LECTURE 5

MULTI-MODAL DISTRIBUTIONS

Saifuddin Syed

TUNING MCMC

TUNING MCMC

- Given an MCMC algorithm there are often hyperparameters $\theta \in \Theta$

TUNING MCMC

- Given an MCMC algorithm there are often hyperparameters $\theta \in \Theta$
  - e.g. step size, mass/covariance matrix, integration time, etc

TUNING MCMC

- Given an MCMC algorithm there are often hyperparameters $\theta \in \Theta$
  - e.g. step size, mass/covariance matrix, integration time, etc
- This defines a family of π-invariant kernels K_θ such that give a valid MCMC algorithm

TUNING MCMC

- Given an MCMC algorithm there are often hyperparameters $\theta \in \Theta$
  - e.g. step size, mass/covariance matrix, integration time, etc
- This defines a family of π-invariant kernels K_θ such that give a valid MCMC algorithm
  - For each θ we can generate a Makrov chain X_t that satisfies the ergodic theorem

$$X_t \sim K_\theta(X_{t-1}, \mathrm{d}x) \qquad \frac{1}{T}\sum_{t=1}^{T} f(X_t) \xrightarrow[T\to\infty]{a.s} \pi[f]$$

TUNING MCMC

- Given an MCMC algorithm there are often hyperparameters $\theta \in \Theta$
  - e.g. step size, mass/covariance matrix, integration time, etc
- This defines a family of π-invariant kernels K_θ such that give a valid MCMC algorithm
  - For each θ we can generate a Makrov chain X_t that satisfies the ergodic theorem

$$X_t \sim K_\theta(X_{t-1}, \mathrm{d}x) \qquad \frac{1}{T}\sum_{t=1}^{T} f(X_t) \xrightarrow[T\to\infty]{a.s} \pi[f]$$

- The efficiency and rates of convergence depend on the paramers and must be turned

TUNING MCMC

- Given an MCMC algorithm there are often hyperparameters $\theta \in \Theta$
  - e.g. step size, mass/covariance matrix, integration time, etc
- This defines a family of π-invariant kernels K_θ such that give a valid MCMC algorithm
  - For each θ we can generate a Makrov chain X_t that satisfies the ergodic theorem

$$X_t \sim K_\theta(X_{t-1}, \mathrm{d}x) \qquad \frac{1}{T}\sum_{t=1}^{T} f(X_t) \xrightarrow[T\to\infty]{a.s} \pi[f]$$

- The efficiency and rates of convergence depend on the paramers and must be turned
- In practice, we do this adaptively and modify the parameters as we simulate the chain

TUNING MCMC

- Given an MCMC algorithm there are often hyperparameters $\theta \in \Theta$
  - e.g. step size, mass/covariance matrix, integration time, etc
- This defines a family of π-invariant kernels K_θ such that give a valid MCMC algorithm
  - For each θ we can generate a Makrov chain X_t that satisfies the ergodic theorem

$$X_t \sim K_\theta(X_{t-1}, \mathrm{d}x) \qquad \frac{1}{T}\sum_{t=1}^{T} f(X_t) \xrightarrow[T\to\infty]{a.s} \pi[f]$$

- The efficiency and rates of convergence depend on the paramers and must be turned
- In practice, we do this adaptively and modify the parameters as we simulate the chain
  - E.g. modify the parameters until a desired step size is reached

TUNING MCMC

- Given an MCMC algorithm there are often hyperparameters $\theta \in \Theta$
  - e.g. step size, mass/covariance matrix, integration time, etc
- This defines a family of π-invariant kernels K_θ such that give a valid MCMC algorithm
  - For each θ we can generate a Makrov chain X_t that satisfies the ergodic theorem

$$X_t \sim K_\theta(X_{t-1}, \mathrm{d}x) \qquad \frac{1}{T}\sum_{t=1}^{T} f(X_t) \xrightarrow[T\to\infty]{a.s} \pi[f]$$

- The efficiency and rates of convergence depend on the paramers and must be turned
- In practice, we do this adaptively and modify the parameters as we simulate the chain
  - E.g. modify the parameters until a desired step size is reached
- **Question:** Is this theoretically justified?

PITFALLS OF ADAPTIVE TUNING

PITFALLS OF ADAPTIVE TUNING

- Formally, we develop a θ_t that depends on the history of $X_1, \ldots, X_t$

$$X_t \sim K_{\theta_t}(X_{t-1}, \mathrm{d}x)$$

PITFALLS OF ADAPTIVE TUNING

- Formally, we develop a θ_t that depends on the history of $X_1, \ldots, X_t$

$$X_t \sim K_{\theta_t}(X_{t-1}, \mathrm{d}x)$$

- X_t is in general no longer a Markov chain, π-invariant or satisfies the ergodic theorem

PITFALLS OF ADAPTIVE TUNING

- Formally, we develop a θ_t that depends on the history of $X_1, \ldots, X_t$

$$X_t \sim K_{\theta_t}(X_{t-1}, \mathrm{d}x)$$

- X_t is in general no longer a Markov chain, π-invariant or satisfies the ergodic theorem
  - Intuitively, we can think of $\theta_t(x_{0:t})$ as a control function governing a dynamical system X_t

PITFALLS OF ADAPTIVE TUNING

- Formally, we develop a θ_t that depends on the history of $X_1, \ldots, X_t$

$$X_t \sim K_{\theta_t}(X_{t-1}, \mathrm{d}x)$$

- X_t is in general no longer a Markov chain, π-invariant or satisfies the ergodic theorem
  - Intuitively, we can think of $\theta_t(x_{0:t})$ as a control function governing a dynamical system X_t
- If not careful, we can nudge the law of the chain towards non-equilibrium states

EXAMPLE: ADAPTIVE MARKOV CHAIN

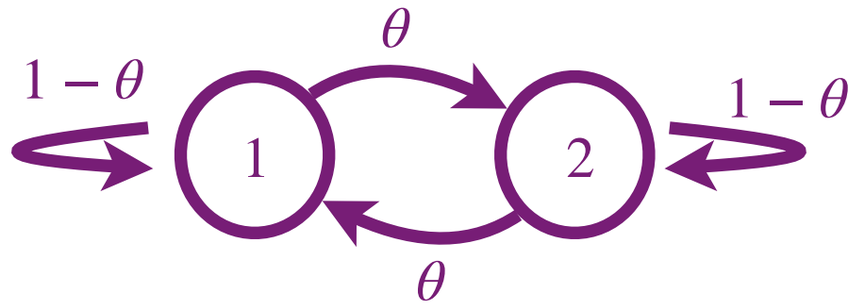
EXAMPLE: ADAPTIVE MARKOV CHAIN

- Suppose $\mathbb{X} = \{1,2\}$ and $\Theta = (0,1)$

EXAMPLE: ADAPTIVE MARKOV CHAIN

- Suppose $\mathbb{X} = \{1,2\}$ and $\Theta = (0,1)$
- Consider the family of Markov kernels K_θ parametrised by $\theta \in \Theta$

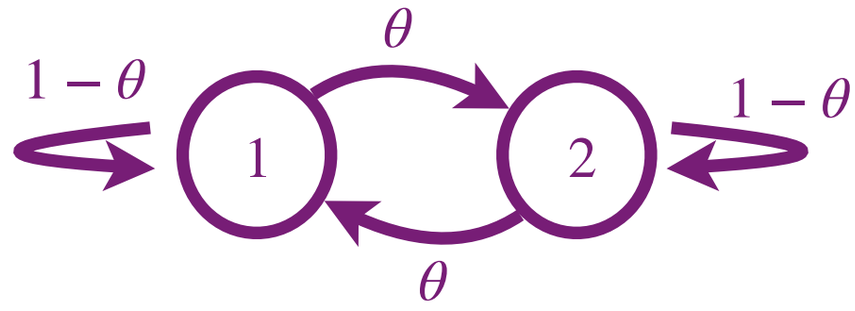
$$K_\theta = \begin{pmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}$$



EXAMPLE: ADAPTIVE MARKOV CHAIN

- Suppose $\mathbb{X} = \{1,2\}$ and $\Theta = (0,1)$
- Consider the family of Markov kernels K_θ parametrised by $\theta \in \Theta$

$$K_\theta = \begin{pmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}$$



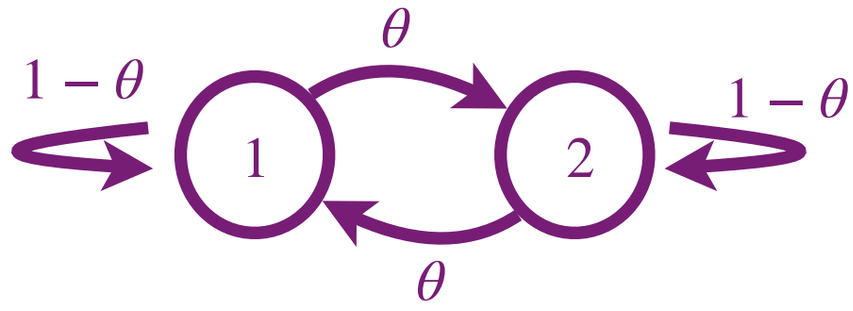
  - For each θ, the chain X_t generated by K_θ is invariant to $\pi(1) = \pi(2) = 0.5$

$\pi K_\theta = \pi$ $\qquad\qquad$ $X_t \sim K_\theta(X_{t-1}, \mathrm{d}x)$

EXAMPLE: ADAPTIVE MARKOV CHAIN

- Suppose $\mathbb{X} = \{1,2\}$ and $\Theta = (0,1)$
- Consider the family of Markov kernels K_θ parametrised by $\theta \in \Theta$

$$K_\theta = \begin{pmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}$$



  - For each θ, the chain X_t generated by K_θ is invariant to $\pi(1) = \pi(2) = 0.5$

$$\pi K_\theta = \pi \qquad\qquad X_t \sim K_\theta(X_{t-1}, \mathrm{d}x)$$

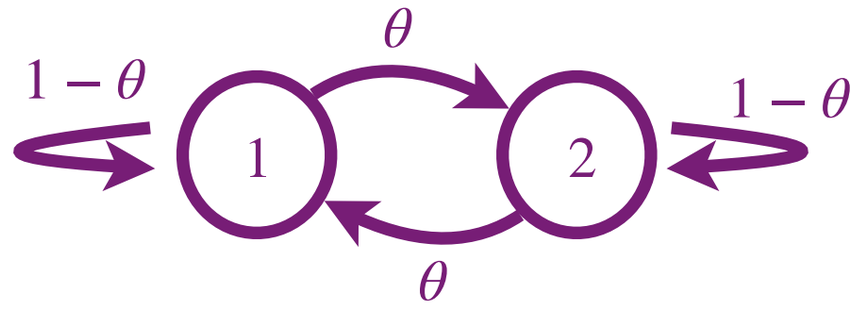
- For $\theta : \mathbb{X} \to [0,1]$ suppose we generate $\hat{X}_t$ adaptively using the kernel at $\theta(\hat{X}_t)$

$$\hat{X}_t \sim K_{\theta(\hat{X}_{t-1})}(\hat{X}_{t-1}, \mathrm{d}x) = \hat{K}(\hat{X}_{t-1}, \mathrm{d}x)$$

EXAMPLE: ADAPTIVE MARKOV CHAIN

- Suppose $\mathbb{X} = \{1,2\}$ and $\Theta = (0,1)$
- Consider the family of Markov kernels K_θ parametrised by $\theta \in \Theta$

$$K_\theta = \begin{pmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}$$



  - For each θ, the chain X_t generated by K_θ is invariant to $\pi(1) = \pi(2) = 0.5$

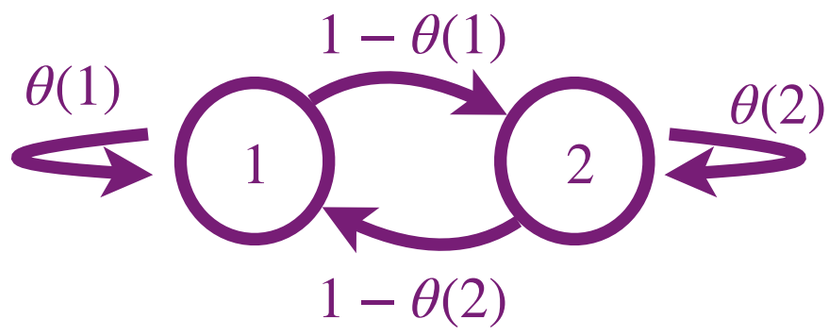
$$\pi K_\theta = \pi \qquad X_t \sim K_\theta(X_{t-1}, \mathrm{d}x)$$

- For $\theta : \mathbb{X} \to [0,1]$ suppose we generate $\hat{X}_t$ adaptively using the kernel at $\theta(\hat{X}_t)$

$$\hat{X}_t \sim K_{\theta(\hat{X}_{t-1})}(\hat{X}_{t-1}, \mathrm{d}x) = \hat{K}(\hat{X}_{t-1}, \mathrm{d}x)$$

- Defines a new Markov chain:

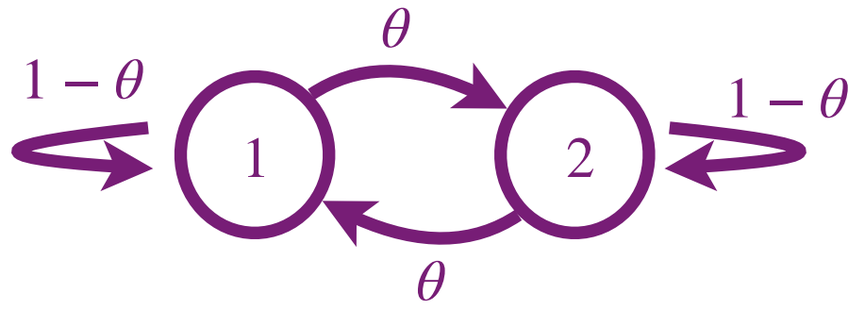
$$\hat{K} = \begin{pmatrix} \theta(1) & 1-\theta(1) \\ 1-\theta(2) & \theta(2) \end{pmatrix}$$



EXAMPLE: ADAPTIVE MARKOV CHAIN

- Suppose $\mathbb{X} = \{1,2\}$ and $\Theta = (0,1)$
- Consider the family of Markov kernels K_θ parametrised by $\theta \in \Theta$

$$K_\theta = \begin{pmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}$$



  - For each θ, the chain X_t generated by K_θ is invariant to $\pi(1) = \pi(2) = 0.5$

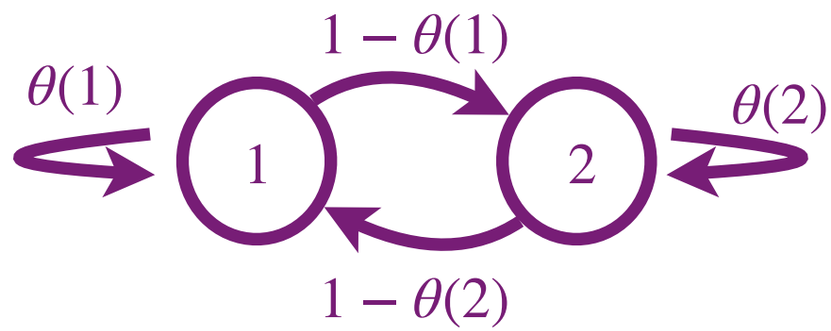
$$\pi K_\theta = \pi \qquad X_t \sim K_\theta(X_{t-1}, \mathrm{d}x)$$

- For $\theta : \mathbb{X} \to [0,1]$ suppose we generate $\hat{X}_t$ adaptively using the kernel at $\theta(\hat{X}_t)$

$$\hat{X}_t \sim K_{\theta(\hat{X}_{t-1})}(\hat{X}_{t-1}, \mathrm{d}x) = \hat{K}(\hat{X}_{t-1}, \mathrm{d}x)$$

- Defines a new Markov chain:

$$\hat{K} = \begin{pmatrix} \theta(1) & 1-\theta(1) \\ 1-\theta(2) & \theta(2) \end{pmatrix}$$



  - The adative chain $\hat{X}_t$ is **not** π-invariant, i.e $\pi K \neq \pi$

PITFALLS OF ADAPTIVE TUNING

PITFALLS OF ADAPTIVE TUNING

- **Problem:** Frequent adaptation never allows enough time for the chain to converge the target

PITFALLS OF ADAPTIVE TUNING

- **Problem:** Frequent adaptation never allows enough time for the chain to converge the target
  - Doing this correctly is delicate to ensure equilibrium is reached

PITFALLS OF ADAPTIVE TUNING

- **Problem:** Frequent adaptation never allows enough time for the chain to converge the target
  - Doing this correctly is delicate to ensure equilibrium is reached
- Generally want to small or infrequent modification to θ

PITFALLS OF ADAPTIVE TUNING

- **Problem:** Frequent adaptation never allows enough time for the chain to converge the target
  - Doing this correctly is delicate to ensure equilibrium is reached
- Generally want to small or infrequent modification to θ
  - Can keep consistency by making **small** modification at each time $\theta_t \approx \theta_{t-1}$

PITFALLS OF ADAPTIVE TUNING

- **Problem:** Frequent adaptation never allows enough time for the chain to converge the target
  - Doing this correctly is delicate to ensure equilibrium is reached
- Generally want to small or infrequent modification to θ
  - Can keep consistency by making **small** modification at each time $\theta_t \approx \theta_{t-1}$
  - Or by making infrequent modifications, for example exponential time between updates

PITFALLS OF ADAPTIVE TUNING

- **Problem:** Frequent adaptation never allows enough time for the chain to converge the target
  - Doing this correctly is delicate to ensure equilibrium is reached
- Generally want to small or infrequent modification to θ
  - Can keep consistency by making **small** modification at each time $\theta_t \approx \theta_{t-1}$
  - Or by making infrequent modifications, for example exponential time between updates
- In practice, we devote a portion of the compute budget for tuning:

PITFALLS OF ADAPTIVE TUNING

- **Problem:** Frequent adaptation never allows enough time for the chain to converge the target
  - Doing this correctly is delicate to ensure equilibrium is reached

- Generally want to small or infrequent modification to θ
  - Can keep consistency by making **small** modification at each time $\theta_t \approx \theta_{t-1}$
  - Or by making infrequent modifications, for example exponential time between updates

- In practice, we devote a portion of the compute budget for tuning:
  - The goal is to learn a learn an optimal hyper paramaters $\theta^\star$ for the sampler

PITFALLS OF ADAPTIVE TUNING

- **Problem:** Frequent adaptation never allows enough time for the chain to converge the target
  - Doing this correctly is delicate to ensure equilibrium is reached
- Generally want to small or infrequent modification to θ
  - Can keep consistency by making **small** modification at each time $\theta_t \approx \theta_{t-1}$
  - Or by making infrequent modifications, for example exponential time between updates
- In practice, we devote a portion of the compute budget for tuning:
  - The goal is to learn a learn an optimal hyper paramaters $\theta^\star$ for the sampler
  - Throw away the samples at the end then use the remaining budget to sample with $\theta^\star$

LOCAL SAMPLERS

LOCAL SAMPLERS

- Most MCMC methods are traditionally designed to be **local**

LOCAL SAMPLERS

- Most MCMC methods are traditionally designed to be **local**
- We often assume the target distribution is continuous in the topological features of $\mathbb{X}$

$$y \approx x \qquad \Longrightarrow \qquad \pi(y) \approx \pi(x)$$

LOCAL SAMPLERS

- Most MCMC methods are traditionally designed to be **local**
- We often assume the target distribution is continuous in the topological features of $\mathbb{X}$

$$y \approx x \qquad \Longrightarrow \qquad \pi(y) \approx \pi(x)$$

- The proposal Q appeals to the topology of the underlying statepace

LOCAL SAMPLERS

- Most MCMC methods are traditionally designed to be **local**
- We often assume the target distribution is continuous in the topological features of $\mathbb{X}$

$$y \approx x \qquad \Longrightarrow \qquad \pi(y) \approx \pi(x)$$

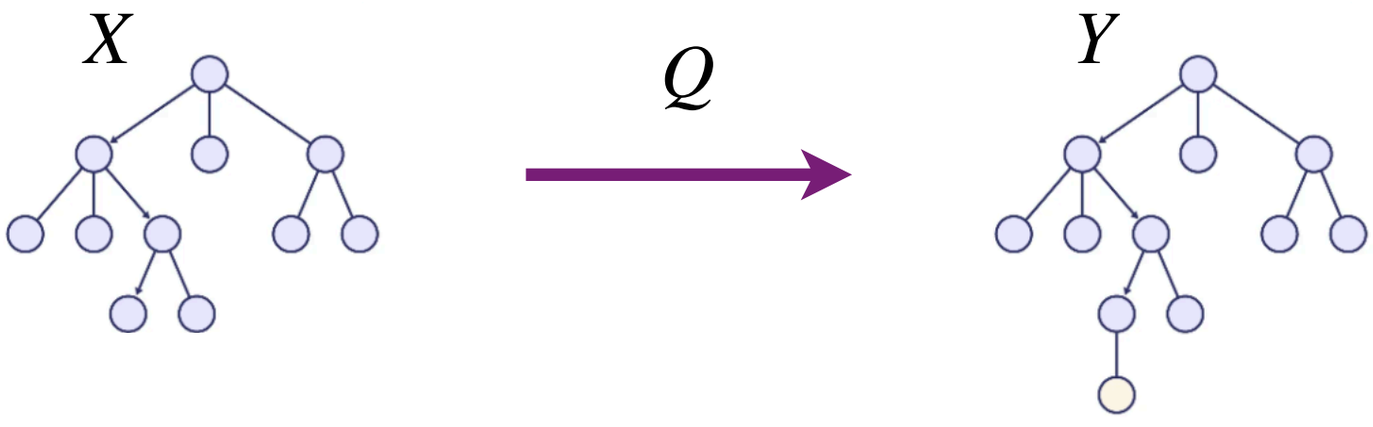
- The proposal Q appeals to the topology of the underlying statepace
  - Given X proposal samples $Y \sim Q(X, \mathrm{d}y)$ within a neighbourhood of X

LOCAL SAMPLERS

- Most MCMC methods are traditionally designed to be **local**
- We often assume the target distribution is continuous in the topological features of $\mathbb{X}$

$$y \approx x \qquad \Longrightarrow \qquad \pi(y) \approx \pi(x)$$

- The proposal Q appeals to the topology of the underlying statepace
  - Given X proposal samples $Y \sim Q(X, \mathrm{d}y)$ within a neighbourhood of X

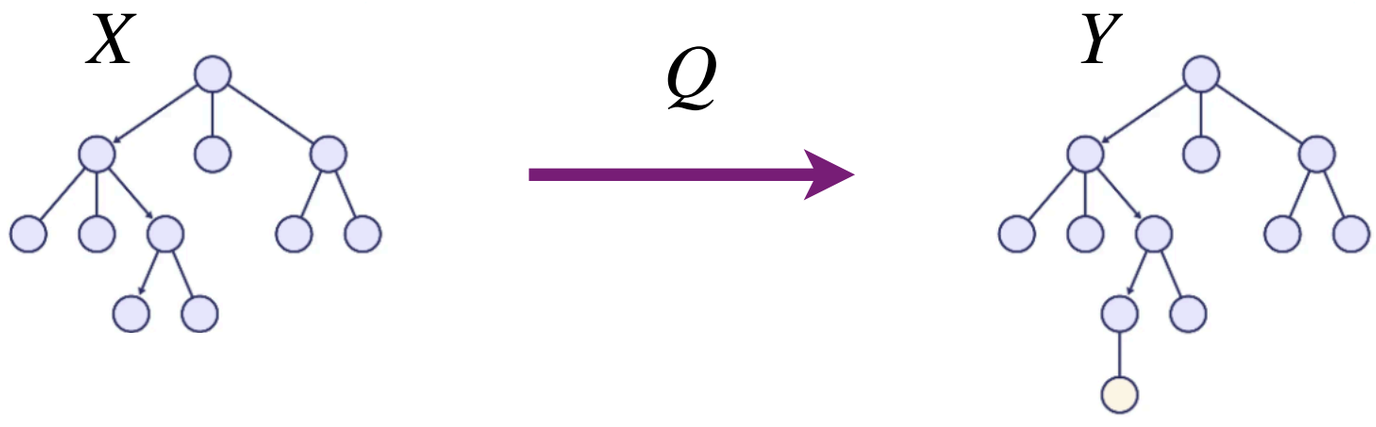


LOCAL SAMPLERS

- Most MCMC methods are traditionally designed to be **local**
- We often assume the target distribution is continuous in the topological features of $\mathbb{X}$

$$y \approx x \qquad \Longrightarrow \qquad \pi(y) \approx \pi(x)$$

- The proposal Q appeals to the topology of the underlying statepace
  - Given X proposal samples $Y \sim Q(X, \mathrm{d}y)$ within a neighbourhood of X

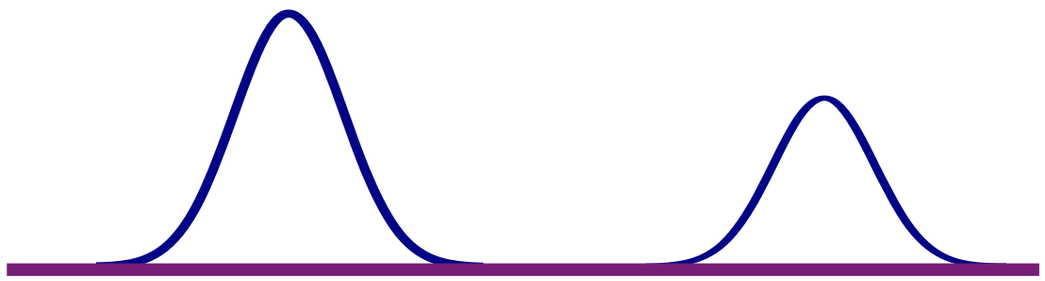


- Often appeals to the dynamics driven by a differential equation such as Langevin or Hamiltionian

REDUCIBILITY OF LOCAL SAMPLERS

REDUCIBILITY OF LOCAL SAMPLERS

- Local samplers prevent the chain from escaping local mode

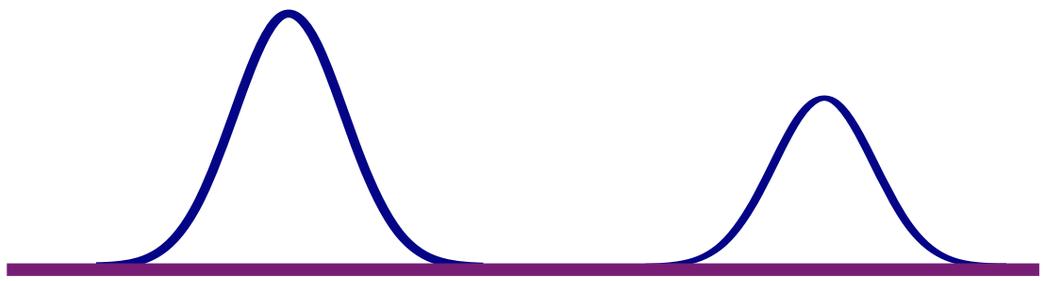


REDUCIBILITY OF LOCAL SAMPLERS

- Local samplers prevent the chain from escaping local mode
- E.g. the continuous Langevin dynamics and Hamiltonian dynamics are not irreducible!

$$\mathrm{d}Y_\tau = -\nabla \log \pi(Y_\tau)\mathrm{d}\tau + \sqrt{2}\mathrm{d}W_\tau$$

$$\mathrm{d}q_t = p_t \mathrm{d}t \qquad \mathrm{d}p_t = \log \nabla \pi(q_t)\mathrm{d}t$$



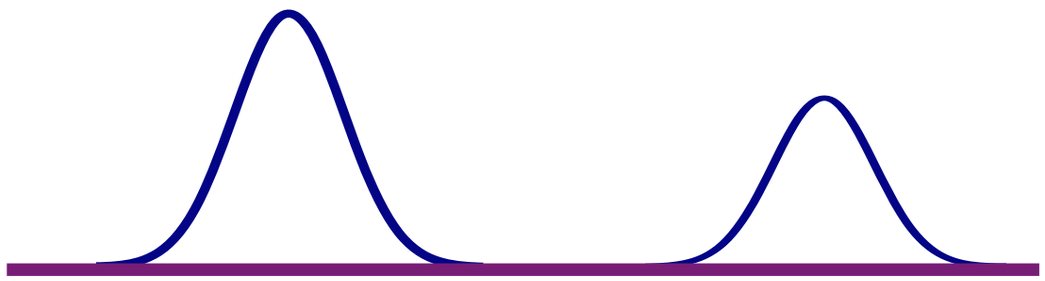
REDUCIBILITY OF LOCAL SAMPLERS

- Local samplers prevent the chain from escaping local mode
- E.g. the continuous Langevin dynamics and Hamiltonian dynamics are not irreducible!

$$\mathrm{d}Y_\tau = -\nabla \log \pi(Y_\tau)\mathrm{d}\tau + \sqrt{2}\mathrm{d}W_\tau$$

$$\mathrm{d}q_t = p_t \mathrm{d}t \qquad \mathrm{d}p_t = \log \nabla \pi(q_t)\mathrm{d}t$$

- A local samplers are generally not irreducible if there exists there exists an open set $U \subset \mathbb{X}$ such that $\pi[\partial U] = 0$ and $\pi[U] < 1$



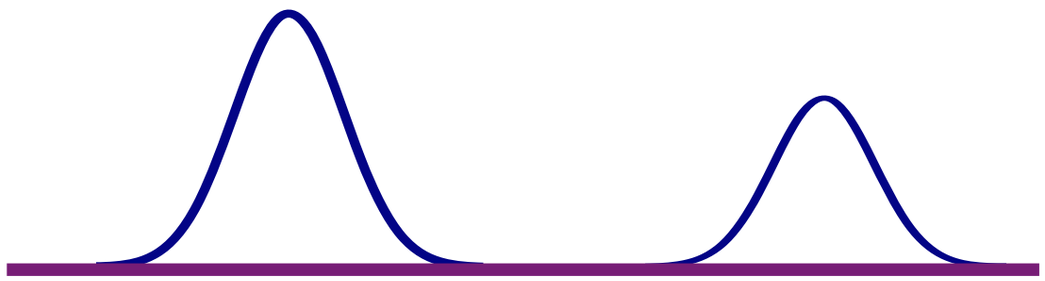
REDUCIBILITY OF LOCAL SAMPLERS

- Local samplers prevent the chain from escaping local mode
- E.g. the continuous Langevin dynamics and Hamiltonian dynamics are not irreducible!

$$\mathrm{d}Y_\tau = -\nabla \log \pi(Y_\tau)\mathrm{d}\tau + \sqrt{2}\mathrm{d}W_\tau$$

$$\mathrm{d}q_t = p_t \mathrm{d}t \qquad \mathrm{d}p_t = \log \nabla \pi(q_t)\mathrm{d}t$$

- A local samplers are generally not irreducible if there exists there exists an open set $U \subset \mathbb{X}$ such that $\pi[\partial U] = 0$ and $\pi[U] < 1$
- In theory, if $\pi(x) > 0$, but $\pi(x) \approx 0$ we can techincally have chains are uni-modal



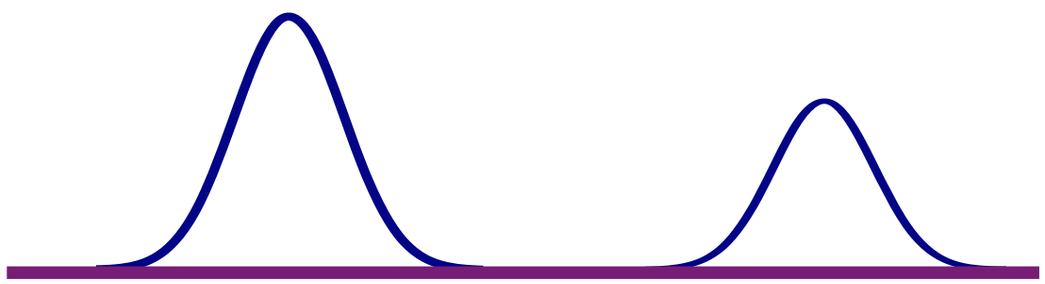
REDUCIBILITY OF LOCAL SAMPLERS

- Local samplers prevent the chain from escaping local mode
- E.g. the continuous Langevin dynamics and Hamiltonian dynamics are not irreducible!

$$\mathrm{d}Y_\tau = -\nabla \log \pi(Y_\tau)\mathrm{d}\tau + \sqrt{2}\mathrm{d}W_\tau$$

$$\mathrm{d}q_t = p_t \mathrm{d}t \qquad \mathrm{d}p_t = \log \nabla \pi(q_t)\mathrm{d}t$$

- A local samplers are generally not irreducible if there exists there exists an open set $U \subset \mathbb{X}$ such that $\pi[\partial U] = 0$ and $\pi[U] < 1$
- In theory, if $\pi(x) > 0$, but $\pi(x) \approx 0$ we can techincally have chains are uni-modal
- In practice, they are *effectively* irreducible



WHAT IS A MODE

WHAT IS A MODE

- What is a mode?

WHAT IS A MODE

- What is a mode?
- Surprisingly ill-defined notion

WHAT IS A MODE

- What is a mode?
- Surprisingly ill-defined notion
  - Intutively its a region of concentrated high probability

WHAT IS A MODE

- What is a mode?
- Surprisingly ill-defined notion
  - Intutively its a region of concentrated high probability
  - local optimum of the target

WHAT IS A MODE

- What is a mode?
- Surprisingly ill-defined notion
  - Intutively its a region of concentrated high probability
  - local optimum of the target
- **Question:** How many modes does this have:

WHAT IS A MODE

- What is a mode?
- Surprisingly ill-defined notion
  - Intutively its a region of concentrated high probability
  - local optimum of the target
- **Question:** How many modes does this have:

- We want a definition that is philosophically aligned with intuitions

WHAT IS A MODE

- What is a mode?
- Surprisingly ill-defined notion
  - Intutively its a region of concentrated high probability
  - local optimum of the target
- **Question:** How many modes does this have:

- We want a definition that is philosophically aligned with intuitions
  - Should not overfit to not assuming extraneous structure of a specifc problem

WHAT IS A MODE

- What is a mode?
- Surprisingly ill-defined notion
  - Intutively its a region of concentrated high probability
  - local optimum of the target
- **Question:** How many modes does this have:

- We want a definition that is philosophically aligned with intuitions
  - Should not overfit to not assuming extraneous structure of a specifc problem
  - e.g. statespace, topology, geometry, smoothnes etc

WHAT IS A MODE

- What is a mode?
- Surprisingly ill-defined notion
  - Intutively its a region of concentrated high probability
  - local optimum of the target
- **Question:** How many modes does this have:

- We want a definition that is philosophically aligned with intuitions
  - Should not overfit to not assuming extraneous structure of a specifc problem
  - e.g. statespace, topology, geometry, smoothnes etc
- Modes are a property of the sampler!

WHAT IS A MODE

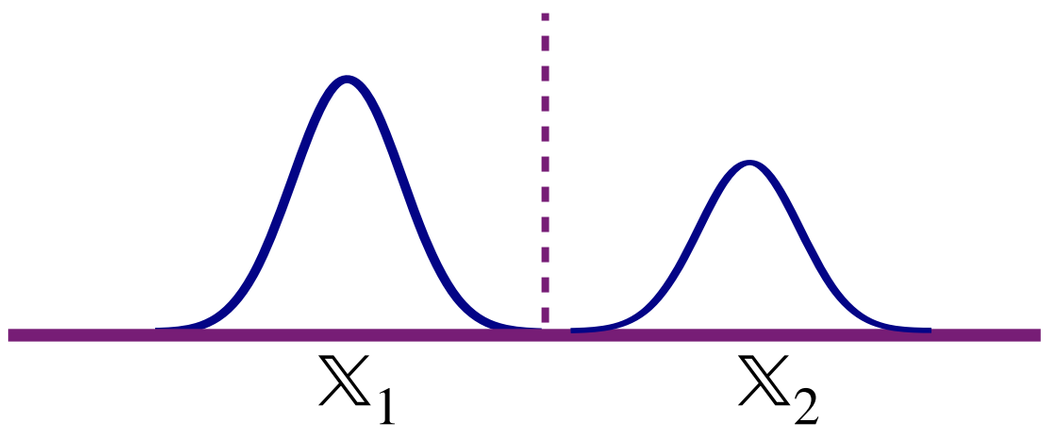
- What is a mode?
- Surprisingly ill-defined notion
  - Intutively its a region of concentrated high probability
  - local optimum of the target
- **Question:** How many modes does this have:

- We want a definition that is philosophically aligned with intuitions
  - Should not overfit to not assuming extraneous structure of a specifc problem
  - e.g. statespace, topology, geometry, smoothnes etc
- Modes are a property of the sampler!
  - π is unimodal if we can efficiently move from one point to another in the support of π

MODES

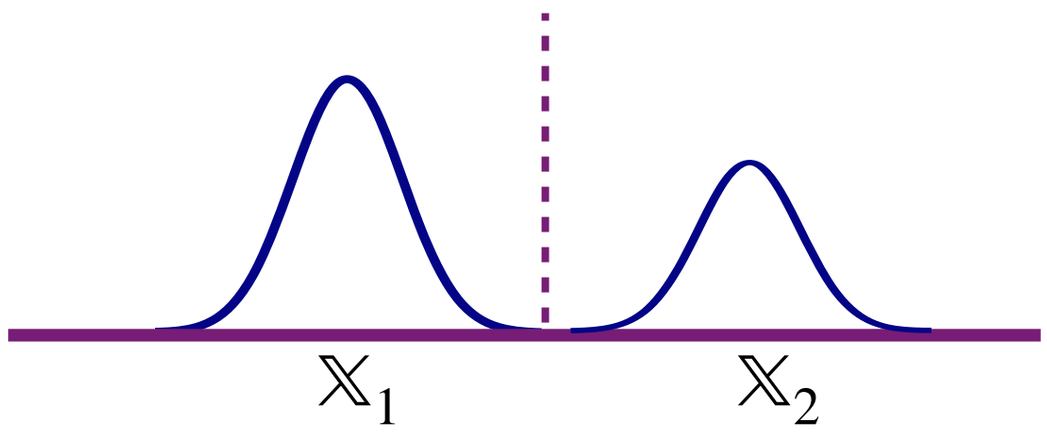
MODES

- Let K be a π-invariant kernel and $x, x' \in \text{supp}(\pi) = \mathbb{X}$, we will say $x \sim_K x'$ if



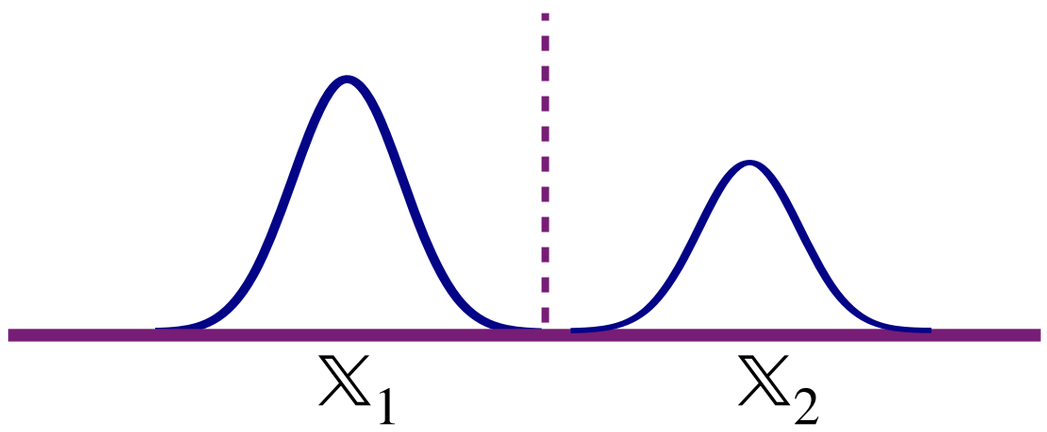
MODES

- Let K be a π-invariant kernel and $x, x' \in \text{supp}(\pi) = \mathbb{X}$, we will say $x \sim_K x'$ if
  - It there is a possible probability to reach x' from x using K and vice-versa



MODES

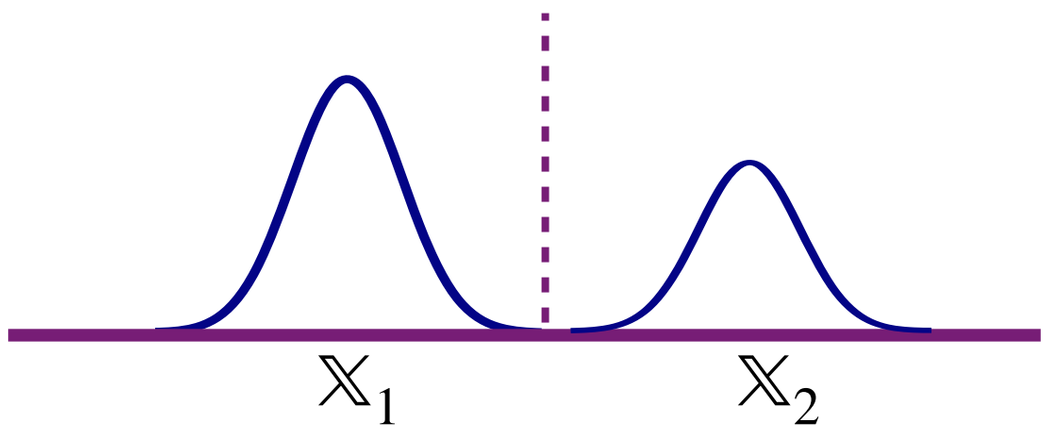
- Let K be a π-invariant kernel and $x, x' \in \text{supp}(\pi) = \mathbb{X}$, we will say $x \sim_K x'$ if
  - It there is a possible probability to reach x' from x using K and vice-versa
  - This defines an equivalent relation on the support of π



MODES

- Let K be a π-invariant kernel and $x, x' \in \text{supp}(\pi) = \mathbb{X}$, we will say $x \sim_K x'$ if
  - It there is a possible probability to reach x' from x using K and vice-versa
  - This defines an equivalent relation on the support of π
- We will say the $\mathbb{X}(x)$ is the mode of x

$$\mathbb{X}(x) = \{x' \in \mathbb{X} : x \sim_K x'\}$$

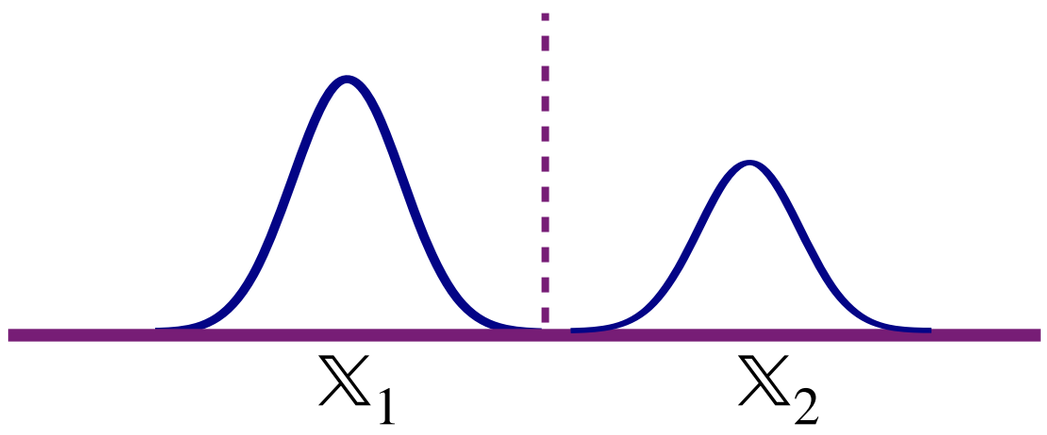


MODES

- Let K be a π-invariant kernel and $x, x' \in \text{supp}(\pi) = \mathbb{X}$, we will say $x \sim_K x'$ if
  - It there is a possible probability to reach x' from x using K and vice-versa
  - This defines an equivalent relation on the support of π
- We will say the $\mathbb{X}(x)$ is the mode of x

$$\mathbb{X}(x) = \{x' \in \mathbb{X} : x \sim_K x'\}$$

  - I.e. the modes correspond to the irreducible components of the kernel

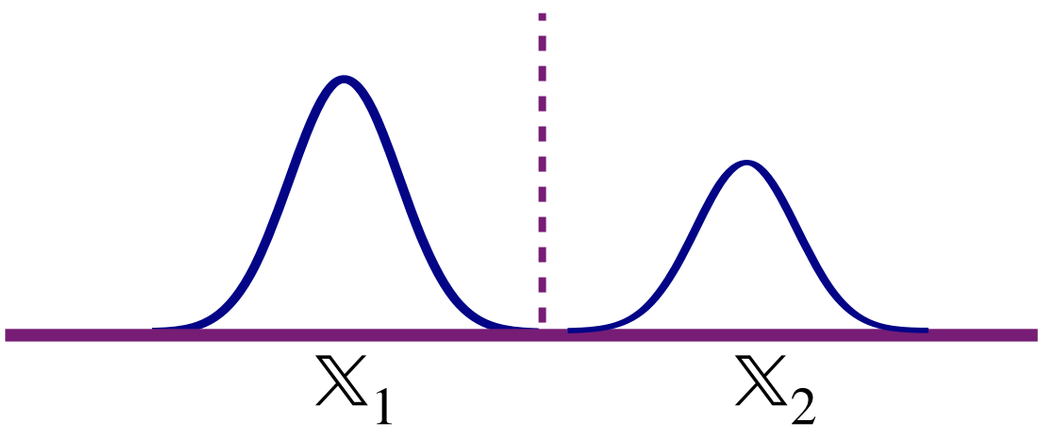


MODES

- Let K be a π-invariant kernel and $x, x' \in \text{supp}(\pi) = \mathbb{X}$, we will say $x \sim_K x'$ if
  - It there is a possible probability to reach x' from x using K and vice-versa
  - This defines an equivalent relation on the support of π
- We will say the $\mathbb{X}(x)$ is the mode of x

$$\mathbb{X}(x) = \{x' \in \mathbb{X} : x \sim_K x'\}$$

  - I.e. the modes correspond to the irreducible components of the kernel
- The number of modes $|\{\mathbb{X}(x) : x \in \mathbb{X}\}|$

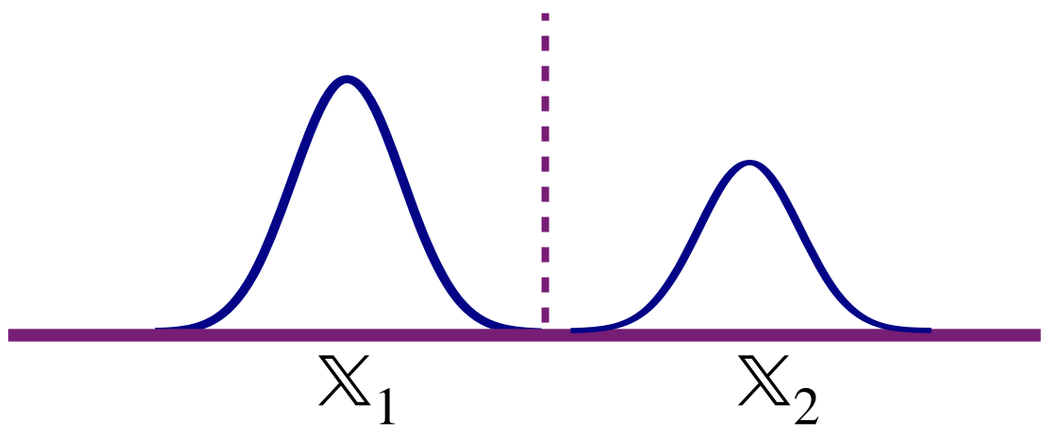


MODES

- Let K be a π-invariant kernel and $x, x' \in \text{supp}(\pi) = \mathbb{X}$, we will say $x \sim_K x'$ if
  - It there is a possible probability to reach x' from x using K and vice-versa
  - This defines an equivalent relation on the support of π
- We will say the $\mathbb{X}(x)$ is the mode of x

$$\mathbb{X}(x) = \{x' \in \mathbb{X} : x \sim_K x'\}$$

  - I.e. the modes correspond to the irreducible components of the kernel
    - The number of modes $|\{\mathbb{X}(x) : x \in \mathbb{X}\}|$
- A distribution is uni-modal iff and only if K is π-irreducible

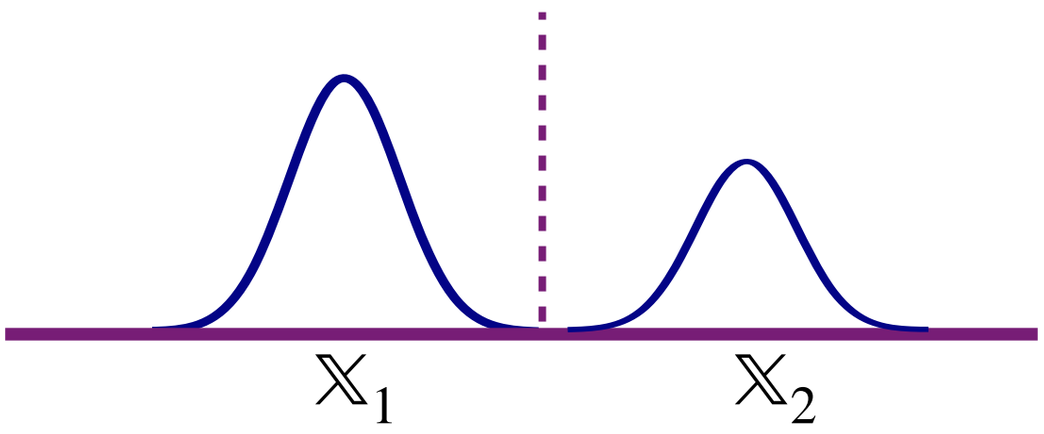


MODES

- Let K be a π-invariant kernel and $x, x' \in \text{supp}(\pi) = \mathbb{X}$, we will say $x \sim_K x'$ if
  - It there is a possible probability to reach x' from x using K and vice-versa
  - This defines an equivalent relation on the support of π
- We will say the $\mathbb{X}(x)$ is the mode of x

$$\mathbb{X}(x) = \{x' \in \mathbb{X} : x \sim_K x'\}$$

  - I.e. the modes correspond to the irreducible components of the kernel
  - The number of modes $|\{\mathbb{X}(x) : x \in \mathbb{X}\}|$
- A distribution is uni-modal iff and only if K is π-irreducible
- Notably π is uni-modal with respect to the independent kernel $K(x, \mathrm{d}x') = \pi(\mathrm{d}x')$



MULTI-MODAL DECOMPOSITION

MULTI-MODAL DECOMPOSITION

- Let $\mathbb{X}_i$ be the modes of the target $\boldsymbol{\pi}$ with respect to $\boldsymbol{K}$

MULTI-MODAL DECOMPOSITION

- Let $\mathbb{X}_i$ be the modes of the target π with respect to K
  - $\mathbb{X}_i \cap \mathbb{X}_j = 0$ for $i \neq j$ and $\mathbb{X} = \cup_i \mathbb{X}_i$

MULTI-MODAL DECOMPOSITION

- Let $\mathbb{X}_i$ be the modes of the target π with respect to K
  - $\mathbb{X}_i \cap \mathbb{X}_j = 0$ for $i \neq j$ and $\mathbb{X} = \cup_i \mathbb{X}_i$
- We have the following decomposition:

MULTI-MODAL DECOMPOSITION

- Let $\mathbb{X}_i$ be the modes of the target π with respect to K
  - $\mathbb{X}_i \cap \mathbb{X}_j = 0$ for $i \neq j$ and $\mathbb{X} = \cup_i \mathbb{X}_i$
- We have the following decomposition:

$$\pi(x) = \sum_i \pi[\mathbb{X}_i]\pi[x \,|\, \mathbb{X}_i]$$

MULTI-MODAL DECOMPOSITION

- Let $\mathbb{X}_i$ be the modes of the target π with respect to K
  - $\mathbb{X}_i \cap \mathbb{X}_j = 0$ for $i \neq j$ and $\mathbb{X} = \cup_i \mathbb{X}_i$
- We have the following decomposition:

$$\pi(x) = \sum_i \pi[\mathbb{X}_i]\pi[x \,|\, \mathbb{X}_i] \quad = \sum_i w_i \pi_i(x)$$

MULTI-MODAL DECOMPOSITION

- Let $\mathbb{X}_i$ be the modes of the target π with respect to K
  - $\mathbb{X}_i \cap \mathbb{X}_j = 0$ for $i \neq j$ and $\mathbb{X} = \cup_i \mathbb{X}_i$
- We have the following decomposition:

$$\pi(x) = \sum_i \pi[\mathbb{X}_i]\pi[x \mid \mathbb{X}_i] \quad = \sum_i w_i \pi_i(x)$$

- Where w_i is the weight of mode i and π_i is π conditional on the mode $\mathbb{X}_i$

MULTI-MODAL DECOMPOSITION

- Let $\mathbb{X}_i$ be the modes of the target π with respect to K
  - $\mathbb{X}_i \cap \mathbb{X}_j = 0$ for $i \neq j$ and $\mathbb{X} = \cup_i \mathbb{X}_i$
- We have the following decomposition:

$$\pi(x) = \sum_i \pi[\mathbb{X}_i]\pi[x \mid \mathbb{X}_i] \quad = \sum_i w_i \pi_i(x)$$

- Where w_i is the weight of mode i and π_i is π conditional on the mode $\mathbb{X}_i$

$$w_i = \pi[\mathbb{X}_i]$$

MULTI-MODAL DECOMPOSITION

- Let $\mathbb{X}_i$ be the modes of the target π with respect to K
  - $\mathbb{X}_i \cap \mathbb{X}_j = 0$ for $i \neq j$ and $\mathbb{X} = \cup_i \mathbb{X}_i$
- We have the following decomposition:

$$\pi(x) = \sum_i \pi[\mathbb{X}_i]\pi[x \mid \mathbb{X}_i] \quad = \sum_i w_i \pi_i(x)$$

- Where w_i is the weight of mode i and π_i is π conditional on the mode $\mathbb{X}_i$

$$w_i = \pi[\mathbb{X}_i] \qquad \pi_i(x) = \pi[x \mid \mathbb{X}_i] = \frac{\pi(x)1[x \in \mathbb{X}_i]}{\pi[\mathbb{X}_i]}$$

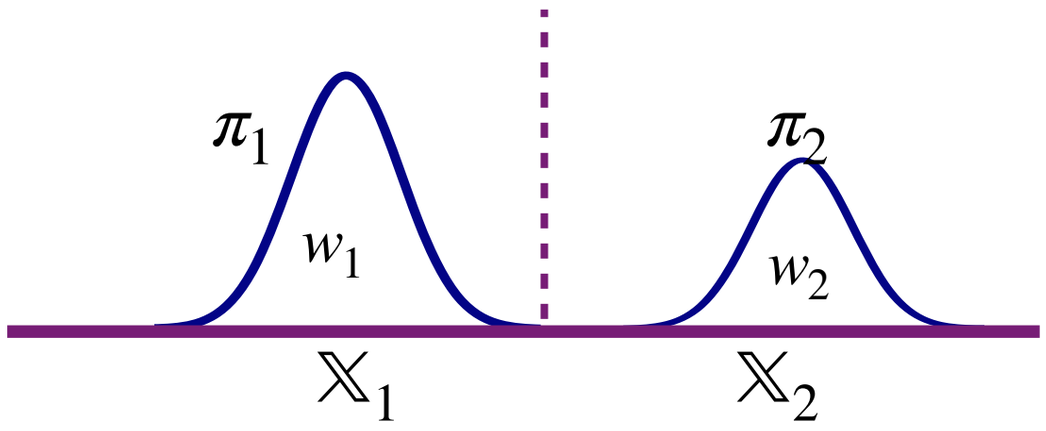
MULTI-MODAL DECOMPOSITION

- Let $\mathbb{X}_i$ be the modes of the target π with respect to K
  - $\mathbb{X}_i \cap \mathbb{X}_j = 0$ for $i \neq j$ and $\mathbb{X} = \cup_i \mathbb{X}_i$
- We have the following decomposition:

$$\pi(x) = \sum_i \pi[\mathbb{X}_i]\pi[x \mid \mathbb{X}_i] \quad = \sum_i w_i \pi_i(x)$$

- Where w_i is the weight of mode i and π_i is π conditional on the mode $\mathbb{X}_i$

$$w_i = \pi[\mathbb{X}_i] \qquad \pi_i(x) = \pi[x \mid \mathbb{X}_i] = \frac{\pi(x)1[x \in \mathbb{X}_i]}{\pi[\mathbb{X}_i]}$$



ERGODIC THEOREM FOR MULTI-MODAL DISTIRBUTIONS

ERGODIC THEOREM FOR MULTI-MODAL DISTIRBUTIONS

- We have the following decomposition multi-modal decomposition π with respect to K

$$\pi = \sum_i w_i \pi_i$$

ERGODIC THEOREM FOR MULTI-MODAL DISTIRBUTIONS

- We have the following decomposition multi-modal decomposition π with respect to K

$$\pi = \sum_i w_i \pi_i$$

- Not that if K is π-invariant, then K is π_i-invariant and π_i-irreducible

ERGODIC THEOREM FOR MULTI-MODAL DISTIRBUTIONS

- We have the following decomposition multi-modal decomposition π with respect to K

$$\pi = \sum_i w_i \pi_i$$

- Not that if K is π-invariant, then K is π_i-invariant and π_i-irreducible

- Let X_t be the Markov chain generated by K, assuming the ergodic theorem hold for each mode:

ERGODIC THEOREM FOR MULTI-MODAL DISTIRBUTIONS

- We have the following decomposition multi-modal decomposition π with respect to K

$$\pi = \sum_i w_i \pi_i$$

- Not that if K is π-invariant, then K is π_i-invariant and π_i-irreducible
- Let X_t be the Markov chain generated by K, assuming the ergodic theorem hold for each mode:
  - If $X_0 \in \mathbb{X}_i$ then we have

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \quad \xrightarrow[T\to\infty]{a.s.} \quad \pi_i[f]$$

ERGODIC THEOREM FOR MULTI-MODAL DISTIRBUTIONS

- We have the following decomposition multi-modal decomposition π with respect to K

$$\pi = \sum_i w_i \pi_i$$

- Not that if K is π-invariant, then K is π_i-invariant and π_i-irreducible
- Let X_t be the Markov chain generated by K, assuming the ergodic theorem hold for each mode:
  - If $X_0 \in \mathbb{X}_i$ then we have

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \quad \xrightarrow[T\to\infty]{a.s.} \quad \pi_i[f]$$

- They can forget the initial condition within a mode, and efficiently sample from the target conditional on the mode

ERGODIC THEOREM FOR MULTI-MODAL DISTIRBUTIONS

- We have the following decomposition multi-modal decomposition π with respect to K

$$\pi = \sum_i w_i \pi_i$$

- Not that if K is π-invariant, then K is π_i-invariant and π_i-irreducible
- Let X_t be the Markov chain generated by K, assuming the ergodic theorem hold for each mode:
  - If $X_0 \in \mathbb{X}_i$ then we have

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \quad \xrightarrow[T\to\infty]{a.s.} \quad \pi_i[f]$$

- They can forget the initial condition within a mode, and efficiently sample from the target conditional on the mode
- **Problem:** They cannot forget the intial mode itself.

ERGODIC THEOREM FOR MULTI-MODAL DISTIRBUTIONS

- We have the following decomposition multi-modal decomposition π with respect to K

$$\pi = \sum_i w_i \pi_i$$

- Not that if K is π-invariant, then K is π_i-invariant and π_i-irreducible
- Let X_t be the Markov chain generated by K, assuming the ergodic theorem hold for each mode:
  - If $X_0 \in \mathbb{X}_i$ then we have

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \quad \xrightarrow[T\to\infty]{a.s.} \quad \pi_i[f]$$

- They can forget the initial condition within a mode, and efficiently sample from the target conditional on the mode
- **Problem:** They cannot forget the intial mode itself.
  - What if we have gradients or do random initialisation or run multiple chains

ERGODIC THEOREM FOR MULTI-MODAL DISTIRBUTIONS

- We have the following decomposition multi-modal decomposition π with respect to K

$$\pi = \sum_i w_i \pi_i$$

- Not that if K is π-invariant, then K is π_i-invariant and π_i-irreducible
- Let X_t be the Markov chain generated by K, assuming the ergodic theorem hold for each mode:
  - If $X_0 \in \mathbb{X}_i$ then we have

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \quad \xrightarrow[T\to\infty]{a.s.} \quad \pi_i[f]$$

- They can forget the initial condition within a mode, and efficiently sample from the target conditional on the mode
- **Problem:** They cannot forget the intial mode itself.
  - What if we have gradients or do random initialisation or run multiple chains
  - Unfortunately none of these work

PITFALLS OF FIRST ORDER METHODS

- In general first-order methods don’t help with multi-modal distributions:

PITFALLS OF FIRST ORDER METHODS

- In general first-order methods don't help with multi-modal distributions:
- We we know each mode up to a normalising constant

$$\pi_i(x) = \frac{\gamma_i(x)}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x) \mathrm{d}x$$

PITFALLS OF FIRST ORDER METHODS

- In general first-order methods don't help with multi-modal distributions:
- We we know each mode up to a normalising constant

$$\pi_i(x) = \frac{\gamma_i(x)}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x) \mathrm{d}x$$

- The normalising constant provides information about releative weighting of modes:

$$\pi(x) \propto \sum_i w_i \gamma_i(x), \qquad Z = \sum_i w_i Z_i$$

PITFALLS OF FIRST ORDER METHODS

- In general first-order methods don't help with multi-modal distributions:
- We we know each mode up to a normalising constant

$$\pi_i(x) = \frac{\gamma_i(x)}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x) \mathrm{d}x$$

- The normalising constant provides information about releative weighting of modes:

$$\pi(x) \propto \sum_i w_i \gamma_i(x), \qquad Z = \sum_i w_i Z_i$$

- The score is agnostic to the normalising constant and relative weight of the modes w_i

$$\nabla \log \pi(x) = \sum_i \nabla \log \gamma_i(x)$$

PITFALLS OF FIRST ORDER METHODS

- In general first-order methods don't help with multi-modal distributions:
- We we know each mode up to a normalising constant

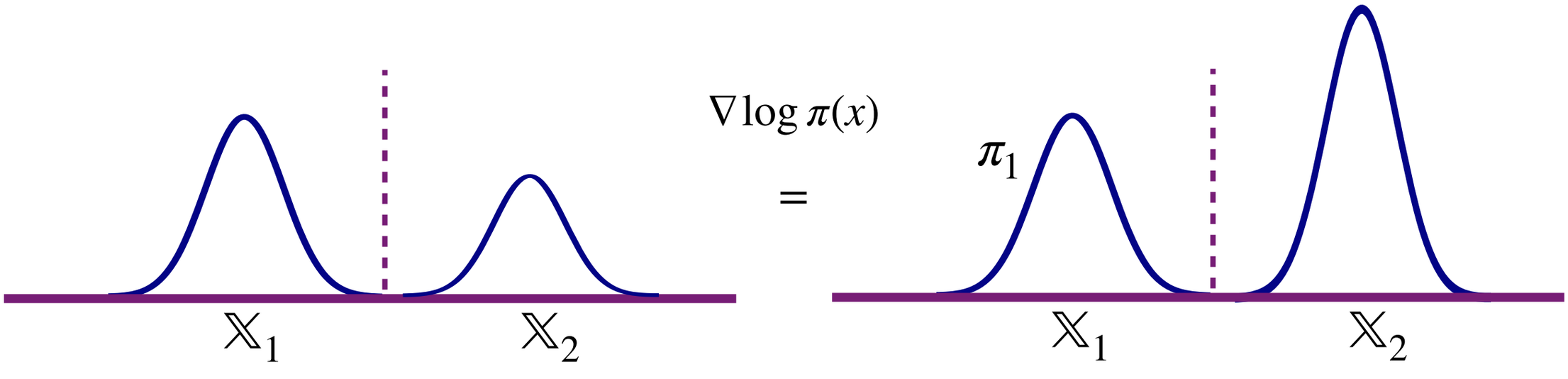
$$\pi_i(x) = \frac{\gamma_i(x)}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x) \mathrm{d}x$$

- The normalising constant provides information about releative weighting of modes:

$$\pi(x) \propto \sum_i w_i \gamma_i(x), \qquad Z = \sum_i w_i Z_i$$

- The score is agnostic to the normalising constant and relative weight of the modes w_i

$$\nabla \log \pi(x) = \sum_i \nabla \log \gamma_i(x)$$



PITFALL OF RANDOM INITIALISATION

PITFALL OF RANDOM INITIALISATION

- Suppose we initialise $X_0 \sim \mu$ for some distribution μ over $\mathbb{X}$

PITFALL OF RANDOM INITIALISATION

- Suppose we initialise $X_0 \sim \mu$ for some distribution μ over $\mathbb{X}$
  - If $X_0 \sim \mu$ then let $\alpha_i = \mu[\mathbb{X}_i]$

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \quad \xrightarrow[T\to\infty]{a.s.} \quad \sum_i \alpha_i \pi_i[f]$$

PITFALL OF RANDOM INITIALISATION

- Suppose we initialise $X_0 \sim \mu$ for some distribution μ over $\mathbb{X}$
  - If $X_0 \sim \mu$ then let $\alpha_i = \mu[\mathbb{X}_i]$

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \quad \xrightarrow[T\to\infty]{a.s.} \quad \sum_i \alpha_i \pi_i[f]$$

- Note that this implies that even if we initialise randomly, we can efficiently sample form

$$\pi_\mu = \sum_i \alpha_i \pi_i$$

PITFALL OF RANDOM INITIALISATION

- Suppose we initialise $X_0 \sim \mu$ for some distribution μ over $\mathbb{X}$
  - If $X_0 \sim \mu$ then let $\alpha_i = \mu[\mathbb{X}_i]$

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \quad \xrightarrow[T\to\infty]{a.s.} \quad \sum_i \alpha_i \pi_i[f]$$

- Note that this implies that even if we initialise randomly, we can efficiently sample form

$$\pi_\mu = \sum_i \alpha_i \pi_i$$

  - $\pi_\mu = \pi$ if and only if $\mu[\mathbb{X}_i] = \pi[\mathbb{X}_i]$

PITFALL OF RANDOM INITIALISATION

- Suppose we initialise $X_0 \sim \mu$ for some distribution μ over $\mathbb{X}$
  - If $X_0 \sim \mu$ then let $\alpha_i = \mu[\mathbb{X}_i]$

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \quad \xrightarrow[T\to\infty]{a.s.} \quad \sum_i \alpha_i \pi_i[f]$$

- Note that this implies that even if we initialise randomly, we can efficiently sample form

$$\pi_\mu = \sum_i \alpha_i \pi_i$$

  - $\pi_\mu = \pi$ if and only if $\mu[\mathbb{X}_i] = \pi[\mathbb{X}_i]$

- If we know the size and location of the modes of π then we can construct μ such that the ergodic average is consistent

PITFALL OF RANDOM INITIALISATION

- Suppose we initialise $X_0 \sim \mu$ for some distribution μ over $\mathbb{X}$
  - If $X_0 \sim \mu$ then let $\alpha_i = \mu[\mathbb{X}_i]$

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \quad \xrightarrow[T\to\infty]{a.s.} \quad \sum_i \alpha_i \pi_i[f]$$

- Note that this implies that even if we initialise randomly, we can efficiently sample form

$$\pi_\mu = \sum_i \alpha_i \pi_i$$

  - $\pi_\mu = \pi$ if and only if $\mu[\mathbb{X}_i] = \pi[\mathbb{X}_i]$

- If we know the size and location of the modes of π then we can construct μ such that the ergodic average is consistent
- In general, we do not know this information, and therefore will get biased results.

PITFALLS OF TARGETTING A DISTRIBUTION

PITFALLS OF TARGETTING A DISTRIBUTION

- The normalising constant encode the information about the relative size of modes

$$\pi_i(x) = \frac{\gamma(x)1[x \in \mathbb{X}_i]}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x)\mathrm{d}x$$

$$\pi(x) \propto \sum_i w_i \gamma_i(x), \qquad Z = \sum_i w_i Z_i$$

PITFALLS OF TARGETTING A DISTRIBUTION

- The normalising constant encode the information about the relative size of modes

$$\pi_i(x) = \frac{\gamma(x)1[x \in \mathbb{X}_i]}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x)\mathrm{d}x$$

$$\pi(x) \propto \sum_i w_i \gamma_i(x), \qquad Z = \sum_i w_i Z_i$$

- For a valid MCMC algorithm should be invariant to the choice of normalising constant

PITFALLS OF TARGETTING A DISTRIBUTION

- The normalising constant encode the information about the relative size of modes

$$\pi_i(x) = \frac{\gamma(x)1[x \in \mathbb{X}_i]}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x)\mathrm{d}x$$

$$\pi(x) \propto \sum_i w_i \gamma_i(x), \qquad Z = \sum_i w_i Z_i$$

- For a valid MCMC algorithm should be invariant to the choice of normalising constant
  - The choice of the mode is determined by the intiatial state X_0

PITFALLS OF TARGETTING A DISTRIBUTION

- The normalising constant encode the information about the relative size of modes

$$\pi_i(x) = \frac{\gamma(x)1[x \in \mathbb{X}_i]}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x)\mathrm{d}x$$

$$\pi(x) \propto \sum_i w_i \gamma_i(x), \qquad Z = \sum_i w_i Z_i$$

- For a valid MCMC algorithm should be invariant to the choice of normalising constant
  - The choice of the mode is determined by the intiatial state X_0
  - The chain $X_0, X_1, \ldots, X_t, \ldots$ generated by K is identical regardless of w_i or Z_i

PITFALLS OF TARGETTING A DISTRIBUTION

- The normalising constant encode the information about the relative size of modes

$$\pi_i(x) = \frac{\gamma(x)1[x \in \mathbb{X}_i]}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x)\mathrm{d}x$$

$$\pi(x) \propto \sum_i w_i\gamma_i(x), \qquad Z = \sum_i w_i Z_i$$

- For a valid MCMC algorithm should be invariant to the choice of normalising constant
  - The choice of the mode is determined by the intiatial state X_0
  - The chain $X_0, X_1, \ldots, X_t, \ldots$ generated by K is identical regardless of w_i or Z_i

$$\pi \propto \sum_i w_i\gamma_i \propto \sum_i w_i'\gamma_i'$$

PITFALLS OF TARGETTING A DISTRIBUTION

- The normalising constant encode the information about the relative size of modes

$$\pi_i(x) = \frac{\gamma(x)1[x \in \mathbb{X}_i]}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x)\mathrm{d}x$$

$$\pi(x) \propto \sum_i w_i\gamma_i(x), \qquad Z = \sum_i w_i Z_i$$

- For a valid MCMC algorithm should be invariant to the choice of normalising constant
  - The choice of the mode is determined by the intiatial state X_0
  - The chain $X_0, X_1, \ldots, X_t, \ldots$ generated by K is identical regardless of w_i or Z_i

$$\pi \propto \sum_i w_i\gamma_i \propto \sum_i w_i'\gamma_i'$$

  - It means we can not estimate Z_i even with a mode

PITFALLS OF TARGETTING A DISTRIBUTION

- The normalising constant encode the information about the relative size of modes

$$\pi_i(x) = \frac{\gamma(x)1[x \in \mathbb{X}_i]}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x)\mathrm{d}x$$

$$\pi(x) \propto \sum_i w_i\gamma_i(x), \qquad Z = \sum_i w_i Z_i$$

- For a valid MCMC algorithm should be invariant to the choice of normalising constant
  - The choice of the mode is determined by the intiatial state X_0
  - The chain $X_0, X_1, \ldots, X_t, \ldots$ generated by K is identical regardless of w_i or Z_i

$$\pi \propto \sum_i w_i\gamma_i \propto \sum_i w_i'\gamma_i'$$

  - It means we can not estimate Z_i even with a mode
- To learn the normalising constant, we need something to compare against

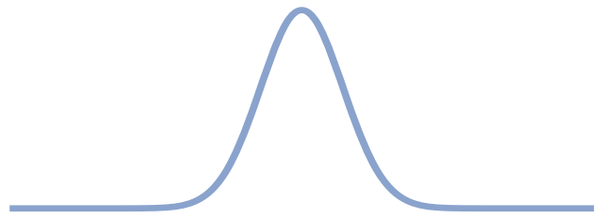
REFERENCE DISTRIBUTIONS

- Suppose we have a target distribution π:

REFERENCE DISTRIBUTIONS

- Suppose we have a target distribution π:

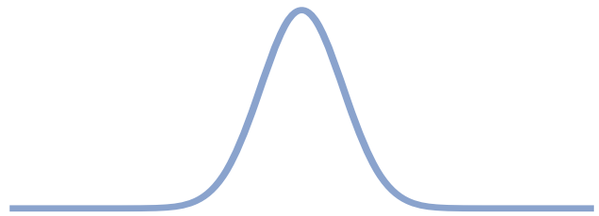
$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x, \quad Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x$$



REFERENCE DISTRIBUTIONS

- Suppose we have a target distribution π:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x, \quad Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x$$

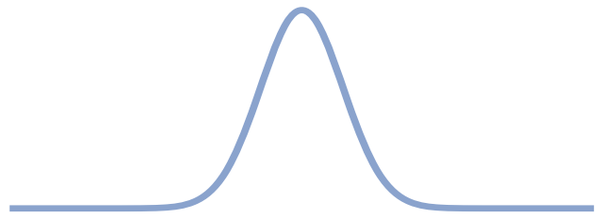


- We will say η is a **reference** distribution for π if:

REFERENCE DISTRIBUTIONS

- Suppose we have a target distribution π:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x, \quad Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x$$

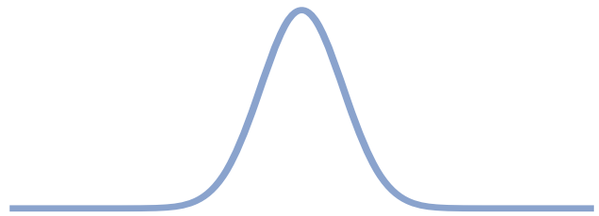


- We will say η is a **reference** distribution for π if:
  - We can sample $X \sim \eta$

REFERENCE DISTRIBUTIONS

- Suppose we have a target distribution π:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x, \quad Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x$$

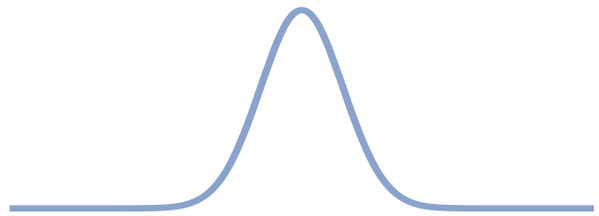


- We will say η is a **reference** distribution for π if:
  - We can sample $X \sim \eta$
  - We can evaluate the normalised density $\eta(x)$, where $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$

REFERENCE DISTRIBUTIONS

- Suppose we have a target distribution π:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x, \quad Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x$$

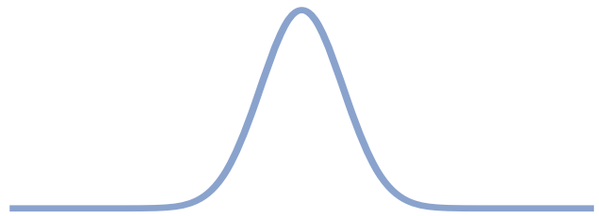


- We will say η is a **reference** distribution for π if:
  - We can sample $X \sim \eta$
  - We can evaluate the normalised density $\eta(x)$, where $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$
  - We have $\pi \ll \eta$ and we can evaluate the weight function $w : \mathbb{X} \to \mathbb{R}_+$defined as the un-normlaised likelihood ratio:

REFERENCE DISTRIBUTIONS

- Suppose we have a target distribution π:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x, \quad Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x$$



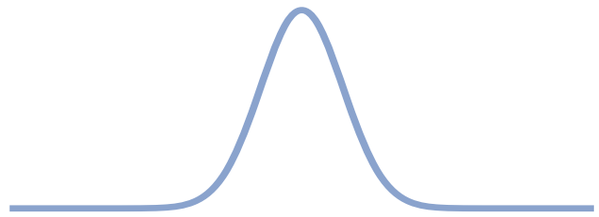
- We will say η is a **reference** distribution for π if:
  - We can sample $X \sim \eta$
  - We can evaluate the normalised density $\eta(x)$, where $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$
  - We have $\pi \ll \eta$ and we can evaluate the weight function $w : \mathbb{X} \to \mathbb{R}_+$defined as the un-normlaised likelihood ratio:

$$w(x) = Z\frac{\mathrm{d}\pi}{\mathrm{d}\eta}(x) = \frac{\gamma(x)}{\eta(x)}$$

REFERENCE DISTRIBUTIONS

- Suppose we have a target distribution π:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x, \quad Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x$$



- We will say η is a **reference** distribution for π if:
  - We can sample $X \sim \eta$
  - We can evaluate the normalised density $\eta(x)$, where $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$
  - We have $\pi \ll \eta$ and we can evaluate the weight function $w : \mathbb{X} \to \mathbb{R}_+$defined as the un-normlaised likelihood ratio:

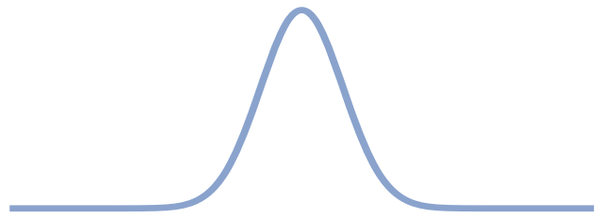
$$w(x) = Z\frac{\mathrm{d}\pi}{\mathrm{d}\eta}(x) = \frac{\gamma(x)}{\eta(x)}$$

- Reference distribution should be simple and tractable:

REFERENCE DISTRIBUTIONS

- Suppose we have a target distribution π:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x, \quad Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x$$



- We will say η is a **reference** distribution for π if:
  - We can sample $X \sim \eta$
  - We can evaluate the normalised density $\eta(x)$, where $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$
  - We have $\pi \ll \eta$ and we can evaluate the weight function $w : \mathbb{X} \to \mathbb{R}_+$defined as the un-normlaised likelihood ratio:

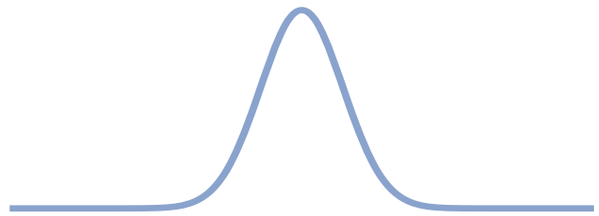
$$w(x) = Z\frac{\mathrm{d}\pi}{\mathrm{d}\eta}(x) = \frac{\gamma(x)}{\eta(x)}$$

- Reference distribution should be simple and tractable:
  - Common choices will be Gaussians, uniform distribution, priors, etc

REFERENCE DISTRIBUTIONS

- Suppose we have a target distribution π:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x, \quad Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x$$



- We will say η is a **reference** distribution for π if:
  - We can sample $X \sim \eta$
  - We can evaluate the normalised density $\eta(x)$, where $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$
  - We have $\pi \ll \eta$ and we can evaluate the weight function $w : \mathbb{X} \to \mathbb{R}_+$defined as the un-normlaised likelihood ratio:

$$w(x) = Z\frac{\mathrm{d}\pi}{\mathrm{d}\eta}(x) = \frac{\gamma(x)}{\eta(x)}$$

- Reference distribution should be simple and tractable:
  - Common choices will be Gaussians, uniform distribution, priors, etc
  - Typically the intialisation for the MCMC chain

REFERENCE DISTRIBUTIONS

REFERENCE DISTRIBUTIONS

- Note that given a reference distributions we can approximate the normalising constant:

REFERENCE DISTRIBUTIONS

- Note that given a reference distributions we can approximate the normalising constant:
- Recall that with a reference distribition:

$$Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x = \int_{\mathbb{X}} w(x)\eta(x)\mathrm{d}x = \eta[w]$$

  - Use Monte Carlo we can approximate:

$$X_1, \ldots, X_N \sim \eta \qquad \hat{Z} = \frac{1}{N}\sum_{n=1}^{N} w(X_n)$$

- It is unbiased and consistent estimator for Z

$$\hat{Z} \xrightarrow[N\to\infty]{a.s.} Z$$

REFERENCE DISTRIBUTIONS

- Note that given a reference distributions we can approximate the normalising constant:
- Recall that with a reference distribition:

$$Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x = \int_{\mathbb{X}} w(x)\eta(x)\mathrm{d}x = \eta[w]$$

  - Use Monte Carlo we can approximate:

$$X_1, \ldots, X_N \sim \eta \qquad \hat{Z} = \frac{1}{N}\sum_{n=1}^{N} w(X_n)$$

    - It is unbiased and consistent estimator for Z

$$\hat{Z} \xrightarrow[N\to\infty]{a.s.} Z$$

- If we have a local sampler K, we can learn a mode

REFERENCE DISTRIBUTIONS

- Note that given a reference distributions we can approximate the normalising constant:
- Recall that with a reference distribition:

$$Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x = \int_{\mathbb{X}} w(x)\eta(x)\mathrm{d}x = \eta[w]$$

  - Use Monte Carlo we can approximate:

$$X_1, \ldots, X_N \sim \eta \qquad \hat{Z} = \frac{1}{N}\sum_{n=1}^{N} w(X_n)$$

    - It is unbiased and consistent estimator for Z

$$\hat{Z} \xrightarrow[N\to\infty]{a.s.} Z$$

- If we have a local sampler K, we can learn a mode
  - Can get consistency if we can initialise in the modes with the right weights

REFERENCE DISTRIBUTIONS

- Note that given a reference distributions we can approximate the normalising constant:
- Recall that with a reference distribition:

$$Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x = \int_{\mathbb{X}} w(x)\eta(x)\mathrm{d}x = \eta[w]$$

  - Use Monte Carlo we can approximate:

$$X_1, \ldots, X_N \sim \eta \qquad \hat{Z} = \frac{1}{N}\sum_{n=1}^{N} w(X_n)$$

    - It is unbiased and consistent estimator for Z

$$\hat{Z} \xrightarrow[N\to\infty]{a.s.} Z$$

- If we have a local sampler K, we can learn a mode
  - Can get consistency if we can initialise in the modes with the right weights
- Reference is able to learn the normalising constant and hence the modes

REFERENCE DISTRIBUTIONS

- Note that given a reference distributions we can approximate the normalising constant:
- Recall that with a reference distribition:

$$Z = \int_{\mathbb{X}} \gamma(x)\mathrm{d}x = \int_{\mathbb{X}} w(x)\eta(x)\mathrm{d}x = \eta[w]$$

  - Use Monte Carlo we can approximate:

$$X_1, \ldots, X_N \sim \eta \qquad \hat{Z} = \frac{1}{N}\sum_{n=1}^{N} w(X_n)$$

    - It is unbiased and consistent estimator for Z

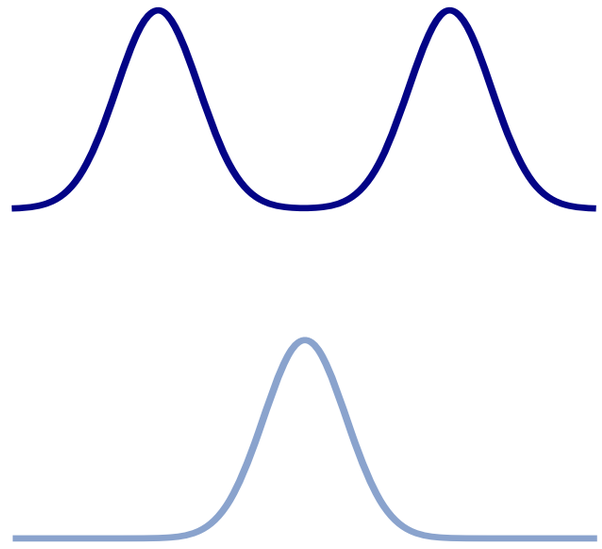
$$\hat{Z} \xrightarrow[N\to\infty]{a.s.} Z$$

- If we have a local sampler K, we can learn a mode
  - Can get consistency if we can initialise in the modes with the right weights
- Reference is able to learn the normalising constant and hence the modes
- Can use the reference to stabilise generated by K

RERERENCE STABILISED MCMC

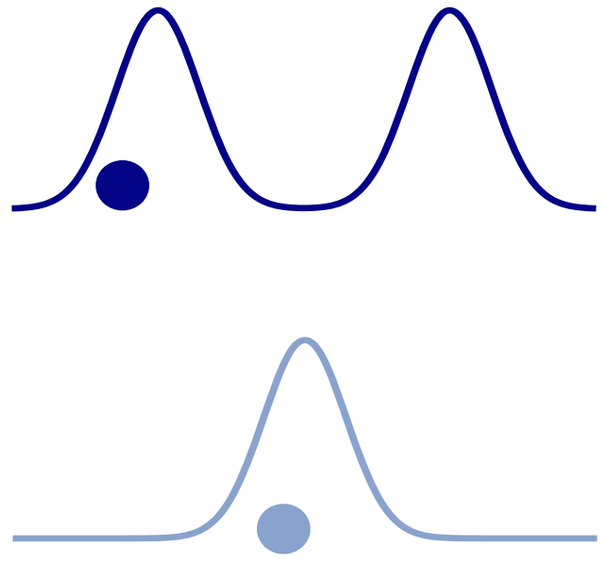
RERERENCE STABILISED MCMC

- Generate a chain in $(X_t, X_t') \in \mathbb{X}^2$ targeting $\pi \otimes \eta$



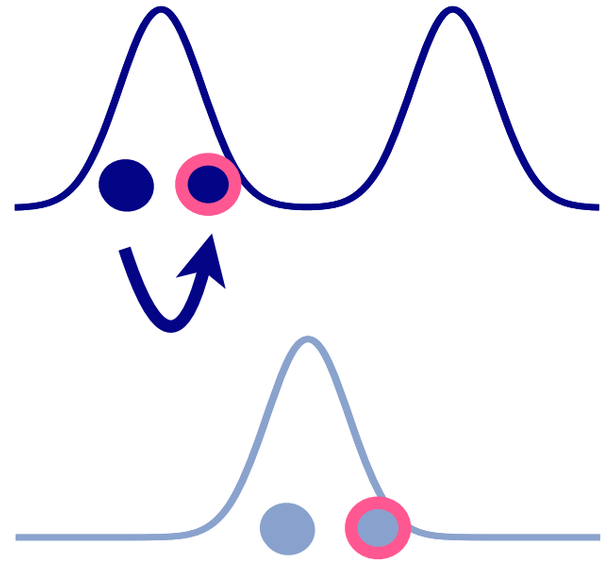
RERERENCE STABILISED MCMC

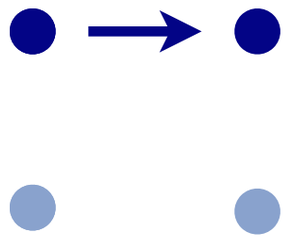
- Generate a chain in $(X_t, X_t') \in \mathbb{X}^2$ targeting $\pi \otimes \eta$
  - **Intitialise:** $X_0, X_0' \sim \eta$



RERERENCE STABILISED MCMC

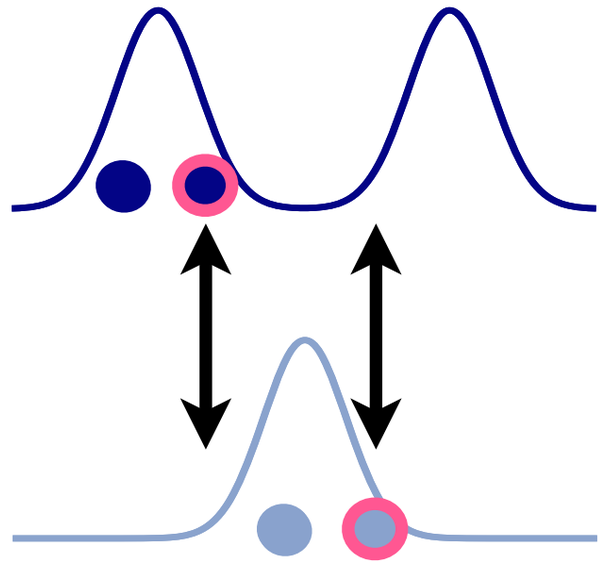
- Generate a chain in $(X_t, X'_t) \in \mathbb{X}^2$ targeting $\pi \otimes \eta$
  - **Intitialise:** $X_0, X'_0 \sim \eta$
  - **Local move:**
    - Propogate target state $X \sim K(X_{t-1}, \mathrm{d}x)$
    - Generate new reference sample $X' \sim \eta$



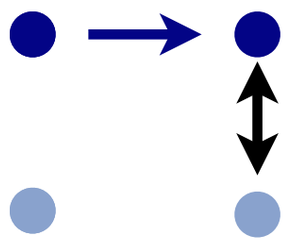


RERERENCE STABILISED MCMC

- Generate a chain in $(X_t, X'_t) \in \mathbb{X}^2$ targeting $\pi \otimes \eta$
  - **Intitialise:** $X_0, X'_0 \sim \eta$
  - **Local move:**
    - Propogate target state $X \sim K(X_{t-1}, \mathrm{d}x)$
    - Generate new reference sample $X' \sim \eta$

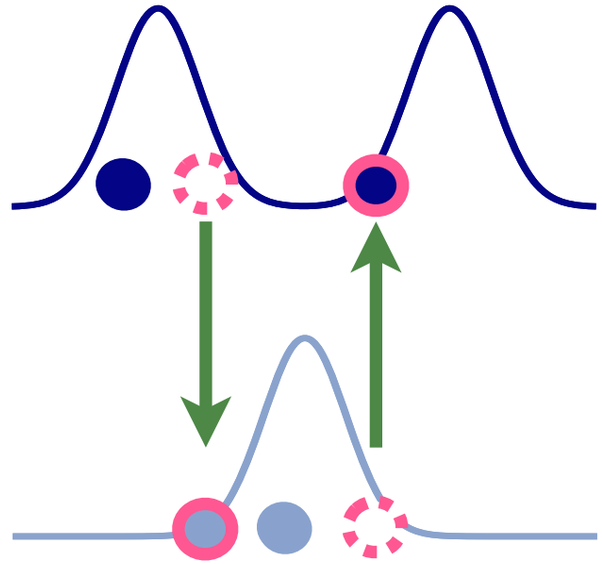


- **Communication:** propose a metropolised swap



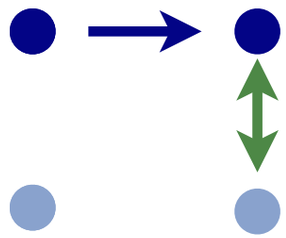
RERERENCE STABILISED MCMC

- Generate a chain in $(X_t, X'_t) \in \mathbb{X}^2$ targeting $\pi \otimes \eta$
  - **Intitialise:** $X_0, X'_0 \sim \eta$
  - **Local move:**
    - Propogate target state $X \sim K(X_{t-1}, \mathrm{d}x)$
    - Generate new reference sample $X' \sim \eta$



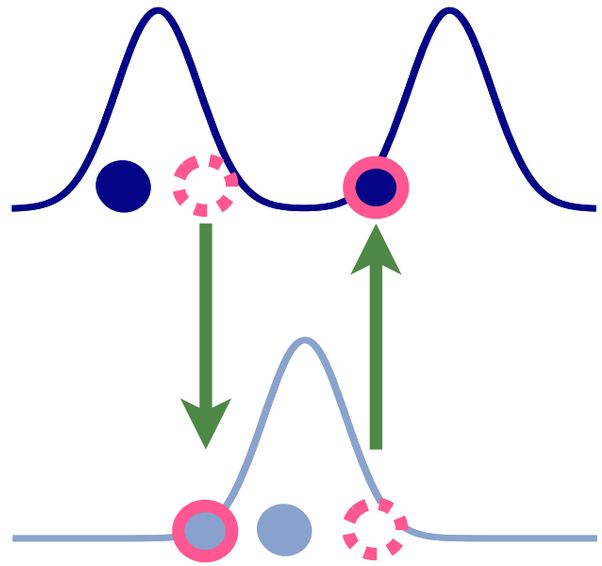
- **Communication:** propose a metropolised swap
  - **Accept** with probability $\alpha(X, X')$

$$(X_t, X'_t) = (X', X)$$



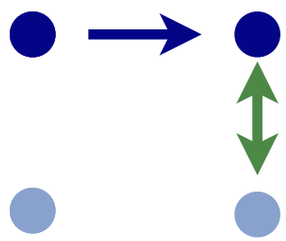
RERERENCE STABILISED MCMC

- Generate a chain in $(X_t, X'_t) \in \mathbb{X}^2$ targeting $\pi \otimes \eta$
  - **Intitialise:** $X_0, X'_0 \sim \eta$
  - **Local move:**
    - Propogate target state $X \sim K(X_{t-1}, \mathrm{d}x)$
    - Generate new reference sample $X' \sim \eta$



- **Communication:** propose a metropolised swap
  - **Accept** with probability $\alpha(X, X')$

$$(X_t, X'_t) = (X', X)$$

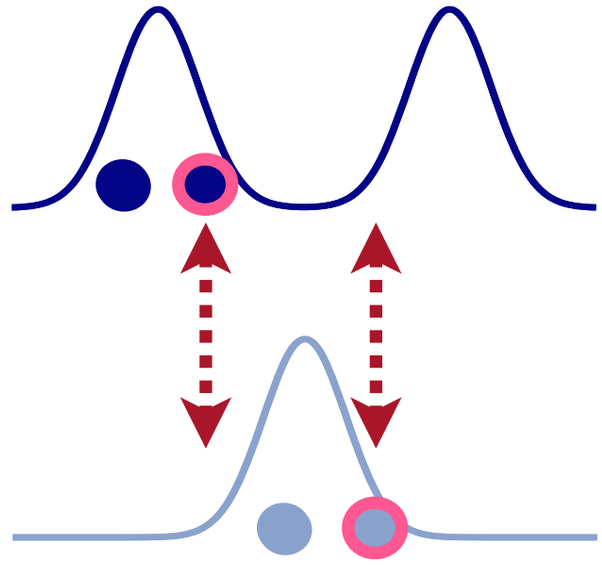


  - Where $\alpha(x, x')$ equals:

$$\alpha(x, x') = 1 \wedge \frac{\pi(x')\eta(x)}{\pi(x)\eta(x')} = 1 \wedge \frac{w(x')}{w(x)}$$

RERERENCE STABILISED MCMC

- Generate a chain in $(X_t, X'_t) \in \mathbb{X}^2$ targeting $\pi \otimes \eta$
  - **Intitialise:** $X_0, X'_0 \sim \eta$
  - **Local move:**
    - Propogate target state $X \sim K(X_{t-1}, \mathrm{d}x)$
    - Generate new reference sample $X' \sim \eta$



- **Communication:** propose a metropolised swap
  - **Accept** with probability $\alpha(X, X')$

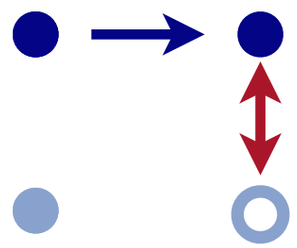
$$(X_t, X'_t) = (X', X)$$

  - **Reject** otherwise and set

$$(X_t, X'_t) = (X, X')$$

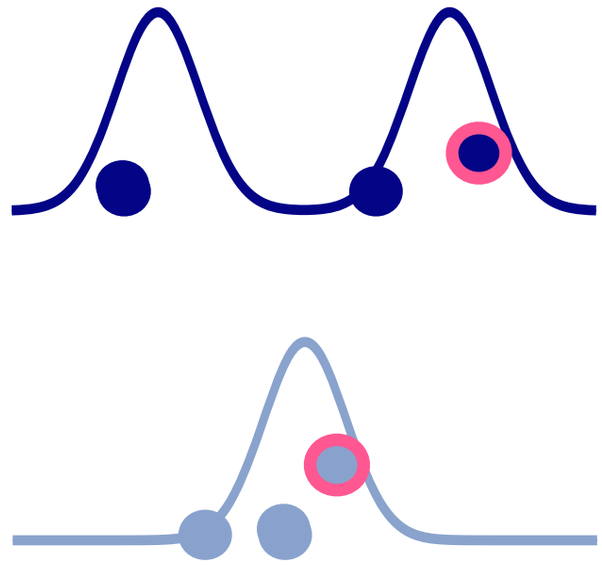
  - Where $\alpha(x, x')$ equals:

$$\alpha(x, x') = 1 \wedge \frac{\pi(x')\eta(x)}{\pi(x)\eta(x')} = 1 \wedge \frac{w(x')}{w(x)}$$



RERERENCE STABILISED MCMC

- Generate a chain in $(X_t, X'_t) \in \mathbb{X}^2$ targeting $\pi \otimes \eta$
  - **Intitialise:** $X_0, X'_0 \sim \eta$
  - **Local move:**
    - Propogate target state $X \sim K(X_{t-1}, \mathrm{d}x)$
    - Generate new reference sample $X' \sim \eta$



- **Communication:** propose a metropolised swap
  - **Accept** with probability $\alpha(X, X')$

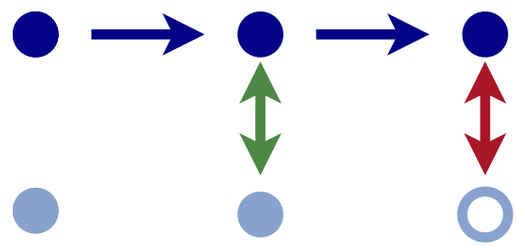
$$(X_t, X'_t) = (X', X)$$

  - **Reject** otherwise and set

$$(X_t, X'_t) = (X, X')$$

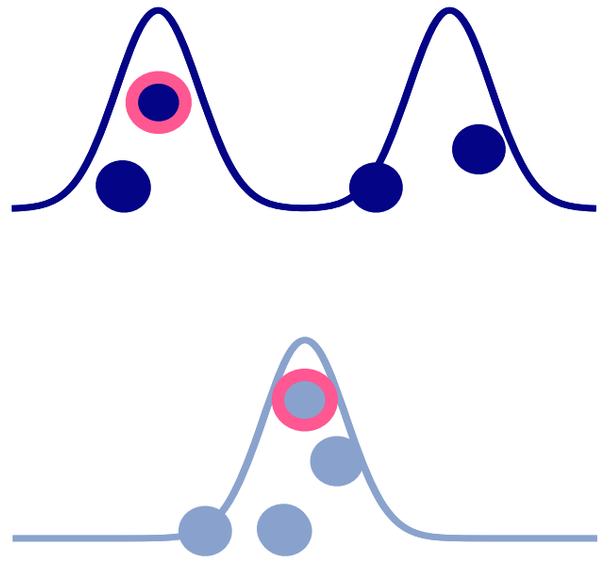
  - Where $\alpha(x, x')$ equals:

$$\alpha(x, x') = 1 \wedge \frac{\pi(x')\eta(x)}{\pi(x)\eta(x')} = 1 \wedge \frac{w(x')}{w(x)}$$



RERERENCE STABILISED MCMC

- Generate a chain in $(X_t, X'_t) \in \mathbb{X}^2$ targeting $\pi \otimes \eta$
  - **Intitialise:** $X_0, X'_0 \sim \eta$
  - **Local move:**
    - Propogate target state $X \sim K(X_{t-1}, \mathrm{d}x)$
    - Generate new reference sample $X' \sim \eta$



- **Communication:** propose a metropolised swap
  - **Accept** with probability $\alpha(X, X')$

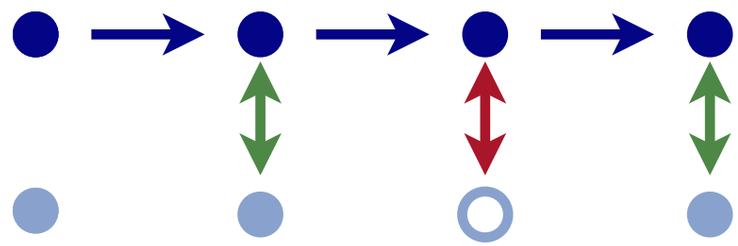
$$(X_t, X'_t) = (X', X)$$

  - **Reject** otherwise and set

$$(X_t, X'_t) = (X, X')$$

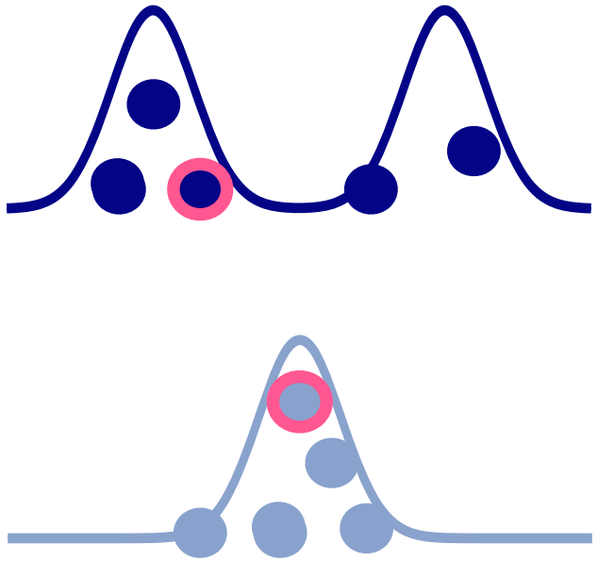
  - Where $\alpha(x, x')$ equals:

$$\alpha(x, x') = 1 \wedge \frac{\pi(x')\eta(x)}{\pi(x)\eta(x')} = 1 \wedge \frac{w(x')}{w(x)}$$



RERERENCE STABILISED MCMC

- Generate a chain in $(X_t, X'_t) \in \mathbb{X}^2$ targeting $\pi \otimes \eta$
  - **Intitialise:** $X_0, X'_0 \sim \eta$
  - **Local move:**
    - Propogate target state $X \sim K(X_{t-1}, \mathrm{d}x)$
    - Generate new reference sample $X' \sim \eta$



- **Communication:** propose a metropolised swap
  - **Accept** with probability $\alpha(X, X')$

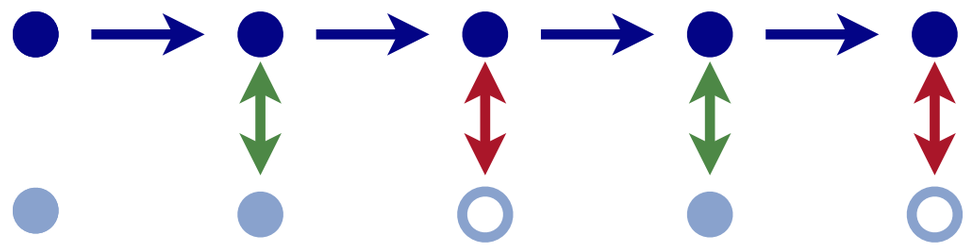
$$(X_t, X'_t) = (X', X)$$

  - **Reject** otherwise and set

$$(X_t, X'_t) = (X, X')$$

  - Where $\alpha(x, x')$ equals:

$$\alpha(x, x') = 1 \wedge \frac{\pi(x')\eta(x)}{\pi(x)\eta(x')} = 1 \wedge \frac{w(x')}{w(x)}$$



REFERENCE DISTRIBUTIONS

REFERENCE DISTRIBUTIONS

- A successful swap acts as a random initialisation that matches the proportionality of the target!

REFERENCE DISTRIBUTIONS

- A successful swap acts as a random initialisation that matches the proportionality of the target!
- Can use the reference samples to also obtain normalising constant estimates

$$\hat{Z} = \frac{1}{T}\sum_{t=1}^{T} w(X'_t), \qquad \mathbb{V}\left[\frac{\hat{Z}}{Z}\right] = \frac{\exp(D(\pi\|\eta)) - 1}{N}$$

$$D(\pi\|\eta) = \log(1 + \chi^2(\pi\|\eta)), \qquad \chi^2(\pi\|\eta) = \mathbb{V}_\eta\left[\frac{\mathrm{d}\pi}{\mathrm{d}\eta}\right]$$

REFERENCE DISTRIBUTIONS

- A successful swap acts as a random initialisation that matches the proportionality of the target!
- Can use the reference samples to also obtain normalising constant estimates

$$\hat{Z} = \frac{1}{T}\sum_{t=1}^{T} w(X'_t), \qquad \mathbb{V}\left[\frac{\hat{Z}}{Z}\right] = \frac{\exp(D(\pi\|\eta)) - 1}{N}$$

$$D(\pi\|\eta) = \log(1 + \chi^2(\pi\|\eta)), \qquad \chi^2(\pi\|\eta) = \mathbb{V}_\eta\left[\frac{\mathrm{d}\pi}{\mathrm{d}\eta}\right]$$

- The swaps and normalising constants are stable if $\eta \approx \pi$ rapidly deterioriates otherwise

REFERENCE DISTRIBUTIONS

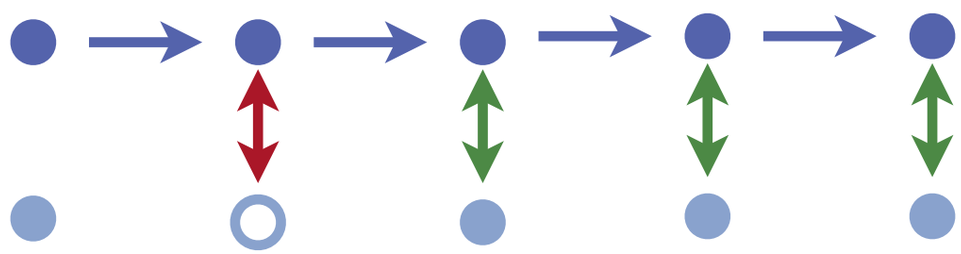
- A successful swap acts as a random initialisation that matches the proportionality of the target!
- Can use the reference samples to also obtain normalising constant estimates

$$\hat{Z} = \frac{1}{T}\sum_{t=1}^{T} w(X'_t), \qquad \mathbb{V}\left[\frac{\hat{Z}}{Z}\right] = \frac{\exp(D(\pi\|\eta)) - 1}{N}$$

$$D(\pi\|\eta) = \log(1 + \chi^2(\pi\|\eta)), \qquad \chi^2(\pi\|\eta) = \mathbb{V}_\eta\left[\frac{\mathrm{d}\pi}{\mathrm{d}\eta}\right]$$

- The swaps and normalising constants are stable if $\eta \approx \pi$ rapidly deterioriates otherwise

$$\lim_{D\to 0^+} \mathbb{V}\left[\frac{\hat{Z}}{Z}\right] \sim \frac{D}{T}$$



REFERENCE DISTRIBUTIONS

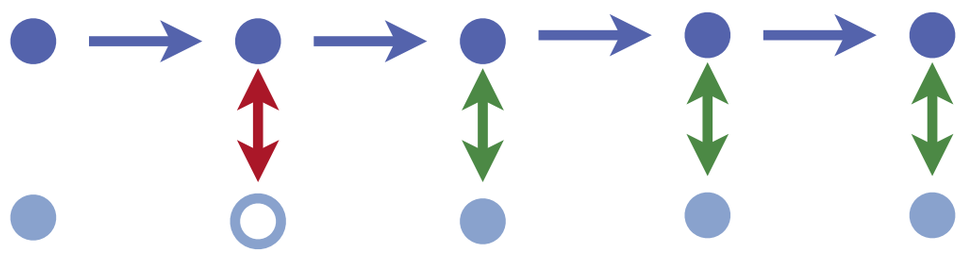
- A successful swap acts as a random initialisation that matches the proportionality of the target!
- Can use the reference samples to also obtain normalising constant estimates

$$\hat{Z} = \frac{1}{T}\sum_{t=1}^{T} w(X'_t), \qquad \mathbb{V}\left[\frac{\hat{Z}}{Z}\right] = \frac{\exp(D(\pi\|\eta)) - 1}{N}$$

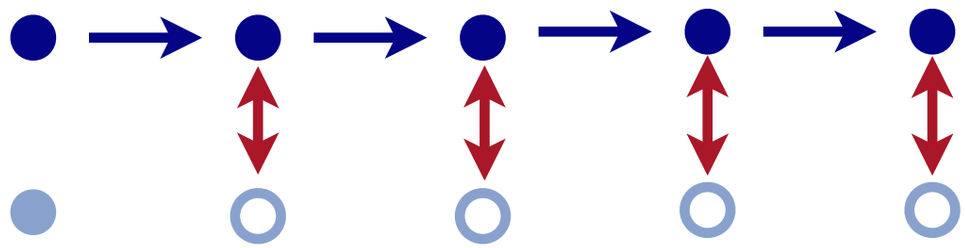
$$D(\pi\|\eta) = \log(1 + \chi^2(\pi\|\eta)), \qquad \chi^2(\pi\|\eta) = \mathbb{V}_\eta\left[\frac{\mathrm{d}\pi}{\mathrm{d}\eta}\right]$$

- The swaps and normalising constants are stable if $\eta \approx \pi$ rapidly deterioriates otherwise

$$\lim_{D\to 0^+} \mathbb{V}\left[\frac{\hat{Z}}{Z}\right] \sim \frac{D}{T}$$



$$\lim_{D\to \infty} \mathbb{V}\left[\frac{\hat{Z}}{Z}\right] \sim \frac{\exp(D)}{T}$$



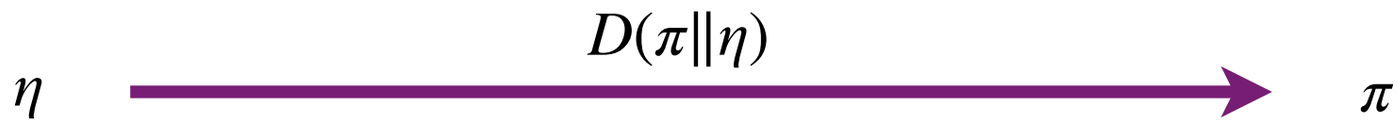
VARIATIONAL INFERENCE

VARIATIONAL INFERENCE

- Given a divergence D and a reference η and target π

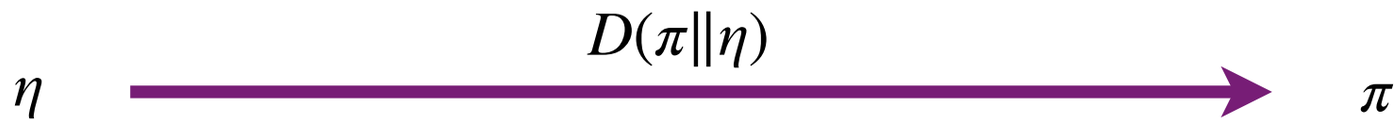
VARIATIONAL INFERENCE

- Given a divergence D and a reference η and target π



VARIATIONAL INFERENCE

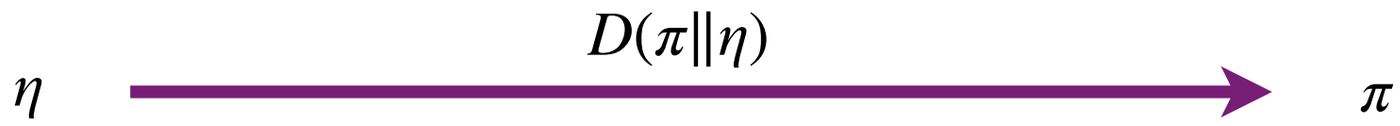
- Given a divergence D and a reference η and target π



- The choice of reference is crutial

VARIATIONAL INFERENCE

- Given a divergence D and a reference η and target π



  - The choice of reference is crutial
    - We can efficiently estimate π using the reference η if and only if $D(\pi||\eta)$ is small

VARIATIONAL INFERENCE

- Given a divergence D and a reference η and target π

$$\eta \xrightarrow{\qquad D(\pi\|\eta) \qquad} \pi$$

- The choice of reference is crutial
  - We can efficiently estimate π using the reference η if and only if $D(\pi\|\eta)$ is small
  - We want a reference that is as close to the target as possible

VARIATIONAL INFERENCE

- Given a divergence D and a reference η and target π

$$\eta \xrightarrow{\quad D(\pi\|\eta) \quad} \pi$$

- The choice of reference is crutial
  - We can efficiently estimate π using the reference η if and only if $D(\pi\|\eta)$ is small
  - We want a reference that is as close to the target as possible

- **Solution 1:** Keep the target fixed and modify the reference (i.e. Variational inference)

VARIATIONAL INFERENCE

- Given a divergence D and a reference η and target π

$$\eta \xrightarrow{\quad D(\pi\|\eta) \quad} \pi$$

- The choice of reference is crutial
  - We can efficiently estimate π using the reference η if and only if $D(\pi\|\eta)$ is small
  - We want a reference that is as close to the target as possible

- **Solution 1:** Keep the target fixed and modify the reference (i.e. Variational inference)

$$\eta_\theta \xrightarrow{\quad D(\pi\|\eta_\theta) \quad} \pi$$

VARIATIONAL INFERENCE

- Given a divergence D and a reference η and target π

$$\eta \xrightarrow{\quad D(\pi\|\eta) \quad} \pi$$

- The choice of reference is crutial
  - We can efficiently estimate π using the reference η if and only if $D(\pi\|\eta)$ is small
  - We want a reference that is as close to the target as possible

- **Solution 1:** Keep the target fixed and modify the reference (i.e. Variational inference)

$$\eta_\theta \xrightarrow{\quad D(\pi\|\eta_\theta) \quad} \pi$$

- Let $\mathcal{Q} = \{\eta_\theta\}_{\theta\in\Theta}$ be a **variational** family of reference distribution

VARIATIONAL INFERENCE

- Given a divergence D and a reference η and target π

$$\eta \xrightarrow{\quad D(\pi\|\eta) \quad} \pi$$

- The choice of reference is crutial
  - We can efficiently estimate π using the reference η if and only if $D(\pi\|\eta)$ is small
  - We want a reference that is as close to the target as possible

- **Solution 1:** Keep the target fixed and modify the reference (i.e. Variational inference)

$$\eta_\theta \xrightarrow{\quad D(\pi\|\eta_\theta) \quad} \pi$$

- Let $\mathcal{Q} = \{\eta_\theta\}_{\theta\in\Theta}$ be a **variational** family of reference distribution
- **Goal:** Find the best reference in $\mathcal{Q}$ that approximages the target

VARIATIONAL INFERENCE

- Given a divergence D and a reference η and target π

$$\eta \xrightarrow{\quad D(\pi\|\eta) \quad} \pi$$

- The choice of reference is crutial
  - We can efficiently estimate π using the reference η if and only if $D(\pi\|\eta)$ is small
  - We want a reference that is as close to the target as possible

- **Solution 1:** Keep the target fixed and modify the reference (i.e. Variational inference)

$$\eta_\theta \xrightarrow{\quad D(\pi\|\eta_\theta) \quad} \pi$$

- Let $\mathcal{Q} = \{\eta_\theta\}_{\theta\in\Theta}$ be a **variational** family of reference distribution
- **Goal:** Find the best reference in $\mathcal{Q}$ that approximages the target

$$\eta \in \arg\min_\theta D(\pi\|\eta_\theta)$$

CHALLENGES WITH VARIATIONAL INFERENCE

$$\eta \in \arg\min_{\theta} D(\pi \| \eta_\theta)$$

CHALLENGES WITH VARIATIONAL INFERENCE

$$\eta \in \arg\min_{\theta} D(\pi \| \eta_\theta)$$

- **Challenge 1:** How do you design the variational family?

CHALLENGES WITH VARIATIONAL INFERENCE

$$\eta \in \arg\min_{\theta} D(\pi \| \eta_\theta)$$

- **Challenge 1:** How do you design the variational family?
  - Extremely problem-specific

CHALLENGES WITH VARIATIONAL INFERENCE

$$\eta \in \arg\min_{\theta} D(\pi \| \eta_\theta)$$

- **Challenge 1:** How do you design the variational family?
  - Extremely problem-specific
  - There is a tradeoff between how flexible your family is and how easy it is to learn

CHALLENGES WITH VARIATIONAL INFERENCE

$$\eta \in \arg\min_{\theta} D(\pi \| \eta_\theta)$$

- **Challenge 1:** How do you design the variational family?
  - Extremely problem-specific
  - There is a tradeoff between how flexible your family is and how easy it is to learn
  - The more flexible the harder it is to optimise over

CHALLENGES WITH VARIATIONAL INFERENCE

$$\eta \in \arg\min_{\theta} D(\pi \| \eta_\theta)$$

- **Challenge 1:** How do you design the variational family?
  - Extremely problem-specific
  - There is a tradeoff between how flexible your family is and how easy it is to learn
  - The more flexible the harder it is to optimise over
- **Challenge 2:** How do you choose the discrepancy D?

CHALLENGES WITH VARIATIONAL INFERENCE

$$\eta \in \arg\min_{\theta} D(\pi \| \eta_\theta)$$

- **Challenge 1:** How do you design the variational family?
  - Extremely problem-specific
  - There is a tradeoff between how flexible your family is and how easy it is to learn
  - The more flexible the harder it is to optimise over
- **Challenge 2:** How do you choose the discrepancy D?
  - Eg. Total variation, forward/reverse KL, Wasserstein, MMD, α-divergence, f-divergence, etc

CHALLENGES WITH VARIATIONAL INFERENCE

$$\eta \in \arg\min_{\theta} D(\pi \| \eta_\theta)$$

- **Challenge 1:** How do you design the variational family?
  - Extremely problem-specific
  - There is a tradeoff between how flexible your family is and how easy it is to learn
  - The more flexible the harder it is to optimise over
- **Challenge 2:** How do you choose the discrepancy D?
  - Eg. Total variation, forward/reverse KL, Wasserstein, MMD, α-divergence, f-divergence, etc
  - Each have their own pros and cons and measure different things

CHALLENGES WITH VARIATIONAL INFERENCE

$$\eta \in \arg\min_{\theta} D(\pi \| \eta_\theta)$$

- **Challenge 1:** How do you design the variational family?
  - Extremely problem-specific
  - There is a tradeoff between how flexible your family is and how easy it is to learn
  - The more flexible the harder it is to optimise over

- **Challenge 2:** How do you choose the discrepancy D?
  - Eg. Total variation, forward/reverse KL, Wasserstein, MMD, α-divergence, f-divergence, etc
  - Each have their own pros and cons and measure different things

- **Challenge 3:** In general, the optimisations can be fragile with limited target information

CHALLENGES WITH VARIATIONAL INFERENCE

$$\eta \in \arg\min_{\theta} D(\pi \| \eta_\theta)$$

- **Challenge 1:** How do you design the variational family?
  - Extremely problem-specific
  - There is a tradeoff between how flexible your family is and how easy it is to learn
  - The more flexible the harder it is to optimise over

- **Challenge 2:** How do you choose the discrepancy D?
  - Eg. Total variation, forward/reverse KL, Wasserstein, MMD, α-divergence, f-divergence, etc
  - Each have their own pros and cons and measure different things

- **Challenge 3:** In general, the optimisations can be fragile with limited target information
  - The more flexible the harder it is to optimise

CHALLENGES WITH VARIATIONAL INFERENCE

$$\eta \in \arg\min_{\theta} D(\pi \| \eta_\theta)$$

- **Challenge 1:** How do you design the variational family?
  - Extremely problem-specific
  - There is a tradeoff between how flexible your family is and how easy it is to learn
  - The more flexible the harder it is to optimise over

- **Challenge 2:** How do you choose the discrepancy D?
  - Eg. Total variation, forward/reverse KL, Wasserstein, MMD, α-divergence, f-divergence, etc
  - Each have their own pros and cons and measure different things

- **Challenge 3:** In general, the optimisations can be fragile with limited target information
  - The more flexible the harder it is to optimise
  - Design choices traditionally appeal towards stability over flexibility

CHALLENGES WITH VARIATIONAL INFERENCE

$$\eta \in \arg\min_{\theta} D(\pi \| \eta_\theta)$$

- **Challenge 1:** How do you design the variational family?
  - Extremely problem-specific
  - There is a tradeoff between how flexible your family is and how easy it is to learn
  - The more flexible the harder it is to optimise over

- **Challenge 2:** How do you choose the discrepancy D?
  - Eg. Total variation, forward/reverse KL, Wasserstein, MMD, α-divergence, f-divergence, etc
  - Each have their own pros and cons and measure different things

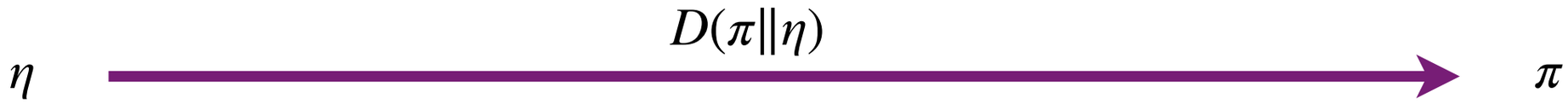
- **Challenge 3:** In general, the optimisations can be fragile with limited target information
  - The more flexible the harder it is to optimise
  - Design choices traditionally appeal towards stability over flexibility

- **Challenge 4:** even the best reference may not be close enough to stabilise the acceptance

ANNEALING

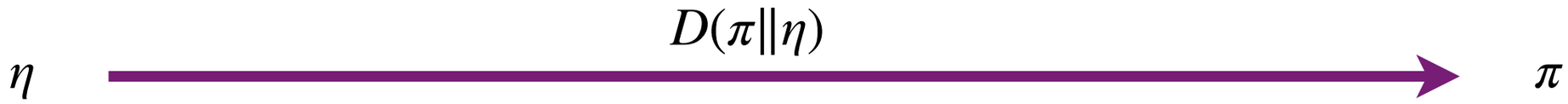
ANNEALING

- Given a divergence D and a reference η and target π



ANNEALING

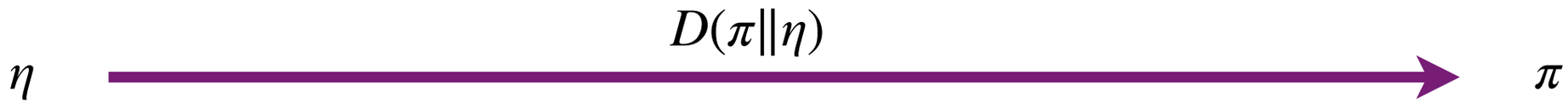
- Given a divergence D and a reference η and target π



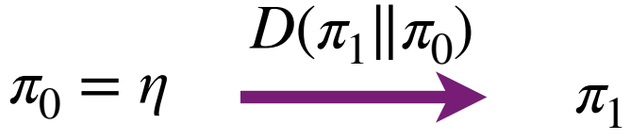
- **Solution 2:** Keep the reference fixed and modify the target

ANNEALING

- Given a divergence D and a reference η and target π



- **Solution 2:** Keep the reference fixed and modify the target
  - We can effieciently propagate inferences from $\eta = \pi_0$ to π_1 if $D(\pi_1\|\pi_0)$ is small



ANNEALING

- Given a divergence D and a reference η and target π

$$\eta \xrightarrow{D(\pi\|\eta)} \pi$$

- **Solution 2:** Keep the reference fixed and modify the target

  - We can effieciently propagate inferences from $\eta = \pi_0$ to π_1 if $D(\pi_1\|\pi_0)$ is small

$$\pi_0 = \eta \xrightarrow{D(\pi_1\|\pi_0)} \pi_1$$

  - We can then efficiently propogate inference from π_1 to π_2 close to π_1

$$\pi_0 = \eta \xrightarrow{D(\pi_1\|\pi_0)} \pi_1 \xrightarrow{D(\pi_2\|\pi_1)} \pi_2$$

ANNEALING

- Given a divergence D and a reference η and target π

$$\eta \xrightarrow{D(\pi\|\eta)} \pi$$

- **Solution 2:** Keep the reference fixed and modify the target

  - We can effieciently propagate inferences from $\eta = \pi_0$ to π_1 if $D(\pi_1\|\pi_0)$ is small

$$\pi_0 = \eta \xrightarrow{D(\pi_1\|\pi_0)} \pi_1$$

  - We can then efficiently propogate inference from π_1 to π_2 close to π_1

$$\pi_0 = \eta \xrightarrow{D(\pi_1\|\pi_0)} \pi_1 \xrightarrow{D(\pi_2\|\pi_1)} \pi_2$$

  - Can repeat until we reach the target

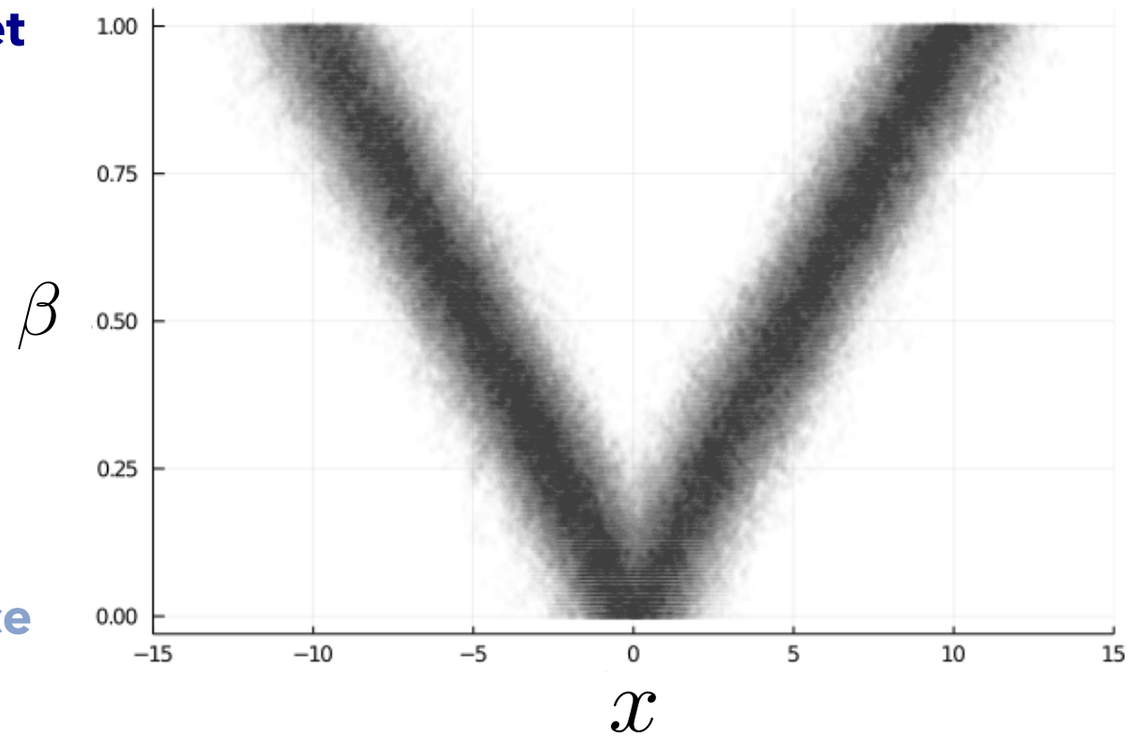
$$\pi_0 = \eta \xrightarrow{D(\pi_1\|\pi_0)} \pi_1 \xrightarrow{D(\pi_2\|\pi_1)} \pi_2 \quad \cdots \quad \pi_{N-1} \xrightarrow{D(\pi_N\|\pi_{N-1})} \pi_N = \pi$$

ANNEALING

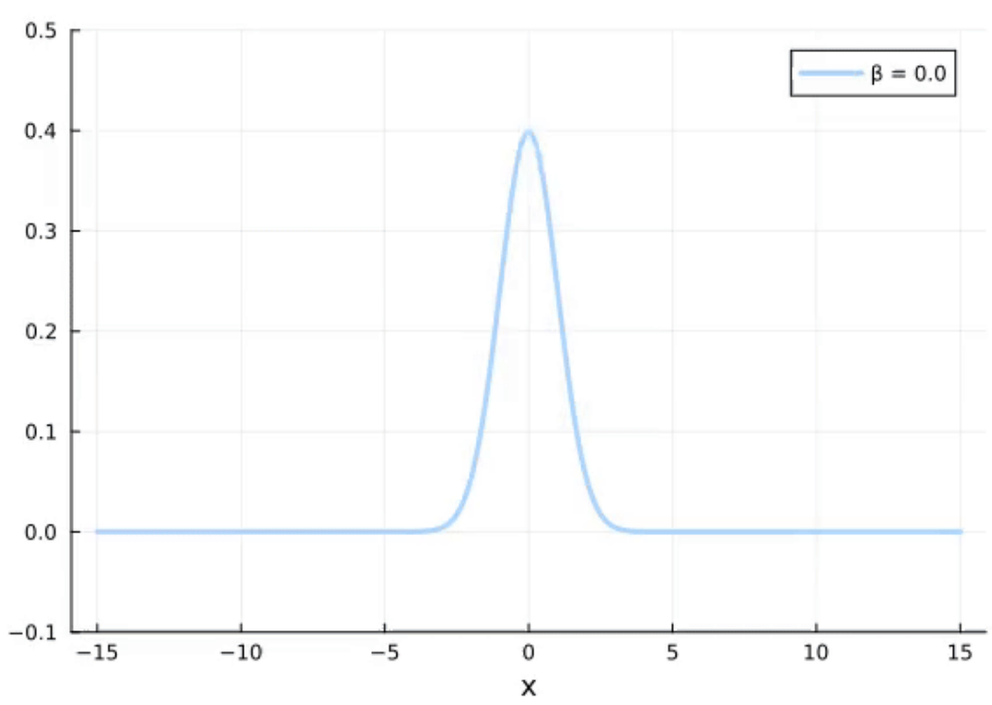
ANNEALING

- For $\beta \in [0,1]$ suppose we hav distributions π_β such that $\pi_0 = \eta$ and $\pi_1 = \pi$
  - π_β is the **annealing distribution** corresponding to the **annealing parameter** β

Target



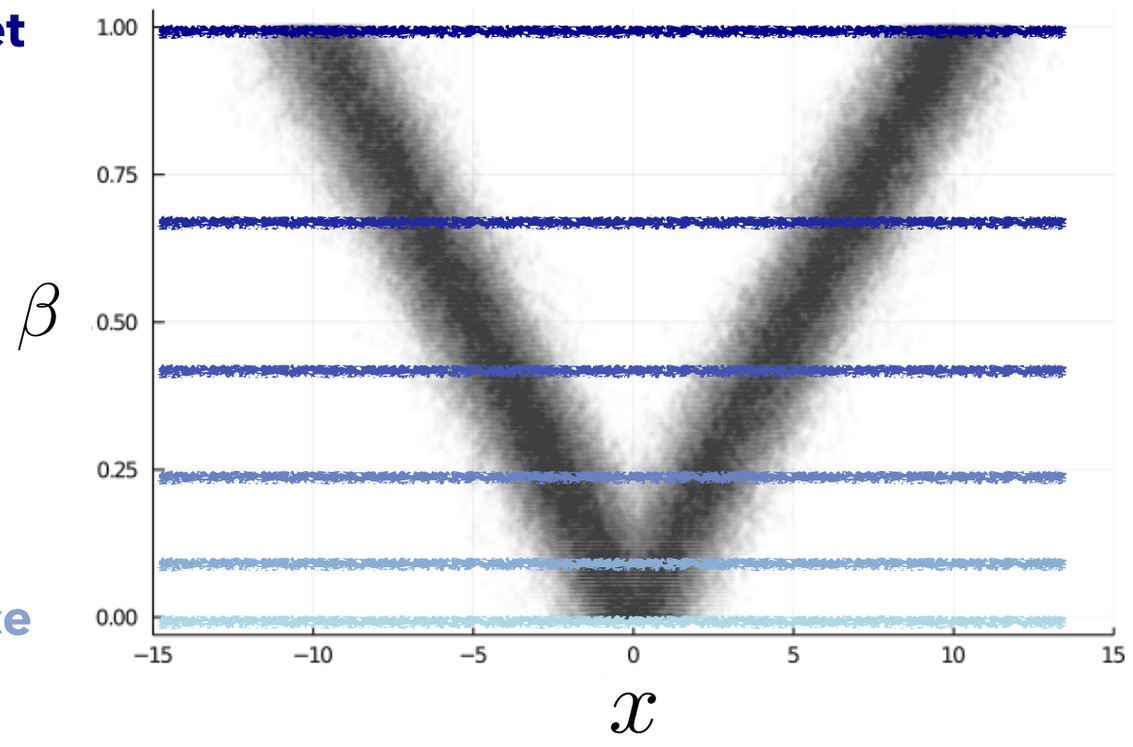
Reference



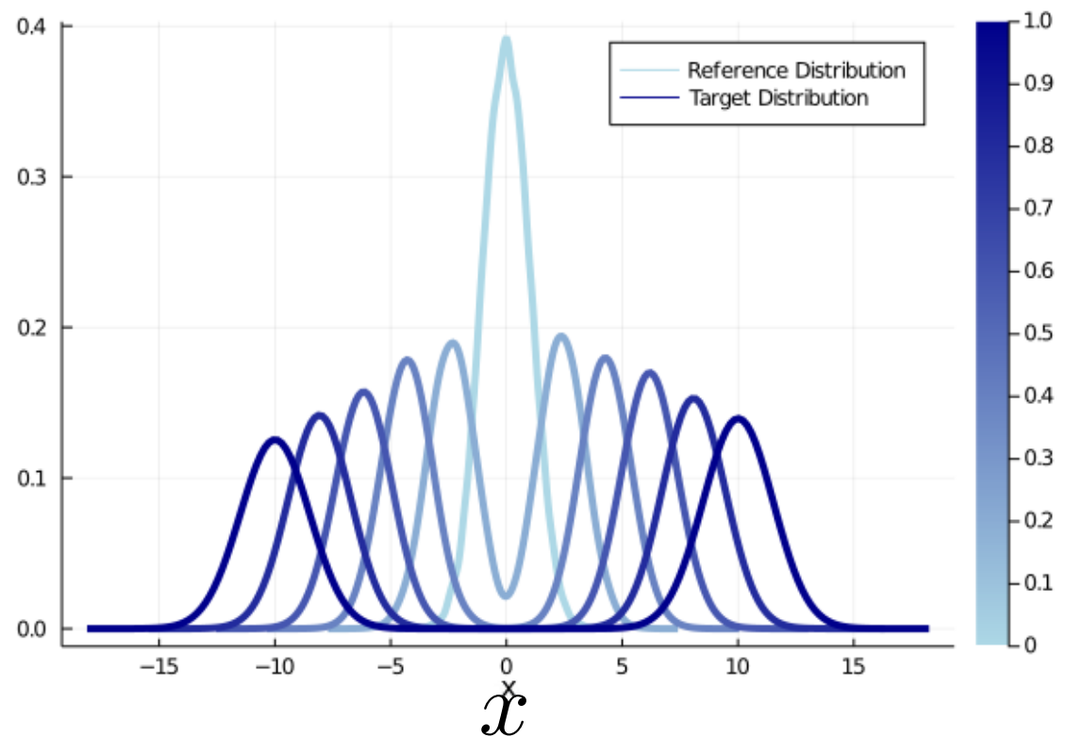
ANNEALING

- For $\beta \in [0,1]$ suppose we hav distributions π_β such that $\pi_0 = \eta$ and $\pi_1 = \pi$
  - π_β is the **annealing distribution** corresponding to the **annealing parameter** β
- We can construct our sequence of annealing distributions by setting $\pi_n = \pi_{\beta_n}$

Target



Reference



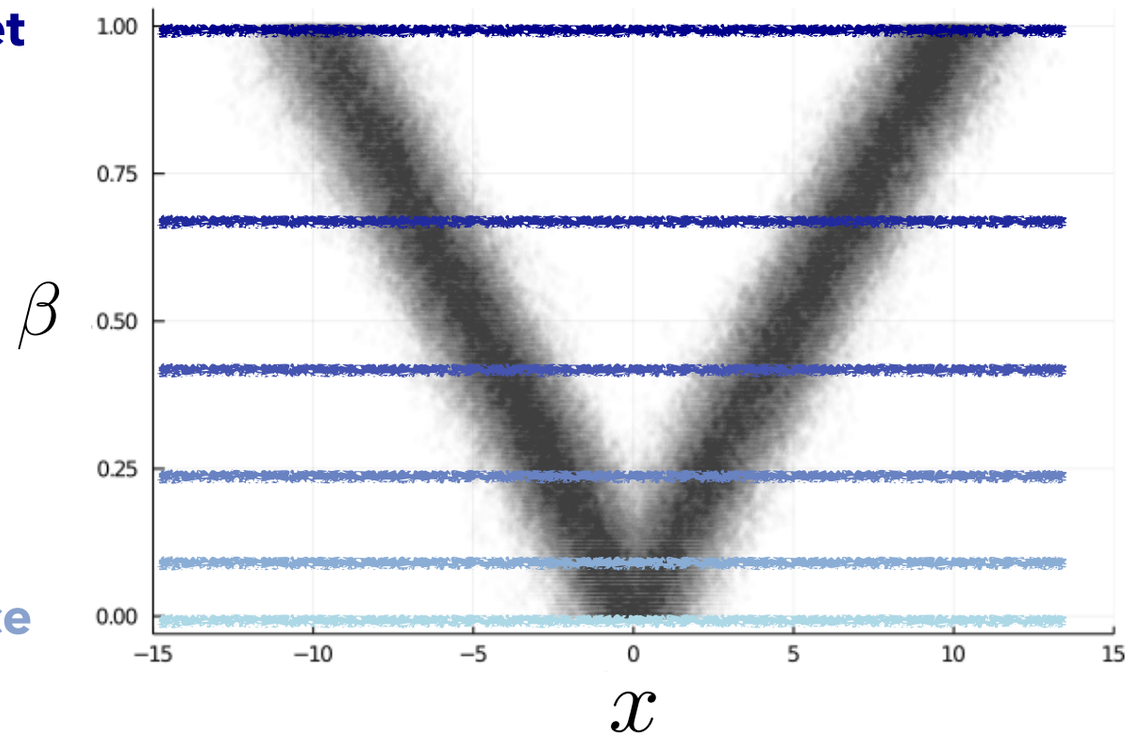
ANNEALING

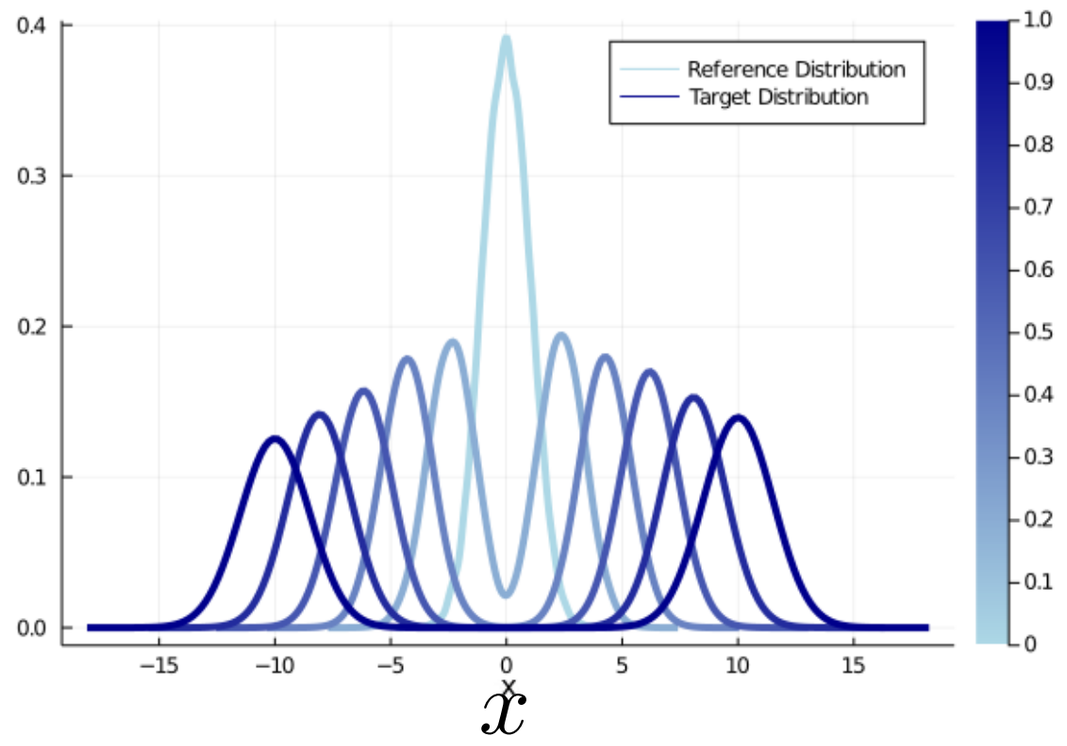
- For $\beta \in [0,1]$ suppose we hav distributions π_β such that $\pi_0 = \eta$ and $\pi_1 = \pi$
  - π_β is the **annealing distribution** corresponding to the **annealing parameter** β
- We can construct our sequence of annealing distributions by setting $\pi_n = \pi_{\beta_n}$
  - Where $\mathscr{B} = \beta_{0:N}$ is the **annealing schedule** satisfying:

$$0 = \beta_0 < \beta_1 < \cdots < \beta_N = 1$$

Target

Reference



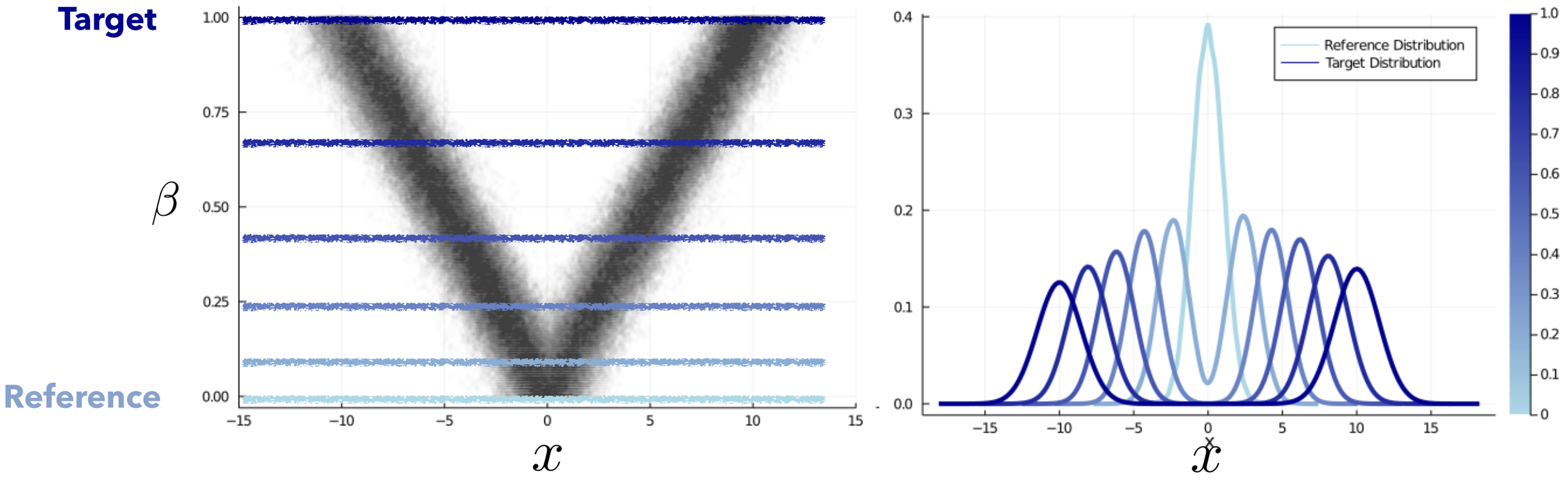


ANNEALING

- For $\beta \in [0,1]$ suppose we hav distributions π_β such that $\pi_0 = \eta$ and $\pi_1 = \pi$
  - π_β is the **annealing distribution** corresponding to the **annealing parameter** β
- We can construct our sequence of annealing distributions by setting $\pi_n = \pi_{\beta_n}$
  - Where $\mathscr{B} = \beta_{0:N}$ is the **annealing schedule** satisfying:

$$0 = \beta_0 < \beta_1 < \cdots < \beta_N = 1$$

- Note that π_{n-1} and π_n are close the schedule if fine and $\beta, \beta' \mapsto D(\pi_{\beta'} \| \pi_\beta)$ is continuous

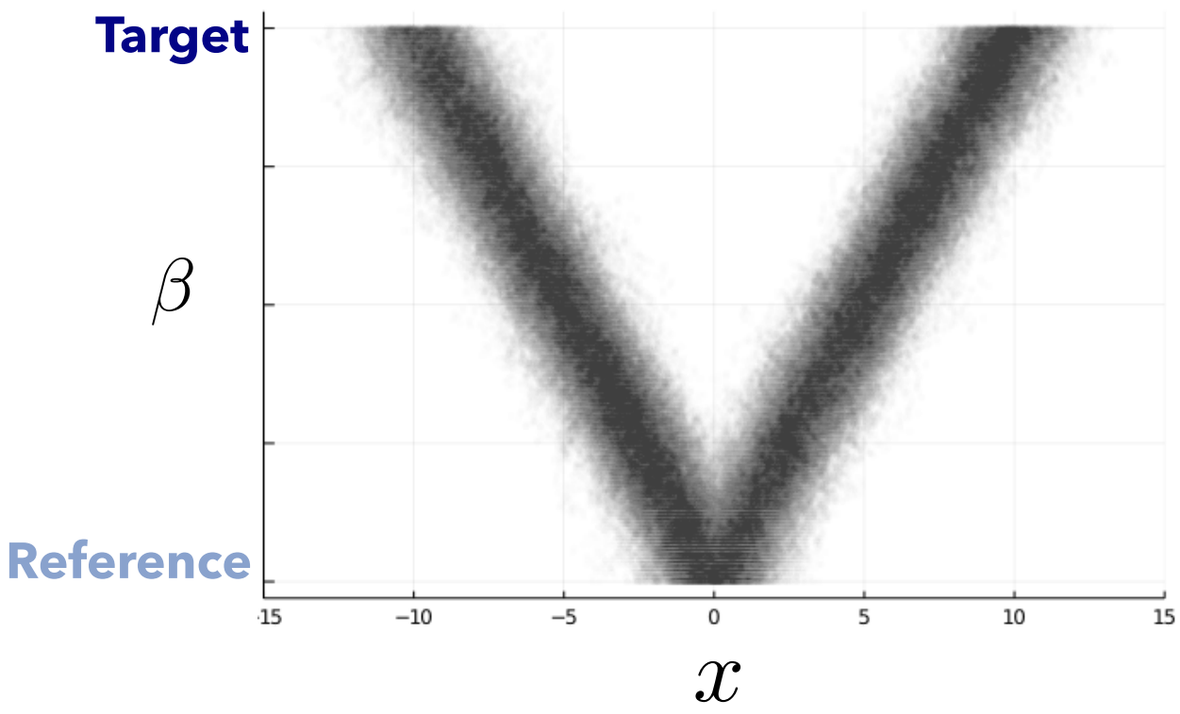


ANNEALING ALGORITHMS

- Annealing is **not** an algorithm but a platform to build algorithms on top of

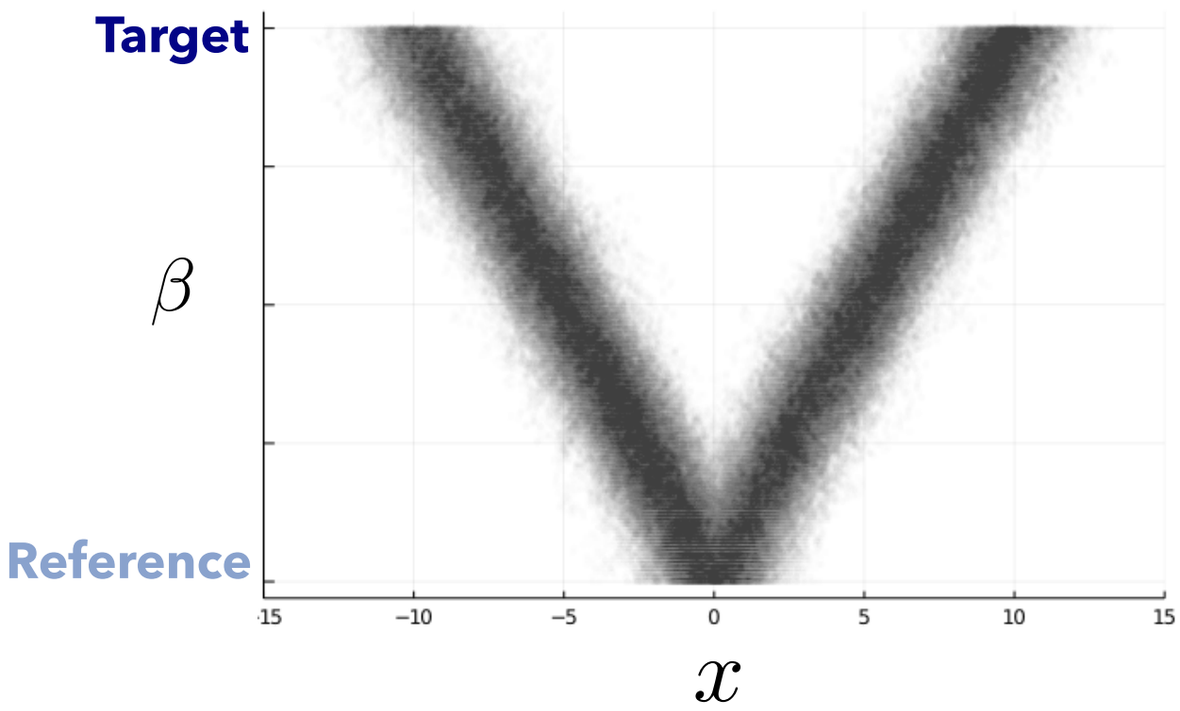
ANNEALING ALGORITHMS

- Annealing is **not** an algorithm but a platform to build algorithms on top of



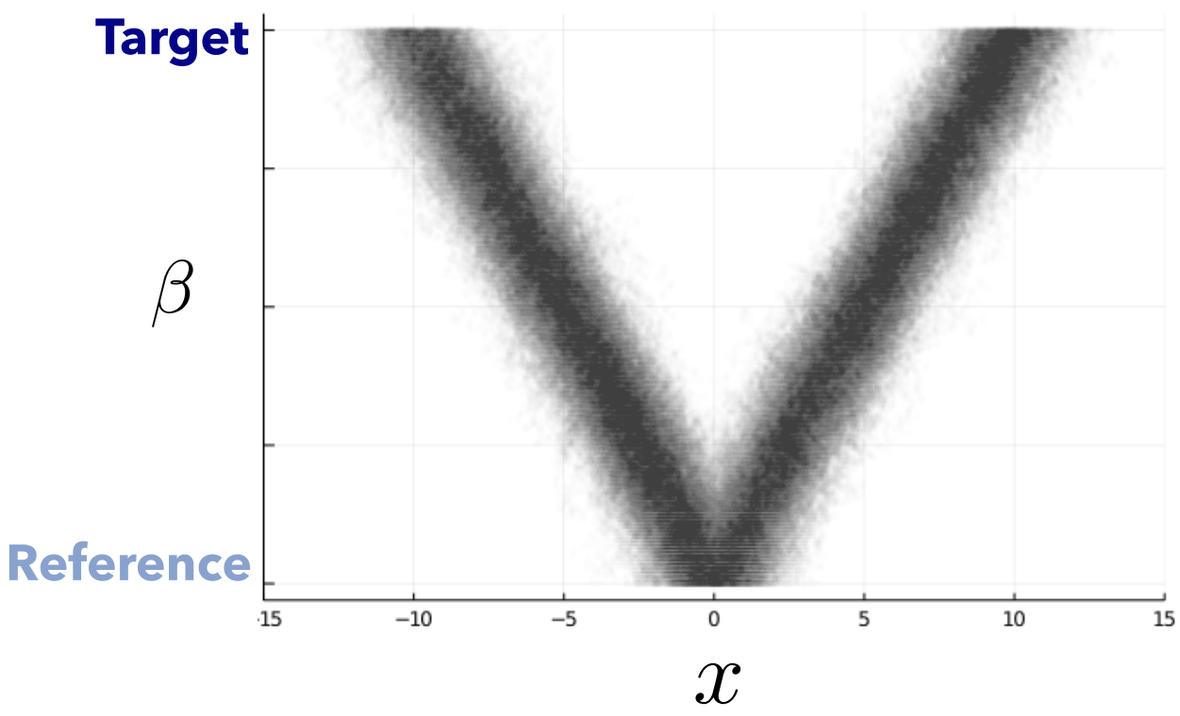
ANNEALING ALGORITHMS

- Annealing is **not** an algorithm but a platform to build algorithms on top of
- **Annealing algorithms** draw inference from the entire path not just the target



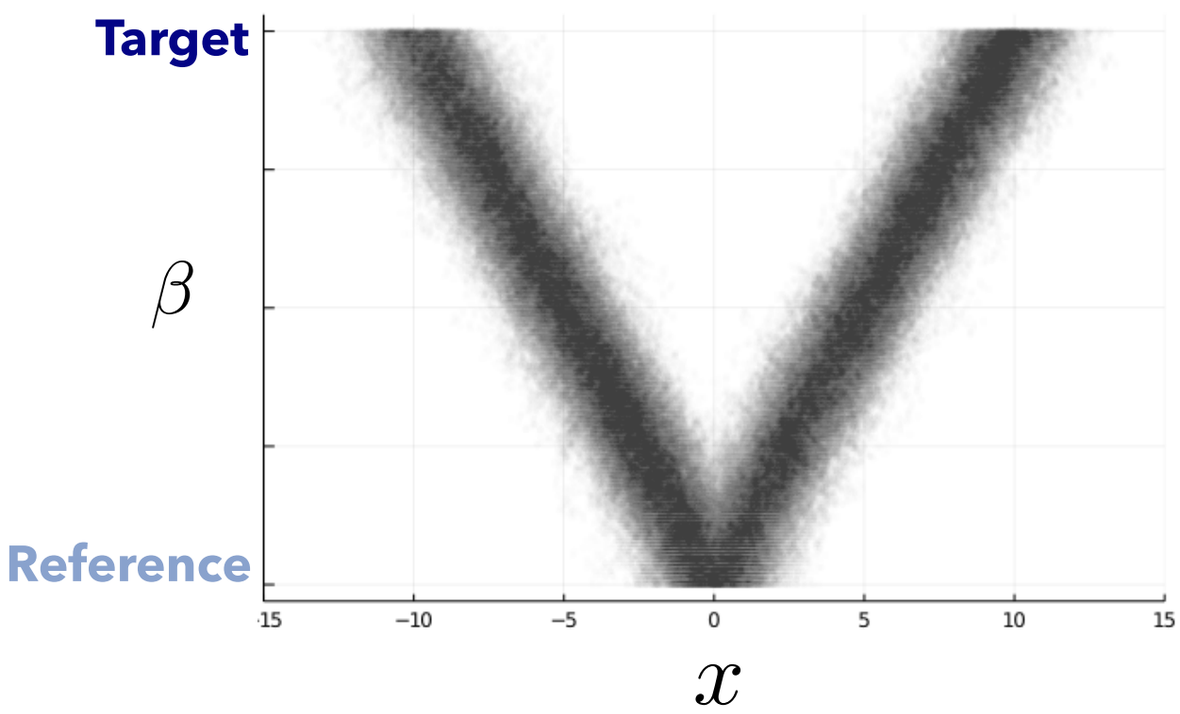
ANNEALING ALGORITHMS

- Annealing is **not** an algorithm but a platform to build algorithms on top of
- **Annealing algorithms** draw inference from the entire path not just the target
- Annealing algorithms are meta algorithms:



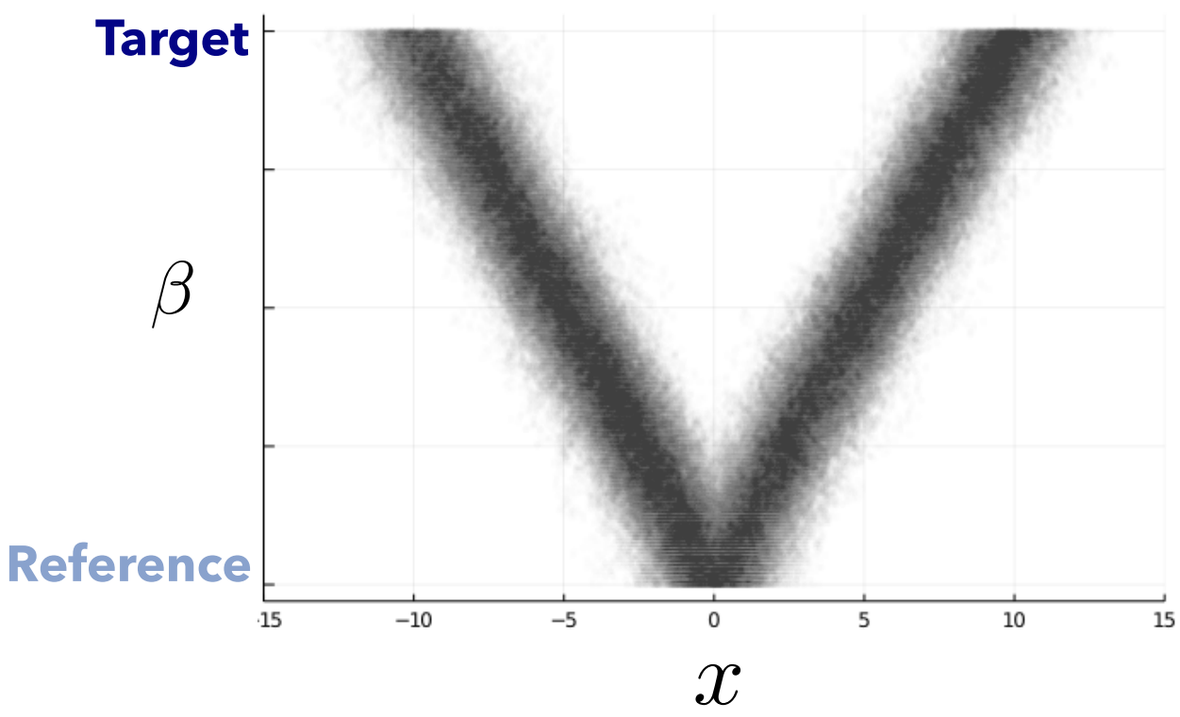
ANNEALING ALGORITHMS

- Annealing is **not** an algorithm but a platform to build algorithms on top of
- **Annealing algorithms** draw inference from the entire path not just the target
- Annealing algorithms are meta algorithms:
  - **Input:** An annealing path π_β + a local efficient inference algorithm for each π_β



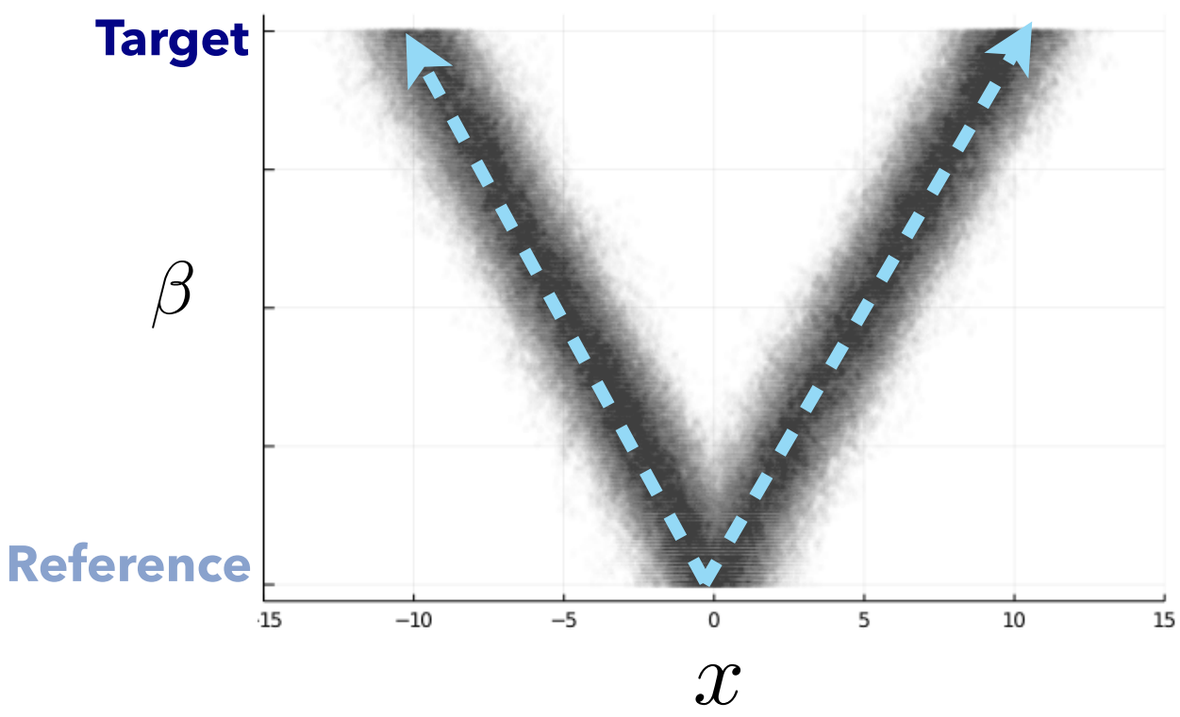
ANNEALING ALGORITHMS

- Annealing is **not** an algorithm but a platform to build algorithms on top of
- **Annealing algorithms** draw inference from the entire path not just the target
- Annealing algorithms are meta algorithms:
  - **Input:** An annealing path π_β + a local efficient inference algorithm for each π_β
  - **Output:** a globally efficient inference algorithm for the path $\beta \mapsto \pi_\beta$



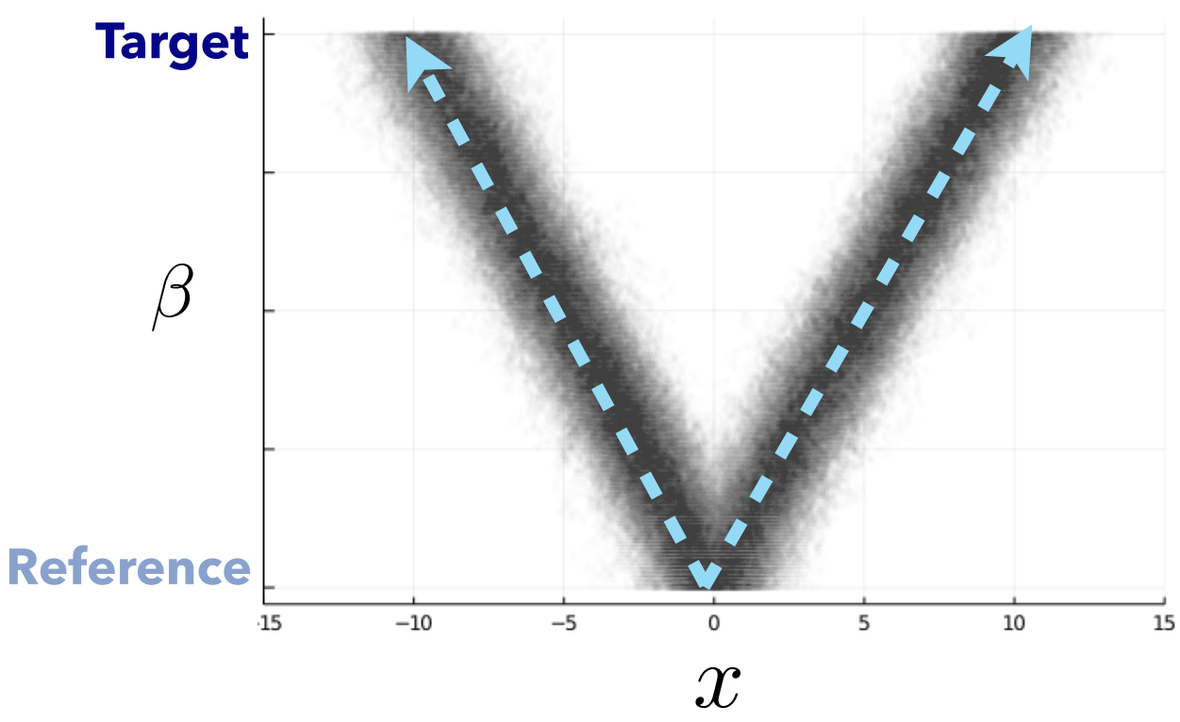
ANNEALING ALGORITHMS

- Annealing is **not** an algorithm but a platform to build algorithms on top of
- **Annealing algorithms** draw inference from the entire path not just the target
- Annealing algorithms are meta algorithms:
  - **Input:** An annealing path π_β + a local efficient inference algorithm for each π_β
  - **Output:** a globally efficient inference algorithm for the path $\beta \mapsto \pi_\beta$
- Transform a d-dimensional multi-modal target into a $d+1$-dimensional unimodal one



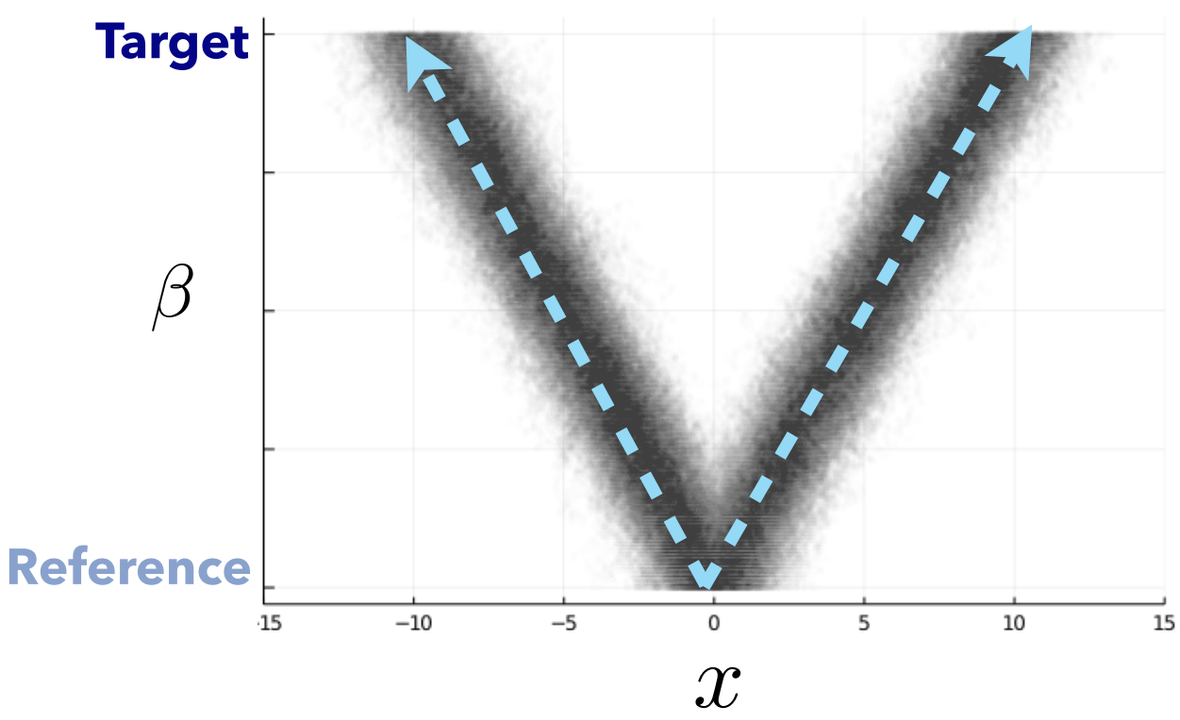
ANNEALING ALGORITHMS

- Annealing is **not** an algorithm but a platform to build algorithms on top of
- **Annealing algorithms** draw inference from the entire path not just the target
- Annealing algorithms are meta algorithms:
  - **Input:** An annealing path π_β + a local efficient inference algorithm for each π_β
  - **Output:** a globally efficient inference algorithm for the path $\beta \mapsto \pi_\beta$
- Transform a d-dimensional multi-modal target into a $d+1$-dimensional unimodal one
- They achieve state-of-the-art performance for hard mult-modal distributions



ANNEALING ALGORITHMS

- Annealing is **not** an algorithm but a platform to build algorithms on top of
- **Annealing algorithms** draw inference from the entire path not just the target
- Annealing algorithms are meta algorithms:
  - **Input:** An annealing path π_β + a local efficient inference algorithm for each π_β
  - **Output:** a globally efficient inference algorithm for the path $\beta \mapsto \pi_\beta$
- Transform a d-dimensional multi-modal target into a $d+1$-dimensional unimodal one
- They achieve state-of-the-art performance for hard mult-modal distributions
  - Can use the reference to assess the quality and ensure robustness



HOW TO CONSTRUCT AN ANNEALING PATH

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η
- In the context of sampling, we are given unnormalised densities:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x$$

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η
- In the context of sampling, we are given unnormalised densities:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x$$

  - We build annealing distribution by curupting the densities:

$$\pi_\beta = \frac{\gamma_\beta(x)}{Z(\beta)}, \qquad Z(\beta) = \int_{\mathbb{X}} \gamma_\beta(x)\mathrm{d}x$$

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η
- In the context of sampling, we are given unnormalised densities:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x$$

  - We build annealing distribution by curupting the densities:

$$\pi_\beta = \frac{\gamma_\beta(x)}{Z(\beta)}, \qquad Z(\beta) = \int_{\mathbb{X}} \gamma_\beta(x)\mathrm{d}x$$

- $\gamma_\beta : \mathbb{X} \to \mathbb{R}$ is the un-normalised density satisfying $\gamma_0 = \eta$ and $\gamma_1 = \gamma$

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η
- In the context of sampling, we are given unnormalised densities:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x$$

  - We build annealing distribution by curupting the densities:

$$\pi_\beta = \frac{\gamma_\beta(x)}{Z(\beta)}, \qquad Z(\beta) = \int_{\mathbb{X}} \gamma_\beta(x)\mathrm{d}x$$

- $\gamma_\beta : \mathbb{X} \to \mathbb{R}$ is the un-normalised density satisfying $\gamma_0 = \eta$ and $\gamma_1 = \gamma$
- Normalinging constants $Z(\beta)$ satisfies $Z(0) = 1$ and $Z(1) = Z$

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η
- In the context of sampling, we are given unnormalised densities:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x$$

  - We build annealing distribution by curupting the densities:

$$\pi_\beta = \frac{\gamma_\beta(x)}{Z(\beta)}, \qquad Z(\beta) = \int_{\mathbb{X}} \gamma_\beta(x)\mathrm{d}x$$

    - $\gamma_\beta : \mathbb{X} \to \mathbb{R}$ is the un-normalised density satisfying $\gamma_0 = \eta$ and $\gamma_1 = \gamma$
    - Normalinging constants $Z(\beta)$ satisfies $Z(0) = 1$ and $Z(1) = Z$
- There is a lot of flexibility in choosing a path. We will study what makes for a good path later.

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η
- In the context of sampling, we are given unnormalised densities:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x$$

  - We build annealing distribution by curupting the densities:

$$\pi_\beta = \frac{\gamma_\beta(x)}{Z(\beta)}, \qquad Z(\beta) = \int_{\mathbb{X}} \gamma_\beta(x)\mathrm{d}x$$

    - $\gamma_\beta : \mathbb{X} \to \mathbb{R}$ is the un-normalised density satisfying $\gamma_0 = \eta$ and $\gamma_1 = \gamma$
    - Normalinging constants $Z(\beta)$ satisfies $Z(0) = 1$ and $Z(1) = Z$
- There is a lot of flexibility in choosing a path. We will study what makes for a good path later.
- The canonical choice is the linear path

$$\gamma_\beta(x) \propto \eta(x)^{1-\beta}\gamma(x)^\beta$$

HOW TO CONSTRUCT AN ANNEALING PATH

- Can think of the annealing path as a continuous curruption of the target π to noise η
- In the context of sampling, we are given unnormalised densities:

$$\pi(\mathrm{d}x) = \frac{\gamma(x)}{Z}\mathrm{d}x$$

  - We build annealing distribution by curupting the densities:

$$\pi_\beta = \frac{\gamma_\beta(x)}{Z(\beta)}, \qquad Z(\beta) = \int_{\mathbb{X}} \gamma_\beta(x)\mathrm{d}x$$

    - $\gamma_\beta : \mathbb{X} \to \mathbb{R}$ is the un-normalised density satisfying $\gamma_0 = \eta$ and $\gamma_1 = \gamma$
    - Normalinging constants $Z(\beta)$ satisfies $Z(0) = 1$ and $Z(1) = Z$
- There is a lot of flexibility in choosing a path. We will study what makes for a good path later.
- The canonical choice is the linear path

$$\gamma_\beta(x) \propto \eta(x)^{1-\beta}\gamma(x)^\beta$$

  - Before designing algorithms, let's build some intuition about annealing paths.