1. Multi-armed bandit algorithms.
   - Consistency: optimal per-round reward.
   - Robbins’ consistent algorithm:
     vanishing exploration implies consistency.
   - Upper confidence bound (UCB) algorithms
     (and a foray into concentration inequalities).
Stochastic bandit problem.

- \( k \) arms.
- Arm \( j \) has unknown reward distribution \( P_{\theta_j} \), for \( \theta_j \in \Theta \).
- Reward: \( X_{j,t} \sim P_{\theta_j} \).
- Mean reward: \( \mu_j = \mathbb{E}X_{j,1} \).
- Best: \( \mu^* = \max_{j^*=1,...,k} \mu_{j^*} \).
- Gap: \( \Delta_j = \mu^* - \mu_j \).
- Number of plays: \( T_j(s) = \sum_{t=1}^s 1[I_t = j] \).
- Pseudo-regret:
  \[
  \overline{R}_n = n \max_{j^*=1,...,k} \mu_{j^*} - \mathbb{E} \sum_{t=1}^n X_{I_t,t} = \sum_{j=1}^k \mathbb{E}T_j(n)\Delta_j.
  \]
Call a strategy consistent if 

\[ \frac{R_n}{n} \rightarrow 0. \]

How might we achieve consistency?

- Explore for a while, then exploit?
  But with positive probability, exploration will mislead us.
  \[ \Rightarrow \] Must explore forever.
Robbin’s strategy.

Fix disjoint exploration sequences

\[ 1 = e_{1}^{1} < e_{2}^{1} < \cdots < e_{n}^{1} < \cdots , \]
\[ 2 = e_{1}^{2} < e_{2}^{2} < \cdots < e_{n}^{2} < \cdots , \]
\[ \vdots \]
\[ k = e_{1}^{k} < e_{2}^{k} < \cdots < e_{n}^{k} < \cdots . \]

At time \( t \), if some \( j, i \) has \( t = e_{i}^{j} \), play \( I_{t} = j \). Otherwise play

\[ I_{t} = \hat{j}_{t} = \arg \max_{j} \frac{1}{T_{j}(t)} \sum_{s=1}^{t} X_{I_{s}, s} 1[I_{s} = j]. \]
Robbin’s strategy.

Since \( e_j^n \rightarrow \infty \), \( T_j(t) \rightarrow \infty \), so the strong law of large numbers shows that

\[
\hat{\mu}_j(t) := \frac{1}{T_j(t)} \sum_{s=1}^{t} X_{I_s,s} 1[I_s = j] \xrightarrow{a.s.} \mu_j,
\]

hence \( \hat{j}_t \rightarrow j^* \).

How often should we explore?

- Explore some fixed proportion of the time?
  But that proportion will always cost us.
  \( \Rightarrow \) Must explore forever, but a vanishing fraction of the time.
Robbin’s strategy.

Vanishing exploration implies consistency:

**Theorem:** If the exploration set up to time \( n \),

\[
E_n := \{ t \leq n : \text{some } j, i \text{ has } t = e_i^j \},
\]

satisfies \( |E_n|/n \to 0 \), then

\[
\frac{R_n}{n} = \sum_{j \neq j^*} \frac{\mathbb{E}T_j(n)}{n} \Delta_j \to 0.
\]
Robbin’s strategy.

**Proof.** With vanishing exploration, if $j \neq j^*$,

$$
\frac{T_j(n)}{n} = \frac{1}{n} \sum_{t=1}^{n} \left( 1[\exists i \text{ s.t. } t = x_i^j] + 1[t \notin E_t, \hat{j}_t = j] \right)
$$

$$
\leq \frac{|E_n|}{n} + \frac{1}{n} \sum_{t=1}^{n} 1[\hat{j}_t = j]
$$

$$
\xrightarrow{a.s.} 0.
$$
Upper Confidence Bounds:
Use data to define an upper bound on $\mu_j$.
Choose the arm with the largest upper bound.

- Optimism in the face of uncertainty.
- Nicely balances exploration (few pulls $\Rightarrow$ loose upper bound $\Rightarrow$ more likely to try it) and exploitation (when confidence intervals are small, the best arm has the best upper bound).
• We want tight upper bounds (or we waste our time on a bad arm), but
• We don’t want the bounds too tight (or we might miss a good arm).
• We shouldn’t leave an arm untried for too long (since if we are misled to wrongfully neglect an arm with a very small probability, that becomes important again after a long period of neglect).

We’ll consider estimates based on sample averages, \( \hat{\mu}_j(t) \), and concentration inequalities in terms of \textit{cumulant generating functions}. So we’ll have a brief digression to look at concentration inequalities...
**Concentration inequalities.**

**Definition:** For a random variable $X$ with mean $\mu$, the moment-generating function is

$$M_{X-\mu}(\lambda) = \mathbb{E} \exp(\lambda(X - \mathbb{E}X)),$$

the cumulant-generating function is

$$\Gamma_{X-\mu}(\lambda) = \log M_{X-\mu}(\lambda).$$
Definition: For a random variable $X$, $\psi : \mathbb{R} \to \mathbb{R}$ is a cumulant generating function upper bound if, for $\lambda > 0$,

$$\psi(\lambda) \geq \max \{ \Gamma_X(\lambda), \Gamma_{-X}(\lambda) \},$$

$$\psi(-\lambda) = \psi(\lambda).$$

The Legendre transform (convex conjugate) of $\psi$ is

$$\psi^*(\epsilon) = \sup_{\lambda \in \mathbb{R}} (\lambda \epsilon - \psi(\lambda)).$$
Concentration Inequalities.

Theorem:

\[ \Gamma_{X+c}(\lambda) = \lambda c + \Gamma_X(\lambda), \]
\[ \Gamma^*_{X+c}(\epsilon) = \Gamma^*_X(\epsilon - c). \]

(Easy to check.)
Concentration Inequalities.

**Theorem:** For $\epsilon \geq 0$, $\mathbb{P}(X - \mathbb{E}X \geq \epsilon) \leq \exp \left( -\psi^*_{X - \mathbb{E}X}(\epsilon) \right)$. 
**Concentration Inequality: Proof.**

\[
\log \mathbb{P} \left( X - \mathbb{E}X \geq \epsilon \right)
\]

\[
= \inf_{\lambda > 0} \log \mathbb{P} \left( \exp \left( \lambda \left( X - \mathbb{E}X - \epsilon \right) \right) \geq 1 \right) \quad \text{(exp is monotonic)}
\]

\[
\leq \inf_{\lambda > 0} \log \mathbb{E} \exp \left( \lambda \left( X - \mathbb{E}X - \epsilon \right) \right) \quad \text{(Markov’s inequality)}
\]

\[
\leq \inf_{\lambda > 0} \left( \psi_{X - \mathbb{E}X} (\lambda) - \lambda \epsilon \right) \quad \text{(cgf bound)}
\]

\[
= \inf_{\lambda \in \mathbb{R}} \left( \psi_{X - \mathbb{E}X} (\lambda) - \lambda \epsilon \right) \quad \text{(from } \epsilon > 0, \text{ definition of } \psi (-\lambda))
\]

\[
= -\psi^*_{X - \mathbb{E}X} (\epsilon).
\]
Concentration Inequalities.

**Theorem:** If $X_1, X_2, \ldots, X_n$ are mean zero, i.i.d. with cgf upper bound $\psi$, then $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ has cgf bound

$$
\psi_{\bar{X}_n}(\lambda) = n \psi \left( \frac{\lambda}{n} \right),
$$

and

$$
\psi^*_{\bar{X}_n}(\epsilon) = n \psi^*(\epsilon),
$$

hence,

$$
\mathbb{P} \left( \bar{X}_n \geq \epsilon \right) \leq \exp \left( -n \psi^*(\epsilon) \right),
$$

(Easy to check.)
Example: Gaussian

For $X \sim N(\mu, \sigma^2)$,

$$
\Gamma_{X-\mu}(\lambda) = \frac{\lambda^2 \sigma^2}{2}, \quad \Gamma^*_{X-\mu}(\epsilon) = \frac{\epsilon^2}{2\sigma^2}.
$$

For $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, it’s easy to check that the bound is tight:

$$
\lim_{n \to \infty} \frac{1}{n} \ln P(\bar{X}_n - \mu \geq \epsilon) = -\frac{\epsilon^2}{2\sigma^2}.
$$
Example: Bounded Support

**Theorem:** [Hoeffding’s Inequality] For a random variable $X \in [a, b]$ with $\mathbb{E}X = \mu$ and $\lambda \in \mathbb{R}$,

$$\ln M_{X-\mu}(\lambda) \leq \frac{\lambda^2(b-a)^2}{8}.$$

Note the resemblance to a Gaussian:

$$\frac{\lambda^2\sigma^2}{2} \text{ vs } \frac{\lambda^2(b-a)^2}{8}.$$

(And since $P$ has support in $[a, b]$, $\text{Var}X \leq (b-a)^2/4$.)
Example: Hoeffding’s Inequality Proof

Define

\[ A(\lambda) = \log (\mathbb{E}e^{\lambda X}) = \log \left( \int e^{\lambda x} dP(x) \right), \]

where \( X \sim P \). Then \( A \) is the log normalization of the exponential family random variable \( X_\lambda \) with reference measure \( P \) and sufficient statistic \( x \).

Since \( P \) has bounded support, \( A(\lambda) < \infty \) for all \( \lambda \), and we know that

\[ A'(\lambda) = \mathbb{E}(X_\lambda), \quad A''(\lambda) = \text{Var}(X_\lambda). \]

Since \( P \) has support in \([a, b]\), \( \text{Var}(X_\lambda) \leq (b - a)^2 / 4 \). Then a Taylor expansion about \( \lambda = 0 \) (at this value of \( \lambda \), \( X_\lambda \) has the same distribution as \( X \), hence the same expectation) gives

\[ A(\lambda) \leq \lambda \mathbb{E}X + \frac{\lambda^2}{2} \frac{(b - a)^2}{4}. \]
**Sub-Gaussian Random Variables**

**Definition:** $X$ is **sub-Gaussian** with parameter $\sigma^2$ if, for all $\lambda \in \mathbb{R}$,

$$\ln M_{X-\mu}(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}.$$ 

Note: Gaussian is sub-Gaussian. $X$ sub-Gaussian iff $-X$ sub-Gaussian. $X$ sub-Gaussian implies $P(X - \mu \geq t) \leq \exp(-t^2/(2\sigma^2))$. 
Theorem: For $X_1, \ldots, X_n$ independent, $\mathbb{E} X_i = \mu$, $X_i$ sub-Gaussian with parameter $\sigma^2$, then for all $t > 0$,

$$P \left( \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \geq t \right) \leq \exp \left( -\frac{nt^2}{2\sigma^2} \right).$$