Online learning for brokerage

François Bachoc

Laboratoire Paul Painlevé
Université de Lille
Institut universitaire de France (IUF)

Joint work with **Nataša Bolić** (University of Ottawa), **Tommaso Cesari** (University of Ottawa) and **Roberto Colomboni** (University of Milan)

November 2025

Outline

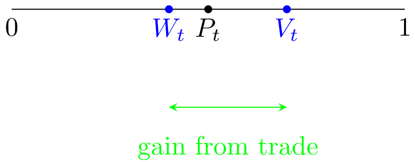
1. Brokerage
2. Contextual linear brokerage
3. Contextual Lipschitz brokerage
4. Full feedback

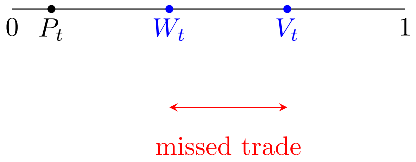
Brokerage problem

- Consider successive pairs of trading agents entering a market at times $t = 1, 2, \ldots$
- At each time t, each agent gives a valuation to a good to be traded.
- Focus on the pairs of valuations $(V_t, W_t)_{t \in \mathbb{N}} \in [0,1]^{\mathbb{N}}$, assumed to be iid.
- We take the role of the trading platform that suggests a trading price $P_t \in [0,1]$ at each time t.
- The gain from trade is

$$\mathrm{gft}(P_t, V_t, W_t) = \underbrace{\mathbb{I}\left\{\min(V_t, W_t) \leq P_t \leq \max(V_t, W_t)\right\}}_{\text{whenever a trade occurs}} \left(\underbrace{[\max(V_t, W_t) - P_t]}_{\text{profit of the buyer}} + \underbrace{[P_t - \min(V_t, W_t)]}_{\text{profit of the seller}} \right)$$

$$= \mathbb{I}\left\{\min(V_t, W_t) \leq P_t \leq \max(V_t, W_t)\right\} \left(\max(V_t, W_t) - \min(V_t, W_t)\right).$$





Brokerage problem as an online learning problem

Online protocol with two-bit feedback

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

1. The learner chooses a price $P_t \in [0, 1]$.
2. Two traders arrive with valuations V_t, W_t and the learner earns the (hidden) gain from trade
$$\mathrm{gft}(P_t, V_t, W_t).$$
3. The learner observes the two-bit feedback
$$\mathbb{I}\{P_t \leq V_t\}, \mathbb{I}\{P_t \leq W_t\}.$$

Cumulated regret from time $t = 1, \ldots, T$

$$R_T := \sup_{p \in [0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{gft}(p, V_t, W_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{gft}(P_t, V_t, W_t)\right].$$

Some related problems

- Dynamic pricing: the learner (seller) posts a price $P_t \in [0,1]$ are receives profit

$$P_t \mathbb{1}\{P_t \leq Y_t\}$$

for a random buyer valuation $Y_t \in [0,1]$ [Kleinberg and Leighton, 2003, Tullii et al., 2024].

- Bilateral trade: the learner (the platform) posts a price $P_t \in [0,1]$ are receives profit

$$\mathbb{1}\{S_t \leq P_t \leq B_t\}(B_t - S_t)$$

for random buyer (B_t) and seller (S_t) valuations in $[0,1]$ [Cesa-Bianchi et al., 2024].
$\Longrightarrow$ Same as our setting but with pre-determined buyer and seller.

Assumption: independent valuations, identical distribution

Assumption

The valuations V_t, W_t are independent with identical distribution.

Identical distribution

- Can be interpreted as a symmetry.
- Removing it makes the problem a variation of bilateral trade [Cesa-Bianchi et al., 2024].

Independence

- Open question: removing this assumption leads to settings of linear regret, similarly as in bilateral trade?
- Two-bit feedback allows to recover the marginal distributions of V_t and W_t but not the joint.

Explicit expressions of the gain from trade

- Write ν for the distribution of V_t and W_t.
- Write $\bar{\nu} = \mathbb{E}\left[V_t\right] = \mathbb{E}\left[W_t\right]$.

Define, for $p \in [0,1]$,

$$\widetilde{\rho}(\nu)(p) := \int_0^p \left(\nu[0,\lambda] + \nu[0,\lambda)\right) \mathrm{d}\lambda + \left(\nu[0,p] + \nu[0,p)\right)(\bar{\nu} - p),$$

$$\rho(\nu)(p) := \widetilde{\rho}(\nu)(p) + \nu\{p\} \left(\int_0^p \nu[0,\lambda] \, \mathrm{d}\lambda + \int_p^1 \nu[\lambda,1] \, \mathrm{d}\lambda \right).$$

Explicit expressions [Bolić et al., 2024].

- In general, we have

$$\mathbb{E}\left[\mathrm{gft}(p, V_t, W_t)\right] = \rho(\nu)(p).$$

- If ν has a density bounded by $M < \infty$, we have

$$0 \leq \rho(\nu)(\bar{\nu}) - \rho(\nu)(p) \leq M \left|\bar{\nu} - p\right|^2.$$

$\Longrightarrow$ Reduces to estimating the mean for continuous distributions.

Algorithm for continuous distributions

To estimate the mean:

$$\mathbb{E}[V_t] = \int_0^1 \mathbb{P}[x \leq V_t] \, \mathrm{d}x.$$

Algorithm Explore then Commit

Input: Exploration time $T_0 \in \mathbb{N}$
for $t = 1, 2, \ldots, T_0$ **do** (explore)
    Post $P_t \leftarrow \frac{t}{T_0}$
    Receive feedback $\mathbb{I}\{P_t \leq V_t\}$ and $\mathbb{I}\{P_t \leq W_t\}$
end for
for $t = T_0 + 1, T_0 + 2, \ldots$ **do** (commit)
    Post $P_t \leftarrow \frac{1}{2T_0} \sum_{s=1}^{T_0} \left(\mathbb{I}\{P_s \leq V_s\} + \mathbb{I}\{P_s \leq W_s\} \right)$
end for

Bounds for continuous distributions

From [Bolić et al., 2024].

Upper bound

With Explore then Commit algorithm, tuning the parameter $T_0 := \lceil \sqrt{MT} \rceil$ yields

$$R_T \leq 2.5 + 2\sqrt{MT} \ .$$

Lower bound

The worst-case regret of any algorithm satisfies, for $T \geq \mathrm{constant}\, M^3$,

$$\sup_{\nu \text{ has density bounded by } M} R_T^\nu \geq \mathrm{constant} \sqrt{MT} \ ,$$

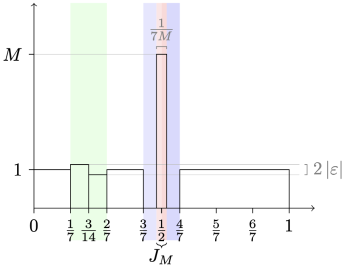
where R_T^ν is the regret at time T under $V_t, W_t \sim \nu$.

Complete tight dependence in M remains open.

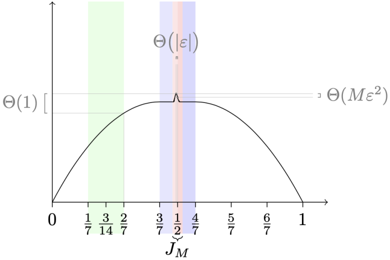
Ideas on the lower bound

From [Bolić et al., 2024] : two close hard instances with $\pm\epsilon$.

Two possible densities



Two possible gains from trade



- Only way to differentiate the two densities: post price in suboptimal region $[\frac{1}{7}, \frac{2}{7}]$.
- Take $\epsilon = (MT)^{-\frac{1}{4}}$.

Impossibility result for general distributions

From [Bolić et al., 2024].

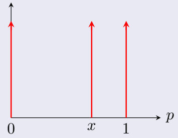
Linear regret for general distributions

For any time horizon T, the worst-case regret of any algorithm satisfies

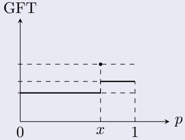
$$\sup_{\nu} R_T^{\nu} \geq \frac{T}{9} ,$$

where the sup is over all distributions ν.

Proof idea: needle in a haystack



(a) Distribution of V_t, W_t (3 weighted Diracs)



(b) Expected gain from trade

Contextual linear setting

In [Bachoc et al., 2025a].

Online protocol with contexts

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

1. A context $c_t \in [0,1]^d$ is revealed.
   - Deterministic adversarial sequence of contexts, unknown in advance.
2. The learner chooses a price $P_t \in [0,1]$.
3. Hidden valuations V_t, W_t and the learner earns the (hidden) gain from trade.
4. The learner observes the two-bit feedback.

Linearity assumpion

The pairs $(V_t, W_t)_{t \in \mathbb{N}}$ are no longer iid.
For all $t \in \mathbb{N}$,

$$\mathbb{E}[V_t] (= \mathbb{E}[W_t]) = c_t^\top \phi$$

for a fixed unknown $\phi \in [0,1]^d$.

Contextual linear setting: regret

Cumulated regret from time $t = 1, \dots, T$

$$R_T := \sup_{p^\star : [0,1]^d \to [0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{gft}(p^\star(c_t), V_t, W_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{gft}(P_t, V_t, W_t)\right].$$

Algorithm

- For densities upper bounded by M.
- Write $\|x\|_Q^2 = x^\top Q x$ for a vector x and a matrix Q.

Algorithm (exploit or learn ridge regression)

Post $P_1 \sim \mathcal{U}([0,1])$, and observe $D_1 := \mathbb{I}\{P_1 \leq V_1\}$

Let $x_1 := [c_1]$, let $Y_1 := [D_1]$ and compute $\hat{\phi}_1 := (x_1 x_1^\top + d^{-1} \mathbf{I}_d)^{-1} x_1 Y_1^\top$ (ridge regression)

for time $t = 2, 3, \ldots$ **do**

  Observe context c_t and define $b_t := \mathbb{I}\left\{ \left\| \sqrt{2} c_t \right\|^2_{(x_{t-1} x_{t-1}^\top + d^{-1} \mathbf{I}_d)^{-1}} > \sqrt{\frac{2d \log(1+2d(T-1))}{MT}} \right\}$

  if $b_t = 1$ **then** (explore - learn ridge regression)

    Post $P_t \sim \mathcal{U}([0,1])$, and observe $D_t := \mathbb{I}\{P_t \leq V_t\}$

    Let $x_t := [x_{t-1} \mid c_t]$, $Y_t := [Y_{t-1} \mid D_t]$ and $\hat{\phi}_t := (x_t x_t^\top + \mathbf{I}_d)^{-1} x_t Y_t^\top$

  else (exploit)

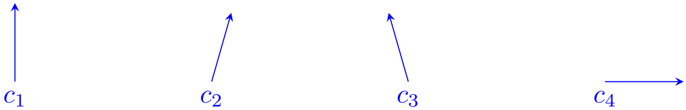
    post $P_t := c_t^\top \hat{\phi}_{t-1}$ and let $x_t := x_{t-1}$, $Y_t := Y_{t-1}$, and $\hat{\phi}_t := \hat{\phi}_{t-1}$

  end if

end for

Illustration of the algorithm

For $d = 2$:



c_1	c_2	c_3	c_4
$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$
explore	explore	exploit	explore
uniform price	uniform price	price $c_3^\top \hat{\phi}$	uniform price
update $\hat{\phi}$	update $\hat{\phi}$	unchanged $\hat{\phi}$	update $\hat{\phi}$
		small regret	

Upper bound

Upper bound

With the previous algorithm (exploit or feed ridge regresion) we have, when the density of V_t and W_t is bounded by M,

$$R_T \leq \mathrm{constant}\sqrt{MdT\log(T)}.$$

A useful tool: **Elliptical potential lemma** (taken from [Lattimore and Szepesvári, 2020])

LEMMA 19.4. *Let $V_0 \in \mathbb{R}^{d\times d}$ be positive definite and $a_1, \dots, a_n \in \mathbb{R}^d$ be a sequence of vectors with $\|a_t\|_2 \leq L < \infty$ for all $t \in [n]$, $V_t = V_0 + \sum_{s\leq t} a_s a_s^\top$. Then,*

$$\sum_{t=1}^{n}\left(1 \wedge \|a_t\|^2_{V_{t-1}^{-1}}\right) \leq 2\log\left(\frac{\det V_n}{\det V_0}\right) \leq 2d\log\left(\frac{\operatorname{trace} V_0 + nL^2}{d\det(V_0)^{1/d}}\right).$$

Lower bound

Lower bound

For any algorithm, for $T \geq \max(4, \mathrm{constant}\, dM^3, 2d)$,

$$\sup_{\substack{\text{settings} \\ \text{contexts}}} R_T^{\text{settings,contexts}} \geq \mathrm{constant}\sqrt{MdT},$$

where the sup is over all settings and context sequence where

- linearity assumption holds,
- V_t and W_t are independent and identically distributed, with density bounded by M.

Open question: ranges of c_t and ϕ

- Currently, we assume $\phi \in [0,1]^d$ and $c_t^\top \phi \in [0,1]$.
- Allows to have tight lower and upper bounds in d, T.
- The lower bound uses $(\phi_1, \ldots, \phi_d) = (\frac{1}{2} \pm o(1), \ldots, \frac{1}{2} \pm o(1))$ and $c_1, \ldots, c_t$ unit vectors.
- Open question: tighter bounds for ϕ sparse or $\|\phi\| \ll \sqrt{d}$.

Contextual Lipschitz setting

- In [Bachoc et al., 2025b].
- Same online protocol with contexts as in the linear case.

Lipschitz assumpion

For all $t, t' \in \mathbb{N}$,

$$\left|\mathbb{E}\left[V_t\right] - \mathbb{E}\left[V_{t'}\right]\right| \leq \left\|c_t - c_{t'}\right\|_{\infty}.$$

(Recall that $\mathbb{E}\left[V_t\right] = \mathbb{E}\left[W_t\right]$, $\forall t$.)

- Same definition of cumulated regret as in the linear case (compete against best deterministic function of contexts).

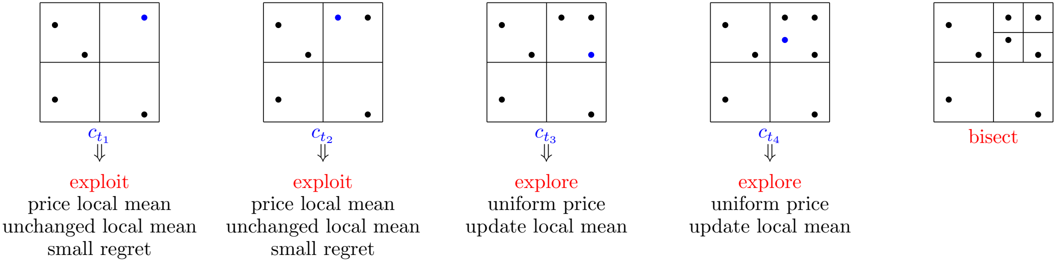
Algorithm

Algorithm Exploit, Explore, and Bisect (main ideas)

- Maintains a set of active cells.
  - Of the form $\prod_{k=1}^{d} (\frac{i_k - 1}{2^j}, \frac{i_k}{2^j}]$, $i_1, \ldots, i_k \in \{1, \ldots, 2^j\}$ with depth j.
  - Starting with just the cell $(0, 1]^d$.
  - Maintaining at all time a partition of $(0, 1]^d$.
- For a cell of depth j, consider the contexts that belong to it, at times when it is active:
  1. Exploit: for the 2^{4j} first contexts, post a local average of two-bit-feedbacks from previously generated uniform prices (in previous exploration times).
  2. Explore: for the 2^{2j} next contexts, generate a uniform price.
  3. Bisect: then make the cell inactive and replace it by its 2^d children by bisection.

Illustration of the algorithm

For $d = 2$: life of an active cell at times $t_1 < t_2 < t_3 < t_4$.



Upper bound

Upper bound

With the previous algorithm (Exploit, Explore, and Bisect) we have, when the density of V_t and W_t is bounded by M,

$$R_T \leq \mathrm{constant}_d M T^{\frac{d+2}{d+4}}.$$

Open question: dependence on an unknown intrinsic context dimension $d_0 < d$?

Lower bound

Lower bound

For any algorithm, for $M \geq 2$,

$$\sup_{\substack{\text{settings} \\ \text{contexts}}} R_T^{\text{settings,contexts}} \geq \mathrm{constant}_d T^{\frac{d+2}{d+4}}$$

where the sup is over all settings and contexts where

- Lipschitz assumption holds,
- V_t and W_t are independent and identically distributed, with density bounded by M.

Tight dependence in M remains open.

Full feedback

Online protocol with full feedback

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

1. The learner chooses a price $P_t \in [0, 1]$.
2. Two traders arrive with valuations V_t, W_t that are observed by the learner.
3. The learner earns the (observed) gain from trade

$$\mathrm{gft}(P_t, V_t, W_t).$$

- The learner does not need to “select” the information it receives.
- Less realistic modeling assumption.
- Motivation:
  - different theoretical behavior,
  - quantify loss of performance due to loss of information in two-bit feedback.

For bounded density

From [Bolić et al., 2024].

Algorithm (follow the mean)

1: Post $P_1 \leftarrow \frac{1}{2}$
2: Receive feedback V_1, W_1
3: **for** $t = 2, 3, \dots$ **do**
4: Post $P_t \leftarrow \frac{\sum_{s=1}^{t-1} V_s + W_s}{2(t-1)}$
5: Receive feedback V_t, W_t
6: **end for**

Upper bound

With this algorithm, we have, when the density of V_t and W_t is bounded by M, for $T \geq 2$,

$$R_T \leq \frac{1}{2} + \frac{M}{4}\big(1 + \log(T - 1)\big).$$

For bounded density

From [Bolić et al., 2024].

Lower bound

The worst-case regret of any algorithm satisfies, for $T \geq \mathrm{constant} M^8$,

$$\sup_{\nu \text{ has density bounded by } M} R_T^\nu \geq \mathrm{constant} M \log(T),$$

where R_T^ν is the regret at time T under $V_t, W_t \sim \nu$.

Complete tight dependence in M remains open.

Without bounded density

From [Bolić et al., 2024].
Recall

$$\mathbb{E}\left[\mathrm{gft}(p, V_t, W_t)\right] = \rho(\nu)(p).$$

Algorithm (follow the ρ)

1: Post $P_1 \leftarrow \frac{1}{2}$
2: Receive feedback V_1, W_1
3: **for** $t = 2, 3, \ldots$ **do**
4: Let $\hat{\nu}_t \leftarrow \frac{1}{2(t-1)} \sum_{s=1}^{t-1} \delta_{V_s} + \delta_{W_s}$
5: Post $P_t \in \mathrm{argmax}_{p \in [0,1]} \; \rho\big(\hat{\nu}_t\big)(p)$
6: Receive feedback V_t, W_t
7: **end for**

Without bounded density

From [Bolić et al., 2024].

Upper bound

With the previous algorithm (follow the ρ), we have

$$R_T \leq 1/2 + 4\left(3\sqrt{\pi} + \sqrt{2}\right)\sqrt{T-1}.$$

Lower bound

The worst-case regret of any algorithm satisfies

$$\sup_{\text{distribution } \nu} R_T^{\nu} \geq \mathrm{constant}\sqrt{T},$$

where R_T^{ν} is the regret at time T under $V_t, W_t \sim \nu$.

A first best of both worlds

From [Bolić et al., 2024].

Algorithm (follow the mean then ρ)

1: **for** $t = 1, 2, \ldots$ **do**
2: Post P_t according to follow the mean
3: **if** $\left|\{V_1, \ldots, V_t, W_1, \ldots, W_t\}\right| < 2t$ (a repetition) **then**
4: $\tau \leftarrow t$
5: **break**
6: **end if**
7: **end for**
8: Run follow the ρ up to time τ without posting prices
9: **for** $t = \tau + 1, \tau + 2, \ldots$ **do**
10: Post P_t according to follow the ρ
11: **end for**

Upper bounds

- If V_t, W_t have density bounded by M: $R_T \leq \frac{1}{2} + \frac{M}{4}\left(1 + \log(T-1)\right)$.
- Otherwise $R_T \leq 7.5 + 6\left(2\sqrt{\pi} + \sqrt{2}\right)\sqrt{T-1}$.

A second best of both worlds

From [Bachoc et al., 2024].

Upper bound: follow the ρ under density bounded by M

We have

$$R_T \leq \mathrm{constant}\left((\log T)^2 + M^2 \log T + M^4\right).$$

Some proof ideas:

- If $\hat{\nu}_t = \frac{1}{t}\sum_{s=1}^{t} \delta_{x_s}$ from a sample $x_1, \ldots, x_t$, with

$$0 \leq x_1 < \cdots < x_k < \frac{1}{t}\sum_{s=1}^{t} x_s < x_{k+1} < \cdots < x_t \leq 1$$

  then $\underset{p\in[0,1]}{\mathrm{argmax}}\ \rho(\hat{\nu}_t)(p) \in \{x_k, x_{k+1}\}$.
- As a consequence, with F the CDF of ν, $\bar{\nu}$ the mean of ν,

$$\mathbb{E}\left[(F(\bar{\nu}) - F(P_t)) \cdot (\bar{\nu} - P_t)\right] \leq \frac{408\log(2t)}{t} + \frac{2 + 33M + 32M^2}{t} + 4\exp\left(\frac{-t}{2048M^4}\right).$$

Conclusion

	No independence	Independence No bounded density	Both
Two-bit feedback (non-contextual)	T?	T	$\sqrt{T}$
Full feedback (non-contextual)		$\sqrt{T}$	$\log(T)$
Two-bit feedback (linear-contextual)	T?	T	$\sqrt{dT}$
Two-bit feedback (Lipschitz-contextual)	T?	T	$T^{\frac{d+2}{d+4}}$

Table: Rates according to assumptions on V_t, W_t and feedback.

- Non-contextual setting: [Bolić et al., 2024].
- Non-contextual setting with additional results (follow the ρ in full-feedback for bounded densities) [Bachoc et al., 2024].
- Contextual linear: [Bachoc et al., 2025a].
- Contextual Lipschitz [Bachoc et al., 2025b].

Open questions:

- tighter dependence with respect to some parameters,
- sparsity, low ambient dimension.

Thank you for your attention!

Bibliography I

Bachoc, F., Bolić, N., Cesari, T., and Colomboni, R. (2024).
An online learning theory of brokerage.
hal-04987905v1.

Bachoc, F., Cesari, T., and Colomboni, R. (2025a).
A parametric contextual online learning theory of brokerage.
In *The Forty-second International Conference on Machine Learning*.

Bachoc, F., Cesari, T., and Colomboni, R. (2025b).
A tight regret analysis of non-parametric repeated contextual brokerage.
In *The 28th International Conference on Artificial Intelligence and Statistics*.

Bolić, N., Cesari, T., and Colomboni, R. (2024).
An online learning theory of brokerage.
In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 216–224.

Bibliography II

Cesa-Bianchi, N., Cesari, T., Colomboni, R., Fusco, F., and Leonardi, S. (2024).
Bilateral trade: A regret minimization perspective.
Mathematics of Operations Research, 49(1):171–203.

Kleinberg, R. and Leighton, T. (2003).
The value of knowing a demand curve: Bounds on regret for online posted-price auctions.
In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE.

Lattimore, T. and Szepesvári, C. (2020).
Bandit algorithms.
Cambridge University Press.

Tullii, M., Gaucher, S., Merlis, N., and Perchet, V. (2024).
Improved algorithms for contextual dynamic pricing.
Advances in Neural Information Processing Systems, 37:126088–126117.