Week 3: Least Squares / Ridge Regr. MLE / over-underfitting

Least Squares:

we have seen the linear regression model with the MSE objective function: $f_w(x) = Xw = y$

and

$$\mathcal{L}(w) = \frac{1}{2N}\sum_{1}^{N}(y_n - x_n^T w) = \frac{1}{2N}\|Y - Xw\|^2 = \frac{1}{2N}e^T e$$

$$\nabla\mathcal{L}(w) = -\frac{1}{N}X^T e$$

- we have seen that we can find w^* with gradient methods (GD, SGD, ...) but this problem has also a closed form solution.

we know by the first-order optimality principle:

if $\mathcal{L}(w)$ is convex then $\nabla\mathcal{L}(w^*) = 0$

given that $\mathcal{L}(w)$ is convex (sum of convex fcts $(\,)^2$)

then $\nabla\mathcal{L}(w^*) = 0 \Leftrightarrow -\frac{1}{N}X^T(y - Xw^*) = 0$

$$\Leftrightarrow X^T X w^* = X^T y$$

(normal equations) $\Leftrightarrow$ $w^* = (X^T X)^{-1} X^T y$

geom. Interp.: $X^T \cdot e = 0$ with X^T rows being the feature vectors

- each feature is $\perp$ to the e
- w^* is given by Proj of y on span(X)

Note: $X^T X$ is invertible iff rank$(X) = D$ (full column rank)

else however, X is ill-cond. and rank deficient (colinear.)

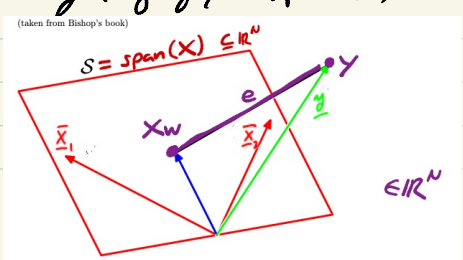
→ use linear solver

Cond. number $= \frac{\lambda_{max}}{\lambda_{min}} \to +\infty$ if ill-cond, $\to 0$

Linear Regression Summary:

- Grid Search
- Iterative Optim. Alg. (GD, SGD ...) $O(ND)$ or $O(D)$ per iteration
- Least Squares (closed form $\mathcal{L}$) $\gg O(D^2N + D^3)$

(taken from Bishop's book)



Maximum Likelihood:

Probabilistic approach for Lin. Reg.

We assume that our data is generated by:

$$y_n = x_n^T w + \underbrace{\varepsilon_n}_{\text{noise}} \quad \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$$

→ given N inputs, the likelihood of y given X:

$$P(y \mid X, w) \overset{\text{iid}}{=} \prod_{n=1}^{N} P(y_n \mid x_n, w)$$

$$(y_n = x_n^T w + \varepsilon_n) \quad = \prod_{n=1}^{N} \mathcal{N}(y_n \mid x_n^T w, \sigma^2)$$

⇒ maximize this likelihood

⇒ maximize the log likelihood

$$\mathcal{L}_{LL}(w) = \log P(y \mid X, w)$$

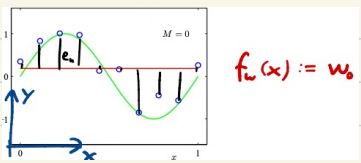
$$= -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - x_n^T w)^2 + \text{const}$$

$$\left(\mathcal{L}_{MSE}(w) = \frac{1}{2N} \sum_{n=1}^{N} (y_n - x_n^T w)^2\right)$$

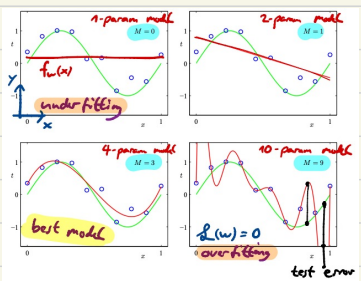
$$\min_w \mathcal{L}_{MSE}(w) = \max_w \mathcal{L}_{LL}(w) = w^*$$

MLE is consistent: $w_{MLE} \to^p w_{True}$

Over- Under fitting :



$f_w(x) := w_0$



A common problem in ML is underfitting and overfitting.

Assume we observe a noisy version of samples of some fct $g(x)$:

$y_n = g(x) + z_n$, where z_n is the noise. (measurement, outlier...)

- a simple linear model (1-param) may overly underfit.

→ One way to overcome this is feature expansion

$\phi(x_n) = \{1, x_n, x_n^2, \dots, x_n^M\}$

$\sin(x_n), \log(x_n)$ ...

- add polynomial basis

M: model complexity

$f_w(x) = f_w(\phi(x))$ (Linear model on "augmented" input)

$= \phi(x)\, w$

→ To avoid overfitting, we can increase the data ($\nearrow N$, M fixed)

overfitting might reduce.

We will see other ways too (Regularization: Ridge Reg.)

Regularization: Ridge Regression & Lasso

We have seen that by augmenting the feature vectors we can make linear models as powerful as we want. However this leads to the problem of overfitting. Regularization is one way to mitigate this undesirable behaviour.

(Occam's razor: simpler models are preferred)

Regularization: penalize complex models

$$\min_w \mathcal{L}(w) + \Omega(w)$$

$\mathcal{L}$: fit; Ω: complexity of model w; Ω: regularizer

L_2-Regularization: (most frequent)

$$\Omega(w) = \lambda \|w\|_2^2$$

λ: trade-off parameter, $\lambda > 0$

- $\lambda \to 0$: no regul. → potential overfit
- $\lambda \to \infty$: underfitting, $\|w\| \to 0$

- When $\mathcal{L}(w)$ is MSE this is called:

Ridge Regression: $\min_w \frac{1}{2N} \sum (y_n - x_n^T w)^2 + \lambda \|w\|_2^2$

($\lambda = 0 \to$ Least Sq.-s.)

→ Explicit Solution:

$$w^* = (X^T X + \lambda' I)^{-1} X^T y \quad (\lambda' = 2N\lambda)$$

First-Order optimality
- $\mathcal{L}$ convex
- $\nabla \mathcal{L}(w) = 0 \Rightarrow w$ optimal

Ridge Regression to fight ill-conditioning: Lifting the eigenvalues

$(X^T X + \lambda' I)$ eigenvalues are at least $\lambda' > 0$ so the inverse exists.

L_1-Regularization: (The Lasso) $\left(\|w\|_1 = \sum_{j=1}^{D} |w_j|\right)$

In combination with MSE for lin. Reg this is called **the Lasso**:

$$\min_w \frac{1}{2N} \sum_{1}^{N} (y_n - x_n^T w)^2 + \lambda \|w\|_1$$

penalized objective

$\Rightarrow$ this encourages sparse w^*