Asymptotically-Optimal Gaussian Bandits with Side Observations

Alexia Atsidakou (UT Austin)[1]

[1]Joint work w/ O. Papadigenopoulos (UT Austin), C.Caramanis (UT Austin), S.Sanghavi (UT Austin and Amazon) & S.Shakkottai (UT Austin)

Gaussian Bandits with Side Observations

Model (introduced in Wu, György, and Szepesvári 2015):

- K Gaussian arms with (unknown) mean rewards $(\mu_1, \ldots, \mu_K)$
- Known feedback matrix $\Sigma = (\sigma_{i,j})_{i,j \in [K]}$
- At each round t, by playing an action $i \in [K]$ the player:
  - collects $X_{i,t} \sim \mathcal{N}(\mu_i, \sigma_{i,i}^2)$
  - observes $X_{j,t} \sim \mathcal{N}(\mu_j, \sigma_{i,j}^2)$ for each arm $j \in [K]$
  - (rewards are realized independently)

Goal: Maximize the total expected reward collected

Gaussian Bandits with Side Observations

Previous related work:

- Wu, György, and Szepesvári 2015: asymptotically optimal regret for the special case where $\sigma_{i,j} \in \{\sigma, \infty\}$
- Graph-structured feedback: Given (directed) graph of arms (nodes), playing arm i reveals the (not necessarily Gaussian) reward of every adjacent arm

Our contribution:

- Asymptotic LP lower bound for the case of general feedback $\Sigma = (\sigma_{i,j})_{i,j \in [K]}$
- An asymptotically optimal LP-based bandit algorithm for the general setting

Linear Programming-based Lower Bound

Formulation: For any reward vector $\mu \in [0, \infty)^K$, we define:

$$C(\mu) = \left\{ c \in [0, \infty)^K : \begin{array}{c} \sum_{j \in [K]} \frac{c_j}{\sigma_{j,i}^2} \geq \frac{2}{\Delta_i^2(\mu)}, \forall i \neq i^*(\mu) \\ \sum_{j \in [K]} \frac{c_j}{\sigma_{j,i}^2} \geq \frac{2}{\Delta_{\min}^2(\mu)}, i = i^*(\mu) \end{array} \right\},$$

where $i^*(\mu) = \text{argmax}_{i \in [K]} \mu_i$, $\Delta_i(\mu) = \max_{j \in [K]} \mu_j - \mu_i$, and $\Delta_{\min}(\mu) = \min_{i \in [K], \Delta_i(\mu) > 0} \Delta_i(\mu)$.

Theorem

For environment $(\mu, \boldsymbol{\Sigma})$, the regret of any consistent policy satisfies

$$\liminf_{T \to \infty} \frac{R_T(\mu)}{\log T} \geq \min_{c \in C(\mu)} \sum_{i \in [K]} c_i \, \Delta_i(\mu).$$

LP-based Algorithm with Asymptotically Optimal Regret

Notation:

- Let $N_i(t)$ the number of samples arm i has been played so far
- Maximum-Likelihood reward estimator at round t:

$$\widehat{\mu}_i(t) = \sum_{\tau=1}^{t-1} \frac{X_{i,\tau}}{\sigma^2_{i_\tau,i}} \Big/ \sum_{\tau=1}^{t-1} \frac{1}{\sigma^2_{i_\tau,i}} \qquad \forall i \in [K],$$

where i_τ the arm played at round τ

Algorithm: At each round t, the algorithm performs one of the following:

- **Greedy exploitation:** Play the arm of best estimated reward
- **Uniform exploration:** Ensure $C(\widehat{\mu})$ is "close" to $C(\mu)$
- **LP-dictated exploration:** Follow the actions indicated by (estimated) LP based on $C(\widehat{\mu})$

LP-based Algorithm with Asymptotically Optimal Regret

At each round t:

Greedy exploitation: If $\left(\frac{N_1(t)}{\log t}, \frac{N_2(t)}{\log t}, \ldots, \frac{N_K(t)}{\log t}\right) \in C(\widehat{\mu})$, then play

$$\boxed{i_t \leftarrow \arg\max_{i \in [K]} \widehat{\mu}_i(t)}$$

LP-based Algorithm with Asymptotically Optimal Regret

At each round t:

n_e: # exploration rounds (initialized at 0)

Uniform exploration: If $\left(\frac{N_1(t)}{\log t}, \frac{N_2(t)}{\log t}, \ldots, \frac{N_K(t)}{\log t}\right) \notin C(\widehat{\mu})$ and

$$\min_{i \in [K]} \sum_{\tau=1}^{t-1} \frac{1}{\sigma^2_{i_\tau, i}} < o(n_e(t)) \text{ (not uniformly explored)}$$

then play

$$\boxed{i_t \leftarrow \arg\min_{k \in [K]} \sigma^2_{k,i}, \text{ where } i = \arg\min_{k \in [K]} \sum_{\tau=1}^{t-1} \frac{1}{\sigma^2_{i_\tau, k}},}$$

and increase n_e by 1

LP-based Algorithm with Asymptotically Optimal Regret

At each round t:

LP-dictated exploration:

If $\left(\frac{N_1(t)}{\log t}, \frac{N_2(t)}{\log t}, \ldots, \frac{N_K(t)}{\log t}\right) \notin C(\widehat{\mu})$ and arms uniformly explored, then

- Compute $c^*(\widehat{\mu}(t)) \leftarrow \arg\min_{c \in C(\widehat{\mu}(t))} \sum_{i \in [K]} c_i \, \Delta_i(\widehat{\mu}(t))$
- Play arm

$$\boxed{i_t = i \text{ with } N_i(t) < c_i^*(\widehat{\mu}(t)) \log t,}$$

and increase n_e by 1

LP-based Algorithm with Asymptotically Optimal Regret

Algorithm: At each round t, the algorithm performs either:

- **Greedy exploitation:** Play the arm of best estimated reward
- **Uniform exploration:** Ensure $C(\widehat{\mu})$ is "close" to $C(\mu)$
- **LP-dictated exploration:** Follow the actions indicated by (estimated) LP based on $C(\widehat{\mu})$

Theorem

The regret of our algorithm satisfies

$$\limsup_{T\to\infty} \frac{R_T(\mu)}{\log T} \leq \sum_{j\in[K]} \Delta_j(\mu) c_j^*(\mu) \qquad \textit{(up to constant factors)}$$

References

Wu, Yifan, András György, and Csaba Szepesvári (2015). "Online Learning with Gaussian Payoffs and Side Observations". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, pp. 1360–1368.