

## Homework #4

RELEASE DATE: 11/25/2021

DUE DATE: 12/09/2021, BEFORE 13:00 on Gradescope

QUESTIONS ARE WELCOMED ON THE NTU COOL FORUM.

*You will use Gradescope to upload your choices and your scanned/printed solutions. For problems marked with (\*), please follow the guidelines on the course website and upload your source code to Gradescope as well. Any programming language/platform is allowed.*

*Any form of cheating, lying, or plagiarism will not be tolerated. Students can get zero scores and/or fail the class and/or be kicked out of school and/or receive other punishments for those kinds of misconducts.*

*Discussions on course materials and homework solutions are encouraged. But you should write the final solutions alone and understand them fully. Books, notes, and Internet resources can be consulted, but not copied from.*

*Since everyone needs to write the final solutions alone, there is absolutely no need to lend your homework solutions and/or source codes to your classmates at any time. In order to maximize the level of fairness in this class, lending and borrowing homework solutions are both regarded as dishonest behaviors and will be punished according to the honesty policy.*

*You should write your solutions in English or Chinese with the common math notations introduced in class or in the problems. We do not accept solutions written in any other languages.*

This homework set comes with 16 problems and a total of 400 points. For each problem, there is one correct choice. If you choose the correct answer, you get 20 points; if you choose an incorrect answer, you get 0 points. For four of the secretly-selected problems, the TAs will grade your detailed solution in terms of the written explanations and/or code based on how logical/clear your solution is. Each of the four problems graded by the TAs counts as additional 20 points (in addition to the correct/incorrect choices you made). In general, each homework (except homework 0) is of a total of 400 points.

## Regularization

1. Consider the augmented error

$$E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N} \mathbf{w}^T \mathbf{w}$$

with some  $\lambda > 0$ . Optimize the error by gradient descent with  $\eta$  as the learning rate. That is,

$$\mathbf{w}(t+1) \leftarrow \mathbf{w}(t) - \eta \nabla E_{\text{aug}}(\mathbf{w}(t))$$

The update rule above is equivalent to multiplying  $\mathbf{w}(t)$  by a shrinkage factor  $s$  before updating with the negative gradient of  $E_{\text{in}}$

$$\mathbf{w}(t+1) \leftarrow s \cdot \mathbf{w}(t) - \eta \nabla E_{\text{in}}(\mathbf{w}(t))$$

$$\frac{\partial E_{\text{aug}}}{\partial w_i} = \frac{\partial E_{\text{in}}}{\partial w_i} + \frac{\partial \frac{\lambda}{N} w^T w}{\partial w_i}$$

What is  $s$ ? Choose the correct answer; explain your answer.

- [a]  $1 - \frac{\eta \lambda}{N}$
- [b]  $1 - \frac{2\lambda}{N}$
- [c]  $1 - \frac{2\eta \lambda}{N}$
- [d]  $1 - \frac{2\eta}{N}$
- [e] none of the other choices

$$\nabla E_{\text{aug}} = \nabla E_{\text{in}} + \frac{2\lambda}{N} \mathbf{w}$$

$$= \frac{\partial E_{\text{in}}}{\partial w_i} + \frac{2\lambda}{N} w_i$$

$$\mathbf{w}(t+1) \leftarrow \mathbf{w}(t) - \eta \nabla E_{\text{aug}}(t)$$

$$= \mathbf{w}(t) - \eta \left( \nabla E_{\text{in}}(t) + \frac{2\lambda}{N} \mathbf{w}(t) \right) = \left( 1 - \frac{2\eta \lambda}{N} \right) \mathbf{w}(t) - \eta \left( \nabla E_{\text{in}}(t) \right)$$

\*

2. Consider  $N$  "labels"  $\{y_n\}_{n=1}^N$  with each  $y_n \in \mathbb{R}$ . Then, solve the following one-variable regularized problem:

$$\min_{w \in \mathbb{R}} \frac{1}{N} \sum_{n=1}^N (w - y_n)^2 + \frac{\lambda}{N} w^2.$$

If the optimal solution to the problem above is  $w^*$ , it can be shown that  $w^*$  is also the optimal solution of

$$\min_{w \in \mathbb{R}} \frac{1}{N} \sum_{n=1}^N (w - y_n)^2 \text{ subject to } w^2 \leq C$$

with  $C = (w^*)^2$ . This allows us to express the relationship between  $C$  in the constrained optimization problem and  $\lambda$  in the augmented optimization problem for any  $\lambda > 0$ . What is the relationship? Choose the correct answer; explain your answer.

[a]  $C = \left( \frac{\sum_{n=1}^N y_n}{N + N\lambda} \right)^2$

[b]  $C = \left( \frac{\sum_{n=1}^N y_n}{N + \lambda} \right)^2$

[c]  $C = \left( \frac{\sum_{n=1}^N y_n^2}{\sum_{n=1}^N y_n + \lambda} \right)^2$

[d]  $C = \left( \frac{\sum_{n=1}^N y_n^2}{N + \lambda} \right)^2$

[e] none of the other choices

$$w^* = [x^T x + \lambda I]^{-1} x^T y$$

$$= (N + \lambda)^{-1} \sum y_n$$

$$= \frac{\sum y_n}{N + \lambda}$$

$$= (w^*)^2 = \left( \frac{\sum y_n}{N + \lambda} \right)^2$$

$$W_{reg} = (x^T x + \lambda I)^{-1} x^T y$$

$$\rightarrow = (\Sigma_1 + \lambda)^{-1} \sum y_n$$

$$= (N + \lambda)^{-1} \sum y_n$$

$$w^2 = \left( \frac{\sum y_n}{N + \lambda} \right)^2$$

(Note: All the choices hint you that a smaller  $\lambda$  corresponds to a bigger  $C$ .)

3. Scaling can affect regularization. Consider a data set  $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$ . Define  $\Phi(\mathbf{x}) = V\mathbf{x}$  where  $V$  is a diagonal matrix with the  $i$ -th diagonal component storing a *positive* value to scale the  $i$ -th feature. Now, conduct L2-regularized linear regression with the transformed data  $\{(\Phi(\mathbf{x}_n), y_n)\}_{n=1}^N$ .

$$\min_{\tilde{\mathbf{w}} \in \mathbb{R}^{d+1}} \frac{1}{N} \sum_{n=1}^N (\tilde{\mathbf{w}}^T \Phi(\mathbf{x}_n) - y_n)^2 + \frac{\lambda}{N} (\tilde{\mathbf{w}}^T \tilde{\mathbf{w}})$$

$$W_{reg} = (Z^T Z + \lambda I)^{-1} Z^T Y$$

$$\text{here } Z = V\mathbf{x}$$

The problem is equivalent to the following regularized linear regression problem on the original data with a different regularizer.

$$\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^T \mathbf{x}_n - y_n)^2 + \frac{\lambda}{N} (\mathbf{w}^T U \mathbf{w})$$

$$\tilde{\mathbf{w}} = (x^T V V^T x + \lambda I)^{-1} V V^T y$$

$$w = V \tilde{w} = V (x^T V V^T x + \lambda I)^{-1} V^T y$$

What is  $U$ ? Choose the correct answer; explain your answer.

[a]  $V$

$$\nabla E_{lin}(w) + \frac{\lambda}{N} \nabla (w^T U w) = 0$$

[b]  $V^2$

$$\frac{2}{N} x^T x w + \frac{\lambda}{N} (S U w) = x^T y$$

[c]  $V^{-1}$

$$w = (x^T x + \lambda U)^{-1} x^T y$$

[d]  $(V^{-1})^2$

[e] none of the other choices

$$U \stackrel{\text{def}}{=} V^{-1} (V^T)^{-1}$$

$$w = 2 \text{ of } 8 [x^T x + \lambda V^T V]^{\frac{1}{2}} x^T y$$

$$= \frac{x^T y}{\sqrt{x^T x + \lambda V^T V}}$$

$$(x^T V V^T x + \lambda I)^{\frac{1}{2}} V^T y$$

$$= \frac{\sqrt{x^T y}}{\sqrt{x^T x + \lambda V^T V}}$$

$$\partial f(\tilde{w}^*) = w^*$$

→ 解上面 等于解下面的  
最优解存在线性關係

通常有的

方向是解  $w = \underbrace{\Box}_{\text{代入下面}} \tilde{w}$

$$\begin{aligned} & \sum ((w^T x_n) - y_n)^2 + \frac{\lambda}{N} (w^T L w) \\ &= \frac{1}{N} \sum ((w^T x_n) - y_n)^2 + \frac{\lambda}{N} (w^T L w) \end{aligned}$$

4. Noise might seem to be an absolute evil at first glance. Can we add noise to combat overfitting, much like how we get vaccinated? Consider a *noisy transform*  $\Phi(\mathbf{x}) = \mathbf{x} + \boldsymbol{\epsilon}$  that adds a noise vector  $\boldsymbol{\epsilon} = (\epsilon_0, \dots, \epsilon_d)$  to  $\mathbf{x}$ . Assume that the noise vector is generated i.i.d. from a multivariate normal distribution  $\mathcal{N}(\mathbf{0}, \sigma^2 I)$ . Consider the following optimization problem that minimizes the expected  $E_{\text{in}}$  of the transformed data:

$$\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \mathbb{E} \left[ \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^T \Phi(\mathbf{x}_n) - y_n)^2 \right]$$

The problem is actually equivalent to L2-regularized linear regression with some  $\lambda$ .

$$\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^T \mathbf{x}_n - y_n)^2 + \frac{\lambda}{N} \|\mathbf{w}\|_2^2$$

What is  $\lambda$ ? Choose the correct answer; explain your answer.

- [a]  $N\sigma^2$
- [b]  $2N\sigma^2$
- [c]  $\frac{N}{2}\sigma^2$
- [d]  $\frac{N}{\sigma^2}$
- [e] none of the other choices

$$\frac{N\sigma^2}{N} = \sigma^2 \|\mathbf{w}\|^2$$

$$y^2 + 2y\mathbf{w} + \mathbf{w}^2$$

$$\frac{\sigma^2}{2} \|\mathbf{w}\|^2$$

$$\frac{1}{2} \|\mathbf{w}\|^2$$

$$2y\mathbf{w}$$

5. Additive smoothing ([https://en.wikipedia.org/wiki/Additive\\_smoothing](https://en.wikipedia.org/wiki/Additive_smoothing)) is a simple yet useful technique in estimating discrete probabilities. Consider the technique for estimating the head probability of a coin. Let  $y_1, y_2, \dots, y_N$  denotes the flip results from a coin, with  $y_n = 1$  represents a head and  $y_n = 0$  represents a tail. Additive smoothing adds  $(\alpha k)$  “virtual flips”, with  $\alpha$  of them being head and the other  $(k-1)\alpha$  being tail. Then, the head probability is estimated by

$$\frac{(\sum_{n=1}^N y_n) + \alpha}{N + \alpha k}$$

The estimate can be viewed as the optimal solution of

$$\min_{y \in \mathbb{R}} \frac{1}{N} \sum_{n=1}^N (y - y_n)^2 + \frac{\alpha k}{N} \Omega(y),$$

where  $\Omega(y)$  is a “regularizer” to this estimation problem. What is  $\Omega(y)$ ? Choose the correct answer; explain your answer.

- [a]  $(y + k)^2$
- [b]  $(y + \frac{1}{k})^2$
- [c]  $(y - k)^2$
- [d]  $(y - \frac{1}{k})^2$
- [e] none of the other choices

$$\begin{aligned} & \text{F} \quad \text{不需要这样写} \\ & \text{d}y \rightarrow \sum (y - y_n)^2 + \frac{\alpha k}{N} (y + k)^2 = 0 \\ & \frac{dy}{dy} \end{aligned}$$

Normal distribution  $N(0, \frac{1}{N} I)$

$$\varepsilon = (\varepsilon_0, \varepsilon_1, \dots, \varepsilon_{d+1})$$

4.  $\sum (\tilde{\mathbf{w}}^T \Phi \mathbf{x}_n - y_n)^2$  缩成 matrix form  $\tilde{\mathbf{x}} = \Phi(\mathbf{x})$   
er.  $\frac{1}{N} \|\tilde{\mathbf{x}} \mathbf{w} - \mathbf{y}\|^2$

(1)  $E[\tilde{\mathbf{x}}^T \tilde{\mathbf{x}}]$  key point

$$E[\tilde{\mathbf{x}}^T \tilde{\mathbf{x}} + \tilde{\mathbf{x}}^T E + E^T \tilde{\mathbf{x}} + E^T E]$$

$$E[\tilde{\mathbf{x}}^T \tilde{\mathbf{x}}]$$

$$= [x_1 \cdot \tilde{\mathbf{e}}_1^T + x_2 \cdot \tilde{\mathbf{e}}_2^T + \dots + x_n \cdot \tilde{\mathbf{e}}_n^T]$$

$$= E[\tilde{\mathbf{x}} \cdot \tilde{\mathbf{e}}_1^T] + E[\tilde{\mathbf{x}} \cdot \tilde{\mathbf{e}}_2^T] + \dots$$

key 取後解法

$$\begin{aligned} & \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \\ \vdots \\ \tilde{\mathbf{x}}_n \end{bmatrix} \\ &= \begin{bmatrix} (x_1, \tilde{\mathbf{e}}_1) \\ (x_2, \tilde{\mathbf{e}}_2) \\ \vdots \\ (x_n, \tilde{\mathbf{e}}_n) \end{bmatrix} \end{aligned}$$

$$\begin{aligned} & \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \\ \vdots \\ \tilde{\mathbf{x}}_n \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} + \begin{bmatrix} \tilde{\mathbf{e}}_1^T \\ \tilde{\mathbf{e}}_2^T \\ \vdots \\ \tilde{\mathbf{e}}_n^T \end{bmatrix} \\ & \mathbf{x} \quad \mathbf{E} \end{aligned}$$

$$\min E\left[ \frac{1}{N} \|(\mathbf{x} + \mathbf{E}) \mathbf{w} - \mathbf{y}\|^2 \right]$$

只有  $\mathbf{E}$  有  $\mathbf{P}_{\mathbf{E}}^T \mathbf{E} = \mathbf{E}$

4. 通过线性代数方法

$$E\left[ \begin{bmatrix} \tilde{\mathbf{x}}^T & \tilde{\mathbf{x}}^T \mathbf{w} \end{bmatrix} \right] = \begin{bmatrix} \mathbf{x}^T & \mathbf{y} \end{bmatrix}$$

after the rule

data:

$$\min_{\mathbf{w} \in \mathbb{R}^{d+1}} E\left[ \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^T \Phi(\mathbf{x}_n) - y_n)^2 \right]$$

$$\left[ \frac{1}{N} \sum (\mathbf{w}^T \Phi(\mathbf{x}_n) - y_n)^2 \right]$$

equivalent to L2-regularized linear regression with some

$$\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^T \mathbf{x}_n - y_n)^2 + \frac{\lambda}{N} \|\mathbf{w}\|_2^2$$

$$\frac{1}{N} \|\Phi(\mathbf{w}) - \mathbf{y}\|^2$$

$$\Phi(\mathbf{x}) = \mathbf{x} + \mathbf{E}$$

$$\Rightarrow E\left[ \frac{1}{N} \|(\mathbf{x} + \mathbf{E}) \mathbf{w} - \mathbf{y}\|^2 \right]$$

$$\begin{aligned} & \frac{1}{N} \|\mathbf{w} \mathbf{x} - \mathbf{y}\|^2 \\ &= \frac{1}{N} \|\mathbf{w}^2 \mathbf{x}^2 - 2\mathbf{w} \mathbf{x} \mathbf{y} + \mathbf{y}^2\| \end{aligned}$$

$$E \frac{1}{N} \| (X+E)w \cdot (X+E)w - 2[X+E]w y + \|y\|^2$$

$$\frac{1}{N} [E\|[(X+E)^T(X+E)]\| + w^T w - 2wy E\|X+E\| + \|y\|^2]$$

$$\frac{1}{N} (E(x^2 + XE^TEx + E^2) \cdot \|w\|^2 - 2xwy - 2Exy + \|y\|^2)$$

$$E(x^2w^2 - 2Exy + \|y\|^2) + \|XE^TEx + E^2\|w^2 - 2Exy$$

$$\frac{1}{N} E(wx-y)^2 + E\|xE\| + E\|Ex\| + E\|\varepsilon^2\|w^2 - 2Exy$$

$$= \frac{1}{N} \sum (wx_n - y_n)^2 + 0 + 0 + \bar{\sigma}^2 \|w\|^2 - 0$$

$$= \frac{1}{N} \sum (wx_n - y_n)^2 + \bar{\sigma}^2 w^2$$

$$= \frac{1}{N} \sum (wx_n - y_n)^2 + \bar{\sigma}^2 w^2$$

0

$$\bar{\sigma}^2 = \frac{\lambda}{N}$$

$$\therefore \lambda = N\bar{\sigma}^2$$

\*

$$\frac{\partial E}{\partial y} = 0 \quad E = \frac{1}{N} \sum_{n=1}^N (y - y_n)^2 + \frac{\alpha k}{N} \sum_{n=1}^N |y - y_n|$$

5.

$$= \frac{1}{N} \sum_{n=1}^N 2(y - y_n) + \frac{\alpha k}{N} \sum_{n=1}^N |y - y_n|$$

$$0 = 2y - 2 \frac{\sum y_n}{N} + \frac{\alpha k}{N} \sum_{n=1}^N |y - y_n|$$

$$\hat{y} = (y + c)^2 \quad \hat{y}'(y) = 2(y + c)$$

$$\frac{\alpha k}{N} \hat{y}'(y) = \frac{2 \sum y_n}{N} \rightarrow y$$

$$\hat{y}'(y) = \frac{2 \sum y_n}{\alpha k} - \frac{2 N y}{\alpha k} = \frac{2}{\alpha k} (\sum y_n - N y)$$

$$2(cy + c) = \frac{2 \sum y_n - 2 N y}{\alpha k}$$

$$c = \frac{\sum y_n - N y - \alpha k y}{\alpha k}$$

$$2c = \frac{2 \sum y_n - 2 N y - 2 \alpha k y}{\alpha k}$$

$$c = \frac{\sum y_n}{\alpha k} - \frac{(N + \alpha k)y}{\alpha k}$$

$$c = \frac{\sum y_n}{\alpha k} - \frac{(N + \alpha k)}{\alpha k} \cdot \frac{\sum y_n + \alpha k}{(N + \alpha k)}$$

$$c = \frac{\sum y_n}{\alpha k} - \frac{\sum y_n + \alpha k}{\alpha k} = \frac{\sum y_n}{\alpha k} - \frac{\sum y_n}{\alpha k} - \frac{1}{k}$$

$$c = -\frac{1}{k} \quad \hat{y} = \left(y - \frac{1}{k}\right)^2$$

X



6. When performing L2-regularization on any twice-differentiable  $E_{\text{in}}(\mathbf{w})$  for some linear model of interest, the optimization problem can be written as:

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^{d+1}} \quad & E_{\text{aug}}(\mathbf{w}) \\ \text{subject to} \quad & E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N} \|\mathbf{w}\|_2^2 \end{aligned}$$

Suppose  $\mathbf{w}^*$  is the minimizer of  $E_{\text{in}}(\mathbf{w})$ . That is,  $\nabla E_{\text{in}}(\mathbf{w}^*) = \mathbf{0}$ . Take the second-order Taylor's expansion of  $E_{\text{in}}(\mathbf{w})$  around  $\mathbf{w}^*$ , we can approximate  $E_{\text{in}}(\mathbf{w})$  by

$$\tilde{E}_{\text{in}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}^*) + \underbrace{(\mathbf{w} - \mathbf{w}^*)^T \nabla E_{\text{in}}(\mathbf{w}^*)}_{0} + \frac{1}{2} (\mathbf{w} - \mathbf{w}^*)^T \mathbf{H} (\mathbf{w} - \mathbf{w}^*)$$

where  $\mathbf{H} \in \mathbb{R}^{(d+1) \times (d+1)}$  is some Hessian matrix. Then,  $E_{\text{aug}}(\mathbf{w})$  can be approximated by

$$\tilde{E}_{\text{aug}}(\mathbf{w}) = \tilde{E}_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N} \|\mathbf{w}\|^2.$$

Which of the following is the minimizer of  $\tilde{E}_{\text{aug}}(\mathbf{w})$ ? Choose the correct answer; explain your answer.

- [a]  $(\mathbf{H} + \lambda \mathbf{I})^{-1} \mathbf{H} \mathbf{w}^*$
- [b]  $(\mathbf{H} + \frac{2\lambda}{N} \mathbf{I})^{-1} \mathbf{H} \mathbf{w}^*$
- [c]  $(\mathbf{H}^T \mathbf{H} + \lambda \mathbf{I})^{-1} \mathbf{H}^T \mathbf{H} \mathbf{w}^*$
- [d]  $(\mathbf{H} + \frac{\lambda}{N} 2\mathbf{I})^{-1} \mathbf{H} \mathbf{w}^*$**
- [e] none of the other choices.

(Note: The result demonstrates the difference between the minimizers of  $E_{\text{in}}$  and  $\tilde{E}_{\text{aug}}$ )

## Validation



7. Consider a binary classification algorithm  $\mathcal{A}_{\text{minority}}$ , which returns a constant classifier that always predicts the minority class (i.e., the class with fewer instances in the data set that it sees). As you can imagine, the returned classifier is the worst- $E_{\text{in}}$  one among all constant classifiers. For a binary classification data set with  $N$  positive examples and  $N$  negative examples, what is  $E_{\text{loocv}}(\mathcal{A}_{\text{minority}})$ ? Choose the correct answer; explain your answer.

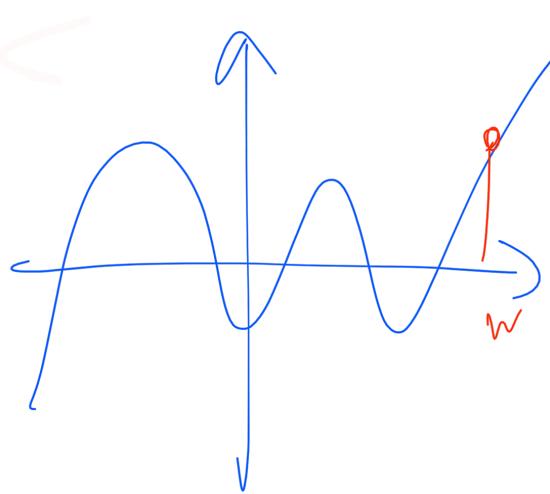
- [a] 0** ∵ 兩組資料相同、取出一個後 (+, -) 預測皆為同樣的 (+, -)
- [b]  $1/N$  ∵ 取出後資料會被劃分到那一側
- [c]  $1/2$  ∵ 取出後資料會被劃分到那一側
- [d]  $(N-1)/N$  ∵ 取出後資料會被劃分到那一側 → 全部誤判 error
- [e] 1 ∵ 取出後資料會被劃分到那一側 → Validation → 全部正確



8. You are given three data points:  $(x_1, y_1) = (2, 0)$ ,  $(x_2, y_2) = (\rho, 2)$ ,  $(x_3, y_3) = (-2, 0)$  with  $\rho \geq 0$ , and a choice between two models: constant (all hypotheses are of the form  $h(x) = w_0$ ) and linear (all hypotheses are of the form  $h(x) = w_0 + w_1 x$ ). For which value of  $\rho$  would the two models be tied using leave-one-out cross-validation with the squared error measure? Choose the closest answer; explain your answer.

- [a] 6.5
- [b] 7.5
- [c] 8.5**
- [d] 9.5
- [e] 10.5

$$E_{\text{loocv}} = 0$$



$$\frac{\partial}{\partial w} \tilde{E}_{\text{aug}}$$

$$= \tilde{F}_{\text{in}}(w)$$

$$\tilde{E}_{\text{in}} \neq w^* \text{ PFA } \}$$

$\Rightarrow$  Taylor Expansion

$$\tilde{E}_{\text{in}} = E_{\text{in}}(w^*) + 0 + \frac{1}{2} (w - w^*)^T H (w - w^*)$$

$$\tilde{E}_{\text{aug}}(w) = \tilde{E}_{\text{in}} + \frac{\lambda}{N} \|w\|^2$$

$$\tilde{E}_{\text{aug}}(w) = E_{\text{in}}(w^*) + \frac{1}{2} (w - w^*)^T H (w - w^*) + \frac{\lambda}{N} \|w\|^2$$

$$\Rightarrow \min \tilde{E}_{\text{aug}}(w) \quad \frac{\partial}{\partial w} \tilde{E}_{\text{aug}}(w) = 0 + (w - w^*)^T H + \frac{2\lambda}{N} w = 0 \quad \begin{aligned} & \text{from 題目} \\ & \frac{1}{2} (w - w^*)^T H (w - w^*) \\ & = \frac{1}{2} \|w - w^*\|^2 \\ & \text{(matrix form)} \end{aligned}$$

$$Hw - Hw^* + \frac{2\lambda}{N} w = 0$$

$$D = \underbrace{- (w - w^*)^T H}_{\text{matrix form}}$$

$$w \left( H + \frac{2\lambda}{N} I \right) = Hw^*$$

$$w = \left( H + \frac{2\lambda}{N} I \right)^{-1} Hw^*$$

$$\text{---} \quad \text{---}$$



8.

$$(z, 0) \quad (p, z) \quad (-z, 0) \quad p \geq 0, \quad \begin{cases} h(x) = w_0 \\ h(x) = w_0 + w_1 x \end{cases}$$

for  $h(x) = w_0$  if  $x_1, y_1, x_2, y_2$

$$E_m = (w_0 - y_1)^2 + (w_0 - y_2)^2 = 2w_0^2 - 2(y_1 + y_2)w_0 + (y_1^2 + y_2^2)$$

$$\rightarrow \left[ w_0 - \left( \frac{y_1 + y_2}{2} \right) \right]^2 + \Delta.$$

if  $w_0 = \frac{y_1 + y_2}{2}$  有誤差、誤差

$$\text{for } (z, 0), (-z, 0) \rightarrow w_0 = 0 \rightarrow \frac{y_1^2 + y_2^2}{2} = (h(x_3) - y_3)^2 = (0 - z)^2 = z^2$$

$$\text{for } (z, 0) (p, z) \rightarrow w_0 = z \rightarrow \frac{y_1^2 + y_2^2}{2} = (h(x_3) - y_3)^2 = (z - z)^2 = 0$$

$$\text{for } (-1, 0) (p, z) \rightarrow w_0 = -1 \rightarrow \frac{y_1^2 + y_2^2}{2} = (h(x_3) - y_3)^2 = (1 - z)^2 = (1 - z)^2 = 1$$

$$\rightarrow \frac{y_1^2 + y_2^2}{2} = \frac{1}{3}(4 + 1 + 1) = 2.$$

for  $h(x) = w_0 + w_1 x$

$$w_1 x_1 + w_0 = y_1 \quad \text{--- ①}$$

$$w_1 x_2 + w_0 = y_2 \quad \text{--- ②}$$

$$\begin{aligned} \text{①} - \text{②} &= (x_1 - x_2) w_1 = y_1 - y_2 \\ w_1 &= \frac{y_1 - y_2}{x_1 - x_2} \end{aligned}$$

$$w_1 x_2 = w_1 x_1 x_2 + w_0 x_2 = y_1 x_2$$

$$\begin{aligned} \text{--- ②} x_1 &= w_1 x_1 x_2 + w_0 x_1 = y_2 x_1 \\ w_0 &= \frac{y_1 x_2 - y_2 x_1}{x_2 - x_1} \end{aligned}$$

for  $(\geq, 0) (-\geq, 0)$

$P_{<2}$

$$w_1 = 0/4 = 0 \quad w_0 = 0$$

$$\rightarrow \frac{w_1^2 + w_0^2}{2} = (h(x) - y)^2$$

$$= (0 - \geq)^2 = \phi_-$$

for  $(\geq, 0) (\leq, 2)$

$(-\geq, 0)$

$$w_1 = \frac{2}{P-2} \quad w_0 = \frac{-4}{P-2}$$

$$\rightarrow \frac{w_1^2 + w_0^2}{2} = \left( \frac{-4}{P-2} + \frac{-4}{P-2} - 0 \right)^2 = \frac{16}{(P-2)^2}$$

for  $(-\geq, 0) (\leq, 2)$

$(\geq, 0)$

$$w_1 = \frac{-2}{-2-P} \quad w_0 = \frac{4}{P+2}$$

$$\rightarrow \frac{w_1^2 + w_0^2}{2} = \left( \frac{4}{P+2} + \frac{4}{P+2} - 0 \right)^2 = \frac{64}{(P+2)^2}$$

$$\rightarrow \text{虚数解} = \frac{1}{3} \left( \frac{64}{(P-2)^2} + \frac{64}{(P+2)^2} + 4 \right) = 2$$

$$\frac{64}{(P-2)^2} + \frac{64}{(P+2)^2} = 2$$

$$\rightarrow P = \pm 2\sqrt{9+4\sqrt{6}}, \pm 2i\sqrt{9+4\sqrt{6}}$$

$$\rightarrow 2\sqrt{9+4\sqrt{6}} = 8.67132$$

- B** 9. For  $N$  labels  $y_1, y_2, \dots, y_N$  generated i.i.d. from some distribution of mean 0 and variance  $\sigma^2$ . Partition the first  $N - K$  labels to be the training data, and the last  $K$  labels to be the validation data. If we “estimate” the mean by the “target” of 0, the expected validation error

$$\mathbb{E} \left( \frac{1}{K} \sum_{n=N-K+1}^N (y_n - 0)^2 \right),$$

where the expectation is taken on the process of generating  $N$  labels, is simply  $\sigma^2$  by the independence of the  $y_n$ 's. Now, assume that we take the average on the first  $N - K$  examples to estimate the mean instead. That is,

$$\bar{y} = \frac{1}{N-K} \sum_{n=1}^{N-K} y_n$$

What is the expected validation error

$$\mathbb{E} \left( \frac{1}{K} \sum_{n=N-K+1}^N (y_n - \bar{y})^2 \right)?$$

Choose the correct answer; explain your answer.

- [a]  $\sigma^2$
- [b]**  $\sigma^2 + \frac{1}{N-K}\sigma^2$
- [c]  $\frac{1}{K(N-K)}\sigma^2$
- [d]  $\sigma^2 + \frac{1}{K(N-K)}\sigma^2$
- [e] none of the other choices

## Learning Principles

- A** 10. In Lecture 9, we talked about the probability to fit data perfectly when the labels are random. For instance, page 5 of Lecture 9 shows that the probability of fitting the data perfectly with positive rays is  $\frac{N+1}{2^N}$ . Consider 4 different points on the unit circle in  $\mathbb{R}^2$  as input vectors  $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$  and  $\mathbf{x}_4$ , and a 2D perceptron model that minimizes  $E_{\text{in}}(\mathbf{w})$  (0/1 error) to the lowest possible value. One way to measure the power of the model is to consider four random labels  $y_1, y_2, y_3, y_4$ , each in  $\pm 1$  and generated by i.i.d. fair coin flips, and then compute

$$\mathbb{E}_{y_1, y_2, y_3, y_4} \left( \min_{\mathbf{w} \in \mathbb{R}^{2+1}} E_{\text{in}}(\mathbf{w}) \right).$$

For a perfect fitting,  $\min E_{\text{in}}(\mathbf{w})$  will be 0; for a less perfect fitting (when the data is not linearly separable),  $\min E_{\text{in}}(\mathbf{w})$  will be some non-zero value. The expectation above averages over all 16 possible combinations of  $y_1, y_2, y_3, y_4$ . What is the value of the expectation? Choose the correct answer; explain your answer.

- [a]** 1/32
- [b] 2/32
- [c] 3/32
- [d] 5/32
- [e] 8/32

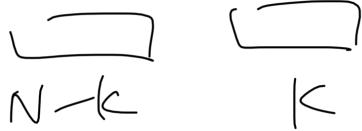
(Note: It can be shown that 1 minus twice the expected value above is the same as the so-called empirical Rademacher complexity of 2D perceptrons. Rademacher complexity, similar to the VC dimension, is another tool to measure the complexity of a hypothesis set. If a hypothesis set shatters some data points, zero  $E_{\text{in}}$  can always be achieved and thus Rademacher complexity is 1; if a hypothesis set cannot shatter some data points, Rademacher complexity provides a soft measure of how “perfect” the hypothesis set is.)

$$E[y_n] = \mu$$

$$E[(y_{n-K})^2] = \sigma^2$$

$$E[y_n^2] = \sigma^2$$

q.



平均值

跟後面的計算無關

$$E\left(\frac{1}{K} \sum_{n=N-K+1}^N (y_n - \bar{y})^2\right)$$

where the expectation is taken on the process of generating  $N$  labels, is simply  $\sigma^2$  by the independence of the  $y_n$ 's. Now, assume that we take the average on the first  $N - K$  examples to estimate the mean instead. That is,

$$\bar{y} = \frac{1}{N-K} \sum_{n=1}^{N-K} y_n$$

What is the expected validation error

$$E\left(\frac{1}{K} \sum_{n=N-K+1}^N (y_n - \bar{y})^2\right)$$

Choose the correct answer; explain your answer.

- [a]  $\sigma^2$
- [b]  $\sigma^2 + \frac{1}{N-K}\sigma^2$
- [c]  $\frac{1}{K(N-K)}\sigma^2$
- [d]  $\sigma^2 + \frac{1}{K(N-K)}\sigma^2$
- [e] none of the other choices

### Learning Principles

10. In Lecture 9, we talked about the probability to fit data perfectly when the labels are random. For instance, page 5 of Lecture 9 shows that the probability of fitting the data perfectly with positive rays is  $\frac{N+1}{2^N}$ . Consider 4 different points on the unit circle in  $\mathbb{R}^2$  as input vectors  $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$  and  $\mathbf{x}_4$ .

$$E\left[\frac{1}{K} \sum_{N-K+1}^N (y_n^2 - 2y_n \cdot \bar{y} + \bar{y}^2)\right]$$

$$\frac{1}{K} \sum_{N-K+1}^N [E(y_n^2) - 2E[y_n \cdot \bar{y}] + E[\bar{y}^2]]$$

$$= \frac{1}{K} \sum_{N-K+1}^N \left( \sigma^2 - 2E[y_n \cdot \frac{1}{N-K} \sum_{n=1}^{N-K} y_n] + E\left[\left(\frac{1}{N-K} \sum_{n=1}^{N-K} y_n\right)^2\right]\right)$$

- [a]  $\sigma^2$
- [b]  $\sigma^2 + \frac{1}{N-K}\sigma^2$
- [c]  $\frac{1}{K(N-K)}\sigma^2$
- [d]  $\sigma^2 + \frac{1}{K(N-K)}\sigma^2$
- [e] none of the other choice

$$E(y_1 + y_2 + \dots + y_N) = 0$$

$$E\left(\sum_i y_i\right) = 0$$

$$E(y_1 + y_2 + \dots + y_N) = 0$$

$$= E\left(\sum_{n=1}^{N-K} y_n\right)$$

$$\sum_{n=1}^{N-K} E(y_n) = (N-K) \cdot 0 = 0$$

$$\begin{aligned}
 &= \left( \bar{y}^2 \right) - \frac{1}{K} \sum_{N-k+1}^N \left[ \left[ 2 \times \frac{1}{N-k} E \left[ y_n \cdot \sum_{j=1}^{N-k} y_j \right] \right] - \frac{1}{(N-k)^2} \left[ \sum_{j=1}^{N-k} y_j \cdot \frac{N-k}{\sum_{j=1}^{N-k} y_j} \right] \right] \\
 &= \bar{y}^2 - \frac{1}{K} \sum_{N-k+1}^N \left[ \frac{2}{N-k} E \left[ y_n \cdot (N-k) E[y_n] \right] - \left( \frac{1}{N-k} \right)^2 E \left[ \left( \sum_{j=1}^{N-k} y_j \right)^2 \right] \right] \\
 &= \bar{y}^2 + \frac{1}{K} \sum_{N-k+1}^N \left( \frac{1}{N-k} \right)^2 E \left[ \left( \sum_{j=1}^{N-k} y_j \right)^2 \right]
 \end{aligned}$$

$$\text{Var}_\text{model} = \frac{1}{N-k} \sum_{j=1}^{N-k} (y_j - \mu)^2 = E[(y_j - \mu)^2] = \bar{y}^2$$

$$\sum_{j=1}^{N-k} (y_j - \mu)^2 = Nk E[(y_j - \mu)^2] = \bar{y}^2$$

$$\mu = 0 \quad \sum_{j=1}^{N-k} (y_j)^2 = (Nk) E[(y_j)^2] = \bar{y}^2$$

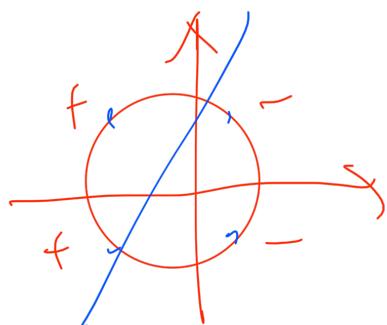
$$= \bar{y}^2 + \frac{1}{K} \cdot \sum_{N-k+1}^N \left( \frac{1}{N-k} \right)^2 (N-k) E(\bar{y}^2)$$

$$= \bar{y}^2 + \frac{1}{K} \cdot K \cdot \left( \frac{1}{N-k} \right)^2 \cdot (N-k) \cdot \bar{y}^2$$

$$= \bar{y}^2 + \frac{1}{N-k} \cdot \bar{y}^2$$

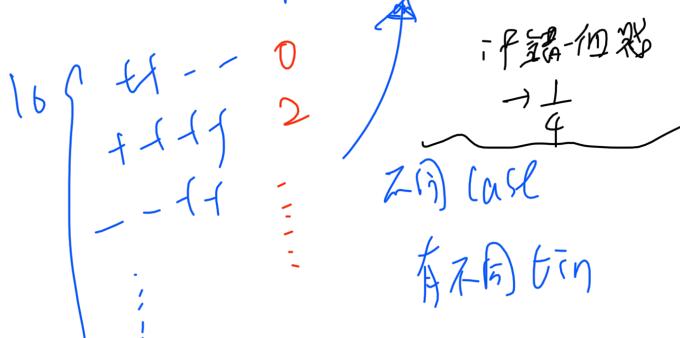
10. 有  $\oplus$  點分布 在 圖上

$$E_{\text{Gauss}} \left( \min_{w \in \mathbb{R}^4} E_m(w) \right)$$



$x$  的 分佈

最優化



$$\nabla E_m = 0$$

對稱的取平均

→ 算是取期望值

右) 正確第 3.

不能分開的

→ 也就是多樣的有



正負一律線皆不能處理

$$\text{this 2 } E_m = \frac{1}{4} \quad \text{other } E_m = 0 \text{ by 單項因}$$

4-1 錯誤其中 1 四錯

$$\rightarrow E(E_m) = \frac{2 \cdot \frac{1}{4} + 14 \cdot 0}{16} = \frac{1}{32} *$$

11. Consider a binary classifier  $g$  such that

$$\begin{aligned} P(g(\mathbf{x}) = -1 | y = +1) &= \epsilon_+ \\ P(g(\mathbf{x}) = +1 | y = -1) &= \epsilon_-. \end{aligned}$$

When deploying the classifier to a test distribution of  $P(y = +1) = P(y = -1) = 1/2$ , we get  $E_{\text{out}}(g) = \frac{1}{2}\epsilon_+ + \frac{1}{2}\epsilon_-$ . Now, if we deploy the classifier to another test distribution  $P(y = +1) = p$  instead of  $1/2$ , the  $E_{\text{out}}(g)$  under this test distribution will then change to a different value. Note that under this test distribution, a constant classifier  $g_-$  that always predicts  $-1$  will suffer from  $E_{\text{out}}(g_-) = p$  as it errors on all the positive examples. At what  $p$ , if its value is between  $[0, 1]$ , will our binary classifier  $g$  be as good as (or as bad as) the constant classifier  $g_-$  in terms of  $E_{\text{out}}$ ? Choose the correct answer; explain your answer.

- [a]  $p = \frac{\epsilon_+}{\epsilon_- - \epsilon_+ + 1}$
- [b]  $p = \frac{\epsilon_-}{\epsilon_- - \epsilon_+ + 1}$**
- [c]  $p = \frac{1 - \epsilon_+}{\epsilon_- - \epsilon_+ + 1}$
- [d]  $p = \frac{1 - \epsilon_-}{\epsilon_- - \epsilon_+ + 1}$
- [e] none of the other choices

$$\frac{1}{2} = C = \frac{2}{N} \cdot \frac{1}{N}$$

$$C\lambda = \frac{1}{2}$$

$$C = \frac{1}{2N}$$

## Experiments with Regularized Logistic Regression

Consider L2-regularized logistic regression with third-order polynomial transformation.

$$\mathbf{w}_\lambda = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{\lambda}{N} \|\mathbf{w}\|^2 + \frac{1}{N} \sum_{n=1}^N \ln(1 + \exp(-y_n \mathbf{w}^T \Phi_3(\mathbf{x}_n))),$$

$$\frac{1}{2} \quad \frac{1}{2N}$$

$$\frac{\lambda}{2N} \quad \frac{1}{2N}$$

Here  $\Phi_3$  is the third-order polynomial transformation introduced in Lecture 6 (with  $Q = 3$ ), defined as

$$\Phi_3(\mathbf{x}) = (1, x_1, x_2, \dots, x_d, x_1^2, x_1x_2, \dots, x_1x_d, x_2^2, x_2x_3, \dots, x_2x_d, \dots, x_d^2, x_1^3, x_1^2x_2, \dots, x_d^3)$$

(2.3.4)

2.3 Next, we will take the following file as our training data set  $\mathcal{D}$ :

[http://www.csie.ntu.edu.tw/~htlin/course/ml21fall/hw4/hw4\\_train.dat](http://www.csie.ntu.edu.tw/~htlin/course/ml21fall/hw4/hw4_train.dat)

and the following file as our test data set for evaluating  $E_{\text{out}}$ : (1), (2, 3, 4), (4, 6, 8, 9, 12, 16)

[http://www.csie.ntu.edu.tw/~htlin/course/ml21fall/hw4/hw4\\_test.dat](http://www.csie.ntu.edu.tw/~htlin/course/ml21fall/hw4/hw4_test.dat)

8, 12, 16, 18, 24, 32

We call the algorithm for solving the problem above as  $A_\lambda$ . The problem guides you to use LIBLINEAR, a machine learning package developed in our university, to solve this problem. In addition to using the default options, what you need to do when running LIBLINEAR are

27, 36, 48, 54

3, 3, 4, 3, 4, 2, 4

- set option  $-s 0$ , which corresponds to solving regularized logistic regression
- set option  $-c C$ , with a parameter value of  $C$  calculated from the  $\lambda$  that you want to use; read README of the software package to figure out how  $C$  and your  $\lambda$  should relate to each other
- set option  $-e 0.000001$ , which corresponds to getting a solution that is really really close to the optimal solution

row = 4  
0, 1, 2, 3, 4

LIBLINEAR can be called from the command line or from major programming languages like python. If you run LIBLINEAR in the command line, please include screenshots of the commands/results; if you run LIBLINEAR from any programming language, please include screenshots of your code.

We will consider the data set as a *binary classification problem* and take the “regression for classification” approach with regularized logistic regression (see Page 6 of Lecture 10). So please evaluate all errors below with the 0/1 error.

$$\text{II. } P(y=-1) = 1 - p \quad \xrightarrow{\text{out}} \text{always predicts -1}$$

$$P = p \cdot E_F + (1-p) E_- \quad \xrightarrow{\text{out}} \begin{cases} E_F & \rightarrow E+ \\ E_- & \rightarrow E- \end{cases}$$

$y = \begin{cases} 1 & g \geq 0 \\ -1 & g < 0 \end{cases}$

$$P(y=-1) = p$$

$$g \rightarrow \begin{cases} 1 & g \geq 0 \\ -1 & g < 0 \end{cases} \quad \rightarrow E_{\text{out}}(g)$$

$$P = \frac{1}{2} \left( 1 - \frac{E_-}{E_F + E_-} \right) = p \cdot E_F + (1-p) E_-$$

$$P = p \cdot E_F + (1-p) E_-$$

$$E_{\text{out}}(-) = p$$

$$-E_- = -p + p \cdot E_F - p E_-$$

$$E_- = p - p \cdot E_F + p E_-$$

$$\frac{E_-}{1 - E_F + E_-} = p$$

$$\frac{E_-}{E_- - E_F + 1} = p \quad \xrightarrow{\text{B}}$$

12. ✓ Select the best  $\lambda^*$  in a cheating manner as

$$\underset{\log_{10} \lambda \in \{-4, -2, 0, 2, 4\}}{\operatorname{argmin}} E_{\text{out}}(\mathbf{w}_\lambda).$$

$$l = \frac{1}{2\lambda} \quad \lambda = 2^0$$

D

Break the tie, if any, by selecting the largest  $\lambda$ . What is  $\log_{10}(\lambda^*)$ ? Choose the closest answer; provide your command/code.

- [a] -4
- [b] -2
- [c] 0
- [d] 2**
- [e] 4

13. Select the best  $\lambda^*$  as

$$\underset{\log_{10} \lambda \in \{-4, -2, 0, 2, 4\}}{\operatorname{argmin}} E_{\text{in}}(\mathbf{w}_\lambda).$$

$$4 = -2 = 100\%$$

- [a] -4
- [b] -2** ✓
- [c] 0
- [d] 2
- [e] 4

14. Now split the given training examples in  $\mathcal{D}$  to two sets: the first 120 examples as  $\mathcal{D}_{\text{train}}$  and 80 as  $\mathcal{D}_{\text{val}}$ . (Ideally, you should randomly do the 120/80 split. Because the given examples are already randomly permuted, however, we would use a fixed split for the purpose of this problem). Run  $\mathcal{A}_\lambda$  on only  $\mathcal{D}_{\text{train}}$  to get  $\mathbf{w}_\lambda^-$  (the weight vector within the  $g^-$  returned), and validate  $\mathbf{w}_\lambda^-$  with  $\mathcal{D}_{\text{val}}$  to get  $E_{\text{val}}(\mathbf{w}_\lambda^-)$ . Select the best  $\lambda^*$  as

$$\underset{\log_{10} \lambda \in \{-4, -2, 0, 2, 4\}}{\operatorname{argmin}} E_{\text{val}}(\mathbf{w}_\lambda^-).$$

C Break the tie, if any, by selecting the largest  $\lambda$ . Then, estimate  $E_{\text{out}}(\mathbf{w}_{\lambda^*}^-)$  with the test set. What is the value of  $E_{\text{out}}(\mathbf{w}_{\lambda^*}^-)$ ? Choose the closest answer; provide your command/code.

- [a] 0.081
- [b] 0.078**
- [c] 0.075
- [d] 0.072
- [e] 0.069

$$\lambda = 100$$

$$92.5 \rightarrow 0.075$$

15. ✓ For the  $\lambda^*$  selected in the previous problem, compute  $\mathbf{w}_{\lambda^*}$  by running  $\mathcal{A}_{\lambda^*}$  with the full training set  $\mathcal{D}$ . Then, estimate  $E_{\text{out}}(\mathbf{w}_{\lambda^*})$  with the test set. What is the value of  $E_{\text{out}}(\mathbf{w}_{\lambda^*})$ ? Choose the closest answer; provide your command/code.

- [a] 0.081
- [b] 0.078
- [c] 0.075
- [d] 0.072**
- [e] 0.069

$$100 \quad 242.49 \quad 700$$

$$0.0725$$

16. Now split the given training examples in  $\mathcal{D}$  to five folds, the first 40 being fold 1, the next 40 being fold 2, and so on. Again, we take a fixed split because the given examples are already randomly permuted. Select the best  $\lambda^*$  as

$$\underset{\log_{10} \lambda \in \{-4, -2, 0, 2, 4\}}{\operatorname{argmin}} E_{\text{cv}}(\mathcal{A}_\lambda).$$

Break the tie, if any, by selecting the largest  $\lambda$ . What is the value of  $E_{\text{cv}}(\mathcal{A}_{\lambda^*})$ ? Choose the closest answer; provide your command/code.

- [a] 0.07
- [b] 0.08**
- [c] 0.09
- [d] 0.10
- [e] 0.11