Masked diffusions: how to choose an optimal schedule and how much is gained?

Hugo Lavenant

Bocconi University

Statistical seminar

ENSAE (France), April 13, 2026

Joint work with:

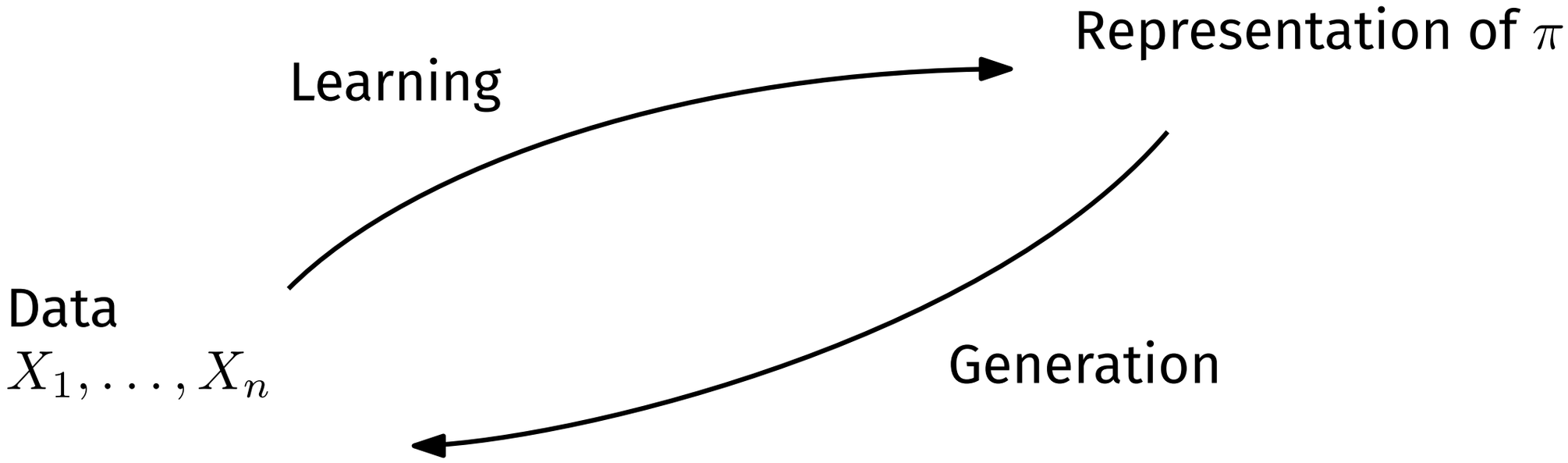
Giacomo Zanella

[Arxiv preprint: 2510.25544]

Generative modelling

Assumption: data generated from π.

Goal: new samples from π.



Generative modelling

Assumption: data generated from π.

Goal: new samples from π.

Learning

Representation of π with masked diffusions

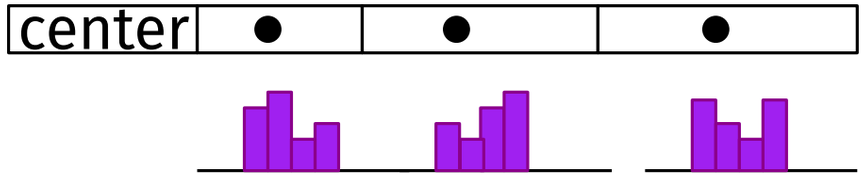
Today

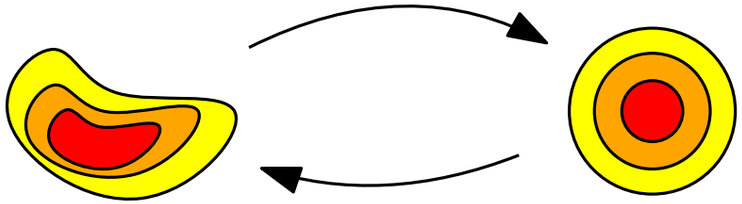
Data : text $X_1, \ldots, X_n$

Generation

Speed up with factorized approximation

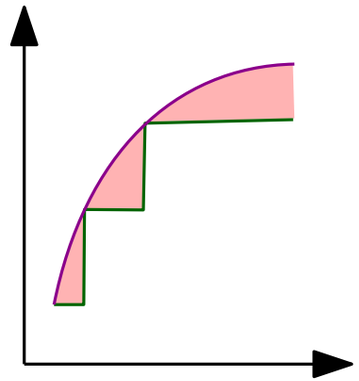
1 - Context: masked diffusion and factorization error



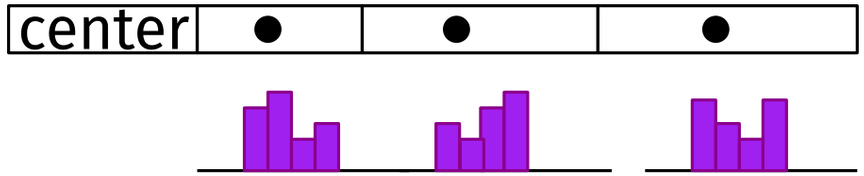


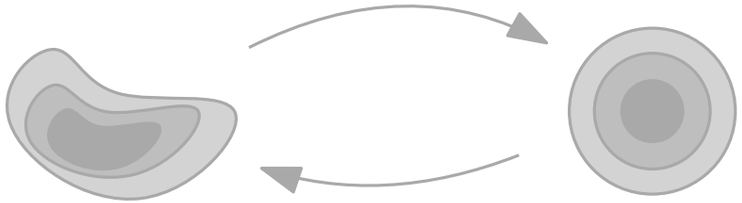
2 - Why ”diffusion”? Analogy with continuous models

3 - Our results: scaling of the factorization error and optimal schedules



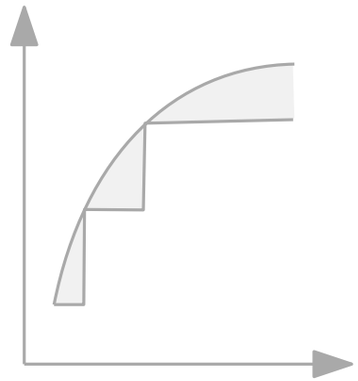
1 - Context: masked diffusion and factorization error



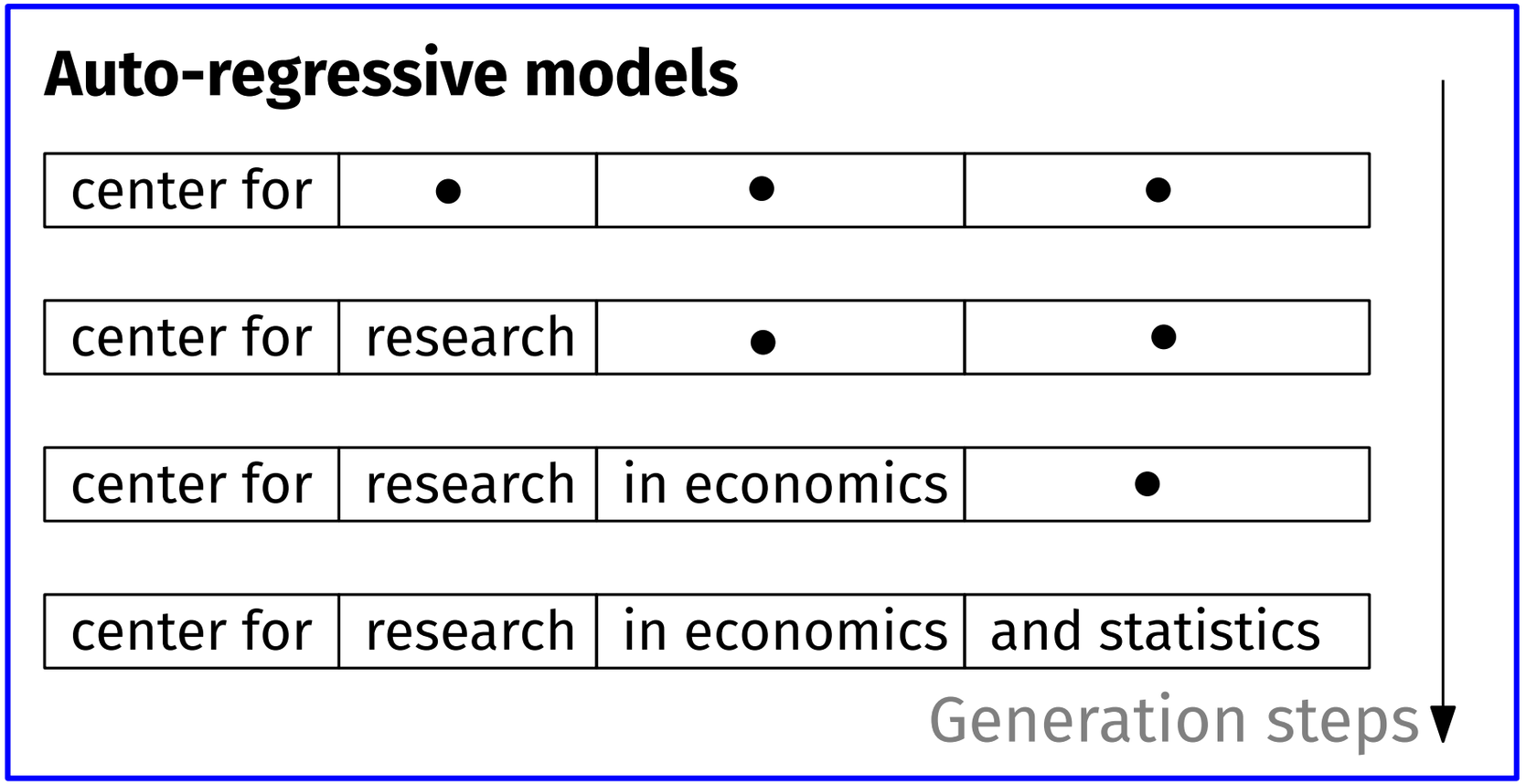


2 - Why ”diffusion”? Analogy with continuous models

3 - Our results: scaling of the factorization error and optimal schedules



Text generation with auto-regressive models



Text generation with auto-regressive models

Auto-regressive models

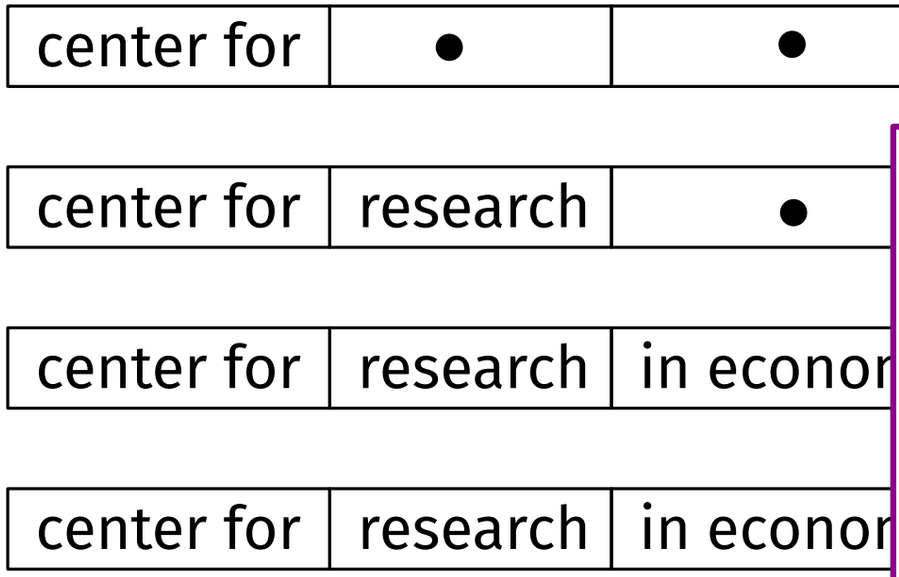
center for	•	•	•
center for	research	•	
center for	research	in econor	
center for	research	in econor	

Computational burden

Generation steps K = Sentence length N

Text generation with auto-regressive models

Auto-regressive models



Computational burden

Generation steps $K =$ Sentence length N

Exact sampling if

$$x_k \sim \pi(x_k | x_1, x_2, \ldots, x_{k-1})$$

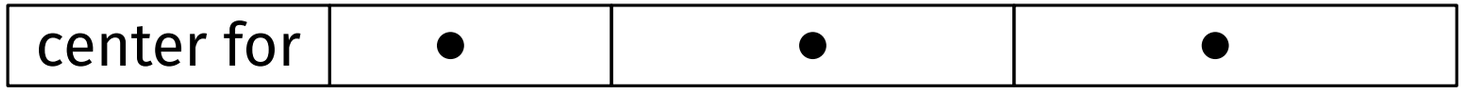
Probabilistic model

$$\pi \in \mathcal{P}(\mathcal{X}^N)$$

N sentence length

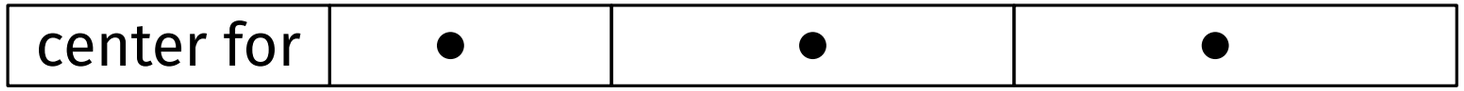
$\mathcal{X}$ alphabet (finite)

Any order models

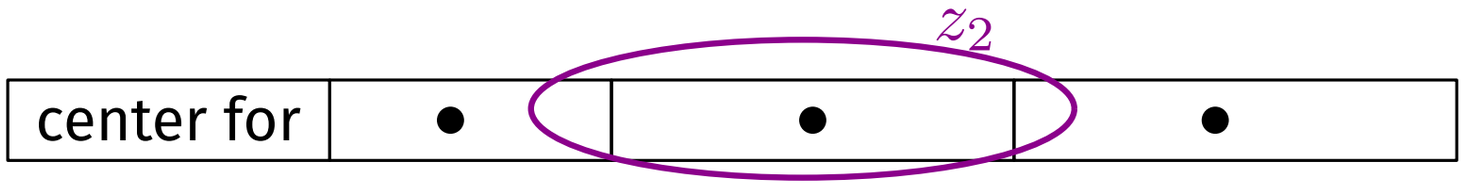


[Kim, Shah, Kontonis, Kakade and Chen, 2025]

Any order models



1. Choose z_k to unmask.



[Kim, Shah, Kontonis, Kakade and Chen, 2025]

Any order models

center for	•	•	•

1. Choose z_k to unmask.
2. Generate x_{z_k}.

Ideally, x_{z_k} sampled from

$$\pi(x_{z_k} | x_{z_1}, \ldots x_{z_{k-1}})$$

[Kim, Shah, Kontonis, Kakade and Chen, 2025]

Any order models

center for	•	•	•

1. Choose z_k to unmask.
2. Generate x_{z_k}.

center for	•	in economics	•

z_2

center for	•	in economics	and statistics

z_3

center for	research	in economics	and statistics

z_4

Generation steps

Ideally, x_{z_k} sampled from

$$\pi(x_{z_k} | x_{z_1}, \ldots x_{z_{k-1}})$$

Again,

generation steps $K =$

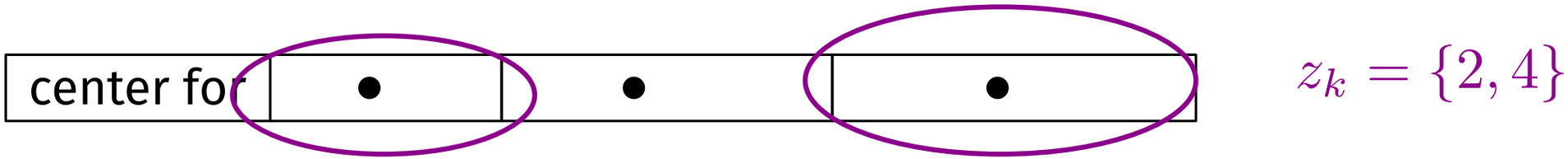
N sentence length

[Kim, Shah, Kontonis, Kakade and Chen, 2025]

Factorized approximation

Iterate for $k = 1, \dots K$

1. Select z_k a set of s_k tokens.

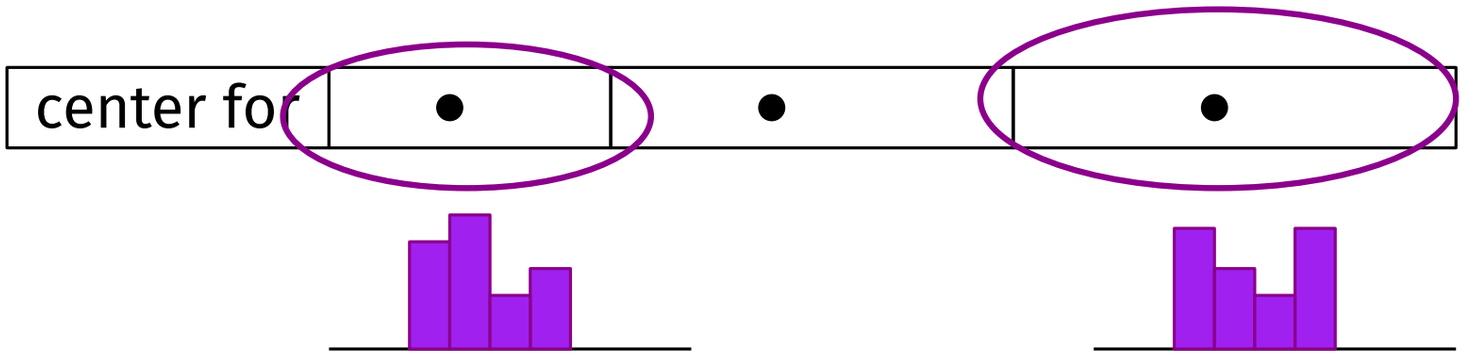


$z_k = \{2, 4\}$

Factorized approximation

Iterate for $k = 1, \dots K$

1. Select z_k a set of s_k tokens.
2. Sample x_i from $\pi(x_i|\mathbf{x}_{z_{<k}})$, independenlty for $i \in z_k$.

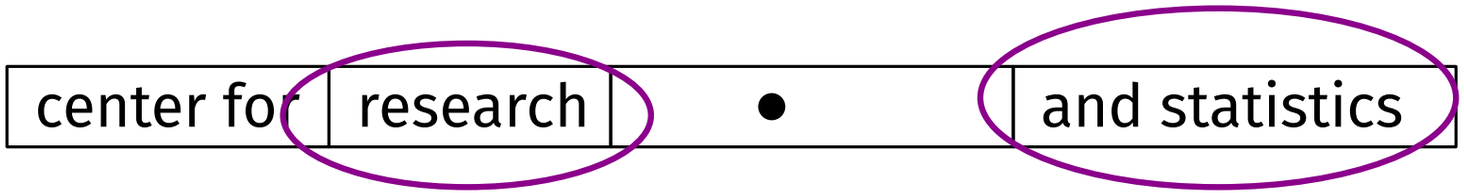


$z_k = \{2, 4\}$

Factorized approximation

Iterate for $k = 1, \dots K$

1. Select z_k a set of s_k tokens.
2. Sample x_i from $\pi(x_i|\mathbf{x}_{z_{<k}})$, independenlty for $i \in z_k$.

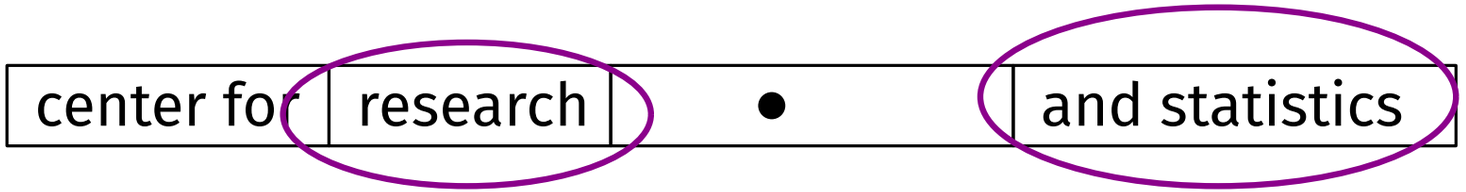


$z_k = \{2, 4\}$

Factorized approximation

Iterate for $k = 1, \dots K$

1. Select z_k a set of s_k tokens.
2. Sample x_i from $\pi(x_i | \mathbf{x}_{z_{<k}})$, independenlty for $i \in z_k$.
3. Set $z_{\leq k} = z_{<k} \cup z_k$ the "unmasked" coordinates.



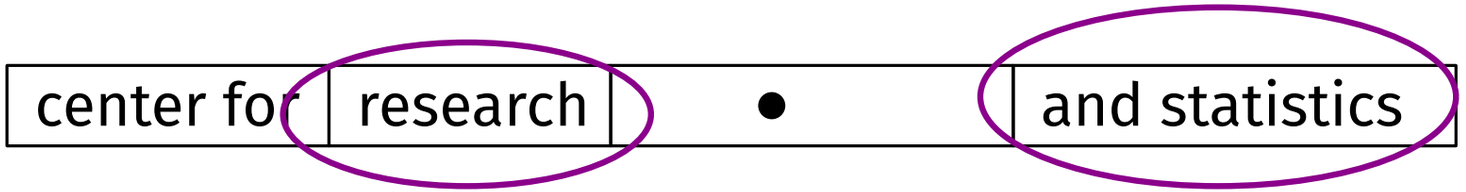
$z_k = \{2, 4\}$

$z_{\leq k} = \{1, 2, 4\}$

Factorized approximation

Iterate for $k = 1, \dots K$

1. Select z_k a set of s_k tokens.
2. Sample x_i from $\pi(x_i | \mathbf{x}_{z_{<k}})$, independenlty for $i \in z_k$.
3. Set $z_{\leq k} = z_{<k} \cup z_k$ the "unmasked" coordinates.



$z_k = \{2, 4\}$

$z_{\leq k} = \{1, 2, 4\}$

✓ K steps $< N$ sentence length

✗ Inexact sampling even if we know conditionals

Wrap-up and setting

$$\pi \in \mathcal{P}(\mathcal{X}^N)$$

N sentence length

$\mathcal{X}$ alphabet (finite)

Wrap-up and setting

Oracle p_θ (obtained via training)

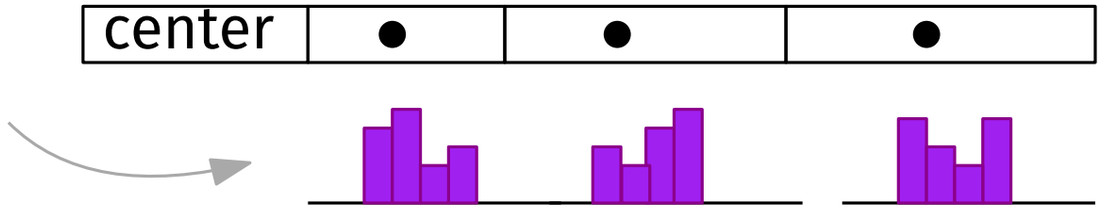
$$p_\theta(x_i|\mathbf{x}_z) \simeq \pi(x_i|\mathbf{x}_z)$$
with $z \subseteq \{1, \dots, N\}$, $i \notin z$.

Same cost to evaluate $p_\theta(\cdot|\mathbf{x}_z) \in \mathbb{R}^{\mathcal{X}}$ for a single i or all $i \notin z$.

$$\pi \in \mathcal{P}(\mathcal{X}^N)$$

N sentence length

$\mathcal{X}$ alphabet (finite)



Wrap-up and setting

Oracle p_θ (obtained via training)

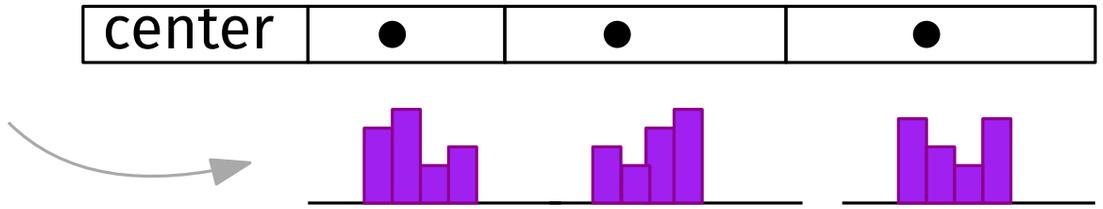
$$p_\theta(x_i|\mathbf{x}_z) \simeq \pi(x_i|\mathbf{x}_z)$$
with $z \subseteq \{1, \dots, N\}$, $i \notin z$.

Same cost to evaluate $p_\theta(\cdot|\mathbf{x}_z) \in \mathbb{R}^{\mathcal{X}}$ for a single i or all $i \notin z$.

$\pi \in \mathcal{P}(\mathcal{X}^N)$

N sentence length

$\mathcal{X}$ alphabet (finite)



1. Initialize $z_0 = \emptyset$.
2. For $k = 1, \dots$, iterate:
   a. Select $z_k \subseteq \{1, \dots, N\} \setminus z_{<k}$.
   b. Sample x_i from $p_\theta(x_i|\mathbf{x}_{z_{<k}})$, independenlty for $i \in z_k$.
   c. Set $z_{\leq k} = z_{<k} \cup z_k$.
3. Until $z_{\leq k} = \{1, \dots\}$. **Return** $\mathbf{x} = (x_1, \dots, x_N)$ and K number of steps.

Wrap-up and setting

Oracle p_θ (obtained via training)

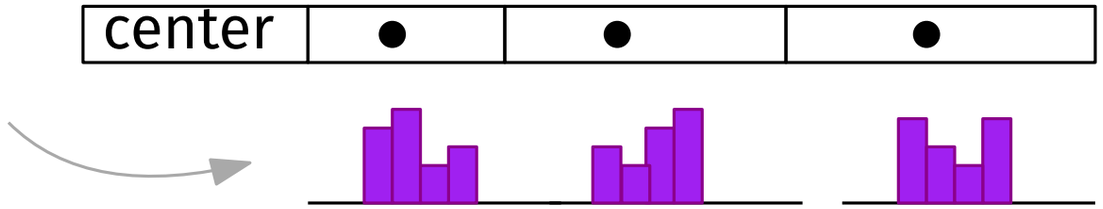
$$p_\theta(x_i|\mathbf{x}_z) \simeq \pi(x_i|\mathbf{x}_z)$$
with $z \subseteq \{1, \dots, N\}$, $i \notin z$.

Same cost to evaluate $p_\theta(\cdot|\mathbf{x}_z) \in \mathbb{R}^{\mathcal{X}}$ for a single i or all $i \notin z$.

$\pi \in \mathcal{P}(\mathcal{X}^N)$

N sentence length

$\mathcal{X}$ alphabet (finite)



1. Initialize $z_0 = \emptyset$.
2. For $k = 1, \dots$, iterate:
   a. Select $z_k \subseteq \{1, \dots, N\} \setminus z_{<k}$.
   b. Sample x_i from $p_\theta(x_i|\mathbf{x}_{z_{<k}})$ independenlty for $i \in z_k$.
   c. Set $z_{\leq k} = z_{<k} \cup z_k$.
3. Until $z_{\leq k} = \{1, \dots\}$. **Return** $\mathbf{x} = (x_1, \dots, x_N)$ and K number of steps.

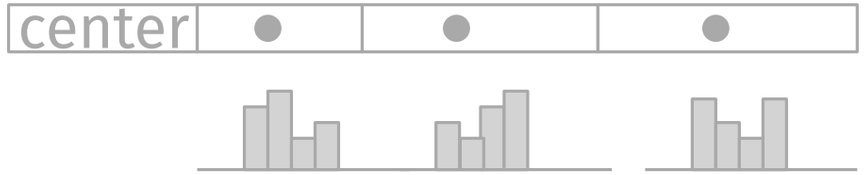
Factorized approximation

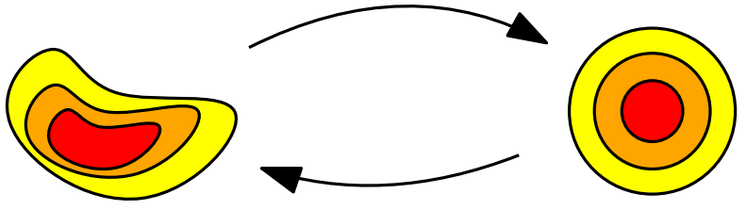
Our questions

Study **computation vs accuracy** trade-off of factorized approximation

- How does it scale in N and K?
- How does it depend on π?
- How to choose the schedule $z_1, \ldots, z_K$ to minimize it?

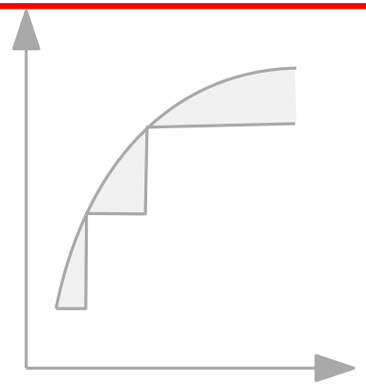
1 - Context: masked diffusion and factorization error





2 - Why ”diffusion”? Analogy with continuous models

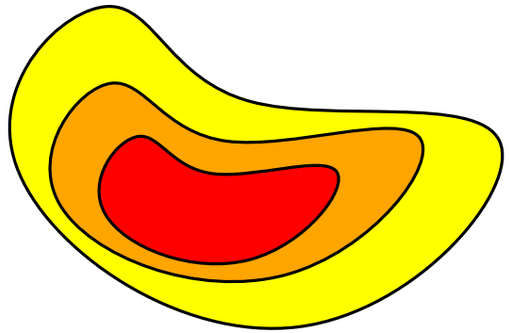
3 - Our results: scaling of the factorization error and optimal schedules



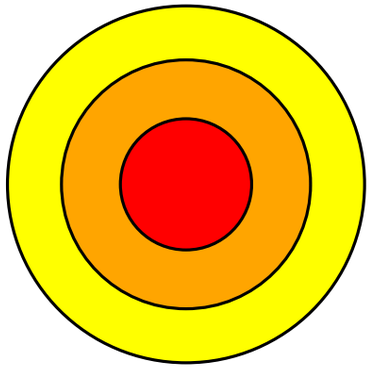
Continuous diffusions

Stochastic process $(X_t)_{t\geq 0}$ with $X_0 \sim \pi$, e.g. SDE

Law π on $\mathbb{R}^d$



noising/forward



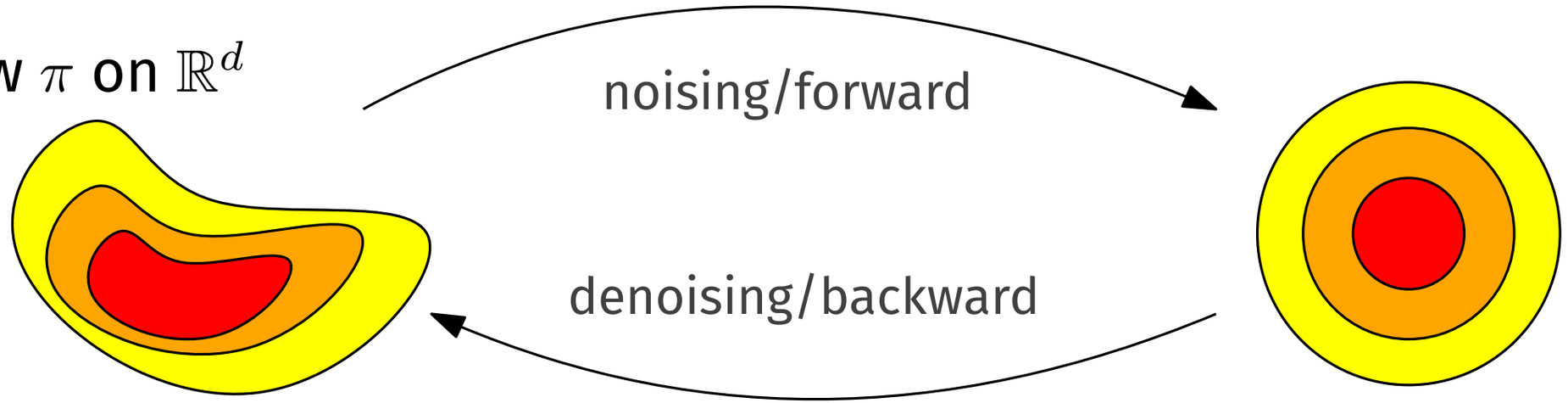
Simple distribution

[Song et al, 2020]

Continuous diffusions

Stochastic process $(X_t)_{t\geq 0}$ with $X_0 \sim \pi$, e.g. SDE

Law π on $\mathbb{R}^d$



Simple distribution

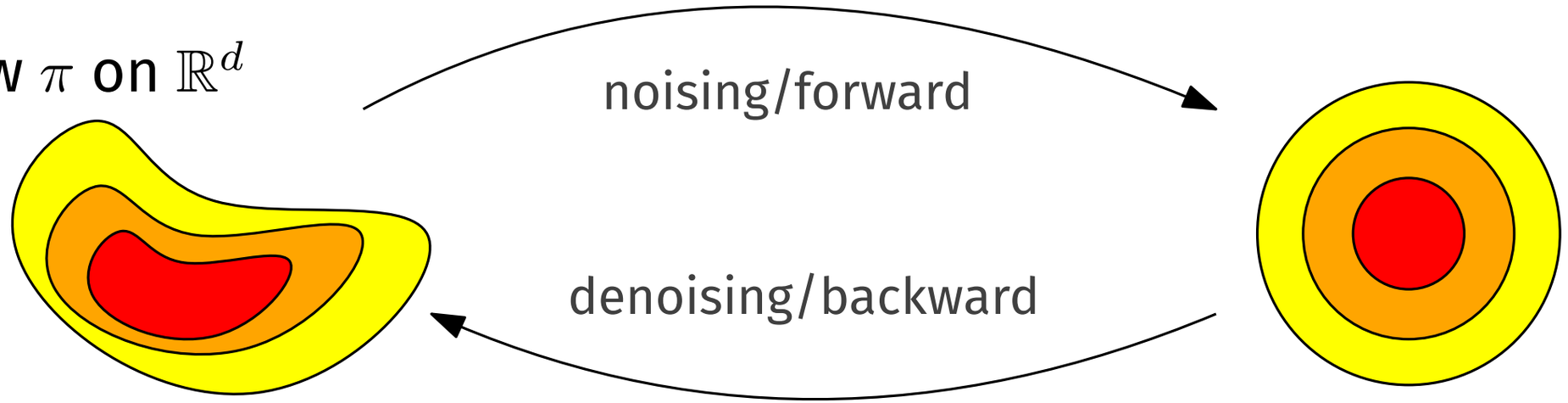
$(Y_t)_{t\geq 0}$ time reversal (in path space) of $(X_t)_{t\geq 0}$

[Song et al, 2020]

Continuous diffusions

Stochastic process $(X_t)_{t\geq 0}$ with $X_0 \sim \pi$, e.g. SDE

Law π on $\mathbb{R}^d$



Simple distribution

$(Y_t)_{t\geq 0}$ time reversal (in path space) of $(X_t)_{t\geq 0}$

Training: learning features of the law of X_t for $t \geq 0$.
Generation: use these features to simulate $(Y_t)_{t\geq 0}$.

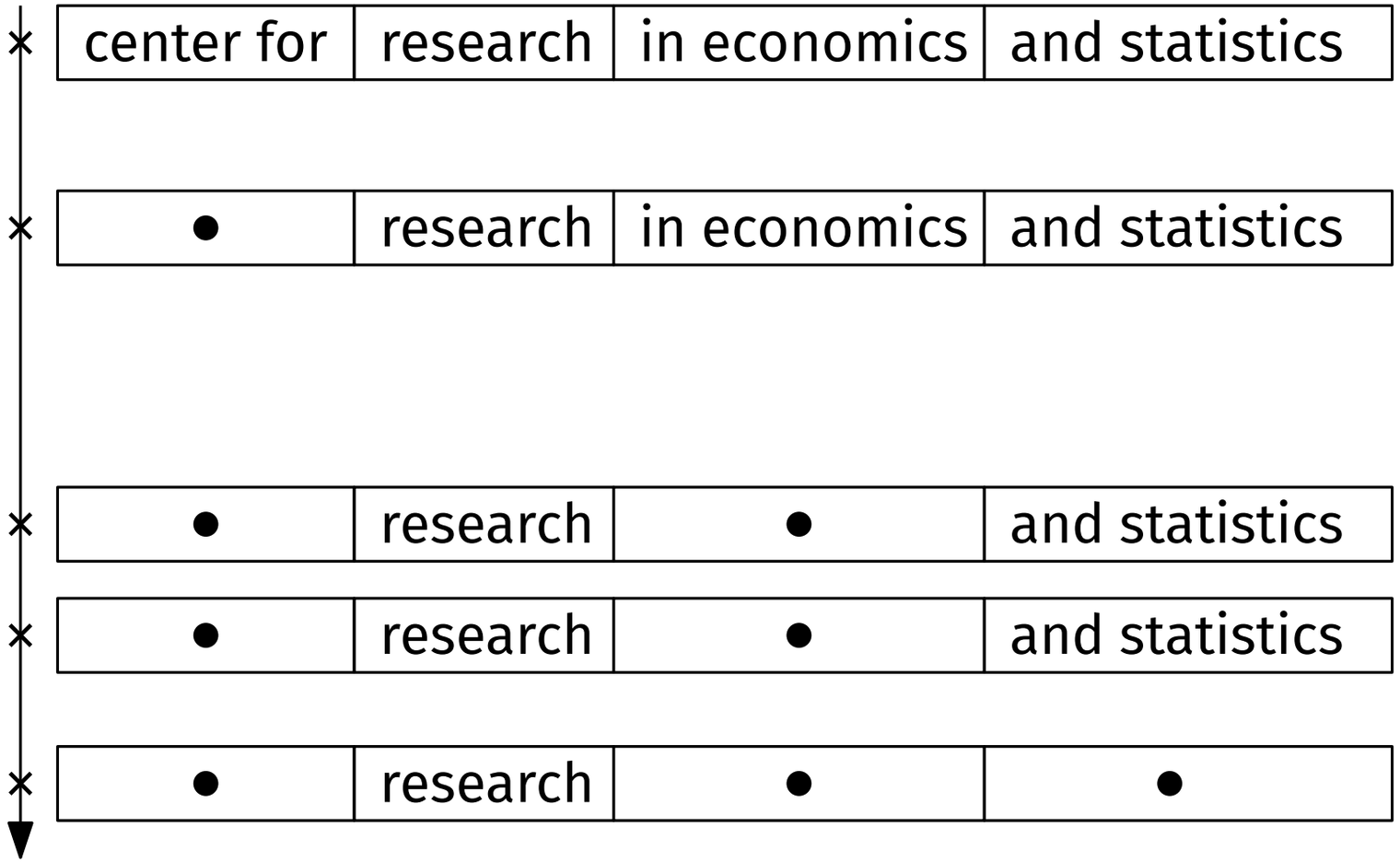
[Song et al, 2020]

Masked discrete diffusions

Noising

1. Initialize $X_0 \sim \pi$ and exponential clock.
2. When clock rings:
   a. Choose i at random,
   b. Substitute i-th component by •,
   c. Start new clock.

Repeat until $\mathbf{x} = (\bullet, \cdots, \bullet)$.



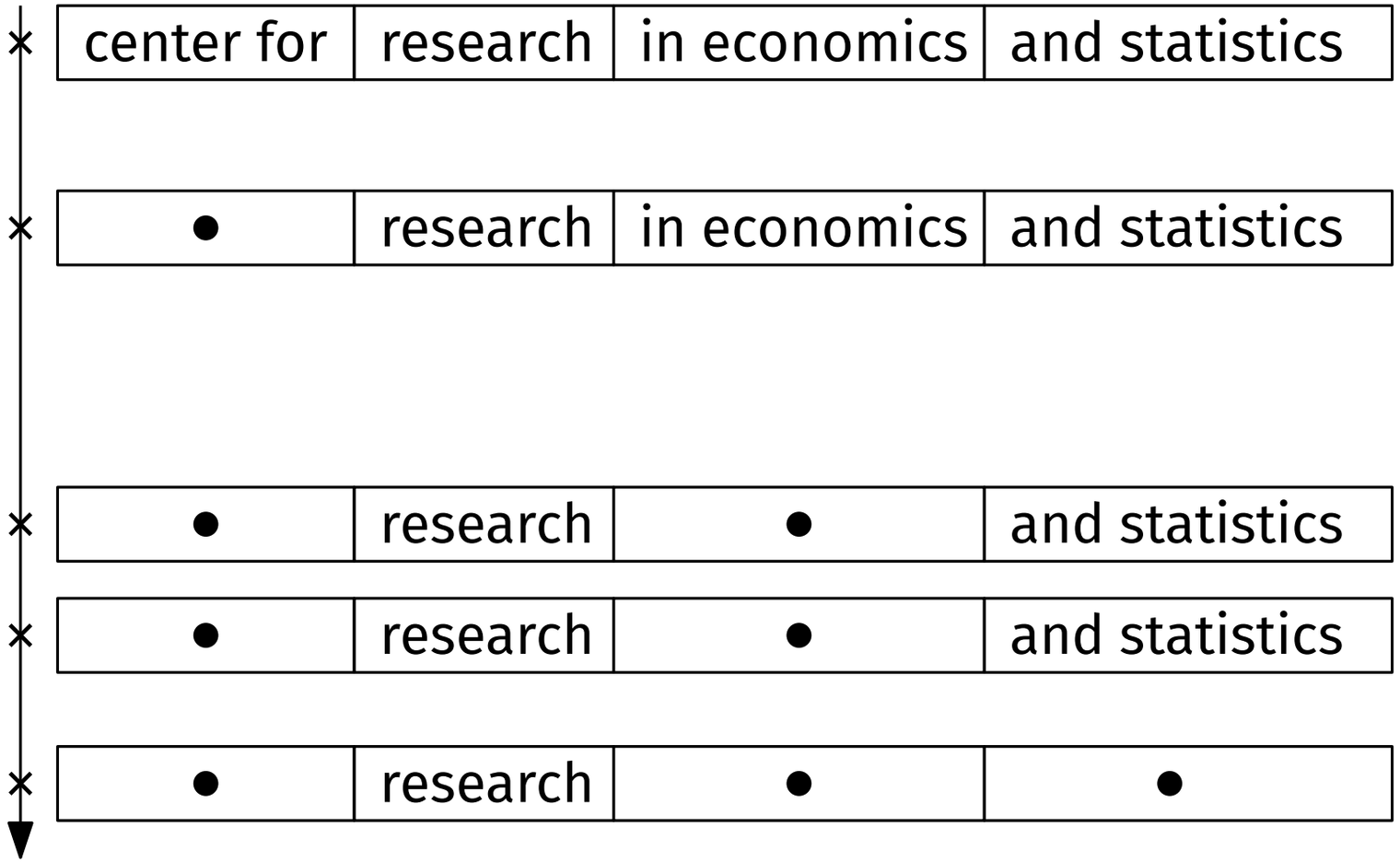
[Austin et al, 2021], [Campbell et al, 2022], [Shi et al, 2024], [Sahoo et al, 2024]...

Masked discrete diffusions

Noising

1. Initialize $X_0 \sim \pi$ and exponential clock.
2. When clock rings:
   a. Choose i at random,
   b. Substitute i-th component by •,
   c. Start new clock.

Repeat until $\mathbf{x} = (\bullet, \cdots, \bullet)$.

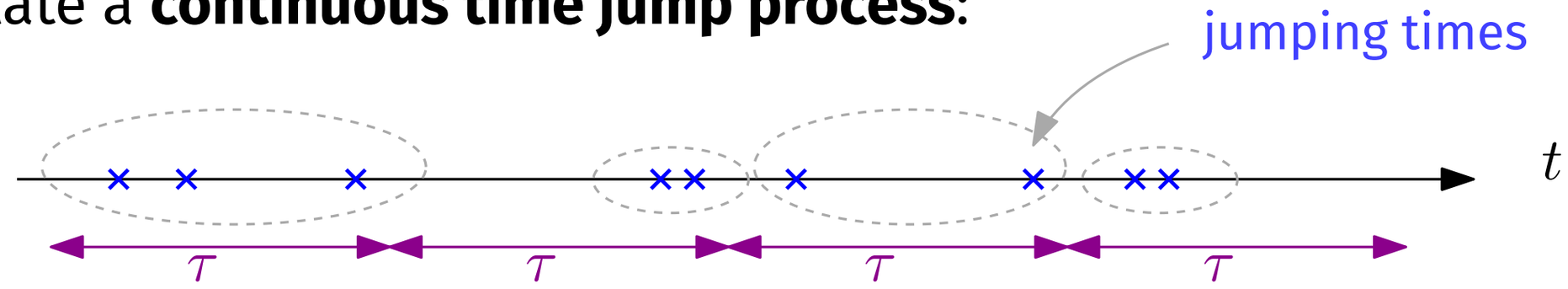


The noising process is a **continuous time jump process over a discrete space**.
The denoising process is the **any-order** generative model $(K = N)$.

[Austin et al, 2021], [Campbell et al, 2022], [Shi et al, 2024], [Sahoo et al, 2024]...

Tau-leaping

To simulate a **continuous time jump process**:

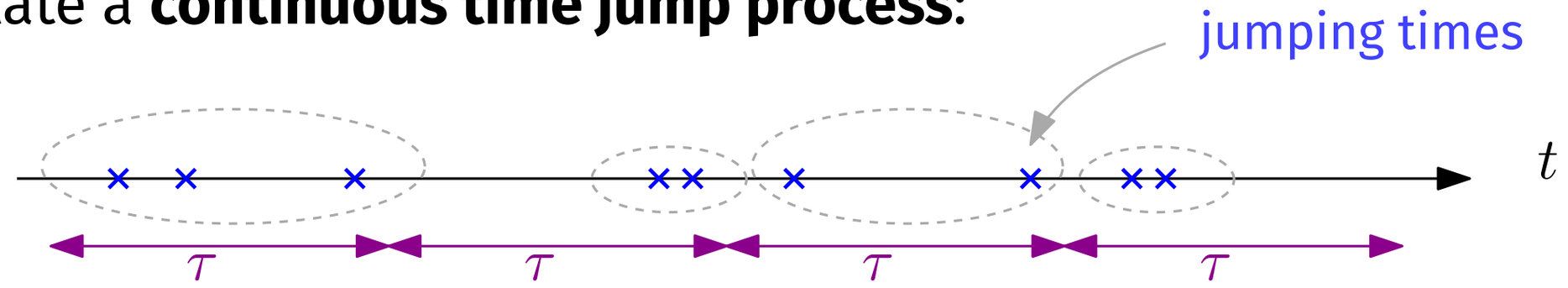


Simulate all jumps in a window of time τ as if they are independent.

[Gillepsie, 2001], [Campbell et al, 2022]

Tau-leaping

To simulate a **continuous time jump process**:



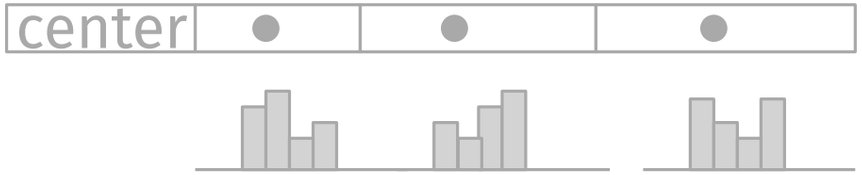
Simulate all jumps in a window of time τ as if they are independent.

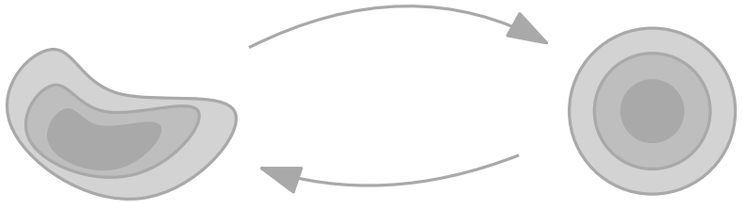
Claim: with τ-leaping in the time reversal of the noising process, you obtain the masked diffusion models with factorized approximation.

Waiting time in forward diffusion and τ give the **schedule**: how many tokens $|z_k|$ are unmasked at each step.

[Gillepsie, 2001], [Campbell et al, 2022]

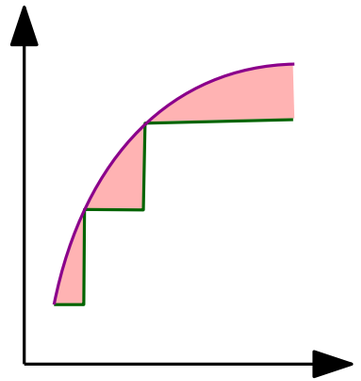
1 - Context: masked diffusion and factorization error





2 - Why ”diffusion”? Analogy with continuous models

3 - Our results: scaling of the factorization error and optimal schedules



Error decomposition

Inputs

- Denoiser $p_\theta(x_i|\mathbf{x}_z)$;
- Planner $\nu_\theta(z_k; \mathbf{x}_{<k}, z_{<k})$.

Two sources of errors

- Denoiser different from $\pi(x_i|\mathbf{x}_z)$;
- Factorized approximation.

Error decomposition

Inputs

- Denoiser $p_\theta(x_i|\mathbf{x}_z)$;
- Planner $\nu_\theta(z_k; \mathbf{x}_{<k}, z_{<k})$.

Two sources of errors

- Denoiser different from $\pi(x_i|\mathbf{x}_z)$;
- Factorized approximation.

Theorem. With $p_{\text{alg}} \in \mathcal{P}(\mathcal{X}^N)$ the output of the algorithm:

$$\mathrm{KL}(\pi|p_{\text{alg}}) \leq E_{\text{learn}} + E_{\text{fact}}$$

$$\mathrm{KL}(\alpha|\beta) = \mathbb{E}_\alpha\left[\log\frac{\alpha}{\beta}\right]$$

Similar results in [Li & Cai, 2025], [Ben-Hamu et al, 2025]

Error decomposition

Inputs

- Denoiser $p_\theta(x_i|\mathbf{x}_z)$;
- Planner $\nu_\theta(z_k; \mathbf{x}_{<k}, z_{<k})$.

Two sources of errors

- Denoiser different from $\pi(x_i|\mathbf{x}_z)$;
- Factorized approximation.

Theorem. With $p_{\text{alg}} \in \mathcal{P}(\mathcal{X}^N)$ the output of the algorithm:

$$\mathrm{KL}(\pi|p_{\text{alg}}) \leq E_{\text{learn}} + E_{\text{fact}}$$

$$E_{\text{learn}} = \mathbb{E}\left[\sum_{k=1}^{K}\sum_{i\in z_k} \mathrm{KL}(\pi(x_i|\mathbf{x}_{z_{<k}})|p_\theta(x_i|\mathbf{x}_{z_{<k}}))\right]$$

$\mathbb{E}$ w.r.t. $\mathbf{x} \sim \pi$ and z_k's given by planner

Similar results in [Li & Cai, 2025], [Ben-Hamu et al, 2025]

Error decomposition

Inputs

- Denoiser $p_\theta(x_i|\mathbf{x}_z)$;
- Planner $\nu_\theta(z_k; \mathbf{x}_{<k}, z_{<k})$.

Two sources of errors

- Denoiser different from $\pi(x_i|\mathbf{x}_z)$;
- Factorized approximation.

Theorem. With $p_{\text{alg}} \in \mathcal{P}(\mathcal{X}^N)$ the output of the algorithm:

$$\mathrm{KL}(\pi|p_{\text{alg}}) \leq E_{\text{learn}} + E_{\text{fact}}$$

$$E_{\text{learn}} = \mathbb{E}\left[\sum_{k=1}^{K}\sum_{i\in z_k} \mathrm{KL}(\pi(x_i|\mathbf{x}_{z_{<k}})|p_\theta(x_i|\mathbf{x}_{z_{<k}}))\right]$$

$$E_{\text{fact}} = \mathbb{E}\left[\sum_{k=1}^{K} \mathrm{TC}_\pi(z_k|\mathbf{x}_{z_{<k}})\right]$$

$\mathbb{E}$ w.r.t. $\mathbf{x}\sim\pi$ and z_k's given by planner

$$\mathrm{TC}_\pi(z_k|\mathbf{x}_{<k}) = \mathrm{KL}\left(\pi(\mathbf{x}_{z_k}|\mathbf{x}_{z_{<k}})\middle|\bigotimes_{i\in z_k}\pi(x_i|\mathbf{x}_{z_{<k}})\right)$$

Similar results in [Li & Cai, 2025], [Ben-Hamu et al, 2025]

Error decomposition

Inputs
- Denoiser $p_\theta(x_i|\mathbf{x}_z)$;
- Planner $\nu_\theta(z_k; \mathbf{x}_{<k}, z_{<k})$.

Two sources of errors
- Denoiser different from $\pi(x_i|\mathbf{x}_z)$;
- Factorized approximation.

Theorem. With $p_{\text{alg}} \in \mathcal{P}(\mathcal{X}^N)$ the output of the algorithm:

$$\mathrm{KL} \ldots$$

$$E_{\text{learn}} = \mathbb{E}\left[\sum_{k=1}^{K}\sum_{i\in z_k} \mathrm{KL}(\pi(x_i \ldots \right]$$

Only depends on π and planner, not denoiser: our focus today

$$E_{\text{fact}} = \mathbb{E}\left[\sum_{k=1}^{K} \mathrm{TC}_\pi(z_k|\mathbf{x}_{z_{<k}})\right]$$

$$\mathrm{TC}_\pi(z_k|\mathbf{x}_{<k}) = \mathrm{KL}\Big(\pi(\mathbf{x}_{z_k}|\mathbf{x}_{z_{<k}})\Big|\bigotimes_{i\in z_k}\pi(x_i|\mathbf{x}_{z_{<k}})\Big)$$

Similar results in [Li & Cai, 2025], [Ben-Hamu et al, 2025]

Worst case bound

Theorem. Schedule with $|z_k|$ constant:

$$E_{\text{fact}} \leq (N-K)D(\pi) \leq (N-K)\log|\mathcal{X}|.$$

$$D(\pi) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_\pi\left[\log\frac{\pi(x_i|\mathbf{x}_{-i})}{\pi(x_i)}\right]$$

sum of total correlation and dual total correlation of π

Worst case bound

Theorem. Schedule with $|z_k|$ constant:

$$E_{\text{fact}} \leq (N-K)D(\pi) \leq (N-K)\log|\mathcal{X}|.$$

$$D(\pi) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_\pi\left[\log\frac{\pi(x_i|\mathbf{x}_{-i})}{\pi(x_i)}\right]$$

sum of total correlation and dual total correlation of π

For any partition $(z_1, \ldots, z_K)$ given, we can build π with equality.

- $x_i = x_j$ under π if i, j belong to same z_k,
- $x_i \perp x_j$ under π if i, j belong to different z_k's.

Not a great bound.

The exchangeable case

Fix only $(s_1, \ldots, s_K)$ with $s_1 + \ldots + s_K = N$ the number of tokens per step.

Exchangeable schedule: choose z_k a subset of $\{1, \ldots, N\} \setminus z_{<k}$ of size s_k uniformly at random.

The exchangeable case

Fix only $(s_1, \ldots, s_K)$ with $s_1 + \ldots + s_K = N$ the number of tokens per step.

Exchangeable schedule: choose z_k a subset of $\{1, \ldots, N\} \setminus z_{<k}$ of size s_k uniformly at random.

Theorem. With $s_1, \ldots, s_K$ deterministic for simplicity,

$$E_{\text{fact}} \leq (\max_k s_k - 1) D(\pi) \leq (\max_k s_k - 1) \log |\mathcal{X}|.$$

scales as N/K

Factor K better than previous bound!

Similar bound in [Li & Cai, 2025]

Elements of proof

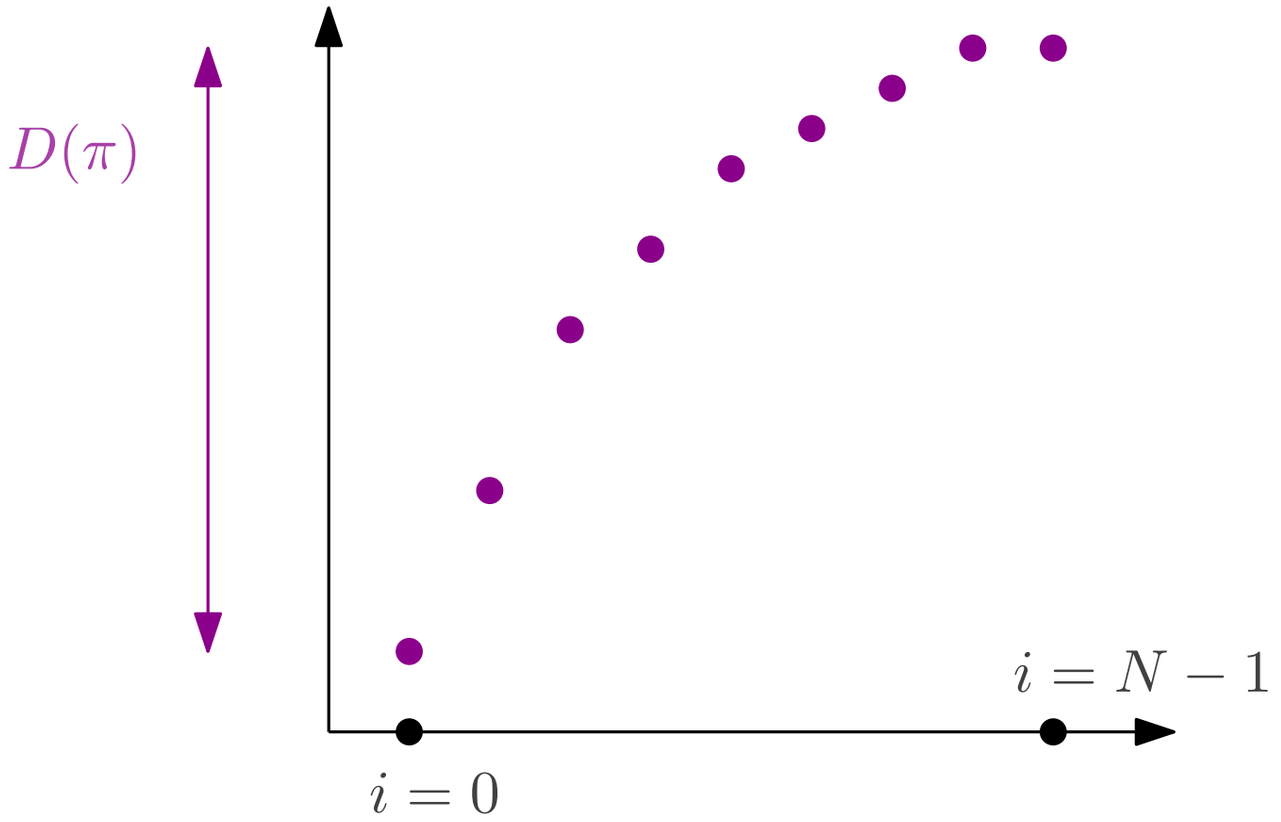
Definition of information profile:

$$f(i) = \mathbb{E}[\log \pi(x_{\sigma_{i+1}} | \mathbf{x}_{\sigma_{\leq i}})]$$

with $\mathbf{x} \sim \pi$, and $\sigma \perp \mathbf{x}$ random permutation of $\{1, \dots, N\}$.

Lemma. f is increasing and

$$f(N-1) - f(0) = D(\pi).$$



Elements of proof

Definition of information profile:

$$f(i) = \mathbb{E}[\log \pi(x_{\sigma_{i+1}} | \mathbf{x}_{\sigma_{\leq i}})]$$

with $\mathbf{x} \sim \pi$, and $\sigma \perp \mathbf{x}$ random permutation of $\{1, \dots, N\}$.

Lemma. f is increasing and

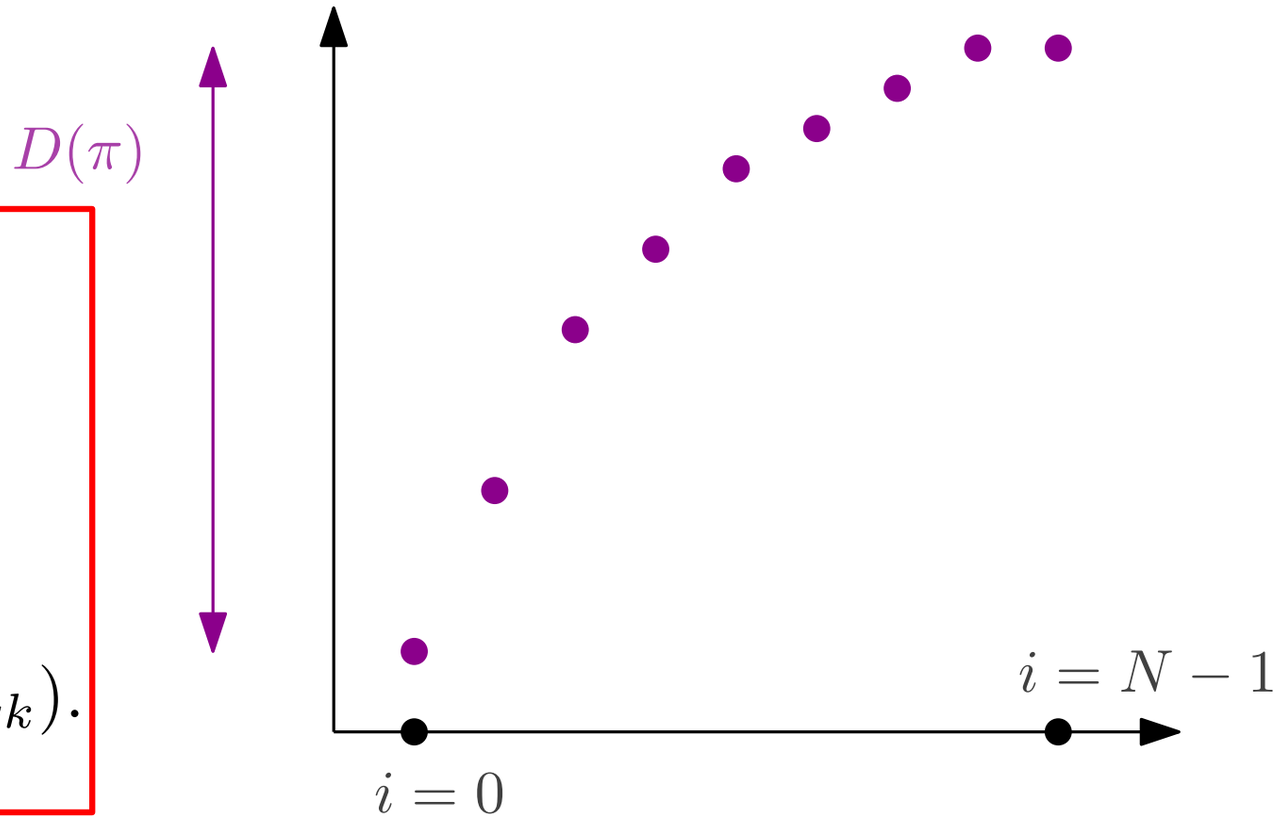
$$f(N-1) - f(0) = D(\pi).$$

Lemma. For exchangeable deterministic schedule, with

$$a_k = |z_{\leq k}| = \sum_{\ell \leq k} s_\ell$$

we have

$$E_{\text{fact}} = \sum_{i=0}^{N-1} f(i) - \sum_{k=0}^{K-1} (a_{k+1} - a_k) f(a_k).$$



Same decomposition in [Chen, Cong & Li, 2025] posted 1 week after ours.

Elements of proof

Definition of information profile:

$$f(i) = \mathbb{E}[\log \pi(x_{\sigma_{i+1}} | \mathbf{x}_{\sigma_{\leq i}})]$$

with $\mathbf{x} \sim \pi$, and $\sigma \perp \mathbf{x}$ random permutation of $\{1, \dots, N\}$.

Lemma. f is increasing and

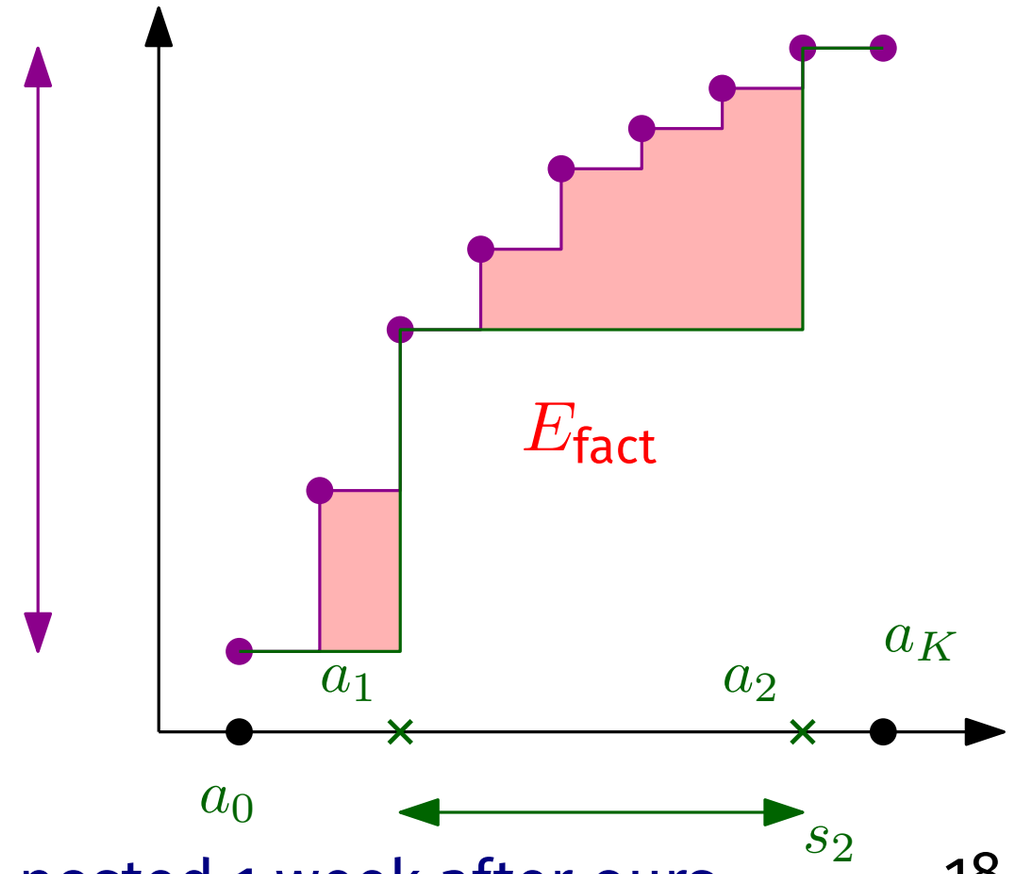
$$f(N-1) - f(0) = D(\pi).$$

Lemma. For exchangeable deterministic schedule, with

$$a_k = |z_{\leq k}| = \textstyle\sum_{\ell \leq k} s_\ell$$

we have

$$E_{\text{fact}} = \sum_{i=0}^{N-1} f(i) - \sum_{k=0}^{K-1} (a_{k+1} - a_k) f(a_k).$$

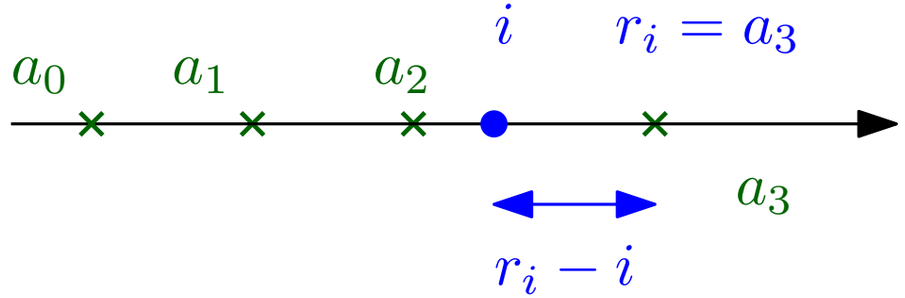


Same decomposition in [Chen, Cong & Li, 2025] posted 1 week after ours.

With this, a one line proof

Algebraic rewriting:

$$E_{\text{fact}} = \sum_{i=1}^{N-1} \Delta f(i)(r_i - i).$$



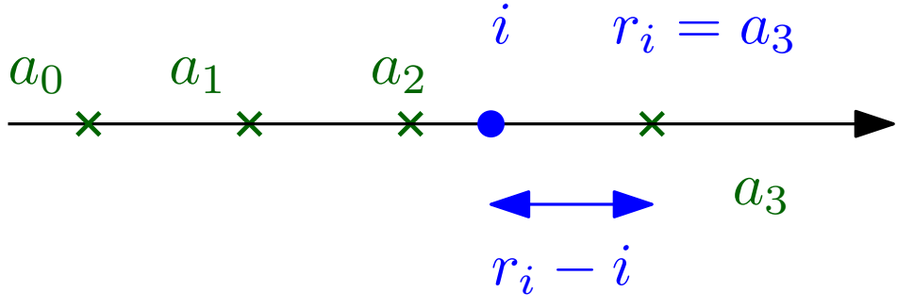
r_i is the first $a_k \geq i$.

$$\Delta f(i) = f(i) - f(i-1)$$

With this, a one line proof

Algebraic rewriting:

$$E_{\text{fact}} = \sum_{i=1}^{N-1} \Delta f(i)(r_i - i).$$



r_i is the first $a_k \geq i$.

$$\Delta f(i) = f(i) - f(i-1)$$

$$0 \leq r_i - i \leq \max_k s_k - 1$$

$$\Delta f \geq 0 \text{ and } \sum_{i=1}^{N-1} \Delta f(i) = D(\pi)$$

$$E_{\text{fact}} \leq (\max_k s_k - 1) D(\pi)$$

Optimizing the schedule

Goal. Given π and K, find $s_1, \ldots, s_K$ with $s_1 + \ldots + s_K = N$ to minimize E_{fact}.

Optimizing the schedule

Goal. Given π and K, find $s_1, \ldots, s_K$ with $s_1 + \ldots + s_K = N$ to minimize E_{fact}.

Recall: with $a_k = \sum_{\ell \leq k} s_\ell$,

$$E_{\text{fact}} = \sum_{i=0}^{N-1} f(i) - \sum_{k=0}^{K-1} (a_{k+1} - a_k) f(a_k).$$

Remark. Knowing f, can be solved in polynomial time with **dynamic programming**.

Optimizing the schedule

Goal. Given π and K, find $s_1, \ldots, s_K$ with $s_1 + \ldots + s_K = N$ to minimize E_{fact}.

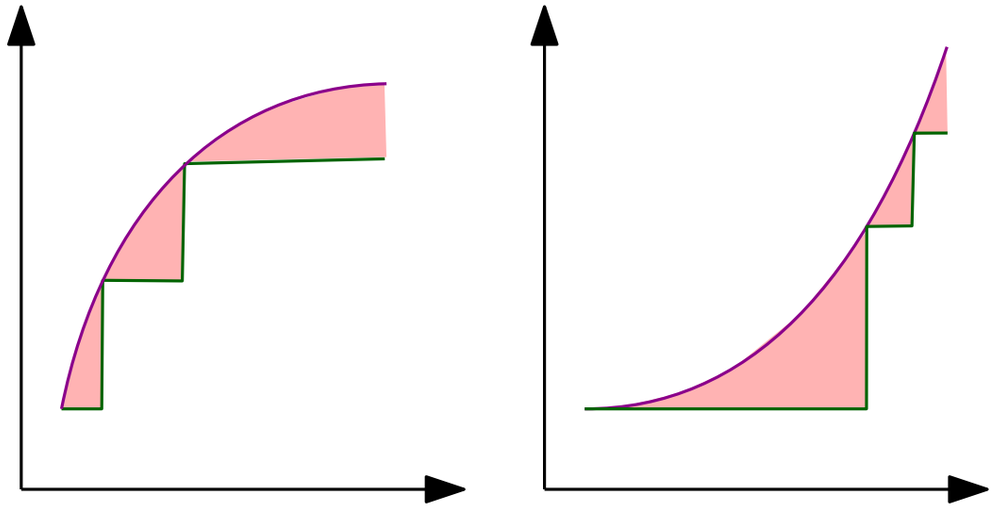
Recall: with $a_k = \sum_{\ell \leq k} s_\ell$,

$$E_{\text{fact}} = \sum_{i=0}^{N-1} f(i) - \sum_{k=0}^{K-1} (a_{k+1} - a_k) f(a_k).$$

Remark. Knowing f, can be solved in polynomial time with **dynamic programming**.

Lemma. Assume f strictly increasing. The optimal $(s_k)_k$ is:

- non-increasing if f convex;
- non-decreasing if f concave.

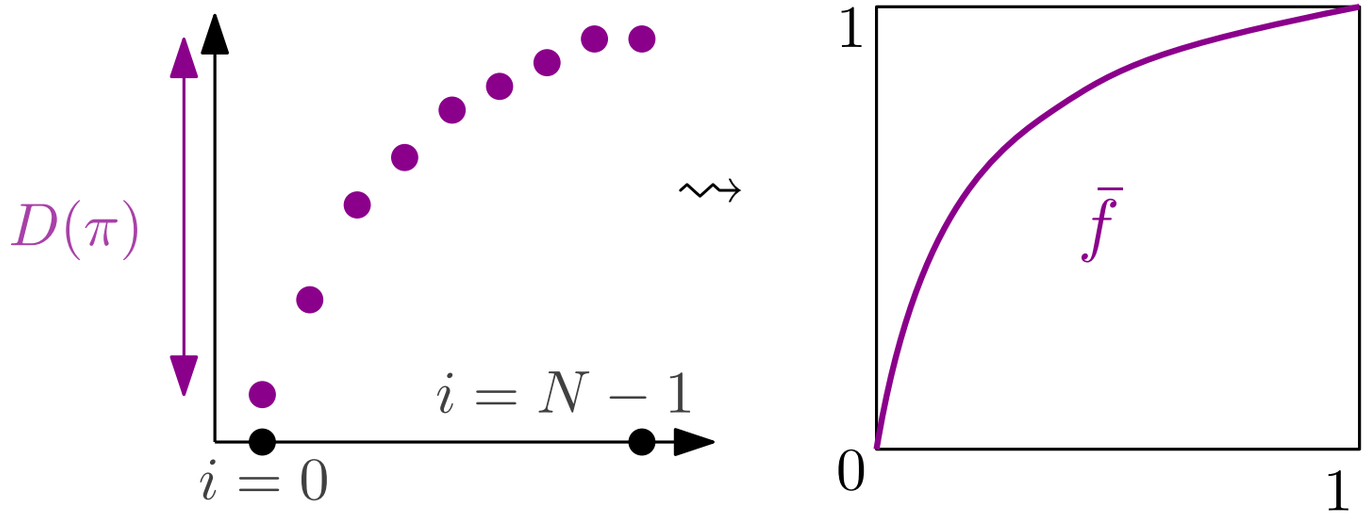


Scaling limit: $N, K \to +\infty$ **with** $N/K \to +\infty$

With $\alpha : [0,1] \to [0,1]$ non-decreasing and C^1, **define** $a_k = \lceil N\alpha_{k/K} \rceil$.

Assume that, in C^1, the rescaled information profile converges:

$$f(i) \simeq f(0) + D(\pi)\bar{f}\left(\frac{i}{N}\right).$$

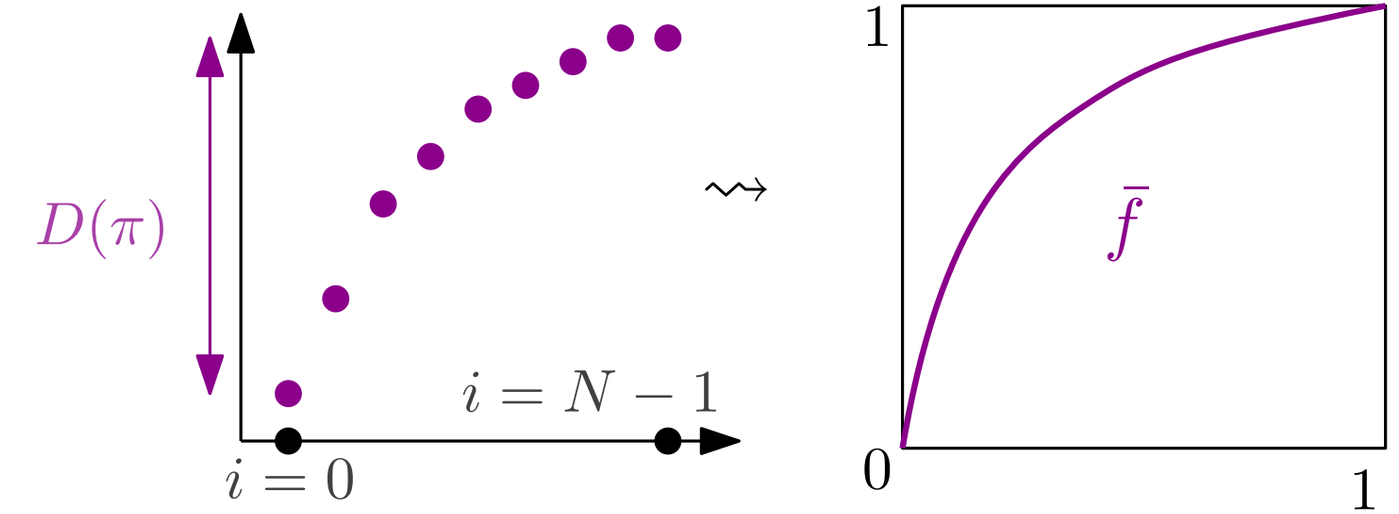


Scaling limit: $N, K \to +\infty$ **with** $N/K \to +\infty$

With $\alpha : [0,1] \to [0,1]$ non-decreasing and C^1, **define** $a_k = \lceil N\alpha_{k/K} \rceil$.

Assume that, in C^1, the rescaled information profile converges:

$$f(i) \simeq f(0) + D(\pi)\bar{f}\left(\frac{i}{N}\right).$$



Theorem. In the limit $N, K \to +\infty$ with $N/K \to +\infty$:

$$E_{\text{fact}} \sim \frac{D(\pi)}{2} \cdot \frac{N}{K} \cdot \int_0^1 \bar{f}'(\alpha_t)\dot{\alpha}_t^2 \, \mathrm{d}t.$$

Complexity of π — Scaling in N, K — Effect of schedule

Consequence: optimal schedule

Proposition. Minimizing $\displaystyle\int_0^1 \bar{f}'(\alpha_t)\dot{\alpha}_t^2 \,\mathrm{d}t$ with $\alpha_0 = 0$ and $\alpha_1 = 1$ yields

$$\alpha_t = G^{-1}(tG(1))$$

with $\displaystyle G(y) = \int_0^y \sqrt{\bar{f}'(z)}\mathrm{d}z$ and the minimal value is $\displaystyle\left(\int_0^1 \sqrt{\bar{f}'(z)}\mathrm{d}z\right)^2$.

Consequence: optimal schedule

Proposition. Minimizing $\displaystyle\int_0^1 \bar{f}'(\alpha_t)\dot{\alpha}_t^2 \,\mathrm{d}t$ with $\alpha_0 = 0$ and $\alpha_1 = 1$ yields

$$\alpha_t = G^{-1}(tG(1))$$

with $G(y) = \displaystyle\int_0^y \sqrt{\bar{f}'(z)}\mathrm{d}z$ and the minimal value is $\displaystyle\left(\int_0^1 \sqrt{\bar{f}'(z)}\mathrm{d}z\right)^2$.

Consequence. Gain from optimizing:

$$\frac{E_{\text{fact}}(\text{optimal sch.})}{E_{\text{fact}}(\text{constant sch.})} \sim \frac{\displaystyle\left(\int_0^1 \sqrt{\bar{f}'(z)}\mathrm{d}z\right)^2}{\displaystyle\int_0^1 \bar{f}'(z)\mathrm{d}z}.$$

In practice if $\hat{f}$ estimator of f,

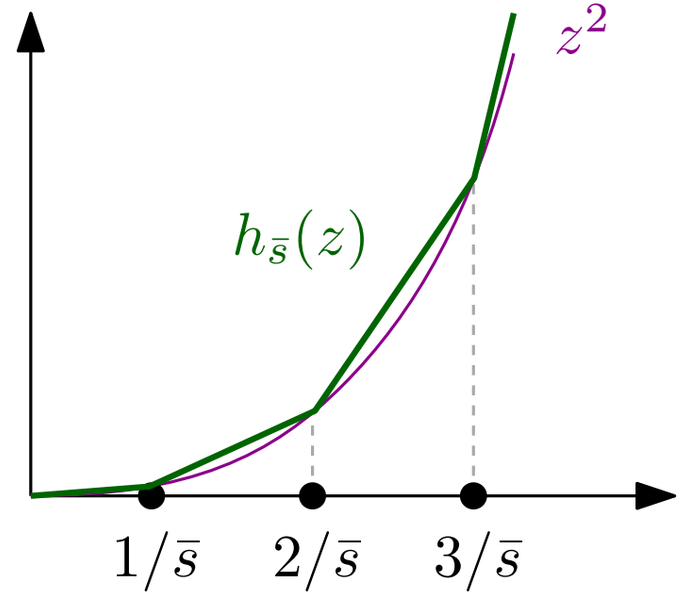
$$s_{k+1} \propto \frac{1}{\sqrt{\Delta\hat{f}(a_k)}}.$$

Remark: the case $N/K \to \bar{s}$

Theorem. In the limit $N, K \to +\infty$ with $N/K \to \bar{s} \in (0, +\infty)$:

$$E_{\text{fact}} \to \frac{D(\pi)}{2} \left(\bar{s} \cdot \int_0^1 \bar{f}'(\alpha_t) h_{\bar{s}}(\dot{\alpha}_t) \, \mathrm{d}t - 1 \right).$$

$h_{\bar{s}}(z)$ piecewise linear and coincide with z^2 for $z \in \{0, 1/\bar{s}, 2/\bar{s}, \ldots\}$



Finding the optimal schedule in practice

π unknown $\rightarrow$ use p_θ to estimate $f(i) = \mathbb{E}[\log \pi(x_{\sigma_{i+1}} | x_{\sigma_{\leq i}})]$.

Ongoing work by:


Carlo Lucibello


Cecilia Secchi


Giacomo Zanella

Finding the optimal schedule in practice

Ongoing work by:

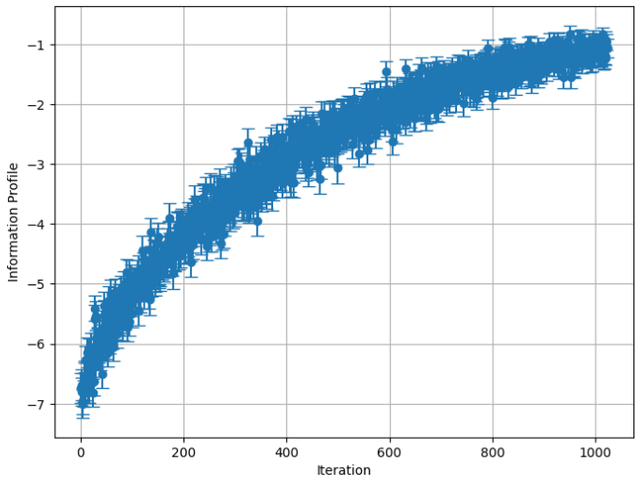

Carlo Lucibello


Cecilia Secchi

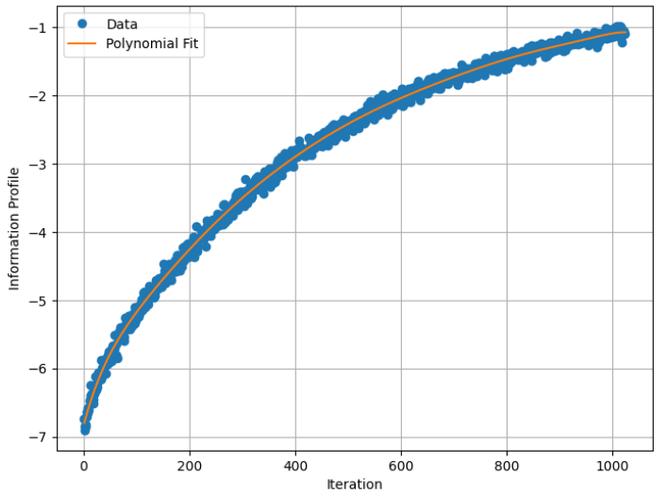

Giacomo Zanella

π unknown $\rightarrow$ use p_θ to estimate $f(i) = \mathbb{E}[\log \pi(x_{\sigma_{i+1}} | x_{\sigma_{\leq i}})]$.

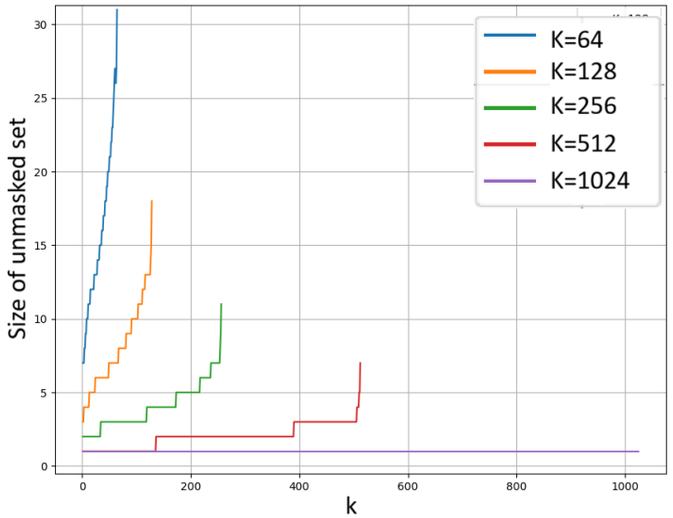
$N = 1024$, language model by Sahoo et al (2024):



Estimating f (100 samples)



Estimating f (1000 samples)

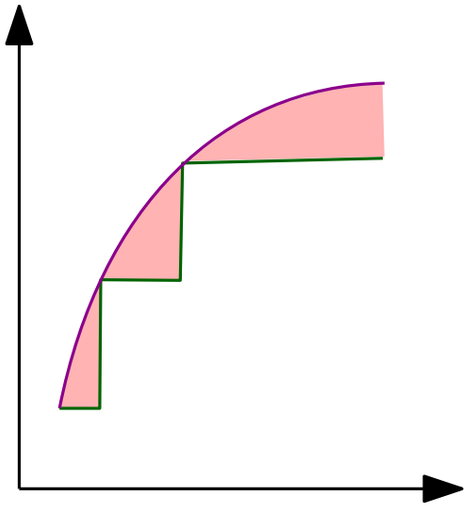


Optimal schedule

Wrap up

What I have presented:

- Simple analysis of masked diffusions.
- Bounds on the factorization error.
- Optimisation of the schedule.



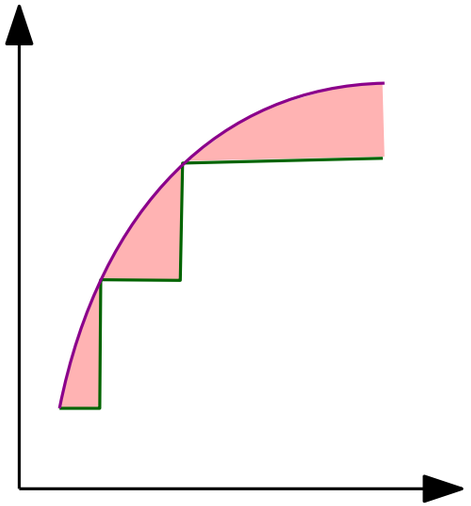
What's next?

- Beyond exchangeable schedules: adaptive planners?
- Error in τ-leaping beyond the masked diffusion?

Wrap up

What I have presented:

- Simple analysis of masked diffusions.
- Bounds on the factorization error.
- Optimisation of the schedule.



What's next?

- Beyond exchangeable schedules: adaptive planners?
- Error in τ-leaping beyond the masked diffusion?

Thank you for your attention