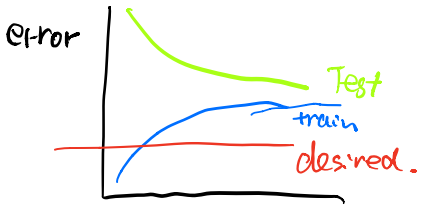
① Overfitting vs. underfitting / Bias vs. variance.

Bias:

- how much your model **underfits** the training model.
- Bias = $E[\hat{f}(x) - f(x)]$: Expected diff between predicted & observation.
- Error < reducible : improve
  irreducible : unmeasured / unmeasurable.
  Bayes Error.



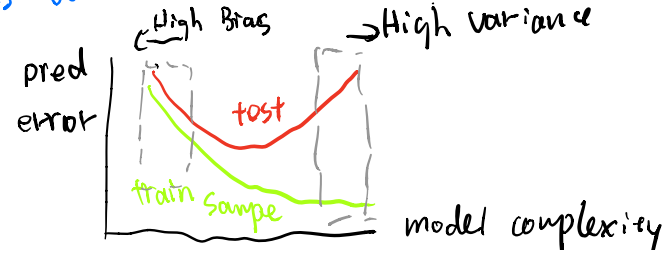
- high bias :
  - more features
  - complex models.

Variance:

- the amount that an algorithm's model will (mover around its mean) **change** if **different data** (new) used.
  ↳ how much you overfit the data.

- high variance:
  - Feature selection
  - Regularization
  - Dimensionality Reduction
  - Bagging (RF).

Bias-variance trade-off.



Overfitting:

- refers to a model that train the data too well.
  - noise/random fluctuations is picked up and learned.
- good performace on train data, poor generalization
- train error << val error. ◦ test error ↑. maybe diff dist.

underfitting:

- can neither model train data nor generalize new
- poor & poor
  - train & val error high.

How to prevent Overfitting

① k-fold cross validation. ← detect. reference. metric

- train/test model k times on diff subset, build up an estimate of performance on unseen data

? if train & val loss small. can we say it's overfitting to the train dataset? or just diff distribution?

② more data

③ feature selection.

④ Early stop

⑤ Regularization.

⑥ Ensembling.

keep modeling with cross-validation to check.

Bagging

Answer:

▲ overfitting/underfitting related to the idea of bias-variance trade-offs.

▲ overfitting means {train too well / capture the detail randomness of current dataset.
Can't be generalized to new data. → low train loss + high val loss

▲ underfitting: Can't capture the info of data.
→ high train & val loss.

▲ To solve:
- under the framework of cross-validation:
  adjust model, until lowest val loss. ⇒ reach the balance point with both low △ { train + val loss. / bias + variance.

  ▲ to solve underfitting:
  - more features
  - complex model.

  ▲ To solve overfitting:
  - ....