Twisted Diffusion Sampling for Accurate Conditional Generation, with Application to Protein Design

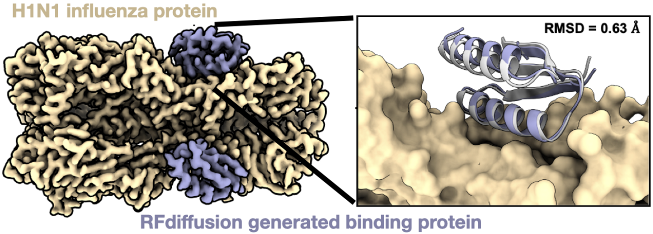
Brian L. Trippe

Columbia University, Department of Statistics

July 19, 2023

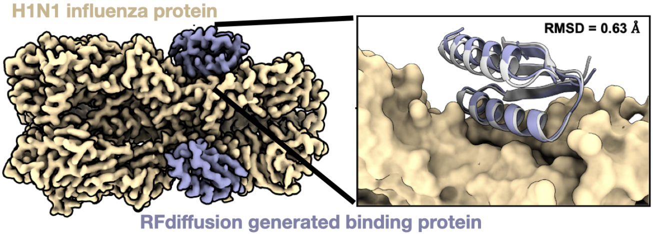
Conditional Sampling for Protein Binder Design

Diffusion models generate protein binders with $10 - 100\times$ higher success rates than previous methods [WJB**T**Y+, 2022]



Conditional Sampling for Protein Binder Design

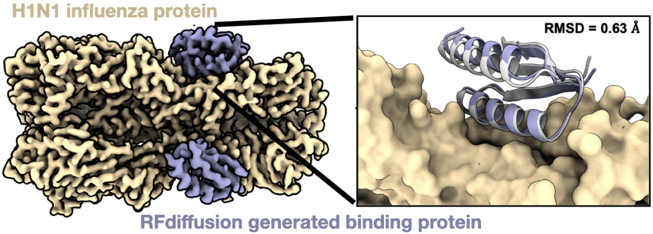
Diffusion models generate protein binders with $10-100\times$ higher success rates than previous methods [WJB**T**Y+, 2022]



Key idea: Learn model of protein structure, design by sampling from conditional distributions

Conditional Sampling for Protein Binder Design

Diffusion models generate protein binders with $10-100\times$ higher success rates than previous methods [WJB**T**Y+, 2022]

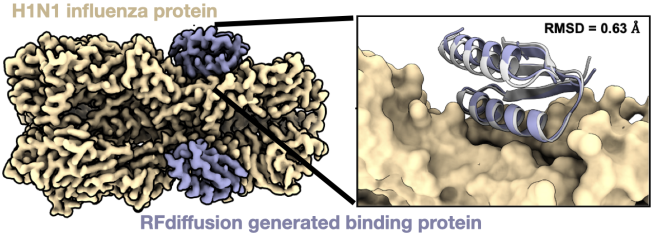


Key idea: Learn model of protein structure, design by sampling from conditional distributions

Challenge: analytical intractability of exact conditionals

Conditional Sampling for Protein Binder Design

Diffusion models generate protein binders with $10 - 100\times$ higher success rates than previous methods [WJB**T**Y+, 2022]



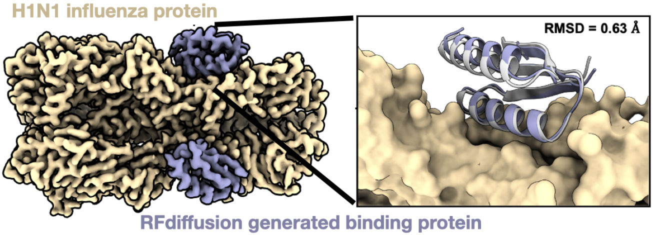
Key idea: Learn model of protein structure, design by sampling from conditional distributions

Challenge: analytical intractability of exact conditionals

- Heuristic guidance [reconstruction & replacement]: inaccurate

Conditional Sampling for Protein Binder Design

Diffusion models generate protein binders with $10 - 100\times$ higher success rates than previous methods [WJB**T**Y+, 2022]



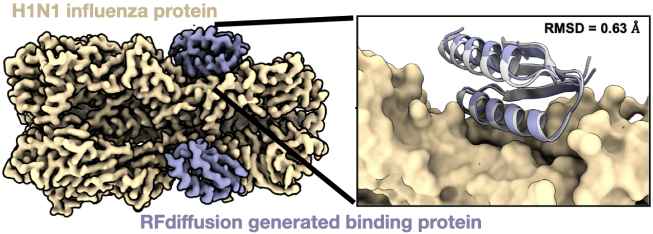
Key idea: Learn model of protein structure, design by sampling from conditional distributions

Challenge: analytical intractability of exact conditionals

- Heuristic guidance [reconstruction & replacement]: inaccurate
- Conditional training: time-consuming

Conditional Sampling for Protein Binder Design

Diffusion models generate protein binders with $10 - 100\times$ higher success rates than previous methods [WJB**T**Y+, 2022]



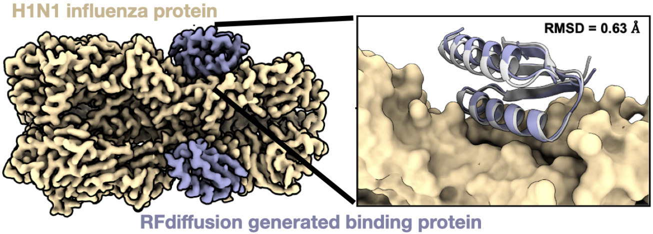
Key idea: Learn model of protein structure, design by sampling from conditional distributions

Challenge: analytical intractability of exact conditionals

- Heuristic guidance [reconstruction & replacement]: inaccurate
- Conditional training: time-consuming, (inaccurate)

Conditional Sampling for Protein Binder Design

Diffusion models generate protein binders with $10 - 100\times$ higher success rates than previous methods [WJB**T**Y+, 2022]



Key idea: Learn model of protein structure, design by sampling from conditional distributions

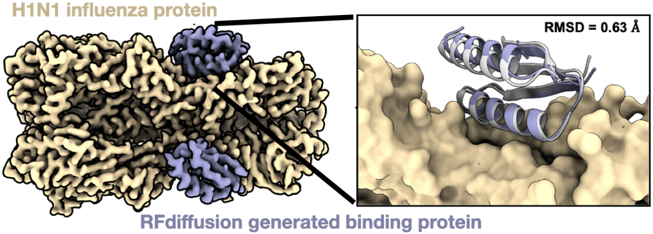
Challenge: analytical intractability of exact conditionals

- Heuristic guidance [reconstruction & replacement]: inaccurate
- Conditional training: time-consuming, (inaccurate)

We provide: Sequential Monte Carlo for conditional generation.

Conditional Sampling for Protein Binder Design

Diffusion models generate protein binders with $10 - 100\times$ higher success rates than previous methods [WJB**T**Y+, 2022]



Key idea: Learn model of protein structure, design by sampling from conditional distributions

Challenge: analytical intractability of exact conditionals

- Heuristic guidance [reconstruction & replacement]: inaccurate
- Conditional training: time-consuming, (inaccurate)

We provide: Sequential Monte Carlo for conditional generation.

- Asymptotically exact (in compute cost), general, and delivers state of the art *in silico* success rates in protein design.

Roadmap

- Diffusion models and conditional generation
- The Twisted Diffusion Sampler (TDS)
- Related work
- Properties of TDS (Theory and Simulations)
- Case study in motif-scaffolding

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0) \prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0)\prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0) \prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q
  - $x^T \sim q(x^T)$
  - $x^{t-1} \sim q(x^{t-1} | x^t)$

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0) \prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q... approximately
  - $x^T \sim q(x^T)$
  - $x^{t-1} \sim q(x^{t-1} | x^t)$

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0)\prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q... approximately
  - $x^T \sim q(x^T) \approx \mathcal{N}(0, T\sigma^2)$
  - $x^{t-1} \sim q(x^{t-1}|x^t)$

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0)\prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q... approximately
  - $x^T \sim q(x^T) \approx \mathcal{N}(0, T\sigma^2)$
  - $x^{t-1} \sim q(x^{t-1}|x^t)$
- Two approximations for $q(x^{t-1} \mid x^t)$:

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0) \prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q... approximately
  - $x^T \sim q(x^T) \approx \mathcal{N}(0, T\sigma^2)$
  - $x^{t-1} \sim q(x^{t-1}|x^t) \approx \mathcal{N}(x^{t-1}|x^t + \sigma^2 \nabla \log q(x^t), \sigma^2)$
- Two approximations for $q(x^{t-1} \mid x^t)$:
  1. Gaussian approximation (via Bayes' rule + Taylor's theorem)

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0)\prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q... approximately
  - $x^T \sim q(x^T) \approx \mathcal{N}(0, T\sigma^2)$
  - $x^{t-1} \sim q(x^{t-1}|x^t) \approx \mathcal{N}(x^{t-1}|x^t + \sigma^2 \nabla \log q(x^t), \sigma^2)$
- Two approximations for $q(x^{t-1} \mid x^t)$:
  1. Gaussian approximation (via Bayes' rule + Taylor's theorem)

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0)\prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q... approximately
  - $x^T \sim q(x^T) \approx \mathcal{N}(0, T\sigma^2)$
  - $x^{t-1} \sim q(x^{t-1}|x^t) \approx \mathcal{N}(x^{t-1}|x^t + \sigma^2 \nabla \log q(x^t), \sigma^2)$
- Two approximations for $q(x^{t-1} \mid x^t)$:
  1. Gaussian approximation (via Bayes' rule + Taylor's theorem)
  2. Tweedie's rule: $\nabla \log q(x^t) = (\mathbb{E}_q[x^0|x^t] - x^t)/t\sigma^2$

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0)\prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q... approximately
  - $x^T \sim q(x^T) \approx \mathcal{N}(0, T\sigma^2)$
  - $x^{t-1} \sim q(x^{t-1}|x^t) \approx \mathcal{N}(x^{t-1}|x^t{+}\sigma^2\nabla \log q(x^t), \sigma^2)$
- Two approximations for $q(x^{t-1} \mid x^t)$:
  1. Gaussian approximation (via Bayes' rule + Taylor's theorem)
  2. Tweedie's rule: $\nabla \log q(x^t) {=} (\mathbb{E}_q[x^0|x^t] {-} x^t)/t\sigma^2$

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0) \prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q... approximately
  - $x^T \sim q(x^T) \approx \mathcal{N}(0, T\sigma^2)$
  - $x^{t-1} \sim q(x^{t-1}|x^t) \approx \mathcal{N}(x^{t-1}|x^t + \sigma^2 \nabla \log q(x^t), \sigma^2)$
- Two approximations for $q(x^{t-1} \mid x^t)$:
  1. Gaussian approximation (via Bayes' rule + Taylor's theorem)
  2. Tweedie's rule: $\nabla \log q(x^t) = (\mathbb{E}_q[x^0|x^t] - x^t)/t\sigma^2$
     - Train $\hat{x}_\theta(x^t, t) \approx \mathbb{E}_q[x^0|x^t]$ (via "denoising score matching")

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0)\prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \dots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q... approximately
  - $x^T \sim q(x^T) \approx \mathcal{N}(0, T\sigma^2)$
  - $x^{t-1} \sim q(x^{t-1}|x^t) \approx \mathcal{N}(x^{t-1}|x^t+\sigma^2\nabla \log q(x^t), \sigma^2)$
- Two approximations for $q(x^{t-1} \mid x^t)$:
  1. Gaussian approximation (via Bayes' rule + Taylor's theorem)
  2. Tweedie's rule: $\nabla \log q(x^t) \approx (\hat{x}_\theta(x^t) - x^t)/t\sigma^2$
     - Train $\hat{x}_\theta(x^t, t) \approx \mathbb{E}_q[x^0|x^t]$ (via "denoising score matching")

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0)\prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \dots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q... approximately
  - $x^T \sim q(x^T) \approx \mathcal{N}(0, T\sigma^2)$
  - $x^{t-1} \sim q(x^{t-1}|x^t) \approx \mathcal{N}(x^{t-1}|x^t + \sigma^2 \nabla \log q(x^t), \sigma^2)$
- Two approximations for $q(x^{t-1} \mid x^t)$:
  1. Gaussian approximation (via Bayes' rule + Taylor's theorem)
  2. Tweedie's rule: $\nabla \log q(x^t) \approx (\hat{x}_\theta(x^t) - x^t)/t\sigma^2 =: s_\theta(x^t)$
     - Train $\hat{x}_\theta(x^t, t) \approx \mathbb{E}_q[x^0|x^t]$ (via "denoising score matching")

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0)\prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q... approximately
  - $x^T \sim q(x^T) \approx \mathcal{N}(0, T\sigma^2)$
  - $x^{t-1} \sim q(x^{t-1}|x^t) \approx \mathcal{N}(x^{t-1}|x^t + \sigma^2 s_\theta(x^t), \sigma^2)$
- Two approximations for $q(x^{t-1} \mid x^t)$:
  1. Gaussian approximation (via Bayes' rule + Taylor's theorem)
  2. Tweedie's rule: $\nabla \log q(x^t) \approx (\hat{x}_\theta(x^t) - x^t)/t\sigma^2 =: s_\theta(x^t)$
     - Train $\hat{x}_\theta(x^t, t) \approx \mathbb{E}_q[x^0|x^t]$ (via "denoising score matching")

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0) \prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by “reversing” q... approximately
  - $x^T \sim q(x^T) \approx \mathcal{N}(0, T\sigma^2)$
  - $x^{t-1} \sim q(x^{t-1} | x^t) \approx \mathcal{N}(x^{t-1} | x^t + \sigma^2 s_\theta(x^t), \sigma^2)$
- Two approximations for $q(x^{t-1} \mid x^t)$:
  1. Gaussian approximation (via Bayes' rule + Taylor's theorem)
  2. Tweedie's rule: $\nabla \log q(x^t) \approx (\hat{x}_\theta(x^t) - x^t)/t\sigma^2 =: s_\theta(x^t)$
     - Train $\hat{x}_\theta(x^t, t) \approx \mathbb{E}_q[x^0 | x^t]$ (via “denoising score matching”)

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0)\prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \dots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q... approximately
  - $x^T \sim q(x^T) \approx \mathcal{N}(0, T\sigma^2)$
  - $x^{t-1} \sim q(x^{t-1}|x^t) \approx \mathcal{N}(x^{t-1}|x^t + \sigma^2 s_\theta(x^t), \sigma^2)$
- Two approximations for $q(x^{t-1} \mid x^t)$:
  1. Gaussian approximation (via Bayes' rule + Taylor's theorem)
  2. Tweedie's rule: $\nabla \log q(x^t) \approx (\hat{x}_\theta(x^t) - x^t)/t\sigma^2 =: s_\theta(x^t)$
     - Train $\hat{x}_\theta(x^t, t) \approx \mathbb{E}_q[x^0|x^t]$ (via "denoising score matching")
- Diffusion model: $p_\theta(x^{0:T}) = p_\theta(x^T)\prod_{t=1}^{T} p_\theta(x^{t-1} \mid x^t)$

Diffusion Set-Up and Notation

- **Goal:** estimation of $q(x^0)$ from samples.
- Noising process: $q(x^{0:T}) = q(x^0)\prod_{t=1}^{T} q(x^t \mid x^{t-1})$
  - $q(x^t \mid x^{t-1}) = \mathcal{N}(x^t \mid x^{t-1}, \sigma^2)$ for $t = 1, \ldots, T$
  - $q(x^t) = \int \mathcal{N}(x^t \mid x^0, t\sigma^2) q(x^0) dx^0$
- Idea: generate new samples by "reversing" q... approximately
  - $x^T \sim q(x^T) \approx \mathcal{N}(0, T\sigma^2) := p_\theta(x^T)$
  - $x^{t-1} \sim q(x^{t-1}|x^t) \approx \mathcal{N}(x^{t-1}|x^t + \sigma^2 s_\theta(x^t), \sigma^2) =: p_\theta(x^{t-1}|x^t)$
- Two approximations for $q(x^{t-1} \mid x^t)$:
  1. Gaussian approximation (via Bayes' rule + Taylor's theorem)
  2. Tweedie's rule: $\nabla \log q(x^t) \approx (\hat{x}_\theta(x^t) - x^t)/t\sigma^2 =: s_\theta(x^t)$
     - Train $\hat{x}_\theta(x^t, t) \approx \mathbb{E}_q[x^0|x^t]$ (via "denoising score matching")
- Diffusion model: $p_\theta(x^{0:T}) = p_\theta(x^T)\prod_{t=1}^{T} p_\theta(x^{t-1} \mid x^t)$

Importance Sampling for Controlled Generation

Goal: Generate x^0 in response to conditioning criteria y

Importance Sampling for Controlled Generation

Goal: Generate x^0 in response to conditioning criteria y

1. Augment p_θ with y, let $p_\theta(x^{0:T}, y) = p_\theta(x^{0:T})p(y \mid x^0)$.

Importance Sampling for Controlled Generation

Goal: Generate x^0 in response to conditioning criteria y

1. Augment p_θ with y, let $p_\theta(x^{0:T}, y) = p_\theta(x^{0:T})p(y \mid x^0)$
2. Given $y \sim p(y|x^0)$, compute $p_\theta(x^{0:T}|y)$.

Importance Sampling for Controlled Generation

Goal: Generate x^0 in response to conditioning criteria y

1. Augment p_θ with y, let $p_\theta(x^{0:T}, y) = p_\theta(x^{0:T})p(y \mid x^0)$
2. Given $y \sim p(y|x^0)$, compute $p_\theta(x^{0:T}|y) \propto p_\theta(x^{0:T})p(y \mid x^0)$.

Importance Sampling for Controlled Generation

Goal: Generate x^0 in response to conditioning criteria y

1. Augment p_θ with y, let $p_\theta(x^{0:T}, y) = p_\theta(x^{0:T})p(y \mid x^0)$
2. Given $y \sim p(y|x^0)$, compute $\underbrace{p_\theta(x^{0:T}|y)}_{target(\nu)} \propto \underbrace{p_\theta(x^{0:T})}_{proposal(\tilde{p})} \underbrace{p(y \mid x^0)}_{weight(w)}$.

Idea: Importance sampling for target $\nu(x)$

- Choose *proposal:* $\tilde{p}(x)$, and *weights:* $w(x) \propto \nu(x)/\tilde{p}(x)$

Importance Sampling for Controlled Generation

Goal: Generate x^0 in response to conditioning criteria y

1. Augment p_θ with y, let $p_\theta(x^{0:T}, y) = p_\theta(x^{0:T})p(y \mid x^0)$
2. Given $y \sim p(y|x^0)$, compute $\underbrace{p_\theta(x^{0:T}|y)}_{target(\nu)} \propto \underbrace{p_\theta(x^{0:T})}_{proposal(\tilde{p})} \underbrace{p(y \mid x^0)}_{weight(w)}$.

Idea: Importance sampling for target $\nu(x)$

- Choose *proposal:* $\tilde{p}(x)$, and *weights:* $w(x) \propto \nu(x)/\tilde{p}(x)$

Key property: Let $\hat{\mathbb{P}}_K = \sum_{k=1}^K w_k \delta_{x_k}$, with $x_k \overset{iid}{\sim} \tilde{p}$ and $w_k = w(x_k)/\sum w(x_{k'})$. If $w(\cdot)$ is bounded, $\hat{\mathbb{P}}_K \to \nu$ as $K \to \infty$.

Importance Sampling for Controlled Generation

Goal: Generate x^0 in response to conditioning criteria y

1. Augment p_θ with y, let $p_\theta(x^{0:T}, y) = p_\theta(x^{0:T})p(y \mid x^0)$
2. Given $y \sim p(y|x^0)$, compute $\underbrace{p_\theta(x^{0:T}|y)}_{target(\nu)} \propto \underbrace{p_\theta(x^{0:T})}_{proposal(\tilde{p})} \underbrace{p(y \mid x^0)}_{weight(w)}$.

Idea: Importance sampling for target $\nu(x)$

- Choose *proposal:* $\tilde{p}(x)$, and *weights:* $w(x) \propto \nu(x)/\tilde{p}(x)$

Key property: Let $\hat{\mathbb{P}}_K = \sum_{k=1}^K w_k \delta_{x_k}$, with $x_k \overset{iid}{\sim} \tilde{p}$ and $w_k = w(x_k)/\sum w(x_{k'})$. If $w(\cdot)$ is bounded, $\hat{\mathbb{P}}_K \to \nu$ as $K \to \infty$. (i.e. $\forall A$, $\lim\limits_{K\to\infty} \hat{\mathbb{P}}_K(A) = \int_A \nu(x)dx$ with prob. 1)

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: # samples needed for class-conditional generation

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: $\#$ samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \dots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: $\#$ samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \ldots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \,||\, p_\theta(x^{0:T}))\}$?

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: $\#$ samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \ldots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \,||\, p_\theta(x^{0:T}))\}$? $\approx 10.$

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: # samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \ldots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \,||\, p_\theta(x^{0:T}))\}$? $\quad \approx 10.$

$$\mathrm{KL}(p_\theta(x^{0:T}|y) \,||\, p_\theta(x^{0:T})) = \int p_\theta(x^{0:T}|y) \log \frac{p_\theta(x^{0:T}|y)}{p_\theta(x^{0:T})} dx^{0:T}$$

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: # samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \dots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \,||\, p_\theta(x^{0:T}))\}$? ≈ 10.

$$\mathrm{KL}(p_\theta(x^{0:T}|y) \,||\, p_\theta(x^{0:T})) = \int p_\theta(x^{0:T}|y) \log \frac{p_\theta(x^{0:T}|y)}{p_\theta(x^{0:T})} dx^{0:T}$$

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: $\#$ samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \dots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \,||\, p_\theta(x^{0:T}))\}$? $\quad \approx 10.$

$$\mathrm{KL}(p_\theta(x^{0:T}|y) \,||\, p_\theta(x^{0:T})) = \int p_\theta(x^{0:T}|y) \log \frac{p_\theta(x^0|y)}{p_\theta(x^0)} dx^{0:T}$$

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \mid\mid \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: # samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \ldots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \mid\mid p_\theta(x^{0:T}))\}$? $\quad \approx 10.$

$$\mathrm{KL}(p_\theta(x^{0:T}|y) \mid\mid p_\theta(x^{0:T})) = \int p_\theta(x^0|y) \log \frac{p_\theta(x^0|y)}{p_\theta(x^0)} dx^0$$

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \mid\mid \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: # samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \ldots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \mid\mid p_\theta(x^{0:T}))\}$? $\approx 10.$

$$\begin{aligned}\mathrm{KL}(p_\theta(x^{0:T}|y) \mid\mid p_\theta(x^{0:T})) &= \int p_\theta(x^0|y) \log \frac{p_\theta(x^0|y)}{p_\theta(x^0)} dx^0 \\ &= \int p_\theta(x^0|y) \log \frac{p(y|x^0)}{p_\theta(y)} dx^0\end{aligned}$$

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: # samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \dots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \,||\, p_\theta(x^{0:T}))\}$? $\quad \approx 10.$

$$\begin{aligned}
\mathrm{KL}(p_\theta(x^{0:T}|y) \,||\, p_\theta(x^{0:T})) &= \int p_\theta(x^0|y) \log \frac{p_\theta(x^0|y)}{p_\theta(x^0)} dx^0 \\
&= \int p_\theta(x^0|y) \log \frac{p(y|x^0)}{p_\theta(y)} dx^0 \\
&= \mathbb{E}_{p_\theta}[\log p(y|x^0) \mid y] - \log p_\theta(y)
\end{aligned}$$

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \mid\mid \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: # samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \ldots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \mid\mid p_\theta(x^{0:T}))\}$? $\quad \approx 10.$

$$\begin{aligned}
\mathrm{KL}(p_\theta(x^{0:T}|y) \mid\mid p_\theta(x^{0:T})) &= \int p_\theta(x^0|y) \log \frac{p_\theta(x^0|y)}{p_\theta(x^0)} dx^0 \\
&= \int p_\theta(x^0|y) \log \frac{p(y|x^0)}{p_\theta(y)} dx^0 \\
&= \mathbb{E}_{p_\theta}[\log p(y|x^0) \mid y] - \log p_\theta(y)
\end{aligned}$$

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: $\#$ samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \ldots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \,||\, p_\theta(x^{0:T}))\}$? $\quad \approx 10.$

$$\begin{aligned}
\mathrm{KL}(p_\theta(x^{0:T}|y) \,||\, p_\theta(x^{0:T})) &= \int p_\theta(x^0|y) \log \frac{p_\theta(x^0|y)}{p_\theta(x^0)} dx^0 \\
&= \int p_\theta(x^0|y) \log \frac{p(y|x^0)}{p_\theta(y)} dx^0 \\
&\approx \mathbb{E}_{p_\theta}[\log p(y|x^0) \mid y] - \log 1/10
\end{aligned}$$

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: # samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \ldots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \,||\, p_\theta(x^{0:T}))\}$? $\approx 10.$

$$
\begin{aligned}
\mathrm{KL}(p_\theta(x^{0:T}|y) \,||\, p_\theta(x^{0:T})) &= \int p_\theta(x^0|y) \log \frac{p_\theta(x^0|y)}{p_\theta(x^0)} dx^0 \\
&= \int p_\theta(x^0|y) \log \frac{p(y|x^0)}{p_\theta(y)} dx^0 \\
&\approx \mathbb{E}_{p_\theta}[\log p(y|x^0) \mid y] + \log 10
\end{aligned}
$$

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: # samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \ldots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \,||\, p_\theta(x^{0:T}))\}$? $\approx 10.$

$$
\begin{aligned}
\mathrm{KL}(p_\theta(x^{0:T}|y) \,||\, p_\theta(x^{0:T})) &= \int p_\theta(x^0|y) \log \frac{p_\theta(x^0|y)}{p_\theta(x^0)} dx^0 \\
&= \int p_\theta(x^0|y) \log \frac{p(y|x^0)}{p_\theta(y)} dx^0 \\
&\approx \mathbb{E}_{p_\theta}[\, \log p(y|x^0) \mid y] + \log 10
\end{aligned}
$$

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: $\#$ samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \ldots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \,||\, p_\theta(x^{0:T}))\}$? $\quad \approx 10.$

$$
\begin{aligned}
\mathrm{KL}(p_\theta(x^{0:T}|y) \,||\, p_\theta(x^{0:T})) &= \int p_\theta(x^0|y) \log \frac{p_\theta(x^0|y)}{p_\theta(x^0)} dx^0 \\
&= \int p_\theta(x^0|y) \log \frac{p(y|x^0)}{p_\theta(y)} dx^0 \\
&\approx \mathbb{E}_{p_\theta}[\log 1] + \log 10
\end{aligned}
$$

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \mid\mid \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: $\#$ samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \dots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \mid\mid p_\theta(x^{0:T}))\}$? $\quad \approx 10.$

$$
\begin{aligned}
\mathrm{KL}(p_\theta(x^{0:T}|y) \mid\mid p_\theta(x^{0:T})) &= \int p_\theta(x^0|y) \log \frac{p_\theta(x^0|y)}{p_\theta(x^0)} dx^0 \\
&= \int p_\theta(x^0|y) \log \frac{p(y|x^0)}{p_\theta(y)} dx^0 \\
&\approx \log 10
\end{aligned}
$$

Efficiency of Importance Sampling

Problem: Need $O(\exp\{\mathrm{KL}(p \,||\, \tilde{p})\})$ samples [Chatterjee and Diaconis, 2018]

Example: # samples needed for class-conditional generation

- Proposal: $\tilde{p}(x) = p_\theta(x^{0:T})$ is MNIST diffusion model
- Weight: $p_\theta(y \mid x^0)$ is classifier for $y \in \{0, \dots, 9\}$,
- Target: $p_\theta(x^0|y)$ is images of class y
- What is $\exp\{\mathrm{KL}(p_\theta(x^{0:T} \mid y) \,||\, p_\theta(x^{0:T}))\}$? $\approx 10.$

$$
\begin{aligned}
\mathrm{KL}(p_\theta(x^{0:T}|y) \,||\, p_\theta(x^{0:T})) &= \int p_\theta(x^0|y) \log \frac{p_\theta(x^0|y)}{p_\theta(x^0)} dx^0 \\
&= \int p_\theta(x^0|y) \log \frac{p(y|x^0)}{p_\theta(y)} dx^0 \\
&\approx \log 10
\end{aligned}
$$

Intuition: Roughly 1 in 10 samples will be digit y

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y)$

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y) \prod_{t=1}^{T} p_\theta(x^{t-1}|x^t, y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y) \prod_{t=1}^{T} p_\theta(x^{t-1}|x^t, y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y) \prod_{t=1}^{T} p_\theta(x^{t-1}|x^t, y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y) \prod_{t=1}^{T} p_\theta(x^{t-1}|x^t, y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y) \prod_{t=1}^{T} p_\theta(x^{t-1}|x^t, y)$

Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y){=}p_\theta(x^T|y)\prod_{t=1}^T p_\theta(x^{t-1}|x^t, y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde{p}_\theta(x^{t-1}|x^t, y){:=}\mathcal{N}(x^{t-1}|x^t{+}\sigma^2[s_\theta(x^t) + \nabla_{x^t} \log \tilde{p}_\theta(y|x^t)], \sigma^2)$
  where $\tilde{p}_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat{x}_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y)=p_\theta(x^T|y)\prod_{t=1}^T p_\theta(x^{t-1}|x^t,y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde{p}_\theta(x^{t-1}|x^t,y) := \mathcal{N}(x^{t-1}|x^t+\sigma^2[s_\theta(x^t)+\nabla_{x^t}\log\tilde{p}_\theta(y|x^t)],\sigma^2)$ where $\tilde{p}_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat{x}_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

Three approximations to derive $\tilde{p}_\theta(x^{t-1} \mid x^t, y)$:
$p_\theta(x^{t-1} \mid x^t, y) = p_\theta(x^{t-1} \mid x^t)p_\theta(y \mid x^{t-1})/p_\theta(y \mid x^t)$

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y)\prod_{t=1}^T p_\theta(x^{t-1}|x^t, y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde{p}_\theta(x^{t-1}|x^t, y) := \mathcal{N}(x^{t-1}|x^t + \sigma^2[s_\theta(x^t) + \nabla_{x^t}\log\tilde{p}_\theta(y|x^t)], \sigma^2)$ where $\tilde{p}_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat{x}_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

Three approximations to derive $\tilde{p}_\theta(x^{t-1} \mid x^t, y)$:
$p_\theta(x^{t-1} \mid x^t, y) = p_\theta(x^{t-1} \mid x^t) p_\theta(y \mid x^{t-1}) / p_\theta(y \mid x^t)$

1. Likelihood approximation

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y)\prod_{t=1}^T p_\theta(x^{t-1}|x^t, y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde{p}_\theta(x^{t-1}|x^t, y) := \mathcal{N}(x^{t-1}|x^t + \sigma^2[s_\theta(x^t) + \nabla_{x^t} \log \tilde{p}_\theta(y|x^t)], \sigma^2)$
  where $\tilde{p}_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat{x}_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

Three approximations to derive $\tilde{p}_\theta(x^{t-1} \mid x^t, y)$:

$$p_\theta(x^{t-1} \mid x^t, y) \overset{1}{\approx} p_\theta(x^{t-1} \mid x^t)\tilde{p}_\theta(y \mid x^{t-1})/\tilde{p}_\theta(y \mid x^t)$$

1. Likelihood approximation

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y)=p_\theta(x^T|y)\prod_{t=1}^T p_\theta(x^{t-1}|x^t,y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde{p}_\theta(x^{t-1}|x^t,y):=\mathcal{N}(x^{t-1}|x^t+\sigma^2[s_\theta(x^t)+\nabla_{x^t}\log\tilde{p}_\theta(y|x^t)],\sigma^2)$ where $\tilde{p}_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat{x}_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

Three approximations to derive $\tilde{p}_\theta(x^{t-1} \mid x^t, y)$:

$$p_\theta(x^{t-1} \mid x^t, y) \overset{1}{\approx} p_\theta(x^{t-1} \mid x^t)\tilde{p}_\theta(y \mid x^{t-1})/\tilde{p}_\theta(y \mid x^t)$$

1. Likelihood approximation ($\tilde{p}_\theta(y|x^0)=p(y|x^0)$ if $\hat{x}_\theta(x^0)=x^0$)

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y)\prod_{t=1}^T p_\theta(x^{t-1}|x^t, y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde{p}_\theta(x^{t-1}|x^t, y) := \mathcal{N}(x^{t-1}|x^t + \sigma^2[s_\theta(x^t) + \nabla_{x^t} \log \tilde{p}_\theta(y|x^t)], \sigma^2)$ where $\tilde{p}_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat{x}_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

Three approximations to derive $\tilde{p}_\theta(x^{t-1} \mid x^t, y)$:

$$p_\theta(x^{t-1} \mid x^t, y) \overset{1}{\approx} p_\theta(x^{t-1} \mid x^t)\tilde{p}_\theta(y \mid x^{t-1})/\tilde{p}_\theta(y \mid x^t)$$

1. Likelihood approximation ($\tilde{p}_\theta(y|x^0) = p(y|x^0)$ if $\hat{x}_\theta(x^0) = x^0$)
2. Taylor expand $\log \frac{\tilde{p}_\theta(y|x^{t-1})}{\tilde{p}_\theta(y|x^t)}$

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y)\prod_{t=1}^{T} p_\theta(x^{t-1}|x^t, y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde{p}_\theta(x^{t-1}|x^t, y) := \mathcal{N}(x^{t-1}|x^t + \sigma^2[s_\theta(x^t) + \nabla_{x^t}\log\tilde{p}_\theta(y|x^t)], \sigma^2)$
  where $\tilde{p}_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat{x}_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

Three approximations to derive $\tilde{p}_\theta(x^{t-1} \mid x^t, y)$:

$$
\begin{aligned}
p_\theta(x^{t-1} \mid x^t, y) &\overset{1}{\approx} p_\theta(x^{t-1} \mid x^t)\tilde{p}_\theta(y \mid x^{t-1})/\tilde{p}_\theta(y \mid x^t) \\
&= p_\theta(x^{t-1} \mid x^t)\exp\{\log\tilde{p}_\theta(y \mid x^{t-1})/\tilde{p}_\theta(y \mid x^t)\}
\end{aligned}
$$

1. Likelihood approximation ($\tilde{p}_\theta(y|x^0) = p(y|x^0)$ if $\hat{x}_\theta(x^0) = x^0$)
2. Taylor expand $\log\frac{\tilde{p}_\theta(y|x^{t-1})}{\tilde{p}_\theta(y|x^t)}$

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y)\prod_{t=1}^T p_\theta(x^{t-1}|x^t, y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde{p}_\theta(x^{t-1}|x^t, y) := \mathcal{N}(x^{t-1}|x^t + \sigma^2[s_\theta(x^t) + \nabla_{x^t} \log \tilde{p}_\theta(y|x^t)], \sigma^2)$
  where $\tilde{p}_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat{x}_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

Three approximations to derive $\tilde{p}_\theta(x^{t-1} \mid x^t, y)$:

$$p_\theta(x^{t-1} \mid x^t, y) \overset{1}{\approx} p_\theta(x^{t-1} \mid x^t)\tilde{p}_\theta(y \mid x^{t-1})/\tilde{p}_\theta(y \mid x^t)$$
$$\overset{2}{\approx} p_\theta(x^{t-1} \mid x^t) \exp\{(x^{t-1} - x^t)\nabla_{x^t} \log \tilde{p}_\theta(y \mid x^t)\}$$

1. Likelihood approximation ($\tilde{p}_\theta(y|x^0) = p(y|x^0)$ if $\hat{x}_\theta(x^0) = x^0$)
2. Taylor expand $\log \frac{\tilde{p}_\theta(y|x^{t-1})}{\tilde{p}_\theta(y|x^t)}$, $O((x^{t-1} - x^t)^2)$ error

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y)\prod_{t=1}^T p_\theta(x^{t-1}|x^t, y)$

Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde{p}_\theta(x^{t-1}|x^t, y) := \mathcal{N}(x^{t-1}|x^t + \sigma^2[s_\theta(x^t) + \nabla_{x^t}\log\tilde{p}_\theta(y|x^t)], \sigma^2)$ where $\tilde{p}_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat{x}_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

Three approximations to derive $\tilde{p}_\theta(x^{t-1} \mid x^t, y)$:

$$p_\theta(x^{t-1} \mid x^t, y) \overset{1}{\approx} p_\theta(x^{t-1} \mid x^t)\tilde{p}_\theta(y \mid x^{t-1})/\tilde{p}_\theta(y \mid x^t)$$
$$\overset{2}{\approx} p_\theta(x^{t-1} \mid x^t)\exp\{(x^{t-1} - x^t)\nabla_{x^t}\log\tilde{p}_\theta(y \mid x^t)\}$$

1. Likelihood approximation ($\tilde{p}_\theta(y|x^0) = p(y|x^0)$ if $\hat{x}_\theta(x^0) = x^0$)
2. Taylor expand $\log\frac{\tilde{p}_\theta(y|x^{t-1})}{\tilde{p}_\theta(y|x^t)}$, $O((x^{t-1} - x^t)^2)$ error

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y)\prod_{t=1}^T p_\theta(x^{t-1}|x^t, y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde{p}_\theta(x^{t-1}|x^t, y) := \mathcal{N}(x^{t-1}|x^t + \sigma^2[s_\theta(x^t) + \nabla_{x^t}\log\tilde{p}_\theta(y|x^t)], \sigma^2)$ where $\tilde{p}_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat{x}_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

Three approximations to derive $\tilde{p}_\theta(x^{t-1} \mid x^t, y)$:

$$p_\theta(x^{t-1} \mid x^t, y) \overset{1}{\approx} p_\theta(x^{t-1} \mid x^t)\tilde{p}_\theta(y \mid x^{t-1})/\tilde{p}_\theta(y \mid x^t)$$
$$\overset{2}{\approx} p_\theta(x^{t-1} \mid x^t)\exp\{(x^{t-1} - x^t)\nabla_{x^t}\log\tilde{p}_\theta(y \mid x^t)\}$$

1. Likelihood approximation ($\tilde{p}_\theta(y|x^0) = p(y|x^0)$ if $\hat{x}_\theta(x^0) = x^0$)
2. Taylor expand $\log\frac{\tilde{p}_\theta(y|x^{t-1})}{\tilde{p}_\theta(y|x^t)}$, $O((x^{t-1} - x^t)^2)$ error
3. "Twist" and complete the square

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y)=p_\theta(x^T|y)\prod_{t=1}^T p_\theta(x^{t-1}|x^t,y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde{p}_\theta(x^{t-1}|x^t,y):=\mathcal{N}(x^{t-1}|x^t+\sigma^2[s_\theta(x^t)+\nabla_{x^t}\log\tilde{p}_\theta(y|x^t)],\sigma^2)$
  where $\tilde{p}_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat{x}_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

Three approximations to derive $\tilde{p}_\theta(x^{t-1} \mid x^t, y)$:

$$
\begin{aligned}
p_\theta(x^{t-1} \mid x^t, y) &\overset{1}{\approx} p_\theta(x^{t-1} \mid x^t)\tilde{p}_\theta(y \mid x^{t-1})/\tilde{p}_\theta(y \mid x^t) \\
&\overset{2}{\approx} p_\theta(x^{t-1} \mid x^t)\exp\{(x^{t-1}-x^t)\nabla_{x^t}\log\tilde{p}_\theta(y \mid x^t)\} \\
&\overset{3}{\approx} \mathcal{N}(x^{t-1}|x^t+\sigma^2[s_\theta(x^t)+\nabla_{x^t}\log\tilde{p}_\theta(y|x^t)],\sigma^2)
\end{aligned}
$$

1. Likelihood approximation ($\tilde{p}_\theta(y|x^0)=p(y|x^0)$ if $\hat{x}_\theta(x^0)=x^0$)
2. Taylor expand $\log\frac{\tilde{p}_\theta(y|x^{t-1})}{\tilde{p}_\theta(y|x^t)}$, $O((x^{t-1}-x^t)^2)$ error
3. "Twist" and complete the square

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y) \prod_{t=1}^T p_\theta(x^{t-1}|x^t, y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde{p}_\theta(x^{t-1}|x^t, y) := \mathcal{N}(x^{t-1}|x^t + \sigma^2[s_\theta(x^t) + \nabla_{x^t} \log \tilde{p}_\theta(y|x^t)], \sigma^2)$
  where $\tilde{p}_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat{x}_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

Three approximations to derive $\tilde{p}_\theta(x^{t-1} \mid x^t, y)$:

$$
\begin{aligned}
p_\theta(x^{t-1} \mid x^t, y) &\overset{1}{\approx} p_\theta(x^{t-1} \mid x^t)\tilde{p}_\theta(y \mid x^{t-1})/\tilde{p}_\theta(y \mid x^t) \\
&\overset{2}{\approx} p_\theta(x^{t-1} \mid x^t) \exp\{(x^{t-1} - x^t)\nabla_{x^t} \log \tilde{p}_\theta(y \mid x^t)\} \\
&\overset{3}{\approx} \mathcal{N}(x^{t-1}|x^t + \sigma^2[s_\theta(x^t) + \nabla_{x^t} \log \tilde{p}_\theta(y|x^t)], \sigma^2)
\end{aligned}
$$

1. Likelihood approximation ($\tilde{p}_\theta(y|x^0) = p(y|x^0)$ if $\hat{x}_\theta(x^0) = x^0$)
2. Taylor expand $\log \frac{\tilde{p}_\theta(y|x^{t-1})}{\tilde{p}_\theta(y|x^t)}$, $O((x^{t-1} - x^t)^2)$ error
3. "Twist" and complete the square, $O(\sigma^2)$ error

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y) = p_\theta(x^T|y)\prod_{t=1}^T p_\theta(x^{t-1}|x^t, y)$
Challenge: intractable, approximate with some $\tilde{p}_\theta(x^{0:T}|y)$:

- $\tilde{p}_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde{p}_\theta(x^{t-1}|x^t, y) := \mathcal{N}(x^{t-1}|x^t + \sigma^2[s_\theta(x^t) + \nabla_{x^t}\log\tilde{p}_\theta(y|x^t)], \sigma^2)$
  where $\tilde{p}_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat{x}_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

Three approximations to derive $\tilde{p}_\theta(x^{t-1} \mid x^t, y)$:

$$
\begin{aligned}
p_\theta(x^{t-1} \mid x^t, y) &\overset{1}{\approx} p_\theta(x^{t-1} \mid x^t)\tilde{p}_\theta(y \mid x^{t-1})/\tilde{p}_\theta(y \mid x^t) \\
&\overset{2}{\approx} p_\theta(x^{t-1} \mid x^t)\exp\{(x^{t-1} - x^t)\nabla_{x^t}\log\tilde{p}_\theta(y \mid x^t)\} \\
&\overset{3}{\approx} \mathcal{N}(x^{t-1}|x^t + \sigma^2[s_\theta(x^t) + \nabla_{x^t}\log\tilde{p}_\theta(y|x^t)], \sigma^2)
\end{aligned}
$$

1. Likelihood approximation ($\tilde{p}_\theta(y|x^0) = p(y|x^0)$ if $\hat{x}_\theta(x^0) = x^0$)
2. Taylor expand $\log\frac{\tilde{p}_\theta(y|x^{t-1})}{\tilde{p}_\theta(y|x^t)}$, $O((x^{t-1} - x^t)^2)$ error
3. "Twist" and complete the square, $O(\sigma^2)$ error

Twisted proposal: $\tilde{p}_\theta(x^{0:T}|y) = \tilde{p}_\theta(x^T|y)\prod_{t=1}^T \tilde{p}_\theta(x^{t-1}|x^t, y)$

A Better *Twisted* Proposal for Importance Sampling

Ideal proposal: $p_\theta(x^{0:T}|y)=p_\theta(x^T|y)\prod_{t=1}^T p_\theta(x^{t-1}|x^t,y)$
Challenge: intractable, approximate with some $\tilde p_\theta(x^{0:T}|y)$:

- $\tilde p_\theta(x^T|y) := p_\theta(x^T) \approx p_\theta(x^T \mid y)$
- $\tilde p_\theta(x^{t-1}|x^t,y):=\mathcal{N}(x^{t-1}|x^t+\sigma^2[s_\theta(x^t)+\nabla_{x^t}\log\tilde p_\theta(y|x^t)],\sigma^2)$
  where $\tilde p_\theta(y \mid x^t) = p_{y|x^0}(y \mid \hat x_\theta(x^t)) \approx p_\theta(y \mid x^t)$.

Three approximations to derive $\tilde p_\theta(x^{t-1} \mid x^t, y)$:

$$
\begin{aligned}
p_\theta(x^{t-1} \mid x^t, y) &\overset{1}{\approx} p_\theta(x^{t-1} \mid x^t)\tilde p_\theta(y \mid x^{t-1})/\tilde p_\theta(y \mid x^t)\\
&\overset{2}{\approx} p_\theta(x^{t-1} \mid x^t)\exp\{(x^{t-1}-x^t)\nabla_{x^t}\log\tilde p_\theta(y \mid x^t)\}\\
&\overset{3}{\approx} \mathcal{N}(x^{t-1}|x^t+\sigma^2[s_\theta(x^t)+\nabla_{x^t}\log\tilde p_\theta(y|x^t)],\sigma^2)
\end{aligned}
$$

1. Likelihood approximation ($\tilde p_\theta(y|x^0)=p(y|x^0)$ if $\hat x_\theta(x^0)=x^0$)
2. Taylor expand $\log\frac{\tilde p_\theta(y|x^{t-1})}{\tilde p_\theta(y|x^t)}$, $O((x^{t-1}-x^t)^2)$ error
3. "Twist" and complete the square, $O(\sigma^2)$ error

Twisted proposal: $\tilde p_\theta(x^{0:T}|y) = \tilde p_\theta(x^T|y)\prod_{t=1}^T \tilde p_\theta(x^{t-1}|x^t,y)$

Importance weights: $w(x^{0:T}) = p_\theta(x^{0:T}, y)/\tilde p_\theta(x^{0:T} \mid y)$

Sequential Monte Carlo & The Twisted Diffusion Sampler

Problem: Error adds up

Sequential Monte Carlo & The Twisted Diffusion Sampler

Problem: Error adds up $\implies \mathrm{KL}(p_\theta(x^0 \mid y) \,||\, \tilde{p}_\theta(x^0 \mid y)) >> 0.$

Sequential Monte Carlo & The Twisted Diffusion Sampler

Problem: Error adds up $\implies \mathrm{KL}(p_\theta(x^0 \mid y) \,||\, \tilde{p}_\theta(x^0 \mid y)) >> 0$.
Idea: Gradually correct error as $t \to 0$

Sequential Monte Carlo & The Twisted Diffusion Sampler

Problem: Error adds up $\implies \mathrm{KL}(p_\theta(x^0 \mid y) \,||\, \tilde{p}_\theta(x^0 \mid y)) >> 0$.
Idea: Gradually correct error as $t \to 0$ (& $\tilde{p}_\theta(y|x^t) \to p_\theta(y|x^t)$)

Sequential Monte Carlo & The Twisted Diffusion Sampler

Problem: Error adds up $\implies \mathrm{KL}(p_\theta(x^0 \mid y) \,||\, \tilde{p}_\theta(x^0 \mid y)) >> 0.$
Idea: Gradually correct error as $t \to 0$ (& $\tilde{p}_\theta(y|x^t) \to p_\theta(y|x^t)$)
Sequential Monte Carlo (SMC) ingredients

- Series of targets: $\nu_t(x^{t:T})$ with $\nu_0(x^{0:T}) = p_\theta(x^{0:T} \mid y)$

Sequential Monte Carlo & The Twisted Diffusion Sampler

Problem: Error adds up $\implies \mathrm{KL}(p_\theta(x^0 \mid y) \mid\mid \tilde{p}_\theta(x^0 \mid y)) >> 0$.
Idea: Gradually correct error as $t \to 0$ (& $\tilde{p}_\theta(y|x^t) \to p_\theta(y|x^t)$)
Sequential Monte Carlo (SMC) ingredients

- Series of targets: $\nu_t(x^{t:T})$ with $\nu_0(x^{0:T}) = p_\theta(x^{0:T} \mid y)$
- Proposals: $\tilde{p}(x^T)$ and $\tilde{p}(x^t \mid x^{t+1})$ for $t = 1, \ldots, T-1$

Sequential Monte Carlo & The Twisted Diffusion Sampler

Problem: Error adds up $\implies \mathrm{KL}(p_\theta(x^0 \mid y) \mid\mid \tilde{p}_\theta(x^0 \mid y)) >> 0$.
Idea: Gradually correct error as $t \to 0$ (& $\tilde{p}_\theta(y|x^t) \to p_\theta(y|x^t)$)
Sequential Monte Carlo (SMC) ingredients

- Series of targets: $\nu_t(x^{t:T})$ with $\nu_0(x^{0:T}) = p_\theta(x^{0:T} \mid y)$
- Proposals: $\tilde{p}(x^T)$ and $\tilde{p}(x^t \mid x^{t+1})$ for $t = 1, \ldots, T-1$
  - "Extend" targets as $\nu_t(x^{0:T}) \propto \nu_t(x^{t:T})\tilde{p}(x^{0:t-1} \mid x^t)$

Sequential Monte Carlo & The Twisted Diffusion Sampler

Problem: Error adds up $\implies \mathrm{KL}(p_\theta(x^0 \mid y) \mid\mid \tilde{p}_\theta(x^0 \mid y)) >> 0$.
Idea: Gradually correct error as $t \to 0$ (& $\tilde{p}_\theta(y|x^t) \to p_\theta(y|x^t)$)
Sequential Monte Carlo (SMC) ingredients

- Series of targets: $\nu_t(x^{t:T})$ with $\nu_0(x^{0:T}) = p_\theta(x^{0:T} \mid y)$
- Proposals: $\tilde{p}(x^T)$ and $\tilde{p}(x^t \mid x^{t+1})$ for $t = 1, \ldots, T-1$
  - "Extend" targets as $\nu_t(x^{0:T}) \propto \nu_t(x^{t:T})\tilde{p}(x^{0:t-1} \mid x^t)$
- *Resample* at each t with weights: $w_T(x^T) \propto \nu_T(x^T)/\tilde{p}(x^T)$ and $w_t(x^t, x^{t+1}) \propto \nu_t(x^t, x^{t+1})/\nu_{t+1}(x^t, x^{t+1})$

Sequential Monte Carlo & The Twisted Diffusion Sampler

Problem: Error adds up $\implies \mathrm{KL}(p_\theta(x^0 \mid y) \,||\, \tilde{p}_\theta(x^0 \mid y)) >> 0.$
Idea: Gradually correct error as $t \to 0$ (& $\tilde{p}_\theta(y|x^t) \to p_\theta(y|x^t)$)
Sequential Monte Carlo (SMC) ingredients

- Series of targets: $\nu_t(x^{t:T})$ with $\nu_0(x^{0:T}) = p_\theta(x^{0:T} \mid y)$
- Proposals: $\tilde{p}(x^T)$ and $\tilde{p}(x^t \mid x^{t+1})$ for $t = 1, \ldots, T-1$
  - "Extend" targets as $\nu_t(x^{0:T}) \propto \nu_t(x^{t:T})\tilde{p}(x^{0:t-1} \mid x^t)$
- *Resample* at each t with weights: $w_T(x^T) \propto \nu_T(x^T)/\tilde{p}(x^T)$ and $w_t(x^t, x^{t+1}) \propto \nu_t(x^t, x^{t+1})/\nu_{t+1}(x^t, x^{t+1})$

Twisted Diffusion Sampler (TDS) is an SMC sampler with:

- $\nu_t(x^{t:T}) \propto p_\theta(x^{t:T})\tilde{p}_\theta(y \mid x^t)$

Sequential Monte Carlo & The Twisted Diffusion Sampler

Problem: Error adds up $\implies \mathrm{KL}(p_\theta(x^0 \mid y) \,||\, \tilde{p}_\theta(x^0 \mid y)) >> 0$.
Idea: Gradually correct error as $t \to 0$ (& $\tilde{p}_\theta(y|x^t) \to p_\theta(y|x^t)$)
Sequential Monte Carlo (SMC) ingredients

- Series of targets: $\nu_t(x^{t:T})$ with $\nu_0(x^{0:T}) = p_\theta(x^{0:T} \mid y)$
- Proposals: $\tilde{p}(x^T)$ and $\tilde{p}(x^t \mid x^{t+1})$ for $t = 1, \ldots, T-1$
  - "Extend" targets as $\nu_t(x^{0:T}) \propto \nu_t(x^{t:T})\tilde{p}(x^{0:t-1} \mid x^t)$
- *Resample* at each t with weights: $w_T(x^T) \propto \nu_T(x^T)/\tilde{p}(x^T)$ and $w_t(x^t, x^{t+1}) \propto \nu_t(x^t, x^{t+1})/\nu_{t+1}(x^t, x^{t+1})$

Twisted Diffusion Sampler (TDS) is an SMC sampler with:

- $\nu_t(x^{t:T}) \propto p_\theta(x^{t:T})\tilde{p}_\theta(y \mid x^t)$
- $\tilde{p}(x^T) = p_\theta(x^T)$ and $\tilde{p}(x^t \mid x^{t+1}) = \tilde{p}_\theta(x^t \mid x^{t+1}, y)$

Sequential Monte Carlo & The Twisted Diffusion Sampler

Problem: Error adds up $\implies \mathrm{KL}(p_\theta(x^0 \mid y) \,||\, \tilde{p}_\theta(x^0 \mid y)) >> 0$.
Idea: Gradually correct error as $t \to 0$ (& $\tilde{p}_\theta(y|x^t) \to p_\theta(y|x^t)$)
Sequential Monte Carlo (SMC) ingredients

- Series of targets: $\nu_t(x^{t:T})$ with $\nu_0(x^{0:T}) = p_\theta(x^{0:T} \mid y)$
- Proposals: $\tilde{p}(x^T)$ and $\tilde{p}(x^t \mid x^{t+1})$ for $t = 1, \ldots, T-1$
  - "Extend" targets as $\nu_t(x^{0:T}) \propto \nu_t(x^{t:T})\tilde{p}(x^{0:t-1} \mid x^t)$
- *Resample* at each t with weights: $w_T(x^T) \propto \nu_T(x^T)/\tilde{p}(x^T)$ and $w_t(x^t, x^{t+1}) \propto \nu_t(x^t, x^{t+1})/\nu_{t+1}(x^t, x^{t+1})$

Twisted Diffusion Sampler (TDS) is an SMC sampler with:

- $\nu_t(x^{t:T}) \propto p_\theta(x^{t:T})\tilde{p}_\theta(y \mid x^t)$
- $\tilde{p}(x^T) = p_\theta(x^T)$ and $\tilde{p}(x^t \mid x^{t+1}) = \tilde{p}_\theta(x^t \mid x^{t+1}, y)$
- $w_t(x^t, x^{t+1}) = p_\theta(x^t|x^{t+1})\tilde{p}_\theta(y|x^t)/[\tilde{p}_\theta(x^t|x^{t+1}, y)\tilde{p}_\theta(y|x^{t+1})]$

The Twisted Diffusion Sampler (TDS)

Algorithm 1: Twisted Diffusion Sampler

$x_k^T \sim \mathcal{N}(0, T\sigma^2)$

$w_k \leftarrow \tilde{p}_k^T = p_{y|x^0}(y \mid \hat{x}_\theta(x_k^T))$

for $t = T, \cdots, 1$ **do**

$\{x_k^t, \tilde{p}_k^t\} \sim \mathrm{Multinomial}\left(\{x_k^t, \tilde{p}_k^t\}; \{w_k\}\right)$

$x_k^{t-1} \sim \tilde{p}_\theta(\cdot|x_k^t, y) = \mathcal{N}\left(x_k^t + \sigma^2[s_\theta(x_k^t) + \nabla_{x_k^t} \log \tilde{p}_k^t], \sigma^2\right)$

$\tilde{p}_k^{t-1} \leftarrow p_{y|x^0}(y \mid \hat{x}_\theta(x_k^{t-1}))$

$w_k \leftarrow [p_\theta(x_k^{t-1} \mid x_k^t) \cdot \tilde{p}_k^{t-1}]/[\tilde{p}_\theta(x_k^{t-1} \mid x_k^t, y) \cdot \tilde{p}_k^t]$

Return $\{w_k\}, \{x_k^0\}$

Roadmap

- Diffusion models and conditional generation
- The Twisted Diffusion Sampler (TDS)
- Related work
- Properties of TDS (Theory and Simulations)
- Case study in motif-scaffolding

Related Work

Previous works do not accurately approximate conditional distributions of unconditional models.

Related Work

Previous works do not accurately approximate conditional distributions of unconditional models.

- **Conditional Training**: Used for text-to-image and protein design.

Related Work

Previous works do not accurately approximate conditional distributions of unconditional models.

- **Conditional Training**: Used for text-to-image and protein design. Expensive: data curation, engineering time, compute.

Related Work

Previous works do not accurately approximate conditional distributions of unconditional models.

- **Conditional Training**: Used for text-to-image and protein design. Expensive: data curation, engineering time, compute.
- **Replacement guidance** [Song et al., 2020]: Widely used for image inpainting.

Related Work

Previous works do not accurately approximate conditional distributions of unconditional models.

- **Conditional Training**: Used for text-to-image and protein design. Expensive: data curation, engineering time, compute.
- **Replacement guidance** [Song et al., 2020]: Widely used for image inpainting. No guarantees, applies only to inpainting

Related Work

Previous works do not accurately approximate conditional distributions of unconditional models.

- **Conditional Training**: Used for text-to-image and protein design. Expensive: data curation, engineering time, compute.
- **Replacement guidance** [Song et al., 2020]: Widely used for image inpainting. No guarantees, applies only to inpainting
- **Reconstruction guidance** [Ho et al., 2022]: Backprop through the denoising model.

Related Work

Previous works do not accurately approximate conditional distributions of unconditional models.

- **Conditional Training**: Used for text-to-image and protein design. Expensive: data curation, engineering time, compute.
- **Replacement guidance** [Song et al., 2020]: Widely used for image inpainting. No guarantees, applies only to inpainting
- **Reconstruction guidance** [Ho et al., 2022]: Backprop through the denoising model. No guarantees, can perform poorly [Zhang et al., 2023]

Related Work

Previous works do not accurately approximate conditional distributions of unconditional models.

- **Conditional Training**: Used for text-to-image and protein design. Expensive: data curation, engineering time, compute.
- **Replacement guidance** [Song et al., 2020]: Widely used for image inpainting. No guarantees, applies only to inpainting
- **Reconstruction guidance** [Ho et al., 2022]: Backprop through the denoising model. No guarantees, can perform poorly [Zhang et al., 2023]
- **Twisted SMC** [Whiteley and Lee, 2014, Guarniero et al., 2017, Heng et al., 2020]: Modify SMC targets and proposals for improved efficiency.

Related Work

Previous works do not accurately approximate conditional distributions of unconditional models.

- **Conditional Training**: Used for text-to-image and protein design. Expensive: data curation, engineering time, compute.
- **Replacement guidance** [Song et al., 2020]: Widely used for image inpainting. No guarantees, applies only to inpainting
- **Reconstruction guidance** [Ho et al., 2022]: Backprop through the denoising model. No guarantees, can perform poorly [Zhang et al., 2023]
- **Twisted SMC** [Whiteley and Lee, 2014, Guarniero et al., 2017, Heng et al., 2020]: Modify SMC targets and proposals for improved efficiency. Requires a twisting function — (choosing it is a contribution of this work)

Related Work

Previous works do not accurately approximate conditional distributions of unconditional models.

- **Conditional Training**: Used for text-to-image and protein design. Expensive: data curation, engineering time, compute.
- **Replacement guidance** [Song et al., 2020]: Widely used for image inpainting. No guarantees, applies only to inpainting
- **Reconstruction guidance** [Ho et al., 2022]: Backprop through the denoising model. No guarantees, can perform poorly [Zhang et al., 2023]
- **Twisted SMC** [Whiteley and Lee, 2014, Guarniero et al., 2017, Heng et al., 2020]: Modify SMC targets and proposals for improved efficiency. Requires a twisting function — (choosing it is a contribution of this work)
- **SMCDiff** [Trippe et al., 2022]: asymptotically accurate for inpaininting.

Related Work

Previous works do not accurately approximate conditional distributions of unconditional models.

- **Conditional Training**: Used for text-to-image and protein design. Expensive: data curation, engineering time, compute.
- **Replacement guidance** [Song et al., 2020]: Widely used for image inpainting. No guarantees, applies only to inpainting
- **Reconstruction guidance** [Ho et al., 2022]: Backprop through the denoising model. No guarantees, can perform poorly [Zhang et al., 2023]
- **Twisted SMC** [Whiteley and Lee, 2014, Guarniero et al., 2017, Heng et al., 2020]: Modify SMC targets and proposals for improved efficiency. Requires a twisting function — (choosing it is a contribution of this work)
- **SMCDiff** [Trippe et al., 2022]: asymptotically accurate for inpaininting. But assumes $p = q$ and doesn't support general likelihoods, or use twisting.

TDS: Asymptotic Accuracy and Small Intermediate KLs

Thm 1: Let $\mathbb{P}_K = (\sum w_k)^{-1} \sum w_k \delta_{x_k^0}$ be the output of K-particle TDS. $\mathbb{P}_K$ converges weakly to $p_\theta(x^0 \mid y)$ as $K \to \infty$.

TDS: Asymptotic Accuracy and Small Intermediate KLs

Thm 1: Let $\mathbb{P}_K = (\sum w_k)^{-1} \sum w_k \delta_{x_k^0}$ be the output of K-particle TDS. $\mathbb{P}_K$ converges weakly to $p_\theta(x^0 \mid y)$ as $K \to \infty$.

- We require $\tilde{p}_\theta(y \mid x^t)$ to be smooth in x^t and bounded.

TDS: Asymptotic Accuracy and Small Intermediate KLs

Thm 1: Let $\mathbb{P}_K = (\sum w_k)^{-1} \sum w_k \delta_{x_k^0}$ be the output of K-particle TDS. $\mathbb{P}_K$ converges weakly to $p_\theta(x^0 \mid y)$ as $K \to \infty$.

- We require $\tilde{p}_\theta(y \mid x^t)$ to be smooth in x^t and bounded.

Thm 2: Set $\sigma^2 = \sigma_*^2 / T$. Then $\max_t \mathrm{KL}(\nu_t \,||\, \nu_{t+1}) < CT^{-1}$.

TDS: Asymptotic Accuracy and Small Intermediate KLs

Thm 1: Let $\mathbb{P}_K = (\sum w_k)^{-1} \sum w_k \delta_{x_k^0}$ be the output of K-particle TDS. $\mathbb{P}_K$ converges weakly to $p_\theta(x^0 \mid y)$ as $K \to \infty$.

- We require $\tilde{p}_\theta(y \mid x^t)$ to be smooth in x^t and bounded.

Thm 2: Set $\sigma^2 = \sigma_*^2 / T$. Then $\max_t \mathrm{KL}(\nu_t \mid\mid \nu_{t+1}) < CT^{-1}$.

- With more steps, fewer particles are required

TDS: Asymptotic Accuracy and Small Intermediate KLs

Thm 1: Let $\mathbb{P}_K = (\sum w_k)^{-1} \sum w_k \delta_{x_k^0}$ be the output of K-particle TDS. $\mathbb{P}_K$ converges weakly to $p_\theta(x^0 \mid y)$ as $K \to \infty$.

- We require $\tilde{p}_\theta(y \mid x^t)$ to be smooth in x^t and bounded.

Thm 2: Set $\sigma^2 = \sigma_*^2 / T$. Then $\max_t \mathrm{KL}(\nu_t \,||\, \nu_{t+1}) < CT^{-1}$.

- With more steps, fewer particles are required

Proof sketch:

$$\mathrm{KL}(\nu_t \,||\, \nu_{t+1}) = \mathbb{E}_{\nu_t}[\log \nu_t(x^{0:T}) / \nu_{t+1}(x^{0:T})]$$

TDS: Asymptotic Accuracy and Small Intermediate KLs

Thm 1: Let $\mathbb{P}_K = (\sum w_k)^{-1} \sum w_k \delta_{x_k^0}$ be the output of K-particle TDS. $\mathbb{P}_K$ converges weakly to $p_\theta(x^0 \mid y)$ as $K \to \infty$.

- We require $\tilde{p}_\theta(y \mid x^t)$ to be smooth in x^t and bounded.

Thm 2: Set $\sigma^2 = \sigma_*^2 / T$. Then $\max_t \mathrm{KL}(\nu_t \parallel \nu_{t+1}) < CT^{-1}$.

- With more steps, fewer particles are required

Proof sketch:

$$\mathrm{KL}(\nu_t \parallel \nu_{t+1}) = \mathbb{E}_{\nu_t}[\log w_t(x^t, x^{t+1})] + \log Z_{\nu_{t+1}} / Z_{\nu_t}$$

TDS: Asymptotic Accuracy and Small Intermediate KLs

Thm 1: Let $\mathbb{P}_K = (\sum w_k)^{-1} \sum w_k \delta_{x_k^0}$ be the output of K-particle TDS. $\mathbb{P}_K$ converges weakly to $p_\theta(x^0 \mid y)$ as $K \to \infty$.

- We require $\tilde{p}_\theta(y \mid x^t)$ to be smooth in x^t and bounded.

Thm 2: Set $\sigma^2 = \sigma_*^2/T$. Then $\max_t \mathrm{KL}(\nu_t \,||\, \nu_{t+1}) < CT^{-1}$.

- With more steps, fewer particles are required

Proof sketch:

$$\mathrm{KL}(\nu_t \,||\, \nu_{t+1}) = \mathbb{E}_{\nu_t}[\log w_t(x^t, x^{t+1})] + \log Z_{\nu_{t+1}}/Z_{\nu_t}$$

- $\log w_t(x^t, x^{t+1}) = O((x^t - x^{t+1})^2 + \sigma^2)$

TDS: Asymptotic Accuracy and Small Intermediate KLs

Thm 1: Let $\mathbb{P}_K = (\sum w_k)^{-1} \sum w_k \delta_{x_k^0}$ be the output of K-particle TDS. $\mathbb{P}_K$ converges weakly to $p_\theta(x^0 \mid y)$ as $K \to \infty$.

- We require $\tilde{p}_\theta(y \mid x^t)$ to be smooth in x^t and bounded.

Thm 2: Set $\sigma^2 = \sigma_*^2 / T$. Then $\max_t \mathrm{KL}(\nu_t \mid\mid \nu_{t+1}) < CT^{-1}$.

- With more steps, fewer particles are required

Proof sketch:

$$\begin{aligned}\mathrm{KL}(\nu_t \mid\mid \nu_{t+1}) &= \mathbb{E}_{\nu_t}[\log w_t(x^t, x^{t+1})] + \log Z_{\nu_{t+1}} / Z_{\nu_t} \\ &= \mathbb{E}_{\nu_t}[O((x^{t+1} - x^t)^2 + \sigma^2)] + \log Z_{\nu_{t+1}} / Z_{\nu_t}\end{aligned}$$

- $\log w_t(x^t, x^{t+1}) = O((x^t - x^{t+1})^2 + \sigma^2)$

TDS: Asymptotic Accuracy and Small Intermediate KLs

Thm 1: Let $\mathbb{P}_K = (\sum w_k)^{-1} \sum w_k \delta_{x_k^0}$ be the output of K-particle TDS. $\mathbb{P}_K$ converges weakly to $p_\theta(x^0 \mid y)$ as $K \to \infty$.

- We require $\tilde{p}_\theta(y \mid x^t)$ to be smooth in x^t and bounded.

Thm 2: Set $\sigma^2 = \sigma_*^2 / T$. Then $\max_t \mathrm{KL}(\nu_t \,||\, \nu_{t+1}) < CT^{-1}$.

- With more steps, fewer particles are required

Proof sketch:

$$\begin{aligned}\mathrm{KL}(\nu_t \,||\, \nu_{t+1}) &= \mathbb{E}_{\nu_t}[\log w_t(x^t, x^{t+1})] + \log Z_{\nu_{t+1}} / Z_{\nu_t} \\ &= \mathbb{E}_{\nu_t}[O((x^{t+1} - x^t)^2 + \sigma^2)] + \log Z_{\nu_{t+1}} / Z_{\nu_t}\end{aligned}$$

- $\log w_t(x^t, x^{t+1}) = O((x^t - x^{t+1})^2 + \sigma^2) \quad Z_{\nu_{t+1}} = Z_{\nu_t} + O(\sigma^2)$

TDS: Asymptotic Accuracy and Small Intermediate KLs

Thm 1: Let $\mathbb{P}_K = (\sum w_k)^{-1} \sum w_k \delta_{x_k^0}$ be the output of K-particle TDS. $\mathbb{P}_K$ converges weakly to $p_\theta(x^0 \mid y)$ as $K \to \infty$.

- We require $\tilde{p}_\theta(y \mid x^t)$ to be smooth in x^t and bounded.

Thm 2: Set $\sigma^2 = \sigma_*^2 / T$. Then $\max_t \mathrm{KL}(\nu_t \mid\mid \nu_{t+1}) < CT^{-1}$.

- With more steps, fewer particles are required

Proof sketch:

$$\begin{aligned}\mathrm{KL}(\nu_t \mid\mid \nu_{t+1}) &= \mathbb{E}_{\nu_t}[\log w_t(x^t, x^{t+1})] + \log Z_{\nu_{t+1}} / Z_{\nu_t} \\ &= \mathbb{E}_{\nu_t}[O((x^{t+1} - x^t)^2 + \sigma^2)] + O(\sigma^2)\end{aligned}$$

- $\log w_t(x^t, x^{t+1}) = O((x^t - x^{t+1})^2 + \sigma^2) \quad Z_{\nu_{t+1}} = Z_{\nu_t} + O(\sigma^2)$

TDS: Asymptotic Accuracy and Small Intermediate KLs

Thm 1: Let $\mathbb{P}_K = (\sum w_k)^{-1} \sum w_k \delta_{x_k^0}$ be the output of K-particle TDS. $\mathbb{P}_K$ converges weakly to $p_\theta(x^0 \mid y)$ as $K \to \infty$.

- We require $\tilde{p}_\theta(y \mid x^t)$ to be smooth in x^t and bounded.

Thm 2: Set $\sigma^2 = \sigma_*^2 / T$. Then $\max_t \mathrm{KL}(\nu_t \mid\mid \nu_{t+1}) < CT^{-1}$.

- With more steps, fewer particles are required

Proof sketch:

$$\begin{aligned}
\mathrm{KL}(\nu_t \mid\mid \nu_{t+1}) &= \mathbb{E}_{\nu_t}[\log w_t(x^t, x^{t+1})] + \log Z_{\nu_{t+1}} / Z_{\nu_t} \\
&= \mathbb{E}_{\nu_t}[O((x^{t+1} - x^t)^2 + \sigma^2)] + O(\sigma^2) \\
&= O(\sigma^2) + O(\sigma^2)
\end{aligned}$$

- $\log w_t(x^t, x^{t+1}) = O((x^t - x^{t+1})^2 + \sigma^2) \quad Z_{\nu_{t+1}} = Z_{\nu_t} + O(\sigma^2)$

TDS: Asymptotic Accuracy and Small Intermediate KLs

Thm 1: Let $\mathbb{P}_K = (\sum w_k)^{-1} \sum w_k \delta_{x_k^0}$ be the output of K-particle TDS. $\mathbb{P}_K$ converges weakly to $p_\theta(x^0 \mid y)$ as $K \to \infty$.

- We require $\tilde{p}_\theta(y \mid x^t)$ to be smooth in x^t and bounded.

Thm 2: Set $\sigma^2 = \sigma_*^2 / T$. Then $\max_t \mathrm{KL}(\nu_t \mid\mid \nu_{t+1}) < CT^{-1}$.

- With more steps, fewer particles are required

Proof sketch:

$$
\begin{aligned}
\mathrm{KL}(\nu_t \mid\mid \nu_{t+1}) &= \mathbb{E}_{\nu_t}[\log w_t(x^t, x^{t+1})] + \log Z_{\nu_{t+1}} / Z_{\nu_t} \\
&= \mathbb{E}_{\nu_t}[O((x^{t+1} - x^t)^2 + \sigma^2)] + O(\sigma^2) \\
&= O(\sigma^2) + O(\sigma^2) = O(\sigma^2)
\end{aligned}
$$

- $\log w_t(x^t, x^{t+1}) = O((x^t - x^{t+1})^2 + \sigma^2) \quad Z_{\nu_{t+1}} = Z_{\nu_t} + O(\sigma^2)$

TDS: Asymptotic Accuracy and Small Intermediate KLs

Thm 1: Let $\mathbb{P}_K = (\sum w_k)^{-1} \sum w_k \delta_{x_k^0}$ be the output of K-particle TDS. $\mathbb{P}_K$ converges weakly to $p_\theta(x^0 \mid y)$ as $K \to \infty$.

- We require $\tilde{p}_\theta(y \mid x^t)$ to be smooth in x^t and bounded.

Thm 2: Set $\sigma^2 = \sigma_*^2 / T$. Then $\max_t \mathrm{KL}(\nu_t \,||\, \nu_{t+1}) < CT^{-1}$.

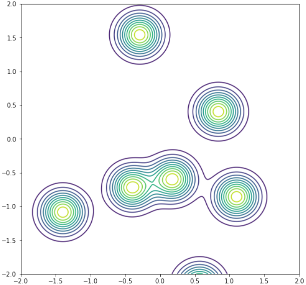
- With more steps, fewer particles are required

Proof sketch:

$$
\begin{aligned}
\mathrm{KL}(\nu_t \,||\, \nu_{t+1}) &= \mathbb{E}_{\nu_t}[\log w_t(x^t, x^{t+1})] + \log Z_{\nu_{t+1}} / Z_{\nu_t} \\
&= \mathbb{E}_{\nu_t}[O((x^{t+1} - x^t)^2 + \sigma^2)] + O(\sigma^2) \\
&= O(\sigma^2) + O(\sigma^2) = O(\sigma^2) = O(T^{-1})
\end{aligned}
$$

- $\log w_t(x^t, x^{t+1}) = O((x^t - x^{t+1})^2 + \sigma^2) \quad Z_{\nu_{t+1}} = Z_{\nu_t} + O(\sigma^2)$

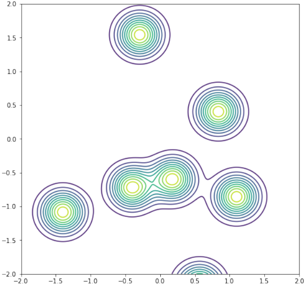
Simulation study



Gaussian mixture $q(x^0)$

- Tractable score & ground truth

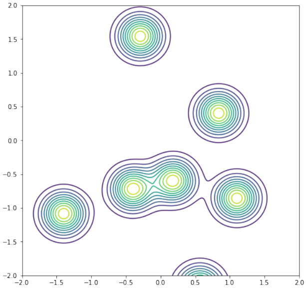
Simulation study



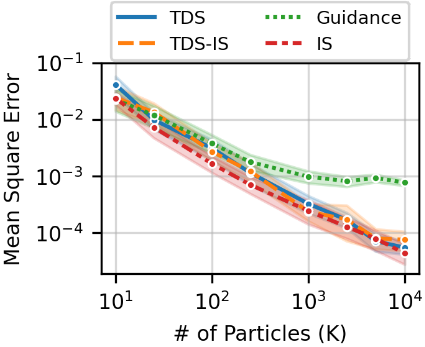
Gaussian mixture $q(x^0)$

- Tractable score & ground truth
- $y \sim \mathrm{Laplace}(\|x^0\|_2, 1)$

Simulation study



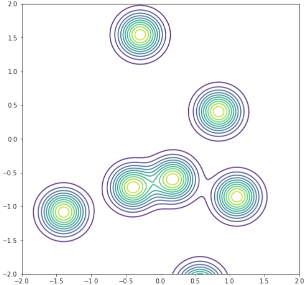
Gaussian mixture $q(x^0)$



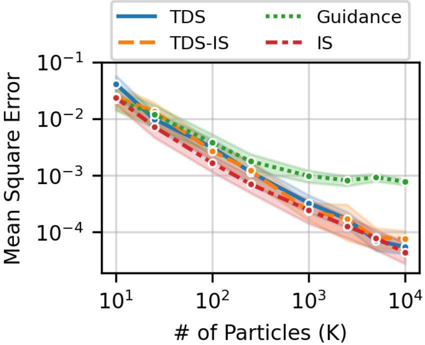
- Tractable score & ground truth
- $y \sim \text{Laplace}(\|x^0\|_2, 1)$

- Estimand: $\mathbb{E}[x^0 \mid y = 0]$

Simulation study



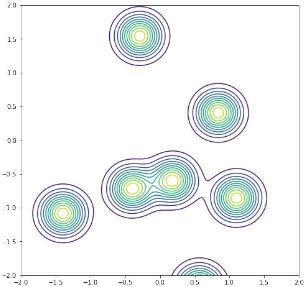
Gaussian mixture $q(x^0)$



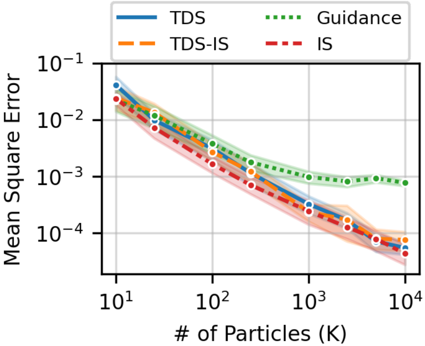
- Tractable score & ground truth
- $y \sim \mathrm{Laplace}(\|x^0\|_2, 1)$

- Estimand: $\mathbb{E}[x^0 \mid y = 0]$
- $O(K^{-1})$ convergence for TDS, and IS

Simulation study



Gaussian mixture $q(x^0)$



- Tractable score & ground truth
- $y \sim \mathrm{Laplace}(\|x^0\|_2, 1)$

- Estimand: $\mathbb{E}[x^0 \mid y = 0]$
- $O(K^{-1})$ convergence for TDS, and IS
- Guidance is biased.

MNIST class-conditional generation

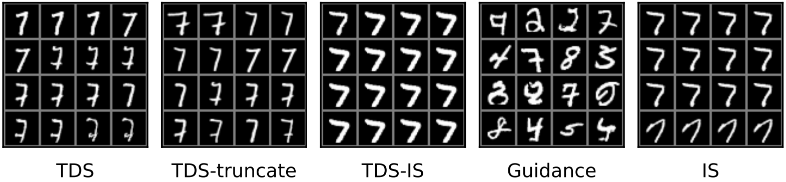
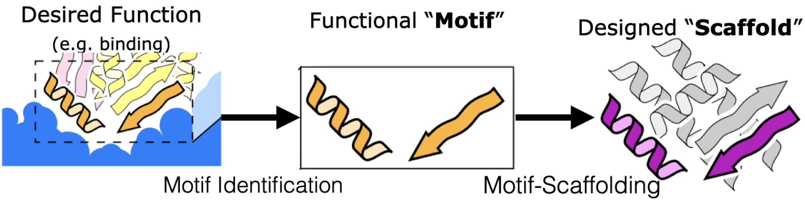


Figure: Approximate conditional samples for class $y = 7$.

Protein Design Case Study: The Motif-Scaffolding Problem

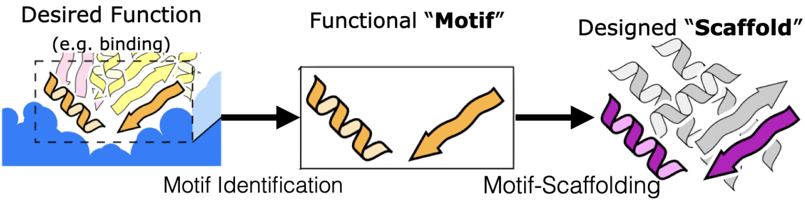
Common protein design workflow[1]



[1] figure credit to David Juergens and Doug Tischer

Protein Design Case Study: The Motif-Scaffolding Problem

Common protein design workflow[1]

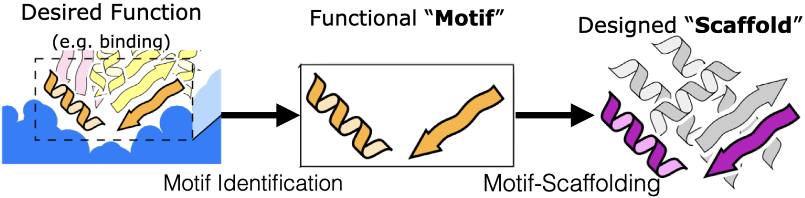


- Used to design vaccines, enzymes [Procko et al., 2014, Jiang et al., 2008]

[1]figure credit to David Juergens and Doug Tischer

Protein Design Case Study: The Motif-Scaffolding Problem

Common protein design workflow[1]

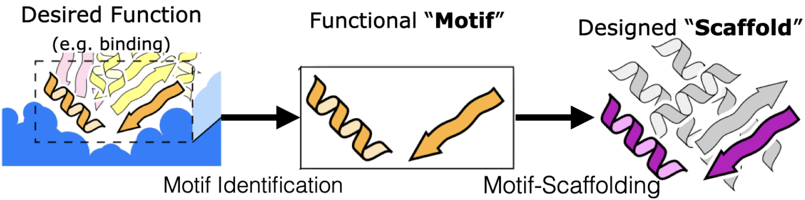


- Used to design vaccines, enzymes [Procko et al., 2014, Jiang et al., 2008]
- AlphaFold provides validation predictive of experimental success [Wang et al., 2022]

[1] figure credit to David Juergens and Doug Tischer

Protein Design Case Study: The Motif-Scaffolding Problem

Common protein design workflow[1]



- Used to design vaccines, enzymes [Procko et al., 2014, Jiang et al., 2008]
- AlphaFold provides validation predictive of experimental success [Wang et al., 2022]
- Recent progress with ML methods [Trippe et al., 2022, Watson et al., 2022], but problem remains open

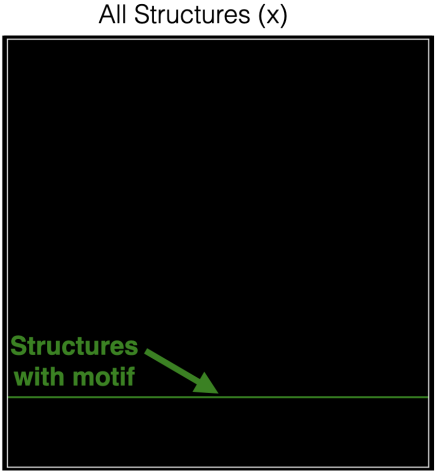
[1] figure credit to David Juergens and Doug Tischer

Protein Design Case Study: The Motif-Scaffolding Problem

What makes this problem hard?

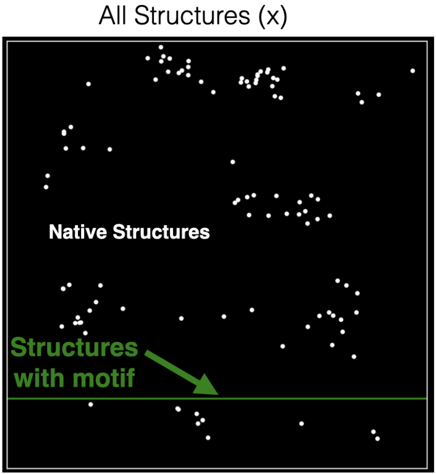
Protein Design Case Study: The Motif-Scaffolding Problem

What makes this problem hard?



Protein Design Case Study: The Motif-Scaffolding Problem

What makes this problem hard?



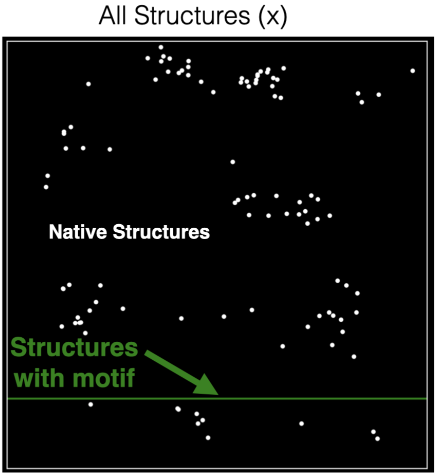
Protein Design Case Study: The Motif-Scaffolding Problem

What makes this problem hard?

Conditional generative modeling approach [Trippe et al., 2022]

1. Fit $p_\theta(x)$ to structures of native proteins.

2. Sample $x \sim p_\theta(x|y)$, for $p_\theta(x,y) = p_\theta(x)\delta_y(x_M)$



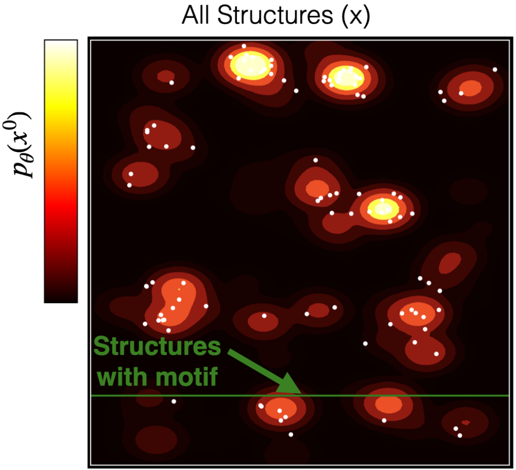
Protein Design Case Study: The Motif-Scaffolding Problem

What makes this problem hard?

Conditional generative modeling approach [Trippe et al., 2022]

1. Fit $p_\theta(x)$ to structures of native proteins.

2. Sample $x \sim p_\theta(x|y)$, for $p_\theta(x, y) = p_\theta(x)\delta_y(x_M)$



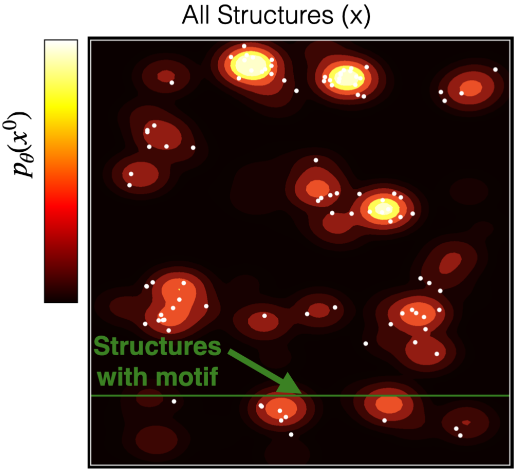
Protein Design Case Study: The Motif-Scaffolding Problem

What makes this problem hard?

Conditional generative modeling approach [Trippe et al., 2022]

1. Fit $p_\theta(x)$ to structures of native proteins.

2. Sample $x \sim p_\theta(x|y)$, for $p_\theta(x, y) = p_\theta(x)\delta_y(x_M)$



Intuition: If $p_\theta(x) > 0$ only if x is a "real" molecule, then $p_\theta(x \mid y) > 0$ only if x is a "real" molecule containing y.

The Motif-Scaffolding Problem Presents Challenges

- Protein structures live in 3D space and must conform to physical constraints

The Motif-Scaffolding Problem Presents Challenges

- Protein structures live in 3D space and must conform to physical constraints → **we use equivariant graph neural networks designed for proteins** [Baek and Baker, 2022, Jumper et al., 2021]

The Motif-Scaffolding Problem Presents Challenges

- Protein structures live in 3D space and must conform to physical constraints $\rightarrow$ **we use equivariant graph neural networks designed for proteins** [Baek and Baker, 2022, Jumper et al., 2021]
- Convenient representations are non-Euclidean

The Motif-Scaffolding Problem Presents Challenges

- Protein structures live in 3D space and must conform to physical constraints $\rightarrow$ **we use equivariant graph neural networks designed for proteins** [Baek and Baker, 2022, Jumper et al., 2021]
- Convenient representations are non-Euclidean $\rightarrow$ Riemannian diffusion [De Bortoli et al., 2022]

The Motif-Scaffolding Problem Presents Challenges

- Protein structures live in 3D space and must conform to physical constraints → **we use equivariant graph neural networks designed for proteins** [Baek and Baker, 2022, Jumper et al., 2021]
- Convenient representations are non-Euclidean → Riemannian diffusion [De Bortoli et al., 2022]
  - **We use an $SE(3)^N$ diffusion model [Yim et al., 2023]**

The Motif-Scaffolding Problem Presents Challenges

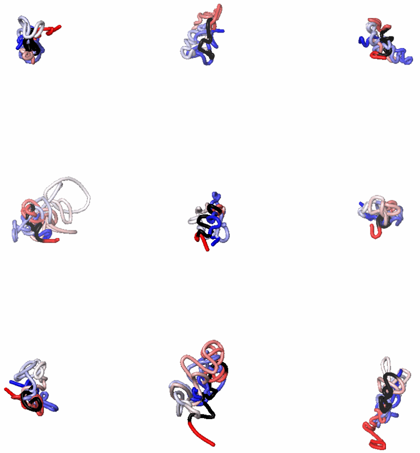
- Protein structures live in 3D space and must conform to physical constraints $\rightarrow$ **we use equivariant graph neural networks designed for proteins** [Baek and Baker, 2022, Jumper et al., 2021]
- Convenient representations are non-Euclidean $\rightarrow$ Riemannian diffusion [De Bortoli et al., 2022]
  - **We use an $SE(3)^N$ diffusion model [Yim et al., 2023]**
- Extra degrees of freedom that are difficult to choose
  - Indices of motif within chain
  - Rotation & translation of motif

The Motif-Scaffolding Problem Presents Challenges

- Protein structures live in 3D space and must conform to physical constraints → **we use equivariant graph neural networks designed for proteins** [Baek and Baker, 2022, Jumper et al., 2021]
- Convenient representations are non-Euclidean → Riemannian diffusion [De Bortoli et al., 2022]
  - **We use an $SE(3)^N$ diffusion model [Yim et al., 2023]**
- Extra degrees of freedom that are difficult to choose
  - Indices of motif within chain
  - Rotation & translation of motif
  - **We marginalize these out in the twisting function**

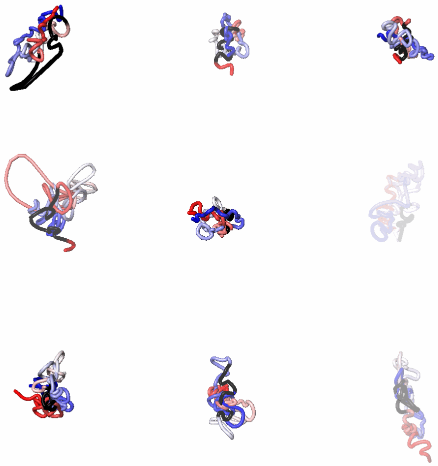
Motif-Scaffolding Problem — Example Trajectory

- View of $\hat{x}_\theta(x^t)$ for $9/64$ particles.
- Most probable motif is in black.



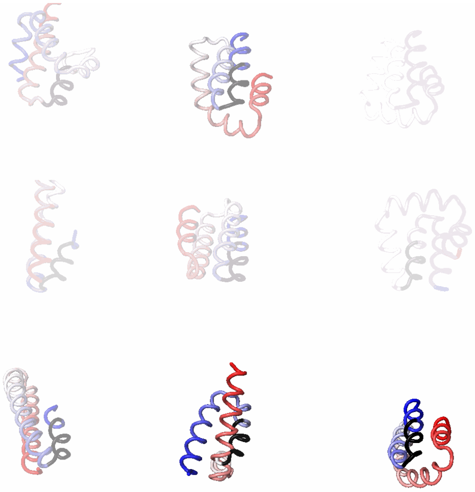
Motif-Scaffolding Problem — Example Trajectory

- View of $\hat{x}_\theta(x^t)$ for $9/64$ particles.
- Most probable motif is in black.



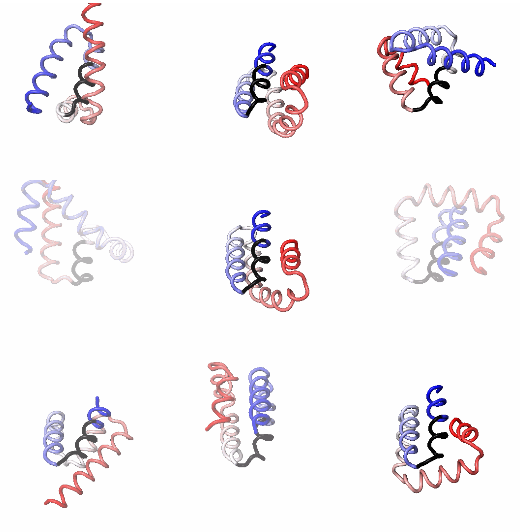
Motif-Scaffolding Problem — Example Trajectory

- View of $\hat{x}_\theta(x^t)$ for $9/64$ particles.
- Most probable motif is in black.

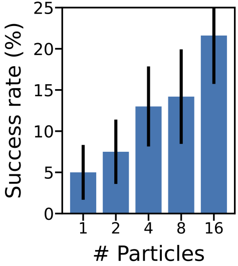


Motif-Scaffolding Problem — Example Trajectory

- View of $\hat{x}_\theta(x^t)$ for 9/64 particles.
- Most probable motif is in black.

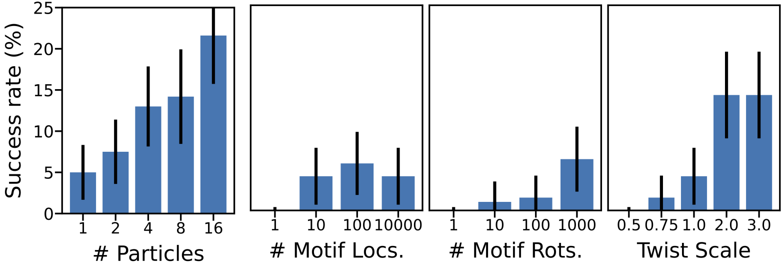


Motif-Scaffolding Problem — Example Results



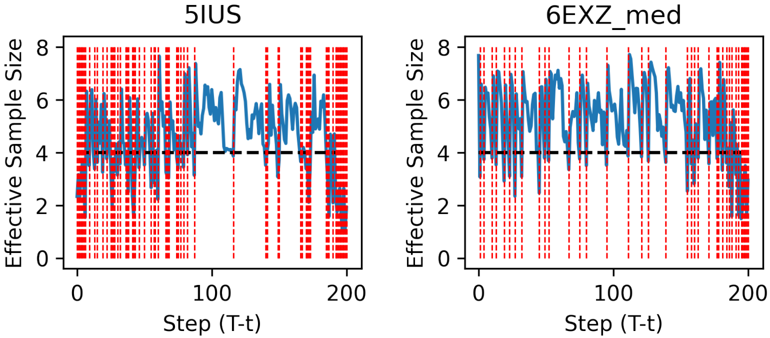
- Up to ∼5X increase in success rate

Motif-Scaffolding Problem — Example Results



- Up to $\sim$5X increase in success rate
- Performance relies on accomodation of degrees of freedom
- Including a multiplicative factor (twist scale) on the twisting function can improve performance
- On benchmark set, state of the art performance on 9/12 problems with short (< 100 residue) scaffolds.

Motif-Scaffolding Problem — Effective Sample Size



Twisted Diffusion Sampling (TDS) — Summary

- Accuracy of conditional sampling is a practical limitation of current controlled generation algorithms for diffusion models.

Twisted Diffusion Sampling (TDS) — Summary

- Accuracy of conditional sampling is a practical limitation of current controlled generation algorithms for diffusion models.
- Sequential Monte Carlo allows asymptotically exact estimation of conditional distributions of diffusion probabilistic models.

Twisted Diffusion Sampling (TDS) — Summary

- Accuracy of conditional sampling is a practical limitation of current controlled generation algorithms for diffusion models.
- Sequential Monte Carlo allows asymptotically exact estimation of conditional distributions of diffusion probabilistic models.
- Via "Twisting," heuristic approximations define proposals that improve efficiency without sacrifing exactness.

Twisted Diffusion Sampling (TDS) — Summary

- Accuracy of conditional sampling is a practical limitation of current controlled generation algorithms for diffusion models.
- Sequential Monte Carlo allows asymptotically exact estimation of conditional distributions of diffusion probabilistic models.
- Via "Twisting," heuristic approximations define proposals that improve efficiency without sacrifing exactness.
- Our implementation, TDS, provides state-of-the-art performance in protein design

Further Information

Trippe, Brian L.*, Luhuan Wu*, Christian A. Naesseth, John P. Cunningham, David Blei. "Practical and Asymptotically Exact Conditional Sampling in Diffusion Models." (2023) *** *equal contribution* `briantrippe.com/TDS_prepreprint.pdf`.

Contact me: blt2114@columbia.edu

References

Minkyung Baek and David Baker. Deep learning and protein structure modeling. *Nature methods*, 19(1):13–14, 2022.

Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

Valentin De Bortoli, Emile Mathieu, Michael John Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems*, 2022.

Pieralberto Guarniero, Adam M Johansen, and Anthony Lee. The iterated auxiliary particle filter. *Journal of the American Statistical Association*, 112 (520):1636–1647, 2017.

Jeremy Heng, Adrian N Bishop, George Deligiannidis, and Arnaud Doucet. Controlled sequential Monte Carlo. *Annals of Statistics*, 48(5), 2020.

Jonathan Ho, Tim Salimans, Alexey A Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022.

Lin Jiang, Eric A Althoff, Fernando R Clemente, Lindsey Doyle, Daniela Rothlisberger, Alexandre Zanghellini, Jasmine L Gallaher, Jamie L Betker, Fujie Tanaka, Carlos F Barbas III, Donald Hilvert, Kendal N Houk, Barry L. Stoddard, and David Baker. De novo computational design of retro-aldol enzymes. *Science*, 319(5868):1387–1391, 2008.

John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael