Problem Set 01 (Graded Portion)

Aman Bhargava — 1005189733

Problem 1.10 (Angles between Word Vectors)

Code can be found in `P1_10.mlx`. Rough work and strategy in `PS01_Jot_Notes`.

Part A: Euclidian Norm

The 7th and 8th titles are closest by Euclidian distance ("James Forder" and "Public image of George W. Bush") while the 9th and 10th titles ("Barack Obama" and "George W. Bush") are closest by Euclidian angle.

These are likely different pairs because the raw number of words in each article may differ drastically. Two articles may have much similarity in word choice, but if one is significantly longer (higher magnitude), their Euclidian distance would be unjustifiably large.

Part B: Normalization with Euclidian Norm

The 9th and 10th titles ("Barack Obama" and "George W. Bush") are closest both by Euclidian angle and distance when the vectors are normalized.

My answers match the initial angle calculation from the previous part. This is because all the vectors are now on the $||\tilde{v}|| = 1$ norm ball. Angle and Euclidian distance on such a ball are loosely interchangable metrics in that situation because $\text{distance} = 1 - \cos\theta$. On the interval $0 \leq \theta \leq \pi$, angle is a monotonically increasing function of distance. Therefore, when mapping between the two distance metrics (Euclidian distance and angle) when all points are on a norm ball, no points will change order. Therefore, it is impossible for there to be different pairs for the closest Euclidian distance and angle respectively.

Part C: Term Frequency Inverse Document Frequency Score

The two articles closest according to the TF-IDF score are the 9th and 10th articles ("Barack Obama" and "George W. Bush") by Euclidian distance and the 8th and 10th articles by Euclidian angle ("Public image of George W. Bush" and "George W. Bush").

Part D: TF-IDF Rationale

The inverse document frequency adjustment is to discount similarity resulting from the use of common words. For instance, a high similarity in the use of words such as 'the' or 'a' is less meaningful than high similarity in the use of words that are extremely rare (such as 'tetrahedron').

Geometrically, it is de-valuing (scaling down) the magnitude of each vector on axes representing extremely common words and scaling up the magnitude of vectors on axes representing extremely uncommon words.

Problem 1.11: $f_i : \mathbb{R}^2 \to \mathbb{R}, i \in [3]$

Code can be found in `P1_11.mlx`*. Rough work and strategy in* `PS01_Jot_Notes`.

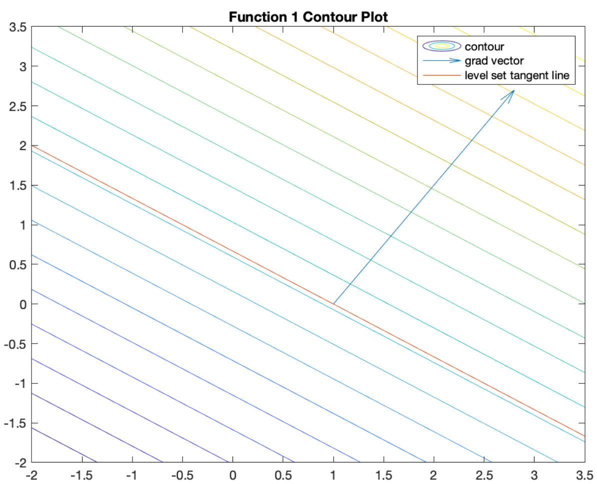
Part A: Closed Form Gradient Solution

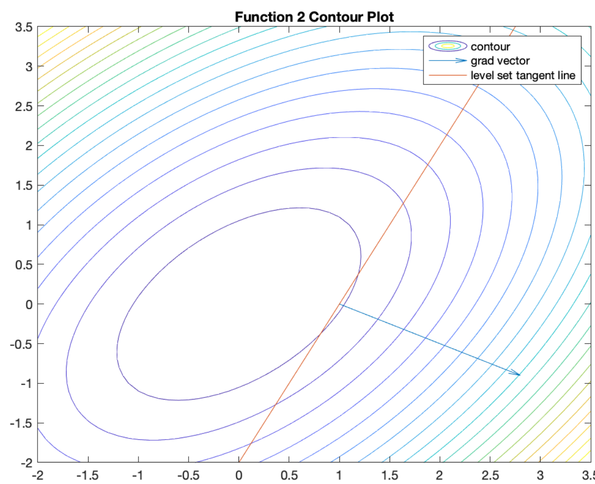
$\nabla f_1 = [2 \;\; 3]^T$

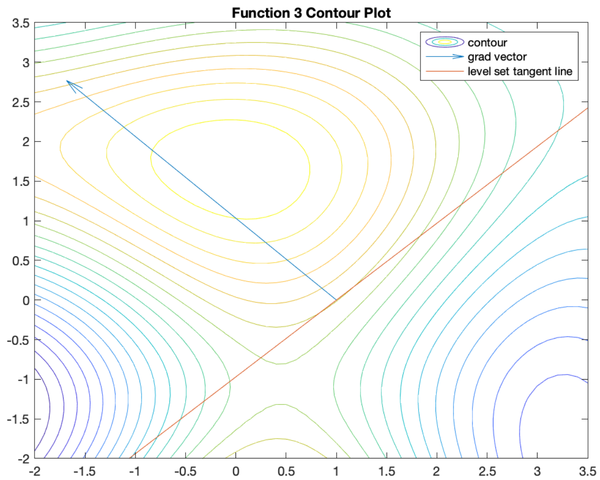
$\nabla f_2 = [2x - y \;\; 2y - x]^T$

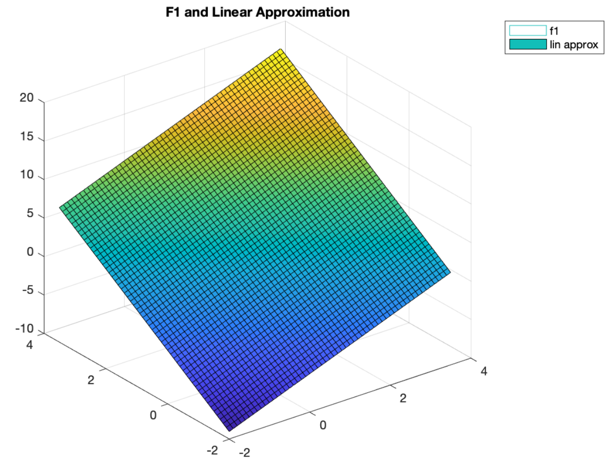
$\nabla f_3 = [\cos(y-5) + (5-y)\cos(x-5) \quad (5-x)\sin(y-5) - \sin(x-5)]^T$

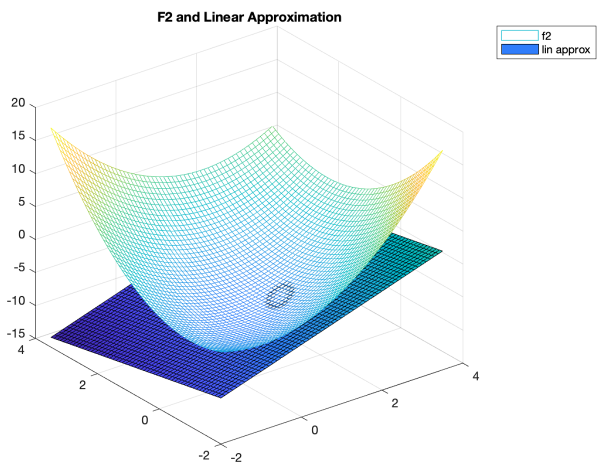
Part B & C: 2-D Contour Plots & 3-D Linear Approximation Plots

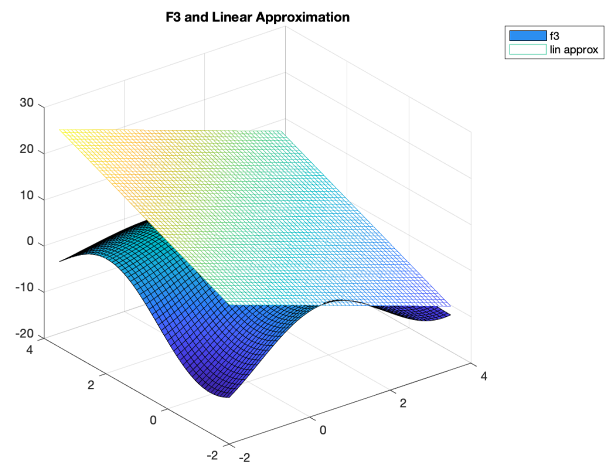








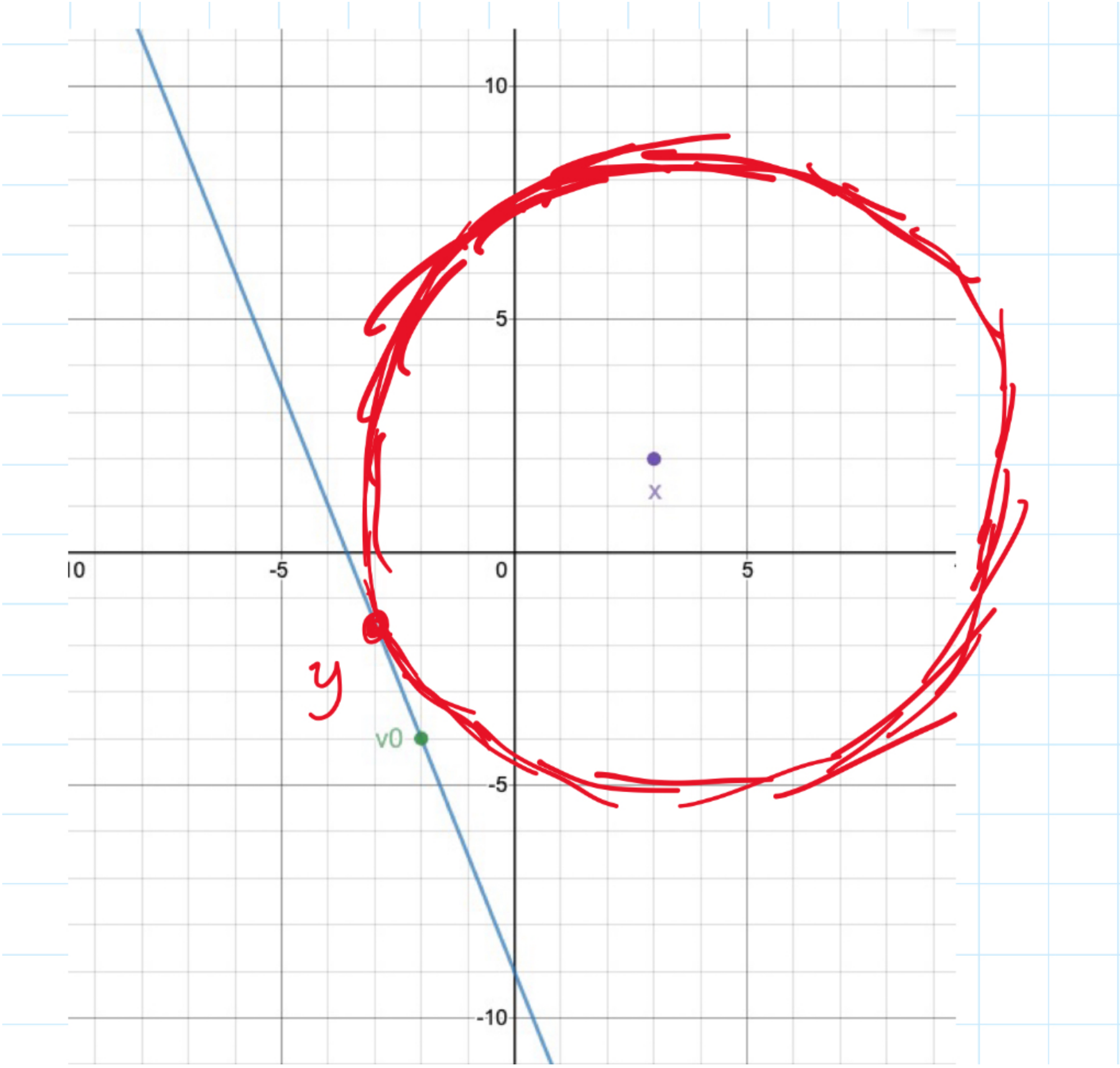




Problem 1.12: Projection with Different Norms

Code can be found in `P1_12.mlx`*. Rough work and strategy in* `PS01_Jot_Notes`.

Part A: Sketching Euclidian Projection

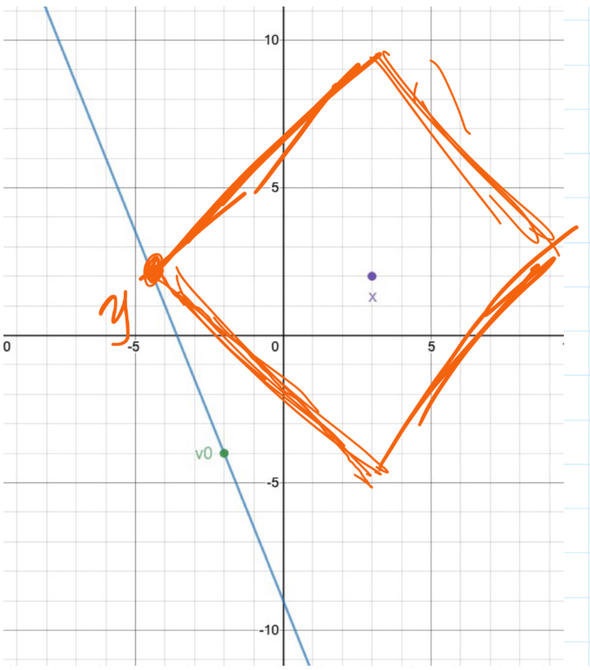


Part B: Closed Form Solution for Euclidian Projection

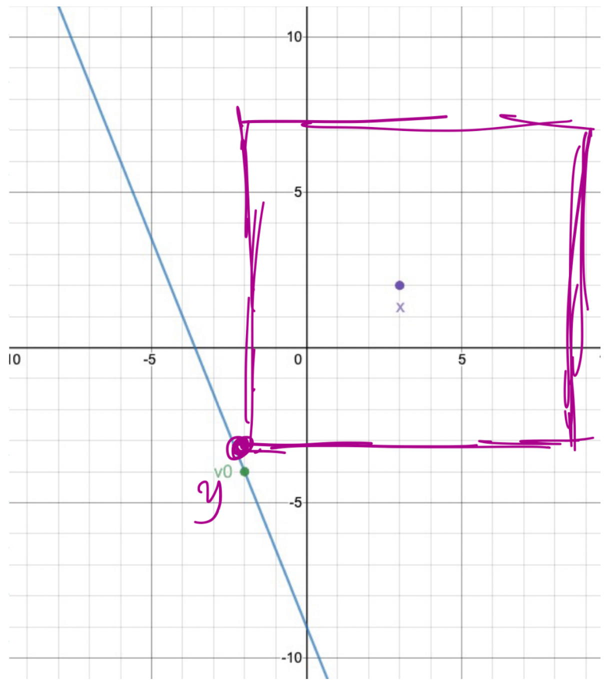
1. We first subtract the offset vector v_0 from y to create $y_{tmp} = y - v_0 = [5 \;\; 6]^T$.
2. Our set of vectors is $S = \{v - v_0\} = \{[-2 \;\; 5]^T\}$. The 'matrix' of inner products in this case is $M = [29]$.
3. Our vector of inner products $\langle v_i, x \rangle$ for each vector in S is $[20]$.

4. The inverse of M is $M_{inv} = [0.034482]$.
5. Solving for the coefficient vector $\alpha = [0.689655]$.
6. Therefore: $y_{proj} = \sum \alpha_i S_i + v_0 = [-3.379 \quad -0.55172]^T$.

Part C: Repeat A for l_1, l_∞



l_1 Norm



l_∞ Norm

Part D: Use `cvx` to Verify Results from B

See code in `P1_12.mlx`.

```
Projected vector y under l2:
   -3.3793
   -0.5517
```

The result from CVX matches with the result from the above calculation.

Part E: Use Function to Solve for l_1, l_∞

See code in `P1_12.mlx`.

```
Projected vector y under l1:
   -4.4000
    2.0000
```

```
Projected vector y under l_infinity:
   -2.2857
   -3.2857
```