Data-Efficient Reinforcement Learning with Probabilistic Model Predictive Control

Sanket Kamthe
Department of Computing
Imperial College London

Marc Peter Deisenroth
Department of Computing
Imperial College London

1 Appendix

1.1 Lipschitz Continuity

Lemma 1. The moment matching mapping f_{MM} is Lipschitz continuous for controls defined over a compact set $\mathcal{U}$.

Proof: Lipschitz continuity requires that the gradient $\partial f_{MM}/\partial \boldsymbol{u}_t$ is bounded. The gradient is

$$\frac{\partial f_{MM}}{\partial \boldsymbol{u}_t} = \frac{\partial \boldsymbol{z}_{t+1}}{\partial \boldsymbol{u}_t} = \left[\frac{\partial \boldsymbol{\mu}_{t+1}}{\partial \boldsymbol{u}_t}, \frac{\partial \boldsymbol{\Sigma}_{t+1}}{\partial \boldsymbol{u}_t}\right]. \tag{1}$$

The derivatives $\left[\frac{\partial \boldsymbol{\mu}_{t+1}}{\partial \boldsymbol{u}_t}, \frac{\partial \boldsymbol{\Sigma}_{t+1}}{\partial \boldsymbol{u}_t}\right]$ can be computed analytically [1].

We first show that the derivative $\partial \boldsymbol{\mu}_{t+1}/\partial \boldsymbol{u}_t$ is bounded. Defining $\boldsymbol{\beta}_d := (\boldsymbol{K}_d + \sigma_{f_d}^2 \boldsymbol{I})^{-1}\boldsymbol{y}_d$, from [1], we obtain for all state dimensions $d = 1, \ldots, D$

$$\mu_{t+1}^d = \sum\nolimits_{i=1}^N \beta_{d_i} q_{d_i}\,, \tag{2}$$

$$q_{d_i} = \sigma_{f_d}^2 |\boldsymbol{I} + \boldsymbol{L}_d^{-1}\widetilde{\boldsymbol{\Sigma}}_t|^{-\frac{1}{2}} \times \tag{3}$$

$$\exp\big(-\tfrac{1}{2}(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{\mu}}_t)^T(\boldsymbol{L}_d + \widetilde{\boldsymbol{\Sigma}}_t)^{-1}(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{\mu}}_t)\big)\,, \tag{4}$$

where N is the size of the training set of the dynamics GP and $\widetilde{\boldsymbol{x}}_i$ the ith training input. The corresponding gradient w.r.t. $\boldsymbol{u}_t$ is given by the last F elements of

$$\frac{\partial \mu_{t+1}^d}{\partial \widetilde{\boldsymbol{\mu}}_t} = \sum\nolimits_{i=1}^N \beta_{d_i} \frac{\partial q_{d_i}}{\partial \widetilde{\boldsymbol{\mu}}_t} \tag{5}$$

$$= \sum\nolimits_{i=1}^N \beta_{d_i} q_{d_i} (\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{\mu}}_t)^T(\widetilde{\boldsymbol{\Sigma}}_t + \boldsymbol{L}_d)^{-1} \in \mathbb{R}^{1\times(D+F)} \tag{6}$$

Let us examine the individual terms in the sum on the rhs in (6): For a given trained GP $\|\boldsymbol{\beta}_d\| < \infty$ is constant. The definition of q_{d_i} in (4) contains an exponentiated negative quadratic term, which is bounded between $[0, 1]$. Since $\boldsymbol{I} + \boldsymbol{L}_d^{-1}\widetilde{\boldsymbol{\Sigma}}_t$ is positive definite, the inverse determinant is defined and bounded. Finally, $\sigma_{f_d}^2 < \infty$, which makes $q_{d_i} < \infty$. The remaining term in (6) is a vector-matrix product. The matrix is regular and its inverse exists and is bounded (and constant as a function of $\boldsymbol{u}_t$. Since $\boldsymbol{u}_t \in \mathcal{U}$ where $\mathcal{U}$ is compact, we can also conclude that the vector difference in (6) is finite, which overall proves that f_{MM} is (locally) Lipschitz continuous and Lemma 1.

1.2 Sequential Quadratic Programming

We can use SQP for solving non-linear optimization problems (NLP) of the form,

$$\begin{aligned} \min_{u} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & b(\boldsymbol{x}) \geq 0 \\ & \boldsymbol{c}(\boldsymbol{x}) = 0. \end{aligned}$$

The Lagrangian $\mathcal{L}$ associated with the NLP is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\sigma}) = f(\boldsymbol{x}) - \boldsymbol{\sigma}^T b(\boldsymbol{x}) - \boldsymbol{\lambda}^T c(\boldsymbol{x}) \tag{7}$$

where, $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma}$ are Lagrange multipliers. Sequential Quadratic Programming (SQP) forms a quadratic (Taylor) approximation of the objective and linear approximation of constraints at each iteration k

$$\begin{aligned} \min_{d} \quad & f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T \boldsymbol{d} + \tfrac{1}{2}\boldsymbol{d}^T \boldsymbol{\nabla}_{xx}^2 \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\sigma})\boldsymbol{d} \\ \text{s.t.} \quad & b(\boldsymbol{x}_k) + \nabla b(\boldsymbol{x}_k)^T \boldsymbol{d} \geq 0 \\ & c(\boldsymbol{x}_k) + \nabla c(\boldsymbol{x}_k)^T \boldsymbol{d} = 0. \end{aligned} \tag{8}$$

The Lagrange multipliers $\boldsymbol{\lambda}$ associated with the equality constraint are same as the ones defined in the control Hamiltonian $\mathcal{H}$ **??**. The Hessian matrix $\boldsymbol{\nabla}_{xx}^2$ can be computed by exploiting the block diagonal structure introduced by the Hamiltonian [**? ?**].

1.3 Moment Matching Approximation [1]

Following the law of iterated expectations, for target dimensions $a = 1, \dots, D$, we obtain the *predictive mean*

$$\begin{aligned}
\mu_t^a &= \mathbb{E}_{\tilde{\boldsymbol{x}}_{t-1}}[\mathbb{E}_{f_a}[f_a(\tilde{\boldsymbol{x}}_{t-1})|\tilde{\boldsymbol{x}}_{t-1}]] = \mathbb{E}_{\tilde{\boldsymbol{x}}_{t-1}}[m_{f_a}(\tilde{\boldsymbol{x}}_{t-1})] \\
&= \int m_{f_a}(\tilde{\boldsymbol{x}}_{t-1})\mathcal{N}\big(\tilde{\boldsymbol{x}}_{t-1}\,|\,\tilde{\boldsymbol{\mu}}_{t-1}, \tilde{\boldsymbol{\Sigma}}_{t-1}\big)d\tilde{\boldsymbol{x}}_{t-1} \\
&= \boldsymbol{\beta}_a^T \boldsymbol{q}_a\,, \qquad (9)
\end{aligned}$$

$$\boldsymbol{\beta}_a = (\boldsymbol{K}_a + \sigma_{w_a}^2)^{-1}\boldsymbol{y}_a \qquad (10)$$

with $\boldsymbol{q}_a = [q_{a_1}, \dots, q_{a_n}]^T$. The entries of $\boldsymbol{q}_a \in \mathbb{R}^n$ are computed using standard results from multiplying and integrating over Gaussians and are given by

$$\begin{aligned}
q_{a_i} &= \int k_a(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_{t-1})\mathcal{N}\big(\tilde{\boldsymbol{x}}_{t-1}\,|\,\tilde{\boldsymbol{\mu}}_{t-1}, \tilde{\boldsymbol{\Sigma}}_{t-1}\big)d\tilde{\boldsymbol{x}}_{t-1} \qquad (11) \\
&= \sigma_{f_a}^2 |\tilde{\boldsymbol{\Sigma}}_{t-1}\boldsymbol{\Lambda}_a^{-1} + \boldsymbol{I}|^{-\frac{1}{2}} \exp\big(-\tfrac{1}{2}\boldsymbol{\nu}_i^T(\tilde{\boldsymbol{\Sigma}}_{t-1} + \boldsymbol{\Lambda}_a)^{-1}\boldsymbol{\nu}_i\big)\,,
\end{aligned}$$

where we define

$$\boldsymbol{\nu}_i := (\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{\mu}}_{t-1}) \qquad (12)$$

is the difference between the training input $\tilde{\boldsymbol{x}}_i$ and the mean of the test input distribution $p(\boldsymbol{x}_{t-1}, \boldsymbol{u}_{t-1})$.

Computing the *predictive covariance matrix* $\boldsymbol{\Sigma}_t \in \mathbb{R}^{D\times D}$ requires us to distinguish between diagonal elements and off-diagonal elements: Using the law of total (co-)variance, we obtain for target dimensions $a, b = 1, \dots, D$

$$\sigma_{aa}^2 = \mathbb{E}_{\tilde{\boldsymbol{x}}_{t-1}}\big[\mathrm{var}_f[x_t^a|\tilde{\boldsymbol{x}}_{t-1}]\big] + \mathbb{E}_{f,\tilde{\boldsymbol{x}}_{t-1}}[(\boldsymbol{x}_t^a)^2] - (\mu_t^a)^2\,, \qquad (13)$$

$$\sigma_{ab}^2 = \mathbb{E}_{f,\tilde{\boldsymbol{x}}_{t-1}}[x_t^a x_t^b] - \mu_t^a \mu_t^b\,, \quad a \neq b\,, \qquad (14)$$

respectively, where μ_t^a is known from (9). The off-diagonal terms do not contain the additional term $\mathbb{E}_{\tilde{\boldsymbol{x}}_{t-1}}[\mathrm{cov}_f[x_t^a, x_t^b|\tilde{\boldsymbol{x}}_{t-1}]]$ because of the conditional independence assumption of the GP models. Different target dimensions do not covary for given $\tilde{\boldsymbol{x}}_{t-1}$.

We start the computation of the covariance matrix with the terms that are common to both the diagonal and the off-diagonal entries: With $p(\tilde{\boldsymbol{x}}_{t-1}) = \mathcal{N}\big(\tilde{\boldsymbol{x}}_{t-1}\,|\,\tilde{\boldsymbol{\mu}}_{t-1}, \tilde{\boldsymbol{\Sigma}}_{t-1}\big)$ and the law of iterated expectations, we obtain

$$\begin{aligned}
\mathbb{E}_{f,\tilde{\boldsymbol{x}}_{t-1}}[x_t^a, x_t^b] &= \mathbb{E}_{\tilde{\boldsymbol{x}}_{t-1}}\big[\mathbb{E}_f[x_t^a|\tilde{\boldsymbol{x}}_{t-1}]\,\mathbb{E}_f[x_t^b|\tilde{\boldsymbol{x}}_{t-1}]\big] \\
&= \int m_f^a(\tilde{\boldsymbol{x}}_{t-1}) m_f^b(\tilde{\boldsymbol{x}}_{t-1}) p(\tilde{\boldsymbol{x}}_{t-1}) d\tilde{\boldsymbol{x}}_{t-1} \qquad (15)
\end{aligned}$$

because of the conditional independence of x_t^a and x_t^b given $\tilde{\boldsymbol{x}}_{t-1}$. Using the definition of the mean function, we obtain

$$\mathbb{E}_{f,\tilde{\boldsymbol{x}}_{t-1}}[x_t^a x_t^b] = \boldsymbol{\beta}_a^T \boldsymbol{Q} \boldsymbol{\beta}_b\,, \qquad (16)$$

$$\boldsymbol{Q} := \int k_a(\tilde{\boldsymbol{x}}_{t-1}, \boldsymbol{X})^T\, k_b(\tilde{\boldsymbol{x}}_{t-1}, \boldsymbol{X}) p(\tilde{\boldsymbol{x}}_{t-1}) d\tilde{\boldsymbol{x}}_{t-1}\,. \qquad (17)$$

Using standard results from Gaussian multiplications and integration, we obtain the entries Q_{ij} of $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$

$$Q_{ij} = \frac{k_a(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{\mu}}_{t-1}) k_b(\tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{\mu}}_{t-1})}{\sqrt{|\boldsymbol{R}|}} \exp\big(\tfrac{1}{2}\boldsymbol{z}_{ij}^T \boldsymbol{T}^{-1} \boldsymbol{z}_{ij}\big) \qquad (18)$$

where we define

$$\begin{aligned}
&\boldsymbol{R} := \tilde{\boldsymbol{\Sigma}}_{t-1}(\boldsymbol{\Lambda}_a^{-1} + \boldsymbol{\Lambda}_b^{-1}) + \boldsymbol{I}\,, \quad \boldsymbol{T} := \boldsymbol{\Lambda}_a^{-1} + \boldsymbol{\Lambda}_b^{-1} + \tilde{\boldsymbol{\Sigma}}_{t-1}^{-1}\,, \\
&\boldsymbol{z}_{ij} := \boldsymbol{\Lambda}_a^{-1}\boldsymbol{\nu}_i + \boldsymbol{\Lambda}_b^{-1}\boldsymbol{\nu}_j\,,
\end{aligned}$$

with $\boldsymbol{\nu}_i$ taken from (12). Hence, the off-diagonal entries of $\boldsymbol{\Sigma}_t$ are fully determined by (9)–(12), (14), (16), (17), and (18).

References

[1] M. P. Deisenroth, D. Fox, and C. E. Rasmussen. Gaussian Processes for Data-Efficient Learning in Robotics and Control. *Transactions on Pattern Analysis and Machine Intelligence*, 37(2):408–23, 2015.