Mathematics for Deep Learning: The Value of Information Theory

Roman V. Belavkin[1] Panos Pardalos[2] Jose Principe[3]

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

[2]Department of Industrial and Systems Engineering, University of Florida, P.O. Box 116595, Gainesville, FL 32611-6595, USA

[3]Department of Electrical & Computer Engineering, University of Florida, P.O. Box 116130, Gainesville, FL 32611-6130, USA

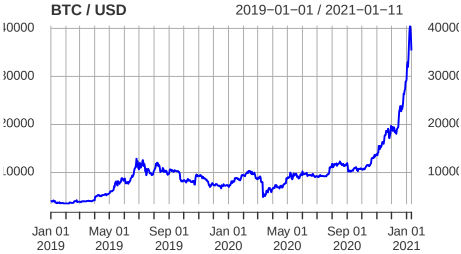
August 26, 2022
ACDL 2022

Motivating Example

Example: Time-Series Prediction

Table: BTC/USD prices $S(t)$

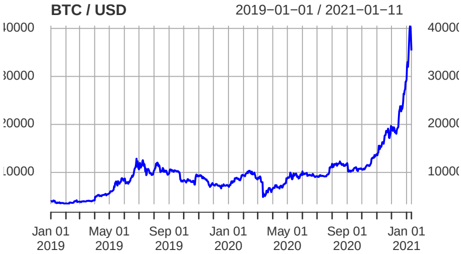
Date	Price(t)	
2019-01-01	3963.1	
2019-01-02	4048.8	
2019-01-03	3924.3	
2019-01-04	3954.9	
2019-01-05	3911.9	
2019-01-06	4168.4	
2019-01-07	4113.9	



Example: Time-Series Prediction

Table: BTC/USD prices $S(t)$

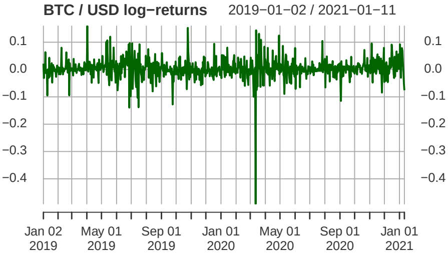
Date	Price(t)	Price(t+1)
2019-01-01	3963.1	4048.8
2019-01-02	4048.8	3924.3
2019-01-03	3924.3	3954.9
2019-01-04	3954.9	3911.9
2019-01-05	3911.9	4168.4
2019-01-06	4168.4	4113.9
2019-01-07	4113.9	?



Example: Time-Series Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

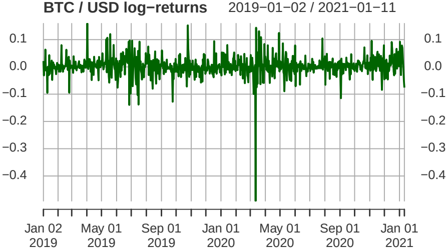
Date	$r(t)$	
2019-01-02	0.021	
2019-01-03	-0.031	
2019-01-04	0.008	
2019-01-05	-0.011	
2019-01-06	0.064	
2019-01-07	-0.013	

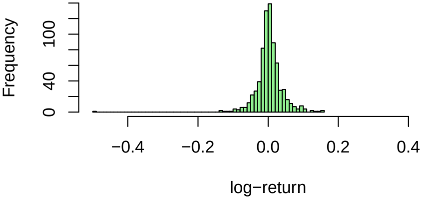


Example: Time-Series Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t)$	$r(t+1)$
2019-01-02	0.021	
2019-01-03	-0.031	
2019-01-04	0.008	
2019-01-05	-0.011	
2019-01-06	0.064	
2019-01-07	-0.013	

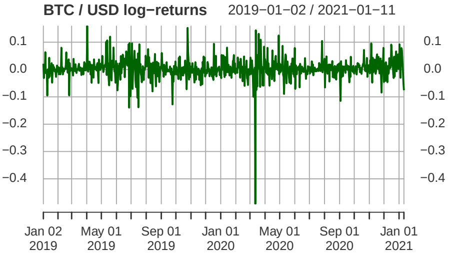


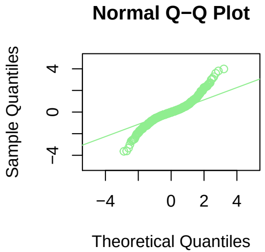


Example: Time-Series Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t)$	$r(t+1)$
2019-01-02	0.021	
2019-01-03	-0.031	
2019-01-04	0.008	
2019-01-05	-0.011	
2019-01-06	0.064	
2019-01-07	-0.013	

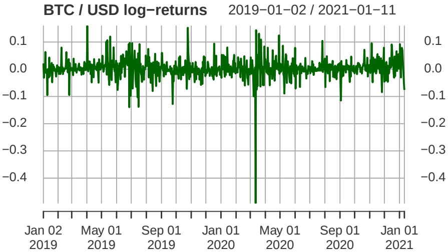


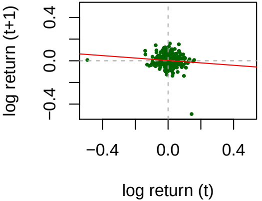


Example: Time-Series Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t)$	$r(t+1)$
2019-01-02	0.021	-0.031
2019-01-03	-0.031	0.008
2019-01-04	0.008	-0.011
2019-01-05	-0.011	0.064
2019-01-06	0.064	-0.013
2019-01-07	-0.013	?

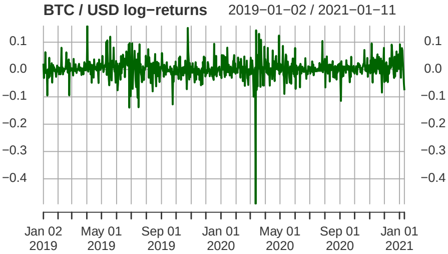


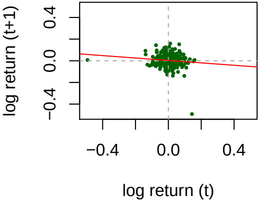


Example: Time-Series Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t)$	$r(t+1)$
2019-01-02	0.021	-0.031
2019-01-03	-0.031	0.008
2019-01-04	0.008	-0.011
2019-01-05	-0.011	0.064
2019-01-06	0.064	-0.013
2019-01-07	-0.013	?

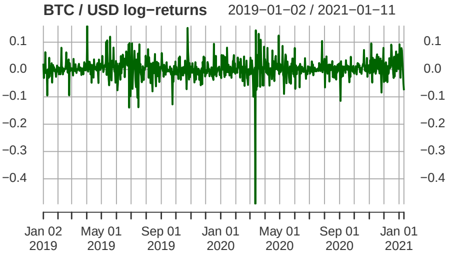




Example: Time-Series Prediction

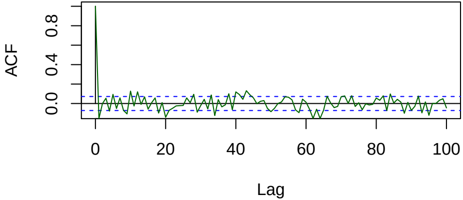
Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t)$	$r(t+1)$
2019-01-02	0.021	-0.031
2019-01-03	-0.031	0.008
2019-01-04	0.008	-0.011
2019-01-05	-0.011	0.064
2019-01-06	0.064	-0.013
2019-01-07	-0.013	?



Predict $r(t+1)$ from $r(t)$:

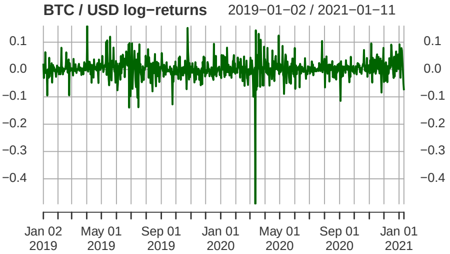
$$f\left(r(t)\right) = y \approx r(t+1)$$



Example: Time-Series Prediction

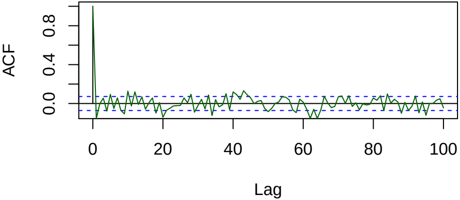
Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t-2)$	$r(t-1)$	$r(t)$	$r(t+1)$
2019-01-06	-0.031	0.008	-0.011	0.064
2019-01-07	0.008	-0.011	0.064	-0.013
2019-01-08	-0.011	0.064	-0.013	-0.0034
2019-01-09	0.064	-0.013	-0.0034	-0.004



Predict $r(t+1)$ from n lags of $r(t)$:

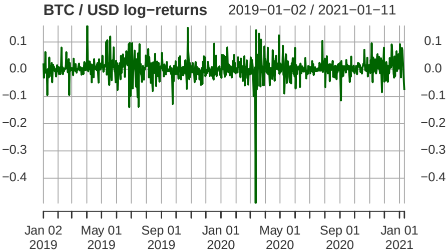
$$f\left(r(t-n), \ldots, r(t)\right) = y \approx r(t+1)$$



Example: Time-Series Prediction

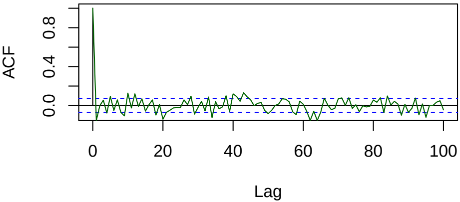
Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t-2)$	$r(t-1)$	$r(t)$	$r(t+1)$
2019-01-06	-0.031	0.008	-0.011	0.064
2019-01-07	0.008	-0.011	0.064	-0.013
2019-01-08	-0.011	0.064	-0.013	-0.0034
2019-01-09	0.064	-0.013	-0.0034	-0.004



Predict $r(t+1)$ from n lags of $r(t)$ for m symbols:

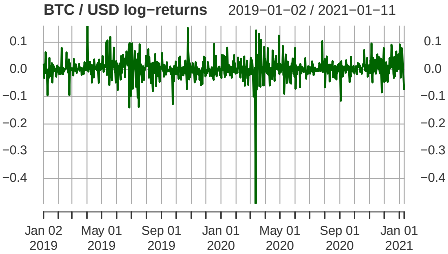
$$f\begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix} = y \approx r(t+1)$$



e.g. symbols: BTC/USD, ETH/USD, IOT/BTC, etc

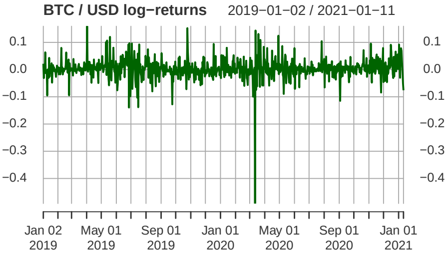
Model Performance

$$f\begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix} = y \approx r(t+1)$$



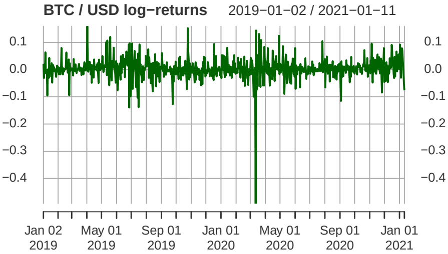
Model Performance

$$f\underbrace{\begin{pmatrix} z_{11} & \cdots & z_{1n} \\ \vdots & \ddots & \vdots \\ z_{m1} & \cdots & z_{mn} \end{pmatrix}}_{\text{predictors}} = y \approx \underbrace{x}_{\text{response}}$$



Model Performance

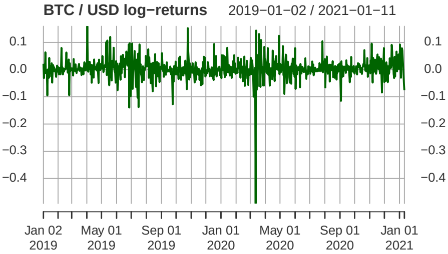
$$f\underbrace{\begin{pmatrix} z_{11} & \cdots & z_{1n} \\ \vdots & \ddots & \vdots \\ z_{m1} & \cdots & z_{mn} \end{pmatrix}}_{\text{predictors}} = y \approx \underbrace{x}_{\text{response}}$$



- Use $n \in [2 : 20]$ lags and $m \in [1 : 5]$ symbols (i.e. $m \times n \in [2 : 100]$).

Model Performance

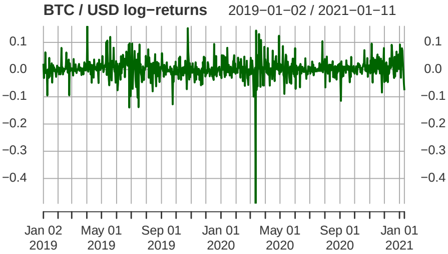
$$f\underbrace{\begin{pmatrix} z_{11} & \cdots & z_{1n} \\ \vdots & \ddots & \vdots \\ z_{m1} & \cdots & z_{mn} \end{pmatrix}}_{\text{predictors}} = y \approx \underbrace{x}_{\text{response}}$$



- Use $n \in [2:20]$ lags and $m \in [1:5]$ symbols (i.e. $m \times n \in [2:100]$).
- Models: linear regression, partial-least squares, neural net.

Model Performance

$$f\underbrace{\begin{pmatrix} z_{11} & \cdots & z_{1n} \\ \vdots & \ddots & \vdots \\ z_{m1} & \cdots & z_{mn} \end{pmatrix}}_{\text{predictors}} = y \approx \underbrace{x}_{\text{response}}$$

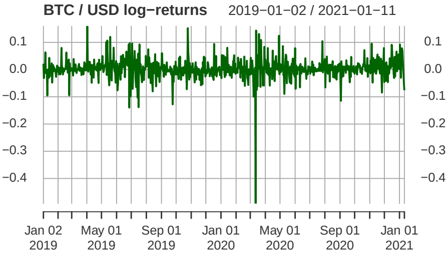


- Use $n \in [2:20]$ lags and $m \in [1:5]$ symbols (i.e. $m \times n \in [2:100]$).
- Models: linear regression, partial-least squares, neural net.
- Root mean-square error

$$\text{RMSE} = \sqrt{\mathbb{E}\{|x-y|^2\}}$$

Model Performance

$$f\underbrace{\begin{pmatrix} z_{11} & \cdots & z_{1n} \\ \vdots & \ddots & \vdots \\ z_{m1} & \cdots & z_{mn} \end{pmatrix}}_{\text{predictors}} = y \approx \underbrace{x}_{\text{response}}$$

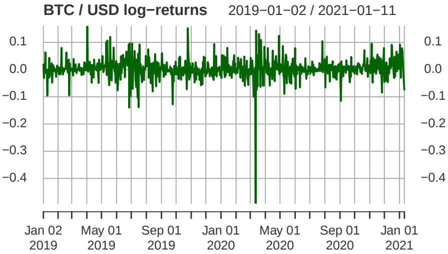


- Use $n \in [2:20]$ lags and $m \in [1:5]$ symbols (i.e. $m \times n \in [2:100]$).
- Models: linear regression, partial-least squares, neural net.
- Root mean-square error

$$\text{RMSE} = \sqrt{\mathbb{E}\{|x-y|^2\}}\,, \qquad R^2 = 1 - \text{RMSE}^2/\sigma_x^2$$

Model Performance

$$f\underbrace{\begin{pmatrix} z_{11} & \cdots & z_{1n} \\ \vdots & \ddots & \vdots \\ z_{m1} & \cdots & z_{mn} \end{pmatrix}}_{\text{predictors}} = y \approx \underbrace{x}_{\text{response}}$$



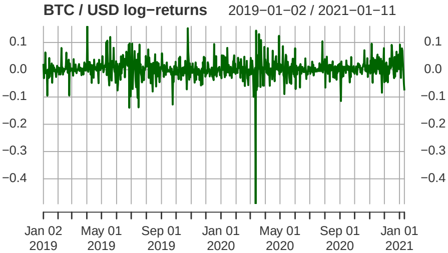
- Use $n \in [2:20]$ lags and $m \in [1:5]$ symbols (i.e. $m \times n \in [2:100]$).
- Models: linear regression, partial-least squares, neural net.
- Root mean-square error

$$\text{RMSE} = \sqrt{\mathbb{E}\{|x-y|^2\}}\,, \qquad R^2 = 1 - \text{RMSE}^2/\sigma_x^2$$

- Is RMSE $= .035$ a good result? ($R^2 \approx .05$)

Model Performance

$$f\underbrace{\begin{pmatrix} z_{11} & \cdots & z_{1n} \\ \vdots & \ddots & \vdots \\ z_{m1} & \cdots & z_{mn} \end{pmatrix}}_{\text{predictors}} = y \approx \underbrace{x}_{\text{response}}$$

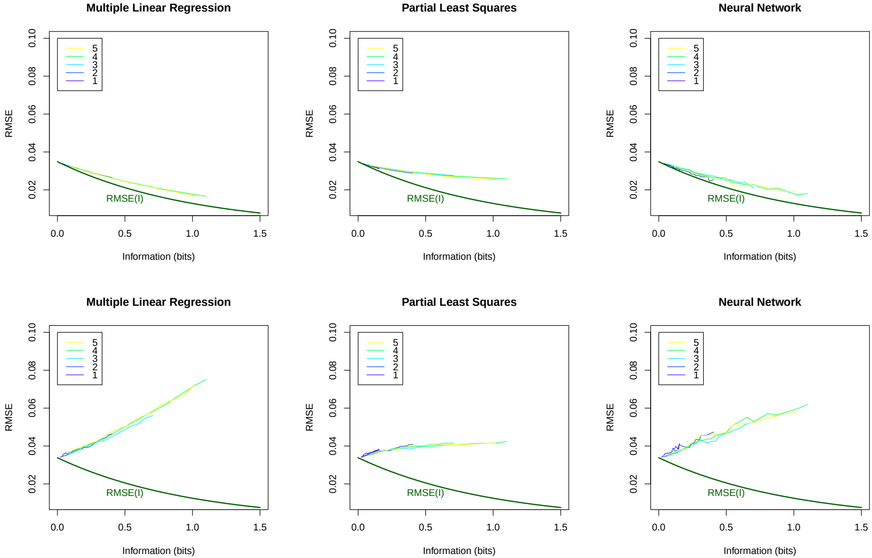


- Use $n \in [2:20]$ lags and $m \in [1:5]$ symbols (i.e. $m \times n \in [2:100]$).
- Models: linear regression, partial-least squares, neural net.
- Root mean-square error

$$\text{RMSE} = \sqrt{\mathbb{E}\{|x-y|^2\}}\,, \qquad R^2 = 1 - \text{RMSE}^2/\sigma_x^2$$

- Is RMSE $= .035$ a good result? $(R^2 \approx .05)$
- What is the smallest possible RMSE here?

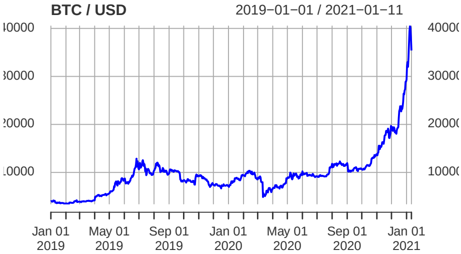
Evaluation of RMSE



Example: Binary Classification and Prediction

Table: BTC/USD prices $S(t)$

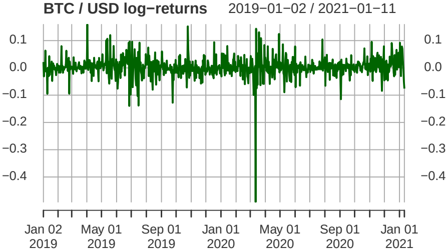
Date	Price(t)	Price(t+1)
2019-01-01	3963.1	4048.8
2019-01-02	4048.8	3924.3
2019-01-03	3924.3	3954.9
2019-01-04	3954.9	3911.9
2019-01-05	3911.9	4168.4
2019-01-06	4168.4	4113.9
2019-01-07	4113.9	?



Example: Binary Classification and Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

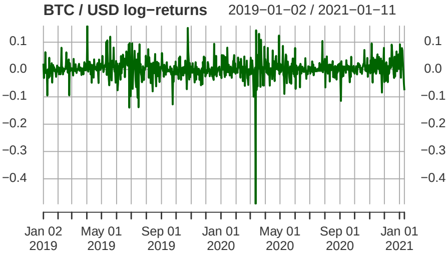
Date	$r(t)$	$r(t+1)$
2019-01-02	0.021	-0.031
2019-01-03	-0.031	0.008
2019-01-04	0.008	-0.011
2019-01-05	-0.011	0.064
2019-01-06	0.064	-0.013
2019-01-07	-0.013	?



Example: Binary Classification and Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

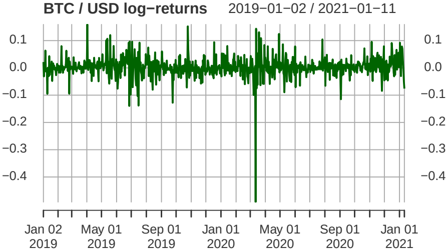
Date	$r(t)$	sign $r(t+1)$
2019-01-02	0.021	-1
2019-01-03	-0.031	1
2019-01-04	0.008	-1
2019-01-05	-0.011	1
2019-01-06	0.064	-1
2019-01-07	-0.013	?



Example: Binary Classification and Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t)$	sign $r(t+1)$
2019-01-02	0.021	-1
2019-01-03	-0.031	1
2019-01-04	0.008	-1
2019-01-05	-0.011	1
2019-01-06	0.064	-1
2019-01-07	-0.013	?



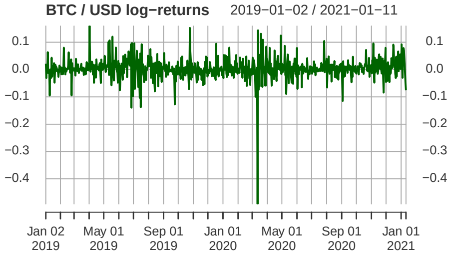
Predict sign of $r(t+1)$ from $r(t)$:

$$f\left(r(t)\right) = y \approx \operatorname{sign}[r(t+1)]$$

Example: Binary Classification and Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t)$	sign $r(t+1)$
2019-01-02	0.021	-1
2019-01-03	-0.031	1
2019-01-04	0.008	-1
2019-01-05	-0.011	1
2019-01-06	0.064	-1
2019-01-07	-0.013	?



Utility $u(x, y)$ is a 2×2 matrix (confusion matrix):

$$\begin{bmatrix} u(x_1, y_1) & u(x_1, y_2) \\ u(x_2, y_1) & u(x_2, y_2) \end{bmatrix}$$

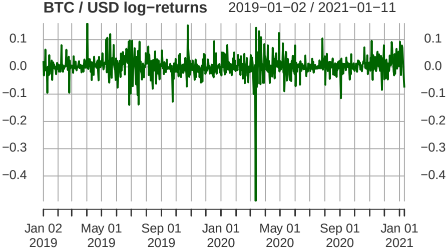
Predict sign of $r(t+1)$ from $r(t)$:

$$f\left(r(t)\right) = y \approx \text{sign}[r(t+1)]$$

Example: Binary Classification and Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t)$	sign $r(t+1)$
2019-01-02	0.021	-1
2019-01-03	-0.031	1
2019-01-04	0.008	-1
2019-01-05	-0.011	1
2019-01-06	0.064	-1
2019-01-07	-0.013	?



Predict sign of $r(t+1)$ from $r(t)$:

$$f\left(r(t)\right) = y \approx \mathrm{sign}[r(t+1)]$$

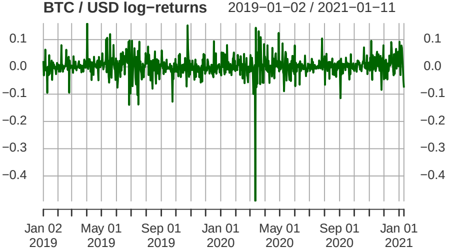
Utility $u(x, y)$ is a 2×2 matrix (confusion matrix):

$$\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix}$$

Example: Binary Classification and Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t)$	sign $r(t+1)$
2019-01-02	0.021	-1
2019-01-03	-0.031	1
2019-01-04	0.008	-1
2019-01-05	-0.011	1
2019-01-06	0.064	-1
2019-01-07	-0.013	?



Predict sign of $r(t+1)$ from $r(t)$:

$$f\left(r(t)\right) = y \approx \mathrm{sign}[r(t+1)]$$

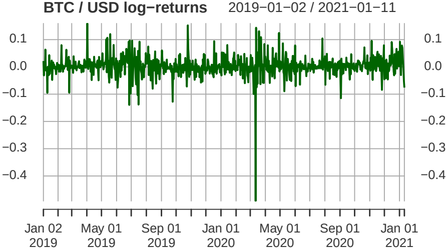
Utility $u(x, y)$ is a 2×2 matrix (confusion matrix):

$$\begin{bmatrix} c_1 + d_1 & c_1 - d_1 \\ c_2 - d_2 & c_2 + d_2 \end{bmatrix}$$

Example: Binary Classification and Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t)$	sign $r(t+1)$
2019-01-02	0.021	-1
2019-01-03	-0.031	1
2019-01-04	0.008	-1
2019-01-05	-0.011	1
2019-01-06	0.064	-1
2019-01-07	-0.013	?



Predict sign of $r(t+1)$ from $r(t)$:

$$f\,(r(t)) = y \approx \operatorname{sign}[r(t+1)]$$

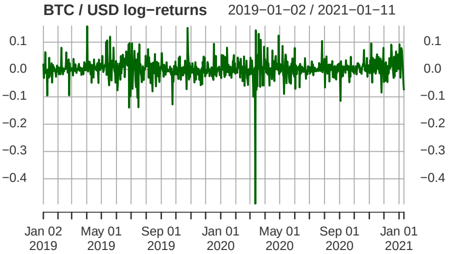
Utility $u(x, y)$ is a 2×2 matrix (confusion matrix):

$$\begin{bmatrix} TP(++) & FN(+-) \\ FP(-+) & TN(--) \end{bmatrix}$$

Example: Binary Classification and Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t)$	sign $r(t+1)$
2019-01-02	0.021	-1
2019-01-03	-0.031	1
2019-01-04	0.008	-1
2019-01-05	-0.011	1
2019-01-06	0.064	-1
2019-01-07	-0.013	?



Predict sign of $r(t+1)$ from $r(t)$:

$$f\,(r(t)) = y \approx \operatorname{sign}[r(t+1)]$$

Utility $u(x, y)$ is a 2×2 matrix (confusion matrix):

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

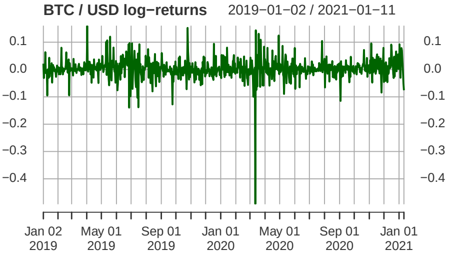
Example: Binary Classification and Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t)$	sign $r(t+1)$
2019-01-02	0.021	-1
2019-01-03	-0.031	1
2019-01-04	0.008	-1
2019-01-05	-0.011	1
2019-01-06	0.064	-1
2019-01-07	-0.013	?

Predict sign of $r(t+1)$ from $r(t)$:

$$f\left(r(t)\right) = y \approx \mathrm{sign}[r(t+1)]$$



Utility $u(x, y)$ is a 2×2 matrix (confusion matrix):

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

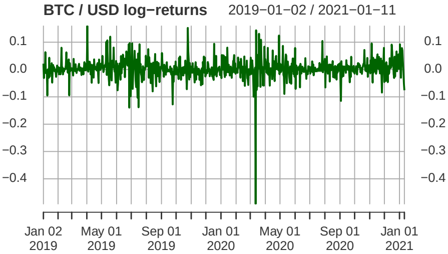
Questions:

Is Accuracy $= .53$ a good result?

Example: Binary Classification and Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t)$	sign $r(t+1)$
2019-01-02	0.021	-1
2019-01-03	-0.031	1
2019-01-04	0.008	-1
2019-01-05	-0.011	1
2019-01-06	0.064	-1
2019-01-07	-0.013	?



Predict sign of $r(t+1)$ from $r(t)$:

$$f(r(t)) = y \approx \text{sign}[r(t+1)]$$

Utility $u(x, y)$ is a 2×2 matrix (confusion matrix):

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Questions:

Is Accuracy $= .53$ a good result?
What is the highest possible accuracy here?

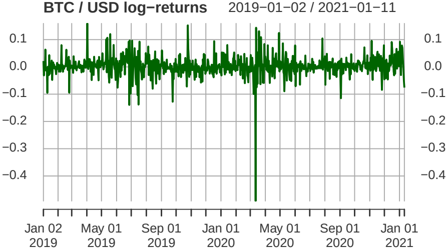
Example: Binary Classification and Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t-1)$	$r(t)$	sign$r(t+1)$
2019-01-06	0.008	-0.011	1
2019-01-07	-0.011	0.064	-1
2019-01-08	0.064	-0.013	-1
2019-01-09	-0.013	-0.0034	-1

Predict sign $r(t+1)$ from n lags:

$$f\left(r(t-n), \dots, r(t)\right) = y \approx \operatorname{sign}[r(t+1)]$$



Utility $u(x, y)$ is a 2×2 matrix (confusion matrix):

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

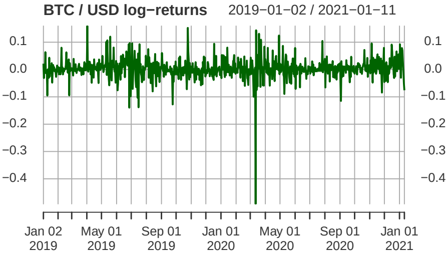
Questions:

Is Accuracy $= .53$ a good result?
What is the highest possible accuracy here?

Example: Binary Classification and Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t-1)$	$r(t)$	sign$r(t+1)$
2019-01-06	0.008	-0.011	1
2019-01-07	-0.011	0.064	-1
2019-01-08	0.064	-0.013	-1
2019-01-09	-0.013	-0.0034	-1



Predict sign $r(t+1)$ from n lags of m symbols (e.g. BTC/USD, ETH/USD, IOT/BTC):

$$f\begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix} = y \approx \mathrm{sign}[r(t+1)]$$

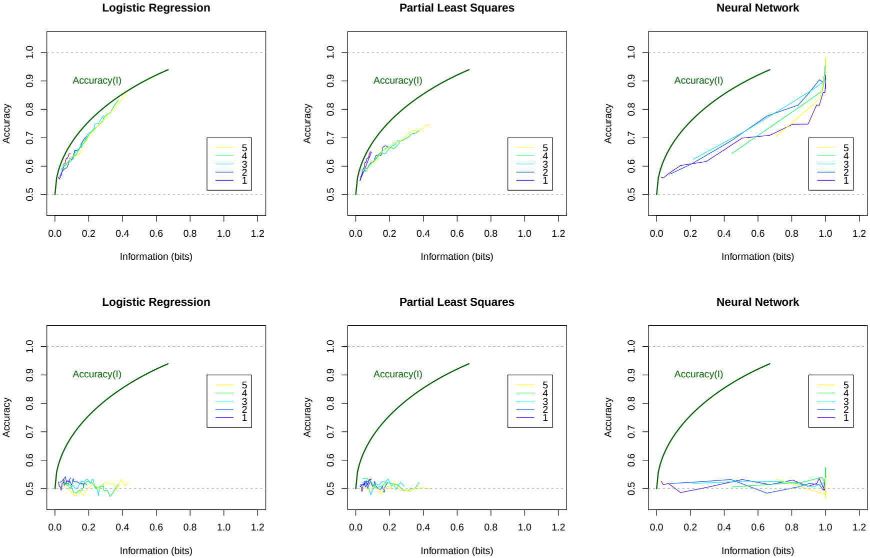
Utility $u(x, y)$ is a 2×2 matrix (confusion matrix):

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

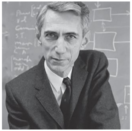
Questions:

Is Accuracy $= .53$ a good result?
What is the highest possible accuracy here?

Evaluation of Accuracy



Information and its Value

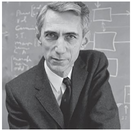


Claude Shannon

$$I_{xy} = \sum_{(x,y)} \left[\ln \frac{P(x \mid y)}{P(x)} \right] P(x,y)$$

(Shannon, 1948)

Information and its Value



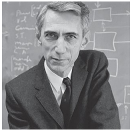
Claude Shannon

$$I_{xy} = \sum_{(x,y)} \left[\ln \frac{P(x \mid y)}{P(x)} \right] P(x,y)$$

(Shannon, 1948)

Ruslan Stratonovich

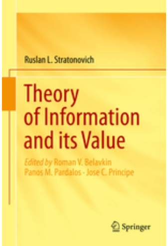
Information and its Value



Claude Shannon

$$I_{xy} = \sum_{(x,y)} \left[\ln \frac{P(x \mid y)}{P(x)} \right] P(x,y)$$

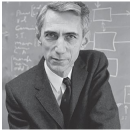
(Shannon, 1948)



(Stratonovich, 1965, 1975, 2020):

Ruslan Stratonovich

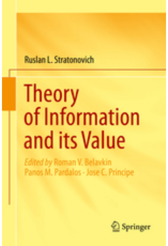
Information and its Value



Claude Shannon

$$I_{xy} = \sum_{(x,y)} \left[\ln \frac{P(x \mid y)}{P(x)} \right] P(x,y)$$

(Shannon, 1948)

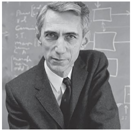


(Stratonovich, 1965, 1975, 2020):

Ruslan Stratonovich

- Belavkin (2013). Optimal measures and Markov transition kernels. *Journal of Global Optimization*, Vol. 55 (387–416).

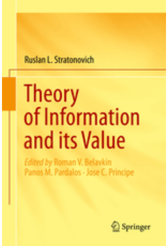
Information and its Value



Claude Shannon

$$I_{xy} = \sum_{(x,y)} \left[\ln \frac{P(x \mid y)}{P(x)} \right] P(x,y)$$

(Shannon, 1948)



(Stratonovich, 1965, 1975, 2020):

Ruslan Stratonovich

- Belavkin (2013). Optimal measures and Markov transition kernels. *Journal of Global Optimization*, Vol. 55 (387–416).
- Belavkin (2018). Relation Between the Kantorovich-Wasserstein Metric and the Kullback-Leibler Divergence. *Information Geometry and Its Applications*, Springer.

Three Types of Information

Definition (Hartley Information)

$$H := \ln |X|$$

Three Types of Information

Definition (Hartley Information)

$$H := \ln |X|$$

Definition (Boltzmann Information)

$$H_P(X) := -\sum_X [\ln P(x)]\, P(x)$$

Three Types of Information

Definition (Hartley Information)

$$H := \ln |X|$$

Definition (Boltzmann Information)

$$H_P(X) := -\sum_X [\ln P(x)]\, P(x) \leq \ln |X|$$

Three Types of Information

Definition (Hartley Information)

$$H := \ln |X|$$

Definition (Boltzmann Information)

$$H_P(X) := -\sum_X [\ln P(x)]\, P(x) \leq \ln |X|$$

Definition (Shannon Information)

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

Three Types of Information

Definition (Hartley Information)

$$H := \ln |X|$$

Definition (Boltzmann Information)

$$H_P(X) := -\sum_X [\ln P(x)]\, P(x) \leq \ln |X|$$

Definition (Shannon Information)

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

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:

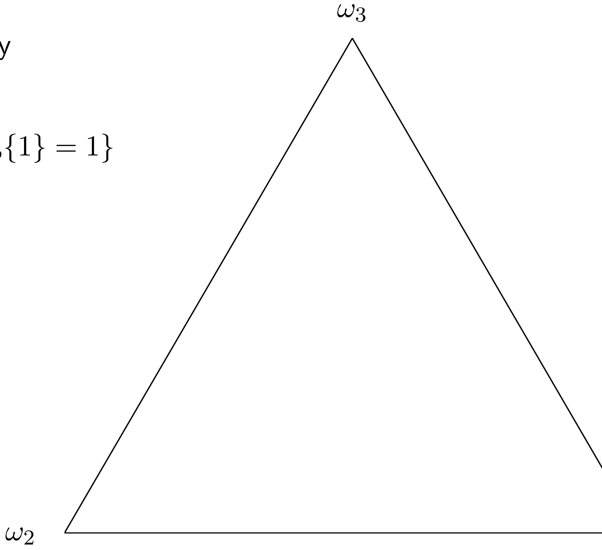
$$0 \le I(X,Y) \le \min[H(X), H(Y)]$$

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

Information-geometruc view

- The set of all probability measures

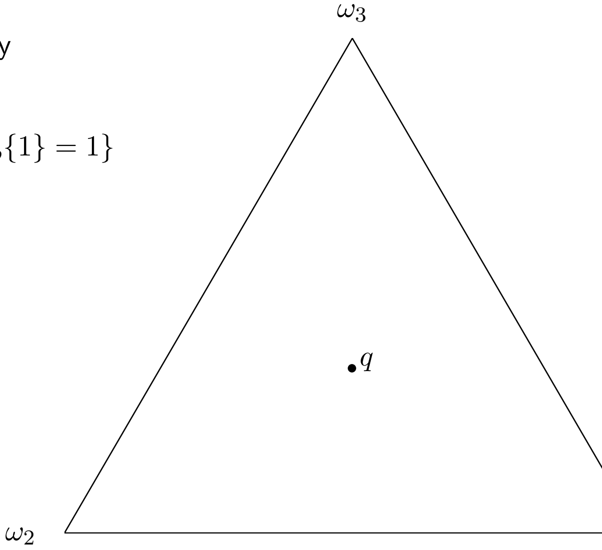
$$\mathcal{P}(\Omega) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$



Information-geometruc view

- The set of all probability measures

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

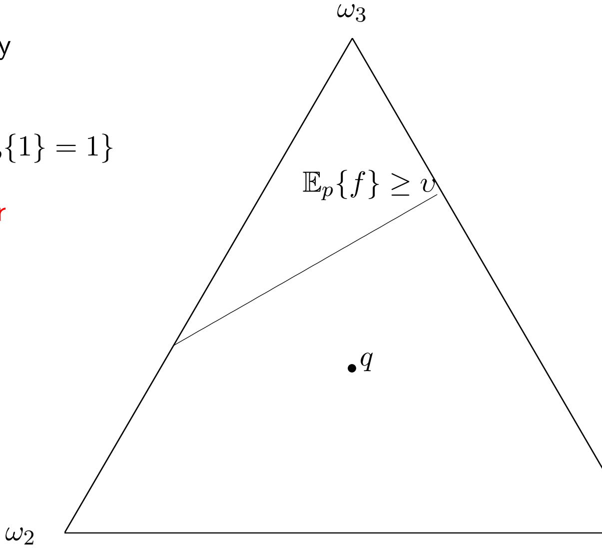


Information-geometruc view

- 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

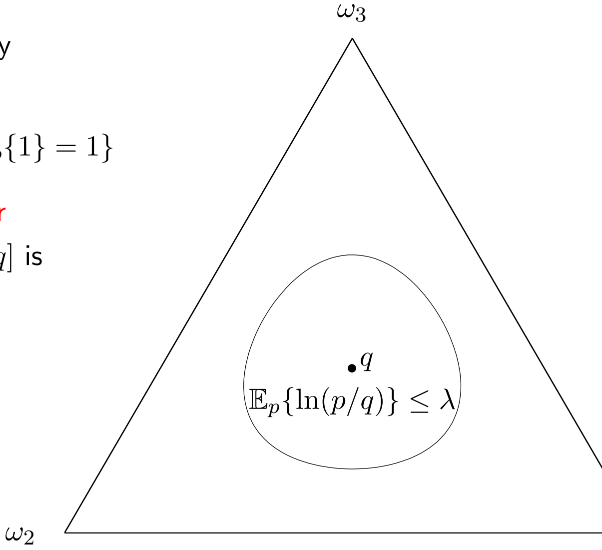


Information-geometruc view

- 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



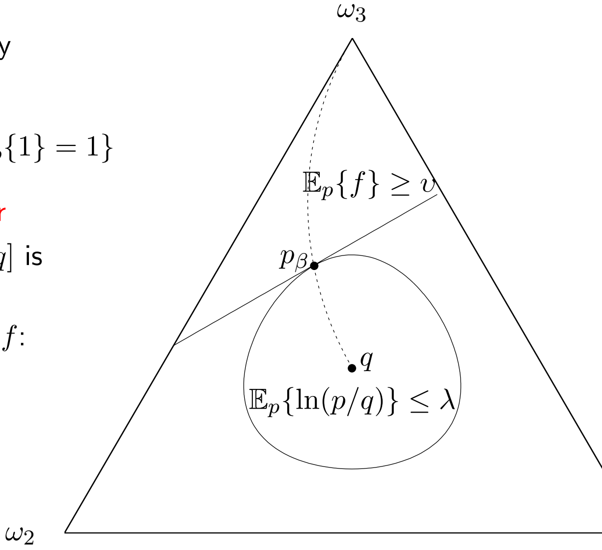
Information-geometruc view

- 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
- $\nabla_p KL[p, q] = \ln \frac{p}{q} = \beta\, f$:

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



Value of Information (Hartley)

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

Value of Information (Hartley)

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

Value of Information (Hartley)

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

$$x = f^{-1}(z) \qquad \mathbb{E}_{P(x)}\{\max_{y(x)} u(x, y)\} =: U(\infty)$$

$$P(x) \qquad \max_{y} \mathbb{E}_{P(x)}\{u(x, y)\} =: U(0)$$

Value of Information (Hartley)

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

$$x = f^{-1}(z) \qquad \mathbb{E}_{P(x)}\{\max_{y(x)} u(x, y)\} =: U(\infty)$$

$$=: U(I)$$

$$P(x) \qquad \max_{y} \mathbb{E}_{P(x)}\{u(x, y)\} =: U(0)$$

Value of Information (Hartley)

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

$$x = f^{-1}(z) \qquad \mathbb{E}_{P(x)}\{\max_{y(x)} u(x, y)\} =: U(\infty)$$

$$=: U(I)$$

$$P(x) \qquad \max_{y} \mathbb{E}_{P(x)}\{u(x, y)\} =: U(0)$$

Definition (Value of Information (Stratonovich, 1965))

$$V(I) := U(I) - U(0)$$

Value of Information (Hartley)

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

$$x = f^{-1}(z) \qquad \mathbb{E}_{P(x)}\{\max_{y(x)} u(x,y)\} =: U(\infty)$$

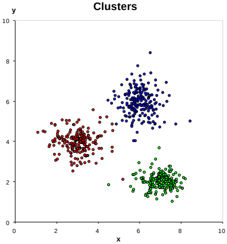
$$x \in f^{-1}(z) \qquad \max_{z(x):\ln|Z|\leq I} \mathbb{E}_{P(z)}\big[\max_{y(z)} \mathbb{E}_{P(x|z)}\{u(x,y) \mid z\}\big] =: U(I)$$

$$P(x) \qquad \max_{y} \mathbb{E}_{P(x)}\{u(x,y)\} =: U(0)$$

Definition (Value of Information (Stratonovich, 1965))

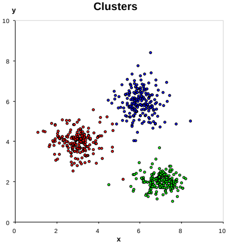
$$V(I) := U(I) - U(0)$$

Example: Mean-Square Minimization



- $P(x)$, $c : X \times X \to \mathbb{R}$

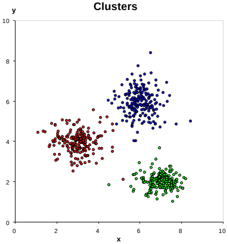
Example: Mean-Square Minimization



- $P(x)$, $c : X \times X \to \mathbb{R}$
- Find $y \in X$ minimizing

$$\mathbb{E}_P\{c(x,y)\} = \sum_x c(x,y)\,P(x)$$

Example: Mean-Square Minimization



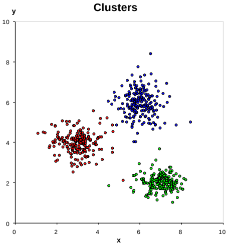
- $P(x)$, $c : X \times X \to \mathbb{R}$
- Find $y \in X$ minimizing

$$\mathbb{E}_P\{c(x,y)\} = \sum_x c(x,y)\, P(x)$$

- Optimal $\hat{y}$ is defined by

$$\sum_x \frac{\partial}{\partial y} c(x,\hat{y})\, P(x) = 0$$

Example: Mean-Square Minimization



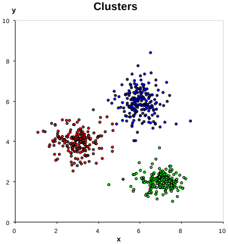
- $P(x)$, $c : X \times X \to \mathbb{R}$
- Find $y \in X$ minimizing

$$\mathbb{E}_P\{c(x,y)\} = \sum_x \frac{1}{2}(x-y)^2\, P(x)$$

- Optimal $\hat{y}$ is defined by

$$\sum_x \frac{\partial}{\partial y}\frac{1}{2}(x-y)^2\, P(x) = 0$$

Example: Mean-Square Minimization



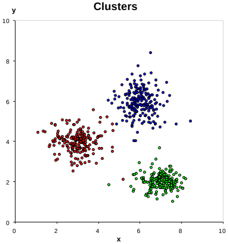
- $P(x)$, $c : X \times X \to \mathbb{R}$
- Find $y \in X$ minimizing

$$\mathbb{E}_P\{c(x,y)\} = \sum_x \frac{1}{2}(x-y)^2\, P(x)$$

- Optimal $\hat{y}$ is defined by

$$\sum_x (x-\hat{y})\, P(x) = 0$$

Example: Mean-Square Minimization



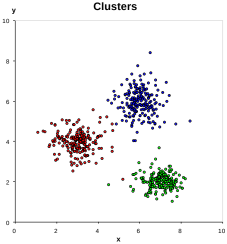
- $P(x)$, $c : X \times X \to \mathbb{R}$
- Find $y \in X$ minimizing

$$\mathbb{E}_P\{c(x,y)\} = \sum_x \frac{1}{2}(x-y)^2 \, P(x)$$

- Optimal $\hat{y}$ is defined by

$$\hat{y} = \sum_x x \, P(x)$$

Example: Mean-Square Minimization



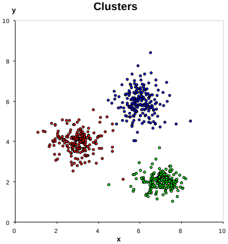
- $P(x)$, $c : X \times X \to \mathbb{R}$
- Find $y \in X$ minimizing

$$\mathbb{E}_P\{c(x,y)\} = \sum_x \frac{1}{2}(x-y)^2\, P(x)$$

- Optimal $\hat{y}$ is defined by

$$\hat{y} = \mathbb{E}\{x\}$$

Example: Mean-Square Minimization



- $P(x)$, $c : X \times X \to \mathbb{R}$
- Find $y \in X$ minimizing

$$\mathbb{E}_P\{c(x,y)\} = \sum_x \frac{1}{2}(x-y)^2\, P(x)$$

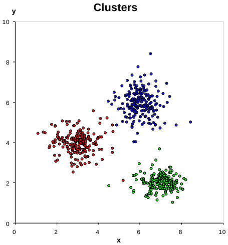
- Optimal $\hat{y}$ is defined by

$$\hat{y} = \mathbb{E}\{x\}$$

k-Means clustering

- Let us partition X into $k = 3$ subsets X_1, X_2, X_3

Example: Mean-Square Minimization



- $P(x)$, $c : X \times X \to \mathbb{R}$
- Find $y \in X$ minimizing

$$\mathbb{E}_P\{c(x,y)\} = \sum_x \frac{1}{2}(x-y)^2\, P(x)$$

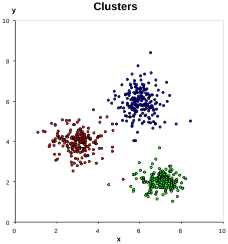
- Optimal $\hat{y}$ is defined by

$$\hat{y} = \mathbb{E}\{x\}$$

k-Means clustering

- Let us partition X into $k = 3$ subsets X_1, X_2, X_3
- This corresponds to some mapping $z(x)$ ($z : X \to \{z_1, z_2, z_3\}$)

Example: Mean-Square Minimization



- $P(x)$, $c : X \times X \to \mathbb{R}$
- Find $y \in X$ minimizing

$$\mathbb{E}_P\{c(x,y)\} = \sum_x \frac{1}{2}(x-y)^2 \, P(x)$$

- Optimal $\hat{y}$ is defined by

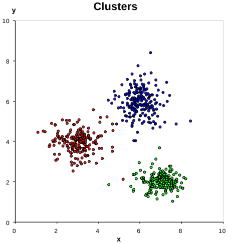
$$\hat{y} = \mathbb{E}\{x\}$$

k-Means clustering

- Let us partition X into $k = 3$ subsets X_1, X_2, X_3
- This corresponds to some mapping $z(x)$ ($z : X \to \{z_1, z_2, z_3\}$)
- Find y_1, y_2, y_3 minimizing

$$\sum_z \mathbb{E}_{P(x|z)}\{c(x,y) \mid z\} \, P(z)$$

Example: Mean-Square Minimization



- $P(x)$, $c : X \times X \to \mathbb{R}$
- Find $y \in X$ minimizing

$$\mathbb{E}_P\{c(x,y)\} = \sum_x \frac{1}{2}(x-y)^2\, P(x)$$

- Optimal $\hat{y}$ is defined by

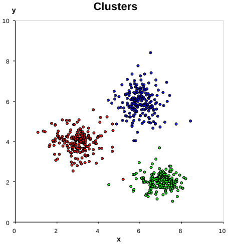
$$\hat{y} = \mathbb{E}\{x\}$$

k-Means clustering

- Let us partition X into $k = 3$ subsets X_1, X_2, X_3
- This corresponds to some mapping $z(x)$ $(z : X \to \{z_1, z_2, z_3\})$
- Find y_1, y_2, y_3 minimizing

$$\sum_z \mathbb{E}_{P(x|z)}\left\{\frac{1}{2}(x-y)^2 \mid z\right\} P(z)\,, \qquad \hat{y}(z) = \sum_x x\, P(x|z)$$

Example: Mean-Square Minimization



- $P(x)$, $c : X \times X \to \mathbb{R}$
- Find $y \in X$ minimizing

$$\mathbb{E}_P\{c(x,y)\} = \sum_x \frac{1}{2}(x-y)^2\, P(x)$$

- Optimal $\hat{y}$ is defined by

$$\hat{y} = \mathbb{E}\{x\}$$

k-Means clustering

- Let us partition X into $k = 3$ subsets X_1, X_2, X_3
- This corresponds to some mapping $z(x)$ $(z : X \to \{z_1, z_2, z_3\})$
- Find y_1, y_2, y_3 minimizing

$$\sum_z \mathbb{E}_{P(x|z)}\left\{\frac{1}{2}(x-y)^2 \mid z\right\} P(z)\,, \qquad \hat{y}(z) = \mathbb{E}\{x \mid z\}$$

Value of Information (Hartley)

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

$$
\begin{aligned}
x = f^{-1}(z) & \qquad \mathbb{E}_{P(x)}\{\max_{y(x)} u(x,y)\} =: U(\infty) \\
x \in f^{-1}(z) & \qquad \max_{z(x):\ln|Z|\leq I} \mathbb{E}_{P(z)}\left[\max_{y(z)} \mathbb{E}_{P(x|z)}\{u(x,y) \mid z\}\right] =: U(I) \\
P(x) & \qquad \max_{y} \mathbb{E}_{P(x)}\{u(x,y)\} =: U(0)
\end{aligned}
$$

Definition (Value of Information (Stratonovich, 1965))

$$V(I) := U(I) - U(0)$$

Value of Information (Shannon)

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

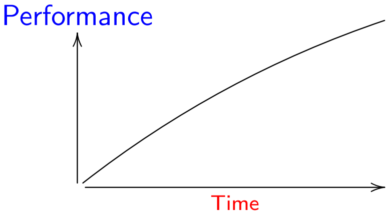
$$
\begin{array}{rcl}
x = f^{-1}(z) & & \mathbb{E}_{P(x)}\{\max_{y(x)} u(x,y)\} =: U(\infty) \\
P(x \mid z) & \max_{P(y|x):I(X,Y)\leq I} & \mathbb{E}_{P(x,y)}\{u(x,y) \mid z\} =: U(I) \\
P(x) & & \max_{y} \mathbb{E}_{P(x)}\{u(x,y)\} =: U(0)
\end{array}
$$

Definition (Value of Information (Stratonovich, 1965))

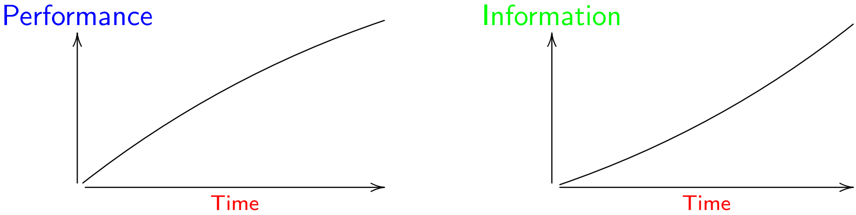
$$V(I) := U(I) - U(0)$$

Learning Systems

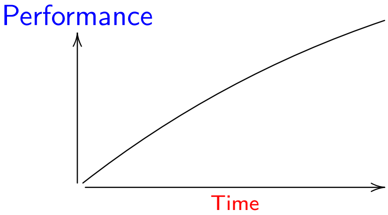
Learning Systems

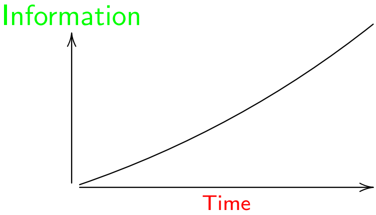


Learning Systems

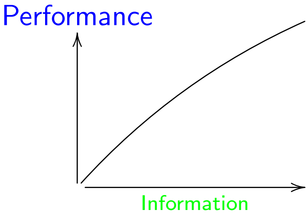


Learning Systems





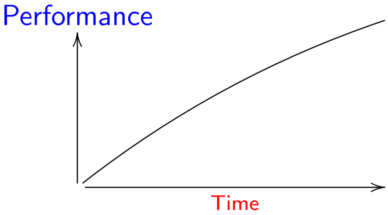
Optimal learning

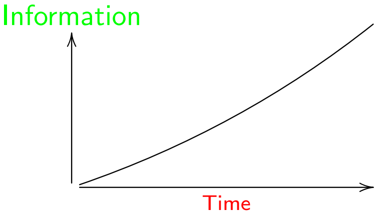


Maximize performance

s.t. information $\leq \lambda$

Learning Systems

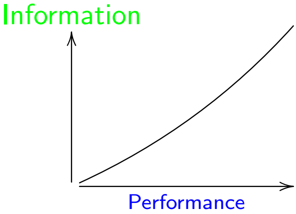




Optimal learning

Minimize information

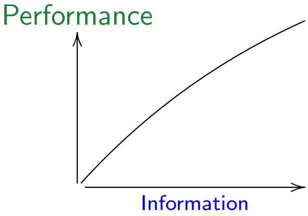
s.t. performance $\geq v$



VoI as Conditional Extremum

- Linear programming problem $U(I)$:

$$\text{maximize} \quad \mathbb{E}_{P(y|x)}\{u(x,y)\} \quad \text{subject to} \quad I(X,Y) \leq I$$



Maximize performance

s.t. information $\leq I$

VoI as Conditional Extremum

- Linear programming problem $U(I)$:

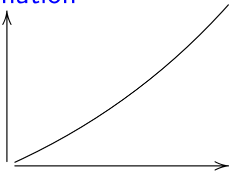
$$\text{maximize} \quad \mathbb{E}_{P(y|x)}\{u(x,y)\} \quad \text{subject to} \quad I(X,Y) \leq I$$

- The inverse convex programming problem $I(U)$:

$$\text{minimize} \quad I(X,Y) \quad \text{subject to} \quad \mathbb{E}_{P(y|x)}\{u(x,y)\} \geq U$$

Minimize information

s.t. performance $\geq V$



Solution

- Lagrange function

$$K(p, \beta) = \mathbb{E}_p\{\ln(p/q)\} + \beta[U - \mathbb{E}_p\{u\}]$$

Solution

- Lagrange function

$$K(p,\beta) = \mathbb{E}_p\{\ln(p/q)\} + \beta[U - \mathbb{E}_p\{u\}]$$

- Necessary and sufficient conditions $\nabla K(p,\beta) = 0$:

$$\begin{aligned}\nabla_p K(p,\beta) &= \ln(p/q) + 1 - \beta U = 0\\ \nabla_\beta K(p,\beta) &= U - \mathbb{E}_p\{u\} = 0\end{aligned}$$

Solution

- Lagrange function

$$K(p, \beta) = \mathbb{E}_p\{\ln(p/q)\} + \beta[U - \mathbb{E}_p\{u\}]$$

- Necessary and sufficient conditions $\nabla K(p, \beta) = 0$:

$$\begin{aligned} \nabla_p K(p, \beta) &= \ln(p/q) + 1 - \beta U = 0 \\ \nabla_\beta K(p, \beta) &= U - \mathbb{E}_p\{u\} = 0 \end{aligned}$$

- Optimal solutions:

$$p(\beta) = e^{\beta\, u - \Psi(\beta)}\, q\,, \qquad \mathbb{E}_{p(\beta)}\{u\} = U \quad \Big(\mathbb{E}_p\{\ln(p/q) = I\}\Big)$$

Solution

- Lagrange function

$$K(p, \beta) = \mathbb{E}_p\{\ln(p/q)\} + \beta[U - \mathbb{E}_p\{u\}]$$

- Necessary and sufficient conditions $\nabla K(p, \beta) = 0$:

$$\begin{aligned} \nabla_p K(p, \beta) &= \ln(p/q) + 1 - \beta U = 0 \\ \nabla_\beta K(p, \beta) &= U - \mathbb{E}_p\{u\} = 0 \end{aligned}$$

- Optimal solutions:

$$p(\beta) = e^{\beta\, u - \Psi(\beta)}\, q\,, \qquad \mathbb{E}_{p(\beta)}\{u\} = U \quad \Big(\mathbb{E}_p\{\ln(p/q) = I\}\Big)$$

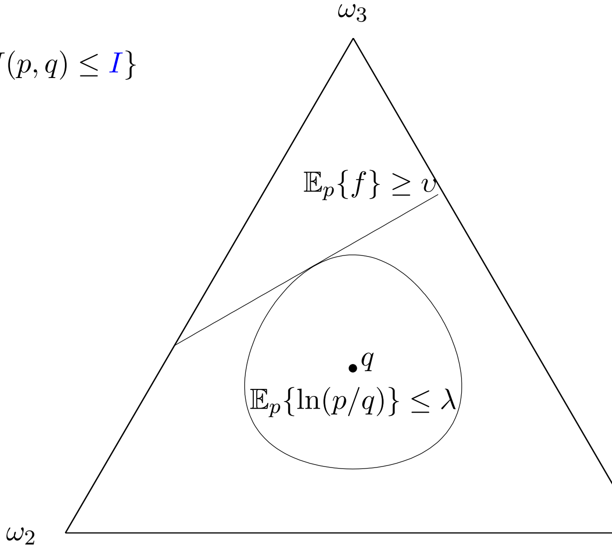
- Optimal *inverse temperature* β:

$$\beta = \frac{dI(U)}{dU} \quad \text{or} \quad \beta^{-1} = \frac{dU(I)}{dI}$$

Optimal Value Functions

- Maximize $\mathbb{E}_p\{u\}$

  $U(I) := \sup\{\mathbb{E}_p\{u\} : I(p, q) \leq I\}$



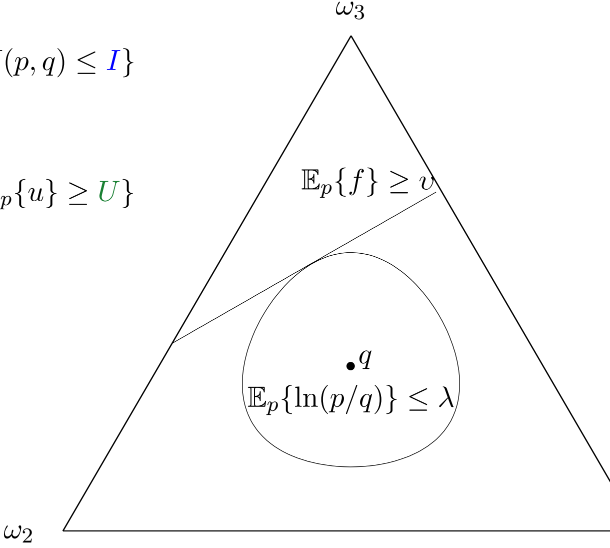
Optimal Value Functions

- Maximize $\mathbb{E}_p\{u\}$

  $U(I) := \sup\{\mathbb{E}_p\{u\} : I(p,q) \leq I\}$

- Minimize $I(p,q)$:

  $I(U) := \inf\{I(p,q) : \mathbb{E}_p\{u\} \geq U\}$



Optimal Value Functions

- Maximize $\mathbb{E}_p\{u\}$

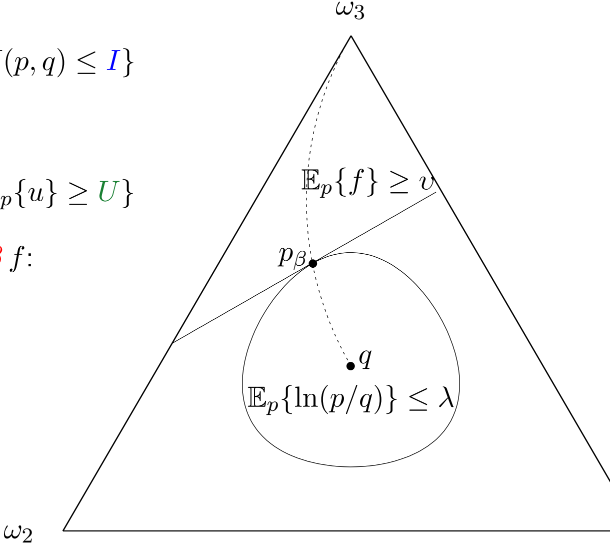
  $$U(I) := \sup\{\mathbb{E}_p\{u\} : I(p,q) \leq I\}$$

- Minimize $I(p,q)$:

  $$I(U) := \inf\{I(p,q) : \mathbb{E}_p\{u\} \geq U\}$$

- $\nabla_p KL[p,q] = \ln \frac{p}{q} = \beta\, f$:

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



Optimal Value Functions

- Maximize $\mathbb{E}_p\{u\}$

  $$U(I) := \sup\{\mathbb{E}_p\{u\} : I(p,q) \leq I\}$$

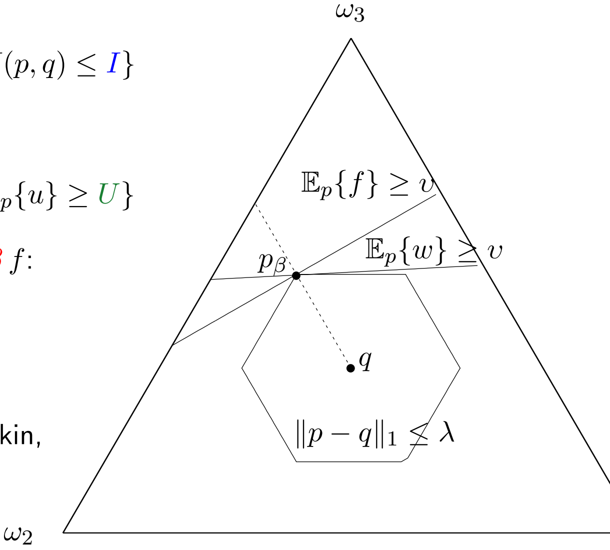
- Minimize $I(p,q)$:

  $$I(U) := \inf\{I(p,q) : \mathbb{E}_p\{u\} \geq U\}$$

- $\nabla_p KL[p,q] = \ln \frac{p}{q} = \beta\, f$:

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

- Generalizations for arbitrary $I(p,q)$ (Belavkin, 2013)



Optimal Value Functions

- Maximize $\mathbb{E}_p\{u\}$

  $U(I) := \sup\{\mathbb{E}_p\{u\} : I(p,q) \leq I\}$

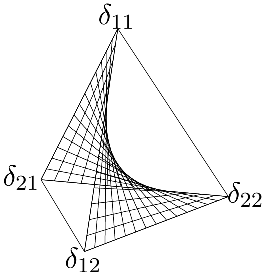
- Minimize $I(p,q)$:

  $I(U) := \inf\{I(p,q) : \mathbb{E}_p\{u\} \geq U\}$

- $\nabla_p KL[p,q] = \ln \frac{p}{q} = \beta\, f$:

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

- Generalizations for arbitrary $I(p,q)$ (Belavkin, 2013)
- $\mathcal{P}(X \otimes Y)$



Solution to the Value of Shannon's Information

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x,y) = P(x)Q(y)e^{\beta u(x,y) - \gamma(\beta,x)}$$

Solution to the Value of Shannon's Information

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x,y) = P(x)Q(y)e^{\beta u(x,y) - \gamma(\beta,x)}$$

- The law of total probability gives two equations:

$$\sum_x e^{\beta\, u(x,y) - \gamma(\beta,x)}\, P(x) = 1\,, \qquad \sum_y e^{\beta\, u(x,y)}\, Q(y) = e^{\gamma(\beta,x)}$$

Solution to the Value of Shannon's Information

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x,y) = P(x)Q(y)e^{\beta u(x,y)-\gamma(\beta,x)}$$

- The law of total probability gives two equations:

$$\sum_x e^{\beta\, u(x,y)-\gamma(\beta,x)}\, P(x) = 1\,, \qquad \sum_y e^{\beta\, u(x,y)}\, Q(y) = e^{\gamma(\beta,x)}$$

- Use the *cumulant generating function*

$$\Gamma(\beta) = \sum_x \gamma(x,\beta)\, P(x)$$

Solution to the Value of Shannon's Information

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x,y) = P(x)Q(y)e^{\beta u(x,y) - \gamma(\beta,x)}$$

- The law of total probability gives two equations:

$$\sum_x e^{\beta\, u(x,y) - \gamma(\beta,x)}\, P(x) = 1\,, \qquad \sum_y e^{\beta\, u(x,y)}\, Q(y) = e^{\gamma(\beta,x)}$$

- Use the *cumulant generating function*

$$\Gamma(\beta) = \sum_x \gamma(x,\beta)\, P(x)$$

- Find $U(I)$ from

$$U(\beta) = \frac{d\Gamma(\beta)}{d\beta}\,, \qquad I(\beta) = \beta\Gamma'(\beta) - \Gamma(\beta)$$

Solution to the Value of Shannon's Information

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x,y) = P(x)Q(y)e^{\beta u(x,y) - \gamma(\beta, x)}$$

- The law of total probability gives two equations:

$$\sum_x e^{\beta\, u(x,y) - \gamma(\beta,x)}\, P(x) = 1\,, \qquad \sum_y e^{\beta\, u(x,y)}\, Q(y) = e^{\gamma(\beta,x)}$$

- Use the *cumulant generating function*

$$\Gamma(\beta) = \sum_x \gamma(x,\beta)\, P(x)$$

- Find $U(I)$ from

$$U(\beta) = \frac{d\Gamma(\beta)}{d\beta}\,, \qquad I(\beta) = \beta\Gamma'(\beta) - \Gamma(\beta)$$

- $\beta^{-1} = U'(I)$ is called *temperature*.

Solution to the Value of Shannon's Information

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x,y) = P(x)Q(y)e^{\beta u(x,y)-\gamma(\beta,x)}$$

- The law of total probability gives two equations:

$$\sum_x e^{\beta\, u(x,y)-\gamma(\beta,x)}\, P(x) = 1\,, \qquad \sum_y e^{\beta\, u(x,y)}\, Q(y) = e^{\gamma(\beta,x)}$$

- Use the *cumulant generating function*

$$\Gamma(\beta) = \sum_x \gamma(x,\beta)\, P(x)$$

- Find $U(I)$ from

$$U(\beta) = \frac{d\Gamma(\beta)}{d\beta}\,, \qquad I(\beta) = \beta\Gamma'(\beta) - \Gamma(\beta)$$

- $\beta^{-1} = U'(I)$ is called *temperature*.
- Note that $Q(y) = \sum_x P(y \mid x)\, P(x)$

Solution to the Value of Shannon's Information

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x \mid y) = P(x)e^{\beta u(x,y) - \gamma(\beta,x)}$$

- The law of total probability gives two equations:

$$\sum_x e^{\beta\, u(x,y) - \gamma(\beta,x)}\, P(x) = 1\,, \qquad \sum_y e^{\beta\, u(x,y)}\, Q(y) = e^{\gamma(\beta,x)}$$

- Use the *cumulant generating function*

$$\Gamma(\beta) = \sum_x \gamma(x,\beta)\, P(x)$$

- Find $U(I)$ from

$$U(\beta) = \frac{d\Gamma(\beta)}{d\beta}\,, \qquad I(\beta) = \beta\Gamma'(\beta) - \Gamma(\beta)$$

- $\beta^{-1} = U'(I)$ is called *temperature*.
- Note that $Q(y) = \sum_x P(y \mid x)\, P(x)$

Computation of VoI

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x,y) = P(x)Q(y)e^{\beta u(x,y) - \gamma(x,\beta)}$$

Computation of VoI

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x,y) = P(x)Q(y)e^{\beta u(x,y) - \gamma(x,\beta)}$$

- If $T(\cdot) = \sum_x e^{\beta\, u(x,y)}(\cdot)$ is invertible, then

$$T(e^{-\gamma(\beta,x)}\, P(x)) = 1 \iff e^{-\gamma(\beta,x)}\, P(x) = T^{-1}(1) =: e^{-\gamma_0(\beta,x)}$$

Computation of VoI

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x,y) = P(x)Q(y)e^{\beta u(x,y)-\gamma(x,\beta)}$$

- If $T(\cdot) = \sum_x e^{\beta\, u(x,y)}(\cdot)$ is invertible, then

$$T(e^{-\gamma(\beta,x)}\, P(x)) = 1 \iff e^{-\gamma(\beta,x)}\, P(x) = T^{-1}(1) =: e^{-\gamma_0(\beta,x)}$$

- The conditional *cumulant generating function* is

$$\Gamma_0(\beta) = \sum_x \gamma_0(\beta,x)\, P(x) = \Gamma(\beta) + H(X)$$

Computation of VoI

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x,y) = P(x)Q(y)e^{\beta u(x,y)-\gamma(x,\beta)}$$

- If $T(\cdot) = \sum_x e^{\beta\, u(x,y)}(\cdot)$ is invertible, then

$$T(e^{-\gamma(\beta,x)}\, P(x)) = 1 \iff e^{-\gamma(\beta,x)}\, P(x) = T^{-1}(1) =: e^{-\gamma_0(\beta,x)}$$

- The conditional *cumulant generating function* is

$$\Gamma_0(\beta) = \sum_x \gamma_0(\beta,x)\, P(x) = \Gamma(\beta) + H(X)$$

- Find $U(I)$ from

$$U(\beta) = \frac{d\Gamma_0(\beta)}{d\beta}\,, \qquad I(\beta) = H(X) - \underbrace{[\Gamma_0(\beta) - \beta\Gamma_0'(\beta)]}_{H(X|Y)}$$

Computation of VoI

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x,y) = P(x)Q(y)e^{\beta u(x,y) - \gamma(x,\beta)}$$

- If $T(\cdot) = \sum_x e^{\beta\, u(x,y)}(\cdot)$ is invertible, then

$$T(e^{-\gamma(\beta,x)}\, P(x)) = 1 \iff e^{-\gamma(\beta,x)}\, P(x) = T^{-1}(1) =: e^{-\gamma_0(\beta,x)}$$

- The conditional *cumulant generating function* is

$$\Gamma_0(\beta) = \sum_x \gamma_0(\beta,x)\, P(x) = \Gamma(\beta) + H(X)$$

- Find $U(I)$ from

$$U(\beta) = \frac{d\Gamma_0(\beta)}{d\beta}\,, \qquad I(\beta) = H(X) - \underbrace{[\Gamma_0(\beta) - \beta\Gamma_0'(\beta)]}$$

Theorem

$$e^{-\gamma_0(\beta,x)} = e^{-\Gamma_0(\beta)}$$

Computation of VoI

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x,y) = P(x)Q(y)e^{\beta u(x,y)-\gamma(x,\beta)}$$

- If $T(\cdot) = \sum_x e^{\beta\, u(x,y)}(\cdot)$ is invertible, then

$$T(e^{-\gamma(\beta,x)}\, P(x)) = 1 \iff e^{-\gamma(\beta,x)}\, P(x) = T^{-1}(1) =: e^{-\gamma_0(\beta,x)}$$

- The conditional *cumulant generating function* is

$$\Gamma_0(\beta) = \sum_x \gamma_0(\beta,x)\, P(x) = \Gamma(\beta) + H(X)$$

- Find $U(I)$ from

$$U(\beta) = \frac{d\Gamma_0(\beta)}{d\beta}\,, \qquad I(\beta) = H(X) - \underbrace{[\Gamma_0(\beta) - \beta\Gamma_0'(\beta)]}$$

Theorem

$$e^{-\gamma_0(\beta,x)} = e^{-\Gamma_0(\beta)} \iff T(1) = \sum_x e^{\beta\, u(x,y)} = e^{\Gamma_0(\beta)}$$

Computation of VoI

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x \mid y) = P(x)e^{\beta u(x,y)-\gamma(x,\beta)}$$

- If $T(\cdot) = \sum_x e^{\beta\, u(x,y)}(\cdot)$ is invertible, then

$$T(e^{-\gamma(\beta,x)}\, P(x)) = 1 \iff e^{-\gamma(\beta,x)}\, P(x) = T^{-1}(1) =: e^{-\gamma_0(\beta,x)}$$

- The conditional *cumulant generating function* is

$$\Gamma_0(\beta) = \sum_x \gamma_0(\beta,x)\, P(x) = \Gamma(\beta) + H(X)$$

- Find $U(I)$ from

$$U(\beta) = \frac{d\Gamma_0(\beta)}{d\beta}\,, \qquad I(\beta) = H(X) - \underbrace{[\Gamma_0(\beta) - \beta\Gamma_0'(\beta)]}$$

Theorem

$$e^{-\gamma_0(\beta,x)} = e^{-\Gamma_0(\beta)} \iff T(1) = \sum_x e^{\beta\, u(x,y)} = e^{\Gamma_0(\beta)}$$

Computation of VoI

- Solutions to $V(I)$ are optimal joint probabilities of the form:

$$P(x \mid y) = e^{\beta u(x,y) - \Gamma_0(\beta)}$$

- If $T(\cdot) = \sum_x e^{\beta\, u(x,y)}(\cdot)$ is invertible, then

$$T(e^{-\gamma(\beta,x)}\, P(x)) = 1 \iff e^{-\gamma(\beta,x)}\, P(x) = T^{-1}(1) =: e^{-\gamma_0(\beta,x)}$$

- The conditional *cumulant generating function* is

$$\Gamma_0(\beta) = \sum_x \gamma_0(\beta,x)\, P(x) = \Gamma(\beta) + H(X)$$

- Find $U(I)$ from

$$U(\beta) = \frac{d\Gamma_0(\beta)}{d\beta}\,, \qquad I(\beta) = H(X) - \underbrace{[\Gamma_0(\beta) - \beta\Gamma_0'(\beta)]}$$

Theorem

$$e^{-\gamma_0(\beta,x)} = e^{-\Gamma_0(\beta)} \iff T(1) = \sum_x e^{\beta\, u(x,y)} = e^{\Gamma_0(\beta)}$$

Free Energy as the Dual of VoI

Entropy of Gibbs state

$$P(z \mid x) = \frac{e^{\beta\, u(x,z)}}{Z(\beta)}$$

β is *inverse temperature* (i.e. β^{-1} is temperature, $Z(\beta)$ is the *partition function* ($\Gamma(\beta) = \ln Z(\beta)$ — the cumulant gen. f-n).

$$H_P(X) := -\sum_X [\ln P(x)]\, P(x)$$

Free Energy as the Dual of VoI

Entropy of Gibbs state

$$P(z \mid x) = \frac{e^{\beta\, u(x,z)}}{Z(\beta)}$$

β is *inverse temperature* (i.e. β^{-1} is temperature, $Z(\beta)$ is the *partition function* ($\Gamma(\beta) = \ln Z(\beta)$ — the cumulant gen. f-n).

$$H_P(X) := -\sum_X [\ln P(x)]\, P(x) = \Gamma(\beta) - \beta\Gamma'(\beta)$$

Free energy and entropy

$$F(\beta^{-1}) := -\beta^{-1}\Gamma(\beta)$$

Free Energy as the Dual of VoI

Entropy of Gibbs state

$$P(z \mid x) = \frac{e^{\beta\, u(x,z)}}{Z(\beta)}$$

β is *inverse temperature* (i.e. β^{-1} is temperature, $Z(\beta)$ is the *partition function* ($\Gamma(\beta) = \ln Z(\beta)$ — the cumulant gen. f-n).

$$H_P(X) := -\sum_X [\ln P(x)]\, P(x) = \Gamma(\beta) - \beta\Gamma'(\beta)$$

Free energy and entropy

$$F(\beta^{-1}) := -\beta^{-1}\Gamma(\beta) \qquad F^*(\lambda) = \inf[\beta^{-1}\lambda - F(\beta^{-1})]$$

Free Energy as the Dual of VoI

Entropy of Gibbs state

$$P(z \mid x) = \frac{e^{\beta\, u(x,z)}}{Z(\beta)}$$

β is *inverse temperature* (i.e. β^{-1} is temperature, $Z(\beta)$ is the *partition function* ($\Gamma(\beta) = \ln Z(\beta)$ — the cumulant gen. f-n).

$$H_P(X) := -\sum_X [\ln P(x)]\, P(x) = \Gamma(\beta) - \beta\Gamma'(\beta)$$

Free energy and entropy

$$F(\beta^{-1}) := -\beta^{-1}\Gamma(\beta) \qquad F^*(\lambda) = \inf[\beta^{-1}\lambda - F(\beta^{-1})]$$

$$F'(\beta^{-1}) = \lambda$$

Free Energy as the Dual of VoI

Entropy of Gibbs state

$$P(z \mid x) = \frac{e^{\beta\, u(x,z)}}{Z(\beta)}$$

β is *inverse temperature* (i.e. β^{-1} is temperature, $Z(\beta)$ is the *partition function* ($\Gamma(\beta) = \ln Z(\beta)$ — the cumulant gen. f-n).

$$H_P(X) := -\sum_X [\ln P(x)]\, P(x) = \Gamma(\beta) - \beta\Gamma'(\beta)$$

Free energy and entropy

$$F(\beta^{-1}) := -\beta^{-1}\Gamma(\beta) \qquad F^*(\lambda) = \inf[\beta^{-1}\lambda - F(\beta^{-1})]$$

$$F'(\beta^{-1}) = \lambda \qquad F^{*\prime}(\lambda) = \beta^{-1}$$

Free Energy as the Dual of VoI

Entropy of Gibbs state

$$P(z \mid x) = \frac{e^{\beta\, u(x,z)}}{Z(\beta)}$$

β is *inverse temperature* (i.e. β^{-1} is temperature, $Z(\beta)$ is the *partition function* ($\Gamma(\beta) = \ln Z(\beta)$ — the cumulant gen. f-n).

$$H_P(X) := -\sum_X [\ln P(x)]\, P(x) = \Gamma(\beta) - \beta\Gamma'(\beta)$$

Free energy and entropy

$$F(\beta^{-1}) := -\beta^{-1}\Gamma(\beta) \qquad F^*(\lambda) = \inf[\beta^{-1}\lambda - F(\beta^{-1})]$$

$$F'(\beta^{-1}) = \lambda = \beta\Gamma'(\beta) - \Gamma(\beta) \qquad F^{*\prime}(\lambda) = \beta^{-1}$$

Free Energy as the Dual of VoI

Entropy of Gibbs state

$$P(z \mid x) = \frac{e^{\beta\, u(x,z)}}{Z(\beta)}$$

β is *inverse temperature* (i.e. β^{-1} is temperature, $Z(\beta)$ is the *partition function* ($\Gamma(\beta) = \ln Z(\beta)$ — the cumulant gen. f-n).

$$H_P(X) := -\sum_X [\ln P(x)]\, P(x) = \Gamma(\beta) - \beta\Gamma'(\beta)$$

Free energy and entropy

$$F(\beta^{-1}) := -\beta^{-1}\Gamma(\beta) \qquad F^*(\lambda) = \inf[\beta^{-1}\lambda - F(\beta^{-1})]$$

$$F'(\beta^{-1}) = \lambda = \underbrace{\beta\Gamma'(\beta) - \Gamma(\beta)}_{-H} \qquad F^{*\prime}(\lambda) = \beta^{-1}$$

Free Energy as the Dual of VoI

Entropy of Gibbs state

$$P(z \mid x) = \frac{e^{\beta\, u(x,z)}}{Z(\beta)}$$

β is *inverse temperature* (i.e. β^{-1} is temperature, $Z(\beta)$ is the *partition function* ($\Gamma(\beta) = \ln Z(\beta)$ — the cumulant gen. f-n).

$$H_P(X) := -\sum_X [\ln P(x)]\, P(x) = \Gamma(\beta) - \beta\Gamma'(\beta)$$

Free energy and entropy

$$F(\beta^{-1}) := -\beta^{-1}\Gamma(\beta) \qquad F^*(\lambda) = \underbrace{\inf[\beta^{-1}\lambda - F(\beta^{-1})]}_{V(\lambda)=\max[\mathbb{E}_P\{u\}:D[p,q]\leq\lambda]}$$

$$F'(\beta^{-1}) = \lambda = \underbrace{\beta\Gamma'(\beta) - \Gamma(\beta)}_{-H} \qquad F^{*\prime}(\lambda) = \beta^{-1}$$

Example: The Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} u(x_1, y_1) & u(x_1, y_2) \\ u(x_2, y_1) & u(x_2, y_2) \end{bmatrix}$$

Example: The Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix}$$

Example: The Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix}$$

- For $P(x) \in \{p, 1-p\}$ the equation $\|e^{\beta\, u(x,y)}\|^T P(x)\, e^{-\gamma(\beta,x)} = 1$ is

$$\begin{bmatrix} e^{\beta\, u_{11}} & e^{\beta\, u_{21}} \\ e^{\beta\, u_{12}} & e^{\beta\, u_{22}} \end{bmatrix} \begin{bmatrix} p\, e^{-\gamma(\beta,x_1)} \\ (1-p)\, e^{-\gamma(\beta,x_2)} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Example: The Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix}$$

- For $P(x) \in \{p, 1-p\}$ the equation $\|e^{\beta\, u(x,y)}\|^T P(x)\, e^{-\gamma(\beta,x)} = 1$ is

$$\begin{bmatrix} p\, e^{-\gamma(\beta,x_1)} \\ (1-p)\, e^{-\gamma(\beta,x_2)} \end{bmatrix} = \frac{1}{\det \|e^{\beta\, u}\|^T} \begin{bmatrix} e^{\beta\, u_{22}} & -e^{\beta\, u_{21}} \\ -e^{\beta\, u_{12}} & e^{\beta\, u_{11}} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Example: The Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix}$$

- For $P(x) \in \{p, 1-p\}$ the equation $\|e^{\beta\, u(x,y)}\|^T P(x)\, e^{-\gamma(\beta,x)} = 1$ is

$$\begin{bmatrix} e^{-\gamma_0(\beta,x_1)} \\ e^{-\gamma_0(\beta,x_2)} \end{bmatrix} = \begin{bmatrix} \frac{e^{\beta\, u_{22}} - e^{\beta\, u_{21}}}{e^{\beta\,(u_{11}+u_{22})} - e^{\beta\,(u_{12}+u_{21})}} \\ \frac{e^{\beta\, u_{11}} - e^{\beta\, u_{12}}}{e^{\beta\,(u_{11}+u_{22})} - e^{\beta\,(u_{12}+u_{21})}} \end{bmatrix}$$

Example: The Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} c_1 + d_1 & c_1 - d_1 \\ c_2 - d_2 & c_2 + d_2 \end{bmatrix}$$

- For $P(x) \in \{p, 1-p\}$ the equation $\|e^{\beta\, u(x,y)}\|^T P(x)\, e^{-\gamma(\beta,x)} = 1$ is

$$\begin{bmatrix} e^{-\gamma_0(\beta,x_1)} \\ e^{-\gamma_0(\beta,x_2)} \end{bmatrix} = \begin{bmatrix} e^{-\beta\, c_1}\, \frac{\sinh(\beta\, d_2)}{\sinh[\beta\,(d_1+d_2)]} \\ e^{-\beta\, c_2}\, \frac{\sinh(\beta\, d_1)}{\sinh[\beta\,(d_1+d_2)]} \end{bmatrix}$$

Example: The Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} c+d & c-d \\ c-d & c+d \end{bmatrix}$$

- For $P(x) \in \{p, 1-p\}$ the equation $\|e^{\beta\, u(x,y)}\|^T P(x)\, e^{-\gamma(\beta,x)} = 1$ is

$$\begin{bmatrix} e^{-\gamma_0(\beta,x_1)} \\ e^{-\gamma_0(\beta,x_2)} \end{bmatrix} = \begin{bmatrix} e^{-\beta\, c_1}\, \frac{\sinh(\beta\, d_2)}{\sinh[\beta\,(d_1+d_2)]} \\ e^{-\beta\, c_2}\, \frac{\sinh(\beta\, d_1)}{\sinh[\beta\,(d_1+d_2)]} \end{bmatrix}$$

- This gives

$$\begin{aligned} \Gamma_0(\beta) &= \beta\, c + \ln\left[2\cosh(\beta\, d)\right] \\ U(\beta) &= c + d\, \tanh(\beta\, d) \\ I(\beta) &= H(X) - [\ln[2\cosh(\beta\, d)] - \beta\, d\, \tanh(\beta\, d)] \end{aligned}$$

Example: The Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} c+d & c-d \\ c-d & c+d \end{bmatrix}$$

- For $P(x) \in \{p, 1-p\}$ the equation $\|e^{\beta\, u(x,y)}\|^T P(x)\, e^{-\gamma(\beta,x)} = 1$ is

$$\begin{bmatrix} e^{-\gamma_0(\beta,x_1)} \\ e^{-\gamma_0(\beta,x_2)} \end{bmatrix} = \begin{bmatrix} e^{-\beta\, c_1}\, \frac{\sinh(\beta\, d_2)}{\sinh[\beta\,(d_1+d_2)]} \\ e^{-\beta\, c_2}\, \frac{\sinh(\beta\, d_1)}{\sinh[\beta\,(d_1+d_2)]} \end{bmatrix}$$

- This gives

$$\Gamma_0(\beta) = \beta\, c + \ln\left[2\cosh(\beta\, d)\right]$$

$$U(\beta) = c + d\, \tanh(\beta\, d)$$

$$I(\beta) = H(X) - \left[\ln[2\cosh(\beta\, d)] - \beta\, d\, \tanh(\beta\, d)\right]$$

- Explicit dependency

$$I(U) = H_2[p] - H_2\left[\frac{1}{2} + \frac{1}{2}\,\frac{U-c}{d}\right]$$

Computation of $Q(y)$ in the Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix}$$

Computation of $Q(y)$ in the Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix}$$

- For $Q(y) \in \{q, 1-q\}$ the equation $\|e^{\beta\, u(x,y)}\| Q(y) = e^{\gamma(\beta, x)}$ is

$$\begin{bmatrix} e^{\beta\, u_{11}} & e^{\beta\, u_{12}} \\ e^{\beta\, u_{21}} & e^{\beta\, u_{22}} \end{bmatrix} \begin{bmatrix} q \\ 1-q \end{bmatrix} = \begin{bmatrix} e^{\gamma(\beta, x_1)} \\ e^{\gamma(\beta, x_2)} \end{bmatrix}$$

Computation of $Q(y)$ in the Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix}$$

- For $Q(y) \in \{q, 1-q\}$ the equation $\|e^{\beta\, u(x,y)}\| Q(y) = e^{\gamma(\beta,x)}$ is

$$\begin{bmatrix} q \\ 1-q \end{bmatrix} = \frac{1}{\det \|e^{\beta\, u}\|} \begin{bmatrix} e^{\beta\, u_{22}} & -e^{\beta\, u_{12}} \\ -e^{\beta\, u_{21}} & e^{\beta\, u_{11}} \end{bmatrix} \begin{bmatrix} e^{\gamma(\beta,x_1)} \\ e^{\gamma(\beta,x_2)} \end{bmatrix}$$

Computation of $Q(y)$ in the Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} c_1 + d_1 & c_1 - d_1 \\ c_2 - d_2 & c_2 + d_2 \end{bmatrix}$$

- For $Q(y) \in \{q, 1-q\}$ the equation $\|e^{\beta\, u(x,y)}\| Q(y) = e^{\gamma(\beta, x)}$ is

$$\begin{bmatrix} q \\ 1-q \end{bmatrix} = \begin{bmatrix} \frac{p}{1-e^{-2\beta\, d_2}} + \frac{1-p}{1-e^{2\beta\, d_1}} \\ \frac{1-p}{1-e^{-2\beta\, d_1}} + \frac{p}{1-e^{2\beta\, d_2}} \end{bmatrix}$$

Computation of $Q(y)$ in the Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} c_1 + d_1 & c_1 - d_1 \\ c_2 - d_2 & c_2 + d_2 \end{bmatrix}$$

- For $Q(y) \in \{q, 1-q\}$ the equation $\|e^{\beta\, u(x,y)}\| Q(y) = e^{\gamma(\beta, x)}$ is

$$\begin{bmatrix} q \\ 1-q \end{bmatrix} = \begin{bmatrix} \frac{p}{1-e^{-2\beta\, d_2}} + \frac{1-p}{1-e^{2\beta\, d_1}} \\ \frac{1-p}{1-e^{-2\beta\, d_1}} + \frac{p}{1-e^{2\beta\, d_2}} \end{bmatrix}$$

- Note that $Q(y) \to P(x)$ as $\beta \to \infty$.

Computation of $Q(y)$ in the Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} c_1 + d_1 & c_1 - d_1 \\ c_2 - d_2 & c_2 + d_2 \end{bmatrix}$$

- For $Q(y) \in \{q, 1-q\}$ the equation $\|e^{\beta\, u(x,y)}\| Q(y) = e^{\gamma(\beta,x)}$ is

$$\begin{bmatrix} q \\ 1-q \end{bmatrix} = \begin{bmatrix} \frac{p}{1-e^{-2\beta\, d_2}} + \frac{1-p}{1-e^{2\beta\, d_1}} \\ \frac{1-p}{1-e^{-2\beta\, d_1}} + \frac{p}{1-e^{2\beta\, d_2}} \end{bmatrix}$$

- Note that $Q(y) \to P(x)$ as $\beta \to \infty$.
- One can check that $Q(y_1) + Q(y_2) = 1$.

Computation of $Q(y)$ in the Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} c_1 + d_1 & c_1 - d_1 \\ c_2 - d_2 & c_2 + d_2 \end{bmatrix}$$

- For $Q(y) \in \{q, 1-q\}$ the equation $\|e^{\beta\, u(x,y)}\|Q(y) = e^{\gamma(\beta,x)}$ is

$$\begin{bmatrix} q \\ 1-q \end{bmatrix} = \begin{bmatrix} \frac{p}{1-e^{-2\beta\, d_2}} + \frac{1-p}{1-e^{2\beta\, d_1}} \\ \frac{1-p}{1-e^{-2\beta\, d_1}} + \frac{p}{1-e^{2\beta\, d_2}} \end{bmatrix}$$

- Note that $Q(y) \to P(x)$ as $\beta \to \infty$.
- One can check that $Q(y_1) + Q(y_2) = 1$.
- $\exists\, \beta_0 \geq 0$ such that $Q(y_1) < 0$ or $Q(y_2) < 0$ for $\beta \in [0, \beta_0)$:

Computation of $Q(y)$ in the Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} c+d & c-d \\ c-d & c+d \end{bmatrix}$$

- For $Q(y) \in \{q, 1-q\}$ the equation $\|e^{\beta\, u(x,y)}\| Q(y) = e^{\gamma(\beta, x)}$ is

$$\begin{bmatrix} q \\ 1-q \end{bmatrix} = \begin{bmatrix} \frac{p}{1-e^{-2\beta\, d_2}} + \frac{1-p}{1-e^{2\beta\, d_1}} \\ \frac{1-p}{1-e^{-2\beta\, d_1}} + \frac{p}{1-e^{2\beta\, d_2}} \end{bmatrix}$$

- Note that $Q(y) \to P(x)$ as $\beta \to \infty$.
- One can check that $Q(y_1) + Q(y_2) = 1$.
- $\exists\, \beta_0 \geq 0$ such that $Q(y_1) < 0$ or $Q(y_2) < 0$ for $\beta \in [0, \beta_0)$:

$$\beta_0 = \frac{1}{2d} \left| \ln \left(\frac{p}{1-p} \right) \right|$$

Computation of $Q(y)$ in the Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} c+d & c-d \\ c-d & c+d \end{bmatrix}$$

- For $Q(y) \in \{q, 1-q\}$ the equation $\|e^{\beta\, u(x,y)}\| Q(y) = e^{\gamma(\beta, x)}$ is

$$\begin{bmatrix} q \\ 1-q \end{bmatrix} = \begin{bmatrix} \frac{p}{1-e^{-2\beta\, d_2}} + \frac{1-p}{1-e^{2\beta\, d_1}} \\ \frac{1-p}{1-e^{-2\beta\, d_1}} + \frac{p}{1-e^{2\beta\, d_2}} \end{bmatrix}$$

- Note that $Q(y) \to P(x)$ as $\beta \to \infty$.
- One can check that $Q(y_1) + Q(y_2) = 1$.
- $\exists\, \beta_0 \geq 0$ such that $Q(y_1) < 0$ or $Q(y_2) < 0$ for $\beta \in [0, \beta_0)$:

$$\beta_0 = \frac{1}{2d} \left| \ln\left(\frac{p}{1-p}\right) \right|$$

- $I(\beta_0) = 0$ and $U(\beta_0) = c + d|2p-1|$.

Computation of $Q(y)$ in the Binary Case

- Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$ and $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} c+d & c-d \\ c-d & c+d \end{bmatrix}$$

- For $Q(y) \in \{q, 1-q\}$ the equation $\|e^{\beta\, u(x,y)}\| Q(y) = e^{\gamma(\beta,x)}$ is

$$\begin{bmatrix} q \\ 1-q \end{bmatrix} = \begin{bmatrix} \frac{p}{1-e^{-2\beta\, d_2}} + \frac{1-p}{1-e^{2\beta\, d_1}} \\ \frac{1-p}{1-e^{-2\beta\, d_1}} + \frac{p}{1-e^{2\beta\, d_2}} \end{bmatrix}$$

- Note that $Q(y) \to P(x)$ as $\beta \to \infty$.
- One can check that $Q(y_1) + Q(y_2) = 1$.
- $\exists\, \beta_0 \geq 0$ such that $Q(y_1) < 0$ or $Q(y_2) < 0$ for $\beta \in [0, \beta_0)$:

$$\beta_0 = \frac{1}{2d} \left| \ln\left(\frac{p}{1-p}\right) \right|$$

- $I(\beta_0) = 0$ and $U(\beta_0) = c + d|2p - 1| = U(I = 0)$.

Example: The Mean-Square Case

- Let $u(x, y) = -\frac{1}{2}|x - y|^2$

Example: The Mean-Square Case

- Let $u(x, y) = -\frac{1}{2}|x - y|^2$
- Optimal transition kernels are Gaussian

$$p(x \mid y) = e^{-\beta \, \frac{1}{2}|x-y|^2 - \Gamma_0(\beta)}$$

Example: The Mean-Square Case

- Let $u(x, y) = -\frac{1}{2}|x - y|^2$
- Optimal transition kernels are Gaussian

$$p(x \mid y) = e^{-\beta \frac{1}{2}|x-y|^2 - \Gamma_0(\beta)}$$

-

$$\Gamma_0(\beta) = \ln \int_{-\infty}^{\infty} e^{-\beta \frac{1}{2}|x-y|^2} \, dx = \frac{1}{2} \ln \frac{2\pi}{\beta}$$

Example: The Mean-Square Case

- Let $u(x, y) = -\frac{1}{2}|x - y|^2$
- Optimal transition kernels are Gaussian

$$p(x \mid y) = e^{-\beta\,\frac{1}{2}|x-y|^2 - \Gamma_0(\beta)}$$

- $$\Gamma_0(\beta) = \ln \int_{-\infty}^{\infty} e^{-\beta\,\frac{1}{2}\,|x-y|^2}\,dx = \frac{1}{2}\ln\frac{2\pi}{\beta}$$

$$U(\beta) = \Gamma_0'(\beta) = -\frac{1}{2\beta}$$

Example: The Mean-Square Case

- Let $u(x,y) = -\frac{1}{2}|x-y|^2$
- Optimal transition kernels are Gaussian

$$p(x \mid y) = e^{-\beta\,\frac{1}{2}|x-y|^2 - \Gamma_0(\beta)}$$

-

$$\Gamma_0(\beta) = \ln \int_{-\infty}^{\infty} e^{-\beta\,\frac{1}{2}\,|x-y|^2}\,dx = \frac{1}{2}\ln\frac{2\pi}{\beta}$$

$$U(\beta) = \Gamma_0'(\beta) = -\frac{1}{2\beta}$$

$$I(\beta) = H(X) - [\Gamma_0(\beta) - \beta\,\Gamma_0'(\beta)] = H(X) - \frac{1}{2}\ln\frac{2\pi\,e}{\beta}$$

Example: The Mean-Square Case

- Let $u(x, y) = -\frac{1}{2}|x - y|^2$
- Optimal transition kernels are Gaussian

$$p(x \mid y) = e^{-\beta \, \frac{1}{2}|x-y|^2 - \Gamma_0(\beta)}$$

-

$$\Gamma_0(\beta) = \ln \int_{-\infty}^{\infty} e^{-\beta \, \frac{1}{2} \, |x-y|^2} \, dx = \frac{1}{2} \ln \frac{2\pi}{\beta}$$

$$U(\beta) = \Gamma_0'(\beta) = -\frac{1}{2\beta}$$

$$I(\beta) = H(X) - [\Gamma_0(\beta) - \beta \, \Gamma_0'(\beta)] = H(X) - \frac{1}{2} \ln \frac{2\pi \, e}{\beta}$$

$$U(I) = -\frac{1}{4\pi \, e} e^{2[H(X) - I]}$$

Example: The Mean-Square Case

- Let $u(x,y) = -\frac{1}{2}|x-y|^2$
- Optimal transition kernels are Gaussian

$$p(x \mid y) = e^{-\beta\,\frac{1}{2}|x-y|^2 - \Gamma_0(\beta)}$$

-

$$\Gamma_0(\beta) = \ln \int_{-\infty}^{\infty} e^{-\beta\,\frac{1}{2}\,|x-y|^2}\,dx = \frac{1}{2}\ln\frac{2\pi}{\beta}$$

$$U(\beta) = \Gamma_0'(\beta) = -\frac{1}{2\beta}$$

$$I(\beta) = H(X) - [\Gamma_0(\beta) - \beta\,\Gamma_0'(\beta)] = H(X) - \frac{1}{2}\ln\frac{2\pi\,e}{\beta}$$

$$U(I) = -\frac{1}{4\pi\,e}e^{2[H(X)-I]}$$

$$V(I) = U(I) - U(0) = \frac{1}{4\pi\,e}e^{2H(X)}\left(1 - e^{-2I}\right)$$

Minimum RMSE

- Using $U(I)$ for $u(x,y) = -\frac{1}{2}|x-y|^2$:

$$\mathrm{RMSE}(I) = \sqrt{-2U(I)} = \frac{1}{\sqrt{2\pi\, e}}\, e^{H(X)-I}$$

Minimum RMSE

- Using $U(I)$ for $u(x,y) = -\frac{1}{2}|x-y|^2$:

$$\text{RMSE}(I) = \sqrt{-2U(I)} = \frac{1}{\sqrt{2\pi\, e}}\, e^{H(X)-I}$$

- For $x \sim \mathcal{N}(\mu, \sigma_x^2)$ we have $H(X) = \frac{1}{2}\ln(2\pi\, e\sigma_x^2)$

$$\text{RMSE}(I) = \sigma_x\, e^{-I}$$

Minimum RMSE

- Using $U(I)$ for $u(x,y) = -\frac{1}{2}|x-y|^2$:

$$\text{RMSE}(I) = \sqrt{-2U(I)} = \frac{1}{\sqrt{2\pi\, e}}\, e^{H(X)-I}$$

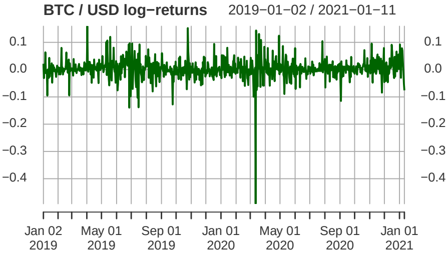
- For $x \sim \mathcal{N}(\mu, \sigma_x^2)$ we have $H(X) = \frac{1}{2}\ln(2\pi\, e\sigma_x^2)$

$$\text{RMSE}(I) = \sigma_x\, e^{-I}\,, \qquad R^2(I) = 1 - e^{-2I}$$

Example: RMSE in Time-Series Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t-2)$	$r(t-1)$	$r(t)$	$r(t+1)$
2019-01-06	-0.031	0.008	-0.011	0.064
2019-01-07	0.008	-0.011	0.064	-0.013
2019-01-08	-0.011	0.064	-0.013	-0.0034
2019-01-09	0.064	-0.013	-0.0034	-0.004



Predict $r(t+1)$ from n lags of $r(t)$ for m symbols (e.g. BTC/USD, ETH/USD, IOT/BTC):

$$f \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix} = y \approx r(t+1)$$

Optimal RMSE using VoI:

$$\text{RMSE}(I) = \sigma_x \, e^{-I}$$

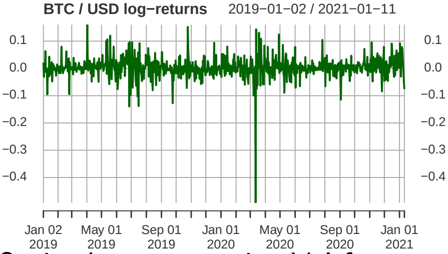
$$R^2(I) = 1 - e^{-2I}$$

Example: Accuracy in Time-Series Prediction

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t-1)$	$r(t)$	sign$r(t+1)$
2019-01-06	0.008	-0.011	1
2019-01-07	-0.011	0.064	-1
2019-01-08	0.064	-0.013	-1
2019-01-09	-0.013	-0.0034	-1

Predict sign $r(t+1)$ from n lags of m symbols (e.g. BTC/USD, ETH/USD, IOT/BTC):



Optimal accuracy using VoI for binary utility as idenity matrix:

$$f \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix} = y \approx \text{sign}[r(t+1)]$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Estimation of Mutual Information

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t-2)$	$r(t-1)$	$r(t)$	$r(t+1)$
2019-01-06	-0.031	0.008	-0.011	0.064
2019-01-07	0.008	-0.011	0.064	-0.013
2019-01-08	-0.011	0.064	-0.013	-0.0034
2019-01-09	0.064	-0.013	-0.0034	-0.004

- Mutual information $I(X, Y) \leq I(X, Z)$ between response x and predictors z.

Estimation of Mutual Information

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

Date	$r(t-2)$	$r(t-1)$	$r(t)$	$r(t+1)$
2019-01-06	-0.031	0.008	-0.011	0.064
2019-01-07	0.008	-0.011	0.064	-0.013
2019-01-08	-0.011	0.064	-0.013	-0.0034
2019-01-09	0.064	-0.013	-0.0034	-0.004

- Mutual information $I(X, Y) \leq I(X, Z)$ between response x and predictors z.
- Here we use Gaussian formula:

$$I_G(X, Z) = \frac{1}{2} \left[\ln \det K_z + \ln \det K_x - \ln \det K_{z \oplus x} \right]$$

where K_i are covariance matrices.

Estimation of Mutual Information

Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

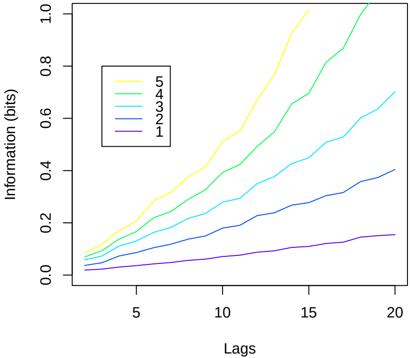
Date	$r(t-2)$	$r(t-1)$	$r(t)$	$r(t+1)$
2019-01-06	-0.031	0.008	-0.011	0.064
2019-01-07	0.008	-0.011	0.064	-0.013
2019-01-08	-0.011	0.064	-0.013	-0.0034
2019-01-09	0.064	-0.013	-0.0034	-0.004

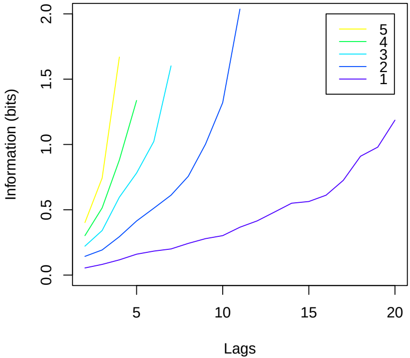
- Mutual information $I(X,Y) \leq I(X,Z)$ between response x and predictors z.
- Here we use Gaussian formula:

$$I_G(X,Z) = \frac{1}{2}\left[\ln\det K_z + \ln\det K_x - \ln\det K_{z\oplus x}\right] \leq I(X,Z)$$

  where K_i are covariance matrices.
- This is sufficient for linear models.

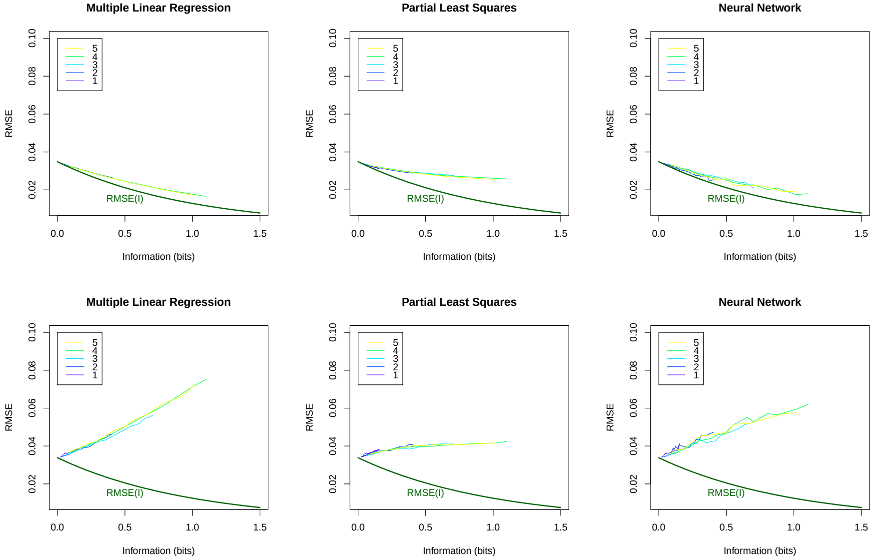
Mutual Information in Training and Testing Sets



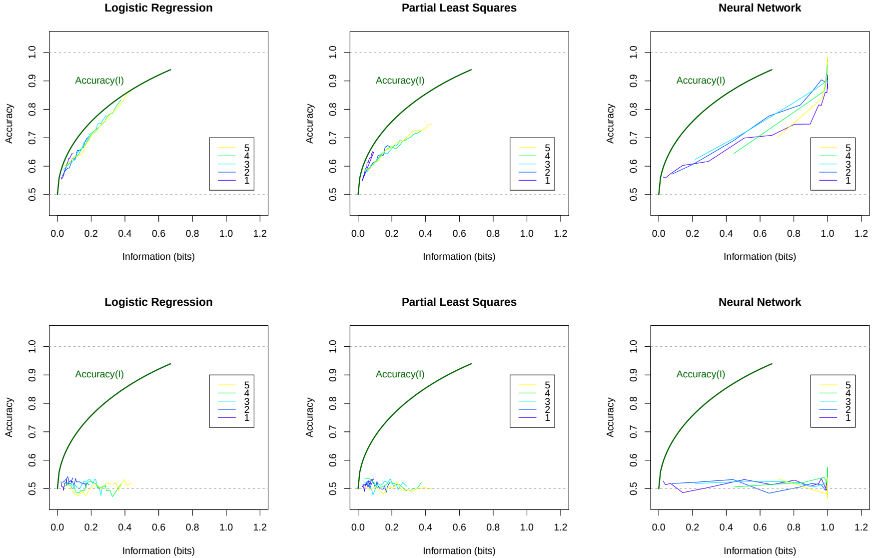


- $n \in [2:20]$ lags.
- $m \in [1:5]$ symbols (BTC/USD, ETH/USD, DAI/BTC, XRP/BTC, IOT/BTC).
- Training / testing sets 100 / 25 days.

Evaluation of RMSE



Evaluation of Accuracy



Other Measures of Model Performance

- Correlation between between prediction y and desired response x:

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y}$$

Other Measures of Model Performance

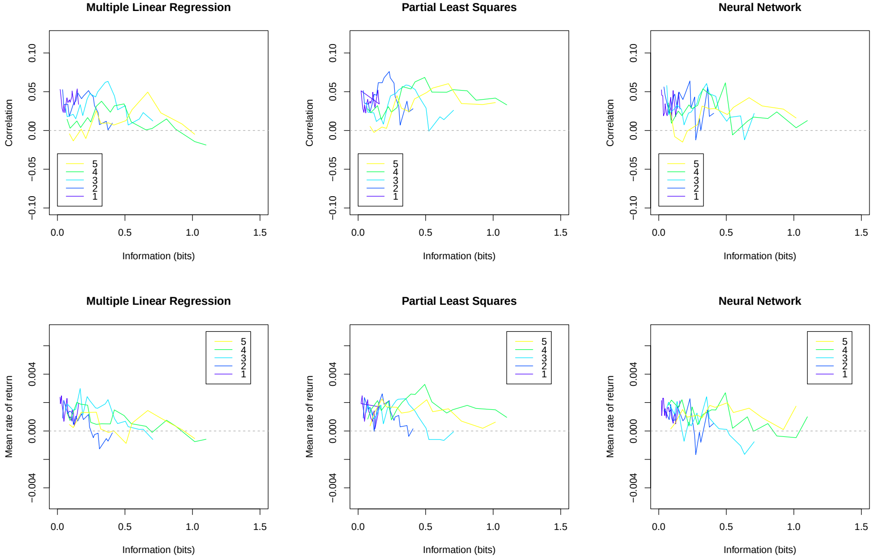
- Correlation between between prediction y and desired response x:

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y}$$

- In the context of day trading, we can estimate daily *Mean Rate of Return* (MRR):

$$\mathrm{MRR} := e^{\mathbb{E}\{\mathrm{sign}(y)\,\mathrm{sign}(x)|x|\}} - 1$$

Evaluation of Correlations and MRRs



Remark

Evolution as an Information Dynamic System

- EPSRC Sandpit '*Math of Life*' (July, 2009):

Evolution as an Information Dynamic System

- EPSRC Sandpit '*Math of Life*' (July, 2009):

- Three year project (2010–13)

Evolution as an Information Dynamic System

- EPSRC Sandpit '*Math of Life*' (July, 2009):

- Three year project (2010–13)
- Followed by two BBSRC project.

Evolution as an Information Dynamic System

- EPSRC Sandpit '*Math of Life*' (July, 2009):

- Three year project (2010–13)
- Followed by two BBSRC project.

Middlesex University : Roman Belavkin

Evolution as an Information Dynamic System

- EPSRC Sandpit '*Math of Life*' (July, 2009):

- Three year project (2010–13)
- Followed by two BBSRC project.

  Middlesex University : Roman Belavkin
  University of Warwick : John Aston

Evolution as an Information Dynamic System

- EPSRC Sandpit '*Math of Life*' (July, 2009):

- Three year project (2010–13)
- Followed by two BBSRC project.

  Middlesex University : Roman Belavkin
  University of Warwick : John Aston
  University of Keele : Alastair Channon & Elizabeth Aston

Evolution as an Information Dynamic System

- EPSRC Sandpit '*Math of Life*' (July, 2009):

- Three year project (2010–13)
- Followed by two BBSRC project.

Middlesex University : Roman Belavkin
University of Warwick : John Aston
University of Keele : Alastair Channon & Elizabeth Aston
University of Manchester : Chris Knight, Rok Krašovec & Danna Gifford

Optimal Mutation Operator

- Optimal solutions achieving $V(I)$ have exponential form, such as:

$$P_{\beta}(b \mid a) = \frac{e^{-\beta\, d(a,b)}}{\sum_z e^{-\beta\, d(a,b)}}$$

Optimal Mutation Operator

- Optimal solutions achieving $V(I)$ have exponential form, such as:

$$P_\beta(b \mid a) = \frac{e^{-\beta\, d(a,b)}}{\sum_z e^{-\beta\, d(a,b)}}$$

- β is called *inverse temperature*, and it is the Lagrange multiplier related to the information constraint:

$$I\{a, b\} \leq I$$

Optimal Mutation Operator

- Optimal solutions achieving $V(I)$ have exponential form, such as:

$$P_\beta(b \mid a) = \frac{e^{-\beta\, d(a,b)}}{\sum_z e^{-\beta\, d(a,b)}}$$

- β is called *inverse temperature*, and it is the Lagrange multiplier related to the information constraint:

$$I\{a,b\} \leq I$$

- The temperature β^{-1} is the slope of $V(I)$:

$$\beta^{-1} = \frac{dV(I)}{dI}$$

Special Case: Hamming Space

Example (Hamming metric)

DNA sequences of length l and alphabet $\{1,\dots,\alpha\}$ are elements of Hamming space $\mathcal{H}^l_\alpha := \{1,\dots,\alpha\}^l$ with Hamming metric

$$d_H(a,b) = \|a-b\|_H = l - \sum_{i=1}^{l} \delta_{a_i}(b_i)$$

Special Case: Hamming Space

Example (Hamming metric)

DNA sequences of length l and alphabet $\{1,\dots,\alpha\}$ are elements of Hamming space $\mathcal{H}^l_\alpha := \{1,\dots,\alpha\}^l$ with Hamming metric

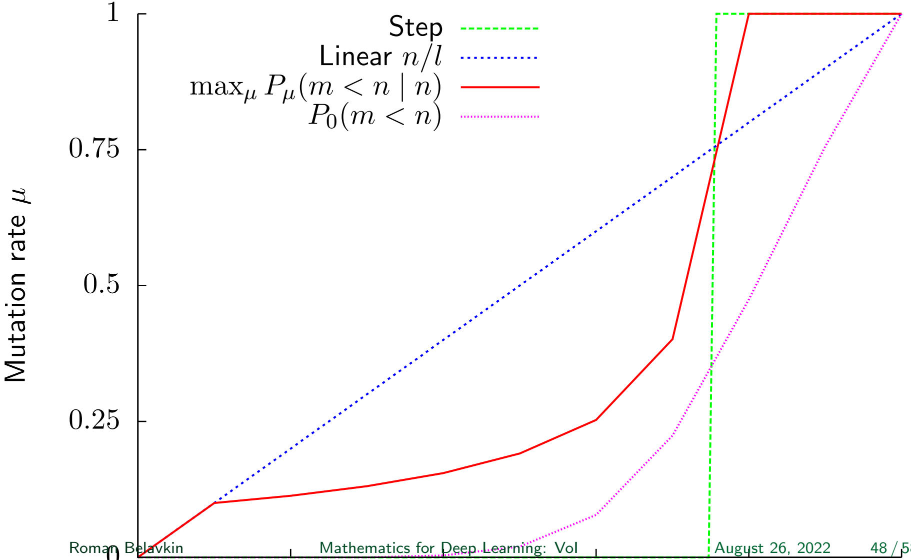
$$d_H(a,b) = \|a-b\|_H = l - \sum_{i=1}^{l} \delta_{a_i}(b_i)$$

Solution

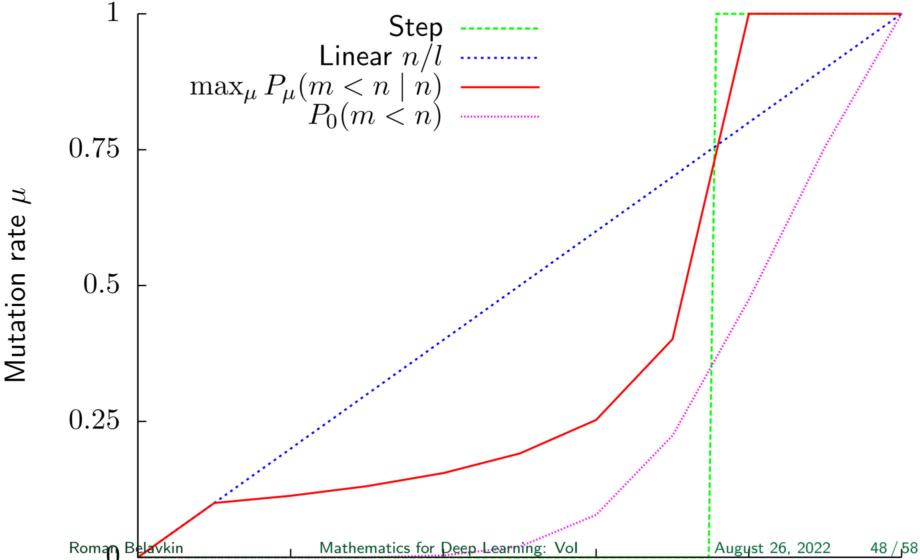
$$P_\beta(b \mid a) = \frac{e^{-\beta\,\|a-b\|_H}}{[1+(\alpha-1)e^{-\beta}]^l} = \prod_{i=1}^{l} \frac{e^{-\beta\,(1-\delta_{a_i}(b_i))}}{1+(\alpha-1)e^{-\beta}}$$

The constraint $\mathbb{E}\{r\} \leq v$ on $r = \|a-b\|_H$ defines $\beta = \ln\left(\mu^{-1}-1\right) + \ln(\alpha-1)$, where $\mu = v/l$ is the mutation rate.

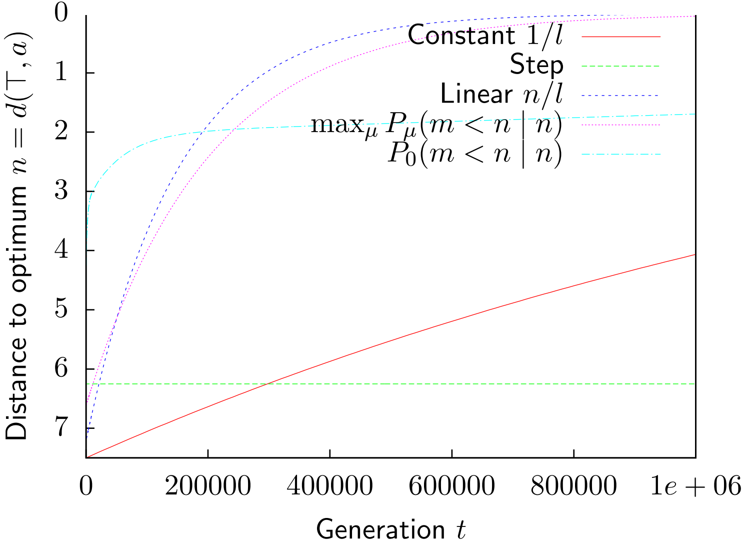
Optimal mutation rate control functions in $\mathcal{H}_4^{10}$



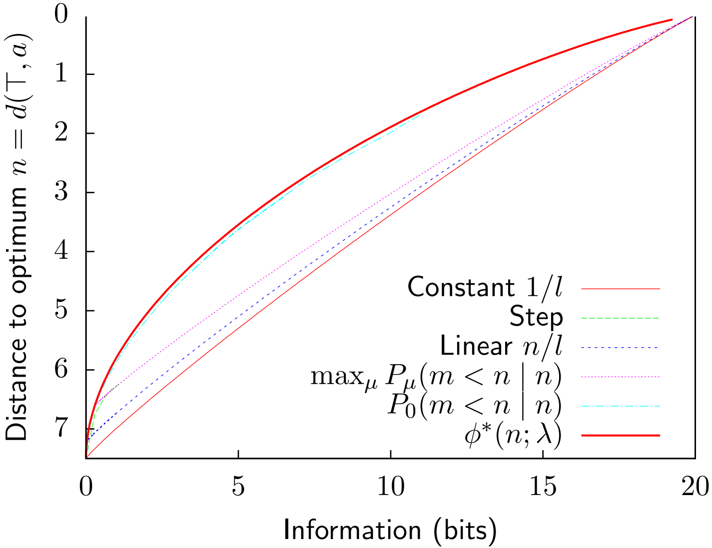
Optimal mutation rate control functions in $\mathcal{H}_4^{10}$



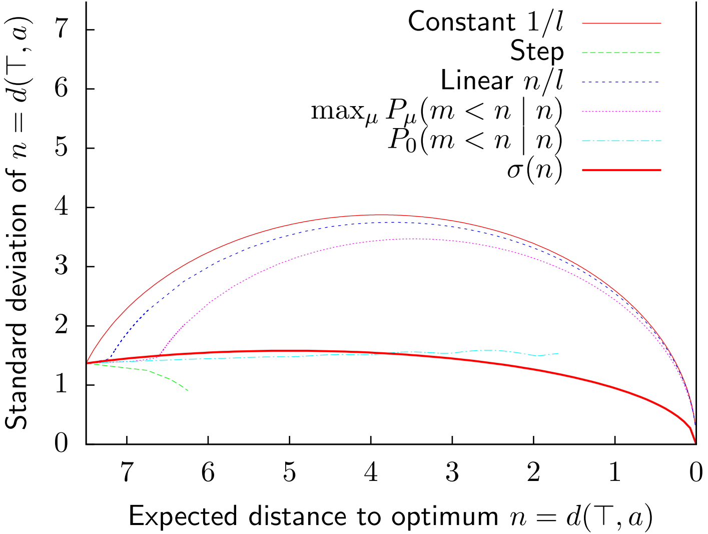
Expected Fitness in Time



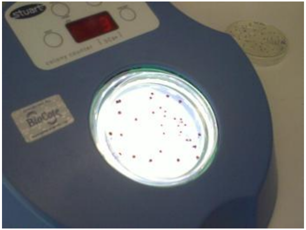
Evolution of Fitness in Information



Fitness Variance and Expectation

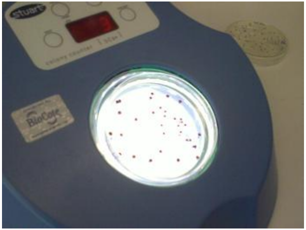


Mutation Rate Control in *E. coli*



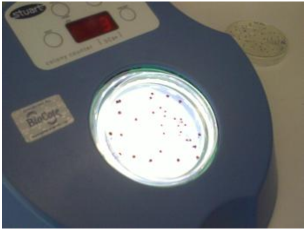
- Used strains of *Escherichia coli* K-12 MG1665

Mutation Rate Control in *E. coli*



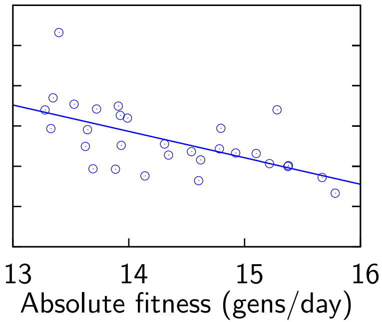
- Used strains of *Escherichia coli* K-12 MG1665
- Fluctuation test using media 50μg/ml of Rifamipicin

Mutation Rate Control in *E. coli*

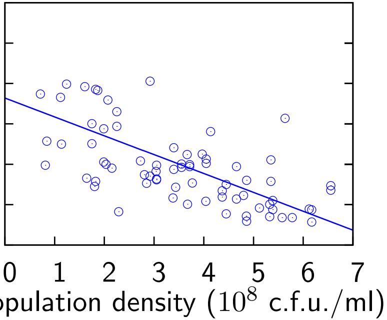


- Used strains of *Escherichia coli* K-12 MG1665
- Fluctuation test using media 50μg/ml of Rifamipicin
- Estimated mutation rates μ in *E.coli* strains grown in Davis minimal medium with different amount of glucose.

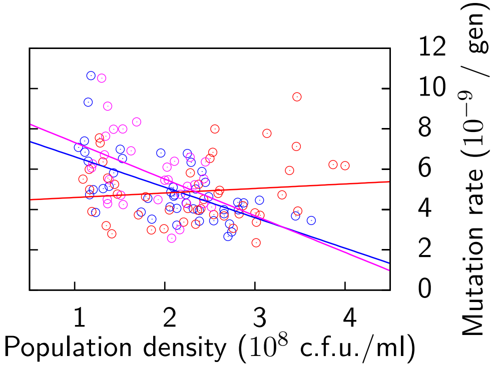
Experimental Results (Krašovec et al., 2014)



Experimental Results (Krašovec et al., 2014)

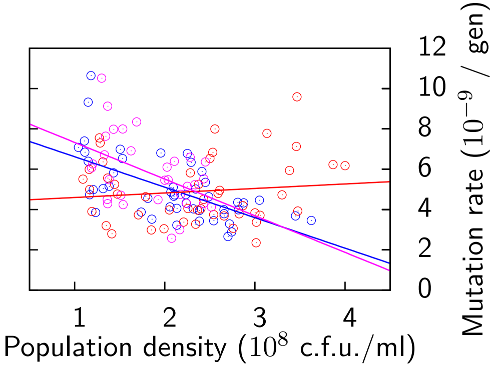


Experimental Results (Krašovec et al., 2014)



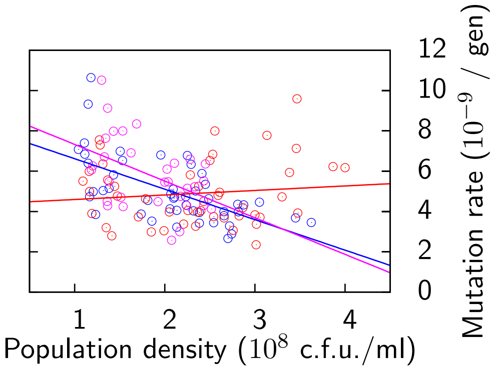
- Strong relationship between μ and density of cells ($p < .0001$).

Experimental Results (Krašovec et al., 2014)



- Strong relationship between μ and density of cells ($p < .0001$).
- No such relationship in the *luxS* quorum sensing mutant ($p = .0234$).

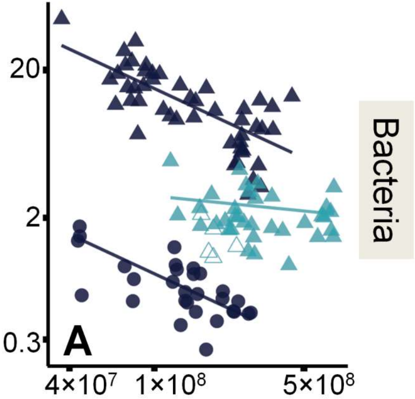
Experimental Results (Krašovec et al., 2014)



- Strong relationship between μ and density of cells ($p < .0001$).
- No such relationship in the *luxS* quorum sensing mutant ($p = .0234$).

Krašovec, R., Belavkin, R., Aston, J., Channon, A., Aston, E., Rash, B., Kadirvel, M., Forbes. S., Knight, C. G. (2014, April). Mutation-rate-plasticity in rifampicin resistance depends on Escherichia coli cell-cell interactions. *Nature Communications*, Vol. 5 (3742).

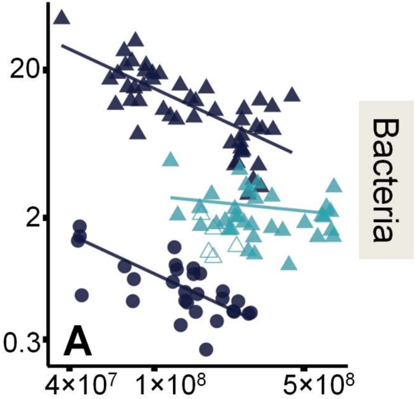
Plastic mutation rates in bacteria (Krašovec et al., 2017)



- *rifampicin* (triangles)

Krašovec, R., Richards, H., Gifford, D. R., Hatcher, C., Faulkner, K. J., Belavkin, R. V., Channon, A., Aston, E., McBain, A. J., Knight, C. G. (2017). Spontaneous mutation rate is a plastic trait associated with population density across domains of life. *PLoS Biology*, 15:8.

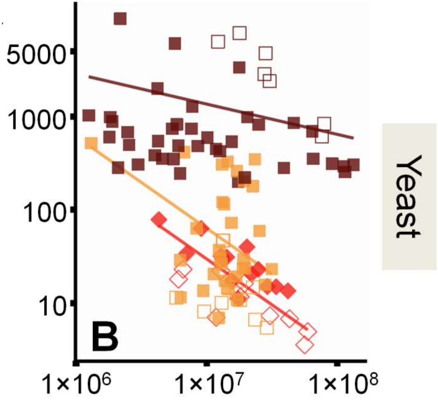
Plastic mutation rates in bacteria (Krašovec et al., 2017)



- *rifampicin* (triangles)
- *nalidixic acid* in *E.coli* (dark circles) and in *P. aeruginosa* (light circles)

Krašovec, R., Richards, H., Gifford, D. R., Hatcher, C., Faulkner, K. J., Belavkin, R. V., Channon, A., Aston, E., McBain, A. J., Knight, C. G. (2017). Spontaneous mutation rate is a plastic trait associated with population density across domains of life. *PLoS Biology*, 15:8.

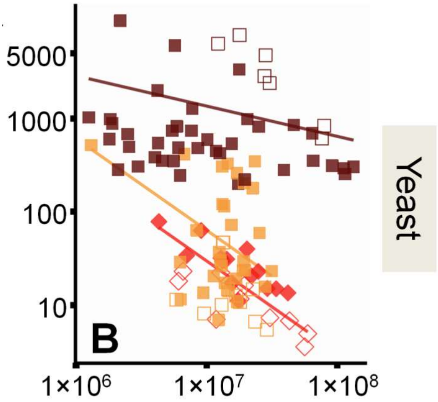
Plastic mutation rates in yeast (Krašovec et al., 2017)



- *hygromycin B* (squares) in *S. cerevisiae*

Krašovec, R., Richards, H., Gifford, D. R., Hatcher, C., Faulkner, K. J., Belavkin, R. V., Channon, A., Aston, E., McBain, A. J., Knight, C. G. (2017). Spontaneous mutation rate is a plastic trait associated with population density across domains of life. *PLoS Biology*, 15:8.

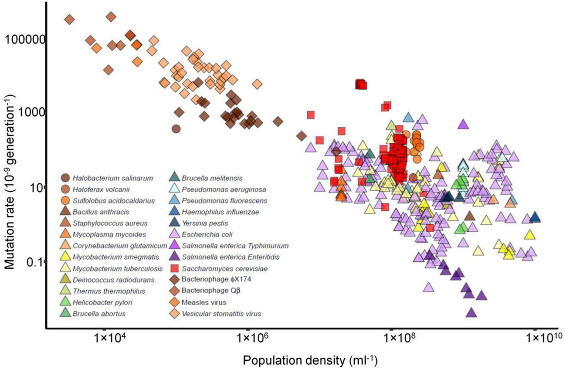
Plastic mutation rates in yeast (Krašovec et al., 2017)



- *hygromycin B* (squares) in *S. cerevisiae*
- *5-FOA* (diamonds)

Krašovec, R., Richards, H., Gifford, D. R., Hatcher, C., Faulkner, K. J., Belavkin, R. V., Channon, A., Aston, E., McBain, A. J., Knight, C. G. (2017). Spontaneous mutation rate is a plastic trait associated with population density across domains of life. *PLoS Biology*, 15:8.

Plastic rates in all domains of life (Krašovec et al., 2017)



>70 years of published data (1943–2016), 67 studies, 26 species.

Krašovec, R., Richards, H., Gifford, D. R., Hatcher, C., Faulkner, K. J., Belavkin, R. V., Channon, A., Aston, E., McBain, A. J., Knight, C. G. (2017). Spontaneous mutation rate is a plastic trait associated with population density across domains of life. *PLoS Biology*, 15:8.

Conclusions

- Presented basic ideas of the value of information theory.

Conclusions

- Presented basic ideas of the value of information theory.
- Used the binary and the mean-square cases to derive formulae for the minimum RMSE and the maximum accuracy of a model as function of information.

Conclusions

- Presented basic ideas of the value of information theory.
- Used the binary and the mean-square cases to derive formulae for the minimum RMSE and the maximum accuracy of a model as function of information.
- VoI gives additional tools to evaluate model performance.

Conclusions

- Presented basic ideas of the value of information theory.
- Used the binary and the mean-square cases to derive formulae for the minimum RMSE and the maximum accuracy of a model as function of information.
- VoI gives additional tools to evaluate model performance.
- The theory provides some deep insights into random phenomena, learning and decisions under uncertainty.

Conclusions

- Presented basic ideas of the value of information theory.
- Used the binary and the mean-square cases to derive formulae for the minimum RMSE and the maximum accuracy of a model as function of information.
- VoI gives additional tools to evaluate model performance.
- The theory provides some deep insights into random phenomena, learning and decisions under uncertainty.
- Control of parameters (mutation rates, learning rates, annealing schedule, exploration-exploitation balance, etc).

Belavkin, R. V. (2013). Optimal measures and Markov transition kernels. *Journal of Global Optimization*, *55*, 387–416.

Belavkin, R. V. (2018). Relation between the Kantorovich-Wasserstein metric and the Kullback-Leibler divergence. In N. Ay, P. Gibilisco, & F. Matúš (Eds.), *Information geometry and its applications* (pp. 363–373). Springer International Publishing.

Krašovec, R., Belavkin, R. V., Aston, J. A. D., Channon, A., Aston, E., Rash, B. M., ... Knight, C. G. (2014, April). Mutation rate plasticity in rifampicin resistance depends on escherichia coli cell-cell interactions. *Nature Communications*, *5*(3742).

Krašovec, R., Richards, H., Gifford, D. R., Hatcher, C., Faulkner, K. J., Belavkin, R. V., ... Knight, C. G. (2017). Spontaneous mutation rate is a plastic trait associated with population density across domains of life. *PLoS Biology*, *15*(8).

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. (1965). On value of information. *Izvestiya of USSR*

Academy of Sciences, Technical Cybernetics, *5*, 3–12. (In Russian)

Stratonovich, R. L. (1975). *Theory of information*. Moscow, USSR: Sovetskoe Radio. (In Russian)

Stratonovich, R. L. (2020). *Theory of information and its value* (R. V. Belavkin, P. M. Pardalos, & J. C. Principe, Eds.). Springer.