CHAPTER 1
PREVIOUS WORK

The intelligent control of distributed aperture optical systems lies at the intersection of remote sensing, optical physics, computer vision, and sequential decision making under uncertainty. This chapter presents a review of relevant foundational and recent work in each of these disciplines, and draws those relations among them that inform the problem formulation given in Ch.2. The chapter is divided broadly into remote sensing (RS) and artificial intelligence (AI), which are further partitioned into relevant sub-disciplines. Within remote sensing, classical image formation, principals of distributed aperture imaging (DAI), and applications of remote sensing to resident space objects (RSOs) are reviewed. Literature from the field of AI is grouped into foundational works, deep learning, computer vision, and reinforcement learning; descriptions are given chronologically and terminate in cross-disciplinary work related to remote sensing. While not comprehensive, the work reviewed in this chapter comprises the foundation upon which both the problem formulation (Ch. 2) and the methods used in this work (Ch. 3) are built. The intersectional hierarchy of topics covered by review is illustrated in Fig. 1.1, in which this work is positioned relative to those topics.

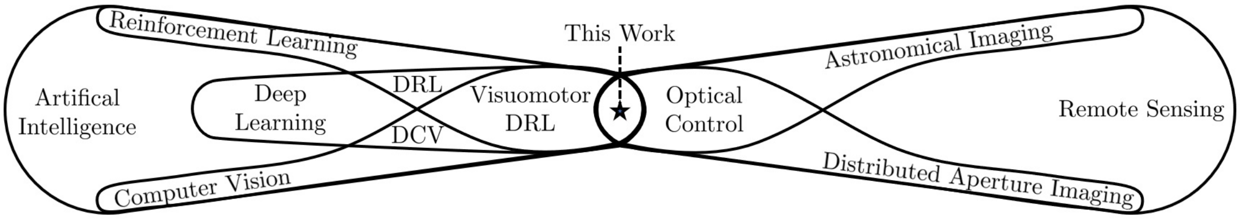

Figure 1.1: A domain overview, in which this work is positioned relative to artificial intelligence, deep learning, reinforcement learning, computer vision, deep computer vision (DCV), deep reinforcement learning (DRL), visuomotor DRL, remote sensing, astronomical imaging, distributed aperture imaging, and optical control.

1.1 Remote Sensing

The term remote sensing was first introduced in 1960 [10], though the first recorded instance of a human activity satisfying the modern definition of the term dates to 1608 [61]. For the purposes of this work, remote sensing is defined as the acquisition of information about an object through analysis of data collected by a device that is not in contact with the object, which is adapted from Lillesand et al. [32]. Clearly, remote sensing is foundational to astronomy, and the sensing of resident space objects generally. The work presented in this dissertation is motivated by two primary remote sensing applications: the discovery of exoplanetary life and the perception of man-made satellites. These applications lie in the antithetical extremities of a spectrum of remote sensing applications. The former involves the sensing of distant, large, relatively homogeneous, natural objects at low spatial resolution, but provides answers to profound questions about the nature of life in the universe; the latter entails sensing of comparatively nearby $(5 - 100 \times 10^6\ m)$, small, highly structured, artificial objects at high spatial resolution, and has immediate practical uses. Between these extrema lie a variety of intermediary applications which may yet be defined.

In this section, relevant works from the field of astronomical remote sensing are reviewed. The fundamental models of optical systems are described, and several key relations are given. A discussion of the contemporary work quantum theoretic approaches to image modeling is provided. The section concludes with a brief discussion of the application of modern computer vision techniques to problems in remote sensing of space objects.

1.1.1 Optical Systems

Like all physical systems, models must be used to describe and engineer optical systems for remote sensing. An image may be formulated as a measure of the intensity, I, of electromagnetic radiation distributed over some space of interest. Typically, this measure will have units of W/m, W/m^2, W/rad, or W/sr. The imaging target, known as the *object*, is modeled as a collection of point sources over some spatial extent and is denoted $O(x)$. Optical diffraction for far-field images is well modeled by Fraunhofer diffraction. For light of a wavelength λ diffracted through a circular aperture of radius a and measured on a Cartesian focal plane at a distance L from the aperture, intensity is modeled by

$$I(x) = I_0 \left[\frac{2J_1(x)}{x} \right]^2, \tag{1.1}$$

where I_0 is the maximum intensity, J_1 is the Bessel function of the first kind, and the distance along the focal plane, x, is given in terms of angular distance from the apertures optical axis, θ, as $x = (2\pi a/\lambda) \sin\theta$ [23]. Eq. 1.1 holds when $L \gg a^2/\lambda$, which is a reasonable assumption throughout this work. The symmetric region bound by the first zero of $I(x)$ is known as the *Airy disk* [23] and is used to quantify the size of the smallest classically resolvable element of an image. An optical system is said to be diffraction limited if the only blurring of incident energy from a distant source is caused by diffraction. For a point source, which is defined as an object of much smaller angular than the Air disk of the optical system imaging that object, the incident energy will be spread across the focal plane exactly as prescribed by Eq. 1.1; as such, Eq. 1.1 is known as the point spread function (PSF) of a diffraction limited optical system. The PSF of an image may possess artifacts from any number of optical, scene, or atmospheric aberrations.

Two point sources are said to be resolvable by a diffraction limited imaging system having a circular aperture with diameter D if their angular separation in radians, θ, satisfies the inequality

$$\theta \geq 1.22 \frac{\lambda}{D}, \tag{1.2}$$

where λ is the wavelength of the source. Eq. 1.2 is known as Rayleigh's criterion [43], and denotes the angular separation at which the two sources are closer to one another, as measured by their apparent separation, than edge of their Airy disk. The definition of resolvability provided by Raleigh's criterion is widely adopted, but inaccurate [64]. Researchers across several disciplines have demonstrated techniques to resolve point sources that do not meet Rayleigh's criterion [41, 42, 8]. Collectively, these techniques are known as subdiffraction limited imaging, or superresolution. While a complete review of superresolution is beyond the scope of this work, Sec. 1.1.3 provides a brief overview of relevant interferometry superresolution techniques.

1.1.2 Image Formation

The work described in the preceding paragraphs provides some constrains on the resolution of an imaging system, but is insufficient to describe image formation. To form a direct image, photons are focused onto a focal plane, on which a collection instrument is positioned. In this work, that instrument will be a scientific charge-coupled device (CCD) camera [9]. CCDs comprise pixels, each of which integrates incident photons by using the incident energy to move electrons from a source line into a local capacitor. These electrons are periodically discharged and their number estimated. The resulting raster of electron counts is known as an image; it would, however, be more precise to describe it as an image *estimate*. Each CCD pixel subtends a solid angle[1] from which photons are integrated. This subtended angle is known as the instantaneous field of view (IFOV) of that pixel. The IFOV is given by the system field of view (FOV), divided by the pixel count of the camera; both of these features are free parameters of the optical system design[2].

Using these concepts, and assuming that the subtended angle (IFOV) is small, the spatial resolving power, s (in meters), of a traditional diffraction-limited optical system with a primary aperture diameter of D over a distance r at a wavelength λ, is given by

$$s = 2r \tan\left(\frac{1.22}{2}\frac{\lambda}{D}\right) \approx 1.22\frac{r\lambda}{D}. \tag{1.3}$$

The aperture diameter required to achieve a given resolving power of a target at a known distance is then

$$D = \frac{1.22}{2}\frac{\lambda}{\arctan(s/2r)} \approx 1.22\frac{r\lambda}{s}. \tag{1.4}$$

This analysis uses Rayleigh's Criterion (Eq. 1.2) to define a resolvable element, and thus represents a classical bound lower bound on resolving power. The small-angle linear approximations used in Eqs. 1.3 and 1.4 are consistent with the conditions of astronomical imaging, for which $D \gg \lambda$.

[1] Strictly, each CCD pixel subtends two solid angles: one in each dimension. For notational convenience, this work assumes square pixels and neglects this nuance.

[2] One helpful framework for conceptualizing optical systems is as "angle-pass filters." A successful optical system sorts incident photons into a set of linearly spaced bins based on their angle of incidence with the primary aperture, while scattering all light that does not fall into an engineered range of incident angles.

The work presented in this dissertation may be framed as an attempt to improve on this classical bound; as such, Eq. 1.4 serves as a useful baseline of comparison.

The optical models used described above are approximations to the underlying quantum mechanical phenomena of image formation. For detailed expositions of the quantum mechanical basis of direct imaging, see Tsang 2016 [57] and Ang 2017 [2]. The direct imaging approaches described in these works require specific experimental conditions and instrumentation, and are not directly applicable to the direct astronomical imaging of extended objects. They do, however, provide a basis for the quantum theoretical limit of resolvability of two sources and, crucially, demonstrate that subdiffraction imaging (as classically defined) is theoretically supported. Tsang 2016 and Ang 2017 provide theoretical proof that correctly designed estimators can image at arbitrarily high precision under certain observational and instrumental constraints. Building on these results, Zhou and Jiang [64] reformulate Rayleigh's criterion in modern (by which they mean quantum mechanical) terms. In doing so, they extend the findings of Tsang 2016 and Ang 2017 to arbitrarily many point sources of arbitrary strength (as measured by source flux or incident energy), in both one- and two-dimensional images. Formally, Theorem 2 of [64] states that the variance of an unbiased estimator for the second moment of source location (in both one and two dimensions) is bounded by a constant[3]. This insight leads naturally to the observation that an interferometric direct imaging system may, in theory, achieve a resolving power (in classical terms) greater than that of a equivalent diameter classical, diffraction-limited optical system, in which the resolving power is limited by Rayleigh's criterion. A directed interference pattern generates a modulation transfer function (MTF), which may be thought of as an estimator[4] of the first and second moments of the locations of the point sources comprising the underlying true image, because an MTF that aligns more closely with the power spectrum of the underlying true image will yield an image that is more similar to the true image [61]. This observation motivates much of this dissertation.

The preceding material is primarily concerned with diffraction-limited imaging scenarios in which the only source of aberrations, or path-length differences from source to focal plane, are

[3] Variance of higher moments increase inverse-polynomially with image size, which may imply interesting tradeoffs between FOV and pixel count design parameters.

[4] The parameters of this estimator correspond to commanded articulations of the secondary apertures in the case of a system like Exolife Finder.

the imaging systems themselves. In practice, astronomical images are almost never free from aberrations [47]. Ground based systems must contend with the atmosphere, and even space based systems have optical aberrations induced by design constraints or introduced inadvertently. The process of approximating the underlying true image from an imperfect sample is known as image recovery [52]. A specialized sub-discipline of image recovery has evolved to address the unique challenges attendant to imaging through an atmosphere over great distances, as is required for astronomical image recovery. Among the most commonly used techniques is multi-frame blind deconvolution (MFBD) [48]. MFBD provides a maximunm likelihood estimate the atmospheric PSF from a series of short exposure images (or specklegrams). This approach was later parallelized for practical use by Matson et al. [35] and the resulting algorithm, Physically Constrained Iterative Deconvolution (PCID), is widely-used for uncompensated imaging of space objects. Recent work by Werth et al. adapted PCID to execution on modern GPUs, resulting in a 11-fold acceleration of image recovery [59]. These postprocessing image recovery approaches are highly applicable to the image recovery subproblem explored later in this dissertation.

The task of designing a scoring criteria or metric for images, generally, without respect to a known true signal, has been treated extensively. See, for example, Stark 2013 Ch. 7. For national security applications the National Imagery Interpretation and Rating System (NIIRS) is widely used [22, 16]; the Space NIIRS (SNIIRS) is applicable to spatially resolved imagery of RSOs, and is therefore directly applicable to one task addressed in this work. Recent work has automated the estimation of the SNIIRS score of ground-based images using convolutional neural networks under a range atmospheric conditions and for a variety of target types [29]. There have also been both statistical [49] and information theoretical [39] treatments of image formation.

1.1.3 Distributed and Shaped Aperture Imaging

Sophisticated methods have been developed to compensate for atmospheric aberrations in ground-based optical systems both actively [18] and passively [48], while space based missions avoid the complication entirely. Yet these approaches often entail prohibitively high costs and complicate sensor operations. Even in the diffraction-limited case, all monolithic optical systems share a

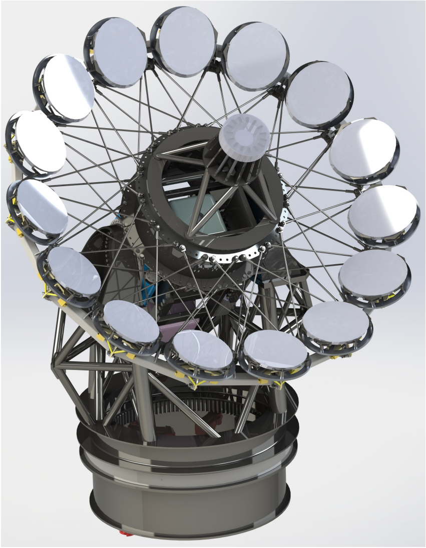

Figure 1.2: A render of the ExoLife Finder telescope design. The primary aperture are structured as an annulus of circular subapertures.

fundamental constraint: to image an object to a chosen level of sensitivity at a fixed distance with higher contrast or angular resolution, one must use a primary aperture with a larger diameter. Thus, direct imaging of very distant objects at high angular resolution requires very large-diameter apertures. Collectively, such telescopes are known as Extremely Large Telescopes (ELTs) [63]. Many ELT designs have been proposed and are under active developed, but facility cost remains a scale-limiting risk even for large-scale international scientific initiatives [4]. Telescope cost for modern glass and steel telescopes is estimated to scale linearly with the area of the primary aperture; for modern segmented arrays, the area/cost proportionality constant is on the order of 10^6 \$/ton [28]. This places ELTs far beyond the limits of practicality for many applications.

The prohibitively high cost of traditional ELTs has motivated research into alternative designs that scale more effectively. The Exo-Life Finder (ELF), show in Fig. 1.2, is an ELT that employs a lightweight tensegrity structure to suspend an annulus of small, low cost mirror segments to achieve an ELT-scale aperture with lower mass and cost [27]. The reference design mission for ELF is the detection of exoplantetary biosignatures [11] using collected spectra [7], but the sensing concept is readily extensible to other applications.

The ELF design reduces cost, but introduces a complication: the design induces path-length difference aberrations. Because the primary aperture segments are not in direct contact with one another, they cannot easily be aligned. The difference in optical path lengths caused by natural oscillations in the structure of the telescope will blur any formed image. To compensate for these aberrations, the ELF design includes articulated secondary apertures which may be rapidly adjusted to correct aberrations caused by the motion of the telescope and the atmosphere. The online control problem associated with choosing and realizing the correct articulations, given the available image data, remains unsolved. Interestingly, formulating ELF secondary articulation as a control problem naturally suggests a related question: if one can realize an aperture control policy that minimizes aberrations[5], could not one also realize a control policy that adapts the PSF to maximize recovery of a spatially extended target? Much of the work in this dissertation addresses this question.

1.2 Artificial Intelligence, Machine Learning, and Computer Vision

Artificial intelligence is a discipline within the field of computer science that is concerned with the creation of intelligent machines. What is meant here by ”intelligent” is the subject of some debate, but for the purposes of this work it will suffice to define a machine as intelligent insofar as it is successful at a task, which may be of arbitrary complexity. Machine learning is an approach to the design of machine intelligence, in which parameterized machines are adapted to data in order to perform tasks.Deep learning is a sub-discipline of machine learning in which deep neural networks are trained using error back-propagation . Deep neural networks are known to be universal function approximators [20]. As such, most recent work in deep learning is concerned with the design of model architectures, regularization, and learning frameworks that improve model training and generalization performance [46]. State of the art performance on for tasks in many previously-distinct fields of study, such as reinforcement learning (RL), natural language processing (NLP), and computer vision (CV), is now achieved almost exclusively by deep learning methods.

While computer vision has been an active field since the mid-1960s [40], the advent of the

[5] The minimization of aberrations results in the most point-like PSF possible, which maximizes recovery of point-sources.

modern era in the field is often dated to the 2012 publication of AlexNet [26]. Krizhevsky et al. contributed the first simultaneous usage of a large-scale dataset [14], convolutional neural networks for efficient weight-sharing [30], and accelerated matrix computation using graphics processing units (GPUs) [38], in the context of open-source frameworks [1] that support automatically differentiated (autodiff) computational graphs for back-propagation [44]. AlexNet marked the beginning of the widespread adoption of deep learning techniques to perceptual automation tasks.[6]

1.2.1 Reinforcement Learning

Supervised and unsupervised learning approaches enjoy broad study and fruitful applications, yet many tasks cannot easily be formulated as either surprised or unsupervised learning applications. For example, some problems involve perceptual tasks and sequential actions, but do not admit direct feedback about the quality of a given action. In this scenario, no label is available for individual perceptual instances by which to evaluate an action, nor is the task unsupervised. This complicates credit assignment, and calls for a new formulation of the learning task. We seek to learn a conditional distribution of actions, given observations of the task, that results in greater eventual reward. The objective, then, of this type of learning is to reinforce those actions that yield reward even if the relationship between the action and the eventual reward are not known. Hence, it is known as reinforcement learning. The following sections render this idea with greater precision.

Reinforcement learning also generalizes supervised and unsupervised learning. It is possible to map any supervised learning task to a specialized reinforcement learning task in which only the reward from the step immediately following an observation is available. Analogously, any unsupervised learning task may be mapped to a reinforcement learning task in which reward is always zero. In practice, mature and specialized approaches to supervised and unsupervised learning tasks outperform the more general reinforcement learning formulation.

[6] One may frame this emergence as the discovery of a new and particularly effective design pattern, sometimes called *differentiable programming*, in which information processing systems are composed of parameterized, composable feature extractors which are trained end-to-end. In this framing, AlexNet represents the discovery of a new abstraction in software engineering as presaged by Hamming in [19].

Sequential Decision Making Under Uncertainty

Many useful tasks can be formulated as the sequential specification of an action, given some information about an environment, in pursuit of a reward. These tasks, collectively known as sequential decision making problems, are often constructed around an agent that observes an environment and decides upon future actions. This formulation admits tasks with sparse and delayed rewards, in which direct credit assignment is difficult or impossible.

Any environment may be modeled as a set of possible states, S, together with a representation of the transition dynamics between states. Transitions between states are represented by a surjective endomorphism on S, P, that is often modeled as a matrix of order $|S| \times |S|$ in which each element p_{ij} represents the probability of transitioning from state s_j to state s_i, denoted $\Pr(s_i|s_j)$. For many tasks, success or failure can be modeled by associating a reward, $R(s)$, with each state, where $R : S \to \mathbb{R}$. Given only this environment model, it is possible to reason about the expected evolution of the environment [50, 21] without intervention by an agent. To model the interaction of an agent and its environment a set, A, is introduced, comprising elements representing actions that can be taken from every state $s \in S$. The influence of agent behavior on the evolution of an environment is modeled by conditioning both the transition function, $P = \Pr(s'|a, s)$, and the reward function, $R(s, a, s')$, on an action $a \in A$, taken in state s, resulting in a subsequent state s'.

Given models of the state, transition dynamics, reward, and actions, it is possible to construct an agent-agnostic representation of a sequential decision making problem. Markov decision processes (MDPs) model the collective valuation of sequential, enacted decisions made by an agent[7] contextualized in an environment [6]. An MDP may be represented as a tuple $\langle S, A, P, R, \rho_0, \gamma \rangle$. MDPs implicitly model the sequential actions: proceeding from the start state, $s_0 \sim \rho_0$ at each time step, t, an agent selects an action $a_t \in A$ given the present state $s_t \in S$. In return, the environment model provides a subsequent state, s_{t+1} and reward, r_{t+1}. This process continues indefinitely, or until some finite horizon, $t = T$, is reached. A discount factor, γ, is included in most modern MDP formulations because it facilitates differential valuation of rewards depending on when those

[7] While agent, actor, and policy are often used interchangeably, they need not be a single entity. For example, in a distributed, multi-agent decision-making system agents may process information without developing a policy, a policy may be constructed without interaction with any agent, and an actor may realize a provided action in the world without performing any information processing.

rewards were received. This is both mathematically convenient for infinite-horizon ($T = \infty$) MDPs [51], and necessary to express certain problem structures [21].

MDPs model fully observable environments: by definition, S is the set of all possible causal states [12] of the environment. Thus, given some $s \in S$, an observer possesses all information about an environment in that state. In practice, few tasks admit this degree of observability. Often, observation is limited to a subset of the state, an indirect representation of the state, or a combination of both. The problem of MDP control under incomplete state information was first introduced by Astrom [3], though this formulation was prefigured by Drake [15]. Over several decades, a model of partially observable Markov decision processes (POMDPs) that generalized MPDs was matured and standardized [60, 24]. POMPDs are formulated as a tuple, $\langle S, A, P, R, \Omega, O, \rho_0, \gamma \rangle$, which is an MDP extended to include a set of possible observations, Ω, and a conditional probability distribution, O, over that set. Of particular relevance to this work, O may represent a statistical model of a image formation, given some underlying sensor, environment, and target object state. This work follows the convention, common in recent reinforcement learning work, of neglecting the distinction between state and observation in analytic expression when describing interactions between the agents and the environment.

By definition, both MDPs and POMDPs conform to the Markov property, which specifies that future states are of independent historical states given the present state [34]. Assumption of the Markov property simplifies the analytic representation of a decision process, but also implies that each state or observation is a sufficient statistic of the history of the environment [13]. This is rarely true in practice, so the design of effective solutions to a problem modeled by an MDP is often predicated upon careful mapping from the problem itself to a model of the problem in which the observation or state is a reasonably sufficient statistic of the history of states. The design and development of a representative environment model enables the formulation of a corresponding decision making agent.

An environment model implicitly specifies the external details of a decision making agent acting in that environment. The observations and actions available to the agent, as well as the reward it will maximize, are determined by O, A, and R from the corresponding MDP, respectively. Many agent-environment interactions may also be represented using only the MDP in which the agent is contextualized. An agent interacting with an environment necessarily traverses a sequence of states and actions, which is known as trajectory, $\tau = (s_0, a_0, s_1, a_1, ...)$. An episode is a finite trajectory from a start state to a terminal state, $\tau = (s_0, a_0, ..., a_{T-1}, s_T)$, where T is the horizon. The set of finite-horizon trajectories possible in an MDP is $\tau_T = S \times (A \times S)^T$. As an agent traverses a trajectory, a sequenced of rewards are accrued. The return of that trajectory is the discounted sum of returns,

$$R(\tau) = \sum_{t=0}^{T} \gamma^t r_t, \tag{1.5}$$

where $T \in \mathbb{Z}$ for the finite-horizon return and $T = \infty$ for the infinite-horizon return.

MDPs provide a means by which to reason precisely and discriminatively about agents making sequential decisions in an environment. They do not, however, include a representation of the internal state of the agent, nor do they specify a means by which to construct agents which achieve high reward. These activities are the principal concern of the discipline of reinforcement learning.

Foundations of Reinforcement Learning

Markov decision processes provide a precise language with which to reason about tasks. Reinforcement learning supplies a compatible[8] formalism, the policy, representing solutions to tasks. Sutton and Barto [54] identify two independent lines of intellectual activity, optimal control and trial-and-error learning, which unify and terminate in modern reinforcement learning. In this section, the expressions and concepts common to many modern applications of reinforcement learning are reviewed.

A deterministic policy, η, is a mapping $\eta : S \rightarrow A$. An agent enacting a policy η chooses action $a_t \in A$ using $a_t = \eta(s_t)$. A stochastic policy, denoted π, represents a conditional probability density function (PDF) over A conditioned upon some $s \in S$. To select an action from a stochastic policy, an agent samples from π. The method of sampling is determined by the structure of the elements in A. If the actions are discrete, a multinomial Bernoulli (categorical) distribution is typically used

[8] The terms policy and agent are often used interchangeably, though it is perhaps more accurate to describe a policy as a realization of the abstraction to which the term agent corresponds

to model π. If the action space is continuous, as is the case in this dissertation, the policy must estimate the sufficient statistics of some distribution class, which may then be sampled. A diagonal multivariate Gaussian is commonly used for this purpose. The policy is sampled by estimating both the mean and standard deviation vectors of the diagonal Gaussian, μ and σ, then sampling the action space using $a_t = \mu(s_t) + \sigma(s_t) \circ z$, where $z \sim \mathcal{N}(\mu = 0, \Sigma = I)$. For convenience, the notation that follows assume a stochastic policy, π.

Because a stochastic policy models a PDF over agent actions, it can be used to represent agent-influenced state transition probabilities, $P(s_{t+1}|s_t, a_t)\pi(a_t|s_t)$. A trajectory is produced by a sequence of agent-influenced state transitions, so this same model of transition probabilities may be extended to compute the probability of a trajectory conditioned upon the policy followed by the agent. Given a finite horizon, the probability of an agent that is following π traversing a trajectory τ is given by

$$P(\tau|\pi) = \rho_0(s_0) \prod_{t=0}^{T-1} P(s_{t+1}|s_t, a_t)\pi(a_t|s_t.) \tag{1.6}$$

Naturally, one may reason about the expected performance over all possible realizations of the MDP using the marginal expectation of $P(\tau|\pi)$ over all values of τ. The expected return of the an agent acting according to π,

$$J(\pi) = \sum_{\tau \in \tau_T} R(\tau)P(\tau|\pi) = \mathop{\mathbb{E}}_{\tau \sim P(\cdot|\pi)}[R(\tau)]. \tag{1.7}$$

With these concepts defined, the objective of reinforcement learning may be expressed precisely: one must construct a policy, π, so as to maximized the expected return, J, in the context of an environment modeled as an MDP. Formulated as an optimization problem, an optimal policy π^* is sought, where

$$\pi^* = \arg\max_{\pi} J(\pi). \tag{1.8}$$

Approaches to this problem abound. A complete review is beyond the scope of this work, but may be found in several comprehensive surveys of the field [25, 55]. In the remainder of this section, the essential concepts upon which modern deep reinforcement learning (DRL) is founded are described.

The simplest method of constructing an optimal policy is to list each trajectory and to select, in each step, that action which is most likely yield the start state of the trajectory with the largest return [9]. This is, of course, infeasible in all non-trivial MDPs. However, this method introduces the idea that the expected return from a state is useful when designing policies. Bellman formalized this insight by introducing what is now known as the value function,

$$V^{\pi}(s) = \mathop{\mathbb{E}}_{\tau \sim \pi}[R(\tau)|s_0 = s], \tag{1.9}$$

which is the expected return obtained by an agent following π that begins in state s. Because $V^{\pi}(s)$ is conditioned on π, this is typically called the on-policy value function. Bellman, in the same work, introduces the principle of optimality: all actions taken by an agent executing an optimal policy are themselves optimal with respect to the preceding actions taken by that agent. This follows directly from self-consistency with the definition of an optimal policy, because the actions of an optimal policy are optimal regardless of when the decision was made. This principle involves both the value of states in the context of optimal policies, and the valuation of state-action pairs. The value function of a state, given an optimal policy can be expressed as

$$V^{*}(s) = \max_{\pi} \mathop{\mathbb{E}}_{\tau \sim \pi^{*}}[R(\tau)|s_0 = s]. \tag{1.10}$$

To represent the expected return of an action taken in a state, Eq. 1.9 is extended to include the specification of an initial action in addition to an initial state. This is known as the action-value function, and is given by

$$Q^{\pi}(s,a) = \mathop{\mathbb{E}}_{\tau \sim \pi}[R(\tau)|s_0 = s, a_0 = a]. \tag{1.11}$$

As with the optimal state value function, the optimal action-value function is obtained by specifying π as that policy which maximizes expected return,

$$Q^{*}(s,a) = \max_{\pi} \mathop{\mathbb{E}}_{\tau \sim \pi^{*}}[R(\tau)|s_0 = s, a_0 = a]. \tag{1.12}$$

[9] Bellman, in the preface to [5], describes this as the ”enumerative” approach, and coins the now-ubiquitous ”curse of dimensionality” to describe its failure mode. He subsequently coins the less pervasive but equally useful terms ”pitfalls of oversimplification” and ”morass of overcomplication.”

With these analytic expressions for value and state-value functions, it is possible articulate Bellman's principal of optimality mathematically. Under a stochastic policy, the optimality principle states that the expected return of an agent following a policy from a state, $V^{\pi}(s)$, is equal to the expected instantaneous reward achieved by following the policy (over the action space PDF implied by π) in that state, plus the γ-discounted expected return of the expected next state (over the state-space PDF implied by the MDP, under the expected action implied by π),

$$V^{\pi}(s) = \underset{a \sim \pi}{\mathbb{E}} \left[r\left(s, a\right) + \gamma \underset{s' \sim P}{\mathbb{E}} \left[V^{\pi}\left(s'\right) \right] \right], \tag{1.13}$$

where $a \sim \pi$ is shorthand for $a \sim \pi(\cdot|s)$ and $s' \sim P$ is shorthand for $s' \sim P(\cdot|a, s)$[10]. The principal of optimally also extends naturally to the state-value function, the only difference being that the instantaneous action is given, and thus is not conditioned upon the policy.

$$Q^{\pi}(s, a) = r\left(s, a\right) + \gamma \underset{s' \sim P}{\mathbb{E}} \left[\underset{a' \sim \pi}{\mathbb{E}} \left[Q^{\pi}\left(s', a'\right) \right] \right]. \tag{1.14}$$

Here, $s' \sim P$ denotes that $s' \sim P(s, a)$ and $a' \sim \pi$ denotes $a' \sim \pi(s')$. These expressions are known as the on-policy Bellman equations for MDPs. Substituting an arbitrary policy, π, with an optimal policy, π^*, yields the Bellman optimality equations for an MDP. The optimal Bellman value function is given by

$$V^{*}(s) = \max_{a} \left[r\left(s, a\right) + \gamma \underset{s' \sim P}{\mathbb{E}} \left[V^{*}\left(s'\right) \right] \right], \tag{1.15}$$

while the optimal Bellman state-value function is given by

$$Q^{*}(s, a) = r\left(s, a\right) + \gamma \underset{s' \sim P}{\mathbb{E}} \left[\max_{a'} \; Q^{*}\left(s', a'\right) \right]. \tag{1.16}$$

The optimal Bellman value and state-value equations for MDPs[11] are the foundation upon which

[10] TODO: somehow cite or discuss the Spinning Up readthedocs, which is almost identical to this - but which I only found AFTER writing it.

[11] Bellman equations and the optimality principle are central to an entire approach to sequential problem solving known as Dynamic Programming. The MDP formulation of the reinformcnent learning problem is only one application within this broader field of study.

most modern DRL methods are built, and were a central feature of the next major advance in reinforcement learning: value iteration and temporal difference methods.

One approach to reinforcement learning is to directly estimate the optimal value or action-value functions for an MDP. Given the optimal value function, one can construct an optimal policy by applying the value function to all reachable states and selecting the action which provides the highest expected probability of transitioning to the state with the highest estimated value. The optimality of the resulting policy is implicit in the definition of $V^*(s)$, assuming that the problem has optimal substructure. An optimal state-value function further simplifies policy construction by providing an estimate of the value of all possible actions from any state. Yet these approaches, collectively known as *value-function methods*, require the optimal state or state-value function which must, in turn, be estimated. To estimate the optimal value function Bellman introduces *value iteration*, an application of dynamic programming. Value iteration refines an estimate of $V^*(s)$ by searching A for the action that maximizes $V(s)$ and setting the estimate of $V^*(s)$ to that value.

Alternatively, one may optimize the policy to improve the expected reward. This family of approaches, known as *policy iteration* methods, was first introduced by Howard [21] shortly after the discovery of dynamic programming. Policy iteration improves the expected reward of an agent conditioned upon a randomly initialized policy by evaluating $V^\pi(s)$, then modifying the policy so as to improve the expected reward. Both value iteration and policy iteration estimate the value function, but policy iteration computes the expected value of $V^\pi(s)$ given the behavior distribution defined by the policy, rather than searching for the action that maximizes $V^*(s)$ directly.

Policy and value iteration are often inapplicable to practical control and decision-making problems because they require the environment dynamics (i.e., P and R) to be known. To move beyond these dependencies, the environment dynamics, sometimes called the environment model, must be estimated from experience; the model estimate may be explicit, such as a matrix representing P, or implicit, as is the case when policy parameters are directly adjusted to improve expected return. To build an estimate of the dynamics of an environment, we may observe an agent contextualized in that environment and record the responses of the environment to the agents actions. Applications leveraging this approach are known as Monte-Carlo (MC) methods which can be realized either on-policy or off-policy and, unlike policy and value iteration, do not involve bootstrapping. MC methods may be used to directly estimate the value and state-value functions by sequentially sampling from the environment and storing the reward achieved at each state, or the return of a full trajectory [36]. MC policy evaluation is widely-used to assess agent model performance in recent RL work, and most methods of interacting with an environment may be formulated as MC sampling from the underlying MDP.

MC methods require each trajectory to proceed until a terminal state is reached, which is often computationally expensive. To retain the benefits of model-free RL without the computational costs attendant to MC methods, Sutton developed temporal-difference (TD) learning, which estimate the value and state-value functions by bootstrapping [53]. TD learning proceeds by estimating a value for each state, which is updated in the direction of the expected return (or target), which is in turn value of the next state in the trajectory. As a TD learning algorithm iterates through a trajectory (or accumulates off-policy transition samples), the observed reward in each state is are partially credited to the preceding state. Intuitively, the uncertainty in the expected return of a trajectory will tend to decrease as the number of remaining steps in that trajectory decreases, as there are fewer opportunities for large changes in the discounted return. In the limiting case (i.e., the final step) there is no uncertainty, and the remaining return is known exactly. TD learning propagates this certainty backward through the states of a trajectory, one state at a time. Sutton also generalized TD learning to bootstrap using an arbitrary number, λ, of steps along the trajectory; this method is known as TD(λ).

TD(λ) serves as the foundational environment sampling approach for most implementations of two classic model-free RL approaches, SARSA and Q-Learning, both of which are extensions of value iteration for unknown environment dynamics. While both methods estimate the state-value function using environment samples, SARSA is on-policy [45]. Q-Learning was the first approach to successfully demonstration off-policy estimation of the action-value function without known environment dynamics [58].

This dissertation builds on the lineage of reinforcement learning approaches known as policy gradient methods. Policies, both deterministic and stochastic, may be represented as parameterized functions. In this formulation, the adaptation of an agent to a problem is done by selecting that agents parameters, so as to maximize its expected return. Policy gradient methods select parameters by computing the gradient of the expected reward and ascending that gradient. Because policy gradient methods directly construct the policy, rather than relying on an estimate of the state or state-value function, they are much more computationally efficient for tasks involving state or action spaces.

Deep Reinforcement Learning

Agents, as defined, admit any realization of a policy, such as finite automata, logic trees, analytic expressions, and parameterized models, or any combination thereof. Most modern applications of reinforcement learning are built using neural networks as parameterized function approximators, collective denoted π_θ, where θ is a parameterization of the policy, and trained using backpropagation of errors. This combination is known as deep reinforcement learning (DRL).

The use of a neural network as a parameterized policy was pioneered by Williams, and is known as the REINFORCE algorithm [62]. This approach introduced introduced the use of neural networks as policies and also provided the first application of policy gradients methods to direct policy optimization. This work formalized the notion of policy gradients, which are defined to be the gradients of the objective function, J, with respect to the parameters, θ, of a parameterized function approximating a policy, π_θ. Later work by Tesauro applied policy gradient methods to develop a backgammon planning agent known as TD-Gammon [56], greatly surpassing the prior state of the art.

The control task addressed in this dissertation requires that a policy map from an image to an articulation command. This problem is intrinsically high-dimensional, which presents numerous challenges to classical approaches to sequential control. Historically, approaches to problems with these features involved extensive hand-designed feature engineering, and rarely succeeded in practice. In 2013, shortly after AlexNet [26] achieved a substantial improvement in the ImageNet large scale visual recognition challenge state of the art using CNNs [14], Mnih et al. demonstrated the viability of deep learning for visual reinforcement learning problems [37]. Mnih et al. introduces the

Deep Q Network (DQN) technique, in which an ϵ-greedy agent is used to construct an off-policy experience replay buffer [33], which is in turn used to estimate the policy gradients with respect to a temporal difference loss function. The temporal difference function is simply the difference between the right and left side of the Bellman state-value function (Eq. 1.16),

$$\delta = \left(r\left(s, a\right) + \gamma \mathop{\mathbb{E}}_{s' \sim P} \left[\mathop{\mathbb{E}}_{a' \sim \pi} \left[Q^{\pi}\left(s', a'\right) \right] \right] \right) - Q^{\pi}(s, a). \tag{1.17}$$

Thus, for a Q-function represented by a network parameterized by θ to be self-consistent, its temporal-difference error should be near zero. Mhin et al. translate this constraint into a loss function,

$$L_i(\theta_i) = \left(r\left(s, a\right) + \gamma \mathop{\mathbb{E}}_{s' \sim P} \left[\max_{a'} \left[Q\left(s', a'\right) \middle| \, \theta_{i-1} \right] \right] \right) - Q(s, a \mid \theta_i), \tag{1.18}$$

which is minimized by batch stochastic gradient descent using (s, a, r, s') tuples stored in the experience replay buffer. This work provided the first demonstration of effective DRL for control tasks that require high-dimensional visual perception, but was limited to tasks of comparatively modest input dimensionality of 84×84, after hand designed preprocessing. While very influential, this early work was not easily extensible to real world tasks because the model learns a behaviour distribution that is based on the assumption that the control dynamics (i.e., the relationship between a command and the change it causes in the environment state) are, themselves, ϵ-greedy; in practice, the control dynamics of real systems can be arbitrarily complex.

Subsequent work by Levine et al. provides further evidence that joint training of the control and perpetual task models improves task performance compared to training both in isolation, explores the practical application of DRL to robotic process automation (RPA) tasks, and scales visual input dimensionality to 240×240 [31]. The authors introduce guided policy search, an approach that combines a controller dynamics model, which learns the conditional control distribution given the system state (i.e., joint articulations and object positions), with a visuomotor model that maps from input images to output controls. Intuitively, the controller model simplifies end-to-end learning by providing a model-based representation of the system dynamics, while on-policy training of the visuomotor policy learns to predict high-value commands given images of the task. During training, the controller requires access to the fully-observable system state, which is not available outside of highly controlled training environments. However, because the visuomotor policy maps images to controls, it may be used after training without the high-quality instrumentation necessary to optimize the the controller model during training. Thus, this method leverages information from a fully-observable environment during training, but only requires a partially observable environment at test time.

As with other application domains of deep learning, DRL relies on benchmarking to assess the progress of the field. Recent work by Duan et al. [17] provides comprehensive benchmarking of several continuous control policy optimization algorithms across 31 continuous control tasks. The tasks are divided into four categories, including classical control problems, higher-dimensional locomotion tasks, partially observable tasks, and hierarchical tasks. The partially observable tasks were constructed by randomizing features of simulated system, inducing a stochastic delay before an action is applied to the environment, and reducing or degrading sensor data available to the model. Task feature randomization and stochastic action latency are both relevant to this dissertation; Duan et al. report an approximately 67% reduction in cumulative expected reward when these task complications are introduced. This suggests that methods to improve generalization to partially observable scenarios may be required. Additionally, this study provides initial benchmark results for hierarchical tasks, in which a low-level task must be solved in pursuit of a higher level goal. Across nine algorithms, each subject to hyperparameter search with five repetitions per configuration, no algorithm produced a policy that achieved a cumulative reward that outperformed a randomly acting agent on any task. Again, hierarchical problem solving is a feature of the task described in this work, which suggests that methods related to hierarchical learning may be applicable.

1.3 Conclusion

The approach and objective of this dissertation may now be state concisely and positioned with respect to prior literature. The work described in this dissertation is a deep reinforcement learning approach to a partially-observable, hierarchical visuomotor task involving high-frequency articulation of optical elements to manipulate optical diffraction, thereby enabling the reconstruction of extended astronomical object imagery.

A review of the literature suggests several potential contributions related to the topic of this dissertation. As deep learning reinforcement learning methods continue to mature, their application to scientific and high contrast imagery is likely to grow. There is comparatively little published work at the intersection of scientific image processing and DRL for visuomotor tasks, but future scientific and industrial application may be enabled by control of this kind. Additionally, the high-frequency and hierarchical nature of the task may lead to new methods in time-varying processing of information at different levels of the task hierarchy.

BIBLIOGRAPHY

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, and Michael Isard. Tensorflow: A system for large-scale machine learning. In *12th ${$USENIX$}$ Symposium on Operating Systems Design and Implementation (${$OSDI$}$ 16)*, pages 265–283, 2016.

[2] Shan Zheng Ang, Ranjith Nair, and Mankei Tsang. Quantum limit for two-dimensional resolution of two incoherent optical point sources. *Physical Review A*, 95(6):063847, June 2017.

[3] Karl Johan Åström. Optimal Control of Markov Processes with Incomplete State Information I. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965.

[4] Anthony J. Beasley, Mark Dickinson, Eric J. Murphy, Sidney Wolff, and Michael H. Wong. Astro2020 APC White Paper Multiwavelength Astrophysics in the Era of the ngVLA and the US ELT Program. 2020.

[5] Richard Bellman. *Dynamic Programming.* Princeton University Press, Princeton, NJ, 1957.

[6] Richard Bellman. A Markovian decision process. *Journal of mathematics and mechanics*, 6(5):679–684, 1957.

[7] S. V. Berdyugina, J. R. Kuhn, M. Langlois, G. Moretto, J. Krissansen-Totton, D. Catling, J. L. Grenfell, T. Santl-Temkiv, K. Finster, J. Tarter, F. Marchis, H. Hargitai, and D. Apai. The Exo-Life Finder (ELF) telescope: New strategies for direct detection of exoplanet biosignatures and technosignatures. In *Ground-Based and Airborne Telescopes VII*, volume 10700, pages 1453–1466. SPIE, October 2018.

[8] E. Bettens, D. Van Dyck, A. J. den Dekker, J. Sijbers, and A. van den Bos. Model-based two-object resolution from observations having counting statistics. *Ultramicroscopy*, 77(1):37–48, May 1999.

[9] W. S. Boyle and G. E. Smith. Charge coupled semiconductor devices. *The Bell System Technical Journal*, 49(4):587–593, April 1970.

[10] James B. Campbell and Randolph H. Wynne. *Introduction to Remote Sensing.* Guilford Press, 2011.

[11] David C. Catling, Joshua Krissansen-Totton, Nancy Y. Kiang, David Crisp, Tyler D. Robinson, Shiladitya DasSarma, Andrew J. Rushby, Anthony Del Genio, William Bains, and Shawn Domagal-Goldman. Exoplanet Biosignatures: A Framework for Their Assessment. *Astrobiology*, 18(6):709–738, June 2018.

[12] James P. Crutchfield and Christopher J. Ellison. The Past and the Future in the Present. *arXiv:1012.0356 [cs, math, nlin, stat]*, December 2010.

[13] James P. Crutchfield, Christopher J. Ellison, and John R. Mahoney. Time's Barbed Arrow: Irreversibility, Crypticity, and Stored Information. *Physical Review Letters*, 103(9):094101, August 2009.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[15] Alvin W. Drake. *Observation of a Markov Process through a Noisy Channel.* PhD thesis, Massachusetts Institute of Technology, 1962.

[16] Ronald G. Driggers, Paul G. Cox, and Michael Kelley. National imagery interpretation rating system and the probabilities of detection, recognition, and identification. *Optical Engineering*, 36(7):1952–1959, 1997.

[17] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking Deep Reinforcement Learning for Continuous Control. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1329–1338. PMLR, June 2016.

[18] R. Q. Fugate, B. L. Ellerbroek, C. H. Higgins, M. P. Jelonek, W. J. Lange, A. C. Slavin, W. J. Wild, D. M. Winker, J. M. Wynia, J. M. Spinhirne, B. R. Boeke, R. E. Ruane, J. F. Moroney, M. D. Oliker, D. W. Swindle, and R. A. Cleis. Two generations of laser-guide-star

adaptive-optics experiments at the Starfire Optical Range. *JOSA A*, 11(1):310–324, January 1994.

[19] Richard R. Hamming. *Art of Doing Science and Engineering: Learning to Learn.* CRC Press, 1997.

[20] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[21] Ronald A. Howard. Dynamic programming and markov processes. 1960.

[22] John M. Irvine. National imagery interpretability rating scales (NIIRS): Overview and methodology. In *Airborne Reconnaissance XXI*, volume 3128, pages 93–103. International Society for Optics and Photonics, 1997.

[23] Francis Arthur Jenkins and Harvey Elliott White. Fundamentals of optics. *Indian Journal of Physics*, 25:265–266, 1957.

[24] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, May 1998.

[25] Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[27] J. R. Kuhn, S. V. Berdyugina, J.-F. Capsal, M. Gedig, M. Langlois, G. Moretto, and K. Thetpraphi. The Exo-Life Finder Telescope (ELF): Design and beam synthesis concepts. In *Ground-Based and Airborne Telescopes VII*, volume 10700, pages 344–350. SPIE, July 2018.

[28] J. R. Kuhn, S. V. Berdyugina, M. Langlois, G. Moretto, E. Thiébaut, C. Harlingten, and D. Halliday. Looking beyond 30m-class telescopes: The Colossus project. In *Ground-Based and Airborne Telescopes V*, volume 9145, pages 533–540. SPIE, July 2014.

[29] Trent Kyono, Jacob Lucas, Michael Werth, Brandoch Calef, Ian McQuaid, and Justin Fletcher. Machine learning for quality assessment of ground-based optical images of satellites. *Optical Engineering*, 59(5):051403, January 2020.

[30] Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[31] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[32] Thomas Lillesand, Ralph W. Kiefer, and Jonathan Chipman. *Remote Sensing and Image Interpretation*. John Wiley & Sons, 2015.

[33] Longxin Lin. Reinforcement learning for robots using neural networks. *undefined*, 1992.

[34] Andrei Andreevich Markov. Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga. *Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete*, 15(135-156):18, 1906.

[35] Charles L. Matson, Kathy Borelli, Stuart Jefferies, Jr Charles C. Beckner, E. Keith Hege, and Michael Lloyd-Hart. Fast and optimal multiframe blind deconvolution algorithm for high-resolution ground-based imaging of space objects. *Applied Optics*, 48(1):A75–A92, January 2009.

[36] Donald Michie and Roger A. Chambers. BOXES: An experiment in adaptive control. *Machine intelligence*, 2(2):137–152, 1968.

[37] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with Deep Reinforcement Learning. *arXiv:1312.5602 [cs]*, December 2013.

[38] Kyoung-Su Oh and Keechul Jung. GPU implementation of neural networks. *Pattern Recognition*, 37(6):1311–1314, June 2004.

[39] J.A. O'Sullivan, R.E. Blahut, and D.L. Snyder. Information-theoretic image formation. *IEEE Transactions on Information Theory*, 44(6):2094–2123, October 1998.

[40] Seymour A. Papert. The Summer Vision Project. July 1966.

[41] Martin Paúr, Bohumil Stoklasa, Jai Grover, Andrej Krzic, Luis L. Sánchez-Soto, Zdeněk Hradil, and Jaroslav Řeháček. Tempering Rayleigh's curse with PSF shaping. *Optica*, 5(10):1177–1180, October 2018.

[42] Sripad Ram, E Sally Ward, and Raimund J Ober. Beyond Rayleigh's criterion: A resolution measure with application to single-molecule microscopy. In *Pnas;103/12/4457*, 2006.

[43] Rayleigh. Investigations in optics, with special reference to the spectroscope. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 8(49):261–274, October 1879.

[44] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[45] Gavin A. Rummery and Mahesan Niranjan. *On-Line Q-learning Using Connectionist Systems*, volume 37. Citeseer, 1994.

[46] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, January 2015.

[47] Daniel J. Schroeder. *Astronomical Optics*. Elsevier, 1999.

[48] Timothy J. Schulz. Multiframe blind deconvolution of astronomical images. *JOSA A*, 10(5):1064–1073, 1993.

[49] M. Shahram and P. Milanfar. Statistical and Information-Theoretic Analysis of Resolution in Imaging. *IEEE Transactions on Information Theory*, 2006.

[50] R. Sittler. Systems Analysis of Discrete Markov Processes. *IRE Transactions on Circuit Theory*, 3(4):257–266, December 1956.

[51] Edward J. Sondik. The Optimal Control of Partially Observable Markov Processes over the Infinite Horizon: Discounted Costs. *Operations Research*, 26(2):282–304, 1978/03//Mar/Apr78.

[52] Henry Stark. *Image Recovery: Theory and Application.* Elsevier, 2013.

[53] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[54] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* MIT press, 2018.

[55] Csaba Szepesvári. Algorithms for Reinforcement Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4(1):1–103, January 2010.

[56] Gerald Tesauro. Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38(3):58–68, 1995.

[57] Mankei Tsang, Ranjith Nair, and Xiao-Ming Lu. Quantum Theory of Superresolution for Two Incoherent Optical Point Sources. *Physical Review X*, 6(3):031033, August 2016.

[58] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

[59] Michael Werth, Brandoch Calef, Kevin Roe, and Amanda Conti. LUCID: Accelerating Image Reconstructions of LEO Satellites Using GPUs. In *2020 IEEE Aerospace Conference*, pages 1–11, March 2020.

[60] Douglas J. White. A survey of applications of Markov decision processes. *Journal of the operational research society*, 44(11):1073–1096, 1993.

[61] Charles Sumner Williams and Orville A. Becklund. *Introduction to the Optical Transfer Function*, volume 112. SPIE Press, 2002.

[62] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992.

[63] Michael H. Wong, Karen J. Meech, Mark Dickinson, Thomas Greathouse, Richard J. Cartwright, Nancy Chanover, and Matthew S. Tiscareno. Transformative Planetary Science with the US ELT Program. *arXiv:2009.08029 [astro-ph]*, September 2020.

[64] Sisi Zhou and Liang Jiang. Modern description of Rayleigh's criterion. *Physical Review A*, 2019.