Linear contextual bandits with global constraints

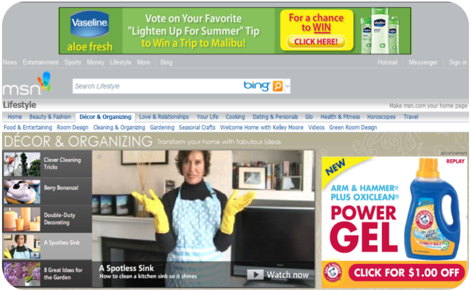
Shipra Agrawal

Industrial Engineering and Operations Research
Columbia University

Based on joint work with Nikhil R. Devanur.

Application: Revenue management in internet advertising

- Operating delivery of ads so that long term revenue from the business is maximized
- Multi-billion dollar annual revenues



Pay-per click advertising

Advertisers specify target user profiles, payment per click

- user opens a page at time t, matches target profile of many ads
- pick one ad
- "if the user clicks" on the shown ad, publisher gets paid

Uncertainty in future user profiles, uncertainty in clicks

"Click-through rate" depends on a combination of user profile and ad features.

Linear regression model

Click-through rates as a linear function of user and ad features.

- Let $x_{t,a}$ be a vector of features of (user t, ad a) combination
- On serving ad a to the user t, the chances of getting clicked is $w^T x_{t,a}$ for some unknown vector w.

Linear contextual bandit problem: explore-exploit in the feature space to learn w quickly.

Linear contextual bandits

In every round t, pick one of the many options (arms) in set A_t.

- For every $a \in A_t$, observe "context vector" $x_{t,a} \in \mathbb{R}^d$ before making the choice.
- On picking option a, observe reward $r_t \in [0, 1]$

Stochastic assumptions

- Reward r_t on picking arm a is i.i.d. from distribution with mean $w^T x_{t,a}$, w is unknown.
- No assumptions on the set A_t or context vectors – could be adversarial

Linear contextual bandits

Goal

- maximize sum of rewards $\sum_t r_t$
- minimize expected regret: compared to best context-dependent policy

$$\mathcal{R}(T) = \sum_t \max_{a \in A_t} w^T x_{t,a} - \mathbb{E}[\sum_t r_t]$$

UCB algorithms

- maintain a confidence ellipsoid around least-square estimate of w, use the most optimistic value $\tilde{w}_t$ in the ellipsoid at time t
- at step t, play $\arg\max_{a \in A_t} \tilde{w}_t^T x_{t,a}$.
- achieve $\tilde{O}(d\sqrt{T})$ regret

Further considerations

Budget constraints!

Maximize the total value while not exceeding the budgets

$$\begin{array}{rl} \text{maximize} & \sum_{t,a\in A_t} r_{t,a} y_{t,a} \\ \forall t, & \sum_{a\in A_t} y_{t,a} \le 1 \\ \forall \text{ads } a, & \sum_{t:a\in A_t} r_{t,a} y_{t,a} \le B_a \end{array}$$

Benchmark: Optimal context dependent policy?

Earlier:

- Maximize $\sum_t w^T x_{t,a_t}$

Benchmark: Optimal context dependent policy?

Earlier:

- Maximize $\sum_t w^T x_{t,a_t}$
- Optimal choice: at every time step choose

$$a_t = \arg\max_{a \in A_t} w^T x_{t,a}$$

Benchmark: Optimal context dependent policy?

Earlier:

- Maximize $\sum_t w^T x_{t,a_t}$
- Optimal choice: at every time step choose

$$a_t = \arg\max_{a \in A_t} w^T x_{t,a}$$

- uncertainty in context set A_t did not matter, if you knew the regression parameter w

Benchmark: Optimal context dependent policy?

Earlier:

- Maximize $\sum_t w^T x_{t,a_t}$
- Optimal choice: at every time step choose

$$a_t = \arg\max_{a \in A_t} w^T x_{t,a}$$

- uncertainty in context set A_t did not matter, if you knew the regression parameter w

Now:

Benchmark: Optimal context dependent policy?

Earlier:

- Maximize $\sum_t w^T x_{t,a_t}$
- Optimal choice: at every time step choose

$$a_t = \arg\max_{a \in A_t} w^T x_{t,a}$$

- uncertainty in context set A_t did not matter, if you knew the regression parameter w

Now:

- Even if you know w, the choice at every step is not obvious

Benchmark: Optimal context dependent policy?

Earlier:

- Maximize $\sum_t w^T x_{t,a_t}$
- Optimal choice: at every time step choose

$$a_t = \arg\max_{a \in A_t} w^T x_{t,a}$$

- uncertainty in context set A_t did not matter, if you knew the regression parameter w

Now:

- Even if you know w, the choice at every step is not obvious
- Ad a or a'?
  - Ad a has highest immediate revenue, but it appears in A_t very frequently
  - Ad a' has smaller immediate revenue, but there may not be another opportunity to use its budget.

Stochastic assumption and Benchmark

Stochastic assumption on A_t:

- Set A_t of context vectors is generated i.i.d. from some distribution $\mathcal{D}$ over collection of sets of context vectors

Benchmark:

Value of best static context-dependent policy $q : A \to \Delta^N$,

$$\mathrm{OPT} = \begin{array}{cc} \max_q & \mathbb{E}[\sum_{t,a \in A_t} r_{t,a}\ q(A_t)_a] \\ \forall \text{ads } a, & \mathbb{E}[\sum_{t:a \in A_t} r_{t,a}\ q(A_t)_a] \leq B_a \end{array}$$

- Expectation over distribution of A_ts, and of $r_{t,a}$ given $w, x_{t,a}$.
- OPT is as good as any adaptive solution that knows w AND the distribution of A_ts.

Further considerations

- Multiple types of feedback – revenue, relevance, cost of serving, click, conversions, demographic targeting
  - Multidimensional reward or value vector

Further considerations

- Multiple types of feedback – revenue, relevance, cost of serving, click, conversions, demographic targeting
  - Multidimensional reward or value vector
- Nonlinear
  - Risk on over-spend, under-delivery
  - Diversity of user profiles
  - Smooth delivery

Further considerations

- Multiple types of feedback – revenue, relevance, cost of serving, click, conversions, demographic targeting
  - Multidimensional reward or value vector
- Nonlinear
  - Risk on over-spend, under-delivery
  - Diversity of user profiles
  - Smooth delivery

Can be modeled as convex constraints and objective

$$\begin{array}{ll} \max & f(\sum_{t,a} \mathbf{v}_{t,a} y_{t,a}) \\ & \sum_{t,a} \mathbf{v}_{t,a} y_{t,a} \in S \\ \forall t, & \sum_a y_{t,a} \leq 1 \end{array}$$

Online decisions with unknown distribution of $\mathbf{v}_{t,a}$!

Linear contextual bandits with global convex constraints and objective

In every round t, pick one of the many options (arms) in set A_t.

- For every $a \in A_t$, observe "context vector" $x_{t,a} \in \mathbb{R}^d$ before making the choice.
- On pulling arm a, observe vector $\mathbf{v}_t \in [0,1]^d$

Stochastic assumptions:

- Given that arm a is pulled, vector $\mathbf{v}_t$ is i.i.d. from distribution with mean $W^T x_{t,a}$, matrix W is unknown.
- Set A_t of context vectors is generated i.i.d. from some distribution over collection of context vectors

Linear contextual bandits with global convex constraints and objective

Goal:

- Maximize $f(\frac{1}{T}\sum_{t=1}^{T}\mathbf{v}_t)$ while ensuring $\frac{1}{T}\sum_{t=1}^{T}\mathbf{v}_t \in S$
- Minimize expected regret:

$$\text{Regret in Objective} = \text{OPT} - f(\frac{1}{T}\sum_{t=1}^{T}\mathbf{v}_t)$$

OPT is the value of best context-dependent policy (?)

$$\text{Regret in constraints} = d(\frac{1}{T}\sum_{t}\mathbf{v}_t, S)$$

$d(\cdot,\cdot)$ is a distance function, e.g. L_1 distance.

Benchmark

Value of best static context-dependent policy

$$\mathrm{OPT} = \begin{array}{ll} \max_q & f\left(\mathbb{E}\left[\left(\sum_{t,a} W^T x_{t,a}\right) q(A_t)\right]\right) \quad \text{such that} \\ & \mathbb{E}\left[\left(\sum_{t,a} W^T x_{t,a}\right) q(A_t)\right] \in S \end{array}$$

- OPT is as good as any adaptive solution that knows W AND the distribution of contexts.

Our results

- $\tilde{O}(dT^{-1/3})$ regret bounds in both objective and distance from constraint set
- $\tilde{O}(d/\sqrt{T})$ regret bound if
  - value of OPT is known to sufficient accuracy.
  - concave objective, no constraints
  - only constraints: feasibility problem
- Important: no dependence on number of arms (possible user+ad types, which is exponential in d)

Main components of the algorithm

Handling unknown W

- On making an observation, update estimate of W using standard linear contextual bandit techniques

Handling uncertainty in contexts: Even with an accurate W, the problem is difficult: "online stochastic convex programming" [Agrawal, Devanur, SODA 2015].

Overview of the algorithm for known W

One dimensional problem, A_t of size 2, objective only.
(W.l.o.g. expected reward $wx_{t,a}$ can be replaced by $x_{t,a}$.)

At time t,

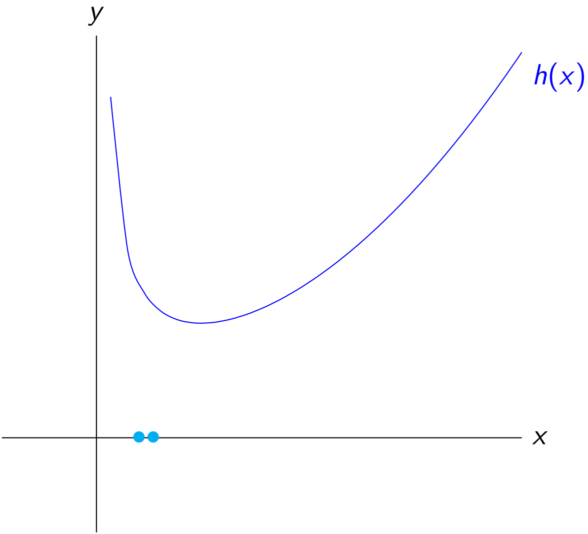
- you see random points $\{x_{t1}, x_{t2}\}$ on x-axis (stochastic assumption).
- Choose one of those points as $x_t^\dagger$.

Overall goal is to minimize $h(\frac{1}{T}\sum_{t=1}^{T} x_t^\dagger)$, where h is convex.

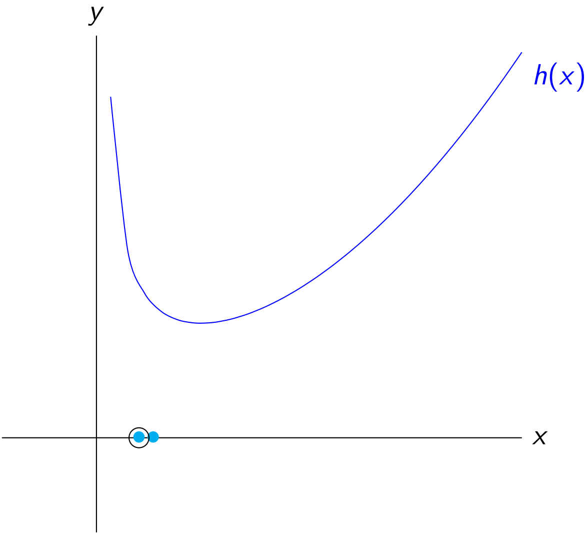
Regret

$$\mathcal{R}(T) = h(\frac{1}{T}\sum_{t=1}^{T} x_t^\dagger) - h(\frac{1}{T}\sum_{t=1}^{T} x_t^*).$$

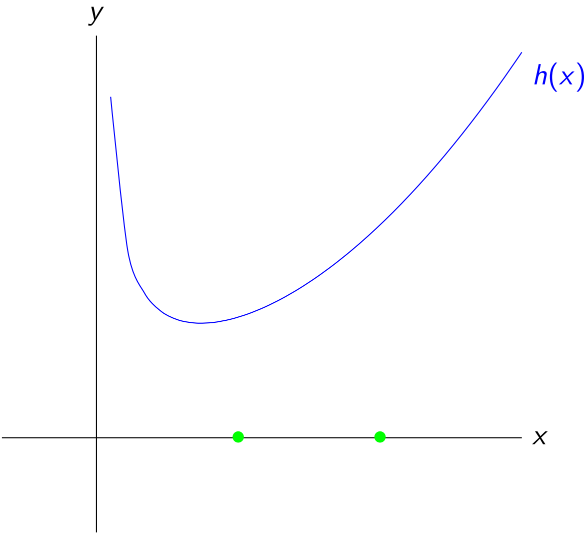
Overview by example



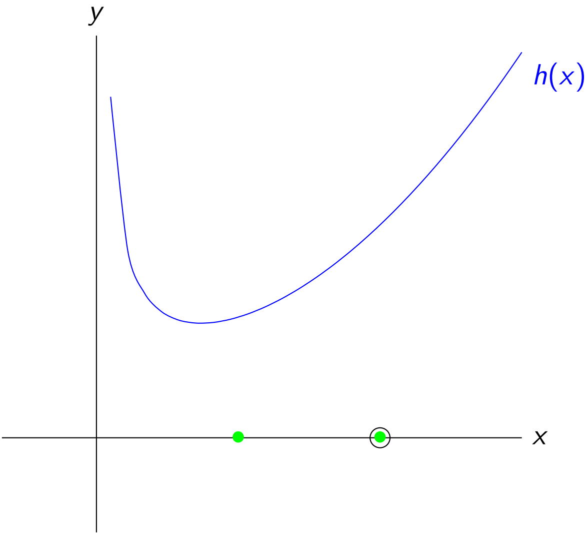
Overview by example



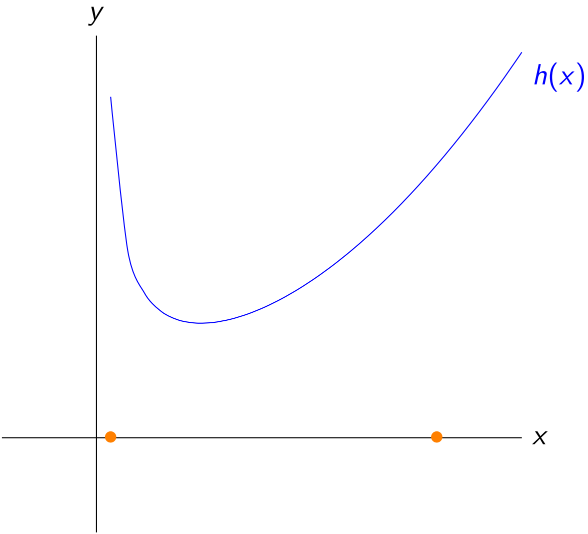
Overview by example



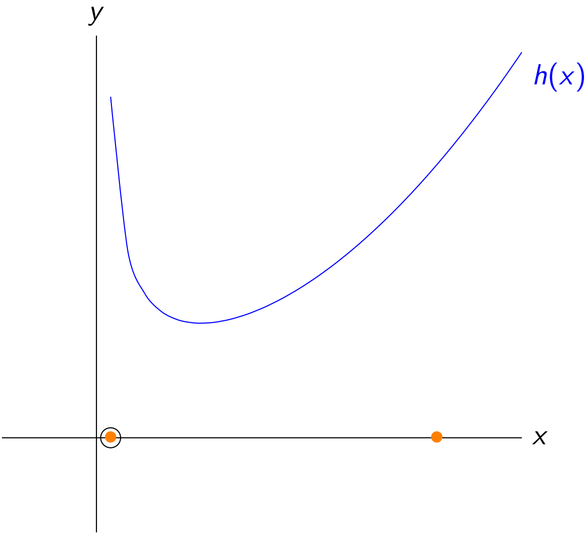
Overview by example



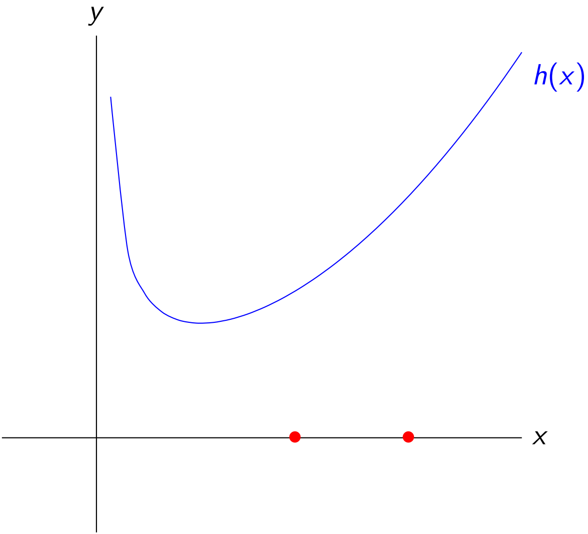
Overview by example



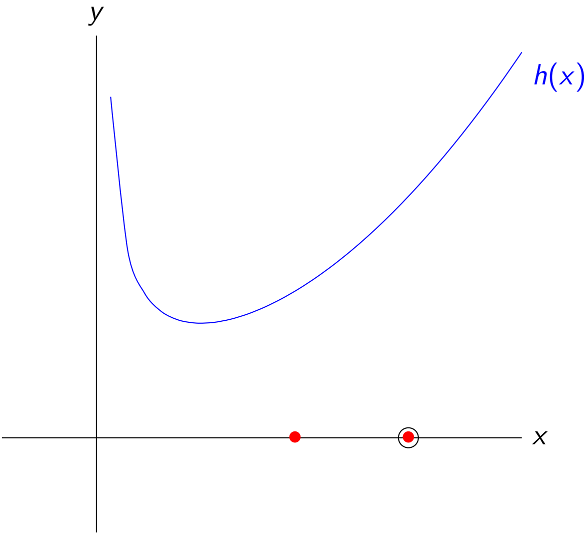
Overview by example



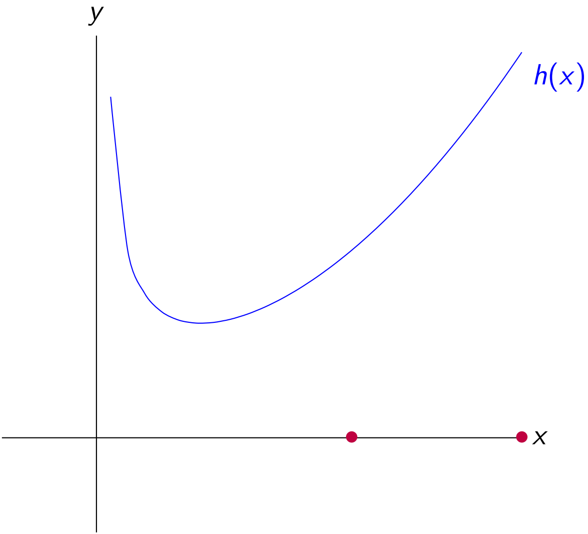
Overview by example



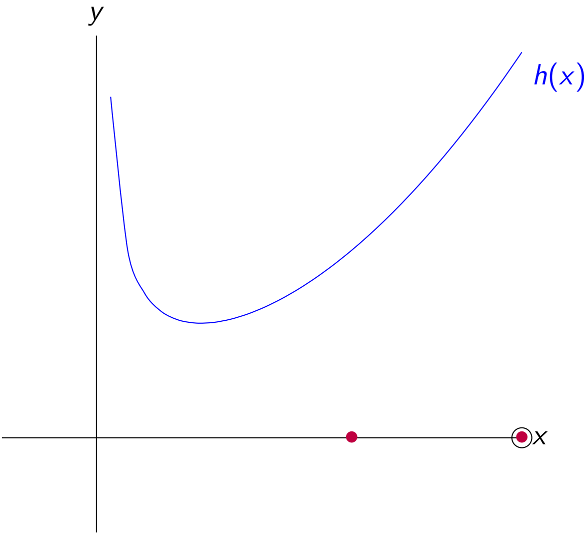
Overview by example



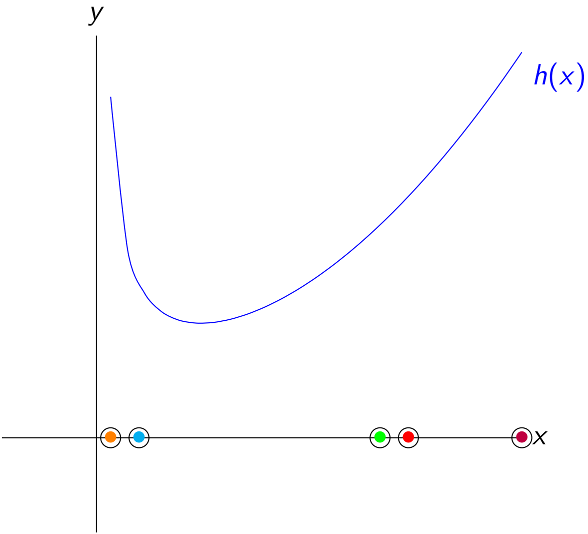
Overview by example



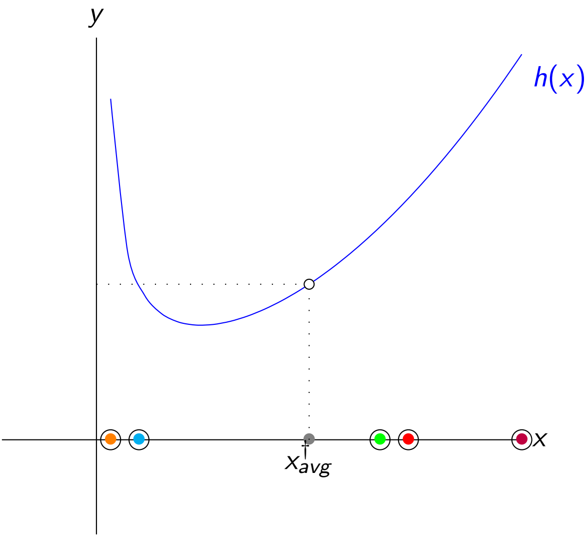
Overview by example



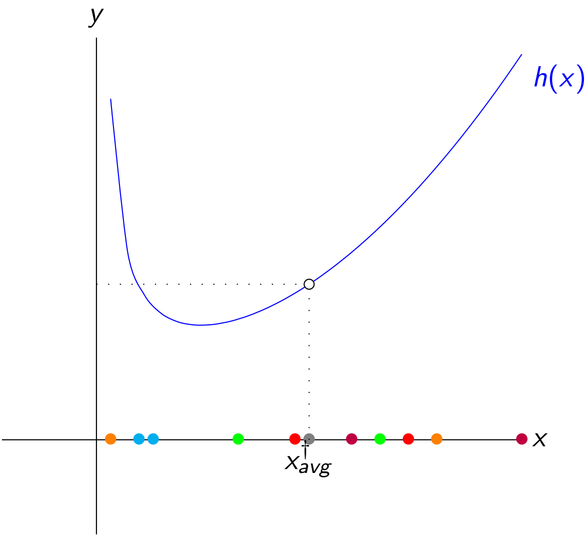
Overview by example



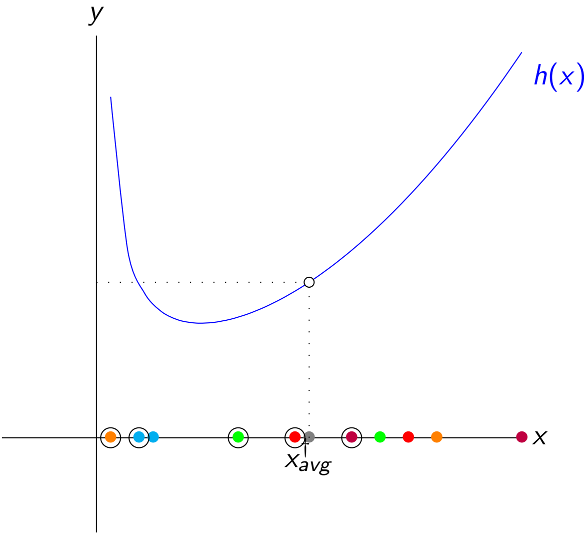
Overview by example



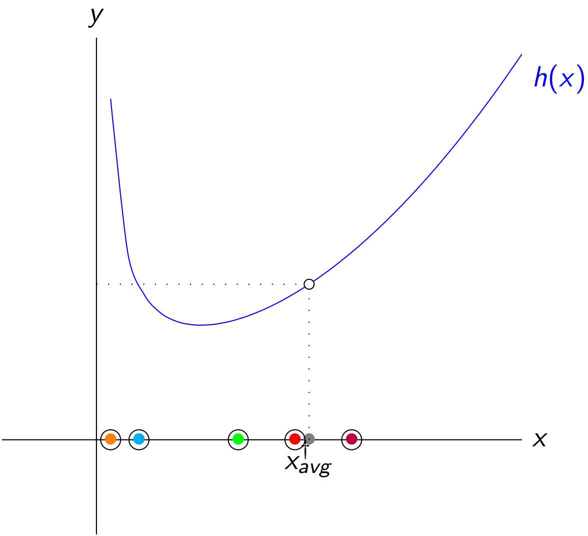
Overview by example



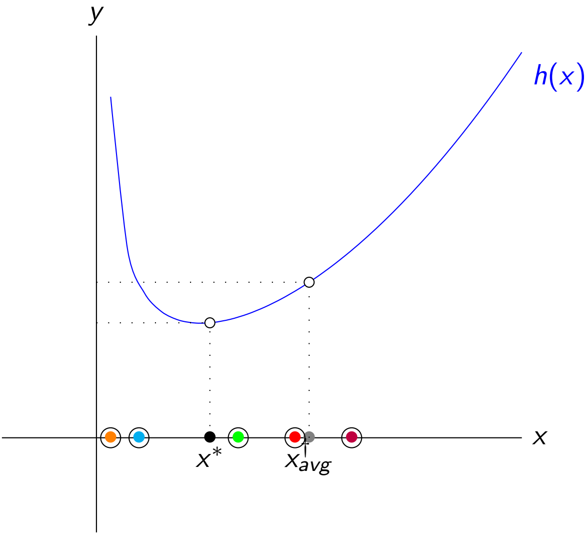
Overview by example



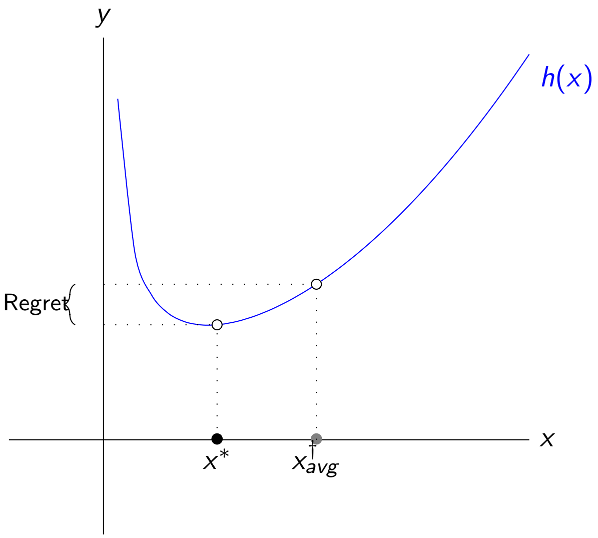
Overview by example



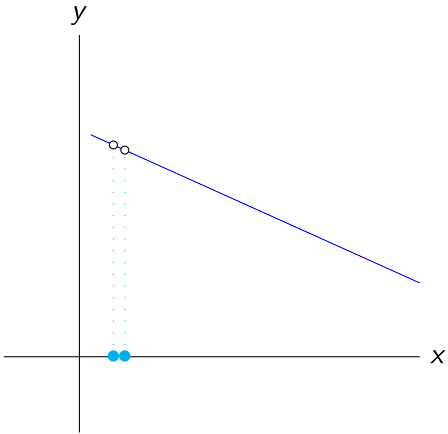
Overview by example



Overview by example

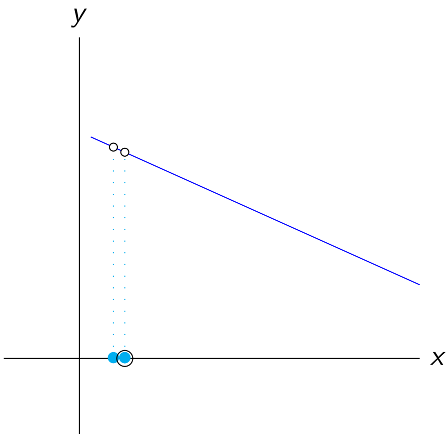


The simpler linear case



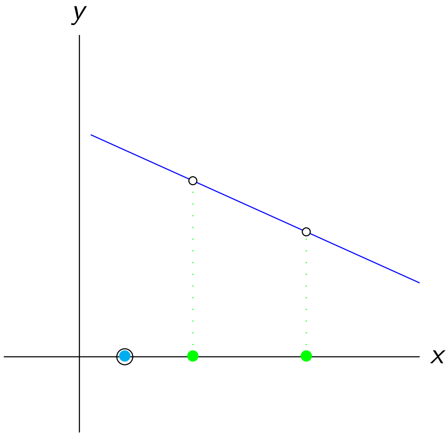
$$h\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) = \frac{1}{T}\sum_t h(x_t)$$

The simpler linear case



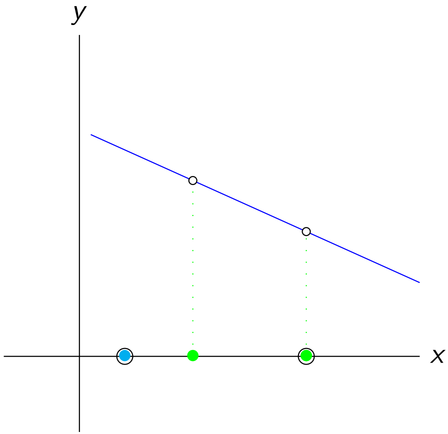
$$h\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) = \frac{1}{T}\sum_t h(x_t)$$

The simpler linear case



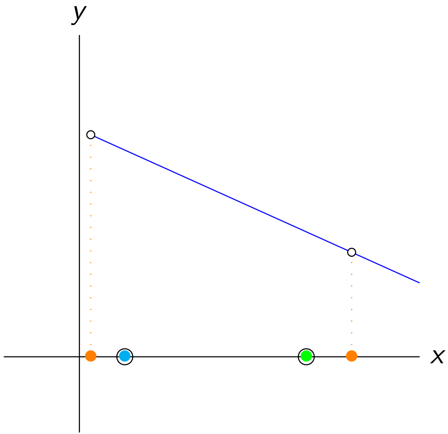
$$h\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) = \frac{1}{T}\sum_t h(x_t)$$

The simpler linear case



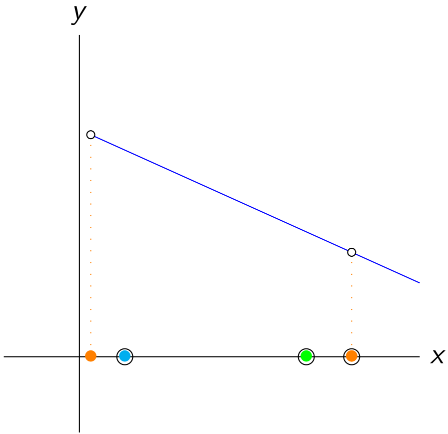
$$h\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) = \frac{1}{T}\sum_t h(x_t)$$

The simpler linear case



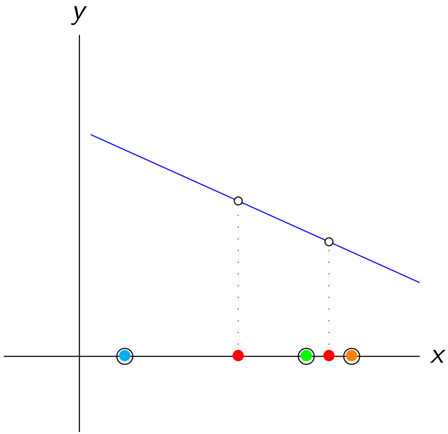
$$h\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) = \frac{1}{T}\sum_t h(x_t)$$

The simpler linear case



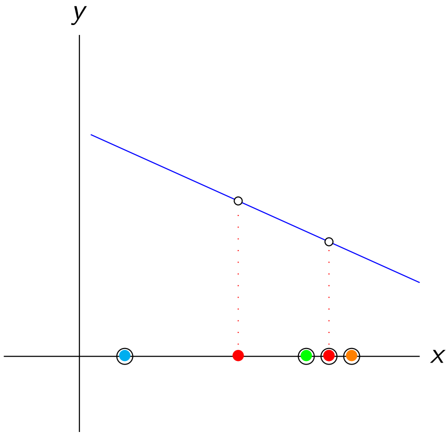
$$h\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) = \frac{1}{T}\sum_t h(x_t)$$

The simpler linear case



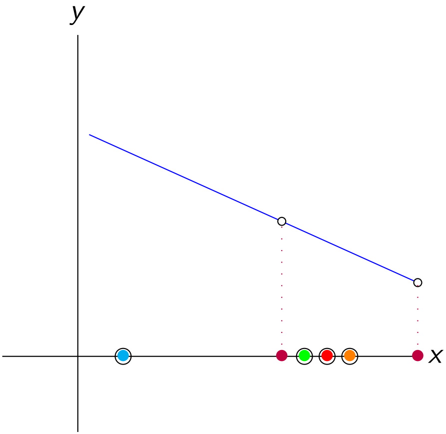
$$h\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) = \frac{1}{T}\sum_t h(x_t)$$

The simpler linear case



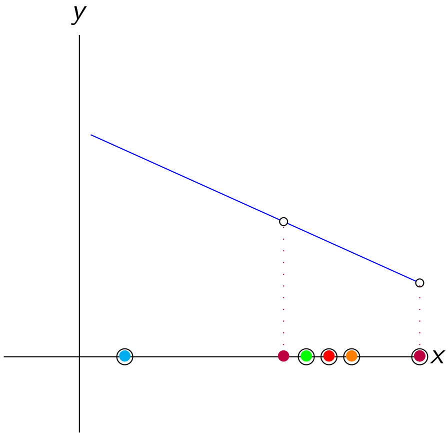
$$h\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) = \frac{1}{T}\sum_t h(x_t)$$

The simpler linear case



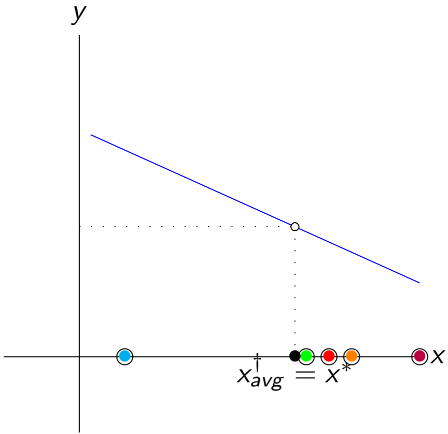
$$h\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) = \frac{1}{T}\sum_t h(x_t)$$

The simpler linear case



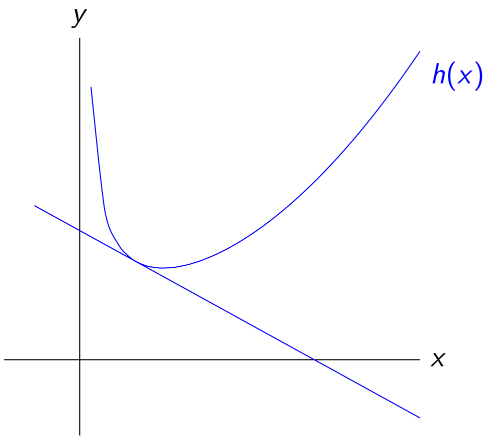
$$h\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) = \frac{1}{T}\sum_t h(x_t)$$

The simpler linear case



$$h\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) = \frac{1}{T}\sum_t h(x_t)$$

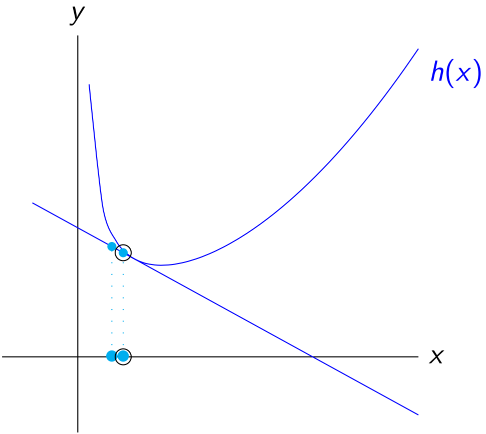
An optimistic algorithm



$$\ell(x^\dagger_{avg}) = \frac{1}{T}\sum_t \ell(x^\dagger_t) \le \frac{1}{T}\sum_t \ell(x^*_t) = \ell(x^*) \le h(x^*)$$

Upper bound on regret: $h(x^\dagger_{avg}) - \ell(x^\dagger_{avg})$

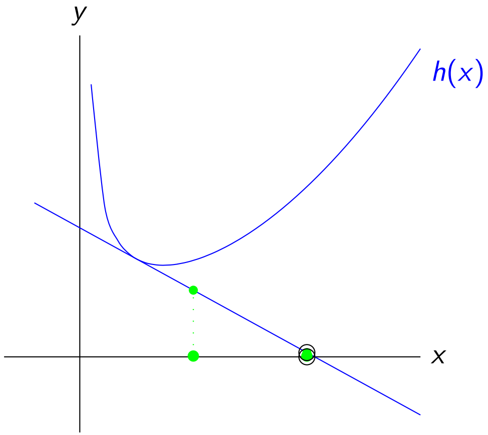
An optimistic algorithm



$$\ell(x^{\dagger}_{avg}) = \frac{1}{T}\sum_{t}\ell(x^{\dagger}_{t}) \le \frac{1}{T}\sum_{t}\ell(x^{*}_{t}) = \ell(x^{*}) \le h(x^{*})$$

Upper bound on regret: $h(x^{\dagger}_{avg}) - \ell(x^{\dagger}_{avg})$

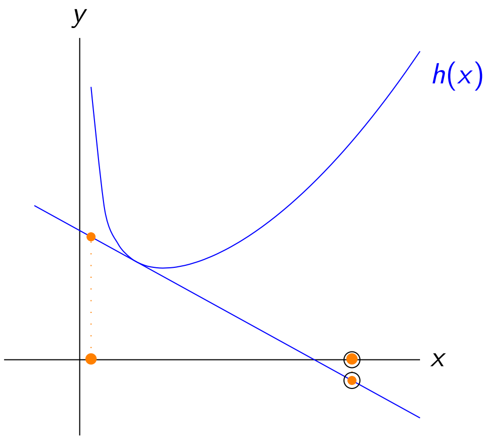
An optimistic algorithm



$$\ell(x^{\dagger}_{avg}) = \frac{1}{T}\sum_t \ell(x^{\dagger}_t) \leq \frac{1}{T}\sum_t \ell(x^{*}_t) = \ell(x^{*}) \leq h(x^{*})$$

Upper bound on regret: $h(x^{\dagger}_{avg}) - \ell(x^{\dagger}_{avg})$

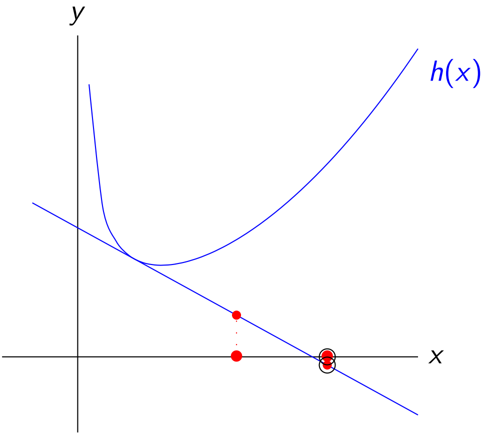
An optimistic algorithm



$$\ell(x^{\dagger}_{avg}) = \frac{1}{T}\sum_t \ell(x^{\dagger}_t) \leq \frac{1}{T}\sum_t \ell(x^{*}_t) = \ell(x^*) \leq h(x^*)$$

Upper bound on regret: $h(x^{\dagger}_{avg}) - \ell(x^{\dagger}_{avg})$

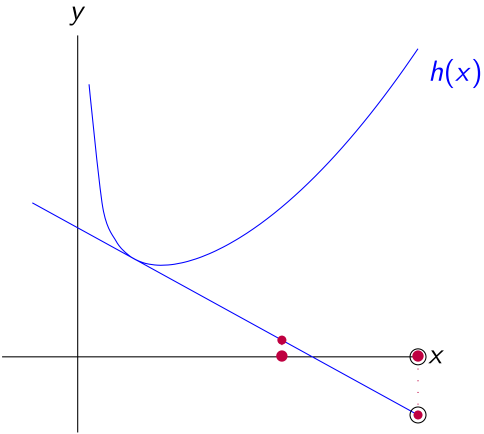
An optimistic algorithm



$$\ell(x^\dagger_{avg}) = \frac{1}{T}\sum_t \ell(x^\dagger_t) \le \frac{1}{T}\sum_t \ell(x^*_t) = \ell(x^*) \le h(x^*)$$

Upper bound on regret: $h(x^\dagger_{avg}) - \ell(x^\dagger_{avg})$

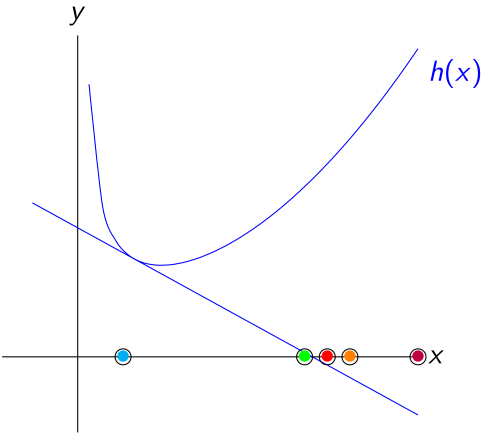
An optimistic algorithm



$$\ell(x^{\dagger}_{avg}) = \frac{1}{T}\sum_t \ell(x^{\dagger}_t) \leq \frac{1}{T}\sum_t \ell(x^{*}_t) = \ell(x^{*}) \leq h(x^{*})$$

Upper bound on regret: $h(x^{\dagger}_{avg}) - \ell(x^{\dagger}_{avg})$

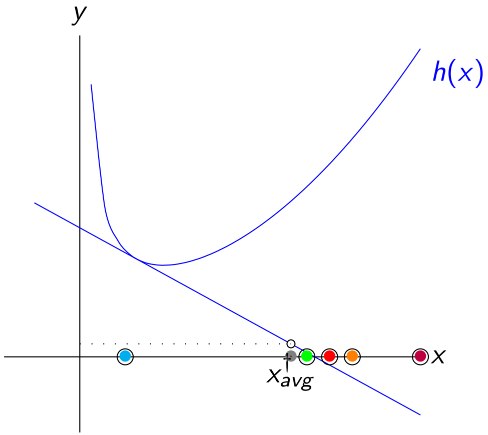
An optimistic algorithm



$$\ell(x^\dagger_{avg}) = \frac{1}{T}\sum_t \ell(x^\dagger_t) \le \frac{1}{T}\sum_t \ell(x^*_t) = \ell(x^*) \le h(x^*)$$

Upper bound on regret: $h(x^\dagger_{avg}) - \ell(x^\dagger_{avg})$

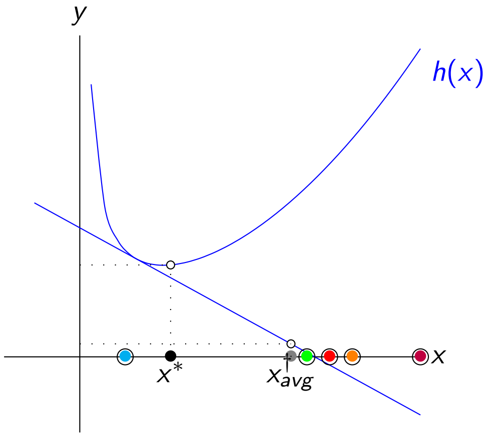
An optimistic algorithm



$$\ell(x^\dagger_{avg}) = \frac{1}{T}\sum_t \ell(x^\dagger_t) \le \frac{1}{T}\sum_t \ell(x^*_t) = \ell(x^*) \le h(x^*)$$

Upper bound on regret: $h(x^\dagger_{avg}) - \ell(x^\dagger_{avg})$

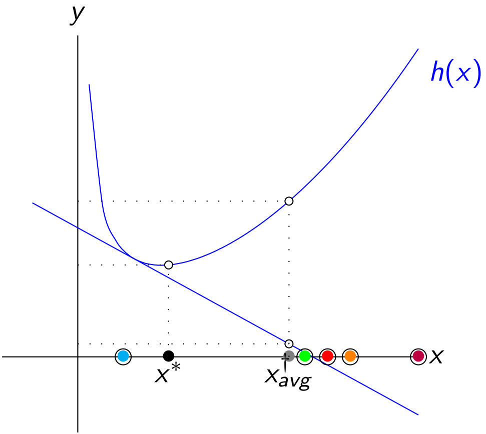
An optimistic algorithm



$$\ell(x^\dagger_{avg}) = \frac{1}{T}\sum_t \ell(x^\dagger_t) \le \frac{1}{T}\sum_t \ell(x^*_t) = \ell(x^*) \le h(x^*)$$

Upper bound on regret: $h(x^\dagger_{avg}) - \ell(x^\dagger_{avg})$

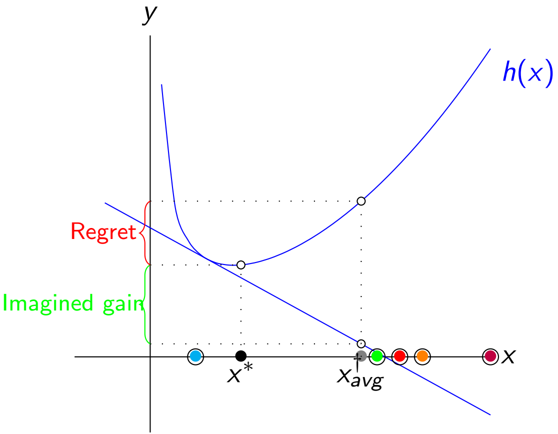
An optimistic algorithm



$$\ell(x^\dagger_{avg}) = \frac{1}{T}\sum_t \ell(x^\dagger_t) \le \frac{1}{T}\sum_t \ell(x^*_t) = \ell(x^*) \le h(x^*)$$

Upper bound on regret: $h(x^\dagger_{avg}) - \ell(x^\dagger_{avg})$

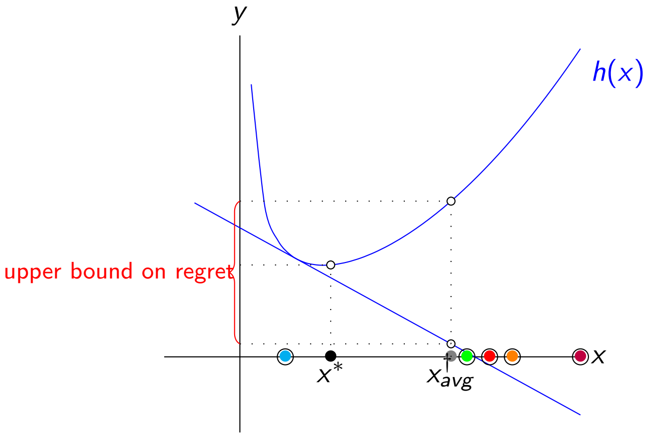
An optimistic algorithm



$$\ell(x^\dagger_{avg}) = \frac{1}{T}\sum_t \ell(x^\dagger_t) \leq \frac{1}{T}\sum_t \ell(x^*_t) = \ell(x^*) \leq h(x^*)$$

Upper bound on regret: $h(x^\dagger_{avg}) - \ell(x^\dagger_{avg})$

An optimistic algorithm

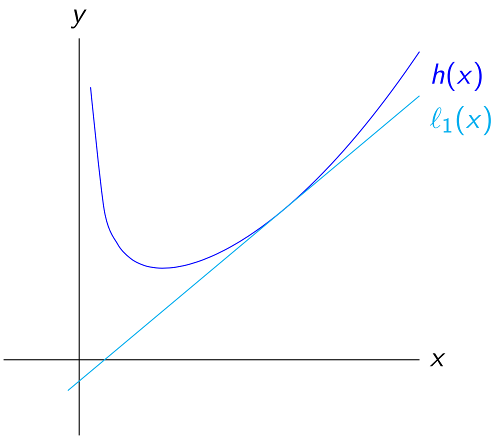


$$\ell(x^\dagger_{avg}) = \frac{1}{T}\sum_t \ell(x^\dagger_t) \le \frac{1}{T}\sum_t \ell(x^*_t) = \ell(x^*) \le h(x^*)$$

Upper bound on regret: $h(x^\dagger_{avg}) - \ell(x^\dagger_{avg})$

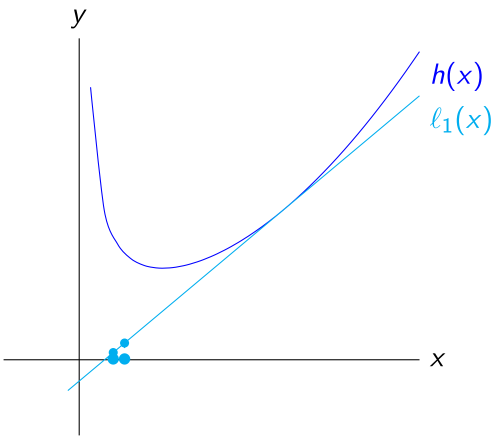
- If $x^{\dagger}_{avg}$ was known, tangent at this point would be a linear function with 0 gap: $\ell(x^{\dagger}_{avg}) = h(x^{\dagger}_{avg})$
- At time t, use current average as a guess for $x^{\dagger}_{avg}$ and take tangent (slope is gradient) at that point.
- Algorithm that uses a different tangent at every step.

Algorithm



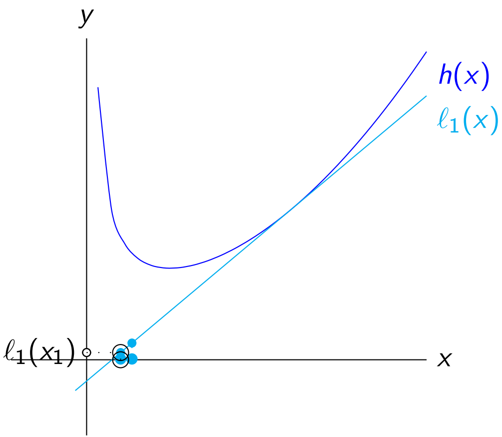
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



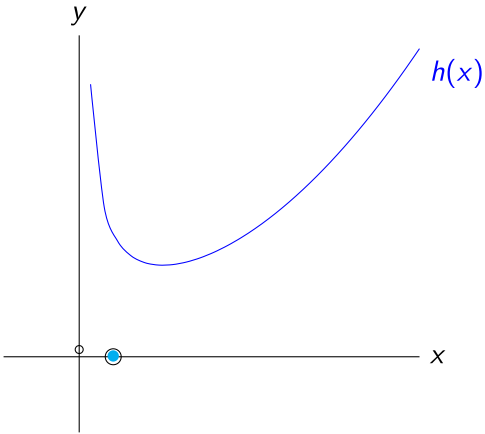
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



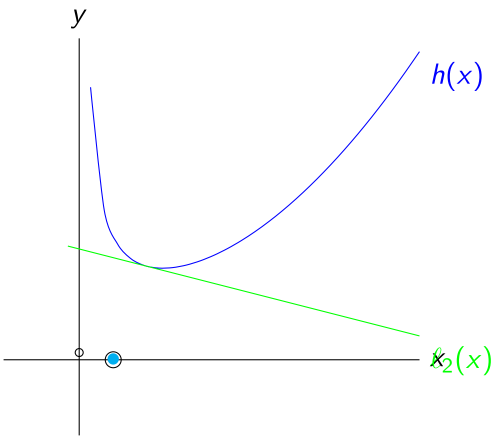
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



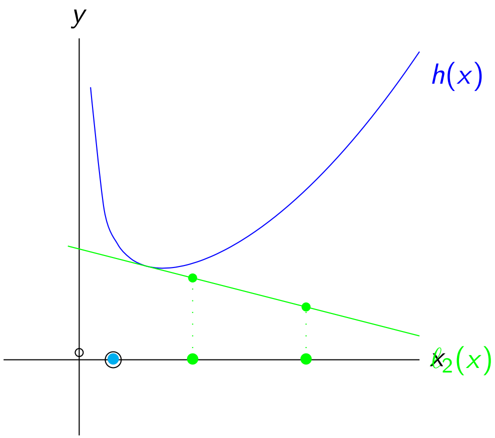
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \le \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \le h(x^*)$$

Algorithm



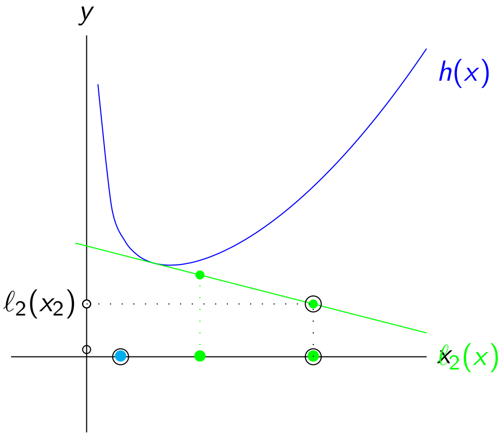
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



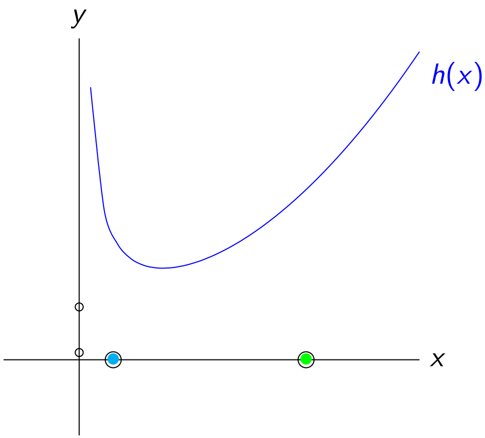
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



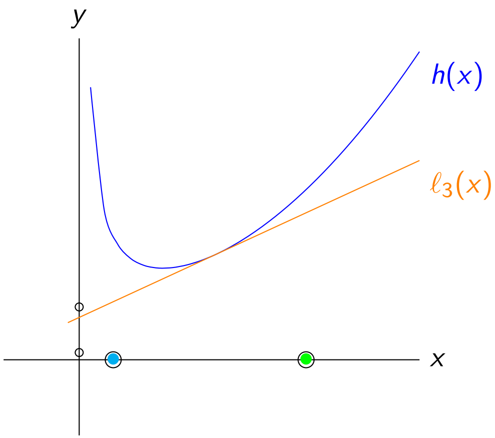
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



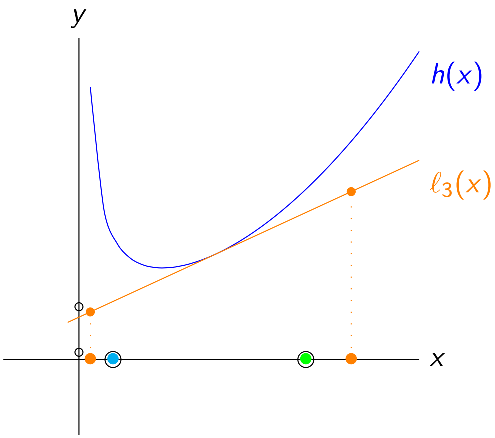
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



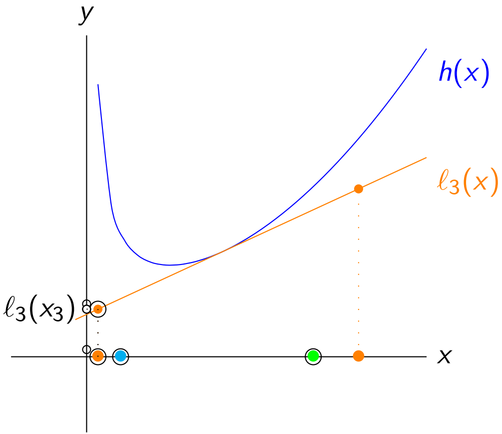
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



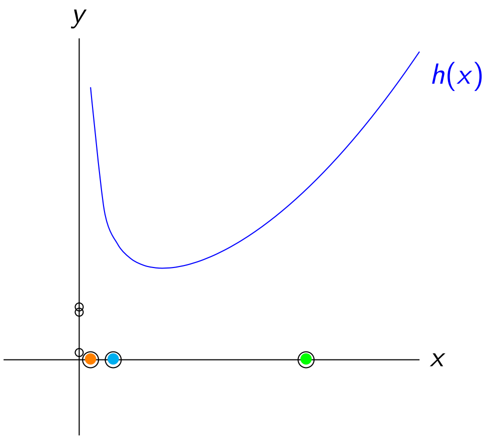
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



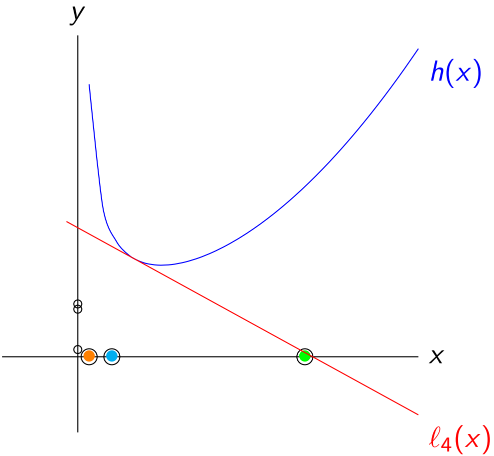
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



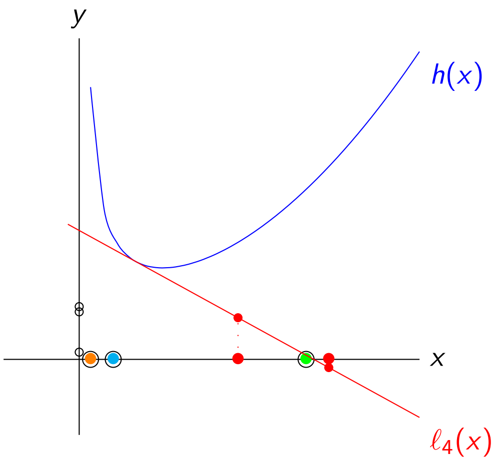
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



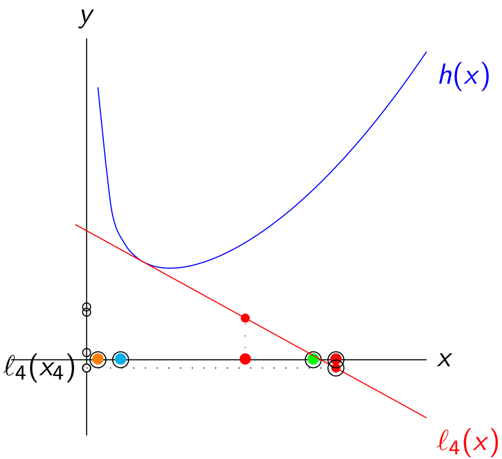
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \le \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \le h(x^*)$$

Algorithm



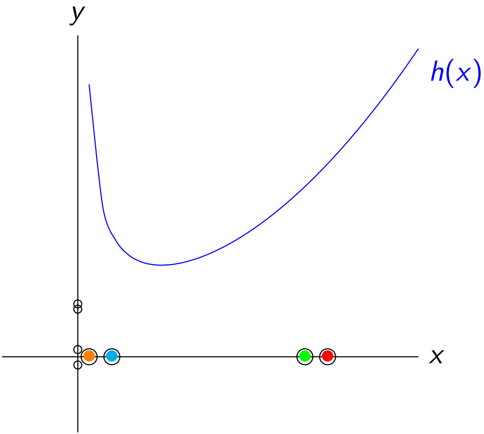
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



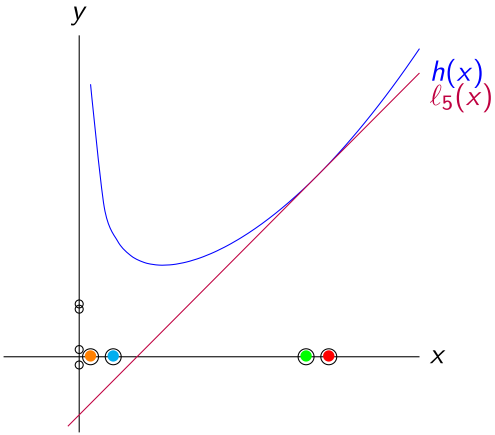
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



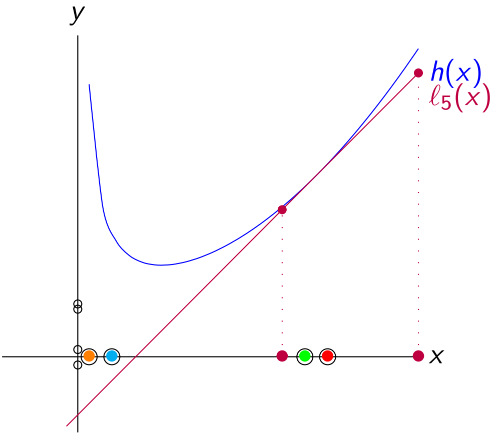
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



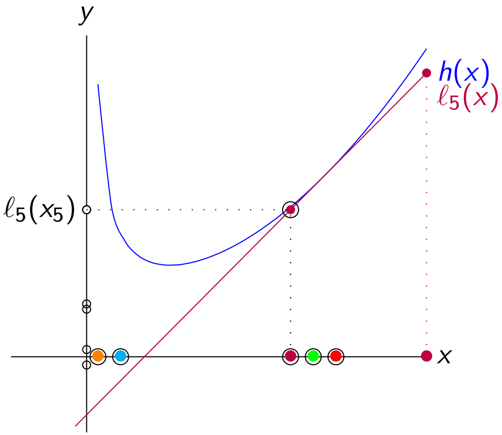
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



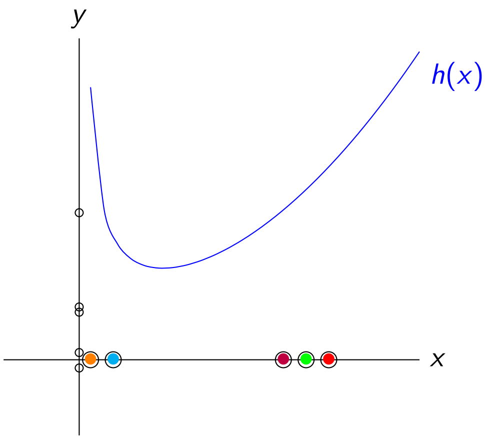
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



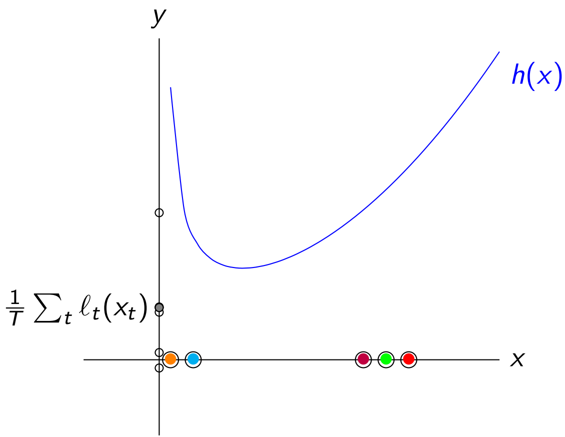
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



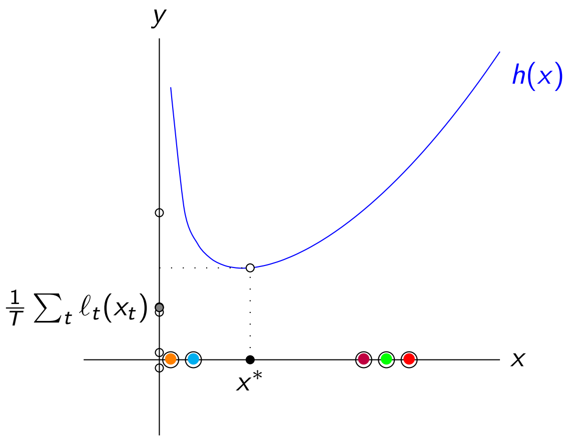
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



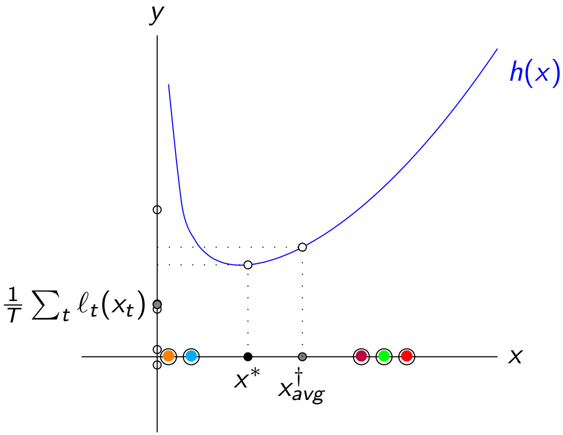
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



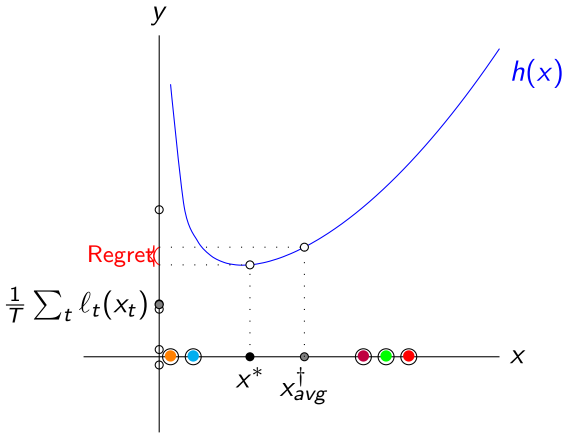
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



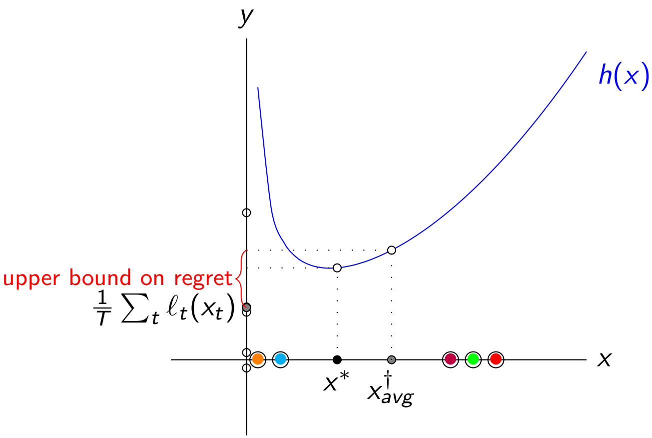
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



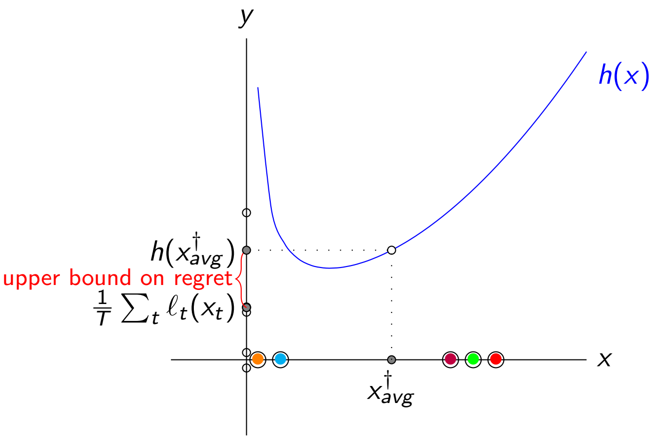
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \le \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \le h(x^*)$$

Algorithm



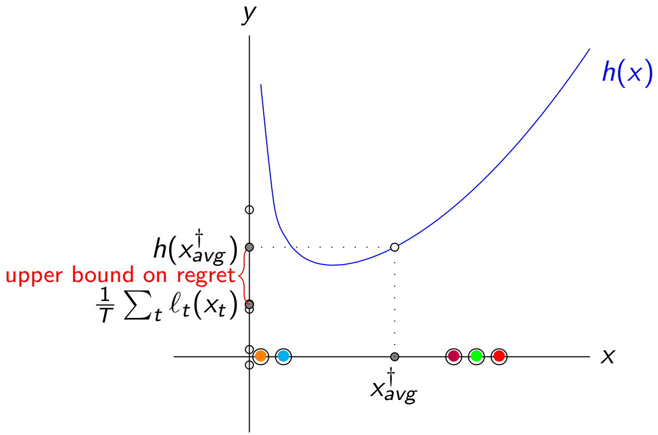
$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



$$\mathbb{E}[\ell_t(x_t^\dagger)|H_{t-1}] \leq \mathbb{E}[\ell_t(x_t^*)|H_{t-1}] = \ell_t(x^*) \leq h(x^*)$$

Algorithm



$$\mathbb{E}[\ell_t(x^\dagger_t)|H_{t-1}] \le \mathbb{E}[\ell_t(x^*_t)|H_{t-1}] = \ell_t(x^*) \le h(x^*)$$

Need to bound the gap $h(x^\dagger_{avg}) - \mathbb{E}[\frac{1}{T}\sum_t \ell_t(x^\dagger_t)]$

Need to bound the gap $h(x^{\dagger}_{avg}) - \mathbb{E}[\frac{1}{T}\sum_t \ell_t(x^{\dagger}_t)]$

- Let $\ell_t(x)$ be tangent at current average. For smooth and convex h,

$$\ell_t(x) := h(x_{avg,t-1}) + \nabla h(x_{avg,t-1})(x - x_{avg,t-1}) \leq h(x)$$

Need to bound the gap $h(x^{\dagger}_{avg}) - \mathbb{E}[\frac{1}{T}\sum_t \ell_t(x^{\dagger}_t)]$

- Let $\ell_t(x)$ be tangent at current average. For smooth and convex h,

$$\ell_t(x) := h(x_{avg,t-1}) + \nabla h(x_{avg,t-1})(x - x_{avg,t-1}) \leq h(x)$$

- Assuming β-smoothness,

$$h(x) \leq h(x_{avg,t-1}) + \nabla h(x_{avg,t-1})(x - x_{avg,t-1}) + \frac{\beta}{2t^2}$$

Need to bound the gap $h(x^\dagger_{avg}) - \mathbb{E}[\frac{1}{T}\sum_t \ell_t(x^\dagger_t)]$

- Let $\ell_t(x)$ be tangent at current average. For smooth and convex h,

$$\ell_t(x) := h(x_{avg,t-1}) + \nabla h(x_{avg,t-1})(x - x_{avg,t-1}) \leq h(x)$$

- Assuming β-smoothness,

$$h(x) \leq h(x_{avg,t-1}) + \nabla h(x_{avg,t-1})(x - x_{avg,t-1}) + \frac{\beta}{2t^2}$$

- Applying smoothness property to $x = x_{avg,t}$, we have a lower bound on $\ell_t(x^\dagger_t)$:

$$\frac{1}{t}\ell_t(x^\dagger_t) \geq h(x_{avg,t}) - \frac{(t-1)}{t}h(x_{avg,t}) - \frac{\beta}{2t^2}$$

Need to bound the gap $h(x^{\dagger}_{avg}) - \mathbb{E}[\frac{1}{T}\sum_t \ell_t(x^{\dagger}_t)]$

- Let $\ell_t(x)$ be tangent at current average. For smooth and convex h,

$$\ell_t(x) := h(x_{avg,t-1}) + \nabla h(x_{avg,t-1})(x - x_{avg,t-1}) \leq h(x)$$

- Assuming β-smoothness,

$$h(x) \leq h(x_{avg,t-1}) + \nabla h(x_{avg,t-1})(x - x_{avg,t-1}) + \frac{\beta}{2t^2}$$

- Applying smoothness property to $x = x_{avg,t}$, we have a lower bound on $\ell_t(x^{\dagger}_t)$:

$$\frac{1}{t}\ell_t(x^{\dagger}_t) \geq h(x_{avg,t}) - \frac{(t-1)}{t}h(x_{avg,t}) - \frac{\beta}{2t^2}$$

- Summing up for $t = 1, \ldots, T$ gives $O(\frac{\beta}{T})$ bound on $h(x_{avg}) - \frac{1}{t}\sum_t \ell_t(x^{\dagger}_t)$. For convex but non-smooth functions, bound degrades to $\tilde{O}\left(1/\sqrt{T}\right)$.

Algorithm Outline

Algorithm 1 Algorithm for minimizing $h(\frac{1}{T}\sum_{t=1}^{T} W\mathbf{x}_{t,a_t})$, with *known* W.

for all $t = 1 \ldots T$ **do**

Observe $\mathbf{x}_{t,a}$ for all $a \in A_t$.

Guess $\ell_t(\cdot)$.

$$a_t := \arg\min_{a \in A_t} \ell_t(W\mathbf{x}_{t,a}).$$

end for

Note:

- Optimistic guess: $\ell_t(W\mathbf{x}_{t,a})$ lower bounds $h(W\mathbf{x}_{t,a})$
- Regret bounded by the gap at played arms:

$$h(\frac{1}{T}\sum_{t} W\mathbf{x}_{t,a_t}) - \frac{1}{T}\sum_{t} \ell_t(W\mathbf{x}_{t,a_t}) \leq \tilde{O}\left(\frac{\log(d)}{\sqrt{T}}\right)$$

Handling unknown W

Replace W by its optimistic estimate: in this case lower confidence bound.

Algorithm 2 Algorithm for unknown W

for all $t = 1 \dots T$ **do**

Observe $\mathbf{x}_{t,a}$ for all $a \in A_t$.

For all $a \in A_t$, compute lower confidence bound (LCB) $\tilde{W}_{t,a}$ as in linear contextual MAB.

Guess tangent $\ell_t(\cdot)$.

Play arm

$$a_t := \arg\min_{a \in A_t} \ell_t(\tilde{W}_{t,a}\mathbf{x}_{t,a}).$$

Observe $\mathbf{v}_t := \mathbf{v}_{t,a_t}$, with expected value $W\mathbf{x}_{t,a_t}$

end for

Additional term added to regret:

$$\left(\frac{1}{T}\sum_t \ell_t(W\mathbf{x}_{t,a_t}) - \frac{1}{T}\sum_t \ell_t(\tilde{W}\mathbf{x}_{t,a_t})\right) \leq \tilde{O}\left(\frac{d}{\sqrt{T}}\right)$$

Further difficulties

So far: algorithm for minimizing a convex function on average decision. How to handle "maximize concave function given constraint set S"

- "Constraints only" case can be handled by posing problem as "minimize distance from the constraint set"
- If OPT known, convert objective into constraint.
- Estimating OPT, requires further exploration, incurring suboptimal regret $dT^{-1/3}$
- Getting $d/\sqrt{T}$ regret (or a tighter lower bound) is open

Thank You