Annealed Flow Transport Monte Carlo

Michael Arbel [*,1,†] Alexander G. D. G. Matthews [*,2] Arnaud Doucet [2]

[*]Equal Contribution

[1]Gatsby Computational Neuroscience Unit, UCL, UK,

[2]DeepMind

[†]Currently at INRIA, Grenoble Rhône-Alpes, France

Part I: Presentation of the method

A particle algorithm for sampling from unnormalized densities.

- ✓ Combines Normalizing Flows (NFs) and Sequential Monte Carlo methods for increased flexibility and adaptivity to the sampling task.
- ✓ Provides consistent estimates when the number of particles increases.
- ✓ Using NFs can provably reduce the asymptotic variance of the estimates.
- ✓ Interpretation of AFT as an optimal control problem for a weighted SDE.
- ✓ Provides a modular plug-and-play implementation.
- ✓ Competitive results compared to challenging benchmarks.

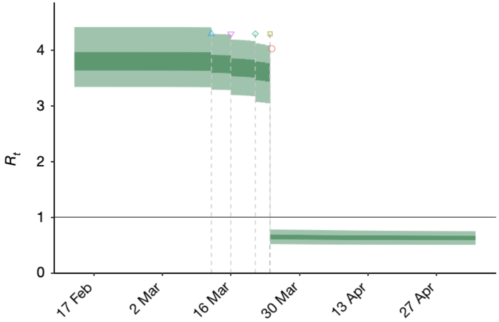
Sampling from intractable densities

Target $\pi(x) = Z^{-1}e^{-V(x)}$

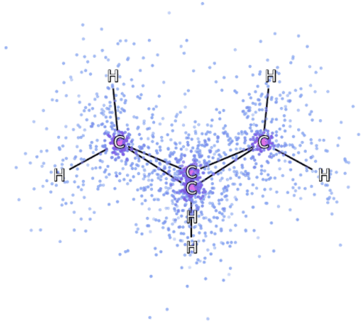
- Goal 1: Sampling from a target density π known up to a normalizing constant Z.
- Goal 2: Estimating the normalizing constant Z.

Sampling from intractable densities: Applications

Bayesian statistics, Compression, Statistical physics, Chemistry.



Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe.
See Flaxman, Mishra, Gandy et al. Nature 2020.



FermiNet project.
See Pfau, Spencer, Matthews and Foulkes.
Physical Review Research 2020.

Sampling from intractable densities: Challenges

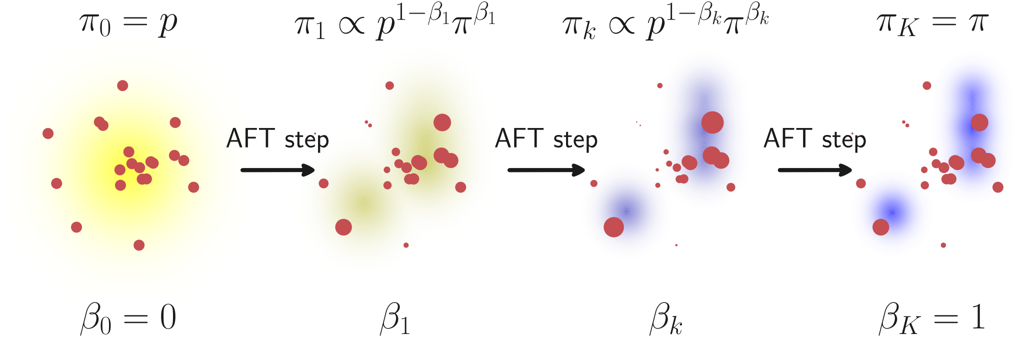
Target $\pi(x) = Z^{-1} e^{-V(x)}$

Challenges:

- Curse of dimensionality.
- Multimodality.
- Limitations of SOTA methods:
  - Accurate estimates require careful design of the algorithms like AIS [Neal, 2001], SMC [Del Moral et al., 2006]
  - Tail under-estimation of flow-based methods [Domke and Sheldon 2018].

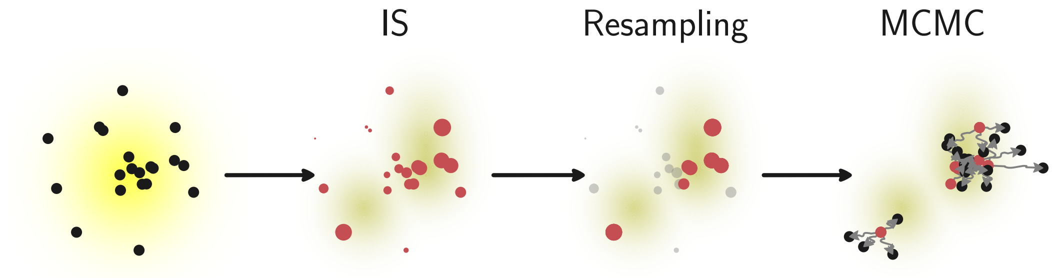
Annealed Flow Transport

We combine SMC methods with NFs to gain the best from both approaches.



- Similarly to SMC: Introduce a sequence of densities π_k interpolating between a proposal p and the target π.
- Sequential sampling: Use samples from π_{k-1} to compute samples from π_k.
- AFT step: combines a Flow transport step followed by standard SMC steps.

Sequential Monte Carlo steps (no flow)



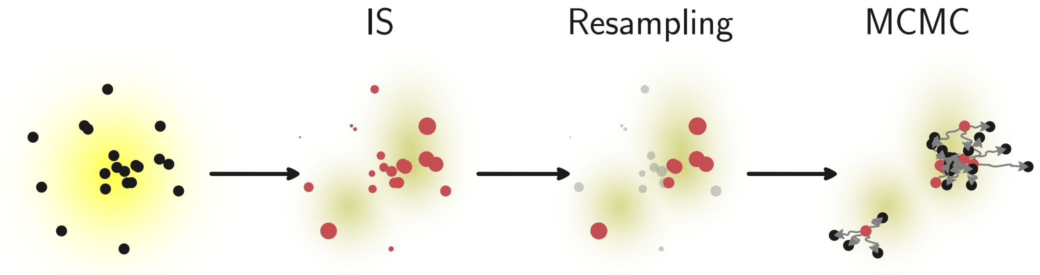
$$\frac{W_k^i}{W_{k-1}^i} \propto \frac{\pi_k(X_{k-1}^i)}{\pi_{k-1}(X_{k-1}^i)}$$

$$\bar{X}_k^i = X_{k-1}^j \qquad j \sim \text{Multi}(W_k)$$

$$X_k^i \sim K_k(\bar{X}_k^i, X_k^i)$$

- Importance Sampling: re-weights particles from $k-1$ proportionally to $\frac{\pi_k}{\pi_{k-1}}$.
- Resampling: duplicate particles with large weights and discard those with small weights. (Recovers AIS (Neal, 2001) if no resampling).
- MCMC step: Move particles according to a Markov Kernel K_k with invariant distribution π_k: (HM, Gibbs-samplers, etc).

Sequential Monte Carlo steps (no flow)

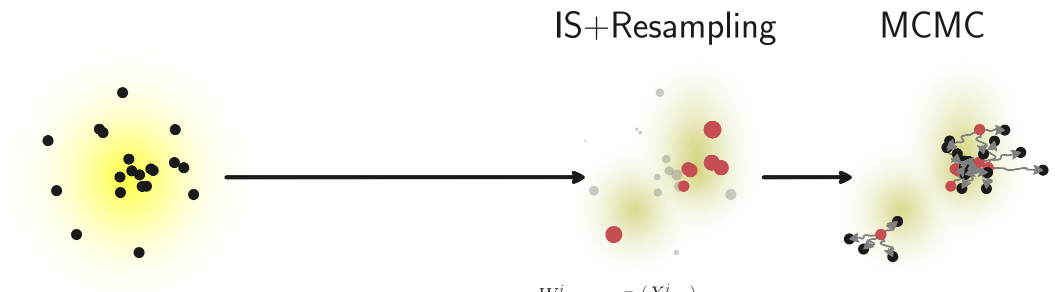


$$\frac{W_k^i}{W_{k-1}^i} \propto \frac{\pi_k(X_{k-1}^i)}{\pi_{k-1}(X_{k-1}^i)} \qquad \begin{array}{c} \bar{X}_k^i = X_{k-1}^j \\ j \sim \text{Multi}(W_k) \end{array} \qquad X_k^i \sim K_k(\bar{X}_k^i, X_k^i)$$

- Estimating normalizing constant Z_k sequentially:

$$Z_k^N := Z_{k-1}^N \left(\sum_{i=1}^{N} W_{k-1}^i \frac{\pi_k\left(X_{k-1}^i\right)}{\pi_{k-1}\left(X_{k-1}^i\right)} \right)$$

Sequential Monte Carlo steps (no flow)



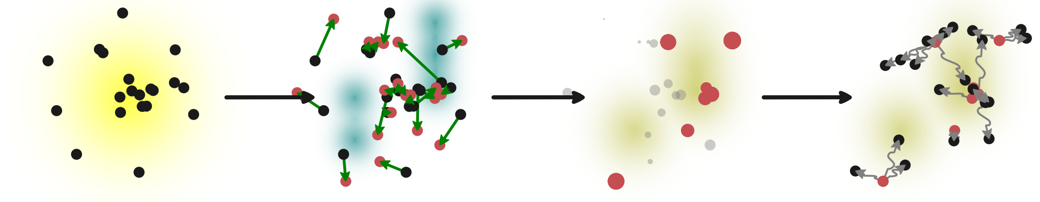
$$\frac{W_k^i}{W_{k-1}^i} \propto \frac{\pi_k(X_{k-1}^i)}{\pi_{k-1}(X_{k-1}^i)}$$

$$\bar{X}_k^i = X_{k-1}^j,\ j \sim \text{Multi}(W_k)$$

$$X_k^i \sim K_k(\bar{X}_k^i, .)$$

Annealed Flow Transport steps (with a flow)

Flow Transport IS + Resampling MCMC



$$\tilde{X}_k^i = T_k\left(X_{k-1}^i\right)$$

$$\frac{W_k^i}{W_{k-1}^i} \propto G_k\left(X_{k-1}^i, \tilde{X}_k^i\right)$$
$$\bar{X}_k^i = \tilde{X}_k^j,\ j \sim \text{Multi}(W_k)$$

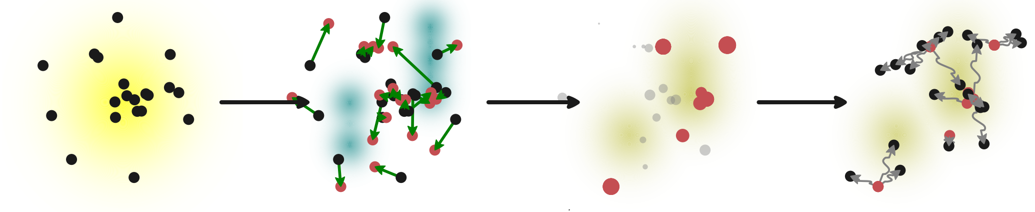
$$X_k^i \sim K_k\left(\bar{X}_k^i, .\right)$$

- Flow Transport T_k moves X_{k-1}^i to new particles $\tilde{X}_k^i$ close to π_k.
- Closed-form expression for the IS weights to correct for inexact flow:

$$G_k(X, Y) = \frac{\pi_k(Y)}{\pi_{k-1}(X)} |\nabla T_k(X)|$$

Annealed Flow Transport steps (with a flow)

Flow Transport IS + Resampling MCMC



$$\tilde{X}_k^i = T_k\left(X_{k-1}^i\right)$$

$$\frac{W_k^i}{W_{k-1}^i} \propto G_k\left(X_{k-1}^i, \tilde{X}_k^i\right)$$

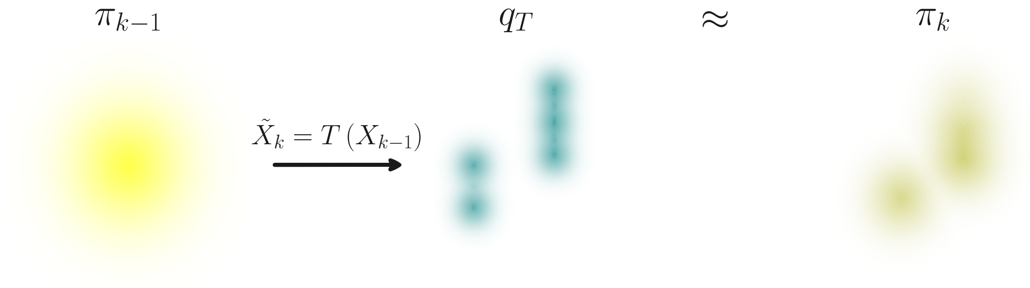
$$\bar{X}_k^i = \tilde{X}_k^j,\ j \sim \text{Multi}(W_k)$$

$$X_k^i \sim K_k\left(\bar{X}_k^i, .\right)$$

- Estimating normalizing constant Z_t sequentially:

$$Z_k^N := Z_{k-1}^N \left(\sum_{i=1}^{N} W_{k-1}^i G_k\left(X_{k-1}^i, X_k^i\right) \right)$$

Learning the Normalizing Flows sequentially

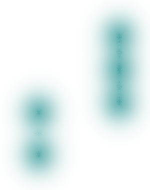


Learning the Normalizing Flows sequentially

π_{k-1}

$\tilde{X}_k = T\left(X_{k-1}\right)$

q_T

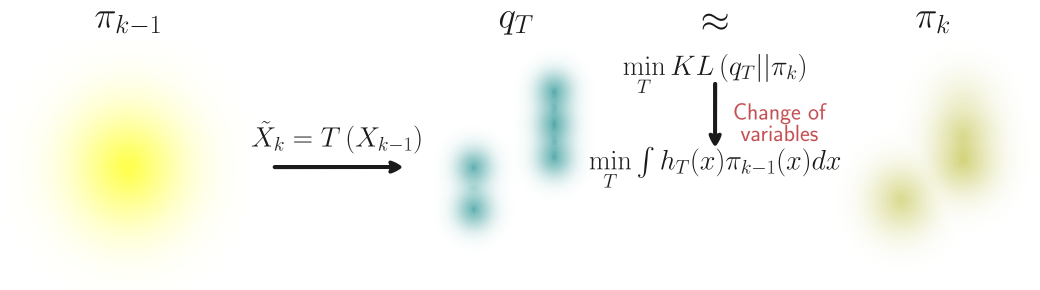


$\approx$

$\min_T KL\left(q_T || \pi_k\right)$

π_k

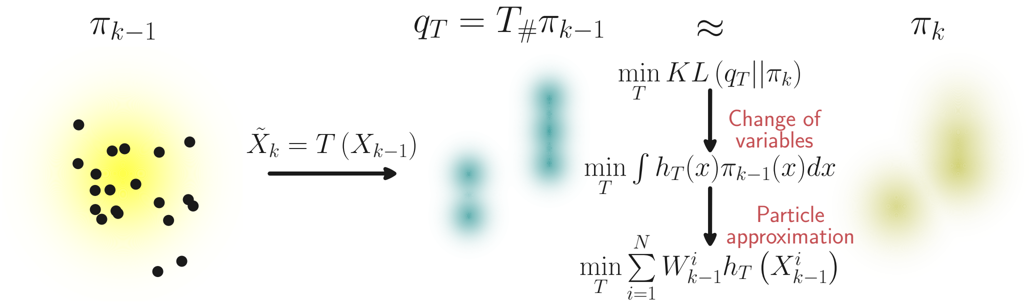
Learning the Normalizing Flows sequentially



- Change of variables: KL as an expectation under π_{k-1} of a function $h_T(x)$

$$h_T(x) = \log \pi_{k-1}(x) - \log \pi_k(T(x)) - \log |\nabla T(x)| + C$$

Learning the Normalizing Flows sequentially



- Change of variables: KL as an expectation under π_{k-1} of a function $h_T(x)$

$$h_T(x) = \log \pi_{k-1}(x) - \log \pi_k(T(x)) - \log |\nabla T(x)| + C$$

- Particle approximation: Use particles X_{k-1}^i and weights W_{k-1}^i to estimate expectation of h_T under π_{k-1}.

Theory: Consistency and Asymptotic Normality

- AFT produces estimates π_K^N and Z_K^N of π and Z using N particles X_K^i and weights W_K^i.
- Consistency:

$$\pi_K^N[f] \xrightarrow{N} \pi[f],$$
$$Z_K^N \xrightarrow{N} Z.$$

- Central Limit theorem:

$$\sqrt{N}\left(\pi_K^N[f] - \pi[f]\right) \xrightarrow{N} \mathcal{N}(0, V^\pi[f])$$
$$\sqrt{N}\left(Z_K^N - Z\right) \xrightarrow{N} \mathcal{N}(0, V^Z)$$

- Extends results of SMC algorithms, but proof involve tools from empirical process theory.
- Variance is optimal if the flows T_k exactly map π_{k-1} to π_k.

Scaling limit: Infinitely many auxiliary densities

- Setting:
  - Population limit: Infinitely many particles $N \to +\infty$
  - Unadjusted Langevin kernel for K_k.
  - Continuous-time limit: Infinitely many auxiliary densities $(\pi_k)_{k=1}^{K} \to (\pi_t)_{[0,1]}$.

Scaling limit: Infinitely many auxiliary densities

- Setting:
  - Population limit: Infinitely many particles $N \to +\infty$
  - Unadjusted Langevin kernel for K_k.
  - Continuous-time limit: Infinitely many auxiliary densities $(\pi_k)_{k=1}^K \to (\pi_t)_{[0,1]}$.
- AFT recovers a weighted controlled diffusion:
  - Sample paths $X_{0,t}$ follows a controlled SDE with control α_t:

$$dX_t = (\alpha_t^\star(X_t) + \nabla \log \pi_t(X_t))dt + \sqrt{2}dB_t$$

Scaling limit: Infinitely many auxiliary densities

- Setting:
  - Population limit: Infinitely many particles $N \to +\infty$
  - Unadjusted Langevin kernel for K_k.
  - Continuous-time limit: Infinitely many auxiliary densities $(\pi_k)_{k=1}^K \to (\pi_t)_{[0,1]}$.
- AFT recovers a weighted controlled diffusion:
  - Sample paths $X_{0,t}$ follows a controlled SDE with control α_t:

$$dX_t = (\alpha_t^\star(X_t) + \nabla \log \pi_t(X_t))dt + \sqrt{2}dB_t$$

  - Sample paths $X_{0,t}$ are re-weighted according to:

$$w_t^{\alpha^\star}(X_{[0,t]}) := \exp\left(\int_0^t g_s^{\alpha^\star}(X_s)ds\right), \qquad g_s^{\alpha}(X_s) := div_x(\alpha_t) + \alpha_t^\top \nabla_x \log \pi_t + \partial_t \log \pi_t$$

Scaling limit: Infinitely many auxiliary densities

- Setting:
  - Population limit: Infinitely many particles $N \to +\infty$
  - Unadjusted Langevin kernel for K_k.
  - Continuous-time limit: Infinitely many auxiliary densities $(\pi_k)_{k=1}^K \to (\pi_t)_{[0,1]}$.
- AFT recovers a weighted controlled diffusion:
  - Sample paths $X_{0,t}$ follows a controlled SDE with control α_t:

$$dX_t = (\alpha_t^\star(X_t) + \nabla \log \pi_t(X_t))dt + \sqrt{2}dB_t$$

  - Sample paths $X_{0,t}$ are re-weighted according to:

$$w_t^{\alpha^\star}(X_{[0,t]}) := \exp\left(\int_0^t g_s^{\alpha^\star}(X_s)ds\right), \qquad g_s^{\alpha}(X_s) := div_x(\alpha_t) + \alpha_t^\top \nabla_x \log \pi_t + \partial_t \log \pi_t$$

  - Weights ensure the marginals of weighted diffusion match π_t exactly.

Scaling limit: Infinitely many auxiliary densities

- Setting:
  - Population limit: Infinitely many particles $N \to +\infty$
  - Unadjusted Langevin kernel for K_k.
  - Continuous-time limit: Infinitely many auxiliary densities $(\pi_k)_{k=1}^K \to (\pi_t)_{[0,1]}$.
- AFT recovers a weighted controlled diffusion:
  - Sample paths $X_{0,t}$ follows a controlled SDE with control α_t:

$$dX_t = (\alpha_t^\star(X_t) + \nabla \log \pi_t(X_t))dt + \sqrt{2}dB_t$$

  - Sample paths $X_{0,t}$ are re-weighted according to:

$$w_t^{\alpha^\star}(X_{[0,t]}) := \exp\left(\int_0^t g_s^{\alpha^\star}(X_s)ds\right), \qquad g_s^\alpha(X_s) := div_x(\alpha_t) + \alpha_t^\top \nabla_x \log \pi_t + \partial_t \log \pi_t$$

  - Weights ensure the marginals of weighted diffusion match π_t exactly.
  - <u>Instantaneous work</u> g_s^α measures how much the density of X_t differs from π_t.

Scaling limit: Infinitely many auxiliary densities

- Setting:
  - Population limit: Infinitely many particles $N \to +\infty$
  - Unadjusted Langevin kernel for K_k.
  - Continuous-time limit: Infinitely many auxiliary densities $(\pi_k)_{k=1}^{K} \to (\pi_t)_{[0,1]}$.
- AFT recovers a weighted controlled diffusion:
  - Sample paths $X_{0,t}$ follows a controlled SDE with control α_t:

$$dX_t = (\alpha_t^\star(X_t) + \nabla \log \pi_t(X_t))dt + \sqrt{2}dB_t$$

  - Sample paths $X_{0,t}$ are re-weighted according to:

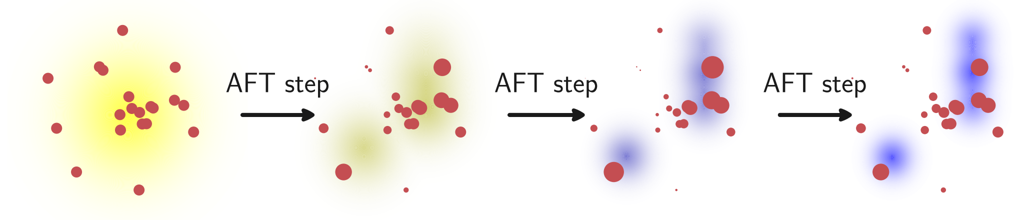
$$w_t^{\alpha^\star}(X_{[0,t]}) := \exp\left(\int_0^t g_s^{\alpha^\star}(X_s)ds\right), \qquad g_s^\alpha(X_s) := div_x(\alpha_t) + \alpha_t^\top \nabla_x \log \pi_t + \partial_t \log \pi_t$$

  - Weights ensure the marginals of weighted diffusion match π_t exactly.
  - Instantaneous work g_s^α measures how much the density of X_t differs from π_t.
  - Optimal control $\alpha^\star$ obtained by minimizing the variance of Instantaneous work:

$$\alpha^\star := \frac{1}{2} \arg\min_\alpha \int_0^1 dt \left(\pi_t[(g_t^\alpha)^2] - \pi_t[g_t^\alpha]^2\right).$$

Annealed Flow Transport

$\pi_0 = p$ $\quad$ $\pi_1 \propto p^{1-\beta_1}\pi^{\beta_1}$ $\quad$ $\pi_k \propto p^{1-\beta_k}\pi^{\beta_k}$ $\quad$ $\pi_K = \pi$



$\beta_0 = 0$ $\quad$ β_1 $\quad$ β_k $\quad$ $\beta_K = 1$

- AFT extends SMC to take advantage of Normalizing flows.
- Known asymptotic behavior
- Known scaling limit