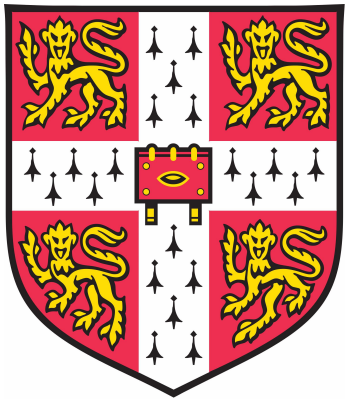
Sequential decision-making under uncertainty: Bandit optimization

Quentin Berthet

The Alan Turing Institute



CCIMI
CANTAB CAPITAL INSTITUTE FOR THE MATHEMATICS OF INFORMATION

CCIMI - Connecting with Industry - 2017

Q. Berthet (Cambridge)

V. Perchet (ENS Paris-Saclay & Criteo)

- **Fast Rates for Bandit Optimization with Upper-Confidence Frank-Wolfe**

  Q. Berthet, V. Perchet, NIPS 2017

  arxiv.org/abs/1702.06917

Multi-armed bandit problems

- **Setting:** K slot machines (one-armed bandits) with reward $\sim \nu_i$ and mean μ_i
- **Sequential aspect:** At each round $t \geq 1$, choice $\pi_t \in [K]$.
- **Bandit feedback:** Observe **only** feedback $X_t \sim \nu_{\pi_t}$ with mean μ_{π_t}
- **Objective:** Maximize reward or expected reward, notion of regret

$$R_T = \sum_{t=1}^{T} X_t \quad \text{or} \quad \sum_{t=1}^{T} \mu_{\pi_t} = \sum_{i \in [K]} \mu_i \, T_i \, .$$

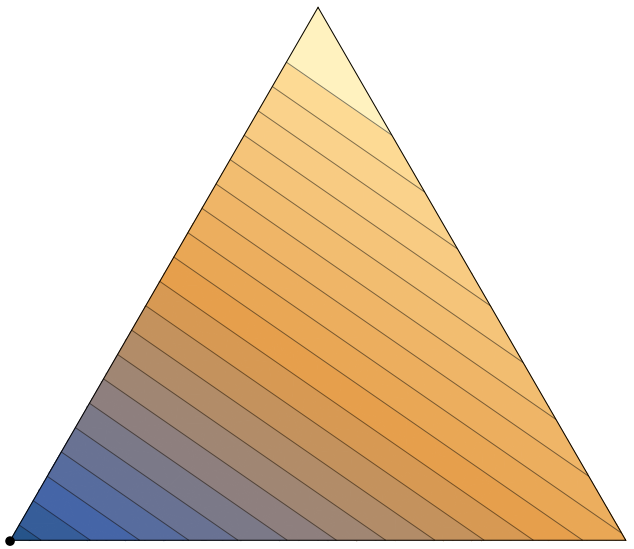
- **Exploitation:** Playing the machines estimated as better (high mean reward).
- **Exploration:** Need to have good estimates $\hat{\mu}_i$ of each mean reward.
- **Stochastic optimization** Maximize function of $(T_1, \ldots, T_K)$ with unknown μ_i.
- **Applications:** Medical experiments, Advertisement, AI in games, $\cdots$

Multi-armed bandits - Stochastic optimization

- **Sequential aspect:** At each round $t \geq 1$, choice $\pi_t \in [K]$.
- **Bandit feedback:** Observe feedback $X_t \sim \nu_{\pi_t}$ with mean loss μ_{π_t}.
- **Objective:** Minimize expected regret, with $L(x) = \langle \mu, x \rangle$

$$R_T = \frac{1}{T}\sum_{t=1}^{T} L(e_{\pi_t}) = \frac{1}{T}\sum_{t=1}^{T} \mu_{\pi_t} = \frac{1}{T}\sum_{i\in[K]} \mu_i T_i = \sum_{i\in[K]} \mu_i \frac{T_i}{T} = L\Big(\frac{1}{T}\sum_{t=1}^{T} e_{\pi_t}\Big)\,.$$

- In this example, we seek to minimize $L(p_t)$, with $p_t = (T_1/T, \ldots, T_K/T)^\top$.
- Stochastic optimization of L on Δ^K, with unknown μ
- Feedback on μ tied to variable p_t
- Optimum at $p_\star = e_\star$.

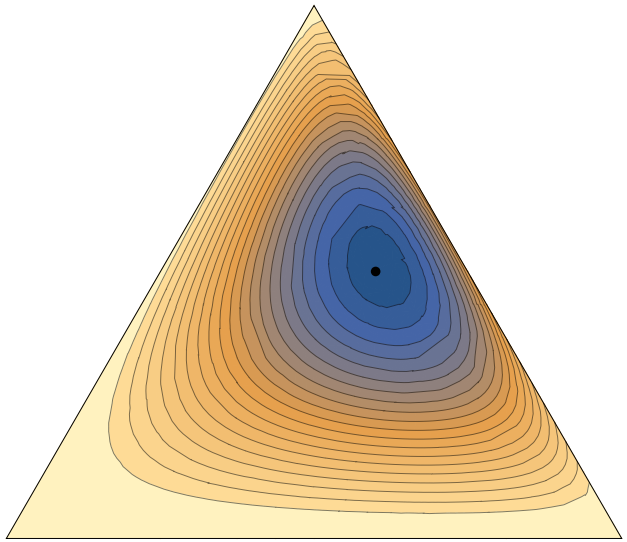


Generalization - Motivation

- **Bandit aspect:** Each round $t \geq 1$, choice $\pi_t \in [K]$, observe bandit feedback.
- **Objective:** Minimize unknown convex loss L in the variable p_t
- **Example:** Basket of K goods, utility maximization with unknown β_i

$$U = \prod_{i=1}^{K} T_i^{\beta_i} = T^B \prod_{i=1}^{K} \left(\frac{T_i}{T}\right)^{\beta_i} \qquad L(p) = -\sum_{i=1}^{K} \beta_i \log(p_i)\,.$$

- Simple example for problem with no optimal action, but optimal strategy
- Stochastic optimization of L on Δ^K, with unknown β
- Feedback on β tied to variable p_t
- Optimum at $p_{\star,i} = \frac{\beta_i}{\sum \beta_j}$.



Generalization - Motivation

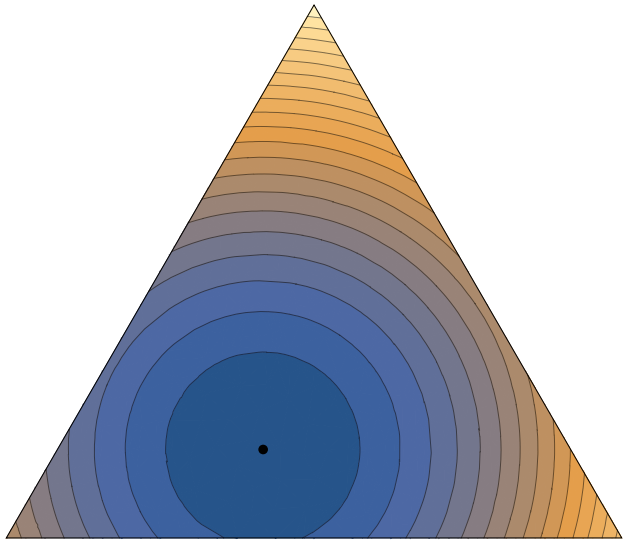
- **Generalization:** Any setting where the loss is a function of p_t (proportions).
- **Applications:** Ressource allocation, experimental design, with uncertainty
  - **Estimation:** Unknown vector $\theta \in \mathbf{R}^K$, K sources with variances σ_i^2

$$\mathbf{E}[\|\hat{\theta} - \theta\|_2^2] = \sum_{i=1}^{K} \frac{\sigma_i^2}{T_i} = \frac{1}{T} \sum_{i=1}^{K} \frac{\sigma_i^2}{T_i/T} \qquad L(p) = \sum_{i=1}^{K} \frac{\sigma_i^2}{p_i} \, .$$

  - **Sequential resource allocation:** $L(p) = \frac{1}{2}\|p - p_\star\|_2^2$, feedback on $p_\star$.

- Stochastic optimization of L on Δ^K, with unknown parameter.
- Feedback on parameter tied to variable p_t.
- Optimum at $p_\star \in \Delta^K$, error

$$L(p_T) - L(p_\star) \, .$$



Problem description

Bandit Optimization:

- Unknown **convex** function $L : \Delta^K \to \mathbf{R}$.
- At each round $t \geq 1$, choice $\pi_t \in [K]$.
- Observe vector $\hat{g}_t$, proxy of the gradient such that w.p. $1 - \delta$

$$|\hat{g}_{t,i} - \nabla_i L(p_t)| \leq \sqrt{\frac{2\log(t/\delta)}{T_i}},$$

  motivated by the parametric setting as a bandit feedback.

Objective: Choosing the actions $\pi_1, \ldots, \pi_T$ in order to minimize $L(p_T)$.

Measure: The performance of any policy is

$$\mathbf{E}[L(p_T)] - L(p_\star)\,.$$

Comparison review

- **Bandit problems**
  - **No regrets:** The loss is not cumulative, and for $e_{\pi_t} \in \Delta^K$

$$\frac{1}{T}\sum_{t=1}^{T} L(e_{\pi_t}) \neq L\Big(\frac{1}{T}\sum_{t=1}^{T} e_{\pi_t}\Big)$$

  - No individual best action, but optimal mixed strategy.
- **Stochastic optimization**
  - Constraint $x_t = p_t$, variable not chosen freely in the domain.
  - Gradient feedback tied to the actions, not independent stochastic.
- **Existing work** Linear bandits with known convex loss

  Agrwal, Devanur, et al., Evan-Dar et al.

"Rules" of bandit problems, "Endgame" of stochastic optimization.

Algorithmic approach

- Problem setting imposes the dynamic on variable $p_t \in \Delta^K$

$$p_{t+1} = \frac{t}{t+1} p_t + \frac{1}{t+1} e_{\pi_{t+1}} = p_t + \frac{1}{t+1}\left(e_{\pi_{t+1}} - p_t\right),$$

  similar to a gradient-type update, in the direction $e_{\pi_{t+1}} - p_t$.

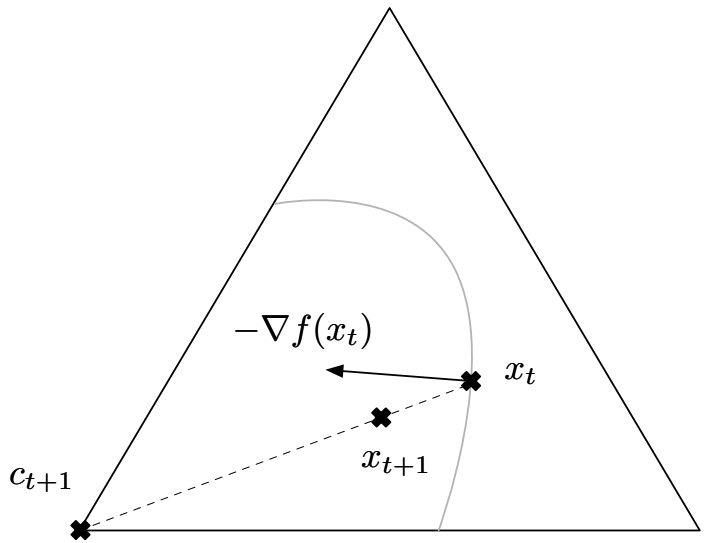
- The Frank-Wolfe algorithm on a convex domain $\mathcal{C}$ follows

$$x_{t+1} = (1-\gamma_t)x_t + \gamma_t c_{t+1} = x_t + \gamma_t\left(c_{t+1} - x_t\right), \text{ for } c_{t+1} \in \operatorname*{argmin}_{u \in \mathcal{C}} \langle \nabla f(x_t), u \rangle .$$

  Frank and Wolfe (56)

- If $\nabla L(p_t)$ were known, we could take

$$e_{\pi_{t+1}} \in \operatorname*{argmin}_{u \in \Delta^K} \langle \nabla L(p_t), u \rangle .$$



Algorithmic approach

- If $\nabla L(p_t)$ were known, applying the Frank-Wolfe algorithm with $\gamma_t = 1/(t+1)$:

$$e_{\pi_{t+1}} \in \operatorname*{argmin}_{u \in \Delta^K} \langle \nabla L(p_t), u \rangle = \operatorname*{argmin}_{i \in [K]} \nabla_i L(p_t)\,.$$

- For a C-smooth function, guarantee of $L(p_T) - L(p_\star) \leq C \log(eT)/T$
- With unknown function, naive idea: using $\hat{g}_t$ as a proxy for the gradient.
- Even for linear functional L (multi-armed bandits), known to be **problematic**.
- Usual fix: correcting for uncertainty. UCB algorithm on $\hat{\mu}_t = \hat{g}_t$.

$$e_{\pi_{t+1}} \in \operatorname*{argmin}_{i \in [K]} \hat{\mu}_{t,i} - \alpha_{t,i} = \operatorname*{argmin}_{u \in \Delta^K} \langle \hat{\mu}_t - \alpha_t, u \rangle\,, \quad \text{Auer et al. (02)}$$

where $\alpha_{t,i}$ is the size of a valid confidence region, order $1/\sqrt{T_i}$.

Upper-confidence Frank-Wolfe algorithm

- Transferring idea of UCB by generalizing it to Frank-Wolfe:

Input: K, $p_0 = \mathbf{1}_{[K]}/K$, sequence $(\delta_t)_{t\geq 0}$;
for $t \geq 0$ **do**
  Observe $\hat{g}_t$, noisy estimate of $\nabla L(p_t)$;
  for $i \in [K]$ **do**
    $\hat{U}_{t,i} = \hat{g}_{t_i} - \sqrt{2\log(t/\delta_t)/T_i}$
  end
  Select $\pi_{t+1} \in \operatorname{argmin}_{i\in[K]} \hat{U}_{t,i}$;
  Update $p_{t+1} = p_t + \frac{1}{t+1}(e_{\pi_{t+1}} - p_t)$
end

- Each coefficient is penalized by uncertainty, promoting exploration
- Runs in time $O(KT)$, omitting gradient computations.
- Proof elements from convex optimization and UCB combined for guarantees.

Results - Slow rate

- Adapting the proof of convergence for known gradients for C-smooth functions

$$L(p_T) - L(p_\star) \leq \underbrace{\frac{1}{T}\sum_{t=1}^{T} \varepsilon_t}_{\text{error of choice}} + \underbrace{\frac{C\log(eT)}{T}}_{\text{opt. term}} .$$

- Together, this yields, for smooth convex functions, the **slow rate**

$$\mathbf{E}[L(p_T)] - L(p_\star) \leq c_2\sqrt{\frac{K\log(T)}{T}} + \frac{C\log(eT) + c_1}{T} .$$

- In the linear case, matches the results of UCB for multi-armed bandits.
- Up to logarithmic terms, matching lower bound in the linear case.

Results - Fast rate

- Instance-dependent fast rates, in multi-armed bandits with gaps $\Delta^{(i)} = \mu_i - \mu_\star$

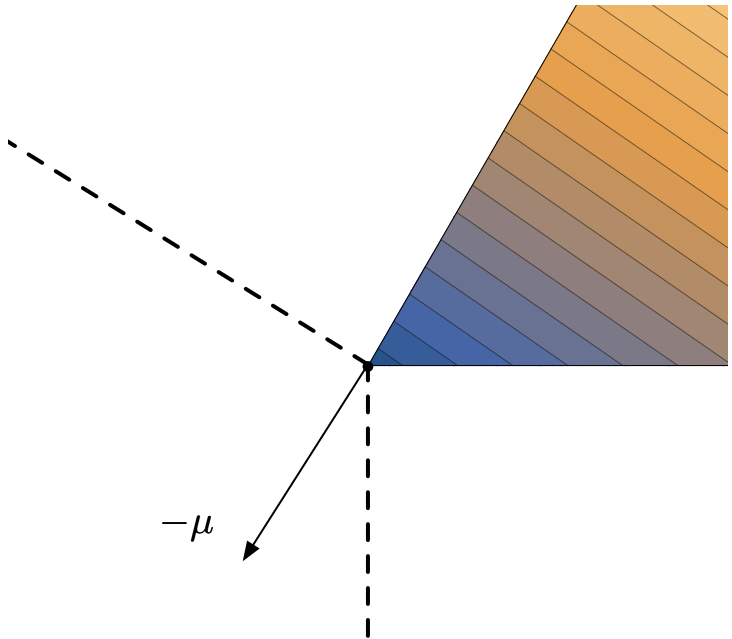
$$\mathbf{E}[L(p_T)] - L(p_\star) \leq \frac{c_2 \log(T)}{T} \sum_{i \neq \star} \frac{1}{\Delta^{(i)}} + \frac{c_1}{T}.$$

- Proof idea: $T(L(p_T) - L(p_\star)) = \sum_{i \neq \star} \Delta^{(i)} T_i$ and

$$\sum_{i \neq \star} \sqrt{T_i} \leq \Big(\sum_{i \neq \star} \Delta^{(i)} T_i \Big)^{1/2} \Big(\sum_{i \neq \star} 1/\Delta^{(i)} \Big)^{1/2}.$$

- General lower bound built on $\Delta \approx 1/\sqrt{T}$

- Geometric interpretation in the simplex corner



Results - Fast rate

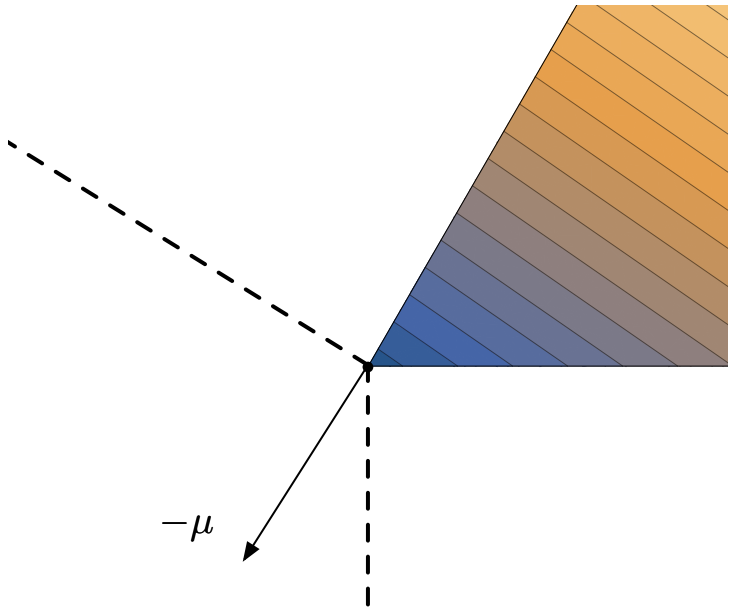
- Possible to extend to gradient gaps $\Delta^{(i)}(L) = \nabla_i L(p_\star) - \nabla_\star L(p_\star)$

$$\mathbf{E}[L(p_T)] - L(p_\star) \leq \frac{c_2 \log(T)}{T} \sum_{i \neq \star} \frac{1}{\Delta^{(i)}(L)} + \frac{c_1}{T} + \frac{C \log(T)}{T}.$$

- Proof idea: $T(L(p_T) - L(p_\star)) \approx \sum_{i \neq \star} \Delta^{(i)}(L)\, T_i$ and

$$\sum_{i \neq \star} \sqrt{T_i} \leq \Big(\sum_{i \neq \star} \Delta^{(i)}(L)\, T_i \Big)^{1/2} \Big(\sum_{i \neq \star} 1/\Delta^{(i)}(L) \Big)^{1/2}.$$

- General lower bound built on $\Delta \approx 1/\sqrt{T}$
- Geometric interpretation in the simplex corner



Results - Fast rate

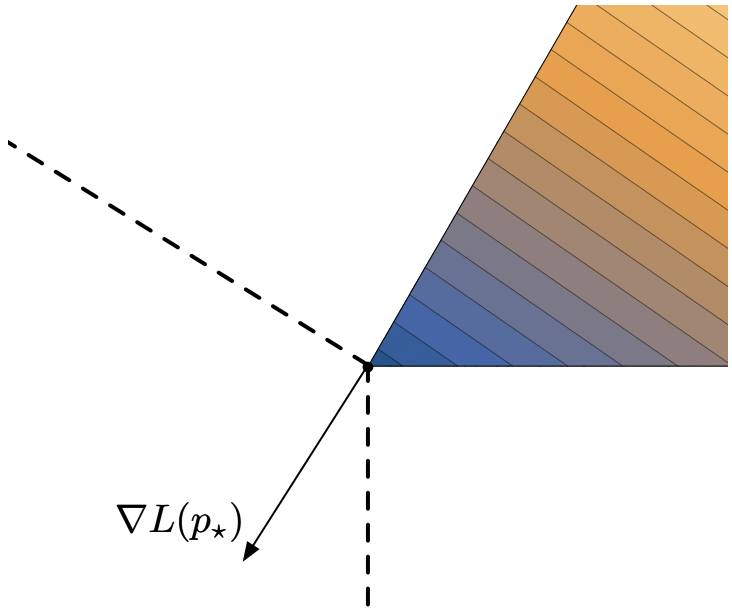
- Possible to extend to gradient gaps $\Delta^{(i)}(L) = \nabla_i L(p_\star) - \nabla_\star L(p_\star)$

$$\mathbf{E}[L(p_T)] - L(p_\star) \leq \frac{c_2 \log(T)}{T} \sum_{i \neq \star} \frac{1}{\Delta^{(i)}(L)} + \frac{c_1}{T} + \frac{C \log(T)}{T}.$$

- Proof idea: $T(L(p_T) - L(p_\star)) \approx \sum_{i \neq \star} \Delta^{(i)} T_i$ and

$$\sum_{i \neq \star} \sqrt{T_i} \leq \Big(\sum_{i \neq \star} \Delta^{(i)} T_i \Big)^{1/2} \Big(\sum_{i \neq \star} 1/\Delta^{(i)} \Big)^{1/2}.$$

- General lower bound built on $\Delta \approx 1/\sqrt{T}$
- Geometric interpretation in the simplex corner



Results - Fast rate

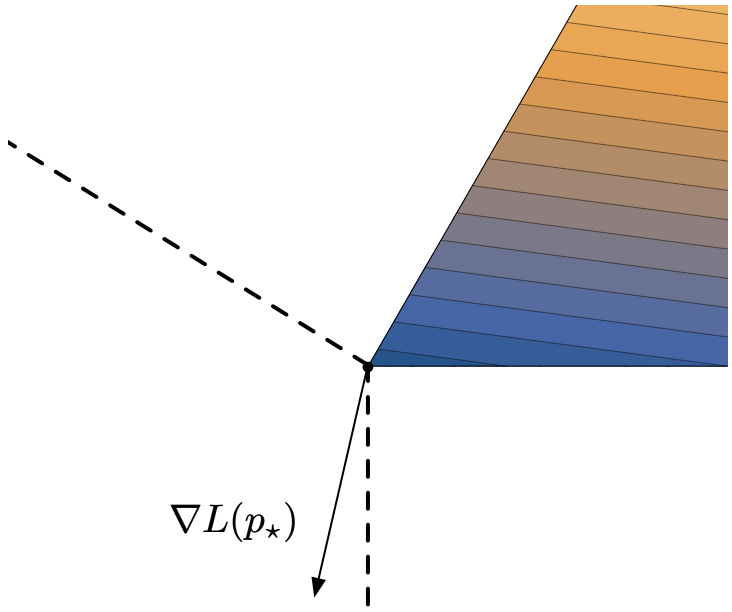
- Possible to extend to gradient gaps $\Delta^{(i)}(L) = \nabla_i L(p_\star) - \nabla_\star L(p_\star)$

$$\mathbf{E}[L(p_T)] - L(p_\star) \leq \frac{c_2 \log(T)}{T} \sum_{i \neq \star} \frac{1}{\Delta^{(i)}(L)} + \frac{c_1}{T} + \frac{C \log(T)}{T}.$$

- Proof idea: $T(L(p_T) - L(p_\star)) \approx \sum_{i \neq \star} \Delta^{(i)} T_i$ and

$$\sum_{i \neq \star} \sqrt{T_i} \leq \Big(\sum_{i \neq \star} \Delta^{(i)} T_i \Big)^{1/2} \Big(\sum_{i \neq \star} 1/\Delta^{(i)} \Big)^{1/2}.$$

- General lower bound built on $\Delta \approx 1/\sqrt{T}$
- Geometric interpretation in the simplex corner



Results - Fast rate

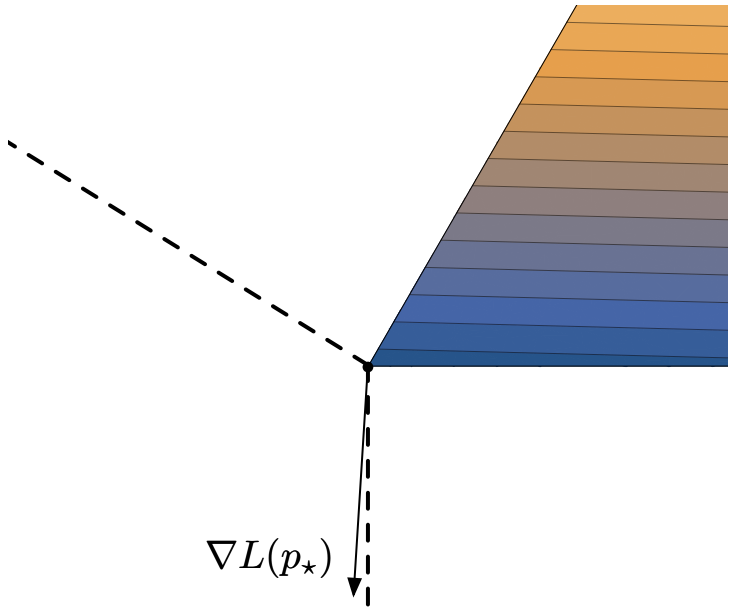
- Possible to extend to gradient gaps $\Delta^{(i)}(L) = \nabla_i L(p_\star) - \nabla_\star L(p_\star)$

$$\mathbf{E}[L(p_T)] - L(p_\star) \leq \frac{c_2 \log(T)}{T} \sum_{i \neq \star} \frac{1}{\Delta^{(i)}(L)} + \frac{c_1}{T} + \frac{C \log(T)}{T}.$$

- Proof idea: $T(L(p_T) - L(p_\star)) \approx \sum_{i \neq \star} \Delta^{(i)} T_i$ and

$$\sum_{i \neq \star} \sqrt{T_i} \leq \Big(\sum_{i \neq \star} \Delta^{(i)} T_i \Big)^{1/2} \Big(\sum_{i \neq \star} 1/\Delta^{(i)} \Big)^{1/2}.$$

- General lower bound built on $\Delta \approx 1/\sqrt{T}$
- Geometric interpretation in the simplex corner



Results - Fast rate

- Slow rate improved to fast rate by adding structure: but mixed strategies?
- We consider μ-strongly convex functions

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|x - y\|_2^2 \,.$$

  with $\text{dist}(p_\star, \partial\Delta^K) = \eta > 0$.

- Classical hypothesis in stochastic optimization and in Frank-Wolfe.

  Polyak and Tsybakov, Garber and Hazan, Jaggi, Lafond et al.

- Not always exploitable, in online and bandit problems.

  Shamir, Jamieson et al.

Results - Fast rate

- For L that is μ-strongly convex and C-smooth with $\text{dist}(p_\star, \partial\Delta^K) = \eta > 0$.
- Improve the convexity bound $L(p_t) - L(p_\star) \leq \nabla L(p_t)^\top (p_t - e_{\star_t})$ to

$$L(p_t) - L(p_\star) \leq \gamma^{-2} |\nabla L(p_t)^\top (p_t - e_{\star_t})|^2 . \quad \text{(Lacoste-Julien and Jaggi)}$$

- Fast rate for μ-strongly convex functions with interior minimum

$$\mathbf{E}[L(p_T)] - L(p_\star) \leq c_1 \frac{\log^2(T)}{T} + c_2 \frac{\log(T)}{T} + c_3 \frac{1}{T} ,$$

  for the same algorithm.

- Proof idea:
  - If ε_t decay fast enough, can replace average error by $\frac{1}{T}\sum \varepsilon_t^2$.
  - The slow rate implies that $T_i(T) \approx p_{\star,i} T$

Upper bounds - Summary

- **Slow rate** for C-smooth functions

$$\mathbf{E}[L(p_T)] - L(p_\star) \lesssim \sqrt{\frac{K \log(T)}{T}}$$

- **Fast rate**
  - With corner condition

$$\mathbf{E}[L(p_T)] - L(p_\star) \lesssim \frac{\log(T)}{T} \sum_{i \neq \star} \frac{1}{\Delta^{(i)}(L)}$$

  - With strong convexity and interior minimum

$$\mathbf{E}[L(p_T)] - L(p_\star) \lesssim c(K, \mu, \eta) \frac{\log^2(T)}{T}$$

- **Remarks**
  - Optimal for the slow rate, lower bound with linear forms.
  - Optimal over restricted classes: with gaps, or strong convexity.

Conclusion

- **Contributions**
  - New model linking stochastic optimization and bandit problem.
  - Natural playground for various applications with sequential decision-making.
  - Solution also bridging these two domains, adaptive fast rates.

- **Open questions**
  - More complicated feedbacks, or side information.
  - Other "objectives" depending on the whole trajectory, not just proportions.
  - Other classes of function: regularity, assumptions.
  - More precise understanding of parameter roles in the complexity.

Work supported by the Isaac Newton Trust and the Alan Turing Institute.