Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

APPROXIMATE DYNAMIC PROGRAMMING ALGORITHMS FOR UNITED STATES AIR FORCE OFFICER SUSTAINMENT

Joseph C. Hoecherl

Headquarters
United States Air Force
1500 W. Perimeter Rd, Rm 4710
Joint Base Andrews, MD 20762-5000, USA

Matthew J. Robbins
Raymond R. Hill
Darryl K. Ahner

Department of Operational Sciences
Air Force Institute of Technology
2950 Hobson Way
Wright-Patterson AFB, OH 45433-7765, USA

ABSTRACT

We consider the problem of making accession and promotion decisions in the United States Air Force officer sustainment system. Accession decisions determine how many officers should be hired into the system at the lowest grade for each career specialty. Promotion decisions determine how many officers should be promoted to the next highest grade. We formulate a Markov decision process model to examine this military workforce planning problem. The large size of the problem instance motivating this research suggests that classical exact dynamic programming methods are inappropriate. As such, we develop and test approximate dynamic programming (ADP) algorithms to determine high-quality personnel policies relative to current practice. Our best ADP algorithm attains a statistically significant 2.8 percent improvement over the sustainment line policy currently employed by the USAF which serves as the benchmark policy.

1 INTRODUCTION

> "The basic manpower problem is the following: Determine the number of personnel and their skills that best meets the future operational requirements of an enterprise." (Gass 1991)

The United States Air Force (USAF) is comprised of approximately 317,000 personnel who enhance national security by providing the distinctive capabilities of air and space superiority, global attack, rapid global mobility, precision engagement, information superiority, and agile combat support to the Department of Defense (DoD). The USAF, like the other branches of the military, is comprised of commissioned officers as well as the enlisted force. These two groups exhibit significantly different behaviors in regards to retention, promotion, and cross-flow between career fields. This research investigates and attempts to discover improved policies regarding management of the commissioned officer corps.

The USAF must recruit, train, and develop its personnel using limited resources. Over the last several years, the draw down from Operations Enduring Freedom and Iraqi Freedom as well as shifting domestic priorities have resulted in significant cuts to current and future outlays for acquisitions, operations, and personnel budgets. Difficult fiscal conditions emphasize the importance of having the correct mix of personnel to field a ready force. The USAF must balance the needs for officers of varying levels of experience within 90 career fields ranging from personnel officers to fighter pilots. Each of these career fields is labeled with an Air Force Specialty Code (AFSC). The USAF has a known set of requirements (i.e., demand) and a known Congressionally-mandated force size constraint.

Officer grades range from O-1 to O-10, with the grades O-7, O-8, O-9, and O-10 corresponding to the ranks of general officers. Our proposed model only considers grades from O-1 to O-6, representing

the vast majority of officers comprising the officer sustainment problem. Current USAF promotion policy includes a nearly 100% promotion rate from O-1 to O-2 and from O-2 to O-3. This policy is primarily due to long training times and a limited performance record with which to differentiate junior officers at these grades. Moreover, officers at the grades of O-1 and O-2 frequently fill O-3 requirements. A complicating feature of the manpower planning problem faced by the USAF is the fact that senior officers are developed from junior officers only, with every recruited officer starting at the grade of O-1.

Due to these characteristics, poor personnel management decisions can have far-reaching impacts and corrective actions can take a substantial amount of time to take effect. Economic factors and changes within the military environment such as operations tempo, salary, and benefits such as health care and combat pay can significantly impact retention rates (Asch, Hosek, Mattock, and Panis 2008, Murray 2004). This can result in a significant level of deviation in retention rates over time, resulting in a uncertain supply of officers to meet USAF personnel requirements. These factors can make predicting how a force structure will develop and progress quite difficult.

The current USAF personnel system determines the number of requirements for each AFSC and grade combination. Additionally, ten years of historical data are used to calculate retention rates for each combination of AFSC and number of commissioned years of service (CYOS). This information informs the development of a retention line. This retention line assumes a deterministic retention based on historical observations and no future deviation. The current mathematical model utilized to determine accessions smooths the total requirements for each AFSC over the projected retention line. Promotion decisions are made independently, so the accession decisions and promotion decisions are not tied together by means of a holistic policy. By examining only accession decisions, such policies are unable to address any deviations from expected outcomes over the 30 year window. Over time, these deviations can become large, which has historically resulted in the application of expensive measures to boost or lower retention such as paying bonuses to retain people (Lakhani 1988, Simon and Warner 2009) or reductions in force (RIFs) to decrease the size of the force. Deviations compounded by changes in the desired force structure during times of build-up or downsizing can exacerbate the level of correction needed.

We formulate a Markov decision process (MDP) model to examine the Air Force's officer sustainment problem. MDPs have several features that make them particularly suitable for this sort of workforce planning problem. An MDP can provide policies that are state-dependent, which allows for a workforce system to dynamically adjust personnel levels over time to attain target personnel levels. State transitions can be modeled stochastically, allowing the MDP model to address the uncertainty inherent in the personnel system.

The state of the system for this problem is found by aggregating individual officers by class descriptors. The three class descriptors for the USAF officer sustainment system are career field (AFSC), commissioned years of service (CYOS), and grade (i.e., rank). The state represents the current total stock of officers, categorized by each possible combination of class descriptors. In each time period, individuals deterministically maintain their career field, stochastically transition either out of the system or to the next CYOS according to a retention parameter, and stochastically transition either to the next grade or remain in their current grade according to promotion rate decisions made within the model. The model provides a policy π that indicates the number of officers to recruit for each career field (i.e., accessions) as well as the percentage of officers within specified promotion windows to be promoted, given the current state of the personnel system. A group of 54 AFSCs (i.e., a single Line of the Air Force competitive category) is examined, so promotion policies apply to officers in a specified promotion window across all career fields within the model. The contribution function imposes a cost for shortages of officers by career field and grade as well as a cost for exceeding the maximum number of allowable officers. These costs are weighted to reflect the criticality of certain AFSC and grade combinations.

The state space of the motivating problem of interest has 9,720 dimensions representing the full 54 Line of the Air Force AFSCs, 30 CYOS groups, and six grades. This level of dimensionality combined with the stochastic nature of the state transitions makes determining a stationary policy computationally intractable.

The size of the problem suggests that development of an exact dynamic programming algorithm to obtain a solution is inappropriate.

To address the large size of the problem, two approximate dynamic programming (ADP) algorithms are developed to obtain sub-optimal but high-quality solutions. The first proposed algorithm uses least squares temporal differences (LSTD) (Bradtke and Barto 1996) with Bellman error minimization in an approximate policy iteration framework to obtain policies. As part of the process, we simulate potential post-decision states and observe the value of a possible outcome of being in that state. After a batch of these observations is simulated, a regression is performed to minimize Bellman error. This algorithm uses a set of basis functions to approximate the value of the post-decision state. The approximation scheme allows the formulation of a non-linear mixed-integer program to solve the inner minimization problem, obtaining optimal actions based on the current approximation. Algorithm variants utilizing instrumental variables regression and Latin hypercube sampling are also examined.

We also implement a variant of the Concave, Adaptive Value Estimation (CAVE) algorithm (Godfrey and Powell 2002) to develop separable, piecewise linear value function approximations that represent the 'cost to go' function for a finite-horizon formulation of the problem. This algorithm simulates potential outcomes of a given policy and uses the outcomes to update the estimate of the gradient of the value function approximation. The algorithm takes advantage of known problem structure to efficiently update the value function approximation for large numbers of policies simultaneously.

2 MDP FORMULATION

In this section, we present the MDP model of the USAF officer sustainment problem. The set of decision epochs is denoted as:

$$\mathscr{T} = \{1, 2, ..., T\},\ T \leq \infty. \tag{1}$$

The state of each officer in the system is defined by an attribute tuple a, composed of three numerical attributes. These attributes are numerical indices representing AFSC, grade, and CYOS. Due to the Air Force officer grade structure, this model consolidates the three initial officer grades (i.e., company grade officers with grades of O-1, O-2, and O-3) into one index of the grade class descriptor, leaving four grades that are explicitly modeled. Additionally, the AFSCs are limited to the Line of the Air Force competitive category. This group consists of 54 AFSCs and contains approximately 80% of the officers in the USAF. The excluded AFSCs include the medical, dental, legal, and chaplain career fields, whose behavior and constraints are sufficiently different to warrant a separate model.

Let the attribute tuple a denote a specified combination of these three attributes, denoted as a_1, a_2, and a_3:

$$a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} \text{AFSC} \\ \text{Grade} \\ \text{CYOS} \end{pmatrix}. \tag{2}$$

The individual sets containing all elements of a single attribute h are annotated as:

$$\mathscr{A}_h = \text{set of all possible officer attributes } a_h. \tag{3}$$

The full attribute set $\mathscr{A} = \mathscr{A}_1 \times \mathscr{A}_2 \times \mathscr{A}_3$ contains the full range of possible combinations of m AFSCs, n grades, and q CYOS. For the full problem examined, these parameter values are 54, 4, and 30, respectively.

$$\mathscr{A} = \text{set of all possible officer attribute vectors } a, \text{ where } |\mathscr{A}| = mnq \tag{4}$$

The state of the USAF personnel system is defined by the number of resources (i.e., officers) of each attribute tuple, $a \in \mathscr{A}$. Let

$$S_{ta} \equiv S_{t,a_1,a_2,a_3} = \text{number of officers at time } t \in \mathscr{T} \text{ possessing attributes defined by the} \tag{5}$$

$$\text{attribute tuple } a = (a_1, a_2, a_3) \in \mathscr{A}. \tag{6}$$

The pre-decision state of the system is a tuple of size $|\mathscr{A}|$:

$$S_t = (S_{ta})_{a\in\mathscr{A}} \equiv (S_{t,1,1,1}, S_{t,1,1,2}, \ldots, S_{t,1,1,q}, S_{t,1,2,1}, \ldots, S_{t,1,n,q}, S_{t,2,1,1}, \ldots, S_{t,m,n,q}). \tag{7}$$

There are $m+n-1$ decisions made annually, defined by the set $\mathscr{D}$. Decisions $\{1,2,\ldots,m\}$ are the accession decisions for each AFSC, determining how many people to commission into the Air Force as junior officers. Decisions $\{m+1, m+2, \ldots, m+n-1\}$ determine the ratio of eligible officers to promote. For the full problem instance of interest in this paper, $m = 54$ and $n = 4$. The action selected is defined by:

$$x_t = (x_{td})_{d\in\mathscr{D}},\ \forall t \in \mathscr{T}. \tag{8}$$

There are upper bounds, β_d, and lower bounds, ζ_d, for each decision, as specified by the decision maker. For the accession decisions, these bounds arise out of pipeline considerations, such as training constraints or minimum training levels to sustain facilities. For the promotion decisions, extremely high or low values can have significant secondary effects on the quality of the force.

$$\beta_d \le x_{td} \le \zeta_d \qquad \forall d \in \mathscr{D},\ t \in \mathscr{T} \tag{9}$$

A deterministic transition is made from pre-decision S_t to post-decision state S_t^x with each non-promotion eligible officer moving to the next CYOS index or exiting the system (i.e., separating or retiring). Accessions fill the first year group (i.e., CYOS) index for each AFSC and promotion decisions are appended to the end of the pre-decision state, increasing the size of S_t^x by $n-1$ elements (as compared to S_t). The promotion decisions are included in the post-decision state because such decisions determine the next transition. The states with the CYOS index associated with the first year of a potential grade (i.e., new promotions) are set to zero, since these will be determined by the stochastic promotion transitions. The post-decision state tuple is then defined by:

$$\begin{aligned} S_t^x =& (x_{t1}, S_{t,1,1,1}, S_{t,1,1,2}, \ldots, S_{t,1,1,q-1}, 0, S_{t,1,2,1}, \ldots, S_{t,1,n,q-1}, x_{t2}, \\ & \ldots, S_{t,m,n,q-1}, x_{t,m+1}, \ldots, x_{t,m+n-1}),\ \forall t \in \mathscr{T}. \end{aligned} \tag{10}$$

The cost function $C(S)$ is defined by the total sum of the shortages by grade and AFSC, and the overages above the maximum number of personnel allowed in the system (i.e., the Congressionally-authorized end strength). To distinguish between the criticality of shortages for various AFSCs, an AFSC criticality coefficient, $(b_{a_1})_{a_1\in\mathscr{A}_1}$ is used to scale the shortage cost for each AFSC. Requirements by AFSC (a_1) and grade (a_2) are annotated as $(R_{a_1,a_2})_{(a_1,a_2)\in\mathscr{A}_1\times\mathscr{A}_2}$ for all combinations of a_1 and a_2. Let F denote the maximum end strength and let f denote the end strength criticality coefficient. The end strength criticality coefficient allows weighting the relative importance of end strength and shortages using decision maker preferences. Air Staff personnel at Headquarters Air Force, in accordance with senior leader preferences, would determine the criticality coefficients. Let C_H and C_E denote the cost due to shortage and end strength, respectively. Then,

$$C_H(S) = \sum_{a_1=1}^{m}\sum_{a_2=1}^{n} b_{a_1}\left(R_{a_1,a_2} - \sum_{a_3=1}^{q} S_a\right)^+, \tag{11}$$

and

$$C_E(S) = f\left(\left(\sum_{a_1=1}^{m}\sum_{a_2=1}^{n}\sum_{a_3=1}^{q} S_a\right) - F\right)^+. \tag{12}$$

Let $C(S)$ denote the total cost associated with a given state, which is simply the sum of the two partial costs, C_H and C_E:

$$C(S) = \sum_{a_1=1}^{m}\sum_{a_2=1}^{n} b_{a_1}\left(R_{a_1,a_2} - \sum_{a_3=1}^{q} S_a\right)^+ + f\left(\left(\sum_{a_1=1}^{m}\sum_{a_2=1}^{n}\sum_{a_3=1}^{q} S_a\right) - F\right)^+. \tag{13}$$

Promotion and retention transitions are modeled utilizing discrete stochastic functions, each following a binomial distribution. The probability an eligible officer transitions to the next highest grade is determined by the promotion decisions x_d for $d \in \{m+1, m+2, ..., m+n-1\}$:

$$\mathbb{P}(S^x_{t,a_1,a_2+1,a_3} = j) = \binom{S^x_{ta}}{j} (x_d)^j (1-x_d)^{S^x_{ta}-j}. \tag{14}$$

Those who are not selected for promotion remain in their current grade. After all promotion transitions are complete, the probability of any individual transitioning to the next year group (CYOS) within the individual's current grade and AFSC is defined by the retention parameter, ρ_a, that can be derived from historical data and an environmental parameter. This additional environmental parameter can be used to scale historical retention rates to reflect beliefs about changing conditions in the future, such as the impact of macroeconomic trends and operations tempo on force retention. The probability that j officers with attribute tuple a remain in the system at time $t+1$ given there are S^x_{ta} officers in the system at time t and decision x_t is made is expressed as:

$$\mathbb{P}(S_{t+1,a_1,a_2,a_3+1} = j) = \binom{S^x_{ta}}{j} (\rho_a)^j (1-\rho_a)^{S^x_{ta}-j}. \tag{15}$$

The objective function is comprised of the cost of the current state as well as the expected future cost of states resulting from a combination of the current state and chosen actions. The actions selected are the accession and promotion decisions, so the cost associated with the combination of the current state and action, typically defined as $C(S,x)$, is not effected in the present by the action x. Thus, the current contribution is defined by $C(S)$, and the consequence of a given decision is captured by the expected value of future states. The objective is to minimize expected total discounted cost:

$$\min_{\pi \in \Pi} \left(\mathbb{E}^{\pi} \left[\sum_{t=0}^{T} \gamma^t C(S_t) \right] \right), \tag{16}$$

where γ is the discount factor, π is a policy, and Π is the set of all possible personnel policies. To determine the optimal policy, we must find a solution to the Bellman equations

$$V_t(S_t) = \min_x \left(C(S_t) + \gamma \mathbb{E}\left[V_{t+1}(S_{t+1}) | S_t, x \right] \right), \tag{17}$$

where $V_t(S_t)$ is the value of the personnel system being in state S_t at time t.

3 APPROXIMATE DYNAMIC PROGRAMMING ALGORITHMS

Having formulated the MDP model of the USAF officer sustainment problem, we proceed by describing our two ADP approaches. In our first approach, an approximate policy iteration (API) algorithmic strategy is utilized to construct high-quality personnel policies based on value function approximations. The value function approximation scheme employs a linear architecture and utilizes a least squares temporal differences updating technique. In our second approach, an approximate value iteration (AVI) algorithmic strategy is adopted. In our AVI algorithm, we construct high-quality personnel policies based again on value function approximations. However, in our AVI approach the value function approximation scheme uses piecewise linear value function approximations. The CAVE algorithm of Godfrey and Powell (2002) is used to update our approximations.

3.1 Approximate Policy Iteration (API): Least Squares Temporal Differences (LSTD)

In our API-LSTD algorithm, we seek to estimate the parameter vector θ using observations that are created from a set of basis functions $\phi_f(S)$, $f \in \mathscr{F}$. The set $\mathscr{F}$ of basis functions allows us to reduce the

dimensionality of the state variable to a selected number of features, $|\mathscr{F}|$. Using the post-decision state, we can write our value function approximation in a form similar to a standard linear regression model

$$\bar{V}_t^x(S_t^x) = \sum_{f \in \mathscr{F}} \theta_f \phi_f(S_t^x). \tag{18}$$

The steps of our API-LSTD algorithm are shown in Algorithm 1. The policy improvement loop of the algorithm begins once N temporal difference sample realizations have been collected. We compactly denote basis function matrices and cost vectors as follows. Let

$$\Phi_{t-1} \triangleq \begin{bmatrix} \phi(S_{t-1,1}^x)^\top \\ \vdots \\ \phi(S_{t-1,N}^x)^\top \end{bmatrix}, \quad \Phi_t \triangleq \begin{bmatrix} \phi(S_{t,1}^x)^\top \\ \vdots \\ \phi(S_{t,N}^x)^\top \end{bmatrix}, \quad C_t \triangleq \begin{bmatrix} C(S_{t,1}) \\ \vdots \\ C(S_{t,N}) \end{bmatrix},$$

where matrices Φ_{t-1} and Φ_t contain rows of basis function evaluations of the sampled post-decision states, and C_t is the cost vector.

Algorithm 1 API-LSTD Algorithm

1: Initialize θ as a vector of zeros.
2: **for** j = 1 **to** M **(Policy Improvement Loop)**
3: Update $\alpha = (M - j + 1)/(5M)$
4: **for** i = 1 **to** N **(Policy Evaluation Loop)**
5: Simulate a random post-decision state $S_{t-1,i}^x$
6: Simulate the transition to the pre-decision state $S_{t,i}$
7: Solve MINLP for optimal decision $X^\pi(S_t|\theta) = \arg\min_x [C(S_t) + \gamma\theta^T \phi(S_t^x)]$
8: Record $C(S_{t,i})$, $\phi(S_{t-1,i}^x)$, and $\phi(S_{t,i}^x)$
9: **End**
10: Compute $\hat{\theta} = [(\Phi_{t-1} - \gamma\Phi_t)^T(\Phi_{t-1} - \gamma\Phi_t)]^{-1}(\Phi_{t-1} - \gamma\Phi_t)^T C_t$
11: Update $\theta = (\alpha)\hat{\theta} + (1-\alpha)\theta$
12: **End**

At each iteration, θ is smoothed according to a step size, α. Experimentation revealed that a decreasing stepsize significantly outperforms a static stepsize. This allowed rapid updates early in the algorithm while benefiting from a more refined estimation as the algorithm converged to a solution.

The selected basis functions (i.e., features) are the interactions between decisions taken to some power ψ and the sums of the states S_a for various combinations of attributes. This selection helps the algorithm relate current states to potential actions and keeps the problem decomposable. For any given pre-decision state, the inner minimization problem becomes:

$$\min_x Z_1 x_1 + Z_2 x_1^2 + \ldots + Z_\psi x_1^\psi + Z_{\psi+1} x_2 + \ldots + Z_{\psi(m+n-1)} x_{m+n-1}^\psi, \tag{19}$$

subject to any pre-defined bounds on decisions, β_d and ζ_d. Each coefficient Z_g is determined by a number of current states and the parameter θ. The inner minimization problem is a large, decomposable mixed-integer nonlinear program (MINLP) with integer decisions (x_{td}) for all $d \leq m$ and continuous decisions x_{td} for all $d > m$. Decomposability allows separation into m small integer nonlinear programs and $n-1$ small continuous nonlinear programs. Each of these problems is solved easily.

Instrumental variables have been demonstrated to significantly improve regression performance for a subset of problems. We consider an adaptation to the API-LSTD algorithm with ordinary least squares regression by changing Step 9 in Algorithm 1 to reflect an alternative regression equation that utilizes

Algorithm 2 API-LSTD Algorithm with Instrumental Variables (API-LSTD-IV)

Initialize θ as a vector of zeros.
for j = 1 **to** M **(Policy Improvement Loop)**
  Update $\alpha = (M - j + 1)/(5M)$
  for i = 1 **to** N **(Policy Evaluation Loop)**
    Simulate a random post-decision state $S^x_{t-1,i}$
    Simulate the transition to the pre-decision state $S_{t,i}$
    Solve MINLP for optimal decision $X^\pi(S_t|\theta) = \arg\min_x [C(S_t) + \gamma\theta^T \phi(S^x_t)]$
    Record $C(S_{t,i})$, $\phi(S^x_{t-1,i})$, and $\phi(S^x_{t,i})$
  End
  Compute $\hat{\theta} = [(\Phi_{t-1})^T(\Phi_{t-1} - \gamma\Phi_t)]^{-1}(\Phi_{t-1})^T C_t$
  Update $\theta = (\alpha)\hat{\theta} + (1-\alpha)\theta$
End

instrumental variables. The interested reader is directed to Scott, Powell, and Moazehi (2013) for further information. The modified algorithm is presented as Algorithm 2:

Many approximate policy iteration algorithm implementations utilize uniform random sampling of possible post-decision states in Step 4. To improve the ability of the parametric regression to separate the effects of each basis function (i.e., feature) on the value function, Algorithm 3 uses Latin Hypercube Sampling (LHS) to generate an improved set of post-decision states. LHS designs help ensure uniform sampling across all possible dimensions thereby improving the ability of the regression to identify which regressors are significantly affecting the cost function (McKay, Beckman, and Conover 1979). The adapted algorithm with both instrumental variables and LHS is shown in Algorithm 3.

Algorithm 3 API-LSTD-IV Algorithm with Latin Hypercube Sampling

Initialize θ as a vector of zeros.
for j = 1 **to** M **(Policy Improvement Loop)**
  Construct an LHS design of N post-decision states, $[S^x_{t-1,1}, S^x_{t-1,2}, ..., S^x_{t-1,N}]$
  Update $\alpha = (M - j + 1)/(5M)$
  for i = 1 **to** N **(Policy Evaluation Loop)**
    Identify the pre-defined post-decision state $S^x_{t-1,i}$
    Simulate the transition to the pre-decision state $S_{t,i}$
    Solve MINLP for optimal decision $X^\pi(S_t|\theta) = \arg\min_x [C(S_t) + \gamma\theta^T \phi(S^x_t)]$
    Record $C(S_{t,i})$, $\phi(S^x_{t-1,i})$, and $\phi(S^x_{t,i})$
  End
  Compute $\hat{\theta} = [(\Phi_{t-1})^T(\Phi_{t-1} - \gamma\Phi_t)]^{-1}(\Phi_{t-1})^T C_t$
  Update $\theta = (\alpha)\hat{\theta} + (1-\alpha)\theta$
End

3.2 Approximate Value Iteration (AVI): Concave Adaptive Value Estimation (CAVE)

A finite horizon formulation of the problem is solved using a version of the general CAVE algorithm proposed by Godfrey and Powell (2002). Godfrey & Powell use this algorithm to develop a piecewise linear value function approximation $\bar{V}(s)$ when the system is in state s. We adapt this convention to develop a value function approximation for each decision, x_{td}, which is equivalent to the corresponding portion of the post-decision state, $S^x_{t,a_1,1,1}$, for $d_t = a_1$ as well as the promotion decisions.

This algorithm uses a series of breakpoints indexed by k_{d_t}, where $k_{d_t} \in \mathcal{K}_{d_t}$, and $\mathcal{K}_{d_t} = \{0, 1, ..., k_{max}\}$. The parameter k_{max} represents the maximum number of allowable breakpoints. These breakpoints are annotated $(v^{k_{d_t}}, u^{k_{d_t}})$, where $v^{k_{d_t}}$ describes the slope of a linear segment projected from $u^{k_{d_t}}$.

The breakpoints $u^{k_{d_t}}$ are ordered such that $u^1 \equiv 0$ and each consecutive point is monotonically increasing:

$$u^0 < u^1 < ... < u^{k_{max}}. \tag{20}$$

The presence of concavity in the problem structure indicates that the slopes are also monotonically decreasing:

$$v^0 > v^1 > ... > v^{k_{max}}. \tag{21}$$

CAVE uses sampling of the gradients for each decision to improve its estimate of the slope for that approximation. This sampling is accomplished by a single simulation forward in time, calculating the sample gradients $\Delta^-_{d_t}(X_{d_t})$ and $\Delta^+_{d_t}(X_{d_t})$ for the segments being evaluated, $k^-_{d_t}$ and $k^+_{d_t}$.

A smoothing interval, Q_{d_t} for each d_t is initially set based on upper and lower interval size parameters, $\varepsilon^-_{d_t}$ and $\varepsilon^+_{d_t}$. This update interval is then expanded to correct any concavity violations. If necessary, new breakpoints are inserted at the ends of the update interval.

After all of these steps are accomplished, the minimum update interval parameters, $\varepsilon^-_{d_t}$, $\varepsilon^+_{d_t}$ can be decreased to allow the algorithm to create a more granular approximation at the next time step. The step size α can also be decreased as iterations are completed to improve value approximation convergence. The CAVE algorithm as adapted from Godfrey and Powell (2001) to our multiple decision, multiple time period problem is shown in Algorithm 4.

4 RESULTS AND ANALYSIS

4.1 Defining Model Inputs and Measures

Four performance measures provide an overview of each algorithm's level of success for each of two problem instances examined. For these problem instances, percentages for the ADP algorithms are reported in terms of percentage improvement (i.e., decrease in cost) over the benchmark policy. We perform 50 simulations of 50 years each, with each simulation beginning at an optimal state (i.e., no shortages or overages). Half-widths are reported at the 95% confidence level to establish statistical significance. Percent reduction in shortages (RIS), percent reduction in overages (RIO), percent reduction in cost (RIC), and percent reduction in total squared deviation (RSD) are reported. RIC is a direct comparison of performance regarding the objective function.

For implementations of the LSTD algorithm, the inner and outer loop parameters are $M = 30$ and $N = 10000$. The discount factor, γ, is set at 0.95. For each of the models, the end strength criticality coefficient, f, and AFSC criticality coefficients, $(b_{a_1})_{a_1 \in \mathcal{A}_1}$, are set to one. The benchmark policy is the sustainment line method currently practiced by Headquarters Air Force, Air Staff (A1 - Manpower, Personnel, and Services).

For the CAVE algorithm, T is set to twice as large as the maximum career length. This helps correctly assess the overages and shortages at the end of the career for the decisions being made. For the decisions at a time epoch to be reasonably representative of the decisions that will actually be made in the future, the impacts of those decisions are measured over a large number of epochs. The decisions made at the end of a finite time horizon model will be biased, since the model's reality is that only a short number of years are relevant, while the USAF has an enduring requirement for officers. Thus, T must be substantially larger than the maximum career length in order to obtain accurate and unbiased stochastic gradient samples.

4.2 Small Problem Instance Definition & Results

A small problem instance with four AFSCs ($m = 4$), three grades ($n = 3$), 15 CYOSs ($q = 15$), and a upper accession limit of 50 accessions, ($\zeta_d = 50$ for $d \leq 4$), was constructed. With multiple grades, promotion

Algorithm 4 CAVE Algorithm

Step 1: Initialization

For each d_t, let $\mathcal{K}_{d_t} = 0$, where $v_{d_t} = 0$, $u_{d_t} = 0$.

Initialize parameters $\varepsilon_{d_t}^-$, $\varepsilon_{d_t}^+$, and α.

for j = 1 **to** M

Step 2: Collect Gradient Information

For each d_t, identify the policy specified by the current value function approximation, $X_{d_t} \geq 0$.

For all decisions simultaneously, sample the gradients $\Delta_{d_t}^-(X_{d_t}, \omega)$ and $\Delta_{d_t}^+(X_{d_t}, \omega)$ over a finite time horizon with random outcome $\omega \in \Omega$

Step 3: Define Smoothing Interval

Let $k_{d_t}^- = \min\{k_{d_t} \in \mathcal{K}_{d_t} : v_{d_t}^{k_{d_t}} \leq (1-\alpha) v_{d_t}^{k_{d_t}+1} + \alpha \Delta_{d_t}^-(X_{d_t})\}$.

Let $k_{d_t}^+ = \max\{k_{d_t} \in \mathcal{K}_{d_t} : (1-\alpha) v_{d_t}^{k_{d_t}-1} + \alpha \Delta_{d_t}^+(X_{d_t}) \leq v_{d_t}^{k_{d_t}}\}$.

Define the smoothing interval $Q_{d_t} = \left[\min\{X_{d_t} - \varepsilon_{d_t}^-, u_{d_t}^{k_{d_t}^-}\}, \max\{X_{d_t} + \varepsilon_{d_t}^+, u_{d_t}^{k_{d_t}^+ + 1}\}\right)$.
If $u_{d_t}^{(k_{d_t}^+ + 1)}$ is undefined, then set $u_{d_t}^{(k_{d_t}^+ + 1)} = \infty$

Create new breakpoints at X_{d_t} and the endpoints of Q_{d_t} as needed. Since a new breakpoint always divides an existing segment, the segment slopes on both sides of the new breakpoint are the same initially.

Step 4: Perform Smoothing

For each segment in the interval Q_{d_t}, update the slope according to $v_{d_t,new}^k = \alpha \Delta_{d_t} + (1-\alpha) v_{d_t,old}^k$, where $\Delta_{d_t} = \Delta_{d_t}^-(X_{d_t})$ if $u_{d_t}^k < X_{d_t}$ and $\Delta_{d_t} = \Delta_{d_t}^+(X_{d_t})$ otherwise.

Adjust $\varepsilon_{d_t}^-$, $\varepsilon_{d_t}^+$, α according to step size rules.

End

decisions were assessed to determine transition rates from one grade to another. We examined $T = 30$ time epochs when implementing the CAVE algorithm.

As this problem instance was examined, repetitions of all variants of the LSTD algorithm produced substantially different θ vectors, with significantly different solution qualities. For each of the ten runs, the produced policy was simulated to examine solution quality, and the algorithm that performed the best in terms of the objective function was selected. The best LSTD algorithm results for each combination of sampling technique, regression technique, and set of basis functions are shown in Table 1.

Table 1: LSTD Percentage Improvement from Benchmark (Small Problem Instance).

	Random Sampling				Latin Hypercube Sampling	
	Ordinary Least Squares		Instrumental Variables		Instrumental Variables	
Basis Functions	RIC	RSD	RIC	RSD	RIC	RSD
4th Order	-32.86	-5.39	-11.31	-65.21	8.29	17.61
3rd Order	-12.44	-12.05	-12.77	5.44	-8.97	-46.37
2nd Order	-12.94	11.59	-11.51	4.83	5.74	21.88
1st Order	-102.42	-1025.4	-90.53	-867.5	-277.51	-4769.85

As expected, the use of instrumental variables improved solution quality significantly. Latin hypercube sampling improved the solution qualities for all sets of basis functions, but improved the solution quality of

complex sets of basis functions by a more significant margin than those with simpler sets. The algorithm implementations with LHS, instrumental variables, and either second or fourth order basis functions are the only LSTD algorithm implementations that are able to provide policies that improve total cost, although several other implementations show improvements in squared deviation.

When observing the policies generated by the algorithms, it becomes apparent that the LSTD algorithm simplifies the problem by generating solutions that are only pseudo-dynamic. In effect, at least one of the decision policies remains static, while the other decision policies are adjusted higher or lower based on the levels of shortages or overages observed. This limitation is likely due to the value function being unable to project onto the span of the basis functions.

Table 2: ADP Percentage Improvement from Benchmark (Small Problem Instance).

	Algorithm	RIC	Half-Width	RIS	Half-Width	RIO	Half-Width	RSD	Half-Width
LSTD	Ordinary Least Squares	-12.44	1.6	-23.14	3.08	20.93	7.86	-12.05	5.76
	Instrumental Variables	-12.77	1.3	-23.84	2.48	21.98	6.88	5.44	4.98
	Instrumental Variables LHS	8.29	1.16	8.86	2.19	0.65	9.2	17.61	4.2
CAVE	Accessions	5.7	1.05	-5.52	2.19	48.96	4.79	35.24	2.86
	Accessions & Promotions	0.0	1.81	55.9	1.69	-190.47	20.1	-68.53	11.66

Two implementations of CAVE were examined. The first utilizes $m = 4$ accession decisions and $n-1 = 2$ promotion decisions; the second implementation only models the $m = 4$ accessions decisions. The accessions-only implementation of the CAVE algorithm utilizes the benchmark policies for the $n-1 = 2$ promotion decisions. The results using these two algorithms are shown in Table 2. Allowing the promotion rates to vary from the benchmark is a relaxation of a problem constraint. Although the relaxation of a constraint indicates that the variant with promotion decisions should be able to outperform the more constrained accessions-only model, the reverse is observed. This can be attributed to a high level of interaction between the promotion and accession decisions that inhibits CAVE's ability to converge to a high-quality solution. This result is a weakness given that non-linear interactions also exist between accession decisions.

For this small problem instance, the LSTD algorithm with fourth order basis functions, instrumental variables Bellman error minimization, and Latin hypercube sampling outperformed all other algorithms tested. This variant showed a statistically significant decrease in shortages and a statistically insignificant decrease in overages, meaning that the improvement was accomplished due to the dynamic nature of the solution without detrimentally impacting overages or shortages. The CAVE algorithm with accession decisions only also outperformed the benchmark policy by a statistically significant margin, with performance comparable to the LSTD algorithm with second order basis functions, instrumental variables, and Latin hypercube sampling.

4.3 Line of the Air Force, Large Problem Instance - Parameterization & Results

For the large problem instance, the most promising algorithm implementations (i.e., LSTD with 2nd and 4th order basis functions and CAVE with accessions only) were applied to a problem of identical size to the Line of the Air Force problem. This problem is formulated with 54 AFSCs ($m = 54$), four grades ($n = 4$), and 30 CYOSs ($q = 30$). For the CAVE algorithm, we let $T = 60$ due to the maximum 30 year career length. Three behavioral profiles were generated to represent significant differences among observed behaviors of different AFSCs, including a high retention profile, a low retention profile, and a standard profile. These profiles represent the varying levels of demand for the skill sets of different career fields within the USAF. Each AFSC was assigned to one of these profiles, then randomly increased or decreased in size (in terms of the number of officers currently in the AFSC) according to a uniform random distribution. Uniformly distributed unbiased noise was then introduced to the retention rates of the AFSC's behavioral profile to generate career fields that are similar, but not identical. This procedure created a heterogeneous mix of AFSCs with different sizes and retention rates.

Table 3: ADP Percentage Improvement from Benchmark (Large Problem Instance).

Algorithm		RIC	Half-Width	RIS	Half-Width	RIO	Half-Width	RSD	Half-Width
LSTD	2nd Order	-49.89	0.79	-49.8	0.8	-2837	3791.96	-41.93	1.5
	4th Order	-61.47	0.87	-30.3	0.71	-63228	8651.6	-779	35.99
CAVE	Accessions	2.82	0.9	-32.68	2.88	95.53	3.21	1.97	1.09

As shown in Table 3, none of the LSTD algorithms tested improved upon the benchmark solution. In addition to the LSTD algorithm failing to generate policies that outperform the benchmark, the subjective quality of the solutions were low. Many observed policies were stationary over the simulated time, indicating that the algorithm was unable to map the value function closely enough to modify the policy dynamically, given the number of observations. This reinforces the earlier observation that the value function does not appear to project onto the span of the basis functions selected. Selection of alternate basis functions may improve this solution quality.

The CAVE algorithm demonstrates a statistically significant improvement over the benchmark policy. For this large problem instance, the total overages were reduced by decreasing the number of accessions for a small number of accessions by one or two. Overages above maximum allowable end strength were nearly eliminated, although shortages increased significantly. Further analysis with alternative algorithm parameter-values should demonstrate potential shortage and overage trade-offs by adjusting the end strength criticality coefficient and the AFSC criticality coefficients.

5 CONCLUSIONS

This paper presents preliminary research concerning the USAF officer sustainment problem. The intent of the research is to determine personnel policies (e.g., officer accession and promotion decisions as a function of current force levels) that decrease the cost of sustaining the officer corps within the USAF in the long term. We constructed an MDP model of the problem and designed, developed, and tested multiple ADP algorithms to obtain high-quality personnel policies relative to a benchmark policy based on current practice at Headquarters, Air Force. While the LSTD algorithm performed well for the small problem instance we examined, it performed poorly for the larger, full problem instance of interest to the USAF. Conversely, the CAVE algorithm performed well for smaller problem instances and the larger problem instance of interest. The CAVE algorithm attained a statistically significant 2.8% improvement in total discounted cost over the benchmark policy in the large problem instance. In future work, we plan to revisit the problem formulation and incorporate more elements of the real-world problem. With respect to solution methodology, we will develop alternative sets of basis functions to further improve the performance of the LSTD algorithm. Concerning CAVE, we plan to examine different state sampling strategies and examine the manner in which we change the smoothing interval in an effort to improve performance. Finally, we will leverage historical data to verify and validate the simulation models that capture the aggregrate behavior of the USAF officer corps and underlie ADP research efforts such as this one.

ACKNOWLEDGMENTS

The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, the Department of Defense, or the United States Government. The authors thank the reviewers for their comments, which helped improve the quality of this paper. The authors are grateful to the Air Staff (A1 - Manpower, Personnel, and Services), Headquarters, United States Air Force for its encouragement and feedback on this line of research.

REFERENCES

Asch, B., J. Hosek, M. Mattock, and C. Panis. 2008. *Assessing Compensation Reform: Research in Support of the 10th Quadrennial Review of Military Compensation*. Santa Monica, CA: RAND Corporation.

Bradtke, S., and A. Barto. 1996. "Linear Least-Squares Algorithms for Temporal Difference Learning". *Machine Learning* 22 (1-3): 33–57.

Gass, S. 1991. "Military Manpower Planning Models". *Computers & Operations Research* 18 (1): 65–73.

Godfrey, G., and W. Powell. 2001. "An Adaptive, Distribution-Free Algorithm for the Newsvendor Problem with Censored Demands, with applications to inventory and distribution". *Management Science* 47 (8): 1101–1112.

Godfrey, G., and W. Powell. 2002. "An Adaptive Dynamic Programming Algorithm for Dynamic Fleet Management, I: Single Period Travel Times". *Transportation Science* 36 (1): 21–39.

Lakhani, H. 1988. "The Effect of Pay and Retention Bonuses on Quit Rates in the US Army". *Industrial and Labor Relations Review* 41 (3): 430–438.

McKay, M., R. Beckman, and W. Conover. 1979. "Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". *Technometrics* 21 (2): 239–245.

Murray, C. 2004. *Military Compensation: Balancing Cash and Noncash Benefits*. Washington, D.C.: Congressional Budget Office.

Scott, W., W. Powell, and S. Moazehi. 2013. "Least Squares Policy Iteration with Instrumental Variables vs. Direct Policy Search: Comparison Against Optimal Benchmarks Using Energy Storage". Technical report, Dept. of Operations Research and Financial Engineering, Princeton University, Princeton, N.J.

Simon, C., and J. Warner. 2009. "The Supply Price of Commitment: Evidence from the Air Force Enlistment Bonus Program". *Defence and Peace Economics* 20 (4): 269–286.

AUTHOR BIOGRAPHIES

JOSEPH C. HOECHERL is an active duty Air Force officer, serving as an Operations Research Analyst at Headquarters Air Force. He holds an M.S. in Operations Research from the Air Force Institute of Technology. His email address is joseph.c.hoecherl.mil@mail.mil.

MATTHEW J. ROBBINS is an active duty Air Force officer, serving as an Associate Professor of Operations Research in the Department of Operational Sciences at the Air Force Institute of Technology. He holds a Ph.D. in Industrial Engineering from the University of Illinois. His research interests include approximate dynamic programming, game theory, simulation, and applications of operations research in the military and public healthcare domains. He has been recognized with a number of awards, most notably winning the 2011 Pritsker Doctoral Dissertation Award (First Place) from the Institute of Industrial Engineers (IIE). He served as the 2013 WSC Military Track Chair. His email address is matthew.robbins@afit.edu.

RAYMOND R. HILL is a Professor of Operations Research in the Department of Operational Sciences at the Air Force Institute of Technology. He holds a Ph.D. in Industrial and Systems Engineering from The Ohio State University. His research interests include simulation and applied statistics. He was the 2013 WSC General Chair and 2016 WSC Military and Homeland Security Track Chair. His email address is rayrhill@gmail.com.

DARRYL K. AHNER is an Associate Professor of Operations Research in the Department of Operational Sciences at the Air Force Institute of Technology. He holds a Ph.D. in Systems Engineering from Boston University. His research interests include the optimization of stochastic models, dynamic programming, test and evaluation, and military operations research applications. He is Director, Scientific Test and Analysis Techniques in Test & Evaluation Center of Excellence, and is an elected Member of the Board for the Military Operations Research Society where he serves as Vice President of Professional Development. His email address is darryl.ahner@afit.edu.