The Framework of Statistical Learning

- $X \subset \mathbb{R}^k$: domain set (e.g. space of images with fixed nr. of pixels)
- $Y \subset \mathbb{R}^9$: target set / set of label (e.g. $Y = \{0,1\}$ binary classification)
- Let $\mathcal{D}$ be a probability distribution on $X \times Y$ $Y = \mathbb{R}^9$: regression
- Assume we are given training set $S = \{(x_i, y_i)_{i=1}^n\}$ $\overset{i.i.d}{\sim} \mathcal{D}$
- Goal: Find a predictor function $h : X \to Y$ that minimizes the expected risk $L_{\mathcal{D}}(h) := \mathbb{E}_{\mathcal{D}}[\ell(Y, h(X))]$ where $\ell : Y \times Y \to Z$ is a given loss/error function

Learning Algorithm:
Specific algorithm that maps S to a specific member function $h_n \in \mathcal{F}$ of a hypothesis space $\mathcal{F} \subset \{h : X \to Y\}$ based on the information of S.

In many cases: $\hat{h}_n = \underset{h \in \mathcal{F}}{\operatorname{argmin}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i))}_{\text{empirical risk}}$

Empirical Risk Minimization

Ex: ▷ Linear Regression: ▷ $\ell(y, z) = \frac{1}{2}(y - z)^2$

▷ $\mathcal{F} = \{ x \longmapsto \beta_0 + \langle x, \beta \rangle \mid \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^k \}$

Bias - Variance Tradeoff:

$$L_{\mathcal{D}}(\hat{h}_n) - \min_h L_{\mathcal{D}}(h) = \underbrace{\left(L_{\mathcal{D}}(\hat{h}_n) - L_{\mathcal{D}}(h_{\mathcal{F}}) \right)}_{\text{estimation error}} + \underbrace{\left(L_{\mathcal{D}}(h_{\mathcal{F}}) - \min_h L_{\mathcal{D}}(h) \right)}_{\text{approximation error}}$$

generalization error

depends on training set S

$h_{\mathcal{F}} = \underset{h \in \mathcal{F}}{\operatorname{argmin}} L_{\mathcal{D}}(h)$

$\mathcal{F}$: hypothesis space

approximation error

▷ smaller if $\mathcal{F}$ is large

$$\hat{h}_n = \underset{h \in \mathcal{F}}{\operatorname{argmin}} \left(\underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i))}_{\text{empirical risk}} + \lambda \underbrace{R(h)}_{\substack{\text{a term that} \\ \text{quantifies "complexity"} \\ \text{of } h \in \mathcal{F}}} \right)$$

$\lambda > 0$ "regularization parameter"

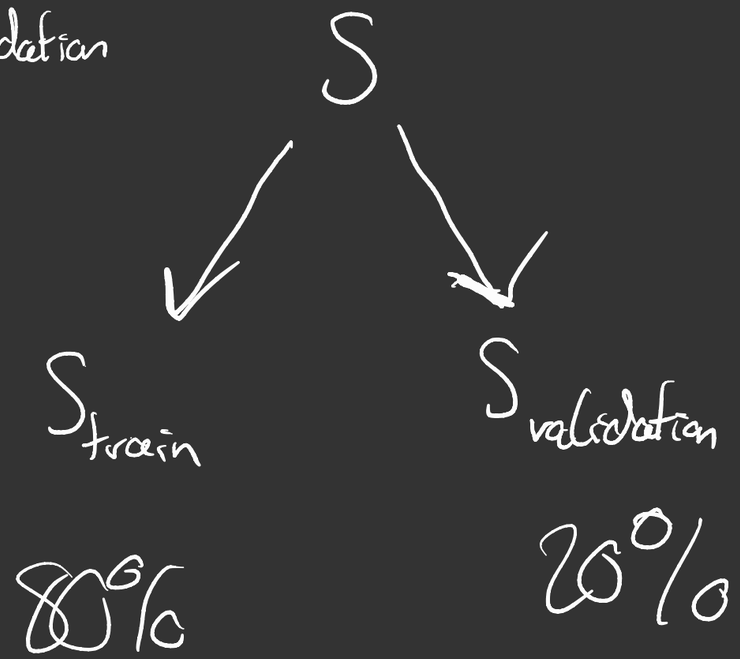
"Regularized loss minimization" (RLM)

Ridge Regression: Linear Regression with $R(h) = \|\beta\|_{\ell_2}^2$

$$\hat{\beta}_n = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} \left\{ \frac{1}{n} \|X\beta - y\|_{\ell_2}^2 + \lambda \|\beta\|_{\ell_2}^2 \right\}$$

- If $\lambda = 0$: $\longrightarrow$ Linear regression
- If $\lambda \longrightarrow +\infty$: coefficients $\hat{\beta}_i$ "shrinked to 0"
- λ in between: Balances fit of linear model to S & size of coefficient vector β

Cross validation



k-fold Cross validation

Preprocessing: If domain set s.t. $X \subset \mathbb{R}^k$, we can define a feature map $\phi: X \to \tilde{X} \subset \mathbb{R}^l$ (often with $k \ll l$) such that $S = (\phi(x_i), y_i)_{i=1}^n$ is used as training set.

Ex: Polynomial features. E.g. if $k=1$, $l=5$:

$$\phi(x) = (x, x^2, x^3, x^4, x^5).$$

$\rightarrow$ Often **improves** expressive power of a learning model!

Sparse Regression

$$\mathcal{F}_k^{\text{sparse}} = \{ h : \mathbb{R}^d \to \mathbb{R} : h(x) = \beta_0 + \langle \beta, x \rangle ; \text{ s.t. } \|\beta\|_0 \leq k \}$$

$$= \sum_{i=1}^{n} \mathbb{1}_{\{|\beta_i| \neq 0\}}(\beta)$$

Idea: Only $k << d$ coefficients of linear model will suffice to explain/predict the outcome $\quad k << d$

Problem: ERM on $\mathcal{F}_k^{\text{sparse}}$ is NP-hard! $\quad = \sum_{i=1}^{n} |\beta_i|$

↳ computational challenges!

- Possible approach: Lasso Regression:

$$\hat{\beta}_n = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{n} \|AB - y\|_2^2 + \lambda \cdot \|\beta\|_1 \right\}$$

regularization parameter, $\lambda > 0$

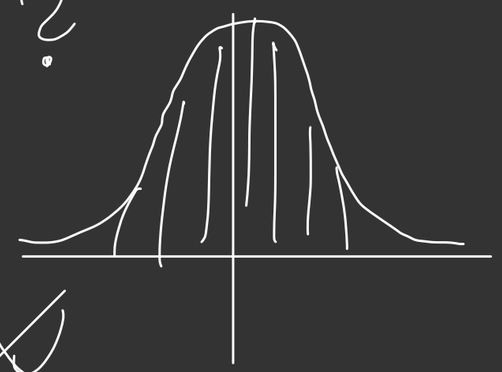
- Computationally, Lasso Regression is
  - more expensive than Ridge or LR, but
  - still convex optimization problem, so polynomial (and efficient) algorithms exist.

- Since Lasso Regression only approximates solution of ERM on $\mathcal{F}_k^{\text{sparse}}$: In addition to estimation and approximation errors, we now also have an optimization error.

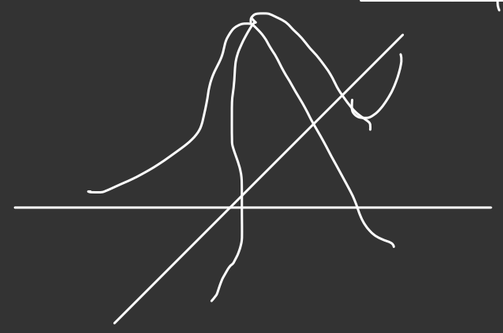
High Dimensional Geometry

Q: How do $x_1, x_2 \in \mathbb{R}^d$, $d >> 1$ relate to each other when "generic"?

- 1-dim Gaussian:



$d = 2$



- d-dim Gaussian:

$$\|X\| \sim \sqrt{d} \pm O(1)$$

$\Rightarrow$ All very far from origin!

- If $x_1, x_2 \in \mathbb{R}^d$ indep. d-dim Gaussian

$\Rightarrow \|x_1 - x_2\|_2 \sim$ close to const. & large.