3011979 Intro to Deep Learning for Medical Imaging

L10 extra: Impact of learning rate on neural network model training

Apr 9th, 2021

Sira Sriswasdi, Ph.D.
Research Affairs, Faculty of Medicine
Chulalongkorn University

Gradient descent in multi-dimension

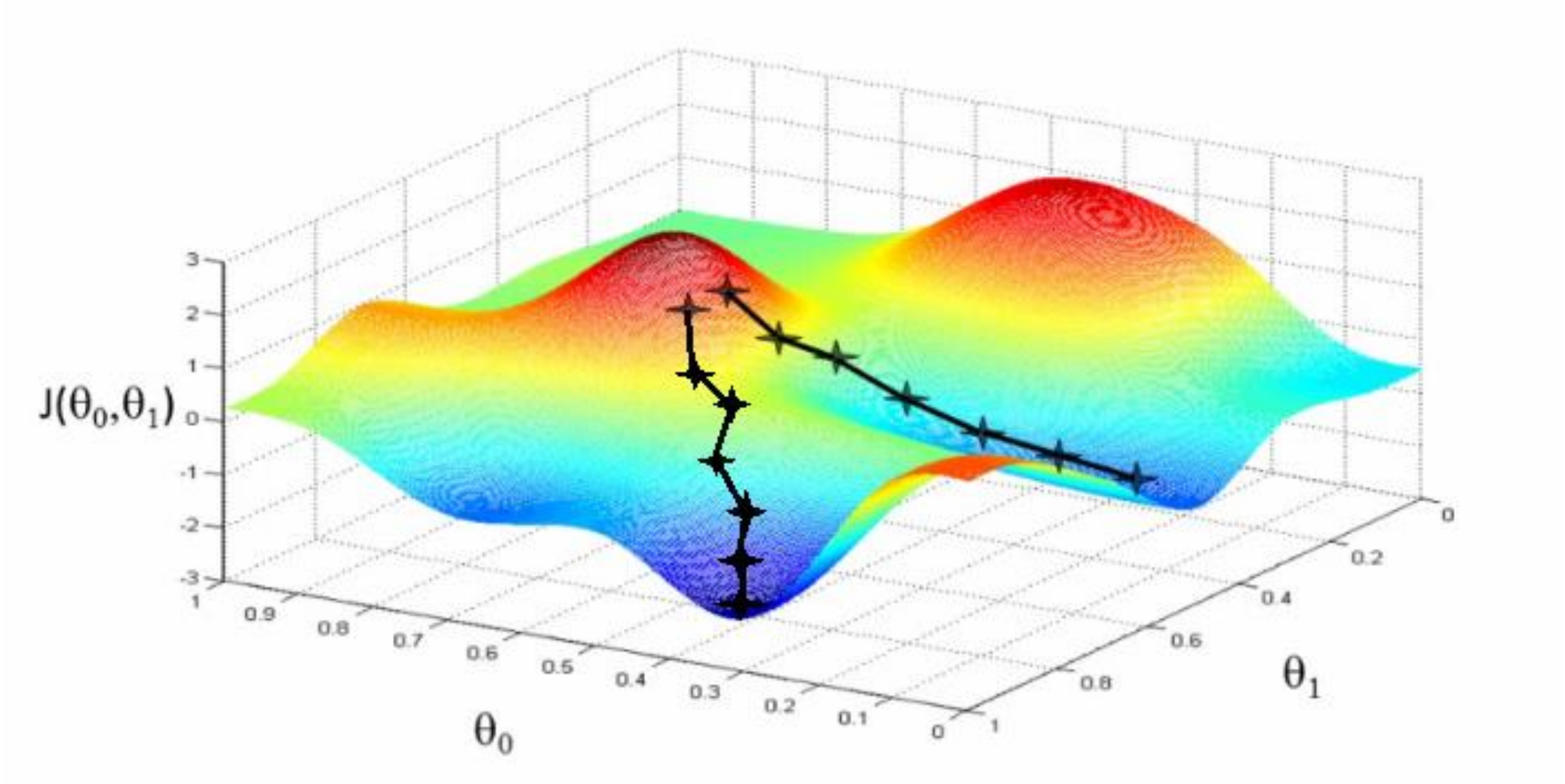
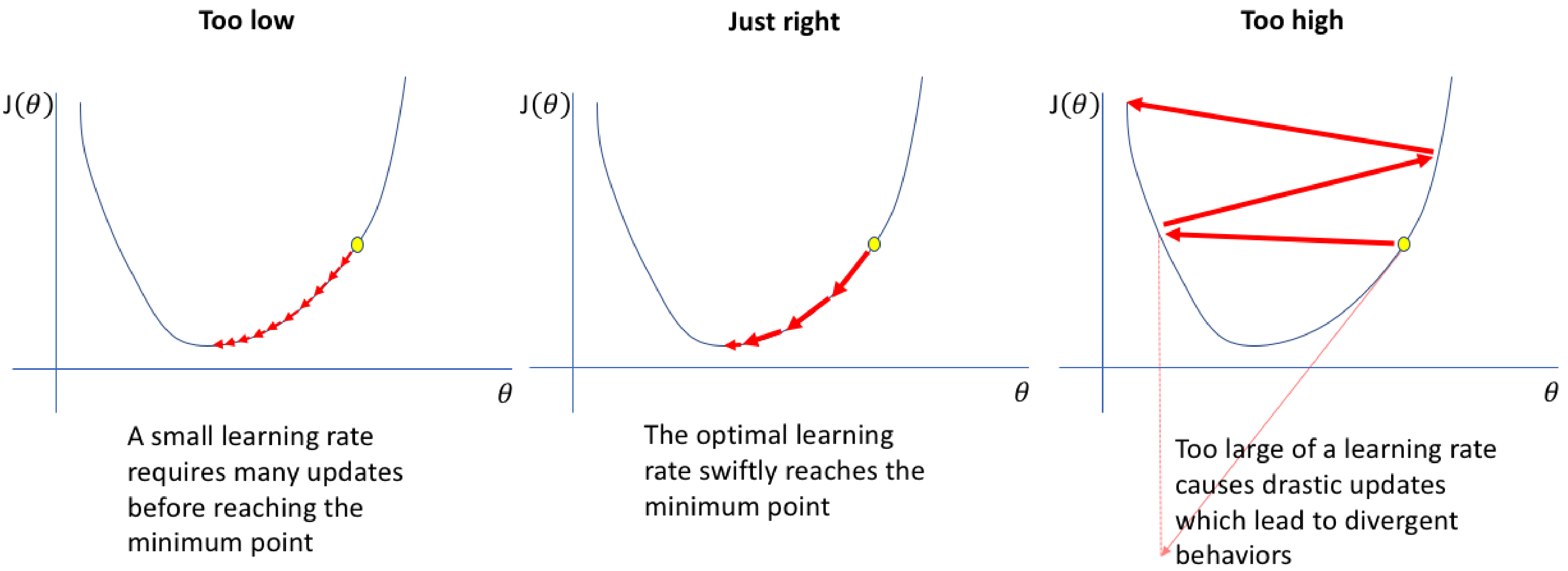


Image from shashank-ojha.github.io

- Starting position can determine which local minima to model converges to
  - When training artificial neural network model, we want to try several random initial weights

Learning rate is a key parameter



Too low

A small learning rate requires many updates before reaching the minimum point

Just right

The optimal learning rate swiftly reaches the minimum point

Too high

Too large of a learning rate causes drastic updates which lead to divergent behaviors

Image from jeremyjordan.me

- We want the model to update in big steps (large learning rate) in the beginning and slow down (small learning rate) once it approaches a local optima
  - Use **ReducedLROnPlateau** callback in Keras/Tensorflow
  - Set **learning_rate** = “adaptive” or “invscaling” in scikit-learn

Loss trend tells you a lot

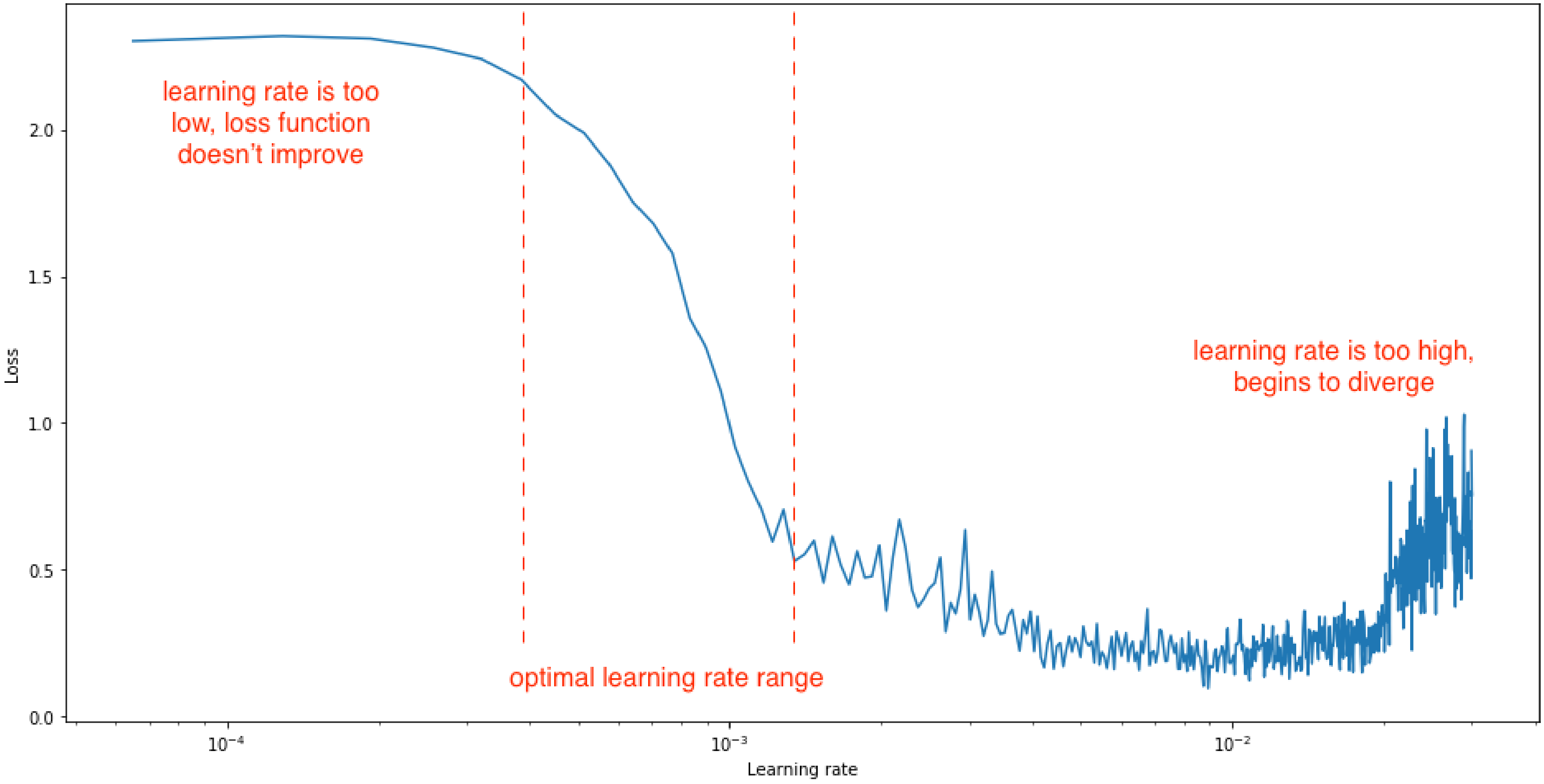


Image from jeremyjordan.me

Learning rate diagnosis through loss trend

- **Large + constant learning rate** = the model will not converge because the weights always change in big step
  - Look for big jumps in training and validation loss trends

- **Small + constant learning rate** = the model can take a very long time to converge
  - Look for slow or no change in training loss trend
  - Ok for small model and if you can wait

- **Large + adaptive learning rate** = best training strategy, the model will make big updates in the beginning and then slow down later
  - Look for big drop in training loss in early epochs and smooth training loss trend overall

An ideal loss trend

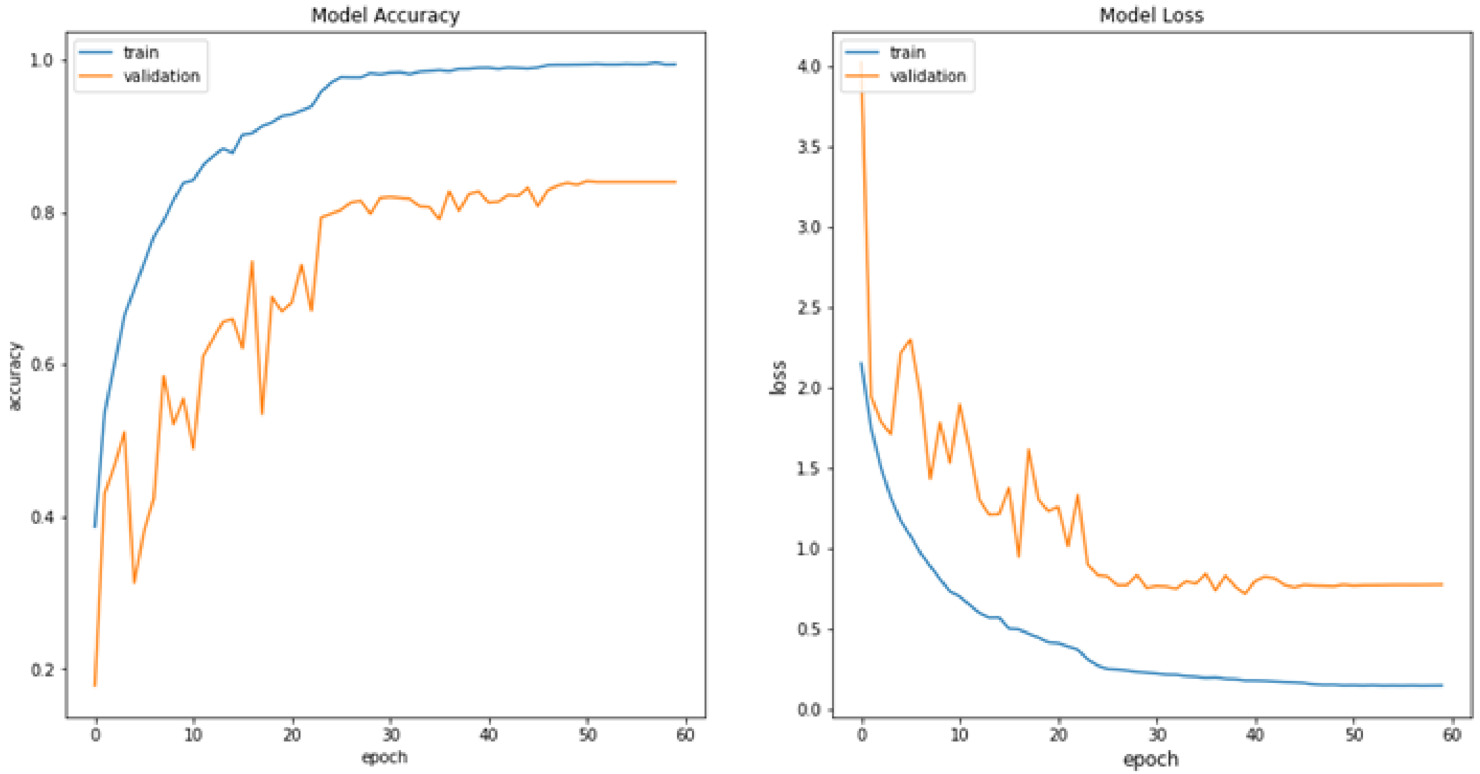


Image from AI Wiki

- Training loss is smooth throughout
- Validation loss makes big changes in the beginning, but the overall trend of loss reduction is clear

A problematic loss trend

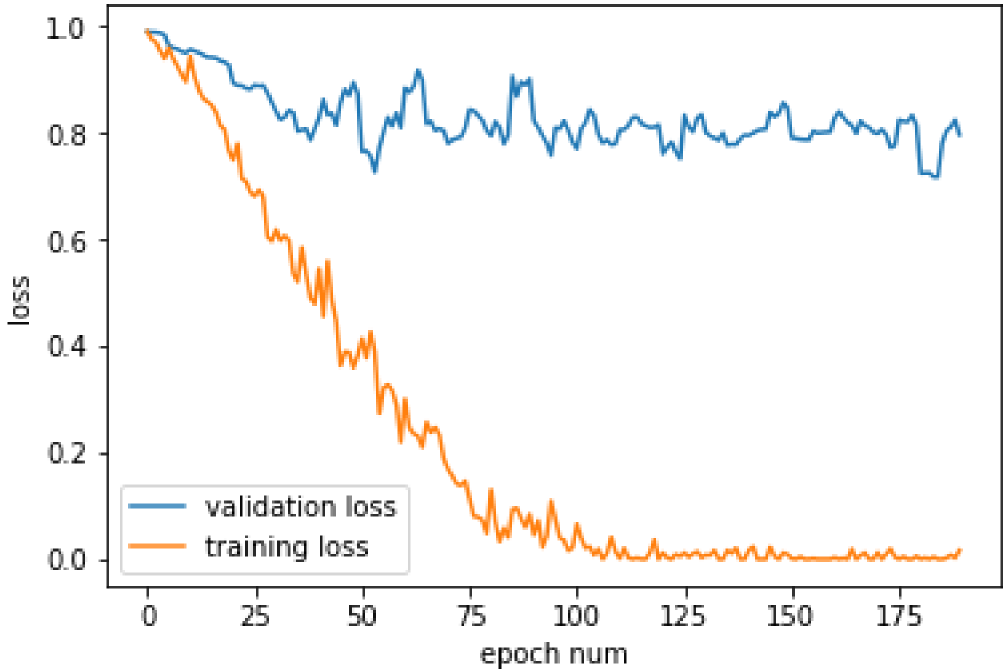


Image from datascience.stackexchange.com

- The model does not improve on validation set
  - Likely overfitting → reduce model complexity
- Validation loss oscillates
  - Validation set does not resemble training set
  - A sign that there is not enough data → collect more

Advanced technique – cyclical learning rate

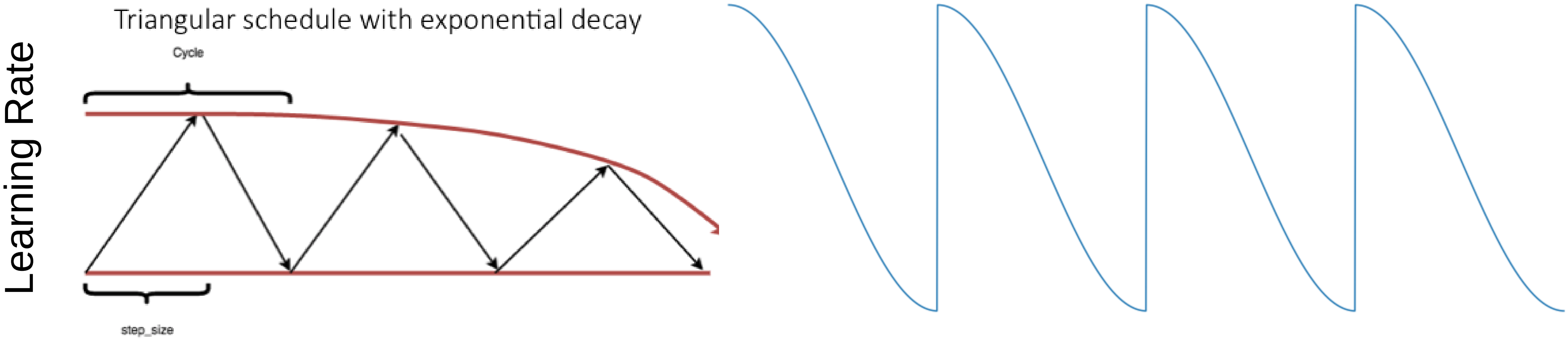


Image from jeremyjordan.me

- Cycles of initial large learning rate + adaptive decay
- Allow the model to jump out of a local optima to explore other regions of the loss surface
- Keep track of the best models with **ModelCheckpoint** in Keras/Tensorflow
  - Can combine multiple models into an ensemble like RandomForest

Cyclical learning rate behavior

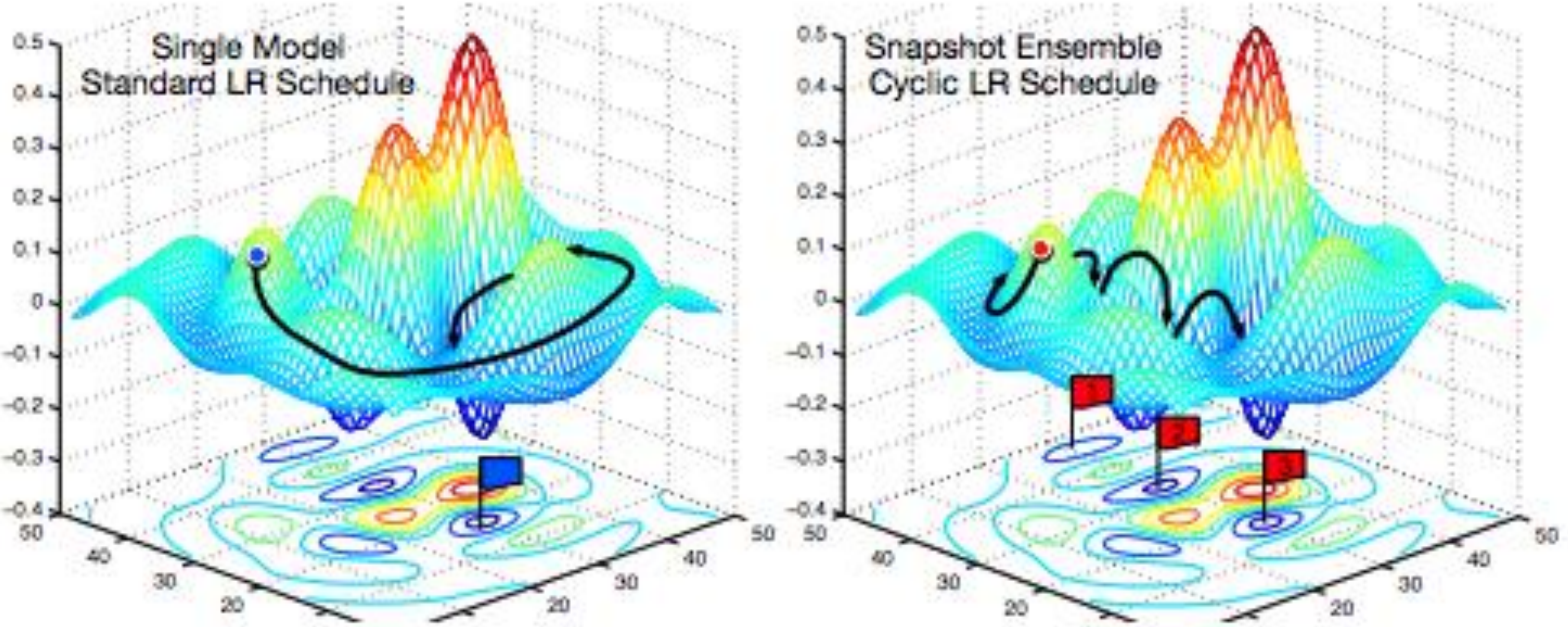


Figure 1: **Left:** Illustration of SGD optimization with a typical learning rate schedule. The model converges to a minimum at the end of training. **Right:** Illustration of Snapshot Ensembling. The model undergoes several learning rate annealing cycles, converging to and escaping from multiple local minima. We take a snapshot at each minimum for test-time ensembling.

- Model can visit multiple local optima