UNIVERSITY OF ALBERTA
CMPUT 267 Winter 2025

Midterm Exam 2

Do Not Distribute

Duration: 60 minutes

Last Name: _____

First Name: _____

Carefully read all of the instructions and questions. Good luck!

---

1. **Do not turn this page** until instructed to begin.
2. This is exam version **0**. Please mark **0 in the special code section in coloum J** of your scantron.
3. Verify that your exam package includes 13 pages (last two are blanck), along with a formula sheet at the end.
4. **Only the scantron will be marked**. All of your answers must be clearly marked on the scantron.
5. Use **pencil only** to fill out the scantron (preferably an HB or #2 pencil).
6. **Erase mistakes completely** on the scantron to avoid misreading by the scanner.
7. **Mark answers firmly and darkly**, filling in the bubbles completely.
8. This exam consists of **15 questions**. Each question is worth **1 mark**. The exam is worth a total of **15 marks**.
9. Some questions may have **multiple correct answers**. To receive **full marks**, you must select **all correct answers**. If you select only **some** of the correct answers, you will receive **partial marks**. Selecting an incorrect option will cancel out a correct one. For example, if you select two answers—one correct and one incorrect—you will receive zero points for that question. If the number of incorrect answers exceeds the correct ones, your score for that question will be zero. **No negative marks** will be given.

Question 1. [1 MARK]

Consider an optimization problem where the goal is to maximize a function $f(w)$ with respect to $w \in \mathbb{R}^d$:

$$\max_{w \in \mathbb{R}^d} \; f(w).$$

Which of the following statements is true?

A. $\max_{w \in \mathbb{R}^d} \; f(w)$ results in the same w^* as $\min_{w \in \mathbb{R}^d} \; [-f(w)]$.

B. $\max_{w \in \mathbb{R}^d} \; f(w)$ results in the same w^* as $\min_{w \in \mathbb{R}^d} \; f(w)$.

C. $\min_{w \in \mathbb{R}^d} \; f(w)$ results in the same optimal value as $\max_{w \in \mathbb{R}^d} \; [-f(w)]$.

D. $\min_{w \in \mathbb{R}^d} \; f(w)$ results in the same optimal value as $-\max_{w \in \mathbb{R}^d} \; [-f(w)]$.

Question 2. [1 MARK]

Consider a polynomial regression model that uses a feature map ϕ_p of degree p. The predictor is given by $f(x) = \phi_p(x)^T w$. Suppose you are given a dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ and the empirical loss is defined as

$$\hat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} \Big(\phi_p(x_i)^T w - y_i \Big)^2.$$

In this question, we are interested in finding $\hat{w} = \arg\min_{w \in \mathbb{R}^{\bar{p}}} \hat{L}(w)$ using first-order gradient descent. Which of the following statements is true?

A. The first-order gradient update rule with a fixed step size is

$$w^{(t+1)} = w^{(t)} - \frac{2}{n} \eta^{(t)} \left[\sum_{i=1}^{n} \Big(\phi_p(x_i)^T w - y_i \Big) \phi_p(x_i) \right].$$

B. If you run first-order gradient descent for infinitly many epochs T, you are guaranteed to converge to the minimizer $w^{(T)} = \hat{w}$, where $\hat{w} = \arg\min_{w \in \mathbb{R}^{\bar{p}}} \hat{L}(w)$.

C. $\hat{L}(w)$ is convex.

D. $\hat{L}(w)$ is not convex.

Question 3. [1 MARK]

Let everything be defined as in the previous question. Now we consider regularized polynomial regression. The regularized empirical loss is defined as

$$\hat{L}_\lambda(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\phi_p(x_i)^T w - y_i\right)^2 + \frac{\lambda}{n}\|w\|^2,$$

where $\lambda > 0$ is the regularization parameter. Which of the following statements is true?

A. The optimization problem $\min_{w\in\mathbb{R}^p} \hat{L}_\lambda(w)$ does not have a closed form solution, since $\hat{L}_\lambda$ is not convex.

B. The regularization term $\frac{\lambda}{n}\|w\|^2$ penalizes large weights, which helps to prevent overfitting.

C. The gradient descent update rule for the regularized loss is

$$w^{(t+1)} = w^{(t)} - \eta^{(t)}\left[\frac{2}{n}\sum_{i=1}^{n}\left(\phi_p(x_i)^T\phi_p(w^{(t)}) - y_i\right)\phi_p(x_i) + \frac{2\lambda}{n}\phi_p(w^{(t)})\right].$$

D. The gradient descent update rule for the regularized loss is

$$w^{(t+1)} = w^{(t)} - \eta^{(t)}\left[\frac{2}{n}\sum_{i=1}^{n}\left(\phi_p(x_i)^T w^{(t)} - y_i\right)\phi_p(x_i) + \frac{2\lambda}{n}w^{(t)}\right].$$

Question 4. [1 MARK]

Consider a convex optimization problem solved by gradient descent $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta^{(t)}\nabla g\left(\mathbf{w}^{(t)}\right)$. Which of the following statements is true?

A. Exponential decaying step sizes, defined by $\eta^{(t)} = \eta_0 \exp(-\lambda t)$, reduce the learning rate more rapidly than inverse decaying step sizes, defined by $\eta^{(t)} = \frac{\eta_0}{1+t}$.

B. The normalized gradient step size, that is defined by $\eta^{(t)} = \frac{\eta}{\epsilon + ||\nabla g(\mathbf{w}^{(t)})||}$, accelerates convergence by removing the influence of the gradient's magnitude from the update direction.

C. The normalized gradient step size, that is defined by $\eta^{(t)} = \frac{\eta}{\epsilon + ||\nabla g(\mathbf{w}^{(t)})||}$, avoids overshooting the minimum by inversly proportionally adapting the step size with respect to the gradient magnitude (gradient magnitude large $\rightarrow$ small step size and vice versa).

D. A constant step size, $\eta^{(t)} = \eta_0$, is generally preferred over any decaying schedule because it maintains consistent update magnitudes throughout the iterations.

Question 5. [1 MARK]

The Poisson distribution is a discrete probability distribution that models the number of events occurring in a fixed interval. Its probability mass function (pmf) is given by

$$p(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

for $k = 0, 1, 2, \ldots$, where $a! = a \times (a-1) \times \cdots \times 1$ is the factorial function.
Now suppose we have data $D = (X_1, X_2, X_3) = (3, 2, 4)$, where each X_i is independently drawn from a Poisson distribution with parameter λ. We want to estimate λ using maximum likelihood estimation (MLE).
Which of the following statements is true?

A. The likelihood function is
$$\frac{\lambda^9 e^{-3\lambda}}{(3!)^2 \cdot 4!}.$$

B. The likelihood function is
$$\frac{\lambda^9 e^{-3\lambda}}{3! \cdot 2! \cdot 4!}.$$

C. The likelihood function is
$$\frac{\lambda^8 e^{-3\lambda}}{3! \cdot 2! \cdot 4!}.$$

D. The likelihood function is
$$\frac{\lambda^9 e^{-2\lambda}}{3! \cdot 2! \cdot 4!}.$$

Question 6. [1 MARK]

Let everything be defined as in the previous question regarding the Poisson distribution. Recall the logarithm property that $\log(x^a) = a\log(x)$ and $\log(\frac{a}{b}) = \log(a) - \log(b)$. Using maximum likelihood estimation (MLE), which of the following statements is true regarding the MLE of λ?

A. The maximum likelihood estimate of λ is $\lambda_{\text{MLE}} = 3$.

B. The maximum likelihood estimate of λ is $\lambda_{\text{MLE}} = \frac{9}{2}$.

C. The maximum likelihood estimate of λ is $\lambda_{\text{MLE}} = \frac{3}{4}$.

D. The maximum likelihood estimate of λ is $\lambda_{\text{MLE}} = 4$.

Question 7. [1 MARK]

Suppose we want to model the number of lightnings appearing in Edmonton each year using a Poisson distribution that is governed by the parameter λ. We place a Gamma prior on λ such that $\lambda \sim \text{Gamma}(a, b)$, with probability density function given by

$$p(\lambda) \propto \lambda^{a-1} e^{-b\lambda},$$

where $\propto$ means proportional to, that is, excluding the constants that do not depend on λ.
We now observe f_1, f_2 and f_3 lightnings in year one, two and three resulting in a likelihood function that is proportional to $\lambda^{f_1+f_2+f_3} e^{-3\lambda}$. What is the posterior distribution of λ given this data?

A. $p(\lambda|\mathcal{D}) \propto \lambda^{f_1+f_2+f_3-a+1} e^{-(b-3)\lambda}$.

B. $p(\lambda|\mathcal{D}) \propto \lambda^{a-1-f_1-f_2-f_3} e^{-(3-b)\lambda}$.

C. $p(\lambda|\mathcal{D}) \propto \lambda^{f_1+f_2+f_3+a-1} e^{-b\lambda}$.

D. $p(\lambda|\mathcal{D}) \propto \lambda^{a-1+f_1+f_2+f_3} e^{-(b+3)\lambda}$.

Question 8. [1 MARK]

Suppose we want to perform ridge regression, where the data is generated from a Gaussian distribution with mean $\mathbf{x}^T\mathbf{w}$ and variance 1. For the bias term w_0, we assume a Gaussian prior with zero mean and a very large variance a: $p(w_0) = \sqrt{\frac{1}{2\pi a}} \exp\left(-\frac{w_0^2}{2a}\right)$. For each regression weight w_j for $j = 1, \ldots, d$, we assume a Gaussian prior with zero mean and variance $1/\lambda$: $p(w_j) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2} w_j^2\right)$, where $\lambda \geq 0$ is the regularization parameter. All weights $w_0, w_1, \ldots, w_d$ are independent. The MAP estimate of the weights is given by

$$\mathbf{w}_{\text{MAP}} = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \mathbf{w})^2 - \log\big(p(\mathbf{w})\big) \right\}.$$

What is the expression for $-\log(p(\mathbf{w}))$?

A. $-\frac{1}{2}\log(\frac{1}{2\pi a}) - \frac{d}{2}\log\left(\frac{\lambda}{2\pi}\right) + \frac{w_0^2}{2a} + \frac{\lambda}{2}\sum_{j=1}^{d} w_j^2$.

B. $-\frac{1}{2}\log(\frac{1}{2\pi a}) - \frac{d}{2}\log\left(\frac{\lambda}{2\pi}\right) + \frac{1}{2a} + \frac{\lambda}{2}\sum_{j=1}^{d} w_j^2$.

C. $\frac{1}{2}\log(\frac{1}{2\pi a}) + \frac{d}{2}\log\left(\frac{\lambda}{2\pi}\right) + \frac{w_0^2}{2a} + \frac{\lambda}{2}\sum_{j=1}^{d} w_j^2$.

D. $-\frac{1}{2}\log(\frac{1}{2\pi a}) + \frac{d}{2}\log\left(\frac{\lambda}{2\pi}\right) - \frac{1}{2a} - \frac{\lambda}{2}\sum_{j=1}^{d} w_j^2$.

Question 9. [1 MARK]

Let the dataset be $\mathcal{D} = ((x_1, y_1), \ldots, (x_n, y_n))$, the mini-batch size $b \in \mathbb{N}$, and $M = \text{floor}(n/b)$ is the number of full batches. In class we learned about mini-batch gradient descent (MBGD). In this question we are interested in the computational efficiency of MBGD compared to batch gradient descent. Which of the following statements are true?

A. MBGD is more efficient than batch gradient descent because it requires less computation for the same number of epochs.

B. MBGD is more efficient than batch gradient descent because the gradient estimate is more precise.

C. MBGD usually finds a better solution in the same number of epochs compared to batch gradient descent.

D. MBGD is often more efficient than batch gradient descent because it updates the weights more frequently.

Question 10. [1 MARK]

Let everything be defined as in the previous question. Which of the following statements are true?

A. If n is divisible by b then there are M mini-batches.

B. If n is not divisible by b then the size of the last mini-batch is $n - Mb$.

C. If n is divisible by b then the estimated loss based on the last mini-batch is

$$\frac{1}{b} \sum_{i=(M-1)b+1}^{Mb} \ell(f(\mathbf{x}_i), y_i).$$

D. The variance of the estimated loss for each mini-batch (not considering the last batch) increases if b decreases.

Question 11. [1 MARK]

Let everything be defined as in the previous two questions. Your friend is trying to implement the version of mini-batch gradient descent discussed in the previous two questions with a constant step size. They have written the following pseudocode and asked you to review it. Please select all possible mistakes in the pseudocode below. Multiple statements may be correct.

Algorithm 1: MBGD Linear Regression Learner (with a constant step size and last mini-batch)

1: **input:** $\mathcal{D} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$, step size η, number of epochs T, mini-batch size b
2: $\mathbf{w} \leftarrow$ random vector in $\mathbb{R}^{d+1}$
3: $M \leftarrow \text{floor}\left(\frac{n}{b}\right)$
4: **for** $m = 1, \ldots, M$ **do**
5: randomly shuffle $\mathcal{D}$
6: **for** $t = 1, \ldots, T$ **do**
7: $\nabla \hat{L}(\mathbf{w}) \leftarrow \frac{2}{b} \sum_{i=(m-1)b+1}^{mb} (\mathbf{x}_i^\top \mathbf{w} - y_i)\mathbf{x}_i$
8: $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \hat{L}(\mathbf{w})$
9: **if** $n > Mb$ **then**
10: $\nabla \hat{L}(\mathbf{w}) \leftarrow \frac{2}{n-Mb} \sum_{i=Mb+1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)\mathbf{x}_i$
11: $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \hat{L}(\mathbf{w})$
12: **return** $\hat{f}(\mathbf{x}) = \mathbf{x}^\top \hat{\mathbf{w}}$

A. There are no mistakes. The pseudocode is correct.

B. The pseudocode is incorrect because the outer loop (line 4) should iterate over t and the inner loop (line 6) over m.

C. The pseudocode is incorrect because the gradient calculation for the last mini-batch is incorrect.

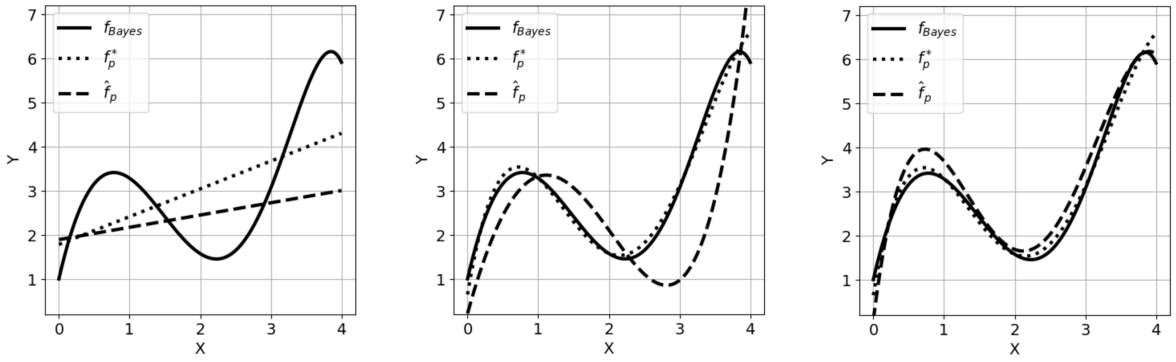
D. The pseudocode is incorrect because we update the parameter twice: in lines 8 and 11.

Question 12. [1 MARK]

Let ϕ_p be the polynomial feature map of degree p, and $\mathcal{F}_p$ the function class containing all polynomials of degree p or less.
Recall that

$$f_{\text{Bayes}} = \underset{f \in \{f | f : \mathbb{R}^{d+1} \to \mathbb{R}\}}{\arg\min} L(f), \quad f_p^* = \underset{f \in \mathcal{F}_p}{\arg\min} L(f), \quad \hat{f}_p = \underset{f \in \mathcal{F}_p}{\arg\min} \hat{L}(f).$$

Below are plots for different values of p and dataset size n. Which of the following statements are true?



(a) Fig 1 (b) Fig 2 (c) Fig 3

A. The data set sizes n does not affect f_p^*.

B. The data set sizes n does not affect f_{Bayes}.

C. For very large data set sizes n, $\hat{f}_p$ in Fig 1 will never be closer to f_{Bayes} than $\hat{f}_p$ in Fig 2, assuming that the polynomial degree p in Fig 1 is smaller than in Fig 2.

D. That $\hat{f}_p$ is much closer to f_p^* in Fig 3 than in Fig 2 might stem from a larger dataset size n.

Question 13. [1 MARK]

Let everything be defined as in the previous question. Which of the following statements are true?

A. The remaining mismatch between f_p^* and f_{Bayes} in Fig 3 is due to the irreducable error.

B. Increasing the polynomial degree p will reduce the approximation and the estimation error.

C. The underlying function that generated the data can be less well represented by quadratic functions than by the function class chosen in Fig 2.

D. $\hat{f}_p$ would match f_{Bayes} perfectly if the approximation and the estimation error are both zero.

Question 14. [1 MARK]

You have access to the true feature-label distribution $\mathbb{P}_{\mathbf{X},Y}$ that generated the data. You are interested in training a model and impatiently start plotting the following figure and start trying to make sense of it. Which of the following statements are true?

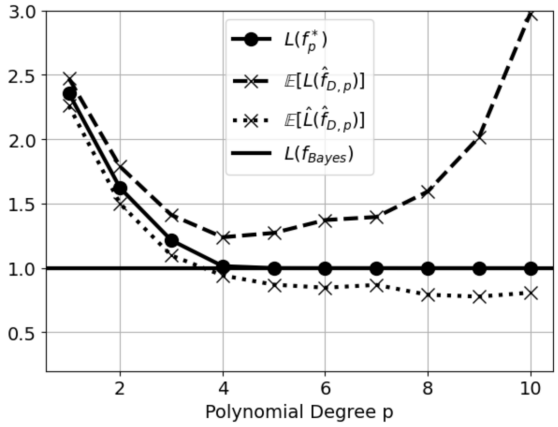


Fig 1

A. $E[\hat{L}(\hat{f}_{D,p})]$ decreases to a lower value than $L(f_{\text{Bayes}})$ for larger p because $\hat{f}_{D,p}$ fits the dataset it was trained on better than f_{Bayes}.

B. f_{Bayes} is a better predictor than $\hat{f}_{D,p}$ for the true feature-label distribution.

C. The expected value of $E[\hat{L}(\hat{f}_{D,p})]$ is calulated with respect to true feature-label distribution $\mathbb{P}_{\mathbf{X},Y}$.

D. The expected value of $E[\hat{L}(\hat{f}_{D,p})]$ is calulated with respect to distribution over the polynomial degree $\mathbb{P}_p$.

Question 15. [1 MARK]

Let ϕ_p be the polynomial feature map of degree p. The function class containing all polynomials of degree p or less is

$$\mathcal{F}_p = \left\{ f \,\middle|\, f : \mathbb{R}^{d+1} \to \mathbb{R}, \text{ and } f(\mathbf{x}) = \phi_p(\mathbf{x})^\top \mathbf{w} \text{ for some } \mathbf{w} \in \mathbb{R}^{\bar{p}} \right\}.$$

Which of the following statements is true?

A. The dimension of $\phi_p(\mathbf{x})$ is $\bar{p} = \binom{p+d}{d}$.

B. $\mathcal{F}_p \subseteq \mathcal{F}_{p+1}$.

C. The dimension of $\mathbf{w}$ is $d+1$.

D. $\mathcal{F}_{25}$ contains exponential functions.

For your notes (1/4)

For your notes (2/4)

For your notes (3/4)

For your notes (4/4)

Formula Sheet

Integration

$$\int_a^b x^d dx = \frac{x^{d+1}}{d+1}\Big|_a^b = \frac{b^{d+1}-a^{d+1}}{d+1} \quad \text{for } d \neq -1$$

Derivatives and Gradient

$$f(x) = x^a, \qquad f'(x) = \frac{df}{dx}(x) = ax^{a-1}$$

$$f(x) = \exp(x), \qquad f'(x) = \frac{df}{dx}(x) = \exp(x)$$

$$f(x) = \ln(x), \qquad f'(x) = \frac{df}{dx}(x) = \frac{1}{x}$$

$$f(x) = g(h(x)), \quad u = h(x) \qquad f'(x) = \frac{df}{dx}(x) = \frac{dg}{du}\frac{du}{dx}(x) = g'(u)h'(x) \quad \triangleright \text{Chain rule}$$

$$f(x) = g(x)h(x), \qquad f'(x) = \frac{df}{dx}(x) = g'(x)h(x) + g(x)h'(x) \quad \triangleright \text{Product rule}$$

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \frac{\partial f}{\partial x_2}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_d}(\mathbf{x})\right)^\top \quad \text{for } \mathbf{x} \in \mathbb{R}^d$$

Probability

Univariate: $\mathbb{P}(X \in \mathcal{E}) \overset{\text{def}}{=} \begin{cases} \sum_{x\in\mathcal{E}} p(x) & \text{if } X \text{ is discrete} \\ \int_{\mathcal{E}} p(x)dx & \text{if } X \text{ is continuous} \end{cases}$

Multivariate: $\mathbb{P}(X \in \mathcal{E}_X, Y \in \mathcal{E}_Y) \overset{\text{def}}{=} \begin{cases} \sum_{x\in\mathcal{E}_x}\sum_{y\in\mathcal{E}_y} p(x,y) & \text{if } X,Y \text{ are discrete} \\ \int_{\mathcal{E}_x}\int_{\mathcal{E}_y} p(x,y)dydx & \text{if } X,Y \text{ are continuous} \\ \int_{\mathcal{E}_x}\sum_{y\in\mathcal{E}_y} p_{Y|X}(y|x)p_X(x)dx & \text{if } X \text{ is continuous, } Y \text{ is discrete} \\ \sum_{x\in\mathcal{E}_x}\int_{\mathcal{E}_y} p_{Y|X}(y|x)p_X(x)dy & \text{if } X \text{ is discrete, } Y \text{ is continuous} \end{cases}$

Marginal pmf: $p_X(x) \overset{\text{def}}{=} \begin{cases} \sum_{y\in\mathcal{Y}} p(x,y) & \text{if } Y \text{ is discrete} \\ \int_{\mathcal{Y}} p(x,y)dy & \text{if } Y \text{ is continuous} \end{cases}$

Marginal: $\mathbb{P}_X(X \in \mathcal{E}_X) \overset{\text{def}}{=} \begin{cases} \sum_{x\in\mathcal{E}_X} p_X(x) & \text{if } X \text{ is discrete} \\ \int_{\mathcal{E}_X} p_X(x)dx & \text{if } X \text{ is continuous} \end{cases}$

Conditional pmf: $p_{Y|X}(y|x) \overset{\text{def}}{=} \frac{p(x,y)}{p_X(x)}$ such that $p_X(x) > 0$

Conditional: $\mathbb{P}_{Y|X}(Y \in \mathcal{E}_Y | X = x) \overset{\text{def}}{=} \begin{cases} \sum_{y\in\mathcal{E}_Y} p_{Y|X}(y|x) & \text{if } Y \text{ is discrete} \\ \int_{\mathcal{E}_Y} p_{Y|X}(y|x)dy & \text{if } Y \text{ is continuous} \end{cases}$

Product Rule: $p(x,y) = p_{Y|X}(y|x)p_X(x)$

Bayes' Rule: $p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)}$

Independence: $p(x_1,\ldots,x_n) = p_{X_1}(x_1)\cdots p_{X_n}(x_n)$

Distribution	Parameters	pmf or pdf	Expectation and Variance
Bernoulli	$\alpha \in [0,1]$	$p(x) = \alpha^x(1-\alpha)^{1-x}$, $x \in \{0,1\}$	$\mathbb{E}[X] = \alpha$, $\mathrm{Var}[X] = \alpha(1-\alpha)$
Discrete Uniform	$n \in \mathbb{N}$	$p(x) = \frac{1}{n}$, $x \in \{1,\ldots,n\}$	$\mathbb{E}[X] = \frac{n+1}{2}$, $\mathrm{Var}[X] = \frac{n^2-1}{12}$
Continuous Uniform	$a,b \in \mathbb{R}, a < b$	$p(x) = \frac{1}{b-a}$, $x \in [a,b]$	$\mathbb{E}[X] = \frac{a+b}{2}$, $\mathrm{Var}[X] = \frac{(b-a)^2}{12}$
Normal	$\mu \in \mathbb{R}, \sigma^2 > 0$	$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, $x \in \mathbb{R}$	$\mathbb{E}[X] = \mu$, $\mathrm{Var}[X] = \sigma^2$
Laplace	$\mu \in \mathbb{R}, b > 0$	$p(x) = \frac{1}{2b}\exp\left(-\frac{\lvert x-\mu\rvert}{b}\right)$, $x \in \mathbb{R}$	$\mathbb{E}[X] = \mu$, $\mathrm{Var}[X] = 2b^2$

Expectation and Variance

Univariate: $\mathbb{E}[X] \stackrel{\text{def}}{=} \begin{cases} \sum_{x\in\mathcal{X}} xp(x) & \text{if } X \text{ is discrete} \\ \int_{\mathcal{X}} xp(x)dx & \text{if } X \text{ is continuous} \end{cases}$

Function: $\mathbb{E}[f(X)] \stackrel{\text{def}}{=} \begin{cases} \sum_{x\in\mathcal{X}} f(x)p(x) & \text{if } X \text{ is discrete} \\ \int_{\mathcal{X}} f(x)p(x)dx & \text{if } X \text{ is continuous} \end{cases}$

Variance: $\mathrm{Var}[X] \stackrel{\text{def}}{=} \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$

Multivariate: $\mathbb{E}[f(X,Y)] \stackrel{\text{def}}{=} \begin{cases} \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} f(x,y)p(x,y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{\mathcal{X}}\int_{\mathcal{Y}} f(x,y)p(x,y)\,dy\,dx & \text{if } X \text{ and } Y \text{ are continuous} \\ \int_{\mathcal{X}}\sum_{y\in\mathcal{Y}} f(x,y)p_{Y|X}(y|x)p_X(x)\,dx & \text{if } X \text{ is continuous, } Y \text{ is discrete} \\ \sum_{x\in\mathcal{X}}\int_{\mathcal{Y}} f(x,y)p_{Y|X}(y|x)p_X(x)\,dy & \text{if } X \text{ is discrete, } Y \text{ is continuous} \end{cases}$

Conditional: $\mathbb{E}[f(Y)|X=x] \stackrel{\text{def}}{=} \begin{cases} \sum_{y\in\mathcal{Y}} f(y)p_{Y|X}(y|x) & \text{if } Y \text{ is discrete} \\ \int_{\mathcal{Y}} f(y)p_{Y|X}(y|x)\,dy & \text{if } Y \text{ is continuous} \end{cases}$

Expectation and Variance Properties

1. $\mathbb{E}[cX] = c\mathbb{E}[X]$
2. $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
3. $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$
4. $\mathrm{Var}[c] = 0$
5. $\mathrm{Var}[cX] = c^2\mathrm{Var}[X]$.

If X and Y are independent:

6. $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$
7. $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

Estimation

Sample Mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ **If X_i are i.i.d.:** $\mathbb{E}[\bar{X}] = \mathbb{E}[X_1], \quad \mathrm{Var}\left[\bar{X}\right] = \frac{\mathrm{Var}[X_1]}{n}$

Estimate of Expected Loss: $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\boldsymbol{X}_i), Y_i)$

Optimization

2nd Order GD: $w^{(t+1)} = w^{(t)} - \frac{g'(w^{(t)})}{g''(w^{(t)})}$ **1st Order GD:** $w^{(t+1)} = w^{(t)} - \eta^{(t)} g'(w^{(t)})$ **Poly Dim:** $\bar{p} = \binom{d+p}{p}$

Evaluation

EE and AE: $\mathbb{E}[L(\hat{f}_D)] = \underbrace{\mathbb{E}[L(\hat{f}_D)] - L(f^*)}_{\text{Estimation Error (EE)}} + \underbrace{L(f^*) - L(f_{\text{Bayes}})}_{\text{Approximation Error (AE)}} + \underbrace{L(f_{\text{Bayes}})}_{\text{Irreducible Error (IE)}}$

Bias and Var: $\mathbb{E}[L(\hat{f}_D)] = \mathbb{E}\Big[\underbrace{\mathbb{E}[(\hat{f}_D(\boldsymbol{X}) - \bar{f}(\boldsymbol{X}))^2|\boldsymbol{X}]}_{\text{Variance}}\Big] + \mathbb{E}\Big[\underbrace{(\bar{f}(\boldsymbol{X}) - f_{\text{Bayes}}(\boldsymbol{X}))^2}_{\text{Bias}}\Big] + \underbrace{L(f_{\text{Bayes}})}_{\text{Irreducible Error}}$

Regularization: $\hat{L}_\lambda(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \vec{w} - y_i)^2 + \frac{\lambda}{n}\sum_{j=1}^{d} w_j^2$

MLE and MAP

MLE: $\arg\max_{w\in\mathcal{W}} p(\mathcal{D}|w)$ **MAP:** $\arg\max_{w\in\mathcal{W}} p(w|\mathcal{D})$