Multi-armed bandits in dynamic pricing

Arnoud den Boer

University of Twente, Centrum Wiskunde & Informatica Amsterdam

Lancaster, January 11, 2016

Dynamic pricing

- A firm sells a product, with abundant inventory, during $T \in \mathbb{N}$ discrete time periods.

Dynamic pricing

- A firm sells a product, with abundant inventory, during $T \in \mathbb{N}$ discrete time periods.
- Each period $t = 1, \dots, T$:
  (i) choose selling price p_t;

Dynamic pricing

- A firm sells a product, with abundant inventory, during $T \in \mathbb{N}$ discrete time periods.
- Each period $t = 1, \ldots, T$:
  (i) choose selling price p_t;
  (ii) observe demand

  $$d_t = \theta_1 + \theta_2 p_t + \epsilon_t,$$

  where $\theta = (\theta_1, \theta_2)$ are unknown parameters in known set Θ,
  ϵ_t unobservable random disturbance term;

Dynamic pricing

- A firm sells a product, with abundant inventory, during $T \in \mathbb{N}$ discrete time periods.
- Each period $t = 1, \ldots, T$:
  (i) choose selling price p_t;
  (ii) observe demand

  $$d_t = \theta_1 + \theta_2 p_t + \epsilon_t,$$

  where $\theta = (\theta_1, \theta_2)$ are unknown parameters in known set Θ,
  ϵ_t unobservable random disturbance term;
  (iii) collect revenue $p_t d_t$.

Dynamic pricing

- A firm sells a product, with abundant inventory, during $T \in \mathbb{N}$ discrete time periods.
- Each period $t = 1, \ldots, T$:
  (i) choose selling price p_t;
  (ii) observe demand

  $$d_t = \theta_1 + \theta_2 p_t + \epsilon_t,$$

  where $\theta = (\theta_1, \theta_2)$ are unknown parameters in known set Θ, ϵ_t unobservable random disturbance term;
  (iii) collect revenue $p_t d_t$.
- Which non-anticipating prices $p_1, \ldots, p_T$ maximize cumulative expected revenue $\min_{\theta \in \Theta} \mathbb{E}\left[\sum_{t=1}^{T} p_t d_t \right]$?

Dynamic pricing

- A firm sells a product, with abundant inventory, during $T \in \mathbb{N}$ discrete time periods.
- Each period $t = 1, \ldots, T$:
  (i) choose selling price p_t;
  (ii) observe demand

  $$d_t = \theta_1 + \theta_2 p_t + \epsilon_t,$$

  where $\theta = (\theta_1, \theta_2)$ are unknown parameters in known set Θ,
  ϵ_t unobservable random disturbance term;
  (iii) collect revenue $p_t d_t$.
- Which non-anticipating prices $p_1, \ldots, p_T$ maximize cumulative expected revenue $\min_{\theta \in \Theta} \mathbb{E}\left[\sum_{t=1}^{T} p_t d_t \right]$?

Intractable problem

Myopic pricing

An intuitive solution:

- Choose arbitrary initial prices $p_1 \neq p_2$.

Myopic pricing

An intuitive solution:

- Choose arbitrary initial prices $p_1 \neq p_2$.
- For each $t \geq 2$:
  (i) determine LS estimate $\hat{\theta}_t$ of θ, based on available sales data;
  (ii) set

$$p_{t+1} = \arg\max_p p \cdot (\hat{\theta}_{t1} + \hat{\theta}_{t2} p) \quad \text{perceived optimal decision}$$

Myopic pricing

An intuitive solution:

- Choose arbitrary initial prices $p_1 \neq p_2$.
- For each $t \geq 2$:
  (i) determine LS estimate $\hat{\theta}_t$ of θ, based on available sales data;
  (ii) set

$$p_{t+1} = \arg\max_p p \cdot (\hat{\theta}_{t1} + \hat{\theta}_{t2} p) \quad \text{perceived optimal decision}$$

- 'Always choose the perceived optimal action'.

Convergence

Does $\hat{\theta}_t$ converge to θ as $t \to \infty$?

Convergence

Does $\hat{\theta}_t$ converge to θ as $t \to \infty$?

No

It *seems* that $\hat{\theta}_t$ always converges, but w.p. zero to the true θ.
Open problem.

Convergence

Does $\hat{\theta}_t$ converge to θ as $t \to \infty$?

No

It *seems* that $\hat{\theta}_t$ always converges, but w.p. zero to the true θ.
Open problem.

Caused by the prevalence of indeterminate equilibria:
Parameter estimates such that the *true* expected demand at the myopic optimal price equals the *predicted* expected demand.

Indeterminate equilibria

If $\hat{\theta}$ suff. close to θ, then $\arg\max_p p \cdot (\hat{\theta}_1 + \hat{\theta}_2 p) = -\hat{\theta}_1/(2\hat{\theta}_2)$.

Then:

$$\text{`True' expected demand: } \theta_1 + \theta_2 \frac{-\hat{\theta}_1}{2\hat{\theta}_2}. \tag{1}$$

$$\text{`Predicted' expected demand: } \hat{\theta}_1 + \hat{\theta}_2 \frac{-\hat{\theta}_1}{2\hat{\theta}_2}. \tag{2}$$

Indeterminate equilibria

If $\hat{\theta}$ suff. close to θ, then $\arg\max_p p \cdot (\hat{\theta}_1 + \hat{\theta}_2 p) = -\hat{\theta}_1/(2\hat{\theta}_2)$.

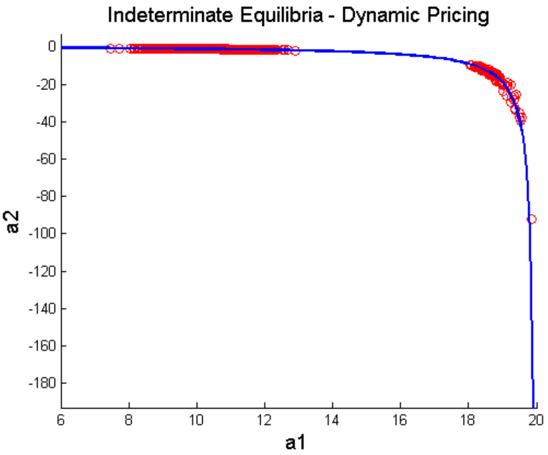
Then:

$$\text{'True' expected demand: } \theta_1 + \theta_2 \frac{-\hat{\theta}_1}{2\hat{\theta}_2}. \tag{1}$$

$$\text{'Predicted' expected demand: } \hat{\theta}_1 + \hat{\theta}_2 \frac{-\hat{\theta}_1}{2\hat{\theta}_2}. \tag{2}$$

If (1) equals (2), then $\hat{\theta}$ is an IE.
Model output 'confirms' correctness of the (incorrect) estimates.

Indeterminate equilibria: example



Back to original problem

Which non-anticipating prices $p_1, \ldots, p_T$ maximize

$$\min_{\theta \in \Theta} \mathbb{E}\Big[\sum_{t=1}^{T} p_t d_t\Big],$$

or, equivalently, minimize the Regret(T)

$$\max_{\theta \in \Theta} \mathbb{E}\Big[T \cdot \max_{p} p \cdot (\theta_1 + \theta_2 p) - \sum_{t=1}^{T} p_t d_t\Big]$$

Back to original problem

Which non-anticipating prices $p_1, \ldots, p_T$ maximize

$$\min_{\theta \in \Theta} \mathbb{E}\Big[\sum_{t=1}^{T} p_t d_t\Big],$$

or, equivalently, minimize the Regret(T)

$$\max_{\theta \in \Theta} \mathbb{E}\Big[T \cdot \max_{p} p \cdot (\theta_1 + \theta_2 p) - \sum_{t=1}^{T} p_t d_t\Big]$$

- Exact solution intractable

Back to original problem

Which non-anticipating prices $p_1, \ldots, p_T$ maximize

$$\min_{\theta \in \Theta} \mathbb{E}\Big[\sum_{t=1}^{T} p_t d_t \Big],$$

or, equivalently, minimize the Regret(T)

$$\max_{\theta \in \Theta} \mathbb{E}\Big[T \cdot \max_{p} p \cdot (\theta_1 + \theta_2 p) - \sum_{t=1}^{T} p_t d_t \Big]$$

- Exact solution intractable
- Myopic pricing not optimal

Back to original problem

Which non-anticipating prices $p_1, \ldots, p_T$ maximize

$$\min_{\theta \in \Theta} \mathbb{E}\Big[\sum_{t=1}^{T} p_t d_t\Big],$$

or, equivalently, minimize the Regret(T)

$$\max_{\theta \in \Theta} \mathbb{E}\Big[T \cdot \max_{p} p \cdot (\theta_1 + \theta_2 p) - \sum_{t=1}^{T} p_t d_t\Big]$$

- Exact solution intractable
- Myopic pricing not optimal
- Let's find asymptotically optimal policies: smallest growth rate of Regret(T) in T.

Asymptotically optimal policy

Important observation: Variation in controls $\Rightarrow$ better estimates.

Asymptotically optimal policy

Important observation: Variation in controls $\Rightarrow$ better estimates.

$$\left|\left|\hat{\theta}_t - \theta\right|\right|^2 = O\left(\frac{\log t}{t\mathrm{Var}(p_1, \ldots, p_t)}\right) \text{ a.s.}$$

Lai and Wei, Annals of Statistics, 1982.

Asymptotically optimal policy

Important observation: Variation in controls $\Rightarrow$ better estimates.

$$\left|\left|\hat{\theta}_t - \theta\right|\right|^2 = O\left(\frac{\log t}{t\mathrm{Var}(p_1, \ldots, p_t)}\right) \text{ a.s.}$$

Lai and Wei, Annals of Statistics, 1982.

To ensure convergence of $\hat{\theta}_t$, some amount of experimentation is necessary.

Asymptotically optimal policy

Important observation: Variation in controls $\Rightarrow$ better estimates.

$$\left|\left|\hat{\theta}_t - \theta\right|\right|^2 = O\left(\frac{\log t}{t\mathrm{Var}(p_1, \ldots, p_t)}\right) \text{ a.s.}$$

Lai and Wei, Annals of Statistics, 1982.

To ensure convergence of $\hat{\theta}_t$, some amount of experimentation is necessary. But, not *too* much.

‘Controlled Variance pricing’

- Choose arbitrary initial prices $p_1 \neq p_2$.
- For each $t \geq 2$:
  (i) determine LS estimate $\hat{\theta}_t$ of θ, based on available sales data;
  (ii) set

$$p_{t+1} = \arg\max_p p \cdot (\hat{\theta}_{t1} + \hat{\theta}_{t2} p)$$

'Controlled Variance pricing'

- Choose arbitrary initial prices $p_1 \neq p_2$.
- For each $t \geq 2$:
  (i) determine LS estimate $\hat{\theta}_t$ of θ, based on available sales data;
  (ii) set

$$p_{t+1} = \arg\max_p p \cdot (\hat{\theta}_{t1} + \hat{\theta}_{t2} p) \quad \text{perceived optimal decision}$$

‘Controlled Variance pricing’

- Choose arbitrary initial prices $p_1 \neq p_2$.
- For each $t \geq 2$:
  (i) determine LS estimate $\hat{\theta}_t$ of θ, based on available sales data;
  (ii) set

$$p_{t+1} = \arg\max_p p \cdot (\hat{\theta}_{t1} + \hat{\theta}_{t2} p) \quad \text{perceived optimal decision}$$

$$\text{s.t. } t \cdot \mathrm{Var}(p_1, \ldots, p_{t+1}) \geq f(t), \quad \text{‘information constraint’}$$

'Controlled Variance pricing'

- Choose arbitrary initial prices $p_1 \neq p_2$.
- For each $t \geq 2$:
  (i) determine LS estimate $\hat{\theta}_t$ of θ, based on available sales data;
  (ii) set

$$p_{t+1} = \arg\max_p p \cdot (\hat{\theta}_{t1} + \hat{\theta}_{t2} p) \quad \text{perceived optimal decision}$$

$$\text{s.t. } t \cdot \mathrm{Var}(p_1, \ldots, p_{t+1}) \geq f(t), \quad \text{`information constraint'}$$

  for some increasing $f : \mathbb{N} \to (0, \infty)$.

'Controlled Variance pricing'

- Choose arbitrary initial prices $p_1 \neq p_2$.
- For each $t \geq 2$:
  (i) determine LS estimate $\hat{\theta}_t$ of θ, based on available sales data;
  (ii) set

$$p_{t+1} = \arg\max_p p \cdot (\hat{\theta}_{t1} + \hat{\theta}_{t2} p) \quad \text{perceived optimal decision}$$

$$\text{s.t. } t \cdot \mathrm{Var}(p_1, \ldots, p_{t+1}) \geq f(t), \quad \text{'information constraint'}$$

  for some increasing $f : \mathbb{N} \to (0, \infty)$.
- 'Always choose the perceived optimal action that induces sufficient experimentation'.

‘Controlled Variance pricing’ - performance

- Regret$(T) = O\left(f(T) + \sum_{t=1}^{T} \frac{\log t}{f(t)}\right)$.

‘Controlled Variance pricing’ - performance

- $\text{Regret}(T) = O\left(f(T) + \sum_{t=1}^{T} \frac{\log t}{f(t)}\right)$.
- f balances between exploration and exploitation.

‘Controlled Variance pricing’ - performance

- $\text{Regret}(T) = O\left(f(T) + \sum_{t=1}^{T} \frac{\log t}{f(t)}\right)$.
- f balances between exploration and exploitation.
- Optimal f gives $\text{Regret}(T) = O(\sqrt{T \log T})$.

'Controlled Variance pricing' - performance

- $\text{Regret}(T) = O\left(f(T) + \sum_{t=1}^{T} \frac{\log t}{f(t)}\right)$.
- f balances between exploration and exploitation.
- Optimal f gives $\text{Regret}(T) = O(\sqrt{T \log T})$.
- No policy beats $\sqrt{T}$.

‘Controlled Variance pricing’ - performance

- $\text{Regret}(T) = O\left(f(T) + \sum_{t=1}^{T} \frac{\log t}{f(t)}\right)$.
- f balances between exploration and exploitation.
- Optimal f gives $\text{Regret}(T) = O(\sqrt{T \log T})$.
- No policy beats $\sqrt{T}$.

Thus, you can characterize asymptotically (near)-optimal amount of experimentation.

‘Controlled Variance pricing’ - performance

- $\text{Regret}(T) = O\left(f(T) + \sum_{t=1}^{T} \frac{\log t}{f(t)}\right)$.
- f balances between exploration and exploitation.
- Optimal f gives $\text{Regret}(T) = O(\sqrt{T \log T})$.
- No policy beats $\sqrt{T}$.

Thus, you can characterize asymptotically (near)-optimal amount of experimentation.

(the optimal ‘constant’ is not yet known, in general).

Extension: multiple products

K products: price vector $\mathbf{p}_t = (p_t(1), \ldots, p_t(K))^\top$,
demand vector $\mathbf{d}_t = \boldsymbol{\theta} \begin{pmatrix} 1 \\ \mathbf{p}_t \end{pmatrix} + \boldsymbol{\epsilon}$, matrix $\boldsymbol{\theta}$, noise-vector $\boldsymbol{\epsilon}$.

Extension: multiple products

K products: price vector $\mathbf{p}_t = (p_t(1), \ldots, p_t(K))^\top$,
demand vector $\mathbf{d}_t = \boldsymbol{\theta} \begin{pmatrix} 1 \\ \mathbf{p}_t \end{pmatrix} + \boldsymbol{\epsilon}$, matrix $\boldsymbol{\theta}$, noise-vector $\boldsymbol{\epsilon}$.

Convergence rates of LS-estimator:

$$\left|\left|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\right|\right|^2 = O\left(\frac{\log t}{\lambda_{\min}(t)}\right) \text{ a.s.,}$$

where $\lambda_{\min}(t)$ is the smallest eigenvalue of the information matrix

$$\sum_{i=1}^{t} \begin{pmatrix} 1 & \mathbf{p}_i^\top \\ \mathbf{p}_i & \mathbf{p}_i \mathbf{p}_i^\top \end{pmatrix}$$

Extension: multiple products

Same type of policy:

$$\mathbf{p}_{t+1} = \arg\max_{\mathbf{p}} \mathbf{p}^\top \hat{\boldsymbol{\theta}}_t \begin{pmatrix} 1 \\ \mathbf{p} \end{pmatrix}$$

Extension: multiple products

Same type of policy:

$$\mathbf{p}_{t+1} = \arg\max_{\mathbf{p}} \mathbf{p}^\top \hat{\boldsymbol{\theta}}_t \begin{pmatrix} 1 \\ \mathbf{p} \end{pmatrix} \quad \text{perceived optimal decision}$$

Extension: multiple products

Same type of policy:

$$\mathbf{p}_{t+1} = \arg\max_{\mathbf{p}} \mathbf{p}^\top \hat{\boldsymbol{\theta}}_t \begin{pmatrix} 1 \\ \mathbf{p} \end{pmatrix} \quad \text{perceived optimal decision}$$

$$\text{s.t. } \lambda_{\min}(t+1) \geq f(t), \quad \text{'information constraint'}$$

Extension: multiple products

Same type of policy:

$$\mathbf{p}_{t+1} = \arg\max_{\mathbf{p}} \mathbf{p}^\top \hat{\boldsymbol{\theta}}_t \begin{pmatrix} 1 \\ \mathbf{p} \end{pmatrix} \quad \text{perceived optimal decision}$$

$$\text{s.t. } \lambda_{\min}(t+1) \geq f(t), \quad \text{`information constraint'}$$

for some increasing $f : \mathbb{N} \to (0, \infty)$.

Extension: multiple products

Same type of policy:

$$\mathbf{p}_{t+1} = \arg\max_{\mathbf{p}} \mathbf{p}^\top \hat{\boldsymbol{\theta}}_t \begin{pmatrix} 1 \\ \mathbf{p} \end{pmatrix} \quad \text{perceived optimal decision}$$

$$\text{s.t. } \lambda_{\text{min}}(t+1) \geq f(t), \quad \text{`information constraint'}$$

for some increasing $f : \mathbb{N} \to (0, \infty)$.

Problem: $\lambda_{\text{min}}(t+1)$ is a complicated object.

Extension: multiple products

Same type of policy:

$$\mathbf{p}_{t+1} = \arg\max_{\mathbf{p}} \mathbf{p}^\top \hat{\boldsymbol{\theta}}_t \begin{pmatrix} 1 \\ \mathbf{p} \end{pmatrix} \quad \text{perceived optimal decision}$$

$$\text{s.t. } \lambda_{\min}(t+1) \geq f(t), \quad \text{`information constraint'}$$

for some increasing $f : \mathbb{N} \to (0, \infty)$.

Problem: $\lambda_{\min}(t+1)$ is a complicated object.

Convertible to non-convex but tractable quadratic constraint.

Extension: multiple products

Same type of policy:

$$\mathbf{p}_{t+1} = \arg\max_{\mathbf{p}} \mathbf{p}^\top \hat{\boldsymbol{\theta}}_t \begin{pmatrix} 1 \\ \mathbf{p} \end{pmatrix} \quad \text{perceived optimal decision}$$
$$\text{s.t. } \lambda_{\min}(t+1) \geq f(t), \quad \text{`information constraint'}$$

for some increasing $f : \mathbb{N} \to (0, \infty)$.

Problem: $\lambda_{\min}(t+1)$ is a complicated object.

Convertible to non-convex but tractable quadratic constraint.

$\text{Regret}(T) = O\left(f(T) + \sum_{t=1}^{T} \frac{\log t}{f(t)}\right)$,
optimal f gives $\text{Regret}(T) = O(\sqrt{T \log T})$.

Many more extensions

- Non-linear demand functions (generalized linear models) $\mathbb{E}[D(p)] = h(\theta_1 + \theta_2 p)$;

Many more extensions

- Non-linear demand functions (generalized linear models) $\mathbb{E}[D(p)] = h(\theta_1 + \theta_2 p)$;
- Time-varying markets (how much data to use for inference?)

Many more extensions

- Non-linear demand functions (generalized linear models) $\mathbb{E}[D(p)] = h(\theta_1 + \theta_2 p)$;
- Time-varying markets (how much data to use for inference?)
- Strategic customer behavior (can you detect this from data?)

Many more extensions

- Non-linear demand functions (generalized linear models) $\mathbb{E}[D(p)] = h(\theta_1 + \theta_2 p)$;
- Time-varying markets (how much data to use for inference?)
- Strategic customer behavior (can you detect this from data?)
- Competition (repeated games with incomplete information? Mean field games with learning?)

Many more extensions

- Non-linear demand functions (generalized linear models) $\mathbb{E}[D(p)] = h(\theta_1 + \theta_2 p)$;
- Time-varying markets (how much data to use for inference?)
- Strategic customer behavior (can you detect this from data?)
- Competition (repeated games with incomplete information? Mean field games with learning?)

den Boer (2015) Surveys in Operations Research and Management Science 20(1)

Why a parametric demand model?

$$d_t = \theta_1 + \theta_2 p_t + \epsilon_t \ldots$$

Why a parametric demand model?

$$d_t = \theta_1 + \theta_2 p_t + \epsilon_t \ldots$$

- Preferred by price managers
- By smartly choosing experimentation prices converging to the optimal price, you can hedge against misspecified linear demand.

Can't this log-term be removed?

$$\text{Regret}(T) = O(\sqrt{T \log T})$$

- Convergence rates of LS estimators: not completely understood
- Does more data lead to better estimators?

Pricing airline tickets

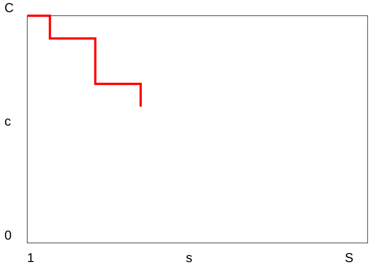
- Sell $C \in \mathbb{N}$ perishable products during (consecutive) selling season of $S \in \mathbb{N}$ periods

Pricing airline tickets

- Sell $C \in \mathbb{N}$ perishable products during (consecutive) selling season of $S \in \mathbb{N}$ periods
- Demand in period t is Bernoulli $h(\beta_0 + \beta_1 p_t)$, unknown β_0, β_1.
- Goal of the firm: maximize total expected revenue.

Full-information solution

If demand distribution known: Markov decision problem.



Optimal prices $\pi^*_\beta(c, s) \in [p_l, p_h]$ for each pair (c, s) of remaining inventory $c \in \{0, 1, \ldots, C\}$ and stage $s \in \{1, \ldots, S\}$.

Pricing airline tickets: incomplete information

Neglecting some technicalities, certainty-equivalent pricing performs well!

I.e., if in period t state is (c_t, s_t), use price $\pi^*_{\hat{\beta}_t}(c_t, s_t)$,

Pricing airline tickets: incomplete information

Neglecting some technicalities, certainty-equivalent pricing performs well!

I.e., if in period t state is (c_t, s_t), use price $\pi^*_{\hat{\beta}_t}(c_t, s_t)$,

Pricing airline tickets: endogenous learning

Reason for good performance: endogenous learning property

Pricing airline tickets: endogenous learning

Reason for good performance: endogenous learning property

- The optimal price $\pi^*_\beta(c, s)$ depends on marginal value of inventory
- This quantity changing throughout the selling season
- Thus, natural price dispersion if π^*_β is used

Pricing airline tickets: endogenous learning

Reason for good performance: endogenous learning property

- The optimal price $\pi^*_\beta(c, s)$ depends on marginal value of inventory
- This quantity changing throughout the selling season
- Thus, natural price dispersion if π^*_β is used
- By continuity arguments: price dispersion if $\hat{\beta}_t$ close to β, for all t in selling season

Pricing airline tickets: endogenous learning

Reason for good performance: endogenous learning property

- The optimal price $\pi^*_\beta(c, s)$ depends on marginal value of inventory
- This quantity changing throughout the selling season
- Thus, natural price dispersion if π^*_β is used
- By continuity arguments: price dispersion if $\hat{\beta}_t$ close to β, for all t in selling season

Endogenous learning causes fast converge of estimates:

$$E\left[\left|\left|\hat{\beta}(t) - \beta^{(0)}\right|\right|^2\right] = O\left(\frac{\log(t)}{t}\right)$$