The Max K-Armed Bandit: A New Model of Exploration Applied to Search Heuristic Selection

Vincent A. Cicirello
Department of Computer Science
Drexel University
Philadelphia, PA 19104
cicirello@cs.drexel.edu

Stephen F. Smith
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
sfs@cs.cmu.edu

Abstract

The multiarmed bandit is often used as an analogy for the tradeoff between exploration and exploitation in search problems. The classic problem involves allocating trials to the arms of a multiarmed slot machine to maximize the expected sum of rewards. We pose a new variation of the multiarmed bandit—the Max K-Armed Bandit—in which trials must be allocated among the arms to maximize the expected best single sample reward of the series of trials. Motivation for the Max K-Armed Bandit is the allocation of restarts among a set of multistart stochastic search algorithms. We present an analysis of this Max K-Armed Bandit showing under certain assumptions that the optimal strategy allocates trials to the observed best arm at a rate increasing double exponentially relative to the other arms. This motivates an exploration strategy that follows a Boltzmann distribution with an exponentially decaying temperature parameter. We compare this exploration policy to policies that allocate trials to the observed best arm at rates faster (and slower) than double exponentially. The results confirm, for two scheduling domains, that the double exponential increase in the rate of allocations to the observed best heuristic outperforms the other approaches.

Introduction

The *K-Armed Bandit* often serves as an analogy for balancing exploration and exploitation in search domains (Berry & Fristedt 1985). The problem is to allocate trials to the arms of a k-armed bandit (i.e., slot machine with k arms, each with a different pay-out distribution) with the goal of maximizing expected total reward. Many have analyzed variations of the bandit problem (e.g., (Agrawal 1995; Auer, Cesa-Bianchi, & Fischer 2002; Auer *et al.* 2002; Berry & Fristedt 1985; Holland 1975)). Others have used bandits as inspiration for, or justification of, exploration strategies—e.g., for genetic algorithms (Holland 1975) and reinforcement learning (Sutton & Barto 1998).

In this paper, a new variation of the multiarmed bandit is posed—the *Max K-Armed Bandit Problem*. The problem, simply stated, is to allocate trials among the k arms so as to maximize the expected best single sample reward. Our motivation is the problem of allocating restarts among multistart stochastic search algorithms to maximize overall search results. Consider an NP-hard combinatorial optimization problem, a stochastic search algorithm that can be biased by a search heuristic, and a set of heuristics which perform differentially on different problem instances. In solving any given problem instance, one would like to dynamically determine and exploit the heuristic that yields the best search performance on this instance. At any point during the search, the goal of future restarts is to find a solution that is better than the current best found. The original multiarmed bandit is concerned with maximizing the expected sum of rewards. However, this does not match the goal in our stochastic search example. In the stochastic search case, we have our current reward (i.e., the best solution found so far) and need to find some reward that is better yet.

Prior research has argued that extreme value theory offers a good model for the distribution of solutions that would be produced across iterations of a heuristic biased stochastic sampling procedure when using the bias of a strong domain heuristic (Cicirello & Smith 2004). Starting from this assumption, we show theoretically that the optimal exploration policy for the Max K-armed Bandit allocates a double exponentially increasing number of trials to the observed best heuristic. We then empirically validate this exploration policy in two complex scheduling domains: (1) weighted tardiness sequencing; and (2) resource constrained project scheduling with time windows.

The K-Armed Bandit: Holland's Analysis

The *k-armed bandit* is a major part of the theoretical underpinning of the genetic algorithm (GA). Holland (1975) uses the k-armed bandit analogy to show that the GA achieves a near-optimal tradeoff of exploration and exploitation.

For the two-armed bandit, the expected reward for arm one is μ_1 with variance σ_1^2 (μ_2 and σ_2^2 for arm two). Furthermore, $\mu_1 \geq \mu_2$, but it is not known which arm is which. The problem is to maximize expected reward for a series of trials. One must determine the optimal tradeoff of exploratory actions (i.e., to discover the payoffs) and exploitation actions (i.e., playing the apparent best). The k-armed bandit is the obvious generalization. Let $R(\mu, \sigma)$ be a reward function that samples a normal distribution with mean μ and standard deviation σ, and let n_i be the number of samples given

the i-th arm. The objective is to allocate the n_i to optimize:

$$\max \sum_{i=1}^{k} n_i R(\mu_i, \sigma_i). \quad (1)$$

If the μ_i are known, then all N trials should be allocated to the arm with the largest μ_i to maximize the expected value of this objective. Without knowledge of the μ_i, it is necessary to perform some exploration to solve the problem.

For the two-armed bandit, Holland showed the optimal policy (to minimize expected loss from trials of the worst arm) allocates n^* trials to the worst arm, and $N - n^*$ to the best arm where in the limit:[1]

$$N - n^* \sim \Theta(\exp(cn^*)), \quad (2)$$

where c is a constant. The trials allocated the observed best arm should increase exponentially relative to the allocation to the observed worst arm. Holland generalized this to the k-arm case, showing the worst-case expected loss for the problem occurs when $\mu_2 = \mu_3 = \ldots = \mu_k$ and $\sigma_2 = \sigma_3 = \ldots = \sigma_k$; and further showing that the best arm should be allocated $N - (k-1)m^*$ trials where N is the total number of trials and where each of the other $k-1$ arms are allocated m^* trials. The optimal number of trials, in the limit, is:

$$N - (k-1)m^* \sim \Theta(\exp(cm^*)). \quad (3)$$

The number of trials allocated to the observed best arm in the optimal allocation should increase exponentially with the number of trials allocated to each of the other $k-1$ arms.

The Max K-Armed Bandit

We now pose a new variation of the multiarmed bandit called the *Max K-Armed Bandit*. In the Max K-Armed Bandit Problem, we are faced with a series of N trials. In any given trial, we can choose any of the k arms. For each of the arms there is an expected payoff according to some probability distribution. The goal is to maximize the value of the *best* single reward received over the N trials. This new objective is to allocate N trials among the arms to optimize:

$$\max \max_{i=1}^{k} \max_{j=1}^{n_i} R_j(D_i), \quad (4)$$

where $R_j(D_i)$ is the reward of the j-th trial of arm i with reward distribution D_i.

In the following subsections, we develop a solution to the Max K-Armed Bandit problem. Under certain assumptions about the distribution of samples of an arm, we show that to maximize the expected max single sample reward over N trials, the number of samples taken from the observed best arm should grow double exponentially in the number of samples taken from the observed second best. We proceed in three steps. First, we make some assumptions about the payoff distributions associated with each arm. Then we consider the special case of two arms. Finally, we generalize this solution to K arms.

[1] See Holland (1975) for complete derivation.

Payoff Distribution Assumptions

To analyze the Max K-Armed Bandit, it is necessary to specify the type of distribution that each of the arms follow. In the classic version of the bandit problem, this is not necessary. Since the classic problem concerns the maximization of the expected sum of rewards, it is sufficient to make assumptions about the means and standard deviations of the arms. In the Max K-Armed Bandit case, we require an expression for the expected max of a series of N trials as a function of N. This necessitates an assumption about the form of the underlying distribution of trials. The extremal types theorem tells us that the distribution of the max of a series of independent and identically distributed trials (as the length of the series grows large) belongs to one of three distribution families independent of the underlying distribution of the trials: the Gumbel, the Fréchet, or the Weibull (Coles 2001). This seems to allow us to carry through with an analysis independent of the form of the distributions of the samples drawn from the arms of the bandit. However, an expression is needed in terms of the length of the series of trials, requiring an assumption on the underlying distribution.

To make an appropriate assumption we consider the target application—allocating restarts among a set of multistart stochastic search heuristics for combinatorial optimization. Cicirello and Smith (2004) argue that a stochastic search procedure that is biased by strong domain heuristics samples from the extreme of the solution quality distribution of the underlying problem space. They showed that such an algorithm generally finds "good" solutions for combinatorial optimization and that "good" solutions are statistically rare in the overall solution space (i.e., extremely low probability of drawing a "good" solution at random). If we randomly sample N solutions, then for large N, the best sample (or maximum element) must follow the extremal types theorem—by definition. The assumption is that the behavior of a stochastic search procedure that is biased by a strong domain heuristic is equivalent to taking the best solution from a sufficiently large series of unbiased random samples. Following extreme value theory, we assume that individual solutions given by the stochastic search are drawn from one of three distribution families: Gumbel, Fréchet, or Weibull (generalized as the Generalized Extreme Value (GEV) distribution).

Since an assumption of the most general GEV distribution prevents a closed form analysis, let us instead assume that each of the arms samples from a type I extreme value distribution (or the Gumbel distribution). This distribution has a cumulative probability of:

$$P(Z \leq z) = G(z) = \exp\left(-\exp\left(-\left(\frac{z-b}{a}\right)\right)\right), \quad (5)$$

where b is the location parameter and a the scale parameter. The probability density function of the Gumbel is:

$$P(Z = z) = \frac{1}{a}\exp\left(-\left(\frac{z-b}{a}\right)\right)\exp\left(-\exp\left(-\left(\frac{z-b}{a}\right)\right)\right). \quad (6)$$

The Max 2-Armed Bandit Case

Theorem 1 *The Two-Armed Double Exponential Sampling Theorem: To optimize the Max 2-Armed Bandit, where the arm samples are drawn from Gumbel distributions, the observed best arm should be sampled at a rate that increases double exponentially relative to the observed second best.*

Let there be two arms, M_1 and M_2, with the rewards of M_i drawn from a Gumbel distribution $G_i(x)$ with location parameter b_i and scale parameter a_i. The mean reward of a single sample of M_i is: $\mu_i = b_i + 0.5772a_i$, where 0.5772 is Euler's number, and the standard deviation is: $\sigma_i = \frac{a_i\pi}{\sqrt{6}}$.

Proof Given that the expected largest sample of a series of trials must be maximized, an expression is needed for the expected value of the maximum of a series of samples. Given N samples $\{X_1, \ldots, X_N\}$ from a distribution, the probability that the maximum of these samples equals x is:

$$P(\max(X_i) = x) = N\,P(X = x)\,P(X \leq x)^{N-1}. \quad (7)$$

With the assumption of samples drawn from a Gumbel distribution, we have:

$$P(\max(X_i) = x) = \frac{N}{a}\exp(-\tfrac{x-b}{a})\exp(-\exp(-\tfrac{x-b}{a}))\ \cdot \exp(-(N-1)\exp(-\tfrac{x-b}{a})) \quad (8)$$

This simplifies to:

$$P(\max(X_i) = x) = \frac{1}{a}\exp(-\tfrac{x-b-a\ln N}{a})\exp(-\exp(-\tfrac{x-b-a\ln N}{a})). \quad (9)$$

From this we see that the distribution of the max of N samples drawn from a Gumbel distribution with location parameter b and scale parameter a is also a Gumbel distribution with location parameter, $b_{\max} = b + a\ln N$ and scale parameter $a_{\max} = a$. Thus the expected max reward of N samples from each of the two arms in the problem is:

$$b_i + 0.5772a_i + a_i \ln N. \quad (10)$$

Consider that M_1 is the better of the two arms in the problem. This necessitates a definition for "better". Let:

$$b_1 + 0.5772a_1 + a_1\ln N > b_2 + 0.5772a_2 + a_2\ln N \quad (11)$$

which implies that $a_1 \geq a_2$. Otherwise, for great enough N this inequality would fail to hold.

In the two-armed problem, where we do not know with certainty which arm is M_1 and which is M_2, the expected max reward if we had access to an omniscient oracle is clearly $b_1 + 0.5772a_1 + a_1\ln N$ —the expected max reward of giving all N trials to the better arm. However, given that we cannot know with certainty which arm is which, some exploration is necessary. Consider that we draw n samples from the observed second best arm, and $N-n$ samples from the observed best arm. Now consider the loss of reward associated with sampling from the second best arm. There are two cases to consider:

1. The observed best is really the best. In this case, the loss comes from giving n less samples to the best arm—with an expected loss equal to: $a_1(\ln N - \ln(N-n))$.
2. The observed best arm is really second best. The loss in this case depends on whether the expected value of giving $N-n$ samples to the second best arm is greater than giving n samples to the best arm. That is, the expected loss is: $\min\{a_1(\ln N - \ln n), (b_1 - b_2) + 0.5772(a_1 - a_2) + a_1\ln N - a_2\ln(N-n)\}$. This form of loss is maximized when the expected value of the max of n samples of the best arm equals that of $N-n$ samples of the second best. This allows us to consider a simplification of the expected loss in this case: $a_1(\ln N - \ln n)$.[2]

Let q be the probability that the observed best arm is really second best. Therefore, $(1-q)$ is the probability that the observed best really is the best. The expected loss of sampling n times from the observed second best and $N-n$ times from the observed best arm, as a function of n is therefore:

$$l(N) = q(a_1(\ln N - \ln n)) + (1-q)(a_1(\ln N - \ln(N-n))). \quad (12)$$

This can be simplified to:

$$l(N) = q(a_1(\ln(N-n) - \ln n)) + a_1(\ln N - \ln(N-n)). \quad (13)$$

To select a value for n that minimizes the expected loss, we need to define q as a function of n. Let M_b be the arm that is perceived as best (i.e., the arm perceived to have the highest expected max single sample reward over a series of N trials) and M_w be the arm that is perceived as second best. The probability q can be stated as the probability that the expected max value of N samples of M_w is greater than the expected max value of N samples of M_b. If we note that the parameters of a Gumbel distribution can be estimated (see (NIST/SEMATECH 2003)) from the data by $\tilde{a} = \frac{s\sqrt{6}}{\pi}$ and $\tilde{b} = \bar{X} - 0.5772\tilde{a}$, where $\bar{X}$ and s are the sample mean and sample standard deviation, then we can define:

$$q(n) = P\left(\begin{array}{l}(\tilde{b}_b + 0.5772\tilde{a}_b + \tilde{a}_b\ln N)\\ -(\tilde{b}_w + 0.5772\tilde{a}_w + \tilde{a}_w\ln N)\\ < 0\end{array}\right) \quad (14)$$

$$= P\left(\begin{array}{l}(\bar{X}_b + \frac{s_b\sqrt{6}}{\pi}\ln N)\\ -(\bar{X}_w + \frac{s_w\sqrt{6}}{\pi}\ln N) < 0\end{array}\right) \quad (15)$$

$$= P\left(\bar{X}_b - \bar{X}_w < (s_w - s_b)\frac{\sqrt{6}}{\pi}\ln N\right). \quad (16)$$

The central limit theorem says that $\bar{X}_b$ approaches a normal distribution with mean μ_b and variance $\frac{\sigma_b^2}{N-n}$. Similarly, $\bar{X}_w$ approaches a normal distribution with mean μ_w and variance $\frac{\sigma_w^2}{n}$. The distribution of $\bar{X}_b - \bar{X}_w$ is the convolution of the distributions $\bar{X}_b$ and $-\bar{X}_w$. The convolution of these distributions is by definition a normal distribution with mean

[2] Alternatively, we could also consider the simplification $(b_1 - b_2) + 0.5772(a_1 - a_2) + a_1\ln N - a_2\ln(N-n)$, but this would unnecessarily complicate the analysis.

$\mu_b - \mu_w$ and variance $\frac{\sigma_b^2}{N-n} + \frac{\sigma_w^2}{n}$. Using an approximation for the tail of a normal distribution, we can define $q(n)$ as:

$$q(n) \lesssim \frac{1}{\sqrt{2\pi}} \frac{\exp(-x^2/2)}{x} \tag{17}$$

where

$$x = \frac{(\mu_b - \mu_w) + \frac{\sqrt{6}}{\pi} \ln(N)(\frac{\sigma_b}{\sqrt{N-n}} - \frac{\sigma_w}{\sqrt{n}})}{\sqrt{\frac{\sigma_b^2}{N-n} + \frac{\sigma_w^2}{n}}}. \tag{18}$$

Given the expressions for $q(n)$ and x, note that $q(n)$ decreases exponentially in n. Using the same simplification made by Holland (1975), note that no matter the value for σ_b, there is a large enough N such that for n close to its optimal value, $\frac{\sigma_b^2}{N-n} \ll \frac{\sigma_w^2}{n}$. This leads to:

$$x \lesssim \frac{(\mu_b - \mu_w)\sqrt{n} - \frac{\sigma_w\sqrt{6}}{\pi} \ln N}{\sigma_w} \tag{19}$$

To select the value of n that will minimize the loss $l(n)$ we begin by taking the derivative of $l(n)$ with respect to n:

$$\frac{dl}{dn} = \begin{array}{l} \frac{dq}{dn}(a_1(\ln(N-n) - \ln n)) \\ -q(n)(\frac{a_1}{N-n} + \frac{a_1}{n}) + \frac{a_1}{N-n} \end{array}, \tag{20}$$

where

$$\frac{dq}{dn} \lesssim -q(n)\frac{x^2+1}{x}\frac{dx}{dn}, \tag{21}$$

and

$$\frac{dx}{dn} \lesssim \frac{\mu_b - \mu_w}{2\sigma_w\sqrt{n}}. \tag{22}$$

The optimal value of n occurs when $\frac{dl}{dn} = 0$ so we can get a bound on the optimal n by solving the following inequality:

$$\begin{aligned} 0 \lesssim\ & \frac{a_1}{N-n} - q(n)\frac{a_1 N}{N-n} \\ & -q(n)\frac{x^2+1}{x}\frac{dx}{dn}(a_1(\ln(N-n) - \ln n)). \end{aligned} \tag{23}$$

We can collect the logarithmic terms on the left to obtain

$$\ln(N-n) - \ln n \lesssim \frac{(1 - q(n)N)x}{(N-n)q(n)\frac{dx}{dn}(x^2+1)} \tag{24}$$

Recalling that $q(n)$ decreases exponentially in n, $(1 - q(n)N)$ rapidly approaches 1. Noting $\frac{x}{x^2+1} \lesssim \frac{1}{x}$, obtain:

$$\ln(N-n) - \ln n \lesssim \frac{1}{(N-n)q(n)\frac{dx}{dn}x} \tag{25}$$

Substituting expressions for $q(n)$ and $\frac{dx}{dn}$ we get:

$$\begin{aligned} &\ln(N-n) - \ln n \lesssim \\ &\frac{\sigma_w\sqrt{8\pi}\sqrt{n}}{(N-n)(\mu_b - \mu_w)} \exp\left(\frac{(\mu_b - \mu_w - \frac{\sigma_w\sqrt{6}}{\pi\sqrt{n}} \ln(N))^2 n}{2\sigma_w^2}\right) \end{aligned} \tag{26}$$

Finally, exponentiate both sides of the inequality, to obtain:

$$\begin{aligned} &N - n \lesssim \\ &\exp\left(\begin{array}{l} \ln(n) + \\ \frac{\sigma_w\sqrt{8\pi}\sqrt{n}}{(N-n)(\mu_b - \mu_w)} \exp\left(\frac{(\mu_b - \mu_w - \frac{\sigma_w\sqrt{6}}{\pi\sqrt{n}} \ln(N))^2 n}{2\sigma_w^2}\right) \end{array}\right) \end{aligned} \tag{27}$$

The question that remains is which term in the exponential dominates the expression. We can take the fraction involving $N - n$ up into the exponential and gain insight into the answer to this question:

$$N - n \lesssim \exp\left(\begin{array}{l} \ln(n) + \\ \exp\left(\begin{array}{l} \frac{(\mu_b - \mu_w - \frac{\sigma_w\sqrt{6}}{\pi\sqrt{n}} \ln(N))^2 n}{2\sigma_w^2} \\ + \ln\left(\frac{\sigma_w\sqrt{8\pi}\sqrt{n}}{(N-n)(\mu_b - \mu_w)}\right) \end{array}\right) \end{array}\right). \tag{28}$$

We must now determine which part of this double exponential dominates as the total number of samples N grows large. Consider the following limits:

$$\lim_{N\to\infty} \frac{\left(\mu_b - \mu_w - \frac{\sigma_w\sqrt{6}}{\pi\sqrt{n}} \ln(N)\right)^2 n}{2\sigma_w^2} = \infty \tag{29}$$

$$\lim_{N\to\infty} \ln\left(\frac{\sigma_w\sqrt{8\pi}\sqrt{n}}{(N-n)(\mu_b - \mu_w)}\right) = -\infty \tag{30}$$

Note that the first expression is dominated by the $(\ln N)^2$ and that within the logarithm of the second expression, the N in the denominator dominates. For large enough N, it is sufficient to consider which of $(\ln N)^2$ and $\ln(1/N)$ dominates. Consider the following:

$$\begin{aligned} &\lim_{N\to\infty}\{(\ln N)^2 + \ln(1/N)\} \\ &\quad = \lim_{N\to\infty}\{(\ln N)^2 + \ln 1 - \ln N\} \\ &\quad = \lim_{N\to\infty}\{\ln N(\ln N - 1)\} \\ &\quad = \infty \end{aligned} \tag{31}$$

Taking this into account and making a few other obvious simplifications, we can arrive at:

$$N - n \sim \Theta(\exp(\exp(cn))) \tag{32}$$

This shows that the number of trials $N - n$ given to the observed best arm should grow double exponentially in n to maximize the expected max single sample reward.

Generalization to the K-Armed Case

Theorem 2 *The Multiarmed Double Exponential Sampling Theorem: To optimize the Max K-Armed Bandit (samples drawn from Gumbel distributions), the observed best arm should be sampled at a rate increasing double exponentially relative to the number of samples given the other $k-1$ arms.*

Proof To make this inductive leap from the result of the two-arm case to the k-arm case, observe the following. The worst case loss in the k-armed case occurs when the $k - 1$

worst arms are identical (as is the case in Holland's analysis of the original k-armed bandit). If these $k-1$ arms are identical then it doesn't matter how we allocate trials among them—the result is equally poor. But, if any of these $k-1$ arms is better than any of the other $k-2$ arms, then we can improve our expected reward by allocating more trials to it. Assume the worst case that the $k-1$ arms are identical. With m^* trials given to each of these $k-1$ worst arms, the analysis of the k-armed case can be considered a special case of the analysis of the two-armed problem. Specifically, we have the observed best arm and a meta-arm comprised of the aggregation of the other $k-1$ arms. The meta-arm is given $n^* = m^*(k-1)$ trials uniformly distributed across the $k-1$ arms. Since the $k-1$ arms are identical in the worst case, the meta-arm behaves identically to the second best arm in the two-arm case. Thus, the number of samples $N - m^*(k-1)$ given the observed best arm should grow double exponentially in $n^* = m^*(k-1)$.

Exploration Strategy

Recall that our goal is to find a good exploration strategy for allocating trials to different heuristics. To design an exploration policy that follows the double exponential sampling theorems, consider Boltzmann exploration (Sutton & Barto 1998). Let the temperature parameter T decay exponentially, choosing heuristic h_i with probability:

$$P(h_i) = \frac{\exp((R_i)/T)}{\sum_{j=1}^{H} \exp((R_j)/T)}. \quad (33)$$

The R_i is some indicator/estimator of the expected max of a series of trials of heuristic h_i. For example, R_i can be an estimator for the expected max for some fixed length series of trials given some distribution assumption. To derive the double exponentially increasing allocation of trials to the observed best arm, the temperature parameter must follow an exponentially decreasing cooling schedule (e.g., $T_j = \exp(-j)$ where j is the trial number). That is, on iteration j choose heuristic h_i with probability:

$$P(h_i|j) = \frac{\exp((R_i)/\exp(-j))}{\sum_{k=1}^{H} \exp((R_k)/\exp(-j))}. \quad (34)$$

Next we present results from NP-hard scheduling domains that contrast the performance of this exploration policy with policies that allocate trials to the observed best heuristic at rates greater (and lesser) than double exponentially.

Weighted Tardiness Scheduling

Problem Formalization: The Weighted Tardiness Scheduling Problem is a sequencing problem. A set of jobs $J = \{j_1, \ldots, j_N\}$ must be sequenced on a single machine. Each of the N jobs j has a weight w_j, duedate d_j, and process time p_j. Preempting a job during processing is not permitted. Only one job at a time can be processed. The objective is to sequence the set of jobs J on a machine to minimize the total weighted tardiness: $T = \sum_{j \in J} w_j T_j = \sum_{j \in J} w_j \max(c_j - d_j, 0)$, where T_j is the tardiness of job j; and c_j, d_j is the completion time and duedate of job j. The completion time of job j is equal to the sum over the process times of all jobs that come before it in the sequence plus that of the job j itself. Specifically, let $\pi(j)$ be the position in the sequence of job j. We can now define c_j as: $c_j = \sum_{i \in J, \pi(i) <= \pi(j)} p_i$.

Table 1: Weighted tardiness: For each number of restarts $[N]$, bold indicates the most best known solutions found.

Algorithm	NB	ARPD	MRPD
D-EXP[400]	94.3	0.12	10.07
EXP[400]	85	0.14	11.28
FASTER[400]	78.7	0.19	13.51
M-DYNA[400]	62	1.66	76.18
D-EXP[800]	100.7	0.12	10.07
EXP[800]	89.3	0.14	11.28
FASTER[800]	83.6	0.17	13.51
M-DYNA[800]	68.3	1.29	74.28
D-EXP[1600]	107.3	0.11	8.47
EXP[1600]	95	0.12	9.75
FASTER[1600]	87.5	0.16	11.28
M-DYNA[1600]	73.3	1.16	71.06

Value-Biased Stochastic Sampling (VBSS): VBSS is an iterative stochastic heuristic search algorithm (Cicirello & Smith 2005). A search heuristic is used to bias a random decision at each decision point. We use VBSS here to generate biased initial configurations for a local search for the weighted tardiness problem known as Multistart Dynasearch (Congram, Potts, & van de Velde 2002). The original Multistart Dynasearch used unbiased initial solutions.

Dispatch Policies as Search Heuristic: Many dispatch policies exist for this problem (Morton & Pentico 1993). A few of the best are used here as candidate search heuristics:

- weighted shortest process time, $\text{WSPT}_i = \frac{w_i}{p_i}$;
- earliest duedate, $\text{EDD}_i = \frac{1}{d_i}$;
- $\text{COVERT}_i(t) = \frac{w_i}{p_i}(1 - \frac{\max(0, d_i - p_i - t)}{k p_i})$, with current time t and parameter k; and
- $\text{R\&M}_i(t) = \frac{w_i}{p_i} \exp(-1 * \frac{\max(0, d_i - p_i - t)}{k \bar{p}})$, with average process time $\bar{p}$.

Experimental Setup: In this experiment, these heuristics are combined across multiple restarts of the dynasearch algorithm. On any given restart, VBSS is used, along with one of these heuristics to construct an initial solution, which is then locally optimized using dynasearch. The following exploration policies are considered: double exponentially increasing rate of allocations to the observed best heuristic (D-Exp); faster than double exponentially increasing allocation rate (Faster); and exponentially increasing allocation rate (Exp). We also compare to multistart dynasearch (M-Dyna) as originally specified by Congram et al.

Results: The results presented here are for the 100 job instances from the benchmark problem set from the OR-Library (Beasley 1998). The set contains 125 instances. Results are shown in Table 1. NB is the number of best known solutions found (no further improvement is made). ARPD (and MRPD) are the average (and maximum) relative percentage deviation from the best known solutions. The results shown are averages of 10 runs for all 125 problem instances.

M-DYNA is the worst of the four variations considered. There is clearly benefit to biasing the initial configurations of the M-DYNA local search, contrary to the untested hypothesis of Congram et al. The trend for any number of iterations considered is that the double exponentially increasing rate of allocations finds the most best known solutions, with smallest percentage deviation from the best knowns. The next best in terms of these criteria is when the observed best heuristic is given an exponentially increasing allocation of trials, followed by the variation with a faster than double exponentially increasing rate of allocations.

Resource Constrained Project Scheduling with Time Windows (RCPSP/max)

Problem Formalization: The RCPSP/max problem is defined as follows. Define $P =< A, \Delta, R >$ as an instance of RCPSP/max. Let A be the set of activities $A = \{a_0, a_1, a_2, \ldots, a_n, a_{n+1}\}$. Activity a_0 is a dummy activity representing the start of the project and a_{n+1} is similarly the project end. Each activity a_j has a fixed duration p_j, a start-time S_j, and a completion-time C_j which satisfy the constraint $S_j + p_j = C_j$. Let Δ be a set of temporal constraints between activity pairs $< a_i, a_j >$ of the form $S_j - S_i \in [T_{i,j}^{\min}, T_{i,j}^{\max}]$. The Δ are generalized precedence relations between activities. The $T_{i,j}^{\min}$ and $T_{i,j}^{\max}$ are minimum and maximum time-lags between the start times of pairs of activities. Let R be the set of renewable resources $R = \{r_1, r_2, \ldots r_m\}$. Each resource r_k has an integer capacity $c_k \geq 1$. Execution of an activity a_j requires one or more resources. For each resource r_k, the activity a_j requires an integer capacity $rc_{j,k}$ for the duration of its execution. An assignment of start-times to activities in A is time-feasible if all temporal constraints are satisfied and is resource-feasible if all resource constraints are satisfied. A schedule is feasible if both sets of constraints are satisfied. The problem is to find a feasible schedule with minimum makespan M where $M(S) = \max\{C_i\}$. That is, find a set of assignments to S such that $S_{\text{sol}} = \arg\min_S M(S)$. The maximum time-lag constraints are the source of difficulty—e.g., finding feasible solutions alone is NP-Hard.

Experimental Setup: We begin with a backtracking CSP heuristic search procedure for the problem (Franck, Neumann, & Schwindt 2001). We modify this algorithm to use VBSS to bias the choice made by the heuristic at each decision point. Five priority rules for the RCPSP/max problem are used as candidate search heuristics:

- smallest "latest start time" first, $\text{LST}_i = \frac{1}{1+LS_i}$;
- "minimum slack time" first, $\text{MST}_i = \frac{1}{1+LS_i-ES_i}$;
- "most total successors" first, $\text{MTS}_i = |\text{Successors}_i|$, where Successors_i is the set of not necessarily immediate successors of a_i in the project network;
- "longest path following" first, $\text{LPF}_i = \text{lpath}(i, n+1)$, where $\text{lpath}(i, n+1)$ is the length of the longest path from a_i to a_{n+1}; and
- "resource scheduling method", $\text{RSM}_i = \frac{1}{1+\max(0, \max_{g \in \text{eligible set}, g \neq i}(ES_i + p_i - LS_g))}$.

LS_i and ES_i are the latest and earliest start times. A few have been redefined from Neumann et al.'s definitions so that the eligible activity with the highest heuristic value is chosen. Eligible activities are those that can be time-feasibly scheduled given constraints involving already scheduled activities. We consider the alternative exploration policies: double exponentially increasing rate of allocations to the observed best heuristic (D-Exp); faster than double exponentially increasing rate (Faster); and exponentially increasing rate (Exp).

Table 2: Summary of the RCPSP/max results.

Algorithm	Δ_{LB}	NO	NF
D-EXP[100]	5.3	649.7	1050.7
EXP[100]	5.3	646.3	1050
FASTER[100]	5.5	617.5	1044
D-EXP[500]	4.8	665.7	1053
EXP[500]	4.8	658.2	1052.6
FASTER[500]	5.2	631	1045
D-EXP[2000]	4.6	675.7	1057
EXP[2000]	4.6	669.9	1057
FASTER[2000]	4.7	654	1051
(Smith & Pyle 2004)	6.8	679	1059
(Cesta, Oddi & Smith 2002)	8.0	670	1057

Results: Table 2 shows the results. We use the benchmark problem instances of Schwindt (2003). Δ_{LB} is the average relative deviation from the known lower bounds. NO and NF are the number of optimal solutions and feasible solutions found. The results are comparable to the first problem domain. The exponential rate of allocations to the observed best heuristic leads to over-exploration and the faster than double exponential rate leads to over-exploitation—both outperformed by the policy that allocates a double exponentially increasing number of trials to the observed best heuristic. These results are competitive with the current best known heuristic approaches to this NP-Hard problem (e.g., (Cesta, Oddi & Smith 2002; Smith & Pyle 2004; Cicirello & Smith 2004)).

Conclusions

In learning domains with a reward structure as in the Max K-Armed Bandit, where the goal is to maximize the best single sample reward received over time, we have seen that the optimal strategy is to allocate a double exponentially increasing number of trials to the observed best action. This

result depends on the assumption that each of the arms follows a Gumbel distribution. This seems restrictive. However, the extremal types theorem tells us that the distribution of the max of a series of trials belongs to one of three distribution families (Gumbel, Fréchet, or Weibull) independent of the underlying distribution of trials. We assumed that the underlying samples were drawn from a Gumbel, but could more generally assume that the distribution of the max of a series of samples from each arm is a Gumbel. The most general assumption that the max of a series of trials from each arm follows a GEV distribution would not have allowed for a closed form analysis, necessitating an approximation.

We showed how the result of the Double Exponential Sampling Theorem can be applied to the problem of allocating iterations of a heuristic guided stochastic sampling algorithm to alternative search heuristics. This approach was validated in two NP-hard scheduling domains, showing that a faster than double exponential rate of allocations to the observed best heuristic results in over exploitation of the model of heuristic performance; while a slower rate results in over exploration. The double exponential increase in the rate of allocation to the observed best heuristic provides the balance between exploration and exploitation that leads to effective problem solving in these domains.

Acknowledgements

This work was supported in part by the National Aeronautics and Space Administration under contract NCC2-1243 and the CMU Robotics Institute.

References

Agrawal, R. 1995. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization* 33(6).

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing* 32(1):48–77.

Auer, P.; Cesa-Bianchi, N.; Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47.

Beasley, J. E. 1998. Weighted tardiness. In *OR-Library*. http://mscmga.ms.ic.ac.uk/jeb/orlib/wtinfo.html.

Berry, D. A., and Fristedt, B. 1985. *Bandit Problems: Sequential Allocation of Experiments*. Chapman-Hall.

Cesta, A; Oddi, A; and Smith, S. F. 2002. A constraint-based method for project scheduling with time windows. *Journal of Heuristics* 8(1).

Cicirello, V. A., and Smith, S. F. 2004. Heuristic selection for stochastic search optimization: Modeling solution quality by extreme value theory. *The 10th Int Conf on Principles and Practice of Constraint Programming*, 197–211.

Cicirello, V. A., and Smith, S. F. 2005. Enhancing stochastic search performance by value-biased randomization of heuristics. *Journal of Heuristics* 11(1). Forthcoming.

Coles, S. 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag.

Congram, R. K.; Potts, C. N.; and van de Velde, S. L. 2002. An iterated dynasearch algorithm for the single-machine total weighted tardiness scheduling problem. *INFORMS Journal on Computing* 14(1):52–67.

Franck, B.; Neumann, K.; and Schwindt, C. 2001. Truncated branch-and-bound, schedule-construction, and schedule-improvement procedures for resource-constrained project scheduling. *OR Spektrum* 23:297–324.

Holland, J. H. 1975. *Adaptation in Natural and Artificial Systems*.

Morton, T. E., and Pentico, D. W. 1993. *Heuristic Scheduling Systems*. John Wiley and Sons.

NIST/SEMATECH. 2003. *e-Handbook of Statistical Methods*. http://www.itl.nist.gov/div898/handbook/.

Schwindt, C. 2003. http://www.wior.uni-karlsruhe.de/LS_Neumann/Forschung/ProGenMax/rcpspmax.html.

Smith, T. B., and Pyle, J. M. 2004. An effective algorithm for project scheduling with arbitrary temporal constraints. *AAAI'04*.

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