IWR AI Summer School

- Bayes' law: conditional and joint probabilities

application ① inversion $p(\theta|y)$ from $p(y|\theta)$

→ normalising flows

② prior induction: $p(\theta|y)$ from $p(\theta, y)$

→ simulation-based inference

Bayes' law: incredibly fundamental

- random experiment: selection from a set Ω of events
- probability measure: $p(\Omega) \to [0,...,1]$
- Kolmogorov-axioms
  ① $P(\Omega) = 1$ (equivalent: $P(\{\}) = 0$)
  ② $P(A) \geq 0$, $A \subset \Omega$
  ③ $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \{\}$

- Laplace probability: look at individual events $w \in \Omega$

K3: $P(\bigcup_i A_i) = \sum_i P(A_i) \to P(A) = \sum_{w \in A} P(\{w\})$

$P(w) = \frac{1}{\#\Omega}$ # number of elements "cardinality"
here: all events w have equal probability

$P(B) = \frac{\#B}{\#\Omega}$ Laplace: ratio of positive over total

- conditional probability $p(B|A) = \frac{\#(A \cap B)}{\#A}$

- Bayes' law: $p(B|A) = \frac{\#(A \cap B)}{\#A} = \frac{\#A \cap B}{\#\Omega} \cdot \frac{\#\Omega}{\#A}$

$= \frac{P(A \cap B)}{P(A)}$

$\to P(A \cap B) = P(B|A) \cdot P(A) = P(A|B) \cdot P(B)$

IWR AI summer school

- applying Bayes' law $A \sim$ data y
  $B \sim$ parameters θ

(transitioning from finite sets to real numbers: replace Kolmogorov axioms with Borel's σ-algebra)

$$\mathcal{L}(y|\theta) \cdot \pi(\theta) = p(y,\theta) = \underset{\text{want to know}}{p(\theta|y)} \cdot p(y)$$

likelihood prior joint posterior evidence

$$\rightarrow p(\theta|y) = \frac{\mathcal{L}(y|\theta) \cdot \pi(\theta)}{p(y)}$$ with $p(y) = \int d\theta\, \mathcal{L}(y|\theta)\pi(\theta)$ to ensure $\int d\theta\, p(\theta|y) = 1$

- where comes the prior from? $\rightarrow$ unsure, sometimes obscure like Benford's law
- what's the meaning of evidence? $\rightarrow$ hypothesis testing on the Jeffrey scale

- random distributions and their entropy

$p(\theta) \sim$ to be interpreted as $\frac{1}{\Delta\theta} \cdot \int_{\theta}^{\theta+\Delta\theta} d\theta\, p(\theta)$

- how "random" are distributions
- are two distributions equal?
- is there a most random distribution?

Shannon-axioms: measurement of the entropy

$S = -\int d\theta\, p(\theta) \ln p(\theta)$ or $S = -\sum_i p_i \ln p_i$

① $p_0 = 1,\ p_{i \neq 0} = 0$: $S = 0$ certain event

② $S = \sum_i \sum_j p_i q_j \ln(p_i q_j) = \sum_i p_i \sum_j q_j (\ln p_i + \ln q_j)$

$$= \sum_i p_i \ln p_i \underbrace{\sum_j q_j}_{=1} + \underbrace{\sum_i p_i}_{=1} \sum_j q_j \ln q_j = S_p + S_q$$

IWR AI summer school

③ $p_i = \frac{1}{n}$ equally likely events,

$$S = -\sum_i p_i \ln p_i = \sum_i \left(\frac{1}{n} \ln n\right) = \frac{1}{n} \ln n \cdot \sum_i 1 = \ln n$$

logarithmic scaling of entropy with n ✓

but other measures can do the same thing! Rényi

$$S_\alpha = -\frac{1}{1-\alpha} \ln \sum p_i^\alpha \quad \text{and} \quad S_\alpha = -\frac{1}{1-\alpha} \ln \int d\theta\, p(\theta)^\alpha$$

$\lim_{\alpha \to 1} S_\alpha = S$ Shannon, with de L'Hôpital's rule

① $p_0 = 1,\ p_{i \neq 0} = 0 \quad S_\alpha = \frac{1}{1-\alpha} \ln \underbrace{\sum_i p_i^\alpha}_{=1} = 0$

② $S_\alpha = \frac{1}{1-\alpha} \ln \sum_i \sum_j (p_i \cdot q_j)^\alpha = \frac{1}{1-\alpha} \ln \left(\sum_i p_i^\alpha \cdot \sum_j q_j^\alpha\right)$

$= \frac{1}{1-\alpha} \ln \sum_i p_i^\alpha - \frac{1}{1-\alpha} \ln \sum_j q_j^\alpha = S_\alpha(p) + S_\alpha(q)$

③ $p_i = \frac{1}{n}$: $\quad S_\alpha = \frac{1}{1-\alpha} \ln \sum_i \left(\frac{1}{n}\right)^\alpha = \frac{1}{1-\alpha} \ln\left(\frac{n}{n^\alpha}\right)$

$= \frac{1}{1-\alpha} \ln n^{1-\alpha} = \ln n$

◎ relative entropies

$$\Delta S = \int d\theta\, p(\theta) \cdot \ln \frac{p(\theta)}{q(\theta)} \quad \text{or} \quad \sum_i p_i \ln \frac{p_i}{q_i}$$

if $p(\theta) = q(\theta)$: $\Delta S = 0$ "Kullback-Leibler divergence"

similarly:

$$\Delta S_\alpha = \frac{1}{1-\alpha} \ln \int d\theta\, p(\theta) \left(\frac{p(\theta)}{q(\theta)}\right)^{\alpha-1} \quad \text{Rényi-divergence}$$

$$\text{or} \quad \frac{1}{1-\alpha} \ln \sum_i p_i \left(\frac{p_i}{q_i}\right)^{\alpha-1}$$

IWR AI summerschool

◎ entropies as functionals: extremising distributions

$$S = \underbrace{-\int_a^b d\theta\, p(\theta) \ln p(\theta)}_{\text{entropy, to be maximised}} - \lambda \underbrace{\left[\int_a^b d\theta\, p(\theta) - 1\right]}_{\text{boundary condition fixes normalisation}}$$

Lagrange-multiplier!

$\delta S = 0$ "Hamilton's principle"

$$\delta S = \int_a^b d\theta \left[\delta p \ln p + p \cdot \frac{\delta p}{p} + \lambda \delta p\right]$$

$$= \int_a^b d\theta \left[\ln p + 1 + \lambda\right] \delta p = 0$$

$\rightarrow \ln p + 1 + \lambda = 0, \quad p \sim \exp\left[-(1+\lambda)\right] = \text{const}$

$\rightarrow$ uniform distribution is entropy-maximising

determine λ: $\int_a^b d\theta \exp(-(1+\lambda)) = \exp(-(1+\lambda))(b-a) = 1$

$\rightarrow -(1+\lambda) = \ln \frac{1}{b-a}, \quad 1+\lambda = \ln(b-a), \quad \lambda = \ln(b-a) - 1$

more boundary conditions

$$S = -\int d\theta\, p \ln p - \lambda \underbrace{\left[\int d\theta\, p - 1\right]}_{\text{fixes normalisation}} - \mu \underbrace{\left[\int d\theta\, p \cdot \theta^2 - \langle \theta^2 \rangle\right]}_{\text{fixes variance}}$$

$$\delta S = \int d\theta \left[\ln p + 1 + \lambda + \mu \theta^2\right] \delta p = 0$$

$\longrightarrow p \sim \exp(-[1 + \lambda + \mu\theta^2])$

$\longrightarrow$ Gaussian distribution is entropy-maximising at fixed variance

○ it is really hard to formulate extremal principles with Rényi-type entropies!

IWR AI summer school

relative entropy and invariance

transformation law of probabilities

"$p(\theta)\,d\theta = p(\alpha)\,d\alpha$"

integration by subs.

$$\int d\theta\, p(\theta) = 1 \quad \text{normalisation} = \int d\alpha \underbrace{\frac{\partial\theta}{\partial\alpha} \cdot p(\theta(\alpha))}_{= p(\alpha)}$$

transformation of KL-divergence

$$S = \int d\theta\, p(\theta) \cdot \ln \frac{p(\theta)}{q(\theta)} = \int d\alpha \frac{\partial\theta}{\partial\alpha} p(\theta(\alpha)) \ln \frac{p(\theta(\alpha))\frac{\partial\theta}{\partial\alpha}}{q(\theta(\alpha))\frac{\partial\theta}{\partial\alpha}}$$

$$= \int d\alpha\, p(\alpha) \cdot \ln \frac{p(\alpha)}{q(\alpha)} \quad \text{invariant!}$$

in contrast: Shannon-entropy

$$S = -\int d\theta\, p(\theta) \ln p(\theta) = -\int d\alpha \frac{\partial\theta}{\partial\alpha} p(\theta(\alpha)) \ln \left[p(\theta(\alpha))\frac{\partial\theta}{\partial\alpha}\right]$$

$$= -\int d\alpha\, p(\alpha) \cdot \ln p(\alpha) - \underbrace{\int d\alpha\, p(\alpha) \cdot \ln \frac{\partial\theta}{\partial\alpha}}_{\text{additional term}}$$

mutual information:

are two random variables independent?

$\langle \theta \cdot \alpha \rangle = \int d\theta \int d\alpha\, p(\theta,\alpha) \cdot \theta \cdot \alpha$ "covariance"

has an implicit assumption about Gaussianity

"marginals"

mutual information

$p(\alpha) = \int d\theta\, p(\theta,\alpha)$

$p(\theta) = \int d\alpha\, p(\theta,\alpha)$

$$I(\theta,\alpha) = \int d\theta \int d\alpha\, p(\theta,\alpha) \ln \frac{p(\theta,\alpha)}{p(\theta) \cdot p(\alpha)}$$

if statistically independent: $p(\theta,\alpha) = p(\theta) \cdot p(\alpha)$

$\rightarrow I(\theta,\alpha) = 0$, because $\dfrac{p(\theta,\alpha)}{p(\theta)\, p(\alpha)} = 1$

IWR AI summer school

◎ mutual information

$$I(\theta,\alpha) = \int d\theta \int d\alpha \, p(\theta,\alpha) \ln \frac{p(\theta,\alpha)}{p(\theta)\,p(\alpha)}$$

$$= \int d\theta \int d\alpha \, p(\theta,\alpha) \ln \frac{p(\theta,\alpha)}{p(\theta)} - \int d\theta \int d\alpha \, p(\theta,\alpha) \ln p(\alpha)$$

$$= \int d\theta \int d\alpha \, p(\alpha|\theta) p(\theta) \ln p(\alpha|\theta) - \int d\theta \int d\alpha \, p(\theta,\alpha) \ln p(\alpha)$$

$$= \int d\theta \, p(\theta) \cdot \underbrace{\int d\alpha \, p(\alpha|\theta) \ln p(\alpha|\theta)} - \int d\alpha \underbrace{\int d\theta \, p(\theta,\alpha)} \ln p(\alpha)$$

$$= -\int d\theta \, p(\theta) \cdot S(\alpha|\theta) \qquad - \int d\alpha \, p(\alpha) \cdot \ln p(\alpha)$$

$$= -S(\alpha|\theta) + S(\alpha)$$

$$\rightarrow \quad S(\alpha) - S(\alpha|\theta) = I(\theta,\alpha) = I(\alpha,\theta) = S(\theta) - S(\theta|\alpha)$$

$$\rightarrow \quad S(\theta|\alpha) + S(\alpha) = S(\alpha|\theta) + S(\theta)$$

entropy-variant of Bayes' law
only possible for Shannon-entropy, not Rényi

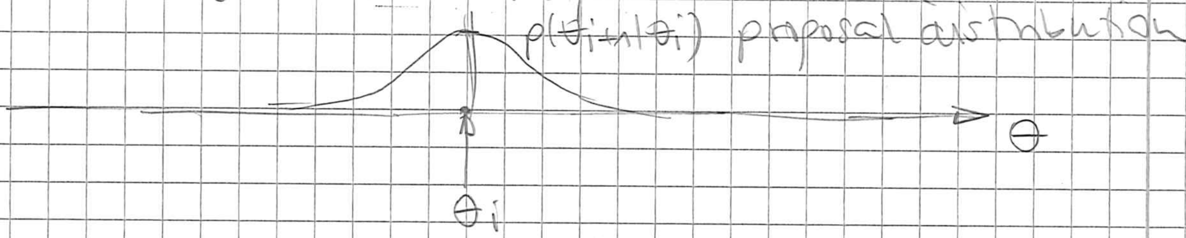
application: ① correlations between (latent) variables
② transformation $p(\alpha)d\alpha = p(\theta)d\theta \rightarrow$
normalising flows
③ target Gaussian $\rightarrow \Delta S =$ KL divergence

IWR AI summer school

- Markov chains: outcome θ_i depends on previous outcomes θ_{i-1} through $p(\theta_i | \theta_{i-1})$

($\text{quantity}: \langle \theta_i \theta_{i-1} \rangle$ or $I(\theta_i, \theta_{i-1})$)

- random walks: set up Markov-chain



draw from proposal distribution generates new sample
for many samples: θ_n is Gaussian distributed (CLT)

Can one "steer" a random walk and get a non-Gaussian distribution → yes: Fokker-Planck

$$dx = \underbrace{\mu(x,t)\,dt}_{\substack{\text{deterministic}\\\text{motion in } dt}} + \underbrace{\sigma(x,t)\,dW}_{\text{random jump in } dt}$$ "Wiener process"

"drift" "diffusion", $D = \frac{\sigma^2}{2}$

for the motion of a single particle →

$$\frac{\partial}{\partial t} p(x,t) = -\frac{\partial}{\partial x}\left[\mu(x,t)\, p(x,t)\right] + \frac{\partial^2}{\partial x^2}\left[D(x,t) \cdot p(x,t)\right]$$

for the 'motion' of a distribution of particles

- steady state solutions: $\frac{\partial}{\partial t} p = 0$

for $\mu = -kx$ and $\frac{\sigma^2}{2} = D = \text{const}$

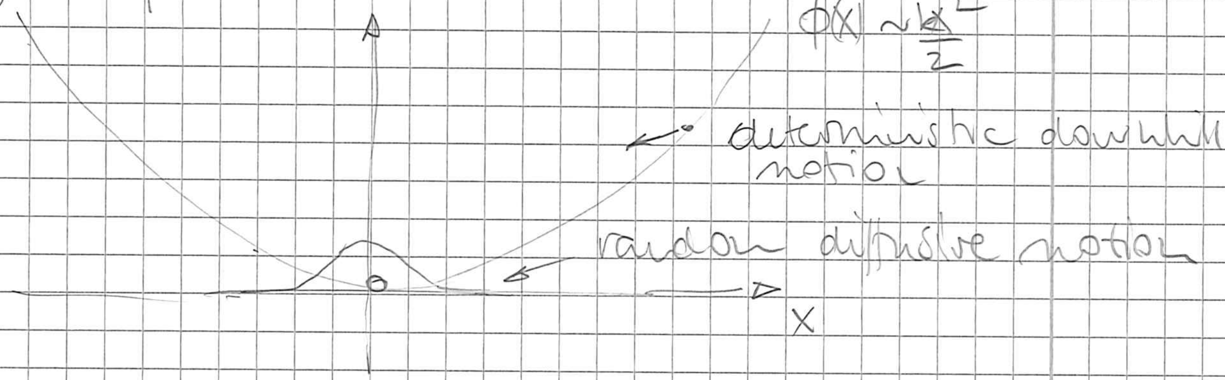
$\Leftrightarrow \phi = \frac{kx^2}{2}$, $\mu = -\frac{\partial \phi}{\partial x}$ parabolic potential

$$\frac{\partial}{\partial x}\left[kx\, p(x)\right] = kp + kx \frac{\partial p}{\partial x} = -\frac{\sigma^2}{2} \frac{\partial^2}{\partial x^2} p = \frac{\partial^2}{\partial x^2}\left[D \cdot p(x)\right]$$

$$\rightarrow p(x) \sim \exp\left(-\frac{kx^2}{2\sigma^2}\right)$$

IWR AI Summer school

physical picture



◎ generalisation to arbitrary potentials

$$\frac{\partial}{\partial t}\rho = -\frac{\partial}{\partial x}(\mu\rho) + \frac{\partial^2}{\partial x^2}\left(\frac{\sigma^2}{2}\rho\right)$$

$$= -\frac{\partial}{\partial x}\left[\mu\rho - \frac{\sigma^2}{2}\frac{\partial}{\partial x}\rho\right]$$ continuity!

$\sim$ probability current

stationary, if $\frac{\partial\rho}{\partial t} = 0 \rightarrow \operatorname{div}\vec{j} = 0$, too

Gauß-law $\int_V dV \operatorname{div}\vec{j} = \int_{\partial V} d\vec{A}\cdot\vec{j} = 0 \rightarrow \vec{j} = 0$ on ∂V

$\mu\rho = \frac{\sigma^2}{2}\frac{\partial\rho}{\partial x}$, with $\mu = -kx$, from $\phi = \frac{kx^2}{2}$

$\frac{\partial}{\partial x}\ln\rho = \frac{-2kx}{\sigma^2} \longrightarrow \ln\rho \sim -\frac{kx^2}{\sigma^2}$, $\rho \sim \exp\left(-\frac{kx^2}{\sigma^2}\right)$

Gauß-distribution is stationary for the Fokker-Planck eqn. in a harmonic potential

◎ dissipation-fluctuation-theorem

$\frac{\sigma^2}{2} = D \cong k_B T$ links diffusivity with temperature

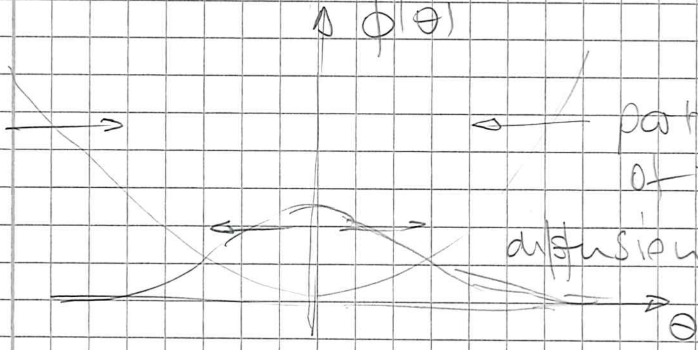
proportionality constant "mobility"

$\longrightarrow \rho \sim \exp\left(-\frac{2kx^2}{k_B T}\right) = \exp\left(-\frac{\phi(x)}{k_B T}\right)$ Boltzmann-distribution

IWR AI Summer School

Markov-chains and sampling

stationary distribution $\frac{\partial \rho}{\partial t}$



particles feel gradient $\partial_i \phi$ of the potential inwards

$\parallel$

diffusion into the outward direction

with Einstein-relation $\rho \sim \exp\left(-\frac{\phi}{T}\right)$ ~ Boltzmann

sampling from $p(\theta)$ → ① construct $\phi \sim -T \ln \phi$

② simulate random walk

Metropolis Hastings-algorithm + many variants

full circle: $p(\theta|y) = \frac{\mathcal{L}(y|\theta)\,\pi(\theta)}{p(y)}$ differentiation

with $p(y) = \int d\theta\, \mathcal{L}(y|\theta)\,\pi(\theta)$

$$Z = \int d\theta\, \exp\Big(-\frac{1}{T}\big(\underbrace{\frac{x^2}{2}+\varphi}_{=\phi(\theta)}\big)\Big) \cdot \exp\left(\frac{1}{T} J\theta\right)$$

→ $\langle \theta^n \rangle = \frac{1}{Z}\frac{\partial}{\partial J} Z \Big|_{J=0,\,T=1}$ moments of $p(\theta|y)$

$F = -T \ln Z$ Helmholtz free energy

$S = -\frac{\partial F}{\partial T} = \ldots = -\int d\theta\, p(\theta|y) \ln p(\theta|y)$ entropy

IWR AI summer school

information geometry

- geometry is the all-powerful language of physics
- ... and of statistics

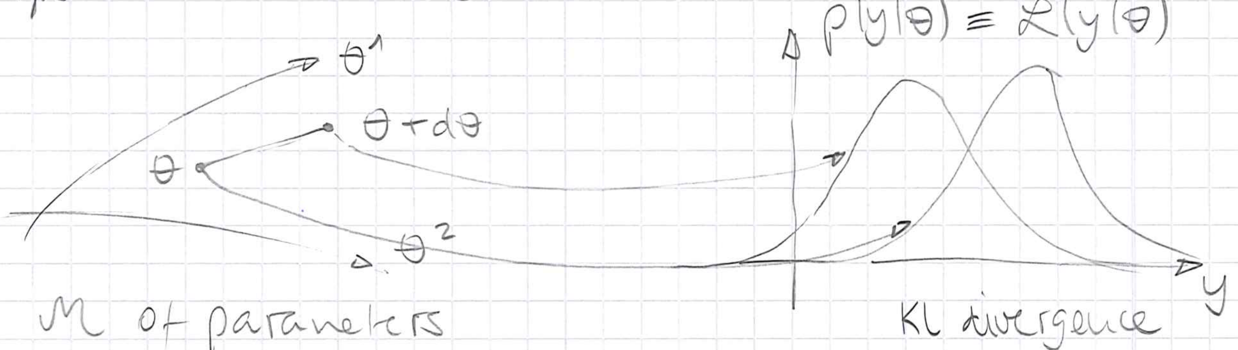
conditional probability $p(y|\theta)$ ← condition in terms of a parameter value θ

infinitesimally different choice $\theta \to \theta + d\theta$: $p(y|\theta + d\theta)$

difference between distributions : Kullback-Leibler

$$\Delta S = \int dy \, p(y|\theta) \cdot \ln \frac{p(y|\theta)}{p(y|\theta + d\theta)} \simeq \frac{1}{2} F_{\mu\nu} d\theta^\mu d\theta^\nu \geq 0$$

$F_{\mu\nu}$ ~ Fisher metric



$\mathcal{M}$ of parameters

↓

has
- metric $F_{\mu\nu}$
- connection ∇_α
- curvature

} differential geometry "Riemannian manifold"