IEMS 402 Statistical Learning - 2025 Winter
HW2

Tuorui Peng[1]

Exercise 1 Adaptive to Anisotropic Smoothnes

We write the $\hat{m}(x)$ as follows:

$$\begin{aligned}\hat{m}(x) =& \frac{\int y\hat{p}(x,y)\,\mathrm{d}y}{\hat{p}(x)}\\ =& \frac{\sum_{i=1}^n h^{-2}K(\frac{X_i-x}{h})\int yK(\frac{Y_i-y}{h})\,\mathrm{d}y}{\sum_{i=1}^n h^{-1}K(\frac{X_i-x}{h})}\\ =& \frac{\sum_{i=1}^n h^{-2}K(\frac{X_i-x}{h})\int (y-Y_i+Y_i)K(\frac{Y_i-y}{h})\,\mathrm{d}y}{\sum_{i=1}^n h^{-1}K(\frac{X_i-x}{h})}\\ =& \frac{\sum_{i=1}^n K(\frac{X_i-x}{h})Y_i}{\sum_{i=1}^n K(\frac{X_i-x}{h})}\end{aligned}$$

which is the same as the kernel regression estimator we defined in class (with window size $h=1$).

$$\hat{m}_{\text{ker}}(x) = \frac{1}{n}\sum_{i=1}^n W(x)Y_i = \frac{\sum_{i=1}^n K(X_i-x)Y_i}{\sum_{i=1}^n K(X_i-x)}$$

Exercise 2 Implicit Bias of Overparameterized Linear Regression

2.(a)

By taking derivative of the empirical risk $R(w)$ with respect to w, we have:

$$\begin{aligned}0 = \frac{\partial}{\partial \mathbf{w}}R(w) =& \frac{\partial}{\partial \mathbf{w}}\frac{1}{2n}\left\|\mathbf{X}\mathbf{w}-\mathbf{t}\right\|_2^2\\ =& \frac{1}{n}\left[\mathbf{X}'(\mathbf{X}\mathbf{w}-\mathbf{t})\right]\\ \Rightarrow \mathbf{X}'\mathbf{X}\mathbf{w} =& \mathbf{X}'\mathbf{t}\end{aligned}$$

which is the expression that the solution $\hat{w}$ satisfies. In underdetermined case, $\mathbf{X}'\mathbf{X}$ is not invertible, and the solution is not unique. We may, however, use e.g. the Moore-Penrose pseudoinverse to find a solution such as

$$\hat{\mathbf{w}}_{\text{Moore}-\text{Penrose}} = (\mathbf{X}'\mathbf{X})^{\dagger}\mathbf{X}'\mathbf{t}$$

2.(b)

With the gradient flow starting from $\mathbf{w}(0)=0$, we have the following ODE:

$$\mathbf{w}(t) = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{\dagger}\mathbf{t}\big(1-\exp(-t/n)\big)$$

[1]TuoruiPeng2028@u.northwestern.edu

- As $t \to \infty$, we have solution

$$\mathbf{w}(t) \to \mathbf{X}'(\mathbf{X}\mathbf{X}')^{\dagger}\mathbf{t} := \mathbf{w}^*$$

- The minimum norm interpolation problem:

$$\arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 \quad \text{subject to} \quad \mathbf{X}\mathbf{w} = \mathbf{t}$$

its Lagrangian is

$$\mathcal{L} = \frac{1}{2}\|\mathbf{w}\|_2^2 + \lambda'(\mathbf{X}\mathbf{w} - \mathbf{t})$$

with minimizer at

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{w}}\mathcal{L} = \mathbf{w} + \mathbf{X}'\lambda = 0 \\
\Rightarrow &\mathbf{X}\mathbf{w} = \mathbf{t} = -\mathbf{X}\mathbf{X}'\lambda \\
\Rightarrow &\lambda = -(\mathbf{X}\mathbf{X}')^{\dagger}\mathbf{t} \\
\Rightarrow &\mathbf{w} = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{\dagger}\mathbf{t}
\end{aligned}$$

which is the same as the solution $\mathbf{w}^*$.

- Here I feel that the relation between the traditional predictor and minimum-norm predictor is similar to the relation between maximum-margin predictor and minimum-norm predictor in SVM. In this relation, the original optimize problem has a tough optimization goal (margin M in SVM, $\|Xw - t\|$ in regression), by transforming to the minimum-norm predictor, we can throw the difficult part (compute margin for SVM, determine degree of freedom for regression) into the regularization term, and solve the problem with a simpler form.

Exercise 3 Benefit of Overparameterization

The result I got is as follows. A similar result as example:

- At low degree of freedom (<7): works well, because the regression task is not complex, a relative small degree of freedom is enough to fit the data;
- At medium degree of freedom (7-70): significant overfitting and we see some terrible deviation from the true function;
- At high degree of freedom (>70): due to the overparameterization, whenever there are some data points at some region, the fitted model would be close to the data points, and thus could be close to the ture model.

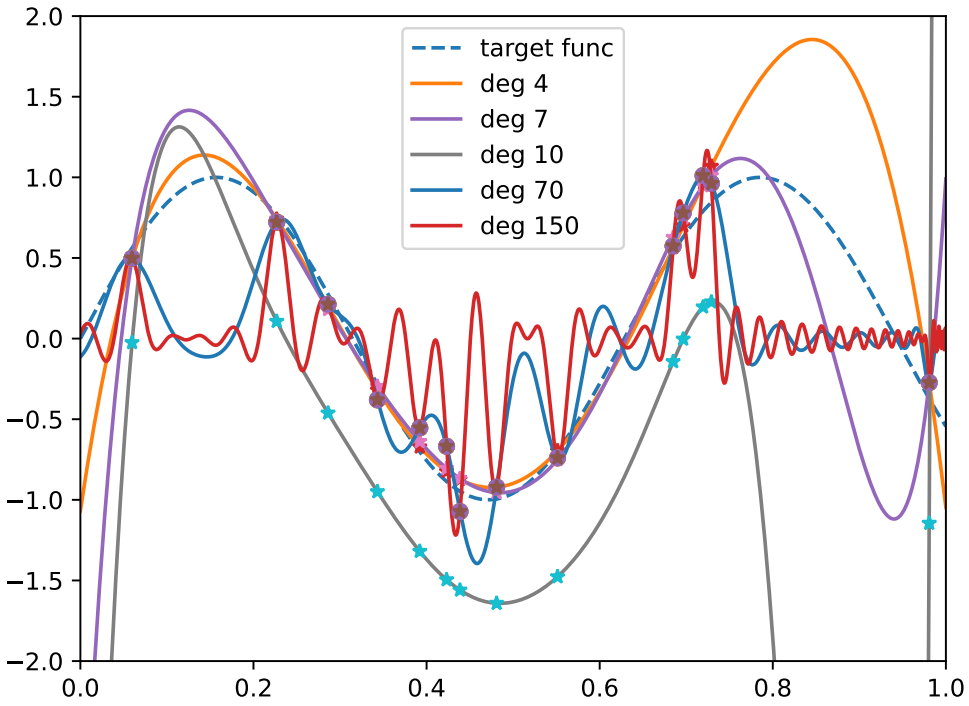

图 1: (Chebyshev) Polynomial Regression: fit v.s. degree of freedom

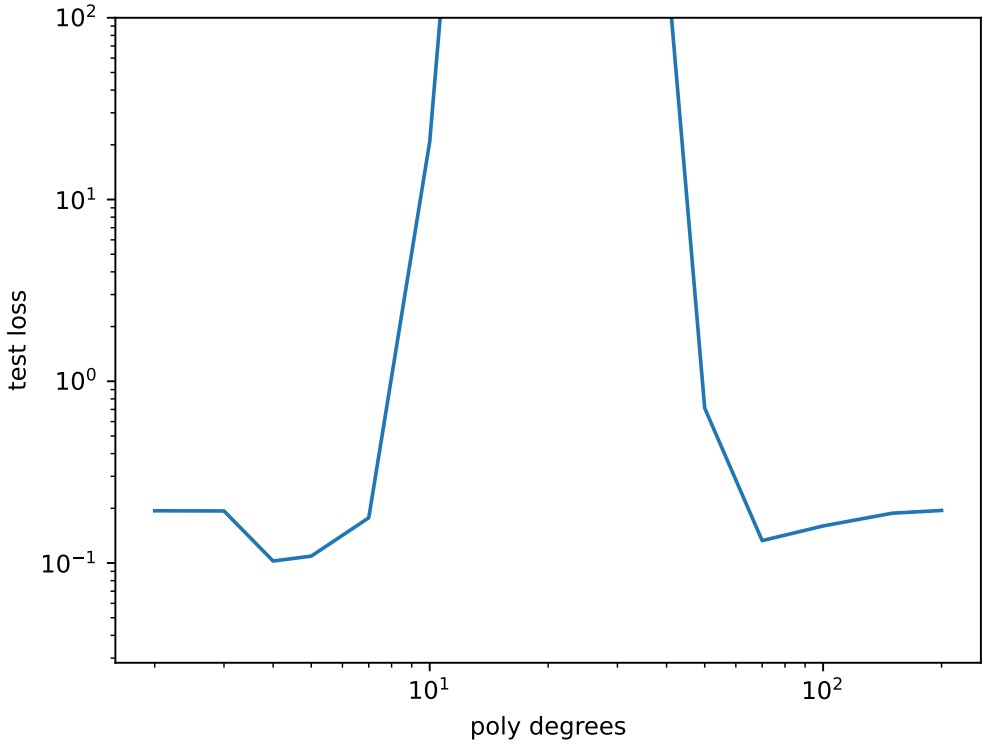

图 2: (Chebyshev) Polynomial Regression: loss v.s. degree of freedom