Is Randomization Necessary?

Mathias Winther Madsen

August 30, 2016

This notes presents and explains the seeming contradiction between the following two observations:

1. the expected reward produced by a random mixture can never be strictly larger than the reward produced by every single mixture component;[1]

2. there are certain decision problems in which performing random actions works strictly better than always performing the same action.[2]

The paradox is explained by noting that the stochastic policies involved in the second observation are, in fact, not freely chosen, but elements from a highly restricted set of distributions. The last section of the note discusses what this means in practice for reinforcement learning.

1 The Sufficiency of Deterministic Policies

The simplest decision problem an agent can be faced with is one in which it chooses an action $a \in \mathcal{A}$ and is paid a reward of $u(a)$. The action a could, for instance, be a deterministic mapping from camera images to motor torques. The payoff u could be an expected performance measure, averaging out the randomness of the external world.

This simple decision problem has a stochastic extension. In the extended form of the problem, the actions are no longer $\mathcal{A}$, but the set of probability distributions over $\mathcal{A}$. When the agent chooses the stochastic action $A \sim P$, it is paid a reward of

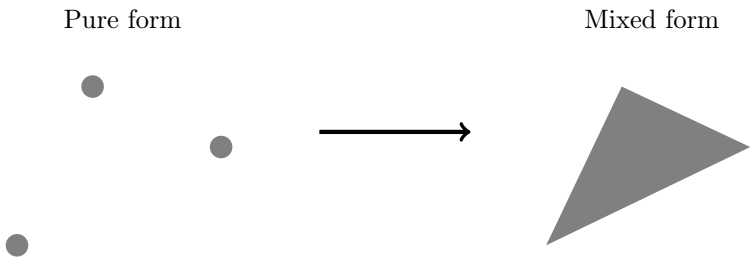
$$\bar{u} \;=\; E_P\left[u(A)\right].$$

We call this stochastic problem the "mixed" form of the game, in constrast to the "pure" deterministic problem. We call the options available to the mixed problem "strategies" instead of "actions."[3]

[1] Blackwell and Girschick: *Theory of Games and Statistical Decisions* (Wiley, 1954), §4.3.

[2] Singh, Jaakkola, and Jordan: "Learning Without State-Estimation in Partially Observable Markovian Decision Processes" (*Proceedings of the ICML-94*, 1994), §3.1.

[3] Blackwell and Girschick (1954), §1.8.



A key fact about mixed problems is that randomization doesn't help:

Theorem 1. *If a mixed-form problem has an optimal strategy, then it has an optimal and deterministic strategy.*

This theorem holds because the expectation of the random variable $u(A)$ cannot equal $\bar{u}$ without satisfying $u(a^*) \geq \bar{u}$ for at least one a^*. It follows that the deterministic strategy $A^* \sim \delta_{a^*}$ (that places all probability mass on a^*) is at least as good as the random strategy A.

This fact about mixed-form decision problems contrasts with the corresponding situation in competitive games. In a game against a malicious opponent, we might indeed have to choose a randomized strategy over the available actions (say, rock–paper–scissors) in order to prevent the opponent from exploiting the predictability of our actions.

This phenomenon, however, is unique to the worst-case assumption that the world is out to get you. If we assume instead that the world is a blind randomization device (and not a malicious reward-minimizer), then randomization never increases average-case performance.

2 Two Necessarily Stochastic Policies

The previous section presented a theoretical argument for why randomizing never leads to strictly higher average-case payoffs. This section presents two examples that suggest the opposite.

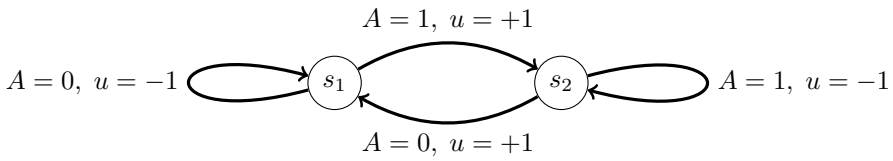
Example 1. Suppose that an agent can choose the bias $\theta \in [0, 1]$ for a bent coin, and that this coin is flipped twice. The agent gets a reward of $u = 1$ if the outcome of the two coin flips are different, and $u = 0$ otherwise.

The solution to the problem is not to make the coin flip deterministic. In fact, the expected utility of the policy with parameter $\theta \in [0, 1]$ is

$$\bar{u} \;=\; 2\theta(1-\theta),$$

since this is the probability of getting two distinct outcomes. This expectation is maximized for $\theta = 1/2$ (where $\bar{u} = 1/2$) and minimized for $\theta = 0$ and $\theta = 1$ (where $\bar{u} = 0$).

Example 2. Consider a partially observable Markov decision problem in which the actions $\mathcal{A} = \{0, 1\}$ produce the following transitions and rewards:



The agent is thus rewarded for changing state — but the action that produces such a change-of-state is not always the same.[4]

We will assume that the agent cannot distinguish between the two states s_1 and s_2. A memoryless policy for this decision problem can therefore be expressed in terms of a single coin flipping parameter $\theta \in [0, 1]$ which expresses how often to perform action $A = 1$ instead of action $A = 0$.

Under a fixed policy parameter $\theta \in [0, 1]$, the expected payoff in the next timestep will be one of the two conditional means

$$\begin{aligned} E[r \,|\, \theta, s_1] &= 1 - 2\theta; \\ E[r \,|\, \theta, s_2] &= 2\theta - 1. \end{aligned}$$

The long-term visiting frequencies of these two states will be θ and $1-\theta$. Hence, the long-term average utility per timestep will be

$$u \;=\; E[r \,|\, \theta] \;=\; -(1 - 2\theta)^2.$$

Again, this expected utility is maximimized at $\theta = 1/2$ (with $u^* = 0$) and minimized at $\theta = 0$ and $\theta = 1$ (with $u = -1$).

Given the result above, how is this possible? Why do the deterministic policies fare strictly worse than the stochastic ones?

3 Problems that Are Not Mixed-Form

The theorem above assumed that the stochastic game was formed by taking the mixed extension of another game with action set $\mathcal{A}$. When this is true, all of the logically possible probability distributions over $\mathcal{A}$ are in the strategy set of the mixed game.

In the two examples above, by contrast, the games were not formed by taking the mixed extension of an already existing game. Instead, we explicitly described the set of probability distributions available to the agent. These sets were, moreover, constructed in such a way that they do not coincide with the complete set of distributions over the sample space $\mathcal{A}$. Some of the distributions are missing.

[4] This example duplicates the graph in Singh, Jaakkola, and Jordan (1994), Figure 1.

Specifically, suppose that the actions available to an agent are sequences,

$$a = (a_1, a_2, \ldots, a_N).$$

If we produced the mixed form of such a decision problem, the resulting strategy set would be the complete set of distributions over $\mathcal{A}^N$. This set contains distributions with all sorts of dependencies and correlations between the sequence elements. Specifically, it also contains all the point-mass distributions that deterministically select a single sequence $(a_1, a_2, \ldots, a_N)$.

On the other hand, in the examples above, we restricted the agent to sampling each element of the sequence independently and from the same (Bernoulli) distribution. The strategy space was thus a space of i.i.d. joint distributions. This excludes all distributions with time dependencies, and almost all the point-mass distributions (although the all-ones and all-zeros distributions are allowed).

In the first example, there are two optimal deterministic strategies,

$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
0	1	0	0

and

$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
0	0	1	0

By the restrictions we placed on the set of strategies, however, the agent was only allowed to select distributions of the form

$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
$(1-\theta)^2$	$\theta(1-\theta)$	$\theta(1-\theta)$	θ^2

for $\theta \in [0,1]$. Neither of the two optimal and deterministic strategies can be expressed in this form.

Similarly, in the partially observable Markov decision problem, one of the many optimal, deterministic strategies is to always pick the alternating sequence

$$0, 1, 0, 1, 0, 1, 0, 1, \ldots$$

However, no sequence of i.i.d. coin flips can deterministically produce this sequence, regardless of the parameter setting.

Again, the set of distributions thus fails to contain one of the point-mass distributions — in fact, it fails to contain all the optimal ones. The set of strategies is not the stochastic extension of the set of sequences.

4 Does Randomness Matter?

The examples in this note show that certain decision problems force us to make a tradeoff between predictability and simplicity.

When our models are expressive to contain the action primitives which actually matter, we do not need to make them stochastic. However, when the model is underexpressive, stochasticity may actually help: it allows us to, once in a while, randomly strike upon very good action combinations that we cannot explicitly formulate in the deterministic fragment of our policy space.

For instance, in the second example discussed above, the agent lived in a Markov world, but was itself memoryless. Its action model was thus misspecified, and randomization could be positively helpful to it. On the other hand, if it had a Markov action policy (that remembered what it did in last round) it could explicitly formulate the optimal deterministic policy

$$0, 1, 0, 1, 0, 1, 0, 1, \ldots$$

and the randomness would be unnecessary.

However, the agent was in fact memoryless. The only two deterministic policies that it can formulate are therefore the ones that sample the sequences

$$0, 0, 0, 0, 0, 0, 0, 0, \ldots$$
$$1, 1, 1, 1, 1, 1, 1, 1, \ldots$$

Neither of these policies are optimal.

Similarly, if the agent had a 2-Markov action policy but lived in a 3-Markov world, randomness might help it. On the other hand, if it had a 2-Markov policy in a 2-Markov world, it would have an optimal, deterministic policy somewhere in its policy space.

In other words, if randomness seems to help you, it is because you are using a misspecified action model. With a more expressive model, you could directly place probability mass on the specific sample values that are responsible for producing the high expected reward under the stochastic policy. However, whether this increase in precision is worth the additional modeling complexity — for instance, the higher dimensionality of the policy space — depends on the specific character of the decision problem.