Recent progress in denoising diffusion models

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

Journées "Probabilités en informatique"
September 2024

Problem set-up
Sampling with iterative algorithms

- **Sampling from probability distribution** $p(x) = \frac{1}{Z} \exp(-f(x))$
  - high-dimensional and "complex"
  - f given (without Z) or f estimated from i.i.d. data

Problem set-up
Sampling with iterative algorithms

- **Sampling from probability distribution** $p(x) = \frac{1}{Z}\exp(-f(x))$
  - high-dimensional and “complex”
  - f given (without Z) or f estimated from i.i.d. data

- **Applications**
  - Image generation $p(x)$
  - Conditional image generation $p(x|y) \propto p(y|x)p(x)$
  - Protein discovery (Frey et al., 2024), etc.

Application to image generation
“Panda riding a bicycle in Paris”

https://stablediffusionweb.com/

Application to image generation
“Mathematicians in the black forest”

https://stablediffusionweb.com/

Problem set-up
Sampling with iterative algorithms

- **Sampling from probability distribution** $p(x) = \frac{1}{Z}\exp(-f(x))$
  - high-dimensional and "complex"
  - f given (without Z) or f estimated from i.i.d. data
- **Applications**
  - Image generation $p(x)$
  - Conditional image generation $p(x|y) \propto p(y|x)p(x)$
  - Protein discovery (Frey et al., 2024), etc.
- **Main difficulty**
  - Multimodal distributions
  - Curse of dimensionality

Problem set-up
Sampling with iterative algorithms

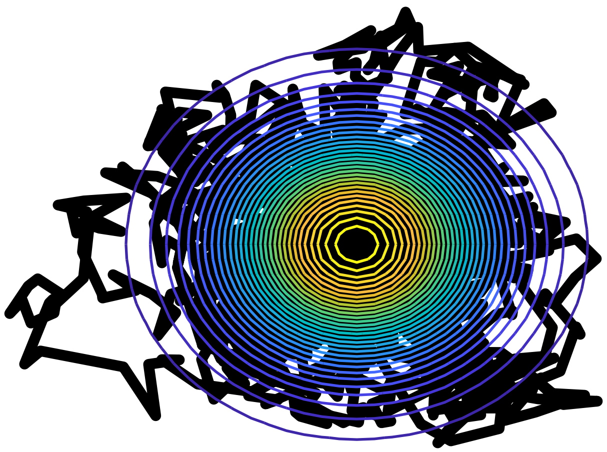
- **Sampling from probability distribution** $p(x) = \frac{1}{Z}\exp(-f(x))$
  - high-dimensional and “complex”
  - f given (without Z) or f estimated from i.i.d. data

Problem set-up
Sampling with iterative algorithms

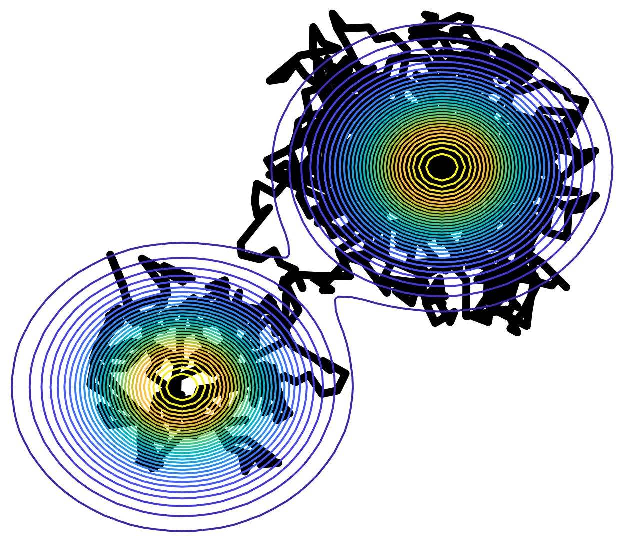
- **Sampling from probability distribution** $p(x) = \frac{1}{Z} \exp(-f(x))$
  - high-dimensional and "complex"
  - f given (without Z) or f estimated from i.i.d. data

- **Langevin algorithms**
  - Discretization of diffusion $dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$:
$$x_{k+1} = x_k - \gamma \nabla f(x_k) + \sqrt{2\gamma} \cdot \mathcal{N}(0, I)$$
  - (slow) convergence (see, e.g., Bakry et al., 2008)
  - fast for smooth log-concave distributions (e.g., f convex) (Dalalyan, 2017, Durmus and Moulines, 2017, Chewi, 2022, etc.)

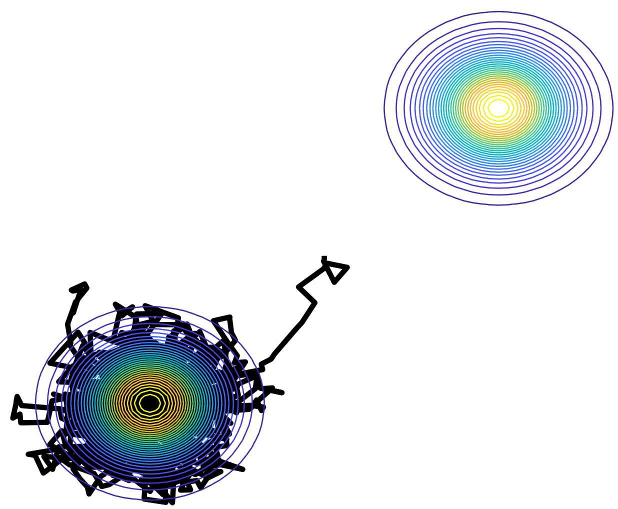
From log-concave to non-log-concave



From log-concave to non-log-concave



From log-concave to non-log-concave



Problem set-up
Sampling with iterative algorithms

- **Sampling from probability distribution** $p(x) \propto \exp(-f(x))$
  - high-dimensional and "complex"
  - f given or f estimated from i.i.d. data

- **Langevin algorithms**
  - Discretization of diffusion $dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$:

$$x_{k+1} = x_k - \gamma \nabla f(x_k) + \sqrt{2\gamma} \cdot \mathcal{N}(0, I)$$

  - (slow) convergence (see, e.g., Bakry et al., 2008)
  - fast for smooth log-concave distributions (e.g., f convex) (Dalalyan, 2017, Durmus and Moulines, 2017, Chewi, 2022, etc.)

- **Going beyond log-concave distributions**

Sampling with denoising diffusion models

Three main ideas

1. **Sampling by denoising (Saremi and Hyvärinen, 2019)**
   - Sampling the noisy data is easier
   - Denoising the data through "empirical Bayes"

Sampling with denoising diffusion models
Three main ideas

1. **Sampling by denoising (Saremi and Hyvärinen, 2019)**
   - Sampling the noisy data is easier
   - Denoising the data through "empirical Bayes"

2. **Learning the denoiser from data (Hyvärinen, 2005, Vincent, 2011)**
   - Multivariate output supervised learning
   - Need samples from the distribution

Sampling with denoising diffusion models
Three main ideas

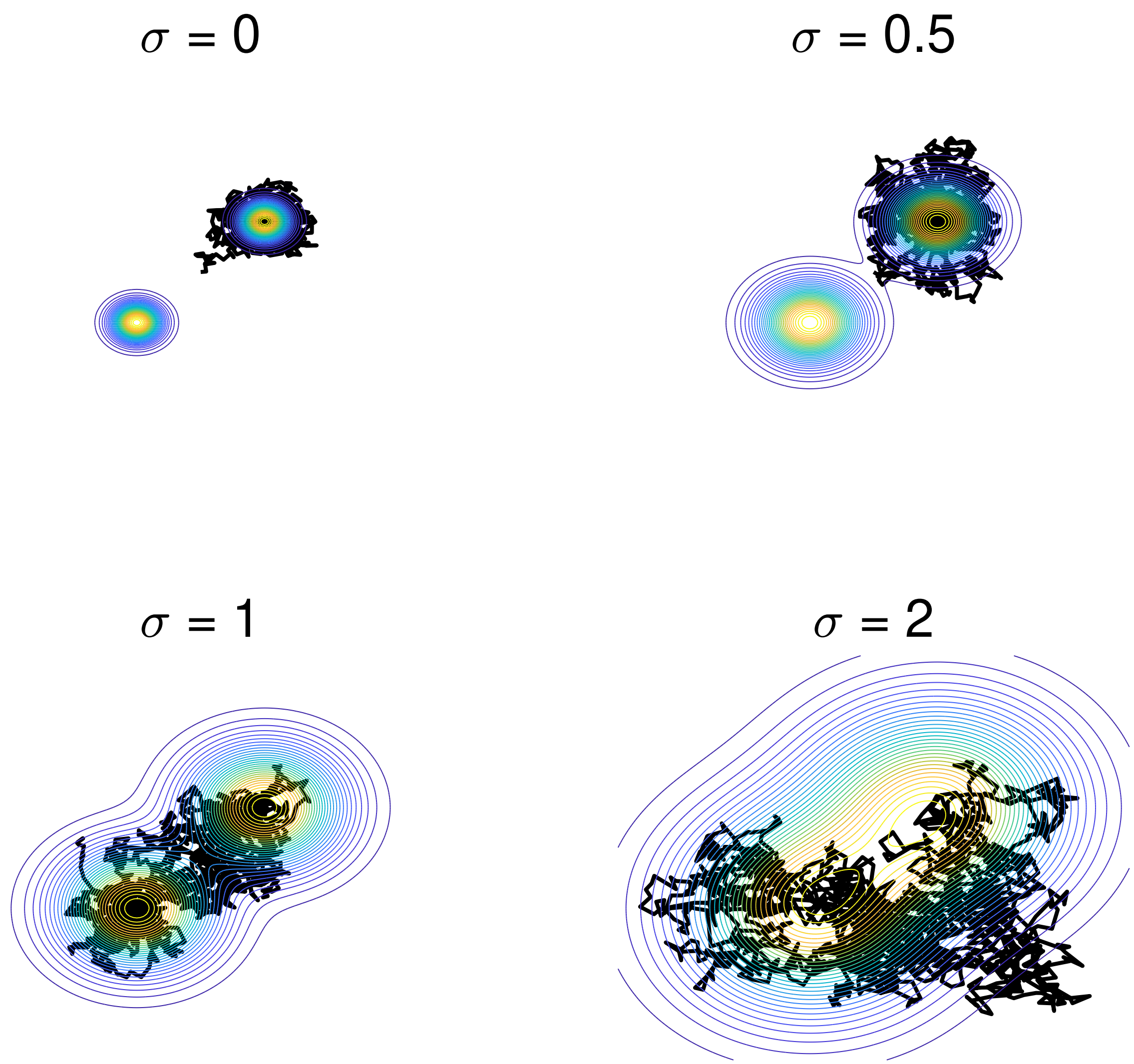
1. **Sampling by denoising (Saremi and Hyvärinen, 2019)**
   - Sampling the noisy data is easier
   - Denoising the data through "empirical Bayes"

2. **Learning the denoiser from data (Hyvärinen, 2005, Vincent, 2011)**
   - Multivariate output supervised learning
   - Need samples from the distribution

3. **Progressive denoising for improved sampling/denoising**
   - Using continuous-time diffusions (Song and Ermon, 2019)
   - Using multiple measurements (Park, Saremi, and B., 2024)

Sampling by denoising (Saremi and Hyvärinen, 2019)

- **Replace** X **by** $Y = X + \sigma \cdot \mathcal{N}(0, I)$**:** sampling is (provably) easier



Sampling by denoising (Saremi and Hyvärinen, 2019)

- **Replace** X **by** $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Sampling (provably) easier
  - Going back from Y to X by "optimal" denoising

Sampling by denoising (Saremi and Hyvärinen, 2019)

- **Replace** X **by** $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Sampling (provably) easier
  - Going back from Y to X by "optimal" denoising

- **Empirical Bayes** (Robbins, 1956, Miyasawa, 1961)
  - Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Key result: $\mathbb{E}\big[X|Y\big] = Y + \sigma^2 \nabla \log q_\sigma(Y)$
  - Proof by integration by parts

Proof by integration by parts

- Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$

$$
\begin{aligned}
\mathbb{E}\big[X|Y=y\big] &= \frac{1}{q_\sigma(y)} \int_{\mathbb{R}^d} x p(x,y) dx = \frac{1}{q_\sigma(y)} \int_{\mathbb{R}^d} x p(y|x) p(x) dx \\
&= \frac{1}{q_\sigma(y)} \int_{\mathbb{R}^d} x \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\|x-y\|^2} p(x) dx \\
&= y + \frac{\sigma^2}{q_\sigma(y)} \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{d/2}} \frac{x-y}{\sigma^2} e^{-\frac{1}{2\sigma^2}\|x-y\|^2} p(x) dx \\
&= y + \frac{\sigma^2}{q_\sigma(y)} \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{d/2}} \big(-\nabla e^{-\frac{1}{2\sigma^2}\|x-y\|^2}\big) p(x) dx \\
&= y + \frac{\sigma^2}{q_\sigma(y)} \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\|x-y\|^2} \nabla p(x) dx \\
&= y + \frac{\sigma^2}{q_\sigma(y)} \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\|y\|^2} \nabla p(x-y) dx \\
&= y + \frac{\sigma^2}{q_\sigma(y)} \nabla q_\sigma(y) = y + \sigma^2 \nabla \log q_\sigma(y)
\end{aligned}
$$

Sampling by denoising (Saremi and Hyvärinen, 2019)

- **Replace** X **by** $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Sampling (provably) easier
  - Going back from Y to X by “optimal” denoising

- **Empirical Bayes** (Robbins, 1956, Miyasawa, 1961)
  - Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Key result: $\mathbb{E}\big[X|Y\big] = Y + \sigma^2 \nabla \log q_\sigma(Y)$
  - Proof by integration by parts

Sampling by denoising (Saremi and Hyvärinen, 2019)

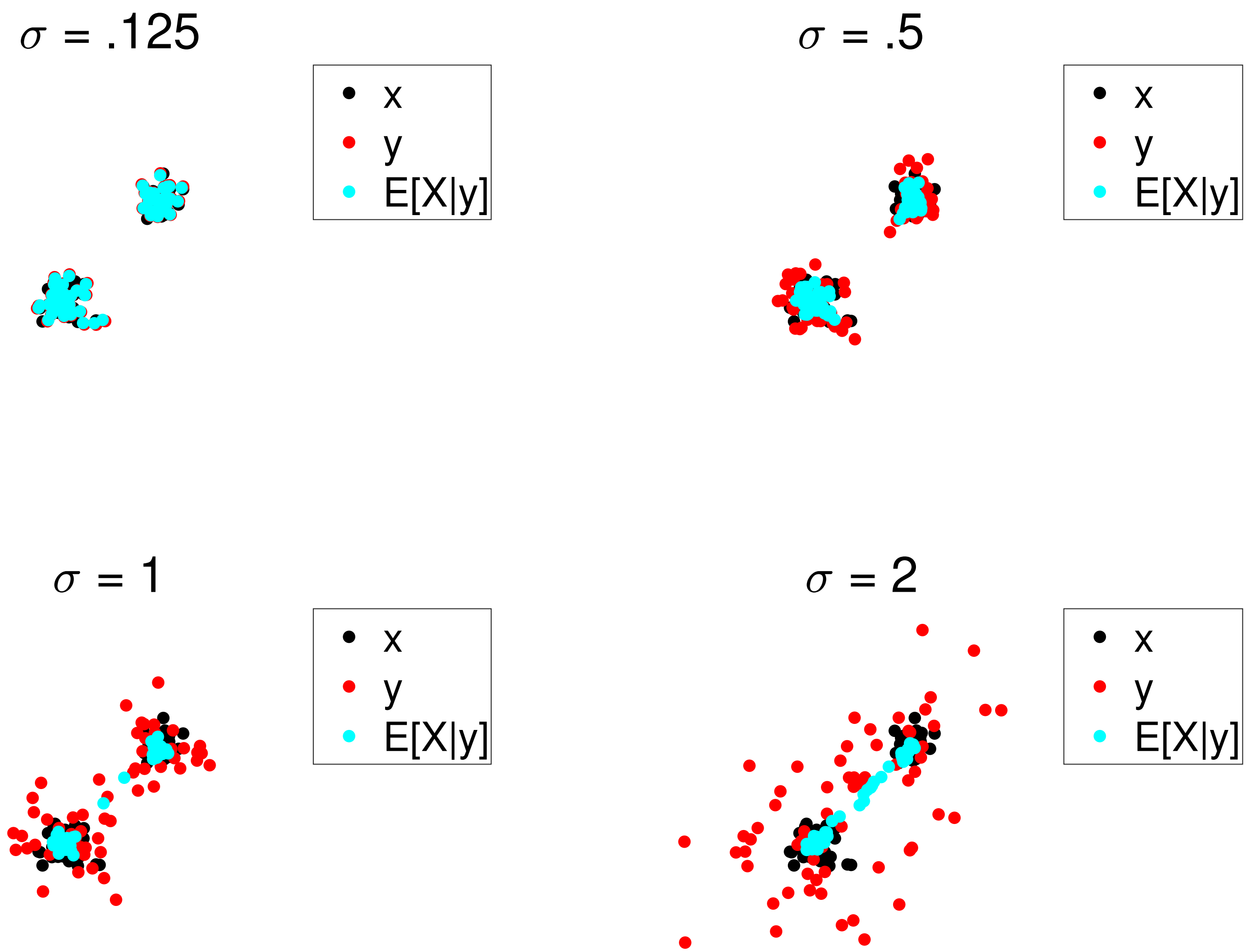
- **Replace** X **by** $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Sampling (provably) easier
  - Going back from Y to X by "optimal" denoising

- **Empirical Bayes** (Robbins, 1956, Miyasawa, 1961)
  - Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Key result: $\mathbb{E}\big[X|Y\big] = Y + \sigma^2 \nabla \log q_\sigma(Y)$
  - Proof by integration by parts
  - No need to know the normalization constant
  - ⚠ "Optimal" does not mean "good" performance

$$W_2\big(\text{law of } X, \text{law of } \mathbb{E}[X|Y]\big)^2 \leqslant \sigma^2 d$$

Sampling by denoising (Saremi and Hyvärinen, 2019)

- **Replace X by $Y = X + \sigma \cdot \mathcal{N}(0, I)$:** effect of denoising



Learning the denoiser from data

- **Empirical Bayes** (Robbins, 1956, Miyasawa, 1961)
  - Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Key result: $\mathbb{E}\big[X|Y\big] = Y + \sigma^2 \nabla \log q_\sigma(Y)$
  - $\nabla \log q_\sigma(Y)$ is the score function

Learning the denoiser from data

- **Empirical Bayes** (Robbins, 1956, Miyasawa, 1961)
  - Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Key result: $\mathbb{E}\big[X|Y\big] = Y + \sigma^2 \nabla \log q_\sigma(Y)$
  - $\nabla \log q_\sigma(Y)$ is the score function

- **Denoising score matching** (Hyvärinen, 2005, Vincent, 2011)
  - Given $x_1, \ldots, x_n$ data sampled from $p(x)$
  - Create noisy samples $y_1, \ldots, y_n \in \mathbb{R}^d$
  - Parameterize $\nabla \log q_\sigma(Y) = f_\theta(Y) \in \mathbb{R}^d$
  - Estimate the density of the noisy variable y by minimizing

$$\frac{1}{n} \sum_{i=1}^{n} \big\| x_i - y_i - \sigma^2 f_\theta(y_i) \big\|^2$$

  - Using classical deep learning models and algorithms

Sampling by denoising (Saremi and Hyvärinen, 2019)

- **Algorithm**

1. Learn score at single scale σ: $Y = X + \sigma \cdot \mathcal{N}(0, I)$
2. Sample Y using Langevin diffusions ("walk")
3. Denoise Y ("jump")

Sampling by denoising (Saremi and Hyvärinen, 2019)

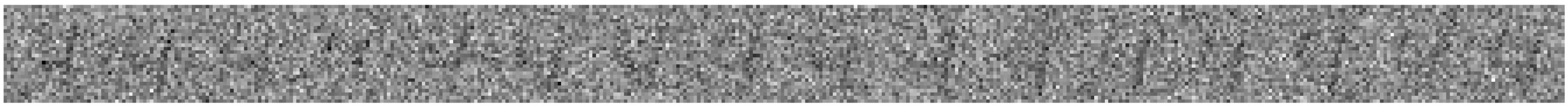
- **Algorithm**

1. Learn score at single scale σ: $Y = X + \sigma \cdot \mathcal{N}(0, I)$
2. Sample Y using Langevin diffusions ("walk")
3. Denoise Y ("jump")

- **Trade-off in choice of σ**
  - σ is too large: Denoising is too "fuzzy"
  - σ is too small: Sampling is difficult

Sampling by denoising (Saremi and Hyvärinen, 2019)

Noisy digits ($\sigma=1$)



Walk-jump sampling ($\sigma=1$)

Walk-jump sampling ($\sigma=1/2$)

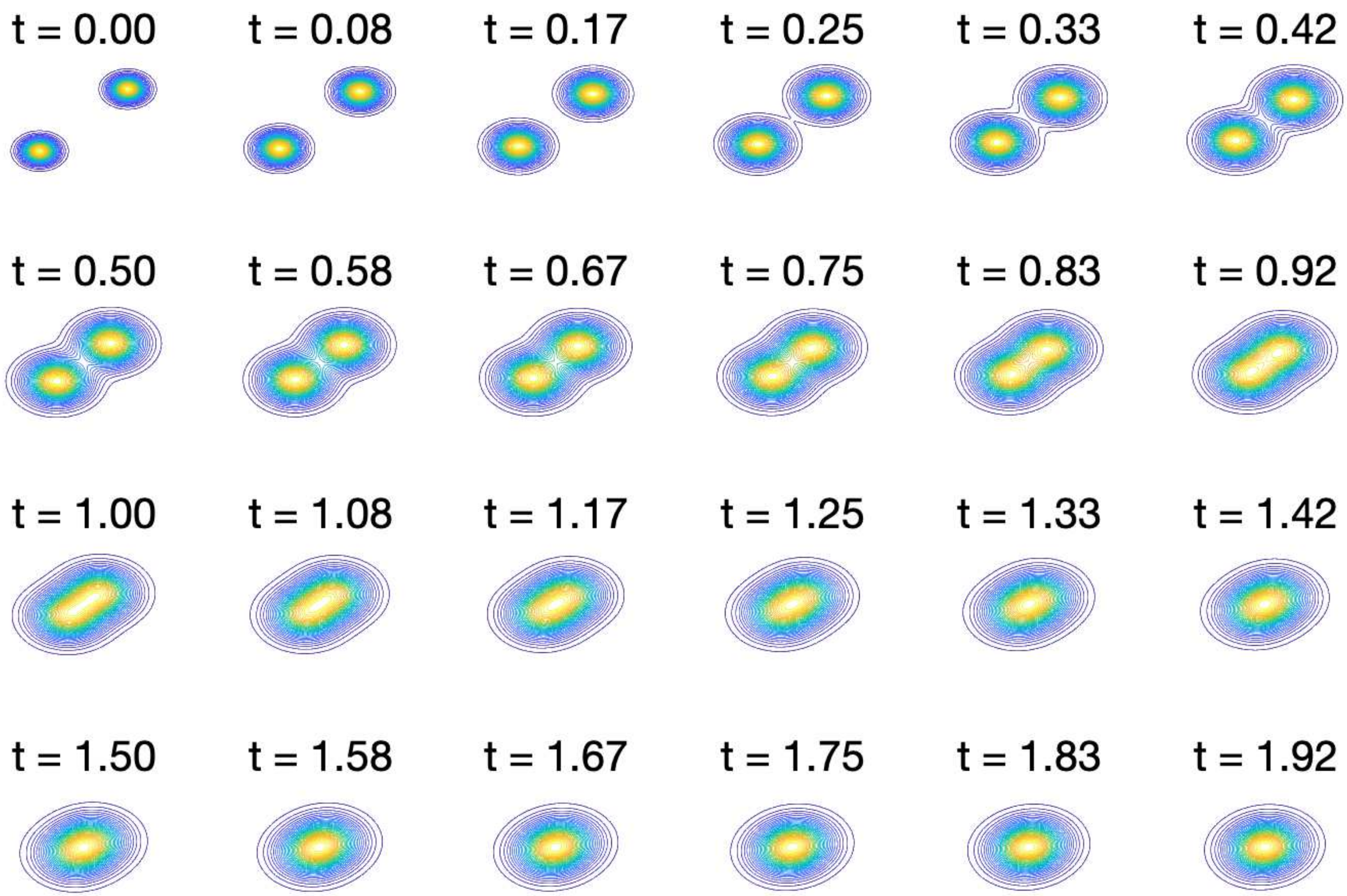
Sampling by denoising
(Saremi, Srivastava, B., 2023)

Denoising diffusion models (Song and Ermon, 2019, Song et al., 2019)

[following expositions from Bortoli (2023) and Peyré (2023)]

- **Forward flow**
  - Ornstein-Uhlenbeck process $dX_t = -X_t dt + \sqrt{2} dB_t$
  - started from $p(x) \propto \exp(-f(x))$ at time $t = 0$
  - marginal distribution: $X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \cdot \mathcal{N}(0, I)$

From data to standard Gaussian



Denoising diffusion models (Song and Ermon, 2019, Song et al., 2019)

[following expositions from Bortoli (2023) and Peyré (2023)]

- **Forward flow**
  - Ornstein-Uhlenbeck process $dX_t = -X_t dt + \sqrt{2} dB_t$
  - started from $p(x) \propto \exp(-f(x))$ at time $t = 0$
  - marginal distribution: $X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \cdot \mathcal{N}(0, I)$

Denoising diffusion models (Song and Ermon, 2019, Song et al., 2019)

[following expositions from Bortoli (2023) and Peyré (2023)]

- **Forward flow**
  - Ornstein-Uhlenbeck process $dX_t = -X_t dt + \sqrt{2} dB_t$
  - started from $p(x) \propto \exp(-f(x))$ at time $t = 0$
  - marginal distribution: $X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \cdot \mathcal{N}(0, I)$

- **Backward flow**
  - For T large, $X_T \approx \mathcal{N}(0, I) \quad \Rightarrow \quad$ backward simulations
  - $Y_t = X_{T-t}$ follows $dY_t = [Y_t + 2\nabla \log r_{T-t}(Y_t)]dt + \sqrt{2} dB_t$ with r_t the density of X_t

Denoising diffusion models
(Song and Ermon, 2019, Song et al., 2019)

[following expositions from Bortoli (2023) and Peyré (2023)]

- **Forward flow**
  - Ornstein-Uhlenbeck process $dX_t = -X_t dt + \sqrt{2} dB_t$
  - started from $p(x) \propto \exp(-f(x))$ at time $t = 0$
  - marginal distribution: $X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \cdot \mathcal{N}(0, I)$

- **Backward flow**
  - For T large, $X_T \approx \mathcal{N}(0, I) \quad \Rightarrow \quad$ backward simulations
  - $Y_t = X_{T-t}$ follows $dY_t = [Y_t + 2\nabla \log r_{T-t}(Y_t)]dt + \sqrt{2} dB_t$ with r_t the density of X_t
  - Simulate the backward SDE using "only" the densities of X_t

$$y_{k+1} = y_k + \gamma y_k + 2\gamma \nabla \log r_{T-\gamma k}(y_k) + \sqrt{2\gamma} \cdot \mathcal{N}(0, I)$$

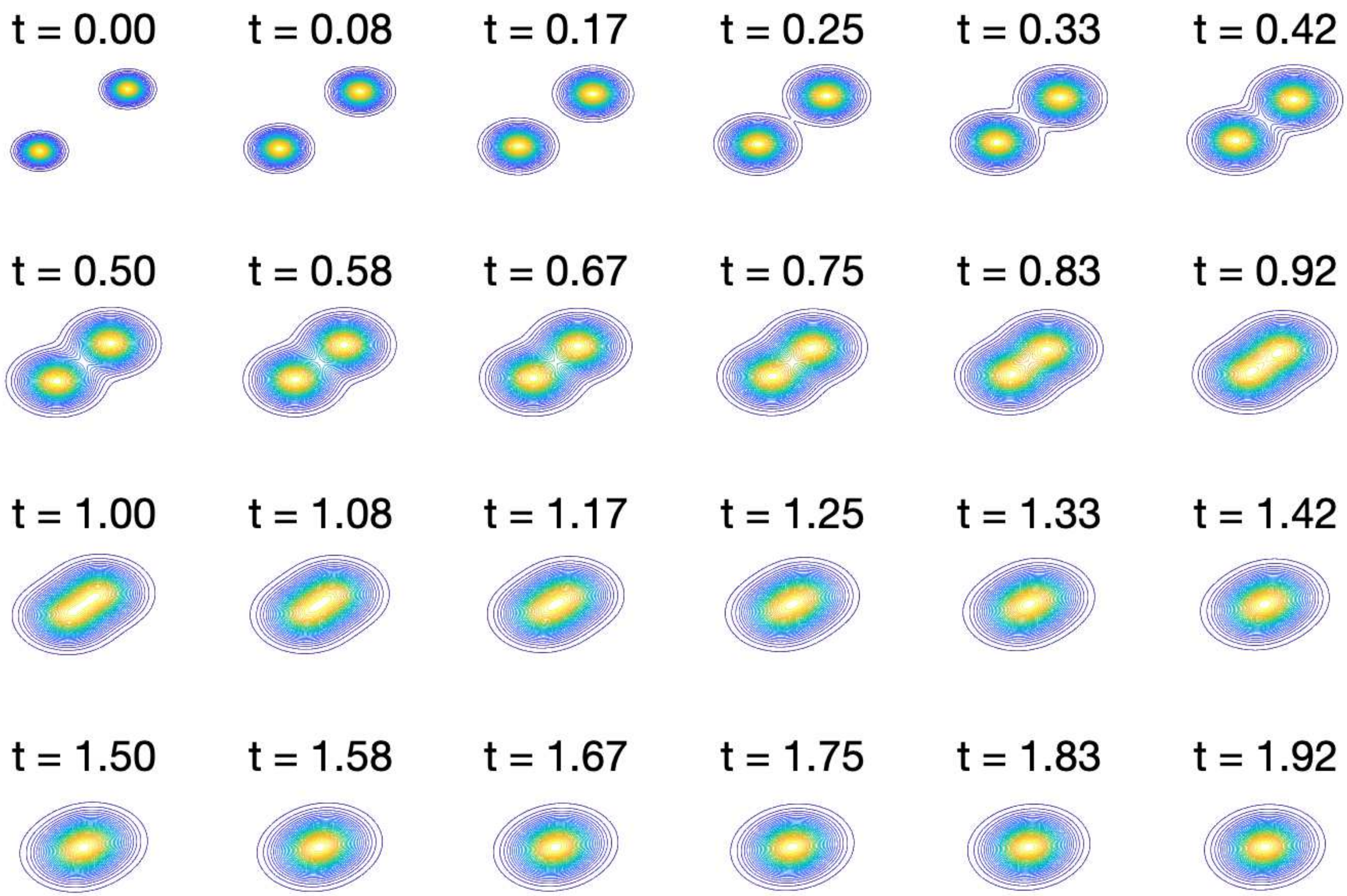
Denoising diffusion models (Song and Ermon, 2019, Song et al., 2019)

- **Learning score functions of noisy samples at various scales**
  - Denoising score matching
- **Denoising diffusion models**
  - Start from T large, $y_0 = X_T$, and discretize the backward SDE

$$y_{k+1} = y_k + \gamma y_k + 2\gamma e^{t_k} \nabla \log q_{\sigma_k}(y_k e^{t_k}) + \sqrt{2\gamma} \cdot \mathcal{N}(0, I)$$

  - with $t_k = T - \gamma k$, and $\sigma_k = e^{T-\gamma k} \sqrt{1 - e^{-2T+2\gamma k}}$

From data to standard Gaussian



Denoising diffusion models (Song and Ermon, 2019, Song et al., 2019)

- **Learning score functions of noisy samples at various scales**
  - Denoising score matching
- **Denoising diffusion models**
  - Start from T large, $y_0 = X_T$, and discretize the backward SDE

$$y_{k+1} = y_k + \gamma y_k + 2\gamma e^{t_k} \nabla \log q_{\sigma_k}(y_k e^{t_k}) + \sqrt{2\gamma} \cdot \mathcal{N}(0, I)$$

  - with $t_k = T - \gamma k$, and $\sigma_k = e^{T-\gamma k}\sqrt{1 - e^{-2T+2\gamma k}}$
- **Alternative view** (Saremi, Park, B., 2023)
  - Diffusion free!

Empirical Bayes with multiple measurements

- **Empirical Bayes** (Robbins, 1956, Miyasawa, 1961)
  - Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Key result: $\mathbb{E}\big[X|Y\big] = Y + \sigma^2 \nabla \log q_\sigma(Y)$

Empirical Bayes with multiple measurements

- **Empirical Bayes** (Robbins, 1956, Miyasawa, 1961)
  - Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Key result: $\mathbb{E}\big[X|Y\big] = Y + \sigma^2 \nabla \log q_\sigma(Y)$

- **Multiple measurements:** $Y_i = X + \varepsilon_i$, $i = 1, \ldots, m$
  - Posterior mean: $\mathbb{E}\big[X|Y_1, \ldots, Y_m\big] = \bar{Y}_{1:m} + \frac{\sigma^2}{m} \nabla \log q_{\sigma/\sqrt{m}}(\bar{Y}_{1:m})$ with $\bar{Y}_{1:m} = \frac{1}{m} \sum_{i=1}^{m} Y_i$

Empirical Bayes with multiple measurements

- **Empirical Bayes** (Robbins, 1956, Miyasawa, 1961)
  - Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Key result: $\mathbb{E}\big[X|Y\big] = Y + \sigma^2 \nabla \log q_\sigma(Y)$

- **Multiple measurements:** $Y_i = X + \varepsilon_i$, $i = 1, \dots, m$
  - Posterior mean: $\mathbb{E}\big[X|Y_1, \dots, Y_m\big] = \bar{Y}_{1:m} + \frac{\sigma^2}{m} \nabla \log q_{\sigma/\sqrt{m}}(\bar{Y}_{1:m})$ with $\bar{Y}_{1:m} = \frac{1}{m} \sum_{i=1}^m Y_i$
  - Increased concentration around the mean (S., P. and B., 2023)

$$W_2\big(\text{law of } X, \text{law of } \mathbb{E}[X|Y_1, \dots, Y_m]\big)^2 \leqslant \frac{\sigma^2 d}{m}$$

  - Improved results with "strong" priors

Empirical Bayes with multiple measurements

- **Empirical Bayes** (Robbins, 1956, Miyasawa, 1961)
  - Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Key result: $\mathbb{E}\big[X|Y\big] = Y + \sigma^2 \nabla \log q_\sigma(Y)$

- **Multiple measurements:** $Y_i = X + \varepsilon_i$, $i = 1, \dots, m$
  - Posterior mean: $\mathbb{E}\big[X|Y_1, \dots, Y_m\big] = \bar{Y}_{1:m} + \frac{\sigma^2}{m} \nabla \log q_{\sigma/\sqrt{m}}(\bar{Y}_{1:m})$ with $\bar{Y}_{1:m} = \frac{1}{m}\sum_{i=1}^m Y_i$
  - Increased concentration around the mean (S., P. and B., 2023)

$$W_2\big(\text{law of } X, \text{law of } \mathbb{E}[X|Y_1, \dots, Y_m]\big)^2 \leqslant \frac{\sigma^2 d}{m}$$

  - Improved results with "strong" priors

- **Idea #1** (Saremi and Srivastava, 2022)
  - Sampling X by sampling $Y_1, \dots, Y_m$ and then Empirical Bayes

Multimeasurement generative models (Saremi and Srivastava, 2022)

x y_1 y_2 y_3 y_4 $\bar{y}_{1:m}$

$\mathbb{E}[x|y_1, \ldots, y_m]$

- **Still hard to sample from** $(y_1, \ldots, y_m)$

Idea #2: Sequential denoising (S., P. and B., 2023)

- **Multiple measurements:** $Y_i = X + \varepsilon_i$, $i = 1, \ldots, m$
- **Algorithm**
  - Sample y_1 from Y_1
  - Iteratively sample y_i from $Y_i | y_1, \ldots, y_{i-1}$, for $i = 1, \ldots, m$

Idea #2: Sequential denoising (S., P. and B., 2023)

- **Multiple measurements:** $Y_i = X + \varepsilon_i$, $i = 1, \ldots, m$

- **Algorithm**
  - Sample y_1 from Y_1
  - Iteratively sample y_i from $Y_i | y_1, \ldots, y_{i-1}$, for $i = 1, \ldots, m$

- **Sampling steps using Langevin algorithms**
  - Feasibility:

$$\nabla_{y_m} \log p(y_m | y_1, \ldots, y_{m-1}) = \frac{1}{\sigma^2}\Big[\bar{y}_{1:m} - y_m + \frac{\sigma^2}{m} \nabla \log q_{\sigma/\sqrt{m}}(\bar{y}_{1:m})\Big]$$

Idea #2: Sequential denoising (S., P. and B., 2023)

- **Multiple measurements:** $Y_i = X + \varepsilon_i$, $i = 1, \ldots, m$

- **Algorithm**
  - Sample y_1 from Y_1
  - Iteratively sample y_i from $Y_i | y_1, \ldots, y_{i-1}$, for $i = 1, \ldots, m$

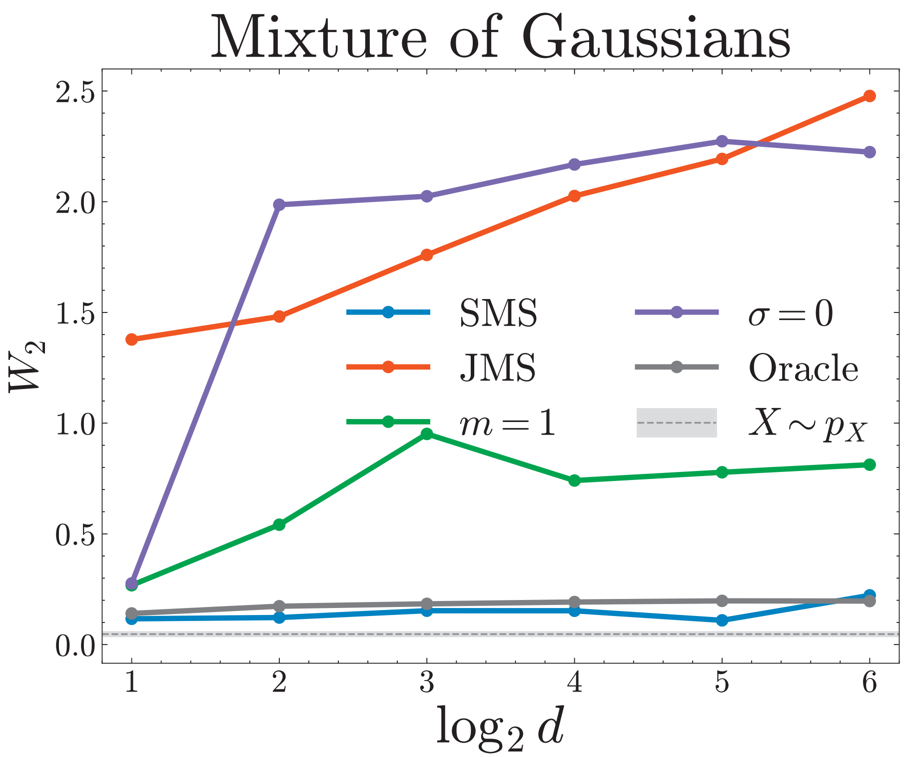
- **Sampling steps using Langevin algorithms**
  - Feasibility:

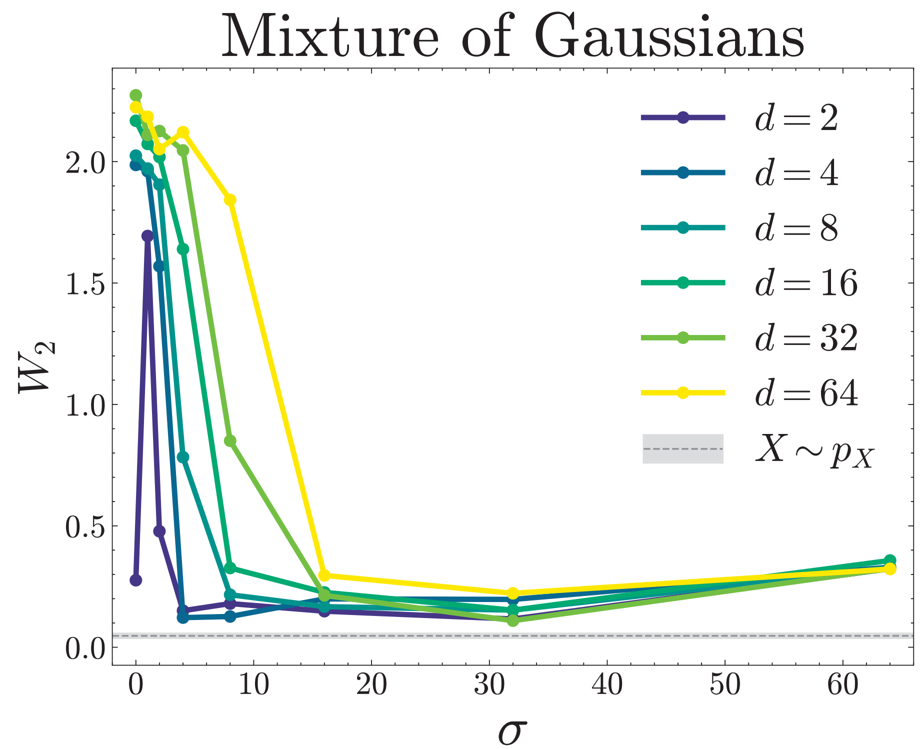
$$\nabla_{y_m} \log p(y_m | y_1, \ldots, y_{m-1}) = \frac{1}{\sigma^2} \Big[\bar{y}_{1:m} - y_m + \frac{\sigma^2}{m} \nabla \log q_{\sigma/\sqrt{m}}(\bar{y}_{1:m}) \Big]$$

- **Main benefit**
  - If σ large enough, only log-concave distributions to sample from
  - If m large enough, $\frac{\sigma}{\sqrt{m}}$ is small enough to obtain clean samples

Synthetic experiments

- **Mixtures of two Gaussians**
  - covariance matrices $\tau^2 I$, $\Delta\mu = 6 \cdot (1, \ldots, 1) \in \mathbb{R}^d$





  - SMS (sequential multimeasurement sampling)
  - JMS (joint multimeasurement sampling)

Discussion

- **Three main ideas**
  - Sampling by denoising
  - Learning the denoiser from data
  - Progressive denoising for improved sampling/denoising
    - Using continuous-time diffusions (Song and Ermon, 2019)
    - Using multiple measurements (Park, Saremi, and B., 2024)

Discussion

- **Three main ideas**
  - Sampling by denoising
  - Learning the denoiser from data
  - Progressive denoising for improved sampling/denoising
    - Using continuous-time diffusions (Song and Ermon, 2019)
    - Using multiple measurements (Park, Saremi, and B., 2024)

- **Key open problems**
  - Theoretical analysis
  - Beyond Gaussians and Euclidean geometry
  - Conditional sampling
  - Rigorous empirical evaluation

References

- Valentin Bortoli. Generative modeling, https://vdeborto.github.io/project/generative_modeling/, 2023.
- Gabriel Peyré. Denoising Diffusion Models, https://mathematical-tours.github.io/book-sources/optim-ml/OptimML-DiffusionModels.pdf, 2023.
- Bakry, D., Cattiaux, P. and Guillin, A. Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré. J. Funct. Anal. 254 727–759, 2008.
- Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., & Zhang, A. R. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. arXiv preprint arXiv:2209.11215. 2022.
- Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. Journal of the Royal Statistical Society. Series B (Statistical Methodology), pp. 651–676, 2017.

- Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. The Annals of Applied Probability, 27(3):1551 – 1587, 2017.

- Frey, N. C., Berenberg, D., Zadorozhny, K., Kleinhenz, J., Lafrance-Vanasse, J., Hotzel, I., . . . , & Saremi, S. Protein discovery with discrete walk-jump sampling. arXiv preprint arXiv:2306.12360, 2023

- Sinho Chewi. Log-Concave Sampling `https://chewisinho.github.io/main.pdf`, 2023.

- Andrea Montanari. Sampling, Diffusions, and Stochastic Localization. arXiv preprint arXiv:2305.10690, 2023.

- Herbert Robbins. An empirical Bayes approach to statistics. In Proc. Third Berkeley Symp., volume 1, pp. 157–163, 1956.

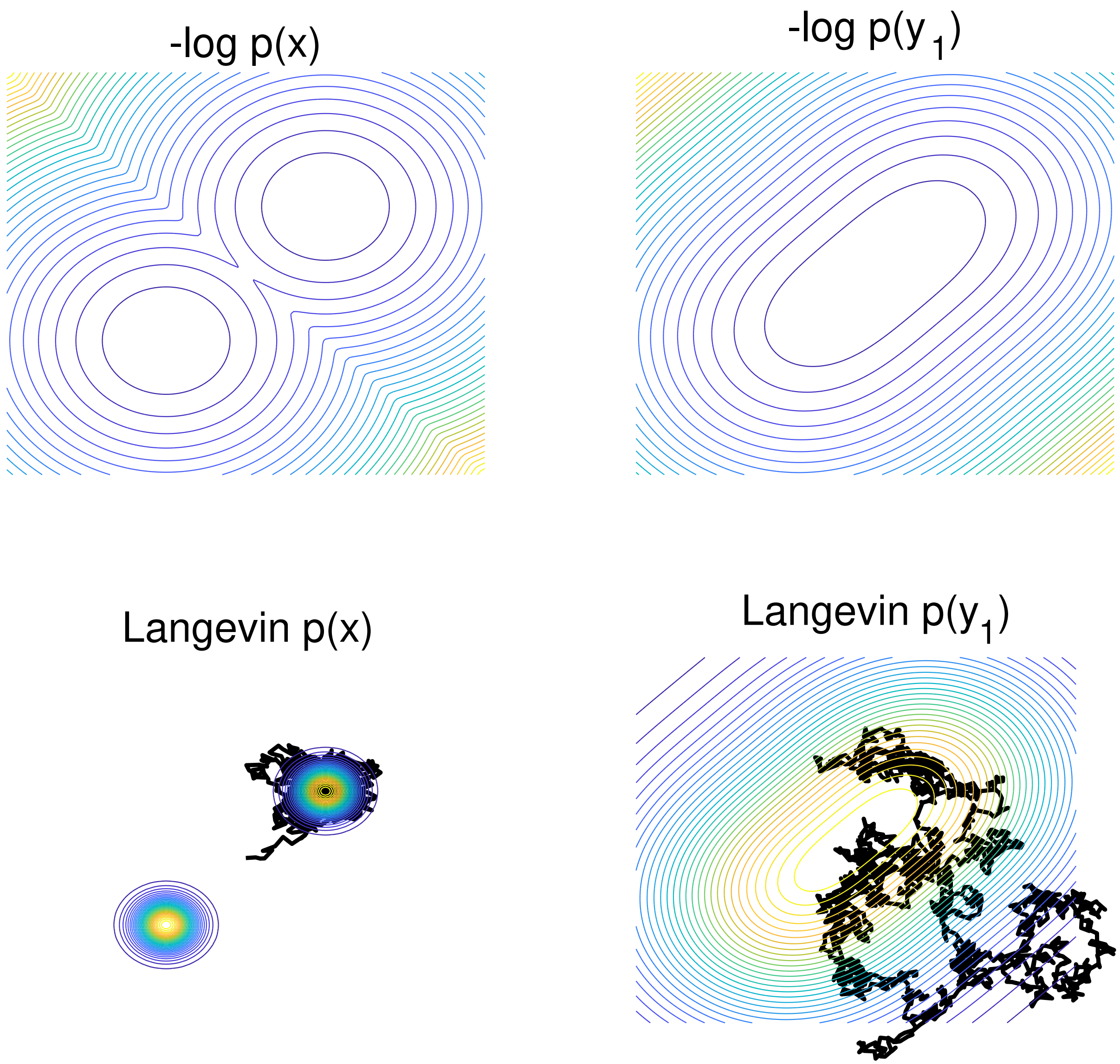
- Koichi Miyasawa. An empirical Bayes estimator of the mean of a normal population. Bulletin of the International Statistical Institute, 38(4):181–188, 1961.

- Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research, 6(Apr):695–709, 2005.
- Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems, 32, 2019.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations, 2021.
- Pascal Vincent. A connection between score matching and denoising autoencoders. Neural computation, 23(7):1661–1674, 2011.
- Saeed Saremi and Rupesh Kumar Srivastava. Multimeasurement generative models. In International Conference on Learning Representations, 2022.
- Conforti, Giovanni, Alain Durmus, and Marta Gentiloni Silveri. Score diffusion models without early stopping: finite Fisher information is all you need. arXiv preprint arXiv:2308.12240, 2023

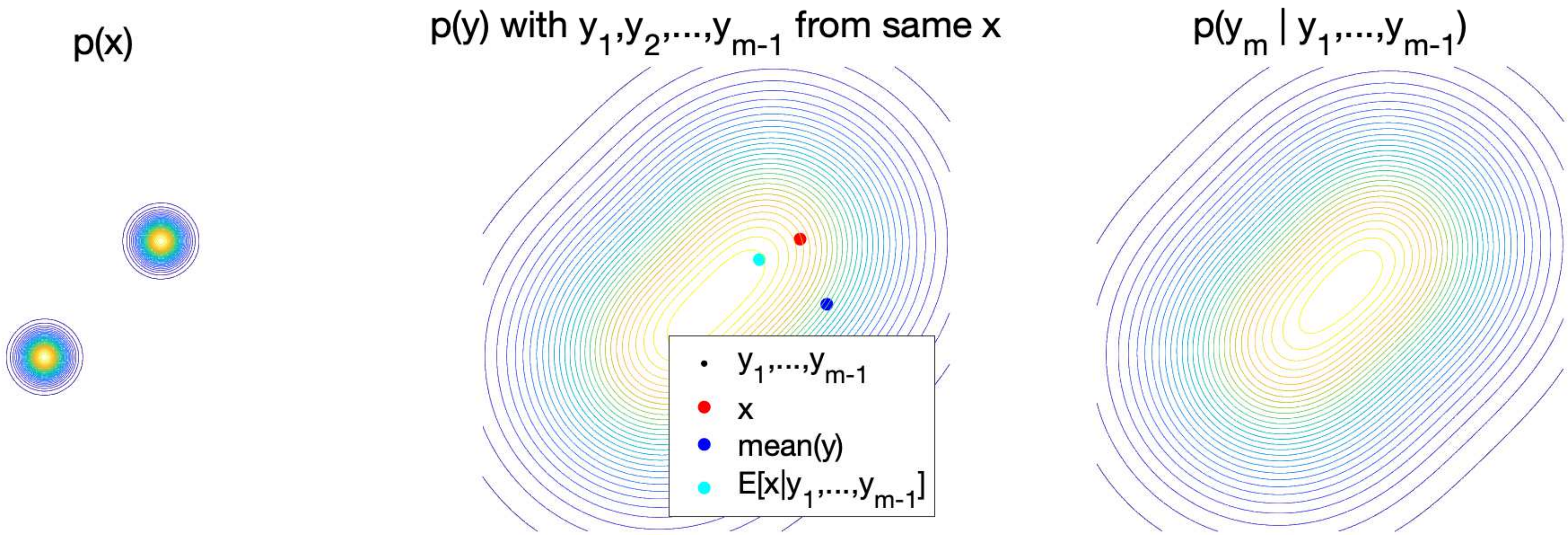
- S. Saremi, R. K. Srivastava, F. Bach. Universal Smoothed Score Functions for Generative Modeling. Technical report, arXiv:2303.11669, 2023.

- S. Saremi, J.-W. Park, F. Bach. Chain of Log-Concave Markov Chains. International Conference on Learning Representations (ICLR), 2024.

First step

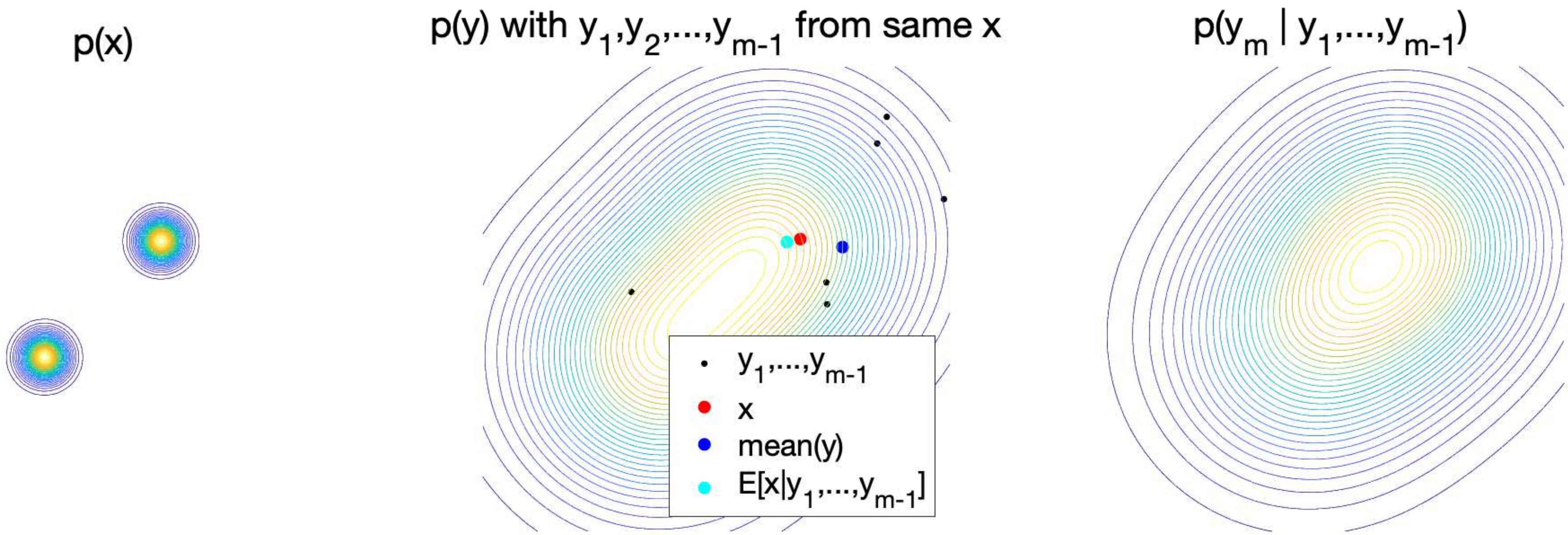


$m = 2$

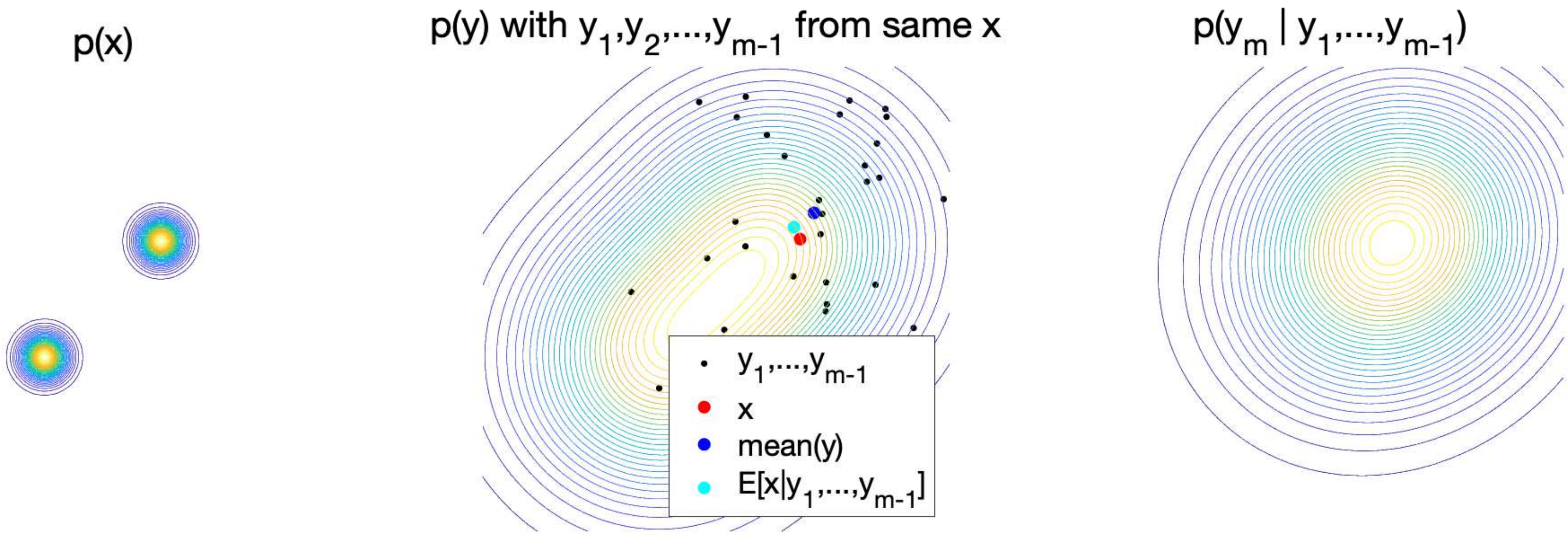


p(x)
p(y) with y_1,y_2,...,y_{m-1} from same x
y_1,...,y_{m-1}
x
mean(y)
E[x|y_1,...,y_{m-1}]
p(y_m | y_1,...,y_{m-1})

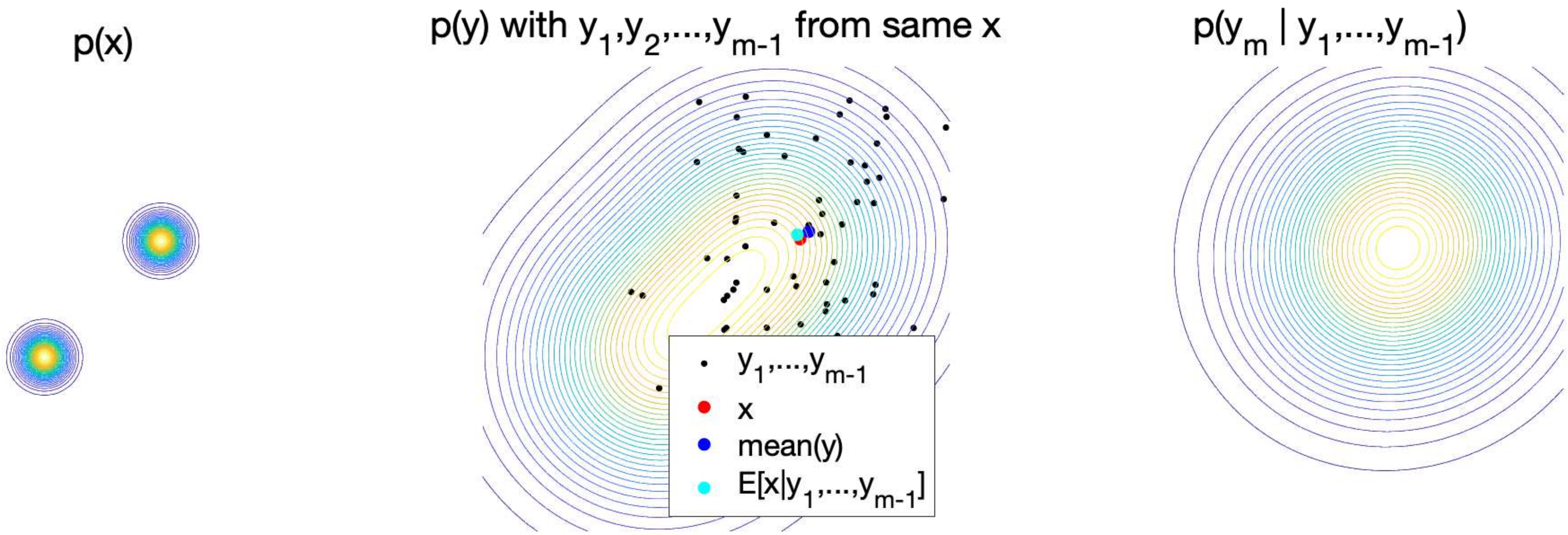
$m = 8$


p(x)
p(y) with $y_1, y_2, \ldots, y_{m-1}$ from same x
$y_1, \ldots, y_{m-1}$
x
mean(y)
$E[x|y_1, \ldots, y_{m-1}]$
$p(y_m \mid y_1, \ldots, y_{m-1})$

$$m = 32$$


p(x)
p(y) with y1,y2,...,ym-1 from same x
p(ym | y1,...,ym-1)
y1,...,ym-1
x
mean(y)
E[x|y1,...,ym-1]

$m = 64$


p(x)
p(y) with y1,y2,...,ym-1 from same x
p(ym | y1,...,ym-1)
y1,...,ym-1
x
mean(y)
E[x|y1,...,ym-1]