Leave-one-out Approximation of Integrated Squared Error:

1. **Approximating means:**
   Consider independent and identically distributed r.v.s \( X_1, X_2, \ldots, X_n \) such that \( X_i \in \{-1, 1\} \) and \( p(X_i = 1) = p \neq p(0) = 1 - p \).
   Then, we define the expected value of \( X_i \) as
   \[
   \mathbb{E}[X_i] = 1 \cdot p(X_i = 1) + (-1) \cdot p(X_i = -1) = p - (1 - p) = 2p - 1.
   \]
   When \( n \) is large, \( p \approx \frac{\sum_{i=1}^{n} X_i}{n} \) and \( 1 - p \approx \frac{\sum_{i=1}^{n} X_i}{n} \).
   The proportion of \( X_i \)'s equal to 1.

   Hence, when \( n \) is large,
   \[
   \mathbb{E}[X_i] \approx 1 \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i = 1) \right) + (-1) \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i = -1) \right)
   = \frac{1}{n} \left( \text{no. of } X_i = 1 \right) + (-1) \left( \text{no. of } X_i = -1 \right).
   \]
   \[
   \approx \frac{1}{n} \sum_{i=1}^{n} X_i.
   \]
   This idea can be generalized. Suppose \( X_1, \ldots, X_n \in \mathbb{B} \) with probability density function \( f(x) \). Then,
   \[
   \mathbb{E}[X_i] = \int_{-\infty}^{\infty} x f(x) \, dx \approx \frac{1}{n} \sum_{i=1}^{n} X_i,
   \]
   \[
   \mathbb{E}[g(X_i)] = \int_{-\infty}^{\infty} g(x) f(x) \, dx \approx \frac{1}{n} \sum_{i=1}^{n} g(X_i),
   \]
   when \( n \) is large, for any function \( g : \mathbb{B} \rightarrow \mathbb{B} \).

2. **Decomposing Integrated Squared Error:**
   \[
   L(w) = \int_{-\infty}^{\infty} (f_m(x) - f(x))^2 \, dx = \int_{-\infty}^{\infty} f_m(x)^2 \, dx - 2 \int_{-\infty}^{\infty} f_m(x) f(x) \, dx + \int_{-\infty}^{\infty} f(x)^2 \, dx
   \]
   approximate this with \( J(w) \) + constant

3. **Leave-one-out Cross-Validation [Rudemo '82]:**
   We would usually approximate \( \int_{-\infty}^{\infty} f_m(x) f(x) \, dx \) as
   \[
   \int_{-\infty}^{\infty} f_m(x) f(x) \, dx \approx \frac{1}{m} \sum_{i=1}^{m} f_m(X_i)
   \]
   where \( X_1, \ldots, X_m \) are samples from \( f(x) \). However, \( f_m \) depends on \( X_i \), which leads to greater "bias." So, we use the alternative approximation
   \[
   \int_{-\infty}^{\infty} f_m(x) f(x) \, dx \approx \frac{1}{m} \sum_{i=1}^{m} h_m(X_i) \leftarrow \text{lower "bias"} \text{ leave-one-out!}
   \]
   where \( h_m \) is the histogram made of samples \( X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m \), and \( h_m \approx f_m \) for large \( m \).
Approximation $J(w)$:

Recall that:

$f_m(x)$ - frequency density

\[
\int_{-\infty}^{\infty} f_m(x)^2 \, dx - 2 \int_{-\infty}^{\infty} f_m(x) f(x) \, dx \approx \sum_{k=1}^{n} \left( \frac{\hat{p}_k}{w(m)} \right)^2 - 2 \frac{1}{w} \sum_{k=1}^{n} \sum_{m=1}^{m} \hat{p}_m \frac{\hat{p}_m}{w(m-1)}
\]

\[
= \frac{1}{w} \sum_{k=1}^{n} \hat{p}_k^2 - 2 \frac{1}{w(m-1)} \sum_{k=1}^{n} \hat{p}_k + \frac{2}{w(m-1)} \sum_{k=1}^{n} \hat{p}_k
\]

\[
= \frac{2}{w(m-1)} - \frac{m+1}{w(m-1)} \sum_{k=1}^{n} \hat{p}_k
\]

Hence, we define $J(w) = \frac{2}{w(m-1)} - \frac{m+1}{w(m-1)} \sum_{k=1}^{n} \hat{p}_k$ and we have:

\[
L(w) \approx J(w) + \int_{-\infty}^{\infty} f(x)^2 \, dx.
\]

constant
1. Properties of mean and variance:

i) For random variables $X$ and $Y$, and constants $a, b, c \in \mathbb{R}$, we have
   \[
   \]

ii) For independent random variables $X$ and $Y$, and constants $a, b, c \in \mathbb{R}$, we have
   \[
   \text{var}(aX + bY + c) = a^2 \text{var}(X) + b^2 \text{var}(Y).
   \]

2. Unbiased estimators of mean and variance:

Let $X_1, \ldots, X_n$ be i.i.d. random variables with expected value $E[X] = \mu$ and variance $\text{var}(X) = \sigma^2$.

Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ be the empirical estimate of $\mu$. Then,
\[
E[\bar{X}] = \frac{1}{n} \sum_{i=1}^{n} E[X] = \mu. \quad \text{unbiased!}
\]

Let $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ be the empirical estimate of $\sigma^2$. Observe that:
\[
E\left[ \sum_{i=1}^{n} (X_i - \bar{X})^2 \right] = E\left[ \sum_{i=1}^{n} X_i^2 - 2\bar{X} \sum_{i=1}^{n} X_i + n\bar{X}^2 \right]
\]
\[
= E\left[ \sum_{i=1}^{n} X_i^2 - 2\bar{X} \sum_{i=1}^{n} X_i + n\bar{X}^2 \right]
\]
\[
= \sum_{i=1}^{n} E[X_i^2] - 2n E[\bar{X}]^2 + n E[\bar{X}^2]
\]
\[
= n E[X^2] - 2n \mu^2 + n \mu^2
\]
\[
= n E[X^2] - n \mu^2
\]

Hence, $E[s^2] = \sigma^2$. \quad \text{unbiased!}

3. Tail bounds:

i) Markov's inequality: For any random variable $X \geq 0$ and any constant $a > 0$,
   \[
P(X \geq a) \leq \frac{E[X]}{a}.
   \]
   \[\text{PF: (Continuous case)} E[X] = \int_{a}^{\infty} f_X(x) dx = \int_{a}^{\infty} \frac{1}{x} dx = \frac{1}{a} \quad \text{pdf of } X \sim \frac{1}{x}
   \]
   \[
   \Rightarrow E[X] \geq a P(X \geq a).
   \]

ii) Chebyshev's inequality: For any random variable $X$ and any constant $a > 0$,
   \[
P(\left| X - E[X] \right| \geq a) \leq \frac{\text{var}(X)}{a^2}.
   \]
   \[\text{PF: Consider the random variable } Y = (X - E[X])^2 \geq 0 \text{ with } E[Y] = \text{var}(X). \text{ By Markov's inequality,}
   \]
   \[
P(Y \geq a^2) = E[Y] P(Y \geq a^2) \leq \frac{E[Y]}{a^2} = \frac{\text{var}(X)}{a^2}.
   \]

Since the event $\{X - E[X] \geq a\}$ is equal to the event $\{Y \geq a^2\}$, we have
\[
P(\left| X - E[X] \right| \geq a) \leq \frac{\text{var}(X)}{a^2}.
   \]
4) Weak Law of Large Numbers (WLLN):

For i.i.d. random variables $X_1, \ldots, X_n$ with mean $\mathbb{E}[X] = \mu$ and variance $\text{var}(X) = \sigma^2$, we have

$$\lim_{n \to \infty} P\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \varepsilon \right) = 0 \quad \text{for all } \varepsilon > 0.$$

(Insulation: $\bar{X}$ is close to $\mathbb{E}[X]$ when $n$ is large with high probability.)

Proof: For any $\varepsilon > 0$, by Chebyshev's inequality,

$$P\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \varepsilon \right) \leq \frac{\text{var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)}{\varepsilon^2} = \frac{\text{var}(X)}{n\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Hence, $\lim_{n \to \infty} P\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \varepsilon \right) = 0.$

5) Central Limit Theorem (CLT):

Let $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ be the standard normal CDF.

Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean $\mathbb{E}[X] = \mu$ and variance $\text{var}(X) = \sigma^2$.

Then, for all $x \in \mathbb{R}$,

$$\lim_{n \to \infty} P\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \leq x \right) = \Phi(x).$$

(Insulation: CDF of this mean $\mu$ and variance $\sigma^2$ random variable looks like normal CDF when $n$ is large.)
Some Probability Results on Induced Distributions

1. Change-of-Variables Formulas (linear case)

Let \( X \) be a continuous random variable with PDF \( f_X \) and CDF \( F_X \). For any constants \( a \neq 0 \) and \( b \in \mathbb{R} \), define the continuous random variable \( Y = aX + b \). Suppose \( Y \) has CDF \( F_Y \) and PDF \( f_Y \). Can we compute \( f_Y \) in terms of \( f_X \)?

\[
F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(aX + b \leq y) = \begin{cases} 
F_X\left(\frac{y-b}{a}\right), & a > 0 \\
1 - F_X\left(\frac{y-b}{a}\right), & a < 0
\end{cases}
\]

\[
f_Y(y) = \begin{cases} 
\frac{d}{dy} F_X\left(\frac{y-b}{a}\right), & a > 0 \\
-\frac{d}{dy} F_X\left(\frac{y-b}{a}\right), & a < 0
\end{cases}
\]

Chain rule

\[
\therefore \text{For all } y \in \mathbb{R}, \quad f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).
\]

2. Convolution: (discrete case)

Let \( X \) and \( Y \) be independent discrete random variables taking values in \( \mathbb{Z} \) with PMFs \( f_X \) and \( f_Y \), respectively. Define the discrete random variable \( Z = X + Y \), which has PMF \( f_Z \) and also takes values in \( \mathbb{Z} \). Can we compute \( f_Z \) in terms of \( f_X \) and \( f_Y \)?

\[
f_Z(k) = \mathbb{P}(Z = k) = \mathbb{P}(X + Y = k) = \sum_{j=-\infty}^{\infty} \mathbb{P}(X = j, Y = k-j) = \sum_{j=-\infty}^{\infty} f_X(j) f_Y(k-j)
\]

Define the convolution of \( f_X \) and \( f_Y \) as \( f_X \ast f_Y(k) = \sum_{j=-\infty}^{\infty} f_X(j) f_Y(k-j) \) for all \( k \in \mathbb{Z} \).

\[
\therefore f_Z = f_X \ast f_Y.
\]
1. **Gradient:**

For a differentiable function \( f : \mathbb{R}^d \rightarrow \mathbb{R} \), \( f(x_1, \ldots, x_d) \), we define its gradient as the vector field \( \nabla f : \mathbb{R}^d \rightarrow \mathbb{R}^d \) given by:

\[
\nabla f(x) = \left[ \begin{array}{c}
\frac{\partial f}{\partial x_1}(x) \\
\frac{\partial f}{\partial x_2}(x) \\
\vdots \\
\frac{\partial f}{\partial x_d}(x)
\end{array} \right]
\]

2. **Gradient of Quadratic Form:**

Define the quadratic form \( f(x) = x^T Ax \) for any \( x \in \mathbb{R}^d \), given a fixed matrix \( A \in \mathbb{R}^{d \times d} \).

**Prop:** \( \forall x \in \mathbb{R}^d, \ \nabla f(x) = (A + A^T)x \).

**Pf:** Observe that \( f(x) = \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix} \begin{bmatrix} A_{11} & \cdots & A_{1d} \\
\vdots & \ddots & \vdots \\
A_{d1} & \cdots & A_{dd} \end{bmatrix} x = \sum_{i=1}^d x_i \sum_{j=1}^d A_{ij} x_j = \sum_{i=1}^d \sum_{j=1}^d x_i x_j A_{ij} \)

For any \( k \in \{1, \ldots, d\} \),

\[
\frac{\partial f}{\partial x_k}(x) = \frac{\partial}{\partial x_k} \left( \sum_{i=1}^d \sum_{j=1}^d x_i x_j A_{ij} \right) = \sum_{i=1}^d x_i A_{ik} + \sum_{j=1}^d x_j A_{jk}
\]

\[
= \begin{bmatrix} A^T \end{bmatrix}_{k1} + \begin{bmatrix} A \end{bmatrix}_{k1}
\]

\[
= \begin{bmatrix} (A + A^T) x \end{bmatrix}_k
\]

3. **Gradient of Linear Form:**

Define the linear form \( f(x) = b^T x \) for any \( x \in \mathbb{R}^d \), given a fixed vector \( b \in \mathbb{R}^d \).

**Prop:** \( \forall x \in \mathbb{R}^d, \ \nabla f(x) = b \).

**Pf:** For any \( k \in \{1, \ldots, d\} \),

\[
\frac{\partial f}{\partial x_k}(x) = \frac{\partial}{\partial x_k} \left( \sum_{i=1}^d b_i x_i \right) = b_k.
\]
Normal Equation for Regression: For any given target values \( y \in \mathbb{R}^N \) & feature matrix \( X \in \mathbb{R}^{N \times (M+1)} \), define the mean-squared error (MSE):

\[
E(\beta) = \frac{1}{N} \| y - X\beta \|^2.
\]

To minimize \( E(\beta) \) over \( \beta \), we use the stationarity condition:

\[
\nabla E(\beta) = 0.
\]

Hence, we have:

\[
0 = \nabla \left( \frac{1}{N} \| y - X\beta \|^2 \right) = \nabla \left( \frac{1}{N} (y - X\beta)^T (y - X\beta) \right)
= \nabla \left( \frac{1}{N} (y^T y - 2(y^T X\beta) + \beta^T X^T X \beta) \right)
= -\frac{2}{N} X^T y + \frac{1}{N} (X^T X + \lambda I) \beta
\]

\[
\iff \quad X^T X \beta = X^T y, \quad \text{Normal Equation}
\]
Mean & Variance of Regression Coefficient

1. Setup:
Assume training samples \((x_1, y_1), \ldots, (x_N, y_N)\) \(y_n \sim N(x_n, \beta + \varepsilon_n)\) are generated according to the simple linear model: 
\[ y_n = \beta + \varepsilon_n \] 
where \(x_1, \ldots, x_n\) are fixed values, \(\beta\) is the true fixed parameters, and \(\varepsilon_1, \ldots, \varepsilon_n\) are i.i.d. noise random variables with mean 0 and variance \(\sigma^2\).

Simple linear regression produces the estimate 
\[ \hat{\beta} = \frac{\sum_{n=1}^{N} x_n y_n - N \bar{x} \bar{y}}{\sum_{n=1}^{N} x_n^2 - N \bar{x}^2} \]

for \(\alpha\), where 
\[ \bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n \] 
and 
\[ \bar{y} = \frac{1}{N} \sum_{n=1}^{N} y_n \].

2. Mean of \(\hat{\beta}\):
Note that 
\[ E[y_n] = \beta + E[\varepsilon_n] = \beta + 0 \] 
and 
\[ \text{var}(y_n) = \sigma^2 \] 
for \(n = 1, \ldots, N\).

Hence,
\[ E[\hat{\beta}] = E \left[ \frac{\sum_{n=1}^{N} (x_n - \bar{x}) y_n}{\sum_{n=1}^{N} x_n^2 - N \bar{x}^2} \right] = \frac{\sum_{n=1}^{N} (x_n - \bar{x})(\beta + \varepsilon_n)}{\sum_{n=1}^{N} x_n^2 - N \bar{x}^2} = \beta \left( \frac{\sum_{n=1}^{N} x_n^2 - N \bar{x}^2}{\sum_{n=1}^{N} x_n^2 - N \bar{x}^2} \right) + \frac{\sigma^2}{\sum_{n=1}^{N} x_n^2 - N \bar{x}^2} \]
\[ \Rightarrow E[\hat{\beta}] = \beta \]
\(\hat{\beta}\) is an unbiased estimator of \(\beta\).

3. Variance of \(\hat{\beta}\):
Note that 
\[ \sum_{n=1}^{N} (x_n - \bar{x})^2 = \sum_{n=1}^{N} x_n^2 + N \bar{x}^2 - 2 \bar{x} \sum_{n=1}^{N} x_n = \sum_{n=1}^{N} x_n^2 - N \bar{x}^2 \]

Hence,
\[ \text{var}(\hat{\beta}) = \text{var} \left( \frac{\sum_{n=1}^{N} (x_n - \bar{x}) y_n}{\sum_{n=1}^{N} (x_n - \bar{x})^2} \right) = \frac{\sigma^2}{\sum_{n=1}^{N} (x_n - \bar{x})^2} \]
\[ \Rightarrow \text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{n=1}^{N} (x_n - \bar{x})^2} \]

4. Standard Error when Testing whether \(\alpha = 0\): (\(\sigma^2\) unknown)
- Use 
\[ s^2 = \frac{1}{N-2} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 \] 

\[ \hat{y}_n = \hat{\beta} x_n + b \]

- \(\hat{y}_n\) is an estimate of \(E_n\) and 
\[ s^2 = \frac{1}{N-2} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 \approx \sigma^2 \]

- Use 
\[ \text{var}(\hat{\beta}) = \frac{s^2}{\sum_{n=1}^{N} (x_n - \bar{x})^2} \] 

to estimate \(\text{var}(\hat{\beta})\)

\[ \text{standard error of } \hat{\beta} \]