An alternative view of denoising diffusion models

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

Joint work with Ji Won Park and Saeed Saremi

October 2023

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

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 $dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$:

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

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

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 $dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$:

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

  - (slow) convergence (see, e.g., Bakry et al., 2008)
  - fast for smooth log-concave distributions
    (Dalalyan, 2014, Durmus and Moulines, 2017, Chewi, 2022, etc.)

- **Going beyond log-concave distributions**

A short introduction to 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: $X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \cdot \mathcal{N}(0, I)$

A short introduction to 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

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

Denoising score matching

- **Score functions after adding noise** $\nabla \log q_t(x) = \frac{\nabla q_t}{q_t}(x)$
  - with q_t density of $X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \cdot \mathcal{N}(0, I)$
  - equivalent to density of $X_0 + e^t \sqrt{1 - e^{-2t}} \cdot \mathcal{N}(0, I) = X_0 + \sigma \cdot \mathcal{N}(0, I)$

Denoising score matching

- **Score functions after adding noise** $\nabla \log q_t(x) = \frac{\nabla q_t}{q_t}(x)$
  - with q_t density of $X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} \cdot \mathcal{N}(0, I)$
  - equivalent to density of $X_0 + e^t \sqrt{1 - e^{-2t}} \cdot \mathcal{N}(0, I) = X_0 + \sigma \cdot \mathcal{N}(0, I)$

- **Empirical Bayes** (Robbins, 1956, Miyasawa, 1961)
  - Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Key result: $\mathbb{E}\big[X|Y\big] = Y + \sigma^2 \nabla \log q_\sigma(Y)$
  - Used within sampling procedure by Saremi and Hyvärinen (2019)

Denoising score matching

- **Empirical Bayes** (Robbins, 1956, Miyasawa, 1961)
  - Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Key result: $\mathbb{E}\big[X|Y\big] = Y + \sigma^2 \nabla \log q_\sigma(Y)$
  - Used within sampling procedure by Saremi and Hyvärinen (2019)

Denoising score matching

- **Empirical Bayes** (Robbins, 1956, Miyasawa, 1961)
  - Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Key result: $\mathbb{E}\big[X|Y\big] = Y + \sigma^2 \nabla \log q_\sigma(Y)$
  - Used within sampling procedure by Saremi and Hyvärinen (2019)

- **Score matching** (Hyvärinen, 2005)
  - Fitting q to p, by mininimizing $\displaystyle\int p(x) \|\nabla \log p(x) - \nabla \log q(x)\|^2 dx$

Denoising score matching

- **Empirical Bayes** (Robbins, 1956, Miyasawa, 1961)
  - Notation: q_σ density of $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Key result: $\mathbb{E}\big[X|Y\big] = Y + \sigma^2 \nabla \log q_\sigma(Y)$
  - Used within sampling procedure by Saremi and Hyvärinen (2019)

- **Score matching** (Hyvärinen, 2005)
  - Fitting q to p, by mininimizing $\displaystyle\int p(x) \|\nabla \log p(x) - \nabla \log q(x)\|^2 dx$

- **Denoising score matching** (Hyvärinen, 2005, Vincent, 2011)
  - Estimate the density of the noisy variable y by minimizing

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

A short introduction to denoising diffusion models (Song and Ermon, 2019, Song et al., 2019)

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

- **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 $\sigma_k = e^{T - \gamma k} \sqrt{1 - e^{-2T + 2\gamma k}}$

A short introduction to denoising diffusion models (Song and Ermon, 2019, Song et al., 2019)

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

- **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)

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 $\bar{Y}_{1:m} = \frac{1}{m}\sum_{i=1}^m Y_i$
  - Increased concentration around the mean (S., P. and B., 2023)

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

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

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

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

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

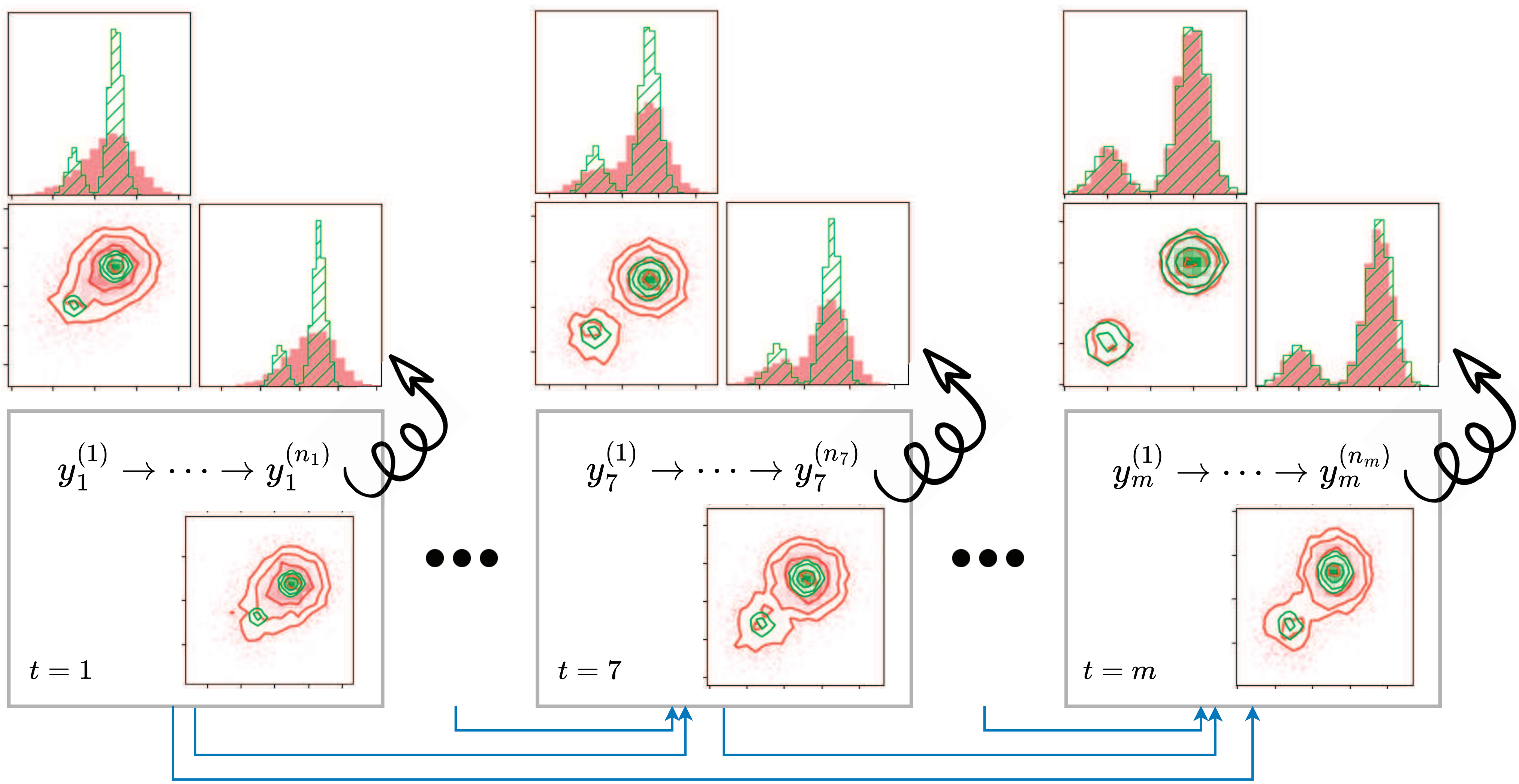
- **Multiple measurements:** $Y_i = X + \varepsilon_i$, $i = 1, \ldots, m$

- **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**
  - Overall non-Markovian
  - Each sampling step Markovian

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



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

More and more log-concave

- **Single measurement:** $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Enough Gaussian blurring leads to unimodality (Loog et al., 2001)
  - Enough Gaussian blurring leads to log-concavity
  - "Proof" (see paper for quantitative statements)

$$\nabla^2 \log q(y) = -\frac{1}{\sigma^2}\Big[I - \frac{1}{\sigma^2}\mathrm{cov}(X|Y = y)\Big]$$

More and more log-concave

- **Single measurement:** $Y = X + \sigma \cdot \mathcal{N}(0, I)$
  - Enough Gaussian blurring leads to unimodality (Loog et al., 2001)
  - Enough Gaussian blurring leads to log-concavity
  - "Proof" (see paper for quantitative statements)

$$\nabla^2 \log q(y) = -\frac{1}{\sigma^2}\Big[I - \frac{1}{\sigma^2}\mathrm{cov}(X|Y = y)\Big]$$

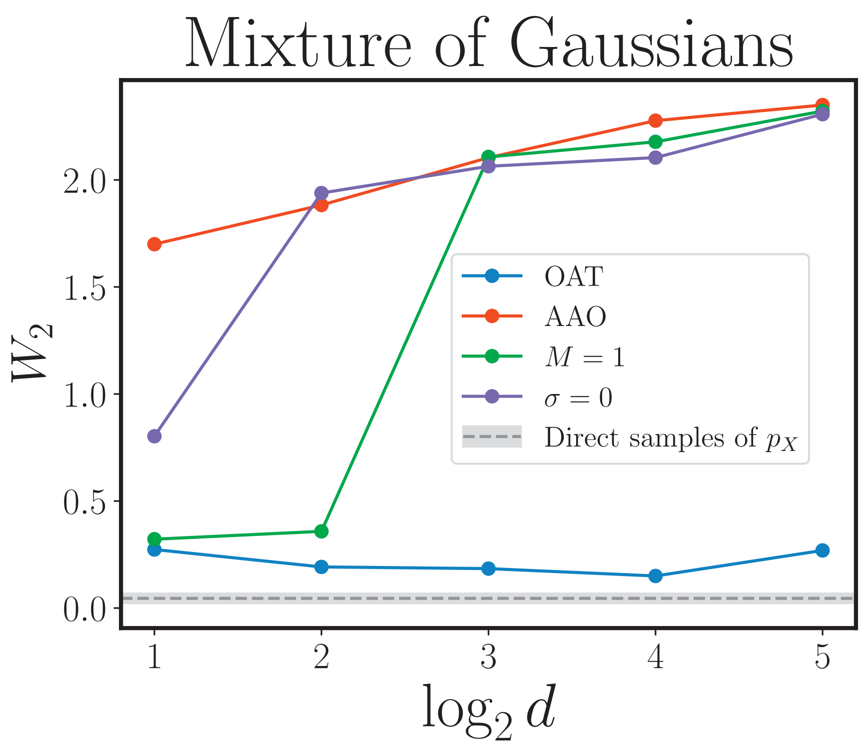
- **Multiple measurements:** $Y_i = X + \sigma \cdot \mathcal{N}(0, I)$, $i = 1, \dots, m$

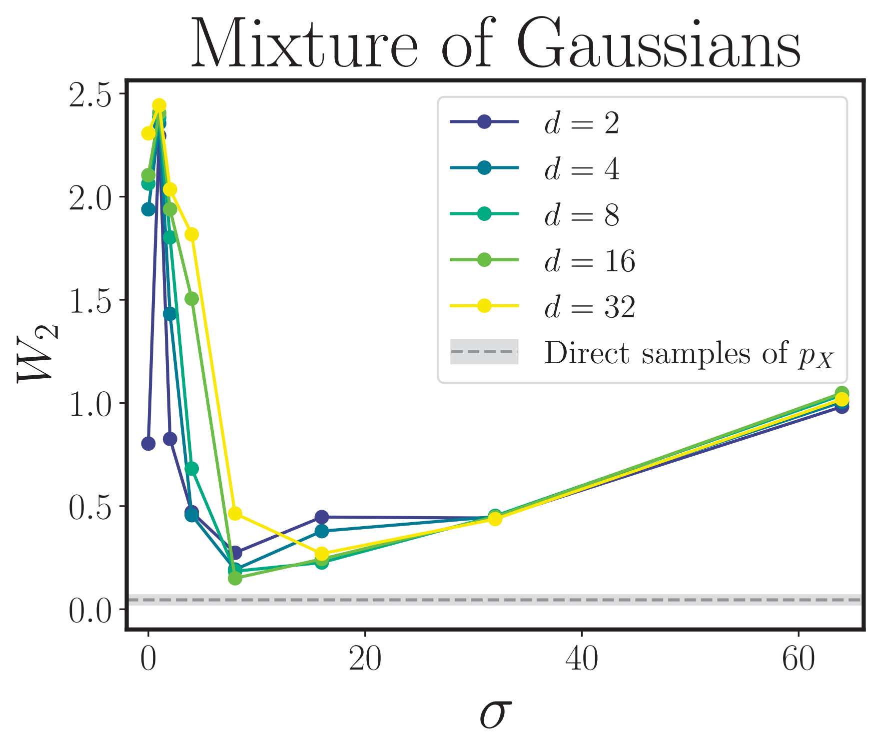
$$\nabla^2_{y_m} \log p(y_m|y_1, \dots, y_{m-1}) = -\frac{1}{\sigma^2}\Big[I - \frac{1}{\sigma^2}\mathrm{cov}(X|y_1, \dots, y_m)\Big]$$

  - Conditioning reduces uncertainty (on average)
  - See precise statements in paper

Synthetic experiments

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





Discussion

- **Sampling from score functions of smoothed densities**
  - Similar steps to denoising diffusion models
  - Clear initialization: σ large enough to obtain log-concavity
  - m large enough to obtain good quality samples
  - Only two hyperparameters: noise level σ and number of measurements m

Discussion

- **Sampling from score functions of smoothed densities**
  - Similar steps to denoising diffusion models
  - Clear initialization: σ large enough to obtain log-concavity
  - m large enough to obtain good quality samples
  - Only two hyperparameters: noise level σ and number of measurements m
- **Extensions**
  - Application to image generation
  - Theoretical analysis (see Conforti et al., 2023)
- **Preprint**
  - Saeed Saremi, Ji Won Park, Francis Bach. Chain of Log-Concave Markov Chains. arXiv:2305.19473, 2023.

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.

- 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.

- Sinho Chewi. Log-Concave Sampling https://chewisinho.github.io/main.pdf, 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, J.-W. Park, F. Bach. Chain of Log-Concave Markov Chains. Technical report, arXiv:2305.19473, 2023.