Accelerating Model-Free Policy Optimization Using Model-Based Gradient: A Composite Optimization Perspective

Yansong Li, Shuo Han, University of Illinois at Chicago

Overview

- A policy optimization algorithm that combines model-based and model-free methods when only an approximate model is available.
- Formulate an optimal control problem for nonlinear systems as a composite optimization problem.

A motivating example

Suboptimal control for nonlinear system:

$$\begin{aligned} \underset{K}{\text{minimize}} \quad & C(K) = \mathbb{E}_{x_0 \sim \mathcal{D}} \left(\sum_{t=0}^{\infty} x_t^\top Q x_t + u_t^\top R u_t \right) \\ \text{s.t.} \quad & x_{t+1} = Ax_t + Bu_t + h\left(x_t, u_t\right) \\ & u_t = -Kx_t \end{aligned}$$

- The matrix A and matrix B are known.
- The function h is unknown.

Composite optimization view

As a composite optimization problem:

$$\underset{K}{\text{minimize}} \quad C(K) = \hat{C}(K) + r(K)$$

- $\hat{C}$: Cost under the approximate model

$$\begin{aligned} \hat{C}(K) \triangleq \mathbb{E}_{x_0 \sim \mathcal{D}} & \left(\sum_{t=0} x_t^\top Q x_t + u_t^\top R u_t \right) \\ \text{s.t.} \quad & x_{t+1} = Ax_t + Bu_t \\ & u_t = -Kx_t \end{aligned}$$

- $r \triangleq C - \hat{C}$: Residual

The way to access the information needed:

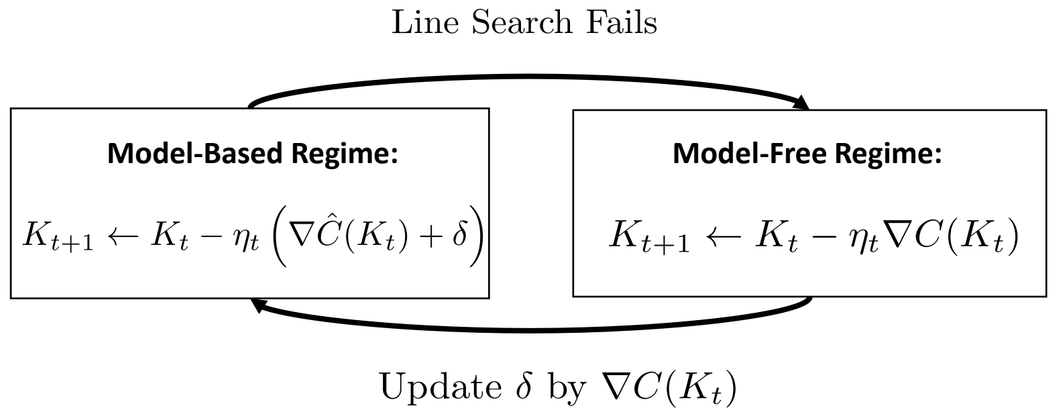
- $\hat{C}, \nabla\hat{C}$: No data collection required
- $C, \nabla C$: Need data collection.

Compensation of gradient

- It is cheap using $\nabla\hat{C}$ as the gradient mapping for policy optimization.
- Using $\nabla\hat{C}$ as the policy gradient may not be accurate enough.
- Add a compensation $\delta \triangleq \nabla C(K) - \nabla\hat{C}(K)$ to make the gradient more accurate near K.
- δ needs to be updated when the optimization variable is far from K.

Gradient compensation algorithm

Idea: Switch between the inexact gradient $\nabla\hat{C}$ and the exact gradient ∇C.



Switching rule:

- From model-based to model-free: Use line search to find out whether $\nabla\hat{C}(K) + \delta$ is a descent direction. Switch to the model-free regime when line search fails.
- From model-free to model-based: Switch to the model-based regime after updating the compensation term δ by the exact gradient.

Theoretical guarantees

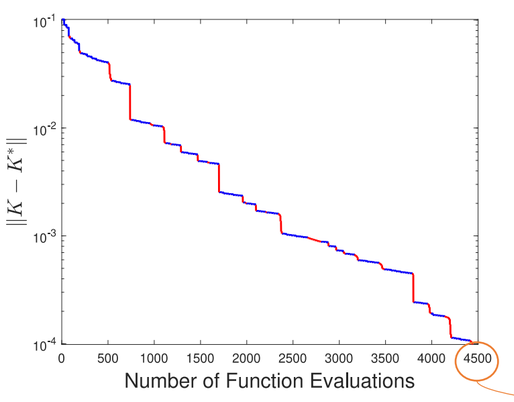
- **Linear convergence:** $C(K_t) - C^\star = \mathcal{O}(\rho^t), 0 < \rho < 1$
- **Lower bound on the number of model-based iterations:** $N \geq \psi(\gamma, \kappa_{\max}, \kappa_{\min}, \eta_{\max}, \eta_{\min}, L_r)$

- The performance is similar to the standard gradient descent.
- Less data collection for reaching the same level of suboptimality.

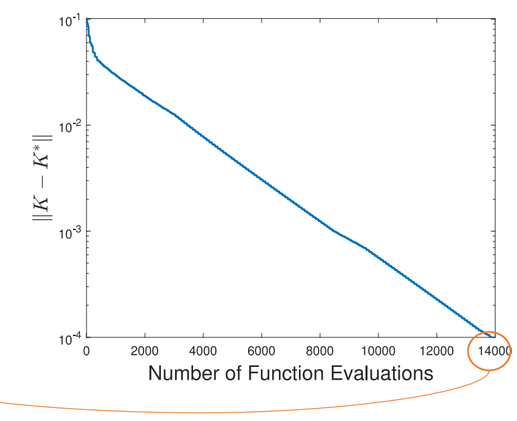
Simulation results

- Blue: model-free regime.
- Red: model-based regime.
- Fewer function evals = Less data collection.

Model-based + Model-free



Purely model-free



The fewer the better