Geometry-Aware Discretization Error of Diffusion Models

Samuel Hurault, Thomas Moreau, Gabriel Peyré

June,19 2026

Generative models

given samples of p_{data}, generate new samples of p_{data}

➡ Choose p_0 easy to sample

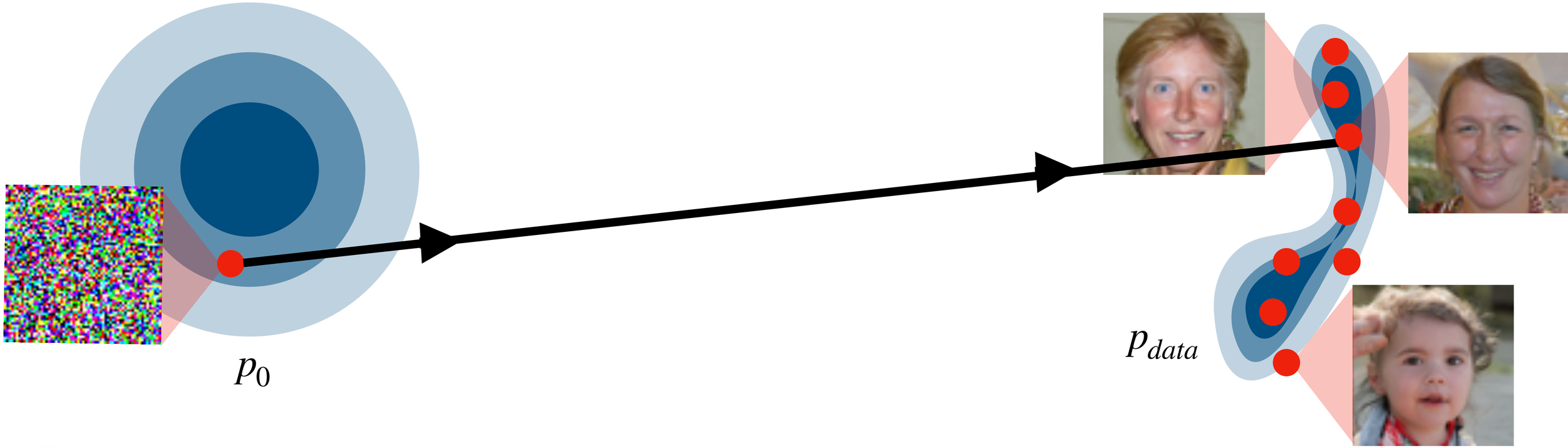
➡ Train T_θ for $T_\theta \# p_0 \approx p_{data}$

$$Y_0 \sim p_0, \quad Y = T_\theta(Y_0)$$

e.g. **Normalizing Flows**

[Rezende, & Mohamed, `15)]

$$\min_\theta KL(T_\theta \# p_0, p_{data})$$

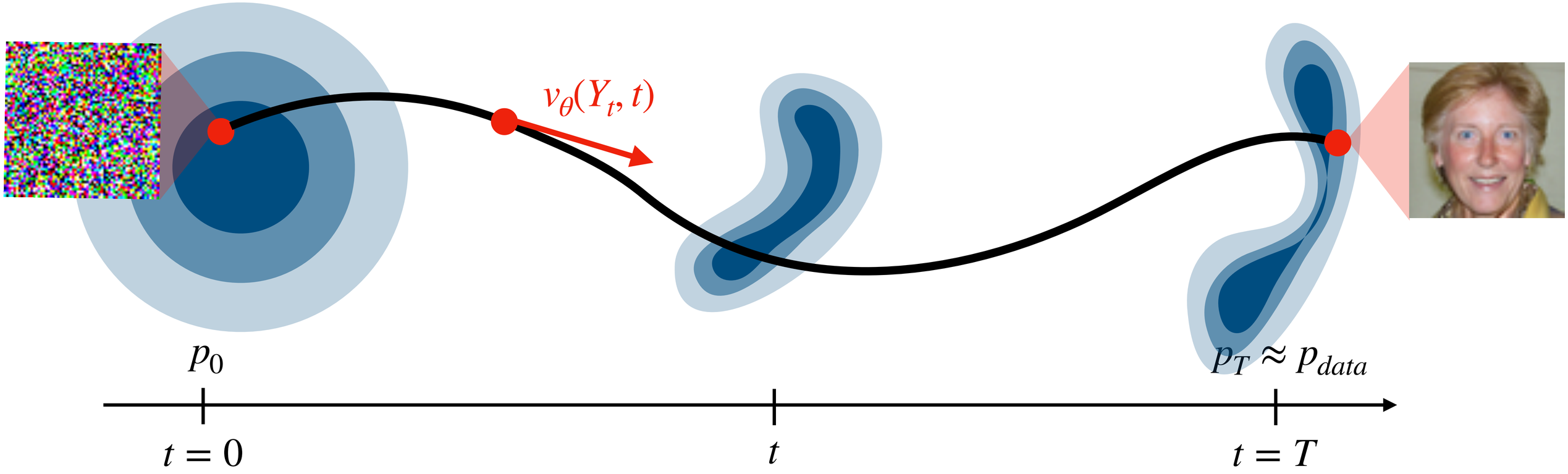


✗ Difficult to train a « single step » model : unstable training, mode collapse …

Generative models

Diffusion models:
[Song et al., `21)]

$$Y_0 \sim p_0, \quad dY_t = v_\theta(Y_t, t)\, dt$$



✓ Continuous transformation
noise → data

✓ Easier sub-training at each time

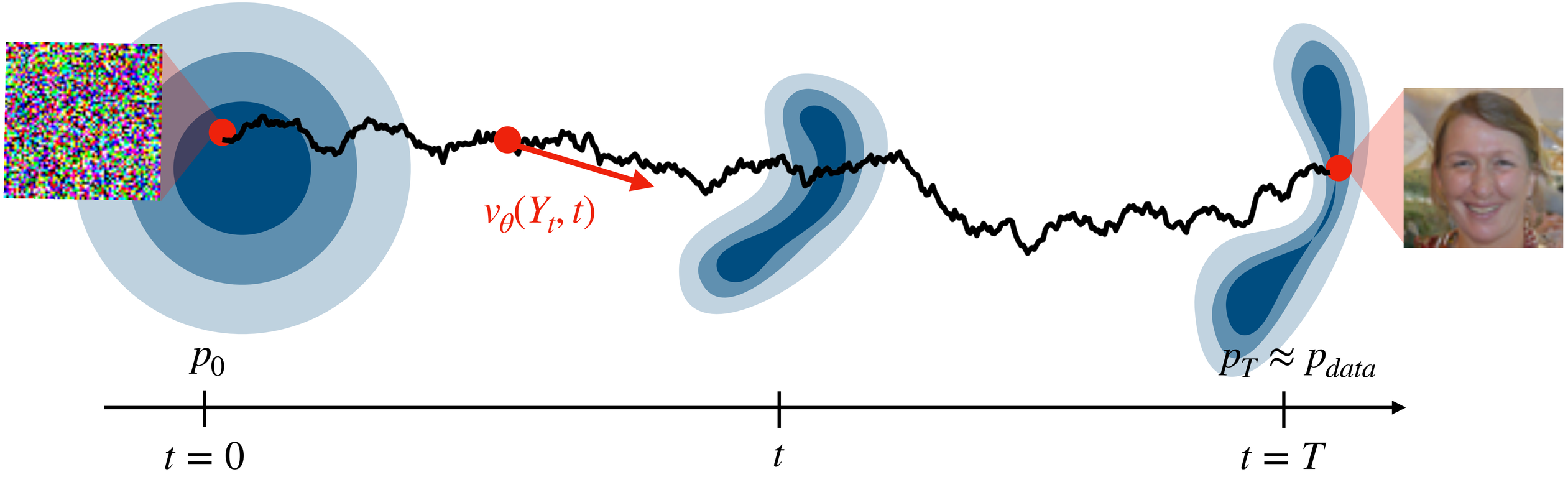
$$\min_\theta \int_0^T \mathbb{E}_{x \sim p_{data}} \| v_\theta(x_t, t) - v^*(x_t, t) \|^2 dt$$

✗ **Discretization error** at inference

Generative models

Diffusion models:
[Song et al., `21)]

$$Y_0 \sim p_0, \quad dY_t = v_\theta(Y_t, t)\,dt + \sqrt{2a_t}\,dW_t$$



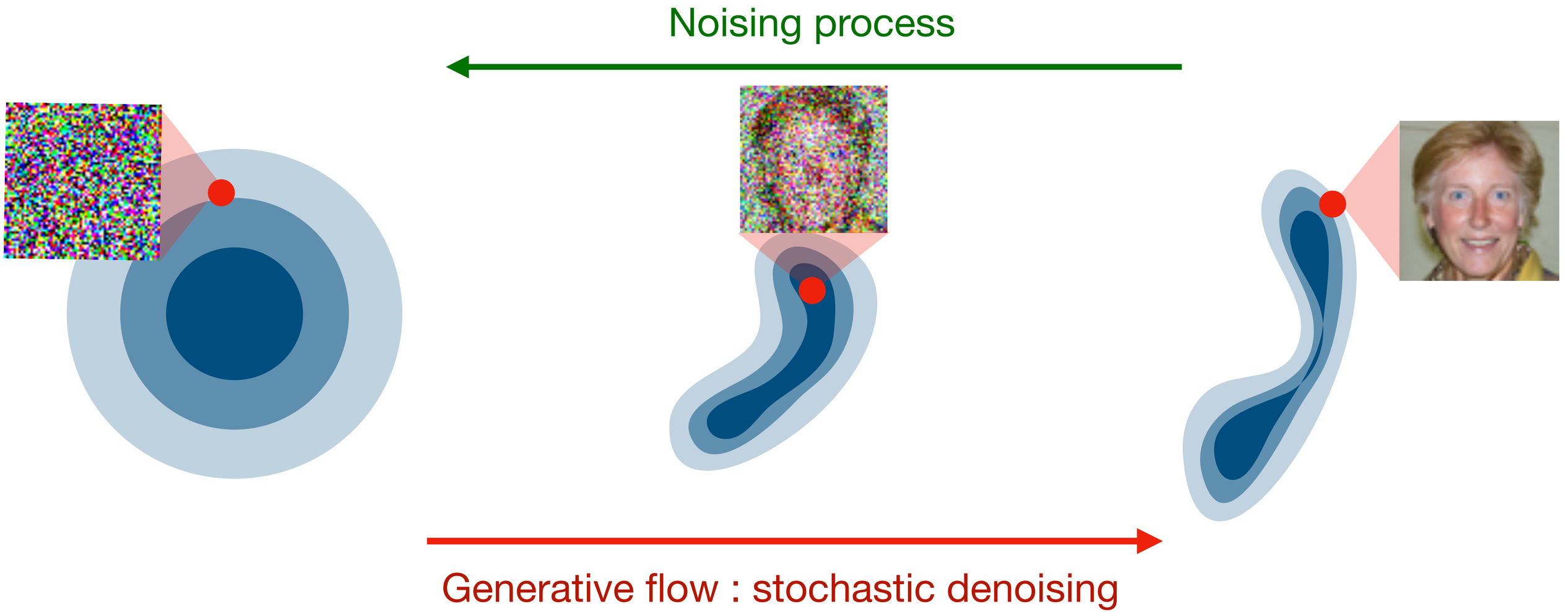
✓ Continuous transformation
noise → data

✓ Easier sub-training at each time

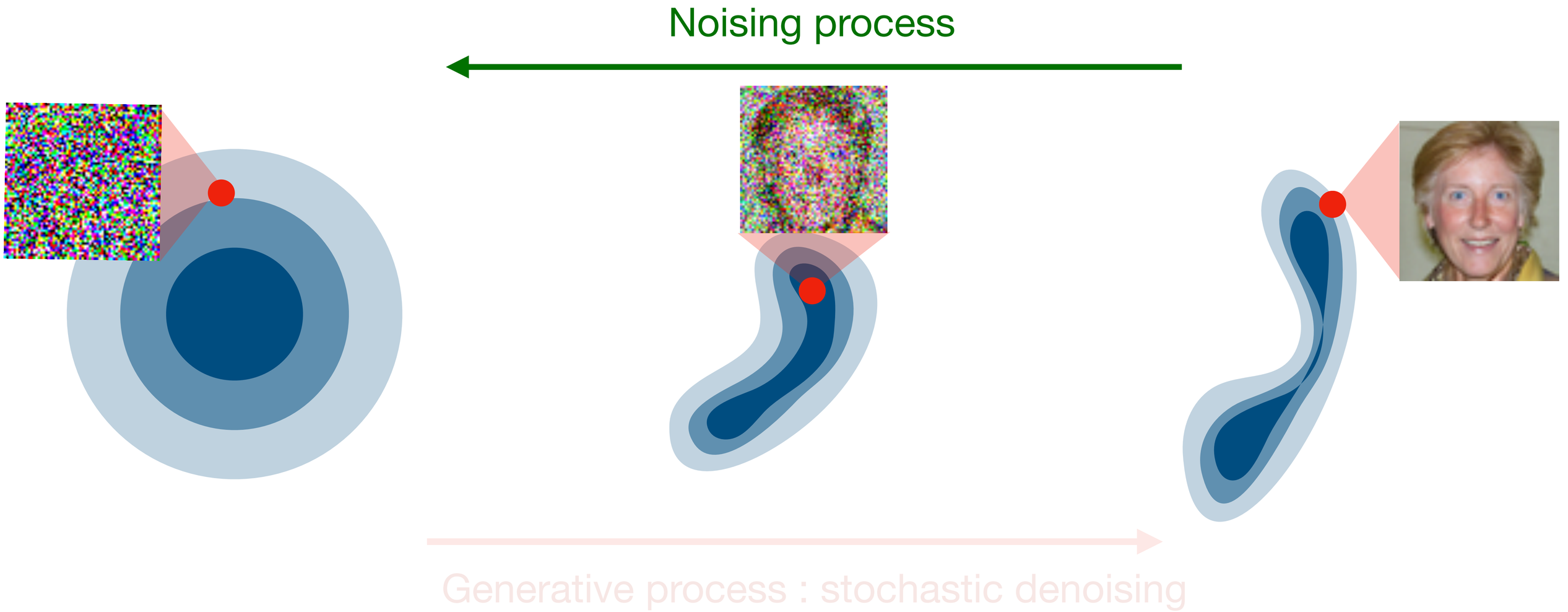
$$\min_\theta \int_0^T \mathbb{E}_{x\sim p_{data}} \|v_\theta(x_t, t) - v^*(x_t, t)\|^2 dt$$

✗ **Discretization error** at inference

Diffusion models



Diffusion models

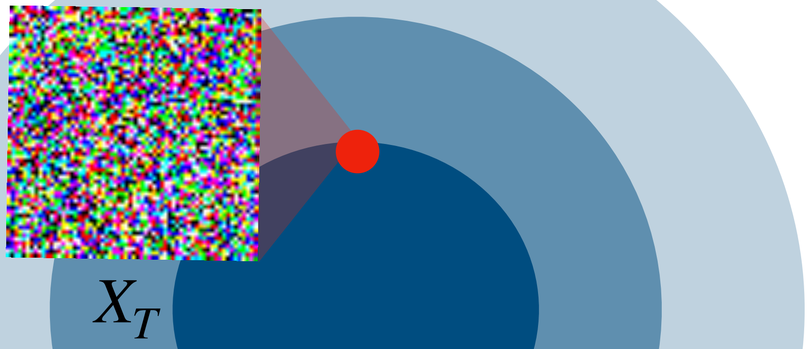


Diffusion models **Noising process** (from data to noise)

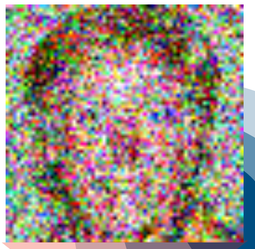
Variance Exploding (VE)

$$X_0 \sim p_{\text{data}}, W_t \sim \mathcal{N}(0, \text{Id}), \quad X_t \overset{d}{=} X_0 + \sigma_t W_t$$

$\sigma_t > 0, \quad \sigma_t \to \infty$



X_T



X_t

X_0

$q_T \approx \mathcal{N}(0, \sigma_T^2 \text{Id})$

$q_t = \text{Law}(X_0 + \sigma_t W_t)$

$q_0 = p_{data}$

$t = T \gg 1$

t

$t = 0$

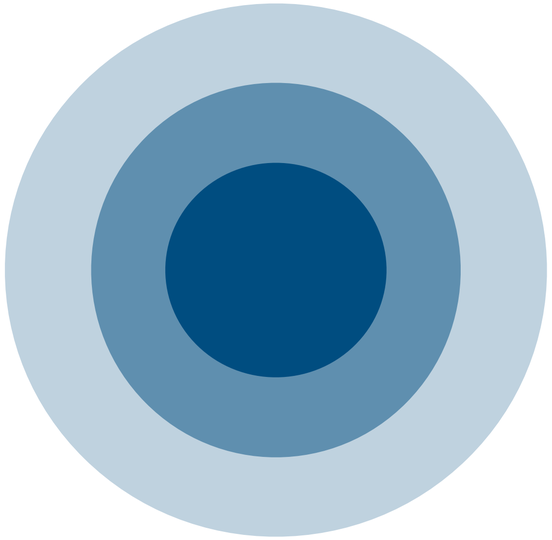
$$\text{Cov}(X_t) = \text{Cov}(X_0) + \sigma_t^2 \text{Id}$$

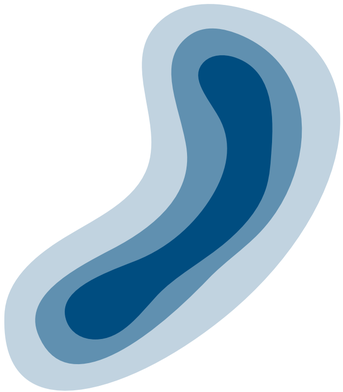
Diffusion models **Noising process** (from data to noise)

Variance Preserving (VP)

$$X_0 \sim p_{\text{data}}, W_t \sim \mathcal{N}(0, \text{Id}), \; X_t \overset{d}{=} \eta_t(X_0 + \sigma_t W_t)$$

$$\sigma_t > 0, \quad \eta_t = \frac{1}{\sqrt{\sigma_t^2 + 1}}$$





$q_T \approx \mathcal{N}(0, \text{Id})$ $\quad$ $q_t = \text{Law}(\eta_t(X_0 + \sigma_t w_t))$ $\quad$ $q_0 = p_{data}$

$t = T \gg 1$ $\quad$ t $\quad$ $t = 0$

$$\text{Cov}(X_t) = \eta_t^2 \, \text{Cov}(X_0) + \left(1 - \eta_t^2\right) \text{Id}$$

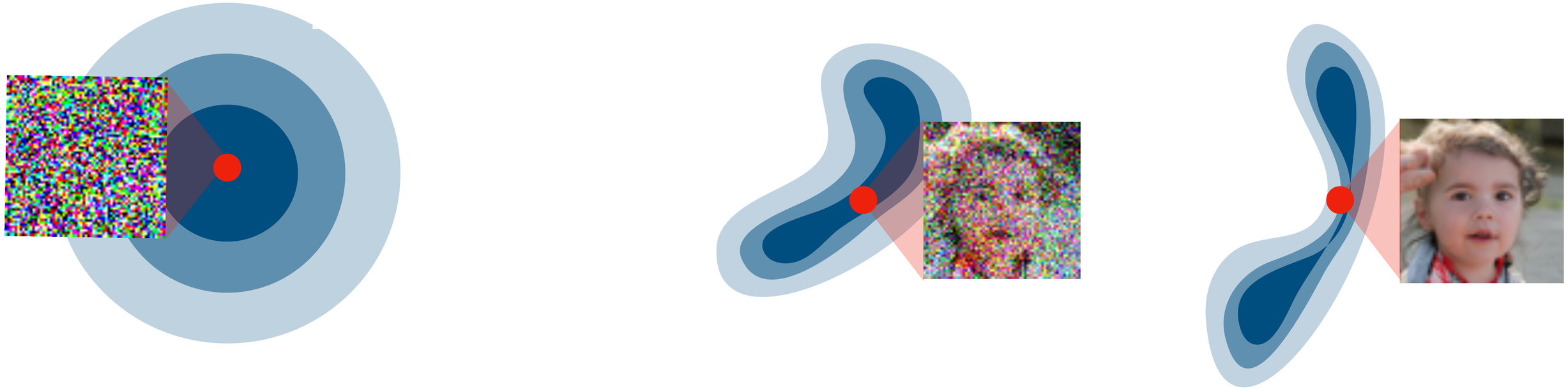
Diffusion models **Noising process** (from data to noise)

General schedule (η_t, σ_t)

$$X_0 \sim p_{\text{data}}, W_t \sim \mathcal{N}(0, \text{Id}), \quad X_t \overset{d}{=} \eta_t(X_0 + \sigma_t W_t)$$

Diffusion models

Noising process



Generative process : stochastic denoising

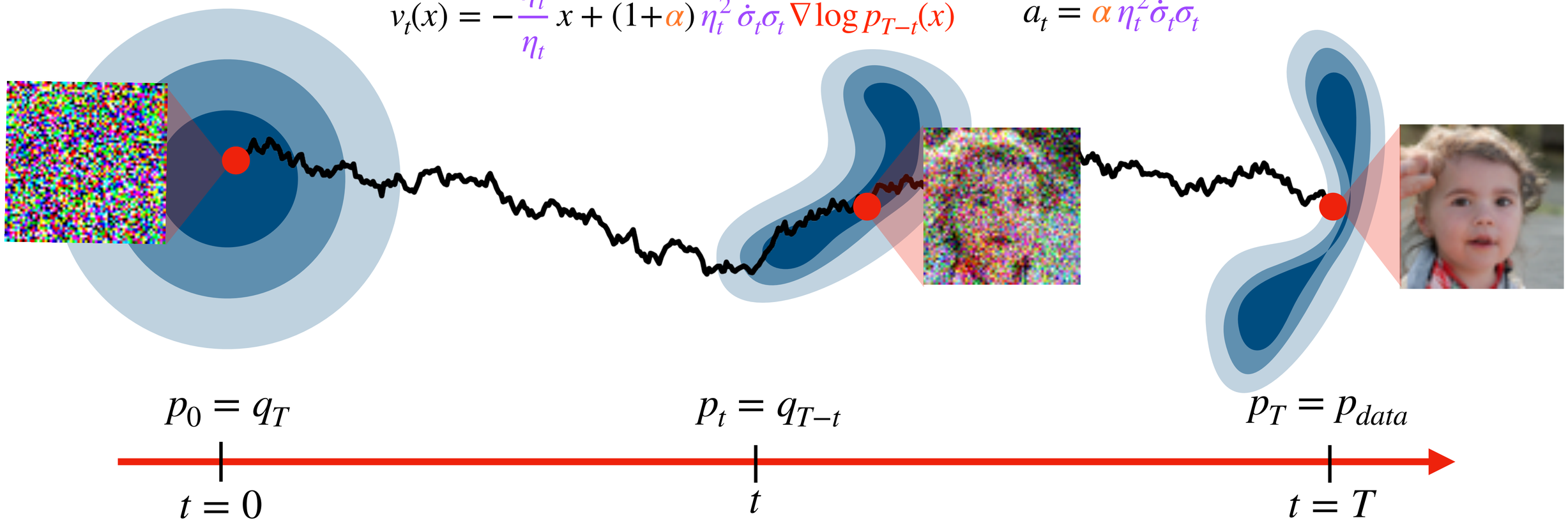
Diffusion models **Generative process** (from noise to data)

$p_t = q_{T-t}$ is sampled from the **SDE**

For $\alpha \geq 0$,

$$Y_0 \sim p_0\,, \quad dY_t = v_t(Y_t)\,dt + \sqrt{2a_t}\,dW_t$$

$$v_t(x) = -\frac{\dot{\eta}_t}{\eta_t}\,x + (1+\alpha)\,\eta_t^2\,\dot{\sigma}_t\sigma_t\,\nabla \log p_{T-t}(x) \qquad a_t = \alpha\,\eta_t^2\dot{\sigma}_t\sigma_t$$



Parameters of the algorithm: diffusion coefficient α, time schedulers η_t, σ_t

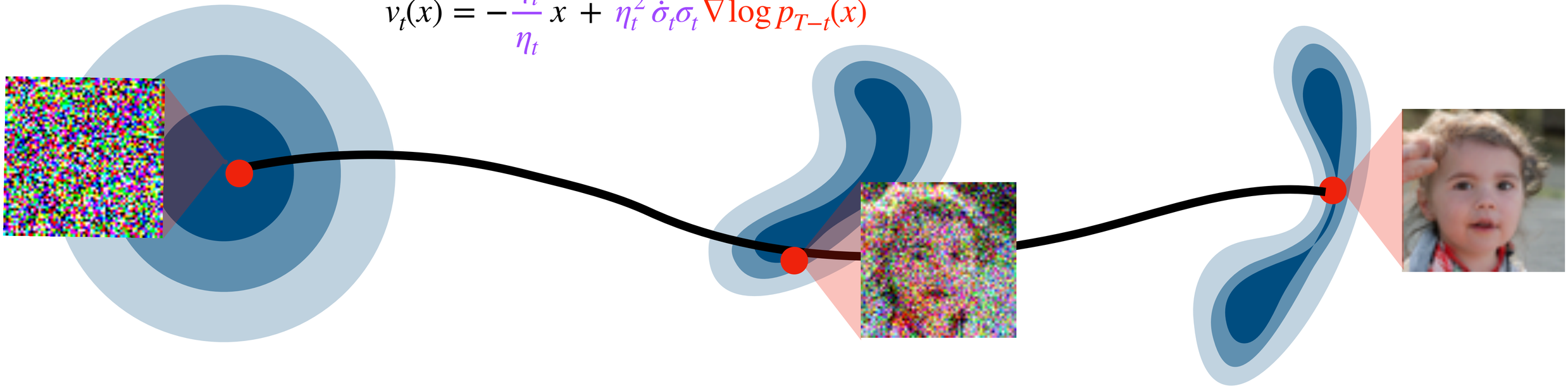
Diffusion models **Generative process** (from noise to data)

$p_t = q_{T-t}$ is sampled from the **ODE**

For $\alpha \geq 0$,

$$Y_0 \sim p_0\,, \quad dY_t = v_t(Y_t)\,dt$$

$$v_t(x) = -\frac{\dot{\eta}_t}{\eta_t}\,x + \eta_t^2\,\dot{\sigma}_t\sigma_t\,\nabla \log p_{T-t}(x)$$



$p_0 = q_T$ $\quad$ $p_t = q_{T-t}$ $\quad$ $p_T = p_{data}$

$t = 0$ $\quad$ t $\quad$ $t = T$

Parameters of the algorithm: diffusion coefficient α, time schedulers η_t, σ_t

Diffusion models errors

$$Y_0 \sim p_0\,, \quad dY_t = v_t(Y_t)\,dt + \sqrt{2a_t}\,dW_t$$

$$v_t(x) = -\frac{\dot{\eta}_{T-t}}{\eta_{T-t}}\,x + (1+\alpha)\,\eta^2_{T-t}\,\dot{\sigma}_{T-t}\sigma_{T-t}\,\nabla \log p_{T-t}(x)$$

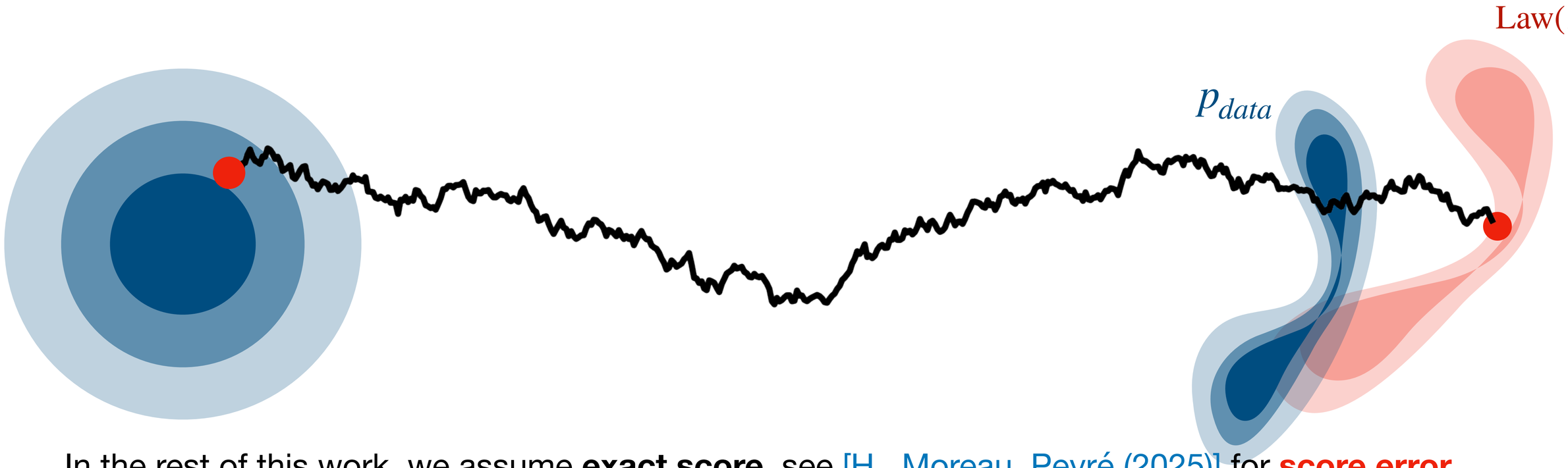
Diffusion models errors: score training error

$$Y_0 \sim p_0\,, \quad dY_t = v_t(Y_t)\,dt + \sqrt{2a_t}\,dW_t$$

$$v_t(x) = -\frac{\dot\eta_t}{\eta_t}\,x + (1+\alpha)\,\eta_t^2\,\dot\sigma_t\sigma_t\,\nabla \log p_{T-t}(x)$$

$$s_t^\theta \approx \nabla \log p_t$$

Trained by denoising Score Matching on a finite dataset $\{x_i \sim p\}_{i \le N_{data}}$

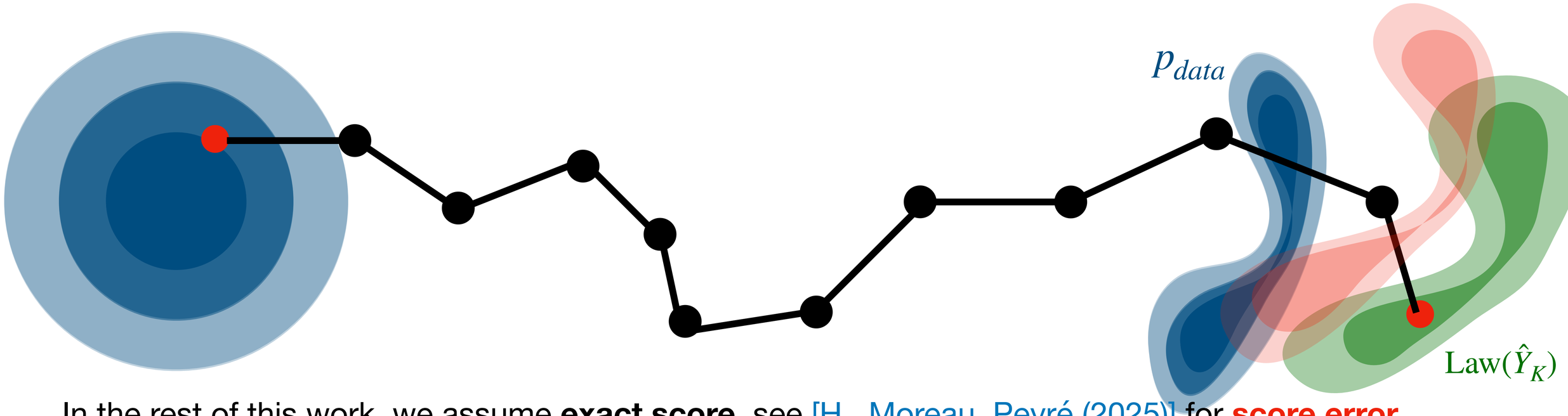


In the rest of this work, we assume **exact score,** see [H., Moreau, Peyré (2025)] for **score error.**

Diffusion models errors: discretization error

Euler discretization: for $\gamma = \frac{T}{K}$,

$$\hat{Y}_0 \sim p_0\,, \quad \hat{Y}_{k+1} = \hat{Y}_k + \gamma\, v_{t_k}(\hat{Y}_k) + \sqrt{2\gamma a_{t_k}}\, W_k$$



In the rest of this work, we assume **exact score,** see [H., Moreau, Peyré (2025)] for **score error.**

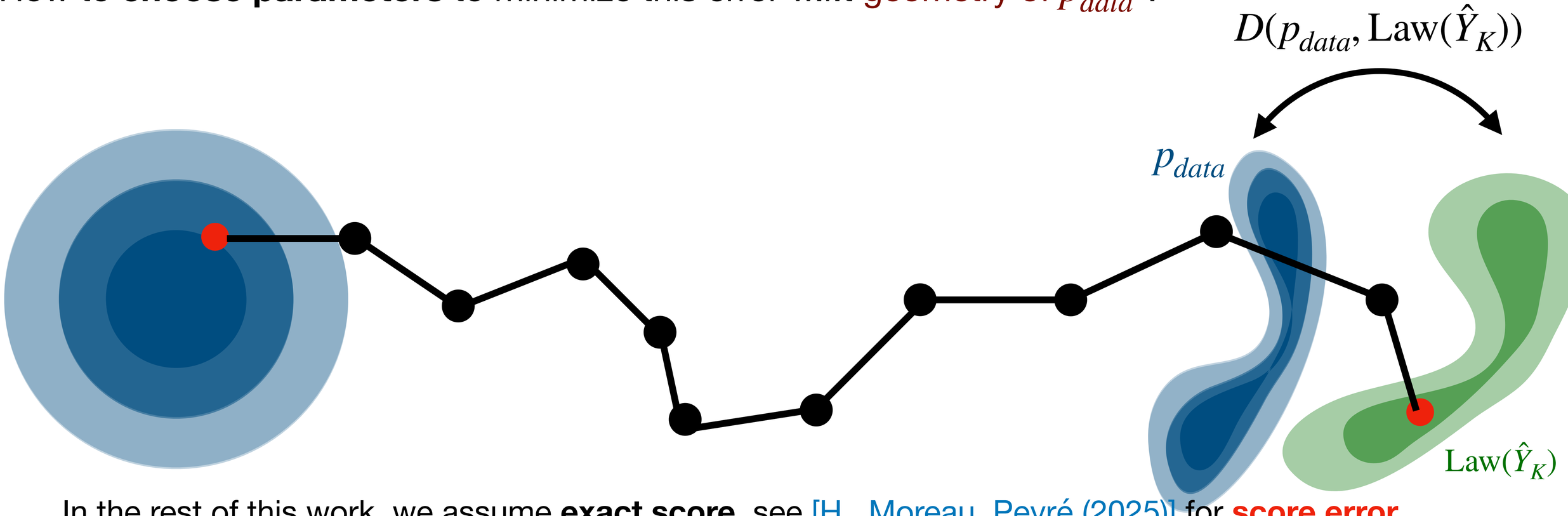
Objective of this work

$$\hat{Y}_0 \sim p_0\,, \quad \hat{Y}_{k+1} = \hat{Y}_k + \gamma\, v_{t_k}(\hat{Y}_k) + \sqrt{2\gamma a_{t_k}}\, W_k$$

We calculate explicitly how **discretization error** affects sampling:

$D(p_{data}, \mathrm{law}(\hat{Y}_K))$ w.r.t. **diffusion parameters** and geometry of p_{data}

- How to **choose parameters** to minimize this error w.r.t geometry of p_{data} ?



In the rest of this work, we assume **exact score,** see [H., Moreau, Peyré (2025)] for **score error.**

Objective of this work

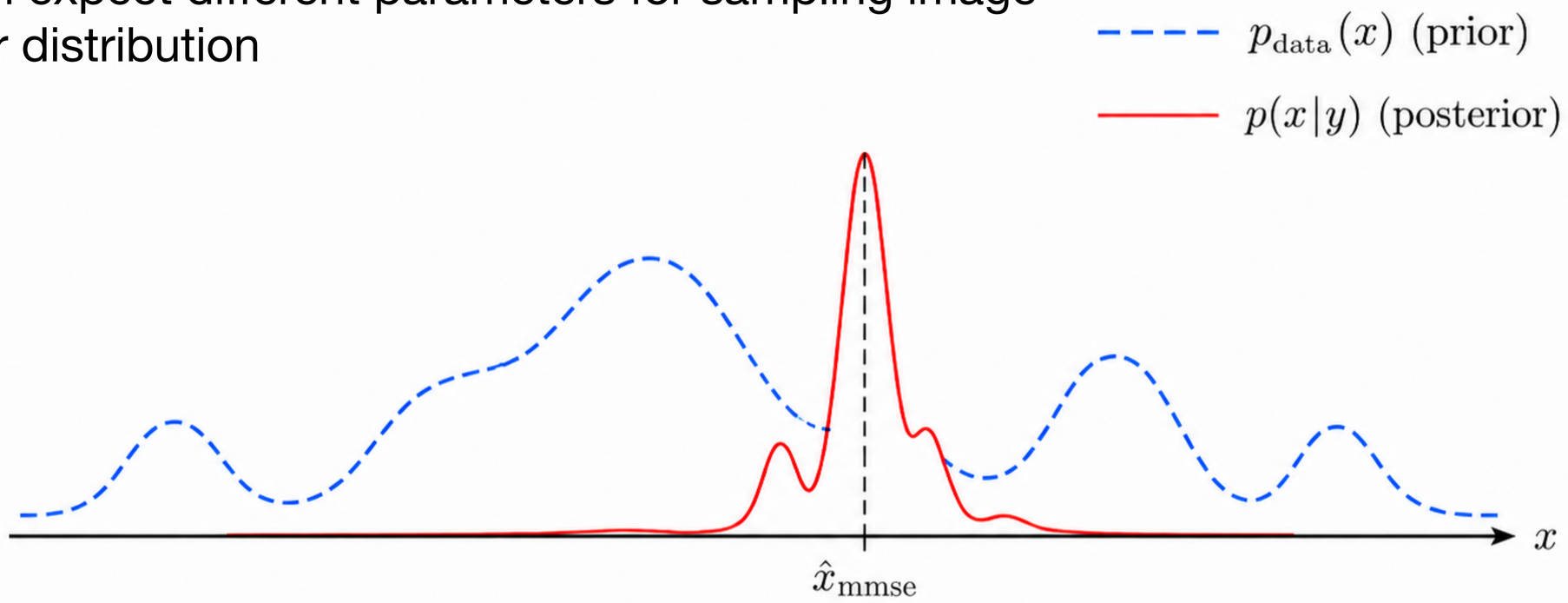
$$\hat{Y}_0 \sim p_0\,, \quad \hat{Y}_{k+1} = \hat{Y}_k + \gamma\, v_{t_k}(\hat{Y}_k) + \sqrt{2\gamma a_{t_k}}\, W_k$$

We calculate explicitly how **discretization error** affects sampling:

$D(p_{data}, \mathrm{law}(\hat{Y}_K))$ w.r.t. **diffusion parameters** and geometry of p_{data}

- How to **choose parameters** to minimize this error w.r.t geometry of p_{data} ?

Example: we can expect different parameters for sampling image prior or posterior distribution



Objective of this work

$$\hat{Y}_0 \sim p_0\,, \quad \hat{Y}_{k+1} = \hat{Y}_k + \gamma\, v_{t_k}(\hat{Y}_k) + \sqrt{2\gamma a_{t_k}}\, W_k$$

We calculate explicitly how **discretization error** affects sampling:

$D(p_{data}, \mathrm{law}(\hat{Y}_K))$ w.r.t. **diffusion parameters** and geometry of p_{data}

- How to **choose parameters** to minimize this error w.r.t geometry of p_{data} ?

Existing works:

*In a **general setting,*** **upper bounds** e.g. [Benton et al.]

$$D(p_{\mathrm{data}}, \mathrm{Law}(\hat{Y}_K)) \lesssim \gamma \mathscr{E}_{disc}$$

$\mathscr{E}_{disc}$ w.r.t. dim, Lipschitz const. of score, bound on p_{data}…

- Not tight
- No influence of geometry of the data.

Not adapted for parameter design

This work:

*Assuming **Gaussian data*** $p_{data} \sim \mathcal{N}(\mu_{data}, \Sigma_{data})$

$$W_2(p_{data}, \mathrm{law}(\hat{Y}_K)) = \gamma \mathscr{E}_{disc} + o(\gamma)$$

with $\mathscr{E}_{disc}$ explicit w.r.t.

- sampling parameters $(\alpha, \sigma_t, \eta_t)$
- geometry of the data : spectrum of Σ_{data}

Adapted for parameter

Discretization error: Gaussian case

Assuming $p_{\text{data}} = \mathcal{N}\left(0, \Sigma_{\text{data}} = U\,\text{Diag}(\lambda_i)U^\top\right)$,

$$\hat{Y}_0 \sim p_0\,, \quad \hat{Y}_{k+1} = \hat{Y}_k + \gamma\, v_{t_k}(\hat{Y}_k) + \sqrt{2\gamma a_{t_k}} W_k$$

$$v_t(x) = H_t x \quad \text{with} \quad H_t = U\,\text{Diag}\Big(\Delta(\lambda_i)\Big)_i U^\top$$

- Linear velocity $v_t \rightarrow$ samples a Gaussian $\text{Law}(\hat{Y}_K) = \mathcal{N}\left(\mathbb{E}[\hat{Y}_K], \text{Cov}[\hat{Y}_K]\right)$
- v_t decomposes per eigendirections

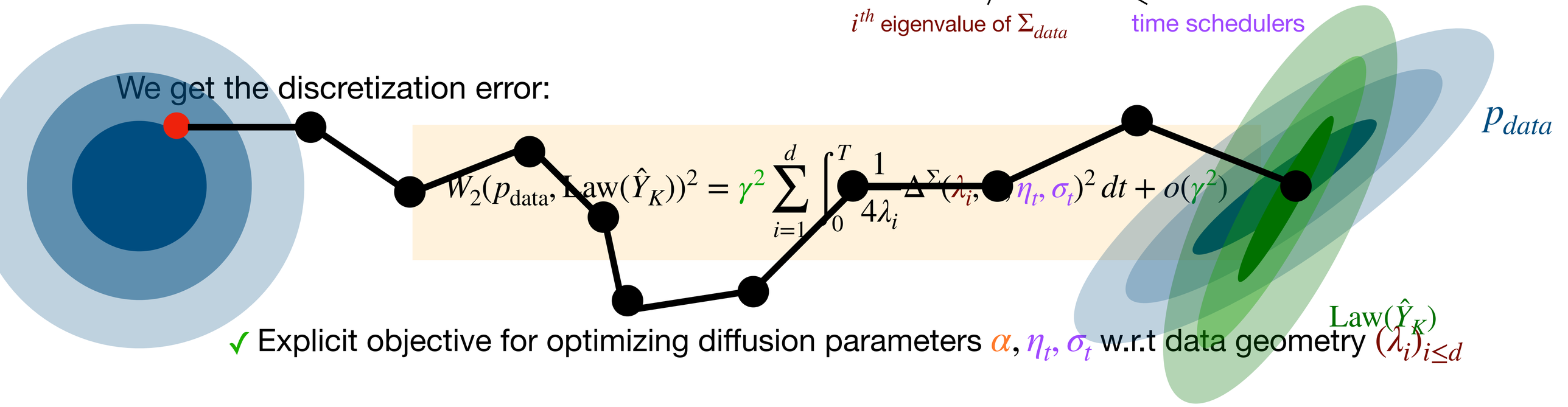
diffusion coefficient

$$\mathbb{E}\left[\hat{Y}_K\right] = \mu_{\text{data}} \qquad \text{Cov}\left[\hat{Y}_K\right] - \Sigma_{\text{data}} = \gamma\, U\,\text{Diag}\Big(\int_0^T \Delta^\Sigma(\lambda_i, \alpha, \eta_t, \sigma_t)\, dt\Big)_i U^\top + o(\gamma)$$

i^{th} eigenvalue of Σ_{data} — time schedulers

We get the discretization error:

$$W_2(p_{\text{data}}, \text{Law}(\hat{Y}_K))^2 = \gamma^2 \sum_{i=1}^{d} \int_0^T \frac{1}{4\lambda_i} \Delta^\Sigma(\lambda_i, \alpha, \eta_t, \sigma_t)^2\, dt + o(\gamma^2)$$

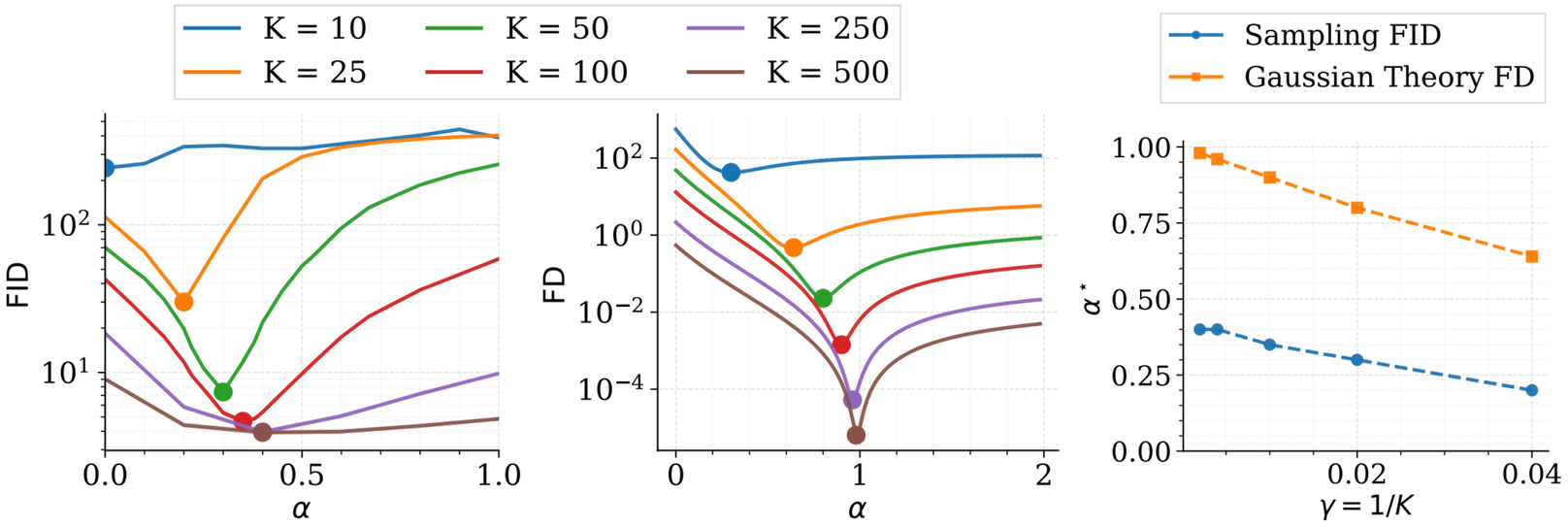


✓ Explicit objective for optimizing diffusion parameters α, η_t, σ_t w.r.t data geometry

Optimization of the diffusion parameter α

Image dataset sampling (FFHQ) : EDM schedule [Karras et al., `23] : *Variance Exploding (VE)* $\eta_t = 1$, noise $\sigma_t = \sigma_{\max} \left(\frac{t}{T}\right)^{\beta}$.

Sampling error w.r.t number of discretization steps K (stepsize $\gamma = T/K$) :



Empirical sampling (FID) **Gaussian theory (FD)** $\alpha^*(\gamma)$

$\alpha^\star$ **decreases** (linearly) with the stepsize γ

For *Variance Exploding (VE)* $\eta_t = 1$, and noise $\sigma_t = \sigma_{\max} \left(\frac{t}{T}\right)^{\beta}$ with $\sigma_{\max} \gg 1$

$$\alpha^\star(\gamma) = 1 - \gamma \sum_{i=1}^{d} w_i \lambda_i^{\frac{-1}{2\beta}} + o(\gamma)$$

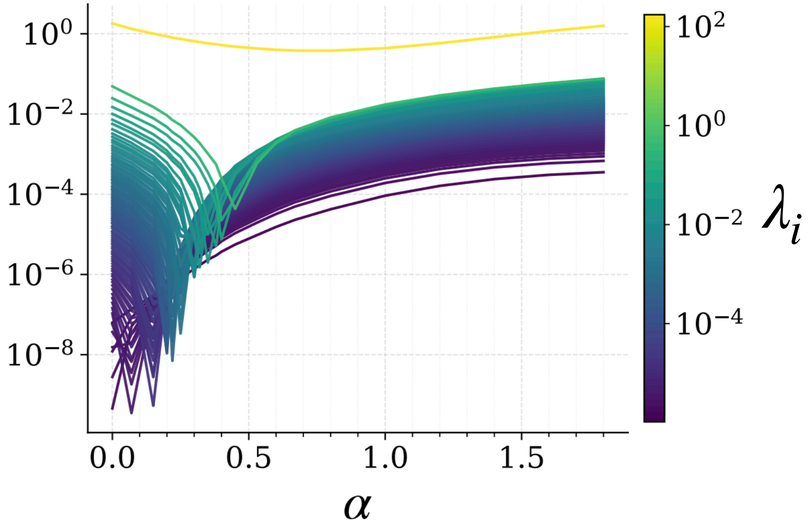
The decrease slope depends on $(\lambda_i)_{i \le d}$: directions with smaller variance λ_i favor smaller values of $\alpha^\star$

Optimization of the diffusion parameter α

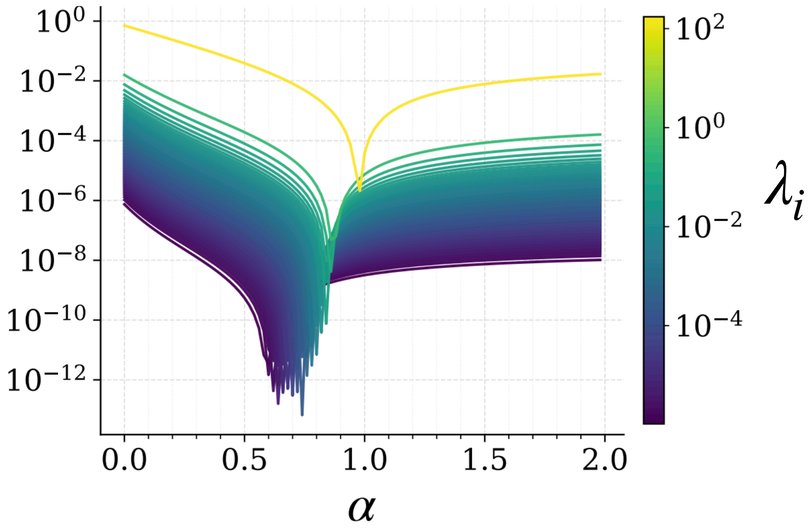
Image dataset sampling (FFHQ) : EDM schedule [Karras et al., `23] : *Variance Exploding (VE)* $\eta_t = 1$, noise $\sigma_t = \sigma_{\max} \left(\frac{t}{T}\right)^{\beta}$.

Error decomposed per eigendirection of Σ_{data}:

Empirical sampling



Gaussian theory



For *Variance Exploding (VE)* $\eta_t = 1$, and noise $\sigma_t = \sigma_{\max} \left(\frac{t}{T}\right)^{\beta}$ with $\sigma_{\max} \gg 1$

$$\alpha^\star(\gamma) = 1 - \gamma \sum_{i=1}^{d} w_i \lambda_i^{\frac{-1}{2\beta}} + o(\gamma)$$

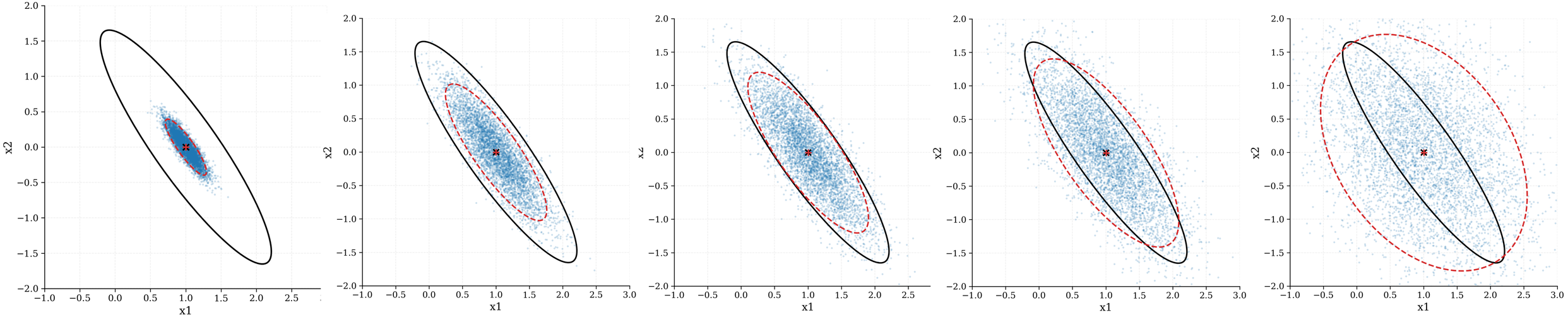
The decrease slope depends on $(\lambda_i)_{i \le d}$: directions with smaller variance λ_i favor smaller values of $\alpha^\star$

Optimization of the diffusion parameter α

Gaussian sampling in 10 steps

Sampled distribution covariance

Target distribution covariance



$\alpha = 0$ $\alpha = 0.25$ $\alpha = 0.5$ $\alpha = 0.75$ $\alpha = 1$

For *Variance Exploding (VE)* $\eta_t = 1$, and noise $\sigma_t = \sigma_{\max} \left(\frac{t}{T}\right)^{\beta}$ with $\sigma_{\max} \gg 1$

$$\alpha^\star(\gamma) = 1 - \gamma \sum_{i=1}^{d} w_i \lambda_i^{\frac{-1}{2\beta}} + o(\gamma)$$

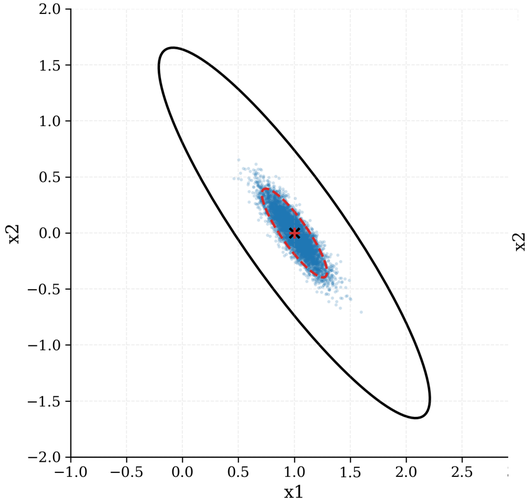
The decrease slope depends on $(\lambda_i)_{i \leq d}$: directions with smaller variance λ_i favor smaller values of $\alpha^\star$

Optimization of the diffusion parameter α

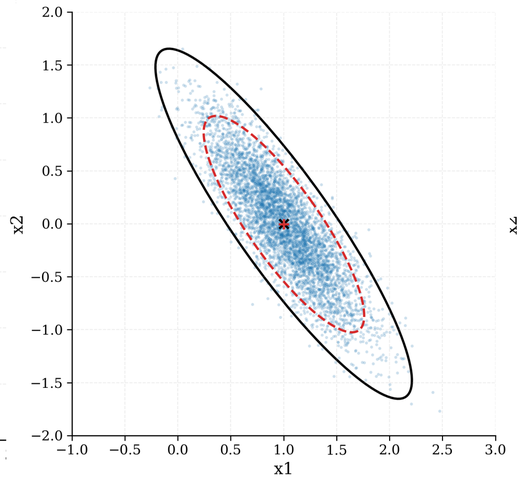
Gaussian sampling in 10 steps

Sampled distribution covariance

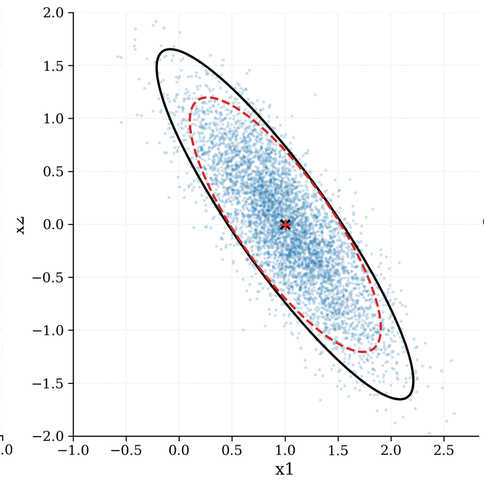
Target distribution covariance



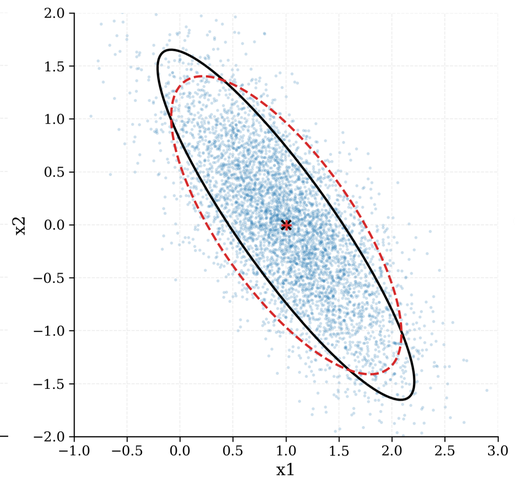
$\alpha = 0$



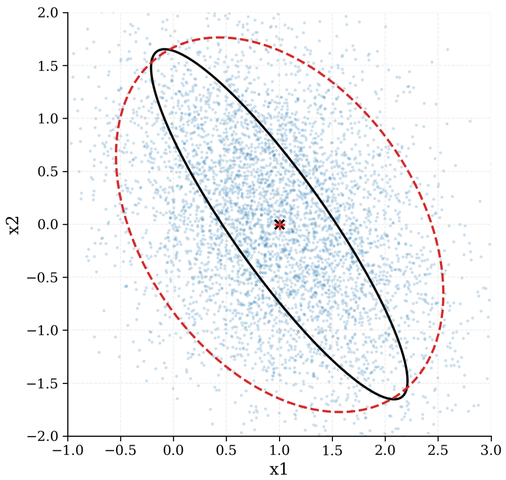
$\alpha = 0.25$



$\alpha = 0.5$



$\alpha = 0.75$



$\alpha = 1$

FFHQ sampling:

$\alpha = 0$

$\alpha = 0.25$

$\alpha = 0.5$

$\alpha = 0.75$

$\alpha = 1$

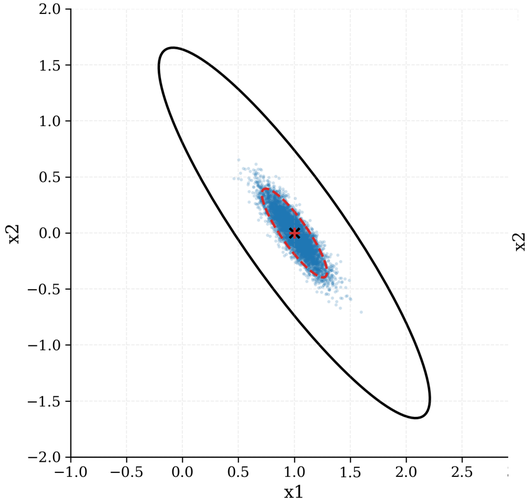
- For small number of steps, choosing $\alpha = 0$ samples close to the data mean

Optimization of the diffusion parameter α

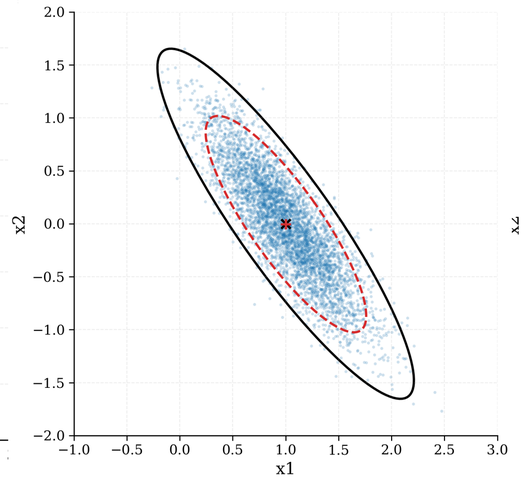
Gaussian sampling in 10 steps

Sampled distribution covariance

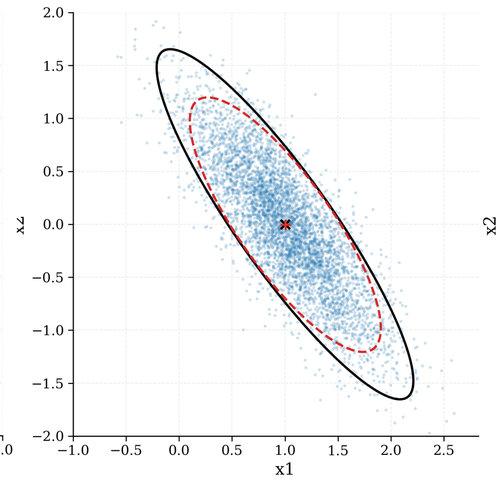
Target distribution covariance



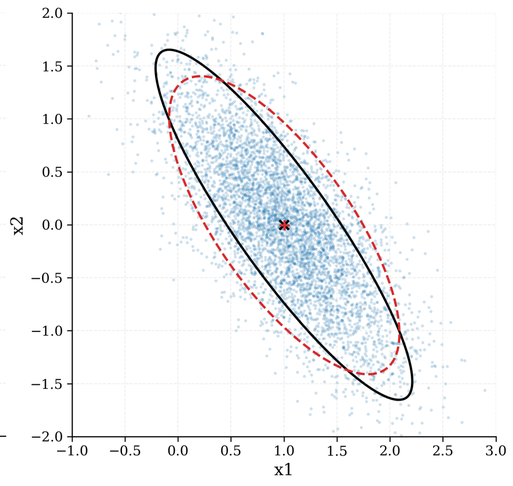
$\alpha = 0$



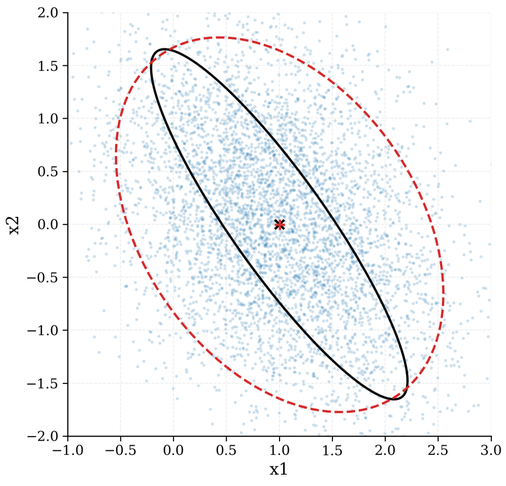
$\alpha = 0.25$



$\alpha = 0.5$



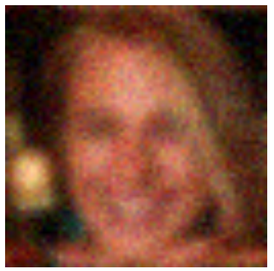
$\alpha = 0.75$



$\alpha = 1$

Posterior sampling: sample the posterior $p(x \mid y)$

x



y

$\alpha = 0$

$\alpha = 0.25$

$\alpha = 0.5$

Discretization error allows **MMSE estimation** using diffusion sampling

- For small number of steps, choosing $\alpha = 0$ samples close to the **MMSE**

Optimization of the rescaling schedule η_t

Classical choices of rescaling schedules:

$$\eta_t^{VE} = 1, \quad \eta_t^{VP} = \sqrt{\frac{1}{1+\sigma_t^2}}$$

We choose the parametrized $\eta_t(c) = \sqrt{\frac{c}{c+\sigma_t^2}}$

- The forward noising process converges to $\mathcal{N}(0, c\,\mathrm{I})$
- Variance Preserving (VP) : $c = 1$
- Variance Exploding (VE) : $c = \infty$

And look for $c^\star = \mathrm{argmin}_{c>0}\, W_2(p_{\mathrm{data}}, \mathrm{Law}(\hat{Y}_K))$

Assuming a power spectrum $\lambda_i = \lambda_{\max} i^{-p}$ and noise $\sigma_t = \sigma_{\max}\left(\frac{t}{T}\right)^\beta$ with large $\sigma_{\max}$ and β

$c^\star = \lambda_{\max}\tau(p)$ with $\tau(p) > 0$ decreasing for $p \lesssim 1$

Optimization of the rescaling schedule η_t

$$\eta_t(c) = \sqrt{\frac{c}{c + \sigma_t^2}}$$

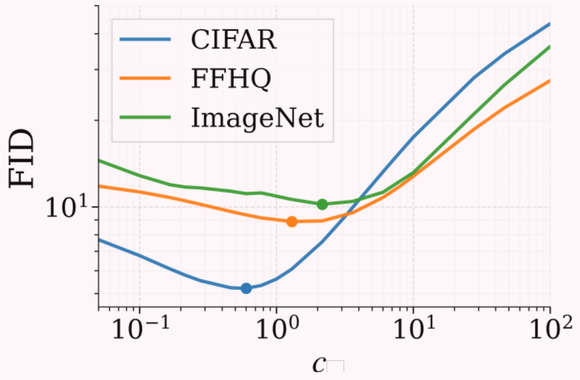
Assuming a power spectrum $\lambda_i = \lambda_{\max} i^{-p}$ and noise $\sigma_t = \sigma_{\max} \left(\frac{t}{T}\right)^{\beta}$ with large $\sigma_{\max}$ and β

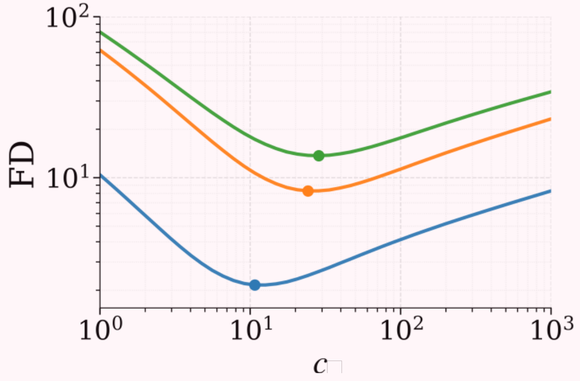
$$c^\star = \lambda_{\max} \tau(p) \quad \text{with } \tau(p) > 0 \text{ decreasing for } p \lesssim 1$$

Full image datasets sampling

Small relative change of p across datasets

Influence of $\lambda_{\max}$





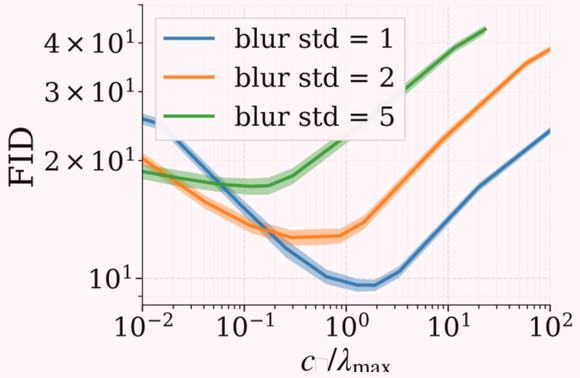
Dataset	$\lambda_{\max}$	p
CIFAR	55.3	1.43
FFHQ	173	1.47
ImageNet	225	1.35

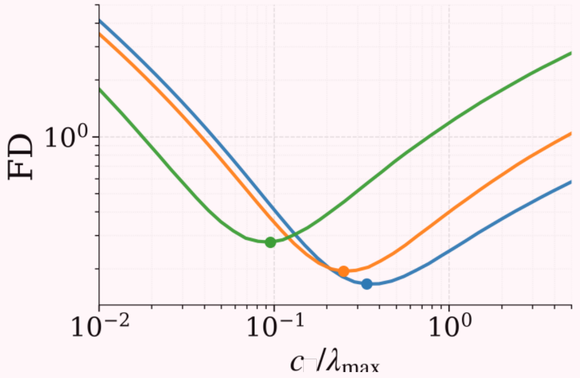
Deblurring Posterior sampling

Sampling of $p(x \mid y = k \star x + w)$

c-axis rescaled by $\lambda_{\max}$

Influence of p





blur std	$\lambda_{\max}$	p
1	0.39	0.57
2	0.86	0.61
5	0.78	0.83

Conclusion

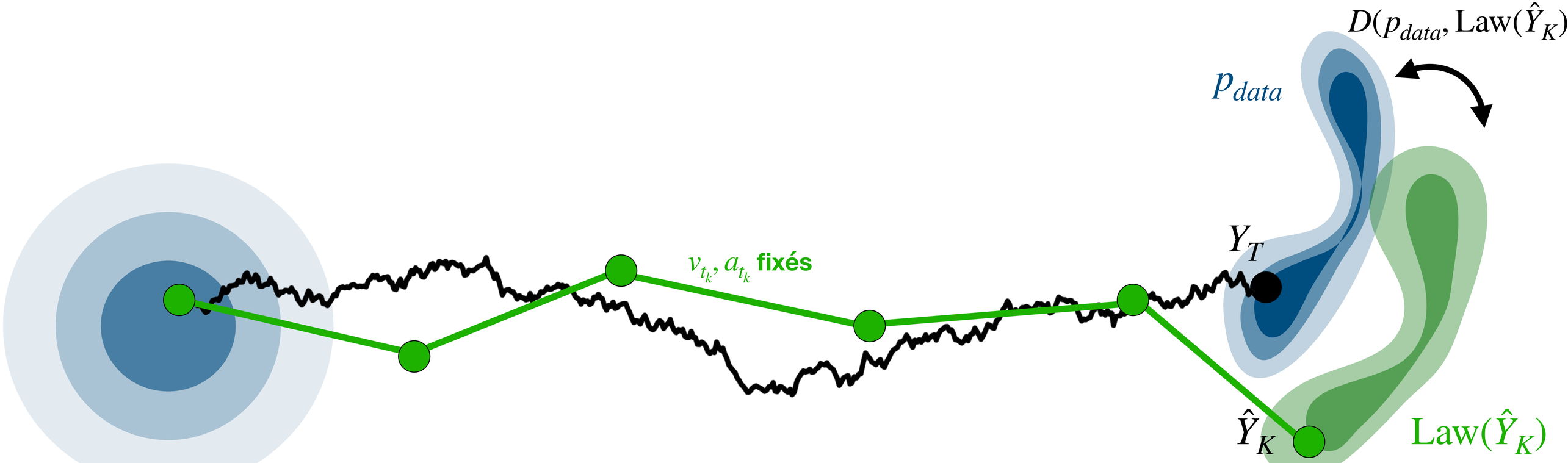
Data geometry matters for optimizing parameters w.r.t discretization error:

- Derived explicit first-order discretization error of diffusion sampling.
- In Gaussian case, gives a tractable objective for tuning diffusion parameters.
- Study how the optimal diffusion parameter α and schedulers η_t, σ_t change with the shape of the data covariance spectrum $(\lambda_i)_{i \leq d}$.
- Gaussian theory predicts trends observed for (non-Gaussian) real image sampling.

Discretization error

Euler discretization: for $\gamma = \frac{T}{K}$,

$$\hat{Y}_0 \sim p_0\,, \quad \hat{Y}_{k+1} = \hat{Y}_k + \gamma\, v_{t_k}(\hat{Y}_k) + \sqrt{2\gamma a_{t_k}}\, W_k$$



Discretization error: general weak error

Euler discretization: for $\gamma = \frac{T}{K}$,

$$\hat{Y}_0 \sim p_0\,, \quad \hat{Y}_{k+1} = \hat{Y}_k + \gamma\, v_{t_k}(\hat{Y}_k) + \sqrt{2\gamma a_{t_k}}\, W_k$$

For a test function $f \in \mathscr{C}^\infty(\mathbb{R}^d, \mathbb{R})$, under mild assumptions on p_{data}

End of discretized process — End of exact (continuous) process

$$\mathbb{E}\left[f(\hat{Y}_K)\right] - \mathbb{E}\left[f(Y_T)\right] = \gamma \int_0^T \mathbb{E}\left[\psi_s(Y_s)\right] ds + o(\gamma)$$

Euler error due to freezing v_s, a_s on $[s, s+\gamma]$

For $a_t = 0$

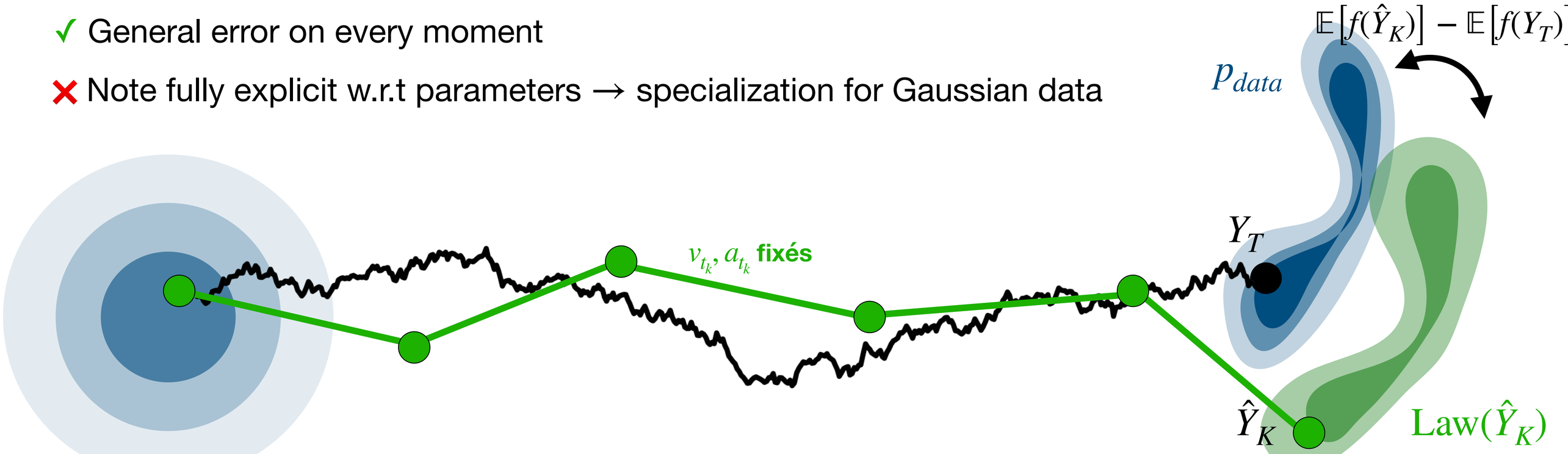
$$\psi_s(Y_s) = -\frac{1}{2}\, J_s(Y)\, D_s v_s(Y_s) \cdot \nabla f(Y_T)$$

Propagation $s \to T$ (flow Jacobian)

local acceleration (material derivative)

✓ General error on every moment

✗ Note fully explicit w.r.t parameters → specialization for Gaussian data



Diffusion models for posterior sampling $p(x \mid y = Ax + w)$

Apply the same algorithm with p_{data} replaced by $p(x \mid y)$

For $\alpha \geq 0$,

$$Y_0 \sim p_0\,, \quad dY_t = v_t(Y_t)\,dt + \sqrt{2a_t}\,dW_t$$

$$v_t(x) = -\frac{\dot{\eta}_{T-t}}{\eta_{T-t}}\,x + (1+\alpha)\,\eta_{T-t}^2\,\dot{\sigma}_{T-t}\sigma_{T-t}\,\nabla \log p_{T-t}(x \mid y)$$

Using Baye's formula, for $x_t \sim p_t$

$$\nabla \log p_t(x_t \mid y) = \nabla \log p_t(x_t) + \nabla \log p_t(y \mid x_t)$$

✓ Unconditional score training

$$= \int p(y \mid x_0) p(x_0 \mid x_t) dx_0$$

✗ Untractable

DPS [Chung et. al, `23], PiGDM [Song et. al, `23], Moment Matching [Rozet et. al, `24] :

Use the Gaussian approximation $p(x_0 \mid x_t) \approx \mathcal{N}(\mathbb{E}[x_0 \mid x_t], \mathbb{V}[x_0 \mid x_t])$

Related to a pertrained denoiser by Tweedie's formulas

Diffusion models

In the **continuous** setting, for fixed σ_t : VE $\eta_t = 1 \Leftrightarrow$ VP $\eta_t = \frac{1}{\sqrt{1+\sigma_t^2}} \Leftrightarrow$ any rescaling η_t

VE schedule $(\sigma_t, \eta_t = 1)$

$$y_0 \sim p_T, \quad dy_t = v_t(y_t)\,dt + \sqrt{2\alpha\xi_t}\,dW_t$$

$$\begin{cases} \beta_t = 0 \\ \xi_t = \dot{\sigma}_{T-t}\sigma_{T-t} \end{cases}$$

$$v_t = \xi_t \nabla \log p_{T-t}$$

$$\hat{y}_t = \eta_t y_t$$

VP schedule (σ_t, η_t)

$$y_0 \sim \hat{p}_T, \quad d\hat{y}_t = v_t(\hat{y}_t)\,dt + \sqrt{2\alpha\xi_t}\,dW_t$$

$$\begin{cases} \beta_t = -\frac{\dot{\eta}_{T-t}}{\eta_{T-t}} \\ \xi_t = \eta_{T-t}^2\,\dot{\sigma}_{T-t}\sigma_{T-t} \end{cases}$$

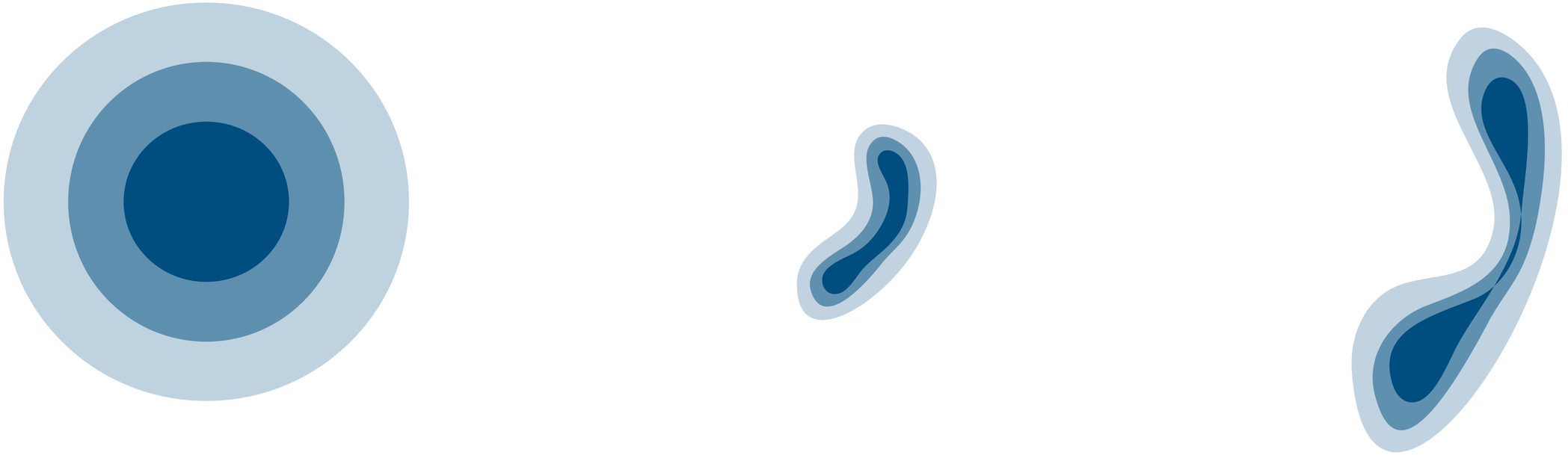
$$v_t = \beta_t\,\mathrm{Id} + \xi_t \nabla \log \hat{p}_{T-t}\left(\frac{\cdot}{\eta_t}\right)$$

Diffusion models **Other Noising Process** (from data to noise)

Flow Matching (FM)

$$X_0 \sim p_{\text{data}}, W_t \sim \mathcal{N}(0, \text{Id}), \quad X_t \overset{d}{=} (1-t)X_0 + tW_t$$

$$\sigma_t = \frac{t}{1-t}, \quad \eta_t = \frac{1}{1+\sigma_t}$$



$q_1 = \mathcal{N}(0, \text{Id})$ $q_t = \text{Law}(\eta_t(X_0 + \sigma_t X_t))$ $q_0 = p_{data}$

$t = T = 1$ t $t = 0$

$$\text{Cov}(X_t) = (1-t)^2 \text{Cov}(X_0) + t^2 \,\text{Id}$$

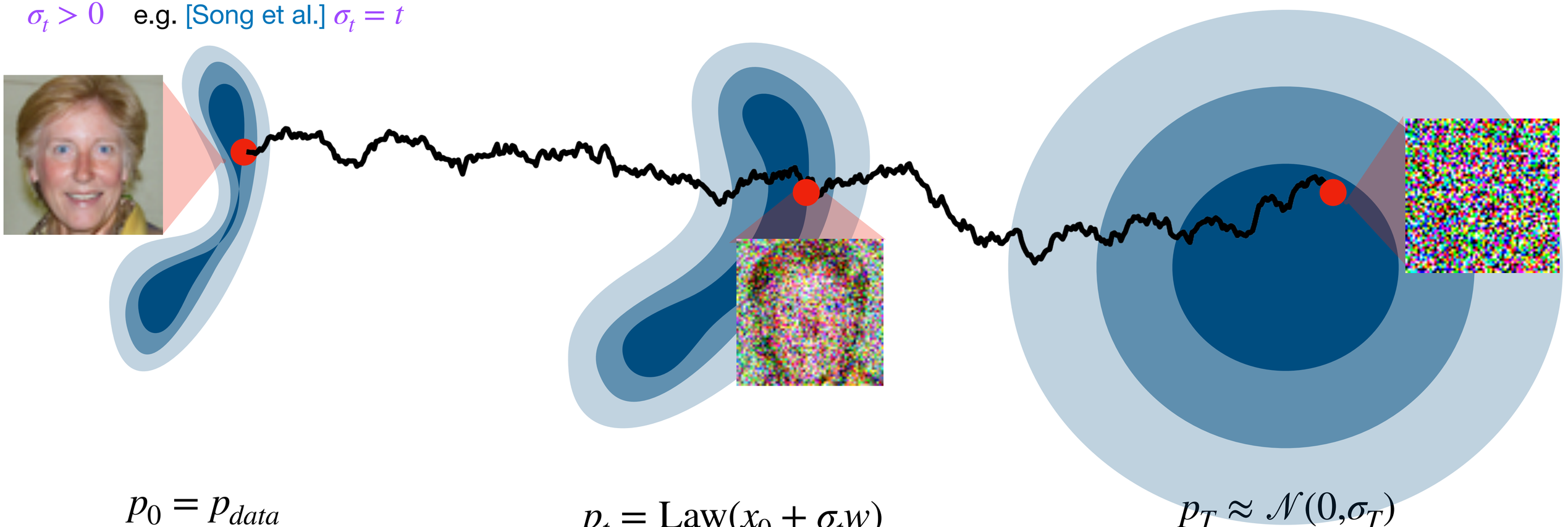
Diffusion models

Forward flow : from data to noise

Variance Exploding (VE)

$x_0 \sim p_{\text{data}}, \quad dx_t = \sqrt{2\dot{\sigma}_t \sigma_t}\, dw_t \quad \Leftrightarrow \quad w_t \sim \mathcal{N}(0, \text{Id}), \quad x_t = x_0 + \sigma_t w_t$

$\sigma_t > 0$ e.g. [Song et al.] $\sigma_t = t$



$p_0 = p_{data}$ $\qquad$ $p_t = \text{Law}(x_0 + \sigma_t w)$ $\qquad$ $p_T \approx \mathcal{N}(0, \sigma_T)$

$t = 0$ $\qquad$ t $\qquad$ $t = T \gg 1$

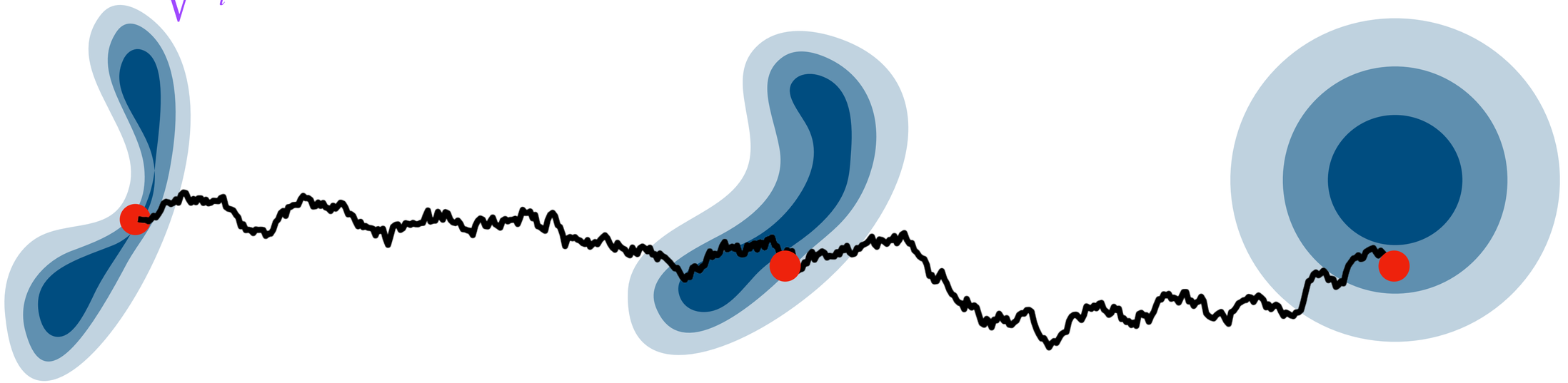
$\text{Cov}(x_0)$ $\qquad$ $\text{Cov}(x_t) = \text{Cov}(x_0) + \sigma_T^2\, \text{Id}$ $\qquad$ $\text{Cov}(x_T) \approx \sigma_T^2\, \text{Id}$

Diffusion models **Forward flow :** from data to noise

Variance Preserving (VP) $x_0 \sim p_{\text{data}}, \quad dx_t = \frac{\dot{\eta}_t}{\eta_t} x_t dt + \eta_t \sqrt{2\dot{\sigma}_t \sigma_t}\, dw_t \quad \Leftrightarrow$

$$w_t \sim \mathcal{N}(0, \mathrm{Id}), \quad x_t = \eta_t(x_0 + \sigma_t w_t)$$

$$\sigma_t > 0, \quad \eta_t = \frac{1}{\sqrt{\sigma_t^2 + 1}}$$



$p_0 = p_{data}$ $\qquad$ $p_t = \mathrm{Law}\left(\eta_t(x_0 + \sigma_t w)\right)$ $\qquad$ $p_T \approx \mathcal{N}(0, \sigma_T)$

$t = 0$ $\qquad$ t $\qquad$ $t = T \gg 1$

$\mathrm{Cov}(x_0)$ $\qquad$ $\mathrm{Cov}(x_t) = \eta_t^2\, \mathrm{Cov}(x_0) + \left(1 - \eta_t^2\right) \mathrm{Id}$ $\qquad$ $\mathrm{Cov}(x_T) \approx \mathrm{Id}$

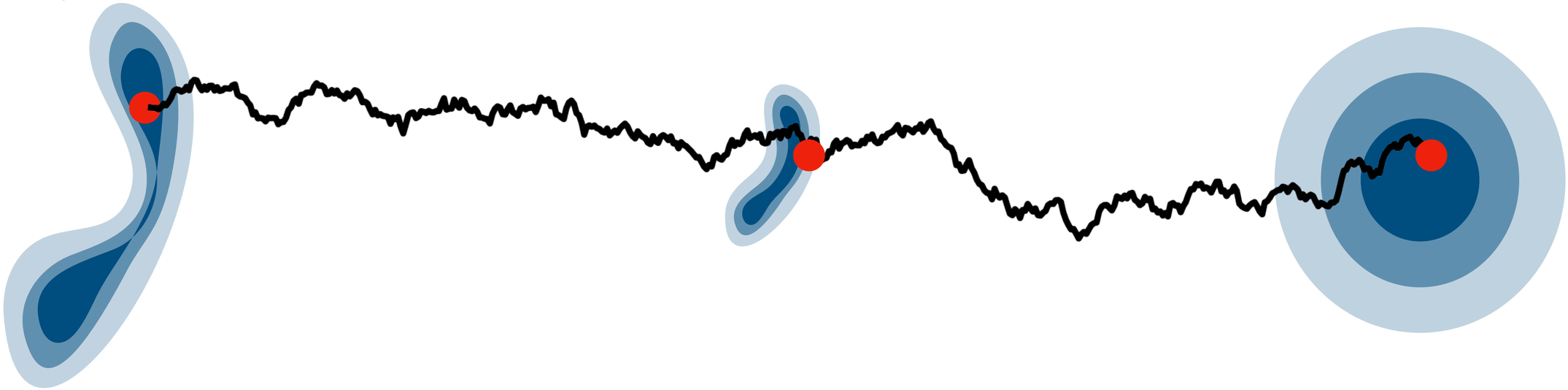
Diffusion models

Forward flow : from data to noise

Flow Matching (FM)

$$w_t \sim \mathcal{N}(0, \mathrm{Id}), \quad x_t = t\, x_0 + (1-t)\, w_t$$

$$\sigma_t = \frac{t}{1-t}, \quad \eta_t = 1 - t \quad (T = 1)$$



$p_0 = p_{data}$ | $p_t = \mathrm{Law}\left(\eta_t(x_0 + \sigma_t w)\right)$ | $p_1 = \mathcal{N}(0, \mathrm{Id})$

$t = 0$ | t | $t = T = 1$

$\mathrm{Cov}(x_0)$ | $\mathrm{Cov}(x_t) = (1-t)^2 \mathrm{Cov}(x_0) + t^2\, \mathrm{Id}$ | $\mathrm{Cov}(x_T) = \mathrm{Id}$

Discretization error: : general weak error

Euler discretization: for $\gamma = \dfrac{T}{K}$,

$$\hat{Y}_0 \sim p_0\,, \quad \hat{Y}_{k+1} = \hat{Y}_k + \gamma\, v_{t_k}(\hat{Y}_k) + a_{t_k} W_k$$

Under mild assumptions on p_{data}

End discretized process — End continuous process

$$\mathbb{E}\left[f(\hat{Y}_K)\right] - \mathbb{E}\left[f(Y_T)\right] = \gamma \int_0^T \mathbb{E}\left[\psi_s(Y_s)\right] ds + o(\gamma)$$

Euler error due to freezing v_s, a_s on $[s, s+\gamma]$

For $a_t = 0$

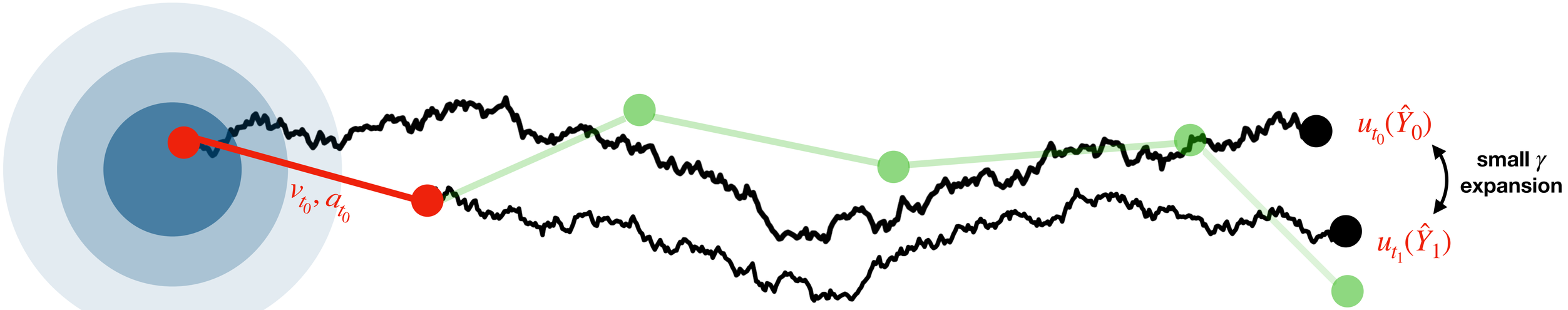
$$\psi_s(Y_s) = -\frac{1}{2}\, J_s(Y)\; D_s v_s(Y_s) \cdot \nabla f(Y_T)$$

Propagation $s \to T$ (flow Jacobian) — local acceleration (material derivative)

Sketch of the proof: Let $u_t(x) = \mathbb{E}[f(Y_T) \mid Y_t = x]$. We have:

$$\mathbb{E}\left[f(\hat{Y}_K)\right] - \mathbb{E}\left[f(Y_T)\right] = \mathbb{E}[u_{t_k}(\hat{Y}_k)] - \mathbb{E}[u_0(\hat{Y}_0)] = \sum_{k=0}^{K} u_{t_{k+1}}(\hat{Y}_{k+1}) - u_{t_k}(\hat{Y}_k) = \sum_{k=0}^{K} \left(\gamma^2 \psi_{t_k}(\hat{Y}_k) + o(\gamma^2)\right) = \gamma \int_0^T \psi_s(Y_s)\, ds + o(\gamma)$$

Telescopic sum — *Expansion for small γ* — *Riemann-sum error*



Discretization error: : general weak error

Euler discretization: for $\gamma = \frac{T}{K}$,

$$\hat{Y}_0 \sim p_0\,, \quad \hat{Y}_{k+1} = \hat{Y}_k + \gamma\, v_{t_k}(\hat{Y}_k) + a_{t_k} W_k$$

Under mild assumptions on p_{data}

End discretized process — End continuous process

$$\mathbb{E}\left[f(\hat{Y}_K)\right] - \mathbb{E}\left[f(Y_T)\right] = \gamma \int_0^T \mathbb{E}\left[\psi_s(Y_s)\right] ds + o(\gamma)$$

Euler error due to freezing v_s, a_s on $[s, s+\gamma]$

For $a_t = 0$

$$\psi_s(Y_s) = -\frac{1}{2}\, J_s(Y)\; D_s v_s(Y_s) \cdot \nabla f(Y_T)$$

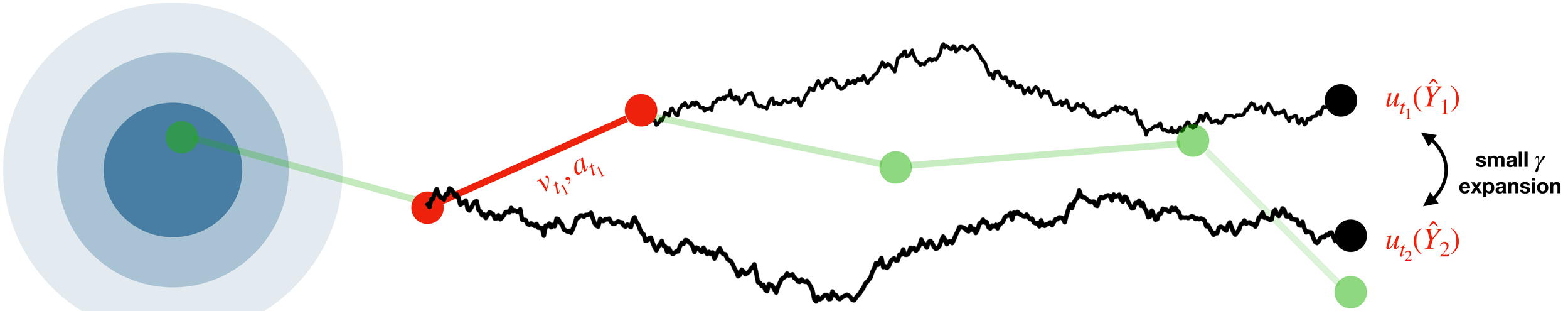
Propagation $s \to T$ (flow Jacobian)

local acceleration (material derivative)

Sketch of the proof: Let $u_t(x) = \mathbb{E}[f(Y_T) \mid Y_t = x]$. We have:

$$\mathbb{E}\left[f(\hat{Y}_K)\right] - \mathbb{E}\left[f(Y_T)\right] = \mathbb{E}[u_{t_k}(\hat{Y}_k)] - \mathbb{E}[u_0(\hat{Y}_0)] = \sum_{k=0}^{K} u_{t_{k+1}}(\hat{Y}_{k+1}) - u_{t_k}(\hat{Y}_k) = \sum_{k=0}^{K} \left(\gamma^2 \psi_{t_k}(\hat{Y}_k) + o(\gamma^2)\right) = \gamma \int_0^T \psi_s(Y_s)\, ds + o(\gamma)$$

Telescopic sum — *Expansion for small γ* — *Riemann-sum error*



Discretization error: : general weak error

Euler discretization: for $\gamma = \frac{T}{K}$,

$$\hat{Y}_0 \sim p_0\,, \quad \hat{Y}_{k+1} = \hat{Y}_k + \gamma\, v_{t_k}(\hat{Y}_k) + a_{t_k} W_k$$

Under mild assumptions on p_{data}

End discretized process — End continuous process

$$\mathbb{E}\left[f(\hat{Y}_K)\right] - \mathbb{E}\left[f(Y_T)\right] = \gamma \int_0^T \mathbb{E}\left[\psi_s(Y_s)\right] ds + o(\gamma)$$

Euler error due to freezing v_s, a_s on $[s, s+\gamma]$

For $a_t = 0$

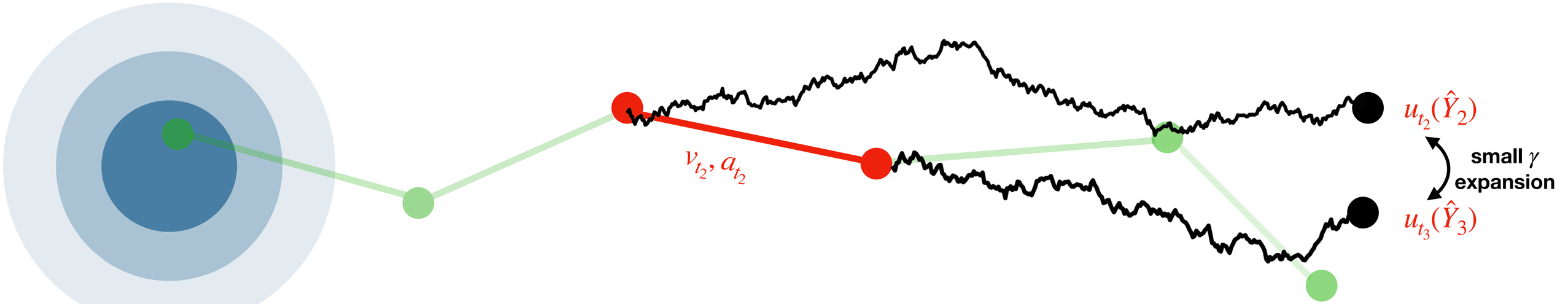
$$\psi_s(Y_s) = -\frac{1}{2}\, J_s(Y)\; D_s v_s(Y_s) \cdot \nabla f(Y_T)$$

Propagation $s \to T$ (flow Jacobian) — local acceleration (material derivative)

Sketch of the proof: Let $u_t(x) = \mathbb{E}[f(Y_T) \mid Y_t = x]$. We have:

$$\mathbb{E}\left[f(\hat{Y}_K)\right] - \mathbb{E}\left[f(Y_T)\right] = \mathbb{E}[u_{t_k}(\hat{Y}_k)] - \mathbb{E}[u_0(\hat{Y}_0)] = \sum_{k=0}^{K} u_{t_{k+1}}(\hat{Y}_{k+1}) - u_{t_k}(\hat{Y}_k) = \sum_{k=0}^{K} \left(\gamma^2 \psi_{t_k}(\hat{Y}_k) + o(\gamma^2)\right) = \gamma \int_0^T \psi_s(Y_s)\, ds + o(\gamma)$$

Telescopic sum — *Expansion for small γ* — *Riemann-sum error*



Discretization error: : general weak error

Euler discretization: for $\gamma = \frac{T}{K}$,

$$\hat{Y}_0 \sim p_0\,, \quad \hat{Y}_{k+1} = \hat{Y}_k + \gamma\, v_{t_k}(\hat{Y}_k) + a_{t_k} W_k$$

Under mild assumptions on p_{data}

End discretized process — End continuous process

$$\mathbb{E}\left[f(\hat{Y}_K)\right] - \mathbb{E}\left[f(Y_T)\right] = \gamma \int_0^T \mathbb{E}\left[\psi_s(Y_s)\right] ds + o(\gamma)$$

Euler error due to freezing v_s, a_s on $[s, s+\gamma]$

For $a_t = 0$

$$\psi_s(Y_s) = -\frac{1}{2}\, J_s(Y)\; D_s v_s(Y_s) \cdot \nabla f(Y_T)$$

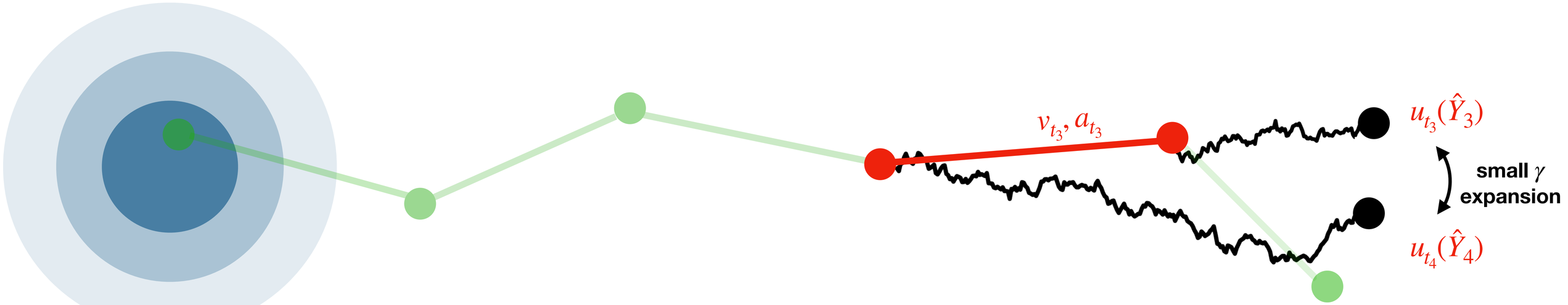
Propagation $s \to T$ (flow Jacobian)

local acceleration (material derivative)

Sketch of the proof: Let $u_t(x) = \mathbb{E}[f(Y_T) \mid Y_t = x]$. We have:

$$\mathbb{E}\left[f(\hat{Y}_K)\right] - \mathbb{E}\left[f(Y_T)\right] = \mathbb{E}[u_{t_k}(\hat{Y}_k)] - \mathbb{E}[u_0(\hat{Y}_0)] = \sum_{k=0}^{K} u_{t_{k+1}}(\hat{Y}_{k+1}) - u_{t_k}(\hat{Y}_k) = \sum_{k=0}^{K} \left(\gamma^2 \psi_{t_k}(\hat{Y}_k) + o(\gamma^2)\right) = \gamma \int_0^T \psi_s(Y_s)\, ds + o(\gamma)$$

Telescopic sum — *Expansion for small γ* — *Riemann-sum error*



Discretization error: : general weak error

Euler discretization: for $\gamma = \frac{T}{K}$,

$$\hat{Y}_0 \sim p_0\,, \quad \hat{Y}_{k+1} = \hat{Y}_k + \gamma\, v_{t_k}(\hat{Y}_k) + a_{t_k} W_k$$

Under mild assumptions on p_{data}

End discretized process — End continuous process

$$\mathbb{E}\left[f(\hat{Y}_K)\right] - \mathbb{E}\left[f(Y_T)\right] = \gamma \int_0^T \mathbb{E}\left[\psi_s(Y_s)\right] ds + o(\gamma)$$

Euler error due to freezing v_s, a_s on $[s, s+\gamma]$

For $a_t = 0$

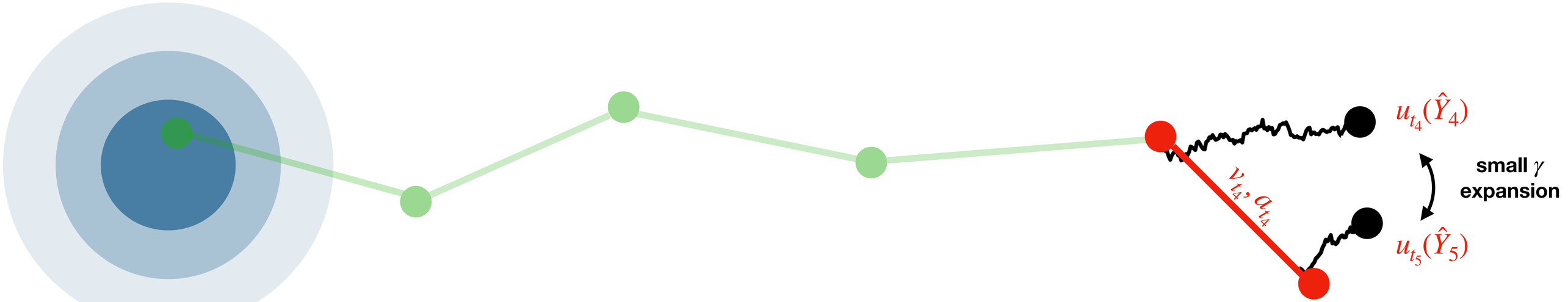
$$\psi_s(Y_s) = -\frac{1}{2}\, J_s(Y)\; D_s v_s(Y_s) \cdot \nabla f(Y_T)$$

Propagation $s \to T$ (flow Jacobian) — local acceleration (material derivative)

Sketch of the proof: Let $u_t(x) = \mathbb{E}[f(Y_T) \mid Y_t = x]$. We have:

$$\mathbb{E}\left[f(\hat{Y}_K)\right] - \mathbb{E}\left[f(Y_T)\right] = \mathbb{E}[u_{t_k}(\hat{Y}_k)] - \mathbb{E}[u_0(\hat{Y}_0)] = \sum_{k=0}^{K} u_{t_{k+1}}(\hat{Y}_{k+1}) - u_{t_k}(\hat{Y}_k) = \sum_{k=0}^{K} \left(\gamma^2 \psi_{t_k}(\hat{Y}_k) + o(\gamma^2)\right) = \gamma \int_0^T \psi_s(Y_s)\, ds + o(\gamma)$$

Telescopic sum — *Expansion for small γ* — *Riemann-sum error*



Optimization of the rescaling schedule η_t

Classical choices of rescaling schedules:

$$\eta_t^{VE} = 1, \quad \eta_t^{VP} = \sqrt{\frac{1}{1+\sigma_t^2}}, \quad \eta_t^{FM} = \frac{1}{1+\sigma_t^2}$$

Gaussian sampling covariance error:

$$\mathrm{Cov}\left[\hat{Y}_K\right] = \Sigma_{\text{data}} + \gamma\, U\, \mathrm{Diag}\left(\int_0^T \Delta^{\Sigma}(\lambda_i, \eta_t, \sigma_t)\, dt\right) U^{\top} + o(\gamma)$$

$$\Delta^{\Sigma}(\lambda_i, \eta_t^{\star}, \sigma_t) = 0 \Leftrightarrow \eta_t^{\star}(\lambda_i) = \sqrt{\frac{\lambda_i}{\lambda_i + \sigma_t^2}} \quad \Rightarrow \mathrm{Var}_{\lambda_i}[X_t] = \mathrm{Var}_{\lambda_i}[X_0]$$

Variance preserving condition on eigendirection λ_i

This is optimal **for a single** eigendirection λ_i .

For multidimensional and anisotropic data, we set $\eta_t(c) = \sqrt{\frac{c}{c+\sigma_t^2}}$

Comes back to VP initialized at $\mathcal{N}(0, c\,\mathrm{I})$ instead of $\mathcal{N}(0, \mathrm{I})$

$$c^{\star} = \mathrm{argmin}_{c>0}\, \mathrm{FD}(p_{\text{data}}, \mathrm{Law}(\hat{Y}_K)) = \mathrm{argmin}_{c>0} \sum_{i=1}^{d} \int_0^T f(\lambda_i, \eta_t(c), \sigma_t)\, dt$$

Assuming a power spectrum $\lambda_i = \lambda_{\max} i^{-p}$ and noise $\sigma_t = \sigma_{\max} \left(\frac{t}{T}\right)^{\beta}$ with large $\sigma_{\max}$ and β

$$c^{\star} = \lambda_{\max} \tau(p) \quad \text{with } \tau(p) > 0 \text{ and decreasing for } p \lesssim 1$$