Making Simulated Annealing Sample Efficient for Discrete Stochastic Optimization

Suhail M. Shah *

September 15, 2020

Abstract

We study the regret of simulated annealing (SA) based approaches to solving discrete stochastic optimization problems. The main theoretical conclusion is that the regret of the simulated annealing algorithm, with either noisy or noiseless observations, depends primarily upon the rate of the convergence of the associated Gibbs measure to the optimal states. In contrast to previous works, we show that SA does not need an increased estimation effort (number of pulls/samples of the selected arm/solution per round for a finite horizon $n$) with noisy observations to converge in probability. By simple modifications, we can make the total number of samples per iteration required for convergence (in probability) to scale as $O(n)$. Additionally, we show that a simulated annealing inspired heuristic can solve the problem of stochastic multi-armed bandits (MAB), by which we mean that it suffers a $O(\log n)$ regret. Thus, our contention is that SA should be considered as a viable candidate for inclusion into the family of efficient exploration heuristics for bandit and discrete stochastic optimization problems.

1 Introduction

Consider the following optimization problem:

$$\min_{a \in A} \mu_a := \mathbb{E}_\omega f(a, \omega),$$

(1)

where $A$ is a finite set (which may have some inherent topological structure), $\omega$ is a random variable and $f : A \to \mathbb{R}$ is a real valued bounded function. The aim in stochastic optimization is to study and solve (1) in an efficient manner. Finding suitable solution to this problem can be quite hard, since one may lack a direct access to the underlying probability distribution to compute $\mu_a$. In such situations, one may be faced with the task of optimizing only through samples $f(a, \omega)$. Given this, the primary concern in stochastic optimization is to obtain a good solution, i.e., $\mu_{avg}$ is as small as possible for any algorithmic output $a_T$.

The above problem can also be studied in a ‘bandit setting’ ([1], [2], [3], [4]) where the emphasis is placed on the total loss or ‘regret’ incurred by the algorithm. A stochastic bandit is a collection of distributions $\nu := (P_a : a \in A)$, where $A$ is the set of available actions with mean-payoff $\mu_a := \int x \, dP_a(x)$. During any time instance $t$, referred to as a round, the learner chooses any action $a_t \in A$, interacts with the environment following which a loss $X_t$ (or for a maximization problem, a reward) is revealed to the learner sampled from $P_{a_t}$. For instance, the learner may select an arm $a_1 \in A$ in the first round and incur the loss $X_1$ drawn at random according to $P_{a_1}$. In any subsequent round $n$, the learner can select an arm $a_n$ based on all prior information available till the current round, i.e. the variables $(a_1, X_1, \cdots, a_{n-1}, X_{n-1})$ and incur a loss $X_n$ independently of $(a_1, X_1, \cdots, a_{n-1}, X_{n-1})$ given

*The author is with the department of Electrical Engineering, Indian Institute of Technology-Bombay, India-400076. (e-mail: suhailshah2005@lgmail.com).
Any strategy (formally a sequence of probability kernels defined on an appropriate probability space, see definition 4.7 in [4]) which instructs on how to select an arm at round \( n \) based on previous information is referred to as a policy. In bandit optimization, the goal of the learner is to construct policies that minimize the expected cumulative regret, which is the difference between the expected loss incurred and the loss incurred (in expectation) by choosing the optimal action. More formally, let \( a^* := \arg \min_{a \in A} \mu_a \) be the optimal loss for any round. At round \( n \), the cumulative (pseudo) regret of a learner will then be

\[
R_n = \mathbb{E} \sum_{t=1}^{n} X_t - n\mu^*_{a^*},
\]

where the expectation is over the learner’s policy (see Section 4.4, [4] for a formal treatment).

Bandit optimization encapsulates the exploration/utilization trade-off encountered in many practical situations. Studying the regret of the algorithm can give us a good idea of the number of estimates/samples an algorithm requires (sample complexity) to reach a suitable solution. The sample complexity may take precedence over other measures of performance, when, for instance, obtaining such samples may entail a lengthy simulation. This is frequently the case for most practical applications of importance, including reinforcement learning.

In this paper, we study the regret of policies based on Simulated Annealing (SA). To the best of our knowledge, there has been no formal study of SA in a bandit setting. Even for the noiseless setting, which is of importance for combinatorial optimization among other things, a regret analysis is missing. Thus, there exists a significant gap of knowledge in understanding how to efficiently utilize function samples when working with noisy estimates in SA. We take a step in this direction and show that SA can indeed search the solution space quite efficiently.

### 1.1 Related Literature and Contributions

In this subsection, we briefly cover the prior work done on simulated annealing and clarify the motivation and contributions of the present work. SA was originally proposed in [5] for finding globally optimum configurations in large NP-complete problems. The foundational work on the theory of simulated annealing was done in the eighties. We briefly mention some of the works here.

In the seminal paper of [7] the remarkable notion of depth was introduced and the algorithm was shown to converge under the assumptions of reversibility. In particular, [7] was the first to pin down the minimum \( \gamma \) for which the algorithm would converge in probability for a cooling schedule of the form \( T_t = \frac{\gamma}{\log(t+1)} \), where \( t \) is the iteration count. [8] provided finite time convergence bounds on the probability of selecting any state in terms of the graph radius. This work, along with [9], was one of the first to use the theory of time inhomogeneous Markov chains to study finite time performance. [10] used the perturbation theory of Markov chains to study the convergence of SA while [11] studied SA using Sobolev inequalities. The latter obtained the same characterization as [7] of the cooling schedule with remarkably concise proofs by using the Dirichlet machinery. We will adopt their approach in the later part of this work.

The program of analysing SA with noisy observations began with [12]. It was established that if the noise variance (in the sample observations) decreased by at least a rate of \( o(T_t^{-1}) = o(\log t) \), the asymptotic behaviour would be unaffected (the cooling schedule bounds remained the same as in [7]). In [13], a convergence in total variation was established with the requirement that the number of samples per iteration increase to the order \( \mathcal{O}(n^{2+\epsilon}) \), \( \epsilon > 0 \). A recent work by [14], clarified that \( n^{\alpha} \), \( \alpha > 0 \) samples could also be sufficient to establish convergence in probability. This was also the first work, to the best of our knowledge, that provided upper bounds on the convergence rate of SA with noisy observations. There are other works, too numerous to mention here, that also study simulated annealing in a noisy setting. We refer the reader to Section 3 in [15] for a more comprehensive account.

\(^1\)For the sake of brevity, this account is incomplete. We refer the reader to [6] (and the references within) for a more detailed history.
Table 1: (i) We assume that the time horizon $n$ is known to the learner. An infinite horizon may be handled via the doubling trick (see Section 2.3, [16]). (ii) The total number of samples is actually random for [14], we quote the expected number (see eq. (28) in ibid). (iii) $\gamma^*$ is the critical depth, see definition 4. (iv) $\Delta_{\min} := \min_a (\mu_a - \mu_{a^*})$.

Contributions of the present work: In this work, we study and establish bounds on the regret of SA with both noisy and noiseless observations to show that it can be used as an efficient exploration heuristic when one is faced with a fixed sample budget. For the noisy setting, Table 1 details the total number of samples required to guarantee convergence in probability for SA as proposed in previous works. We note that all of these works require that the number of samples required during any round keep on increasing (sometimes as much as $O(t^2)$). This does not paint a flattering portrait of SA and makes it seem unappealing in comparison to the state of the art. Nevertheless, as we shall see, SA can be a lot more sample efficient (in fact, linearly dependent on the horizon $n$), with simple modifications like keeping track of an empirical average of past samples at each arm and allocating a very small fraction of the budget to uniformly search all the arms. In defence of the previous works, the analysis is done assuming a large (but finite) action space for which the suggested modifications to SA may not be entirely suitable (e.g., uniformly searching the sample space or a finite horizon). But in such situations, where one usually has some form of regularity on the cost function, one can use SA as a sub-routine in a more structured approach to searching continuous arm sets (see remark 9), or else switch to a continuous version of SA. The finite time horizon issue can be handled via the doubling trick (see Section 2.3, [16]). Our contributions are as follows:

(i) In section 3, we establish an upper bound on the regret of SA with exact observations. This, in particular, is relevant for many combinatorial optimization problems where SA is used routinely to obtain (globally) optimal solutions ([17], [18], [19]). We show that there are two components in the regret bound: (i) A fast decaying transient component which depends upon the energy landscape of the function being minimized. (ii) A ‘steady state’ component whose decay is governed by how fast the Gibbs measure converges to the state of the optimal states.

(ii) In Section 4, we propose an SA inspired heuristic to solve the multi armed bandit problem. We show that the regret can be upper bounded by $O(\log n)$, where $n$ is the time horizon. This shows that SA can solve the multi armed bandit problem with optimal rates within constant factors. We note that this setting can be considered as a fully connected setting for SA, so one should intuitively expect better performance.

(iii) In Section 5, we study SA in the noisy case for general graphical structures (with some assumptions). This is of importance for graph based optimization and bandit optimization with graphical constraints, wherein during any round the choice of the next arm is constrained to a certain subset of the arm set depending upon the current arm. Our main theoretical conclusion is that the regret of the algorithm is primarily dependent upon the rate of convergence of the associated Gibbs measure to the optimal states if each arm is allocated a small fraction of the budget beforehand.
The theory is further substantiated with numerical experiments reported in Appendix A. The performance is benchmarked against standard bandit algorithms known to achieve optimal regret.

**Notation**: We have recalled all the notation used throughout the paper in Appendix B.

## 2 Problem Formulation and the Algorithm

The basic procedure of simulated annealing consists of:

- A finite action/arm/solution set \( A \) with cardinality \( |A| = k \), to which is associated a real valued, bounded cost function \( \mu : A \to \mathbb{R} \). The action with the least cost is assumed to be unique and denoted by \( a^* \). In the noisy setting, one does not know the exact value of \( \mu_a \) and has only access to sample observations (drawn according to \( P_a \)).

- There is a bijective correspondence between the elements of \( A \), the set of all the possible configurations of the optimization problem and the nodes of the graph. We denote the latter by \( G := \{A, E\} \), where \( E \) is the set of edges. Accordingly, for each arm \( a \in A \), there is a set \( N(a) \subset A/\{a\} \) called the neighbourhood set of \( a \). It is assumed that \( a \in N(a') \) if and only if \( a' \in N(a) \).

- For any \( a \in A \), a collection of positive co-efficients \( g(a, a'), a' \in N(a) \) such that

\[
\frac{g(a, a')}{g(a)} = 1, \quad \text{where} \quad g(a) := \sum_{a' \in N(a)} g(a, a').
\]

Here, \( \frac{g(a, a')}{g(a)} \) represents the probability of selecting a neighbouring candidate arm \( a' \) for transition, when the present state of the algorithm is \( a \).

- A cooling schedule \( T : \mathbb{N} \to (0, \infty) \), assumed to be non increasing. \( T(n) \) is also referred to as the temperature of the algorithm.

Given these parameters, the SA algorithm consists of a discrete time inhomogeneous Markov chain, whose transition mechanism (for the noiseless case), denoted by \( P(n) \) for temperature \( T_n \), can be described as:

\[
P_{x,y}(n) = \begin{cases} 
0, & \text{if } y /\notin N(x) \\
\frac{g(x, y)}{g(x)} \exp\left\{ -\frac{(\mu_y - \mu_x)^+}{T_n} \right\}, & \text{otherwise}
\end{cases}
\]

and

\[
P_{x,x}(n) = 1 - \sum_{i \in N(x)} P_{x,i}(n),
\]

where \((x)^+ := \max(0, x)\). It is quite natural to assume a connected graphical structure for modelling problems to be solved via SA (see Section 3, [6]). We recall that \( g(a, a')/\sum_{i \in N(a)} g(a, i) \) represents the probability of selecting the state \( a' \in N(a) \), given that the current state is \( a \). A common practice is to set this equal to \( \frac{1}{|N(a)|} \) for all \( a' \in N(a) \). Let

\[
\pi_i(n) := \frac{g(i) \exp(-\frac{\mu_i}{T_n})}{\sum_{j} g(j) \exp(-\frac{\mu_j}{T_n})},
\]

We assume that

\[
g(a, a') = g(a', a) = 1
\]

This assumption is without any loss of generality for all the results in this work.
whenever $a$ and $a'$ are neighbours. The above condition guarantees that $\pi^0$ is the stationary distribution of $P(n)$, if we freeze the temperature at $T_n$ (see Proposition 3.1, [8]). Additionally, this also makes sure that the detailed balance equation holds, so the time homogeneous Markov chain $P(n)$ is time reversible. We recall all the assumptions taken thus far in the following:

**Assumptions:**

(i) We assume that there exists a unique arm $a^*$, such that

$$a^* = \arg \min_{j \in A} \mu_j.$$

(ii) The directed graph $G$ is connected. For all neighbours $a, a' \in A$, we have $g(a, a') = g(a', a)$.

(iii) The noise in the observations is assumed to be sub-Gaussian. This means that for any $\epsilon > 0$,

$$\mathbb{P}(X \geq \epsilon) \leq \exp \left( -\frac{\epsilon^2}{2\sigma^2} \right)$$

for some $\sigma > 0$, where $X$ is a generic noise variable. The family of sub-Gaussian bandits with $k$-arms having variance $\sigma^2$ is denoted by $\mathcal{E}_{SG}^k(\sigma^2)$.

**2.1 An asymptotic perspective of SA with noisy observations**

It is instructive here, before we proceed to perform a theoretical analysis, to provide some intuition as to why SA with noisy observations should in principle work. Consider any two arms $i$ and $j \in N(i)$ with distinct means $\mu_i < \mu_j$ and set $\Delta_{ij} := \mu_j - \mu_i$. Since we do not know $\mu_i$ and $\mu_j$, we will use the empirical mean estimates in their stead in the transition mechanism (2). Accordingly, let $\hat{\mu}_i(n)$ and $\hat{\mu}_j(n)$ be the calculated empirical mean at arm $i$ and $j$. Consider the event $G_n$ that $i$ and $j$ have been pulled $T_i(n)$ and $T_j(n)$ times respectively up to round $n$. Then, since our arms are sub-Gaussian, the Cramer Chernoff method gives the following bound,

$$\mathbb{P}\left( \left| \hat{\mu}_p(n) - \mu_p \right| \geq \frac{\Delta_{ij}}{2} \right) \leq 2e^{-\frac{T_i(n)\Delta_{ij}^2}{16\sigma^2}} \quad \text{for } p = i, j,$$

where $T_i(n) = \min\{T_i(n), T_j(n)\}$. Suppose that $T_i(n) = \Omega\left( \frac{16\sigma^2 \log n}{\Delta_{ij}^2} \right)$, then

$$\mathbb{P}(G_n) := \mathbb{P}\left( \left| \hat{\mu}_p(n) - \mu_p \right| \leq \frac{\Delta_{ij}}{2} \right) \geq 1 - \frac{2}{n} \quad \text{for } p = i, j.$$

Then, for event $G_n$, which occurs with high probability for large $n$, we have

$$P(a_{n+1} = j | a_n = i) = \exp \left( -\frac{(\hat{\mu}_j - \hat{\mu}_i)^+}{T_n} \right)$$

$$\sim \exp \left( -\frac{(\mu_j - \Delta_{ij}^2 - \mu_i + \Delta_{ij}^2)}{2T_n} \right)$$

$$\sim \exp \left( -\frac{(\mu_j - \mu_i)}{T_n} \right).$$

Conversely, $P(a_{n+1} = i | a_n = j) = 1$ with high probability. Thus, for fixed time horizon $n$, $\Omega(\log n)$ pulls for each arm appear to be sufficient to distinguish the arms with high probability and one could guess that the convergence behaviour of SA would be somewhat similar to a noiseless setting. Admittedly, while this two arm bandit instance may only present a highly simplified view of the task.
Figure 1: The event $\hat{\mu}_p(n) \in [\mu_p + \Delta_{ij}/2, \mu_p + \Delta_{ij}/2]$, $p = i, j$, occurs with high probability for $T_p(n) = \log n$.

At hand, the above facts do to some extent (especially for the standard MAB) underpin the main aspects of the theory. It is not surprising the original study by [12] came to the conclusion that the post sampling noise variance should decrease at a minimum rate of $o(\log n)$ (see proposition 3.1 in [12]) for SA to succeed despite noisy observations.

The SA algorithm can be thought of as a (time inhomogeneous) random walk on $A$, which gradually gets reinforced, via the Glauber dynamics, to prefer the low energy (mean) states (arms). So, it lends itself naturally to be analyzed as a vertex reinforced random walk (VRRW) [20], [21]. In what follows, we look at its workings through the lens of stochastic approximation. This will help us relate its asymptotic behaviour to the widely studied field of evolutionary games which may have broader implications for analysis of algorithms besides SA. Let $\{a_n\}_{n \geq 0}$ denote the time inhomogeneous Markov chain at hand. Also, let $a_n \in A_n$ denote the value of the arm pulled at time $n$ and $z_i(n)$ denote the total loss incurred by taking action $i$ up to time $n$, so that

$$z_i(n) = \sum_{m=1}^{n} I\{a_m = i\} \tilde{\mu}_i,$$

where $\tilde{\mu}_i = \mu_i + \xi_i$ with $\xi_i$ being sub-Gaussian. Let $x(n)$ denote the empirical frequency, i.e. the $i$'th component of $x(n)$ is the fraction of the total loss till time $n$, so that

$$x(n) = \frac{z(n)}{n},$$

We analyze the frequency with which any arm is visited by studying $x(n)$ as a stochastic approximation scheme. Accordingly, we re-write the above iteration component-wise as:

$$x_i(n) = \left(1 - \frac{1}{n}\right)x_i(n-1) + \frac{I\{a_{n-1} = i\}}{n} \tilde{\mu}_i \quad (4)$$

or equivalently,

$$x_i(n) = \left(1 - \frac{1}{n}\right)x_i(n-1) + \frac{1}{n} \left(I(a_{n-1} = i)\mu_i + M_n\right),$$

where

$$M_n := I\{a_{n-1} = i\} \tilde{\mu}_i - I(a_{n-1} = i)\mu_i = I\{a_{n-1} = i\}(\tilde{\mu}_i - \mu_i).$$
is a Martingale difference sequence. The probability transition mechanism in (4) is given by,

\[ P(a_{n+1} = j | a_n = i) = \mathbb{I}\{j \in A(a_n)\} \left( \frac{g(i,j)}{g(i)} \exp \left( - \frac{(\hat{\mu}_j - \hat{\mu}_i)^+}{T_n} \right) \right). \]  

(5)

The scaled version of the previous iteration can be written as follows:

\[ y_i(n) = \left(1 - \frac{1}{n}\right)y_i(n-1) + \frac{1}{n} \left( \mathbb{1}(a_{n-1} = i) + M_i(n) \right), \]  

(6)

where \( y_i := \frac{x_i}{\mu_i} \) and \( M_i(n) := \frac{\bar{m}_i}{\mu_i} \). Let

\[ \epsilon_n := P(a_n = j | a_{n-1} = i) - \bar{P}(\bar{a}_n = j | \bar{a}_{n-1} = i), \]

where \( \bar{P}(\bar{a}_n = j | \bar{a}_{n-1} = i) \) is given by (2) and \( \{\bar{a}_n\} \) is the process generated by the transition mechanism \( \bar{P} \). To be more precise, \( \bar{P}(\bar{a}_n = j | \bar{a}_{n-1} = i) \) is (5) with noiseless estimates (i.e., \( \hat{\mu}_i = \mu_i \), for all \( i \in A \)). So, (6) can be written as:

\[ y_i(n) = \left(1 - \frac{1}{n}\right)y_i(n-1) + \frac{1}{n} \left( \mathbb{1}(\bar{a}_{n-1} = i) + \epsilon_n + M_i(n) \right). \]

The difference of the above equation from (6) is that (provided \( \epsilon_n \to 0 \), see section 5.1 for a bound on this quantity), the probability of transition is governed by \( \bar{P}(\cdot | \cdot) \) instead of \( P(\cdot | \cdot) \). For SA to succeed, we should have \( y(n) \to y^* \), where

\[ y^*_a = 1 \quad \text{and} \quad y^*_a = 0 \quad \forall a \neq a^*. \]  

(7)

For a fixed \( T \), by the standard analysis of stochastic approximation with Markov noise ([22], Chapter 6), the sequence \( y(n) \) will almost surely track the asymptotic behavior of the o.d.e.,

\[ \dot{y}(t) = \pi_T(y(t)) - y(t), \]  

(8)

where \( \pi_T(y) \) is the \( y \)-dependant stationary distribution of time homogeneous chain (2). The dynamics (8) is a special case of the replicator dynamics of evolutionary biology [23], [24], evolving on the unit simplex \( \Delta^k \). Plugging the value of (quasi) stationary distribution for (2) in the above shows that any equilibrium point, say \( y^* \), satisfies:

\[ y^*_i \propto \exp \left( - \frac{y^*_i}{T} \right) \sum_{j \in A} \exp \left( - \frac{y^*_j}{T} \right). \]

Letting \( T \to 0 \) will lead to (7). This again leads us to the conclusion that if the noise variance goes to zero (i.e. \( \epsilon_n \to 0 \) here), the noisy observations will not affect the algorithm.

### 3 Noiseless Case

In this section we perform a regret analysis of the simulated annealing algorithm with exact observations. We first briefly review the mode of convergence for time inhomogeneous Markov Chains and introduce some theoretical constructs that will also be of use to us later. We presently recall the notion of weak ergodic convergence. Let

\[ P(m, n) := \prod_{t=m}^{n-1} P(t) \]

\[ ^3\text{Actually, the time schedule does not have to be constant, the claim will hold for } T(n) = \frac{\gamma}{\log n}. \]
for any time inhomogeneous chain with transition mechanism \( \{P(t)\}_{t \geq 0} \).

**Definition 1 (Weak Ergodic Convergence)** A time-inhomogeneous Markov chain is said to be weakly ergodic if, for all \( m \)[4]
\[
\lim_{n \to \infty} \sup_{\nu(0), \bar{\nu}(0)} \|\nu(0)P(m, n) - \bar{\nu}(0)P(m, n)\| = 0.
\]

The above definition implies a tendency towards equality of the rows of \( P(m, n) \), i.e., a ‘loss of memory’ of the initial conditions. The investigation of conditions under which weak ergodicity holds is aided by the introduction of the following quantity:

**Definition 2 (Coefficient of Ergodicity)** Given a stochastic matrix \( P \), its coefficient of ergodicity, denoted by \( \kappa(P) \), is defined to be:
\[
\kappa(P) = \frac{1}{2} \max_{i,j} \sum_{s=1}^{k} |P_{is} - P_{js}| = 1 - \min_{i,j} \sum_{s=1}^{k} \min\{P_{is}, P_{js}\}.
\]

We recommend the reader to [25] or [26] for a detailed discussion of the importance of using ergodicity coefficients to study time inhomogeneous Markov chains. The main two properties which we will use are (Theorem V.2.4, [25]):
\[
\kappa(PQ) \leq \kappa(P)\kappa(Q) \quad \text{and} \quad \|PQ\| \leq \|P\|\|Q\| \tag{9}
\]
for any stochastic matrices \( P, Q \). We next define the concept of critical depth, first introduced in [7]:

**Definition 3 (Critical Depth)** Let \( p = (i_0, \cdots, i_n) \) be any path in \( \mathcal{G} \), i.e., for any \( i_p \in \mathcal{P}, i_{p+1} \in \mathcal{N}(i_p) \). Let
\[
\text{Elev}(p) := \max_{0 \leq m \leq n} \mu_{i_m} \quad \text{and} \quad e(p) := \text{Elev}(p) - \mu_{i_0} - \mu_{i_n}.
\]

Let \( \mathcal{E}(p) = \sup_{p \in \mathcal{P}} e(p) \), for any choice of path \( \mathcal{P} \). Also, let \( e \) be the minimum value of \( \mathcal{E}(p) \) as \( \mathcal{P} \) runs over all selections of allowable paths. Then, the **critical depth**, denoted by \( \gamma^* \), is defined to be \( \gamma^* := e + \mu_{a^*} \).

The formal statement of critical depth is a bit hard to digest. A high level intuition for \( \gamma^* \) is that it can be thought of as the least upper bound on the energy barrier that one has to climb in order to reach the optimal arm. To be more precise, let us say that arm \( i \) communicates with the optimal arm \( a^* \) at height \( h \), if there exists a path starting at \( i \) and ending at \( a^* \) and such that the largest value of any arm encountered along the path is \( \mu_i + h \). Then, \( \gamma^* \) is the smallest number such that every \( i \in \mathcal{A} \) communicates with \( a^* \) at height \( \gamma^* \). We will later see how the ergodicity co-efficient is related to the critical depth (see eq. (19)). Let
\[
L := \max_{i \in \mathcal{A}} \max_{j \in \mathcal{N}(i)} |\mu_j - \mu_i| \quad \text{and} \quad R := \min_{i \in \mathcal{A}} \min_{j \in \mathcal{N}(i)} g(i, j) \frac{g(i, j)}{g(i)}
\]

We have for any \( i, j \in \mathcal{N}(i) \) in (2),
\[
P_{ij}(m) \geq Re^{-\frac{L}{m}}. \tag{10}
\]

From the definition of ergodicity coefficient, we have
\[
\kappa(P_{ij}(m - \tau, m)) \leq 1 - \min_{i,j} \left\{ \sum_{s=1}^{k} \min \left\{ P_{is}(m - \tau, m), P_{js}(m - \tau, m) \right\} \right\}
\leq 1 - \max_{i} \min_{s} P_{is}(m - \tau, m).
\]

---

[4] \( \|x\| \) is always the \( \ell_1 \) norm and \( \|A\| \) the infinity norm in this paper (see Appendix B for definitions).
We have,

\[ \kappa(P_{ij}(m - \tau, m)) \leq 1 - R^\tau \exp(-\frac{\tau L}{T_{m-1}}). \]  

(11)

If for \( T_m := \frac{\gamma}{\log m} \), the condition \( \frac{\tau L}{\gamma} \leq 1 \) is satisfied, then the chain corresponding to (2) is weakly ergodic (see Theorem 5.1, [8]). Our aim will be to get an upper bound on how fast such a convergence takes place and its dependence on \( \gamma \), by establishing a regret bound. To do so, we first prove a preliminary result to bound the probability of selecting a sub-optimal arm.

**Lemma 1.** Let the cooling schedule be of the parametric form \( T_i = \frac{\gamma}{\log((i+1))} \). Then, there exists a constant \( m_0 < \infty \) such that the probability of selecting a sub-optimal arm for any \( t \geq m_0 \) is bounded by

\[ P(a_t \in A/A^*) \leq \frac{2 \exp\left(\frac{R^\tau}{\gamma} \left(1 - \frac{1}{m} + 1\right)^{1 - \frac{1}{m}}\right)}{\exp\left(\frac{R^\tau}{\gamma} \left(1 - \frac{1}{m} + 1\right)^{1 - \frac{1}{m}}\right)} + \sum_{a \in A/A^*} \frac{4g(a)}{g(a^*)} (t + 1)^{-\frac{m - m^*}{\gamma}}, \]

(12)

so that,

\[ P(a_t \in A/A^*) = O\left(\exp\left(-\frac{R^\tau}{\gamma} \left(1 - \frac{1}{m} + 1\right)^{1 - \frac{1}{m}}\right) + \sum_{a \in A/A^*} \frac{1}{(t + t_0)^{\frac{m - m^*}{\gamma}}}\right). \]

**Proof.** Let \( \nu(0) \) denote the initial probability distribution. The probability of selecting a bad arm can be written as:

\[ P(a_t \in A/A^*) = \sum_{i \in A/A^*} [\nu(0)P(0, t)]_i, \]

\[ \leq \|\nu(0)P(0, t) - e^*\|, \]

where \( e^* \) is the unit vector with all entries zero except the one corresponding to \( a^* \). To prove the result we upper bound the term \( \|\nu(0)P(0, t) - e^*\| \). Accordingly, we first decompose this quantity as:

\[ \|\nu(0)P(0, t) - e^*\| \leq \|\nu(0)P(0, t) - \nu(0)Q(t)\| + \|\nu(0)Q(t) - e^*\|. \]

(13)

where:

\[ Q(t) = \begin{bmatrix} \vdots \\ \pi(t) \\ \vdots \end{bmatrix} \]

with \( \pi(t) \) being the quasi-stationary distribution defined in (3).

**Term I:** We have,

\[ \|\nu(0)P(0, t) - \nu(0)Q(t)\| \leq \|P(0, t) - Q(t)\|. \]

Furthermore, we have for any \( 0 < \tilde{m} \leq t \), the following decomposition

\[ \|P(0, t) - Q(t)\| \leq \|P(0, t) - Q(\tilde{m})P(\tilde{m}, t)\| + \|Q(\tilde{m})P(\tilde{m}, t) - Q(t)\|. \]

(14)

We deal with each of these terms separately. Consider the first term:

\[ \|P(0, t) - Q(\tilde{m})P(\tilde{m}, t)\| \leq \|P(0, \tilde{m}) - Q(\tilde{m})\| \kappa(P(\tilde{m}, t)) \]

\[ \leq 2 \prod_{k=\tau_0}^{\frac{\gamma}{\gamma} - 1} \kappa(P(k\tau_0, (k + 1))) \]

9
for some \( t_0 \tau = \tilde{m} \), assuming \( t_0 \in \mathbb{Z}^+ \) w.l.o.g. Since \( \kappa(P(m - \tau, m)) \leq 1 - R^\tau \exp\left(-\frac{\tau L}{T_{m-1}}\right) \) (see eq. (11)), we get

\[
\|P(0, t) - Q(\tilde{m})P(\tilde{m}, t)\| \leq 2 \prod_{k=t_0}^{\tilde{m}-1} \left[ 1 - R^\tau \exp\left(-\frac{\tau L}{T_{k\tau + \tau - 1}}\right) \right]
\]

\[
= 2 \prod_{k=t_0}^{\tilde{m}-1} \left[ 1 - R^\tau \exp\left(-\frac{\tau L}{\gamma \log(k\tau + \tau)}\right) \right]
\]

\[
= 2 \prod_{k=t_0}^{\tilde{m}-1} \left[ 1 - \frac{R^\tau}{(k\tau + \tau)^{\frac{1}{\gamma}}} \right]
\]

\[
\leq 2 \exp\left(-R^\tau \sum_{k=t_0}^{\tilde{m}-1} \frac{1}{(k+1)\tau^{\frac{1}{\gamma}}} \right)
\]

\[
= 2 \exp\left(-\frac{R^\tau}{\tau^{\frac{1}{\gamma}}} \sum_{k=t_0}^{\tilde{m}-1} \frac{1}{(k+1)\tau^{\frac{1}{\gamma}}} \right)
\]

\[
\leq 2 \exp\left(-\frac{R^\tau}{\tau^{\frac{1}{\gamma}}(1 - \frac{1}{\gamma})} \left( \left(\frac{t}{\tau}\right)^{1 - \frac{1}{\gamma}} - (t_0 + 1)^{1 - \frac{1}{\gamma}} \right) \right)
\]

\[
\|P(0, t) - Q(\tilde{m})P(\tilde{m}, t)\| \leq \frac{2 \exp\left(-\frac{R^\tau}{\tau^{\frac{1}{\gamma}}(1 - \frac{1}{\gamma})} \left( \left(\frac{t}{\tau}\right)^{1 - \frac{1}{\gamma}} - (t_0 + 1)^{1 - \frac{1}{\gamma}} \right) \right)}{\exp\left(-\frac{R^\tau}{\tau^{\frac{1}{\gamma}}t^{1 - \frac{1}{\gamma}}} \right)}. \tag{15}
\]

We now consider the second term in (14). We note that stationarity of \( \pi_t \) implies \( Q(t)P(t) = Q(t) \).

So,

\[
Q(l)P(l, m) = Q(l)P(l, m) - Q(l + 1)P(l + 1, m) + Q(l + 1)P(l + 1, m)
\]

\[
= (Q(l) - Q(l + 1))P(l + 1, m) + Q(l + 1)P(l + 1, m).
\]

Recursion on the above equation gives

\[
Q(l)P(l, m) = (Q(l) - Q(l + 1))P(l + 1, m) + (Q(l + 1) - Q(l + 2))P(l + 2, m)
\]

\[
+ Q(l + 2)P(l + 2, m)
\]

\[
= \sum_{t=l}^{m-1} (Q(t) - Q(t + 1)) + Q(m).
\]

Take \( l = \tilde{m} \) and \( m = t \) in the above to get:

\[
\|Q(\tilde{m})P(\tilde{m}, t) - Q(t)\| \leq \sum_{t=\tilde{m}}^{t-1} \|Q(t) - Q(t + 1)\|.
\]

To bound this we have to study the decay of the transition probabilities w.r.t cooling schedule. Accordingly, one can verify the following relation in a straightforward manner:

\[
\frac{g(a)T^2}{\pi_a^2(T)} \pi_a(T) = - \sum_{a' \in \mathcal{A}} g(a')(\mu_{a'} - \mu_a) \exp\left(-\frac{\mu_{a'} - \mu_a}{T}\right)
\]

\[
= \sum_{a': \mu_{a'} < \mu_a} g(a')(\mu_a - \mu_{a'}) \exp\left(\frac{\mu_a - \mu_{a'}}{T}\right) - \sum_{a': \mu_{a'} > \mu_a} g(a')(\mu_{a'} - \mu_a) \exp\left(-\frac{\mu_{a'} - \mu_a}{T}\right). \tag{16}
\]
For \( a = a^* \), it can be easily seen that \( \dot{\pi}(T) < 0 \), so that \( \pi_{a^*}(t+1) > \pi_{a}(t) \) (since temperature \( T \) is a decreasing function of \( t \)). For \( a \in \mathcal{A}/a^* \), we have \( \lim_{T \to 0} \dot{\pi}(T) > 0 \). Thus, there exists a \( m_0 < \infty \), such that

\[
\dot{\pi}(t) > 0, \quad \forall t \geq m_0, \ a \in \mathcal{A}/a^* 
\]

which implies \( \pi_{a}(t+1) < \pi_{a}(t) \), for all \( t \geq m_0 \).

Then, for all \( t \geq m_0 \)

\[
\|\pi(t) - \pi(t+1)\| = \sum_{a \in \mathcal{A}/a^*} |\pi_{a}(t) - \pi_{a}(t+1)| = \sum_{a \in \mathcal{A}/a^*} \pi_{a}(t) - \pi_{a}(t) = 2 \sum_{a \in \mathcal{A}/a^*} \pi_{a}(t+1) - \pi_{a}(t),
\]

where we have used the fact that \( \pi_{a^*}(t) = 1 - \sum_{a \in \mathcal{A}/a^*} \pi_{a}(t) \). Thus, we have, with \( \tilde{m} = m_0 \), the following bound

\[
\sum_{t = m_0}^{t-1} \|\pi(t+1) - \pi(t)\| \leq 2 \sum_{t = m_0}^{t-1} \sum_{a \in \mathcal{A}/a^*} \pi_{a}(t+1) - \pi_{a}(t) \leq 2 \sum_{a \in \mathcal{A}/a^*} \pi_{a}(t). \tag{17}
\]

The above term will be bounded in the same way as Term II as follows:

**Term II**: We have

\[
\|\nu(0)Q(t) - e^*\| \leq \|\pi(t) - e^*\| = 1 - \pi_{a^*}(t) + \sum_{a \in \mathcal{A}/a^*} \pi_{a}(t) = 2 \sum_{a \in \mathcal{A}/a^*} \pi_{a}(t).
\]

Recalling the definition of \( \pi_{a}(t) \), we have

\[
\|\nu(0)Q(t) - e^*\| \leq 2 \frac{\sum_{a \in \mathcal{A}/a^*} g(a) \exp\left(-\frac{\mu_a - \mu_{a^*}}{T}\right)}{\sum_{a \in \mathcal{A}} g(a) \exp\left(-\frac{\mu_a}{T}\right)} = 2 \frac{\sum_{a \in \mathcal{A}/a^*} \frac{g(a)}{g(a^*)} \exp\left(-\frac{(\mu_a - \mu_{a^*})}{T}\right)}{1 + \sum_{a \in \mathcal{A}/a^*} \frac{g(a)}{g(a^*)} \exp\left(-\frac{(\mu_a - \mu_{a^*})}{T}\right)} \leq 2 \sum_{a \in \mathcal{A}/a^*} \frac{g(a)}{g(a^*)} (t+1)\frac{\mu_a - \mu_{a^*}}{T}. \tag{18}
\]

To conclude the proof of the lemma, bound Term I in \( \ref{eq:term_i} \) using \( \ref{eq:term_i_bound} \) with \( m_0 = \tilde{m} = t_0 \), and \( \ref{eq:term_i_bound} \) in \( \ref{eq:term_i_bound_2} \). Term II has been bounded in \( \ref{eq:term_ii} \).

**Remark 2**. We note that the constant \( m_0 \) marks the onset of monotonic decrease of all but the quasi-stationary probabilities of the optimal arm. This value can be calculated explicitly from \( \ref{eq:term_i_bound} \) by setting it to 0 for (any) next to least cost arm and solving for \( T \) the resulting equation (for a detailed explanation see for example, eq. 3.15, \[8\]).
We recall that $\tau$ is the smallest integer such that $\max_i \min_j P_{ij}(m - \tau, m) > 0$. It was proved in [27] using the results of [28] (see also Proposition 7.2, [29]) that the smallest such $\tau$ satisfies:

$$\gamma^* \leq \tau L$$

(19)

Thus, we have for $\gamma > \tau L \geq \gamma^*$,

$$0 < 1 - \frac{\tau L}{\gamma} \leq 1 - \frac{\gamma^*}{\gamma}$$

to ensure the decay of the first term in (12). Thus, we recover the characterization of the cooling schedule of [7] for finite time convergence bounds. We can use Lemma 1 to upper bound the regret.

**Theorem 3.** Let $\Delta := \max_{i \in A \setminus \{a^*\}} \{\mu_i - \mu_{a^*}\}$ and $\gamma > \tau L \geq \gamma^*$. The regret of SA is bounded by:

$$R_n \leq 2\Delta \exp\left(\frac{R^*}{\tau L} \left(\frac{m_0}{\gamma} + 1\right)^{1 - \frac{\gamma^*}{\gamma}}\right) \left(1 + \left[\frac{1}{1 - \frac{\gamma^*}{\gamma}}\right]\right) \Gamma\left(1 + \left[\frac{1}{1 - \frac{\gamma^*}{\gamma}}\right]\right) + \sum_{a \in A \setminus \{a^*\}} \frac{4\Delta g(a)}{g(a^*) (1 - \frac{\mu_{a^*} - \mu_{a^*}}{\gamma}) (n + 1)^{1 - \frac{\mu_{a^*} - \mu_{a^*}}{\gamma}}} ,$$

for $n \geq m_0$, where $\Gamma$ is the gamma function. So, we have,

$$R_n = O(n^{1 - \frac{\mu_{a^*} - \mu_{a^*}}{\gamma}})$$

**Proof.** Let $\Delta_i := \mu_i - \mu_{a^*}$. We recall the standard regret decomposition identity,

$$R_n = \sum_{i=1}^{k} \Delta_i E[T_i(n)] = \sum_{i=1}^{k} \Delta_i \left(\sum_{t=1}^{n} P(a_t = i)\right).$$

The regret can thus be bounded by (since $\Delta_{a^*} = 0$)

$$R_n \leq \Delta \sum_{i=1}^{n} P(a_t \in A / \{a^*\}).$$

The result is proved by using the previous lemma. The bound for the second term is obvious while the first term can be handled by noting that it can bounded by an integral of the form $\int_{0}^{\infty} e^{-cx^{a}} dx$ (with $a = 1 - \frac{\gamma}{\gamma}$ and $c = \frac{R^*}{\tau L}$), which in turn can be bounded as:

$$\int_{0}^{\infty} e^{-cx^{a}} dx = \frac{1}{c^a a} \int_{0}^{\infty} u^\frac{1}{a} - 1 e^{-u} du \quad (set \ cx^a = u)$$

$$= \frac{1}{c^a a} \Gamma\left(\frac{1}{a}\right)$$

$$= \frac{1}{c^a a} \Gamma\left(1 + \frac{1}{a}\right).$$

\[\Box\]

### 4 Stochastic Multi-Armed Bandit

In this section, we consider the standard stochastic MAB problem. To be more precise, we do not assume a graphical structure on $A$, so that one may think of the graph as being fully connected. It is,
Algorithm 1 SA Bandits

Input: Arm set \( A = \{k\} \); Time horizon \( n \); Temperature Parameter \( \gamma \).

Initialization: (i) Initialize the number of visits to any arm till time \( t \), denoted by \( T_a(t) \), to 0. (ii) Initialize the empirical mean estimate \( \hat{\mu}_{i,T_a(0)} = -\infty \), for all \( i \in [k] \).

for \( t = 0, \cdots, n \) do

(a) If an arm has not been played for \( \lceil 16 \log \frac{n}{\gamma^2} \rceil \) times, play it or else select arm \( a_{\min} \) according to:

\[
a_{\min} = \arg \min_{i \in [k]} \hat{\mu}_{i,T_a(t)}.
\]

(b) Uniformly randomly select an arm \( a' \in A/\{a_{\min}\} \). Accept the transition according to

\[
P_{a_{\min},a'} = \exp\left( -\frac{\left(\hat{\mu}_{a',T_a(t)} - \hat{\mu}_{a_{\min},T_a(t)}\right)^+}{T_t} \right).
\]

(c) Pull the currently selected arm \( a_{t+1} \) to get the sample \( \mu'_{a_{t+1}} \), and update the mean according to

\[
\hat{\mu}_{a_{t+1},T_{a_{t+1}}(t)+1} = \left(1 - \frac{1}{T_{a_{t+1}}(t) + 1}\right)\hat{\mu}_{a_{t+1},T_{a_{t+1}}(t)} + \frac{\mu'_{a_{t+1}}}{T_{a_{t+1}}(t) + 1}.
\]

(d) Update the cooling schedule by setting \( T_{t+1} = \frac{\gamma}{\log(t+1)} \).

end

Output: Resulting Policy: \( \nu(n) \)

However, inadvisable to carry out a straightforward naive implementation of the simulated annealing algorithm. Instead, we propose a slight modification, which takes into account the additional structure of a fully connected setting.

We devote this current section for formally proposing and analysing this algorithm. The pseudo code is provided in Algorithm 1. A key difference from traditional SA lies in how we select a root node from which the ensuing exploratory move may be performed. One can note that the exploration process is no longer Markovian in nature. However, selecting the \( \arg \min_i \hat{\mu}_i(\cdot) \) in step (a) is not necessary as discussed in the following remark.

Remark 4. We can retain the original structure of the SA algorithm, i.e. use the previously selected state in place of \( \arg \min_i \hat{\mu}_i(\cdot) \) to perform the exploratory move. The regret will be still logarithmic but the \((\Delta_i - \gamma)^2\) term in the denominator of the second term in \((20)\) in Theorem 5 will change to \( \min(\mu_i - \mu_{i-1}, \mu_{i+1} - \mu_i) - \gamma \), if the initial number of pulls for each arm is \( \lceil 16 \log \frac{n}{\gamma^2} \rceil \).

Remark 5. The requirement of initial pulls (equal to \( \lceil 16 \log \frac{n}{\gamma^2} \rceil \) here) is not unusual (see for example Algorithm UCB1-Normal, Figure 4, [2]) and we assume it to make the proofs concise. It is certainly not necessary from an empirical standpoint and all the experiments reported here forego this step. One can also use a different strategy along the lines of \( \epsilon \)-greedy algorithm in [2] (see Figure 3 in ibid), where the exploration parameter \( \epsilon \) is kept equal to 1 for the initial few iterations. Accordingly, once an arm is uniformly selected in Step (b) for our case, we can assume that the probability transition mechanism is overridden by keeping \( T_1 \) arbitrarily large, so that the arm is pulled with probability one, if \( T_a(t) \leq \lceil 16 \log \frac{n}{\gamma^2} \rceil \). This introduces an additional (constant) term in the regret which does not arise for Algorithm 1. Yet another strategy to avoid the initial pull requirement, would be the strategy of pulling both the candidate arm and the selected arm for transition if the condition \( T_a(t) \geq \lceil 16 \log \frac{n}{\gamma^2} \rceil \) is not met.
The next theorem gives the upper bound on the regret of Algorithm 1. This turns out to be logarithmic, thereby establishing the efficacy of Algorithm 1 in solving the MAB problem. Without loss of generality, we assume that the first arm is optimal, so that \( \mu_{a^*} = \mu_1 \), and \( \mu_1 \leq \mu_2 \leq \cdots \leq \mu_k \) and \( \Delta_i := \mu_i - \mu_1 \).

**Theorem 6.** If \( \gamma \in (0, \Delta_2) \), then the regret of Algorithm 1 on any bandit \( P \in \mathcal{E}_{SG}^k(1) \) environment, is bounded by

\[
R_n \leq \left( 2k + \frac{\log n}{k-1} \right) \sum_{i=1}^{k} \Delta_i + \sum_{i=1}^{k} \frac{16\Delta_i}{(\Delta_i - \gamma)^2} \log n,
\]

where \( n \) is the number of rounds and \( k \) is the number of arms.

**Proof.** We recall the equation,

\[
R_n = \sum_{i=1}^{k} \Delta_i \mathbb{E}[T_i(n)].
\]

We proceed by considering a suitably defined event \( G \) and subsequently bounding \( \mathbb{E}[T_i(n)] \) for each sub-optimal arm \( i \in A/\{a_1\} \) on \( G \) (and its complement). Accordingly, the event \( G \) is defined to be:

\[
G := \left\{ \mu_1 \leq \min_{s \in [n]} \{ \hat{\mu}_{1,s} + \epsilon_{1,s} \} \right\} \cap \left\{ \max_{s \in [n]} \{ \hat{\mu}_{1,s} - \epsilon_{1,s} \} \leq \mu_1 \leq \min_{s \in [n]} \{ \hat{\mu}_{2,s} + \epsilon_{2,s} \} \right\}...
\]

\[
\ldots \cap \left\{ \max_{s \in [n]} \{ \hat{\mu}_{k-1,s} - \epsilon_{k-1,s} \} \leq \mu_{k-1} \leq \min_{s \in [n]} \{ \hat{\mu}_{k,s} + \epsilon_{k,s} \} \right\}...
\]

\[
\ldots \cap \left\{ \min_{s \in [n]} \{ \hat{\mu}_{k,s} - \epsilon_{k,s} \} \leq \mu_k \right\},
\]

where,

\[
\epsilon_{i,s} = \sqrt{\frac{2 \log \frac{1}{\delta}}{s}}.
\]

We specify the value of \( \delta \) below. We essentially proceed by demonstrating two facts which will allow us to bound \( \mathbb{E}[T_i(n)] \):

1. The complement event \( G^c \) occurs with sufficiently low probability if we set \( \delta = \frac{1}{n^{2}} \).

2. On the event \( G \), if each arm has been been played at least \( u_i = \frac{16 \log n}{(\Delta_i - \gamma)^2} \) times, the probability of transitioning to a sub-optimal arm diminishes as \( O\left(\frac{1}{T}\right) \) for round \( t \).

We have:

\[
\mathbb{E}[T_i(n)] = \mathbb{E}[\mathbb{I}\{G\}T_i(n)] + \mathbb{E}[\mathbb{I}\{G^c\}T_i(n)].
\]

We first consider fact 1. Let us define an event \( G^c_i \) as

\[
G^c_i = \left\{ \mu_i \geq \min_{s \in [n]} \{ \hat{\mu}_{i,s} + \epsilon_{i,s} \} \right\} \cup \left\{ \mu_i \leq \max_{s \in [n]} \{ \hat{\mu}_{i,s} - \epsilon_{i,s} \} \right\}.
\]

We remark in passing that by definition, \( G^c = \cup_i G^c_i \). We have

\[
G^c_i \subset \bigcup_{s \in [n]} \left\{ \left| \mu_i - \hat{\mu}_{i,s} \right| \geq \epsilon_{i,s} \right\}.
\]
Since we assume sub-Gaussian bandits, the tail decay can be bounded using the Cramer-Chernoff bound, which gives
\[
P(\lvert \mu_i - \hat{\mu}_{i,s} \rvert \geq \epsilon_{i,s}) \leq 2 \exp \left( - \frac{s \epsilon_{i,s}^2}{2} \right).
\]

By definition of \( \epsilon_{i,s} \), we have
\[
P(G^c_i) \leq \sum_{s=1}^{n} P(\lvert \mu_i - \hat{\mu}_{i,s} \rvert \leq \epsilon_{i,s}) \leq \sum_{s=1}^{n} \sqrt{\frac{2 \log \frac{1}{s}}{s}} \leq 2n\delta.
\]

Using this estimate we have
\[
E[I\{G c_i\}T_i(n)] \leq n P(G^c) \leq 2n^2\delta k \leq 2k
\]
using the fact that \( \delta = \frac{1}{n^2} \).

Next, we prove fact 2. We note that on the event \( G \), \( \text{argmin}_i \hat{\mu}_{i,T_i(t)} = \mu_1 \) is guaranteed (since \( \epsilon_{i,T_i(t)} \leq \frac{\Delta_i - \gamma}{2} \) for all \( i \), when \( T_i(t) \geq \left\lceil \frac{16 \log n}{(\Delta_i - \gamma)^2} \right\rceil \)). So, we need only consider the probability of transitioning from the first arm to any other sub-optimal arm. To avoid notational clutter, let \( u_i = T_i(t) \) in what follows. For any round \( t \) and \( i \neq 1 \),
\[
\exp \left( - \frac{(\hat{\mu}_{1,ui} - \hat{\mu}_{1,u1})^+}{T_t} \right) = \exp \left( - \frac{\hat{\mu}_{1,ui} - \hat{\mu}_{1,u1}}{T_t} \right) \leq \exp \left( \frac{\mu_1 + \epsilon_{1,u1} - \mu_i + \epsilon_{i,ui}}{T_t} \right).
\]

We use the fact, on the event \( G \), \( \hat{\mu}_{1,s} \leq \mu_1 + \epsilon_{1,s} \) and \( \hat{\mu}_{i,s} \geq \mu_i - \epsilon_{i,s} \). Furthermore, by our choice of \( \epsilon_{i,s} \), we have
\[
\epsilon_{i,ui} \leq \frac{\Delta_i - \gamma}{2}.
\]

So, at any round \( t \), we have the probability of transitioning to a sub-optimal arm is bounded by:
\[
\exp \left( - \frac{\hat{\mu}_{1,ui} - \hat{\mu}_{1,u1}}{T_t} \right) \leq \exp \left( \frac{\epsilon_{1,u1} + \epsilon_{1,ui}}{T_t} \right) \leq \frac{\epsilon_{1,u1} + \epsilon_{1,ui}}{T_t} \leq \frac{\Delta_i - \gamma}{2T_t} \leq \frac{1}{T_t}.
\]

The above estimate can be used to bound \( E[T_i(n)] \) on \( G \). We have, on event \( G \), the following holds
\[
E[I\{G\}T_i(n)] \leq \left\lceil \frac{16 \log n}{(\Delta_i - \gamma)^2} \right\rceil + \frac{1}{k-1} \sum_{t \in [n]} \frac{1}{t} \leq \left\lceil \frac{16 \log n}{(\Delta_i - \gamma)^2} \right\rceil + \frac{\log n}{k-1}.
\]
It remains to bring together our estimates of $E[T_i(n)]$ on events $G$ and $G^c$ to finish the proof. Using (22) and (23) in (21), we have

$$R_n = \sum_i \Delta_i E[T_i(n)] \leq \sum_i \Delta_i \left( 2k + \left\lceil \frac{16n \log n}{(\Delta_i - \gamma)^2} \right\rceil + \log n \right) \leq \left( 2k + \frac{\log n}{k - 1} \right) \sum_{i=2}^k \Delta_i + \sum_{i=2}^k \frac{\Delta_i \times 16 \log n}{(\Delta_i - \gamma)^2}.$$  

We have stated Algorithm 1 with $\gamma := \frac{\Delta}{n}$ in Step(a). The regret bound obtained above can be turned into one that is independent of the reciprocal of the sub-optimality gaps, $\Delta_i$. A straightforward calculation in the above regret bound yields (see e.g. Theorem 7.2, [4]):

$$R_n \leq 16\sqrt{n}k \log n + \left( 2k + \frac{\log n}{k - 1} \right) \sum_{i=1}^k \Delta_i.$$

5 Noisy Simulated Annealing

In this section we analyze SA with noisy observations in full generality. Accordingly, we only assume a connected graph (see assumption(ii)) in place of the fully connected setting of the previous section. The formal procedure is detailed in Algorithm 2.

We remark that, although we can guarantee the conditions stipulated in step (a) for $T_i(t_0)$, we stress that from an empirical standpoint, step (a) is unnecessary since one rarely anneals the temperature to zero. To satisfy the condition for small to medium sized arm sets, one can execute a simple depth first search algorithm (which has a worst case complexity of $O(k + e)$, with $e := |\mathcal{E}|$ denoting the number of edges) to allocate $T_i(t_0)$ budget to each arm. This will incur the least regret. For larger sized arm sets, a better strategy would be to just initialize the empirical mean to $0$. To satisfy the condition for a certain pre-determined time duration. Since our graph is connected, we are guaranteed to visit all arms and allocate the preliminary budget to them. One could also use the previously mentioned strategy of pulling both arms if the condition for $T_i(t_0)$ is not met. Our analysis builds upon the framework established in [30] to study SA.

5.1 Noise Perturbation Bounds

To establish a regret bound, we first study how noisy observations affect the transition probability. To do this, we construct a related stochastic process $\{\bar{a}_n\}_{n \geq 1}$, undisturbed by any noise so that to select the next arm, we use the actual mean difference. To be more precise, we recall the definition of the probability matrix of noiseless SA:

$$\hat{P}_{a,a'}(n) := P(\bar{a}_{n+1} = a' | \bar{a}_n = a) := \begin{cases} 0, & \text{if } a' \notin \mathcal{N}(a), a' \neq a, \\ \frac{g(a,a')}{g(a)} \exp \left\{ \frac{-|\mu_{a'} - \mu_a|}{T_n} \right\}, & \text{otherwise,} \end{cases}$$

for any $a, a' \in \mathcal{A}$. Let $\{a_n\}_{n \geq 0}$ index the original process. Our aim is to bound the difference:

$$\epsilon_n := \bar{P}(\bar{a}_{n+1} = a | \bar{a}_n = a') - P(a_{n+1} = a | a_n = a') .$$

Towards this end, we note that the transition probability for the noisy process, denoted by $\{a_n\}$, can be written as (when $a_{n+1} \neq a_n$):

$$P_{a,a'}(n) = P(a_{n+1} = a | a_n = a') = \frac{g(a, a')}{g(a)} \exp \left\{ \frac{-|\mu_{a'} - \mu_a + \lambda(n)|}{T_n} \right\} ,$$
In obtaining the above inequality, we have used the easily verifiable relation: 
\[ (a + b)^+ - b^+ \leq |a| \] for any \( a, b \in \mathbb{R} \). Setting \( a = \lambda(n) \) and \( b = \mu_{n+1} - \mu_n \) in this relation gives the required inequality. Denote \( \beta_n := \frac{1}{T_n} \), so that
\[
\epsilon_n \leq \mathbb{E}_\lambda(n) e^{\beta_n \left( \frac{1}{T_n} \sum_{i=1}^{T_n} a_i - \frac{1}{T_n} \sum_{j=1}^{T_n} \hat{a}_j \right)} - 1.
\]
where \( \tilde{u}_i \) are sub-Gaussian with variance \( \sigma^2 \). Let \( u_n \) denote the term inside the brackets in the above expression. Using the relation \( \mathbb{E} X = \int_{x \geq 0} \mathbb{P}(X \geq x) dx \) for any non-negative random variable \( X \), we have

\[
e_n \leq \int_0^1 \mathbb{P}(\beta_n|u_n| > \log x) dx + \int_1^{\infty} \mathbb{P}(\beta_n|u_n| > \log x) dx - 1
= \int_1^{\infty} \mathbb{P}(\beta_n|u_n| > \log x) dx.
\]

Let \( N_n := \min\{T_a(n), T_{a'}(n)\} \). Using the sub-Gaussian property of \( u_n \) (see eq. (24)), we have

\[
\epsilon_n \leq 2\alpha(n)e^{\alpha(n)^2} \int_{\infty}^{0} e^{-s} ds.
\]

To finish, we note that,

\[
\int_{-\infty}^{\infty} \exp(-u^2) du \leq \sqrt{\pi}.
\]

Thus we have :

\[
\epsilon_n \leq 2\sqrt{\pi} \alpha(n)e^{\alpha(n)^2}.
\]

We state this result in the following lemma:

**Lemma 7.** For any \( a, a' \in A \), we have

\[
\bar{P}_{a,a'}(n) - P_{a,a'}(n) \leq 2\sqrt{\pi} \alpha(n)e^{\alpha(n)^2},
\]

where \( \alpha(n) := \frac{\sigma \beta_n}{\sqrt{N_n}} \) with \( \sigma \) being the variance of the noise, \( \beta_n := \frac{1}{T_n} \) the inverse temperature and \( N_n := \min\{T_a(n), T_{a'}(n)\} \).

We note that if \( T_a(n) = O((\log n)^{-(2+\epsilon)}) \), \( \epsilon > 0, a \in A \), then \( \alpha(n) = O((\log n)^{-\frac{1}{2}}) \to 0 \), as \( n \to \infty \). Then one can establish the asymptotic convergence of algorithm 2 by using Theorem 3.1 of [12] in conjunction with the above result.

### 5.2 Construction of the Markov Process

To establish the regret bound, it is convenient to switch to a continuous time analysis. This will allow us to use the concept of Dirichlet form which is central to the main arguments of the proofs. Before doing so, it is instructive to go through the standard construction of the Markov process since handling time inhomogeneity requires some care. We follow the construction of [30] here. Let \( \{X(t) : t \in \mathbb{R}^+\} \) be the required Markov process having the transition mechanism \( P(s,t) \) corresponding to the SA chain.
Let $\nu(0)$ be the initial distribution of the chain. So, we want to define a process \{\(X(t) : t \in \mathbb{R}^+\)\} which satisfies:

\[
\mathbb{P}(X(0) = i) = \nu_i(0) \quad P(s, t)_{X(\sigma)} := \mathbb{P}(X(t) = j | X(\sigma), \sigma \in [0, s]).
\] (25)

A naive construction of \{\(X(t)\)\} with the above transition mechanism would potentially mean dealing with an uncountable number of random variables for each \((t, i) \in [0, \infty) \times \mathcal{A}\). To avoid this, we let \(S(t, i, j) := \frac{1}{g(i)} \sum_{m=1}^j g(i, m) e^{-\beta(t)(\hat{\mu}_m(t) - \tilde{\mu}_i(t))} + \) where \(\hat{\mu}_i(t)\) is defined below. Define:

\[
\Psi(t, i, u) = \begin{cases} j & \text{if } \frac{S(t, i, j-1)}{S(t, i, k)} \leq u < \frac{S(t, i, j)}{S(t, i, k)} \\ i & \text{if } u > 1. \end{cases}
\] (26)

Also, we determine the function \(T\) by:

\[
\int_s^{s+T(s, i, \xi)} S(\tau, i, k)^{-1} d\tau = \xi.
\]

Having defined the above functions, the rest of the construction is the same as that of a standard cadlag jump process: Let \(X(0)\) be an \(\mathcal{A}\)-valued random variable with distribution \(\nu(0)\) and \(\{E_n : n > 1\}\) be a sequence of i.i.d. mean 1 exponential random variables. Furthermore, let \(\{U_n : n > 1\}\) be a sequence of i.i.d. random variables uniformly distributed on \([0, 1]\) and also independent of the sigma algebra \(\sigma(X(0) \cup \{E_n : n > 1\})\). Finally, set \(J_0 = 0\) and \(X(0) = X_0\), and, when \(n > 1\), we inductively define the following

\[
J_n - J_{n-1} = T(J_{n-1}, X(J_{n-1}), E_n), \quad X(J_n) = \Psi(J_n, X(J_{n-1}), U_n), \quad X(t) := X(J_{n-1}), \text{ for } J_{n-1} \leq t < J_n,
\]

\[
\hat{\mu}_i(t) := \frac{1}{T_i(t)} \sum_{m=1}^n I\{X(t') = i, t' \in [J_{m-1}, J_m]\} \hat{\mu}_i \text{ for } J_{n-1} \leq t < J_n, \quad i \in \mathcal{A},
\]

where \(\mathbb{E} \hat{\mu}_i = \mu_i\) and \(T_i(t) := \sum_{m=1}^n I\{X(t') = i, t' \in [J_{m-1}, J_m]\}\) is the number of visits to arm \(i\) till time \(t \in [J_{n-1}, J_n]\). A routine argument can show that for the above process, (25) holds. The Q-matrix for the SA process with a continuous time temperature parameter \(T_i : \mathbb{R} \rightarrow \mathbb{R}\) can be defined as:

\[
Q_{ij}(t) := \begin{cases} \frac{g(i, j)}{g(i)} \exp(-\beta(t)(\hat{\mu}_j(t) - \hat{\mu}_i(t))) & \text{for } j \neq i \\ -\sum_{j \neq i} Q_{ij}(t) & \end{cases}
\] (27)

We note that our Q-matrix is time dependent which potentially poses a problem towards deploying the standard Markov machinery (more precisely, the Kolmogorov forward equation). We therefore show that the forward equation holds almost everywhere because \(Q(t)\) is continuous almost everywhere. We will prove the latter in what follows, i.e., we prove

\[
\frac{d}{dt} P(s, t) = P(s, t)Q(t) \text{ almost surely on } (s, \infty) \text{ with } P(s, s) = I.
\] (28)

To prove (28), we use an approximation argument. Accordingly, let

\[
Q^N(t) := Q([t]_N) \quad \text{where } [t]_N = \frac{J_n}{N} \text{ for } t \in \left[\frac{J_n}{N}, \frac{J_{n+1}}{N}\right) \text{ for } t > s
\]

It is obvious, the solution to (28) with \(Q(t)\) replaced with \(Q^N(t)\), is given by

\[
P^N(s, t) = P^N(s, \min\{s \lor [t]_N\}) \exp((t-s)Q([t]_N)).
\]
We note that:

\[ \|P^N(s,t) - P^M(s,t)\| \leq \int_s^t \|Q([\tau]_N) - Q([\tau]_M)\| \|P^N(s,\tau)\| d\tau + \int_s^t \|Q([\tau]_M)\| \|P^N(s,\tau) - P^M(s,\tau)\| d\tau. \]

Let \( G \) denote the matrix with entries \( g_{ij} := \frac{g(i,j)}{\sum_{j \in N(A)} g(i,j)} \) and \( g_{ii} = 0 \). So, the above inequality gives

\[ \|P^N(s,t) - P^M(s,t)\| \leq \int_s^t \|Q([\tau]_N) - Q([\tau]_M)\| d\tau + \|G\| \int_s^t \|P^N(s,\tau) - P^M(s,\tau)\| d\tau. \]

An application of Gronwall’s lemma gives

\[ \sup_{0 \leq s \leq t \leq T} \|P^N(s,t) - P^M(s,t)\| = O\left( e^{GT} \int_0^T \|Q([\tau]_N) - Q([\tau]_M)\| d\tau \right). \]

We note that as \( N, M \to \infty \), then \( |\beta([\tau]_N) - \beta([\tau]_M)| \to 0 \) implying that \( \|Q([\tau]_N) - Q([\tau]_M)\| \to 0 \). So the above equation guarantees that \( P^N(s,t) \) converges to some \( (s,t) \to P(s,t) \) on finite intervals. In particular, this implies that

\[ P(s,t) = I + \int_s^t P(s,\tau)Q(\tau) d\tau \tag{29} \]

which is the integrated form of (28). It can then be seen by routine arguments that

\[ \dot{\nu}(t) = \nu(t)Q(t) \ \text{for} \ t \in [s, \infty) \iff \nu(t) = \nu(s)P(s,t) \ \text{for} \ t \in [s, \infty). \]

for \( P(s,t) \) satisfying (29). This completes the construction of the process \( \{X(t)\} \).

5.3 Regret Bound

To begin establishing the regret bound, we briefly recall the details of some technical concepts related to Markov processes. Let \( f \) denote a column vector determined by any function \( f : A \to \mathbb{R}^k \). Also, let \( \pi(t) \) denote the stationary distribution of \( P(t) \), determined by the continuous cooling schedule \( T(t) = \frac{\gamma}{1 + \log(t+t_0)} \), where \( t \in \mathbb{R} \). We will use the following notation throughout this section:

\[ \|f\|_\pi = \sqrt{\langle f^2 \rangle_\pi} \quad \text{where} \quad \langle g \rangle_\pi := \sum_{i \in A} g_i \pi_i. \]

Since \( \pi_i(t) > 0 \) for each \( i \in A \), we note that \( \| \cdot \|_\pi \) is a complete norm. The inner product between any two functions \( f, g : A \to \mathbb{R}^k \) is denoted by \( \langle f, g \rangle_\pi := \sum_{i \in A} \pi_i f_i g_i \).

**Definition 5 (Dirichlet Forms and Poincaré inequality)** Let \( L^2(\pi) \) denote the space of \( f \) for which \( \|f\|_\pi < \infty \). This space is actually a Hilbert space for which the transition matrix \( P \) acts as a self adjoint contraction assuming the detailed balance equations are satisfied. The variance of \( f \) is then defined to be:

\[ \text{Var}_\pi(f) := \|f - \langle f \rangle_\pi\|_\pi^2. \]

The Dirichlet form is defined as:

\[ \mathcal{E}_t(f,f) := \frac{1}{2} \sum_{i,j \in A} \pi_i Q_{ij}(t)(f_j - f_i)^2. \tag{30} \]
Using the Dirichlet form, the Poincaré constant (assuming fixed $t$) can be defined as

$$\xi_+ = \inf \{ \mathcal{E}_t(f,f) : f \in L^2(\pi) \text{ and } \text{Var}_\pi(f) = 1 \}$$

which gives us the well known Poincaré inequality:

$$\xi_+ \text{Var}_\pi(f) \leq \mathcal{E}_t(f,f), \ f \in L^2(\pi).$$

For our purposes, we need the definition of $\xi_+$ when $\pi := \pi(t)$. We denote this by

$$\lambda(t) := \inf \{ \mathcal{E}_t(f,f) : \text{Var}_{\pi(t)}(f) = 1 \}.$$  

The upper (and lower) bound on $\lambda(t)$ has been provided in (Theorem 5.4.11, [30]):

$$\pi_- e^{-\beta(t)\gamma^*} \leq \lambda(t) \leq \pi_+ e^{-\beta(t)\gamma^*},$$  

where $\pi_-, \pi_+$ are bounded constants, $\gamma^*$ is the critical depth and $\beta(t) := \frac{1}{t}$ is the inverse temperature.

We use the noise perturbation bound derived in the previous section to establish related upper bounds for the $Q$-matrix perturbation with respect to the noiseless process. Accordingly, let

$$\sigma_{ij}(t) := \tilde{Q}_{ij}(t) - Q_{ij}(t)$$

where $\tilde{Q}(t) := Q$-matrix corresponding to the continuous time version of the undisturbed process $\{a_n\}$. Then, by Lemma [7] we have

$$\sigma_{ij}(t) \leq 2\sqrt{\pi}c(t)e^{\alpha(t)^2},$$

where $\alpha(t) := \frac{\pi^2(t)}{\min\{T_i(t), T_j(t)\}}$, $\beta(t) = \frac{\log(t+1)}{1}$ and $t_0 := J_{t_0}$ (see step(a) of Algorithm 2). By step (a) of the algorithm, we have $\alpha(t) \leq \frac{\pi^2(t)}{\sqrt{T(t_0)}}$, where $T(t_0) := \min_{a \in \mathcal{A} T_a(J_{t_0})}$. Set $\alpha(t) := c\beta(t)$, where $c := \frac{\sigma}{\sqrt{T(t_0)}}$. Suppose $T_i(t_0) := O(t^2(\log t)^2)$ for any $\epsilon > 0$. Then, we have for any $t \geq t_0$ that $c\beta(t) \leq \frac{\sigma}{\gamma t^2}$, which in turn implies

$$\sigma_{ij}(t) \leq \frac{2\sqrt{\pi}c(e^{\frac{\sigma^2}{\gamma t^2}})}{\gamma},$$

for $t \geq t_0$. So,

$$\int_{t_0}^{t} \sigma(t)dt \leq \int_{t_0}^{t} 2\sqrt{\pi}c\beta(t)e^{(c\beta(t))^2}dt \leq \frac{2\sqrt{\pi}c e^{\frac{\sigma^2}{\gamma(1 - \frac{\sigma}{\gamma t^2})}}}{\gamma}t^{1 - \frac{\sigma}{\gamma}}.$$  

We will use the above bound later. Just as in the noiseless case, in order to establish a regret bound, we first bound the probability of the event $X(t) \in \mathcal{A}/\{a^*\}$.

**Lemma 8.** Let $\beta(t) = \frac{\log(1+at)}{\gamma}$, where $a = \frac{\pi_-(\gamma)}{\mu_{\max}}$ and $\mu_{\max} = \max_i \mu_i$. Suppose that $T_i(t_0) := O(t^2(\log t)^2)$ for any $\epsilon > 0$. Then, for $\gamma > \frac{2\sigma^*}{\epsilon}$, we have for any $t > t_0 + \tau_0$, where $\tau_0 = \left(\frac{4\sqrt{\pi}c e^{\frac{\sigma^2}{\gamma}}}{{\gamma \pi}_{\min}}\right)^{\frac{1}{2 - \frac{\sigma}{\gamma}}}$, that

$$P\left( X(t) \notin \mathcal{A}/\{a^*\} \right) \leq \sqrt{\frac{2g(k+4)}{(1+at)^{\Delta_{\min}}}},$$

where $g := 2 \sum_{i \in \mathcal{A}/a^*} \frac{g(a)}{g(a)^2}$, $\Delta_{\min} := \min_{i \in \mathcal{A}} (\mu_{a_i} - \mu_{a^*})$ and $k$ is the number of arms.
Proof. Let $\nu(t) = \nu(t_0)P(t_0, t)$ for any $t \geq t_0$. We denote the Radon-Nikodym derivative of $\nu(t)$ w.r.t $\pi(t)$ by:

$$f_i(t) = \frac{\nu_i(t)}{\pi_i(t)}, \forall t \geq t_0, i \in \mathcal{A}.$$  

To get an estimate on the probability of selecting a bad arm, we note that

$$P(A(t) \notin \mathcal{A}/\{a^*\}) = \sum_{i \in \mathcal{A}/\{a^*\}} \nu_i(t) = \sum_{i \in \mathcal{A}/\{a^*\}} \pi_i(t) f_i(t) = \langle f(t), \mathbb{1}\{i \in \mathcal{A}/a^*\}\rangle_{\pi(t)} \leq \|f(t)\|_{\pi(t)} \sqrt{\sum_{i \in \mathcal{A}/\{a^*\}} \pi_i(t)}.$$ 

Suppose that $\|f(t)\|_{\pi(t)} \leq C$ for some constant $C > 0$, then we would have

$$P(A(t) \notin \mathcal{A}/\{a^*\}) \leq \frac{Cg}{(1 + at)^{\frac{\min_i \mu_i}{\pi(t)}}},$$

where we have used the continuous version of the previously established estimate of the decay rate of the Gibbs measure to the Dirac measure on $a^*$ (see (18)). The rest of the proof bounds $\|f(t)\|_{\pi(t)}$ by providing an upper bound on the constant $C$.

Let $Z(t) := \sum_{a \in \mathcal{A}} g(a)e^{-\beta(t)\mu_a}$ denote the partition function, so that $\pi_a(t) = g(a)e^{-\beta(t)\mu_a}/Z(t)$. Then, we have by the easily verifiable relation $\dot{Z}(t) = -\ddot{\beta}(t)Z(t)(\mu - \langle \mu \rangle_{\pi(t)})$, that

$$\frac{d}{dt}\|f(t)\|^2_{\pi(t)} = \frac{d}{dt}\left(\frac{1}{Z(t)} \sum_{i \in \mathcal{A}} (f_i(t))^2 e^{-\beta(t)\mu_i}\right) = 2 \frac{\dot{Z}(t)}{Z(t)} \sum_{i \in \mathcal{A}} f_i(t) \dot{f}_i(t) e^{-\beta(t)\mu_i} - \ddot{\beta}(t) \sum_{i \in \mathcal{A}} \mu_i f_i^2(t) e^{-\beta(t)\mu_i} - \frac{1}{Z^2(t)} \sum_{i \in \mathcal{A}} \dot{Z}(t) f_i^2(t) e^{-\beta(t)\mu_i}
= 2\langle f(t), \dot{f}(t) \rangle_{\pi(t)} - \ddot{\beta}(t)(\mu - \langle \mu \rangle_{\pi(t)}), f^2(t)\rangle_{\pi(t)}.$$  

(35)

We can establish a different upper bound on $\frac{d}{dt}\|f(t)\|^2_{\pi(t)}$ by noting that

$$\|f(t)\|^2_{\pi(t)} = \langle f(t), f(t) \rangle_{\pi(t)} = \sum_{j \in \mathcal{A}} \left(\sum_{i \in \mathcal{A}} \nu_i(t_0) P_{ij}(t_0, t)\right) f_j(t) = \sum_{i \in \mathcal{A}} \nu_i(t_0) \left(\sum_{j \in \mathcal{A}} P_{ij}(t_0, t) f_j(t)\right) = \langle P(t_0, t) f(t), f(t) \rangle_{\pi(t_0)}.$$  

22
Differentiating the above equality:

\[
\frac{d}{dt} \|f\|_{\pi(t)}^2 = \left< P(t_0, t)Q(t)f(t) \right>_{\nu(t_0)} + \left< P(t_0, t)\dot{f}(t) \right>_{\nu(t_0)} = \left( \sum_{i \in A} \left( \sum_{p \in A} \nu_p(t_0)P_{pi}(t_0, t) \right) \left( \sum_{j \in A} q_{ij}(t)f_j(t) \right) \right) + \sum_{i \in A} \left( \sum_{p \in A} \nu_p(t_0)P_{pi}(t_0, t) \right) \dot{f}_i(t)
\]

\[
= \sum_{i \in A} f_i(t)\pi_i(t) \left( \sum_{j \in A} q_{ij}(t)f_j(t) \right) + \sum_{i \in A} f_i(t)\pi_i(t)\dot{f}_i(t)
\]

\[
= \sum_{i \in A} \pi_i(t)\left( - \sum_{j \in A/\{i\}} q_{ij}(t)f_j^2(t) + \sum_{j \in A/\{i\}} q_{ij}(t)f_i(t) f_j(t) \right) + \langle f(t), \dot{f}(t) \rangle_{\pi(t)}
\]

\[
= -\mathcal{E}_t(f(t), f(t)) + \langle f(t), \dot{f}(t) \rangle_{\pi(t)}.
\]  

(36)

where we have used (28) in the first term of the first equation, (27) in the fourth equation and the fact that \(\pi_i(t)q_{ij}(t) = \pi_j(t)q_{ij}(t)\) along with the definition of (30) in deriving the last equation. Define

\[
\mathcal{E}_t(f(t), f(t)) := \frac{1}{2} \sum_{i \in A, j \neq i} \pi_i(t)Q_{ij}(t) (f(j) - f(i))^2.
\]

Combining estimates (35) and (36) to remove the \(\langle f(t), \dot{f}(t) \rangle\) term, we have:

\[
\frac{d}{dt} \|f(t)\|_{\pi(t)}^2 = -2\mathcal{E}_t(f(t), f(t)) + \dot{\beta}(t)(\mu - \langle \mu \rangle_{\pi(t)}, f^2(t))_{\pi(t)}
\]

\[
= -2\mathcal{E}_t(f(t), f(t)) + \dot{\beta}(t)(\mu - \langle \mu \rangle_{\pi(t)}, f^2(t))_{\pi(t)} + 2\mathcal{E}_t(f(t), f(t)) - 2\mathcal{E}_t(f(t), f(t)),
\]

\[
\leq -2\lambda(t)\|f(t)\|_{\pi(t)}^2 + 2\lambda(t) + \dot{\beta}(t)\mu_{\max}\|f^2(t)\|_{\pi(t)} + 2\mathcal{E}_t(f(t), f(t)) - 2\mathcal{E}_t(f(t), f(t))
\]

\[
\leq -\lambda(t)\|f(t)\|_{\pi(t)}^2 + 2\lambda(t) + 2\mathcal{E}_t(f(t), f(t)) - 2\mathcal{E}_t(f(t), f(t)),
\]  

(37)

where we have used the facts: (i) \(\mathcal{E}_t(f(t), f(t)) \geq \text{Var}_{\pi(t)}(f(t))\lambda(t) = (\|f\|_{\pi(t)}^2 - 1)\lambda(t)\) in the first term in the third inequality, and, (ii) \(\mu_{\max}\dot{\beta}(t) \leq \lambda - e^{-\beta(t)}\gamma^* \leq \lambda(t)\) by definition of \(\beta(t)\) in the last inequality.

We also have (see eq. (32))

\[
\mathcal{E}_t(f(t), f(t)) - \mathcal{E}_t(f(t), f(t)) = \frac{1}{2} \sum_{i \in A, j \neq i} \pi_i(t)\pi_{ij}(t) (f(j) - f(i))^2,
\]

so that

\[
\mathcal{E}_t(f(t), f(t)) - \mathcal{E}_t(f(t), f(t)) \leq \frac{1}{2} \max_{i,j: i \neq j} \sum_{i \in A, j \neq i} \pi_i(t)(f(j) - f(i))^2.
\]

Let \(\sigma(t) = \max_{i,j: i \neq j} \pi_{ij}(t)\). The above inequality gives

\[
\mathcal{E}_t(f(t), f(t)) - \mathcal{E}_t(f(t), f(t)) \leq k\sigma(t)\|f(t)\|_{\pi(t)}^2.
\]

Plugging this back in (37) and w.l.o.g. absorbing the factor of \(k\) into the \(\sigma(t)\) term, we have

\[
\frac{d}{dt} \|f(t)\|_{\pi(t)}^2 \leq -\left( \lambda(t) - \sigma(t) \right)\|f(t)\|_{\pi(t)}^2 + 2\lambda(t)
\]

\[
\implies \frac{d}{dt} \|f(t)\|_{\pi(t)}^2 + (\lambda(t) - \sigma(t))\|f(t)\|_{\pi(t)}^2 \leq 2\lambda(t).
\]

23
Set $\rho(t) = \int_{t_0}^{t} \lambda(t) dt - \int_{t_0}^{t} \sigma(t) dt$ and multiply both sides of the above inequality by $e^{\rho(t)}$ to get
\[
\frac{d}{dt} e^{\rho(t)} \| f(t) \|^2 \pi(t) \leq 2\lambda(t) e^{\rho(t)}.
\]
Thus,
\[
\| f(t) \|^2 \pi(t) \leq \frac{1}{e^{\rho(t)}} \| f(t_0) \|^2 \pi(t_0) + \frac{2}{e^{\rho(t)}} \int_{t_0}^{t} \lambda(\tau) e^{\rho(\tau)} d\tau.
\] (38)

We first establish conditions on $\gamma$ for $\int_{t_0}^{t} \lambda(t) dt > \int_{t_0}^{t} \sigma(t) dt$ to hold. We recall that $\lambda(t) \geq \pi_- e^{-\beta t} = \pi_-(1 + at)^{-\frac{\gamma}{\pi}}$. Also, as discussed previously, $\int_{t_0}^{t} \sigma(t) dt = O(t^{1-\frac{\gamma}{2}})$ (see eq. (34)). So, for the required condition to hold, we must have:
\[
\int_{t_0}^{t} \lambda(\tau) d\tau > \int_{t_0}^{t} \sigma(\tau) d\tau \iff \frac{\pi_-}{1 - \frac{\gamma}{\pi}}(1 + at)^{1-\frac{\gamma}{2}} > \int_{t_0}^{t} \sigma(\tau) d\tau = \frac{2\sqrt{\pi\sigma e^{\gamma^2}}}{\gamma(1 - \frac{\gamma}{2})} t^{1-\frac{\gamma}{2}}.
\]

The above condition will be satisfied for $\gamma > \frac{2\gamma}{t}$ for $t \geq t' = O(1)$. We assume $t' = t_0$ without loss of generality. We next consider the integral term $\int_{t_0}^{t} \lambda(\tau) e^{\rho(\tau)} d\tau$:
\[
\int_{t_0}^{t} \lambda(\tau) e^{\rho(\tau)} d\tau = \int_{t_0}^{t} \frac{\lambda(\tau) - \sigma(\tau)}{1 - \frac{\sigma(\tau)}{\lambda(\tau)}} e^{\int_{t_0}^{\tau} \lambda(\tau') d\tau'} e^{\int_{t_0}^{\tau} \sigma(\tau') d\tau'} d\tau
\]
\[
= \int_{t_0}^{t} \frac{1}{1 - \frac{\sigma(\tau)}{\lambda(\tau)}} \left( \frac{d}{d\tau} e^{\rho(\tau)} \right) d\tau.
\]

Consider any $\tau \geq t_0$. Then, we have $\frac{\sigma(\tau)}{\lambda(\tau)} \leq \frac{2\sqrt{\pi\sigma e^{\gamma^2}}}{\gamma(1 - \frac{\gamma}{2})} t^{1-\frac{\gamma}{2}}$ (see eq. (33)) and $\lambda(\tau) \geq \pi_- e^{-\beta(\tau)\gamma^*}$ (see eq. (31)). So,
\[
\frac{\sigma(\tau)}{\lambda(\tau)} \leq \frac{2\sqrt{\pi\sigma e^{\gamma^2}}}{\gamma(1 - \frac{\gamma}{2})} t^{1-\frac{\gamma}{2}}
\]
\[
\leq \frac{2\sqrt{\pi\sigma e^{\gamma^2}}}{\gamma(1 - \frac{\gamma}{2})} \left( \frac{1}{t^{1-\frac{\gamma}{2}}} \right)
\]
\[
\leq \frac{1}{t^{1-\frac{\gamma}{2}}},
\]
if $t \geq \left( \frac{4\sqrt{\pi\sigma e^{\gamma^2}}}{\gamma(1 - \frac{\gamma}{2})} \right)^{1-\frac{\gamma}{2}} := t_0$. So, we have
\[
\int_{t_0}^{t} \lambda(\tau) e^{\rho(\tau)} d\tau \leq \int_{t_0}^{t} \left( e^{\rho(t)} - e^{\rho(t_0)} \right).
\]

Plugging this back in (38), we get
\[
\| f(t) \|^2 \pi(t) \leq \frac{\| f(t_0) \|^2 \pi(t_0)}{e^{\rho(t)}} + 4 \left( 1 - \frac{e^{\rho(t_0)}}{e^{\rho(t)}} \right) \leq 4 + \| f(t_0) \|^2 \pi(t_0) \leq 4 + k.
\]
This bounds $\| f(t) \|^2 \pi(t)$ and hence concludes the proof.

One can deduce from the previous result, that the convergence rate is upper bounded by $t^{-\frac{\Delta_{\min}}{2\gamma}}$. Depending upon the energy landscape (through the dependence of $\gamma$ on $\gamma^*$), this can be arbitrarily bad compared to the fully connected setting.
To establish the regret, we use can proceed the same way as in Theorem 2. We first recover the discrete time process \( \{a_n\}_{n \geq 1} \) from \( \{X(t)\}_{t > 0} \) by using the following identification: \( a_n = X(t) \), for \( t \in [J_n, J_{n+1}) \). For simplicity we assume that the initial exploration phase is performed by the breadth first search algorithm, which has complexity of \( O(e + k) \), where \( e \) is the number of edges. Then, we have,

**Theorem 9.** Assume the conditions of lemma 8 are fulfilled. Then, the regret of Algorithm 2 is bounded by

\[
R_n \leq \left( \frac{4\sqrt{\pi}e \pi^2}{\gamma \pi} \right)^{\frac{1}{\gamma}} + (e + k)n^\gamma (\log n)^2 + \frac{\sqrt{g(4 + k)}}{a(1 - \Delta_{\min}^2 \gamma)} (an + 1)^{1 - \Delta_{\min}^2 \gamma}
\]

for any \( \epsilon > 0 \) and \( \gamma > 2 \gamma^* \).

We consider the implications of the initial exploration phase which requires \( T_i(t_o) \geq n^\gamma (\log n)^2 \), \( \epsilon > 0 \). It is obvious that for large action spaces (or continuous ones), such a condition may be infeasible or even unnecessary in case of some metric structure on \( A \). We discuss this in the following remark:

**Remark 10.** In making simulated annealing more efficient for solving stochastic/bandit optimization with large/continuous action sets, one must be cognizant of the fact that the Markov chain structure of discrete simulated annealing naturally arises in many combinatorial problems which can have local or global least cost configurations. For stochastic optimization, a graphical structure on the domain set may seem contrived and one may as well assume a fully connected setting, i.e. any solution point can be accessed from the current one. Given this, we conjecture that a more structured approach towards selecting the possible candidate for transition can help SA perform better. One may use Algorithm 1 as a subroutine in conjunction with a doubling trick to tackle continuous action sets where the pay-off function has some regularity (see Algorithm CAB1 in [31]). Another possible approach could be to use the tree based search technique of [32]. However, in general, any search method employed for optimization on Euclidean spaces has the disadvantage of the convergence rate depending on the dimension of the search space (see e.g. Chapter 2 in [33]).

**Remark 11.** The monotonically decreasing cooling schedule of simulated annealing has been a point of debate ever since it was proposed. The case of noisy observations, as one can guess, inherits these problems. In fact, one can even question the wisdom of decreasing \( T(t) \to 0 \) for the noisy case, given that one needs to visit each state infinitely often to get exact convergence. It is a common practice to employ a constant time schedule in bandit algorithms (see e.g. the detailed experiments performed in [34]), even though the theoretical guarantees may not be exact. We can suggest the following alternative here: Let \( T_{e_{ij}}(t) \) denote the number of visits to edge \( e_{ij} \in A \times A \) determined by \( i, j \), then we defined the cooling schedule as:

\[
\beta(t) := \frac{\log T_{e_{an,a'}}(t)}{\gamma},
\]

where \( a' \in N(a_n) \) is the candidate arm uniformly selected from the neighbourhood of the current arm \( a_n \). This makes the cooling schedule depend on the state of the algorithm. By a limiting pigeon-hole argument for a finite action set (which implies that at least one edge in \( G \) will be visited infinitely often), one can see that \( \limsup \beta(t) \to \infty \). Although, the previous results may not hold with the same guarantees since the time dependence in the upper bound will be through \( T_{e_{an,a'}}(t) \) instead of \( t \), we conjecture that this time step may lead to a more adaptive approach to exploration.
Appendix A. Numerical Experiments

In this appendix, we provide empirical confirmation of the efficiency of the proposed algorithm for solving the stochastic MAB. An instance of the bandit problem is characterized by \( k \) (number of arms) and the reward distributions (assumed to be normal with variance \( \sigma^2 \)). We consider a reward maximization problem with the reward means calculated according to:

\[
\mu_a = \begin{cases} 
0, & \text{if } a = a^* \\
-\Delta_{\text{min}} - |\mathcal{N}(0, 0.1)|, & \text{if } a \in A \setminus \{a^*\}
\end{cases}
\]

The value \( \Delta_{\text{min}} \) has been set to 0.1 for all experiments performed here.

We briefly describe the methodology of the algorithms we use for performance comparison:

(i) \( \epsilon \)-Greedy: At round \( t \), the probability \( P_a(t) \) of selecting arm \( a \) is given by:

\[
P_a(t) = \begin{cases} 
1 - \epsilon + \frac{\epsilon}{k}, & \text{if } a = \arg \max_{a' \in A} \hat{\mu}_{a'}(t) \\
\epsilon \frac{1}{k}, & \text{otherwise}.
\end{cases}
\]

(ii) UCB: The UCB family of algorithms incorporate the idea of optimism in the face of uncertainty to determine the policy. Initially, each arm is played once and subsequently for any round \( t \), the algorithm greedily selects arm \( a_t \) according to:

\[
a_t \in \arg \max_{a' \in A} \left\{ \hat{\mu}_{a'}(t) + \sqrt{\frac{2 \ln t}{T_{a'}(t)}} \right\}
\]

(iii) Boltzmann Exploration/Softmax: This algorithm is a softmax method where the probability of picking the arm is decided by the Boltzmann distribution, i.e. the rprobability of selecting an arm is proportional to its current empirical mean:

\[
P_a(t + 1) := \frac{e^{\hat{\mu}_a(t)}}{\sum_{a' \in A} e^{\hat{\mu}_{a'}(t)}},
\]

where \( \tau \) controls the randomness of the choice.

The results have been plotted for two criterion: (i) Fraction of optimal arm plays (ii) Regret accumulated till time \( n \). The parameters for all the comparison algorithm have been set according to [34]. As is evident from the plotted results, the performance of the tested algorithms varies to a great degree depending on \( k \) and \( \sigma^2 \). The broad conclusion that one can draw about the proposed algorithm is that it has the most robust performance across all the metrics we have tested against. For large variances, the performance of the algorithms do not differ to any significant degree except for the case of 2-armed bandit where the UCB algorithm dominates for large variances but the performance degrades significantly for low variances (this is consistent with the observations of [34]).
Appendix B. List of Notations

References

[1] H. Robbins, “Some aspects of the sequential design of experiments,” *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[3] S. Bubeck and N. Cesa-Bianchi, “Regret analysis of stochastic and nonstochastic multi-armed bandit problems,” arXiv preprint arXiv:1204.5721, 2012.

[4] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[5] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[6] D. Bertsimas, J. Tsitsiklis, et al., “Simulated annealing,” *Statistical science*, vol. 8, no. 1, pp. 10–15, 1993.

[7] B. Hajek, “Cooling schedules for optimal annealing,” *Mathematics of operations research*, vol. 13, no. 2, pp. 311–329, 1988.

[8] D. Mitra, F. Romeo, and A. Sangiovanni-Vincentelli, “Convergence and finite-time behavior of simulated annealing,” in *1985 24th IEEE Conference on Decision and Control*, pp. 761–767, IEEE, 1985.

[9] B. Gidas, “Nonstationary markov chains and convergence of the annealing algorithm,” *Journal of Statistical Physics*, vol. 39, no. 1-2, pp. 73–131, 1985.

[10] J. N. Tsitsiklis, “Markov chains with rare transitions and simulated annealing,” *Mathematics of Operations Research*, vol. 14, no. 1, pp. 70–90, 1989.

[11] R. Holley and D. Stroock, “Simulated annealing via sobolev inequalities,” *Communications in Mathematical Physics*, vol. 115, no. 4, pp. 553–569, 1988.

[12] S. B. Gelfand and S. K. Mitter, “Simulated annealing with noisy or imprecise energy measurements,” *Journal of Optimization Theory and Applications*, vol. 62, no. 1, pp. 49–62, 1989.

[13] W. J. Gutjahr and G. C. Pflug, “Simulated annealing for noisy cost functions,” *Journal of global optimization*, vol. 8, no. 1, pp. 1–13, 1996.

[14] C. Bouttier and I. Gavra, “Convergence rate of a simulated annealing algorithm with noisy observations,” *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 127–171, 2019.

[15] J. Branke, S. Meisel, and C. Schmidt, “Simulated annealing in the presence of noise,” *Journal of Heuristics*, vol. 14, no. 6, pp. 627–654, 2008.

[16] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.

[17] D. S. Johnson, C. R. Aragon, L. A. McGeoch, and C. Schevon, “Optimization by simulated annealing: An experimental evaluation; part i, graph partitioning,” *Operations research*, vol. 37, no. 6, pp. 865–892, 1989.
[18] D. S. Johnson, C. R. Aragon, L. A. McGeoch, and C. Schevon, “Optimization by simulated annealing: an experimental evaluation; part ii, graph coloring and number partitioning,” *Operations research*, vol. 39, no. 3, pp. 378–406, 1991.

[19] C. Koulamas, S. Antony, and R. Jaen, “A survey of simulated annealing applications to operations research problems,” *Omega*, vol. 22, no. 1, pp. 41–56, 1994.

[20] R. Pemantle, “Vertex-reinforced random walk,” *Probability Theory and Related Fields*, vol. 92, no. 1, pp. 117–136, 1992.

[21] M. Benaïm *et al.*, “Vertex-reinforced random walks and a conjecture of pemantle,” *The Annals of Probability*, vol. 25, no. 1, pp. 361–392, 1997.

[22] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*, vol. 48. Springer, 2009.

[23] J. Hofbauer, K. Sigmund, *et al.*, *Evolutionary games and population dynamics*. Cambridge university press, 1998.

[24] W. H. Sandholm, *Population games and evolutionary dynamics*. MIT press, 2010.

[25] D. L. Isaacson and R. W. Madsen, *Markov chains theory and applications*. 1976.

[26] E. Seneta, *Non-negative matrices and Markov chains*. Springer Science & Business Media, 2006.

[27] T.-S. Chiang and Y. Chow, “On eigenvalues and annealing rates,” *Mathematics of operations research*, vol. 13, no. 3, pp. 508–511, 1988.

[28] A. D. Venttsel, “On the asymptotics of eigenvalues of matrices with elements of order e^{- Vij/(2ε^2)},” in *Doklady Akademii Nauk*, vol. 202, pp. 263–265, Russian Academy of Sciences, 1972.

[29] J. N. Tsitsiklis, “A survey of large time asymptotics of simulated annealing algorithms,” in *Stochastic Differential Systems, Stochastic Control Theory and Applications*, pp. 583–599, Springer, 1988.

[30] D. W. Stroock, *An introduction to Markov processes*, vol. 230. Springer Science & Business Media, 2013.

[31] R. D. Kleinberg, “Nearly tight bounds for the continuum-armed bandit problem,” in *Advances in Neural Information Processing Systems*, pp. 697–704, 2005.

[32] S. Bubeck, G. Stoltz, C. Szepesvári, and R. Munos, “Online optimization in x-armed bandits,” in *Advances in Neural Information Processing Systems*, pp. 201–208, 2009.

[33] S. Bubeck, “Convex optimization: Algorithms and complexity,” *arXiv preprint arXiv:1405.4980*, 2014.

[34] V. Kuleshov and D. Precup, “Algorithms for multi-armed bandit problems,” *arXiv preprint arXiv:1402.6028*, 2014.
Parameters \((k = 2, \sigma \in \{0.01, 0.1, 1\}, \Delta_{\text{min}} = 0.1)\):

\(\sigma = 0.01\) : SA, \(\gamma^{-1} = 0.002\); Softmax, \(\tau = 0.001\); \(\epsilon\)-greedy, \(\epsilon = 0.005\).

\(\sigma = 0.1\) : SA, \(\gamma^{-1} = 0.01\); Softmax, \(\tau = 0.01\); \(\epsilon\)-greedy, \(\epsilon = 0.001\).

\(\sigma = 1\) : SA, \(\gamma^{-1} = 0.2\); Softmax, \(\tau = 0.1\); \(\epsilon\)-greedy, \(\epsilon = 0.05\).
Parameters ($k = 10$, $\sigma \in \{0.01, 0.1, 1\}$, $\Delta_{\text{min}} = 0.1$):

- $\sigma = 0.01$ : SA, $\gamma^{-1} = 0.001$; Softmax, $\tau = 0.001$; $\epsilon$-greedy, $\epsilon = 0.001$.
- $\sigma = 0.1$ : SA, $\gamma^{-1} = 0.01$; Softmax, $\tau = 0.01$; $\epsilon$-greedy, $\epsilon = 0.005$.
- $\sigma = 1$ : SA, $\gamma^{-1} = 0.05$; Softmax, $\tau = 0.05$; $\epsilon$-greedy, $\epsilon = 0.1$. 

30
Parameters ($k = 50$, $\sigma \in \{0.01, 0.1, 1\}$, $\Delta_{\text{min}} = 0.1$):

$\sigma = 0.01$: SA, $\gamma^{-1} = 0.001$; Softmax, $\tau = 0.001$; $\epsilon$-greedy, $\epsilon = 0.005$.

$\sigma = 0.1$: SA, $\gamma^{-1} = 0.01$; Softmax, $\tau = 0.01$; $\epsilon$-greedy, $\epsilon = 0.005$.

$\sigma = 1$: SA, $\gamma^{-1} = 0.01$; Softmax, $\tau = 0.007$; $\epsilon$-greedy, $\epsilon = 0.01$. 

31
| Notation | Description |
|----------|-------------|
| $\|x\|$, $x \in \mathbb{R}^k$ | $\sum_{i \in k} |x_i|$ |
| $\|P\|$, $P \in \mathbb{R}^{m \times n}$ | $\max_{1 \leq i \leq m} \sum_{j=1}^n |P_{ij}|$ |
| $\|f\|_\pi$, $f : \mathcal{A} \to \mathbb{R}^k$ | $\sqrt{\langle |f|^2 \rangle_\pi}$ where $\langle g \rangle_\pi := \sum_{i \in \mathcal{A}} g_i \pi_i$ |
| $\mathcal{A}$ | Arm Set. |
| $\mathcal{G} = \{\mathcal{A}, \mathcal{E}\}$ | Graph with edge set $\mathcal{E}$. |
| $n$ | Total number of rounds. |
| $k$ | Total number of arms. |
| $\mu$ | Loss vector with component $\mu_i$ denoting the loss of arm $i$. |
| $\mathcal{E}^k_{SG}(\sigma^2)$ | Sub Gaussian Bandits with variance $\sigma^2$ and $|\mathcal{A}| = k$. |
| $\hat{\mu}_i(t)$, $\hat{\mu}_i, T_a(t)$ | Empirical average of the loss at arm $i$ at round $t$ or with $T_a(t)$ samples. |
| $\Delta_i$ | $\mu_i - \mu_{a^*}$. |
| $\nu(n)$ | Probability distribution of selecting the arms (i.e. policy) at time $n$. |
| $T_a(t)$ | Total number of pulls of arm $i$ at round $t$. |
| $\gamma^*$ | Critical Depth (Definition 3). |
| $\beta(t)$ | $1/T_i$. |
| $P(t)$ | Transition probability at round $t$, see (2). |
| $\mathbf{P}(m, n)$ | $\prod_{t=m}^{n-1} P(t)$. |
| $L$ | $\max_{i \in \mathcal{A}} \max_{j \in N(i)} |\mu_j - \mu_i|$. |
| $R$ | $\min_{i \in \mathcal{A}} \min_{j \in N(i)} \frac{g(i,j)}{g(i)g(j)}$. |
| $g(i,j)/g(i)$ | Probability of selecting $j$ while in node $i$. |
| $g(i)$ | $\sum_{j \in N(i)} g(t,j)$ |
| $\pi(t)$ | Quasi stationary distribution of $P(t)$, see eq. (3). |
| $\kappa(P)$ | Ergodicity co-efficient of matrix $P$. |
| $\text{Var}_{\pi}(f)$ | $\|f - \langle f \rangle_{\pi}\|_2^2$. |
| $Q(t)$ | $Q$-matrix. |
| $\mathcal{E}_a(f, f)$ | $\frac{1}{2} \sum_{j \neq i} \pi_j Q_{ij}(t)(f_j - f_i)^2$ |
| $\lambda(t)$ | $\inf\{\mathcal{E}_a(f, f) : \text{Var}_{\pi(t)}(f) = 1\}$. |