On the Complexity of Best Arm Identification with Fixed Confidence

Discrete Optimization with Noise

Aurélien Garivier[†], Emilie Kaufmann[⋆]

COLT, June 23th 2016, New York

[†] Institut de Mathématiques de Toulouse
LabeX CIMI
Université Paul Sabatier, France

[⋆] Université Lille, CNRS UMR 9189
Laboratoire CRIStAL
F-59000 Lille, France

The Problem

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

ν_1

ν_2

ν_3

ν_4

ν_5

At round t, you may:

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

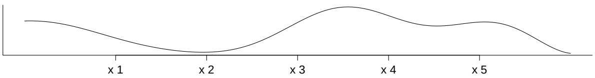
so as to identify the best arm $a^* = \text{argmax}_a\, \mu_a$ and $\mu^* = \max_a \mu_a$
as fast as possible: stopping time τ .

Fixed-budget setting	Fixed-confidence setting
given $\tau = T$ minimize $\mathbb{P}(\hat{a}_\tau \neq a^*)$	minimize $\mathbb{E}[\tau]$ under constraint $\mathbb{P}(\hat{a}_\tau \neq a^*) \leq \delta$

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 arm $a^* = \mathrm{argmax}_a\, \mu_a$ and $\mu^* = \max_a \mu_a$

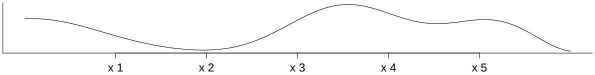
as fast as possible: stopping time τ .

Fixed-budget setting	Fixed-confidence setting
given $\tau = T$ minimize $\mathbb{P}(\hat{a}_\tau \neq a^*)$	minimize $\mathbb{E}[\tau]$ under constraint $\mathbb{P}(\hat{a}_\tau \neq a^*) \le \delta$

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

so as to identify the best arm $a^* = \text{argmax}_a\ \mu_a$ and $\mu^* = \max_a \mu_a$

as fast as possible: stopping time τ_δ .

Fixed-budget setting	Fixed-confidence setting
given $\tau = T$	minimize $\mathbb{E}[\tau_\delta]$
minimize $\mathbb{P}(\hat{a}_\tau \neq a^*)$	under constraint $\mathbb{P}(\hat{a}_\tau \neq a^*) \le \delta$

Intuition

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

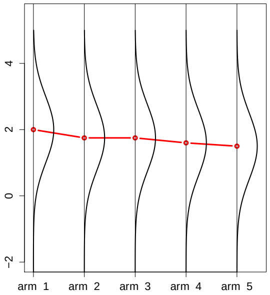
$$\nu_a = \mathcal{N}(\mu_a, 1)$$

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

At time t:

➜ you have sampled n_a times the option a

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



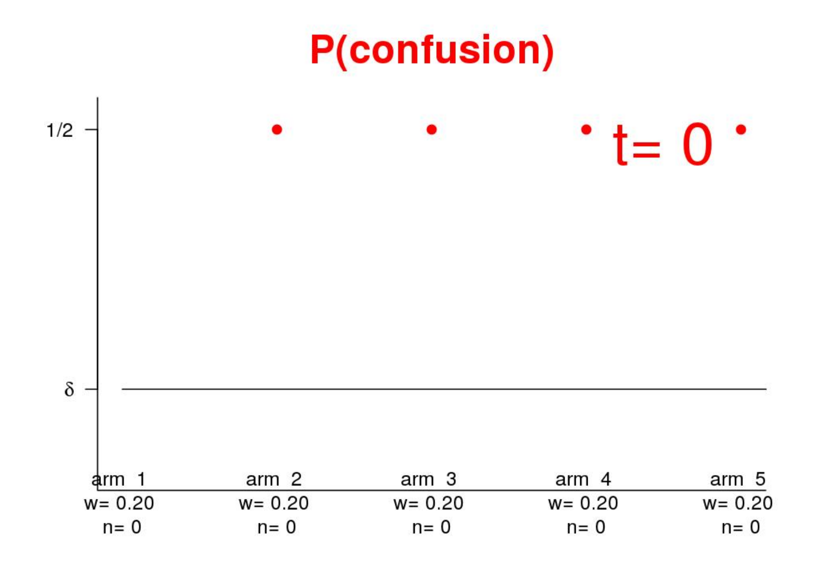
$\longrightarrow$ if you stop at time t, your probability of prefering arm $a \geq 2$ to arm $a^* = 1$ is:

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

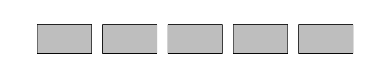
$$= \bar{\Phi}\left(\frac{\mu_1 - \mu_a}{\sqrt{1/n_1 + 1/n_a}}\right)$$

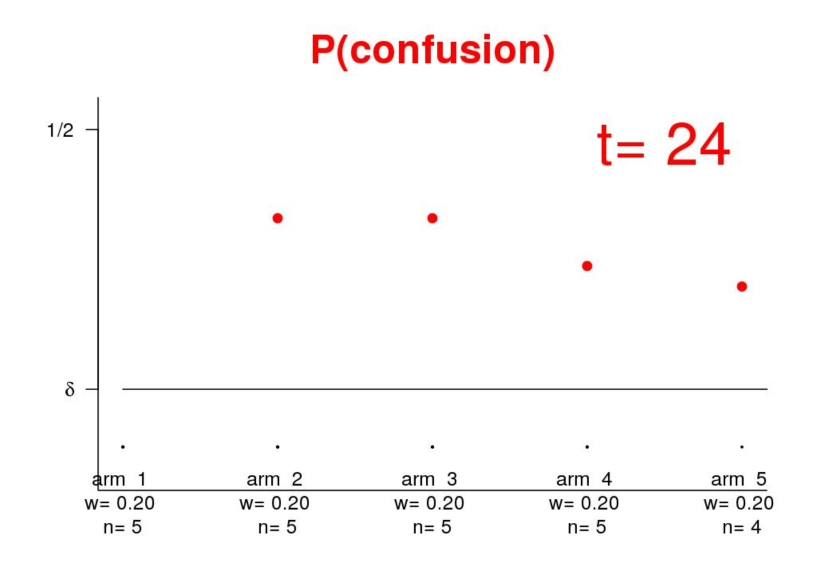
where $\bar{\Phi}(u) = \int_u^\infty \frac{e^{-u^2/2}}{\sqrt{2\pi}} du$

Uniform Sampling

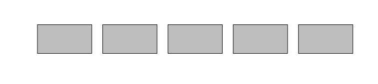


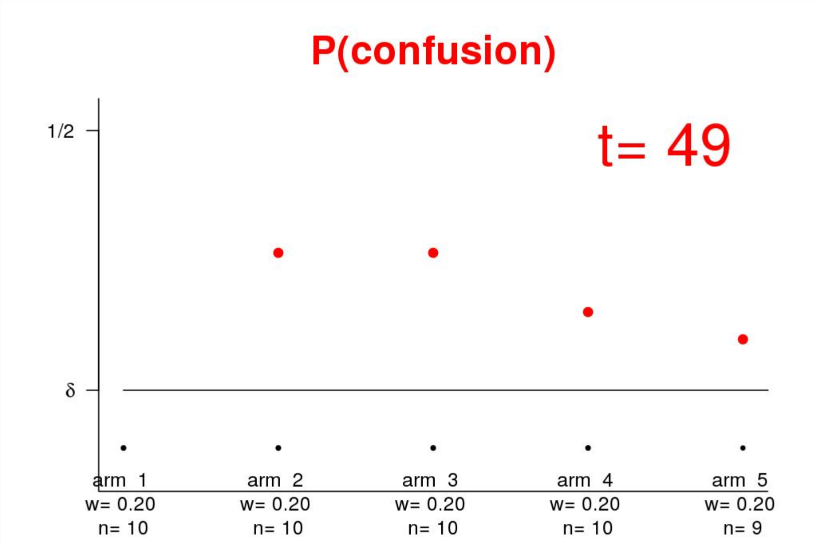
Uniform Sampling



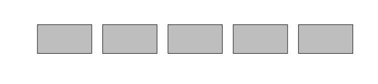


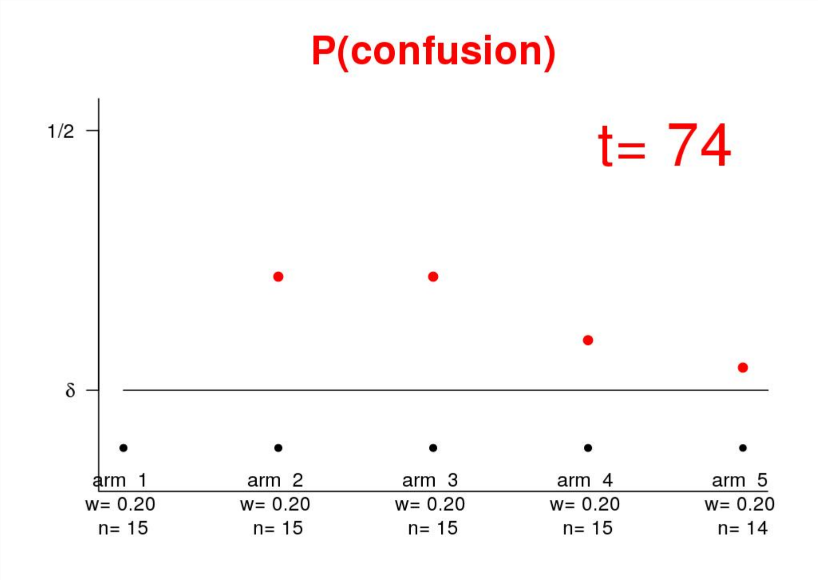
Uniform Sampling



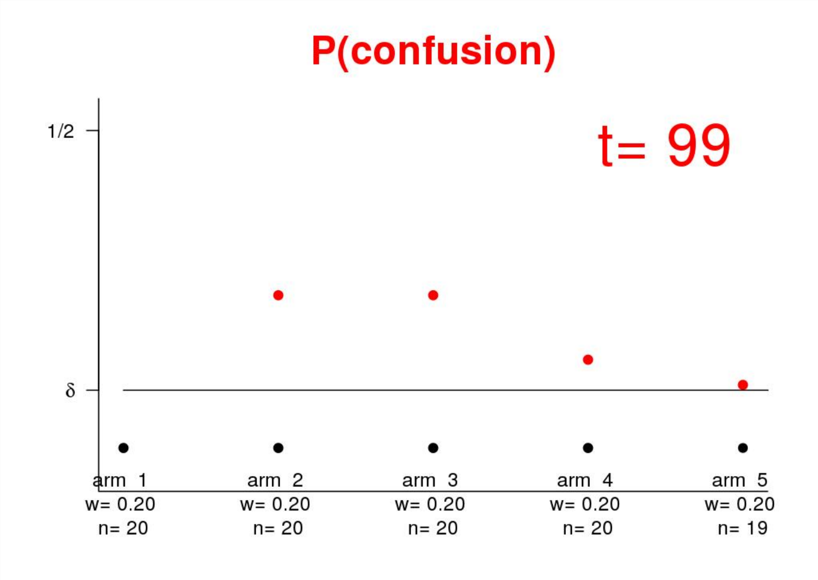


Uniform Sampling

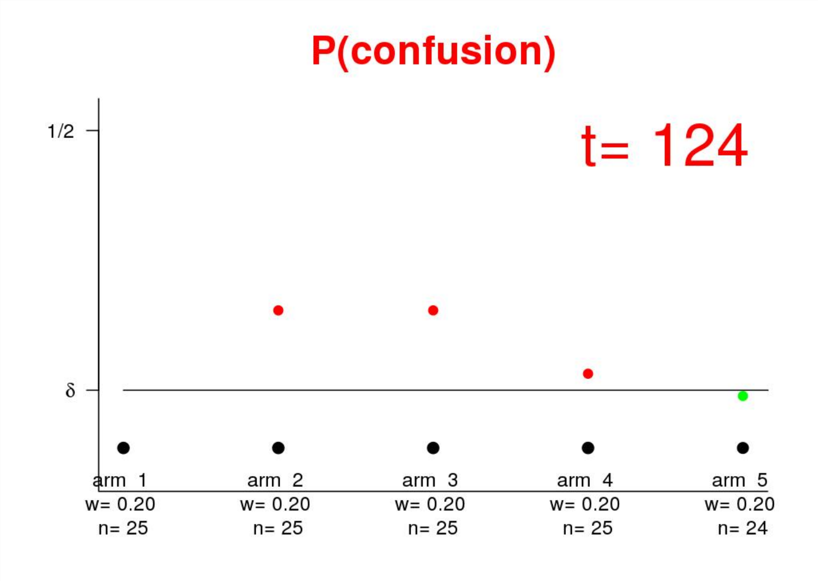




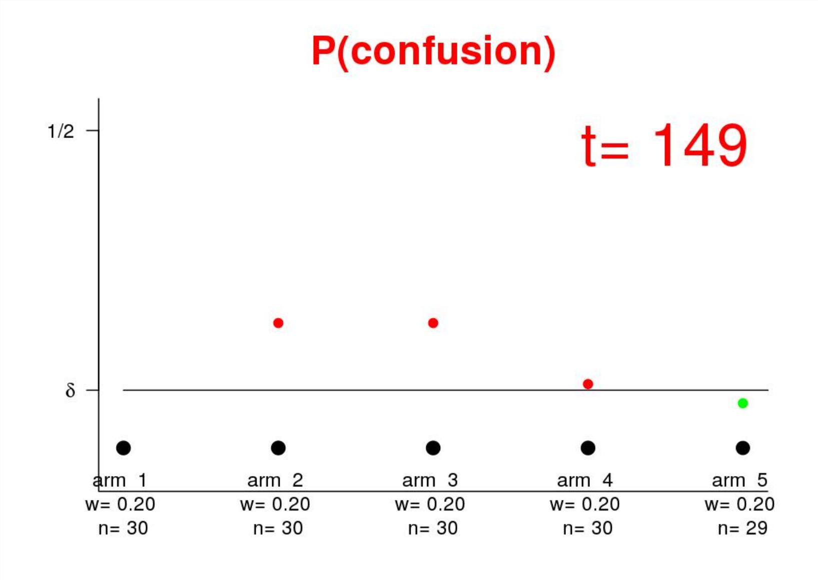
Uniform Sampling



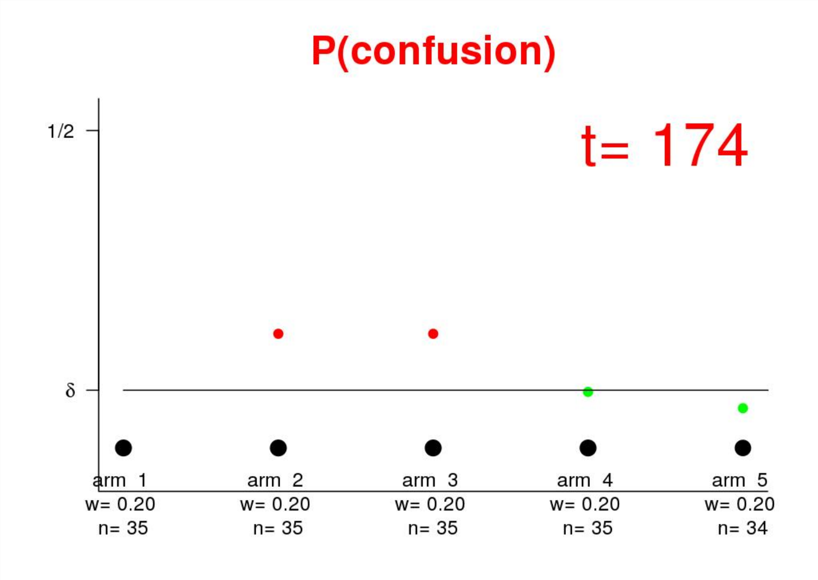
Uniform Sampling



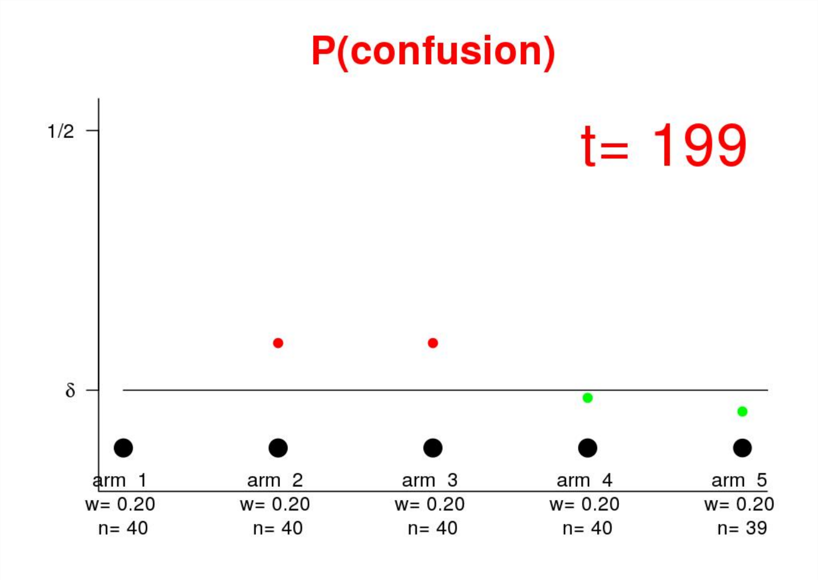
Uniform Sampling



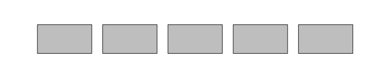
Uniform Sampling

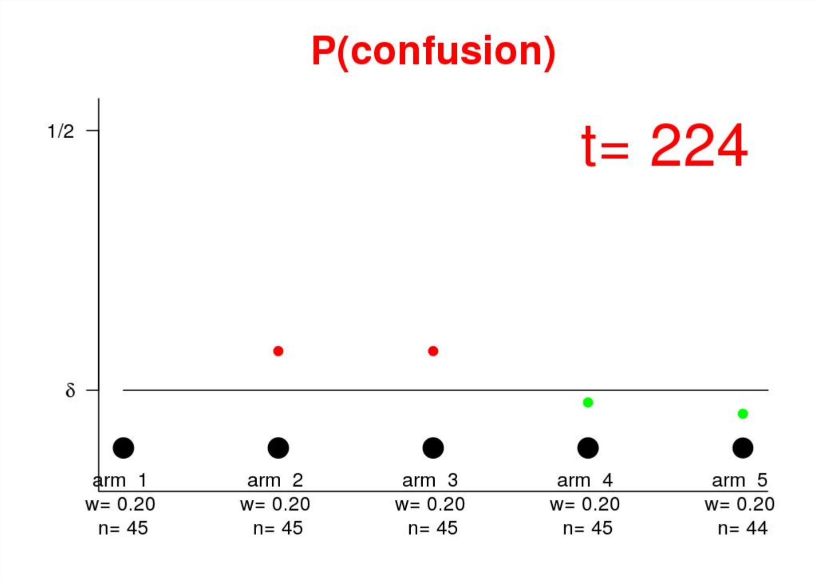


Uniform Sampling

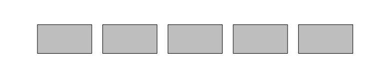


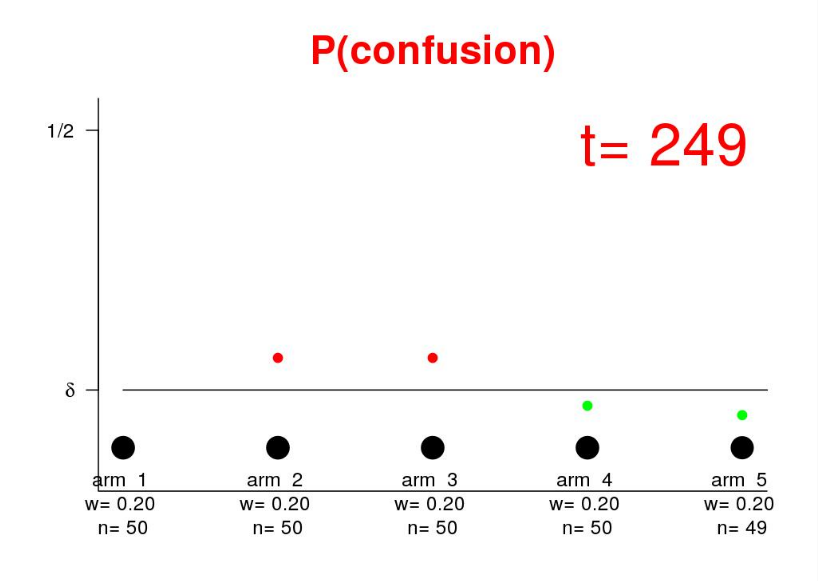
Uniform Sampling



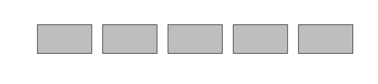


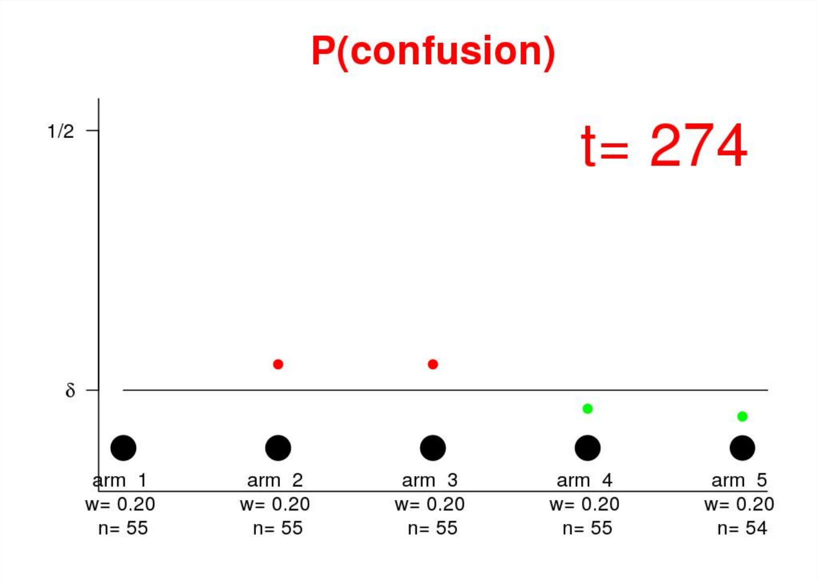
Uniform Sampling



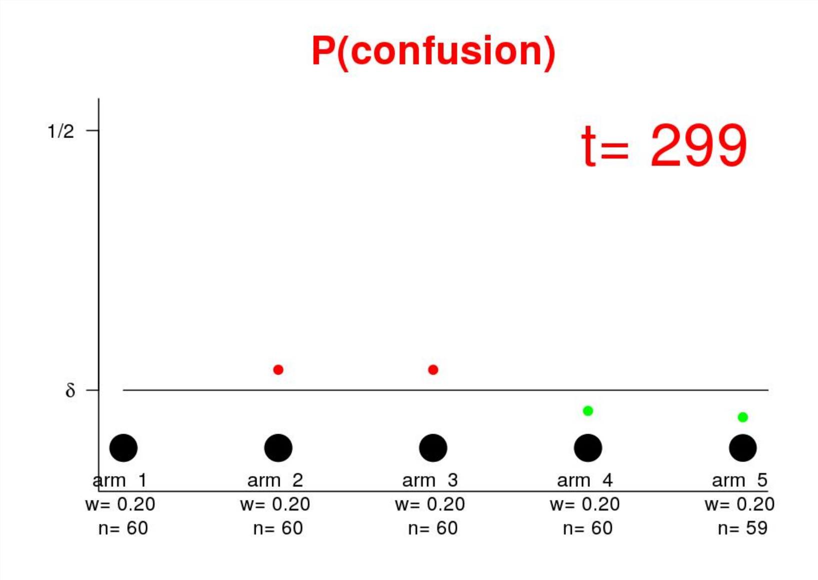


Uniform Sampling

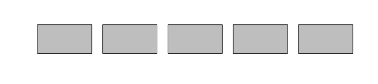
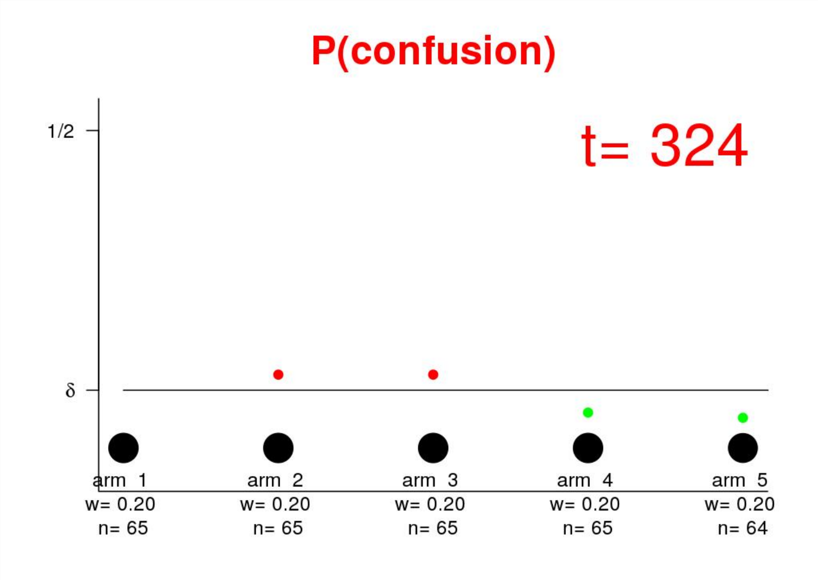




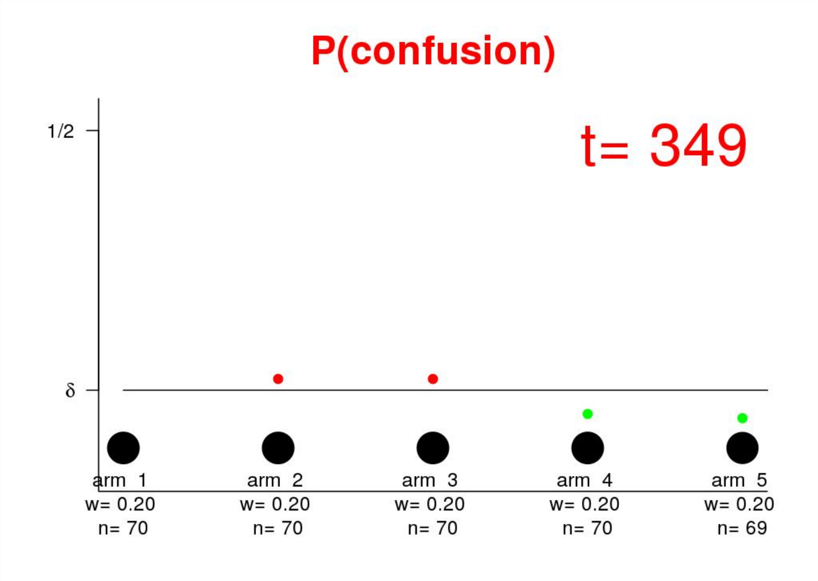
Uniform Sampling



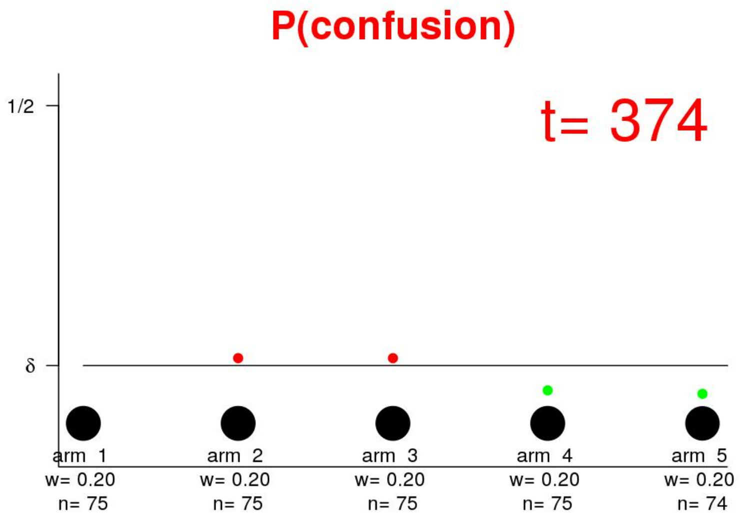
Uniform Sampling

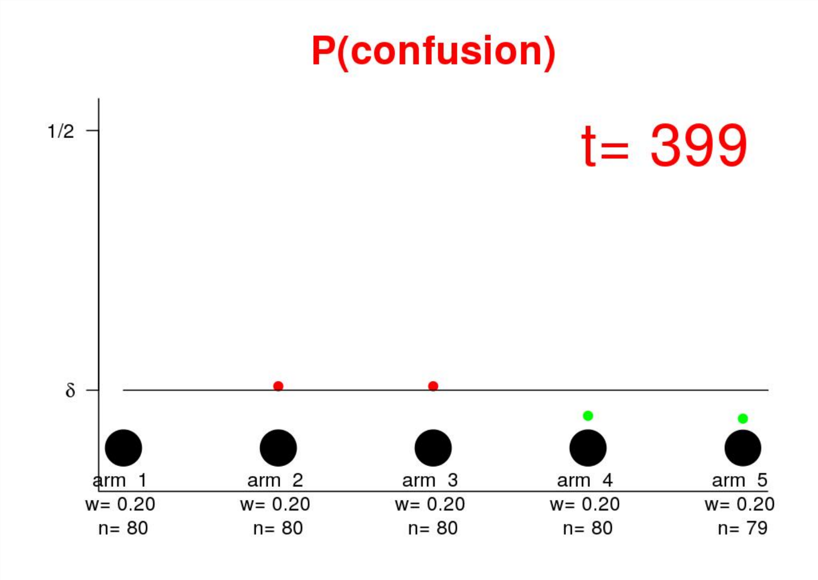
Uniform Sampling



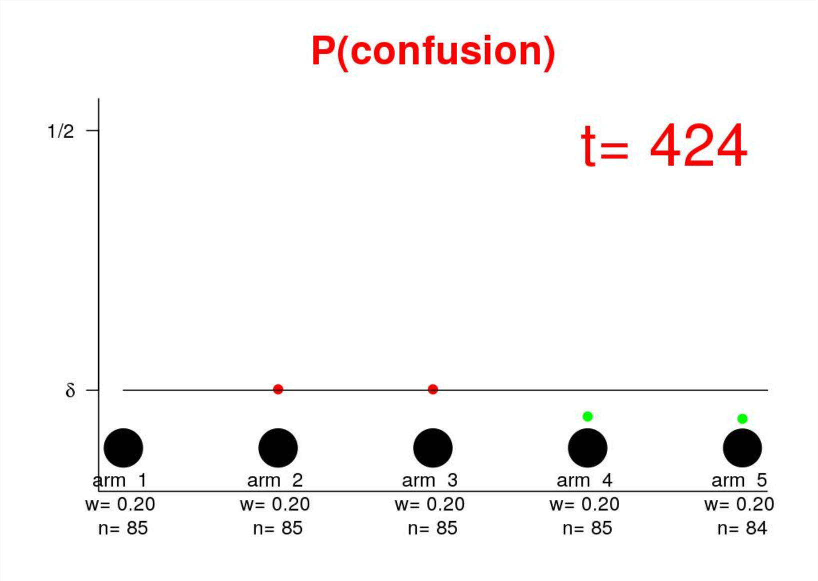
Uniform Sampling



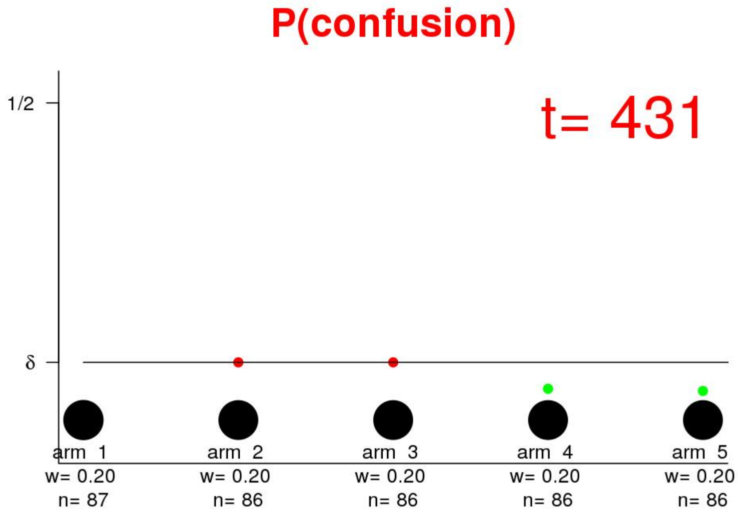
Uniform Sampling



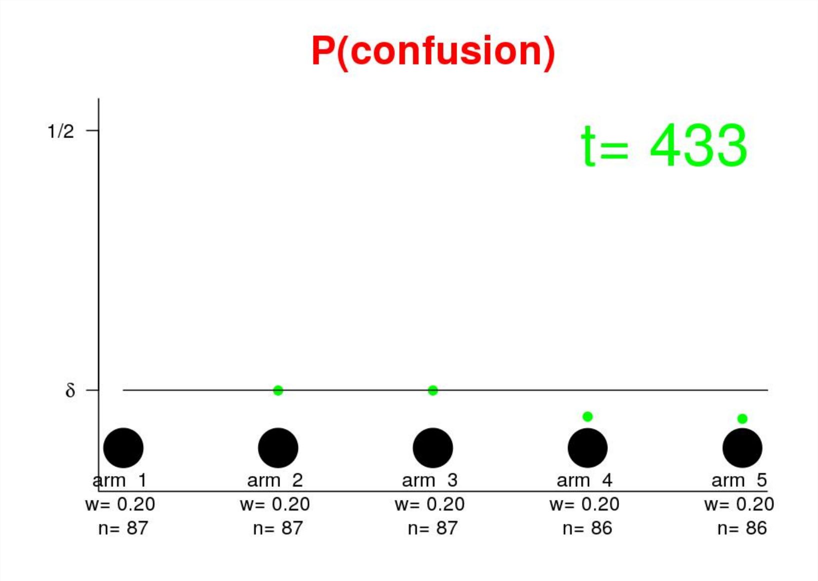
Uniform Sampling



Uniform Sampling



Uniform Sampling

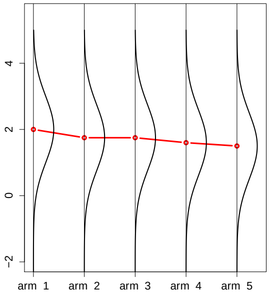


Intuition: Equalizing the Probabilities of Confusion

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 $w_1 + \dots + w_K = 1$.**

At time t:

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

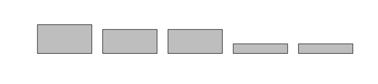
➜ your empirical average is $\bar{X}_{a,n_a}$.

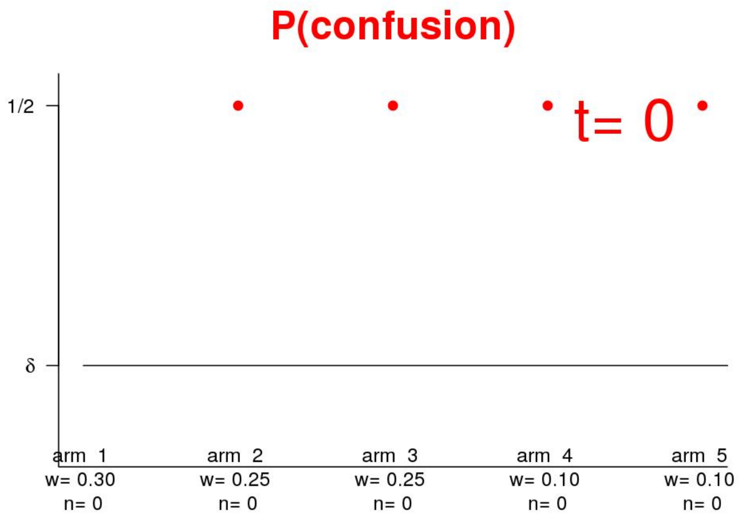
$\longrightarrow$ if you stop at time t, your probability of prefering arm $a \geq 2$ to arm $a^* = 1$ is:

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

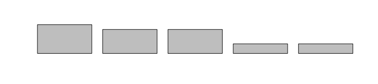
$$= \bar{\Phi}\left(\frac{\mu_1 - \mu_a}{\sqrt{1/n_1 + 1/n_a}}\right)$$

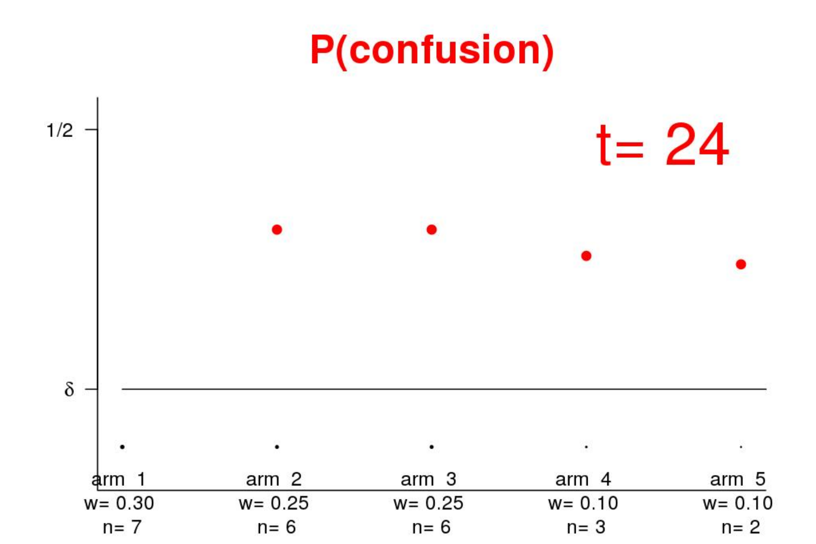
Improving: trial 1



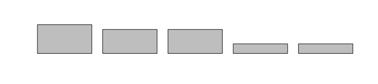


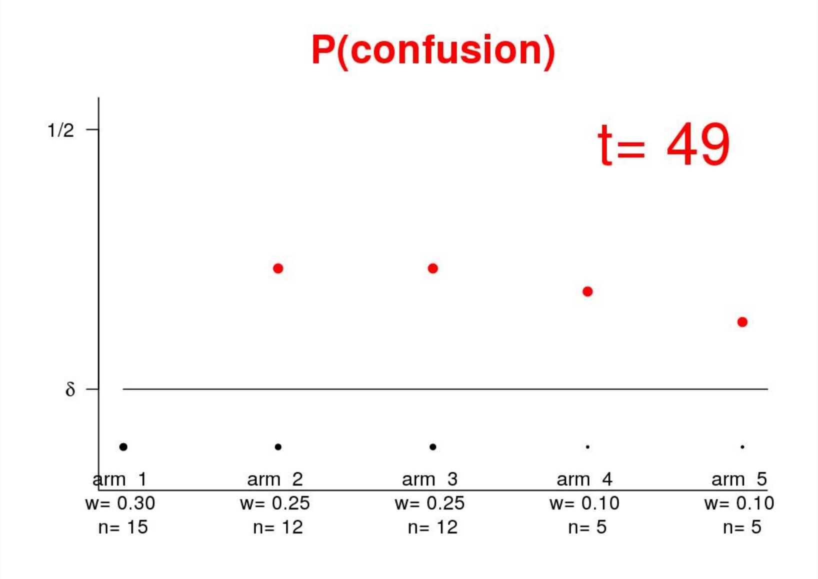
Improving: trial 1



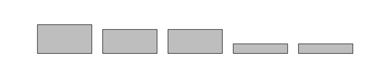


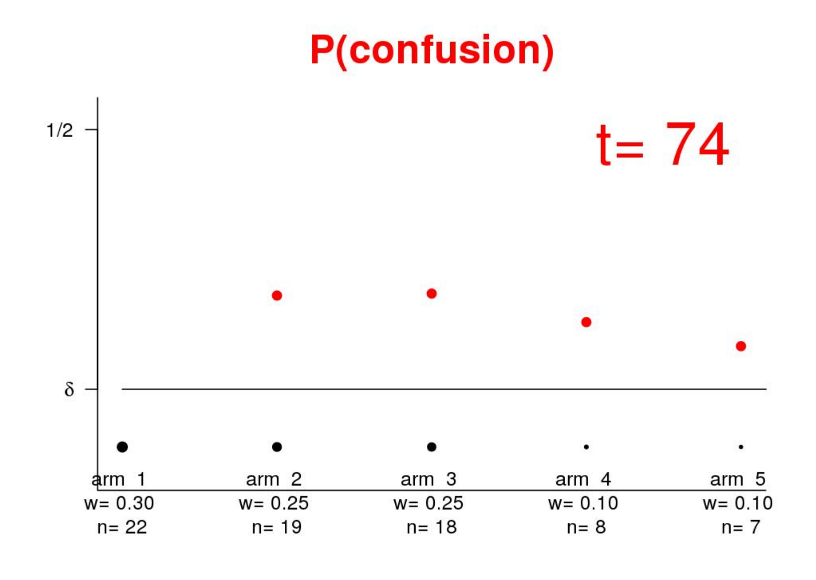
Improving: trial 1



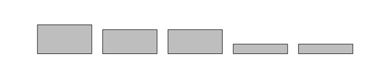


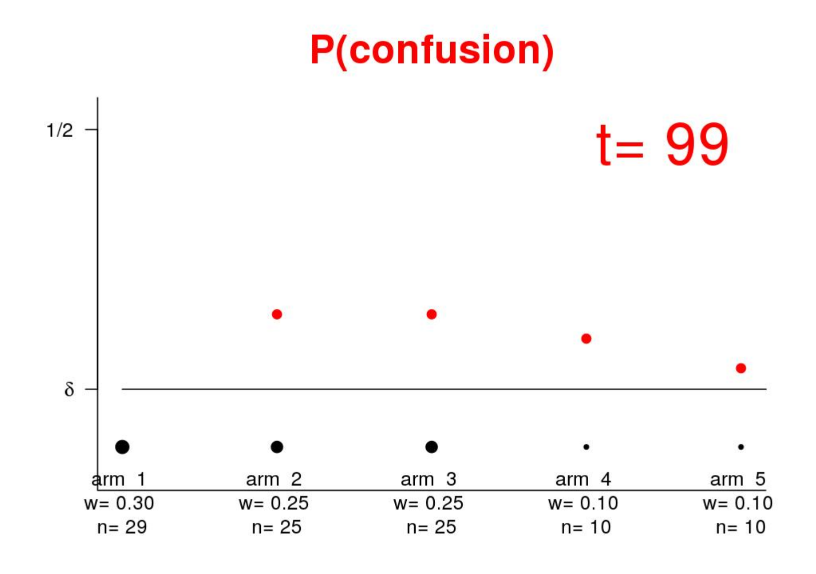
Improving: trial 1



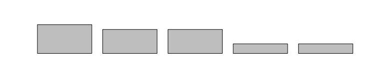


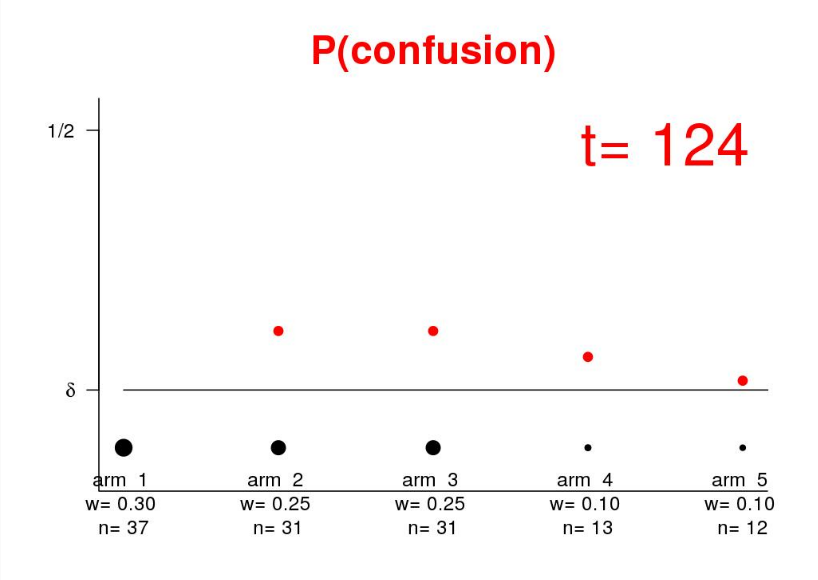
Improving: trial 1



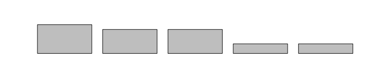


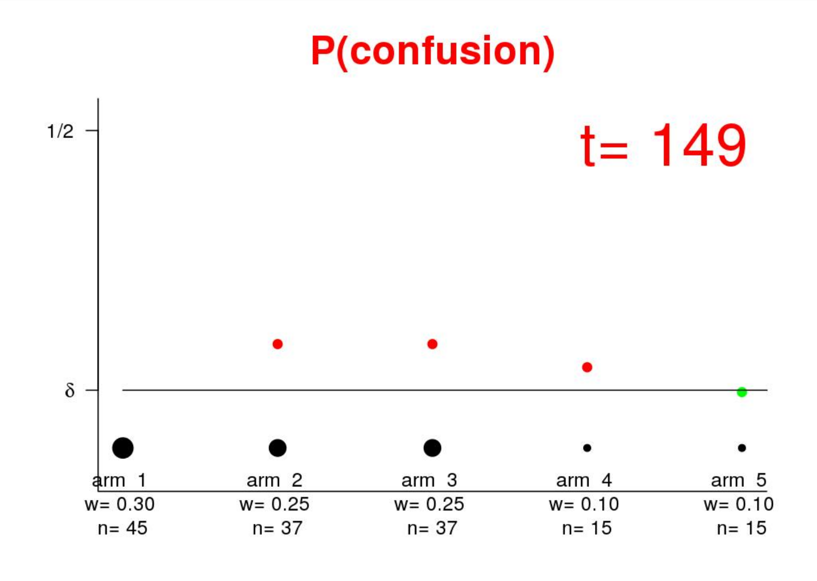
Improving: trial 1



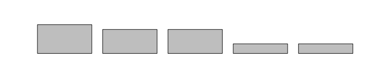


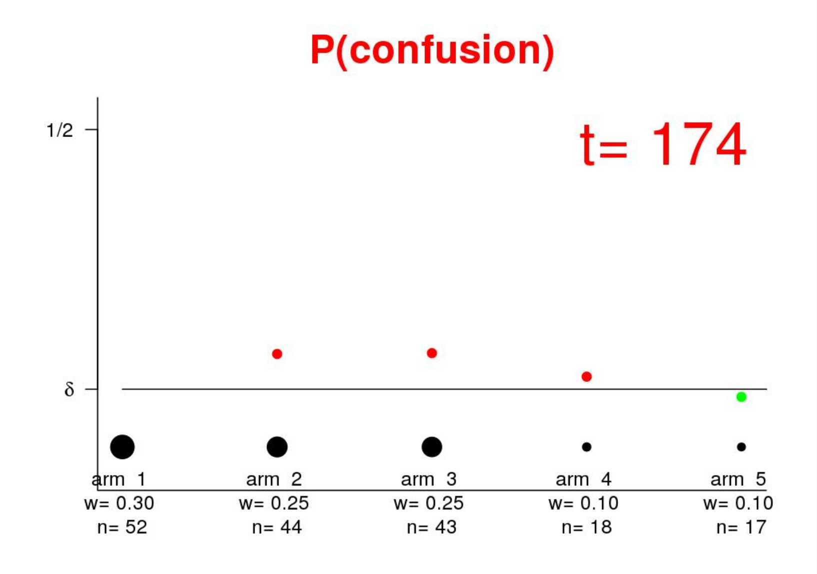
Improving: trial 1



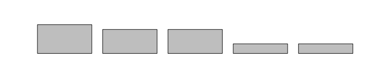


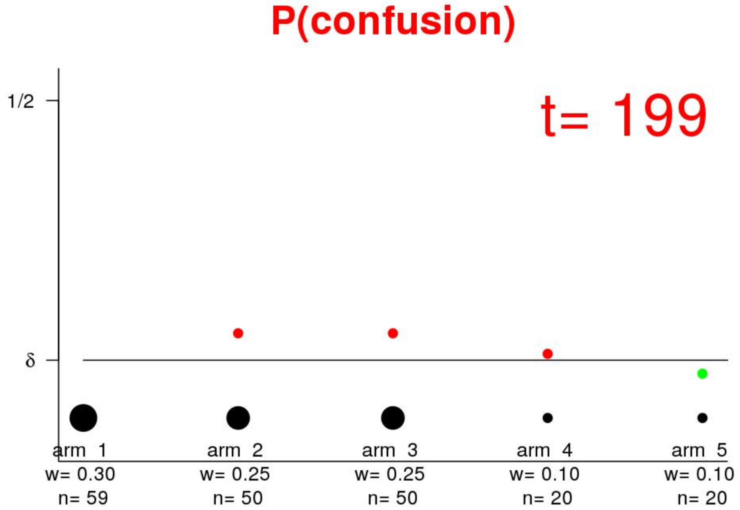
Improving: trial 1



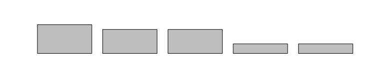


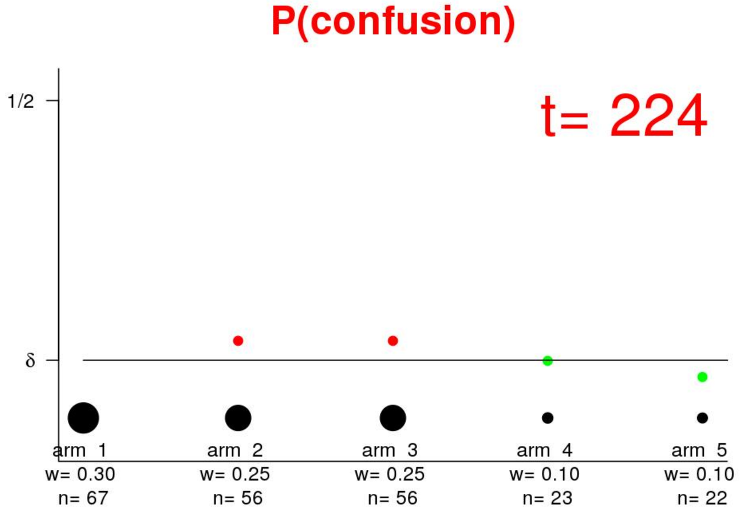
Improving: trial 1



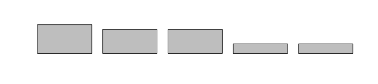


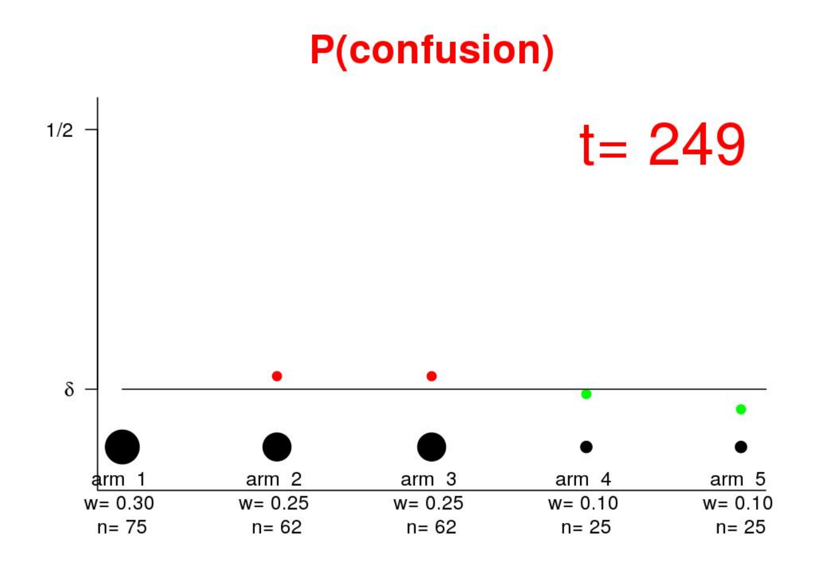
Improving: trial 1



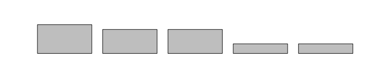


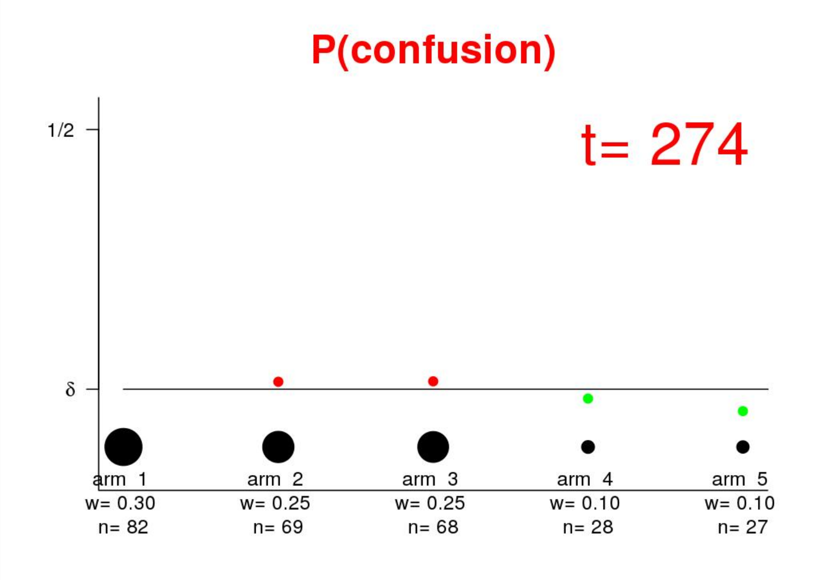
Improving: trial 1



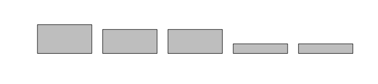


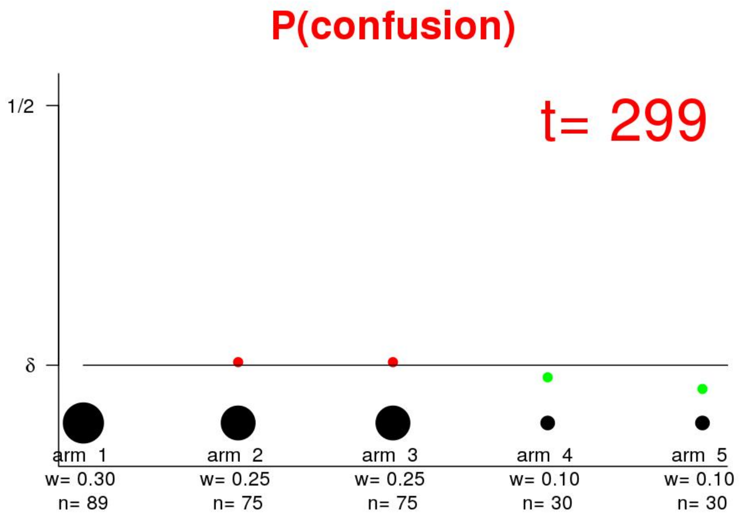
Improving: trial 1



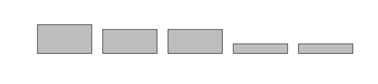


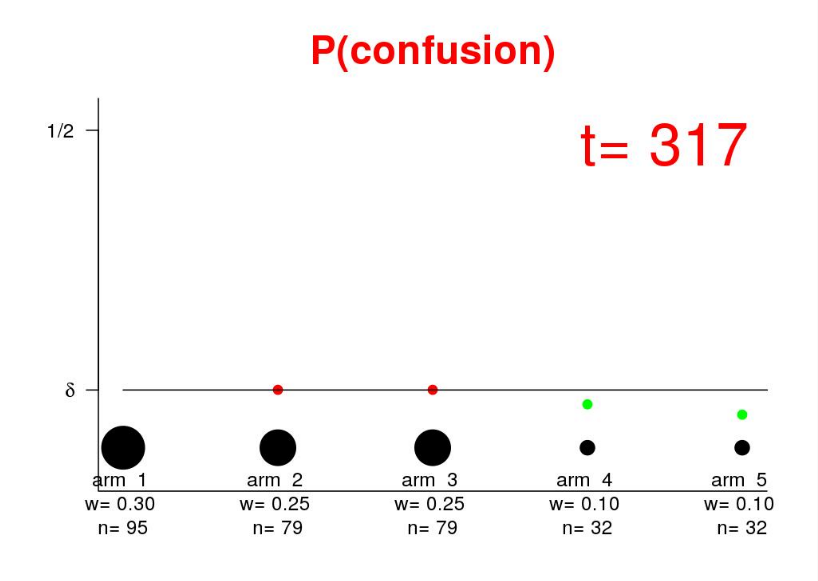
Improving: trial 1



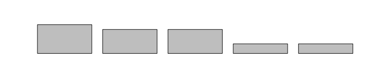


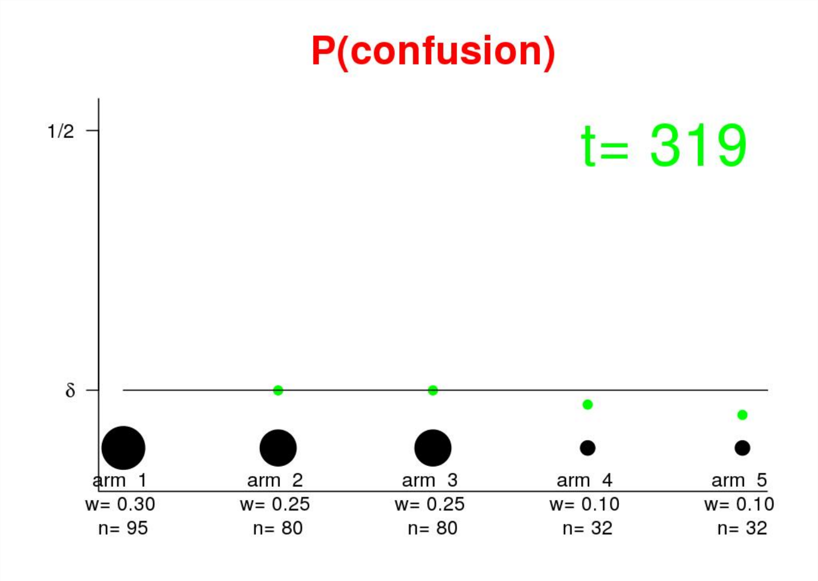
Improving: trial 1



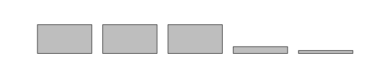


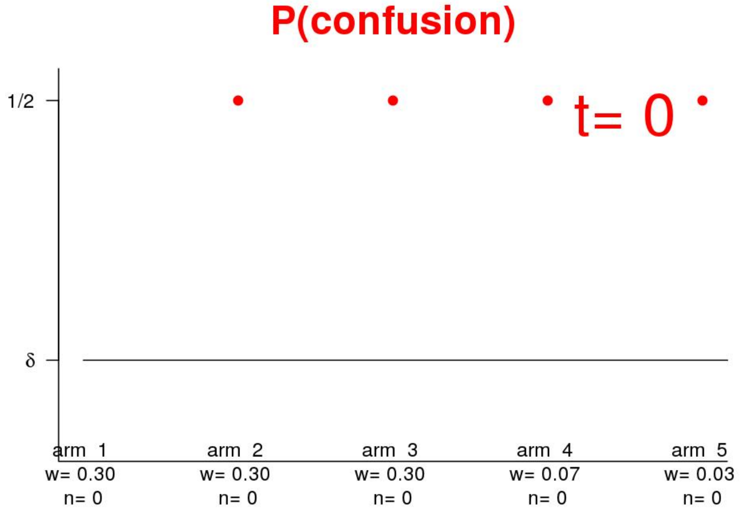
Improving: trial 1



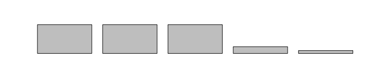


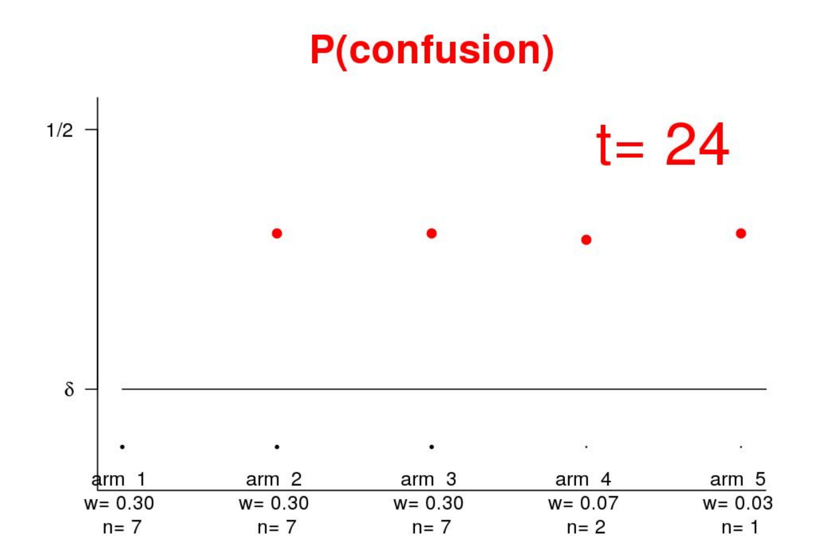
Improving: trial 2



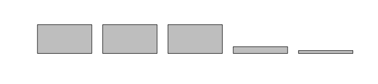


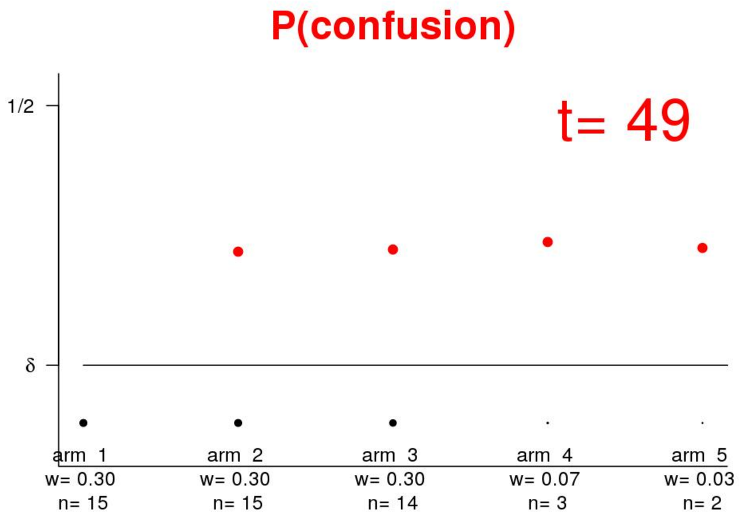
Improving: trial 2



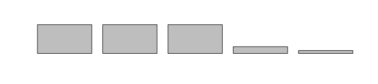


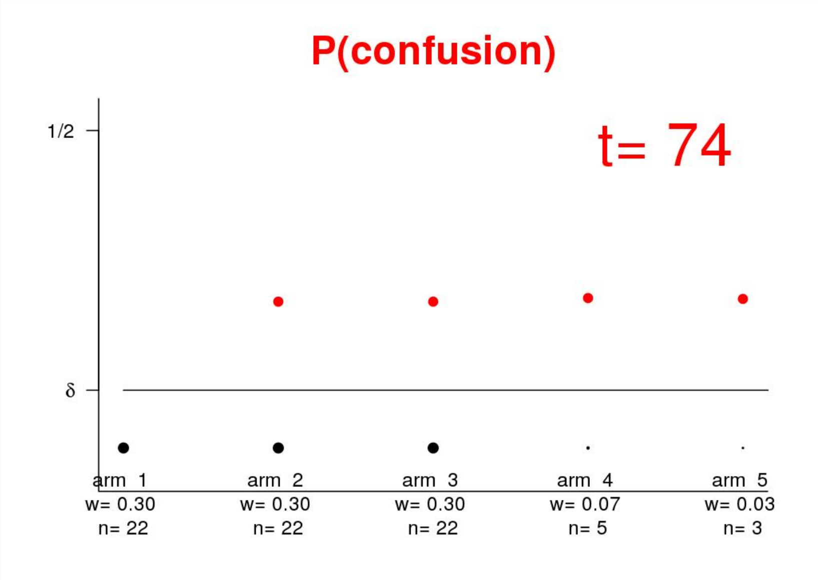
Improving: trial 2



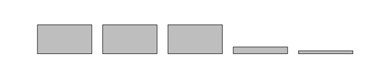


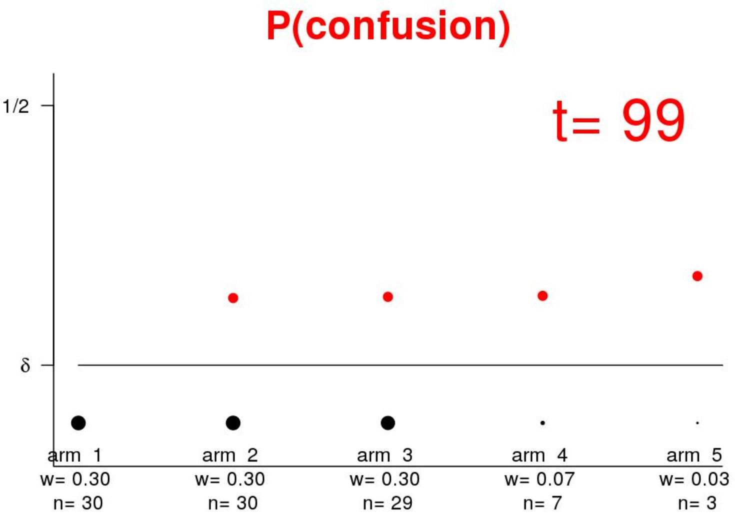
Improving: trial 2



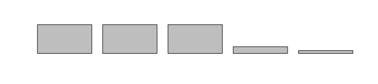


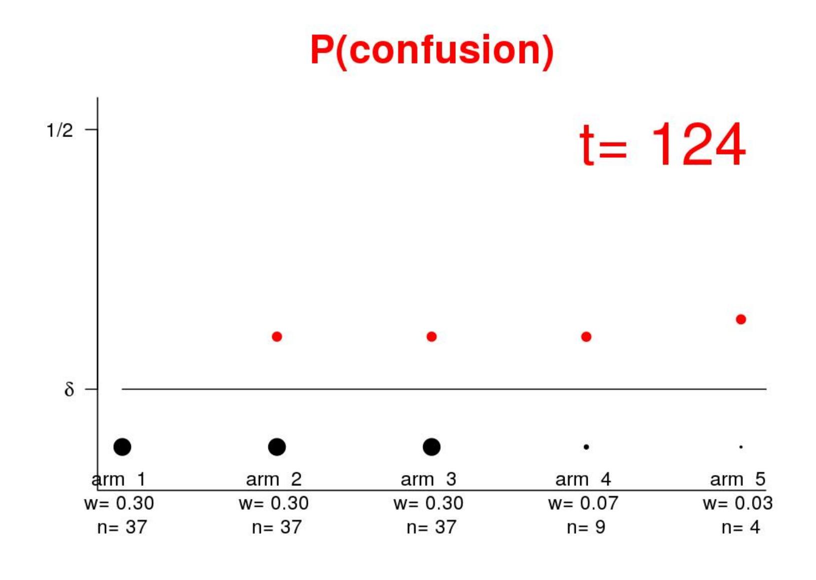
Improving: trial 2



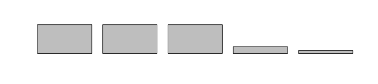


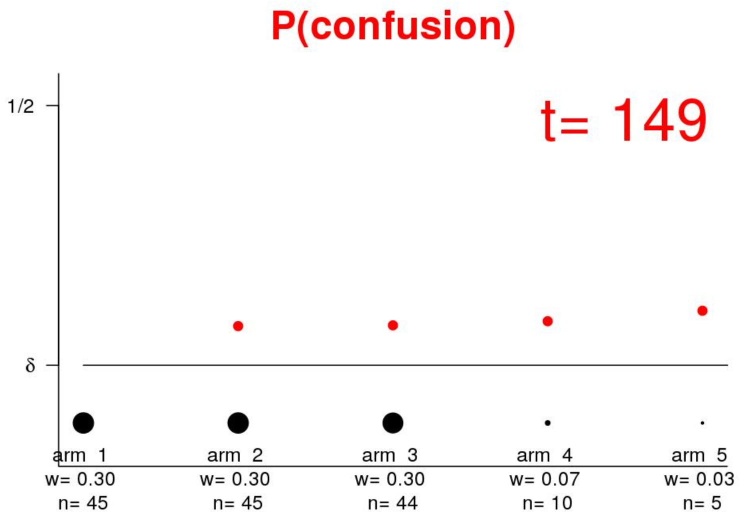
Improving: trial 2



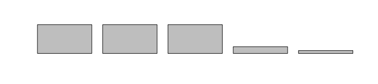


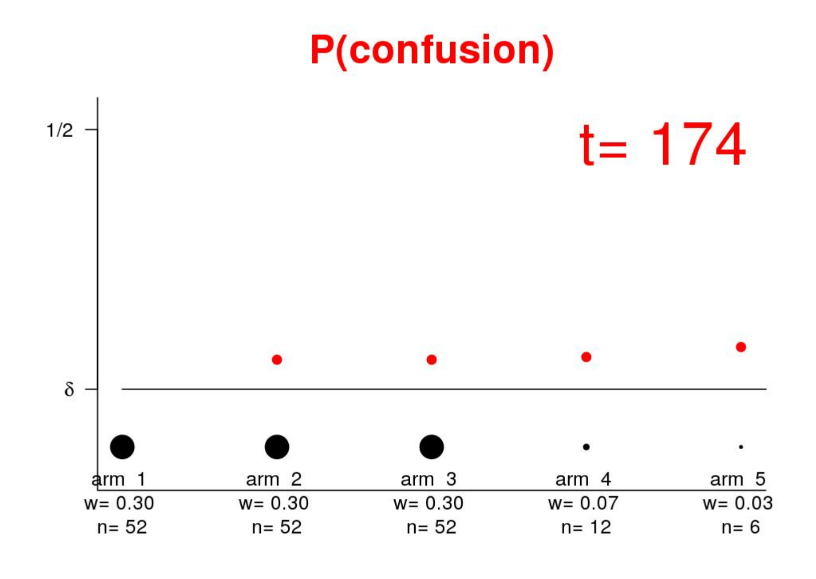
Improving: trial 2



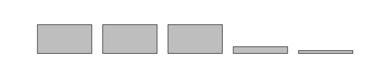


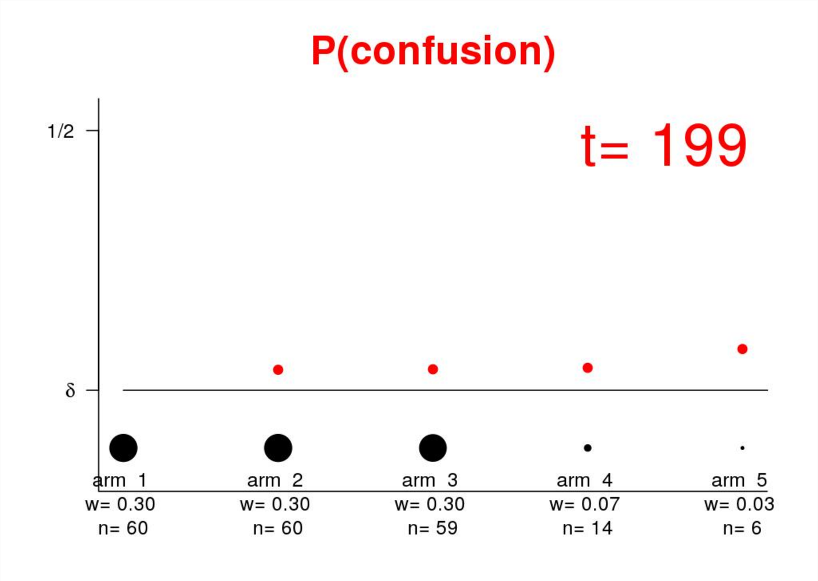
Improving: trial 2



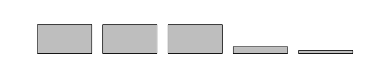


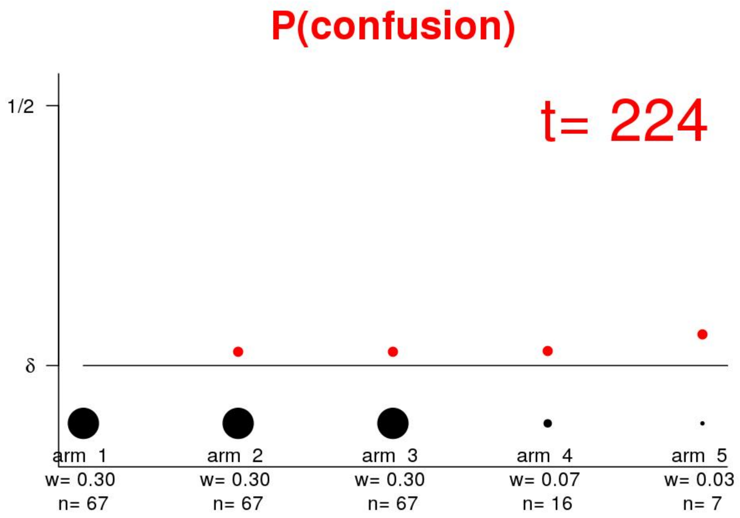
Improving: trial 2



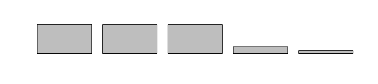


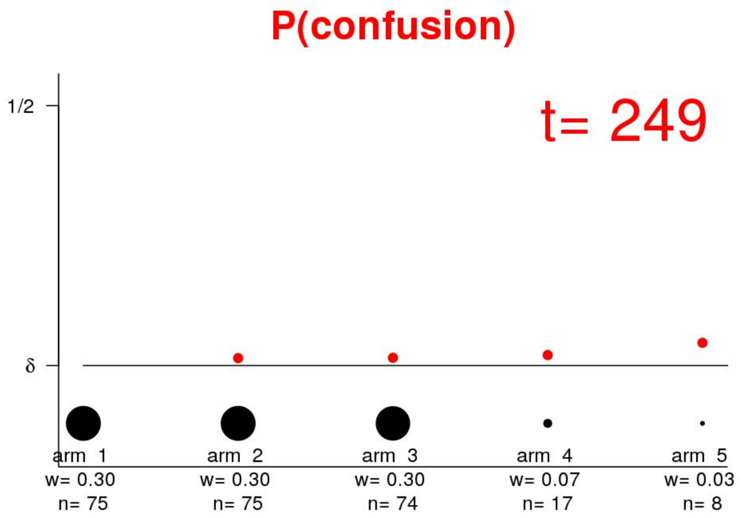
Improving: trial 2



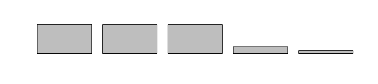


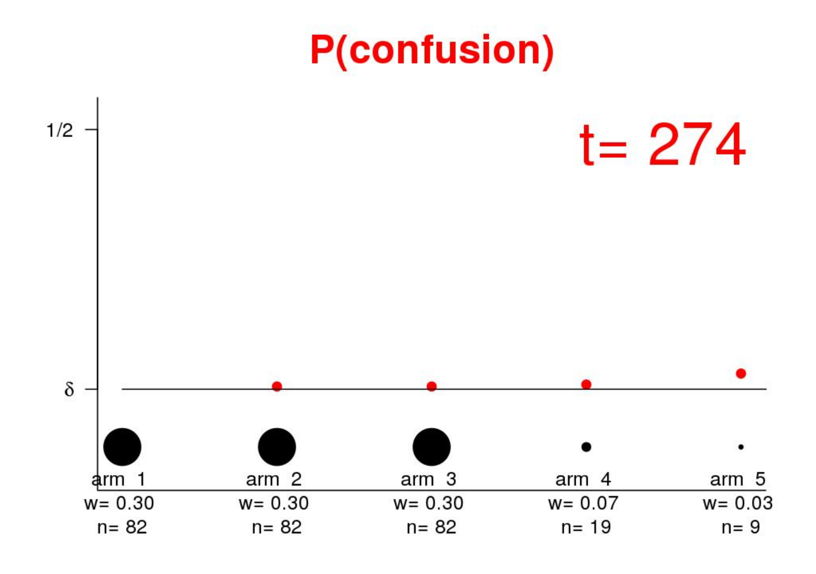
Improving: trial 2



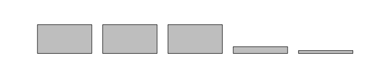


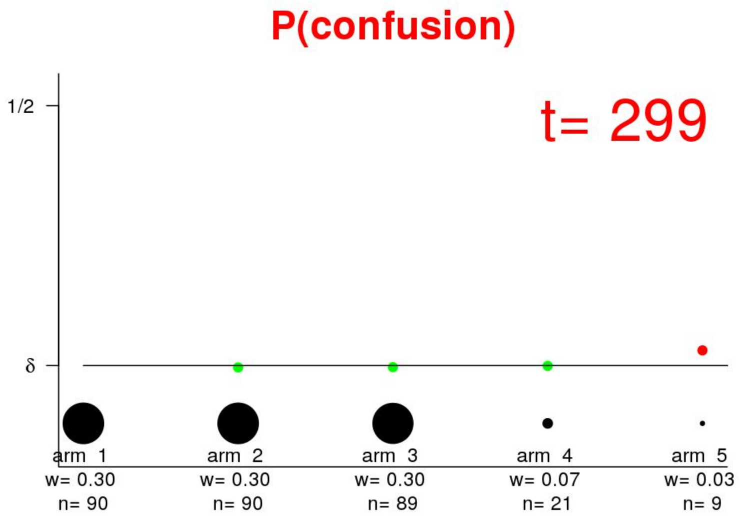
Improving: trial 2



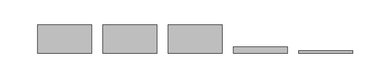


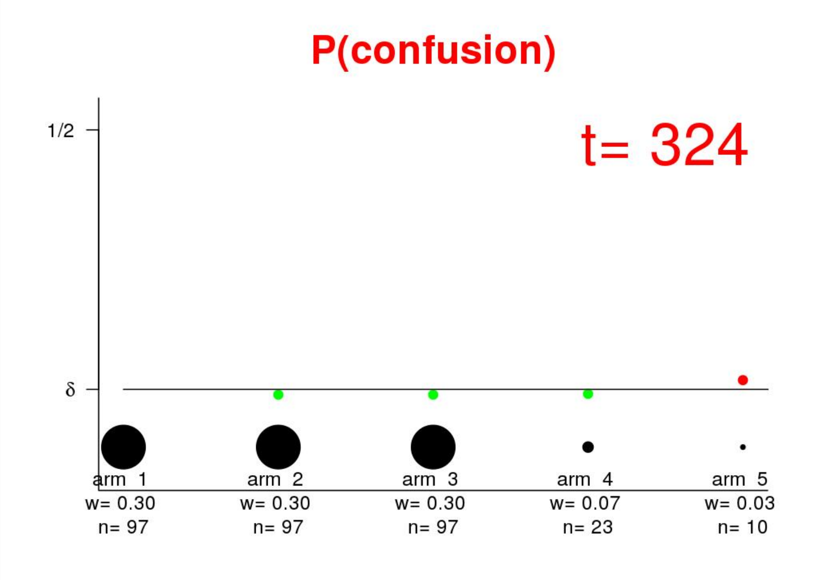
Improving: trial 2



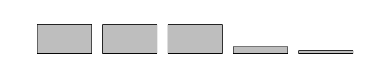


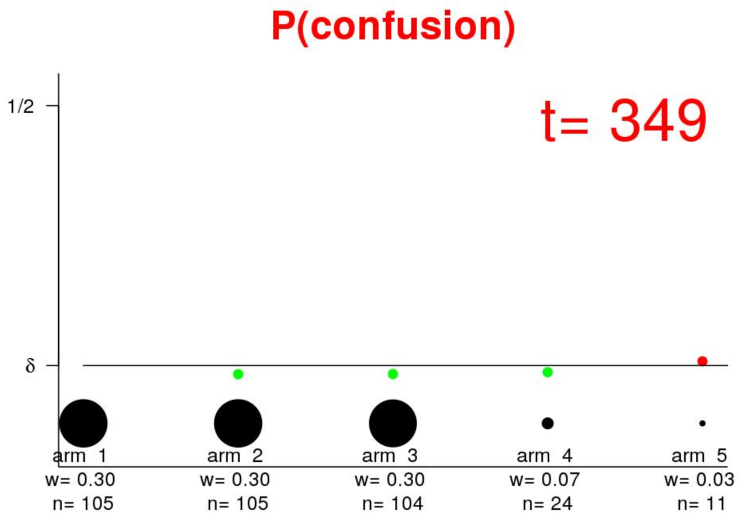
Improving: trial 2



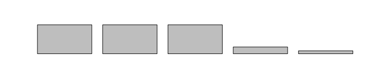


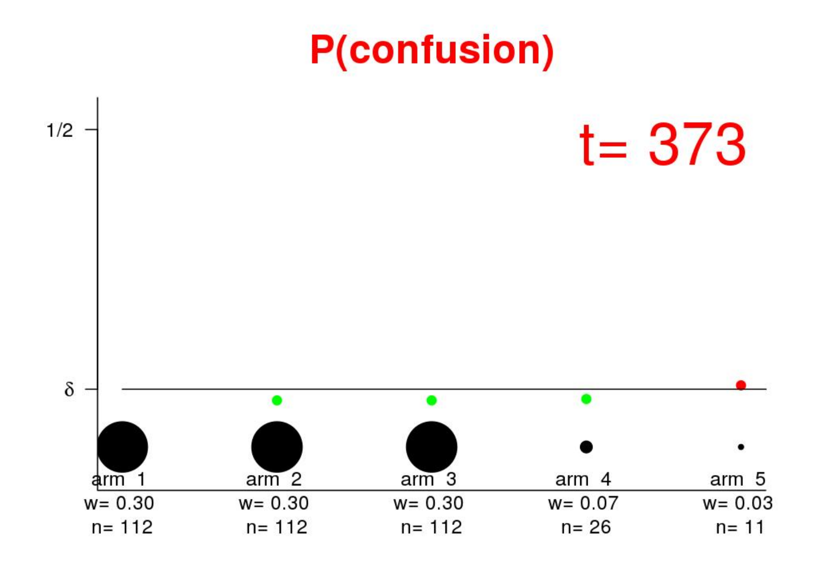
Improving: trial 2



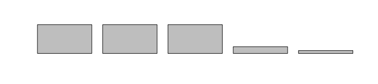


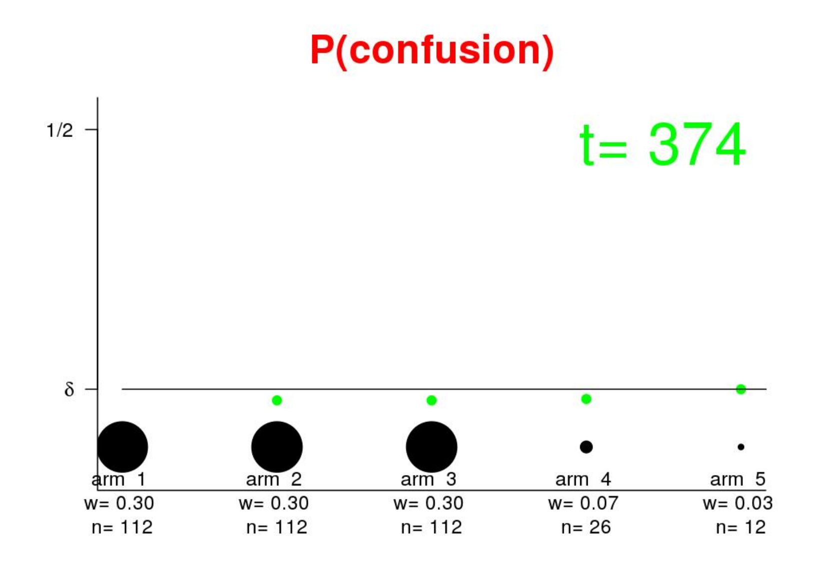
Improving: trial 2



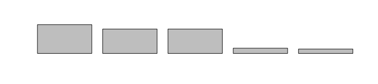


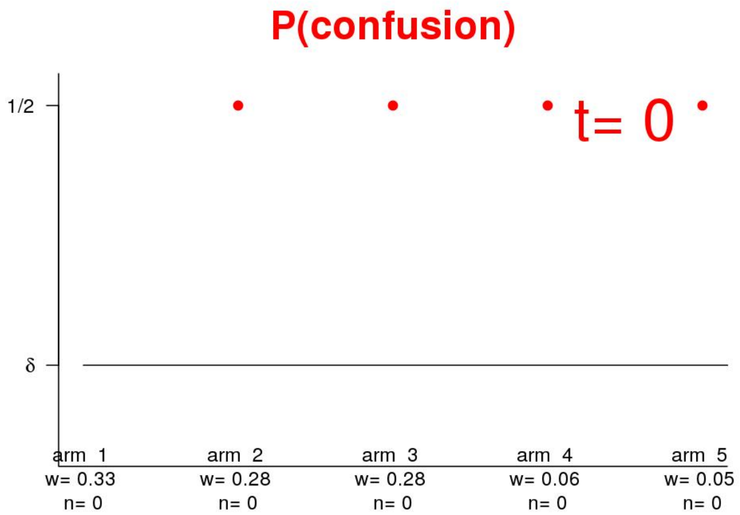
Improving: trial 2



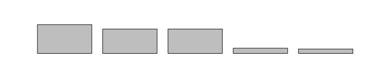


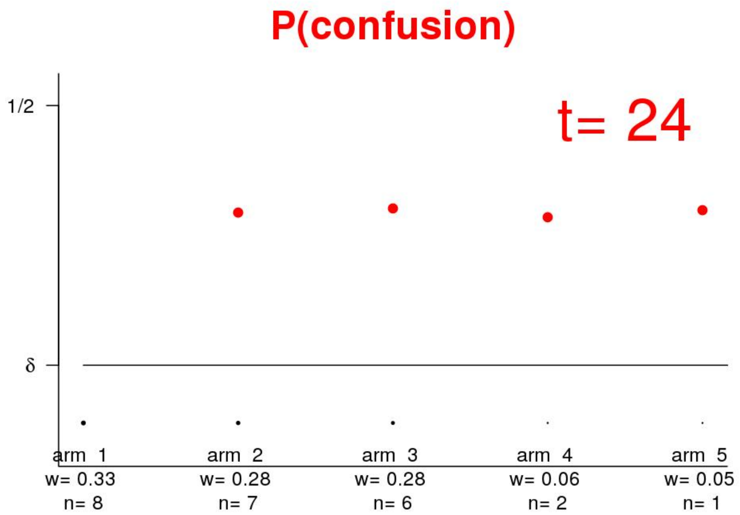
Improving: trial 3



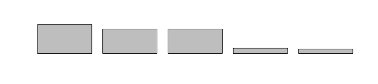


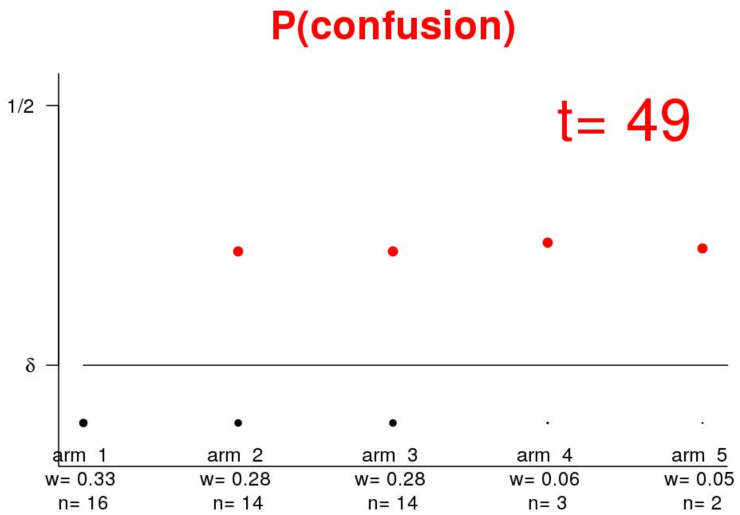
Improving: trial 3



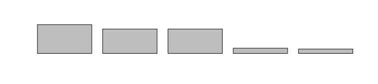


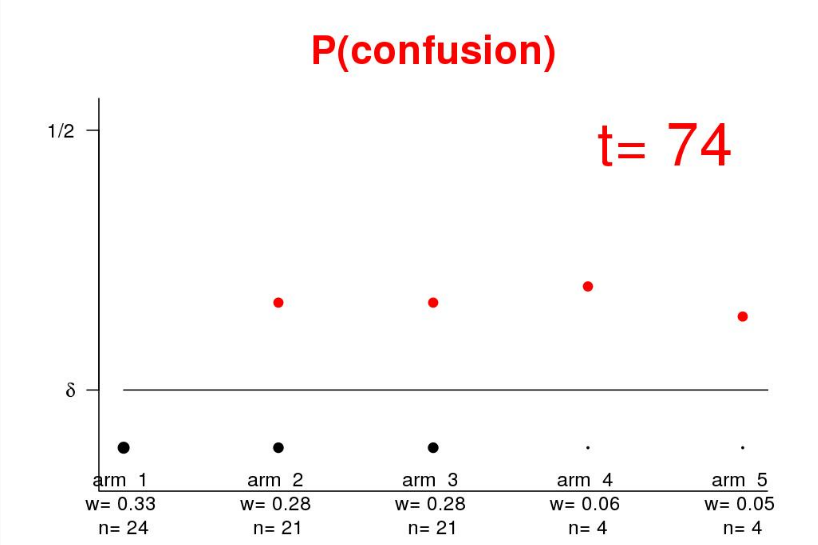
Improving: trial 3



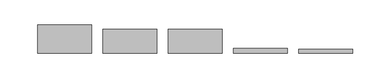


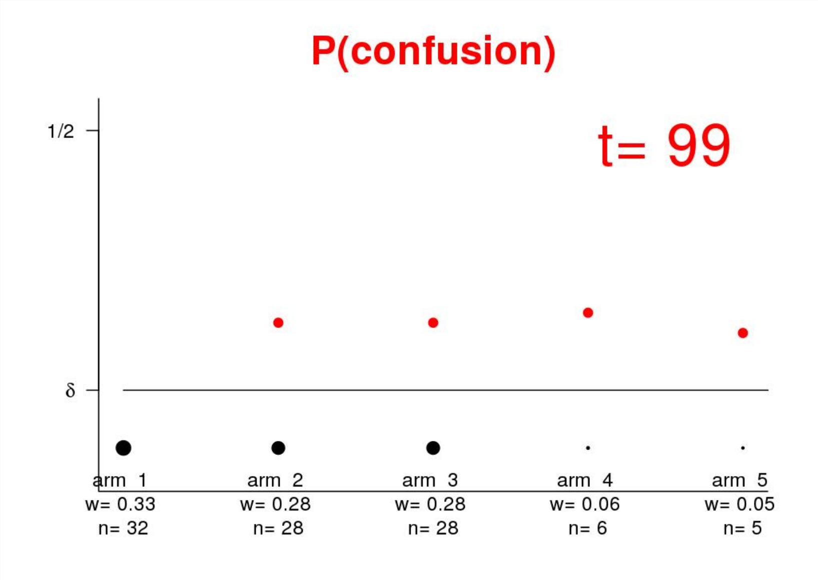
Improving: trial 3



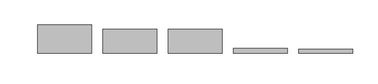


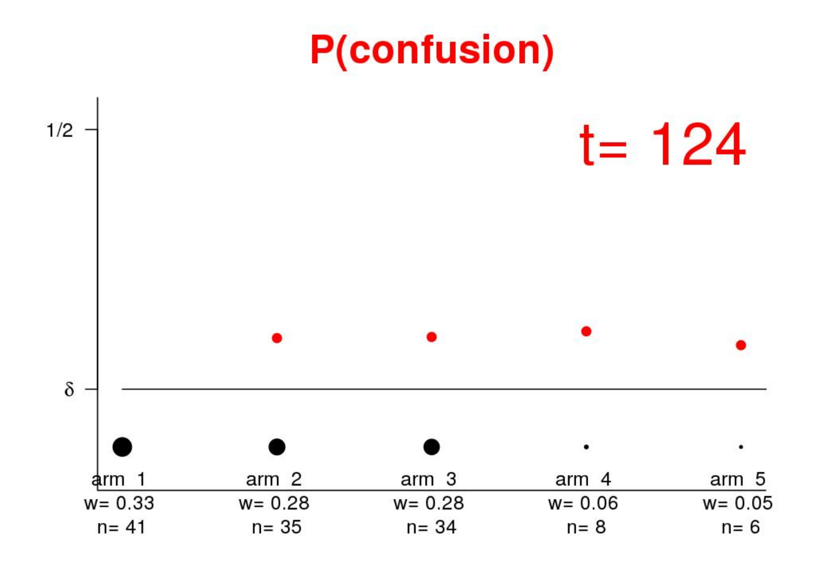
Improving: trial 3



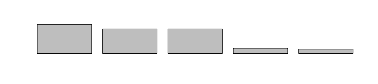


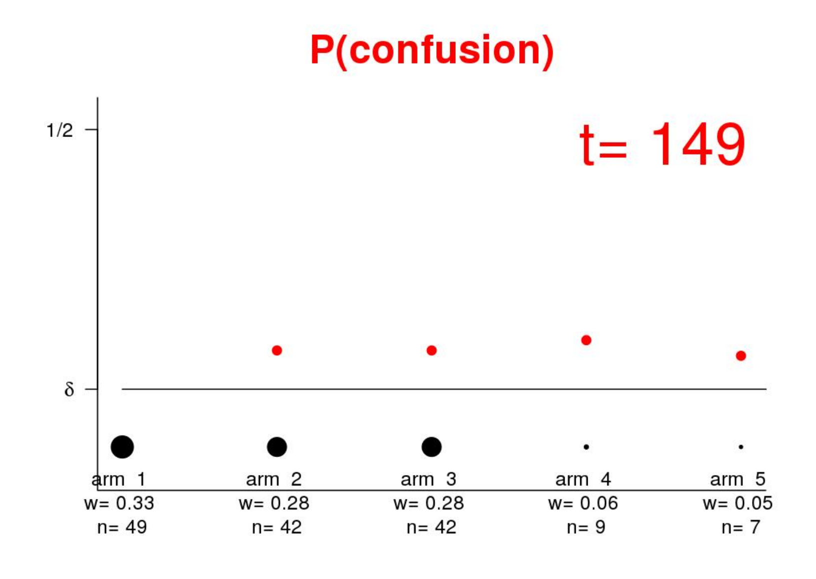
Improving: trial 3



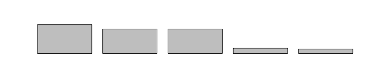


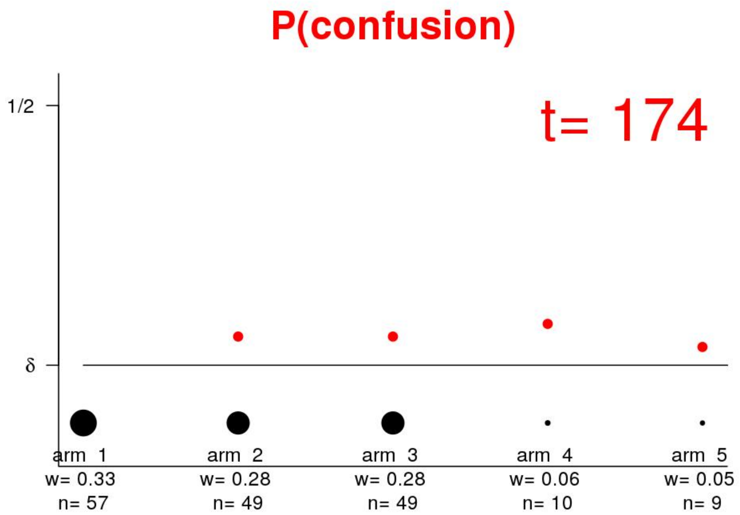
Improving: trial 3



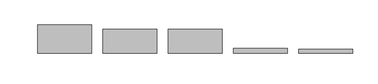


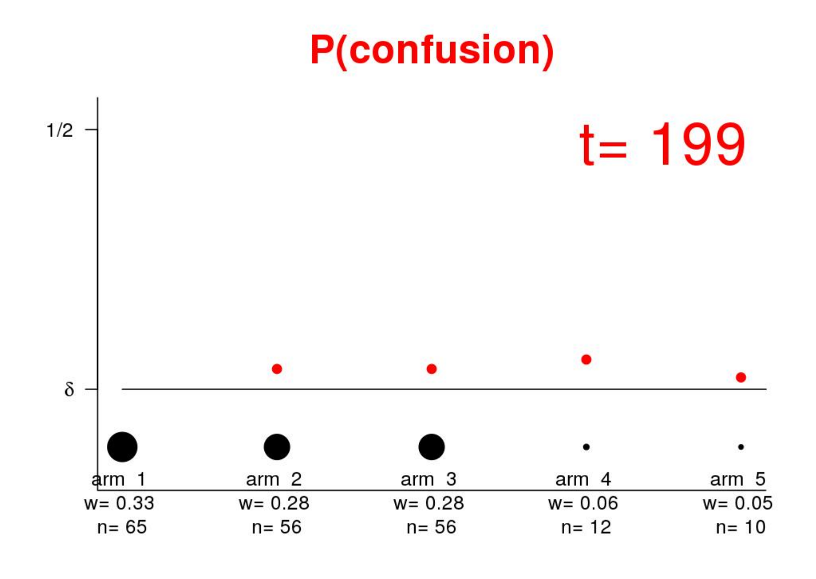
Improving: trial 3



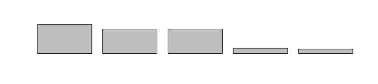


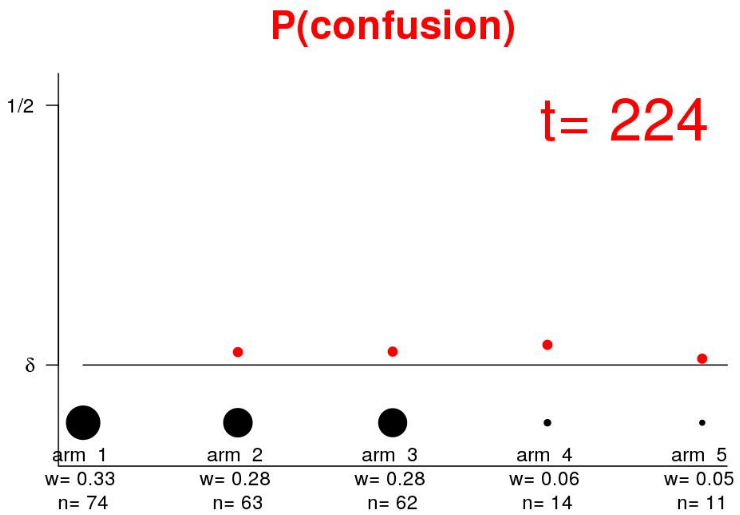
Improving: trial 3



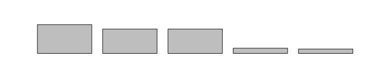


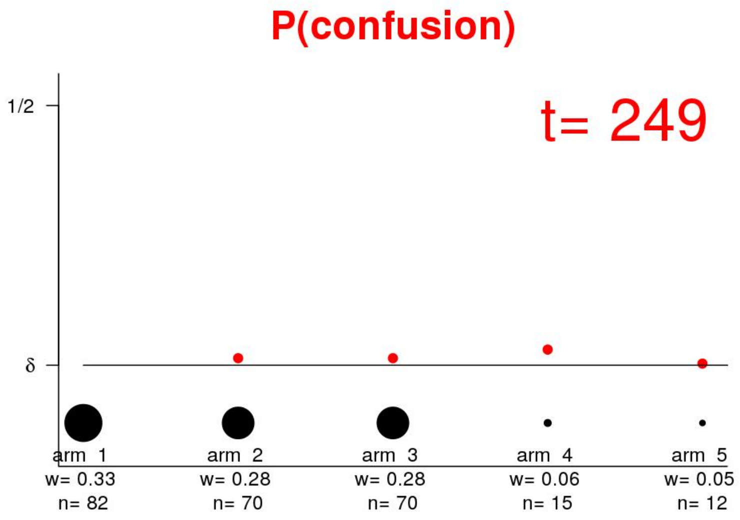
Improving: trial 3



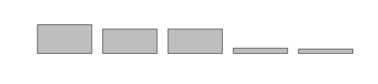


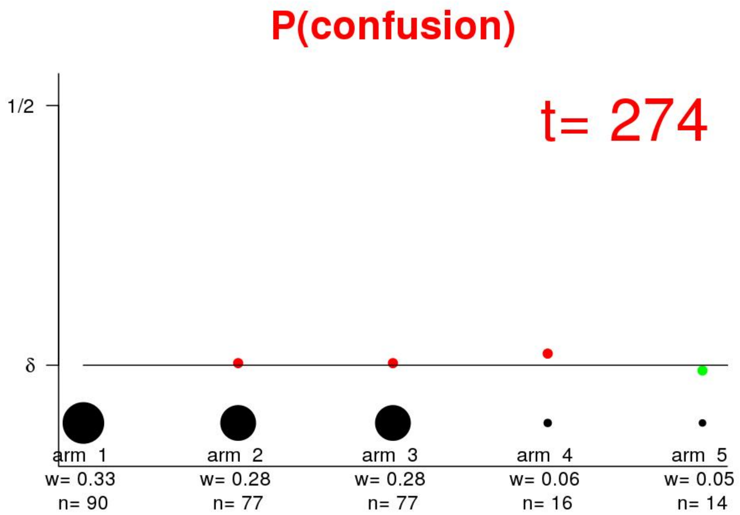
Improving: trial 3



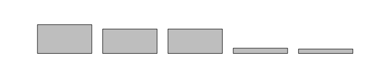


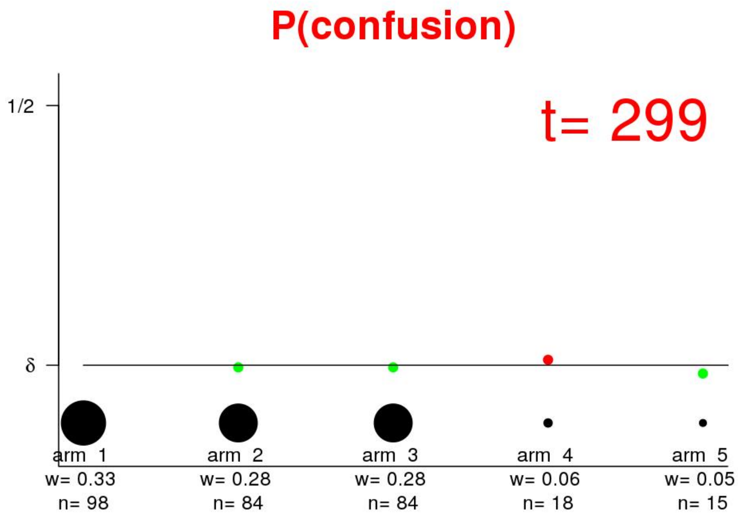
Improving: trial 3



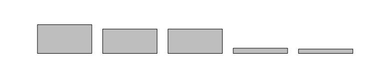


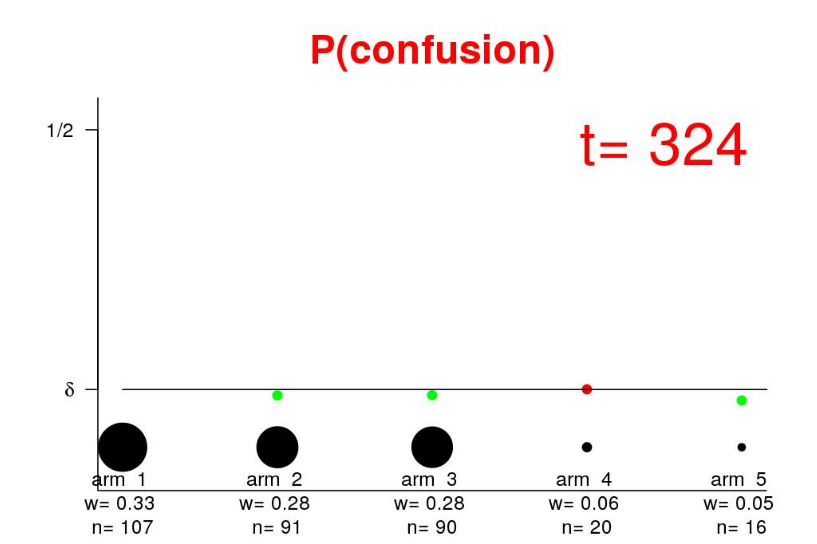
Improving: trial 3



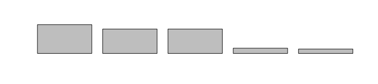


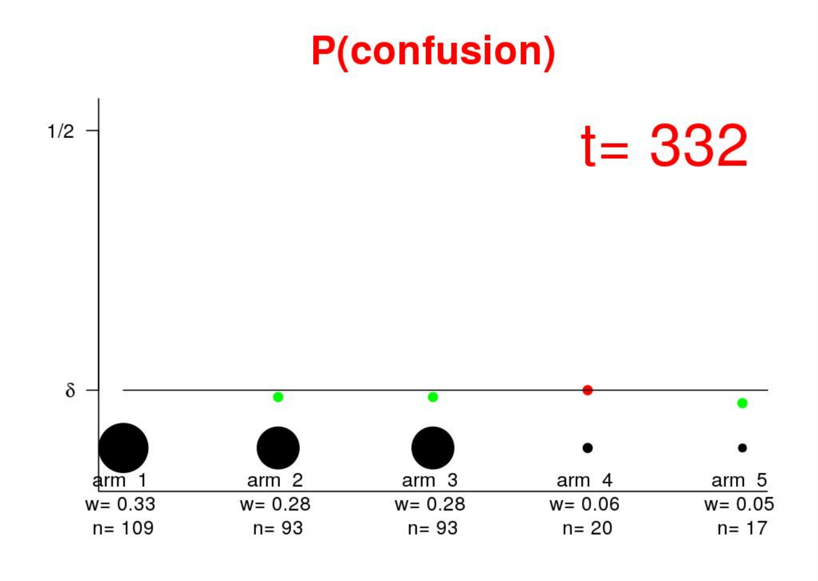
Improving: trial 3



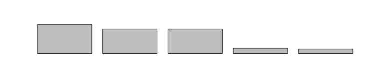


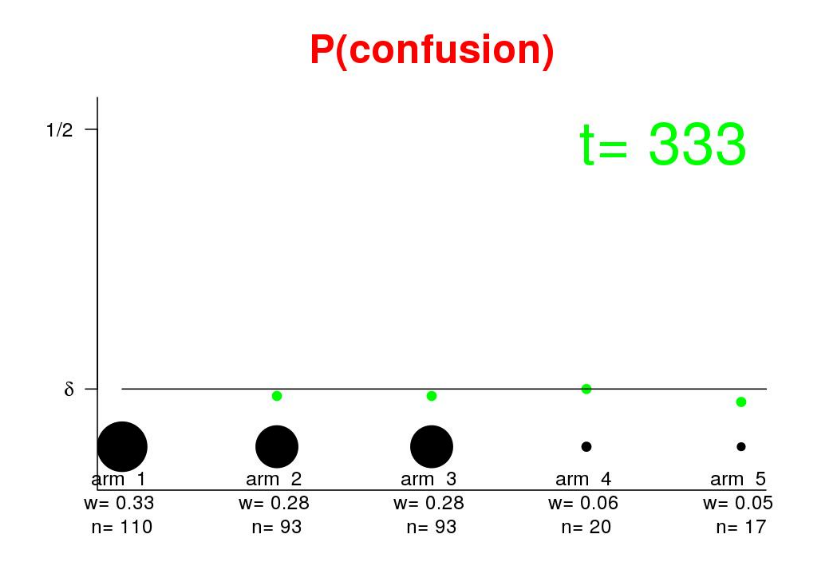
Improving: trial 3



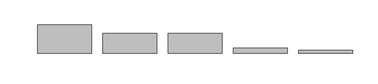


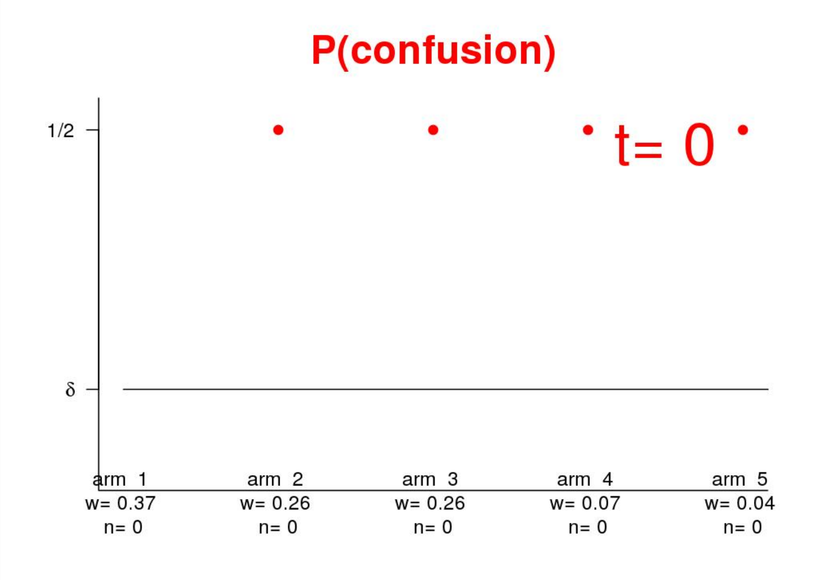
Improving: trial 3



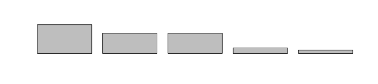


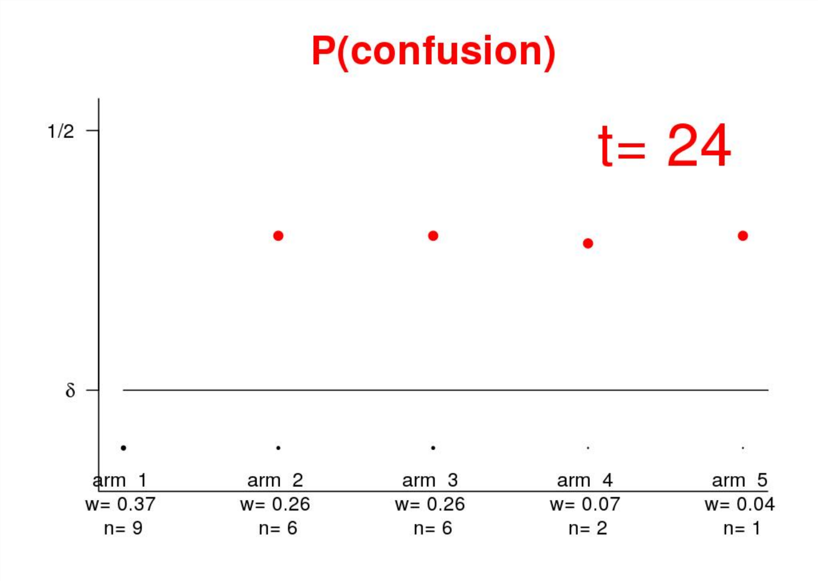
Optimal Proportions



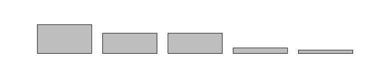


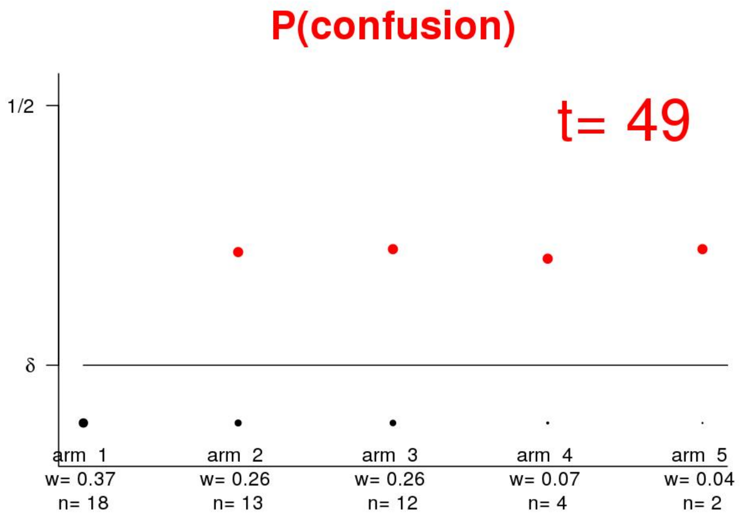
Optimal Proportions



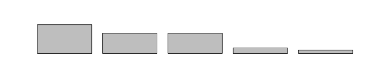


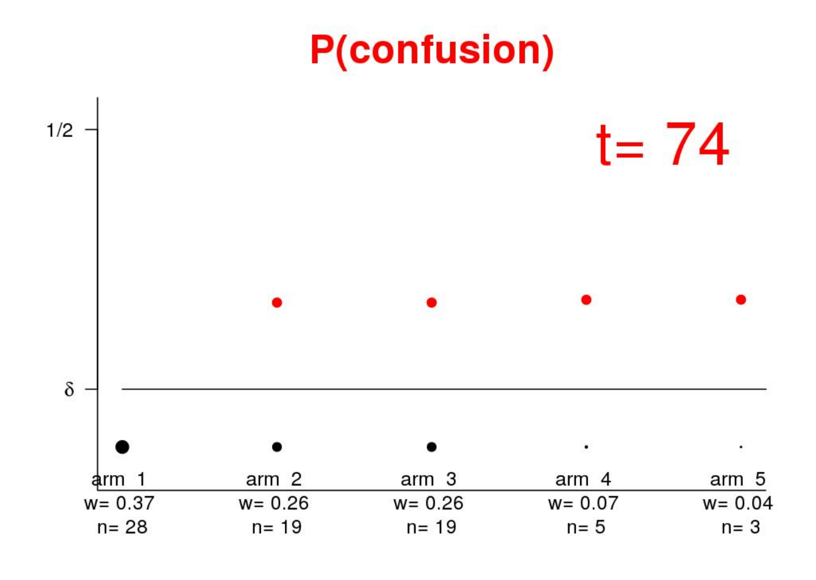
Optimal Proportions



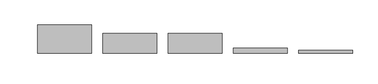


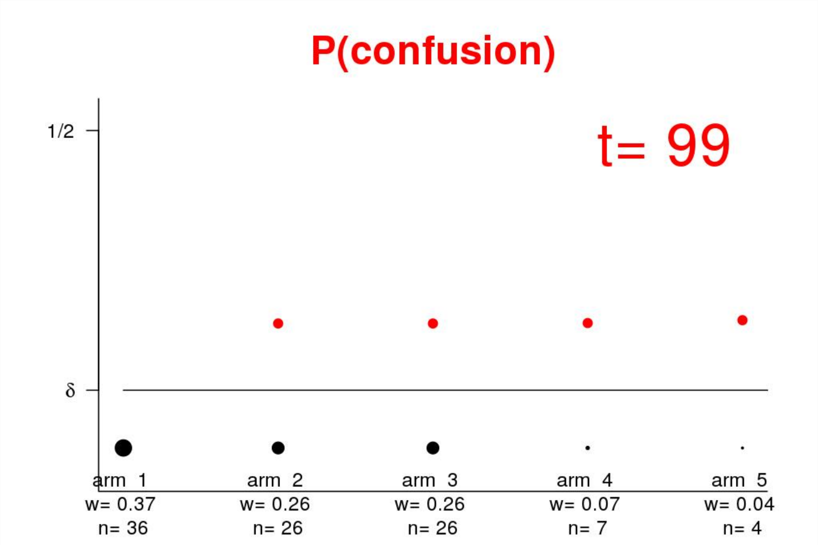
Optimal Proportions



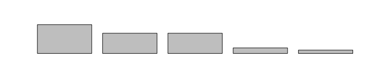


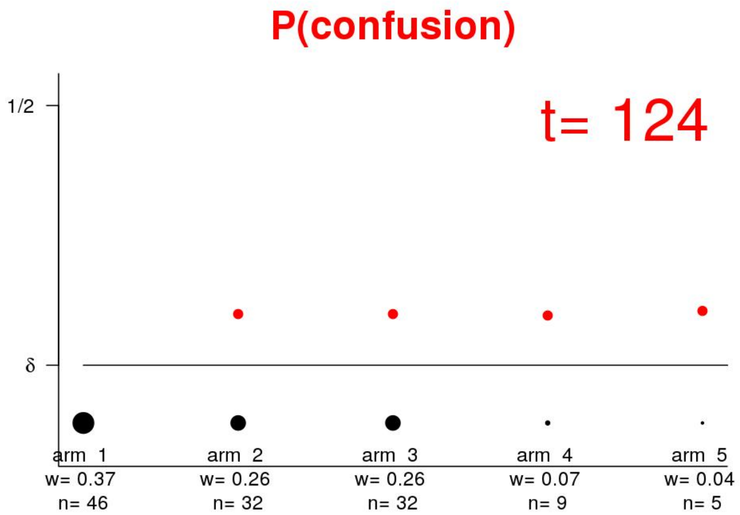
Optimal Proportions



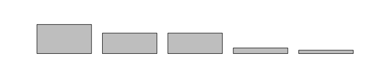


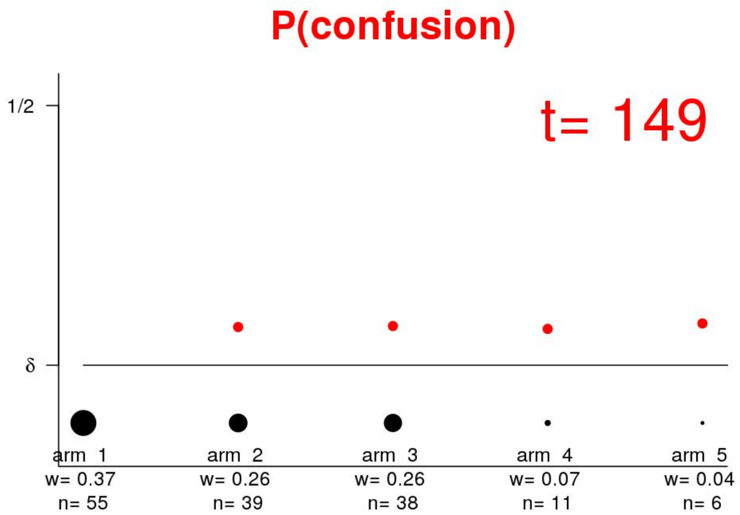
Optimal Proportions



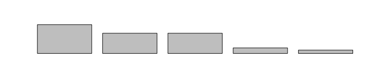


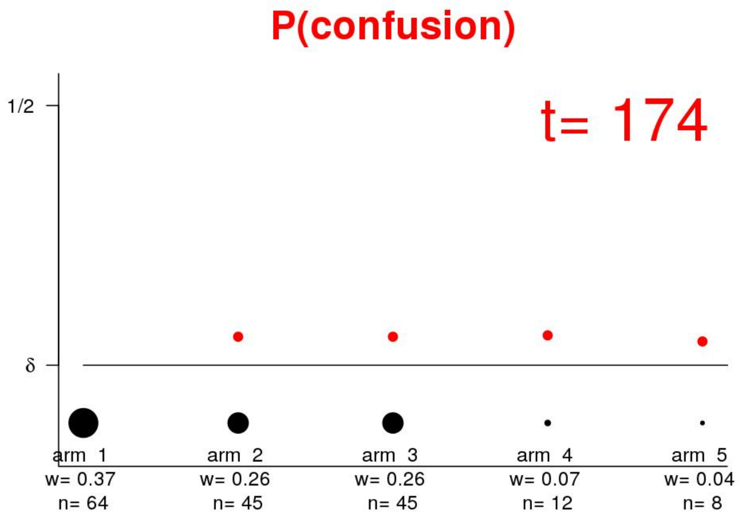
Optimal Proportions



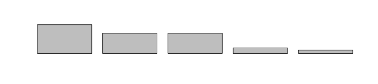


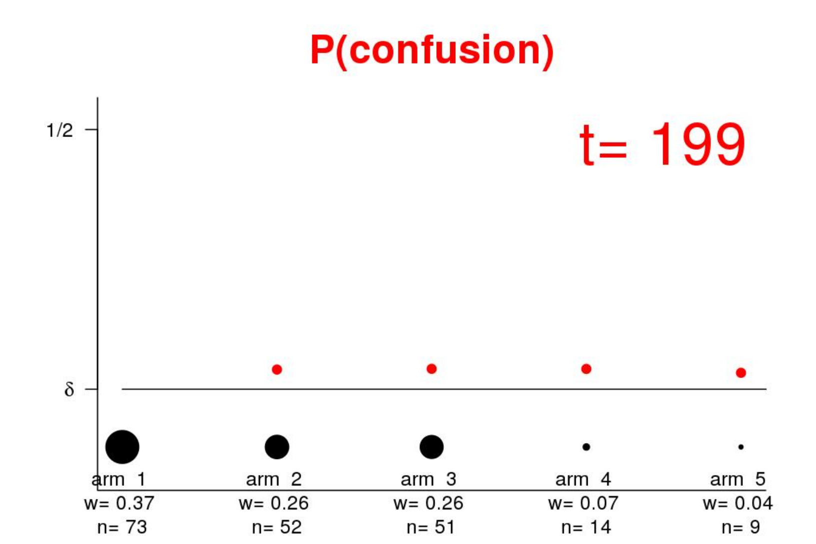
Optimal Proportions



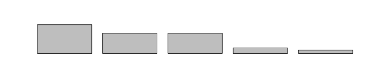


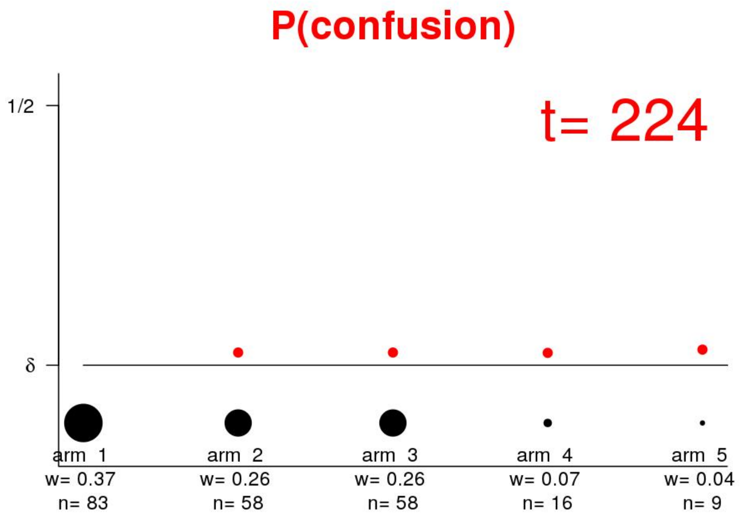
Optimal Proportions



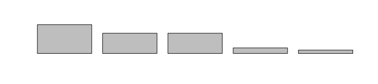


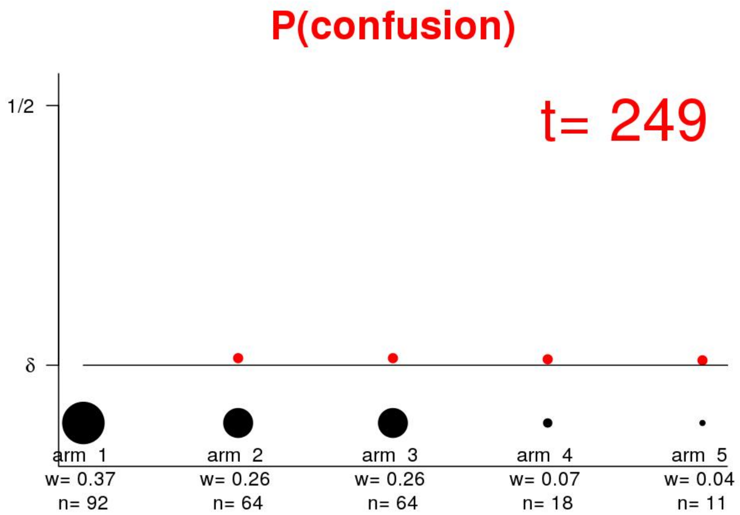
Optimal Proportions



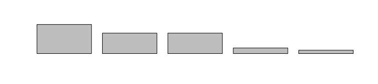


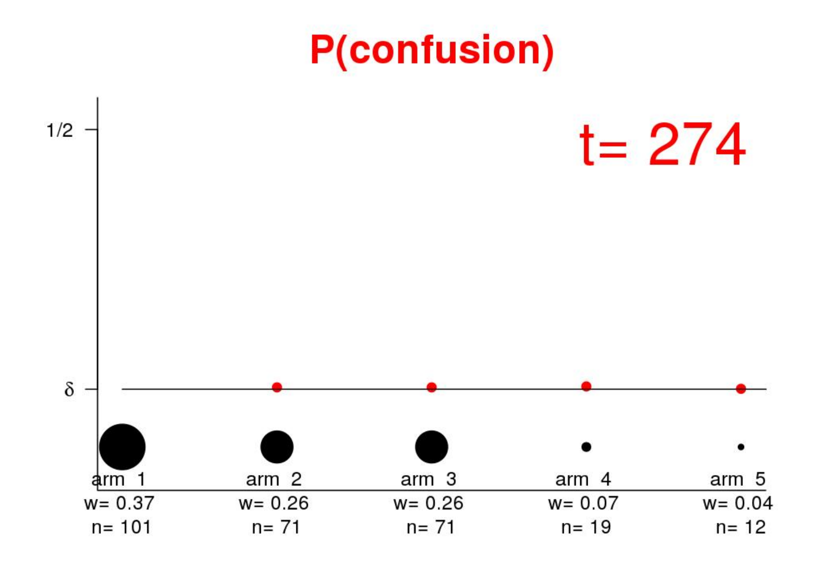
Optimal Proportions



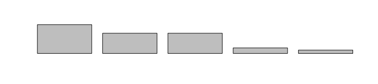


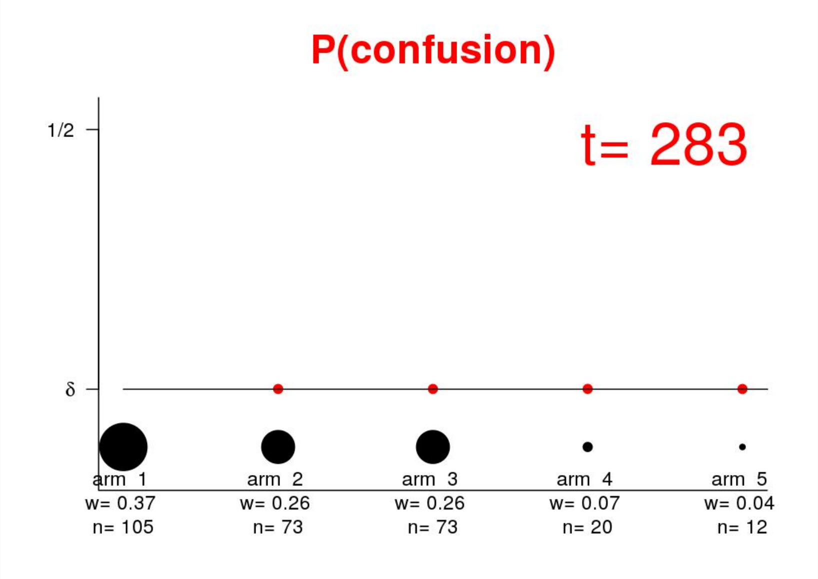
Optimal Proportions



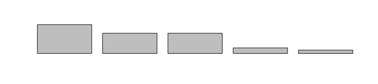


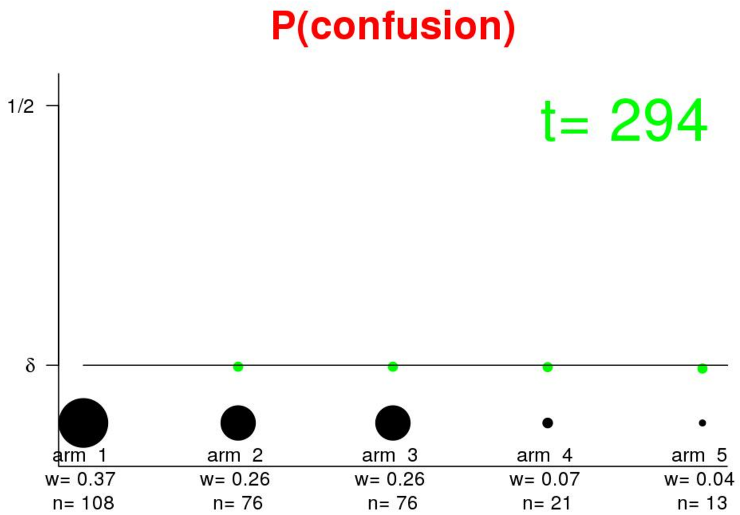
Optimal Proportions





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

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

Back to: Equalizing the Probabilities of Confusion

Most simple setting: for all $a \in \{1, \dots, K\}$, $\nu_a = \mathcal{N}(\mu_a, 1)$.

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

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

At time t, you have sampled $n_a \approx w_a\, t$ times option a and the 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:

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

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

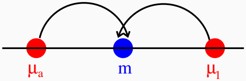
$$\leq e^{-\frac{(\mu_1 - \mu_a)^2}{2\left(1/n_1 + 1/n_a\right)}} \qquad \text{Chernoff's bound}$$

Back to: Equalizing the Probabilities of Confusion

Most simple setting: for all $a \in \{1, \ldots, K\}$, $\nu_a = \mathcal{N}(\mu_a, 1)$.

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

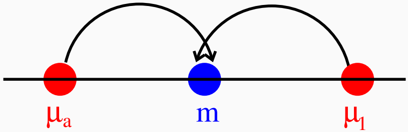
$$\begin{aligned}
\mathbb{P}\left(\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right) &= \mathbb{P}\left(\frac{\bar{X}_{a,n_a} - \mu_a - \left(\bar{X}_{1,n_1} - \mu_1\right)}{\sqrt{1/n_1 + 1/n_a}} > \frac{\mu_1 - \mu_a}{\sqrt{1/n_1 + 1/n_a}}\right) \\
&= \bar{\Phi}\left(\frac{\mu_1 - \mu_a}{\sqrt{1/n_1 + 1/n_a}}\right) \\
&\approx e^{-\frac{(\mu_1 - \mu_a)^2}{2\left(1/n_1 + 1/n_a\right)}} \qquad \text{Large Deviation Principle} \\
&= e^{-\frac{n_1(\mu_1 - m)^2}{2}} \times e^{-\frac{n_a(\mu_a - m)^2}{2}} \qquad \text{with} \quad m = \frac{n_1\mu_1 + n_1\mu_a}{n_1 + n_a} \\
&\approx \mathbb{P}\left(\bar{X}_{1,n_1} < m\right) \times \mathbb{P}\left(\bar{X}_{a,n_a} \geq m\right) \\
&= \max_{\mu_a < m < \mu_1} \mathbb{P}\left(\bar{X}_{1,n_1} < m\right) \times \mathbb{P}\left(\bar{X}_{a,n_a} \geq m\right) \qquad \text{(cf. Sanov)}
\end{aligned}$$



Entropic Method for Large Deviations Lower Bounds

Let $d(\mu, \mu') = \mathrm{KL}\big(\mathcal{N}(\mu, 1), \mathcal{N}(\mu', 1)\big) = \frac{(x-y)^2}{2}$,
$\mathcal{KL}(Y, Z) = \mathrm{KL}\big(\mathcal{L}(Y), \mathcal{L}(Z)\big)$, $\quad \epsilon > 0$, $\quad \mu_a \le m \le \mu_1$ and

- $X_{1,1}, \dots, X_{1,n_1} \overset{iid}{\sim} \mathcal{N}(\mu_1, 1)$
- $X_{a,1}, \dots, X_{a,n_a} \overset{iid}{\sim} \mathcal{N}(\mu_a, 1)$
- $X'_{1,1}, \dots, X'_{1,n_1} \overset{iid}{\sim} \mathcal{N}(m - \epsilon, 1)$
- $X'_{a,1}, \dots, X'_{a,n_a} \overset{iid}{\sim} \mathcal{N}(m + \epsilon, 1)$



$$n_1\, d(m-\epsilon, \mu_1) + n_a\, d(m+\epsilon, \mu_a) = \mathcal{KL}\big((X'_{a,i})_{a,i}, (X_{a,i})_{a,i}\big)$$

$\mathrm{KL}(P \otimes P', Q \otimes Q') = \mathrm{KL}(P, Q) + \mathrm{KL}(P', Q')$

$$\ge \mathcal{KL}\Big(\mathbb{1}\left\{\bar{X}'_{a,n_a} > \bar{X}'_{1,n_1}\right\}, \mathbb{1}\left\{\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right\}\Big)$$

contraction of entropy
= data-processing inequality

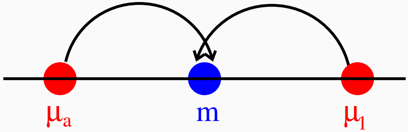
$$= \mathrm{kl}\Big(\mathbb{P}\left(\bar{X}'_{a,n_a} > \bar{X}'_{1,n_1}\right), \mathbb{P}\left(\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right)\Big)$$

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

$$\ge \mathbb{P}\left(\bar{X}'_{a,n_a} > \bar{X}'_{1,n_1}\right) \log \frac{1}{\mathbb{P}\Big(\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\Big)} - \log(2)$$

$\mathrm{kl}(p, q) \ge p \log \frac{1}{q} - \log 2$

Entropic Method for Large Deviations Lower Bounds



$$n_1\, d(m-\epsilon,\mu_1) + n_a\, d(m+\epsilon,\mu_a) = \mathcal{KL}\big((X'_{a,i})_{a,i},(X_{a,i})_{a,i}\big)$$

$\mathrm{KL}(P\otimes P', Q\otimes Q') = \mathrm{KL}(P,Q) + \mathrm{KL}(P',Q')$

$$\geq \mathcal{KL}\left(\mathbb{1}\left\{\bar{X}'_{a,n_a} > \bar{X}'_{1,n_1}\right\},\ \mathbb{1}\left\{\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right\}\right)$$

contraction of entropy
= data-processing inequality

$$= \mathrm{kl}\left(\mathbb{P}\left(\bar{X}'_{a,n_a} > \bar{X}'_{1,n_1}\right),\ \mathbb{P}\left(\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right)\right)$$

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

$$\geq \mathbb{P}\left(\bar{X}'_{a,n_a} > \bar{X}'_{1,n_1}\right)\log\frac{1}{\mathbb{P}\left(\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right)} - \log(2)$$

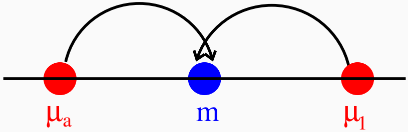
$\mathrm{kl}(p,q) \geq p\log\frac{1}{q} - \log 2$

$$\mathbb{P}\left(\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right) \geq \max_{\mu_1\leq m\leq\mu_a} \exp\left(-\frac{n_1\, d(m-\epsilon,\mu_1) + n_a\, d(m+\epsilon,\mu_a) + \log(2)}{1-e^{-(n_1+n_a)\epsilon^2/2}}\right)$$

$$= \exp\left(-\frac{\frac{(\mu_1-\mu_a+\epsilon)^2}{1/n_1+1/n_a} + \log(2)}{2\left(1-e^{-(n_1+n_a)\epsilon^2/2}\right)}\right)$$

$m = \frac{n_1\mu_1 + n_a\mu_a}{n_1+n_a}$

Entropic Method for Large Deviations Lower Bounds



$$n_1\, d(m-\epsilon,\mu_1) + n_a\, d(m+\epsilon,\mu_a) = \mathcal{KL}\Big((X'_{a,i})_{a,i}, (X_{a,i})_{a,i}\Big)$$

$\mathrm{KL}(P\otimes P', Q\otimes Q') = \mathrm{KL}(P,Q) + \mathrm{KL}(P',Q')$

$$\geq \mathcal{KL}\left(\mathbb{1}\left\{\bar{X}'_{a,n_a} > \bar{X}'_{1,n_1}\right\}, \mathbb{1}\left\{\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right\}\right)$$

contraction of entropy
= data-processing inequality

$$= \mathrm{kl}\left(\mathbb{P}\left(\bar{X}'_{a,n_a} > \bar{X}'_{1,n_1}\right), \mathbb{P}\left(\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right)\right)$$

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

$$\geq \mathbb{P}\left(\bar{X}'_{a,n_a} > \bar{X}'_{1,n_1}\right)\log\frac{1}{\mathbb{P}\left(\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right)} - \log(2)$$

$\mathrm{kl}(p,q) \geq p\log\frac{1}{q} - \log 2$

$$\implies T\Big(w_1\, d(m-\epsilon,\mu_1) + w_a\, d(m+\epsilon,\mu_a)\Big) \gtrsim \log\frac{1}{\mathbb{P}\left(\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right)}$$

➜ if you want to have $\mathbb{P}\left(\bar{X}_{a,n_a} > \bar{X}_{1,n_1}\right) \leq \delta$ then you need

$$T \gtrsim \frac{\log(1/\delta)}{w_1 d(m,\mu_1) + w_a\, d(m,\mu_a)}$$

Lower Bound

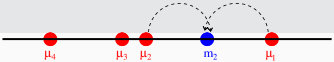
Lower-Bounding the Sample Complexity

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

Uniform δ-PAC Constraint **[Kaufmann, Cappé, G. '15]**

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

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



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

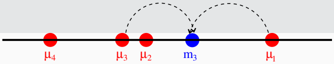
$$\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]\, d(\mu_1, m_2 - \epsilon) + \mathbb{E}_{\boldsymbol{\mu}}[N_2(\tau_\delta)]\, d(\mu_2, m_2 + \epsilon) \quad \geq \quad \mathrm{kl}(\delta, 1-\delta)$$

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

Uniform δ-PAC Constraint [Kaufmann, Cappé, G. '15]

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

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



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

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

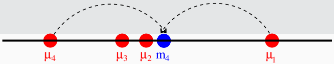
Lower-Bounding the Sample Complexity

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

Uniform δ-PAC Constraint **[Kaufmann, Cappé, G. '15]**

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

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



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

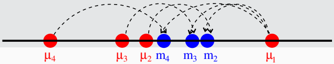
$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]\, d(\mu_1, m_2 - \epsilon) + \mathbb{E}_{\boldsymbol{\mu}}[N_2(\tau_\delta)]\, d(\mu_2, m_2 + \epsilon) &\geq \mathrm{kl}(\delta, 1-\delta) \\
\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]\, d(\mu_1, m_3 - \epsilon) + \mathbb{E}_{\boldsymbol{\mu}}[N_3(\tau_\delta)]\,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

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

Uniform δ-PAC Constraint [Kaufmann, Cappé, G. '15]

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

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



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

For any δ-PAC 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

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

➜ they do not depend on δ

PAC-BAI as a Game

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

- 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

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

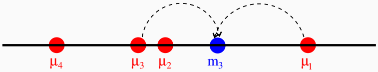
- $T^*(\boldsymbol{\mu}) \, \mathrm{kl}(\delta, 1-\delta)$ = value of the game
  $\boldsymbol{w}^*$ = optimal action for the statistician

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

$$T w_1\, d(\mu_1, \lambda_a - \epsilon) + T w_a\, d(\mu_a, \lambda_a + \epsilon) \geq \mathrm{kl}(\delta, 1-\delta)$$

  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, G. '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

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,

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

Sequential Generalized Likelihood Test

High values of the Generalized Likelihood Ratio

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

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

$-Z_{b,a}(t)$ 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:

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

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

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

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

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:

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

Chernoff stopping rule [Chernoff '59]

Two other possible interpretations of the stopping rule:

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

stop when $Z(t) = t \, F\left(\frac{N_a(t)}{t}, \hat{\mu}(t)\right) \geq \beta(t, \delta)$ 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)

Asymptotic Optimality of the T&S strategy

Theorem

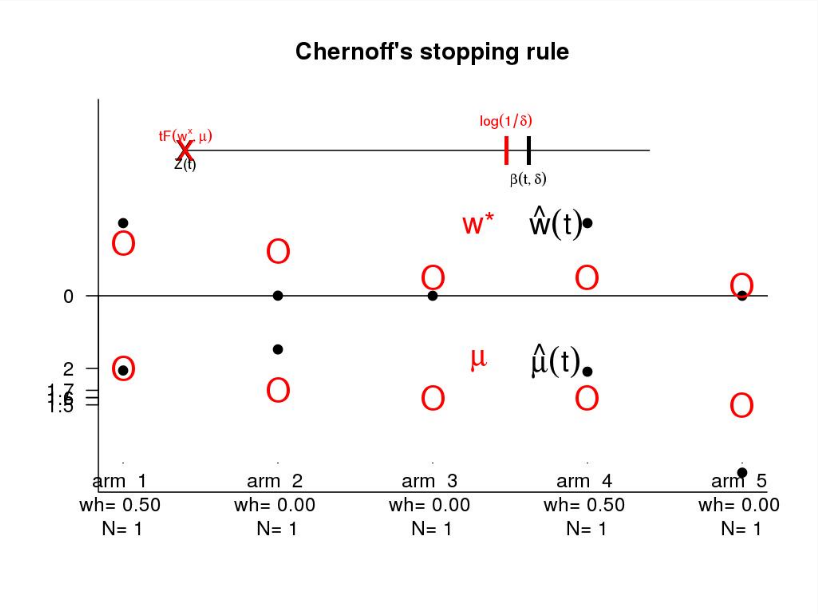
The Track-and-Stop strategy, that uses

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

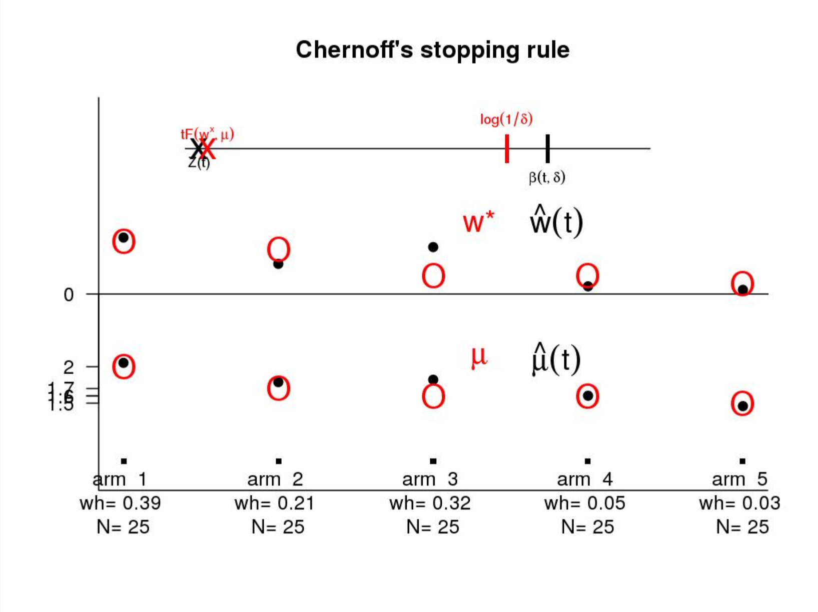
is δ-PAC for every $\delta \in (0,1)$ and satisfies

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

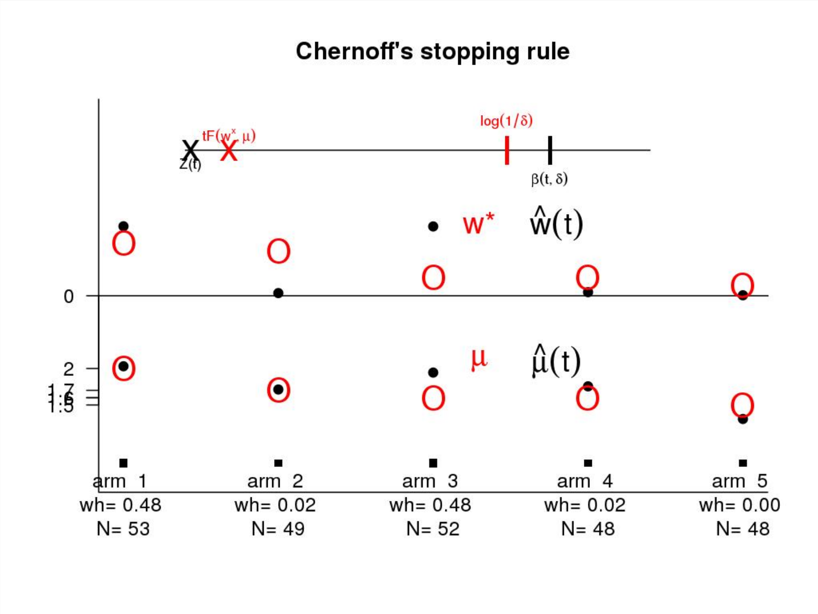
Why is the T&S Strategy asymptotically Optimal?



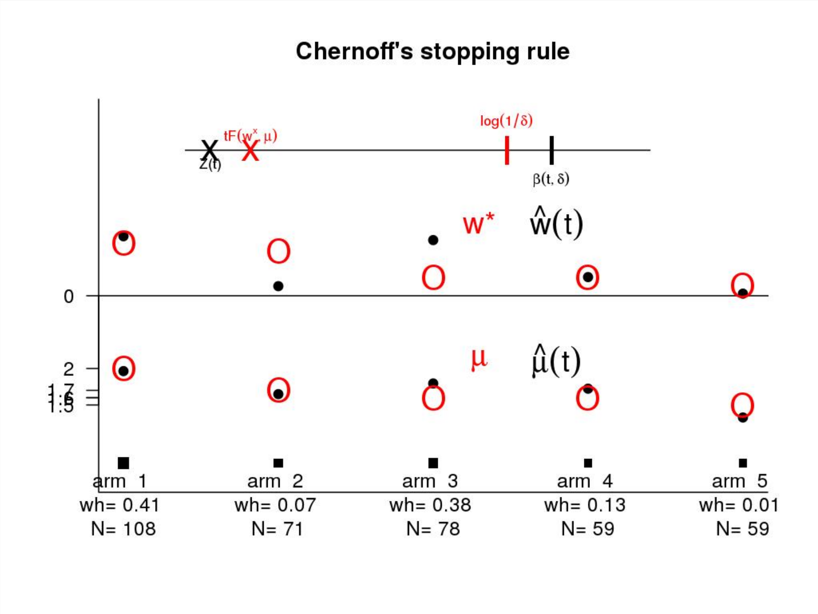
Why is the T&S Strategy asymptotically Optimal?



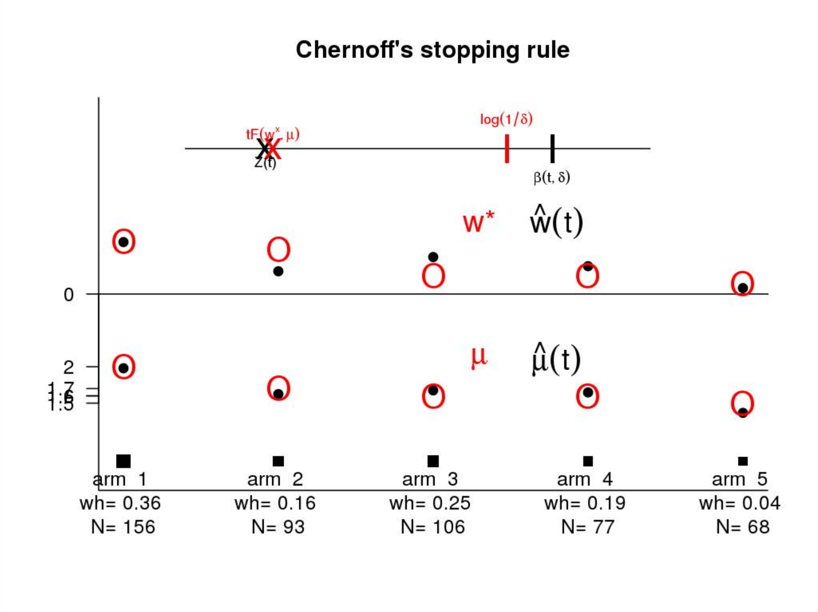
Why is the T&S Strategy asymptotically Optimal?



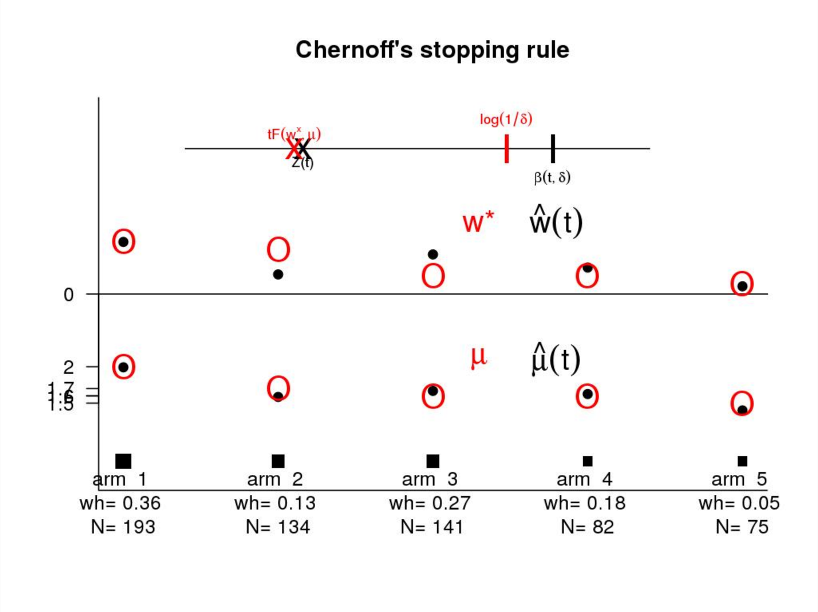
Why is the T&S Strategy asymptotically Optimal?



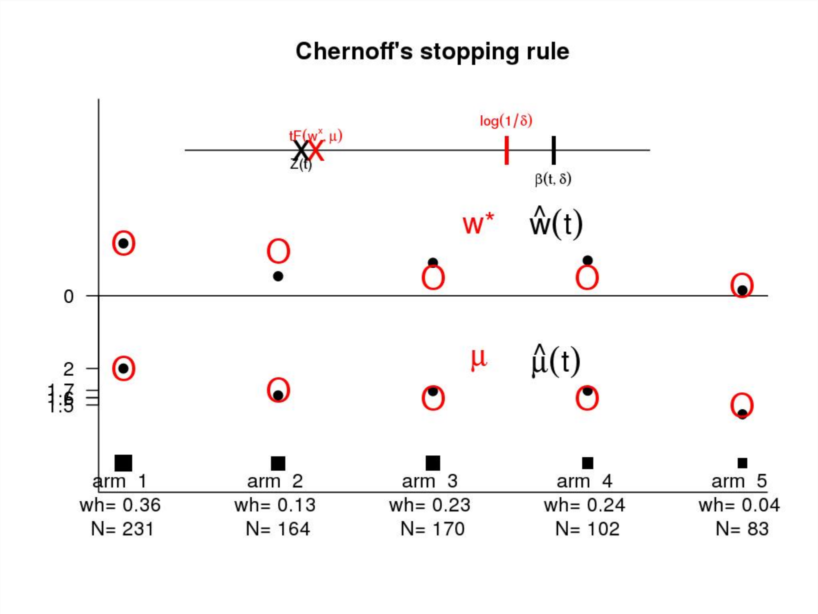
Why is the T&S Strategy asymptotically Optimal?



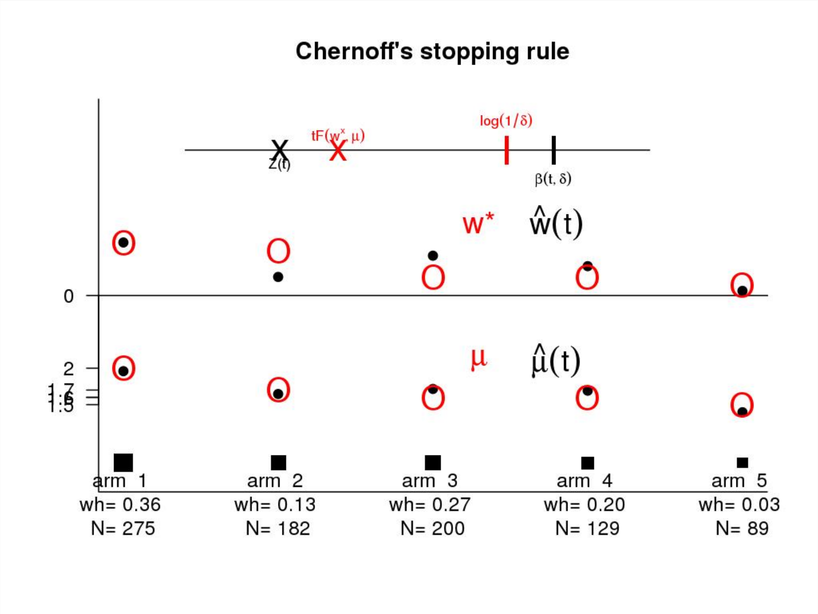
Why is the T&S Strategy asymptotically Optimal?



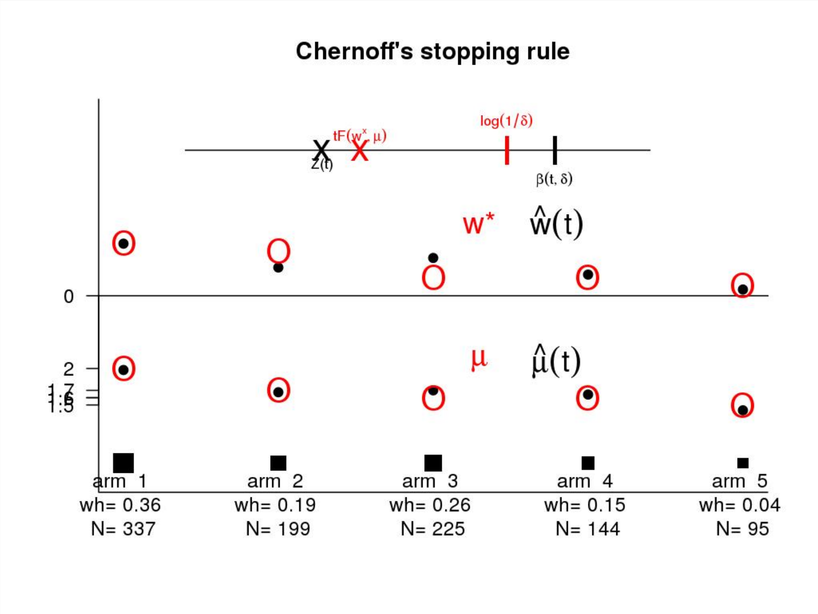
Why is the T&S Strategy asymptotically Optimal?



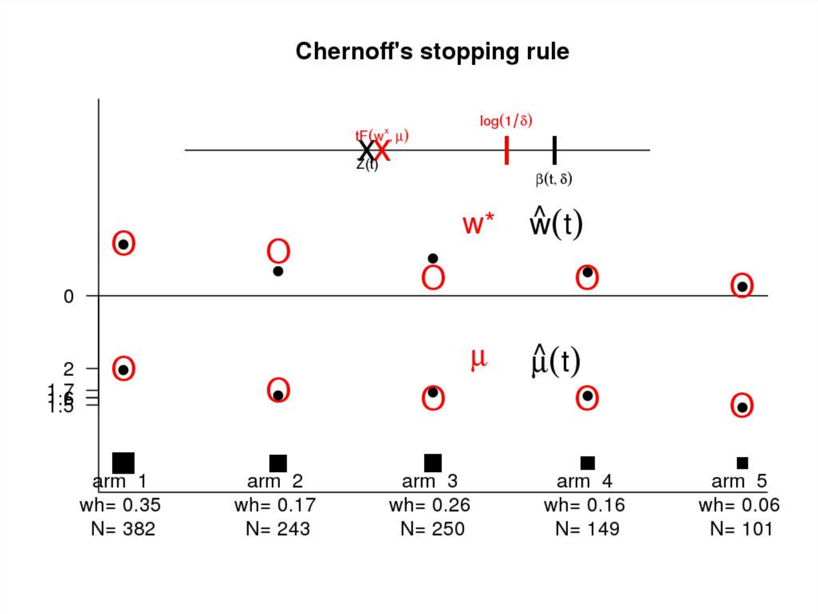
Why is the T&S Strategy asymptotically Optimal?



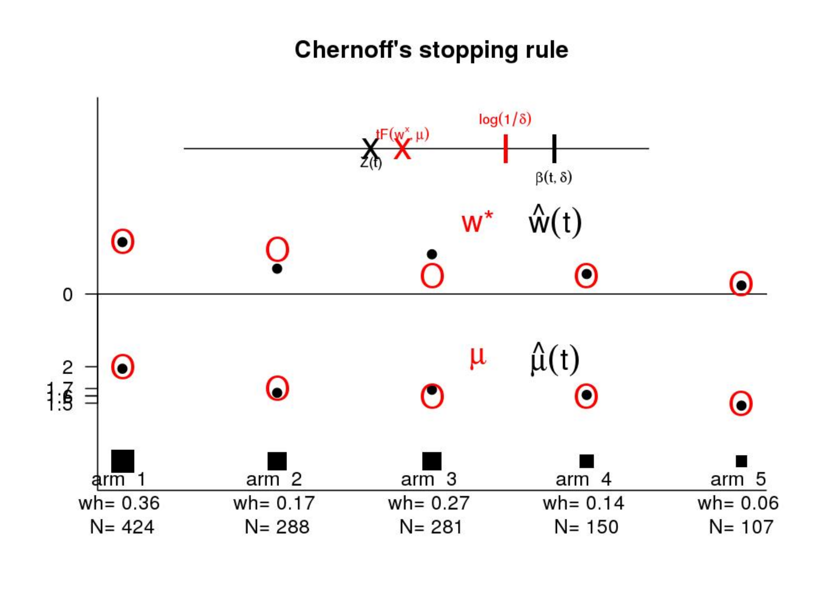
Why is the T&S Strategy asymptotically Optimal?



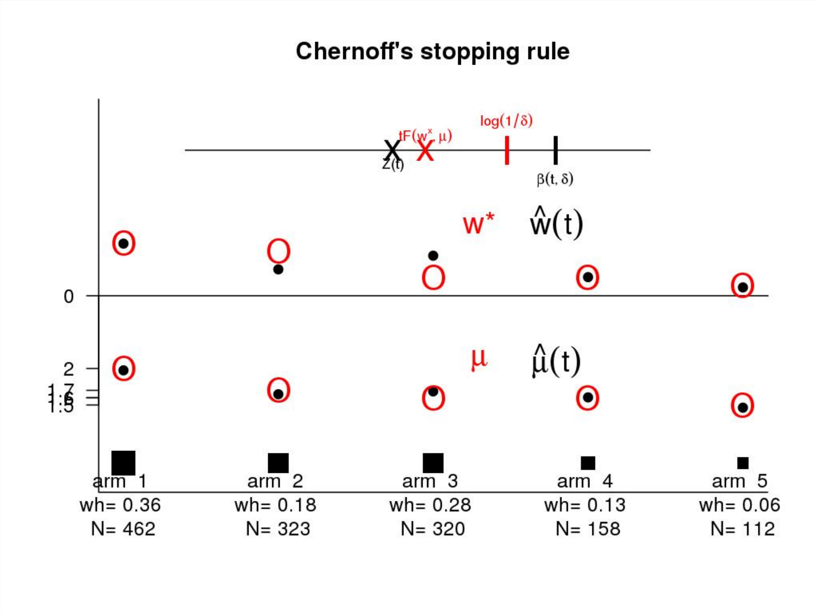
Why is the T&S Strategy asymptotically Optimal?



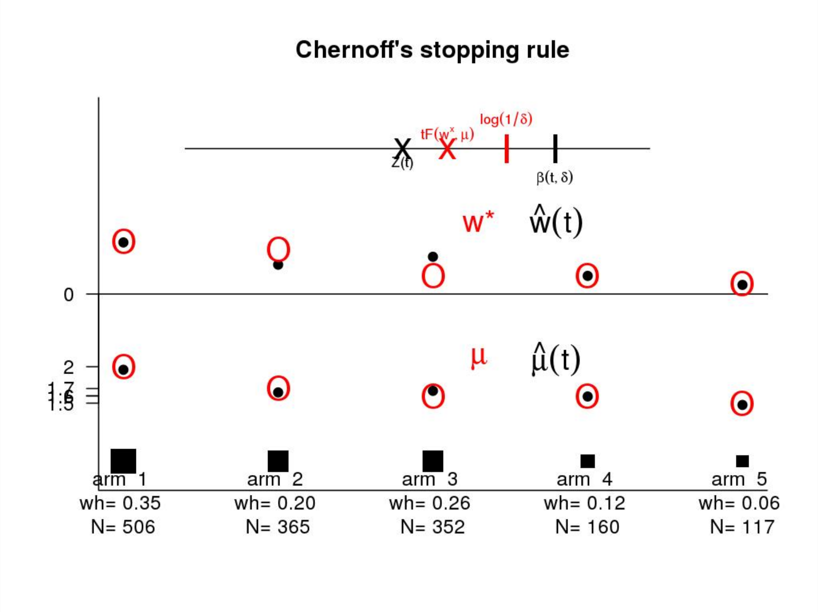
Why is the T&S Strategy asymptotically Optimal?



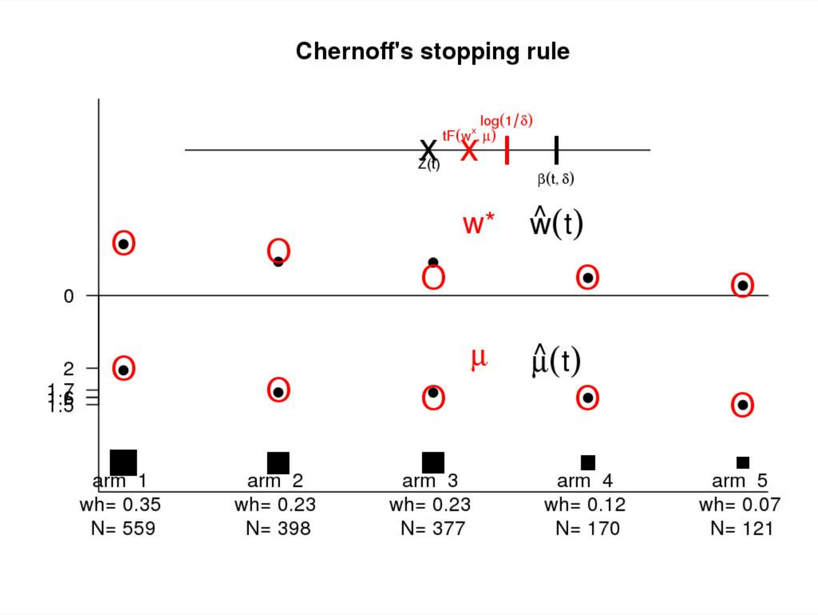
Why is the T&S Strategy asymptotically Optimal?



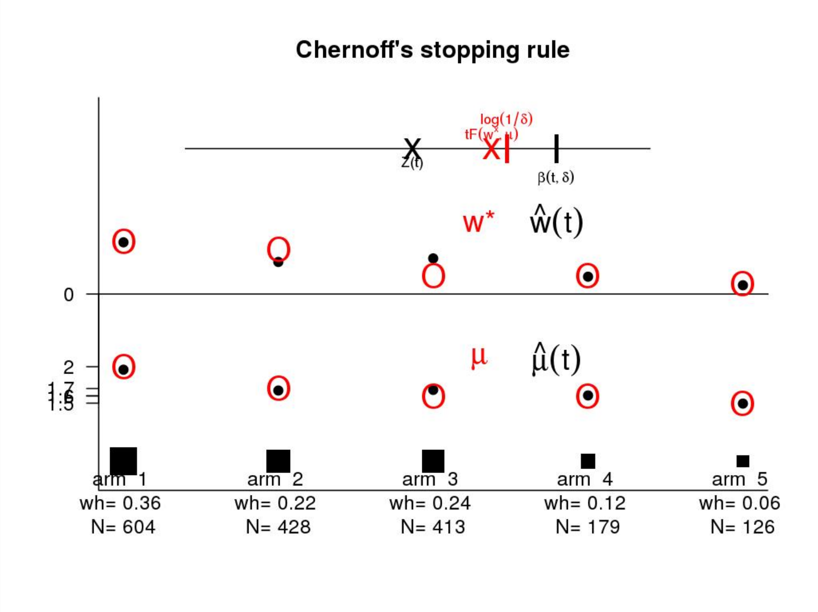
Why is the T&S Strategy asymptotically Optimal?



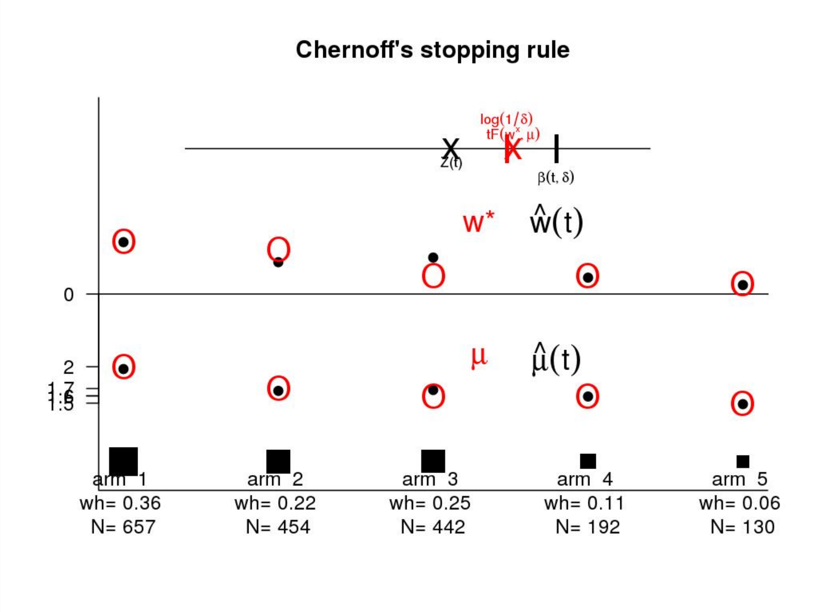
Why is the T&S Strategy asymptotically Optimal?



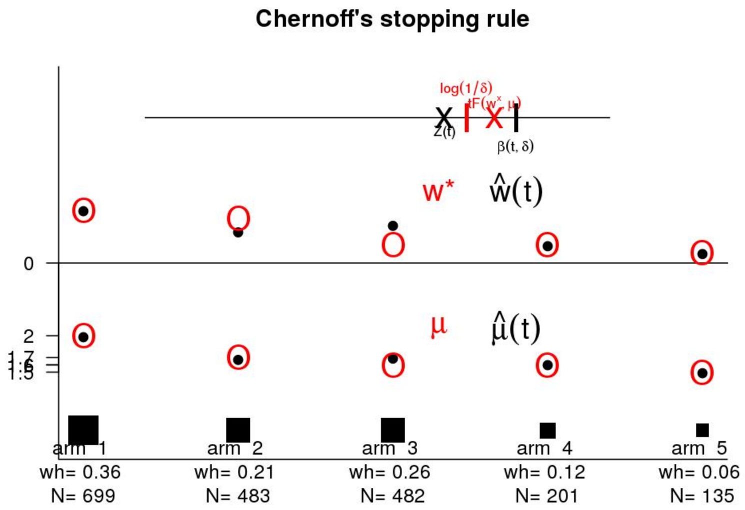
Why is the T&S Strategy asymptotically Optimal?



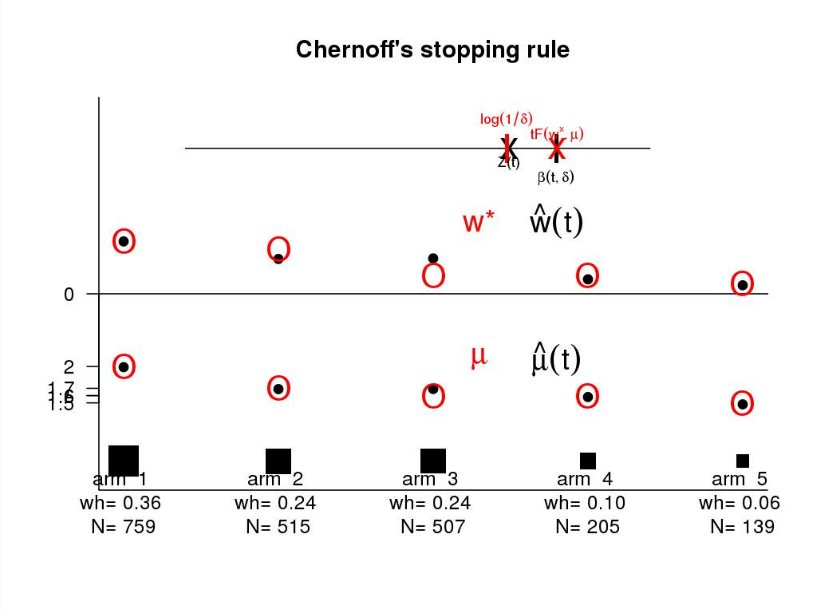
Why is the T&S Strategy asymptotically Optimal?



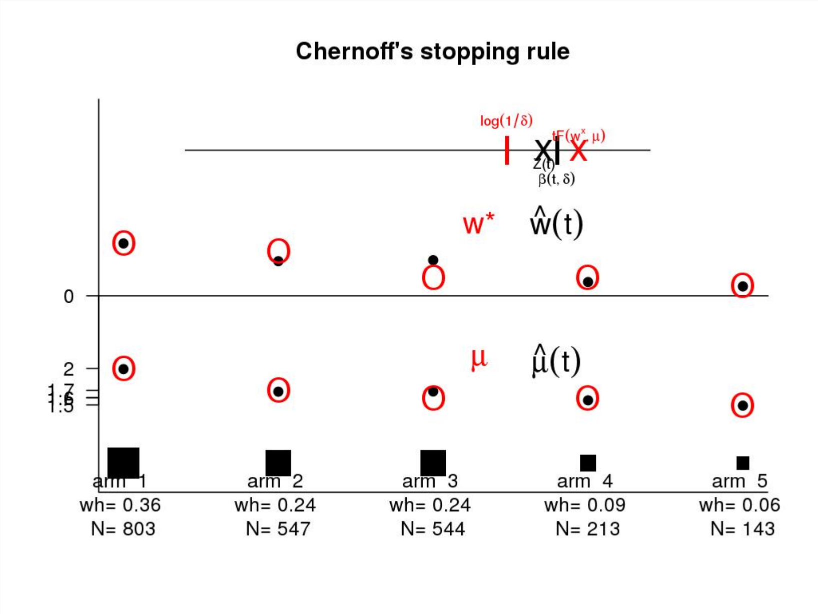
Why is the T&S Strategy asymptotically Optimal?



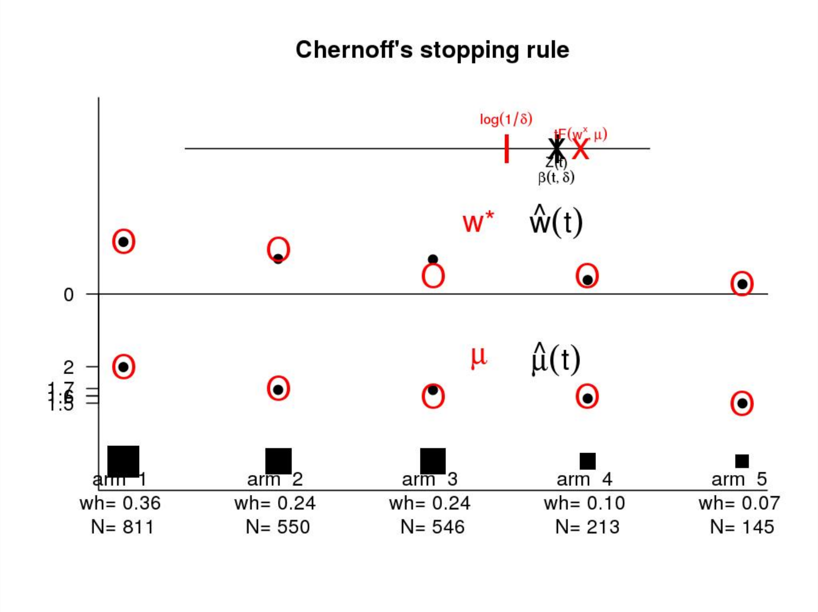
Why is the T&S Strategy asymptotically Optimal?



Why is the T&S Strategy asymptotically Optimal?



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 $a \in \{1, \dots, K\}$

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

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

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

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

➜ $Z(t) = t \times F\left(\hat{\boldsymbol{\mu}}(t), (N_a(t)/t)_{a=1}^{K}\right) \sim t \times F(\mu, \mathbf{w}^*) = t \times T^*(\mu)^{-1}$
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 $\quad \limsup_{\delta \to 0} \dfrac{\tau_\delta}{\log(1/\delta)} \leq (1+\epsilon) T^*(\boldsymbol{\mu}) \quad a.s.$

Numerical Experiments

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

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

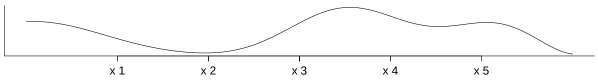
For best arm identification, we showed that

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

and provided an efficient strategy asymptotically matching this bound.

Future work:

- (easy) anytime stopping ➜ gives a confidence level
- (easy) find an ϵ-optimal arm
- (easy) 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 and continuous settings



References

- O. Cappé, A. Garivier, O-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 2013.
- H. Chernoff. Sequential design of Experiments. The Annals of Mathematical Statistics, 1959.
- E. Even-Dar, S. Mannor, Y. Mansour, Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *JMLR*, 2006.
- T.L. Graves and T.L. Lai. Asymptotically Efficient adaptive choice of control laws in controlled markov chains. SIAM Journal on Control and Optimization, 35(3):715743, 1997.
- S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi- armed bandits. ICML, 2012.
- E. Kaufmann, O. Cappé, A. Garivier. On the Complexity of Best Arm Identification in Multi-Armed Bandit Models. *JMLR*, 2015
- A. Garivier, E. Kaufmann. Optimal Best Arm Identification with Fixed Confidence, COLT'16, New York, arXiv:1602.04589
- A. Garivier, P. Ménard, G. Stoltz. Explore First, Exploit Next: The True Shape of Regret in Bandit Problems.
- E. Kaufmann, S. Kalyanakrishnan. The information complexity of best arm identification, COLT 2013
- T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 1985.
- D. Russo. Simple Bayesian Algorithms for Best Arm Identification, COLT 2016
- N.K. Vaidhyan and R. Sundaresan. Learning to detect an oddball target. arXiv:1508.05572, 2015.