Distributed Risk Governance as an Online Non-Parametric Reinforcement Learning Problem

John Benjamin Cassel

Introduction

The purpose of this paper is to introduce the possibility of moderately realistic, but still generic, models for early-stage distributed risk governance that can be analyzed with the methods of machine learning and decision theoretic planning. This paper extracts and refines the core machine learning models of a larger project aimed at a more general audience [Cassel, 2011].

Risk governance is a difficult challenge. Risk governance concerns itself with all aspects of engaging risks [Renn, 2008], including

- **Risk Assessment** which takes into account all assessments of all risks posed to all involved, from the physical risks to workers and the surrounding population, to the financial risks to owners and backers.
- **Risk Perception** which attends to the perceptions of those stakeholders, both of the risks and of other stakeholders, and the effects of these perceptions on stakeholder preferences and actions.
- **Risk Management** which engages in monitoring risk conditions, controlling risk levels, and maintaining risk response capabilities, thus successfully executing any necessary interventions.
- **Risk Communication** which understands and participates in technical, legal, financial, and cultural channels of communication about risk.
- **Risk Participation** which engages and incorporates all stakeholders in the deliberation of risk.

Distributed risk governance is a sub-domain of risk governance activities when the underlying risk is either not localized or not yet known to be localized. When instituting a policy, whether regulating an activity, creating risk-sharing initiatives (such as moving an activity into insurability), or prohibiting a particular activity, one simultaneously affects broad regions and multiple industries. Further, those subject to the results of these broad actions will often interact with those not subject to them (for example, manufacturers in other nations are not subject to the same body of environmental regulations). Unlike the construction of a given facility, it is presumed our interaction with stakeholders is limited to a sample, and thus educational and participatory initiatives cannot have any substantial coverage except at great cost. Therefore, anticipating and accommodating the initial reaction to proposals and communications is critical to informing early design stages of distributed risk governance.

In order to capture this initial reaction, we will want to sample the population of potential stakeholders to determine both their concerns and knowledge about the subjects. One particular challenge is that the stakeholder's knowledge about the domain, including the physical behavior of the risks being addressed, as well as the composition and concerns of other stakeholders, may not be accurate, and further that this inaccuracy can not necessarily be determined within the confines of a domain-general modeling process.

Therefore, the overall objective of stakeholder modeling is to provide a timely overall strategic assessment of the risk situation and its game-theoretic elements, while correctly estimating the risk of its own remaining faults. This assessment should forecast projected stakeholder actions in a way that allows one to either affirm that those actions are in line with their strategic interests or find a better course of action. This assessment may be undertaken with respect to either all stakeholders or a specific subset.

The formulation of this work is draws from classic decision problems and reinforcement learning approaches [LaValle, 2006] [Sutton and Barto, 1998] [Bertsekas and Tsitsiklis, 1996]. However, this formulation does not take as a given most[1] of the state space, criteria, po-

[1] There must be some initial prompt question that will broadly

tential observations, other agents, actions, or rewards as givens, but learns even these primitives through sampling. However, it is assumed that all of these elements are present, providing an inductive constraint that allows these parameters to be constructed non-parametrically, using the approach that Kemp et al. [2010] applies to causal learning. Later, we will see that other aspects of the problem suggest a fit to the Bayesian non-parametric approach.

Let us start by elaborating some key characteristics of the model:

- **Online** Stakeholders will continue to undertake actions, changing the current situation as well as the perception of other stakeholders.
- **State-neutral samples** Sampling from the stakeholders can be modeled as having no effect on the overall population, using open-ended protocols that attempt to evade issues of framing, and also from the assumption that the sample does not prompt a large-scale communication to other potential stakeholders.
- **Samples are time dependent** Making queries is modeled as taking time, usually the setup costs in engaging stakeholders. This time may very by stakeholder or stakeholder category. This time dependency means that events or actions may change the situation before model-informed strategic advice can be provided.
- **Samples have a cost** Making a sample is modeled as a cost, namely the amount necessary to compensate the querying researcher, and possibly also compensating the queried stakeholder, which can be stakeholder dependent.

The Sampling Process

Each sample is taken from the viewpoint of a particular stakeholder in the overall set of stakeholders, $sk \in Sk$. We will say that each stakeholder is drawn from a stakeholder mixture model[2], $\overline{Sk}$. What this means is that a random process induces a partitions over the stakeholders into categories, and then each member of that category generates features according to its own distribution. This realistically models a set of independently undertaken interviews, as even individuals with similar knowledge and concerns for all matters of substance will differ on the details they provide in any particular engagement. In particular, due to the way we have constructed the problem, we can assume that there are an infinite number of potential stakeholder categories, but many of them are rarer than others. If we can arrange our queries so that they are exchangeable, which is to say so that it does not matter in which order we make our queries, then as we query we expect that we will continue to discover new categories, but progressively more rarely, and generally will find that new stakeholders fit into the categories in the proportions that we have observed before. Thus, one reasonable model for the stakeholder category discovery process is the Dirichlet process (also called the Chinese Restaurant Process or CRP) with a concentration parameter γ. So, we say that given a particular γ, stakeholder categories are distributed according to the CRP:

$$P(\overline{Sk}), \gamma_{Sk} \sim CRP(\gamma_{Sk})$$

Of course, we know that there are not actually an infinite number of stakeholders. Just as a linked-list is a data structure that in theory models an infinite amount of data, but generally works nicely in the practical limits of computer memory, so here we do not take theoretically infinite capacity, but merely use it to flexibly fit the data we do choose to gather [Jordan, 2010].

The Contents of a Given Sample

Each stakeholder will report on both the dynamics of the phenomena in a positive sense as well as the impacts across multiple criteria in a normative sense. We will say that these states of affairs are what bring about benefits and harms, and therefore the agents experiencing those effects will seek to act to cause events that transition to more desirable states of affairs.

Criteria Each stakeholder has some number of criteria ($c \in C$) to which they respond and are motivated by. These criteria are incommensurate, such that any tradeoff between them leads to regret. Each criteria has a preferred optimization, whether to maximize, minimize, or to pursue a certain value. Every criteria also has a margin of indifference, ϵ_c, for which every stakeholder is indifferent to changes between, and a discount parameter α_c, which specifies how much less a given benefit or loss should be taken into account if it occurs later in time.

probe the subject being discussed.

[2]Following the notation of Kemp et al. [2010], we use an overline ($\overline{x}$) to indicate a category model

The Structure of States-of-affairs The overall state-of-affairs at any given point in time, $x \in X$, is only partially provided in any sample, and will consist of some set of fragments called structures, $s \in S$, such that $x \subset S$. This paper will leave the choice of how to represent structures open, although Cassel [2011] provides an open and expressive but not unduly complicated choice. Often, many structures will share sufficient 'structure' as to have similar consequences, so that it is handy instead work with a structural predicate $sp \in Sp$. Without a loss of generality, we can specify the time associated with a given structure or state-of-affairs, by $s[t]$ or $x[t]$, respectively. We will also say that the given set of structures occurring at a given time in a simulation is $S_{sim}[t]$, that the time a sample is taken is 0, and that a sequence of states in time starting from a given point is $\overrightarrow{S[t]}$, such that all the structure that might occur in a given simulation run is $\overrightarrow{S_{sim}[0]}$, and that this notation for indexing time will be in place throughout.

When asked about a given condition, stakeholders may or may not describe the same state-of-affairs, but would reasonably may often answer similarly to others asked. If asked in an exchangeable way, we should expect that they answer in similar proportions to what was seen before, with diminishing returns in discovering new answers, such that the elicited state-of-affairs could reasonably be modeled as a CRP.

$$P(\overline{X}), \gamma_x \sim CRP(\gamma_x)$$

However, we should also expect variation, given the description of a particular state-of-affairs, in what the structures the stakeholder chooses to describe. We might expect that some features are more salient, while others are less so, given the infinite number of possible facts that hold at any given time. Therefore, we expect to see certain combinations of features, in roughly similar percentage over time, with diminishing returns in terms of discovering new features. If samples are taken exchangeably, then this corresponds well to the Beta Process, which is also called the Indian Buffet Process (IBP) [Griffiths and Ghahramani, 2005] [Thibaux and Jordan, 2007] which also takes a concentration parameter γ, so we might model structures as being distributed as IBP within a given state-of-affairs.

$$P(\overline{S}|\overline{X}), \gamma_{X,S} \sim IBP(\gamma_{X,S})$$

In general, we will want to talk about a particular set of elements occurring together, in the same way that structures together form a state-of-affairs or that a set of structural predicates might form an equivalence class over states-of-affairs. In a given sample, key elements will be forgotten, while superfluous elements will be recalled. Given some arbitrary element of this model that occurs in a set, $elem \in Elem$, let $\Phi(\overline{Elem})$ be the underlying generator of that set, leading to various reports distributed according to the IBP. For example, we can say that $\Phi(\overline{S}) = \overline{X}$ and that $\gamma_{\Phi(S,S)} = \gamma_{(X,S)}$.

$$P(\overline{S}|\Phi(\overline{S})), \gamma_{\Phi(S,S)} \sim IBP(\gamma_{\Phi(S,S)})$$

Of course, particular structures being the case make other structures certain, or impossible, or more or less likely, or in other words structures depend on each other. In general, dependency is something we should expect to see in a number of different places, so let's establish it generally now. Dependencies are when some set of elements determine the distribution over some other set of elements. Formally, we can describe this as $d \in D, D : \mathcal{P}(Dt) \rightarrow \Omega(\mathcal{P}(Dt))$, where $\mathcal{P}$ indicates a power set and $dt \in Dt$ is the set of dependency terms, such that $S \subset Dt$. However, structures are only one kind of dependent term, and usually it is only sensible to declare dependencies between similar items, so let us say, without loss of generality, that D_S is the set of dependencies restricted to structures as terms.

The Impact of Current Conditions Each stakeholder is impacted by a different set of goods and harms in different ways. We represent the total impact of any given structure, to any particular stakeholder, for any particular criteria, as a reward function yielding a real number $R : S \times Sk \times C \rightarrow \Re$. Impacts can also be specified relative to the scale of the weights in the structure, which implies that impacts can also be supplied qualitatively.

A given observer will report some subset of the overall impacts: $R_{ob} \subset R$. It is important to note that each observer does not only report their own overall impact, nor is each observer's reported impact taken as the final word.

In this case, we are interested in the likelihood a given set of circumstances has a relation to, or *indicates*, a given criteria for a particular stakeholder ($P_{\mathbf{1}}$), for which we will take a Beta prior, and if so, what the the distribution of that impact is. In other words: does a stakeholder experience an impact relative to a concern in a particular situation? We use a reward indicator function to indicate that their might be.

$$P_{\mathbf{1}}(C|\overline{S}, \overline{Sk}), \gamma_{\mathbf{1}_R} \sim \mathrm{B}(\gamma_{\mathbf{1}_R}, \gamma_{\mathbf{1}_R})$$

Given that an impact exists, what makes sense as a default distribution of impacts? This is a matter is unresolved here and which could appropriately draw from other research. However, let us attempt to reason about

it. In a given situation, benefits and harms are not independent of each other. Often, a favorable situation allows one to effectively engage in other pursuits, improving the situation further. Similarly, when things are going wrong, these effects do not necessarily sum up, but can influence each other, multiplying their consequences. As a result, we might wish to distribute the assessment of a given criteria across situations as log-normal, scaled in some criteria-specific way, σ_C.

$$P(value(R)_C|\overline{S},\overline{Sk},\mathbf{1}(C,\overline{S},\overline{Sk})),\sigma_C \sim Log\mathcal{N}(0,\sigma_C)$$

The same state can cause a stakeholder both benefits and harms, by different criteria, such that both acting to cause the state and acting to avoid it cause regret, reflecting the difficulty of some real-world policy choices.

The Observability of Structures Each of these structures may or may not be directly observable. Should they not be, then there may be corresponding observations that correspond more or less well to that structure being the case, $o \in O \subset S$, where some unknown distribution determines which observations will emerge given those underlying conditions $\theta \in \Theta : \mathcal{P}(S) \rightarrow \Omega(\mathcal{P}(O))$. As different observation functions, or sensings, are provided by different observers, we designate the observation functions elicited from a particular observer as $\Theta_{ob} \subset \Theta$. As in the case of rewards, we are interested in the likelihood a given set of observations has a relation to, or *indicates*, a given set of structures ($P_{\mathbf{1}}$), for which we will again take a Beta prior. However, conditioned on that relation existing, we are also interested in what the distribution of that relationship is ($P_{\mathbf{R}}$), for which we take a uniform prior.

$$P_{\mathbf{1}}(\Phi(\overline{O})|\Phi(\overline{S})),\gamma_{\mathbf{1}_\Theta} \sim \mathrm{B}(\gamma_{\mathbf{1}_\Theta},\gamma_{\mathbf{1}_\Theta})$$

$$P_{\mathbf{R}}(\Phi(\overline{O})|\Phi(\overline{S}),\mathbf{1}(\Phi(\overline{O})|\Phi(\overline{S}))) \sim U(0,1)$$

From here on, for ease of reading, we will adopt the notation of $\mathbf{1}(\cdot)$ to mean that the indicating condition has been satisfied, such that we can write the above line as:

$$P_{\mathbf{R}}(\Phi(\overline{O})|\Phi(\overline{S}),\mathbf{1}(\cdot)) \sim U(0,1)$$

The Effects of Events Given that we understand how to represent a given state, let us now look into understanding how a state can change. Events change one state of affairs into another over a period of time according to some probability distribution, such that $E : X \rightarrow \Omega(X \times T)$. This is accomplished in terms of its structures, such that each event only changes some portions of some of the structures, leaving the rest unchanged. So, we can say that each event will change structures in the state that match some structure expressions into structures that match other structural predicates, or $E : \mathcal{P}(Sp) \rightarrow \Omega(\mathcal{P}(Sp) \times T)$. This implies that, for all the states-of-affairs that might be affected by the event in the same way, and all of the substructures which satisfy the expressions in the same way, we only need to specify a given event once. A given observer will report some subset of events, which we designate as $E_{ob} \subset E$. For a given event ($e \in E$), let us describe the set of structural predicates forming the precondition as $Pre(e) \subset \mathcal{P}(Sp)$, and postconditions as $Post(e) \subset \mathcal{P}(Sp)$. Symmetrically, we can talk about the subsets of events that have a given expression ($sp \in Sp$) as a precondition $PreSet(sp) \in E$ or postcondition $PostSet(sp) \in E$. Let $p_{e,t}$ be the distribution of the duration of event e, given that it does occur.

Given a sampled set of events and a sampled state of affairs, we consider what states of affairs are indicated as a result, and if indicated, a relationship distribution.

$$P_{\mathbf{1}}(\overline{X}|\Phi(\overline{E}),\overline{X}),\gamma_{\mathbf{1}_E} \sim \mathrm{B}(\gamma_{\mathbf{1}_E},\gamma_{\mathbf{1}_E})$$

$$P_{\mathbf{R}}(\overline{X}|\Phi(\overline{E}),\overline{X},\mathbf{1}(\cdot)) \sim U(0,1)$$

Events are dependent upon other events, such that D_E is the set of dependencies between them, which we also take to have both an indicator distribution, and if indicated, a relationship distribution.

$$P_{\mathbf{1}}(\Phi(\overline{E})|\Phi(\overline{E})),\gamma_{\mathbf{1}_D} \sim \mathrm{B}(\gamma_{\mathbf{1}_D},\gamma_{\mathbf{1}_D})$$

$$P_{\mathbf{R}}(\Phi(\overline{E})|\Phi(\overline{E}),\mathbf{1}(\cdot)) \sim U(0,1)$$

The Nature of Stakeholders Stakeholders themselves have some current condition, and that condition is part of the overall state-of-affairs. For that reason, it makes sense to think of stakeholders as structures, $Sk \subset S$. This implies that the same stakeholder can experience different rewards and harms based upon their current condition. Although the ability of events to change desires or preferences is not typically explicitly represented in computational planning approaches, to do so is the basic operation of analytical sociology [Hedstrom, 2005], and is essential to these models being sociologically plausible. There are a number of ways the stake of a stakeholder can change. They may be persuaded, or the environment of the stakeholder can change, in response to which the stakeholder adapts their preferences, in a kind of cognitive dissonance. The preferences of a stakeholder may also change due to changes in their material stakes; for example, a farmer who has had their farm repossessed by a bank would certainly have less reason to care about agricultural policies.

The Actions of Stakeholders Stakeholders do not

merely observe their environment but also participate in it. As such, we generally designate undertaking actions as another kind of structure $a \in A \subset S$. Stakeholders will often anticipate that they or others could respond particular ways in a given situation, or in other words that they will undertake a particular policy. We model these expected policies as saying that if the current conditions match some structural expressions, then particular stakeholders will undertake actions with some distribution, or more formally $\pi \in \Pi, \Pi : X \times Sk \to \Omega(A)$. It will often be convenient to specify the set of actions that apply through the specification of a basic action and particular structures identified by a structural predicate to which that action is applied, such as 'drink' and 'from the cup of coffee I just poured'.

Given that we can talk about states-of-affairs, we would like to be able to talk about which actions are possible and relevant for stakeholders in those states-of-affairs, no matter how likely they are. The policy indicator function indicates whether a given action could possibly be taken, no matter how likely it is.

$$P_{\mathbf{1}}(\overline{A}|\overline{X}, \overline{Sk}, \gamma_{\mathbf{1}_\Pi}) \sim \mathrm{B}(\gamma_{\mathbf{1}_\Pi}, \gamma_{\mathbf{1}_\Pi})$$

If a given action is possible, how likely is to to be undertaken? We would say that it is likely that if the observer knows how to specify it as a possibility, they also likely have a suspicion of whether or not it will be undertaken. For that reason, we can say that the distribution is almost uniform, but not quite, and thus γ_Π is less than, but close to, one.

$$P_{\mathbf{R}}(\overline{A}|\overline{X}, \overline{Sk}, \mathbf{1}(\cdot)), \gamma_\Pi \sim \mathrm{B}(\gamma_\Pi, \gamma_\Pi)$$

Of course, as established before, actions do not stand alone, but they are anticipated to have their compliments and substitutes, so there might be a dependence between sets of policies, where were need to specify both a likelihood of the dependence (or the indicator) and the likelihood given the dependence.

$$P_{\mathbf{1}}(\Phi(\overline{\Pi})|\Phi(\overline{\Pi})), \gamma_{\mathbf{1}_{D_\Pi}} \sim \mathrm{B}(\gamma_{\mathbf{1}_{D_\Pi}}, \gamma_{\mathbf{1}_{D_\Pi}})$$

$$P_{\mathbf{R}}(\Phi(\overline{\Pi})|\Phi(\overline{\Pi}), \mathbf{1}(\cdot)) \sim U(0,1)$$

In this framework, a particular combination of actions is an event. Given a categorical state-of-affairs and some combination of actions, we can ask if another state-of-affairs is the result of that event, and if so, how likely that is:

$$P_{\mathbf{1}}(\overline{X}|\Phi(\overline{\Pi}), \overline{X}), \gamma_{\mathbf{1}_E} \sim \mathrm{B}(\gamma_{\mathbf{1}_E}, \gamma_{\mathbf{1}_E})$$

$$P_{\mathbf{R}}(\overline{D}(\overline{X}|\Phi(\overline{\Pi}), \overline{X}, \mathbf{1}(\cdot))) \sim U(0,1)$$

Knowledge Considerations Between Stakeholders Finally, a category of deferences means that over any set of model variables ($V = X \cup \Pi \cup E \cup D \cup S \cup R \cup \Theta \cup Sk \cup O$) a deference may be given to a particular category of stakeholders. This deference represents either the trust or suspicion a stakeholder has about the knowledge of another stakeholder. These deferences are represented as a likelihood, where zero is complete distrust, one is complete trust, and one-half represents no opinion. Altogether, $df \in Df, Df : V \times Sk \times Sk \to [0,1]$, where $[0,1]$ is set of real numbers from zero to one, inclusive. We parameterize the model for model elements using the same Beta and Uniform distributions as used throughout.

$$P_{\mathbf{1}}(\Phi(\overline{V})|\Phi(\overline{Sk})), \gamma_{\mathbf{1}_{Df}} \sim \mathrm{B}(\gamma_{\mathbf{1}_{Df}}, \gamma_{\mathbf{1}_{Df}})$$

$$P_{\mathbf{R}}(\Phi(\overline{V})|\Phi(\overline{Sk}), \mathbf{1}(\cdot)) \sim U(0,1)$$

Overall, each sample provides a particular account of the beliefs of a given stakeholder, including the states of affairs, how events affect these states, the kinds of impacts stakeholders may experience as a result of these changes, the kinds of actions stakeholders may take to change the state of affairs, and how knowledge about these matters is perceived to be distributed. Together, these structural elements represent an object-level causal model of multiple impacts, M (such that $M = (S, X, Sk, R, C, E, D, A, O, \Theta, \Pi, Df)$). We can see that M is a causal model (in the sense of Pearl [2000]), with S playing the role of variables, and E, Π, Θ, and Df playing the role of functions. What we have actually sample these elements from is $\overline{M}$, the distribution generating the actual model at work.

Parameterization and Policy Considerations

This model has left open a number of free parameters as priors. Some of them have intuitive settings, but others are directly tied to the policy questions which motivate this problem. What are we to suppose for the expected rates for discovering new stakeholders (γ_{Sk}), states-of-affairs (γ_x), criteria[3], and actions[4]? Is it not the case that, if we take unknown harms to unknown stakeholders, in unknown conditions, that we must, as a matter of precaution, assume that these parameters are very large and that, as we appear to hit diminishing returns we are merely unlucky and need to persevere?

[3] The discovery rate of criteria can be analyzed as an IBP mixture over stakeholders

[4] The action discovery rate can be analyzed in the same way as criteria

Or is it instead the case that, in representing the interests of any significant group of the general public, we merely have to capture the most generally held values and stakes, as well as the conditions and actions that could possibly disrupt them, and that we must let marginal concerns be marginal? Whether one pursues diversity or representativeness is a topic of debate in the framing of deliberation and its purposes [Renn, 2008], so the question is very likely insoluble. However, framing this choice as picking between discovery rates may allow machine learning to provide some policy guidance:

1. At all times following the initial sampling, we can assess the rate at which the number of these factors has been growing. Although it may be made necessary by process constraints to assume a rate to gage the number of initial participants, there is no reason not to reassess the most likely rate of discovery, given some examples. Further, even if the exact sampling rate is not trusted, pre-agreed models provide Bayesian guidelines for the amount of minimum belief change.

2. Some information gathering activities are much less expensive than others. For this reason, generative activities such as brainstorming should initially aim for quantity, while more costly elicitation activities may expect this to be smaller. Machine learning helps by assessing how well what is learned from low-cost, low-engagement activities transfers to what is discovered through samples.

Overall Model Process

Now that we have described the model produced from a given sample, let us return to the overall problem. Over the course of sampling, we develop a mixture model of stakeholder viewpoints that can themselves be sampled and combined in various ways to form agent simulations. A complete simulation run (`simrun`, Algorithm 1) within a given mixture model looks the same whether the samples were undertaken or are generated to simulate the result of later sampling.

By looking at the expected simulated difference in a model provided by a single stakeholder versus a mixture of stakeholders, one can discover mistakes in the policy of a particular stakeholder group due to mistaken perceptions about the circumstances or preferences of other stakeholders. These policy mistakes are opportunities for interventions that substitute an improved policy (π^*) for one that would have been taken by default (π_d). The simplest default policy is the mixture of policies reported by stakeholder members. However, there may be reasons not to trust the self-reported default policy, as stakeholders can change their reported preferences due to persuasion or experience. Further, stakeholders may be motivated to misreport their actions to seem more cooperative than they actually are. A more realistic model may be to take a mixture over the reported policies by that stakeholder as elicited from all stakeholders, weighted by deferences.

However, given the opportunity to generate more samples, and run this augmented agent simulation, means we might overturn the results of the model from our samples so far. If the advantage of that intervention is robust to further samplingeffectively, then the intervention should be undertaken. However, if the advantage of the intervention might be overturned by further samples, then we have two options. If there is sufficient time and to do so is cost effective, more samples can be made as to verify or refute the appropriateness of candidate interventions. On the other hand, if the time needed to make the required samples is beyond the horizon of a potential policy change or establishing it is not

Algorithm 1 *simrun* (Scenario Model Simulation Run)

(Get the initial conditions)
$S_{sim}[0] = sample(\mathcal{P}(S[0]))$
$t = 0$
(Keep going while the possible absolute discounted risk is greater than some small quantity of indifference for any criteria)
while $\exists C, \alpha_c^t \sum_r^{R_c} |value(r)| > \epsilon_c$ **do**
(Evaluate the rewards and losses for each stakeholder)
$R_{sim}[t] = R(S_{sim}[t], Sk_{sim}[t], C)$
(Sample the actions that stakeholders will take)
$\{Sk, A, Se\}[t] = sample(\Pi(S_{sim}[t]))$
(Sample the events that result from the current state and those actions)
$E_{sim}[t] = sample(E(\{Sk, A, Se\}, S_{sim})[t])$
(Based on those events, determine the resulting state-of-affairs)
$S_{sim}[t+1] = E_{sim}(S_{sim})[t]$
$t = t + 1$
end while
(Give the user everything that happened in the simulation)
return $\overrightarrow{S_{sim}[0]}, \overrightarrow{R_{sim}[0]}, \overrightarrow{\{Sk, A, Se\}[0]}, \overrightarrow{E_{sim}[0]}$

cost-effective, we are better off deciding not to make an intervention and slowing the sampling process to a rate more appropriate to monitor the developing the situation instead of filling a known information need. In short, for every intervention, we have an optimal stopping problem for sampling with two terminal actions, act or withdraw.

Fortunately, optimal stopping is a respectably understood problem within the reinforcement learning framework [Tsitsiklis and Roy, 1999], [Bertsekas and Tsitsiklis, 1996], [Roy, 2010]. Let us now undertake to express the problem described so far in forms amenable to that approach. Let us say that $\rho_c(sk, \pi)$ is the expected loss in a given criteria for a given stakeholder if they take a given policy, such that $\rho \in \mathbf{P} : Sk \times \Pi \times C \rightarrow \Re$. This ρ holds with respect to some model, m, but we can also talk about the expected loss over a distribution of multiple models for which the criteria is still salient and a subset of the policy is still viable, or $\overline{\rho}$. Let us say that if we are comparing $\overline{\rho}$ while not making reference to any particular stakeholder or criteria, we are looking at the Pareto difference between them, remembering for the determination of $\overline{\rho}$ we aren't *only* interested in C, the sampled set of criteria, but $\overline{C}$, or the projected generated set of criteria given the number of samples, and likewise for stakeholders and policies.

It is important not to forget the costs of undertaking governance. Also, for each sample, we have a cost function (sc). There is the equivocation of how to compare the sample cost with the amount of stakeholder harm per criteria. Let us say that such a conversion factor is given on a per criteria basis while including the percentage of the addressed stakeholder population to which this harm applies. Let us call these conversion factors the *mandate* of the party applying this analysis. There are also costs to both intervention and non-intervention, but as these can vary in criteria including communications cost, possible credibility loss, and mandate loss, it is better to fully model policy changing actions as a stakeholder activity. Therefore, let us call $\hat{\pi}$ the intervention policy aimed at producing π.

Let us say that $\overline{\delta}_{N,k}(\hat{\pi}_1, \hat{\pi}_2)$ is the expected additional loss incurred within $\overline{\rho}$ given that intervention policy $\hat{\pi}_1$ was undertaken instead of $\hat{\pi}_2$ as evaluated under N real samples and k samples generated from the $\overline{M}$ of N. Also, let us say that $\hat{\pi}[t]^*$ best model-averaged intervention policy that can be undertaken at time t, which is indexed by the amount of time necessary to take t samples. The best expected difference between policies at k projected samples is $(\overline{\delta}^*(k))$ follows naturally as $\overline{\delta}^*(k) = \overline{\delta}_{N,k}(\hat{\pi}[k]^*, \hat{\pi}_d)$.

Therefore, for distinguishing whether we should , sample further, or not intervene, we are presented with the dynamic programming operator T such that

$$(T\overline{\delta})^*(k) = min\{\overline{\delta}^*(k), sc() + \overline{\delta}^*(k+1), 0\}$$

While it is certain that this cannot be computed optimally, there are no known obstacles to approximation approaches being applicable.

Future Work

As the primary purpose of this paper is to introduce risk governance as a problem that can be subject to approaches from machine learning, this work invites a broad variety of improvements, implementations, and applications. As such, it is getting ahead of matters to anticipate the core issues.

However, there is one issue so clearly central that it must be highlighted. The probabilistic process we've described requires , the property that samples must have been able to be taken in any order. As these models sample events as they are ongoing, the occurrence of any event separates the samples before and after. As a result, we need a way to determine under which events other events remain exchangeable, and where not exchangeable, how those processes should be modeled. Another constraint on exchangeability is if the samples are actually exchangeable or if the participants from whom the samples are elicited are communicating. However, unlike the first issue, participant communication can be addressed through the elicitation process methodology.

Conclusion

This paper has provided a non-parametric reinforcement learning model for risk governance. As a result of such a model, we can examine the potential loss caused by an inappropriate level of information gathering being undertaken before undertaking an intervention. We can also use such a model to answer questions about, given the information we have, what kinds of stakeholder populations we have in front of us, how those populations might act in a game-theoretic way, and how yet these populations may behave differently given what we don't know. The presence of such models allows us to engage in risk analysis skeptically, aware of at least some of the risks caused by our level of knowledge.

References

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Ma.

Cassel, J. B. (2011). Addressing risk governance deficits through scenario modeling practices. Master's thesis, OCAD University. Available from http://john-benjamin-cassel.com/FinalProject.pdf.

Griffiths, T. and Ghahramani, Z. (2005). Infinite latent feature models and the indian buffet process. *Advances in Neural Information Processing Systems*, 18(17).

Hedstrom, P. (2005). *Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge University Press, Cambridge.

Jordan, M. I. (2010). Bayesian nonparametric learning: Expressive priors for intelligent systems. In *Heuristics, Probability, and Causality: A Tribute to Judea Pearl*, chapter 10. College Publications, http://www.collegepublications.co.uk.

Kemp, C., Goodman, N. D., and Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, 34 (7):1185–1243.

LaValle, S. (2006). *Planning Algorithms*. Cambridge University Press, New York.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York.

Renn, O. (2008). *Risk Governance: Coping with Uncertainty in a Complex World*. Earthscan, London.

Roy, B. V. (2010). On regression-based stopping times. *Discrete Event Dynamic Systems*, 20:307–324.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, Ma.

Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the indian buffet process. In Meila, M. and Shen, X., editors, *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, volume 11, Madison, WI. Omnipress.

Tsitsiklis, J. N. and Roy, B. V. (1999). Optimal stopping of markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. *IEEE Transactions on Automatic Control*, 44:1840–1851.

