Sequential decisions under uncertainty

KTH/EES PhD course

Lecture 10

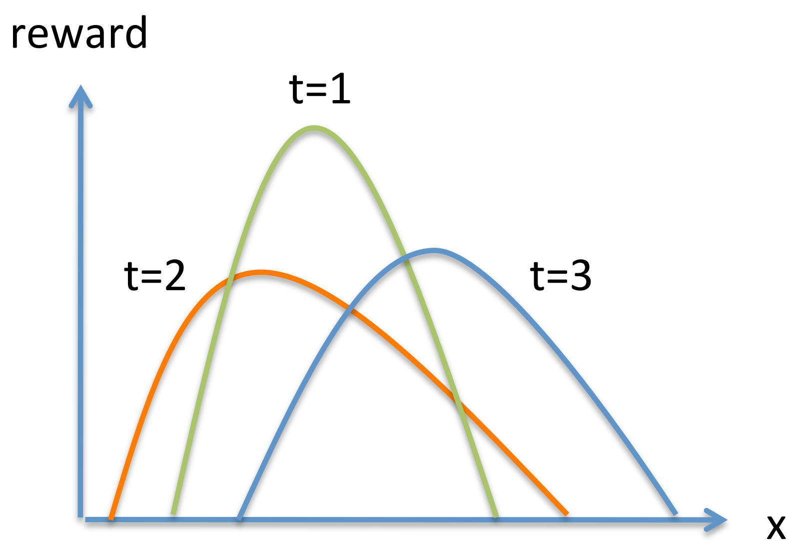
Lecture 10

- Online optimization
  - Full information
  - Bandit setting

Based on

- Online convex programming and generalized infinitesimal gradient ascent. Zinkevich. ICML03
- Online convex optimization in the bandit seting: gradient descent without a gradient. Flaxman, Kalai, McMahan. SODA05.

A motivating example



At the beginning of each year, Volvo has to select a vector x (in a convex set) representing the relative efforts in producing various models (S60, V70, …). The reward is an arbitrarily varying and unknown concave function of x. How to maximize reward over say 50 years?

Model

- Online convex optimization
  - A feasible convex set of actions *X*
  - A sequence of convex cost functions on *X*: $c_1, c_2, \ldots$
- Decision maker
  - Time horizon *N*
  - At step t, selected action x_t
  - Cost: $c_t(x_t)$
  - Feedback. Full information: $\nabla c_t(x_t)$

    Bandit: $c_t(x_t)$

Regret

- Cumulative cost: $\sum_{t=1}^{N} c_t(x_t)$

- Cumulative cost of the best action: $\sum_{t=1}^{N} c_t(x^\star)$

$$x^\star \in \arg\max_{x \in X} \sum_{t=1}^{N} c_t(x)$$

- Regret: $R(N) = \sum_{t=1}^{N} c_t(x_t) - \sum_{t=1}^{N} c_t(x^\star)$

- Goal: minimize regret

Full information

- Online gradient descent

$$w_{t+1} = x_t - \eta \nabla c_t(x_t)$$

$$x_{t+1} = \arg\min_{x \in X} \|x - w_{t+1}\|_2^2$$

Full information

Theorem

Assume that $\mathrm{diam}(X) \leq R$

$$\|\nabla c_t(x)\|_2^2 \leq G, \quad \forall x \in X, \forall t = 1, ..., N$$

Then under the online gradient descent algorithm:

$$R(N) \leq RG\sqrt{N}$$

Bandit setting

- Online convex optimization
  - A feasible convex set of actions *X*
  - A sequence of convex cost functions on *X*: $c_1, c_2, \ldots$
- Decision maker
  - Time horizon *N*
  - At step t, selected action x_t
  - Cost: $c_t(x_t)$

Bandit setting

- Idea: one sample estimate of the gradient

$$\hat{f}(x) = \mathbb{E}_{v \in B}[f(x + \delta v)] \qquad B = \{x : \|x\|_2 \leq 1\}$$

$$\mathbb{E}_{u \in S}[f(x + \delta u)u] = \frac{\delta}{d} \nabla \hat{f}(x) \qquad S = \{x : \|x\|_2 = 1\}$$

- Simulated gradient descent algorithm

u_t uniformly chosen in *B*

$$x_t = y_t + \delta u_t$$

$$y_{t+1} = P_{(1-\alpha)X}(y_t - \nu c_t(x_t) u_t)$$

Bandit setting

Theorem

Assume that $r \leq \mathrm{diam}(X) \leq R$

$$\|\nabla c_t(x)\|_2^2 \leq G, \quad \forall x \in X, \forall t = 1, ..., N$$

$$c_t(x) \leq C, \quad \forall x \in X, \forall t$$

If $N \geq (\frac{3Rd}{2r})^2$, $\nu = \frac{R}{C\sqrt{N}}$, $\delta = (\frac{rR^2d^2}{12N})^{1/3}$, $\quad \alpha = (\frac{3Rd}{2r\sqrt{N}})^{1/3}$

Then under the online gradient descent algorithm:

$$\mathbb{E}[R(N)] \leq 3CN^{5/6}(dR/r)^{1/3}$$