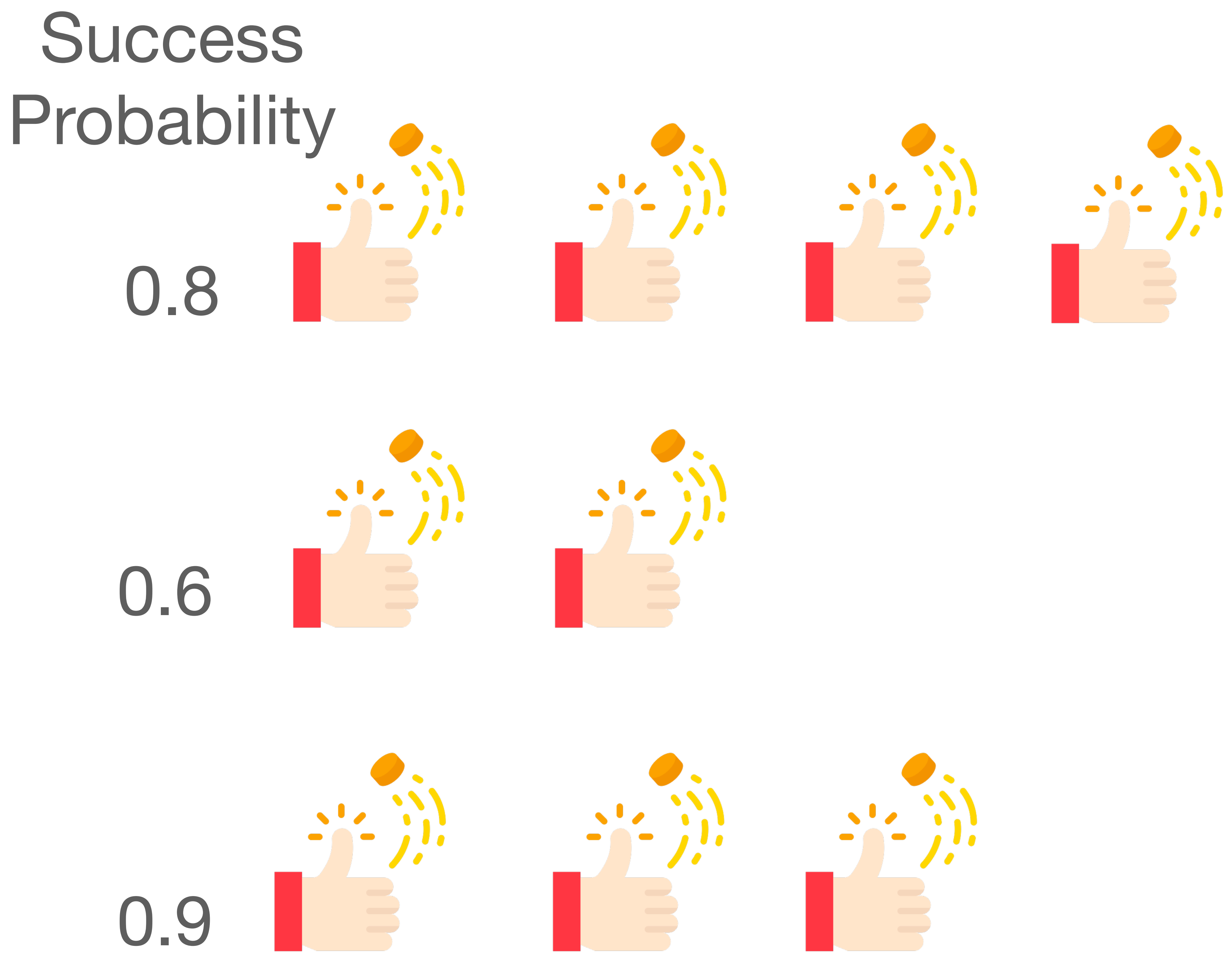
Flipping Coins to Win!

Sandeep Juneja, Ashoka

IIT Madras
I Conf on Stochastic Calc and Finance
June 4, 2024



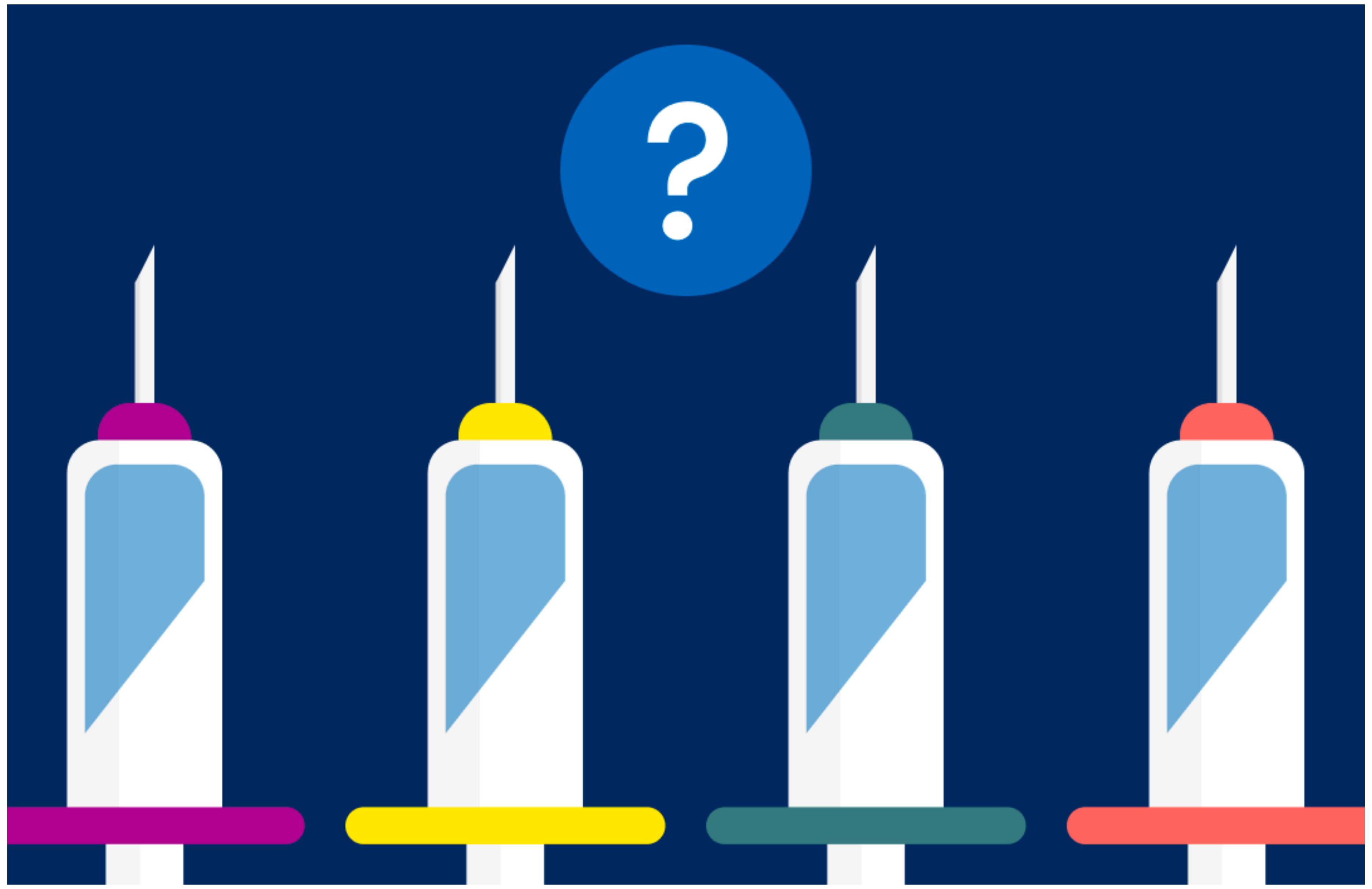
Multi armed bandits: What do we discuss

- Sequentially generate samples from a number of arms
- To maximise long term stochastic reward (optimally manage explore and exploit trade-off)
- Simple and yet interesting setting to illustrate the underlying conceptual ideas
- A large number or practical applications esp. in online settings fit these settings with some adjustments

Which coin do you sample next?

To maximise expected reward

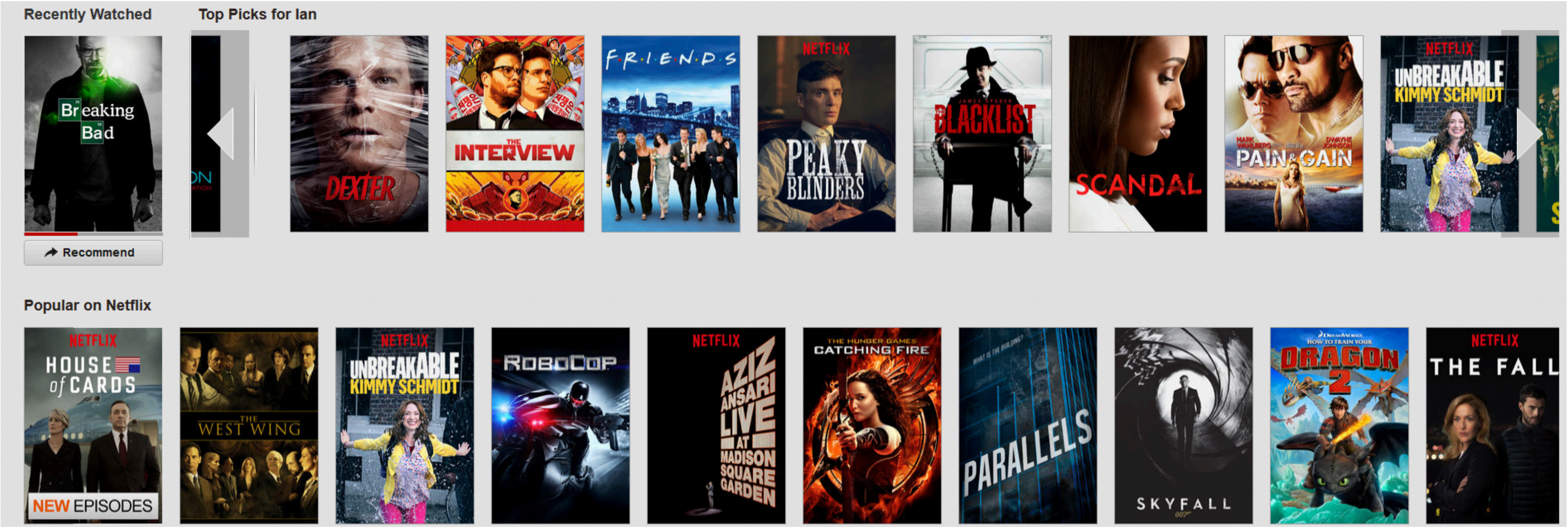
Applications: Clinical trials



- Four vaccines (or experimental drugs). Which ones to give to patients
- 'it seems apparent that a considerable saving of individuals otherwise sacrificed to the inferior (drug) treatment might be effected' Thompson, 1933
-

Applications

- Placing advertisements on a Google search
- Web construction amongst many options
- Recommendation systems
    - Movies/products to recommend
    - Facebook posts to show
    - News paper articles to bring to your attention
    - Price to offer for a digital good
- Travel route to recommend amongst many



Maximise expected reward

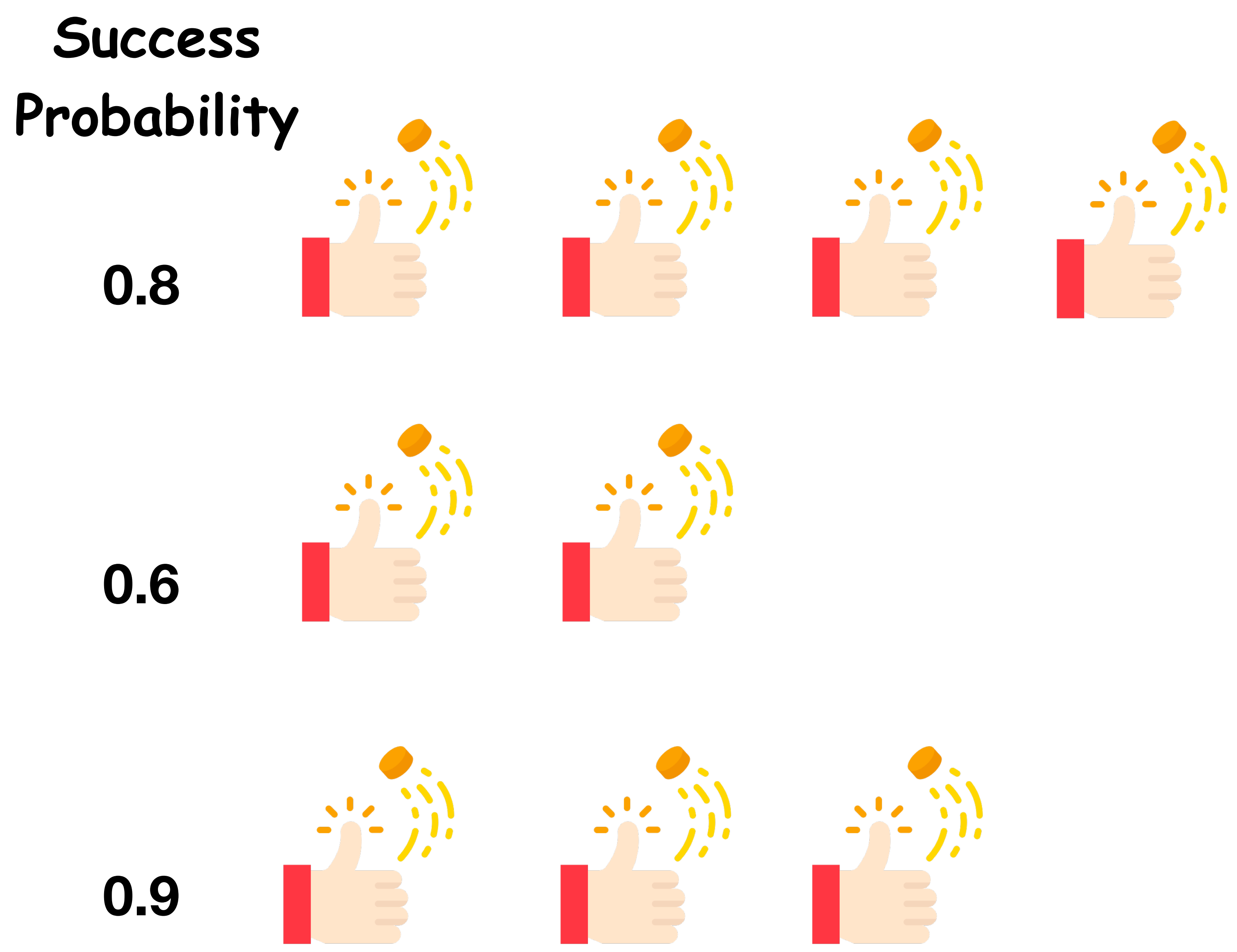
or

Minimise expected regret

Stochastic regret minimisation problem (Lai and Robbins 1985)

Given K unknown probability distributions (Coins) that can be sampled from, sample to maximise expected reward or, equivalently, minimise the expected regret in n steps.

What is the best explore and exploit trade-off



Stochastic regret minimisation

K Bernoulli arms with unknown means $(\mu_1, \mu_2, \ldots, \mu_K)$.

W.l.o.g. $\mu_1 > \max_{a \geq 2} \mu_a$. Expected regret $\mu_1 - \mu_a$ is suffered every time a sub-optimal arm a is pulled

Algorithm generates samples sequentially

Aim: Max $\sum_{t=1}^{T} EX_t$

equivalently Min $ER_T = T \times \mu_1 - \sum_{t=1}^{T} EX_t$

or Min $ER_T = \sum_{a=1}^{K} (\mu_1 - \mu_a) \times EN_a(T)$

Stochastic regret minimisation problem

Some simple strategies

Egalitarian principle: Equal samples to all

Each arm is given T/K samples

Regret equals $$\frac{T}{K}\sum_{a}(\mu_1 - \mu_a)$$

Linear in time T !

Greedy strategy: Follow the leader

Pull arm with the largest sample mean thus far

Consider one coin heads w.p. 0.9.

Other heads w.p. 0.6

Regret at least 0.06 T, so linear

Is sub linear regret possible?

Explore then commit when $\mu_1 - \mu_2$ is known

Sample each arm m times.

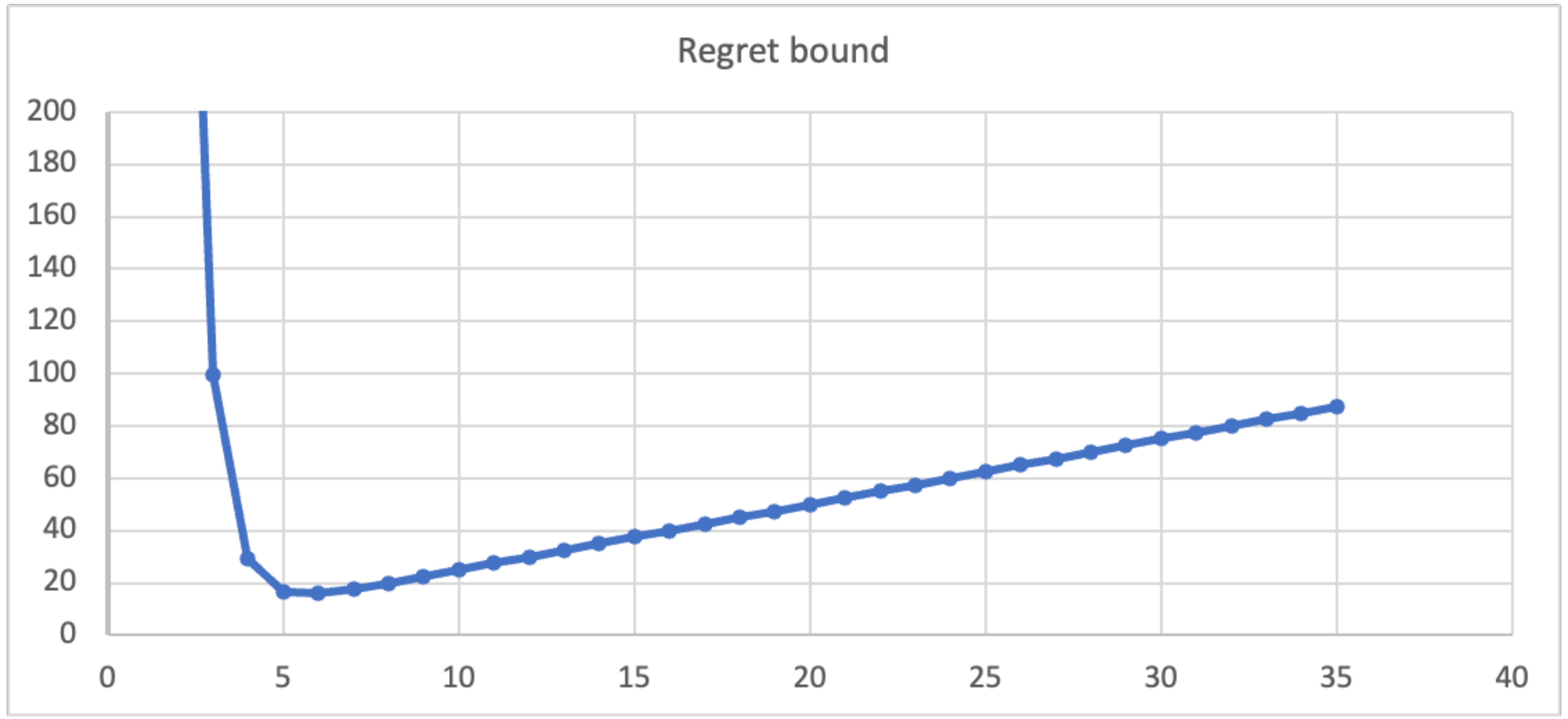
Thereafter sample the empirical winner for remaining T-Km trials

Regret in two arms $N(\mu_1,1)$ and $N(\mu_2,1)$ setting, $\mu_1 > \mu_2$ setting

$$m(\mu_1 - \mu_2) + (T - 2m) \times \exp(-m(\mu_1 - \mu_2)^2/4))$$

Explore then commit strategy

Minimum at $m = \Theta(\log T)$



$T = 10{,}000, \quad \mu_1 - \mu_2 = 2.5$

Regret $\leq \Theta(\log T)$

Logarithmic regret! Requires knowledge of T and $(\mu_1 - \mu_2)$.

Successive elimination algo (Bounded [0, 1] rv) $\alpha(t) = \sqrt{\frac{2\log T}{t}}$

1. Sample each active arm once. Compute indexes

$$UCB_a(t) = \bar{X}_{a,t} + \alpha(t) \text{ and } LCB_a(t) = \bar{X}_{a,t} - \alpha(t)\,.$$

2. Eliminate arms for which $UCB_a(t) < \max_a LCB_a(t)$

4. If a single arm remains, then assign remaining samples to this arm.

5. Increment t and repeat

Our friend: Hoeffding

Each $X_i \in [-1,1]$ are independent, identically distributed with zero mean

Law of large numbers, Central limit theorem

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx 0 + \frac{1}{\sqrt{n}} N(0,1)$$

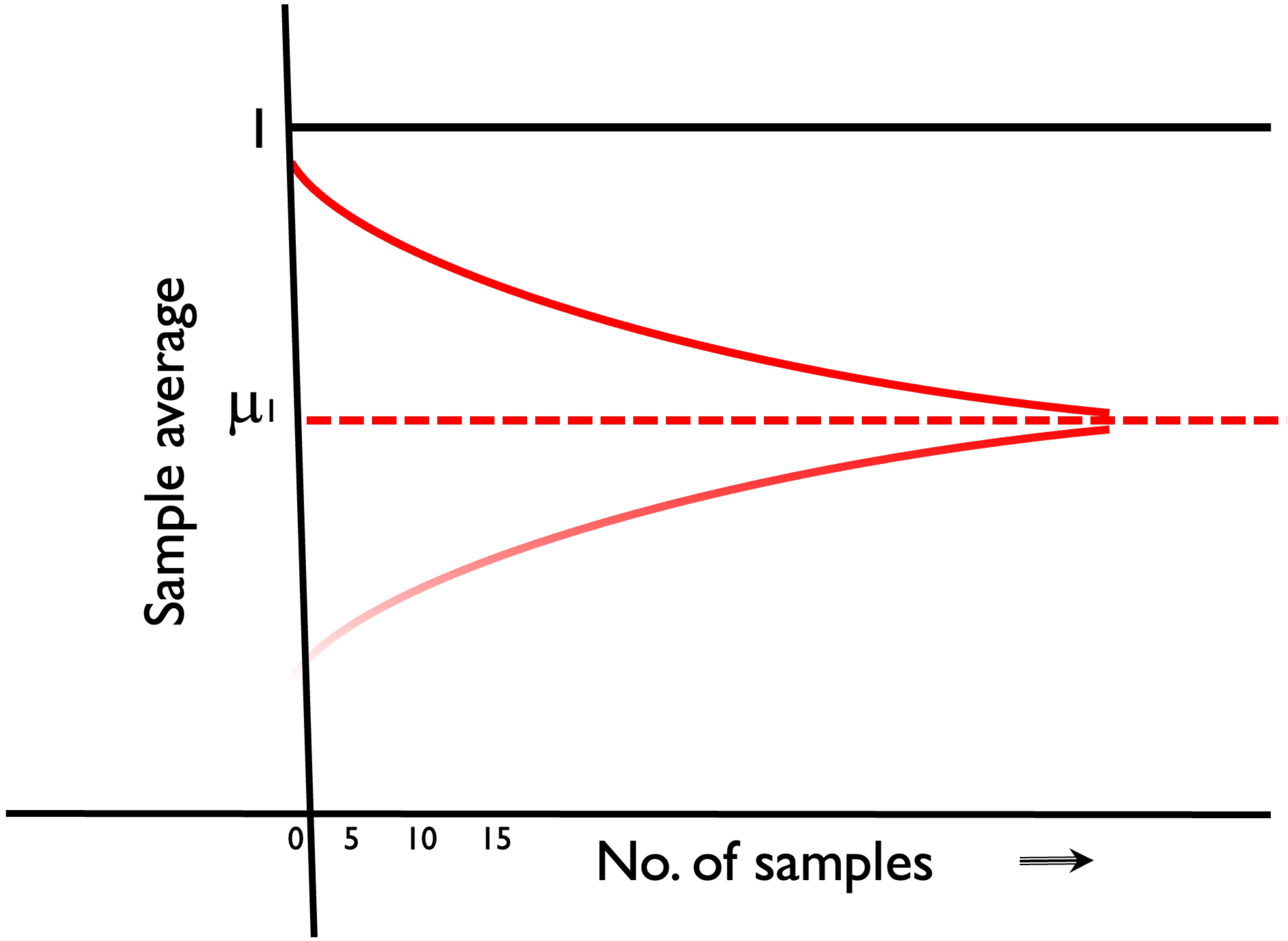
Hoeffding's Inequality captures large deviations -

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \epsilon\right) \leq \exp(-n\epsilon^2/2)\,.$$

Concentration inequality

$$\alpha_t = \sqrt{\frac{2\log T}{t}}$$

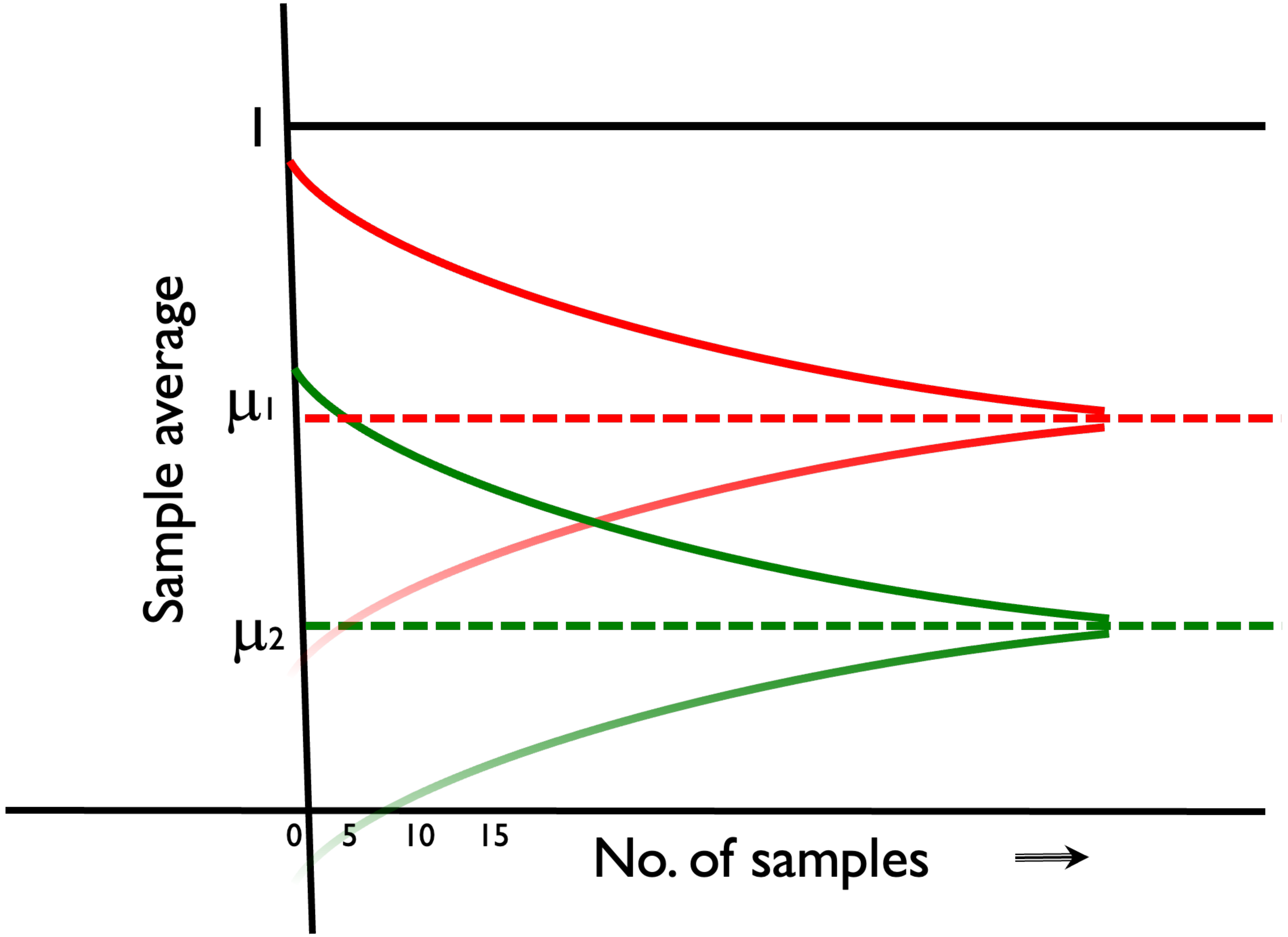
$$P\left(\text{there exists } t \le T: \ \frac{1}{t}\sum_{i=1}^{t} X_i \ge \alpha_t\right) \le \sum_{t=1}^{T} P\left(\frac{1}{t}\sum_{i=1}^{t} X_i \ge \alpha_t\right) \le \quad 1/T$$



$\bar{X}_t \in \mu \pm \alpha_t$ for all t
with probability 1-1/T

Successive elimination algorithm

$$\alpha(t) = \sqrt{\frac{2\log T}{t}}$$



$$UCB_a(t) = \bar{X}_{a,t} + \alpha(t)$$
$$LCB_a(t) = \bar{X}_{a,t} - \alpha(t)$$

If

$$UCB_a(t) < \max_a LCB_a(t)$$

eliminate

Instance dependent regret

Best arm, arm 1 will never be rejected on the good set. Arm 1 loses if

$$\bar{X}_{1,t} < \bar{X}_{a,t} - 2\alpha(t)$$

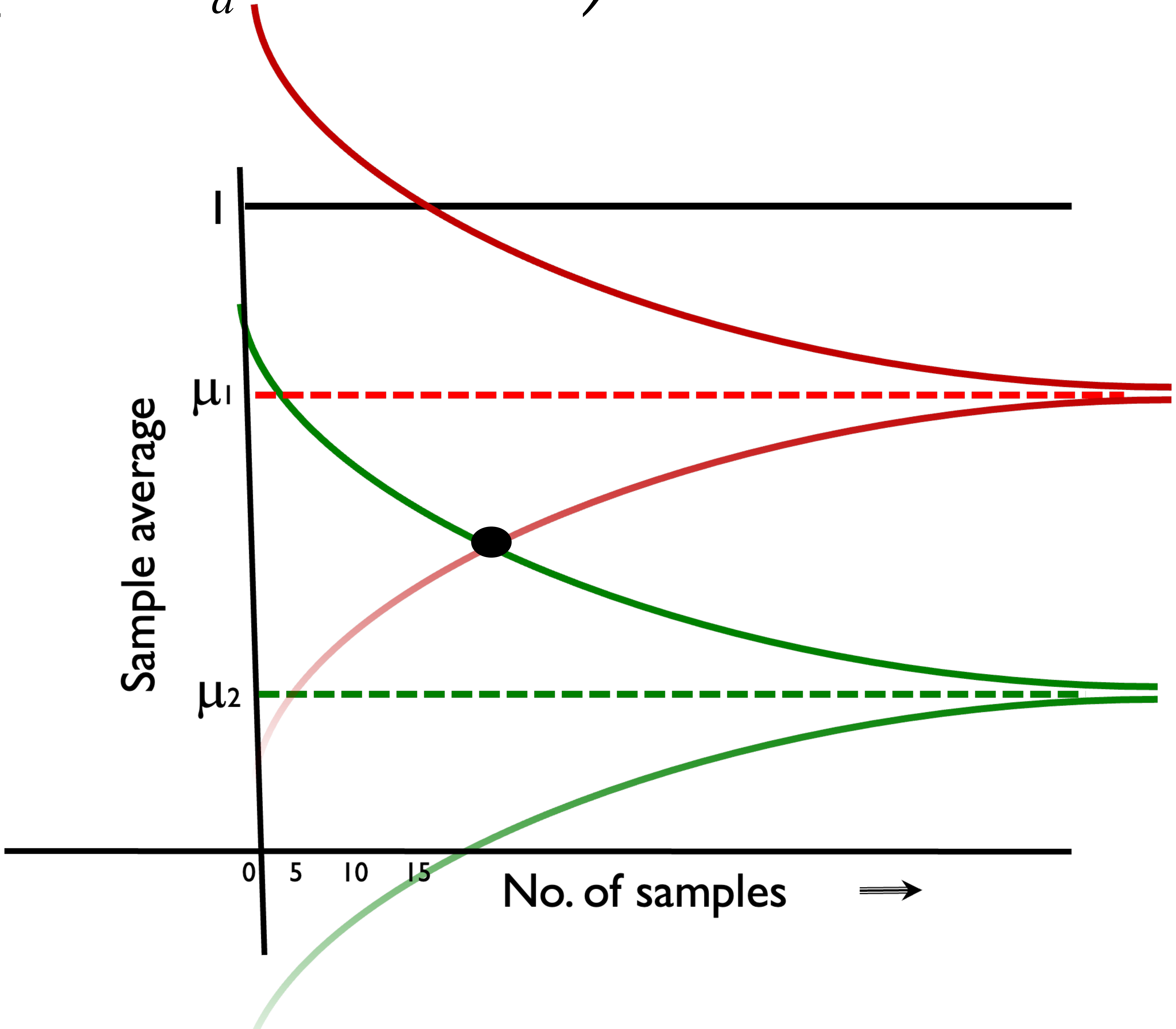
But on a good set

$$\bar{X}_{1,t} \geq \mu_1 - \alpha(t) \geq \mu_a - \alpha(t) \geq \bar{X}_{a,t} - 2\alpha(t)$$

Exp. regret $$O\left(\log(T)\sum_a \frac{1}{(\mu_{\max}-\mu_a)}\right)$$

Consider tubes

$$\bar{X}_t \in \mu \pm 2\alpha_t$$



Instance dependent regret

Suppose arm a rejected after sampled $t+1$ times.

$$\mu_a + 2\alpha(t) \geq \bar{X}_{a,t} + \alpha(t) \geq \bar{X}_{1,t} - \alpha(t) \geq \mu_1 - 2\alpha(t)$$

Thus, $(\mu_1 - \mu_a) \leq 4\alpha(t)$, or $t \leq 32(\mu_1 - \mu_a)^{-2} \log T$

So the total expected regret from the good set as well as from the rogue set is bounded from above by $K + 1 + 32 \log T \sum_{a \geq 2} (\mu_1 - \mu_a)^{-1}$.

Upper Confidence Bound Algorithm (Auer et al. 2002)

Form an optimistic upper confidence bound (UCB) on each arm

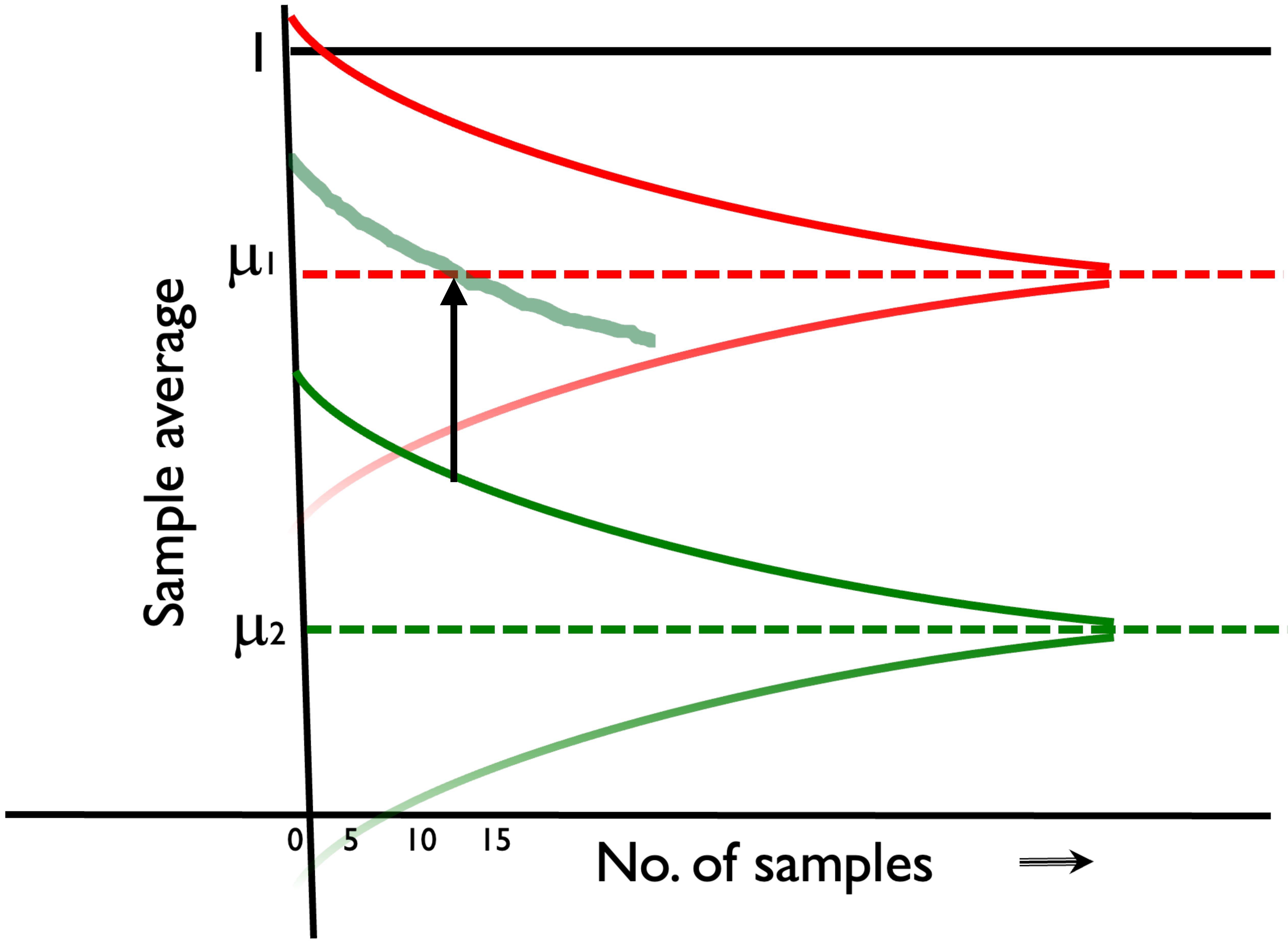
This UCB is greater than the sample average but converges to it as the number of samples increase

It increases if arm is not sampled for a long time - encouraging exploration

Algorithm simply involves sampling the arm with the largest UCB `Index'

Upper Confidence Bound Algorithm (Auer et al. 2002)

Adaptive arm selection



At each step t+1 select an arm with the largest value of index

$$\bar{X}_a(t) + \sqrt{\frac{2\log t}{N_a(t)}}$$

Upper Confidence Bound Algorithm

UCB does a good trade-off between explore and exploit.

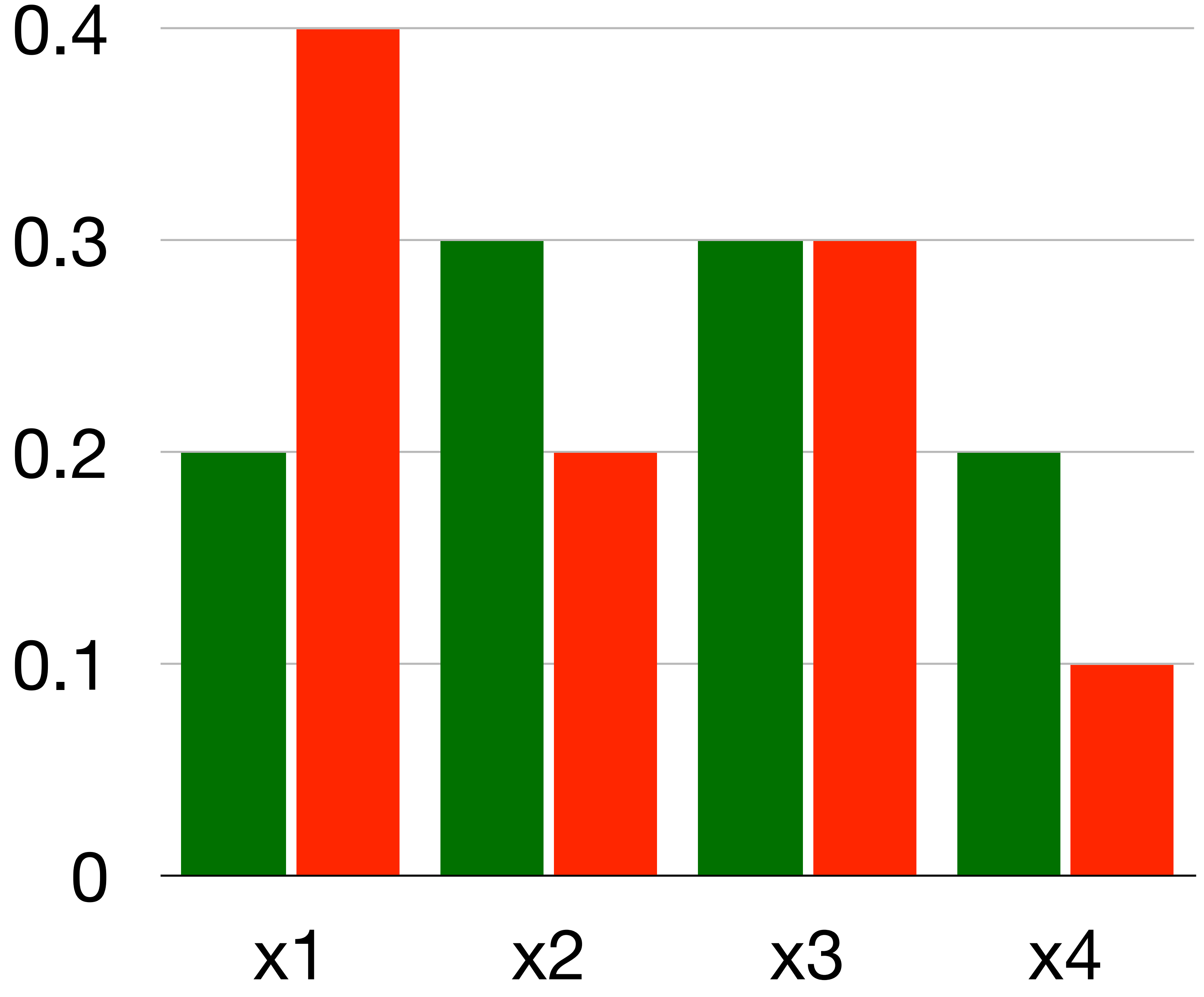
$$EN_a(T) \leq \frac{8 \log T}{\Delta_a^2} + 1 + \frac{\pi^2}{3}$$

Better than successive rejection

Lower bounds and algorithms that match even the constant in the lower bounds - general distributions

Large deviations result (Sanov's Thm.)

Green true dist ν. Red empirical dist μ (based on generated samples $(X_1, X_2, \ldots, X_n)$)



Prob of seeing emp dist μ when the true dist is ν

$$\approx \exp(-nKL(\mu \,|\, \nu))$$

where $KL(\mu \,|\, \nu) = \sum_{i=1}^{4} \mu_i \log\left(\frac{\mu_i}{\nu_i}\right)$

Lower bounds

Lai and Robbins 88, Burneta Katehakis 96

$$\liminf_{T \to \infty} \frac{EN_a(T)}{\log T} \geq \frac{1}{KL_{inf}(\mu_a, m(\mu_1))}$$

where $KL_{inf}(\mu_a, x) = \inf_{\nu \in \mathcal{L} : m(\nu) > x} KL(\mu_a, \nu)$

Heuristic argument for lower bound: Using Sanov's Thm.

For arm a and 1, generated samples with h.p. close to true dist.

Algorithm concerned that data of arm a coming from dist ν with $m(\nu) > m(\mu_1)$, and current data a large deviations leading to wrong conclusion.

Evidence needed so regret from potential error is small.

If m samples given to arm a. Chance that arm a is from dist ν and emp dist looks like μ_a

$$\approx \exp(-mKL(\mu_a | \nu)).$$

- Want m so error prob is order 1/T

So

$$m \geq \frac{\log T}{KL(\mu_a | \nu)}$$

Want this for all ν with $m(\nu) > m(\mu_1)$, hence

$$m \geq \frac{\log T}{KL_{\text{inf}}(\mu_a | m(\mu_1))}$$

Arm 1 gets most of T samples. Its large deviations not a concern

The Data Processing Inequality

$$KL(P_X | Q_X) \geq KL(P_{g(X)} | Q_{g(X)})$$

$$KL(P_\mu(X) | P_\nu(X)) \geq KL(P_\mu(I_E) | P_\nu(I_E))$$

$$KL(P_\mu(X) | P_\nu(X)) = \sum_{a=1}^{K} E_{P_\mu} N_a(T) KL(\mu_a | \nu_a)$$

KL-UCB Algorithm

We restrict arm distributions to

$$\mathscr{L} := \{ \text{ Probability measures } \eta : \mathbb{E}_{X\sim\eta}(|X|^{1+\epsilon} \leq B\}$$

Some conditions on the underlying distributions are necessary

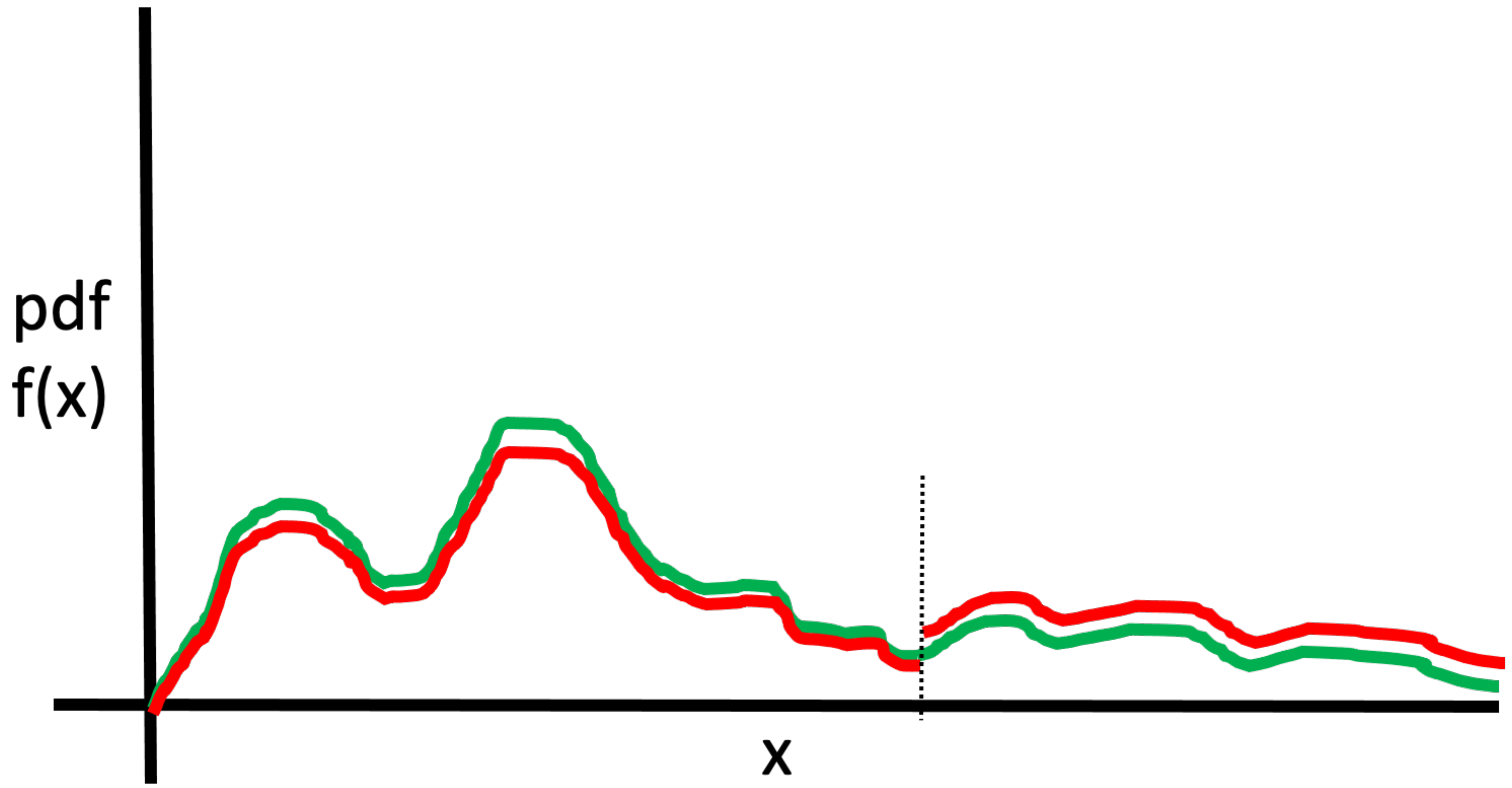
Glynn and J 2015

Easy to find two distributions

whose

KL distance is arbitrarily close

but means are arbitrarily far



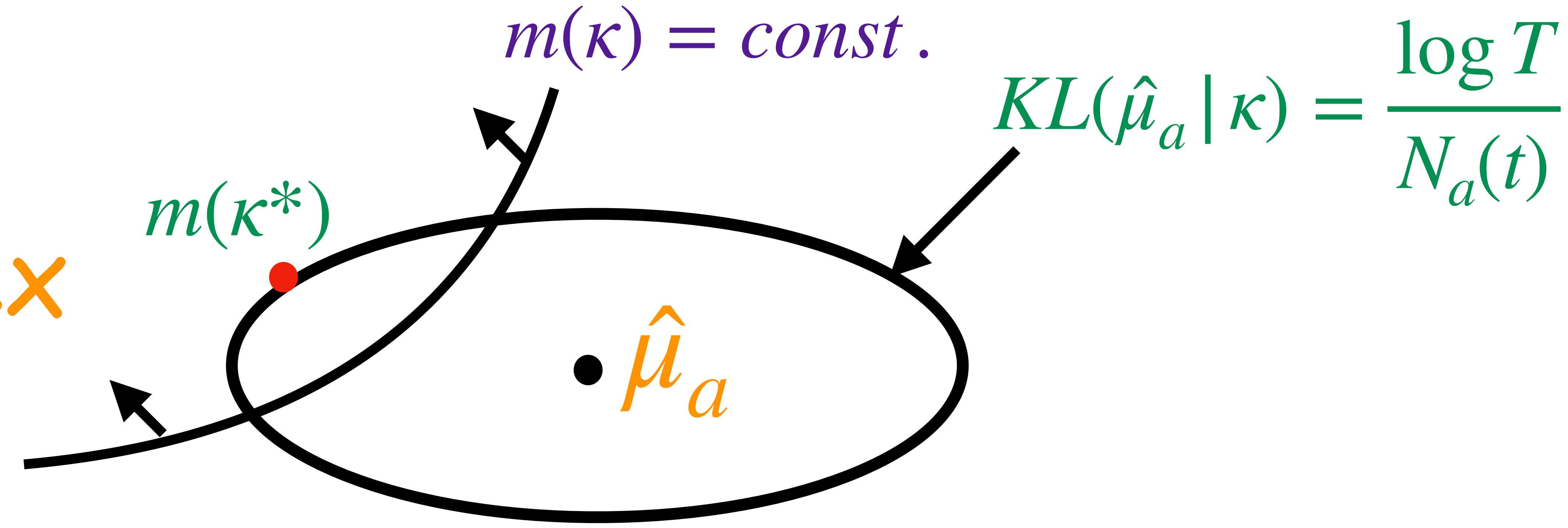
Intermission

https://www.jimmycarr.com/

KL-UCB Algorithm: Index based (Garivier, Cappe 2011, Agrawal, J, Koolen 2021)

A disc around empirical distribution

Largest mean in that disc is the index



$$U_a(t) = \sup\left\{ m(\kappa), \kappa \in \mathscr{L}, KL(\hat{\mu}_a | \kappa) \le \frac{\log T}{N_a(t)} \right\} = \sup\left\{ x : KL_{inf}(\hat{\mu}_a | x) \le \frac{\log T}{N_a(t)} \right\}$$

Matches the lower bound!

Heuristic argument on why the algorithm works

All indexes typically dominate their mean

At least one arm gets $\geq t/K$ samples. So its index close to its mean

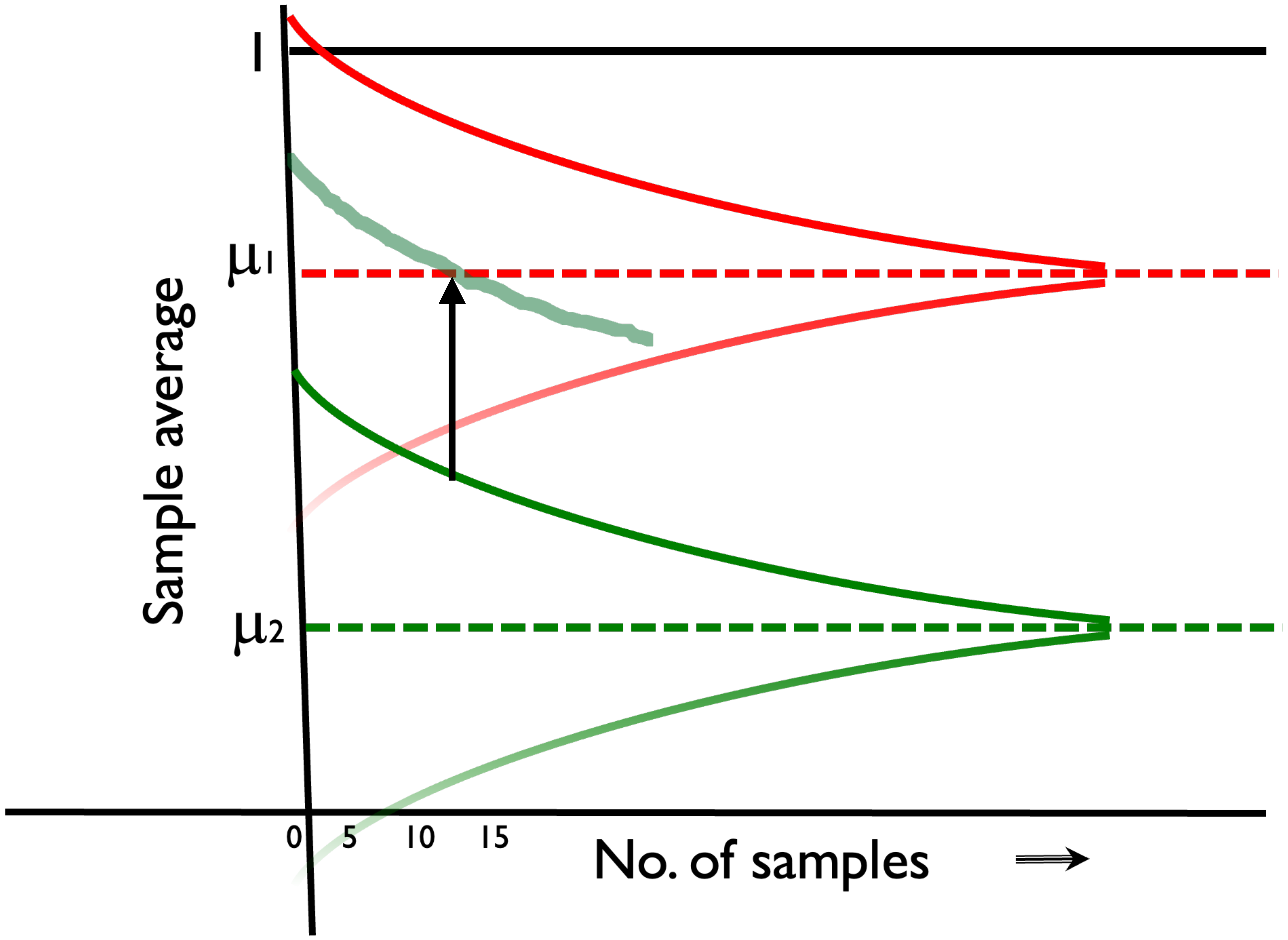
So arm 1 must get most of the samples

Every time arm $a \neq 1$ wins, its index just exceeds index of arm 1. Thus,

$$N_a(t) \approx \frac{\log t}{KL_{\inf}(\mu_a \,|\, m(\mu_1))}$$

KL Upper Confidence Bound Algorithm (for Bernoulli's)

Adaptive arm selection



Index

$$\sup \left\{ x : KL_{inf}(\hat{\mu}_2 \,|\, x) \leq \frac{\log T}{n} \right\}$$

Rigorous analysis requires bounding the times sub-optimal arms are pulled (Agrawal, J Glynn, 2020, Agrawal, J, Koolen 2021)

This relies on controlling probabilities such as

$$\mathbf{P}(\exists t \in \mathbf{N} : N_a(t) KL_{\inf}(\hat{\mu}_a(t), m(\mu_a)) \geq x)$$

Dual representations, exponential concave inequalities and mixture martingales cleverly used for this

Understanding $KL_{\inf}(\eta, x)$

It equals $\inf_{\kappa} \sum_i \log\left(\frac{\eta_i}{\kappa_i}\right)\eta_i$ such that

$$\sum_i |y_i|^{1+\epsilon}\kappa_i \leq B, \quad \sum_i y_i\kappa_i \geq x \text{ and } \sum_i \kappa_i = 1.$$

This is a convex program and is solved through Lagrangian duality.

Using duality, $KL_{\inf}(\eta, x)$ can be seen to equal

$$\max_{(\lambda_1,\lambda_2)\in\mathcal{R}_2} E_\eta \log(1-(X-x)\lambda_1-(B-|X|^{1+\epsilon}\lambda_2),$$ where

For empirical distribution $\hat{\mu}_a(n)$ we have $KL_{\inf}(\hat{\mu}_a(t), m(\mu_a))$ equals

$$\max_{(\lambda_1,\lambda_2)\in\mathcal{R}_2} \frac{1}{N_a(n)} \sum_{i=1}^{N_a(n)} \log(1-(X_i-m(\mu_1))\lambda_1-(B-|X_i|^{1+\epsilon})\lambda_2)).$$

In developing concentration inequality for this, the maximum function poses difficulties.
We observe that inside the maximum we have a sum of exp-concave functions.

Sum of exp concave functions: a useful inequality

Let $\Lambda \subseteq \Re^d$ be a compact and convex subset and q be the uniform distribution on Λ. Let $g_t : \Lambda \to \Re$ be any series of exp-concave functions. Then

$$\max_{\lambda \in \Lambda} \sum_{t=1}^{T} g_t(\lambda) \leq \log E_{\lambda \sim q} e^{\sum_{t=1}^{T} g_t(\lambda)} + d \log(T+1) + 1.$$

Thus $\max_{\lambda \in \Lambda} \exp\left(\sum_{t=1}^{T} g_t(\lambda)\right)$ is close to the expectation $E_{\lambda \sim q} e^{\sum_{t=1}^{T} g_t(\lambda)}$.

The latter is a mixture of super-martingales and hence is a super martingale.

Ville’s inequality

Ville's inequality: For a non-negative super martingale $(M_n : n \geq 0)$,

$$P(\exists n : M_n \geq x) \leq \frac{EM_0}{x}.$$

Donsker Varadhan Representation of KL Divergence

Let μ and ν be any probability measures on a common space. Then,

$$KL(\mu \,|\, \nu) = \sup_{g} \left(E_{\mu} g - \log E_{\nu} e^{g} \right).$$

Conclusion

- Introduced the regret minimisation problem along with practical applications
- Discussed many naive and then sensible rules for arm selection and analysed their performance
- Arrived at a lower bound on the samples needed
- Introduced KL_UCB algorithm that is optimal for general distributions