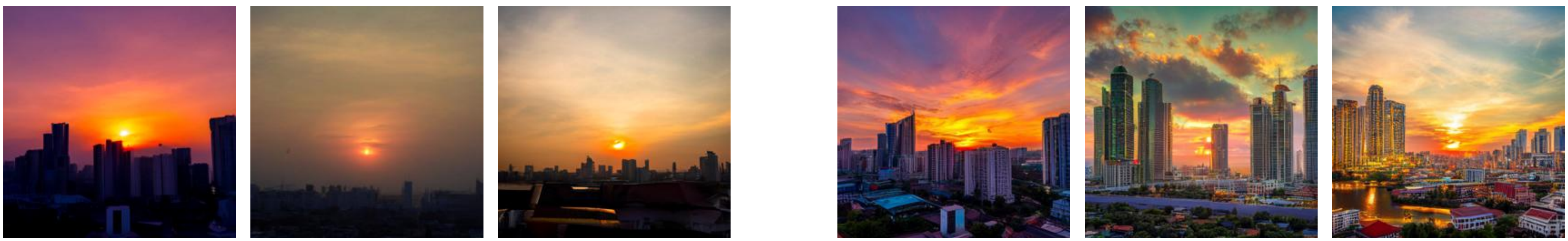
Reward fine-tuning for flow and diffusion models

Carles Domingo-Enrich,
Nov. 19 2024

Main reference: *Adjoint Matching: Fine-tuning Flow and Diffusion Generative Models with Memoryless Stochastic Optimal Control*. C. Domingo-Enrich, M. Drozdzal, B. Karrer, R. T. Q. Chen, ICLR 2025. https://arxiv.org/abs/2409.08861

Open questions

- What is the best way to perform RLHF on diffusion and flow models?
- How do we solve stochastic optimal control problems in high dimensions?



“Beautiful colorful sunset midst of building in Bangkok Thailand”

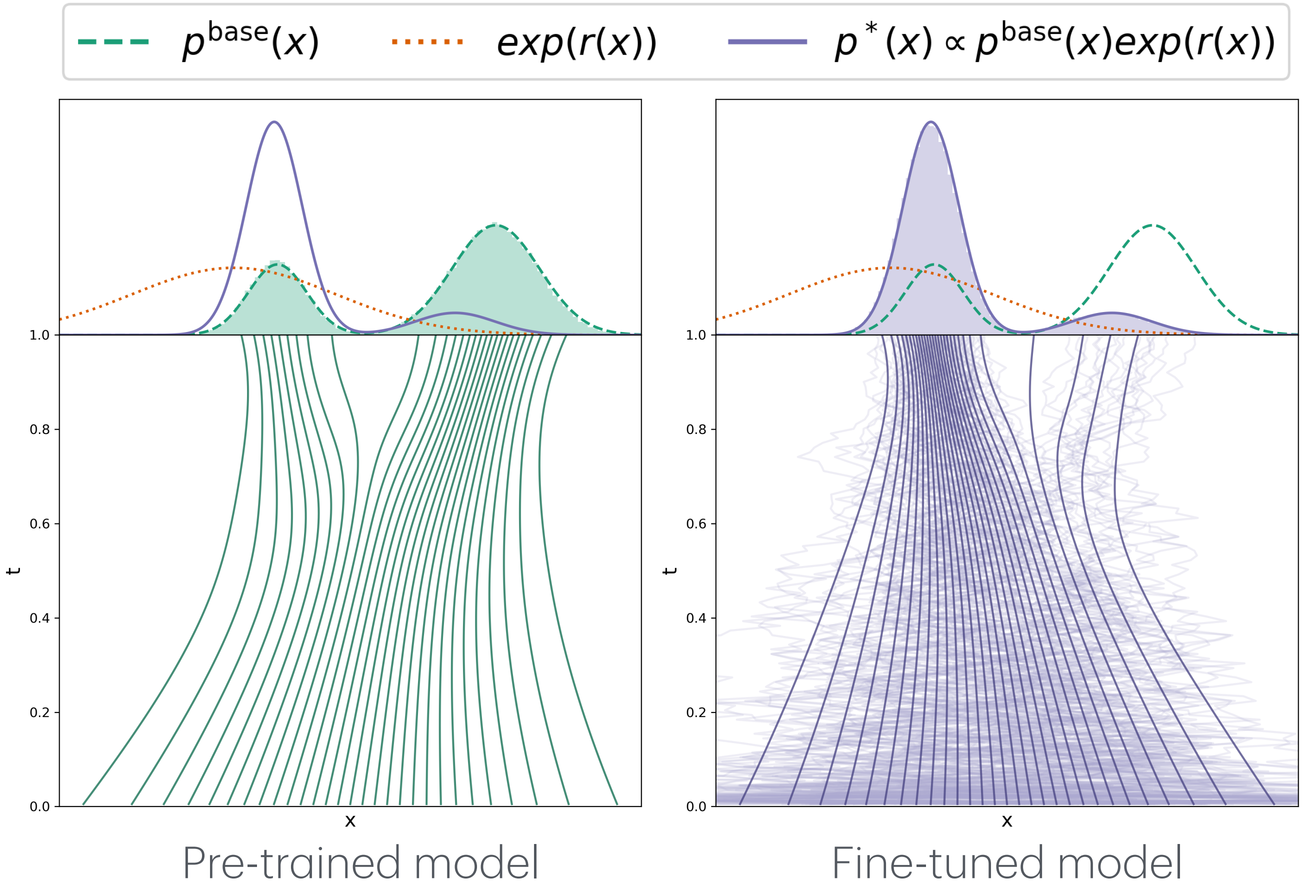
“Beautiful grandma and granddaughter are mixing salad and smiling while cooking in kitchen”

Fine-tuning setup

- Pre-trained diffusion or flow matching model that generates distribution p^{base}

- Reward model $r(x)$, trained using human preferences or encoding conditional sampling:
  $r(x) = \log p(o \mid x)$

- Goal: modify pre-trained diffusion model such that it generates
  $p^*(x) \propto p(x)\exp(r(x))$



Flow Matching and diffusion models: notation

- $\bar{X}_0 \sim p_0 = N(0,I)$ and $\bar{X}_1 \sim p_{\text{data}}$
- Reference flow $\bar{\mathbf{X}} = (\bar{X}_t)_{t\in[0,1]}$, where

$$\bar{X}_t = \beta_t \bar{X}_0 + \alpha_t \bar{X}_1 \,.$$

- Generative flow $\mathbf{X} = (X_t)_{t\in[0,1]}$, where

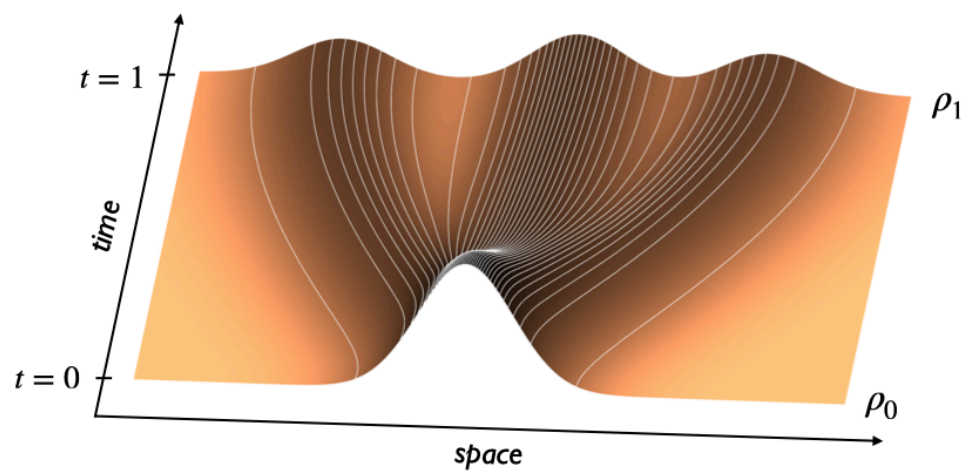
$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = v(X_t, t), \quad X_0 \sim N(0,I)$$

- $\bar{X}_t$ and X_t are equally distributed, for all $t \in [0,1]$

- How do we train the vector field v?

  By matching the derivative of the reference flow (*Flow Matching*):

$$v(\bar{X}_t, t) = \operatorname{argmin}_{\hat{v}} \mathbb{E} \left\| \hat{v}(\bar{X}_t, t) - \frac{\mathrm{d}\bar{X}_t}{\mathrm{d}t} \right\|^2$$



Source: Albergo & Vanden-Eijnden, ICLR 2023

The Flow Matching Generative SDE

- Let p_t be the distribution of the reference flow $\bar{X}_t = \beta_t \bar{X}_0 + \alpha_t \bar{X}_1$. Score function $\mathfrak{s}(x,t) := \nabla \log p_t(x)$.
- FM vector field $v(x,t)$ in terms of $\mathfrak{s}(x,t)$: $\quad v(x,t) = \frac{\dot{\alpha}_t}{\alpha_t} x + \beta_t (\frac{\dot{\alpha}_t}{\alpha_t}\beta_t - \dot{\beta}_t)\mathfrak{s}(x,t)\,.$
- Flow Matching Generative SDE: $\mathrm{d}X_t = \left(v(X_t,t) + \sigma(t)^2 \mathfrak{s}(X_t,t)\right)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}B_t$, or equivalently,

$$\mathrm{d}X_t = \left(v(X_t,t) + \frac{\sigma(t)^2}{2\beta_t(\frac{\dot{\alpha}_t}{\alpha_t} - \dot{\beta}_t)} \left(v(X_t,t) - \frac{\dot{\alpha}_t}{\alpha_t} X_t \right)\right)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}B_t, \qquad X_0 \sim N(0,I)$$

  where the *noise schedule* $\sigma(t)$ is arbitrary. X_t still has the same distribution as $\bar{X}_t$, for all $t \in [0,1]$!

- The FM Generative SDE reads: $\mathrm{d}X_t = b(X_t,t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}B_t, \qquad X_0 \sim \mathcal{N}(0,I),$

  where $b(x,t) = \kappa_t x + \left(\frac{\sigma(t)^2}{2} + \eta_t\right)\mathfrak{s}(x,t), \quad \kappa_t = \frac{\dot{\alpha}_t}{\alpha_t}, \quad \eta_t = \beta_t\left(\frac{\dot{\alpha}_t}{\alpha_t}\beta_t - \dot{\beta}_t\right)$

Stochastic optimal control

Definition of the problem

$$\min_u \overbrace{\mathbb{E}\left[\int_0^1 \left(\underbrace{\frac{1}{2}\|u(X_t^u,t)\|^2}_{\text{control cost}} + \underbrace{f(X_t^u,t)}_{\text{state cost}}\right) dt + \underbrace{g(X_1^u)}_{\text{terminal cost}}\right]}^{\text{control objective}},$$

$$s.t.\ \ \mathrm{d}X_t^u = (\underbrace{b(X_t^u,t)}_{\substack{\text{base}\\ \text{drift}}} + \sigma(t)\underbrace{u(X_t^u,t)}_{\text{control}})\,\mathrm{d}t + \underbrace{\sigma(t)}_{\substack{\text{noise}\\ \text{schedule}}}\ \mathrm{d}B_t, \qquad X_0^u \sim p_0.$$

Value function:

$$V(x,t) := \min_{u\in\mathcal{U}} \mathbb{E}_{X^u}\left[\int_t^1 \left(\frac{1}{2}\|u(X_s^u,s)\|^2 + f(X_s^u,s)\right)\,\mathrm{d}s + g(X_1^u)\ \Big|\ X_t^u = x\right] = -\log \mathbb{E}_X\left[\exp\left(-\int_t^1 f(X_s,s)\mathrm{d}s - g(X_1)\right)\ \Big|\ X_t = x\right].$$

SOC = KL-regularized RL

$\mathbb{P}^u(\,\cdot\,|\,X_0)$, $\mathbb{P}(\,\cdot\,|\,X_0)$: conditional probability measures of the controlled and uncontrolled processes.

From the Girsanov theorem: $D_{\mathrm{KL}}(\mathbb{P}^u(\,\cdot\,|\,X_0)\|\mathbb{P}(\,\cdot\,|\,X_0)) = \mathbb{E}_{X^u}\left[\int_0^1 \frac{1}{2}\|u(X_t^u,t)\|^2 \mathrm{d}t\right].$

$$\Longrightarrow \max_{u\in\mathcal{U}} \mathbb{E}_{X_0\sim p_0}\Bigg[\underbrace{\mathbb{E}_{X^u\sim\mathbb{P}^u(\cdot|X_0)}\left[\int_0^1 -f(X_t^u,t)\mathrm{d}t - g(X_1^u)\right] - D_{\mathrm{KL}}(\mathbb{P}^u(\,\cdot\,|\,X_0)\|\mathbb{P}(\,\cdot\,|\,X_0))}_{\text{Solution}:\quad \mathbb{P}^{u^*}(X|X_0)\propto\mathbb{P}(X|X_0)\exp\left(-\int_0^1 f(X_t,t)\,\mathrm{d}t - g(X_1)\right).}\Bigg],$$

The tilted path measure

$$\max_{u \in \mathcal{U}} \mathbb{E}_{X_0 \sim p_0} \underbrace{\left[\mathbb{E}_{X^u \sim \mathbb{P}^u(\cdot|X_0)} \left[\int_0^1 - f(X_t^u, t) \mathrm{d}t - g(X_1^u) \right] - D_{\mathrm{KL}}(\mathbb{P}^u(\,\cdot\,|X_0) \| \mathbb{P}(\,\cdot\,|X_0)) \right]}_{\text{Solution :} \quad \mathbb{P}^{u^*}(X|X_0) \propto \mathbb{P}(X|X_0) \exp\left(- \int_0^1 f(X_t, t)\,\mathrm{d}t - g(X_1) \right).},$$

Normalization constant of $\mathbb{P}(X|X_0)\exp\big(- \int_0^1 f(X_t, t)\,\mathrm{d}t - g(X_1)\big)$:

$$\mathbb{E}_{X \sim \mathbb{P}(\cdot|X_0)} \left[\exp\left(- \int_0^1 f(X_t, t)\,\mathrm{d}t - g(X_1) \right) \right] = \exp\big(- V(X_0, 0)\big),$$

where $V(x, t) = -\log \mathbb{E}_{X \sim \mathbb{P}} \left[\exp\left(- \int_t^1 f(X_s, s) \mathrm{d}s - g(X_1) \right) \,\middle|\, X_t = x \right]$ is the value function.

$$\Longrightarrow \quad \mathbb{P}^{u^*}(X) = \mathbb{P}(X) \exp\left(- \int_0^1 f(X_t, t)\,\mathrm{d}t - g(X_1) + V(X_0, 0) \right).$$

Memoryless generative processes

Setting $f(x,t)=0$ and $g(x)=-r(x)$, we obtain $\mathbb{P}^{u^*}(X)=\mathbb{P}(X)\exp\big(r(X_1)+V(X_0,0)\big)$.

$$\implies \mathbb{P}^{u^*}(X_0,X_1)=\mathbb{P}(X_0,X_1)\exp\big(r(X_1)+V(X_0,0)\big)\ .$$

Definition: A generative process is **memoryless** if X_0 and X_1 are independent, i.e. $\mathbb{P}(X_0,X_1)=\mathbb{P}(X_0)\mathbb{P}(X_1)$.

Then, $\mathbb{P}^{u^*}(X_1)=\displaystyle\int \mathbb{P}^{u^*}(X_0,X_1)\mathrm{d}X_0=\int \mathbb{P}(X_0)\mathbb{P}(X_1)\exp\big(r(X_1)+V(X_0,0)\big)\mathrm{d}X_0 \propto \mathbb{P}(X_1)\exp\big(r(X_1)\big)$.

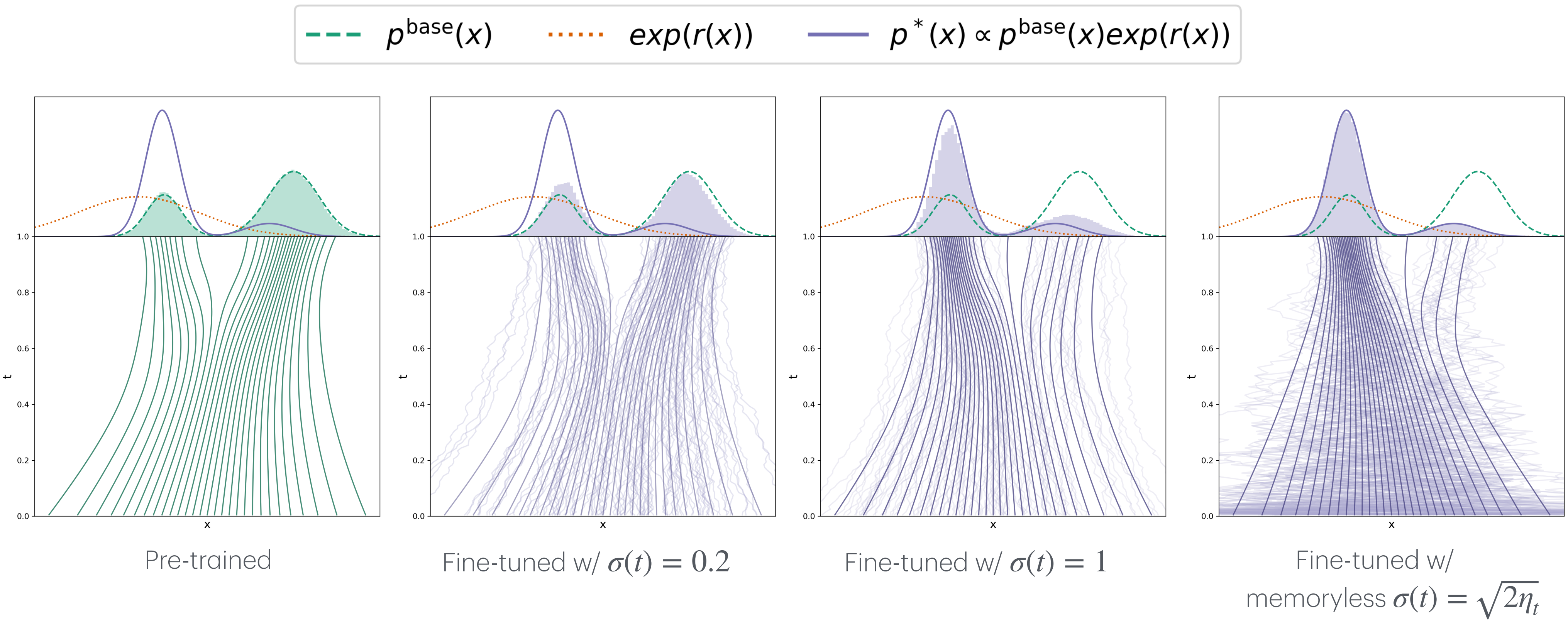
Recall: $\mathrm{d}X_t=b(X_t,t)\,\mathrm{d}t+\sigma(t)\,\mathrm{d}B_t, \qquad X_0\sim\mathcal{N}(0,I),$

where $b(x,t)=\kappa_t x+\Big(\dfrac{\sigma(t)^2}{2}+\eta_t\Big)\mathfrak{s}(x,t), \quad \kappa_t=\dfrac{\dot\alpha_t}{\alpha_t}, \quad \eta_t=\beta_t\Big(\dfrac{\dot\alpha_t}{\alpha_t}\beta_t-\dot\beta_t\Big)$

Proposition: X is memoryless iff

$$\sigma(t)^2=2\eta_t+\chi(t), \text{ where } \chi:[0,1]\to\mathbb{R} \text{ is s.t. } \forall t\in(0,1], \quad \lim_{t'\to 0^+}\alpha_{t'}\exp\Big(-\int_{t'}^{t}\frac{\chi(s)}{2\beta_s^2}\,\mathrm{d}s\Big)=0.$$

Fine-tuning with different noise schedules



Fine-tuning recipe

to sample the fine-tuned model with arbitrary noise schedules

Recall: $\mathrm{d}X_t = b(X_t, t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}B_t, \qquad X_0 \sim \mathcal{N}(0,I),$

where $b(x,t) = \kappa_t x + \left(\frac{\sigma(t)^2}{2} + \eta_t\right)\mathfrak{s}(x,t), \quad \kappa_t = \frac{\dot{\alpha}_t}{\alpha_t}, \quad \eta_t = \beta_t\left(\frac{\dot{\alpha}_t}{\alpha_t}\beta_t - \dot{\beta}_t\right)$

Theorem: In order to allow the use of arbitrary noise schedules and still generate samples according to the tilted distribution, the fine-tuning problem must be done with the memoryless noise schedule $\sigma(t) = \sqrt{2\eta_t}$.

Implicit parameterization of the control

Recall: $b(x,t) = v(X_t,t) + \dfrac{\sigma(t)^2}{2\beta_t(\frac{\dot\alpha_t}{\alpha_t}\beta_t - \dot\beta_t)}\left(v(X_t,t) - \dfrac{\dot\alpha_t}{\alpha_t}X_t\right)$

$$\sigma(t) = \sqrt{2\eta_t} \implies b(x,t) = 2v(x,t) - \frac{\dot\alpha_t}{\alpha_t}x$$

$$\left.\begin{aligned} b(x,t) &= 2v^{\text{base}}(x,t) - \frac{\dot\alpha_t}{\alpha_t}x \\ b(x,t) + \sigma(t)u(x,t) &= 2v^{\text{finetune}}(x,t) - \frac{\dot\alpha_t}{\alpha_t}x \end{aligned}\right\} \implies u(x,t) = \sqrt{\frac{2}{\beta_t(\frac{\dot\alpha_t}{\alpha_t}\beta_t - \dot\beta_t)}}\,(v^{\text{finetune}}(x,t) - v^{\text{base}}(x,t))$$

Methods for SOC: the adjoint method

Discrete adjoint method: $\mathcal{L}(u; X^u) := \int_0^1 \left(\frac{1}{2} \|u(X_t^u, t)\|^2 + f(X_t^u, t) \right) \mathrm{d}t + g(X_1^u)\,.$

Continuous Adjoint method / Basic Adjoint Matching:

$$\mathcal{L}_{\mathrm{Basic-Adj-Match}}(u; X^u) := \frac{1}{2} \int_0^1 \left\| u(X_t, t) + \sigma(t)^\top a(t; X^{\bar{u}}, \bar{u}) \right\|^2 \mathrm{d}t, \qquad \bar{u} = \mathrm{stopgrad}(u),$$

where the adjoint state satisfies the adjoint ODE:

$$\frac{\mathrm{d}}{\mathrm{d}t} a(t; X^u, u) = -\left[\left(\nabla_x (b(X_t^u, t) + \sigma(t) u(X_t^u, t)) \right)^\top a(t; X^u, u) + \nabla_x \left(f(X_t^u, t) + \frac{1}{2} \|u(X_t^u, t)\|^2 \right) \right],$$

$$a(1; X^u, u) = \nabla g(X_1^u)\,.$$

Basic Adjoint Matching

$$\mathcal{L}_{\mathrm{Basic-Adj-Match}}(u; X^u) := \frac{1}{2}\int_0^1 \left\| u(X_t, t) + \sigma(t)^\top a(t; X^{\bar{u}}, \bar{u}) \right\|^2 \mathrm{d}t, \qquad \bar{u} = \mathrm{stopgrad}(u)$$

Interpretation: $a(t; X^u, u) := \nabla_{X_t^u}\left(\int_t^1 \left(\frac{1}{2}\|u(X_{t'}^u, t')\|^2 + f(X_{t'}^u, t') \right) \mathrm{d}t' + g(X_1^u) \right),$

And the optimal control satisfies:

$$u^*(x,t) = -\,\sigma(t)^\top \nabla V(x,t) = -\,\sigma(t)^\top \nabla_x \mathbb{E}\left[\int_t^1 \left(\frac{1}{2}\|u^*(X_s^{u^*}, s)\|^2 + f(X_s^{u^*}, s) \right) \mathrm{d}s + g(X_1^{u^*}) \,\middle|\, X_t^{u^*} = x \right].$$

Guarantee: The only critical point of $\mathbb{E}[\mathcal{L}_{\mathrm{Basic-Adj-Match}}]$ is the optimal control u^*.

Proof idea: u^* is the only solution of the fixed point equation above.

Adjoint Matching

Adjoint ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}a(t;X^u,u) = -\left[\left(\nabla_{X_t}(b(X_t,t)+\sigma(t)u(X_t,t))\right)^\top a(t;X^u,u) + \nabla_{X_t}\left(f(X_t,t)+\frac{1}{2}\|u(X_t,t)\|^2\right)\right],$$

$$a(1;X^u,u) = \nabla_x g(X_1^u)\,.$$

Lean Adjoint ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{a}(t;X^u) = -\left(\nabla_x b(X_t^u,t)^\top \tilde{a}(t;X^u) + \nabla_x f(X_t^u,t)\right),$$

$$\tilde{a}(1;X^u) = \nabla_x g(X_1^u)\,.$$

Key observation:

$$\mathbb{E}\left[\nabla_{X_t}(\sigma(t)u^*(X_t^{u^*},t))^\top a(t;X^{u^*},u^*) + \nabla_{X_t}\frac{1}{2}\|u(X_t^{u^*},t)\|^2\right] = 0$$

Adjoint Matching loss:

$$\mathscr{L}_{\mathrm{Adj-Match}}(u;X^u) := \frac{1}{2}\int_0^1 \left\|u(X_t,t)+\sigma(t)^\top\tilde{a}(t;X^{\bar{u}})\right\|^2 \mathrm{d}t\,.$$

The only critical point of $\mathbb{E}[\mathscr{L}_{\mathrm{Adj-Match}}]$ is the optimal control u^*

Recap

- New algorithm: **Adjoint Matching**
- Sample trajectories $\mathbf{X}$ according to the SDE $\mathrm{d}X_t^u = (b(X_t^u, t) + \sigma(t)u(X_t^u, t))\,\mathrm{d}t + \sigma(t)\,\mathrm{d}B_t$
- For each trajectory, compute:

$$\mathscr{L}_{\text{Adj-Match}}(u_\theta; \mathbf{X}) := \frac{1}{2}\int_0^1 \left\| u_\theta(X_t, t) + \sigma(t)\tilde{a}(t; \mathbf{X}) \right\|^2 \mathrm{d}t,$$

where $\tilde{a}$ is the solution to the *Lean Adjoint ODE*:

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{a}(t; \mathbf{X}) = -\nabla_x b(X_t, t)^\top \tilde{a}(t; \mathbf{X}), \qquad \tilde{a}(1; \mathbf{X}) = -\nabla_x r(X_1)\,.$$

- Compute $\nabla_\theta \mathscr{L}_{\text{Adj-Match}}(u_\theta; \mathbf{X})$ and update θ using e.g. Adam.

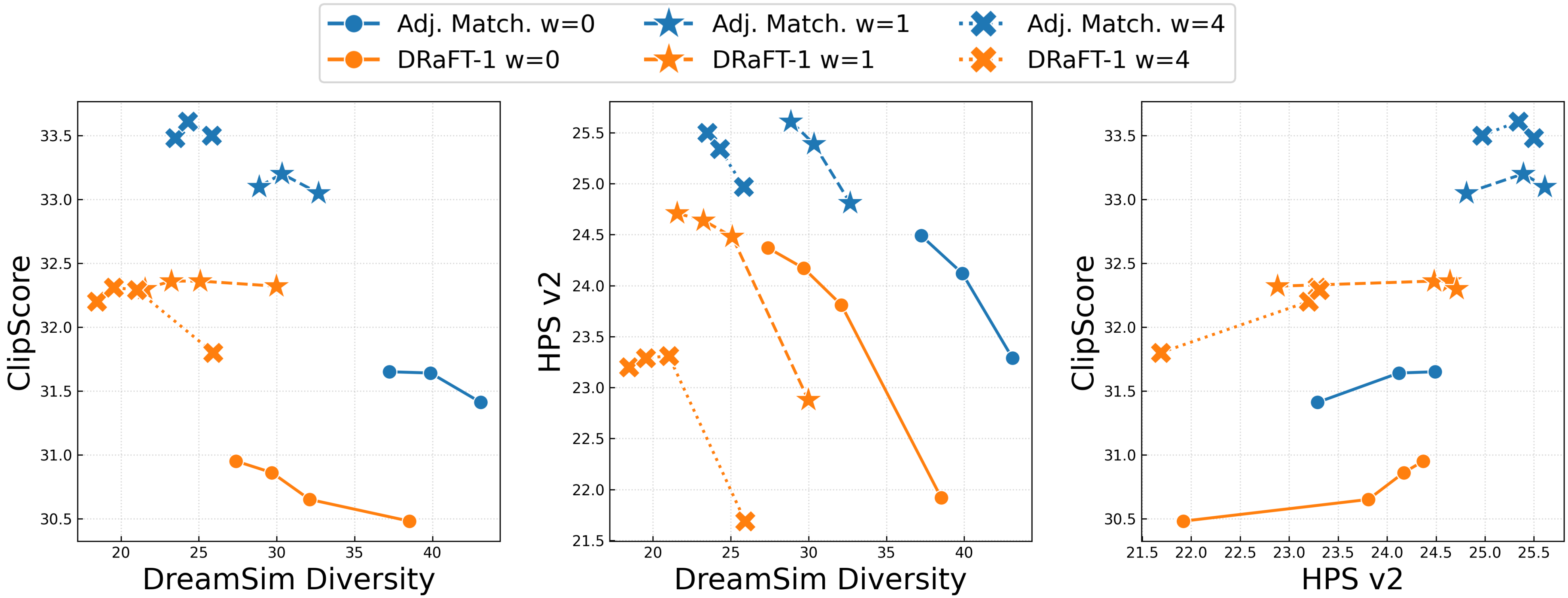
Comparison to non-SOC method

Adjoint Matching

"Handsome Smiling man in blue jacket portrait"

DRaFT-1 (Clark. et al., 2024)

Comparison to non-SOC method



Tradeoffs between different aspects of generative models: text-to-image consistency (ClipScore), sample diversity for each prompt (DreamSim Diversity), and generalization to unseen human preferences (HPS v2).

Thank you!

Backup: Stochastic optimal control

Important objects

Cost functional:

$$J(u; x, t) := \mathbb{E}_{X^u}\left[\int_t^1 \left(\frac{1}{2}\|u(X_s^u, s)\|^2 + f(X_s^u, s)\right) \mathrm{d}s + g(X_1^u) \;\middle|\; X_t^u = x\right].$$

Value function: $V(x, t) := \min_{u \in \mathcal{U}} J(u; x, t) = J(u^*; x, t),$

Path integral representation of the value function:

$$V(x, t) = -\log \mathbb{E}_X\left[\exp\left(-\int_t^1 f(X_s, s)\mathrm{d}s - g(X_1)\right) \;\middle|\; X_t = x\right].$$

Optimal control: $u^*(x, t) = -\sigma(t)^\top \nabla_x V(x, t) = -\sigma(t)^\top \nabla_x J(u^*, x, t)\,.$

Reminder: the SOC problem

$$\min_u \; \mathbb{E}\left[\int_0^1 \left(\frac{1}{2}\|u(X_t^u, t)\|^2 + f(X_t^u, t)\right) dt + g(X_1^u)\right],$$

$$s.t. \;\; \mathrm{d}X_t^u = (b(X_t^u, t) + \sigma(t)u(X_t^u, t))\,\mathrm{d}t + \sigma(t)\,\mathrm{d}B_t,$$

$$X_0^u \sim p_0\,.$$