ *with probability zero*.

So be careful out there.

What is going on with infinite-dimensional $\mathcal{H}_K$?

Wrote down statements like: $\theta \sim \mathcal{N}(0, I)$ and $f \sim \mathcal{N}(0, I)$. The space $\mathcal{H}_K$ can be written

$$\mathcal{H}_K = \{f : \|f\|^2_{\mathcal{H}_K} < \infty\} = \{\theta^T \phi(\cdot) : \sum_{i=1}^{\infty} \theta_i^2 < \infty\}$$

Difference between finite and infinite: not every θ yields a function $\theta^T \phi(\cdot)$ in $\mathcal{H}_K$.

A hint: $\theta \sim \mathcal{N}(0, I) \Rightarrow \mathbb{E}\|\theta^T \phi(\cdot)\|^2_{\mathcal{H}_K} = \infty$.

In fact: $\theta \sim \mathcal{N}(0, I) \Rightarrow \theta^T \phi(\cdot) \in \mathcal{H}_K$ *with probability zero*.

So be careful out there.

What is going on with infinite-dimensional $\mathcal{H}_K$?

Wrote down statements like: $\theta \sim \mathcal{N}(0, I)$ and $f \sim \mathcal{N}(0, I)$. The space $\mathcal{H}_K$ can be written

$$\mathcal{H}_K = \{f : \|f\|^2_{\mathcal{H}_K} < \infty\} = \{\theta^T \phi(\cdot) : \sum_{i=1}^{\infty} \theta_i^2 < \infty\}$$

Difference between finite and infinite: not every θ yields a function $\theta^T \phi(\cdot)$ in $\mathcal{H}_K$.

A hint: $\theta \sim \mathcal{N}(0, I) \Rightarrow \mathbb{E}\|\theta^T \phi(\cdot)\|^2_{\mathcal{H}_K} = \infty$.
In fact: $\theta \sim \mathcal{N}(0, I) \Rightarrow \theta^T \phi(\cdot) \in \mathcal{H}_K$ *with probability zero*.

So be careful out there.

A hint that things are amiss

Assume: $\theta \sim \mathcal{N}(0, I)$, $\{\phi_i\}_{i=1}^{\infty}$ orthonormal basis in $\mathcal{H}_K$.

$$
\begin{aligned}
\mathbb{E}\|\theta^T \phi\|_{\mathcal{H}_K}^2 &= \mathbb{E}\|\sum_{i=1}^{\infty} \theta_i \phi_i\|_{\mathcal{H}_K}^2 \\
&= \mathbb{E}\sum_{i=1}^{\infty} \theta_i^2 \\
&= \sum_{i=1}^{\infty} 1 \\
&= \infty
\end{aligned}
$$