

# Model-free Learning for Risk-constrained Linear Quadratic Regulator with Structured Feedback in Networked Systems

Kyung-bin Kwon, Lintao Ye, Vijay Gupta, and Hao Zhu

**Abstract**—We develop a model-free learning algorithm for the infinite-horizon linear quadratic regulator (LQR) problem. Specifically, (risk) constraints and structured feedback are considered, in order to reduce the state deviation while allowing for a sparse communication graph in practice. By reformulating the dual problem as a nonconvex-concave minimax problem, we adopt the gradient descent max-oracle (GDmax), and for model-free setting, the stochastic (S)GDmax using zero-order policy gradient. By bounding the Lipschitz and smoothness constants of the LQR cost using specifically defined sublevel sets, we can design the stepsize and related parameters to establish convergence to a stationary point (at with high probability). Numerical tests in a networked microgrid control problem have validated the convergence of our proposed SGDmax algorithm while demonstrating the effectiveness of risk constraints. The SGDmax algorithm has attained a satisfactory optimality gap compared to the classical LQR control, especially for the full feedback case.

## I. INTRODUCTION

The linear quadratic regulator (LQR) problem is one of the most fundamental problems in optimal control theory [1], [2]. Recently, there is significant interest in model-free learning of the standard LQR problem using gradient-based approaches [3]–[5], with connection to the popular reinforcement learning (RL) methods. Nonetheless, model-free learning and convergence analysis for general LQR problems are still lacking such as constrained LQR and structured feedback design.

Constrained formulations have attracted recent interest for the LQ [6]–[8] and general RL problems [9], [10]. Constraints can improve the safety of resultant policies while potentially increasing the learning rates as a regularization. In particular, recent work [6], [8] has considered the mean-variance risk for the LQR problems, that can effectively mitigate the random state deviation from its mean due to noisy disturbance. More interestingly, [6] has shown that this risk measure is equivalent to a quadratic constraint function that is similar to the LQR cost. In addition, [8] has developed a dual-ascent based double-loop algorithm by utilizing the global convergence of LQR learning [3], [4] for the inner-loop. Nonetheless, this double-loop procedure

may be complicated to implement in practice because the inner-loop convergence is in a probabilistic way due to the stochastic gradient.

Meanwhile, decentralized control problems [11]–[13] arise in various real-world applications where sensors and actuators are distributed in a networked system. For example, it is very useful for power system control designs such as wide-area damping control [14], [15] or networked microgrid control [16], [17]. In decentralized LQR problems, a sparse communication graph leads to structured feedback gain, which has also been considered in recent gradient-based learning approaches [3], [18]. In general, the stabilizable region of structured LQR is disconnected with a complex geometry [19], and thus it is difficult to analyze. While gradient-based learning for structured LQR does not lead to global convergence as in the unstructured case [3], [4], it is easy for implementation as the gradient can be simply performed over the non-zero entries [3], [18].

Our goal is to develop model-free learning algorithms for risk-constrained LQR problem under sparse feedback structure that arises in networked systems. The structured feedback is incorporated by considering the sparse non-zero entries only, and thus the gradient computation and updates can be performed without accounting for such structured constraint. Nonetheless, it leads to convergence to only a stationary point. As for the constraint function, it is similar to the LQR cost with the mean-variance risk as a special case as shown by [6], [7]. To deal with this constraint, we consider the dual problem which shares the stationary point (SP) with the minimax problem for the Lagrangian function. The resultant nonconvex-concave minimax reformulation motivates us to adopt Gradient-Descent max-oracle (GDmax) and the stochastic (S)GDmax algorithms in [20] to solve the outer minimization problem via GD updates. More specifically, the SGDmax relies on the zero-order policy gradient (ZOPG) [21] which has bounded noise variance.

Nonetheless, the key challenge in establishing the convergence lies in the LQR cost function, which is shown to exhibit local-only Lipschitz and smoothness with location-dependent constants [3], [4]. To tackle this, we introduce a compact sublevel set within which the upper bounds of Lipschitz and smoothness constants hold. Such analysis enables us to carefully design the stepsize and related parameters to establish the convergence to SP, while the convergence of SGDmax in a model-free setting can be attained with a high probability. Numerical results have validated the convergence of our algorithms and demonstrated the impact of having risk constraint and structured feedback in learning LQR policy.

This work has been partially supported by NSF Grants 1802319, 1952193, and 2130706.

K. Kwon and H. Zhu are with the Department of Electrical & Computer Engineering, The University of Texas at Austin, 2501 Speedway, Austin, TX, 78712, USA; Emails: {kwon8908kr, haozhu}@utexas.edu.

L. Ye is with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan City, Hubei Province, 430074, China; Email: yelintao93@hust.edu.cn

V. Gupta is with the Department of Electrical and Computer Engineering, Purdue University, 465 Northwestern Ave, West Lafayette, IN, 47907, USA; Email: gupta869@purdue.edu

The SGDmax algorithm have attained satisfactory optimality gap compared to the classical LQR control, especially for the full feedback case.

The remainder of this paper is organized as follows. Sec. II formulates the infinite-horizon risk-constrained LQR with structured feedback. Sec. III introduces the dual-related minimax reformulation and analyzes the convergence of the Gradient Descent with max-oracle (GDmax) algorithm. Sec. IV extends it to model-free learning by developing the Stochastic (S)GDmax via zero-order policy gradient. Sec. V presents the numerical results in a networked load frequency control (LFC) problem, while the paper is wrapped up in Sec. VI.

**Notations:** Let  $\|\cdot\|$  denote the  $L_2$ -norm,  $\nabla_{\mathcal{K}}\mathcal{L}$  the gradient of  $\mathcal{L}$  that admits the structure defined in  $\mathcal{K}$ ,  $\{X^j\}$  a sequence of  $\{X^0, X^1, \dots\}$ ,  $\mathcal{P}_{\mathcal{Y}}(\cdot)$  the projection onto the set  $\mathcal{Y}$ , and the operator  $\otimes$  the Kronecker product of matrices. Last,  $\mathbb{E}(\cdot)$  denotes the expectation while  $\mathbb{P}(\cdot)$  the probability of an event.

## II. PROBLEM FORMULATION

We consider the infinite-horizon LQR problem for a linear time-invariant system given by

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad t = 0, 1, \dots \quad (1)$$

with the state  $x_t \in \mathbb{R}^n$ , action  $u_t \in \mathbb{R}^m$ , and random noise  $w_t \in \mathbb{R}^n$  that is uncorrelated across time. In addition, the model parameters  $A \in \mathbb{R}^{n \times n}$  and  $B \in \mathbb{R}^{n \times m}$  can be unknown. The constrained LQR problem with structured feedback aims to find an optimal linear feedback gain  $K \in \mathbb{R}^{m \times n}$  for the control policy  $u_t = -Kx_t$  to:

$$\begin{aligned} \min_{K \in \mathcal{K}} R_0(K) &= \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \sum_{t=0}^{T-1} [x_t^\top Q x_t + u_t^\top R u_t] \\ \text{s.t. } R_i(K) &= \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \sum_{t=0}^{T-1} [x_t^\top Q_i x_t + u_t^\top R_i u_t] \leq c_i, \forall i \end{aligned} \quad (2)$$

where matrices  $\{Q, R\}$  and  $\{Q_i, R_i\}_{i \in \mathcal{I}}$  are all positive semi-definite, with  $\mathcal{I}$  representing the set of the constraints. The feasible set  $\mathcal{K}$  enforces a structured policy as

$$\mathcal{K} = \{K : K_{a,b} = 0 \text{ if and only if } (a, b) \notin \mathcal{E}\}. \quad (3)$$

Here, the structure pattern  $\mathcal{E}$  is specified by the edges of a given communication or information-exchange graph. Hence, the action for agent  $a$ , denoted as  $u_{a,t}$ , is determined as  $u_{a,t} = -K_a x_{a,t}$ , where  $K_a$  is a row vector with only non-zero elements in  $a$ -th row of  $K$  and  $x_{a,t}$  is a sub-vector of  $x_t$  according to  $\mathcal{E}$ . The structured  $\mathcal{K}$  is motivated by a multi-agent setting for networked control, where individual agents can access partial feedback only depending on communication links. Notably, this structured constraint will lead to a complicated geometry of the feasible region [3], [19]. While the structured  $\mathcal{K}$  makes the analysis more difficult than the full feedback case, it does not increase the complexity of computing the gradient as denoted by  $\nabla_{\mathcal{K}}$

later on. This is because one can represent the cost as a function of only non-zero entries in  $K$  which can eliminate this structured constraint [3]. Accordingly, the  $\nabla_{\mathcal{K}}$  operation needs no projection onto  $\mathcal{K}$ , and can be thought of as the gradient for an unstructured  $K$ . Therefore, gradient-based methods are ideal for learning a structured policy.

As for the quadratic constraint in (2), one can consider the mean-variance risk as a special instance, represented by

$$R_c(K) = \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \sum_{t=0}^{T-1} (x_t^\top Q x_t - \mathbb{E}[x_t^\top Q x_t | h_t])^2 \leq \delta$$

with the system trajectory  $h_t := \{x_0, u_0, \dots, x_{t-1}, u_{t-1}\}$  and a risk tolerance  $\delta$ . This risk measure limits the deviation from the expected cost given the past trajectory, and thus can mitigate extreme scenarios due to the uncertainty in the noisy dynamics. Interestingly, under a finite fourth-order moment of noise  $w_t$ , [6], [7] has developed a tractable reformulation  $R_c(K)$ , as

$$R_c(K) = \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \sum_{t=0}^{T-1} (4x_t^\top Q W Q x_t + 4x_t^\top Q M_3) \leq \bar{\delta} \quad (4)$$

with  $\bar{\delta} = \delta - m_4 + 4\text{tr}\{(WQ)^2\}$  and the (weighted) noise statistics given as

$$\bar{w} = \mathbb{E}[w_t], \quad (5)$$

$$W = \mathbb{E}[(w_t - \bar{w})(w_t - \bar{w})^\top], \quad (6)$$

$$M_3 = \mathbb{E}[(w_t - \bar{w})(w_t - \bar{w})^\top Q (w_t - \bar{w})], \quad (7)$$

$$m_4 = \mathbb{E}[(w_t - \bar{w})^\top Q (w_t - \bar{w}) - \text{tr}(WQ)]^2. \quad (8)$$

With known noise statistics, this risk constraint shares the quadratic form in (2) with an additional linear term, which does not affect our proposed gradient-based learning. Note that we will consider the general quadratic constraints  $R_i(K)$  in Section III, which can be applied to the risk-constraint in (4) with the exact same manner. The ensuing section first develops the deterministic algorithm for problem (2), which can provide insights on the model-free extension later on.

## III. A PRIMAL GRADIENT DESCENT (GD) APPROACH

To deal with constraints in (2), consider its Lagrangian function by introducing the multiplier vector  $\lambda = \{\lambda_i \geq 0\}$ ,

$$\begin{aligned} \mathcal{L}(K, \lambda) &= R_0(K) + \sum_{i \in \mathcal{I}} \lambda_i [R_i(K) - c_i] \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \sum_{t=0}^{T-1} [x_t^\top Q_\lambda x_t + u_t^\top R_\lambda u_t] - c_\lambda \end{aligned} \quad (9)$$

where we define  $Q_\lambda := Q + \sum_{i \in \mathcal{I}} \lambda_i Q_i$ , and likewise for  $R_\lambda$  and  $c_\lambda$ . Clearly,  $\mathcal{L}(K, \lambda)$  shares the same structure as an unconstrained LQR cost which is suitable for first-order algorithms. For simplicity, consider that the problem (2) is feasible and thus  $\lambda$  is finite [22, Sec. 5.2]. We consider the bounded set  $\mathcal{Y} := [0, \Lambda]^{|\mathcal{I}|}$  for  $\lambda$  with a large enough  $\Lambda \in \mathbb{R}$ , which can be set based on a feasible  $K^0$ . Using the dual function  $\mathcal{D}(\lambda) := \min_{K \in \mathcal{K}} \mathcal{L}(K, \lambda)$ , the dual problem becomes

$$\max_{\lambda \in \mathcal{Y}} \mathcal{D}(\lambda) = \max_{\lambda \in \mathcal{Y}} \min_{K \in \mathcal{K}} \mathcal{L}(K, \lambda). \quad (10)$$

As  $\mathcal{L}(K, \lambda)$  is related to LQR cost, the inner minimization problem is not convex. Recent works [3], [4], [18] have extensively analyzed the LQR cost which can be used to establish the local Lipschitz and smoothness properties of  $\mathcal{L}(K, \lambda)$ . Specifically, it is possible to find related constants that hold within a subset  $\mathcal{G}^0 \subset \mathcal{K}$ . This compact sublevel set will be defined later on, but is first introduced here for bounding the constants as stated below.

**Lemma 1** (Lipschitz and smoothness). *For any  $\lambda$  and  $K \in \mathcal{G}^0$ , the function  $\mathcal{L}(K, \lambda)$  is locally  $L_0$ -Lipschitz within a radius  $\psi_K$ ; i.e., for  $\forall K' \in \mathcal{G}^0$  such that  $\|K - K'\| \leq \psi_K$ , we have  $\|\mathcal{L}(K, \lambda) - \mathcal{L}(K', \lambda)\| \leq L_0\|K - K'\|$ . In addition, it is also locally  $\ell_0$ -smooth within a radius  $\beta_K$ , such that for  $\forall K' \in \mathcal{G}^0$  that satisfies  $\|K - K'\| \leq \beta_K$ , we have  $\|\nabla \mathcal{L}_K(K, \lambda) - \nabla \mathcal{L}_K(K', \lambda)\| \leq \ell_0\|K - K'\|$ .*

Strictly speaking, the recent LQR analysis [4], [18] asserts that Lipschitz and smoothness are only local properties, and thus the corresponding constants  $L_K$  and  $\ell_K$  depend on  $K$ . Nonetheless, using a compact set  $\mathcal{G}^0$ , we can obtain the bounds that can hold for any  $K \in \mathcal{G}^0$ , as given by

$$L_0 := \sup_{K \in \mathcal{G}^0} L_K, \text{ and } \ell_0 := \sup_{K \in \mathcal{G}^0} \ell_K. \quad (11)$$

We can also determine a general neighborhood radius as

$$\rho_0 := \inf_{K \in \mathcal{G}^0} \min\{\beta_K, \psi_K\} \quad (12)$$

that holds for any  $K \in \mathcal{G}^0$  as well.

Interestingly, the KKT conditions for problem (10) is related to the stationary point (SP) of a reformulated minimax problem. Recent results have shown that nonconvex-concave minimax problems can be solved using the so-termed Gradient Descent with max-oracle (GDmax) algorithm [23]. To this end, consider the problem

$$\min_{K \in \mathcal{K}} \Phi(K) \text{ where } \Phi(K) := \max_{\lambda \in \mathcal{Y}} \mathcal{L}(K, \lambda), \quad (13)$$

which is essentially the minimax counterpart of problem (10). As the Lagrangian function is linear in  $\lambda$ , it is possible to directly find the best  $\lambda$  in (13). Specifically, its  $i$ -th element, namely  $\lambda_i$ , depends on the feasibility of constraint  $i$  under given  $K$ ; i.e.,  $\lambda_i$  equals to 0 if constraint  $i$  is satisfied and  $\Lambda$  otherwise. Unfortunately, the function  $\Phi(K)$  is not differentiable everywhere. To tackle this issue, we consider its Moreau envelope  $\Phi_\mu(\cdot)$  for a given  $\mu > 0$ , defined as

$$\Phi_\mu(K) := \min_{K' \in \mathcal{K}} \Phi(K') + \frac{1}{2\mu} \|K' - K\|^2, \quad \forall K \in \mathcal{K}. \quad (14)$$

It can be used for defining the SP of the non-differentiable  $\Phi(K)$ , following from [20, Lemma 3.6].

**Lemma 2.** *For function  $\Phi(K)$  that is  $\ell_0$ -weakly convex and  $L_0$ -Lipschitz within the compact set  $\mathcal{G}^0$ , its Moreau envelope  $\Phi_{\mu_0}(K)$  is convex by setting  $\mu_0 := 1/(2\ell_0)$ . In addition, the  $\epsilon$ -SP of  $\Phi(K)$ , namely  $K_\epsilon$ , satisfies  $\|\nabla \Phi_{\mu_0}(K_\epsilon)\| \leq \epsilon$ .*

The properties of  $\Phi(K)$  in Lemma 2 follow from its relation to  $\mathcal{L}(K, \lambda)$ , as detailed in [20]. Even though it is non-differentiable, one can define the SP here based on  $\Phi_{\mu_0}(K)$

---

**Algorithm 1:** Gradient Descent with max-oracle (GDmax)

---

- 1 **Inputs:** A feasible policy  $K^0$ , upper bound  $\Lambda$  for  $\lambda$ , threshold  $\epsilon$ , and the initial iteration index  $j = 0$ .
  - 2 Determine  $L_0, \ell_0$ , and  $\rho_0$  using the set  $\mathcal{G}^0$  and compute the stepsize as in (15).
  - 3 **while**  $\|\nabla \mathcal{L}_K(K^j, \lambda^j)\| > \epsilon$  **do**
  - 4     Obtain  $\lambda^j \leftarrow \arg \max_{\lambda \in \mathcal{Y}} \mathcal{L}(K^j, \lambda)$
  - 5     Update  $K^{j+1} \leftarrow K^j - \eta \nabla \mathcal{L}_K(K^j, \lambda^j)$ ;
  - 6     Set  $j \leftarrow j + 1$ .
  - 7 **end**
  - 8 **Return:** the final iterate  $K^j$ .
- 

which will be used for the convergence analysis of GD updates later on. Notably, the  $\epsilon$ -SP of  $\Phi(K)$  is equivalently related to the stationarity conditions for  $\mathcal{L}(K, \lambda)$ . According to [20, Prop. 4.12], one can utilize  $K_\epsilon$  from Lemma 2 to generate the following pair  $(\tilde{K}_\epsilon, \tilde{\lambda}_\epsilon)$  by performing an additional  $O(\epsilon^{-2})$  number of gradient updates:

$$\begin{aligned} \|\nabla \mathcal{L}_K(\tilde{K}_\epsilon, \tilde{\lambda}_\epsilon)\| &\leq \epsilon \\ \left\| \mathbb{P}_{\mathcal{Y}} \left( \tilde{\lambda}_\epsilon + (1/\ell_0) \nabla_{\lambda} \mathcal{L}(\tilde{K}_\epsilon, \tilde{\lambda}_\epsilon) \right) - \tilde{\lambda}_\epsilon \right\| &\leq \epsilon/\ell_0 \end{aligned}$$

where  $\mathbb{P}_{\mathcal{Y}}$  stands for the projection onto  $\mathcal{Y}$ . Clearly, when  $\epsilon \rightarrow 0$  this represents the Lagrangian optimality conditions for problem (10), and thus the pair  $(\tilde{K}_\epsilon, \tilde{\lambda}_\epsilon)$  can be viewed as the  $\epsilon$ -SP for  $\mathcal{L}(K, \lambda)$ .

We can solve (13) using iterative GD updates, as tabulated in Algorithm 1. With an initial  $K^0$ , we need to find the subgradient of  $\Phi(K^j)$  at every iteration  $j$ . Interestingly, this is equivalent to the gradient of  $\mathcal{L}$  over  $K^j$  [20]; i.e.,  $\partial \Phi(K^j) = \nabla \mathcal{L}_K(K^j, \lambda^j)$  with  $\lambda^j$  being the optimal multiplier for the given  $K^j$ . Hence, the Lagrangian  $\mathcal{L}$  will be used to perform the GD updates for  $\Phi(K)$  minimization. The convergence of Algorithm 1 can be established below with the detailed proof in [24, Appendix A].

**Theorem 1.** *With an initial  $K^0 \in \mathcal{K}$  and by setting stepsize*

$$\eta \leq \min \left\{ \frac{\epsilon^2}{4\ell_0 L_0^2}, \rho_0 \right\}, \quad (15)$$

*Algorithm 1 is guaranteed to converge to  $K_\epsilon$  for  $\Phi(K)$ , which can be used to obtain an  $\epsilon$ -SP for the dual problem (10). The number of iterations required for attaining  $K_\epsilon$  is  $O(\ell_0 L_0^2 \Phi_{\mu_0}(K^0)/\epsilon^4)$ .*

As discussed in [24, Appendix A], we can bound the iterative changes in  $\Phi_{\mu_0}(K^j)$ , which ensures that the sequence  $\{\Phi_{\mu_0}(K^j)\}$  is non-increasing. Thus, if we define the sublevel set to be

$$\mathcal{G}^0 := \{K \in \mathcal{K} | \Phi_{\mu_0}(K) \leq \Phi_{\mu_0}(K^0)\}, \quad (16)$$

then the iterates  $\{K^j\}$  are guaranteed to be within  $\mathcal{G}^0$ . This is exactly how one can bound the constants  $L_0$  and  $\ell_0$  as given by (11). Of course, the choice of  $\mu_0$  in the sublevel set  $\mathcal{G}_0$  depends on  $\ell_0$ , which may not be known before  $\mathcal{G}^0$  is constructed. This issue is discussed in the following remark.

---

**Algorithm 2:** Zero-Order Policy Gradient (ZOPG)

---

- 1 **Inputs:** smoothing radius  $r$ , the policy  $K$  and its perturbation  $U \in \mathcal{S}_K$ , both of  $n_K$  non-zeros.
  - 2 Obtain  $\lambda' \leftarrow \arg \max_{\lambda \in \mathcal{Y}} \mathcal{L}(K + rU, \lambda)$ ;
  - 3 Estimate the gradient  
 $\hat{\nabla}_K \mathcal{L}(K; U) = \frac{n_K}{r} \mathcal{L}(K + rU, \lambda') U$ .
  - 4 **Return:**  $\hat{\nabla}_K \mathcal{L}(K; U)$ .
- 

**Remark 1** (Sublevel set). *With initial  $K^0$  given, the set  $\mathcal{G}^0$  is defined with the value  $\mu_0$ , which depends on the upper bound of  $\ell_K$  within  $\mathcal{G}^0$  as shown in (11). This dependence can be addressed by determining the value of  $\mu_0$  in an adaptive fashion. Starting with a rough estimate of  $\ell_0$  and  $\mu_0$ , one can first construct a  $\mathcal{G}^0$  and compare the resultant bound with the original estimate on  $\ell_0$ . If the latter is larger, then  $\mathcal{G}^0$  works well. Otherwise, one can gradually increase the  $\ell_0$  estimate to achieve that condition. Our experimental experience suggests some conservative choice of stepsize can ensure the convergence in practice.*

#### IV. STOCHASTIC GD FOR MODEL-FREE LEARNING

To account for unknown system dynamics, we extend the GDmax approach to a model-free setting. The iterative gradient will be obtained via the zero-order optimization [21]. Unfortunately, this stochastic gradient update can complicate the convergence analysis as detailed later, mainly due to the aforementioned issue on local properties of LQR cost.

Zero-order policy gradient (ZOPG) has been popularly developed in recent years for model-free gradient-based learning. It provides an unbiased gradient estimate in an efficient manner. For the function  $\Phi(K)$ , ZOPG aims to evaluate the function value at any  $K$  under a structured, random perturbation from the set  $\mathcal{S}_K = \{U \in \mathcal{K} : \|U\| = 1\}$ , as detailed in Algorithm 2. Note that the structure of perturbation  $U$  is the same as that of  $K$  with non-zero entries randomly sampled from e.g., the uniform distribution, followed by a normalization step to ensure unity norm. Given a smoothing radius  $r > 0$ , the ZOPG is estimated using the resultant  $\Phi(K + rU)$  from this perturbation by finding the corresponding optimal  $\lambda$  in (13). We denote  $n_K$  as the total number of nonzero entries in  $K$ , which is used to scale the gradient estimate. Since the estimated  $\hat{\nabla}_K \mathcal{L}$  follows from matrix  $U$ , it maintains the same sparse structure given by  $K$ .

The stochastic ZOPG will make it more difficult to maintain the iterative updates to stay within a sublevel set, and likewise for bounding Lipschitz and smoothness constants. Fortunately, [4] has developed an approach to attain this condition with a high probability. Specifically, one can set up a ten-fold sublevel set, given by

$$\mathcal{G}^1 := \{K \in \mathcal{K} | \Phi_{\mu_0}(K) \leq 10 \Phi_{\mu_0}(K^0)\}. \quad (17)$$

Using  $\mathcal{G}^1$ , one can determine  $L_0, \ell_0$ , and  $\rho_0$  over the set  $\mathcal{G}^1$  similar to (11)-(12), and they will be used for the convergence analysis. Note that the choice of  $\mu_0$  in  $\mathcal{G}^1$  depends on the  $\ell_0$  value, which can be addressed as discussed in Remark 1.

---

**Algorithm 3:** Stochastic Gradient Descent with max-oracle (SGDmax)

---

- 1 **Inputs:** A feasible policy  $K^0$ , upper bound  $\Lambda$  for  $\lambda$ , threshold  $\epsilon$ , and number of ZOPG samples  $M$ .
  - 2 Determine  $L_0, \ell_0$ , and  $\rho_0$  with the set  $\mathcal{G}^1$  and compute  $r, \eta$ , and  $J$  as in (18);
  - 3 **for**  $j = 0, 1, \dots, J - 1$  **do**
  - 4     **for**  $s = 1, \dots, M$  **do**
  - 5         Sample the random  $U_s \in \mathcal{S}_K$ ;
  - 6         Use Algorithm 2 to return  $\hat{\nabla} \mathcal{L}_K(K^j; U_s)$ .
  - 7     **end**
  - 8     Update  $K^{j+1} \leftarrow K^j - \eta \left( \frac{1}{M} \sum_{s=1}^M \hat{\nabla} \mathcal{L}(K^j; U_s) \right)$ .
  - 9 **end**
  - 10 **Return:** the final iterate  $K^J$ .
- 

Algorithm 3 tabulates the ZOPG-based model-free learning approach for solving (10), termed as the Stochastic Gradient Descent with max-oracle (SGDmax) [23]. Its convergence guarantee can be established with the detailed proof in [24, Appendix B].

**Theorem 2.** *With an initial  $K^0 \in \mathcal{K}$  and a given  $\epsilon > 0$ , we can set the parameters as*

$$r \leq \min \left\{ \rho_0, \frac{L_0 \sqrt{M}}{\ell_0} \right\}, \quad \eta \leq \frac{\epsilon^2}{\alpha \ell_0 (L_0^2 + \ell_0^2 r^2 / M)},$$
$$\text{and } J = \frac{2\sqrt{10}\alpha\Phi_{\mu_0}(K^0)}{\eta\epsilon^2} \quad (18)$$

with  $L_0, \ell_0$  and  $\rho_0$  being specified using  $\mathcal{G}^1$ , and a large constant  $\alpha$ . Then, Algorithm 1 converges to the  $\epsilon$ -SP  $K_\epsilon$  with probability at least  $(0.9 - \frac{4}{\alpha} - \frac{4}{\sqrt{10}\alpha})$ .

Last, the proposed algorithms can be easily extended to the case of full feedback  $K$ , with computational advantages over existing solutions as discussed below.

**Remark 2** (Full feedback  $K$ ). *For the full feedback case, we can directly implement the proposed Algorithms 1-3 by dropping the structured set  $\mathcal{K}$ . This setting has been considered in [8] by using a dual-ascent based double-loop scheme where the inner-loop minimizes  $K$  till convergence for any fixed  $\lambda$ . In contrast, our proposed algorithms eliminate this inner-loop, which is more computationally efficient. Investigating the global convergence property of our proposed SGDmax algorithm for the full feedback case constitutes as an interesting future direction.*

#### V. NUMERICAL TESTS

To demonstrate the effectiveness of the proposed model-free learning approach, we consider the load frequency control (LFC) problem in a low-inertia networked microgrid (MG) system with a risk constraint on the frequency states. Fig. 1 depicts a radially connected system with  $N = 6$  MGs, while the detailed model parameters are given in [17]. Consider the communication graph to be the same as the MG network show in Fig. 1. Thus, each MG  $a$  can only exchange information with their neighboring MGs that are

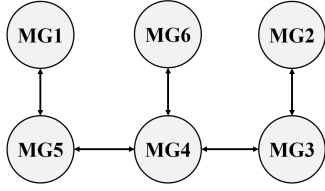


Fig. 1: A radially connected networked microgrid system.

physically connected by tie-lines, and the structured feedback  $\mathcal{K}$  is specified accordingly.

Each MG  $a$  is assumed to follow linearized power-frequency dynamics including turbine swing and primary control based on the automatic generation control (AGC) signal. Thus, the following symbols all correspond to the deviation from steady-state values as denoted by  $\Delta$ . First, the primary frequency control in each MG  $a$  is proportional to frequency deviation as  $\Delta P_{f,a} = -(1/R_a)\Delta f_a$  based on the given droop  $R_a$ . Second, the secondary AGC signal  $\Delta P_{C,a}$  constitutes as the control action  $u_t$  in (1) to be designed. The two controls jointly determine the power output of MG  $a$  as denoted by  $\Delta P_{G,a}$ . Last,  $\Delta f_a$  is also affected by the unknown load demand deviation  $\Delta P_{L,a}$  and the total power inflow  $\Delta P_{tie,a}$ , in addition to  $\Delta P_{G,a}$ . Note that  $\Delta P_{tie,a}$  is the total tie-line power inflow from all neighboring MGs due to their frequency differences, as

$$\Delta P_{tie,a} = \int \sum_{a \leftrightarrow b} K_{tie,a}(\Delta f_a - \Delta f_b) dt, \quad (19)$$

where  $a \leftrightarrow b$  indicates two MGs are connected to each other. In addition to the MG dynamics, the Area Control Error (ACE) defined as  $z_a := \beta_a \Delta f_a + \Delta P_{tie,a}$  is also a state variable as an integral control input with the bias factor  $\beta_a = D_a + 1/R_a$  [25].

Hence, MG  $a$  has the state vector  $x_a = [\Delta f_a, \Delta P_{G,a}, \Delta P_{tie,a}, \int z_a]^\top$  and the control action  $u_a = \Delta P_{C,a}$ , with load disturbance  $w_a = \Delta P_{L,a}$ . Assuming all MGs having the same parameter values, we can drop the parameter index  $a$  and represent the aggregated network dynamics by:

$$\dot{x} = (I_N \otimes A_1 + L \otimes A_2)x + (I_N \otimes B_u)u + (I_N \otimes B_w)\tilde{w}$$

with each variable collecting all MGs' respective state, action, and disturbance while the system matrices  $A_1, A_2, B_u$  and  $B_w$  are given in [24]. For the aggregated dynamics, the LQR objective cost is specified by

$$Q = I_{N_L} \otimes Q_a, \text{ and } R = I_{N_L} \otimes R_a$$

where the matrices  $Q_a$  and  $R_a$  are same for every MG  $a$  and aim to penalize the deviation of both state and action from steady-state values. As discussed in Section II, we further consider a risk constraint  $R_c(\cdot)$  in (4) for reducing the mean-variance risk in order to improve frequency regulation.

We consider the following three cases to demonstrate the impact of structured  $K$  along with the risk constraint:

- Case 1): Structured  $K$  with risk constraint
- Case 2): Full  $K$  with risk constraint
- Case 3): Full  $K$  without risk constraint

For cases 1 and 2, we implemented Algorithm 3 using SGDmax while a simple ZOPG-based algorithm [4] was

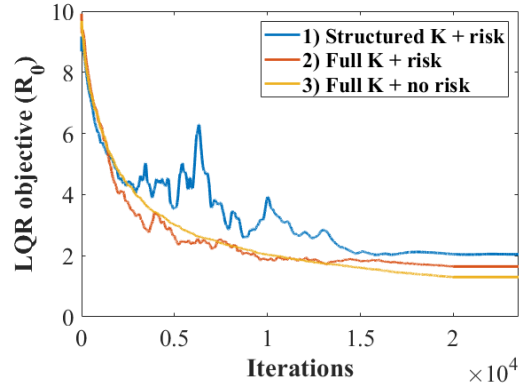


Fig. 2: Comparison of LQR objective trajectories for the three cases.

used for case 3. For all algorithms, we set  $\alpha = 10^4$  with a smoothing radius  $r = 1$  and  $M = 100$  samples for ZOPG, which makes the probability of convergence to be at least 0.887 and the stepsize  $\eta \leq 1.24 \times 10^{-4}$  as given in Theorem 2. We picked the stepsize of  $\eta = 10^{-4}$ . All three cases have shown to converge to a steady-state with sufficient updates, as shown by Fig. 2. Case 1 demonstrates the highest steady-state LQR cost out of the three, as it has the most restrictive conditions. However, the minimum LQR cost by case 1 is still pretty close to that by case 3, implying some good optimality gap. Notably, case 1 has shown some large fluctuations along the learning process, indicating a complicated geometry that the problem may have. This issue has affected its convergence speed. Interestingly, all three cases converge at the similar number of iterations.

We also test the converged policy by each case by generating a scenario that all six MGs have some random load changes in a 20-second window. Each area experiences a step load change at a random time. Fig. 3 compares the frequency deviation and the total power inflow for MG 2. Clearly, Fig. 3(a) demonstrates that the risk constraint can effectively reduce the frequency deviation, as case 2 has the smallest deviation among all three. With the risk constraint, case 1 tends to exhibit great frequency performance as well, but also shows some small oscillations possibly due to the structured feedback policy. This observation points out that limited information exchange can potentially affect the control performance. Similar patterns have been observed in Fig. 3(b). While case 1 can maintain the tie-line inflow at the same level as case 2, it still has more noticeable oscillations. As the power inflow is proportional to frequency difference, reducing the risk of frequency deviation can enhance the performance in maintaining the level of power inflow.

To sum up, our numerical tests have validated the convergence performance of the proposed SGDmax based policy gradient method for risk-constrained LQR problem with structured policy. The effectiveness of risk constraint in mitigating large state deviation have been verified, while the sparse structure of  $K$  has shown to save communication overhead at the cost of transient oscillations.

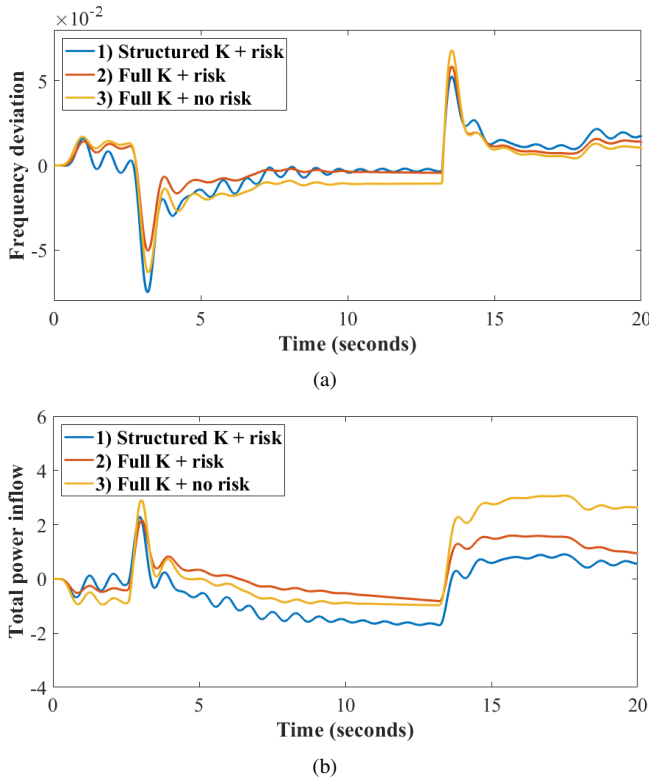


Fig. 3: Comparison of the (a) frequency deviation and (b) total power inflow at MG 2 for the three cases.

## VI. CONCLUSIONS

The paper developed a model-free learning framework for risk-constrained LQR problem under structured feedback in a networked setting. By dualizing the risk constraint, we consider the minimax reformulation of the dual problem and leverage the stochastic (S)GDmax algorithms to approach the stationary points (SPs). Specifically, the SGDmax algorithm relies on the ZOPG-based updates, making it suitable for model-free learning. Using the recent results on the local Lipschitz and smoothness of LQR cost, convergence of the (S)GDmax algorithms can be established by properly bounding the related constants for choosing the stepsize. Notably, for SGDmax the convergence can only be shown with a high probability, due to the additional noise in the gradient estimate. Numerical tests on a networked microgrid system have validated the convergence of our proposed algorithms while demonstrating the impact of risk and structured constraints for the LQR problem. Exciting future research directions open up on investigating the landscape for the converged SP in the structured feedback case and establishing the global convergence for the full feedback case.

## REFERENCES

- [1] R. E. Kalman *et al.*, "Contributions to the theory of optimal control," *Bol. soc. mat. mexicana*, vol. 5, no. 2, pp. 102–119, 1960.
- [2] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*. USA: Prentice-Hall, Inc., 1990.
- [3] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, "LQR through the lens of first order methods: Discrete-time case," *arXiv preprint arXiv:1907.08921*, 2019.
- [4] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright, "Derivative-free methods for policy optimization: Guarantees for linear quadratic systems," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 2916–2925.
- [5] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," 2018.
- [6] A. Tsiamis, D. S. Kalogerias, L. F. O. Chamon, A. Ribeiro, and G. J. Pappas, "Risk-Constrained Linear-Quadratic Regulators," in *2020 59th IEEE Conference on Decision and Control (CDC)*, 2020, pp. 3040–3047.
- [7] A. Tsiamis, D. S. Kalogerias, A. Ribeiro, and G. J. Pappas, "Linear Quadratic Control with Risk Constraints," *arXiv preprint arXiv:2112.07564*, 2021.
- [8] F. Zhao, K. You, and T. Başar, "Global Convergence of Policy Gradient Primal-dual Methods for Risk-constrained LQRs," *arXiv preprint arXiv:2104.04901*, 2021.
- [9] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained reinforcement learning has zero duality gap," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [10] D. Ding, K. Zhang, T. Başar, and M. Jovanovic, "Natural policy gradient primal-dual method for constrained markov decision processes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8378–8390, 2020.
- [11] P. Shah and P. A. Parrilo, "H2-optimal decentralized control over posets: A state-space solution for state-feedback," *IEEE Transactions on Automatic Control*, vol. 58, no. 12, pp. 3084–3096, 2013.
- [12] M. Rotkowitz and S. Lall, "A Characterization of Convex Problems in Decentralized Control," *IEEE Transactions on Automatic Control*, vol. 51, no. 2, pp. 274–286, 2006.
- [13] L. Ye, H. Zhu, and V. Gupta, "On the sample complexity of decentralized linear quadratic regulator with partially nested information structure," *arXiv preprint arXiv:2110.07112*, 2021.
- [14] F. Dörfler, M. R. Jovanović, M. Chertkov, and F. Bullo, "Sparse and optimal wide-area damping control in power networks," in *2013 American Control Conference*. IEEE, 2013, pp. 4289–4294.
- [15] A. Chakraborty, "Wide-area control of power systems: Employing data-driven, hierarchical reinforcement learning," *IEEE Electrification Magazine*, vol. 9, no. 1, pp. 45–52, 2021.
- [16] C. Wang, J. Duan, B. Fan, Q. Yang, and W. Liu, "Decentralized High-Performance Control of DC Microgrids," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3355–3363, 2019.
- [17] E. E. Vlahakis, L. D. Dritsas, and G. D. Halikias, "Distributed LQR design for identical dynamically coupled systems: Application to Load Frequency Control of multi-area Power Grid," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 4471–4476.
- [18] Y. Li, Y. Tang, R. Zhang, and N. Li, "Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach," *IEEE Transactions on Automatic Control*, 2021.
- [19] H. Feng and J. Lavaei, "On the exponential number of connected components for the feasible set of optimal decentralized control problems," in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 1430–1437.
- [20] T. Lin, C. Jin, and M. Jordan, "On gradient descent ascent for nonconvex-concave minimax problems," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6083–6093.
- [21] J. C. Spall, "A one-measurement form of simultaneous perturbation stochastic approximation," *Automatica*, vol. 33, no. 1, pp. 109–112, 1997.
- [22] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
- [23] C. Jin, P. Netrapalli, and M. Jordan, "What is local optimality in nonconvex-nonconcave minimax optimization?" in *International Conference on Machine Learning*. PMLR, 2020, pp. 4880–4889.
- [24] K. Kwon, L. Ye, V. Gupta, and H. Zhu, "Model-free learning for risk-constrained linear quadratic regulator with structured feedback in networked systems," 2022. [Online]. Available: <https://arxiv.org/abs/2204.01779>
- [25] E. Vlahakis, L. Dritsas, and G. Halikias, "Distributed LQR design for a class of large-scale multi-area power systems," *Energies*, vol. 12, no. 14, p. 2664, 2019.