SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of AI &DS

Course Name – 19AD602 DEEP LEARNING

III Year / VI Semester

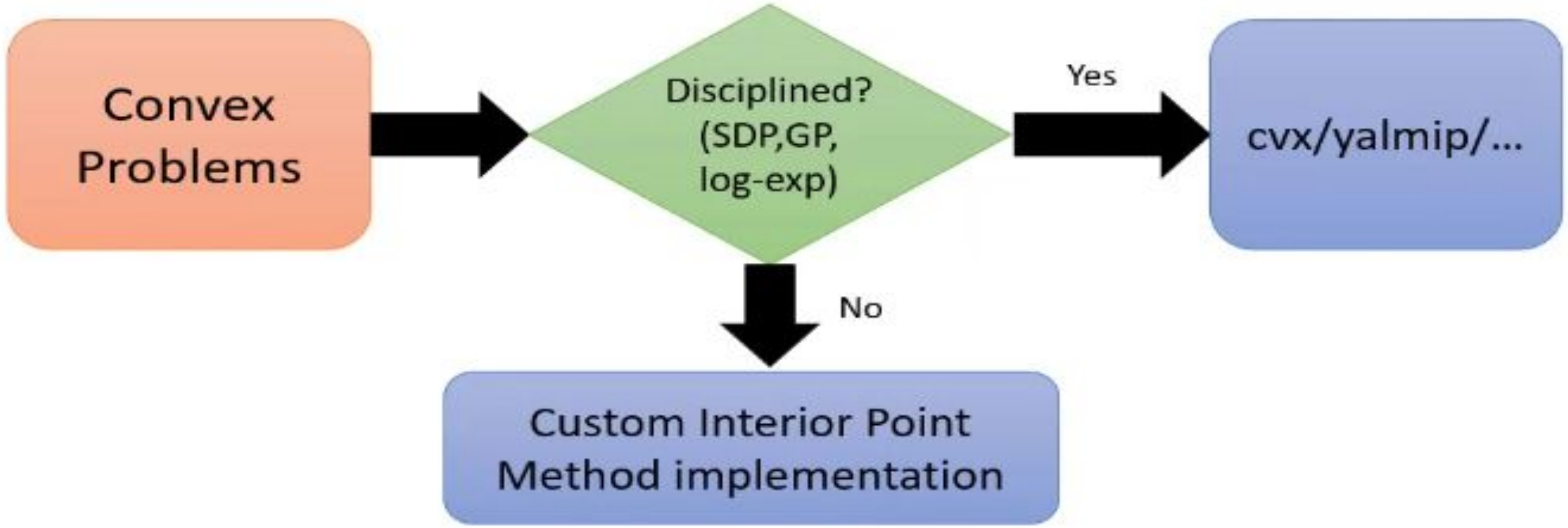
UNIT-4 OPTIMIZATION AND GENERALIZATION

Topic: Non-convex optimization for deep networks AND Stochastic Gradient Descent

CASE STUDY:

A company trains a deep learning model for image recognition without optimizations. The model takes 10 hours to train, achieves only 70% accuracy, and suffers from overfitting, leading to poor generalization on new data.

Convex Optimization



Non-convex Problems: Exact Solution

- Discover "hidden" convexity
  - Via introduction of variables, elimination, epigraph, Schur's complement, etc.
  - Example non-convex QCQP with special structure may turn out to be GP (Problem 9, Assignment 5)

- Prove strong duality despite the problem being non-convex
  - Solve the dual (which is always convex problem)
  - Recover the primal
  - Eg: QCQP with single inequality (See: Boyd's book, Appendix B)

Non-convex Problems: Exact Solution

- Quasiconvexity or parameterizable in single variable
  - Eg: Minimization of quasiconvex function
  - Solving a series of convex problems

Non-convex Problems: Approximate Solution

- Relaxation
  - l0 norm to l1 norm,
  - rank to nuclear norm,
  - SDP relaxation
- Block coordinate Descent for Block Convex Problems
- Successive Convex Approximation/Majorization Minimization

Stochastic Gradient Descent (SGD) is a variant of the Gradient Descent algorithm that is used for optimizing machine learning models. It addresses the computational inefficiency of traditional Gradient Descent methods when dealing with large datasets in machine learning projects.

In SGD, instead of using the entire dataset for each iteration, only a single random training example (or a small batch) is selected to calculate the gradient and update the model parameters. This random selection introduces randomness into the optimization process, hence the term "stochastic" in stochastic Gradient Descent

The advantage of using SGD is its computational efficiency, especially when dealing with large datasets. By using a single example or a small batch, the computational cost per iteration is significantly reduced compared to traditional Gradient Descent methods that require processing the entire dataset.

Stochastic Gradient Descent Algorithm

- Initialization: Randomly initialize the parameters of the model.
- Set Parameters: Determine the number of iterations and the learning rate (alpha) for updating the parameters.
- Stochastic Gradient Descent Loop: Repeat the following steps until the model converges or reaches the maximum number of iterations:
  - Shuffle the training dataset to introduce randomness.
  - Iterate over each training example (or a small batch) in the shuffled order.
  - Compute the gradient of the cost function with respect to the model parameters using the current training example (or batch).
  - Update the model parameters by taking a step in the direction of the negative gradient, scaled by the learning rate.
  - Evaluate the convergence criteria, such as the difference in the cost function between iterations of the gradient.
- Return Optimized Parameters: Once the convergence criteria are met or the maximum number of iterations is reached, return the optimized model parameters.

In SGD, since only one sample from the dataset is chosen at random for each iteration, the path taken by the algorithm to reach the minima is usually noisier than your typical Gradient Descent algorithm. But that doesn't matter all that much because the path taken by the algorithm does not matter, as long as we reach the minimum and with a significantly shorter training time.

thank you