Signaled Bandits

Carlos Gonzalez

University of Oxford - Department of Economics

Econometrics Lunch Seminar

June 10th, 2025

Small Recap

	Arm 1		Arm 2	
Signaled Bandit	μ_1	$N(\mu_1, \sigma_r^2)$, $N(\mu_2, \sigma_s^2)$	μ_2	$N(\mu_1, \sigma_s^2)$, $N(\mu_2, \sigma_r^2)$
Bandit	μ_1	$N(\mu_1, \sigma^2)$, $N(\mu_2, \infty)$	μ_2	$N(\mu_1, \infty)$, $N(\mu_2, \sigma^2)$
Experts Problem	μ_1	$N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$	μ_2	$N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$
Reversed Bandit	μ_1	$N(\mu_1, \infty)$, $N(\mu_2, \sigma^2)$	μ_2	$N(\mu_1, \sigma^2)$, $N(\mu_2, \infty)$

Small Recap

	Arm 1		Arm 2	
Signaled Bandit	μ_1	$N(\mu_1, \sigma_r^2)$, $N(\mu_2, \sigma_s^2)$	μ_2	$N(\mu_1, \sigma_s^2)$, $N(\mu_2, \sigma_r^2)$
Bandit	μ_1	$N(\mu_1, \sigma^2)$, $N(\mu_2, \infty)$	μ_2	$N(\mu_1, \infty)$, $N(\mu_2, \sigma^2)$
Experts Problem	μ_1	$N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$	μ_2	$N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$
Reversed Bandit	μ_1	$N(\mu_1, \infty)$, $N(\mu_2, \sigma^2)$	μ_2	$N(\mu_1, \sigma^2)$, $N(\mu_2, \infty)$

Interpretation

- I was asked to provide a motivation for this model
- **Reinterpret** the problem
  - Learner gets for **free** a **low quality signal** with noise σ_s^2 for all arms in every period
  - When selecting arm k (i) learner gets the **returns** of arm k, and (ii) she **improves the signal quality** from $\sigma_s^2 \to \sigma_r^2$
- I will now discuss two general application frameworks through specific toy examples

Model Application I: Batched Treatment Effects

- A bus company owns a fleet of n buses
- According to its license it must operate at least $\bar{n}$ buses in each of the G routes, such that $n > G\bar{n}$
- There are CRS, i.e. each coach in route g gives the same expected return μ_g

Model Application I: Batched Treatment Effects

- It must decide between a grid of batch allocations, i.e. $\{n - G\bar{n}, \bar{n}, \bar{n}, \ldots, \}, \{\bar{n}, n - G\bar{n}, \bar{n}, \ldots\}, \{n/G, n/G, \ldots\}$
- Because the learner does not optimize (learning nor reward wise) over $\bar{n}$ buses in each route, she can take these buses as the free low quality signal

Model Application I: Batched Treatment Effects

- When selecting batch k
  - The bus company obtains an (expected) return of $\mu_k = \sum_g^G n_k^g \mu_g$
  - It observes n_k^g realizations of each route reward Y_i^g
  - These realizations are enough to recover μ_k for all k (as they are enough to recover μ_g)
  - BUT, selecting route k remains most informative about μ_k and no other route is more informative about μ_k than k, i.e. $\sigma_r^2 \leq \sigma_s^2$
- Both sources of information can be combined à la GLS and my results apply

Model Application I: Batched Treatment Effects

- When selecting batch k
  - The bus company obtains an (expected) return of $\mu_k = \sum_g^G n_k^g \mu_g$
  - It observes n_k^g realizations of each route reward Y_i^g
  - These realizations are enough to recover μ_k for all k
  - BUT, selecting route k remains most informative about μ_k and no other route is more informative about μ_k than k, i.e. $\sigma_r^2 \leq \sigma_s^2$
- Both sources of information can be combined à la GLS and my results apply

Model Application II: Parametric Treatment Effects

- There are 2 hospitals h with nursing teams of equal quality. The health officer is trying to assign a single cardiologist to one of them
- Relevant variable is heart-attack risk $Y_i^h \sim \mathrm{Log}(\mu_h, s)$ with unkown μ_h
- The cardiologist holds an advantage wrt nursing teams in two dimensions
  - nurses can only detect if health risk $Y_i^h \geq \bar{y}$ while cardiologist can observe Y_i^h
  - cardiologist can treat severely ill patients better, so, you want to have him in the hospital with highest μ_h

Model Application II: Parametric Treatment Effects

- Under parametric (logistic) assumptions, μ_h is identified from the low quality signal $\mathbb{P}(Y_i^h \geq \bar{y})$ (via MLE for instance)
- But $\mathrm{Var}(\hat{\mu}_h^{\mathrm{MLE}}) > \mathrm{Var}(\hat{\mu}_h)$ (as it uses less information)
- Both estimators can be combined à la GLS and our results apply (approximately)

Proof for `GUCB`

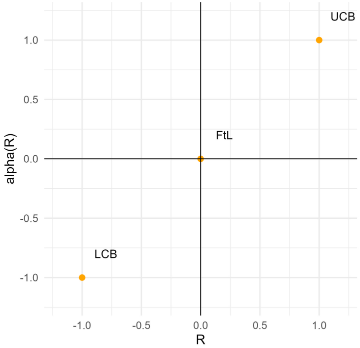
- The proof for an upper bound on the GLS-enchanced versions of `UCB`/`LCB` was missing
- I show that `GUCB`
  - Achieves a regret $\leq \sqrt{24\sigma_r^2 N \ln N \frac{\sigma_s^2}{\sigma_s^2+\sigma_r^2}}$ in the **general signaled bandit game**
  - When $\sigma_s^2 = \infty$, it matches the upper bound of `UCB` in the bandit setting
  - When $\sigma_s^2 = \sigma_r^2$, it saves a factor of $\sqrt{2}$

Interpolating Mechanism

- Having the interpolating mechanism as a function of (σ_s^2/σ_r^2) was very inconvenient (non-symmetric)

- Instead, more natural to have it as a function of $R = (\sigma_s^2 - \sigma_r^2)/\max\{\sigma_r^2, \sigma_s^2\}$

Problem	R	Algorithm	α
Bandit ($\sigma_s^2 = \infty$)	1	`UCB`: $\hat{\mu} + B$	1
Expert ($\sigma_s^2 = \sigma_r^2$)	0	`FtL`: $\hat{\mu}$	0
Rev Bandit ($\sigma_r^2 = \infty$)	-1	`LCB`: $\hat{\mu} - B$	-1

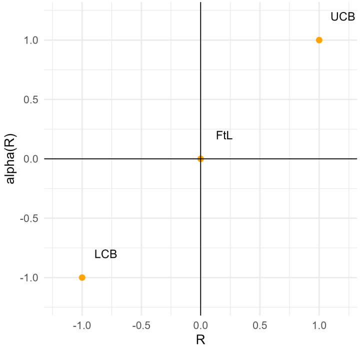


Interpolating Mechanism

- Having the interpolating mechanism as a function of (σ_s^2/σ_r^2) was very inconvenient (non-symmetric and difficult to interpret)

- Instead, more natural to have it as a function of $R = (\sigma_s^2 - \sigma_r^2)/\max\{\sigma_r^2, \sigma_s^2\}$

Problem	R	Algorithm	α
Bandit $(\sigma_s^2 = \infty)$	1	`UCB`: $\hat{\mu} + B$	1
Expert $(\sigma_s^2 = \sigma_r^2)$	0	`FtL`: $\hat{\mu}$	0
Rev Bandit $(\sigma_r^2 = \infty)$	-1	`LCB`: $\hat{\mu} - B$	-1



Tuning the Interpolating Mechanism

- Natural candidate to interpolate a function which goes through these three points is R^z,
  - but how to select the optimal z? (the convexity/concavity of the function)
  - I can't use insights from the proof yet (because I don't have it)
  - But I have an interesting heuristic! (and very strong simulation evidence)
- $\mathscr{R}(\texttt{FtGL}) \leq \sqrt{8N(\sigma_r^2 + \sigma_s^2)}$ $\qquad$ $\mathscr{R}(\texttt{GUCB}) \leq \sqrt{24\sigma_r^2 N \ln N^{\frac{\sigma_s^2}{\sigma_s^2+\sigma_r^2}}}$
- Let $\tilde{R}$ be the R which matches the upper bounds on regret, i.e. $\text{Upper}(\texttt{FtGL}) = \text{Upper}(\texttt{GUCB})$

Tuning the Interpolating Mechanism

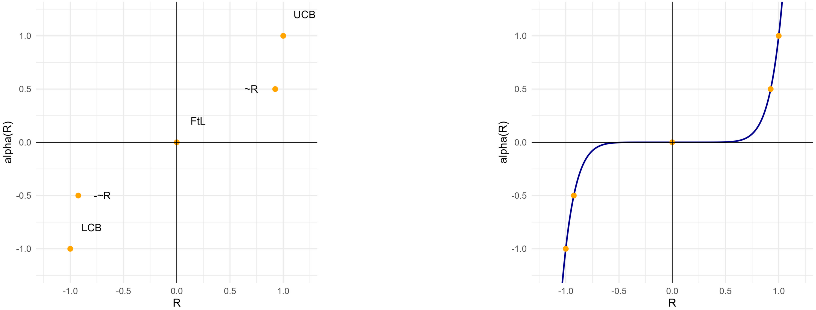
- Natural candidate to interpolate a function which goes through these three points is R^z,
  - but how to select the optimal z? (the convexity/concavity of the function)
  - I haven't proved a sharp upper bound on regret for the interpolating algorithm yet, so I can't use the insights from the proof to motivate z
  - I have an interesting heuristic! (and very strong simulation evidence)

- $\mathcal{R}(\texttt{FtGL}) \leq \sqrt{8N(\sigma_r^2 + \sigma_s^2)}$ $\qquad$ $\mathcal{R}(\texttt{GUCB}) \leq \sqrt{24\sigma_r^2 N \ln N \frac{\sigma_s^2}{\sigma_s^2+\sigma_r^2}}$

- Let $\tilde{R}$ be the R which matches the upper bounds on regret, i.e. $\text{Upper}(\texttt{FtGL}) = \text{Upper}(\texttt{GUCB})$

Tuning the Interpolating Mechanism

- Impose $f(\tilde{R}) = 1/2$
- Select z^* such that $\tilde{R}^{z*} \approx 1/2$, i.e. $z^* = \log(1/2)/\log(\tilde{R})$



$$N = 200, z^* \approx 8.83$$

Tuning the Interpolating Mechanism

- Simulation wise this selection does very well, but
  - slightly bigger $z > z*$ seems to do even better
  - it does not really outperform `FtGL` in cases with $\sigma_s^2 \approx \sigma_r^2$ (I might now why this is the case)

Moving Forward

- Crack proof for `Inter-CB`
- Find real world applications. Shouldn't be very difficult (?!) Any suggestions are welcome
- Reconcile existing literature
- Technical stuff (K arms, mixed variance, adversarial feedback, etc.)