Model-Free Trajectory Optimization for Reinforcement Learning (Supplementary Material)

Riad Akrour[1] AKROUR@IAS.TU-DARMSTADT.DE
Abbas Abdolmaleki[3] ABBAS.A@UA.PT
Hany Abdulsamad[2] ABDULSAMAD@IAS.TU-DARMSTADT.DE
Gerhard Neumann[1] NEUMANN@IAS.TU-DARMSTADT.DE
1: CLAS, 2: IAS, TU Darmstadt, Darmstadt, Germany
3: IEETA, University of Aveiro, Aveiro, Portugal

1. Dual Function

Recall the quadratic form of the Q-Function $\tilde{Q}_t(s,a)$ in the action a and state s

$$\tilde{Q}_t(s,a) = \frac{1}{2}a^T Q_{aa} a + a^T Q_{as} s + a^T q_a + q(s). \quad (1)$$

The new policy $\pi'_t(a|s)$ solution of the constrained maximization problem is again of linear-Gaussian form and given by

$$\pi'_t(a|s) = \mathcal{N}(a|FLs + Ff, F(\eta^* + \omega^*)),$$

such that the gain matrix, bias and covariance matrix of π'_t are function of matrices F and L and vector f where

$$F = (\eta^*\Sigma_t^{-1} - Q_{aa})^{-1}, \quad L = \eta^*\Sigma_t^{-1}K_t + Q_{as},$$
$$f = \eta^*\Sigma_t^{-1}k_t + q_a.$$

With η^* and ω^* the optimal Lagrange multipliers related to the KL and entropy constraints, obtained by minimizing the dual function

$$g_t(\eta,\omega) = \eta\epsilon - \omega\beta + (\eta+\omega)\int \tilde{\rho}_t(s)$$
$$\log\left(\int \pi(a|s)^{\eta/(\eta+\omega)} \exp\left(\tilde{Q}_t(s,a)/(\eta+\omega)\right)\right) \mathrm{d}\mathbf{s}.$$

From the quadratic form of $\tilde{Q}_t(s,a)$ and by additionally assuming that the state distribution is approximated by $\tilde{\rho}_t(\mathbf{s}) = \mathcal{N}(\mathbf{s}|\mu_\mathbf{s}, \Sigma_\mathbf{s})$, the dual function simplifies to

$$g_t(\eta,\omega) = \eta\epsilon - \omega\beta + \mu_\mathbf{s}^T M \mu_\mathbf{s} + \mathrm{tr}(\Sigma_s M) + \mu_\mathbf{s}^T m + m_0.$$

Where M, m and m_0 are defined by

$$M = \frac{1}{2}\left(L^T F L - \eta K_t^T \Sigma_t^{-1} K_t\right), \; m = L^T F f - \eta K_t^T \Sigma_t^{-1} k_t,$$
$$m_0 = \frac{1}{2}(f^T F f - \eta k_t^T \Sigma_t^{-1} k_t - \eta \log|2\pi\Sigma_t|$$
$$+ (\eta+\omega)\log|2\pi(\eta+\omega)F|).$$

The convex dual function g_t can be efficiently minimized by gradient descent and the policy update is performed upon the computation of η^* and ω^*. The gradient w.r.t. η and ω is given by[1]

$$\frac{\partial g_t(\eta,\omega)}{\partial \eta} = \mathrm{cst} + \mathrm{lin} + \mathrm{quad}$$
$$\mathrm{cst} = \epsilon - \frac{1}{2}\left(k_t - Ff\right)^T \Sigma_t^{-1}\left(k_t - Ff\right) - \frac{1}{2}[\log|2\pi\Sigma_t|$$
$$- \log|2\pi(\eta+\omega)F| + (\eta+\omega)\mathrm{tr}(\Sigma_t^{-1}F) - d_a].$$
$$\mathrm{lin} = \left((K_t - FL)\mu_s\right)^T \Sigma_t^{-1}(Ff - k_t).$$
$$\mathrm{quad} = \mu_s^T (K_t + FL)^T \Sigma_t^{-1}(K_t + FL)\mu_s$$
$$+ \mathrm{tr}(\Sigma_s (K_t + FL)^T \Sigma_t^{-1}(K_t + FL))$$
$$\frac{\partial g_t(\eta,\omega)}{\partial \omega} = -\beta + \frac{1}{2}(d_a + \log|2\pi(\eta+\omega)F|).$$

[1] cst, lin, quad, F, L and f all depend on η and ω. We dropped the dependency from the notations for compactness. d_a is the dimensionality of the action.