Interpretable Policy Optimization Algorithms

Anonymous authors
Paper under double-blind review

Abstract

To make machine learning widely applicable in today's society, we need to create models that are understood by laypeople. Creating techniques that provide decision support would be highly beneficial in areas such as healthcare, finance, and robotics. Towards that goal, we design a model-free, interpretable version of the Trust Region Policy Optimization method. The method works by creating policies that have the structure of a differentiable decision tree. We also provide convergence bounds and intuition on how to tune the parameters to get this algorithm to work effectively. We lay out the construction of the method and list the subproblems that have to be solved at each leaf and internal node.

Part I

Model-Free Reinforcement Learning

In the upcoming chapter, we wish to develop model-free interpretable reinforcement learning algorithms, where we fit a decision tree based policy. If we assume a decision tree of a fixed depth d and a structure is given, the goal of parametric model-free reinforcement learning would be to find the splitting constraints at each of the nodes, and the random variables over the space of actions for each of the leaves.

The traditional problem with fitting a decision-tree based policy is **non-differentiability** of the return. To overcome this obstacle we define a differentiable decision tree where the direction to go is chosen in a differentiable manner. In what proceeds, we would assume access to a binary tree $\mathcal{T}$ and the objective would be to fit constraint conditions on the nodes and action space random variables on the leaves. The following parameterization is taken from **?**.

1 Differentiable Decision Binary Trees

Given a binary tree $\mathcal{T}$, we wish to create maps from the state space S_t to the action space A_t at each node η. To do this we attach a differentiable constraint condition $\mu_\eta(x; a_\eta, \beta_\eta, \phi_\eta)$ to each of the nodes η, and a random variable α_η over the action space A to each of the leaves η. With this setting, given a binary tree $\mathcal{T}$, we would be able to define a map $\mathcal{T}_\eta$ from each node η of the tree to the action space A_t. If we are at a node η and we have $B_\eta(x) \sim Ber(\mu_\eta(x; a_\eta, \beta_\eta, \phi_\eta))$. The definition is as follows:

$$\begin{aligned} T_\eta(x) &= \alpha_\eta \text{ if } \eta \text{ is leaf} \\ &= B_\eta(x) T_{\mathcal{L}(\eta)}(x) + (1 - B_\eta(x)) T_{\mathcal{R}(\eta)}(x) \text{ otherwise} \end{aligned} \tag{1}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In the above and elsewhere we have that $\mathcal{L}(\eta)$ is the left of the given node η, and $\mathcal{R}(\eta)$ denotes the node on the right of the given node η. This means at the node η we go left or right probabilistically after the flipping of a coin with probability $\mu_\eta(x)$.

Since an indicator based splitting function for $\mu_\eta(x; a_\eta, \beta_\eta, \phi_\eta)$ is not differentiable with respect to x, we define the splitting condition $\mu_\eta(x; a_\eta, \beta_\eta, \phi_\eta)$ as:

$$\mu_\eta(x; a_\eta, \beta_\eta, \phi_\eta) := \frac{1}{(1 + \exp(-a_\eta(\beta_\eta^T x - \phi_\eta))} \tag{2}$$

In this setting, given the skeleton of the tree with the splitting conditions and leaf-action random variables, we define Random Action Binary Tree (RABT) maps below:

Random Action Binary Tree maps : Given a tree T with head node η_0, constraint condition maps $(\mu_\eta(a_\eta, \beta_\eta, \phi_\eta))_{\eta \in N(\mathcal{T})}$, where we have $N(\mathcal{T})$ to be the set of nodes of $\mathcal{T}$; $(\alpha_l)_{l \in L(\mathcal{T})}$;, where α_l denotes the action random variable over A for $l \in Leaves(T)$. We define the $\mathbf{RABT - map}$ T_{η_0} which is defined to satisfy 1 and the constraint condition μ_η satisfies 2.

Parameter Space : The parameter space for a decision tree policy with tree $\mathcal{T}$ is given by $\Theta := \{(a_\eta, \beta_\eta, \phi_\eta) | \eta \in Nodes(\mathcal{T})\} \sqcup \{\alpha_l \in \Delta(A) | l \in Leaves(\mathcal{T}))\}$.

Decision tree based policies : For a fixed binary tree $\mathcal{T}$ and action space A, we define the set of decision tree based policies as $\Pi(\mathcal{T}) := \{T_{\eta_0} | [\mu(a_\eta, \beta_\eta, \phi_\eta) \in Const(T_{\eta_0}), \alpha_l \in \Delta(A)]; a_\eta \in \mathbb{R}, \beta_\eta \in \mathbb{R}^d, \phi_\eta \in \mathbb{R}, \eta \in Nodes(T_{\eta_0}), l \in Leaves(T)\}$, where $Const(T)$ is the space of constraint condition maps for a given tree T. For a fixed parameter $\theta \in \Theta$ and a fixed tree T, another notation we introduce for T_{η_0} is given by T_θ. Note that in the following $T_\theta(s)$ is a random variable that indicates the random action taken at state s.

In the next paragraph, we review the classification problem for fitting a differentiable decision tree to a dataset $\mathcal{D}$.

2 Classification With Decision Trees

Given the classifying dataset $[(X_i, Y_i) \forall i \in [N]]$, where we have $X_i \in \mathbb{R}^d, Y_i \in \{0, 1\}$, where $i \in [N]$, how would we find a decision tree function parameterized as above that fits this data? Assume in the following that $B_\eta(x)$ is the Bernoulli random variable which is 1 with probability $\mu_\eta(x; a_\eta, b_\eta, \phi_\eta))$. With the above notation and using a mean-squared error objective we see that we can describe the optimization problem at a node η as follows:

$$\mathcal{L}_\eta((X_i, Y_i)_{i=1}^N) := \sum_{i=1}^N \mathbb{E}\left[\left(B_\eta(X_i) T_{L(\eta)}(X_i) + (1 - B_\eta(X_i)) T_{R(\eta)}(X_i) - Y_i\right)^2\right] \tag{3}$$

We expand upon the above expectation as follows:

$$\sum_{i=1}^N \mathbb{E}\left[\left(B_\eta(X_i)(T_{\mathcal{L}(\eta)}(X_i) - T_{\mathcal{R}(\eta)}(X_i)) + (T_{\mathcal{R}(\eta)}(X_i) - Y_i)\right)^2\right] \tag{4}$$

Taking the square of the above term we get the following:

$$\sum_{i=1}^N \mathbb{E}\Bigg[\Bigg(B_\eta^2(X_i)(T_{\mathcal{L}(\eta)}(X_i) - T_{\mathcal{R}(\eta)}(X_i))^2 + (T_{\mathcal{R}(\eta)}(X_i) - Y_i)^2\Bigg] + \tag{5}$$

$$+ 2B_\eta(X_i)(T_{\mathcal{L}(\eta)}(X_i) - T_{\mathcal{R}(\eta)}(X_i))(T_{\mathcal{R}(\eta)}(X_i) - Y_i)\Bigg)\Bigg] \tag{6}$$

We have that $B_\eta(X_i) = B_\eta(X_i)^2$, and we have $\mathbb{E}[B_\eta(X_i)] = \mu_\eta(X_i; a_\eta, b_\eta, \phi_\eta)$ so the above expression after taking the expectation looks like:

$$\sum_{i=1}^{N} \Bigg(\mu_\eta(X_i; a_\eta, b_\eta, \phi_\eta)(T_{\mathcal{L}(\eta)}(X_i) - T_{\mathcal{R}(\eta)}(X_i))^2 + (T_{\mathcal{R}(\eta)}(X_i) - Y_i)^2 + \tag{7}$$

$$+2\mu_\eta(X_i; a_\eta, b_\eta, \phi_\eta)(T_{\mathcal{L}(\eta)}(X_i) - T_{\mathcal{R}(\eta)}(X_i))(T_{\mathcal{R}(\eta)}(X_i) - Y_i) \Bigg) \tag{8}$$

Further expansion of the above gives us:

$$\sum_{i=1}^{N} \Bigg(\mu_\eta(X_i; a_\eta, b_\eta, \phi_\eta) \Bigg(T_{\mathcal{L}(\eta)}(X_i) - T_{\mathcal{R}(\eta)}(X_i) \Bigg) [T_{\mathcal{L}(\eta)}(X_i) + T_{\mathcal{R}(\eta)}(X_i) - 2Y_i] + (T_{\mathcal{R}(\eta)}(X_i) - Y_i)^2 \Bigg) \tag{9}$$

In the above we have $B_\eta(x) \sim Ber(\mu_\eta(x; a_\eta, \beta_\eta, \phi_\eta))$. To compute the optimal parameters $a_\eta, \beta_\eta, \phi_\eta$ for the node η we note the differential of $\mu_\eta(x)$ with respect to $a_\eta, \beta_\eta, \phi_\eta$. This gives us:

$$\frac{\partial \mu_\eta(x)}{\partial a_\eta} = \frac{(\beta_\eta^\top x - \phi_\eta) \exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)}{(1 + \exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big))^2}. \tag{10}$$

When we differentiate the loss function $\mathcal{L}_\eta((X_i, Y_i)_{i=1}^N)$ with respect to the parameter a_η, β_η, or ϕ_η we see that the expression evaluates to the following:

$$\frac{\partial}{\partial a_\eta} \mathcal{L}_\eta(X_i, Y_i) := \sum_{i=1}^{N} \Bigg(\frac{\partial \mu_\eta(X_i)}{\partial a_\eta} \Bigg(T_{\mathcal{L}(\eta)}(X_i) - T_{\mathcal{R}(\eta)}(X_i) \Bigg) [T_{\mathcal{L}(\eta)}(X_i) + T_{\mathcal{R}(\eta)}(X_i) - 2Y_i] \Bigg) \tag{11}$$

Differentiation of μ_η with respect to β_η, we get the following to be true:

$$\frac{\partial \mu_\eta(x)}{\partial \beta_\eta} = \frac{a_\eta x \exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)}{(1 + \exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big))^2}. \tag{12}$$

We differentiate $\mathcal{L}_\eta((X_i, Y_i)_{i=1}^N)$ once again with respect to β_η and this gives us the following expression:

$$\frac{\partial}{\partial \beta_\eta} \mathcal{L}_\eta(X_i, Y_i) := \sum_{i=1}^{N} \Bigg(\frac{\partial \mu_\eta(X_i)}{\partial \beta_\eta} \Bigg(T_{\mathcal{L}(\eta)}(X_i) - T_{\mathcal{R}(\eta)}(X_i) \Bigg) [T_{\mathcal{L}(\eta)}(X_i) + T_{\mathcal{R}(\eta)}(X_i) - 2Y_i] \Bigg) \tag{13}$$

Using terminology as in the neural network literature, we can forward propagate the inputs to store the values $T_\eta(X_i)$ and then perform backward propagation notation to store the $\frac{\partial \mu}{\partial \beta_\mu}$ values. We are also allowed to use different loss functions such as the cross-entropy to measure the loss in the differentiable decision tree.

$$\mathcal{L}_\eta((X_i, Y_i)_{i=1}^N) := \tag{14}$$

$$-\Bigg(\sum_{i=1}^{N} \mathbb{E}\Bigg[Y_i \log(B_\eta(X_i) T_{L(\eta)}(X_i) + (1 - B_\eta(X_i)) T_{R(\eta)}(X_i)) \tag{15}$$

$$+(1 - Y_i) \log(1 - B_\eta(X_i)) T_{L(\eta)}(X_i) - (1 - B_\eta(X_i)) T_{R(\eta)}(X_i)) \Bigg] \Bigg) \tag{16}$$

We can use the above loss function in the reinforcement learning setting to obtain an estimate of the policy function. Here we would have the trajectories $[(X_i, Y_i)_{i=1}^N]$ to be given by the $[(S_t, A_t)_{t=1}^N]$ and we would have to deductively reason the policy which gave rise to this set of policies.

3 Decision Tree Trust Region Policy Optimization

In this section, we recall the Trust Region Policy Optimization (TRPO) Algorithm. This algorithm works by optimizing a surrogate objective function while ensuring that the optimized policy and previous policy are close to each other in the KL-divergence. We proceed to compute the version of TRPO which uses a decision tree based final policy, where the parameters we optimize for are given by constraint conditions for each of the nodes and random variables over the space of actions for each of the leaves. We compute the non-linear objective and KL divergence conditions for each of the nodes and the leaves. We then proceed to compute the approximated version of this algorithm by taking a Taylor series expansion. We end with computation of the pseudocode for the method and list the various subproblems at the head and other nodes for the problem.

3.1 Trust Region Policy Optimization

The Trust Region Policy Optimization (TRPO) algorithm is a reinforcement learning algorithm where we optimize for a surrogate objective while assuming that the new and old policies are close to one another in KL-divergence. In particular, let π_θ denote a policy with parameters θ. The theoretical TRPO update is:

$$\theta_{k+1} = \arg\max_{\theta} \mathcal{L}(\theta_k, \theta) \tag{17}$$

Where we have:

$$\bar{D}_{KL}(\theta||\theta_k) \leq \delta \tag{18}$$

where $\mathcal{L}(\theta_k, \theta)$ is the surrogate advantage, a measure of how policy π_θ performs relative to the old policy π_{θ_k} using data from the old policy:

$$\mathcal{L}(\theta_k, \theta) = \mathbb{E}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s, a)\right] \tag{19}$$

where in the above we have:

$$\bar{D}_{KL}(\theta||\theta_k) = \mathbb{E}_{s\sim\theta_k}[D_{KL}(\pi_\theta(.|s), \pi_{\theta_k}(.|s))] \tag{20}$$

It is important to note that the original paper **?** considers π_θ to be the policy we optimize for all the timesteps $t \in [T]$.

3.2 Decision Tree Trust Region Policy Optimization

In this subsection, we consider the Trust Region Policy Optimization algorithm (TRPO) where the final policy is given by a parametrized decision tree as above.

Assume we have a fixed tree $\mathcal{T}$ of a certain depth where the leaf and node structure is given. We assume π_θ is the given policy where $\theta := \{(a_\eta, \beta_\eta, \phi_\eta)|\eta \in [T]\} \sqcup \{\alpha_l | l \in \text{ leaves(T)}\}$. The goal of this subsection would be to compute the optimal parameters $(a_\eta, \beta_\eta, \phi_\eta)$ and action random variables $\alpha_l \forall l \in \text{Leaves}(\mathcal{T})$ for the TRPO objective and constraint.

In this setting, what would be the KL divergence between the policies $\pi_\theta(.|s)$ and $\pi_{\theta_k}(.|s)$?

We want to bound/simplify the following $\bar{D}_{KL}(\theta||\theta_k)$ which is given by:

$$\bar{D}_{KL}(\theta||\theta_k) = \mathbb{E}_{s\sim\theta_k}[D_{KL}(\pi_\theta(.|s), \pi_{\theta_k}(.|s))] \tag{21}$$

The above written in full is given as:

$$\int_{S_t} \int_{A_t} \left(\pi_\theta(a|s) \log \frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} da\right) ds \tag{22}$$

For a fixed decision tree policy π_θ, we define the random variables $(\alpha_l)_{l \in L(T)}$ to be those random variables over the action space A. Below we redefine the notation for the evaluation of the probabilities and we evaluate $\pi_\theta(a|s)$ as follows:

$$\pi_\theta(a|s) := \mathbb{P}[T_\theta(s) = a] \tag{23}$$

$$= \sum_{l \in L(T_\theta)} \mathbb{P}_{T_\theta}[s \to l]\mathbb{P}[\eta_l = a] \tag{24}$$

In the above, we regard $T_\theta : S \to A$ as a random map from the state space to the action space. In the above, we have that $\mathbb{P}_{T_\theta}[s \to l]$ is the probability that we land in leaf l if we follow the path dictated by T_θ.

If the decision tree structure is deterministic, we see that there exists a unique leaf $l' \in Leaves(T)$ such that $\mathbb{P}_{T_\theta}[s \to l'] = 1$, in that case $\pi_\theta(a|s) = \mathbb{P}[\alpha_{l'} = a]$. In this case, we can define $\eta_\theta(s)$ and $\eta_{\theta_k}(s)$ to be the leaves we end at, when we follow $T_\theta(s)$ and $T_{\theta_k}(s)$.

Then we can rewrite 22 as $\mathbb{E}[D_{KL}(\pi_\theta(.|s), \pi_{\theta_k}(.|s))]$ as the following:

$$\frac{1}{|S|} \sum_{s \in S} D_{KL}(\pi_\theta(.|s), \pi_{\theta_k}(.|s)) := \tag{25}$$

$$\frac{1}{|S|} \sum_{s \in S} \sum_{a \in A} \mathbb{P}[T_{\theta(s)} = a] \log \left(\frac{\mathbb{P}[T_{\theta(s)} = a]}{\mathbb{P}[T_{\theta_k(s)} = a]} \right) \tag{26}$$

How would we simplify the above? The above can be interpreted as the average KL divergence between the leaves of the trees with θ and θ_k as a parameter. Can we further simplify the objective in this scenario; recall the objective is given by:

$$\mathcal{L}(\theta_k, \theta) = \mathbb{E}_{s,a \sim \pi_{\theta_k}} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s, a) \right] \tag{27}$$

In the above, we have that the advantage function is given by:

$$A^{\pi_{\theta_k}}(s, a) := Q_{\pi_{\theta_k}}(s, a) - V_{\pi_{\theta_k}}(s) \tag{28}$$

The equation [27] can be written as:

$$\mathcal{L}(\theta_k, \theta) = \mathbb{E}_{s,a \sim \pi_{\theta_k}} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s, a) \right] \tag{29}$$

A simplification of the above is given by:

$$\mathbb{E}_{s,a \sim \pi_{\theta_k}} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} [Q_{\pi_{\theta_k}}(s, a) - V_{\pi_{\theta_k}}(s)] \right] \tag{30}$$

Recall that the TRPO computes the objective over a rollout of the final policy:

$$\sum_{t=0}^{T-1} \sum_{i=1}^{N} \frac{\pi_\theta(a_t^i|s_t^i)}{\pi_{\theta_k}(a_t^i|s_t^i)} \left[Q_{\pi_{\theta_k}}(s_t^i, a_t^i) - V_{\pi_{\theta_k}}(s_t^i) \right] \tag{31}$$

The online notation for 30 is given as follows:

$$\arg\max_{\theta \in \Theta} \frac{1}{T} \mathbb{E}_{(s_t,a_t) \sim \tau_t} \left[\sum_{t=1}^{T} \frac{\pi_\theta(a_t^i|s_t^i)}{\pi_{\theta_k}(a_t^i|s_t^i)} \left[Q_{\pi_{\theta_k}}(s_t, a_t) - V_{\pi_{\theta_k}}(s_t) \right] \right] \tag{32}$$

In the PPO paper from **?**, we see that different objectives are introduced for the 33. In particular we use $r_t(\theta) := \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$:

1. No clipping or penalty : $L_t(\theta) = r_t(\theta)\hat{A}_t$
2. Clipping : $L_t(\theta) = \min(r_t(\theta)\hat{A}_t, \text{clip}(1-\epsilon, 1+\epsilon))\hat{A}_t$
3. KL penalty : $r_t(\theta)\hat{A}_t - \beta KL[\pi_{old}, \pi_\theta]$

The full optimization problem is now given by:

$$\arg\max_{\theta\in\Theta} \frac{1}{T}\mathbb{E}_{(s_t,a_t)\sim\tau_t}\left[\sum_{t=1}^{T}\frac{\mathbb{P}[T_\theta(s_t)=a_t]}{\mathbb{P}[T_{\theta_k}(s_t)=a_t]}\left[Q_{\pi_{\theta_k}}(s_t,a_t) - V_{\pi_{\theta_k}}(s_t)\right]\right] \tag{33}$$

The approximated TRPO update is now given by:

$$\mathcal{L}(\theta_k,\theta) \simeq g^T(\theta-\theta_k) \tag{34}$$

$$\bar{D}_{KL}(\theta||\theta_k) \simeq \frac{1}{2}(\theta-\theta_k)^T H(\theta-\theta_k) \tag{35}$$

In the above, we have that :

$$g := \nabla_\theta J(\theta) \tag{36}$$

$$Hx := \nabla_\theta((\nabla_\theta \bar{D}_{KL}(\theta||\theta_k))^T x) \tag{37}$$

What would $\mathbb{E}[\nabla_\theta J(\theta)]$ be in the differentiable decision tree setting and what about $\bar{D}_{KL}(\theta||\theta_k)$, how would that look like?

Recall the grad-log-prob lemma from **?**:

$$\nabla_\theta J_\pi(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)R(\tau)] \tag{38}$$

The above can be simplified as follows:

$$\nabla_\theta J_\pi(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T}\nabla_\theta \log \mathbb{P}[T_\theta(s_t)=a_t]R(\tau)\right] \tag{39}$$

How does $\nabla_\theta \log \mathbb{P}[T_\theta(s_t)=a_t]$ behave with respect to θ? Recall that it can be written as:

$$\nabla_\theta\left[\log \mathbb{P}[T_\theta(s_t)=a_t]\right] = \tag{40}$$

$$\frac{\frac{\partial}{\partial\theta}\mathbb{P}[T_\theta(s_t)=a_t]|_{\theta=\theta_0}}{\mathbb{P}[T_\theta(s_t)=a_t]} \tag{41}$$

At a node η we have the splitting function to be given as follows:

$$\mu_\eta(x;a_\eta,\beta_\eta,\phi_\eta) := \left(1+\exp\left(-a_\eta(\beta_\eta^T x-\phi_\eta)\right)\right)^{-1} \tag{42}$$

The splitting equation at a node η is given by:

$$\begin{aligned} T_\eta(x) &= y_\eta \text{ if leaf} \\ &= B_\eta(x)T_{\mathcal{L}(\eta)}(x) + (1-B_\eta(x))T_{\mathcal{R}(\eta)}(x) \text{ otherwise} \end{aligned} \tag{43}$$

The below gives the derivatives of μ_η with respect to it's parameters:

$$\frac{\partial\mu_\eta(x)}{\partial a_\eta} = \frac{(\beta_\eta^\top x-\phi_\eta)\exp\big(-a_\eta(\beta_\eta^\top x-\phi_\eta)\big)}{(1+\exp\big(-a_\eta(\beta_\eta^\top x-\phi_\eta)\big))^2}. \tag{44}$$

$$\frac{\partial \mu_\eta(x)}{\partial \beta_\eta} = \frac{a_\eta x \exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)}{(1+\exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big))^2}. \tag{45}$$

For a non-leaf node (η) we have $\mathbb{P}[T_\eta(x)=a]$, to be given by:

$$\mathbb{P}[T_\eta(x)=a] = \mathbb{P}[B_\eta = 1]\mathbb{P}[T_{\mathcal{L}(\eta)}(x)=a] + \tag{46}$$
$$\mathbb{P}[B_\eta = 0]\mathbb{P}[T_{\mathcal{R}(\eta)}(x)=a] \tag{47}$$

The above can be expanded to give:

$$\mathbb{P}[T_\eta(x)=a] = \mu_\eta(x)\mathbb{P}[T_{\mathcal{L}(\eta)}(x)=a] + \tag{48}$$
$$(1-\mu_\eta(x))\mathbb{P}[T_{\mathcal{R}(\eta)}(x)=a] \tag{49}$$

Let the differential of the above with respect to a node level parameter such as $a_\eta, b_\eta, \phi_\eta$ be denoted as $\mu'_\eta(x)$, then we have that the above is given by:

$$\mathbb{P}'[T_\eta(x)=a] = \mu'_\eta(x)(\mathbb{P}[T_{\mathcal{L}(\eta)}(x)=a] - \mathbb{P}[T_{\mathcal{R}(\eta)}(x)=a]) \tag{50}$$

After the above calculation for θ representing non-leaf parameters, we see that the expression for:

$$\nabla_\theta \Bigg[\log \mathbb{P}[T_\theta(s_t)=a_t]\Bigg] = \tag{51}$$

$$\frac{\mu'_\eta(s_t)\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t)=a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t)=a_t]\Bigg)}{\Bigg[\mu_\eta(s_t)\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t)=a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t)=a_t]\Bigg) + \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t)=a_t]\Bigg]} \tag{52}$$

For node level parameters θ, we evaluate $\nabla_\theta J_\pi(\theta)$ which would be splitting the parameters into node and leaf level parameters, we then have the first equality $\nabla_\theta J(\theta)$ as given by **?**:

$$\nabla_\theta J_\pi(\theta) := \tag{53}$$

$$\mathbb{E}_{\tau\sim\pi_\theta}\Bigg[\sum_{t=0}^{T}\nabla_\theta\Bigg[\log\mathbb{P}[T_\theta(s_t)=a_t]\Bigg]R(\tau)\Bigg] := \tag{54}$$

$$\mathbb{E}_{\tau\sim\pi_\theta, P_t}\Bigg[\sum_{t=0}^{T}\frac{\mu'_\eta(s_t)\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t)=a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t)=a_t]\Bigg)R(\tau)}{\Bigg[\mu_\eta(s_t)\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t)=a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t)=a_t]\Bigg) + \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t)=a_t]\Bigg]}\Bigg] \tag{55}$$

On the other end, for a leaf node l, $\mathbb{P}[T_\eta(x)=a]$ evaluates as:

$$\mathbb{P}[T_l=a] = p_{al} \forall l \in Leaves(T), \forall a \in |A|-1 \tag{56}$$

For the final action $a_{|A|}$, we have that the probability of taking action $a_{|A|}$ looks like:

$$\mathbb{P}[T_l=a_{|A|}] = 1 - \sum_{a\in|A|-1} p_{al} \forall l \in Leaves(T) \tag{57}$$

For the leaf node l, we have that the differential of the log-probs to be given as below:

$$\nabla_{p_l}\Bigg[\log\mathbb{P}[T_l(s_t)=a_t]\Bigg] = \frac{1}{\mathbb{P}[T_l(s_t)=a_t]}\Bigg(\frac{\partial}{\partial p_{al}}\mathbb{P}[T_l(s_t)=a_t]\Bigg)_{a\in|A|} \tag{58}$$

Simplification of the differential works as follows:

$$\frac{\partial}{\partial p_{al}}\mathbb{P}[T_l(s_t) = a_t] = 1 \text{ if } a_t = a \tag{59}$$

Otherwise we have:

$$\frac{\partial}{\partial p_{al}}\mathbb{P}[T_l(s_t) = a_t] = -1 \text{ if } a_t = a_{|A|} \tag{60}$$

Hence the differential of the log-probs evaluates to:

$$\nabla_{p_l}\Bigg[\log \mathbb{P}[T_l(s_t) = a_t]\Bigg] = \frac{1}{\mathbb{P}[T_l(s_t) = a_t]}\Bigg(\mathbb{I}[a_t = a] - \mathbb{I}[a_t = a_{|A|}]\Bigg)_{a \in |A|-1} \tag{61}$$

For leaf level parameters we have that the parameter of interest is given by θ_l, here we have $\nabla_{\theta_l} J_\pi(\theta_l)$. In this case the gradient of the return is given by:

$$\nabla_{\theta_l} J_\pi(\theta_l) := \tag{62}$$

$$\mathbb{E}_{\tau \sim \pi_{\theta_l}}\Bigg[\sum_{t=0}^{T} \nabla_{\theta_l}\Bigg[\log \mathbb{P}[T_{\theta_l}(s_t) = a_t]\Bigg] R(\tau)\Bigg] := \tag{63}$$

$$\mathbb{E}_{\tau \sim \pi_{\theta_l}, P_t}\Bigg[\sum_{t=0}^{T} \frac{\nabla_{\theta_l}\mathbb{P}[T_{\theta_l}(s_t) = a_t]}{\mathbb{P}[T_{\theta_l}(s_t) = a_t]} R(\tau)\Bigg] \tag{64}$$

We expand the above for leaf level parameters l, this gives us:

$$\nabla_{\theta_l} J_\pi(\theta_l) := \tag{65}$$

$$\mathbb{E}_{\tau \sim \pi_{\theta_l}}\Bigg[\sum_{t=0}^{T} \nabla_{\theta_l}\Bigg[\log \mathbb{P}[T_{\theta_l}(s_t) = a_t]\Bigg] R(\tau)\Bigg] := \tag{66}$$

$$\mathbb{E}_{\tau \sim \pi_{\theta_l}, P_t}\Bigg[\sum_{t=0}^{T} \frac{1}{p_{a_t l}}\Bigg(\mathbb{I}[a_t = a] - \mathbb{I}[a_t = a_{|A|}]\Bigg)_{a \in |A|-1} R(\tau)\Bigg] \tag{67}$$

For node-level parameters $\theta \in \Theta$. The expression for the matrix product is given by:

$$Hx := \nabla_\theta(\nabla_\theta \bar{D}_{KL}(\theta||\theta_k)^T x) \tag{68}$$

We can simplify the above as follows:

$$\bar{D}_{KL}(\theta||\theta_k) = \mathbb{E}_{s \sim \theta_k}\Bigg[\sum_{a \in A} \mathbb{P}[T_\theta(s) = a] \log\Bigg(\frac{\mathbb{P}[T_\theta(s) = a]}{\mathbb{P}[T_{\theta_k}(s_t) = a_t]}\Bigg)\Bigg] \tag{69}$$

If we take the derivative ∇_θ of the above expression we get the following:

$$\mathbb{E}_{s \sim \theta_k}\Bigg[\sum_{a \in A}\Bigg(\nabla_\theta \mathbb{P}[T_\theta(s) = a]\Bigg(\log \mathbb{P}[T_\theta(s) = a] - \log \mathbb{P}[T_{\theta_k}(s_t) = a_t]\Bigg) \tag{70}$$

$$+\mathbb{P}[T_\theta(s) = a]\nabla_\theta \log \mathbb{P}[T_\theta(s) = a]\Bigg)\Bigg] \tag{71}$$

Simplification of the above gives us the following:

$$\mathbb{P}'[T_\eta(s) = a] = \mu'_\eta(s)\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s) = a] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s) = a]\Bigg) \tag{72}$$

Expansion of the above gives us:

$$\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}(\nabla_\theta\mu_\eta(\theta))\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg) \tag{73}$$

$$\Bigg(\log\frac{\mathbb{P}[T_\theta(s)=a]}{\mathbb{P}[T_{\theta_k}(s_t)=a_t]}+1\Bigg)\Bigg] \tag{74}$$

The log fractional term below can be further simplified as:

$$\log\frac{\mathbb{P}[T_\theta(s)=a]}{\mathbb{P}[T_{\theta_k}(s_t)=a_t]}= \tag{75}$$

$$\log\frac{\mu_\theta(s)[\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}=a]]+\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]}{\mu_{\theta_k}(s_t)[\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t)=a_t]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s_t)=a_t]]+\mathbb{P}[T_{\mathcal{R}(\eta)}(s_t)=a_t]} \tag{76}$$

We simplify the above as follows:

$$\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\nabla_\theta[\mu_\eta(\theta)]\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg) \tag{77}$$

$$\Bigg(1+\log\frac{\mu_\theta(s)[\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}=a]]+\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]}{\mu_{\theta_k}(s_t)[\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t)=a_t]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s_t)=a_t]]+\mathbb{P}[T_{\mathcal{R}(\eta)}(s_t)=a_t]}\Bigg)\Bigg)\Bigg] \tag{78}$$

The total expression for $\nabla_\theta\bar{D}_{KL}(\theta||\theta_k)$ is given by:

$$\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\nabla_\theta[\mu_\eta(\theta)]\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg)\Bigg(1+\log\frac{\mathbb{P}[T_\theta(s)=a]}{\mathbb{P}[T_{\theta_k}(s_t)=a_t]}\Bigg)\Bigg)\Bigg] \tag{79}$$

We expand out the above expression such that it gives us:

$$\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\nabla_\theta[\mu_\eta(\theta)]\sum_{a\in A}\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg)\Bigg(\log\frac{\mathbb{P}[T_{\theta(s)}=a]}{\mathbb{P}[T_{\theta_k}(s_t)=a_t]}\Bigg)\Bigg)\Bigg] \tag{80}$$

Since we would need to find the double derivative ∇^2_θ of the term $\bar{D}_{KL}(\theta||\theta_k)$, this is given as:

$$\nabla_\theta[\mu_\eta(\theta)]\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg)\Bigg(\log\frac{\mathbb{P}[T_\theta(s)=a]}{\mathbb{P}[T_{\theta_k}(s_t)=a_t]}\Bigg)\Bigg)\Bigg] \tag{81}$$

We would need to compute the ∇_θ derivative for the following; this looks as follows:

$$\nabla^2_\theta[\mu_\eta(\theta)]\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg)\Bigg(\log\frac{\mathbb{P}[T_\theta(s)=a]}{\mathbb{P}[T_{\theta_k}(s_t)=a_t]}\Bigg)\Bigg)\Bigg] \tag{82}$$

$$+\nabla_\theta[\mu_\eta(\theta)]\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg)\frac{\nabla_\theta\mathbb{P}[T_\theta(s)=a]}{\mathbb{P}[T_\theta(s)=a]}\Bigg)\Bigg] \tag{83}$$

We simplify the above expression as below:

$$\nabla^2_\theta[\mu_\eta(\theta)]\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg)\Bigg(\log\frac{\mathbb{P}[T_\theta(s)=a]}{\mathbb{P}[T_{\theta_k}(s_t)=a_t]}\Bigg)\Bigg)\Bigg] \tag{84}$$

$$+\nabla_\theta[\mu_\eta(\theta)]\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg)\nabla_\theta\Bigg(\log\mathbb{P}[T_\theta(s)=a]\Bigg)\Bigg] \tag{85}$$

We simplify the above term as follows:

$$\nabla^2_\theta[\mu_\eta(\theta)]\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg)\Bigg(\log\frac{\mathbb{P}[T_\theta(s)=a]}{\mathbb{P}[T_{\theta_k}(s_t)=a_t]}\Bigg)\Bigg)\Bigg] \tag{86}$$

$$+\nabla_\theta[\mu_\eta(\theta)]\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg)\frac{\nabla_\theta\mathbb{P}[T_\theta(s)=a]}{\mathbb{P}[T_\theta(s)=a]}\Bigg] \tag{87}$$

We have the following to be true for the probability of choosing an action a for a tree T_η.

$$\mathbb{P}[T_\eta(x)=a]=\mu_\eta(x)\mathbb{P}[T_{\mathcal{L}(\eta)}(x)=a]+ \tag{88}$$

$$(1-\mu_\eta(x))\mathbb{P}[T_{\mathcal{R}(\eta)}(x)=a] \tag{89}$$

We also have the derivative $\nabla_\theta\mathbb{P}[T_\theta(s)=a]$ to be given as:

$$\mathbb{P}'[T_\theta(s)=a]=\mu'_\eta(s)\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg) \tag{90}$$

The final expression then reads:

$$\nabla^2_\theta[\mu_\eta(s,\theta)]\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg)\Bigg(\log\frac{\mathbb{P}[T_\theta(s)=a]}{\mathbb{P}[T_{\theta_k}(s_t)=a_t]}\Bigg)\Bigg)\Bigg] \tag{91}$$

$$+(\nabla_\theta[\mu_\eta(s,\theta)])(\nabla_\theta[\mu_\eta(s,\theta)])^T\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Bigg)^2\frac{1}{\mathbb{P}[T_\theta(s)=a]}\Bigg)\Bigg] \tag{92}$$

Similarly for leaf-level parameters $\theta'\in\Theta$. The expression for the matrix product is given by:

$$Hx:=\nabla_{\theta'}(\nabla_{\theta'}\bar{D}_{KL}(\theta'||\theta'_k)^Tx) \tag{93}$$

Expansion of ∇_θ of the KL-divergence from 79 gives us the following:

$$\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\Bigg(\nabla_{\theta'}\mathbb{P}[T_{\theta'}(s)=a]\Bigg(\log\mathbb{P}[T_{\theta'}(s)=a]-\log\mathbb{P}[T_{\theta'_k}(s_t)=a_t]\Bigg) \tag{94}$$

$$+\mathbb{P}[T_{\theta'}(s)=a]\nabla_{\theta'}\log\mathbb{P}[T_{\theta'}(s)=a]\Bigg)\Bigg] \tag{95}$$

We simplify the above as follows:

$$\mathbb{E}_{s\sim\theta'_k}\Bigg[\sum_{a\in A}\Bigg(\nabla_{\theta'}\mathbb{P}[T_{\theta'}(s)=a]\Bigg(\log\frac{\mathbb{P}[T_{\theta'}(s)=a_t]}{\mathbb{P}[T_{\theta'_k}(s_t)=a_t]}+1\Bigg)\Bigg)\Bigg] \tag{96}$$

We recall the expressions for:

$$\nabla_{\theta'}\Bigg[\mathbb{P}[T_{\theta'}(s_t)=a]\Bigg]=\Bigg(\mathbb{I}[a_t=a]-\mathbb{I}[a_t=a_{|A|}]\Bigg)_{a\in|A|-1} \tag{97}$$

The probability $\mathbb{P}[T_{\theta'}(s_t)=a_t]$ at leaf l is given by:

$$p_{a_tl}\mathbb{I}[a_t=a_i\forall i\neq|A|]+(1-\sum_{a'\in A-a_{|A|}}p_{a'l})\mathbb{I}[a_t=a_{|A|}] \tag{98}$$

The probability $\mathbb{P}[T_{\theta'}(s_t) = a_t]$ at leaf l is also written as:

$$\mathbb{P}[T_{\theta'}(s_t) = a_t] = p_{a_t l} \sum_{i=1}^{|A|-1} \mathbb{I}[a_t = a_i] + (1 - \sum_{a' \in A - a_{|A|}} p_{a' l}) \mathbb{I}[a_t = a_{|A|}] \tag{99}$$

The above probability $\mathbb{P}[T_{\theta'}(s_t) = a_t]$ at leaf l can also be written as:

$$p_{a_t l}(1 - \mathbb{I}[a_t = a_{|A|}]) + (1 - \sum_{a' \in A - a_{|A|}} p_{a' l}) \mathbb{I}[a_t = a_{|A|}] \tag{100}$$

We also have the following to be true for $\mathbb{P}[T_{\theta'}(s_t) = a_t]$ at leaf l :

$$\mathbb{E}_{s \sim \theta'_k}\Big[\sum_{a \in A}\Big(\mathbb{I}[a_t = a] - \mathbb{I}[a_t = a_{|A|}]\Big)\Big(\log \frac{p_{a_t l}\mathbb{I}[a_t = a_i \forall i \neq |A|] + (1 - \sum_{a' \in A - a_{|A|}} p_{a' l})\mathbb{I}[a_t = a_{|A|}]}{p'_{a_t l}\mathbb{I}[a_t = a_i \forall i \neq |A|] + (1 - \sum_{a' \in A - a_{|A|}} p'_{a' l})\mathbb{I}[a_t = a_{|A|}]} + 1\Big)\Big] \tag{101}$$

To compute the Hessian $\nabla^2_{\theta'} \bar{D}_{KL}(\theta' || \theta'_k)$ we take the gradient $\nabla_{\theta'}$ for the above expression; this is given by:

$$\Big(\mathbb{E}_{s \sim \theta'_k}\Big[\sum_{a \in A}\Big(\mathbb{I}[a_t = a] - \mathbb{I}[a_t = a_{|A|}]\Big)\Big(\frac{p'_{a_t l}\mathbb{I}[a_t = a_i \forall i \neq |A|] + (1 - \sum_{a' \in A - a_{|A|}} p'_{a' l})\mathbb{I}[a_t = a_{|A|}]}{p_{a_t l}\mathbb{I}[a_t = a_i \forall i \neq |A|] + (1 - \sum_{a' \in A - a_{|A|}} p_{a' l})\mathbb{I}[a_t = a_{|A|}]}\Big) \tag{102}$$

$$\Big(\mathbb{I}[a_t = a_j] - \mathbb{I}[a_t = a_{|A|}]\Big)\Big]\Big)_{j=1}^{|A|-1} \tag{103}$$

We can simplify the above as follows:

$$\Big(\mathbb{E}_{s \sim \theta'_k}\Big[\Big(\frac{p'_{a_t l}\mathbb{I}[a_t = a_i \forall i \neq |A|] + (1 - \sum_{a' \in A - a_{|A|}} p'_{a' l})\mathbb{I}[a_t = a_{|A|}]}{p_{a_t l}\mathbb{I}[a_t = a_i \forall i \neq |A|] + (1 - \sum_{a' \in A - a_{|A|}} p_{a' l})\mathbb{I}[a_t = a_{|A|}]}\Big)\Big(\mathbb{I}[a_t = a_j] - \mathbb{I}[a_t = a_{|A|}]\Big) \tag{104}$$

$$\sum_{a \in A}\Big(\mathbb{I}[a_t = a] - \mathbb{I}[a_t = a_{|A|}]\Big)\Big]\Big)_{j=1}^{|A|-1} \tag{105}$$

Further simplification of the above gives us:

$$\Big(\mathbb{E}_{s \sim \theta'_k}\Big[\Big(\frac{p'_{a_t l}\mathbb{I}[a_t = a_i \forall i \neq |A|] + (1 - \sum_{a' \in A - a_{|A|}} p'_{a' l})\mathbb{I}[a_t = a_{|A|}]}{p_{a_t l}\mathbb{I}[a_t = a_i \forall i \neq |A|] + (1 - \sum_{a' \in A - a_{|A|}} p_{a' l})\mathbb{I}[a_t = a_{|A|}]}\Big)\Big(\mathbb{I}[a_t = a_j] - \mathbb{I}[a_t = a_{|A|}]\Big) \tag{106}$$

$$\Big(1 - |A|\mathbb{I}[a_t = a_{|A|}]\Big)\Big]\Big)_{j=1}^{|A|-1} \tag{107}$$

We aim to take ∇_p of the expression in 106

$$-\Bigg(\mathbb{E}_{s\sim\theta'_k}\Bigg[\frac{p'_{a_t l}\mathbb{I}[a_t=a_i\forall i\neq|A|]+(1-\sum_{a'\in A-a_{|A|}}p'_{a'l})\mathbb{I}[a_t=a_{|A|}]}{\Big(p_{a_t l}\mathbb{I}[a_t=a_i\forall i\neq|A|]+(1-\sum_{a'\in A-a_{|A|}}p_{a'l})\mathbb{I}[a_t=a_{|A|}]\Big)^2}\Big(\mathbb{I}[a_t=a_j]-\mathbb{I}[a_t=a_{|A|}]\Big) \tag{108}$$

$$\Big(\mathbb{I}[a_t=a_i]-\mathbb{I}[a_t=a_{|A|}]\Big)\Big(1-|A|\mathbb{I}[a_t=a_{|A|}]\Big)\Bigg]\Bigg)_{i=1,j=1}^{|A|-1,|A|-1} \tag{109}$$

4 Objective functions for decision tree TRPO

In this section, we list the objective and constraint maps for the trust region policy optimization algorithm with a policy parameterized using a decision tree. For the problem over an internal node η, we assume the previous parameters to be given by θ_{old}, and we would have to optimize for the new parameters θ. Further we assume that for the problem over a leaf node l, the parameters to optimize for are the probabilities of taking an action a at leaf node l which is denoted by p'_{al} which is the initial set of parameters

The non-linear objective for an internal node η as follows:

$$\arg_\theta\max\Bigg(\mathbb{E}_{\tau\sim\pi_\theta,P_t}\Bigg[\sum_{t=0}^{T}\frac{\nabla_\theta[\mu_\eta(s_t,\theta)].(\theta-\theta_{old})\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t)=a_t]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s_t)=a_t]\Big)R(\tau)}{\Big[\mu_\eta(s_t,\theta)\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t)=a_t]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s_t)=a_t]\Big)+\mathbb{P}[T_{\mathcal{R}(\eta)}(s_t)=a_t]\Big]}\Bigg]\Bigg) \tag{110}$$

The final constraint for the internal node η is given by:

$$\frac{1}{2}(\theta-\theta_{old})^T\Bigg(\nabla^2_\theta[\mu_\eta(s,\theta)]\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Big)\Big(\log\frac{\mathbb{P}[T_\theta(s)=a]}{\mathbb{P}[T_{\theta_k}(s)=a]}\Big)\Big)\Bigg] \tag{111}$$

$$+(\nabla_\theta[\mu_\eta(s,\theta)])(\nabla_\theta[\mu_\eta(s,\theta)])^T\mathbb{E}_{s\sim\theta_k}\Bigg[\sum_{a\in A}\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Big)^2\frac{1}{\mathbb{P}[T_\theta(s)=a]}\Big)\Bigg]\Bigg)(\theta-\theta_{old})\leq\delta \tag{112}$$

The linear multiplier for a leaf node l as follows:

$$\mathbb{E}_{\tau\sim\pi_{\theta_l},P_t}\Bigg[\sum_{t=0}^{T}\frac{1}{p_{a_t l}}\Big(\mathbb{I}[a_t=a]-\mathbb{I}[a_t=a_{|A|}]\Big).\Big(p_{al}-p'_{al}\Big)R(\tau)\Bigg] \tag{113}$$

The KL-divergence constraint condition for the leaf node l is given by:

$$-\frac{1}{2}\Big(p_{al}-p'_{al}\Big)^T\Bigg(\mathbb{E}_{s\sim\theta'_k}\Bigg[\frac{p'_{a_tl}\mathbb{I}[a_t=a_i\forall i\neq|A|]+(1-\sum_{a'\in A-a_{|A|}}p'_{a'l})\mathbb{I}[a_t=a_{|A|}]}{\Big(p_{a_tl}\mathbb{I}[a_t=a_i\forall i\neq|A|]+(1-\sum_{a'\in A-a_{|A|}}p_{a'l})\mathbb{I}[a_t=a_{|A|}]\Big)^2} \tag{114}$$

$$\Big(\mathbb{I}[a_t=a_j]-\mathbb{I}[a_t=a_{|A|}]\Big)\Big(\mathbb{I}[a_t=a_i]-\mathbb{I}[a_t=a_{|A|}]\Big) \tag{115}$$

$$\Big(1-|A|\mathbb{I}[a_t=a_{|A|}]\Big)\Bigg]\Bigg)\Big(p_{al}-p'_{al}\Big)\leq\delta \tag{116}$$

5 Interpretable TRPO Empirical Formulations

We list the empirical TRPO objective and constraints which looks as follows: The non-linear objective for an internal node η as follows:

$$\arg_\theta\max\Bigg(\sum_{i=1}^{N}\frac{1}{N}\Bigg[\sum_{t=0}^{T}\frac{\nabla_\theta[\mu_\eta(s_t^i,\theta)].(\theta-\theta_{old})\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i)=a_t^i]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a_t^i]\Big)R(\tau^i)}{\Big[\mu_\eta(s_t^i,\theta)\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i)=a_t^i]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a_t^i]\Big)+\mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a_t^i]\Big]}\Bigg]\Bigg) \tag{117}$$

The empirical evaluation for the constraint condition is given by:

$$\frac{1}{2}(\theta-\theta_{old})^T\Bigg(\sum_{t=1}^{T}\sum_{i=1}^{N}\nabla_\theta^2[\mu_\eta(s_t^i,\theta)]\Bigg[\sum_{a\in A}\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a]\Big)\Big(\log\frac{\mathbb{P}[T_\theta(s_t^i)=a]}{\mathbb{P}[T_{\theta_k}(s_t^i)=a]}\Big)\Big) \tag{118}$$

$$+\Bigg[(\nabla_\theta[\mu_\eta(s_t^i,\theta)])(\nabla_\theta[\mu_\eta(s_t^i,\theta)])^T\sum_{a\in A}\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a]\Big)^2\frac{1}{\mathbb{P}[T_\theta(s_t^i)=a]}\Big)\Bigg]\Bigg)(\theta-\theta_{old})\leq\delta \tag{119}$$

The objective and constraint formulation for a leaf node l is given by:

$$\sum_{t=0}^{T}\sum_{i=1}^{N}\frac{1}{\mathbb{P}[T_{\theta_l}(s_t^i)=a_t^i]}\Big(\mathbb{I}[a_t=a]-\mathbb{I}[a_t=a_{|A|}]\Big)_{a\in[A]}.\Big(p_{al}-p'_{al}\Big)_{a\in[a]}R(\tau^i) \tag{120}$$

The KL-divergence constraint condition for the leaf node l is given by:

$$-\frac{1}{2}\Big(p_{al}-p'_{al}\Big)^T_{a\in|A|-1}\Bigg(\sum_{t=1}^{T}\Bigg[\frac{p'_{a_tl}\mathbb{I}[a_t=a_i\forall i\neq|A|]+(1-\sum_{a'\in A-a_{|A|}}p'_{a'l})\mathbb{I}[a_t=a_{|A|}]}{\Big(p_{a_tl}\mathbb{I}[a_t=a_i\forall i\neq|A|]+(1-\sum_{a'\in A-a_{|A|}}p_{a'l})\mathbb{I}[a_t=a_{|A|}]\Big)^2} \tag{121}$$

$$\Big(\mathbb{I}[a_t=a_j]-\mathbb{I}[a_t=a_{|A|}]\Big)\Big(\mathbb{I}[a_t=a_i]-\mathbb{I}[a_t=a_{|A|}]\Big) \tag{122}$$

$$\Big(1-|A|\mathbb{I}[a_t=a_{|A|}]\Big)\Bigg]\Bigg)^{|A|-1,|A|-1}_{i,j=1}\Big(p_{al}-p'_{al}\Big)_{a\in|A|-1}\leq\delta \tag{123}$$

6 Interpretable TRPO Approximation

To implement the TRPO algorithm we must take a linear approximation of the objective and a double derivative approximation of the constraint condition. Recall the approximated TRPO has the following objective and constraint equations:

$$\max_{\theta} \nabla_{\theta} L_{\theta_{old}}(\theta)|_{\theta=\theta_{old}}.(\theta - \theta_{old}) \tag{124}$$

$$Subj : \frac{1}{2}(\theta_{old} - \theta)^T A(\theta_{old})(\theta_{old} - \theta) \leq \delta \tag{125}$$

In the above, we have:

$$A(\theta_{old})_{ij} = \frac{\partial}{\partial \theta_i}\frac{\partial}{\partial \theta_j}\mathbb{E}_{s\sim\rho_{\pi}}\Big[D_{KL}(\pi(.|s,\theta_{old})||\pi(.|s,\theta)\Big]\Big|_{\theta=\theta_{old}} \tag{126}$$

The updates for the above are then given by:

$$\theta_{new} = \theta_{old} + \frac{1}{\lambda}A(\theta_{old})^{-1}\nabla_{\theta}L(\theta)|_{\theta=\theta_{old}} \tag{127}$$

In the above, we have that λ is potentially a time dependent parameter.

7 Interpretable TRPO pseudocode

In this section, we list the pseudocode for the interpretable TRPO algorithm. The steps for the method are as follows:

1. Assume we have trajectories $\tau^n = [(s_0^n, a_0^n, r_0^n), (s_1^n, a_1^n, r_1^n), (s_2^n, a_2^n, r_2^n), ..., (s_T^n, a_T^n, r_T^n)]$ for $n \in [N]$.
2. Initialize a tree T of a corresponding height which has internal and leaf nodes.
3. Initialize constraint condition maps $\mu_{\eta}(x; a_{\eta}, b_{\eta}, \phi_{\eta})$ for each internal node η and random variables α_l over the action space A for each leaf node. In this step, we randomly initialize a_{η}, b_{η}, ϕ_{η}, and α_l for each leaf $l \in T$.
4. Perform the forward evaluation of the input; At each internal node η we perform the following calculations:
   (a) $T_{\eta}(s_t^n) = B_{\eta}(s_t^n)T_{\mathcal{L}(\eta)}(s_t^n) + (1 - B_{\eta}(s_t^n))T_{\mathcal{R}(\eta)}(s_t^n)$.
   In the above we have $B_{\eta}(s_t^n) \sim Ber(\mu_{\eta}(s_t^n; a_{\eta}, b_{\eta}, \phi_{\eta}))$;
   For each internal node η, We store the values the $\mu_{\eta}(s), B_{\eta}(s), T_{\eta}(s), T_{\mathcal{L}(\eta)}(s), T_{\mathcal{R}(\eta)}(s), \nabla_{\theta}\mu_{\eta}(s_t, \theta)$.
   (b) We reverse recursively store the probabilities; i.e we have:

$$\mathbb{P}[T_{\eta}(s) = a] = \mu_{\theta}(s)\mathbb{P}[T_{\mathcal{L}(\eta)}(s) = a] + (1 - \mu_{\theta}(s))\mathbb{P}[T_{\mathcal{R}(\eta)}(s) = a] \tag{128}$$

   (c) Store the average sum of rewards for each trajectory $R(\tau^i) := \sum_{t=1}^{T} \gamma^t r_t^i(s_t, a_t)$
5. When we are at a leaf node l, we sample from the random variable α_l which has probabilities p_{al} to pick the action of our choosing.

6. For each internal node η we calculate the objective function in a depth first manner which is given as:

$$\arg_\theta \max \Bigg(\sum_{i=1}^{N} \frac{1}{N} \Bigg[\sum_{t=0}^{T} \frac{\nabla_\theta[\mu_\eta(s_t^i,\theta)].(\theta-\theta_{old}) \Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i)=a_t^i] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a_t^i] \Bigg) R(\tau^i)}{\Bigg[\mu_\eta(s_t^i,\theta) \Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i)=a_t^i] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a_t^i] \Bigg) + \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a_t^i] \Bigg]} \Bigg] \Bigg) \tag{129}$$

7. Loop over each node η in a depth first manner, the constraint condition is evaluated:

$$\frac{1}{2}(\theta-\theta_{old})^T \Bigg(\sum_{t=1}^{T} \sum_{i=1}^{N} \nabla_\theta^2[\mu_\eta(s_t^i,\theta)] \Bigg[\sum_{a\in A} \Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i)=a] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a] \Bigg) \Bigg(\log \frac{\mathbb{P}[T_\theta(s_t^i)=a]}{\mathbb{P}[T_{\theta_k}(s_t^i)=a]} \Bigg) \Bigg] \tag{130}$$

$$+ \Bigg[(\nabla_\theta[\mu_\eta(s_t^i,\theta)])(\nabla_\theta[\mu_\eta(s_t^i,\theta)])^T \sum_{a\in A} \Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i)=a] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a] \Bigg)^2 \frac{1}{\mathbb{P}[T_\theta(s_t^i)=a]} \Bigg) \Bigg] \Bigg) \Bigg] (\theta-\theta_{old}) \leq \delta \tag{131}$$

8. Replace $\theta_{\eta,old}$ with the newly obtained parameters θ after we maximize the objective and obtain parameters which are in the appropriate trust region.
9. For each leaf node l, we calculate the objective which is given as:

$$\sum_{t=0}^{T} \sum_{i=1}^{N} \frac{1}{p_{a_t^i l}} \Bigg(\mathbb{I}[a_t^i=a] - \mathbb{I}[a_t=a_{|A|}] \Bigg)_{a\in[A]} . \Bigg(p_{al} - p'_{al} \Bigg)_{a\in[a]} R(\tau^i) \tag{132}$$

10. The constraint condition for leaf l is given by:

$$-\frac{1}{2} \Bigg(p_{al} - p'_{al} \Bigg)^T \Bigg(\mathbb{E}_{s\sim\theta'_k} \Bigg[\frac{p'_{a_t l} \mathbb{I}[a_t = a_i \forall i \neq |A|] + (1 - \sum_{a'\in A-a_{|A|}} p'_{a'l}) \mathbb{I}[a_t = a_{|A|}]}{\Bigg(p_{a_t l} \mathbb{I}[a_t = a_i \forall i \neq |A|] + (1 - \sum_{a'\in A-a_{|A|}} p_{a'l}) \mathbb{I}[a_t = a_{|A|}] \Bigg)^2} \tag{133}$$

$$\Bigg(\mathbb{I}[a_t=a_j] - \mathbb{I}[a_t=a_{|A|}] \Bigg) \Bigg(\mathbb{I}[a_t=a_i] - \mathbb{I}[a_t=a_{|A|}] \Bigg) \tag{134}$$

$$\Bigg(1 - |A| \mathbb{I}[a_t=a_{|A|}] \Bigg) \Bigg] \Bigg) \Bigg(p_{al} - p'_{al} \Bigg) \leq \delta \tag{135}$$

11. Replace $\theta_{l,old}$ with the newly obtained parameters θ after we solve maximize the objective and obtain parameters which are in the appropriate trust region.
12. Return to step 4. to restart the interpretable TRPO loop.

8 Solving the subproblems

In this section, we present simplifications of the subproblems present in steps 6-7 of the pseudocode, and steps 9-10.

8.1 Internal node subproblem

For each internal node η, calculate the objective function given as:

$$\arg_\theta \max \Bigg(\sum_{i=1}^{N} \frac{1}{N} \Bigg[\sum_{t=0}^{T} \frac{\nabla_\theta[\mu_\eta(s_t^i,\theta)].(\theta-\theta_{old}) \Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i)=a_t^i] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a_t^i] \Big) R(\tau^i)}{\Big[\mu_\eta(s_t^i,\theta) \Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i)=a_t^i] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a_t^i] \Big) + \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a_t^i] \Big]} \Bigg] \Bigg) \tag{136}$$

We have the following to be true for the activation selection function:

$$\mu_\eta(x;a_\eta,\beta_\eta,\phi_\eta) := \Big(1+\exp\Big(-a_\eta(\beta_\eta^T x-\phi_\eta) \Big) \Big)^{-1} \tag{137}$$

We also have:

$$\frac{\partial \mu_\eta(x)}{\partial a_\eta} = \frac{(\beta_\eta^\top x-\phi_\eta)\exp\big(-a_\eta(\beta_\eta^\top x-\phi_\eta)\big)}{(1+\exp\big(-a_\eta(\beta_\eta^\top x-\phi_\eta)\big))^2}. \tag{138}$$

Further derivatives include:

$$\frac{\partial \mu_\eta(x)}{\partial \beta_\eta} = \frac{a_\eta x\exp\big(-a_\eta(\beta_\eta^\top x-\phi_\eta)\big)}{(1+\exp\big(-a_\eta(\beta_\eta^\top x-\phi_\eta)\big))^2}. \tag{139}$$

$$\frac{\partial \mu_\eta(x)}{\partial \phi_\eta} = \frac{-a_\eta\exp\big(-a_\eta(\beta_\eta^\top x-\phi_\eta)\big)}{(1+\exp\big(-a_\eta(\beta_\eta^\top x-\phi_\eta)\big))^2}. \tag{140}$$

In the below, we have the matrix product to be given by:

$$\nabla_\theta[\mu_\eta(s_t^i,\theta)].(\theta-\theta_{old}) := \tag{141}$$

$$\frac{(\beta_\eta^\top s_t^i-\phi_\eta)\exp\big(-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)\big)}{(1+\exp\big(-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)\big))^2}(\alpha_\eta-\alpha_{\eta_{old}})+ \tag{142}$$

$$\frac{a_\eta s_t^i\exp\big(-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)\big)}{(1+\exp\big(-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)\big))^2}(\beta_\eta-\beta_{\eta_{old}})+ \tag{143}$$

$$\frac{-a_\eta\exp\big(-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)\big)}{(1+\exp\big(-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)\big))^2}(\phi_\eta-\phi_{\eta_{old}}) \tag{144}$$

The notation for the above works as follows:

$$c_L(s_t^i,a_t^i) = \mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i)=a_t^i] \tag{145}$$

$$c_R(s_t^i,a_t^i) = \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a_t^i] \tag{146}$$

$$\tag{147}$$

After using the above variables as input, the numerator for the objective is given by:

$$\Big((\beta_\eta^\top s_t^i-\phi_\eta)\exp\big(-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)\big)(\alpha_\eta-\alpha_{\eta_{old}})+ \tag{148}$$

$$a_\eta s_t^i\exp\big(-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)\big)(\beta_\eta-\beta_{\eta_{old}})+ \tag{149}$$

$$-a_\eta\exp\big(-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)\big)(\phi-\phi_{old})\Big)\Big(c_L(s_t^i,a_t^i)-c_R(s_t^i,a_t^i)\Big)R(\tau^i) \tag{150}$$

We simplify the above as follows:

$$\exp\big(-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)\big)\Bigg((\beta_\eta^\top s_t^i-\phi_\eta)(\alpha_\eta-\alpha_{\eta_{old}})+ \tag{151}$$

$$a_\eta s_t^i(\beta_\eta-\beta_{\eta_{old}})+ \tag{152}$$

$$-a_\eta(\phi-\phi_{old})\Bigg)\Bigg(c_L(s_t^i,a_t^i)-c_R(s_t^i,a_t^i)\Bigg)R(\tau^i) \tag{153}$$

The denominator would look like:

$$\Bigg(1+\exp\Bigg(-a_\eta(\beta_\eta^T s_t^i-\phi_\eta)\Bigg)\Bigg)^2\Bigg[\frac{1}{(1+\exp(-a_\eta(\beta_\eta^T s_t^i-\phi_\eta)))}(c_L(s_t^i,a_t^i)-c_R(s_t^i,a_t^i))+c_R(s_t^i,a_t^i)\Bigg] \tag{154}$$

The next simplification step for the above would look like:

$$\Bigg(1+\exp\Bigg(-a_\eta(\beta_\eta^T s_t^i-\phi_\eta)\Bigg)\Bigg)\Bigg[(c_L(s_t^i,a_t^i)-c_R(s_t^i,a_t^i))+c_R(s_t^i,a_t^i)(1+\exp(-a_\eta(\beta_\eta^T s_t^i-\phi_\eta)))\Bigg] \tag{155}$$

The denominator would then be given as:

$$\Bigg(1+\exp\Bigg(-a_\eta(\beta_\eta^T s_t^i-\phi_\eta)\Bigg)\Bigg)\Bigg[(c_L(s_t^i,a_t^i)+c_R(s_t^i,a_t^i)\exp\Bigg(-a_\eta(\beta_\eta^T s_t^i-\phi_\eta)\Bigg)\Bigg) \tag{156}$$

In total the numerator divided by the denominator $F(a_\eta,\beta_\eta,\phi_\eta)$ reads as follows:

$$\begin{aligned}\arg_{a_\eta,\beta_\eta,\phi_\eta}\max\sum_{i=1}^{N}\sum_{t=1}^{T}&\frac{e^{-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)}}{\big(1+e^{-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)}\big)\big(c_L+c_R e^{-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)}\big)}\\ &\times\Big[(\beta_\eta^\top s_t^i-\phi_\eta)(\alpha_\eta-\alpha_{\eta_{\text{old}}})+a_\eta s_t^i(\beta_\eta-\beta_{\eta_{\text{old}}})-a_\eta(\phi_\eta-\phi_{\eta_{\text{old}}})\Big]\\ &\times(c_L-c_R)\,R(\tau^i),\end{aligned} \tag{157}$$

We simplify the constraint conditions for the above objective:

We have the double derivative $\nabla_\theta^2[\mu_\eta(s_t^i,\theta)]$ to be given by a (3×3) matrix whose structure is as follows:

$$\begin{pmatrix}\frac{\partial^2\mu_\eta(s,\theta)}{\partial a_\eta^2} & \frac{\partial^2\mu_\eta(s,\theta)}{\partial a_\eta\partial\beta_\eta} & \frac{\partial^2\mu_\eta(s,\theta)}{\partial a_\eta\partial\phi_\eta}\\ \frac{\partial^2\mu_\eta(s,\theta)}{\partial a_\eta\partial\beta_\eta} & \frac{\partial^2\mu_\eta(s,\theta)}{\partial\beta_\eta^2} & \frac{\partial^2\mu_\eta(s,\theta)}{\partial\beta_\eta\partial\phi_\eta}\\ \frac{\partial^2\mu_\eta(s,\theta)}{\partial a_\eta\partial\phi_\eta} & \frac{\partial^2\mu_\eta(s,\theta)}{\partial\beta_\eta\partial\phi_\eta} & \frac{\partial^2\mu_\eta(s,\theta)}{\partial\phi_\eta^2}\end{pmatrix}$$

We compute each of the above expressions in the double derivative; to compute $\frac{\partial}{\partial a_\eta^2}$ we see that:

$$\frac{\partial\mu_\eta(x)}{\partial a_\eta}=\frac{z_\eta\exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2} \tag{158}$$

Where in the above we have $z_\eta=\beta^T x-\phi_\eta$; the double derivative is given as:

$$\frac{\partial^2\mu_\eta(x)}{\partial a_\eta^2}=\frac{-z_\eta^2\exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2}+z_\eta\exp(-a_\eta z_\eta)\Bigg(\frac{\partial}{\partial a_\eta}(1+\exp(-a_\eta z_\eta))^{-2}\Bigg) \tag{159}$$

Simplification of the above gives us:

$$\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta^2} = \frac{-z_\eta^2 \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2} - 2z_\eta^2 \exp(-2a_\eta z_\eta)\Big((1+\exp(-a_\eta z_\eta))^{-3}\Big) \tag{160}$$

The final step gives us:

$$\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta^2} = \frac{-z_\eta^2 \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2}\left[1 + \frac{2\exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))}\right] \tag{161}$$

Finally, the above gives us:

$$\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta^2} = \frac{-z_\eta^2 \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2}\left[\frac{1+3\exp(-a_\eta z_\eta)}{1+\exp(-a_\eta z_\eta)}\right] \tag{162}$$

Where in the above we have $z_\eta = \beta_\eta^T x - \phi_\eta$; We now compute the cross derivatives, recall we have $\frac{\partial^2}{\partial a_\eta \partial \beta_\eta}$

$$\frac{\partial}{\partial \beta_\eta}\frac{\partial \mu_\eta(x)}{\partial a_\eta} = \frac{\partial}{\partial \beta_\eta}\frac{z_\eta \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2} \tag{163}$$

Simplification of the above gives us:

$$\frac{\partial}{\partial \beta_\eta}\frac{\partial \mu_\eta(x)}{\partial a_\eta} = \frac{x \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2} - \frac{z_\eta x a_\eta \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2} - \frac{2x a_\eta z_\eta (\exp(-a_\eta z_\eta))^2}{(1+\exp(-a_\eta z_\eta))^3} \tag{164}$$

Simplification of the above gives us:

$$\frac{\partial}{\partial \beta_\eta}\frac{\partial \mu_\eta(x)}{\partial a_\eta} = \frac{x \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2}\left[1 - z_\eta a_\eta - \frac{2a_\eta z_\eta (\exp(-a_\eta z_\eta))}{(1+\exp(-a_\eta z_\eta))}\right] \tag{165}$$

Final simplification of the above gives us:

$$\frac{\partial}{\partial a_\eta}\left(\frac{\partial \mu_\eta(x)}{\partial \beta_\eta}\right) = \frac{x\ \exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)}{\big(1+\exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)\big)^3} \tag{166}$$

$$\times \Bigg[\big(1 - a_\eta(\beta_\eta^\top x - \phi_\eta)\big)\big(1+\exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)\big) \tag{167}$$

$$- 2a_\eta(\beta_\eta^\top x - \phi_\eta)\exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)\Bigg]. \tag{168}$$

We now compute the ϕ_η derivative for $\frac{\partial}{\partial a_\eta}$; recall

$$\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta \partial \phi_\eta} = \frac{\partial}{\partial \phi_\eta}\frac{z_\eta \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2} \tag{169}$$

The above is now simplified as:

$$\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta \partial \phi_\eta} = -1\frac{\exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2} + \frac{z_\eta a_\eta \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2} \tag{170}$$

$$- \frac{2a_\eta z_\eta (\exp(-a_\eta z_\eta))^2}{(1+\exp(-a_\eta z_\eta))^3} \tag{171}$$

When simplified the above term gives us:

$$\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta \partial \phi_\eta} = \frac{\exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2}\bigg(-1 + z_\eta a_\eta - \frac{2a_\eta z_\eta \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))}\bigg) \tag{172}$$

Final simplification for the above gives us the following:

$$\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta \partial \phi_\eta} = \frac{\exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^3}\bigg(-(1+\exp(-a_\eta z_\eta)) - z_\eta a_\eta (1-\exp(-a_\eta z_\eta))\bigg) \tag{173}$$

We now compute the double derivatives; the initial double derivative we compute is $\frac{\partial}{\partial^2 \beta}$:

$$\frac{\partial}{\partial \beta_\eta}\frac{\partial \mu_\eta(x)}{\partial \beta_\eta} = \frac{\partial}{\partial \beta_\eta}\frac{a_\eta x \exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)}{(1+\exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big))^2}. \tag{174}$$

The above double derivative ∇^2_β is finally given as:

$$\frac{2a_\eta^2 x x^T \exp(-a_\eta(\beta^T x - \phi_\eta))}{(1+\exp(-a_\eta(\beta^T x - \phi_\eta)))^3} - \frac{a_\eta^2 x x^T \exp(-a_\eta(\beta^T x - \phi_\eta))}{(1+\exp(-a_\eta(\beta_\eta^T x - \phi_\eta)))^2} \tag{175}$$

A final simplification of the double derivative ∇^2_β is given as:

$$\frac{a_\eta^2 x x^T \exp(-a_\eta(\beta^T x - \phi_\eta))(1-\exp(-a_\eta(\beta^T x - \phi_\eta)))}{(1+\exp(-a_\eta(\beta^T x - \phi_\eta)))^3} \tag{176}$$

We now compute the mixed derivative with β and ϕ_η; the β derivative looks as follows:

$$\frac{\partial}{\partial \phi_\eta}\frac{\partial \mu_\eta(x)}{\partial \beta_\eta} = \frac{\partial}{\partial \phi_\eta}\frac{a_\eta x \exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)}{(1+\exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big))^2}. \tag{177}$$

The final joint derivative $\frac{\partial^2}{\partial \phi_\eta \beta_\eta}$ looks as follows:

$$\frac{a_\eta^2 x \exp(-a_\eta z_\eta)(1+\exp(-a_\eta z_\eta))}{(1+\exp(-a_\eta z_\eta))^3} + \tag{178}$$

$$-\frac{2a_\eta^2 x (\exp(-a_\eta z_\eta))^2}{(1+\exp(-a_\eta z_\eta))^3} \tag{179}$$

Simplification of the $\frac{\partial}{\partial \phi_\eta \beta_\eta}$ term gives us:

$$\frac{a_\eta^2 x \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^3}\bigg[1+\exp(-a_\eta z_\eta) - 2\exp(-a_\eta z_\eta)\bigg] \tag{180}$$

The derivative $\frac{\partial}{\partial \phi_\eta \beta_\eta}$ looks as follows:

$$\frac{a_\eta^2 x \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^3}\bigg[1-\exp(-a_\eta z_\eta)\bigg] \tag{181}$$

The final mixed derivative we wish to evaluate is given by $\frac{\partial^2}{\partial^2 \phi_\eta}$:. The single derivative $\frac{\partial}{\partial^2 \phi_\eta}$

$$\frac{\partial^2 \mu_\eta(x)}{\partial \phi_\eta^2} = \frac{\partial}{\partial \phi_\eta} \frac{-a_\eta \exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)}{(1+\exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big))^2}. \tag{182}$$

The simplification of the above looks like:

$$\frac{\partial^2 \mu_\eta(x)}{\partial \phi_\eta^2} = \frac{-a_\eta^2 \exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)}{(1+\exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big))^2} \tag{183}$$

$$+2\frac{a_\eta^2(\exp(-a_\eta z_\eta))^2}{(1+\exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big))^3} \tag{184}$$

The above simplified gives us:

$$\frac{\partial^2 \mu_\eta(x)}{\partial \phi_\eta^2} = \frac{a_\eta^2 \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^2}\left[2\frac{\exp(-a_\eta z_\eta)}{1+\exp(-a_\eta z_\eta)} - 1\right] \tag{185}$$

The final expression for $\frac{\partial^2}{\partial \phi_\eta^2}$:

$$\frac{\partial^2 \mu_\eta(x)}{\partial \phi_\eta^2} = \frac{a_\eta^2 \exp(-a_\eta z_\eta)(\exp(-a_\eta z_\eta) - 1))}{(1+\exp(-a_\eta z_\eta))^3} \tag{186}$$

Our goal now would be to input the values of $\nabla_\theta^2[\mu_\eta(s_t^i,\theta)], \nabla_\theta \mu_\eta(s_t^i,\theta)$, and we introduce the following notations:

$$c_L(s_t^i, a_t^i) = \mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i) = a_t^i] \tag{187}$$

$$c_R(s_t^i, a_t^i) = \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i) = a_t^i] \tag{188}$$

$$p_\theta(s_t^i, a) = \mathbb{P}[T_\theta(s_t^i) = a] \tag{189}$$

$$p_{\theta_k}(s_t^i, a) = \mathbb{P}[T_{\theta_k}(s_t^i) = a] \tag{190}$$

Recall the constraint condition is given by:

$$\frac{1}{2}(\theta - \theta_{old})^T \Bigg(\sum_{t=1}^{T}\sum_{i=1}^{N} \nabla_\theta^2[\mu_\eta(s_t^i,\theta)]\Bigg[\sum_{a\in A}\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i)=a] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a]\Bigg)\Bigg(\log\frac{\mathbb{P}[T_\theta(s_t^i)=a]}{\mathbb{P}[T_{\theta_k}(s_t^i)=a]}\Bigg)\Bigg] \tag{191}$$

$$+\Bigg[(\nabla_\theta[\mu_\eta(s_t^i,\theta)])(\nabla_\theta[\mu_\eta(s_t^i,\theta)])^T \sum_{a\in A}\Bigg(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t^i)=a] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t^i)=a]\Bigg)^2 \frac{1}{\mathbb{P}[T_\theta(s_t^i)=a]}\Bigg)\Bigg]\Bigg](\theta - \theta_{old}) \leq \delta \tag{192}$$

The Hessian matrix which consists of the double derivatives of μ_η with respect to $a_\eta, b_\eta, \phi_\eta$ looks as follows; we first compute :

$$\sum_{i=1,j=1}^{3,3} v_i \frac{\partial \mu_\eta(s_t^i,\theta)}{\partial x_i \partial x_j} v_j \tag{193}$$

With the notation as per our setting, we have the non-mixed derivatives for 193 to be given as follows:

$$\Delta a \frac{\partial^2 \mu_\eta}{\partial a_\eta^2}\Delta a + \Delta\beta \frac{\partial^2 \mu_\eta}{\partial \beta_\eta^2}\Delta\beta + \Delta\phi \frac{\partial^2 \mu_\eta}{\partial \phi_\eta^2}\Delta\phi \tag{194}$$

$$+2\Delta a_\eta \frac{\partial^2 \mu_\eta}{\partial a_\eta \partial \beta_\eta}\Delta\beta_\eta + 2\Delta\beta_\eta \frac{\partial^2 \mu_\eta}{\partial \beta_\eta \partial \phi_\eta}\Delta\phi_\eta + 2\Delta\phi_\eta \frac{\partial^2 \mu_\eta}{\partial a_\eta \partial \phi_\eta}\Delta a_\eta \tag{195}$$

The evaluation of $(\theta - \theta_{old})^T \nabla^2_\theta [\mu_\eta(s^i_t, \theta)](\theta - \theta_{old})$ is given by the following:

$$\begin{aligned}\Delta^\top \nabla^2 \mu_\eta(x)\, \Delta = \frac{\exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^3} \Bigg\{ & - z_\eta^2 (1 + 3\exp(-a_\eta z_\eta)(\Delta a)^2 \\ & + a_\eta^2 \big(1 - \exp(-a_\eta z_\eta)\big) \Big[(x^\top \Delta\beta)^2 - (\Delta\phi)^2 \Big] \\ & - 2\Delta a\, \Delta\phi \Big[1 + \exp(-a_\eta z_\eta) + a_\eta z_\eta \big(1 - \exp(-a_\eta z_\eta)\big) \Big] \\ & + 2\Delta a (x^\top \Delta\beta) \Big[(1 - a_\eta z_\eta)\big(1 + \exp(-a_\eta z_\eta)\big) - 2 a_\eta z_\eta \exp(-a_\eta z_\eta) \Big] \\ & + 2 a_\eta^2 \Delta\phi (x^\top \Delta\beta)\big(1 - \exp(-a_\eta z_\eta)\big) \Bigg\}. \end{aligned} \tag{196}$$

We then evaluate the $(\theta - \theta_{old})^T \nabla_\theta \mu_\eta (\nabla_\theta \mu_\eta)^T (\theta - \theta_{old})$ which is:

$$(\theta - \theta_{\text{old}})^T \nabla_\theta \mu_\eta(x) (\nabla_\theta \mu_\eta(x))^T (\theta - \theta_{\text{old}}) = \frac{\exp(-2a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^4} \left(z_\eta \Delta a_\eta + a_\eta x^\top \Delta\beta_\eta - a_\eta \Delta\phi_\eta \right)^2, \tag{197}$$

Simplification of the constraint equation 191 now gives us the following; we aim to collect the coefficients of the squared terms $(\Delta a)^2, (\Delta\beta)^2, (\Delta\phi)^2$, and the cross-terms which are $(\Delta a \Delta\phi), (\Delta\beta \Delta a), (\Delta\phi \Delta\beta)$ separately. We see that we get the following to be true after we collect the terms.

1. The $(\Delta a)^2$ coefficient is given by:

$$\frac{e}{(1+e)^3} \Bigg[\sum_{i=1}^{N} \sum_{t=1}^{T} -z_\eta^2 (1 + 3e^{-a_\eta z_\eta}) \sum_a \Big(c_L(s^i_t, a) - c_R(s^i_t, a) \Big) \log \left(\frac{p_\theta(s^i_t, a)}{p_{\theta_k}(s^i_t, a)} \right) + \tag{198}$$

$$\frac{e^2}{(1+e)^4} z_\eta^2 \sum_a \frac{\Big(c_L(s^i_t, a) + c_R(s^i_t, a) - 2c_L(s^i_t, a) c_R(s^i_t, a) \Big)}{p_\theta(s^i_t, a)} \Bigg] \tag{199}$$

2. The $(x^T \Delta\beta)^2$ coefficient is given by:

$$\frac{e}{(1+e)^3} \sum_{i=1}^{N} \sum_{t=1}^{T} \Big[a_\eta^2 \Big(1 - \exp(-a_\eta z_\eta) \Big) \Big] + \tag{200}$$

$$\frac{e^2}{(1+e)^4} a_\eta^2 \sum_a \frac{\Big(c_L(s^i_t, a) + c_R(s^i_t, a) - 2c_L(s^i_t, a) c_R(s^i_t, a) \Big)}{p_\theta(s^i_t, a)} \tag{201}$$

3. The $(\Delta\phi)^2$ coefficient is given by:

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \frac{e}{(1+e)^3} [-a_\eta^2 (1 - e)] + \tag{202}$$

$$\frac{e^2}{(1+e)^4} a_\eta^2 \tag{203}$$

4. The $\Delta a x \Delta\beta$ coefficient is given by:

$$\frac{2e}{(1+e)^3} \Big[1 - a_\eta z_\eta + \exp(-a_\eta z_\eta) - 3 a_\eta z_\eta \exp(-a_\eta z_\eta) \Big] \tag{204}$$

$$+ \frac{2e^2 z_\eta a_\eta}{(1+e)^4} \tag{205}$$

5. The $\Delta a \Delta \phi_\eta$ coefficient is given as:

$$\frac{2e}{(1+e)^3}\left[1 - a_\eta z_\eta + \exp(-a_\eta z_\eta) - 3a_\eta z_\eta\right] \tag{206}$$

$$-\frac{2e^2 z_\eta a_\eta}{(1+e)^4} \tag{207}$$

6. The $\Delta \phi x \Delta \beta$ coefficient is given by:

$$\frac{2e}{(1+e)^3}\left[a_\eta^2\left(1-e\right)\right] \tag{208}$$

$$-\frac{2e^2 a_\eta^2}{(1+e)^4} \tag{209}$$

8.1.1 Approximate constraint condition

In this section, we come up with a special case where the constraint condition seems to be true.

Notation. Let

$$\Delta \coloneqq \theta - \theta_{\text{old}}, \qquad s \equiv s_t^i,$$

and define the action–difference weights

$$\Delta p_s(a) \coloneqq \mathbb{P}[T_{\mathcal{L}(\eta)}(s) = a] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s) = a],$$

as well as the log-ratio

$$\ell_s(a) \coloneqq \log \frac{\mathbb{P}[T_\theta(s) = a]}{\mathbb{P}[T_{\theta_k}(s) = a]}.$$

Quadratic form. Consider

$$\frac{1}{2}\Delta^T \sum_{t=1}^{T}\sum_{i=1}^{N}\Bigg[\nabla_\theta^2 \mu_\eta(s_t^i) \sum_{a\in A} \Delta p_\theta(s_t^i, a)\, \ell_{s_t^i}(a) + \nabla_\theta \mu_\eta(s_t^i) \nabla_\theta \mu_\eta(s_t^i)^T \sum_{a\in A} \frac{\Delta p_\theta(s_t^i, a)^2}{\mathbb{P}[T_\theta(s_t^i) = a]}\Bigg]\Delta. \tag{210}$$

Bounding the Hessian term. Assume the Hessian is uniformly bounded in operator norm:

$$\left\|\nabla_\theta^2 \mu_\eta(s)\right\|_{\text{op}} \le C_H.$$

Then

$$\Delta^T \nabla_\theta^2 \mu_\eta(s) \Delta \le C_H \|\Delta\|^2,$$

and therefore

$$\left|\Delta^T \nabla_\theta^2 \mu_\eta(s) \Delta \sum_{a\in A} \Delta p_s(a)\, \ell_s(a)\right| \le C_H \|\Delta\|^2 \sum_{a\in A} \left|\Delta p_s(a)\right|\left|\ell_s(a)\right|. \tag{211}$$

—

Bounding the gradient outer-product term. From the explicit gradient,

$$\nabla_\theta \mu_\eta(s) = \frac{\exp(-a_\eta z_\eta(s))}{(1+\exp(-a_\eta z_\eta(s)))^2} \begin{bmatrix} z_\eta(s) \\ a_\eta x \\ -a_\eta \end{bmatrix},$$

we obtain

$$\begin{aligned} \Delta^T \nabla_\theta \mu_\eta(s) \nabla_\theta \mu_\eta(s)^T \Delta &= \big(\Delta^T \nabla_\theta \mu_\eta(s)\big)^2 \\ &\le \frac{e^2}{(1+e)^4} \Big(z_\eta(s)\Delta a_\eta + a_\eta x^\top \Delta\beta_\eta - a_\eta \Delta\phi_\eta \Big)^2. \end{aligned} \tag{212}$$

Hence,

$$\begin{aligned} &\Delta^T \nabla_\theta \mu_\eta(s) \nabla_\theta \mu_\eta(s)^T \Delta \sum_{a\in A} \frac{\Delta p_s(a)^2}{\mathbb{P}[T_\theta(s)=a]} \\ &\quad \le \frac{e^2}{(1+e)^4} \Big(z_\eta(s)\Delta a_\eta + a_\eta x^\top \Delta\beta_\eta - a_\eta \Delta\phi_\eta \Big)^2 \sum_{a\in A} \frac{\Delta p_s(a)^2}{\mathbb{P}[T_\theta(s)=a]}. \end{aligned} \tag{213}$$

—

Combined bound. Putting both bounds together yields

$$\begin{aligned} &\frac{1}{2}\Delta^T \sum_{t=1}^{T}\sum_{i=1}^{N} \left[\nabla^2_\theta \mu_\eta(s_t^i) \sum_{a\in A} \Delta p_{s_t^i}(a)\, \ell_{s_t^i}(a) + \nabla_\theta \mu_\eta(s_t^i) \nabla_\theta \mu_\eta(s_t^i)^T \sum_{a\in A} \frac{\Delta p_{s_t^i}(a)^2}{\mathbb{P}[T_\theta(s_t^i)=a]} \right] \Delta \\ &\quad \le \frac{1}{2} \sum_{t=1}^{T}\sum_{i=1}^{N} \left[C_H \|\Delta\|^2 \sum_{a\in A} \big|\Delta p_{s_t^i}(a)\big| \big|\ell_{s_t^i}(a)\big| \right. \\ &\qquad \left. + \frac{e^2}{(1+e)^4} \Big(z_\eta(s_t^i)\Delta a_\eta + a_\eta x^\top \Delta\beta_\eta - a_\eta \Delta\phi_\eta \Big)^2 \sum_{a\in A} \frac{\Delta p_{s_t^i}(a)^2}{\mathbb{P}[T_\theta(s_t^i)=a]} \right]. \end{aligned} \tag{214}$$

Norm-based simplification. Using Cauchy-Schwarz, there exists $C_\mu > 0$ such that

$$\Big(z_\eta(s)\Delta a_\eta + a_\eta x^\top \Delta\beta_\eta - a_\eta \Delta\phi_\eta \Big)^2 \le C_\mu \|\Delta\|^2.$$

The LHS for the above can be written as a dot product between $v = [z_\eta(s), a_\eta x, -a_\eta]$ and $w = [\Delta a_\eta, \Delta\beta_\eta, \Delta\phi_\eta]$. The full expression would then read:

$$| < v, w > |^2 \le ||v||^2 ||w||^2 \tag{215}$$

For our setting, we would have $v = [z_\eta(s), a_\eta x, -a_\eta]$. In this setting we would have the bounds to read the following:

$$||v||^2 = z_\eta^2 + a_\eta^2 ||x||^2 + a_\eta^2 \tag{216}$$

The above implies that:

$$||v|| \le \sqrt{\bar{Z}^2 + A^2 ||x||^2 + A^2} \tag{217}$$

The above would imply that $C_\mu := \bar{Z}^2 + A^2||x||^2 + A^2$

Therefore,

$$\text{LHS} \le \frac{\|\Delta\|^2}{2} \sum_{t=1}^{T}\sum_{i=1}^{N} \left[C_H \sum_{a\in A} \big|\Delta p_{s_t^i}(a)\big| \big|\ell_{s_t^i}(a)\big| + C_\mu \frac{e^2}{(1+e)^4} \sum_{a\in A} \frac{(\Delta p_{s_t^i}(a))^2}{\mathbb{P}[T_\theta(s_t^i)=a]} \right]. \tag{218}$$

The term after the $||\Delta^2||$ can be bounded by assuming bounds for different quantities. We first assume:

$$\Delta p_{s_t^i}(a) \leq 2 \tag{219}$$

We attach the bounds from 219 into 218. This gives us the following:

$$\sum_{t=1}^{T}\sum_{i=1}^{N}\left[C_H \sum_{a\in A}\left|\ell_{s_t^i}(a)\right| + C_\mu \frac{4e^2}{(1+e)^4}\sum_{a\in A}\frac{1}{p_\theta(s_t^i,a)}\right]. \tag{220}$$

The final constraint reads as follows:

$$\frac{||\Delta||^2}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\left[C_H \sum_{a\in A}\left|\log\frac{p_\theta(s_t^i,a)}{p_{\theta_k}(s_t^i,a)}\right| + C_\mu \frac{4e^2}{(1+e)^4}\sum_{a\in A}\frac{1}{p_\theta(s_t^i,a)}\right]. \leq \delta \tag{221}$$

In the above we have:

$$p_\theta(s_t^i,a_t^i) = \mu_\eta(s_t^i;\theta)c_L(s_t^i,a_t^i) + (1-\mu_\eta(s_t^i,a_t^i;\theta))c_R(s_t^i,a_t^i) \tag{222}$$

We can impose the above bounds as in 221 by breaking 221 into $\sqrt{\delta}\times\sqrt{\delta}$ on the RHS. This gives us the following expressions:

$$\frac{||\Delta||^2}{2} \leq \sqrt{\delta} \tag{223}$$

$$C_H\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{a\in A}\left|\log\frac{p_\theta(s_t^i,a)}{p_{\theta_k}(s_t^i,a)}\right| + C_\mu\frac{4e^2}{(1+e)^4}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{a\in A}\frac{1}{p_\theta(s_t^i,a)} \leq \sqrt{\delta} \tag{224}$$

The $||\Delta||^2$ bound is then simplified as:

$$||\Delta|| \leq \sqrt{2}\delta^{\frac{1}{4}} \tag{225}$$

We can have the initial summation bounds broken up as:

$$C_H\left|\log\frac{p_\theta(s_t^i,a)}{p_{\theta_k}(s_t^i,a)}\right| \leq \frac{\sqrt{\delta}}{2NT|A|} \tag{226}$$

The above is now simplified as:

$$p_\theta(s_t^i,a) \leq \exp\left(\frac{\sqrt{\delta}}{2NT|A|C_H}\right)p_{\theta_k}(s_t^i,a) \tag{227}$$

The next part of the summation bounds is written as:

$$C_\mu\frac{4e^2}{(1+e)^4}\left|\frac{1}{p_\theta(s_t^i,a)}\right| \leq \frac{\sqrt{\delta}}{2NT|A|} \tag{228}$$

The above is simplified as:

$$\left|\frac{1}{p_\theta(s_t^i,a)}\right| \leq \frac{(1+e)^4\sqrt{\delta}}{8e^2NT|A|C_\mu} \tag{229}$$

The final simplification for the above is given as:

$$\left|p_\theta(s_t^i,a)\right| \geq \frac{8e^2NT|A|C_\mu}{(1+e)^4\sqrt{\delta}} \tag{230}$$

Below we establish bounds for C_H and C_μ.

Hessian block structure. Let the parameter vector be

$$\theta = (a_\eta, \beta_\eta, \phi_\eta) \in \mathbb{R}^{d+2},$$

and denote the margin

$$z_\eta(x) := \beta_\eta^\top x - \phi_\eta.$$

The Hessian of $\mu_\eta(x)$ with respect to θ admits the block decomposition

$$\nabla^2 \mu_\eta(x) = \begin{pmatrix} H_{aa} & H_{a\beta} & H_{a\phi} \\ H_{\beta a} & H_{\beta\beta} & H_{\beta\phi} \\ H_{\phi a} & H_{\phi\beta} & H_{\phi\phi} \end{pmatrix},$$

where $H_{aa} \in \mathbb{R}$, $H_{\beta\beta} \in \mathbb{R}^{d\times d}$, $H_{\phi\phi} \in \mathbb{R}$, and the remaining blocks are conformable mixed derivatives.

Assumptions. Assume that the covariates and parameters are uniformly bounded:

$$\|x\|_2 \le B, \qquad |a_\eta| \le A, \qquad |\beta_i| \le \bar{\beta}, \qquad |\phi_\eta| \le \bar{\phi}$$

for fixed constants $A, B, \bar{\beta}, \bar{\phi} > 0$.

With the above setup, we can bound:

$$|z_\eta| \le |\beta^T x - \phi| \le dB\bar{\beta} + \bar{\phi} \tag{231}$$

In the above we have that the dimension of the space is d. We wish to bound the exponential term as follows:

$$\exp(-a_\eta z_\eta) \le \exp(A(dB\bar{\beta} + \bar{\phi})) \tag{232}$$

We also have the following bounds for $(1 + \exp(-a_\eta z_\eta))^{-1}$ where we can show that:

$$(1 + \exp(-a_\eta z_\eta))^{-1} \le (1 + \exp(-A(dB\bar{\beta} + \bar{\phi})))^{-1} \tag{233}$$

Recall that we have $\frac{\partial^2}{\partial a_\eta^2}$ to be given as follows:

$$\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta^2} = \frac{-z_\eta^2 \exp(-a_\eta z_\eta)}{(1 + \exp(-a_\eta z_\eta))^2} \left[\frac{1 + 3\exp(-a_\eta z_\eta)}{1 + \exp(-a_\eta z_\eta)}\right] \tag{234}$$

We can bound the above as follows:

$$\left|\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta^2}\right| \le \frac{(\bar{\phi} + dB\bar{\beta})^2 \exp(A(dB\bar{\beta} + \bar{\phi}))}{(1 + \exp(-A(dB\bar{\beta} + \bar{\phi})))^2} \left[\frac{1 + 3\exp(A(dB\bar{\beta} + \bar{\phi}))}{1 + \exp(-A(dB\bar{\beta} + \bar{\phi}))}\right] \tag{235}$$

Simplification of the above when we expand out the reciprocal of $(1 + \exp(-a_\eta z_\eta))$ gives us the following:

$$\left|\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta^2}\right| \le \frac{(\bar{\phi} + dB\bar{\beta})^2 [\exp(A(dB\bar{\beta} + \bar{\phi}))]^4}{(1 + \exp(A(dB\bar{\beta} + \bar{\phi})))^2} \left[\frac{1 + 3\exp(A(dB\bar{\beta} + \bar{\phi}))}{1 + \exp(A(dB\bar{\beta} + \bar{\phi}))}\right] \tag{236}$$

We introduce more notation to obtain cleaner bounds, this is given by:

$$\left|\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta^2}\right| \le \frac{(\bar{\phi} + dB\bar{\beta})^2 [\exp(A(dB\bar{\beta} + \bar{\phi}))]^4}{(1 + \exp(A(dB\bar{\beta} + \bar{\phi})))^2} \left[\frac{1 + 3\exp(A(dB\bar{\beta} + \bar{\phi}))}{1 + \exp(A(dB\bar{\beta} + \bar{\phi}))}\right] \tag{237}$$

In the above we assume:

$$\bar{Z} := (\bar{\phi} + dB\bar{\beta}) \tag{238}$$
$$\bar{E} := \exp(A\bar{Z}) \tag{239}$$
$$\tag{240}$$

With the above notation, we see that the bounds for $\frac{\partial^2}{\partial a^2}$ is given by:

$$\left| \frac{\partial^2 \mu_\eta(x)}{\partial a_\eta^2} \right| \leq \frac{\bar{Z}^2 \bar{E}^4}{(1+\bar{E})^2} \left[\frac{1+3\bar{E}}{1+\bar{E}} \right] \tag{241}$$

The double derivative for $\frac{\partial}{\partial \beta_\eta^2}$ is expanded as:

$$\frac{a_\eta^2 x x^T \exp(-a_\eta(\beta^T x - \phi_\eta))(1 - \exp(-a_\eta(\beta^T x - \phi_\eta)))}{(1 + \exp(-a_\eta(\beta^T x - \phi_\eta)))^3} \tag{242}$$

We expand the 243 expression which gives us the following:

$$\frac{a_\eta^2 x x^T (\exp(a_\eta(\beta^T x - \phi_\eta)) - 1)}{\exp(a_\eta(\beta^T x - \phi_\eta))^2 \frac{(1+\exp(a_\eta(\beta^T x - \phi_\eta)))^3}{\exp(a_\eta(\beta^T x - \phi_\eta))^3}} \tag{243}$$

The above would then give the following:

$$\frac{a_\eta^2 x x^T (\exp(a_\eta(\beta^T x - \phi_\eta)) - 1) \exp(a_\eta(\beta^T x - \phi_\eta))}{(1 + \exp(a_\eta(\beta^T x - \phi_\eta)))^3} \tag{244}$$

We bound the above with the following:

$$\frac{a_\eta^2 ||x||^2 (\bar{E} - 1)\bar{E}}{(1 + \exp(a_\eta(\beta^T x - \phi_\eta)))^3} \tag{245}$$

The final bounds for $\frac{\partial}{\partial \beta_\eta^2}$ is given by:

$$\frac{A^2 ||x||^2 (\bar{E} - 1)\bar{E}}{(1 + \frac{1}{\bar{E}})^3} \tag{246}$$

The above expanded gives us the following upper bound:

$$\frac{A^2 ||x||^2 (\bar{E} - 1)\bar{E}^4}{(1 + \bar{E})^3} \tag{247}$$

The final bounds for $\frac{\partial^2}{\partial \phi^2}$ is given as:

$$\frac{\partial^2 \mu_\eta(x)}{\partial \phi_\eta^2} = \frac{a_\eta^2 \exp(-a_\eta z_\eta)(\exp(-a_\eta z_\eta) - 1))}{(1 + \exp(-a_\eta z_\eta))^3} \tag{248}$$

The above $\frac{\partial^2}{\partial \phi^2}$ term can be simplified as : final bounds for $\frac{\partial^2}{\partial \phi^2}$ is given as:

$$\frac{\partial^2 \mu_\eta(x)}{\partial \phi_\eta^2} = \frac{a_\eta^2(1-\exp(a_\eta z_\eta))}{\exp(a_\eta z_\eta)^2 \frac{(1+\exp(a_\eta z_\eta))^3}{\exp(a_\eta z_\eta)^3}} \tag{249}$$

The final simplification for the above is given by:

$$\frac{\partial^2 \mu_\eta(x)}{\partial \phi_\eta^2} = \frac{a_\eta^2(1-\exp(a_\eta z_\eta))\exp(a_\eta z_\eta)}{(1+\exp(a_\eta z_\eta))^3} \tag{250}$$

We can bound the above with the following quantity:

$$\frac{\partial^2 \mu_\eta(x)}{\partial \phi_\eta^2} = \frac{a_\eta^2(1+\bar{E})\bar{E}}{(1+\frac{1}{\bar{E}})^3} \tag{251}$$

We expand the above which gives us the following:

$$\frac{\partial^2 \mu_\eta(x)}{\partial \phi_\eta^2} = \frac{a_\eta^2 \bar{E}^4}{(1+\bar{E})^2} \tag{252}$$

We can compute an upper bound for 252 which is given by:

$$\frac{\partial^2 \mu_\eta(x)}{\partial \phi_\eta^2} \leq \frac{A^2 \bar{E}^4}{(1+\bar{E})^2} \tag{253}$$

We have the $\frac{\partial^2}{\partial a_\eta \partial \beta_\eta}$ term given as:

$$\frac{\partial}{\partial a_\eta}\left(\frac{\partial \mu_\eta(x)}{\partial \beta_\eta}\right) = \frac{x\,\exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)}{\big(1+\exp\big(-a_\eta(\beta_\eta^\top x-\phi_\eta)\big)\big)^3} \tag{254}$$

$$\times \Bigg[\big(1-a_\eta(\beta_\eta^\top x-\phi_\eta)\big)\big(1+\exp\big(-a_\eta(\beta_\eta^\top x-\phi_\eta)\big)\big) \tag{255}$$

$$-2a_\eta(\beta_\eta^\top x-\phi_\eta)\exp\big(-a_\eta(\beta_\eta^\top x-\phi_\eta)\big)\Bigg]. \tag{256}$$

We aim to bound the above as follows:

$$\frac{\partial}{\partial a_\eta}\left(\frac{\partial \mu_\eta(x)}{\partial \beta_\eta}\right) \leq \frac{||x||\,\bar{E}}{\big(1+\frac{1}{\bar{E}}\big)^3} \tag{257}$$

$$\times \Bigg[\big(1+\bar{Z}\big)(1+E) \tag{258}$$

$$+2\bar{Z}E\Bigg]. \tag{259}$$

The above can be bounded by:

$$\frac{\partial}{\partial a_\eta}\left(\frac{\partial \mu_\eta(x)}{\partial \beta_\eta}\right) \leq \frac{||x||\,\bar{E}^4}{\big(1+\bar{E}\big)^3} \tag{260}$$

$$\times \Bigg[1+\bar{Z}+E+3E\bar{Z}\Bigg]. \tag{261}$$

Recall the $\frac{\partial}{\partial a_\eta \phi_\eta}$ derivative term is given as:

$$\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta \partial \phi_\eta} = \frac{\exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^3}\bigg(-(1+\exp(-a_\eta z_\eta)) - z_\eta a_\eta (1-\exp(-a_\eta z_\eta))\bigg) \tag{262}$$

We can bound the above as :

$$|\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta \partial \phi_\eta}| \leq \frac{\bar{E}}{(1+\frac{1}{\bar{E}})^3}\bigg(1+\exp(-a_\eta z_\eta) - z_\eta a_\eta + z_\eta a_\eta \exp(-a_\eta z_\eta))\bigg) \tag{263}$$

We bound the above as follows:

$$|\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta \partial \phi_\eta}| \leq \frac{\bar{E}}{(1+\frac{1}{\bar{E}})^3}\bigg(1+\bar{E}+\bar{Z}A+\bar{Z}A\bar{E}\bigg) \tag{264}$$

The above can be further bounded as:

$$|\frac{\partial^2 \mu_\eta(x)}{\partial a_\eta \partial \phi_\eta}| \leq \frac{\bar{E}^4}{(1+\bar{E})^2}\bigg((\bar{Z}A+1)\bigg) \tag{265}$$

We can bound the $\beta_\eta \phi_\eta$ derivative term with the following expression:

$$\frac{\partial^2 \mu_\eta(x)}{\partial \beta_\eta \phi_\eta} \leq \frac{a_\eta^2 x \exp(-a_\eta z_\eta)}{(1+\exp(-a_\eta z_\eta))^3}\bigg[1-\exp(-a_\eta z_\eta)\bigg] \tag{266}$$

We bound the above term as follows:

$$\left\| \frac{\partial^2 \mu_\eta(x)}{\partial \beta_\eta \, \partial \phi_\eta} \right\| \;\leq\; \frac{A^2 \|x\| \, \bar{E}}{\left(1+\frac{1}{\bar{E}}\right))^3} \left[1+\bar{E}\right]. \tag{267}$$

We can bound the above as follows:

$$\left\| \frac{\partial^2 \mu_\eta(x)}{\partial \beta_\eta \, \partial \phi_\eta} \right\| \;\leq\; \frac{A^2 \|x\| \, \bar{E}^4}{\left(1+\bar{E}\right)^2} \tag{268}$$

Global spectral norm bound. Using the triangle inequality for operator norms of block matrices,

$$\|\nabla^2 \mu_\eta(x)\|_2 \leq \sum_{i,j} \|H_{ij}\|_2, \tag{269}$$

We obtain the following bounds for the

$$\begin{aligned}\|\nabla^2 \mu_\eta(x)\|_2 \leq \frac{\bar{Z}^2\bar{E}^4}{(1+\bar{E})^2}\left[\frac{1+3\bar{E}}{1+\bar{E}}\right] + \frac{A^2\bar{E}^4}{(1+\bar{E})^2} + \frac{A^2||x||^2(\bar{E}-1)\bar{E}^4}{(1+\bar{E})^3} \\ +2\frac{A^2\|x\|\,\bar{E}^4}{\left(1+\bar{E}\right)^2} + \frac{2\bar{E}^4}{(1+\bar{E})^2}\bigg(\bar{Z}A+1\bigg) + \frac{||x||\,\bar{E}^4}{\left(1+\bar{E}\right)^3}\bigg[1+\bar{Z}+E+3E\bar{Z}\bigg]\end{aligned}$$

In the above we have that:

$$\begin{aligned}C_H := \frac{\bar{Z}^2\bar{E}^4}{(1+\bar{E})^2}\left[\frac{1+3\bar{E}}{1+\bar{E}}\right] + \frac{A^2\bar{E}^4}{(1+\bar{E})^2} + \frac{A^2||x||^2(\bar{E}-1)\bar{E}^4}{(1+\bar{E})^3} \\ +2\frac{A^2\|x\|\,\bar{E}^4}{\left(1+\bar{E}\right)^2} + \frac{2\bar{E}^4}{(1+\bar{E})^2}\bigg(\bar{Z}A+1\bigg) + \frac{||x||\,\bar{E}^4}{\left(1+\bar{E}\right)^3}\bigg[1+\bar{Z}+E+3E\bar{Z}\bigg]\end{aligned}$$

8.2 Interpretable TRPO node problem

The full optimization problem for an internal node η is given by:

$$\begin{aligned}\arg_{a_\eta,\beta_\eta,\phi_\eta}\max \sum_{i=1}^{N}\sum_{t=1}^{T} &\frac{e^{-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)}}{\big(1+e^{-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)}\big)\big(c_L+c_R e^{-a_\eta(\beta_\eta^\top s_t^i-\phi_\eta)}\big)}\\ &\times\Big[(\beta_\eta^\top s_t^i-\phi_\eta)(a_\eta-a_{\eta_{\text{old}}})+a_\eta s_t^i(\beta_\eta-\beta_{\eta_{\text{old}}})-a_\eta(\phi_\eta-\phi_{\eta_{\text{old}}})\Big]\\ &\times(c_L-c_R)\,R(\tau^i),\end{aligned} \tag{270}$$

The above is subject to the following constraints which govern how much p_θ and the constraints θ are allowed to be modified.

$$\Big|p_\theta(s_t^i,a)\Big| \le \frac{NT|A|C_\mu}{8\sqrt{\delta}} \tag{271}$$

$$p_\theta(s_t^i,a) \le \exp\left(\frac{\sqrt{\delta}}{NT|A|C_H}\right)p_{\theta_k}(s_t^i,a) \tag{272}$$

$$\Delta a_\eta^2+\Delta\beta_\eta^2+\Delta\phi_\eta^2 \le \sqrt{\delta} \tag{273}$$

To solve the above, we would employ a gradient ascent technique which is described below:

Gradient-based optimization of the objective. **Step 1: Compact reformulation.** Define

$$\theta_\eta := (a_\eta,\beta_\eta,\phi_\eta), \qquad z_t^i := \beta_\eta^\top s_t^i-\phi_\eta.$$

Let the objective be written as

$$J(\theta_\eta)=\sum_{i=1}^{N}\sum_{t=1}^{T} w_t^i(\theta_\eta)\,\Delta_t^i(\theta_\eta),$$

where

$$w_t^i(\theta_\eta)=\frac{e^{-a_\eta z_t^i}}{\big(1+e^{-a_\eta z_t^i}\big)\big(c_L(s_t^i,a_t^i)+c_R(s_t^i,a_t^i)e^{-a_\eta z_t^i}\big)}\big(c_L(s_t^i,a_t^i)-c_R(s_t^i,a_t^i)\big)R(\tau^i),$$

and

$$\Delta_t^i(\theta_\eta)=z_t^i\big(a_\eta-a_{\eta_{\text{old}}}\big)+a_\eta\, s_t^{i\top}\big(\beta_\eta-\beta_{\eta_{\text{old}}}\big)-a_\eta\big(\phi_\eta-\phi_{\eta_{\text{old}}}\big).$$

Step 2: Gradient computation. The gradient of J is obtained using the product rule,

$$\nabla_{\theta_\eta}J(\theta_\eta)=\sum_{i=1}^{N}\sum_{t=1}^{T}\Big(\nabla_{\theta_\eta}w_t^i(\theta_\eta)\,\Delta_t^i(\theta_\eta)+w_t^i(\theta_\eta)\,\nabla_{\theta_\eta}\Delta_t^i(\theta_\eta)\Big).$$

The required partial derivatives of Δ_t^i are

$$\begin{aligned}\nabla_{a_\eta}\Delta_t^i &= z_t^i+s_t^{i\top}(\beta_\eta-\beta_{\eta_{\text{old}}})-(\phi_\eta-\phi_{\eta_{\text{old}}}),\\ \nabla_{\beta_\eta}\Delta_t^i &= a_\eta s_t^i,\\ \nabla_{\phi_\eta}\Delta_t^i &= -a_\eta.\end{aligned}$$

Moreover, since

$$\frac{\partial z_t^i}{\partial \beta_\eta} = s_t^i, \qquad \frac{\partial z_t^i}{\partial \phi_\eta} = -1$$

To compute the gradients of w_t^i with respect to $a_\eta, \beta_\eta, \phi_\eta$ we rewrite $w_t^i(\theta_\eta)$ as follows:

$$w_t^i(\theta_\eta) = \frac{\exp(a_\eta z_t^i)}{(1+\exp(a_\eta z_t^i))(c_R(s_t^i,a_t^i)+c_L(s_t^i,a_t^i)\exp(a_\eta z_t^i))}(c_L(s_t^i,a_t^i)-c_R(s_t^i,a_t^i))R(\tau^i) \tag{274}$$

The differential of $w_t^i(\theta_\eta)$ with respect to a_η is given by:

$$\nabla_{a_\eta} w_t^i(\theta_\eta) = (c_L(s_t^i,a_t^i)-c_R(s_t^i,a_t^i))R(\tau^i)\frac{z_t^i\exp(a_\eta z_t^i)}{(1+\exp(a_\eta z_t^i))(c_R(s_t^i,a_t^i)+c_L(s_t^i,a_t^i)\exp(a_\eta z_t^i))} \tag{275}$$

$$-(c_L(s_t^i,a_t^i)-c_R(s_t^i,a_t^i))R(\tau^i)\exp(a_\eta z_t^i)\frac{z_t^i\exp(a_\eta z_t^i)(c_R(s_t^i,a_t^i)+c_L(s_t^i,a_t^i)\exp(a_\eta z_t^i))+(1+\exp(a_\eta z_t^i))c_L(s_t^i,a_t^i)z_t^i}{(1+\exp(a_\eta z_t^i))(c_R(s_t^i,a_t^i)+c_L(s_t^i,a_t^i)\exp(a_\eta z_t^i))^2} \tag{276}$$

We simplify the above as follows:

$$\nabla_{a_\eta} w_t^i(\theta_\eta) = \frac{(c_L(s_t^i,a_t^i)-c_R(s_t^i,a_t^i))R(\tau^i)z_t^i\exp(a_\eta z_t^i)}{(1+\exp(a_\eta z_t^i))(c_R(s_t^i,a_t^i)+c_L(s_t^i,a_t^i)\exp(a_\eta z_t^i))}\Bigg[1-\frac{(c_R(s_t^i,a_t^i)+c_L(s_t^i,a_t^i)\exp(a_\eta z_t^i))+(1+\exp(a_\eta z_t^i))c_L(s_t^i,a_t^i)}{(1+\exp(a_\eta z_t^i))(c_R(s_t^i,a_t^i)+c_L(s_t^i,a_t^i)\exp(a_\eta z_t^i))} \tag{277}$$

We simplify the above expression which gives us:

$$\nabla_{a_\eta} w_t^i(\theta_\eta) = \frac{(c_L(s_t^i,a_t^i)-c_R(s_t^i,a_t^i))R(\tau^i)z_t^i\exp(a_\eta z_t^i)}{(1+\exp(a_\eta z_t^i))(c_R(s_t^i,a_t^i)+c_L(s_t^i,a_t^i)\exp(a_\eta z_t^i))}\Bigg[1-\frac{(c_R(s_t^i,a_t^i)+c_L(s_t^i,a_t^i)\exp(a_\eta z_t^i))+(1+\exp(a_\eta z_t^i))c_L(s_t^i,a_t^i)}{(1+\exp(a_\eta z_t^i))(c_R(s_t^i,a_t^i)+c_L(s_t^i,a_t^i)\exp(a_\eta z_t^i))} \tag{278}$$

Simplification of the above yields the following:

$$\nabla_{a_\eta} w_t^i(\theta_\eta) = \frac{(c_L(s_t^i,a_t^i)-c_R(s_t^i,a_t^i))R(\tau^i)z_t^i\exp(a_\eta z_t^i)}{((1+\exp(a_\eta z_t^i))(c_R(s_t^i,a_t^i)+c_L(s_t^i,a_t^i)\exp(a_\eta z_t^i)))^2} \tag{279}$$

$$\Bigg[\exp(a_\eta z_t^i)(c_R(s_t^i,a_t^i)-c_L(s_t^i,a_t^i))+c_L(s_t^i,a_t^i)(\exp^2(a_\eta z_t^i)-1)\Bigg] \tag{280}$$

For w_t^i we now compute $\nabla_{\beta_\eta} w_t^i$; this gives us the following:

$$\nabla_{\beta_\eta} w_t^i(\theta_\eta) = \frac{(c_L(s_t^i,a_t^i)-c_R(s_t^i,a_t^i))R(\tau^i)a_\eta x\exp(a_\eta z_t^i)}{((1+\exp(a_\eta z_t^i))(c_R(s_t^i,a_t^i)+c_L(s_t^i,a_t^i)\exp(a_\eta z_t^i)))^2} \tag{281}$$

$$\Bigg[\exp(a_\eta z_t^i)(c_R(s_t^i,a_t^i)-c_L(s_t^i,a_t^i))+c_L(s_t^i,a_t^i)(\exp^2(a_\eta z_t^i)-1)\Bigg] \tag{282}$$

We also compute $\nabla_{\phi_\eta} w_t^i$ which gives us the following:

$$\nabla_{\phi_\eta} w_t^i(\theta_\eta) = \frac{-(c_L(s_t^i,a_t^i)-c_R(s_t^i,a_t^i))R(\tau^i)a_\eta\exp(a_\eta z_t^i)}{((1+\exp(a_\eta z_t^i))(c_R(s_t^i,a_t^i)+c_L(s_t^i,a_t^i)\exp(a_\eta z_t^i)))^2} \tag{283}$$

$$\Bigg[\exp(a_\eta z_t^i)(c_R(s_t^i,a_t^i)-c_L(s_t^i,a_t^i))+c_L(s_t^i,a_t^i)(\exp^2(a_\eta z_t^i)-1)\Bigg] \tag{284}$$

We now recall that the total gradient is given by:

$$\nabla_{\theta_\eta} J(\theta_\eta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \Big(\nabla_{\theta_\eta} w_t^i(\theta_\eta)\, \Delta_t^i(\theta_\eta) + w_t^i(\theta_\eta)\, \nabla_{\theta_\eta} \Delta_t^i(\theta_\eta) \Big) \tag{285}$$

Step 3: Gradient ascent algorithm. Starting from an initial parameter vector $\theta_\eta^{(0)} = (a_\eta^{(0)}, \beta_\eta^{(0)}, \phi_\eta^{(0)})$ with $a_\eta^{(0)} > 0$, iterate for $k = 0, 1, 2, \ldots$

$$\begin{aligned} a_\eta^{(k+1)} &= a_\eta^{(k)} + \gamma_k \nabla_{a_\eta} J(\theta_\eta^{(k)}), \\ \beta_\eta^{(k+1)} &= \beta_\eta^{(k)} + \gamma_k \nabla_{\beta_\eta} J(\theta_\eta^{(k)}), \\ \phi_\eta^{(k+1)} &= \phi_\eta^{(k)} + \gamma_k \nabla_{\phi_\eta} J(\theta_\eta^{(k)}), \end{aligned}$$

where $\gamma_k > 0$ is a step size.

To enforce the constraint $a_\eta > 0$, either project

$$a_\eta^{(k+1)} \leftarrow \max\{a_\eta^{(k+1)}, \varepsilon\}, \qquad \varepsilon > 0,$$

or reparameterize $a_\eta = \exp(\tilde{a}_\eta)$ and perform unconstrained updates on $\tilde{a}_\eta$.

The iterations are terminated when a convergence criterion is satisfied, and the final iterate

$$(\hat{a}_\eta, \hat{\beta}_\eta, \hat{\phi}_\eta) = \theta_\eta^{(k)}$$

is returned as an approximate solution to

$$\arg \max_{a_\eta, \beta_\eta, \phi_\eta} J(\theta_\eta).$$

We would run the above procedure 8.2 for T iterations.

8.3 Leaf Node Subproblem

In this subsection, we list the subproblem we get at the level of the leaf and devise strategies to solve this for the interpretable TRPO.

1. For each leaf node l, we calculate the objective which is given as:

$$\sum_{t=0}^{T} \sum_{i=1}^{N} \frac{1}{p_{a_t^i l}} \Big(\mathbb{I}[a_t^i = a] - \mathbb{I}[a_t^i = a_{|A|}] \Big)_{a \in [A]} . \Big(p_{al} - p'_{al} \Big)_{a \in [a]} R(\tau^i) \tag{286}$$

2. The constraint condition for leaf l is given by:

$$-\frac{1}{2} \Big(p_{al} - p'_{al} \Big)_{a \in |A|-1}^{T} \Bigg(\sum_{i=1}^{N} \sum_{t=1}^{T} \Bigg[\frac{p'_{a_t^i l} \mathbb{I}[a_t^i = a_k \forall k \neq |A|] + (1 - \sum_{a \in A - a_{|A|}} p'_{a_t^i l}) \mathbb{I}[a_t^i = a_{|A|}]}{\Big(p_{a_t^i l} \mathbb{I}[a_t^i = a_k \forall k \neq |A|] + (1 - \sum_{a' \in A - a_{|A|}} p_{a' l}) \mathbb{I}[a_t^i = a_{|A|}] \Big)^2} \tag{287}$$

$$\Big(\mathbb{I}[a_t^i = a_j] - \mathbb{I}[a_t^i = a_{|A|}] \Big) \Big(\mathbb{I}[a_t = a_i] - \mathbb{I}[a_t = a_{|A|}] \Big) \tag{288}$$

$$\Big(1 - |A| \mathbb{I}[a_t = a_{|A|}] \Big) \Bigg] \Bigg)_{i,j=1}^{|A|-1, |A|-1} \Big(p_{al} - p'_{al} \Big)_{a \in |A|-1} \leq \delta \tag{289}$$

We begin the calculations by simplification of the objective 321. After expansion of the dot product $(p_{al} - p'_{al}).\Big(\mathbb{I}[a_t = a] - \mathbb{I}[a_t = a_{|A|}]\Big)$ and after we input $\mathbb{P}[T_\theta(s_t^i) = a_t^i] = p_{a_t^i l}$, we get the following leaf objective:

$$\sum_{t=0}^{T}\sum_{i=1}^{N}\frac{1}{p_{a_t^i l}}\sum_{j=1}^{|A|}\Big(\mathbb{I}[a_t^i = a_j] - \mathbb{I}[a_t^i = a_{|A|}]\Big).\Big(p_{a_j l} - p'_{a_j l}\Big)R(\tau^i) \tag{290}$$

We simplify the dot product above as:

$$\sum_{j=1}^{|A|}\Big(\mathbb{I}[a_t^i = a_j] - \mathbb{I}[a_t^i = a_{|A|}]\Big).\Big(p_{a_j l} - p'_{a_j l}\Big) = \tag{291}$$

$$\sum_{j=1}^{|A|}\Big(\mathbb{I}[a_t^i = a_j]\Big(p_{a_j l} - p'_{a_j l}\Big)\Big) \tag{292}$$

This implies that the modified leaf objective now looks like:

$$\sum_{t=0}^{T}\sum_{i=1}^{N}\frac{1}{p_{a_{tl}^i}}R(\tau^i)\sum_{j=1}^{|A|}\Big(\mathbb{I}[a_t^i = a_j]\Big(p_{a_j l} - p'_{a_j l}\Big)\Big) \tag{293}$$

We would again employ a gradient based optimization technique where we take the differential of 320 with respect to $\partial_{p_{a_j l}}$ in order to find the maxima of 320 with respect to $\nabla_{p_{a_j l}}$. The differential of 320 is given by:

$$\frac{\partial}{\partial p_{a_j l}}\Big[\sum_{t=0}^{T}\sum_{i=1}^{N}R(\tau^i)\sum_{j=1}^{|A|}\Big(\mathbb{I}[a_t^i = a_j]\frac{p_{a_j l} - p'_{a_j l}}{p_{a_j l}}\Big)\Big] \tag{294}$$

Expansion of the differential above gives us the following:

$$\Big[2\sum_{t=0}^{T}\sum_{i=1}^{N}R(\tau^i)\sum_{j=1}^{|A|}\Big(\mathbb{I}[a_t^i = a_j]\frac{1}{p_{a_j l}^2}\Big)\Big] \tag{295}$$

Traditionally one would perform traditional gradient ascent to compute the modified p_{al} parameter; a potential update would look like:

$$p_{al} \to p_{al} + \Big[2\sum_{t=0}^{T}\sum_{i=1}^{N}R(\tau^i)\sum_{j=1}^{|A|}\Big(\mathbb{I}[a_t^i = a_j]\frac{1}{p_{a_j l}^2}\Big)\Big] \tag{296}$$

However, in our setting we have that $p_{.l}$ lies on the simplex and the updates must remember the same; i.e we always have:

$$\sum_{a} p_{al} = 1 \tag{297}$$

We recall the mirror descent on the simplex Δ_n; Here we have to perform :

$$\max_{p\in\Delta_d} J(p) \tag{298}$$

where we have $\Delta_d := \{x : \sum_{i=1}^{d} x_i = 1\}$. Here we would also assume J is concave/convex and we have bounded gradients; i.e $||\nabla J(p)||_\infty \leq G$. Recall the mirror ascent step to be given by:

$$p_i^{(k+1)} = \frac{p_i^{(k)} \exp(\eta g_i^{(k)})}{\sum_j p_j^{(k)} \exp(\eta g_j^{(k)})} \tag{299}$$

In the above we have:

$$g_i^{(k)} := \frac{\partial J(p^{(k)})}{\partial p_i} \tag{300}$$

For the reinforcement learning setup, we have the objective function being given as $J(p) = J_\pi(\theta_l)$; where we have $\theta_l = (p_{al})_{a \in A}$. With this definition, we see that:

$$\nabla_{p_l}\Bigg(\sum_{t=0}^{T}\sum_{i=1}^{N}\frac{1}{p_{a_t^i l}}\Bigg(\mathbb{I}[a_t^i = a] - \mathbb{I}[a_t^i = a_{|A|}]\Bigg)_{a\in[A]}.\Bigg(p_{al} - p'_{al}\Bigg)_{a\in[a]} R(\tau^i)\Bigg) \tag{301}$$

The derivative of $J(\theta_l)$ with respect to ∇_{θ_l} is given by:

$$\nabla_{p_l}\Bigg(\sum_{t=0}^{T}\sum_{i=1}^{N}\frac{1}{p_{a_t^i l}}\sum_{a\in A}\mathbb{I}[a_t^i = a]\Bigg(p_{al} - p'_{al}\Bigg)R(\tau^i)\Bigg) \tag{302}$$

We further simplify the above expression to give us:

$$\nabla_{p_l}\Bigg(\sum_{t=0}^{T}\sum_{i=1}^{N}\sum_{a\in A}\mathbb{I}[a_t^i = a]\Bigg(1 - \frac{p'_{al}}{p_{al}}\Bigg)R(\tau^i)\Bigg) \tag{303}$$

Taking the derivative inside, we get the following expression for derivative:

$$2\Bigg(\sum_{t=0}^{T}\sum_{i=1}^{N}\sum_{a\in A}\mathbb{I}[a_t^i = a]\Bigg(\frac{p'_{al}}{p_{al}^2}\Bigg)R(\tau^i)\Bigg) \tag{304}$$

We recall from the mirror ascent update equartions 299 and 300 from before, we see that $\frac{\partial J(p^{(k)})}{\partial p_i}$ evaluated at the previous point of update:

$$2\Bigg(\sum_{t=0}^{T}\sum_{i=1}^{N}\sum_{a\in A}\mathbb{I}[a_t^i = a]\Bigg(\frac{1}{p'^{(k)}_{al}}\Bigg)R(\tau^i)\Bigg) \tag{305}$$

The update in the above setting works as follows:

$$p_i^{(k+1)} = \frac{p_i^{(k)} \exp\Bigg(2\eta\Bigg(\sum_{t=0}^{T}\sum_{n=1}^{N}\mathbb{I}[a_t^n = a_i]\Bigg(\frac{1}{p^{(k)}_{a_i l}}\Bigg)R(\tau^n)\Bigg)\Bigg)}{\sum_j p_j^{(k)} \exp\Bigg(2\eta\Bigg(\sum_{t=0}^{T}\sum_{n=1}^{N}\mathbb{I}[a_t^n = a_j]\Bigg(\frac{1}{p^{(k)}_{a_j l}}\Bigg)R(\tau^n)\Bigg)\Bigg)\Bigg)} \tag{306}$$

We now recall the empirical version of the interpretable TRPO constraint condition.

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}\Bigg(\frac{p_i^{'(k)}\exp\Bigg(2\eta\Bigg(\sum_{t=0}^{T}\sum_{n=1}^{N}\mathbb{I}[a_t^n=a_i]\Big(\frac{1}{p_{a_i l}^{'(k)}}\Big)R(\tau^n)\Bigg)\Bigg)}{\sum_q p_q^{'(k)}\exp\Bigg(2\eta\Bigg(\sum_{t=0}^{T}\sum_{n=1}^{N}\mathbb{I}[a_t^n=a_q]\Big(\frac{1}{p_{a_q l}^{'(k)}}\Big)R(\tau^n)\Bigg)\Bigg)\Bigg)}-p'_{a_i l}\Bigg) \tag{307}$$

$$\Bigg(\sum_{n=1}^{N}\sum_{t=1}^{T}\Bigg[\frac{p'_{a_t^n l}\mathbb{I}[a_t^n=a_k\forall k\neq|A|]+(1-\sum_{a\in A-a_{|A|}}p'_{al})\mathbb{I}[a_t^n=a_{|A|}]}{\Big(p_{a_t^n l}\mathbb{I}[a_t^n=a_k\forall k\neq|A|]+(1-\sum_{a\in A-a_{|A|}}p_{al})\mathbb{I}[a_t^n=a_{|A|}]\Big)^2} \tag{308}$$

$$\Big(\mathbb{I}[a_t^n=a_j]-\mathbb{I}[a_t^n=a_{|A|}]\Big)\Big(\mathbb{I}[a_t^n=a_i]-\mathbb{I}[a_t^n=a_{|A|}]\Big) \tag{309}$$

$$\Big(1-|A|\mathbb{I}[a_t^n=a_{|A|}]\Big)\Bigg]\Bigg)_{i,j=1}^{|A|-1,|A|-1} \tag{310}$$

$$\Bigg(\frac{p_j^{'(k)}\exp\Bigg(2\eta\Bigg(\sum_{t=0}^{T}\sum_{n=1}^{N}\mathbb{I}[a_t^n=a_j]\Big(\frac{1}{p_{a_j l}^{'(k)}}\Big)R(\tau^n)\Bigg)\Bigg)}{\sum_q p_q^{'(k)}\exp\Bigg(2\eta\Bigg(\sum_{t=0}^{T}\sum_{n=1}^{N}\mathbb{I}[a_t^n=a_q]\Big(\frac{1}{p_{a_q l}^{'(k)}}\Big)R(\tau^n)\Bigg)\Bigg)\Bigg)}-p'_{a_j l}\Bigg)\leq\delta \tag{311}$$

After I make the substitution of 306 into 322 we get that the update looks like the above.

8.4 Interpretable TRPO leaf problem

In this subsection, we list the interpretable subproblem for the leaves in the TRPO algorithm.

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{i=1}^{|A|-1}\sum_{j=1}^{|A|-1}\Bigg(\frac{p_i^{'(k)}\exp\Big(2\eta\sum_{t=0}^{T}\sum_{n=1}^{N}\mathbb{I}[a_t^n=a_i]\Big(\frac{1}{p_{a_i l}^{'(k)}}\Big)R(\tau^n)\Big)}{\sum_q p_q^{'(k)}\exp\Big(2\eta\sum_{t=0}^{T}\sum_{n=1}^{N}\mathbb{I}[a_t^n=a_q]\Big(\frac{1}{p_{a_q l}^{'(k)}}\Big)R(\tau^n)\Big)}-p'_{a_i l}\Bigg) \tag{312}$$

$$\tag{313}$$

$$\times\Bigg[\frac{p'_{a_t^n l}\,\mathbb{I}[a_t^n=a_k\;\forall k\neq|A|]+\Big(1-\sum_{a\in A\setminus a_{|A|}}p'_{al}\Big)\mathbb{I}[a_t^n=a_{|A|}]}{\Big(p_{a_t^n l}\,\mathbb{I}[a_t^n=a_k\;\forall k\neq|A|]+\Big(1-\sum_{a\in A\setminus a_{|A|}}p_{al}\Big)\mathbb{I}[a_t^n=a_{|A|}]\Big)^2} \tag{314}$$

$$\tag{315}$$

$$\cdot\big(\mathbb{I}[a_t^n=a_j]-\mathbb{I}[a_t^n=a_{|A|}]\big)\big(\mathbb{I}[a_t^n=a_i]-\mathbb{I}[a_t^n=a_{|A|}]\big)\big(1-|A|\mathbb{I}[a_t^n=a_{|A|}]\big)\Bigg] \tag{316}$$

$$\tag{317}$$

$$\times\Bigg(\frac{p_j^{'(k)}\exp\Big(2\eta\sum_{t=0}^{T}\sum_{n=1}^{N}\mathbb{I}[a_t^n=a_j]\Big(\frac{1}{p_{a_j l}^{'(k)}}\Big)R(\tau^n)\Big)}{\sum_q p_q^{'(k)}\exp\Big(2\eta\sum_{t=0}^{T}\sum_{n=1}^{N}\mathbb{I}[a_t^n=a_q]\Big(\frac{1}{p_{a_q l}^{'(k)}}\Big)R(\tau^n)\Big)}-p'_{a_j l}\Bigg)\;\leq\;\delta \tag{318}$$

The above is the constraint condition we place for iteration of the p vectors.

The update in the above setting works as follows:

$$p_i^{(k+1)} = \frac{p_i^{(k)} \exp\left(2\eta\left(\sum_{t=0}^{T}\sum_{n=1}^{N}\mathbb{I}[a_t^n = a_i]\left(\frac{1}{p_{a_i l}^{(k)}}\right)R(\tau^n)\right)\right)}{\sum_j p_j^{(k)} \exp\left(2\eta\left(\sum_{t=0}^{T}\sum_{n=1}^{N}\mathbb{I}[a_t^n = a_j]\left(\frac{1}{p_{a_j l}^{(k)}}\right)R(\tau^n)\right)\right)\right)} \tag{319}$$

The objective in this setting is given by:

$$\sum_{t=0}^{T}\sum_{i=1}^{N}\frac{1}{p_{a_{tl}^i}}R(\tau^i)\sum_{j=1}^{|A|}\left(\mathbb{I}[a_t^i = a_j]\left(p_{a_j l} - p'_{a_j l}\right)\right) \tag{320}$$

9 TRPO bounds from the literature

We list different sources of literature that give bounds for the TRPO with the goal of obtaining a bound for the most general version of this algorithm.

For each leaf node l, we calculate the objective which is given as:

$$\sum_{t=0}^{T}\sum_{i=1}^{N}\frac{1}{\mathbb{P}[T_{\theta_l}(s_t^i) = a_t^i]}\left(\mathbb{I}[a_t = a] - \mathbb{I}[a_t = a_{|A|}]\right)_{a\in[A]}.\left(p_{al} - p'_{al}\right)_{a\in[a]}R(\tau^i) \tag{321}$$

The constraint condition for leaf l is given by:

$$-\frac{1}{2}\left(p_{al} - p'_{al}\right)^T\left(\mathbb{E}_{s\sim\theta'_k}\left[\frac{p'_{a_t l}\mathbb{I}[a_t = a_i \forall i \neq |A|] + (1 - \sum_{a'\in A - a_{|A|}} p'_{a'l})\mathbb{I}[a_t = a_{|A|}]}{\left(p_{a_t l}\mathbb{I}[a_t = a_i \forall i \neq |A|] + (1 - \sum_{a'\in A - a_{|A|}} p_{a'l})\mathbb{I}[a_t = a_{|A|}]\right)^2}\right. \tag{322}$$

$$\left(\mathbb{I}[a_t = a_j] - \mathbb{I}[a_t = a_{|A|}]\right)\left(\mathbb{I}[a_t = a_i] - \mathbb{I}[a_t = a_{|A|}]\right) \tag{323}$$

$$\left.\left(1 - |A|\mathbb{I}[a_t = a_{|A|}]\right)\right]\left(p_{al} - p'_{al}\right) \leq \delta \tag{324}$$

The first result we quote is from :
1. Kakade and Langford: Assume we update the policies as per the TRPO updates, we then have that :

$$\eta(\pi_{new}) \geq \mathcal{L}_{\pi_{old}}(\pi_{new}) - \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2 \tag{325}$$

In the above, we have that η denotes the average discounted return which is given by:

$$\mathcal{L}_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_\pi(s)\sum_a \tilde{\pi}(a|s)A_\pi(s,a) \tag{326}$$

Here we have that:

$$A_\pi(s,a) := Q_\pi(s,a) - V_\pi(s) \tag{327}$$

In the above, we have π_{new} to be given as:

$$\pi_{new}(a|s) := (1-\alpha)\pi_{old}(a|s) + \alpha\pi'(a|s) \tag{328}$$

In the above, we have π' to be given as:

$$\pi' = \arg_\pi \max \mathcal{L}_{\pi_{old}}(\pi) \tag{329}$$

We also have the discounted visitation frequencies $\rho_\pi(s)$ to be given by:

$$\rho_\pi(s) := \sum_i \gamma^i \mathbb{P}[s_i = s] \tag{330}$$

Expanding the 326 gives us the following:

$$\eta(\pi_{new}) - \eta(\pi_{old}) \geq \sum_s \rho_{\pi_{old}}(s) \sum_a \pi_{new}(a|s) A_{\pi_{old}}(s,a) - \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2 \tag{331}$$

When we expand the following for each of the terms we get:

$$\eta(\pi_{new}) - \eta(\pi_{old}) \geq \sum_s \sum_i \gamma^i \mathbb{P}[S_i = s] \sum_a \Big[(1-\alpha)\pi_{old}(a|s) + \alpha\pi'(a|s)\Big][Q_{\pi_{old}}(s,a) - V_{\pi_{old}}(s)] \tag{332}$$

$$- \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2 \tag{333}$$

When we use the expansion with $\mathbb{P}[S_i = s] = \pi_0 P^i(s)$ below, we get:

$$\eta(\pi_{new}) - \eta(\pi_{old}) \geq \sum_s \sum_i \gamma^i \pi_0 P^i(s) \sum_a \Big[(1-\alpha)\pi_{old}(a|s) + \alpha\pi'(a|s)\Big][Q_{\pi_{old}}(s,a) - V_{\pi_{old}}(s)] \tag{334}$$

$$- \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2 \tag{335}$$

We expand the summation over the actions part which gives us the following equation:

$$\sum_a \Big[(1-\alpha)\pi_{old}(a|s) + \alpha\pi'(a|s)\Big][Q_{\pi_{old}}(s,a) - V_{\pi_{old}}(s)] \tag{336}$$

We expand the above which gives us:

$$(1-\alpha)\sum_a \pi_{old}(a|s)[Q_{\pi_{old}}(s,a) - V_{\pi_{old}}(s)] + \tag{337}$$

$$\sum_a Q_{\pi_{old}}(s,a)[(1-\alpha)\pi_{old}(a|s) + \alpha\pi'(a|s)] - V_{\pi_{old}}(s) \tag{338}$$

Further expansion of the above gives us:

$$(1-\alpha)\sum_a \pi_{old}(a|s)[Q_{\pi_{old}}(s,a) - V_{\pi_{old}}(s)] + \tag{339}$$

$$\sum_a Q_{\pi_{old}}(s,a)[(1-\alpha)\pi_{old}(a|s) + \alpha\pi'(a|s)] - V_{\pi_{old}}(s) \tag{340}$$

Expansion of the first portion of 339 gives us:

$$(1-\alpha)\Bigg(\sum_a \pi_{old}(a|s)Q_{\pi_{old}}(s,a) - V_{\pi_{old}}(s)\Bigg) \tag{341}$$

$$+(1-\alpha)\sum_a Q_{\pi_{old}}(s,a)\pi_{old}(a|s)+ \tag{342}$$

$$\alpha\sum_a Q_{\pi_{old}}(s,a)\pi'(a|s) - V_{\pi_{old}}(s) \tag{343}$$

The final expansion of 341 gives us:

$$\alpha\sum_a Q_{\pi_{old}}(s,a)\pi'(a|s)+ \tag{344}$$

$$2(1-\alpha)\sum_a Q_{\pi_{old}}(s,a)\pi_{old}(a|s) \tag{345}$$

$$-(2-\alpha)V_{\pi_{old}}(s) \tag{346}$$

We notice that $\sum_a Q_{\pi_{old}}(s,a)\pi_{old}(a|s) = V_{\pi_{old}}(s)$, since we have:

$$\sum_a Q_{\pi_{old}}(s,a)\pi_{old}(a|s) \tag{347}$$

$$= \sum_a \mathbb{E}[r + \gamma P^a V(s)]\pi_{old}(a|s) \tag{348}$$

$$= V_{\pi_{old}}(s) \tag{349}$$

The above makes sense intuitively since $Q_{\pi_{old}}(s,a)$ is defined as the expected return when at state s and we take action a and we sum this over all the possible policy action values for $\pi_{old}(a|s)$. Substituting the same in 344 we get:

$$\alpha\sum_a Q_{\pi_{old}}(s,a)\pi'(a|s)+ \tag{350}$$

$$2(1-\alpha)V_{\pi_{old}}(s) \tag{351}$$

$$-(2-\alpha)V_{\pi_{old}}(s) \tag{352}$$

The next step gives us the following :

$$\alpha\sum_a Q_{\pi_{old}}(s,a)\pi'(a|s) - \alpha\sum_a Q_{\pi_{old}}(s,a)\pi_{old}(a|s) \tag{353}$$

Final simplification of the above gives us the following:

$$\alpha\sum_a Q_{\pi_{old}}(s,a)\Bigg(\pi'(a|s) - \pi_{old}(a|s)\Bigg) \tag{354}$$

Recall that, we have:

$$\eta(\tilde{\pi}) := \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s)\sum_a \tilde{\pi}(a|s)A_\pi(s,a) \tag{355}$$

$$L_\pi(\tilde{\pi}) := \eta(\pi) + \sum_s \rho_{\pi}(s)\sum_a \tilde{\pi}(a|s)A_\pi(s,a) \tag{356}$$

We claim here that $\tilde{\pi} = \pi_{old}$:

$$\pi' := \arg\max_{\tilde{\pi}} \Bigg[\sum_s \rho_\pi(s)\sum_a \tilde{\pi}(a|s)A_\pi(s,a)\Bigg] \tag{357}$$

In the above, we also see that $A_\pi(s,a) := Q_\pi(s,a) - V_\pi$.

In order to maximize 357, we can maximize 357 for each of the states s. This means that we would need to pick action $a \in A$ which maximizes the objective for each:

$$s \to \rho_\pi(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a) \tag{358}$$

This implies that for a state s we pick the action a which maximizes the objective $A_\pi(s,a) := Q_\pi(s,a) - V_\pi(s)$. Here we have:

$$\pi(s) := \arg_{a \in A} \max[Q_\pi(s,a) - V_\pi(s)] \tag{359}$$

So it seems that $\pi' := \arg_{\tilde{\pi}} \max \mathcal{L}_\pi(\tilde{\pi})$.

Can you further simplify the following? Recheck the calculations once again since we get 0 here otherwise. So this approach does not seem to give the needed bounds. i.e if we substitute $0 \leftrightarrow \sum_a \Big[(1-\alpha)\pi_{old}(a|s) + \alpha\pi'(a|s)\Big]$:

$$\eta(\pi_{new}) - \eta(\pi_{old}) \geq \sum_s \sum_i \gamma^i \pi_0 P^i(s) \sum_a \Big[(1-\alpha)\pi_{old}(a|s) + \alpha\pi'(a|s)\Big][Q_{\pi_{old}}(s,a) - V_{\pi_{old}}(s)] \tag{360}$$

$$-\frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2 \tag{361}$$

The substitution of 0 into the sum over the actions a gives us $0 - \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2$.

---

1. Further simplifications for the expressions; Does $\sum_a Q_{\pi_{old}}(s,a)\pi_{old}(s,a)$ remind you of something? Yes

2. Plug in $\pi_\theta(a|s) = \mathbb{P}[T_\theta(s) = a]$

3. Simplification of $V_{\pi_{old}}(s), \pi'(a|s)$. Assume that we have a decision tree policy with parameter θ_{old}

4. Difference between the L operator and the policy difference lemma [Do this]

Other bounding ideas

II. Elseiver TRPO paper with algorithms and implementation direction details.

III. Neural PPO how to convert this to TRPO.

IV. Trust region policy optimization via entropy regularization for Kullback–Leibler divergence constraint

10 Time dependent TRPO

How does the TRPO algorithm behave when the policies π_t are dynamic?

We observe that the following result continues to hold:

$$\eta(\tilde{\pi}_t) = \eta(\pi_t) + \mathbb{E}_{s_0,a_0,\ldots,}\left[\sum_{t=0}^{\infty}\gamma^t A_{\pi_t}(s_t,a_t)\right] \tag{362}$$

We have the following result:

$$\eta(\tilde{\pi}_t) = \tag{363}$$

$$\eta(\pi_t) + \sum_{t=0}^{\infty}\sum_{s}\mathbb{P}[s_t = s|\tilde{\pi}]\sum_{a}\tilde{\pi}_t(a|s)\gamma^t A_{\pi_t}(s,a) \tag{364}$$

However the full expansion which is:

$$\eta(\tilde{\pi}_t) := \eta(\pi_t) + \sum_{s}\rho_{\tilde{\pi}}(s)\sum_{a}\tilde{\pi}(a|s)A_{\pi}(s,a) \tag{365}$$

will fail to be true as from the original paper since in order to remove the $\sum_a$ term, we cannot group together all the terms depending on t. In this case, we have the $\mathcal{L}_{\pi}(\tilde{\pi})$ to be defined as:

$$L_{\pi_t}(\tilde{\pi}_t) := \eta(\pi_t) + \sum_{s}\rho_{\pi_t}(s)\sum_{a}\tilde{\pi_t}(a|s)A_{\pi_t}(s,a) \tag{366}$$

Lemma 1 holds

Does the theorem hold? The coupling proof continues to be true.

Since the return substitution formulae looks like:

$$\eta(\tilde{\pi}_\theta^t) = \eta(\pi_\theta^t) + \sum_{t=0}^{\infty}\sum_{s}\mathbb{P}[S_t = s|\tilde{\pi}_\theta^t]\sum_{a}\tilde{\pi}_\theta^t(a|s)A_{old}^t(s,a) \tag{367}$$

With the addition of t as a subscript to the TRPO objective and constraint we see that the optimization problem in the dynamic finite horizon is given by:

$$\eta(\tilde{\pi}_\theta^t) = \eta(\pi_\theta^t) + \sum_{t=0}^{\infty}\sum_{s}\mathbb{P}[S_t = s|\tilde{\pi}_\theta^t]\sum_{a}\tilde{\pi}_\theta^t(a|s)A_{old}^t(s,a) \tag{368}$$

The optimization problem for the finite time setup is given by:

$$\max_{(\theta_t)_{t\in[T]}}\mathbb{E}_{s\sim\rho_{\theta_{old}}^t}\mathbb{E}_{a\sim q_t}\left[\frac{\pi_\theta^t(a|s_n)}{q_t(a|s_n)}A_{\theta_{old}}^t(s,a)\right] \tag{369}$$

The constraint condition is given by the same with an addition of t to the subscript. It looks like:

$$\mathbb{E}_{s\sim\rho_{\theta_{old}}}\left[D_{KL}(\pi_{\theta_{old}}^t(.|s)||\pi_\theta^t(.|s))\right] \leq \delta \tag{370}$$

11 T-timestep TRPO

In this section, we list the T timestep TRPO decision tree algorithm. The above algorithm uses the same objective in the main and the leaf nodes, however we have t as a subscript to each of the terms in the objective and the constraint and we have to solve T independent problems. How can we find a version of the TRPO that works well until a finite time horizon T and relate the problems to each other per timestep?

12 Trust Region Policy Optimization + Linear Approximation Bounds

Here we aim to quantify the error when we use the linear version of the TRPO algorithm. Recall the normal version of the TRPO which is given as:

$$\mathcal{L}(\theta, \theta_k) := \mathbb{E}_{s,a\sim\pi_{\theta_k}}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}A^{\pi_{\theta_k}}(s,a)\right] \tag{371}$$

$$\bar{D}_{KL}(\theta||\theta_k) = \mathbb{E}_{s\sim\pi_{\theta_k}}[D_{KL}(\pi_\theta(.|s), \pi_{\theta_k}(.|s)] \tag{372}$$

We have the linearized TRPO to be given by:

$$\mathcal{L}(\theta_k, \theta) \approx g^T(\theta - \theta_k) \tag{373}$$

$$\bar{D}_{KL}(\theta||\theta_k) \approx \frac{1}{2}(\theta - \theta_k)^T H(\theta - \theta_k) \tag{374}$$

Recall in the linear version of the TRPO algorithm we have that:

$$g = \nabla_\theta J(\pi_\theta)|_{\theta=\theta_0} \tag{375}$$

$$H = \left(\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\mathbb{E}_{s\sim\pi_{\theta_k}}\left[D_{KL}(\pi_{\theta_{old}}(.|s)||D_{KL}(\pi_\theta(.|s)))\right]_{i,j}\right)|_{\theta=\theta_0} \tag{376}$$

What would be the approximation error when we linearize the objective and approximate the constraints with a quadratic equation? We find that we need to solve the following series of problems:

$$\theta_{k+1} = \arg\max_\theta \mathcal{L}(\theta_k, \theta) \tag{377}$$

$$\bar{D}_{KL}(\theta||\theta_k) \leq \delta \tag{378}$$

$$\theta_{k+1} = \arg\max_\theta g^T(\theta - \theta_k) \tag{379}$$

$$\bar{D}_{KL}(\theta||\theta_k) \leq \delta \tag{380}$$

$$\theta_{k+1} = \arg\max_\theta g^T(\theta - \theta_k) \tag{381}$$

$$\frac{1}{2}(\theta - \theta_k)^T H(\theta - \theta_k) \leq \delta \tag{382}$$

How does the Taylor expansion of $\mathcal{L}(\theta, \theta_k)$ around θ_k look like?

$$\mathcal{L}_{\theta_k}(\theta) = \eta(\pi_{\theta_k}) + \sum_s \rho_{\pi_{\theta_k}}(s)\sum_a \pi_\theta(a|s)A_{\pi_{\theta_k}}(s,a) \tag{383}$$

We notice that we would need to Taylor expand $\pi_\theta(s)$ around θ_k

$$\pi_\theta(a|s) = \pi_{\theta_k}(a|s) + (\theta - \theta_k)\frac{\partial\pi_\theta}{\partial\theta}|_{\theta=\theta_k}(s,a) \tag{384}$$

$$+\frac{1}{2}(\theta - \theta_k)^T\frac{\partial^2\pi_\theta}{\partial\theta^2}|_{\theta=\theta_k}(s,a)(\theta - \theta_k) + ... \tag{385}$$

We have that the following is true for the Taylor expansion of π_θ around θ_k:

$$\pi_\theta(a|s) = \sum_{|\alpha|\leq k}\frac{D^\alpha\pi_\theta(a|s)}{\alpha!}(\theta - \theta_k)^\alpha + \sum_{|\beta|=k+1} R_\beta(x)(\theta - \theta_k)^\beta \tag{386}$$

In the above, we have that:

$$|R_\beta(x)| \leq \frac{1}{\beta!} \max_{|\alpha|=\beta} \max_{y \in \beta} |D^\alpha \pi_\theta(a|s)| \tag{387}$$

The final expression for $\mathcal{L}(\theta, \theta_k)$ is given by:

$$\mathcal{L}_{\theta_k}(\theta) = \mathcal{L}_{\theta_k}(\theta_k) + \sum_{|\alpha| \leq k} \left(\sum_s \sum_a \frac{D^\alpha \pi_\theta(a|s)}{\alpha!} \right) A_{\pi_{\theta_k}}(s,a)(\theta - \theta_k)^\alpha \rho_{\pi_{\theta_k}}(s) \tag{388}$$

$$+ \sum_{|\beta|=k+1} \left(\sum_s \sum_a \rho_{\pi_{\theta_k}}(s)) A_{\pi_{\theta_k}}(s,a)(\theta - \theta_k)^\beta R_\beta(s) \right) \tag{389}$$

For the linear approximation, we have that the error is given by:

$$|\mathcal{L}_{\theta_k}(\theta) - g^T.(\theta - \theta_k)| \leq \sum_{|\beta|=2} \left(\sum_s \sum_a \rho_{\pi_{\theta_k}}(s)) A_{\pi_{\theta_k}}(s,a)(\theta - \theta_k)^2 \left| \bar{D}^2 \pi_\theta(a|s) \right| \right) \tag{390}$$

With the knowledge of the above, how would we relate the final error between the different optimization problems involving the linear and the non-linear approximations?

At the first timestep, we recall that the optimization problems are given by:

$$\theta_{k+1} = \arg\max_\theta \mathcal{L}(\theta_k, \theta) \tag{391}$$

$$\bar{D}_{KL}(\theta||\theta_k) \leq \delta \tag{392}$$

$$\theta_{k+1} = \arg\max_\theta g^T(\theta - \theta_k) \tag{393}$$

$$\bar{D}_{KL}(\theta||\theta_k) \leq \delta \tag{394}$$

Given functions f and g over a domain S such that $|f(\theta) - g(\theta)| \leq Ch(\theta)$; Assume the minimum of f is attained at θ_f and the minimum of g is attained at θ_g. We now have the following result that says $||f(\theta_f) - g(\theta_g)|| \leq L|\theta_f - \theta_g| + Ch(\theta_g)$; here we have that L denotes the Lipschitz constant for the function f.

How would we compute $\arg_\theta \max g^T(\theta - \theta_k)$? I would assume the maxima are attained for when each coordinate of θ takes extrema values.

How would the landscape for θ look like? We believe that this satisfies:

$$\mathbb{E}_{s \sim \pi_{\theta_k}} \left[D_{KL}(\pi_{\theta_k}(.|s)||\pi_\theta(.|s)]) \right] \tag{395}$$

$$\bar{D}_{KL}(\theta||\theta_k) \leq \delta \tag{396}$$

$$\tag{397}$$

To observe the landscape above for the distribution of θ, we expand out the above inequality which gives us:

$$\mathbb{E}_{s \sim \pi_{\theta_k}} \left[\int_A \pi_\theta(a|s) \log \frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} da \right] \leq \delta \tag{398}$$

$$\tag{399}$$

Fitting to the case where we have $\pi_\theta(.|s)$ to be a decision tree function, we see that the constraint condition is then given by:

$$\frac{1}{|S|}\sum_{s\in S}\sum_{a\in A}\mathbb{P}[T_\theta(s)=a]\log\left(\frac{\mathbb{P}[T_\theta(s)=a]}{\mathbb{P}[T_{\theta_k}(s)=a]}\right)\leq\delta \tag{400}$$

The above would then restrict $\eta_{\theta(s)} \ll \eta_{\theta_k(s)}$. Further if we make the following notation, we see that:

$$\eta(\theta(s),a)\equiv\mathbb{P}[T_\theta(s)=a] \tag{401}$$

$$\eta(\theta_k(s),a)\equiv\mathbb{P}[T_{\theta_k}(s)=a] \tag{402}$$

Finally we have:

$$\sum_{s\in S}\sum_{a\in A}\eta(\theta(s),a)\log\left(\frac{\eta(\theta(s),a)}{\eta(\theta_k(s),a)}\right)\leq\delta|S| \tag{403}$$

Assume we restrict to changes in the node activation coefficients for a single node; for example at node η we have:

$$\mathbb{P}[T_\theta(s)=a]=\mu_\eta(s)\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Big)+\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a] \tag{404}$$

If θ_k differs from θ only at node η^*, we see that for the children η of η^* we have that:

$$\eta(\theta_k(s),a)=\eta(\theta(s),a) \tag{405}$$

This implies that the only potential change in the estimation of the probabilities occurs at the ancestor nodes over the modified node.

12.1 Head Node Analysis

Assume in the initial case that we make a modification to the activation function only at the top node; in that case we have the constraint being given by:

$$\sum_{a\in A}\eta(\theta(h),a)\log\left(\frac{\eta(\theta(h),a)}{\eta(\theta_k(h),a)}\right) \tag{406}$$

The above occurs since $\eta(\theta(s),a)=\eta(\theta_k(s),a)\forall s\neq h$.
The log term in the above can be expanded as:

$$\log\left(\frac{\mu_\theta(s)\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Big)+\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]}{\mu_{\theta_k}(s)\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Big)+\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]}\right) \tag{407}$$

The above can be simplified by taking $\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Big)$ in common in the numerator and denominator as:

$$\log\left(\frac{\mu_\theta(s)+\frac{\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]}{\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Big)}}{\mu_{\theta_k}(s)+\frac{\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]}{\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\Big)}}\right) \tag{408}$$

The log terms are now simplified as follows:

$$\log\left(\mu_{\theta_k}(s) + C_a + (\mu_\theta(s) - \mu_{\theta_k}(s))\right) \tag{409}$$

$$-\log\left(\mu_{\theta_k}(s) + C_a\right) \tag{410}$$

If we take $C_a := \frac{\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]}{\left(\mathbb{P}[T_{\mathcal{L}(\eta)}(s)=a]-\mathbb{P}[T_{\mathcal{R}(\eta)}(s)=a]\right)}$ Further simplification of the above gives us:

$$\log\left(1 + \frac{\mu_\theta(s) - \mu_{\theta_k}(s)}{\mu_{\theta_k}(s) + C_a}\right) = \frac{\mu_\theta(s) - \mu_{\theta_k}(s)}{\mu_{\theta_k}(s) + C_a} - \tag{411}$$

$$\frac{1}{2}\left(\frac{\mu_\theta(s) - \mu_{\theta_k}(s)}{\mu_{\theta_k}(s) + C_a}\right)^2 + \frac{1}{3}\left(\frac{\mu_\theta(s) - \mu_{\theta_k}(s)}{\mu_{\theta_k}(s) + C_a}\right)^3 \tag{412}$$

$$\tag{413}$$

We assume in the above that $\mu_{\theta_k}(s) + C_a >> 0$ and proceed with the calculation of the bounds;

12.2 Big denominator bounds

We recall the expansion from 411, here we see that the following holds true:

$$\log\left(1 + \frac{\mu_\theta(s) - \mu_{\theta_k}(s)}{\mu_{\theta_k}(s) + C_a}\right) = \frac{\mu_\theta(s) - \mu_{\theta_k}(s)}{\mu_{\theta_k}(s) + C_a} - \tag{414}$$

$$\frac{1}{2}\left(O(\mu_\theta(s) - \mu_{\theta_k}(s))^2\right) \tag{415}$$

How does the approximation look like?

$$\mu_\theta(x) - \mu_{\theta_k}(x) := \left(1 + \exp(-a_\eta(\beta_\eta^T x - \phi_\eta))\right)^{-1} - \tag{416}$$

$$\left(1 + \exp(-a_\eta(\beta_{k,\eta}^T x - \phi_{k,\eta}))\right)^{-1} \tag{417}$$

We have that the sigmoid function is given by:

$$\sigma(z) := \frac{1}{(1 + e^{-z})} \tag{418}$$

The above is a difference of the sigmoid functions which under Taylor expansion looks like:

$$\sigma(z) - \sigma(z_k) \approx \sigma'(z_k)(z - z_k) + \frac{1}{2}O(\sigma''(z)(z - z_k)^2) \tag{419}$$

In the above, we have the differences $(z - z_k)$ which looks like :

$$(z - z_k) := (\beta_\eta^T x - \phi_\eta)\Delta a + a_{k,\eta}(x^T \Delta\beta - \Delta\phi)) \tag{420}$$

We also have the following to be true for $\sigma'(z)$, where we have $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ and we have that $\sigma''(z) = \sigma(z)(1 - \sigma(z))(1 - 2\sigma(z))$ so that we have $|\sigma''(z)| \leq 1 \; \forall z \in \mathbb{R}$.

The total bounds are given by:

$$\mu_\theta(x) - \mu_{\theta_k}(x) \approx \frac{\mu_{\theta_k}(x)\big(1-\mu_{\theta_k}(x)\big)}{\mu_{\theta_k}(x)+C_a}\Big[(\beta_{k,\eta}^\top x - \phi_{k,\eta})(a_\eta - a_{k,\eta}) + a_{k,\eta}\big(x^\top(\beta_\eta - \beta_{k,\eta}) - (\phi_\eta - \phi_{k,\eta})\big)\Big] + O\big((z-z_k)^2\big). \tag{421}$$

With this definition in place, we have that the expansion of the log reads as follows:

$$\log\left(1 + \frac{\mu_\theta(s) - \mu_{\theta_k}(s)}{\mu_{\theta_k}(s)+C_a}\right) = \frac{\mu_\theta(s) - \mu_{\theta_k}(s)}{\mu_{\theta_k}(s)+C_a} - \tag{422}$$

$$\frac{1}{2}\left(O(\mu_\theta(s) - \mu_{\theta_k}(s))^2\right) \tag{423}$$

In the calculations below we aim to remove all the terms involving the products of differences:

$$\frac{\mu_{\theta_k}(x)\big(1-\mu_{\theta_k}(x)\big)}{\mu_{\theta_k}(x)+C_a}\Big[(\beta_{k,\eta}^\top x - \phi_{k,\eta})(a_\eta - a_{k,\eta}) + a_{k,\eta}\big(x^\top(\beta_\eta - \beta_{k,\eta}) - (\phi_\eta - \phi_{k,\eta})\big)\Big] + O((z-z_k)^2) \tag{424}$$

The first order approximation of the KL divergence between the two policies is given by:

$$\sum_{s\in S}\sum_{a\in A}\eta(\theta(s),a)\log\left(\frac{\eta(\theta(s),a)}{\eta(\theta_k(s),a)}\right) \tag{425}$$

The log-differences are estimated by:

$$\sum_{x\in S}\sum_{a\in A}\eta(\theta(x),a)\frac{\mu_{\theta_k}(x)\big(1-\mu_{\theta_k}f(x)\big)}{\mu_{\theta_k}(x)+C_a}\left[(\beta_{k,\eta}^\top x - \phi_{k,\eta})(a_\eta - a_{k,\eta}) + a_{k,\eta}\big(x^\top(\beta_\eta - \beta_{k,\eta}) - (\phi_\eta - \phi_{k,\eta})\big)\right] + \tag{426}$$

$$O((z-z_k)^2) \tag{427}$$

Expecting that C_a to be small for all actions $a \in A$ might be too big of an assumption. We now proceed to simplify 426; this looks as follows:

$$\sum_{x\in S}\mu_{\theta_k}(x)\left(1-\mu_{\theta_k}(x)\right)\left[(\beta_{k,\eta}^\top x - \phi_{k,\eta})(a_\eta - a_{k,\eta}) + a_{k,\eta}\big(x^\top(\beta_\eta - \beta_{k,\eta}) - (\phi_\eta - \phi_{k,\eta})\big)\right]\sum_{a\in A}\frac{\eta(\theta(x),a)}{C_a+\mu_{\theta_k}(x)} \tag{428}$$

Assume we are permitted changes in only the coefficient a_η, we then see that the above simplifies to a linear range in the a values; the exact range of values is given as follows:

$$\sum_{x\in S}\mu_{\theta_k}(x)\left(1-\mu_{\theta_k}(x)\right)\left[(\beta_{k,\eta}^\top x - \phi_{k,\eta})(a_\eta - a_{k,\eta})\right]\sum_{a\in A}\frac{\eta(\theta(x),a)}{C_a+\mu_{\theta_k}(x)} \leq \delta \tag{429}$$

The difference in $(a_\eta - a_{k,\eta})$ values is now written as:

$$(a_\eta - a_{k,\eta}) \leq \frac{\delta}{\sum_{a\in A}\sum_{x\in S}\frac{\eta(\theta(x),a)\mu_{\theta_k}(x)(1-\mu_{\theta_k}(x))}{C_a+\mu_{\theta_k}(x)}(\beta_{k,\eta}^\top x - \phi_{k,\eta})} \tag{430}$$

So the final optimization problems under a first-order restriction of the KL-values is given as:

$$a_\eta = \arg\max_{a_\eta}\mathcal{L}(a_{k,\eta}, a_\eta) \tag{431}$$

$$(a_\eta - a_{k,\eta}) \leq \frac{\delta}{\sum_{a\in A}\sum_{x\in S}\frac{\eta(\theta(x),a)\mu_{\theta_k}(x)(1-\mu_{\theta_k}(x))}{C_a+\mu_{\theta_k}(x)}(\beta_{k,\eta}^\top x - \phi_{k,\eta})} \tag{432}$$

For the above, we recall the non-linear objective being given as:

$$\arg_\theta \max \Bigg(\mathbb{E}_{\tau \sim \pi_\theta, P_t} \Bigg[\sum_{t=0}^{T} \frac{\nabla_\theta [\mu_\eta(s_t, \theta)].(\theta - \theta_{old}) \Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big) R(\tau)}{\Big[\mu_\eta(s_t, \theta) \Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big) + \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big]} \Bigg] \Bigg) \tag{433}$$

In the above we see that we wish to evaluate $\nabla_\theta[\mu_\eta(x, \theta)]$ at the node η and for the a value; this is given as:

$$\frac{\partial \mu_\eta(x)}{\partial a_\eta} = \frac{(\beta_\eta^\top x - \phi_\eta) \exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)}{1 + \exp\big(-a_\eta(\beta_\eta^\top x - \phi_\eta)\big)}. \tag{434}$$

To the non-linear objective we add in the expressions as per above, which gives us the following:

$$\arg_{a_\eta} \max \Bigg(\mathbb{E}_{\tau \sim \pi_\theta, P_t} \Bigg[\sum_{t=0}^{T} \frac{(a_\eta - a_{\eta_k}) \frac{\partial \mu_\eta(s_t, \theta)}{\partial a_\eta} \Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big) R(\tau)}{\Big[\mu_\eta(s_t, \theta) \Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big) + \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big]} \Bigg] \Bigg) \tag{435}$$

In the above, we aim to eliminate the $a_{\eta,k}$ terms, this can be done in the denominator as follows:

$$\mu_\eta(s_t, \theta) = \mu_\eta(x) \mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] + \tag{436}$$

$$(1 - \mu_\eta(x)) \mathbb{P}[T_{\mathcal{R}(\eta)} = a_t] \tag{437}$$

We expand the above as follows:

$$\arg_\theta \max \Bigg(\mathbb{E}_{\tau \sim \pi_\theta, P_t} \Bigg[\sum_{t=0}^{T} \frac{(a_\eta - a_{\eta_k}) \frac{\partial \mu_\eta(s_t, \theta)}{\partial a_\eta} \Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big) R(\tau)}{\Big[\mu_\eta(s_t, \theta) \Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big) + \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big]} \Bigg] \Bigg) \tag{438}$$

We simplify the above by noting that :

$$\arg_\theta \max \Bigg(\mathbb{E}_{\tau \sim \pi_\theta, P_t} \Bigg[\sum_{t=0}^{T} \frac{(a_\eta - a_{\eta_k}) \frac{(\beta_\eta^\top s_t - \phi_\eta) \exp\left(-a_\eta(\beta_\eta^\top s_t - \phi_\eta)\right)}{1 + \exp\left(-a_\eta(\beta_\eta^\top s_t - \phi_\eta)\right)} \Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big) R(\tau)}{\Big[\Big(1 + \exp(-a_\eta(\beta_\eta^T s_t - \phi_\eta)) \Big)^{-1} \Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big) + \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t} \tag{439}$$

The above is now fully written as:

$$\arg_\theta \max \Bigg(\mathbb{E}_{\tau \sim \pi_\theta, P_t} \Bigg[\sum_{t=0}^{T} \frac{(a_\eta - a_{\eta_k})(\beta_\eta^\top s_t - \phi_\eta) \exp\big(-a_\eta(\beta_\eta^\top s_t - \phi_\eta)\big) \Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big) R(\tau)}{\Big[\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big) + \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big(1 + \exp(-a_\eta(\beta_\eta^T s_t - \phi_\eta)) \Big) \Big]} \tag{440}$$

The above is further simplified as follows:

$$\arg_\theta \max \left(\mathbb{E}_{\tau \sim \pi_\theta, P_t} \left[\sum_{t=0}^{T} \frac{(a_\eta - a_{\eta_k})(\beta_\eta^\top s_t - \phi_\eta) \exp\big(-a_\eta(\beta_\eta^\top s_t - \phi_\eta)\big) \Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big) R(\tau)}{\left[\Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] + \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \exp(-a_\eta(\beta_\eta^T s_t - \phi_\eta) \Big) \right]} \right] \right) \tag{441}$$

In order to compute the argmax for the above, we plan to compute the derivative of the above term; recall we have

$$N(a_\eta) := (a_\eta - a_{\eta_k})(\beta_\eta^\top s_t - \phi_\eta) \exp\big(-a_\eta(\beta_\eta^\top s_t - \phi_\eta)\big) \Big(\mathbb{P}[T_{\mathcal{L}(\eta)}(s_t) = a_t] - \mathbb{P}[T_{\mathcal{R}(\eta)}(s_t) = a_t] \Big) R(\tau) \tag{442}$$

We make the following definitions for the different values of the constants:

$$c_{\beta,t} = (\beta_\eta^\top s_t - \phi_\eta) \tag{443}$$

$$c_{L,t} = \mathbb{P}[T_{\mathcal{L}(\eta)} = a_t] \tag{444}$$

$$c_{R,t} = \mathbb{P}[T_{\mathcal{R}(\eta)} = a_t] \tag{445}$$

Rewriting $F(a_\eta)$ with the new notation we get the following to be true:

$$F(a_\eta) := \frac{(a_\eta - a_{\eta_k}) c_{\beta,t} \exp(-a_\eta c_{\beta,t})(c_{L,t} - c_{R,t})}{(c_{L,t} + c_{R,t} \exp(-a_\eta c_{\beta,t}))} \tag{446}$$

We have the numerator $N(a_\eta)$ to be given as:

$$N(a_\eta) := (a_\eta - a_{\eta_k}) c_{\beta,t} \exp(-a_\eta c_{\beta,t})(c_{L,t} - c_{R,t}) \tag{447}$$

We have the denominator $D(a_\eta)$ to be given by:

$$D(a_\eta) = \Big[c_{L,t} + c_{R,t} \exp(-a_\eta c_{\beta,t}) \Big] \tag{448}$$

We have that the differential of N with respect to a_η is given by:

$$N'(a_\eta) := c_{\beta,t} \exp(-a_\eta c_{\beta,t})(c_{L,t} - c_{R,t}) - c_{\beta,t}(a_\eta - a_{\eta_k}) \exp(-a_\eta c_{\beta,t})(c_{L,t} - c_{R,t}) \tag{449}$$

We see that the differential of D with respect to a_η is given by:

$$D'(a_\eta) = -c_{R,t} c_{\beta,t} \exp(-a_\eta c_{\beta,t}) \tag{450}$$

The differential of $F(a_\eta)$ is given as follows:

$$\frac{N'(a_\eta) D(a_\eta) - N(a_\eta) D'(a_\eta)}{(D(a_\eta))^2} \tag{451}$$

In total we write the numerator above is given as follows:

$$\Big[c_{\beta,t} \exp(-a_\eta c_{\beta,t})(c_{L,t} - c_{R,t}) - c_{\beta,t}(a_\eta - a_{\eta_k}) \exp(-a_\eta c_{\beta,t})(c_{L,t} - c_{R,t}) \Big] \Big[c_{L,t} + c_{R,t} \exp(-a_\eta c_{\beta,t}) \Big] \tag{452}$$

$$+ (a_\eta - a_{\eta_k}) c_{\beta,t} \exp(-a_\eta c_{\beta,t})(c_{L,t} - c_{R,t}) c_{R,t} c_{\beta,t} \exp(-a_\eta c_{\beta,t}) \tag{453}$$

Simplification of the numerator gives us the following:

$$(c_{L,t} - c_{R,t}) \exp(-a_\eta c_{\beta,t}) c_{\beta,t} \Big(\Big[1 - a_\eta + a_{\eta_k} \Big] \Big[c_{L,t} + c_{R,t} \exp(-a_\eta c_{\beta,t}) \Big] \tag{454}$$

$$+ c_{R,t} c_{\beta,t} (a_\eta - a_{\eta_k}) \exp(-a_\eta c_{\beta,t}) \Big) \tag{455}$$

The denominator when squared is then given as follows:

$$\left[c_{L,t} + c_{R,t} \exp\left(-a_\eta c_{\beta,t}\right)\right]^2 \tag{456}$$

Taking things in common for the above, we get that the expansion looks like:

$$c_{\beta,t}(c_{L,t} - c_{R,t}) \exp(-a_\eta c_{\beta,t}) \Bigg[(1 - c_{\beta,t})(c_{L,t} + c_{R,t} \exp(-a_\eta c_{\beta,t})) \tag{457}$$

$$-(a_\eta - a_{\eta_k})(c_{R,t}(-c_{\beta,t}) \exp(-a_\eta c_{\beta,t}))\Bigg] \tag{458}$$

Expanding out the expressions above we get:

$$c_{\beta,t}(c_{L,t} - c_{R,t}) \exp(-a_\eta c_{\beta,t}) \Bigg[(1 - c_{\beta,t})c_{L,t} + c_{R,t} \exp(-a_\eta c_{\beta,t})(1 - c_{\beta,t} + ((a_\eta - a_{\eta_k})c_{\beta,t}))\Bigg] \tag{459}$$

Equating the above to zero, we get:

$$\Bigg[(1 - c_{\beta,t})(c_{L,t} + c_{R,t} \exp(-a_\eta c_{\beta,t})) + c_{R,t} \exp(-a_\eta c_{\beta,t})(a_\eta - a_{\eta_k})c_{\beta,t} = 0\Bigg] \tag{460}$$

The full expansion of the above yields the following:

$$\Bigg[((1 - c_{\beta,t})c_{L,t} + c_{R,t} \exp(-a_\eta c_{\beta,t})[(1 - c_{\beta,t}) - a_{\eta_k} c_{R,t}] \tag{461}$$

$$+a_\eta \exp(-a_\eta c_{\beta,t})(c_{\beta,t} c_{R,t})\Bigg] = 0 \tag{462}$$

Expanding the above we get:

$$\Bigg(c_1 + c_2 \exp(-a_\eta c_{\beta,t}) + c_3 a_\eta \exp(-a_\eta c_{\beta,t})\Bigg) = 0 \tag{463}$$

Step 1: Factor the exponential term. We begin with

$$c_1 + c_2 e^{-a_\eta c_{\beta,t}} + c_3 a_\eta e^{-a_\eta c_{\beta,t}} = 0.$$

Factoring out the exponential term gives

$$c_1 + e^{-a_\eta c_{\beta,t}}\big(c_2 + c_3 a_\eta\big) = 0.$$

Rearranging,

$$e^{-a_\eta c_{\beta,t}}\big(c_2 + c_3 a_\eta\big) = -c_1.$$

Step 2: Normalize and change variables. Dividing both sides by c_3,

$$\left(a_\eta + \frac{c_2}{c_3}\right) e^{-a_\eta c_{\beta,t}} = -\frac{c_1}{c_3}.$$

Define the change of variables

$$u := a_\eta + \frac{c_2}{c_3}.$$

Then

$$u\, e^{-c_{\beta,t} u} = -\frac{c_1}{c_3} \exp\left(-\frac{c_2 c_{\beta,t}}{c_3}\right).$$

Step 3: Apply the Lambert W function. Multiplying both sides by $-c_{\beta,t}$,

$$(-c_{\beta,t}u)\, e^{-c_{\beta,t}u} = c_{\beta,t}\frac{c_1}{c_3}\exp\left(-\frac{c_2 c_{\beta,t}}{c_3}\right).$$

By definition of the Lambert W function,

$$-c_{\beta,t}u = W\left(c_{\beta,t}\frac{c_1}{c_3}\exp\left(-\frac{c_2 c_{\beta,t}}{c_3}\right)\right).$$

Step 4: Solve for a_η. Substituting back for u, we obtain

$$a_\eta = -\frac{1}{c_{\beta,t}}W\left(c_{\beta,t}\frac{c_1}{c_3}\exp\left(-\frac{c_2 c_{\beta,t}}{c_3}\right)\right) - \frac{c_2}{c_3}.$$

In the above, we must have that a_η must be in 464.

The differences in the a values is given as:

$$(a_\eta - a_{k,\eta}) \leq \frac{\delta}{\sum_{a\in A}\sum_{x\in S}\frac{\eta(\theta(x),a)\mu_{\theta_k}(x)(1-\mu_{\theta_k}(x))}{C_a+\mu_{\theta_k}(x)}(\beta_{k,\eta}^\top x - \phi_{k,\eta})} \tag{464}$$

We have the following upper bound for the Lambert-W function; this is given as:

$$\log(x) - \log\log(x) \leq W(x) \leq \log(x) - \frac{1}{2}\log\log(x) \tag{465}$$

We expand the upper bound above as follows:

$$\log\left(c_{\beta,t}\frac{c_1}{c_3}\exp\left(-\frac{c_2 c_{\beta,t}}{c_3}\right)\right) - \log\log\left(c_{\beta,t}\frac{c_1}{c_3}\exp\left(-\frac{c_2 c_{\beta,t}}{c_3}\right)\right) \leq W(x) \leq \tag{466}$$

$$\log\left(c_{\beta,t}\frac{c_1}{c_3}\exp\left(-\frac{c_2 c_{\beta,t}}{c_3}\right)\right) - \frac{1}{2}\log\log\left(c_{\beta,t}\frac{c_1}{c_3}\exp\left(-\frac{c_2 c_{\beta,t}}{c_3}\right)\right) \tag{467}$$

In the above, we have that:

$$c_1 = \left(1 - c_{\beta,t}\right)c_{L,t} \tag{468}$$

$$c_2 = c_{R,t} \tag{469}$$

$$c_3 = c_{\beta,t}c_{R,t} \tag{470}$$

Does the computed a_η lie in the confidence interval of our choosing?

For each of the different coefficient values; $a_\eta, \phi_\eta, \beta_\eta$, if we see that under a linear approximation the KL divergence condition results in a linear constraint.

Recall that at the first timestep the constraint conditions would match and the objectives would differ. In particular we would be starting from different θ_k values in the linear constraint and different θ_k values in the non-linear objective; however for future timesteps we see that since we search in a region close to the previous optimal values, the search spaces would differ.

References