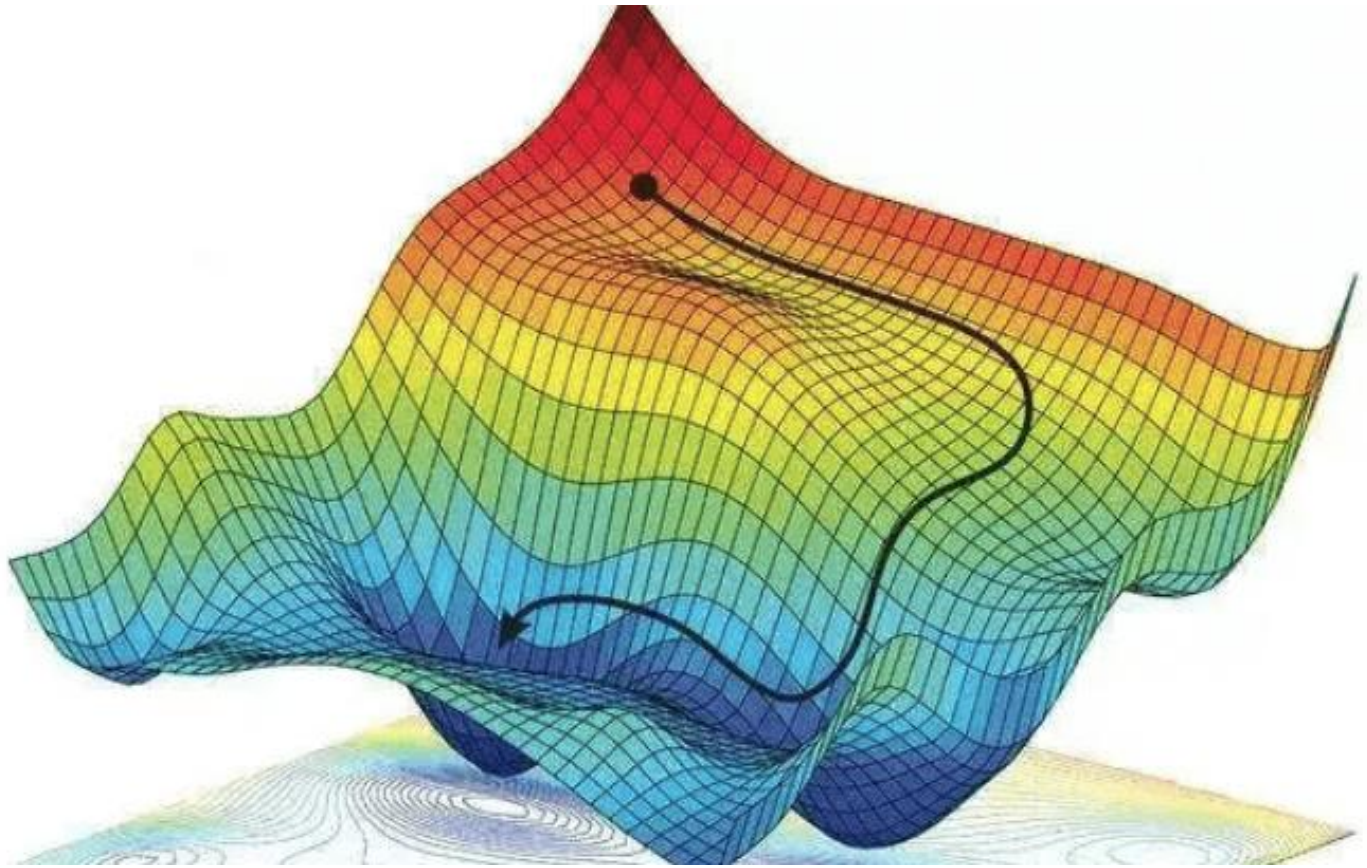
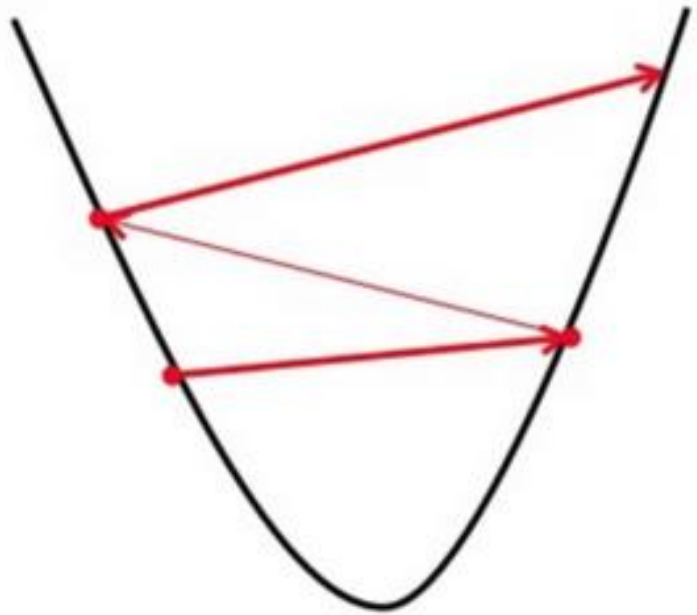
Some basics of Machine Learning

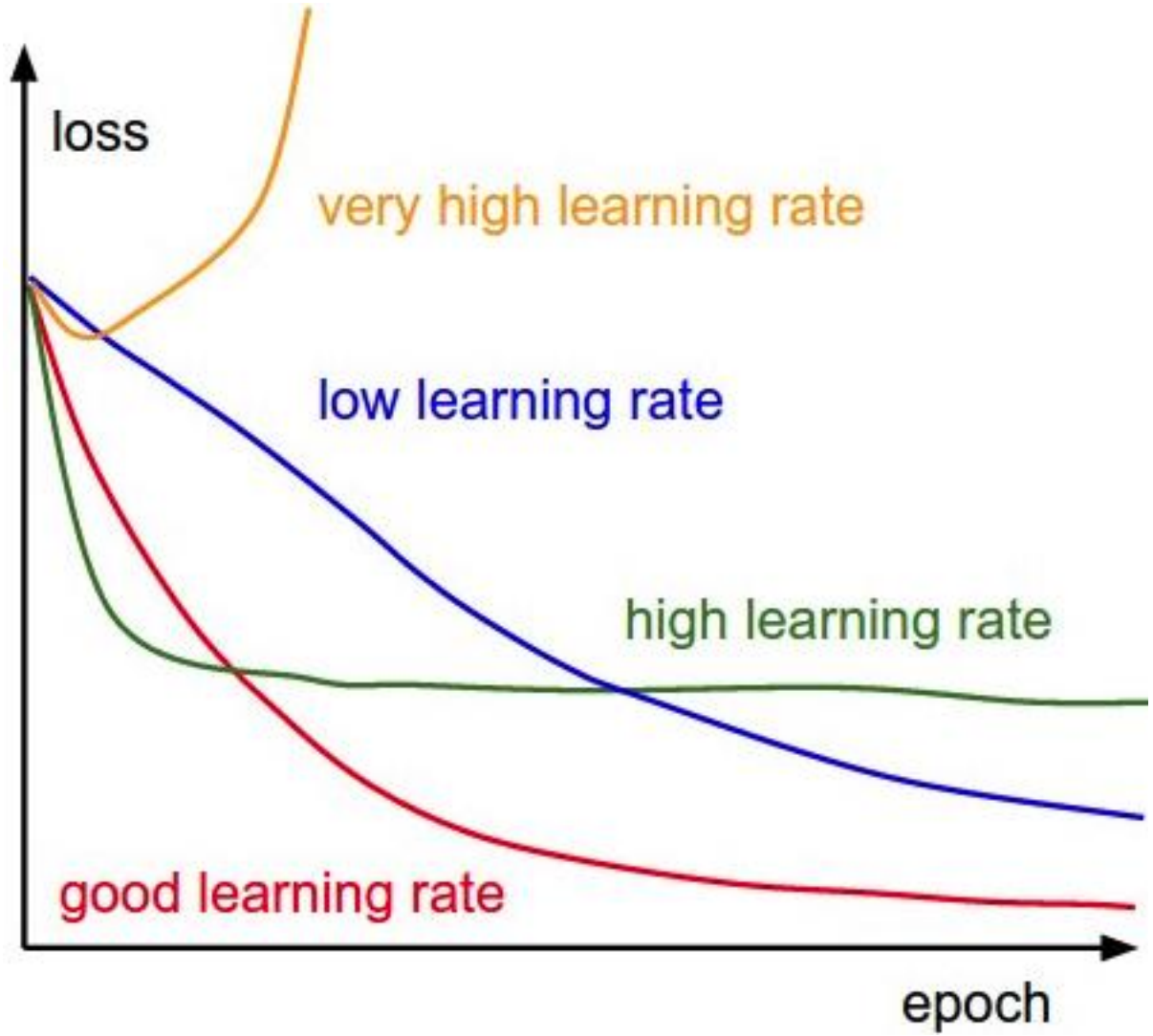
With a Focus on the importance of Hyperparameters

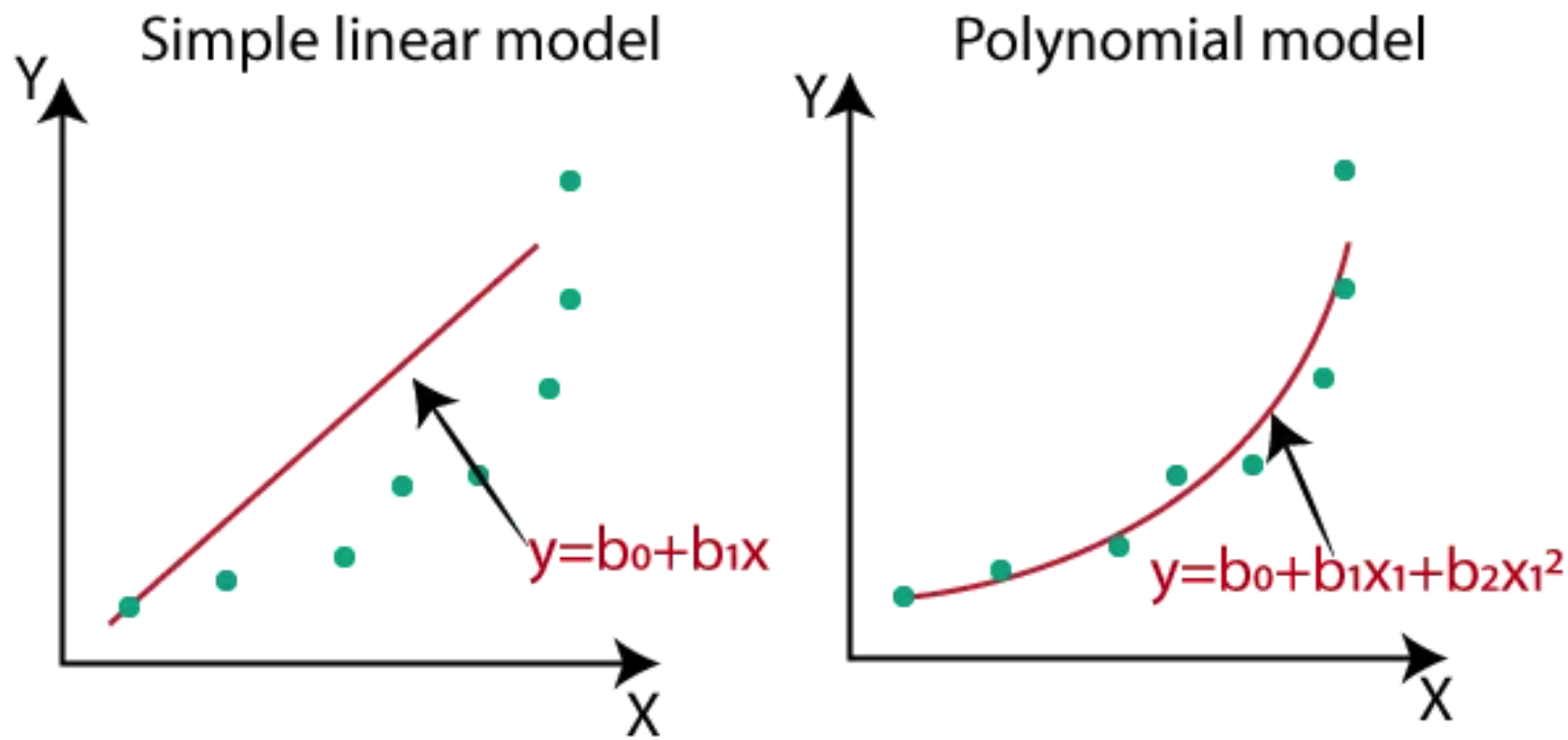


Big learning rate

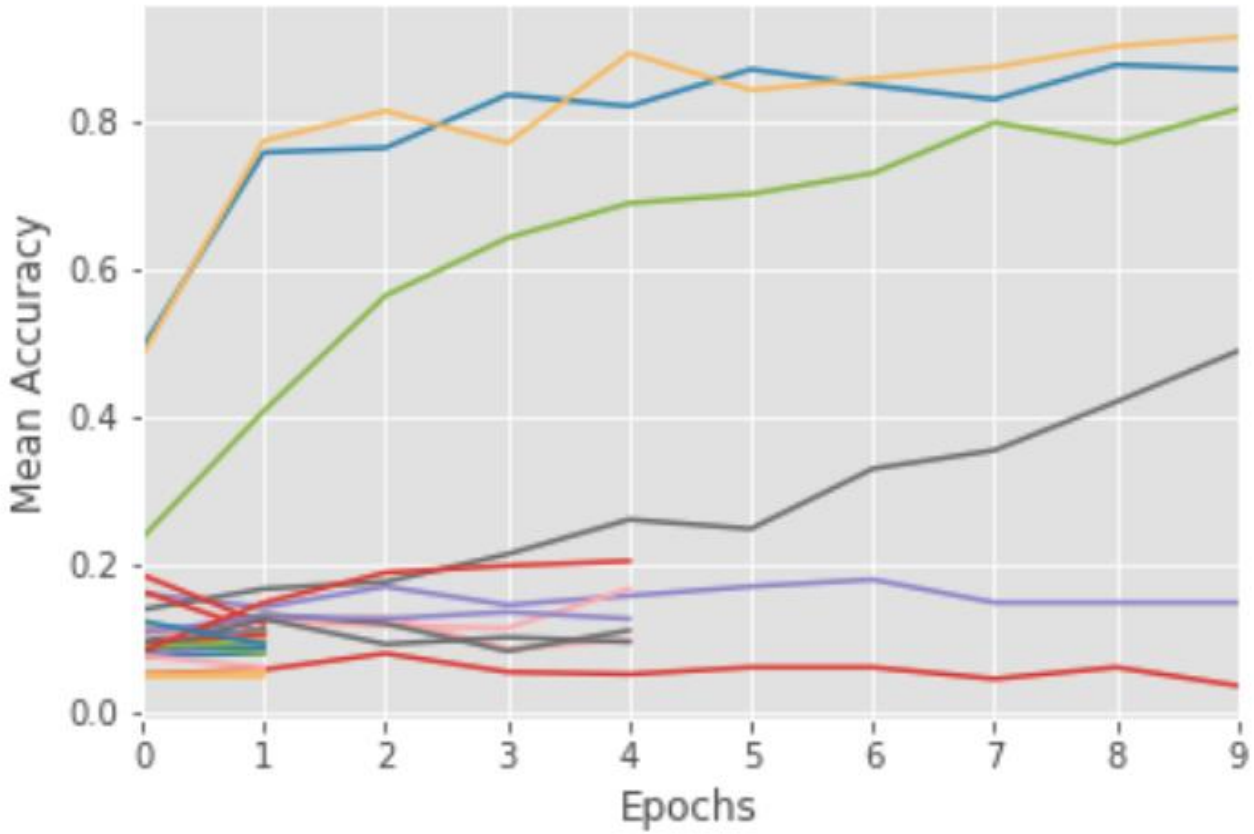


Small learning rate

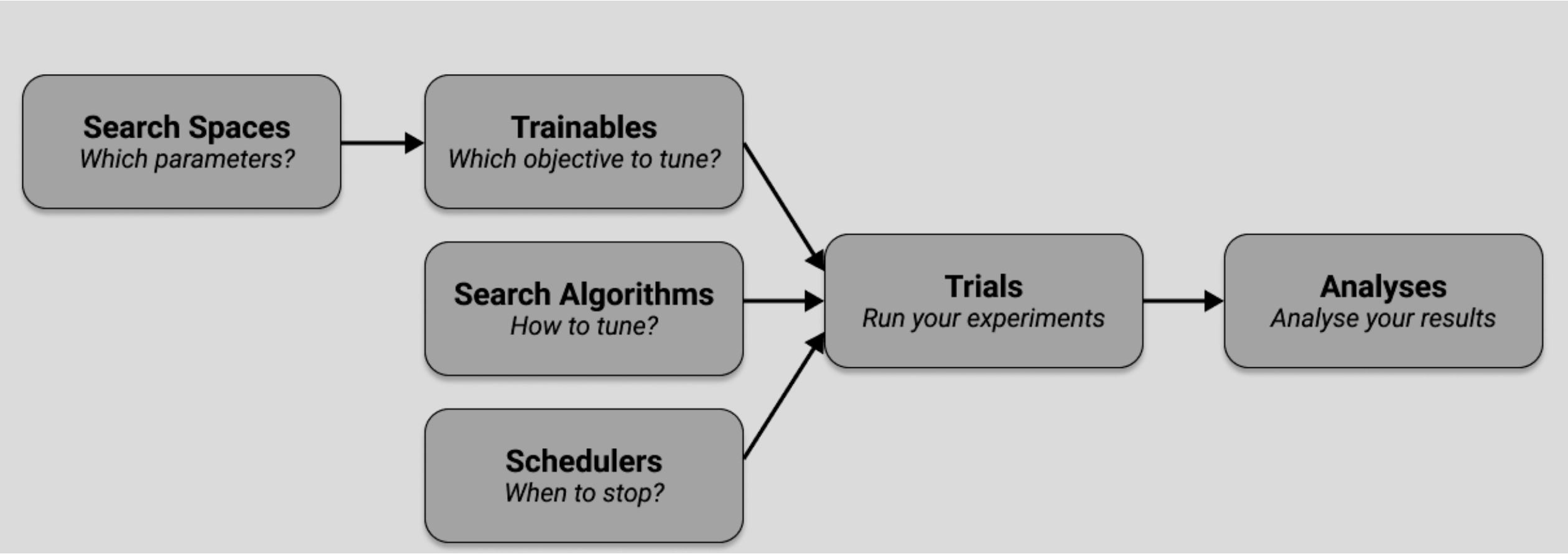

loss
very high learning rate
low learning rate
high learning rate
good learning rate
epoch


Simple linear model
Y
X
y=b0+b1x
Polynomial model
Y
X
y=b0+b1x1+b2x1²

Hyperparameter	Approximate sensitivity
Learning rate	High
Optimizer choice	Low
Other optimizer params (e.g., Adam beta1)	Low
Batch size	Low
Weight initialization	Medium
Loss function	High
Model depth	Medium
Layer size	High
Layer params (e.g., kernel size)	Medium
Weight of regularization	Medium
Nonlinearity	Low

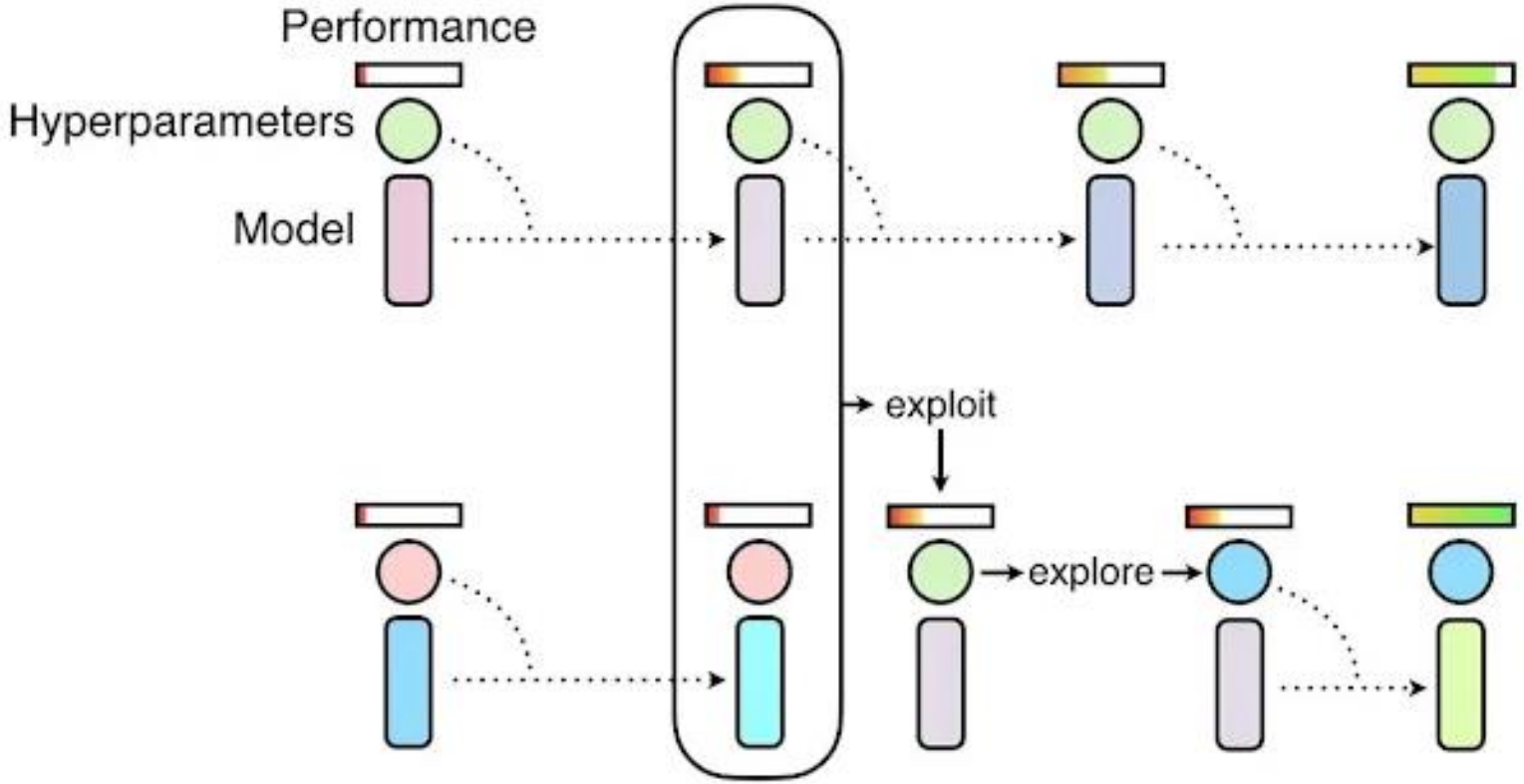

Mean Accuracy
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
6
7
8
9
Epochs

Key Concepts of Ray Tune



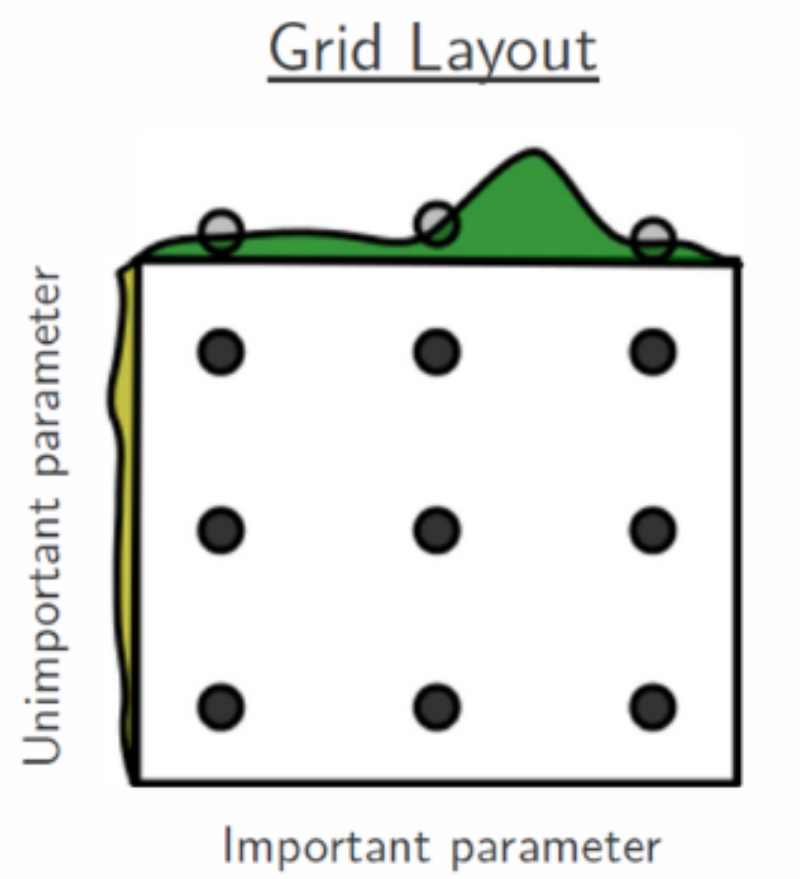
Example: Population Based Trainer

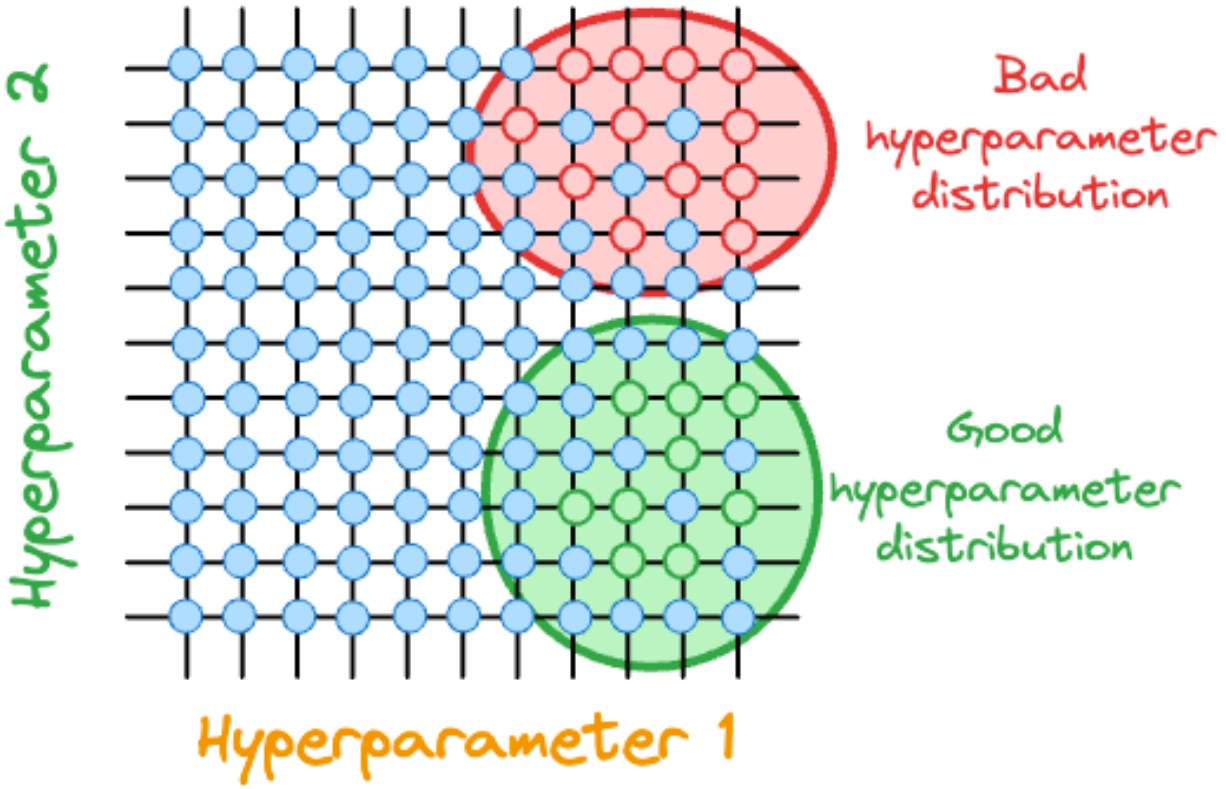
Search Algorithms
How to tune?



Example: Bayesian Optimization

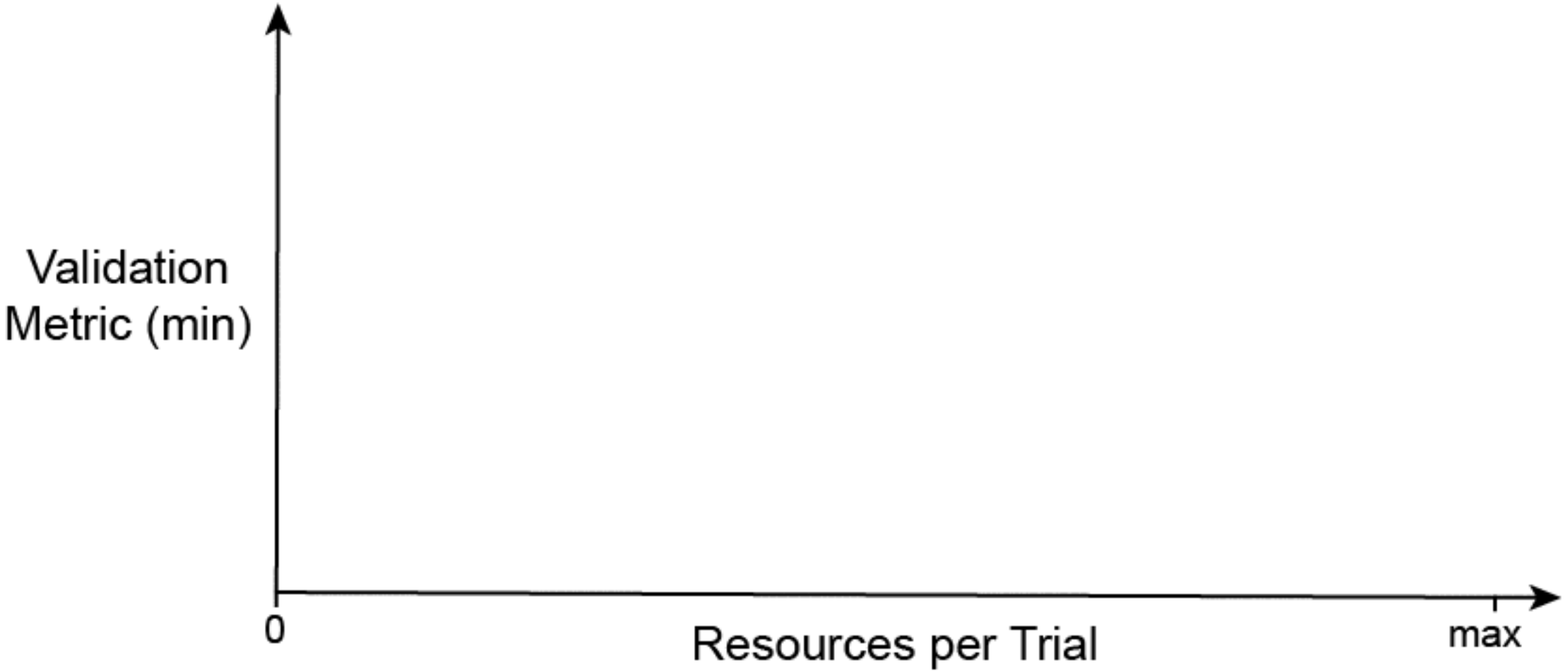
Search Algorithms
How to tune?





Example: ASHA

Schedulers
When to stop?



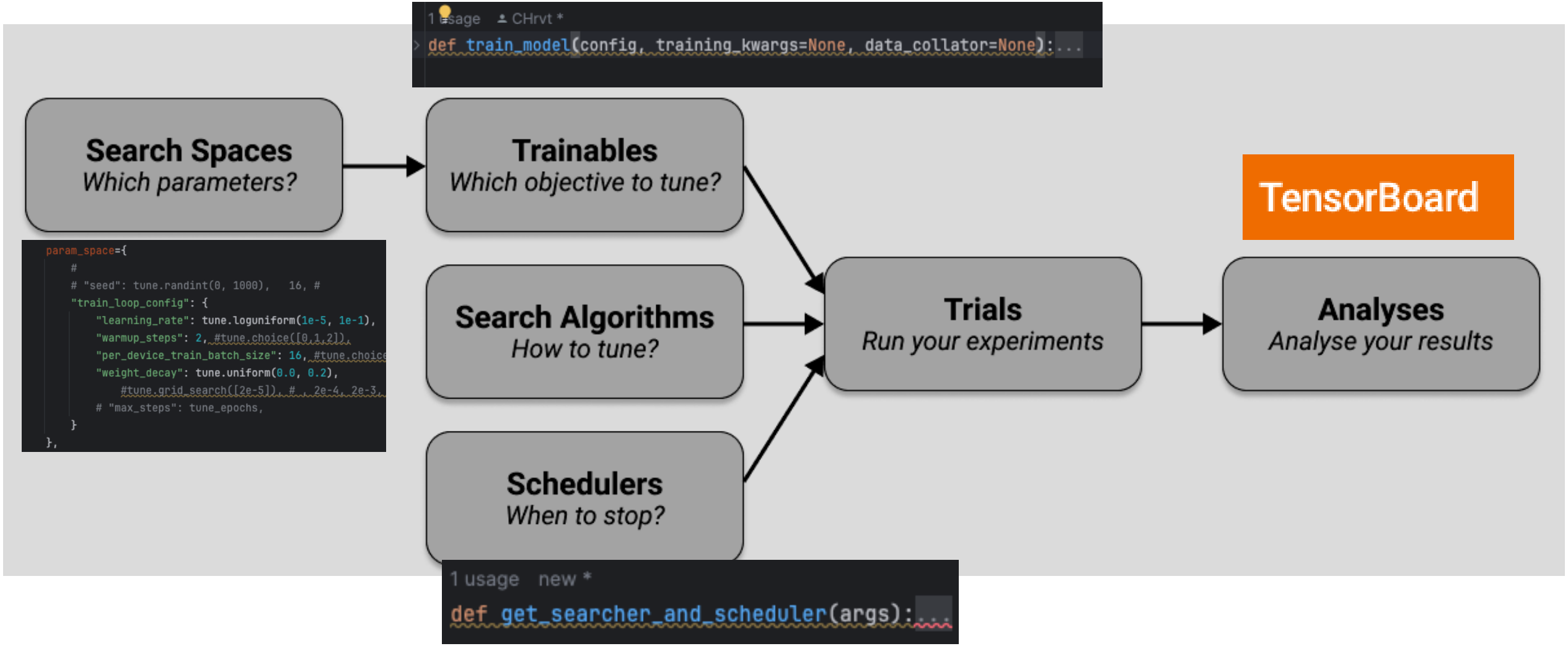
Code base overview

```
/asr-christian
├── /output                    # output folder of training results
├── /main_ray_for_HF.py        # main train function
├── /utils.py                  # Reusable utility functions
├── /configs                   # config files for different settings
└── /fine_tune.sh              # bash script for cluster
```

main_ray_for_HF.py

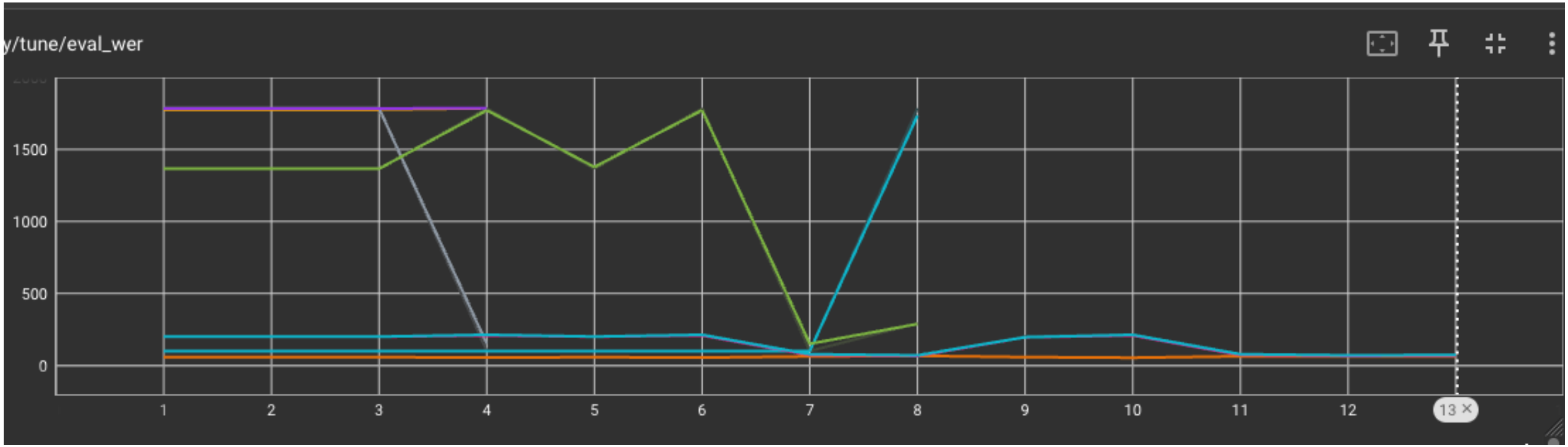
```python
def parse_args():
    """ Parses command line arguments for the training """
    parser = configargparse.ArgumentParser()

    # Plotting

    # Training settings for Seq2SeqTrainingArguments
    parser.add_argument("--per_device_train_batch_size", type=int, default=16, help="Batch size per device")
    parser.add_argument("--gradient_accumulation_steps", type=int, default=1, help="increase by 2x for every 2x decrease in batch size")
    parser.add_argument("--output_tag", type=str,
                        default="whisper-tiny-de",
                        help="Base directory where model is save.")
    parser.add_argument("--max_steps", type=int, default=1000, help="Max Number of gradient steps")
    # parser.add_argument("--warmup_steps", type=int, default=500, help="Warumup gradient steps")
    # parser.add_argument("--lr", type=float, default=1e-5, help="Initial learning rate.")
    parser.add_argument("--generation_max_length", type=int, default=225, help="Max length of token output")
    parser.add_argument("--save_steps", type=int, default=1000, help="After how many steps to save the model?")
    parser.add_argument("--eval_steps", type=int, default=1000, help="After how many steps to evaluate model")
    parser.add_argument("--logging_steps", type=int, default=25, help="After how many steps to do some logging")



    # model settings
    parser.add_argument("--model_type", type=str, default="openai/whisper-tiny", help="Model to optimize")
```

main_ray_for_HF.py



TensorBoard

Trial name	status	...fig/learning_rate	...nfig/weight_decay	iter	total time (s)	loss	grad_norm	learning_rate	epoch
TorchTrainer_87d48704	TERMINATED	1.45961e-05	0.0795039	13	304.626	0.0739	1.66526	0	24.04
TorchTrainer_16b8b31e	TERMINATED	1.67555e-05	0.148985	13	300.36	0.0425	0.723585	0	24.04
TorchTrainer_63e42d0a	TERMINATED	0.0170589	0.0362905	4	123.133	89.7637	200.942	0.0106618	9.04
TorchTrainer_f506cf47	TERMINATED	0.0725014	0.0665699	4	120.779	357.009	220.22	0.0453134	9.04
TorchTrainer_59920c70	TERMINATED	0.0140648	0.000794956	8	203.404	44.9807	201.13	0.00293017	19.04
TorchTrainer_ac708801	TERMINATED	0.000125898	0.197719	13	302.116	0.0004	0.0060018	0	24.04
TorchTrainer_58f979ec	TERMINATED	0.0150517	0.194179	4	116.54	77.3033	218.035	0.00940729	9.04
TorchTrainer_f69fed3d	TERMINATED	0.0147345	0.149292	8	193.819	51.1019	49.0424	0.00306968	19.04



Challenges:

- Find the right resource allocations for cluster computing
  - Multi nodes, efficient GPU utilization => might depend on choosen search algorithm
- whisper v3 could not be trained on gtx 1080 with BS 1
  - Different partitions, e.g. scanvenger which have better GPUs
- Which model to fine-tune? v2, v3, v3-german, medium, all?
- Faster Whisper conversion
- Conditional vs. non-conditional training?