

Aligning Consistency Models by Preference Optimization


Deep Generative Models

Amirabbas afzali
Zahra Maleki

contact 
contact 

Borna Khodabandeh
Asemaneh Nafe

contact 
contact 

Ashkan majidi contact 



Consistency models as an RL policy

There are two approaches we can take regarding modelling consistency models as an RL policy. first, we can adapt the multi step approach while formulating our Markov Decision Process (MDP).

$$\begin{aligned} \mathbf{s}_t &\triangleq (\mathbf{x}_{\tau_t}, \tau_t, \mathbf{c}) & \pi(\mathbf{a}_t | \mathbf{s}_t) &\triangleq f_{\theta}(\mathbf{x}_{\tau_t}, \tau_t, \mathbf{c}) + Z & P(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t) &\triangleq (\delta_{\mathbf{x}_{\tau_{t+1}}}, \delta_{\tau_{t+1}}, \delta_{\mathbf{c}}) \\ \mathbf{a}_t &\triangleq \mathbf{x}_{\tau_{t+1}} & \mu &\triangleq (\mathcal{N}(0, I), \delta_{\tau_0}, p(\mathbf{c})) & R_H(\mathbf{s}_H) &= r(f_{\theta}(\mathbf{x}_{\tau_H}, \tau_H, \mathbf{c}), \mathbf{c}) \end{aligned}$$

where is $Z = \sqrt{\tau_t^2 - \tau_H^2} \mathbf{z}$ is noise. Further, where $r(\cdot, \cdot)$ is the reward function that we are using to align the model and R_H is the reward at timestep H . At other timesteps, we let the reward be 0.

Diffusion Model Alignment Using DPO

We have a discrete-time reverse process with a Markov structure, and is trained using the ELBO minimization associated with the model.

$$p_{\theta}(x_{1:T}) = \prod_{t=1}^T p_{\theta}(x_{t-1} | x_t) \quad p_{\theta}(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_{\theta}(x_t), \sigma_{t|t-1}^2 \cdot \frac{\sigma_{t-1}^2}{\sigma_t^2} \mathbf{I}) \quad (1)$$

$$\mathcal{L} = \mathbb{E}_{x_0, \epsilon, t, x_t} [\omega(\lambda_t) \|\epsilon - \epsilon_{\theta}(x_t; t)\|_2^2] \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \quad t \sim \mathcal{U}(0, T) \quad (2)$$

$$x_t \sim q(x_t | x_0) = \mathcal{N}(x_t; \alpha_t x_0, \sigma_t^2 \mathbf{I}) \quad \lambda_t = \frac{\alpha_t^2}{\sigma_t^2} \quad (3)$$

Using RLHF to guide these models is pretty straightforward.

$$\mathcal{L}_{\text{BT}}(\phi) = -\mathbb{E}_{c, x_0^w, x_0^l \sim \mathcal{D}} [\log \sigma(r_{\phi}(c, x_0^w) - r_{\phi}(c, x_0^l))] \quad (4)$$

$$\max_{p_{\theta}} \mathbb{E}_{c \sim \mathcal{D}_c, x_0 \sim p_{\theta}(x_0 | c)} [r(c, x_0)] - \beta \mathbb{D}_{\text{KL}}[p_{\theta}(x_0 | c) \| p_{\text{ref}}(x_0 | c)] \quad (5)$$

Prior work like DDPO, utilize the multistep nature of Diffusion models more effectively. and solve the problem from above. we can bypass the reward model in this framework as well.

$$r(c, x_0) = \mathbb{E}_{p_{\theta}(x_{1:T} | x_0, c)} [R(c, x_{0:T})] \quad (6)$$

$$\max_{p_{\theta}} \mathbb{E}_{c \sim \mathcal{D}_c, x_{0:T} \sim p_{\theta}(x_{0:T} | c)} [R(c, x_{0:T})] - \beta \mathbb{D}_{\text{KL}}[p_{\theta}(x_{0:T} | c) \| p_{\text{ref}}(x_{0:T} | c)] \quad (7)$$

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{c, x_0^w, x_0^l \sim \mathcal{D}} \left[\log \sigma \left(\beta \mathbb{E}_{\substack{x_{1:T}^w \sim p_{\theta}(x_{1:T}^w | x_0^w, c) \\ x_{1:T}^l \sim p_{\theta}(x_{1:T}^l | x_0^l, c)}} \left[\log \frac{p_{\theta}(x_{0:T}^w)}{p_{\text{ref}}(x_{0:T}^w)} - \log \frac{p_{\theta}(x_{0:T}^l)}{p_{\text{ref}}(x_{0:T}^l)} \right] \right) \right] \quad (8)$$

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{c, x_0^w, x_0^l \sim \mathcal{D}} \left[\log \sigma \left(\beta \mathbb{E}_{\substack{x_{1:T}^w \sim p_{\theta}(x_{1:T}^w | x_0^w, c) \\ x_{1:T}^l \sim p_{\theta}(x_{1:T}^l | x_0^l, c)}} \left[\sum_{t=1}^T \log \frac{p_{\theta}(x_t^w | x_{t-1}^w)}{p_{\text{ref}}(x_t^w | x_{t-1}^w)} - \log \frac{p_{\theta}(x_t^l | x_{t-1}^l)}{p_{\text{ref}}(x_t^l | x_{t-1}^l)} \right] \right) \right] \quad (9)$$

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{c, x_0^w, x_0^l \sim \mathcal{D}} \left[\log \sigma \left(\beta T \mathbb{E}_{t \sim \mathcal{U}(1, T), x_{1:T}^w \sim p_{\theta}(x_{1:T}^w | x_0^w, c)} \left[\log \frac{p_{\theta}(x_t^w | x_{t-1}^w)}{p_{\text{ref}}(x_t^w | x_{t-1}^w)} - \log \frac{p_{\theta}(x_t^l | x_{t-1}^l)}{p_{\text{ref}}(x_t^l | x_{t-1}^l)} \right] \right) \right] \quad (10)$$

$$\mathcal{L}_{\text{DPO}} \leq -\mathbb{E}_{c, x_0^w, x_0^l \sim \mathcal{D}} \mathbb{E}_{t \sim \mathcal{U}(1, T), x_{1:T}^w \sim p_{\theta}(x_{1:T}^w | x_0^w, c)} \left[\log \sigma \left(\beta T \left[\log \frac{p_{\theta}(x_t^w | x_{t-1}^w, c)}{p_{\text{ref}}(x_t^w | x_{t-1}^w, c)} - \log \frac{p_{\theta}(x_t^l | x_{t-1}^l, c)}{p_{\text{ref}}(x_t^l | x_{t-1}^l, c)} \right] \right) \right] \quad (11)$$

$$\mathcal{L}_{\text{DPO}} \leq -\mathbb{E}_{c, x_0^w, x_0^l \sim \mathcal{D}} \mathbb{E}_{\substack{t \sim \mathcal{U}(1, T) \\ x_{t,t-1}^w \sim p_{\theta}(x_{t,t-1}^w | x_0^w, c) \\ x_{t,t-1}^l \sim p_{\theta}(x_{t,t-1}^l | x_0^l, c)}} \left[\log \sigma \left(\beta T \left[\log \frac{p_{\theta}(x_t^w | x_{t-1}^w, c)}{p_{\text{ref}}(x_t^w | x_{t-1}^w, c)} - \log \frac{p_{\theta}(x_t^l | x_{t-1}^l, c)}{p_{\text{ref}}(x_t^l | x_{t-1}^l, c)} \right] \right) \right] \quad (12)$$

This is all good, however sampling from $p_{\theta}(x_{t,t-1} | x_0, c)$ is still intractable, here they approximate $p(x_{t,t-1} | x_0, c)$ with $q(x_{t,t-1} | x_0, c) = q(x_t | x_0, c)q(x_{t-1} | x_t, x_0, c)$ or $p_{\theta}(x_{t-1} | x_t, c)q(x_t | x_0, c)$ arriving at some alternatives bounds.

Using the gaussian nature of the forward and recess process, the bound can be rewritten in terms of $\epsilon_\theta(x_t, t)$ and the noise prediction model.

Consistency Models

Rewriting the RLHF objective, we arrive at the following problem.

$$\max_{p_\theta} \mathbb{E}_{c \sim \mathcal{D}_c, x_{\tau_H} \sim p_\theta(x_{\tau_H}|c)} [r(c, f_\theta(x_{\tau_H}, \tau_H, c))] - \beta \mathbb{D}_{\text{KL}} [p_\theta(x_0|c) \| p_{\text{ref}}(x_0|c)] \quad (13)$$

Note that we require the distributions of $x_0 = f_\theta(x_{\tau_H}, \tau_H, c)$, $x_{\tau_H} \sim p_\theta(x_{\tau_H}|c)$ to remain similar, not x_{τ_H} . One simple approach could be to approximate $x_{\tau_H} \approx x_0$, if $\tau_H \approx 0$, and to perform preference optimization with respect to the noisy outputs, for now we absorb the final operation of $x_0 = f_\theta(x_{\tau_H}, \tau_H, c)$ into the policy itself, using $\pi(a_{\tau_H}|s_{\tau_H})$ which would be a deterministic transform, and $\tau_{H+1} = 0$.

We can imitate the derivation of Diffusion DPO, to find the following objective. (We have a huge problem with the last deterministic step in the chain, let's ignore it for now, and address it later)

$$p_\theta(x_{\tau_0:H+1}|c) = \prod_{t=0}^H p_\theta(x_{\tau_{t+1}}|x_{\tau_t}, c), \quad p_\theta(x_{\tau_{t+1}}|x_{\tau_t}, c) = \mathcal{N}(x_{\tau_{t+1}}; f_\theta(x_{\tau_t}, \tau_t, c), (\tau_t^2 - \tau_{H+1}^2)\mathbf{I}) \quad (14)$$

$$r(c, x_0) = \mathbb{E}_{p_\theta(x_{\tau_0:H}|x_{\tau_{H+1}}=x_0, c)} [R(c, x_{\tau_0:H+1})] \quad (15)$$

$$\max_{p_\theta} \mathbb{E}_{c \sim \mathcal{D}_c, x_{\tau_0:H+1} \sim p_\theta(x_{\tau_0:H+1}|c)} [R(c, x_{\tau_0:H+1})] - \beta \mathbb{D}_{\text{KL}} [p_\theta(x_{\tau_0:H+1}|c) \| p_{\text{ref}}(x_{\tau_0:H+1}|c)] \quad (16)$$

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{c, x_{\tau_{H+1}}^w, x_{\tau_{H+1}}^l \sim \mathcal{D}} \left[\log \sigma \left(\beta \mathbb{E}_{\substack{x_{\tau_0:H}^w \sim p_\theta(x_{\tau_0:H}^w | x_{\tau_{H+1}}^w, c) \\ x_{\tau_0:H}^l \sim p_\theta(x_{\tau_0:H}^l | x_{\tau_{H+1}}^l, c)}} \left[\log \frac{p_\theta(x_{\tau_0:H+1}^w | c)}{p_{\text{ref}}(x_{\tau_0:H+1}^w | c)} - \log \frac{p_\theta(x_{\tau_0:H+1}^l | c)}{p_{\text{ref}}(x_{\tau_0:H+1}^l | c)} \right] \right) \right] \quad (17)$$

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{c, x_{\tau_{H+1}}^w, x_{\tau_{H+1}}^l \sim \mathcal{D}} \left[\log \sigma \left(\beta \mathbb{E}_{\substack{x_{\tau_0:H}^w \sim p_\theta(x_{\tau_0:H}^w | x_{\tau_{H+1}}^w, c) \\ x_{\tau_0:H}^l \sim p_\theta(x_{\tau_0:H}^l | x_{\tau_{H+1}}^l, c)}} \left[\sum_{t=0}^H \log \frac{p_\theta(x_{\tau_{t+1}}^w | x_{\tau_t}^w, c)}{p_{\text{ref}}(x_{\tau_{t+1}}^w | x_{\tau_t}^w, c)} - \log \frac{p_\theta(x_{\tau_{t+1}}^l | x_{\tau_t}^l, c)}{p_{\text{ref}}(x_{\tau_{t+1}}^l | x_{\tau_t}^l, c)} \right] \right) \right] \quad (18)$$

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{c, x_{\tau_{H+1}}^w, x_{\tau_{H+1}}^l \sim \mathcal{D}} \left[\log \sigma \left(\sum_{t=0}^H \beta \mathbb{E}_{\substack{x_{\tau_t, t+1}^w \sim p_\theta(x_{\tau_t, t+1}^w | x_{\tau_{H+1}}^w, c) \\ x_{\tau_t, t+1}^l \sim p_\theta(x_{\tau_t, t+1}^l | x_{\tau_{H+1}}^l, c)}} \left[\log \frac{p_\theta(x_{\tau_{t+1}}^w | x_{\tau_t}^w, c)}{p_{\text{ref}}(x_{\tau_{t+1}}^w | x_{\tau_t}^w, c)} - \log \frac{p_\theta(x_{\tau_{t+1}}^l | x_{\tau_t}^l, c)}{p_{\text{ref}}(x_{\tau_{t+1}}^l | x_{\tau_t}^l, c)} \right] \right) \right] \quad (19)$$

Since H is small, we will keep the sum in this case, taking the whole trajectory into count. the remaining problem is that sampling from $p(x_{\tau_t, t+1}|x_{\tau_{H+1}})$ remains intractable, the overarching problem is that unlike diffusion models, there is not forward-backward process, and that consistency models take huge leaps in between states, hindering our approximations. To minimize errors, we first calculate any tractable part of the previous equations.

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{c, x_{\tau_{H+1}}^w, x_{\tau_{H+1}}^l \sim \mathcal{D}} \left[\log \sigma \left(\sum_{t=0}^H \beta \mathbb{E}_{\substack{x_{\tau_t, t+1}^w \sim p_\theta(x_{\tau_t, t+1}^w | x_{\tau_{H+1}}^w, c) \\ x_{\tau_t, t+1}^l \sim p_\theta(x_{\tau_t, t+1}^l | x_{\tau_{H+1}}^l, c)}} \left[\log \frac{p_\theta(x_{\tau_{t+1}}^w | x_{\tau_t}^w, c)}{p_{\text{ref}}(x_{\tau_{t+1}}^w | x_{\tau_t}^w, c)} - \log \frac{p_\theta(x_{\tau_{t+1}}^l | x_{\tau_t}^l, c)}{p_{\text{ref}}(x_{\tau_{t+1}}^l | x_{\tau_t}^l, c)} \right] \right) \right] \quad (20)$$

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{c, x_{\tau_{H+1}}^w, x_{\tau_{H+1}}^l \sim \mathcal{D}} \left[\log \sigma \left(\sum_{t=0}^H \beta \mathbb{E}_{\substack{x_{\tau_t, t+1}^w \sim p_\theta(x_{\tau_t, t+1}^w | x_{\tau_t}^w, c), x_{\tau_t}^w \sim p_\theta(x_{\tau_t}^w | x_{\tau_{H+1}}^w, c) \\ x_{\tau_t, t+1}^l \sim p_\theta(x_{\tau_t, t+1}^l | x_{\tau_t}^l, c), x_{\tau_t}^l \sim p_\theta(x_{\tau_t}^l | x_{\tau_{H+1}}^l, c)}} \left[\log \frac{p_\theta(x_{\tau_{t+1}}^w | x_{\tau_t}^w, c)}{p_{\text{ref}}(x_{\tau_{t+1}}^w | x_{\tau_t}^w, c)} - \log \frac{p_\theta(x_{\tau_{t+1}}^l | x_{\tau_t}^l, c)}{p_{\text{ref}}(x_{\tau_{t+1}}^l | x_{\tau_t}^l, c)} \right] \right) \right] \quad (21)$$

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{c, x_{\tau_{H+1}}^w, x_{\tau_{H+1}}^l \sim \mathcal{D}} \left[\log \sigma \left(\sum_{t=0}^H \beta \mathbb{E}_{\substack{x_{\tau_{t+1}}^w \sim p_\theta(x_{\tau_{t+1}}^w | x_{\tau_t}^w, c), x_{\tau_t}^w \sim p_\theta(x_{\tau_t}^w | x_{\tau_{H+1}}^w, c) \\ x_{\tau_{t+1}}^l \sim p_\theta(x_{\tau_{t+1}}^l | x_{\tau_t}^l, c), x_{\tau_t}^l \sim p_\theta(x_{\tau_t}^l | x_{\tau_{H+1}}^l, c)}} \left[\log \frac{p_\theta(x_{\tau_{t+1}}^w | x_{\tau_t}^w, c)}{p_{\text{ref}}(x_{\tau_{t+1}}^w | x_{\tau_t}^w, c)} - \log \frac{p_\theta(x_{\tau_{t+1}}^l | x_{\tau_t}^l, c)}{p_{\text{ref}}(x_{\tau_{t+1}}^l | x_{\tau_t}^l, c)} \right] \right) \right] \quad (22)$$

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{c, x_{\tau_{H+1}}^w, x_{\tau_{H+1}}^l \sim \mathcal{D}} \left[\log \sigma \left(\sum_{t=0}^H \beta \mathbb{E}_{\substack{x_{\tau_t}^w \sim p_\theta(x_{\tau_t}^w | x_{\tau_{H+1}}^w, c) \\ x_{\tau_t}^l \sim p_\theta(x_{\tau_t}^l | x_{\tau_{H+1}}^l, c)}} \left[\mathbb{D}_{\text{KL}}(p_\theta(x_{\tau_{t+1}}^w | x_{\tau_t}^w, c) \| p_{\text{ref}}(x_{\tau_{t+1}}^w | x_{\tau_t}^w, c)) \right. \right. \right. \\ \left. \left. \left. - \mathbb{D}_{\text{KL}}(p_\theta(x_{\tau_{t+1}}^l | x_{\tau_t}^l, c) \| p_{\text{ref}}(x_{\tau_{t+1}}^l | x_{\tau_t}^l, c)) \right] \right) \right] \quad (23)$$

The objective has been simplified, however sampling from $p_\theta(x_{\tau_t}|x_{\tau_{H+1}}, c)$ remains intractable.

— The challenge regarding deterministic mappings

First, let us calculate the KL-divergence between the two gaussian distributions.

$$\mathbb{D}_{\text{KL}}(p_\theta(x_{\tau_{t+1}}|x_{\tau_t}, c) \| p_{\text{ref}}(x_{\tau_{t+1}}|x_{\tau_t}, c)) = \frac{\|f_\theta(x_{\tau_t}, \tau_t, c) - f_{\text{ref}}(x_{\tau_t}, \tau_t, c)\|^2}{2(\tau_t^2 - \tau_H^2)} = \frac{d(f_\theta(x_{\tau_t}, \tau_t, c), f_{\text{ref}}(x_{\tau_t}, \tau_t, c))}{2(\tau_t^2 - \tau_H^2)} \quad (24)$$

Here, we can see the problem regarding the last iteration, since they are exact mappings with zero variance, we would arrive at:

$$\mathbb{D}_{\text{KL}}(p_\theta(x_{\tau_{H+1}}^w|x_{\tau_H}^w, c) \| p_{\text{ref}}(x_{\tau_{H+1}}^w|x_{\tau_H}^w, c)) = \begin{cases} 0 & f_\theta(x_{\tau_H}, \tau_H, c) = f_{\text{ref}}(x_{\tau_H}, \tau_H, c) \\ \infty & f_\theta(x_{\tau_H}, \tau_H, c) \neq f_{\text{ref}}(x_{\tau_H}, \tau_H, c) \end{cases} \quad (25)$$

We can approach this problem in many ways, including

- Ignoring this final term all together, since $f_\theta(x_{\tau_H}, \tau_H, c) \approx f_{\text{ref}}(x_{\tau_H}, \tau_H, c)$ if τ_H is sufficiently small.
- Smoothing, instead of an indicator function, we can smoothen this divergence by using a smoothened L^2 divergence of $\frac{(f_\theta(x_{\tau_H}, \tau_H, c) - f_{\text{ref}}(x_{\tau_H}, \tau_H, c))^2}{2\epsilon^2} = \frac{d(f_\theta(x_{\tau_H}, \tau_H, c), f_{\text{ref}}(x_{\tau_H}, \tau_H, c))}{2\epsilon^2}$ for some sufficiently small ϵ .
- Instead of using the KL-regularized MDP for our RLHF optimization, we can use an alternative f -Divergence regularized MDP with an f -Divergence that is not domain sensitive.
- We can consider that the indicator function is equivalent to $\max_{\nu \in \mathbb{R}} \{\nu(f_\theta(x_{\tau_H}, \tau_H, c) - f_{\text{ref}}(x_{\tau_H}, \tau_H, c))\}$
- We can treat it as a constraint on the problem, and either use primal-dual methods to satisfy the constraint, or again smoothen the constraint by adding terms to the loss function such as $\lambda(\|f_\theta(x_{\tau_H}, \tau_H, c) - f_{\text{ref}}(x_{\tau_H}, \tau_H, c)\|^2) = \lambda d(f_\theta(x_{\tau_H}, \tau_H, c), f_{\text{ref}}(x_{\tau_H}, \tau_H, c))$, This technique can be effectively coupled with the consistency model loss.
- If possible, find an alternative bound for $\mathbb{D}_{\text{KL}}(p_\theta(x_0|c) \| p_{\text{ref}}(x_0|c))$ than the divergence of the entire chain.

— The challenge regarding intractability of $p(x_{\tau_t}|x_{\tau_{H+1}}, c)$

For diffusion models, we were able to approximate samples from $p(x_t|x_0, c)$ with samples from the forward diffusion process $q(x_t|x_0)$, which would be equivalent to adding gaussian noise to the input.

In general, the following relation holds, which would simplify matters since $p(x_{\tau_t}|x_{\tau_{t+1}}, c)$ is easier to work with, since it involves two consecutive time steps rather than only the end of the chain.

$$\begin{aligned} p(x_{\tau_t}|x_{\tau_{H+1}}, c) &= \mathbb{E}_{x_{\tau_{t+1}} \sim p(x_{\tau_{t+1}}|x_{\tau_{H+1}}, c)} [p(x_{\tau_t}|x_{\tau_{H+1}}, x_{\tau_{t+1}}, c)] = \sum_{x_{\tau_{t+1}}} p(x_{\tau_t}|x_{\tau_{H+1}}, x_{\tau_{t+1}}, c) p(x_{\tau_{t+1}}|x_{\tau_{H+1}}, c) \\ &= \sum_{x_{\tau_{t+1}}} p(x_{\tau_t}|x_{\tau_{t+1}}, c) p(x_{\tau_{t+1}}|x_{\tau_{H+1}}, c) \quad \text{Markov chain} \Rightarrow x_{\tau_t}|x_{\tau_{t+1}} \perp x_{\tau_{H+1}} \\ &= \sum_{x_{\tau_{t+1}}} p(x_{\tau_t}|x_{\tau_{t+1}}, c) \sum_{x_{\tau_{t+2}}} p(x_{\tau_{t+1}}|x_{\tau_{t+2}}, c) \sum_{x_{\tau_{t+2}}} \cdots \sum_{x_{\tau_H}} p(x_{\tau_{H-1}}|x_{\tau_H}, c) p(x_{\tau_H}|x_{\tau_{H+1}}, c) \end{aligned} \quad (26)$$

For diffusion models, we can approximate $p(x_t|x_{t-1}, c)$ with $q(x_t|x_{t-1})$ since only a small amount of noise has been removed between the two stages. arriving at the final bound of $q(x_t|x_0)$. for consistency models we can utilize the following approximation, effectively treating $f(x_{\tau_t}, \tau_t, c)$ as the inverse of adding a noise $\sqrt{\tau_t^2 - \tau_H^2}z$ and vice versa.

$$p(x_{\tau_{t+1}}|x_{\tau_t}, c) = f(x_{\tau_t}, \tau_t, c) + \sqrt{\tau_t^2 - \tau_H^2}Z \Rightarrow p(x_{\tau_t}|x_{\tau_{t+1}}, c) \approx f(x_{\tau_{t+1}}, \tau_{t+1}, c) + \sqrt{\tau_{t-1}^2 - \tau_H^2}Z \triangleq q_1(x_{\tau_t}|x_{\tau_{t+1}}, c) \quad (27)$$

$$p(x_{\tau_t}|x_{\tau_{t+1}}, c) \approx x_{\tau_{t+1}} + \sqrt{\tau_{t-1}^2 - \tau_t^2}Z = q_2(x_{\tau_t}|x_{\tau_{t+1}}, c) \quad (28)$$

Incorporating q_1 or q_2 into the previous equation, we arrive at the following, with q admitting a markov structure.

$$p(x_{\tau_t}|x_{\tau_{H+1}}, c) \approx q(x_{\tau_t}|x_{\tau_{H+1}}, c), \quad q(x_{\tau_{t,H+1}}|x_{\tau_{H+1}}, c) = \prod_{h=H}^t q(x_{\tau_h}|x_{\tau_{h+1}}, c) \quad (29)$$

Since our expectation in equation 23 only relies on samples from $p(x_{\tau_t}|x_{\tau_{H+1}}, c)$ and is in a summation form of $\sum_{t=0}^H \mathbb{E}_{x_{\tau_t}^*} [h(x_{\tau_t})]$, We can utilize the same chain for each time step, transforming it into $\mathbb{E}_{x_{\tau_0:H}^*} \sum_{t=0}^H h(x_{\tau_t})$, and we can use samples from the entire chain for our monte-carlo estimation.

Final form

Ignoring the problem caused by deterministic mappings in our writings, by estimating samples and calculating the KL-Divergence, we arrive at the following preference optimization objective.

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{c, x_{\tau_{H+1}}^w, x_{\tau_{H+1}}^l \sim \mathcal{D}} \left[\log \sigma \left(\beta \mathbb{E}_{\substack{x_{\tau_{0:H}}^w \sim q_\theta(x_{\tau_{0:H}}^w | x_{\tau_{H+1}}^w, c) \\ x_{\tau_{0:H}}^l \sim q_\theta(x_{\tau_{0:H}}^l | x_{\tau_{H+1}}^l, c)}} \left[\sum_{t=0}^H w(t) (d(f_\theta(x_{\tau_t}^w, \tau_t, c), f_{\text{ref}}(x_{\tau_t}^w, \tau_t, c)) - d(f_\theta(x_{\tau_t}^l, \tau_t, c), f_{\text{ref}}(x_{\tau_t}^l, \tau_t, c))) \right] \right) \right] \quad (30)$$

Using Jensen's Inequality, we can reduce this to a single expectation.

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(c, x_{\tau_{H+1}}^w, x_{\tau_{H+1}}^l) \sim \mathcal{D}, x_{\tau_{0:H}}^w \sim q_\theta(x_{\tau_{0:H}}^w | x_{\tau_{H+1}}^w, c), x_{\tau_{0:H}}^l \sim q_\theta(x_{\tau_{0:H}}^l | x_{\tau_{H+1}}^l, c)} \left[\log \sigma \left(\beta \sum_{t=0}^H w(t) (d(f_\theta(x_{\tau_t}^w, \tau_t, c), f_{\text{ref}}(x_{\tau_t}^w, \tau_t, c)) - d(f_\theta(x_{\tau_t}^l, \tau_t, c), f_{\text{ref}}(x_{\tau_t}^l, \tau_t, c))) \right) \right] \quad (31)$$

Where $w(t) = \frac{1}{\max\{\tau_t^2 - \tau_H^2, \epsilon^2\}}$, with some sufficiently small $\epsilon > 0$, and q being an appropriate approximate. The loss can then be appropriately approximated using monte-carlo methods.

While the estimation q_θ relies on the model parameters θ , the optimization is relatively straightforward as we can use the reparametrization trick for optimization, and take the expectation with respect to samples from standard gaussian distributions.

And finally, to ensure that the model remains consistent, and to partially mitigate the problem regarding deterministic mappings, our training process can use the final objective function bellow.

$$\mathcal{L} = \mathcal{L}_{\text{DPO}} + \lambda \mathcal{L}_{\text{con}} \quad (32)$$

Reward estimate

If we consider the equivalence between equation 30 and the bradley terry loss in equation 4, we can find the following reward estimate.

$$r(c, x_0^w) - r(c, x_0^l) = \beta h_{\pi_\theta}(c, x_0^w, x_0^l) = \beta \mathbb{E}_{\substack{x_{\tau_{0:H}}^w \sim q_\theta(x_{\tau_{0:H}}^w | x_{\tau_{H+1}}^w, c) \\ x_{\tau_{0:H}}^l \sim q_\theta(x_{\tau_{0:H}}^l | x_{\tau_{H+1}}^l, c)}} \left[\sum_{t=0}^H w(t) (d(f_\theta(x_{\tau_t}^w, \tau_t, c), f_{\text{ref}}(x_{\tau_t}^w, \tau_t, c)) - d(f_\theta(x_{\tau_t}^l, \tau_t, c), f_{\text{ref}}(x_{\tau_t}^l, \tau_t, c))) \right] \quad (33)$$

And therefore we can extract a reward function from the policy.

$$r(c, x_0^w) = \beta \mathbb{E}_{\substack{x_{\tau_{0:H}}^w \sim q_\theta(x_{\tau_{0:H}}^w | x_{\tau_{H+1}}^w, c) \\ x_{\tau_{0:H}}^l \sim q_\theta(x_{\tau_{0:H}}^l | x_{\tau_{H+1}}^l, c)}} \left[\sum_{t=0}^H w(t) d(f_\theta(x_{\tau_t}^w, \tau_t, c), f_{\text{ref}}(x_{\tau_t}^w, \tau_t, c)) \right] + \beta \log Z(c) \quad (34)$$

Using this reward estimate, we can extend this formulation to any preference optimization method.