Estimate to Decide: Matrix Completion driven Smoothed Online Quadratic Optimization

Neelkamal Bhuyan[1] **Debankur Mukherjee**[1] **Adam Wierman**[2]

[1]Georgia Institute of Technology [2]California Institute of Technology
nbhuyan3@gatech.edu
debankur.mukherjee@isye.gatech.edu
adamw@caltech.edu

Abstract

This work tackles the problem of **blind online optimization with movement costs**, where a player must make sequential decisions to balance an unknown dynamic hitting cost $f_t(x)$ against a metric penalty $c(x_t, x_{t-1})$ for changing actions between consecutive rounds, while requiring to estimate f_t's structure. We study this problem for general quadratic costs under a restrictive, noisy bandit feedback model. In this setting, the player only observes the location of the hitting cost before taking an action and receives a single, noisy value of the cost it suffers post-action. To address this challenge, we provide the first algorithm for this setting that provably achieves a **sub-linear dynamic regret**, by combining online matrix estimation and the dynamic balancing of hitting and switching costs, within a principled exploration-exploitation framework.

1 Introduction

Incorporating dynamic cost-function estimation in online sequential decision-making is critical for enhancing the robustness and realism of models for complex behaviors. This capability is crucial across diverse fields: in robotic manipulation for learning cost functions from human demonstrations [26, 51, 11, 52], in cognitive and neural modeling for inferring the objectives behind biological motion [57, 56, 46, 63], and in industrial automation for system identification [33].

A canonical framework for modeling such sequential decisions is smoothed online quadratic optimization (SOQO). Here, over T rounds, a player selects an action $x_t \in \mathbb{R}^d$ to minimize a quadratic hitting cost $f_t(x_t)$, while also needing to account for an additional penalty of $\frac{1}{2}\|x_t - x_{t-1}\|_2^2$ for switching actions between consecutive rounds. This simple but powerful model helps model decision-making in diverse systems including large scale of data-centers and grids [29, 61, 47, 7, 45, 38, 37, 44], online video transmission [28, 17], and chip thermal management [72, 73].

Despite its wide applicability, algorithms for SOQO and the broader field of smoothed online convex optimization (SOCO) rely on a 'full information access' paradigm. Specifically, the existing frameworks predominantly assume (i) full knowledge of the current cost function $f_t(\cdot)$ [15, 23, 74, 19, 21, 53, 9, 10], or (ii) partial structure [48, 31], or (iii) access to a gradient oracle $\nabla f_t(\cdot)$[58, 59].

In contrast, this work focuses on an information-agnostic setup, where the player is aware of only the '*location parameter*' v_t of the hitting cost at each round without any information about the matrix A involved in the quadratic form. Thus, the player's knowledge about the hitting cost structure at time t must be inferred solely through (a) the location trajectory $\{v_\tau\}_{\tau=1}^{t}$, (b) previous actions taken $\{x_\tau\}_{\tau=1}^{t-1}$, and (c) the corresponding incurred *noisy* hitting costs observed, up to round t. This information model where the underlying structure of hitting costs is never revealed, with the player receiving only the (noisy) value of the penalty $f_t(x_t)$ after taking an action x_t, is known as ***rank-*1 *measurement*** model in statistical estimation literature [18, 12] and as (noisy) ***bandit feedback*** in the online algorithms community.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

2 Motivation and Related Works

The past two decades have seen significant advances in the field of online optimization [43, 38, 2, 8, 3, 5, 19, 53, 9]. Clever techniques, for handling switching costs, that tap into adjacent fields of convex body chasing [8, 23], optimal transport [19, 21] and distributed optimization [39, 30, 9] have been developed, thanks to the structural assumptions on hitting costs and/or information-access model, the absence of which leads to provably opaque negative results [4, 15].

Recent applications [17] of SOQO/SOCO, however, require foregoing of certain assumptions, specifically the apriori knowledge of hitting cost function for making an online decision. Existing algorithms rely either on (i) full information model [13, 14, 15, 24, 25, 20, 21, 53, 9, 10] or (ii) a weaker gradient oracle model but with boundedness assumptions on the action space [75, 16, 55, 54]. Few works consider limited or delayed feedback from the environment [49, 31, 58, 59] but still assume noiseless oracle access to hitting cost structure or gradient.

In the closely related field of online control and reinforcement learning, there has been a recent surge in learning the underlying cost function parameters that are typically unknown in applications like robot bio-physics [50, 22, 41, 69] and human-in-the-loop behavior learning [66, 62, 65, 32]. Significant progress has been made in recent years in the context of Linear Quadratic Regulation (LQR), where the Q and R matrices of the state and action cost functions, respectively, are learned through Inverse Optimal Control (IOC) [27, 36, 71, 34, 35]. However, even the state-of-the-art algorithms rely on the use of hindsight ***optimal action trajectory*** [27, 36, 71, 34, 35, 1, 64, 67, 70, 6]. Similar limitations exist beyond LQR in recent literature, with works focusing on *offline* data-driven approaches to learn underlying reward/cost function [68, 60, 42, 6].

SOQO and LQR are connected through their shared structure of balancing immediate costs against switching costs between consecutive actions. In fact, LQR optimal control can be framed as an instance of SOQO [25, 40, 9], linking the theory and methods of SOQO closely with LQR. With applications of both SOQO and LQR demanding simultaneous cost function estimation and online optimization, we aim to answer the following question in this work:

> *"How can an online policy perform* ***online estimation*** *of an unknown cost function from only noisy bandit feedback while guaranteeing minimal* ***dynamic regret****?"*

The challenge in addressing this question lies in a fundamental trade-off between online optimization and data-driven learning as elaborated in Section 3.1. In the next section, we formally introduce our problem set-up and highlight the technical challenges.

3 Model and Preliminaries

Consider an online game in an action space $\mathbb{R}^d$, $d \geq 1$, over a finite time horizon of T rounds. In each round the player chooses an action x_t and suffers a hitting cost of $f_t(x_t) = \frac{1}{2}(x_t - v_t)^T A(x_t - v_t)$, where A is a positive definite $d \times d$ matrix, and v_t is a *location parameter* that is revealed in an online fashion. In the following we will assume that $\{v_t\}_{t=1}^T$ forms a martingale sequence. Additionally, in each round, the player incurs a switching cost of $\frac{1}{2}\|x_t - x_{t-1}\|_2^2$ for transitioning between actions.

Prior works [25, 9] predominantly assume that the player has complete knowledge of the matrix A. We weaken this assumption in our information model in two distinct manners: In each round t (i) the player is only aware of the location v_t of quadratic hitting cost prior to choosing x_t that additionally needs to account for switching costs $\frac{1}{2}\|x_t - x_{t-1}\|_2^2$ (ii) after which it receives a noisy value of the hitting cost incurred, that is, $\frac{1}{2}(x_t - v_t)^T A(x_t - v_t) + \eta_t$, where η_t can be random or adversarial. Under this information model, the player computes an action x_t^{ALG} at each round t to solve the following optimization problem:

$$\underset{x_1,\ldots,x_T}{\operatorname{argmin}} \sum_{t=1}^{T} \frac{1}{2}(x_t - v_t)^T A(x_t - v_t) + \frac{1}{2}\|x_t - x_{t-1}\|_2^2 \tag{3.1}$$

The online sequence of actions $(x_t^{\text{ALG}})_{t=1}^T$ has to satisfy two objectives:

1. Minimize the total objective (3.1) without knowing A, which leads to the second objective,

2. High fidelity estimation of underlying unknown matrix A at each round t, using noisy past measurements $\left\{\frac{1}{2}(x_\tau^{\text{ALG}} - v_\tau)^T A(x_\tau^{\text{ALG}} - v_\tau) + \eta_\tau\right\}_{\tau=1}^{t-1}$.

Performance metric: We consider the following **(dynamic) regret** as the performance metric: For any online algorithm ALG define

$$\text{Regret}_{\text{ALG}}[1, T] := \mathbb{E}[\text{Cost}_{\text{ALG}}[1, T]] - \mathbb{E}[\text{Cost}^*[1, T]],$$

where $\mathbb{E}[\text{Cost}^*[1, T]]$ is the total cost of the online optimal algorithm that knows the A matrix beforehand, established in [10]. Online algorithms have a particularly difficult time maintaining a sub-linear dynamic regret, as seen in [76, 77, 75, 9] as the player needs to continuously track and analyze the hitting cost. Maintaining such a guarantee becomes especially tricky in our bandit feedback model. We illustrate this through an example in the following subsection.

3.1 A Fundamental Trade-off

Consider the action space $\mathbb{R}^d$ where the player starts at the origin. The time horizon is fixed as $T = 2d-1$ and underlying A matrix is diagonal with d distinct positive entries. Now, the environment supplies a sequence of minimizers $\{v_t\}_{t=1}^{2d-1}$ such that for the first d rounds $(v_t)_2 = (v_t)_3 = \ldots = (v_t)_d = 0$, and starting with round $(d+1)$, at $t = (d+i)^{th}$ round, v_t is such that $v_t - x_{t-1}$ is parallel to e_{i+1} for $i \in \{1, \ldots, d-1\}$. Such a sequence of minimizers can occur adversarially or stochastically $\big($for example, $v_{d+i} \sim \mathcal{N}(x_{t-1}, e_{i+1}e_{i+1}^T)\big)$.

Most online algorithms in the SOCO literature [23, 25, 74, 53, 9, 10, 39] have the general form:

$$x_t^{\text{ALG}} = \operatorname*{argmin}_{x \in \mathbb{R}^d} f_t(x) + c(x, x_{t-1}) + g(x, v_t, x_{t-1}),$$

which for quadratic cost functions, place x_t on the line between x_{t-1} and v_t. This means that any robust online algorithm ALG dictates that the player take a sequence of actions $(x_1^{\text{ALG}}, \ldots, x_d^{\text{ALG}})$ parallel to e_1 for the first d rounds. During these rounds, the bandit feedback model collects information:

$$\left\{(1/2) \cdot (x_k^{\text{ALG}} - v_k)^T A(x_k^{\text{ALG}} - v_k)\right\}_{k=1}^d = \{c_k A_{1,1}\}_{k=1}^d .$$

At round $(d+1)$, x_{d+1}^{ALG} is supposed to be on the line between x_d and v_{d+1}, which is parallel to e_2. Consequently, x_{d+1} has **direct dependence** on $A_{2,2}$, the second diagonal entry of the unknown matrix A. The player now resorts to the above rank-1 data collected so far, as that is the only information it has on matrix A.

However, as it turns out, the first d rounds only generated information about $A_{1,1,}$, forcing the player to make a ***blind guess*** regarding x_{d+1}. In fact it gets worse, as this will happen repeatedly, with round $d+i$ requiring the value of $A_{i+1,i+1}$ for computing a robust action but the player having knowledge of only $\{A_{1,1,}, \ldots, A_{i,i}\}$. Alternatively, the player could have spent rounds $\{2, 3 \ldots, d\}$ probing the hitting cost along the rest of the directions, collecting information on the entire matrix A. This *data collection* operation, however, leads to high hitting and switching costs, due to significant deviation from the minimizer trajectory.

> "*Collecting high-fidelity rank-*1 *data for matrix estimation incurs high online costs but potentially pays off in the long-run. Solely following a robust online algorithm might have initial benefit but leaves the player vulnerable should it require knowledge of the matrix A.*"

In particular, the bandit feedback model allows only *Following the Minimizer* (FTM) as a possible candidate for a robust online algorithm, where in each round t, $x_t^{\text{FTM}} = v_t$. Although it incurs low cost initially, it has high regret accumulation over the horizon as established in [9]:

Remark 3.1. *For a martingale minimizer sequence $\{v_t\}_t$,* FTM *incurs an $\Omega(T)$ regret.*

This illustrates that SOQO and data-driven learning are orthogonal tasks and are extremely difficult to combine. To tackle this issue, we will be approaching this problem from an exploration-exploitation trade-off, en-route to a dual-purpose algorithm.

4 Randomized Exploration, Estimation, and Exploitation

We introduce Algorithm 1 as the first blind online algorithm that balances (i) live data collection, (ii) matrix estimation and, (iii) trajectory optimization, simultaneously to ensure the following *sub-linear* regret guarantee using only noisy rank-1 oracle:

Theorem 4.1. *Consider the quadratic hitting costs $f_t(x) = \frac{1}{2}(x - v_t)^T A(x - v_t)$, where the minimizer sequence $\{v_t\}_{t=1}^T$ forms a martingale sequence that is revealed in an online manner. Under noisy bandit feedback $\{f_t(x_t) + \eta_t\}_t$, Algorithm 1 has the following regret guarantee:*

$$\textit{Regret}[1, T] = \Theta(\sqrt{T - c_1 d^2})$$

with high probability $(1 - \exp(-C_0 m))$, where C_0, c_1 are universal constants from matrix estimation theory [18, 12]. The constants depend on noise upper bar $\sqrt{\bar{\eta}}$, smallest singular value of A, that is $\underline{\sigma_d^A}$, martingale process variance $\mathbb{E}[(v_t - v_{t-1})(v_t - v_{t-1})^T] = \Sigma$ and dimension d.

Algorithm 1 SCaLE

Input: noise cap $\bar{\eta}$, rank r, floor $\underline{\sigma_r^A}$, horizon T

Initialize: $m = c_1 r d$, $\hat{C}_{T+1} = I_{d\times d}$, $\gamma^2 = \sqrt{\bar{\eta}} \max\left\{T^{1/2}, \frac{1}{\underline{\sigma_r^A}}\right\}$

for $t = 1, 2, \ldots, m$ **do**
  $z_t \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$
  $x_t \leftarrow v_t + \gamma z_t$
  $y_t \leftarrow f_t(x_t) + \eta_t$
end for
$\hat{A}^{\texttt{SCaLE}} \leftarrow \underset{\substack{M \succeq 0 \\ \|Y - \mathcal{A}(M)\|_1 \leq \bar{\eta} m}}{\operatorname{argmin}} Tr(M)$
Regenerate candidate sequence $\{x_t^{\text{LAI}(\hat{A})}\}_{t=1}^m$
$x_{m+1} \leftarrow \hat{C}_{m+1} x_m^{\text{LAI}(\hat{A}^{\texttt{SCaLE}})} + (I - \hat{C}_m) v_m$
for $t = m + 2, \ldots, T$ **do**
  $x_t \leftarrow \hat{C}_t x_{t-1} + (I - \hat{C}_t) v_t$
end for

The condition $T > c_1 d^2$ above is a direct consequence of the following fundamental limit in matrix recovery [12]:

Remark 4.2. *At least $m = \Theta(d^2)$ rank-1 (noisy) measurements are required to ensure statistical consistency of the trace-minimizing estimator. In its absence, any estimator $\hat{A}$ exhibits,*

$$\inf_{\hat{A}} \sup_{\substack{A \in \mathbb{R}^{d\times d} \\ rank(A) = r}} \mathbb{E}\|\hat{A} - A\|_F^2 = \infty.$$

Notice that in such black-box online optimization scenarios, one typically resorts to Follow the Minimizer approach due to lack of an information-rich oracle, suffering from $\Omega(T)$ regret. However, a careful combination of data collection, estimation and then timely switch to online optimization, results in a sub-linear dynamic regret.

The key to achieving such guarantees is to strike a delicate balance between the two opposing processes in play: (i) useful rank-1 data collection under noise and, (ii) minimizing online hitting and switching costs. Algorithm 1 achieves it by prioritizing rank-1 data collection for the first m rounds. Although it suffers from a linearly increasing regret during those rounds, it gets compensated by near-optimal estimate $\hat{A}^{(m)}$ achieved, which it plugs in into the structure of the online-optimal LAI algorithm, to closely follow the benchmark henceforth.

Proving the sub-linear regret in Theorem 4.1 entailed establishing a tight relationship between the dynamic gap $\|\hat{A}^{(t)} - A\|$ and the dynamic regret against LAI. To that end, we quantify erroneous knowledge of the A matrix in the stochastic SOQO problem by establishing a regret guarantee for the general class of *adaptive interpolation* algorithms introduced in [9]:

Theorem 4.3. *The sequence of actions with $\{\hat{C}_t : \hat{C}_t^{-1} = 2I + \hat{A}^{(t)} - \hat{C}_{t+1} \; \forall\, 1 \leq t \leq T\}$*

$$x_t = \hat{C}_t x_{t-1} + (I - \hat{C}_t) v_t$$

has the following regret guarantee

$$\textit{Regret}[1, T] \leq \frac{\sigma^2(\lambda_{\min}^A + 1)^2}{2(\lambda_{\min}^A + 2)\lambda_{\min}^A} \sum_{s=1}^{T-1} \|M_s\|_{op}$$

where $\|M_s\|_{op} \propto \|\hat{A}^{(s)} - A\|_{op}$.

References

[1] P. Ahmadi, M. Rahmani, and A. Shahmansoorian. Model-free inverse optimal control for completely unknown nonlinear systems by adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2025.

[2] L. Andrew, S. Barman, K. Ligett, M. Lin, A. Meyerson, A. Roytman, and A. Wierman. A tale of two metrics: Simultaneous bounds on competitiveness and regret. volume 30, pages 741–763. PMLR, 7 2013.

[3] A. Antoniadis, N. Barcelo, M. Nugent, K. Pruhs, K. Schewior, and M. Scquizzato. Chasing Convex Bodies and Functions. In E. Kranakis, G. Navarro, and E. Chávez, editors, *LATIN 2016: Theoretical Informatics*, pages 68–81, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg.

[4] A. Antoniadis and K. Schewior. A tight lower bound for online convex optimization with switching costs. In *Proc. WAOA '21: International Workshop on Approximation and Online Algorithms*, pages 164–175. Springer, 2017.

[5] K. A. Antonios and Schewior. A tight lower bound for online convex optimization with switching costs. pages 164–175. Springer International Publishing, 2018.

[6] H. J. Asl and E. Uchibe. Data-driven inverse optimal control for continuous-time nonlinear systems. *arXiv preprint arXiv:2503.09090*, 2025.

[7] M. Badiei, N. Li, and A. Wierman. Online convex optimization with ramp constraints. pages 6730–6736, 2015.

[8] N. Bansal, A. Gupta, R. Krishnaswamy, K. Pruhs, K. Schewior, and C. Stein. A 2-competitive algorithm for online convex optimization with switching costs. volume 40, pages 96–109. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2015. Keywords: Stochastic, Scheduling.

[9] N. Bhuyan, D. Mukherjee, and A. Wierman. Best of both worlds guarantees for smoothed online quadratic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 3850–3888. PMLR, 21–27 Jul 2024.

[10] N. Bhuyan, D. Mukherjee, and A. Wierman. Optimal decentralized smoothed online convex optimization. *arXiv preprint arXiv:2411.08355*, 2024.

[11] A. Byravan and D. Fox. Graph-based inverse optimal control for robot manipulation. In *Robotics: Science and Systems (RSS)*, 2016.

[12] T. T. Cai and A. Zhang. ROP: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102 – 138, 2015.

[13] N. Chen, A. Agarwal, A. Wierman, S. Barman, and L. L. H. Andrew. Online convex optimization using predictions. In *Proc. SIGMETRICS '15*, pages 191–204, 2015.

[14] N. Chen, J. Comden, Z. Liu, A. Gandhi, and A. Wierman. Using predictions in online optimization: Looking forward with an eye on the past. *ACM SIGMETRICS Perf.Eval. Rev.*, 44(1):193–206, 2016.

[15] N. Chen, G. Goel, and A. Wierman. Smoothed online convex optimization in high dimensions via online balanced descent. volume 75, pages 1574–1594. PMLR, 7 2018.

[16] S. Chen, W.-W. Tu, P. Zhao, and L. Zhang. Optimistic online mirror descent for bridging stochastic and adversarial online convex optimization. *arXiv preprint arXiv:2302.04552*, 2023.

[17] T. Chen, Y. Lin, N. Christianson, Z. Akhtar, S. Dharmaji, M. Hajiesmaili, A. Wierman, and R. K. Sitaraman. Soda: An adaptive bitrate controller for consistent high-quality video streaming. In *ACM SIGCOMM 2024*, 2024.

[18] Y. Chen, Y. Chi, and A. J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

[19] N. Christianson, T. Handina, and A. Wierman. Chasing convex bodies and functions with black-box advice. volume 178, pages 867–908. PMLR, 7 2022.

[20] N. Christianson, T. Handina, and A. Wierman. Chasing convexbodies and functions with black-box advice. In *Proc. COLT '22*, 2022.

[21] N. Christianson, J. Shen, and A. Wierman. Optimal robustness-consistency tradeoffs for learning-augmented metrical task systems. volume 206, pages 9377–9399. PMLR, 7 2023.

[22] H. El-Hussieny and J.-H. Ryu. Inverse discounted-based lqr algorithm for learning human movement behaviors. *Applied Intelligence*, 49(4):1489–1501, 2019.

[23] G. Goel, Y. Lin, H. Sun, and A. Wierman. Beyond online balanced descent: An optimal algorithm for smoothed online optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[24] G. Goel, Y. Lin, H. Sun, and A. Wierman. Beyond online balanced descent: An optimal algorithm for smoothed online optimization. *Adv. Neural Inf. Process. Syst*, 32:1875–1885, 2019.

[25] G. Goel and A. Wierman. An online algorithm for smoothed regression and lqr control. volume 89, pages 2504–2513. PMLR, 7 2019.

[26] S. M. S. M. S. Hassan et al. Robotic arm manipulation with inverse reinforcement learning and td-mpc. *arXiv preprint arXiv:2407.12941*, 2023.

[27] W. Jin, D. Kulić, J. F.-S. Lin, S. Mou, and S. Hirche. Inverse optimal control for multiphase cost functions. *IEEE Transactions on Robotics*, 35(6):1387–1398, 2019.

[28] V. Joseph and G. de Veciana. Jointly optimizing multi-user rate adaptation for video transport over wireless systems: Mean-fairness-variability tradeoffs. pages 567–575, 2012.

[29] S.-J. Kim and G. B. Giannakis. An online convex optimization approach to real-time energy pricing for demand response. *IEEE Transactions on Smart Grid*, 8:2784–2793, 2017.

[30] P. Li, J. Yang, A. Wierman, and S. Ren. Learning-augmented decentralized online convex optimization in networks. *arXiv preprint arXiv:2306.10158*, 2023.

[31] P. Li, J. Yang, A. Wierman, and S. Ren. Robust learning for smoothed online convex optimization with feedback delay. *Advances in Neural Information Processing Systems*, 36:16889–16916, 2023.

[32] W.-H. Li and H.-N. Wu. Human behavior learning for a class of noisy discrete-time nonlinear hitl systems via moving horizon estimation and state-dependent riccati equation. *Available at SSRN 5245076*, 2025.

[33] Y. Li and C. Yu. Resilient inverse optimal control for tracking: Overcoming process noise challenges. *Journal of the Franklin Institute*, 361(4):2288–2310, 2024.

[34] Y. Li, C. Yu, H. Fang, and J. Chen. Inverse optimal control for linear quadratic tracking with unknown target states. *arXiv preprint arXiv:2402.17247*, 2024.

[35] B. Lian, W. Xue, F. L. Lewis, and A. Davoudi. Inverse value iteration and q-learning: Algorithms, stability, and robustness. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):6970–6980, 2024.

[36] B. Lian, W. Xue, Y. Xie, F. L. Lewis, and A. Davoudi. Off-policy inverse q-learning for discrete-time antagonistic unknown systems. *Automatica*, 155:111171, 2023.

[37] M. Lin, Z. Liu, A. Wierman, and L. L. H. Andrew. Online algorithms for geographical load balancing. pages 1–10, 2012.

[38] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska. Dynamic right-sizing for power-proportional data centers. pages 1098–1106, 2011.

[39] Y. Lin, J. Gan, G. Qu, Y. Kanoria, and A. Wierman. Decentralized online convex optimization in networked systems. In *International Conference on Machine Learning*, pages 13356–13393. PMLR, 2022.

[40] Y. Lin, Y. Hu, G. Shi, H. Sun, G. Qu, and A. Wierman. Perturbation-based regret analysis of predictive control in linear time varying systems. *Advances in Neural Information Processing Systems*, 34:5174–5185, 2021.

[41] Y.-H. Lin, S.-W. Chiu, Y.-C. Lin, C.-C. Lin, and L.-K. Pan. Inverse problem algorithm application to semi-quantitative analysis of 272 patients with ischemic stroke symptoms: Carotid stenosis risk assessment for five risk factors. *Journal of Mechanics in Medicine and Biology*, 20(09):2040021, 2020.

[42] S. Liu and M. Zhu. In-trajectory inverse reinforcement learning: Learn incrementally before an ongoing trajectory terminates. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[43] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew. Geographical load balancing with renewables. *ACM SIGMETRICS Perform. Eval. Rev.*, 39(3):62–66, 2011.

[44] T. Lu, M. Chen, and L. L. H. Andrew. Simple and effective dynamic provisioning for power-proportional data centers. *IEEE Transactions on Parallel and Distributed Systems*, 24:1161–1171, 2013.

[45] S. Mei, Y. Wang, and Z. Sun. Robust economic dispatch considering renewable generation. pages 1–5, 2011.

[46] M. Mistry et al. An inverse optimal control approach to explain human arm movements. *Scientific Reports*, 8:4736, 2018.

[47] B. Narayanaswamy, V. K. Garg, and T. S. Jayram. Online optimization for the smart (micro) grid. pages 1–10, 2012.

[48] W. Pan, G. Shi, Y. Lin, and A. Wierman. Online optimization with feedback delay and nonlinear switching cost. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(1):1–34, 2022.

[49] W. Pan, G. Shi, Y. Lin, and A. Wierman. Online optimization with feedback delay and nonlinear switching cost. *Proc. ACM Meas. Anal. Comput. Syst.*, 6(1):1–34, 2022.

[50] M. C. Priess, R. Conway, J. Choi, J. M. Popovich, and C. Radcliffe. Solutions to the inverse lqr problem with application to biological systems analysis. *IEEE Transactions on control systems technology*, 23(2):770–777, 2014.

[51] J. R. Rebula, S. Schaal, J. Finley, and L. Righetti. A robustness analysis of inverse optimal control of bipedal walking. *arXiv preprint arXiv:2104.12042*, 2021.

[52] J. Rosell et al. Reasoning for robot manipulation. *Institute of Industrial and Control Engineering (IOC), Polytechnic University of Catalonia*, 2021.

[53] D. Rutten, N. Christianson, D. Mukherjee, and A. Wierman. Smoothed online optimization with unreliable predictions. *Proc. ACM Meas. Anal. Comput. Syst.*, 7, 3 2023.

[54] S. Sachs, H. Hadiji, T. van Erven, and C. Guzman. Accelerated rates between stochastic and adversarial online convex optimization. *arXiv preprint arXiv:2303.03272*, 2023.

[55] S. Sachs, H. Hadiji, T. van Erven, and C. Guzmán. Between stochastic and adversarial online convex optimization: Improved regret bounds via smoothness. volume 35, pages 691–702. Curran Associates, Inc., 2022.

[56] M. Schmittwilken, M. Schultheis, and C. A. Rothkopf. Putting perception into action with inverse optimal control for continuous psychophysics. *eLife*, 11:e76635, 2022.

[57] M. Schultheis et al. Inverse optimal control adapted to the noise characteristics of the human sensorimotor system. *PLOS Computational Biology*, 17(3):e1008893, 2021.

[58] S. Senapati and R. Vaze. Online convex optimization with switching cost and delayed gradients. *Performance Evaluation*, 162:102371, 2023.

[59] H. Shah, P. Chandrasekhar, and R. Vaze. Online convex optimization with switching cost with only one single gradient evaluation. *arXiv preprint arXiv:2507.04133*, 2025.

[60] Z. Sun and G. Jia. Inverse reinforcement learning by expert imitation for the stochastic linear–quadratic optimal control problem. *Neurocomputing*, 633:129758, 2025.

[61] H. Wang, J. Huang, X. Lin, and H. Mohsenian-Rad. Exploring smart grid and data center interactions for electric power load balancing. *SIGMETRICS Perform. Eval. Rev.*, 41:89–94, 1 2014.

[62] M. Wang and H.-N. Wu. Adaptive inverse optimal control for linear human-in-the-loop systems with completely unknown dynamics. *IEEE Transactions on Automation Science and Engineering*, 2024.

[63] K. Westermann, J. F.-S. Lin, and D. Kulić. Inverse optimal control with time-varying objectives: application to human jumping movement analysis. *Scientific reports*, 10(1):11174, 2020.

[64] H. Wu, Q. Hu, J. Zheng, F. Dong, Z. Ouyang, and D. Li. Discounted inverse reinforcement learning for linear quadratic control. *IEEE Transactions on Cybernetics*, 2025.

[65] H.-N. Wu, W.-H. Li, and M. Wang. A finite-horizon inverse linear quadratic optimal control method for human-in-the-loop behavior learning. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(6):3461–3470, 2024.

[66] H.-N. Wu and M. Wang. Human-in-the-loop behavior modeling via an integral concurrent adaptive inverse reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):11359–11370, 2023.

[67] J. Wu, W. Xue, F. L. Lewis, and B. Lian. Data-driven optimization-based cost and optimal control inference. *IEEE Control Systems Letters*, 2025.

[68] W. Xue, B. Lian, J. Fan, T. Chai, and F. L. Lewis. Inverse reinforcement learning for trajectory imitation using static output feedback control. *IEEE Transactions on Cybernetics*, 54(3):1695–1707, 2024.

[69] H. Yu, A. Ramadan, J. Cholewicki, J. M. Popovich Jr, N. P. Reeves, J. S. H. You, and J. Choi. Inferring human control intent using inverse linear quadratic regulator with output penalty versus gain penalty: Better fit but similar intent. *Journal of Dynamic Systems, Measurement, and Control*, 146(6):061103, 2024.

[70] M. Yu, L. Feng, L. Jiang, and Y.-H. Ni. A convex optimization approach to model-free inverse optimal control with provable convergence. *arXiv preprint arXiv:2507.19965*, 2025.

[71] M. Yu and Y. Ni. Inverse reinforcement learning for discrete-time linear quadratic systems. In *2024 14th Asian Control Conference (ASCC)*, pages 1027–1032. IEEE, 2024.

[72] F. Zanini, D. Atienza, L. Benini, and G. D. Micheli. Multicore thermal management with model predictive control. pages 711–714, 2009.

[73] F. Zanini, D. Atienza, G. D. Micheli, and S. P. Boyd. Online convex optimization-based algorithm for thermal management of mpsocs. pages 203–208. Association for Computing Machinery, 2010.

[74] L. Zhang, W. Jiang, S. Lu, and T. Yang. Revisiting smoothed online learning. *arXiv:2102.06933*, 2021.

[75] L. Zhang, W. Jiang, J. Yi, and T. Yang. Smoothed online convex optimization based on discounted-normal-predictor. volume 35, pages 4928–4942. Curran Associates, Inc., 2022.

[76] L. Zhang, T. Yang, J. Yi, R. Jin, and Z.-H. Zhou. Improved dynamic regret for non-degenerate functions. volume 30. Curran Associates, Inc., 2017.

[77] Y. Zhang, R. J. Ravier, M. M. Zavlanos, and V. Tarokh. A distributed online convex optimization algorithm with improved dynamic regret. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2449–2454, 2019.