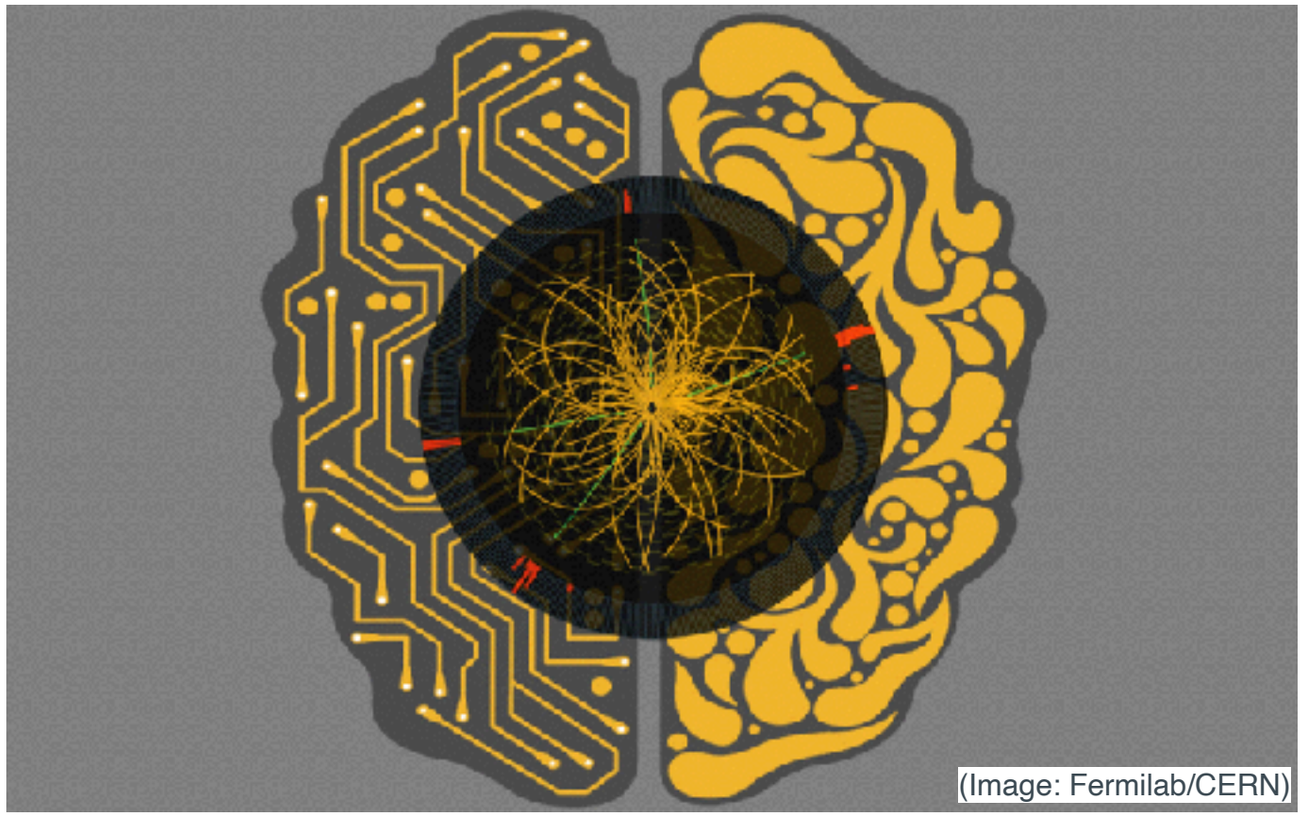
PHY 835: Collider Physics Phenomenology

Machine Learning in Fundamental Physics

Gary Shiu, UW-Madison



Lecture 4: Linear Regression

Recap of Lecture 3

- What is Gradient Descent?
- Comparing gradient descent vs Newton's method
- Limitations of Gradient Descent
- Stochastic Gradient Descent
- How can it be modified? E.g. adding momentum
- Second order methods (RMSProp and ADAM)

Outline for today

- Linear Regression
- Least Square regression Regularization
- Ridge regression
- Lasso regression
- MLE and MAP
- Linear Regression on 1D Ising model

References: 1803.08823, chapter 5 and 7 Goodfellow et al.

Linear Regression

Setting up the Problem

- Given a dataset: $\{(y_i, \mathbf{x}^{(i)}), i = 1, \ldots, n\}$, $\mathbf{x}^{(i)} \in \mathbb{R}^p$

response **observation vector** **p features**

- Assume the true function/model that generates these samples:

$$y_i = f(\mathbf{x}^{(i)}; \omega_{\text{true}}) + \epsilon_i$$

i.i.d. white noise with zero mean and finite variance

- Compactly cast all samples into an $X \in \mathbb{R}^{n \times p}$ **design matrix**:

$$X = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdot & \cdot & x_p^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdot & \cdot & x_p^{(2)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_1^{(n)} & x_2^{(n)} & \cdot & \cdot & \cdot\, x_p^{(n)} \end{pmatrix}$$

Setting up the Problem

- The function $f(\mathbf{x}^{(i)}; \omega_{true})$ is never known to us explicitly. For linear regression, we assume:

$$y_i = f(\mathbf{x}^{(i)}) + \epsilon_i = \hat{\omega}_{\text{true}}^T \mathbf{x}^{(i)} + \epsilon_i$$

- Replace $\mathbf{x}^{(i)}$ by $\phi(\mathbf{x}^{(i)})$ and $\omega^T \mathbf{x}^{(i)}$ by $\omega^T \phi(\mathbf{x}^{(i)})$: basis function expansion.

- **Goal:** find $g(\mathbf{x}^{(i)}; \hat{\omega})$ known as **predictor** which best approximates f.

- For later purposes, define the L^k norm ($1 \leq k \in \mathbb{Z}$) of a vector $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ as

$$||\mathbf{x}||_k = \left(|x_1|^k + |x_2|^k + \ldots |x_d|^k \right)^{1/k}$$

Least Square Regression

- Ordinary least squares linear regression (OLS):

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} ||\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}||_2^2 = \min_{\boldsymbol{w}\in\mathbb{R}^p} \sum_{i=1}^{n} (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^{(i)} - y_i)^2.$$

- The solution denoted by $\hat{\boldsymbol{w}}_{\mathrm{LS}} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^p} ||\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}||_2^2$, can be obtained by differentiation:

$$0 = \frac{\partial}{\partial_m}(X_{ij}\omega_j - y_i)(X_{ik}\omega_k - y_i) = 2(X_{ij}\omega_j - y_i)X_{im}$$

$$\Longrightarrow \qquad \hat{\boldsymbol{w}}_{\mathrm{LS}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}. \qquad \text{(if } \mathbf{X}^T\mathbf{X} \text{ is invertible)}$$

- If $\mathrm{rank}(\mathbf{X}) = p$, $\mathbf{X}^T\mathbf{X}$ is invertible and $\hat{\omega}_{LS}$ is unique.

- If $\mathrm{rank}(\mathbf{X}) < p$, $\mathbf{X}^T\mathbf{X}$ is singular, and $\hat{\omega}_{LS}$ has infinitely many solutions:

$$\omega_0 + \eta \text{ where } \mathbf{X}\eta = 0 \qquad \rightarrow \text{pick one solution}$$

OLS Performance

- One can show (experiment with the Juypter notebooks):

$$\bar{E}_{\text{in}} = \sigma^2 \left(1 - \frac{p}{n}\right)$$
$$\bar{E}_{\text{out}} = \sigma^2 \left(1 + \frac{p}{n}\right)$$

- **Average generalization error**:

$$|\bar{E}_{\text{in}} - \bar{E}_{\text{out}}| = 2\sigma^2 \frac{p}{n}$$

- If $p \gg n$ (higher dim. data), generalization error is very large meaning that the model is not learning.

- Even if $p \approx n$, still may not learn well due to the **intrinsic noise**.

- Can we do better? → **Regularization**

Ridge Regression

- Add L^2 norm of the parameter vector as penalty in loss function:

$$\hat{\boldsymbol{w}}_{\text{Ridge}}(\lambda) = \underset{\boldsymbol{w}\in\mathbb{R}^p}{\arg\min} \left(||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2 + \lambda ||\boldsymbol{w}||_2^2 \right).$$

$$\Leftrightarrow \quad \hat{\boldsymbol{w}}_{\text{Ridge}}(t) = \underset{\boldsymbol{w}\in\mathbb{R}^p:\ ||\boldsymbol{w}||_2^2 \le t}{\arg\min} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2.$$

- **Regularizer:** effectively constraining the magnitude of parameters.

- Solving this constrained minimization problem by differentiation:

$$0 = 2(X_{ij}\omega_j - y_i)X_{im} + 2\lambda\omega_m$$

$$\Rightarrow \quad \hat{\boldsymbol{w}}_{\text{Ridge}}(\lambda) = (\boldsymbol{X}^{\text{T}}\boldsymbol{X} + \lambda I_{p\times p})^{-1}\boldsymbol{X}^{\text{T}}\boldsymbol{y} \overset{\uparrow}{=} \frac{\hat{\boldsymbol{w}}_{\text{LS}}}{1+\lambda},$$

orthogonal X

Ridge Regression

- What is the relation between $\hat{y}_{Ridge}$ and $\hat{y}_{LS}$?

- **Singular value decomposition (SVD):**

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}} \qquad \text{where} \qquad \begin{array}{l} \boldsymbol{U} \in \mathbb{R}^{n\times p} \text{ and } \boldsymbol{V} \in \mathbb{R}^{p\times p} \\ \boldsymbol{D} \in \mathbb{R}^{p\times p} = \mathrm{diag}(d_1, d_2, \ldots, d_p) \end{array}$$

- U and V are **(semi)-orthogonal** matrices:

$$\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{1}\,, \quad \text{but only } \mathbf{U}^T\mathbf{U} = 1 \text{ since } p \leq n$$

- The diagonal values of D:

$$d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$$

- X is **singular** if at least one $d_j \geq 0$.

Ridge Regression

- Recast the Ridge estimator:

$$\hat{\boldsymbol{w}}_{\text{Ridge}} = \boldsymbol{V}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^{\text{T}}\boldsymbol{y}$$

- The Ridge predictor is then:

$$\begin{aligned}
\hat{\boldsymbol{y}}_{\text{Ridge}} &= \boldsymbol{X}\hat{\boldsymbol{w}}_{\text{Ridge}} \\
&= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^{\text{T}}\boldsymbol{y} \\
&= \sum_{j=1}^{p} \boldsymbol{U}_{:,j} \frac{d_j^2}{d_j^2 + \lambda} \boldsymbol{U}_{:j}^{\text{T}}\boldsymbol{y} \\
&\leq \boldsymbol{U}\boldsymbol{U}^{\text{T}}\boldsymbol{y}
\end{aligned}$$

$$\begin{aligned}
\hat{y}_{LS} &= \mathbf{X}\hat{\omega}_{LS} \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{U}\mathbf{U}^T\mathbf{y}
\end{aligned}$$

- Both regressions project $\mathbf{y}$ to the column space of $\mathbf{X}$. Ridge regression shrinks each basis component by a factor of $d_j^2/(d_j^2 + \lambda)$.

LASSO Regression

- Add L^1 norm of the parameter vector as penalty in loss function:

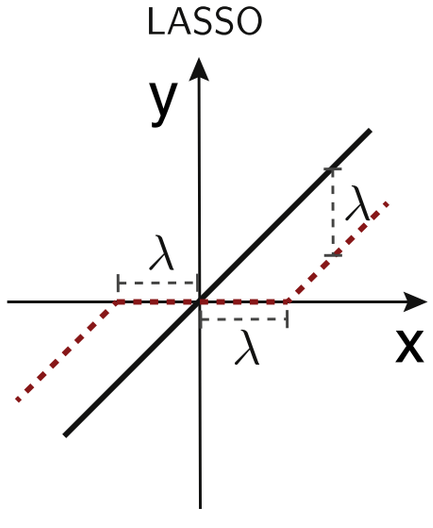
$$\hat{\boldsymbol{w}}_{\text{LASSO}}(\lambda) = \underset{\boldsymbol{w}\in\mathbb{R}^p}{\arg\min} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2 + \lambda ||\boldsymbol{w}||_1$$

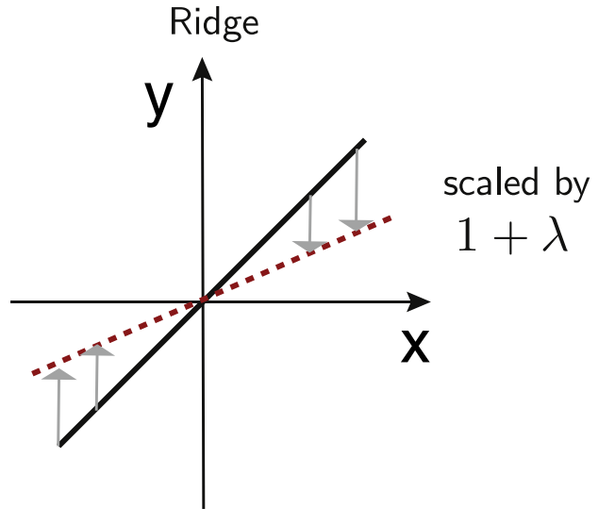
$$\Leftrightarrow \qquad \hat{\boldsymbol{w}}_{\text{LASSO}}(t) = \underset{\boldsymbol{w}\in\mathbb{R}^p:\ ||\boldsymbol{w}||_1 \le t}{\arg\min} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2$$

- The L^1 regularizer is not everywhere differentiable so analytic solution is harder, but LASSO is a convex problem ($\Rightarrow$ optimization).

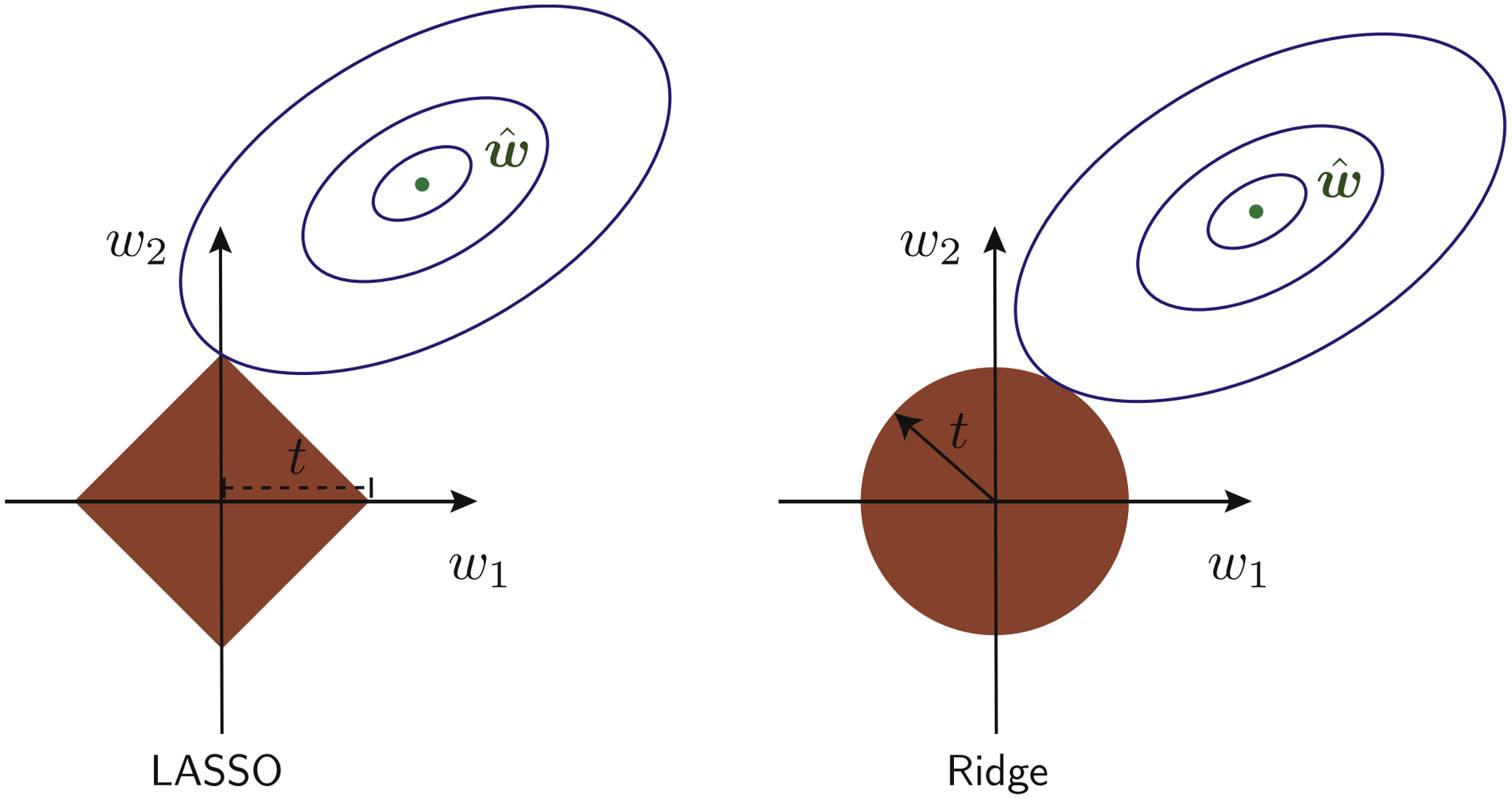
- For orthogonal X: $\hat{w}_j^{\text{LASSO}}(\lambda) = \text{sign}(\hat{w}_j^{\text{LS}})(|\hat{w}_j^{\text{LS}}| - \lambda)_+$

"soft-thresholding"





LASSO vs Ridge



LASSO gives sparse solutions: many components of $\hat{\omega}_{LASSO}$ are zero.

Bayesian Formulation of Linear Regression

Bayesian Formulation

- Formulate least square regression from a Bayesian point of view.
- Regularization corresponds to a choice of prior.
- A regression model is defined by a **conditional probability**:

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x})).$$

- For linear regression:

$$\mu = \boldsymbol{x}^T \boldsymbol{w}, \qquad \sigma^2(\boldsymbol{x}) = \sigma^2, \text{ then } \boldsymbol{\theta} = (\boldsymbol{w}, \sigma^2).$$

- **Maximum likelihood estimation (MLE)** for θ is the one that minimizes the **mean square error** used in OLS.

Bayesian Formulation

- Maximizing the log likelihood:

$$\hat{\boldsymbol{\theta}} \equiv \arg\max_{\boldsymbol{\theta}} \log p(\mathcal{D}|\boldsymbol{\theta})$$

- Assuming that the samples are i.i.d.:

$$l(\boldsymbol{\theta}) \equiv \log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(y_i|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}).$$

- Using the Bayesian representation of $p(\mathbf{y} \mid \mathbf{x}, \theta)$:

$$\begin{aligned} l(\boldsymbol{\theta}) &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}^{(i)}\right)^2 - \frac{n}{2} \log\left(2\pi\sigma^2\right) \\ &= -\frac{1}{2\sigma^2} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2 + \text{const.} \end{aligned}$$

Maximum a Posteriori Probability (MAP)

- **Bayes' rule:**

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int \mathrm{d}\boldsymbol{\theta}'\, p(\boldsymbol{X}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}.$$

- MAP amounts to maximizing the log posteriori:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} \equiv \arg\max_{\boldsymbol{\theta}} \log p(\mathcal{D}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}).$$

- Consider a Gaussian distribution for the prior: $p(\boldsymbol{w}) = \prod_j \mathcal{N}(w_j|0, \tau^2)$

- **Ridge regression "=" Putting Gaussian prior on weights:**

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} &\equiv \arg\max_{\boldsymbol{\theta}} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^{(i)})^2 - \frac{1}{2\tau^2} \sum_{j=1}^{n} w_j^2 \right] \\ &= \arg\max_{\boldsymbol{\theta}} \left[-\frac{1}{2\sigma^2} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2 - \frac{1}{2\tau^2} ||\boldsymbol{w}||_2^2 \right].\end{aligned}$$

with hyperparameter λ in the regularizer corresponds to $\lambda \equiv \sigma^2/\tau^2$:

Example: 1D Ising Model

- Ensemble of spin configurations and their energy generated from:

$$H = -J \sum_{j=1}^{L} S_j S_{j+1} \qquad S_j \in \{\pm 1\} \qquad \mathcal{D} = (\{S_j\}_{i=1}^{L}, E_j)$$

- **Goal:** to learn a model that predicts E_j from the spin configurations.

- **Ansatz:** pairwise interactions

$$H_{\text{model}}[S^i] = -\sum_{j=1}^{L}\sum_{k=1}^{L} J_{j,k} S_j^i S_k^i,$$

- This problem can be cast as a linear regression problem:

$$H_{\text{model}}[S^i] = \mathbf{X}^i \cdot \mathbf{J}.$$

$\mathbf{X}^i$: $\{S_j^i S_k^i\}_{j,k=1}^{L}$

Example: 1D Ising Model

- How can we measure performance?

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left| y_i^{\text{true}} - y_i^{\text{pred}} \right|^2}{\sum_{i=1}^{n} \left| y_i^{\text{true}} - \frac{1}{n} \sum_{i=1}^{n} y_i^{\text{pred}} \right|^2}.$$

$R^2 = 1$ best performance

$R^2 < 0$ possible

- We want to compare Mean Square Error, LASSO & Ridge regression

- Experiment with the Juypter notebook:

https://physics.bu.edu/%7Epankajm/MLnotebooks.html

Example: 1D Ising Model

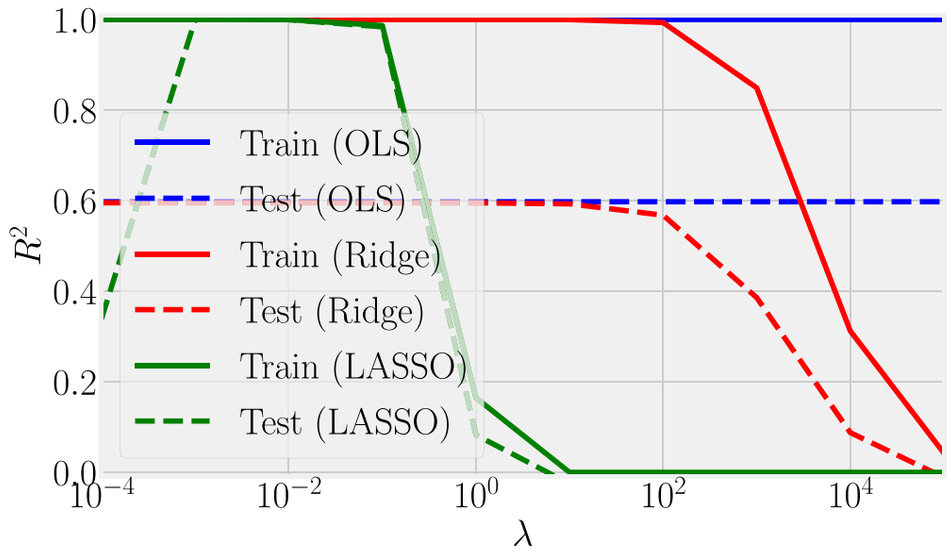
- Performance depends on hyperparameter λ. Tuning λ is known as **hyperparameter tuning.**

- There can be **optimal values** for λ

- Observed different solutions for Ridge and LASSO.

- Using regularizer can lead to better results.

- Regularization restricts parameter space (less complex model class).



Summary

- Linear regression
- Regularization (Ridge, LASSO)
- MLE
- MAP
- Relation of MLE and MAP with Least/Square and Ridge regression
- Linear regression will be replaced by more complicated/non-linear models
- Regression on the 1D Ising model