Handling Heterogeneous Curvatures in Bandit LQR Control

Yu-Hu Yan, Jing Wang, Peng Zhao (*Spotlight*)

LAMDA Group, Nanjing University, China

ICML International Conference On Machine Learning

NANJING UNIVERSITY

LAMDA Learning And Mining from DatA

Our Paper (OpenReview) Homepage (Yu-Hu Yan)

Problem Setup

Linear Quadratic Regulator (LQR):

LQR aims to control the following *linear dynamical system*:

$$\begin{aligned}\mathbf{x}_{t+1} &= A\mathbf{x}_t + B\mathbf{u}_t + \boldsymbol{\xi}_t \\ \mathbf{y}_t &= C\mathbf{x}_t + \mathbf{e}_t \\ &\textit{(partial observation)}\end{aligned}$$

- A, B: known system transition matrices
- $\mathbf{x}_t, \mathbf{u}_t$: state and action
- $\boldsymbol{\xi}_t$: disturbances (noises)

LQR aims to minimize the *cumulative cost*:

$$\min_{\pi\in\Pi}\sum_{t=1}^{T} c_t(\mathbf{x}_t^\pi, \mathbf{u}_t^\pi) \triangleq \sum_{t=1}^{T} {\mathbf{x}_t^\pi}^\top Q_t \mathbf{x}_t^\pi + \sum_{t=1}^{T} {\mathbf{u}_t^\pi}^\top R_t \mathbf{u}_t^\pi$$

(*time-varying* PSD Q_t, R_t)

In ***online*** LQR problem, we focus on *policy regret*: [Dekel et al., ICML 2012]

$$\sum_{t=1}^{T} c_t(\mathbf{x}_t, \mathbf{u}_t) - \min_{\pi\in\Pi}\sum_{t=1}^{T} c_t(\mathbf{x}_t^\pi, \mathbf{u}_t^\pi)$$

π: a specific policy chosen from a predefined policy class Π

Bandit cost: Learner only observes a scalar value of $c_t(\mathbf{y}_t, \mathbf{u}_t)$.

Oblivious adversary: Costs and disturbances are chosen ***in advance***.

Assumptions:

① **Stability:** The system is stable, i.e., the spectral radius $\rho(A) < 1$.

(A standard assumption that can be extended to strongly stabilizable systems.)

② **Disturbances:** *semi-adversarial* disturbances [Simchowitz et al., COLT 2020]

$$\boldsymbol{\xi}_t = \boldsymbol{\xi}_t^{\text{adv}} + \boldsymbol{\xi}_t^{\text{sto}}, \quad \mathbf{e}_t = \mathbf{e}_t^{\text{adv}} + \mathbf{e}_t^{\text{sto}}$$

Adversarial condition: $\|\boldsymbol{\xi}_t\|_2, \|\mathbf{e}_t\|_2 \le W$.

Stochastic condition: *semi-adversarial* = *adversarial* + *stochastic*

$$\mathbb{E}[\boldsymbol{\xi}_t^{\text{sto}}] = \mathbb{E}[\mathbf{e}_t^{\text{sto}}] = \mathbf{0},\ \mathbb{E}[\boldsymbol{\xi}_t^{\text{sto}}{\boldsymbol{\xi}_t^{\text{sto}}}^\top] \succeq \text{Var}_{\boldsymbol{\xi}}\cdot I,\ \mathbb{E}[\mathbf{e}_t^{\text{sto}}{\mathbf{e}_t^{\text{sto}}}^\top] \succeq \text{Var}_{\mathbf{e}}\cdot I.$$

*(A standard assumption when dealing with **strongly convex** cost functions.)*

③ **Costs:** *heterogeneous* quadratic costs

quadratic, *heterogeneous*: $Q_t, R_t \succeq \alpha_t I$ smooth: $\nabla^2 c_t(\cdot,\cdot) \preceq \beta_c I$

Lipschitz: $|c_t(\mathbf{y},\mathbf{u}) - c_t(\mathbf{y}',\mathbf{u}')| \le L_c R_c \|(\mathbf{y}-\mathbf{y}', \mathbf{u}-\mathbf{u}')\|_2$

Motivation: Why heterogeneous curvatures?

(i) adaptive to true curvatures of costs: capable of using more curvature information (since $\alpha_t \ge \alpha_{\min}$) for better performance. $\alpha_{\min}$ is hard to obtain before algorithms start.

(ii) robust to corrupted quadratic costs: allowing $\alpha_t = 0$ for some rounds where algorithm using only $\alpha_{\min} = 0$ will perform poorly.

Key Question: *Is it possible design an algorithm that is*
① adaptive to heterogeneous curvatures for better performance and
② robust to (possibly) corrupted cost functions?

Our Results

A general theorem for bandit LQR with heterogeneous curvatures

Theorem 1. *By choosing the regularization coefficients $\{\lambda_t\}_{t=0}^T$ and step sizes $\{\eta_t\}_{t=1}^T$ properly, for Lipschitz and α_t-quadratic costs, we obtain*

$$\mathbb{E}[\text{Reg}_T] \le \widetilde{\mathcal{O}}\left(\inf_{\{\lambda_1^\star,\ldots,\lambda_T^\star\}}\left\{T^{1/3} + \lambda_{1:T-1}^\star + \sum_{t=1}^{T-1}(\alpha_{1:t} + \lambda_{0:t}^\star)^{-1/3}\right\}\right).$$

For Lipschitz, smooth, and α_t-quadratic costs, we obtain

$$\mathbb{E}[\text{Reg}_T] \le \widetilde{\mathcal{O}}\left(\inf_{\{\lambda_1^\star,\ldots,\lambda_T^\star\}}\left\{\sqrt{T} + \lambda_{1:T-1}^\star + \sum_{t=1}^{T-1}(\alpha_{1:t} + \lambda_{0:t}^\star)^{-1/2}\right\}\right),$$

where $\{\lambda_t^\star\}_{t=1}^T$ represent the optimal regularization coefficients.

- Coefficients $\{\lambda_t^\star\}_{t=1}^T$ are only for analysis, so one may plug in any feasible $\{\lambda_t^\dagger\}_{t=1}^T$.
- By choosing different $\{\lambda_t^\dagger\}_{t=1}^T$ *in analysis*, we achieve different guarantees (pure convexity, pure quadraticity, corrupted quadraticity, decaying quadraticity).

	Convexity ($\alpha_t = 0$)	Quadraticity ($\alpha_t = \alpha > 0$)	Corrupted Quadraticity ($\alpha_t = \alpha\cdot\mathbb{1}_{t\notin\mathcal{T}}$)	Decaying Quadraticity ($\alpha_t = t^{-\gamma}$)
Lipschitz Functions	$\widetilde{\mathcal{O}}(T^{3/4})$ (Gradu et al., 2020)	N/A	N/A	N/A
	$\widetilde{\mathcal{O}}(T^{3/4})$	$\widetilde{\mathcal{O}}(T^{2/3})$	(when $\lvert\mathcal{T}\rvert = T^{8/9}$) $\widetilde{\mathcal{O}}(T^{2/3})$	$\begin{cases}\widetilde{\mathcal{O}}(T^{2/3+\gamma/3}), & \gamma\in[0,1/4]\\ \widetilde{\mathcal{O}}(T^{3/4}), & \gamma\in(1/4,1]\end{cases}$
Smooth Functions	$\widetilde{\mathcal{O}}(T^{2/3})$ (Cassel and Koren, 2020)	$\widetilde{\mathcal{O}}(\sqrt{T})$ (Sun et al., 2023)	N/A	N/A
	$\widetilde{\mathcal{O}}(T^{2/3})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	(when $\lvert\mathcal{T}\rvert = T^{3/4}$) $\widetilde{\mathcal{O}}(\sqrt{T})$	$\begin{cases}\widetilde{\mathcal{O}}(T^{1/2+\gamma/2}), & \gamma\in[0,1/3]\\ \widetilde{\mathcal{O}}(T^{2/3}), & \gamma\in(1/3,1]\end{cases}$

Why not consider more general ***strongly convex*** functions?

*Our results rely on the **with-history reduction** scheme [Sun et al., NeurIPS 2023], which needs quadraticity.*

Reduction to BCO with Memory

Technique I: With-History Reduction to BCO with Memory

A common parametrization: ***Nature's y*** and ***disturbance-response policy***

Nature's y: $\mathbf{y}_t^{\text{nat}} \triangleq \mathbf{e}_t + \sum_{i=1}^{t-1} CA^{i-1}\boldsymbol{\xi}_{t-i}$

Disturbance-Response Policy: $\mathbf{u}_t(M) = \sum_{i=0}^{m-1} M^{[i]}\mathbf{y}_{t-i}^{\text{nat}}$ [Simchowitz et al., COLT 2020]

Previous works use *truncation* to define a with-memory function:

⇨ ***Drawback:** truncation error is hard to control in this problem.*

Ours: The with-history loss function is defined as [Sun et al., NeurIPS 2023] .

$$f_t(N_0,\ldots,N_H) \triangleq c_t\left(\mathbf{y}_t^{\text{nat}} + \sum_{i=1}^{H} G^{[i]}\mathbf{u}_{t-i}(N_{H-i}) + \sum_{i=H+1}^{t-1} G^{[i]}\mathbf{u}_{t-i}(M_{t-i}), \mathbf{u}_t(N_H)\right).$$

history more than H steps before

- ***Essential idea:** **absorbing** history into c_t to form the "with-history" function, leading to a **lossless** reduction.*
- ***Benefit with price**: leading to **non-oblivious** $f_t(\cdot)$ for BCO problems, but fortunately can be handled thanks to **obliviousness of** $c_t(\cdot)$.*

Handling Heterogeneous Curvatures

Technique II: Stability Analysis using Newton Decrement

BCO with heterogeneous curvatures: [Luo et al., COLT 2022]

Consider a general setup with online functions $\{h_t(\cdot)\}_{t=1}^T$ and decisions $\{w_t\}_{t=1}^T$.

$$\text{Reg}_T \triangleq \mathbb{E}\left[\sum_{t=1}^{T} h_t(w_t) - \min_{w\in\mathcal{W}}\sum_{t=1}^{T} h_t(w)\right]$$

heterogeneous curvatures: $\nabla^2 h_t(\cdot) \succeq \sigma_t I$

- For heterogeneous curvatures, this work follows AOGD (OGD with carefully designed *time-varying step sizes*) [Bartlett et al., NIPS 2008].
- For bandit feedback with strongly convex functions, *FTRL* is adopted.

⇨ *FTRL* with *time-varying step sizes* for BCO with heterogeneous curvatures.

Issue: A crucial term in regret analysis is $\|\nabla\psi(w)\|_{\nabla^{-2}\psi(w)}$. ($\psi(\cdot)$: FTRL regularizer)

Solution: ***lifting*** the domain to $\boldsymbol{\mathcal{W}} \triangleq \{\boldsymbol{w} = (w,1) \mid w\in\mathcal{W}\}$ and $\|\nabla\psi(w)\|_{\nabla^{-2}\psi(w)}$ can be bounded by normal barrier (a special self-concordant barrier).

Algorithm 1 Subroutine of Luo et al. (2022)

Input: bandit value $h_t(w_t)$, curvature σ_t, last-round decision $\bar{w}_t$, step size η_{t+1}, regularization coefficient λ_t

1: Estimate gradient $\boldsymbol{g}_t \triangleq d(h_t(w_t) + \frac{\lambda_t}{2}\|w_t\|_2^2)\boldsymbol{H}_t^{1/2}\boldsymbol{s}_t$

2: Update as $\bar{w}_{t+1} = \arg\min_{w\in\boldsymbol{\mathcal{W}}} \boldsymbol{F}_{t+1}(\boldsymbol{w})$ — *gradient estimator*

3: Compute $\boldsymbol{H}_{t+1} \triangleq \nabla^2\Psi(\bar{\boldsymbol{w}}_{t+1}) + \eta_{t+1}(\sigma_{1:t} + \lambda_{0:t})\boldsymbol{I}$ — *key update step*

4: Draw $\boldsymbol{s}_{t+1}$ randomly from $\mathbb{S}^{d+1}\cap(\boldsymbol{H}_{t+1}^{-1/2}\boldsymbol{e}_{d+1})^\perp$

5: Perturb $\boldsymbol{w}_{t+1} = (w_{t+1}, 1) = \bar{\boldsymbol{w}}_{t+1} + \boldsymbol{H}_{t+1}^{-1/2}\boldsymbol{s}_{t+1}$ — *perturbation step*

$$\boldsymbol{F}_{t+1}(\boldsymbol{w}) \triangleq \sum_{s=1}^{t}\left(\boldsymbol{g}_s^\top\boldsymbol{w} + \frac{\sigma_s}{2}\|\boldsymbol{w} - \bar{\boldsymbol{w}}_s\|_2^2\right) \quad \text{(APPROX)}$$

$$+\sum_{s=1}^{t}\frac{\lambda_s}{2}\|\boldsymbol{w} - \bar{\boldsymbol{w}}_s\|_2^2 + \frac{\lambda_0}{2}\|\boldsymbol{w}\|_2^2 \quad \text{(REGLR-I)}$$

$$+\frac{1}{\eta_{t+1}}\psi(\boldsymbol{w}). \quad \text{(REGLR-II)}$$

(notions in the lifted domain use bold italic symbols)

BCO with switching costs and heterogeneous curvatures:

Memory terms can be further transformed into ***switching costs***:

$$\|w_t - w_{t+1}\|_2 \text{ and } \|w_t - w_{t+1}\|_2^2$$

We identify a ***key stability*** term in both ***regret*** and ***switching cost*** analysis.

$\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_{\boldsymbol{H}_t}$ — $\{\bar{\boldsymbol{w}}_t\}_{t=1}^T$ are online decisions; $\boldsymbol{H}_t$ is the perturbation matrix, i.e., $\boldsymbol{w}_t = \bar{\boldsymbol{w}}_t + \boldsymbol{H}_t^{1/2}\boldsymbol{s}_t$

- Important in ***regret***: closely related to $\langle g_t, \bar{w}_t - \bar{w}_{t+1}\rangle$, a crucial term in FTRL analysis.
- Important in ***memory***: closely related to *switching cost* terms of $\|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t+1}\|_2$.

Newton decrement: $\lambda(w,f) \triangleq \|\nabla f(w)\|_{\nabla^{-2}f(w)}$ [Nesterov and Nemirovskii, 1994]

Proposition 1. *For a self-concordant barrier f, if the Newton decrement $\lambda(w,f) \le 1/2$, then it follows that $\|w - w^\star\|_{\nabla^2 f(w)} \le 2\lambda(w,f)$, where $w^\star = \arg\min_w f(w)$.*

Proof sketch: As $\bar{\boldsymbol{w}}_{t+1}$ is a minimizer of $\boldsymbol{F}_{t+1}(\cdot)$, we transform $\boldsymbol{H}_t$ to $\nabla^2\boldsymbol{F}_{t+1}(\cdot)$.

Using Newton decrement, we are able to:

(i) ***greatly simplify*** the proof for BCO with heterogeneous curvatures.

(ii) handle heterogeneous curvatures in BCO ***with memory***.

*Note that our analysis for heterogeneous curvatures in BCO-M holds for **general strongly convex functions**.*

Key References

[1] Luo et al., Adaptive Bandit Convex Optimization with Heterogeneous Curvature, COLT 2022

[2] Sun et al., Optimal Rates for Bandit Nonstochastic Control, NeurIPS 2023