58 Example (Line Segment)
The set of all the points, $\mathbf{x}$, on the line segment connecting point $\mathbf{x}_1$ to point $\mathbf{x}_2$ can be specified as

$$\mathbf{x} = \alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \text{ for } 0 \le \alpha \le 1.$$

---

59 Example (Convex vs Nonconvex Functions)
Draw some examples of single variable functions that are convex and that are nonconvex.

---

60 Example (Minimum Value of Convex Functions)
(a) Can a strictly convex function have more than one minimum value?

(b) Can a strictly convex function have inflection points?

---

61 Definition (L^2 Regularization)
L^2 **regularization** is a procedure for preventing over-fitting. The term $\alpha \mathbf{w}^\top \mathbf{w}$ is added to the loss function, $\ell(\mathbf{w})$. This term prevents the weights, $\mathbf{w}$, from becoming too large.

$$\begin{aligned} \ell(\mathbf{w}) &= \text{MSE}(\mathbf{w}) + \alpha \mathbf{w}^\top \mathbf{w} \\ &= \frac{1}{n}(\hat{\mathbf{y}} - \mathbf{y})^\top (\hat{\mathbf{y}} - \mathbf{y}) + \alpha \mathbf{w}^\top \mathbf{w} \end{aligned}$$

If the value of the hyper-parameter α is too large, we will have under-fitting, resulting in a large training error. If the value is too small, we may or may not have over-fitting. Over-fitting is characterized by a small training error, but a large test error. The optimal value of α is chosen to minimizes the test error.

---

62 Example (L^2 Regularization of Linear Regression)
Recall that for linear regression $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$. To apply L^2 regularization, we must minimize

$$\ell(\mathbf{w}) = \frac{1}{n}(\hat{\mathbf{y}} - \mathbf{y})^\top (\hat{\mathbf{y}} - \mathbf{y}) + \alpha \mathbf{w}^\top \mathbf{w}.$$

$$\begin{aligned} \frac{d\ell}{d\mathbf{w}} &= \left(\frac{2}{n}(\hat{\mathbf{y}} - \mathbf{y})^\top \frac{d\hat{\mathbf{y}}}{d\mathbf{w}} + 2\alpha \mathbf{w}^\top \right)^\top \\ &= \left(\frac{2}{n}(\mathbf{X}\mathbf{w} - \mathbf{y})^\top \mathbf{X} + 2\alpha \mathbf{w}^\top \right)^\top \\ &= \frac{2}{n}\mathbf{X}^\top (\mathbf{X}\mathbf{w} - \mathbf{y}) + 2\alpha \mathbf{w} \end{aligned}$$

Setting $\frac{d\ell}{d\mathbf{w}} = 0$ implies

$$\begin{aligned} \mathbf{X}^\top \mathbf{X}\mathbf{w} + n\alpha \mathbf{w} &= \mathbf{X}^\top \mathbf{y} \\ (\mathbf{X}^\top \mathbf{X} + \alpha_o \mathbf{I})\mathbf{w} &= \mathbf{X}^\top \mathbf{y} \end{aligned}$$

where $\alpha_o = n\alpha$. Therefore, we must repeatedly solve

$$\mathbf{A}(\alpha_o)\mathbf{w} = \mathbf{b}$$

where $\mathbf{A}(\alpha_o) = \mathbf{X}^\top \mathbf{X} + \alpha_o \mathbf{I}$ and $\mathbf{b} = \mathbf{X}^\top \mathbf{y}$ in order to determine α_o that minimize the test error.

---

4 Logistic Regression

- Prediction problems with continuous target variables are called *regression* problems.
- Prediction problems with discrete target variables are called *classification* problems.

Logistic regression is used for classification problems.

63 Example (Regression vs Classification)
Regression or Classification?

(a) Use a person's age to predict their resting heart rate.

(b) Use a person's handwriting to predict if they are male or female.

(c) Spam Filter