Reward Tilted Rectified Flow via Path Following

Qiang Liu

Let $X_0 \sim \rho_0$ be a base distribution and $X_1 \sim \pi_1^\alpha$ be a target distribution depending on a scalar parameter α. Define the linear interpolation $X_t = (1-t)X_0 + tX_1$, and the rectified flow velocity field

$$v_t^\alpha(x) = \mathbb{E}_{(X_0,X_1)\sim\gamma^\alpha}[X_1 - X_0 \mid X_t = x], \qquad t \in [0,1),$$

where $(X_0, X_1) \sim \gamma^\alpha$ is a coupling of π_0 and π_1, typically the independent coupling $\gamma(x_0, x_1) = \pi_0(x) \times \pi_1^\alpha(x)$.

Assume (X_0, X_1) has a joint density $\gamma^\alpha(x_0, x_1)$ that is differentiable in α. It could be of interest to compute the derivative $\partial_\alpha v_t^\alpha(x)$. Knowing this derivative allows for smooth path-following updates, enabling one to transition from one rectified flow to another in a continuous manner.

Theorem 0.1 (Derivative of the RF velocity via conditional score)**.** *Assume* $\log \gamma^\alpha$ *is differentiable in* α *and differentiation can be exchanged with integration/conditioning. Then for any* $t \in [0,1)$ *and* x*,*

$$\partial_\alpha v_t^\alpha(x) = \mathrm{Cov}_{\gamma^\alpha}\Big(X_1 - X_0,\ \partial_\alpha \log \gamma^\alpha(X_0, X_1) \,\Big|\, X_t = x\Big).$$

In particular, in the case of independent coupling $\gamma^\alpha(x_0, x_1) = \pi_0(x_0)\pi_1^\alpha(x_1)$*, we have*

$$\partial_\alpha v_t^\alpha(x) = \mathrm{Cov}_{\gamma^\alpha}\Big(X_1 - X_0,\ \partial_\alpha \log \gamma^\alpha(X_0, X_1) \,\Big|\, X_t = x\Big).$$

Proof. Fix $t \in [0,1)$ and x. Note

$$v_t^\alpha(x) = \frac{m_t^\alpha(x) - x}{1-t}, \qquad m_t^\alpha(x) = \mathbb{E}[X_1 \mid X_t = x].$$

So it suffices to differentiate $m_t^\alpha(x)$.

Since $X_t = (1-t)X_0 + tX_1$, conditioning on $X_t = x$ and $X_1 = x_1$ pins down

$$X_0 = \frac{x - tx_1}{1-t}.$$

A convenient way to write the conditional density of X_1 given $X_t = x$ is

$$p_t^\alpha(x_1 \mid x) = \frac{\gamma^\alpha\Big(\frac{x-tx_1}{1-t},\, x_1\Big) \cdot \frac{1}{1-t}}{\rho_t^\alpha(x)}, \qquad \rho_t^\alpha(x) = \int \gamma^\alpha\Big(\frac{x - tx_1}{1-t},\, x_1\Big) \cdot \frac{1}{1-t}\, dx_1.$$

Note that $m_t^\alpha(x) = \int x_1\, p_t^\alpha(x_1 \mid x)\, dx_1$. Differentiate under the integral:

$$\partial_\alpha m_t^\alpha(x) = \int x_1\, \partial_\alpha p_t^\alpha(x_1 \mid x)\, dx_1 = \int x_1\, p_t^\alpha(x_1 \mid x)\, \partial_\alpha \log p_t^\alpha(x_1 \mid x)\, dx_1 = \mathbb{E}\big[X_1\, \partial_\alpha \log p_t^\alpha(X_1 \mid x) \,\big|\, X_t = x\big].$$

Now expand the conditional score. Up to an α-independent constant,

$$\log p_t^\alpha(x_1 \mid x) = \log \gamma^\alpha\Big(\frac{x - tx_1}{1-t},\, x_1\Big) - \log \rho_t^\alpha(x) + \mathrm{const}(t).$$

Hence

$$\partial_\alpha \log p_t^\alpha(x_1 \mid x) = s^\alpha\Big(\frac{x - tx_1}{1-t},\, x_1\Big) - \partial_\alpha \log \rho_t^\alpha(x).$$

Using the Fisher identity for the marginal $\rho_t^\alpha(x)$,

$$\partial_\alpha \log \rho_t^\alpha(x) = \mathbb{E}\Big[s^\alpha(X_0, X_1) \,\Big|\, X_t = x\Big].$$

So

$$\partial_\alpha \log p_t^\alpha(X_1 \mid x) = s^\alpha(X_0, X_1) - \mathbb{E}\Big[s^\alpha(X_0, X_1) \,\Big|\, X_t = x\Big].$$

Plugging back,

$$\partial_\alpha m_t^\alpha(x) = \mathbb{E}\Big[X_1\Big(s^\alpha(X_0, X_1) - \mathbb{E}[s^\alpha(X_0, X_1) \mid X_t = x]\Big) \,\Big|\, X_t = x\Big] = \mathrm{Cov}\Big(X_1,\ s^\alpha(X_0, X_1) \,\Big|\, X_t = x\Big).$$

Finally,

$$\partial_\alpha v_t^\alpha(x) = \frac{1}{1-t}\,\partial_\alpha m_t^\alpha(x).$$

If instead we view $v_t^\alpha(x) = \mathbb{E}[X_1 - X_0 \mid X_t = x]$, the same calculation gives directly

$$\partial_\alpha v_t^\alpha(x) = \mathrm{Cov}_{\gamma^\alpha}\Big(X_1 - X_0,\ s^\alpha(X_0, X_1) \,\Big|\, X_t = x\Big).$$

□

[I wanted to present the following as an application instance of the formula above. show case reward titlting as an application, and say that we can apply iterative "gradient descent" update following a grid on α; taighting the presentation that I had below]

Reward Tilting Now, assume we are given a reward function $r(x)$, and we are interested in sampling from the reward-tiled target distribution:

$$\pi_1^\alpha(x) = \frac{\pi_1(x)\exp(\alpha r(x))}{Z_\alpha}, \qquad Z_\alpha = \int \pi_1(x)\exp(\alpha r(x))\,\mathrm{d}x,$$

and the RF velocity field associated with π_1^α is

$$v_t^\alpha(x) = \mathbb{E}_{(X_0, X_1)\sim \pi_0 \times \pi_1^\alpha}[X_1 - X_0 \mid X_t = x].$$

As we change α from 0 to 1, we gradually change the velocity field from v_t^0 to v_t^1.

Assume the independent coupling $\gamma^\alpha(x_0, x_1) = \rho_0(x_0)\pi_1^\alpha(x_1)$. Then $\partial_\alpha \log \gamma^\alpha(x_0, x_1) = \partial_\alpha \log \pi_1^\alpha(x_1) = r(x_1) - \partial_\alpha \log Z_\alpha$, so the constant term drops out in the conditional covariance and

$$\partial_\alpha v_t^\alpha(x) = \mathrm{Cov}_\alpha\big(X_1 - X_0,\ r(X_1) \,\big|\, X_t = x\big).$$

With this, we can gradually update v_t^0 towards v_t^1 by following a grid on α:

$$v_t^{\alpha_{k+1}} = v_t^{\alpha_k} + \partial_\alpha v_t^{\alpha_k} \cdot (\alpha_{k+1} - \alpha_k).$$

1 Connection to Diffusion NFT

DiffusionNFT [1] can be viewed as learning a reward-tilted update around a pretrained velocity field. Assume we have a reward $r(x) \in [0,1]$. Let v_t^p be a pretrained velocity field trained on pairs $(X_0, X_1) \sim \rho_0 \times \pi_1$, with the standard linear interpolation $X_t = (1-t)X_0 + tX_1$.

DiffusionNFT introduces a learnable field $\mu_t(\cdot)$ and forms two symmetric perturbations around v_t^p:

$$v_t^+(x) := v_t^p(x) + \beta\big(\mu_t(x) - v_t^p(x)\big), \qquad v_t^-(x) := v_t^p(x) - \beta\big(\mu_t(x) - v_t^p(x)\big),$$

where $\beta > 0$ controls how far we move away from the pretrained model. Then it fits μ by a reward-weighted regression:

$$L(\mu) := \mathbb{E}\Big[r(X_1)\left\|v_t^+(X_t) - (X_1 - X_0)\right\|^2 + (1 - r(X_1))\left\|v_t^-(X_t) - (X_1 - X_0)\right\|^2\Big].$$

Intuitively, high-reward samples pull the policy toward v_t^+, while low-reward samples pull it toward v_t^-. After training, sampling uses the updated policy v_t^+.

Let v_t^* be the minimum of $L(\mu)$, and $v_t^{*\pm}$ the corresponding perturbations. As it turns out, v_t^* can be viewed as approximating the RF vector field of $\pi^{2/\beta}(x) = \pi_1(x)\exp(2r(x)/\beta)/Z$ with a linearization:

$$v_t^*(x) = v_t^0(x) + \frac{2}{\beta}\partial_\alpha v_t^\alpha(x)\ |_{\alpha=0},$$

and $v_t^{*+}(x)$ approximates $\pi^2(x) = \pi_1(x)\exp(2r(x))/Z$ via

$$v_t^{*+}(x) = v_t^0(x) + 2\partial_\alpha v_t^\alpha(x)\ |_{\alpha=0}.$$

Remark We can rewrite Diffusion NFT as a variant of velocity field extrapolation:

$$v_t^{*+}(x) = v_t^0(x) + 2\hat{m}_t(x)(\hat{v}_t(x) - v_t^0(x)),$$

where

$$\hat{v}_t(x) = \frac{\mathbb{E}[r(X_1)(X_1 - X_0) \mid X_t = x]}{\mathbb{E}[r(X_1) \mid X_t = x]}, \qquad \hat{m}_t(x) = \mathbb{E}[r(X_1) \mid X_t = x].$$

Remark Compare this with the negative RF for signed measure $(2r(x) - 1)\pi_1(x)$, which is

$$\begin{aligned} v_t(x) &= \frac{\mathbb{E}[(2r(x) - 1)(X_1 - X_0) \mid X_t = x]}{\mathbb{E}[(2r(x) - 1) \mid X_t = x]} \\ &= \frac{2\hat{m}_t(x)\hat{v}_t(x) - v_t^0(x)}{2\hat{m}_t(x) - 1} \\ &= v_t^0(x) + \lambda_t(x)(\hat{v}_t(x) - v_t^0(x)), \end{aligned}$$

where

$$\lambda_t(x) = \frac{2\hat{m}_t(x)}{2\hat{m}_t(x) - 1}.$$

Therefore, Diffusion NFT can be viewed as replacing $\lambda_t(x) = \frac{2\hat{m}_t(x)}{2\hat{m}_t(x)-1}$ with $\hat{m}_t(x)$.

Theorem 1.1. *Following the set up above, we have*

$$v_t^{NFT}(x) = v_t^0(x) + 2\,\partial_\alpha v_t^\alpha(x)\big|_{\alpha=0},$$

where by Theorem 0.1 (independent coupling) we have

$$\partial_\alpha v_t^\alpha(x)\ |_{\alpha=0} = \mathrm{Cov}_{\gamma^0}\big(X_1 - X_0,\ r(X_1)\,\big|\,X_t = x\big).$$

Proof. Write $y := X_1 - X_0$. For fixed (t, x), minimizing $L(\mu)$ over functions $\mu_t(\cdot)$ is pointwise in x and amounts to minimizing the conditional objective

$$\mathbb{E}\Big[r(X_1)\|v_t^+(x) - y\|^2 + (1 - r(X_1))\|v_t^-(x) - y\|^2\,\Big|\,X_t = x\Big]$$

with $v_t^\pm(x) = v_t^p(x) \pm \beta(\mu_t(x) - v_t^p(x))$. This is a strictly convex quadratic in $\mu_t(x)$, and the first-order condition yields

$$\begin{aligned}\mu_t^\star(x) &= \frac{1}{\beta}\,\mathbb{E}\Big[(2r(X_1) - 1)y + (1 + \beta - 2r(X_1))\,v_t^p(x)\,\Big|\,X_t = x\Big]\\ &= v_t^p(x) + \frac{1}{\beta}\,\mathbb{E}\Big[(2r(X_1) - 1)(y - v_t^p(x))\,\Big|\,X_t = x\Big].\end{aligned}$$

Assume $v_t^p(x) = \mathbb{E}[y \mid X_t = x]$, then the second term above equals the X_t-conditioned covariance between $2r(x)$ and y:

$$\mu_t^\star(x) = v_t^p(x) + \frac{2}{\beta}\mathrm{cov}_{\gamma^0}(y, 2r(X_1) \mid X_t = x).$$

Plugging into $v_t^{*+}(x) = v_t^p(x) + \beta(\mu_t^*(x) - v_t^p(x))$ gives

$$v_t^{*+}(x) = v_t^p(x) + 2\mathrm{cov}_{\gamma^0}(y, r(X_1) \mid X_t = x).$$

□

Lemma 1.2. *Let $\mu^\pm = \mu \pm \beta(\mu - v)$, and*

$$\ell(\mu) = r\left\|\mu^+ - y\right\|^2 + (1 - r)\left\|\mu^- - y\right\|^2,$$

where $\mu, v, y \in \mathbb{R}^d$, $r \in \mathbb{R}$. Then, we can rewrite $\ell(\mu)$ into

$$\ell(\mu) = \|\beta\mu - \beta v - (2r - 1)(y - v)\|^2 + const,$$

where const is a term that does not depend on μ.

Hence, the minimum of $\ell(\mu)$ is achieved at

$$\mu^* = v + \frac{1}{\beta}(2r - 1)(y - v).$$

Proof. Expand the loss:

$$\begin{aligned}\ell(\mu) &= r\left\|\mu^+ - y\right\|^2 + (1 - r)\left\|\mu^- - y\right\|^2\\ &= r\left\|(1 - \beta)v + \beta\mu - y\right\|^2 + (1 - r)\left\|(1 + \beta)v - \beta\mu - y\right\|^2\\ &= \beta^2\mu^2 + 2\beta\mu(r(1 - \beta)v - ry + (1 - r)y - (1 - r)(1 + \beta)v) + const\\ &= \beta^2\mu^2 + 2\beta\mu(-\beta v + (1 - 2r)(y - v)) + const\\ &= \|\beta\mu - \beta v - (2r - 1)(y - v)\|^2 + const.\end{aligned}$$

Hence, the minimum is achieved at

$$\mu^* = \frac{1}{\beta}(\beta v - (1-2r)(y-v)) = v + \frac{1}{\beta}(2r-1)(y-v).$$

□

Discussion However, this still does not fully explain *why* DiffusionNFT is a good choice.

- We do not necessarily need the Taylor approximation. In principle, we can try to estimate the velocity field of a reward-tilted distribution directly via reward-weighted least squares.

  Let $\epsilon = 1/N$ be a small number. 1) Given the current ODE, draw samples $X_1^{(i)}$ from it. Update the ODE by fitting a weighted LS:

  $$\min_v \sum_i \exp(\epsilon r(X_1^{(i)}))\mathbb{E}\left[\|v_t(X_t^{(i)}) - (X_1^{(i)} - X_0^{(i)})\|^2\right].$$

  Get the new ODE.

  2) Sample from the new ODE and repeat.

  Usually, people would do $\pi_1 \to \pi_1(x)\exp(r(x))$ in a single step (effectively $N=1$). But if r is very sharp, then the importance weights have large variance, and the update can be inaccurate.

  But if we break the path into $N = 1/\epsilon$ pieces, each step only reweights by $\exp(\epsilon r(x))$, which reduces variance.

- The choice of reward is somewhat arbitrary. So exactly sampling from the exponential-tilted distribution is not necessarily the "best" choice; it depends on how we define r. (Different monotone transforms of r induce different tilts and potentially different tradeoffs.)

- Approximating $\exp(2r(x))$ by a first-order Taylor expansion gives weights $\hat{w}(x) = 1 + 2r(x)$. If we use $\hat{w}(X_1)$ to reweight the flow-matching regression, the resulting conditional minimizer is the ratio

  $$\hat{v}_t(x) = \frac{v_t^0(x) + 2\mathbb{E}[r(X_1)(X_1 - X_0)|X_t = x]}{1 + \mathbb{E}[2r(X_1)|X_t = x]}.$$

  Equivalently, $\hat{v}_t$ is the minimizer of $\mathbb{E}[\hat{w}(X_1)\|(X_1 - X_0) - v_t(X_t)\|^2]$ with $\hat{w} = 1 + 2r$. It can also be written as

  $$\hat{v}_t(x) = v_t^0(x) + \frac{1}{1 + m_t(x)}\Big(2\,\mathbb{E}[r(X_1)(X_1 - X_0) \mid X_t = x] - m_t(x)\,v_t^0(x)\Big),$$

  where

  $$m_t(x) = \mathbb{E}[2r(X_1)|X_t = x].$$

  If we further linearize this ratio-type estimator, we recover a DiffusionNFT-style first-order update. But this raises the key question: why is this particular linearization (and its implicit stepsize) a good choice in practice?

References

[1] Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., Zhu, J., and Liu, M.-Y. (2025). Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint.*