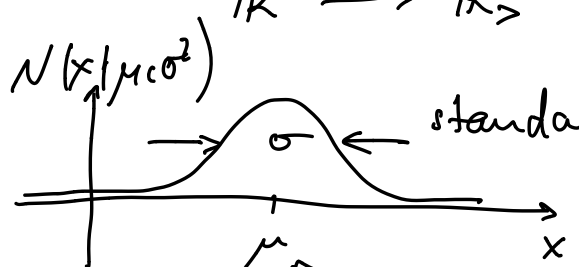
Gaussian Mixture Models

- Gaussian distribution in 1d

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$x \longmapsto \mathcal{N}(x \mid \mu, \sigma^2)$$
$$\mathbb{R} \longrightarrow \mathbb{R}_{>}$$



σ^2 : variance

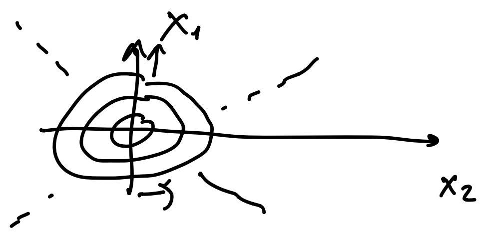
→ normalized: $\int_{-\infty}^{\infty} dx \; \mathcal{N}(x \mid \mu, \sigma^2) = 1$

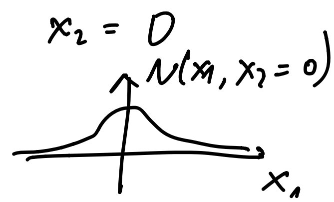
$\mathcal{N}(x) > 0 \;\; \forall x \Rightarrow$ probability distr.

→ multivariate Gaussian distribution

$\vec{x} \in \mathbb{R}^d$

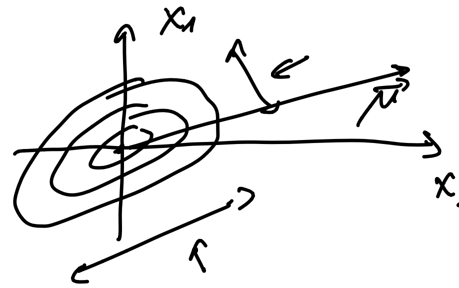
e.g. $d = 2$





one possibility: $\mathcal{N}(\vec{x} \mid \mu_1, \mu_2 ; \sigma_1^2, \sigma_2^2) = \mathcal{N}(x_1 \mid \mu_1, \sigma_1)\, \mathcal{N}(x_2 \mid \mu_2, \sigma_2)$

what if distr. is tilted / squeezed?



→ need coupling between x_1, x_2
→ no longer have product of $N \times N$

<u>def</u>: covariance matrix : $\Sigma \in \mathbb{R}^{d \times d}$, positive definite

$$\boxed{N(\vec{x} \mid \vec{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^t \Sigma^{-1} (\vec{x}-\vec{\mu})\right]}$$

$$\neq e^{-c_1 x_1^2} \times e^{-c_2 x_2^2}$$

have $e^{-\Sigma_{12} x_1 x_2}$ ← quadratic

generalizes previous example:

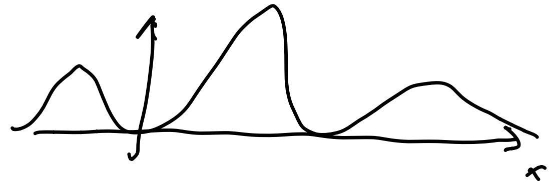
e.g. $\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$

$\vec{\mu} = (\mu_1, \mu_2)^t$

- Gaussian distrs. have a single / unique maximum at $\vec{x}_* = \vec{\mu}$
  - sometimes called unimodal

– Gaussian mixture models

want a distr. that can describe multiple maxima, e.g.

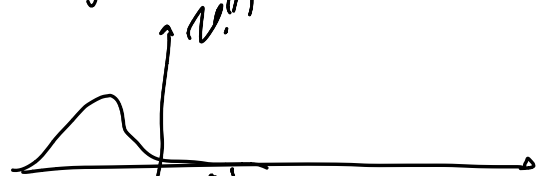


"multimodal distribution"

→ back to 1d

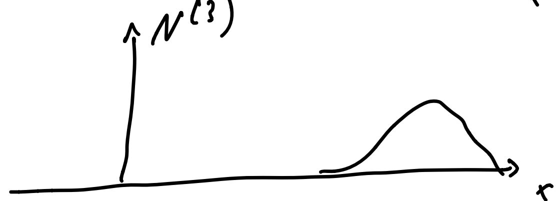
<u>idea</u>: shift different Gaussian distrs. & add them up

$$N^{(1)}(x \mid \mu^{(1)}, \sigma^{(1)}) =$$



$$N^{(2)}(x \mid \mu^{(2)}, \sigma^{(2)}) =$$

$$N^{(3)}(x \mid \mu^{(3)}, \sigma^{(3)}) =$$



$$N(x \mid \{\mu^{(i)}, \sigma^{(i)2}\}_{i=1}^{3}) = \sum_{i=1}^{3} \alpha^{(i)} N(x \mid \mu^{(i)}, \sigma^{(i)2})$$

impose normalization condition:

$$\int_{-\infty}^{\infty} dx\, N(x \mid \{\mu^{(i)}, \sigma^{(i)2}\}) = \sum_{i=1}^{3} \alpha^{(i)} \times 1 \overset{!}{=} 1$$

$$\Rightarrow \boxed{\sum_{i=1}^{K} \alpha^{(i)} = 1}, \quad K \in \mathbb{N}$$

<u>Def</u> (Gaussian mixture model) in 1d

$$N(x \mid \{\mu^{(i)}, \sigma^{(i)2}\}_{i=1}^{K}) = \sum_{i=1}^{K} \alpha^{(i)} N(x \mid \mu^{(i)}, \sigma^{(i)2})$$

→ if data $\vec{x} \in \mathbb{R}^d$, $d > 1$

$$N(\vec{x} \mid \{\vec{\mu}^{(i)}, \Sigma^{(i)}\}_{i=1}^{K}) = \sum_{i=1}^{K} \alpha^{(i)} N(\vec{x} \mid \vec{\mu}^{(i)}, \Sigma^{(i)})$$

where $\vec{\mu}^{(i)} \in \mathbb{R}^d$ & $\Sigma^{(i)} \in \mathbb{R}^{d \times d}$ pos. def. $\forall i = 1, \ldots, K$

& $\sum_{i=1}^{K} \alpha^{(i)} = 1$

— Applications of Gaussian mixtures

1) clustering : e.g. K-mean, density estimation

2) as a variational ansatz for a more complex prob. distr.

→ approximate with Gaussian mixture

learn variational parameters:

$\vec{\mu}^{(i)}, \Sigma^{(i)}, \alpha^{(i)}$

Markov Processes

- **def**: a stochastic (or random) process which satisfies the Markov property:

→ Markov property / "Markovianity"
- predictions regarding future outcomes depend only on current state and NOT on the process history
- "memorylessness"
- can make future predictions without knowing the history

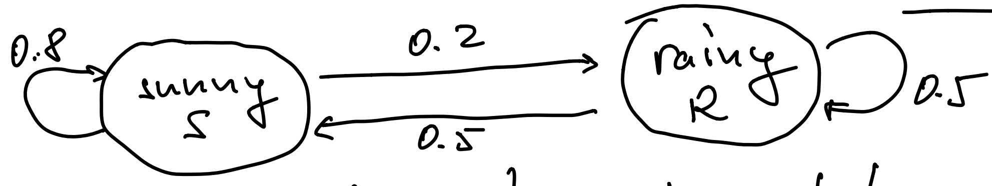
→ **formally**: let $X_1, X_2, \dots,$ be random variables

$X_1, X_2, \dots$ satisfy the Markov property, if

$$P(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P(X_{n+1} = x \mid X_n = x_n)$$

- **Example**: weather model

Markov chain



$X_i \in \{"S", "R"\}$ weather state

e.g. $X_0, X_1, X_2, \dots = S, R, R, S$ etc.

$P(X_1 = S \mid X_0 = S) = 0.8 \qquad P(X_1 = R \mid X_0 = S) = 0.2$

$P(X_1 = S \mid X_0 = R) = 0.5 \qquad P(X_1 = R \mid X_0 = R) = 0.5$

→ <u>transition matrix</u>:

$$P = \begin{array}{cc} S & R \\ \left(\begin{array}{cc} 0.8 & 0.2 \\ 0.5 & 0.5 \end{array}\right) \end{array} \begin{array}{c} S \\ R \end{array}$$

→ governs the dynamics of Markov chain

<u>e.g.</u> today is raining. $\Rightarrow X_0 = R$

what is the probability to have a rainy day in one week?

$\vec{x}_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \Leftrightarrow R$ today

→ distr. after 7 days: $P^7 = \overbrace{P \cdot P \dots P}^{7 \text{ times}}$

$$\vec{x}_7 = \vec{x}_0^{\,t} P^7 = (0, 1) \begin{pmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{pmatrix}^7 \approx \begin{pmatrix} 0.7141 \\ 0.2859 \end{pmatrix}$$

<u>eg.</u> today is sunny: $\vec{x}_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$$\vec{x}_7 = (1, 0) P^7 \approx \begin{pmatrix} 0.7143 \\ 0.2857 \end{pmatrix}$$

almost the same outcome irrespective of initial condition

– stationary distribution: (of a Markov chain)

$$\boxed{\vec{\pi}^t P = \vec{\pi}}$$ fix point distribution

→ solve for $\vec{\pi}$: $\vec{\pi} = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix}$

$$(\pi_1, \pi_2)\left[\begin{pmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right] = 0$$

$$(\pi_1, \pi_2)\begin{pmatrix} -0.2 & 0.2 \\ 0.5 & -0.5 \end{pmatrix} = 0$$

$$\Rightarrow \quad -0.2\pi_1 + 0.5\pi_2 = 0$$

$$\left.\begin{array}{l} \rightarrow \quad \pi_2 = 0.4\pi_1 \\ \text{normalization: } \pi_1 + \pi_2 \overset{!}{=} 1 \end{array}\right\} \Rightarrow \vec{\pi} = \begin{pmatrix} 5/7 \\ 2/7 \end{pmatrix}$$

<u>def</u>: $$\boxed{\lim_{n\to\infty} \vec{x}_0^{\,t} P^n = \vec{\pi}}$$

$\vec{\pi}$: stable long-time attractor, i.e. dynamics of Markov chain eventually approaches $\vec{\pi}$

– Markov process: transition prob. (P matrix)

$$S_0 \xrightarrow{P} S_1 \xrightarrow{P} S_2 \xrightarrow{P} \dots$$

states (e.g. "S", "R")

Ising Models

- spin configurations/states
  - binary variable $s_j \in \{\pm 1\}$

e.g. $S = \overbrace{+ - - - + + - + + - - \ldots -}^{=L}$

$= \uparrow \downarrow \downarrow \downarrow \uparrow \uparrow \downarrow \uparrow \uparrow \downarrow \downarrow \ldots \downarrow$

$= 1\ 0\ 0\ 0\ 1\ 1\ 0\ 1\ 1\ 0\ 0 \ldots 0$

applications of Ising models:

- unsupervised learning: energy-based models (Restricted Boltzmann machines (RBMs), Variational Autoencoders (VAEs)) ..
- bar code / morse code
- magnets, electron spin, etc ...

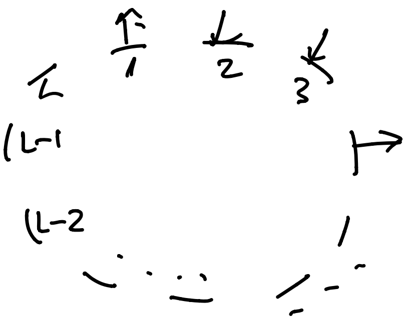
$s = \{ S_j \}_{j=1}^{L}$, $j = 1, \ldots, L$: lattice sites

open 1d lattice: $\underline{\uparrow}\ \underline{\uparrow}\ \underline{\downarrow}\ \underline{\uparrow}\ \_\ \_\ \_\ \_\ \_$

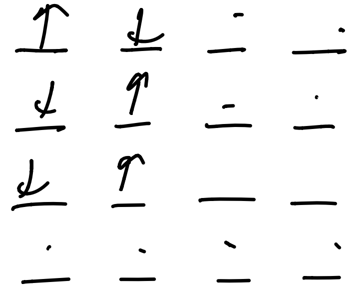
sites

$j = 1 \quad 2 \quad 3 \quad \ldots \quad L$

closed/periodic lattice:



· 2d lattice; e.g. square:
open



· 2d lattice with periodic boundary conditions:

- <u>def</u>: Ising model: a function (energy)

$$H : \{\pm 1\}^L \longrightarrow \mathbb{R}$$

$$S \longmapsto H(S)$$

<u>examples</u>:

→ single-body: $H(S)$ is linear in S

$$H(S) = \sum_{j=1}^{L} h_j S_j \quad , \; h_j \in \mathbb{R}$$

→ two-body / interacting

$$H(S) = \sum_{i=1}^{L} \sum_{j=1}^{L} J_{ij} S_i S_j + \sum_i h_j S_j$$

$J_{ij} \in \mathbb{R}$, $h_j \in \mathbb{R}$

→ measure correlations between spins

→ multi-body:

$$H(S) = \sum_{i_1, \dots, i_k = 1}^{L} J_{i_1, i_2, \dots, i_k} S_{i_1} S_{i_2} \cdots S_{i_k}$$

- ground state configuration: S_{GS}

$H(S_{GS}) \leq H(S) \quad \forall S$

S_{GS} minimizes the energy $H(S)$

$E_{GS} = H(S_{GS})$: ground state energy

—

→ a system can be in any config. S with probability $p(S)$

- which configuration is more likely?

<u>Boltzman distribution</u>: $p(S) := \dfrac{e^{-\beta H(S)}}{Z}$

$\beta = 1/T$ inverse temperature

$H(s)$: energy function

Z : partition function / sum [normalization constant]

$$\sum_{\{s\}} \frac{e^{-\beta H(s)}}{Z} = 1 \iff Z_\beta := \sum_{\{s\}} e^{-\beta H(s)}$$

↑ all possible configurations $\longleftrightarrow$ N lattice sites

2^N configurations

→ exponentially many !!!

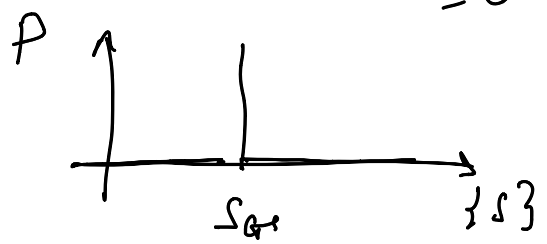
— role of temperature T $(\beta = 1/T)$

1) zero temperature : $T = 0 \iff \beta = \infty$

$$p(s) \sim e^{-\beta H(s)} = \frac{e^{-\beta(H(s) - E_{GS})}}{e^{\beta E_{GS}}} \sim$$

↘ put in Z

$$\sim e^{-\beta \underbrace{(H(s) - E_{GS})}_{\geq 0}} \overset{\beta = \infty}{=} \begin{cases} 1, & \text{if } s = s_{GS} \\ 0, & \text{otherwise} \end{cases}$$



2) infinite temperature : $T = \infty \iff \beta = 0$

$$p(S) \sim e^{-\beta H(S)} \overset{\beta \to 0}{\sim} 1$$

→ system can be in any state S with equal probability.

$$p(S) := \frac{1}{2^N}$$

3) intermediate temperature: $0 < \beta < \infty$

→ there is a <u>finite</u> prob. for the system to occupy <u>any</u> state

→ which states are more likely?

→ how can we find likely-to-occur states?

Monte Carlo Methods

- <u>broad def</u>: a class of computational algos that rely on repeated random sampling
- widely used in
  - optimization (→ RL)
  - generating samples from a prob. distr.
  - numerical (high-dimensional) integration
- <u>Markov Chain Monte Carlo</u> (MCMC)

<u>idea</u>: design a Markov chain, whose stationary distr. π is the target distribution p which we want to sample

→ obtain a sample from p by recording the states in the Markov chain

$$X_{t-1} \xrightarrow[p(x_t | x_{t-1})]{} X_t \xrightarrow[p(x_{t+1} | x_t)]{} X_{t+1} \quad \dots$$

$P(x|x')$: transition probability

1) existence of π: sufficient (not necessary) condition is detailed balance:
for every pair x, x' the transition $x \to x'$ is reversible;
in other words: the prob. of being in x & transitioning to x' is equal to the prob. of being in x' & transitioning to x: $x' \to x$.

$$\boxed{\underbrace{P(x'|x)}_{\text{transiton } x \to x'}\underbrace{\pi(x)}_{\text{being in } x} = P(x|x')\pi(x')}$$

2) uniqueness of π: is guaranteed if every state is:

a) aperiodic: system does not return to same state at fixed time intervals

b) positive recurrent: system will revisit the state in finite number of steps

Metropolis-Hastings algorithm:

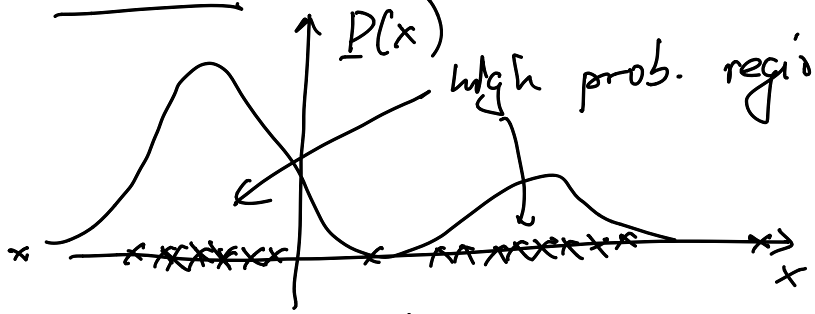
- want to sample from $\underline{P}(x)$
- assume we know $f(x) \sim \underline{P}(x)$

e.g. $\underline{P}(x) = \frac{1}{Z} e^{-\beta H(x)}$

but Z is not known

yet, $f(x) = e^{-\beta H(x)} \sim \underline{P}(x)$

→ intuition:



want sample from $\underline{P}$: $\{x_1, x_2, x_3, \dots, \}$

- Algorithm:

detailed balance: $p(x'|x)\, \underline{P}(x) = p(x|x')\, \underline{P}(x')$

$$\Rightarrow \frac{p(x'|x)}{p(x|x')} = \frac{\underline{P}(x')}{\underline{P}(x)} \underset{\underline{P} \sim f}{=} \frac{f(x')}{f(x)}$$

idea, separate transition $x \to x'$ in two steps:

1) proposal step: let $g(x'|x)$ define the conditional prob. of proposing state x' given state x

2) acceptance / rejection step:
let $A(x', x)$ be the acceptance distr. to accept the proposed state x'.

$\Rightarrow$ transition prob.: $p(x'|x) = g(x'|x)\,A(x', x)$

$\rightarrow$ plugging into detailed balance:

$$\frac{A(x', x)}{A(x, x')} = \frac{g(x|x')\,\underline{P}(x')}{g(x'|x)\,\underline{P}(x)} \qquad \text{(can replace } \underline{P} \text{ with } f\text{)}$$

$\rightarrow$ view as def. of A

$\rightarrow$ choose acceptance ratio $A(x, x')$ which satisfies this condition:

$$A(x', x) = \min\left(1, \frac{g(x|x')}{g(x'|x)} \frac{\underline{P}(x')}{\underline{P}(x)}\right)$$

either: $A(x, x') = 1$ or $A(x', x) = 1$

– pseudo code: (generates samples as Markov chain)

1. Initialize:
   a) pick any initial state x_0
   b) set $t = 0$

2. Iterate:

a) generate a random candidate state according to $g(x'|x_t)$

b) compute acceptance prob.

$$A(x', x_t) = \min\left(1, \frac{g(x_t|x')\,\underline{P}(x')}{g(x'|x_t)\,\underline{P}(x_t)}\right)$$

c) accept/reject move:

i) generate random number $\varepsilon \in [0,1]$

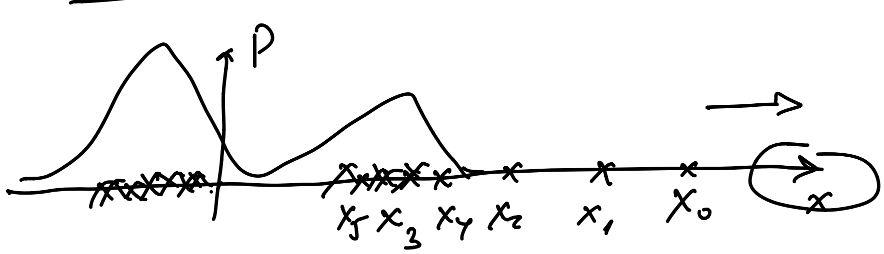
ii) if $\varepsilon \leq A(x', x_t)$ $\rightarrow$ accept state & set $x_{t+1} = x'$

iii) if $\varepsilon > A(x', x_t)$ $\rightarrow$ reject state & set $x_{t+1} = x_t$

d) increment $t \rightarrow t+1$

$\Rightarrow$ sequence $\{x_0, x_1, \dots, x_T\}$ gives an empirical sample from $\underline{P}(x)$

– intuition: (for A)



– Caveats:

1) thermalization / burn-in time

the number of steps required for the Markov chain to enter the high-prob. parts of configuration space

2) auto-correlation time:

states: $x_{t-1}, x_t, x_{t+1}, \dots$ are correlated

→ avoid this by "measuring", i.e. recording a state in the sample, every $N_{auto\text{-}corr}$ steps

$\boxed{x_0}, x_1, \dots, x_{N_{autocorr}}, \boxed{x_{N_{autocorr}+1}}, \dots$

• $N_{auto\text{-}corr}$ is observable dependent, i.e. if we use $\{x_0, x_1, \dots, x_T\}$ to estimate an observable O:

$$\langle O \rangle = \int \underline{P}(x)\, O(x)\, dx$$

$$\approx \frac{1}{T} \sum_{i=1}^{T} O(x_i)$$