Mathematics for Deep Learning

Roman V. Belavkin

Faculty of Science and Technology
Middlesex University, London NW4 4BT, UK

August 25, 2022
ACDL 2022

Introduction

Bayesian estimation and optimization

Functional analysis, duality, convex analysis

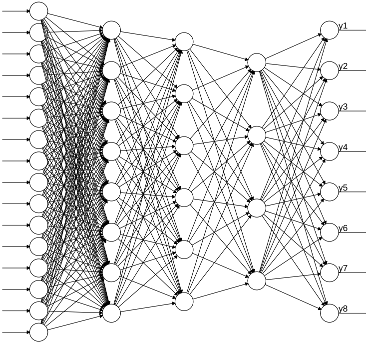
Information theory and information geometry

Probability and analysis

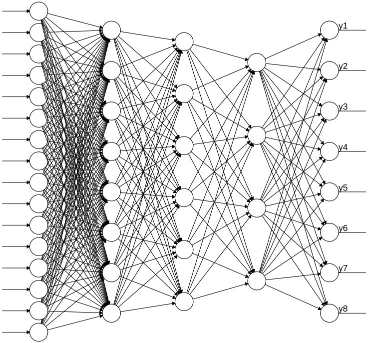
Stochastic filtering equations

Tensor algebras

Feed-forward neural networks



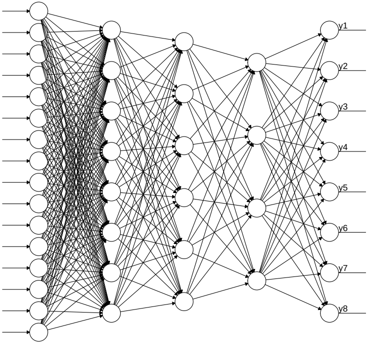
Feed-forward neural networks



Supervised training

1. Initialize the weights w_{ij} (e.g. at random).
2. Repeat

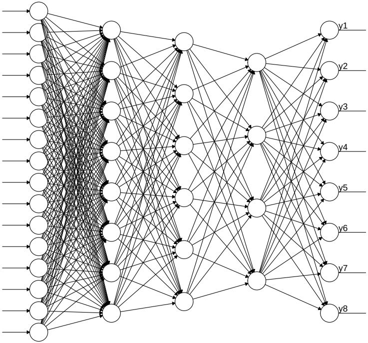
Feed-forward neural networks



Supervised training

1. Initialize the weights w_{ij} (e.g. at random).
2. Repeat
    1. Feed the network with an input z from a training set.

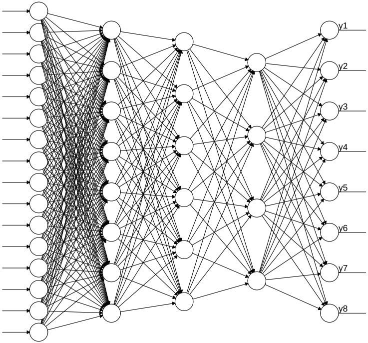
Feed-forward neural networks



Supervised training

1. Initialize the weights w_{ij} (e.g. at random).
2. Repeat
    1. Feed the network with an input z from a training set.
    2. Compute the network's output $y(z)$.

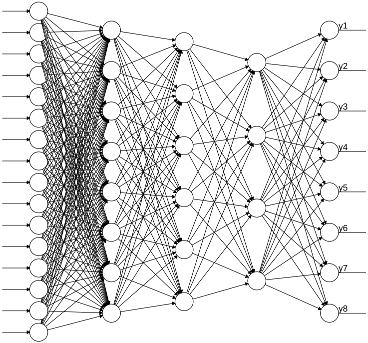
Feed-forward neural networks



Supervised training

1. Initialize the weights w_{ij} (e.g. at random).
2. Repeat
   1. Feed the network with an input z from a training set.
   2. Compute the network's output $y(z)$.
   3. Compute the error $x - y(z)$.

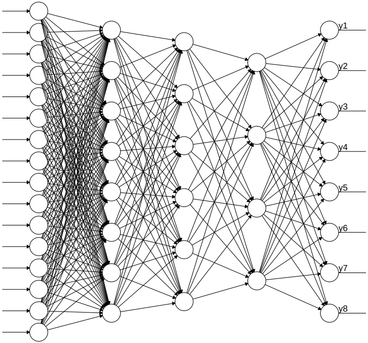
Feed-forward neural networks



Supervised training

1. Initialize the weights w_{ij} (e.g. at random).
2. Repeat
   1. Feed the network with an input z from a training set.
   2. Compute the network's output $y(z)$.
   3. Compute the error $x - y(z)$.
   4. Change the weights w_{ij} to minimise the error's cost $c(x, y(z))$.

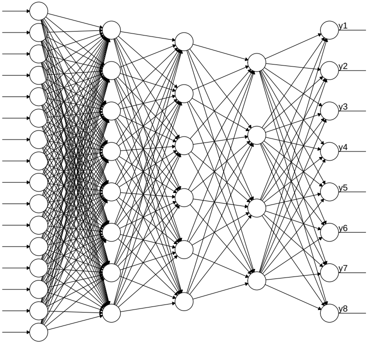
Feed-forward neural networks



Supervised training

1. Initialize the weights w_{ij} (e.g. at random).
2. Repeat
   1. Feed the network with an input z from a training set.
   2. Compute the network's output $y(z)$.
   3. Compute the error $x - y(z)$.
   4. Change the weights w_{ij} to minimise the error's cost $c(x, y(z))$.
3. Until the mean cost is small.

Feed-forward neural networks



Supervised training

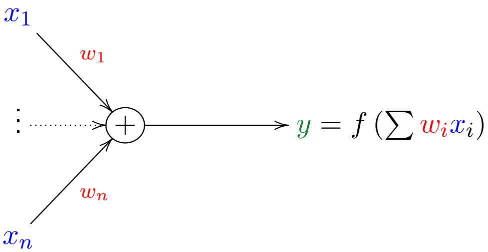
1. Initialize the weights w_{ij} (e.g. at random).
2. Repeat
    1. Feed the network with an input z from a training set.
    2. Compute the network's output $y(z)$.
    3. Compute the error $x - y(z)$.
    4. Change the weights w_{ij} to minimise the error's cost $c(x, y(z))$.
3. Until the mean cost is small.

$$R(I) := \inf_{z(x)} \mathbb{E}_{P(z)} \left\{ \inf_{y(z)} \mathbb{E}_{P(x|z)} \{c(x, y) \mid z\} \right\}$$

Subject to $z(x) : \ln |Z| \leq I$

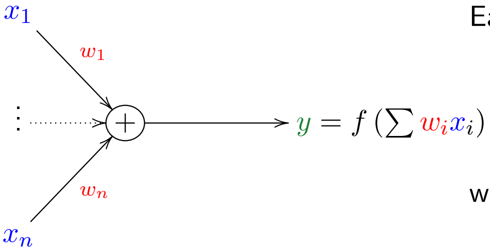
Integrate and Fire Neurons

McCulloch and Pitts (1943)



Integrate and Fire Neurons

McCulloch and Pitts (1943)



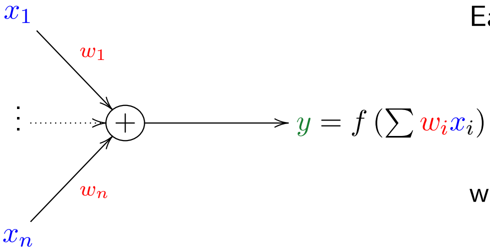
Each node computes a weighted sum:

$$v = \sum_{i=1}^{n} w_i x_i$$

which reminds us a linear model.

Integrate and Fire Neurons

McCulloch and Pitts (1943)



Each node computes a weighted sum:

$$v = \sum_{i=1}^{n} w_i x_i$$

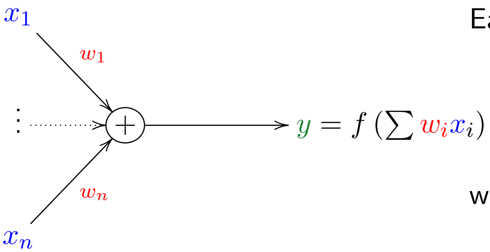
which reminds us a linear model.

Activation function:

$$f(v) = \begin{cases} 1 & \text{if } v \geq a \\ 0 & \text{otherwise} \end{cases}$$

Integrate and Fire Neurons

McCulloch and Pitts (1943)



Each node computes a weighted sum:

$$v = \sum_{i=1}^{n} w_i x_i$$

which reminds us a linear model.

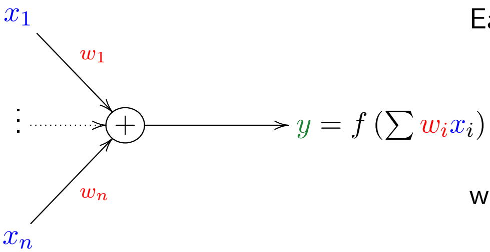
Activation function:

$$f(v) = \begin{cases} 1 & \text{if } v \geq a \\ 0 & \text{otherwise} \end{cases}$$

- Each node partitions the input space into two halves.

Integrate and Fire Neurons

McCulloch and Pitts (1943)



Each node computes a weighted sum:

$$v = \sum_{i=1}^{n} w_i x_i$$

which reminds us a linear model.

Activation function:

$$f(v) = \begin{cases} 1 & \text{if } v \geq a \\ 0 & \text{otherwise} \end{cases}$$

- Each node partitions the input space into two halves.
- With several nodes the perceptron acts as a classifier.

Single layer perceptrons

- The weighted sum for 2 inputs defines a line on (x_1, x_2)-plane:

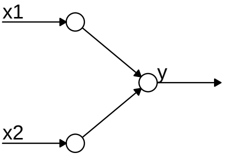
$$a = w_1 x_1 + w_2 x_2$$

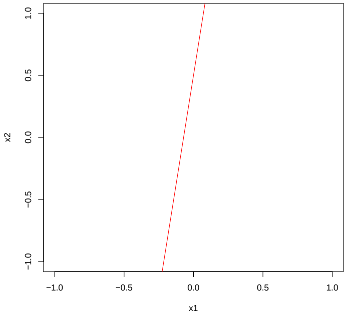
Single layer perceptrons

- The weighted sum for 2 inputs defines a line on (x_1, x_2)-plane:

$$a = w_1 x_1 + w_2 x_2 \quad \Longrightarrow \quad x_2 = \frac{a}{w_2} - \frac{w_1}{w_2} x_1$$

- The line splits the plane into 2 halfspaces.



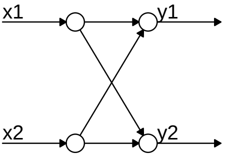


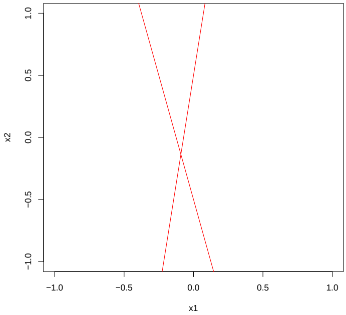
Single layer perceptrons

- The weighted sum for 2 inputs defines a line on (x_1, x_2)-plane:

$$a = w_1 x_1 + w_2 x_2 \qquad \Longrightarrow \qquad x_2 = \frac{a}{w_2} - \frac{w_1}{w_2} x_1$$

- The line splits the plane into 2 halfspaces.



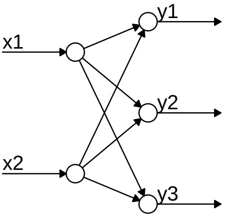


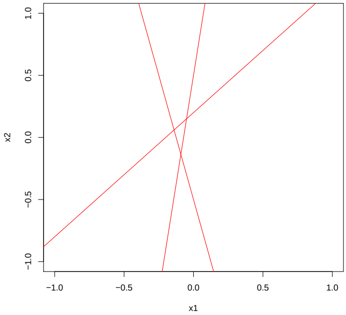
Single layer perceptrons

- The weighted sum for 2 inputs defines a line on (x_1, x_2)-plane:

$$a = w_1 x_1 + w_2 x_2 \qquad \Longrightarrow \qquad x_2 = \frac{a}{w_2} - \frac{w_1}{w_2} x_1$$

- The line splits the plane into 2 halfspaces.



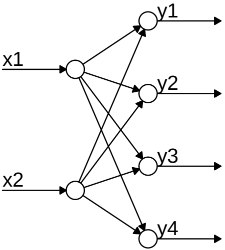


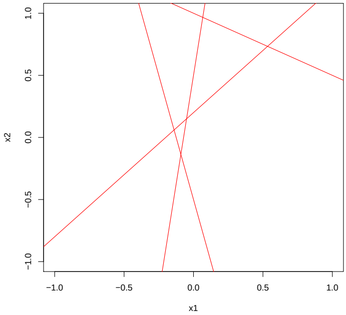
Single layer perceptrons

- The weighted sum for 2 inputs defines a line on (x_1, x_2)-plane:

$$a = w_1 x_1 + w_2 x_2 \qquad \Longrightarrow \qquad x_2 = \frac{a}{w_2} - \frac{w_1}{w_2} x_1$$

- The line splits the plane into 2 halfspaces.



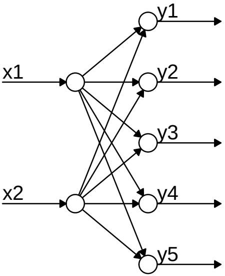


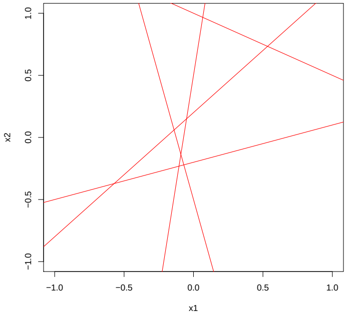
Single layer perceptrons

- The weighted sum for 2 inputs defines a line on (x_1, x_2)-plane:

$$a = w_1 x_1 + w_2 x_2 \qquad \Longrightarrow \qquad x_2 = \frac{a}{w_2} - \frac{w_1}{w_2} x_1$$

- The line splits the plane into 2 halfspaces.



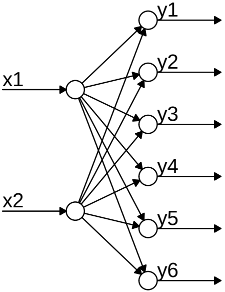


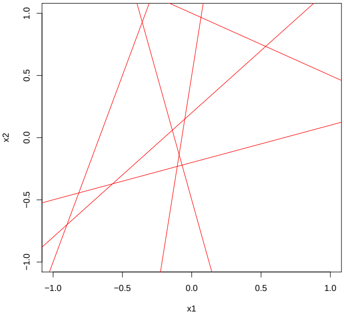
Single layer perceptrons

- The weighted sum for 2 inputs defines a line on (x_1, x_2)-plane:

$$a = w_1 x_1 + w_2 x_2 \quad \Longrightarrow \quad x_2 = \frac{a}{w_2} - \frac{w_1}{w_2} x_1$$

- The line splits the plane into 2 halfspaces.



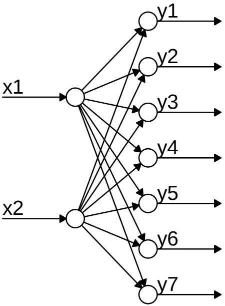


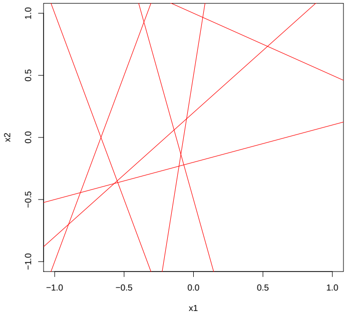
Single layer perceptrons

- The weighted sum for 2 inputs defines a line on (x_1, x_2)-plane:

$$a = w_1 x_1 + w_2 x_2 \quad \Longrightarrow \quad x_2 = \frac{a}{w_2} - \frac{w_1}{w_2} x_1$$

- The line splits the plane into 2 halfspaces.



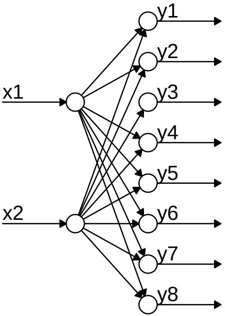


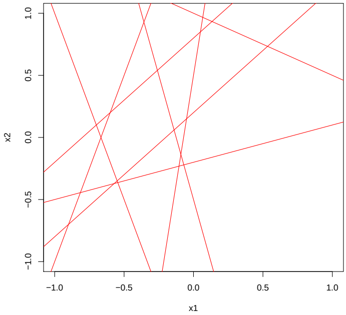
Single layer perceptrons

- The weighted sum for 2 inputs defines a line on (x_1, x_2)-plane:

$$a = w_1 x_1 + w_2 x_2 \quad \Longrightarrow \quad x_2 = \frac{a}{w_2} - \frac{w_1}{w_2} x_1$$

- The line splits the plane into 2 halfspaces.



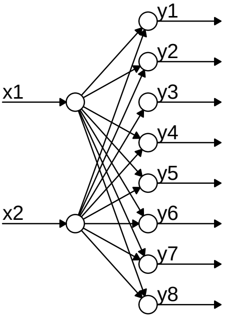


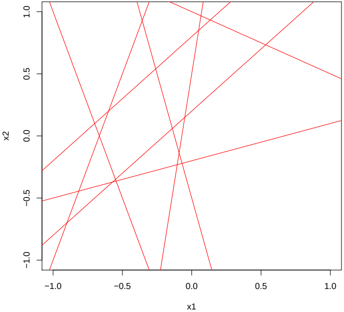
Single layer perceptrons

- The weighted sum for 2 inputs defines a line on (x_1, x_2)-plane:

$$a = w_1 x_1 + w_2 x_2 \quad \Longrightarrow \quad x_2 = \frac{a}{w_2} - \frac{w_1}{w_2} x_1$$

- The line splits the plane into 2 halfspaces.



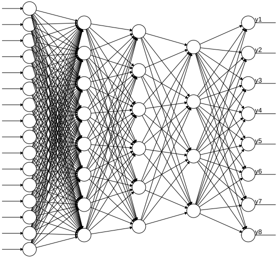


- n nodes partition the space into 2^n subsets.

Learning neural models and the value of information (VoI)

$$R(I) := \inf_{z(x)} \mathbb{E}_{P(z)} \left\{ \inf_{y(z)} \mathbb{E}_{P(x|z)} \{c(x,y) \mid z\} \right\}$$

Subject to $z(x) : \ln |Z| \leq I$

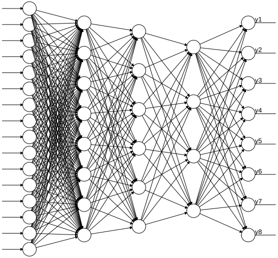


- Some features of deep NNs training including:

Learning neural models and the value of information (VoI)

$$R(I) := \inf_{z(x)} \mathbb{E}_{P(z)} \left\{ \inf_{y(z)} \mathbb{E}_{P(x|z)} \{c(x,y) \mid z\} \right\}$$

Subject to $z(x) : \ln |Z| \leq I$

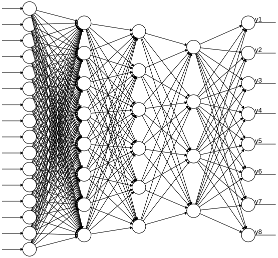


- Some features of deep NNs training including:
  - Partial connectivity.

Learning neural models and the value of information (VoI)

$$R(I) := \inf_{z(x)} \mathbb{E}_{P(z)} \left\{ \inf_{y(z)} \mathbb{E}_{P(x|z)} \{c(x,y) \mid z\} \right\}$$

Subject to $z(x) : \ln |Z| \leq I$

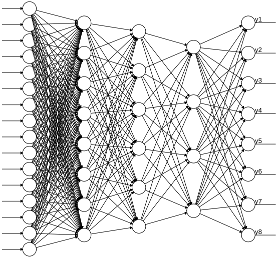


- Some features of deep NNs training including:
  - Partial connectivity.
  - Randomization techniques, such as dropout.

Learning neural models and the value of information (VoI)

$$R(I) := \inf_{z(x)} \mathbb{E}_{P(z)} \left\{ \inf_{y(z)} \mathbb{E}_{P(x|z)} \{c(x,y) \mid z\} \right\}$$

Subject to $z(x) : \ln |Z| \leq I$

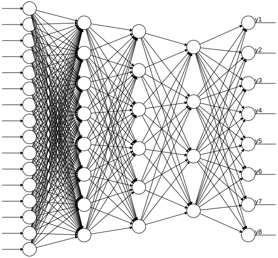


- Some features of deep NNs training including:
  - Partial connectivity.
  - Randomization techniques, such as dropout.
  - Specialized layers (e.g. convolution, $\max$-pooling).

Learning neural models and the value of information (VoI)

$$R(I) := \inf_{z(x)} \mathbb{E}_{P(z)} \left\{ \inf_{y(z)} \mathbb{E}_{P(x|z)} \{c(x, y) \mid z\} \right\}$$

Subject to $z(x) : \ln |Z| \leq I$

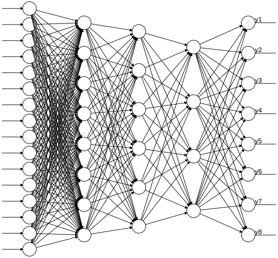


- Some features of deep NNs training including:
  - Partial connectivity.
  - Randomization techniques, such as dropout.
  - Specialized layers (e.g. convolution, $\max$-pooling).
  - Weight regularization.

Learning neural models and the value of information (VoI)

$$R(I) := \inf_{z(x)} \mathbb{E}_{P(z)} \left\{ \inf_{y(z)} \mathbb{E}_{P(x|z)} \{c(x,y) \mid z\} \right\}$$

Subject to $z(x) : \ln |Z| \leq I$

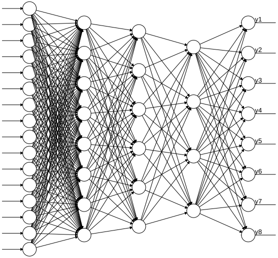


- Some features of deep NNs training including:
    - Partial connectivity.
    - Randomization techniques, such as dropout.
    - Specialized layers (e.g. convolution, $\max$-pooling).
    - Weight regularization.
    - Sublinear activation functions (e.g. ReLU $= \max\{0, v\}$).

Learning neural models and the value of information (VoI)

$$R(I) := \inf_{z(x)} \mathbb{E}_{P(z)} \left\{ \inf_{y(z)} \mathbb{E}_{P(x|z)} \{c(x,y) \mid z\} \right\}$$

Subject to $z(x) : \ln |Z| \leq I$

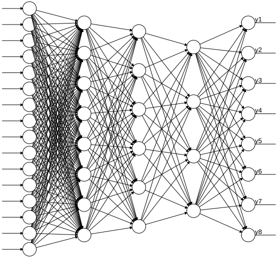


- Some features of deep NNs training including:
  - Partial connectivity.
  - Randomization techniques, such as dropout.
  - Specialized layers (e.g. convolution, $\max$-pooling).
  - Weight regularization.
  - Sublinear activation functions (e.g. ReLU $= \max\{0, v\}$).
  - Crossentropy as cost function.

Learning neural models and the value of information (VoI)

$$R(I) := \inf_{z(x)} \mathbb{E}_{P(z)} \left\{ \inf_{y(z)} \mathbb{E}_{P(x|z)} \{c(x,y) \mid z\} \right\}$$

Subject to $z(x) : \ln |Z| \leq I$



- Some features of deep NNs training including:
  - Partial connectivity.
  - Randomization techniques, such as dropout.
  - Specialized layers (e.g. convolution, $\max$-pooling).
  - Weight regularization.
  - Sublinear activation functions (e.g. ReLU $= \max\{0, v\}$).
  - Crossentropy as cost function.
- Learning weights in neural networks can be seens as a process of maximization of the value of information.

Optimization under uncetainty

- $(\Omega, \mathcal{A}, P)$ — probability space, $x, y, z : \Omega \to \mathbb{R}$ — random variables.

Optimization under uncetainty

- $(\Omega, \mathcal{A}, P)$ — probability space, $x, y, z : \Omega \to \mathbb{R}$ — random variables.
- x — desired response (hidden), y — model response, z — data.

Optimization under uncetainty

- $(\Omega, \mathcal{A}, P)$ — probability space, $x, y, z : \Omega \to \mathbb{R}$ — random variables.
- x — desired response (hidden), y — model response, z — data.
- $u(x, y)$ — utility (or cost $c = -u$).

Optimization under uncetainty

- $(\Omega, \mathcal{A}, P)$ — probability space, $x, y, z : \Omega \to \mathbb{R}$ — random variables.
- x — desired response (hidden), y — model response, z — data.
- $u(x, y)$ — utility (or cost $c = -u$).
- Risk and expected utility

$$R[y(z)] = \int_Z \int_X c(x, y(z)) p(x, z) \, dx \, dz$$

Optimization under uncetainty

- $(\Omega, \mathcal{A}, P)$ — probability space, $x, y, z : \Omega \to \mathbb{R}$ — random variables.
- x — desired response (hidden), y — model response, z — data.
- $u(x, y)$ — utility (or cost $c = -u$).
- Risk and expected utility

$$U[y(z)] = \int_Z \int_X u(x, y(z)) p(x, z) \, dx \, dz$$

Optimization under uncetainty

- $(\Omega, \mathcal{A}, P)$ — probability space, $x, y, z : \Omega \to \mathbb{R}$ — random variables.
- x — desired response (hidden), y — model response, z — data.
- $u(x, y)$ — utility (or cost $c = -u$).
- Risk and expected utility

$$U[y(z)] = \int_Z \int_X u(x, y(z)) p(x, z)\, dx\, dz$$
$$= \int_Z U[y \mid z]\, p(z)\, dz$$

where $U[y \mid z] = \int_X u(x, y)\, p(x \mid z)\, dx$ — conditional (posterior) EU.

Optimization under uncetainty

- $(\Omega, \mathcal{A}, P)$ — probability space, $x, y, z : \Omega \to \mathbb{R}$ — random variables.
- x — desired response (hidden), y — model response, z — data.
- $u(x, y)$ — utility (or cost $c = -u$).
- Risk and expected utility

$$\begin{aligned} U[y(z)] &= \int_Z \int_X u(x, y(z)) p(x, z)\, dx\, dz \\ &= \int_Z U[y \mid z]\, p(z)\, dz \end{aligned}$$

where $U[y \mid z] = \int_X u(x, y)\, p(x \mid z)\, dx$ — conditional (posterior) EU.

Theorem

$$\max_{y(z)} U[y(z)] = \int_Z \max_{y(z)} U[y \mid z]\, p(z)\, dz$$

Optimal estimation / decision

- Optimal $y(z)$ is defined from the condition:

$$\frac{\partial}{\partial y}U[y \mid z] = \int_X \frac{\partial}{\partial y}u(x,y)\,p(x \mid z)\,dx = 0$$

Optimal estimation / decision

- Optimal $y(z)$ is defined from the condition:

$$\frac{\partial}{\partial y}U[y \mid z] = \int_X \frac{\partial}{\partial y}u(x,y)\,p(x \mid z)\,dx = 0$$

Example (Mean-square cost)

For $u(x,y) = -\frac{1}{2}(x-y)^2$ we have $\frac{\partial}{\partial y}u(x,y) = y - x$, so that

$$\int_X (y-x)p(x \mid z)\,dx = 0 \qquad \Longleftrightarrow$$

Optimal estimation / decision

- Optimal $y(z)$ is defined from the condition:

$$\frac{\partial}{\partial y}U[y \mid z] = \int_X \frac{\partial}{\partial y}u(x,y)\,p(x \mid z)\,dx = 0$$

Example (Mean-square cost)

For $u(x,y) = -\frac{1}{2}(x-y)^2$ we have $\frac{\partial}{\partial y}u(x,y) = y - x$, so that

$$\int_X (y-x)p(x \mid z)\,dx = 0 \quad \Longleftrightarrow \quad \hat{y}(z) = \int_X x\,p(x \mid z)\,dx = \mathbb{E}\{x \mid z\}$$

Optimal estimation / decision

- Optimal $y(z)$ is defined from the condition:

$$\frac{\partial}{\partial y}U[y \mid z] = \int_X \frac{\partial}{\partial y}u(x,y)\,p(x \mid z)\,dx = 0$$

Example (Binary cost)

For $u(x,y) = \delta(x-y)$ we have

$$\int_X \frac{\partial}{\partial y}\delta(x-y(z))p(x \mid z)\,dx = \frac{\partial}{\partial x}p(x \mid z) = 0$$

so that $\hat{y}(z) = \arg\max p(x \mid z)$.

Optimal estimation / decision

- Optimal $y(z)$ is defined from the condition:

$$\frac{\partial}{\partial y}U[y \mid z] = \int_X \frac{\partial}{\partial y}u(x, y)\, p(x \mid z)\, dx = 0$$

Example (Cross-entropy cost)

For $u(x, y) = x \ln y + (1 - x) \ln(1 - y)$ for $x, y \in (0, 1)$.

Optimal estimation / decision

- Optimal $y(z)$ is defined from the condition:

$$\frac{\partial}{\partial y}U[y \mid z] = \int_X \frac{\partial}{\partial y}u(x,y)\,p(x \mid z)\,dx = 0$$

Example (Cross-entropy cost)

For $u(x,y) = x\,\ln y + (1-x)\,\ln(1-y)$ for $x, y \in (0,1)$.
Then $\frac{\partial}{\partial y}u(x,y) = \frac{x-y}{y(1-y)}$, and

$$\hat{y}(z) = \int_X x\,p(x \mid z)\,dx = \mathbb{E}\{x \mid z\}$$

Optimal estimation / decision

- Optimal $y(z)$ is defined from the condition:

$$\frac{\partial}{\partial y} U[y \mid z] = \int_X \frac{\partial}{\partial y} u(x, y)\, p(x \mid z)\, dx = 0$$

Example (Cross-entropy cost)

For $u(x, y) = x \ln y + (1 - x) \ln(1 - y)$ for $x, y \in (0, 1)$.
Then $\frac{\partial}{\partial y} u(x, y) = \frac{x-y}{y(1-y)}$, and

$$\hat{y}(z) = \int_X x\, p(x \mid z)\, dx = \mathbb{E}\{x \mid z\}$$

Equivalent to minimization of $KL[x, y] = \sum x[\ln x - \ln y]$

Duality: Observables $\in X \leftarrow \langle\cdot,\cdot\rangle \rightarrow Y \ni$ States

$\langle\cdot,\cdot\rangle : X \times Y \to \mathbb{C}$

$$\langle x, y\rangle := \sum x_i y_i \,, \qquad \langle x, y\rangle := \int x\, dy \,, \qquad \langle x, y\rangle := \operatorname{tr}\{xy\}$$

Duality: Observables $\in X \leftarrow \langle\cdot,\cdot\rangle \rightarrow Y \ni$ States

$\langle\cdot,\cdot\rangle : X \times Y \to \mathbb{C}$

$$\langle x, y\rangle := \sum x_i y_i\,, \qquad \langle x, y\rangle := \int x\,dy\,, \qquad \langle x, y\rangle := \operatorname{tr}\{xy\}$$

- X is a $*$-algebra with $1 \in X$

Duality: Observables $\in X \leftarrow \langle\cdot,\cdot\rangle \rightarrow Y \ni$ States

$\langle\cdot,\cdot\rangle : X \times Y \to \mathbb{C}$

$$\langle x, y\rangle := \sum x_i y_i\,, \qquad \langle x, y\rangle := \int x\,dy\,, \qquad \langle x, y\rangle := \mathrm{tr}\,\{xy\}$$

- X is a $*$-algebra with $1 \in X$
- Y dual of X

Duality: Observables $\in X \leftarrow \langle\cdot,\cdot\rangle \rightarrow Y \ni$ States

$\langle\cdot,\cdot\rangle : X \times Y \to \mathbb{C}$

$$\langle x, y\rangle := \sum x_i y_i \,, \qquad \langle x, y\rangle := \int x\, dy \,, \qquad \langle x, y\rangle := \operatorname{tr}\{xy\}$$

- X is a $*$-algebra with $1 \in X$
- Involution $(x^* z)^* = z^* x$
- Y dual of X

Duality: Observables $\in X \leftarrow \langle\cdot,\cdot\rangle \rightarrow Y \ni$ States

$\langle\cdot,\cdot\rangle : X \times Y \to \mathbb{C}$

$$\langle x, y\rangle := \sum x_i y_i\,, \qquad \langle x, y\rangle := \int x\,dy\,, \qquad \langle x, y\rangle := \operatorname{tr}\{xy\}$$

- X is a $*$-algebra with $1 \in X$
- Involution $(x^* z)^* = z^* x$
- Y dual of X
- Involution $\langle x, y^*\rangle = \langle x^*, y\rangle^*$

Duality: Observables $\in X \leftarrow \langle\cdot,\cdot\rangle \rightarrow Y \ni$ States

$\langle\cdot,\cdot\rangle : X \times Y \to \mathbb{C}$

$$\langle x, y\rangle := \sum x_i y_i\,, \qquad \langle x, y\rangle := \int x\,dy\,, \qquad \langle x, y\rangle := \operatorname{tr}\{xy\}$$

- X is a $*$-algebra with $1 \in X$
- Involution $(x^* z)^* = z^* x$
- $X_+ := \{x : z^* z = x,\ \exists\, z \in X\}$

- Y dual of X
- Involution $\langle x, y^*\rangle = \langle x^*, y\rangle^*$

Duality: Observables $\in X \leftarrow \langle\cdot,\cdot\rangle \rightarrow Y \ni$ States

$\langle\cdot,\cdot\rangle : X \times Y \to \mathbb{C}$

$$\langle x, y\rangle := \sum x_i y_i\,, \qquad \langle x, y\rangle := \int x\,dy\,, \qquad \langle x, y\rangle := \operatorname{tr}\{xy\}$$

- X is a $*$-algebra with $1 \in X$
- Involution $(x^* z)^* = z^* x$
- $X_+ := \{x : z^* z = x,\ \exists\, z \in X\}$

- Y dual of X
- Involution $\langle x, y^*\rangle = \langle x^*, y\rangle^*$
- $Y_+ := \{y : \langle x, y\rangle \geq 0,\ \forall\, x \in X_+\}$

Duality: Observables $\in X \leftarrow \langle\cdot,\cdot\rangle \rightarrow Y \ni$ States

$\langle\cdot,\cdot\rangle : X \times Y \to \mathbb{C}$

$$\langle x, y\rangle := \sum x_i y_i\,, \qquad \langle x, y\rangle := \int x\,dy\,, \qquad \langle x, y\rangle := \operatorname{tr}\{xy\}$$

- X is a $*$-algebra with $1 \in X$
- Involution $(x^* z)^* = z^* x$
- $X_+ := \{x : z^* z = x,\ \exists\, z \in X\}$
- Observables $x = x^*$

- Y dual of X
- Involution $\langle x, y^*\rangle = \langle x^*, y\rangle^*$
- $Y_+ := \{y : \langle x, y\rangle \geq 0,\ \forall\, x \in X_+\}$

Duality: Observables $\in X \leftarrow \langle\cdot,\cdot\rangle \rightarrow Y \ni$ States

$\langle\cdot,\cdot\rangle : X \times Y \to \mathbb{C}$

$$\langle x, y\rangle := \sum x_i y_i\,, \qquad \langle x, y\rangle := \int x\,dy\,, \qquad \langle x, y\rangle := \mathrm{tr}\,\{xy\}$$

- X is a $*$-algebra with $1 \in X$
- Involution $(x^* z)^* = z^* x$
- $X_+ := \{x : z^* z = x,\ \exists\, z \in X\}$
- Observables $x = x^*$

- Y dual of X
- Involution $\langle x, y^*\rangle = \langle x^*, y\rangle^*$
- $Y_+ := \{y : \langle x, y\rangle \geq 0,\ \forall\, x \in X_+\}$
- States $y \geq 0$, $\langle 1, y\rangle = 1$

Duality: Observables $\in X \leftarrow \langle\cdot,\cdot\rangle \rightarrow Y \ni$ States

$\langle\cdot,\cdot\rangle : X \times Y \to \mathbb{C}$

$$\langle x, y\rangle := \sum x_i y_i\,, \qquad \langle x, y\rangle := \int x\,dy\,, \qquad \langle x, y\rangle := \operatorname{tr}\{xy\}$$

- X is a $*$-algebra with $1 \in X$
- Involution $(x^* z)^* = z^* x$
- $X_+ := \{x : z^* z = x,\ \exists\, z \in X\}$
- Observables $x = x^*$

- Y dual of X
- Involution $\langle x, y^*\rangle = \langle x^*, y\rangle^*$
- $Y_+ := \{y : \langle x, y\rangle \geq 0,\ \forall\, x \in X_+\}$
- States $y \geq 0$, $\langle 1, y\rangle = 1$

- The base of Y_+ is the set of all states (statistical manifold):

$$\mathcal{P}(X) := \{p \in Y_+ : \langle 1, p\rangle = 1\}$$

Duality: Observables $\in X \leftarrow \langle\cdot,\cdot\rangle \rightarrow Y \ni$ States

$\langle\cdot,\cdot\rangle : X \times Y \to \mathbb{C}$

$$\langle x, y\rangle := \sum x_i y_i\,, \qquad \langle x, y\rangle := \int x\,dy\,, \qquad \langle x, y\rangle := \operatorname{tr}\{xy\}$$

- X is a $*$-algebra with $1 \in X$
- Involution $(x^* z)^* = z^* x$
- $X_+ := \{x : z^* z = x,\ \exists\, z \in X\}$
- Observables $x = x^*$

- Y dual of X
- Involution $\langle x, y^*\rangle = \langle x^*, y\rangle^*$
- $Y_+ := \{y : \langle x, y\rangle \geq 0,\ \forall\, x \in X_+\}$
- States $y \geq 0$, $\langle 1, y\rangle = 1$

- The base of Y_+ is the set of all states (statistical manifold):

$$\mathcal{P}(X) := \{p \in Y_+ : \langle 1, p\rangle = 1\}$$

- Transposition: $\forall\, z \in X\ \exists\, z' \in Y$: $\langle zx, y\rangle = \langle x, z'y\rangle$

Duality: Observables $\in X \leftarrow \langle\cdot,\cdot\rangle \rightarrow Y \ni$ States

$\langle\cdot,\cdot\rangle : X \times Y \to \mathbb{C}$

$$\langle x, y\rangle := \sum x_i y_i\,, \qquad \langle x, y\rangle := \int x\,dy\,, \qquad \langle x, y\rangle := \operatorname{tr}\{xy\}$$

- X is a $*$-algebra with $1 \in X$
- Involution $(x^* z)^* = z^* x$
- $X_+ := \{x : z^* z = x,\ \exists\, z \in X\}$
- Observables $x = x^*$

- Y dual of X
- Involution $\langle x, y^*\rangle = \langle x^*, y\rangle^*$
- $Y_+ := \{y : \langle x, y\rangle \geq 0,\ \forall\, x \in X_+\}$
- States $y \geq 0$, $\langle 1, y\rangle = 1$

- The base of Y_+ is the set of all states (statistical manifold):

$$\mathcal{P}(X) := \{p \in Y_+ : \langle 1, p\rangle = 1\}$$

- Transposition: $\forall\, z \in X\ \exists\, z' \in Y$: $\langle zx, y\rangle = \langle x, z'y\rangle$
- Y is a left (resp. right) module over $X \subseteq Y$ w.r.t. $z'y$ (resp. $yz^{*\prime *}$).

Exponents and Logarithms

- Define by the power series

$$e^x := \sum_{n=0}^{\infty} \frac{x^n}{n!}\,, \qquad \ln y := \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}\,(y-1)^n$$

Exponents and Logarithms

- Define by the power series

$$e^x := \sum_{n=0}^{\infty} \frac{x^n}{n!}\,, \qquad \ln y := \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}\,(y-1)^n$$

- Group homomorphisms for $xz = zx$ and $yz = zy$:

$$e^{x+z} = e^x e^z \qquad \text{and} \qquad \ln(yz) = \ln y + \ln z$$

Exponents and Logarithms

- Define by the power series

$$e^x := \sum_{n=0}^{\infty} \frac{x^n}{n!}\,, \qquad \ln y := \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}\,(y-1)^n$$

- Group homomorphisms for $xz = zx$ and $yz = zy$:

$$e^{x+z} = e^x e^z \qquad \text{and} \qquad \ln(yz) = \ln y + \ln z$$

- Group homomorphisms for tesnor product $\otimes$ and Kronecker $\oplus$:

$$e^{x \oplus z} = e^x \otimes e^z \qquad \text{and} \qquad \ln(y \otimes z) = \ln y \oplus \ln z$$

Exponents and Logarithms

- Define by the power series

$$e^x := \sum_{n=0}^{\infty} \frac{x^n}{n!}\,, \qquad \ln y := \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}\,(y-1)^n$$

- Group homomorphisms for $xz = zx$ and $yz = zy$:

$$e^{x+z} = e^x e^z \qquad \text{and} \qquad \ln(yz) = \ln y + \ln z$$

- Group homomorphisms for tesnor product $\otimes$ and Kronecker $\oplus$:

$$e^{x \oplus z} = e^x \otimes e^z \qquad \text{and} \qquad \ln(y \otimes z) = \ln y \oplus \ln z$$

- Because $X \subseteq Y$, we can consider

$$\exp : X \to Y \qquad \text{and} \qquad \ln : Y \to X$$

Random variables and probability measures

Duality

- $\mathcal{C}_c(\Omega)$ linear space of continuous functions with compact support.

Random variables and probability measures

Duality

- $\mathcal{C}_c(\Omega)$ linear space of continuous functions with compact support.
- $\mathcal{C}'_c(\Omega) =: \mathcal{M}(\Omega)$ linear space of signed Radon measures on Ω.

Random variables and probability measures

Duality

- $\mathcal{C}_c(\Omega)$ linear space of continuous functions with compact support.
- $\mathcal{C}'_c(\Omega) =: \mathcal{M}(\Omega)$ linear space of signed Radon measures on Ω.
- The pairing $\langle \cdot, \cdot \rangle$ of $\mathcal{C}_c(\Omega)$ and $\mathcal{M}(\Omega)$ is w.r.t.:

$$\langle f, \mu \rangle = \sum_{\Omega} f(\omega)\, \mu(\omega)\,, \quad \langle f, \mu \rangle = \int_{\Omega} f(\omega)\, d\mu(\omega)\,, \quad \langle f, \mu \rangle = \mathrm{tr}\,\{f\,\mu\}$$

Random variables and probability measures

Duality

- $\mathcal{C}_c(\Omega)$ linear space of continuous functions with compact support.
- $\mathcal{C}'_c(\Omega) =: \mathcal{M}(\Omega)$ linear space of signed Radon measures on Ω.
- The pairing $\langle \cdot, \cdot \rangle$ of $\mathcal{C}_c(\Omega)$ and $\mathcal{M}(\Omega)$ is w.r.t.:

$$\langle f, \mu \rangle = \sum_{\Omega} f(\omega)\, \mu(\omega)\,, \quad \langle f, \mu \rangle = \int_{\Omega} f(\omega)\, d\mu(\omega)\,, \quad \langle f, \mu \rangle = \operatorname{tr}\{f\, \mu\}$$

Random variables and probability measures

Duality

- $\mathcal{C}_c(\Omega)$ linear space of continuous functions with compact support.
- $\mathcal{C}_c'(\Omega) =: \mathcal{M}(\Omega)$ linear space of signed Radon measures on Ω.
- The pairing $\langle \cdot, \cdot \rangle$ of $\mathcal{C}_c(\Omega)$ and $\mathcal{M}(\Omega)$ is w.r.t.:

$$\langle f, \mu \rangle = \sum_{\Omega} f(\omega)\,\mu(\omega)\,, \quad \langle f, \mu \rangle = \int_{\Omega} f(\omega)\,d\mu(\omega)\,, \quad \langle f, \mu \rangle = \mathrm{tr}\,\{f\,\mu\}$$

Algebraic structure

- $\mathcal{C}_c(\Omega)$ is a $*$-algebra, and $\mathcal{C}_c''(\Omega)$ is a unital von-Neumann algebra:

$$f + g\,, \qquad f \cdot g\,, \qquad f \mapsto f^*\,, \qquad \|f^* f\|_\infty = \|f\|_\infty^2$$

- Events $E \subseteq \Omega$ are indicator functions $\chi_E(\omega)$.

Random variables and probability measures

Duality

- $\mathcal{C}_c(\Omega)$ linear space of continuous functions with compact support.
- $\mathcal{C}'_c(\Omega) =: \mathcal{M}(\Omega)$ linear space of signed Radon measures on Ω.
- The pairing $\langle \cdot, \cdot \rangle$ of $\mathcal{C}_c(\Omega)$ and $\mathcal{M}(\Omega)$ is w.r.t.:

$$\langle f, \mu \rangle = \sum_{\Omega} f(\omega)\,\mu(\omega)\,, \quad \langle f, \mu \rangle = \int_{\Omega} f(\omega)\,d\mu(\omega)\,, \quad \langle f, \mu \rangle = \mathrm{tr}\,\{f\,\mu\}$$

Algebraic structure

- $\mathcal{C}_c(\Omega)$ is a $*$-algebra, and $\mathcal{C}''_c(\Omega)$ is a unital von-Neumann algebra:

$$f + g\,, \qquad f \cdot g\,, \qquad f \mapsto f^*\,, \qquad \|f^* f\|_\infty = \|f\|_\infty^2$$

- Events $E \subseteq \Omega$ are indicator functions $\chi_E(\omega)$.
- Random variables (observables) are $f \in \mathcal{C}''_c(\Omega)$.

Probability measures

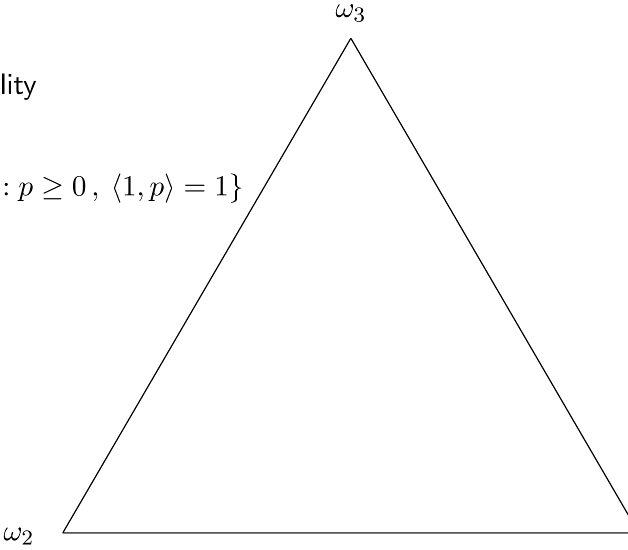
- The set of all probability measures

  $$\mathcal{P}(\Omega) := \{p \in \mathcal{M}(\Omega) : p \geq 0\,,\ \langle 1, p\rangle = 1\}$$

Probability measures

- The set of all probability measures

$$\mathcal{P}(\Omega) := \{p \in \mathcal{M}(\Omega) : p \geq 0\,,\ \langle 1, p \rangle = 1\}$$



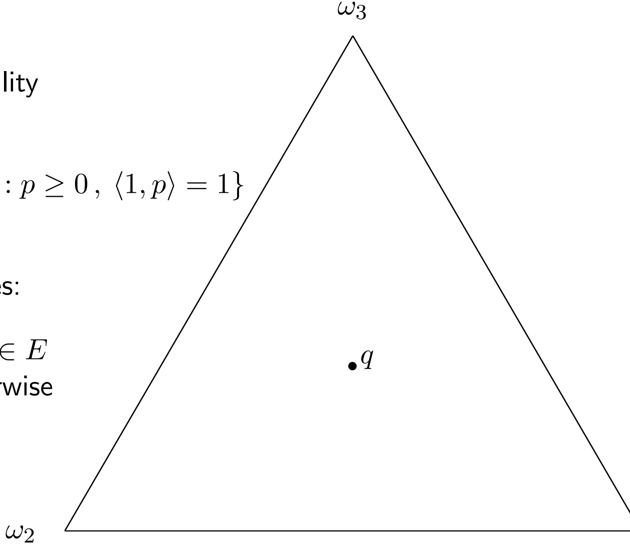
Probability measures

- The set of all probability measures

  $$\mathcal{P}(\Omega) := \{p \in \mathcal{M}(\Omega) : p \geq 0\,,\ \langle 1, p\rangle = 1\}$$

- $\delta \in \operatorname{ext} \mathcal{P}$ are Dirac (elementary) measures:

  $$\delta_\omega(E) = \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{otherwise} \end{cases}$$



Choice under uncertainty

- Choice of $\omega \in (\Omega, \lesssim)$ is equivalent to choice of $\delta \in \operatorname{ext} \mathcal{P} \equiv \Omega$

Choice under uncertainty

- Choice of $\omega \in (\Omega, \lesssim)$ is equivalent to choice of $\delta \in \operatorname{ext} \mathcal{P} \equiv \Omega$
- Utility $u : \Omega \to \mathbb{R}$ is

$$u(a) = \int_E u(\omega)\, \delta_a(\omega)$$

Choice under uncertainty

- Choice of $\omega \in (\Omega, \lesssim)$ is equivalent to choice of $\delta \in \mathrm{ext}\ \mathcal{P} \equiv \Omega$
- Utility $u : \Omega \to \mathbb{R}$ is

$$u(a) = \int_E u(\omega)\, \delta_a(\omega)$$

- Choice of different lotteries on Ω is a choice of $p \in \mathcal{P}$

Choice under uncertainty

- Choice of $\omega \in (\Omega, \lesssim)$ is equivalent to choice of $\delta \in \operatorname{ext} \mathcal{P} \equiv \Omega$
- Utility $u : \Omega \to \mathbb{R}$ is

$$u(a) = \int_E u(\omega)\, \delta_a(\omega)$$

- Choice of different lotteries on Ω is a choice of $p \in \mathcal{P}$
- The expected utility is

$$\mathbb{E}_p\{u\} := \int_\Omega u(\omega)\, dp(\omega)$$

Choice under uncertainty

- Choice of $\omega \in (\Omega, \lesssim)$ is equivalent to choice of $\delta \in \operatorname{ext} \mathcal{P} \equiv \Omega$
- Utility $u : \Omega \to \mathbb{R}$ is

$$u(a) = \int_E u(\omega)\, \delta_a(\omega)$$

- Choice of different lotteries on Ω is a choice of $p \in \mathcal{P}$
- The expected utility is

$$\mathbb{E}_p\{u\} := \int_\Omega u(\omega)\, dp(\omega)$$

Expected Utility (von Neumann & Morgenstern, 1944)

$(\mathcal{P}, \lesssim)$ satisfies axioms (continuity, independence and Archimedian) if and only if there exists the expected utility representation of $(\mathcal{P}, \lesssim)$:

$$q \lesssim p \qquad \iff \qquad \mathbb{E}_q\{u\} \leq \mathbb{E}_p\{u\}$$

Entropy and information

- Surprise: $-\ln P(x)$

Entropy and information

- Surprise: $-\ln P(x)$
- Entropy is expected surprise

$$H(X) := \mathbb{E}_P\{-\ln P(x)\}$$

Entropy and information

- Surprise: $-\ln P(x)$
- Entropy is expected surprise

$$H(X) := \mathbb{E}_P\{-\ln P(x)\}$$

- Shannon (1948)'s mutual information between x and y:

$$I(X,Y) := H(X) - H(X \mid Y)$$

Entropy and information

- Surprise: $-\ln P(x)$
- Entropy is expected surprise

$$H(X) := \mathbb{E}_P\{-\ln P(x)\}$$

- Shannon (1948)'s mutual information between x and y:

$$\begin{aligned} I(X,Y) &:= H(X) - H(X \mid Y) \\ &= \sum_{X\times Y} \left[\ln \frac{w(x,y)}{q(x)\,p(y)}\right] P(x,y) \end{aligned}$$

Entropy and information

- Surprise: $-\ln P(x)$
- Entropy is expected surprise

$$H(X) := \mathbb{E}_P\{-\ln P(x)\}$$

- Shannon (1948)'s mutual information between x and y:

$$\begin{aligned} I(X,Y) &:= H(X) - H(X \mid Y) \\ &= \sum_{X \times Y} \left[\ln \frac{w(x,y)}{q(x)\,p(y)}\right] P(x,y) \end{aligned}$$

- Entropy as self-information:

$$I(X,X) = H(X)$$

Entropy and information

- Surprise: $-\ln P(x)$
- Entropy is expected surprise

$$H(X) := \mathbb{E}_P\{-\ln P(x)\}$$

- Shannon (1948)'s mutual information between x and y:

$$\begin{aligned} I(X,Y) &:= H(X) - H(X \mid Y) \\ &= \sum_{X\times Y}\left[\ln\frac{w(x,y)}{q(x)\,p(y)}\right] P(x,y) \end{aligned}$$

- Entropy as self-information:

$$I(X,X) = H(X)$$

- Information upper bound:

$$0 \le I(X,Y) \le \min[H(X),H(Y)]$$

Entropy and information

- Surprise: $-\ln P(x)$
- Entropy is expected surprise

$$H(X) := \mathbb{E}_P\{-\ln P(x)\}$$

- Shannon (1948)'s mutual information between x and y:

$$\begin{aligned} I(X,Y) &:= H(X) - H(X \mid Y) \\ &= \sum_{X\times Y} \left[\ln \frac{w(x,y)}{q(x)\,p(y)}\right] P(x,y) \end{aligned}$$

- Entropy as self-information:

$$I(X,X) = H(X)$$

- Information upper bound:

$$0 \leq I(X,Y) \leq \min[H(X),H(Y)]$$

- Kullback-Leibler divergence: $KL[p,q] := \mathbb{E}_p\{\ln(p/q)\}$

Entropy and information

- Surprise: $-\ln P(x)$
- Entropy is expected surprise

$$H(X) := \mathbb{E}_P\{-\ln P(x)\} = r(X) - KL[p, r/r(X)]$$

- Shannon (1948)'s mutual information between x and y:

$$\begin{aligned} I(X,Y) &:= H(X) - H(X \mid Y) \\ &= \sum_{X\times Y} \left[\ln \frac{w(x,y)}{q(x)\,p(y)}\right] P(x,y) \end{aligned}$$

- Entropy as self-information:

$$I(X,X) = H(X)$$

- Information upper bound:

$$0 \leq I(X,Y) \leq \min[H(X), H(Y)]$$

- Kullback-Leibler divergence: $KL[p,q] := \mathbb{E}_p\{\ln(p/q)\}$

Entropy and information

- Surprise: $-\ln P(x)$
- Entropy is expected surprise

$$H(X) := \mathbb{E}_P\{-\ln P(x)\} = r(X) - KL[p, r/r(X)]$$

- Shannon (1948)'s mutual information between x and y:

$$\begin{aligned} I(X,Y) &:= H(X) - H(X \mid Y) \\ &= \sum_{X \times Y} \left[\ln \frac{w(x,y)}{q(x)\,p(y)}\right] P(x,y) = KL[w, q \otimes p] \end{aligned}$$

- Entropy as self-information:

$$I(X,X) = H(X)$$

- Information upper bound:

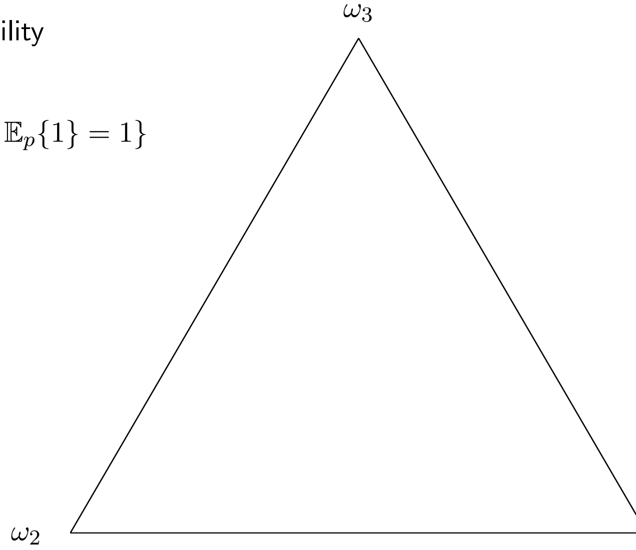
$$0 \leq I(X,Y) \leq \min[H(X), H(Y)]$$

- Kullback-Leibler divergence: $KL[p,q] := \mathbb{E}_p\{\ln(p/q)\}$

Information-geometruc view

- The set of all probability measures

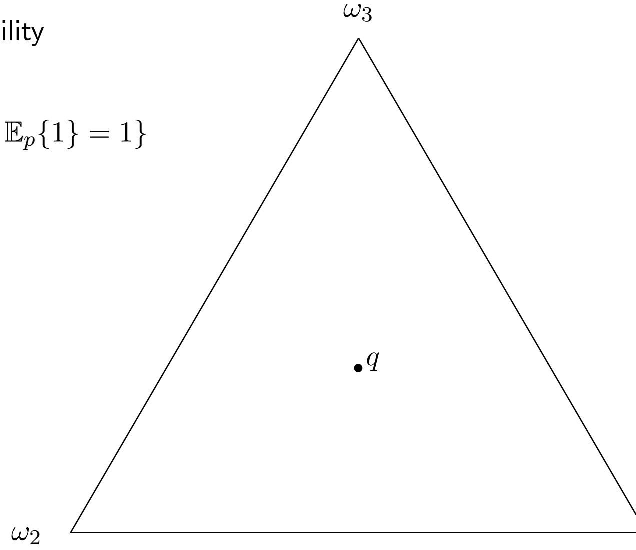
$$\mathcal{P}(\Omega) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$



Information-geometruc view

- The set of all probability measures

$$\mathcal{P}(\Omega) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

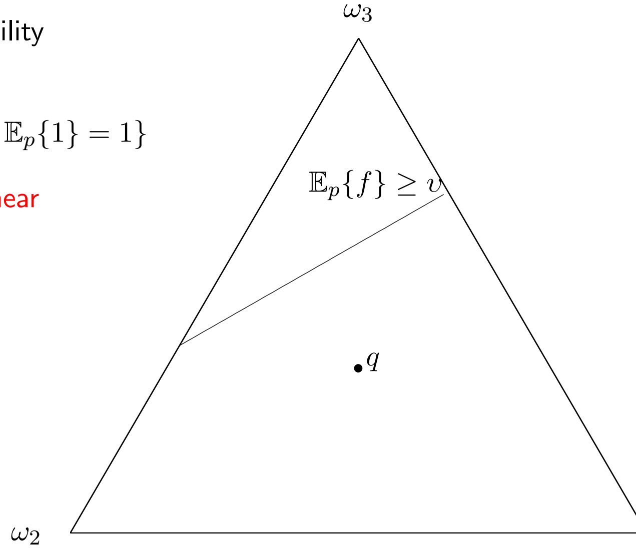


Information-geometruc view

- The set of all probability measures

$$\mathcal{P}(\Omega) := \{p : p \geq 0\,,\, \mathbb{E}_p\{1\} = 1\}$$

- $\mathbb{E}_p\{f\} := \langle f, p \rangle$ is linear

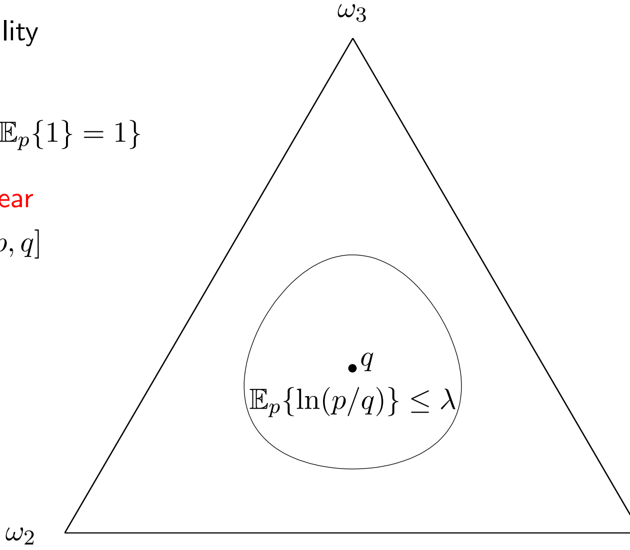


Information-geometruc view

- The set of all probability measures

$$\mathcal{P}(\Omega) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- $\mathbb{E}_p\{f\} := \langle f, p \rangle$ is linear
- $\mathbb{E}_p\{\ln(p/q)\} =: KL[p, q]$ is convex



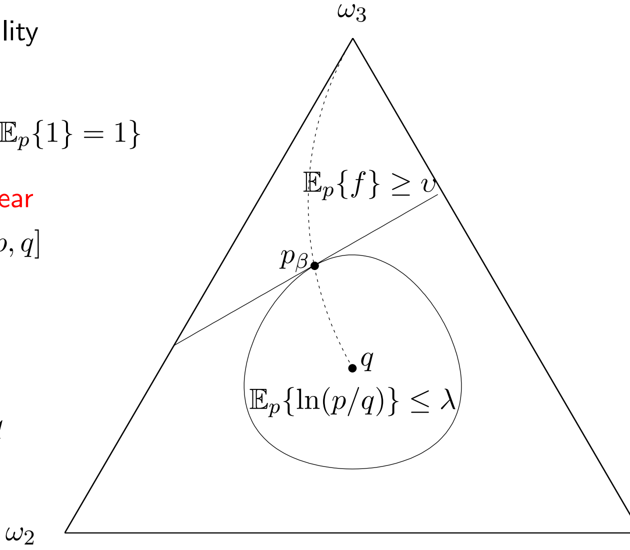
Information-geometruc view

- The set of all probability measures

$$\mathcal{P}(\Omega) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- $\mathbb{E}_p\{f\} := \langle f, p \rangle$ is linear
- $\mathbb{E}_p\{\ln(p/q)\} =: KL[p, q]$ is convex
- $\nabla_p KL[p, q] = \ln \frac{p}{q} = \beta\, f$:

$$p(\beta) = e^{\beta\, f - \Gamma(\beta)}\, q$$



Generating Functions

- $P(\omega)$ gives us all moments of $x : \mathcal{A} \subseteq 2^{\Omega} \to \mathbb{R}$:

$$\mathbb{E}_P\{x\}\,,\ \mathbb{E}_P\{x^2\}\,,\ \mathbb{E}_P\{x^3\}\ldots$$

Generating Functions

- $P(\omega)$ gives us all moments of $x : \mathcal{A} \subseteq 2^\Omega \to \mathbb{R}$:

$$\mathbb{E}_P\{x\}\,,\ \mathbb{E}_P\{x^2\}\,,\ \mathbb{E}_P\{x^3\}\dots$$

- Note that

$$\mathbb{E}\{x^n\} = \frac{1}{i^n}\frac{\partial^n \Theta(u)}{\partial u^n}\bigg|_{u=0}$$

of the characteristic function $\Theta(u) = \mathbb{E}_P\{e^{iux}\}$.

Generating Functions

- $P(\omega)$ gives us all moments of $x : \mathcal{A} \subseteq 2^\Omega \to \mathbb{R}$:

$$\mathbb{E}_P\{x\}\,,\ \mathbb{E}_P\{x^2\}\,,\ \mathbb{E}_P\{x^3\}\dots$$

- Note that

$$\mathbb{E}\{x^n\} = \frac{1}{i^n}\frac{\partial^n\Theta(u)}{\partial u^n}\bigg|_{u=0}$$

of the characteristic function $\Theta(u) = \mathbb{E}_P\{e^{iux}\}$.

- $\Theta(u)$ is the Fourier transform of P, so that

$$P(x) = \frac{1}{2\pi}\int_U \Theta(u)e^{-ixu}\,du$$

Generating Functions

- $P(\omega)$ gives us all moments of $x : \mathcal{A} \subseteq 2^\Omega \to \mathbb{R}$:

$$\mathbb{E}_P\{x\}\,,\ \mathbb{E}_P\{x^2\}\,,\ \mathbb{E}_P\{x^3\}\ldots$$

- Note that

$$\mathbb{E}\{x^n\} = \frac{1}{i^n}\frac{\partial^n \Theta(u)}{\partial u^n}\bigg|_{u=0}$$

  of the characteristic function $\Theta(u) = \mathbb{E}_P\{e^{iux}\}$.

- $\Theta(u)$ is the Fourier transform of P, so that

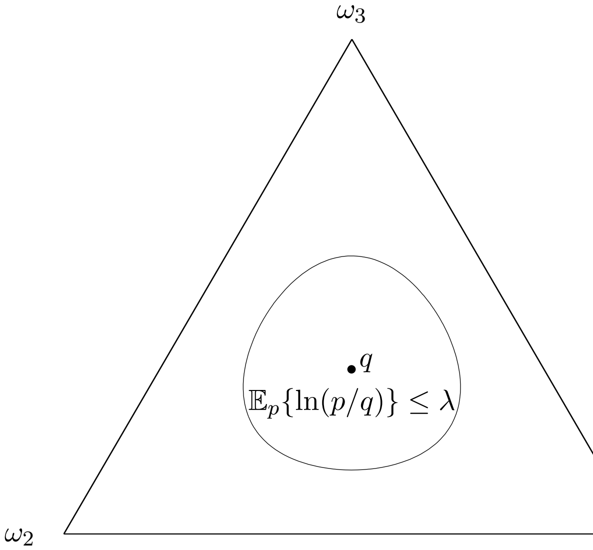
$$P(x) = \frac{1}{2\pi}\int_U \Theta(u)e^{-ixu}\,du$$

- $\Gamma[u] := \ln\Theta(u)$ is the kumulant generating function.

The KL-divergence and its Dual $\Gamma[u] = \ln \Theta(u)$

- The KL-divergence between p, $q \in \mathcal{P}(\Omega)$:

$$KL[p,q] := \mathbb{E}_P\{\ln(p/q)\}$$



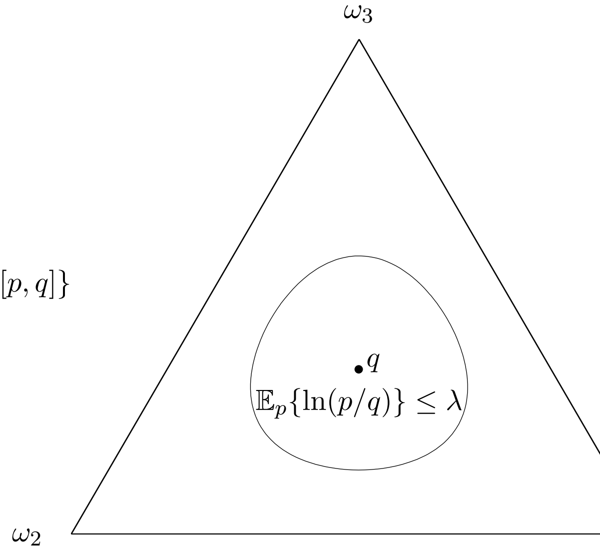
The KL-divergence and its Dual $\Gamma[u] = \ln\Theta(u)$

- The KL-divergence between p, $q \in \mathcal{P}(\Omega)$:

$$KL[p,q] := \mathbb{E}_P\{\ln(p/q)\}$$

- Its Legendre-Fenchel transform is $\Gamma[u] := \ln\Theta(u)$:

$$\Gamma[u] = \sup_p\{\mathbb{E}_p\{u\} - KL[p,q]\}$$
$$= \ln\mathbb{E}_q\{e^u\}$$



The KL-divergence and its Dual $\Gamma[u] = \ln \Theta(u)$

- The KL-divergence between p, $q \in \mathcal{P}(\Omega)$:

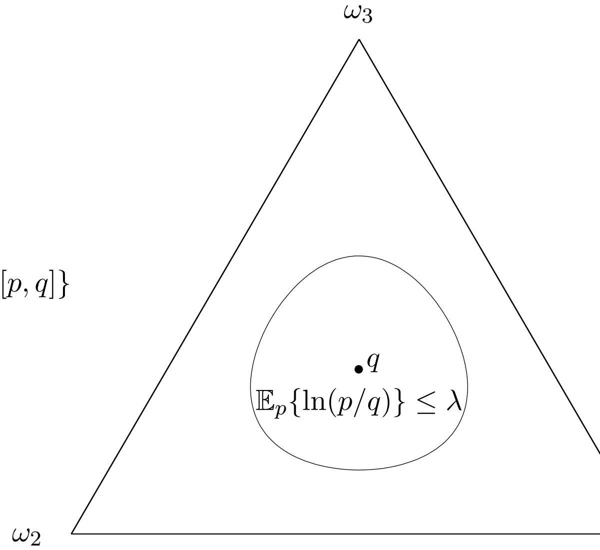
$$KL[p,q] := \mathbb{E}_P\{\ln(p/q)\}$$

- Its Legendre-Fenchel transform is $\Gamma[u] := \ln \Theta(u)$:

$$\Gamma[u] = \sup_p\{\mathbb{E}_p\{u\} - KL[p,q]\}$$
$$= \ln \mathbb{E}_q\{e^u\}$$

- Free energy:

$$F(\beta^{-1}) = -\beta^{-1}\Gamma[\beta\, u]$$



Markov Processes and PDE

Consider a transformation of $p(x)$ into $p_\tau(x_\tau)$ after time τ:

$$p_\tau(x_\tau) = \int_X p(x_\tau \mid x)\, p(x)\, dx$$

Markov Processes and PDE

Consider a transformation of $p(x)$ into $p_\tau(x_\tau)$ after time τ:

$$\begin{aligned} p_\tau(x_\tau) &= \int_X p(x_\tau \mid x)\, p(x)\, dx \\ &= \int_X \left[\frac{1}{2\pi}\int_U \Theta(u,x)\, e^{-iu(x_\tau - x)}\, du\right] p(x)\, dx \end{aligned}$$

Markov Processes and PDE

Consider a transformation of $p(x)$ into $p_\tau(x_\tau)$ after time τ:

$$\begin{aligned}
p_\tau(x_\tau) &= \int_X p(x_\tau \mid x)\, p(x)\, dx \\
&= \int_X \left[\frac{1}{2\pi} \int_U \Theta(u, x)\, e^{-iu(x_\tau - x)}\, du \right] p(x)\, dx \\
&= \int_X \left[\frac{1}{2\pi} \int_U \sum_{n=0}^{\infty} \frac{m_n(x)}{n!}\, (iu)^n e^{-iu(x_\tau - x)}\, du \right] p(x)\, dx
\end{aligned}$$

Markov Processes and PDE

Consider a transformation of $p(x)$ into $p_\tau(x_\tau)$ after time τ:

$$\begin{aligned}
p_\tau(x_\tau) &= \int_X p(x_\tau \mid x)\, p(x)\, dx \\
&= \int_X \left[\frac{1}{2\pi}\int_U \Theta(u,x)\, e^{-iu(x_\tau - x)}\, du\right] p(x)\, dx \\
&= \int_X \left[\frac{1}{2\pi}\int_U \sum_{n=0}^{\infty} \frac{m_n(x)}{n!}\, (iu)^n e^{-iu(x_\tau - x)}\, du\right] p(x)\, dx \\
&= \int_X \left[\frac{1}{2\pi}\int_U \sum_{n=0}^{\infty} \frac{m_n(x)}{n!} \left(-\frac{\partial}{\partial x_\tau}\right)^n e^{-iu(x_\tau - x)}\, du\right] p(x)\, dx
\end{aligned}$$

Markov Processes and PDE

Consider a transformation of $p(x)$ into $p_\tau(x_\tau)$ after time τ:

$$\begin{aligned}
p_\tau(x_\tau) &= \int_X p(x_\tau \mid x)\, p(x)\, dx \\
&= \int_X \left[\frac{1}{2\pi}\int_U \Theta(u,x)\, e^{-iu(x_\tau - x)}\, du\right] p(x)\, dx \\
&= \int_X \left[\frac{1}{2\pi}\int_U \sum_{n=0}^{\infty} \frac{m_n(x)}{n!}\,(iu)^n e^{-iu(x_\tau - x)}\, du\right] p(x)\, dx \\
&= \int_X \left[\frac{1}{2\pi}\int_U \sum_{n=0}^{\infty} \frac{m_n(x)}{n!}\left(-\frac{\partial}{\partial x_\tau}\right)^n e^{-iu(x_\tau - x)}\, du\right] p(x)\, dx \\
&= \int_X \sum_{n=0}^{\infty} \frac{1}{n!}\left(-\frac{\partial}{\partial x_\tau}\right)^n [m_n(x)\, p(x)] \underbrace{\left[\frac{1}{2\pi}\int_U e^{-iu(x_\tau - x)}\, du\right]}_{\delta(x_\tau - x)} dx
\end{aligned}$$

Markov Processes and PDE

Consider a transformation of $p(x)$ into $p_\tau(x_\tau)$ after time τ:

$$
\begin{aligned}
p_\tau(x_\tau) &= \int_X p(x_\tau \mid x)\, p(x)\, dx \\
&= \int_X \left[\frac{1}{2\pi}\int_U \Theta(u,x)\, e^{-iu(x_\tau - x)}\, du\right] p(x)\, dx \\
&= \int_X \left[\frac{1}{2\pi}\int_U \sum_{n=0}^{\infty} \frac{m_n(x)}{n!}\, (iu)^n e^{-iu(x_\tau - x)}\, du\right] p(x)\, dx \\
&= \int_X \left[\frac{1}{2\pi}\int_U \sum_{n=0}^{\infty} \frac{m_n(x)}{n!}\, \left(-\frac{\partial}{\partial x_\tau}\right)^n e^{-iu(x_\tau - x)}\, du\right] p(x)\, dx \\
&= \int_X \sum_{n=0}^{\infty} \frac{1}{n!}\left(-\frac{\partial}{\partial x_\tau}\right)^n [m_n(x)\, p(x)] \underbrace{\left[\frac{1}{2\pi}\int_U e^{-iu(x_\tau - x)}\, du\right]}_{\delta(x_\tau - x)} dx \\
&= p(x_\tau) + \sum_{n=1}^{\infty} \frac{1}{n!}\left(-\frac{\partial}{\partial x_\tau}\right)^n [m_n(x_\tau)\, p(x_\tau)]
\end{aligned}
$$

The Kinetic equation

- Thus, p_τ turns out to be ‘expanded’ at p:

$$p_\tau(x_\tau) = p(x_\tau) + \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x_\tau} \right)^n [m_n(x_\tau)\, p(x_\tau)]$$

The Kinetic equation

- Thus, p_τ turns out to be 'expanded' at p:

$$p_\tau(x_\tau) = p(x_\tau) + \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x_\tau}\right)^n [m_n(x_\tau)\, p(x_\tau)]$$

- The quotient of $p_\tau - p$ and τ is

$$\frac{p_\tau(x_\tau) - p(x_\tau)}{\tau} = \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x_\tau}\right)^n \left[\frac{m_n(x_\tau)}{\tau}\, p(x_\tau)\right]$$

The Kinetic equation

- Thus, p_τ turns out to be 'expanded' at p:

$$p_\tau(x_\tau) = p(x_\tau) + \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x_\tau}\right)^n [m_n(x_\tau)\, p(x_\tau)]$$

- The quotient of $p_\tau - p$ and τ is

$$\frac{p_\tau(x_\tau) - p(x_\tau)}{\tau} = \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x_\tau}\right)^n \left[\frac{m_n(x_\tau)}{\tau}\, p(x_\tau)\right]$$

- In the limit $\tau \to 0$, we obtain the kinetic equation:

$$\dot{p}(x) = \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x}\right)^n [K_n(x)\, p(x)]$$

where $K_n(x) := \lim_{\tau \to 0} \frac{m_n(x)}{\tau}$

The Fokker-Planck equation

Definition (Continuous Markov process)

- A Markov process, for which $K_n = 0$ for all $n > 2$.

The Fokker-Planck equation

Definition (Continuous Markov process)

- A Markov process, for which $K_n = 0$ for all $n > 2$.
- In this case, the kinetic equation is

$$\dot{p}(x) = -\frac{\partial}{\partial x}[K_1(x)\,p(x)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[K_2(x)\,p(x)]$$

The Fokker-Planck equation

Definition (Continuous Markov process)

- A Markov process, for which $K_n = 0$ for all $n > 2$.
- In this case, the kinetic equation is

$$\dot{p}(x) = -\frac{\partial}{\partial x}[K_1(x)\,p(x)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[K_2(x)\,p(x)]$$

- Coefficient $K_1(x)$ is called drift.

The Fokker-Planck equation

Definition (Continuous Markov process)

- A Markov process, for which $K_n = 0$ for all $n > 2$.
- In this case, the kinetic equation is

$$\dot{p}(x) = -\frac{\partial}{\partial x}[K_1(x)\,p(x)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[K_2(x)\,p(x)]$$

- Coefficient $K_1(x)$ is called drift.
- Coefficient $K_2(x)$ is called diffusion.

The Fokker-Planck equation

Definition (Continuous Markov process)

- A Markov process, for which $K_n = 0$ for all $n > 2$.
- In this case, the kinetic equation is

$$\dot{p}(x) = -\frac{\partial}{\partial x}[K_1(x)\,p(x)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[K_2(x)\,p(x)]$$

- Coefficient $K_1(x)$ is called drift.
- Coefficient $K_2(x)$ is called diffusion.

Example (Diffusion equation)

If the drift $K_1 = 0$ and diffusion $K_2 = 1$, then

$$\frac{\partial p(x,t)}{\partial t} = \frac{1}{2}\frac{\partial^2 p(x,t)}{\partial x^2}$$

This corresponds to the Wiener process $\{x(t)\}_{t\in(0,T)}$.

Noisy observation of a Hidden Markov process

- $\{x(t)\}_{t\in[0,T]}$ — unobserved Markov process with $\pi(x_m \mid x_{m-1})$.

Noisy observation of a Hidden Markov process

- $\{x(t)\}_{t\in[0,T]}$ — unobserved Markov process with $\pi(x_m \mid x_{m-1})$.
- $\{z(t)\}_{t\in[0,T]}$ — observed process related to x by $z = \varphi(x, \xi)$, where ξ and x are independent, and $\xi = \varphi^{-1}(x, z)$.

Noisy observation of a Hidden Markov process

- $\{x(t)\}_{t\in[0,T]}$ — unobserved Markov process with $\pi(x_m \mid x_{m-1})$.
- $\{z(t)\}_{t\in[0,T]}$ — observed process related to x by $z = \varphi(x, \xi)$, where ξ and x are independent, and $\xi = \varphi^{-1}(x, z)$.
- Denote the posterior density by

$$w_m(x_m) := p_m(x_m \mid z_m, \dots, z_1)$$

Noisy observation of a Hidden Markov process

- $\{x(t)\}_{t\in[0,T]}$ — unobserved Markov process with $\pi(x_m \mid x_{m-1})$.
- $\{z(t)\}_{t\in[0,T]}$ — observed process related to x by $z = \varphi(x,\xi)$, where ξ and x are independent, and $\xi = \varphi^{-1}(x,z)$.
- Denote the posterior density by

$$w_m(x_m) := p_m(x_m \mid z_m, \dots, z_1)$$

- $w_m(x_m)$ defintes *filtering* of x_m from observations $(z_1, \dots, z_m)$.

Noisy observation of a Hidden Markov process

- $\{x(t)\}_{t\in[0,T]}$ — unobserved Markov process with $\pi(x_m \mid x_{m-1})$.
- $\{z(t)\}_{t\in[0,T]}$ — observed process related to x by $z = \varphi(x, \xi)$, where ξ and x are independent, and $\xi = \varphi^{-1}(x, z)$.
- Denote the posterior density by

$$w_m(x_m) := p_m(x_m \mid z_m, \dots, z_1)$$

- $w_m(x_m)$ defintes *filtering* of x_m from observations $(z_1, \dots, z_m)$.
- Let $L_m(x_m, \dots, x_1) = p(z_m, \dots, z_1 \mid x_m, \dots, x_1)$ denote the *likelihood function*.

Noisy observation of a Hidden Markov process

- $\{x(t)\}_{t\in[0,T]}$ — unobserved Markov process with $\pi(x_m \mid x_{m-1})$.
- $\{z(t)\}_{t\in[0,T]}$ — observed process related to x by $z = \varphi(x, \xi)$, where ξ and x are independent, and $\xi = \varphi^{-1}(x, z)$.
- Denote the posterior density by

$$w_m(x_m) := p_m(x_m \mid z_m, \dots, z_1)$$

- $w_m(x_m)$ defintes *filtering* of x_m from observations $(z_1, \dots, z_m)$.
- Let $L_m(x_m, \dots, x_1) = p(z_m, \dots, z_1 \mid x_m, \dots, x_1)$ denote the *likelihood function*.
- The aim is to define a recursive relation between $w_m(x_m)$ and $w_{m-1}(x_{m-1})$.

Noisy observation of a Hidden Markov process

- $\{x(t)\}_{t\in[0,T]}$ — unobserved Markov process with $\pi(x_m \mid x_{m-1})$.
- $\{z(t)\}_{t\in[0,T]}$ — observed process related to x by $z = \varphi(x, \xi)$, where ξ and x are independent, and $\xi = \varphi^{-1}(x, z)$.
- Denote the posterior density by

$$w_m(x_m) := p_m(x_m \mid z_m, \ldots, z_1)$$

- $w_m(x_m)$ defintes *filtering* of x_m from observations $(z_1, \ldots, z_m)$.
- Let $L_m(x_m, \ldots, x_1) = p(z_m, \ldots, z_1 \mid x_m, \ldots, x_1)$ denote the *likelihood function*.
- The aim is to define a recursive relation between $w_m(x_m)$ and $w_{m-1}(x_{m-1})$.
- This will allow us to derive a (stochastic) differential equation for $w(x, t)$.

Recursive posterior densities

Theorem (Stratonovich (1959b, 1959a, 1960))

If the likelihood function is multiplicative $L_m(x_m,\dots,x_1)=p(z_m,\dots,z_1\mid x_m,\dots,x_1)=\prod_{t=1}^m p(z_t\mid x_t)$, *then*

$$w_m(x_m)=C_m L_m(x_m)\int \pi(x_m\mid x_{m-1})w_{m-1}(x_{m-1})\,dx_{x-1}$$

Recursive posterior densities

Theorem (Stratonovich (1959b, 1959a, 1960))

If the likelihood function is multiplicative $L_m(x_m,\dots,x_1) = p(z_m,\dots,z_1 \mid x_m,\dots,x_1) = \prod_{t=1}^m p(z_t \mid x_t)$, *then*

$$w_m(x_m) = C_m L_m(x_m) \int \pi(x_m \mid x_{m-1}) w_{m-1}(x_{m-1})\, dx_{x-1}$$

Proof.

$$p(x_m,\dots,x_1 \mid z_m,\dots,z_1) = C_m p(z_m,\dots,z_1 \mid x_m,\dots,x_1) p(x_m,\dots,x_1)$$

Recursive posterior densities

Theorem (Stratonovich (1959b, 1959a, 1960))

If the likelihood function is multiplicative $L_m(x_m, \ldots, x_1) = p(z_m, \ldots, z_1 \mid x_m, \ldots, x_1) = \prod_{t=1}^{m} p(z_t \mid x_t)$, *then*

$$w_m(x_m) = C_m L_m(x_m) \int \pi(x_m \mid x_{m-1}) w_{m-1}(x_{m-1})\, dx_{x-1}$$

Proof.

$$p(x_m, \ldots, x_1 \mid z_m, \ldots, z_1) = C_m L_m(x_m) \left[\prod_{t=1}^{m-1} p(z_t \mid x_t) \pi(x_{t+1} \mid x_t) \right] \pi(x_1)$$

Recursive posterior densities

Theorem (Stratonovich (1959b, 1959a, 1960))

If the likelihood function is multiplicative $L_m(x_m, \dots, x_1) = p(z_m, \dots, z_1 \mid x_m, \dots, x_1) = \prod_{t=1}^{m} p(z_t \mid x_t)$, *then*

$$w_m(x_m) = C_m L_m(x_m) \int \pi(x_m \mid x_{m-1}) w_{m-1}(x_{m-1})\, dx_{x-1}$$

Proof.

$$p(x_m, \dots, x_1 \mid z_m, \dots, z_1) = C_m L_m(x_m) \left[\prod_{t=1}^{m-1} p(z_t \mid x_t) \pi(x_{t+1} \mid x_t) \right] \pi(x_1)$$

Then

$$\frac{p(x_m, \dots, x_1 \mid z_m, \dots, z_1)}{p(x_{m-1}, \dots, x_1 \mid z_{m-1}, \dots, z_1)} = \frac{C_m}{C_{m-1}} L_m(x_m) \pi(x_m \mid x_{m-1})$$

Recursive posterior densities

Theorem (Stratonovich (1959b, 1959a, 1960))

If the likelihood function is multiplicative
$L_m(x_m, \dots, x_1) = p(z_m, \dots, z_1 \mid x_m, \dots, x_1) = \prod_{t=1}^{m} p(z_t \mid x_t)$, *then*

$$w_m(x_m) = C_m L_m(x_m) \int \pi(x_m \mid x_{m-1}) w_{m-1}(x_{m-1})\, dx_{x-1}$$

Proof.

$$p(x_m, \dots, x_1 \mid z_m, \dots, z_1) = C_m L_m(x_m) \left[\prod_{t=1}^{m-1} p(z_t \mid x_t) \pi(x_{t+1} \mid x_t) \right] \pi(x_1)$$

Then

$$\frac{p(x_m, x_{m-1} \mid z_m, \dots, z_1)}{p(x_{m-1} \mid z_{m-1}, \dots, z_1)} = \frac{C_m}{C_{m-1}} L_m(x_m) \pi(x_m \mid x_{m-1})$$

The Filtering Equation

Theorem (Stratonovich-Kushner-Zakai)

Let $\{x(t)\}_{t\in[0,T]}$ *be a continuous Markov process, and* $\{z(t)\}_{t\in[0,T]}$ *be the observed process given by*

$$dx(t) = a(x,t)dt + b(x,t)\,dw(t)\,, \qquad dz(t) = s(x,t)dt + \sqrt{N_0/2}\,dv(t)$$

where v *is the standard Gaussian white noise.*

The Filtering Equation

Theorem (Stratonovich-Kushner-Zakai)

Let $\{x(t)\}_{t\in[0,T]}$ be a continuous Markov process, and $\{z(t)\}_{t\in[0,T]}$ be the observed process given by

$$dx(t) = a(x,t)dt + b(x,t)\,dw(t)\,, \qquad dz(t) = s(x,t)dt + \sqrt{N_0/2}\,dv(t)$$

where v is the standard Gaussian white noise.
Then the (normalized) posterior density $w(x,t) := p(x_t \mid z_t, \dots, z_1)$ and the unnormalized measure $u(x,t) = w(x,t)/C(t)$ satisfy

$$dw(x,t) = \mathcal{L}[w(x,t)]\,dt + \frac{2}{N_0}w(x,t)[s(x,t) - \langle s(x,t)\rangle][dz - \langle s(x,t)\rangle\,dt]$$

$$du(x,t) = \mathcal{L}[u(x,t)]\,dt + \frac{2}{N_0}u(x,t)s(x,t)\,dz$$

The Filtering Equation

Theorem (Stratonovich-Kushner-Zakai)

Let $\{x(t)\}_{t\in[0,T]}$ be a continuous Markov process, and $\{z(t)\}_{t\in[0,T]}$ be the observed process given by

$$dx(t) = a(x,t)dt + b(x,t)\,dw(t)\,, \qquad dz(t) = s(x,t)dt + \sqrt{N_0/2}\,dv(t)$$

where v is the standard Gaussian white noise.
Then the (normalized) posterior density $w(x,t) := p(x_t \mid z_t,\dots,z_1)$ and the unnormalized measure $u(x,t) = w(x,t)/C(t)$ satisfy

$$dw(x,t) = \mathcal{L}[w(x,t)]\,dt + \frac{2}{N_0}w(x,t)[s(x,t) - \langle s(x,t)\rangle][dz - \langle s(x,t)\rangle\,dt]$$

$$du(x,t) = \mathcal{L}[u(x,t)]\,dt + \frac{2}{N_0}u(x,t)s(x,t)\,dz$$

where $\mathcal{L}[\cdot] := -\frac{\partial}{\partial x}[a(x,t)\cdot] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[b(x,t)\cdot]$ is the Kolmogorov-Fokker-Planck operator.

The Filtering Equation

Theorem (Stratonovich-Kushner-Zakai)

Let $\{x(t)\}_{t\in[0,T]}$ be a continuous Markov process, and $\{z(t)\}_{t\in[0,T]}$ be the observed process given by

$$dx(t) = a(x,t)dt + b(x,t)\,dw(t)\,, \qquad dz(t) = s(x,t)dt + \sqrt{N_0/2}\,dv(t)$$

where v is the standard Gaussian white noise.
Then the (normalized) posterior density $w(x,t) := p(x_t \mid z_t, \ldots, z_1)$ and the unnormalized measure $u(x,t) = w(x,t)/C(t)$ satisfy

$$dw(x,t) = \mathcal{L}[w(x,t)]\,dt + \frac{2}{N_0}w(x,t)[s(x,t) - \langle s(x,t)\rangle]\,\underbrace{[dz - \langle s(x,t)\rangle\,dt]}_{\textit{Innovation}}$$

$$du(x,t) = \mathcal{L}[u(x,t)]\,dt + \frac{2}{N_0}u(x,t)s(x,t)\,dz$$

where $\mathcal{L}[\cdot] := -\frac{\partial}{\partial x}[a(x,t)\cdot] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[b(x,t)\cdot]$ is the Kolmogorov-Fokker-Planck operator.

Proof sketch.

Consider variations $\delta z_m := \int_{\Delta t_m} z(x,t)\,dt = s_m \Delta t + \sqrt{N_0/2}\delta v_m$ during small Δt and use recursion:

$$\frac{w_m(x_m)}{C_m} = L_m(x_m) \int \pi(x_m \mid x_{m-1}) w_{m-1}(x_{m-1})\, dx_{m-1}$$

$\square$

Proof sketch.

Consider variations $\delta z_m := \int_{\Delta t_m} z(x,t)\,dt = s_m \Delta t + \sqrt{N_0/2}\delta v_m$ during small Δt and use recursion:

$$\frac{w_m(x_m)}{C_m} = L_m(x_m)\int \pi(x_m \mid x_{m-1}) w_{m-1}(x_{m-1})\,dx_{m-1}$$
$$= L_m(x_m)\int \{\delta(x_m - x_{m-1}) + \Delta t \mathcal{L}[\delta(x_m - x_{m-1})]\}\, w_{m-1}(x_{m-1})$$

$\square$

Proof sketch.

Consider variations $\delta z_m := \int_{\Delta t_m} z(x,t)\,dt = s_m \Delta t + \sqrt{N_0/2}\delta v_m$ during small Δt and use recursion:

$$\begin{aligned}\frac{w_m(x_m)}{C_m} &= L_m(x_m) \int \pi(x_m \mid x_{m-1}) w_{m-1}(x_{m-1})\, dx_{m-1} \\ &= L_m(x_m) \int \{\delta(x_m - x_{m-1}) + \Delta t \mathcal{L}[\delta(x_m - x_{m-1})]\}\, w_{m-1}(x_{m-1}) \\ &= L_m(x_m) \{w_{m-1}(x_m) + \Delta t \mathcal{L}[w_{m-1}(x_m)]\}\end{aligned}$$

$\square$

Proof sketch.

Consider variations $\delta z_m := \int_{\Delta t_m} z(x,t)\,dt = s_m\Delta t + \sqrt{N_0/2}\delta v_m$ during small Δt and use recursion:

$$\begin{aligned}
\frac{w_m(x_m)}{C_m} &= L_m(x_m)\int \pi(x_m \mid x_{m-1}) w_{m-1}(x_{m-1})\,dx_{m-1} \\
&= L_m(x_m)\int \{\delta(x_m - x_{m-1}) + \Delta t\mathcal{L}[\delta(x_m - x_{m-1})]\}\, w_{m-1}(x_{m-1}) \\
&= e^{-\frac{1}{N_0\Delta t}[\delta z_m - s_m\Delta t]^2}\,\{w_{m-1}(x_m) + \Delta t\mathcal{L}[w_{m-1}(x_m)]\}
\end{aligned}$$

$\square$

Proof sketch.

Consider variations $\delta z_m := \int_{\Delta t_m} z(x,t)\,dt = s_m \Delta t + \sqrt{N_0/2}\delta v_m$ during small Δt and use recursion:

$$\begin{aligned}
\frac{w_m(x_m)}{C_m} &= L_m(x_m) \int \pi(x_m \mid x_{m-1}) w_{m-1}(x_{m-1})\,dx_{m-1} \\
&= L_m(x_m) \int \{\delta(x_m - x_{m-1}) + \Delta t \mathcal{L}[\delta(x_m - x_{m-1})]\}\, w_{m-1}(x_{m-1}) \\
&= \left[1 + \frac{2}{N_0}\delta z_m s_m + o(\Delta t)\right] \{w_{m-1}(x_m) + \Delta t \mathcal{L}[w_{m-1}(x_m)]\}
\end{aligned}$$

$\square$

Proof sketch.

Consider variations $\delta z_m := \int_{\Delta t_m} z(x,t)\,dt = s_m \Delta t + \sqrt{N_0/2}\delta v_m$ during small Δt and use recursion:

$$\frac{w_m(x_m)}{C_m} = L_m(x_m) \int \pi(x_m \mid x_{m-1}) w_{m-1}(x_{m-1})\,dx_{m-1}$$

$$= L_m(x_m) \int \{\delta(x_m - x_{m-1}) + \Delta t \mathcal{L}[\delta(x_m - x_{m-1})]\}\, w_{m-1}(x_{m-1})$$

$$= \left[1 + \frac{2}{N_0}\delta z_m s_m + o(\Delta t)\right] \{w_{m-1}(x_m) + \Delta t \mathcal{L}[w_{m-1}(x_m)]\}$$

$$= w_{m-1}(x_m) + \Delta t \mathcal{L}[w_{m-1}(x_m)] + \frac{2}{N_0}\delta z_m s_m w_{m-1}(x_m) + o(\Delta t)$$

□

Proof sketch.

Consider variations $\delta z_m := \int_{\Delta t_m} z(x,t)\,dt = s_m \Delta t + \sqrt{N_0/2}\delta v_m$ during small Δt and use recursion:

$$\begin{aligned}\frac{w_m(x_m)}{C_m} &= L_m(x_m)\int \pi(x_m \mid x_{m-1}) w_{m-1}(x_{m-1})\,dx_{m-1} \\ &= L_m(x_m)\int \{\delta(x_m - x_{m-1}) + \Delta t \mathcal{L}[\delta(x_m - x_{m-1})]\}\, w_{m-1}(x_{m-1}) \\ &= \left[1 + \frac{2}{N_0}\delta z_m s_m + o(\Delta t)\right]\{w_{m-1}(x_m) + \Delta t \mathcal{L}[w_{m-1}(x_m)]\} \\ &= w_{m-1}(x_m) + \Delta t \mathcal{L}[w_{m-1}(x_m)] + \frac{2}{N_0}\delta z_m s_m w_{m-1}(x_m) + o(\Delta t)\end{aligned}$$

where we used the Levy's rule $\mathbb{E}\{\delta z^2\} = (N_0/2)\Delta t + o(\Delta t)$.

$\square$

Proof sketch.

Consider variations $\delta z_m := \int_{\Delta t_m} z(x,t)\,dt = s_m\Delta t + \sqrt{N_0/2}\delta v_m$ during small Δt and use recursion:

$$\begin{aligned}
\frac{w_m(x_m)}{C_m} &= L_m(x_m)\int \pi(x_m \mid x_{m-1})w_{m-1}(x_{m-1})\,dx_{m-1} \\
&= L_m(x_m)\int \{\delta(x_m - x_{m-1}) + \Delta t\mathcal{L}[\delta(x_m - x_{m-1})]\}\,w_{m-1}(x_{m-1}) \\
&= \left[1 + \frac{2}{N_0}\delta z_m s_m + o(\Delta t)\right]\{w_{m-1}(x_m) + \Delta t\mathcal{L}[w_{m-1}(x_m)]\} \\
&= w_{m-1}(x_m) + \Delta t\mathcal{L}[w_{m-1}(x_m)] + \frac{2}{N_0}\delta z_m s_m w_{m-1}(x_m) + o(\Delta t)
\end{aligned}$$

where we used the Levy's rule $\mathbb{E}\{\delta z^2\} = (N_0/2)\Delta t + o(\Delta t)$.

$$C_m^{-1} = 1 + \frac{2}{N_0}\delta z_m\langle s_m\rangle + o(\Delta t)$$

□

Proof sketch.

Consider variations $\delta z_m := \int_{\Delta t_m} z(x,t)\,dt = s_m \Delta t + \sqrt{N_0/2}\delta v_m$ during small Δt and use recursion:

$$\begin{aligned}\frac{w_m(x_m)}{C_m} &= L_m(x_m)\int \pi(x_m \mid x_{m-1}) w_{m-1}(x_{m-1})\,dx_{m-1}\\ &= L_m(x_m)\int \{\delta(x_m - x_{m-1}) + \Delta t \mathcal{L}[\delta(x_m - x_{m-1})]\}\, w_{m-1}(x_{m-1})\\ &= \left[1 + \frac{2}{N_0}\delta z_m s_m + o(\Delta t)\right]\{w_{m-1}(x_m) + \Delta t \mathcal{L}[w_{m-1}(x_m)]\}\\ &= w_{m-1}(x_m) + \Delta t \mathcal{L}[w_{m-1}(x_m)] + \frac{2}{N_0}\delta z_m s_m w_{m-1}(x_m) + o(\Delta t)\end{aligned}$$

where we used the Levy's rule $\mathbb{E}\{\delta z^2\} = (N_0/2)\Delta t + o(\Delta t)$.

$$C_m = 1 - \frac{2}{N_0}\delta z_m \langle s_m \rangle + \frac{2}{N_0}\langle s_m \rangle^2 + o(\Delta t)$$

□

Multiway data and tensors

- Data is often represented by matrices.

Multiway data and tensors

- Data is often represented by matrices.
- One of the main reasons is SVD that allows one to approximate a matrix of high rank by a matrix of a smaller rank (compression).

Multiway data and tensors

- Data is often represented by matrices.
- One of the main reasons is SVD that allows one to approximate a matrix of high rank by a matrix of a smaller rank (compression).
- A lot of data is multiway and is more naturally represented by tensors (e.g. video, stocks).

Multiway data and tensors

- Data is often represented by matrices.
- One of the main reasons is SVD that allows one to approximate a matrix of high rank by a matrix of a smaller rank (compression).
- A lot of data is multiway and is more naturally represented by tensors (e.g. video, stocks).
- There was no analogous SVD for tensors until recently, so tensors are usually converted to matrixes (matricization) leading to a loss of some information.

Multiway data and tensors

- Data is often represented by matrices.
- One of the main reasons is SVD that allows one to approximate a matrix of high rank by a matrix of a smaller rank (compression).
- A lot of data is multiway and is more naturally represented by tensors (e.g. video, stocks).
- There was no analogous SVD for tensors until recently, so tensors are usually converted to matrixes (matricization) leading to a loss of some information.
- Recent work by (Kilmer et al., 2021) defined new tensor-tensor algebra with new t-products and t-SVDs.

Multiway data and tensors

- Data is often represented by matrices.
- One of the main reasons is SVD that allows one to approximate a matrix of high rank by a matrix of a smaller rank (compression).
- A lot of data is multiway and is more naturally represented by tensors (e.g. video, stocks).
- There was no analogous SVD for tensors until recently, so tensors are usually converted to matrixes (matricization) leading to a loss of some information.
- Recent work by (Kilmer et al., 2021) defined new tensor-tensor algebra with new t-products and t-SVDs.
- The main result proving optimality of such decompositions.

Kilmer, M. E., Horesh, L., Avron, H., & Newman, E. (2021). Tensor-tensor algebra for optimal representation and compression of multiway data. *PNAS*, *118*(28).

McCulloch, W., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115-133.

Shannon, C. E. (1948, July and October). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423 and 623–656.

Stratonovich, R. L. (1959a). On the theory of optimal non-linear filtration of random functions. *Theory of Probability and its Applications*, *4*, 223–225. (English translation)

Stratonovich, R. L. (1959b). Optimum nonlinear systems which bring about a separation of a signal with constant parameters from noise. *Radiofizika*, *2*(6), 892–901.

Stratonovich, R. L. (1960). Conditional Markov processes. *Theory of Probability and its Applications (USSR)*, *5*(2), 156–178.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.