Proceedings of Machine Learning Research vol 144:1–13, 2021

Primal-dual Learning for the Model-free Risk-constrained Linear Quadratic Regulator

Feiran Zhao ZHAOFR18@MAILS.TSINGHUA.EDU.CN
Keyou You YOUKY@TSINGHUA.EDU.CN
Department of Automation and BNRist, Tsinghua University

Abstract

Risk-aware control, though with promise to tackle unexpected events, requires a known exact dynamical model. In this work, we propose a model-free framework to learn a risk-aware controller of a linear system. We formulate it as a discrete-time infinite-horizon LQR problem with a state predictive variance constraint. Since its optimal policy is known as an affine feedback, i.e., $u^*(x) = -Kx + l$, we alternatively optimize the gain pair (K, l) by designing a primal-dual learning algorithm. First, we observe that the Lagrangian function enjoys an important local gradient dominance property. Based on it, we then show that there is no duality gap despite the non-convex optimization landscape. Furthermore, we propose a primal-dual algorithm with global convergence to learn the optimal policy-multiplier pair. Finally, we validate our results via simulations.

Keywords: Risk-aware control; Policy optimization; Reinforcement learning; Optimal control; Constrained Markov decision process.

1. Introduction

Stochastic optimal control (Åström, 2012) is a well-studied framework that deals with inherent random noises in the dynamical system. Its classical formulation targets to minimize an expected long-term cost, which is risk-neutral as it only optimize the expectation without explicit considerations on the variability of the state. Thus, the system behaviours may be easily influenced by less probable but large noises, leading to catastrophic consequences for the safety-critical systems. In decades, the risk-aware controllers have been proposed to tackle the extreme noises with a slight sacrifice of average performance (Sopasakis et al., 2019; Jacobson, 1973; Moore et al., 1997; Bäuerle and Rieder, 2014; Roulet et al., 2020). For example, the risk is typically addressed by replacing the cost with its exponentiation (Speyer et al., 1992; Moore et al., 1997) or optimizing the risk measure (Chapman et al., 2019) e.g., Conditional Value-at-Risk (CVaR) (Rockafellar et al., 2000). However, most of them are model-based (Chapman et al., 2019; Moore et al., 1997; Speyer et al., 1992) and hence not directly applicable when the exact dynamical model is unknown.

Model-free reinforcement learning (RL) (Sutton et al., 1998; Bertsekas, 2019) has achieved tremendous progress recently in the continuous control field (Mnih et al., 2015; Lillicrap et al., 2016). Instead of identifying the underlying dynamical model first, it approaches the control problem by directly searching for an optimal policy that minimizes the estimated cost function. Under the RL framework, the prevalent risk-averse methods (Wen and Topcu, 2018; Prashanth L and Fu, 2018; Borkar and Jain, 2014) take the risk into consideration by, e.g., adding a risk-related cumulative cost constraint to the Markov decision process (MDP) (Paternain et al., 2019; Chow et al., 2017; Yu et al., 2019; Tessler et al., 2018), or formulating the risk as an adversary (Pan et al., 2019).

Though empirically successful on the continuous control benchmarks (Pan et al., 2019; Tessler et al., 2018), they typically lack strong theoretical guarantees, hampering their physical-world applications.

Recent advances in the context of policy optimization (PO) for the linear quadratic regulator (LQR) (Bertsekas, 1995), including policy gradient (Fazel et al., 2018; Bu et al., 2019; Zhang et al., 2019, 2020) and random search methods (Malik et al., 2019; Mohammadi et al., 2020), have been shown to enjoy the global convergence in spite of the non-convex nature of the optimization landscape. Some works focus on the LQR variants e.g., robust control with multiplicative noises (Gravell et al., 2020), distributed LQR (Li et al., 2019) and Markov jump linear systems(Jansch-Porto et al., 2020). In particular, the PO for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantees is analyzed in Zhang et al. (2020) for a risk-sensitive linear exponential quadratic Gaussian (LEQG) (Whittle, 1981) instance. However, to the best of our knowledge, there is no such analysis for the risk-aware formulation with a risk constraint explicitly concerned.

In this paper, we consider the learning problem for the model-free risk-aware controller. Inspired by Tsiamis et al. (2020), we formulate it as a discrete-time infinite-horizon LQR problem with a one-step predicted state variability constraint. By Zhao et al. (2021a), the solution to it is an affine state feedback policy. Thus, we can alternatively optimize over the stabilizing affine policy set. Nevertheless, in contrast to LQR, three challenges exist in our setting. Firstly, the constraint optimization problem is non-convex in that the objective function, the risk constraint and the stabilizing policy set are all non-convex. Moreover, the optimization variable in LQR is a single feedback gain (Fazel et al., 2018), while in our case it is a gain pair and hence the optimization landscape is not clear yet. Finally, the first-order optimization methods cannot be used since the dynamical model is unknown.

This work proposes a primal-dual learning framework to solve the risk-constrained LQR problem. Alongside, we take an initial step towards understanding the theoretical aspects of PO for the constrained LQR. Our contributions are summarized below. Firstly, in spite of the constrained non-convex optimization nature, we show that the strong duality holds. Secondly, we study the optimization landscape of the Lagrangian function over the stabilizing affine policy set. In particular, we find that it enjoys two favourable properties, i.e., the local gradient dominance and Lipschitz property. Thirdly, we propose a primal-dual algorithm to learn the optimal policy-multiplier pair and show its global convergence.

2. Problem Formulation

In the standard setup of LQR, we consider a time-invariant discrete linear stochastic system with full state observations,

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{1}$$

where the next state x_{t+1} is a linear combination of the current state $x_t \in \mathbb{R}^n$, the control $u_t \in \mathbb{R}^m$, and the random noise $w_t \in \mathbb{R}^d$. The model parameters are denoted as $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$.

The goal of infinite-horizon LQR is to find a control policy π which minimizes an average long-term cost, i.e.,

$$\text{minimize} \quad \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}(x_t^\top Q x_t + u_t^\top R u_t) \tag{2}$$

where $u_t = \pi(h_t)$ with the history trajectory $h_t = \{x_0, u_0, \cdots, x_{t-1}, u_{t-1}\}$ and the expectation is taken with respect to the random noise w_t. Throughout the paper, we make the following assumption standard in the control theory (Bertsekas, 1995).

Assumption 1 *Q is positive semi-definite and R is positive definite. The pair (A, B) is stabilizable and $(A, Q^{1/2})$ is observable.*

Under Assumption 1, solving (2) yields a unique linear state feedback policy $u_t = -Kx_t$ when w_t has zero mean. Clearly, the classical LQR is risk-neutral as it aims to minimize only the expected cost. Thus, the state may be largely influenced by the low-probability but large noises, especially those with heavy-tailed distributions.

In this paper, we study the infinite-horizon risk-constrained LQR in Zhao et al. (2021a) and solve it in a model-free approach. That is,

$$
\begin{aligned}
&\text{minimize} \quad \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}(x_t^\top Q x_t + u_t^\top R u_t) \\
&\text{subject to} \quad \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}(x_t^\top Q x_t - \mathbb{E}[x_t^\top Q x_t | h_t])^2 \leq \rho
\end{aligned}
\tag{3}
$$

where $\rho > 0$ is a user-defined risk tolerance constant. In contrast to standard LQR (2), we do not require the noise w_t to be zero-mean. Instead, we only assume a finite 4th-order moment of w_t (Tsiamis et al., 2020).

In our recent work Zhao et al. (2021a), we have shown that the optimal policy to (3) is an affine state feedback, i.e., $u^*(x) = -K^*x + l^*$, which is also able to stabilize the system. Exploiting this affine structure, we can alternatively optimize the gain pair (K, l). Define the mean $\bar{w} = \mathbb{E}[w_t]$, the covariance $W = \mathbb{E}[(w_t - \bar{w})(w_t - \bar{w})^\top] > 0$, higher-order weighted statistics $M_3 = \mathbb{E}[(w_i - \bar{w})(w_i - \bar{w})^\top Q (w_i - \bar{w})]$ and $m_4 = \mathbb{E}[(w_i - \bar{w})^\top Q (w_i - \bar{w}) - \mathrm{tr}(WQ)]^2$ of the noise w_t. Given that w_t has a finite 4-order moment, (3) can be reformulated by Zhao et al. (2021a) as

$$
\begin{aligned}
&\text{minimize} \quad J(K,l) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}(x_t^\top Q x_t + u_t^\top R u_t) \\
&\text{subject to} \quad J_c(K,l) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}(4x_t^\top QWQx_t + 4x_t^\top QM_3) \leq \bar{\rho}
\end{aligned}
\tag{4}
$$

with $u_t = -Kx_t + l$ and $\bar{\rho} = \rho - m_4 + 4\,\mathrm{tr}\{(WQ)^2\}$. In (4), (K, l) is the optimization variable.

The direct PO for the risk-neutral formulation (2) has been well studied and typically enjoys the convergence guarantee, including random search and policy gradient methods Fazel et al. (2018). However, (4) is not only a non-convex constrained optimization problem, but also differs from (2) in that the optimization variable is a gain pair (K, l). In this paper, we study the analytical property of the constrained optimization problem (4). Furthermore, we propose a convergent primal-dual algorithm to solve it exactly by solely using data.

3. Primal-dual Optimization for Risk-constrained LQR

In this section, we introduce the primal-dual method for solving the risk-constrained LQR problem in (4). In contrast to Fazel et al. (2018), its Lagrangian function is only locally gradient dominated and locally Lipschitz with respect to the policy. Moreover, we establish the strong duality for the non-convex constrained optimization problem (4).

3.1. Primal-dual method

In the rest of the paper, we use the augmented matrix $X = [K\ l]$ to denote the optimization variable. Define $\mathcal{S} = \{X = [K\ l] | \rho(A - BK) < 1, K \in \mathbb{R}^{n\times m}, l \in \mathbb{R}^n\}$, where $\rho(\cdot)$ denotes the spectral radius. Clearly, we have $J(X) < +\infty$ and $J_c(X) < +\infty$ if and only if $X \in \mathcal{S}$. Let $\mu \geq 0$ denote the Lagrange multiplier and $Q_\mu = Q + 4\mu QWQ$ and $S = 2\mu QM_3$. We define the Lagrangian function of (4) as

$$\mathcal{L}(X, \mu) = J(X) + \mu(J_c(X) - \bar{\rho}) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1} c_\mu(x_t, u_t), \tag{5}$$

where $c_\mu(x_t, u_t) = x_t^\top Q_\mu x_t + 2x_t^\top S + u_t^\top R u_t - \mu\bar{\rho}$, which is a reshaped cost with a risk weight μ that balances the objective and the risk. Accordingly, we define the dual function $D(\mu) = \min_X \mathcal{L}(X, \mu)$ and the dual problem

$$\max_{\mu\geq 0}\ D(\mu) = \max_{\mu\geq 0}\ \min_{X\in\mathcal{S}} \mathcal{L}(X, \mu). \tag{6}$$

Our primal-dual method is iteratively given as

$$X_j \in \operatorname*{argmin}_{X\in\mathcal{S}}\ \mathcal{L}(X, \mu_j), \tag{7}$$

$$\mu_{j+1} = [\mu_j + \xi_j \cdot \omega(\mu_j)]_+, \tag{8}$$

where the stepsize $\xi_j > 0$, $\omega(\mu_j)$ is a subgradient of $D(\mu)$ at μ_j and $[x]_+ = \max\{0, x\}$ for any $x \in \mathbb{R}$.

To guarantee the global convergence of the primal-dual method, the strong duality between the primal problem and dual problem is essential. However, the constrained optimization problem (4) is non-convex and, therefore, the strong duality does not trivially follow. Moreover, the primal-dual method requires to solve (7) under a fixed multiplier μ. Though for LQR problems some model-free algorithms are guaranteed to find an optimal gain K (Fazel et al., 2018; Malik et al., 2019; Mohammadi et al., 2020), they cannot be directly applied as our optimization variable is a gain pair (K, l). In particular, these algorithms exploit favourable properties of the objective function such as gradient dominance (Fazel et al., 2018) and Lipschitz continuity (Malik et al., 2019), which are unclear for $\mathcal{L}(K, l, \mu)$. In what follows, we work towards addressing these problems.

3.2. Closed-form of the Lagrangian Function and its Gradient

We first derive the closed-form of $\mathcal{L}(X, \mu)$. It follows from (5) that $\mathcal{L}(X, \mu)$ is finite if and only if $X \in \mathcal{S}$. For a stabilizing policy $X \in \mathcal{S}$, the state has a stationary distribution, the mean $\bar{x}_{K,l}$ of which satisfies $\bar{x}_{K,l} = (A - BK)\bar{x}_{K,l} + Bl + \bar{w}$, and its correlation matrix can be solved through a Lyapunov equation

$$\Sigma_K = W + (A - BK)\Sigma_K(A - BK)^\top. \tag{9}$$

Suppose that $P_K \geq 0$ is the solution of the Lyapunov equation

$$P_K = Q_\mu + K^\top RK + (A - BK)^\top P_K(A - BK)$$

and let $E_K = (R + B^\top P_K B)K - B^\top P_K A$ and $V = (I - (A - BK))^{-1}$.

Proposition 1 (Closed-form expression) *The Lagrangian function $\mathcal{L}(X,\mu)$ is given by*

$$\mathcal{L}(K,l,\mu)=\operatorname{tr}\{P_K(W+(Bl+\bar{w})(Bl+\bar{w})^\top)\}+g_{K,l}^\top(Bl+\bar{w})+l^\top Rl-\mu\bar{\rho}. \tag{10}$$

where $g_{K,l}^\top=2(-l^\top E_K+S^\top+\bar{w}^\top P_K(A-BK))V$ and $z_{K,l}$ is a constant.

Proposition 2 (Policy gradient expression) *The gradient of $\mathcal{L}(X,\mu)$ with respect to X is given by $\nabla\mathcal{L}(X,\mu)=2\,[E_K\;\;G_{K,l}]\,\Phi_{K,l}$, where $G_{K,l}=(R+B^\top P_KB)l+B^\top P_K\bar{w}+\frac{1}{2}B^\top g_{K,l}$ and $\Phi_{K,l}$ is the correlation matrix*

$$\Phi_{K,l}=\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}\begin{bmatrix}x_t\\-1\end{bmatrix}\begin{bmatrix}x_t\\-1\end{bmatrix}^\top=\begin{bmatrix}\Sigma_K+\bar{x}_{K,l}\bar{x}_{K,l}^\top & -\bar{x}_{K,l}\\-\bar{x}_{K,l}^\top & 1\end{bmatrix}>0. \tag{11}$$

Since $\Phi_{K,l}$ is positive definite, the stationary point of $\mathcal{L}(X,\mu)$ can be uniquely solved by setting the gradients to zero as $X^*(\mu)=[K^*(\mu)\;l^*(\mu)]$ with

$$\begin{aligned}K^*(\mu)&=(R+B^\top P_{K^*(\mu)}B)^{-1}B^\top P_{K^*(\mu)}A,\\ l^*(\mu)&=-(R+B^\top P_{K^*(\mu)}B)^{-1}B^\top V^\top(P_{K^*(\mu)}\bar{w}+S).\end{aligned} \tag{12}$$

3.3. Properties of the Lagrangian Function

The minimization on $\mathcal{L}(X,\mu)$ in (10) is a non-convex optimization problem, in that both the objective function and the stabilizing policy set $\mathcal{S}$ are non-convex, which poses challenges in solving (7) with standard policy gradient-based methods. In the PO for classical LQR problems (2) (Fazel et al., 2018), this is alleviated by observing that the objective function is globally gradient dominated. For a differentiable function $f(x):\mathbb{R}^n\to\mathbb{R}$ with a finite global minimum f^*, it is *globally gradient dominated* if

$$f(x)-f^*\le\lambda\|\nabla f(x)\|^2,\;\;\forall x\in\operatorname{dom}(f)\subseteq\mathbb{R}^n \tag{13}$$

where $\lambda\ge 0$ is a gradient dominance constant. Clearly, it implies that a stationary point must be the global minimizer. Hence, if $f(z)$ is also Lipschitz smooth, one would expect that the gradient-based algorithms converge at a linear rate to the global minimum (Malik et al., 2019).

We show that $\mathcal{L}(X,\mu)$ enjoys a local gradient dominance property, which is weaker than the more common global one in the sense that it only holds locally over a compact set. Before formalizing it, we note that the compact set can be constructed by observing that $\mathcal{L}(X,\mu)$ is coercive.

Lemma 3 (Coercivity) *Under a fixed $\mu>0$, the Lagrangian $\mathcal{L}(X,\mu)$ is coercive in X in the sense that $\lim_{X\to\partial\mathcal{S}}\mathcal{L}(X,\mu)=+\infty$, where $\partial\mathcal{S}$ denotes the boundary of $\mathcal{S}$. Moreover, it has a compact α-sublevel set*

$$\mathcal{S}_\alpha=\{X\in\mathbb{R}^{m\times(n+1)}|\mathcal{L}(X,\mu)\le\alpha\}. \tag{14}$$

Then, we obtain the local gradient dominance property of $\mathcal{L}(X,\mu)$ over $\mathcal{S}_\alpha$.

Lemma 4 (Local Gradient Dominance) *$\mathcal{L}(X,\mu)$ is gradient dominated locally over the compact set $\mathcal{S}_\alpha$ in (14), namely,*

$$\mathcal{L}(X,\mu)-\mathcal{L}(X^*(\mu),\mu)\le\lambda_\alpha\operatorname{tr}\{\nabla\mathcal{L}^\top\nabla\mathcal{L}\},$$

where $\lambda_\alpha=\frac{\|\Phi^*\|}{4\sigma_{min}(R)\cdot\phi_\alpha^2}>0$ is a constant related to $\mathcal{S}_\alpha$ and $\phi_\alpha=\min_{X\in\mathcal{S}_\alpha}\sigma_{min}(\Phi_{K,l})>0$.

By the local gradient dominance and the coercivity, we can determine the global minimizer of the Lagrangian.

Theorem 5 *The critial point $X^*(\mu)$ in (12) is the unique global minimizer of $\mathcal{L}(X,\mu)$.*

Finally, we show that both $\mathcal{L}(X,\mu)$ and its gradient $\nabla\mathcal{L}$ are locally Lipschitz.

Lemma 6 (Locally Lipschitz) *There exist positive scalars $(\zeta_X,\beta_X,\gamma_X)$ that depends on the current policy X, such that for all policies $X' \in \mathcal{S}$ satisfying $\|X'-X\| \le \gamma_X$, we have*

$$|\mathcal{L}(X',\mu)-\mathcal{L}(X,\mu)| \le \zeta_X\|X'-X\| \text{ and}$$
$$\|\nabla\mathcal{L}(X',\mu)-\nabla\mathcal{L}(X,\mu)\| \le \beta_X\|X'-X\|.$$

Note that the scalars ζ_X,β_X,γ_X in Lemma 6 are functions of X as well as the problem parameters e.g., (A,B,Q,R).

3.4. Strong Duality

The duality analysis is generally difficult for a non-convex constrained optimization problem. Nevertheless, we show that the strong duality between the primal problem (4) and dual problem (6) indeed holds, by leveraging the established properties of the Lagrangian.

Theorem 7 *Suppose that the Slater's condition in (4) holds, i.e., there exists a policy $\widetilde{X} \in \mathcal{S}$ such that $J_c(\widetilde{X}) < \bar{\rho}$, then there is no duality gap for the primal problem (4) and the dual problem (6).*

4. Primal-dual Learning Algorithm for the Risk-constrained LQR

In the model-free setting, (A,B) is unknown and the gradient $\nabla\mathcal{L}(X,\mu)$ cannot be computed directly. Thus, we estimate the gradient via noisy samples of the Lagrangian. By focusing on a sublevel set, we can leverage the gradient dominance and smoothness to develop a random search method to solve (7). Moreover, we propose a primal-dual algorithm to find an optimal pair (X^*,μ^*) where an estimation of the subgradient is also used for the dual ascent in (8).

4.1. Random Search for (7)

Assume that we have a cost oracle, which returns a noisy evaluation of $\mathcal{L}(X,\mu)$ and $J_c(X)$ as

$$\widehat{\mathcal{L}}(X,\mu) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1} c_\mu(x_t,u_t) \text{ and } \widehat{J}_c(X) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(4x_t^\top QWQx_t + 4x_t^\top QM_3),$$

respectively. In practice, T is selected to be sufficiently large since the estimation error in the cost decreases quickly as $T\to\infty$ (Malik et al., 2019). Clearly, the oracle is weaker than the commonly assumed state-input trajectories in the model-free setting.

We develop a stochastic zero-order algorithm to solve (7) in Algorithm 1. The difficulties in the convergence analysis of Algorithm 1 hinge on that (a) both the objective function $\mathcal{L}(X,\mu)$ and its feasible set $\mathcal{S}$ are non-convex; (b) unlike in Fazel et al. (2018), the gradient dominance property does not hold globally; (c) $\mathcal{L}(X,\mu)$ is infinite for $X \notin \mathcal{S}$ therefore the step size η must be chosen

Algorithm 1: Random search algorithm to solve (7)

Input: Initial policy X_0, number of iterations N, smoothing radius r, step size η, multiplier μ.
for $i = 0, 1, \cdots, N-1$ **do**
  Sample a perturbation $U \in \mathbb{R}^{m \times n}$ uniformly form a unit ball and apply $X = X_{(i)} + rU$;
  Obtain a noisy Lagrangian function $\widehat{\mathcal{L}}(X, \mu)$ from the oracle;
  Compute the stochastic gradient $\widehat{\nabla \mathcal{L}} = \widehat{\mathcal{L}}(X, \mu)\frac{n}{r^2}U$;
  Update $X_{(i+1)} = X_{(i)} - \eta\widehat{\nabla \mathcal{L}}$;
end

carefully. Motivated by Malik et al. (2019), we address these problems by analysing over a compact set and showing that the algorithm remains in it with large probabilities.

Define the gap between an initial policy $X_{(0)}$ and $X^*(\mu)$ as $\Delta_0 = \mathcal{L}(X_{(0)}, \mu) - \mathcal{L}(X^*(\mu), \mu)$ as well as the compact set

$$\mathcal{S}_0 = \left\{X \mid \mathcal{L}(X, \mu) - \mathcal{L}(X_{(0)}, \mu) \leq 10\Delta_0\right\} \subset \mathcal{S}. \tag{15}$$

We denote the gradient dominance constant over $\mathcal{S}_0$ as λ_0. Also, define the Lipschitz constants

$$\beta_0 = \sup_{X \in \mathcal{S}_0} \beta_X, \qquad \zeta_0 = \sup_{X \in \mathcal{S}_0} \zeta_X, \qquad \gamma_0 = \inf_{X \in \mathcal{S}_0} \gamma_X. \tag{16}$$

By doing so, the properties hold globally on $\mathcal{S}_0$. To ensure the step size η not too large, it should be set according to the variance of the estimated gradient $\widehat{\nabla \mathcal{L}}$. The gradient norm is defined as

$$G_\infty = \sup_{X \in \mathcal{S}_0} \|\widehat{\nabla \mathcal{L}}\|_2, \quad \text{and} \quad G_2 = \sup_{X \in \mathcal{S}_0} \mathbb{E}\|\widehat{\nabla \mathcal{L}} - \mathbb{E}[\widehat{\nabla \mathcal{L}}|X]\|_2^2. \tag{17}$$

We make the following assumption for the noise w_t to ensure the existence of the gradient norm.

Assumption 2 *The noise w_t is uniformly bounded, i.e., $\|w_t\| \leq v$ where $v > 0$ is a constant.*

The following theorem shows that with a large probability, Algorithm 1 converges and $\{X_{(i)}\}$ always remain in $\mathcal{S}_0$. For notional convenience, we denote $\theta_0 = \min\{\frac{1}{2\beta_0}, \frac{\gamma_0}{\zeta_0}\}$.

Theorem 8 *Suppose that the step-size and smoothing radius are chosen such that*

$$\eta \leq \min\{\frac{\epsilon}{240\lambda_0\beta_0 G_2}, \frac{1}{2\beta_0}, \frac{\gamma_0}{G_\infty}\} \quad \textit{and} \quad r \leq \min\{\frac{\theta_0}{8\lambda_0\beta_0}\sqrt{\frac{\epsilon}{15}}, \frac{1}{2\beta_0}\sqrt{\frac{\epsilon}{30\lambda_0}}, \gamma_0\}.$$

Then, for any error tolerance ϵ such that $\epsilon \log(120\Delta_0/\epsilon) < \frac{10}{3}\Delta_0$ and $N = \frac{4\lambda_0}{\eta}\log\left(\frac{120\Delta_0}{\epsilon}\right)$, with probability greater than $\frac{3}{4}$ the iterations in Algorithm 1 yield a controller X_N such that

$$\mathcal{L}(X_{(N)}, \mu) - \mathcal{L}(X^*, \mu) \leq \epsilon.$$

The proof follows from Theorem 1 in Malik et al. (2019) and we extend it in that we derive the gradient norm bound for the average cost setting. In view of Furieri et al. (2020), the convergence probability in Theorem 8 can be improved to $1 - \delta$ for any $0 < \delta < 1$ by working on $\mathcal{S}_\delta = \{X \mid \mathcal{L}(X, \mu) - \mathcal{L}(X_{(0)}, \mu) \leq 10\delta^{-1}\Delta_0\} \subset \mathcal{S}$. For simplicity, we adopt the methodology in Malik et al. (2019).

Algorithm 2: Primal-dual learning algorithm for the risk-constrained LQR

Input: Initial multiplier μ_1, step size ξ_j, $j \in \{1, 2, \dots\}$.
for $j = 1, 2, \dots$ **do**
  Step 1: learning the dual function
  Learn a policy $X_j \in \operatorname*{argmin}_{X \in \mathcal{S}} \mathcal{L}(X, \mu_j)$ by Algorithm 1;
  Step 2: dual ascent
  Obtain a noisy sample $\widehat{J}_c(X_j)$ from the oracle;
  Estimate the subgradient $\hat{\omega}(\mu_j)$ by (19);
  Update the dual variable by $\mu_{j+1} = [\mu_j + \xi_j \hat{\omega}(\mu_j)]_+$;
end

4.2. Primal-dual Algorithm

By dual theory (Nesterov, 2013; Nedić and Ozdaglar, 2009), the subgradient of $D(\mu)$ is given as

$$\omega(\mu) = J_c(X^*(\mu)) - \bar{\rho}, \tag{18}$$

However, $J_c(X^*(\mu))$ cannot be computed directly as we do not have a dynamical model. To this end, we estimate it by a noisy sample from the oracle. The subgradient $\omega(\mu)$ is approximated as

$$\widehat{\omega}(\mu) = \widehat{J}_c(X^*(\mu)) - \bar{\rho} \tag{19}$$

We present our complete primal-dual algorithm in Algorithm 2. In general, there is no guarantee that a primal variable sequence will converge to the optimal solution unless the subdifferential at the dual variables is a singleton (Bertsekas, 1997; Boyd et al., 2004). Fortunately, this is indeed the case for (4) as minimizing the Lagrangian function yields a unique solution, which implies that the subgradient in (18) is actually a gradient. Furthermore, we analyze its convergence by leveraging the boundedness of the gradient norm $\|\hat{\omega}(\mu)\|$, which is evidenced by the fact that a stabilizing policy $X^*(\mu)$ yields a finite cost.

Theorem 9 *Let $\mathbb{E}\|\hat{\omega}(\mu_j)\| \le b$ and $\mathbb{E}\|\mu_j\| \le e$ with $b, e > 0$. Define $\bar{\mu}_j = \frac{1}{j}\sum_{i=1}^{j} \mu_i$. Then, by selecting a diminishing step size $\xi_j = \frac{1}{be}\sqrt{\frac{2}{j}}$, Algorithm 2 satisfies*

$$D^* - \mathbb{E}[D(\bar{\mu}_j)] \le \frac{3be}{\sqrt{j}}.$$

5. Simulation Results

In the experiment, we consider an unmanned aerial vehicle (UAV) that operates in a 2-D x-y plane. The discrete-time dynamical model is given by a double integrator as

$$x_{k+1} = \begin{bmatrix} 1 & 0.5 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0 & 1 \end{bmatrix} x_k + \begin{bmatrix} 0.125 & 0 \\ 0.5 & 0 \\ 0 & 0.125 \\ 0 & 0.5 \end{bmatrix} (u_k + w_k), \tag{20}$$

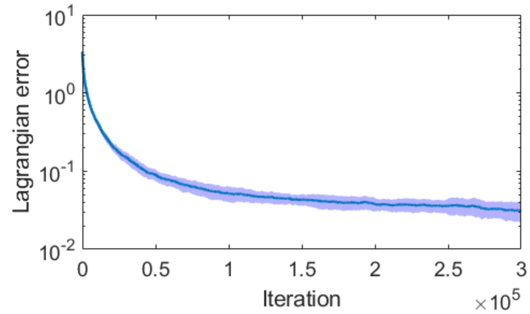
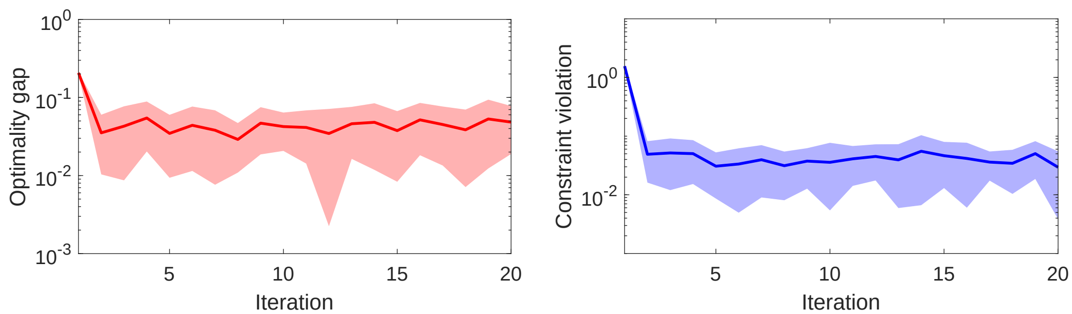


Figure 1: Relative Lagrangian error $(\mathcal{L}(X_i, \mu) - D(\mu))/D(\mu)$ of Algorithm 1 for a fixed μ. The bold centreline denotes the mean of 20 trials and the shaded region demonstrates their standard deviation.



(*a*) Optimality gap $|J(X_j) - J(X^*)|/J(X^*)$. (*b*) Risk constraint violation $|J_c(X_j) - \bar{\rho}|/\bar{\rho}$.

Figure 2: Convergence of the primal-dual learning method. The centreline denotes the mean of 20 trials and the shaded region demonstrates their standard deviation.

where $(x_{k,1}, x_{k,3})$ and $(x_{k,2}, x_{k,4})$ denote the position and velocity, respectively, u_k represents the acceleration and w_k is the input disturbance from the wind. Suppose that the gust $w_{k,1}$ in the direction of $x_{k,1}$ is subject to a mixed Gaussian distribution of $\mathcal{N}(3, 30)$ and $\mathcal{N}(8, 60)$ with weights 0.2 and 0.8, respectively. In contrast, the gust $w_{k,2}$ in the orthogonal direction satisfies $w_{k,2} \sim \mathcal{N}(0, 0.01)$. We set the penalty matrix in (4) as $Q = \text{diag}(1, 0.1, 2, 0.2)$ and $R = \text{diag}(1, 1)$.

We verify the convergence of the proposed primal-dual learning method by examining the optimality gap and the risk constraint violation. Since the system (20) is open-loop unstable, we select an initial policy

$$K_0 = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \end{bmatrix} \text{ and } l_0 = \begin{bmatrix} -6 \\ 0 \end{bmatrix},$$

which is readily stabilizing. The risk tolerance in (4) is set as $\bar{\rho} = 15$.

We first perform the random search in Algorithm 1 to solve (7). We set the initial multiplier as $\mu_1 = 0$, the smoothing radius as $r = 0.2$ and the sample horizon of the oracle as $T = 100$. We empirically select the step size $\eta = 1 \times 10^{-5}$ to yield a good performance, which is common in the policy optimization of LQR problems. We perform 20 independent trials and display the relative

Lagrangian error when $\mu = 2$ in Fig. 1. Clearly, Algorithm 1 converges to relative error 3% within 3×10^5 iterations and exhibits small variance.

Then, we conduct our primal-dual learning method in Algorithm 2 with 20 independent trials. The horizon of the risk oracle is set as $T = 10^4$ to reduce the variance of subgradient. Denote the optimal value of (4) as $J(X^*)$. Since there is an inevitable error in the Lagrangian (around 3%) per iteration, the optimality gap and constraint violation finally converge to 5%, see Fig. 2. The variance of them originates from the primal iteration and the subgradient estimation.

6. Conclusion

In this paper, we have proposed a primal-dual learning framework for the model-free risk-constrained LQR. In particular, we have shown that the Lagrangian function is both locally gradient dominated and Lipschitz, based on which the strong duality is established. Furthermore, we have shown the global convergence of the proposed primal-dual learning algorithm.

This work only considers the gradient descent method in a stochastic form. However, the optimization landscape of natural gradient and Gauss-Newton method for the risk-constrained LQR, even in the model-based setting, is still unclear. We have considered the policy gradient primal-dual method in the model-based setting in Zhao et al. (2021b). Also note that there is only one constraint in our optimization problem. It is also interesting to study the PO for LQR with multiple constraints, which will be our future work.

Acknowledgments

We would like to thank Mr. Kaiqing Zhang from University of Illinois at Urbana-Champaign for his constructive suggestions, and Mr. Jiaqi Zhang from Tsinghua University for his advice on writing. This research was supported by National Natural Science Foundation of China under Grant no. 62033006.

References

Karl J Åström. *Introduction to stochastic control theory*. Courier Corporation, 2012.

Nicole Bäuerle and Ulrich Rieder. More risk-sensitive markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334, 1997.

Dimitri P Bertsekas. *Reinforcement learning and optimal control*. Athena Scientific, 2019.

Vivek Borkar and Rahul Jain. Risk-constrained markov decision processes. *IEEE Transactions on Automatic Control*, 59(9):2574–2579, 2014.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Jingjing Bu, Afshin Mesbahi, Maryam Fazel, and Mehran Mesbahi. LQR through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019.

Margaret P Chapman, Jonathan Lacotte, Aviv Tamar, Donggun Lee, Kevin M Smith, Victoria Cheng, Jaime F Fisac, Susmit Jha, Marco Pavone, and Claire J Tomlin. A risk-sensitive finite-time reachability approach for safety of stochastic dynamic systems. In *American Control Conference*, pages 2958–2963, 2019.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476, 2018.

Luca Furieri, Yang Zheng, and Maryam Kamgarpour. Learning the globally optimal distributed LQ regulator. In *Learning for Dynamics and Control*, pages 287–297, 2020.

B. Gravell, P. Mohajerin Esfahani, and T. H. Summers. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, 2020. ISSN 1558-2523. doi: 10.1109/TAC.2020.3037046.

David Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic control*, 18(2):124–131, 1973.

Joao Paulo Jansch-Porto, Bin Hu, and Geir Dullerud. Convergence guarantees of policy optimization methods for markovian jump linear systems. *arXiv preprint arXiv:2002.04090*, 2020.

Yingying Li, Yujie Tang, Runyu Zhang, and Na Li. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. *arXiv preprint arXiv:1912.09135*, 2019.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2916–2925, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Hesameddin Mohammadi, Mahdi Soltanolkotabi, and Mihailo R Jovanović. On the linear convergence of random search for discrete-time LQR. *IEEE Control Systems Letters*, 5(3):989–994, 2020.

John B Moore, Robert J Elliott, and Subhrakanti Dey. Risk-sensitive generalizations of minimum variance estimation and control. *Journal of Mathematical Systems Estimation and Control*, 7: 123–126, 1997.

Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Xinlei Pan, Daniel Seita, Yang Gao, and John Canny. Risk averse robust adversarial reinforcement learning. In *International Conference on Robotics and Automation*, pages 8522–8528, 2019.

Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. In *Advances in Neural Information Processing Systems*, pages 7555–7565, 2019.

A Prashanth L and Michael Fu. Risk-sensitive reinforcement learning: A constrained optimization viewpoint. *arXiv*, pages arXiv–1810, 2018.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Vincent Roulet, Maryam Fazel, Siddhartha Srinivasa, and Zaid Harchaoui. On the convergence of the iterative linear exponential quadratic gaussian algorithm to stationary points. In *American Control Conference*, pages 132–137, 2020.

Pantelis Sopasakis, Mathijs Schuurmans, and Panagiotis Patrinos. Risk-averse risk-constrained optimal control. In *2019 18th European Control Conference (ECC)*, pages 375–380. IEEE, 2019.

Jason L Speyer, C-H Fan, and Ravi N Banavar. Optimal stochastic estimation with exponential cost criteria. In *Proceedings of the 31st IEEE Conference on Decision and Control*, pages 2293–2299, 1992.

Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.

A. Tsiamis, D. S. Kalogerias, L. F. O. Chamon, A. Ribeiro, and G. J. Pappas. Risk-constrained linear-quadratic regulators. In *59th IEEE Conference on Decision and Control (CDC)*, pages 3040–3047, 2020. doi: 10.1109/CDC42340.2020.9303967.

Min Wen and Ufuk Topcu. Constrained cross-entropy method for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 7450–7460, 2018.

Peter Whittle. Risk-sensitive linear quadratic gaussian control. *Advances in Applied Probability*, pages 764–777, 1981.

Ming Yu, Zhuoran Yang, Mladen Kolar, and Zhaoran Wang. Convergent policy optimization for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 3127–3139, 2019.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pages 11598–11610, 2019.

Kaiqing Zhang, Bin Hu, and Tamer Basar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. In *Learning for Dynamics and Control*, pages 179–190, 2020.

Feiran Zhao, Keyou You, and Tamer Başar. Infinite-horizon risk-constrained linear quadratic regulator with average cost. *arXiv preprint arXiv:2103.15363*, 2021a.

Feiran Zhao, Keyou You, and Tamer Başar. Global convergence of policy gradient primal-dual methods for risk-constrained LQRs. *arXiv preprint arXiv:2104.04901*, 2021b.