[EECS 545] HW1

Huy Le

January 2026

1. Derivation and Proof

(a) We are given the loss function

$$L = \frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - h(x^{(i)}))^2,$$

where $h(x) = w_1 x + w_0$. To find optimal parameters, we want to constrain $\nabla L = 0$.

First, we want to find optimal w_0.

$$\begin{aligned}
\frac{\partial L}{\partial w_0} &= \frac{\partial}{\partial w_0}\left[\frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - w_1 x^{(i)} - w_0)^2\right] \\
&= -\sum_{i=1}^{N}(y^{(i)} - w_1 x^{(i)} - w_0) = 0 \\
\implies \sum_{i=1}^{N} y^{(i)} - w_1 \sum_{i=1}^{N} x^{(i)} - N w_0 &= 0 \\
\implies N w_0 &= \sum_{i=1}^{N} y^{(i)} - w_1 \sum_{i=1}^{N} x^{(i)} \\
\implies w_0 &= \frac{1}{N}\sum_{i=1}^{N} y^{(i)} - w_1 \cdot \frac{1}{N}\sum_{i=1}^{N} x^{(i)}
\end{aligned}$$

Therefore, $w_0 = \overline{Y} - w_1 \overline{X}$.

Then, we find optimal w_1. We have

$$\begin{aligned}
\frac{\partial L}{\partial w_1} &= \frac{\partial}{\partial w_1}\left[\frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - w_1 x^{(i)} - w_0)^2\right] \\
&= -\sum_{i=1}^{N} x^{(i)}(y^{(i)} - w_1 x^{(i)} - w_0) = 0.
\end{aligned}$$

$\implies \sum_{i=1}^{N} x^{(i)} y^{(i)} - w_1 \sum_{i=1}^{N} (x^{(i)})^2 - w_0 \sum_{i=1}^{N} x^{(i)} = 0$

Substituting $w_0 = \overline{Y} - w_1\overline{X}$, we have

$$\begin{aligned}
\sum_{i=1}^{N} x^{(i)}y^{(i)} - w_1 \sum_{i=1}^{N}(x^{(i)})^2 - (\overline{Y} - w_1\overline{X}) \sum_{i=1}^{N} x^{(i)} = 0 \\
\implies \sum_{i=1}^{N} x^{(i)}y^{(i)} - w_1 \sum_{i=1}^{N}(x^{(i)})^2 - \overline{Y} \cdot N\overline{X} + w_1\overline{X} \cdot N\overline{X} = 0 \\
\implies \sum_{i=1}^{N} x^{(i)}y^{(i)} - N\overline{Y}\,\overline{X} = w_1 \left[\sum_{i=1}^{N}(x^{(i)})^2 - N\overline{X}^2\right] \\
\implies w_1 = \frac{\sum_{i=1}^{N} x^{(i)}y^{(i)} - N\overline{Y}\,\overline{X}}{\sum_{i=1}^{N}(x^{(i)})^2 - N\overline{X}^2} \\
\implies w_1 = \frac{\frac{1}{N}\sum_{i=1}^{N} x^{(i)}y^{(i)} - \overline{Y}\,\overline{X}}{\frac{1}{N}\sum_{i=1}^{N}(x^{(i)})^2 - \overline{X}^2}.
\end{aligned}$$

(b)
i. Proof that a real symmetric $d \times d$ matrix $\mathbf{A}$ is positive definite if and only if all eigenvalues $\lambda_i > 0$.
($\Rightarrow$) Suppose $\mathbf{A}$ is positive definite. For each eigenvector $\mathbf{u}_i$ with eigenvalue λ_i, we have $\mathbf{A}\mathbf{u}_i = \lambda_i\mathbf{u}_i$. Since eigenvectors are nonzero, we have that

$$\mathbf{u}_i^\top \mathbf{A}\mathbf{u}_i = \mathbf{u}_i^\top(\lambda_i\mathbf{u}_i) = \lambda_i\|\mathbf{u}_i\|^2 > 0.$$

As $\|\mathbf{u}_i\|^2 > 0$, $\lambda_i > 0$ for all i.
($\Leftarrow$) Assume all $\lambda_i > 0$. For any nonzero $\mathbf{z} \in \mathbb{R}^d$, we have

$$\mathbf{z}^\top\mathbf{A}\mathbf{z} = \mathbf{z}^\top\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top\mathbf{z} = (\mathbf{U}^\top\mathbf{z})^\top\boldsymbol{\Lambda}(\mathbf{U}^\top\mathbf{z}).$$

Let $\mathbf{c} = \mathbf{U}^\top\mathbf{z} = [c_1, c_2, \ldots, c_d]^\top$. Since $\mathbf{U}$ is orthogonal, $\mathbf{c} \neq \mathbf{0}$ iff $\mathbf{z} \neq \mathbf{0}$. Thus

$$\mathbf{z}^\top\mathbf{A}\mathbf{z} = \sum_{i=1}^{d} \lambda_i c_i^2 > 0$$

since $\lambda_i > 0$ for all i and at least one $c_i \neq 0$.
ii. Let $\mathbf{u}_i$ be an eigenvector of $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ with eigenvalue λ_i,

$$(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \beta\mathbf{I})\mathbf{u}_i = \boldsymbol{\Phi}^\top\boldsymbol{\Phi}\mathbf{u}_i + \beta\mathbf{u}_i = \lambda_i\mathbf{u}_i + \beta\mathbf{u}_i = (\lambda_i + \beta)\mathbf{u}_i.$$

Therefore, $\mathbf{u}_i$ remains an eigenvector of $\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \beta\mathbf{I}$ with eigenvalue $\lambda_i + \beta$.
Since $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ is positive semi-definite, all $\lambda_i \geq 0$. Therefore, the eigenvalues of $\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \beta\mathbf{I}$ are $\lambda_i + \beta > 0$ for any $\beta > 0$.
From property proved in part i, this implies $\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \beta\mathbf{I}$ is positive definite.

(c) We show that maximizing the log-likelihood is equivalent to minimizing the given loss function.

$$\begin{aligned}
\log P(y^{(n)}|\mathbf{x}^{(n)}) = \mathbb{I}(y^{(n)} = 1)\log P(y^{(n)} = 1|\mathbf{x}^{(n)}) \\
+ \mathbb{I}(y^{(n)} = -1)\log P(y^{(n)} = -1|\mathbf{x}^{(n)}).
\end{aligned}$$

Since probabilities must sum to 1, we have

$$P(y = -1|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}) = 1 - \frac{1}{1 + \exp(-\mathbf{w}^\top\phi(\mathbf{x}))} = \frac{1}{1 + \exp(\mathbf{w}^\top\phi(\mathbf{x}))}$$

We can express $P(y|\mathbf{x}) = \dfrac{1}{1 + \exp(-y\mathbf{w}^\top\phi(\mathbf{x}))}$ for $y \in \{-1, +1\}$

We have

$$\log P(y^{(n)}|\mathbf{x}^{(n)}) = -\log(1+\exp(-y^{(n)}\mathbf{w}^\top\phi(\mathbf{x}^{(n)})))$$

$$\implies \sum_{n=1}^{N}\log P(y^{(n)}|\mathbf{x}^{(n)}) = -\sum_{n=1}^{N}\log(1+\exp(-y^{(n)}\mathbf{w}^\top\phi(\mathbf{x}^{(n)})))$$

So maximizing the log-likelihood of the logistic regression is equivalent to minimizing the loss function

$$\sum_{n=1}^{N}\log(1+\exp(-y^{(n)}\mathbf{w}^\top\phi(\mathbf{x}^{(n)})))$$

2.1 GD and SGD

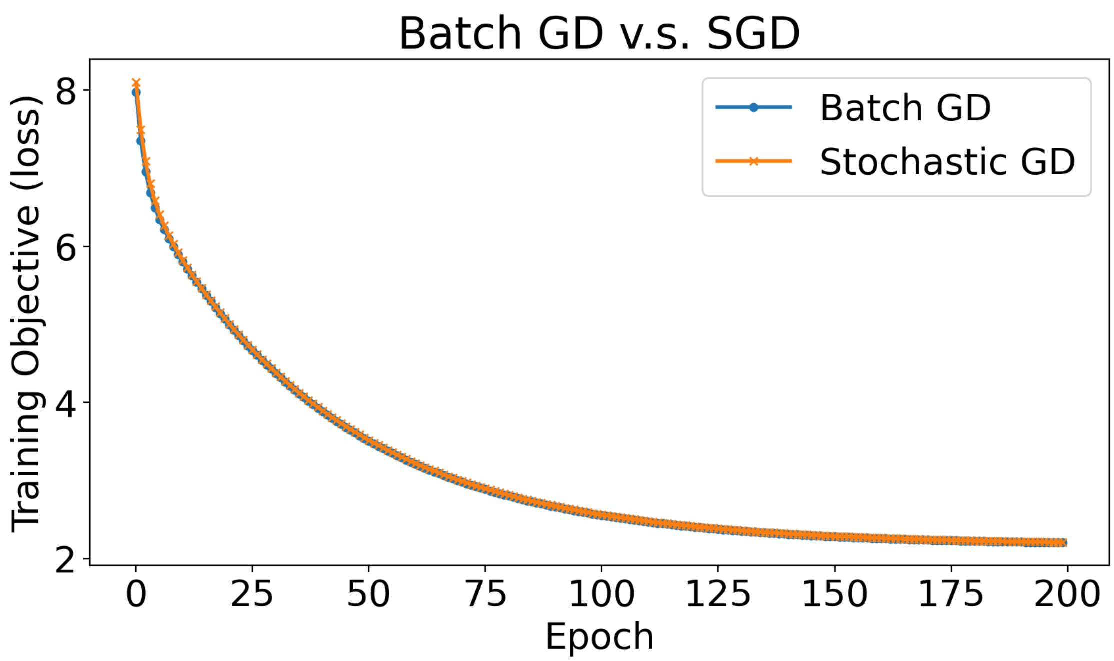

Figure 1: Training objective (loss) versus epoch for Batch GD and SGD.

From the printed runtimes, Batch GD is faster (about 0.00 s vs. 0.02 s for SGD). Meanwhile, SGD achieves the lower test objective $E(\mathbf{w}_{\text{test}})$ (about 0.1340 vs. 0.1351 for Batch GD).

2.2 Over-fitting study

(b) The RMS error curves (training and test) for polynomial feature dimension $M \in \{1, \ldots, 10\}$ are shown below.

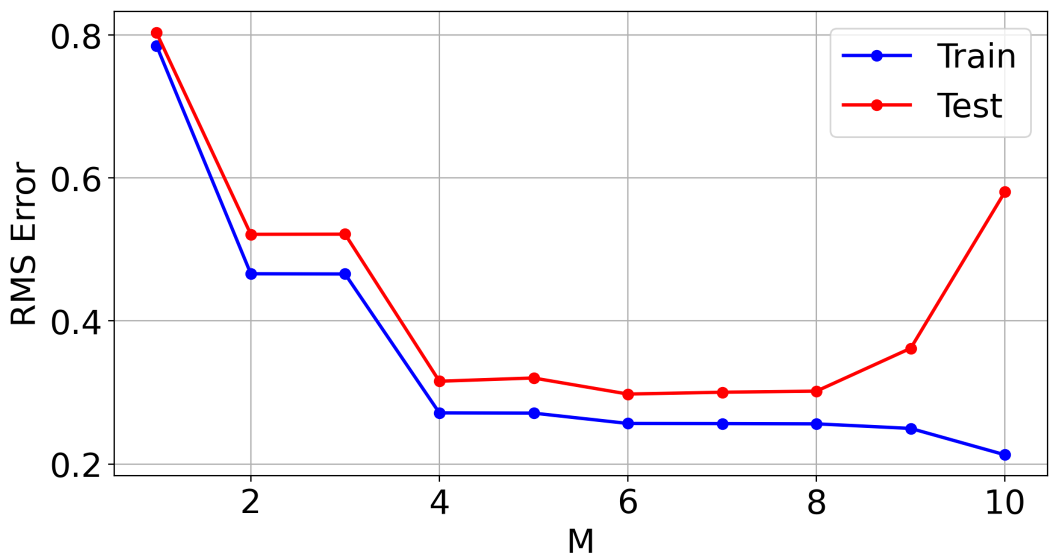

Figure 2: RMS error versus number of features M on training and test sets

(c) From the plot, the training RMS generally decreases as M increases, while the test RMS decreases at first and then increases for larger M. The smallest test RMS is achieved around $M = 6$ while the training RMS is also relatively small, so a polynomial with degree of 6 would best fit the data. For small M ($M = 1, 2$), both training and test RMS are relatively high which shows underfitting, whereas for large M (10) the training RMS becomes very small but the test RMS increases, which is evident to overfitting.

2.3 Regularization (Ridge Regression)

(b) The training and test RMS errors as a function of the regularization factor λ are shown below.

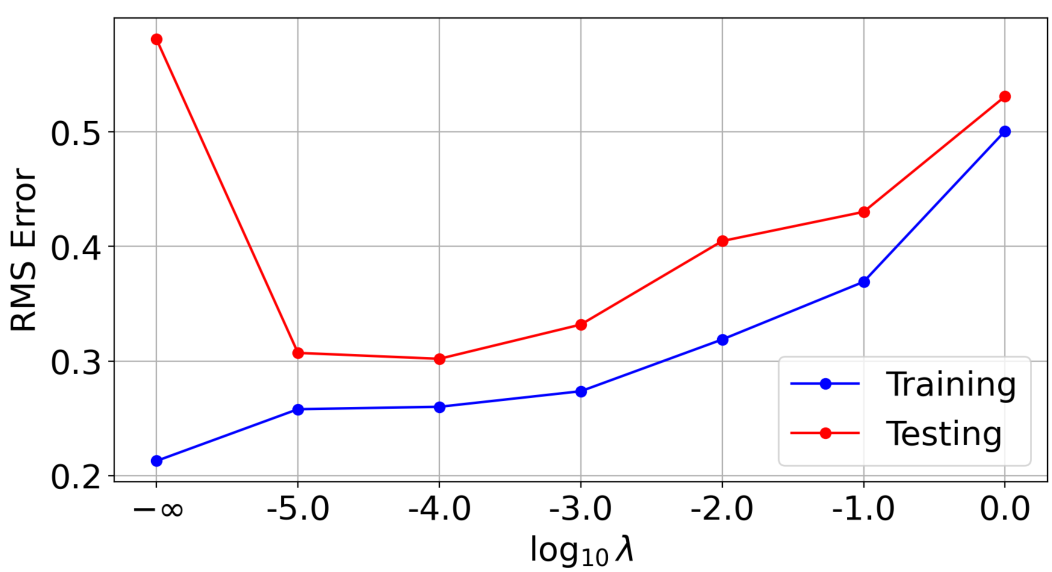

Figure 3: RMS error versus regularization factor λ (ridge regression) on training and test sets

(c) From the plot, the training RMS is slightly lower at $\log_{10}(\lambda) = -5$ than that of $\log_{10}(\lambda) = -4$, but the test RMS achieves its minimum at $\log_{10}(\lambda) = -4$. Hence, $\lambda = 10^{-4}$ stands out as the best choice for the degree-9 polynomial.

3. Locally weighted linear regression

Consider the objective

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N} r^{(i)}\big(\mathbf{w}^\top \mathbf{x}^{(i)} - y^{(i)}\big)^2$$

(a) Let $\mathbf{X} \in \mathbb{R}^{D\times N}$ be the data matrix whose i-th column is $\mathbf{x}^{(i)}$, and let $\mathbf{y} \in \mathbb{R}^{N\times 1}$ be the vector whose i-th entry is $y^{(i)}$. We define the diagonal matrix

$$\mathbf{R} \triangleq \frac{1}{2}\,\mathrm{diag}\big(r^{(1)}, \ldots, r^{(N)}\big) \in \mathbb{R}^{N\times N}$$

Then $\mathbf{w}^\top\mathbf{X} - \mathbf{y}^\top$ is a $1 \times N$ row vector whose i-th entry equals $\mathbf{w}^\top\mathbf{x}^{(i)} - y^{(i)}$, and hence

$$E_D(\mathbf{w}) = (\mathbf{w}^\top\mathbf{X} - \mathbf{y}^\top)\,\mathbf{R}\,(\mathbf{w}^\top\mathbf{X} - \mathbf{y}^\top)^\top$$

(b) Given

$$E_D(\mathbf{w}) = (\mathbf{X}^\top\mathbf{w} - \mathbf{y})^\top\mathbf{R}(\mathbf{X}^\top\mathbf{w} - \mathbf{y}),$$

and the fact that $\mathbf{R}$ is symmetric, we have

$$\nabla_{\mathbf{w}} E_D(\mathbf{w}) = 2\mathbf{X}\mathbf{R}(\mathbf{X}^\top\mathbf{w} - \mathbf{y})$$

Setting the gradient to zero gives the weighted normal equation

$$\mathbf{X}\mathbf{R}\mathbf{X}^\top\mathbf{w} = \mathbf{X}\mathbf{R}\mathbf{y}$$

Assuming $\mathbf{X}\mathbf{R}\mathbf{X}^\top$ is invertible, then the closed form solution is

$$\mathbf{w}^* = (\mathbf{X}\mathbf{R}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{R}\mathbf{y}.$$

(c) The conditional likelihood for each example is

$$p\big(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}\big) = \frac{1}{\sqrt{2\pi}\,\sigma^{(i)}}\exp\left(-\frac{\big(y^{(i)} - \mathbf{w}^\top\mathbf{x}^{(i)}\big)^2}{2\big(\sigma^{(i)}\big)^2}\right)$$

Assuming the N examples are independent, the likelihood of the full dataset is given by

$$\mathcal{L}(\mathbf{w}) = \prod_{i=1}^{N} p\big(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}\big).$$

Hence,

$$\begin{aligned}
-\log\mathcal{L}(\mathbf{w}) &= -\sum_{i=1}^{N}\log p\big(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}\big) \\
&= \sum_{i=1}^{N}\left[\log(\sqrt{2\pi}\,\sigma^{(i)}) + \frac{\big(y^{(i)} - \mathbf{w}^\top\mathbf{x}^{(i)}\big)^2}{2\big(\sigma^{(i)}\big)^2}\right] \\
&= \underbrace{\sum_{i=1}^{N}\log(\sqrt{2\pi}\,\sigma^{(i)})}_{\text{constant w.r.t. } \mathbf{w}} + \sum_{i=1}^{N}\frac{\big(y^{(i)} - \mathbf{w}^\top\mathbf{x}^{(i)}\big)^2}{2\big(\sigma^{(i)}\big)^2}
\end{aligned}$$

Therefore, dropping the constant term (independent of $\mathbf{w}$), maximizing the likelihood is equivalent to minimizing

$$\sum_{i=1}^{N} \frac{\left(y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)}\right)^2}{2\left(\sigma^{(i)}\right)^2} = \frac{1}{2}\sum_{i=1}^{N} r^{(i)}\left(\mathbf{w}^\top \mathbf{x}^{(i)} - y^{(i)}\right)^2,$$

where

$$r^{(i)} = \frac{1}{\left(\sigma^{(i)}\right)^2}$$

(d)

ii.

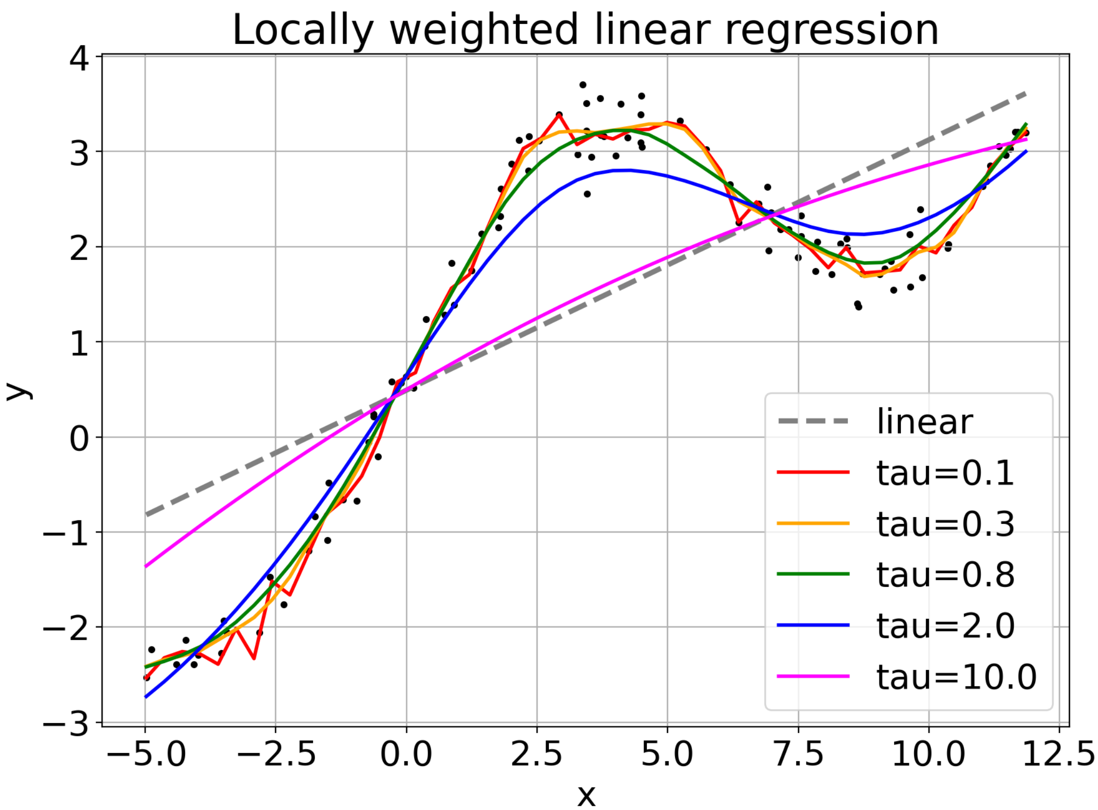

Figure 4: Locally weighted linear regression fits for different bandwidths τ

iii. When τ is very small, the weights concentrate on points extremely close to the query x, so the fit has high variance and can overfit. When τ is very large, however, the weights become nearly uniform, so the fit approaches ordinary least squares and becomes overly smooth, which corresponds to high bias and can underfit.