The idea is to compose a policy with following form:

$$\pi_\theta(a|s) = \mathcal{N}(a \mid \text{Mean}_\theta(s), \text{DiagVar}_\theta) \quad (0)$$

$$\text{where } \text{Mean}_\theta(s) = \alpha_{TV} \cdot a^{ref}(s) + (1-\alpha_{TV}) \cdot a_\theta^{explore}(s) \quad (1)$$

- $a^{ref}(s)$ comes directly from a planning method on dynamics model, while $a_\theta^{explore}(s)$ is a neural network policy to be aggregated to the reference policy $a^{ref}(s)$ with a weight of $1-\alpha_{TV}$.
- α_{TV} denotes a measure of Trust Value of reference policy.
- Trust Value should be dependent on the discrepency between actual reward in real environment $r(s,a)$ and simulated reward from dynamics model $\hat{r}(s,a)$. So, α_{TV} should depend on $r^{diff} = |r(s,a) - \hat{r}(s,a)|$, which can be trained online in a supervised way.

- Supervised learning for r_ϕ^{diff} with parameter ϕ
- Then, α_{TV}, for example, could be formulated as:

$$\alpha_{TV}(s,a) = \exp\left(-r_\phi^{diff}(s,a)\right) \tag{2}$$

- r_{Label}^{diff} could be defined as normalized:

$$r_{Label}^{diff}(s,a) = \frac{|r(s,a) - \hat{r}(s,a)|}{|\hat{r}(s,a)|} \tag{3}$$

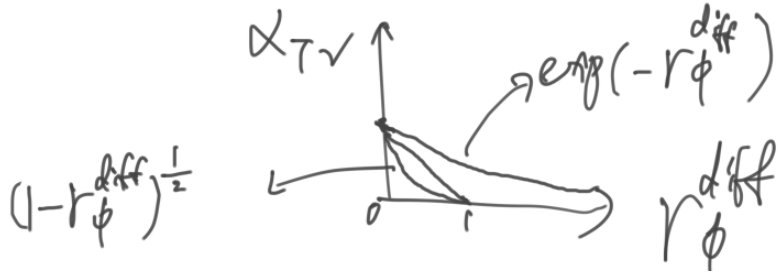
- α_{TV} could also be other functions of $r_\phi^{diff}(s,a)$.
- In summary, total loss could be:

$$\min_{\theta,\phi} L(\theta,\phi) = -L(\theta) + \beta \cdot L(\phi), \text{ where } \beta \text{ is a hyperparameter.}$$

$$\begin{cases} L(\theta) = \mathbb{E}_{s\sim\rho_{\theta_{old}},\, a\sim\pi_{\theta_{old}}}\left[\dfrac{\pi_{\theta_{new}}(a|s)}{\pi_{\theta_{old}}(a|s)}\, A_{\theta_{old}}(s,a)\right] & (4) \\ L(\phi) = \| r_\phi^{diff}(s,a) - r_{Label}^{diff}(s,a) \|_2 & (5) \end{cases}$$

Discussions of $r_\phi^{diff}(s,a)$

- ϕ Network learns $|r(s,a) - \hat{r}(s,a)|$?
  or by normalizing $\frac{|r(s,a) - \hat{r}(s,a)|}{|\hat{r}(s,a)|}$?
  if normalize, what is denominator? $|\hat{r}(s,a)|$ or $|r(s,a)|$ or?
- if $r(s,a) > \hat{r}(s,a)$, should we assume $\alpha_{TV}(r_\phi^{diff}) = 1$?
- Should we consider $r_\phi^{diff} \in [0, +\infty)$, or $\in [0, 1]$, or $\in (-\infty, +\infty)$
- Exact function α_{TV} of r_ϕ^{diff} ? $\alpha_{TV} = \exp(-r_\phi^{diff})$?
  or other functions such as
  $\alpha_{TV} = 1 - r_\phi^{diff}$?
  $\alpha_{TV} = (1 - r_\phi^{diff})^p$



Discussions of $\alpha_{TV} \cdot a^{ref}$

- more generally, linear transformation of a^{ref}? i.e. $\alpha_{TV} \cdot \underline{\underline{T}} \cdot a^{ref}$ where T is a transformation matrix.
- Most generally, non linear. $\alpha_{TV} \cdot f_w(a^{ref})$, where f_w is a NN with w as parameter.