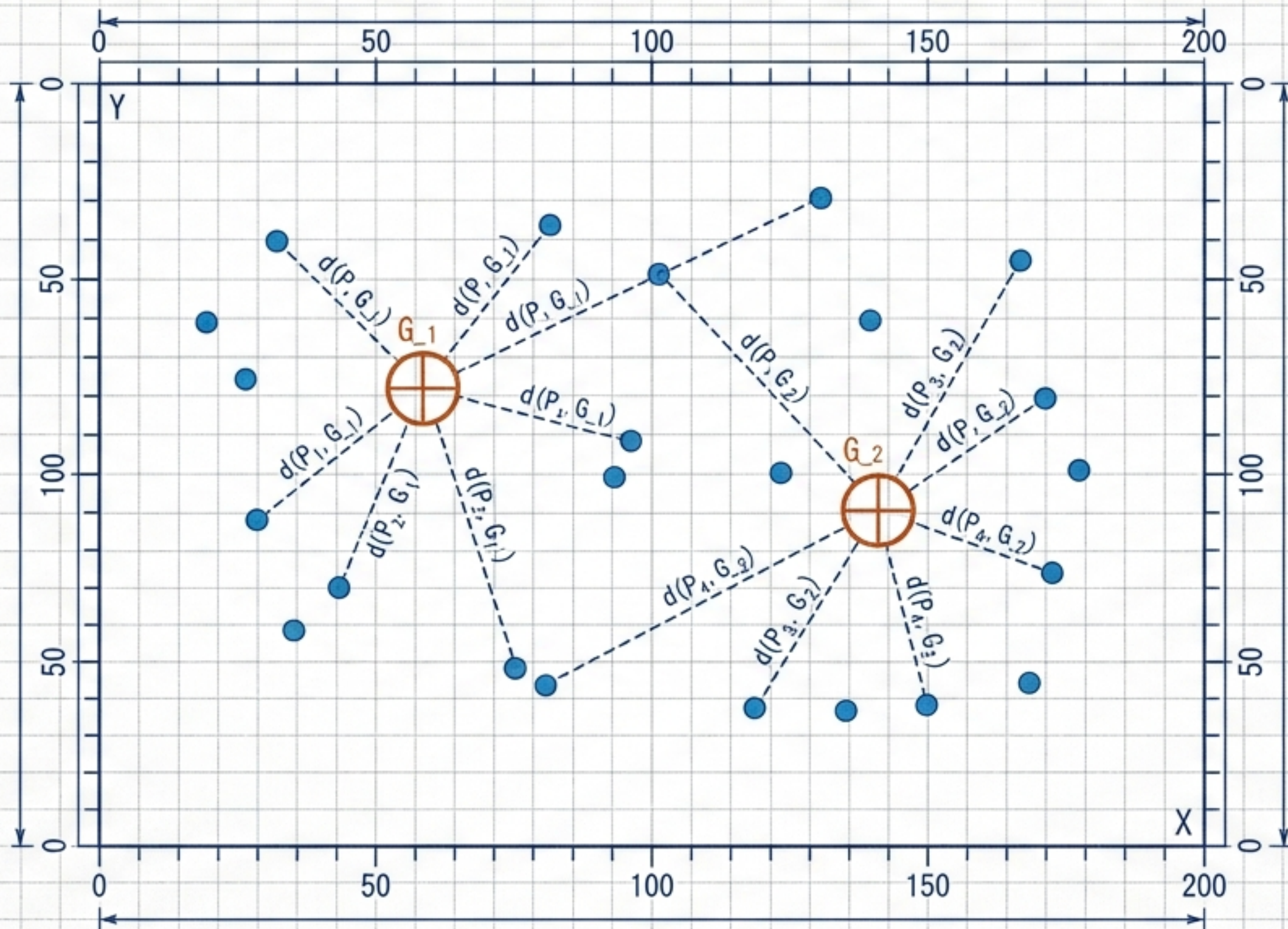


内部ロジックの視覚的解読：k-means法と初期値最適化

応用情報技術者試験（令和3年春期 午後問3）のコード実装と確率的アプローチを「設計図」として読み解く。解答は提示せず、自ら導き出すための視点を提供する。



アルゴリズムの3フェーズと本資料の探索領域

Step 1: 初期化 (Initialization)

ランダム、または確率的に
初期コアを空間に配置する。



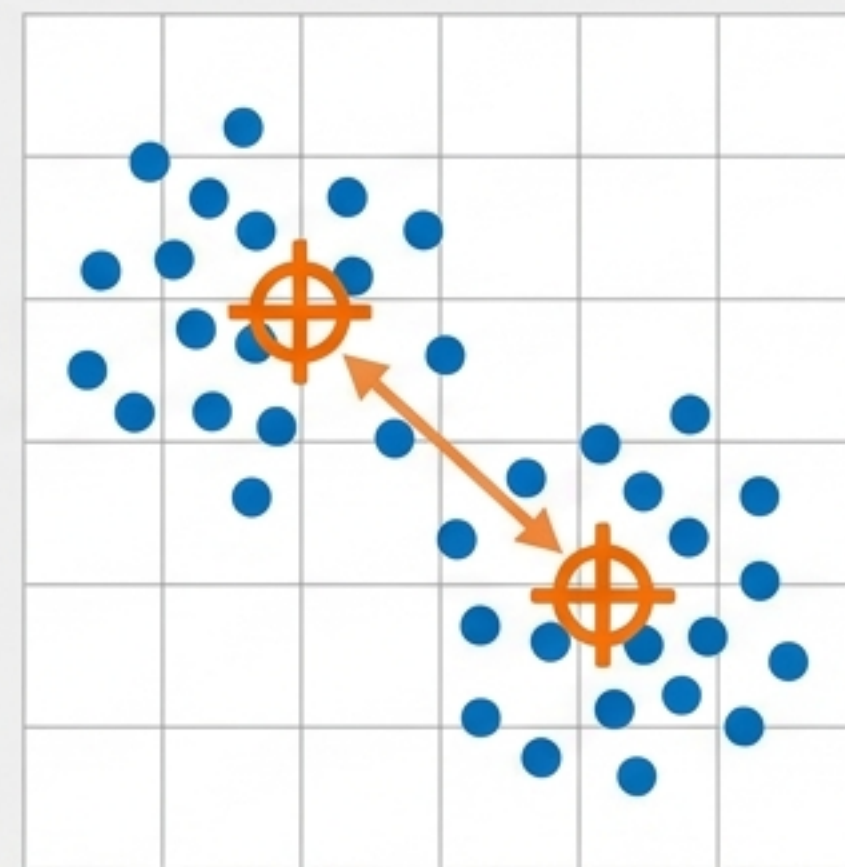
Step 2: 所属の更新 (Assignment)

各データ点が、最も近い重
心のクラスタに所属する。



Step 3: 重心の更新 (Update)

新しい所属データに基づき、
重心の座標を再計算する。



対象コードの解剖：所属クラスタを見直す反復処理

pseudocode

```
for ( s を 1 から N まで 1 ずつ増やす )  
  for ( t を 1 から K まで 1 ずつ増やす )  
    grav_length[t] <- grav_dist(s, t)  
  endfor  
  [ オ ]  
endfor
```

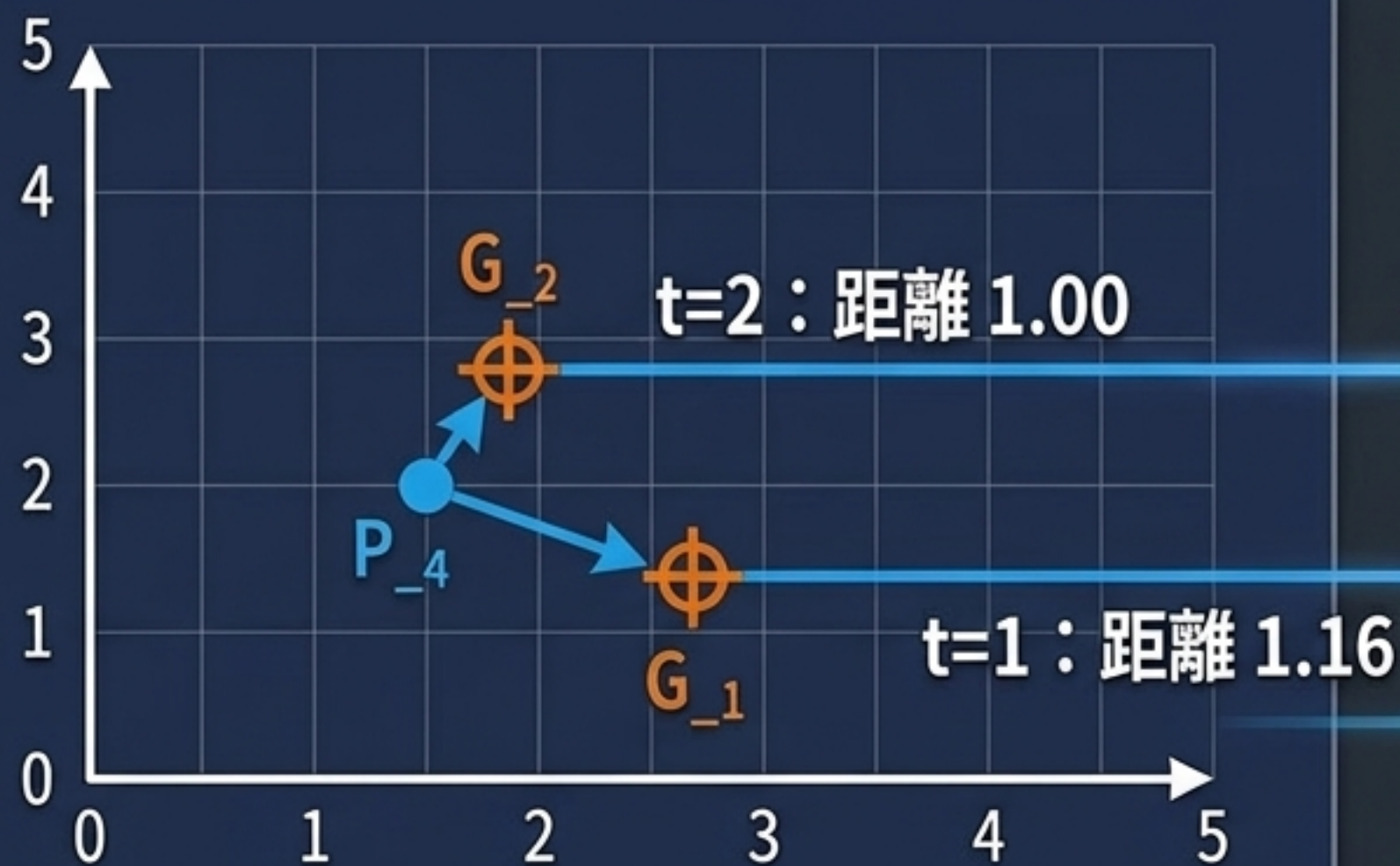
分析の焦点

内側のループ（変数t）が終了した直後、プログラムは点sに関する『ある重要なデータ』を全て集め終わっている。

次にすべき行動は何か？

ミクロな視点：ループ内で実行されている「測距」プロセス

座標空間 (Physical Space)



メモリ空間 (Memory Space)

grav_length

1.16

1.00

[1]

[2]

結論：内側ループ終了時、配列 `grav_length` には『点 s から全重心への距離リスト』が完成している。

空欄[オ]のミッション：距離リストから所属を決定せよ

ピースA (関数):
`min_index()`

配列の中で、最小値が
格納されている『添字』
を返す。

点sの距離リスト

grav_length

1.16

1.00

ピースB (配列):
`cluster[s]`

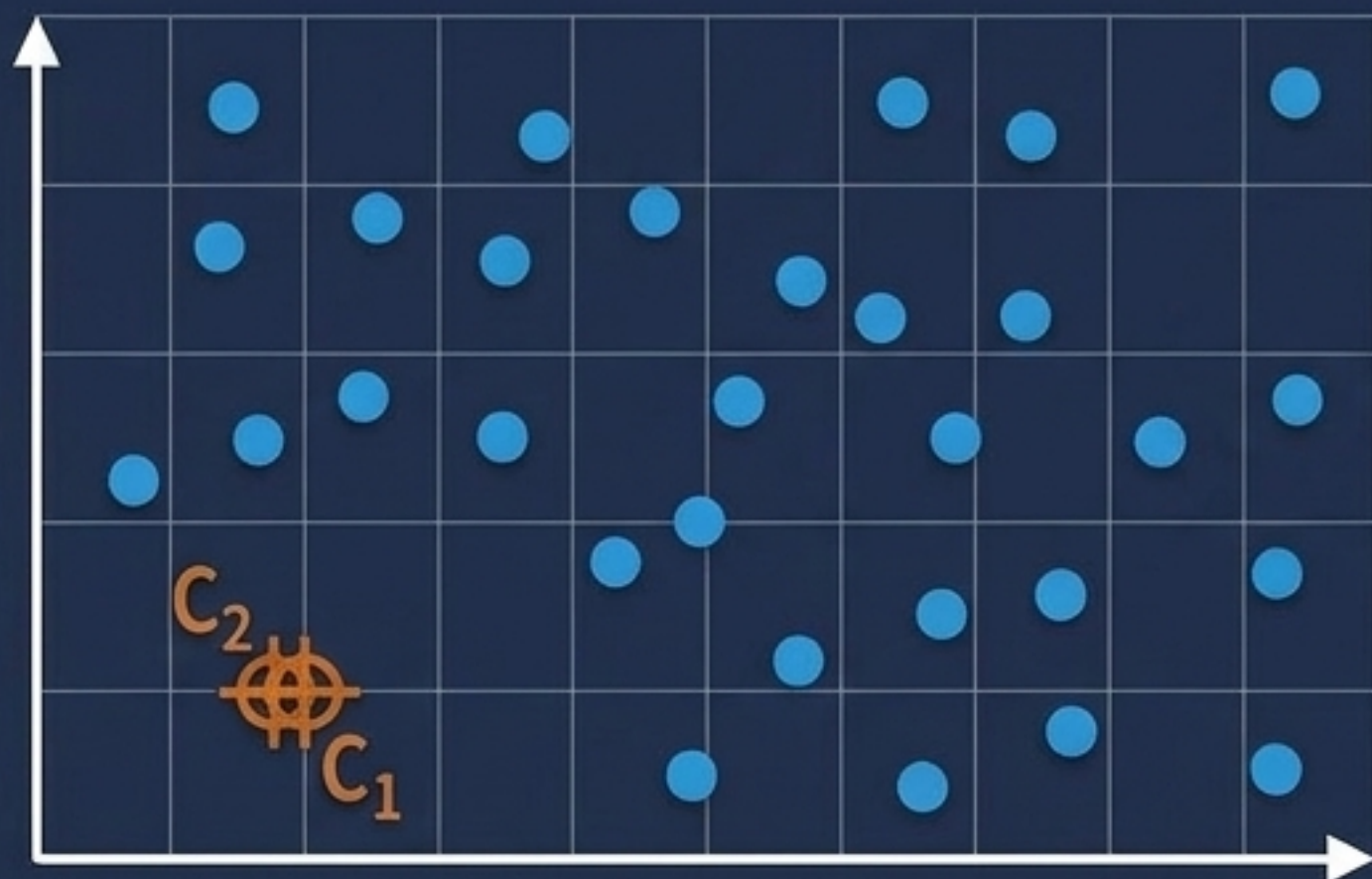
点sが所属する
『クラス番号』を
格納する。

思考プロセス：

1. 「最も近い重心」を選ぶには、距離リストから何を探すべきか？
2. 見つけた「重心の番号」を、点sのデータとしてどこに記録すべきか？

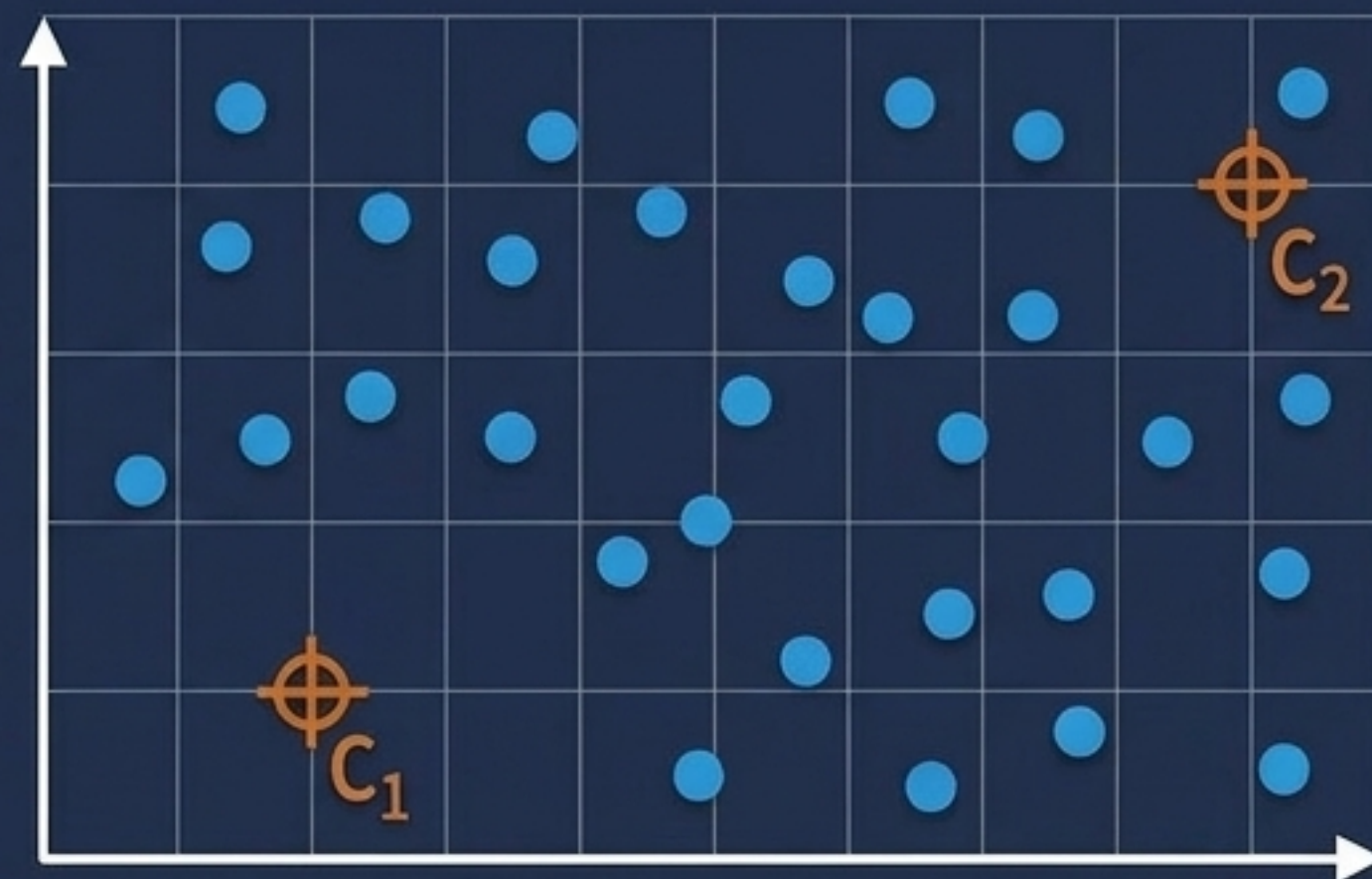
マクロな欠陥：なぜ「ランダムな初期化」は危険なのか？

ランダムな悲劇 (Bad Initialization)



問題点：コアが近すぎると、空間の分割に偏りが生じ、収束に膨大な時間がかかる。あるいは不自然な境界線で誤った局所解に陥る。

理想的な配置 (Good Initialization)

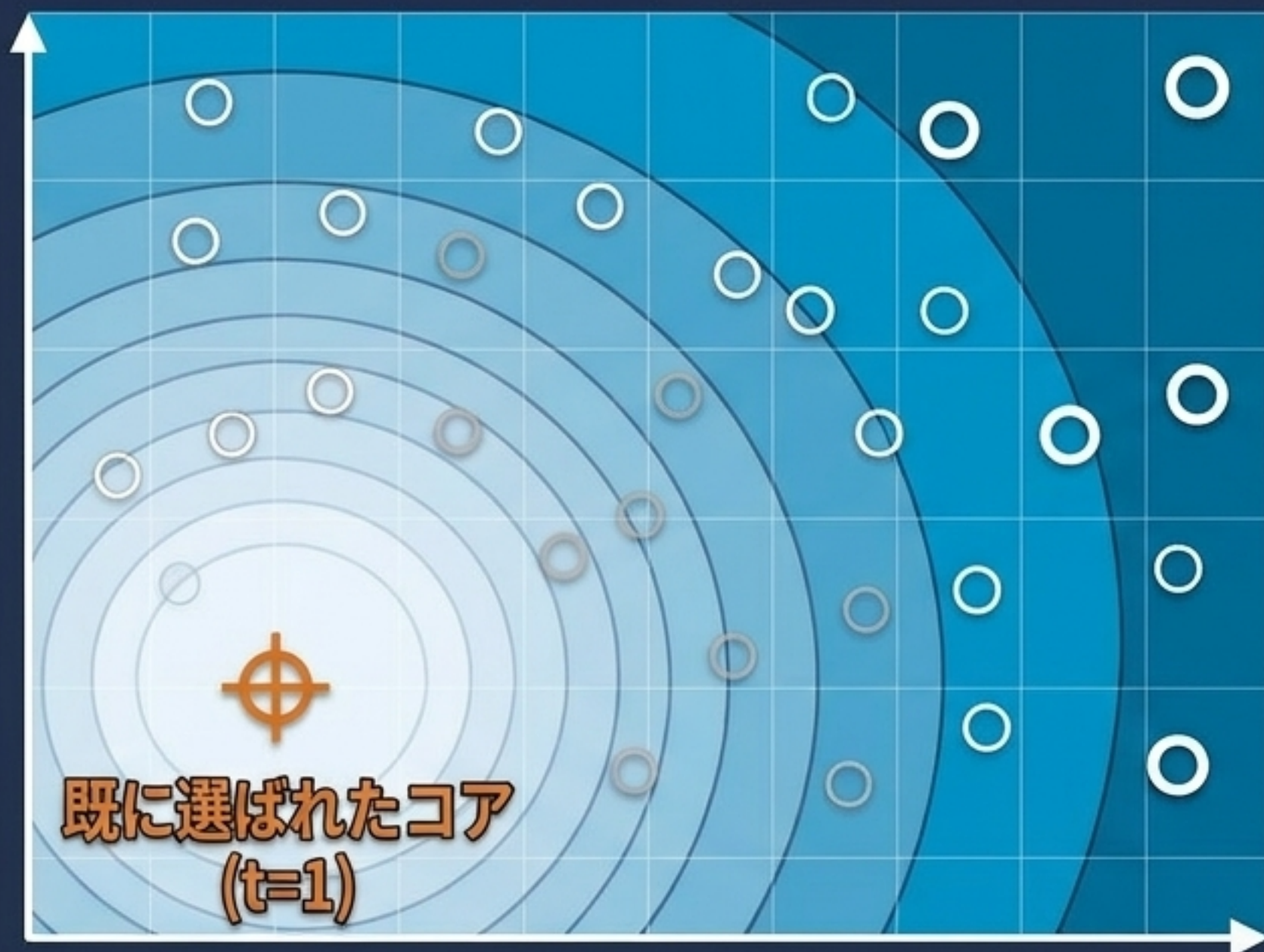


利点：最初から空間を大きく分割できるため、少ない計算回数で自然な分類が完成する。アルゴリズムが安定する。

改良戦略：既存のコアから「遠い点」ほど選ばれやすくする

k-means++の基本思想

- 1つ目のコアはランダムに選択する。
- 2つ目以降のコアは、すでに決まったコアから『遠く離れている点』ほど、高いウェイト（確率）を与えてルーレットを回す。



距離の総合スコア『T_s』の視覚化



計算式：

$$T_s = d_1 + d_2 + d_3$$

定義：

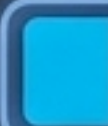
T_s とは、特定の候補点 P_s が、既存のコア集団全体から『トータルでどれくらい離れているか』を示す指標である。

空欄[カ]のロジック：選ばれるべきはどちらの点か？

ケースA：既存コアの近くの点

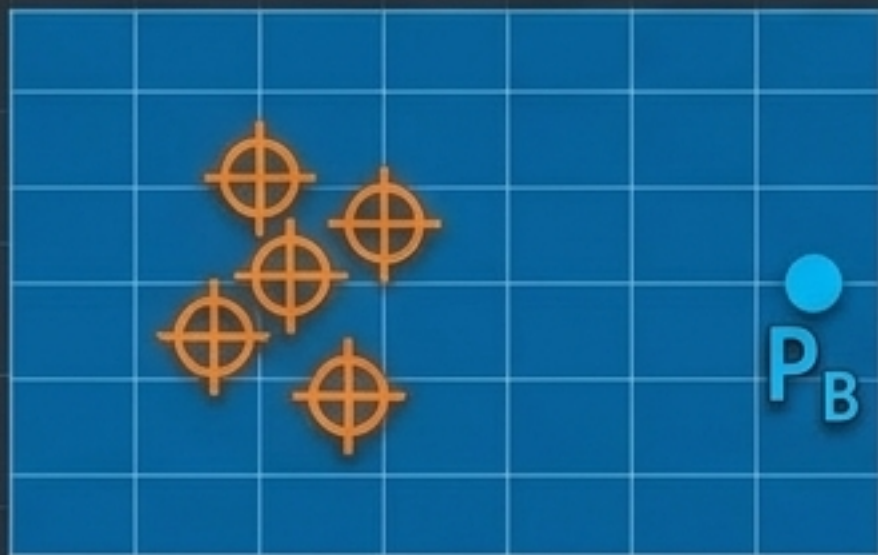


距離の和 T_A

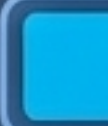


10%

ケースB：既存コアから遠く離れた孤立点



距離の和 T_B



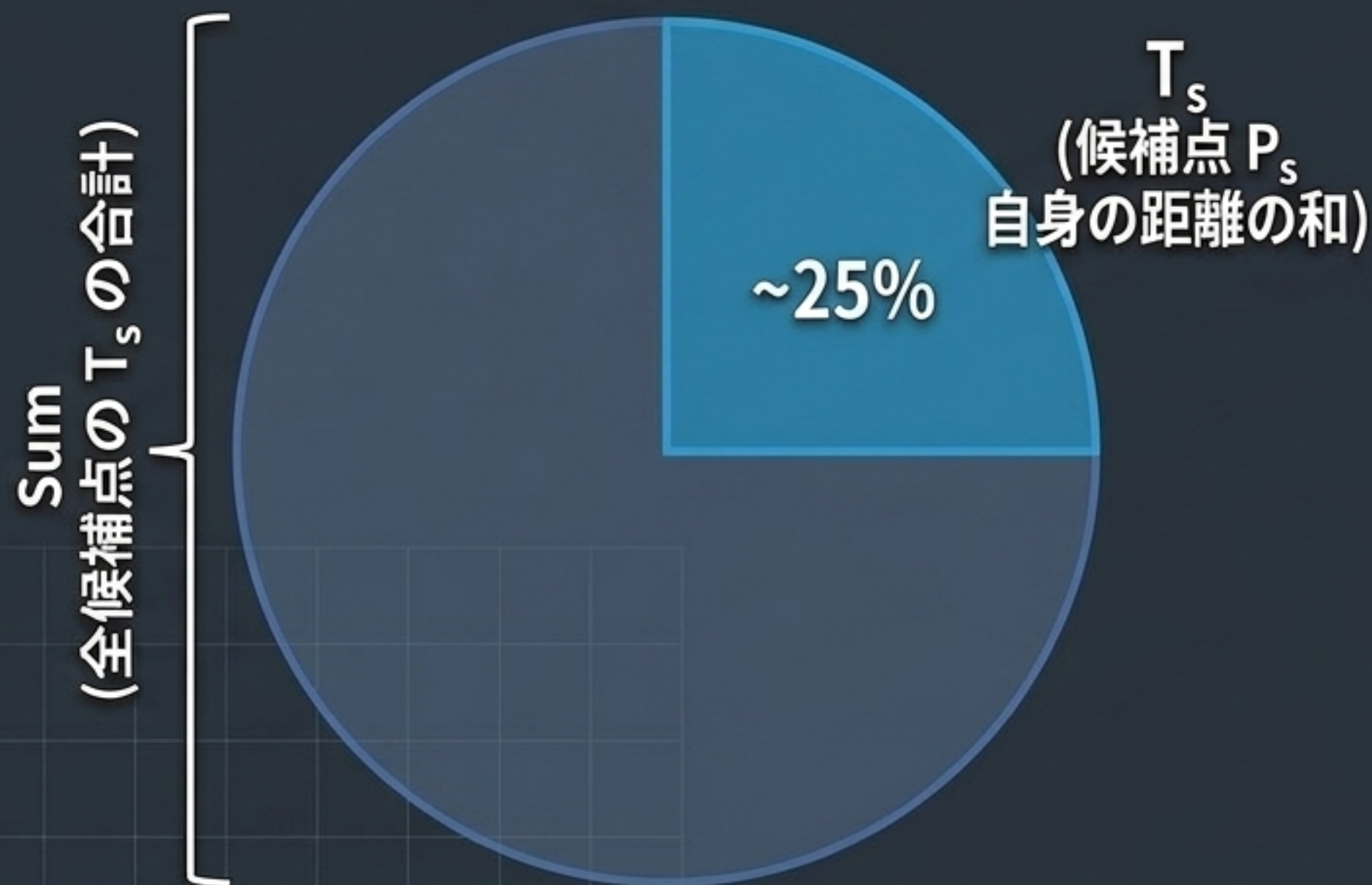
90%

設問の条件と目的：
改良アルゴリズムの目的は
『遠い点を優先する』こと。

では、点 P_s が新しいコアとして選ばれる確率を高く設定すべきなのは、 T_s の値がどのような状態のときか？

空欄[キ]のロジック：確率を数式（ルーレット）に変換する

確率ルーレット



確率の数式構築プロセス：

- 基本原則：事象が起こる確率 = (自分の取り分) ÷ (全体の総量)
- 自分の取り分 = 候補点 P_s の距離の和
- 全体の総量 = 全ての候補点の距離の和の合計

結論：ある点 P_s が選ばれる確率は、部分と全体の変数をどう組み合わせれば表現できるか？

統合 (Synthesis) : 最適化されたアルゴリズムの真の姿

設問2 : ミクロな確実性

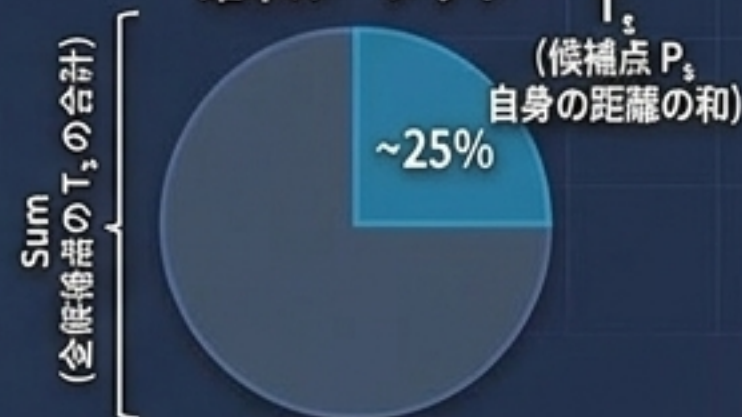
min_index のアレ例



厳密な距離計算による
妥協のない所属判定

設問3 : マクロな確率論

確率ルーレット



距離に比例した確率ウェイト
による大局的な初期配置

堅牢な k-means++ 法の完成

コードの羅列は、空間を制御するための設計図に過ぎない。
背後にある空間的・数学的な意図を視覚化したとき、空欄の答えは必然として浮かび上がる。