Adaptive Monte Carlo via Bandit Allocation

James Neufeld JNEUFELD@UALBERTA.CA
András György GYORGY@UALBERTA.CA
Dale Schuurmans DAES@UALBERTA.CA
Csaba Szepesvári CSABA.SZEPESVARI@UALBERTA.CA
Department of Computing Science, University of Alberta, Edmonton, AB, Canada T6G 2E8

Abstract

We consider the problem of sequentially choosing between a set of unbiased Monte Carlo estimators to minimize the mean-squared-error (MSE) of a final combined estimate. By reducing this task to a *stochastic multi-armed bandit* problem, we show that well developed allocation strategies can be used to achieve an MSE that approaches that of the best estimator chosen in retrospect. We then extend these developments to a scenario where alternative estimators have different, possibly stochastic costs. The outcome is a new set of adaptive Monte Carlo strategies that provide stronger guarantees than previous approaches while offering practical advantages.

1. Introduction

Monte Carlo methods are a pervasive approach to approximating complex integrals, which are widely deployed in all areas of science. Their widespread adoption has led to the development dozens of specialized Monte Carlo methods for any given task, each having their own tunable parameters. Consequently, it is usually difficult for a practitioner to know which approach and corresponding parameter setting might be most effective for a given problem.

In this paper we develop algorithms for sequentially allocating calls between a set of unbiased estimators to minimize the expected squared error (MSE) of a combined estimate. In particular, we formalize a new class of adaptive estimation problem: *learning to combine Monte Carlo estimators*. In this scenario, one is given a set of Monte Carlo estimators that can each approximate the expectation of some function of interest. We assume initially that each estimator is unbiased but has *unknown* variance. In practice, such estimators could include any unbiased method and/or variance reduction technique, such as unique instantiations of importance, stratified, or rejection sampling; antithetic variates; or control variates (Robert & Casella, 2005). The problem is to design a sequential allocation procedure that can interleave calls to the estimators and combine their outputs to produce a combined estimate whose MSE decreases as quickly as possible. To analyze the performance of such a meta-strategy we formalize the notion of *MSE-regret*: the time-normalized excess MSE of the combined estimate compared to the best estimator selected in hindsight, i.e., with knowledge of the distribution of estimates produced by each base estimator.

Our first main contribution is to show that this meta-task can be reduced to a *stochastic multi-armed bandit problem*, where bandit arms are identified with base estimators and the payoff of an arm is given by the negative square of its sampled estimate. In particular, we show that the MSE-regret of *any* meta-strategy is equal to its bandit-regret when the procedure is used to play in the corresponding bandit problem. As a consequence, we conclude that existing bandit algorithms, as well as their bounds on bandit-regret, can be immediately applied to achieve new results for adaptive Monte Carlo estimation. Although the underlying reduction is quite simple, the resulting adaptive allocation strategies provide novel alternatives to traditional *adaptive* Monte Carlo strategies, while providing strong finite-sample performance guarantees.

Second, we consider a more general case where the alternative estimators require different (possibly random) costs to produce their sampled estimates. Here we develop a suitably designed bandit formulation that yields bounds on the MSE-regret for cost-aware estimation. We develop new algorithms for this generalized form of adaptive Monte Carlo, provide explicit bounds on their MSE-regret, and compare their performance to a state-of-the-art adaptive Monte Carlo method. By instantiating a set of viable base estimators and selecting from them algorithmically, rather than tuning parameters manually, we discover that both computation and experimentation time can be reduced.

This work is closely related, and complementary to work on adaptive stratified sampling (Carpentier & Munos, 2011), where a strategy is designed to allocate samples between fixed strata to achieve MSE-regret bounds relative

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

to the best allocation proportion chosen in hindsight. Such work has since been extended to optimizing the number (Carpentier & Munos, 2012a) and structure (Carpentier & Munos, 2012b) of strata for differentiable functions. The method proposed in this paper, however, can be applied more broadly to any set of base estimation strategies and potentially even in combination with these approaches.

2. Background on Bandit Problems

The multi-armed bandit (MAB) problem is a sequential allocation task where an agent must choose an action at each step to maximize its long term payoff, when only the payoff of the selected action can be observed (Cesa-Bianchi & Lugosi, 2006; Bubeck & Cesa-Bianchi, 2012). In the *stochastic* MAB problem (Robbins, 1952) the payoff for each action $k \in \{1, 2, \dots, K\}$ is assumed to be generated independently and identically (i.i.d.) from a fixed but unknown distribution ν_k. The performance of an allocation policy can then by analyzed by defining the *cumulative regret* for any sequence of n actions, given by

$$R_n \doteq \max_{1 \le k \le K} \mathbb{E}\Big[\sum_{t=1}^{n} X_{k,t} - \sum_{t=1}^{n} X_{I_t, T_{I_t}(t)}\Big], \qquad (1)$$

where $X_{k,t} \in \mathbb{R}$ is the random variable giving the tth payoff of action k, $I_t \in \{1, \dots, K\}$ denotes the action taken by the policy at time-step t, and $T_k(t) \doteq \sum_{s=1}^{t} \mathbb{I}\{I_s = k\}$ denotes the number of times action k is chosen by the policy up to time t. Here, $\mathbb{I}\{p\}$ is the indicator function, set to 1 if the predicate p is true, 0 otherwise. The objective of the agent is to maximize the total payoff, or equivalently to minimize the cumulative regret. By rearranging (1) and conditioning, the regret can be rewritten

$$R_n = \sum_{k=1}^{K} \mathbb{E}[T_k(n)](\mu^* - \mu_k), \qquad (2)$$

where $\mu_k \doteq \mathbb{E}[X_{k,t}]$ and $\mu^* \doteq \max_{j=1,\dots,K} \mu_j$.

The analysis of the stochastic MAB problem was pioneered by Lai & Robbins (1985) who showed that, when the payoff distributions are defined by a single parameter, the asymptotic regret of any sub-polynomially consistent policy (i.e., a policy that selects non-optimal actions only sub-polynomially many times in the time horizon) is lower bounded by $\Omega(\log n)$. In particular, for Bernoulli payoffs

$$\liminf_{n \to \infty} \frac{R_n}{\log n} \ge \sum_{k:\Delta_k > 0} \frac{\Delta_k}{\mathrm{KL}(\mu_k, \mu^*)}, \qquad (3)$$

where $\Delta_k \doteq \mu^* - \mu_k$ and $\mathrm{KL}(p, q) \doteq p \log(p/q) + (1 - p) \log(\frac{1-p}{1-q})$ for $p, q \in [0, 1]$. Lai & Robbins (1985) also presented an algorithm based on upper confidence bounds (UCB), which achieves a regret asymptotically matching the lower bound (for certain parametric distributions).

Later, Auer et al. (2002a) proposed UCB1 (Algorithm 1), which broadens the practical use of UCB by dropping the requirement that payoff distributions fit a particular parametric form. Instead, one need only make the much weaker assumption that the rewards are bounded; in particular, we let $X_{k,t} \in [0, 1]$. Auer et al. (2002a) proved that, for any finite number of actions n, UCB1's regret is bounded by

$$R_n \le \sum_{k:\Delta_k > 0} \frac{8 \log n}{\Delta_k} + \Big(1 + \frac{\pi^2}{3}\Big)\Delta_k. \qquad (4)$$

Various improvements of the UCB1 algorithm have since been proposed. One approach of particular interest is the UCB-V algorithm (Audibert et al., 2009), which takes the empirical variances into account when constructing confidence bounds. Specifically, UCB-V uses the bound

$$\bar{\mu}_{k,t} + \sqrt{\frac{2\mathbb{V}_{k,T_k(t-1)}\mathcal{E}_{T_k(t-1),t}}{T_k(t-1)}} + c\frac{3\mathcal{E}_{T_k(t-1),t}}{T_k(t-1)},$$

where $\mathbb{V}_{k,s}$ denotes the empirical variance of arm k's payoffs after s pulls, $c > 0$, and $\mathcal{E}_{s,t}$ is an *exploration function* required to be a non-decreasing function of s or t (typically $\mathcal{E}_{s,t} = \zeta \log(t)$ for a fixed constant $\zeta > 0$). The UCB-V procedure can then be constructed by substituting the above confidence bound into Algorithm 1, which yields a regret bound that scales with the true variance of each arm

$$R_n \le c_\zeta \sum_{k:\Delta_k > 0} \Big(\frac{\mathbb{V}(X_{k,1})}{\Delta_k} + 2\Big) \log n. \qquad (5)$$

Here c_ζ is a constant relating to ζ and c. In the worst case, when $\mathbb{V}(X_{k,1}) = 1/4$ and $\Delta_k = 1/2$, this bound is slightly worse than UCB1's bound; however, it is usually better in practice, particularly if some k has small Δ_k and $\mathbb{V}(X_{k,1})$.

A more recent algorithm is KL-UCB (Cappé et al., 2013), where the confidence bound for arm k is based on solving

$$\sup\Big\{\mu : \mathrm{KL}\left(\bar{\mu}_{k,t}, \mu\right) \le \frac{f(t)}{T_k(t)}\Big\},$$

for a chosen increasing function $f(\cdot)$, which can be solved efficiently since $\mathrm{KL}(p, \cdot)$ is smooth and increasing on $[p, 1]$. By choosing $f(t) = \log(t) + 3\log\log(t)$ for $t \ge 3$ (and $f(1) = f(2) = f(3)$), KL-UCB achieves a regret bound

$$R_n \le \sum_{k:\Delta_k > 0} \Big(\frac{\Delta_k}{\mathrm{KL}\left(\mu_k, \mu^*\right)}\Big) \log n + O(\sqrt{\log(n)}) \qquad (6)$$

Algorithm 1 UCB1 (Auer et al., 2002a)

for $k \in \{1, \dots, K\}$
  Play k, observe $X_{k,1}$, set $\bar{\mu}_{k,1} := X_{k,1}$; $T_k(1) := 1$.
end for
for $t \in \{K+1, K+2, \dots\}$
  Play action k that maximizes $\bar{\mu}_{j,t-1} + \sqrt{\frac{2\log t}{T_j(t-1)}}$; set $T_k(t) = T_k(t-1) + 1$ and $T_j(t) = T_j(t-1)$ for $j \neq k$, observe payoff $X_{k,T_k(t)}$, and compute $\bar{\mu}_{k,t} = (1 - 1/T_k(t))\bar{\mu}_{k,t-1} + X_{k,T_k(t)}/T_k(t)$.
end for

Algorithm 2 Thompson Sampling (Agrawal & Goyal, 2012)

Require: Prior parameters α and β
Initialize: $S_{1:K}(0) := 0$, $F_{1:K}(0) := 0$
for $t \in \{1, 2, \ldots\}$
  Sample $\theta_{t,k} \sim \mathcal{B}(S_k(t-1)+\alpha, F_k(t-1)+\beta)$, $k = 1 \ldots K$
  Play $k = \arg\max_{j=1,\ldots,K} \theta_{t,j}$; observe $X_t \in [0, 1]$
  Sample $\hat{X}_t \sim \text{Bernoulli}(X_t)$
  if $\hat{X}_t = 1$ **then** set $S_k(t) = S_k(t-1) + 1$ **else** set $F_k(t) = F_k(t) + 1$
end for

for $n \geq 3$, with explicit constants for the "higher order" terms (Cappé et al., 2013, Corollary 1). Apart from the higher order terms, this bound matches the lower bound (3). In general, KL-UCB is expected to be better than UCB-V except for large sample sizes and small variances. Note that, given any set of UCB algorithms, one can apply the tightest upper confidence from the set, via the union bound, at the price of a small additional constant in the regret.

Another approach that has received significant recent interest is Thompson sampling (TS) (Thompson, 1933): a Bayesian method where actions are chosen randomly in proportion to the posterior probability that their mean payoff is optimal. TS is known to outperform UCB-variants when payoffs are Bernoulli distributed (Chapelle & Li, 2011; May & Leslie, 2011). Indeed, the finite time regret of TS under Bernoulli payoff distributions closely matches the lower bound (3) (Kaufmann et al., 2012):

$$R_n \leq (1+\varepsilon) \sum_{k:\Delta_k>0} \frac{\Delta_k(\log(n) + \log\log(n))}{\mathrm{KL}(\mu_k, \mu^*)} + C(\varepsilon, \mu_{1:K}),$$

for every $\varepsilon > 0$, where C is a problem-dependant constant. However, since it is not possible to have Bernoulli distributed payoffs with the same mean and different variances, this analysis is not directly applicable to our setting. Instead, we consider a more general version of Thompson sampling (Agrawal & Goyal, 2012) that converts real-valued to Bernoulli-distributed payoffs through a resampling step (Algorithm 2), which has been shown to obtain

$$R_n \leq \Big(\sum_{k:\Delta_k>0} \frac{1}{\Delta_k^2} \Big)^2 \log(n). \tag{7}$$

3. Combining Monte Carlo Estimators

We now formalize the main problem we consider in this paper. Assume we are given a finite number of Monte Carlo estimators, $k = 1, \ldots, K$, where base estimator k produces a sequence of real-valued random variables $(X_{k,t})_{(t=1,2,\ldots)}$ whose mean converges to the unknown target quantity, $\mu \in \mathbb{R}$. Observations from the different estimators are assumed to be independent from each other. We assume, initially, that drawing a sample from each estimator takes constant time, hence the estimators differ only in terms of how fast their respective sample means $\overline{X}_{k,n} = \frac{1}{n}\sum_{t=1}^{n} X_{k,t}$ converge to μ. The goal is to design a *sequential estimation procedure* that works in discrete time steps: For each round $t = 1, 2, \ldots$, based on the previous observations, the procedure selects one estimator $I_t \in \{1, \ldots, K\}$, whose observation is used by an outer procedure to update the estimate $\hat{\mu}_t \in \mathbb{R}$ based on the values observed so far.

As is common in the Monte Carlo literature, we evaluate accuracy by the mean-squared error (MSE). That is, we define the loss of the sequential method $\mathcal{A}$ at the end of round n by $L_n(\mathcal{A}) = \mathbb{E}\left[(\hat{\mu}_n - \mu)^2\right]$. A reasonable goal is to then compare the loss, $L_{k,n} = \mathbb{E}\left[(\overline{X}_{k,n} - \mu)^2\right]$, of each base estimator to the loss of $\mathcal{A}$. In particular, we propose to evaluate the performance of $\mathcal{A}$ by the (normalized) regret

$$R_n(\mathcal{A}) = n^2 \Big(L_n(\mathcal{A}) - \min_{1\leq k\leq K} L_{k,n} \Big),$$

which measures the excess loss of $\mathcal{A}$ due to its initial ignorance of estimator quality. Implicit in this definition is the assumption that $\mathcal{A}$'s time to select the next estimator is negligible compared to the time to draw an observation. Note also that the excess loss is multiplied by n^2, which ensures that, in standard settings, when $L_{k,n} \propto 1/n$, a sublinear regret (i.e., $|R_n(\mathcal{A}|)/n \to 0$ as $n \to \infty$) implies that the loss of $\mathcal{A}$ asymptotically matches that of the best estimator.

In the next two sections we will adopt a simple strategy for combining the values returned from the base estimators: $\mathcal{A}$ simply returns their (unweighted) average as the estimate $\hat{\mu}_n$ of μ. A more sophisticated approach may be to weight each of these samples inversely proportional to their respective (sample) variances. However, if the adaptive procedure can quickly identify and ignore highly suboptimal arms the savings from the weighted estimator will diminish rapidly. Interestingly, this argument does not immediately translate to the nonuniform cost case considered in Section 5 as will be shown empirically in Section 6.

4. Combining Unbiased I.I.D. Estimators

Our main assumption in this section will be the following:

Assumption 4.1. *Each estimator produces a sequence of i.i.d. random observations with common mean μ and finite variance; values from different estimators are independent.*

Let ψ_k denote the distribution of samples from estimator k. Note that $\Psi = (\psi_k)_{1\leq k\leq K}$ completely determines the sequential estimation problem. Since the samples coming from estimator k are i.i.d., we have $\mathbb{V}(X_{k,1}) = \mathbb{V}(X_{k,t})$. Let $V_k = \mathbb{V}(X_{k,1})$ and $V^* = \min_{1\leq k\leq K} V_k$. Furthermore, let $L_{k,t} = \mathbb{V}(X_{k,1})/t$, hence $\min_{1\leq k\leq K} L_{k,t} = V^*/t$. We then have the first main result of this section.

Theorem 1 (Regret Identity)**.** *Consider K estimators for which Assumption 4.1 holds, and let $\mathcal{A}$ be an arbitrary allocation procedure. Then, for any $n \geq 1$, the MSE-regret*

of the estimation procedure $\mathcal{A}^{\text{avg}}$, estimating μ using the sample-mean of the observations obtained by $\mathcal{A}$, satisfies

$$R_n(\mathcal{A}^{\text{avg}}) = \sum_{k=1}^{K} \mathbb{E}\left[T_k(n)\right](V_k - V^*)\,. \tag{8}$$

The proof follows from a simple calculation given in Appendix A. Essentially, one can rewrite the loss as $L_n(\mathcal{A}) = \frac{1}{n^2}\mathbb{E}\left[\sum_{k=1}^{K} S_{k,n}^2 + 2\sum_{k\neq j} S_{k,n}S_{j,n}\right]$, where $S_{k,n}$ is the centered sum of observations for arm k. The cross-terms can be shown to cancel by independence, and Wald's second identity with some algebra gives the result.

The tight connection between sequential estimation and bandit problems revealed by (8) allows one to reduce sequential estimation to the design of bandit strategies and vice versa. Furthermore, regret bounds transfer both ways.

Theorem 2 (Reduction)**.** *Let Assumption 4.1 hold for Ψ. Define a corresponding bandit problem (ν_k) by assigning ν_k as the distribution of $-X_{k,1}^2$. Given an arbitrary allocation strategy $\mathcal{A}$, let $\text{Bandit}(\mathcal{A})$ be the bandit strategy that consults $\mathcal{A}$ to select the next arm after obtaining reward Y_t (assumed nonpositive), based on feeding observations $(-Y_t)^{1/2}$ to $\mathcal{A}$ and copying $\mathcal{A}$'s choices. Then, the bandit-regret of $\text{Bandit}(\mathcal{A})$ in bandit problem (ν_k) is the same as the MSE-regret of $\mathcal{A}$ in estimation problem Ψ. Conversely, given an arbitrary bandit strategy $\mathcal{B}$, let $\text{MC}(\mathcal{B})$ be the allocation strategy that consults $\mathcal{B}$ to select the next estimator after observing $-Y_t^2$, based on feeding rewards Y_t to $\mathcal{B}$ and copying $\mathcal{B}$'s choices. Then the MSE-regret of $\text{MC}(\mathcal{B})$ in estimation problem Ψ is the same as the bandit-regret of $\mathcal{B}$ in bandit problem (ν_k) (where $\text{MC}(\mathcal{B})$ uses the average of observations as its estimate).*

Proof of Theorem 2. The result follows from Theorem 1 since $V_k = \mathbb{E}[X_{k,1}^2] - \mu^2$ and $V^* = \mathbb{E}[X_{k^*,1}^2] - \mu^2$ where k^* is the lowest variance estimator, hence $V_k - V^* = \mathbb{E}[X_{k,1}^2] - \min_{1\le k'\le K}\mathbb{E}[X_{k',1}^2]$. Furthermore, the bandit problem (ν_k) ensures the regret of a procedure that chooses arm k $T_k(n)$ times is $\sum_{k=1}^{K}\mathbb{E}[T_k(n)]\Delta_k$, where $\Delta_k = \max_{1\le k'\le K}\mathbb{E}[-X_{k',1}^2] - \mathbb{E}[-X_{k,1}^2] = V_k - V^*$. □

From this theorem one can also derive a lower bound. First, let $\mathbb{V}(\psi)$ denote the variance of $X\sim\psi$ and let $\mathbb{V}^*(\Psi) = \min_{1\le k\le K}\mathbb{V}(\psi_k)$. For a family $\mathcal{F}$ of distributions over the reals, let $D_{\inf}(\psi, v, \mathcal{F}) = \inf_{\psi'\in\mathcal{F}:\mathbb{V}(\psi')<v} D(\psi,\psi')$, where $D(\psi,\phi) = \int \log\frac{d\psi}{d\phi}(x)\,d\psi(x)$, if the Radon-Nikodym derivative $d\psi/d\phi$ exists, and ∞ otherwise. Note that $D_{\inf}(\psi, v, \mathcal{F})$ measures how distinguishable ψ is from distributions in $\mathcal{F}$ having smaller variance than v. Further, we let $R_n(\mathcal{A}, \Psi)$ denote the regret of $\mathcal{A}$ on the estimation problem specified using the distributions Ψ.

Theorem 3 (MSE-Regret Lower Bound)**.** *Let $\mathcal{F}$ be the set of distributions supported on $[0,1]$ and assume that $\mathcal{A}$ allocates a subpolynomial fraction to suboptimal estimators for any $\Psi\in\mathcal{F}^K$: i.e., $\mathbb{E}_\Psi[T_k(n)] = O(n^a)$ for all $a>0$ and k such that $\mathbb{V}(\psi_k) > \mathbb{V}^*(\Psi)$. Then, for any $\Psi\in\mathcal{F}$ where not all variances are equal and $0 < D_{\inf}(\psi_k, \mathbb{V}^*(\Psi), \mathcal{F}) < \infty$ holds whenever $\mathbb{V}(\psi_k) > \mathbb{V}^*(\Psi)$, we have*

$$\liminf_{n\to\infty}\frac{R_n(\mathcal{A},\Psi)}{\log n} \ge \sum_{k:\mathbb{V}(\psi_k)>\mathbb{V}^*(\Psi)} \frac{\mathbb{V}(\psi_k) - \mathbb{V}^*(\Psi)}{D_{\inf}(\psi_k, \mathbb{V}^*(\Psi), \mathcal{F})}.$$

Proof. The result follows from Theorem 2 and (Burnetas & Katehakis, 1996, Proposition 1). □

Using Theorem 2 we can also establish bounds on the MSE-regret for the algorithms mentioned in Section 2.

Theorem 4 (MSE-Regret Upper Bounds)**.** *Let Assumption 4.1 hold for $\Psi = (\psi_k)$ where (for simplicity) we assume each ψ_k is supported on $[0,1]$. Then, after n rounds, $\text{MC}(\mathcal{B})$ achieves the MSE-Regret bound of: (4) when using $\mathcal{B}$ = UCB1; (5) when using $\mathcal{B}$ = UCB-V with $c_\zeta = 10$; (6) when using $\mathcal{B}$ = UCB-KL; and (7) when using $\mathcal{B}$ = TS.*[1]

Additionally, due to Theorem 2, one can also obtain bounds on the minimax MSE-regret by exploiting the lower bound for bandits in (Auer et al., 2002b). In particular, the UCB-based bandit algorithms above can all be shown to achieve the minimax rate $O(\sqrt{Kn})$ up to logarithmic factors, immediately implying that the minimax MSE-regret of $\text{MC}(\mathcal{B})$ for $\mathcal{B}\in\{\text{UCB1}, \text{UCB-V}, \text{KL-UCB}\}$ is of order $L_n(\text{MC}(\mathcal{B})) - L_n^* = \tilde{O}(K^{1/2}n^{-(1+\frac{1}{2})})$.[2]

Appendix B provides a discussion of how alternative ranges on the observations can be handled, and how the above methods can still be applied when the payoff distribution is unbounded but satisfies moment conditions.

5. Non-uniform Estimator Costs

Next, we consider the case when the base estimators can take *different* amounts of time to generate observations. A consequence of non-uniform estimator times, which we refer to as *non-uniform costs*, is that the definitions of the loss and regret must be modified accordingly. Intuitively, if an estimator takes more time to produce an observation, it is less useful than another estimator that produces observations with (say) identical variance but in less time.

To develop an appropriate notion of regret for this case, we introduce some additional notation. Let $D_{k,m}$ denote the time needed by estimator k to produce its mth observation, $X_{k,m}$. As before, we let $I_m\in\{1,\dots,K\}$ denote the index of the estimator that $\mathcal{A}$ chooses in round m. Let J_m denote the time when $\mathcal{A}$ observes the mth sample,

[1] Note that to apply the regret bounds from Section 2, one has to feed the bandit algorithms with $1 - Y_t^2$ instead of $-Y_t^2$ in Theorem 2. This modification has no effect on the regret.

[2] $\tilde{O}$ denotes the order up to logarithmic factors. To remove such factors one can exploit MOSS (Audibert & Bubeck, 2010).

$Y_m = X_{I_m,T_{I_m}(m)}$; thus, $J_1 = D_{I_1,1}$, $J_2 = J_1 + D_{I_2,T_{I_2}(2)}$, and $J_{m+1} = J_m + D_{I_m,T_{I_m}(m)} = \sum_{s=1}^{m} D_{I_s,T_{I_s}(s)}$. For convenience, define $J_0 = 0$. Note that round m starts at time J_{m-1} with $\mathcal{A}$ choosing an estimator, and finishes at time J_m when the observation is received and $\mathcal{A}$ (instantaneously) updates it estimate. Thus, at time J_m a new estimate $\hat{\mu}_m$ becomes available: the estimate is "renewed". Let $\hat{\mu}(t)$ denote the estimate available at time $t \geq 0$. Assuming $\mathcal{A}$ produces a default estimate $\hat{\mu}_0$ before the first observation, we have $\hat{\mu}(t) = \hat{\mu}_0$ on $[0, J_1)$, $\hat{\mu}(t) = \hat{\mu}_1$ on $[J_1, J_2)$, etc. If $N(t)$ denotes the round index at time t (i.e., $N(t) = 1$ on $[0, J_1)$, $N(t) = 2$ on $[J_1, J_2)$, etc.) then $\hat{\mu}(t) = \hat{\mu}_{N(t)-1}$. The MSE of $\mathcal{A}$ at time $t \geq 0$ is

$$L(\mathcal{A}, t) = \mathbb{E}\left[(\hat{\mu}(t) - \mu)^2\right] .$$

By comparison, the estimate for a single estimator k at time t is $\hat{\mu}_k(t) = \mathbb{I}\{N_k(t) > 1\} \frac{\sum_{m=1}^{N_k(t)-1} X_{k,m}}{N_k(t)-1}$, where $N_k(t) = 1$ on $[0, D_{k,1})$, $N_k(t) = 2$ on $[D_{k,1}, D_{k,1} + D_{k,2})$, etc. We set $\hat{\mu}_k(t) = 0$ on $[0, D_{k,1})$ to let $\hat{\mu}_k(t)$ be well-defined on $[0, D_{k,1})$. Thus, at time $t \geq 0$ the MSE of estimator k is

$$L_k(t) = \mathbb{E}\left[(\hat{\mu}_k(t) - \mu)^2\right] .$$

Given these definitions, it is natural to define the regret as

$$R(\mathcal{A}, t) = t^2 \Big(L(\mathcal{A}, t) - \min_{1 \leq k \leq K} L_k(t) \Big) .$$

As before, the t^2 scaling is chosen so that, under the condition that $L_k(t) \propto 1/t$, a sublinear regret implies that $\mathcal{A}$ is "learning". Note that this definition generalizes the previous one: if $D_{k,m} = 1 \, \forall k, m$, then $R_n(\mathcal{A}) = R(\mathcal{A}, n)$.

In this section we make the following assumption.

Assumption 5.1. *For each k, $(X_{k,m}, D_{k,m})_{(m=1,2,\ldots)}$ is an i.i.d. sequence such that $\mathbb{E}[X_{k,1}] = \mu$, $V_k \doteq \mathbb{V}(X_{k,1}) < \infty$, $\mathbb{P}(D_{k,m} > 0) = 1$ and $\delta_k \doteq \mathbb{E}[D_{k,1}] = \mathbb{E}[D_{k,m}] < \infty$. Furthermore, we assume that the sequences for different k are independent of each other.*

Note that Assumption 5.1 allows $D_{k,m}$ to be a deterministic value; a case that holds when the estimators use deterministic algorithms to produce observations. Another situation arises when $D_{k,m}$ is stochastic (i.e., the estimator uses a randomized algorithm) and $(X_{k,m}, D_{k,m})$ are correlated. In which case $\hat{\mu}_k(t)$ may be a biased estimate of μ. However, if $(X_{k,m})_m$ and $(D_{k,m})_m$ are independent and $\mathbb{P}(N_k(t) > 1) = 1$ then $\hat{\mu}_k(t)$ is unbiased. Indeed, in such a case, $(N_k(t))_t$ is independent of the partial sums $(\sum_{m=1}^{n} X_{k,m})_n$, hence $\mathbb{E}[\hat{\mu}_k(t)] = \mathbb{E}\left[\frac{\sum_{m=1}^{N_k(t)-1} X_{k,m}}{N_k(t)-1}\right] = \sum_{n=2}^{\infty} \mathbb{P}(N_k(t) = n) \mathbb{E}\left[\frac{\sum_{m=1}^{n-1} X_{k,m}}{n-1} \middle| N_k(t) = n\right] = \mathbb{P}(N_k(t) > 1)\, \mathbb{E}[X]$, because $\hat{\mu}_k(t) = 0$ when $N_k(t) \leq 1$.

Using Assumption 5.1, a standard argument of *renewal reward processes* gives $L_k(t) \sim V_k/(t/\delta_k) = V_k \delta_k / t$, where $f(t) \sim g(t)$ means $\lim_{t\to\infty} f(t)/g(t) = 1$. Intuitively, estimator k will produce approximately t/δ_k independent observations during $[0, t)$; hence, the variance of their average is approximately $V_k/(t/\delta_k)$ (made more precise in the proof of Theorem 5). This implies $\min_{1\leq k\leq K} L_k(t) \sim \min_{1\leq k\leq K} \frac{\delta_k V_k}{t}$. Thus, any allocation strategy $\mathcal{A}$ competing with the best estimator must draw most of its observations from k satisfying $\delta_k V_k = \delta^* V^* \doteq \min_{1\leq k\leq K} \frac{\delta_k V_k}{t}$. For simplicity, we assume $k^* = \arg\min_{1\leq k\leq K} \delta_k V_k$ is unique, with $\delta^* = \delta_{k^*}$ and $V^* = V_{k^*}$.

As before, we will consider adaptive strategies that estimate μ using the mean of the observations: $\hat{\mu}_m = \frac{S_m}{m}$ such that $S_m \doteq \sum_{s=1}^{m} Y_s$. Hence, the estimate at time $t \geq J_1$ is

$$\hat{\mu}(t) = \frac{S(t)}{N(t) - 1}, \text{ where } S(t) = S_{N(t)-1} . \tag{9}$$

Our aim is to bound regret of the overall algorithm by bounding the number of times the allocation strategy chooses suboptimal estimators. We will do so by generalizing (2) to the nonuniform-cost case, but unlike the equality obtained in Theorem 1, here we provide an upper bound.

Theorem 5. *Let Assumption 5.1 hold and assume that $(X_{k,m})$ are bounded and k^* is unique. Let the estimate of $\mathcal{A}$ at time t be defined by the sample mean $\hat{\mu}(t)$. Let $t > 0$ be such that $\mathbb{E}[N(t) - 1] > 0$ and $\mathbb{E}[N_{k^*}(t)] > 0$, and assume that for any $k \neq k^*$, $\mathbb{E}[T_k(N(t))] \leq f(t)$ for some $f : (0, \infty) \to [1, \infty)$ such that $f(t) \leq c_f t$ for some $c_f > 0$ and any $t > 0$. Assume furthermore that $\mathbb{P}(D_{k,1} > t) \leq C_D t^{-2}$ and $\mathbb{E}\left[N_{k^*}(t)^2\right] \leq C_N t^2$ for all $t > 0$. Then, for any $c < \sqrt{t/(8\delta_{\max})}$ where $\delta_{\max} = \max_k \delta_k$, the regret of $\mathcal{A}$ at time t is bounded by*

$$\begin{aligned} R(\mathcal{A}, t) \;\leq\; & (c + C)\sqrt{t} \;+\; C' f(t) \\ & +\; C'' t^2\, \mathbb{P}\Big(N_{k^*}(t) > \mathbb{E}[N_{k^*}(t)] + c\sqrt{\mathbb{E}[N_{k^*}(t) - 1]} \Big) \\ & +\; C''' t^2\, \mathbb{P}\Big(N(t) < \mathbb{E}[N(t)] - c\sqrt{\mathbb{E}[N(t) - 1]} \Big), \end{aligned} \tag{10}$$

for some appropriate constants $C, C', C'', C''' > 0$ that depend on the problem parameters δ_k, V_k, the upper bound on $|X_{k,m}|$, and the constants c_f, C_D and C_N.

The proof of the theorem is given in Appendix C. Several comments are in order. First, recall that the optimal regret rate is order $\sqrt{t}$ in this setting, up to logarithmic factors. To obtain such a rate, one need only achieve $f(t) = O(\sqrt{t})$, which can be attained by stochastic or even adversarial bandit algorithms (Bubeck & Cesa-Bianchi, 2012) receiving rewards with expectation $-\delta_k V_k$ and a well-concentrated number of samples. The moment condition on $N_{k^*}(t)$ is also not restrictive; for example, if the estimators are rejection samplers, their sampling times will have a geometric distribution that satisfies the polynomial tail condition. Furthermore, if $D_{k,m} \geq \delta^-$ for some $\delta^- > 0$ then $N_k(t) < t/\delta^-$ for all k, which ensures the moment condition on $N_{k^*}(t)$.

Although it was sufficient to use the negative second moment $-X_{k,m}^2$ instead of variance as the bandit reward under uniform costs, this simplification is no longer possible when costs are nonuniform, since $\delta_k V_k = \delta_k(\mathbb{E}[X_{k,1}^2] - \mu^2)$ now involves the unknown expectation μ. Several strategies can be followed to bypass this difficulty. For example, given independent costs and observations, one can use each bandit algorithm decision twice, feeding rewards $r_{k,m} = -\frac{1}{4}(D_{k,2m} + D_{k,2m+1})(X_{k,2m} - X_{k,2m+1})^2$ whose expectation is $\delta_k V_k$. Similar constructions using slightly more data can be used for the dependent case.

Note that ensuring $\mathbb{E}[T_k(N(t))] \leq f(t)$ can be nontrivial. Typical guarantees for UCB-type algorithms ensure that the expected number of pulls to a suboptimal arm k in n rounds is bounded by a function $g_k(n)$. However, due to their dependence, $\mathbb{E}[T_k(N(t))]$ cannot generally be bounded by $g_k(\mathbb{E}[N(t)])$. Nevertheless, if, for example, $D_{k,m} \geq \delta^-$ for some $\delta^- > 0$, then $N(t) - 1 \leq t/\delta^-$, hence $f_k(t) = g_k(t/\delta^-)$ can be used.

Finally, we need to ensure that the last two terms in (10) remain small, which follows if $N(t)$ and $N_{k^*}(t)$ concentrate around their means. In general, $\mathbb{P}(N < \mathbb{E}[N] - C\sqrt{\mathbb{E}[N]\log(1/\delta)}) \leq \delta$ for some constant C, therefore $c = C\sqrt{\log(1/\delta)}$ can be chosen to achieve $t^2\mathbb{P}(N < \mathbb{E}[N] - c\sqrt{\mathbb{E}[N]}) \leq t^2\delta$, hence by chosing δ to be $O(t^{-3/2})$ we achieve $\tilde{O}(\sqrt{t})$ regret. However, to ensure concentration, the allocation strategy must also select the optimal estimator most of the time. For example, Audibert et al. (2009) show that with default parameters, UCB1 and UCB-V will select suboptimal arms with probability $\Omega(1/n)$, making $t^2\mathbb{P}(N < \mathbb{E}[N] - c\sqrt{\mathbb{E}[N]}) = \Omega(t)$. However, by increasing the constant 2 in UCB1 and the parameter ζ in UCB-V, it follows from (Audibert et al., 2009) that the chance of using *any* suboptimal arm more than $c\log(t)\sqrt{t}$ times can be made smaller than c/t (where c is some problem-dependent constant). Outside of this small probability event, the optimal arm is used $t - cK\log(t)\sqrt{t}$ times, which is sufficient to show concentration of $N(t)$. In summary, we conclude that $\tilde{O}(\sqrt{t})$ regret can be achieved in Theorem 5 under reasonable assumptions.

6. Experiments

We conduct experimental investigations in a number of scenarios to better understand the effectiveness of multi-armed bandit algorithms for adaptive Monte Carlo estimation.

6.1. Preliminary Investigation: A 2-Estimator Problem

We first consider the performance of allocation strategies on a simple 2-estimator problem. Note that this evaluation differs from standard evaluations of stochastic bandits through the absence of single-parameter payoff distributions, such as the Bernoulli, which cannot have identical means yet different variances. This is an important detail, since stochastic bandit algorithms such as KL-UCB and TS are often evaluated on single-parameter payoff distributions, but their advantages in such scenarios might not extend to adaptive Monte Carlo estimation.

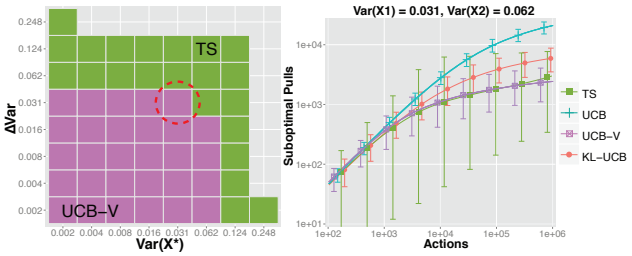

Figure 1. **Left**: Tile plot indicating which approach achieved lowest regret (averaged over 2000 independent runs) in the 2-estimator scaled-Bernoulli setting, at time 10^6. X-axis is the variance of the optimal estimator, and Y-axis is the *additional* variance on the second estimator. **Right**: Log-plot illustrating the expected number of suboptimal selections for the highlighted case (dashed red circle). Error bars indicate 99% empirical percentiles.

In particular, we consider problems when $X_{k,t} = \mu + s_k(Z_t - \frac{1}{2})$, where Z_t is standard Bernoulli and $s_k \in (0, 1)$ is a separate scale parameter for $k \in \{1, 2\}$. This design permits the maximum range for variance around a mean within a bounded interval. We evaluated the four bandit strategies detailed in Section 2: UCB1, UCB-V, KL-UCB, and TS, where for UCB-V we used the same settings as (Audibert et al., 2009), and for TS we used the uniform Beta prior, i.e., $\alpha_0 = 1$ and $\beta_0 = 1$.

The relative performance of these approaches is reported in Figure 1. TS appears best suited for scenarios where *either estimator* has high variance, whereas UCB-V is more effective when faced with medium or low variance estimators. Additionally, KL-UCB out-performs UCB-V in high variance settings, but in all such cases was eclipsed by TS.

6.2. Option Pricing

We next consider a more practical application of adaptive Monte Carlo estimation to the problem of pricing financial instruments. In particular, following (Douc et al., 2007; Arouna, 2004), we consider the problem of pricing *European call options* under the assumption that the interest rate evolves in time according to the Cox-Ingersoll-Ross (CIR) model (Cox et al., 1985), a popular model in mathematical finance (details provided in Appendix F). In a nutshell, this model assumes that the interest rate $r(t)$, as a function of time $t > 0$, follows a *square root diffusion model*. The price of a European caplet option with "strike price" $K > 0$, "nominee amount" $M > 0$ and "maturity" $T > 0$ is then given by $P = M\exp(-\int_0^T r(t)dt)\max(r(T) - K, 0)$. The problem is to determine the expected value of P.

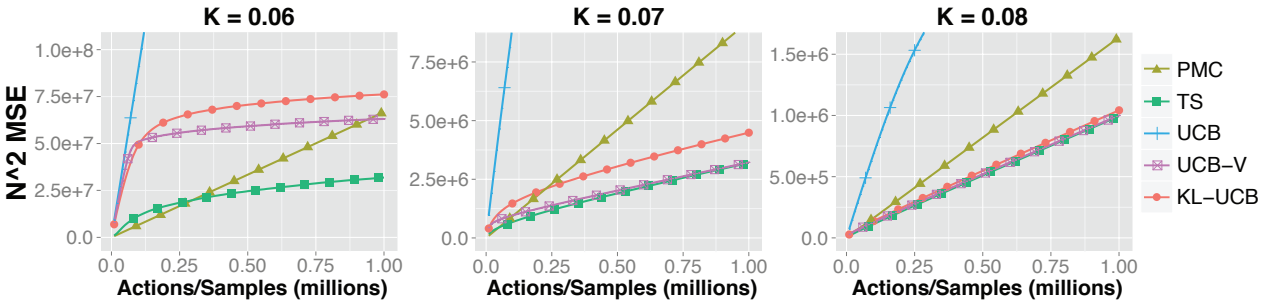

Figure 2. Plots showing the normalized MSE of different adaptive strategies for estimating the price of a European caplet option under the Cox-Ingersol-Ross interest rate model, using different strike prices (K). All results are statistically significant up to visual resolution.

A naive approach to estimating $\mathbb{E}[P]$ is to simulate independent realizations of $r(t)$ for $0 < t \leq T$. However, any simulation where the interest rate $r(T)$ lands below the strike price can be ignored since the payoff is zero. Therefore, a common estimation strategy is to use importance sampling by introducing a "drift" parameter $\theta > 0$ into the proposal density for $r(t)$, with $\theta = 0$ meaning no drift; this encourages simulations with higher interest rates. The importance weights for these simulations can then be efficiently calculated as a function of θ (see Appendix F).

Importantly, the task of adaptively allocating trials between different importance sampling estimators has been previously studied on this problem, using an unrelated technique known as *d-kernel Population Monte Carlo (PMC)* (Douc et al., 2007). Space restrictions prevent us from providing a full description of the PMC method, but, roughly speaking, the method defines the proposal density as mixture over the set $\{\theta_k\}$ of drift parameters considered. At each time step, PMC samples a new drift value according to this mixture and then simulates an interest rate. After a fixed number of samples, say G (the *population* size), the mixture coefficient α_k of each drift parameter θ_k is adjusted by setting it to be proportional to the sum of importance weights sampled from that parameter: $\alpha_k = \frac{\sum_{t=1}^{G} w_t \mathbb{I}\{I_t = k\}}{\sum_{t=1}^{G} w_t}$. The new proposal is then used to generate the next population.

We approximated the option prices under the same parameter settings as (Douc et al., 2007), namely, $\nu = 0.016$, $\kappa = 0.2$, $r_0 = 0.08$, $T = 1$, $M = 1000$, $\sigma = 0.02$, and $n = 100$, for different strike prices $K = \{0.06, 0.07, 0.08\}$ (see Appendix F). However, we consider a wider set of proposals given by $\theta_k = k/10$ for $k \in \{0, 1, ..., 15\}$. The results averaged over 1000 simulations are given in Figure 2.

These results generally indicate that the more effective bandit approaches are significantly better suited to this allocation task than the PMC approach, particularly in the longer term. Among the bandit based strategies, TS is the clear winner, which, given the conclusions from the previous experiment, is likely due to high level of variance introduced by the option pricing formula. Despite this strong showing for the bandit methods, PMC remains surprisingly competitive at this task, doing uniformly better than UCB, and better than all other bandit allocation strategies early on for $K = 0.06$. However, we believe that this advantage of PMC stems from the fact that PMC explore the entire space of mixture distribution (rather than single θ_k). It remains an interesting area for future work in bandit-based allocation strategies to extend the existing methods to continuously parameterized settings.

6.3. Adaptive Annealed Importance Sampling

Many important applications of Monte Carlo estimation occur in Bayesian inference, where a particularly challenging problem is evaluating the *model evidence* of a latent variable model. Evaluating such quantities is useful for a variety of purposes, such as Bayesian model comparison and testing/training set evaluation (Robert, 2012). However, the desired quantities are notoriously difficult to estimate in many important settings, due in no small part to the fact that popular high-dimensional Monte Carlo strategies, such as Markov Chain Monte Carlo (MCMC) methods, cannot be directly applied (Neal, 2005).

Nevertheless, a popular approach for approximating such values is *annealed importance sampling* (AIS) (Neal, 2001) (or more generally sequential Monte Carlo samplers (Del Moral et al., 2006)). In a nutshell, AIS combines the advantages of importance sampling with MCMC by defining a proposal density through a sequence of MCMC transitions applied to a sequence of *annealed* distributions, which slowly blend between the proposal (prior) and the target (un-normalized posterior). While such a technique can offer impressive practical advantages, it often requires considerable effort to set parameters; in particular, the practitioner must specify the number of annealing steps, the annealing rate or "schedule", the underlying MCMC method (and its parameters), and the number of MCMC transitions to execute at annealing step. Even when these parameters have been appropriately tuned on preliminary data, there is no assurance that these choices will remain effective when deployed on larger or slightly different data sets.

Here we consider the problem of approximating the normalization constant for a Bayesian logistic regression

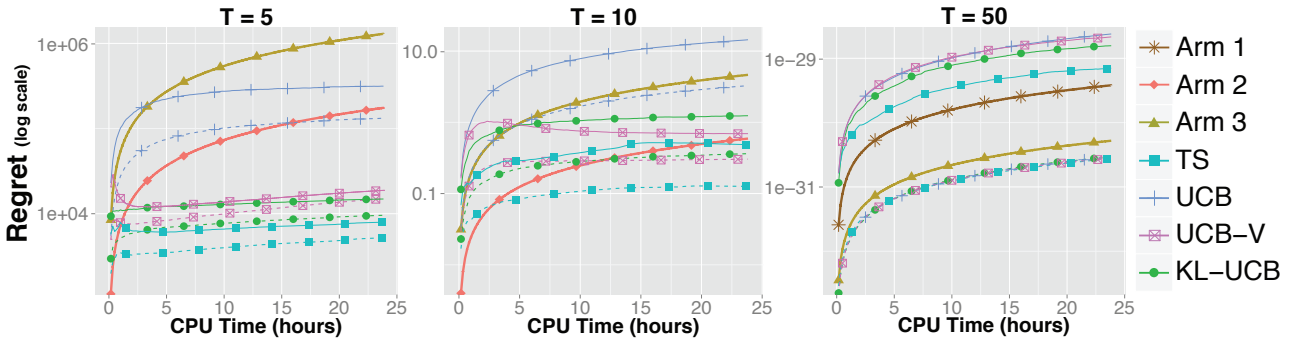

Figure 3. Plots showing the average regret (over 20 independent runs) of bandit allocators on the Logistic Regression model. T is training sample size, "Arm 1/2/3" indicates each fixed estimator; and the one missing from a figure is the best for that setting. The solid lines indicate the performance of combining observations uniformly, whereas the dashed lines indicate the performance of combining observations using *inverse variance weights*.

model on different sized subsets of the 8-dimensional *Pima Indian diabetes* UCI data set (Bache & Lichman, 2013). We consider the problem of allocating resources between three AIS estimators that differ only in the number of annealing steps they use; namely, 400, 2000, and 8000 steps. In each case, we fix the annealing schedule using the *power of 4* heuristic suggested by (Kuss & Rasmussen, 2005), with a single slice sampling MCMC transition used at each step (Neal, 2003) (this entails one "step" in each of the 8 dimensions); see Appendix G for further details.

A key challenge in this scenario is that the computational costs associated with each arm differ substantially, and, because slice sampling uses an internal rejection sampler, these costs are stochastic. To account for these costs we directly use *elapsed CPU-time* when drawing a sample from each estimator, as reported by the JAVA VM. This choice reflects the true underlying cost and is particularly convenient since it does not require the practitioner to implement special accounting functionality. Since we do not expect this cost to correlate with the sample returns, we use the independent costs payoff formulation from Section 5: $-\frac{1}{4}(D_{k,2m} + D_{k,2m+1})(X_{k,2m} - X_{k,2m+1})^2$.

The results for the different allocation strategies for training sets of size 5, 10, and 50 are shown in Figure 3. Perhaps the most striking result is the performance improvement achieved by the *nonuniformly combined estimators*, which are indicated by the dashed lines. These estimators do not change the underlying allocation; instead they improve the final combined estimate by weighting each observation inversely proportional to the sample variance of the estimator that produced it. This performance improvement is an artifact of the nonuniform cost setting, since arms that are very close in terms of $V_k\delta_k$ can still have considerably different variances, which is especially true for AIS. Also observe that no one arm is optimal for all three training set sizes, consequently, we can see that bandit allocation (Thompson sampling in particular) is able to outperform any static strategy. In practice, this implies that even after exhaustive parameter tuning, automated adaptation can still offer considerable benefits simply due to changes in the data.

7. Conclusion

In this paper we have introduced a new sequential decision making strategy for competing with the best consistent Monte Carlo estimator in a finite pool. When each base estimator produces unbiased values at the same cost, we have shown that the sequential estimation problem maps to a corresponding bandit problem, allowing future improvements in bandit algorithms to be transfered to combining unbiased estimators. We have also shown a weaker reduction for problems where the different estimators take different (possibly random) time to produce an observation.

We expect this work to inspire further research in the area. For example, one may consider combining not only finitely many, but infinitely many estimators using appropriate bandit techniques (Bubeck et al., 2011), and/or exploit the fact that the observation from one estimator may reveal information about the variance of others. This is the case for example when the samplers use importance sampling, leading to the (new) stochastic variant of the problem known as "bandits with side-observations" (Mannor & Shamir, 2011; Alon et al., 2013). However, much work remains to be done, such studying in detail the use variance weighted estimators, dealing with continuous families of estimators, or a more thorough empirical investigation of the alternatives available.

Acknowledgements

This work was supported by the Alberta Innovates Technology Futures and NSERC. Part of this work was done while Cs.Sz. was visiting Technion, Haifa and MSR, Redmond, whose support and hospitality are greatly acknowledged.

References

Agrawal, S. and Goyal, N. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pp. 39.1–39.26, 2012.

Alon, N., Cesa-Bianchi, N., Gentile, C., and Mansour, Y. From bandits to experts: A tale of domination and independence. *NIPS*, pp. 1610–1618, 2013.

Arouna, B. Adaptive Monte Carlo technique, a variance reduction technique. *Monte Carlo Methods and Applications*, 2004.

Audibert, J. and Bubeck, S. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010.

Audibert, J., Munos, R., and Szepesvári, Cs. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002a.

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

Bache, K. and Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, Cs. X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, June 2011.

Burnetas, A. and Katehakis, M. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17:122–142, 1996.

Cappé, O., Garivier, A., Maillard, O., Munos, R., and Stoltz, G. Kullback-Leibler upper confidence bounds for optimal sequential decision making. *Annals of Statistics*, 41(3):1516–1541, 2013.

Carpentier, A. and Munos, R. Finite time analysis of stratified sampling for Monte Carlo. In *NIPS-24*, pp. 1278–1286, 2011.

Carpentier, A. and Munos, R. Minimax number of strata for online stratified sampling given noisy samples. In *Algorithmic Learning Theory*, pp. 229–244, 2012a.

Carpentier, A. and Munos, R. Adaptive stratified sampling for Monte-Carlo integration of differentiable functions. In *NIPS*, pp. 251–259, 2012b.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Chapelle, O. and Li, L. An empirical evaluation of Thompson sampling. In *NIPS-24*, pp. 2249–2257, 2011.

Cox, J., Ingersoll Jr, J., and Ross, S. A theory of the term structure of interest rates. *Econometrica: Journal of the Econometric Society*, pp. 385–407, 1985.

Del Moral, P., Doucet, A., and Jasra, A. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:411–436, 2006.

Douc, R., Guillin, A., Marin, J., and Robert, C. Minimum variance importance sampling via population Monte Carlo. *ESAIM: Probability and Statistics*, 11:427–447, 2007.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite time analysis. In *Algorithmic Learning Theory*, pp. 199–213, 2012.

Kuss, M. and Rasmussen, C. Assessing approximate inference for binary gaussian process classification. *The Journal of Machine Learning Research*, 6:1679–1704, 2005.

Lai, T. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Mannor, S. and Shamir, O. From bandits to experts: On the value of side-observations. In *NIPS*, pp. 684–692, 2011.

May, B. and Leslie, D. Simulation studies in optimistic Bayesian sampling in contextual-bandit problems. Technical report, 11: 02, Statistics Group, Department of Mathematics, University of Bristol, 2011.

Neal, R. Annealed importance sampling. Technical report, University of Toronto, 2001.

Neal, R. Slice sampling. *Annals of statistics*, pp. 705–741, 2003.

Neal, R. Estimating ratios of normalizing constants using linked importance sampling. Technical report, University of Toronto, 2005.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58: 527–535, 1952.

Robert, C. *Bayesian computational methods*. Springer, 2012.

Robert, C. and Casella, G. *Monte Carlo Statistical Methods*. Springer-Verlag New York, 2005.

Thompson, W. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.