Contextual Stochastic Bandits with Budget Constraints and Fairness Application

Gilles Stoltz

Laboratoire de mathématiques d'Orsay

Joint work with Evgenii Chzhen, Christophe Giraud, and Zhen Li

Contextual Stochastic Bandits with Budget Constraints and Fairness Application

1. Stochastic bandits
2. Contextual stochastic bandits
3. Contextual stochastic bandits with budget constraints
   - Application to fairness: small budgets

K–armed stochastic bandits

Simplest possible framework

K probability distributions $\nu_1, \ldots, \nu_K$ in a model $\mathcal{D}$
with expectations $\mu_1, \ldots, \mu_K$ $\longrightarrow$ $\mu^\star = \max_{a \in [K]} \mu_a$

At each round $t = 1, 2, \ldots,$

1. Statistician picks arm $A_t \in [K]$
2. She gets a reward Y_t drawn according to ν_{A_t}
3. This is the only feedback she receives

$\longrightarrow$ Exploration–exploitation dilemma
estimate the ν_a vs. get high rewards Y_t

Goal:
Maximize expected cumulative rewards $\longleftrightarrow$ Minimize regret

$$R_T = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = \sum_{a \in [K]} \bigl(\mu^\star - \mu_a\bigr)\,\mathbb{E}\bigl[N_a(T)\bigr]$$

$\longleftrightarrow$ Control the $\mathbb{E}\bigl[N_a(T)\bigr]$ for suboptimal arms a

Setting:
Distributions $\nu_1, \ldots, \nu_K$ with expectations $\mu_1, \ldots, \mu_K$
At each round $t \geqslant 1$, pick arm $A_t \in [K]$, get and observe $Y_t \sim \nu_{A_t}$

Proof of the rewriting of regret

Tower rule: $\mathbb{E}[Y_t \,|\, A_t] = \mu_{A_t}$ thus $\mathbb{E}[Y_t] = \mathbb{E}[\mu_{A_t}]$

$$R_T = \sum_{t=1}^{T} \bigl(\mu^\star - \mathbb{E}[Y_t]\bigr) = \sum_{t=1}^{T} \bigl(\mu^\star - \mathbb{E}[\mu_{A_t}]\bigr)$$

$$= \sum_{t=1}^{T} \sum_{a \in [K]} \bigl(\mu^\star - \mu_a\bigr)\, \mathbb{E}\bigl[\mathbb{I}_{\{A_t = a\}}\bigr] = \sum_{a \in [K]} \bigl(\mu^\star - \mu_a\bigr)\, \mathbb{E}\bigl[N_a(T)\bigr]$$

where $N_a(T) = \sum_{t=1}^{T} \mathbb{I}_{\{A_t = a\}}$

Model: $\nu_1, \ldots, \nu_K$ are distributions over $[0,1]$

A popular strategy: UCB [upper confidence bound]

Auer, Cesa-Bianchi and Fisher [2002]

For $t \geqslant K$, pick $$A_{t+1} \in \mathop{\arg\max}_{a \in [K]} \left\{ \widehat{\mu}_a(t) + \sqrt{\frac{2 \ln t}{N_a(t)}} \right\}$$

Exploitation: cf. empirical mean $\widehat{\mu}_a(t) = \dfrac{1}{N_a(t)} \displaystyle\sum_{s=1}^{t} Y_s \, \mathbb{I}_{\{A_s = a\}}$

Exploration: cf. $\sqrt{2 \ln t / N_a(t)}$ favors arms a not pulled often

Regret bounds (suboptimal) of two types

– Distribution-dependent bound: $$R_T \lesssim \sum_{a : \mu_a < \mu^\star} \frac{8 \ln T}{\mu^\star - \mu_a}$$

– Distribution-free bound: $$\sup_{\nu_1, \ldots, \nu_K} R_T \lesssim \sqrt{8 K T \ln T}$$

Proof of $R_T \lesssim \sum_{a:\mu_a<\mu^\star} \dfrac{8 \ln T}{\mu^\star - \mu_a}$

Hoeffding–Azuma: $$\mathbb{P}\left\{ \big|\mu_a - \widehat{\mu}_a(t)\big| \leqslant \sqrt{\frac{2\ln t}{N_a(t)}} \right\} \geqslant 1 - 2t^{-3}$$

Indeed, by optional skipping:

$$\begin{aligned}
&\mathbb{P}\left\{ \big|\mu_a - \widehat{\mu}_a(t)\big| > \sqrt{\frac{2\ln t}{N_a(t)}} \right\} \\
&= \sum_{n=1}^{t} \mathbb{P}\left\{ \big|\mu_a - \widehat{\mu}_{a,n}\big| > \sqrt{\frac{2\ln t}{n}} \ \text{ and } \ N_a(t) = n \right\} \\
&\leqslant \sum_{n=1}^{t} \underbrace{\mathbb{P}\left\{ \big|\mu_a - \widehat{\mu}_{a,n}\big| > \sqrt{\frac{\ln(1/t^{-4})}{2n}} \right\}}_{\leqslant 2t^{-4}}
\end{aligned}$$

where $\widehat{\mu}_{a,n}$ denotes the average of n–sample with distribution ν_a

Proof of $R_T \lesssim \sum_{a:\mu_a<\mu^\star} \frac{8 \ln T}{\mu^\star - \mu_a}$

Hoeffding–Azuma: $$\mathbb{P}\left\{ \left| \mu_a - \widehat{\mu}_a(t) \right| \leqslant \sqrt{\frac{2 \ln t}{N_a(t)}} \right\} \geqslant 1 - 2t^{-3}$$

If $A_t = b$ is not an optimal arm $a^\star$, then by design

$$\widehat{\mu}_{a^\star}(t) + \sqrt{\frac{2 \ln t}{N_{a^\star}(t)}} \leqslant \widehat{\mu}_b(t) + \sqrt{\frac{2 \ln t}{N_b(t)}}$$

thus w.h.p. $$\mu^\star \leqslant \mu_b + 2\sqrt{\frac{2 \ln t}{N_b(t)}}$$

which imposes $$N_b(t) \leqslant \frac{8 \ln T}{(\mu^\star - \mu_a)^2}$$

Conclude with $$R_T = \sum_{a \in [K]} \left(\mu^\star - \mu_a \right) \mathbb{E}\left[N_a(T) \right]$$

Proof of $\sup_{\nu_1,\ldots,\nu_K} R_T \lesssim \sqrt{8KT \ln T}$

We proved $$\mathbb{E}\big[N_b(t)\big] \lesssim \frac{8 \ln T}{(\mu^\star - \mu_a)^2}$$

Thus $$R_T = \sum_{a \in [K]} (\mu^\star - \mu_a) \sqrt{\mathbb{E}\big[N_a(T)\big]} \sqrt{\mathbb{E}\big[N_a(T)\big]}$$
$$\leqslant \sqrt{8 \ln T} \sum_{a \in [K]} \sqrt{\mathbb{E}\big[N_a(T)\big]}$$
$$\leqslant \sqrt{8KT \ln T}$$

Contextual stochastic bandits with K arms

Linear modeling + Logistic modeling

At each round $t = 1, 2, \ldots,$

0. A context $\mathbf{x}_t \in \mathbb{R}^d$ is determined by the environment
1. Statistician picks arm $A_t \in [K]$
2. She gets a reward Y_t with conditional expectation $r(\mathbf{x}_t, A_t)$
3. This is the only feedback she receives

Goal:
Maximize expected rewards $\longleftrightarrow$ Minimize expected regret

$$R_T = \sum_{t=1}^{T} \text{targets?} - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$$

Structural assumptions handy! E.g., linearity:

$$r(\mathbf{x}, a) = \varphi(\mathbf{x}, a)^{\mathsf{T}} \theta_\star \qquad \rightsquigarrow \qquad \text{targets} \quad \max_{a \in [K]} \varphi(\mathbf{x}_t, a)^{\mathsf{T}} \theta_\star$$

Transfer function $\varphi : \mathbb{R}^d \times [K] \to \mathbb{R}^m$ known,
But parameters $\theta_\star \in \mathbb{R}^d$ unknown

Setting: contexts $\mathbf{x}_t \in \mathbb{R}^d$, pick arms $A_t \in [K]$, get rewards Y_t

Regret
$$R_T = \sum_{t \leqslant T} \max_{a \in [K]} \varphi(\mathbf{x}_t, a)^{\mathsf{T}} \theta_\star - \sum_{t \leqslant T} \mathbb{E}\big[\varphi(\mathbf{x}_t, A_t)^{\mathsf{T}} \theta_\star\big]$$

Key: learn $\theta_\star$ (= estimate it while playing)

LinUCB with regularization $\lambda > 0$ for bounded contexts

Abbasi-Yadkori, Pál, Szepesvári [2011]

Based on the idea $\displaystyle\sum_{s=1}^{t-1} \varphi(\mathbf{x}_s, A_s) Y_s \approx \sum_{s=1}^{t-1} \varphi(\mathbf{x}_s, A_s) \varphi(\mathbf{x}_s, A_s)^{\mathsf{T}} \theta_\star$

Statement: let
$$M_{t-1} = \lambda \, \mathrm{Id} + \sum_{s=1}^{t-1} \varphi(\mathbf{x}_s, A_s) \varphi(\mathbf{x}_s, A_s)^{\mathsf{T}}$$

and
$$\widehat{\theta}_{t-1} = \big(M_{t-1}\big)^{-1} \sum_{s=1}^{t-1} \varphi(\mathbf{x}_s, A_s) Y_s$$

Setting: bounded contexts $\mathbf{x}_t \in \mathbb{R}^d$, arms $A_t \in [K]$, rewards Y_t

reward function $r(\mathbf{x}, a) = \varphi(\mathbf{x}, a)^{\mathsf{T}}\theta_\star$, with $\mathbb{E}\big[Y_t \,\big|\, A_t, x_t\big] = \varphi(\mathbf{x}_t, A_t)^{\mathsf{T}}\theta_\star$

$$\widehat{\theta}_{t-1} = \big(M_{t-1}\big)^{-1}\sum_{s=1}^{t-1}\varphi(\mathbf{x}_s, A_s)Y_s \quad \text{where} \quad M_{t-1} = \lambda\,\mathrm{Id} + \sum_{s=1}^{t-1}\varphi(\mathbf{x}_s, A_s)\varphi(\mathbf{x}_s, A_s)^{\mathsf{T}}$$

Confidence region on $\theta_\star$:

$$\mathbb{P}\Big\{\,\big\|\widehat{\theta}_{t-1} - \theta_\star\big\|_{M_{t-1}} \lesssim \square\sqrt{\ln(t/\delta)}\Big\} = 1 - \delta$$

where $\|u\|_M = \sqrt{u^{\mathsf{T}}Mu}$ and provided that λ is well set

Complex proof based on "Laplace's method of mixtures"

Simultaneous confidence intervals on the $r(\mathbf{x}, a)$: based on

$$\big|\varphi(\mathbf{x}, a)^{\mathsf{T}}\widehat{\theta}_{t-1} - \varphi(\mathbf{x}, a)^{\mathsf{T}}\theta_\star\big| \leqslant \big\|\widehat{\theta}_{t-1} - \theta_\star\big\|_{M_{t-1}}\,\big\|\varphi(\mathbf{x}, a)\big\|_{(M_{t-1})^{-1}}$$

$$\leqslant \underbrace{\square\sqrt{\ln(t/\delta)}\,\big\|\varphi(\mathbf{x}, a)\big\|_{(M_{t-1})^{-1}}}_{=\varepsilon_{t-1,\delta}(\mathbf{x},a)}$$

where $\displaystyle\sum_{t=1}^{T}\varepsilon_{t-1,\delta}(\mathbf{x}_t, A_t) \lesssim \sqrt{T}\ln(T/\delta)$ by linear algebra

Setting: bounded contexts $\mathbf{x}_t \in \mathbb{R}^d$, arms $A_t \in [K]$, rewards Y_t
reward function $r(\mathbf{x}, a) = \varphi(\mathbf{x}, a)^{\mathsf{T}}\theta_\star$, with $\mathbb{E}\big[Y_t \,\big|\, A_t, x_t\big] = \varphi(\mathbf{x}_t, A_t)^{\mathsf{T}}\theta_\star$

Simultaneous confidence intervals: $\big|\hat{r}_{t-1}(\mathbf{x}, a) - r(\mathbf{x}, a)\big| \leqslant \varepsilon_{t-1,\delta}(\mathbf{x}, a)$

where $\displaystyle\sum_{t \leqslant T} \varepsilon_{t-1,\delta}(\mathbf{x}_t, A_t) \lesssim \sqrt{T}\ln(T/\delta)$

Optimistic choice: $$A_t \in \mathop{\arg\max}_{a \in [K]} \Big\{ \hat{r}_{t-1}(\mathbf{x}_t, a) + \varepsilon_{t-1,\delta}(\mathbf{x}_t, a) \Big\}$$

Regret bound: $$R_T = \sum_{t=1}^{T} \max_{a \in [K]} r(\mathbf{x}_t, a) - \sum_{t=1}^{T} Y_t \leqslant \widetilde{\mathcal{O}}\big(\sqrt{T}\big)$$

In high-probability (but algorithm depends on δ)
Or in expectation (set $\delta = t^{-4}$, e.g.)

We could also have obtained high-probability bounds based on the UCB strategy in the non-contextual case

Logistic bandits

Extended from Faury, Abeille, Calauzènes, Fercoq [2020]

At each round $t = 1, 2, \ldots,$

0. A context $\mathbf{x}_t \in \mathbb{R}^d$ is determined by the environment
1. Statistician picks arm $A_t \in [K]$
2. The outcome $Y_t \in \{0, 1\}$ is drawn with probability $P(\mathbf{x}_t, A_t)$
3. This is the only feedback Statistician receives
4. Statistician gets the reward $r(\mathbf{x}_t, A_t)\, Y_t$

Conversion rate P unknown but reward function r known

Structural assumption:

$$P(\mathbf{x}, a) = \eta\big(\varphi(\mathbf{x}, a)^{\mathsf{T}} \theta_\star\big) \qquad \text{where} \qquad \eta(x) = \frac{1}{1 + \mathrm{e}^{-x}}$$

Similar results may be achieved as for linear bandits

Estimation based on maximum likelihood

Contextual stochastic bandits with K arms

And now, with budget constraints!

At each round $t = 1, 2, \ldots,$

0. A context $\mathbf{x}_t \sim \mathbb{Q}$ is drawn at random
1. Statistician picks arm $A_t \in [K]$
2. She gets a reward Y_t with conditional expectation $r(\mathbf{x}_t, A_t)$
3. She also suffers costs $\mathbf{Z}_t$ with conditional expectation $\mathbf{c}(\mathbf{x}_t, A_t)$
4. Her feedback is Y_t and $\mathbf{Z}_t$

Vector-valued costs: possibly several constraints

Goals:

Maximize $\displaystyle\sum_{t \leqslant T} Y_t$ while ensuring $\displaystyle\sum_{t \leqslant T} \mathbf{Z}_t \leqslant T\mathbf{B}$

Known: budget $T\mathbf{B}$

Unknown: reward function r, cost function $\mathbf{c}$, distribution $\mathbb{Q}$

but structural assumptions to be issued on r and $\mathbf{c}$

Setting called CBwK – contextual bandits with knapsacks

First reference for CBwK: Badanidiyuru, Langford, Slivkins [2014]

State of the art = $T\mathbf{B}$ at best $T^{3/4}$: Agrawal and Devanur [2016], Han et al. [2022]

Fairness application

Inspired from Chohlas-Wood, Coots, Zhu, Brunskill, Goel [2021]

Fair budget spending among groups: Z'_t first component of $\mathbf{Z}_t$

$$\sum_{t=1}^{T} Z'_t \leqslant T B_{\text{total}}$$

$$\text{and} \qquad \forall g \in \mathcal{G}, \quad \left| \frac{1}{T\gamma_g} \sum_{t=1}^{T} Z'_t \, \mathbb{I}_{\{\mathrm{gr}(\mathbf{x}_t)=g\}} - \frac{1}{T} \sum_{t=1}^{T} Z'_t \right| \leqslant \tau$$

where $\gamma_g = \mathbb{Q}\{\mathrm{gr}(\,\cdot\,) = g\}$
and τ is a tolerance factor, ideally $\sim 1/\sqrt{T}$, i.e., $T\tau$ of order $\sqrt{T}$

$\mathbf{B}$ contains a B_{total} component, as well as components $\pm\gamma_g\tau$

Setting: context $\mathbf{x}_t \sim \mathbb{Q}$, arm $A_t \in [K]$, reward Y_t and costs $\mathbf{Z}_t$

Conditional expectations: $r(\mathbf{x}_t, A_t)$ and $\mathbf{c}(\mathbf{x}_t, A_t)$

Total budget constraints $T\mathbf{B}$, where some components are as small as $\sqrt{T}$

Benchmark: static policies $\pi : \mathbf{x} \mapsto \big(\pi_a(\mathbf{x})\big)_{a\in[K]} \in \mathcal{P}\big([K]\big)$

We assume feasibility, and actually do so for $\mathbf{B} - \varepsilon\mathbf{1}$ (OK if a null-cost action exists)

$$\mathrm{opt}(r, \mathbf{c}, \mathbf{B}) = \sup_{\pi} \left\{ \mathbb{E}_{\mathbf{X}\sim\mathbb{Q}}\left[\sum_{a\in[K]} r(\mathbf{X}, a)\,\pi_a(\mathbf{X})\right] \right.$$
$$\left. \text{under} \quad \mathbb{E}_{\mathbf{X}\sim\mathbb{Q}}\left[\sum_{a\in[K]} \mathbf{c}(\mathbf{X}, a)\,\pi_a(\mathbf{X})\right] \leqslant \mathbf{B} \right\}$$

Regret: $$R_T = T\,\mathrm{opt}(r, \mathbf{c}, \mathbf{B}) - \sum_{t\leqslant T} Y_t$$

Hard constraint: $$\sum_{t\leqslant T} \mathbf{Z}_t \leqslant T\mathbf{B}$$

Regret: Minimize $R_T = T\,\mathrm{opt}(r, \mathbf{c}, \mathbf{B}) - \sum_{t \leqslant T} Y_t$ where

$$\mathrm{opt}(r, \mathbf{c}, \mathbf{B})$$

$$= \sup_{\pi} \left\{ \mathbb{E}_{\mathbf{X}\sim\mathbb{Q}}\left[\sum_{a\in[K]} r(\mathbf{X}, a)\,\pi_a(\mathbf{X}) \right] : \ \mathbb{E}_{\mathbf{X}\sim\mathbb{Q}}\left[\sum_{a\in[K]} \mathbf{c}(\mathbf{X}, a)\,\pi_a(\mathbf{X}) \right] \leqslant \mathbf{B} \right\}$$

$$= \sup_{\pi} \inf_{\boldsymbol{\lambda}\geqslant\mathbf{0}} \mathbb{E}_{\mathbf{X}\sim\mathbb{Q}}\left[\sum_{a\in[K]} r(\mathbf{X}, a)\,\pi_a(\mathbf{X}) - \left\langle \boldsymbol{\lambda}, \sum_{a\in[K]} \mathbf{c}(\mathbf{X}, a)\,\pi_a(\mathbf{X}) - \mathbf{B} \right\rangle \right]$$

$$= \min_{\boldsymbol{\lambda}\geqslant\mathbf{0}} \mathbb{E}_{\mathbf{X}\sim\mathbb{Q}}\left[\max_{a\in[K]} \left\{ r(\mathbf{X}, a) - \langle \mathbf{c}(\mathbf{X}, a) - \mathbf{B}, \boldsymbol{\lambda} \rangle \right\} \right]$$

$\rightarrow$ Suffices to learn r and c, as well as $\lambda^\star$ $\rightsquigarrow$ parametric problems[†]
Cf. $\mathbf{x}_t \sim \mathbb{Q}$ observed at each round

Learn r and c: via [†]structural assumptions (linearity or logistic)
Uniform bounds available

Target: $$\mathrm{opt}(r, \mathbf{c}, \mathbf{B}) = \min_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \mathbb{E}_{\mathbf{X}\sim\mathbb{Q}}\left[\max_{a\in[K]}\left\{r(\mathbf{X}, a) - \langle \mathbf{c}(\mathbf{X}, a) - \mathbf{B}, \boldsymbol{\lambda}\rangle\right\}\right]$$

→ Gradient descent on dual / best response for primal variable(s)

Algorithm with fixed step size γ

For $t = 1, 2, \ldots, T$:

1. Play $A_t \in \operatorname*{arg\,max}_{a\in[K]}\left\{\hat{r}_{t-1}(\mathbf{x}_t, a) - \left\langle \hat{\mathbf{c}}_{t-1}(\mathbf{x}_t, a) - (\mathbf{B} - b\mathbf{1}), \boldsymbol{\lambda}_{t-1}\right\rangle\right\}$
2. Make gradient step $\boldsymbol{\lambda}_t = \Big(\boldsymbol{\lambda}_{t-1} + \gamma\big(\hat{\mathbf{c}}_{t-1}(\mathbf{x}_t, a) - (\mathbf{B} - b\mathbf{1})\big)\Big)_+$
3. Update estimates $\hat{r}_t$ and $\hat{\mathbf{c}}_t$ of functions r and $\mathbf{c}$

Optimistic estimates: $\hat{r}_t$ upper bounds r and $\hat{\mathbf{c}}_t$ lower bounds $\mathbf{c}$

Idea already in Agrawal and Devanur [2016]
But the key to handle smaller budgets is the tuning of γ

From Chzhen, Giraud, Li, Stoltz [2023]

Strategy with fixed γ

1. Play $A_t \in \arg\max_{a \in [K]} \Big\{ \hat{r}_{t-1}(\mathbf{x}_t, a) - \big\langle \hat{\mathbf{c}}_{t-1}(\mathbf{x}_t, a) - (\mathbf{B} - b\mathbf{1}), \, \boldsymbol{\lambda}_{t-1} \big\rangle \Big\}$
2. Make gradient step $\boldsymbol{\lambda}_t = \Big(\boldsymbol{\lambda}_{t-1} + \gamma \big(\hat{\mathbf{c}}_{t-1}(\mathbf{x}_t, a) - (\mathbf{B} - b\mathbf{1}) \big) \Big)_+$

Analysis for fixed γ: the projected-gradient descent entails

$$\left\| \left(\sum_{t=1}^{T} \mathbf{Z}_t - T(\mathbf{B} - b\mathbf{1}) \right)_+ \right\| \leqslant \widetilde{\mathcal{O}}\left(\frac{1 + \|\lambda^\star\|}{\gamma} \right)$$

Cost margin Tb should be of the same order $\big(1 + \|\lambda^\star\|\big)/\gamma$

That margin adds a term of order $\|\lambda^\star\| \big(Tb + \sqrt{T} \big)$ to regret

$\rightarrow$ Oracle choices $b \sim 1/\sqrt{T}$ and $\gamma \sim \big(1 + \|\lambda^\star\|\big)/\sqrt{T}$

lead to $\big(1 + \|\lambda^\star\|\big)\sqrt{T}$ regret

Estimating $\|\lambda^\star\|$ on $\sqrt{T}$ exploration rounds (see, e.g.: Agrawal and Devanur [2016], Han et al. [2022]) imposes $\min \mathbf{B} \geqslant T^{-1/4}$

We use instead a careful doubling trick $\gamma_k = 2^k/\sqrt{T}$

Breaking condition based on budget controls

Theorem:

If $\min \mathbf{B}$ is larger than $1/\sqrt{T}$ up to poly-log terms, then w.h.p.,

$$\sum_{t \leqslant T} \mathbf{Z}_t \leqslant T\mathbf{B} \qquad \text{and} \qquad R_T \lesssim \widetilde{\mathcal{O}}\big(1 + \|\lambda^\star\|\big)\sqrt{T}$$

Note: $\|\lambda^\star\| \leqslant \dfrac{2\,\mathrm{opt}(r, \mathbf{c}, \mathbf{B})}{\min \mathbf{B}}$ if null-cost action