If machine learning is to be of any use, can it...

Sai Ravela
Learn Physics?

• Let’s re-examine the Poisson Equation problem

\[ \nabla^2 \phi = f \]

On a periodic domain
In 1D

\[ \phi = D^{-1}f \]

\[ \phi = Lf \]

Machine Learning problem:
Estimate \( D, L \)
Given: \( \{\phi_n, f_n\}, \) the training data

Sound Easy?
Step 1

\[ J(L^\#) := \frac{1}{2N} \sum_{i=1}^{N} \| \phi - L^\# f \|_2^2 \]

How many samples?  Why L_2?  Matrix form of operator \( L \)

\( L^\# \) degrees of freedom: \( n^2 \)
Step 2, OLS

- \( J(L^o) := \frac{1}{2N} \sum_{i=1}^{N} \| \phi_i - L^o f_i \|_2^2 \)

- \( L^o \) has \( n \) degrees of freedom; block circulant matrix for our problem

- Oracle:
  - Let’s invoke invariance as a principle
  - \( L^\# \Rightarrow L^o \)

\[
L^o = \begin{pmatrix}
l_1 & \cdots & l_n \\
\vdots & \ddots & \vdots \\
l_2 & \cdots & l_1
\end{pmatrix}
-- \text{special Toeplitz matrix}
Equivalently

- \( J(l) \overset{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} J_i(l) \);

- \( \Rightarrow \hat{l} = (\sum_{i=1}^{N} F_i^T F_i)^{-1} \left( \sum_{j=1}^{N} F_j^T \phi_j \right) \)

- Where, \( l = \begin{bmatrix} l_1 \\ \vdots \\ l_n \end{bmatrix} \) and \( F_i = \begin{pmatrix} f_{i,1} & \cdots & f_{i,n} \\ \vdots & \ddots & \vdots \\ f_{i,2} & \cdots & f_{i,1} \end{pmatrix} \)

- \( J_i(l) \overset{\text{def}}{=} \frac{1}{2} \left\| \phi_i - F_i l \right\|^2 \)

- Local solution is: \( \hat{l}_i = (F_i^T F_i)^{-1} F_i^T \phi_i \)

- Can estimate \( \hat{l} = (F_i^T F_i)^{-1} F_i^T \phi_i \) work?

What is the approximation?
What kind of training data?
How many samples N/n?
Regularization

\[ J_i(l) \overset{\text{def}}{=} \frac{1}{2} \| \phi_i - F_i l \|^2_2 + \frac{\lambda}{2} \| l \|^2_2 \]

\[ \hat{l}_i = (F_i^T F_i + \lambda I)^{-1} F_i^T \phi_i \]

\[ F_i = U_i S_i U_i^T \]

\[ U_{\#i} \overset{\text{def}}{=} U_i(:, 1:S < N \ll n) \]

• Leading S directions

\[ \hat{l}_i = U_{\#i} [S_{\#i}^2 + \lambda I]^{-1} S_{\#i} U_{\#i}^T \phi_i \]

Why L_2 ?

What does \( \lambda \) do?

Why this form? What are they?

How many “leading” dimensions?

Try Matlab Code
Recursion

• $J(l) \overset{\text{def}}{=} \frac{1}{2} \sum_{i=1}^{N} \| \phi_i - F_i l \|_2^2$

• $Q_1 = F_1^T F_1$ and $\hat{l}_1 = [F_1^T F_1]^{-1} F_1^T [\phi_1]$

• $\hat{l}_{i+1} = \hat{l}_i + [Q_i + F_{i+1}^T F_{i+1}]^{-1} F_{i+1}^T [\phi_{i+1} - F_{i+1} \hat{l}_i]$

• $Q_{i+1} = Q_i + F_{i+1}^T F_{i+1}$ (cumulant;)

• Assignment 1: Derive this result.
• Compare with batch solution
• Compare with naïve average
• Optional: Compare with Kalman Filter
Estimation Beyond OLS

- Maximum Likelihood Estimation (MLE)
  
  Let \( J_i(l) \) \( \text{def} \frac{1}{2} \left\| \phi_i - F_i l \right\|^2_R = \frac{1}{2} \left[ \phi_i - F_i l \right]^T R^{-1} \left[ \phi_i - F_i l \right] \)

  then \( \hat{l}_i = (F_i^T R^{-1} F_i)^{-1} F_i^T R^{-1} \phi_i \)

- Maximum A Posteriori Estimation (MAP):
  
  Let \( J_i(l) \) \( \text{def} \frac{1}{2} \left\| \phi_i - F_i l \right\|^2_R + \frac{1}{2} \left\| l - \bar{l} \right\|^2_C \)

  \( \hat{l}_i = (F_i^T R^{-1} F_i + C^{-1})^{-1} [F_i^T R^{-1} \phi_i + C^{-1} \bar{l}] \)

- Can you combine this with regularization? Truncation?

- Why are these called MAP and MLE? – Next class from a Bayesian perspective.
The best result did not come by
- Increasing sample size
- Using L2 regularization

It immediately emerged when we chose a “subspace”

The Kernel ($\nabla^2$) is sparse, so $L^0$ is non-sparse. $U$ is the Fourier Transform (in the continuous case), and therefore truncating $U_#$ projects the residual onto very low frequencies, promoting sparse $D$.

What we assumed was
- Linearity, Invariance, small magnitude perturbations.