The Issue of Data Drifting in PPO

Qiang Liu

In general, proximal point updates of policy optimization can be written as

$$\max_{\pi\in\Delta} \mathbb{E}_{s\sim d_0,\, a\sim\pi^{\rm ref}(\cdot|s)} \left[A(s,a)\, w_\pi(s,a) - \beta\, \phi(w_\pi(s,a))\right], \qquad \text{with} \quad w_\pi(s,a) = \frac{\pi(a \mid s)}{\pi^{\rm ref}(a \mid s)},$$

where Δ is the set of valid policy distributions, $w_\pi(s,a)$ is the density ratio, and $\phi\colon [0,+\infty) \to \mathbb{R}$ is a penalty function that attains a unique minimum at $w=1$, thereby encouraging uniform density ratios. The common case of a KL divergence penalty is recovered by setting

$$\phi_{\rm KL}(w) = w\log w - w + 1,$$

since

$$\begin{aligned}\mathrm{KL}(\pi(\cdot|s)\,\|\,\pi^{\rm ref}(\cdot|s)) &= \mathbb{E}_{a\sim\pi^{\rm ref}(\cdot|s)}\left[w_\pi(s,a)\log w_\pi(s,a)\right] \\ &= \mathbb{E}_{a\sim\pi^{\rm ref}(\cdot|s)}\left[w_\pi(s,a)\log w_\pi(s,a) - w_\pi(s,a) + 1\right].\end{aligned}$$

In the KL case, the optimal policy has a closed form:

$$\pi^{\rm new}(a|s) = \frac{1}{Z(s)}\pi^{\rm ref}(a|s)\cdot\exp\left(\frac{A(s,a)}{\beta}\right),$$

where $Z(s) = \sum_a \pi^{\rm ref}(a|s)\cdot\exp\left(\frac{A(s,a)}{\beta}\right)$ is the normalization constant.

For a general strictly convex function ϕ with minimum at $w=1$, the optimal policy satisfies

$$w_{\pi^*}(s,a) = \max\left((\nabla\phi)^{-1}\left(\frac{A(s,a)-\lambda(s)}{\beta}\right),\, 0\right),$$

where $(\nabla\phi)^{-1}$ denotes the inverse function of $\nabla\phi$, and $\lambda(s)$ is chosen to enforce the normalization condition $\sum_a w_{\pi^*}(s,a)\pi^{\rm ref}(a \mid s) = 1$. This constraint corresponds to the Lagrange multiplier associated with $\sum_a \pi(a \mid s) = 1$ in the optimization.

However, in practice, data drift often occurs: the data used to train models may not be generated by $\pi^{\rm ref}$. It may be drawn from a mixture of different models or come from human annotators, whose distributions may be unknown or lack analytic form.

Let $\pi^{\rm data}$ denote the actual distribution of the training data. Then the objective we are effectively optimizing is

$$\max_{\pi\in\Delta} \mathbb{E}_{s\sim d_0,\, a\sim\pi^{\rm data}(\cdot|s)} \left[A(s,a)\, w_\pi(s,a) - \beta\, \phi(w_\pi(s,a))\right], \tag{1}$$

where the density ratio $w_\pi(s,a) = \frac{\pi(a|s)}{\pi^{\rm ref}(a|s)}$ is still defined relative to $\pi^{\rm ref}$, but the data distribution $\pi^{\rm data}$ differs.

What is the effect of using different π^{data} and π^{ref}?

As it turns out, the discrepancy between π^{data} and π^{ref} introduces a subtle bias in the solution. One can show that the optimal solution to (1) is

$$\pi^{\text{new}}(a|s) = \pi^{\text{ref}}(a|s) \cdot \max\left\{(\nabla\phi)^{-1}\left(\frac{1}{\beta}\left[A(s,a) - \lambda(s)\frac{\pi^{\text{ref}}(a|s)}{\pi^{\text{data}}(a|s)}\right]\right),\ 0\right\}, \tag{2}$$

where the density ratio of data vs reference, $w_{\text{ref/data}}(s,a) = \frac{\pi^{\text{ref}}(a|s)}{\pi^{\text{data}}(a|s)}$, appears in the Lagrangian term $\lambda(s)$.

In particular, the KL penalty yields

$$\pi^{\text{new}}(a \mid s) = \pi^{\text{ref}}(a \mid s)\ \exp\left(\frac{1}{\beta}\left(A(s,a) - w_{\text{ref/data}}(s,a)\lambda(s)\right)\right).$$

How does this discrepancy bias the solution?

It depends on the sign of the Lagrange multiplier $\lambda(s)$. If $\lambda(s) \geq 0$, the term $-\lambda(s)w_{\text{ref/data}}(s,a)$ in the exponent acts as a penalty on actions where $w_{\text{ref/data}}(s,a)$ is large (i.e., actions that are rare in the data relative to the reference). Consequently, the resulting policy π^{new} is systematically biased toward actions that are more common in the data distribution π^{data}.

This condition ($\lambda(s) \geq 0$) is indeed the typical case. It holds whenever the unnormalized partition function satisfies

$$\sum_a \pi^{\text{ref}}(a \mid s) \exp\left(\frac{A(s,a)}{\beta}\right) \geq 1.$$

Assume the advantage function $A(s,a)$ is centered such that $\mathbb{E}_{a\sim\pi^{\text{ref}}(\cdot|s)}[A(s,a)] = 0$, Jensen's inequality guarantees that

$$\mathbb{E}_{a\sim\pi^{\text{ref}}(\cdot|s)}\left[\exp\left(\frac{A(s,a)}{\beta}\right)\right] \geq \exp\left(\mathbb{E}_{a\sim\pi^{\text{ref}}(\cdot|s)}\left[\frac{A(s,a)}{\beta}\right]\right) = \exp(0) = 1.$$

Is it possible to eliminate this bias?

Yes. As shown by the proof below, the bias arises from the normalization constraint $\sum_a \pi(a \mid s) = 1$. We can eliminate it by making w unconstrained. This can be achieved by introducing a scalar function V:

$$\max_{\pi\in\Delta,\, V\in\mathcal{V}} \mathbb{E}_{s\sim d_0,\, a\sim\pi^{\text{ref}}(\cdot|s)}\left[A(s,a)\, w_{\pi,V}(s,a) - \beta\phi(w_{\pi,V}(s,a))\right], \qquad \text{with} \quad w_{\pi,V}(s,a) = V(s)\frac{\pi(a \mid s)}{\pi^{\text{ref}}(a \mid s)}. \tag{3}$$

where V is optimized in the set of all functions of s. This removes the normalization constraint on w, yielding

$$w^*(s,a) = \max\left\{(\nabla\phi)^{-1}\left(\frac{A(s,a)}{\beta}\right),\quad 0\right\}, \tag{4}$$

and therefore:

$$\pi^{\text{new}}(a \mid s) = \frac{1}{V(s)}\pi^{\text{ref}}(a \mid s)w^*(s,a), \qquad V(s) = \sum_a \pi^{\text{ref}}(a \mid s)w^*(s,a). \tag{5}$$

Here, V plays the role of a value function. Therefore, we jointly learn the policy and value function.

1 Proofs

Theorem 1.1. *Assume ϕ: $[0,\infty) \to \mathbb{R}$ is strictly convex and differentiable. The solution of (1) satisfies the condition in (2).*

Proof. Let $w(s,a) = \frac{\pi(a|s)}{\pi^{\text{ref}}(a|s)}$. The optimization problem with respect to w is:

$$\max_w \sum_{a,s} d_0(s)\pi^{\text{data}}(a|s)\left[A(s,a)w(s,a) - \beta\phi(w(s,a))\right]$$

subject to the constraints:

$$w(s,a) \geq 0, \quad \text{and} \quad \sum_a w(s,a)\pi^{\text{ref}}(a|s) = 1.$$

The Lagrangian for this problem is:

$$\mathcal{L}(w,\lambda) = \sum_{a,s} d_0(s)\pi^{\text{data}}(a|s)\left[[A(s,a)w(s,a) - \beta\phi(w(s,a))] - \lambda(s)(\sum_a w(s,a)\pi^{\text{ref}}(a|s) - 1)\right].$$

Taking the derivative with respect to $w(s,a)$:

$$\frac{\partial \mathcal{L}}{\partial w(s,a)} = d_0(s)(\pi^{\text{data}}(a|s)A(s,a) - \beta\pi^{\text{data}}(a|s)\nabla\phi(w(s,a)) - \lambda(s)\pi^{\text{ref}}(a|s)).$$

Setting the derivative to zero for the optimal solution w^* (assuming an interior solution $w^* > 0$):

$$\beta\pi^{\text{data}}(a|s)\nabla\phi(w^*(s,a)) = \pi^{\text{data}}(a|s)A(s,a) - \lambda(s)\pi^{\text{ref}}(a|s).$$

Dividing by $\beta\pi^{\text{data}}(a|s)$ (assuming $\pi^{\text{data}}(a|s) > 0$):

$$\nabla\phi(w^*(s,a)) = \frac{1}{\beta}\left(A(s,a) - \lambda(s)\frac{\pi^{\text{ref}}(a|s)}{\pi^{\text{data}}(a|s)}\right).$$

Inverting $\nabla\phi$ and applying the non-negativity constraint gives:

$$w^*(s,a) = \max\left\{(\nabla\phi)^{-1}\left(\frac{1}{\beta}\left[A(s,a) - \lambda(s)\frac{\pi^{\text{ref}}(a|s)}{\pi^{\text{data}}(a|s)}\right]\right), 0\right\}.$$

Finally, substituting $w^*(s,a)$ back into $\pi^{\text{new}}(a|s) = w^*(s,a)\pi^{\text{ref}}(a|s)$ yields the result. □

Theorem 1.2. *Assume ϕ: $[0,\infty) \to \mathbb{R}$ is strictly convex. The solution of (3) is given by (4)-(5).*

Proof. In the unconstrained formulation defined in (3), we optimize over both π and V. Note that the term $w_{\pi,V}(s,a) = V(s)\frac{\pi(a|s)}{\pi^{\text{ref}}(a|s)}$ can be treated as a free optimization variable for each (s,a), denoted simply as $w(s,a)$, because $V(s)$ absorbs the normalization constant. Specifically, for any non-negative function $w(s,a)$, we can recover a valid π and V by setting $V(s) = \sum_a w(s,a)\pi^{\text{ref}}(a|s)$ and $\pi(a|s) = \frac{w(s,a)\pi^{\text{ref}}(a|s)}{V(s)}$.

Thus, the optimization decouples for each (s,a) under the reference distribution:

$$\max_{w(s,a)\geq 0} \sum_{a,s} d_0(s)\pi^{\text{data}}(a \mid s)\left[A(s,a)w(s,a) - \beta\phi(w(s,a))\right].$$

Taking the derivative with respect to $w(s,a)$ and setting it to zero:

$$A(s,a) - \beta \nabla \phi(w^*(s,a)) = 0 \implies \nabla \phi(w^*(s,a)) = \frac{A(s,a)}{\beta}.$$

Inverting $\nabla \phi$ gives:

$$w^*(s,a) = (\nabla \phi)^{-1} \left(\frac{A(s,a)}{\beta} \right).$$

Applying the non-negativity constraint yields Eq. (4). The expressions for π^{new} and V in Eq. (5) follow immediately from the definition of $w_{\pi,V}$. □

References