Concentration:
Introduction to Sequential Analysis

Master 2 Mathematics and Computer Science

Guillaume Aubrun, Aurélien Garivier, Rémi Gribonval

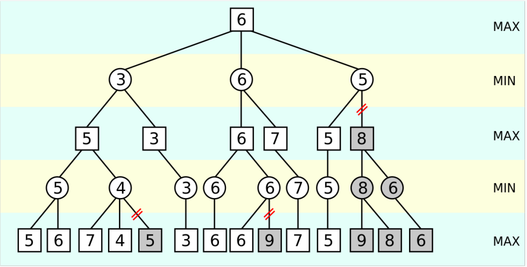
2020-2021

Table of contents

Sequential Decision Problems

Solving a Game

2 player game with finite number of actions

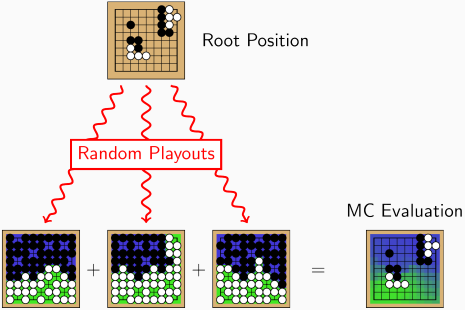


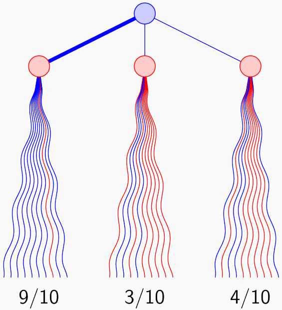
src: `wikipedia.org`

but too deep for exhaustive search of minimax action (by alpha-beta)
Example: Go ($\approx 10^{171}$ possible configurations)

Monte Carlo Tree Search

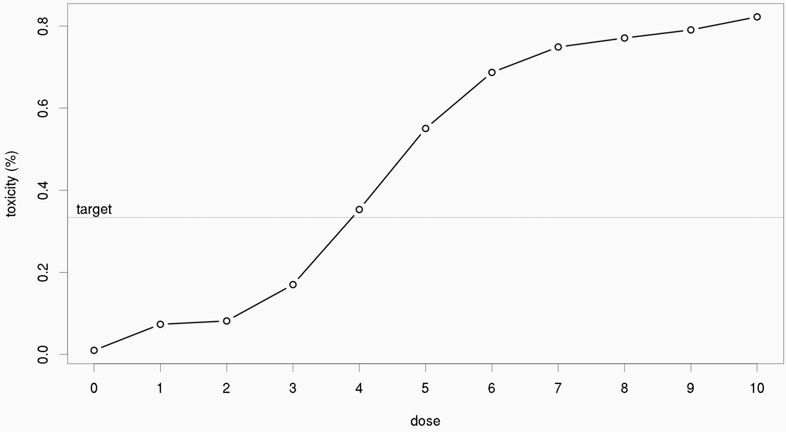
Heuristic search algorithm using random playouts / rollouts



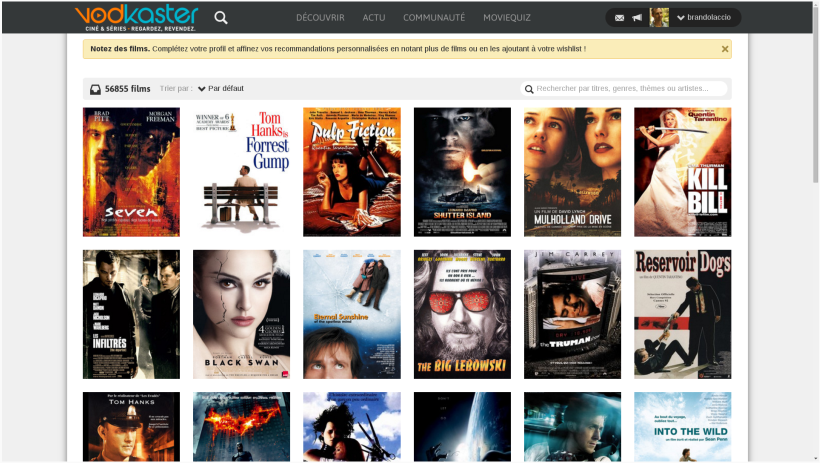


src: `https://www.remi-coulom.fr/`

Dose Finding

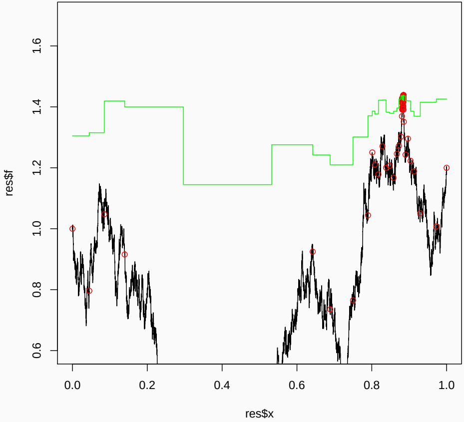


Content Recommandation



Optimization

- Goal : maximize function $f : C \subset \mathbb{R}^d \to \mathbb{R}$ possibly observed with noise
- Applications: computer experiment



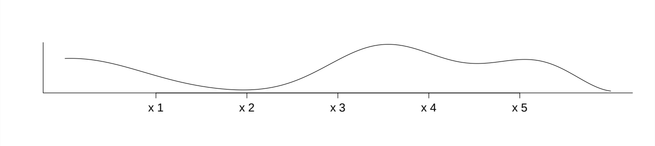
- Model: f comes from a Gaussian Process, or when it has a small norm in the induced RKHS.

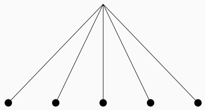
The Simple Bandit Model

Best-Arm Identification with Fixed Confidence

K options = probability distributions $\boldsymbol{\nu} = (\nu_a)_{1 \leq a \leq K}$

$\nu_a \in \mathcal{F}$ exponential family parameterized by its expectation μ_a





At round t, you may:

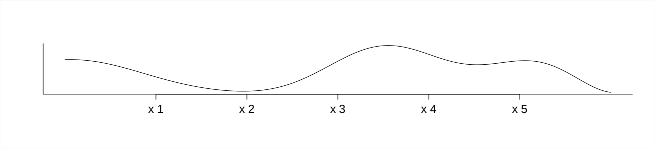
- choose an option $A_t = \phi_t(A_1, X_1, \dots, A_{t-1}, X_{t-1})) \in \{1, \dots, K\}$
- observe a new independent sample $X_t \sim \nu_{A_t}$

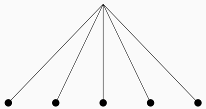
so as to identify the best option $a^* = \text{argmax}_a \, \mu_a$ and $\mu^* = \max_a \mu_a$
as fast as possible: stopping time τ .

Best-Arm Identification with Fixed Confidence

K options = probability distributions $\boldsymbol{\nu} = (\nu_a)_{1 \le a \le K}$

$\nu_a \in \mathcal{F}$ exponential family parameterized by its expectation μ_a





At round t, you may:

- choose an option $A_t = \phi_t(A_1, X_1, \ldots, A_{t-1}, X_{t-1}) \in \{1, \ldots, K\}$
- observe a new independent sample $X_t \sim \nu_{A_t}$

so as to identify the best option $a^* = \text{argmax}_a \, \mu_a$ and $\mu^* = \max_a \mu_a$
as fast as possible: stopping time τ_δ .

Intuition: a Simple Example

Most simple setting: for all $a \in \{1, \ldots, K\}$,

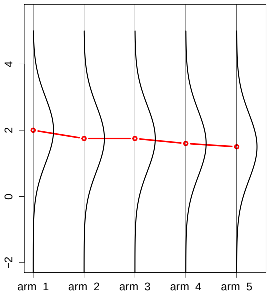
$$\nu_a = \mathcal{N}(\mu_a, 1)$$

For example: $\mu = [2, 1.75, 1.75, 1.6, 1.5]$.

At time t:

➜ you have sampled n_a times the option a

➜ your empirical average is $\bar{X}_{a,n_a}$.



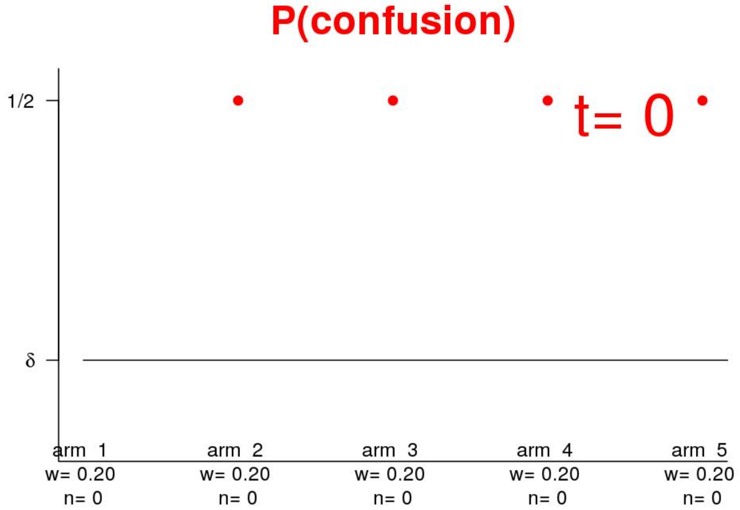
$\longrightarrow$ if you stop at time t, your probability of prefering arm $a \geq 2$ to arm $a^* = 1$ is:

$$\mathbb{P}\left(\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right) = \mathbb{P}\left(\frac{\bar{X}_{a,n_a} - \mu_a - \left(\bar{X}_{1,n_1} - \mu_1\right)}{\sqrt{1/n_1 + 1/n_a}} > \frac{\mu_1 - \mu_a}{\sqrt{1/n_1 + 1/n_a}}\right)$$

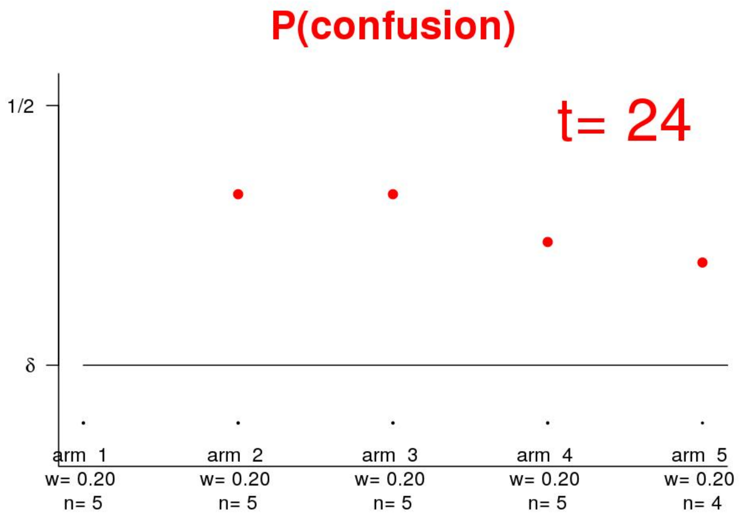
$$= \bar{\Phi}\left(\frac{\mu_1 - \mu_a}{\sqrt{1/n_1 + 1/n_a}}\right)$$

where $\bar{\Phi}(u) = \int_u^\infty \frac{e^{-u^2/2}}{\sqrt{2\pi}} du$

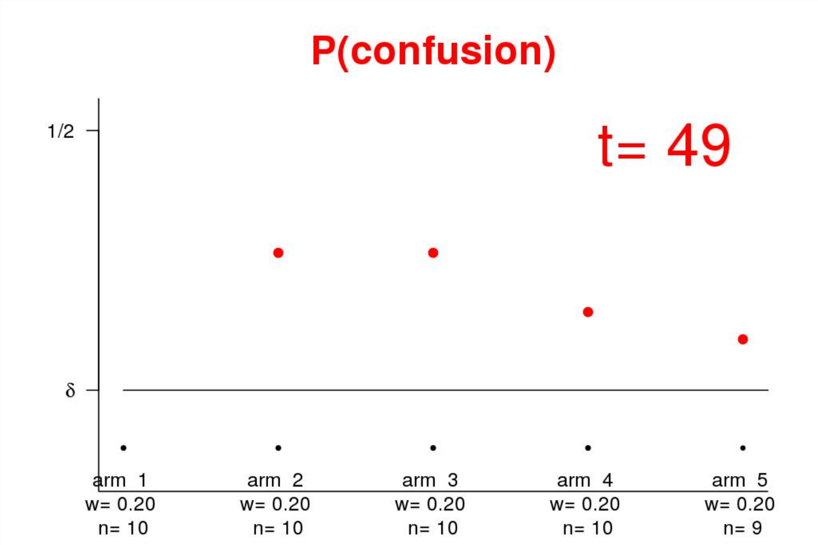
Uniform Sampling



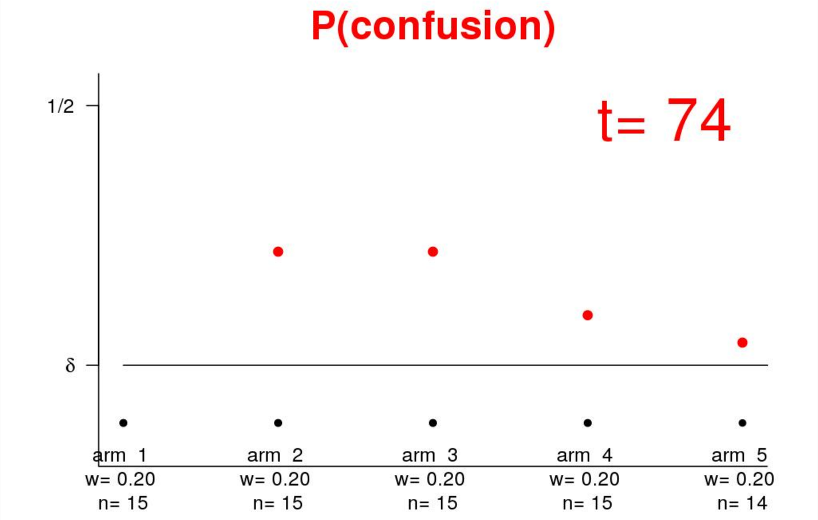
Uniform Sampling



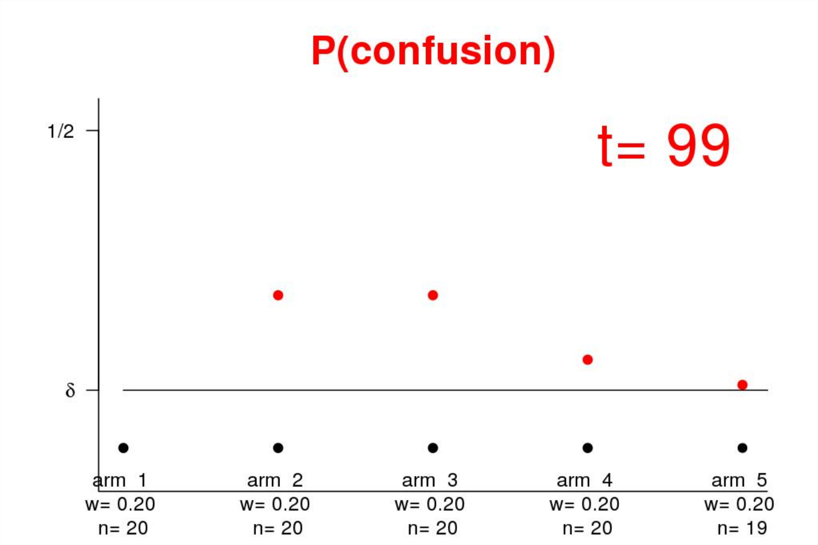
Uniform Sampling



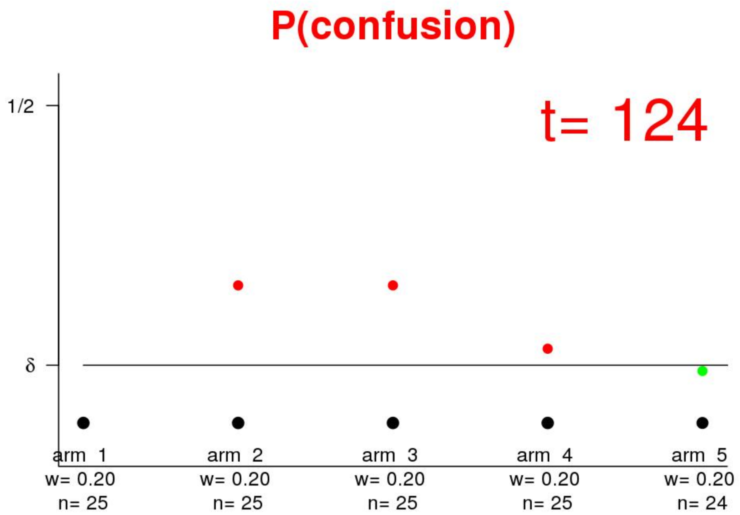
Uniform Sampling



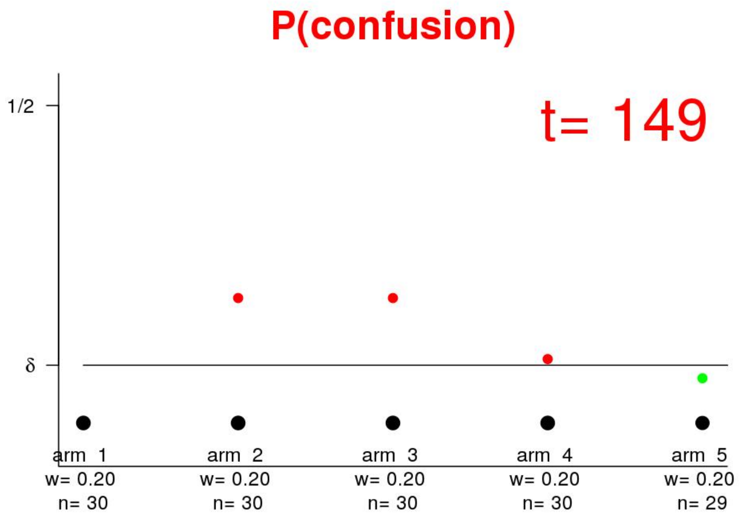
Uniform Sampling



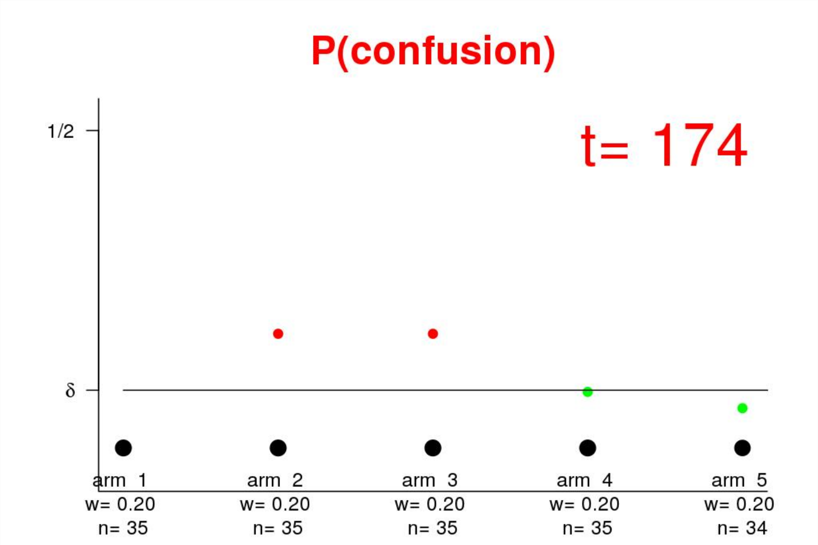
Uniform Sampling



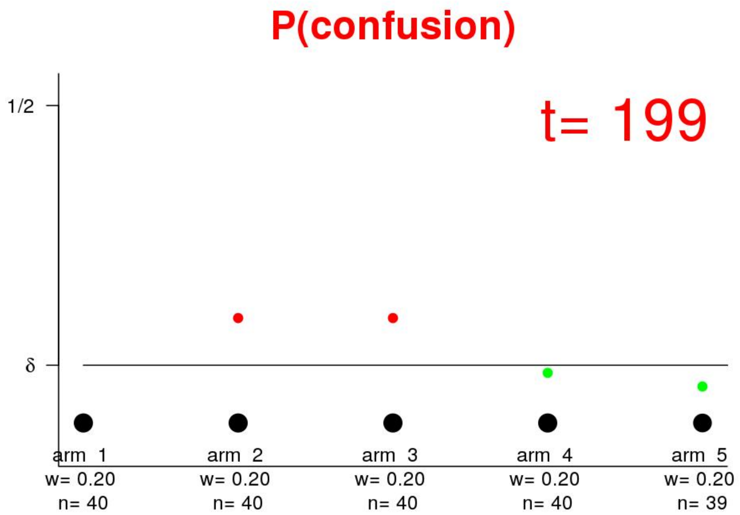
Uniform Sampling



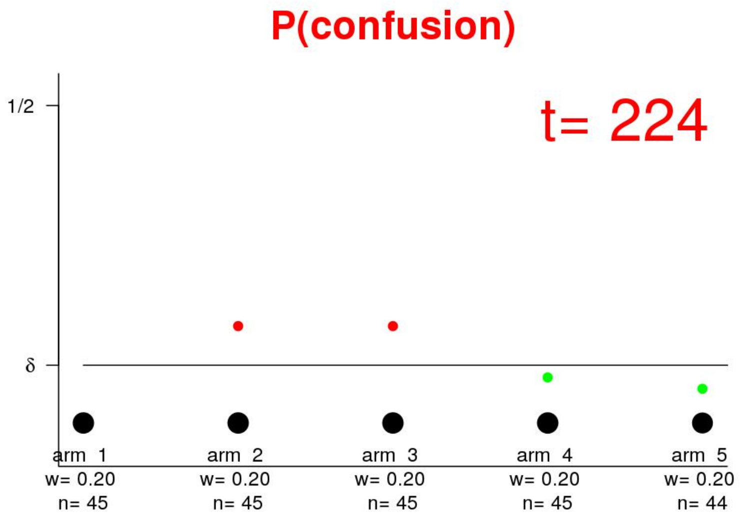
Uniform Sampling



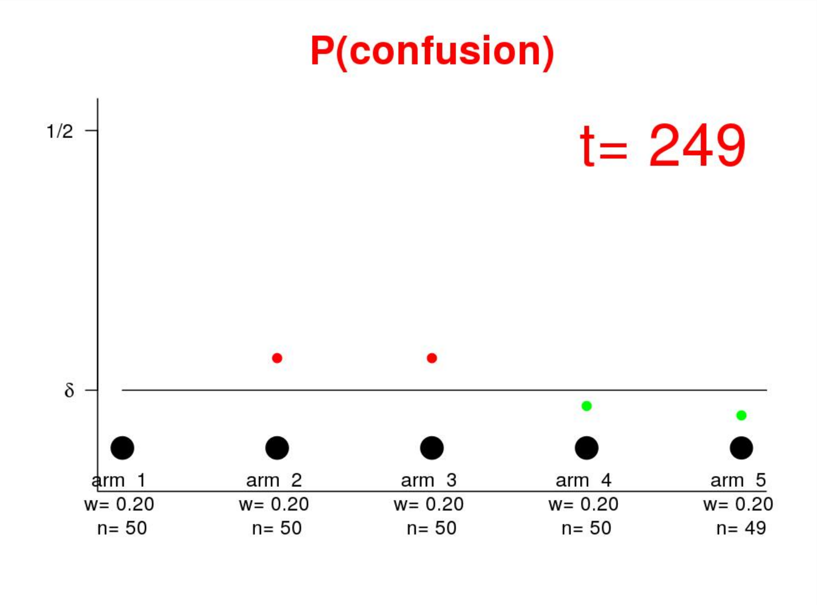
Uniform Sampling



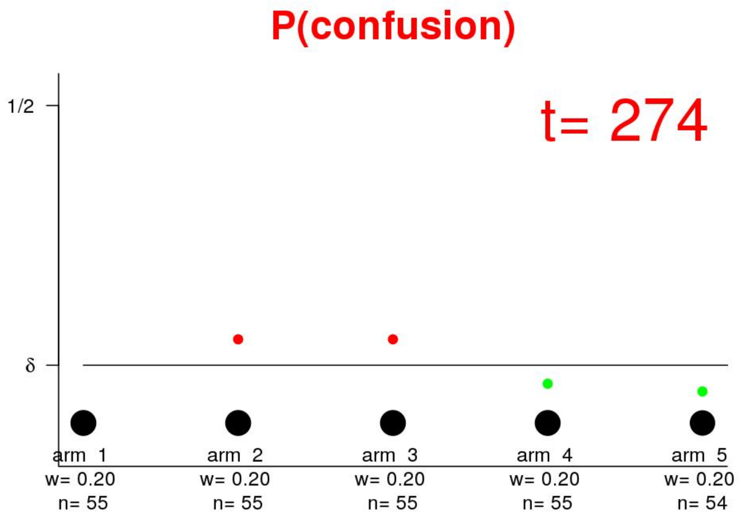
Uniform Sampling



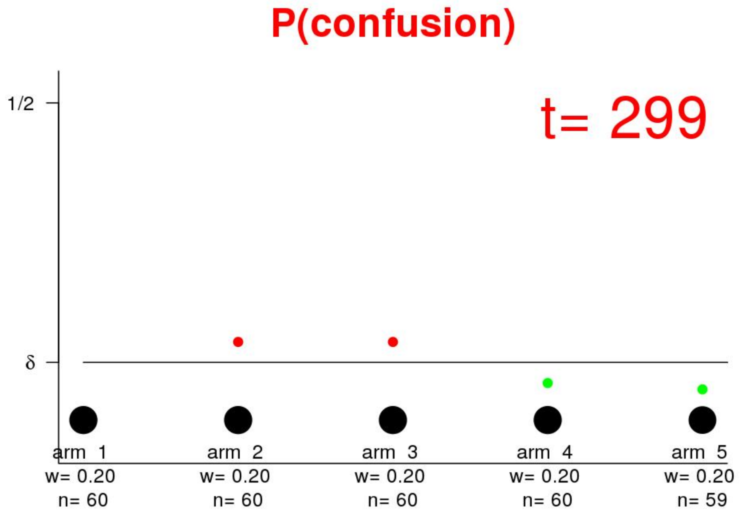
Uniform Sampling



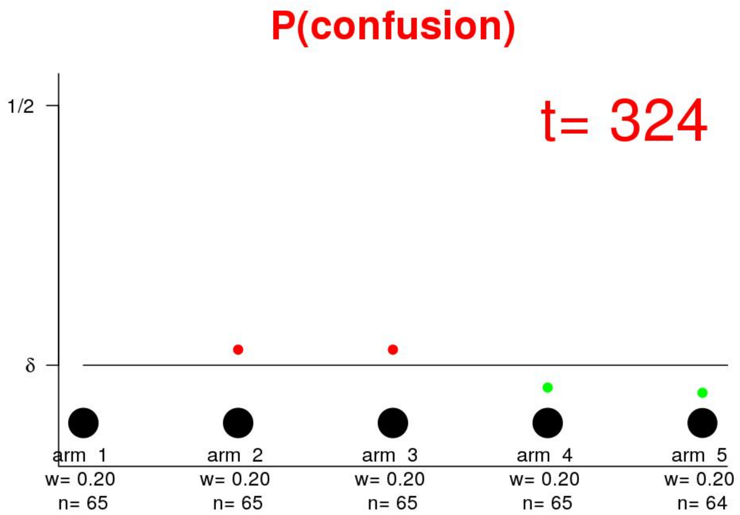
Uniform Sampling



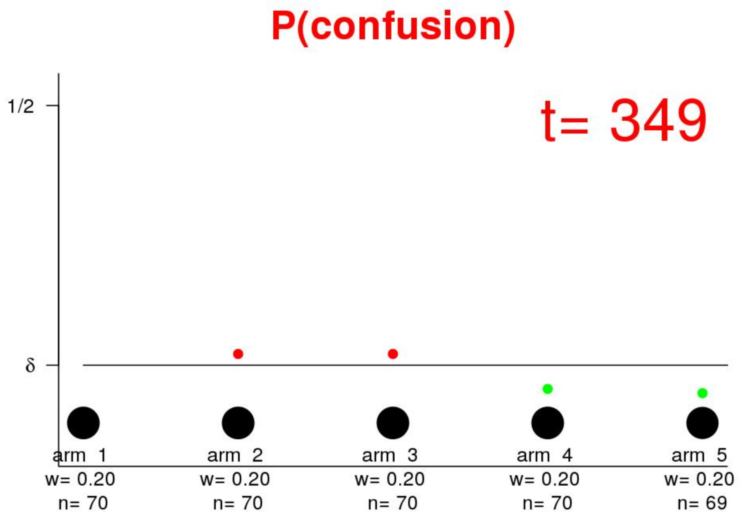
Uniform Sampling

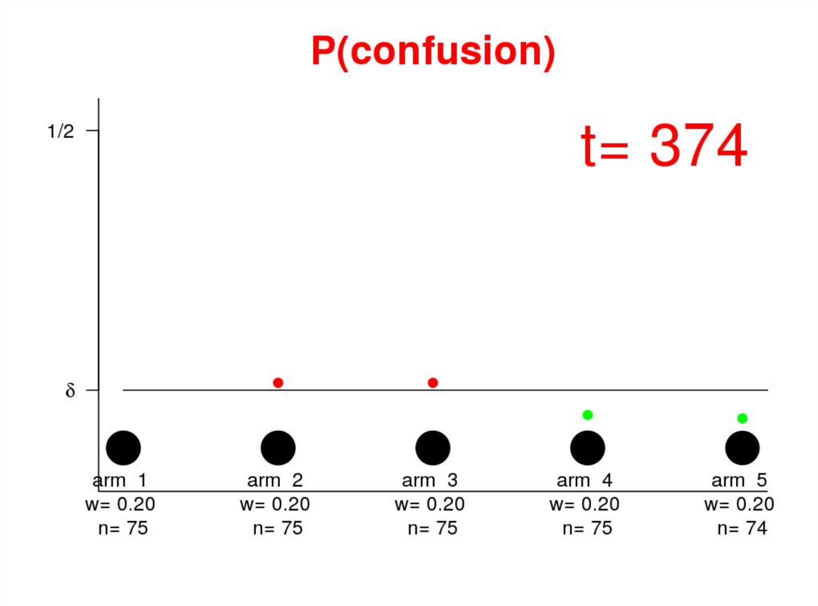


Uniform Sampling

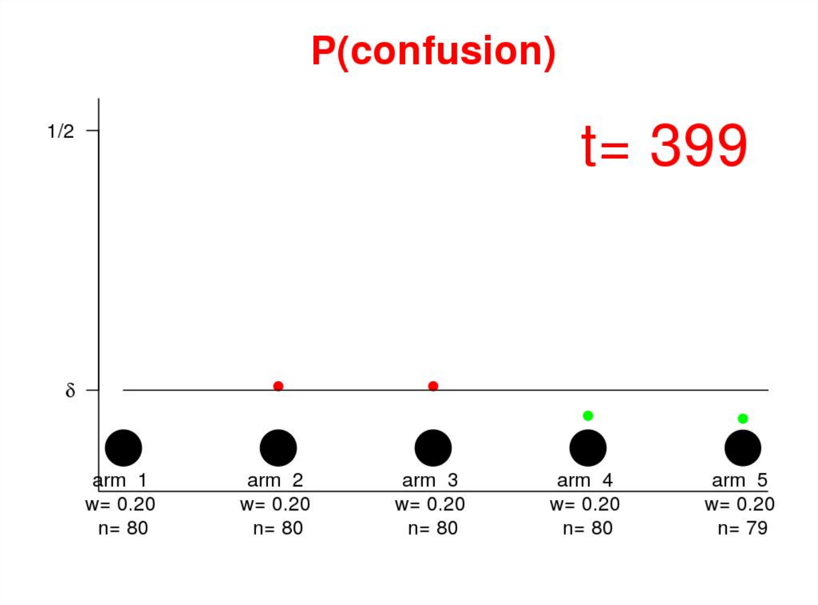


Uniform Sampling

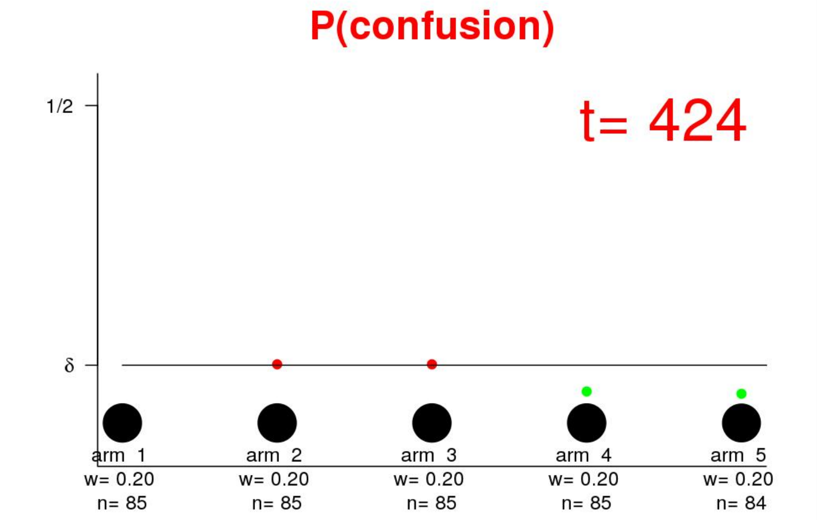



P(confusion)
t= 374
1/2
δ
arm 1
w= 0.20
n= 75
arm 2
w= 0.20
n= 75
arm 3
w= 0.20
n= 75
arm 4
w= 0.20
n= 75
arm 5
w= 0.20
n= 74

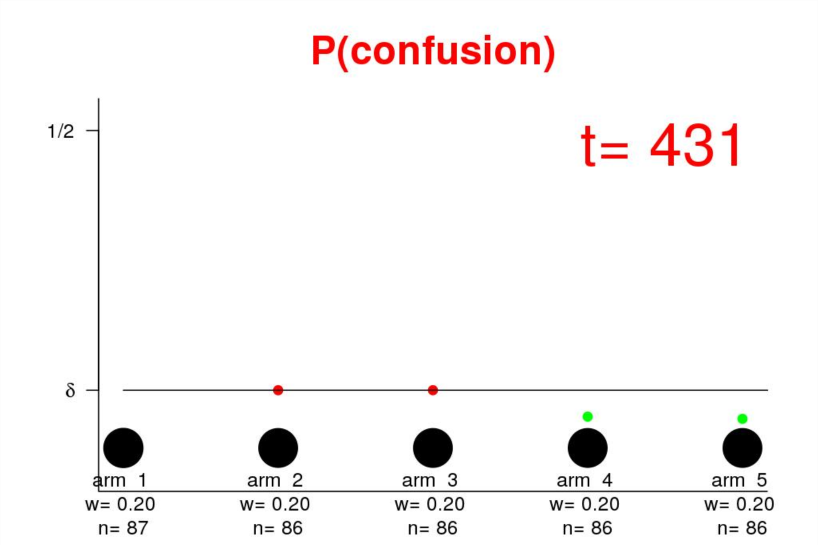
Uniform Sampling



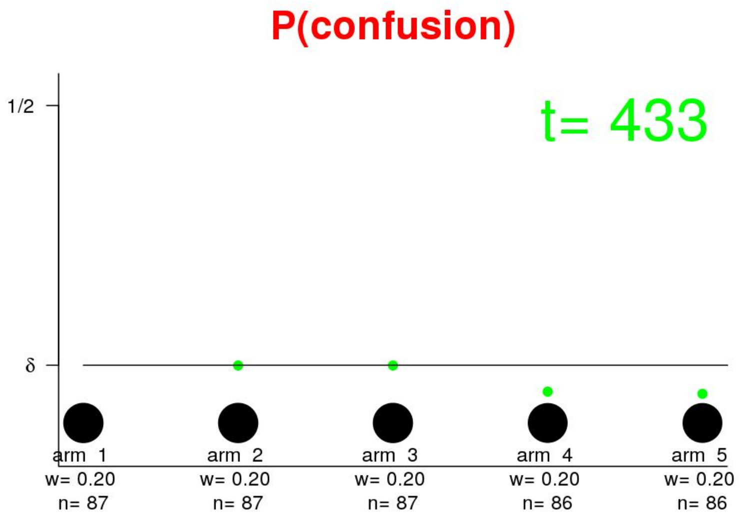
Uniform Sampling



Uniform Sampling



Uniform Sampling

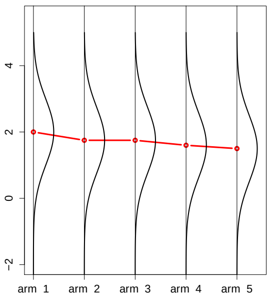


Intuition: Equalizing the Probabilities of Confusion

Most simple setting: for all $a \in \{1, \dots, K\}$,

$$\nu_a = \mathcal{N}(\mu_a, 1)$$

For example: $\mu = [2, 1.75, 1.75, 1.6, 1.5]$.



Active Learning

➜ You allocate a relative budget w_a to option a, with $w_1 + \cdots + w_K = 1$.

At time t:

➜ you have sampled $\mathbf{n_a} \approx \mathbf{w_a t}$ times the option a

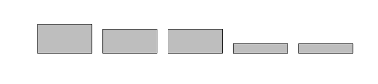
➜ your empirical average is $\bar{X}_{a,n_a}$.

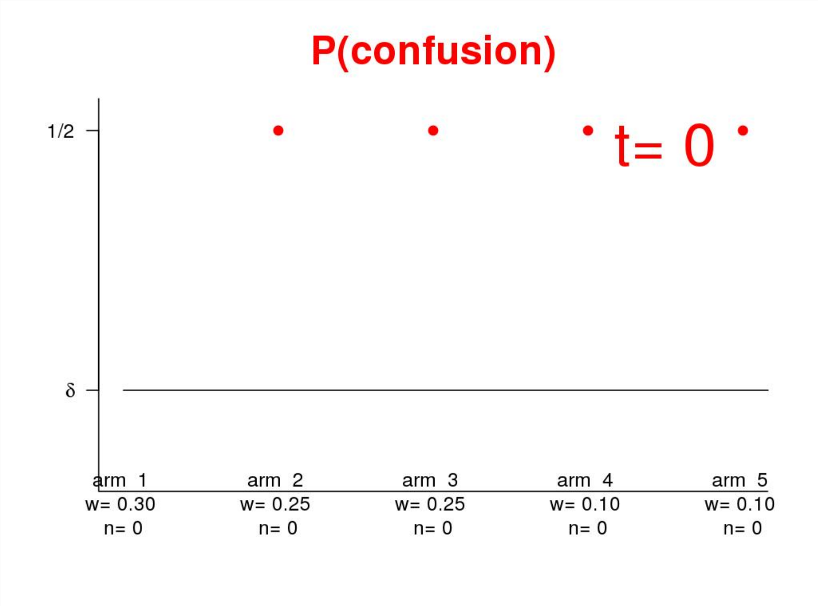
$\longrightarrow$ if you stop at time t, your probability of prefering arm $a \geq 2$ to arm $a^* = 1$ is:

$$\mathbb{P}\left(\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right) = \mathbb{P}\left(\frac{\bar{X}_{a,n_a} - \mu_a - (\bar{X}_{1,n_1} - \mu_1)}{\sqrt{1/n_1 + 1/n_a}} > \frac{\mu_1 - \mu_a}{\sqrt{1/n_1 + 1/n_a}}\right)$$

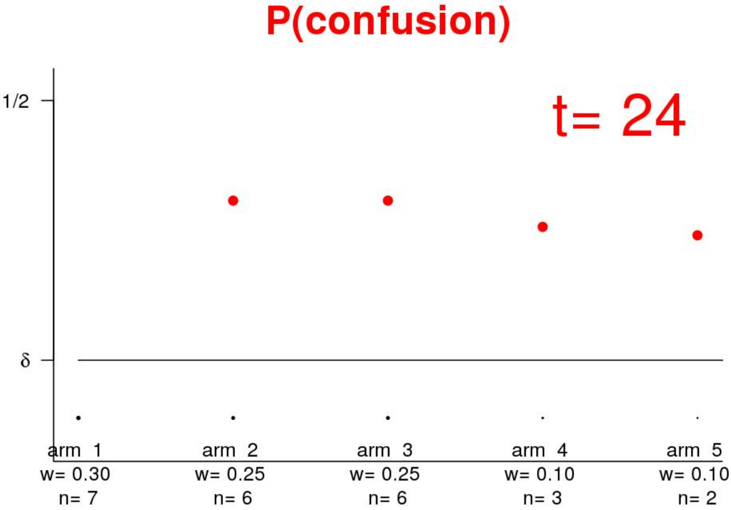
$$= \bar{\Phi}\left(\frac{\mu_1 - \mu_a}{\sqrt{1/n_1 + 1/n_a}}\right)$$

Improving: trial 1

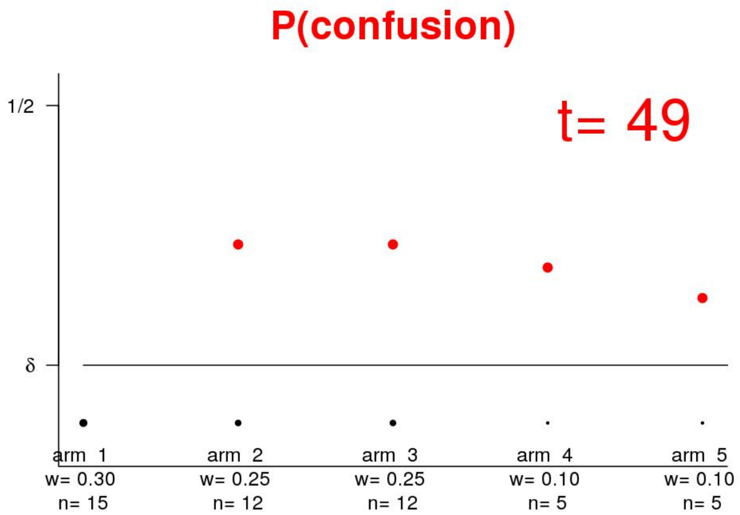




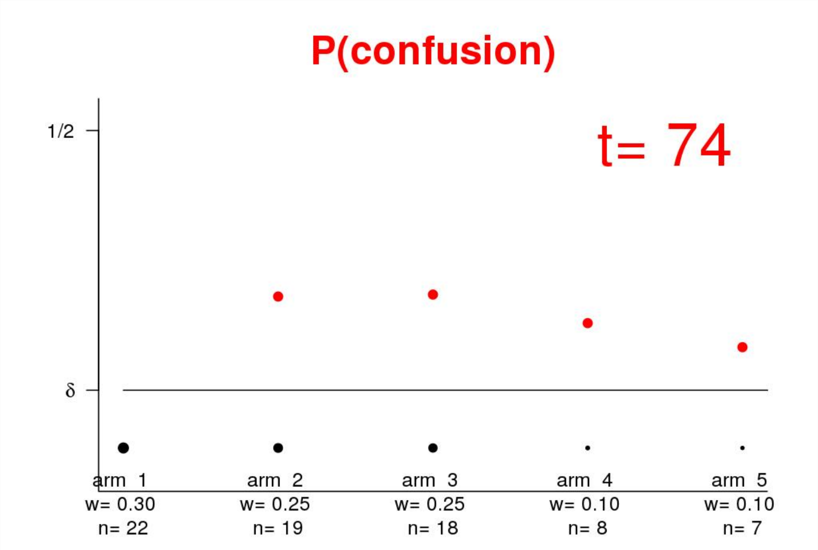
Improving: trial 1



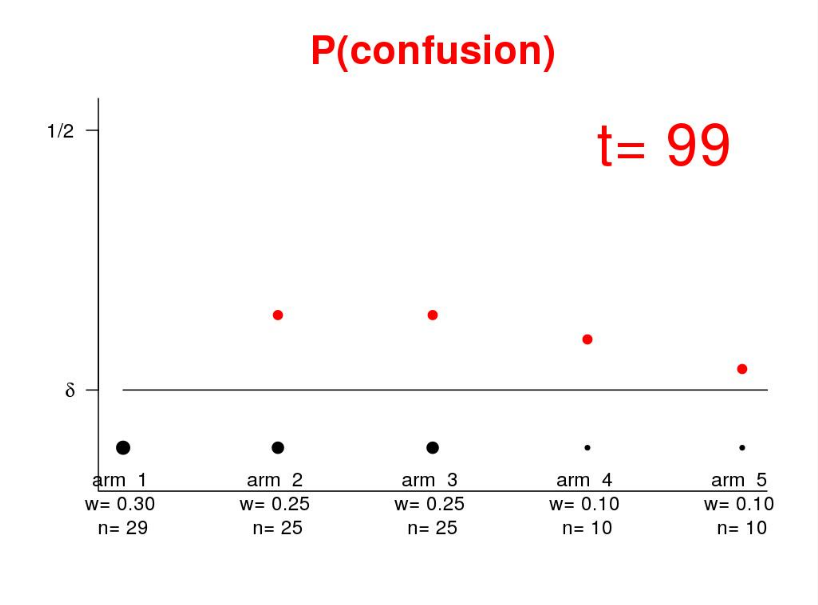
Improving: trial 1



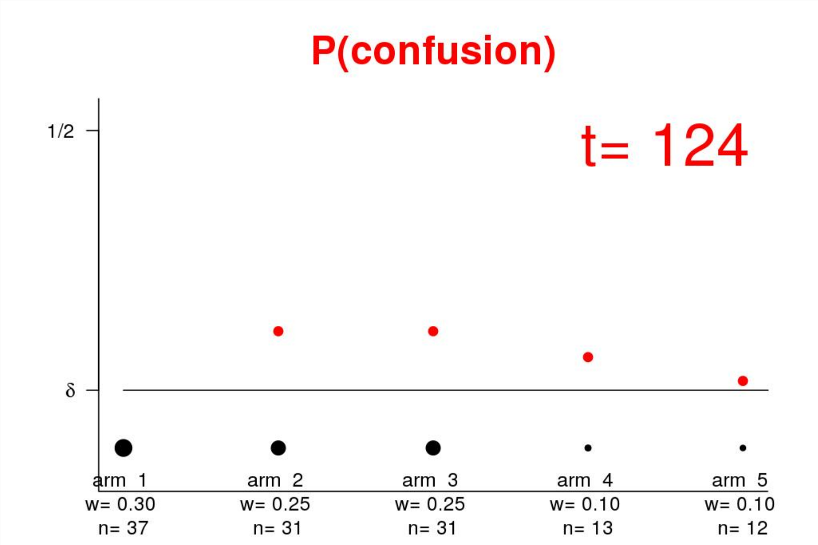
Improving: trial 1



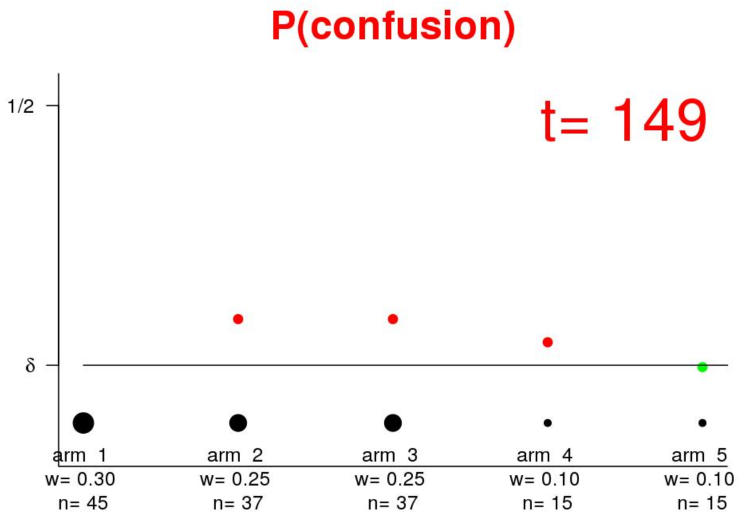
Improving: trial 1



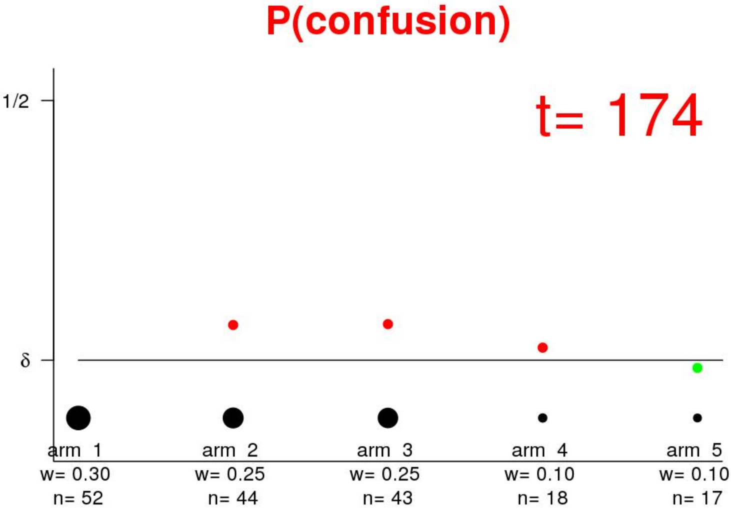
Improving: trial 1



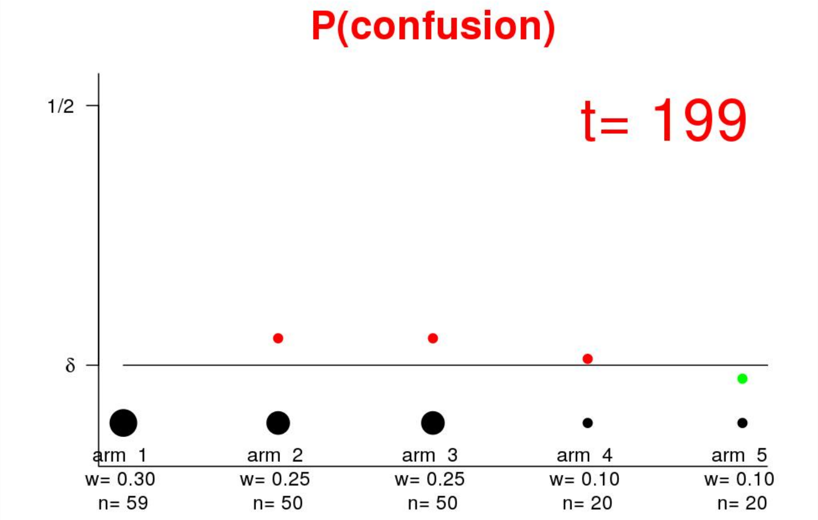
Improving: trial 1



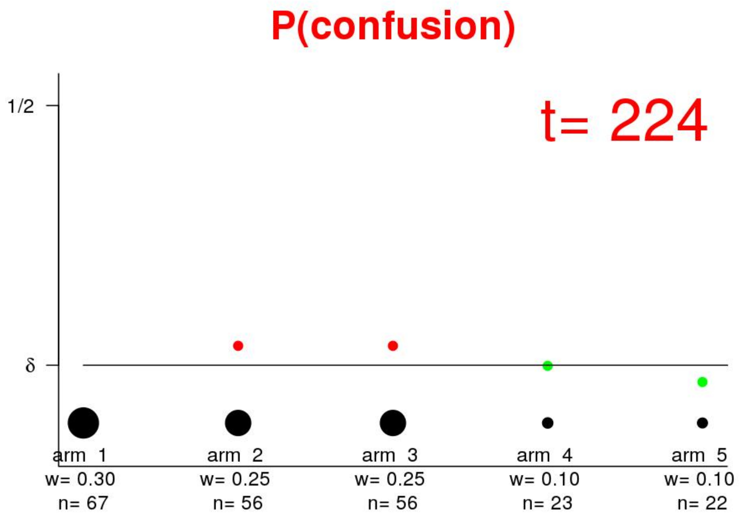
Improving: trial 1



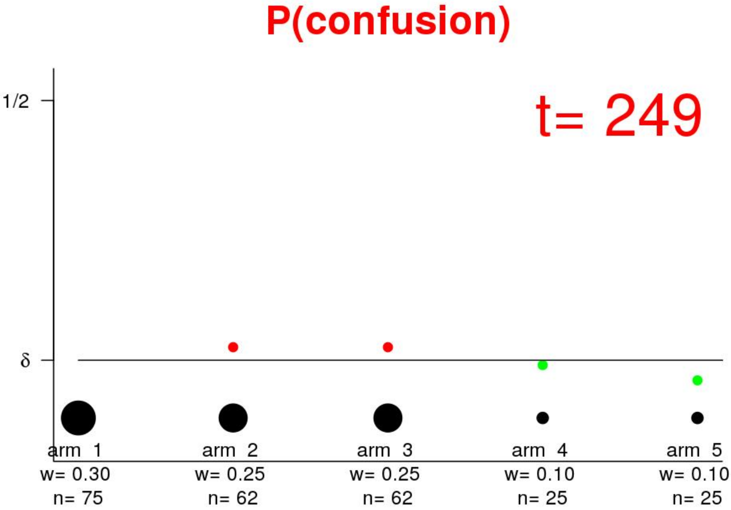
Improving: trial 1



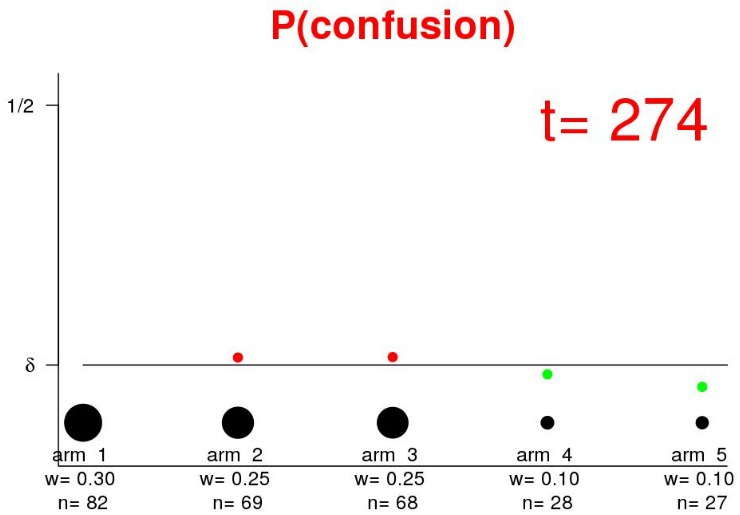
Improving: trial 1



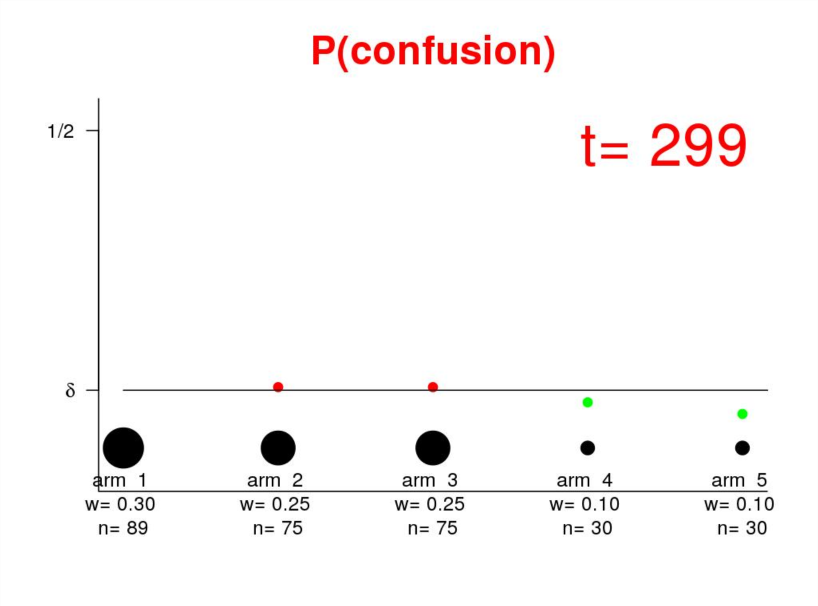
Improving: trial 1



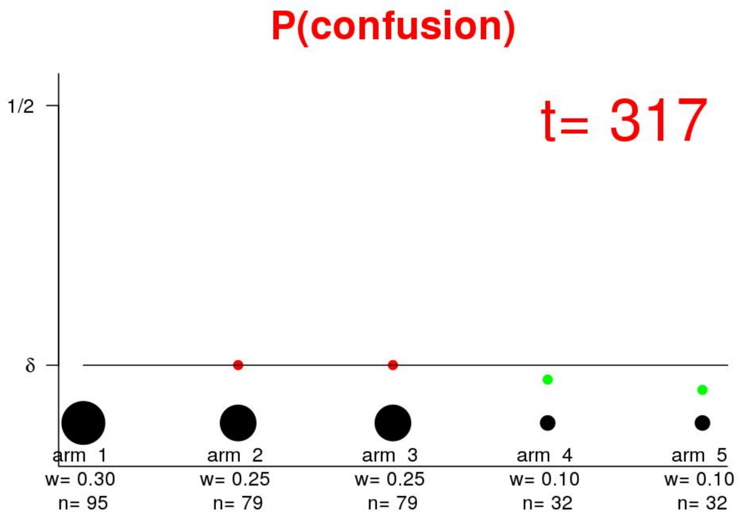
Improving: trial 1



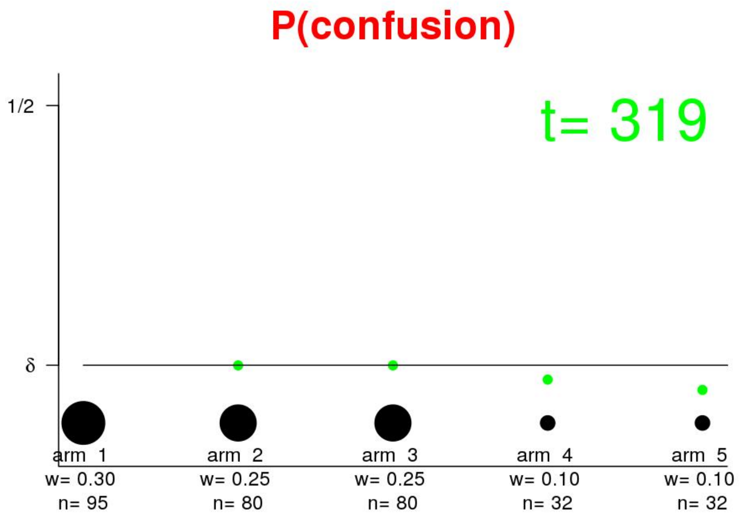
Improving: trial 1



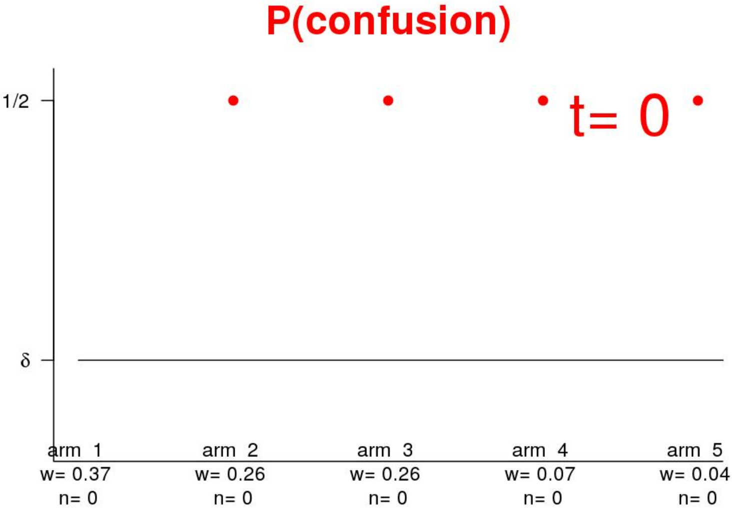
Improving: trial 1



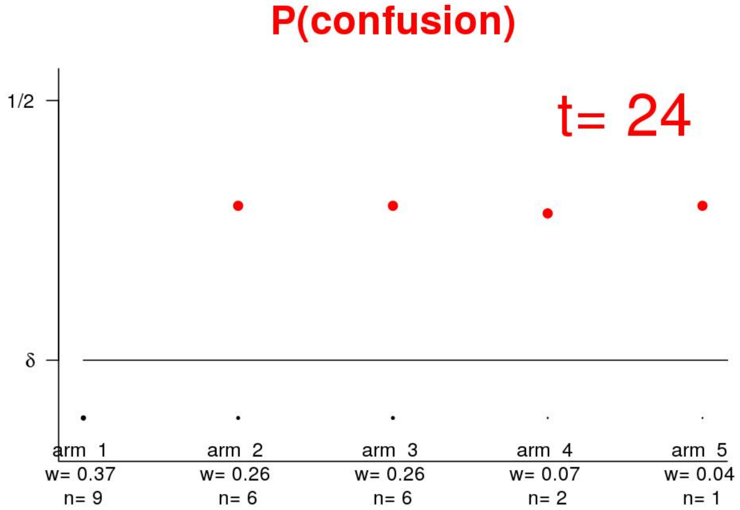
Improving: trial 1



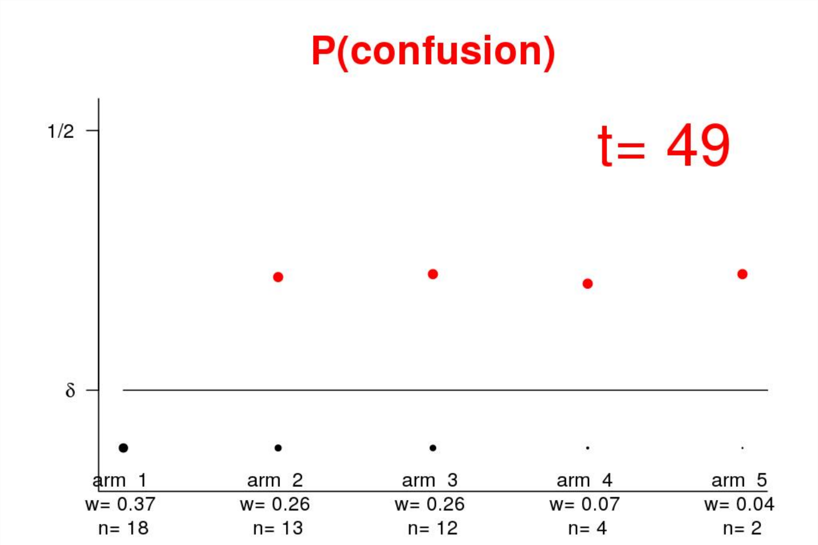
Optimal Proportions



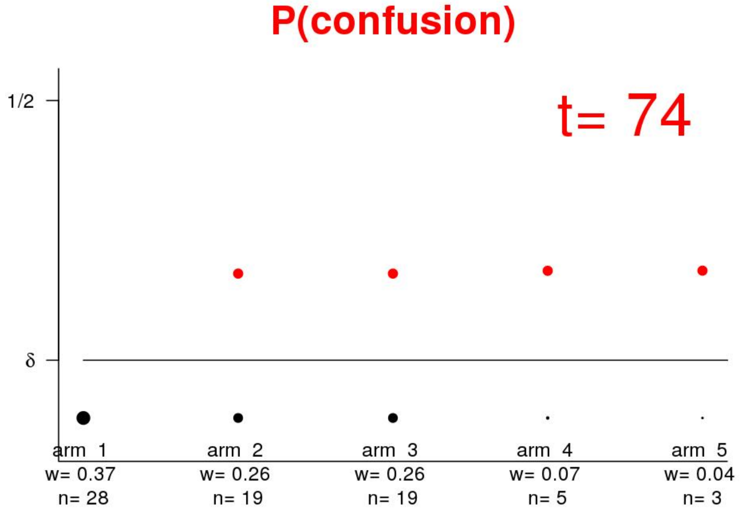
Optimal Proportions



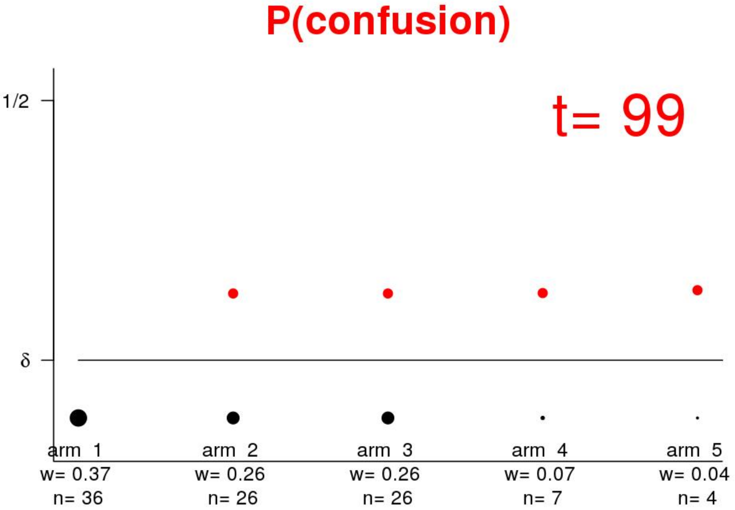
Optimal Proportions



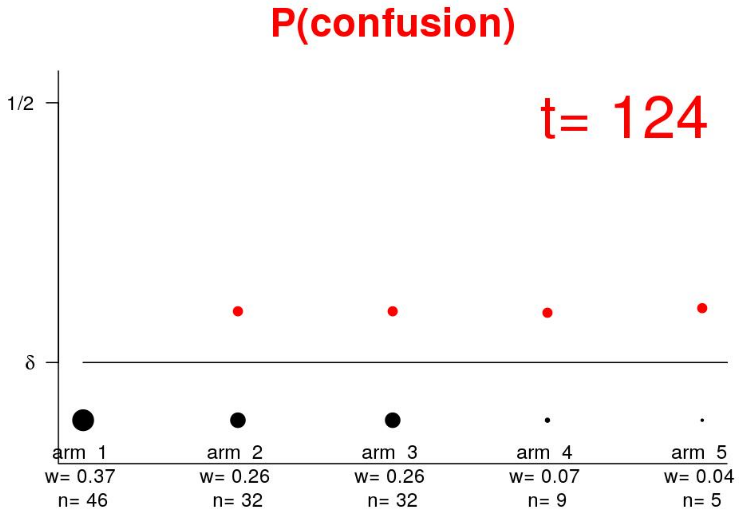
Optimal Proportions



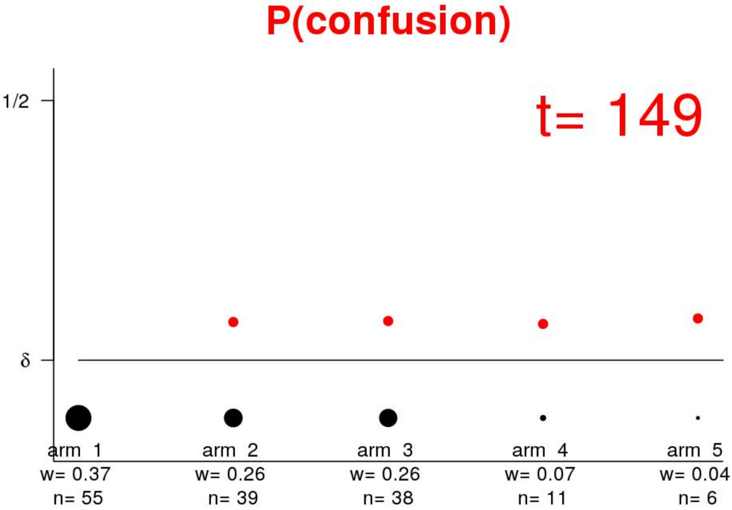
Optimal Proportions



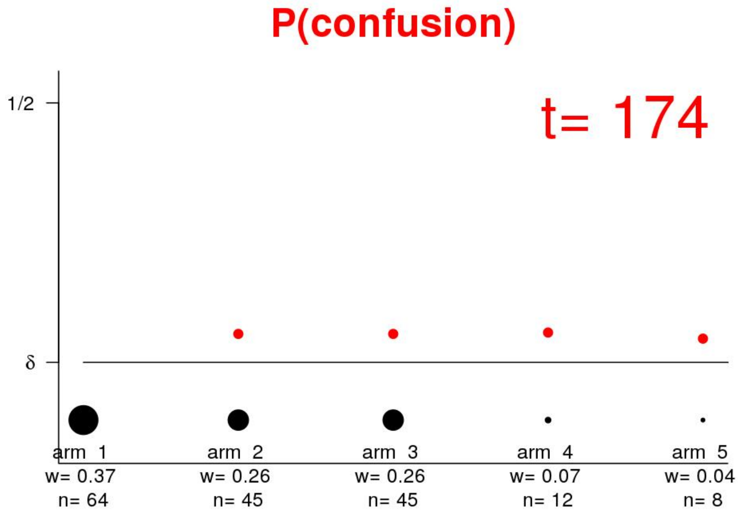
Optimal Proportions



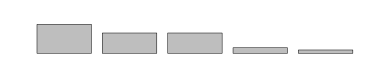
Optimal Proportions

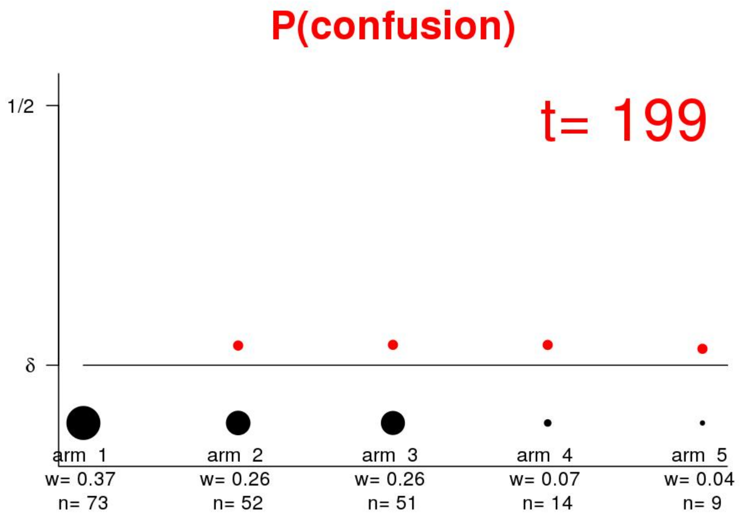


Optimal Proportions

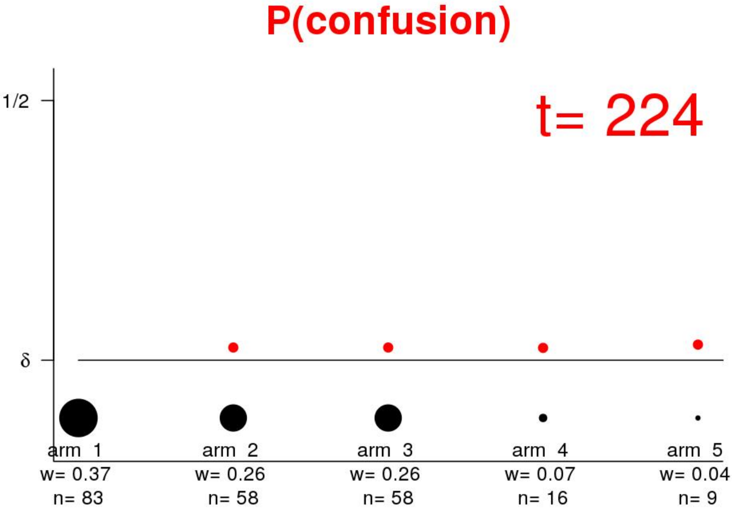


Optimal Proportions

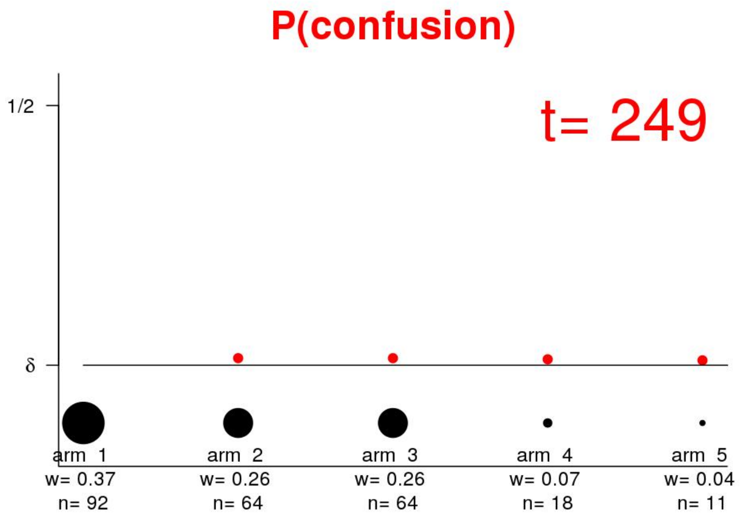




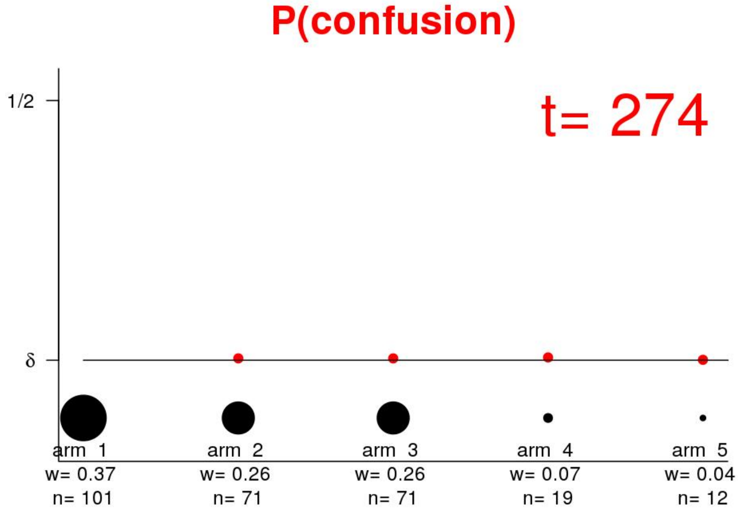
Optimal Proportions



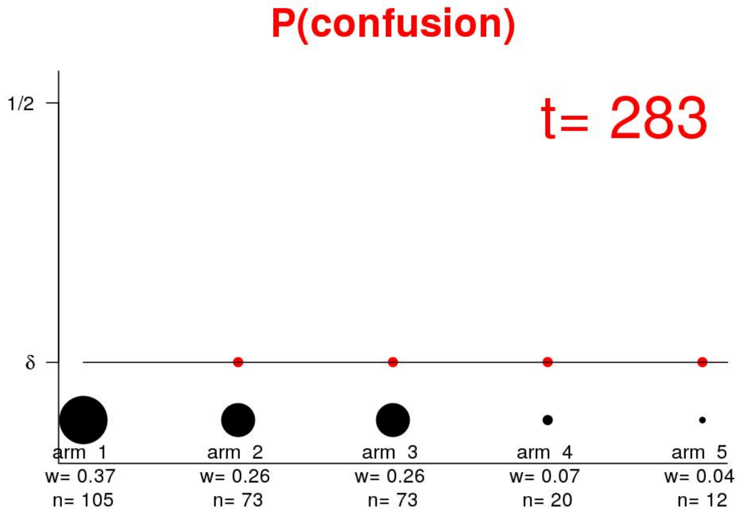
Optimal Proportions



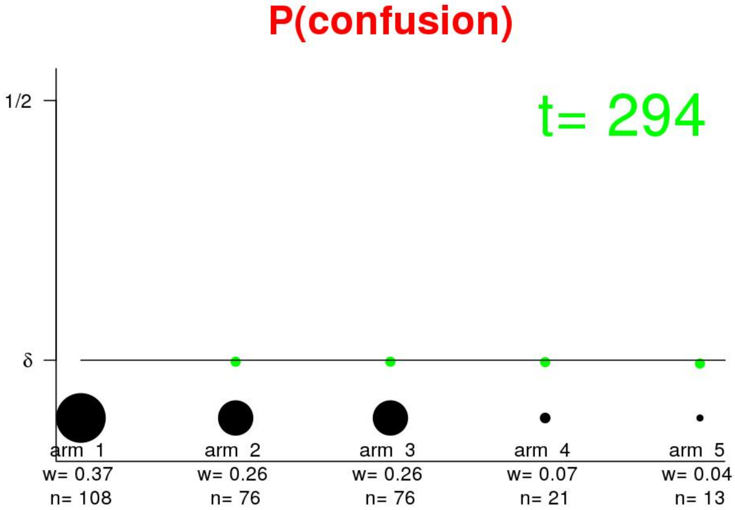
Optimal Proportions



Optimal Proportions



Optimal Proportions



How to Turn this Intuition into a Theorem?

- The arms are not Gaussian (no formula for probability of confusion)
  $\longrightarrow$ large deviations (Sanov, KL)
- You do not allocate a relative budget at first, but you use sequential sampling
  $\longrightarrow$ no fixed-size samples: *sequential experiment*
  $\longrightarrow$ tracking lemma
- How to compute the optimal proportions?
  $\longrightarrow$ lower bound, game
- The parameters of the distribution are unknown
  $\longrightarrow$ (sequential) estimation
- When should you stop?
  $\longrightarrow$ Chernoff's stopping rule

Exponential Families

$\nu_1, \ldots, \nu_K$ belong to a one-dimensional exponential family

$$\mathbb{P}_{\lambda,\Theta,b} = \left\{\nu_\theta, \theta \in \Theta : \nu_\theta \text{ has density } f_\theta(x) = \exp(\theta x - b(\theta)) \; w.r.t.\ \lambda\right\}$$

Example: Gaussian, Bernoulli, Poisson distributions...

- ν_θ can be parametrized by its mean $\mu = \dot{b}(\theta)$: $\nu^\mu := \nu_{\dot{b}^{-1}(\mu)}$

Notation: Kullback-Leibler divergence

For a given exponential family,

$$d(\mu, \mu') := \mathrm{KL}(\nu^\mu, \nu^{\mu'}) = \mathbb{E}_{X\sim\nu^\mu}\left[\log \frac{d\nu^\mu}{d\nu^{\mu'}}(X)\right]$$

is the KL-divergence between the distributions of mean μ and μ'.

We identify $\nu = (\nu^{\mu_1}, \ldots, \nu^{\mu_K})$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ and consider

$$\mathcal{S} = \left\{\boldsymbol{\mu} \in (\dot{b}(\Theta))^K : \exists a \in \{1, \ldots, K\} : \mu_a > \max_{i \neq a} \mu_i\right\}$$

Classical strategies

LUCB: Lower-Upper Confidence Bounds

- Build confidence bounds $L_a(t)$ and $U_a(t)$ such that with probability at least $1-\delta$, for all times $t \geq 1$ and all arms $a \in \{1, \ldots, K\}$:

$$\mu_a \in [L_a(t), U_a(t)] \, ,$$

- Sample alternately

$$\hat{a}(t) = \operatorname*{argmax}_{a \in \{1,\ldots,K\}} L_a(t) \qquad \text{and} \qquad \operatorname*{argmax}_{b \neq \hat{a}(t)} U_b(t)$$

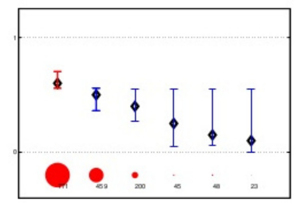
- Stopping time τ_δ = the first time t when

$$\exists \hat{a} \in \{1, \ldots, K\} : \forall a \neq \hat{a}, U_a(t) < L_{\hat{a}}(t)$$

Analysis: δ-correct by nature, and with probability at least $1-\delta$:

$$\tau_\delta \leq C \sum_{a \neq a^*} \frac{1}{(\mu^* - \mu_a)^2}$$

for some constant C.

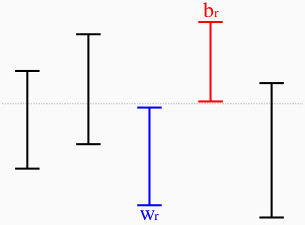


Racing: Successive Eliminations

- Proceed in rounds where, at each round, all active arms are sample once
- Keep a list of active arms = those which have not been eliminated
- At the end of each round, eliminate the arms which are provably suboptimal (with a global risk δ)

Analysis: similarly, one finds a constant C such that

$$\mathbb{E}\left[\tau_\delta\right] \leq C \sum_{a \neq a^*} \frac{1}{(\mu^* - \mu_a)^2} \, .$$



Lower Bound

Information in Bandit Models

Theorem [see Garivier, Ménard and Stoltz, M.O.R. to appear]

For all bandit problems μ and λ, all stopping time τ and $\sigma(\mathcal{F}_\tau)$–measurable random variables Z with values in $[0,1]$,

$$\sum_{a=1}^{K} \mathbb{E}_\mu\big[N_a(\tau)\big]\, d(\mu_a, \lambda_a) \geq \mathrm{kl}\big(\mathbb{E}_\mu[Z], \mathbb{E}_\lambda[Z]\big)\,.$$

Proof: if $I_\tau = (A_1, X_{A_1,1}, \dots, A_\tau, X_{A_\tau, N_{A_\tau}(\tau)})$,

$$\sum_{a=1}^{K} \mathbb{E}_\mu\big[N_a(\tau)\big]\, d(\mu_a, \lambda_a) = \mathrm{KL}\big(\mathbb{P}_\mu^{I_\tau}, \mathbb{P}_\lambda^{I_\tau}\big) \geq \mathrm{KL}\big(\mathbb{P}_\mu^{Z}, \mathbb{P}_\lambda^{Z}\big) \geq \mathrm{kl}\big(\mathbb{E}_\mu[Z], \mathbb{E}_\lambda[Z]\big)$$

by *tensorization* and *contraction* of entropy (and small lemma).

Lower-Bounding the Sample Complexity

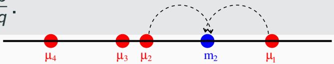
Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ be two elements of $\mathcal{S}$.

Uniform δ-correct Constraint [Kaufmann, Cappé, Garivier '15]

If $a^*(\boldsymbol{\mu}) \neq a^*(\boldsymbol{\lambda})$, any δ-correct algorithm satisfies

$$\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}\left[N_a(\tau_\delta)\right] d(\mu_a, \lambda_a) \geq \mathrm{kl}(\delta, 1-\delta)$$

where $\mathrm{kl}(p,q) = p \log \frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$.



Let $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}$. Take: $\lambda_1 = m_2 - \epsilon \quad \lambda_2 = m_2 + \epsilon$

$$\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]\, d(\mu_1, m_2 - \epsilon) + \mathbb{E}_{\boldsymbol{\mu}}[N_2(\tau_\delta)]\, d(\mu_2, m_2 + \epsilon) \quad \geq \quad \mathrm{kl}(\delta, 1-\delta)$$

Lower-Bounding the Sample Complexity

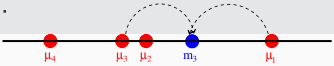
Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$ be two elements of $\mathcal{S}$.

Uniform δ-correct Constraint [Kaufmann, Cappé, Garivier '15]

If $a^*(\boldsymbol{\mu}) \neq a^*(\boldsymbol{\lambda})$, any δ-correct algorithm satisfies

$$\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}\left[N_a(\tau_\delta)\right] d(\mu_a, \lambda_a) \geq \mathrm{kl}(\delta, 1-\delta)$$

where $\mathrm{kl}(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$.



Let $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}$. Take: $\quad \lambda_1 = m_3 - \epsilon \quad \lambda_3 = m_3 + \epsilon$

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]\, d(\mu_1, m_2 - \epsilon) + \mathbb{E}_{\boldsymbol{\mu}}[N_2(\tau_\delta)]\, d(\mu_2, m_2 + \epsilon) &\geq \mathrm{kl}(\delta, 1-\delta) \\
\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]\, d(\mu_1, m_3 - \epsilon) + \mathbb{E}_{\boldsymbol{\mu}}[N_3(\tau_\delta)]\, d(\mu_3, m_3 + \epsilon) &\geq \mathrm{kl}(\delta, 1-\delta)
\end{aligned}$$

Lower-Bounding the Sample Complexity

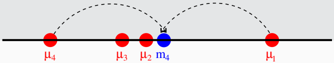
Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ be two elements of $\mathcal{S}$.

Uniform δ-correct Constraint [Kaufmann, Cappé, Garivier '15]

If $a^*(\boldsymbol{\mu}) \neq a^*(\boldsymbol{\lambda})$, any δ-correct algorithm satisfies

$$\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}\left[N_a(\tau_\delta)\right] d(\mu_a, \lambda_a) \geq \mathrm{kl}(\delta, 1-\delta)$$

where $\mathrm{kl}(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$.



Let $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}$. Take: $\lambda_1 = m_4 - \epsilon \quad \lambda_4 = m_4 + \epsilon$

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]\, d(\mu_1, m_2 - \epsilon) + \mathbb{E}_{\boldsymbol{\mu}}[N_2(\tau_\delta)]\, d(\mu_2, m_2 + \epsilon) &\geq \mathrm{kl}(\delta, 1-\delta) \\
\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]\, d(\mu_1, m_3 - \epsilon) + \mathbb{E}_{\boldsymbol{\mu}}[N_3(\tau_\delta)]\, d(\mu_3, m_3 + \epsilon) &\geq \mathrm{kl}(\delta, 1-\delta) \\
\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]\, d(\mu_1, m_4 - \epsilon) + \mathbb{E}_{\boldsymbol{\mu}}[N_4(\tau_\delta)]\, d(\mu_4, m_4 + \epsilon) &\geq \mathrm{kl}(\delta, 1-\delta)
\end{aligned}$$

Lower-Bounding the Sample Complexity

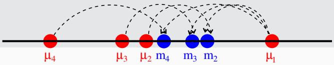
Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ be two elements of $\mathcal{S}$.

Uniform δ-correct Constraint [Kaufmann, Cappé, Garivier '15]

If $a^*(\boldsymbol{\mu}) \neq a^*(\boldsymbol{\lambda})$, any δ-correct algorithm satisfies

$$\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}\left[N_a(\tau_\delta)\right] d(\mu_a, \lambda_a) \geq \mathrm{kl}(\delta, 1-\delta)$$

where $\mathrm{kl}(p, q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$.



Let $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}$.

$$\begin{aligned}
\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)]\, d(\mu_a, \lambda_a) &\geq \mathrm{kl}(\delta, 1-\delta) \\
\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \times \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}\, d(\mu_a, \lambda_a) &\geq \mathrm{kl}(\delta, 1-\delta) \\
\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \times \left(\sup_{w \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a\, d(\mu_a, \lambda_a) \right) &\geq \mathrm{kl}(\delta, 1-\delta)
\end{aligned}$$

Lower Bound: the Complexity of BAI

Theorem [Garivier and Kaufmann 2016]

For any δ-correct algorithm,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^*(\boldsymbol{\mu}) \, \mathrm{kl}(\delta, 1-\delta) \, ,$$

where

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{\mathbf{w} \in \Sigma_K} \inf_{\lambda \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} w_a \, d(\mu_a, \lambda_a) \right) .$$

- $\mathrm{kl}(\delta, 1-\delta) \sim \log(1/\delta)$ when $\delta \to 0$, $\mathrm{kl}(\delta, 1-\delta) \geq \log\left(1/(2.4\delta)\right)$
- cf. [Graves and Lai 1997, Vaidhyan and Sundaresan, 2015]

➜ the optimal proportions of arm draws are

$$\mathbf{w}^*(\boldsymbol{\mu}) = \underset{\mathbf{w} \in \Sigma_K}{\mathrm{argmax}} \inf_{\lambda \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \right)$$

➜ they do not depend on δ

PAC-BAI as a Game

Given a parameter $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$:

- the statistician chooses proportions of arm draws $\boldsymbol{w} = (w_a)_a$
- the opponent chooses an alternative model $\boldsymbol{\lambda}$
- the payoff is the minimal number $T = T(\boldsymbol{w}, \boldsymbol{\lambda})$ of draws necessary to ensure that he does not violate the δ-PAC constraint

$$\sum_{a=1}^{K} T w_a \, d(\mu_a, \lambda_a) \geq \mathrm{kl}(\delta, 1-\delta)$$

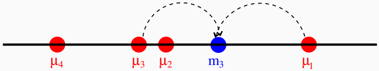
- $T^*(\boldsymbol{\mu}) \, \mathrm{kl}(\delta, 1-\delta) =$ value of the game
  $\boldsymbol{w}^* =$ optimal action for the statistician

PAC-BAI as a Game

Given a parameter $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ such that $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$:

- the statistician chooses proportions of arm draws $\boldsymbol{w} = (w_a)_a$
- the opponent chooses an arm $a \in \{2, \ldots, K\}$ and

  $\lambda_a = \arg\min_\lambda w_1\, d(\mu_1, \lambda) + w_a\, d(\mu_a, \lambda)$



- the payoff is the minimal number $T = T(\boldsymbol{w}, a, \delta)$ of draws necessary to ensure that

$$Tw_1\, d(\mu_1, \lambda_a - \epsilon) + Tw_a\, d(\mu_a, \lambda_a + \epsilon) \geq \mathrm{kl}(\delta, 1-\delta)$$

  that is $T(\boldsymbol{w}, a, \delta) = \dfrac{\mathrm{kl}(\delta, 1-\delta)}{w_1\, d(\mu_1, \lambda_a - \epsilon) + w_a\, d(\mu_a, \lambda_a + \epsilon)}$

- $T^*(\boldsymbol{\mu})\, \mathrm{kl}(\delta, 1-\delta)$ = value of the game

  $\boldsymbol{w}^*$ = optimal action for the statistician

Properties of $T^*(\mu)$ and $\mathbf{w}^*(\mu)$

1. Unique solution, solution of scalar equations only
2. For all $\boldsymbol{\mu} \in \mathcal{S}$, for all a, $w_a^*(\boldsymbol{\mu}) > 0$
3. $\mathbf{w}^*$ is continuous in every $\boldsymbol{\mu} \in \mathcal{S}$
4. If $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$, one has $w_2^*(\boldsymbol{\mu}) \geq \cdots \geq w_K^*(\boldsymbol{\mu})$

   (one may have $w_1^*(\boldsymbol{\mu}) < w_2^*(\boldsymbol{\mu})$)
5. Case of two arms [Kaufmann, Cappé, Garivier '14]

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq \frac{\mathrm{kl}(\delta, 1-\delta)}{d_*(\mu_1, \mu_2)} \, .$$

   where d_* is the 'reversed' Chernoff information

$$d_*(\mu_1, \mu_2) := d(\mu_1, \mu_*) = d(\mu_2, \mu_*) \, .$$

6. Gaussian arms : algebraic equation but no simple formula for $K \geq 3$.

$$\sum_{a=1}^{K} \frac{2\sigma^2}{\Delta_a^2} \leq T^*(\boldsymbol{\mu}) \leq 2 \sum_{a=1}^{K} \frac{2\sigma^2}{\Delta_a^2} \, .$$

The Track-and-Stop Strategy

Outline

Sampling rule: Tracking the optimal proportions

$\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t), \dots, \hat{\mu}_K(t))$: vector of empirical means

Introducing

$$U_t = \left\{ a : N_a(t) < \sqrt{t} \right\},$$

the arm sampled at round $t+1$ is

$$A_{t+1} \in \begin{cases} \underset{a \in U_t}{\operatorname{argmin}}\ N_a(t) \quad \text{if } U_t \neq \emptyset & (\textit{forced exploration}) \\ \underset{1 \leq a \leq K}{\operatorname{argmax}}\ t\, w_a^*(\hat{\boldsymbol{\mu}}(t)) - N_a(t) & (\textit{tracking}) \end{cases}$$

Lemma

Under the Tracking sampling rule,

$$\mathbb{P}_{\boldsymbol{\mu}} \left(\lim_{t \to \infty} \frac{N_a(t)}{t} = w_a^*(\boldsymbol{\mu}) \right) = 1.$$

Outline

Sequential Generalized Likelihood Test

High values of the Generalized Likelihood Ratio

$$Z_{a,b}(t) := \log \frac{\max_{\{\boldsymbol{\lambda}:\lambda_a \geq \lambda_b\}} dP_{\boldsymbol{\lambda}}(X_1, \dots, X_t)}{\max_{\{\boldsymbol{\lambda}:\lambda_a \leq \lambda_b\}} dP_{\boldsymbol{\lambda}}(X_1, \dots, X_t)}$$

$$= N_a(t)\, d\big(\hat{\mu}_a(t), \hat{\mu}_{a,b}(t)\big) + N_b(t)\, d\big(\hat{\mu}_b(t), \hat{\mu}_{a,b}(t)\big) \quad \text{if } \hat{\mu}_a(t) > \hat{\mu}_b(t)$$

$$-Z_{b,a}(t) \text{ otherwise}$$

reject the hypothesis that $(\mu_a \leq \mu_b)$.

We stop when one arm is assessed to be significantly larger than all other arms, according to a GLR test:

$$\tau_\delta = \inf\big\{t \in \mathbb{N} : \exists a \in \{1, \dots, K\}, \forall b \neq a, Z_{a,b}(t) > \beta(t, \delta)\big\}$$

$$= \inf\left\{t \in \mathbb{N} : \quad Z(t) := \max_{a \in \{1,\dots,K\}} \min_{b \neq a} Z_{a,b}(t) > \beta(t, \delta)\right\}$$

Chernoff stopping rule [Chernoff '59]

Two other possible interpretations of the stopping rule:

➜ MDL:

$$Z_{a,b}(t) = \big(N_a(t) + N_b(t)\big) H\big(\hat{\mu}_{a,b}(t)\big) - \Big[N_a(t) H\big(\hat{\mu}_a(t)\big) + N_b(t) H\big(\hat{\mu}_b(t)\big)\Big]$$

Sequential Generalized Likelihood Test

High values of the Generalized Likelihood Ratio

$$Z_{a,b}(t) := \log \frac{\max_{\{\boldsymbol{\lambda}:\lambda_a \geq \lambda_b\}} dP_{\boldsymbol{\lambda}}(X_1, \ldots, X_t)}{\max_{\{\boldsymbol{\lambda}:\lambda_a \leq \lambda_b\}} dP_{\boldsymbol{\lambda}}(X_1, \ldots, X_t)}$$

reject the hypothesis that $(\mu_a \leq \mu_b)$.

We stop when one arm is assessed to be significantly larger than all other arms, according to a GLR test:

$$\tau_\delta = \inf \left\{ t \in \mathbb{N} : \quad Z(t) := \max_{a \in \{1,\ldots,K\}} \min_{b \neq a} Z_{a,b}(t) > \beta(t, \delta) \right\}$$

Chernoff stopping rule [Chernoff '59]

Two other possible interpretations of the stopping rule:

➜ plug-in complexity estimate: if $F(w, \mu) := \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\mu)} \sum_{a=1}^{K} w_a \, d\big(\mu_a, \lambda_a\big)$,

stop when $Z(t) = t\, F\Big(\frac{N_a(t)}{t}, \hat{\mu}(t)\Big) \geq \beta(t, \delta)$ instead of the lower bound $\frac{t}{T^*(\mu)} = t\, F(\mathbf{w}^*, \mu) \geq \mathrm{kl}(\delta, 1-\delta)$.

Calibration

Theorem

The Chernoff rule is δ-PAC for $\beta(t, \delta) = \log\left(\frac{2(K-1)t}{\delta}\right)$

Lemma

If $\mu_a < \mu_b$, whatever the sampling rule,

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : Z_{a,b}(t) > \log\left(\frac{2t}{\delta}\right)\right) \leq \delta$$

The proof uses:

➜ Barron's lemma (change of distribution)

➜ and Krichevsky-Trofimov's universal distribution

(very information-theoretic ideas)

Outline

Asymptotic Optimality of the T&S strategy

Theorem [Garivier and Kaufmann 2016]

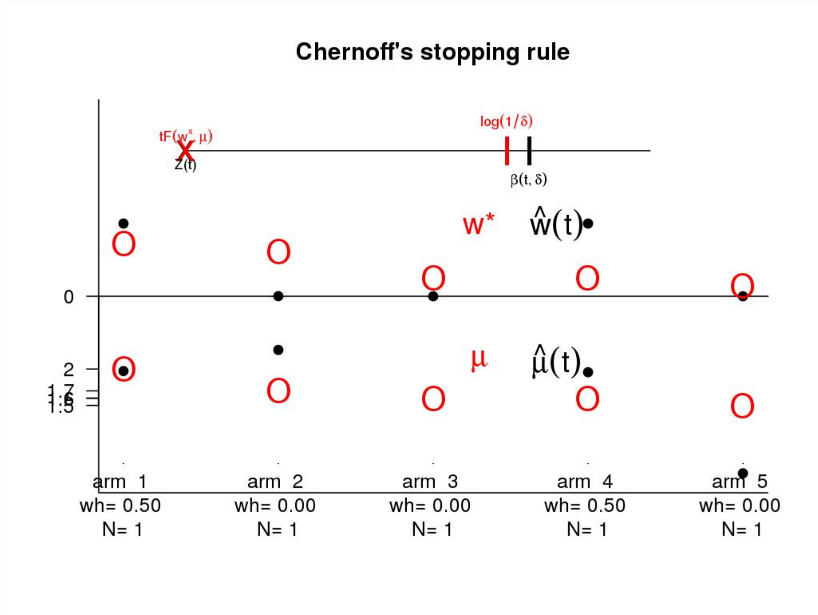
The Track-and-Stop strategy, that uses

- the Tracking sampling rule
- the Chernoff stopping rule with $\beta(t, \delta) = \log\left(\frac{2(K-1)t}{\delta}\right)$
- and recommends $\hat{a}_{\tau_\delta} = \underset{a=1\dots K}{\text{argmax}}\ \hat{\mu}_a(\tau_\delta)$

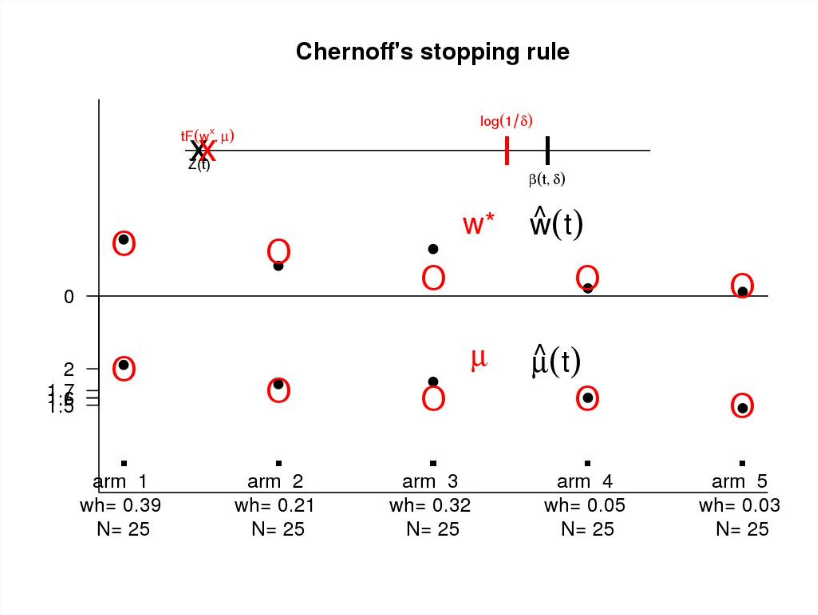
is δ-PAC for every $\delta \in (0, 1)$ and satisfies

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} = T^*(\boldsymbol{\mu}).$$

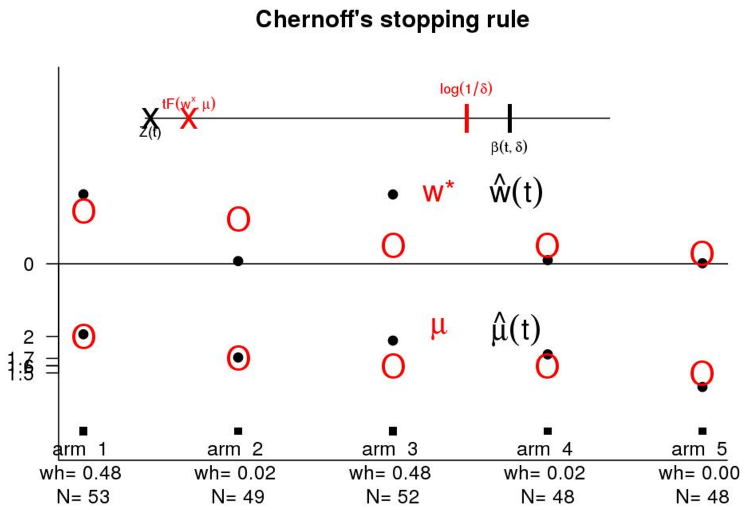
Why is the T&S Strategy asymptotically Optimal?

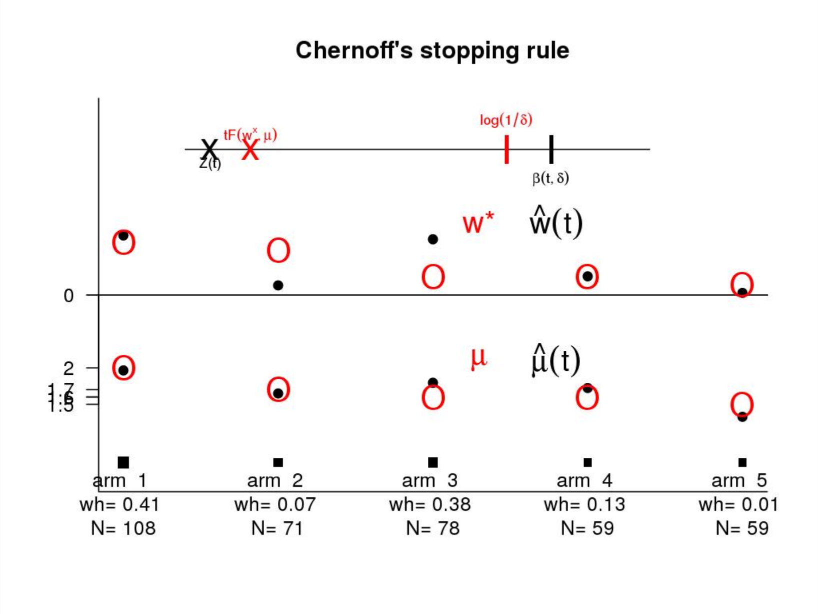


Why is the T&S Strategy asymptotically Optimal?

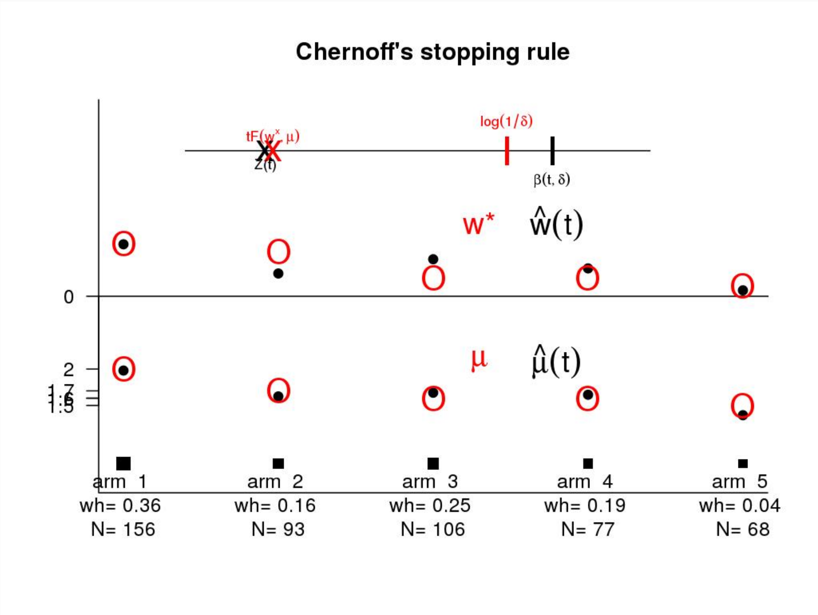


Why is the T&S Strategy asymptotically Optimal?

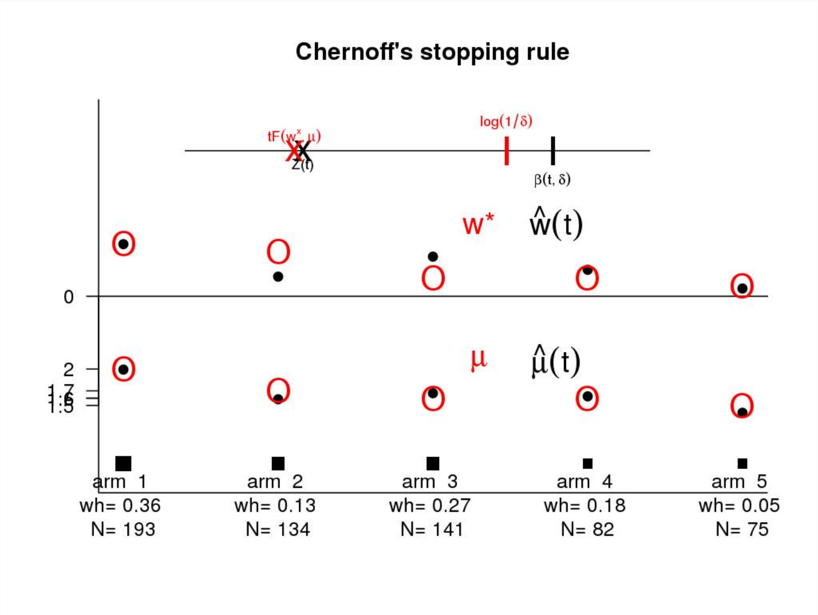



Chernoff's stopping rule
tF(w*, μ)
Z(t)
log(1/δ)
β(t, δ)
w*
ŵ(t)
0
μ
μ̂(t)
2
1.7
1.6
1.5
arm 1
wh= 0.41
N= 108
arm 2
wh= 0.07
N= 71
arm 3
wh= 0.38
N= 78
arm 4
wh= 0.13
N= 59
arm 5
wh= 0.01
N= 59

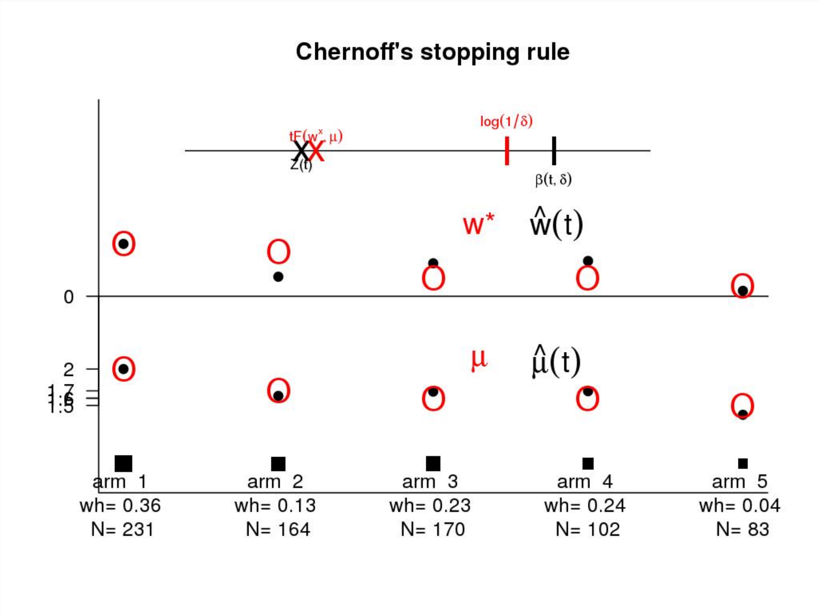
Why is the T&S Strategy asymptotically Optimal?



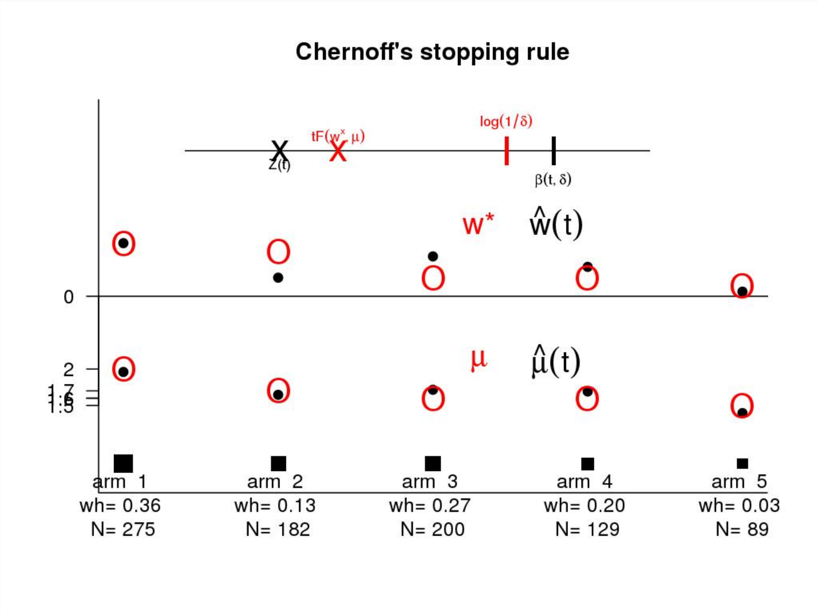
Why is the T&S Strategy asymptotically Optimal?



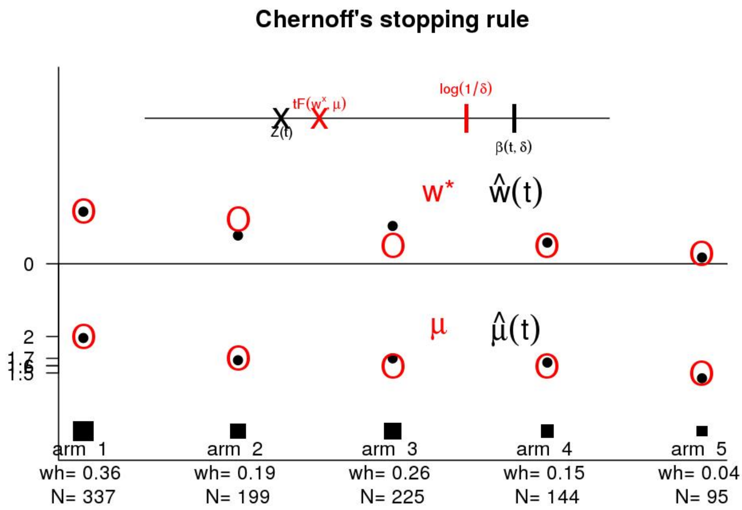
Why is the T&S Strategy asymptotically Optimal?



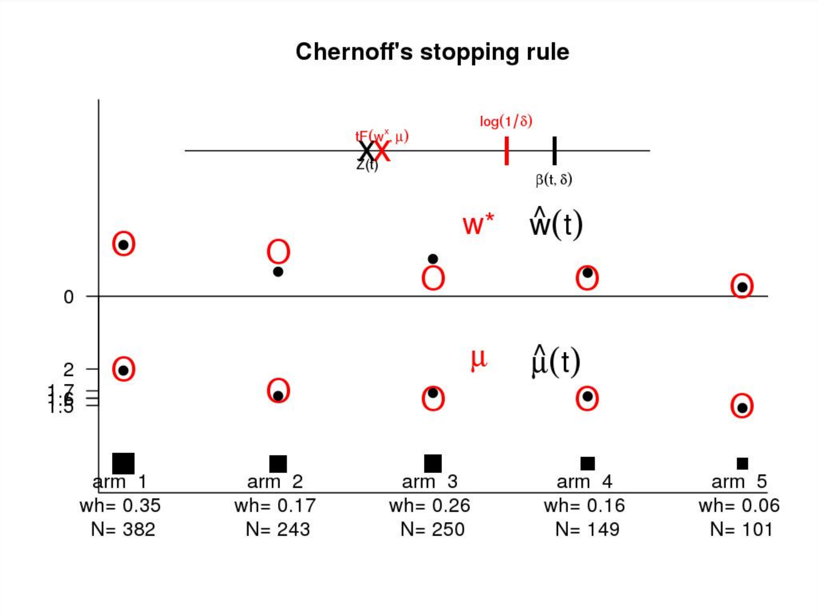
Why is the T&S Strategy asymptotically Optimal?



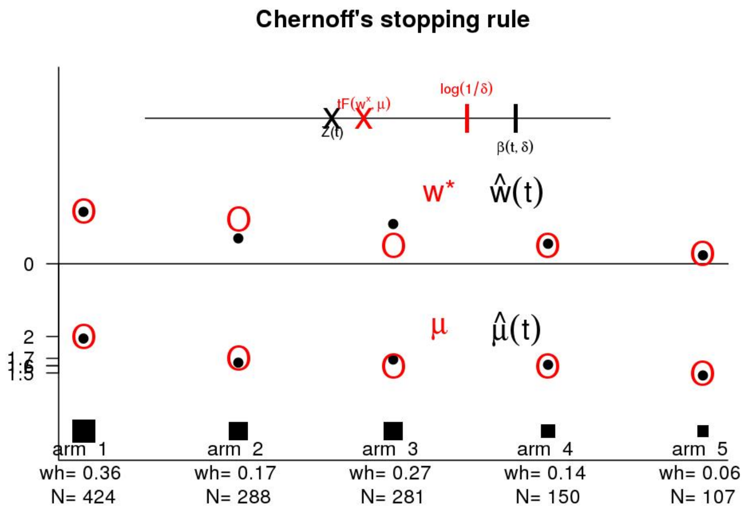
Why is the T&S Strategy asymptotically Optimal?



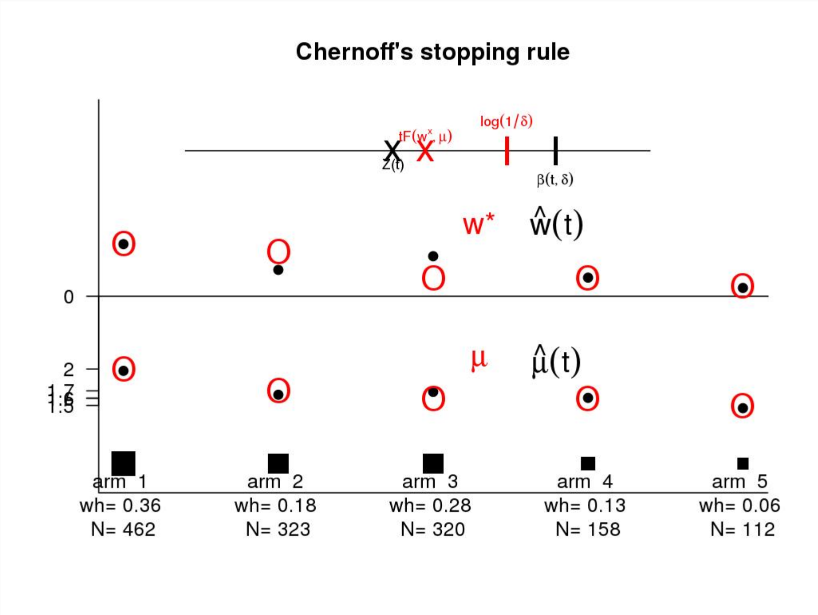
Why is the T&S Strategy asymptotically Optimal?



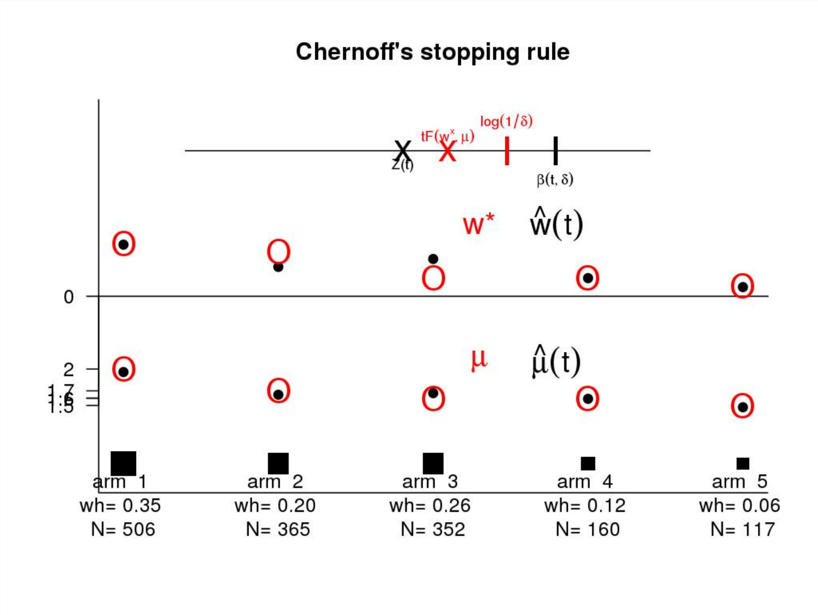
Why is the T&S Strategy asymptotically Optimal?



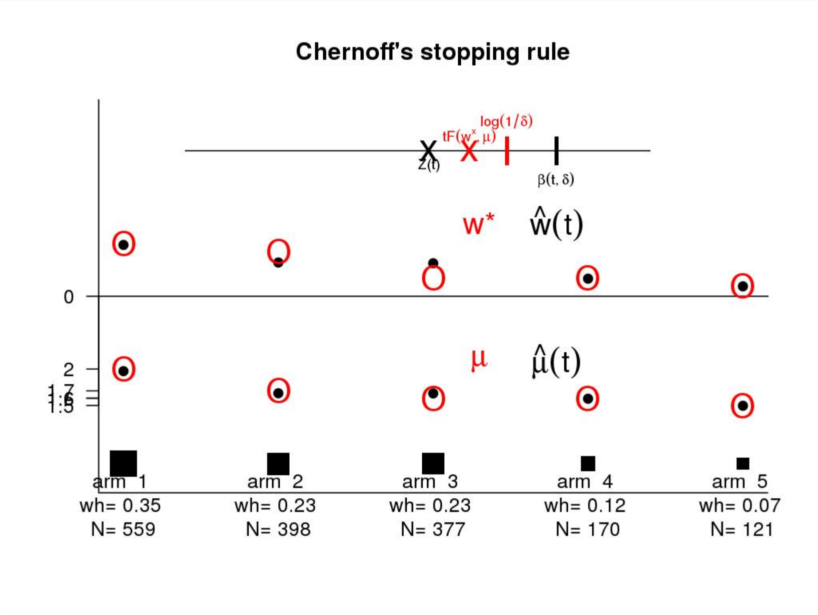
Why is the T&S Strategy asymptotically Optimal?



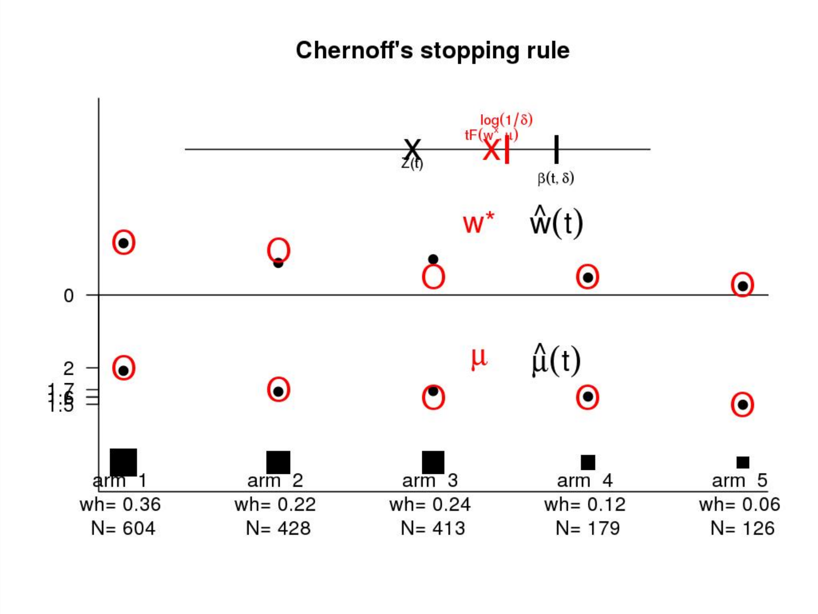
Why is the T&S Strategy asymptotically Optimal?



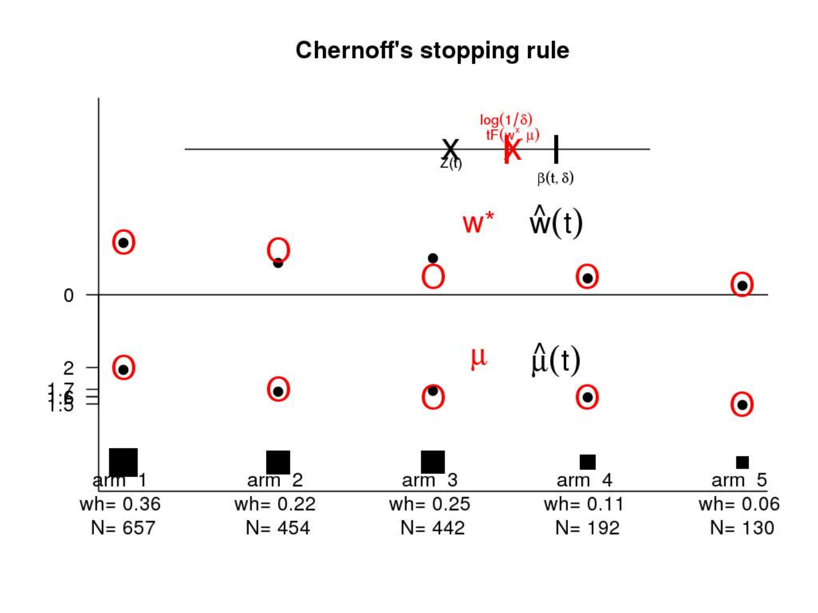
Why is the T&S Strategy asymptotically Optimal?



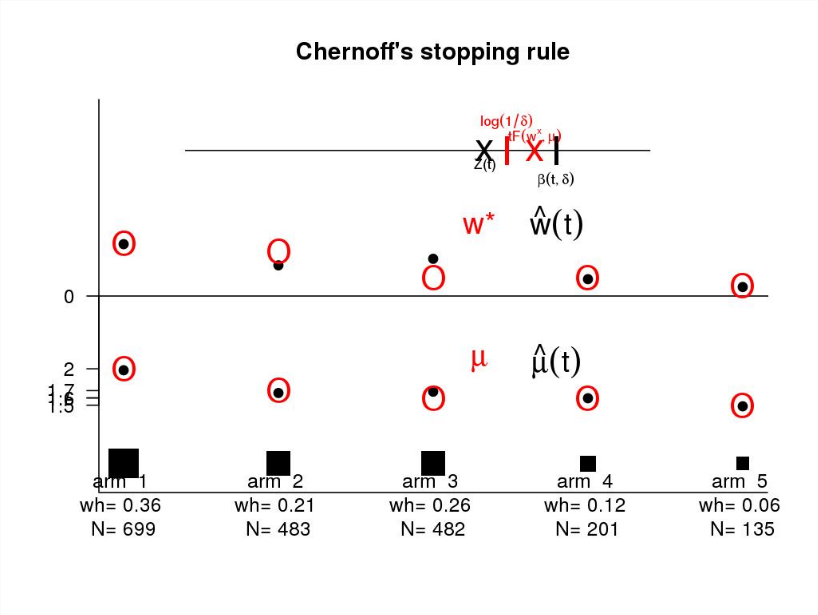
Why is the T&S Strategy asymptotically Optimal?



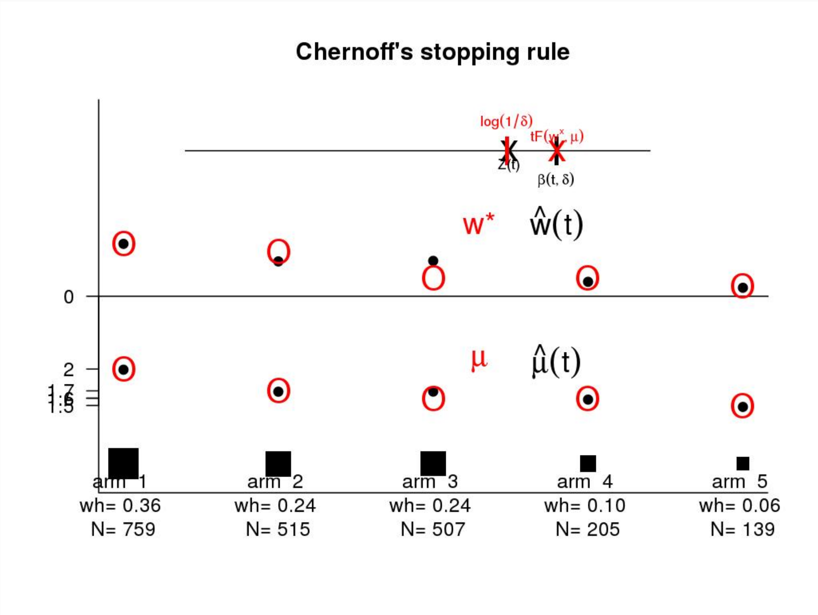
Why is the T&S Strategy asymptotically Optimal?



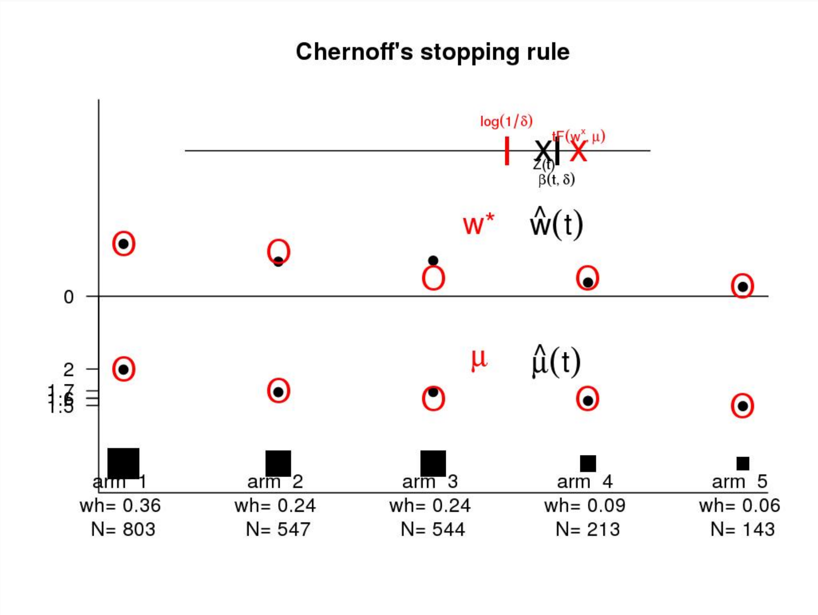
Why is the T&S Strategy asymptotically Optimal?



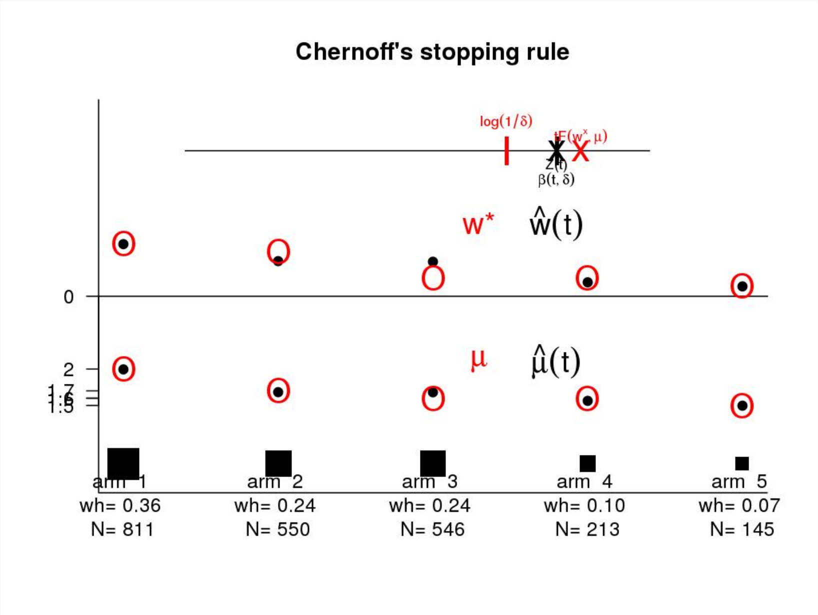
Why is the T&S Strategy asymptotically Optimal?



Why is the T&S Strategy asymptotically Optimal?



Why is the T&S Strategy asymptotically Optimal?



Sketch of proof (almost-sure convergence only)

- forced exploration $\implies$ $N_a(t) \to \infty$ a.s. for all $a \in \{1, \dots, K\}$

➜ $\hat{\boldsymbol{\mu}}(t) \to \boldsymbol{\mu}$ a.s.

➜ $\boldsymbol{w}^*\big(\hat{\boldsymbol{\mu}}(t)\big) \to \boldsymbol{w}^*$ a.s.

➜ tracking rule: $\dfrac{N_a(t)}{t} \underset{t\to\infty}{\to} w_a^*$ a.s.

- but the mapping $F : (\boldsymbol{\mu}', w) \mapsto \inf\limits_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\mu}')} \sum\limits_{a=1}^{K} w_a d(\mu'_a, \lambda_a)$ is continuous at $(\boldsymbol{\mu}, w^*(\boldsymbol{\mu}))$:

➜ $Z(t) = t \times F\Big(\hat{\boldsymbol{\mu}}(t), (N_a(t)/t)_{a=1}^{K}\Big) \sim t \times F(\mu, \mathbf{w}^*) = t \times T^*(\mu)^{-1}$
and for every $\epsilon > 0$ there exists t_0 such that

$$t \geq t_0 \ \Rightarrow \ Z(t) \geq t \times (1+\epsilon)^{-1} T^*(\boldsymbol{\mu})^{-1}$$

$\implies$ Thus $\tau_\delta \leq t_0 \wedge \inf\left\{t \in \mathbb{N} : (1+\epsilon)^{-1} T^*(\boldsymbol{\mu})^{-1} t \geq \log(2(K-1)t/\delta)\right\}$
and $\limsup\limits_{\delta\to 0} \dfrac{\tau_\delta}{\log(1/\delta)} \leq (1+\epsilon) T^*(\boldsymbol{\mu})$ *a.s.*

Numerical Experiments

- $\boldsymbol{\mu}_1 = [0.5\ 0.45\ 0.43\ 0.4] \quad ➜ \quad w^*(\boldsymbol{\mu}_1) = [0.42\ 0.39\ 0.14\ 0.06]$
- $\boldsymbol{\mu}_2 = [0.3\ 0.21\ 0.2\ 0.19\ 0.18] ➜ w^*(\boldsymbol{\mu}_2) = [0.34\ 0.25\ 0.18\ 0.13\ 0.10]$

In practice, set the threshold to $\beta(t,\delta) = \log\left(\frac{\log(t)+1}{\delta}\right)$ (δ-PAC OK)

	Track-and-Stop	Chernoff-Racing	KL-LUCB	KL-Racing
$\boldsymbol{\mu}_1$	4052	4516	8437	9590
$\boldsymbol{\mu}_2$	1406	3078	2716	3334

Table 1: Expected number of draws $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ for $\delta = 0.1$, averaged over $N = 3000$ experiments.

➜ Empirically good even for 'large' values of the risk δ

➜ Racing is sub-optimal in general, because it plays $w_1 = w_2$

➜ LUCB is sub-optimal in general, because it plays $w_1 = 1/2$

Perspectives

For best arm identification, we showed that

$$\limsup_{\delta \to 0} \inf_{\delta\text{-correct strategy}} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} = \left(\sup_{w \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \right) \right)^{-1}$$

and provided an efficient strategy asymptotically matching this bound.

Future work:

- $*$ anytime stopping ➜ gives a confidence level
- $**$ find an ϵ-optimal arm (PAC-setting)
- $*$ find the m-best arms
- $***$ design and analyze more stable algorithm (hint: optimism)
- $***$ give a simple algorithm with a finite-time analysis
  candidate: play action maximizing the expected increase of $Z(t)$
- $***$ extend to structured (dose, MCTS) and continuous settings