# Data, Information and its Value: Applications to Model Selection and Parameter Control

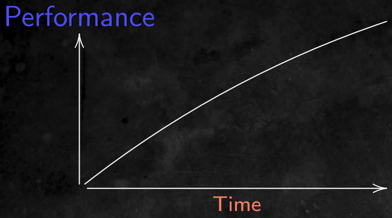
Roman V. Belavkin (Faculty of Science and Technology, Middlesex University, London NW4 4BT, UK )

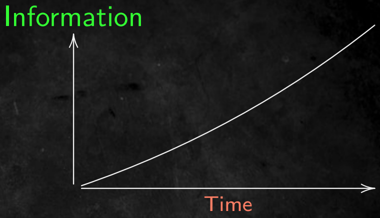
May 24, 20024

XVI Summer School on Operational Research, Data and Decision Making
National Research University Higher School of Economics,
Nizhny Novgorod
Faculty of Informatics, Mathematics and Computer Science
Laboratory of Algorithms and technologies for network analysis

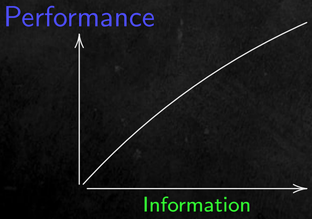
1. Overview of the Value of Information (VoI) Theory
    - Motivation: Learning Systems
    - Example: Mean-Square Minimization
    - Variational Problems
2. Solution for Shannon's VoI
3. Computation of VoI
    - The Binary Case
    - The Mean-Square Case
4. Applications of VoI
    - Performance of regression models
    - Performance of classification models
    - Performance of evolutionary systems

# Learning Systems



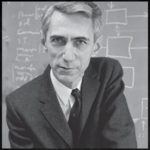

# Optimal learning




Maximize performance

s.t. information $\leq \lambda$

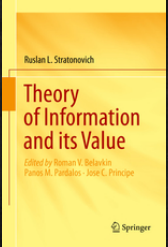
# Information and its Value



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):

Ruslan Stratonovich

- R. V. Belavkin (2013). Optimal measures and Markov transition kernels. *Journal of Global Optimization*, Vol. 55 (387–416).
- R. V. Belavkin (2010). Utility and Value of Information in Cognitive Science, Biology and Quantum Theory. *Quantum Bio-Informatics III*, 2009.

# Three Types of Information

## Definition (Hartley Information)

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

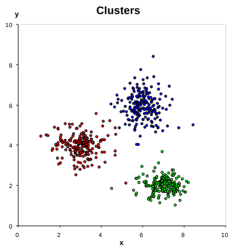
## Definition (Boltzmann Information)

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

## Definition (Shannon Information)

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

# 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 Hartley's information

- Define the following quantities:

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

$$U(I) := \sup_{z(x)} \mathbb{E}_{P(z)}\left\{\sup_{y(z)} \mathbb{E}_{P(x|z)}\{u(x,y) \mid z\}\right\}$$

Subject to $|Z| \leq e^I$

- The *value of Hartley information* (Stratonovich, 1965):

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

### Remark

*We are looking for optimal function* $y(x) = y \circ z(x)$ *subject to* $|Y| \leq |Z| \leq e^I$.

## Value of Shannon's information

- Define the following quantities:

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

$$U(I) := \sup_{P(y|x)} \mathbb{E}_{P(x,y)}\{u(x,y)\}$$

Subject to $I(X,Y) \leq I$

- The *value of Shannon's information* (Stratonovich, 1965):

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

### Remark

*Instead of functions* $y(x) = y \circ z(x)$*, we are looking for optimal* $P(y \mid x)$ *subject to* $I(X,Y) \leq I$*.*
*See R. V. Belavkin, 2018 for a relation to optimal transport.*

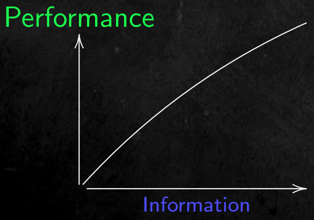
## VoI as Conditional Extremum

- Linear programming problem $V(I)$:

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

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

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




Maximize performance
s.t. information $\leq I$

## Outline of 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:

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

- Optimal *inverse temperature* $\beta$:

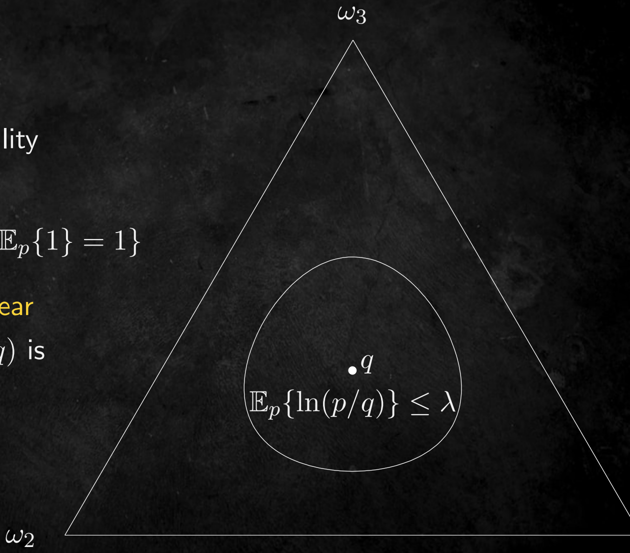
$$\beta=\frac{dI(U)}{dU}\quad\text{or}\quad\beta^{-1}=\frac{dU(I)}{dI}\,\beta\frac{d}{d\beta}\Gamma(\beta)-\Gamma(\beta)=I$$

## Information-Geometric View

- The set of all probability measures

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

- $\mathbb{E}_p\{u\} := \langle u, p\rangle$ is linear
- $\mathbb{E}_p\{\ln(p/q)\} =: I(p,q)$ is convex



## Information-Geometruc View (cont)

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

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

- Minimize $I(p,q)$:

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

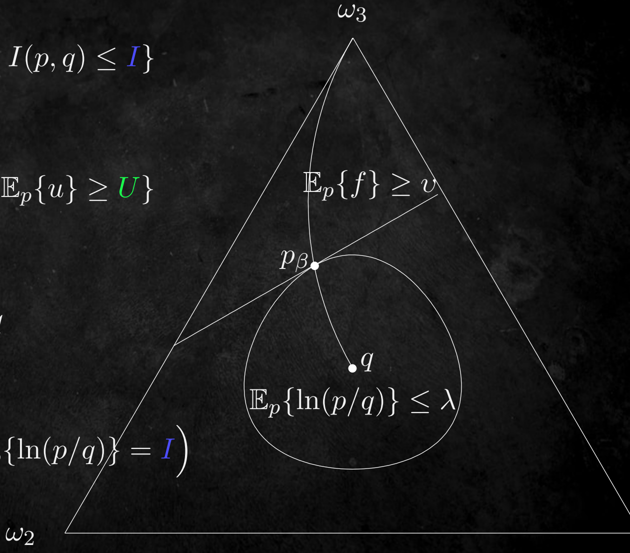
- Optimal solutions:

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

- Constraints

$$\mathbb{E}_{p(\beta)}\{u\} = U, \quad \left(\mathbb{E}_p\{\ln(p/q)\} = I\right)$$

- $\mathcal{P}(X \otimes Y)$



## Solution to the Value of Shannon's Information

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

$$P(x \mid y) =$$

- 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:

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

- 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

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

### Theorem

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

## Example: The Binary Case (R. Belavkin, Pardalos, & Principe, 2022)

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

$$\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix} \begin{bmatrix} c_1 + d_1 & c_1 - d_1 \\ c_2 - d_2 & c_2 + d_2 \end{bmatrix} \begin{bmatrix} c + d & c - d \\ c - d & c + d \end{bmatrix}$$

- 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 $u : X \times Y \to \mathbb{R}$:

$$\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix} \begin{bmatrix} c_1 + d_1 & c_1 - d_1 \\ c_2 - d_2 & c_2 + d_2 \end{bmatrix} \begin{bmatrix} c + d & c - d \\ c - d & c + d \end{bmatrix}$$

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

## Computation of VoI for Translation Invariant Utility

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

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

- For translation invariant $u(x,y) = u(x+a, y+a)$ one can show:

$$\Gamma_0(\beta) = \ln \int_X e^{\beta\, u(x,y)}\, dx$$

- $\beta$ is the *inverse temperature* related to the constraint $I(X,Y) \leq I$:

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

- Find $U(I)$ from

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

## Example: The Mean-Square Case (R. Belavkin, Pardalos, & Principe, 2023)

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

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

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

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

(R. Belavkin et al., 2023)

# Example: Time-Series Prediction

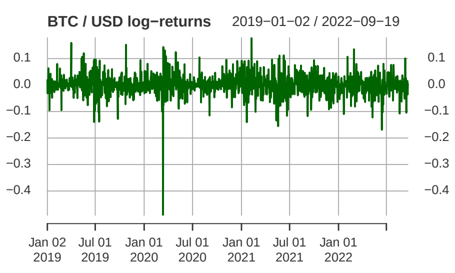



Table: log-returns $r(t) = \log \frac{S(t+1)}{S(t)}$

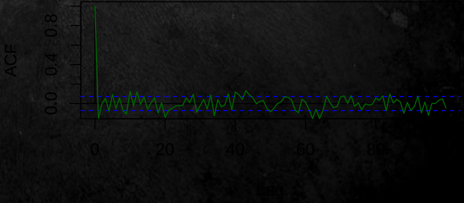
| Date | $r(t-2)$ | $r(t-1)$ | $r(t)$ | $r(t+1)$ |
|---|---|---|---|---|
| 2019-01-06 | -0.031 | 0.008 | -0.011 | 0.064 |
| 2019-01-07 | 0.008 | -0.011 | 0.064 | -0.013 |
| 2019-01-08 | -0.011 | 0.064 | -0.013 | -0.0034 |
| 2019-01-09 | 0.064 | -0.013 | -0.0034 | -0.004 |

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

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

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

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



## Estimation of Mutual Information

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

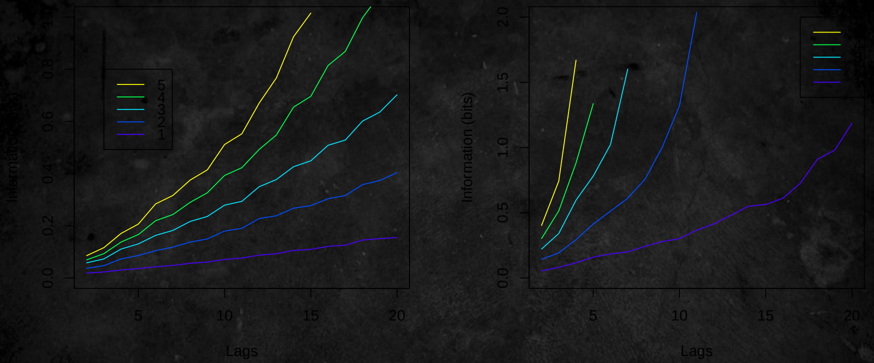
| Date | $r(t-2)$ | $r(t-1)$ | $r(t)$ | $r(t+1)$ |
|---|---|---|---|---|
| 2019-01-06 | -0.031 | 0.008 | -0.011 | 0.064 |
| 2019-01-07 | 0.008 | -0.011 | 0.064 | -0.013 |
| 2019-01-08 | -0.011 | 0.064 | -0.013 | -0.0034 |
| 2019-01-09 | 0.064 | -0.013 | -0.0034 | -0.004 |

- 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.

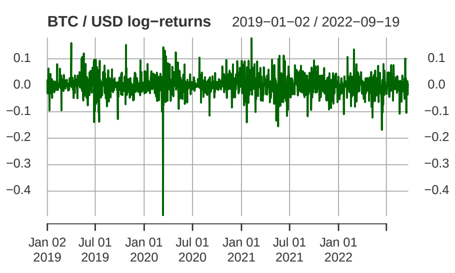
## Mutual Information in Training and Testing Sets




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

## Model Performance

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

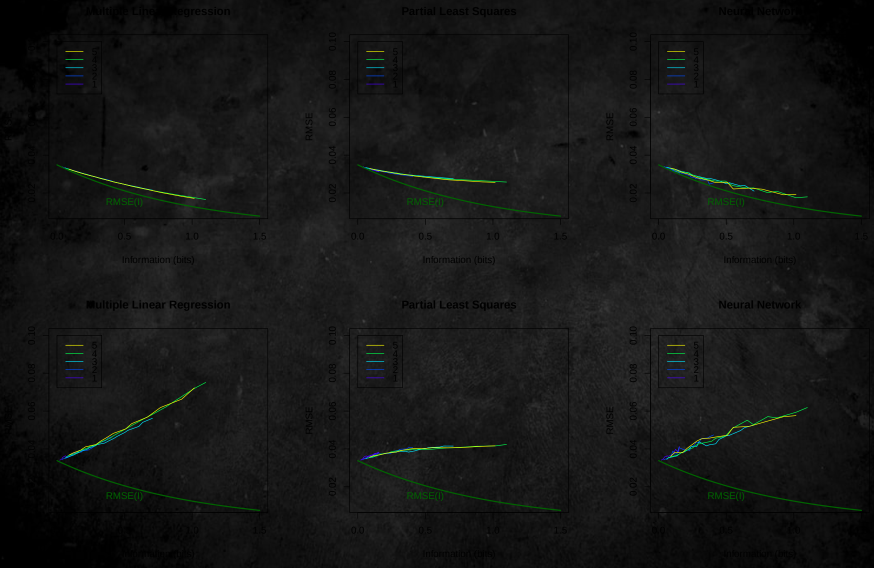



- Use $n \in [2:20]$ lags and $m \in [1:5]$ symbols (i.e. $m \times n \in [2:100]$).
- Models: linear regression, partial-least squares, neural net.
- Root mean-square error

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

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

# Evaluation of RMSE

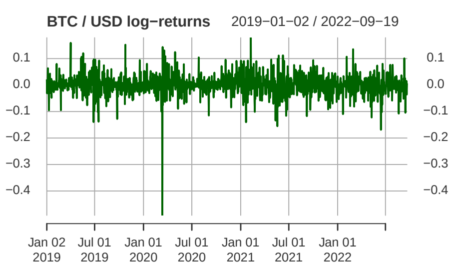



(R. Belavkin et al., 2023)

# 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):

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

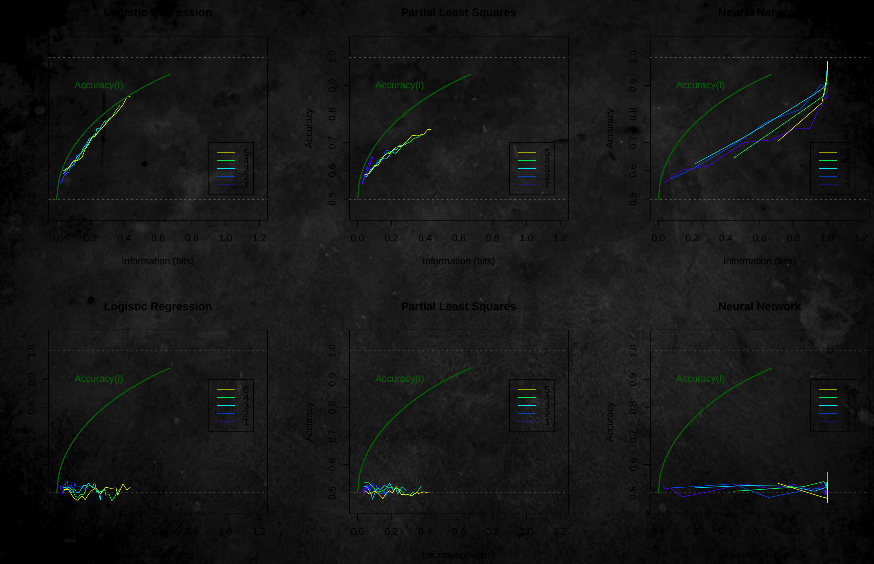
Utility $u(x,y)$ is a $2 \times 2$ matrix (confusion matrix):

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

## Questions:

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

# Evaluation of Accuracy




(R. Belavkin et al., 2022)

## 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:

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

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

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

- The temperature $\beta^{-1}$ 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

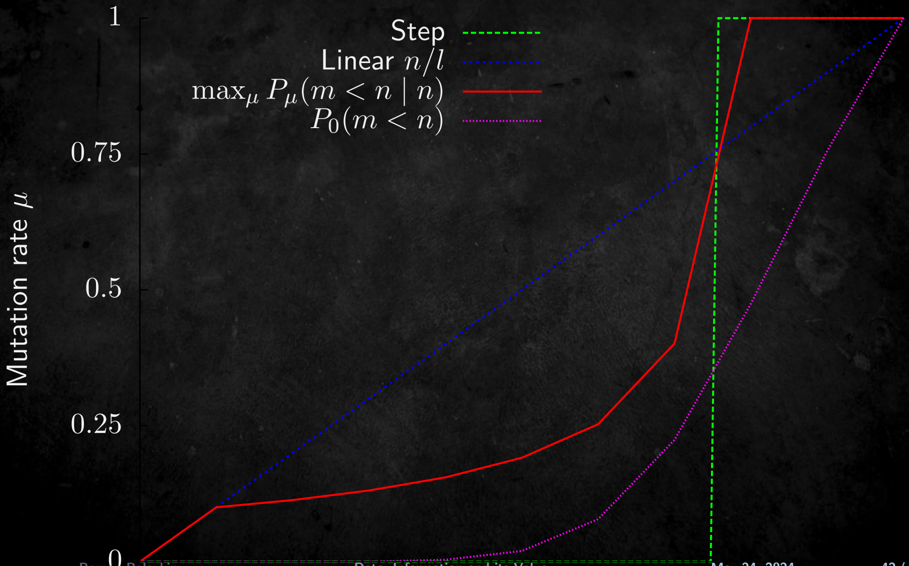
$$d_H(a,b) = \|a-b\|_H = l - \sum_{i=1}^{l} \delta_{a_i}(b_i)$$

### Solution

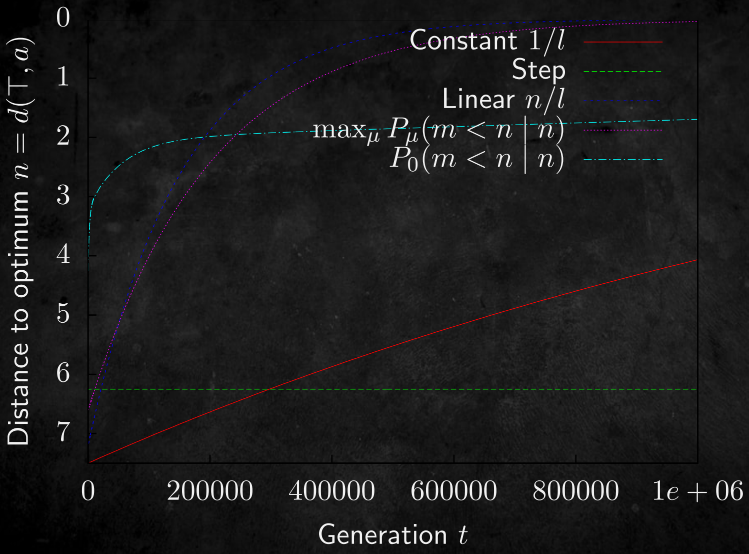
$$P_\beta(b \mid a) = \frac{e^{-\beta\,\|a-b\|_H}}{[1+(\alpha-1)e^{-\beta}]^l} = \prod_{i=1}^{l} \frac{e^{-\beta\,(1-\delta_{a_i}(b_i))}}{1+(\alpha-1)e^{-\beta}}$$

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

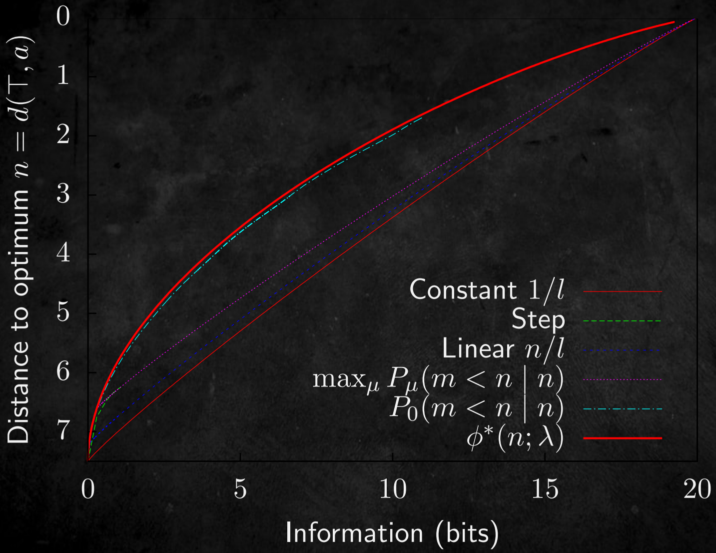
## Optimal mutation rate control functions in $\mathcal{H}_4^{10}$



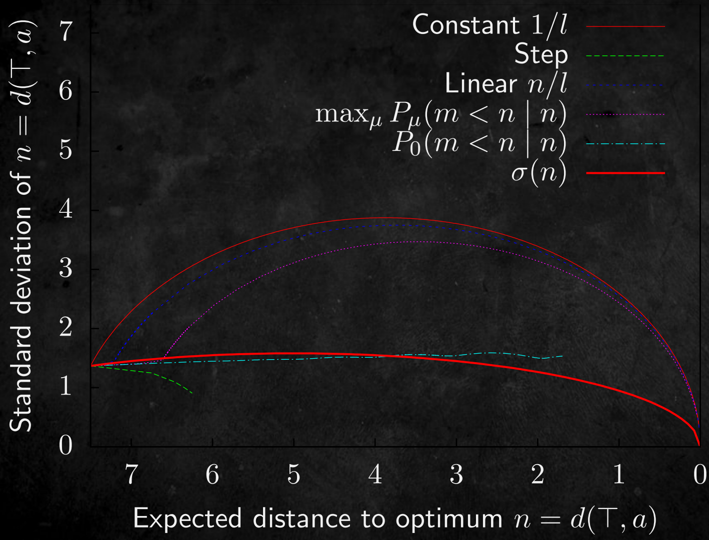
## Expected Fitness in Time



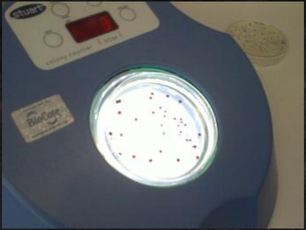
## Evolution of Fitness in Information



## Fitness Variance and Expectation

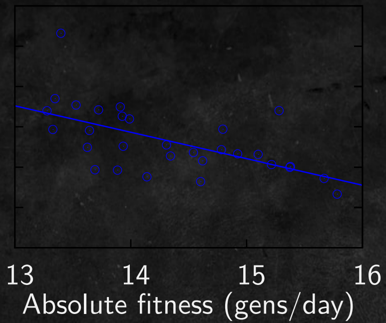


# Mutation Rate Control in E. coli




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

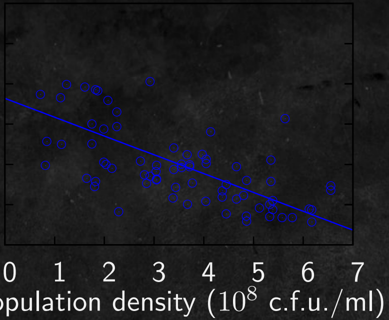
# Experimental Results (Krašovec et al., 2014)




- Strong relationship between $\mu$ 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).

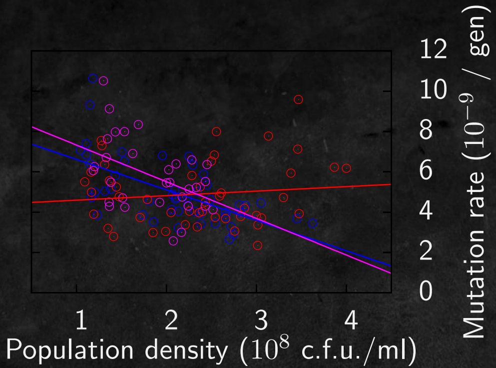
# Experimental Results (Krašovec et al., 2014)




- Strong relationship between $\mu$ 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).

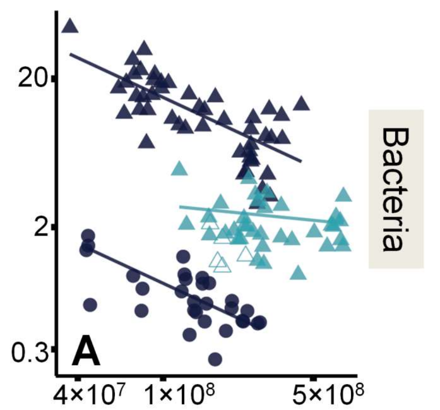
## Experimental Results (Krašovec et al., 2014)




- Strong relationship between $\mu$ 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).

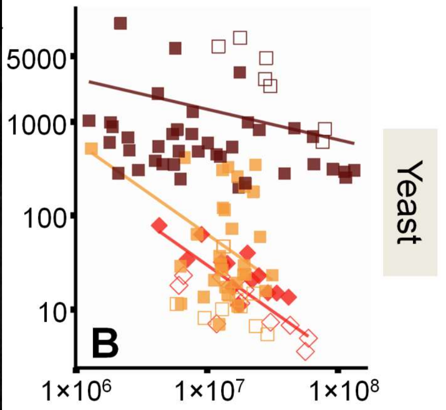
## Plastic mutation rates in bacteria (Krašovec et al., 2017)




- *rifampicin* (triangles)
- *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.

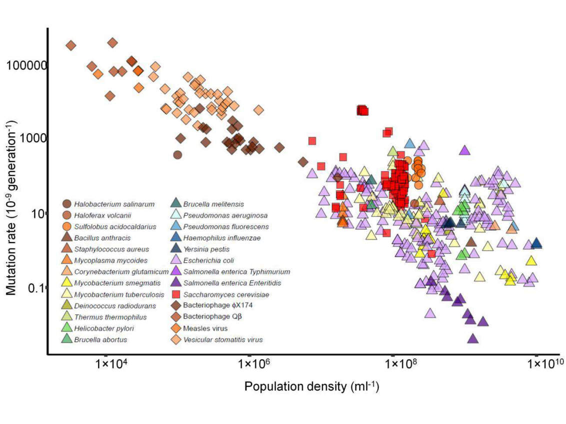
# Plastic mutation rates in yeast (Krašovec et al., 2017)




- *hygromycin B* (squares) in *S. cerevisiae*
- *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.
- 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., Pardalos, P., & Principe, J. (2022). Value of information in the binary case and confusion matrix. *Physical Sciences Forum*, *5*(1). Retrieved from `https://www.mdpi.com/2673-9984/5/1/8` doi: 10.3390/psf2022005008

Belavkin, R., Pardalos, P., & Principe, J. (2023). Value of information in the mean-square case and its application to the analysis of financial time-series forecast. In D. E. S. et al. (Ed.), *Learning and intelligent optimization* (pp. 1–15). Switzerland AG: Springer Nature.

Belavkin, R. V. (2010). Utility and value of information in cognitive science, biology and quantum theory. In L. Accardi, W. Freudenberg, & M. Ohya (Eds.), *Quantum Bio-Informatics III* (Vol. 26). World Scientific.

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)