

Markov Modeling of Component Fault Growth Over A Derived Domain of Feasible Output Control Effort Modifications

Brian Bole¹, Kai Goebel², George Vachtsevanos³

^{1,3} *Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA.*

bbole3@gatech.edu

gfv@ece.gatech.edu

² *NASA Ames Research Center, Moffett Field, CA 94035, USA.*

kai.goebel@nasa.gov

ABSTRACT

This paper introduces a novel Markov process formulation of stochastic fault growth modeling, in order to facilitate the development and analysis of prognostics-based control adaptation. A metric representing the relative deviation between the nominal output of a system and the net output that is actually enacted by an implemented prognostics-based control routine, will be used to define the action space of the formulated Markov process. The state space of the Markov process will be defined in terms of an abstracted metric representing the relative health remaining in each of the system's components. The proposed formulation of component fault dynamics will conveniently relate feasible system output performance modifications to predictions of future component health deterioration.

1. INTRODUCTION

Continuous improvements in fault diagnostic and fault growth prognostic technologies have spawned a prolific growth in the application domain for these technologies, and a growing research interest in the development of techniques for using available diagnostic and prognostic information to improve system control and maintenance. Presently, a clear analytical process may be applied to implement and evaluate the effectiveness of tools for fault classification and fault growth prediction; however, the same cannot generally be said for the follow-on task of making intelligent control decisions based on available diagnostic and prognostic information. In general, the application of an analytical approach to the implementation and evaluation of prognostics-based decision making techniques will be complicated by the potential for high uncertainty in estimating the future effects of avail-

able control actions, and the need to define a computationally tractable space of present and future control decisions to be optimized over.

Several recent publications have suggested metrics for quantifying prognostic uncertainty and evaluating the ability of control modifications to affect meaningful change on fault growth predictions (Saxena et al., 2008; P. Wang et al., 2012; Edwards et al., 2010). This paper describes a process for abstracting uncertain models of environmental and fault growth dynamics into the generalized notation of a non-deterministic Markov process, in order to promote an application independent analysis of prognostics-based control strategies.

Markov decision processes have been widely applied to the representation of problems involving decision making in the presence of uncertain or stochastic modeling information in the contexts of economics (Haurie & Moresino, 2006), supply chain management (Parlara et al., 1995), scheduled maintenance (Smilowitz & Madanat, 1994), health care (Sonnenberg & Beck, 1993), and many other disciplines, in addition to being a widely used tool for describing fault-adaptive and robust control problems (Zhang & Jiang, 2008). A formal description of fault growth modeling and remaining useful life (RUL) estimation in terms of Markov process models, as well as a survey of similar stochastic modeling techniques, are given in Banjevic and Jardine (2006).

The state transition probabilities in a Markov process description of fault dynamics may be chosen to approximate an analytical formulation of a stochastic process, such as a Gaussian process model of fault growth dynamics, as described in Sankararaman et al. (2009); alternatively, state transition probabilities may be defined purely based on empirical observations of the fault growth process, as is the case with hidden Markov model learning techniques (Baruah & Chinnam, 2005), or they may be derived from a combination of a priori and empirical information.

Brian Bole et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

The derivation of stochastic fault growth models for a particular application is not specifically addressed in this publication; however, an effort is made to clearly identify how various sources of uncertain modeling information would be incorporated into the Markov process representation of fault dynamics. Arguments are provided for defining the state space of the fault growth process in terms of a metric representing the relative health of system components, as well as for defining the action space of the fault growth process in terms of a metric representing the relative deviation between the system's nominal output response and the net system output that is actually enacted by an implemented prognostics-based control routine at each control time-step. Operational constraints on minimum acceptable system output performance and maximum acceptable fault growth risk will be formulated as bounds on the action space available to implemented controllers.

2. GENERALIZED STOCHASTIC MODELING OF THE FAULT GROWTH PROCESS

Consider a discretized fault growth process of the form:

$$\gamma_l \in S_l, \quad S_l = \{\alpha_l, \alpha_l + \Delta_l, \alpha_l + 2\Delta_l, \dots, \alpha_l + m_l \Delta_l\} \quad (1)$$

$$p_{i,j}^l(\mathbf{x}(k), \mathbf{a}(k), \mathbf{w}(k)) = P(\gamma_l(k+1) = s_j | \gamma_l(k) = s_i, \mathbf{a} = \mathbf{a}(k), \mathbf{w} = \mathbf{w}(k)), \quad (2)$$

$$s_i, s_j \in S_l, \quad \mathbf{a}(k) \in \mathbf{A}(k), \quad \mathbf{w}(k) \in \mathbf{W}(k), \quad k \in \mathbb{N}$$

$$\sum_{j=0}^m p_{i,j}^l = 1, \quad \forall i \in \{0, 1, \dots, m\} \quad (3)$$

where $\gamma_l(k)$ is a random variable representing the magnitude of the l^{th} component fault mode at time-index k , S_l represents a uniformly quantized state space of potential fault magnitudes, and $p_{i,j}^l(\mathbf{x}, \mathbf{a}, \mathbf{w})$ represents the probability of transitioning from damage state s_i to damage state s_j , given estimates of the system state \mathbf{x} , a set of low-level control commands \mathbf{a} , and the states of environmental and other exogenous inputs to the system \mathbf{w} . Eq. (3) specifies that the sum of all transition probabilities defined at each system state must always be equal to one.

The Markov process notation given in Eqs. (1)-(3) may be used to describe fault growth processes for all systems in which the following assumptions are satisfied:

Assumption 1. State transition probabilities $p_{i,j}^l$ are only dependent on the current states of γ_l , \mathbf{x} , \mathbf{a} , and \mathbf{w} , and not on any states or inputs occurring previously. This is referred to as the memoryless assumption, or the Markov assumption. For cases in which the fault growth process is not completely memoryless, but future states only depend on a finite number m of previous states, the Markov assumption can be satisfied by defining the state space of the process to be the ordered m -

tuple of the current state and the m previously visited states (H. S. Wang & Chang, 1996).

Assumption 2. State transition probabilities are considered to be time invariant; although, it may be the case that fault growth models are not precisely known a priori and must be adapted online using techniques such as particle filtering (Orchard et al., 2008) or Bayesian learning (Saha et al., 2009). Online adaptation of the prognostic model on the basis of past observations will technically violate Assumption 1; however, the error induced by this effect is typically ignored because model adaptation transients are generally difficult to characterize and they will die out as the model is adapted.

Assumption 3. At all discrete time-steps, the state space \mathbf{S} , the action space \mathbf{A} , and the space of environmental and other exogenous inputs to the system \mathbf{W} are adequately represented by a finite set of states, which will be bounded from above by the availability of computational resources. In the event that fault growth must be modeled as a continuous time process, such as the Poisson process (Shetty et al., 2008), a representation of fault growth modeling similar to that given in Eq. (2) may be expressed in terms of a continuous time Markov process (Serfozo, 1979) or a semi-Markov process (Dong & He, 2007).

The required assumptions are mild enough to allow a wide array of fault growth processes to be described in terms of the Markov process notation given in Eqs. (1)-(3) (Guidaa & Pulcini, 2011; Tang et al., 2009).

2.1. Formulating the Markov process in terms of component health rather than fault magnitude

Component fault magnitudes can generally be described by a real number corresponding to a measurable physical property such as crack length, spall width, or pitting depth; although, in many cases, faults cannot be directly measured in situ and diagnostic routines are needed to approximate current fault magnitudes based on the secondary effects observed in available sensor measurements. Sensor noise and modeling uncertainties will often result in significant diagnostic uncertainty, and it is common practice for diagnostic estimates to be reported in terms of a probability distribution over the potential fault magnitudes that could correspond to a given set observations.

The notation introduced in this subsection will add a layer of abstraction to the fault magnitude estimates produced by online fault diagnostic and fault growth prognostic routines. Rather than formulating the Markov process in terms of component fault magnitudes, a transformation is assumed to exist that will express the fault growth process in terms of an application independent metric representing the state of health (SOH) of each system component.

The SOH for component l at time t will be represented in

terms of a probability distribution over a uniformly quantized set of component health percentages between 0% and 100%;

$$\begin{aligned} \gamma_l(t) \in S, \quad S = \{s_1, s_2, \dots, s_m\}, \\ s_1 = 0\%, \quad s_i = s_{i-1} + \Delta\%, \quad 1 < i \leq 100\%/\Delta\% \end{aligned} \quad (4)$$

where $\Delta\%$ is a quantization step-size for the state space of γ_l . The notation given in Eq. (2) is now considered to define the probability of transitioning from one component SOH to another, on the basis of the current states of \mathbf{x} , \mathbf{a} , and \mathbf{w} at time-index k .

Component fault magnitudes are generally expected to monotonically increase with time; correspondingly, component SOH should monotonically decrease with time. A mandate of monotonically decreasing component health is incorporated into the Markov process notation as follows:

$$p_{i,j}^l = 0, \text{ if } j > i, \quad \sum_{j=0}^m p_{i,j}^l = 1 \quad \forall i \in \{0, 1, \dots, m\} \quad (5)$$

This constraint will be problematic for techniques that model process uncertainty with an analytical distribution that lacks an explicit lower bound. For example, in the case of fault growth prognostic techniques such as Kalman filtering and Gaussian process modeling, an assumption of Gaussian uncertainty will introduce some probability that the fault mode does not monotonically increase. In such cases, it would be necessary to assure that the probability attributed to the non-realizable outcomes, $P(\gamma_l(\tau) < \gamma_l(t))$ for $\tau > t$, will be acceptably small. It will not always be the case that component health is strictly monotonically decreasing; however, consideration of this constraint serves to illustrate the flexibility provided by representing modeling uncertainties in terms of a quantized probability mass function (pmf), when compared to techniques that assume a continuous probability distribution function (pdf).

2.2. Formulating the action space of the fault growth process in terms of commanded and applied loads

Consider a process model for component damage accumulation that is expressed in terms of a metric representing the load or stress applied to a component at each control time-step:

$$\begin{aligned} p_{i,j}^l(u_l) = P(\gamma_l(k+1) = s_j | \gamma_l(k) = s_i, u_l = u_l(k)), \\ s_i, s_j \in S, \quad u_l \in U_l(k) \end{aligned} \quad (6)$$

where u_l represents the load applied to component l and $U_l(k)$ represents the domain of feasible component load allocations for component l at time-index k . The component loading variable may represent pressure, force, torque, or a wide variety of other stressors that drive component damage.

Determination of $U_l(k)$ at present and future control time-

steps will require a mapping function to translate system hardware limitations and estimated environmental loading conditions into the component loading domain. A mapping between available low-level control actions and feasible component loadings, as well as an inverse mapping are both assumed to be known:

$$\begin{aligned} F(\mathbf{a}(k), \mathbf{x}(k), \boldsymbol{\gamma}(k), \mathbf{w}(k)) : \quad \mathbf{A}(k) \rightarrow \mathbf{U}(k) \\ F(\mathbf{u}(k), \mathbf{x}(k), \boldsymbol{\gamma}(k), \mathbf{w}(k))^{-1} : \quad \mathbf{U}(k) \rightarrow \mathbf{A}(k) \end{aligned} \quad (7)$$

where F represents a mapping from a given system state, health state, exogenous demand state, and a set of low-level control actions available at time-index k , onto the domain of feasible component load allocations available at time-index k .

2.2.1. Performance constraints

In addition to being defined in terms of the hardware limitations of control effectors, the domain of feasible component load allocations will generally also be bounded by operational constraints on minimum allowable output performance and maximum allowable fault growth risk.

Consider a constraint on minimum allowable system performance that is defined in terms of a maximum allowable deviation from some commanded system state:

$$|\mathbf{x}_c(k) - \mathbf{x}(k)|_i \leq \Delta_i(k), \quad i \in \{1, 2, \dots, n\} \quad (8)$$

where \mathbf{x} and \mathbf{x}_c are n dimensional vectors that represent the actual and commanded output states of a system respectively, and Δ_i specifies a maximum acceptable error between the i^{th} dimensions of \mathbf{x} and \mathbf{x}_c .

If the system's kinematics are known, then Newton's laws of motion can be applied to express the system's dynamics in terms of the instantaneous forces on its constituent components,

$$\dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{u}, \mathbf{w}) \quad (9)$$

If the system is overactuated, then formulating the system's dynamics in terms of instantaneous component loads will result in actuation redundancies being identified by the nullspace of \mathbf{u} . Consider the following generic representation of nonlinear system dynamics:

$$\dot{\mathbf{x}} = A(\mathbf{x}, \mathbf{w}) + B(\mathbf{x}) \mathbf{u} \quad (10)$$

where $A(\mathbf{x}, \mathbf{w}) \in \mathbb{R}^n$, $B(\mathbf{x}) \in \mathbb{R}^{n \times q}$, $\mathbf{x}(t) \in \mathbb{R}^n$, and $\mathbf{u}(t) \in \mathbb{R}^q$. If $B(\mathbf{x})$ does not have full column rank, i.e., $\text{rank}\{B(\mathbf{x})\} = p < q \quad \forall \mathbf{x}$, then the system is overactuated, and $B(\mathbf{x})$ can be factorized as:

$$B(\mathbf{x}) = B_\nu(\mathbf{x}) B_u(\mathbf{x}) \quad (11)$$

where $B_\nu(\mathbf{x}) \in \mathbb{R}^{n \times p}$ and $B_u(\mathbf{x}) \in \mathbb{R}^{p \times q}$ both have rank p .

The system can now be rewritten as:

$$\begin{aligned}\dot{\mathbf{x}} &= A(\mathbf{x}, \mathbf{w}) + B_\nu(\mathbf{x}) \boldsymbol{\nu} \\ \boldsymbol{\nu} &= B_u(\mathbf{x}) \mathbf{u}\end{aligned}\quad (12)$$

where $\boldsymbol{\nu}(t) \in \mathbb{R}^p$ represents the net output control effort produced by the system's q components.

Inverting the dynamics given in Eq. (12) enables the performance constraint given in Eq. (8) to be expressed as a maximum allowable deviation from a given output control effort profile:

$$\begin{aligned}\mathbf{r} &= B_\nu(\mathbf{x}_c)^{-1} \cdot (\dot{\mathbf{x}}_c - A(\mathbf{x}_c, \mathbf{w})) \\ |\nu_i(k) - r_i(k)| &\leq \tilde{\Delta}_i(k), \quad i \in \{1, 2, \dots, p\}\end{aligned}\quad (13)$$

where $\boldsymbol{\nu}$ and \mathbf{r} are p dimensional vectors that represent an actual and a desired net output force to be exerted by the system at a given time, and $\tilde{\Delta}_i(k)$ specifies a maximum acceptable error between the i^{th} dimensions of $\boldsymbol{\nu}$ and \mathbf{r} .

Because $B_\nu(\mathbf{x})$ has full column rank, a system response, $\dot{\mathbf{x}}$, is uniquely determined by $\boldsymbol{\nu}(t)$; however, if the system is overactuated, then the allocation of load among functionally redundant components may be specified by minimizing the aggregate component damages corresponding to load allocations in the nullspace of $B_u(\mathbf{x})$. Prognostics-based control in terms of component load allocations has been analyzed for an overactuated electro-mechanical actuator and an unmanned ground vehicle in previous publications (Bole et al., 2010, 2011).

2.2.2. Prognostic constraints

Prognostic constraints are typically specified in terms of a lower bound on the failure time of system components. Constraints on the minimum acceptable component failure time may be specified in terms of a maximum acceptable probability that the component will reach 0% health by a given time:

$$P(\gamma_l(t_M) = 0\% | \gamma_l(t_p), \mathbf{w}(t_p), u_l(t_p)) < \alpha_l \quad (14)$$

where t_p is the time at which the fault growth prediction is made and α_l is an upper bound on the probability that component l is failed at time t_M .

If u_l was known over the domain $t = [t_p, t_M]$, then forward induction could be used to evaluate Eq. (14) from Eq. (6). Many publications on the topic of prognostics-based control opt to simplify the prognostics problem by assuming that component loadings will be unvarying over the prediction horizon:

$$u(t) = u(t_p) \quad \forall t \in [t_p, t_M] \quad (15)$$

However, in most cases, time-varying environmental loading conditions and time-varying component health estimates are expected to result in time-varying loadings on a system's

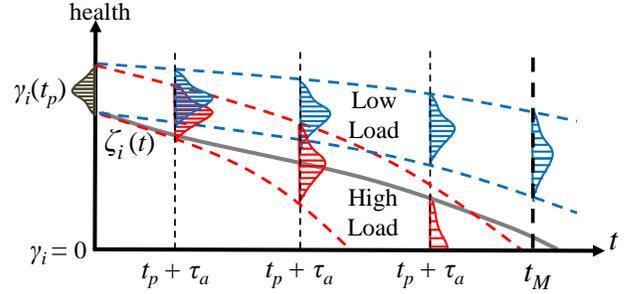


Figure 1. Uncertainty in fault growth predictions for high and low loads at various prognostic horizons

components. In such cases, the production of prognostic estimates with the highest degree of realism will require future component loadings to be modeled as a stochastic process that accounts for the statistics of all parameters affecting component load allocations within the controlled system.

Consider the drawing in Figure 1, the growth of uncertainty in component health estimates at prognostic horizons of increasing length is shown for two potential loading conditions, denoted as 'high load' and 'low load'. As shown in the figure, fault growth estimates at different loading conditions are expected to diverge over time. In cases where prognostic uncertainty becomes very large, due to high uncertainty on predictions of future component loading conditions, or high uncertainty on predictions of component damage as a function of loading profiles, the magnitude of prognostic uncertainty may be managed by limiting the length of the prediction horizon.

The specification of a lower bound on acceptable value at risk (VaR) assessments of system health over the range $t = [t_p, t_M]$ may be used to enforce constraints on component longevity, using fixed horizon prognostic predictions. The VaR of a random variable X at a confidence level ψ is defined as:

$$VaR_\psi(X) = \inf \{x \in \mathbb{R} : P(X < x) > \psi\} \quad (16)$$

A finite horizon prognosis constraint is written as follows:

$$VaR_{\beta_l}(\gamma_l(t_p + \tau)) > \zeta_l(t_p + \tau) \quad (17)$$

where τ specifies a time horizon at which prognostic constraints will be evaluated, ζ_l specifies a lower bound on the depletion of component health over the range $t = [t_p, t_M]$, and β_l defines the maximum acceptable probability that the health of component l is less than ζ_l at time $t_p + \tau$.

An example definition of $\zeta_l(t)$ is given in Figure 1. The specification of an appropriate profile for $\zeta_l(t)$ will be left as a design choice. Some general considerations for the specification of an appropriate profile include:

1. If component health is expected to be monotonically decreasing, then $\zeta_l(t)$ must also be a monotonically de-

creasing function. Additionally, if it is ever the case that $VaR_{\beta_l}(\gamma_l(t_p)) < \zeta_l(t_p + \tau)$, then the prognostic constraint is unsatisfiable.

2. The greater the difference between $VaR_{\beta_l}(\gamma_l(t_p))$ and $\zeta_l(t_p + \tau)$, the greater the control freedom allowed under the prognostic constraint. Online or a priori information could be used to make adjustments to $\zeta_l(t)$ so as to allow for greater control freedom during higher priority time-periods at the expense of potentially incurring greater component health deterioration over those time-periods.

2.3. Fault prediction in terms of relative deviations from nominal system outputs

Consider the existence of a nominal control system, which would adequately control a system in the absence of any component degradation modes. This section introduces a metric to represent the relative deviation between the net output control effort that would have been exerted by the system on its environment if a nominal control law were used, and the net output control effort that is actually exerted by a given prognostics-based control routine at each control time-step.

The proposed metric is defined for each of the system's output degrees of freedom as:

$$\rho_l(k) = \frac{|\nu_l(k)|}{|\nu_l^*(k)|}, \quad l \in \{1, 2, \dots, p\} \quad (18)$$

where ν^* and ν represent the net output control effort that would have been commanded by a nominal control law and the net control effort output that is actually commanded by an implemented control routine at time-index k .

The Markov process defined in Eq. (6) is rewritten in terms of this new metric as:

$$\begin{aligned} p_{s_i, s_j}^l(\rho(k)) &= P(\gamma_l(k+1) = s_j | \gamma_l(k) = s_i, \rho = \rho(k)) \\ &= \int_{\nu_r \in N_r(k)} P(\nu_r(k) | \rho_r = \rho_r(k)) \cdot \end{aligned} \quad (19)$$

$$\begin{aligned} P(\gamma_l(k+1) = s_j | \gamma_l(k) = s_i, \mathbf{u}(k) = H(\nu(k))) & du \\ s_i, s_j \in S, u_l \in U_l(k), r \in \{1, 2, \dots, p\} \end{aligned}$$

where $P(\nu_r(k) | \rho_r)$ can be estimated using available stochastic modeling of the future net output control effort demands on the system, and $H(\nu(k))$ represents a mapping from a net system output force vector to a component loading vector. As described in Section 2.2.1, if no overactuation is present in the system, then component loadings are uniquely specified by a net system output force profile; however, if the system is overactuated, then the nullspace of the component loading domain can be resolved by an optimization routine that seeks to minimize the aggregate loss

of health among functionally redundant degrees of freedom. Note that all modeling of internal and external dynamics that affect component degradation are now described by the probability transition matrix $p_{s_i, s_j}^l(\rho)$. This formulation of fault growth dynamics provides a convenient means for analyzing the prognostics-based control problem, because it directly relates modifications to system output performance to predictions of component degradations.

The performance constraint, defined in terms of allowable net system output control effort in Eq. (13), is now expressed in terms of ρ as:

$$|1 - \rho_i(k)| \leq \bar{\Delta}_i(k), \quad i \in \{1, 2, \dots, p\} \quad (20)$$

where $\bar{\Delta}_i$ defines a constraint on the maximum allowable deviation from a system's nominal control effort output in dimension i at time-index k .

A uniformly quantized state space for ρ_i under the performance constraint is defined as:

$$\begin{aligned} \rho_i(k) \in \Theta(k), \quad \Theta(k) &= [\theta_1, \theta_2, \dots, \theta_m], \\ \theta_1 &= 1 - \bar{\Delta}_i(k), \quad \theta_m = 1 + \bar{\Delta}_i(k) \end{aligned} \quad (21)$$

2.3.1. An example of output control effort regulation

A simple example is considered here to examine the regulation of a system's net control effort output using the performance metric ρ . The example system to be regulated is a linear actuator attached to a simple mass-spring-damper system, defined by a mass m , a spring constant k , and a damping coefficient c :

$$m\ddot{x} = -kx - c\dot{x} + \nu \quad (22)$$

A nominal control law for the system is represented by the following proportional feedback equation:

$$\nu^* = k_p \cdot (x - x_c) \quad (23)$$

where x and x_c represent an actual and a commanded actuator position respectively, ν and ν^* represent the net actuator output force and that commanded by the nominal control law respectively, and k_p is a gain coefficient within the nominal controller. Diagrams of the mass-spring-damper system and the proposed control law are shown in Figure 2. Values for all variables in the controlled mass-spring-damper system are given in Table 1.

In this example, the net output force exerted by the controlled effectors on the environment and the performance metric used to regulate that output are both one dimensional:

$$\nu = \rho\nu^* \quad (24)$$

Figure 3 shows the behavior of the system as ρ is linearly varied over the domain $[0.2, 1]$. It should generally be expected that smaller values of ρ will induce greater errors in tracking

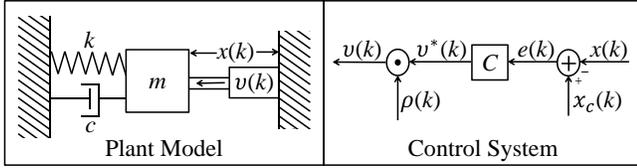


Figure 2. Diagram of mass-spring-damper plant model and control system

Coefficient	Value
m	1
k	2
c	5
k_p	15

Table 1. Coefficients used in mass-spring-damper simulation

a desired output profile, and these induced errors will result in greater net actuator output force demands by the nominal control law. Both of these features are clearly visible in the experimental results given in Figure 3. Another observation that can be made from the results given in Figure 3 is that while $\rho < 1$ will always result in an apparent reduction in the net output control effort that would have been commanded by a nominal control law at a given control time-step, the error dynamics that are induced by lowering ρ will not necessarily result in lowered net system output control effort over a finite window of observation. A marginally increasing trend in the peak-to-peak actuator loads over a cycle is observed as ρ is lowered; this type of behavior illustrates the fact that achieving a good tradeoff between the reduction of component loading and induced errors in trajectory tracking will generally require closed loop regulation of ρ .

3. THE PROGNOSTICS-BASED CONTROL PROBLEM

The prognostics-based control problem can generally be viewed as an optimization problem, in which implemented control routines will select values of ρ_i at each control time-step in an attempt to minimize the risks posed by the application of load to degrading components, while also minimizing any deviation from a system's nominal performance. The problem of specifying appropriate metrics for assessing the risk posed by probabilistic prognostic predictions of future component health deterioration may generally be considered independently from the problem of building prognostic models. Assuming that discrete Markov modeling will provide a sufficiently accurate approximation of a system's fault dynamics, then the Markov process notation described in Section 2 could be used to evaluate any given risk metric for use in any given prognostics-based control method. Future work will address the specification of appropriate risk metrics and the use of this Markov process formulation for deriving and evaluating prognostics-based control policies on sample applications.

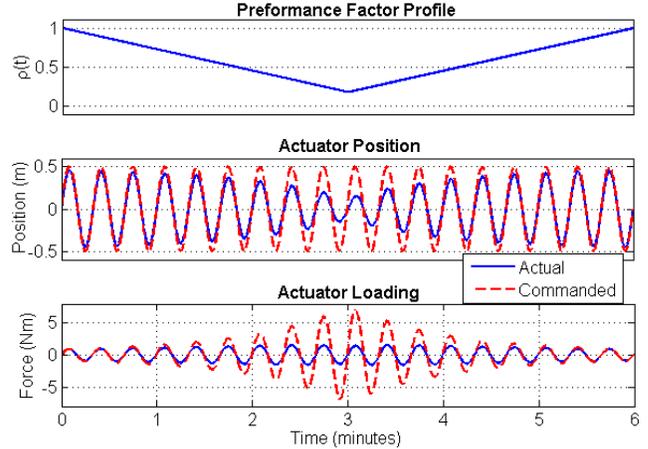


Figure 3. Plots showing position profile tracking and actuator loading dynamics as ρ is varied over the range $[0.2, 1]$

If aversion to the potential degradation of a system's nominal output loading performance and aversion to the potential degradation of component SOH are adequately expressed in terms of an expectation of accumulated state transition costs, then the search for a control policy that optimizes a stochastic system of the form disclosed in Eq. (19) is expressible as a Markov decision process (MDP). MDPs are commonly used to analyze problems involving decision making in the presence of uncertain or stochastic information, and optimizing control policies may be identified using well studied MDP optimization techniques such as backwards induction for finite horizon optimization problems, and linear programming, value iteration, and policy iteration for discounted and average-reward infinite horizon optimization problems. The requirement that fault risk must be expressed as an expectation of accumulated state transition costs over a finite or infinite horizon may seem to be an overly restrictive constraint on the general problem of quantifying risk; however, over the past several decades much has been published on the theory of encoding various forms of risk aversion into the specification of MDP state transition costs (Hernandez & Marcus, 1996; Ruszczyrski, 2009).

4. CONCLUSIONS

This paper introduced a novel Markov process representation of component health dynamics. A metric representing the relative deviation between instantaneous samples of the net output force that would have been enacted by a nominally controlled system, and the net system outputs that are actually enacted by an implemented prognostics-based control routine at each time-step, was used to define the action space of the Markov process. The state space of the proposed Markov process formulation was defined in terms of an abstracted metric representing the relative health remaining in each of the system's components. Operational constraints on minimum

acceptable system output performance and maximum acceptable fault growth risk were formalized, and the mappings necessary to impose those constraints on the domain of feasible control actions available to implemented prognostics-based control routines were specified. Arguments were provided for the potential convenience and robustness of the proposed notation for evaluating prognostics-based controllers.

ACKNOWLEDGMENT

This work was supported by the United States National Aeronautics and Space Administration (NASA) under the Graduate Student Research Program (GSRP). The GSRP provides funding for Brian Bole to perform graduate research concerning the development and implementation of prognostics based reasoning techniques, under the direction of Dr George Vachtsevanos, Professor Emeritus at the Georgia Institute of Technology, and Dr Kai Goebel, Director of the Prognostics Center of Excellence at NASA AMES.

REFERENCES

- Banjevic, D., & Jardine, A. (2006). Calculation of reliability function and remaining useful life for a Markov failure time process. *IMA Journal of Management Mathematics*, 17, 115-130.
- Baruah, P., & Chinnam, R. B. (2005). HMMs for diagnostics and prognostics in machining processes. *International Journal of Production Research*, 43(6), 1275-1293.
- Bole, B., Brown, D. W., Pei, H.-L., Goebel, K., Tang, L., & Vachtsevanos, G. (2010, Oct.). Fault adaptive control of overactuated systems using prognostic estimation. In *Annual conference of the prognostics and health management society*.
- Bole, B., Tang, L., Goebel, K., & Vachtsevanos, G. (2011). Adaptive load-allocation for prognosis-based risk management. In *Annual conference of the prognostics and health management society*.
- Dong, M., & He, D. (2007). Hidden semi-Markov model-based methodology for multi-sensor equipment health diagnosis and prognosis. *European Journal of Operational Research*, 178, 858-878.
- Edwards, D., Orchard, M., Tang, L., Goebel, K., & Vachtsevanos, G. (2010). Impact of input uncertainty on failure prognostic algorithms: Extending the remaining useful life of nonlinear systems. In *Annual conference of the prognostics and health management society*.
- Guidaa, M., & Pulcini, G. (2011). A continuous-state Markov model for age- and state-dependent degradation processes. *Structural Safety*, 33(6), 354-366.
- Hauriea, A., & Moresino, F. (2006). A stochastic control model of economic growth with environmental disaster prevention. *Automatica*, 42(8), 1417-1428.
- Hernandez, D., & Marcus, S. (1996). Risk sensitive control of Markov processes in countable state space. *Systems & Control Letters*, 29, 147-155.
- Orchard, M., Kacprzyński, G., Goebel, K., Saha, B., & Vachtsevanos, G. (2008). Advances in uncertainty representation and management for particle filtering applied to prognostics. In *Annual conference of the prognostics and health management society*.
- Parlara, M., Wang, Y., & Gerchak, Y. (1995). A periodic review inventory model with Markovian supply availability. *International Journal of Production Economics*, 42(2), 131-136.
- Ruszczyski, A. (2009). Risk-averse dynamic programming for Markov decision processes. In *20th international symposium on mathematical programming*.
- Saha, B., Goebel, K., Poll, S., & Christophersen, J. (2009). Prognostics methods for battery health monitoring using a Bayesian framework. *IEEE Transactions on Instrumentation and Measurement*, 58(2), 291-296.
- Sankararaman, S., Ling, Y., Shantz, C., & Mahadevan, S. (2009). Uncertainty quantification in fatigue damage prognosis. In *Annual conference of the prognostics and health management society*.
- Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., & Schwabacher, M. (2008). Metrics for evaluating performance of prognostic techniques. In *Annual conference of the prognostics and health management society*.
- Serfozo, R. F. (1979). An equivalence between continuous and discrete time Markov decision processes. *Operations Research*, 27, 616-620.
- Shetty, P., Mylaraswamy, D., & Ekambaram, T. (2008). A hybrid prognostic model formulation and health estimation of auxiliary power units. *Journal of engineering for gas turbines and power*, 130(2).
- Smilowitz, K., & Madanat, S. (1994). Optimal inspection and repair policies for infrastructure facilities. *Transportation science*, 28, 55-62.
- Sonnenberg, F., & Beck, R. (1993). Markov models in medical decision making. *Medical Decision Making*, 13(4), 322-338.
- Tang, L., Kacprzyński, G. J., Goebel, K., & Vachtsevanos, G. (2009). Methodologies for uncertainty management in prognostics. In *IEEE aerospace conference*.
- Wang, H. S., & Chang, P.-C. (1996). On verifying the first-order Markovian assumption for a Rayleigh fading channel model. *IEEE Transactions on Vehicular Technology*, 45(2), 353-357.
- Wang, P., Youn, B., & Hu, C. (2012). A generic probabilistic framework for structural health prognostic and uncertainty management. *Mechanical Systems and Signal Processing*, 28, 622-637.
- Zhang, Y., & Jiang, J. (2008). Bibliographical review on reconfigurable fault-tolerant control systems. *Annual Reviews in Control*, 32(2), 229-252.

BIOGRAPHIES



Brian M. Bole graduated from the FSU-FAMU School of Engineering in 2008 with a B.S. in Electrical and Computer Engineering and a B.S. in Applied Math. Brian received a M.S. degree in Electrical Engineering from the Georgia Institute of Technology in 2011, and he is currently pursuing a Ph.D. Brian's research interests

include: analysis of stochastic processes, risk analysis, and optimization of stochastic systems. Brian is currently investigating the use of risk management and stochastic optimization techniques for optimal adaptation of active component load allocations in robotic and aviation applications. In a previous project, Brian worked with the Georgia Tech EcoCar team to develop an energy management controller for optimizing the fuel economy of a charge sustaining hybrid electric vehicle.



Kai Goebel received the degree of Diplom-Ingenieur from the Technische Universität München, Germany in 1990. He received the M.S. and Ph.D. from the University of California at Berkeley in 1993 and 1996, respectively. Dr. Goebel is a senior scientist at NASA Ames Research Center where he leads the Diagnostics & Prognostics groups

in the Intelligent Systems division. In addition, he directs the Prognostics Center of Excellence and he is the Associate Principal Investigator for Prognostics of NASA's Integrated Vehicle Health Management Program. He worked at General

Electric's Corporate Research Center in Niskayuna, NY from 1997 to 2006 as a senior research scientist. He has carried out applied research in the areas of artificial intelligence, soft computing, and information fusion. His research interest lies in advancing these techniques for real time monitoring, diagnostics, and prognostics. He holds ten patents and has published more than 100 papers in the area of systems health management.



George J. Vachtsevanos is a Professor Emeritus of Electrical and Computer Engineering at the Georgia Institute of Technology. He was awarded a B.E.E. degree from the City College of New York in 1962, a M.E.E. degree from New York University in 1963 and the Ph.D. degree in Electrical Engineering from the City University of New

York in 1970. He directs the Intelligent Control Systems laboratory at Georgia Tech where faculty and students are conducting research in intelligent control, neurotechnology and cardiotechnology, fault diagnosis and prognosis of large-scale dynamical systems and control technologies for Unmanned Aerial Vehicles. His work is funded by government agencies and industry. He has published over 240 technical papers and is a senior member of IEEE. Dr. Vachtsevanos was awarded the IEEE Control Systems Magazine Outstanding Paper Award for the years 2002-2003 (with L. Wills and B. Heck). He was also awarded the 2002-2003 Georgia Tech School of Electrical and Computer Engineering Distinguished Professor Award and the 2003-2004 Georgia Institute of Technology Outstanding Interdisciplinary Activities Award.