On the Smoothness of Linear Value Function Approximations

Branislav Kveton
Intelligent Systems Program
University of Pittsburgh
bkveton@cs.pitt.edu

Milos Hauskrecht
Department of Computer Science
University of Pittsburgh
milos@cs.pitt.edu

Abstract

Markov decision processes (MDPs) with discrete and continuous state and action components can be solved efficiently by hybrid approximate linear programming (HALP). The main idea of the approach is to approximate the optimal value function by a set of basis functions and optimize their weights by linear programming. It is known that the solution to this convex optimization problem minimizes the $\mathcal{L}_1$-norm distance in between the optimal value function and its approximation. In this paper, we relate this measure to the max-norm error of the same value function. We believe that this theoretical analysis may help to understand the quality of HALP approximations in continuous domains.

Introduction

Markov decision processes (MDPs) (Bellman 1957; Puterman 1994) provide an elegant mathematical framework for solving sequential decision problems in the presence of uncertainty. However, traditional techniques for solving MDPs are computationally infeasible in real-world domains, which are factored and represented by both discrete and continuous state and action variables. Approximate linear programming (ALP) (Schweitzer & Seidmann 1985) has recently emerged as a promising approach to address these challenges (Kveton & Hauskrecht 2006).

Our paper centers around hybrid ALP (HALP) (Guestrin, Hauskrecht, & Kveton 2004), which is an established framework for solving large factored MDPs with discrete and continuous state and action variables. The main idea of the approach is to approximate the optimal value function by a linear combination of basis functions and optimize it by linear programming (LP). The combination of factored reward and transition models with the linear value function approximation permits the scalability of the approach.

The quality of HALP solutions inherently depends on the choice of basis functions. Therefore, it is often assumed that these are provided as a part of the problem definition, which is unrealistic. The goal of this paper is to analyze the quality of HALP approximations. Based on the analysis, we provide a simple advice for selecting basis functions.

Hybrid factored MDPs

Discrete-state factored MDPs (Boutilier, Dearden, & Goldszmidt 1995) permit a compact representation of stochastic decision problems by exploiting their structure. In this work, we consider hybrid factored MDPs with exponential-family transition models (Kveton & Hauskrecht 2006). This model extends discrete-state factored MDPs to the domains of discrete and continuous state and action variables.

A *hybrid factored MDP with an exponential-family transition model (HMDP)* (Kveton & Hauskrecht 2006) is given by a 4-tuple $\mathcal{M} = (\mathbf{X}, \mathbf{A}, P, R)$, where $\mathbf{X} = \{X_1, \ldots, X_n\}$ is a state space characterized by a set of discrete and continuous variables, $\mathbf{A} = \{A_1, \ldots, A_m\}$ is an action space represented by action variables, $P(\mathbf{X}' \mid \mathbf{X}, \mathbf{A})$ is an exponential-family transition model of state dynamics conditioned on the preceding state and action choice, and R is a reward model assigning immediate payoffs to state-action configurations.[1] In the remainder of the paper, we assume that the quality of a policy is measured by the *infinite horizon discounted reward* $\mathrm{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$, where $\gamma \in [0, 1)$ is a *discount factor* and r_t is the reward obtained at the time step t.

Hybrid ALP

Value iteration, policy iteration, and linear programming are the most fundamental dynamic programming (DP) methods for solving MDPs (Puterman 1994; Bertsekas & Tsitsiklis 1996). Unfortunately, none of these methods is suitable for solving hybrid factored MDPs. First, their complexity grows exponentially in the number of state variables if the variables are discrete. Second, these methods assume a finite support for the optimal value function or policy, which may not exist if continuous variables are present. As a result, any feasible approach to solving arbitrary HMDPs is likely to be approximate. To compute these approximate solutions, Munos and Moore (2002) proposed an adaptive non-uniform discretization of continuous-state spaces and Feng *et al.* (2004) used DP backups of piecewise constant and piecewise linear value functions.

Linear value function model: Since a factored representation of an MDP may not guarantee a structure in the optimal value function or policy (Koller & Parr 1999), we resort to *linear value function approximation* (Bellman, Kalaba, &

[1] *General state and action space MDP* is an alternative name for a hybrid MDP. The term *hybrid* does not refer to the dynamics of the model, which is discrete-time.

Kotkin 1963; Van Roy 1998):

$$V^{\mathbf{w}}(\mathbf{x}) = \sum_i w_i f_i(\mathbf{x}). \quad (1)$$

This approximation restricts the form of the value function $V^{\mathbf{w}}$ to the linear combination of $|\mathbf{w}|$ basis functions $f_i(\mathbf{x})$, where $\mathbf{w}$ is a vector of tunable weights. Every basis function can be defined over the complete state space $\mathbf{X}$, but often is restricted to a subset of state variables $\mathbf{X}_i$ (Bellman, Kalaba, & Kotkin 1963; Koller & Parr 1999).

Similarly to discrete-state ALP (Schweitzer & Seidmann 1985), *hybrid ALP (HALP)* (Guestrin, Hauskrecht, & Kveton 2004) optimizes the linear value function approximation (Equation 1). Therefore, it transforms an initially intractable problem of estimating V^* in the hybrid state space $\mathbf{X}$ into a lower dimensional space $\mathbf{w}$. The HALP formulation is given by a linear program:

$$\begin{aligned} &\text{minimize}_{\mathbf{w}} \sum_i w_i \alpha_i \quad (2)\\ &\text{subject to: } \sum_i w_i F_i(\mathbf{x},\mathbf{a}) - R(\mathbf{x},\mathbf{a}) \geq 0 \quad \forall\, \mathbf{x},\mathbf{a}; \end{aligned}$$

where $\mathbf{w}$ represents the variables in the LP, α_i denotes *basis function relevance weight*:

$$\begin{aligned} \alpha_i &= \mathrm{E}_{\psi(\mathbf{x})}[f_i(\mathbf{x})] \quad (3)\\ &= \sum_{\mathbf{x}_D} \int_{\mathbf{x}_C} \psi(\mathbf{x}) f_i(\mathbf{x})\, \mathrm{d}\mathbf{x}_C, \end{aligned}$$

$\psi(\mathbf{x})$ is a *state relevance density function* weighting the approximation, and $F_i(\mathbf{x},\mathbf{a}) = f_i(\mathbf{x}) - \gamma g_i(\mathbf{x},\mathbf{a})$ is the difference between the basis function $f_i(\mathbf{x})$ and its discounted *backprojection*:

$$\begin{aligned} g_i(\mathbf{x},\mathbf{a}) &= \mathrm{E}_{P(\mathbf{x}'|\mathbf{x},\mathbf{a})}[f_i(\mathbf{x}')] \quad (4)\\ &= \sum_{\mathbf{x}'_D} \int_{\mathbf{x}'_C} P(\mathbf{x}' \mid \mathbf{x},\mathbf{a}) f_i(\mathbf{x}')\, \mathrm{d}\mathbf{x}'_C. \end{aligned}$$

Vectors $\mathbf{x}_D$ ($\mathbf{x}'_D$) and $\mathbf{x}_C$ ($\mathbf{x}'_C$) are the discrete and continuous components of value assignments $\mathbf{x}$ ($\mathbf{x}'$) to all state variables $\mathbf{X}$ ($\mathbf{X}'$). The HALP formulation is feasible if the set of basis functions contains a constant function $f_0(\mathbf{x}) \equiv 1$. We assume that such a basis function is always present.

In the remainder of this paper, we analyze the quality of HALP approximations. Please refer to Hauskrecht and Kveton (2004), Guestrin *et al.* (2004), Kveton and Hauskrecht (2005), and Kveton and Hauskrecht (2006) for information on how to apply and solve HALP formulations.

Existing work

De Farias and Van Roy (2003) analyzed the quality of ALP. Based on their work, we may conclude that optimization of the objective function $\mathrm{E}_\psi[V^{\mathbf{w}}]$ in HALP is identical to minimizing the $\mathcal{L}_1$-norm error $\|V^* - V^{\mathbf{w}}\|_{1,\psi}$. This equivalence can be proved from the following proposition.

Proposition 1 *Let $\widetilde{\mathbf{w}}$ be a solution to the HALP formulation (2). Then $V^{\widetilde{\mathbf{w}}} \geq V^*$.*

Proof: The Bellman operator $\mathcal{T}^*$ is a contraction mapping. Based on its monotonicity, $V \geq \mathcal{T}^* V$ implies $V \geq \mathcal{T}^* V \geq \cdots \geq V^*$ for any value function V. Since constraints in the HALP formulation (2) enforce $V^{\widetilde{\mathbf{w}}} \geq \mathcal{T}^* V^{\widetilde{\mathbf{w}}}$, we conclude $V^{\widetilde{\mathbf{w}}} \geq V^*$. ■

Based on Proposition 1, we know that HALP optimizes the linear value function model with respect to the weighted $\mathcal{L}_1$-norm error $\|V^* - V^{\mathbf{w}}\|_{1,\psi}$. The following theorem bounds the quality of a greedy policy for the value function $V^{\widetilde{\mathbf{w}}}$.

Theorem 1 *Let $\widetilde{\mathbf{w}}$ be an optimal solution to the HALP formulation (2). Then the expected error of a greedy policy:*

$$u(\mathbf{x}) = \arg\sup_{\mathbf{a}} \left[R(\mathbf{x},\mathbf{a}) + \gamma \mathrm{E}_{P(\mathbf{x}'|\mathbf{x},\mathbf{a})} \left[V^{\widetilde{\mathbf{w}}}(\mathbf{x}') \right] \right]$$

can be bounded as:

$$\|V^* - V^u\|_{1,\nu} \leq \frac{1}{1-\gamma} \left\| V^* - V^{\widetilde{\mathbf{w}}} \right\|_{1,\mu_{u,\nu}},$$

where $\|\cdot\|_{1,\nu}$ and $\|\cdot\|_{1,\mu_{u,\nu}}$ are weighted $\mathcal{L}_1$-norms, V^u is a value function for the greedy policy u, and $\mu_{u,\nu}$ denotes the expected frequency of state visits generated by following the policy u given the initial state distribution ν.

Based on Theorem 1, the state relevance density function ψ should resemble the expected frequency of state visits $\mu_{u,\nu}$. Unfortunately, $\mu_{u,\nu}$ is unknown unless $V^{\widetilde{\mathbf{w}}}$ is known, which is optimized with respect to the unknown distribution $\mu_{u,\nu}$. To break this cycle, de Farias and Van Roy (2003) suggested an iterative scheme that resolves several LPs and adapts $\mu_{u,\nu}$ accordingly. Alternatively, real-world control problems often exhibit a lot of structure, which permits guessing of $\mu_{u,\nu}$.

Error bounds

This section demonstrates how to bound the max-norm error $\|V^* - V^{\mathbf{w}}\|_\infty$ of a linear approximation $V^{\mathbf{w}}$ in terms of its $\mathcal{L}_1$-norm error $\|V^* - V^{\mathbf{w}}\|_{1,\psi}$. This result is a step towards understanding the quality of HALP approximations. For instance, based on the work of Williams and Baird III (1993), we can bound the loss of acting greedily with respect to the value function $V^{\mathbf{w}}$ by its max-norm error $\|V^* - V^{\mathbf{w}}\|_\infty$. In combination with our work (Theorems 2 and 3), we can derive max-norm bounds on the quality of greedy policies for HALP approximations. Note that Theorem 1 only provides bounds on the $\mathcal{L}_1$-norm errors of greedy policies.

For discrete-state factored MDPs, we can easily prove the following proposition.

Proposition 2 *Let $\widetilde{\mathbf{w}}$ be an optimal solution to the HALP formulation (2) with discrete state variables. Then the max-norm error of $V^{\widetilde{\mathbf{w}}}$ can be bounded as:*

$$\left\| V^* - V^{\widetilde{\mathbf{w}}} \right\|_{\infty,\psi} \leq \left\| V^* - V^{\widetilde{\mathbf{w}}} \right\|_{1,\psi},$$

where $\|\cdot\|_{1,\psi}$ and $\|\cdot\|_{\infty,\psi}$ are $\mathcal{L}_1$ and infinity norms weighted by the state relevance density function ψ.

Proof: The claim directly follows from the definition of the norms $\|\cdot\|_{1,\psi}$ and $\|\cdot\|_{\infty,\psi}$. ■

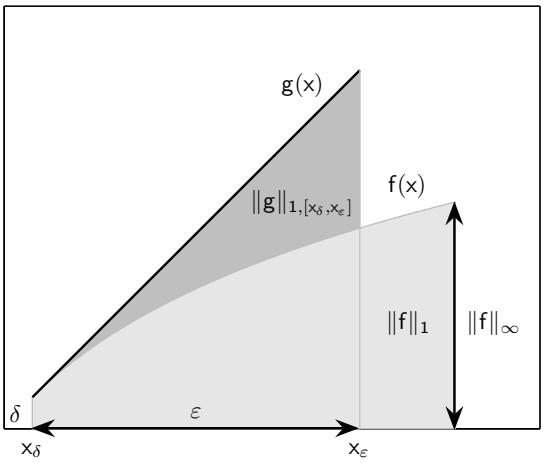

Figure 1: A graphical representation of the bound from Theorem 2 in a single dimension. Light and dark gray regions correspond to the integrals $\|f\|_1$ and $\|g\|_{1,[\mathbf{x}_\delta,\mathbf{x}_\varepsilon]}$.

Unfortunately, this bound is loose and not of much practical interest. In the remainder of this section, we prove a tighter bound for continuous-state spaces. The notion of continuity is captured through the Lipschitz condition.

Definition 1 *The function $f(\mathbf{x})$ is* Lipschitz continuous *if:*

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq K \|\mathbf{x} - \mathbf{x}'\|_\infty \quad \forall\, \mathbf{x}, \mathbf{x}'; \tag{5}$$

where K is referred to as a Lipschitz constant.

In the rest of the paper, it is useful to think of the max-norm ($\mathcal{L}_1$-norm) as being the supremum (integral) of a function.

Theorem 2 *Let $\tilde{\mathbf{w}}$ be an optimal solution to the HALP formulation (2) with continuous state variables. If the function:*

$$f(\mathbf{x}) = \psi(\mathbf{x}) \left| V^*(\mathbf{x}) - V^{\tilde{\mathbf{w}}}(\mathbf{x}) \right|$$

is Lipschitz continuous, and there exists a state $\mathbf{x}_\delta$ such that $f(\mathbf{x}_\delta) \leq \delta$, the max-norm error of $V^{\tilde{\mathbf{w}}}$ can be bounded as:

$$\left\| V^* - V^{\tilde{\mathbf{w}}} \right\|_{\infty,\psi} \leq \delta + K \min \left\{ \sqrt[n]{\frac{\|V^* - V^{\tilde{\mathbf{w}}}\|_{1,\psi}}{\delta C}},\ \sqrt[n+1]{\frac{2\,\|V^* - V^{\tilde{\mathbf{w}}}\|_{1,\psi}}{KC}} \right\},$$

where $\|\cdot\|_{1,\psi}$ and $\|\cdot\|_{\infty,\psi}$ are $\mathcal{L}_1$ and infinity norms weighted by the state relevance density function ψ, n is the number of state variables, K represents the Lipschitz constant of $f(\mathbf{x})$, and C is a problem-specific constant.

Proof: To prove the theorem, we define a function:

$$g(\mathbf{x}) = \delta + K \|\mathbf{x} - \mathbf{x}_\delta\|_1 . \tag{6}$$

It follows that the function $g(\mathbf{x})$ is an upper bound on $f(\mathbf{x})$ because $f(\mathbf{x}_\delta) \leq \delta$, $\|\mathbf{x} - \mathbf{x}_\delta\|_1 \geq \|\mathbf{x} - \mathbf{x}_\delta\|_\infty$, and K is the Lipschitz constant of $f(\mathbf{x})$. Furthermore, $g(\mathbf{x})$ is increasing faster than $f(\mathbf{x})$ in every dimension. As a result, there exists a point $\mathbf{x}_\varepsilon$ such that $g(\mathbf{x}_\varepsilon) \geq \|f\|_\infty$, and the integral of $g(\mathbf{x})$ between $\mathbf{x}_\delta$ and $\mathbf{x}_\varepsilon$ is smaller or equal to $\|f\|_1$. A graphical interpretation of this situation in a single dimension is shown in Figure 1.

In general, the integral $\|g\|_{1,[\mathbf{x}_\delta,\mathbf{x}_\varepsilon]}$ can be computed as:

$$\|g\|_{1,[\mathbf{x}_\delta,\mathbf{x}_\varepsilon]} = \left[\delta + \frac{K}{2}\varepsilon\right] \prod_{i=1}^{n} \varepsilon_i,$$

where $\varepsilon_i = |x_{\varepsilon i} - x_{\delta i}|$, $\varepsilon = \|\mathbf{x}_\varepsilon - \mathbf{x}_\delta\|_1$, and n denotes the number of state variables $\mathbf{X}$. Since $\varepsilon = \sum_{i=1}^{n} \varepsilon_i$, we rewrite the equation as:

$$\|g\|_{1,[\mathbf{x}_\delta,\mathbf{x}_\varepsilon]} = C \left[\delta + \frac{K}{2}\varepsilon\right] \varepsilon^n,$$

where $C = \prod_{i=1}^{n} (\varepsilon_i / \varepsilon)$ is a problem-specific constant that guarantees $g(\mathbf{x}_\varepsilon) \geq \|f\|_\infty$. The constant C is bounded from above by n^{-n}. Finally, we recognize that C, δ, K, and ε are always nonnegative, which leads to the conclusion:

$$\varepsilon \leq \min \left\{ \sqrt[n]{\frac{\|f\|_1}{\delta C}},\ \sqrt[n+1]{\frac{2\,\|f\|_1}{KC}} \right\} \tag{7}$$

assuming $\|g\|_{1,[\mathbf{x}_\delta,\mathbf{x}_\varepsilon]} \leq \|f\|_1$. Direct combination of Equations 6 and 7 yields our final result. ■

To make the bound in Theorem 2 practical, we have to assure a low Lipschitz factor K and the existence of a state $\mathbf{x}_\delta$ such that $f(\mathbf{x}_\delta) \leq \delta$. We cannot guarantee the existence of such a state yet. However, we can affect the factor K by the choice of basis functions and state relevance densities. In particular, to achieve a low value K, we should use basis functions that yield close approximations to V^*. In practice, this condition cannot be guaranteed unless we know V^*. Furthermore, the Lipschitz factor of V^* may be large itself. To address these concerns, we generalize Theorem 2 to an arbitrary partitioning of the state space $\mathbf{X}$.

Theorem 3 *Let $\tilde{\mathbf{w}}$ be an optimal solution to the HALP formulation (2) with continuous state variables. If:*

$$\Omega = \{\omega_1, \ldots, \omega_{|\Omega|}\}$$

is a mutually-exclusive partitioning of the state space $\mathbf{X}$, the function:

$$f(\mathbf{x}) = \psi(\mathbf{x}) \left| V^*(\mathbf{x}) - V^{\tilde{\mathbf{w}}}(\mathbf{x}) \right|$$

is Lipschitz continuous on each partition, and there exists a state $\mathbf{x}_{\delta_\omega}$ for every ω such that $f(\mathbf{x}_{\delta_\omega}) \leq \delta_\omega$, the max-norm error of $V^{\tilde{\mathbf{w}}}$ can be bounded as:

$$\left\| V^* - V^{\tilde{\mathbf{w}}} \right\|_{\infty,\psi} \leq \max_{\omega \in \Omega} \left\{ \delta_\omega + K_\omega \min \left\{ \sqrt[n]{\frac{\|V^* - V^{\tilde{\mathbf{w}}}\|_{1,\psi_\omega}}{\delta_\omega C_\omega}},\ \sqrt[n+1]{\frac{2\,\|V^* - V^{\tilde{\mathbf{w}}}\|_{1,\psi_\omega}}{K_\omega C_\omega}} \right\} \right\},$$

where the explanation of symbols is identical to Theorem 2. All subscripted symbols are partition-specific.

Proof: Based on the definition of the max-norm $\|\cdot\|_{\infty,\psi}$, we conclude:

$$\left\|V^* - V^{\tilde{\mathbf{w}}}\right\|_{\infty,\psi} = \max_{\omega\in\Omega} \left\|V^* - V^{\tilde{\mathbf{w}}}\right\|_{\infty,\psi_\omega},$$

where $\psi_\omega(\mathbf{x}) = \psi(\mathbf{x})\mathbf{1}_{\mathbf{x}\in\omega}(\mathbf{x})$, and $\mathbf{1}_{\mathbf{x}\in\omega}(\mathbf{x})$ is the indicator function of the partition ω. The final result is a consequence of bounding each $\left\|V^* - V^{\tilde{\mathbf{w}}}\right\|_{\infty,\psi_\omega}$ by Theorem 2. ■

Theorem 2 provides an insight into the relation between the $\mathcal{L}_1$-norm objective $\left\|V^* - V^{\tilde{\mathbf{w}}}\right\|_{1,\psi}$ and the max-norm error $\left\|V^* - V^{\tilde{\mathbf{w}}}\right\|_{\infty,\psi}$. The max-norm error can be minimized by lowering $\mathcal{L}_1$-norm errors $\left\|V^* - V^{\tilde{\mathbf{w}}}\right\|_{1,\psi_\omega}$ if the growth rate of K_ω and δ_ω is controlled. This result leads to an intuitive advice for choosing basis functions. If the shape of the value function V^* is not known, we should prefer smooth approximations. These are not likely to inflate Lipschitz constants K_ω where V^* is smooth.

Conclusions

Development of efficient methods for solving large factored MDPs is a challenging problem. In this paper, we analyzed the quality of linear approximations and bounded their max-norm error by the objective value in HALP. We believe that this analysis can help us to understand the quality of HALP approximations in continuous domains.

Acknowledgment

During the academic years 2004-06, the first author was supported by two Andrew Mellon Predoctoral Fellowships. The first author recognizes support from Intel Corporation in the summer 2005. This research was also partially supported by two National Science Foundation grants CMS-0416754 and ANI-0325353. We thank anonymous reviewers for providing comments that led to the improvement of the paper.

References

Bellman, R.; Kalaba, R.; and Kotkin, B. 1963. Polynomial approximation – a new computational technique in dynamic programming: Allocation processes. *Mathematics of Computation* 17(82):155–161.

Bellman, R. 1957. *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Bertsekas, D., and Tsitsiklis, J. 1996. *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific.

Boutilier, C.; Dearden, R.; and Goldszmidt, M. 1995. Exploiting structure in policy construction. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1104–1111.

de Farias, D. P., and Van Roy, B. 2003. The linear programming approach to approximate dynamic programming. *Operations Research* 51(6):850–856.

Feng, Z.; Dearden, R.; Meuleau, N.; and Washington, R. 2004. Dynamic programming for structured continuous Markov decision problems. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 154–161.

Guestrin, C.; Hauskrecht, M.; and Kveton, B. 2004. Solving factored MDPs with continuous and discrete variables. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 235–242.

Hauskrecht, M., and Kveton, B. 2004. Linear program approximations for factored continuous-state Markov decision processes. In *Advances in Neural Information Processing Systems 16*, 895–902.

Koller, D., and Parr, R. 1999. Computing factored value functions for policies in structured MDPs. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 1332–1339.

Kveton, B., and Hauskrecht, M. 2005. An MCMC approach to solving hybrid factored MDPs. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 1346–1351.

Kveton, B., and Hauskrecht, M. 2006. Solving factored MDPs with exponential-family transition models. In *Proceedings of the 16th International Conference on Automated Planning and Scheduling*.

Munos, R., and Moore, A. 2002. Variable resolution discretization in optimal control. *Machine Learning* 49:291–323.

Puterman, M. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY: John Wiley & Sons.

Schweitzer, P., and Seidmann, A. 1985. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications* 110:568–582.

Van Roy, B. 1998. *Planning Under Uncertainty in Complex Structured Environments*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Williams, R., and Baird III, L. 1993. Tight performance bounds on greedy policies based on imperfect value functions. Technical Report NU-CCS-93-14, Northeastern University.