Feature-Based Aggregation and Deep Reinforcement Learning: A Survey and Some New Implementations

Dimitri P. Bertsekas†

Abstract

In this paper we discuss policy iteration methods for approximate solution of a finite-state discounted Markov decision problem, with a focus on feature-based aggregation methods and their connection with deep reinforcement learning schemes. We introduce features of the states of the original problem, and we formulate a smaller “aggregate” Markov decision problem, whose states relate to the features. The optimal cost function of the aggregate problem, a nonlinear function of the features, serves as an architecture for approximation in value space of the optimal cost function or the cost functions of policies of the original problem. We discuss properties and possible implementations of this type of aggregation, including a new approach to approximate policy iteration. In this approach the policy improvement operation combines feature-based aggregation with reinforcement learning based on deep neural networks, which is used to obtain the needed features. We argue that the cost function of a policy may be approximated much more accurately by the nonlinear function of the features provided by aggregation, than by the linear function of the features provided by deep reinforcement learning, thereby potentially leading to more effective policy improvement.

Contents

1. Introduction
   1.1 Alternative Approximate Policy Iteration Methods
   1.2 Terminology

2. Approximate Policy Iteration: An Overview
   2.1 Direct and Indirect Approximation
   2.2 Indirect Methods Based on Projected Equations
   2.3 Indirect Methods Based on Aggregation
   2.4 Implementation Issues

3. Approximate Policy Iteration Based on Neural Networks

4. Feature-Based Aggregation Framework
   4.1 The Aggregate Problem
   4.2 Feature Formation by Using Scoring Functions
   4.3 Illustrative Examples
   4.4 Multistep Aggregation

5. Policy Iteration with Feature-Based Aggregation and a Neural Network

6. Concluding Remarks

7. References
1. INTRODUCTION

We consider a discounted infinite horizon dynamic programming (DP) problem with \( n \) states, which we denote by \( i = 1, \ldots, n \). State transitions \((i, j)\) under control \( u \) occur at discrete times according to transition probabilities \( p_{ij}(u) \), and generate a cost \( \alpha^k g(i, u, j) \) at time \( k \), where \( \alpha \in (0, 1) \) is the discount factor. We consider deterministic stationary policies \( \mu \) such that for each \( i \), \( \mu(i) \) is a control that belongs to a constraint set \( U(i) \). We denote by \( J_\mu(i) \) the total discounted expected cost of \( \mu \) over an infinite number of stages starting from state \( i \), and by \( J^*(i) \) the minimal value of \( J_\mu(i) \) over all \( \mu \). We denote by \( J_\mu \) and \( J^* \) the vectors of \( \mathbb{R}^n \) that have components \( J_\mu(i) \) and \( J^*(i) \), \( i = 1, \ldots, n \), respectively. As is well known, \( J_\mu \) is the unique solution of the Bellman equation for policy \( \mu \):

\[
J_\mu(i) = \sum_{j=1}^{n} p_{ij}(\mu(i)) \left( g(i, \mu(i), j) + \alpha J_\mu(j) \right), \quad i = 1, \ldots, n, \tag{1.1}
\]

while \( J^* \) is the unique solution of the Bellman equation

\[
J^*(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) \left( g(i, u, j) + \alpha J^*(j) \right), \quad i = 1, \ldots, n. \tag{1.2}
\]

In this paper, we survey several ideas from aggregation-based approximate DP and deep reinforcement learning, all of which have been essentially known for some time, but are combined here in a new way. We will focus on methods of approximate policy iteration (PI for short), whereby we evaluate approximately the cost vector \( J_\mu \) of each generated policy \( \mu \). Our cost approximations use a feature vector \( F(i) \) of each state \( i \), and replace \( J_\mu(i) \) with a function that depends on \( i \) through \( F(i) \), i.e., a function of the form

\[
\hat{J}_\mu(F(i)) \approx J_\mu(i), \quad i = 1, \ldots, n.
\]

We refer to such \( \hat{J}_\mu \) as a feature-based approximation architecture.

At the typical iteration of our approximate PI methodology, the cost vector \( J_\mu \) of the current policy \( \mu \) is approximated using a feature-based architecture \( \hat{J}_\mu \), and a new policy \( \hat{\mu} \) is then generated using a policy "improvement" procedure; see Fig. 1.1. The salient characteristics of our approach are two:

(a) The feature vector \( F(i) \) may be obtained using a neural network or similar scheme that automatically constructs features.

(b) The policy improvement, which generates \( \hat{\mu} \) is based on a DP problem that involves feature-based aggregation.

By contrast, the standard policy improvement method is based on the one-step lookahead minimization

\[
\hat{\mu}(i) \in \arg \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) \left( g(i, u, j) + \alpha \hat{J}_\mu(F(j)) \right), \quad i = 1, \ldots, n; \tag{1.3}
\]
Approximate Policy Evaluation

Policy Improvement

(May Involve a Neural Network)

(May Involve Aggregation)

Figure 1.1 Schematic view of feature-based approximate PI. The cost $J_\mu(i)$ of the current policy $\mu$ starting from state $i$ is replaced by an approximation $\hat{J}_\mu(F(i))$ that depends on $i$ through its feature vector $F(i)$. The feature vector is assumed independent of the current policy $\mu$ in this figure, but in general could depend on $\mu$.

or alternatively, on multistep lookahead, possibly combined with Monte-Carlo tree search. We will argue that our feature-based aggregation approach has the potential to generate far better policies at the expense of a more computation-intensive policy improvement phase.

1.1 Alternative Approximate Policy Iteration Methods

A survey of approximate PI methods was given in 2011 by the author [Ber11a], and focused on linear feature-based architectures. These are architectures where $F(i)$ is an $s$-dimensional vector

$$F(i) = (F_1(i), \ldots, F_s(i)) \in \mathbb{R}^s,$$

and $\hat{J}_\mu$ is a linear function of $F$, i.e.,

$$\hat{J}_\mu(F(i)) = \sum_{\ell=1}^{s} F_\ell(i)r_\ell, \quad i = 1, \ldots, n,$$

for some scalar weights $r_1, \ldots, r_s$.† We considered in [Ber11a] two types of methods:

(a) Projected equation methods, including temporal difference methods, where policy evaluation is based on simulation-based matrix inversion methods such as LSTD($\lambda$), or stochastic iterative methods such as TD($\lambda$), or variants of $\lambda$-policy iteration such as LSPE($\lambda$)].

(b) General aggregation methods (not just the feature-based type considered here).

These methods will be briefly discussed in Section 2. The present paper is complementary to the survey [Ber11a], and deals with approximate PI with nonlinear feature-based architectures, including some that are generated with the aid of neural networks.

† In our notation, $\mathbb{R}^s$ denotes the Euclidean space of $s$-dimensional real vectors, and $\mathbb{R}$ denotes the real line.
An important advantage of linear feature-based architectures is that given the form of the feature vector $F(\cdot)$, they can be trained with linear least squares-type methods. However, determining good features may be a challenge in general. Neural networks resolve this challenge through training that constructs automatically features and simultaneously combines the components of the features linearly with weights. This is commonly done by cost fitting/nonlinear regression, using a large number of state-cost sample pairs, which are processed through a sequence of alternately linear and nonlinear layers (see Section 3). The outputs of the final nonlinear layer are the features, which are then processed by a final linear layer that provides a linear combination of the features as a cost function approximation.

The idea of representing cost functions in terms of features of the state in a context that we may now call approximate DP goes back to the work of Shannon on chess [Sha50]. The work of Samuel [Sam59], [Sam67] on checkers extended some of Shannon’s algorithmic schemes and introduced temporal difference ideas that motivated much subsequent research. The use of neural networks to simultaneously extract features of the optimal or the policy cost functions, and construct an approximation to these cost functions was also investigated in the early days of reinforcement learning; some of the original contributions that served as motivation for much subsequent work are Werbos [Wer77], Barto, Sutton, and Anderson [BSA83], Christensen and Korf [ChK86], Holland [Hol86], and Sutton [Sut88]. The use of a neural network as a cost function approximator for a challenging DP problem was first demonstrated impressively in the context of the game of backgammon by Tesauro [Tes92], [Tes94], [Tes95], [Tes02]. In Tesauro’s work the parameters of the network were trained by using a form of temporal differences (TD) learning, and the features constructed by the neural network were supplemented by some handcrafted features.† Following Tesauro’s work, the synergistic potential of approximations using neural network or other architectures, and DP techniques had become apparent, and it was laid out in an influential survey paper by Barto, Bradtke and Singh [BBS95]. It was then systematically developed in the neuro-dynamic programming book by Bertsekas and Tsitsiklis [BeT96], and the reinforcement learning book by Sutton and Barto [SuB98]. Subsequent books on approximate DP and reinforcement learning, which discuss approximate PI, among other techniques, include Cao [Cao07], Busoniu et. al. [BBD10], Szepesvari [Sze10], Powell [Pow11], Chang, Fu, Hu, and

† Tesauro also constructed a different backgammon player, trained by a neural network, but with a supervised learning approach, which used examples from human expert play [Tes89] (he called this approach “comparison learning”). However, his TD-based algorithm performed substantially better, and its success has been replicated by others, in both research and commercial programs. Tesauro and Galperin [TeG96] proposed still another approach to backgammon, based on a rollout strategy, which resulted in an even better playing program (see [Ber17] for an extensive discussion of rollout as a general approximate DP approach). At present, rollout-based backgammon programs are viewed as the most powerful in terms of performance, but are too time-consuming for real-time play. They have been used in a limited diagnostic way to assess the performance of other programs. A list of articles on computer backgammon may be found at http://www.bkgm.com/articles/page07.html.
Marcus [CFH13], Vrabie, Vamvoudakis, and Lewis [VVL13], and Gosavi [Gos15]. To these, we may add the edited collections by Si, Barto, Powell, and Wunsch [SBP04], Lewis, Liu, and Lendaris [LLL08], and Lewis and Liu [LeL12], which contain several survey papers.

The original ideas on approximate PI were enriched by further methodological research such as rollout (Abramson [Abr90], Tesauro and Galperin [TeG96], Bertsekas, Tsitsiklis, and Wu [BTW97], Bertsekas and Castanon [BeC99]; see the surveys in [Ber13], [Ber17]), adaptive simulation and Monte Carlo tree search (Chang, Hu, Fu, and Marcus [CFH05], [CFH13], Coulom [Cou06]; see the survey by Browne et al. [BPW12]), and deep neural networks (which are neural networks with many and suitably specialized layers; see for the example the book by Goodfellow, Bengio, and Courville [GBC16], the textbook discussion in [Ber17], Ch. 6, and the recent surveys by Schmidhuber [Sch15], Arulkumaran et al. [ADB17], Liu et al. [LWL17], and Li [Li17]).

A recent impressive success of the deep neural network-based approximate PI methodology is the AlphaZero program, which attained a superhuman level of play for the games of chess, Go, and others (see Silver et al. [SHS17]). A noteworthy characteristic of this program is that it does not use domain-specific knowledge (i.e., handcrafted features), but rather relies entirely on the deep neural network to construct features for cost function approximation (at least as reported in [SHS17]). Whether it is advisable to rely exclusively on the neural network to provide features is an open question, as earlier investigations, including the ones by Tesauro noted earlier, suggest that using additional problem-specific hand-crafted features can be very helpful in the context of approximate DP. Except for the use of deep rather than shallow neural networks (which are used in backgammon), the AlphaZero algorithm is similar to several other algorithms that have been proposed in the literature and/or have been implemented in the past. It can be viewed as a conceptually straightforward implementation of approximate PI, using Monte Carlo tree search and a single neural network to construct a cost and policy approximation, and does not rely on any fundamentally new ideas or insightful theoretical analysis. Conceptually, it bears considerable similarity to Tesauro’s TD-Gammon program. Its spectacular success may be attributed to the skillful implementation of an effective mix of known ideas, coupled with great computational power.

We note that the ability to simultaneously extract features and optimize their linear combination is not unique to neural networks. Other approaches that use a multilayer architecture have been proposed (see the survey by Schmidhuber [Sch15]), including the Group Method for Data Handling (GMDH), which is principally based on the use of polynomial (rather than sigmoidal) nonlinearities. The GMDH method was investigated extensively in the Soviet Union starting with the work of Ivakhnenko in the late 60s; see e.g., [Iva68]. It has been used in a large variety of applications, and its similarities with the neural network methodology have been noted (see the survey by Ivakhnenko [Iva71], and the large literature summary at the web site http://www.gmdh.net). Most of the GMDH research relates to inference-type problems. We
are unaware of any application of GMDH in the context of approximate PI, but we believe this to be a
fruitful area of investigation. In any case, the feature-based PI ideas of the present paper apply equally well
in conjunction with GMDH networks as with the neural networks described in Section 3.

While automatic feature extraction is a critically important aspect of neural network architectures, the
linearity of the combination of the feature components at the final layer may be a limitation. A nonlinear
alternative is based on aggregation, a dimensionality reduction approach to address large-scale problems.
This approach has a long history in scientific computation and operations research (see for example Bean,
Birge, and Smith [BBS87], Chatelin and Miranker [ChM82], Douglas and Douglas [DoD93], and Rogers et.
al. [RPW91]). It was introduced in the simulation-based approximate DP context, mostly in the form of
value iteration; see Singh, Jaakkola, and Jordan [SJJ95], Gordon [Gor95], Tsitsiklis and Van Roy [TsV96],
and the neuro-dynamic programming book [BeT96], Sections 3.1.2 and 6.7. Aggregation architectures based
on features were discussed in Section 3.1.2 of the neuro-dynamic programming book [BeT96], and in Section
6.5 of the author’s DP book [Ber12], including the feature-based architecture that is the focus of the present
paper. They have the capability to produce policy cost function approximations that are nonlinear functions
of the feature components, thus yielding potentially more accurate approximations.

Feature-based aggregation assumes a given form of feature vector, so for problems where good features
are not apparent, it needs to be modified or to be supplemented by a method that can construct features from
training data. Motivated by the reported successes of deep reinforcement learning with neural networks, we
propose a two-stage process: first use a neural network or related scheme to construct good features for cost
approximation, and then use there features to construct a nonlinear feature-based aggregation architecture.
In effect we are proposing a new way to implement neural network-based approximate PI: retain the policy
evaluation phase which uses a neural network, but replace the policy improvement phase with the solution of
an aggregate DP problem. This DP problem involves the features that are generated by the neural network
(possibly together with other handcrafted features). Its dimension may be reduced to a manageable level by
sampling, and cost function values are generalized to the entire feature space by linear interpolation. Our
alternative policy improvement phase may be more complicated, but may be far more powerful as it relies
on the potentially more accurate function approximation provided by a nonlinear combination of features.

Aside from the power brought to bear by nonlinearly combining features, let us also note some other
advantages that are generic to aggregation. In particular:

(a) Aggregation aims to solve an “aggregate” DP problem, itself an approximation of the original DP
problem, in the spirit of coarse-grid discretization of large state space problems. As a result, aggregation
methods enjoy the stability and policy convergence guarantee of exact PI. By contrast, temporal
difference-based and other PI methods can suffer from convergence difficulties such as policy oscil-
lations and chattering (see e.g., [BeT96], [Ber11a], [Ber12]). A corollary to this is that when an
aggregation scheme performs poorly, it is easy to identify the cause: it is the quantization error due to approximating a large state space with a smaller “aggregate” space. The possible directions for improvement (at a computational cost of course) are then clear: introduce additional aggregate states, and increase/improve these features.

(b) Aggregation methods are characterized by error bounds, which are generic to PI methods that guarantee the convergence of the generated policies. These error bounds are better by a factor $1 - \alpha$ compared to the corresponding error bounds for methods where policies need not converge, such as generic temporal difference methods with linear cost function approximation [see Eqs. (2.2) and (2.3) in the next section].

Let us finally note that the idea of using a deep neural network to extract features for use in another approximation architecture is central in the Deepchess program by David, Netanyahu, and Wolf [DNW16], which was estimated to perform at the level of a strong grandmaster, and at the level of some of the strongest computer chess programs. In this work the features were used, in conjunction with supervised learning and human grandmaster play selections, to train a deep neural network to compare any pair of legal moves in a given chess position, in the spirit of Tesauro’s comparison training approach [Tes89]. By contrast in our proposal the features are used to formulate an aggregate DP problem, which can be solved by exact methods, including some that are based on simulation.

The paper is organized as follows. In Section 2, we provide context for the subsequent developments, and summarize some of the implementation issues in approximate PI methods. In Section 3, we review some of the central ideas of approximate PI based on neural networks. In Section 4, we discuss PI ideas based on feature-based aggregation, assuming good features are known. Finally, in Section 5, we describe some of the ways to combine the feature extraction capability of deep neural networks with the nonlinear approximation possibilities offered by aggregation.

1.2 Terminology

The success of approximate DP in addressing challenging large-scale applications owes much to an enormously beneficial cross-fertilization of ideas from decision and control, and from artificial intelligence. The boundaries between these fields are now diminished thanks to a deeper understanding of the foundational issues, and the associated methods and core applications. Unfortunately, however, there have been substantial discrepancies of notation and terminology between the artificial intelligence and the optimization/decision/control fields, including the typical use of maximization/value function/reward in the former field and the use of minimization/cost function/cost per stage in the latter field. The notation and terminology used in this paper is standard in DP and optimal control, and in an effort to forestall confusion of readers that are accustomed to either the reinforcement learning or the optimal control terminology, we provide a list of selected terms commonly used in reinforcement learning (for example in the popular book by Sutton and Barto [SuB98],
and its 2018 on-line 2nd edition), and their optimal control counterparts.

(a) Agent = Controller or decision maker.

(b) Action = Control.

(c) Environment = System.

(d) Reward of a stage = (Opposite of) Cost of a stage.

(e) State value = (Opposite of) Cost of a state.

(f) Value (or state-value) function = (Opposite of) Cost function.

(g) Maximizing the value function = Minimizing the cost function.

(h) Action (or state-action) value = Q-factor of a state-control pair.

(i) Planning = Solving a DP problem with a known mathematical model.

(j) Learning = Solving a DP problem in model-free fashion.

(k) Self-learning (or self-play in the context of games) = Solving a DP problem using policy iteration.

(l) Deep reinforcement learning = Approximate DP using value and/or policy approximation with deep neural networks.

(m) Prediction = Policy evaluation.

(n) Generalized policy iteration = Optimistic policy iteration.

(o) Episodic task or episode = Finite-step system trajectory.

(p) Continuing task = Infinite-step system trajectory.

(q) Afterstate = Post-decision state.

2. APPROXIMATE POLICY ITERATION: AN OVERVIEW

Many approximate DP algorithms are based on the principles of PI: the policy evaluation/policy improvement structure of PI is maintained, but the policy evaluation is done approximately, using simulation and some approximation architecture. In the standard form of the method, at each iteration, we compute an approximation $\tilde{J}_\mu(\cdot, r)$ to the cost function $J_\mu$ of the current policy $\mu$, and we generate an “improved” policy.
\[ \hat{\mu} \text{ using } \]
\[ \hat{\mu}(i) \in \arg \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) \left( g(i,u,j) + \alpha \tilde{J}_\mu(j,r) \right), \quad i = 1, \ldots, n. \] (2.1)

Here \( \tilde{J}_\mu \) is a function of some chosen form (the approximation architecture), which depends on the state and on a parameter vector \( r = (r_1, \ldots, r_s) \) of relatively small dimension \( s \).

The theoretical basis for the method was discussed in the neuro-dynamic programming book [BeT96], Prop. 6.2. It was shown there that if the policy evaluation is accurate to within \( \delta \) (in the sup-norm sense), then for an \( \alpha \)-discounted problem, the method, while not convergent, is stable in the sense that it will yield in the limit (after infinitely many policy evaluations) stationary policies that are optimal to within

\[ \frac{2\alpha \delta}{(1-\alpha)^2}, \] (2.2)

where \( \alpha \) is the discount factor. If the generated sequence of policies actually converges to some \( \bar{\mu} \), then \( \bar{\mu} \) is optimal to within

\[ \frac{2\alpha \delta}{1-\alpha} \] (2.3)

(see [BeT96], Section 6.4.1); this is a significantly improved error bound. In general, policy convergence may not be guaranteed, although it is guaranteed for the aggregation methods of this paper. Experimental evidence indicates that these bounds are often conservative, with just a few policy iterations needed before most of the eventual cost improvement is achieved.

### 2.1 Direct and Indirect Approximation

Given a class of functions \( J \) that defines an approximation architecture, there are two general approaches for approximating the cost function \( J_\mu \) of a fixed policy \( \mu \) within \( J \). The most straightforward approach, referred to as direct (or cost fitting), is to find a \( \tilde{J}_\mu \in J \) that matches \( J_\mu \) in some least squares error sense, i.e.,

\[ \tilde{J}_\mu \in \arg \min_{J \in J} \| \tilde{J} - J_\mu \|^2. \] (2.4)

Typically \( \| \cdot \| \) is some weighted Euclidean norm with positive weights \( \xi_i, i = 1, \ldots, n \), while \( J \) consists of a parametrized class of functions \( \tilde{J}(i,r) \) where \( r = (r_1, \ldots, r_s) \in \mathbb{R}^s \) is the parameter vector, i.e.,

\[ J = \{ \tilde{J}(\cdot, r) \mid r \in \mathbb{R}^s \}. \]

† The minimization in the policy improvement phase may also be done approximately through Q-factor approximations or a Monte Carlo tree search, but we will not focus much on this possibility. Our discussion extends straightforwardly to schemes that include approximate policy improvement.
Then the minimization problem in Eq. (2.4) is written as

\[
\min_{r \in \mathbb{R}^s} \sum_{i=1}^{n} \xi_i \left( \tilde{J}(i, r) - J_{\mu}(i) \right)^2,
\]

(2.5)

and can be viewed as an instance of nonlinear regression.

In simulation-based methods, the preceding minimization is usually approximated by a least squares minimization of the form

\[
\min_{r \in \mathbb{R}^s} \sum_{m=1}^{M} \left( \tilde{J}(i_m, r) - \beta_m \right)^2,
\]

(2.6)

where \((i_m, \beta_m), m = 1, \ldots, M\), are a large number of state-cost sample pairs, i.e., for each \(m\), \(i_m\) is a sample state and \(\beta_m\) is equal to \(J_{\mu}(i_m)\) plus some simulation noise. Under mild statistical assumptions on the sample collection process, the sample-based minimization (2.6) is equivalent in the limit to the exact minimization (2.5). Neural network-based approximation, as described in Section 3, is an important example of direct approximation that uses state-cost training pairs.

A common choice is to take \(\mathcal{J}\) to be the subspace \(\{ \Phi r \mid r \in \mathbb{R}^s \}\) that is spanned by the columns of an \(n \times s\) matrix \(\Phi\), which can be viewed as basis functions (see the left side of Fig. 2.1). Then the approximation problem (2.6) becomes the linear least squares problem

\[
\min_{(r_1, \ldots, r_s) \in \mathbb{R}^s} \sum_{m=1}^{M} \left( \sum_{\ell=1}^{s} \phi_{i_m \ell} r_\ell - \beta_m \right)^2,
\]

(2.7)

where \(\phi_{i \ell}\) is the \(i\ell\)th entry of the matrix \(\Phi\) and \(r_\ell\) is the \(\ell\)th component of \(r\). The solution of this problem can be obtained analytically and can be written in closed form (see e.g., [BeT96], Section 3.2.2). Note that the \(i\)th row of \(\Phi\) may be viewed as a feature vector of state \(i\), and \(\Phi r\) may be viewed as a linear feature-based architecture.

In Section 3, we will see that neural network-based policy evaluation combines elements of both a linear and a nonlinear architecture. The nonlinearity is embodied in the features that the neural network constructs through training, but once the features are given, the neural network can be viewed as a linear feature-based architecture.

An often cited weakness of simulation-based direct approximation is excessive simulation noise in the cost samples \(\beta_m\) that are used in the least squares minimization (2.7) or (2.7). This has motivated alternative approaches for policy evaluation that inherently involve less noise. A major approach of this type, referred to as indirect (or equation fitting), is to approximate Bellman’s equation for the policy \(\mu\),

\[
J(i) = \sum_{j=1}^{n} p_{ij}(\mu(i)) \left( g(i, \mu(i), j) + \alpha J(j) \right), \quad i = 1, \ldots, n,
\]

(2.8)

with another equation that is defined on the set \(\mathcal{J}\). The solution of the approximate equation is then used as an approximation of the solution of the original. The most common indirect methods assume a
Direct Method: Projecting the cost vector $J_\mu$

Indirect Method: Solving a projected form of Bellman’s equation

Figure 2.1 Two methods for approximating the cost function $J_\mu$ as a linear combination of basis functions. The approximation architecture is the subspace $\mathcal{J} = \{\Phi_r \mid r \in \mathbb{R}^s\}$, where $\Phi$ is matrix whose columns are the basis functions. In the direct method (see the figure on the left), $J_\mu$ is projected on $\mathcal{J}$. In an example of the indirect method, the approximation is obtained by solving the projected form of Bellman’s equation $\Phi_r = \Pi T_\mu \Phi_r$, where $T_\mu \Phi_r$ is the vector with components

$$(T_\mu \Phi_r)(i) = \sum_{j=1}^n p_{ij}(\mu(i)) \left( g(i, \mu(i), j) + \alpha \Phi_r(j) \right), \quad i = 1, \ldots, n,$$

and $(\Phi_r)(j)$ is the jth component of the vector $\Phi_r$ (see the figure on the right).

linear approximation architecture, i.e., $\mathcal{J}$ is the subspace $\mathcal{J} = \{\Phi_r \mid r \in \mathbb{R}^s\}$, and approximate Bellman’s equation with another equation with fewer variables, the $s$ parameters $r_1, \ldots, r_s$. Two major examples of this approach are projected equation methods and aggregation methods, which we proceed to discuss.

2.2 Indirect Methods Based on Projected Equations

Approximation using projected equations has a long history in numerical computation (e.g., partial differential equations) where it is known as Galerkin approximation [see e.g., [KVZ72], [Fle84], [Saa03], [Kir11]]. The projected equation approach is a special case of the so called Bubnov-Galerkin method, as noted in the papers [Ber11a], [Ber11b], and [YuB10]. In the context of approximate DP it is connected with temporal difference methods, and it is discussed in detail in many sources (see e.g., [BeT96], [BBD10], [Ber12], [Gos15]).

To state the projected equation, let us introduce the transformation $T_\mu$, which is defined by the right-hand side of the Bellman equation (2.8); i.e., for any $J \in \mathbb{R}^n$, $T_\mu J$ is the vector of $\mathbb{R}^n$ with components

$$(T_\mu J)(i) = \sum_{j=1}^n p_{ij}(\mu(i)) \left( g(i, \mu(i), j) + \alpha J(j) \right), \quad i = 1, \ldots, n. \quad (2.9)$$

Note that $T_\mu$ is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$, and in fact in compact vector-matrix notation, it is
written as

\[ T_\mu J = g_\mu + \alpha P_\mu J, \quad J \in \mathbb{R}^n, \]  

(2.10)

where \( P_\mu \) is the transition probability matrix of \( \mu \), and \( g_\mu \) is the expected cost vector of \( \mu \), i.e., the vector with components

\[ \sum_{j=1}^{n} p_{ij}(\mu(i))g(i, \mu(i), j), \quad i = 1, \ldots, n. \]

Moreover the Bellman equation (2.8) is written as the fixed point equation

\[ J = T_\mu J. \]

Let us denote by \( \Pi J \) the projection of a vector \( J \in \mathbb{R}^n \) onto \( J \) with respect to some weighted Euclidean norm, and consider \( \Pi T_\mu \Phi r \), the projection of \( T_\mu \Phi r \) (here \( T_\mu \Phi r \) is viewed as a vector in \( \mathbb{R}^n \), and \( \Pi \) is viewed as an \( n \times n \) matrix multiplying this vector). Then the projected equation takes the form

\[ \Phi r = \Pi T_\mu \Phi r; \]  

(2.11)

see the right-hand side of Fig. 2.1. With this equation we want to find a vector \( \Phi r \) of \( J \), which when transformed by \( T_\mu \) and then projected back onto \( J \), yields itself. This is an overdetermined system of linear equations (\( n \) equations in the \( s \) unknowns \( r_1, \ldots, r_s \)), which in matrix-vector notation is written as

\[ \sum_{\ell=1}^{s} \phi_{i\ell}r_\ell = \sum_{m=1}^{n} \pi_{im} \sum_{j=1}^{n} p_{mj}(\mu(m)) \left( g(m, \mu(m), j) + \alpha \sum_{\ell=1}^{s} \phi_{j\ell}r_\ell \right), \quad i = 1, \ldots, n; \]  

(2.12)

here \( \phi_{i\ell} \) is the \( i\ell \)th component of the matrix \( \Phi \) and \( \pi_{im} \) is the \( im \)th component of the projection matrix \( \Pi \).

The system can be shown to have a unique solution, assuming that the Markov chain corresponding to the policy \( \mu \) has a unique steady-state distribution, and that \( \Phi \) has linear independent columns (see e.g., [Ber12], Section 6.3).

An important extension is to replace the projected equation (2.11) with the equation

\[ \Phi r = \Pi T^{(\lambda)}_\mu \Phi r, \]  

(2.13)

where \( \lambda \) is a scalar with \( 0 \leq \lambda < 1 \), and the transformation \( T^{(\lambda)}_\mu \) is defined by

\[ (T^{(\lambda)}_\mu J)(i) = (1 - \lambda) \sum_{\ell=0}^{\infty} \lambda^\ell (T^{(\lambda)}_\mu J)(i), \quad i = 1, \ldots, n, \quad J \in \mathbb{R}^n, \]

and \( T^{(\lambda)}_\mu J \) is the \( \ell \)-fold composition of \( T_\mu \) applied to the vector \( J \). This approach to the approximate solution of Bellman’s equation is supported by extensive theory and practical experience (see the textbooks noted earlier). In particular, the TD(\( \lambda \)) algorithm, and other related temporal difference methods, such as LSTD(\( \lambda \)) and LSPE(\( \lambda \)), aim to solve by simulation the projected equation (2.13). The choice of \( \lambda \) embodies the
important bias-variance tradeoff: larger values of $\lambda$ lead to better approximation of $J_\mu$, but require a larger number of simulation samples because of increased simulation noise (see the discussion in Section 6.3.6 of [Ber12]). An important insight is that the operator $T_\mu^{(\lambda)}$ is closely related to the proximal operator of convex analysis (with $\lambda$ corresponding to the penalty parameter of the proximal operator), as shown in the author’s paper [Ber16a] (see also the monograph [Ber18], Section 1.2.5).

A major issue in projected equation methods is whether the linear transformation $\Pi T_\mu$ [or $\Pi T_\mu^{(\lambda)}$] is a contraction mapping, in which case Eq. (2.11) [or Eq. (2.13), respectively] has a unique solution, which may be obtained by iterative fixed point algorithms. This depends on the projection norm, and it turns out that there are special norms for which $\Pi T_\mu^{(\lambda)}$ is a contraction (these are related to the steady-state distribution of the system’s Markov chain under $\mu$; see the discussion of [Ber11a] or Section 6.3 of [Ber12]). An important fact is that for any projection norm, $\Pi T_\mu^{(\lambda)}$ is a contraction provided $\lambda$ is sufficiently close to 1. Still the contraction issue regarding $\Pi T_\mu^{(\lambda)}$ is significant and affects materially the implementation of the corresponding approximate PI methods.

Another important concern is that the projection matrix $\Pi$ may have some negative entries [i.e., some of the components $\pi_{im}$ in Eq. (2.12) may be negative], and as a result the linear transformations $\Pi T_\mu$ and $\Pi T_\mu^{(\lambda)}$ may lack the monotonicity property that is essential for the convergence of the corresponding approximate PI method. Indeed the lack of monotonicity (the possibility that we may not have $\Pi T_\mu J \geq \Pi T_\mu J'$ for two vectors $J, J'$ with $J \geq J'$) is the mathematical reason for policy oscillations in PI methods that are based on temporal differences (see [Ber11a], [Ber12]). We refer to the literature for further details and analysis regarding the projected equations (2.11) and (2.13), as our focus will be on aggregation methods, which we discuss next.

### 2.3 Indirect Methods Based on Aggregation

Aggregation is another major indirect approach for policy evaluation. It has a long history in numerical computation. Simple examples of aggregation involve finite-dimensional approximations of infinite dimensional equations, coarse grid approximations of linear systems of equations defined over a dense grid, and other related methods for dimensionality reduction of high-dimensional systems. In the context of DP, the aggregation idea is implemented by replacing the Bellman equation $J = T_\mu J$ [cf. Eq. (2.8)] with a lower-dimensional “aggregate” equation, which is defined on an approximation subspace $\mathcal{J} = \{\Phi r \mid r \in \mathbb{R}^s\}$. The aggregation counterpart of the projected equation $\Phi r = \Pi T_\mu \Phi r$ is

$$\Phi r = \Phi DT_\mu \Phi r,$$

(2.14)

where $\Phi$ and $D$ are some matrices, and $T_\mu$ is the linear transformation given by Eq. (2.9).\footnote{It turns out that under some widely applicable conditions, including the assumptions of Section 4, the projected}
vector-matrix notation for the linear system of \( n \) equations in the \( s \) variables \( r_1, \ldots, r_s \)

\[
\sum_{k=1}^{s} \phi_{ik} r_k = \sum_{k=1}^{s} \phi_{ik} \sum_{m=1}^{n} d_{km} \sum_{j=1}^{n} p_{mj}(\mu(m)) \left( g(m, \mu(m), j) + \alpha \sum_{\ell=1}^{s} \phi_{j\ell} r_\ell \right), \quad i = 1, \ldots, n,
\]

where \( \phi_{il} \) is the \( il \)th component of the matrix \( \Phi \) and \( d_{km} \) is the \( km \)th component of the matrix \( D \).

A key restriction for aggregation methods as applied to DP is that the rows of \( D \) and \( \Phi \) should be probability distributions. These distributions usually have intuitive interpretations in the context of specific aggregation schemes; see [Ber12], Section 6.5 for a discussion. Assuming that \( \Phi \) has linearly independent columns, which is true for the most common types of aggregation schemes, Eq. (2.14) is equivalent to

\[
r = DT_\mu \Phi r,
\]

or

\[
r_k = \sum_{m=1}^{n} d_{km} \sum_{j=1}^{n} p_{mj}(\mu(m)) \left( g(m, \mu(m), j) + \alpha \sum_{\ell=1}^{s} \phi_{j\ell} r_\ell \right), \quad k = 1, \ldots, s.
\]

In most of the important aggregation methods, including the one of Section 4, \( D \) and \( \Phi \) are chosen so that the product \( D\Phi \) is the identity:

\[
D\Phi = I.
\]

Assuming that this is true, the operator \( I - DT_\mu \Phi \) of the aggregation equation (2.15) is obtained by pre-multiplying and post-multiplying the operator \( I - T_\mu \) of the Bellman equation with \( D \) and \( \Phi \), respectively. Mathematically, this can be interpreted as follows:

(a) We replace the variables \( J(j) \) of the Bellman equation \( J = T_\mu J \) with convex combinations of the variables \( r_\ell \) of the system (2.14), using the rows \( (\phi_{j1}, \ldots, \phi_{js}) \) of \( \Phi \):

\[
J(j) \approx \sum_{\ell=1}^{s} \phi_{j\ell} r_\ell.
\]

(b) We form the equations of the aggregate system by taking convex combinations of the components of Bellman equation using the rows of \( D \).

We can now associate the aggregate system of Eq. (2.16) with a discounted DP problem that has \( s \) states, called the aggregate states in what follows. At an abstract level, the aggregate states may be viewed and aggregation equations are closely related. In particular, it can be proved under these conditions that the matrix \( \Phi D \) that appears in the aggregation equation (2.14) is a projection with respect to a suitable weighted Euclidean seminorm (see [YuB12], Section 4, or the book [Ber12]; it a norm projection in the case of hard aggregation). Aside from establishing the relation between the two major indirect approximation methods, projected equation and aggregation, it provides the basis for transferring the rich methodology of temporal differences methods such as TD(\( \lambda \)) to the aggregation context.
as entities associated with the \( s \) rows of \( D \) or the \( s \) columns of \( \Phi \). Since \( T_\mu \) has the form \( T_\mu J = g_\mu + \alpha P_\mu J \) [cf. Eq. (2.10)], the aggregate system (2.16) becomes

\[
r = \hat{g}_\mu + \alpha \hat{P}_\mu r,
\]

where

\[
\hat{g}_\mu = Dg_\mu, \quad \hat{P}_\mu = DP_\mu \Phi.
\]

It is straightforward to verify that \( \hat{P}_\mu \) is a transition probability matrix, since the rows of \( D \) and \( \Phi \) are probability distributions. This means that the aggregation equation (2.17) [or equivalently Eq. (2.16)] represents a policy evaluation/Bellman equation for the discounted problem with transition matrix \( \hat{P}_\mu \) and cost vector \( \hat{g}_\mu \). This problem will be called the *aggregate DP problem* associated with policy \( \mu \) in what follows. The corresponding aggregate state costs are \( r_1, \ldots, r_s \). Some important consequences of this are:

(a) The aggregation equation (2.17)-(2.18) inherits the favorable characteristics of the Bellman equation \( J = T_\mu J \), namely its monotonicity and contraction properties, and its uniqueness of solution.

(b) Exact DP methods, such as PI and Q-learning, may also be used to solve the aggregate DP problem. These methods often have more regular behavior than their counterparts based on projected equations.

(c) Approximate DP methods, such as variants of simulation-based PI, may be used to solve the aggregate DP problem.

The preceding characteristics of the aggregation approach may be turned to significant advantage, and may counterbalance the restriction on the structure of \( D \) and \( \Phi \) (their rows must be probability distributions, as stated earlier).

### 2.4 Implementation Issues

The implementation of approximate PI methods involves several delicate issues, which have been extensively investigated but have not been fully resolved, and are the subject of continuing research. We discuss briefly some of these issues in what follows in this section. We preface this discussion by noting that all of these issues are addressed more easily and effectively within the direct approximation and the aggregation frameworks, than within the temporal difference/projected equation framework, because of the deficiencies relating to the lack of monotonicity and contraction of the operator \( \Pi T_\mu \), which we noted in Section 2.2.

#### The Issue of Exploration

An important generic difficulty with simulation-based PI is that in order to evaluate a policy \( \mu \), we may need to generate cost samples using that policy, but this may bias the simulation by underrepresenting states that
are unlikely to occur under $\mu$. As a result, the cost-to-go estimates of these underrepresented states may be highly inaccurate, causing potentially serious errors in the calculation of the improved control policy $\hat{\mu}$ via the policy improvement equation (2.1).

The situation just described is known as inadequate exploration of the system’s dynamics. It is a particularly acute difficulty when the system is deterministic [i.e., $p_{ij}(u)$ is equal to 1 for a single successor state $j$], or when the randomness embodied in the transition probabilities of the current policy is “relatively small,” since then few states may be reached from a given initial state when the current policy is simulated.

One possibility to guarantee adequate exploration of the state space is to break down the simulation to multiple short trajectories (see [Ber11c], [Ber12], [YuB12]) and to ensure that the initial states employed form a rich and representative subset. This is naturally done within the direct approximation and the aggregation frameworks, but less so in the temporal difference framework, where the theoretical convergence analysis relies on the generation of a single long trajectory.

Another possibility for exploration is to artificially introduce some extra randomization in the simulation of the current policy, by occasionally generating random transitions using some policy other than $\mu$ (this is called an off-policy approach and its implementation has been the subject of considerable discussion; see the books [SuB98], [Ber12]). A Monte Carlo tree search implementation may naturally provide some degree of such randomization, and has worked well in game playing contexts, such as the AlphaZero architecture for playing chess, Go, and other games (Silver et al., [SHS17]). Other related approaches to improve exploration based on generating multiple short trajectories are discussed in Sections 6.4.1 and 6.4.2 of [BeT12].

**Limited Sampling/Optimistic Policy Iteration**

In the approximate PI approach discussed so far, the evaluation of the current policy $\mu$ must be fully carried out. An alternative is optimistic PI, where relatively few simulation samples are processed between successive policy changes and corresponding parameter updates. Optimistic PI with cost function approximation is frequently used in practical applications. However, the associated theoretical convergence properties are not fully understood. In particular, optimistic PI can exhibit fascinating and counterintuitive behavior, including a natural tendency for a phenomenon called chattering, whereby the generated parameter sequence converges, while the generated policy sequence oscillates because the limit of the parameter sequence corresponds to multiple policies. This phenomenon is explained in Section 6.4.2 of [BeT96] (see also Section 6.4.3 of [Ber12]), and can be potentially very damaging, because there is no guarantee that the policies involved in the oscillation are “good” policies, and there is often no way to verify how well they compare to the optimal.

On the other hand optimistic PI may in some cases deal better with the problem of exploration discussed earlier. The reason is that with rapid changes of policy, there may be less tendency to bias the simulation towards particular states that are favored by any single policy.
The Issue of Policy Oscillations

Contrary to exact PI, which converges to an optimal policy in a fairly regular manner, approximate PI may oscillate. By this we mean that after a few iterations, policies tend to repeat in cycles. The associated parameter vectors \( r \) may also tend to oscillate, although it is possible, in optimistic approximate PI methods, that there is convergence in parameter space and oscillation in policy space.

We note that oscillations can be avoided and approximate PI can be shown to converge under special conditions that arise in particular when aggregation is used for policy evaluation. These conditions involve certain monotonicity assumptions [e.g., the nonnegativity of the components \( \pi_{im} \) of the projection matrix in Eq. (2.12)], which are fulfilled in the case of aggregation (see [Ber11a]). However, for temporal difference methods, policy oscillations tend to occur generically, and often for very simple problems, involving few states (a two-state example is given in [Ber11a], and in [Ber12], Section 6.4.3).

Model-Free Implementation

In many problems a mathematical model [the transition probabilities \( p_{ij}(u) \) and the cost vector \( g \)] is unavailable or hard to construct, but instead the system and cost structure can be simulated far more easily. In particular, let us assume that there is a computer program that for any given state \( i \) and control \( u \), simulates sample probabilistic transitions to a successor state \( j \) according to \( p_{ij}(u) \), and generates the transition cost \( g(i, u, j) \).

As noted earlier, the direct and indirect approaches to approximate evaluation of a single policy may be implemented in model-free fashion, simply by generating the needed cost samples by simulation. However, given the result \( \hat{J}_\mu(\cdot, r) \) of the approximate policy evaluation, the policy improvement minimization of Eq. (2.1) requires the transition probabilities \( p_{ij}(u) \), so it is not model-free. To provide a model-free version, suppose that the values of \( \hat{J}_\mu \) and corresponding transition costs can be generated in a model-free fashion when needed. Then we can introduce a parametric family/approximation architecture of Q-factor functions, \( \tilde{Q}_\mu(i, u, w) \), where \( w \) is the parameter vector, and use a regularized least squares fit/regression to approximate the expected value that is minimized in Eq. (2.1). The steps are as follows:

(a) Use the simulator to collect a large number of “representative” sample state-control pairs \((i_m, u_m)\) and successor states \(j_m, m = 1, \ldots, M\), and corresponding sample Q-factors

\[
\beta_m = g(i_m, u_m, j_m) + \alpha \hat{J}_\mu(j_m, r), \quad m = 1, \ldots, M.
\]  

(b) Determine the parameter vector \( \tilde{w} \) by the least-squares minimization

\[
\tilde{w} \in \arg \min_w \sum_{m=1}^{M} (\tilde{Q}_\mu(i_m, u_m, w) - \beta_m)^2
\]
(or a regularized minimization whereby a quadratic regularization term is added to the above quadratic objective).

(c) Use the policy

$$\hat{\mu}(i) \in \arg\min_{u \in U(i)} \tilde{Q}_\mu(i, u, \tilde{w}), \quad i = 1, \ldots, n. \quad (2.21)$$

This policy may be generated on-line when the control constraint set $U(i)$ contains a reasonably small number of elements. Otherwise an approximation in policy space is needed to represent the policy $\hat{\mu}$ using a policy approximation architecture. Such an architecture could be based on a neural network, in which case it is commonly called an “action network” or “actor network” to distinguish from its cost function approximation counterpart, which is called a “value network” or “critic network.”

Note some important points about the preceding approximation procedure:

1. It does not need the transition probabilities $p_{ij}(u)$ to generate the policy $\hat{\mu}$ through the minimization (2.21). The simulator to collect the samples (2.19) suffices.

2. The policy $\hat{\mu}$ obtained through the minimization (2.21) is not the same as the one obtained through the minimization (2.1). There are two reasons for this. One is the approximation error introduced by the Q-factor architecture $\tilde{Q}_\mu$, and the other is the simulation error introduced by the finite-sample regression (2.20). We have to accept these sources of error as the price to pay for the convenience of not requiring a mathematical model.

3. Two approximations are potentially required: One to compute $\tilde{J}_\mu$, which is needed for the samples $\beta_m$ [cf. Eq. (2.19)], and another to compute $\tilde{Q}_\mu$ through the least squares minimization (2.20), and the subsequent policy generation formula (2.21). The approximation methods to obtain $\tilde{J}$ and $\tilde{Q}$ may not be the same and may be unrelated.

An alternative to first computing $\tilde{J}_\mu(i, r)$ and then computing subsequently $\tilde{Q}_\mu(i, u, w)$ via the procedure (2.19)-(2.21) is to introduce a parametric approximation architecture for the policy Q-factor, $\tilde{Q}_\mu(i, u, r)$, where $r$ is the parameter vector. We may then train this Q-factor architecture directly, using state-control-Q-factor samples, and either the direct or the indirect approach. Algorithms for approximating policy cost functions can be adapted to approximating policy Q-factor functions. This adaptation is in principle straightforward, although the implementation may be challenging, as it often the case in approximate PI methods.

3. APPROXIMATE POLICY ITERATION BASED ON NEURAL NETWORKS

In this section we will describe some of the basic ideas of the neural network methodology as it applies to
the approximation of the cost vector $J_\mu$ of a fixed policy $\mu$. Since $\mu$ is fixed throughout this section, we drop the subscript $\mu$ is what follows. A neural network provides an architecture of the form

$$\tilde{J}(i, v, r) = \sum_{\ell=1}^{s} F_\ell(i, v) r_\ell$$

that depends on a parameter vector $v$ and a parameter vector $r = (r_1, \ldots, r_s)$. Here for each state $i$, $\tilde{J}(i, v, r)$ approximates $J_\mu(i)$, while the vector

$$F(i, v) = (F_1(i, v), \ldots, F_s(i, v))$$

may be viewed as a feature vector of the state $i$. Notice the different roles of the two parameter vectors: $v$ parametrizes $F(i, v)$, and $r$ is a vector of weights that combine linearly the components of $F(i, v)$. The idea is to use training to obtain simultaneously both the features and the linear weights.

Consistent with the direct approximation framework of Section 2.1, to train a neural network, we generate a training set that consists of a large number of state-cost pairs $(i_m, \beta_m)$, $m = 1, \ldots, M$, and we find $(v, r)$ that minimizes

$$\sum_{m=1}^{M} \left(\tilde{J}(i_m, v, r) - \beta_m\right)^2.$$  

The training pairs $(i_m, \beta_m)$ are generated by some kind of calculation or simulation, and they may contain noise, i.e., $\beta_m$ is the cost of the policy starting from state $i_m$ plus some error.†

The simplest type of neural network is the single layer perceptron; see Fig. 3.1. Here the state $i$ is encoded as a vector of numerical values $y(i)$ with components $y_1(i), \ldots, y_k(i)$, which is then transformed linearly as

$$Ay(i) + b,$$

where $A$ is an $m \times k$ matrix and $b$ is a vector in $\mathbb{R}^m$. Some of the components of $y(i)$ may be known interesting features of $i$ that can be designed based on problem-specific knowledge or prior training experience. This transformation will be referred to as the linear layer of the neural network. We view the components of $A$ and $b$ as parameters to be determined, and we group them together into the parameter vector $v = (A, b)$.

Each of the $s$ scalar output components of the linear layer,

$$\left(Ay(i) + b\right)_\ell, \quad \ell = 1, \ldots, s,$$

† There are also neural network implementations of the indirect/projected equation approximation approach, which make use of temporal differences, such as for example nonlinear versions of TD($\lambda$). We refer to the textbook literature on the subject, e.g., [SuB98]. In this paper, we will focus on neural network training that is based on the quadratic cost function (3.2).
becomes the input to a nonlinear differentiable function $\sigma$ that maps scalars to scalars. Typically $\sigma$ is differentiable and monotonically increasing. A simple and popular possibility is the rectified linear unit, which is simply the function $\max\{0, \xi\}$, “rectified” to a differentiable function by some form of smoothing operation; for example

$$\sigma(\xi) = \ln(1 + e^\xi).$$

Other functions, used since the early days of neural networks, have the property

$$-\infty < \lim_{\xi \to -\infty} \sigma(\xi) < \lim_{\xi \to \infty} \sigma(\xi) < \infty.$$ 

Such will be referred to as sigmoids, and some common choices are the hyperbolic tangent function

$$\sigma(\xi) = \tanh(\xi) = \frac{e^\xi - e^{-\xi}}{e^\xi + e^{-\xi}},$$

and the logistic function

$$\sigma(\xi) = \frac{1}{1 + e^{-\xi}}.$$ 

In what follows, we will ignore the character of the function $\sigma$ (except for the differentiability requirement), and simply refer to it as a “nonlinear unit” and to the corresponding layer as a “nonlinear layer.”

At the outputs of the nonlinear units, we obtain the scalars

$$F_\ell(i, v) = \sigma((A_y(i) + b)_\ell), \quad \ell = 1, \ldots, s.$$ 

One possible interpretation is to view these scalars as feature components of state $i$, which are linearly combined using weights $r_\ell$, $\ell = 1, \ldots, s$, to produce the final output

$$\sum_{\ell=1}^s F_\ell(i, v)r_\ell = \sum_{\ell=1}^s \sigma((A_y(i) + b)_\ell)r_\ell. \quad (3.3)$$
Note that each value $F_\ell(i, v)$ depends on just the $\ell$th row of $A$ and the $\ell$th component of $b$, not on the entire vector $v$. In some cases this motivates placing some constraints on individual components of $A$ and $b$ to achieve special problem-dependent “handcrafted” effects.

Given a set of state-cost training pairs $(i_m, \beta_m)$, $m = 1, \ldots, M$, the parameters of the neural network $A$, $b$, and $r$ are obtained by solving the training problem (3.2), i.e.,

$$
\min_{A,b,r} \sum_{m=1}^{M} \left( \sum_{\ell=1}^{s} \sigma\left((Ay(i_m) + b)_{\ell}\right) r_{\ell} - \beta_m \right)^2.
$$

(3.4)

The cost function of this problem is generally nonconvex, so there may exist multiple local minima.

It is common to augment the cost function of this problem with a regularization function, such as a quadratic in the parameters $A$, $b$, and $r$. This is customary in least squares problems in order to make the problem easier to solve algorithmically. However, in the context of neural network training, regularization is primarily important for a different reason: it helps to avoid overfitting, which refers to a situation where a neural network model matches the training data very well but does not do as well on new data. This is a well known difficulty in machine learning, which may occur when the number of parameters of the neural network is relatively large (roughly comparable to the size of the training set). We refer to machine learning and neural network textbooks for a discussion of algorithmic questions regarding regularization and other issues that relate to the practical implementation of the training process. In any case, the training problem (3.4) is an unconstrained nonconvex differentiable optimization problem that can in principle be addressed with standard gradient-type methods.

Let us now discuss briefly two issues regarding the neural network formulation and training process just described:

(a) A major question is how to solve the training problem (3.4). The salient characteristic of the cost function of this problem is the sum of a potentially very large number $M$ of component functions. This structure can be exploited with a variant of the gradient method, called incremental,† which computes just the gradient of a single squared error component

$$
\left( \sum_{\ell=1}^{s} \sigma\left((Ay(i_m) + b)_{\ell}\right) r_{\ell} - \beta_m \right)^2
$$

of the sum in Eq. (3.4) at each iteration, and then changes the current iterate in the opposite direction of this gradient using some stepsize; the books [Ber15], [Ber16b] provide extensive accounts, and theoretical analyses including the connection with stochastic gradient methods are given in the book.

† Sometimes the more recent name “stochastic gradient descent” is used in reference to this method. However, once the training set has been generated, possibly by some deterministic process, the method need not have a stochastic character, and it also does not guarantee cost function descent at each iteration.
(b) Another important question is how well we can approximate the cost function of the policy with a neural network architecture, assuming we can choose the number of the nonlinear units \( s \) to be as large as we want. The answer to this question is quite favorable and is provided by the so-called universal approximation theorem. Roughly, the theorem says that assuming that \( i \) is an element of a Euclidean space \( X \) and \( y(i) \equiv i \), a neural network of the form described can approximate arbitrarily closely (in an appropriate mathematical sense), over a closed and bounded subset \( S \subset X \), any piecewise continuous function \( J : S \rightarrow \mathbb{R} \), provided the number \( s \) of nonlinear units is sufficiently large. For proofs of the theorem at different levels of generality, we refer to Cybenko [Cyb89], Funahashi [Fun89], Hornik, Stinchcombe, and White [HSW89], and Leshno et al. [LLP93]. For intuitive explanations we refer to Bishop ([Bis95], pp. 129-130) and Jones [Jon90].

While the universal approximation theorem provides some assurance about the adequacy of the neural network structure, it does not predict how many nonlinear units we may need for “good” performance in a given problem. Unfortunately, this is a difficult question to even pose precisely, let alone to answer adequately. In practice, one is reduced to trying increasingly larger numbers of units until one is convinced that satisfactory performance has been obtained for the task at hand. Experience has shown that in many cases the number of required nonlinear units and corresponding dimension of \( A \) can be very large, adding significantly to the difficulty of solving the training problem. This has motivated various suggestions for modifications of the neural network structure. One possibility is to concatenate multiple single layer perceptrons so that the output of the nonlinear layer of one perceptron becomes the input to the linear layer of the next, as we will now discuss.

**Multilayer and Deep Neural Networks**

An important generalization of the single layer perceptron architecture is deep neural networks, which involve multiple layers of linear and nonlinear functions. The number of layers can be quite large, hence the “deep” characterization. The outputs of each nonlinear layer become the inputs of the next linear layer; see Fig. 3.2. In some cases it may make sense to add as additional inputs some of the components of the state \( i \) or the state encoding \( y(i) \).

The training problem for multilayer networks has the form

\[
\min_{v,r} \sum_{m=1}^{M} \left( \sum_{\ell=1}^{s} F_{\ell}(i, v)r_{\ell} - \beta_m \right)^2,
\]

where \( v \) represents the collection of all the parameters of the linear layers, and \( F_{\ell}(i, v) \) is the \( \ell \)th feature.
component produced at the output of the final nonlinear layer. Various types of incremental gradient methods can also be applied here, specially adapted to the multi-layer structure and they are the methods most commonly used in practice, in combination with techniques for finding good starting points, etc. An important fact is that the gradient with respect to $v$ of each feature component $F_\ell(i, v)$ can be efficiently calculated using a special procedure known as backpropagation, which is just a computationally efficient way to apply the chain rule of differentiation. We refer to the specialized literature for various accounts (see e.g., [Bis95], [BeT96], [HOT06], [Hay08], [LZX17]).

In view of the universal approximation property, the reason for having multiple nonlinear layers is not immediately apparent. A commonly given explanation is that a multilayer network provides a hierarchical sequence of features, where each set of features in the sequence is a function of the preceding set of features in the sequence. In the context of specific applications, this hierarchical structure can be exploited in order to specialize the role of some of the layers and to enhance particular characteristics of the state. Another reason commonly given is that with multiple linear layers, one may consider the possibility of using matrices $A$ with a particular sparsity pattern, or other structure that embodies special linear operations such as convolution. When such structures are used, the training problem often becomes easier, because the number of parameters in the linear layers may be drastically decreased.

Deep neural networks also have another advantage, which is important for our aggregation-related purposes in this paper: the final features obtained as output of the last nonlinear layer tend to be more complex, so their number can be made smaller as the number of nonlinear layers increases. This tends to facilitate the implementation of the feature-based aggregation schemes that we will discuss in what follows.

4. FEATURE-BASED AGGREGATION FRAMEWORK

In this section, we will specialize the general aggregation framework of Section 2.3 by introducing features in the definition of the matrices $D$ and $\Phi$. The starting point is a given feature mapping, i.e., a function $F$
that maps a state $i$ into its feature vector $F(i)$. We assume that $F$ is constructed in some way (including hand-crafted, or neural network-based), but we leave its construction unspecified for the moment.

We will form a lower-dimensional DP approximation of the original problem, and to this end we introduce disjoint subsets $S_1, \ldots, S_q$ of state-feature pairs $(i, F(i))$, which we call aggregate states. The subset of original system states $I_\ell$ that corresponds to $S_\ell$,

$$I_\ell = \{ i \mid (i, F(i)) \in S_\ell \}, \quad \ell = 1, \ldots, q,$$

(4.1)

is called the disaggregation set of $S_\ell$. An alternative and equivalent definition, given $F$, is to start with disjoint subsets of states $I_\ell$, $\ell = 1, \ldots, q$, and define the aggregates states $S_\ell$ by

$$S_\ell = \{ (i, F(i)) \mid i \in I_\ell \}, \quad \ell = 1, \ldots, q.$$

(4.2)

Mathematically, the aggregate states $S_\ell$ are the restrictions of the feature mapping $F$ on the disaggregation sets $I_\ell$; see Fig. 4.1.

To preview our framework, we will aim to construct an aggregate DP problem whose states will be the aggregate states $S_1, \ldots, S_q$, and whose optimal costs, denoted $r_1^*, \ldots, r_q^*$, will be used to construct a function approximation $\tilde{J}$ to the optimal cost function $J^*$ (or in a related context, the cost function of a policy). This approximation will be constant over each disaggregation set; see Fig. 4.1.† Our ultimate objective is that $\tilde{J}$ approximates closely $J^*$, from which it follows as a general guideline that the aggregate states should be selected so that $J^*$ is nearly constant over each of the disaggregation sets $I_1, \ldots, I_q$. This will also be brought out by our subsequent analysis.

† Strictly speaking, in the general version of our scheme, the approximate cost of a state $i$ will depend on $F$ nonlinearly, but may depend on values $F(j)$ of some other states $j \neq i$. However, dependence on just $F(i)$ holds in the most commonly used feature-based aggregation schemes, including the hard aggregation scheme to be introduced later in this section, while in other schemes it can be made to hold after a redefinition of $F$. 

**Figure 4.1** Illustration of aggregate states and a corresponding cost approximation, which is constant over each disaggregation set. Here there are three aggregate states, with disaggregation sets denoted $I_1, I_2, I_3$. 

#25
To formulate an aggregation model that falls within the framework of Section 2.3, we need to specify the matrices $\Phi$ and $D$. We refer to the row of $D$ that corresponds to aggregate state $S_\ell$ as the *disaggregation distribution of* $S_\ell$ and to its elements $d_{\ell 1}, \ldots, d_{\ell n}$ as the *disaggregation probabilities of* $S_\ell$. Similarly, we refer to the row of $\Phi$ that corresponds to state $j$, $\{\phi_{j\ell} \mid \ell = 1, \ldots, q\}$, as the *aggregation distribution of* $j$, and to its elements as the *aggregation probabilities of* $j$. We impose some restrictions on the components of $D$ and $\Phi$, which we describe next.

**Definition of a Feature-Based Aggregation Architecture:**

Given the collection of aggregate states $S_1, \ldots, S_q$ and the corresponding disaggregation sets $I_1, \ldots, I_q$, the aggregation and disaggregation probabilities satisfy the following:

(a) The disaggregation probabilities map each aggregate state onto its disaggregation set. By this we mean that the row of the matrix $D$ that corresponds to an aggregate state $S_\ell$ is a probability distribution $(d_{\ell 1}, \ldots, d_{\ell n})$ over the original system states that assigns zero probabilities to states that are outside the disaggregation set $I_\ell$:

$$d_{\ell i} = 0, \quad \forall \ i \notin I_\ell, \ \ell = 1, \ldots, q. \quad (4.3)$$

(For example, in the absence of special problem-specific considerations, a reasonable and convenient choice would be to assign equal probability to all states in $I_\ell$.)

(b) The aggregation probabilities map each original system state that belongs to a disaggregation set onto the aggregate state of that set. By this we mean that the row $\{\phi_{j\ell} \mid \ell = 1, \ldots, q\}$ of the matrix $\Phi$ that corresponds to an original system state $j$ is specified as follows:

(i) If $j$ belongs to some disaggregation set, say $I_\ell$, then

$$\phi_{j\ell} = 1, \quad (4.4)$$

and $\phi_{j\bar{\ell}} = 0$ for all $\bar{\ell} \neq \ell$.

(ii) If $j$ does not belong to any disaggregation set, the row $\{\phi_{j\ell} \mid \ell = 1, \ldots, q\}$ is an arbitrary probability distribution.

The requirement $d_{\ell i} = 0$ for all $i \notin I_\ell$, cf. Eq. (4.3), leaves a lot of room for choice of the disaggregation probabilities. Simple examples show that the values of these probabilities can affect significantly the quality of aggregation-based approximations. Thus, finding a good or “optimal” set of disaggregation probabilities...
is an interesting issue; the paper by Van Roy [Van06] provides a relevant discussion. Similarly, the choice of
the aggregation probabilities $\phi_{j\ell}$, subject to the constraint (4.4), is important. Generally, problem-specific
knowledge and intuition can be very helpful in designing aggregation schemes in various contexts, but we
will not address this issue further in this paper.

There are several possible methods to choose the aggregate states. Generally, as noted earlier, the
idea will be to form disaggregation sets over which the cost function values $[J^*(i) \text{ or } J_\mu(i) \text{, depending on
the situation}]$ vary as little as possible. We list three general approaches below, and we illustrate these
approaches later with examples:

(a) **Feature-based approach:** Start with a collection of disjoint subsets $F_\ell$, $\ell = 1, \ldots, q$, of the set of all
possible feature values

$$F = \{ F(i) \mid i = 1, \ldots, n \},$$

compute in some way disjoint state subsets $I_1, \ldots, I_q$ such that

$$F(i) \in F_\ell, \quad \forall \ i \in I_\ell, \ \ell = 1, \ldots, q,$$

and obtain the aggregate states

$$S_\ell = \{ (i, F(i)) \mid i \in I_\ell \}, \quad \ell = 1, \ldots, q,$$

with corresponding disaggregation sets $I_1, \ldots, I_q$. This scheme is appropriate for problems where it can
be implemented so that each disaggregation set $I_\ell$ consists of states with similar cost function values.

(b) **State and feature-based approach:** Sample in some way the set of original system states $i$, compute the
corresponding feature vectors $F(i)$, and divide the pairs $(i, F(i))$ thus obtained into subsets $S_1, \ldots, S_q$.
Some problem-specific knowledge may be used to organize the state sampling, with proper consideration
given to issues of sufficient exploration and adequate representation of what is viewed as important
parts of the state space. This scheme is suitable for problems where states with similar feature vectors
have similar cost function values, and is ordinarily the type of scheme that we would use in conjunction
with neural network-constructed features (see Section 5).

(c) **State-based approach:** Start with a collection of disjoint subsets of states $I_1, \ldots, I_q$, and introduce an
artificial feature vector $F(i)$ that is equal to the index $\ell$ for the states $i \in I_\ell$, $\ell = 1, \ldots, q$, and to some
default index, say 0, for the states that do not belong to $\cup^q_\ell I_\ell$. Then use as aggregate states the subsets

$$S_\ell = \{ (i, \ell) \mid i \in I_\ell \}, \quad \ell = 1, \ldots, q,$$

with $I_1, \ldots, I_q$ as the corresponding disaggregation sets. In this scheme, the feature vector plays a
subsidiary role, but the idea of using disaggregation subsets with similar cost function values is still
central, as we will discuss shortly. (The scheme where the aggregate states are identified with subsets $I_1, \ldots, I_q$ of original system states has been called “aggregation with representative features” in [Ber12], Section 6.5, where its connection with feature-based aggregation has been discussed.)

The approaches of forming aggregate states just described cover most of the aggregation schemes that have been used in practice. Two classical examples of the state-based approach are the following:

**Hard Aggregation:**

The starting point here is a partition of the state space that consists of disjoint subsets $I_1, \ldots, I_q$ of states with $I_1 \cup \cdots \cup I_q = \{1, \ldots, n\}$. The feature vector $F(i)$ of a state $i$ identifies the set of the partition that $i$ belongs to:

$$F(i) = \ell, \quad \forall i \in I_\ell, \ell = 1, \ldots, q. \quad (4.5)$$

The aggregate states are the subsets

$$S_\ell = \{(i, \ell) \mid i \in I_\ell\}, \quad \ell = 1, \ldots, q,$$

and their disaggregation sets are the subsets $I_1, \ldots, I_q$. The disaggregation probabilities $d_{i\ell}$ are positive only for states $i \in I_\ell$ [cf. Eq. (4.3)]. The aggregation probabilities are equal to either 0 or 1, according to

$$\phi_{j\ell} = \begin{cases} 1 & \text{if } j \in I_\ell, \\ 0 & \text{otherwise,} \end{cases} \quad j = 1, \ldots, n, \ell = 1, \ldots, q, \quad (4.6)$$

[cf. Eq. (4.4)].

The following aggregation example is typical of a variety of schemes arising in discretization or coarse grid schemes, where a smaller problem is obtained by discarding some of the original system states. The essence of this scheme is to solve a reduced DP problem, obtained by approximating the discarded state costs by interpolation using the nondiscarded state costs.

**Aggregation with Representative States:**

The starting point here is a collection of states $i_1, \ldots, i_q$ that we view as “representative.” The costs of the nonrepresentative states are approximated by interpolation of the costs of the representative states, using the aggregation probabilities. The feature mapping is

$$F(i) = \begin{cases} \ell & \text{if } i = i_\ell, \ell = 1, \ldots, q, \\ 0 & \text{otherwise.} \end{cases} \quad (4.7)$$
The aggregate states are $S_\ell = \{(i_\ell, \ell)\}, \ell = 1, \ldots, q$, the disaggregation sets are $I_\ell = \{i_\ell\}, \ell = 1, \ldots, q$, and the disaggregation probabilities are equal to either 0 or 1, according to

$$d_{\ell i} = \begin{cases} 1 & \text{if } i = i_\ell, \\ 0 & \text{otherwise,} \end{cases} \quad i = 1, \ldots, n, \; \ell = 1, \ldots, q,$$

[cf. Eq. (4.3)]. The aggregation probabilities must satisfy the constraint $\phi_{j \ell} = 1$ if $j = i_\ell, \ell = 1, \ldots, q$ [cf. Eq. (4.4)], and can be arbitrary for states $j \notin \{i_1, \ldots, i_q\}$.

4.1 The Aggregate Problem

Given a feature-based aggregation framework (i.e., the aggregate states $S_1, \ldots, S_q$, the corresponding disaggregation sets $I_1, \ldots, I_q$, and the aggregation and disaggregation distributions), we can consider an aggregate DP problem that involves transitions between aggregate states. In particular, the transition probabilities $p_{ij}(u)$, and the disaggregation and aggregation probabilities specify a controlled dynamic system involving both the original system states and the aggregate states (cf. Fig. 4.2).†

(i) From aggregate state $S_\ell$, we generate a transition to original system state $i$ according to $d_{\ell i}$ (note that

† We will consider the aggregate problem for the case where there are multiple possible controls at each state. However, it is also possible to consider the aggregate problem for the purpose of finding an approximation to the cost function $J_\mu$ of a given policy $\mu$; this is the special case where the control constraint set $U(i)$ consists of the single control $\mu(i)$ for every state $i$.}
\( i \) must belong to the disaggregation set \( I_\ell \), because of the requirement that \( d_{\ell i} > 0 \) only if \( i \in I_\ell \).

(ii) From original system state \( i \), we generate a transition to original system state \( j \) according to \( p_{ij}(u) \), with cost \( g(i, u, j) \).

(iii) From original system state \( j \), we generate a transition to aggregate state \( S_\ell \) according to \( \phi_{j\ell} \) [note here the requirement that \( \phi_{j\ell} = 1 \) if \( j \in I_\ell \); cf. Eq. (4.4)].

This is a DP problem with an enlarged state space that consists of two copies of the original state space \( \{1, \ldots, n\} \) plus the \( q \) aggregate states. We introduce the corresponding optimal vectors \( \tilde{J}_0 \), \( \tilde{J}_1 \), and \( r^* = \{r^*_1, \ldots, r^*_q\} \) where:

- \( r^*_\ell \) is the optimal cost-to-go from aggregate state \( S_\ell \).
- \( \tilde{J}_0(i) \) is the optimal cost-to-go from original system state \( i \) that has just been generated from an aggregate state (left side of Fig. 4.3).
- \( \tilde{J}_1(j) \) is the optimal cost-to-go from original system state \( j \) that has just been generated from an original system state (right side of Fig. 4.3).

Note that because of the intermediate transitions to aggregate states, \( \tilde{J}_0 \) and \( \tilde{J}_1 \) are different.

These three vectors satisfy the following three Bellman’s equations:

\[
\begin{align*}
    r^*_\ell &= \sum_{i=1}^{n} d_{\ell i} \tilde{J}_0(i), \quad \ell = 1, \ldots, q, \quad (4.8) \\
    \tilde{J}_0(i) &= \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) (g(i, u, j) + \alpha \tilde{J}_1(j)), \quad i = 1, \ldots, n, \quad (4.9) \\
    \tilde{J}_1(j) &= \sum_{\ell=1}^{q} \phi_{j\ell} r^*_\ell, \quad j = 1, \ldots, n. \quad (4.10)
\end{align*}
\]

By combining these equations, we see that \( r^* \) satisfies

\[
    r^* = H r^*, \quad (4.11)
\]

where \( H \) is the mapping the vector \( r \) to the vector \( Hr \) with components defined by

\[
    (Hr)(\ell) = \sum_{i=1}^{n} d_{\ell i} \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) \left( g(i, u, j) + \alpha \sum_{\ell=1}^{q} \phi_{j\ell} r^*_\ell \right), \quad \ell = 1, \ldots, q. \quad (4.12)
\]

It can be seen that \( H \) is a contraction mapping with respect to the sup-norm and thus has \( r^* \) as its unique fixed point. This follows from standard contraction arguments and the fact that \( d_{\ell i} \), \( p_{ij}(u) \), and \( \phi_{j\ell} \) are probabilities. Note the nature of \( r^*_\ell \): it is the optimal cost of the aggregate state \( S_\ell \), which is the restriction of the feature mapping \( F \) on the disaggregation set \( I_\ell \), so that \( \tilde{J}_1 \) depends on the state through \( F \).
Solution of the Aggregate Problem

While the aggregate problem involves more states than the original DP problem, it is in fact easier in some important ways. The reason is that it can be solved with algorithms that execute over the smaller space of aggregate states. In particular, exact and approximate simulation-based algorithms, can be used to find the lower-dimensional vector \( r^* \) without computing the higher-dimensional vectors \( \tilde{J}_0 \) and \( \tilde{J}_1 \). Section 6.5.2 of [Ber12] describes the application of the powerful LSTD and LSPE methods for policy evaluation to find the vector \( r_\mu \) of cost values of the aggregate states that correspond to a given policy \( \mu \). These methods can be used to implement a simulation-based PI algorithm for finding \( r^* \), exactly in a finite number of iterations or approximately in fewer iterations. We also refer to the same reference and other literature regarding the adaptation of Q-learning to solve directly the aggregate problem, without going through PI. The simulator used for these methods is based on Figs. 4.2 and 4.3: transitions to and from the aggregate states are generated using the aggregation and disaggregation probabilities, respectively, while transitions \((i, j)\) between original system states are generated using a simulator of the original system (which is assumed to be available).

Once \( r^* \) is found, the optimal-cost-to-go of the original problem may be approximated by the vector \( \tilde{J}_1 \) of Eq. (4.10), and a suboptimal policy may be found through the minimization (4.9) that defines \( \tilde{J}_0 \). Note that \( \tilde{J}_1 \) is a “piecewise linear” cost approximation of \( J^* \): it is constant over each of the disaggregation sets \( I_\ell, \ell = 1, \ldots, q \) [and equal to the optimal cost \( r_\ell^* \) of the aggregate state \( S_\ell \); cf. Eqs. (4.4) and (4.10)], and it is interpolated/linear outside the disaggregation sets [cf. Eq. (4.10)]. In the case where \( \bigcup_{\ell=1}^q I_\ell = \{1, \ldots, n\} \) (e.g., in hard aggregation), the disaggregation sets \( I_\ell \) form a partition of the original system state space, and \( \tilde{J}_1 \) is piecewise constant. Figure 4.4 illustrates a simple example of approximate cost function \( \tilde{J}_1 \).

Finally, let us note that for the purposes of using feature-based aggregation to improve a given policy \( \mu \), it is not essential to solve the aggregate problem to completion. Instead, we may perform one or just a few PIs and adopt the final policy obtained as a new “improved” policy. The quality of such a policy depends on how well the aggregate problem approximates the original DP problem. While it is not easy to quantify
Figure 4.4. Schematic illustration of the approximate cost function $\tilde{J}_1$. Here the original states are the integers between 1 and 50. In this figure there are three aggregate states numbered 1, 2, 3. The corresponding disaggregation sets are $I_1 = \{1, \ldots, 10\}$, $I_2 = \{20, \ldots, 30\}$, $I_3 = \{40, \ldots, 50\}$ are shown in the figure. The values of the approximate cost function $\tilde{J}_1(i)$ are constant within each disaggregation set $I_\ell$, $\ell = 1, 2, 3$, and are obtained by linear interpolation for states $i$ that do not belong to any one of the sets $I_\ell$. If the sets $I_\ell$, $\ell = 1, 2, 3$, include all the states 1, \ldots, 50, we have a case of hard aggregation. If each of the sets $I_\ell$, $\ell = 1, 2, 3$, consist of a single state, we have a case of aggregation with representative states.

the relevant approximation, generally a small error can be achieved if:

(a) The feature mapping $F$ “conforms” to the optimal cost function $J^*$ in the sense that $F$ varies little in regions of the state space where $J^*$ also varies little.

(b) The aggregate states are selected so that $F$ varies little over each of the disaggregation sets $I_1, \ldots, I_q$.

This is intuitive and is supported by the subsequent discussion and analysis.

Error Bounds

Intuitively, if the disaggregation sets nearly cover the entire state space (in the sense that $\bigcup_{\ell=1,\ldots,q} I_\ell$ contains “most” of the states 1, \ldots, $n$) and $J^*$ is nearly constant over each disaggregation set, then $\tilde{J}_0$ and $\tilde{J}_1$ should be close to $J^*$. In particular, in the case of hard aggregation, we have the following error bound, due to Tsitsiklis and VanRoy [TsV96]. We adapt their proof to the notation and terminology of this paper.

Proposition 4.1: In the case of hard aggregation, where $\bigcup_{\ell=1,\ldots,q} I_\ell = \{1, \ldots, n\}$, and Eqs. (4.5), (4.6) hold, we have

$$|J^*(i) - r^*_\ell| \leq \frac{\epsilon}{1 - \alpha}, \quad \forall \ i \text{ such that } i \in I_\ell, \ \ell = 1, \ldots, q,$$

(4.13)

where

$$\epsilon = \max_{\ell=1,\ldots,q} \max_{i,j \in I_\ell} |J^*(i) - J^*(j)|.$$  

(4.14)
Proof: Consider the mapping $H$ defined by Eq. (4.12), and consider the vector $\mathbf{r}$ with components defined by

$$\mathbf{r}_\ell = \min_{i \in I_\ell} J^*(i) + \frac{\epsilon}{1 - \alpha}, \quad \ell \in 1, \ldots, q.$$ 

Denoting by $\ell(j)$ the index of the disaggregation set to which $j$ belongs, i.e., $j \in I_{\ell(j)}$, we have for all $\ell$,

$$(H\mathbf{r})(\ell) = \sum_{i=1}^{n} d_{\ell i} \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) \left( g(i, u, j) + \alpha \mathbf{r}_{\ell(j)} \right)$$

$$\leq \sum_{i=1}^{n} d_{\ell i} \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) \left( g(i, u, j) + \alpha J^*(j) + \frac{\alpha \epsilon}{1 - \alpha} \right)$$

$$= \sum_{i=1}^{n} d_{\ell i} \left( J^*(i) + \frac{\epsilon}{1 - \alpha} \right)$$

$$\leq \min_{i \in I_\ell} \left( J^*(i) + \epsilon \right) + \frac{\epsilon}{1 - \alpha}$$

$$= \min_{i \in I_\ell} J^*(i) + \frac{\epsilon}{1 - \alpha}$$

$$= \mathbf{r}_\ell,$$

where for the second equality we used the Bellman equation for the original system, which is satisfied by $J^*$, and for the second inequality we used Eq. (4.14). Thus we have $H\mathbf{r} \leq \mathbf{r}$, from which it follows that $r^* \leq \mathbf{r}$ (since $r^* = \lim_{k \to \infty} H^k \mathbf{r}$ and $H$ is monotone). This proves one side of the desired error bound. The other side follows similarly. Q.E.D.

The scalar $\epsilon$ of Eq. (4.14) is the maximum variation of optimal cost within the sets of the partition of the hard aggregation scheme. Thus the meaning of the preceding proposition is that if the optimal cost function $J^*$ varies by at most $\epsilon$ within each set of the partition, the hard aggregation scheme yields a piecewise constant approximation to the optimal cost function that is within $\epsilon/(1 - \alpha)$ of the optimal. We know that for every approximation $\tilde{J}$ of $J^*$ that is constant within each disaggregation set, the error

$$\max_{i=1, \ldots, n} |J^*(i) - \tilde{J}(i)|$$

is at least equal to $\epsilon/2$. Based on the bound (4.13), the actual value of this error for the case where $\tilde{J}$ is obtained by hard aggregation involves an additional multiplicative factor that is at most equal to $2/(1 - \alpha)$, and depends on the disaggregation probabilities. In practice the bound (4.13) is typically conservative, and no examples are known where it is tight. Moreover, even for hard aggregation, the manner in which the error $J^* - \tilde{J}_1$ depends on the disaggregation distributions is unclear and is an interesting research question.

The following proposition extends the result of the preceding proposition to the case where the aggregation probabilities are all either 0 or 1, in which case the cost function $\tilde{J}_1$ obtained by aggregation is a piecewise constant function, but the disaggregation sets need not form a partition of the state space. Examples of this type of scheme include cases where the aggregation probabilities are generated by a “nearest
neighbor” scheme, and the cost $\tilde{J}_1(j)$ of a state $j \notin \cup_{\ell=1}^q I_{\ell}$ is taken to be equal to the cost of the “nearest” state within $\cup_{\ell=1}^q I_{\ell}$.

**Proposition 4.2:** Assume that each aggregation probability $\phi_{j\ell}$, $j = 1, \ldots, n$, $\ell = 1, \ldots, q$, is equal to either 0 or 1, and consider the sets

$$\hat{I}_\ell = \{ j \mid \phi_{j\ell} = 1 \}, \quad \ell = 1, \ldots, q.$$  

Then we have

$$|J^*(i) - r^*_\ell| \leq \frac{\epsilon}{1 - \alpha}, \quad \forall i \in \hat{I}_\ell, \; \ell = 1, \ldots, q,$$

where

$$\epsilon = \max_{\ell=1, \ldots, q} \max_{i,j \in I_{\ell}} |J^*(i) - J^*(j)|.$$

**Proof:** We first note that by the definition of a feature-based aggregation scheme, we have $I_{\ell} \subset \hat{I}_\ell$ for all $\ell = 1, \ldots, q$, while the sets $\hat{I}_\ell$, $\ell = 1, \ldots, q$, form a partition of the original state space, in view of our assumption on the aggregation probabilities. Let us replace the feature vector $F$ with another feature vector of the form

$$\hat{F}(i) = \ell, \quad \forall i \in \hat{I}_\ell, \; \ell = 1, \ldots, q.$$  

Since the aggregation probabilities are all either 0 or 1, the resulting aggregation scheme with $I_{\ell}$ replaced by $\hat{I}_\ell$, and with the aggregation and disaggregation probabilities remaining unchanged, is a hard aggregation scheme. When the result of Prop. 4.1 is applied to this hard aggregation scheme, the result of the present proposition follows.  

Q.E.D.

The preceding propositions suggest the principal guideline for a feature-based aggregation scheme. It should be designed so that states that belong to the same disaggregation set have nearly equal optimal costs. The next section elaborates on schemes that are based on this idea.

### 4.2 Feature Formation by Using Scoring Functions

Clearly the feature mapping $F$ and the method to obtain aggregate states is critical for the success of feature-based aggregation. In the subsequent Section 5 we will discuss how deep neural network architectures can be used for this purpose. In what follows in this section we consider some simple forms of feature mappings.
that can be used when we already have a reasonable estimate of the optimal cost function $J^*$ or the cost function $J_\mu$ of some policy $\mu$, which we can use to group together states with similar estimated optimal cost. Then the aggregation approach can provide an improved piecewise constant or piecewise linear cost approximation. We provide some simple illustrative examples of this approach in Section 4.3.

In particular, suppose that we have obtained in some way a real-valued scoring function $V(i)$ of the state $i$, which serves as an index of undesirability of state $i$ as a starting state (smaller values of $V$ are assigned to more desirable states, consistent with the view of $V$ as some form of “cost” function). For example, $V$ may be an approximation of the cost function of some “good” (e.g., near-optimal) policy. In particular, $V$ may be obtained with any one of a variety of techniques. One possibility is problem approximation (replacing the original problem with a simpler problem that can be solved analytically or computationally; see [Ber17], Section 6.2). Another possibility is to obtain $V$ by training a neural network or other architecture using samples of state-cost pairs obtained by using a software or human expert, and some supervised learning technique, such as for example Tesauro’s comparison learning scheme [Tes89], [Tes01]. Still another possibility is to compute $V$ using some form of policy evaluation algorithm like TD($\lambda$); the role of the algorithm will be to provide a feature mapping that is used for feature-based aggregation.

Given the scoring function $V$, we will construct a feature mapping that groups together states $i$ with roughly equal scores $V(i)$. In particular, we let $R_1, \ldots, R_q$ be $q$ disjoint intervals that form a partition of the set of possible values of $V$ [i.e., are such that for any state $i$, there is a unique interval $R_\ell$ such that $V(i) \in R_\ell$]. We define a feature vector $F(i)$ of the state $i$ according to

$$F(i) = \ell, \quad \forall \ i \text{ such that } V(i) \in R_\ell.$$

(4.15)

This feature vector in turn defines a partition of the state space into the sets

$$I_\ell = \{i \mid F(i) = \ell\} = \{i \mid V(i) \in R_\ell\}, \quad \ell = 1, \ldots, q.$$

Assuming that all the sets $I_\ell$ are nonempty, we thus obtain a hard aggregation scheme, with aggregation probabilities defined by Eq. (4.6).

In practice, the intervals $R_\ell$ and the disaggregation probabilities may be obtained by forming a fairly large sample set of states $\{i_m \mid m = 1, \ldots, M\}$, by computing their corresponding scores

$$\{V(i_m) \mid m = 1, \ldots, M\},$$

and by suitably dividing the range of these scores to form the intervals $R_\ell$. This is a form of “discretization” of the original state space based on the score values of the states. Simultaneously we obtain subsets of states $\hat{I}_\ell \subset I_\ell$ to which we can assign positive disaggregation probabilities. Figure 4.5 illustrates this idea for the case where each subset $\hat{I}_\ell$ consists of a single (representative) state. As the figure indicates, the role of
the scoring function is to assist in forming a set of states that is small (to keep the aggregate DP problem computations manageable) but representative (to provide sufficient detail in the approximation of $J^*$, i.e., be dense in the parts of the state space where $J^*$ varies a lot, and sparse in other parts).

We have the following proposition, which illustrates the important role of the quantization error, defined as

$$
\delta = \max_{\ell=1,\ldots,q} \max_{i,j \in I_\ell} |V(i) - V(j)|. \tag{4.16}
$$

It represents the maximum error that can be incurred by approximating $V$ within each set $I_\ell$ with a single value from its range within the subset.

\begin{prop}
Consider the hard aggregation scheme defined by a scoring function $V$ as described above. Assume that the variations of $J^*$ and $V$ over the sets $I_1, \ldots, I_q$ are within a factor $\beta \geq 0$ of each other, i.e., that

$$
|J^*(i) - J^*(j)| \leq \beta |V(i) - V(j)|, \quad \forall i, j \in I_\ell, \ell = 1, \ldots, q.
$$

(a) We have

$$
|J^*(i) - r^*_\ell| \leq \frac{\beta \delta}{1 - \alpha}, \quad \forall i \in I_\ell, \ell = 1, \ldots, q,
$$

where $\delta$ is the quantization error of Eq. (4.16).
\end{prop}

Figure 4.5. Schematic illustration of aggregation based on sampling states and using a scoring function $V$ to form a representative set $i_1, \ldots, i_q$. A piecewise linear approximation of $J^*$ is obtained using the corresponding aggregate costs $r^*_1, \ldots, r^*_q$ and the aggregation probabilities.
(b) Assume that there is no quantization error, i.e., $V$ and $J^*$ are constant within each interval $I_\ell$.

Then the aggregation scheme yields the optimal cost function $J^*$ exactly, i.e.,

$$J^*(i) = r^*_\ell, \quad \forall \ i \in I_\ell, \ \ell = 1, \ldots, q.$$  

Proof: (a) Since we are dealing with a hard aggregation scheme, the result of Prop. 4.1 applies. By our assumptions, the maximum variation of $J^*$ over the disaggregation sets $I_\ell$ is bounded by $\epsilon = \beta\delta$, and the result of part (a) follows from Prop. 4.1.

(b) This is a special case of part (a) with $\delta = \epsilon = 0$. Q.E.D.

Examples of scoring functions that may be useful in various settings are cost functions of nearly optimal policies, or approximations to such cost functions, provided for example by a neural network or other approximation schemes. Another example, arising in the adaptive aggregation scheme proposed by Bertsekas and Castanon [BeC89], is to use as $V(i)$ the residual vector $(TJ)(i) - J(i)$, where $J$ is some approximation to the optimal cost function $J^*$, or the residual vector $(T_\mu J)(i) - J(i)$, where $J$ is some approximation to the cost function of a policy $\mu$. Note that it is not essential that $V$ approximates well $J^*$ or $J_\mu$. What is important is that states with similar values of $J^*$ or $J_\mu$ also have similar values of $V$, so they are grouped together into aggregate states.

**Variants of Scoring Function Schemes**

The approach of forming features using a single scoring function can be extended to the case where we have a vector of scoring functions $V(i) = (V_1(i), \ldots, V_s(i))$. Then we can partition the set of possible values of $V(i)$ into $q$ disjoint subsets $R_1, \ldots, R_q$, define a feature vector $F(i)$ according to

$$F(i) = \ell, \quad \forall \ i \text{ such that } V(i) \in R_\ell, \ \ell = 1, \ldots, q,$$

and proceed as in the case of a scalar scoring function, i.e., obtain a hard aggregation scheme with disaggregation sets given by

$$I_\ell = \{i \mid F(i) = \ell\} = \{i \mid V(i) \in R_\ell\}, \quad \ell = 1, \ldots, q.$$

One possibility to obtain multiple scoring functions is to start with a single fairly simple scoring function, obtain aggregate states as described above, solve the corresponding aggregate problem, and use the optimal cost function of that problem as an additional scoring function. This is reminiscent of feature iteration, an idea that has been suggested in several approximate DP works.
Let us finally mention another scheme that is based on a single scoring function $V$, which is defined separately on each one of a collection of disjoint subsets $C_1, \ldots, C_m$ that form a partition of the state space. We define a feature vector $F(i)$ that depends not only on the value of $V(i)$ but also on the membership of $i$ in the subsets of the partition. In particular, for each $\theta = 1, \ldots, m$, let $R_{1\theta}, \ldots, R_{q\theta}$ be $q$ disjoint intervals that form a partition of the set of possible values of $V$ over the set $C_\theta$. We then define

$$F(i) = (\ell, \theta), \quad \forall i \in C_\theta \text{ such that } V(i) \in R_{\ell\theta}. \quad (4.18)$$

This feature vector in turn defines a partition of the state space into the $qm$ sets

$$I_{\ell\theta} = \{i \mid F(i) = (\ell, \theta)\} = \{i \in C_\theta \mid V(i) \in R_{\ell\theta}\}, \quad \ell = 1, \ldots, q, \theta = 1, \ldots, m,$$

which represent the disaggregation sets of the resulting hard aggregation scheme. In this scheme the aggregate states depend not only on the range of values of $V$ but also on the subset $C_\theta$ of the partition.

### 4.3 Illustrative Examples

We will now consider two simple illustrative examples, which were presented in the author’s paper [Ber95] as instances of poor performance of TD($\lambda$) (see also the book [BeT96], Section 6.3.2). In these examples the cost function of a policy will be approximated using feature-based aggregation and a scoring function. The approximate cost function thus obtained will be compared with the results of the TD(1) and TD(0) algorithms.

Both examples belong to the class of stochastic shortest path (SSP for short) problems, where there is no discounting and in addition to the states $1, \ldots, n$, there is an additional cost-free and absorbing termination state, denoted 0 (the text references given earlier discuss in detail such problems). Our aggregation methodology of this section extends straightforwardly to SSP problems. The principal change needed is that the termination state 0 must form an aggregate state by itself with corresponding disaggregation set $\{0\}$. As before there are also other aggregate states $S_1, \ldots, S_q$ whose disaggregation sets $I_1, \ldots, I_q$ are subsets of $\{1, \ldots, n\}$. With this special handling of the termination state 0, the aggregate problem becomes a standard SSP problem whose termination state is 0. The Bellman equation of the aggregate problem for the aggregate state costs is given by

$$r_\ell = \sum_{i=1}^{n} d_{i\ell} \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) \left( g(i, u, j) + \sum_{\bar{\ell}=1}^{q} \phi_{j\bar{\ell}} r_{\bar{\ell}} \right), \quad \ell = 1, \ldots, q, \quad (4.19)$$

[cf. Eq. (4.12)]. It has a unique solution under some well-known conditions that date to the paper by Bertsekas and Tsitsiklis [BeT91]. In particular, these conditions are satisfied when the termination state 0 is reached with probability 1 from every state and under all stationary policies (i.e., all policies are proper
in the terminology of [BeT91], [BeT96], [Ber12], which also introduce some less stringent conditions). This is true for both the original and the aggregate problem in the examples to be presented.

Our examples involve a problem with a single policy \( \mu \) where the corresponding Markov chain is deterministic with \( n \) states plus a termination state 0. Under \( \mu \), when at state \( i = 1, \ldots, n \), we move to state \( i - 1 \) at a cost \( g_i \). Thus starting at state \( i \) we traverse each of the states \( i - 1, \ldots, 1 \) and terminate at state 0 at costs \( g_i, \ldots, g_1 \), respectively, while accumulating the total cost

\[
J_\mu(i) = g_i + \cdots + g_1, \quad i = 1, \ldots, n,
\]

with \( J_\mu(0) = 0 \). We consider a linear approximation to this cost function, which we denote by \( V \):

\[
V(i) = ri, \quad i = 1, \ldots, n,
\]

where \( r \) is a scalar parameter. This parameter may be obtained by using any of the simulation-based methods that are available for training linear architectures, including TD(\( \lambda \)). In the subsequent discussion we will assume that TD(\( \lambda \)) is applied in an idealized form where the simulation samples contain no noise.

The TD(1) algorithm is based on minimizing the sum of the squares of the differences between \( J_\mu \) and \( V \) over all states, yielding the approximation

\[
\hat{V}_1(i) = \hat{r}_1 i, \quad i = 0, 1, \ldots, n,
\]

where

\[
\hat{r}_1 \in \arg \min_{r \in \mathbb{R}} \sum_{i=1}^{n} (J_\mu(i) - r i)^2.
\] (4.20)

Here, consistent with our idealized setting of noise-free simulation, we assume that \( J_\mu(i) \) is computed exactly for all \( i \). The TD(0) algorithm is based on minimizing the sum of the squares of the errors in satisfying the Bellman equation \( V(i) = g_i + V(i - 1) \) (or temporal differences) over all states, yielding the approximation

\[
\hat{V}_0(i) = \hat{r}_0 i, \quad i = 0, 1, \ldots, n,
\]

where

\[
\hat{r}_0 \in \arg \min_{r \in \mathbb{R}} \sum_{i=1}^{n} (g_i + r(i - 1) - ri)^2.
\] (4.21)

Again, we assume that the temporal differences \( (g_i + r(i - 1) - ri) \) are computed exactly for all \( i \).

The straightforward solution of the minimization problems in Eqs. (4.20) and (4.21) yields

\[
\hat{r}_1 = \frac{n(g_1 + \cdots + g_n) + (n - 1)(g_1 + \cdots + g_{n-1}) + \cdots + g_1}{n^2 + (n - 1)^2 + \cdots + 1},
\]

and

\[
\hat{r}_0 = \frac{ng_n + (n - 1)g_{n-1} + \cdots + g_1}{n + (n - 1) + \cdots + 1}.
\]
Consider now two different choices of the one stage costs $g_i$:

(a) $g_1 = 1$ and $g_i = 0$ for all $i \neq 1$.

(b) $g_n = -(n - 1)$ and $g_i = 1$ for all $i \neq n$.

Figures 4.6 and 4.7 provide plots of $J_\mu(i)$, and the approximations $\hat{V}_1(i)$ and $\hat{V}_0(i)$ for these two cases (these plots come from [Ber95] where the number of states used was $n = 50$).

We will now consider a hard aggregation scheme based on using $\hat{V}_1$ and $\hat{V}_0$ as scoring functions. The aggregate states of such a scheme in effect consist of disaggregation subsets $I_1, \ldots, I_q$ with $\bigcup_{\ell=1}^q I_\ell = \{1, \ldots, n \}$ plus the subset $\{0\}$ that serves as the termination state of the aggregate problem. With either $\hat{V}_1$ or $\hat{V}_0$ as the scoring function, the subsets $I_1, \ldots, I_q$ consist of contiguous states. In order to guarantee that the termination state is eventually reached in the aggregate problem, we assume that the disaggregation probability of the smallest state within each of the subsets $I_1, \ldots, I_q$ is strictly positive; this is a very mild restriction, which is naturally satisfied in typical schemes that assign equal probability to all the states in a disaggregation set.

Consider first case (a) (cf. Fig. 4.6). Then, because the policy cost function $J_\mu$ is constant within each of the subsets $I_1, \ldots, I_q$, the scalar $\epsilon$ in Prop. 4.1 is equal to 0, implying that the hard aggregation scheme yields the optimal cost function, i.e., $r^*_i = J_\mu(i)$ for all $i \in I_\ell$. To summarize, in case (a) the TD(0) approach yields a very poor linear cost function approximation, the TD(1) approach yields a poor linear cost function approximation, but the aggregation scheme yields exactly the nonlinear policy cost function $J_\mu$. 

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.6.pdf}
\caption{Form of $J_\mu(i)$ and the linear approximations $\hat{V}_1(i)$ and $\hat{V}_0(i)$ for case (a): $g_1 = 1$ and $g_i = 0$ for all $i = 2, \ldots, n$.}
\end{figure}
Figure 4.7. Form of $J_\mu(i)$ and the linear approximations $\hat{V}_1(i)$ and $\hat{V}_0(i)$ for case (b): $g_n = -(1-n)$ and $g_i = 1$ for all $i = 1, \ldots, n-1$.

Consider next case (b) (cf. Fig. 4.7). Then, the hard aggregation scheme yields a piecewise constant approximation to the optimal cost function. The quality of the approximation is degraded by quantization effects. In particular, as the variations of $J_\mu$, and $\hat{V}_1$ or $\hat{V}_0$ increase over the disaggregation sets $I_1, \ldots, I_q$, the quality of the approximation deteriorates, as predicted by Prop. 4.3. Similarly, as the number of states in the disaggregation sets $I_1, \ldots, I_q$ is reduced, the quality of the approximation improves, as illustrated in Fig. 4.8. In the extreme case where there is only one state in each of the disaggregation sets, the aggregation scheme yields exactly $J_\mu$.

To summarize, in case (b) the TD(0) approach yields a very poor linear cost function approximation, the TD(1) approach yields a reasonably good linear cost function approximation, while the aggregation scheme yields a piecewise constant approximation whose quality depends on the coarseness of the quantization that is implicit in the selection of the number $q$ of disaggregation subsets. The example of case (b) also illustrates how the quality of the scoring function affects the quality of the approximation provided by the aggregation scheme. Here both $\hat{V}_1$ and $\hat{V}_0$ work well, despite their very different form, because states with similar values of $J_\mu$ also have similar values of $\hat{V}_1$ as well as $\hat{V}_0$ (cf. Prop. 4.3).
4.4 Multistep Aggregation

The aggregation methodology discussed so far is based on the Markov chain of Fig. 4.2, which returns to an aggregate state after a single transition of the original chain. We may obtain alternative aggregation frameworks by considering a different Markov chain. One possibility, suggested in [Ber11a] and illustrated in Fig. 4.9, is specified by disaggregation and aggregation probabilities as before, but involves \( k > 1 \) transitions between original system states in between transitions from and to aggregate states.

The aggregate DP problem for this scheme involves \( k + 1 \) copies of the original state space, in addition to the aggregate states. We accordingly introduce vectors \( \tilde{J}_0, \tilde{J}_1, \ldots, \tilde{J}_k \), and \( r^* = \{ r^*_1, \ldots, r^*_q \} \) where:

- \( r^*_\ell \) is the optimal cost-to-go from aggregate state \( S_\ell \).
- \( \tilde{J}_0(i) \) is the optimal cost-to-go from original system state \( i \) that has just been generated from an aggregate state (left side of Fig. 4.9).
- \( \tilde{J}_1(j_1) \) is the optimal cost-to-go from original system state \( j_1 \) that has just been generated from an original system state \( i \).
- \( \tilde{J}_m(j_m), m = 2, \ldots, k \), is the optimal cost-to-go from original system state \( j_m \) that has just been generated from an original system state \( j_{m-1} \).
These vectors satisfy the following set of Bellman equations:

$$ r^*_\ell = \sum_{i=1}^{n} d_{i\ell} \tilde{J}_0(i), \quad \ell = 1, \ldots, q, $$

$$ \tilde{J}_0(i) = \min_{u \in U(i)} \sum_{j_1=1}^{n} p_{ij_1}(u) (g(i, u, j_1) + \alpha \tilde{J}_1(j_1)), \quad i = 1, \ldots, n, $$

$$ \tilde{J}_m(j_m) = \min_{u \in U(j_m)} \sum_{j_{m+1}=1}^{n} p_{jmj_{m+1}}(u) (g(j_m, u, j_{m+1}) + \alpha \tilde{J}_{m+1}(j_{m+1})), \quad j_m = 1, \ldots, n, \ m = 1, \ldots, k - 1, $$

$$ \tilde{J}_k(j_k) = \sum_{\ell=1}^{q} \phi_{j_k \ell} r^*_\ell, \quad j_k = 1, \ldots, n. $$

By combining these equations, we obtain an equation for $r^*$:

$$ r^* = DT^k(\Phi r^*), $$

where $T$ is the usual DP mapping of the original problem [cf. Eqs. (4.11)-(4.12), which can be written as $r^* = DT(\Phi r^*)$ and correspond to $k = 1$]. As earlier, it can be seen that the associated mapping $DT^k\Phi$ is a contraction mapping with respect to the sup-norm, but its contraction modulus is $\alpha^k$ rather than $\alpha$.

There is a similar mapping corresponding to a fixed policy and it can be used to implement a PI algorithm, which evaluates a policy through calculation of a corresponding parameter vector $r$ and then improves it. However, there is a major difference from the single-step aggregation case: a policy involves a set of $k$ control functions $\{\mu_0, \ldots, \mu_{k-1}\}$, and while a known policy can be easily simulated, its improvement
involves multistep lookahead using the minimizations of Eqs. (4.22)-(4.24), and may be costly. Thus the preceding implementation of multistep aggregation-based PI is a useful idea only for problems where the cost of this multistep lookahead minimization (for a single given starting state) is not prohibitive.

On the other hand, from a theoretical point of view, a multistep aggregation scheme provides a means of better approximation of the true optimal cost vector $J^*$, independent of the use of a large number of aggregate states. This can be seen from Eqs. (4.22)-(4.24), which by classical value iteration convergence results, show that $\tilde{J}_0(i) \to J^*(i)$ as $k \to \infty$, regardless of the choice of aggregate states. Moreover, because the modulus of the underlying contraction is $\alpha^k$, we can verify an improved error bound in place of the bound (4.13) of Prop. 4.1, which corresponds to $k = 1$:

$$|J^*(i) - J^*_k| \leq \frac{\epsilon}{1 - \alpha^k}, \quad \forall i \text{ such that } i \in I_\ell, \ell = 1, \ldots, q,$$

where $\epsilon$ is given by Eq. (4.14). The proof is very similar to the one of Prop. 4.1.

**$\lambda$-Aggregation**

Another possibility, which is related to multistep aggregation and is called $\lambda$-aggregation, is to introduce a parameter $\lambda \in (0, 1)$ and consider a Markov chain that makes a transition with probability $1 - \lambda$ from an original system state to an aggregate state at each step, rather than with certainty after $k$ steps as in Fig. 4.2. The cost vector of a given stationary policy $\mu$, may be evaluated approximately by $\Phi r(\mu)$, where $r(\mu)$ is the solution of the equation

$$r = DT^\mu_\lambda(\Phi r), \quad (4.25)$$

in the spirit of the projected equation (2.13). This equation has a unique solution because the mapping $DT^\mu_\lambda \Phi$ can be shown to be a contraction mapping with respect to the sup-norm.

As noted earlier, the aggregation equation $\Phi r = \Phi DT^\mu_\lambda(\Phi r)$ is a projected equation because $\Phi D$ is a projection mapping with respect to a suitable Euclidean projection norm. Similarly, the $\lambda$-aggregation equation $\Phi r = \Phi DT^\lambda_\mu(\Phi r)$ is a projected equation, which is related to the proximal algorithm [Ber16a], and may be solved by using temporal differences. Thus we may use exploration-enhanced versions of the LSTD($\lambda$) and LSPE($\lambda$) methods in an approximate PI scheme to solve the multistep aggregation equation $\Phi r = \Phi DT^\lambda(\Phi r)$. We refer to [Ber12] for further discussion of $\lambda$-aggregation.

### 5. POLICY ITERATION WITH FEATURE-BASED AGGREGATION AND A NEURAL NETWORK

We noted in Section 3 that neural networks can be used to construct features at the output of the last nonlinear layer. The neural network training process also yields linear weighting parameters for the feature vector $F(i)$ at the output of the last layer, thus obtaining an approximation $\tilde{J}_\mu(F(i))$ to the cost function
Neural Network Features Approximate Cost

Policy Improvement

µ

Approximately Improved Policy ˜

Current Policy µ

F(i)

Feature Vector

Policy Improvement

Improved Policy ˆ µ

Approximately Improved Policy ˆ µ

Neural Network Features Approximate Cost

Figure 5.1. Schematic illustration of PI using a neural network-based cost approximation. Starting with a training set of state-cost pairs generated using the current policy µ, the neural network yields a set of features and an approximate cost evaluation ˆ Jµ using a linear combination of the features. This is followed by policy improvement using ˆ Jµ to generate the new policy ˆ µ.

of a given policy µ. Thus given the current policy µ, the typical PI produces the new policy ˆ µ using the approximate policy improvement operation (1.3) or a multistep variant, as illustrated in Fig. 5.1.

A similar PI scheme can be constructed based on feature-based aggregation with features supplied by the same neural network; see Fig. 5.2. The main idea is to replace the (approximate) policy improvement operation with the solution of an aggregate problem, which provides the (approximately) improved policy ˆ µ. This is a more complicated policy improvement operation, but computes the new policy ˆ µ based on a more accurate cost function approximation: one that is a nonlinear function of the features rather than linear. Moreover, ˆ µ not only aspires to be an improved policy relative to µ, but also to be an optimal policy based on the aggregate problem, an approximation itself of the original DP problem. In particular, suppose that the neural network approximates Jµ perfectly. Then the scheme of Fig. 5.1 will replicate a single step of the PI algorithm starting from µ, while the aggregation scheme of Fig. 5.2, with sufficient number of aggregate states, will produce a policy that is arbitrarily close to optimal.

Let us now explain each of the steps of the PI procedure of Fig. 5.2, starting with the current policy µ.

(a) Feature mapping construction: We train the neural network using a training set of state-cost pairs that are generated using the current policy µ. This provides a feature vector F(i) as described in Section 3.

(b) Sampling to obtain the disaggregation sets: We sample the state space, generating a subset of states I ⊂ {1, . . . , n}. We partition the corresponding set of state-feature pairs \{(i, F(i)) | i ∈ I\} into a collection of subsets S1, . . . , Sq. We then consider the aggregation framework with S1, . . . , Sq as the aggregate states, and the corresponding aggregate problem as described in Section 4. The sampling to obtain the set of states I may be combined with exploration to ensure that a sufficiently representative set of states is included.

(c) Aggregate problem solution: The aggregate DP problem is solved by using a simulation-based method such as LSTD(λ), LSPE(λ), or Q-learning, to yield (perhaps approximately) the values rℓ, ℓ = 1, . . . , q.

(d) Definition of the improved policy: The “improved” policy is simply the optimal policy of the aggregate
Figure 5.2. Illustration of PI using feature-based aggregation with features supplied by a neural network. Starting with a training set of state-cost pairs generated using the current policy $\mu$, the neural network yields a set of features, which are used to construct a feature-based aggregation framework. The optimal policy of the corresponding aggregate problem is used as the new policy $\hat{\mu}$.

Let us also note that there are several options for implementing the algorithmic ideas of this section.

(1) The neural network-based feature construction process may be used successively any number of times, each time generating a new set of features, corresponding aggregate problems, and corresponding policies. Alternatively, the neural network training and feature construction process may be done only once, followed by the (perhaps approximate) solution of the corresponding feature-based aggregate problem.

(2) Several deep neural network-based PI cycles may be performed, a subset of the features thus generated may be selected, and the corresponding aggregate problem is solved just once, as a way of improving the final policy generated by the deep reinforcement learning process.

(3) Following each cycle of neural network-based feature evaluation, the generated features may be supplemented with additional problem-specific handcrafted features, and/or features from previous cycles. This is a form of feature iteration that was noted in the preceding section.

Finally, let us mention a potential weakness of using the features obtained at the output of the last nonlinear layer of the neural network in the context of aggregation: the sheer number of these features may be so large that the resulting number of aggregate states may become excessive. To address this situation one may consider pruning some of the features, or reducing their number using some form of regression, at the potential loss of some approximation accuracy. In this connection let us also emphasize a point made earlier in connection with an advantage of deep (rather than shallow) neural networks: because with each additional
layer, the generated features tend to be more complex, their number at the output of the final nonlinear layer of the network can be made smaller as the number of layers increases. An extreme case is to use the cost function approximation obtained at the output of the neural network as a single feature/scoring function, in the spirit of Section 4.2.

6. CONCLUDING REMARKS

We have surveyed some aspects of approximate PI methods with a focus on the idea of combining feature-based aggregation with deep neural networks. The policy improvement phase is based on aggregation using a feature vector provided by a neural network, perhaps in combination with handcrafted features. We have argued that this type of policy improvement, while more time-consuming, may yield more effective policies, owing to the DP character of the aggregate problem. The algorithmic idea of this paper seems to work well on small examples. However, tests with challenging problems are needed to fully evaluate its merits, particularly since solving the aggregate DP problem is more time-consuming than the standard one-step lookahead policy improvement scheme of Eq. (2.1) or its multistep lookahead variants.

In this paper we have focused on finite-state discounted Markov decision problems, but our approach clearly extends to other types of finite-state DP involving stochastic uncertainty, including stochastic shortest path problems and semi-Markov decision problems. It is also worth considering extensions of the feature-based aggregation approach to infinite-state problems, including those arising in the context of partially observed Markov decision problems. Finally, one may combine the aggregation approach of this paper with any methodology that automatically constructs good features or scoring functions for approximating the optimal cost function or the cost functions of policies.

7. REFERENCES


[Bis95] Bishop, C. M, 1995. Neural Networks for Pattern Recognition, Oxford University Press, N. Y.


