Information Evolution of Optimal Learning

Roman V. Belavkin

School of Computing Science, Middlesex University, London NW4 4BT, UK

May 29, 2008

1. Introduction
   - Historic remarks
   - Optimisation & information in learning
   - Choice and optimisation
   - Preliminaries
   - Probability triangle (simplex)
   - Convex functions
2. Polar sets and their functions
   - Support and distance of a convex body
   - Functions of polar sets
   - Necessary conditions of extrema
   - Characterisation of the solutions
   - Generalised characteristic potentials
3. Evolution of optimal learning
   - Dynamics
   - Information potential fields
   - Information brachistochrone

Historic remarks

- Bayesian estimation, choice under uncertainty (Wald, 1950; Neumann & Morgenstern, 1944; Savage, 1954; Anscombe & Aumann, 1963).

Historic remarks

- Bayesian estimation, choice under uncertainty (Wald, 1950; Neumann & Morgenstern, 1944; Savage, 1954; Anscombe & Aumann, 1963).
- Variational analysis, optimal control, maximum principle (Pontryagin et al., 1962; Gelfand & Fomin, 1963)

Historic remarks

- Bayesian estimation, choice under uncertainty (Wald, 1950; Neumann & Morgenstern, 1944; Savage, 1954; Anscombe & Aumann, 1963).
- Variational analysis, optimal control, maximum principle (Pontryagin et al., 1962; Gelfand & Fomin, 1963)
- Dynamic programming (Bellman, 1957)

Historic remarks

- Bayesian estimation, choice under uncertainty (Wald, 1950; Neumann & Morgenstern, 1944; Savage, 1954; Anscombe & Aumann, 1963).
- Variational analysis, optimal control, maximum principle (Pontryagin et al., 1962; Gelfand & Fomin, 1963)
- Dynamic programming (Bellman, 1957)
- Conditional Markov processes (Stratonovich, 1959, 1960):
  - Stochastic optimal control
  - Optimal non–linear filtering
  - Linear version (Kalman & Bucy, 1961)

Historic remarks

- Bayesian estimation, choice under uncertainty (Wald, 1950; Neumann & Morgenstern, 1944; Savage, 1954; Anscombe & Aumann, 1963).
- Variational analysis, optimal control, maximum principle (Pontryagin et al., 1962; Gelfand & Fomin, 1963)
- Dynamic programming (Bellman, 1957)
- Conditional Markov processes (Stratonovich, 1959, 1960):
  - Stochastic optimal control
  - Optimal non–linear filtering
  - Linear version (Kalman & Bucy, 1961)
- Information theory (Shannon, 1948)
- Information value theory (Kolmogorov, 1957; Jaynes, 1957; Stratonovich, 1965)

Historic remarks

- Bayesian estimation, choice under uncertainty (Wald, 1950; Neumann & Morgenstern, 1944; Savage, 1954; Anscombe & Aumann, 1963).
- Variational analysis, optimal control, maximum principle (Pontryagin et al., 1962; Gelfand & Fomin, 1963)
- Dynamic programming (Bellman, 1957)
- Conditional Markov processes (Stratonovich, 1959, 1960):
  - Stochastic optimal control
  - Optimal non–linear filtering
  - Linear version (Kalman & Bucy, 1961)
- Information theory (Shannon, 1948)
- Information value theory (Kolmogorov, 1957; Jaynes, 1957; Stratonovich, 1965)
- Learning and adaptation?

Optimisation & information in learning

Proposition

There is no need to learn, if there is nothing to optimise

Optimisation & information in learning

Proposition

There is no need to learn, if there is nothing to optimise

Proposition

There is nothing new to learn, if one has full information

Choice and optimisation

Definition

A preference relation is $\lesssim \subseteq A^2$ which is

1. Total: $(a_1 \lesssim a_2) \vee (a_2 \lesssim a_1) = A^2$
2. Transitive: $(a_1 \lesssim a_2) \wedge (a_2 \lesssim a_3) \Longrightarrow (a_1 \lesssim a_3)$

Choice and optimisation

Definition

A preference relation is $\lesssim \subseteq A^2$ which is

1. Total: $(a_1 \lesssim a_2) \vee (a_2 \lesssim a_1) = A^2$
2. Transitive: $(a_1 \lesssim a_2) \wedge (a_2 \lesssim a_3) \Longrightarrow (a_1 \lesssim a_3)$

Definition

Given $(A, \lesssim)$, Alexandrov topology is $T \subset 2^A$ such that $\emptyset$, A and all upper sets are in T.

Choice and optimisation

Definition

A preference relation is $\lesssim \subseteq A^2$ which is

1. Total: $(a_1 \lesssim a_2) \vee (a_2 \lesssim a_1) = A^2$
2. Transitive: $(a_1 \lesssim a_2) \wedge (a_2 \lesssim a_3) \Longrightarrow (a_1 \lesssim a_3)$

Definition

Given $(A, \lesssim)$, Alexandrov topology is $T \subset 2^A$ such that $\emptyset$, A and all upper sets are in T.

Lemma

Every set with preference relation is Borel measurable.

Preference morphisms

Lemma

For any $f : A \to B$ *the following are equivalent:*

1. f *is monotone:* $a_1 \lesssim_A a_2 \implies f(a_1) \lesssim_B f(a_2)$;
2. f *is continuous:* $f(a) = b \implies f(O_a) \subset O_b$;
3. f *is measurable:* $f^{-1}(B' \in \mathcal{B}) = A' \in \mathcal{A}$

Preference morphisms

Lemma

For any $f : A \to B$ the following are equivalent:

1. *f is monotone: $a_1 \lesssim_A a_2 \implies f(a_1) \lesssim_B f(a_2)$;*
2. *f is continuous: $f(a) = b \implies f(O_a) \subset O_b$;*
3. *f is measurable: $f^{-1}(B' \in \mathcal{B}) = A' \in \mathcal{A}$*

Definition

Function $u : A \to \overline{\mathbb{R}}$ is called a utility representation of preference relation $(A, \lesssim)$ if $\forall\, a_1,\, a_2 \in A$

$$a_1 \lesssim a_2 \implies u(a_1) \leq u(a_2)$$

Preliminaries

Dual spaces X and Y of functions $x : A \to \overline{\mathbb{R}}$ and $y : A \to \overline{\mathbb{R}}$

Preliminaries

Dual spaces X and Y of functions $x : A \to \overline{\mathbb{R}}$ and $y : A \to \overline{\mathbb{R}}$

$x \in X$ and $y \in Y$ are coupled in H (i.e. $X \subset H \subset Y$) by a bilinear form $(\cdot, \cdot) : X \times Y \to \overline{\mathbb{R}}$, such as the inner product

$$(x, y) = \int_A x(a)y(a)da$$

Preliminaries

Dual spaces X and Y of functions $x : A \to \overline{\mathbb{R}}$ and $y : A \to \overline{\mathbb{R}}$

$x \in X$ and $y \in Y$ are coupled in H (i.e. $X \subset H \subset Y$) by a bilinear form $(\cdot, \cdot) : X \times Y \to \overline{\mathbb{R}}$, such as the inner product

$$(x, y) = \int_A x(a)y(a)da$$

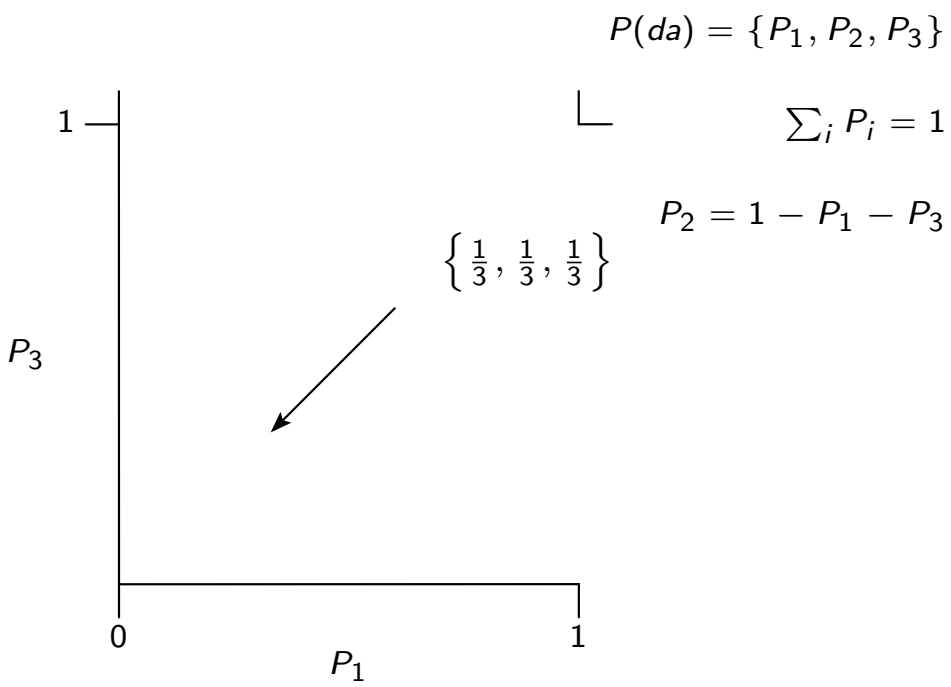
Expected value

$$E\{x\} = \int_A x(a)P(da) = (x, y)$$

where $y(a) \geq 0$, $\int_A y(a)da = P(A) = 1$

Probability triangle (simplex)

$E\{x\}$ is a unique representation of $(\mathcal{P}, \lesssim_{\mathcal{P}})$, where $\mathcal{P}$ is the set of all probability measures in Y (Neumann & Morgenstern, 1944).



Convex functions

Definition

$F : X \to \mathbb{R}_\infty$ is convex if $\forall x \in X$

$$F((1-t)x_1 + t\,x_2) \le (1-t)F(x_1) + t\,F(x_2), \qquad t \in [0,1]$$

F is proper if $\mathrm{dom}(F) \equiv \{x \in X : F(x) < \infty\} \neq \emptyset$

Convex functions

Definition

$F : X \to \mathbb{R}_\infty$ is convex if $\forall x \in X$

$$F((1-t)x_1 + t\,x_2) \le (1-t)F(x_1) + t\,F(x_2), \qquad t \in [0,1]$$

F is proper if $\mathrm{dom}(F) \equiv \{x \in X : F(x) < \infty\} \neq \emptyset$

Information (KL) divergence (relative entropy)

$$I(y, y_0) = \int_A \ln \frac{y}{y_0}\, y(a)da = \int_A \ln \frac{d\mu}{d\nu}\, d\mu(a)$$

Here $y(a) = \mu(da)/da$ is the density function.

Convex functions

Proposition

If $F : X \to \mathbb{R}_\infty$ is proper, convex and l.s.c. on Banach X, then F is continuous on $\mathrm{Int}(\mathrm{dom}(F))$.(e.g. Showalter, 1997)

Convex functions

Proposition

If $F : X \to \mathbb{R}_\infty$ is proper, convex and l.s.c. on Banach X, then F is continuous on $\mathrm{Int}(\mathrm{dom}(F))$.(e.g. Showalter, 1997)

Proposition

Strictly convex F has unique minimum on $\mathrm{dom}(F)$.

Proof.

If not, then $F((1-t)x_1 + tx_2) < (1-t)F(x_1) + tF(x_2) = C$ □

Convex functions

Proposition

If $F : X \to \mathbb{R}_\infty$ is proper, convex and l.s.c. on Banach X, then F is continuous on $\mathrm{Int}(\mathrm{dom}(F))$.(e.g. Showalter, 1997)

Proposition

Strictly convex F has unique minimum on $\mathrm{dom}(F)$.

Proof.

If not, then $F((1-t)x_1 + tx_2) < (1-t)F(x_1) + tF(x_2) = C$ □

Definition

Subdifferential $\partial F(x) \equiv \{y \in Y : (y, x' - x) \leq F(x') - F(x)\}$

Support and distance of a convex body

K is a convex body if $x_1, x_2 \in K$ implies $(1-t)x_1 + tx_2 \in K$

Support and distance of a convex body

K is a convex body if $x_1, x_2 \in K$ implies $(1-t)x_1 + tx_2 \in K$

Definition

Support of K_x is $F(y) = \sup\{(y, x) : x \in K_x\}$

$F(y)$ is convex and 1st order homogeneous $F(tx) = tF(x)$

Support and distance of a convex body

K is a convex body if $x_1, x_2 \in K$ implies $(1-t)x_1 + tx_2 \in K$

Definition

Support of K_x is $F(y) = \sup\{(y,x) : x \in K_x\}$

$F(y)$ is convex and 1st order homogeneous $F(tx) = tF(x)$

Definition

Distance in K_x is $\tilde{F}(x) = \inf\{D : x \in DK_x\}$

Support and distance of a convex body

K is a convex body if $x_1, x_2 \in K$ implies $(1-t)x_1 + tx_2 \in K$

Definition

Support of K_x is $F(y) = \sup\{(y,x) : x \in K_x\}$

$F(y)$ is convex and 1st order homogeneous $F(tx) = tF(x)$

Definition

Distance in K_x is $\tilde{F}(x) = \inf\{D : x \in DK_x\}$

Definition

K_x and K_y are polar if support of K_x is distance in K_y.

$$\begin{aligned} F(y) &= \sup\{(y,x) : x \in K_x\} = \inf\{C : y \in CK_y\} \\ \tilde{F}(x) &= \sup\{(y,x) : y \in K_y\} = \inf\{D : x \in DK_x\} \end{aligned}$$

Functions of polar sets

- Replace conditions $x \in K_x$, $y \in K_y$ by $\tilde{F}(x) \leq D < \infty$, $F(y) \leq C < \infty$

$$F(y) = \sup\{(y,x) : \tilde{F}(x) \leq D\}, \quad \tilde{F}(x) = \sup\{(y,x) : F(y) \leq C\}$$

Functions of polar sets

- Replace conditions $x \in K_x$, $y \in K_y$ by $\tilde{F}(x) \leq D < \infty$, $F(y) \leq C < \infty$

  $$F(y) = \sup\{(y,x) : \tilde{F}(x) \leq D\}, \quad \tilde{F}(x) = \sup\{(y,x) : F(y) \leq C\}$$

- Dual problems

  $$D(C) = \inf\{\tilde{F}(x) : (y,x) \geq C\}, \quad C(D) = \inf\{F(y) : (x,y) \geq D\}$$

Functions of polar sets

- Replace conditions $x \in K_x$, $y \in K_y$ by $\tilde{F}(x) \le D < \infty$, $F(y) \le C < \infty$

  $$F(y) = \sup\{(y,x) : \tilde{F}(x) \le D\}, \quad \tilde{F}(x) = \sup\{(y,x) : F(y) \le C\}$$

- Dual problems

  $$D(C) = \inf\{\tilde{F}(x) : (y,x) \ge C\}, \quad C(D) = \inf\{F(y) : (x,y) \ge D\}$$

- Legendre–Fenchel transform

  $$F(y) = \sup\{(y,x) - \tilde{F}(x)\}, \qquad \tilde{F}(x) = \sup\{(x,y) - F(y)\}$$
  $$x \in \partial F(y), \qquad y \in \partial \tilde{F}(x)$$

  In fact, $\partial F = (\partial \tilde{F})^{-1}$

Necessary conditions of extrema

Theorem

The extremum $y^* \in K_y$ *for* $\sup\{(x, y) : F(y) \leq C\} = D$ *satisfies*

$$\beta x \in \partial F(y^*), \quad F(y^*) = C, \quad \beta^{-1}(C) = D'(C)$$

Necessary conditions of extrema

Theorem

The extremum $y^* \in K_y$ *for* $\sup\{(x,y) : F(y) \le C\} = D$ *satisfies*

$$\beta x \in \partial F(y^*), \quad F(y^*) = C, \quad \beta^{-1}(C) = D'(C)$$

Proof.

Lagrangian function $L(y, \beta^{-1}) = (x,y) + \beta^{-1}[C - F(y)]$

$$\partial_y L(y, \beta^{-1}) = x - \beta^{-1}\partial F(y) \ni 0, \quad \Rightarrow \quad \beta x \in \partial F(y^*)$$

$$\partial_{\beta^{-1}} L(y, \beta^{-1}) = C - F(y) \ni 0, \quad \Rightarrow \quad F(y^*) = C$$

$$D'(C) = \beta^{-1} \qquad (C'(D) = \beta)$$

Necessary conditions of extrema

Theorem

The extremum $y^* \in K_y$ *for* $\sup\{(x,y) : F(y) \le C\} = D$ *satisfies*

$$\beta x \in \partial F(y^*), \quad F(y^*) = C, \quad \beta^{-1}(C) = D'(C)$$

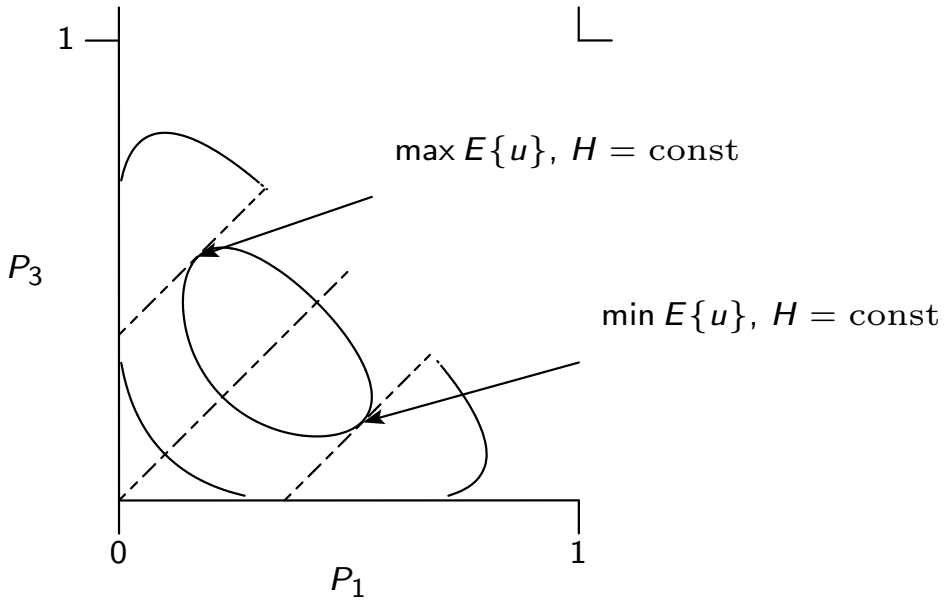
When $y \in Y$ are probabilities, and $F(y)$ is information divergence

$$\beta x - \gamma(\beta) = \ln \frac{y^*}{y_0} \quad \Longrightarrow \quad y^* = y_0 e^{\beta x - \gamma(\beta)}$$

If $x = -|t|^2$, then y^* is Gaussian.

Why 'extrema'?

There is one extremum of $F(y)$ (minimum), but there are two extrema of (x, y): maximum ($\beta \geq 0$) and minimum ($\beta < 0$).



Characterisation of the solution at ∞

Consider $a^* \in da \neq 0$ such that $x(a^*) = \pm\infty$

Characterisation of the solution at ∞

Consider $a^* \in da \neq 0$ such that $x(a^*) = \pm\infty$

Theorem

Linear fnl. $(x, y) \neq \pm\infty$ *iff* $|(x, y)| < \infty$ *for* $|x(a)|$, $|y(a)| < \infty$ *and*

$$x(a^*) = \pm\infty \implies y(a^*) = 0 \qquad x(a^*) = 0 \impliedby y(a^*) = \pm\infty$$

Characterisation of the solution at ∞

Consider $a^* \in da \neq 0$ such that $x(a^*) = \pm\infty$

Theorem

Linear fnl. $(x, y) \neq \pm\infty$ *iff* $|(x, y)| < \infty$ *for* $|x(a)|$, $|y(a)| < \infty$ *and*

$$x(a^*) = \pm\infty \implies y(a^*) = 0 \qquad x(a^*) = 0 \impliedby y(a^*) = \pm\infty$$

Proof.

($\Rightarrow$) By convention $\pm\infty \cdot 0 = 0$ we have

$$(x, y) = (x, y)|_{x,y\neq\pm\infty} + (x, 0)|_{x=\pm\infty} + (0, y)|_{y=\pm\infty} = D + 0 = D \neq \infty$$

($\Leftarrow$) By convention $\infty - \infty = -\infty$

$$(x, y) = (x, y)|_{x\neq\pm\infty} + (x, y)|_{x=\pm\infty} = D \mp \infty = \mp\infty$$

Characterisation of the solution at 0

Theorem

Let $\tilde{F}(x)$, $F(y)$ *be dual convex fnls, and* $\beta x = \partial F(y^*)$. *Then* $\forall x \in X$ *s.t.* $(x, y) \neq \text{const}$, $\partial \tilde{F}(x_0) \ni 0$

$$y^*(a) = 0 \implies \beta x(a) = x_0(a)$$

Characterisation of the solution at 0

Theorem

Let $\tilde{F}(x)$, $F(y)$ *be dual convex fnls, and* $\beta x = \partial F(y^*)$. *Then* $\forall x \in X$ *s.t.* $(x, y) \neq \text{const}$, $\partial \tilde{F}(x_0) \ni 0$

$$y^*(a) = 0 \implies \beta x(a) = x_0(a)$$

Proof.

$$\partial F(y^*) \ni \beta x \;\Rightarrow\; y^*(a) = 0 = \partial \tilde{F}(\beta x(a)) \in \partial \tilde{F}(x_0(a))$$ □

Characterisation of the solution at 0

Theorem

Let $\tilde{F}(x)$, $F(y)$ be dual convex fnls, and $\beta x = \partial F(y^*)$. Then $\forall x \in X$ s.t. $(x, y) \neq \text{const}$, $\partial \tilde{F}(x_0) \ni 0$

$$y^*(a) = 0 \implies \beta x(a) = x_0(a)$$

If $F(y)$ is information divergence, then its dual is

$$\tilde{F}(\beta x) = \ln \int_A y_0 e^{\beta x(a)} da$$

which achieves minimum at $x_0 = -\infty$.
Thus, $y^*(a_0) = 0$ implies $x(a_0) = -\infty$

Generalised characteristic potentials

Definition

Given $x \in X$ and $y^* = \partial\tilde{F}(\beta x)$ for constraints $(x, y^*) = D$ (or $F(y^*) = C$), let

$$\Psi(\beta) \equiv \sup[\beta D - C(D)] \qquad \Phi(\beta^{-1}) \equiv \inf[\beta^{-1} C - D(C)]$$

Generalised characteristic potentials

Definition

Given $x \in X$ and $y^* = \partial\tilde{F}(\beta x)$ for constraints $(x, y^*) = D$ (or $F(y^*) = C$), let

$$\Psi(\beta) \equiv \sup[\beta D - C(D)] \qquad \Phi(\beta^{-1}) \equiv \inf[\beta^{-1} C - D(C)]$$

- One can check that $\Phi(\beta^{-1}) = -\beta^{-1}\Psi(\beta)$

Generalised characteristic potentials

Definition

Given $x \in X$ and $y^* = \partial\tilde{F}(\beta x)$ for constraints $(x, y^*) = D$ (or $F(y^*) = C$), let

$$\Psi(\beta) \equiv \sup[\beta D - C(D)] \qquad \Phi(\beta^{-1}) \equiv \inf[\beta^{-1} C - D(C)]$$

- One can check that $\Phi(\beta^{-1}) = -\beta^{-1}\Psi(\beta)$
- Thus, $\Phi(\beta^{-1})$ is both convex ($\beta < 0$) and concave ($\beta \geq 0$)

Generalised characteristic potentials

Definition

Given $x \in X$ and $y^* = \partial\tilde{F}(\beta x)$ for constraints $(x, y^*) = D$ (or $F(y^*) = C$), let

$$\Psi(\beta) \equiv \sup[\beta D - C(D)] \qquad \Phi(\beta^{-1}) \equiv \inf[\beta^{-1} C - D(C)]$$

- One can check that $\Phi(\beta^{-1}) = -\beta^{-1}\Psi(\beta)$
- Thus, $\Phi(\beta^{-1})$ is both convex ($\beta < 0$) and concave ($\beta \geq 0$)
-

$$\begin{aligned} C(D) &= \sup[\beta D - \Psi(\beta)], & D(\beta) &= \Psi'(\beta) \\ D(C) &= \inf[\beta^{-1} C - \Phi(\beta^{-1})], & C(\beta^{-1}) &= \Phi'(\beta^{-1}) \end{aligned}$$

Examples

Let $F(y) = I(y, da) = \int_A \ln \frac{y}{da} y(a) da$, where da is Lebesgue measure. Thus, $F(y)$ is minus entropy, and

$$\Psi(\beta) = \ln \int_A e^{\beta x(a)} da = \ln Z(\beta) = \Gamma(\beta), \quad \Phi(\beta^{-1}) = -\beta^{-1}\Gamma(\beta)$$

Examples

Let $F(y) = I(y, da) = \int_A \ln \frac{y}{da} y(a) da$, where da is Lebesgue measure. Thus, $F(y)$ is minus entropy, and

$$\Psi(\beta) = \ln \int_A e^{\beta x(a)} da = \ln Z(\beta) = \Gamma(\beta), \quad \Phi(\beta^{-1}) = -\beta^{-1}\Gamma(\beta)$$

Example (Finite set)

Let $A = \{a_1, a_2\}$, $da = 1$, and $x : A \to \{c - d, c + d\}$.

$$\begin{aligned} Z(\beta) &= e^{\beta(c-d)} + e^{\beta(c+d)} = 2e^{\beta c} \cosh(\beta d) \\ \Gamma(\beta) &= \ln 2 + \beta c + \ln \cosh(\beta d) \\ D(\beta) &= \Gamma'(\beta) = c + d \tanh(\beta d) \end{aligned}$$

Examples

Let $F(y) = I(y, da) = \int_A \ln \frac{y}{da} y(a) da$, where da is Lebesgue measure. Thus, $F(y)$ is minus entropy, and

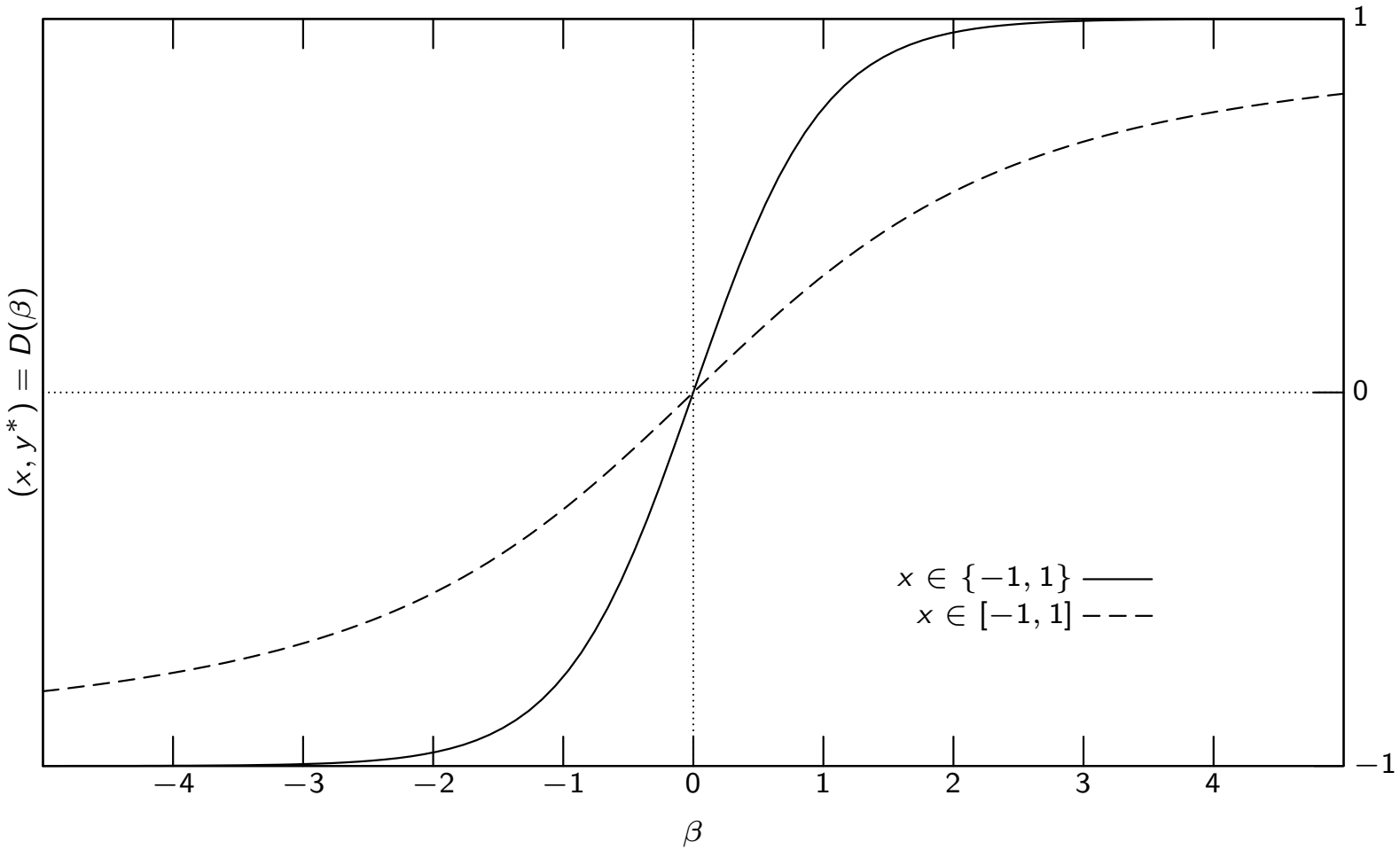
$$\Psi(\beta) = \ln \int_A e^{\beta x(a)} da = \ln Z(\beta) = \Gamma(\beta), \quad \Phi(\beta^{-1}) = -\beta^{-1}\Gamma(\beta)$$

Example (Uncountable set)

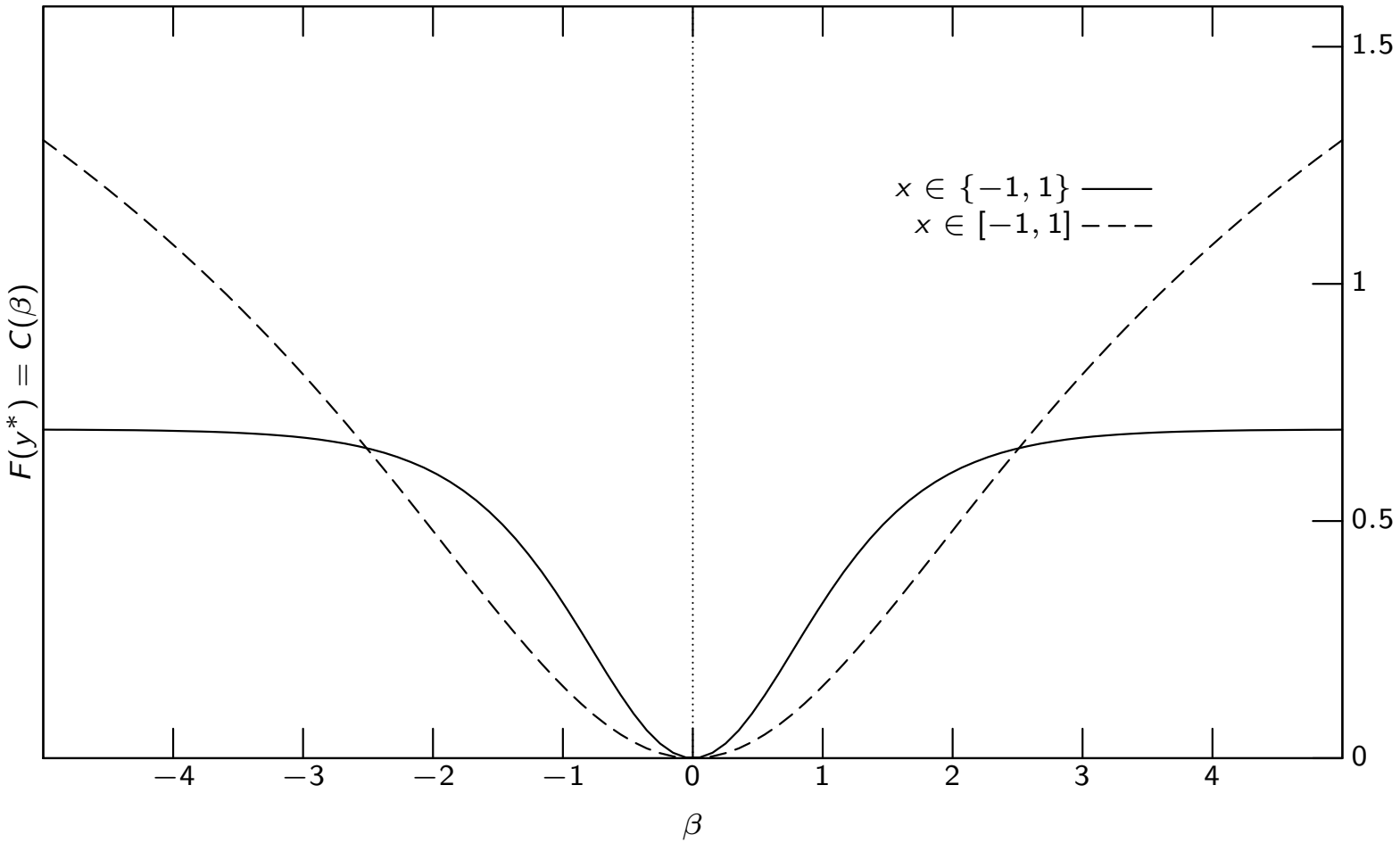
Let A be a continuum, $da = \mathrm{const}$, $x : A \to [c-d, c+d] \subseteq \overline{\mathbb{R}}$.

$$\begin{aligned} Z(\beta) &= \int_{c-d}^{c+d} e^{\beta x(a)} da = 2e^{\beta c} \beta^{-1} \sinh(\beta d) \\ \Gamma(\beta) &= \beta c - \ln|\beta| d + \ln|\sinh(\beta d)| \\ D(\beta) &= c - \beta^{-1} + d \coth(\beta d) \end{aligned}$$

Expected utility



Information



Dynamics

Theorem

Let $C = C(t)$, $D = D(t)$ be monotone functions describing constraints $F(y) \leq C$ or $(x, y) \geq D$ on interval $t \in [t_1, t_2]$.

Dynamics

Theorem

Let $C = C(t)$, $D = D(t)$ be monotone functions describing constraints $F(y) \leq C$ or $(x, y) \geq D$ on interval $t \in [t_1, t_2]$. Then

$$\int_{t_1}^{t_2} (x, y(t))\, dt \quad \leq \quad \Psi(\beta_2) - \Psi(\beta_1)$$

Dynamics

Theorem

Let $C = C(t)$, $D = D(t)$ be monotone functions describing constraints $F(y) \leq C$ or $(x, y) \geq D$ on interval $t \in [t_1, t_2]$. Then

$$\begin{aligned} \int_{t_1}^{t_2} (x, y(t))\, dt &\leq \Psi(\beta_2) - \Psi(\beta_1) \\ \int_{t_1}^{t_2} F(y(t))\, dt &\geq \Phi(\beta_1^{-1}) - \Phi(\beta_2^{-1}) \end{aligned}$$

where β_1, β_2 are determined from

$$\beta^{-1} = (\Phi')^{-1}(C), \qquad \beta = (\Psi')^{-1}(D)$$

Dynamics

Proof.

An integral is maxi-(mini)-mised if each integrand is maxi-(mini)-mised:

$$\int_{t_1}^{t_2} (x, y(t))\, dt \quad \leq \quad \int_{\beta_1}^{\beta_2} D(\beta) d\beta = \int_{\beta_1}^{\beta_2} \Psi'(\beta) d\beta = \Psi|_{\beta_1}^{\beta_2}$$

Dynamics

Proof.

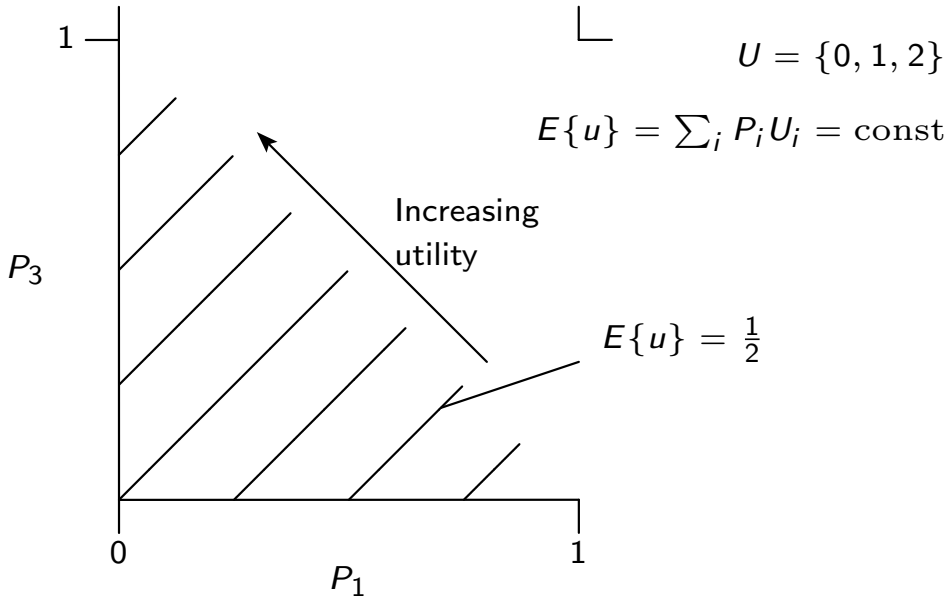
An integral is maxi-(mini)-mised if each integrand is maxi-(mini)-mised:

$$\int_{t_1}^{t_2}(x,y(t))\,dt \leq \int_{\beta_1}^{\beta_2} D(\beta)d\beta = \int_{\beta_1}^{\beta_2}\Psi'(\beta)d\beta = \Psi|_{\beta_1}^{\beta_2}$$

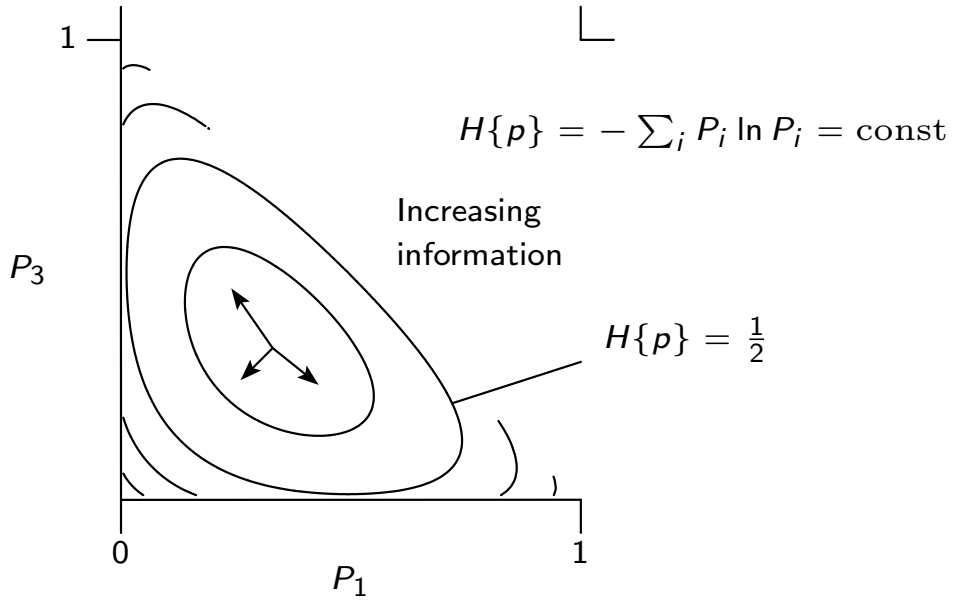
$$\int_{t_1}^{t_2} F(y(t))\,dt \geq \int_{\beta_2^{-1}}^{\beta_1^{-1}} C(\beta^{-1})d\beta^{-1} = \Phi(\beta^{-1})\Big|_{\beta_2^{-1}}^{\beta_1^{-1}}$$

□

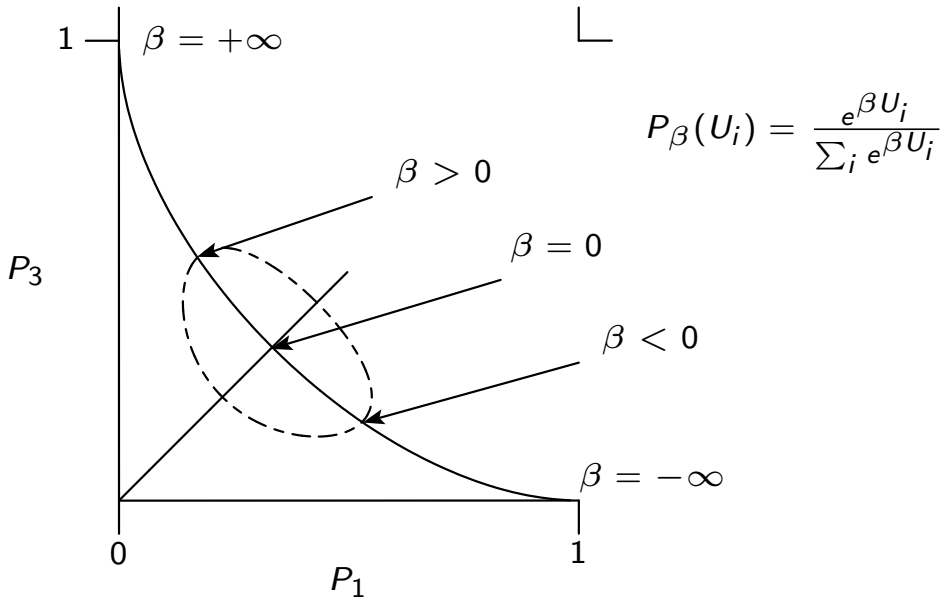
Expected utility



Information divergence



Information brachistochrone



Why exponential and logarithmic functions?

- Utility functions and probabilities do not form a linear spaces in standard algebra $(+, \times)$ (e.g. multiply by negative scalar).

Why exponential and logarithmic functions?

- Utility functions and probabilities do not form a linear spaces in standard algebra $(+, \times)$ (e.g. multiply by negative scalar).
- They do form linear spaces in non–standard algebras over semi–rings $(A, +, \times)$ with idempotent $+$:

	A	$+$	$\times$	0	1
$\mathbb{R}_{\max}$	$\mathbb{R} \cup -\infty$	$\max$	$+$	$-\infty$	0
$\mathbb{R}_{+}$	$\mathbb{R}_{+}$	$\max$	$\times$	0	1

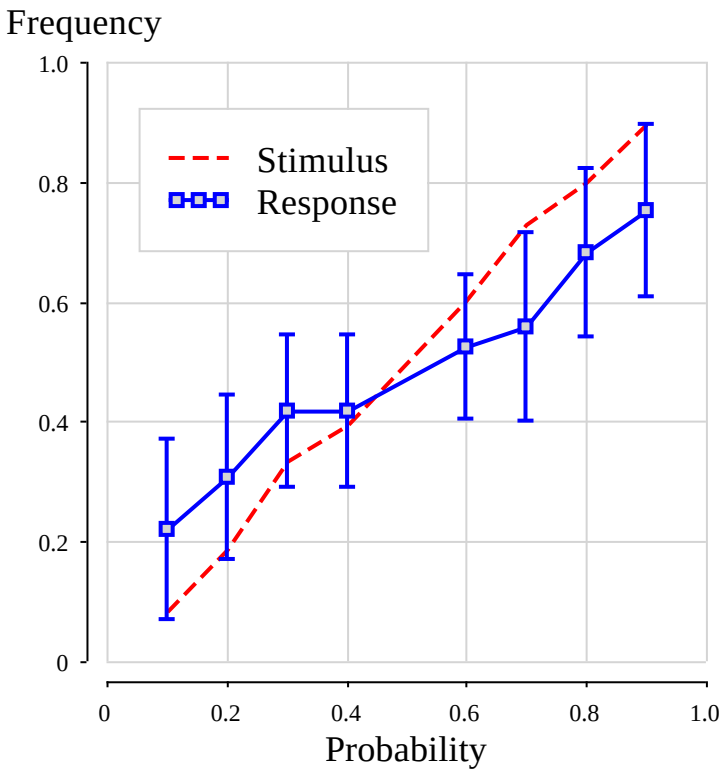
Why exponential and logarithmic functions?

- Utility functions and probabilities do not form a linear spaces in standard algebra $(+, \times)$ (e.g. multiply by negative scalar).
- They do form linear spaces in non–standard algebras over semi–rings $(A, +, \times)$ with idempotent $+$:

	A	$+$	$\times$	0	1
$\mathbb{R}_{\max}$	$\mathbb{R} \cup -\infty$	max	$+$	$-\infty$	0
$\mathbb{R}_{+}$	$\mathbb{R}_{+}$	max	$\times$	0	1

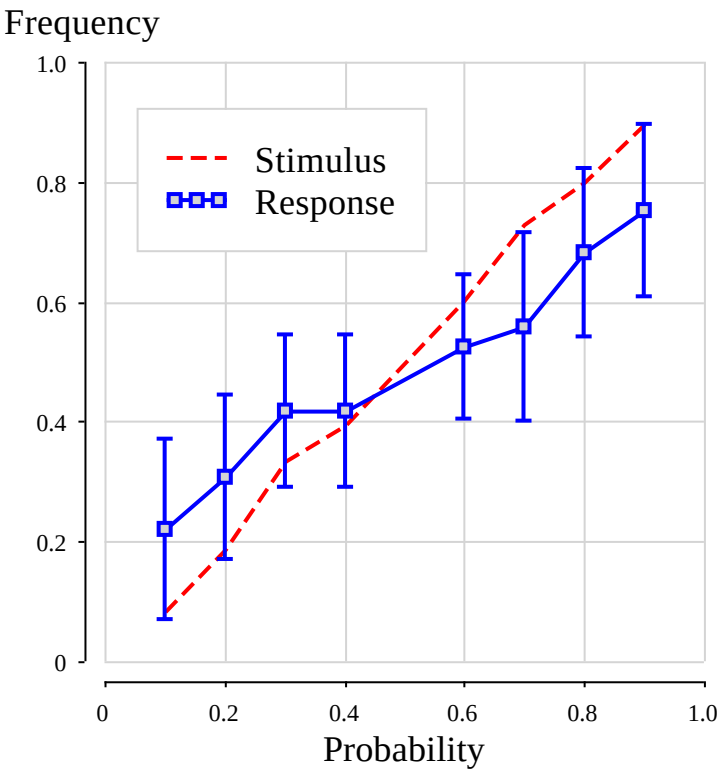
- The exponential and logarithmic functions are isomorphisms between $\mathbb{R}_{\max}$ and $\mathbb{R}_{+}$ is $x \mapsto e^x$ $(y \mapsto \ln y)$.

Probability matching

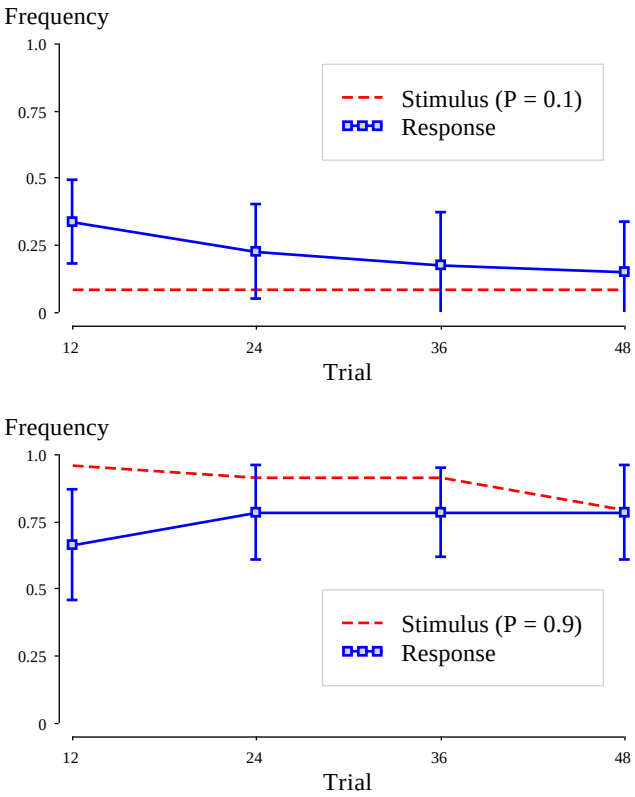


(Friedman et al., 1964)

Probability matching

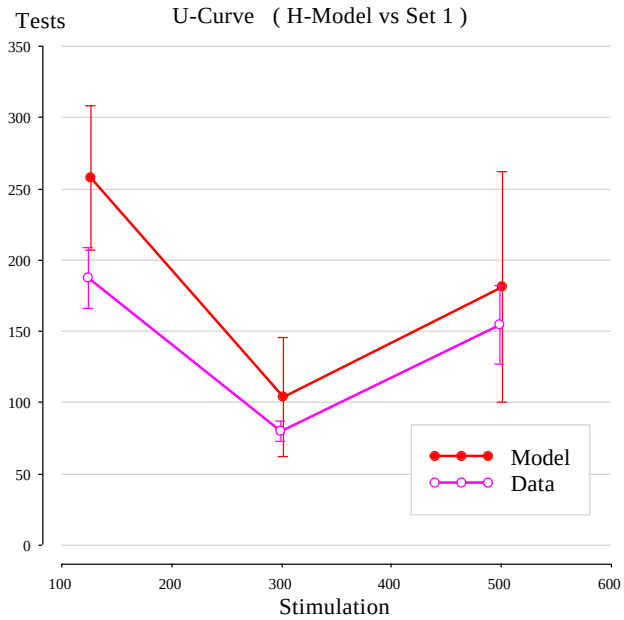


(Friedman et al., 1964)

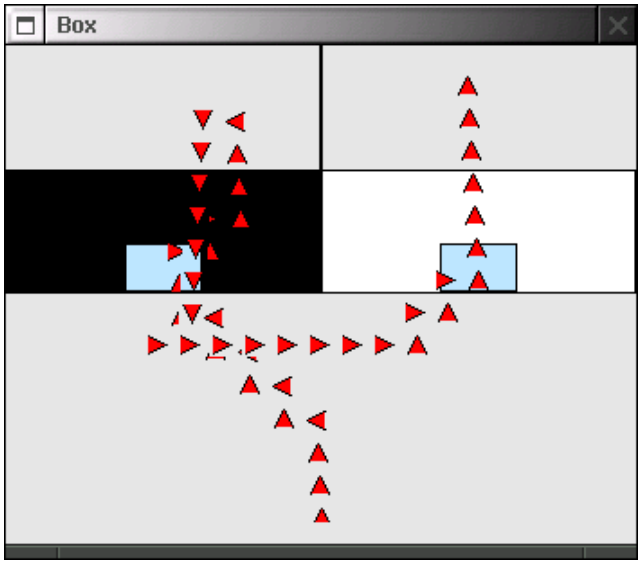


(Belavkin, 2001)

The inverted–U effect

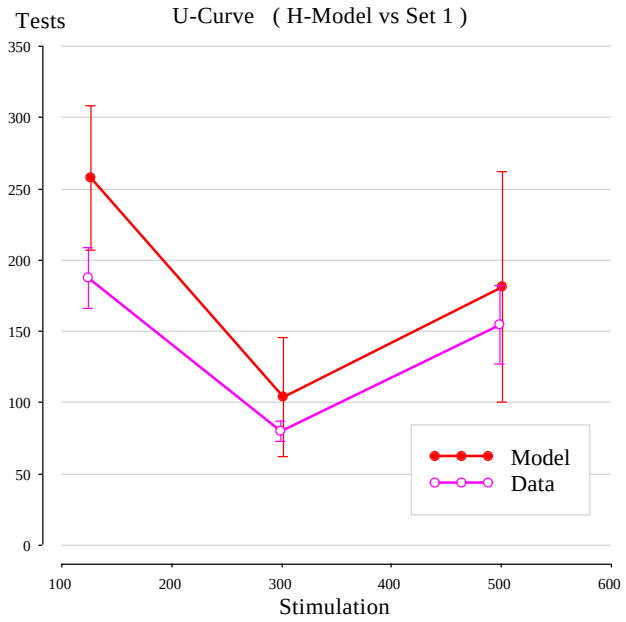


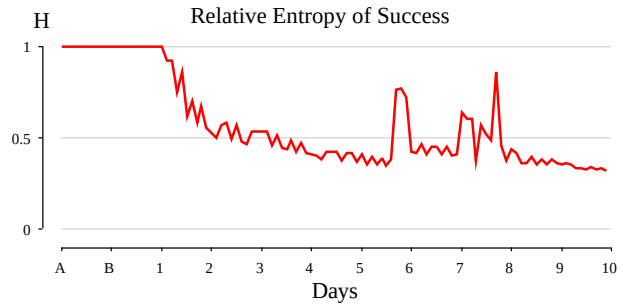
(Yerkes & Dodson, 1908)



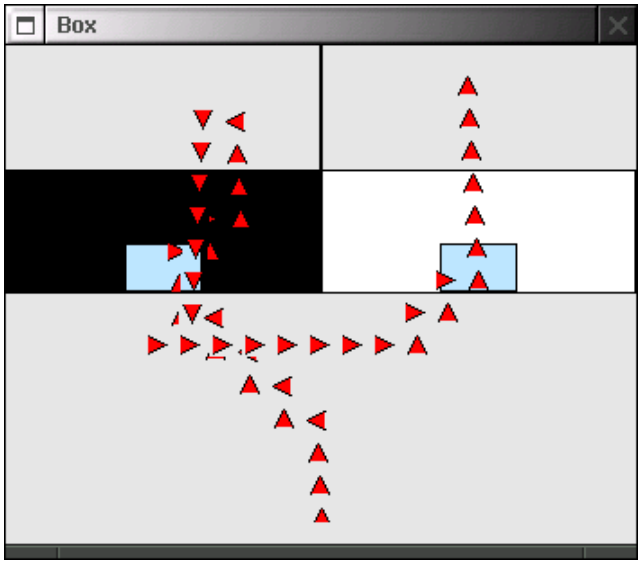
(Belavkin & Ritter, 2003, 2004)

The inverted–U effect



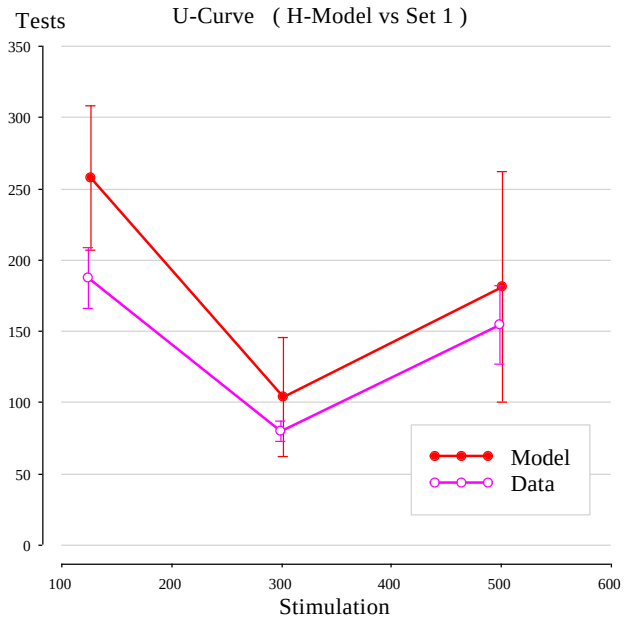


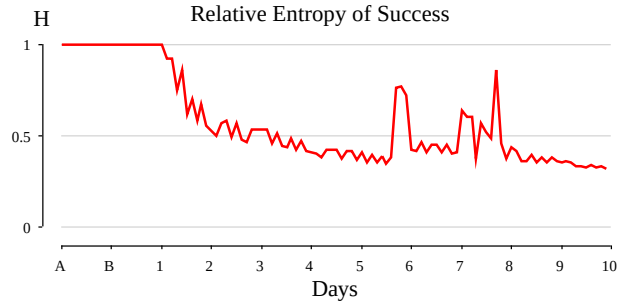
(Yerkes & Dodson, 1908)



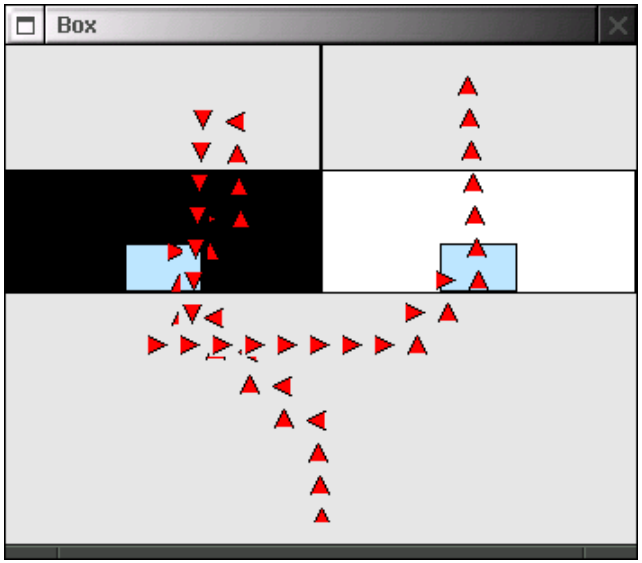
(Belavkin & Ritter, 2003, 2004)

The inverted–U effect

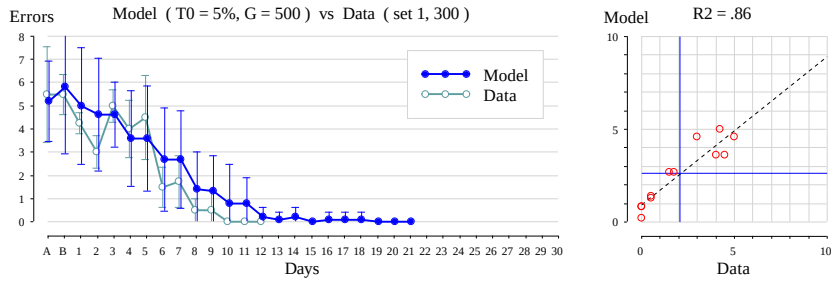




(Yerkes & Dodson, 1908)

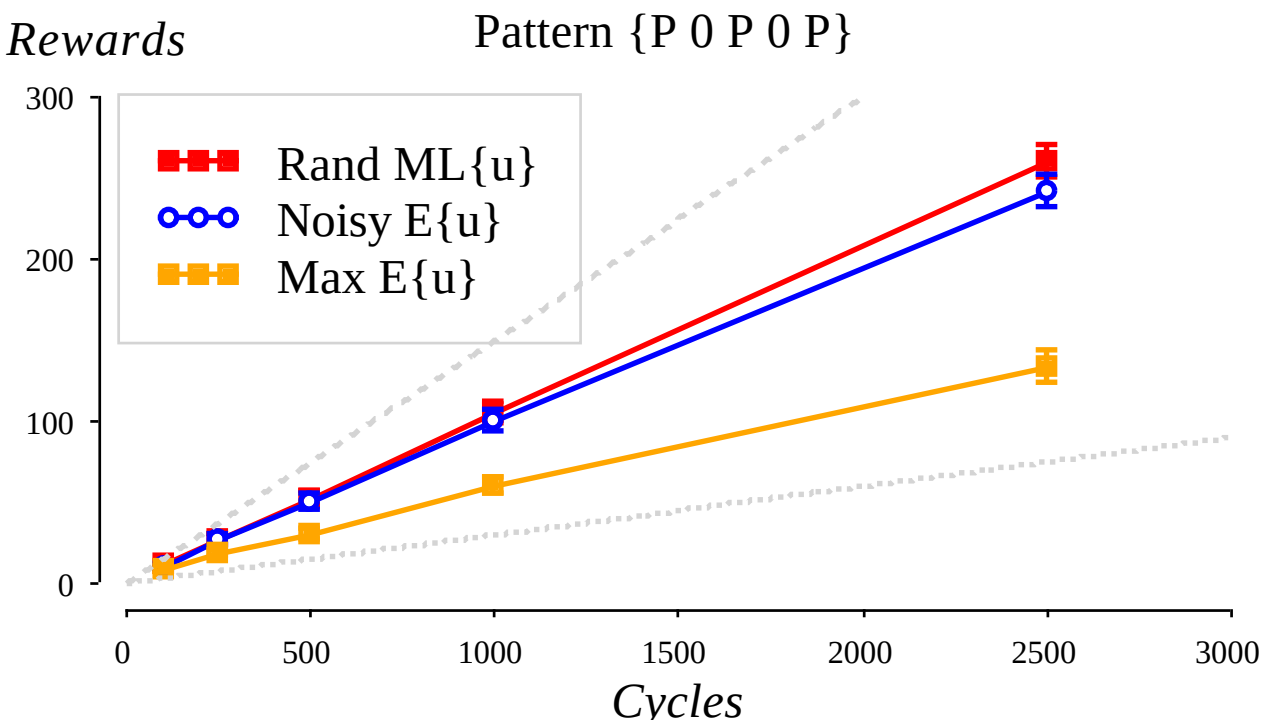


(Belavkin & Ritter, 2003, 2004)



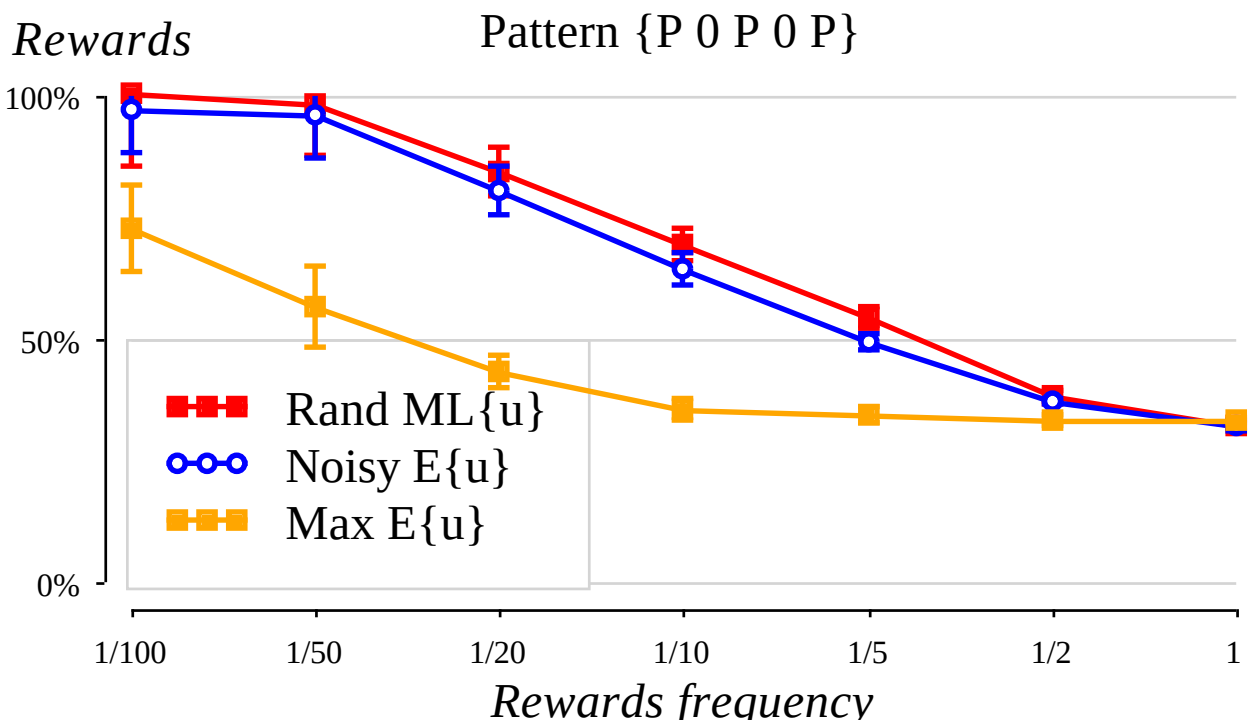
Optimal action selection

(Belavkin, 2005)



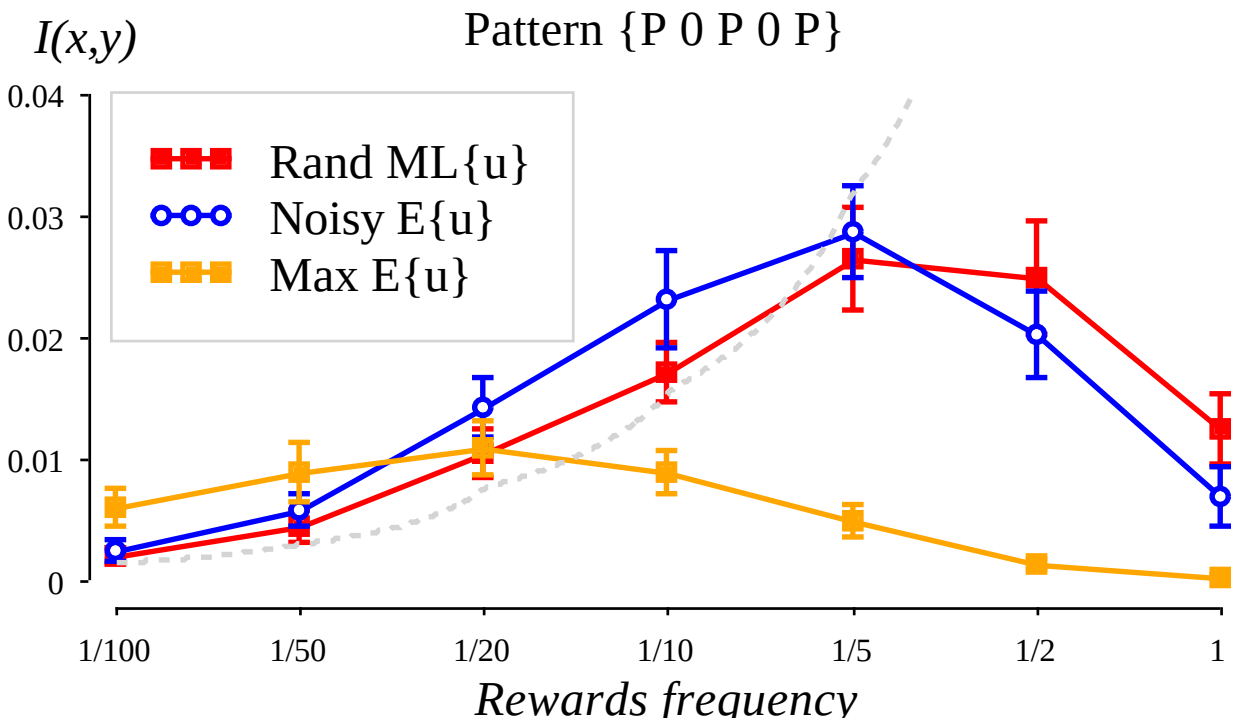
Optimal action selection

(Belavkin, 2005)



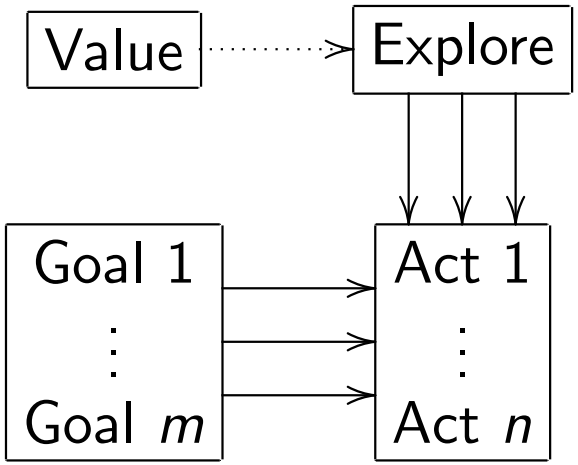
Optimal action selection

(Belavkin, 2005)



Learning in neural cell assemblies

(Belavkin & Huyck, 2008)



Conclusions

Questions?

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *Annals of Mathematical Statistics*, *34*, 199–205.

Belavkin, R. V. (2001, March). The role of emotion in problem solving. In C. Johnson (Ed.), *Proceedings of the AISB'01 Symposium on Emotion, Cognition and Affective Computing* (pp. 49–57). Heslington, York, England: AISB. (ISBN 1-902956-19-7)

Belavkin, R. V. (2005, December). Acting irrationally to improve performance in stochastic worlds. In M. Bramer, F. Coenen, & T. Allen (Eds.), *Proceedings of AI–2005, the 25th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence* (Vol. XXII, pp. 305–316). Cambridge: Springer. (ISBN 1-84628-225-X)

Belavkin, R. V., & Huyck, C. (2008). Emergence of rules in cell assemblies of flif neurons. In *The 18th European Conference*

on Artificial Intelligence.

Belavkin, R. V., & Ritter, F. E. (2003, April). The use of entropy for analysis and control of cognitive models. In F. Detje, D. Dörner, & H. Schaub (Eds.), *Proceedings of the Fifth International Conference on Cognitive Modeling* (pp. 21–26). Bamberg, Germany: Universitäts–Verlag Bamberg. (ISBN 3-933463-15-7)

Belavkin, R. V., & Ritter, F. E. (2004). Optimist: A new conflict resolution algorithm for ACT–R. In *Proceedings of the Sixth International Conference on Cognitive Modeling* (pp. 40–45). Mahwah, NJ: Lawrence Erlbaum. (ISBN 0-8058-5426-6)

Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Friedman, M. P., Burke, C. J., Cole, M., Keller, L., Millward, R. B., & Estes, W. K. (1964). Two–choice behaviour under extended training with shifting probabilities of reinforcement.

In R. C. Atkinson (Ed.), *Studies in mathematical psychology* (pp. 250–316). Stanford, CA: Stanford University Press.

Gelfand, I. M., & Fomin, S. V. (1963). *Calculus of variations*. Prentice–Hall. (Translated from Russian)

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106, 108*, 620–630, 171–190.

Kalman, R. E., & Bucy, R. S. (1961, December). New results in linear filtering and prediction theory. *Transactions ASME, Series D–J. Basic Engineering.*, *83*, 94–107.

Kolmogorov, A. N. (1957). The theory of information transmission. In *Meeting of the USSR Academy of Sciences on Scientific Problems of Production Automatisation, 1956* (pp. 66–99). USSR Academy of Sciences. (In Russian)

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

Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., & Mishchenko, E. F. (1962). *The mathematical theory of optimal processes*. John Wiley and Sons. (Translated from Russian)

Savage, L. (1954). *The foundations of statistics*. New York: John Wiley & Sons.

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

Showalter, R. E. (1997). *Monotone operators in Banach space and nonlinear partial differential equations* (Vol. 49). American Mathematical Society.

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

Stratonovich, R. L. (1960). Conditional Markov processes. *Theory*

of Probability and its Applications (USSR), *5*(2), 156–178.

Stratonovich, R. L. (1965). O tsennosti informatsii [On utility of information]. *Izvestiya of USSR Academy of Sciences, Technical Cybernetics*, *5*, 3–12. (In Russian)

Wald, A. (1950). *Statistical decision functions*. New York: John Wiley & Sons.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit formation. *Journal of Comparative Neurology and Psychology*, *18*, 459–482.