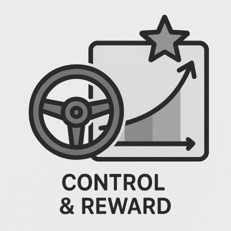
Reward Alignment

Are you doing $\texttt{Rectify}(\texttt{Align}(\{X_t\}))$ or $\texttt{Align}(\texttt{Rectify}(\{X_t\}))$?

Reward Alignment

- Given a reward function $r(x)$, how to modify the flow/diffusion model to favor r?
- Applications:
  - Classifier free/classifier guidance
  - Reward alignment
  - Control, editing
  - Sampling from density + data
  - Constrained domains
  - Reinforcement learning



Tilted Distributions and Processes

- Let $r(x)$ be a non-negative reward
- $X_1^r \sim \pi^r$ is the r-**tilted distribution** of $X_1 \sim \pi_r$ if

$$\pi_1^r(x) = \frac{\pi_1(x) \times r(x)}{A_r}, \qquad A_r = \mathbb{E}_{\pi_1}[r(x)].$$

- $\{X_t^r\} \sim \mathrm{P}^r$ is the r-**tilted process** of $\{X_t\} \sim \mathrm{P}^r$ if

$$\mathrm{P}^r(\{x_t\}) = \frac{\mathrm{P}(\{x_t\}) \times r(x_1)}{A_r}.$$

  Reweight paths based on the reward $r(x_1)$ at the end point.

- Yields a variational representation:

$$\min_{\mathrm{Q}} \mathrm{KL}(\mathrm{Q} \mid\mid P) - \mathbb{E}_Q[\log r(X_1)].$$

Two Approaches

Tilt the interpolation process, then derive rectified flow

$$\texttt{Rectify}(\texttt{Tilt}(\{X_t\}))$$

Leveraging special properties of rectified flow

Examples: DPS [CKM+22], RewardSDS [CYB25], Importance Weighted Score Matching [WZC+25], ORW-CFM-W2 [FSC+25]

Tilt or steer the learned ODE/SDE.

$$\texttt{Tilt}(\texttt{Rectify}(\{X_t\}))$$

Yields generic control and RL problems of ODE/SDE.

Examples: Readout Guidance [LDW+24], Universal Guidance [BCS+23], Autoguidance [KAK+24], Diffusion Policy Optimization (DRWR) [RLA+24]

Nice recent reviews on RL/alignment of diffusion: [UZBL24], [UZW+25].

Tilted Rectified Flow: General Case

- Define $\{Z_t^r\}$ be the rectified flow induced from the titled interpolation process:

$$\{Z_t^r\} = \texttt{Rectify}(\{X_t^r\}), \qquad \{X_t^r\} = r\text{-}\texttt{tilt}(X).$$

- **Velocity field of v_t^r:** reward-weighted average of the original slope $\dot{X}_t$.

$$v_t^r(x) = \frac{\mathbb{E}[r(X_1)\dot{X}_t \mid X_t = x]}{\mathbb{E}[r(X_1) \mid X_t = x]}$$

- Can be estimated by reward-weighted mean square:

$$\min_{\mu} \mathbb{E}_{(X_0, X_1, t)}[r(X_1)\|\dot{X}_t - v(X_t)\|^2]$$

- **Initial Distribution:** For an independent coupling (X_0, X_1), the initial distribution is unchanged: $\rho_0^r = \rho_0$.

Titled Rectified Flow: Score Functions

- **Marginal Distribution $X_t^r \sim \rho_t^r$:**

$$\rho_t^r(x) = \rho_t(x)\frac{\mathbb{E}[r(X_1)|X_t = x]}{\mathbb{E}[r(X_1)]}$$

- **Score Function:**

$$\nabla \log \rho_t^r(x) = \nabla \log \rho_t(x) + \nabla \log \mathbb{E}[r(X_1) \mid X_t = x]$$

- For independent Gaussian noise X_0, we have affine relation:

$$v^r(x) - v_t(x) = \frac{1-t}{t}(\nabla \log \rho_t^r(x) - \nabla \log \rho_t(x)).$$

This yields

$$v_t^r(x) = v_t(x) + \frac{1-t}{t}\nabla_x \log \mathbb{E}\left[r(X_1) \mid X_t = x\right].$$

Tilted Rectified Flow: Independent Gaussian Noise

When $X_0|X_1 \sim \texttt{Normal}(0, I)$, velocity of tilted RF is

$$v_t^r(x) = v_t(x) + \frac{1-t}{t}\nabla_x \log \mathbb{E}\left[r(X_1) \mid X_t = x\right].$$

- Using Taylor Approximation ($\mathbb{E}\left[r(X_1)|X_t\right]) \approx r(\mathbb{E}\left[X_1|X_t\right])$:

$$v_t^r(x) = v_t(x) + \frac{1-t}{t}\nabla_x \log r(\mu_{1,t}(x)).$$

where $\mu_{1,t}(x)$ is the expected prediction of the target X_1 :

$$\mu_{1,t}(x) = \mathbb{E}\left[X_1 \mid X_t = x\right] = x + (1-t)v_t(x).$$

- Examples: Classifier Guidance [DN21], DPS [CKM+22], RF-inversion [RCR+24], etc.

Reward Tilting: Variational Approaches

Variational Approach: π_1^r is the solution of

$$\min_{\varrho_1} \mathrm{KL}\left(\varrho_1 \,\|\, \rho_1\right) - \mathbb{E}_{\varrho_1}[r(X_1)]$$

- With independent Gaussian noise, using KL-velocity formula:

$$\min_{v} \int \frac{t}{1-t} \mathbb{E}\left[\left\|\mu_t(X_t^r) - v_t(X_t^r)\right\|^2 - \log r(X_1^r)\right].$$

- Same as Variational Score Distillation but with a reward. Examples: Diff-Instruct++ [Luo24], Reward-Instruct [LHL$^+$25], etc.

Takeaways

$\texttt{Rectify}(\texttt{Tilt}(\{X_t\}))$

Leveraging special properties of rectified flow: General, score function, variational forms.

$\texttt{Tilt}(\texttt{Rectify}(\{X_t\}))$

General control/RL tools: policy gradient, preference optimization, Q-learning, optimal control, Doob's h-transform, etc.