This written part of HW0 covers some of the fundamental concepts in probability theory and linear algebra, the prerequisites of the course. In addition to this written part, there are programming exercises for dynamic programming and PyTorch on Colab.

Problem 1

(conditional probability, Bayes' rule)

There is a multiple choice exam with 5 choices for each question. Suppose there is a 0.5 probability of a student knowing the answer, and a 0.25 probability that they can eliminate a choice, otherwise all 5 choices seem equally plausible. When a student does not know the answer, they would guess randomly from the 4 or 5 choices. If a student answers a question correctly, what is the probability they knew the answer?

Problem 2

(joint probability, marginal probability, independence, expectation)

Let $X \sim Ber(0.3)$ and $Y \sim Ber(0.7)$ be independent random variables. Define S and T by: $S = X + Y$ and $T = X - Y$.

(a) Find the joint and marginal PMFs for S and T.

(b) Are S and T independent?

(c) Find $\mathbb{E}[S]$ and $\mathrm{Var}(2T)$.

Problem 3

In this problem we will use the problem of linear least squares to review several concepts from linear algebra. We are given a data matrix $X \in \mathbb{R}^{n \times d}$ and the corresponding labels $y \in \mathbb{R}^n$, and the ordinary least squares problem is defined as:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T w)^2 = \min_{w \in \mathbb{R}^d} \frac{1}{n} \|Xw - y\|^2.$$

The above formulation is associated with the linear system

$$Xw = y.$$

(a) (rank) Assume that the columns of X are linearly independent. What is the rank of X?

(b) (solutions of linear systems) Following (a), suppose that $n > d$. In this case, how many solutions can we obtain for the linear system? Please explain.

(c) (singular value decomposition) Now assume that $n < d$ and the rows of X are linearly independent. In this case, the linear system will have infinitely many solutions. A classic way is to select the solution with minimal norm, which would be $w = X^{\dagger}y = X^T(XX^T)^{-1}y$ (we won't prove this here). $X^{\dagger}$ is also called the Moore-Penrose inverse of X, which in practice can be calculated by the singular value decomposition (SVD). Recall that the SVD finds a decomposition $X = U\Sigma V^T$ with orthogonal matrices $U \in \mathbb{R}^{n\times n}, V \in \mathbb{R}^{d\times d}$ and rectangular diagonal matrix $\Sigma \in \mathbb{R}^{n\times d}$. Show with SVD that the solution can be written in the form of

$$w = X^{\dagger}y = V\Sigma^{\dagger}U^T y,$$

where $\Sigma^{\dagger} \in \mathbb{R}^{d\times n}$ has reciprocals of Σ in its diagonal entries.

(d) (matrix calculus) Another way to solve linear least squares is with (batch) gradient descent (GD), which is an iterative optimization algorithm for finding the local minimum of a function. Its stochastic approximation, the (mini-batch) stochastic gradient descent (SGD), is commonly used in training modern neural networks. For linear least squares, the objective is convex and differentiable, so we can use the update rule

$$w_t = w_{t-1} - \gamma \nabla_w \mathcal{L}(w),$$

where $\mathcal{L}(w) = \frac{1}{n}\|Xw - y\|^2$, with a γ that is small enough to achieve the optimal solution. Derive $\nabla_w \mathcal{L}$ and write down the gradient descent iteration.

Problem 1

(conditional probability, Bayes' rule)

There is a multiple choice exam with 5 choices for each question. Suppose there is a 0.5 probability of a student knowing the answer, and a 0.25 probability that they can eliminate a choice, otherwise all 5 choices seem equally plausible. When a student does not know the answer, they would guess randomly from the 4 or 5 choices. If a student answers a question correctly, what is the probability they knew the answer?

Let: $A = \{\text{know the answer}\}$

$B = \{\neg\text{know the answer} \wedge \text{can eliminate 1 option}\}$

$C = \{\neg\text{know the answer} \wedge \neg\text{can eliminate 1 option}\}$

$$P(\text{correct answer}) = P(\text{correct answer} \mid A) \cdot P(A) + P(\text{correct answer} \mid B) \cdot P(B) + P(\text{correct answer} \mid C) \cdot P(C)$$

$$P(\text{correct answer}) = 1 \cdot 0.5 + 0.25 \cdot 0.25 + 0.2 \cdot 0.25$$

$$P(\text{correct answer}) = 0.5 + 0.0625 + 0.05$$

$$P(\text{correct answer}) = 0.6125$$

By Bayes' rule:

$$P(\text{know the answer} \mid \text{correct answer}) = \frac{P(\text{correct answer} \mid A) \cdot P(A)}{P(\text{correct answer})}$$

$$= \frac{0.5}{0.6125}$$

$$P(\text{know the answer} \mid \text{correct answer}) \approx 0.816$$

Problem 2

(joint probability, marginal probability, independence, expectation)

Let $X \sim Ber(0.3)$ and $Y \sim Ber(0.7)$ be independent random variables. Define S and T by: $S = X + Y$ and $T = X - Y$.

(a) Find the joint and marginal PMFs for S and T.

I = Independence

* $P(\cap) = 0$. events are disjoint

Marginal S

$P(S=0) = P(X=0, Y=0) \overset{I}{=} 0.3 \cdot 0.7 = 0.21$

$P(S=1) \overset{*}{=} P(X=1, Y=0) + P(X=0, Y=1) \overset{I}{=} 0.3^2 + 0.7^2 = 0.58$

$P(S=2) = P(X=1, Y=1) \overset{I}{=} 0.7 \cdot 0.3 = 0.21 \; (= 1 - P(S=0) - P(S=1))$

Marginal T

$P(T=-1) = P(X=0, Y=1) \overset{I}{=} 0.7 \cdot 0.7 = 0.49$

$P(T=0) \overset{*}{=} P(X=0, Y=0) + P(X=1, Y=1) \overset{I}{=} 0.7 \cdot 0.3 + 0.3 \cdot 0.7 = 0.42$

$P(T=1) = P(X=1, Y=0) \overset{I}{=} 0.3^2 = 0.09$

Joint S, T

$(S=0)$ $P(S=0, T=-1) = P(S=0, T=1) = 0$ | $P(S=0, T=0) = P(X=0, Y=0) = 0.21$
$(\neq P(S=0) \cdot P(T=0) = 0{,}0882)$

$(S=1)$ $P(S=1, T=0) = 0$ | $P(S=1, T=-1) = P(X=0, Y=1) = 0.49$ | $P(S=1, T=1) = P(X=1, Y=0) = 0{,}09$
$(\neq P(S=1) \cdot P(T=1) = 0{,}0522)$

$(S=2)$ $P(S=2, T=-1) = P(S=2, T=1) = 0$ | $P(S=2, T=0) = P(X=1, Y=1) = 0.21$
$(\neq P(S=2) \cdot P(T=0) = 0{,}0882)$

(b) Are S and T independent? ¡NO! $P(S=s, T=t) \neq P(S=s, T=t)$ for some (s,t).

(c) Find $\mathbb{E}[S]$ and $\mathrm{Var}(2T)$. $E[S] = 0.58 + 0.42 = 1$. (symmetric wrt 1).

$\mathrm{VAR}(2T) = 4\,\mathrm{VAR}(T) = 4\left[E[T^2] - E[T]^2\right] = 4\left[0.58 - (-0.49 + 0.09)^2\right] = 4 \cdot [0.58 - 0{,}16] = 1{,}68$

Problem 3

In this problem we will use the problem of linear least squares to review several concepts from linear algebra. We are given a data matrix $X \in \mathbb{R}^{n\times d}$ and the corresponding labels $y \in \mathbb{R}^n$, and the ordinary least squares problem is defined as:

$$\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T w)^2 = \min_{w\in\mathbb{R}^d} \frac{1}{n}\|Xw - y\|^2.$$

The above formulation is associated with the linear system

$$Xw = y.$$

$X = n \begin{bmatrix} & & \end{bmatrix}_d$

(a) (rank) Assume that the columns of X are linearly independent. What is the rank of X?

$X : \mathbb{R}^d \to \mathbb{R}^n$.

Rank is d. If X has LI columns then $n \geq d$. (otherwise is impossible). If X has LI columns then X is injective as linear map, thus: $\dim \operatorname{Im}(X) = d - \dim \ker(X) = d$ (Rank – nullity theorem)

(b) (solutions of linear systems) Following (a), suppose that $n > d$. In this case, how many solutions can we obtain for the linear system? Please explain.

Either one or NONE. From a LA pov. Since RANK $X < n$, then as a linear map, X is not surjective. Thus, for $y \in \mathbb{R}^n$, we have two options: either $y \in \operatorname{Im}(X)$ or $y \in \operatorname{Im}(X)^c$ $(\neq \emptyset)$. If $y \in \operatorname{Im}(X)$ then $y = Xw$ for some unique w (because X is injective) otherwise $y \notin \operatorname{Im}(X)^c$ and the system has no solutions.

From a matrix pov. the system $Xw = y$ is equivalent to finding vectors $w \in \mathbb{R}^d$ who lie in the intersection of the sets

$$\{X_i v = y_i\} \quad i \in \{1, \dots, n\}$$

where X_i is the ith row of X and y_i is the ith element of y. It suffices to show that it is not possible that the intersection has dim ≥ 1. If that was the case, we would have solutions w_1, w_2 distinct. But then

$$X(w_1 - w_2) = 0$$

which contradicts the columns being LI

(c) (singular value decomposition) Now assume that $n < d$ and the rows of X are linearly independent. In this case, the linear system will have infinitely many solutions. A classic way is to select the solution with minimal norm, which would be $w = X^\dagger y = X^T(XX^T)^{-1}y$ (we won't prove this here). $X^\dagger$ is also called the Moore-Penrose inverse of X, which in practice can be calculated by the singular value decomposition (SVD). Recall that the SVD finds a decomposition $X = U\Sigma V^T$ with orthogonal matrices $U \in \mathbb{R}^{n\times n}, V \in \mathbb{R}^{d\times d}$ and rectangular diagonal matrix $\Sigma \in \mathbb{R}^{n\times d}$. Show with SVD that the solution can be written in the form of

$$w = X^\dagger y = V\Sigma^\dagger U^T y,$$

where $\Sigma^\dagger \in \mathbb{R}^{d\times n}$ has reciprocals of Σ in its diagonal entries.

WHT:

$$X^\dagger y = X^T (XX^T)^{-1} y$$

$$= V\Sigma^T U^T (U\Sigma V^T V \Sigma^T U^T)^{-1} y$$

$$= V\Sigma^T U^T (U\Sigma\Sigma^T U^T)^{-1} y$$

note that

$$(U\Sigma\Sigma^T U^T)\cdot(U(\Sigma^\dagger)^T \Sigma^\dagger U^T) = I_n$$

Hence

$$X^\dagger y = V\Sigma^T U^T U (\Sigma^\dagger)^T \Sigma^\dagger U^T y$$

$$= V \underbrace{\Sigma^T(\Sigma^T)^\dagger}_{I_d} \Sigma^\dagger U^T y$$

$$= V\Sigma^\dagger U^T y.$$

(d) (matrix calculus) Another way to solve linear least squares is with (batch) gradient descent (GD), which is an iterative optimization algorithm for finding the local minimum of a function. Its stochastic approximation, the (mini-batch) stochastic gradient descent (SGD), is commonly used in training modern neural networks. For linear least squares, the objective is convex and differentiable, so we can use the update rule

$$w_t = w_{t-1} - \gamma \nabla_w \mathcal{L}(w),$$

where $\mathcal{L}(w) = \frac{1}{n}\|Xw - y\|^2$, with a γ that is small enough to achieve the optimal solution. Derive $\nabla_w \mathcal{L}$ and write down the gradient descent iteration.

WHT:

$$\mathcal{L}(w) = \frac{1}{n} \sum_{k=1}^{n} (x_k^T w - y_k)^2$$

AND

$$\frac{\partial \mathcal{L}}{\partial w_i} = \frac{2}{n} \sum_{k=1}^{n} x_{ki} (x_k^T w - y_k) = \frac{2}{n} X_{-,i}^T (Xw - y)$$

Hence.

$$\nabla_w \mathcal{L} = \frac{2}{n} X^T (Xw - y)$$

And thus:

$$\begin{aligned} w_t &= w_{t-1} - \gamma \nabla_w \mathcal{L}(w_{t-1}) \\ &= w_{t-1} - \frac{2\gamma}{n} X^T (Xw - y) \end{aligned}$$