Greed is Good: Guided Generation from a Greedy Perspective

Zander W. Blasingame Chen Liu

Clarkson University

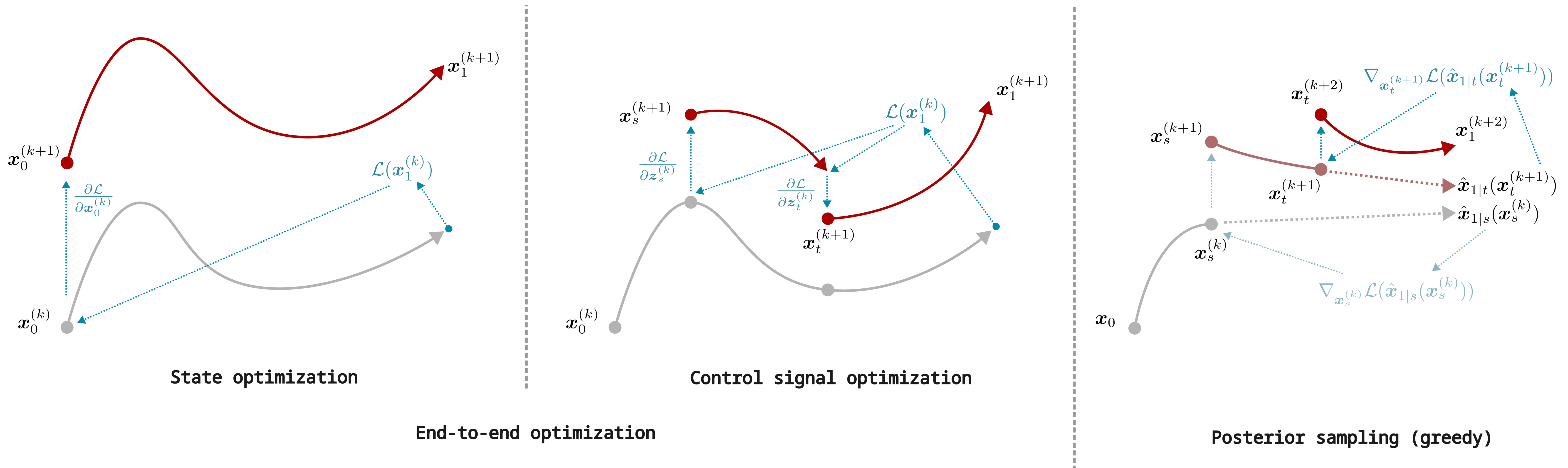

Figure 1. Visual comparison of different gradient-based training-free guided generation techniques. For state optimization techniques (left), the gray curve represents the previous solution trajectory $\{\boldsymbol{x}_t^{(k)}\}_{t\in[0,1]}$ at the k-th iteration, and the red curve shows the updated solution trajectory at the $(k+1)$-th iteration via the gradients in cyan. For control signal optimization techniques (middle), the solution trajectory is updated by changing the control signal $\boldsymbol{z}_t$ via the gradients $\partial\mathcal{L}/\partial\boldsymbol{z}_t$ calculated throughout the discretization scheme. For posterior sampling techniques (right), we use dull colors to denote the steps need to compute the $(k+1)$-th iteration and bright colors to denote the steps $(k+2)$-th iteration.

Training-free guided generation

Let $(\boldsymbol{X}_0, \boldsymbol{X}_1) \sim p(\boldsymbol{x}_0)q(\boldsymbol{x}_1)$ where $q(\boldsymbol{x})$ is the target distribution and $p(\boldsymbol{x}_0)$ is the prior distribution. Define $\boldsymbol{X}_t$ as

$$\boldsymbol{X}_t = \alpha_t \boldsymbol{X}_1 + \sigma_t \boldsymbol{X}_0, \tag{1}$$

for schedule (α_t, σ_t). Then, the vector field of the *affine conditional flow* $\Phi_t(\boldsymbol{x}|\boldsymbol{x}_1) = \alpha_t\boldsymbol{x}_1 + \sigma_t\boldsymbol{x}$ is given by

$$\boldsymbol{u}_t(\boldsymbol{x}) = \mathbb{E}[\dot\alpha_t \boldsymbol{X}_1 + \dot\sigma_t \boldsymbol{X}_0 | \boldsymbol{X}_t = \boldsymbol{x}]. \tag{2}$$

Assume that $\boldsymbol{u}_t^\theta$ is trained to zero loss, so $\boldsymbol{u}_t^\theta = \boldsymbol{u}_t$.

Problem statement. Find the optimal trajectory, *i.e.*, given a continuously differentiable loss function, $\mathcal{L} \in \mathcal{C}^1(\mathbb{R}^d; \mathbb{R})$, find the minimizer

$$\min_{\boldsymbol{x}_0} \mathcal{L}\left(\boldsymbol{x}_0 + \int_0^1 \boldsymbol{u}_\tau^\theta(\boldsymbol{x}_\tau)\,\mathrm{d}\tau\right). \tag{3}$$

A natural solution is to solve the *continuous adjoint equations* to find the gradients $\boldsymbol{a}_{\boldsymbol{x}}(t) \coloneqq \partial\mathcal{L}/\partial\boldsymbol{x}_t$, *i.e.*, solve

$$\boldsymbol{a}_{\boldsymbol{x}}(1) = \frac{\partial\mathcal{L}}{\partial\boldsymbol{x}_1}, \qquad \frac{\mathrm{d}\boldsymbol{a}_{\boldsymbol{x}}}{\mathrm{d}t}(t) = -\boldsymbol{a}_{\boldsymbol{x}}(t)^\top \frac{\partial\boldsymbol{u}_t^\theta}{\partial\boldsymbol{x}}(\boldsymbol{x}_t). \tag{4}$$

We call such an approach *optimize-then-discretize* (OTD) and vanilla backprop is referred to as *discretize-then-optimize* (DTO) [1].

However, this approach has a few difficulties:

Truncation errors. The trajectories $\{\boldsymbol{x}_n\}$ and $\{\tilde{\boldsymbol{x}}_n\}$ solved in reverse-time are not guaranteed to match.

Stability concerns. Stability in forward-time of the ODE instability in reverse-time for the adjoint equations.

Computational cost. Requires many *network function evaluations* NFEs.

N.B., for the first two issues *algebraically reversible* solvers offer a promising alternative [2, 3].

The greedy strategy

Rather than performing the full optimization back to $\boldsymbol{x}_0$, what if we greedily took an optimal step at each $\boldsymbol{x}_t$ instead?

Recall the target prediction model $\boldsymbol{x}_{1|t}^\theta(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{X}_1|\boldsymbol{X}_t = \boldsymbol{x}]$. Let $\gamma_t \coloneqq \alpha_t/\sigma_t$.

Proposition 1 (Exact solution of affine probability paths)**.** *Given an initial value of $\boldsymbol{x}_s$ at time $s \in [0,1]$ the solution $\boldsymbol{x}_t$ at time $t \in [0,1]$ of an ODE governed by the vector field in Eq. (2) is:*

$$\boldsymbol{x}_t = \frac{\sigma_t}{\sigma_s}\boldsymbol{x}_s + \sigma_t \int_{\gamma_s}^{\gamma_t} \boldsymbol{x}_{1|\gamma}^\theta(\boldsymbol{x}_\gamma)\,\mathrm{d}\gamma. \tag{5}$$

The flow can be written as the Taylor expansion.

$$\boldsymbol{x}_t = \frac{\sigma_t}{\sigma_s}\boldsymbol{x}_s + \sigma_t \sum_{n=0}^{k-1} \frac{\mathrm{d}^n}{\mathrm{d}\gamma^n}\left[\boldsymbol{x}_{1|\gamma}^\theta(\boldsymbol{x}_\gamma)\right]_{\gamma=\gamma_s} \frac{h^{n+1}}{n!} + \mathcal{O}(h^{k+1}). \tag{6}$$

Theoretical results

Theorem 2 (Greedy as an explicit Euler scheme in DTO)**.** *For some trajectory state $\boldsymbol{x}_t$ at time t, the greedy gradient given by $\nabla_{\boldsymbol{x}_t}\mathcal{L}(\boldsymbol{x}_{1|t}^\theta(\boldsymbol{x}_t))$ is the DTO scheme with explicit Euler discretization of flow model with step size $h = \gamma_1 - \gamma_t$.*

Proof sketch. Take the Taylor expansion of Eq. (6) and let $k=1$, then in the limit of $t \to 1$, we have $\tilde{\boldsymbol{x}}_1 = \boldsymbol{x}_{1|s}^\theta(\boldsymbol{x}_s)$. Thus the greedy strategy is the same as calculating the gradients via DTO with an explicit Euler scheme. □

Theorem 3 (Greedy as an implicit Euler method in OTD)**.** *For some trajectory state $\boldsymbol{x}_t$ at time t, the greedy gradient given by $\nabla_{\boldsymbol{x}_t}\mathcal{L}(\hat{\boldsymbol{x}}_{1|t}(\boldsymbol{x}_t))$ is an implicit Euler discretization of the continuous adjoint equations for the true gradients with step size $h = \gamma_1 - \gamma_t$.*

Proof sketch. First we use the technique of exponential integrators to simplify the continuous adjoint equations. Then we perform a first-order Taylor expansion around γ_t which is equivalent to an implicit Euler scheme as we are calculating the gradient flow from 1 to t. □

Theorem 4 (Greedy convergence)**.** *For affine probability paths, if there exists a sequence of states $\boldsymbol{x}_n$ at time t such that it converges to the locally optimal solution $\hat{\boldsymbol{x}}_{1|t}(\boldsymbol{x}_n) \to \boldsymbol{x}_1^*$. Then, $\|\Phi_{1|t}(\boldsymbol{x}_n) - \boldsymbol{x}_1^*\|$ is $\mathcal{O}(h^2)$ as $n \to \infty$.*

Proof sketch. First, we take a first-order Taylor expansion of the flow described in Eq. (5) which as shown above simplifies to the denoiser plus $\mathcal{O}(h^2)$ truncation error terms. Clearly, if $\hat{\boldsymbol{x}}_{1|t}(\boldsymbol{x}_n) \to \boldsymbol{x}_1^*$ then the first term of the Taylor expansion does as well. □

Consider optimizing the control signal as in [4, 5], *i.e.*, inject an additional control signal, $\boldsymbol{z} \in \mathcal{C}^1(\mathbb{R}; \mathbb{R}^d)$, to the vector field, $\boldsymbol{u}_t^\theta$, such that

$$\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{u}_t^\theta(\boldsymbol{x}_t) + \boldsymbol{z}(t). \tag{7}$$

Theorem 5 (Continuous adjoint equations for the control term)**.** *Let $\boldsymbol{u}_\theta \in \mathcal{C}^{1,1}([0,1] \times \mathbb{R}^d; \mathbb{R}^d)$ be a parameterization of some time-dependent vector field of a neural ODE that is Lipschitz continuous in its second argument, and let $\boldsymbol{z} \in \mathcal{C}^1([0,1]; \mathbb{R}^d)$ be an additional control signal such that the new dynamics are given by Eq. (7). Let $\boldsymbol{a}_{\boldsymbol{z}}(t) \coloneqq \partial\mathcal{L}/\partial\boldsymbol{z}(t)$ then*

$$\boldsymbol{a}_{\boldsymbol{z}}(t) = -\int_1^t \boldsymbol{a}_{\boldsymbol{x}}(s)\,\mathrm{d}s. \tag{8}$$

References

[1] P. Kidger, *On Neural Differential Equations*. Ph.d. thesis, Oxford University, 2022.

[2] S. McCallum and J. Foster, “Efficient, accurate and stable gradients for neural odes,” *arXiv preprint arXiv:2410.11648*, 2024.

[3] Z. W. Blasingame and C. Liu, “A reversible solver for diffusion SDEs,” in *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2025.

[4] L. Wang, C. Cheng, Y. Liao, Y. Qu, and G. Liu, “Training free guided flow-matching with optimal control,” in *The Thirteenth International Conference on Learning Representations*, 2025.

[5] X. Liu, L. Wu, S. Zhang, C. Gong, W. Ping, and Q. Liu, “Flowgrad: Controlling the output of generative odes with gradients,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24335–24344, 2023.

Correspondence to blasinzw@clarkson.edu