Homework 1 Review

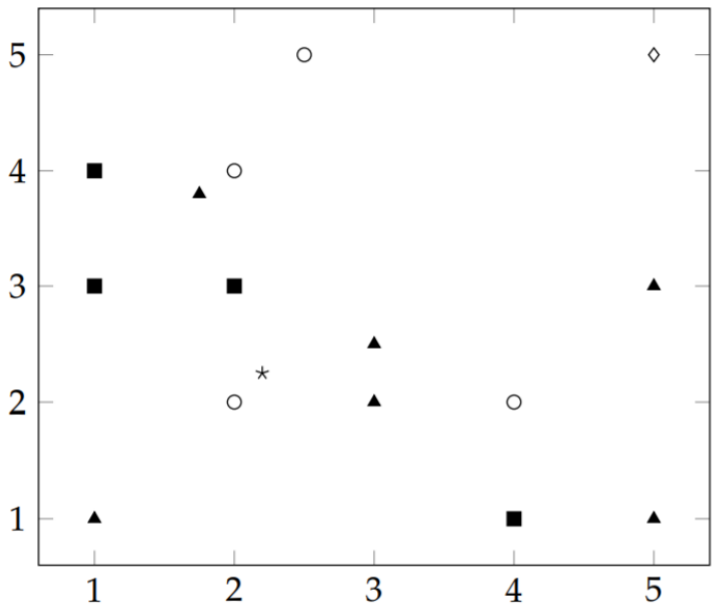
Chung-Wei Lee

September 16

Outline

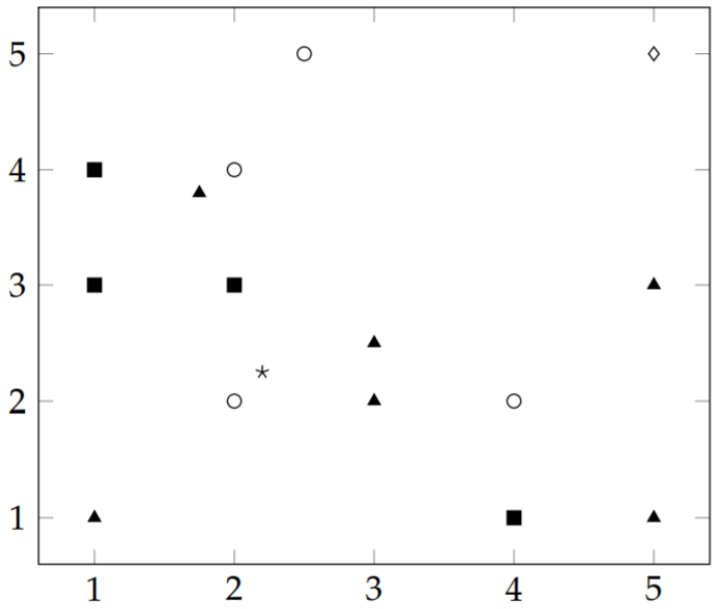
Problem 1

We denote the total number of training points as N and consider K-nearest-neighbor (KNN) classifier with L2 distance.



Problem 1

We denote the total number of training points as N and consider K-nearest-neighbor (KNN) classifier with L2 distance.

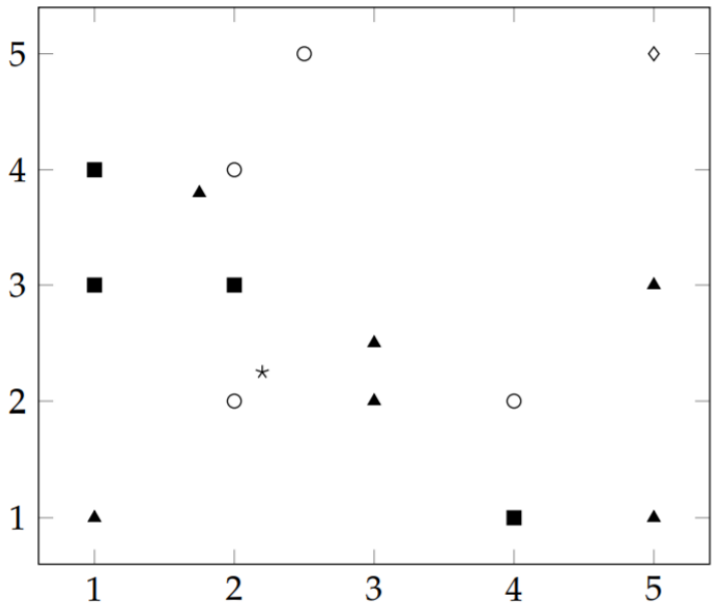


(a)

What is the prediction for the test point star when $K = 4$? Explain why

Problem 1

We denote the total number of training points as N and consider K-nearest-neighbor (KNN) classifier with L2 distance.

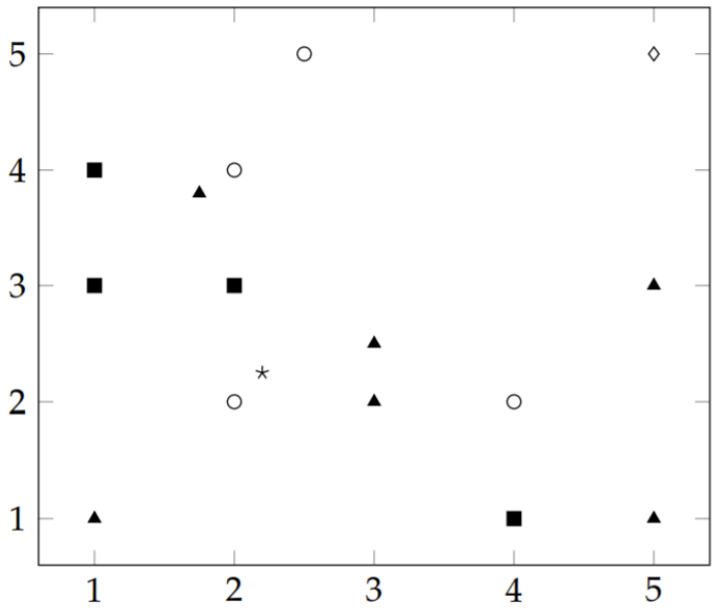


(a)

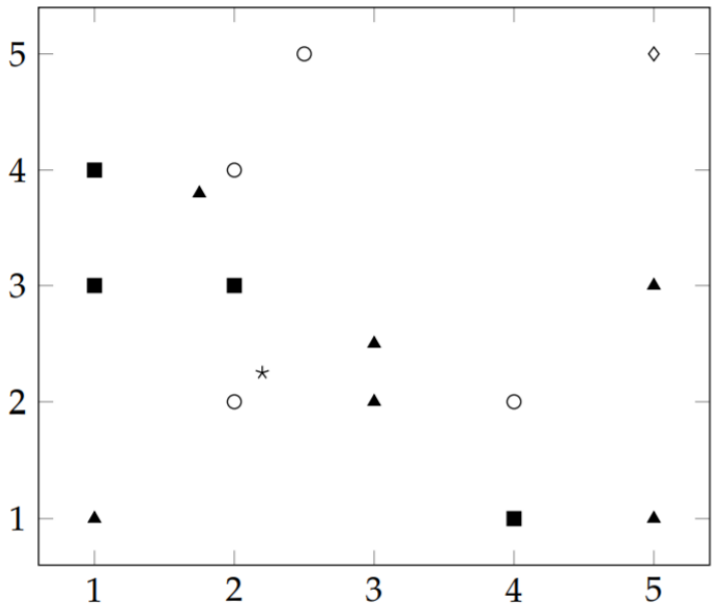
What is the prediction for the test point star when $K = 4$? Explain why

Triangle. The 4 nearest neighbors: 1 square, 1 circle, and 2 triangles.

Problem 1



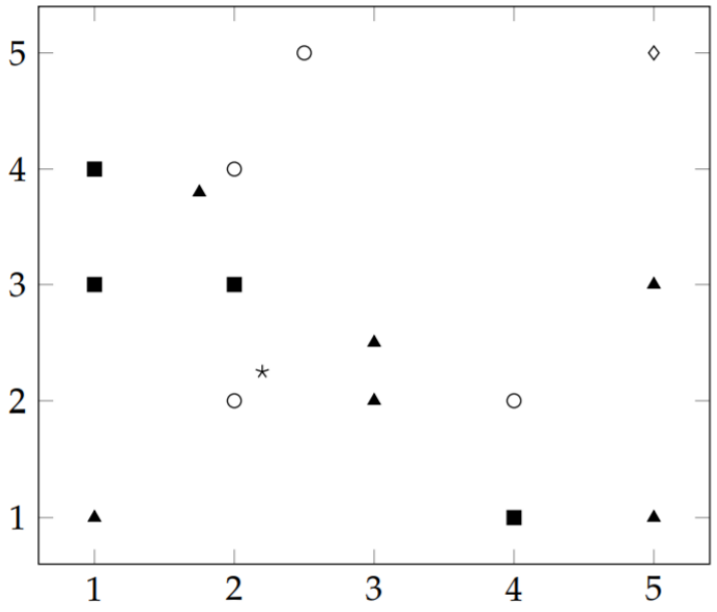
Problem 1



(b)

What is the diamond classified as for $K = N$? Explain why.

Problem 1

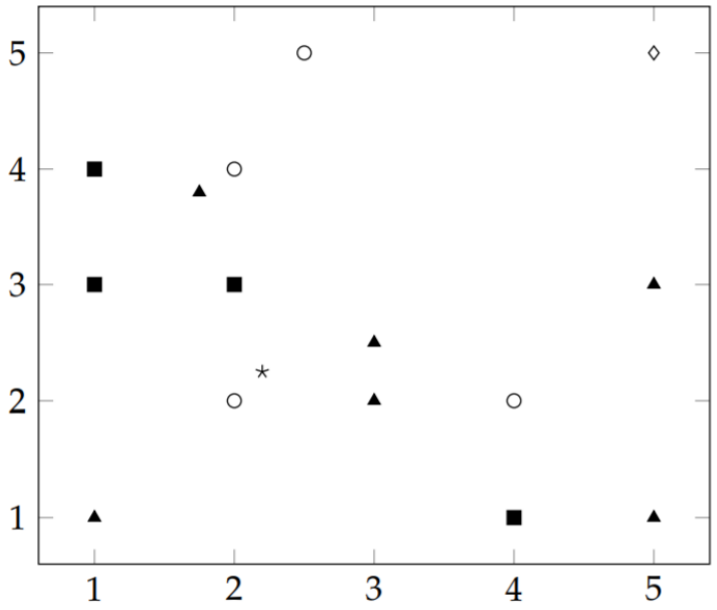


(b)

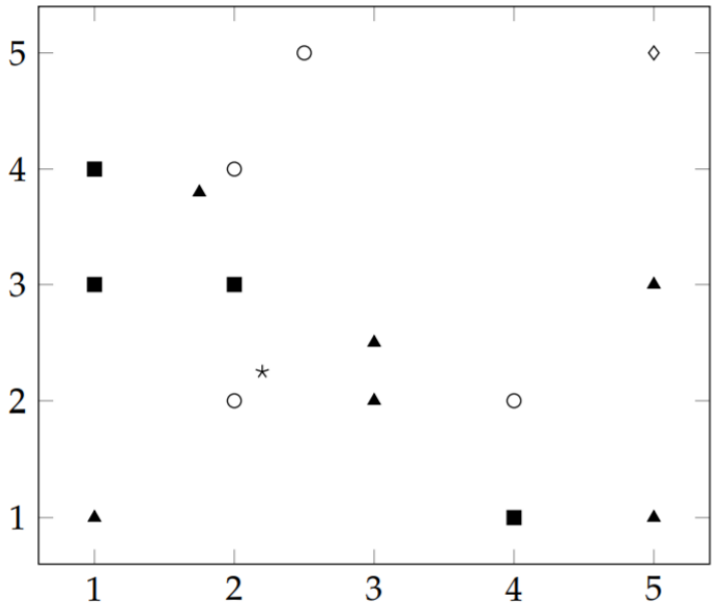
What is the diamond classified as for $K = N$? Explain why.

Triangle. When $K = N$ the prediction is always the majority label of the training set. Triangle is the majority in this example.

Problem 1



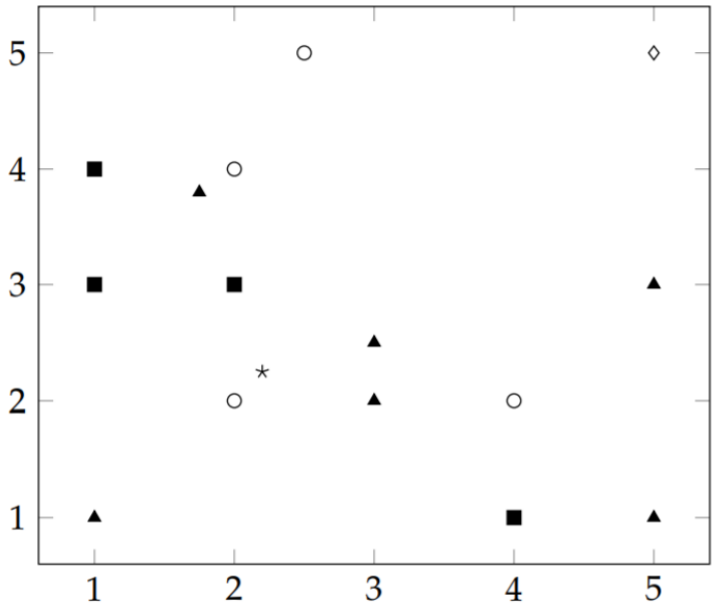
Problem 1



(c)

Suppose one performs leave-one-out validation (that is, N-fold cross validation) to choose the best K. List the triangles that are correctly classified (as a validation point) in this process for the run with $K = 1$.

Problem 1

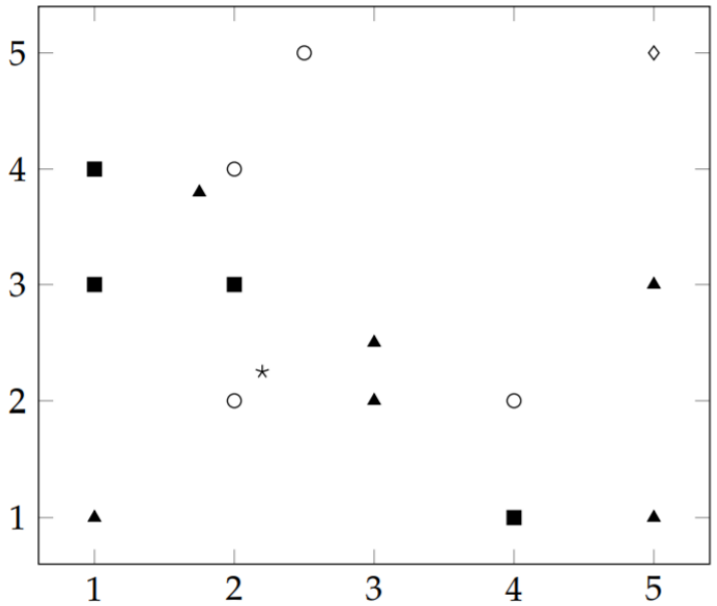


(c)

Suppose one performs leave-one-out validation (that is, N-fold cross validation) to choose the best K. List the triangles that are correctly classified (as a validation point) in this process for the run with $K = 1$.

Recall leave-one-out validation: use each point in turn as a validation dataset and use the others as a training dataset.

Problem 1



(c)

Suppose one performs leave-one-out validation (that is, N-fold cross validation) to choose the best K. List the triangles that are correctly classified (as a validation point) in this process for the run with $K = 1$.

Recall leave-one-out validation: use each point in turn as a validation dataset and use the others as a training dataset. The two triangles with x-value 3 are correctly classified.

Problem 2.1

In this problem we generalize the L2 regularization by

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^D} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$$

where $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ is a positive definite matrix. Find the closed form of $\boldsymbol{w}'_*$.

Problem 2.1

In this problem we generalize the L2 regularization by

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^D} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$$

where $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ is a positive definite matrix. Find the closed form of $\boldsymbol{w}'_*$.

Set $f(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$.

Problem 2.1

In this problem we generalize the L2 regularization by

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^D} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$$

where $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ is a positive definite matrix. Find the closed form of $\boldsymbol{w}'_*$.

Set $f(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$. Then we have

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}} \ f(\boldsymbol{w})$$

Problem 2.1

In this problem we generalize the L2 regularization by

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^D} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$$

where $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ is a positive definite matrix. Find the closed form of $\boldsymbol{w}'_*$.

Set $f(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$. Then we have

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}} \; f(\boldsymbol{w}) \;\Rightarrow\; \nabla f(\boldsymbol{w}'_*) = 0$$

Problem 2.1

In this problem we generalize the L2 regularization by

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^D} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$$

where $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ is a positive definite matrix. Find the closed form of $\boldsymbol{w}'_*$.

Set $f(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$. Then we have

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}} \ f(\boldsymbol{w}) \ \Rightarrow \ \nabla f(\boldsymbol{w}'_*) = 0$$

Since $\nabla f(\boldsymbol{w}) = 2\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + 2\boldsymbol{M}\boldsymbol{w}$, we get

Problem 2.1

In this problem we generalize the L2 regularization by

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^D} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$$

where $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ is a positive definite matrix. Find the closed form of $\boldsymbol{w}'_*$.

Set $f(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$. Then we have

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}} \ f(\boldsymbol{w}) \ \Rightarrow \ \nabla f(\boldsymbol{w}'_*) = 0$$

Since $\nabla f(\boldsymbol{w}) = 2\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + 2\boldsymbol{M}\boldsymbol{w}$, we get

$$\nabla f(\boldsymbol{w}'_*) = 2\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w}'_* - \boldsymbol{y}) + 2\boldsymbol{M}\boldsymbol{w}'_* = 0$$

Problem 2.1

In this problem we generalize the L2 regularization by

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^D} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$$

where $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ is a positive definite matrix. Find the closed form of $\boldsymbol{w}'_*$.

Set $f(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$. Then we have

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}} \; f(\boldsymbol{w}) \;\Rightarrow\; \nabla f(\boldsymbol{w}'_*) = 0$$

Since $\nabla f(\boldsymbol{w}) = 2\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + 2\boldsymbol{M}\boldsymbol{w}$, we get

$$\nabla f(\boldsymbol{w}'_*) = 2\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w}'_* - \boldsymbol{y}) + 2\boldsymbol{M}\boldsymbol{w}'_* = 0 \;\Rightarrow\; \left(\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{M}\right)\boldsymbol{w}'_* = \boldsymbol{X}^\top\boldsymbol{y}$$

Problem 2.1

In this problem we generalize the L2 regularization by

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^D} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$$

where $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ is a positive definite matrix. Find the closed form of $\boldsymbol{w}'_*$.

Set $f(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$. Then we have

$$\boldsymbol{w}'_* = \arg\min_{\boldsymbol{w}} \; f(\boldsymbol{w}) \;\Rightarrow\; \nabla f(\boldsymbol{w}'_*) = 0$$

Since $\nabla f(\boldsymbol{w}) = 2\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + 2\boldsymbol{M}\boldsymbol{w}$, we get

$$\nabla f(\boldsymbol{w}'_*) = 2\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w}'_* - \boldsymbol{y}) + 2\boldsymbol{M}\boldsymbol{w}'_* = 0 \;\Rightarrow\; \left(\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{M}\right)\boldsymbol{w}'_* = \boldsymbol{X}^\top\boldsymbol{y}$$

As $\boldsymbol{M}$ is positive definite, we conclude $\boldsymbol{w}'_* = \left(\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{M}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$

Matrix Calculus Example

Consider $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{A} \in \mathbb{R}^{D\times D}$, we have

Matrix Calculus Example

Consider $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{A} \in \mathbb{R}^{D\times D}$, we have

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_D} \end{bmatrix}^\top$$

Matrix Calculus Example

Consider $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{A} \in \mathbb{R}^{D\times D}$, we have

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_D} \end{bmatrix}^\top$$

Consider entry i, we have

Matrix Calculus Example

Consider $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{A} \in \mathbb{R}^{D\times D}$, we have

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_D} \end{bmatrix}^\top$$

Consider entry i, we have

$$\frac{\partial f}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right)$$

Matrix Calculus Example

Consider $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{A} \in \mathbb{R}^{D\times D}$, we have

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_D} \end{bmatrix}^\top$$

Consider entry i, we have

$$\begin{aligned} \frac{\partial f}{\partial x_i} &= \frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) \\ &= \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{D} x_j \cdot (\boldsymbol{A}\boldsymbol{x})_j\right) \end{aligned}$$

Matrix Calculus Example

Consider $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{A} \in \mathbb{R}^{D\times D}$, we have

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_D} \end{bmatrix}^\top$$

Consider entry i, we have

$$\begin{aligned} \frac{\partial f}{\partial x_i} &= \frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) \\ &= \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{D} x_j \cdot (\boldsymbol{A}\boldsymbol{x})_j\right) \\ &= \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{D} x_j \sum_{k=1}^{D} A_{jk}x_k\right) \end{aligned}$$

Matrix Calculus Example

Consider $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{A} \in \mathbb{R}^{D\times D}$, we have

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_D} \end{bmatrix}^\top$$

Consider entry i, we have

$$\begin{aligned}
\frac{\partial f}{\partial x_i} &= \frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) \\
&= \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{D} x_j \cdot (\boldsymbol{A}\boldsymbol{x})_j\right) \\
&= \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{D} x_j \sum_{k=1}^{D} A_{jk}x_k\right) \\
&= \frac{\partial}{\partial x_i}\left(x_i \sum_{k=1}^{D} A_{ik}x_k + \sum_{j\neq i} x_j \sum_{k=1}^{D} A_{jk}x_k\right)
\end{aligned}$$

Matrix Calculus Example

$$\frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) = \frac{\partial}{\partial x_i}\left(x_i \sum_{k=1}^{D} A_{ik}x_k + \sum_{j\neq i} x_j \sum_{k=1}^{D} A_{jk}x_k\right)$$

Matrix Calculus Example

$$\frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) = \frac{\partial}{\partial x_i}\left(x_i \sum_{k=1}^{D} A_{ik}x_k + \sum_{j\neq i} x_j \sum_{k=1}^{D} A_{jk}x_k\right)$$

$$= x_i A_{ii} + \sum_{k=1}^{D} A_{ik}x_k + \frac{\partial}{\partial x_i}\left(\sum_{j\neq i} x_j \sum_{k=1}^{D} A_{jk}x_k\right)$$

Matrix Calculus Example

$$\begin{aligned}
\frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) &= \frac{\partial}{\partial x_i}\left(x_i \sum_{k=1}^{D} A_{ik}x_k + \sum_{j\neq i} x_j \sum_{k=1}^{D} A_{jk}x_k\right) \\
&= x_i A_{ii} + \sum_{k=1}^{D} A_{ik}x_k + \frac{\partial}{\partial x_i}\left(\sum_{j\neq i} x_j \sum_{k=1}^{D} A_{jk}x_k\right) \\
&= x_i A_{ii} + \sum_{k=1}^{D} A_{ik}x_k + \sum_{j\neq i} x_j A_{ji}
\end{aligned}$$

Matrix Calculus Example

$$\frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) = \frac{\partial}{\partial x_i}\left(x_i \sum_{k=1}^{D} A_{ik}x_k + \sum_{j\neq i} x_j \sum_{k=1}^{D} A_{jk}x_k\right)$$

$$= x_i A_{ii} + \sum_{k=1}^{D} A_{ik}x_k + \frac{\partial}{\partial x_i}\left(\sum_{j\neq i} x_j \sum_{k=1}^{D} A_{jk}x_k\right)$$

$$= x_i A_{ii} + \sum_{k=1}^{D} A_{ik}x_k + \sum_{j\neq i} x_j A_{ji}$$

$$= \sum_{k=1}^{D} A_{ik}x_k + \sum_{j=1}^{D} x_j A_{ji}$$

Matrix Calculus Example

$$
\begin{aligned}
\frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) &= \frac{\partial}{\partial x_i}\left(x_i \sum_{k=1}^{D} A_{ik}x_k + \sum_{j\neq i} x_j \sum_{k=1}^{D} A_{jk}x_k\right) \\
&= x_i A_{ii} + \sum_{k=1}^{D} A_{ik}x_k + \frac{\partial}{\partial x_i}\left(\sum_{j\neq i} x_j \sum_{k=1}^{D} A_{jk}x_k\right) \\
&= x_i A_{ii} + \sum_{k=1}^{D} A_{ik}x_k + \sum_{j\neq i} x_j A_{ji} \\
&= \sum_{k=1}^{D} A_{ik}x_k + \sum_{j=1}^{D} x_j A_{ji} \\
&= (\boldsymbol{A}\boldsymbol{x})_i + (\boldsymbol{A}^\top \boldsymbol{x})_i
\end{aligned}
$$

Matrix Calculus Example

Using $\frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) = (\boldsymbol{A}\boldsymbol{x})_i + (\boldsymbol{A}^\top \boldsymbol{x})_i$, we conclude

Matrix Calculus Example

Using $\frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) = (\boldsymbol{A}\boldsymbol{x})_i + (\boldsymbol{A}^\top \boldsymbol{x})_i$, we conclude

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_D} \end{bmatrix}$$

Matrix Calculus Example

Using $\frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) = (\boldsymbol{A}\boldsymbol{x})_i + (\boldsymbol{A}^\top \boldsymbol{x})_i$, we conclude

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_D} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{A}\boldsymbol{x})_1 + (\boldsymbol{A}^\top \boldsymbol{x})_1 \\ (\boldsymbol{A}\boldsymbol{x})_2 + (\boldsymbol{A}^\top \boldsymbol{x})_2 \\ \vdots \\ (\boldsymbol{A}\boldsymbol{x})_D + (\boldsymbol{A}^\top \boldsymbol{x})_D \end{bmatrix}$$

Matrix Calculus Example

Using $\frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) = (\boldsymbol{A}\boldsymbol{x})_i + (\boldsymbol{A}^\top \boldsymbol{x})_i$, we conclude

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_D} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{A}\boldsymbol{x})_1 + (\boldsymbol{A}^\top \boldsymbol{x})_1 \\ (\boldsymbol{A}\boldsymbol{x})_2 + (\boldsymbol{A}^\top \boldsymbol{x})_2 \\ \vdots \\ (\boldsymbol{A}\boldsymbol{x})_D + (\boldsymbol{A}^\top \boldsymbol{x})_D \end{bmatrix} = \left(\boldsymbol{A} + \boldsymbol{A}^\top\right)\boldsymbol{x}$$

Matrix Calculus Example

Using $\frac{\partial}{\partial x_i}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) = (\boldsymbol{A}\boldsymbol{x})_i + (\boldsymbol{A}^\top \boldsymbol{x})_i$, we conclude

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_D} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{A}\boldsymbol{x})_1 + (\boldsymbol{A}^\top \boldsymbol{x})_1 \\ (\boldsymbol{A}\boldsymbol{x})_2 + (\boldsymbol{A}^\top \boldsymbol{x})_2 \\ \vdots \\ (\boldsymbol{A}\boldsymbol{x})_D + (\boldsymbol{A}^\top \boldsymbol{x})_D \end{bmatrix} = \left(\boldsymbol{A} + \boldsymbol{A}^\top\right)\boldsymbol{x}$$

When A is symmetric, we have $\nabla f(\boldsymbol{x}) = \left(\boldsymbol{A} + \boldsymbol{A}^\top\right)\boldsymbol{x} = 2\boldsymbol{A}\boldsymbol{x}$

Problem 2.2

Given a training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N) \in \mathbb{R}^D \times \mathbb{R}$, where each outcome $y_n = \boldsymbol{w}_*^{\mathrm{T}} \boldsymbol{x}_n + \epsilon_n$ with ϵ_n being an independent Gaussian noise with zero mean and variance σ^2 for some $\sigma > 0$. In other words, the probability of seeing any outcome $y \in \mathbb{R}$ given $\boldsymbol{x}_n$ is

$$\Pr\left(y \mid \boldsymbol{x}_n; \boldsymbol{w}_*, \sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(y - \boldsymbol{w}_*^{\mathrm{T}} \boldsymbol{x}_n\right)^2}{2\sigma^2}\right)$$

Problem 2.2

Given a training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N) \in \mathbb{R}^D \times \mathbb{R}$, where each outcome $y_n = \boldsymbol{w}_*^{\mathrm{T}} \boldsymbol{x}_n + \epsilon_n$ with ϵ_n being an independent Gaussian noise with zero mean and variance σ^2 for some $\sigma > 0$. In other words, the probability of seeing any outcome $y \in \mathbb{R}$ given $\boldsymbol{x}_n$ is

$$\Pr\left(y \mid \boldsymbol{x}_n; \boldsymbol{w}_*, \sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(y - \boldsymbol{w}_*^{\mathrm{T}} \boldsymbol{x}_n\right)^2}{2\sigma^2}\right)$$

Problem 2.2.1

Assume σ is fixed and given and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is invertible, find the maximum likelihood estimation for $\boldsymbol{w}_*$. In other words, first write down the probability of seeing the outcomes $y_1, \ldots, y_N$ given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ as a function of the value of $\boldsymbol{w}_*$; then find the value of $\boldsymbol{w}_*$ that maximizes this probability.

Review of Maximum Likelihood Estimation (MLE)

- The likelihood function $P(\boldsymbol{w})$: the probability of seeing the input $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ as a function of $\boldsymbol{w}$:

$$P(\boldsymbol{w}) = \prod_{n=1}^{N} \mathbb{P}\left(y_n \mid \boldsymbol{x_n}; \boldsymbol{w}\right)$$

Review of Maximum Likelihood Estimation (MLE)

- The likelihood function $P(\boldsymbol{w})$: the probability of seeing the input $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ as a function of $\boldsymbol{w}$:

$$P(\boldsymbol{w}) = \prod_{n=1}^{N} \mathbb{P}\left(y_n \mid \boldsymbol{x_n}; \boldsymbol{w}\right)$$

- Maximum Likelihood Estimation (MLE): find $\boldsymbol{w}^*$ that maximizes the probability $P(\boldsymbol{w})$:

Review of Maximum Likelihood Estimation (MLE)

- The likelihood function $P(\boldsymbol{w})$: the probability of seeing the input $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ as a function of $\boldsymbol{w}$:

$$P(\boldsymbol{w}) = \prod_{n=1}^{N} \mathbb{P}\left(y_n \mid \boldsymbol{x_n}; \boldsymbol{w}\right)$$

- Maximum Likelihood Estimation (MLE): find $\boldsymbol{w}^*$ that maximizes the probability $P(\boldsymbol{w})$:

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmax}}\ P(\boldsymbol{w}) = \underset{\boldsymbol{w}}{\operatorname{argmax}} \prod_{n=1}^{N} \mathbb{P}\left(y_n \mid \boldsymbol{x_n}; \boldsymbol{w}\right)$$

In this case, we have

In this case, we have

$$P(\boldsymbol{w}) = \prod_{n=1}^{N} \Pr\left(y_n \mid \boldsymbol{x}_n; \boldsymbol{w}, \sigma\right) = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2}{2\sigma^2}\right)$$

In this case, we have

$$P(\boldsymbol{w}) = \prod_{n=1}^{N} \Pr\left(y_n \mid \boldsymbol{x}_n; \boldsymbol{w}, \sigma\right) = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2}{2\sigma^2}\right)$$

Taking the negative log, this becomes

In this case, we have

$$P(\boldsymbol{w}) = \prod_{n=1}^{N} \Pr\left(y_n \mid \boldsymbol{x}_n; \boldsymbol{w}, \sigma\right) = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2}{2\sigma^2}\right)$$

Taking the negative log, this becomes

$$-\ln(P(\boldsymbol{w})) = \sum_{n=1}^{N} -\ln\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2}{2\sigma^2}\right)\right)$$

In this case, we have

$$P(\boldsymbol{w}) = \prod_{n=1}^{N} \Pr\left(y_n \mid \boldsymbol{x}_n; \boldsymbol{w}, \sigma\right) = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2}{2\sigma^2}\right)$$

Taking the negative log, this becomes

$$\begin{aligned} -\ln(P(\boldsymbol{w})) &= \sum_{n=1}^{N} -\ln\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2}{2\sigma^2}\right)\right) \\ &= \sum_{n=1}^{N} \ln\left(\sigma\sqrt{2\pi}\right) + \left(\frac{\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2}{2\sigma^2}\right) \end{aligned}$$

In this case, we have

$$P(\boldsymbol{w}) = \prod_{n=1}^{N} \Pr\left(y_n \mid \boldsymbol{x}_n; \boldsymbol{w}, \sigma\right) = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2}{2\sigma^2}\right)$$

Taking the negative log, this becomes

$$\begin{aligned}
-\ln(P(\boldsymbol{w})) &= \sum_{n=1}^{N} -\ln\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2}{2\sigma^2}\right)\right) \\
&= \sum_{n=1}^{N} \ln\left(\sigma\sqrt{2\pi}\right) + \left(\frac{\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2}{2\sigma^2}\right) \\
&= N \ln\sqrt{2\pi} + N \ln\sigma + \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2
\end{aligned}$$

In this case, we have

$$P(\boldsymbol{w}) = \prod_{n=1}^{N} \Pr\left(y_n \mid \boldsymbol{x}_n; \boldsymbol{w}, \sigma\right) = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2}{2\sigma^2}\right)$$

Taking the negative log, this becomes

$$\begin{aligned}
-\ln(P(\boldsymbol{w})) &= \sum_{n=1}^{N} -\ln\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2}{2\sigma^2}\right)\right) \\
&= \sum_{n=1}^{N} \ln\left(\sigma\sqrt{2\pi}\right) + \left(\frac{\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2}{2\sigma^2}\right) \\
&= N\ln\sqrt{2\pi} + N\ln\sigma + \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2 \\
&= N\ln\sqrt{2\pi} + N\ln\sigma + \frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2
\end{aligned}$$

Therefore, maximizing $P(\boldsymbol{w})$ is the same minimizing

$$-\ln(P(\boldsymbol{w})) = N \ln \sqrt{2\pi} + N \ln \sigma + \frac{1}{2\sigma^2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

Therefore, maximizing $P(\boldsymbol{w})$ is the same minimizing

$$-\ln(P(\boldsymbol{w})) = N \ln \sqrt{2\pi} + N \ln \sigma + \frac{1}{2\sigma^2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

As $\boldsymbol{w}$ only appears in the last term, the problem is equivalent to minimizing $\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$, the same objective as for least square regression.

Therefore, maximizing $P(\boldsymbol{w})$ is the same minimizing

$$-\ln(P(\boldsymbol{w})) = N \ln \sqrt{2\pi} + N \ln \sigma + \frac{1}{2\sigma^2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

As $\boldsymbol{w}$ only appears in the last term, the problem is equivalent to minimizing $\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$, the same objective as for least square regression. Therefore, the MLE for is exactly the same as the least square solution:

$$\boldsymbol{w}_* = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

Problem 2.2.2

Now consider σ as a parameter of the probabilistic model too, that is, the model is specified by both $\boldsymbol{w}_*$ and σ. Find the maximum likelihood estimation for $\boldsymbol{w}_*$ and σ.

Problem 2.2.2

Now consider σ as a parameter of the probabilistic model too, that is, the model is specified by both $\boldsymbol{w}_*$ and σ. Find the maximum likelihood estimation for $\boldsymbol{w}_*$ and σ.

Similar to 1., maximizing $P(\boldsymbol{w}, \sigma)$ is the same minimizing

$$-\ln(P(\boldsymbol{w}, \sigma)) = N \ln \sqrt{2\pi} + N \ln \sigma + \frac{1}{2\sigma^2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

Problem 2.2.2

Now consider σ as a parameter of the probabilistic model too, that is, the model is specified by both $\boldsymbol{w}_*$ and σ. Find the maximum likelihood estimation for $\boldsymbol{w}_*$ and σ.

Similar to 1., maximizing $P(\boldsymbol{w}, \sigma)$ is the same minimizing

$$-\ln(P(\boldsymbol{w}, \sigma)) = N \ln \sqrt{2\pi} + N \ln \sigma + \frac{1}{2\sigma^2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

We first fix σ and minimize over $\boldsymbol{w}$, which leads to the same MLE for $\boldsymbol{w}*$:

$$\boldsymbol{w}_* = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

Problem 2.2.2

Now consider σ as a parameter of the probabilistic model too, that is, the model is specified by both $\boldsymbol{w}_*$ and σ. Find the maximum likelihood estimation for $\boldsymbol{w}_*$ and σ.

Similar to 1., maximizing $P(\boldsymbol{w}, \sigma)$ is the same minimizing

$$-\ln(P(\boldsymbol{w}, \sigma)) = N \ln \sqrt{2\pi} + N \ln \sigma + \frac{1}{2\sigma^2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

We first fix σ and minimize over $\boldsymbol{w}$, which leads to the same MLE for $\boldsymbol{w}*$:

$$\boldsymbol{w}_* = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

Next we minimize $-\ln(P(\boldsymbol{w}_*, \sigma))$ as function of σ by setting the derivative w.r.t. σ to be 0 :

$$\frac{\partial - \ln(P(\boldsymbol{w}_*, \sigma))}{\partial \sigma} = \frac{N}{\sigma} - \frac{1}{\sigma^3} \|\boldsymbol{X}\boldsymbol{w}_* - \boldsymbol{y}\|_2^2 = 0$$

Similar to 1., maximizing $P(\boldsymbol{w}, \sigma)$ is the same minimizing

$$-\ln(P(\boldsymbol{w}, \sigma)) = N \ln \sqrt{2\pi} + N \ln \sigma + \frac{1}{2\sigma^2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

We first fix σ and minimize over w, which leads to the same MLE for $\boldsymbol{w}*$:

$$\boldsymbol{w}_* = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

Next we minimize $-\ln(P(\boldsymbol{w}_*, \sigma))$ as function of σ by setting the derivative w.r.t. σ to be 0 :

$$\frac{\partial - \ln(P(\boldsymbol{w}_*, \sigma))}{\partial \sigma} = \frac{N}{\sigma} - \frac{1}{\sigma^3} \|\boldsymbol{X}\boldsymbol{w}_* - \boldsymbol{y}\|_2^2 = 0$$

Similar to 1., maximizing $P(\boldsymbol{w}, \sigma)$ is the same minimizing

$$-\ln(P(\boldsymbol{w}, \sigma)) = N \ln \sqrt{2\pi} + N \ln \sigma + \frac{1}{2\sigma^2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

We first fix σ and minimize over w, which leads to the same MLE for $\boldsymbol{w}*$:

$$\boldsymbol{w}_* = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

Next we minimize $-\ln(P(\boldsymbol{w}_*, \sigma))$ as function of σ by setting the derivative w.r.t. σ to be 0 :

$$\frac{\partial - \ln(P(\boldsymbol{w}_*, \sigma))}{\partial \sigma} = \frac{N}{\sigma} - \frac{1}{\sigma^3} \|\boldsymbol{X}\boldsymbol{w}_* - \boldsymbol{y}\|_2^2 = 0$$

Solving for σ and plugging $\boldsymbol{w}_* = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ give the MLE estimate:

$$\sigma = \frac{1}{\sqrt{N}} \|\boldsymbol{X}\boldsymbol{w}_* - \boldsymbol{y}\|_2 = \frac{1}{\sqrt{N}} \left\|\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}\right\|_2$$

Problem 3

For a binary classification dataset of N data points, where every $\mathbf{x}_i$ has a corresponding label $y_i \in \{-1, 1\}$ and is normalized: $\|\mathbf{x}_i\| = \sqrt{\mathbf{x}_i^T \mathbf{x}_i} = 1, \forall i \in \{1, 2, \ldots, N\}$, the perceptron algorithm proceeds as below: Suppose there exists an unknown hyperplane $\mathbf{w}_{\text{opt}}$ such that $\|\mathbf{w}_{\text{opt}}\| = 1$ and $y_i \mathbf{w}_{opt}^T \mathbf{x}_i \geq \gamma, \forall i \in [N]$ for some positive value γ.

Problem 3

For a binary classification dataset of N data points, where every $\mathbf{x}_i$ has a corresponding label $y_i \in \{-1, 1\}$ and is normalized: $\|\mathbf{x}_i\| = \sqrt{\mathbf{x}_i^T \mathbf{x}_i} = 1, \forall i \in \{1, 2, \ldots, N\}$, the perceptron algorithm proceeds as below: Suppose there exists an unknown hyperplane $\mathbf{w}_{\text{opt}}$ such that $\|\mathbf{w}_{\text{opt}}\| = 1$ and $y_i \mathbf{w}_{opt}^T \mathbf{x}_i \geq \gamma, \forall i \in [N]$ for some positive value γ.

Algorithm 1 Perceptron

Initialize $\mathbf{w}_1 = \mathbf{0}$
for k = 1, 2, . . . , **do**
  Pick a data point $\mathbf{x}_i$ randomly
  Make a prediction $y = sign(\mathbf{w}_k^T \mathbf{x}_i)$ using current $\mathbf{w}_k$
  if $y \neq y_i$ **then**
    $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y_i \mathbf{x}_i$
  else
    $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k$

Problem 3

For a binary classification dataset of N data points, where every $\mathbf{x}_i$ has a corresponding label $y_i \in \{-1, 1\}$ and is normalized: $\|\mathbf{x}_i\| = \sqrt{\mathbf{x}_i^T \mathbf{x}_i} = 1, \forall i \in \{1, 2, \ldots, N\}$, the perceptron algorithm proceeds as below: Suppose there exists an unknown hyperplane $\mathbf{w}_{\text{opt}}$ such that $\|\mathbf{w}_{\text{opt}}\| = 1$ and $y_i \mathbf{w}_{opt}^T \mathbf{x}_i \geq \gamma, \forall i \in [N]$ for some positive value γ.

Algorithm 1 Perceptron

Initialize $\mathbf{w}_1 = \mathbf{0}$
for k = 1, 2, . . . , **do**
  Pick a data point $\mathbf{x}_i$ randomly
  Make a prediction $y = sign(\mathbf{w}_k^T \mathbf{x}_i)$ using current $\mathbf{w}_k$
  if $y \neq y_i$ **then**
    $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y_i \mathbf{x}_i$
  else
    $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k$

Problem 3.1

In iteration k such that $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, prove $\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma$

Problem 3.1

In iteration k such that $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, prove $\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma$

Problem 3.1

In iteration k such that $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, prove $\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma$

When $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have

$$\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$$

Problem 3.1

In iteration k such that $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, prove $\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma$

When $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have

$$\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$$

Taking inner product with $\mathbf{w}_{opt}$ on the both sides, we get

$$\mathbf{w}_{k+1}^T \mathbf{w}_{opt} = \mathbf{w}_k^T \mathbf{w}_{opt} + y_i \mathbf{x}_i^T \mathbf{w}_{opt}$$

Problem 3.1

In iteration k such that $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, prove $\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma$

When $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have

$$\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$$

Taking inner product with $\mathbf{w}_{opt}$ on the both sides, we get

$$\mathbf{w}_{k+1}^T \mathbf{w}_{opt} = \mathbf{w}_k^T \mathbf{w}_{opt} + y_i \mathbf{x}_i^T \mathbf{w}_{opt}$$

Since $y_i \mathbf{x}_i^T \mathbf{w}_{opt} \geq \gamma$ guaranteed by the problem, we conclude

$$\mathbf{w}_{k+1}^T \mathbf{w}_{opt} = \mathbf{w}_k^T \mathbf{w}_{opt} + y_i \mathbf{x}_i^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma$$

Problem 3.2

In iteration k such that $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, prove $\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + 1$

Problem 3.2

In iteration k such that $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, prove $\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + 1$

When $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have

$$\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$$

Problem 3.2

In iteration k such that $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, prove $\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + 1$

When $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have

$$\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$$

Therefore,

$$\|\mathbf{w}_{k+1}\|^2 = \mathbf{w}_{k+1}^T \mathbf{w}_{k+1} = \left(\mathbf{w}_k + y_i \mathbf{x}_i\right)^T \left(\mathbf{w}_k + y_i \mathbf{x}_i\right)$$

Problem 3.2

In iteration k such that $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, prove $\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + 1$

When $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have

$$\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$$

Therefore,

$$\begin{aligned} \|\mathbf{w}_{k+1}\|^2 = \mathbf{w}_{k+1}^T \mathbf{w}_{k+1} &= \left(\mathbf{w}_k + y_i \mathbf{x}_i\right)^T \left(\mathbf{w}_k + y_i \mathbf{x}_i\right) \\ &= \|\mathbf{w}_k\|^2 + 2 y_i \mathbf{w}_k^T \mathbf{x}_i + y_i^2 \mathbf{x}_i^T \mathbf{x}_i \end{aligned}$$

Problem 3.2

In iteration k such that $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, prove $\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + 1$

When $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have

$$\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$$

Therefore,

$$\begin{aligned} \|\mathbf{w}_{k+1}\|^2 = \mathbf{w}_{k+1}^T \mathbf{w}_{k+1} &= \left(\mathbf{w}_k + y_i \mathbf{x}_i\right)^T \left(\mathbf{w}_k + y_i \mathbf{x}_i\right) \\ &= \|\mathbf{w}_k\|^2 + 2 y_i \mathbf{w}_k^T \mathbf{x}_i + y_i^2 \mathbf{x}_i^T \mathbf{x}_i \end{aligned}$$

Since $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have $2 y_i \mathbf{w}_k^T \mathbf{x}_i \leq 0$.

Problem 3.2

In iteration k such that $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, prove $\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + 1$

When $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have

$$\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$$

Therefore,

$$\begin{aligned} \|\mathbf{w}_{k+1}\|^2 = \mathbf{w}_{k+1}^T \mathbf{w}_{k+1} &= \left(\mathbf{w}_k + y_i \mathbf{x}_i\right)^T \left(\mathbf{w}_k + y_i \mathbf{x}_i\right) \\ &= \|\mathbf{w}_k\|^2 + 2 y_i \mathbf{w}_k^T \mathbf{x}_i + y_i^2 \mathbf{x}_i^T \mathbf{x}_i \end{aligned}$$

Since $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have $2 y_i \mathbf{w}_k^T \mathbf{x}_i \leq 0$. Moreover, by the assumption on input, we have $y_i^2 \mathbf{x}_i^T \mathbf{x}_i = y_i^2 \|\mathbf{x}_i\|^2 = 1$, so

Problem 3.2

In iteration k such that $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, prove $\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + 1$

When $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have

$$\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$$

Therefore,

$$\begin{aligned} \|\mathbf{w}_{k+1}\|^2 = \mathbf{w}_{k+1}^T \mathbf{w}_{k+1} &= \left(\mathbf{w}_k + y_i \mathbf{x}_i\right)^T \left(\mathbf{w}_k + y_i \mathbf{x}_i\right) \\ &= \|\mathbf{w}_k\|^2 + 2 y_i \mathbf{w}_k^T \mathbf{x}_i + y_i^2 \mathbf{x}_i^T \mathbf{x}_i \end{aligned}$$

Since $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have $2 y_i \mathbf{w}_k^T \mathbf{x}_i \leq 0$. Moreover, by the assumption on input, we have $y_i^2 \mathbf{x}_i^T \mathbf{x}_i = y_i^2 \|\mathbf{x}_i\|^2 = 1$, so

$$\|\mathbf{w}_{k+1}\|^2 = \|\mathbf{w}_k\|^2 + 2 y_i \mathbf{w}_k^T \mathbf{x}_i + y_i^2 \mathbf{x}_i^T \mathbf{x}_i$$

Problem 3.2

In iteration k such that $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, prove $\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + 1$

When $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have

$$\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$$

Therefore,

$$\begin{aligned} \|\mathbf{w}_{k+1}\|^2 = \mathbf{w}_{k+1}^T \mathbf{w}_{k+1} &= \left(\mathbf{w}_k + y_i \mathbf{x}_i\right)^T \left(\mathbf{w}_k + y_i \mathbf{x}_i\right) \\ &= \|\mathbf{w}_k\|^2 + 2 y_i \mathbf{w}_k^T \mathbf{x}_i + y_i^2 \mathbf{x}_i^T \mathbf{x}_i \end{aligned}$$

Since $y_i \neq \text{sign}\left(\mathbf{w}_k^T \mathbf{x}_i\right)$, we have $2 y_i \mathbf{w}_k^T \mathbf{x}_i \leq 0$. Moreover, by the assumption on input, we have $y_i^2 \mathbf{x}_i^T \mathbf{x}_i = y_i^2 \|\mathbf{x}_i\|^2 = 1$, so

$$\begin{aligned} \|\mathbf{w}_{k+1}\|^2 &= \|\mathbf{w}_k\|^2 + 2 y_i \mathbf{w}_k^T \mathbf{x}_i + y_i^2 \mathbf{x}_i^T \mathbf{x}_i \\ &\leq \|\mathbf{w}_k\|^2 + y_i^2 \mathbf{x}_i^T \mathbf{x}_i = \|\mathbf{w}_k\|^2 + 1 \end{aligned}$$

Problem 3.3

For any iteration $k+1$, with M being the total number of mistakes the algorithm has made so far for the first k iterations, we have

$$M\gamma \leq \|\mathbf{w}_{k+1}\| \leq \sqrt{M}$$

Problem 3.3

For any iteration $k+1$, with M being the total number of mistakes the algorithm has made so far for the first k iterations, we have

$$M\gamma \leq \|\mathbf{w}_{k+1}\| \leq \sqrt{M}$$

By repeatedly applying results from Problem 3.1 we have

$$\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma$$

Problem 3.3

For any iteration $k+1$, with M being the total number of mistakes the algorithm has made so far for the first k iterations, we have

$$M\gamma \leq \|\mathbf{w}_{k+1}\| \leq \sqrt{M}$$

By repeatedly applying results from Problem 3.1 we have

$$\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma \geq \cdots \geq \mathbf{w}_1^T \mathbf{w}_{opt} + M\gamma$$

Problem 3.3

For any iteration $k+1$, with M being the total number of mistakes the algorithm has made so far for the first k iterations, we have

$$M\gamma \leq \|\mathbf{w}_{k+1}\| \leq \sqrt{M}$$

By repeatedly applying results from Problem 3.1 we have

$$\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma \geq \cdots \geq \mathbf{w}_1^T \mathbf{w}_{opt} + M\gamma = M\gamma$$

Problem 3.3

For any iteration $k+1$, with M being the total number of mistakes the algorithm has made so far for the first k iterations, we have

$$M\gamma \leq \|\mathbf{w}_{k+1}\| \leq \sqrt{M}$$

By repeatedly applying results from Problem 3.1 we have

$$\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma \geq \cdots \geq \mathbf{w}_1^T \mathbf{w}_{opt} + M\gamma = M\gamma$$

Using $\mathbf{a}^T\mathbf{b} \leq \|\mathbf{a}\|\|\mathbf{b}\|$ and $\|\mathbf{w}_{opt}\| = 1$, we have

$$M\gamma \leq \mathbf{w}_{k+1}^T \mathbf{w}_{opt} \leq \|\mathbf{w}_{k+1}\| \, \|\mathbf{w}_{opt}\| = \|\mathbf{w}_{k+1}\|$$

Problem 3.3

For any iteration $k+1$, with M being the total number of mistakes the algorithm has made so far for the first k iterations, we have

$$M\gamma \leq \|\mathbf{w}_{k+1}\| \leq \sqrt{M}$$

By repeatedly applying results from Problem 3.1 we have

$$\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma \geq \cdots \geq \mathbf{w}_1^T \mathbf{w}_{opt} + M\gamma = M\gamma$$

Using $\mathbf{a}^T\mathbf{b} \leq \|\mathbf{a}\|\|\mathbf{b}\|$ and $\|\mathbf{w}_{opt}\| = 1$, we have

$$M\gamma \leq \mathbf{w}_{k+1}^T \mathbf{w}_{opt} \leq \|\mathbf{w}_{k+1}\| \, \|\mathbf{w}_{opt}\| = \|\mathbf{w}_{k+1}\|$$

Similarly, using results of Problem 3.2 repeatedly, we have

$$\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + 1$$

Problem 3.3

For any iteration $k+1$, with M being the total number of mistakes the algorithm has made so far for the first k iterations, we have

$$M\gamma \leq \|\mathbf{w}_{k+1}\| \leq \sqrt{M}$$

By repeatedly applying results from Problem 3.1 we have

$$\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma \geq \cdots \geq \mathbf{w}_1^T \mathbf{w}_{opt} + M\gamma = M\gamma$$

Using $\mathbf{a}^T\mathbf{b} \leq \|\mathbf{a}\|\|\mathbf{b}\|$ and $\|\mathbf{w}_{opt}\| = 1$, we have

$$M\gamma \leq \mathbf{w}_{k+1}^T \mathbf{w}_{opt} \leq \|\mathbf{w}_{k+1}\| \, \|\mathbf{w}_{opt}\| = \|\mathbf{w}_{k+1}\|$$

Similarly, using results of Problem 3.2 repeatedly, we have

$$\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + 1 \leq \cdots \leq \|\mathbf{w}_1\|^2 + M$$

Problem 3.3

For any iteration $k+1$, with M being the total number of mistakes the algorithm has made so far for the first k iterations, we have

$$M\gamma \leq \|\mathbf{w}_{k+1}\| \leq \sqrt{M}$$

By repeatedly applying results from Problem 3.1 we have

$$\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma \geq \cdots \geq \mathbf{w}_1^T \mathbf{w}_{opt} + M\gamma = M\gamma$$

Using $\mathbf{a}^T\mathbf{b} \leq \|\mathbf{a}\|\|\mathbf{b}\|$ and $\|\mathbf{w}_{opt}\| = 1$, we have

$$M\gamma \leq \mathbf{w}_{k+1}^T \mathbf{w}_{opt} \leq \|\mathbf{w}_{k+1}\| \, \|\mathbf{w}_{opt}\| = \|\mathbf{w}_{k+1}\|$$

Similarly, using results of Problem 3.2 repeatedly, we have

$$\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + 1 \leq \cdots \leq \|\mathbf{w}_1\|^2 + M = M$$

Problem 3.4

Using result of Problem 3.3, conclude $M \leq \gamma^{-2}$

Problem 3.4

Using result of Problem 3.3, conclude $M \leq \gamma^{-2}$

From Problem 3.3, we have

$$M\gamma \leq \sqrt{M}$$

Problem 3.4

Using result of Problem 3.3, conclude $M \leq \gamma^{-2}$

From Problem 3.3, we have

$$M\gamma \leq \sqrt{M}$$

This implies

$$\sqrt{M} \leq \frac{1}{\gamma}$$

Problem 3.4

Using result of Problem 3.3, conclude $M \leq \gamma^{-2}$

From Problem 3.3, we have

$$M\gamma \leq \sqrt{M}$$

This implies

$$\sqrt{M} \leq \frac{1}{\gamma} \;\Rightarrow\; M \leq \frac{1}{\gamma^2}$$