Stochastic Optimal Control with Linearized Dynamics

Stochastisch optimale Regelung mit linearisierten Modellen
Master-Thesis von Hany Abdulsamad
Tag der Einreichung:

1. Gutachten: Prof. Gerhard Neumann
2. Gutachten: Prof. Jan Peters
3. Gutachten: Prof. Ulrich Konigorski

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Stochastic Optimal Control with Linearized Dynamics
Stochastisch optimale Regelung mit linearisierten Modellen

Vorgelegte Master-Thesis von Hany Abdulsamad

1. Gutachten: Prof. Gerhard Neumann
2. Gutachten: Prof. Jan Peters
3. Gutachten: Prof. Ulrich Konigorski

Tag der Einreichung:

Erklärung zur Master-Thesis

Hiermit versichere ich, die vorliegende Master-Thesis ohne Hilfe Dritter nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 1. März 2016

_____

(Hany Abdulsamad)

Abstract

Policy Search is a powerful class for learning optimal control policies of complex systems. By allowing a very broad description of a task, they are suitable for solving challenging robotic applications. Although model-free Policy Search approaches require the least amount of knowledge about the environment, they often suffer from the disadvantage of having to draw a large number of samples from the system. Therefore, in cases where it is feasible to reconstruct the system dynamics, it is advantageous to include as much prior knowledge about the learning setting as possible. In this work we consider this insight as motivation for exploring model-based Policy Search algorithms.

A recent approach, Guided Policy Search, has combined the strengths of a powerful model-based trajectory optimization technique, Stochastic Optimal Control, with Relative Entropy Policy Search to learn policies of complicated tasks like bipedal walking. By alternating between linearizing the system dynamics and optimizing local policies, it follows the main scheme of iterative methods like Differential Dynamic Programing and Iterative Linear Quadratic Gaussian. The novelty is, however, the introduction of a relative entropy bound on the trajectory distribution in order to preserve the locality of the linearization and improve the robustness of convergence.

In this work we will examine and reformulate Guided Policy Search in order to highlight its main contribution. We will show that the bound on the trajectory distribution is equivalent to a bound on the change of the policy. Moreover, we will motivate and propose a new constraint that would strictly bound the state distribution between iterations, and further ensure the validity of the linearization, while allowing us to preform larger steps on the policy updates. In addition, we will introduce a bound on the entropy of the policy, which allows to control the ability the controller to explore the action space and prevent premature convergence.

We will present results and compare all variants of the proposed algorithms on highly non-linear systems, such as a swing-up task on a torque- and angle-constrained double and quad pendulum. As supplementary material, we will provide a full and detailed mathematical derivation of our methods.

Acknowledgments

I would especially like to thank Prof. Gerhard Neumann, Head of the Computational Learning and Autonomous Systems (CLAS) group, for introducing me the to ideas covered in this thesis and for his patient supervision, open-door policy and the countless hours of his time, that have resulted in many informative discussions for me. I also thank M.Sc. Oleg Arenz, who co-supervised me and always took the time to share his insights and experience.
I owe a debt of gratitude to Prof. Jan Peters, Head of the Intelligent Autonomous Systems (IAS) group, and all IAS and CLAS members, who constantly engage and motivate their students. During my time at IAS and CLAS, I've had the pleasure of working closely with Alexandros Paraschos and Simone Parisi. I am deeply grateful for their help and support.
Finally, I thank Prof. Ulrich Konigorski and M.Sc. Zhongyi Gong from the Institut für Regelungstechnik und Mechatronik (RTM) at the Electrical Engineering Department, who agreed to co-supervise me and showed interest in my work.

Contents

Figures and Tables

List of Figures

List of Algorithms

Abbreviations

List of Abbreviations

Notation	Description
DDP	Differential Dynamic Programming
DP	Dynamic Programming
ESAPS	Entropy State-Action Bound Policy Search
GPS	Guided Policy Search
iLQG	Iterative Linear Quadratic Gaussian
KLD	Kullback-Leibler Divergence
LQG	Linear Quadratic Gaussian
MDP	Markov Decision Process
REPS	Relative Entropy Policy Search
RL	Reinforcement Learning
SAPS	State-Action Bound Policy Search
SOC	Stochastic Optimal Control

1 Introduction

Recent advancements in the field of robotics have resulted in a considerable growth in robotic applications and tasks. The introduction of new platforms such as high dimensional humanoids and high velocity/torque manipulators gives the promise of making headway in solving major tasks like bipedal locomotion and grasping. However, with this progress comes a sharp rise in the complexity and non-linearity of the dynamical systems in question, which poses several challenges from a control and planning point of view.
Trajectory Optimization methods set out to solve Optimal Control problems under general time, energy and spacial constraints. Stochastic Optimal Control (SOC) with linearized dynamics, in particular, is a powerful approach to obtain optimal control laws for non-linear systems. Fundamental work on Stochastic Optimal Control includes Differential Dynamic Programming (DDP) (Mayne, 1966) (Jacobson and Mayne, 1970), Iterative Linear Quadratic Gaussian (Todorov and Li, 2005) (Tassa et al., 2012), Approximate Inference Control (AICO) (Toussaint, 2009a) (Rawlik et al., 2010) and Robust Policy Updates for Stochastic Optimal Control (RSOC) (Rueckert et al., 2014).

1.1 Locality and Validity of Linearization

Stochastic Optimal Control algorithms implement an iterative scheme using linearized dynamics to locally optimize the current trajectory. A key element in the stability of such procedure is a mechanism to control the step size of the update of the controller in a principled manner. The linearized dynamics are only accurate in the vicinity of the linearization point. Solutions that stray too far from this linearization point have to be avoided, as they may cause oscillations or even instabilities. A recent approach, Guided Policy Search (GPS) (Levine and Koltun, 2014) (Levine and Abbeel, 2014), addresses this issue by introducing a relative entropy bound on the update of the trajectory distribution between iterations.
In the course of this thesis, we will evaluate the aforementioned approach and extend it by proposing new bounds. Therefore, we will argue for a an explicit bound on the state distribution, instead of the trajectory distribution, as it is crucial to the linearization. This bound should provide a stronger guarantee for the validity of the linearization between iterations, hence, allowing more aggressive updates of the policy. Moreover, we will suggest an entropy constraint that would allow us to control the exploration rate of the policy, thus preventing premature convergence issues that have been observed in GPS.

1.2 Reinforcement Learning vs. Motion Planning

At this point it is necessary to draw an important distinction between two categories of Optimal Control methods. Namely, Motion Planning algorithms and Reinforcement Learning methods (RL) (Sutton and Barto, 1998). In Motion Planning, a complete model of the environment, a mapping from system dynamics to reward, is available and can exploited to optimize the expected return of the trajectory. State-of-the-art algorithms in this area are CHOMP (Ratliff et al., 2009), STOMP (Kalakrishnan et al., 2011) and TRJOPT (Schulman et al., 2013). Whereas, in a Reinforcement Learning setting such models are either learned online, as in PILCO (Deisenroth and Rasmmussen, 2011) and GPS (Levine and Abbeel, 2014), or completely circumvented, as in REPS (Peters et al., 2010), in the process of finding the optimal policy. The focus of this work will be devoted to model-based Reinforcement Learning algorithms with state-of-the-art GPS as the center piece.

1.3 Preliminaries

1.3.1 Markov Decision Process

A Markov Decision Process (MDP) is a mathematical model for sequential decision making in a Motion Planning or Reinforcement Learning setting. An MDP sequence is time discrete and can stretch over a finite or infinite time horizon. MPDs derive their name from the Markov property, which stipulates that in a Markovian system, the distribution over the future state $\mathbf{s}'$ of the environment depends only on the current state $\mathbf{s}$ and next action $\mathbf{a}$. The transition to a future state $\mathbf{s}'$ is governed by the stochastic system dynamics $\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})$. Because decision making has to be rationalized by some quantifiable measure, MDPs also specify a reward function $R_t(\mathbf{s},\mathbf{a})$, that rates the quality of state-action pairs $(\mathbf{s},\mathbf{a})$. Since we are interested in time-constrained trajectories, we will, for the remainder of this thesis, always consider finite-horizon MDPs. Furthermore, we assume the dynamics to be of linear-Gaussian nature with time-variant quadratic reward functions,

$$\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) = \mathcal{N}(\mathbf{s}'|\mathbf{A}_t\mathbf{s} + \mathbf{b}_t\mathbf{a} + \mathbf{c}_t, \mathbf{\Sigma}_{\mathbf{s}'}),$$
$$R_t(\mathbf{s},\mathbf{a}) = \mathbf{s}^T\mathbf{M}_t\mathbf{s} + \mathbf{a}^T\mathbf{H}_t\mathbf{a}.$$

We dub these conditions on the dynamics and reward function as the *LQG assumptions*.

1.3.2 Stochastic Optimal Control

Stochastic Optimal Control's objective is to find a control policy $\pi_t(\mathbf{a}|\mathbf{s})$ that maximizes a reward measure $R_t(\mathbf{s},\mathbf{a})$ along a trajectory. General solutions are rare and restricted to a small class of systems. Aside from discrete systems, the most important exception are systems with linear-Gaussian dynamics. It has be shown, that the optimal controller for a system adhering to the LQG assumptions can be computed in closed-form by applying Dynamic Programming (DP) (Bellman, 1957). DP introduces the concept of state value function $V_t(\mathbf{s})$ and state-action value function $Q_t(\mathbf{s},\mathbf{a})$. $V_t(\mathbf{s})$ is defined as the expected reward-to-go under a certain policy $\pi_t(\mathbf{a}|\mathbf{s})$ starting from a state $\mathbf{s}$, whereas $Q_t(\mathbf{s},\mathbf{a})$ is the expected reward-to-go after executing an action $\mathbf{a}$ and subsequently following a policy $\pi_t(\mathbf{a}|\mathbf{s})$. DP implements a backward induction algorithm, that recursively computes both value functions starting from the end time point T, where $V_t(\mathbf{s}) = R_T(\mathbf{s})$. The Bellman equations, here given in continuous form, define the relation between $V_t(\mathbf{s})$ and $Q_t(\mathbf{s},\mathbf{a})$ as follows

$$Q_t(\mathbf{s},\mathbf{a}) = R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) V_{t+1}(\mathbf{s}') d\mathbf{s}',$$
$$V_t(\mathbf{s}) = \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) Q_t(\mathbf{s},\mathbf{a}) d\mathbf{a}.$$

Following these definitions, determining the optimal policy $\pi_t(\mathbf{a}|\mathbf{s})$ is reduced to finding a function that maximizes the state-action value function $Q_t(\mathbf{s},\mathbf{a})$ at each time step of the trajectory

$$\pi_t^*(\mathbf{a}|\mathbf{s}) = \underset{\mathbf{a}}{\operatorname{argmax}}\ Q_t(\mathbf{s},\mathbf{a}).$$

1.3.3 Information Theoretic Bounds

In this section we introduce the theoretical background to the entropy and relative entropy bounds that we will encounter in the course of this thesis.

Differential Entropy

The entropy of a distribution p over a random variable $\mathbf{s}$ is a measure of the variance of that distribution and , thus, also a measure of the average amount of information embedded in it. Relevant to this work is the entropy of probability distribution over continuous random variables, also called differential entropy. In that case the entropy $\mathcal{H}$ of a distribution $p(\mathbf{s})$ is defined as

$$\mathcal{H} = -\int_{\mathbf{s}} p(\mathbf{s}) \log p(\mathbf{s}) d\mathbf{s}.$$

We will use the entropy as a measure of the stochasticity of the control policy, which in turn would allow us to judge and control its capability in exploring the state-action space.

Relative Entropy

The relative entropy, also known as the Kullback-Leibler divergence, $D_{\mathrm{KL}}(p||q)$ between two probability distribution $p(\mathbf{s})$ and $q(\mathbf{s})$, is a non negative measure of information loss when $p(\mathbf{s})$ is used to approximate $q(\mathbf{s})$ and is defined as

$$D_{\mathrm{KL}}(p(\mathbf{s})||q(\mathbf{s})) = \int_{\mathbf{s}} p(\mathbf{s}) \log \frac{p(\mathbf{s})}{q(\mathbf{s})} d\mathbf{s}.$$

The relative entropy of two conditionals $p(\mathbf{a}|\mathbf{s})$ and $q(\mathbf{a}|\mathbf{s})$, or their expected KL divergence, is defined analogously

$$D_{\mathrm{KL}}(p(\mathbf{a}|\mathbf{s})||q(\mathbf{a}|\mathbf{s})) = \int_{\mathbf{s}} p(\mathbf{s}) \int_{\mathbf{a}} p(\mathbf{a}|\mathbf{s}) \log \frac{p(\mathbf{a}|\mathbf{s})}{q(\mathbf{a}|\mathbf{s})} d\mathbf{a} d\mathbf{s}.$$

The measure of relative entropy is central to the ideas of this work. We will use it to define bounds over different distributions with the aim of limiting their change between iterations, thus ensuring the stability of the algorithm.

2 Related Work

2.1 Iterative Local Methods for Non-Linear Systems

In our introduction of Stochastic Optimal Control, we have discussed the limitations of its framework, in which tractable solution are exclusive to discrete and linear systems. These restrictions may seem to completely eliminate the possibility of applying Stochastic Optimal Control to non-linear systems. However, it is possible to apply SOC in an iterative scheme with the following structure,

- Starting from an initial state, apply an initial control sequence to the non-linear dynamics to obtain a finite state sequence.
- Linearize the dynamics around each point of the retrieved trajectory and quadratize the reward function.
- Formulate and solve a local LQG problem with respect to state-action deviations to get a new locally optimal policy.
- Execute the new policy on the non-linear system to obtain a new trajectory.

In the coming sections we will introduce Differential Dynamic Programming (DDP) and Iterative Linear Quadratic Gaussian (iLQG), two algorithms that follow this iterative cycle.

2.1.1 Differential Dynamic Programming

Differential Dynamic Programing was introduced in (Mayne, 1966) (Jacobson and Mayne, 1970). It follows the main scheme described above. Starting from the objective of maximizing the expected reward along the trajectory $\tau = \{\mathbf{s}_1, \mathbf{a}_1, ..., \mathbf{s}_T, \mathbf{a}_T\}$

$$J(\mathbf{s}_1, \mathbf{A}) = \sum_{t=0}^{T-1} R_t(\mathbf{s}, \mathbf{a}) + R_T(\mathbf{s}).$$

Maximizing J is equivalent to finding the optimal state value function $V_t(\mathbf{s})$ that maximizes the reward-to-go for each state $\mathbf{s}$ and time step t

$$V_t(\mathbf{s}) \equiv \max_{\mathbf{A}} \; J(\mathbf{s}_1, \mathbf{A}).$$

By setting $V_T(\mathbf{s}) = R_T(\mathbf{s})$ and applying Dynamic Programing, we can reduce the maximization over the whole control sequence to a sequence of maximizations over a single control

$$\begin{aligned} V_t(\mathbf{s}) &= \max_{\mathbf{a}} \; [R_t(\mathbf{a}|\mathbf{s}) + \sum_{s'} \mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a}) V_{t+1}(\mathbf{s}')] \\ &= \max_{\mathbf{a}} \; [R_t(\mathbf{s}, \mathbf{a}) + V_{t+1}(\mathcal{P}_t(\mathbf{s}, \mathbf{a}))], \end{aligned}$$

where $\mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ are the linearized dynamics at each time step of the current trajectory.
By moving to a notation that describes the perturbations around each state-action pair $(\mathbf{s}, \mathbf{a})$, we are able to reformulate the argument of the maximization problem as

$$Q_t(\delta\mathbf{s}, \delta\mathbf{a}) = R_t(\mathbf{s} + \delta\mathbf{s}, \mathbf{a} + \delta\mathbf{a}) - R_t(\mathbf{s}, \mathbf{a}) + V_{t+1}(\mathcal{P}_t(\mathbf{s} + \delta\mathbf{s}, \mathbf{a} + \delta\mathbf{a})) - V_{t+1}(\mathcal{P}_t(\mathbf{s}, \mathbf{a})).$$

After expanding to second order, the Jacobians and Hessians of the dynamics can be determined. The subscripts denote the derivatives with respect to state $\mathbf{s}$ and action $\mathbf{a}$

$$\begin{aligned}
Q_{s,t} &= R_{s,t} + \mathcal{P}_{s,t}^T V_{s,t+1}, \\
Q_{a,t} &= R_{a,t} + \mathcal{P}_{s,t}^T V_{s,t+1}, \\
Q_{ss,t} &= R_{ss,t} + \mathcal{P}_{s,t}^T V_{ss,t+1} \mathcal{P}_{s,t} + V_{s,t+1} \mathcal{P}_{ss,t}, \\
Q_{aa,t} &= R_{aa,t} + \mathcal{P}_{a,t}^T V_{ss,t+1} \mathcal{P}_{s,t} + V_{s,t+1} \mathcal{P}_{aa,t}, \\
Q_{as,t} &= R_{as,t} + \mathcal{P}_{a,t}^T V_{ss,t+1} \mathcal{P}_{s,t} + V_{s,t+1} \mathcal{P}_{as,t}.
\end{aligned}$$

For the optimal local control sequence $\delta\mathbf{a}^*$, we maximize the function $Q_t(\delta\mathbf{s}, \delta\mathbf{a})$ and get a policy $\pi_t(\delta\mathbf{a}|\delta\mathbf{s})$ that resembles a linear controller

$$\begin{aligned}
\delta a^* &= \underset{\delta\mathbf{a}}{\operatorname{argmax}}\ Q_t(\delta\mathbf{s}, \delta\mathbf{a}) \\
&= -Q_{aa,t}^{-1}(Q_{a,t} + Q_{as,t}\delta\mathbf{s}) = \mathbf{k}_t + \mathbf{K}_t \delta\mathbf{s}.
\end{aligned}$$

Substituting the policy $\pi_t(\delta a|\delta s)$ into $Q_t(\delta s, \delta a)$ leads to a quadratic value function

$$\begin{aligned}
\Delta V_t &= -\frac{1}{2} Q_{a,t} Q_{aa,t}^{-1} Q_{a,t}, \\
V_{s,t} &= Q_{s,t} - Q_{a,t} Q_{aa,t}^{-1} Q_{as,t}, \\
V_{ss,t} &= Q_{ss,t} - Q_{sa,t} Q_{aa,t}^{-1} Q_{as,t}.
\end{aligned}$$

Applying the new policy to the non-linear system to get a new trajectory completes one cycle of DDP. The main problem with this formulation, is that it greedily exploits the local dynamics and produces policies that be can be arbitrarily different between iterations, undermining the locality and validity of the linearization. In most cases this leads to divergence or oscillations. The authors addressed this issue by introducing a regularization to the action-reward Hessian

$$\tilde{Q}_{aa,t} = Q_{aa,t} + \mu I.$$

which is equivalent to adding a reward for stying close to the last policy and not straying. This regularization is helpful under the assumption that small changes in the policy imply small changes in the state space and, thus, preserve the validity of the linearization.

2.1.2 Iterative Linear Quadratic Gaussian

Iterative Linear Quadratic Gaussian sets out to correct the shortcomings of DDP by offering several improvements on the regularization and a line search algorithm.
In (Tassa et al., 2012) the authors present a new regularization on the state reward, that would force the new trajectory to be stay close to the last one and results in modified state and action Hessian matrices

$$\begin{aligned}
Q_{aa,t} &= R_{aa,t} + \mathcal{P}_{a,t}^T (V_{ss,t+1} + \mu I) \mathcal{P}_{s,t} + V_{s,t+1} \mathcal{P}_{aa,t}, \\
Q_{as,t} &= R_{as,t} + \mathcal{P}_{a,t}^T (V_{ss,t+1} + \mu I) \mathcal{P}_{s,t} + V_{s,t+1} \mathcal{P}_{as,t}.
\end{aligned}$$

which also results in a new quadratic value function that takes into account the new regularization

$$\begin{aligned}
\Delta V_t &= -\frac{1}{2} k_t^T Q_{aa,t}^{-1} k_t + k_t^T Q_{a,t}, \\
V_{s,t} &= Q_{s,t} - K_t^T Q_{aa,t}^{-1} k_t + K_t^T Q_{a,t} + Q_{as,t}^T k_t, \\
V_{ss,t} &= Q_{ss,t} - K_t^T Q_{aa,t}^{-1} K_t + K_t^T Q_{as,t} + Q_{as,t}^T K_t.
\end{aligned}$$

Furthermore, with aim of bounding the trajectory change even more, and preventing highly non-linear systems from diverging, the authors also introduce a scaler α to the policy parameters

$$\hat{\mathbf{a}}_t = \mathbf{a}_t + \alpha \mathbf{k}_t + \mathbf{K}_t \delta \mathbf{s}.$$

This scaler is optimized by a line search method based on the expected improvement of the reward.

2.2 Relative Entropy Policy Search

Relative Entropy Policy Search (REPS) is a model-free Reinforcement Learning approach (Peters et al., 2010). The novelty of REPS is the introduction of a new type of bounds, that can be imposed between updates. The bound resembles a relative entropy measure, or a Kullback-Leibler divergence, on the state-action distribution. In a Reinforcement Learning environment this constraint is crucial to convergence, as it preserves the experience contained in the last policy and last state distribution, that has developed over multiple iterations and constrains the algorithm from jumping arbitrarily to new unexplored regions of the state space. The optimization problem under REPS is given as

$$\underset{\pi(\mathbf{a}|\mathbf{s})\mu(\mathbf{s})}{\operatorname{argmax}} \quad \sum_{\mathbf{s},\mathbf{a}} R(\mathbf{s},\mathbf{a})\mu(\mathbf{s})\pi(\mathbf{a}|\mathbf{s}), \tag{2.5a}$$

$$\text{s.t.} \quad \sum_{\mathbf{s}'} \mu(\mathbf{s}')\Phi(\mathbf{s}') = \sum_{\mathbf{s},\mathbf{a}} \mu(\mathbf{s})\pi(\mathbf{a}|\mathbf{s})\mathcal{P}(\mathbf{s}'|\mathbf{s},\mathbf{a})\Phi(\mathbf{s}'), \tag{2.5b}$$

$$\sum_{\mathbf{s},\mathbf{a}} \mu(\mathbf{s})\pi(\mathbf{a}|\mathbf{s}) \log \frac{\mu(\mathbf{s})\pi(\mathbf{a}|\mathbf{s})}{q(\mathbf{s},\mathbf{a})} \leq \epsilon, \tag{2.5c}$$

$$\sum_{\mathbf{s},\mathbf{a}} \mu(\mathbf{s})\pi(\mathbf{a}|\mathbf{s}) = 1. \tag{2.5d}$$

where the objective 2.5a maximizes the reward with respect to the joint distribution over the state $\mu(\mathbf{s})$ and conditional action $\pi(\mathbf{a}|\mathbf{s})$ and Equation 2.5c ensures that the state-action distribution $\mu(\mathbf{s})\pi(\mathbf{a}|\mathbf{s})$ stays close the old one $q(\mathbf{s},\mathbf{a})$. Under this formulation the optimal policy is a normalized exponential

$$\pi(\mathbf{a}|\mathbf{s}) \propto \exp\Big[\frac{1}{\eta}\Big(\eta \log q(\mathbf{s},\mathbf{a}) + R(\mathbf{s},\mathbf{a}) + \sum_{s'} \mathcal{P}(\mathbf{s}'|\mathbf{s},\mathbf{a})\boldsymbol{\theta}^T\Phi(\mathbf{s}') - \boldsymbol{\theta}^T\Phi(\mathbf{s})\Big)\Big].$$

The parameters $\boldsymbol{\theta}$ and η are the Lagrangian multipliers corresponding to Equations 2.5b and 2.5c and can be optimized by a gradient descent method.

3 Guided Policy Search

Guided Policy Search (GPS) was developed over multiple publications (Levine and Koltun, 2013, 2014; Levine and Abbeel, 2014). The idea as presented in (Levine and Koltun, 2013) is to introduce a set of guiding trajectories generated under locally optimal Differential Dynamic Programming (DDP) and weighted by Importance Sampling (IS) to exploit regions in the state space with high rewards and to "guide" and speed-up convergence. In (Levine and Koltun, 2014) the algorithm was further modified to ensure the usefulness of the guiding trajectories. This improvement is done by alternating between optimizing a set of trajectories for high rewards (Trajectory Optimization), while constraining the policy to match the actions in each trajectory thus constraining the policy update from straying into unexplored regions of the state space (Policy Search). While the contributions in (Levine and Koltun, 2013, 2014) are interesting in their own standing, this thesis will concentrate on the core of Guided Policy Search in its latest and most refined version as presented in (Levine and Abbeel, 2014), which imposes a KL-divergence bound on the trajectory distribution between iterations.

3.1 Optimization Problem

In their work the authors adopt a trajectory-based notation (Levine and Abbeel, 2014)

$$\underset{p(\tau)}{\operatorname{argmax}} \quad \int_\tau R(\tau)p(\tau)d\tau, \tag{3.1a}$$

$$\text{s.t.} \quad \int_\tau p(\tau)\log\frac{p(\tau)}{q(\tau)}d\tau \leq \epsilon, \tag{3.1b}$$

$$p(\tau) = p(\mathbf{s}_1)\prod_{t=1}^{T-1}\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})\pi_t(\mathbf{a}|\mathbf{s}). \tag{3.1c}$$

where the objective 3.1a maximizes the reward $R(\tau)$ along a trajectory $\tau = \{\mathbf{s}_1, a_1, ..., s_T, a_T\}$, while Equation 3.1b provides the KL-bound on the current and last trajectory distribution $p(\tau)$ and $q(\tau)$. Equation 3.1c propagates the state s along the trajectory under the local linear dynamics $\mathcal{P}(\mathbf{s}'|\mathbf{s},\mathbf{a})$ and the Gaussian policy $\pi(\mathbf{a}|\mathbf{s})$ starting from the state distribution $p(\mathbf{s}_1)$.
We find this notation to be somewhat unclear, therefore we transform the problem to its step-based equivalent. Thus, we are able to show that the KL-divergence bound imposed on the trajectory distribution $p(\tau)$ can be, in fact, simplified to a bound set on the policy $\pi(\mathbf{a}|\mathbf{s})$. For the purpose of clarity we preform this transformation explicitly. By substituting the dynamics constraint 3.1c into the KL-bound 3.1b and replacing trajectories τ with state-action pairs $(\mathbf{s},\mathbf{a})$ we can rewrite the integral in 3.1b

$$D_{\mathrm{KL}}(p(\tau)||q(\tau)) = \int_\tau p(\tau)\log\frac{p(\mathbf{s}_1)\prod_{t=1}^{T-1}\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})\pi_t(\mathbf{a}|\mathbf{s})}{p(\mathbf{s}_1)\prod_{t=1}^{T-1}\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})q_t(\mathbf{a}|\mathbf{s})}d\tau \tag{3.2a}$$

$$= \sum_{t=1}^{T-1}\int_{\mathbf{s}}\int_{\mathbf{a}} p_t(\mathbf{s},\mathbf{a})\log\frac{\pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{a}|\mathbf{s})}d\mathbf{a}d\mathbf{s} \tag{3.2b}$$

$$= \sum_{t=1}^{T-1}\int_{\mathbf{s}} p_t(\mathbf{s})\int_{\mathbf{a}}\pi_t(\mathbf{a}|\mathbf{s})\log\frac{\pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{a}|\mathbf{s})}d\mathbf{a}d\mathbf{s}. \tag{3.2c}$$

From Equation 3.2c, it is clear that KL-bound on the trajectory distribution is equivalent to an expected bound on the policy at each time step. At this point we are able to rewrite the whole problem with our new state-action-pairs notation

$$\underset{\pi_t(\mathbf{a}|\mathbf{s})}{\operatorname{argmax}} \quad \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s}, \tag{3.3a}$$

$$\text{s.t.} \quad \forall \mathbf{s}', \forall t > 1 \quad \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_{t-1}(\mathbf{s}) \pi_{t-1}(\mathbf{a}|\mathbf{s}) \mathcal{P}_{t-1}(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} = \mu_t(\mathbf{s}'), \tag{3.3b}$$

$$\forall t < T \quad \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{a}|\mathbf{s})} d\mathbf{a} d\mathbf{s} \leq \epsilon, \tag{3.3c}$$

$$\forall t < T, \forall \mathbf{s} \quad \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} = 1, \tag{3.3d}$$

$$\forall \mathbf{s} \quad \mu_1(\mathbf{s}) = p_1(\mathbf{s}). \tag{3.3e}$$

where the reward $R(\mathbf{s},\mathbf{a})$ is to be maximized with respect to the state-action distribution, given by the policy $\pi_t(\mathbf{a}|\mathbf{s})$ and its induced state distribution $\mu_t(\mathbf{s})$, while under the system dynamics constraint 3.3b, that propagates the initial state distribution through time and is referred to as a *forward pass*. Equation 3.3c is a constraint on the expected KL-bound on the policy for each time step, whereas Equation 3.3d ensures the policy is a distribution, and Equation 3.3e specifies the initial state distribution $\mu_1(\mathbf{s})$.

3.2 Dual Problem

For the purpose of this thesis we produce a complete derivation of the closed-form solution of Guided Policy Search under the assumptions of linear dynamics, Gaussian noise and quadratic rewards, see Appendix A. We start by applying the method of Lagrangian multipliers to formulate the so called primal problem, which introduces a new Lagrangian multiplier per constraint and time step. The state-dependent Lagrangian multipliers $V_t(\mathbf{s})$ are associated with the dynamics constraint 3.3b and will later resemble the state value function, while α_t are associated with the KL-bound given in Equation 3.3c. By solving for the optimal policy $\pi_t(\mathbf{a}|\mathbf{s})$ we obtain a normalized exponential of the state-action value function $Q_t(\mathbf{s},\mathbf{a})$

$$\pi_t(\mathbf{a}|\mathbf{s}) \propto \exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t q_t(\mathbf{a}|\mathbf{s}) + R_t(\mathbf{s},\mathbf{a}) + \int_{s'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}'\Big)\Big]. \tag{3.4}$$

By plugging Equation 3.4 into the primal problem we arrive at the Lagrangian dual $L(\mu_t, V_t, \alpha_t)$

$$\begin{aligned} L(\mu_t, V_t, \alpha_t) = &\int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} - \sum_{t=1}^{T} \int_{\mathbf{s}'} V_t(\mathbf{s}') \mu_t(\mathbf{s}') d\mathbf{s}' + \sum_{t=1}^{T-1} \alpha_t \epsilon \\ &+ \sum_{t=1}^{T-1} \int_{\mathbf{s}} \alpha_t \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} q_t(\mathbf{a}|\mathbf{s}) \exp\Big[\frac{1}{\alpha_t}\Big(R_t(\mathbf{s},\mathbf{a}) + \int_{s'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}'\Big)\Big] d\mathbf{a} d\mathbf{s}. \end{aligned} \tag{3.5}$$

The dual L is a function of the state distribution $\mu_t(\mathbf{s})$ and the Lagrangian multiplier $V_t(\mathbf{s})$ and α_t. By exploiting the duality of this optimization, we are able to maximize the primal problem by minimizing the dual function (Boyd and Vandenberghe, 2009). Therefore, we take the partial derivatives of L and apply dual descent in their respective directions,

$$\frac{\partial L}{\partial \mu_t} = \begin{cases} R_T(\mathbf{s}) - V_T(\mathbf{s}) & , t = T \\ V_t(\mathbf{s}) - \alpha_t \log \int_{\mathbf{a}} \exp\Big[\frac{\alpha_t \log q_t(\mathbf{a}|\mathbf{s}) + R_t(\mathbf{s},\mathbf{a}) + \int_{s'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})}{\alpha_t} \Big] & , t < T \end{cases}, \tag{3.6a}$$

$$\frac{\partial L}{\partial V_t} = \begin{cases} p_1(\mathbf{s}) - \mu_1(\mathbf{s}) & , t = 1 \\ \mu_t(\mathbf{s}) - \int_{\hat{\mathbf{s}}} \int_{\mathbf{a}} \pi_{t-1}(\mathbf{a}|\hat{\mathbf{s}}) \mu_{t-1}(\hat{\mathbf{s}}) \mathcal{P}_{t-1}(\mathbf{s}|\hat{\mathbf{s}}, a) d\mathbf{a} d\hat{\mathbf{s}} & , t > 1 \end{cases}, \tag{3.6b}$$

$$\frac{\partial L}{\partial \alpha_t} = \epsilon - \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{a}|\mathbf{s})} d\mathbf{a} d\mathbf{s}. \tag{3.6c}$$

Setting the derivatives in Equations 3.6a and 3.6b to zero delivers two optimality conditions for the state value function $V_t(\mathbf{s})$ and the state distribution $\mu_t(\mathbf{s})$, that correspond to a *backward pass* (backward propagation of future reward) and *forward pass* (forward propagation of the state distribution) respectively

$$V_t(\mathbf{s}) = \begin{cases} R_T(\mathbf{s}) & , t = T \\ \alpha_t \log \int_{\mathbf{a}} \exp\Big[\frac{\alpha_t \log q_t(\mathbf{a}|\mathbf{s}) + R_t(\mathbf{s},\mathbf{a}) + \int_{s'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})}{\alpha_t} \Big] & , t < T \end{cases}, \tag{3.7a}$$

$$\mu_t(\mathbf{s}) = \begin{cases} p_1(\mathbf{s}) & , t = 1 \\ \int_{\hat{s}} \int_{\mathbf{a}} \pi_{t-1}(\mathbf{a}|\hat{\mathbf{s}}) \mu_{t-1}(\hat{\mathbf{s}}) \mathcal{P}_{t-1}(\mathbf{s}|\hat{\mathbf{s}}, \mathbf{a}) d\mathbf{a} d\hat{\mathbf{s}} & , t > 1 \end{cases}. \tag{3.7b}$$

Under the LQG assumptions, these passes can be computed in closed form, whereas α_t have to be optimized by gradient descent. Considering the partial derivative of L with respect to α_t, it is worth noting, that at the optimal point, the KL-constraint given in Equation 3.3c is met exactly at the bound ϵ, because the gradient in Equation 3.6c becomes zero. Finally, by plugging Equations 3.7a and 3.7b into Equation 3.8 the dual simplifies to

$$L(\mu_t, V_t, \alpha_t) = \int_{\mathbf{s}} V_1(\mathbf{s}) \mu_1(\mathbf{s}) d\mathbf{s} + \sum_{t=1}^{T-1} \alpha_t \epsilon. \tag{3.8}$$

3.3 Policy Dependent Reward

An interesting insight into the state value function $V_t(\mathbf{s})$, which stands for the expected reward-to-go and is defined in Equation 3.7a, is the emergence of new terms that augment the immediate reward to include a policy-related term $q_t(\mathbf{a}|\mathbf{s})$ in addition to the standard state-action reward as provided by a time-variant function $R_t(\mathbf{s},\mathbf{a})$ in a setting analog to DDP and iLQG

$$r_t(\mathbf{s},\mathbf{a}) = R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{a}|\mathbf{s}). \tag{3.9}$$

Under linear-Gaussian dynamics $\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) = \mathcal{N}(\mathbf{s}'|\mathbf{A}_t\mathbf{s} + \mathbf{b}_t\mathbf{a} + \mathbf{c}_t, \mathbf{\Sigma}_{s'})$ and quadratic reward $R_t(\mathbf{s},\mathbf{a}) = (\mathbf{z}-\mathbf{s})^T \mathbf{M}_t (\mathbf{z}-\mathbf{s}) + \mathbf{a}^T \mathbf{H}_t \mathbf{a}$, we show that the overall reward $r_t(\mathbf{s},\mathbf{a})$ is also quadratic

$$r_t(\mathbf{s},\mathbf{a}) = \mathbf{s}^T \mathbf{R}_{ss.t} \mathbf{s} + \mathbf{a}^T \mathbf{R}_{aa.t} \mathbf{a} + \mathbf{a}^T \mathbf{R}_{sa.t}^T \mathbf{s} + \mathbf{s}^T \mathbf{R}_{sa.t} \mathbf{a} + \mathbf{s}^T \mathbf{r}_{s.t} + \mathbf{a}^T \mathbf{r}_{s.t} + r_{0,t}. \tag{3.10}$$

$$\mathbf{R}_{ss,t} = \mathbf{M}_t - \frac{\alpha_t}{2}(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{K}_t^q, \tag{3.11a}$$

$$\mathbf{R}_{aa,t} = \mathbf{H}_t - \frac{\alpha_t}{2}(\boldsymbol{\Sigma}_{a,t}^q)^{-1}, \tag{3.11b}$$

$$\mathbf{R}_{sa,t} = \frac{\alpha_t}{2}(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}, \tag{3.11c}$$

$$\mathbf{r}_{s,t} = -\alpha_t(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q - 2\mathbf{M}_t\mathbf{z}, \tag{3.11d}$$

$$\mathbf{r}_{a,t} = \alpha_t(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q, \tag{3.11e}$$

$$r_{0,t} = \mathbf{z}^T\mathbf{M}_t\mathbf{z} - \alpha_t \log\sqrt{\left|2\pi\boldsymbol{\Sigma}_{a,t}^q\right|} - \frac{\alpha_t}{2}(\mathbf{k}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q. \tag{3.11f}$$

A quadratic reward function $r_t(\mathbf{s},\mathbf{a})$, by definition, forces a quadratic state value function $V_t(\mathbf{s})$

$$V_t(\mathbf{s}) = \mathbf{s}^T\mathbf{V}_t\mathbf{s} + \mathbf{s}^T\mathbf{v}_t + \nu_t. \tag{3.12}$$

In turn and by considering Equation 3.4, a quadratic state value function gives rise to a time-variant linear-Gaussian optimal policy

$$\pi_t(\mathbf{a}|\mathbf{s}) = \mathcal{N}(\mathbf{a}|\mathbf{k}_t^\pi + \mathbf{K}_t^\pi\mathbf{s}, \boldsymbol{\Sigma}_{a,t}^\pi). \tag{3.13}$$

3.4 Implementation

In this section we describe the structure of our version of Guided Policy Search as we have implemented it. For the purpose of brevity, we do not consider the process of linearization. Generally, linearization is done by sampling full trajectories from the non-linear system under the current policy and fitting linear-Gaussian dynamics at each time step. The implementation as discussed here, focuses on the optimization step, and presupposes the existence of the linearized dynamics.

Based on the derivation of the dual function from the previous sections, we have transformed the problem into a convex minimization problem over three parameters per time step $V_t(\mathbf{s})$, $\mu_t(\mathbf{s})$ and α_t. However, since Equations 3.7a and 3.7b deliver closed-form solutions to the optimal state value function $V_t(\mathbf{s})$ and state distribution $\mu_t(\mathbf{s})$ as functions of α_t, the problem is reduced to a minimization of the dual with respect to α_t and can be iteratively solved by a gradient descent scheme. In this case, the whole procedure can be seen as a batch-coordinate-descent optimization with respect to $V_t(\mathbf{s})$, $\mu_t(\mathbf{s})$ and α_t. Algorithm 1 shows the step by step sequence of the minimization.

Although a gradient descent implementation is a straight forward procedure, it is recommended to use more sophisticated optimizers as provided by *Mathworks MATLAB* or *Non-Linear Optimization Library (NLopt)* (Johnson, 2016), because they provide advanced heuristics of modulating the step size along the gradient and a numerical estimate of the second degree derivative (Hessian), generally leading to faster convergence and less computation cost.

For reasons related to computational stability and efficiency, all our algorithms will be implemented in the framework of the *Armadillo Linear Algebra Library* (Sanderson, 2010).

input : T ; /* time horizon */
$\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})$; /* linearized dynamics */
$\mu_1(\mathbf{s})$; /* initial state distribution */
$q_t(\mathbf{a}|\mathbf{s})$; /* last policy */
$\mathbf{M}_t, \mathbf{H}_t, \mathbf{z}_t$; /* reward matrices and goal state */

output: $\pi_t(\mathbf{a}|\mathbf{s})$; /* optimal policy */
$V_t(\mathbf{s})$; /* optimal state value function */
$\mu_t(\mathbf{s})$; /* state distribution under optimal policy */
α_t ; /* optimal Lagrangian parameters α_t */

initialize α_t ; /* initial guess of α_t */

/* minimizing the dual by gradient descent */
while $L(\mu_t, V_t, \alpha_t)$ not at minimum **do**

/* compute augmented reward function using Equation 3.10 */
$r_t(\mathbf{s},\mathbf{a}) \leftarrow \text{overall\_reward}(\mathbf{M}_t, \mathbf{H}_t, \mathbf{z}_t, q_t(\mathbf{a}|\mathbf{s}), \alpha_t)$;

/* compute value function and policy using Equations 3.7a and 3.4 */
$[V_t(\mathbf{s}), \pi_t(\mathbf{a}|\mathbf{s})] \leftarrow \text{backward\_pass}(r_t(\mathbf{s},\mathbf{a}), \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}), \alpha_t)$;

/* compute the state distribution using Equation 3.7b */
$\mu_t(\mathbf{s}) \leftarrow \text{forward\_pass}(\mu_1(\mathbf{s}), \pi_t(\mathbf{a}|\mathbf{s}), \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}))$;

/* update Lagrange dual value with Equation 3.8 */
$L(\mu_t, V_t, \alpha_t) \leftarrow \text{update\_dual}(V_1(\mathbf{s}), \mu_1(\mathbf{s}), \alpha_t, \epsilon)$;

/* compute Lagrange dual gradient with respect to α_t using Equation 3.6c */
$\frac{\partial L}{\partial \alpha_t} \leftarrow \text{dual\_alpha\_gradient}(\mu_t(\mathbf{s}), \pi_t(\mathbf{a}|\mathbf{s}), q_t(\mathbf{a}|\mathbf{s}), \epsilon)$;

/* update α_t along the gradient with step λ_t */
$\alpha_t = \alpha_t - \lambda_t \frac{\partial L}{\partial \alpha_t}$;

Algorithm 1: Guided Policy Search in Pseudo-Code

4 State-Action Bound Policy Search

At the beginning of this thesis we introduced the general scheme of applying Stochastic Optimal Control to non-linear systems. The main challenge is the absence of a theoretical guarantee on the improvement of the induced trajectory after each iteration. This shortcoming is due to the restricted validity of the local dynamics to a small region around the linearization point. A greedy exploitation of the linearized dynamics may lead to policies that force the non-linear system into regions of the state space that are "far away" from what is expected under the linearized model making the optimization step under the model meaningless. Therefore, it is crucial to maintain a bound on the state distribution between iterations in order to ensure the validity of the locally optimized controller.
Iterative Linear Quadratic Gaussian (iLQG) tries to solve this problem by introducing a scalar to the policy parameters which is optimized by a backtracking line-search scheme that increases or reduces the step size based on the improvement in the expected reward. Guided Policy Search follows a similar logic; by introducing a relative entropy bound on the change of the stochastic policy, the induced state distribution becomes implicitly bounded. However, for highly dynamical systems this condition would require imposing very small steps on the policy, which might dramatically slow down convergence and cost a considerable extra amount of samples on the real system.
In this chapter we aim to address this issue. We propose the introduction of an explicit relative entropy bound on the state-action distribution and set out to show that such a bound would allow taking larger steps in the policy space while preventing the state distribution from diverging, thus reducing the number of needed iterations and overall samples.

4.1 Optimization Problem

We take a similar formulation to Guided Policy Search, but replace the KL-bound on the policy distribution by a bound on the state-action distribution

$$\underset{\pi_t(\mathbf{a}|\mathbf{s})}{\operatorname{argmax}} \quad \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s}, \tag{4.1a}$$

$$\text{s.t.} \quad \forall \mathbf{s}', \forall t > 1 \quad \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_{t-1}(\mathbf{s}) \pi_{t-1}(\mathbf{a}|\mathbf{s}) \mathcal{P}_{t-1}(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} = \mu_t(\mathbf{s}'), \tag{4.1b}$$

$$\forall t < T \quad \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s},\mathbf{a})} d\mathbf{a} d\mathbf{s} \leq \epsilon, \tag{4.1c}$$

$$\forall \mathbf{s}, \forall t < T \quad \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} = 1, \tag{4.1d}$$

$$\forall \mathbf{s}, t = 1 \quad \mu_1(\mathbf{s}) = p_1(\mathbf{s}). \tag{4.1e}$$

The objective in 4.1a seeks to maximize the reward under the final state-action distribution $p_t(\mathbf{s},\mathbf{a}) = \mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})$, while 4.1b keeps the state distribution $\mu_t(\mathbf{s})$ under the constraint of the linearized system dynamics. Our novelty state-action bound is introduced in 4.1c, with $q_t(\mathbf{s},\mathbf{a})$ representing the state-action distribution of the last linearization. The remaining constraints 4.1d and 4.1e ensure that the policy is a distribution and specify the initial state distribution respectively.

4.2 Dual Problem

As in our derivation of Guided Policy Search in Chapter 3, we apply the method of Lagrangian multipliers to formulate the primal problem with one Lagrangian multiplier per constraint and time step. The full derivation under the LQG assumptions is listed in Appendix B. In this case, the optimal policy is also a normalized exponential of the state-action value function $Q_t(\mathbf{s}, \mathbf{a})$

$$\pi_t(\mathbf{a}|\mathbf{s}) \propto \exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t \log q_t(\mathbf{s}, \mathbf{a}) + R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a})d\mathbf{s}'\Big)\Big]. \tag{4.2}$$

We obtain the Lagrangian dual function $L(\mu_t, V_t, \alpha_t)$ by substituting the optimal policy Equation 4.2 into the primal problem

$$\begin{aligned} L = &\int_{\mathbf{s}} \mu_T(\mathbf{s})R_T(\mathbf{s})d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s})p_1(\mathbf{s})d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}')\mu_T(\mathbf{s}')d\mathbf{s}' \\ &- \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_t(\mathbf{s}')\mu_t(\mathbf{s}')d\mathbf{s}' + \sum_{t=1}^{T-1} \alpha_t \epsilon - \sum_{t=1}^{T-1} \int_{\mathbf{s}} \alpha_t \mu_t(\mathbf{s}) \log \mu_t(\mathbf{s})d\mathbf{s} \\ &+ \sum_{t=1}^{T-1} \int_{\mathbf{s}} \alpha_t \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t \log q_t(\mathbf{s}, \mathbf{a}) + R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a})d\mathbf{s}'\Big)\Big]d\mathbf{a}d\mathbf{s}. \end{aligned} \tag{4.3}$$

According to the principle of duality, minimizing the dual function is equivalent to maximizing the primal problem (Boyd and Vandenberghe, 2009). Therefore, we minimize L by taking its partial derivatives

$$\frac{\partial L}{\partial \mu_t} = \begin{cases} R_T(\mathbf{s}) - V_T(\mathbf{s}) & , t = T \\ V_t(\mathbf{s}) - \alpha_t \log \int_{\mathbf{a}} \exp\Big[\frac{\alpha_t \log q_t(\mathbf{s}, \mathbf{a}) - \alpha_t \log \mu_t(\mathbf{s}) - \alpha_t + R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a})}{\alpha_t}\Big] & , t < T \end{cases}, \tag{4.4a}$$

$$\frac{\partial L}{\partial V_t} = \begin{cases} p_1(\mathbf{s}) - \mu_1(\mathbf{s}) & , t = 1 \\ \mu_t(\mathbf{s}) - \int_{\hat{\mathbf{s}}} \int_{\mathbf{a}} \pi_{t-1}(\mathbf{a}|\hat{\mathbf{s}})\mu_{t-1}(\hat{\mathbf{s}})\mathcal{P}_{t-1}(\mathbf{s}|\hat{\mathbf{s}}, a)d\mathbf{a}d\hat{\mathbf{s}} & , t > 1 \end{cases}, \tag{4.4b}$$

$$\frac{\partial L}{\partial \alpha_t} = \epsilon - \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s}, \mathbf{a})} d\mathbf{a}d\mathbf{s}. \tag{4.4c}$$

At the optimal point of L the partial derivatives are equal to zero, which can be seen as optimality conditions for the state value function $V_t(\mathbf{s})$ and the state distribution $\mu_t(\mathbf{s})$

$$V_t(\mathbf{s}) = \begin{cases} R_T(\mathbf{s}) & , t = T \\ \alpha_t \log \int_{\mathbf{a}} \exp\Big[\frac{\alpha_t \log q_t(\mathbf{s}, \mathbf{a}) - \alpha_t \log \mu_t(\mathbf{s}) - \alpha_t + R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a})}{\alpha_t}\Big] & , t < T \end{cases}, \tag{4.5a}$$

$$\mu_t(\mathbf{s}) = \begin{cases} p_1(\mathbf{s}) & , t = 1 \\ \int_{\hat{\mathbf{s}}} \int_{\mathbf{a}} \pi_{t-1}(\mathbf{a}|\hat{\mathbf{s}})\mu_{t-1}(\hat{\mathbf{s}})\mathcal{P}_{t-1}(\mathbf{s}|\hat{\mathbf{s}}, a)d\mathbf{a}d\hat{\mathbf{s}} & , t > 1 \end{cases}. \tag{4.5b}$$

Analog to Guided Policy Search in Chapter 3, the optimality conditions resemble a *backward pass* and a *forward pass* that can be computed in closed-form in an LQG environment. Furthermore, the KL-constraint 4.1c is being met exactly at the bound ϵ due to Equation 4.4c becoming equal to zero at the optimal point. Also, using Equations 4.5a and 4.5b, we can further simplify the Lagrange dual $L(\mu_t, V_t, \alpha_t)$

$$L(\mu_t, V_t, \alpha_t) = \int_{\mathbf{s}} V_1(\mathbf{s})\mu_1(\mathbf{s})d\mathbf{s} + \sum_{t=1}^{T-1} \alpha_t(\epsilon + 1). \tag{4.6}$$

4.3 State-Action Dependent Reward

The introduction of the state-action constraint 4.1c results in an augmented reward function. The new terms not only account for distance to the last policy $q_t(\mathbf{a}|\mathbf{s})$, but also weight the distance between $\mu_t(\mathbf{s})$, the current state distribution, and $q_t(\mathbf{s})$, the state distribution under the last policy around which the system was linearized

$$\begin{aligned} r_t(\mathbf{s},\mathbf{a}) &= R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) - \alpha_t \log \mu_t(\mathbf{s}) - \alpha_t \\ &= R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{a}|\mathbf{s}) + \alpha_t \log q_t(\mathbf{s}) - \alpha_t \log \mu_t(\mathbf{s}) - \alpha_t. \end{aligned} \tag{4.7}$$

By substituting linear-Gaussians dynamics $\mathcal{P}(\mathbf{s}'|\mathbf{s},\mathbf{a}) = \mathcal{N}(\mathbf{s}|\boldsymbol{\tau}^q_{s,t}, \boldsymbol{\Sigma}^q_{s,t})$, a Gaussian state-action distribution $q_t(\mathbf{s},\mathbf{a}) = \mathcal{N}(\mathbf{s},\mathbf{a}|\boldsymbol{\tau}^q_{s,a,t}, \boldsymbol{\Sigma}^q_{s,a,t})$ and a quadratic reward function $R_t(\mathbf{s},\mathbf{a}) = (\mathbf{z}-\mathbf{s})^T M_t(\mathbf{z}-) + \mathbf{a}^T H_t \mathbf{a}$, the overall reward $r_t(\mathbf{s},\mathbf{a})$ becomes also quadratic

$$r_t(\mathbf{s},\mathbf{a}) = \mathbf{s}^T \mathbf{R}_{ss,t}\mathbf{s} + \mathbf{a}^T \mathbf{R}_{aa,t}\mathbf{a} + \mathbf{a}^T \mathbf{R}^T_{sa,t}\mathbf{s} + \mathbf{s}^T \mathbf{R}_{sa,t}\mathbf{a} + \mathbf{s}^T \mathbf{r}_{s,t} + \mathbf{a}^T \mathbf{r}_{a,t} + r_{0,t}, \tag{4.8a}$$

$$\mathbf{R}_{ss,t} = \mathbf{M}_t - \frac{\alpha_t}{2}(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{K}^q_t - \frac{\alpha_t}{2}(\boldsymbol{\Sigma}^q_{s,t})^{-1} + \frac{\alpha_t}{2}(\boldsymbol{\Sigma}^p_{s,t})^{-1}, \tag{4.8b}$$

$$\mathbf{R}_{aa,t} = \mathbf{H}_t - \frac{\alpha_t}{2}(\boldsymbol{\Sigma}^q_{a,t})^{-1}, \tag{4.8c}$$

$$\mathbf{R}_{sa,t} = \frac{\alpha_t}{2}(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}, \tag{4.8d}$$

$$\mathbf{r}_{s,t} = -\alpha_t(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t + \alpha_t(\boldsymbol{\Sigma}^q_{s,t})^{-1}\boldsymbol{\tau}^q_{s,t} - \alpha_t(\boldsymbol{\Sigma}^p_{s,t})^{-1}\boldsymbol{\tau}^p_{s,t} - 2\mathbf{M}_t\mathbf{z}, \tag{4.8e}$$

$$\mathbf{r}_{a,t} = \alpha_t(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t, \tag{4.8f}$$

$$r_{0,t} = \mathbf{z}^T\mathbf{M}_t\mathbf{z} - \frac{\alpha_t}{2}\log\left|2\pi\boldsymbol{\Sigma}^q_{a,t}\right| - \frac{\alpha_t}{2}(\mathbf{k}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t \tag{4.8g}$$

$$- \frac{\alpha_t}{2}\log|2\pi\boldsymbol{\Sigma}^q_{s,t}| - \frac{\alpha_t}{2}(\boldsymbol{\tau}^q_{s,t})^T(\boldsymbol{\Sigma}^q_{s,t})^{-1}\boldsymbol{\tau}^q_{s,t} - \alpha_t \tag{4.8h}$$

$$+ \frac{\alpha_t}{2}\Big(\log|2\pi\boldsymbol{\Sigma}^p_{s,t}| + (\boldsymbol{\tau}^p_{s,t})^T(\boldsymbol{\Sigma}^p_{s,t})^{-1}\boldsymbol{\tau}^p_{s,t}\Big). \tag{4.8i}$$

4.4 Implementation

In this section we present the implementation of State-Action Bound Policy Search (SAPS). We ignore the linearization step and focus on the convex minimization problem of the dual $L(\mu_t, V_t, \alpha_t)$ as presented in the previous sections.

4.4.1 Circular Dependency of $V_t(s)$ and $\mu_t(s)$

The equations of the *backward pass* 4.5a and *forward pass* 4.5b introduce a new algorithmic challenge that did not occur under Guided Policy Search. The emergence of new state-distribution-dependent terms in the augmented reward function $r_t(\mathbf{s},\mathbf{a})$ of the state value function $V_t(\mathbf{s})$, generates a circular dependency between $V_t(\mathbf{s})$ and the state distribution $\mu_t(\mathbf{s})$. This relation becomes clear when we recognize that the state distribution $\mu_t(\mathbf{s})$ is a function of the policy $\pi_t(\mathbf{a}|\mathbf{s})$, Equation 4.5b, and that $\pi_t(\mathbf{a}|\mathbf{s})$ is in its self a function of the state value function $V_t(\mathbf{s})$, Equation 4.2.

4.4.2 Block Descent over $V_t(s)$ and $\mu_t(s)$

At this point we propose a new approach to calculate the state value function $V_t(\mathbf{s})$ and state distribution $\mu_t(\mathbf{s})$. The Equations 4.5a and 4.5b still offer optimality conditions and can be used iteratively in a

block-descent scheme on the dual $L(\mu_t, V_t, \alpha_t)$. Starting with an initial and broad guess of the state distribution $p_t(\mathbf{s})$, we iteratively apply the *backward pass*, to compute $V_t(\mathbf{s})$ and $\pi(\mathbf{a}|\mathbf{s})$, and *forward pass*, to compute $\mu_t(\mathbf{s})$, and update $p_t(\mathbf{s})$ by interpolating in the direction of $\mu_t(\mathbf{s})$ until both distributions match. Algorithm 2 provides a detailed view of this procedure.

input : T ; /* time horizon */
$\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})$; /* linearized dynamics */
$\mu_1(\mathbf{s})$; /* initial state distribution */
$q_t(\mathbf{a}|\mathbf{s})$; /* last policy */
$q_t(\mathbf{s})$; /* last state distribution */
α_t ; /* current Lagrangian parameters α_t */
$\mathbf{M}_t, \mathbf{H}_t, \mathbf{z}_t$; /* reward matrices and goal state */

output: $\pi_t(\mathbf{a}|\mathbf{s})$; /* policy under current α_t */
$V_t(\mathbf{s})$; /* state value function under current α_t */
$\mu_t(\mathbf{s})$; /* state distribution under current α_t */

initialize $p_t(\mathbf{s})$; /* initial guess of state distribution */
$L(\mu_t, V_t, \alpha_t)$; /* initial dual value */
γ_t ; /* interpolation step size */

/* minimizing the dual with respect to $V_t(\mathbf{s})$ and $\mu_t(\mathbf{s})$ */
while $p_t(\mathbf{s}) \neq \mu_t(\mathbf{s})$ **do**
  /* compute augmented reward function using Equation 4.7 */
  $r_t(\mathbf{s},\mathbf{a}) \leftarrow$ overall_reward$(\mathbf{M}_t, \mathbf{H}_t, \mathbf{z}_t, q_t(\mathbf{a}|\mathbf{s}), q_t(\mathbf{s}), p_t(\mathbf{s}), \alpha_t)$;
  /* compute value function and policy using Equations 4.5a and 4.2 */
  $[V_t(\mathbf{s}), \pi_t(\mathbf{a}|\mathbf{s})] \leftarrow$ backward_pass$(r_t(\mathbf{s},\mathbf{a}), \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}), \alpha_t)$;
  /* compute the state distribution using Equation 4.5b */
  $\mu_t(\mathbf{s}) \leftarrow$ forward_pass$(\mu_1(\mathbf{s}), \pi_t(\mathbf{a}|\mathbf{s}), \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}))$;
  /* check KL-divergence between $p_t(\mathbf{s})$ and $\mu_t(\mathbf{s})$ */
  if $D_{KL}(p_t(\mathbf{s}), \mu_t(\mathbf{s})) <$ threshold **then**
    break;
  /* interpolate $p_t(\mathbf{s})$ in the direction of $\mu_t(\mathbf{s})$ with step size γ_t */
  $\tilde{p}_t(\mathbf{s}) \leftarrow$ interpolate_distribution$(p_t(\mathbf{s}), \mu_t(\mathbf{s}))$;
  /* update Lagrange dual value with Equation 4.6 */
  $\tilde{L}(\mu_t, V_t, \alpha_t) \leftarrow$ update_dual$(V_1(\mathbf{s}), \tilde{p}_1(\mathbf{s}), \alpha_t, \epsilon)$;
  /* check if the dual reached a lower value */
  if $\tilde{L} < L$ **then**
    $L = \tilde{L}$;
    $p_t(\mathbf{s}) = \tilde{p}_t(\mathbf{s})$;
  else
    $\gamma_t = 0.5 \cdot \gamma_t$;

Algorithm 2: State-Action Policy Search: Dual Block Descent over $V_t(\mathbf{s})$ and $\mu_t(\mathbf{s})$ in Pseudo-Code

4.4.3 Gradient Descent over α_t

input : T ; /* time horizon */
$\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})$; /* linearized dynamics */
$\mu_1(\mathbf{s})$; /* initial state distribution */
$q_t(\mathbf{a}|\mathbf{s})$; /* last policy */
$q_t(\mathbf{s})$; /* last state distribution */
$\mathbf{M}_t, \mathbf{H}_t, \mathbf{z}_t$; /* reward matrices and goal state */

output: $\pi_t(\mathbf{a}|\mathbf{s})$; /* optimal policy */
$V_t(\mathbf{s})$; /* optimal state value function */
$\mu_t(\mathbf{s})$; /* optimal state distribution */

initialize α_t ; /* initial guess of α_t */

/* minimizing the dual by gradient descent */
while $L(\mu_t, V_t, \alpha_t)$ not at minimum **do**
  /* do block-descent to compute $V_t(\mathbf{s})$ and $\mu_t(\mathbf{s})$ */
  $[V_t(\mathbf{s}), \pi_t(\mathbf{a}|\mathbf{s}), \mu_t(\mathbf{s})] \leftarrow \text{block\_descent}(\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}), q_t(\mathbf{a}|\mathbf{s}), q_t(\mathbf{s}), p_t(\mathbf{s}), \mathbf{M}_t, \mathbf{H}_t, \mathbf{z}_t, \alpha_t)$;
  /* update Lagrange dual value with Equation 4.6 */
  $L(\mu_t, V_t, \alpha_t) \leftarrow \text{update\_dual}(V_1(\mathbf{s}), \mu_1(\mathbf{s}), \alpha_t, \epsilon)$;
  /* compute Lagrange dual gradient with respect to α_t using Equation 4.4c */
  $\frac{\partial L}{\partial \alpha_t} \leftarrow \text{dual\_alpha\_gradient}(\mu_t(\mathbf{s}), \pi_t(\mathbf{a}|\mathbf{s}), q_t(\mathbf{a}|\mathbf{s}), q_t(\mathbf{s}), \epsilon)$;
  /* update α_t along the gradient with step λ_t */
  $\alpha_t = \alpha_t - \lambda_t \frac{\partial L}{\partial \alpha_t}$;

Algorithm 3: State-Action Policy Search: Dual Gradient Descent over α_t in Pseudo-Code

4.4.4 Block Coordinate Descent

A significant draw back of Algorithm 3 is the computation cost of preforming the block descent over $V_t(\mathbf{s})$ and $\mu_t(\mathbf{s})$ for every gradient-descent step of α_t. Therefore, we suggest a modified algorithm, that implements a different block-coordinate-descent with respect to $V_t(\mathbf{s})$, $\mu_t(\mathbf{s})$ and α_t. By holding the state value function $V_t(\mathbf{s})$ constant while optimizing α_t, and vice versa, we are able to optimize both separately and reduce computation time dramatically. However, that would require us to reconsider the optimality condition of $\mu_t(\mathbf{s})$ when optimizing α_t. Thus, we need to retake the partial derivative of Equation 4.6 with respect to $\mu_t(\mathbf{s})$, we arrive at a different closed-form condition for $\mu_t(\mathbf{s})$

$$\mu_t(\mathbf{s}) = \mathcal{N}(\mathbf{s}|V_t(\mathbf{s}), \hat{V}_t(\mathbf{s}), \alpha_t), \tag{4.9}$$

where $\hat{V}_t(\mathbf{s})$ is a term that resembles an α_t-dependent state value function

$$\hat{V}_t(\mathbf{s}) = \log \int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t \log q_t(\mathbf{s},\mathbf{a}) + R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big]d\mathbf{a}. \tag{4.10}$$

A full derivation of the coordinate-descent scheme can be found in Appendix B.

input : T ; /* time horizon */
$\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})$; /* linearized dynamics */
$\mu_1(\mathbf{s})$; /* initial state distribution */
$q_t(\mathbf{a}|\mathbf{s})$; /* last policy */
$q_t(\mathbf{s})$; /* last state distribution */
$\mathbf{M}_t, \mathbf{H}_t, \mathbf{z}_t$; /* reward matrices and goal state */

output: $\pi_t(\mathbf{a}|\mathbf{s})$; /* optimal policy */
$V_t(\mathbf{s})$; /* optimal state value function */
$\mu_t(\mathbf{s})$; /* optimal state distribution */

initialize α_t ; /* initial guess of α_t */

/* minimizing the dual by coordinate descent */
while $L(\mu_t, V_t, \alpha_t)$ not at minimum **do**
  /* do block-descent to compute $V_t(\mathbf{s})$ and $\mu_t(\mathbf{s})$ */
  $[V_t(\mathbf{s}), \pi_t(\mathbf{a}|\mathbf{s}), \sim] \leftarrow \text{block\_descent}(\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}), q_t(\mathbf{a}|\mathbf{s}), q_t(\mathbf{s}), p_t(\mathbf{s}), \mathbf{M}_t, \mathbf{H}_t, \mathbf{z}_t, \alpha_t)$;

  /* minimize Lagrange dual with respect to α_t */
  while $L(\mu_t, \alpha_t)$ not at minimum **do**
    /* compute $\hat{V}_t(\mathbf{s})$ with Equation 4.10 */
    $[\hat{V}_t(\mathbf{s}), \hat{\pi}_t(\mathbf{s})] \leftarrow \text{co\_descent\_backward\_pass}(\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}), V_t(\mathbf{s}), q_t(\mathbf{a}|\mathbf{s}), q_t(\mathbf{s}), \mathbf{M}_t, \mathbf{H}_t, \mathbf{z}_t, \alpha_t)$;

    /* compute state distribution $\hat{\mu}_t(\mathbf{s})$ with Equation 4.9 */
    $\hat{\mu}_t(\mathbf{s}) \leftarrow \text{co\_descent\_state\_distribution}(\hat{V}_t(\mathbf{s}), V_t(\mathbf{s}), \alpha_t)$;

    /* update Lagrange dual value with Equation 4.3 */
    $L(\mu_t, \alpha_t) \leftarrow \text{update\_dual}(V_t(\mathbf{s}), \hat{V}_t(\mathbf{s}), \hat{\mu}_t(\mathbf{s}), \alpha_t, \epsilon)$;

    /* compute Lagrange dual gradient with respect to α_t using Equation 4.4c */
    $\frac{\partial L}{\partial \alpha_t} \leftarrow \text{dual\_alpha\_gradient}(\hat{\mu}_t(\mathbf{s}), \hat{\pi}_t(\mathbf{a}|\mathbf{s}), q_t(\mathbf{a}|\mathbf{s}), q_t(\mathbf{s}), \epsilon)$;

    /* update α_t along the gradient with step λ_t */
    $\alpha_t = \alpha_t - \lambda_t \frac{\partial L}{\partial \alpha_t}$;

Algorithm 4: State-Action Policy Search: Dual Coordinate Descent in Pseudo-Code

5 Entropy State-Action Bound Policy Search

The introduction of a stochastic policy to the classical Markov Decision Process formulation of Optimal Control, poses challenges similar to problems that occur in general Stochastic Search settings (Abdolmaleki et al., 2015). These issues boil down to the problem of exploration vs. exploitation. The stochasticity of a policy adds to the ability of an algorithm to explore the state-action space. The challenge lies in systematically controlling the variance of the policy in a way that allows for exploration but also converges to a mean controller that maximizes the expected reward. Algorithms like Guided Policy Search and State-Action Bound Policy Search can suffer from premature convergence, because of the nature of their relative entropy bounds. The KL-divergence acts on the mean and variance of a distribution and may result in the algorithm opting to greedily maximizing its reward by rapidly shrinking the variance and barely exploring in the direction of mean actions. To counteract this dynamic, we introduce a new constraint on the entropy of the policy that aims to maintain a lower bound of stochasticity and, thus, forces exploration in the action space.

5.1 Optimization Problem

The new optimization problem is analog to that of State-Action Bound Policy Search with the addition of an entropy constraint in Equation 5.1d

$$\underset{\pi_t(\mathbf{a}|\mathbf{s})}{\operatorname{argmax}} \quad \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s}, \tag{5.1a}$$

$$\text{s.t.} \quad \forall \mathbf{s}', \forall t > 1 \quad \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_{t-1}(\mathbf{s}) \pi_{t-1}(\mathbf{a}|\mathbf{s}) \mathcal{P}_{t-1}(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} = \mu_t(\mathbf{s}'), \tag{5.1b}$$

$$\forall t < T \quad \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s},\mathbf{a})} d\mathbf{a} d\mathbf{s} \leq \epsilon, \tag{5.1c}$$

$$\forall t < T \quad \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \log \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} \leq \delta, \tag{5.1d}$$

$$\forall \mathbf{s}, \forall t < T \quad \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} = 1, \tag{5.1e}$$

$$\forall \mathbf{s}, t = 1 \quad \mu_1(\mathbf{s}) = p_1(\mathbf{s}). \tag{5.1f}$$

The hyperparameter δ can be chosen in such a way, for example, to maintain or increase the variance or entropy of the last policy $q_t(\mathbf{a}|\mathbf{s})$ by some factor.

5.2 Dual Problem

Just as in GPS and SAPS, we transform the primal problem to its dual equivalent by solving for $\pi_t(\mathbf{a}|\mathbf{s})$. The introduction of the entropy constraint 5.1d results in a new Lagrangian variable for each time step β_t. A complete derivation of Entropy State-Action Bound Policy Search is in Appendix C

$$\pi_t(\mathbf{a}|\mathbf{s}) \propto \exp\Big[\frac{1}{\alpha_t + \beta_t}\Big(R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}'\Big)\Big]. \tag{5.2}$$

We substitute $\pi_t(\mathbf{a}|\mathbf{s})$ into the primal problem to get the dual function $L(\mu_t, V_t, \alpha_t, \beta_t)$

$$\begin{aligned} L = &\int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}') \mu_T(\mathbf{s}') d\mathbf{s}' \\ &- \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_t(\mathbf{s}') \mu_t(\mathbf{s}') d\mathbf{s}' + \sum_{t=1}^{T-1} \alpha_t \epsilon + \sum_{t=1}^{T-1} \beta_t \delta - \sum_{t=1}^{T-1} \int_{\mathbf{s}} \alpha_t \mu_t(\mathbf{s}) \log \mu_t(\mathbf{s}) d\mathbf{s} \\ &+ \sum_{t=1}^{T-1} (\alpha_t + \beta_t) \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} \exp\Big[\frac{R_t(\mathbf{s}, \mathbf{a}) + \alpha_t \log q_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a}) d\mathbf{s}'}{\alpha_t + \beta_t} \Big] d\mathbf{a} d\mathbf{s}. \end{aligned} \tag{5.3}$$

For dual minimization, we take the partial derivatives of $L(\mu_t, V_t, \alpha_t, \beta_t)$ and set them to zero to get the optimality conditions of the state value function $V_t(\mathbf{s})$ and state distribution $\mu_t(\mathbf{s})$

$$\frac{\partial L}{\partial \mu_t} = \begin{cases} R_T(\mathbf{s}) - V_T(\mathbf{s}) & , t = T \\ V_t(\mathbf{s}) - (\alpha_t + \beta_t) \log \int_{\mathbf{a}} \exp\Big[\frac{\alpha_t \log q_t(\mathbf{s}, \mathbf{a}) - \alpha_t \log \mu_t(\mathbf{s}) - \alpha_t + R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a})}{\alpha_t + \beta_t} \Big] & , t < T \end{cases}, \tag{5.4a}$$

$$\frac{\partial L}{\partial V_t} = \begin{cases} p_1(\mathbf{s}) - \mu_1(\mathbf{s}) & , t = 1 \\ \mu_t(\mathbf{s}) - \int_{\hat{\mathbf{s}}} \int_{\mathbf{a}} \pi_{t-1}(\mathbf{a}|\hat{\mathbf{s}}) \mu_{t-1}(\hat{\mathbf{s}}) \mathcal{P}_{t-1}(\mathbf{s}|\hat{\mathbf{s}}, a) d\mathbf{a} d\hat{\mathbf{s}} & , t > 1 \end{cases}, \tag{5.4b}$$

$$\frac{\partial L}{\partial \alpha_t} = \epsilon - \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s}, \mathbf{a})} d\mathbf{a} d\mathbf{s} \tag{5.4c}$$

$$\frac{\partial L}{\partial \beta_t} = \delta - \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \log \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s}. \tag{5.4d}$$

By plugging these optimality conditions into Equation 5.3 we get a simplified dual $L(\mu_t, V_t, \alpha_t, \beta_t)$

$$L(\mu_t, V_t, \alpha_t, \beta_t) = \int_{\mathbf{s}} V_1(\mathbf{s}) \mu_1(\mathbf{s}) d\mathbf{s} + \sum_{t=1}^{T-1} \alpha_t (\epsilon + 1) + \sum_{t=1}^{T-1} \beta_t \delta. \tag{5.5}$$

5.3 Augmented Reward

From Equations 5.4b and 5.4a, it is clear that the reward function $r_t(\mathbf{s}, \mathbf{a})$ is similar to that of the State-ACtion Bound Policy Search in Equation 4.7. However, the temperature parameter of the state-action value function $Q_t(\mathbf{s}, \mathbf{a})$ and the weighting of the state value function $V_t(\mathbf{s})$ have the added value of β_t

$$Q_t(\mathbf{s}, \mathbf{a}) = \frac{1}{\alpha_t + \beta_t} \Big(r_t(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\mathcal{P}}[V_{t+1}(\mathbf{s}')] \Big), \tag{5.6a}$$

$$V_t(\mathbf{s}) = (\alpha_t + \beta_t) \log \int_{\mathbf{a}} \exp\Big[Q_t(\mathbf{s}, \mathbf{a}) \Big] d\mathbf{a}. \tag{5.6b}$$

In Appendix C, we do full derivation of ESAPS under linear Gaussian dynamics $\mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a}) = \mathcal{N}(\mathbf{s}'|\mathbf{A}_t \mathbf{s} + \mathbf{b}_t \mathbf{a} + \mathbf{c}_t, \mathbf{\Sigma}_{s'})$ and time variant quadratic rewards $R_t(\mathbf{s}, \mathbf{a}) = (\mathbf{z} - \mathbf{s})^T \mathbf{M}_t (\mathbf{z} - \mathbf{s}) + \mathbf{a}^T \mathbf{H}_t \mathbf{a}$ and show the resulting value functions $Q_t(\mathbf{s}, \mathbf{a})$ and $V_t(\mathbf{s})$ are also quadratic and the policy $\pi_t(\mathbf{a}|\mathbf{s})$ is a linear-Gaussian distribution.

5.4 Implementation

The implementation of ESAPS is similar in its structure to SAPS with an additional optimization over β_t. Algorithm 5 shows the details of the coordinate-descent scheme.

input : T ; /* time horizon */
$\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})$; /* linearized dynamics */
$\mu_1(\mathbf{s})$; /* initial state distribution */
$q_t(\mathbf{a}|\mathbf{s})$; /* last policy */
$q_t(\mathbf{s})$; /* last state distribution */
$\mathbf{M}_t, \mathbf{H}_t, \mathbf{z}_t$; /* reward matrices and goal state */

output: $\pi_t(\mathbf{a}|\mathbf{s})$; /* optima policy */
$V_t(\mathbf{s})$; /* optimal state value function */
$\mu_t(\mathbf{s})$; /* optimal state distribution */

initialize α_t, β_t ; /* initial guess of α_t, β_t */

/* minimizing the dual by coordinate descent */
while $L(\mu_t, V_t, \alpha_t, \beta_t)$ not at minimum **do**
  /* do block-descent to compute $V_t(\mathbf{s})$ and $\mu_t(\mathbf{s})$ */
  $[V_t(\mathbf{s}), \pi_t(\mathbf{a}|\mathbf{s}), \sim] \leftarrow$ block_descent$(\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}), q_t(\mathbf{a}|\mathbf{s}), q_t(\mathbf{s}), p_t(\mathbf{s}), \mathbf{M}_t, \mathbf{H}_t, \mathbf{z}_t, \alpha_t, \beta_t)$;

  /* minimize Lagrange dual with respect to α_t */
  while $L(\mu_t, \alpha_t)$ not at minimum **do**
    /* compute state value function $\hat{V}_t(\mathbf{s})$ */
    $[\hat{V}_t(\mathbf{s}), \hat{\pi}_t(\mathbf{s})] \leftarrow$
    co_descent_backward_pass$(\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}), V_t(\mathbf{s}), q_t(\mathbf{a}|\mathbf{s}), q_t(\mathbf{s}), \mathbf{M}_t, \mathbf{H}_t, \mathbf{z}_t, \alpha_t, \beta_t)$;

    /* compute state distribution $\hat{\mu}_t(\mathbf{s})$ */
    $\hat{\mu}_t(\mathbf{s}) \leftarrow$ co_descent_state_distribution$(\hat{V}_t(\mathbf{s}), V_t(\mathbf{s}), \alpha_t, \beta_t)$;

    /* update Lagrange dual value with Equation 5.3 */
    $L(\mu_t, \alpha_t, \beta_t) \leftarrow$ update_dual$(V_t(\mathbf{s}), \hat{V}_t(\mathbf{s}), \hat{\mu}_t(\mathbf{s}), \alpha_t, \beta_t, \epsilon, \delta)$;

    /* compute Lagrange dual gradient with respect to α_t using Equation 5.4c */
    $\frac{\partial L}{\partial \alpha_t} \leftarrow$ dual_alpha_gradient$(\hat{\mu}_t(\mathbf{s}), \hat{\pi}_t(\mathbf{a}|\mathbf{s}), q_t(\mathbf{a}|\mathbf{s}), q_t(\mathbf{s}), \epsilon)$;

    /* compute Lagrange dual gradient with respect to β_t using Equation 5.4d */
    $\frac{\partial L}{\partial \beta_t} \leftarrow$ dual_beta_gradient$(\hat{\mu}_t(\mathbf{s}), \hat{\pi}_t(\mathbf{a}|\mathbf{s}), \delta)$;

    /* update α_t β_t along the gradient with step λ_t and ζ_t */
    $\alpha_t = \alpha_t - \lambda_t \frac{\partial L}{\partial \alpha_t}$; $\beta_t = \beta_t - \zeta_t \frac{\partial L}{\partial \beta_t}$

Algorithm 5: Entropy State-Action Policy Search: Dual Coordinate Descent in Pseudo-Code

6 Evaluation

6.1 Double Pendulum Task

The double pendulum task is a setup with a fully actuated two link arm under the influence of gravity. The objective of the learners is to do a full swing up of the pendulum starting from the down-right position and try to stabilize the tail of the trajectory around the up-right posture. To make the task harder, we introduce friction to the joints and shift the center of mass towards the end of the second link. Furthermore, we limit the allowed torque by applying sharp non-linear constraints. The number of samples used for linearization is 25 per iteration.

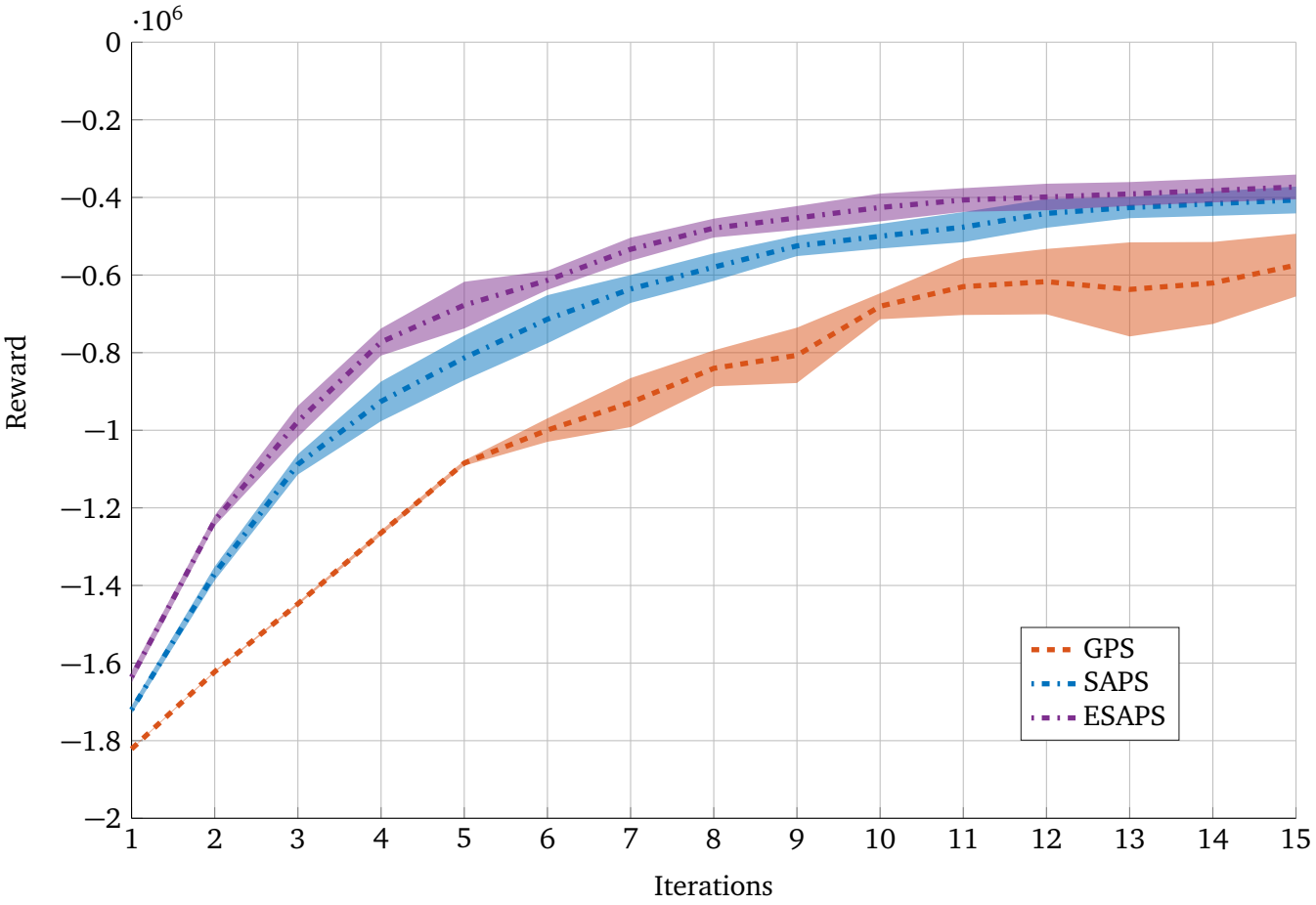

Figure 6.1.: Double Pendulum Task: The total expected reward of GPS, SAPS and ESAPS in comparison during a swing-up task. Each learner is given 25 iterations per trial to find the best policy. To account for the stochasticity of the setup, 10 trails were preformed and averaged. The hyperparameters of each learner were optimized separately to reflect its best performance.

Figure 6.1 shows a direct comparison between GPS, SAPS and ESAPS after independent optimization of the respective hyperparameters. After 25 iterations, GPS reaches the lowest reward and demonstrates the highest rate of oscillation during the last 5 iteration, which is due to the system prematurely running

into the torque limits. SAPS and ESAPS both out preform GPS by reaching the same reward level after only have the number of iterations or less.

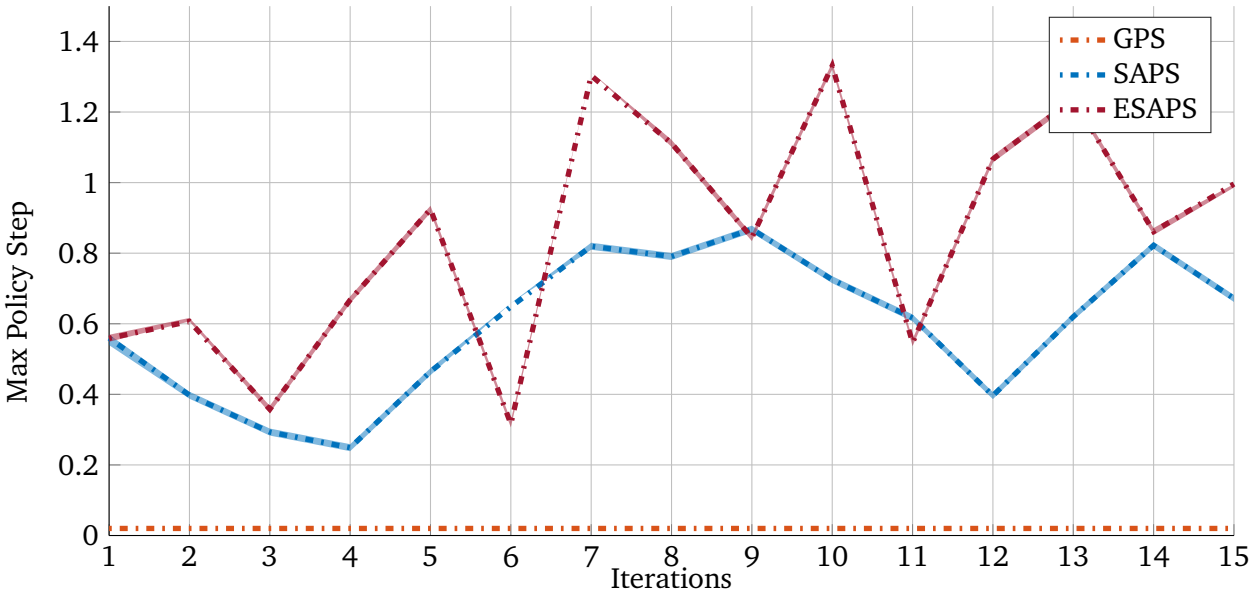

Figure 6.2.: Double Pendulum Task: The maximum change in the policy for each iteration of GPS, SAPS and ESAPS. GPS has a constant step that is equal its KL-bound. SAPS takes significantly bigger steps while maintaining the upper bound on the state-action distribution. ESAPS is able to take the largest steps due to its ability to maintain a larger variance

Figure 6.2 illustrates the maximum KL-divergence of the policy after each iteration. The results validate our assumption, that by bounding the state-action distribution in SAPS and ESAPS, we are able take larger steps in the policy space without the risk of leaving the vicinity of the linearized dynamics. Also, by maintaining a significant portion of its entropy, ESAPS is capable of taking larger steps in the direction of the mean action.

6.2 Quad Pendulum Task

The quad pendulum task is similar to that of the double pendulum, albeit with much higher complexity in the dynamics. The pendulum is fully actuated and has to be swung-up and stabilized in the up-right position. We only specify the end-points of the trajectory for stabilization and forgo the specification of any other via-points. The number of samples used for linearization is 100 per iteration.
Figure 6.3 offers a comparison of the total expected reward of ESAPS against GPS. The hyperparameters of both algorithms were optimized independently. It is clear that ESAPS outperforms GPS by a very large margin, reaching similar reward levels after only 25 iterations compared to 50 iterations for GPS.
A justification for this difference in performance is found in Figure 6.4, that compares the maximum policy step that both algorithms can take without risking divergence. ESAPS can, at least for some time steps, take steps 6-7 times the step of GPS without compromising the integrity of the linearization.

6.3 Discussion

Based on the results we have presented, it is clear that our assumptions have been validated to some extent. In direct comparison to GPS, we were able to show the impact of bounding the state distribution to preserve the validity of the linearization, as it allowed us to execute larger steps in the policy space and significantly reduce the number of iterations and samples. Also, the existence of the an entropy lower bound has contributed to maintaining exploration and, thus, reaching better end policies.

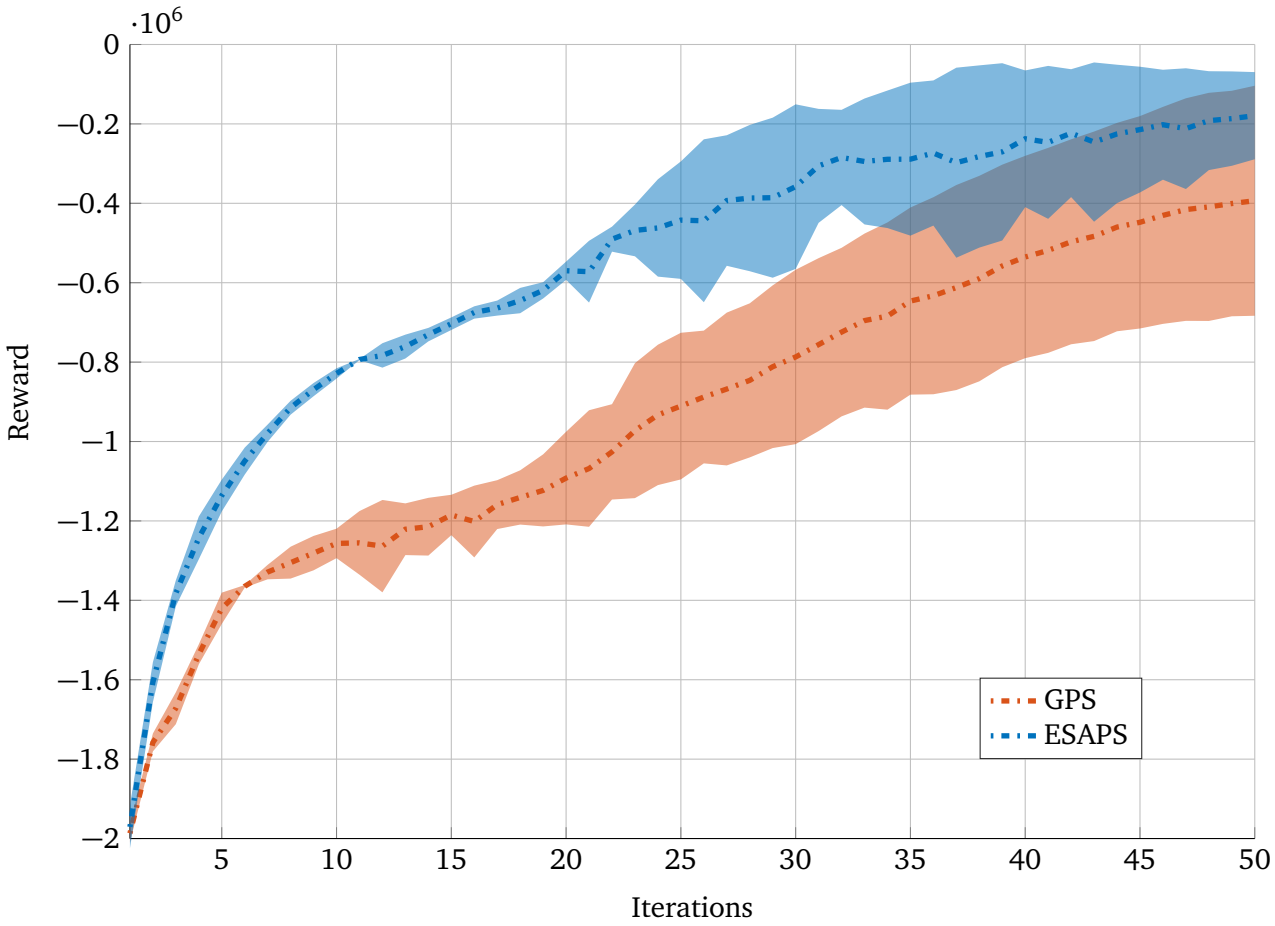

Figure 6.3.: Quad Pendulum Task: The expected reward of GPS and ESAPS. Each learner is given 50 iterations. For a statistical mean of the expected reward, 10 trails were preformed and averaged. The hyperparameters of each learner were optimized separately to reflect its best performance. The final result shows ESAPS outperforming GPS significantly.

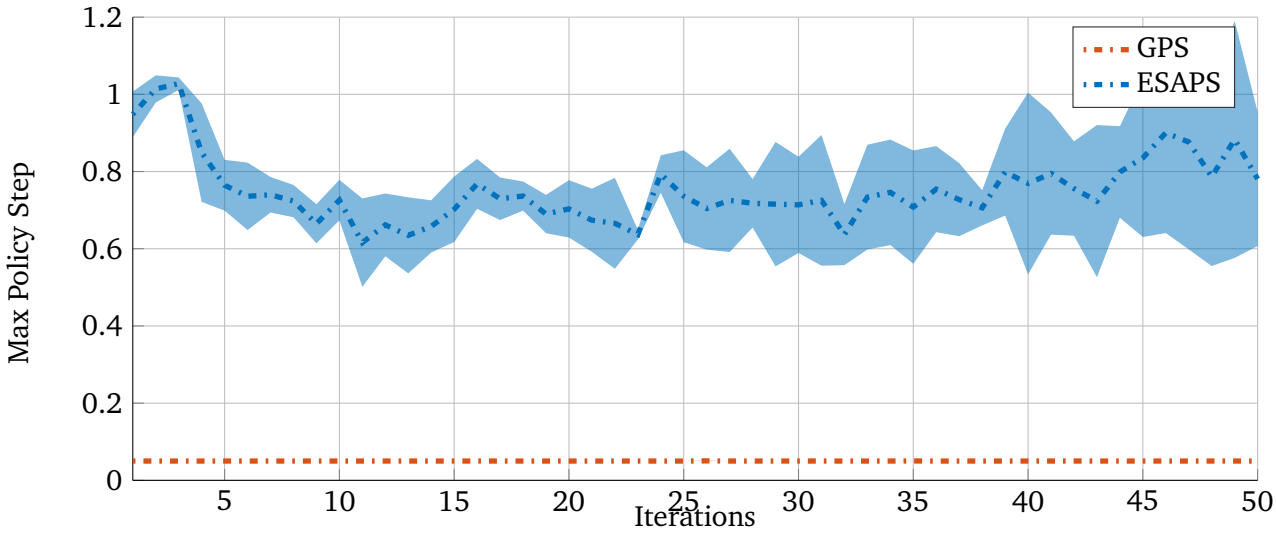

Figure 6.4.: Quad Pendulum Task: The maximum step in the policy space for each iteration of GPS and ESAPS. The step of GPS, per definition, is constant and equal its KL-bound. ESAPS, however, modulates the maximum step size based on the state-action bound

7 Future Work

In this chapter we suggest a possible list of improvements and areas of further research, based on the encouraging results we have presented.

7.1 Separate Bounds on State and Action

Our main contribution in this thesis has been the introduction of an upper bound on the change of state distribution in iterative Stochastic Optimal Control methods. We have chosen to achieve that by bounding the state-action distribution. However, it is conceivable that two separate bounds, one on the policy and one the state distribution, may carry some advantages, such as being able to set independent upper or lower bounds on the policy change.

7.2 Comparison to Full Gradient Descent

In our derivations we have shown that we are able to compute the optimal value function and state distribution in closed-form based on two optimality conditions from the dual partial derivatives. This formulation reduces the minimization of the dual function to a gradient descent problem over the Lagrangian multipliers associated with relative entropy constraint. In the future we plan to analyze the possibility of applying full gradient descent on the value function and state distribution and comparing the search direction to that of the optimality conditions.

7.3 Principled Control of Policy Entropy

By adding the entropy constraint on the policy in Entropy Bound State Action Policy Search, we were able to prevent the decay of the policy variance, allowing us to explore the action space for a larger number of iterations. A possible extension is the introduction of some heuristics that would not only maintain the variance but also increase it. Such ability to manipulate the entropy would help in escaping shallow local minima that might result of sub-optimal initialization of the policy.

7.4 Reformulation for Deterministic Policies

By concentrating on the formulation of Guided Policy Search, we are limited to a class of algorithms that try to optimize a stochastic policy. However, the original formulation of the Markov Decision Process does not necessarily require such a policy. In fact, it states that the optimal policy is a deterministic controller. Based on this insight, it may be interesting to reformulate the problem along the lines of Differential Dynamic Programming and Iterative Quadratic Gaussian and exploring equivalent regularizations that correspond to what we have introduces in this thesis.

7.5 Further Evaluations on Larger and Real Systems

Although our results are promising, further comparisons to other state-of-the-art algorithms are still needed for stronger validation. Also the application on high dimensional and real systems would help us understand the scalability of computation time and feasibility in regard to the number of samples.

8 Conclusion

Stochastic Optimal Control with linearized dynamics is a powerful technique for learning optimal control policies of highly non-linear systems. In this thesis we have investigated and introduced several variations of state-of-the-art algorithms in this field.
In our introduction we have discussed a major issue in this class of algorithms, which is its dependency on the validity of the model around the linearization point. Hence, it is crucial to provide guarantees that would prevent a greedy exploitation of the local dynamics.
In Chapter 3, we went on to analyze a recent approach, Guided Policy Search, that addresses this issue, by forcing a relative entropy bound on the trajectory distribution between iterations. We succeeded in reformulating GPS and were able to show that its proposed constraint is equivalent to bounding the policy update at each time step. We have also argued that such an approach only implicitly bounds the state distribution around which the system is linearized. Thus, to avoid divergence in highly dynamical systems, the algorithm is limited to very small updates on the policy, which would, in turn, increase the number of iteration and samples needed.
In Chapter 4, relying on these insights, we proposed a new constraint that explicitly imposes a relative entropy bound on the state distribution by bounding the state-action distribution instead of the policy. This addition has resulted in a number of new algorithmic challenges, which we were able to deal with. The main issue was the emergence of new reward terms that encode the distance between the current and last state distribution, which has lead to a circular dependency between the value function and state distribution, which we were able to solve by applying a block-coordinate-descent scheme.
By concentrating on a class of algorithms that require a stochastic policy and due to the nature of the relative entropy bounds we have introduced, we were inevitably confronted with problems of trade-off between exploration and exploitation. We addressed this issue, in Chapter 5, through an additional constraint on the differential entropy of the policy, thus, allowing us to control the stochasticity of policy as the algorithm advances after each iteration.
As proof of concept of our contribution, we have compared our algorithms with GPS by preforming a swing-up task on the highly non-linear double and quad pendulums. The results validate our view, that a bound on the state-action distribution allows for more aggressive updates of the policy, while setting an upper bound on the divergence of the state distribution.
Finally, we have discussed ways to improve and extend our contribution, such as introducing separate bounds on the state and action, developing a principled approach for manipulating the entropy of the policy and preforming evaluations on higher dimensional and real systems.

References

Abbas Abdolmaleki, Rudolf Lioutikov, Nuno Lau, Luis Reis, Jan Peters, and Gerhard Neumann. *Model-Based Relative Entropy Stochastic Search*. Advances in Neural Information Processing Systems (NIPS), 2015.

Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2009.

Marc Deisenroth and Carl Rasmmussen. *PILCO: A Model-Based and Data-Efficient Approach to Policy Search*. International Conference on Machine Learning (ICML), 2011.

Marc Deisenroth, Gerhard Neumann, and Jan Peters. *A Survey on Policy Search for Robotics*. Foundations and Trends in Robotics, 2013.

David Jacobson and David Mayne. *Differential Dynamic Programming*. American Elsevier Publishing Company, 1970.

Steven G. Johnson. *The NLopt Nonlinear-Optimization Package*. 2016.

Mrinal Kalakrishnan, Sachin Chitta, Evangelos Theodorou, Peter Pastor, and Stefan Schaal. *STOMP: Stochastic Trajectory Optimization for Motion Planning*. International Conference on Robotics and Automation (ICRA), 2011.

Sergey Levine and Pieter Abbeel. *Learning Neural Network Policies with Guided Policy Search under Unknown Dyamics*. Advances in Neural Information Processing Systems (NIPS), 2014.

Sergey Levine and Vladeln Koltun. *Guided Policy Search*. International Conference on Machine Learning (ICML), 2013.

Sergey Levine and Vladeln Koltun. *Learning Complex Neural Network Policies with Trajectory Opitmization*. International Conference on Machine Learning (ICML), 2014.

David Mayne. *A Second-order Gradient Method for Determining Optimal Trajectories of Non-linear Discrete-time Systems*. International Journal of Control, Vol. 3, Iss. 1, 1966.

Jan Peters, Katharina Mülling, and Yasemin Altun. *Relative Entropy Policy Search*. National Conference on Artificial Intelligence (AAAI), 2010.

Kaare Petersen and Michael Pedersen. *The Matrix Cookbook*. 2012.

Nathan Ratliff, Matt Zucker, Andrew Bagnell, and Siddhartha Srinivasa. *CHOMP: Gradient Optimization Techniques for Efficient Motion Planning*. International Conference on Robotics and Automation (ICRA), 2009.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. *Approximate Inference and Stochastic Optimal Control*. e-Print arXiv:1009.3958, 2010.

Elmar Rueckert, Max Mindt, Jan Peters, and Gerhard Neumann. *Robust Policy Updates for Stochastic Optimal Control*. IEEE/RAS International Conference on Humanoid Robots (HUMANOIDS), 2014.

Conrad Sanderson. *Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments*. NICTA, 2010.

John Schulman, Jonathan Ho, Alex Lee, Henry Awwal, Ibrahim Bradlow, and Pieter Abbeel. *Finding Locally Optimal, Collision-Free Trajectories with Sequential Convex Optimization*. International Conference on Robotics and Automation (ICRA), 2013.

Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Yuval Tassa, Tom Erez, and Emanuel Todorov. *Synthesis and Stabilization of Complex Behaviors through Online Trajectory Optimization*. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2012.

Emanuel Todorov and Weiwei Li. *A Generalized Iterative LQG Method for Locally-Optimal Feedback Control of Constrained Nonlinear Stochastic Systems*. American Control Conference (ACC), 2005.

Marc Toussaint. *Robot Trajectory Optimization using Approximate Inference*. International Conference on Machine Learning (ICML), 2009a.

Marc Toussaint. *Stochastic Optimal Control*. Lecture Notes, 2009b.

A Derivation of Guided Policy Search

$$\underset{\pi_t(\mathbf{a}|\mathbf{s})}{\operatorname{argmax}} \quad \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s}$$

$$\forall \mathbf{s}, \forall t < T \quad \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} = 1$$

$$\forall \mathbf{s}', \forall t > 1 \quad \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_{t-1}(\mathbf{s}) \pi_{t-1}(\mathbf{a}|\mathbf{s}) \mathcal{P}_{t-1}(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} = \mu_t(\mathbf{s}')$$

$$\forall \mathbf{s}, t = 1 \quad \mu_1(\mathbf{s}) = p_1(\mathbf{s})$$

$$\forall t < T \quad \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{a}|\mathbf{s})} d\mathbf{a} d\mathbf{s} \leq \epsilon$$

Primal Problem:

$$\begin{aligned}
L(\pi_t, \mu_t, V_t, \lambda_t, \alpha_t) &= \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) \Big(1 - \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a}\Big) d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) \Big(p_1(\mathbf{s}) - \mu_1(\mathbf{s})\Big) d\mathbf{s} \\
&\quad + \sum_{t=2}^{T} \int_{\mathbf{s}'} V_t(\mathbf{s}') \Big(\int_{\mathbf{s}} \int_{\mathbf{a}} \mu_{t-1}(\mathbf{s}) \pi_{t-1}(\mathbf{a}|\mathbf{s}) \mathcal{P}_{t-1}(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} - \mu_t(\mathbf{s}') \Big) d\mathbf{s}' \\
&\quad + \sum_{t=1}^{T-1} \Big(\alpha_t \epsilon - \alpha_t \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{a}|\mathbf{s})} d\mathbf{a} d\mathbf{s} \Big) \\
L(\pi_t, \mu_t, V_t, \lambda_t, \alpha_t) &= \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) d\mathbf{s} - \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}') \mu_T(\mathbf{s}') d\mathbf{s}' \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} d\mathbf{s}' - \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_t(\mathbf{s}') \mu_t(\mathbf{s}') d\mathbf{s}' \\
&\quad + \sum_{t=1}^{T-1} \Big(\alpha_t \epsilon - \alpha_t \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{a}|\mathbf{s})} d\mathbf{a} d\mathbf{s} \Big) \\
\frac{\partial L(\pi_t, \mu_t, V_t, \lambda_t, \alpha_t)}{\partial \pi_t} &= R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) - \lambda_t(\mathbf{s}) + \int_{\mathbf{s}'} \mu_t(\mathbf{s}) V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' \\
&\quad - \alpha_t \mu_t(\mathbf{s}) \log \frac{\pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{a}|\mathbf{s})} - \alpha_t \mu_t(\mathbf{s}) = 0 \\
\Rightarrow \pi_t(\mathbf{a}|\mathbf{s}) &= q_t(\mathbf{a}|\mathbf{s}) \exp\Big[\frac{1}{\alpha_t} \Big(R_t(\mathbf{s},\mathbf{a}) - \frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' - \alpha_t \Big) \Big]
\end{aligned}$$

Dual Problem:

$$
\begin{aligned}
L(\mu_t, V_t, \lambda_t, \alpha_t) &= \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{s} d\mathbf{a} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) d\mathbf{s} - \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}') \mu_T(\mathbf{s}') d\mathbf{s}' \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \mathcal{P}(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} d\mathbf{s}' - \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_t(\mathbf{s}') \mu_t(\mathbf{s}') d\mathbf{s}' \\
&\quad + \sum_{t=1}^{T-1} \Big(\alpha_t \epsilon - \alpha_t \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \ldots \\
&\quad \ldots \log \frac{q_t(\mathbf{a}|\mathbf{s}) \exp\Big[\frac{1}{\alpha_t} \Big(R_t(\mathbf{s},\mathbf{a}) - \frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' - \alpha_t \Big) \Big]}{q_t(\mathbf{a}|\mathbf{s})} d\mathbf{a} d\mathbf{s} \Big) \\
&= \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} + \sum_{t=1}^{T-1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}') \mu_T(\mathbf{s}') d\mathbf{s}' \\
&\quad - \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_t(\mathbf{s}') \mu_t(\mathbf{s}') d\mathbf{s}' + \sum_{t=1}^{T-1} \alpha_t (\epsilon + 1)
\end{aligned}
$$

Solve for λ_t:

$$
\begin{aligned}
1 &= \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} \\
1 &= \int_{\mathbf{a}} q_t(\mathbf{a}|\mathbf{s}) \exp\Big[\frac{1}{\alpha_t} \Big(R_t(\mathbf{s},\mathbf{a}) - \frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' - \alpha_t \Big) \Big] d\mathbf{a} \\
1 &= \int_{\mathbf{a}} q_t(\mathbf{a}|\mathbf{s}) \exp\Big[\frac{1}{\alpha_t} \Big(R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' \Big) d\mathbf{s}' \Big] \exp\Big[-\frac{\lambda_t(\mathbf{s})}{\alpha_t \mu_t(\mathbf{s})} - 1 \Big] d\mathbf{a} \\
\exp\Big[\frac{\lambda_t(\mathbf{s})}{\alpha_t \mu_t(\mathbf{s})} + 1 \Big] &= \int_{\mathbf{a}} q_t(\mathbf{a}|\mathbf{s}) \exp\Big[\frac{1}{\alpha_t} \Big(R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' \Big) d\mathbf{s}' \Big] d\mathbf{a} \\
\lambda_t(\mathbf{s}) &= \alpha_t \mu_t(\mathbf{s}) \Big(-1 + \log \int_{\mathbf{a}} q_t(\mathbf{a}|\mathbf{s}) \exp\Big[\frac{1}{\alpha_t} \Big(R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' \Big) \Big] d\mathbf{a} \Big)
\end{aligned}
$$

$$
\begin{aligned}
L(\mu_t, V_t, \alpha_t) &= \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}') \mu_T(\mathbf{s}') d\mathbf{s}' \\
&\quad - \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_t(\mathbf{s}') \mu_t(\mathbf{s}') d\mathbf{s}' + \sum_{t=1}^{T-1} \alpha_t \epsilon \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}} \alpha_t \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} q_t(\mathbf{a}|\mathbf{s}) \exp\Big[\frac{1}{\alpha_t} \Big(R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' \Big) \Big] d\mathbf{a} d\mathbf{s}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial \mu_t} &= -V_t(\mathbf{s}) + \alpha_t \log \int_{\mathbf{a}} q_t(\mathbf{a}|\mathbf{s}) \exp\Big[\frac{1}{\alpha_t}\Big(R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big]d\mathbf{a} \\
0 &= -V_t(\mathbf{s}) + \alpha_t \log \int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t \log q_t(\mathbf{a}|\mathbf{s}) + R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big]d\mathbf{a} \\
\frac{\partial L}{\partial \mu_T} &= R_T(\mathbf{s}) - V_T(\mathbf{s})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial V_t} &= -\mu_t(\mathbf{s}) + \int_{\hat{\mathbf{s}}} \mu_{t-1}(\hat{\mathbf{s}}) \int_{\mathbf{a}} \frac{q_{t-1}(\mathbf{a}|\hat{\mathbf{s}}) \exp\Big[\frac{1}{\alpha_{t-1}}\Big(R_{t-1}(\hat{\mathbf{s}},\mathbf{a}) + \int_{\mathbf{s}'} V_t(\mathbf{s}')\mathcal{P}_{t-1}(\mathbf{s}'|\hat{\mathbf{s}},\mathbf{a})d\mathbf{s}'\Big)\Big]}{\int_{\mathbf{a}} q_{t-1}(\mathbf{a}|\hat{\mathbf{s}}) \exp\Big[\frac{1}{\alpha_{t-1}}\Big(R_{t-1}(\hat{\mathbf{s}},\mathbf{a}) + \int_{\mathbf{s}'} V_t(\mathbf{s}')\mathcal{P}_{t-1}(\mathbf{s}'|\hat{\mathbf{s}},\mathbf{a})d\mathbf{s}'\Big)\Big]d\mathbf{a}} \mathcal{P}_{t-1}(\mathbf{s}|\hat{\mathbf{s}},\mathbf{a})d\mathbf{a}d\hat{\mathbf{s}} \\
0 &= -\mu_t(\mathbf{s}) + \int_{\hat{\mathbf{s}}} \int_{\mathbf{a}} \pi_{t-1}(\mathbf{a}|\hat{\mathbf{s}})\mu_{t-1}(\hat{\mathbf{s}})\mathcal{P}_{t-1}(\mathbf{s}|\hat{\mathbf{s}},\mathbf{a})d\mathbf{a}d\hat{\mathbf{s}} \\
\frac{\partial L}{\partial V_1} &= p_1(\mathbf{s}) - \mu_1(\mathbf{s})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial \alpha_t} &= \epsilon + \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} q_t(\mathbf{a}|\mathbf{s}) \exp\Big[\frac{1}{\alpha_t}\Big(R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big]d\mathbf{a}d\mathbf{s} \\
&\quad - \int_{\mathbf{s}} \mu_t(\mathbf{s})\alpha_t \int_{\mathbf{a}} \frac{q_t(\mathbf{a}|\mathbf{s}) \exp\Big[\frac{1}{\alpha_t}\Big(R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big]}{\int_{\mathbf{a}} q_t(\mathbf{a}|\mathbf{s}) \exp\Big[\frac{1}{\alpha_t}\Big(R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big]d\mathbf{a}} \ldots \\
&\quad \ldots \frac{1}{\alpha_t^2}\Big(R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)d\mathbf{a}d\mathbf{s} \\
0 &= \epsilon + \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} q_t(\mathbf{a}|\mathbf{s}) \exp\Big[\frac{1}{\alpha_t}\Big(R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big]d\mathbf{a}d\mathbf{s} \\
&\quad - \frac{1}{\alpha_t} \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})\Big(R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)d\mathbf{a}d\mathbf{s} \\
&= \epsilon - \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{a}|\mathbf{s})} d\mathbf{a}d\mathbf{s}
\end{aligned}
$$

$$
L(\mu_t, V_t, \alpha_t) = \int_{\mathbf{s}} V_1(\mathbf{s})p_1(\mathbf{s})d\mathbf{s} + \sum_{t=1}^{T-1} \alpha_t \epsilon
$$

Plug in Gaussians:

$$
\begin{aligned}
\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) &= \mathcal{N}(\mathbf{s}'|\mathbf{A}_t\mathbf{s}+\mathbf{b}_t\mathbf{a}+\mathbf{c}_t,\boldsymbol{\Sigma}_{s',t}) \\
q_t(\mathbf{a}|\mathbf{s}) &= \mathcal{N}(\mathbf{a}|\mathbf{K}_t^q\mathbf{s}+\mathbf{k}_t^q,\boldsymbol{\Sigma}_{a,t}^q) \\
\mu_t(\mathbf{s}) &= \mathcal{N}(\mathbf{s}|\boldsymbol{\tau}_{s,t}^\mu,\boldsymbol{\Sigma}_{s,t}^\mu) \\
R_t(\mathbf{s},\mathbf{a}) &= (\mathbf{z}-\mathbf{s})^T\mathbf{M}_t(\mathbf{z}-\mathbf{s})+\mathbf{a}^T\mathbf{H}_t\mathbf{a} \\
R_T &= (\mathbf{z}-\mathbf{s})^T\mathbf{M}_T(\mathbf{z}-\mathbf{s}) \\
\int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}' &= \mathbb{E}_{\mathcal{P}}[V_{t+1}(\mathbf{s}')] \\
V_{t+1}(\mathbf{s}) &= \mathbf{s}^T\mathbf{V}_{t+1}\mathbf{s}+\mathbf{s}^T\mathbf{v}_{t+1}+v_{t+1}
\end{aligned}
$$

$$
\begin{aligned}
r_t(\mathbf{s},\mathbf{a}) &= R_t(\mathbf{s},\mathbf{a})+\alpha_t\log q_t(\mathbf{a}|\mathbf{s}) \\
&= (\mathbf{z}-\mathbf{s})^T\mathbf{M}_t(\mathbf{z}-\mathbf{s})+\mathbf{a}^T\mathbf{H}_t\mathbf{a}-\alpha_t\log\sqrt{\left|2\pi\boldsymbol{\Sigma}_{a,t}^q\right|}-\frac{\alpha_t}{2}(\mathbf{a}-\mathbf{K}_t^q s-\mathbf{k}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}(\mathbf{a}-\mathbf{K}_t^q s-\mathbf{k}_t^q) \\
&= (\mathbf{z}-\mathbf{s})^T\mathbf{M}_t(\mathbf{z}-\mathbf{s})+\mathbf{a}^T\mathbf{H}_t\mathbf{a}-\alpha_t\log\sqrt{\left|2\pi\boldsymbol{\Sigma}_{a,t}^q\right|}-\frac{\alpha_t}{2}\Big(\mathbf{a}^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{a}-\mathbf{a}^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{K}_t^q s-\mathbf{a}^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q \\
&\quad -\mathbf{s}^T(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{a}+\mathbf{s}^T(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{K}_t^q\mathbf{s}+\mathbf{s}^T(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q-(\mathbf{k}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{a} \\
&\quad +(\mathbf{k}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{K}_t^q\mathbf{s}+(\mathbf{k}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q\Big) \\
&= \mathbf{z}^T\mathbf{M}_t\mathbf{z}-\mathbf{z}^T\mathbf{M}_t\mathbf{s}-\mathbf{s}^T\mathbf{M}_t\mathbf{z}+\mathbf{s}^T\mathbf{M}_t\mathbf{s}+\mathbf{a}^T\mathbf{H}_t\mathbf{a}-\alpha_t\log\sqrt{\left|2\pi\boldsymbol{\Sigma}_{a,t}^q\right|} \\
&\quad -\frac{\alpha_t}{2}\Big(\mathbf{a}^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{a}-\mathbf{a}^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{K}_t^q\mathbf{s}-\mathbf{a}^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q-\mathbf{s}^T(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{a} \\
&\quad +\mathbf{s}^T(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{K}_t^q\mathbf{s}+\mathbf{s}^T(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q-(\mathbf{k}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{a} \\
&\quad +(\mathbf{k}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{K}_t^q\mathbf{s}+(\mathbf{k}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q\Big) \\
&= \mathbf{s}^T\mathbf{R}_{ss,t}\mathbf{s}+\mathbf{a}^T\mathbf{R}_{aa,t}\mathbf{a}+\mathbf{a}^T\mathbf{R}_{sa,t}^T\mathbf{s}+\mathbf{s}^T\mathbf{R}_{sa,t}\mathbf{a}+\mathbf{s}^T\mathbf{r}_{s,t}+\mathbf{a}^T\mathbf{r}_{a,t}+r_{0,t} \\
\mathbf{R}_{ss,t} &= \mathbf{M}_t-\frac{\alpha_t}{2}(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{K}_t^q \\
\mathbf{R}_{aa,t} &= \mathbf{H}_t-\frac{\alpha_t}{2}(\boldsymbol{\Sigma}_{a,t}^q)^{-1} \\
\mathbf{R}_{sa,t} &= \frac{\alpha_t}{2}(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1} \\
\mathbf{r}_{s,t} &= -\alpha_t(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q-2\mathbf{M}_t\mathbf{z} \\
\mathbf{r}_{a,t} &= \alpha_t(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q \\
r_{0,t} &= \mathbf{z}^T\mathbf{M}_t\mathbf{z}-\alpha_t\log\sqrt{\left|2\pi\boldsymbol{\Sigma}_{a,t}^q\right|}-\frac{\alpha_t}{2}(\mathbf{k}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q
\end{aligned}
$$

Compute Q-Function:

$$
\begin{aligned}
Q_t(\mathbf{s},\mathbf{a}) &= \frac{1}{\alpha_t}\Big(r_t(\mathbf{s},\mathbf{a}) + \mathbb{E}_{\mathcal{P}}[V_{t+1}(\mathbf{s}')]\Big) \\
&= \frac{1}{\alpha_t}\Big(\mathbf{s}^T\mathbf{R}_{ss,t}\mathbf{s} + \mathbf{a}^T\mathbf{R}_{aa,t}\mathbf{a} + \mathbf{a}^T\mathbf{R}_{sa,t}^T\mathbf{s} + \mathbf{s}^T\mathbf{R}_{sa,t}\mathbf{a} + \mathbf{s}^T\mathbf{r}_{s,t} + \mathbf{a}^T\mathbf{r}_{a,t} + r_{0,t} + \mathbb{E}_{\mathcal{P}}[V_{t+1}(\mathbf{s}')]\Big) \\
&= \frac{1}{\alpha_t}\Big(\mathbf{s}^T\mathbf{R}_{ss,t}\mathbf{s} + \mathbf{a}^T\mathbf{R}_{aa,t}\mathbf{a} + \mathbf{a}^T\mathbf{R}_{sa,t}^T\mathbf{s} + \mathbf{s}^T\mathbf{R}_{sa,t}\mathbf{a} + \mathbf{s}^T\mathbf{r}_{s,t} + \mathbf{a}^T\mathbf{r}_{a,t} + r_{0,t} \\
&\quad +(\mathbf{A}_t\mathbf{s} + \mathbf{b}_t\mathbf{a} + \mathbf{c}_t)^T\mathbf{V}_{t+1}(\mathbf{A}_t\mathbf{s} + \mathbf{b}_t\mathbf{a} + \mathbf{c}_t) + \mathbf{Tr}(\mathbf{V}_{t+1}\mathbf{\Sigma}_{s',t}) + \mathbf{v}_{t+1}^T(\mathbf{A}_t\mathbf{s} + \mathbf{b}_t\mathbf{a} + \mathbf{c}_t) - \nu_{t+1}\Big) \\
&= \frac{1}{\alpha_t}\Big(\mathbf{s}^T\mathbf{R}_{ss,t}\mathbf{s} + \mathbf{a}^T\mathbf{R}_{aa,t}\mathbf{a} + \mathbf{a}^T\mathbf{R}_{sa,t}^T\mathbf{s} + \mathbf{s}^T\mathbf{R}_{sa,t}\mathbf{a} + \mathbf{s}^T\mathbf{r}_{s,t} + \mathbf{a}^T\mathbf{r}_{a,t} + r_{0,t} + \mathbf{Tr}(\mathbf{V}_{t+1}\mathbf{\Sigma}_{s',t}) \\
&\quad +\mathbf{s}^T\mathbf{A}_t^T\mathbf{V}_{t+1}\mathbf{A}_t\mathbf{s} + \mathbf{s}^T\mathbf{A}_t^T\mathbf{V}_{t+1}\mathbf{b}_t\mathbf{a} + \mathbf{s}^T\mathbf{A}_t^T\mathbf{V}_{t+1}\mathbf{c}_t + \mathbf{a}^T\mathbf{b}_t^T\mathbf{V}_{t+1}\mathbf{A}_t\mathbf{s} + \mathbf{a}^T\mathbf{b}_t^T\mathbf{V}_{t+1}\mathbf{b}_t\mathbf{a} + \mathbf{a}^T\mathbf{b}_t^T\mathbf{V}_{t+1}\mathbf{c}_t \\
&\quad +\mathbf{c}_t^T\mathbf{V}_{t+1}\mathbf{A}_t\mathbf{s} + \mathbf{c}_t^T\mathbf{V}_{t+1}\mathbf{b}_t\mathbf{a} + \mathbf{c}_t^T\mathbf{V}_{t+1}\mathbf{c}_t + \mathbf{v}_{t+1}^T\mathbf{A}_t\mathbf{s} + \mathbf{v}_{t+1}^T\mathbf{b}_t\mathbf{a} + \mathbf{v}_{t+1}^T\mathbf{c}_t + \nu_{t+1}\Big) \\
&= \mathbf{s}^T\mathbf{Q}_{ss,t}\mathbf{s} + \mathbf{a}^T\mathbf{Q}_{aa,t}\mathbf{a} + \mathbf{s}^T\mathbf{Q}_{s,t} + \mathbf{a}^T\mathbf{Q}_{a,t} + \mathbf{s}^T\mathbf{Q}_{sa,t}\mathbf{a} + \mathbf{a}^T\mathbf{Q}_{sa,t}^T\mathbf{s} + Q_{0,t} \\
\mathbf{Q}_{ss,t} &= \frac{1}{\alpha_t}(\mathbf{R}_{ss,t} + \mathbf{A}_t^T\mathbf{V}_{t+1}\mathbf{A}_t) \\
\mathbf{Q}_{aa,t} &= \frac{1}{\alpha_t}(\mathbf{R}_{aa,t} + \mathbf{b}_t^T\mathbf{V}_{t+1}\mathbf{b}_t) \\
\mathbf{Q}_{sa,t} &= \frac{1}{\alpha_t}(\mathbf{R}_{sa,t} + \mathbf{A}_t^T\mathbf{V}_{t+1}\mathbf{b}_t) \\
\mathbf{Q}_{s,t} &= \frac{1}{\alpha_t}(\mathbf{r}_{s,t} + 2\mathbf{A}_t^T\mathbf{V}_{t+1}\mathbf{c}_t + \mathbf{A}_t^T\mathbf{v}_{t+1}) \\
\mathbf{Q}_{a,t} &= \frac{1}{\alpha_t}(\mathbf{r}_{a,t} + 2\mathbf{b}_t^T\mathbf{V}_{t+1}\mathbf{c}_t + \mathbf{b}_t^T\mathbf{v}_{t+1}) \\
Q_{0,t} &= \frac{1}{\alpha_t}\Big(\mathbf{c}_t^T\mathbf{V}_{t+1}\mathbf{c}_t + \mathbf{Tr}(\mathbf{V}_{t+1}\mathbf{\Sigma}_{s',t}) + \mathbf{v}_{t+1}^T\mathbf{c}_t + \nu_{t+1} + r_{0,t}\Big) \\
\mathbf{x} &= \begin{pmatrix}\mathbf{s}\\ \mathbf{a}\end{pmatrix} \qquad \mathbf{W}_t = \begin{pmatrix}-2\mathbf{Q}_{ss,t} & -2\mathbf{Q}_{sa,t}\\ -2\mathbf{Q}_{sa,t}^T & -2\mathbf{Q}_{aa,t}\end{pmatrix} \qquad \mathbf{w}_t = \begin{pmatrix}\mathbf{Q}_{s,t}\\ \mathbf{Q}_{a,t}\end{pmatrix} \\
Q_t(\mathbf{s},\mathbf{a}) &= -\frac{1}{2}\mathbf{x}^T\mathbf{W}_t\mathbf{x} + \mathbf{x}^T\mathbf{w}_t + w_t
\end{aligned}
$$

Compute V-Function:

$$\begin{aligned}\exp\Big[Q(\mathbf{s},\mathbf{a})\Big] &= \exp\Big[-\frac{1}{2}\mathbf{x}^T\mathbf{W}_t\mathbf{x}+\mathbf{x}^T\mathbf{w}_t+w_t\Big] \\ &= \frac{\sqrt{\Big|2\pi\mathbf{W}_t^{-1}\Big|}}{\exp\Big[-\frac{1}{2}\mathbf{w}_t^T\mathbf{W}_t^{-1}\mathbf{w}_t-w_t\Big]}\mathcal{N}[\mathbf{x}|\mathbf{w}_t,\mathbf{W}_t]\end{aligned}$$

$$\begin{aligned}V_t(\mathbf{s}) &= \alpha_t\log\int_{\mathbf{a}}\exp\Big[Q_t(\mathbf{s},\mathbf{a})\Big]d\mathbf{a} \\ &= \alpha_t\log\Big[\frac{\sqrt{\Big|2\pi\mathbf{W}_t^{-1}\Big|}}{\exp\Big[-\frac{1}{2}\mathbf{w}_t^T\mathbf{W}_t^{-1}\mathbf{w}_t-w_t\Big]}\int_{\mathbf{a}}\mathcal{N}[\mathbf{x}|\mathbf{w}_t,\mathbf{W}_t]d\mathbf{a}\Big] \\ &= \alpha_t\log\Big[\frac{\sqrt{\Big|2\pi\mathbf{W}_t^{-1}\Big|}}{\exp\Big[-\frac{1}{2}\mathbf{w}_t^T\mathbf{W}_t^{-1}\mathbf{w}_t-w_t\Big]}\cdots \\ &\quad \cdots\int_{\mathbf{a}}\mathcal{N}[\mathbf{s}|\mathbf{Q}_{s,t}-\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t},-2\mathbf{Q}_{ss,t}+2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T]\mathcal{N}[\mathbf{a}|\mathbf{Q}_{a,t}+2\mathbf{Q}_{sa,t}^T\mathbf{s},-2\mathbf{Q}_{aa,t}]d\mathbf{a}\Big] \\ &= \alpha_t\log\Big[\frac{\sqrt{\Big|2\pi\mathbf{W}_t^{-1}\Big|}}{\exp\Big[-\frac{1}{2}\mathbf{w}_t^T\mathbf{W}_t^{-1}\mathbf{w}_t-w_t\Big]}\mathcal{N}[\mathbf{s}|\mathbf{Q}_{s,t}-\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t},-2\mathbf{Q}_{ss,t}+2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T] \\ &= \alpha_t\log\Big[\frac{\sqrt{\Big|2\pi\mathbf{W}_t^{-1}\Big|}}{\exp\Big[-\frac{1}{2}\mathbf{w}_t^T\mathbf{W}_t^{-1}\mathbf{w}_t-w_t\Big]}\cdots \\ &\quad \ldots\mathcal{N}(\mathbf{s}|(-2\mathbf{Q}_{ss,t}+2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)^{-1}(\mathbf{Q}_{s,t}-\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t}),(-2\mathbf{Q}_{ss,t}+2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)^{-1})\Big] \\ &= \alpha_t\log\frac{\sqrt{\Big|2\pi\mathbf{W}_t^{-1}\Big|}}{\exp\Big[-\frac{1}{2}\mathbf{w}_t^T\mathbf{W}_t^{-1}\mathbf{w}_t-w_t\Big]}-\alpha_t\log\sqrt{\Big|2\pi(-2\mathbf{Q}_{ss,t}+2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)^{-1}\Big|} \\ &\quad -\frac{\alpha_t}{2}\mathbf{s}^T(-2\mathbf{Q}_{ss,t}+2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)\mathbf{s}+\alpha_t\mathbf{s}^T(\mathbf{Q}_{s,t}-\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t}) \\ &\quad -\frac{\alpha_t}{2}(\mathbf{Q}_{s,t}-\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t})^T(-2\mathbf{Q}_{ss,t}+2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)^{-1}(\mathbf{Q}_{s,t}-\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t}) \\ &= \mathbf{s}^T\mathbf{V}_t\mathbf{s}+\mathbf{s}^T\mathbf{v}_t+v_t\end{aligned}$$

$$\begin{aligned}\mathbf{V}_t &= \alpha_t(\mathbf{Q}_{ss,t}-\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T) \\ \mathbf{v}_t &= \frac{\alpha_t}{2}(\mathbf{Q}_{s,t}-\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t})\end{aligned}$$

$$
\begin{aligned}
\nu_t &= \alpha_t \log \frac{\sqrt{\left|2\pi \mathbf{W}_t^{-1}\right|}}{\exp\left[-\frac{1}{2}\mathbf{w}_t^T \mathbf{W}_t^{-1}\mathbf{w}_t - w_t\right]} - \alpha_t \log \sqrt{\left|2\pi(-2\mathbf{Q}_{ss,t} + 2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)^{-1}\right|} \\
&\quad -\frac{\alpha_t}{2}(\mathbf{Q}_{s,t} - \mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t})^T(-2\mathbf{Q}_{ss,t} + 2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)^{-1}(\mathbf{Q}_{s,t} - \mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t}) \\
&= -\alpha_t \log \frac{\sqrt{(2\pi)^{N_\mathbf{s}}\left|(-2\mathbf{Q}_{ss,t} + 2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)^{-1}\right|}}{\sqrt{(2\pi)^{N_\mathbf{a}+N_\mathbf{s}}\left|\mathbf{W}_t^{-1}\right|}} + \frac{\alpha_t}{2}\mathbf{w}_t^T\mathbf{W}_t^{-1}\mathbf{w}_t + \alpha_t w_t + \mathbf{v}_t^T\mathbf{V}_t^{-1}\mathbf{v}_t \\
&= \frac{\alpha_t N_\mathbf{a}}{2}\log(2\pi) - \frac{\alpha_t}{2}\log\frac{\left|\frac{\alpha_t}{2}\mathbf{V}_t^{-1}\right|}{\left|\mathbf{W}_t^{-1}\right|} + \frac{\alpha_t}{2}\mathbf{w}_t^T\mathbf{W}_t^{-1}\mathbf{w}_t + \alpha_t w_t + \mathbf{v}_t^T\mathbf{V}_t^{-1}\mathbf{v}_t \\
&= \frac{\alpha_t}{2}\begin{pmatrix}\mathbf{Q}_{s,t}\\ \mathbf{Q}_{a,t}\end{pmatrix}^T\begin{pmatrix}-2\mathbf{Q}_{ss,t} & -2\mathbf{Q}_{sa,t}\\ -2\mathbf{Q}_{sa,t}^T & -2\mathbf{Q}_{aa,t}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{Q}_{s,t}\\ \mathbf{Q}_{a,t}\end{pmatrix} + \alpha_t w_t + \mathbf{v}_t^T\mathbf{V}_t^{-1}\mathbf{v}_t + \frac{\alpha_t N_\mathbf{a}}{2}\log(2\pi) - \frac{\alpha_t}{2}\log\frac{\left|\frac{\alpha_t}{2}\mathbf{V}_t^{-1}\right|}{\left|\mathbf{W}_t^{-1}\right|} \\
&= \begin{matrix}\frac{\alpha_t}{2}\end{matrix}\begin{pmatrix}\mathbf{Q}_{s,t}\\ \mathbf{Q}_{a,t}\end{pmatrix}^T\begin{pmatrix}(-2\mathbf{Q}_{ss,t} + 2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)^{-1} & -(-2\mathbf{Q}_{ss,t} + 2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}\mathbf{Q}_{sa,t}^T)^{-1}\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\\ -\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T(-2\mathbf{Q}_{ss,t} + 2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)^{-1} & \mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T(-2\mathbf{Q}_{ss,t} + 2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)^{-1}\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1} - 2\mathbf{Q}_{aa,t}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{Q}_{s,t}\\ \mathbf{Q}_{a,t}\end{pmatrix} \\
&\quad + \alpha_t w_t + \mathbf{v}_t^T\mathbf{V}_t^{-1}\mathbf{v}_t + \frac{\alpha_t N_\mathbf{a}}{2}\log(2\pi) - \frac{\alpha_t}{2}\log\frac{\left|\frac{\alpha_t}{2}\mathbf{V}_t^{-1}\right|}{\left|\mathbf{W}_t^{-1}\right|} \\
&= \frac{\alpha_t}{2}\begin{pmatrix}\mathbf{Q}_{s,t}\\ \mathbf{Q}_{a,t}\end{pmatrix}^T\begin{pmatrix}\frac{\alpha_t}{2}\mathbf{V}_t^{-1} & -\frac{\alpha_t}{2}\mathbf{V}_t^{-1}\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\\ -\frac{\alpha_t}{2}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T\mathbf{V}_t^{-1} & \frac{\alpha_t}{2}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T\mathbf{V}_t^{-1}\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1} - \frac{1}{2}\mathbf{Q}_{aa,t}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{Q}_{s,t}\\ \mathbf{Q}_{a,t}\end{pmatrix} \\
&\quad + \alpha_t w_t + \mathbf{v}_t^T\mathbf{V}_t^{-1}\mathbf{v}_t + \frac{\alpha_t N_\mathbf{a}}{2}\log(2\pi) - \frac{\alpha_t}{2}\log\frac{\left|\frac{\alpha_t}{2}\mathbf{V}_t^{-1}\right|}{\left|\mathbf{W}_t^{-1}\right|} \\
&= \frac{\alpha_t}{2}\Big[\frac{\alpha_t}{2}\mathbf{Q}_{s,t}^T\mathbf{V}_t^{-1}\mathbf{Q}_{s,t} - \frac{\alpha_t}{2}\mathbf{Q}_{a,t}^T\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T\mathbf{V}_t^{-1}\mathbf{Q}_{s,t} - \frac{\alpha_t}{2}\mathbf{Q}_{s,t}^T\mathbf{V}_t^{-1}\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t} \\
&\quad + \frac{\alpha_t}{2}\mathbf{Q}_{a,t}^T\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T\mathbf{V}_t^{-1}\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t} - \frac{1}{2}\mathbf{Q}_{a,t}^T\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t}\Big] \\
&\quad + \alpha_t w_t + \mathbf{v}_t^T\mathbf{V}_t^{-1}\mathbf{v}_t + \frac{\alpha_t N_\mathbf{a}}{2}\log(2\pi) - \frac{\alpha_t}{2}\log\frac{\left|\frac{\alpha_t}{2}\mathbf{V}_t^{-1}\right|}{\left|\mathbf{W}_t^{-1}\right|} \\
&= -\frac{\alpha_t}{4}\mathbf{Q}_{a,t}^T\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t} + \frac{\alpha_t}{2}\Big[\frac{\alpha_t}{2}(\mathbf{Q}_{s,t}^T - \mathbf{Q}_{a,t}^T\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)\mathbf{V}_t^{-1}\mathbf{Q}_{s,t} - \frac{\alpha_t}{2}(\mathbf{Q}_{s,t}^T - \mathbf{Q}_{a,t}^T\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)\mathbf{V}_t^{-1}\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t}\Big] \\
&\quad + \alpha_t w_t + \mathbf{v}_t^T\mathbf{V}_t^{-1}\mathbf{v}_t + \frac{\alpha_t N_\mathbf{a}}{2}\log(2\pi) - \frac{\alpha_t}{2}\log\frac{\left|\frac{\alpha_t}{2}\mathbf{V}_t^{-1}\right|}{\left|\mathbf{W}_t^{-1}\right|}
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{\alpha_t}{4}\mathbf{Q}_{a,t}^T\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t} + \frac{\alpha_t}{2}\mathbf{v}_t^T\mathbf{V}_t^{-1}\mathbf{Q}_{s,t} - \frac{\alpha_t}{2}\mathbf{v}_t^T\mathbf{V}_t^{-1}\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t} + \alpha_t w_t + \mathbf{v}_t^T\mathbf{V}_t^{-1}\mathbf{v}_t \\
&\quad + \frac{\alpha_t N_{\mathbf{a}}}{2}\log(2\pi) - \frac{\alpha_t}{2}\log\frac{\left|\frac{\alpha_t}{2}\mathbf{V}_t^{-1}\right|}{\left|\mathbf{W}_t^{-1}\right|} \\
&= -\frac{\alpha_t}{4}\mathbf{Q}_{a,t}^T\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t} - \mathbf{v}_t^T\mathbf{V}_t^{-1}\mathbf{v}_t + \alpha_t w_t + \mathbf{v}_t^T\mathbf{V}_t^{-1}\mathbf{v}_t - \frac{\alpha_t N_{\mathbf{a}}}{2}\log(2\pi) + \frac{\alpha_t}{2}\log\frac{\left|\frac{\alpha_t}{2}\mathbf{V}_t^{-1}\right|}{\left|\mathbf{W}_t^{-1}\right|} \\
v_t &= -\frac{\alpha_t}{4}\mathbf{Q}_{a,t}^T\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t} + \alpha_t w_t + \frac{\alpha_t N_{\mathbf{a}}}{2}\log(2\pi) - \frac{\alpha_t}{2}\log\frac{\left|\frac{\alpha_t}{2}\mathbf{V}_t^{-1}\right|}{\left|\mathbf{W}_t^{-1}\right|} \\
\pi(\mathbf{a}|\mathbf{s}) &= \frac{q_t(\mathbf{a}|\mathbf{s})\exp\left[\frac{1}{\alpha_t}\left(R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\right)\right]}{\int_{\mathbf{a}} q_t(\mathbf{a}|\mathbf{s})\exp\left[\frac{1}{\alpha_t}\left(R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\right)\right]d\mathbf{a}} \\
&= \frac{\exp\left[Q_t(\mathbf{s},\mathbf{a})\right]}{\int_{\mathbf{a}}\exp\left[Q_t(\mathbf{s},\mathbf{a})\right]d\mathbf{a}} \\
&= \frac{\frac{\sqrt{\left|2\pi\mathbf{W}_t^{-1}\right|}}{\exp\left[-\frac{1}{2}\mathbf{w}_t^T\mathbf{W}_t^{-1}\mathbf{w}_t - w_t\right]}\mathcal{N}[\mathbf{x}|\mathbf{w}_t,\mathbf{W}_t]}{\frac{\sqrt{\left|2\pi\mathbf{W}_t^{-1}\right|}}{\exp\left[-\frac{1}{2}\mathbf{w}_t^T\mathbf{W}_t^{-1}\mathbf{w}_t - w_t\right]}\mathcal{N}[\mathbf{x}|\mathbf{w}_t,\mathbf{W}_t]d\mathbf{a}} \\
&= \frac{\mathcal{N}[\mathbf{x}|\mathbf{w}_t,\mathbf{W}_t]}{\int_{\mathbf{a}}\mathcal{N}[\mathbf{x}|\mathbf{w}_t,\mathbf{W}_t]d\mathbf{a}} \\
&= \frac{\mathcal{N}[\mathbf{s}|\mathbf{Q}_{s,t} - \mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t}, -2\mathbf{Q}_{ss,t} + 2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T]\mathcal{N}[\mathbf{a}|\mathbf{Q}_{a,t} + 2\mathbf{Q}_{sa,t}^T\mathbf{s}, -2\mathbf{Q}_{aa,t}]}{\mathcal{N}(\mathbf{s}|(-2\mathbf{Q}_{ss,t} + 2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)^{-1}(\mathbf{Q}_{s,t} - \mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t}), (-2\mathbf{Q}_{ss,t} + 2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T)^{-1})} \\
&= \frac{\mathcal{N}[\mathbf{s}|\mathbf{Q}_{s,t} - \mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t}, -2\mathbf{Q}_{ss,t} + 2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T]\mathcal{N}[\mathbf{a}|\mathbf{Q}_{a,t} + 2\mathbf{Q}_{sa,t}^T\mathbf{s}, -2\mathbf{Q}_{aa,t}]}{\mathcal{N}[\mathbf{s}|\mathbf{Q}_{s,t} - \mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t}, -2\mathbf{Q}_{ss,t} + 2\mathbf{Q}_{sa,t}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T]} \\
&= \mathcal{N}[\mathbf{a}|\mathbf{Q}_{a,t} + 2\mathbf{Q}_{sa,t}^T\mathbf{s}, -2\mathbf{Q}_{aa,t}] \\
&= \mathcal{N}(\mathbf{a}|-\frac{1}{2}\mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{a,t} - \mathbf{Q}_{aa,t}^{-1}\mathbf{Q}_{sa,t}^T\mathbf{s}, -\frac{1}{2}\mathbf{Q}_{aa,t}^{-1}) \\
&= \mathcal{N}(\mathbf{a}|\mathbf{k}_t^\pi + \mathbf{K}_t^\pi\mathbf{s}, \mathbf{\Sigma}_{a,t}^\pi)
\end{aligned}
$$

$$
\begin{aligned}
\int_{\mathbf{s}} \mu_t(\mathbf{s}) D_{\mathrm{KL}}(\pi_t(\mathbf{a}|\mathbf{s})||q_t(\mathbf{a}|\mathbf{s}))d\mathbf{s} &= \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{a}|\mathbf{s})} d\mathbf{a}d\mathbf{s} \\
&= \int_{\mathbf{s}} \mathcal{N}(\mathbf{s}|\boldsymbol{\tau}^{\mu}_{s,t}, \boldsymbol{\Sigma}^{\mu}_{s,t}) \Big(\frac{1}{2} \log \frac{|\boldsymbol{\Sigma}^{q}_{a,t}|}{|\boldsymbol{\Sigma}^{\pi}_{a,t}|} + \frac{1}{2} \mathbf{Tr}\big((\boldsymbol{\Sigma}^{q}_{a,t})^{-1} \boldsymbol{\Sigma}^{\pi}_{a,t}\big) - \frac{1}{2} N_{\mathbf{a}} \\
&\quad + \frac{1}{2} \Big((\mathbf{K}^{q}_{t} - \mathbf{K}^{\pi}_{t}) s - (-\mathbf{k}^{q}_{t} + \mathbf{k}^{\pi}_{t}) \Big)^{T} (\boldsymbol{\Sigma}^{q}_{a,t})^{-1} \Big((\mathbf{K}^{q}_{t} - \mathbf{K}^{\pi}_{t}) s - (-\mathbf{k}^{q}_{t} + \mathbf{k}^{\pi}_{t}) \Big) \Big) d\mathbf{s} \\
&= \int_{\mathbf{s}} \mathcal{N}(\mathbf{s}|\boldsymbol{\tau}^{\mu}_{s,t}, \boldsymbol{\Sigma}^{\mu}_{s,t}) \Big(\frac{1}{2} \log \frac{|\boldsymbol{\Sigma}^{q}_{a,t}|}{|\boldsymbol{\Sigma}^{\pi}_{a,t}|} + \frac{1}{2} \mathbf{Tr}\big((\boldsymbol{\Sigma}^{q}_{a,t})^{-1} \boldsymbol{\Sigma}^{\pi}_{a,t}\big) - \frac{1}{2} N_{\mathbf{a}} \\
&\quad + \frac{1}{2} \Big(\mathbf{s}^{T} (\mathbf{K}^{q}_{t} - \mathbf{K}^{\pi}_{t})^{T} (\boldsymbol{\Sigma}^{q}_{a,t})^{-1} (\mathbf{K}^{q}_{t} - \mathbf{K}^{\pi}_{t}) s \\
&\quad - 2\mathbf{s}^{T} (\mathbf{K}^{q}_{t} - \mathbf{K}^{\pi}_{t})^{T} (\boldsymbol{\Sigma}^{q}_{a,t})^{-1} (-\mathbf{k}^{q}_{t} + \mathbf{k}^{\pi}_{t}) \\
&\quad + (-\mathbf{k}^{q}_{t} + \mathbf{k}^{\pi}_{t})^{T} (\boldsymbol{\Sigma}^{q}_{a,t})^{-1} (-\mathbf{k}^{q}_{t} + \mathbf{k}^{\pi}_{t}) \Big) d\mathbf{s} \\
&= \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}^{q}_{a,t}|}{|\boldsymbol{\Sigma}^{\pi}_{a,t}|} + \frac{1}{2} \mathbf{Tr}\big((\boldsymbol{\Sigma}^{q}_{a,t})^{-1} \boldsymbol{\Sigma}^{\pi}_{a,t}\big) - \frac{1}{2} N_{\mathbf{a}} \\
&\quad + \frac{1}{2} \mathbf{Tr}\big((\mathbf{K}^{q}_{t} - \mathbf{K}^{\pi}_{t})^{T} (\boldsymbol{\Sigma}^{q}_{a,t})^{-1} (\mathbf{K}^{q}_{t} - \mathbf{K}^{\pi}_{t}) \boldsymbol{\Sigma}^{\mu}_{s,t} \big) \\
&\quad + \frac{1}{2} (\boldsymbol{\tau}^{\mu}_{s,t})^{T} (\mathbf{K}^{q}_{t} - \mathbf{K}^{\pi}_{t})^{T} (\boldsymbol{\Sigma}^{q}_{a,t})^{-1} (\mathbf{K}^{q}_{t} - \mathbf{K}^{\pi}_{t}) \boldsymbol{\tau}^{\mu}_{s,t} \\
&\quad - (\boldsymbol{\tau}^{\mu}_{s,t})^{T} (\mathbf{K}^{q}_{t} - \mathbf{K}^{\pi}_{t})^{T} (\boldsymbol{\Sigma}^{q}_{a,t})^{-1} (-\mathbf{k}^{q}_{t} + \mathbf{k}^{\pi}_{t}) \\
&\quad + \frac{1}{2} (-\mathbf{k}^{q}_{t} + \mathbf{k}^{\pi}_{t})^{T} (\boldsymbol{\Sigma}^{q}_{a,t})^{-1} (-\mathbf{k}^{q}_{t} + \mathbf{k}^{\pi}_{t})
\end{aligned}
$$

$$
\begin{aligned}
\mu_t(\mathbf{s}) &= \int_{\hat{\mathbf{s}}} \int_{\mathbf{a}} \pi_{t-1}(\mathbf{a}|\hat{\mathbf{s}}) \mu_{t-1}(\hat{\mathbf{s}}) \mathcal{P}_{t-1}(\mathbf{s}|\hat{\mathbf{s}}, \mathbf{a}) d\mathbf{a}d\hat{\mathbf{s}} \\
&= \int_{\hat{\mathbf{s}}} \mu_{t-1}(\hat{\mathbf{s}}) \int_{\mathbf{a}} \pi_{t-1}(\mathbf{a}|\hat{\mathbf{s}}) \mathcal{P}_{t-1}(\mathbf{s}|\hat{\mathbf{s}}, \mathbf{a}) d\mathbf{a}d\hat{\mathbf{s}} \\
\mathcal{N}(\mathbf{s}|\boldsymbol{\tau}^{\mu}_{s,t}, \boldsymbol{\Sigma}^{\mu}_{s,t}) &= \int_{\hat{\mathbf{s}}} \mathcal{N}(\hat{\mathbf{s}}|\boldsymbol{\tau}^{\mu}_{s,t-1}, \boldsymbol{\Sigma}^{\mu}_{s,t-1}) \int_{\mathbf{a}} \mathcal{N}(\mathbf{s}|\mathbf{A}_{t-1}\hat{\mathbf{s}} + \mathbf{b}_{t-1}\mathbf{a}_{t-1} + \mathbf{c}_{t-1}, \boldsymbol{\Sigma}_{s,t-1}) ... \\
&\quad ... \mathcal{N}(\mathbf{a}_{t-1}|\mathbf{k}^{\pi}_{t-1} + \mathbf{K}^{\pi}_{t-1}\hat{\mathbf{s}}, \boldsymbol{\Sigma}^{\pi}_{a,t-1}) d\mathbf{a}d\hat{\mathbf{s}} \\
&= \int_{\hat{\mathbf{s}}} \mathcal{N}(\mathbf{s}|\mathbf{A}_{t-1}\hat{\mathbf{s}} + \mathbf{c}_{t-1} + \mathbf{b}_{t-1}(\mathbf{k}^{\pi}_{t-1} + \mathbf{K}^{\pi}_{t-1}\hat{\mathbf{s}}), \boldsymbol{\Sigma}_{s,t-1} + \mathbf{b}_{t-1}\boldsymbol{\Sigma}^{\pi}_{a,t-1}\mathbf{b}^{T}_{t-1}) ... \\
&\quad ... \mathcal{N}(\hat{\mathbf{s}}|\boldsymbol{\tau}^{\mu}_{s,t-1}, \boldsymbol{\Sigma}^{\mu}_{s,t-1}) d\hat{\mathbf{s}} \\
&= \mathcal{N}(\mathbf{s}|\mathbf{c}_{t-1} + \mathbf{b}_{t-1}\mathbf{k}^{\pi}_{t-1} + (\mathbf{A}_{t-1} + \mathbf{b}_{t-1}\mathbf{K}^{\pi}_{t-1})\boldsymbol{\tau}^{\mu}_{s,t-1}, \\
&\quad \boldsymbol{\Sigma}_{s,t-1} + \mathbf{b}_{t-1}\boldsymbol{\Sigma}^{\pi}_{a,t-1}\mathbf{b}^{T}_{t-1} + (\mathbf{A}_{t-1} + \mathbf{b}_{t-1}\mathbf{K}^{\pi}_{t-1})\boldsymbol{\Sigma}^{\mu}_{s,t-1}(\mathbf{A}_{t-1} + \mathbf{b}_{t-1}\mathbf{K}^{\pi}_{t-1})^{T})
\end{aligned}
$$

$$
\begin{aligned}
V_T &= (\mathbf{z} - \mathbf{s})^{T} \mathbf{M}_{T} (\mathbf{z} - \mathbf{s}) = \mathbf{s}^{T} \mathbf{M}_{T} s - 2\mathbf{s}^{T} \mathbf{M}_{T} \mathbf{z} + \mathbf{z}^{T} \mathbf{M}_{T} \mathbf{z} \\
L(\mu_t, V_t, \alpha_t) &= \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} + \sum_{t=1}^{T-1} \alpha_t \epsilon \\
&= \int_{\mathbf{s}} \mathcal{N}(\mathbf{s}|\boldsymbol{\tau}^{\mu}_{s,1}, \boldsymbol{\Sigma}^{\mu}_{s,1}) (\mathbf{s}^{T}\mathbf{V}_1\mathbf{s} + \mathbf{s}^{T}\mathbf{v}_1 + \vartheta_1) d\mathbf{s} + \sum_{t=1}^{T-1} \alpha_t \epsilon \\
&= (\boldsymbol{\tau}^{\mu}_{s,1})^{T} \mathbf{V}_1 \boldsymbol{\tau}^{\mu}_{s,1} + (\boldsymbol{\tau}^{\mu}_{s,1})^{T} \mathbf{v}_1 + \vartheta_1 + \mathbf{Tr}(\mathbf{V}_1 \boldsymbol{\Sigma}^{\mu}_{s,1}) + \sum_{t=1}^{T-1} \alpha_t \epsilon
\end{aligned}
$$

B Derivation of State-Action Bound Policy Search

$$\underset{\pi_t(\mathbf{a}|\mathbf{s})}{\operatorname{argmax}} \quad \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s}$$

$$\forall \mathbf{s}, \forall t < T \quad \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} = 1$$

$$\forall \mathbf{s}', \forall t > 1 \quad \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_{t-1}(\mathbf{s}) \pi_{t-1}(\mathbf{a}|\mathbf{s}) \mathcal{P}_{t-1}(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} = \mu_t(\mathbf{s}')$$

$$\forall \mathbf{s}, t = 1 \quad \mu_1(\mathbf{s}) = p_1(\mathbf{s})$$

$$\forall t < T \quad \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s},\mathbf{a})} d\mathbf{a} d\mathbf{s} \leq \epsilon$$

Primal Problem:

$$\begin{aligned}
L(\pi_t, \mu_t, V_t, \lambda_t, \alpha_t) &= \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) \Big[1 - \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} \Big] d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) \Big[p_1(\mathbf{s}) - \mu_1(\mathbf{s}) \Big] d\mathbf{s} \\
&\quad + \sum_{t=2}^{T} \int_{\mathbf{s}'} V_t(\mathbf{s}') \Big(\int_{\mathbf{s}} \int_{\mathbf{a}} \mu_{t-1}(\mathbf{s}) \pi_{t-1}(\mathbf{a}|\mathbf{s}) \mathcal{P}_{t-1}(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} - \mu_t(\mathbf{s}') \Big) d\mathbf{s}' \\
&\quad + \sum_{t=1}^{T-1} \Big(\alpha_t \epsilon - \alpha_t \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s},\mathbf{a})} d\mathbf{a} d\mathbf{s} \Big) \\
L(\pi_t, \mu_t, V_t, \lambda_t, \alpha_t) &= \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) d\mathbf{s} - \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}') \mu_T(\mathbf{s}') d\mathbf{s}' \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} d\mathbf{s}' - \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_t(\mathbf{s}') \mu_t(\mathbf{s}') d\mathbf{s}' \\
&\quad + \sum_{t=1}^{T-1} \Big(\alpha_t \epsilon - \alpha_t \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s},\mathbf{a})} d\mathbf{a} d\mathbf{s} \Big) \\
\frac{\partial L(\pi_t, \mu_t, V_t, \lambda_t, \alpha_t)}{\partial \pi_t} &= R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) - \lambda_t(\mathbf{s}) + \int_{\mathbf{s}'} \mu_t(\mathbf{s}) V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' \\
&\quad - \alpha_t \mu_t(\mathbf{s}) \log \frac{\mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s},\mathbf{a})} - \alpha_t \mu_t(\mathbf{s}) = 0 \\
\Rightarrow \pi_t(\mathbf{a}|\mathbf{s}) &= \frac{q_t(\mathbf{s},\mathbf{a})}{\mu_t(\mathbf{s})} \exp\Big[\frac{1}{\alpha_t} \Big(R_t(\mathbf{s},\mathbf{a}) - \frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' - \alpha_t \Big) \Big]
\end{aligned}$$

Dual Problem:

$$
\begin{aligned}
L(\mu_t, V_t, \lambda_t, \alpha_t) &= \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{s} d\mathbf{a} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) d\mathbf{s} - \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}') \mu_T(\mathbf{s}') d\mathbf{s}' \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \mathcal{P}(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} d\mathbf{s}' - \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_t(\mathbf{s}') \mu_t(\mathbf{s}') d\mathbf{s}' \\
&\quad + \sum_{t=1}^{T-1} \Big(\alpha_t \epsilon - \alpha_t \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \ldots \\
&\quad \ldots \log \frac{q_t(\mathbf{s},\mathbf{a}) \exp\Big[\frac{1}{\alpha_t} \Big(R_t(\mathbf{s},\mathbf{a}) - \frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' - \alpha_t \Big) \Big]}{q_t(\mathbf{s},\mathbf{a})} d\mathbf{a} d\mathbf{s} \Big) \\
&= \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} + \sum_{t=1}^{T-1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}') \mu_T(\mathbf{s}') d\mathbf{s}' \\
&\quad - \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_t(\mathbf{s}') \mu_t(\mathbf{s}') d\mathbf{s}' + \sum_{t=1}^{T-1} \alpha_t (\epsilon + 1)
\end{aligned}
$$

Solve for λ_t:

$$
\begin{aligned}
1 &= \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} \\
1 &= \int_{\mathbf{a}} \frac{q_t(\mathbf{s},\mathbf{a})}{\mu_t(\mathbf{s})} \exp\Big[\frac{1}{\alpha_t} \Big(R_t(\mathbf{s},\mathbf{a}) - \frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' - \alpha_t \Big) \Big] d\mathbf{a} \\
1 &= \int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t} \Big(\alpha_t \log q_t(\mathbf{s},\mathbf{a}) - \alpha_t \log \mu_t(\mathbf{s}) + R_t(\mathbf{s},\mathbf{a}) - \frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} \\
&\quad + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' - \alpha_t \Big) \Big] d\mathbf{a} \\
1 &= \int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t} \Big(\alpha_t \log q_t(\mathbf{s},\mathbf{a}) + R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' \Big) d\mathbf{s}' \Big] \ldots \\
&\quad \ldots \exp\Big[- \frac{\lambda_t(\mathbf{s})}{\alpha_t \mu_t(\mathbf{s})} - 1 - \log \mu_t(\mathbf{s}) \Big] d\mathbf{a} \\
\exp\Big[\frac{\lambda_t(\mathbf{s})}{\alpha_t \mu_t(\mathbf{s})} + 1 + \log \mu_t(\mathbf{s}) \Big] &= \int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t} \Big(\alpha_t \log q_t(\mathbf{s},\mathbf{a}) + R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' \Big) d\mathbf{s}' \Big] d\mathbf{a} \\
\lambda_t(\mathbf{s}) &= \alpha_t \mu_t(\mathbf{s}) \Big[-1 - \log \mu_t(\mathbf{s}) + \log \int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t} \Big(\alpha_t \log q_t(\mathbf{s},\mathbf{a}) + R_t(\mathbf{s},\mathbf{a}) \\
&\quad + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' \Big) \Big] d\mathbf{a} \Big]
\end{aligned}
$$

$$
\begin{aligned}
L(\mu_t, V_t, \alpha_t) &= \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}') \mu_T(\mathbf{s}') d\mathbf{s}' \\
&\quad - \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_t(\mathbf{s}') \mu_t(\mathbf{s}') d\mathbf{s}' + \sum_{t=1}^{T-1} \alpha_t \epsilon - \sum_{t=1}^{T-1} \int_{\mathbf{s}} \alpha_t \mu_t(\mathbf{s}) \log \mu_t(\mathbf{s}) d\mathbf{s} \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}} \alpha_t \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t \log q_t(\mathbf{s}, \mathbf{a}) + R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a}) d\mathbf{s}'\Big)\Big] d\mathbf{a} d\mathbf{s}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial \mu_t} &= -V_t(\mathbf{s}) - \alpha_t \log \mu_t(\mathbf{s}) - \alpha_t + \alpha_t \log \int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t \log q_t(\mathbf{s}, \mathbf{a}) + R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a}) d\mathbf{s}'\Big)\Big] d\mathbf{a} \\
0 &= -V_t(\mathbf{s}) - \alpha_t + \alpha_t \log \int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t \log q_t(\mathbf{s}, \mathbf{a}) - \alpha_t \log \mu_t(\mathbf{s}) + R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a}) d\mathbf{s}'\Big)\Big] d\mathbf{a} \\
0 &= -V_t(\mathbf{s}) + \alpha_t \log \int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t \log q_t(\mathbf{s}, \mathbf{a}) - \alpha_t \log \mu_t(\mathbf{s}) - \alpha_t + R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a}) d\mathbf{s}'\Big)\Big] d\mathbf{a} \\
\frac{\partial L}{\partial \mu_T} &= R_T(\mathbf{s}) - V_T(\mathbf{s})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial V_t} &= -\mu_t(\mathbf{s}) + \int_{\hat{\mathbf{s}}} \mu_{t-1}(\hat{\mathbf{s}}) \int_{\mathbf{a}} \frac{\exp\Big[\frac{1}{\alpha_{t-1}}\Big(\alpha_{t-1} \log q_{t-1}(\hat{\mathbf{s}}, \mathbf{a}) + R_{t-1}(\hat{\mathbf{s}}, \mathbf{a}) + \int_{\mathbf{s}'} V_t(\mathbf{s}') \mathcal{P}_{t-1}(\mathbf{s}'|\hat{\mathbf{s}}, \mathbf{a}) d\mathbf{s}'\Big)\Big]}{\int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_{t-1}}\Big(\alpha_{t-1} \log q_{t-1}(\hat{\mathbf{s}}, \mathbf{a}) + R_{t-1}(\hat{\mathbf{s}}, \mathbf{a}) + \int_{\mathbf{s}'} V_t(\mathbf{s}') \mathcal{P}_{t-1}(\mathbf{s}'|\hat{\mathbf{s}}, \mathbf{a}) d\mathbf{s}'\Big)\Big] d\mathbf{a}} \mathcal{P}_{t-1}(\mathbf{s}|\hat{\mathbf{s}}, \mathbf{a}) d\mathbf{a} d\hat{\mathbf{s}} \\
0 &= -\mu_t(\mathbf{s}) + \int_{\hat{\mathbf{s}}} \int_{\mathbf{a}} \pi_{t-1}(\mathbf{a}|\hat{\mathbf{s}}) \mu_{t-1}(\hat{\mathbf{s}}) \mathcal{P}_{t-1}(\mathbf{s}|\hat{\mathbf{s}}, \mathbf{a}) d\mathbf{a} d\hat{\mathbf{s}} \\
\frac{\partial L}{\partial V_1} &= p_1(\mathbf{s}) - \mu_1(\mathbf{s})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial \alpha_t} &= \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t \log q_t(\mathbf{s}, \mathbf{a}) + R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a}) d\mathbf{s}'\Big)\Big] d\mathbf{a} d\mathbf{s} - \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \mu_t(\mathbf{s}) d\mathbf{s} \\
&\quad \epsilon - \int_{\mathbf{s}} \mu_t(\mathbf{s}) \alpha_t \int_{\mathbf{a}} \frac{\exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t \log q_t(\mathbf{s}, \mathbf{a}) + R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a}) d\mathbf{s}'\Big)\Big]}{\int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t \log q_t(\mathbf{s}, \mathbf{a}) + R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a}) d\mathbf{s}'\Big)\Big] d\mathbf{a}} \ldots \\
&\quad \ldots \frac{1}{\alpha_t^2}\Big(R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a}) d\mathbf{s}'\Big) d\mathbf{a} d\mathbf{s} \\
0 &= \epsilon + \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t \log q_t(\mathbf{s}, \mathbf{a}) + R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a}) d\mathbf{s}'\Big)\Big] d\mathbf{a} d\mathbf{s} \\
&\quad - \frac{1}{\alpha_t} \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \Big(R_t(\mathbf{s}, \mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s}, \mathbf{a}) d\mathbf{s}'\Big) d\mathbf{a} d\mathbf{s} - \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \mu_t(\mathbf{s}) d\mathbf{s} \\
&= \epsilon - \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s}, \mathbf{a})} d\mathbf{a} d\mathbf{s}
\end{aligned}
$$

$$
L(\mu_t, V_t, \alpha_t) = \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} + \sum_{t=1}^{T-1} \alpha_t (\epsilon + 1)
$$

Plug in Gaussians:

$$
\begin{aligned}
\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) &= \mathcal{N}(\mathbf{s}'|\mathbf{A}_t\mathbf{s}+\mathbf{b}_t\mathbf{a}+\mathbf{c}_t,\boldsymbol{\Sigma}_{\mathbf{s}'}) \\
q_t(\mathbf{s}) &= \mathcal{N}(\mathbf{s}|\boldsymbol{\tau}^q_{s,t},\boldsymbol{\Sigma}^q_{s,t}) \\
q_t(\mathbf{a}|\mathbf{s}) &= \mathcal{N}(\mathbf{a}|\mathbf{K}_t\mathbf{s}+\mathbf{k}_t,\boldsymbol{\Sigma}^q_{a,t}) \\
q_t(\mathbf{s},\mathbf{a}) &= q_t(\mathbf{a}|\mathbf{s})q_t(\mathbf{s}) = \mathcal{N}\Big(\begin{bmatrix}\mathbf{s}\\ \mathbf{a}\end{bmatrix}|\boldsymbol{\tau}^q_{s,a,t},\boldsymbol{\Sigma}^q_{s,a,t}\Big) \\
&= \mathcal{N}\Big(\begin{bmatrix}\mathbf{s}\\ \mathbf{a}\end{bmatrix}\Big|\begin{bmatrix}\boldsymbol{\tau}^q_{s,t}\\ \mathbf{k}_t+\mathbf{K}_t\boldsymbol{\tau}^q_{s,t}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}^q_{s,t} & (\mathbf{K}_t\boldsymbol{\Sigma}^q_{s,t})^T\\ \mathbf{K}_t\boldsymbol{\Sigma}^q_{s,t} & \boldsymbol{\Sigma}^q_{a,t}+\mathbf{K}_t(\mathbf{K}_t\boldsymbol{\Sigma}^q_{s,t})^T\end{bmatrix}\Big) \\
\mu_t(\mathbf{s}) &= \mathcal{N}(\mathbf{s}|\boldsymbol{\tau}^p_{s,t},\boldsymbol{\Sigma}^p_{s,t}) \\
R_t(\mathbf{s},\mathbf{a}) &= (\mathbf{z}-\mathbf{s})^T\mathbf{M}_t(\mathbf{z}-\mathbf{s})+\mathbf{a}^T\mathbf{H}_t\mathbf{a} \\
R_T &= (\mathbf{z}-\mathbf{s})^T\mathbf{M}_T(\mathbf{z}-\mathbf{s}) \\
\int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}' &= \mathbb{E}_{\mathcal{P}}[V_{t+1}(\mathbf{s}')] \\
V_{t+1}(\mathbf{s}) &= \mathbf{s}^T\mathbf{V}_{t+1}\mathbf{s}+\mathbf{s}^T\mathbf{v}_{t+1}+v_{t+1}
\end{aligned}
$$

$$
\begin{aligned}
r_t(\mathbf{s},\mathbf{a}) &= R_t(\mathbf{s},\mathbf{a})+\alpha_t\log q_t(\mathbf{s},\mathbf{a})-\alpha_t\log\mu_t(\mathbf{s})-\alpha_t \\
&= R_t(\mathbf{s},\mathbf{a})+\alpha_t\log q_t(\mathbf{a}|\mathbf{s})+\alpha_t\log q_t(\mathbf{s})-\alpha_t\log\mu_t(\mathbf{s})-\alpha_t \\
&= \mathbf{z}^T\mathbf{M}_t\mathbf{z}-\mathbf{z}^T\mathbf{M}_t\mathbf{s}-\mathbf{s}^T\mathbf{M}_t\mathbf{z}+\mathbf{s}^T\mathbf{M}_t\mathbf{s}+\mathbf{a}^T\mathbf{H}_t\mathbf{a}-\frac{\alpha_t}{2}\log\Big|2\pi\boldsymbol{\Sigma}^q_{a,t}\Big| \\
&\quad -\frac{\alpha_t}{2}\Big(\mathbf{a}^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{a}-\mathbf{a}^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{K}^q_t\mathbf{s}-\mathbf{a}^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t-\mathbf{s}^T(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{a}+\mathbf{s}^T(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{K}^q_t\mathbf{s} \\
&\quad +\mathbf{s}^T(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t-(\mathbf{k}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{a}+(\mathbf{k}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{K}^q_t\mathbf{s}+(\mathbf{k}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t\Big)-\frac{\alpha_t}{2}\log|2\pi\boldsymbol{\Sigma}^q_{s,t}| \\
&\quad -\frac{\alpha_t}{2}(\mathbf{s}-\boldsymbol{\tau}^q_{s,t})^T(\boldsymbol{\Sigma}^q_{s,t})^{-1}(\mathbf{s}-\boldsymbol{\tau}^q_{s,t})-\alpha_t+\alpha_t\Big(\frac{1}{2}\log|2\pi\boldsymbol{\Sigma}^p_{s,t}|+\frac{1}{2}(\mathbf{s}-\boldsymbol{\tau}^p_{s,t})^T(\boldsymbol{\Sigma}^p_{s,t})^{-1}(\mathbf{s}-\boldsymbol{\tau}^p_{s,t})\Big) \\
&= \mathbf{z}^T\mathbf{M}_t\mathbf{z}-\mathbf{z}^T\mathbf{M}_t\mathbf{s}-\mathbf{s}^T\mathbf{M}_t\mathbf{z}+\mathbf{s}^T\mathbf{M}_t\mathbf{s}+\mathbf{a}^T\mathbf{H}_t\mathbf{a}-\frac{\alpha_t}{2}\log\Big|2\pi\boldsymbol{\Sigma}^q_{a,t}\Big| \\
&\quad -\frac{\alpha_t}{2}\Big(\mathbf{a}^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{a}-\mathbf{a}^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{K}^q_t\mathbf{s}-\mathbf{a}^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t-\mathbf{s}^T(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{a} \\
&\quad +\mathbf{s}^T(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{K}^q_t\mathbf{s}+\mathbf{s}^T(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t-(\mathbf{k}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{a} \\
&\quad +(\mathbf{k}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{K}^q_t\mathbf{s}+(\mathbf{k}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t\Big)-\frac{\alpha_t}{2}\log|2\pi\boldsymbol{\Sigma}^q_{s,t}| \\
&\quad -\frac{\alpha_t}{2}\mathbf{s}^T(\boldsymbol{\Sigma}^q_{s,t})^{-1}\mathbf{s}+\alpha_t\mathbf{s}^T(\boldsymbol{\Sigma}^q_{s,t})^{-1}\boldsymbol{\tau}^q_{s,t}-\frac{\alpha_t}{2}(\boldsymbol{\tau}^q_{s,t})^T(\boldsymbol{\Sigma}^q_{s,t})^{-1}\boldsymbol{\tau}^q_{s,t}-\alpha_t \\
&\quad +\frac{\alpha_t}{2}\Big(\log|2\pi\boldsymbol{\Sigma}^p_{s,t}|+\mathbf{s}^T(\boldsymbol{\Sigma}^p_{s,t})^{-1}\mathbf{s}-2\mathbf{s}^T(\boldsymbol{\Sigma}^p_{s,t})^{-1}\boldsymbol{\tau}^p_{s,t}+(\boldsymbol{\tau}^p_{s,t})^T(\boldsymbol{\Sigma}^p_{s,t})^{-1}\boldsymbol{\tau}^p_{s,t}\Big) \\
&= \mathbf{s}^T\mathbf{R}_{ss,t}\mathbf{s}+\mathbf{a}^T\mathbf{R}_{aa,t}\mathbf{a}+\mathbf{a}^T\mathbf{R}^T_{sa,t}\mathbf{s}+\mathbf{s}^T\mathbf{R}_{sa,t}\mathbf{a}+\mathbf{s}^T\mathbf{r}_{s,t}+\mathbf{a}^T\mathbf{r}_{a,t}+r_{0,t}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{R}_{ss,t} &= \mathbf{M}_t - \frac{\alpha_t}{2}(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{K}_t^q - \frac{\alpha_t}{2}(\boldsymbol{\Sigma}_{s,t}^q)^{-1} + \frac{\alpha_t}{2}(\boldsymbol{\Sigma}_{s,t}^p)^{-1} \\
\mathbf{R}_{aa,t} &= \mathbf{H}_t - \frac{\alpha_t}{2}(\boldsymbol{\Sigma}_{a,t}^q)^{-1} \\
\mathbf{R}_{sa,t} &= \frac{\alpha_t}{2}(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1} \\
\mathbf{r}_{s,t} &= -\alpha_t(\mathbf{K}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q + \alpha_t(\boldsymbol{\Sigma}_{s,t}^q)^{-1}\boldsymbol{\tau}_{s,t}^q - \alpha_t(\boldsymbol{\Sigma}_{s,t}^p)^{-1}\boldsymbol{\tau}_{s,t}^p - 2\mathbf{M}_t\mathbf{z} \\
\mathbf{r}_{a,t} &= \alpha_t(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q \\
r_{0,t} &= \mathbf{z}^T\mathbf{M}_t\mathbf{z} - \frac{\alpha_t}{2}\log\left|2\pi\boldsymbol{\Sigma}_{a,t}^q\right| - \frac{\alpha_t}{2}(\mathbf{k}_t^q)^T(\boldsymbol{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q \\
&\quad -\frac{\alpha_t}{2}\log|2\pi\boldsymbol{\Sigma}_{s,t}^q| - \frac{\alpha_t}{2}(\boldsymbol{\tau}_{s,t}^q)^T(\boldsymbol{\Sigma}_{s,t}^q)^{-1}\boldsymbol{\tau}_{s,t}^q - \alpha_t \\
&\quad +\frac{\alpha_t}{2}\Big(\log|2\pi\boldsymbol{\Sigma}_{s,t}^p| + (\boldsymbol{\tau}_{s,t}^p)^T(\boldsymbol{\Sigma}_{s,t}^p)^{-1}\boldsymbol{\tau}_{s,t}^p\Big)
\end{aligned}
$$

Coordinate Descent:

$$
\begin{aligned}
L(\mu_t,\alpha_t) &= \int_{\mathbf{s}}\mu_T(\mathbf{s})R_T(\mathbf{s})d\mathbf{s} + \int_{\mathbf{s}}V_1(\mathbf{s})p_1(\mathbf{s})d\mathbf{s} - \int_{\mathbf{s}'}V_T(\mathbf{s}')\mu_T(\mathbf{s}')d\mathbf{s}' \\
&\quad -\sum_{t=1}^{T-1}\int_{\mathbf{s}'}V_t(\mathbf{s}')\mu_t(\mathbf{s}')d\mathbf{s}' + \sum_{t=1}^{T-1}\alpha_t\epsilon - \sum_{t=1}^{T-1}\int_{\mathbf{s}}\alpha_t\mu_t(\mathbf{s})\log\mu_t(\mathbf{s})d\mathbf{s} \\
&\quad +\sum_{t=1}^{T-1}\int_{\mathbf{s}}\alpha_t\mu_t(\mathbf{s})\log\int_{\mathbf{a}}\exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t\log q_t(\mathbf{s},\mathbf{a}) + R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'}V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big]d\mathbf{a}d\mathbf{s} \\
\frac{\partial L(\mu_t,\alpha_t)}{\partial\mu_t} &= -V_t(\mathbf{s}) - \alpha_t\log\mu_t(\mathbf{s}) - \alpha_t \\
&\quad +\alpha_t\log\int_{\mathbf{a}}\exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t\log q_t(\mathbf{s},\mathbf{a}) + R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'}V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big]d\mathbf{a} = 0 \\
\log\mu_t(\mathbf{s}) &= -\frac{1}{\alpha_t}V_t(\mathbf{s}) - 1 + \log\int_{\mathbf{a}}\exp\Big[\frac{1}{\alpha_t}\Big(\alpha_t\log q_t(\mathbf{s},\mathbf{a}) + R_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'}V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big]d\mathbf{a} \\
&= -\frac{1}{\alpha_t}(\mathbf{s}^T\mathbf{V}_t\mathbf{s} + \mathbf{s}^T\mathbf{v}_t + v_t + \alpha_t) + \frac{1}{\alpha_t}(\mathbf{s}^T\hat{\mathbf{V}}_t\mathbf{s} + \mathbf{s}^T\hat{\mathbf{v}}_t + \hat{v}_t) \\
&= -\frac{1}{2}\Big(\mathbf{s}^T\frac{2}{\alpha_t}(\mathbf{V}_t - \hat{\mathbf{V}}_t)\mathbf{s} + \mathbf{s}^T\frac{2}{\alpha_t}(\mathbf{v}_t - \hat{\mathbf{v}}_t) + \frac{2}{\alpha_t}(v_t - \hat{v}_t + \alpha_t)\Big) \\
\log\mathcal{N}(s|a,A) &= -\frac{1}{2}\Big(\log|2\pi A| + (s-a)^TA^{-1}(s-a)\Big) \\
&= -\frac{1}{2}\Big(\log|2\pi A| + s^TA^{-1}s - 2s^TA^{-1}a + a^TA^{-1}a\Big) \\
A^{-1} &= \frac{2}{\alpha_t}(\mathbf{V}_t - \hat{\mathbf{V}}_t) \\
A^{-1}a &= -\frac{1}{\alpha_t}(\mathbf{v}_t - \hat{\mathbf{v}}_t) \\
a &= -\frac{1}{2}(\mathbf{V}_t - \hat{\mathbf{V}}_t)^{-1}(\mathbf{v}_t - \hat{\mathbf{v}}_t) \\
\mu_t(\mathbf{s}) &= \mathcal{N}(s|a,A) \\
&= \mathcal{N}(\mathbf{s}|-\frac{1}{2}(\mathbf{V}_t - \hat{\mathbf{V}}_t)^{-1}(\mathbf{v}_t - \hat{\mathbf{v}}_t), \frac{\alpha_t}{2}(\mathbf{V}_t - \hat{\mathbf{V}}_t)^{-1})
\end{aligned}
$$

C Derivation of Entropy State-Action Bound Policy Search

$$\underset{\pi_t(\mathbf{a}|\mathbf{s})}{\operatorname{argmax}} \quad \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{s} d\mathbf{a} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s}$$

$$\forall \mathbf{s}, \forall t < T \quad \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} = 1$$

$$\forall \mathbf{s}', \forall t > 1 \quad \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_{t-1}(\mathbf{s}) \pi_{t-1}(\mathbf{a}|\mathbf{s}) \mathcal{P}_{t-1}(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} = \mu_t(\mathbf{s}')$$

$$\forall \mathbf{s}, t = 1 \quad \mu_1(\mathbf{s}) = p_1(\mathbf{s})$$

$$\forall t < T \quad \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s},\mathbf{a})} d\mathbf{a} d\mathbf{s} \leq \epsilon$$

$$\forall t < T \quad \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \log \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} \leq \delta$$

Primal Problem:

$$\begin{aligned}
L(\pi_t, \mu_t, V_t, \lambda_t, \alpha_t, \beta_t) \;=\; & \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} \\
& + \sum_{t=1}^{T-1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) \Big[1 - \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} \Big] d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) \Big(p_1(\mathbf{s}) - \mu_1(\mathbf{s}) \Big) d\mathbf{s} \\
& + \sum_{t=2}^{T} \int_{\mathbf{s}'} V_t(\mathbf{s}') \Big(\int_{\mathbf{s}} \int_{\mathbf{a}} \mu_{t-1}(\mathbf{s}) \pi_{t-1}(\mathbf{a}|\mathbf{s}) \mathcal{P}_{t-1}(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} - \mu_t(\mathbf{s}') \Big) d\mathbf{s}' \\
& + \sum_{t=1}^{T-1} \Big(\alpha_t \epsilon - \alpha_t \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s},\mathbf{a})} d\mathbf{a} d\mathbf{s} \Big) \\
& + \sum_{t=1}^{T-1} \Big(\beta_t \delta - \beta_t \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \log \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} \Big)
\end{aligned}$$

$$\begin{aligned}
L(\pi_t, \mu_t, V_t, \lambda_t, \alpha_t, \beta_t) \;=\; & \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} \\
& + \sum_{t=1}^{T-1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) d\mathbf{s} - \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}') \mu_T(\mathbf{s}') d\mathbf{s}' \\
& + \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} d\mathbf{s}' - \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_t(\mathbf{s}') \mu_t(\mathbf{s}') d\mathbf{s}' \\
& + \sum_{t=1}^{T-1} \Big(\alpha_t \epsilon - \alpha_t \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s},\mathbf{a})} d\mathbf{a} d\mathbf{s} \Big) \\
& + \sum_{t=1}^{T-1} \Big(\beta_t \delta - \beta_t \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \log \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} \Big)
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial L(\pi_t, \mu_t, V_t, \lambda_t, \alpha_t, \beta_t)}{\partial \pi_t} &= R_t(\mathbf{s},\mathbf{a})\mu_t(\mathbf{s}) - \lambda_t(\mathbf{s}) + \int_{\mathbf{s}'} \mu_t(\mathbf{s}) V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' \\
&\quad -\alpha_t \mu_t(\mathbf{s}) \log \frac{\mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s},\mathbf{a})} - \alpha_t \mu_t(\mathbf{s}) - \beta_t \mu_t(\mathbf{s}) \log \pi_t(\mathbf{a}|\mathbf{s}) - \beta_t \mu_t(\mathbf{s}) \\
&= R_t(\mathbf{s},\mathbf{a})\mu_t(\mathbf{s}) - \lambda_t(\mathbf{s}) + \int_{\mathbf{s}'} \mu_t(\mathbf{s}) V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' - \alpha_t \mu_t(\mathbf{s}) \log \mu_t(\mathbf{s}) \\
&\quad -\alpha_t \mu_t(\mathbf{s}) \log \pi_t(\mathbf{a}|\mathbf{s}) + \alpha_t \mu_t(\mathbf{s}) \log q_t(\mathbf{s},\mathbf{a}) - \alpha_t \mu_t(\mathbf{s}) - \beta_t \log \pi_t(\mathbf{a}|\mathbf{s}) - \beta_t \\
(\alpha_t + \beta_t)\mu_t(\mathbf{s}) \log \pi_t(\mathbf{a}|\mathbf{s}) &= R_t(\mathbf{s},\mathbf{a})\mu_t(\mathbf{s}) - \lambda_t(\mathbf{s}) + \int_{\mathbf{s}'} \mu_t(\mathbf{s}) V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}' - \alpha_t \mu_t(\mathbf{s}) \log \mu_t(\mathbf{s}) \\
&\quad +\alpha_t \mu_t(\mathbf{s}) \log q_t(\mathbf{s},\mathbf{a}) - \alpha_t \mu_t(\mathbf{s}) - \beta_t \mu_t(\mathbf{s})
\end{aligned}
$$

$$
\Rightarrow \pi_t(\mathbf{a}|\mathbf{s}) = \exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) - \frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) - \alpha_t(\log \mu_t(\mathbf{s}) + 1) - \beta_t + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}'}{\alpha_t + \beta_t} \Big]
$$

Dual Problem:

$$
\begin{aligned}
L &= \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a}) \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{s} d\mathbf{a} + \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) d\mathbf{s} - \sum_{t=1}^{T\text{-}1} \int_{\mathbf{s}} \lambda_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}') \mu_T(\mathbf{s}') d\mathbf{s}' \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \mathcal{P}(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{a} d\mathbf{s} d\mathbf{s}' - \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_t(\mathbf{s}') \mu_t(\mathbf{s}') d\mathbf{s}' \\
&\quad + \sum_{t=1}^{T-1} \alpha_t \epsilon + \sum_{t=1}^{T-1} \beta_t \delta \\
&\quad - \sum_{t=1}^{T-1} \alpha_t \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \ldots \\
&\qquad \ldots \log \frac{\mu_t(\mathbf{s}) \exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) - \frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) - \alpha_t(\log \mu_t(\mathbf{s}) + 1) - \beta_t + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}'}{\alpha_t + \beta_t} \Big]}{q_t(\mathbf{s},\mathbf{a})} d\mathbf{a} d\mathbf{s} \\
&\quad - \sum_{t=1}^{T-1} \beta_t \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s}) \pi_t(\mathbf{a}|\mathbf{s}) \ldots \\
&\qquad \ldots \Big[\frac{R_t(\mathbf{s},\mathbf{a}) - \frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) - \alpha_t(\log \mu_t(\mathbf{s}) + 1) - \beta_t + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}') \mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}'}{\alpha_t + \beta_t} \Big]
\end{aligned}
$$

$$
\begin{aligned}
= \quad & \sum_{t=1}^{T\text{-}1}\int_{\mathbf{s}}\int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a})\mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})d\mathbf{s}d\mathbf{a} + \int_{\mathbf{s}}\mu_T(\mathbf{s})R_T(\mathbf{s})d\mathbf{s} \\
& + \sum_{t=1}^{T-1}\int_{\mathbf{s}}\lambda_t(\mathbf{s})d\mathbf{s} - \sum_{t=1}^{T\text{-}1}\int_{\mathbf{s}}\lambda_t(\mathbf{s})\int_{\mathbf{a}}\pi_t(\mathbf{a}|\mathbf{s})d\mathbf{a}d\mathbf{s} + \int_{\mathbf{s}}V_1(\mathbf{s})p_1(\mathbf{s})d\mathbf{s} - \int_{\mathbf{s}'}V_T(\mathbf{s}')\mu_T(\mathbf{s}')d\mathbf{s}' \\
& + \sum_{t=1}^{T-1}\int_{\mathbf{s}'}V_{t+1}(\mathbf{s}')\int_{\mathbf{s}}\int_{\mathbf{a}}\mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})\mathcal{P}(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{a}d\mathbf{s}d\mathbf{s}' - \sum_{t=1}^{T-1}\int_{\mathbf{s}'}V_t(\mathbf{s}')\mu_t(\mathbf{s}')d\mathbf{s}' + \sum_{t=1}^{T-1}\alpha_t\epsilon + \sum_{t=1}^{T-1}\beta_t\delta \\
& - \sum_{t=1}^{T-1}\alpha_t\int_{\mathbf{s}}\int_{\mathbf{a}}\mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})\log\mu_t(\mathbf{s})d\mathbf{a}d\mathbf{s} + \sum_{t=1}^{T-1}\alpha_t\int_{\mathbf{s}}\int_{\mathbf{a}}\mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})\log q_t(\mathbf{s},\mathbf{a})d\mathbf{a}d\mathbf{s} \\
& - \sum_{t=1}^{T-1}(\alpha_t+\beta_t)\int_{\mathbf{s}}\int_{\mathbf{a}}\mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})\ldots \\
& \ldots\Big(\frac{R_t(\mathbf{s},\mathbf{a}) - \frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} + \alpha_t\log q_t(\mathbf{s},\mathbf{a}) - \alpha_t(\log\mu_t(\mathbf{s})+1) - \beta_t + \int_{\mathbf{s}'}V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'}{\alpha_t+\beta_t}\Big)d\mathbf{a}d\mathbf{s}
\end{aligned}
$$

$$
\begin{aligned}
L(\mu_t,V_t,\lambda_t,\alpha_t,\beta_t) \quad = \quad & \sum_{t=1}^{T\text{-}1}\int_{\mathbf{s}}\int_{\mathbf{a}} R_t(\mathbf{s},\mathbf{a})\mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})d\mathbf{s}d\mathbf{a} + \int_{\mathbf{s}}\mu_T(\mathbf{s})R_T(\mathbf{s})d\mathbf{s} \\
& + \sum_{t=1}^{T-1}\int_{\mathbf{s}}\lambda_t(\mathbf{s})d\mathbf{s} - \sum_{t=1}^{T\text{-}1}\int_{\mathbf{s}}\lambda_t(\mathbf{s})\int_{\mathbf{a}}\pi_t(\mathbf{a}|\mathbf{s})d\mathbf{a}d\mathbf{s} + \int_{\mathbf{s}}V_1(\mathbf{s})p_1(\mathbf{s})d\mathbf{s} - \int_{\mathbf{s}'}V_T(\mathbf{s}')\mu_T(\mathbf{s}')d\mathbf{s}' \\
& + \sum_{t=1}^{T-1}\int_{\mathbf{s}'}V_{t+1}(\mathbf{s}')\int_{\mathbf{s}}\int_{\mathbf{a}}\mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})\mathcal{P}(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{a}d\mathbf{s}d\mathbf{s}' - \sum_{t=1}^{T-1}\int_{\mathbf{s}'}V_t(\mathbf{s}')\mu_t(\mathbf{s}')d\mathbf{s}' \\
& + \sum_{t=1}^{T-1}\alpha_t\epsilon + \sum_{t=1}^{T-1}\beta_t\delta - \sum_{t=1}^{T-1}\alpha_t\int_{\mathbf{s}}\int_{\mathbf{a}}\mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})\log\mu_t(\mathbf{s})d\mathbf{a}d\mathbf{s} \\
& + \sum_{t=1}^{T-1}\alpha_t\int_{\mathbf{s}}\int_{\mathbf{a}}\mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})\log q_t(\mathbf{s},\mathbf{a})d\mathbf{a}d\mathbf{s} \\
& - \sum_{t=1}^{T-1}\int_{\mathbf{s}}\int_{\mathbf{a}}\mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})\Big(R_t(\mathbf{s},\mathbf{a}) - \frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} + \alpha_t\log q_t(\mathbf{s},\mathbf{a}) - \alpha_t\log\mu_t(\mathbf{s}) - \alpha_t - \beta_t \\
& + \int_{\mathbf{s}'}V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)d\mathbf{a}d\mathbf{s} \\
= \quad & \int_{\mathbf{s}}\mu_T(\mathbf{s})R_T(\mathbf{s})d\mathbf{s} + \sum_{t=1}^{T-1}\int_{\mathbf{s}}\lambda_t(\mathbf{s})d\mathbf{s} + \int_{\mathbf{s}}V_1(\mathbf{s})p_1(\mathbf{s})d\mathbf{s} - \int_{\mathbf{s}'}V_T(\mathbf{s}')\mu_T(\mathbf{s}')d\mathbf{s}' \\
& - \sum_{t=1}^{T-1}\int_{\mathbf{s}'}V_t(\mathbf{s}')\mu_t(\mathbf{s}')d\mathbf{s}' + \sum_{t=1}^{T-1}\alpha_t(\epsilon+1) + \sum_{t=1}^{T-1}\beta_t(\delta+1)
\end{aligned}
$$

Solve for λ_t:

$$
\begin{aligned}
1 &= \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} \\
1 &= \int_{\mathbf{a}} \exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) - \frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) - \alpha_t(\log \mu_t(\mathbf{s}) + 1) - \beta_t + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}'}{\alpha_t + \beta_t}\Big] d\mathbf{a} \\
1 &= \int_{\mathbf{a}} \exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}'}{\alpha_t + \beta_t}\Big] \exp\Big[\frac{-\frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} - \alpha_t(\log \mu_t(\mathbf{s}) + 1) - \beta_t}{\alpha_t + \beta_t}\Big] d\mathbf{a}
\end{aligned}
$$

$$
\begin{aligned}
\exp\Big[\frac{\frac{\lambda_t(\mathbf{s})}{\mu_t(\mathbf{s})} + \alpha_t(\log \mu_t(\mathbf{s}) + 1) + \beta_t}{\alpha_t + \beta_t}\Big] &= \int_{\mathbf{a}} \exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}'}{\alpha_t + \beta_t}\Big] d\mathbf{a} \\
\lambda_t(\mathbf{s}) &= \mu_t(\mathbf{s})\Big(-\alpha_t(\log \mu_t(\mathbf{s}) + 1) - \beta_t + (\alpha_t + \beta_t)... \\
&\quad ... \log \int_{\mathbf{a}} \exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}'}{\alpha_t + \beta_t}\Big] d\mathbf{a}\Big)
\end{aligned}
$$

$$
\begin{aligned}
L(\mu_t, V_t, \alpha_t, \beta_t) &= \int_{\mathbf{s}} \mu_T(\mathbf{s}) R_T(\mathbf{s}) d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}') \mu_T(\mathbf{s}') d\mathbf{s}' \\
&- \sum_{t=1}^{T-1} \int_{\mathbf{s}'} V_t(\mathbf{s}') \mu_t(\mathbf{s}') d\mathbf{s}' + \sum_{t=1}^{T-1} \alpha_t \epsilon + \sum_{t=1}^{T-1} \beta_t \delta - \sum_{t=1}^{T-1} \int_{\mathbf{s}} \alpha_t \mu_t(\mathbf{s}) \log \mu_t(\mathbf{s}) d\mathbf{s} \\
&+ \sum_{t=1}^{T-1} (\alpha_t + \beta_t) \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} \exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}'}{\alpha_t + \beta_t}\Big] d\mathbf{a} d\mathbf{s}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial \mu_t} &= -V_t(\mathbf{s}) - \alpha_t(\log \mu_t(\mathbf{s}) + 1) + (\alpha_t + \beta_t) \log \int_{\mathbf{a}} \exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}'}{\alpha_t + \beta_t}\Big] d\mathbf{a} \\
0 &= -V_t(\mathbf{s}) + (\alpha_t + \beta_t) \log \int_{\mathbf{a}} \exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) - \alpha_t(\log \mu_t(\mathbf{s}) + 1) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) d\mathbf{s}'}{\alpha_t + \beta_t}\Big] d\mathbf{a} \\
\frac{\partial L}{\partial \mu_T} &= R_T(\mathbf{s}) - V_T(\mathbf{s}) \\
\frac{\partial L}{\partial V_t} &= -\mu_t(\mathbf{s}) + \int_{\hat{\mathbf{s}}} (\alpha_{t-1} + \beta_{t-1}) \mu_{t-1}(\hat{\mathbf{s}}) ... \\
&\quad ... \int_{\mathbf{a}} \frac{\exp\Big[\frac{1}{\alpha_{t-1}+\beta_{t-1}}\Big(R_{t-1}(\hat{\mathbf{s}},\mathbf{a}) + \alpha_{t-1} \log q_{t-1}(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_t(\mathbf{s}')\mathcal{P}_{t-1}(\mathbf{s}'|\hat{\mathbf{s}},\mathbf{a}) d\mathbf{s}'\Big)\Big]}{\int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_{t-1}+\beta_{t-1}}\Big(R_{t-1}(\hat{\mathbf{s}},\mathbf{a}) + \alpha_{t-1} \log q_{t-1}(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_t(\mathbf{s}')\mathcal{P}_{t-1}(\mathbf{s}'|\hat{\mathbf{s}},\mathbf{a}) d\mathbf{s}'\Big)\Big] d\mathbf{a}} \mathcal{P}_{t-1}(\mathbf{s}|\hat{\mathbf{s}},\mathbf{a}) d\mathbf{a} d\hat{\mathbf{s}} \\
0 &= -\mu_t(\mathbf{s}) + \int_{\hat{\mathbf{s}}} \int_{\mathbf{a}} \pi_{t-1}(\mathbf{a}|\hat{\mathbf{s}}) \mu_{t-1}(\hat{\mathbf{s}}) \mathcal{P}_{t-1}(\mathbf{s}|\hat{\mathbf{s}},\mathbf{a}) d\mathbf{a} d\hat{\mathbf{s}} \\
\frac{\partial L}{\partial V_1} &= p_1(\mathbf{s}) - \mu_1(\mathbf{s})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial \alpha_t} &= \epsilon + \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} \exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'}{\alpha_t+\beta_t}\Big] d\mathbf{a} d\mathbf{s} \\
&\quad - \int_{\mathbf{s}} (\alpha_t+\beta_t)\mu_t(\mathbf{s}) \int_{\mathbf{a}} \frac{\exp\Big[\frac{1}{\alpha_t+\beta_t}\Big(R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big]}{\int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t+\beta_t}\Big(R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big] d\mathbf{a}} \ldots \\
&\quad \ldots \frac{1}{(\alpha_t+\beta_t)^2}\Big(R_t(\mathbf{s},\mathbf{a}) - \beta_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big) d\mathbf{a} d\mathbf{s} - \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \mu_t(\mathbf{s}) d\mathbf{s} \\
&= \epsilon + \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} \exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'}{\alpha_t+\beta_t}\Big] d\mathbf{a} d\mathbf{s} \\
&\quad - \frac{1}{\alpha_t+\beta_t} \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})\Big(R_t(\mathbf{s},\mathbf{a}) - \beta_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big) d\mathbf{a} d\mathbf{s} \\
&\quad - \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \mu_t(\mathbf{s}) d\mathbf{s} \\
&= \epsilon - \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s}) \log \frac{\mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})}{q_t(\mathbf{s},\mathbf{a})} d\mathbf{a} d\mathbf{s} \\
\frac{\partial L}{\partial \beta_t} &= \delta + \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} \exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'}{\alpha_t+\beta_t}\Big] d\mathbf{a} d\mathbf{s} \\
&\quad - \int_{\mathbf{s}} (\alpha_t+\beta_t)\mu_t(\mathbf{s}) \int_{\mathbf{a}} \frac{\exp\Big[\frac{1}{\alpha_t+\beta_t}\Big(R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big]}{\int_{\mathbf{a}} \exp\Big[\frac{1}{\alpha_t+\beta_t}\Big(R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big)\Big] d\mathbf{a}} \ldots \\
&\quad \ldots \frac{1}{(\alpha_t+\beta_t)^2}\Big(R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big) d\mathbf{a} d\mathbf{s} \\
&= \delta + \int_{\mathbf{s}} \mu_t(\mathbf{s}) \log \int_{\mathbf{a}} \exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'}{\alpha_t+\beta_t}\Big] d\mathbf{a} d\mathbf{s} \\
&\quad - \frac{1}{\alpha_t+\beta_t} \int_{\mathbf{s}} \int_{\mathbf{a}} \mu_t(\mathbf{s})\pi_t(\mathbf{a}|\mathbf{s})\Big(R_t(\mathbf{s},\mathbf{a}) + \alpha_t \log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'\Big) d\mathbf{a} d\mathbf{s} \\
&= \delta - \int_{\mathbf{s}} \mu_t(\mathbf{s}) \int_{\mathbf{a}} \pi_t(\mathbf{a}|\mathbf{s}) \log \pi_t(\mathbf{a}|\mathbf{s}) d\mathbf{a} d\mathbf{s}
\end{aligned}
$$

$$
L(\mu_t, V_t, \alpha_t) = \int_{\mathbf{s}} V_1(\mathbf{s}) p_1(\mathbf{s}) d\mathbf{s} + \sum_{t=1}^{T-1} \alpha_t(\epsilon+1) + \sum_{t=1}^{T-1} \beta_t \delta
$$

Plug in Gaussians:

$$
\begin{aligned}
\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a}) &= \mathcal{N}(\mathbf{s}'|\mathbf{A}_t\mathbf{s}+\mathbf{b}_t\mathbf{a}+\mathbf{c}_t,\Sigma_{\mathbf{s}'}) \\
q_t(\mathbf{s}) &= \mathcal{N}(\mathbf{s}|\boldsymbol{\tau}^q_{s,t},\boldsymbol{\Sigma}^q_{s,t}) \\
q_t(\mathbf{a}|\mathbf{s}) &= \mathcal{N}(\mathbf{a}|\mathbf{K}_t\mathbf{s}+\mathbf{k}_t,\boldsymbol{\Sigma}^q_{a,t}) \\
q_t(\mathbf{s},\mathbf{a}) &= q_t(\mathbf{a}|\mathbf{s})q_t(\mathbf{s}) \\
&= \mathcal{N}(\begin{bmatrix}\mathbf{s}\\ \mathbf{a}\end{bmatrix}|\boldsymbol{\tau}^q_{s,a,t},\boldsymbol{\Sigma}^q_{s,a,t}) = \mathcal{N}\Big(\begin{bmatrix}\mathbf{s}\\ \mathbf{a}\end{bmatrix}\Big|\begin{bmatrix}\boldsymbol{\tau}^q_{s,t}\\ \mathbf{k}_t+\mathbf{K}_t\boldsymbol{\tau}^q_{s,t}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}^q_{s,t} & (\mathbf{K}_t\boldsymbol{\Sigma}^q_{s,t})^T\\ \mathbf{K}_t\boldsymbol{\Sigma}^q_{s,t} & \boldsymbol{\Sigma}^q_{a,t}+\mathbf{K}_t(\mathbf{K}_t\boldsymbol{\Sigma}^q_{s,t})^T\end{bmatrix}\Big) \\
\mu_t(\mathbf{s}) &= \mathcal{N}(\mathbf{s}|\boldsymbol{\tau}^p_{s,t},\boldsymbol{\Sigma}^p_{s,t}) \\
R_t(\mathbf{s},\mathbf{a}) &= (\mathbf{z}_t-\mathbf{s})^T\mathbf{M}_t(\mathbf{z}_t-\mathbf{s})+\mathbf{a}_t^T\mathbf{H}_t\mathbf{a}_t \\
R_T &= (\mathbf{z}_T-\mathbf{s}_T)^T\mathbf{M}_t(\mathbf{z}_T-\mathbf{s}_T) \\
\int_{\mathbf{s}'} V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}' &= \mathbb{E}_{\mathcal{P}}[V_{t+1}(\mathbf{s}')] \\
V_{t+1}(\mathbf{s}) &= \mathbf{s}^T\mathbf{V}_{t+1}\mathbf{s}+\mathbf{v}_{t+1}^T\mathbf{s}+v_{t+1}
\end{aligned}
$$

$$
\begin{aligned}
r_t(\mathbf{s},\mathbf{a}) &= R_t(\mathbf{s},\mathbf{a})+\alpha_t\log q_t(\mathbf{s},\mathbf{a})-\alpha_t\log\mu_t(\mathbf{s})-\alpha_t \\
&= R_t(\mathbf{s},\mathbf{a})+\alpha_t\log q_t(\mathbf{a}|\mathbf{s})+\alpha_t\log q_t(\mathbf{s})-\alpha_t\log\mu_t(\mathbf{s})-\alpha_t \\
&= \mathbf{z}^T\mathbf{M}_t\mathbf{z}-\mathbf{z}^T\mathbf{M}_t\mathbf{s}-\mathbf{s}^T\mathbf{M}_t\mathbf{z}+\mathbf{s}^T\mathbf{M}_t\mathbf{s}+\mathbf{a}^T\mathbf{H}_t\mathbf{a}-\frac{\alpha_t}{2}\log\Big|2\pi\boldsymbol{\Sigma}^q_{a,t}\Big| \\
&\quad -\frac{\alpha_t}{2}\Big(\mathbf{a}^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{a}-\mathbf{a}^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{K}^q_t s-\mathbf{a}^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t-\mathbf{s}^T(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{a}+\mathbf{s}^T(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{K}^q_t\mathbf{s} \\
&\quad +\mathbf{s}^T(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t-(\mathbf{k}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{a}+(\mathbf{k}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{K}^q_t\mathbf{s}+(\mathbf{k}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t\Big)-\frac{\alpha_t}{2}\log|2\pi\boldsymbol{\Sigma}^q_{s,t}| \\
&\quad -\frac{\alpha_t}{2}(\mathbf{s}-\boldsymbol{\tau}^q_{s,t})^T(\boldsymbol{\Sigma}^q_{s,t})^{-1}(\mathbf{s}-\boldsymbol{\tau}^q_{s,t})-\alpha_t+\alpha_t\Big(\frac{1}{2}\log|2\pi\boldsymbol{\Sigma}^p_{s,t}|+\frac{1}{2}(\mathbf{s}-\boldsymbol{\tau}^p_{s,t})^T(\boldsymbol{\Sigma}^p_{s,t})^{-1}(\mathbf{s}-\boldsymbol{\tau}^p_{s,t})\Big) \\
&= \mathbf{z}^T\mathbf{M}_t\mathbf{z}-\mathbf{z}^T\mathbf{M}_t\mathbf{s}-\mathbf{s}^T\mathbf{M}_t\mathbf{z}+\mathbf{s}^T\mathbf{M}_t\mathbf{s}+\mathbf{a}^T\mathbf{H}_t\mathbf{a}-\frac{\alpha_t}{2}\log\Big|2\pi\boldsymbol{\Sigma}^q_{a,t}\Big| \\
&\quad -\frac{\alpha_t}{2}\Big(\mathbf{a}^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{a}-\mathbf{a}^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{K}^q_t\mathbf{s}-\mathbf{a}^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t-\mathbf{s}^T(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{a} \\
&\quad +\mathbf{s}^T(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{K}^q_t\mathbf{s}+\mathbf{s}^T(\mathbf{K}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t-(\mathbf{k}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{a}+(\mathbf{k}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{K}^q_t\mathbf{s}+(\mathbf{k}^q_t)^T(\boldsymbol{\Sigma}^q_{a,t})^{-1}\mathbf{k}^q_t\Big) \\
&\quad -\frac{\alpha_t}{2}\log|2\pi\boldsymbol{\Sigma}^q_{s,t}|-\frac{\alpha_t}{2}\mathbf{s}^T(\boldsymbol{\Sigma}^q_{s,t})^{-1}\mathbf{s}+\alpha_t\mathbf{s}^T(\boldsymbol{\Sigma}^q_{s,t})^{-1}\boldsymbol{\tau}^q_{s,t}-\frac{\alpha_t}{2}(\boldsymbol{\tau}^q_{s,t})^T(\boldsymbol{\Sigma}^q_{s,t})^{-1}\boldsymbol{\tau}^q_{s,t}-\alpha_t \\
&\quad +\frac{\alpha_t}{2}\Big(\log|2\pi\boldsymbol{\Sigma}^p_{s,t}|+\mathbf{s}^T(\boldsymbol{\Sigma}^p_{s,t})^{-1}\mathbf{s}-2\mathbf{s}^T(\boldsymbol{\Sigma}^p_{s,t})^{-1}\boldsymbol{\tau}^p_{s,t}+(\boldsymbol{\tau}^p_{s,t})^T(\boldsymbol{\Sigma}^p_{s,t})^{-1}\boldsymbol{\tau}^p_{s,t}\Big) \\
&= \mathbf{s}^T\mathbf{R}_{ss,t}\mathbf{s}+\mathbf{a}^T\mathbf{R}_{aa,t}\mathbf{a}+\mathbf{a}^T\mathbf{R}^T_{sa,t}\mathbf{s}+\mathbf{s}^T\mathbf{R}_{sa,t}\mathbf{a}+\mathbf{s}^T\mathbf{r}_{s,t}+\mathbf{a}^T\mathbf{r}_{a,t}+r_{0,t}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{R}_{ss,t} &= \mathbf{M}_t - \frac{\alpha_t}{2}(\mathbf{K}_t^q)^T(\mathbf{\Sigma}_{a,t}^q)^{-1}\mathbf{K}_t^q + \frac{\alpha_t}{2}(\mathbf{\Sigma}_{s,t}^q)^{-1} + \frac{\alpha_t}{2}(\mathbf{\Sigma}_{s,t}^p)^{-1} \\
\mathbf{R}_{aa,t} &= \mathbf{H}_t - \frac{\alpha_t}{2}(\mathbf{\Sigma}_{a,t}^q)^{-1} \\
\mathbf{R}_{sa,t} &= +\frac{\alpha_t}{2}(\mathbf{K}_t^q)^T(\mathbf{\Sigma}_{a,t}^q)^{-1} \\
\mathbf{r}_{s,t} &= -\alpha_t(\mathbf{K}_t^q)^T(\mathbf{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q + \alpha_t(\mathbf{\Sigma}_{s,t}^q)^{-1}\boldsymbol{\tau}_{s,t}^q - \alpha_t(\mathbf{\Sigma}_{s,t}^p)^{-1}\boldsymbol{\tau}_{s,t}^p - 2\mathbf{M}_t\mathbf{z} \\
\mathbf{r}_{a,t} &= \alpha_t(\mathbf{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q \\
r_{0,t} &= \mathbf{z}^T\mathbf{M}_t\mathbf{z} - \frac{\alpha_t}{2}\log\left|2\pi\mathbf{\Sigma}_{a,t}^q\right| - \frac{\alpha_t}{2}(\mathbf{k}_t^q)^T(\mathbf{\Sigma}_{a,t}^q)^{-1}\mathbf{k}_t^q - \frac{\alpha_t}{2}\log|2\pi\mathbf{\Sigma}_{s,t}^q| \\
&\quad -\frac{\alpha_t}{2}(\boldsymbol{\tau}_{s,t}^q)^T(\mathbf{\Sigma}_{s,t}^q)^{-1}\boldsymbol{\tau}_{s,t}^q - \alpha_t + \frac{\alpha_t}{2}\Big(\log|2\pi\mathbf{\Sigma}_{s,t}^p| + (\boldsymbol{\tau}_{s,t}^p)^T(\mathbf{\Sigma}_{s,t}^p)^{-1}\boldsymbol{\tau}_{s,t}^p\Big)
\end{aligned}
$$

Coordinate Descent:

$$
\begin{aligned}
L(\mu_t, V_t, \alpha_t, \beta_t) &= \int_{\mathbf{s}} \mu_T(\mathbf{s})R_T(\mathbf{s})d\mathbf{s} + \int_{\mathbf{s}} V_1(\mathbf{s})p_1(\mathbf{s})d\mathbf{s} - \int_{\mathbf{s}'} V_T(\mathbf{s}')\mu_T(\mathbf{s}')d\mathbf{s}' \\
&- \sum_{t=1}^{T-1}\int_{\mathbf{s}'} V_t(\mathbf{s}')\mu_t(\mathbf{s}')d\mathbf{s}' + \sum_{t=1}^{T-1}\alpha_t\epsilon + \sum_{t=1}^{T-1}\beta_t\delta - \sum_{t=1}^{T-1}\int_{\mathbf{s}}\alpha_t\mu_t(\mathbf{s})\log\mu_t(\mathbf{s})d\mathbf{s} \\
&+ \sum_{t=1}^{T-1}\int_{\mathbf{s}}(\alpha_t+\beta_t)\mu_t(\mathbf{s})\log\int_{\mathbf{a}}\exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) + \alpha_t\log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'}V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'}{\alpha_t+\beta_t}\Big]d\mathbf{a}d\mathbf{s} \\
\frac{\partial L(\mu_t, V_t, \alpha_t, \beta_t)}{\partial \mu_t} &= -V_t(\mathbf{s}) - \alpha_t\log\mu_t(\mathbf{s}) - \alpha_t \\
&\quad + \frac{(\alpha_t+\beta_t)}{\alpha_t}\log\int_{\mathbf{a}}\exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) + \alpha_t\log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'}V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'}{\alpha_t+\beta_t}\Big]d\mathbf{a} = 0 \\
\log\mu_t(\mathbf{s}) &= -\frac{1}{\alpha_t}V_t(\mathbf{s}) - 1 \\
&\quad + \frac{(\alpha_t+\beta_t)}{\alpha_t}\log\int_{\mathbf{a}}\exp\Big[\frac{R_t(\mathbf{s},\mathbf{a}) + \alpha_t\log q_t(\mathbf{s},\mathbf{a}) + \int_{\mathbf{s}'}V_{t+1}(\mathbf{s}')\mathcal{P}_t(\mathbf{s}'|\mathbf{s},\mathbf{a})d\mathbf{s}'}{\alpha_t+\beta_t}\Big]d\mathbf{a} \\
&= -\frac{1}{\alpha_t}(\mathbf{s}^T\mathbf{V}_t\mathbf{s} + \mathbf{s}^T\mathbf{v}_t + v_t + \alpha_t) + \frac{1}{\alpha_t}(\mathbf{s}^T\hat{\mathbf{V}}_t\mathbf{s} + \mathbf{s}^T\hat{\mathbf{v}}_t + \hat{v}_t) \\
&= -\frac{1}{2}\Big(\mathbf{s}^T\frac{2}{\alpha_t}(\mathbf{V}_t - \hat{\mathbf{V}}_t)\mathbf{s} + \mathbf{s}^T\frac{2}{\alpha_t}(\mathbf{v}_t - \hat{\mathbf{v}}_t) + \frac{2}{\alpha_t}(\hat{v}_t - v_t + \alpha_t)\Big) \\
\log\mathcal{N}(s|a,A) &= -\frac{1}{2}\Big(\log|2\pi A| + (s-a)^TA^{-1}(s-a)\Big) \\
&= -\frac{1}{2}\Big(\log|2\pi A| + s^TA^{-1}s - 2s^TA^{-1}a + a^TA^{-1}a\Big) \\
A^{-1} &= \frac{2}{\alpha_t}(\mathbf{V}_t - \hat{\mathbf{V}}_t) \\
A^{-1}a &= -\frac{1}{\alpha_t}(\mathbf{v}_t - \hat{\mathbf{v}}_t) \\
a &= -\frac{1}{2}(\mathbf{V}_t - \hat{\mathbf{V}}_t)^{-1}(\mathbf{v}_t - \hat{\mathbf{v}}_t) \\
\mu_t(\mathbf{s}) &= \mathcal{N}(s|a,A) \\
&= \mathcal{N}(\mathbf{s}|-\frac{1}{2}(\mathbf{V}_t - \hat{\mathbf{V}}_t)^{-1}(\mathbf{v}_t - \hat{\mathbf{v}}_t), \frac{\alpha_t}{2}(\mathbf{V}_t - \hat{\mathbf{V}}_t)^{-1})
\end{aligned}
$$