# Recent interactions between online learning and active statistics

*Vianney Perchet*

Colloquium Paris 13
Paris, June 2020

Crest, ENSAE
& Criteo Research, Paris

# Original Motivations

# The Original Motivating Problem

**Heterogenous Source Estimation**

- $d$ different sources $X_k(t) \sim \mathcal{N}(\mu_k, \sigma_k^2)$ to estimate
- Total of $N$ samples to allocate: $(N_1, N_2, \ldots, N_d)$
- Minimization of $\mathbb{E}\|\hat{\mu} - \mu\|^2 = \sum_k \frac{\sigma_k^2}{N_k} = \frac{1}{N} \sum_k \frac{\sigma_k^2}{p_k}$

**Loss defined on Proportions**

$$\mathcal{L}(p_1, \ldots, p_K) = \sum_k \frac{\sigma_k^2}{p_k}, \quad \text{with } p \in \Delta_d$$

# The solution ?

Min. of $\mathcal{L}(p_1, \ldots, p_d) = \sum_k \frac{\sigma_k^2}{p_k}$, constraint to $p \in \Delta_d$

- Easy to solve, $p_k^* = \frac{\sigma_k}{\sum \sigma_j}$ with error $\mathcal{L}(p^*) = (\sum \sigma_k)^2 \simeq \sigma^2 d^2$

**The Question**

What if the $\sigma_k$ are also unknown?

- Sequentially estimate $\widehat{\sigma}_k^2 = \frac{1}{N_k} \sum_{t=1}^{N_k} \left( X_k(t) - \overline{X}_k(t) \right)^2$
  - Bigger $N_k$, better estimation of $\sigma_k^2$
  - Do not overshoot ! Smaller $\sigma_k^2$, smaller $N_k$

Sequential (simultaneous) Estimation **vs.** Optimization

# More complex: Linear Regression

**Standard Linear Regression:** $Y_i = X_i^\top \beta + \varepsilon_i$

Homoscedastic case

- **Design Matrix:** $\mathbb{X} = (X_1, \ldots, X_N)^\top$
- **Unbiased Estimate:** $\widehat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top Y = \beta + (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \varepsilon$
- **Expected Error:** $\mathbb{E}\|\widehat{\beta} - \beta\|^2 = \sigma^2 \operatorname{Tr}(\mathbb{X}^\top \mathbb{X})^{-1}$ if $\varepsilon \sim \mathcal{N}(0, \sigma^2 \operatorname{Id})$.

Heteroscedastic case

- **Known variance:** $\operatorname{Var}(\varepsilon) = \Omega$
- **Unbiased Estimate:** $\widehat{\beta} = (\mathbb{X}^\top \Omega^{-1} \mathbb{X})^{-1} \mathbb{X}^\top \Omega^{-1} Y$
- **Expected Error:** $\mathbb{E}\|\widehat{\beta} - \beta\|^2 = \operatorname{Tr}(\mathbb{X}^\top \Omega^{-1} \mathbb{X})^{-1}$.

# Active Linear Regression

- Fixed Design: $\mathbb{X} \subset \mathbb{R}^{N \times d}$ is fixed and given
- Random Design: $X_i \in \mathbb{R}^d$ are iid $\sim \mathcal{M}(\mathbb{R}^d)$

**Active Design:** From a given set $\{X^{(1)}, \ldots, X^{(K)}\} \subset \mathbb{R}^d$

  - **Choose** $X_i \in \{X^{(1)}, \ldots, X^{(K)}\}$ to sample and **Observe** $Y_i = X_i^\top \beta + \varepsilon(X_i)$
  - Sample $X_{i+1}$, observe $Y_{i+1}$, etc.
  - **Estimate** $\beta$ from $Y_1, \ldots, Y_N$ and $\mathbb{X}$

- **Easy cases:** Homoscedastic or known variance

**Active Heteroscedastic Linear Regression ??**

- "Optimization of design matrix" vs "Estimation of variance"
- **Minimize** $\text{Tr}(\mathbb{X}^\top \Omega^{-1} \mathbb{X})^{-1}$ and **estimate** $\widehat{\Omega}$

# Best solution in hindsight

**Minimize** $\mathrm{Tr}(\mathbb{X}^\top \Omega^{-1} \mathbb{X})^{-1}$ with $X_i \in \{X^{(1)}, \dots, X^{(K)}\}$

- Assume $\varepsilon_t$ independent, Gaussian $\mathcal{N}(0, \sigma^2(X^{(k)}))$
- **Total number N of samples** allowed
  - **Optimal allocation** $\mathbf{N}^{(1)}, \dots, \mathbf{N}^{(K)}$ s.t., $\sum \mathbf{N}^{(k)} = N$.
  - Discretization errors. Consider proportion $\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(K)}$

$$\mathbb{X}^\top \Omega^{-1} \mathbb{X} = \mathbf{N} \sum_k \mathbf{p}^{(k)} \frac{X^{(k)} (X^{(k)})^\top}{\sigma_k^2}$$

- Asymptotically, it boils down to

**Min.** over "sampling simplex" $\mathrm{Tr}(\sum_k \mathbf{p}^{(k)} \frac{X^{(k)} (X^{(k)})^\top}{\sigma_k^2})^{-1}$

# Related questions

**Activification** of Statistical Procedures

- Heterogenous Source Estimation
- Linear regression
- Estimation of Gaussian mixtures
- Clustering
- ..

Sounds like **Exploration** vs **Exploitation**

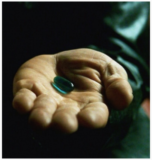
and **multi-armed bandits**

# An intro to multi-armed Bandit

## Classical Examples of Bandits Problems

- Size of data: $n$ patients with some proba of getting cured
- Choose one of two treatments to prescribe

or



- Patients **cured** or **dead**

1) **Inference:** Find the best treatment between the red and blue
2) **Cumul:** Save as many patients as possible

# Classical Examples of Bandits Problems

- Size of data: $n$ banners with some proba of click
- Choose one of two ads to display

or

- Banner **clicked** or **ignored**

1) **Inference:** Find the best ad between the red and blue
2) **Cumul:** Get as many clicks as possible

# Classical Examples of Bandits Problems

- Size of data: $n$ auctions with some expected revenue
- Choose one of two strategies(bid/opt out) to follow

or

- Auction **won** or **lost**

1) **Inference:** Find the best strategy between the red and blue
2) **Cumul:** Win as many profitable auctions as possible

# Classical Examples of Bandits Problems

- Size of data: *n* mails with some proba of spam
- Choose one of two actions: spam or ham


or

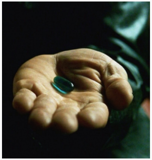

- Mail **correctly** or **incorrectly** classified

1) **Inference:** Find the best strategy between the red and blue
2) **Cumul:** as possibleMinimize number of errors

# Classical Examples of Bandits Problems

- Size of data: *$n$ patients* with some proba of getting cured
- Choose one of two treatments to prescribe

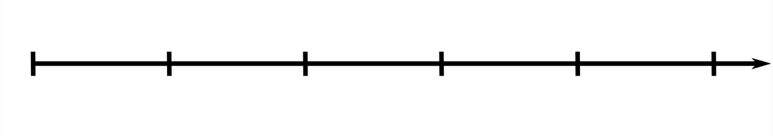
or



- Patients **cured** ♡ or **dead** ☠

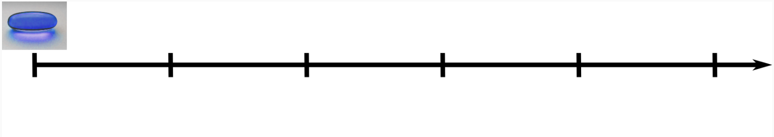
1) **Inference:** Find the best treatment between the red and blue
2) **Cumul:** Save as many patients as possible

# Two-Armed Bandit



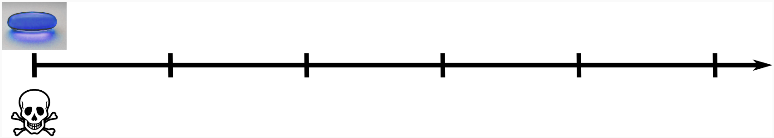
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



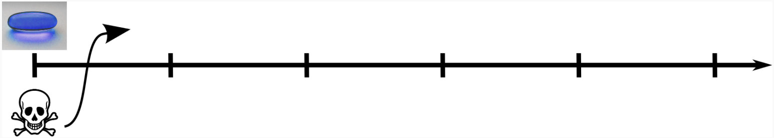
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



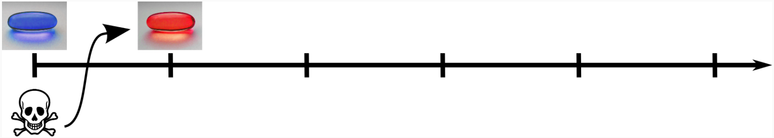
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



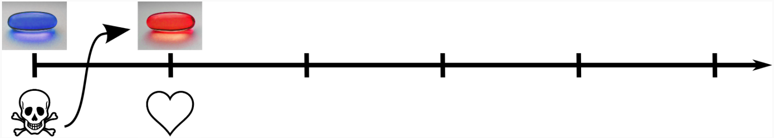
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



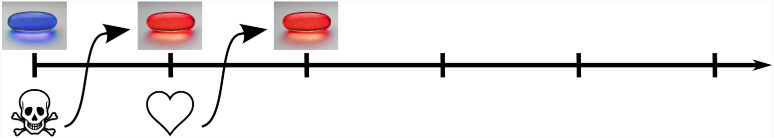
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



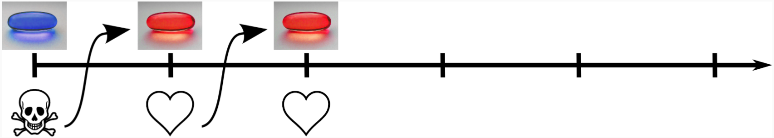
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



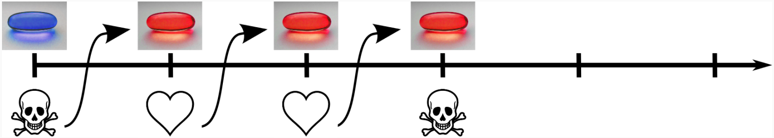
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



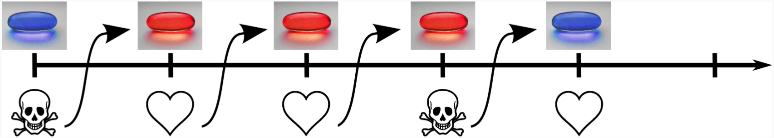
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



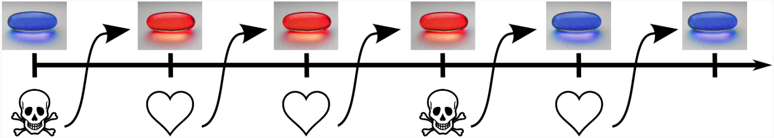
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



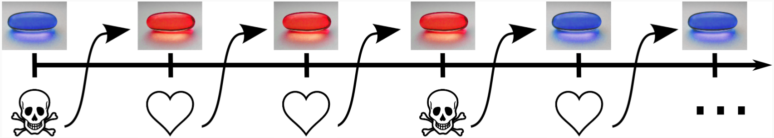
– Patients arrive and are treated sequentially.

# Two-Armed Bandit



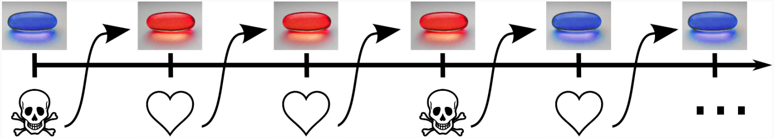
– Patients arrive and are treated sequentially.

# Two-Armed Bandit




– Patients arrive and are treated sequentially.

# Two-Armed Bandit



- Patients arrive and are treated sequentially.
- Save as many as possible.

# Estimation of Means

Discrete-time proc.: $X_n^{(1)}$ in $[0,1]$
"The efficiency of treatment 1 on patient $n$"

**Estimate the mean** $\mu_1$

**Hoeffding inequality: exponential decay**

$\left|\overline{X}_n^{(k)} - \mu_1\right| > \varepsilon$ with proba at most $2\exp\left(-2n\varepsilon^2\right)$.

Finite number of mistakes:

$$\mathbb{E}\sum_{n\in\mathbb{N}} \mathbb{1}\left\{\left|\overline{X}_n^{(k)} - \mu_1\right| > \varepsilon\right\} \leq \frac{1}{\varepsilon^2}$$

# Regret Minimization

- Choose **one** ad to display $k_n$. Reward: $X_n^{(k_n)}$

  Maximize cumulative reward $\sum_{m=1}^{n} X_m^{(k_m)}$ or $\sum_{m=1}^{n} \mu^{(k_m)}$

**Minimize Regret** [Hannan'56]

$$R_n = n\mu^\star - \sum_{m=1}^{n} \mu_{k_m}, \quad \text{with} \quad \mu^\star = \max\{\mu_k\}$$

- Equivalent formulation with $\Delta_k = \mu^\star - \mu_k$:

$$R_n = \sum_k \Delta_k \sum_{m=1}^{n} \mathbb{1}\{k_m = k \neq \star\}$$

# Stochastic & Full Monitoring

- Full Monitoring: all values $X_n^{(k)}$ observed.
- Optimal algorithm: $k_n = \arg\max \overline{X}_n^{(k)}$:

$$\mathbb{E}R_n \leq \sum_k \frac{1}{\Delta_k} \qquad \text{and for small } n,\ \mathbb{E}R_N \leq n \max \Delta_k$$

  **Bounded** regret, **uniformly** in $n$!
- **Given** $n$, worst $\Delta$ is $\sqrt{\frac{d}{n}}$ and $\mathbb{E}R_n \leq \sqrt{dn}$
- But in the examples, **only** $X_n^{(k_n)}$ is observed (bandit monitoring)!

# Stochastic & Bandit Monitoring

– $\overline{X}_n^{(k)} = \frac{1}{n}\sum_{m=1}^n X_m^{(k)}$ **not** available, **only** $\widehat{X}_n^{(k)} = \frac{\sum_{m:k_m=k} X_m^{(k)}}{\sharp\{m : k_m = k\}}$

– with $k_n = \arg\max \widehat{X}_n^{(k)}$, $\mathbb{E}R_n = \Theta(n)$.

because $\mathbb{E}[\overline{X}_n^{(k)}] \leq \mu_k$ **negatively** biased

– **Positive** (vanishing) bias ? Tradeoff Exploitation/Exploration

**Upper Confidence Bound** [Auer,Cesa-Bianchi,Fischer'02]

$$k_n = \arg\max \widehat{X}_n^{(k)} + \sqrt{\frac{2\log(n)}{\sharp\{m : k_m = k\}}}$$

Regret: $\mathbb{E}R_n \leq \sum_k \frac{\log(n)}{\Delta_k}$

# An active linear optim on Multi-Armed Bandits

- $d$ different sources $X_k(t) \sim \mathcal{N}(\mu_k, \sigma_k^2)$
- Total of $N$ samples to sequentially allocate: $(N_1, N_2, \ldots, N_d)$
- Minimization of $\frac{1}{N}\sum_k N_k\mu_k = \sum_k p_k\mu_k$

**Loss defined on Proportions**

$$\mathcal{L}(p_1, \ldots, p_d) = \sum_k p_k\mu_k = p^\top \mu, \text{ with } p \in \Delta_d$$

- Let's take $\sigma_k^2 = 1$ in bandits to simplify

# Back to UCB

**Upper-Confidence Bound - algorithm**

1) Estimate $\mu_k$ by $\overline{\mu}_k(t) = \frac{1}{N_k(t)} \sum_{s=1}^{N_k(t)} X_k(s)$, but **biased**
2) "Positively-bias it" with $\overline{\mu}_k(t) - \sqrt{2\frac{\log(t)}{N_k(t)}}$
3) Sample/pull the "arm" with smallest "unbiased" estimate

**UCB-algo**

$$\pi_{t+1} = \arg\min_k \left\{ \overline{\mu}_k(t) - \sqrt{2\frac{\log(t)}{N_k(t)}} \right\}$$

4) Enjoy Optimization error / "regret"

$$\mathcal{L}(p_N) - \mathcal{L}(p^*) \lesssim \frac{\log(N)}{N} \sum_k \frac{1}{\mu_k - \mu_{k^*}}$$

# Ugly & useless but insightful 1 page proof

$$\mathcal{L}(p_{t+1}) - \mathcal{L}(p^*) = \mathcal{L}(p_t + \frac{1}{t+1}(e_{\pi_{t+1}} - p_t)) - \mathcal{L}(p^*)$$

# Ugly & useless but insightful 1 page proof

$$\begin{aligned}\mathcal{L}(p_{t+1}) - \mathcal{L}(p^*) &= \mathcal{L}\left(p_t + \frac{1}{t+1}(e_{\pi_{t+1}} - p_t)\right) - \mathcal{L}(p^*) \\ &= \mathcal{L}(p_t) - \mathcal{L}(p^*) + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p_t)\end{aligned}$$

# Ugly & useless but insightful 1 page proof

$$
\begin{aligned}
\mathcal{L}(p_{t+1}) - \mathcal{L}(p^*) &= \mathcal{L}\left(p_t + \frac{1}{t+1}(e_{\pi_{t+1}} - p_t)\right) - \mathcal{L}(p^*) \\
&= \mathcal{L}(p_t) - \mathcal{L}(p^*) + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p_t) \\
&= \mathcal{L}(p_t) - \mathcal{L}(p^*) + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(p^* - p_t) \\
&\qquad\qquad\qquad\qquad + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p^*)
\end{aligned}
$$

# Ugly & useless but insightful 1 page proof

$$\begin{aligned}
\mathcal{L}(p_{t+1}) - \mathcal{L}(p^*) &= \mathcal{L}\Big(p_t + \frac{1}{t+1}(e_{\pi_{t+1}} - p_t)\Big) - \mathcal{L}(p^*) \\
&= \mathcal{L}(p_t) - \mathcal{L}(p^*) + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p_t) \\
&= \mathcal{L}(p_t) - \mathcal{L}(p^*) + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(p^* - p_t) \\
&\qquad + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p^*) \\
&\leq \Big(1 - \frac{1}{t+1}\Big)\Big[\mathcal{L}(p_t) - \mathcal{L}(p^*)\Big] + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p^*)
\end{aligned}$$

# Ugly & useless but insightful 1 page proof

$$\begin{aligned}
\mathcal{L}(p_{t+1}) - \mathcal{L}(p^*) &= \mathcal{L}\left(p_t + \frac{1}{t+1}(e_{\pi_{t+1}} - p_t)\right) - \mathcal{L}(p^*) \\
&= \mathcal{L}(p_t) - \mathcal{L}(p^*) + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p_t) \\
&= \mathcal{L}(p_t) - \mathcal{L}(p^*) + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(p^* - p_t) \\
&\qquad\qquad + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p^*) \\
&\leq \left(1 - \frac{1}{t+1}\right)\left[\mathcal{L}(p_t) - \mathcal{L}(p^*)\right] + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p^*) \\
&\leq \frac{t}{t+1}\left[\mathcal{L}(p_t) - \mathcal{L}(p^*)\right] + \frac{1}{t+1}\underbrace{(\mu_{\pi_{t+1}} - \mu_{k^*})}_{:=\varepsilon_{t+1}}
\end{aligned}$$

# Ugly & useless but insightful 1 page proof

$$\begin{aligned}
\mathcal{L}(p_{t+1}) - \mathcal{L}(p^*) &= \mathcal{L}\Big(p_t + \frac{1}{t+1}(e_{\pi_{t+1}} - p_t)\Big) - \mathcal{L}(p^*) \\
&= \mathcal{L}(p_t) - \mathcal{L}(p^*) + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p_t) \\
&= \mathcal{L}(p_t) - \mathcal{L}(p^*) + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(p^* - p_t) \\
&\qquad\qquad + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p^*) \\
&\leq \Big(1 - \frac{1}{t+1}\Big)\Big[\mathcal{L}(p_t) - \mathcal{L}(p^*)\Big] + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p^*) \\
&\leq \frac{t}{t+1}\Big[\mathcal{L}(p_t) - \mathcal{L}(p^*)\Big] + \frac{1}{t+1}\underbrace{(\mu_{\pi_{t+1}} - \mu_{k^*})}_{:=\varepsilon_{t+1}}
\end{aligned}$$

$$\mathcal{L}(p_N) - \mathcal{L}(p^*) \leq \frac{1}{N}\sum_{t=1}^{N}\mu_{\pi_t} - \mu_{k^*} := \frac{1}{N}\sum_{t=1}^{N}\varepsilon_t$$

# Still that proof !

$$\mathcal{L}(p_N) - \mathcal{L}(p^*) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon_t = \frac{1}{N}\sum_{k} N_k(\mu_k - \mu_{k^*})$$

- $\pi_{t+1} = k$ if $\overline{X}_k(t) - \sqrt{\frac{\log(t)}{N_k(t)}} \simeq \mu_k - \sqrt{\frac{\log(N)}{N_k(t)}} \leq \mu_{k^*} \Rightarrow \varepsilon_t \lesssim \sqrt{\frac{\log(t)}{N_k(t)}}$

**Slow rate of convergence**

$$\mathcal{L}(p_N) - \mathcal{L}(p^*) \lesssim \frac{1}{N}\sum_k \sum_{s=1}^{N_k} \sqrt{\frac{\log(N)}{s}}$$
$$\lesssim \frac{1}{N}\sum_k \sqrt{\log(N) N_k} \leq \sqrt{\frac{d\log(N)}{N}}$$

# From slow to fast rates

– Start from the slow rate

$$\frac{1}{N}\sum_{k\neq k^*} N_k(\mu_k - \mu_{k^*}) = \mathcal{L}(p_N) - \mathcal{L}(p^*) \lesssim \frac{1}{N}\sum_k \sqrt{\log(N) N_k}$$

– Enforce $\mu_k - \mu_{k^*}$ and Cauchy-Schwartz

$$\sum_{k\neq k^*} N_k(\mu_k - \mu_{k^*}) \lesssim \sqrt{\sum_{k\neq k^*} N_k(\mu_k - \mu_{k^*})}\sqrt{\sum_{k\neq k^*} \frac{\log(N)}{\mu_k - \mu_{k^*}}}$$

– Enjoy your fast rates !

$$\mathcal{L}(p_N) - \mathcal{L}(p^*) \leq \frac{\log(N)}{N} \sum_k \frac{1}{\mu_k - \mu_{k^*}}$$

# What did we learn with UCB ?

1. **Optimistic Estimation of $\nabla\mathcal{L}(p)$** or "positively-biased"

$$\overline{\mu}_k(t) - \sqrt{2\frac{\log(t)}{N_k(t)}} = \widehat{\nabla}_k^- \mathcal{L}(p_t) \quad \text{and} \quad e_{t+1} = \arg\min_{p\in\Delta_d} \widehat{\nabla}^- \mathcal{L}(p_t)^\top p$$

2. **Variant of Frank-Wolfe:** $p_{t+1} = (1-\gamma_t)p_t + \gamma_t \arg\min_{p\in\Delta_d} \nabla\mathcal{L}(p_t)^\top p$

$$\begin{aligned} p_{t+1} &= (1 - \tfrac{1}{t+1})p_t + \tfrac{1}{t+1} e_{t+1} \\ &= (1 - \tfrac{1}{t+1})p_t + \tfrac{1}{t+1} \arg\min_{p\in\Delta_d} \widehat{\nabla}^- \mathcal{L}(p_t)^\top p \end{aligned}$$

3. **From Slow to Fast Rates** with some simple algebra

$$\mathcal{L}(p_N) - \mathcal{L}(p^*) \lesssim \sqrt{\frac{\log(N)}{N}} \text{ vs. } \frac{\log(N)}{N}$$

# Links with active statistics

# More General Model

Optimization of convex loss $\mathcal{L}(p_N)$ on $\Delta_d$, think of $\mathcal{L}(p) = \sum_k \frac{\sigma_k^2}{p_k}$

- Typical parametric form: $\mathcal{L}_\theta(p) = \sum_k f_k(\theta_k, p_k)$ with $\theta_k$ unknown

**Main assumption (typical case)**

$f_k$ is smooth w.r.t. $p$ and $\theta$

- $\|\nabla f_k(\theta_k, p_k) - \nabla f_k(\theta'_k, p'_k)\| \leq C|p_k - p'_k| + C'\|\theta_k - \theta'_k\|$
- At stage $t$, choose $e_{\pi_t}$ and observe $X_{\pi_t}(t) \sim \mathcal{N}(\theta_{\pi_t}, 1)$

  After $N_k(t)$ observations, $\overline{X}_k(t) \simeq \theta_k \pm \sqrt{\frac{\log(t/\delta)}{N_k(t)}}$
- Noisy information on $\nabla_k \mathcal{L}(\cdot)$ only when sampling process $k$

# Other examples

- **Utility maximization** Optim. **basket of substitutes goods**

# Other examples

- **Utility maximization** Optim. **basket of substitutes goods**
  - V. thinks of "cardio", "bench-press" and "squats" for fitness training



# Other examples

– **Utility maximization** Optim. **basket of substitutes goods**

- V. thinks of “cardio”, “bench-press” and “squats” for fitness training



- Q. thinks of “wine”, “bread” and “cheese” for his breakfast



# Other examples

- **Utility maximization** Optim. **basket of substitutes goods**
  - V. thinks of “cardio”, “bench-press” and “squats” for fitness training



  - Q. thinks of “wine”, “bread” and “cheese” for his breakfast



  - Kobb-Douglas utility $\mathcal{U}(x_1, \dots, x_d) = x_1^{\beta_1} x_2^{\beta_2} \dots x_d^{\beta_d}$
  - Use/buy one good (same price 1), estimate log-utility increase
- **online Markovitz portfolio optimization**
  - Optimize $\mathcal{L}(p) = p^\top \Sigma p - \lambda \mu^\top p$ with $\Sigma$ known, $\mu$ unknown
- **General Case**
  - $\mathcal{L}$ is $C$-smooth w.r.t. $p$ and $\left|\widehat{\nabla}_k \mathcal{L}(p) - \nabla_k \mathcal{L}(p)\right| \leq C' \sqrt{\frac{\log(t/\delta)}{N_k(t)}}$

# The algorithm - Stochastic/Online optimization

UC-FW: Upper Confident Frank-Wolfe

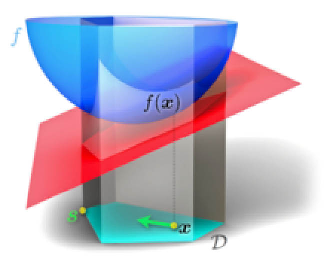
- Optimistic/Unbiased grad. $\widehat{\nabla}_k^- \mathcal{L}(p) = \widehat{\nabla}_k \mathcal{L}(p) - C' \sqrt{\frac{\log(t/\delta)}{N_k(t)}}$
- Frank-Wolfe: $e_{\pi_{t+1}} = \arg\min_{p \in \Delta_k} p^\top \widehat{\nabla}_k^- \mathcal{L}(p_t)$, with $\delta = 1/t$

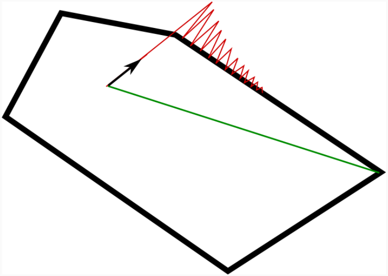
**First result (rather easy) Slow Rate of FwUC**

$$\mathbb{E}\mathcal{L}(p_N) - \mathcal{L}(p^*) \lesssim \sqrt{\frac{d \log(N)}{N}} + \frac{\log(N)}{N}$$

- **Proof ?** (almost) identical to UCB !

# Frank-Wolfe vs Gradient Descent



- For linear functions:

  Projected gradient descent (in red) can converge slowly

  Frank-Wolfe goes straight to the minimum

# The proof. Identical !!

$$
\begin{aligned}
\mathcal{L}(p_{t+1}) - \mathcal{L}^* &= \mathcal{L}(p_t + \frac{1}{t+1}(e_{\pi_{t+1}} - p_t)) - \mathcal{L}^* \\
&\leq \mathcal{L}(p_t) - \mathcal{L}^* + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p_t) + \frac{C}{(t+1)^2} \\
&= \mathcal{L}(p_t) - \mathcal{L}^* + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(p^* - p_t) \\
&\qquad + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p^*) + \frac{C}{(t+1)^2} \\
&\leq \frac{t}{t+1}\Big[\mathcal{L}(p_t) - \mathcal{L}^*\Big] + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p^*) + \frac{C}{(t+1)^2} \\
&\leq \frac{t}{t+1}\Big[\mathcal{L}(p_t) - \mathcal{L}^*\Big] + \frac{1}{t+1}\varepsilon_t + \frac{C}{(t+1)^2}
\end{aligned}
$$

$$
\mathcal{L}(p_N) - \mathcal{L}^* \leq \frac{1}{N}\sum_{t=1}^{N}\varepsilon_t + C\frac{\log(N)}{N} \text{ and } \sum \varepsilon_t \simeq C\sum_k \sum_t \sqrt{\frac{\log(t)}{N_k(t)}}
$$

# Similar results/techniques

- **Stochastic Frank Wolfe** (errors independent of algorithms)
  - [Jaggi], [Lacoste-Julien et al.], [Lafond et al.]
- **Global Cost.** Specific $\mathcal{L}(p) = f(\theta^\top p)$ with $\theta$ unknown, $f$ known
  - **Adversarial:** [Even-Dar et al.], [Blackwell], [Mannor et al.], [Rakhlin et al.]etc.
  - **Stochastic:** [Agrawal and Devanur], [Agrawal et al] Also use stochastic Frank Wolfe
- **Specific Cases.** with pb tailored algorithm
  - [Carpentier et al.], [the bandit community]

# Fast Rates !

# Slow to Fast rates ?

As in **bandit** ? $\sqrt{\frac{d\log(N)}{N}}$ transformed into $\frac{d\log(N)}{N}$ ?

1) Slow rate: $\mathcal{L}(p_N) - \mathcal{L}(p^*) \lesssim \frac{1}{N}\sum_k \sqrt{\log(N) N_k}$

2) Lower bound the convex functions

$$\mathcal{L}(p_N) - \mathcal{L}(p^*) \gtrsim (p_N - p^*)\nabla\mathcal{L}(p^*) \gtrsim \frac{1}{N}\sum_{k \neq k^*} N_k(\nabla_k\mathcal{L}(p^*) - \nabla_{k^*}\mathcal{L}(p^*))$$

3) Cauchy-Schwartz

$$\frac{1}{N}\sum_{k \neq k^*} N_k(\nabla_k\mathcal{L}(p^*) - \nabla_{k^*}\mathcal{L}(p^*)) \lesssim \frac{\log(N)}{N}\sum_{k \neq k^*}\frac{1}{\nabla_k\mathcal{L}(p^*) - \nabla_{k^*}\mathcal{L}(p^*)}$$

4) Another lower bound: fast rate !

$$\mathcal{L}(p_N) - \mathcal{L}(p^*) \lesssim \left(1 + \frac{CK}{\min_k \nabla_k\mathcal{L}(p^*) - \nabla_{k^*}\mathcal{L}(p^*)}\right).\text{lhs} \lesssim O\left(\frac{\log(N)}{N}\right)$$

# What about interior minimized functions ?

- **General Case.** Can we do the same ?
  - Without more assumption, **no**.
  - Maybe with strong convexity

**Strong convexity**

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \mu \|y - x\|^2$$

- **Positive Results.** Fast rates sometimes possible
  - without noise [Garben and Hazan] [Jaggi][...]
  - with decaying noise [Lafond et al.]
  - in online convex optim. [Polyak-Tsybakov], [Bach-P.][...]
- **Negative Results**
  - Cannot leverage strong convexity in online convex optim. [Shamir], [Jamieson et al.]
  - No choice of parameter in FW, has to be $\frac{1}{t+1}$

# The model for fast rates

- On top of the previous assumptions

**Assumptions**

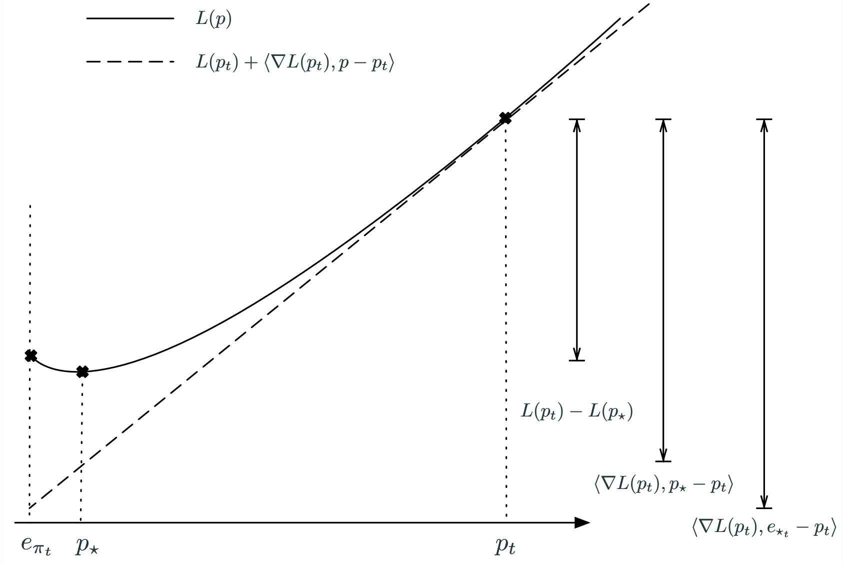
$\mathcal{L}$ is $\mu$-strongly convex and minimized in the interior of $\Delta_d$

$\eta := d(\partial\Delta_d, p^*)$ will play a role [Lacoste-Julien & Jaggi]

$$\mathcal{L}(p) - \mathcal{L}(p^*) \leq \frac{1}{2\mu\eta^2}|\nabla\mathcal{L}(p)^\top(e_{\star,p} - p)|^2$$

where $e_{\star,p} = \arg\min_{q\in\Delta_d} \mathcal{L}(p)^\top q$

# The model for fast rates



# The model for fast rates

- On top of the previous assumptions

**Assumptions**

$\mathcal{L}$ is $\mu$-strongly convex and minimized in the interior of $\Delta_d$

$\eta := d(\partial\Delta_d, p^*)$ will play a role [Lacoste-Julien & Jaggi]

$$\mathcal{L}(p) - \mathcal{L}(p^*) \leq \frac{1}{2\mu\eta^2} |\nabla\mathcal{L}(p)^\top (e_{\star,p} - p)|^2$$

where $e_{\star,p} = \arg\min_{q\in\Delta_d} \mathcal{L}(p)^\top q$

- **Main idea - change in proofs**
  - Before $\frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\star,p_t} - p_t) \leq -\frac{1}{t+1}(\mathcal{L}(p_t) - \mathcal{L}(p^*))$
  - Now $\frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\star,p_t} - p_t) \leq -\frac{\sqrt{2\mu\eta^2}}{t+1}\sqrt{(\mathcal{L}(p_t) - \mathcal{L}(p^*))}$

# Fast rates, our result

FwUC: $e_{\pi_{t+1}} = \arg\min_{p \in \Delta_k} p^\top \widehat{\nabla}_k^- \mathcal{L}(p_t)$, with $\delta = 1/t$

**Assumptions:**

- $C$-smoothness/gradient estimation,
- $\mu$-strong convexity,
- $\eta$-interior minimum

**Main result, Fast rates of FwUC**

$$\mathbb{E}\mathcal{L}(p_N) - \mathcal{L}(p^*) \leq c_1 \frac{\log^2(N)}{N} + c_2 \frac{\log(N)}{N} + c_3 \frac{1}{N}$$

with $c_1 = 3\frac{d(C')^2}{\mu\eta^2}$, $c_2 = 3\frac{dC'\|\mathcal{L}\|_\infty}{(\mu\eta^2)^3}$, $c_3 = dC'\|L\|_\infty + C$

# Some remarks

- **FwUC** **Fully adaptive** to
  - The strong/non-strong convexity **and** the parameter $\mu$
  - The horizon $N$
  - And any other constants/parameters except $C'$
- Parameters **dependencies** (**Leading Term**)
  - **Linear** in the ambiant dimension $d$
  - **inverse-Linear** in the strongly-convexity parameter $\mu$
  - **inverse-square** in the distance to the boundary $\eta$ (but $\frac{1}{d}$ on $\Delta_d$)
- **Generalizations**
  - Gradients errors $\left(\frac{\log(t/\delta)}{N_k(t)}\right)^{\beta}$ with $\beta \leq 1/2$

    Slow rate $\left(\frac{\log(N)}{N}\right)^{\beta}$, and fast rates $\frac{\log(N)}{N^{2\beta}}$
  - (Non-strongly convex) without interior minimum but $\nabla\mathcal{L}(p^*) \ll 0$
- **Lower bounds** matching in $N$ (classic in bandits/stoc. optim)

# Ideas of proof

Objective: $\mathcal{L}(p_t) - \mathcal{L}^* \leq \frac{\sum_{t=1}^{T} \varepsilon_t^2}{T} \simeq \frac{1}{T}\sum_t \frac{\log(t)}{t} \simeq \frac{\log^2(T)}{T}$

$$\begin{aligned}\mathcal{L}(p_{t+1}) - \mathcal{L}^* &\leq \mathcal{L}(p_t) - \mathcal{L}^* + \frac{1}{t+1}\nabla\mathcal{L}(p_t)^\top(e_{\pi_{t+1}} - p_t) + \frac{C}{(t+1)^2} \\ &\leq \mathcal{L}(p_t) - \mathcal{L}^* - \frac{\sqrt{2\mu\eta^2}}{t+1}\sqrt{\mathcal{L}(p_t) - \mathcal{L}^*} + \frac{\varepsilon_t}{t+1} + \frac{C}{(t+1)^2}\end{aligned}$$

- Introducing $\rho_t = \mathcal{L}(p_t) - \mathcal{L}^*$ and $\psi(x) = x - \sqrt{\alpha x}$, we get

$$(t+1)\rho_{t+1} \leq t\rho_t + \left[\psi(\rho_t) - \psi(\tfrac{\varepsilon_t^2}{\alpha})\right] + \frac{\varepsilon_t^2}{\alpha} + \frac{C}{t+1}$$

- if $\psi(\rho_t) - \psi(\frac{\varepsilon_t^2}{\alpha}) \leq 0$ then ok. **but not always...**

  more or less only asymptotically, if everything goes right.

# Some details (again from slow to fast)

– If $\rho_t \le \frac{\varepsilon_t^2}{\alpha}$ then $(t+1)\rho_{t+1} \le t\rho_t + \frac{\varepsilon_t^2}{\alpha} + \frac{C}{t+1}$

$$T\rho_T \le \frac{\tau\varepsilon_\tau^2}{\alpha} + \frac{1}{\alpha}\sum_{t=\tau+1}^{T} \varepsilon_s^2 + C\log(eT)$$

- $\tau\varepsilon_\tau^2 \simeq \dfrac{\log(T)}{p_\tau(\pi_\tau)}$, with $p_\tau(\pi_\tau)$ the current proportion of action $\pi_t$

– **Use again the slow rates !** and strong cvx + interior minimum

$$\|p_\tau - p_*\|^2 \le \frac{1}{\mu}\big(L(p_\tau) - L(p_*)\big) \le \frac{1}{\mu}\frac{\sum_{s=1}^{\tau}\varepsilon_s}{\tau} \le \sqrt{\frac{d\log(T)}{\mu^2 T}}$$

– **Conclude:** $p_\tau(\pi_\tau) \simeq p_\star(\pi_\tau) - \frac{1}{T^{1/4}} > \frac{p_\star(\pi_\tau)}{2}$ is a constant !

# Back to heterogeneous estimation

$$\mathcal{L}(p) = \sum_k \frac{\sigma_k^2}{p_k}, \quad p_k^* = \frac{\sigma_k}{\sum_j \sigma_j}, \quad , \mathcal{L}(p^*) = \big(\sum_j \sigma_j\big)^2$$

- **Main issue:** $\mathcal{L}$ **not smooth in** $p$ nor $\sigma^2$...
  - But smooth "**around**" $p^*$, with $C' \simeq \frac{\sum \sigma_j}{\sigma_{\min}}$ and $C \simeq \frac{(\sum \sigma_j)^3}{\sigma_{\min}}$
- **First phase of rough estimation of** $\sigma_k^2$
  - Difficult to estimate $\sigma_k^2 \pm \varepsilon$, **easy** for $[\frac{\sigma_k^2}{2}, \frac{3\sigma_k^2}{2}]$
  - $X_t \sim \mathcal{N}(\theta_k, 1)$, sample as long as $\overline{X}_\tau \leq \sqrt{\frac{\log(T/\delta)}{\tau}}$
  - Need roughly $\frac{\log(T/\delta)}{\theta^2} = o(T)$ samples
- **Second phase of sampling linear time**
  - $k$ sampled $N\frac{\widehat{\sigma}_k/2}{\sum_j 3\widehat{\sigma}_j/2} \leq N_k$ times
- **Third phase of optimization, using FwUC**
  - $p_t$ far from boundary, close to $p^*$. Valid upper-bounds on $C, C'$