Chapter 1
Decision-Making

> "*Philosophically minded students of probability nimbly skip among these different ideas* [frequentist and Bayesian]*, and take pains to say which probability concept they are employing at the moment. The vast majority of the practitioners of probability do no such thing. They go on talking of probability, doing their statistics and their decision theory oblivious to all this accumulated subtlety. [...] Extremists of one school or another argue vigorously that the distinction is a sham, for there is only one kind of probability.*"
>
> Hacking (2006, pp. 14)

Humans make decisions constantly: "What to eat for dinner?", "Which university to attend?", "What is a good rule-of-thumb when arriving in a foreign place?", etc. Artificial intelligence (AI) systems can also benefit from human-like decision-making.

To formulate decision-making processes, decision theory has been developed (Wald, 1949). Let $\mathcal{D}$ be ***data*** and f a ***latent variable/parameter***, generated under an *unknown* joint distribution $p(\mathcal{D}, f)$. Furthermore, let $a \in \mathcal{A}$ be the set of all possible ***actions*** that can be taken, and let $u(f, a)$ be a ***utility function*** that measures the "compatibility" of a and f.[1]

Example 1.1. *$\mathcal{D}$ could be a list of symptoms and f a disease. The set of $\mathcal{A}$ contains possible drugs that can be taken. The utility function u indicates the effectiveness of a drug against a disease.*

There are two, non-mutually-exclusive ways to view statistical decision-making: Bayesian and frequentist. As indicated in the epigraph of this chapter, in this text, we accept that they are both valid and useful for different purposes.

1.1 Bayesian decision theory

Bayesian statistics assumes that probabilities represent *beliefs*: "I am 50% sure", "I am quite confident that this will work", etc. Decision-making under this framework is then used to take an action based on one's (e.g., a model's) belief about the unknown. Indeed, ***Bayesian decision***

[1] An alternative view, common in the frequentist formalism, is to replace the utility function with a *loss function*, which, w.l.o.g., can be taken as $-u$.

theory assumes that f is *unknown* and $\mathcal{D}$ is *observed*, and concerns in finding the best action a_* under the *posterior belief* $p(f \mid \mathcal{D})$ and the utility function u:

$$a_* = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \, \mathbb{E}_{f \sim p(f \mid \mathcal{D})}[u(f, a)]. \tag{1.1}$$

Example 1.2. *Under the setting of Example 1.1, a doctor can assess their belief $p(f \mid \mathcal{D})$ of a patient having a disease f after observing the patient's symptoms $\mathcal{D}$. The doctor's utility function u encodes their preferences, e.g. whether they are rather risk-averse or risk-taking. The doctor then prescribe a drug a_* that maximizes their expected utility under posterior belief.*

Example 1.3 (Pascal's Wager). *Let $\mathcal{A} = \{$"Believe in God", "Don't believe in God"$\}$ and let $f \in \{$"God exists", "God doesn't exist"$\}$. Suppose our utility weighs the potential **eternal** "reward" or "punishment" in heaven and hell, respectively. It makes sense, therefore, for someone to have the utility function $u(f, a)$ s.t.:*

- *u("God exists", "Believe in God") $= \infty$,*
- *u("God doesn't exist", "Believe in God") $= a$ where $-\infty < a < 0$; notice that while this is undesirable (negative utility), a is finite,*
- *u("God exists", "Don't believe in God") $= -\infty$,*
- *u("God doesn't exist", "Don't believe in God") $= b$ where $0 < b < \infty$; where the argument here is that we gain something that is finite in our lifetime, but nothing more.*

In this case, the optimal action a_* is to "believe in God", even if a posteriori, *$p($"God exists" $\mid \mathcal{D})$ is very small. Do note that, different utility function will yield different a_*.*

1.2 Frequentist decision theory

Frequentist statistics assumes that probabilities represent *long-running relative frequency*: "What is the occurrence of a disease in given population?", "the error rate of this program is 1%", etc. That is, if we sample the data again and again, what is the proportion of a case of interest. Notice that, here, the data is therefore assumed to be random.

Thus, in contrast to the previous section, ***frequentist decision theory*** assumes that f, while still unknown, is fixed and $\mathcal{D}$ is generated through $\mathcal{D} \sim p(\mathcal{D} \mid f)$. Notice that, here, the roles of $\mathcal{D}$ and f is reversed compared to their roles in the Bayesian decision theory. Since $\mathcal{D}$ is now random, we aim to find an optimal *function* δ_* that maps a realization of data to an action:

$$\delta_* = \underset{\delta}{\operatorname{argmax}} \, \mathbb{E}_{\mathcal{D} \sim p(\mathcal{D} \mid f)}[u(f, \delta(\mathcal{D}))]. \tag{1.2}$$

or, equivalently,

$$\delta_* = \underset{\delta}{\operatorname{argmin}} \, \mathbb{E}_{\mathcal{D} \sim p(\mathcal{D} \mid f)}[\ell(f, \delta(\mathcal{D}))], \tag{1.3}$$

where $\ell = -u$ is the so-called ***loss function***. That is, we want to find the best "policy" δ_* that works on various situations $\mathcal{D} \sim p(\mathcal{D} \mid f)$. The function u is interpreted as measuring how good it is to do an action $\delta(\mathcal{D})$ under the data $\mathcal{D}$ given that f is the underlying parameter that generates $\mathcal{D}$.

Example 1.4. *Under the setting of Example 1.1, suppose a public health organization wants to recommend a policy/rule—a "what-to-do" guideline—for the general population. The goal is then to find a policy δ_* such that when presented with a set of symptoms $\mathcal{D}$, it recommends a drug a for treating the underlying, unknown disease f.*

However, notice that u depends on f which we have assumed to be unknown. It follows that we cannot even compute $u(f, \delta(\mathcal{D}))$ and thus we cannot perform the maximization. We do not have such a problem in the Bayesian case since we have a belief about f.

To circumvent this issue, we need to take f out of the equation. One way to do so is as follows. Let $R(f, \delta) = \mathbb{E}_{\mathcal{D} \sim p(\mathcal{D}|f)}[\ell(f, \delta(\mathcal{D}))]$ be the ***risk*** function. Then, we find the optimal decision function δ_* that minimizes the worst-case risk:

$$\delta_* = \operatorname*{argmin}_{\delta} \max_{f} R(f, \delta). \tag{1.4}$$

Continuing the previous example, the minimax decision function has the interpretation that it is the one that is optimal under the worst-case risk when we consider various plausible alternatives of the underlying diseases f.

Example 1.5. *If we think f could be "cold", "malaria", and "COVID-19", then we want to provide a treatment guideline that would be relatively effective for every possible f. Note that, δ_* might not be the best possible guideline for each individual f.*

Remark 1.6. Should you be Bayesian or frequentist? *Both!* Hopefully the epigraph and the discussion in this chapter convinced you that being both is the correct move. They are different tools, for different situations and goals.

Chapter 2
Concentration Inequalities

Let X be a random variable (r.v.). It is often useful to know how such a r.v. *concentrated* around a value. ***Concentration inequalities*** are the bread-and-butter for theoretical analyses in machine learning (and other fields!).

Example 2.1. *Let X indicate the grade of a random student in a given population (e.g. in a class). A concentration inequality is useful to answer the following question: "What is the prevalence of students with a grade at least (e.g.)* 80*?"*

Let $\mathbb{P}$ denote the long-running relative frequency of its random *event* argument. That is, let $(X_i)_{i=1}^n$ be a sequence of samples of X and A be an event that depends on X such as $X \geq a$ for some value a, then we can write $\mathbb{P}(A) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \mathbb{I}[A_i]$, where $\mathbb{I}[\cdot]$ denotes the indicator function, which equals one if its event argument happens, and zero otherwise. We call $\mathbb{P}(A)$ the ***probability*** of observing the event A.

Example 2.2. *Let X indicate the grade of a random student in a given population. Let Z be a random event indicating $X \geq 80$, i.e. the event when a random student has a grade greater than or equals* 80*. Then $\mathbb{P}(Z)$ indicates the proportion of students with grade ≥ 80 when we pick and measure people at random many times.*

Remark 2.3. *The notion of probability here differs from Bayesian statistics, where it denotes* degree of beliefs *about some event.*

Here, we will describe some useful concentration inequality. The proofs are omitted since they are standard and we focus on their applications.

Theorem 2.4 (Markov's Inequality). Let X be a nonnegative r.v. and assume that $\mathbb{E}(X)$, the expected value of X exists (i.e., $\mathbb{E}(X) < \infty$). Then,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a} \tag{2.1}$$

for all $a > 0$. Moreover,

$$\mathbb{P}(X \geq b\mathbb{E}(X)) \leq \frac{1}{b} \tag{2.2}$$

for all $b > 1$.

Markov's inequality is useful if we only know the expected value $\mathbb{E}(X)$ of X.

Example 2.5. *Continuing Example 2.2, suppose we know that on average, students have grade* 50*. Then, the proportion of people in the population weighing* ≥ 80 *is at most:* $\mathbb{P}(X \geq 80) \leq \frac{50}{80} = 0.625$.

However, notice that the bound on the probability decreases *linearly* in a. We can obtain a tighter bound if we know further information about X, namely its *variance* $\mathrm{Var}(X)$.

Theorem 2.6 (Chebyshev's Inequality). Let X be a r.v. and assume that both $\mathbb{E}(X)$ and $\mathrm{Var}(X)$ exist. Then,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2} \tag{2.3}$$

for all $a > 0$. Moreover,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq b|\mathbb{E}(X)|) \leq \frac{\mathrm{Var}(X)}{b^2(\mathbb{E}(X))^2} \tag{2.4}$$

for all $b > 1$.

Chebyshev's inequality is useful to bound the proportion of some measurements deviating from the population's mean.

Example 2.7. *Let* X *and* $\mathbb{E}(X)$ *be as in Example 2.5. Suppose* $\mathrm{Var}(X) = 10$*. Then,*

$$\mathbb{P}(|X - 50| \geq 30) \leq \frac{10}{30^2} = 0.0111.$$

That is, it is quite rare that a random student's grade deviates by 30 *or more from the average in a class with* $\mathbb{E}(X) = 50$ *and* $\mathrm{Var}(X) = 10$*.*

Recall that Chebyshev's inequality is tighter than Markov's inequality and they differ in how they use the moments (mean, variance) of the r.v. It is thus logical to ask: Can we get a tighter bound if we consider higher moments? The answer is the Chernoff bound. First, let us state its special version for Bernoulli random variables.

Theorem 2.8 (Chernoff Bound—Bernoulli). Let $(X_i)_{i=1}^n$ be independent Bernoulli r.v.s. with their respective expectations $(p_i)_{i=1}^n$. Let $X = \sum_{i=1}^n X_i$. Then,

(i) For every $0 < \delta \leq 1$, it holds that

$$\mathbb{P}(X \geq (1+\delta)\mathbb{E}(X)) \leq \exp(-\mathbb{E}(X)\delta^2/3). \tag{2.5}$$

(ii) For every $0 < \delta < 1$, it holds that

$$\mathbb{P}(X \leq (1-\delta)\mathbb{E}(X)) \leq \exp(-\mathbb{E}(X)\delta^2/2). \tag{2.6}$$

Notice that the bound of (this version) of Chernoff bound is exponential in the constant a. Let us compare the "strength" of Markov's, Chebyshev's, and Chernoff's bounds in the following example.

Example 2.9. *Consider $(X_i)_{i=1}^n$ be n independent tosses of a fair coin—$X_i = 1$ if head and $X_i = 0$ otherwise. Let $X = \sum_{i=1}^n X_i$ denote the number of heads we see. It's expected value is thus $\mathbb{E}(X) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i) = \frac{n}{2}$. We would like to see the frequency of the event where the number of heads $\geq \frac{3}{4}n$. With Markov's inequality, we see that*

$$\mathbb{P}\left(X \geq \frac{3}{4}n\right) \leq \frac{n/2}{(3/4)n} = \frac{2}{3}.$$

Notice the constant bound. Meanwhile, with Chebyshev's inequality, we obtain

$$\mathbb{P}\left(X \geq \frac{3}{4}n\right) \leq \mathbb{P}\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq \frac{\mathrm{Var}(X)}{(\frac{n}{4})^2} = \frac{n/4}{n^2/16} = \frac{4}{n}.$$

by noting that $\mathrm{Var}(X) = \frac{n}{4}$. This bound is indeed stronger than Markov's since it decreases as n increases. Finally, for the Chernoff bound, we let $\delta = 1/2$ since then $(1+\delta)\mathbb{E}(X) = \frac{3}{4}n$, and obtain

$$\mathbb{P}\left(X \geq \frac{3}{4}n\right) \leq \exp(-\mathbb{E}(X)\delta^2/3) = \exp\left(-\frac{n}{2}\frac{1}{4}\frac{1}{3}\right) = \exp(-n/24).$$

Notice that the Chernoff bound decreases exponentially in the number of tosses.

The following is the general version of the Chernoff bound. Recall that $M(t) = \mathbb{E}(\exp(tX))$ is the moment-generating function of X.

Theorem 2.10 (Chernoff Bound). Let X be a random variable and let a be an arbitrary value of X. Then,

(i) For every $t > 0$, it holds that

$$\mathbb{P}(X \geq a) \leq \mathbb{E}(\exp(tX))\exp(-ta). \tag{2.7}$$

(ii) For every $t < 0$, it holds that

$$\mathbb{P}(X \leq a) \leq \mathbb{E}(\exp(tX))\exp(-ta). \tag{2.8}$$

Next, we have a similar, exponentially decreasing bound in the form of ***Hoeffding's inequality*** which requires us to know the upper and lower bounds of the values of the r.v.s.

Theorem 2.11 (Hoeffding's Inequality). Let $(X_i)_{i=1}^n$ be i.i.d. r.v.s. with mean μ, where for each i we have $l \leq X_i \leq h$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ be their sample mean. Then,

$$\mathbb{P}\left(|\bar{X} - \mu| \geq a\right) \leq 2\exp\left(-\frac{2na^2}{(h-l)^2}\right), \tag{2.9}$$

for all $a > 0$.

2.1 Gaussian Tail Bounds

For Gaussian random variables, we have the following theorem (Srinivas et al., 2010):

Theorem 2.12 (Gaussian Tail Bound). Let X be a Gaussian r.v. with mean μ and variance σ^2. For any $\beta > 0$,

$$\mathbb{P}(|X - \mu| \geq \beta\sigma) \leq \exp\big(-\beta^2/2\big). \tag{2.10}$$

Here is another useful property for Gaussian r.v.s. with nonpositive means:

Theorem 2.13 (Gaussian Tail with Nonpositive Mean). Let X be a Gaussian r.v. with mean $\mu \leq 0$ and variance σ^2. Then,

$$\mathbb{E}(X\,\mathbb{I}(X \geq 0)) = \frac{\sigma}{\sqrt{2\pi}} \exp\left(\frac{-\mu^2}{2\sigma^2}\right). \tag{2.11}$$

The expression $X\mathbb{I}(X \geq 0)$ means we are looking at the Gaussian r.v. X where it takes values ≥ 0 and ignore everywhere else.

2.2 Other Useful Inequalities

In theoretical analysis, concentration inequalities are often paired with other inequalities. Here, we shall see some of the commonly-used useful inequalities. The simplest is the ***union bound***.

Theorem 2.14 (Union Bound). Let $(A_i)_{i=1}^n$ be a sequence of random events. Then,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i). \tag{2.12}$$

Example 2.15. *Suppose the probability of a student getting a perfect* 100 *grade is at most* 0.001. *Denote* A_i *to be the event a student* i *gets grade* 100. *Then the probability of at least one student obtaining the perfect grade in a class of size* 50 *is*

$$\mathbb{P}\left(\bigcup_{i=1}^{12} A_i\right) \leq \sum_{i=1}^{50} \mathbb{P}(A_i) \leq \sum_{i=1}^{50} 0.001 = 0.05.$$

That is, there is at most 5% *chance/relative frequency that a student will get* 100 *in this setting.*

Another useful inequality is ***Jensen's inequality*** which allows us to swap an expectation operator with a convex/concave function.

Theorem 2.16 (Jensen's Inequality). Let X be a random variable taking value in $\mathbb{R}^n$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex or concave function. Then,

(i) if f is **convex**: $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$,
(ii) if f is **concave**: $f(\mathbb{E}(X)) \geq \mathbb{E}(f(X))$.

Moreover, both inequalities also hold for empirical means.

Chapter 3
Frequentist Bandits

In K-armed bandit problem, we have K different actions $a_t \in \mathcal{A} := \{1, \ldots, K\}$ we can perform at each time step $t = 1, \ldots, T$. After performing an action $a \in \mathcal{A}$, the we observe a reward value $r(a) \in [0, 1]$ distributed as an *unknown* reward distribution $p(r \mid a)$.

Let $\mu(a) = \mathbb{E}(r(a))$ be the unknown expected reward of action a. Let us also denote $a_* = \text{argmax}_{a \in \mathcal{A}} \mu(a)$ to be the action with the highest expected reward. We can define

$$R_T = \sum_{t=1}^{T} r(a_*) - r(a_t), \tag{3.1}$$

called the ***regret*** over a run of an algorithm where we select a sequence of actions $(a_t)_{t=1}^T$. This measures "how far away" our actions deviate from the optimal actions.

Since each a_t in (3.1) is a random variable that depends on an algorithm's run, R_T is also a r.v. Thus, it makes sense to study the ***expected regret***

$$\mathbb{E}(R_T) = \sum_{t=1}^{T} \mu(a_*) - \mu(a_t) = T\mu(a_*) - \sum_{t=1}^{T} \mu(a_t). \tag{3.2}$$

Ideally, an algorithm has ***no regret***, i.e., $\lim_{T \to \infty} \mathbb{E}(R_T)/T = 0$. Our goal is to construct an algorithm for picking sequences of actions that minimize the expected regret and asymptotically have no regret. The algorithm shall leverage *frequentist* technique, e.g. using the sample mean to estimate μ and making a decision based on this estimate.

3.1 Explore-Then-Exploit

The simplest algorithm is to explore for $NK < T$ rounds and exploit for the remaining $T - N$ rounds. Exploration here means that we try each action N times. Meanwhile, exploitation means that we use our estimate of the expected reward of each action, $\hat{\mu}(a) = 1/N \sum_{t=1}^{N} r_t(a)$, to pick our estimate of best action $\hat{a}_* = \text{argmax}_{a \in \mathcal{A}} \hat{\mu}(a)$, and always pick this action. The algorithm is summarized in Algorithm 1.

Theorem 3.1. *With $N = (T/K)^{2/3}(\log T)^{1/3}$, the explore-then-exploit algorithm has regret of*

$$\mathbb{E}(R_T) \leq \mathcal{O}\Big((K T^2 \log T)^{1/3}\Big) \qquad \textit{with probability} \geq 1 - \frac{2K}{T^4}.$$

That is, it has no regret as $T \to \infty$ with high probability.

Algorithm 1 Explore-Then-Exploit

Input: Time horizon T, set of K actions $\mathcal{A}$, number of tries per action N.
Output: Cumulative reward r_{total}

$r_{\text{total}} = 0$
for $t = 1, \dots, N$ **do**
  for all $a \in \mathcal{A}$ **do**
    $r_{ta} = \texttt{do\_action}(a)$
    $r_{\text{total}} = r_{\text{total}} + r_{ta}$
  end for
  $\hat{\mu}(a) = \frac{1}{N} \sum_{t=1}^{N} r_{ta}$ for each $a \in \mathcal{A}$
  $\hat{a}_* = \text{argmax}_{a \in \mathcal{A}} \hat{\mu}(a)$
  for $t = 1, \dots, T - NK$ **do**
    $r_{ta} = \texttt{do\_action}(\hat{a}_*)$
    $r_{\text{total}} = r_{\text{total}} + r_{ta}$
  end for
end for
return r_{total}

Proof. Let $\hat{\mu}(a) = 1/N \sum_{t=1}^{N} r_t(a)$ be empirical average reward of action a. Define $\varepsilon = \sqrt{(2 \log T)/N}$. Also, define an event $E = \{|\hat{\mu}(a) - \mu(a)| \leq \varepsilon; \forall a \in \mathcal{A}\}$ where all actions' estimates are within ε distance to the respective true values.

Assume that E holds. Recall that $\hat{a}_* = \text{argmax}_{a \in \mathcal{A}} \hat{\mu}(a)$ and $a_* = \text{argmax}_{a \in \mathcal{A}} \mu(a)$. So, by definition,

$$\hat{\mu}(\hat{a}_*) \geq \hat{\mu}(a_*) \quad \text{and} \quad \mu(\hat{a}_*) \leq \mu(a_*).$$

Now, since E holds, $\hat{\mu}(\hat{a}_*) - \mu(\hat{a}_*) \leq \varepsilon$ and $\mu(a_*) - \hat{\mu}(a_*) \leq \varepsilon$. (Notice the absolute value in $\mathbb{E}$.) And so,

$$\mu(\hat{a}_*) + \varepsilon \geq \hat{\mu}(\hat{a}_*) \quad \text{and} \quad \hat{\mu}(a_*) \geq \mu(a_*) - \varepsilon.$$

Altogether they imply

$$\begin{aligned} &\mu(\hat{a}_*) + \varepsilon \geq \hat{\mu}(\hat{a}_*) \geq \hat{\mu}(a_*) \geq \mu(a_*) - \varepsilon \\ \iff &\mu(\hat{a}_*) + \varepsilon \geq \mu(a_*) - \varepsilon \\ \iff &2\varepsilon \geq \mu(a_*) - \mu(\hat{a}_*). \end{aligned}$$

Hence, we have $\mu(a_*) - \mu(\hat{a}_*) \leq 2\sqrt{(2 \log T)/N}$. This is the bound on the regret during the exploitation phase assuming that E holds. The upper bound on the regret during the exploration phase is trivially NK since $\mu(\cdot) \in [0, 1]$. Thus, under E,

$$\mathbb{E}(R_T) \leq NK + \sum_{t=1}^{T-NK} 2\sqrt{\frac{2 \log T}{N}} = NK + 2(T - NK)\sqrt{\frac{2 \log T}{N}}.$$

Substituting $N = (T/K)^{2/3} (\log T)^{1/3}$, we obtain $\mathbb{E}(R_T) \leq \mathcal{O}((K T^2 \log T)^{1/3})$. It is clear that $\lim_{T \to \infty} \mathbb{E}(R_T)/T = 0$ since $(T^2 \log T)^{1/3}/T = (\log T)/T$.

Now we compute the probability of the event E. Using Hoeffding's inequality (Theorem 2.11), we can bound the deviation $|\hat{\mu}(a) - \mu(a)|$ of our estimate to the true expected reward of action a:

$$\mathbb{P}(|\hat{\mu}(a) - \mu(a)| \geq \varepsilon) \leq 2\exp\left(-2N\varepsilon^2\right) = \frac{2}{T^4}.$$

The complement of E is $E^c = \{|\hat{\mu}(a) - \mu(a)| \geq \varepsilon; \exists a \in \mathcal{A}\}$. By the union bound (Theorem 2.14), we have

$$\mathbb{P}(E^c) = \mathbb{P}\left(\bigcup_{a=1}^{K} \{|\hat{\mu}(a) - \mu(a)| \geq \varepsilon\}\right) \leq \frac{2K}{T^4}.$$

So, $\mathbb{P}(E) = 1 - \mathbb{P}(E^c) \geq 1 - \frac{2K}{T^4}$. This is the probability of attaining the regret bound below. Note that we can ignore the event E^c since it occurs with such a low probability. □

Remark 3.2. *In summary, the proof strategy boils down to*

1. *defining a event E that encompass "nice" properties for our analysis,*
2. *bounding the expected regret $\mu(a_*) - \mu(a_t)$ at each time step t under E,*
3. *extending it to the bound of the cumulative expected regret $\mathbb{E}(R_T)$ by summing them,*
4. *reasoning about its expected value using the union bound and concentration inequality,*
5. *arguing that the probability of the event E is high.*

To get the value for N in the hypothesis, one can aim to solve for the bound w.r.t. N s.t. the bound is minimized (i.e. tighter). If we only care about showing the no-regret property, we can pick N such the bound is sublinear *in T. Because, then, $\mathbb{E}(R_T)$ will grow slower than T and thus $\mathbb{E}(R_T)/T$ will converge to $0 \implies$ no regret.*

3.2 Upper Confidence Bound (UCB)

Let us now consider the following decision rule: At each time $t = 1, \ldots, T$, we pick an action that maximizes the function

$$\mathrm{UCB}_t(a) = \mu_t(a) + \sqrt{(2\log T)/N_t(a)}, \tag{3.3}$$

where $\mu_t(a) = 1/N_t(a) \sum_{i=1}^{t} r_i(a_i)\mathbb{I}(a_i = a)$ is the empirical mean estimate of $\mu(a)$ after t rounds, and $N_t(a) = \sum_{i=1}^{t} \mathbb{I}(a_t = a)$ is the number of times the action a has been selected. The algorithm is summarized in Algorithm 2.

Intuitively, we maintain both our estimate of μ in the form of μ_t, *and* our "confidence"—not to be confused with the definition of confidence in the Bayesian setting—about that estimate. This "confidence" is essentially an error bar around μ_t, the standard error around the sample mean. If our estimate of an action is high and the error bar is wide, we will therefore tend to pick that action (exploration). As t increases, the values of N_t will increase, and hence the error bars will decrease. We can then be confident that our estimate μ_t is very close to μ and we can simply pick the best action every time (exploitation).

Algorithm 2 UCB

Input: Time horizon T, set of K actions $\mathcal{A}$
Output: Cumulative reward r_{total}

$r_{\text{total}} = 0$
for $t = 1, \dots, T$ **do**
  Count $N_t(a)$ for each $a \in \mathcal{A}$
  Compute $\mu_t(a)$ for each $a \in \mathcal{A}$
  $a_t = \text{argmax}_{a \in \mathcal{A}} \mu_t(a) + \sqrt{(2 \log T)/N_t(a)}$
  $r_{ta} = \texttt{do\_action}(a_t)$
  $r_{\text{total}} = r_{\text{total}} + r_{ta}$
end for
return r_{total}

Theorem 3.3. *At each round $t = 1, \dots, T$, the UCB algorithm has expected regret of*

$$\mathbb{E}(R_t) \leq \mathcal{O}\Big(\sqrt{Kt \log T}\Big) \qquad \textit{with probability} \geq 1 - \frac{2K}{T^3}.$$

Thus, the UCB algorithm has no regret w.h.p.

Proof. Define $\varepsilon_t(a) = \sqrt{(2 \log T)/N_t(a)}$. Suppose the event $E = \{|\hat{\mu}_t(a) - \mu(a)| \leq \varepsilon_t(a); \forall a \in \mathcal{A}, \forall t = 1, \dots, T\}$ holds. Let a_* and a_t be the (unknown) optimal arm and the selected arm at time t, respectively. Since a_t is selected at time t, then by the algorithm, $\text{UCB}_t(a_t) \geq \text{UCB}_t(a_*)$. Since E holds, $\mu(a_t) + \varepsilon_t(a_t) \geq \hat{\mu}(a_t)$. Moreover, by definition, $\text{UCB}_t(a_*) \geq \mu(a_*)$. Therefore,

$$\mu(a_t) + 2\varepsilon_t(a_t) \geq \hat{\mu}(a_t) + \varepsilon_t(a_t) = \text{UCB}_t(a_t) \geq \text{UCB}_t(a_*) \geq \mu(a_*).$$

Rearranging, we have

$$\Delta_t(a_t) := \mu(a_*) - \mu(a_t) \leq 2\varepsilon_t(a_t) = 2\sqrt{(2 \log T)/N_t(a_t)}.$$

We will use this bound to obtain the bound for $\mathbb{E}(R_t)$.

Since we pick a single action at each time step, first we note that $t = \sum_{a \in \mathcal{A}} N_t(a)$. Moreover, the expected total regret $\mathbb{E}(R_t)$ can be decomposed over actions:

$$\begin{aligned}
\mathbb{E}(R_t) &= \sum_{i=1}^{t} \Delta_t(a_i) = \sum_{a \in \mathcal{A}} \sum_{j=1}^{N_t(a)} \Delta_t(a) \\
&= \sum_{a \in \mathcal{A}} 2\sqrt{(2 \log T)/N_t(a)} N_t(a) \\
&= 2\sqrt{(2 \log T)} \sum_{a \in \mathcal{A}} \sqrt{N_t(a)}.
\end{aligned}$$

Now, notice that $\sqrt{\cdot}$ is a concave function. By Jensen's inequality, we can then bound the average of $\sqrt{N_t}$ by (recall that $|\mathcal{A}| = K$)

$$\frac{1}{K}\sum_{a\in\mathcal{A}}\sqrt{N_t(a)} \leq \sqrt{\frac{1}{K}\sum_{a\in\mathcal{A}} N_t(a)} = \sqrt{\frac{t}{K}}.$$

This implies that $\sum_{a\in\mathcal{A}}\sqrt{N_t(a)} \leq K\sqrt{t/K} = \sqrt{Kt}$

Therefore, we can bound $\mathbb{E}(R_t)$ by

$$\mathbb{E}(R_t) \leq 2\sqrt{2}\sqrt{\log T}\sqrt{Kt} = \mathcal{O}\Big(\sqrt{Kt\log T}\Big).$$

Taking $t = T$, we clearly see that $\mathbb{E}(R_T)$ is sublinear. Thus the UCB algorithm has no regret.

The last thing we need to show is the probability that the results above hold. I.e., we want to show that E holds with high probability. By Hoeffding's inequality and subsituting in $\varepsilon_t(a)$, we obtain $\mathbb{P}(|\hat{\mu}_t(a) - \mu(a)| \geq \varepsilon_t(a)) \leq 2/T^4$. Then, by the union bound over a and t, we obtain $\mathbb{P}(E^c) \leq (2KT)/T^4$. Therefore, $\mathbb{P}(E) \geq 1 - 2K/T^3$. That is, our analysis below will hold with high probability.

□

Chapter 4
Gaussian Processes

Let $f : \mathcal{X} \to \mathbb{R}$ be a function. When X is finite, one can think of f as a collection of function values $(f(x))_{x\in\mathcal{X}}$ computed across ***evaluation/context points*** $\mathcal{X}$. The same intuitive image can be useful to think of f in the infinite case.

A ***Gaussian process (GP)*** can be seen as a probability distribution on a function space $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R}\}$. The defining property of a GP is that any finite collection of evaluation points $(x_i)_{i=1}^n \subset \mathcal{X}$, the probability distribution over $(f(x_i))_{i=1}^n$ is *multivariate Gaussian*. A GP is fully characterized by its ***mean function*** $\mu : \mathcal{X} \to \mathbb{R}$ and its ***covariance function*** $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

The covariance function, expressed through a (positive-definite) ***kernel*** $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the property it is symmetric in its two arguments and

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0 \tag{4.1}$$

holds for all $(x_i \in \mathcal{X})_{i=1}^n$, $(c_i \in \mathbb{R})_{i=1}^n$, and $n \in \mathbb{N}$. The latter can be expressed through linear algebra: Let $(\boldsymbol{K})_{ij} = k(x_i, x_j)$ be the matrix with coefficients equal all evaluations of k under $(x_i)_{i=1}^n$. Then (4.1) is equivalent as saying that $\boldsymbol{K}$ is positive semi-definite.

An example of commonly-used covariance functions is the ***Matèrn kernel*** with smoothness parameter ν. This class of kernels induces a GP over space of functions that is up to k-times differentiable for $k < \nu$. So, with $\nu = 5/2$, the GP is over space of functions that are twice differentiable. Another example is the ***radial basis function (RBF) kernel***, also known as the ***squared exponential kernel***. This can be seen as a the limit of the Matérn kernel when $\nu \to \infty$. It thus induces a GP on C^∞. See standard Gaussian process textbooks, e.g. Williams & Rasmussen (2006), for definitions.

4.1 Posterior Inference

GPs are useful to make prediction about an unknown function f. Let $\mathcal{D} := \{(x_i, f(x_i))\}_{i=1}^n$ be a dataset. Assuming a GP prior[1] $p(f) = \mathcal{GP}(0, k)$ over f, the GP posterior is described through the updated mean and covariance functions $\mu(\cdot \mid \mathcal{D}) : \mathcal{X} \to \mathbb{R}$ and $k(\cdot, \cdot \mid \mathcal{D}) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, respectively. They are characterized by

$$\mu(x \mid \mathcal{D}) = k(x, X)\big(k(X, X) + \sigma_n^2 I\big)^{-1} Y \tag{4.2}$$

[1] In practical applications, μ is often simply set to the zero function.

$$k(x \mid \mathcal{D}) = k(x,x) - k(x,X)\big(k(X,X) + \sigma_n^2 I\big)^{-1} k(X,x), \tag{4.3}$$

where we have defined shorthands $X := (x_i)_{i=1}^n$ and $Y := (f(x_i))_{i=1}^n \in \mathbb{R}^n$. Also, $k(x,X)$, $k(X,X)$, and $k(X,x)$ are the matrix representations of the kernel under those evaluation points. Finally, $\sigma_n^2 > 0$ is a measurement noise assumed in evaluating $f(x)$, i.e. $y = f(x) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$.

4.2 Reproducing Kernel Hilbert Space

As mentioned before, a GP defines a probability distribution on a function space. What exactly is that function space? Inspecting (4.2), we see that $\mathcal{GP}(0,k)$ describes a set of posterior means

$$x \mapsto \sum_{i=1}^n \alpha_i k(x_i, x) \qquad \text{where } \alpha_i = \big(k(X,X) + \sigma_n^2 I\big)^{-1} Y, \tag{4.4}$$

for under all possible dataset $\mathcal{D}$. Note that this set of functions is fully characterized by the choice of the kernel of a GP. Indeed, $(k(x_i,\cdot) \in \mathbb{R}^n)_{i=1}^n$, seen as vectors, act as a basis of the resulting functions. This basis vectors vary depending on the evaluation points X.

We define the ***reproducing kernel Hilbert space (RKHS)*** $\mathcal{H}_k$ of $\mathcal{GP}(0,k)$ to be the completion of the space of functions above. It is endowed with the inner product

$$\langle f, f' \rangle = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \alpha'_j k(x_i, x'_j) \tag{4.5}$$

for $f = \sum_{i=1}^n \alpha_i k(x_i,\cdot)$ and $f' = \sum_{j=1}^m \alpha'_j k(x'_j,\cdot)$.

The RKHS inner product induced a norm $\|\cdot\|_{H_k}$ that tells us about the "complexity" of a function in H_k. Under this norm, we can define the ***RKHS ball*** of radius r by

$$\mathcal{H}_k[b] := \{f \in H_k \text{ such that } \|f\|_{H_k} \le b\}, \tag{4.6}$$

which contains all possible GP posterior means under a kernel k with "complexity" at most b.

4.3 Information Capacity

Since GPs are useful for learning an unknown function f through a dataset $\mathcal{D}$, it is useful to know how well we can learn f with a GP prior $\mathcal{GP}(0,k)$ through noisy observations of f with noise variance σ_n^2. This notion is termed ***information capacity***. Intuitively, the information encoded in the GP prior through the covariance function k determines the information content of f, while the noise level σ_n^2 limits the amount of information provided by observations.

The information regarding f expressed through $\mathcal{D}$ can be described by the ***mutual information***, also known as the ***information gain***:

$$\mathrm{MI}(Y, f) := \frac{1}{2} \log\det(I + \sigma_n^{-2} K(X,X)). \tag{4.7}$$

The information capacity is then defined as the maximum information gain through a dataset $\mathcal{D} = (X, Y)$ of size T:

$$\gamma_T(f) := \sup_{|\mathcal{D}|=T} \mathrm{MI}(Y, f). \tag{4.8}$$

As a motivating example, $\mathcal{D}$ could be obtained through a sequential decision-making process, and we want to know how well we have learned about an unknown function f under some observation noise σ_n^2 after T steps. If the function f is clear from the context, one can also simply write this quantity as γ_T.

For compact $\mathcal{X} \subset \mathbb{R}^d$ and a fixed σ_n, we have the following, depending on the covariance function k:

- Matérn with smoothness parameter ν: $\gamma_t = \mathcal{O}(T^\alpha (\log T)^{1-\alpha})$ where $\alpha = d/(2\nu + d)$.
- RBF: $\gamma_T = \mathcal{O}((\log T)^{d+1})$.

See Srinivas et al. (2010) for the detailed discussion. The intuition is as follows: The smoother the function f is (i.e., as ν increases), the less information we gain through new data points, since we can already easily predict the function values on the other regions of $\mathcal{X}$. Put another way, smooth functions have less "surprise".

The following result is an important application of the maximum information gain. We will use it extensively in the subsequent chapters. Suppose we have selected T observations at context points $(x_t)_{t=1}^T$. Let $(\sigma_t^2(x_t))_{t=1}^T$ be the predictive variance of x_t's under the GP at each time step t. Through the chain rule for mutual information, the information capacity (4.8) can be written as

$$\mathrm{MI}(Y, f) = \frac{1}{2} \sum_{t=1}^{T} \log\left(1 + \frac{\sigma_t^2(x_t)}{\sigma_n^2}\right). \tag{4.9}$$

Theorem 4.1 (Srinivas et al., 2010). Given $m \in \mathbb{R}$, let $k(x, x) \leq m$ for all $x \in \mathcal{X}$. Then $\sum_{t=1}^T \sigma_t^2(x_t) = \mathcal{O}(\gamma_T)$. More specifically, $\sum_{t=1}^T \sigma_t^2(x_t) \leq \frac{2m}{\log(1+\sigma_n^{-2}m)} \gamma_T$.

4.4 Useful Inequalities

The following result, known as (some variant of) the Borell-TIS inequality (Van Der Vaart et al., 1996), is useful to bound the frequency of the supremum of GP sample paths.

Theorem 4.2 (Borell-TIS Inequality). Let $\mathcal{X}$ be a topological space and let $f \sim \mathcal{GP}(0, k)$ be a sample path of a centered Gaussian process on $\mathcal{X}$. If $\sup_{x\in\mathcal{X}} |f(x)|$ finite, then for every $\lambda > 0$,

$$\mathbb{P}(\sup_{x\in\mathcal{X}} |f(x)| \geq \lambda) \leq 2 \exp\left(\frac{-\lambda^2}{8\mathbb{E}(\sup_{x\in\mathcal{X}} |f(x)|)^2}\right). \tag{4.10}$$

Chapter 5
Discrete Bayesian Optimization

To start off, we assume the search space (action space in the bandit lingo) $\mathcal{X}$ is finite. This is practically very relevant, e.g. in drug and materials discovery applications.

In ***Bayesian optimization (BO)***, we want to (w.l.o.g.) maximize an *unknown* function $f : \mathcal{X} \to \mathbb{R}$.[1] This implies that the maximizer $x_* = \text{argmax}_{x \in \mathcal{X}} f(x)$ and is also unknown. While we do not know f holistically, we assume we can *evaluate* $f(x)$ for any $x \in \mathcal{X}$. Note however that this evaulation is in general very costly and we want to find the maximium with as few evaluations as possible.

Since f is unknown, we define a prior $p(f) = \mathcal{GP}(\mu, k)$ with a mean function $\mu : \mathcal{X} \to \mathbb{R}$ and kernel/covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. At each iteration $t = 1, \ldots, T$, a BO algorithm will select an evaluation point[2] $x_t \in \mathcal{X}$ through an acquisition function $\alpha(x; D_t) = \mathbb{E}_{p(f|\mathcal{D}_t)}(u(x, f))$ where u is a utility function and $p(f \mid \mathcal{D}_t)$ is the posterior belief over f after observing previously gathered data points $\mathcal{D}_t = \{(x_i, f(x_i))\}_{i=1}^{t-1}$. More specifically, the algorithm will select $x_t = \text{argmax}_{x \in \mathcal{X}} \alpha(x; D_t)$ and evaluate $f(x_t)$. This process is repeated until termination at time T; see Algorithm 3

Algorithm 3 Discrete GP-UCB for BO

Input: Time budget T, GP prior $\mathcal{GP}(\mu, k)$, unknown function f
Output: Maximum of f found after T steps

for $t = 1, \ldots, T$ **do**
  Count $N_t(a)$ for each $a \in \mathcal{A}$
  Compute $\mu_t(a)$ for each $a \in \mathcal{A}$
  $a_t = \text{argmax}_{a \in \mathcal{A}} \mu_t(a) + \sqrt{(2 \log T)/N_t(a)}$
  $r_{ta} = \texttt{do\_action}(a_t)$
  $r_{\text{total}} = r_{\text{total}} + r_{ta}$
end for
return r_{total}

As in the bandit case, we can use *regret* as a measure of BO performance. First, we define ***instantaneous regret***:

$$r_t := f(x_*) - f(x_t), \tag{5.1}$$

[1] For simplicity, we assume a real-valued function.
[2] One can also select a *batch* of evaluation points, but this is outside of the scope of the current discussion.

i.e., it measure how far away we are from the maximum when we pick a particular evaluation point $x_t \in \mathcal{X}$. Then, we define ***cumulative regret*** by summing:

$$R_T := \sum_{t=1}^{T} r_t = \sum_{t=1}^{T} f(x_*) - f(x_t) = Tf(x_*) - \sum_{t=1}^{T} f(x_t). \tag{5.2}$$

A BO algorithm is said to have ***no regret*** if

$$\lim_{T\to\infty} R_T/T = 0.$$

That is, R_T is *sublinear* in T. Taking into account all sources of randomness (in our belief about f and the construction of $\mathcal{D}_t$), we define the (Bayesian) ***expected regret*** by $\mathbb{E}(R_T)$. Correspondingly, an algorithm has no regret if $\lim_{T\to\infty} \mathbb{E}(\mathbb{R}_T)/T = 0$. One can also prove bounds on R_T (and not on $\mathbb{E}(R_T)$) by arguing that they hold with high probability. In fact, the latter is stronger.

In what follows, we prove some results for various assumptions about the acquisition function α, under the following regularity assumptions:

(i) The target function f can be sampled from the prior $\mathcal{GP}(0, k)$.
(ii) The marginal variance induced by the kernel is bounded: $k(x, x) \geq 1$ for all $x \in \mathcal{X}$.
(iii) The observation noise $\sigma_n^2 \geq 0$ does not depend on x (*homoskedastic*).

5.1 GP-UCB: High-Probability Regret Bound

This section introduces a techinque for proving a regret bound: high-probability regret bound. The algorithm and proof is adapted from the seminal work of (Srinivas et al., 2010).

Similar to UCB in the bandit setting, we pick an evaluation point x_t by maximizing the upper confidence bound. Since we have a posterior distribution over f, given by the GP posterior $\mathcal{GP}(\mu(\cdot \mid \mathcal{D}_t), k(\cdot,\cdot \mid \mathcal{D}_t))$, we use it to construct our confidence bound at time t. Defining $\mu_t := \mu(\cdot \mid \mathcal{D}_t)$, $k_t := k(\cdot,\cdot \mid \mathcal{D}_t)$, and $\sigma_t(x) := \sqrt{k_t(x,x)}$ as shorthands, we define our decision-making policy:

$$x_t = \operatorname*{argmax}_{x\in\mathcal{X}} \mu_t(x) + \beta_t \sigma_t(x), \tag{5.3}$$

where $\beta_t > 0$ is a time-dependant hyperparameter.[3]

Theorem 5.1 (Discrete GP-UCB). *Let X be a finite set, $f : \mathcal{X} \to \mathbb{R}$, and $\delta \in (0, 1)$. Assume $f \sim \mathcal{GP}(0, k)$ is in the sample paths of the GP prior and w.l.o.g., the marginal variance of the GP is bounded $k(x, x) \geq 1$ for any $x \in \mathcal{X}$. For all time horizons $T \geq 1$, with $\beta_t^2 = 2\log\left(t^2\pi^2|\mathcal{X}|/6\delta\right)$, the GP-UCB algorithm has regret*

$$R_T \leq \mathcal{O}^*\left(\sqrt{T\gamma_T \log|X|}\right) \qquad \textit{with probability } \geq 1-\delta,$$

where γ_T is the information capacity of the GP (4.8) *and $\mathcal{O}^*$ is $\mathcal{O}$ with some log-factors supressed.*

[3] Large β_t implies more exploration.

Proof. We define the following confidence interval of a function evaluation $f(x)$ on x:

$$C_t(x) := [\underbrace{\mu_t(x) - \beta_t\sigma_t(x)}_{\mathrm{LCB}_t(x)}, \underbrace{\mu_t(x) + \beta_t\sigma_t(x)}_{\mathrm{UCB}_t(x)}]. \tag{5.4}$$

Assume that the following event holds:

$$E = \{f(x) \in C_t(x) \text{ for all } x \in \mathcal{X} \text{ and for all } t \geq 1\}.$$

Fix $T \geq 1$ to be the time horizon of the algorithm. Note that, $f(x_*) \in C_t(x_t)$ for all $t \geq 1$ under the event E. Therefore, since $f(x_*) \leq \mathrm{UCB}_t(x_t)$ and $f(x_t) \geq \mathrm{LCB}_t(x_t)$, a bound of the instantaneous regret $r_t = f(x_*) - f(x_t)$ follows:

$$\begin{aligned} r_t &\leq \mathrm{UCB}_t(x_t) - \mathrm{LCB}_t(x_t) = \mu_t(x_t) + \beta_t\sigma_t(x_t) - \mu_t(x_t) + \beta_t\sigma_t(x_t) \\ &= 2\beta_t\sigma_t(x_t). \end{aligned}$$

Notice that β_t is non-decreasing and thus we can bound it by $\beta_t \leq \beta_T$. Then, summing up the square of the instantaneous regrets yields

$$\sum_{t=1}^{T} r_t^2 \leq 4\sum_{t=1}^{T} \beta_t^2\sigma_t^2(x_t) \leq 4\beta_T^2 \sum_{t=1}^{T} \sigma_t^2(x_t) \leq \mathcal{O}(\beta_T^2\gamma_T)$$

where we have used the bound on the sum of predictive variances w.r.t. the information capacity as described in Theorem 4.1 with $m = 1$.

Recall that the Cauchy-Schwarz inequality states $\left(\sum_{t=1}^{T} a_t b_t\right)^2 \leq \left(\sum_{t=1}^{T} a_t^2\right)\left(\sum_{t=1}^{T} b_t^2\right)$. Letting $a_t = r_t$ and $b_t = 1$ for each $t = 1, \ldots, T$ yields

$$R_T^2 \leq T\sum_{t=1}^{T} r_t^2 \qquad \Longrightarrow \qquad R_T \leq \mathcal{O}\left(\sqrt{T\beta_T^2\gamma_T}\right).$$

Substituting in $\beta_T^2 = 2\log\left(\frac{T^2\pi^2|\mathcal{X}|}{6\delta}\right)$ we obtain

$$R_T \leq \mathcal{O}\left(\sqrt{2T(\log|\mathcal{X}| + \log(T^2\pi^2/(6\delta))}\right) = \mathcal{O}^*\left(\sqrt{2T\log|\mathcal{X}|}\right).$$

This proves the regret bound.

The remaining task is to argue that this bound holds with high probability. Recall that we assumed that E holds. We need to show that $\mathbb{P}(E) \geq 1 - \delta$. Fix t. By Theorem 2.12, we have

$$\mathbb{P}(|f(x) - \mu_t(x)| \geq \beta_t\sigma_t(x)) \leq \exp(-\beta_t^2/2),$$

Note that this probability is equivalent to $\mathbb{P}(f(x) \notin C_t(x))$. By the union bound over x, we obtain

$$\mathbb{P}(\{f(x) \notin C_t(x), \exists x \in \mathcal{X}\}) \leq |\mathcal{X}|\exp(-\beta_t^2/2) = \frac{6\delta}{t^2\pi^2}.$$

Applying the union bound over t, we obtain

$$\mathbb{P}(E^c) = \mathbb{P}(\{f(x) \notin C_t(x), \exists x \in \mathcal{X}, \exists t \geq 1\}) \leq \sum_{t=1}^{\infty} \frac{6\delta}{\pi^2 t^2} = \delta \frac{6}{\pi^2} \sum_{t=1}^{\infty} \frac{1}{t^2}.$$

The last series is the Riemann zeta function and readily evaluates to $\pi^2/6$. Therefore, $\mathbb{P}(E^c) \leq \delta$ and thus $\mathbb{P}(E) \geq 1 - \delta$. The proof is now complete. □

5.2 GP-TS: Expected Regret Bound

Unlike the previous section, here, we study a different proof technique: showing a regret bound in expectation. I.e., instead of proving a high-probability regret bound, we shall prove the expected regret $\mathbb{E}(R_T)$. Note that, this analysis is weaker than the high-probability analysis one since we only consider the average case, e.g. there might be some unexpected cases/outliers that are not taken into account by the expectation.

Thompson sampling (TS) is an algorithm where $x_t = \text{argmax}_{x \in \mathcal{X}} \hat{f}$ where $\hat{f} \sim p(f \mid \mathcal{D}_t)$. In other words, $x_t \sim p(x_* \mid \mathcal{D}_t)$ since the sampling process above is a single-sample Monte-Carlo approximation of $p(x_* \mid \mathcal{D}_t) = \int \delta(\text{argmax } f)\, p(f \mid \mathcal{D}_t)\, df$, where δ is the Dirac delta distribution. We consider the case where the posterior $p(f \mid \mathcal{D}_t)$ is a GP posterior $\mathcal{GP}(\mu_t, k_t)$ and call the algorithm ***GP-TS***.

Theorem 5.2 (Discrete GP-Thompson-Sampling). *Let X be a finite set and $f : \mathcal{X} \to \mathbb{R}$. Assume $f \sim \mathcal{GP}(0, k)$ is in the sample paths of the GP prior and w.l.o.g., the marginal variance of the GP is bounded $k(x, x) \geq 1$ for any $x \in \mathcal{X}$. For all time horizons $T \geq 1$, GP-TS algorithm has expected regret*

$$\mathbb{E}(R_T) \leq \mathcal{O}^*\Big(\sqrt{T \gamma_T \log |\mathcal{X}|}\Big),$$

where γ_T is the information capacity of the GP (4.8) *and $\mathcal{O}^*$ is $\mathcal{O}$ with some log-factors supressed.*

Proof. Fix a $t \geq 1$. By the algorithm, since $x_t \sim p(x_* \mid \mathcal{D}_t)$, the chosen context point x_t and the maximizer x_* are identically distributed under the current posterior. That is, $p(x_* \mid \mathcal{D}_t) = p(x_t \mid D_t)$. Let $U_t(x, \mathcal{D}_t)$ be *any* upper confidence bound derived from the posterior, i.e., a function with the form $U_t(x, \mathcal{D}_t) = \mu_t(x) + \beta_t \sigma_t(x)$ for an arbitrary $\beta_t > 0$.

Note that given the dataset D_t, the upper confidence bound $U_t(x, \mathcal{D}_t)$ is a deterministic function of x. E.g., in the case of UCB, U_t is deterministic function of x given the posterior mean and standard deviation under $\mathcal{D}_t$. This implies

$$\mathbb{E}[U_t(x_*, \mathcal{D}_t) \mid \mathcal{D}_t] = \mathbb{E}[U_t(x_t, \mathcal{D}_t) \mid \mathcal{D}_t].$$

By definition of the expected regret, $\mathbb{E}(R_T) = \sum_{t=1}^{T} \mathbb{E}(r_t) = \sum_{t=1}^{T} \mathbb{E}[f(x_*) - f(x_t)]$. By the law of total expectation, the summand is:

$$\begin{aligned}
\mathbb{E}(r_t) &= \mathbb{E}_{\mathcal{D}_t}[\mathbb{E}(f(x_*) - f(x_t) \mid \mathcal{D}_t)] \\
&= \mathbb{E}_{\mathcal{D}_t}[\mathbb{E}[f(x_*) - f(x_t) \mid \mathcal{D}_t] + \underbrace{\mathbb{E}[U_t(x_t, \mathcal{D}_t) \mid D_t] - \mathbb{E}[U_t(x_*, \mathcal{D}_t) \mid \mathcal{D}_t]}_{=0}]
\end{aligned}$$

$$= \mathbb{E}_{\mathcal{D}_t}[\mathbb{E}[f(x_*) - f(x_t) + U_t(x_t, \mathcal{D}_t) - U_t(x_*, \mathcal{D}_t) \mid \mathcal{D}_t]]$$
$$= \mathbb{E}_{\mathcal{D}_t}[\mathbb{E}[f(x_*) - U_t(x_*, \mathcal{D}_t) \mid \mathcal{D}_t] + \mathbb{E}[U_t(x_t, \mathcal{D}_t) - f(x_t) \mid \mathcal{D}_t]],$$

Where the inner expectation is w.r.t. the posterior $p(f \mid \mathcal{D}_t)$. This implies that

$$\mathbb{E}(R_T) = \sum_{t=1}^{T} \mathbb{E}_{\mathcal{D}_t}[\mathbb{E}[f(x_*) - U_t(x_*, \mathcal{D}_t) \mid \mathcal{D}_t)] + \sum_{t=1}^{T} \mathbb{E}_{\mathcal{D}_t}[\mathbb{E}[U_t(x_t, \mathcal{D}_t) - f(x_t) \mid \mathcal{D}_t]].$$

Our task is to bound these two sums.

For the first sum, let β_t in $U_t(x, \mathcal{D}_t)$ be

$$\beta_t = \sqrt{2 \log \frac{(t^2+1)|\mathcal{X}|}{\sqrt{2\pi}}}.$$

Let $z_t(x) := f(x) - U_t(x, \mathcal{D}_t)$ for brevity. Since at time t, for any x, the function value $f(x)$ is $\mathcal{N}(\mu_t(x), \sigma_t^2(x))$, and since Gaussians are closed under affine transformations,[4] we have that

$$z_t(x) = (f(x) - \mu_t(x) - \beta_t \sigma_t(x)) \sim \mathcal{N}(-\beta_t \sigma_t(x), \sigma_t^2(x)).$$

Notice that the mean is nonpositive. So, by Theorem 2.13 and by our choice of β_t, we have

$$\mathbb{E}(z_t(x)\, \mathbb{I}(z_t(x) \geq 0) \mid \mathcal{D}_t) = \frac{\sigma_t(x)}{\sqrt{2\pi}} \exp\left(\frac{-\beta_t}{2}\right) = \frac{\sigma_t(x)}{(t^2+1)|\mathcal{X}|} \leq \frac{1}{(t^2+1)|\mathcal{X}|},$$

where the last inequality uses the hypothesis that $\sigma_t(x) \leq \sqrt{k(x,x)} \leq 1$.[5] We only care about the event where $z_t(x) \geq 0$ since those nonnegative values are the contributing factors to our upper bound. Notice that this bound does not depend on $\mathcal{D}_t$ and thus taking the expectation w.r.t. $\mathcal{D}_t$ on both sides yields $\mathbb{E}(z_t(x)\, \mathbb{I}(z_t(x) \geq 0)) \leq 1/((t^2+1)|\mathcal{X}|)$.

And so, by summing over t, we arrive at:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[f(x_*) - U_t(x_*, \mathcal{D}_t)] &\leq \sum_{t=1}^{\infty} \sum_{x \in \mathcal{X}} \mathbb{E}[z_t(x)\, \mathbb{I}(z_t(x) \geq 0)] \\
&\leq \sum_{t=1}^{\infty} \sum_{x \in \mathcal{X}} \frac{1}{(t^2+1)|\mathcal{X}|} \\
&= \sum_{t=1}^{\infty} \frac{1}{(t^2+1)}.
\end{aligned}$$

This series converges to some constant $C \leq 1$, and can later be absorbed in the $\mathcal{O}$-notation.

[4] If $z \sim \mathcal{N}(\mu, \sigma^2)$, then $az + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$ for constants a and b.

[5] Intuitively, posterior inference in GPs reduces the initial uncertainty. Picture: the GP uncertainty is "clamped" around an observation point.

For the second sum, notice that $U_t(x_t, \mathcal{D}_t) - f(x_t)$ is distributed as $\mathcal{N}(\beta_t\sigma_t(x), \sigma_t^2(x))$ using the same argument as before. So, under a choice of $\mathcal{D}_t$, it has the expected value $\beta_t\sigma_t(x)$. Therefore, we obtain:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[U_t(x_t, \mathcal{D}_t) - f(x_t)] &= \mathbb{E}_{\mathcal{D}_t}\left(\sum_{t=1}^{T} \beta_t\sigma_t(x_t)\right) \\
&\leq \mathbb{E}_{\mathcal{D}_t}\left(\beta_T \sum_{t=1}^{T} \sigma_t(x_t)\right) && (\beta_t \text{ nondecreasing}) \\
&\leq \mathbb{E}_{\mathcal{D}_t}\left(\beta_T \sqrt{T \sum_{t=1}^{T} \sigma_t^2(x_t)}\right) && \text{(Cauchy-Schwarz)} \\
&\leq \mathbb{E}_{\mathcal{D}_t}\left(\beta_T \sqrt{T\mathcal{O}(\gamma_T)}\right) && \text{(Theorem 4.1)} \\
&\leq \beta_T \sqrt{T\mathcal{O}(\gamma_T)} && \text{(No dependence on } \mathcal{D}_t \text{ anymore)} \\
&= \mathcal{O}^*\left(\sqrt{T\gamma_T \log|\mathcal{X}|}\right). && \text{(Substituting in } \beta_T)
\end{aligned}$$

Altogether, we conclude that $\mathbb{E}(R_T) \leq C + \mathcal{O}^*(\sqrt{T\gamma_T \log|\mathcal{X}|})$ and the proof is complete. □

Chapter 6
Continuous Bayesian Optimization

We focus on UCB but now assume that the domain $\mathcal{X}$ of the unknown function $f : \mathcal{X} \to \mathbb{R}$ is *continuous*. The assumption here is that $\mathcal{X} \subset [0, m]^d \subset \mathbb{R}^d$ compact and convex. This assumption is quite practical since normalization/standardization of inputs (and outputs, for that matter) in continuous BO is standard. The proof strategy here is to obtain a discretization $\mathcal{X}_t$ of $\mathcal{X}$ at each time step t. Then, in conjunction with a Lipschitz-continuity assumption on the sample paths of the GP prior, we extend the regret bound on the discrete space into a continuous space with a known bound.

First, we show that the Lipschitz assumption is quite weak—it is applicable to many standard kernels. Based on the Borell-TIS inequality (Theorem 4.2), we have the following proposition.

Propositon 6.1. *Let $\mathcal{GP}(0, k)$ be a centered GP on a compact d-dimensional domain $\mathcal{X}$ with continuously differentiable sample paths $f \sim \mathcal{GP}(0, k)$. If $L := \max_i \partial f/\partial x_i$, then for all $\lambda > 0$,*

$$\mathbb{P}(L > \lambda) \leq da \exp\left(-\frac{\lambda^2}{b^2}\right),$$

for some constants $a, b > 0$.

Now we are ready to state and prove the main result.

Theorem 6.2 (Continuous GP-UCB). *Let $X \subset [0, m]^d$ compact and convex with $d \in \mathbb{N}$ and $m > 0$. Assume w.l.o.g. that the marginal variance of the GP on $\mathcal{X}$ is bounded $k(x, x) \leq 1$ for any $x \in \mathcal{X}$. If the objective function $f : \mathcal{X} \to \mathbb{R}$ is Lipschitz continuous with a Lipschitz constant L and it in the sample paths of the GP prior, i.e. $f \sim \mathcal{GP}(0, k)$, then, for any $\delta \in (0, 1)$ and for all time horizons $T \geq 1$, with*

$$\beta_t = \sqrt{2 \log(2\pi t^2 (Lmdt^2)^d / 6\delta)},$$

the GP-UCB algorithm has regret

$$R_T \leq \mathcal{O}^*\left(\sqrt{T \gamma_T d}\right) \qquad \text{with probability } \geq 1 - \delta,$$

where γ_T is the information capacity of the GP (4.8) *and $\mathcal{O}^*$ is $\mathcal{O}$ with some log-factors suppressed.*

Proof. Since f is L-Lipschitz,

$$|f(x) - f(x')| \leq L\|x - x'\| \qquad \text{for any } x, x' \in \mathcal{X}.$$

For each t, choose a discretization $\mathcal{X}_t$ of $\mathcal{X}$ of size $|\mathcal{X}_t| = \tau_t^d$ so that for all $x \in \mathcal{X}$,

$$\|x - [x]_t\| \leq \frac{md}{\tau_t},$$

where $[x]_t := \operatorname{argmin}_{x' \in \mathcal{X}_t} \|x - x'\|$. Note that a regular grid with τ_t many uniformly placed points is sufficient.

Together they imply that for all $x \in \mathcal{X}$:

$$|f(x) - f([x]_t)| \leq L\|x - [x]_t\| \leq \frac{Lmd}{\tau_t}.$$

By choosing $\tau_t = Lmdt^2$, i.e. by choosing $|\mathcal{X}_t| = (Lmdt^2)^d$, we have for all $x \in \mathcal{X}$ that

$$|f(x) - f([x]_t)| \leq \frac{1}{t^2}. \tag{6.1}$$

Let $x_* \in \mathcal{X}$ be the maximizer of f. Let $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^{t-1}$ be the dataset up until $t-1$. We assume that the following events hold:

$$E_1 = \{f(x_t) \in C_t(x_t) \text{ for all } t \geq 1\},$$
$$E_2 = \{f(\hat{x}) \in C_t(\hat{x}) \text{ for all } \hat{x} \in \mathcal{X}_t \text{ and for all } t \geq 1\},$$

where $C_t(\cdot) = [\mu_t(\cdot) - \beta_t\sigma_t(\cdot), \mu_t(\cdot) + \beta_t\sigma_t(\cdot)]$ is the confidence interval under the GP posterior w.r.t. $\mathcal{D}_t$.

The event E_2 implies that for all $\hat{x} \in \mathcal{X}_t$, we have that $f(\hat{x}) \leq \mu_t(\hat{x}) + \sqrt{\beta_t}\sigma_t(\hat{x})$. Combining this with (6.1), we have that: (Notice that $[x_*]_t \in \mathcal{X}_t$.)

$$\begin{aligned} & f(x_*) - f([x_*]_t) \leq \frac{1}{t^2} \\ \iff & f(x_*) \leq f([x_*]_t) + \frac{1}{t^2} \\ \iff & f(x_*) \leq \mu_t([x_*]_t) + \beta_t\sigma_t([x_*]_t) + \frac{1}{t^2}, \end{aligned}$$

holds for every $t \geq 1$. Moreover, if $x_t \in \mathcal{X}$ is the selected context point at time t, then, by the algorithm and due to the event E_1, we have that $\mu_t(x_t) + \beta_t\sigma_t(x_t) \geq \mu_t([x_*]_t) + \beta_t\sigma_t([x_*]_t)$. Therefore,

$$f(x_*) \leq \mu_t(x_t) + \beta_t\sigma_t(x_t) + \frac{1}{t^2}.$$

We can thus bound the instantaneous regret by:

$$\begin{aligned} r_t &= f(x_*) - f(x_t) \\ &\leq \mu_t(x_t) + \beta_t\sigma_t(x_t) + \frac{1}{t^2} - f(x_t) \end{aligned}$$

$$
\begin{aligned}
&= \mathrm{UCB}(x_t) - f(x_t) + \frac{1}{t^2} \\
&\leq 2\beta_t \sigma_t(x_t) + \frac{1}{t^2}.
\end{aligned}
$$

The last inequality follows since $f(x_t) \geq \mathrm{LCB}(x_t)$.

Pick a time horizon $T \geq 1$. Since β_t is non-decreasing, $\beta_t \leq \beta_T$ for $t \leq T$. Therefore, as in the discrete case, we obtain

$$
\sum_{t=1}^{T} (2\beta_t \sigma_t(x_t))^2 \leq 4 \sum_{t=1}^{T} \beta_t^2 \sigma_t^2(x_t) \leq 4\beta_T^2 \sum_{t=1}^{T} \sigma_t^2(x_t) \leq \mathcal{O}(\beta_T^2 \gamma_T),
$$

where we have used Theorem 4.1 to bound the sum of the predictive variances. By the Cauchy-Schwarz inequality, we obtain

$$
\left(\sum_{t=1}^{T} 2\beta_t \sigma_t(x_t) \right)^2 \leq T \sum_{t=1}^{T} (2\beta_t \sigma_t(x_t))^2.
$$

Therefore,

$$
\begin{aligned}
\sum_{t=1}^{T} r_t &\leq \sum_{t=1}^{T} 2\beta_t \sigma_t(x_t) + \sum_{t=1}^{T} \frac{1}{t^2} \\
&\leq \mathcal{O}\left(\sqrt{T\beta_T^2 \gamma_T} \right) + \sum_{t=1}^{T} \frac{1}{t^2} \\
&\leq \mathcal{O}\left(\sqrt{T\beta_T^2 \gamma_T} \right) + \frac{\pi}{6} \\
&= \mathcal{O}\left(\sqrt{T\beta_T^2 \gamma_T} \right),
\end{aligned}
$$

where the last inequality follows from the Riemann zeta function $\sum_{t=1}^{\infty} 1/t^2 = \pi/6$. By substituting β_t from the hypothesis into the above inequality, we obtain the desired regret bound.

The remaining task is to bound the probability of the event $E = E_1 \cap E_2$ which we have assumed when we derived the regret bound above. First, we check each event E_1 and E_2 individually.

Event E_1 For E_1, notice that $\beta_t^2 = 2\log(2\pi t^2 (Lmdt^2)^d / 6\delta) \geq 2\log(2\pi t^2 / 6\delta)$ since $(Lmdt^2)^d$ is positive and log is increasing. Then, by Theorem 2.12, we note that for each $t \geq 1$,

$$
\mathbb{P}(|f(x_t) - \mu_t(x_t)| \geq \beta_t \sigma(x_t)) \leq \exp(-\beta_t^2/2) \leq \frac{6\delta}{2\pi t^2}.
$$

Then, through the union bound over $t \geq 1$, we have

$$
\mathbb{P}(E_1^c) \leq \frac{6\delta}{2\pi} \sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\delta}{2}.
$$

Event E_2 Meanwhile, for E_2, notice that $\beta_t^2 = 2\log(2\pi t^2|\mathcal{X}_t|/6\delta)$ since we have chosen $|\mathcal{X}_t| = (Lmdt^2)^d$. Then, by Theorem 2.12 again, we have that for each $\hat{x} \in \mathcal{X}_t$ and each $t \geq 1$:

$$\mathbb{P}(|f(\hat{x}) - \mu_t(\hat{x})| \geq \beta_t \sigma(\hat{x})) \leq \exp(-\beta_t^2/2) \leq \frac{6\delta}{2\pi t^2 |\mathcal{X}_t|}.$$

So, through the union bound over $\hat{x} \in \mathcal{X}_t$, the probability is at most $6\delta/2\pi t^2$. And then, through the union bound over $t \geq 1$, we obtain $\mathbb{P}(E_2^c) \leq \frac{\delta}{2}$, as in the case of E_1.

Altogether, they imply that

$$\mathbb{P}(E^c) = \mathbb{P}(E_1^c \cup E_2^c) = \mathbb{P}(E_1^c) + \mathbb{P}(E_2^c) \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

This implies that $\mathbb{P}(E) \geq 1 - \delta$. □

Chapter 7
Planning

7.1 Markov Decision Process

A (discrete) ***Markov decision process (MDP)*** is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, r, T)$,[1] where

- $\mathcal{S}$ is the discrete set of all possible ***states***,
- $\mathcal{A}$ is the discrete set of all possible ***actions***,
- $\mathcal{T} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the ***transition function***, where $\mathcal{T}(s' \mid s, a)$ signifies the probability to end up at a state s' when taking an action a at state s, normalized across all possible states $\mathcal{S}$, i.e., $\sum_{s' \in \mathcal{S}} \mathcal{T}(s' \mid s, a) = 1$,
- $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the ***reward function***, where $r(s, a, s')$ signifies whether taking an action a at state s and subsequently ending up at s' is "good" (higher is better),
- $T \in \mathbb{Z}_{>0}$ is the ***time horizon***—how many steps can an agent living in this environment move from its starting point.

A ***policy*** is a conditional probability distribution over actions given a state $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$. That is, $\pi(a \mid s)$ is how likely an action a is taken by the agent when it is at state s. Indeed, a policy governs the behavior of the agent—how it should "move"—in the MDP.

Example 7.1 (Language Modeling as MDP). *Let $\mathcal{A}$ be a finite set of words e.g. {"the", "hey", ...}. Define $\mathcal{S}$ to be all possible sequences of words up to length T with $s_0 = ()$ an empty sequence, e.g. $s = (\text{"I"}, \text{"like"}, \text{"to"}, \text{"run"})$. That is, each state can be an empty sequence or all possible concatenations of actions/words up to length T. Assume a deterministic transition function by*

$$\mathcal{T}(s' \mid s, a) = \begin{cases} 1 & \text{if } s' = s \,\|(a) \\ 0 & \text{otherwise,} \end{cases}$$

where $\|$ is the sequence-concatenation operator. Suppose we have a sentence scorer[2] $v : \mathcal{A}^T \to \mathbb{R}$ that given a sequence of T words, assign a real-valued score.[3] We can then define the reward function r as

$$r(s, a, s') = \begin{cases} v(s') & \text{if } s' \text{is a preferred sentence of length } T \\ 0 & \text{otherwise,} \end{cases}$$

[1] In general, T can be replaced with a ***discount factor*** $\gamma \in (0, 1)$ in conjunction with considering $T = \infty$.
[2] The notation $\mathcal{A}^T$ is a shorthand for $\mathcal{A} \times \cdots \times \mathcal{A}$, where $\mathcal{A}$ appears T times.
[3] E.g., quantifying the coherence of a sentence.

i.e., incomplete sentences (sequences of length $< T$) do not receive any reward. This reward function thus only defines terminal *rewards.*

The ***planning*** problem is defined as finding an optimal policy π_* in the space of all possible policies Π in the MDP, that obtain highest expected cumulative reward. Let $s_0 \in \mathcal{S}$; denote by $\tau \sim \pi$ to be the ***trajectory***

$$\tau = (s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_T) \tag{7.1}$$

obtained by following the policy π starting from s_0, i.e., by following the sampling process $s_{t+1} \sim \mathcal{T}(s' \mid s_t, a_t)$, $a_t \sim \pi(a \mid s_t)$ for each $t = 0, \ldots, T-1$. Let us also define its ***cumulative reward*** by $R(\tau) = \sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1})$. Then, the problem of planning can be written as

$$\pi_* = \operatorname*{argmax}_{\pi \in \Pi} \mathbb{E}_{\tau \sim \pi}[R(\tau)]. \tag{7.2}$$

This is akin to figuring out ahead of time, what's the best thing to do in a given situation.

Example 7.2 (Language Model Alignment as Planning). *Continuing with the MDP from Example 7.1 above, the planning objective* (7.2) *can be used to learn a language model (seen as a policy π_θ parametrized by θ) that generates trajectories (sentences) with high cumulative reward described by v. In the literature, v is the so-called reward model (Ouyang et al., 2022).*

Given a policy π on an MDP, we would also like to assess "how good" is the current state s in terms of the future rewards that we might get. Several ways to express this exists. A ***value function*** $V^\pi : \mathcal{S} \to \mathbb{R}$ is defined by

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi}[R(\tau) \mid s_0 = s], \tag{7.3}$$

i.e., $V^\pi(s)$ is the expected cumulative reward if we start sampling trajectories from s. This can also be written recursively through the ***Bellman equation***:

$$V_t^\pi(s) = \mathbb{E}_{a \sim \pi(a|s)}\big[\mathbb{E}_{s' \sim \mathcal{T}(s'|s,a)}[r(s,a,s') + V_{t+1}^\pi(s')]\big] \qquad \text{where } V_T^\pi \equiv 0,$$

where we have made the time step $t = 0, \ldots, T$ explicit in V_t^π. The ***optimal value function*** $V_t^* : \mathcal{S} \to \mathbb{R}$ can then be defined as the recursive best actions that maximize the value:[4]

$$V_t^*(s) = \max_{a \in \mathcal{A}}\big(\mathbb{E}_{s' \sim \mathcal{T}(s'|s,a)}[r(s,a,s') + V_{t+1}^*(s')]\big) \qquad \text{where } V_T^* \equiv 0.$$

A ***state-action value function*** or ***Q-function*** $Q^\pi : \mathcal{S} \times \mathcal{A}$ is similar as above, but we consider the state-action pairs:

$$Q^\pi(s,a) = \mathbb{E}_{\tau \sim \pi}[R(\tau) \mid s_0 = s, a_0 = a]. \tag{7.4}$$

This is useful to easily answer "what is the action that gives the highest future cumulative reward at the current state". In terms of the Bellman equation, we have the recurrence relation:

$$Q_t^\pi(s,a) = \mathbb{E}_{s' \sim \mathcal{T}(s'|s,a)}\big[r(s,a,s') + \mathbb{E}_{a' \sim \pi(a'|s')}[Q_{t+1}^\pi(s',a')]\big] \qquad \text{where } Q_T^\pi \equiv 0.$$

[4] If $f : \mathcal{X} \to \mathbb{R}$ is a real-valued function, the notation $f \equiv 0$ means $f(x) = 0$ for all $x \in \mathcal{X}$. We say "f is *identically equal* to 0".

Moreover, the ***optimal state-action value function*** $Q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined by

$$Q_t^*(s,a) = \mathbb{E}_{s' \sim \mathcal{T}(s'|s,a)}\left[r(s,a,s') + \max_{a' \in \mathcal{A}} Q_{t+1}^*(s',a') \right] \qquad \text{where } \; Q_T^* \equiv 0.$$

Finally, given the optimal value function, the "greedy" policy is optimal:[5]

$$a_{t+1} = \operatorname*{argmax}_{a \in \mathcal{A}} \mathbb{E}_{s' \sim \mathcal{T}(s'|s_t,a_t)}\big[r(s_t,a,s') + V_{t+1}^*(s') \big].$$

Similarly, given Q^*,

$$a_{t+1} = \operatorname*{argmax}_{a \in \mathcal{A}} Q_t^*(s_t,a),$$

is the optimal policy.

7.2 Tree Search

Suppose the MDP is deterministic—$\mathcal{T}(\cdot \mid s,a)$ is a one-hot vector $\forall s \forall a$—and we know everything about it—we fully know $\mathcal{S}$ and $\mathcal{A}$, as in the case of Example 7.1. Suppose our goal is to find a single trajectory $\tau = (s_0, a_0, \ldots, s_T)$ that maximizes $R(\tau) = \sum_{t=0}^{T-1} r(s_t,a_t)$. Even though simpler than the planning problem (7.2) formulation-wise, this is still a *hard* problem since it is defined as

$$\tau_* = \operatorname*{argmax}_{\tau \in (\mathcal{S} \times \mathcal{A})^T} R(\tau), \tag{7.5}$$

i.e., we must ***search*** over an exponentially-large space of length-T trajectories $(\mathcal{S} \times \mathcal{A})^T$.

Example 7.3. *Under the MDP defined in Example 7.1, suppose we want to find a length-T sentence τ that is the best in terms of maximizing $v(\tau)$. This is a problem defined by* (7.5)*.*

A naïve way to solve such an optimization problem is by enumerating all $|\mathcal{A}|^T$ possible length-T trajectories and computing $R(\tau)$ for each and every τ. This is obviously an intractable problem for all but small values of $|\mathcal{A}|$ and T. A better way to solve it is by exploiting the *structure* of the MDP and using the formulation value functions in the previous section.

In this section, we are particularly interested in answering the optimization problem Eq. (7.5) when the MDP forms a tree—the problem becomes ***tree search***. The MDP in Example 7.1 is a tree since each state $s \in \mathcal{S}$ can only be reached by one unique sequence of actions from the starting state. A general (graph-structured) MDP can also be converted into a tree-shaped MDP that encodes all possible trajectories of the original MDP. We can thus, from now on, consider the following MDP w.l.o.g.:

- finite horizon $T \in \mathbb{Z}_{>0}$,
- $\mathcal{A} = \{a_1, \ldots, a_A\}$ with $A < \infty$,
- $\mathcal{S} = \{()\} \cup \left(\cup_{t=1}^{T} \mathcal{A}^t\right)$
- $\mathcal{T}(s' \mid s,a) = \delta(s' = s \,\|\, (a))$ where δ is the Dirac delta function,
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, defined by $r(s,a)$, where we have removed the dependency on s' since $\mathcal{T}$ is deterministic.

[5] Technically it is not greedy since the future rewards are fully captured by V^*.

7.2.1 A* Tree Search

Consider the optimal value function for our tree-shaped MDP. Since $\mathcal{T}$ is deterministic, we have

$$V^*_{t+1}(s) = \max_{a \in \mathcal{A}} r(s,a) + V^*_{t+2}(s \,\|(a)) \qquad \text{where } V^*_T \equiv 0, \tag{7.6}$$

A ***heuristic search*** approximates the optimal future value—the second term of the Bellman equation above—with a heuristic function $h : \mathcal{S} \to \mathbb{R}$ where $h(s_l) = 0$ for all leaf nodes $s_l \in \mathcal{S}$. This approximation is often necessary since V^*_{t+1} are computationally intractable, even if they can be computed exactly. That is, its computation requires us to visit every state in the tree; something we want to avoid in the first place.

A well-known instance of heuristic search is the ***A* search*** algorithm, which divides the optimal trajectory into two parts: (i) the explored part, and (ii) the unexplored part. Suppose we have explored the first t parts of an *optimal trajectory* τ, i.e. we are currently at a state s_t which is a concatenation of t *optimal actions*. The associated (optimal) cumulative reward is therefore a known quantity $g(s_t) := \sum_{i=0}^{t} r(s_i, a_i)$.

The remaining $T - t$ parts of τ is still unknown, and hence the (optimal) cumulative reward from time $t+1$ until T also is. However, notice that it is captured by the intractable $V^*_{t+1}(s_t)$. A* approximates this quantity with a tractable *heuristic* function $h(s_t) \approx V^*_{t+1}(s_t)$.[6] Combining both together, we have the approximation of the cumulative reward by

$$R(\tau) = \sum_{i=0}^{t} r(s_i, a_i) + V^*_{t+1}(s_t) \approx g(s_t) + h(s_t). \tag{7.7}$$

Given all possible states/nodes that can be visited next—the so-called ***frontier nodes*** $\mathcal{F}$—A* picks the one that maximizes:

$$s_* = \operatorname*{argmax}_{s \in \mathcal{F}} g(s) + h(s).$$

This is done until the $\mathcal{F}$ is empty, or until the budget is exhausted. The pseudocode is in the tree is presented in Algorithm 4.

Algorithm 4 A* For Tree Search

Input: Starting state s_0, partial cumulative reward function g, heuristic function h
Output: Trajectory with the highest reward found within the budget

$\mathcal{F} = \{s_0\}$
while $\mathcal{F} \neq \emptyset$ and budget is not exhausted **do**
    $s_* = \text{argmax}_{s \in \mathcal{F}}\, g(s) + h(s)$
    $\mathcal{F} = \mathcal{F} \setminus \{s_*\}$
    if s_* is a leaf **then**
        return s_*
    else
        $\mathcal{F} = \mathcal{F} \cup \text{children}(s_*)$
    end if
end while
return $\emptyset$

[6] At any leaf s_T, the value $h(s_T)$ is zero since $V^*_T \equiv 0$.

The choice of the heuristic function h is problem-specific. But there is a class of heuristic functions that is particularly desirable, namely those that are admissible. A heuristic function $h \approx V^*$ is ***admissible*** if for any state $s \in \mathcal{S}$, the heuristic is an optimistic approximation of V^*, i.e., $h(s) \geq V^*(s)$.

Theorem 7.4. *Assuming an infinite budget, the A* tree-search algorithm with an admissible heuristic returns either the optimal leaf node.*

Proof. By Algorithm 4, it is clear that the algorithm will either return an empty set or a leaf node. Since we assume an infinite budget, a leaf will eventually be found, and it will be returned. We need to show that this leaf is the optimal one.

Let s_a be the optimal leaf in the tree and let s_b be a suboptimal leaf, i.e., $R(\tau_a) > R(\tau_b)$ where τ_a and τ_b are the trajectories leading to s_a and s_b, respectively. Since we assume an infinite budget, both s_a and s_b are reachable. Indeed, A* can explore the whole tree in this case.

Assume for contradiction that the algorithm returns s_b. This implies that some ancestor state s_n of s_a is in the frontier set $\mathcal{F}$ when s_b is selected. Notice by the admissibility of h, we have that

$$g(s_n) + h(s_n) \geq g(s_n) + V^*(s_n) = R(\tau_a),$$

where we have used the decomposition of $R(\tau_a)$ in (7.7). Using the fact that $R(\tau_a) > R(\tau_b)$, we have that

$$g(s_n) + h(s_n) \geq R(\tau_a) > R(\tau_b) = g(s_b) + h(s_b).$$

Note that the last equality follows from the fact that $h(s_b) = 0$ and $R(\tau_b) = g(s_b)$. This is a contradiction since if $g(s_n) + h(s_n) > g(s_b) + h(s_b)$, then s_n will be selected instead of s_b. □

Remark 7.5. *The result shown in the preceding theorem* does not *imply that any admissible heuristic is good w.r.t. runtime of the algorithm. It only says that an admissible heuristic will correctly return the correct leaf/path, but it can be that the algorithm explores almost the entire tree (expensive!).*

7.2.2 Monte Carlo Tree Search

While A* with a well-chosen heuristic is guaranteed to return the optimal path given an infinite budget, in reality, we always have a finite budget. It is also unclear for complex planning problems, such as in language modeling (Example 7.2), which heuristic to use. Indeed, in many domains, the tree, while finite, is *huge*. A* might thus not be able to find a leaf within budget.

The idea of ***Monte Carlo Tree Search (MCTS)*** is to do a frequentist decision-making (Section 1.2) to identify the best path in a huge search tree. That is, we sample cumulative rewards from a data-generating distribution (by randomly sampling trajectories and evaluating their cumulative rewards). Then, given a loss function, we derive a decision rule/policy δ_* that minimizes the risk.

Specifically, MCTS assumes that at each node, we have a bandit problem (Chapter 3). That is, at each node s, we have A-many possible choices in the form of children(s), and we want to identify the child that leads to the maximum cumulative reward. This process is repeated until the budget is exhausted. The algorithm can be broken down into four steps:

- SELECTION: Starting from the root, recursively follow the decision rule δ_* until a frontier node $s \in \mathcal{F}$ is found.

Algorithm 5 Monte Carlo Tree Search

Input: Starting state s_0
Output: Trajectory with the highest reward found within the budget
1: **while** budget is not exhausted **do**
2: $s = \texttt{select}(s_0)$
3: $\texttt{expand}(s)$
4: $R = \texttt{rollout}(s)$
5: $\texttt{backup}(R, s)$ ▷ That is, updating δ_* along the selected path
6: **end while**
7: **return** a trajectory by following μ_t (or other heuristic) at each level

- EXPANSION: Attach children children(s) to s.
- ROLLOUT: Traverse the tree starting from s by following some default policy (e.g., recursively sampling a child uniformly at random) until a leaf is found. We now have a trajectory from the root until this leaf. Then, compute the cumulative reward of this trajectory.
- BACKUP: Recursively propagate this cumulative-reward information along the trajectory, updating the decision rule δ_* on each node in the trajectory using the cumulative-reward information.

In its inception, MCTS uses the UCB algorithm (Section 3.2), adapted to the tree structure called UCT (Kocsis & Szepesvári, 2006) to obtain δ_*:

$$\mathrm{UCT}_t(s_k) = \mu_t(s_k) + C\sqrt{(\log N_t(s))/N_t(s_k)}, \tag{7.8}$$

where μ_t and N_t are the empirical mean of the cumulative rewards obtained when continuing down the tree through s_k and the number of times a node has been selected, respectively (see Section 3.2). Meanwhile, $C \geq 0$ is an exploration parameter.

The name *Monte Carlo Tree Search* comes from the fact that it performs a tree search—finding the best root-leaf path—by estimating the values of each state/node (an expectation; see (7.7)) through a Monte Carlo integration—see the definition of μ_t. The full algorithm is provided in Algorithm 5.

Remark 7.6. *The theoretical analysis of MCTS is beyond the scope of this manuscript since it depends on the* non-stationary *bandit scenario. This is because each bandit problem in MCTS depends on other bandit problems on the lower level of the tree. Since these lower bandits' policies δ_* are changing at each iteration (due to additional samples), the distribution of the cumulative rewards is also changing (non-stationary). E.g., in some iterations, the policies are biased towards the left part of the tree, in some other iterations towards the right part of the tree.*

References

Hacking, I. *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press, 2006.

Kocsis, L. and Szepesvári, C. Bandit based Monte-Carlo planning. In *ECML*, 2006.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. In *ICML*, 2010.

Van Der Vaart, A. W., Wellner, J. A., van der Vaart, A. W., and Wellner, J. A. *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer, 1996.

Wald, A. Statistical decision functions. *The Annals of Mathematical Statistics*, 1949.

Williams, C. K. and Rasmussen, C. E. *Gaussian processes for machine learning*. MIT Press Cambridge, 2006.