Parallel sampling via autospeculation

Thuy-Duong “June” Vuong

UC San Diego

Caltech Combinatorics Seminar, December 2025

Based on joint work with
Nima Anari, Carlo Baronio, CJ Chen, Alireza Haqi, Frederic Koehler and Anqi Li

Generative modeling

Autoregression (ChatGPT,…)

- $\mu : [q]^n \to \mathbb{R}_{\geq 0}$
- Sample via conditional marginal oracle
$$\mu(X_i = x_i | X_S = x_S)$$

$$x_1 \sim \mu(X_1), x_2 \sim \mu(X_2 | X_1 = x_1),$$
$$x_3 \sim \mu(X_2 | X_{1:2} = x_{1:2}), \ldots$$

$$Law(x_1 x_2 \ldots x_n) = \mu \to \mu$$

Denoising diffusion (DALL-E,…)

- $\mu : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$
- Sample via conditional mean oracle
$$f(t, x) = \mathbb{E}_{Y \sim \mu, g \sim \mathcal{N}(0, \frac{1}{t} \cdot I)} [Y | Y + g = \frac{x}{t}]$$

Continuous diffusion process:
$$d\, \bar{X}_t = f(t, \bar{X}_t) dt + dB_t$$
$$Law\left(\frac{\overline{X_t}}{t}\right) = \mu * \mathcal{N}\left(0, \frac{1}{t} \cdot I\right) \to \mu$$

To implement, need to truncate & discretize

Generative modeling

Autoregression (ChatGPT,…)

- $\mu : [q]^n \to \mathbb{R}_{\geq 0}$
- Sample via conditional marginal oracle

$$\mu(X_i = x_i | X_S = x_S)$$

Denoising diffusion (DALL-E,…)

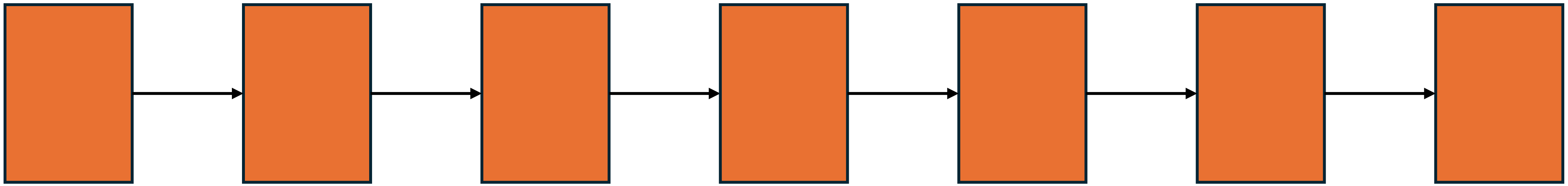
- $\mu : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$
- Sample via conditional mean oracle

$$f(t, x) = \mathbb{E}_{Y \sim \mu, g \sim \mathcal{N}(0, \frac{1}{t} \cdot I)} [Y | Y + g = \frac{x}{t}]$$

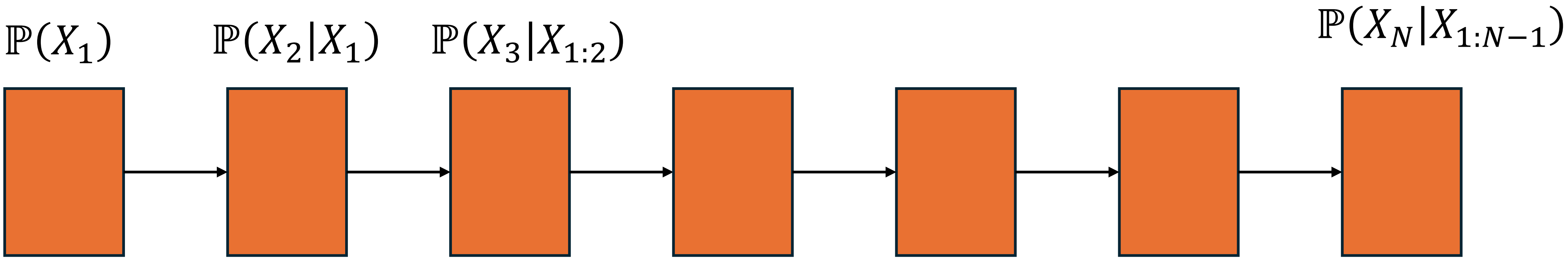
- In ML, oracle is learned from neural nets
- Theoretically, can evaluate/approximate oracles for certain distributions.
  Ex: determinantal distribution (spanning trees, planar perfect matchings)

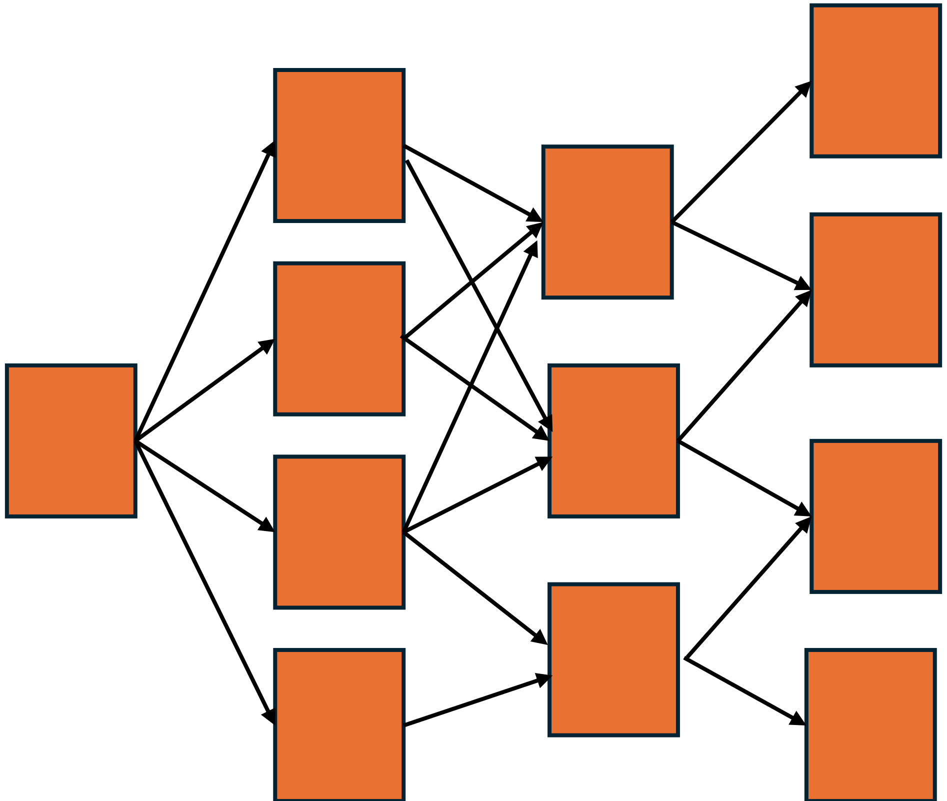
Generative modeling

$\mathbb{P}(X_1)$ $\mathbb{P}(X_2|X_1)$ $\mathbb{P}(X_3|X_{1:2})$ $\mathbb{P}(X_N|X_{1:N-1})$



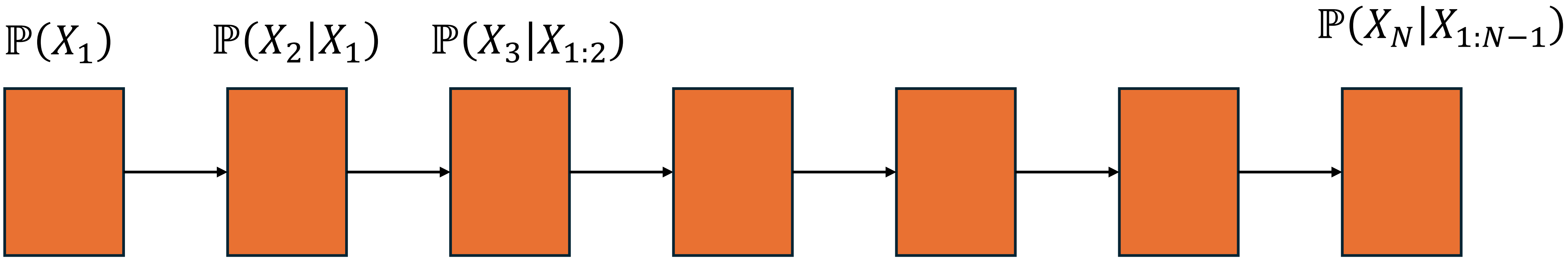
Generative modeling

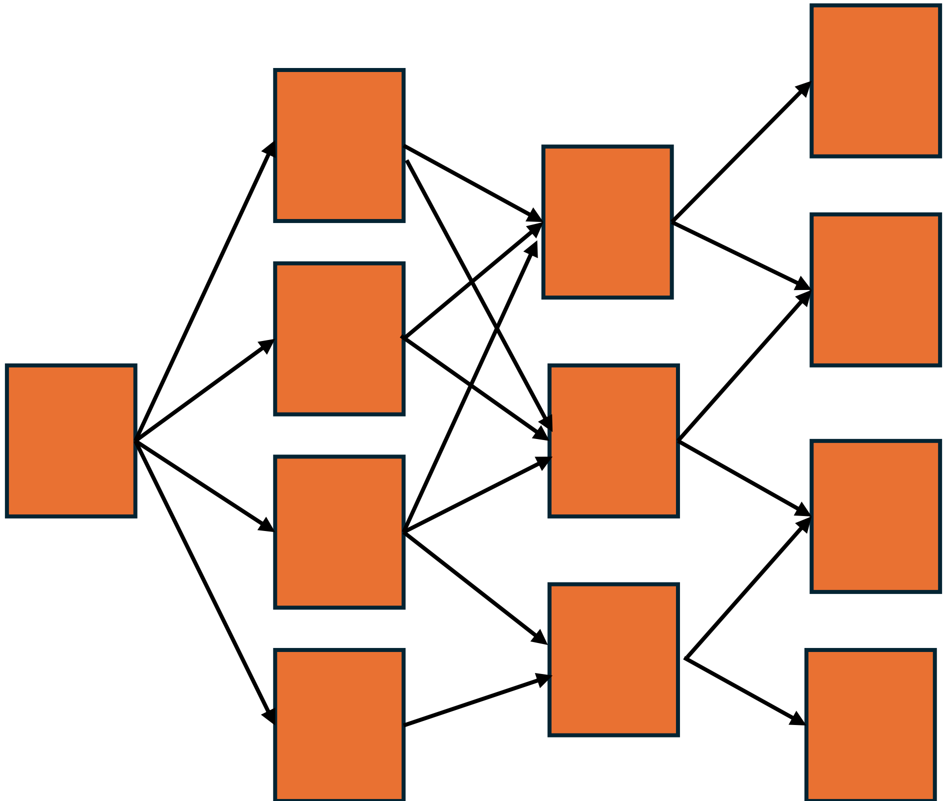




- Appears to be inherently sequential processes
- Parallelization?

Generative modeling





- Appears to be inherently sequential processes
- Parallelization?
- Parallel computing model:
  - Many parallel processes per round
  - Input of a round depends on previous rounds
- Our results: autoregression and diffusion can be parallelized in $\tilde{O}(n^{\frac{1}{2}})$ rounds

Autoregression (ChatGPT,…)

- Folklore: $\Theta(n)$ oracle queries/total work
- [AGR24] :
  - Upper bound: $\tilde{O}(n^{\frac{2}{3}})$ rounds & $O(n \log n)$ queries/total work
  - Lower bound: $\Omega(n^{\frac{1}{3}})$ rounds $\forall$ poly-queries algorithm
- This work:
  - $\tilde{O}(n^{\frac{1}{2}})$ rounds & $O(n \log n)$ queries/total work
  - Application: $\tilde{O}(n^{\frac{1}{4}})$ rounds algorithm to sample planar perfect matching

Denoising diffusion (DALL-E,…)

- Guarantee: $TV\left(\text{output}, \mu * \mathcal{N}(0, \delta \cdot I)\right) \leq \epsilon_{\mathrm{TV}}$
- [BDDD23]: $\widetilde{\mathrm{O}}(n\epsilon_{TV}^{-2})$ queries, for $Var(\mu) \leq R^2$
- [HDSA25] :$\widetilde{\mathrm{O}}\left(n^{2/3}\epsilon_{TV}^{-4/3}\right)$ rounds & $\mathrm{O}(n\epsilon_{TV}^{-2})$queries, for $Var(\mu) \leq R^2$
- This work:$\widetilde{\mathrm{O}}\left(n^{1/2}\right)$ rounds &$\widetilde{\mathrm{O}}(\max(n\epsilon_{TV}^{-2}, R^4\delta^{-2}))$queries,

  for $\sup \left|\left|X - \mathbb{E}_\mu[X]\right|\right|^2 \leq R^2$

Denoising diffusion (DALL-E,…)

- Guarantee: $TV\left(\text{output}, \mu * \mathcal{N}(0, \delta \cdot I)\right) \leq \epsilon_{\text{TV}}$
- [BDDD23]: $\text{O}\left(n\epsilon_{TV}^{-2}\, polylog\left(\frac{1}{\delta}, \frac{1}{\epsilon_{TV}}, n, R\right)\right)$ queries, for $Var(\mu) \leq R^2$
- [HDSA25] : $\text{O}\left(n^{2/3}\epsilon_{TV}^{-4/3}\, polylog\left(\frac{1}{\delta}, \frac{1}{\epsilon_{TV}}, n, R\right)\right)$ rounds &

  $\text{O}\left(n\epsilon_{TV}^{-2}\, polylog\left(\frac{1}{\delta}, \frac{1}{\epsilon_{TV}}, n, R\right)\right)$queries, for $Var(\mu) \leq R^2$
- This work: $\text{O}\left(n^{1/2}\, polylog\left(\frac{1}{\delta}, \frac{1}{\epsilon_{TV}}, n, R\right)\right)$ rounds &

  $\text{O}\left(\max(n\epsilon_{TV}^{-2}, R^4\delta^{-2})polylog\left(\frac{1}{\delta}, \frac{1}{\epsilon_{TV}}, n, R\right)\right)$queries, for $\sup \left\|X - \mathbb{E}_\mu[X]\right\|^2 \leq R^2$

Algorithm

Rejection sampling

μ: target distribution

ν: reference distribution

The following exactly samples from μ:

1. Sample $\boldsymbol{x} \sim \nu$
2. Accept and return $\boldsymbol{x}$ with probability $\frac{d\mu}{d\nu}(\boldsymbol{x})$

$$\rightarrow \mathbb{P}[\text{output x}] = d\nu(\boldsymbol{x}) \cdot \frac{d\mu}{d\nu}(\boldsymbol{x}) = d\mu(x)$$

Rejection sampling

μ: target distribution

ν: reference distribution

The following exactly samples from μ:

1. Sample $\boldsymbol{x} \sim \nu$
2. Accept and return $\boldsymbol{x}$ with probability $\frac{d\mu}{d\nu}(\boldsymbol{x})$ What if $\frac{d\mu}{d\nu}(\boldsymbol{x}) > 1$?

$$\rightarrow \mathbb{P}[\text{output x}] = d\nu(\boldsymbol{x}) \cdot \frac{d\mu}{d\nu}(\boldsymbol{x}) = d\mu(x)$$

Speculative rejection sampling

μ: target distribution

ν: reference distribution

The following exactly samples from μ:

1. Sample $\boldsymbol{x} \sim \nu$
2. Accept and return $\boldsymbol{x}$ with probability $\min\{1, \frac{d\mu}{d\nu}(\boldsymbol{x})\}$

$$\rightarrow \mathbb{P}[\text{output x}] = \min\{d\mu(x), d\nu(x)\}$$

Speculative rejection sampling

μ: target distribution

ν: reference distribution

The following exactly samples from μ:

1. Sample $\boldsymbol{x} \sim \nu$
2. Accept and return $\boldsymbol{x}$ with probability $\min\{1, \frac{d\mu}{d\nu}(\boldsymbol{x})\}$
   $\rightarrow \mathbb{P}[\text{output x}] = \min\{d\mu(x), d\nu(x)\}$
3. While not accepted do:
   Sample $\boldsymbol{y} \sim \mu$
   Accept and return $\boldsymbol{y}$ with probability $\max\{0,1 - \frac{d\nu}{d\mu}(\boldsymbol{y})\}$

$\rightarrow \mathbb{P}[\text{output y}] = \max\{0, d\mu(y) - d\nu(y)\}$

Recursive speculative rejection sampling

μ: target distribution

ν: reference distribution

The following exactly samples from μ: How to implement?

1. Sample $\boldsymbol{x} \sim \nu$
2. Accept and return $\boldsymbol{x}$ with probability $\min\{1, \frac{d\mu}{d\nu}(\boldsymbol{x})\}$
3. While not accepted do:
   Sample $\boldsymbol{y} \sim \mu$ by: $y_{1:n/2} \sim \mu(X_{1:n/2}), y_{n/2+1:n} \sim \mu(X_{n/2+1:n} | X_{1:n/2} = y_{1:n/2})$
   Accept and return $\boldsymbol{y}$ with probability $\max\{0, 1 - \frac{d\nu}{d\mu}(\boldsymbol{y})\}$

Recursive speculative rejection sampling

μ: target distribution

ν: reference distribution that admits fast sampler e.g. in $O(1)$ rounds
& can be built w/ conditional marginal/mean oracles for μ

The following exactly samples from μ: How to implement?

1. Sample $\boldsymbol{x} \sim \nu$
2. Accept and return $\boldsymbol{x}$ with probability $\min\{1, \frac{d\mu}{d\nu}(\boldsymbol{x})\}$
3. While not accepted do:
   ...

- Fully sequential: no gain
- Fully parallel: need infinite #processors

Recursive speculative rejection sampling

μ: target distribution

ν: reference distribution

The following exactly samples from μ:

How to implement?

1. Sample $\boldsymbol{x} \sim \nu$
2. Accept and return $\boldsymbol{x}$ with probability $\min\{1, \frac{d\mu}{d\nu}(\boldsymbol{x})\}$
3. While not accepted do:

   …

- Fully sequential: no gain
- Fully parallel: need infinite #processors
- Key idea: sequentially process batches of geometrically increasing size, where each batch is processed in parallel

Recursive speculative rejection sampling

μ: target distribution

ν: reference distribution

The following exactly samples from μ:

1. Sample $\boldsymbol{x} \sim \nu$
2. Accept and return $\boldsymbol{x}$ with probability $\min\{1, \frac{d\mu}{d\nu}(\boldsymbol{x})\}$
3. For r=0,1,2,... do:

   For $i = \lceil(1+\rho)^r\rceil, \lceil(1+\rho)^r\rceil + 1, \cdots, \lceil(1+\rho)^{r+1}\rceil - 1$ do in parallel

   Sample $\boldsymbol{y} \sim \mu$ by:

   $$y_{1:n/2} \sim \mu(X_{1:n/2}), y_{n/2+1:n} \sim \mu(X_{n/2+1:n} | X_{1:n/2} = y_{1:n/2})$$

   Accept and return $\boldsymbol{y}$ with probability $\max\{0, 1 - \frac{d\nu}{d\mu}(\boldsymbol{y})\}$

Recursive speculative rejection sampling

- $S = \{a+1, \cdots, b\}$
- μ_S: target distribution
- ν_S: reference distribution

In O(1) rounds and O(|S|) queries, can:

- Compute $\frac{d\mu_S}{d\nu_S}$
- Sample ν_S

The following exactly samples from μ_S:

1. Sample $\boldsymbol{x} \sim \nu_S$
2. Accept and return $\boldsymbol{x}$ with probability $\min\{1, \frac{d\mu_S}{d\nu_S}(\boldsymbol{x})\}$
3. For r=0,1,2,... do:

   For $i = \lceil (1+\rho)^r \rceil, \lceil (1+\rho)^r \rceil + 1, \cdots, \lceil (1+\rho)^{r+1} \rceil - 1$ do in parallel:

   Sample $\boldsymbol{y} \sim \mu_S$ by: Set $\mathrm{L}(S) = \left\{a+1, \ldots, \frac{a+b}{2}\right\}, \mathrm{R}(\mathrm{S}) = \{\frac{a+b}{2} + 1, \ldots, b\}$

   $$y_{L(S)} \sim \mu_{L(S)}, y_{R(S)} \sim \mu_{R(S)}(\cdot \mid X_{<R(S)} = . . || y_{L(S)})$$

   Accept and return $\boldsymbol{y}$ with probability $\max\{0, 1 - \frac{d\nu_S}{d\mu_S}(\boldsymbol{y})\}$

Analysis

Technical challenge

- Process in parallel → runtime determined by the slowest thread

#rounds #rounds in threads generated by the while loop

$$T^S \approx 1 + \mathbf{1}[\boldsymbol{x}_S \text{ not accepted}] \cdot \max\{T^{L(\mathrm{S}),\text{thread } \#t} + T^{R(\mathrm{S}),\text{thread } \#t}\}$$

- Bounding $\mathbb{E}[T]$ means bounding expectation of max
- Seems difficult

Analysis: high level

- Process in parallel → runtime determined by slowest thread

$$T^S \approx 1 + \mathbf{1}[\boldsymbol{x}_S \text{ not accepted}] \cdot \max\{T^{L(\mathrm{S}),\text{thread }\#t} + T^{R(\mathrm{S}),\text{thread }\#t}\}$$

- Key idea: expanding T as sum of max{indicators}
- $\mathbb{E}[T] = \sum \mathbb{E}[\max\{indicators\}] \approx \sum \mathbb{E}[\sum indicators] \approx \sum TV(\nu_S, \mu_S)$

$$\approx \tilde{O}\left(\sqrt{n}\right)$$

Pinning lemma

Analysis: step 1

Pinning lemma

$$\mathbb{E}[T] = \sum \mathbb{E}[\max\{indicators\}] \approx \sum \mathbb{E}[\sum indicators] \approx \sum TV(\nu_S, \mu_S) \approx \tilde{O}\,(\sqrt{n})$$

$$T^S \approx 1 + \mathbf{1}[\boldsymbol{x}_S \text{ not accepted}] \cdot \max\{T^{\mathrm{L(S)},\#1} + T^{R(S),\#1}, T^{\mathrm{L(S)},\#2} + T^{R(S),\#2}, \dots\}$$

$$\begin{tikzcd}
 & \boxed{S = \{a+1, \cdots, b\}} \arrow[dl] \arrow[dr] & \\
\boxed{\mathrm{L}(S) = \left\{a+1, \dots, \frac{a+b}{2}\right\}} & & \boxed{\mathrm{R(S)} = \{\frac{a+b}{2} + 1, \dots, b\}}
\end{tikzcd}$$

Analysis: step 1

Pinning lemma

$$\mathbb{E}[T] = \sum \mathbb{E}[\max\{indicators\}] \approx \sum \mathbb{E}[\sum indicators] \approx \sum TV(\nu_S, \mu_S) \approx \tilde{O}\,(\sqrt{n})$$

$$T^S \approx 1 + \mathbf{1}[\boldsymbol{x}_S \text{ not accepted}] \cdot \max\{T^{\mathrm{L(S)},\#1} + T^{R(S),\#1}, T^{\mathrm{L(S)},\#2} + T^{R(S),\#2}, \ldots\}$$

$$\approx 1 + I^S \cdot \max\{1 + I^{L(S),\#t_0} \max\{\ldots\} + 1 + I^{R(S),\#t_0} \max\{\ldots\}\}$$

$$\approx 1 + I^S + \sum_{\substack{S_1:\text{child of S} \\ t_0:\text{thread index}}} I^S \cdot I^{S_1,\#t_0} + \cdots$$

Analysis: step 2

Pinning lemma

$$\mathbb{E}[T] = \sum \mathbb{E}[\max\{indicators\}] \approx \sum \mathbb{E}[\sum indicators] \approx \sum TV(\nu_S, \mu_S) \approx \tilde{O}\,(\sqrt{n})$$

$$\mathbb{E}[T^S] \approx 1 + \mathbb{E}[I^S] + \sum_{S_1\text{:child of S}} \mathbb{E}[\sum_{t_0\text{:thread index}} I^S \cdot I^{S_1,\#t_0}] + \cdots$$

Recursive speculative rejection sampling

- $S = \{a+1, \cdots, b\}$, μ_S: target distribution, ν_S: reference distribution

1. Sample $\boldsymbol{x} \sim \nu_S$
2. Accept and return $\boldsymbol{x}$ with probability $\min\{1, \frac{d\mu_S}{d\nu_S}(\boldsymbol{x})\}$ $\rightarrow \mathbb{E}[\mathrm{I}^S] = \mathrm{TV}(\nu_S, \mu_S)$
3. For r=0,1,2,... do:

   For $i = \lceil (1+\rho)^r \rceil, \lceil (1+\rho)^r \rceil + 1, \cdots, \lceil (1+\rho)^{r+1} \rceil - 1$ do in parallel:

   Sample $\boldsymbol{y} \sim \mu_S$ by: Set $\mathrm{L}(S) = \left\{a+1, \ldots, \frac{a+b}{2}\right\}$, $\mathrm{R(S)} = \{\frac{a+b}{2} + 1, \ldots, b\}$

   $$y_{L(S)} \sim \mu_{L(S)}, y_{R(S)} \sim \mu_{R(S)}(\cdot \mid X_{<R(S)} = .. \| y_{L(S)})$$

   Accept and return $\boldsymbol{y}$ with probability $\max\{0, 1 - \frac{d\nu_S}{d\mu_S}(\boldsymbol{y})\}$

Analysis: step 2

Pinning lemma

$$\mathbb{E}[T] = \sum \mathbb{E}[\max\{indicators\}] \approx \sum \mathbb{E}[\sum indicators] \approx \sum TV(\nu_S, \mu_S) \approx \tilde{O}(\sqrt{n})$$

$$\mathbb{E}[T^S] \approx 1 + \underbrace{\mathbb{E}[I^S]}_{TV(\nu_S, \mu_S)} + \sum_{S_1:\text{child of S}} \mathbb{E}[\sum_{t_0:\text{thread index}} I^S \cdot I^{S_1, \# t_0}] + \cdots$$

Analysis: step 2

Pinning lemma

$$\mathbb{E}[T] = \sum \mathbb{E}[\max\{indicators\}] \approx \sum \mathbb{E}[\sum indicators] \approx \sum TV(\nu_S, \mu_S) \approx \tilde{O}\left(\sqrt{n}\right)$$

$$\mathbb{E}[T^S] \approx 1 + \mathbb{E}[I^S] + \sum_{S_1\text{:child of S}} \mathbb{E}[\sum_{t_0\text{:thread index}} I^S \cdot I^{S_1,\#t_0}] + \cdots$$

$$\mathbb{E}[\sum_{t_0\text{:thread index}} I^S \cdot I^{S_1,\#t_0}] = \mathbb{E}\left[\, \mathbb{E}[I^S] \times \mathbb{E}[\#\text{threads } \mathrm{t}_0] \times \mathbb{E}[TV(\nu_{S_1}, \mu_{S_1})] \,\right]$$

Analysis: step 2

Pinning lemma

$$\mathbb{E}[T] = \sum \mathbb{E}[\max\{indicators\}] \approx \sum \mathbb{E}[\sum indicators] \approx \sum TV(\nu_S, \mu_S) \approx \tilde{O}\,(\sqrt{n})$$

$$\mathbb{E}[T^S] \approx 1 + \mathbb{E}[I^S] + \sum_{S_1\text{:child of S}} \mathbb{E}[\sum_{t_0\text{:thread index}} I^S \cdot I^{S_1,\#t_0}] + \cdots$$

$$\mathbb{E}[\sum_{t_0\text{:thread index}} I^S \cdot I^{S_1,\#t_0}] = \mathbb{E}\Big[\,\mathbb{E}[I^S] \times \mathbb{E}[\#\text{threads } \mathrm{t}_0] \times \mathbb{E}[TV(\nu_{S_1}, \mu_{S_1})]\,\Big]$$

Next, bound product of the first two terms

Recursive speculative rejection sampling

- $S = \{a+1, \cdots, b\}$, μ_S: target distribution, ν_S: reference distribution

1. Sample $\boldsymbol{x} \sim \nu_S$
2. Accept and return $\boldsymbol{x}$ with probability $\min\{1, \frac{d\mu_S}{d\nu_S}(\boldsymbol{x})\}$ $\quad \rightarrow \mathbb{E}[\mathrm{I}^S] = \mathrm{TV}(\nu_S, \mu_S)$
3. For r=0,1,2,... do:

   For $i = \lceil (1+\rho)^r \rceil, \lceil (1+\rho)^r \rceil + 1, \cdots, \lceil (1+\rho)^{r+1} \rceil - 1$ do in parallel:

   Sample $\boldsymbol{y} \sim \mu_S$ by: Set $\mathrm{L}(S) = \left\{a+1, \ldots, \frac{a+b}{2}\right\}$, $\mathrm{R(S)} = \{\frac{a+b}{2} + 1, \ldots, b\}$

   $$y_{L(S)} \sim \mu_{L(S)}, y_{R(S)} \sim \mu_{R(S)}(\cdot \mid X_{<R(S)} = ..||y_{L(S)})$$

   Accept and return $\boldsymbol{y}$ with probability $\max\{0, 1 - \frac{d\nu_S}{d\mu_S}(\boldsymbol{y})\}$

$\rightarrow \mathbb{P}[\boldsymbol{y} \text{ accepted}] = \mathrm{TV}(\nu_S, \mu_S) \qquad \Rightarrow \mathbb{E}[\#\text{threads } \mathrm{t}_0] = \mathbb{E}[\#y's] \leq \frac{1+\rho}{\mathrm{TV}(\nu_S, \mu_S)} = \frac{1}{\mathbb{E}[\mathrm{I}^S]}$

Analysis: step 2

Pinning lemma

$$\mathbb{E}[T] = \sum \mathbb{E}[\max\{indicators\}] \approx \sum \mathbb{E}[\sum indicators] \approx \sum TV(\nu_S, \mu_S) \approx \tilde{O}(\sqrt{n})$$

$$\mathbb{E}[T^S] \approx 1 + \mathbb{E}[I^S] + \sum_{S_1\text{:child of S}} \mathbb{E}[\sum_{t_0\text{:thread index}} I^S \cdot I^{S_1,\#t_0}] + \cdots$$

$$\mathbb{E}[\sum_{t_0\text{:thread index}} I^S \cdot I^{S_1,\#t_0}] = \mathbb{E}[\underbrace{\mathbb{E}[I^S] \times \mathbb{E}[\#\text{threads } \mathrm{t}_0]}_{\leq (1+\rho)} \times \mathbb{E}[TV(\nu_{S_1}, \mu_{S_1})]]$$

Analysis: step 2

Pinning lemma

$$\mathbb{E}[T] = \sum \mathbb{E}[\max\{indicators\}] \approx \sum \mathbb{E}[\sum indicators] \approx \sum TV(\nu_S, \mu_S) \approx \tilde{O}(\sqrt{n})$$

$$\mathbb{E}[T^S] \approx 1 + \mathbb{E}[I^S] + \sum_{S_1:\text{child of S}} \mathbb{E}[\sum_{t_0:\text{thread index}} I^S \cdot I^{S_1,\#t_0}] + \cdots$$

$$\mathbb{E}[\sum_{t_0:\text{thread index}} I^S \cdot I^{S_1,\#t_0}] = \mathbb{E}\Big[\underbrace{\mathbb{E}[I^S] \times \mathbb{E}[\#\text{threads } \mathrm{t}_0]}_{\leq (1+\rho)} \times \mathbb{E}[TV(\nu_{S_1}, \mu_{S_1})]\Big]$$

$$\leq (1+\rho)\mathbb{E}[TV\big(\nu_{S_1}, \mu_{S_1}\big)]$$

Analysis: step 2

Pinning lemma

$$\mathbb{E}[T] = \sum \mathbb{E}[\max\{indicators\}] \approx \sum \mathbb{E}\Big[\sum indicators\Big] \approx \sum TV(\nu_S, \mu_S) \approx \tilde{O}\left(\sqrt{n}\right)$$

$$\begin{aligned}
&\mathbb{E}[T^S] \\
&\approx 1 + \mathbb{E}[I^S] + (1+\rho) \sum_{S_1\text{:child of S}} \mathbb{E}\left[TV\left(\nu_{S_1}, \mu_{S_1}\right)\right] \\
&+ (1+\rho)^2 \sum_{S_2\text{:grandchild of S}} \mathbb{E}\left[TV\left(\nu_{S_2}, \mu_{S_2}\right)\right] \ldots
\end{aligned}$$

Set $\rho = \frac{1}{\text{recursion depth}}$: $(1+\rho)^m \leq O(1)$

Analysis: step 3

Pinning lemma

$$\mathbb{E}[T] = \sum \mathbb{E}[\max\{indicators\}] \approx \sum \mathbb{E}[\sum indicators] \approx \sum TV(\nu_S, \mu_S) \approx \tilde{O}\,(\sqrt{n})$$

$$\mathbb{E}[T^S] \approx 1 + \mathbb{E}[I^S] + \sum_{S_1\text{:child of S}} \mathbb{E}\left[TV\left(\nu_{S_1}, \mu_{S_1}\right)\right] + \sum_{S_2\text{:grandchild of S}} \mathbb{E}\left[TV\left(\nu_{S_2}, \mu_{S_2}\right)\right] \ldots$$

Reference & target distributions (ν_S&μ_S)?

Autoregression (ChatGPT,…)

- $\mu : [q]^n \to \mathbb{R}_{\geq 0}$
- Sample via conditional marginal oracle

$$\mu(X_i = x_i | X_S = x_S)$$

$$x_1 \sim \mu(X_1), x_2 \sim \mu(X_2 | X_1 = \sigma_1), \\ x_3 \sim \mu(X_2 | X_{1:2} = \sigma_{1:2}), \ldots$$

Denoising diffusion (DALL-E,…)

- $\mu : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$
- Sample via conditional mean oracle

$$f(t, x) = \mathbb{E}_{Y \sim \mu, g \sim \mathcal{N}(0, \frac{1}{t} \cdot I)} [Y | Y + g = \frac{x}{t}]$$

Continuous diffusion process:

$$d\, \bar{X}_t = f(t, \bar{X}_t) dt + dB_t$$

$$Law\left(\frac{\overline{X_t}}{t}\right) = \mu * \mathcal{N}\left(0, \frac{1}{t} \cdot I\right) \to \mu$$

Autoregression (ChatGPT,...)

- $\mu : [q]^n \to \mathbb{R}_{\geq 0}$
- Sample via conditional marginal oracle $\mu(X_i = x_i | X_S = x_S)$
- 1st attempt:
  - $S = \{a + 1, \ldots, b\}$
  - $\mu_S = \mu(X_S | X_{<S}) = \mu(X_{a+1:b} | X_{1:a})$
  - $\nu_S =$ product distribution with same marginal as μ_S

    $= \mu(X_{a+1} | X_{1:a}) \otimes \mu(X_{a+2} | X_{1:a}) \otimes \cdots \otimes \mu(X_b | X_{1:a})$

In O(1) rounds, can:

- Compute $\frac{d\mu_S}{d\nu_S}$
- Sample ν_S

Autoregression (ChatGPT,...)

- $\mu : [q]^n \to \mathbb{R}_{\geq 0}$
- Sample via conditional marginal oracle $\mu(X_i = x_i | X_S = x_S)$
- 1st attempt:
  - $S = \{a+1, .., b\}$
  - $\mu_S = \mu(X_S | X_{<S}) = \mu(X_{a+1:b} | X_{1:a})$
  - ν_S = product distribution with same marginal as μ_S
    $= \mu(X_{a+1} | X_{1:a}) \otimes \mu(X_{a+2} | X_{1:a}) \otimes \cdots \otimes \mu(X_b | X_{1:a})$
  - Unfortunately, $\sum_{S_\ell:\text{ descendant of root}} \mathbb{E}\left[TV\left(\nu_{S_\ell}, \mu_{S_\ell}\right)\right] = \Omega(n)$
  - Ex: $\mu = \text{Uniform}(\{x \in \{0,1\}^n : x_1 = x_2, x_3 = x_4, \ldots\})$
    $\mathbb{E}[TV(\nu_{1:2}, \mu_{1:2}) + TV(\nu_{3:4}, \mu_{3:4}) + \cdots] = \Omega(n)$

For fixed set S,
$TV(\mu_S, \nu_S) = \Omega(1)$
b/c dependencies

Autoregression (ChatGPT,…)

- $\mu : [q]^n \to \mathbb{R}_{\geq 0}$
- Sample via conditional marginal oracle $\mu(X_i = x_i | X_S = x_S)$
- 1st attempt:
  - $\mathrm{S} = \{a+1, .., b\}$
  - $\mu_S = \mu(X_S | X_{<S}) = \mu(X_{a+1:b} | X_{1:a})$
  - $\nu_S =$ product distribution with same marginal as μ_S
    $= \mu(X_{a+1} | X_{1:a}) \otimes \mu(X_{a+2} | X_{1:a}) \otimes \cdots \otimes \mu(X_b | X_{1:a})$
  - Unfortunately, $\sum_{S_\ell:\text{ descendant of root}} \mathbb{E}\left[TV\left(\nu_{S_\ell}, \mu_{S_\ell}\right)\right] = \Omega(n)$
- Can produce random subsets by taking a random recursive partition of [n]

For fixed set S,
$TV(\mu_S, \nu_S) = \Omega(1)$
b/c dependencies

Pinning lemma:
For random small subset S,
$TV(\mu_S, \nu_S) = o(1)$

Autoregression (ChatGPT,...)

- $\mu : [q]^n \to \mathbb{R}_{\geq 0}$
- Sample via conditional marginal oracle $\mu(X_i = x_i | X_S = x_S)$
- Solution: Sample $\sigma \sim \text{Uniform}(S_n)$
  - $S = \{a+1, \ldots, b\}$
  - $\mu_S = \mu\left(X_{\sigma(S)} \middle| X_{\sigma(<S)}\right) = \mu(X_{\sigma(a+1:b)} | X_{\sigma(1:a)})$
  - $\nu_S =$ product distribution with same marginal as μ_S

$$= \mu\left(X_{\sigma(a+1)} \middle| X_{\sigma(1:a)}\right) \otimes \cdots \otimes \mu\left(X_{\sigma(b)} \middle| X_{\sigma(1:a)}\right)$$

- Can produce random subsets by taking a random recursive partition of [n]

For fixed set S,
$TV(\mu_S, \nu_S) = \Omega(1)$
b/c dependencies

Pinning lemma:
For random set S,
$TV(\mu_S, \nu_S) = o(1)$

Autoregression (ChatGPT,…)

- $\mu : [q]^n \to \mathbb{R}_{\geq 0}$
- Sample via conditional marginal oracle $\mu(X_i = x_i | X_S = x_S)$
- Solution: Sample $\sigma \sim \text{Uniform}(S_n)$
  - $S = \{a+1, \ldots, b\}$
  - $\mu_S = \mu\left(X_{\sigma(S)} \middle| X_{\sigma(<S)}\right) = \mu(X_{\sigma(a+1:b)} | X_{\sigma(1:a)})$
  - ν_S = product distribution with same marginal as μ_S

    $= \mu\left(X_{\sigma(a+1)} \middle| X_{\sigma(1:a)}\right) \otimes \cdots \otimes \mu\left(X_{\sigma(b)} \middle| X_{\sigma(1:a)}\right)$
- $\sum_{S_\ell:\text{level}-\ell \text{ descendant of root}} \mathbb{E}\left[TV\left(\nu_{S_\ell}, \mu_{S_\ell}\right)\right] = O(\sqrt{n \log q})$

For fixed set S,
$TV(\mu_S, \nu_S) = \Omega(1)$
b/c dependencies

Pinning lemma:
For random set S,
$TV(\mu_S, \nu_S) = o(1)$

Denoising diffusion (DALL-E,...)

- $\mu : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$
- Sample via conditional mean oracle: $f(t,x) = \mathbb{E}_{Y\sim\mu, g\sim\mathcal{N}(0,\frac{1}{t}\cdot I)}[Y | Y + g = \frac{x}{t}]$
- Continuous diffusion process: $d\,\bar{X}_t = f(t,\bar{X}_t)dt + dB_t$,

$$Law\left(\frac{\bar{X}_t}{t}\right) = \mu * \mathcal{N}\left(0, \frac{1}{t}\cdot I\right) \to \mu$$

- Discretized process w/ endpoints $t_0 < t_1 < t_2 < \cdots < \mathrm{t_N} = \frac{1}{\delta}$:

$$X_{t_{i+1}} = X_{t_i} + f\left(t_i, X_{t_i}\right)(\mathrm{t}_{i+1} - t_i) + \mathcal{N}(0, (t_{i+1} - t_i)\cdot I)$$

- With appropriate discretization, $TV(\bar{X}_t, X_t) \leq \epsilon_{TV}$

Denoising diffusion (DALL-E,...)

- $\mu : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$
- Sample via conditional mean oracle: $f(t,x) = \mathbb{E}_{Y\sim\mu, g\sim\mathcal{N}(0,\frac{1}{t}\cdot I)}\left[Y | Y + g = \frac{x}{t}\right]$
- Continuous diffusion process: $d\,\bar{X}_t = f(t, \bar{X}_t)dt + dB_t$,

$$Law\left(\frac{\overline{X_t}}{t}\right) = \mu * \mathcal{N}\left(0, \frac{1}{t}\cdot I\right) \to \mu$$

- Discretized process w/ endpoints $t_0 < t_1 < t_2 < \cdots < \mathrm{t_N} = \frac{1}{\delta}$:

$$X_{t_{i+1}} = X_{t_i} + f\left(t_i, X_{t_i}\right)(\mathrm{t}_{i+1} - t_i) + \mathcal{N}(0, (t_{i+1} - t_i)\cdot I)$$

- $\mathrm{S} = \{a+1, .., b\}$,
- $\mu_S = Law\left(\left(X_{t_i} - X_{t_{i-1}}\right)_{i=a+1}^{b}\right)$
- $\nu_S = Law((f\left(t_a, X_{t_a}\right)(t_i - t_{i-1}) + \mathcal{N}(0, (t_{i+1} - t_i)\cdot I)_{i=a+1}^{b})$

In O(1) rounds, can:

- Compute $\frac{d\mu_S}{d\nu_S}$
- Sample ν_S

Denoising diffusion (DALL-E,…)

- $\mu : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ w/ oracle $f(t,x) = \mathbb{E}_{Y\sim\mu, g\sim\mathcal{N}(0,\frac{1}{t}\cdot I)}\,[Y|Y+g=\frac{x}{t}]$
- Continuous diffusion process: $d\,\bar{X}_t = f(t,\bar{X}_t)dt + dB_t$,
$$Law\left(\frac{\bar{X}_t}{t}\right) = \mu * \mathcal{N}\left(0,\frac{1}{t}\cdot I\right) \to \mu$$
- Discretized process w/ endpoints $t_0 < t_1 < t_2 < \cdots < \mathrm{t_N} = \frac{1}{\delta}$:
$$X_{t_{i+1}} = X_{t_i} + f\left(t_i, X_{t_i}\right)(\mathrm{t}_{i+1} - t_i) + \mathcal{N}(0,(t_{i+1}-t_i)\cdot I)$$
- $\mathrm{S} = \{a+1,\ldots,b\}, \mu_S = Law\left(\left(X_{t_i} - X_{t_{i-1}}\right)_{i=a+1}^{b}\right)$, $\nu_S = Law((f\left(t_a, X_{t_a}\right)(t_i - t_{i-1}) + \mathcal{N}(0,(t_{i+1}-t_i)\cdot I)_{i=a+1}^{b})$
$$TV(\mu_S,\nu_S)^2 \leq \mathbb{E}[\sum_{i=a}^{b-1}(t_{i+1}-t_i)\left\|f\left(t_i,X_{t_i}\right) - f\left(t_a,X_{t_a}\right)\right\|^2]$$

Denoising diffusion (DALL-E,…)

- $\mu : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ w/ oracle $f(t,x) = \mathbb{E}_{Y\sim\mu, g\sim\mathcal{N}(0,\frac{1}{t}\cdot I)}\, [Y|Y+g = \frac{x}{t}]$
- Continuous diffusion process: $d\,\bar{X}_t = f(t,\bar{X}_t)dt + dB_t \ \to Law\left(\frac{\overline{X_t}}{t}\right) = \mu * \mathcal{N}\left(0, \frac{1}{t}\cdot I\right) \to \mu$
- Discretized process w/ endpoints $t_0 < t_1 < t_2 < \cdots < \mathrm{t_N} = \frac{1}{\delta}$:

$$X_{t_{i+1}} = X_{t_i} + f\left(t_i, X_{t_i}\right)(\mathrm{t}_{i+1} - t_i) + \mathcal{N}(0, (t_{i+1} - t_i)\cdot I)$$

- $\mathrm{S} = \{a+1,..,b\}, \mu_S = Law\left(\left(X_{t_i} - X_{t_{i-1}}\right)_{i=a+1}^{b}\right)$, $\nu_S = Law((f\left(t_a, X_{t_a}\right)(t_i - t_{i-1}) + \mathcal{N}(0, (t_{i+1} - t_i)\cdot I)_{i=a+1}^{b})$

$$TV(\mu_S, \nu_S)^2 \leq \mathbb{E}\left[\sum_{i=a}^{b-1} (t_{i+1} - t_i)\left\|f\left(t_i, X_{t_i}\right) - f\left(t_a, X_{t_a}\right)\right\|^2\right] \leq TV(\bar{\mu}_S, \nu_S)^2 + R^4(t_b - t_a)^2$$

$$TV(\overline{\mu_S}, \nu_S)^2 \ \leq \mathbb{E}\left[\sum_{i=a}^{b-1}(t_{i+1} - t_i)\left\|f\left(t_i, \bar{X}_{t_i}\right) - f\left(t_a, \bar{X}_{t_a}\right)\right\|^2\right] = \mathbb{E}[(t_b - t_a)(tr(\Sigma_a) - tr(\Sigma_b))]$$

Denoising diffusion (DALL-E,…)

- $\mu : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ w/ oracle $f(t,x) = \mathbb{E}_{Y\sim\mu, g\sim\mathcal{N}(0,\frac{1}{t}\cdot I)}\left[Y \middle| Y+g=\frac{x}{t}\right]$
- $S=\{a+1,..,b\}, \mu_S = Law\left(\left(X_{t_i}-X_{t_{i-1}}\right)_{i=a+1}^{b}\right), \nu_S = Law((f\left(t_a,X_{t_a}\right)(t_i-t_{i-1})+\mathcal{N}(0,(t_{i+1}-t_i)\cdot I))_{i=a+1}^{b})$
- $TV(\mu_S,\nu_S)^2 \leq \mathbb{E}\left[\sum_{i=a}^{b-1}(t_{i+1}-t_i)\left\|f\left(t_i,X_{t_i}\right)-f\left(t_a,X_{t_a}\right)\right\|^2\right] \leq TV(\bar{\mu}_S,\nu_S)^2 + R^4(t_b-t_a)^2 \preccurlyeq R^4(t_b-t_a)^2$
- $TV(\overline{\mu_S},\nu_S)^2 \leq \mathbb{E}\left[\sum_{i=a}^{b-1}(t_{i+1}-t_i)\left\|f\left(t_i,\bar{X}_{t_i}\right)-f\left(t_a,\bar{X}_{t_a}\right)\right\|^2\right] = \mathbb{E}[(t_b-t_a)(tr(\Sigma_a)-tr(\Sigma_b))]$

$$\mathbb{E}\left[\sum_{S_\ell:\text{level}-\ell\text{ descendant of root}} TV\left(\nu_{S_\ell},\mu_{S_\ell}\right)\right] \leq d_{TV}\left((X_t)_t,(\overline{X_t})_t\right)\text{sup}.. + \mathbb{E}\left[\sum_{S_\ell:\text{level}-\ell\text{ descendant of root}} TV\left(\nu_{S_\ell},\bar{\mu}_{S_\ell}\right)\right]$$

$$\leq \sqrt{\frac{n}{N}}\cdot\frac{R^2}{\delta} + O\left(tr(\Sigma_0)(T_1-T_0)+tr\left(\Sigma_{T_1}\right)(T_2-T_1)+\ \dots\right)$$

$$tr(\Sigma_t) \leq \frac{n}{t+R^{-2}}$$

$$\leq \tilde{O}(\sqrt{n})$$

Open question

- Get tight bound for autoregression.
  - Currently, upper bound $O(n^{\frac{1}{2}})$ but lower bound $\Omega(n^{\frac{1}{3}})$
- Show lower bound for denoising diffusion.
  - Currently, upper bound $O(n^{\frac{1}{2}})$ but no lower bound