On Global Optimality of Deterministic and Non-Deterministic Transformations

Roman V. Belavkin

Middlesex University

July 4, 2011

Markov Transition Kernels

- A, B two measurable sets (i.e. $(A, \mathcal{A})$, $(B, \mathcal{B})$).

Markov Transition Kernels

- A, B two measurable sets (i.e. $(A, \mathcal{A})$, $(B, \mathcal{B})$).
- $\mathcal{P}(A)$ the set of all probability measures $P(A_i)$ on A.

Markov Transition Kernels

- A, B two measurable sets (i.e. $(A, \mathcal{A})$, $(B, \mathcal{B})$).
- $\mathcal{P}(A)$ the set of all probability measures $P(A_i)$ on A.

Definition (Transition kernel)

A conditional probability $P(A_i \mid b) \in \mathcal{P}(A)$ that is $\mathcal{B}$-measurable $\forall\, A_i \in \mathcal{A}$.

Markov Transition Kernels

- A, B two measurable sets (i.e. $(A, \mathcal{A})$, $(B, \mathcal{B})$).
- $\mathcal{P}(A)$ the set of all probability measures $P(A_i)$ on A.

Definition (Transition kernel)

A conditional probability $P(A_i \mid b) \in \mathcal{P}(A)$ that is $\mathcal{B}$-measurable $\forall\, A_i \in \mathcal{A}$.

Independence

If $P(A_i \mid b) = P(A_i)$ for all $A_i \in \mathcal{A}$, $b \in B$. Thus,

$$P(A_i \cap B_j) = P(A_i)\, P(B_j)$$

Markov Transition Kernels

- A, B two measurable sets (i.e. $(A, \mathcal{A})$, $(B, \mathcal{B})$).
- $\mathcal{P}(A)$ the set of all probability measures $P(A_i)$ on A.

Definition (Transition kernel)

A conditional probability $P(A_i \mid b) \in \mathcal{P}(A)$ that is $\mathcal{B}$-measurable $\forall\, A_i \in \mathcal{A}$.

Independence

If $P(A_i \mid b) = P(A_i)$ for all $A_i \in \mathcal{A}$, $b \in B$. Thus,

$$P(A_i \cap B_j) = P(A_i)\, P(B_j)$$

Deterministic dependency

Represented by a measurable $f : B \to A$ or by $\delta_{f(b)}(A_i)$:

$$P(A_i \mid b) = \delta_{f(b)}(A_i) := \begin{cases} 1 & \text{if } f(b) = a \in A_i \\ 0 & \text{otherwise} \end{cases}$$

Simplex $\mathcal{P}(A \times B)$

- $\mathcal{P}(A \times B)$ the set of all joint probability measures $P(A_i \cap B_j)$.

Simplex $\mathcal{P}(A \times B)$

- $\mathcal{P}(A \times B)$ the set of all joint probability measures $P(A_i \cap B_j)$.
- Each $p \in \mathcal{P}(A \times B)$ uniquely defines $P(A_i \mid b)$ and $P(B_j)$.

Simplex $\mathcal{P}(A \times B)$

- $\mathcal{P}(A \times B)$ the set of all joint probability measures $P(A_i \cap B_j)$.
- Each $p \in \mathcal{P}(A \times B)$ uniquely defines $P(A_i \mid b)$ and $P(B_j)$.
- Recall Bayes formula for $P(B_j) > 0$ and marginalization:

$$P(A_i \mid B_j) = \frac{P(A_i \cap B_j)}{P(B_j)}\,, \qquad P(B_j) = \sum_{a \in A} P(A \cap B_j)$$

Simplex $\mathcal{P}(A \times B)$

- $\mathcal{P}(A \times B)$ the set of all joint probability measures $P(A_i \cap B_j)$.
- Each $p \in \mathcal{P}(A \times B)$ uniquely defines $P(A_i \mid b)$ and $P(B_j)$.
- Recall Bayes formula for $P(B_j) > 0$ and marginalization:

$$P(A_i \mid B_j) = \frac{P(A_i \cap B_j)}{P(B_j)}\,, \qquad P(B_j) = \sum_{a \in A} P(A \cap B_j)$$

- All measurable $f : B \to A$ correspond to $p_f \in \mathcal{P}(A \times B)$.

Simplex $\mathcal{P}(A \times B)$

- $\mathcal{P}(A \times B)$ the set of all joint probability measures $P(A_i \cap B_j)$.
- Each $p \in \mathcal{P}(A \times B)$ uniquely defines $P(A_i \mid b)$ and $P(B_j)$.
- Recall Bayes formula for $P(B_j) > 0$ and marginalization:

$$P(A_i \mid B_j) = \frac{P(A_i \cap B_j)}{P(B_j)}, \qquad P(B_j) = \sum_{a \in A} P(A \cap B_j)$$

- All measurable $f : B \to A$ correspond to $p_f \in \mathcal{P}(A \times B)$.

Remark (Interior of $\mathcal{P}(A \times B)$)

$P_f(A_i \cap B_j) = \delta_{f(b)}(A_i) P(B_j) = 0$ if $f(b) \notin A_i$.

- *Thus, $p_f \in \partial\mathcal{P}(A \times B)$.*

Simplex $\mathcal{P}(A \times B)$

- $\mathcal{P}(A \times B)$ the set of all joint probability measures $P(A_i \cap B_j)$.
- Each $p \in \mathcal{P}(A \times B)$ uniquely defines $P(A_i \mid b)$ and $P(B_j)$.
- Recall Bayes formula for $P(B_j) > 0$ and marginalization:

$$P(A_i \mid B_j) = \frac{P(A_i \cap B_j)}{P(B_j)}, \qquad P(B_j) = \sum_{a \in A} P(A \cap B_j)$$

- All measurable $f : B \to A$ correspond to $p_f \in \mathcal{P}(A \times B)$.

Remark (Interior of $\mathcal{P}(A \times B)$)

$P_f(A_i \cap B_j) = \delta_{f(b)}(A_i) P(B_j) = 0$ *if* $f(b) \notin A_i$.

- *Thus,* $p_f \in \partial \mathcal{P}(A \times B)$.
- $p \in \mathrm{Int}(\mathcal{P}(A \times B))$ *implies* p *is non-deterministic.*

Composite Systems in Applications

- In communication and cryptography, A and B are Alice and Bob.

Composite Systems in Applications

- In communication and cryptography, A and B are Alice and Bob.
- In estimation and control, A — outputs (actions); B — inputs.

Composite Systems in Applications

- In communication and cryptography, A and B are Alice and Bob.
- In estimation and control, A — outputs (actions); B — inputs.
- In computation, A — answers; B — problems.

Composite Systems in Applications

- In communication and cryptography, A and B are Alice and Bob.
- In estimation and control, A — outputs (actions); B — inputs.
- In computation, A — answers; B — problems.

Example (Deterministic and Non-deterministic algorithms)

- Words $w \in \{1, \dots, \alpha\}^n$ from finite alphabet $\{1, \dots, \alpha\}$

Composite Systems in Applications

- In communication and cryptography, A and B are Alice and Bob.
- In estimation and control, A — outputs (actions); B — inputs.
- In computation, A — answers; B — problems.

Example (Deterministic and Non-deterministic algorithms)

- Words $w \in \{1, \dots, \alpha\}^n$ from finite alphabet $\{1, \dots, \alpha\}$
- Γ — a system of computations transforming initial w_0 to final w_t.

Composite Systems in Applications

- In communication and cryptography, A and B are Alice and Bob.
- In estimation and control, A — outputs (actions); B — inputs.
- In computation, A — answers; B — problems.

Example (Deterministic and Non-deterministic algorithms)

- Words $w \in \{1, \dots, \alpha\}^n$ from finite alphabet $\{1, \dots, \alpha\}$
- Γ — a system of computations transforming initial w_0 to final w_t.
- $\gamma(w_t) = w_{t+1}$ is *direct processing* operator (Kolmogorov & Uspenskii, 1958).

Composite Systems in Applications

- In communication and cryptography, A and B are Alice and Bob.
- In estimation and control, A — outputs (actions); B — inputs.
- In computation, A — answers; B — problems.

Example (Deterministic and Non-deterministic algorithms)

- Words $w \in \{1, \dots, \alpha\}^n$ from finite alphabet $\{1, \dots, \alpha\}$
- Γ — a system of computations transforming initial w_0 to final w_t.
- $\gamma(w_t) = w_{t+1}$ is *direct processing* operator (Kolmogorov & Uspenskii, 1958).
- An algorithm can

Composite Systems in Applications

- In communication and cryptography, A and B are Alice and Bob.
- In estimation and control, A — outputs (actions); B — inputs.
- In computation, A — answers; B — problems.

Example (Deterministic and Non-deterministic algorithms)

- Words $w \in \{1, \dots, \alpha\}^n$ from finite alphabet $\{1, \dots, \alpha\}$
- Γ — a system of computations transforming initial w_0 to final w_t.
- $\gamma(w_t) = w_{t+1}$ is *direct processing* operator (Kolmogorov & Uspenskii, 1958).
- An algorithm can
  1. Terminate $\Gamma(w_0) = (w_1, \dots, w_t)$ with final w_t (answer).

Composite Systems in Applications

- In communication and cryptography, A and B are Alice and Bob.
- In estimation and control, A — outputs (actions); B — inputs.
- In computation, A — answers; B — problems.

Example (Deterministic and Non-deterministic algorithms)

- Words $w \in \{1, \dots, \alpha\}^n$ from finite alphabet $\{1, \dots, \alpha\}$
- Γ — a system of computations transforming initial w_0 to final w_t.
- $\gamma(w_t) = w_{t+1}$ is *direct processing* operator (Kolmogorov & Uspenskii, 1958).
- An algorithm can
  1. Terminate $\Gamma(w_0) = (w_1, \dots, w_t)$ with final w_t (answer).
  2. Terminate $\Gamma(w_0) = (w_1, \dots, w_t)$ with non-final w_t (no answer).

Composite Systems in Applications

- In communication and cryptography, A and B are Alice and Bob.
- In estimation and control, A — outputs (actions); B — inputs.
- In computation, A — answers; B — problems.

Example (Deterministic and Non-deterministic algorithms)

- Words $w \in \{1, \dots, \alpha\}^n$ from finite alphabet $\{1, \dots, \alpha\}$
- Γ — a system of computations transforming initial w_0 to final w_t.
- $\gamma(w_t) = w_{t+1}$ is *direct processing* operator (Kolmogorov & Uspenskii, 1958).
- An algorithm can
  1. Terminate $\Gamma(w_0) = (w_1, \dots, w_t)$ with final w_t (answer).
  2. Terminate $\Gamma(w_0) = (w_1, \dots, w_t)$ with non-final w_t (no answer).
  3. Not terminate $\Gamma(w_0) = (w_1, \dots, w_t, \dots)$.

Composite Systems in Applications

- In communication and cryptography, A and B are Alice and Bob.
- In estimation and control, A — outputs (actions); B — inputs.
- In computation, A — answers; B — problems.

Example (Deterministic and Non-deterministic algorithms)

- Words $w \in \{1, \dots, \alpha\}^n$ from finite alphabet $\{1, \dots, \alpha\}$
- Γ — a system of computations transforming initial w_0 to final w_t.
- $\gamma(w_t) = w_{t+1}$ is *direct processing* operator (Kolmogorov & Uspenskii, 1958).
- An algorithm can
  1. Terminate $\Gamma(w_0) = (w_1, \dots, w_t)$ with final w_t (answer).
  2. Terminate $\Gamma(w_0) = (w_1, \dots, w_t)$ with non-final w_t (no answer).
  3. Not terminate $\Gamma(w_0) = (w_1, \dots, w_t, \dots)$.
- Let $A := \prod_{t=1}^{\infty}\{w_t\}$ (output sequences), $B := \{w_0\}$ (input words).

Composite Systems in Applications

- In communication and cryptography, A and B are Alice and Bob.
- In estimation and control, A — outputs (actions); B — inputs.
- In computation, A — answers; B — problems.

Example (Deterministic and Non-deterministic algorithms)

- Words $w \in \{1, \dots, \alpha\}^n$ from finite alphabet $\{1, \dots, \alpha\}$
- Γ — a system of computations transforming initial w_0 to final w_t.
- $\gamma(w_t) = w_{t+1}$ is *direct processing* operator (Kolmogorov & Uspenskii, 1958).
- An algorithm can
  1. Terminate $\Gamma(w_0) = (w_1, \dots, w_t)$ with final w_t (answer).
  2. Terminate $\Gamma(w_0) = (w_1, \dots, w_t)$ with non-final w_t (no answer).
  3. Not terminate $\Gamma(w_0) = (w_1, \dots, w_t, \dots)$.
- Let $A := \prod_{t=1}^{\infty} \{w_t\}$ (output sequences), $B := \{w_0\}$ (input words).
- $p \in \mathcal{P}(A \times B)$ represent all algorithms (deterministic or not).

Optimality

- A utility representation of $\lesssim$ on $A \times B$ is $x : A \times B \to \mathbb{R}$:

$$(a_1, b_1) \lesssim (a_2, b_2) \quad \iff \quad x(a_1, b_1) \leq x(a_2, b_2)$$

Optimality

- A utility representation of $\lesssim$ on $A \times B$ is $x : A \times B \to \mathbb{R}$:

$$(a_1, b_1) \lesssim (a_2, b_2) \quad \iff \quad x(a_1, b_1) \leq x(a_2, b_2)$$

- $l : A \times B \to \mathbb{R}$ is a loss function iff $x(a, b) = -l(a, b)$ is utility.

Optimality

- A utility representation of $\lesssim$ on $A \times B$ is $x : A \times B \to \mathbb{R}$:

$$(a_1, b_1) \lesssim (a_2, b_2) \iff x(a_1, b_1) \leq x(a_2, b_2)$$

- $l : A \times B \to \mathbb{R}$ is a loss function iff $x(a, b) = -l(a, b)$ is utility.
- Given $p \in \mathcal{P}(A \times B)$, the expected utility is

$$\mathbb{E}_p\{x\} := \sum_{(a,b)\in A\times B} x(a, b)\, P(a \cap b)$$

Optimality

- A utility representation of $\lesssim$ on $A \times B$ is $x : A \times B \to \mathbb{R}$:

$$(a_1, b_1) \lesssim (a_2, b_2) \quad \iff \quad x(a_1, b_1) \leq x(a_2, b_2)$$

- $l : A \times B \to \mathbb{R}$ is a loss function iff $x(a, b) = -l(a, b)$ is utility.
- Given $p \in \mathcal{P}(A \times B)$, the expected utility is

$$\mathbb{E}_p\{x\} := \sum_{(a,b) \in A \times B} x(a, b)\, P(a \cap b)$$

- $\mathbb{E}_p\{x\}$ represents $\lesssim$ on $\mathcal{P}(A \times B)$: $q \lesssim p \iff \mathbb{E}_q\{x\} \leq \mathbb{E}_p\{x\}$.

Optimality

- A utility representation of $\lesssim$ on $A \times B$ is $x : A \times B \to \mathbb{R}$:

$$(a_1, b_1) \lesssim (a_2, b_2) \quad \iff \quad x(a_1, b_1) \leq x(a_2, b_2)$$

- $l : A \times B \to \mathbb{R}$ is a loss function iff $x(a, b) = -l(a, b)$ is utility.
- Given $p \in \mathcal{P}(A \times B)$, the expected utility is

$$\mathbb{E}_p\{x\} := \sum_{(a,b)\in A\times B} x(a, b)\, P(a \cap b)$$

- $\mathbb{E}_p\{x\}$ represents $\lesssim$ on $\mathcal{P}(A \times B)$: $q \lesssim p \iff \mathbb{E}_q\{x\} \leq \mathbb{E}_p\{x\}$.

Example (Time and utility of computation)

- Computational cost of Γ can be defined

$$l(\Gamma(w_0), w_0) := \begin{cases} t & \text{if } \Gamma(w_0) = (w_1, \dots, w_t) \text{ and } w_t \text{ is a final word} \\ \infty & \text{otherwise} \end{cases}$$

Optimality

- A utility representation of $\lesssim$ on $A \times B$ is $x : A \times B \to \mathbb{R}$:

$$(a_1, b_1) \lesssim (a_2, b_2) \quad \Longleftrightarrow \quad x(a_1, b_1) \leq x(a_2, b_2)$$

- $l : A \times B \to \mathbb{R}$ is a loss function iff $x(a,b) = -l(a,b)$ is utility.
- Given $p \in \mathcal{P}(A \times B)$, the expected utility is

$$\mathbb{E}_p\{x\} := \sum_{(a,b)\in A\times B} x(a,b)\, P(a \cap b)$$

- $\mathbb{E}_p\{x\}$ represents $\lesssim$ on $\mathcal{P}(A \times B)$: $q \lesssim p \iff \mathbb{E}_q\{x\} \leq \mathbb{E}_p\{x\}$.

Example (Time and utility of computation)

- Computational cost of Γ can be defined

$$l(\Gamma(w_0), w_0) := \begin{cases} t & \text{if } \Gamma(w_0) = (w_1, \dots, w_t) \text{ and } w_t \text{ is a final word} \\ \infty & \text{otherwise} \end{cases}$$

- Boolean utility $x(\Gamma(w_0), w_0) = 1 - \delta_\infty(l(\Gamma(w_0), w_0))$

Information

Definition (Information resource (distance))

A closed (lower semicontinuous) functional $F : \mathcal{P} \to \mathbb{R} \cup \{\infty\}$ ($I : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+ \cup \{\infty\}$). We usually put $F(p) = I(p, q)$.

Information

Definition (Information resource (distance))

A closed (lower semicontinuous) functional $F : \mathcal{P} \to \mathbb{R} \cup \{\infty\}$ ($I : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+ \cup \{\infty\}$). We usually put $F(p) = I(p, q)$.

Example (Kullback-Leibler divergence (Kullback, 1959))

$$I_{KL} := \mathbb{E}_p\{\ln(p/q)\}$$

Additive: $I_{KL}(p_1 p_2, q_1, q_2) = I_{KL}(p_1, q_1) + I_{KL}(p_2, q_2)$.

Information

Definition (Information resource (distance))

A closed (lower semicontinuous) functional $F : \mathcal{P} \to \mathbb{R} \cup \{\infty\}$ $(I : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+ \cup \{\infty\})$. We usually put $F(p) = I(p, q)$.

Example (Kullback-Leibler divergence (Kullback, 1959))

$$I_{KL} := \mathbb{E}_p\{\ln(p/q)\}$$

Additive: $I_{KL}(p_1 p_2, q_1, q_2) = I_{KL}(p_1, q_1) + I_{KL}(p_2, q_2)$.

Example (Total variation and Fisher information metrics)

$$I_V(p, q) = \|p - q\|_1 \,, \quad I_F(p, q) = 2 \arccos\langle 1, p^{1/2} q^{1/2} \rangle$$

Variational Problems

- $\mathbb{E}_p\{x\} = \langle x, p \rangle$ is linear.

Variational Problems

- $\mathbb{E}_p\{x\} = \langle x, p \rangle$ is linear.
- $I(p, q)$ is closed (lower semicontinuous), possibly convex.

Variational Problems

- $\mathbb{E}_p\{x\} = \langle x, p \rangle$ is linear.
- $I(p, q)$ is closed (lower semicontinuous), possibly convex.

Problems

- Linear programming problem:

$$\text{maximize (minimize)} \quad \mathbb{E}_p\{x\} \quad \text{subject to} \quad \mathbb{E}_p\{\ln(p/q)\} \leq \lambda$$

Variational Problems

- $\mathbb{E}_p\{x\} = \langle x, p \rangle$ is linear.
- $I(p, q)$ is closed (lower semicontinuous), possibly convex.

Problems

- Linear programming problem:

$$\text{maximize (minimize)} \quad \mathbb{E}_p\{x\} \quad \text{subject to} \quad \mathbb{E}_p\{\ln(p/q)\} \leq \lambda$$

- The inverse convex programming problem:

$$\text{minimize} \quad \mathbb{E}_p\{\ln(p/q)\} \quad \text{subject to} \quad \mathbb{E}_p\{x\} \geq \upsilon \quad \left(\mathbb{E}_p\{x\} \leq \upsilon\right)$$

Variational Problems

- $\mathbb{E}_p\{x\} = \langle x, p \rangle$ is linear.
- $I(p, q)$ is closed (lower semicontinuous), possibly convex.

Problems

- Linear programming problem:

$$\text{maximize (minimize)} \quad \mathbb{E}_p\{x\} \quad \text{subject to} \quad I(p, q) \leq \lambda$$

- The inverse convex programming problem:

$$\text{minimize} \quad I(p, q) \quad \text{subject to} \quad \mathbb{E}_p\{x\} \geq \upsilon \quad \Big(\mathbb{E}_p\{x\} \leq \upsilon\Big)$$

Variational Problems

- $\mathbb{E}_p\{x\} = \langle x, p \rangle$ is linear.
- $I(p,q)$ is closed (lower semicontinuous), possibly convex.

Problems

- Linear programming problem:

$$\text{maximize (minimize)} \quad \mathbb{E}_p\{x\} \quad \text{subject to} \quad F(p) \leq \lambda$$

Variational Problems

- $\mathbb{E}_p\{x\} = \langle x, p \rangle$ is linear.
- $I(p, q)$ is closed (lower semicontinuous), possibly convex.

Problems

- Linear programming problem:

$$\text{maximize (minimize)} \quad \mathbb{E}_p\{x\} \quad \text{subject to} \quad F(p) \leq \lambda$$

- The inverse convex programming problem:

$$\text{minimize} \quad F(p) \quad \text{subject to} \quad \mathbb{E}_p\{x\} \geq \upsilon \quad \left(\mathbb{E}_p\{x\} \leq \upsilon\right)$$

Variational Problems

- $\mathbb{E}_p\{x\} = \langle x, p\rangle$ is linear.
- $I(p, q)$ is closed (lower semicontinuous), possibly convex.

Problems

- Linear programming problem:

$$\text{maximize (minimize)} \quad \mathbb{E}_p\{x\} \quad \text{subject to} \quad F(p) \leq \lambda$$

- The inverse convex programming problem:

$$\text{minimize} \quad F(p) \quad \text{subject to} \quad \mathbb{E}_p\{x\} \geq \upsilon \quad \left(\mathbb{E}_p\{x\} \leq \upsilon\right)$$

- For $p \in \mathcal{P}(A \times B)$, each $p := P(A \cap B) = P(A \mid B)\, P(B)$.

Variational Problems

- $\mathbb{E}_p\{x\} = \langle x, p \rangle$ is linear.
- $I(p, q)$ is closed (lower semicontinuous), possibly convex.

Problems

- Linear programming problem:

$$\text{maximize (minimize)} \quad \mathbb{E}_p\{x\} \quad \text{subject to} \quad F(p) \leq \lambda$$

- The inverse convex programming problem:

$$\text{minimize} \quad F(p) \quad \text{subject to} \quad \mathbb{E}_p\{x\} \geq \upsilon \quad \left(\mathbb{E}_p\{x\} \leq \upsilon\right)$$

- For $p \in \mathcal{P}(A \times B)$, each $p := P(A \cap B) = P(A \mid B)\,P(B)$.

  type II : Find optimal $P(B)$ for fixed $P(A \mid b)$.

Variational Problems

- $\mathbb{E}_p\{x\} = \langle x, p\rangle$ is linear.
- $I(p, q)$ is closed (lower semicontinuous), possibly convex.

Problems

- Linear programming problem:

$$\text{maximize (minimize)} \quad \mathbb{E}_p\{x\} \quad \text{subject to} \quad F(p) \leq \lambda$$

- The inverse convex programming problem:

$$\text{minimize} \quad F(p) \quad \text{subject to} \quad \mathbb{E}_p\{x\} \geq \upsilon \quad \left(\mathbb{E}_p\{x\} \leq \upsilon\right)$$

- For $p \in \mathcal{P}(A \times B)$, each $p := P(A \cap B) = P(A \mid B)\, P(B)$.

  type II : Find optimal $P(B)$ for fixed $P(A \mid b)$.

  type III : Find optimal $P(A \mid b)$ for fixed $P(B)$.

Optimal Solutions

Necessary and sufficient conditions

- p_β maximizes $\langle x, p \rangle$ on $\{p : F(p) \leq \lambda\}$ iff

$$p_\beta \in \partial F^*(\beta x)\,, \qquad F(p_\beta) = \lambda$$

Optimal Solutions

Necessary and sufficient conditions

- p_β maximizes $\langle x, p\rangle$ on $\{p : F(p) \leq \lambda\}$ iff

$$p_\beta \in \partial F^*(\beta x)\,, \qquad F(p_\beta) = \lambda$$

- p_β minimizes $F(p)$ on $\{p : \langle x, p\rangle \geq \upsilon\}$ iff

$$p_\beta \in \partial F^*(\beta x)\,, \qquad \langle x, p_\beta\rangle = \upsilon$$

Optimal Solutions

Necessary and sufficient conditions

- p_β maximizes $\langle x, p\rangle$ on $\{p : F(p) \leq \lambda\}$ iff

$$p_\beta \in \partial F^*(\beta x)\,, \qquad F(p_\beta) = \lambda$$

- p_β minimizes $F(p)$ on $\{p : \langle x, p\rangle \geq \upsilon\}$ iff

$$p_\beta \in \partial F^*(\beta x)\,, \qquad \langle x, p_\beta\rangle = \upsilon$$

- Lagrange multiplier β^{-1} (or β) is defined by λ (υ).

Optimal Solutions

Necessary and sufficient conditions

- p_β maximizes $\langle x, p\rangle$ on $\{p : F(p) \leq \lambda\}$ iff

$$p_\beta \in \partial F^*(\beta x)\,, \qquad F(p_\beta) = \lambda$$

- p_β minimizes $F(p)$ on $\{p : \langle x, p\rangle \geq \upsilon\}$ iff

$$p_\beta \in \partial F^*(\beta x)\,, \qquad \langle x, p_\beta\rangle = \upsilon$$

- Lagrange multiplier β^{-1} (or β) is defined by λ (υ).

- For $F(p) = I_{KL}(p, q) = \mathbb{E}_p\{\ln(p/q)\}$, we have

$$\partial F^*(\beta x) = \{p_\beta \propto e^{\beta x}\, q\}$$

Optimal Solutions

Necessary and sufficient conditions

- p_β maximizes $\langle x, p\rangle$ on $\{p : F(p) \leq \lambda\}$ iff

$$p_\beta \in \partial F^*(\beta x)\,, \qquad F(p_\beta) = \lambda$$

- p_β minimizes $F(p)$ on $\{p : \langle x, p\rangle \geq \upsilon\}$ iff

$$p_\beta \in \partial F^*(\beta x)\,, \qquad \langle x, p_\beta\rangle = \upsilon$$

- Lagrange multiplier β^{-1} (or β) is defined by λ (υ).

- For $F(p) = I_{KL}(p, q) = \mathbb{E}_p\{\ln(p/q)\}$, we have

$$\partial F^*(\beta x) = \{p_\beta \propto e^{\beta x}\, q\}$$

- With normalization $\|p\|_1 = 1$:

$$p_\beta = e^{\beta\, x - \Psi_x(\beta)}\, q\,, \qquad p_0 = q$$

Exponential kernels

- Let $p := P(a \cap b)$, $q := P(a)P(b)$ (i.e. p, $q \in \mathcal{P}(A \times B)$)

Exponential kernels

- Let $p := P(a \cap b)$, $q := P(a)P(b)$ (i.e. p, $q \in \mathcal{P}(A \times B)$)
- Then $I_{KL}(p, q) := \mathbb{E}_p\{\ln(p/q)\}$ is Shannon mutual information:

$$
\begin{aligned}
I_S\{a, b\} &:= \sum_{A \times B} \ln\left[\frac{P(a \cap b)}{P(a)\,P(b)}\right] P(a \cap b) \\
&= \sum_{B} P(b) \sum_{A} \ln\left[\frac{P(a \mid b)}{P(a)}\right] P(a \mid b)
\end{aligned}
$$

Exponential kernels

- Let $p := P(a \cap b)$, $q := P(a)P(b)$ (i.e. p, $q \in \mathcal{P}(A \times B)$)
- Then $I_{KL}(p, q) := \mathbb{E}_p\{\ln(p/q)\}$ is Shannon mutual information:

$$\begin{aligned} I_S\{a, b\} &:= \sum_{A \times B} \ln\left[\frac{P(a \cap b)}{P(a)\,P(b)}\right] P(a \cap b) \\ &= \sum_{B} P(b) \sum_{A} \ln\left[\frac{P(a \mid b)}{P(a)}\right] P(a \mid b) \end{aligned}$$

- The optimal conditional probabilities are

$$P_\beta(a \mid b) = e^{\beta\, x(a,b) - \Psi_x(\beta, b)}\, P(a)$$

Exponential kernels

- Let $p := P(a \cap b)$, $q := P(a)P(b)$ (i.e. p, $q \in \mathcal{P}(A \times B)$)
- Then $I_{KL}(p, q) := \mathbb{E}_p\{\ln(p/q)\}$ is Shannon mutual information:

$$
\begin{aligned}
I_S\{a, b\} &:= \sum_{A \times B} \ln\left[\frac{P(a \cap b)}{P(a)\, P(b)}\right] P(a \cap b) \\
&= \sum_{B} P(b) \sum_{A} \ln\left[\frac{P(a \mid b)}{P(a)}\right] P(a \mid b)
\end{aligned}
$$

- The optimal conditional probabilities are

$$P_\beta(a \mid b) = e^{\beta\, x(a,b) - \Psi_x(\beta, b)}\, P(a)$$

- Observe that for $\beta < \infty$

$$P_\beta(a \mid b) > 0 \qquad \Longleftrightarrow \qquad P(a) > 0$$

Exponential kernels

- Let $p := P(a \cap b)$, $q := P(a)P(b)$ (i.e. p, $q \in \mathcal{P}(A \times B)$)
- Then $I_{KL}(p, q) := \mathbb{E}_p\{\ln(p/q)\}$ is Shannon mutual information:

$$\begin{aligned} I_S\{a, b\} &:= \sum_{A \times B} \ln \left[\frac{P(a \cap b)}{P(a)\,P(b)} \right] P(a \cap b) \\ &= \sum_{B} P(b) \sum_{A} \ln \left[\frac{P(a \mid b)}{P(a)} \right] P(a \mid b) \end{aligned}$$

- The optimal conditional probabilities are

$$P_\beta(a \mid b) = e^{\beta\, x(a,b) - \Psi_x(\beta, b)}\, P(a)$$

- Observe that for $\beta < \infty$

$$P_\beta(a \mid b) > 0 \qquad \iff \qquad P(a) > 0$$

- Thus, $p_\beta := P_\beta(a \mid b)\, P(b) \in \mathrm{Int}(\mathcal{P}(A \times B))$ is non-deterministic.

Non-existence Theorem (General Case)

Theorem (Belavkin, Accepted)

- *Let* $\{p_\beta\}_x$ *be a family of* $p_\beta \in \mathcal{P}(A \times B)$ *maximizing* $\mathbb{E}_p\{x\}$ *on sets* $\{p : F(p) \leq \lambda\}$ *for all values* $\lambda = F(p)$.

Non-existence Theorem (General Case)

Theorem (Belavkin, Accepted)

- *Let* $\{p_\beta\}_x$ *be a family of* $p_\beta \in \mathcal{P}(A \times B)$ *maximizing* $\mathbb{E}_p\{x\}$ *on sets* $\{p : F(p) \leq \lambda\}$ *for all values* $\lambda = F(p)$.
- *Let* F *be minimized at* $p_0 \in \partial F^*(0) \subset \mathrm{Int}(\mathcal{P}(A \times B))$.

Non-existence Theorem (General Case)

Theorem (Belavkin, Accepted)

- *Let* $\{p_\beta\}_x$ *be a family of* $p_\beta \in \mathcal{P}(A \times B)$ *maximizing* $\mathbb{E}_p\{x\}$ *on sets* $\{p : F(p) \leq \lambda\}$ *for all values* $\lambda = F(p)$.
- *Let* F *be minimized at* $p_0 \in \partial F^*(0) \subset \mathrm{Int}(\mathcal{P}(A \times B))$.
- *If* $F^*(x)$ *is strictly convex, then*

Non-existence Theorem (General Case)

Theorem (Belavkin, Accepted)

- *Let* $\{p_\beta\}_x$ *be a family of* $p_\beta \in \mathcal{P}(A \times B)$ *maximizing* $\mathbb{E}_p\{x\}$ *on sets* $\{p : F(p) \leq \lambda\}$ *for all values* $\lambda = F(p)$.
- *Let* F *be minimized at* $p_0 \in \partial F^*(0) \subset \mathrm{Int}(\mathcal{P}(A \times B))$.
- *If* $F^*(x)$ *is strictly convex, then*
  1. p_β *is deterministic iff* $\lambda \geq \sup^x F$ *(i.e.* $\beta \to \infty$*).*

Non-existence Theorem (General Case)

Theorem (Belavkin, Accepted)

- *Let* $\{p_\beta\}_x$ *be a family of* $p_\beta \in \mathcal{P}(A \times B)$ *maximizing* $\mathbb{E}_p\{x\}$ *on sets* $\{p : F(p) \leq \lambda\}$ *for all values* $\lambda = F(p)$.
- *Let* F *be minimized at* $p_0 \in \partial F^*(0) \subset \mathrm{Int}(\mathcal{P}(A \times B))$.
- *If* $F^*(x)$ *is strictly convex, then*
  1. p_β *is deterministic iff* $\lambda \geq \sup^x F$ *(i.e.* $\beta \to \infty$*).*
  2. *If* $p_f \in \mathcal{P}(A \times B)$ *is deterministic and* $F(p_f) = F(p_\beta)$*, then*

$$\mathbb{E}_{p_f}\{x\} < \mathbb{E}_{p_\beta}\{x\}$$

Non-existence Theorem (General Case)

Theorem (Belavkin, Accepted)

- *Let* $\{p_\beta\}_x$ *be a family of* $p_\beta \in \mathcal{P}(A \times B)$ *maximizing* $\mathbb{E}_p\{x\}$ *on sets* $\{p : F(p) \leq \lambda\}$ *for all values* $\lambda = F(p)$.
- *Let* F *be minimized at* $p_0 \in \partial F^*(0) \subset \mathrm{Int}(\mathcal{P}(A \times B))$.
- *If* $F^*(x)$ *is strictly convex, then*
  1. p_β *is deterministic iff* $\lambda \geq \sup^x F$ *(i.e.* $\beta \to \infty$*).*
  2. *If* $p_f \in \mathcal{P}(A \times B)$ *is deterministic and* $F(p_f) = F(p_\beta)$*, then*
  $$\mathbb{E}_{p_f}\{x\} < \mathbb{E}_{p_\beta}\{x\}$$
  3. *If* $p_f \in \mathcal{P}(A \times B)$ *is deterministic and* $\mathbb{E}_{p_f}\{x\} = \mathbb{E}_{p_\beta}\{x\}$*, then*
  $$F(p_f) > F(p_\beta)$$

Discussion

- Strict convexity of F^* ensures that Gâteaux derivative of $F(p)$ exists at any $p \in \mathrm{Int}(\mathrm{dom}F)$.

Discussion

- Strict convexity of F^* ensures that Gâteaux derivative of $F(p)$ exists at any $p \in \text{Int}(\text{dom}F)$.
- If F^* is not strictly convex, then p_β may maximize $\mathbb{E}_p\{x\}$ and $\mathbb{E}_p\{w\}$ with utilities x and w representing different $\lesssim$ on $A \times B$ (i.e. $p \in \mathcal{P}(A \times B)$ do not separate some variational problems).

Discussion

- Strict convexity of F^* ensures that Gâteaux derivative of $F(p)$ exists at any $p \in \mathrm{Int}(\mathrm{dom} F)$.
- If F^* is not strictly convex, then p_β may maximize $\mathbb{E}_p\{x\}$ and $\mathbb{E}_p\{w\}$ with utilities x and w representing different $\lesssim$ on $A \times B$ (i.e. $p \in \mathcal{P}(A \times B)$ do not separate some variational problems).
- Additivity of $I(p,q)$ implies strict convexity of $I^*(x,q)$ (i.e. $\nabla I(p,q) \propto \ln(p/q) \iff \exp(x)\, q = \nabla I^*(x,q)$).

Discussion

- Strict convexity of F^* ensures that Gâteaux derivative of $F(p)$ exists at any $p \in \mathrm{Int}(\mathrm{dom} F)$.
- If F^* is not strictly convex, then p_β may maximize $\mathbb{E}_p\{x\}$ and $\mathbb{E}_p\{w\}$ with utilities x and w representing different $\lesssim$ on $A \times B$ (i.e. $p \in \mathcal{P}(A \times B)$ do not separate some variational problems).
- Additivity of $I(p, q)$ implies strict convexity of $I^*(x, q)$ (i.e. $\nabla I(p, q) \propto \ln(p/q) \iff \exp(x)\, q = \nabla I^*(x, q)$).
- One can construct examples such that for any deterministic p_f satisfying any constraint $F(p) \leq \lambda < \sup^x F(p)$ (or any $\mathbb{E}_p\{x\} \geq \upsilon > \upsilon_0$)

$$\mathbb{E}_{p_f}\{x\} = -\infty \qquad (\text{or } F(p_f) = \infty)$$

Discussion

- Strict convexity of F^* ensures that Gâteaux derivative of $F(p)$ exists at any $p \in \mathrm{Int}(\mathrm{dom} F)$.
- If F^* is not strictly convex, then p_β may maximize $\mathbb{E}_p\{x\}$ and $\mathbb{E}_p\{w\}$ with utilities x and w representing different $\lesssim$ on $A \times B$ (i.e. $p \in \mathcal{P}(A \times B)$ do not separate some variational problems).
- Additivity of $I(p,q)$ implies strict convexity of $I^*(x,q)$ (i.e. $\nabla I(p,q) \propto \ln(p/q) \iff \exp(x)\, q = \nabla I^*(x,q)$).
- One can construct examples such that for any deterministic p_f satisfying any constraint $F(p) \leq \lambda < \sup^x F(p)$ (or any $\mathbb{E}_p\{x\} \geq \upsilon > \upsilon_0$)

$$\mathbb{E}_{p_f}\{x\} = -\infty \qquad (\text{or } F(p_f) = \infty)$$

- Strict inequalities for solutions p_β: $-\infty < \mathbb{E}_{p_\beta}\{x\}$ or $\infty > F(p_\beta)$.

Belavkin, R. V. (Accepted). Optimal measures and transition kernels. *Journal of Global Optimization*.

Kolmogorov, A. N., & Uspenskii, V. A. (1958). On the definition of an algorithm. *Uspekhi Mat. Nauk*, *13*(4), 3–28. (In Russian)

Kullback, S. (1959). *Information theory and statistics*. John Wiley and Sons.

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