2.2 核心抑制：去你梦里的彼岸

众所周知，强化学习是一种最优控制的类似物。这里介绍一些使用强化学习框架对（图像）生成模型进行微调的方法。

2.2.1 Score Matching 的策略梯度

首先我们来复习标准的 Policy Gradient 算法。设强化学习中的策略被 θ 参数化，从而每条轨迹的概率 $\mathbb{P}_\theta$ 也依赖于 θ。我们希望最大化所有轨迹上奖励的期望 $\mathbb{E}_{\tau\sim\mathbb{P}_\theta}[R[\tau]]$，$R[\tau]$ 是奖励泛函。我们希望通过梯度上升的方式提升这个奖励，并且只能从 $\mathbb{P}_\theta$ 中取样。标准的 Policy Gradient / REINFORCE 算法是：

$$\begin{aligned}\nabla_\theta \mathbb{E}_{\tau\sim\mathbb{P}_\theta}[R[\tau]] &= \nabla_\theta \int \mathbb{P}_\theta[\tau]R[\tau]\,\mathrm{d}\tau \\ &= \int \nabla_\theta \mathbb{P}_\theta[\tau]R[\tau]\,\mathrm{d}\tau \\ &= \int \mathbb{P}_\theta[\tau]\nabla_\theta \log \mathbb{P}_\theta[\tau]R[\tau]\,\mathrm{d}\tau \\ &= \mathbb{E}_{\tau\sim\mathbb{P}_\theta}\left[\nabla_\theta \log \mathbb{P}_\theta[\tau]R[\tau]\right].\end{aligned}$$

现在考虑用 Policy Gradient 对扩散模型进行微调，此时 $R[\tau]$ 通常只与轨迹的终点有关。$\mathbb{P}_\theta$ 很容易计算。设生成时使用的 SDE 是：

$$\mathrm{d}X_t = b_\theta(X_t, t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t.$$

无穷小传播子

$$\mathbb{P}(x \to x+\Delta x,\ t \to t+\Delta t) = \frac{1}{Z}\exp\left[-\frac{\|\Delta x - b_\theta(x,t)\Delta t\|^2}{2\sigma^2\Delta t}\right].$$

为了得到一条路径的概率，只需连乘传播子：

$$\begin{aligned}\mathbb{P}[\tau] &= \mathbb{P}\left[x_0,\ x_1 \sim x_1 + \mathrm{d}x_1,\ \ldots,\ x_N \sim x_N + \mathrm{d}x_N\right] \\ &= \frac{1}{\mathcal{Z}}\exp\left[-\sum_{i=1}^N \frac{\|\Delta x_i - b_\theta(x_i,t)\Delta t\|^2}{2\sigma^2\Delta t}\right] \\ &= \frac{1}{\mathcal{Z}}\exp\left[-\sum_{i=1}^N \frac{\|\Delta x_i - b_\theta(x_i,t)\Delta t\|^2}{2\sigma^2(\Delta t)^2}\Delta t\right] \\ &= \frac{1}{\mathcal{Z}}\exp\left[-\sum_{i=1}^N \frac{\left\|\frac{\Delta x_i}{\Delta t} - b_\theta(x_i,t)\right\|^2}{2\sigma^2}\Delta t\right] \\ &= \frac{1}{\mathcal{Z}}\exp\left[-\sum_{i=1}^N \frac{\|\dot{x}_i - b_\theta(x_i,t)\|^2}{2\sigma^2}\Delta t\right] \\ &\approx \frac{1}{\mathcal{Z}}\exp(-S).\end{aligned}$$

作用量：

$$S = \frac{1}{2\sigma^2}\int \|\dot{x} - b_\theta(x,t)\|^2\,\mathrm{d}t.$$

因此，忽略常数（反正常数会在求梯度的过程中被扔掉），获得

$$\log \mathbb{P}_\theta[\tau] \sim -\frac{1}{2\sigma^2}\int \|\dot{x}_t - b_\theta(x_t,t)\|^2\,\mathrm{d}t.$$

利用之前的 Policy Gradient 结论计算梯度即可：

$$\begin{aligned}\nabla_\theta \mathbb{E}_{\tau\sim\mathbb{P}_\theta}[R[\tau]] &= \mathbb{E}_{\tau\sim\mathbb{P}_\theta}\left[\nabla_\theta \log \mathbb{P}_\theta[\tau] R[\tau]\right] \\ &\sim \mathbb{E}_{\tau\sim\mathbb{P}_\theta}\left[\nabla_\theta\left(-\frac{1}{2\sigma^2}\int \|\dot{x}_t - b_\theta(x_t,t)\|^2\,\mathrm{d}t\right) R[\tau]\right] \\ &\sim \mathbb{E}_{\tau\sim\mathbb{P}_\theta}\left[\frac{1}{\sigma^2}\left(\int \left(\frac{\partial b_\theta}{\partial\theta}\right)^{\mathsf{T}}\left(\dot{x}_t - b_\theta(x_t,t)\right)\mathrm{d}t\right) R[\tau]\right] \\ &\sim \mathbb{E}_{\tau\sim\mathbb{P}_\theta}\left[\frac{1}{\sigma^2}\left(\int \left(\frac{\partial b_\theta}{\partial\theta}\right)^{\mathsf{T}}\left(\mathrm{d}X_t - b_\theta(X_t,t)\,\mathrm{d}t\right)\right) R[\tau]\right].\end{aligned}$$

这就完成了推导。

值得注意的是，这里的损失可以被转化为一个 Flow Matching 风味的回归损失。固定旧参数 $\bar{\theta}$，记：

$$F_t[\tau] := \frac{R[\tau]}{\sigma^2}\left(\dot{x}_t - b_{\bar{\theta}}(x_t,t)\right).$$

上面的策略梯度写为：

$$\nabla_\theta \mathbb{E}_{\tau\sim\mathbb{P}_\theta}[R[\tau]]\big|_{\theta=\bar{\theta}} \sim \mathbb{E}_{\tau\sim\mathbb{P}_{\bar{\theta}}}\left[\int \left(\frac{\partial b_\theta}{\partial\theta}\bigg|_{\theta=\bar{\theta}}\right)^{\mathsf{T}} F_t[\tau]\,\mathrm{d}t\right].$$

假如我们要构造 Flow Matching 风味损失：

$$\mathcal{L}(\theta;\bar{\theta}) = \frac{1}{2}\mathbb{E}_{\mathbb{P}_{\bar{\theta}}}\left[\int_0^1 \|b_\theta(x_t,t) - \hat{b}_t\|^2\,\mathrm{d}t\right].$$

我们希望用 $\bar{\theta}$ 相关的东西构造这里的目标速度 $\hat{b}_t$，求导：

$$-\nabla_\theta \mathcal{L} = \mathbb{E}_{\mathbb{P}_{\bar{\theta}}}\left[\int_0^1 (\partial_\theta b_\theta)^{\mathsf{T}}\left(\hat{b}_t - b_\theta(x_t,t)\right)\mathrm{d}t\right].$$

只需取：

$$\begin{aligned}&\hat{b}_t - b_{\bar{\theta}}(x_t,t) = F_t[\tau] \\ \Longrightarrow\qquad &\hat{b}_t = b_{\bar{\theta}}(x_t,t) + \frac{R[\tau]}{\sigma^2}\left(\dot{x}_t - b_{\bar{\theta}}(x_t,t)\right).\end{aligned}$$

得到损失：

$$\mathcal{L}(\theta,\bar{\theta}) = \frac{1}{2}\mathbb{E}_{\mathbb{P}_{\bar{\theta}}}\mathbb{E}_t\left\|b_\theta(x_t,t) - \mathrm{sg}\left[b_{\bar{\theta}}(x_t,t) + \frac{R[\tau]}{\sigma^2}\left(\dot{x}_t - b_{\bar{\theta}}(x_t,t)\right)\right]\right\|^2.$$

2.2.2 Flow Matching 的影子价格

考虑一个更简单的 Setting，我们使用流模型生成数据，具体而言，生成的过程是：

$$\frac{\mathrm{d}X_t^\theta}{\mathrm{d}t} = v_\theta(X_t^\theta,t), \qquad X_0^\theta = \xi, \qquad \xi\sim\rho_0.$$

我们希望直接得到终点奖励 $R(X_1^\theta)$ 对流模型参数的导数 θ，取定从起点采样的噪声 ξ：

$$\nabla_\theta R(X_1^\theta) = \left(\frac{\partial X_1^\theta}{\partial\theta}\right)^{\mathsf{T}}\nabla_x R(X_1^\theta) =: M_1^{\mathsf{T}} a_1.$$

其中

$$M_t = \frac{\partial X_t^\theta}{\partial\theta}, \qquad a_t = \nabla_x\left[R\left(\Phi_{t\to1}^\theta(x)\right)\right]\big|_{x=X_t^\theta}.$$

分别为 t 时刻粒子位置 X_t^θ 对参数 θ 的响应，以及终点奖励对 t 时刻粒子位置 X_t^θ 的响应。我们希望计算 $M_1^{\mathsf{T}}a_1$，但是其实知道 $M_0 = 0$，因此我们思考能不能计算：

$$\frac{\mathrm{d}(M_t^{\mathsf{T}}a_t)}{\mathrm{d}t} = \dot{M}_t^{\mathsf{T}}a_t + M_t^{\mathsf{T}}\dot{a}_t.$$

分别计算这两部分：

$$\begin{aligned}\frac{\mathrm{d}M_t}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial X_t^\theta}{\partial\theta}\right)\\ &= \frac{\partial}{\partial\theta}\frac{\mathrm{d}X_t^\theta}{\mathrm{d}t}\\ &= \frac{\partial}{\partial\theta}v_\theta(X_t^\theta,t)\\ &= \left.\frac{\partial v_\theta(x,t)}{\partial\theta}\right|_{x=X_t^\theta} + \left.\frac{\partial v_\theta(x,t)}{\partial x}\right|_{x=X_t^\theta}\frac{\partial X_t^\theta}{\partial\theta}\\ &=: B_t + A_t M_t.\end{aligned}$$

其中

$$B_t = \left.\frac{\partial v_\theta(x,t)}{\partial\theta}\right|_{x=X_t^\theta}, \qquad A_t = \left.\frac{\partial v_\theta(x,t)}{\partial x}\right|_{x=X_t^\theta}.$$

对于 a_t 的计算，需要使用前面在确定性最优控制中推出协态变量 $p(t)$ 时相同的技巧（在时间上展开一步）：

$$\begin{aligned}a_t &= \nabla_x\left[R\big(\Phi_{t\to1}^\theta(x)\big)\right]\big|_{x=X_t^\theta}\\ &=: \frac{\partial R(X_1^\theta)}{\partial X_t^\theta}\\ &= \left(\frac{\partial X_{t+\Delta t}^\theta}{\partial X_t^\theta}\right)^{\mathsf{T}}\frac{\partial R(X_1^\theta)}{\partial X_{t+\Delta t}^\theta}\\ &\simeq \left(\frac{\partial X_t^\theta + v_\theta(X_t^\theta,t)\Delta t}{\partial X_t^\theta}\right)^{\mathsf{T}}\frac{\partial R(X_1^\theta)}{\partial X_{t+\Delta t}^\theta}\\ &= (I + A_t\Delta t)^{\mathsf{T}}a_{t+\Delta t}.\end{aligned}$$

进而推出：

$$a_{t+\Delta t} - a_t \sim -A_t^{\mathsf{T}}a_{t+\Delta t}\Delta t.$$

两侧同除 Δt 得到：

$$\dot a_t = -A_t^{\mathsf{T}}a_t.$$

把以上算出来的两个时间导数合起来：

$$\begin{aligned}\frac{\mathrm{d}(M_t^{\mathsf{T}}a_t)}{\mathrm{d}t} &= \dot M_t^{\mathsf{T}}a_t + M_t^{\mathsf{T}}\dot a_t\\ &= (B_t + A_tM_t)^{\mathsf{T}}a_t - M_t^{\mathsf{T}}A_t^{\mathsf{T}}a_t\\ &= B_t^{\mathsf{T}}a_t.\end{aligned}$$

进而推出：

$$M_1^{\mathsf{T}}a_1 = \int_0^1\left(\left.\frac{\partial v_\theta(x,t)}{\partial\theta}\right|_{x=X_t^\theta}\right)^{\mathsf{T}}a_t\,\mathrm{d}t.$$

其中影子价格 a_t 按照 Adjoint ODE 演化：

$$\dot a_t = -A_t^{\mathsf{T}}a_t, \qquad a_1 = \nabla_x R(x)|_{x=X_1^\theta}.$$

实际训练时，需再对初始数据 ξ 期望：

$$\nabla_\theta\mathbb{E}_{\tau\sim\mathbb{P}_\theta}[R[\tau]] \sim \mathbb{E}_{\xi\sim\rho_0}\mathbb{E}_t\left[\left(\left.\frac{\partial v_\theta(x,t)}{\partial\theta}\right|_{x=X_t^\theta}\right)^{\mathsf{T}}a_t\right].$$

2.2.3 Flow Matching 的终端分布倾斜

微调 Flow Model 时，我们的奖励通常只在终点结算，因此一种既能提高奖励，又使得数据不大幅偏离预训练模型的终点分布的方式是直接对终点分布进行倾斜：设原模型生成的终点分布是 $Y \sim q_{\text{old}}$，通常选择非负函数 $w(y)$，定义倾斜分布：

$$q_\star(y) = \frac{w(y)}{Z} q_{\text{old}}(y).$$

例如，如果奖励非负，可以选择 $w(y) = R(y)$；若奖励是实数，可以选择 $w(y) = \exp\big(R(y)/\tau\big)$。为了使得后训练完成的速度场能够将初始分布推前到倾斜分布 $q_\star$，应当采取的损失是：

$$\mathcal{L}_w = \frac{1}{2}\mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon,\,t}\left[w(Y)\|v_\theta(X_t,t) - U\|^2\right], \qquad X_t = (1-t)\epsilon + tY, \quad U = Y - \epsilon.$$

我们尝试逐点求解回归出的速度场。取定 $X_t := x$，时间 t，损失：

$$_w(v;x,t) = \frac{1}{2}\mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon}\left[w(Y)\|v - U\|^2 \,\middle|\, X_t = x\right].$$

由于现在的问题是一个逐点优化问题，直接对 v 求导：

$$\begin{aligned}
\nabla_v w(v;x,t) &= \frac{1}{2}\nabla_v \mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon}\left[w(Y)\|v - U\|^2 \mid X_t = x\right] \\
&= \mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon}\left[w(Y)(v - U) \mid X_t = x\right] \\
&= v\,\mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon}\left[w(Y) \mid X_t = x\right] - \mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon}\left[Uw(Y) \mid X_t = x\right].
\end{aligned}$$

进而推出：

$$v_\star(x,t) = \frac{\mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon}\left[Uw(Y) \mid X_t = x\right]}{\mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon}\left[w(Y) \mid X_t = x\right]}.$$

可以考虑将旧模型的速度场 $v_{\text{old}}(X_t,t)$ 放入损失，记：

$$d_\theta(X_t,t) = v_\theta(X_t,t) - v_{\text{old}}(X_t,t), \qquad h = U - v_{\text{old}}(X_t,t).$$

损失将被分解为：

$$\begin{aligned}
\mathcal{L}_w &= \frac{1}{2}\mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon,\,t}\left[w(Y)\|v_\theta(X_t,t) - U\|^2\right] \\
&= \frac{1}{2}\mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon,\,t}\left[w(Y)\|d_\theta - h\|^2\right] \\
&\sim \frac{1}{2}\mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon,\,t}\left[w(Y)\|d_\theta\|^2\right] - \mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon,\,t}\left[w(Y)d_\theta^{\mathsf{T}}h\right].
\end{aligned}$$

这里的 d_θ 是速度的改变量。仍然取定 $X_t = x$ 和时间 t，进行逐点优化得到速度场的最优改变量：

$$d_\star(x,t) = \frac{\mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon}\left[w(Y)h \,|\, X_t = x\right]}{\mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon}\left[w(Y) \mid X_t = x\right]}.$$

这里的分子上计算了旧模型预测的后验速度和条件速度的偏差 $h = U - v_{\text{old}}$ 与终点奖励加权 $w(Y)$ 的相关性。

很多 Heuristic 的方法都有类似的设计（损失中的一项用于惩罚后训练的模型不能离开初始模型太远，另一项作为引导），例如 Diffusion NFT 使用的损失：

$$\begin{aligned}
\mathcal{L}_{\text{NFT}} &= \frac{1}{2}\mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon,\,t}\left[r\|\beta d_\theta - h\|^2 + (1-r)\|-\beta d_\theta - h\|^2\right] \\
&= \mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon,\,t}\left[\frac{\beta^2}{2}\|d_\theta\|^2 - \beta(2r-1)d_\theta^{\mathsf{T}}h + \frac{1}{2}\|h\|^2\right] \\
&\sim \mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon,\,t}\left[\frac{\beta^2}{2}\|d_\theta\|^2 - \beta(2r-1)d_\theta^{\mathsf{T}}h\right].
\end{aligned}$$

通过取定 $X_t = x$ 和时间 t，逐点求导得到：

$$d_\star^{\text{NFT}}(x,t) = \frac{1}{\beta}\mathbb{E}_{Y\sim q_{\text{old}},\,\epsilon}\left[(2r-1)h \,|\, X_t = x\right].$$