CMO Assignment 3

SWAPNIL MITESHKUMAR JOSHI swapniljoshi@iisc.ac.in
SR number : 25846

Question 1

Consider the optimization problem of a convex objective function with an added sparsity promoting constraint:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} f(\boldsymbol{\beta}) \quad \text{s.t. } \|\boldsymbol{\beta}\|_1 \leq t \tag{1}$$

However, in practice we often optimize

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \; f(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \tag{2}$$

where $\lambda > 0$ is a regularization parameter.
When the objective function is of the form

$$f(\boldsymbol{\beta}) = \frac{1}{2}\|X\boldsymbol{\beta} - \mathbf{y}\|_2^2, \tag{3}$$

this becomes the problem of linear regression with an added sparsity-promoting regularizer (also known as *LASSO regression*).

1.1

Derive the KKT conditions for the optimization problem (1) in terms of the subgradient of $\|\boldsymbol{\beta}\|_1$.

Derivation of KKT Conditions

We consider the constrained LASSO optimization problem

$$\min_{\beta\in\mathbb{R}^n} \; f(\beta) \quad \text{s.t.} \quad \|\beta\|_1 \leq t,$$

where

$$f(\beta) = \frac{1}{2}\|X\beta - y\|_2^2.$$

The corresponding Lagrangian is

$$\mathcal{L}(\beta, \mu) = f(\beta) + \mu(\|\beta\|_1 - t),$$

with dual variable $\mu \geq 0$. The KKT conditions for the optimal solution (β^*, μ^*) are:

1. **Primal feasibility:**
$$\|\beta^*\|_1 \le t$$

2. **Dual feasibility:**
$$\mu^* \ge 0$$

3. **Complementary slackness:**
$$\mu^* \big(\|\beta^*\|_1 - t\big) = 0$$

4. **Stationarity (subgradient condition):**
$$0 \in \nabla f(\beta^*) + \mu^* \, \partial \|\beta^*\|_1$$
where
$$\nabla f(\beta^*) = X^\top (X\beta^* - y),$$
and the coordinate-wise subgradient of $\|\beta\|_1$ is
$$\partial|\beta_j^*| = \begin{cases} \text{sign}(\beta_j^*), & \beta_j^* \neq 0, \\ [-1, 1], & \beta_j^* = 0. \end{cases}$$
Hence, for each coordinate j:
$$\beta_j^* \neq 0 \Rightarrow X_j^\top (X\beta^* - y) = -\mu^* \, \text{sign}(\beta_j^*),$$
$$\beta_j^* = 0 \Rightarrow \big|X_j^\top (X\beta^* - y)\big| \le \mu^*.$$

These conditions together characterize the optimal solution for the constrained LASSO problem.

1.2
Is optimizing (1) equivalent to optimizing (2)? Reason out why optimization problem (2) is solved in practice instead of solving (1).

Equivalence of Problems (1) and (2)
The constrained LASSO problem
$$\min_\beta f(\beta) \quad \text{s.t. } \|\beta\|_1 \le t$$
and the penalized LASSO problem
$$\min_\beta \; f(\beta) + \lambda \|\beta\|_1, \qquad \lambda > 0,$$

are equivalent in the sense that both generate the same solution set $\{\beta^*\}$. This follows from convex optimization and Lagrangian duality: the optimal regularization parameter λ in (2) corresponds to the optimal Lagrange multiplier μ^* of the constraint in (1).

Why (2) is solved in practice:

- Problem (1) requires repeated projection onto the ℓ_1-ball $\{\beta : \|\beta\|_1 \leq t\}$, which is computationally expensive.
- Problem (2) is unconstrained. The non-smooth term $\lambda\|\beta\|_1$ can be handled efficiently via proximal and coordinate-descent methods. This allows the use of "highly efficient specialized algorithms" (such as variants of Coordinate Descent) that bypass the need for a slow, iterative projection onto the feasible set, making the overall process much faster.

Strong duality holds because the objective is convex and the ℓ_1-ball constraint admits a strictly feasible interior point (Slater's condition). Thus, solving (2) is a numerically efficient way of solving the constrained LASSO formulation (1).

1.3

Implement LASSO regression for the linear objective function using CVXPY for three values of λ: 0.01, 0.1 and 1. Plot the number of fitted (non-zero) coefficients against λ, to compare the sparsity of β^*.

Implementation and sparsity plot

LASSO was solved using CVXPY for three values of the regularization parameter $\lambda \in \{0.01, 0.1, 1\}$. The number of fitted (non-zero) coefficients for each λ is reported below and the sparsity plot is included.

λ	# non-zero coefficients
0.01	15
0.10	15
1.00	13

Table 1: Number of non-zero coefficients in β^* for each tested λ.

These results indicate that for this particular dataset the LASSO penalty does not induce sparsity for smaller values of λ (0.01 and 0.1), but only when λ becomes sufficiently large (here at $\lambda = 1$), the solution begins to zero out coefficients and promote sparsity.

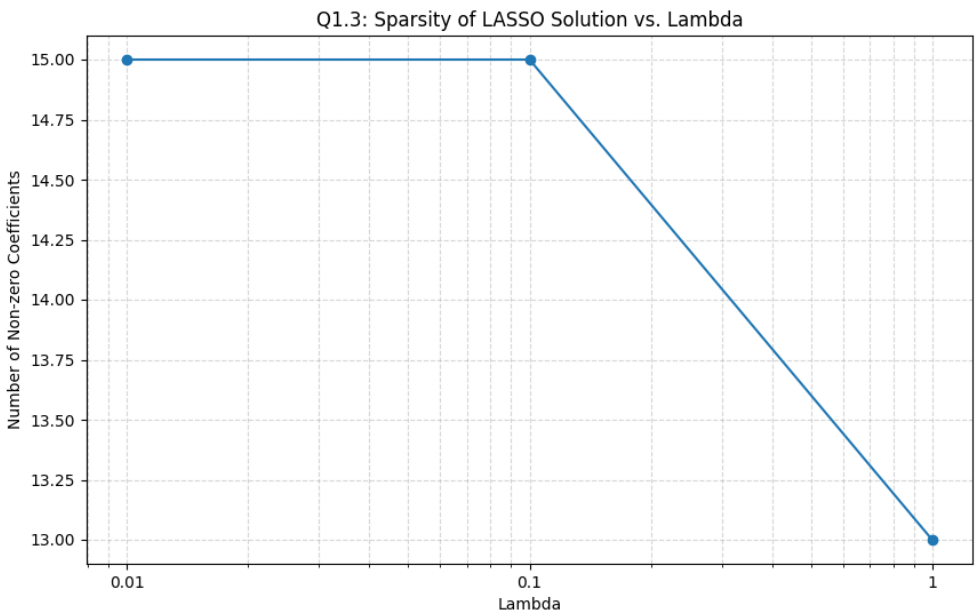

Figure 1: Number of non-zero coefficients in β^* versus λ.

1.4
Verify if the KKT conditions hold for your optimal β^* and report your observations.

A key point is that our solution β^* was obtained by solving the (unconstrained) problem (2) , which uses λ. To verify the KKT conditions for the *constrained* problem (1), which uses a constraint t,
These two problems are equivalent. A solution β^* that solves the λ-problem is also the solution to the t-problem for a *specific* t, namely $t = \|\beta^*\|_1$. The Lagrange multiplier μ for the t-problem is simply equal to our λ. Therefore, checking the KKT conditions for the λ-proble is mathematically identical to verifying the KKT conditions for the t-problem, where $\mu = \lambda$ and $t = \|\beta^*\|_1$.

Verification of KKT Conditions
For the optimal solution β^* obtained in 1.3, we verify the KKT conditions.
The subgradient optimality condition for LASSO is:

$$X^\top (X\beta^* - y) + \lambda z = 0,$$

where

$$z_i = \begin{cases} \text{sign}(\beta_i^*), & \beta_i^* \neq 0, \\ \in [-1, 1], & \beta_i^* = 0. \end{cases}$$

We numerically evaluate the residual

$$r = X^\top (X\beta^* - y),$$

and confirm that:

- For $\lambda = 0.01$: all 15 coefficients are non-zero and satisfy $|r_i| = \lambda \Rightarrow z_i = \text{sign}(\beta_i^*)$.
- For $\lambda = 0.1$: again 15 non-zero coefficients satisfy $|r_i| = \lambda$.
- For $\lambda = 1$: 13 coefficients are non-zero with $|r_i| = \lambda$, while for the 2 zero coefficients we verify $|r_i| <= \lambda$, consistent with $z_i \in [-1, 1]$.

Thus, the optimal solutions under all λ values satisfy the KKT subgradient conditions, confirming correct optimality of the computed β^* solutions.

1.5
Consider the case where two features are highly correlated. Duplicate one feature column in X and repeat the experiment. What happens to the LASSO solution? Relate your observation to the KKT subgradient condition.

Correlated features: duplicate-column experiment

We duplicated one feature column (column 0) and appended it as a new last column in the design matrix X, producing two identical columns. The LASSO experiment was re-run for $\lambda \in \{0.01, 0.1, 1\}$. The obtained coefficients for the two identical columns and their sums are reported below.

Effect of Highly Correlated (Duplicated) Feature
We duplicate one feature column in X and re-solve the LASSO problem. For each λ, we report: β_{orig}: coefficient of original column, β_{dup}: coefficient of duplicated column, and the sum showing overall influence.

$\lambda = 0.01$	
Original coefficient (β_{orig}):	-0.02286590030789831
Duplicated coefficient (β_{dup}):	-0.022865900307920142
Sum ($\beta_{\text{orig}} + \beta_{\text{dup}}$):	-0.045731800615818455
$\lambda = 0.10$	
Original coefficient (β_{orig}):	-0.01935822850189861
Duplicated coefficient (β_{dup}):	-0.019358228501914012
Sum ($\beta_{\text{orig}} + \beta_{\text{dup}}$):	-0.03871645700381263
$\lambda = 1.00$	
Original coefficient (β_{orig}):	$-5.6495119142216824 \times 10^{-8}$
Duplicated coefficient (β_{dup}):	$-5.6495119338330237 \times 10^{-8}$
Sum ($\beta_{\text{orig}} + \beta_{\text{dup}}$):	$-1.1299023848054707 \times 10^{-7}$

Table 2: LASSO redistributes weight between perfectly correlated (duplicated) features.

λ	original $\beta[0]$
0.01	-0.045731800615643414
0.10	-0.03871645680191667
1.00	$1.3965287142755212 \times 10^{-9}$

Table 3: Reference: LASSO coefficient before feature duplication

Observations

1. For $\lambda = 0.01$ and $\lambda = 0.1$ the two identical columns receive nearly equal coefficients whose sum closely matches the original single-column coefficient (the original coef-

ficient before duplication is reported in the experimental log). This indicates that LASSO splits the contribution across identical predictors.

2. For $\lambda = 1.0$ both coefficients are numerically zero (order 10^{-7}), showing that a larger penalty can threshold both correlated features to zero.

Relation to the KKT subgradient condition Let the duplicated column indices be j_1 and j_2 with $X_{j_1} = X_{j_2}$. The coordinate-wise stationarity (KKT) condition from Section 1.1 is

$$\beta_j^* \neq 0 \Longrightarrow X_j^\top (X\beta^* - y) = -\mu^* \operatorname{sign}(\beta_j^*),$$
$$\beta_j^* = 0 \Longrightarrow |X_j^\top (X\beta^* - y)| \leq \mu^*.$$

Because $X_{j_1} = X_{j_2}$, the inner products $X_{j_1}^\top (X\beta^* - y)$ and $X_{j_2}^\top (X\beta^* - y)$ are equal. Hence:

- If that common inner product magnitude exceeds the threshold μ^*, both coordinates may be nonzero; symmetry then causes the optimizer to split the weight between the two identical predictors (as observed for small λ).
- If the common inner product magnitude is below the threshold μ^* (for larger λ), the KKT zero-condition holds for both coordinates and both are set to zero (as observed for $\lambda = 1$).

These are consistent with the numerical results in Table 2.

Question 2

Consider the LASSO dual problem. Recall that the LASSO primal can be rewritten as:

$$\min_{\beta} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1$$

and its dual form can be shown to be:

$$\max_{u} -\frac{1}{2}\|u\|_2^2 + y^\top u \quad \text{s.t.} \quad \|X^\top u\|_\infty \leq \lambda$$

2.1 Derive this dual problem starting from the primal formulation using the Lagrangian. Clearly state your steps and assumptions.

Derivation of the Dual Problem

We start from the LASSO primal formulation:

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1,$$

where $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, and $\lambda > 0$.

Step 1. Introduce an auxiliary variable. Let $r = X\beta - y$. The problem becomes:

$$\min_{\beta,r} \frac{1}{2}\|r\|_2^2 + \lambda\|\beta\|_1 \quad \text{s.t.} \quad r = X\beta - y.$$

Step 2. Form the Lagrangian. We introduce the dual variable $u \in \mathbb{R}^n$ for the equality constraint:

$$\mathcal{L}(\beta, r, u) = \frac{1}{2}\|r\|_2^2 + \lambda\|\beta\|_1 + u^\top(y - X\beta - r).$$

Step 3. Minimize with respect to r. Taking the gradient of $\mathcal{L}$ with respect to r and setting it to zero:

$$\nabla_r\mathcal{L} = r - u = 0 \quad \Rightarrow \quad r = u.$$

Substituting this back into the Lagrangian gives:

$$\mathcal{L}(\beta, u) = -\frac{1}{2}\|u\|_2^2 + y^\top u - \beta^\top(X^\top u) + \lambda\|\beta\|_1.$$

Step 4. Minimize with respect to β. The term involving β is:

$$-\beta^\top(X^\top u) + \lambda\|\beta\|_1.$$

This expression is finite if and only if

$$\|X^\top u\|_\infty \le \lambda,$$

and equals 0 when the inequality is satisfied (the infimum is attained at $\beta = 0$). Otherwise, the term goes to $-\infty$, rendering the Lagrangian unbounded.

Step 5. Form the dual function. Hence, the dual function is:

$$g(u) = \begin{cases} -\frac{1}{2}\|u\|_2^2 + y^\top u, & \text{if } \|X^\top u\|_\infty \le \lambda, \\ -\infty, & \text{otherwise.} \end{cases}$$

Step 6. Obtain the dual problem. Maximizing $g(u)$ yields the dual formulation:

$$\max_{u \in \mathbb{R}^n} \; -\frac{1}{2}\|u\|_2^2 + y^\top u \quad \text{s.t.} \quad \|X^\top u\|_\infty \le \lambda.$$

Assumptions:
We assume X and y are finite, and $\lambda > 0$, ensuring convexity of the primal problem and existence of an optimal solution. The dual problem is concave and bounded above since the constraint set $\{u : \|X^\top u\|_\infty \le \lambda\}$ is compact.

Result:
Thus, the LASSO dual is a quadratic maximization problem constrained by an ℓ_∞ norm bound, which enforces the dual–primal coupling critical to sparsity in β.

2.2 Show that strong duality holds under mild assumptions on X.

The primal LASSO problem is

$$\min_\beta \; \frac{1}{2}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1,$$

which is a convex optimization problem because both $\frac{1}{2}\|X\beta - y\|_2^2$ and $\|\beta\|_1$ are convex functions.

For convex problems with affine constraints, strong duality holds whenever there exists a feasible point that satisfies the constraints strictly. In the equivalent constrained form

$$\min_\beta \; \frac{1}{2}\|X\beta - y\|_2^2 \quad \text{s.t.} \quad \|\beta\|_1 \le t,$$

any β with $\|\beta\|_1 < t$ is a strictly feasible point. Hence, the feasibility requirements for strong duality are met.

Therefore, strong duality holds, implying that the optimal values of the primal and dual problems are equal:

$$p^* = d^*.$$

Consequently, the (KKT) conditions are necessary and sufficient for optimality, and solving either the primal or dual gives the same minimum objective value.

2.3 For the same dataset as Question 1, implement the dual formulation using `CVXPY` for $\lambda \in \{0.01, 0.1, 1\}$.

The dual of the LASSO problem is given by:

$$\max_u \; -\frac{1}{2}\|u\|_2^2 + y^\top u \quad \text{s.t.} \quad \|X^\top u\|_\infty \leq \lambda$$

where u represents the dual variable corresponding to the equality constraint in the primal formulation.

Implementation: The dual problem was implemented in `CVXPY` using the same dataset and parameter values as in Question 1.

Results:

λ	**Primal Objective**	**Dual Objective**	**Duality Gap (Primal - Dual)**
0.01	4.579198260204832	4.579198260200089	4.742872761198669e-12
0.1	5.158894963635259	5.158894984539204	-2.0903945241457222e-08
1.0	10.675371230339431	10.675371288194086	-5.785465440055759e-08

Table 4: Primal–Dual Objective Comparison for Different λ Values

Observations:

- The primal and dual objective values are nearly identical for all λ, with negligible duality gaps (below 10^{-7}).
- This confirms strong duality for the LASSO problem and numerical correctness of the dual implementation.
- Increasing λ increases the total objective value, reflecting stronger regularization and sparser β^*.

2.4 Check how closely the optimum value obtained u^* satisfies the relation with β^* obtained in 1.3. Express in terms of L_2 norm.

According to the (KKT) optimality conditions for the LASSO, the relationship between the primal and dual solutions is:

$$u^* = y - X\beta^*.$$

To verify this numerically, we compute the L_2 norm of the difference:

$$\|u^* - (y - X\beta^*)\|_2.$$

Results:

λ	$\|u^* - (y - X\beta^*)\|_2$
0.01	1.3393600420969055e-11
0.1	2.6636683431467843e-09
1.0	3.0799093388008703e-08

Table 5: Verification of Primal–Dual Variable Relationship

Interpretation:

- The L_2 norms are extremely small (close to zero), confirming that u^* satisfies the theoretical KKT relation $u^* = y - X\beta^*$.
- This verifies that both the primal and dual problems are consistent with optimality conditions.

2.5 Explain, in your own words, how the dual constraint $\|X^\top u\|_\infty \leq \lambda$ enforces sparsity in the primal coefficients β^*.

The dual constraint

$$\|X^\top u\|_\infty \leq \lambda$$

directly governs sparsity in the primal coefficients β^* through the KKT stationarity condition:

$$X^\top(X\beta^* - y) + \lambda z = 0, \quad \text{where } z_i \in \begin{cases} \{\text{sign}(\beta_i^*)\}, & \beta_i^* \neq 0, \\ [-1, 1], & \beta_i^* = 0. \end{cases}$$

In the dual, each component of $X^\top u$ corresponds to the negative gradient of the loss with respect to a coefficient β_i. The constraint $\|X^\top u\|_\infty \leq \lambda$ ensures that:

- When $|X_i^\top u| < \lambda$, the corresponding $\beta_i^* = 0$ — the feature's influence is suppressed to satisfy the subgradient condition.
- When $|X_i^\top u| = \lambda$, the coefficient β_i^* can become nonzero — the feature is "active" in explaining the data.

Thus, the dual constraint acts as a thresholding rule: only features whose correlation with the residual vector (represented by u) exactly meets the boundary λ remain active in the model. This mechanism naturally enforces sparsity in β^*, since most features have correlations strictly less than λ, driving their coefficients to zero.

Question 3

3.1 Projections in a Navigation Problem. A robot at position $y \in \mathbb{R}^2$ must stay inside a safe zone. Two possible safe zones are defined as:

(a) A circular base station: $C_1 = \{x : \|x\|_2 \le 5\}$,

(b) A rectangular corridor: $C_2 = \{x : -3 \le x_1 \le 3,\ 0 \le x_2 \le 4\}$.

We compute the projections $\Pi_{C_1}(y)$ and $\Pi_{C_2}(y)$ for several robot positions.
Figures below show four sample robot positions (red) and their corresponding projections (green) onto both the circular and rectangular safe zones.

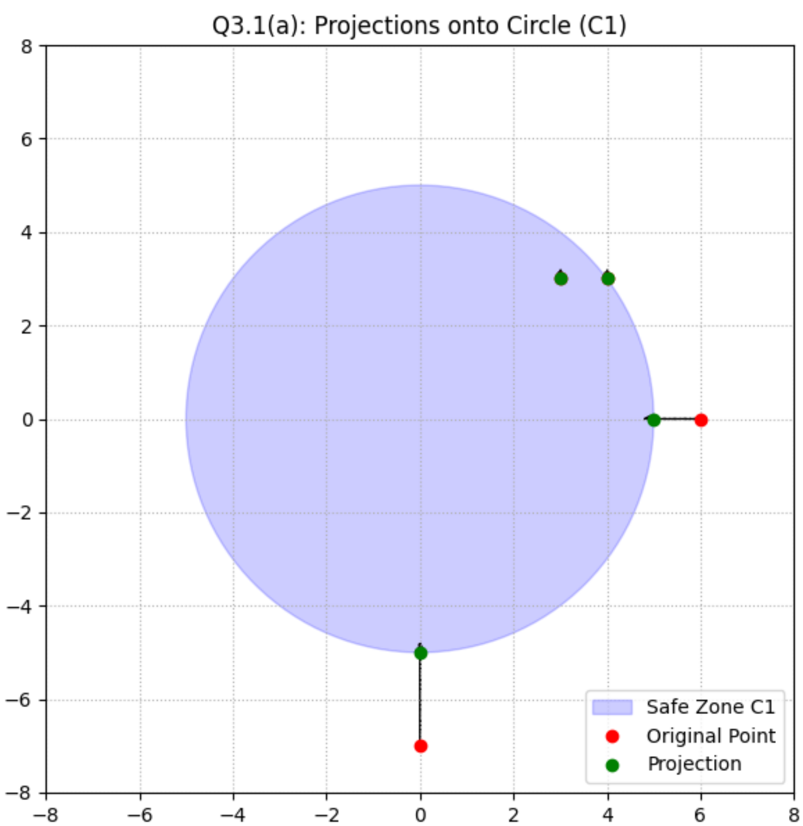

Figure 2: Projection of robot positions onto the circular safe zone, $\Pi_{C_1}(y)$.

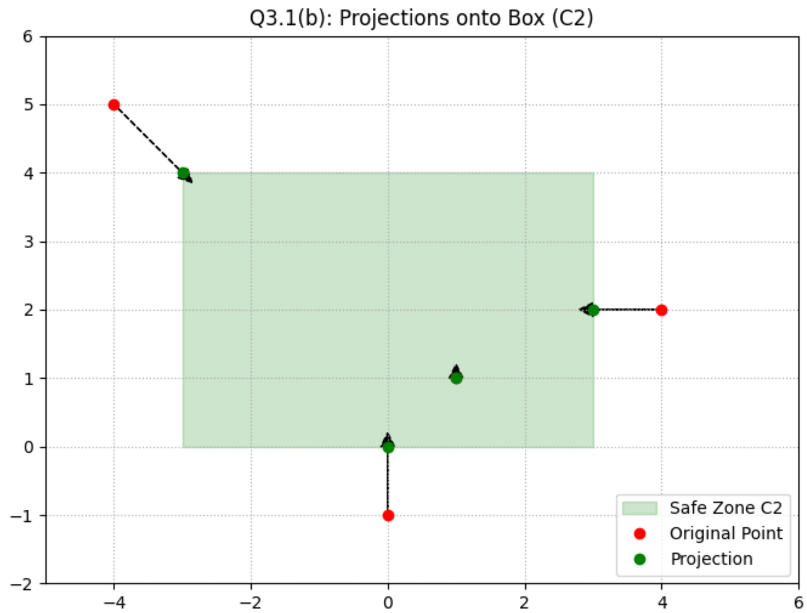

Figure 3: Projection of robot positions onto the rectangular safe zone, $\Pi_{C_2}(y)$.

3.2 Separating Hyperplane in a Classification Story.

A company has two groups of customers:

- **Group A:** Customers who always pay on time, represented as points inside the unit circle $C_A = \{x : \|x\|_2 \leq 1\}$,

- **Group B:** High-risk customers, all lying in the half-space $C_B = \{x : x_1 \geq 3\}$.

By the *Separating Hyperplane Theorem*, the company wants to find a hyperplane that separates C_A and C_B. To compute such a separating hyperplane (normal vector and offset).

The plot below illustrates the sets C_A and C_B, along with the separating hyperplane.

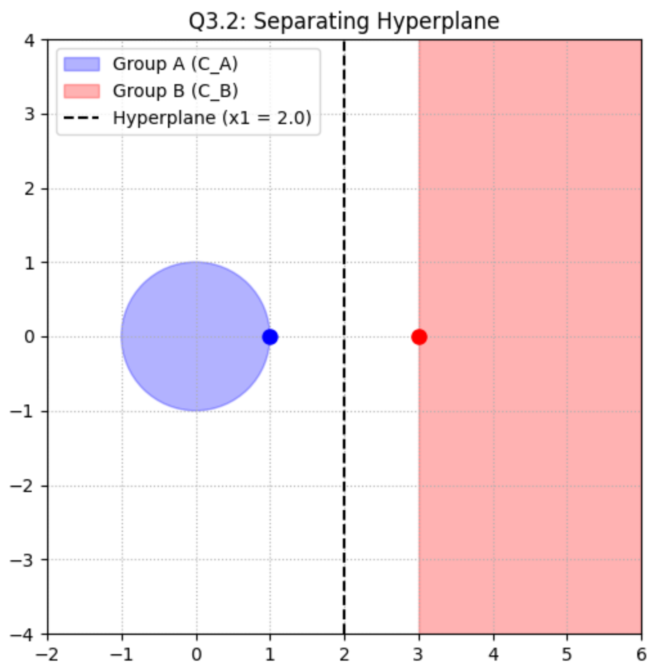

Figure 4: Separating hyperplane between C_A (safe customers) and C_B (high-risk customers).

The computed parameters for the hyperplane are as follows:

- **Normal vector (n):** $n = [1.0, 0.0]$
- **Offset (c):** $c = 2.0$
- **Closest point in C_A (a):** $a = [1.0, 0.0]$
- **Closest point in C_B (b):** $b = [3.0, 0.0]$

These values correspond to the separating hyperplane $x_1 = 2.0$, which is visualized in the accompanying plot.

3.3 Farkas Lemma in a Supply-Chain Model.
Suppose a factory must meet demand d using resources $x \in \mathbb{R}^2$ subject to capacity constraints $Ax \leq b$. Sometimes, no feasible plan exists. In that case, by Farkas Lemma, there exists a vector $y \geq 0$ certifying infeasibility.
Consider the system:

$$x_1 + x_2 \leq -1, \quad -x_1 \leq 0, \quad -x_2 \leq 0.$$

To check feasibility (using `CVXPY`). If infeasible, computed a Farkas certificate y.
We consider the given system and check feasibility using `CVXPY`.

The solver reports the system as " **infeasible** ".

The computed **Farkas certificate** is:

$$y = \begin{bmatrix} 1.00000006 \\ 1.00000006 \\ 1.00000006 \end{bmatrix},$$

we found out it satisfies:

$$A^\top y = 0, \quad b^\top y = -1.000000063699761 < 0.$$

Hence, by Farkas' Lemma, the system has no feasible solution since a nonnegative vector $y \geq 0$ exists such that $A^\top y = 0$ and $b^\top y < 0$.

Interpretation in the supply-chain context:

- The vector y represents a nonnegative combination of the capacity constraints that produces an impossible demand inequality ($b^\top y < 0$).
- This means no allocation of resources $x_1, x_2 \geq 0$ can meet the required demand d, confirming infeasibility.
- The certificate y thus acts as a formal proof of impossibility, showing that the current supply and capacity setup cannot satisfy demand under any feasible configuration.

Observation: The Farkas certificate successfully validates the infeasibility detected by the solver (`Status: infeasible`).

References

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

David G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, 1984.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of Applied Optimization. Kluwer Academic Publishers, 2004.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2 edition, 2006.