Statistical and online learning

Ideas from learning theory

Geelon So, agso@ucsd.edu
December 15, 2023

Abstract

At a high level, LEARNING is the process of extracting knowledge, skills, and useful behaviors out of past experience. To give machines the capability to learn, we need to operationalize this definition. The learning theorist offers two general models: STATISTICAL LEARNING and ONLINE LEARNING.

In this talk, I'll introduce these two classical frameworks and discuss some of their central questions, technical ideas, and limitations.

Learning from experience

A general schematic of learning:

$$\mathcal{A}_n : (Z_1, \ldots, Z_n) \mapsto \hat{\theta}_n.$$

- $(Z_1, \ldots, Z_n)$, data/experience
- $\hat{\theta}_n$, knowledge/decision

Learning from experience

A general schematic of learning:

$$\mathcal{A}_n : (Z_1, \ldots, Z_n) \mapsto \hat{\theta}_n.$$

- $(Z_1, \ldots, Z_n)$, data/experience
- $\hat{\theta}_n$, knowledge/decision

Goal: derive better knowledge/decisions from more data/experience

Classical statistics

Question: how to infer population-level knowledge from individual samples?

Classical statistics

Question: how to infer population-level knowledge from individual samples?

- hypothesis testing
- parameter estimation

Classical statistics

Question: how to infer population-level knowledge from individual samples?

- hypothesis testing
- parameter estimation

Example setting: making business decisions based on population statistics

Parameter estimation

We aim to learn some parameter θ^* about a population.

Parameter estimation

We aim to learn some parameter θ^* about a population.

- The true parameter may be too costly or impossible to measure.

Parameter estimation

We aim to learn some parameter θ^* about a population.

- The true parameter may be too costly or impossible to measure.
- Instead, we construct an estimate from data:

$$\hat{\theta}_n \equiv \mathcal{A}_n(Z_1, \ldots, Z_n).$$

Parameter estimation

We aim to learn some parameter θ^* about a population.

- The true parameter may be too costly or impossible to measure.
- Instead, we construct an estimate from data:

$$\hat{\theta}_n \equiv \mathcal{A}_n(Z_1, \ldots, Z_n).$$

- Questions in statistics and experimental design:

Parameter estimation

We aim to learn some parameter θ^* about a population.

- The true parameter may be too costly or impossible to measure.
- Instead, we construct an estimate from data:

$$\hat{\theta}_n \equiv \mathcal{A}_n(Z_1, \ldots, Z_n).$$

- Questions in statistics and experimental design:
  - unbiasedness: is the estimator correct in expectation?

Parameter estimation

We aim to learn some parameter θ^* about a population.

- The true parameter may be too costly or impossible to measure.
- Instead, we construct an estimate from data:

$$\hat{\theta}_n \equiv \mathcal{A}_n(Z_1, \ldots, Z_n).$$

- Questions in statistics and experimental design:
  - unbiasedness: is the estimator correct in expectation?
  - consistency: does the estimator converge to the true parameter with more data?

Parameter estimation

We aim to learn some parameter θ^* about a population.

- The true parameter may be too costly or impossible to measure.
- Instead, we construct an estimate from data:

$$\hat{\theta}_n \equiv \mathcal{A}_n(Z_1, \ldots, Z_n).$$

- Questions in statistics and experimental design:
  - unbiasedness: is the estimator correct in expectation?
  - consistency: does the estimator converge to the true parameter with more data?
  - efficiency: how well does the estimator recover information about θ^* in data?

Parameter estimation

We aim to learn some parameter θ^* about a population.

- The true parameter may be too costly or impossible to measure.
- Instead, we construct an estimate from data:

$$\hat{\theta}_n \equiv \mathcal{A}_n(Z_1, \ldots, Z_n).$$

- Questions in statistics and experimental design:
  - unbiasedness: is the estimator correct in expectation?
  - consistency: does the estimator converge to the true parameter with more data?
  - efficiency: how well does the estimator recover information about θ^* in data?
  - robustness: how sensitive is the estimator to outliers?

Mean estimation

Example. Let p be a distribution over $[0, 1]$.

Mean estimation

Example. Let p be a distribution over $[0, 1]$.

- Goal: estimate the mean $\theta^* = \mathbb{E}_{Z \sim p}[Z]$.

Mean estimation

Example. Let p be a distribution over $[0, 1]$.

- Goal: estimate the mean $\theta^* = \mathop{\mathbb{E}}_{Z\sim p}[Z]$.
- Sample mean estimator:

$$\hat{\theta}_n(Z_1, \ldots, Z_n) := \frac{1}{n}\sum_{i=1}^{n} Z_i,$$

where $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} p$.

Mean estimation

Example. Let p be a distribution over $[0, 1]$.

- Goal: estimate the mean $\theta^* = \mathop{\mathbb{E}}_{Z \sim p}[Z]$.
- Sample mean estimator:

$$\hat{\theta}_n(Z_1, \ldots, Z_n) := \frac{1}{n} \sum_{i=1}^{n} Z_i,$$

  where $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} p$.
- Law of large numbers:

$$\hat{\theta}_n \to \theta^*.$$

Concentration of measure phenomenon

Hoeffding's inequality. Let $\hat{\theta}_n$ be as before. Then:

$$\Pr\left(\left|\hat{\theta}_n - \theta^*\right| < \varepsilon\right) \geq 1 - \delta_n,$$

Concentration of measure phenomenon

Hoeffding's inequality. Let $\hat{\theta}_n$ be as before. Then:

$$\Pr\left(\left|\hat{\theta}_n - \theta^*\right| < \varepsilon\right) \geq 1 - \delta_n,$$

where $\delta_n = 2\exp\left(-2n\varepsilon^2\right)$.

Concentration of measure phenomenon

Hoeffding's inequality. Let $\hat{\theta}_n$ be as before. Then:

$$\Pr\left(\left|\hat{\theta}_n - \theta^*\right| < \varepsilon\right) \geq 1 - \delta_n,$$

where $\delta_n = 2\exp\left(-2n\varepsilon^2\right)$.

- A finite-sample version of the law of large numbers.

Statistical learning theory

Artificial intelligence

Question: can machines perform tasks that traditionally require human cognition?

Artificial intelligence

Question: can machines perform tasks that traditionally require human cognition?

- Hard: figuring out the mechanism of human cognition

Artificial intelligence

Question: can machines perform tasks that traditionally require human cognition?

- Hard: figuring out the mechanism of human cognition
- Easier: approximating human cognition functionally

$$f : \mathcal{X} \to \mathcal{Y}$$

Artificial intelligence

Question: can machines perform tasks that traditionally require human cognition?

- Hard: figuring out the mechanism of human cognition
- Easier: approximating human cognition functionally

$$f : \mathcal{X} \to \mathcal{Y}$$

- $\mathcal{X}$ is the set of problem/task instances
- $\mathcal{Y}$ is the set of possible decisions/responses

Learning from examples

Supervised learning paradigm.

Learning from examples

Supervised learning paradigm. Teach a machine by giving it examples (training data):

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

Learning from examples

Supervised learning paradigm. Teach a machine by giving it examples (training data):

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

where ideally each $Y_i = f(X_i)$ is an example of how to perform task X_i.

Learning from examples

Supervised learning paradigm. Teach a machine by giving it examples (training data):

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

where ideally each $Y_i = f(X_i)$ is an example of how to perform task X_i.

- Specify a loss function:

$$\ell(x, y, y') = \text{how much worse the response } y' \text{ is than } y \text{ for the instance } x.$$

Learning from examples

Supervised learning paradigm. Teach a machine by giving it examples (training data):

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

where ideally each $Y_i = f(X_i)$ is an example of how to perform task X_i.

- Specify a loss function:

$$\ell(x, y, y') = \text{how much worse the response } y' \text{ is than } y \text{ for the instance } x.$$

- Specify a class of functions that could perform task: $\mathcal{F} \equiv \{f_\theta : \theta \in \Theta\}$.

Learning from examples

Supervised learning paradigm. Teach a machine by giving it examples (training data):

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

where ideally each $Y_i = f(X_i)$ is an example of how to perform task X_i.

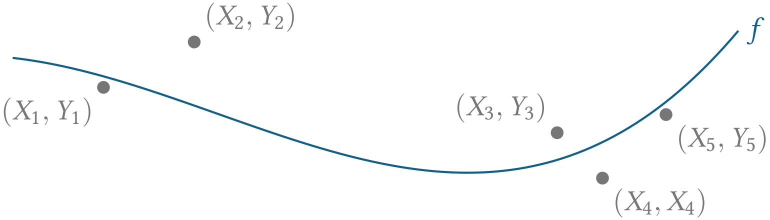
- Specify a loss function:

$$\ell(x, y, y') = \text{how much worse the response } y' \text{ is than } y \text{ for the instance } x.$$

- Specify a class of functions that could perform task: $\mathcal{F} \equiv \left\{f_\theta : \theta \in \Theta\right\}$.
- Solve the optimization problem: $\hat{\theta}_n \leftarrow \underset{f_\theta \in \mathcal{F}}{\arg\min} \sum_{i=1}^{n} \ell\big(X_i, Y_i, f_\theta(X_i)\big).$

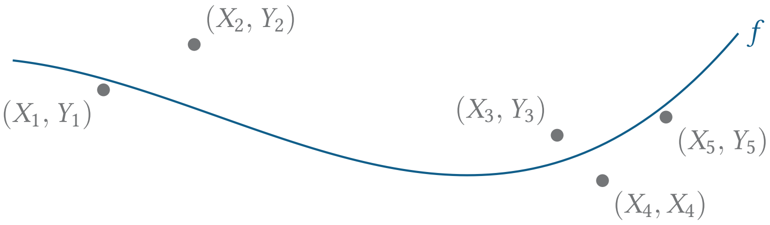
Learning from examples

A new goal: generalize well to previously unseen tasks not in training data



Learning from examples

A new goal: generalize well to previously unseen tasks not in training data



Intuitive justification of optimization approach: the model that performs best on the train set will also perform best overall.

- Possible justification: extension of method of least squares

Theoretical justification: empirical risk minimization

Why should good performance on the train set lead to good overall performance?

Theoretical justification: empirical risk minimization

Why should good performance on the train set lead to good overall performance?

- Statistical assumption: each training example and future tasks are drawn i.i.d. from the same distribution p over $\mathcal{X} \times \mathcal{Y}$

Theoretical justification: empirical risk minimization

Why should good performance on the train set lead to good overall performance?

- Statistical assumption: each training example and future tasks are drawn i.i.d. from the same distribution p over $\mathcal{X} \times \mathcal{Y}$
- This allows us to use tools from classical statistics to justify approach.

Theoretical justification: empirical risk minimization

Why should good performance on the train set lead to good overall performance?

- Statistical assumption: each training example and future tasks are drawn i.i.d. from the same distribution p over $\mathcal{X} \times \mathcal{Y}$
- This allows us to use tools from classical statistics to justify approach.
- In some sense, very general: no further assumptions on p.

Risk minimization

What is being optimized?

- Define the *risk* of f_θ as its average (population) loss:

$$R(f_\theta) := \mathop{\mathbb{E}}_{(X,Y)\sim p}\left[\ell(X, Y, f_\theta(X))\right].$$

Risk minimization

What is being optimized?

- Define the *risk* of f_θ as its average (population) loss:

$$R(f_\theta) := \mathop{\mathbb{E}}_{(X,Y)\sim p}\left[\ell(X, Y, f_\theta(X))\right].$$

  - Natural goal: find $f_\theta \in \mathcal{F}$ minimizing risk.

Risk minimization

What is being optimized?

- Define the *risk* of f_θ as its average (population) loss:

$$R(f_\theta) := \mathop{\mathbb{E}}_{(X,Y)\sim p}\left[\ell(X, Y, f_\theta(X))\right].$$

  - Natural goal: find $f_\theta \in \mathcal{F}$ minimizing risk.
  - Suppose we only have access to p through samples.

Risk minimization

What is being optimized?

- Define the *risk* of f_θ as its average (population) loss:

$$R(f_\theta) := \mathop{\mathbb{E}}_{(X,Y)\sim p}\left[\ell(X, Y, f_\theta(X))\right].$$

  - Natural goal: find $f_\theta \in \mathcal{F}$ minimizing risk.
  - Suppose we only have access to p through samples.
- Estimate this statistical quantity using samples:

$$\hat{R}_n(f_\theta) := \frac{1}{n}\sum_{i=1}^{n}\ell\big(X_i, Y_i, f_\theta(X_i)\big).$$

Risk minimization

What is being optimized?

- Define the *risk* of f_θ as its average (population) loss:

$$R(f_\theta) := \mathop{\mathbb{E}}_{(X,Y)\sim p}\left[\ell(X, Y, f_\theta(X))\right].$$

    - Natural goal: find $f_\theta \in \mathcal{F}$ minimizing risk.
    - Suppose we only have access to p through samples.
- Estimate this statistical quantity using samples:

$$\hat{R}_n(f_\theta) := \frac{1}{n}\sum_{i=1}^{n}\ell\big(X_i, Y_i, f_\theta(X_i)\big).$$

We call this the *empirical risk* (note: this is a random variable).

Risk minimization

What is being optimized?

- Define the *risk* of f_θ as its average (population) loss:

$$R(f_\theta) := \mathop{\mathbb{E}}_{(X,Y)\sim p}\left[\ell(X, Y, f_\theta(X))\right].$$

  - Natural goal: find $f_\theta \in \mathcal{F}$ minimizing risk.
  - Suppose we only have access to p through samples.
- Estimate this statistical quantity using samples:

$$\hat{R}_n(f_\theta) := \frac{1}{n}\sum_{i=1}^{n}\ell\big(X_i, Y_i, f_\theta(X_i)\big).$$

  We call this the *empirical risk* (note: this is a random variable).
- The function $f_{\hat{\theta}_n}$ minimizing $\hat{R}_n$ is called the *empirical risk minimizer* (ERM).

Sketch of empirical risk minimization

- Because of the i.i.d. assumption, the empirical risk exhibits concentration:

$$\Pr\left(\left|\hat{R}_n(f_\theta) - R(f_\theta)\right| < \varepsilon\right) > 1 - \delta.$$

Sketch of empirical risk minimization

- Because of the i.i.d. assumption, the empirical risk exhibits concentration:

$$\Pr\left(\left|\hat{R}_n(f_\theta) - R(f_\theta)\right| < \varepsilon\right) > 1 - \delta.$$

- Suppose that there are N functions $\mathcal{F} = \{f_{\theta_1}, \ldots, f_{\theta_N}\}$; by union bound,

$$\Pr\left(\max_{f_\theta \in \mathcal{F}} \left|\hat{R}_n(f_\theta) - R(f_\theta)\right| < \varepsilon\right) > 1 - \boldsymbol{N}\delta.$$

Sketch of empirical risk minimization

- Because of the i.i.d. assumption, the empirical risk exhibits concentration:

$$\Pr\left(\left|\hat{R}_n(f_\theta) - R(f_\theta)\right| < \varepsilon\right) > 1 - \delta.$$

- Suppose that there are N functions $\mathcal{F} = \{f_{\theta_1}, \ldots, f_{\theta_N}\}$; by union bound,

$$\Pr\left(\max_{f_\theta \in \mathcal{F}} \left|\hat{R}_n(f_\theta) - R(f_\theta)\right| < \varepsilon\right) > 1 - N\delta.$$

- If the event in the probability holds, then the empirical risk minimizer satisfies:

$$R(f_{\hat{\theta}_n}) < \min_{f_\theta \in \mathcal{F}} R(f_\theta) + 2\varepsilon.$$

Sketch of empirical risk minimization

- Because of the i.i.d. assumption, the empirical risk exhibits concentration:

$$\Pr\left(\left|\hat{R}_n(f_\theta) - R(f_\theta)\right| < \varepsilon\right) > 1 - \delta.$$

- Suppose that there are N functions $\mathcal{F} = \{f_{\theta_1}, \ldots, f_{\theta_N}\}$; by union bound,

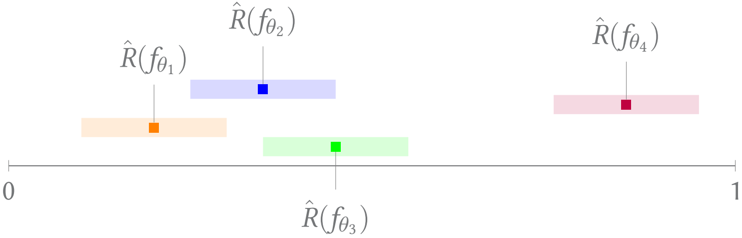
$$\Pr\left(\max_{f_\theta \in \mathcal{F}} \left|\hat{R}_n(f_\theta) - R(f_\theta)\right| < \varepsilon\right) > 1 - N\delta.$$

- If the event in the probability holds, then the empirical risk minimizer satisfies:

$$R(f_{\hat{\theta}_n}) < \min_{f_\theta \in \mathcal{F}} R(f_\theta) + 2\varepsilon.$$

  - The ERM cannot be much worse than the best-in-class function.

Sketch of empirical risk minimization



Model capacity

Notice that the union bound is only meaningful if ε is sufficiently large so that $N\delta \ll 1$

$$\Pr\left(\max_{f_\theta \in \mathcal{F}} \left|\hat{R}_n(f_\theta) - R(f_\theta)\right| < \varepsilon\right) > 1 - N\delta.$$

Model capacity

Notice that the union bound is only meaningful if ε is sufficiently large so that $N\delta \ll 1$

$$\Pr\left(\max_{f_\theta \in \mathcal{F}} \left|\hat{R}_n(f_\theta) - R(f_\theta)\right| < \varepsilon\right) > 1 - N\delta.$$

Since N is the size of the model $\mathcal{F}$, this leads to the classic bias-variance tradeoff:

$$R(f_{\hat{\theta}_n}) < \underbrace{\min_{f_\theta \in \mathcal{F}} R(f_\theta)}_{\text{bias}} + \underbrace{2\varepsilon}_{\text{variance}},$$

Model capacity

Notice that the union bound is only meaningful if ε is sufficiently large so that $N\delta \ll 1$

$$\Pr\left(\max_{f_\theta \in \mathcal{F}} \left|\hat{R}_n(f_\theta) - R(f_\theta)\right| < \varepsilon\right) > 1 - N\delta.$$

Since N is the size of the model $\mathcal{F}$, this leads to the classic bias-variance tradeoff:

$$R(f_{\hat{\theta}_n}) < \underbrace{\min_{f_\theta \in \mathcal{F}} R(f_\theta)}_{\text{bias}} + \underbrace{2\varepsilon}_{\text{variance}},$$

where the bias term decreases (expressivity increases) when N increases,

Model capacity

Notice that the union bound is only meaningful if ε is sufficiently large so that $N\delta \ll 1$

$$\Pr\left(\max_{f_\theta \in \mathcal{F}} \left|\hat{R}_n(f_\theta) - R(f_\theta)\right| < \varepsilon\right) > 1 - N\delta.$$

Since N is the size of the model $\mathcal{F}$, this leads to the classic bias-variance tradeoff:

$$R(f_{\hat{\theta}_n}) < \underbrace{\min_{f_\theta \in \mathcal{F}} R(f_\theta)}_{\text{bias}} + \underbrace{\quad 2\varepsilon \quad}_{\text{variance}},$$

where the bias term decreases (expressivity increases) when N increases, but the variance term increases with N.

Bias-variance tradeoff

Intuition: how the bias of $\mathcal{F}$ relates to the capacity of $\mathcal{F}$.

Bias-variance tradeoff

Intuition: how the bias of $\mathcal{F}$ relates to the capacity of $\mathcal{F}$.

- **Small capacity:** if $\mathcal{F}$ cannot perform many tasks, none might be well-suited for the task at hand. This leads to a large bias term.

Bias-variance tradeoff

Intuition: how the bias of $\mathcal{F}$ relates to the capacity of $\mathcal{F}$.

- **Small capacity:** if $\mathcal{F}$ cannot perform many tasks, none might be well-suited for the task at hand. This leads to a large bias term.
- **Large capacity:** if the examples can correspond to many different tasks in $\mathcal{F}$, which is the right one? This leads to a large variance term.

Classic generalization result

Generalization theory tends to give us bounds on the estimation error term so that:

$$R(f_{\hat{\theta}}) \leq \sqrt{\frac{\text{capacity of the model}}{\text{number of training examples}}} + \text{bias of the model.}$$

Overall logic

- Computationally: learning as an optimization problem

Overall logic

- Computationally: learning as an optimization problem
  - finding the model that best fits training data

Overall logic

- Computationally: learning as an optimization problem
  - finding the model that best fits training data
- Theoretically: learning as a parameter estimation problem

Overall logic

- Computationally: learning as an optimization problem
  - finding the model that best fits training data
- Theoretically: learning as a parameter estimation problem
  - introduce statistical assumption

Limitations of classical statistical learning theory

- Standard supervised learning framed as learning from examples

Limitations of classical statistical learning theory

- Standard supervised learning framed as learning from examples
  - What are other modes of learning?

Limitations of classical statistical learning theory

- Standard supervised learning framed as learning from examples
  - What are other modes of learning?
- It imposes a strong distributional assumption

Limitations of classical statistical learning theory

- Standard supervised learning framed as learning from examples
  - What are other modes of learning?
- It imposes a strong distributional assumption
  - How do we think about learning problems that are not naturally statistical?

Limitations of classical statistical learning theory

- Standard supervised learning framed as learning from examples
  - What are other modes of learning?
- It imposes a strong distributional assumption
  - How do we think about learning problems that are not naturally statistical?
- It does not seem to explain the successes of deep learning

Online learning

Online learning framework

Setting. For $t = 1, 2, \ldots$

Online learning framework

Setting. For $t = 1, 2, \ldots$

- receive instance $X_t \in \mathcal{X}$

Online learning framework

Setting. For $t = 1, 2, \ldots$

- receive instance $X_t \in \mathcal{X}$
- make prediction $\hat{Y}_t \in \mathcal{Y}$

Online learning framework

Setting. For $t = 1, 2, \ldots$

- receive instance $X_t \in \mathcal{X}$
- make prediction $\hat{Y}_t \in \mathcal{Y}$
- receive label Y_t

Online learning framework

Setting. For $t = 1, 2, \ldots$

- receive **instance** $X_t \in \mathcal{X}$
- make **prediction** $\hat{Y}_t \in \mathcal{Y}$
- receive **label** Y_t
- incur **loss** $\ell(X_t, Y_t, \hat{Y}_t)$

Online learning framework

Setting. For $t = 1, 2, \ldots$

- receive instance $X_t \in \mathcal{X}$
- make prediction $\hat{Y}_t \in \mathcal{Y}$
- receive label Y_t
- incur loss $\ell(X_t, Y_t, \hat{Y}_t)$

View this as a repeated zero-sum game against Nature.

Online learning framework

Setting. For $t = 1, 2, \ldots$

- receive instance $X_t \in \mathcal{X}$
- make prediction $\hat{Y}_t \in \mathcal{Y}$
- receive label Y_t
- incur loss $\ell(X_t, Y_t, \hat{Y}_t)$

View this as a repeated zero-sum game against Nature.

- To learn means being able to exploit a suboptimal strategy.

Online learning framework

Setting. For $t = 1, 2, \ldots$

- receive instance $X_t \in \mathcal{X}$
- make prediction $\hat{Y}_t \in \mathcal{Y}$
- receive label Y_t
- incur loss $\ell(X_t, Y_t, \hat{Y}_t)$

View this as a repeated zero-sum game against Nature.

- To learn means being able to exploit a suboptimal strategy.
- Goal: if Nature plays $(X_t, Y_t) \overset{\text{i.i.d.}}{\sim} p$, learner should be able to recover results from statistical setting.

Online learning framework

Difference with statistical learning

- There is no test-train split.

Online learning framework

Difference with statistical learning

- There is no test-train split.
- There is no statistical assumptions, e.g. (X_t, Y_t) could be arbitrary.

Online learning framework

Difference with statistical learning

- There is no test-train split.
- There is no statistical assumptions, e.g. (X_t, Y_t) could be arbitrary.
- The goal is *regret minimization* instead of risk minimization.

Goal in online learning

The cumulative loss at time T:

$$L_T = \sum_{t=1}^{T} \ell(X_t, Y_t, \hat{Y}_t).$$

Goal in online learning

The cumulative loss at time T:

$$L_T = \sum_{t=1}^{T} \ell(X_t, Y_t, \hat{Y}_t).$$

Goal: we would like to minimize the cumulative loss, but this is an extremely tall order

Goal in online learning

The cumulative loss at time T:

$$L_T = \sum_{t=1}^{T} \ell(X_t, Y_t, \hat{Y}_t).$$

Goal: we would like to minimize the cumulative loss, but this is an extremely tall order

- regret analysis: compare against some restricted class of predictors

Goal in online learning

The cumulative loss at time T:

$$L_T = \sum_{t=1}^{T} \ell(X_t, Y_t, \hat{Y}_t).$$

Goal: we would like to minimize the cumulative loss, but this is an extremely tall order

- regret analysis: compare against some restricted class of predictors
  - regret: looking back in hindsight and realizing there was a simple strategy

Prediction with experts

We can think of $\mathcal{F}$ as a set of experts.

Prediction with experts

We can think of $\mathcal{F}$ as a set of experts.

- Each round, the expert θ recommends predicting $\hat{Y}_t = f_\theta(X_t)$.

Prediction with experts

We can think of $\mathcal{F}$ as a set of experts.

- Each round, the expert θ recommends predicting $\hat{Y}_t = f_\theta(X_t)$.
- Each expert incurs cumulative loss:

$$L_{T;\theta} = \sum_{t=1}^{T} \ell\big(X_t, Y_t, f_\theta(X_t)\big).$$

Prediction with experts

We can think of $\mathcal{F}$ as a set of experts.

- Each round, the expert θ recommends predicting $\hat{Y}_t = f_\theta(X_t)$.
- Each expert incurs cumulative loss:

$$L_{T;\theta} = \sum_{t=1}^{T} \ell\big(X_t, Y_t, f_\theta(X_t)\big).$$

- The regret of deciding to predict $(\hat{Y}_t)_t$ instead is defined:

$$R_T = \sum_{t=1}^{T} \ell\big(X_t, Y_t, \hat{Y}_t\big) - \min_{f_\theta \in \mathcal{F}} \underbrace{\sum_{t=1}^{T} \ell\big(X_t, Y_t, f_\theta(X_t)\big)}_{L_{T;\theta}}.$$

Prediction with experts

We can think of $\mathcal{F}$ as a set of experts.

- Each round, the expert θ recommends predicting $\hat{Y}_t = f_\theta(X_t)$.
- Each expert incurs cumulative loss:

$$L_{T;\theta} = \sum_{t=1}^{T} \ell\big(X_t, Y_t, f_\theta(X_t)\big).$$

- The regret of deciding to predict $(\hat{Y}_t)_t$ instead is defined:

$$R_T = \sum_{t=1}^{T} \ell\big(X_t, Y_t, \hat{Y}_t\big) - \min_{f_\theta \in \mathcal{F}} \underbrace{\sum_{t=1}^{T} \ell\big(X_t, Y_t, f_\theta(X_t)\big)}_{L_{T;\theta}}.$$

  - In hindsight, this is the regret for not listening to the best expert.

Meaning of achieving low regret

In online learning, the aim is sublinear regret $R_T = o(T)$,

$$\lim_{T\to\infty} \frac{1}{T} \underbrace{\left(\sum_{t=1}^{T} \ell\big(X_t, Y_t, \hat{Y}_t\big) - \min_{f_\theta \in \mathcal{F}} \sum_{t=1}^{T} \ell\big(X_t, Y_t, f_\theta(X_t)\big) \right)}_{R_T} \leq 0.$$

Meaning of achieving low regret

In online learning, the aim is sublinear regret $R_T = o(T)$,

$$\lim_{T\to\infty} \frac{1}{T} \underbrace{\left(\sum_{t=1}^{T} \ell\big(X_t, Y_t, \hat{Y}_t\big) - \min_{f_\theta \in \mathcal{F}} \sum_{t=1}^{T} \ell\big(X_t, Y_t, f_\theta(X_t)\big) \right)}_{R_T} \leq 0.$$

- In the i.i.d. case, this means:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \ell(X_t, Y_t, \hat{Y}_t) \leq \min_{f_\theta \in \mathcal{F}} R(f_\theta).$$

Meaning of achieving low regret

In online learning, the aim is sublinear regret $R_T = o(T)$,

$$\lim_{T \to \infty} \frac{1}{T} \underbrace{\left(\sum_{t=1}^{T} \ell\big(X_t, Y_t, \hat{Y}_t\big) - \min_{f_\theta \in \mathcal{F}} \sum_{t=1}^{T} \ell\big(X_t, Y_t, f_\theta(X_t)\big) \right)}_{R_T} \leq 0.$$

- In the i.i.d. case, this means:

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \ell(X_t, Y_t, \hat{Y}_t) \leq \min_{f_\theta \in \mathcal{F}} R(f_\theta).$$

- Regret minimization is meaningful when the best expert achieves low loss.

Follow the leader algorithm

At round t, follow the advice of the expert who made the fewest mistakes so far:

Follow the leader algorithm

At round t, follow the advice of the expert who made the fewest mistakes so far:

$$\text{listen to expert } \theta_t^* = \arg\min_{\theta} L_{t-1;\theta},$$

Follow the leader algorithm

At round t, follow the advice of the expert who made the fewest mistakes so far:

$$\text{listen to expert } \theta_t^* = \arg\min_{\theta} L_{t-1;\theta},$$

breaking ties arbitrarily.

Follow the leader algorithm

At round t, follow the advice of the expert who made the fewest mistakes so far:

$$\text{listen to expert } \theta_t^* = \arg\min_\theta L_{t-1;\theta},$$

breaking ties arbitrarily.

- This is the natural extension of empirical risk minimization to the online setting.

Follow the leader algorithm

At round t, follow the advice of the expert who made the fewest mistakes so far:

$$\text{listen to expert } \theta_t^* = \arg\min_{\theta} L_{t-1;\theta},$$

breaking ties arbitrarily.

- This is the natural extension of empirical risk minimization to the online setting.
- An adversary can force this algorithm to make a mistake almost every round.

Follow the leader algorithm

At round t, follow the advice of the expert who made the fewest mistakes so far:

$$\text{listen to expert } \theta_t^* = \arg\min_\theta L_{t-1;\theta},$$

breaking ties arbitrarily.

- This is the natural extension of empirical risk minimization to the online setting.
- An adversary can force this algorithm to make a mistake almost every round.
- On the other hand, the best expert is correct at least on average $1/N$ of times.

Hedge algorithm

At round t, follow a random expert's advice:

$$\text{probability of listening to expert } \theta \propto \exp\left(-\eta L_{t-1;\theta}\right).$$

Hedge algorithm

At round t, follow a random expert's advice:

$$\text{probability of listening to expert } \theta \propto \exp\left(-\eta L_{t-1;\theta}\right).$$

If there are N experts and $\eta = \sqrt{\frac{\log N}{8T}}$, then this achieves sublinear regret:

$$R_T \leq \sqrt{\frac{T \log N}{2}}.$$

Hedge algorithm

At round t, follow a random expert's advice:

$$\text{probability of listening to expert } \theta \propto \exp\left(-\eta L_{t-1;\theta}\right).$$

If there are N experts and $\eta = \sqrt{\frac{\log N}{8T}}$, then this achieves sublinear regret:

$$R_T \le \sqrt{\frac{T \log N}{2}}.$$

- This turns out to be optimal, with matching lower bound.

Online convex optimization (OCO)

Setting. Let $K \subset \mathbb{R}^N$ be a convex, compact set.
For $t = 1, 2, \ldots$

- make decision $z_t \in K \subset \mathbb{R}^N$
- receive convex loss function $\ell_t : K \to \mathbb{R}$
- incur loss $\ell_t(z_t)$

Goal. Minimize regret:

$$R_T = \sum_{t=1}^{T} \ell_t(z_t) - \inf_{z \in K} \sum_{t=1}^{T} \ell_t(z).$$

Meaning of low regret in OCO

In the case that $\ell_t \equiv \ell$ is a fixed convex loss function:

- Let $\bar{z}_T = \frac{1}{T} \sum z_i$ be the average iterate.
- Let $z^* = \arg\min \ell(z)$.

Jensen's inequality implies:

$$\ell(\bar{z}_T) - \ell(z^*) \leq \frac{1}{T} \sum_{t=1}^{T} \ell(z_t) - \ell(z^*) \leq \frac{R_T}{T}.$$

Prediction with experts as OCO

Let $K = \Delta^{N-1}$ be the probability simplex over the N experts $\mathcal{F} = \{f_{\theta_1}, \ldots, f_{\theta_2}\}$.

- Define the linear loss ℓ_t generated by X_t, Y_t by:

$$\ell_t(z) := \sum_{i=1}^{N} z_i \cdot \ell\big(X_t, Y_t, f_{\theta_i}(X_t)\big)$$

  - This is the expected loss incurred by choosing to listen to a random expert, where the probability of listening to expert θ_i is z_i.
- By convexity, the cumulative loss is achieved on a vertex of the simplex:

$$\min_{i \in [N]} \sum_{t=1}^{T} \ell\big(X_t, Y_t, f_{\theta_i}(X_t)\big) = \min_{z \in K} \sum_{t=1}^{T} \ell_t(z).$$

- Hedge is equivalent to online mirror descent.

Negative result for online learning

Consider learning a threshold function on the interval $[0, 1]$,

$$f_\theta(x) = \mathbf{1}\{x \geq \theta\}.$$

Negative result for online learning

Consider learning a threshold function on the interval $[0, 1]$,

$$f_\theta(x) = \mathbf{1}\{x \geq \theta\}.$$

- **Realizable setting**: adversary may select any $(X_t, Y_t)_t$ as long as some θ^* satisfies:

$$Y_t = f_{\theta^*}(X_t), \qquad \forall t.$$

Negative result for online learning

Consider learning a threshold function on the interval $[0, 1]$,

$$f_\theta(x) = \mathbf{1}\{x \geq \theta\}.$$

- **Realizable setting**: adversary may select any $(X_t, Y_t)_t$ as long as some θ^* satisfies:

$$Y_t = f_{\theta^*}(X_t), \qquad \forall t.$$

- **Negative result**: no online learner achieves sublinear regret in the worst case.

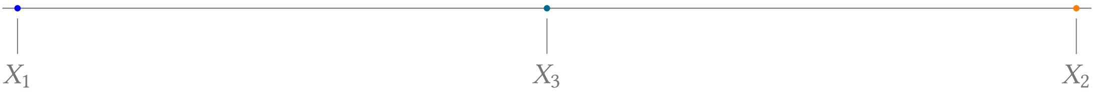
Construction of hard case

BINARY SEARCH SAMPLING ALGORITHM

Construction of hard case

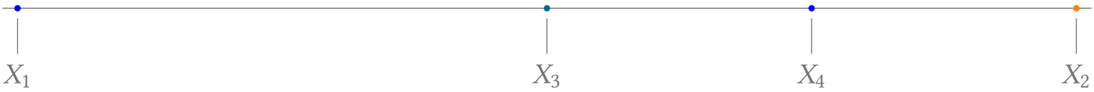
BINARY SEARCH SAMPLING ALGORITHM

Construction of hard case



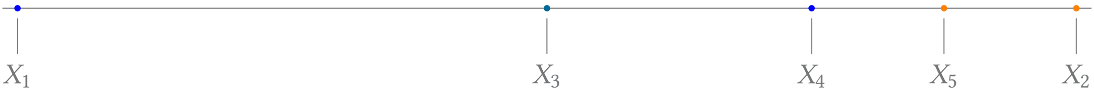
BINARY SEARCH SAMPLING ALGORITHM

Construction of hard case



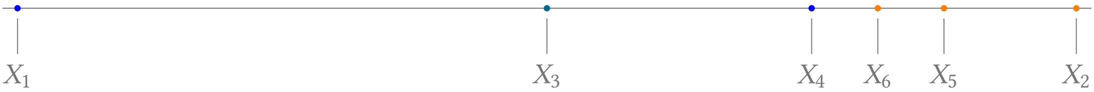
BINARY SEARCH SAMPLING ALGORITHM

Construction of hard case



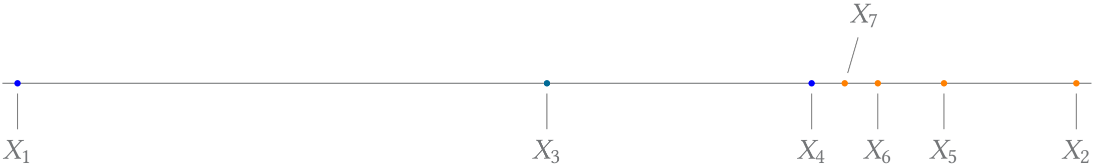
BINARY SEARCH SAMPLING ALGORITHM

Construction of hard case



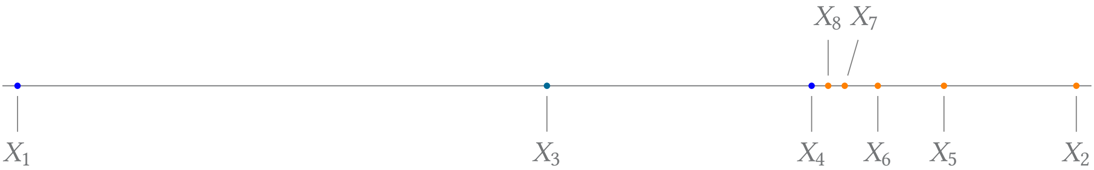
BINARY SEARCH SAMPLING ALGORITHM

Construction of hard case



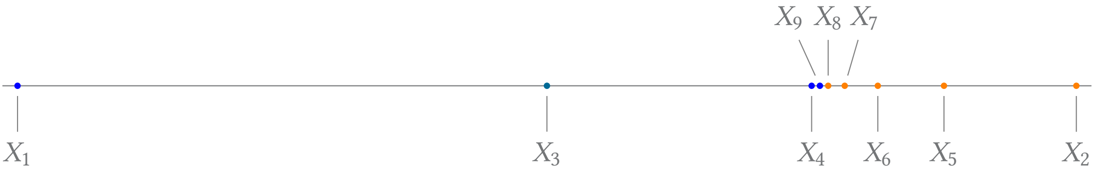
BINARY SEARCH SAMPLING ALGORITHM

Construction of hard case



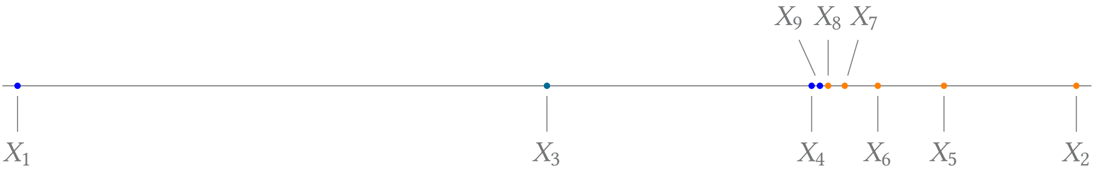
BINARY SEARCH SAMPLING ALGORITHM

Construction of hard case



BINARY SEARCH SAMPLING ALGORITHM

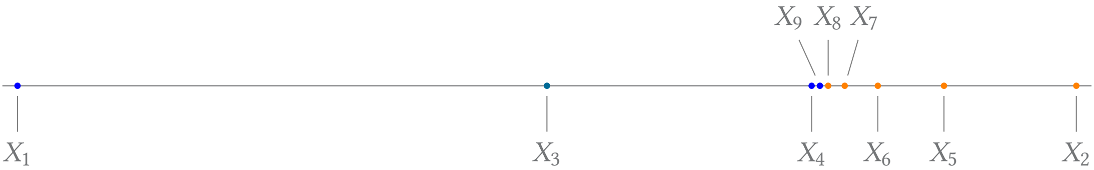
Construction of hard case



BINARY SEARCH SAMPLING ALGORITHM

For $t = 1, 2, \ldots$

Construction of hard case

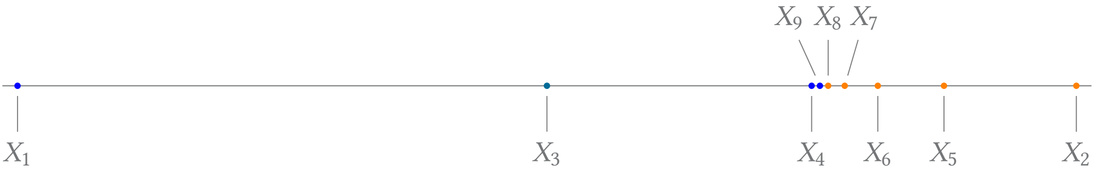


BINARY SEARCH SAMPLING ALGORITHM

For $t = 1, 2, \ldots$

- $X_- \leftarrow$ max **negative** data point in data set

Construction of hard case

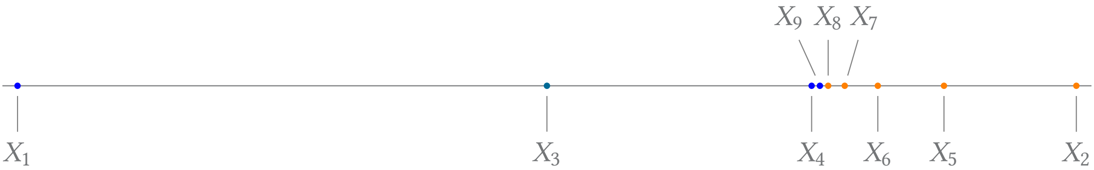


BINARY SEARCH SAMPLING ALGORITHM

For $t = 1, 2, \ldots$

- $X_- \leftarrow$ max negative data point in data set
- $X_+ \leftarrow$ min positive data point in data set

Construction of hard case



BINARY SEARCH SAMPLING ALGORITHM

For $t = 1, 2, \ldots$

- $X_- \leftarrow$ max negative data point in data set
- $X_+ \leftarrow$ min positive data point in data set
- $X_{t+1} \leftarrow \mathrm{mean}(X_-, X_+)$ and $Y_{t+1} \sim \mathrm{Ber}(\frac{1}{2})$

Learning thresholds: statistical vs. online

There's a big gap in hardness of learning thresholds:

- statistical setting: $R(f_{\hat{\theta}_n}) \asymp \frac{1}{n}$
- worst-case online setting: $\frac{1}{T} R_T \asymp \frac{1}{2}$

Limitations of online learning model

- The worst-case online setting is too hard.

Limitations of online learning model

- The worst-case online setting is too hard.
    - Reality may be much more average case or structured.

Limitations of online learning model

- The worst-case online setting is too hard.
  - Reality may be much more average case or structured.
  - In certain settings, the theory may not help us design/understand learning algorithms.

Some contributions of these frameworks

- Tools for learning problems with a statistical component
- Tools for learning problems with an adversarial component

Recap

- Statistical learning and online learning empirically very successful

Recap

- Statistical learning and online learning empirically very successful
- Much of learning left unsaid by these frameworks

Recap

- Statistical learning and online learning empirically very successful
- Much of learning left unsaid by these frameworks
- Much of learning cannot be approached by these frameworks

Recap

- Statistical learning and online learning empirically very successful
- Much of learning left unsaid by these frameworks
- Much of learning cannot be approached by these frameworks
- What's next?