GMM Bandits

Carlos Gonzalez

University of Oxford - Department of Economics

Econometrics Lunch Seminar

April 29, 2025

Outline

Outline

Signaled Bandits

- **Learner** faces K **arms** (options) for N periods. She selects one arm per period

- When playing arm k she receives an **unobserved reward** μ_k and observes a **signal about all arms** $k' \in \{1, \ldots, K\}$ drawn from distribution $N(\mu_{k'}, \sigma^2_{k'k})$

- Her goal is to find a policy which maximizes the cumulative sum of rewards, or, equivalently, **minimizes regret**

  - **Regret:** The expected difference between the rewards of the sequence of arms selected by the policy and the rewards of the best arm

Today's Talk

- Today,
  - $K = 2$
    - **rewards** to the signals of selected arms about themselves, and
    - **signals** to the signals of the selected arms about the other arms
  - Normal signals (only subgauss is required, i.e. $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2 \sigma^2 / 2)$)
  - Symmetric Variances: $\sigma_k^2 = \sigma_{k'}^2 = \sigma_r^2$ and $\sigma_{k'k}^2 = \sigma_{kk'}^2 = \sigma_s^2$
  - Variances well-separated from 0, i.e. $\sigma_r^2, \sigma_s^2 \geq 1$

Limiting Regimes

	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	$N(\mu_1, \sigma^2)$, $N(\mu_2, \infty)$

Limiting Regimes

	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	$N(\mu_1, \sigma^2)$, $N(\mu_2, \infty)$

Limiting Regimes

	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	$N(\mu_1, \sigma^2)$, $N(\mu_2, \infty)$

Limiting Regimes

	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	$N(\mu_1, \sigma^2)$, $N(\mu_2, \infty)$

Outline

Introduction to Lower Bounds

- A problem has a **lower bound** of $O(N^\alpha)$ if there exists an environment (a collection of μ_k) which induces a regret of at least $O(N^\alpha)$ on **any policy**

- It boils down to find a **difficult** environment in the game

	Regret when $A_i = 2$	Difficulty to Identify Optimal Arm
$\mu_1 >> \mu_2$	High	Low
$\mu_1 \approx \mu_2$	Low	High

- The (reversed) **bandit** problem is **easier**, and the **experts'** problem is **harder** than the signaled bandit game. Thus, we expect

$$L(\text{Experts}) \leq L(\text{Signaled Bandit}) \leq L(\text{Bandit})$$

Lower Bound on Signaled Bandit

- Small problem. In the general K arm game

$$L(\text{Experts}) = C\sqrt{N\sigma^2} \quad L(\text{Bandit}) = C\sqrt{N(K-1)\sigma^2}$$

- Bounds match for the case $K = 2$. To avoid this problem, assume that every policy queries the bad arm less than $N/\bar{K}$ (with $\bar{K} \geq 1$), then

Theorem 3.5. Lower Bounds on Signaled Bandits

For any policy π there exists a two-arm signaled-bandit problem P st

$$\mathcal{R}_N(\pi, P) \geq \frac{1}{27}\sqrt{\frac{N\bar{K}\sigma_r^2\sigma_s^2}{(\bar{K}-1)\sigma_r^2 + \sigma_s^2}}$$

Lower Bounds on Signaled Bandits

- The general signaled bandit game (with $\sigma_s^2, \sigma_r^2 \leq \infty$) is indeed easier than bandit game and more difficult than the experts' game

	Lower Bound
Experts Problem	$C\sqrt{N\sigma^2}$
Bandit Problem	$C\sqrt{N\bar{K}\sigma_r^2}$
Signaled Bandit Problem	$\left[C\sqrt{N\sigma_r^2}, C\sqrt{N\bar{K}\sigma_r^2}\right]$

Lower Bound Proof

Outline

Upper Bounds on Limiting Regimes

	Algorithm	**Index**	**Regret**
Experts' Problem	`FtL`	$\hat{\mu}_{ki}$	$O\left(\sqrt{N}\right)$
Bandit Problem	`UCB`	$\hat{\mu}_{ki} + \sqrt{\frac{2\ln f(i)}{N_k(i)}}$	$O\left(\sqrt{N \ln N}\right)$
Reversed Bandit	`LCB`	$\hat{\mu}_{ki}\left(N_{k'(i)}\right) - \sqrt{\frac{2\ln f(i)}{N_{k'}(i)}}$	$O\left(\sqrt{N \ln N}\right)$

Intermediate Regimes and GMM

- We first need a mechanism to combine information **efficiently** (coming from signals and rewards)
- **GMM is the perfect candidate!**

$$\hat{\mu}_k^{\text{GMM}} = \frac{\frac{N_k}{\sigma_r^2}\hat{\mu}_{kr} + \frac{N_{k'}}{\sigma_s^2}\hat{\mu}_{ks}}{\frac{N_k}{\sigma_r^2} + \frac{N_{k'}}{\sigma_s^2}}$$

Pure Exploitation in Intermediate Regimes

- Do we need any exploration when $\sigma_s^2, \sigma_r^2 < \infty$?
- Follow the GMM Leader `FtGL` $\arg\max_k \hat{\mu}_k^{\text{GMM}}$

Theorem 5.1 Upper Bound on FtGL

For any two-arm stochastic signaled-bandit problem, the regret of `FtGL` satisfies

$$\mathscr{R}_N^{\texttt{FtGL}} \leq \sqrt{8N(\sigma_r^2 + \sigma_s^2)}$$

- For fixed σ_s^2 it remains asymptotically optimal with regret $O\left(\sqrt{N}\right)$, but for $\sigma_s^2 \geq \ln N$, a confidence bound based algorithm like `UCB` must dominate `FtGL`

GMM based Confidence Bound Algorithms

- High confidence bound for a GMM-based Confidence Bound Algorithm and adapt limiting-regimes optimal algorithms

Algorithm	Index	Regret
`GUCB`	$\hat{\mu}_k^{\mathrm{GMM}} + \sqrt{\frac{2\ln f(i)\sigma_r^2\sigma_s^2}{N_k\sigma_s^2+N_{k'}\sigma_r^2}}$	**?**
`GLCB`	$\hat{\mu}_k^{\mathrm{GMM}} - \sqrt{\frac{2\ln f(i)\sigma_r^2\sigma_s^2}{N_k\sigma_s^2+N_{k'}\sigma_r^2}}$	**?**

- For $\sigma_s^2 \to \infty$, `GUCB` $\to$ `UCB`   For $\sigma_r^2 \to \infty$, `GLCB` $\to$ `LCB`

- Conjecture excessive exploration in high and full-information regimes (?)

Small Recap

- `FtGL` is asymptotically-optimal in the signaled bandit game but suffers badly for large σ_s^2 or σ_r^2
- `GUCB` (`GLCB`) has the potential to outperform `FtGL` in regimes with high σ_s^2 (high σ_r^2)
- Can we get an algorithm which is near-optimal across all regimes?

An Optimal Algorithm in the Signaled Bandit Game

- Look like GUCB for $\sigma_s^2 >> \sigma_r^2$, like FtGL for $\sigma_s^2 \approx \sigma_r^2$, and like GLCB for $\sigma_s^2 << \sigma_r^2$

- GMM Confidence Bound Algorithm GMM-CB

$$\hat{\mu}_{ki}^{\text{GMM}} + g(\sigma_r^2/\sigma_s^2) \cdot \sqrt{\frac{2 \ln f(i) \sigma_r^2 \sigma_s^2}{N_{k'i}\sigma_r^2 + N_k \sigma_s^2}}$$

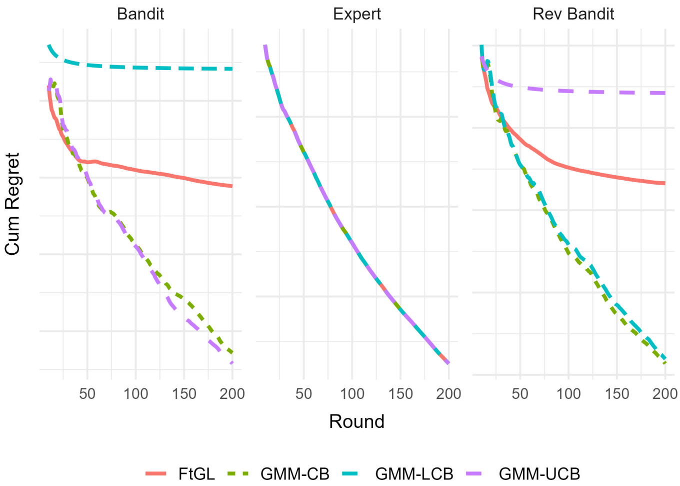
- where $g(\cdot)$
  - $\lim_{R \to 0} g(R) = 1$, $g(1) = 0$ and $\lim_{R \to \infty} g(R) = -1$
  - is a continuous weakly decreasing function
  - $g(\sigma_r^2/\sigma_s^2) = -g(\sigma_s^2/\sigma_r^2)$

Selection of g

Alternative Approach

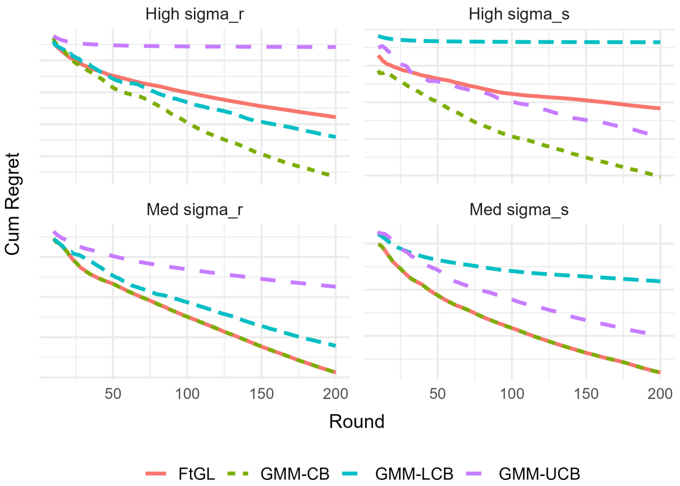
Outline

Limiting Regimes



$$\mu_1 = 1 > 0.9 = \mu_2$$

Intermediate Regimes

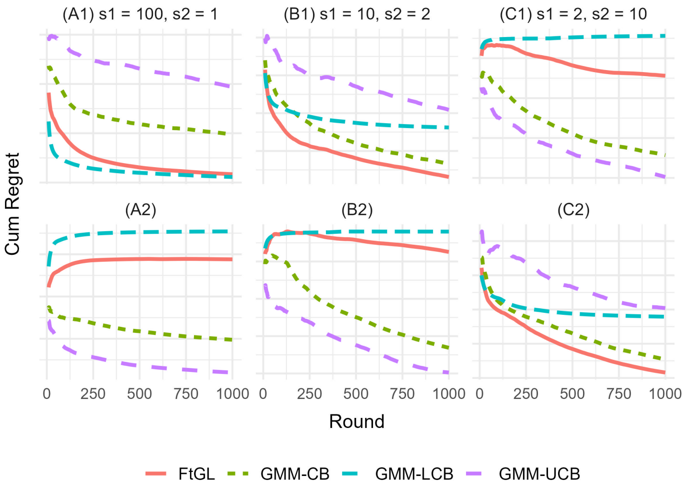


$\mu_1 = 1 > 0.9 = \mu_2$ High $\sigma^2 = 10$, Medium $\sigma^2 = 2$

Mixed Variance

- In regimes with $\sigma_{12}^2 \neq \sigma_{21}^2$, weight symmetry is broken
  - **Example:** $\sigma_r^2 = 1, \sigma_{12}^2 = 1, \sigma_{21}^2 = \infty$. Then weight $g(\sigma_1^2/\sigma_{12}^2) = g(1) = 0$ and $g(\sigma_2^2/\sigma_{21}^2) = g(0) = 1$. Thus arm 2 gets oversampled regardless μ_1, μ_2
- **Fix:** Restore symmetry by setting $g = (g_1 + g_2)/2$
- Good empirical results but no theory whatsoever for this regime. Left in the dark as there is no general theory of limiting-optimal algorithms in "one-sided bandits"

Mixed Variance



$$(\mathrm{X1})\ \mu_1 = 1 > 0.9 = \mu_2,\ (\mathrm{X2})\ \mu_1 = 0.9 < 1 = \mu_2,\ \sigma_{r1}^2 = \sigma_{r2}^2 = 1$$

Outline

Conclusion

- **Unified** Bandit and Experts' Problems in a single framework through Signaled Bandits
- Explored **Reversed Bandits** and derived a near-optimal algorithm `LCB`
- Derived a **lower bound** for the Signaled Bandit game and showed that $L(\text{Experts}) \leq L(\text{Signaled}) \leq L(\text{Bandit})$
- Introduced the GMM logic to Signaled Bandit games and mainstream algorithms `FtGL`, `GUCB`, `GLCB`
- Derived an upper bound for `FtGL`

Next Steps

- Upper Bound for `GUCB` and `GLCB` in the Signaled Bandit game
- Upper Bound for `GMM-CB` in the Signaled Bandit game (and optimality claims)
- Extensions to the $K > 2$ arm game
- Asymmetric variances
- Applications
- Your feedback

Proof Lower Bound

$$P = (P_1, P_2) = ((N(\Delta, \sigma_r^2), N(0, \sigma_s^2)), (N(0, \sigma_r^2), N(\Delta, \sigma_s^2)))$$
$$P' = (P_1', P_2') = ((N(\Delta, \sigma_r^2), N(2\Delta, \sigma_s^2)), (N(2\Delta, \sigma_r^2), N(\Delta, \sigma_s^2)))$$

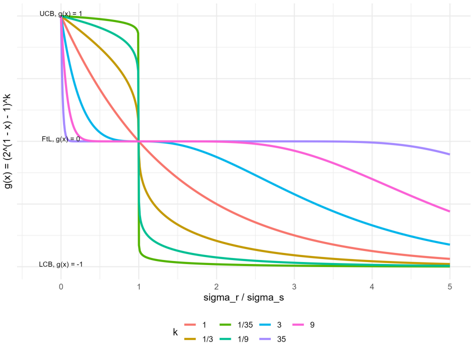
$$\begin{aligned}
&\mathcal{R}_N(\pi, P) + \mathcal{R}_N(\pi, P') \\
&> \mathbb{P}_P(N_1(N) \leq N/2) \cdot \Delta \frac{N}{2} + \mathbb{P}_{P'}(N_1(N) > N/2) \cdot \Delta \frac{N}{2} \\
&> \Delta \frac{N}{4} \exp(-D(\mathbb{P}_P, \mathbb{P}_{P'})) && \text{BH Ineq} \\
&= \Delta \frac{N}{4} \exp\left(- \sum_k \mathbb{E}_P[N_k(N)] \cdot \left(D(P_{kr}, P_{kr}') + D(P_{k's}, P_{k's}') \right) \right) && \text{Div Dec}
\end{aligned}$$

Algebra + Worst Case Selection of Δ

Return

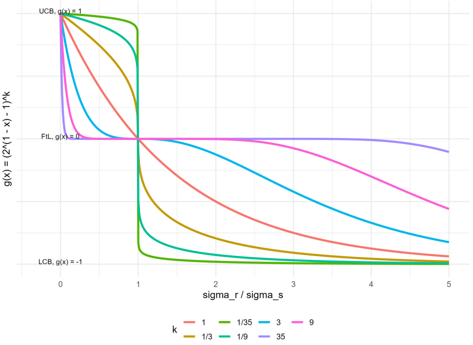
$g(\cdot)$ Selection

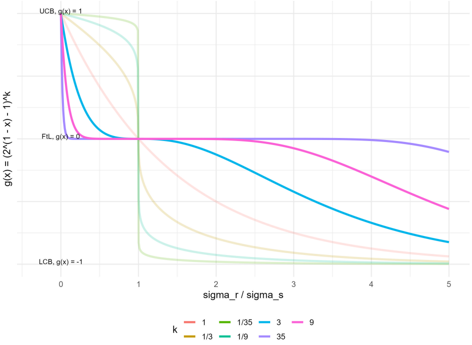
$$g(x) = (2^{1-x} - 1)^k$$



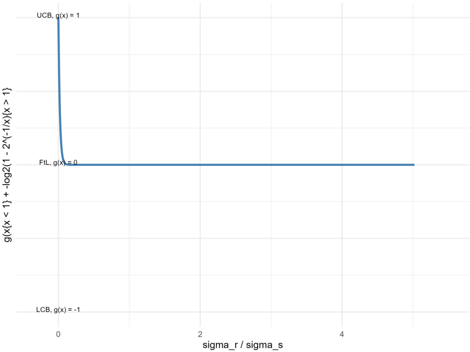
$g(\cdot)$ Selection

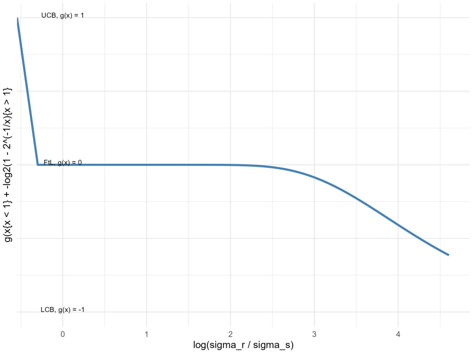
$$g(x) = (2^{1-x} - 1)^k$$





$g(\cdot)$ Selection





$$g(x) = (2^{1-f(x)} - 1)^{35} \text{ with } f(x) = x\{x \leq 1\} - \log_2(1 - 2^{-1/x})\{x > 1\}$$

Return

An Alternative Approach

$$\hat{\mu}_{ki}^{\text{GMM}} + h(\sigma_r^2/\sigma_s^2) \cdot \sqrt{\frac{2 \ln f(i) \sigma_r^2 \sigma_s^2}{N_{k'i}\sigma_r^2 + N_k \sigma_s^2}} + (1 - h(\sigma_r^2/\sigma_s^2)) \cdot \sqrt{\frac{2 \ln f(i) \sigma_r^2 \sigma_s^2}{N_{ki}\sigma_r^2 + N_{k'} \sigma_s^2}}$$

- where $h(\cdot)$
  - $\lim_{R \to 0} h(R) = 1$, $h(1) = 1/2$ and $\lim_{R \to \infty} h(R) = 0$
  - is a continuous weakly decreasing function
  - $h(\sigma_r^2/\sigma_s^2) = 1 - h(\sigma_s^2/\sigma_r^2)$
  - Candidate $h(R) = 1/(1+R)$

Return