Lecture 1: Introduction and learning theory overview

Akshay Krishnamurthy
akshay@cs.umass.edu

January 24, 2022

1 Introduction

This course is primarily about reinforcement learning or sequential decision making. This is a mathematical framework for learning to make decisions in an unknown environment to accomplish some task. These kinds of problems come up in a broad range of domains including robotics, aerospace engineering, online education, personalization, precision medicine, and elsewhere. Additionally, reinforcement learning has also been studied in neuroscience as a framework for how humans and biological organisms learn. With these applications as motivation and with recent high-profile empirical successes of reinforcement learning, it has become a central topic of study in the machine learning and artificial intelligence research communities.

In its most abstract form, the reinforcement learning protocol involves an agent repeatedly interacting with an environment in the schematic displayed in Figure 1. At each time step, the environment is in some configuration

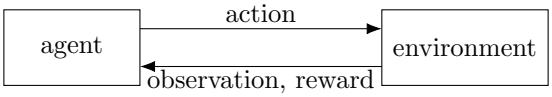

Figure 1: The reinforcement learning protocol

and the agent has received some information from the environment, in the form of observations and rewards. Using this information, the agent makes a decision, which we typically call an action. This action is executed in the environment, and, in response, the environment updates its internal state and sends the agent a new (observation, reward) tuple. We are interested in designing agents that learn to take actions that result in accumulating a lot of reward.

An example is a navigation task. We may have an agent in a maze trying to find its way out. The environment configuration is just the location of the agent in the maze. At each time the agent can try to move north, south, east, or west, and the environment will update the agent's location accordingly. Perhaps the agent only gets a reward if it successfully exits the maze. For observations, an easier setting is where the agent gets a top-down view of the maze. This is like how we might solve mazes in puzzle books. A harder setting is where the agent only gets a first-person view of the environment, like corn-field mazes you might visit around halloween.

To solve such problems, an agent needs to address three challenges:

- **Generalization.** The agent may have to leverage experience gathered in one situation to make good decisions in another situation. Indeed, when operating with a first-person view, the observations are extremely complicated, to the point where the agent may never see the same observation twice. The ability to generalize captures how we may be able to learn from complex inputs.
- **Exploration.** The agent may have to carefully make decisions to identify what task it is trying to solve, or to even see any reward. For example, in the maze scenario, if we act completely randomly, it would take us an exponentially long time to exit the maze by chance. Of course by doing this we will spend most of our time around where we started, which may be very wasteful. So the agent must be much more deliberate or strategic about how it takes actions so that it can keep visiting new areas of its environment.

- **Credit Assignment.** When the agent finally does see a reward, it then needs to understand how its past actions contributed toward this outcome. For example if we do walk around randomly and eventually get out of the maze, can we then say that turning left or right at some intersection was a good idea? Attributing some good or bad outcome to the previous actions taken allows us to solve problems with long horizons, which are prevalent in practice.

These challenges are summarized in Figure 2. Actually, addressing one or two of these challenges in isolation is already quite interesting both in theory and practice, and there is a large body of work studying each combination. Almost all of these topics have entire textbooks (and in some cases many textbooks) dedicated to them. We will

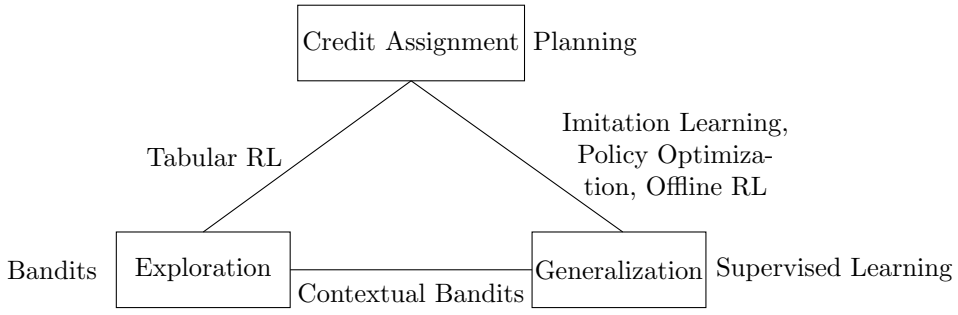

Figure 2: The challenges in reinforcement learning and the sub-topics/sub-problems.

visit essentially all of these problems in this course, building up our techniques from the simpler problems toward the more complex ones. In the last part of the course we will study how to address all three challenges.

2 Probability basics

This course will be quite heavy in statistical reasoning, so let us discuss some of the concepts we will see repeatedly.

Random variables. A probability space Ω is just a set of possible outcomes, and a *random variable* $X : \Omega \to \mathbb{R}$ is a mapping from outcomes to the reals. If P is a distribution over outcomes and $A \subset \mathbb{R}$, we can write $P(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A)$, and we typically use $X \sim P$ to denote that P is the distribution over outcomes. If the range of X is discrete then we define the *probability mass function* $p(x) = P(X = x)$. Otherwise we define the *probability density function* to satisfy $\forall A \subset \mathbb{R} : P(X \in A) = \int_A p(x)dx$. If we have two random variables X, Y over the same outcome space, they have a joint distribution $P(X \in A, Y \in B)$ and joint probability density/mass functions $p(x, y)$, defined in the obvious way. The marginal density for X is defined as $p(x) = \int_{\mathbb{R}} p(x, y)dy$ and the conditional density is $p(y \mid x) = p(x, y)/p(x)$.

An *event* is simply a binary-valued random variable. These can equivalently be described as subsets of the outcome space, $A \subset \Omega$. The probability (under P) of the event is $P(\omega \in A)$.

Often the outcome space Ω is not explicitly described and instead the distribution of the random variable is specified. Even this distribution may be implicit, and I use the notation $\mathbb{P}[\cdot]$ to refer to probabilities of events in such cases. For our purposes, this sort of informality will be quite common so that we don't get caught up in measure-theoretic considerations.

Expectations. If X is a random variable and $g : \mathbb{R} \to \mathbb{R}$ is a function, the expected value is $\mathbb{E}[g(X)] = \int g(X(\omega))dP(\omega) = \int g(x)p(x)dx$ where the integral is replaced by a sum in the discrete case. For a pair of random variables (X, Y) the conditional expectation $\mathbb{E}[Y|X]$ is a random variable, whose value when $X = x$ is given by $\mathbb{E}[Y|X = x] = \int yp(y|x)dy$.

Some useful properties:

1. *Linearity.* $\mathbb{E}[\sum_j c_j g_j(X)] = \sum_j c_j \mathbb{E}[g_j(X)]$, where c_j's are not random variables.

2. *Independence.* If X, Y are independent random variables, meaning that $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for all A, B, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
3. *Iterated Expectation.*

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \int \mathbb{E}[Y|X = x]p(x)dx$$

Variance. The variance of a random variable is simply

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

The variance also inherits some useful properties from the expectation:

1. *Independence.* If $X_1, \ldots, X_n$ are independent, then $\mathrm{Var}[\sum_i a_i X_i] = \sum_i a_i^2 \mathrm{Var}[X_i]$.
2. *Law of total variance.* $\mathrm{Var}[Y] = \mathrm{Var}[\mathbb{E}[Y|X]] + \mathbb{E}[\mathrm{Var}[Y|X]]$

Moment generating function. The "tail behavior" of a random variable (or distribution) describes how likely it is that the random variable takes extreme values, or values far away from its expectation. This will be quite important for us as we discuss how to generalize. We will soon see that the tail behavior is governed by the *moment generating function* which is defined as $M_X(t) = \mathbb{E}[e^{tX}]$.

It's a simple exercise to show that

$$\left.\frac{\partial^k M_X(t)}{\partial t^k}\right|_{t=0} = \mathbb{E}[X^k].$$

The RHS is called the k^{th} non-central moment, which is why the moment generating function is named as it is.

Markov's inequality. To understand tail behavior and generalization, we need two elementary tools:

Proposition 1 (Markov's inequality)**.** *If X is a non-negative random variable, then for any $\epsilon > 0$*

$$\mathbb{P}[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon}$$

Proof. By definition,

$$\mathbb{P}[X \geq \epsilon] = \int_\epsilon^\infty p(x)dx = \int_\epsilon^\infty \frac{x}{x}p(x)dx \leq \frac{1}{\epsilon}\int_\epsilon^\infty xp(x)dx \leq \frac{1}{\epsilon}\int_0^\infty xp(x)dx = \frac{\mathbb{E}[X]}{\epsilon}$$

□

We'll use this to build up some more powerful tools for understanding the fluctuations of random variables.

Union bound. Lastly, we will often have to work with many events that each describe something undesirable happening. We'll want to ensure that nothing undesirable happens, which requires understanding the probability of the union of these bad events. For this, the following bound is useful:

$$\mathbb{P}[\omega \in A \cup B] \leq \mathbb{P}[\omega \in A] + \mathbb{P}[\omega \in B].$$

Of course this bound could be quite loose, but we'll see that in many cases it is good enough.

3 Concentration inequalities

A natural question that comes up in statistics and machine learning is: Suppose that $X_1, \ldots, X_n$ are a sequence of independent and identically distributed random variables with $\mathbb{E}[X_i] = \mu$, $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and we define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. How close is $\bar{X}_n$ to μ?

Foreshadowing, in the machine learning context, the random variable X_i may be the indicator that your classifier made a mistake on some input (aka the "loss"). We want to be sure that given a small amount of data/samples, we can get an accurate measurement of how we expect the classifier to perform in the future.

How to use Markov's inequality. We can answer this question using Markov's inequality. The obstacle is that the random variable $\bar{X}_n - \mu$ may not be non-negative. However we can first square everything and then apply Markov's inequality afterwards:

$$\mathbb{P}[|\bar{X}_n - \mu| \geq \epsilon] = \mathbb{P}[(\bar{X}_n - \mu)^2 \geq \epsilon^2] \leq \frac{\mathbb{E}[(\bar{X}_n - \mu)^2]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

The last line follows by independence and using the properties of variance we discussed above. To see what this bound says about $|\bar{X}_n - \mu|$ it is helpful to re-arrange this inequality as follows: We choose $\delta \in (0,1)$ and want to set $\epsilon = \sqrt{\sigma^2/(n\delta)}$ so that the right hand side is at most δ. Then we can say that with probability at least $1 - \delta$

$$|\bar{X}_n - \mu| \leq \sqrt{\frac{\sigma^2}{n\delta}}$$

Looking only at the dependence on the number of samples n, we see that $\bar{X}_n$ converges to μ at a $1/\sqrt{n}$ rate, which is the correct behavior (e.g., as predicted by the central limit theorem). The dependence on the variance parameter is also sharp. However, the dependence on the "failure probability" parameter δ is quite bad for our application since we will want to take a union bound over many events of this type. Having δ appear polynomially precludes us from doing so.

On the other hand, we made quite weak assumptions about the random variable X, namely that it has finite variance. In most cases, we have much more information we can leverage.

Chernoff method. If we have information about higher moments, it is better to apply Markov's inequality to the moment generating function:

Proposition 2 (Chernoff method)**.**

$$\mathbb{P}[X \geq \epsilon] \leq \inf_{t>0} \exp(-t\epsilon) M_X(t).$$

For many distributions you can often just look up or calculate the moment generating function directly. For example, for Gaussian distribution $\mathcal{N}(0,1)$ we have $M_X(t) = \exp(t^2/2)$ and so Chernoff's method gives

$$\mathbb{P}[X \geq \epsilon] \leq \inf_{t>0} \exp(-t\epsilon + t^2/2) = \exp(-\epsilon^2/2).$$

If we do the same re-arranging as we did previously, we'll see that the tail behavior now depends on $\sqrt{\log(1/\delta)}$ where δ is the failure probability. So this is a much better bound than what we can get from Chebyshev's inequality.

The next result extends this technique to a large family of distributions.

Lemma 3 (Hoeffding's lemma, worse constant)**.** *Let X be a random variable with mean 0 such that $a \leq X \leq b$ almost surely, then*

$$M_X(t) \leq \exp(t^2(b-a)^2/2)$$

Proof. Let us first study a rademacher random variable $\sigma \sim \text{Unif}(\{-1,1\})$:

$$\mathbb{E}[e^{t\sigma}] = \frac{1}{2}\left[e^t + e^{-t}\right] = \frac{1}{2}\left[\sum_{k=0}^{\infty} \frac{t^k}{k!} + \frac{(-t)^k}{k!}\right] = \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{t^{2k}}{2^k k!} = \exp(t^2/2)$$

This is stronger (by a factor of 4) than what we want, but it only applies to Rademacher random variables. For a general bounded random variable X we use *symmetrization*. Let X' denote an independent copy of X, which is also mean 0. Then

$$\begin{aligned}\mathbb{E}e^{tX} &= \mathbb{E}\exp(t(X - \mathbb{E}[X'])) \leq \mathbb{E}_{X,X'}\exp(t(X - X')) = \mathbb{E}_{X,X',\sigma}\exp(t\sigma(X - X')) \\ &\leq \mathbb{E}_{X,X'}\exp(t^2(X-X')^2/2) \leq \exp(t^2(a-b)^2/2)\end{aligned}$$

The first inequality is Jensen's inequality, due to the convexity of $\exp(\cdot)$. Then we use that $X - X'$ and $X' - X$ have the same distribution, so introducing the Rademacher random variable σ has no effect. Finally, thinking of X, X' as fixed, we use the MGF bound for Rademacher r.v.s and then finally plug in the worst case values for X, X'. □

Actually one can prove a sharper version of this lemma with the 2 in the denominator replaced by an 8, but we will not do this here. Combining the sharper version with the Chernoff method gives:

Theorem 4 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ by iid random variables with mean μ such that $a_i \leq X \leq b_i$ almost surely for all i. Then with $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, we have*

$$\mathbb{P}\left[\bar{X}_n - \mu \geq \epsilon\right] \leq \exp\left\{\frac{-2n^2\epsilon^2}{\sum_{i=1}^n (a_i - b_i)^2}\right\}$$

This bound gives the right behavior in terms of both the number of samples n and the failure probability parameter δ.

While there is much more to concentration inequalities than we can cover here, we should be able to get through most of the course and cover the most basic results using Hoeffding's inequality and what we have already seen.

4 PAC learning and uniform convergence

Let us put the probabilistic tools to use in a machine learning setup. In a basic supervised learning task we have

1. An instance space $\mathcal{X}$, consisting of the items/features we want to use to make predictions.
2. A label space $\mathcal{Y}$, which describes the predictions we can make.
3. A loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that we use to measure the quality of our predictions.

We consider a probabilistic setup where there is a distribution P over instance-label pairs (X, Y) and we want find a prediction rule $f : \mathcal{X} \to \mathcal{Y}$ that makes predictions with low loss under P. We measure this via the risk

$$R(f) := \mathbb{E}_{(X,Y)\sim P}\ell(f(X), Y).$$

For example, in binary classification we take $\mathcal{Y} = \{0, 1\}$ and are typically interested in the 0/1 loss $\ell(y, y') = \mathbf{1}\{y \neq y'\}$. In this case, the risk is $R(f) = \mathbb{P}[f(X) \neq Y]$ which is the probability of misclassification.

Unfortunately, in most cases we do not know the distribution P so we cannot compute the function f that minimizes the risk directly. Instead we have access to a training set $S = \{(X_i, Y_i)\}_{i=1}^n$ of labeled examples drawn iid from P. Given the training set, we want to find a good predictor f_S. A natural way to do this is called *empirical risk minimization*. Here we commit to some function class $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ and then we find the function $f \in \mathcal{F}$ with the lowest loss on the training set:

$$\hat{f}_{\text{ERM}} \leftarrow \operatorname*{argmin}_{f\in\mathcal{F}} \underbrace{\frac{1}{n}\sum_{i=1}^n \ell(f(X_i), Y_i)}_{=:\hat{R}_n(f)}$$

A central question in statistical learning is: how well does $\hat{f}_{\text{ERM}}$ *generalize* to the underlying distribution P? The probabilistic tools we have developed already allow us to establish a basic generalization result.

Theorem 5 (Uniform convergence, finite $|\mathcal{F}|$)**.** *Suppose that for all $y, y' \in \mathcal{Y}$ we have $0 \leq \ell(y, y') \leq B$ and that $|\mathcal{F}| < \infty$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$*

$$\forall f \in \mathcal{F} : \quad |\hat{R}_n(f) - R(f)| \leq \sqrt{\frac{B^2 \log(2|\mathcal{F}|/\delta)}{2n}}.$$

Consequently

$$R(\hat{f}_{ERM}) \leq \min_{f\in\mathcal{F}} R(f) + \sqrt{\frac{2B^2 \log(2|\mathcal{F}|/\delta)}{n}}.$$

This theorem states that, given n training samples, we can find a function with similar test performance to the best function in our class, where the sub-optimality is $O(\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}})$. Essentially, minimizing the training error is a good strategy for ensuring good test performance!

However, there are a couple of things to unpack here. First, this bound demonstrates a trade off between *estimation error* and *approximation error*. Indeed, note that we are only comparing to the performance of the best function in our class $\min_{f\in\mathcal{F}} R(f)$. If $\mathcal{F}$ is chosen poorly, or perhaps if it is quite small, then maybe there are no good predictors in the class. Thus we would incur a large amount of *approximation error*, which is what we pay for using a simple function class. One strategy to try to mitigate this is to choose a large function class, but observe that the sub-optimality term scales with $\log|\mathcal{F}|$, so we have higher estimation error with larger function classes. This captures the overfitting phenomenon you may have seen elsewhere, or in practice.

Let us prove this result:

Proof. As the loss is bounded in $[0, B]$ we may apply Hoeffding's inequality. Fix a function $f \in \mathcal{F}$ and observe that $\ell(f(X_i), Y_i)$ has mean $R(f)$ so Hoeffding's inequality directly gives

$$\mathbb{P}[|\hat{R}_n(f) - R(f)| \geq \epsilon] \leq 2\exp\left\{-2n\epsilon^2/B^2\right\}$$

(Note we are taking one union bound here already, as we are controlling both upper and lower tails.) Now, the union bound yields

$$\mathbb{P}[\exists f \in \mathcal{F} : |\hat{R}_n(f) - R(f)| \geq \epsilon] \leq \sum_{f\in\mathcal{F}} \mathbb{P}[|\hat{R}_n(f) - R(f)| \geq \epsilon] \leq 2|\mathcal{F}|\exp\left\{-2n\epsilon^2/B^2\right\}$$

Re-arranging this inequality proves the first statement. The reason we cannot just apply Hoeffding's inequality directly to $\hat{f}_{\mathrm{ERM}}$ is that it is also a random variable dependent on the training set. So $\hat{R}_n(\hat{f}_{\mathrm{ERM}})$ does not decompose into a sum of independent random variables.

For the second statement, let $\bar{f} = \mathrm{argmin}_{f\in\mathcal{F}} R(f)$. Observe that as $\hat{f}_{\mathrm{ERM}}$ minimizes the empirical risk:

$$R(\hat{f}_{\mathrm{ERM}}) \leq \hat{R}_n(\hat{f}_{\mathrm{ERM}}) + \Delta_n \leq \hat{R}_n(\bar{f}) + \Delta_n \leq R(\bar{f}) + 2\Delta_n = \min_{f\in\mathcal{F}} R(f) + 2\Delta_n,$$

where Δ_n is the right hand side of the uniform convergence bound. □

5 Martingales

Later in the course we will deal with learning algorithms that interact with an environment in a sequential feedback driven loop. While concentration will continue to play a central role, we will need some tools to capture random variables with temporal dependencies. The basic object here is an *martingale* which is a (joint distribution over a) sequence of random variables $X_1, X_2, X_3, \ldots$, such that

$$\forall n : \mathbb{E}[X_{n+1} \mid X_{1:n}] = X_n$$

The example to keep in mind is to think of $Y_1, \ldots, Y_n$ as independent random variables and then set $X_n = \sum_{i=1}^n Y_n$ to be the cumulative sums. In this case, the above property says that each of the Y_i's should be mean zero. But this property can hold even if the Y_i's are highly dependent, which we will see happens often in sequential decision making applications.

While martingales are fairly subtle, for our purposes, we will mostly see that they behave quite like independent random variables. Indeed, the Azuma-Hoeffding inequality describes an analogous concentration of measure phenomenon to Hoeffding's inequality in the iid case:

Theorem 6 (Azuma-Hoeffding inequality)**.** *Let $\{X_i\}_{i=0}^n$ be a martingale with $X_0 = 0$ and with increments $|X_i - X_{i-1}| \leq b_i$ almost surely. Then*

$$\mathbb{P}\left[X_n \geq \epsilon\right] \leq \exp\left\{\frac{-\epsilon^2}{2\sum_{i=1}^n b_i^2}\right\}$$

We may use this inequality later in the course.