Value of Information and Geometry of Optimal Decision-making and Learning

Roman V. Belavkin

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

July 22, 2018
ACDL 2018

Optimal learning

First variational problem

Value of information

Iterative algorithms as Markov morphisms

Optimal transition kernels

Optimal control of mutation rate

Optimal learning

Learning, Optimization, Self-Adaptation

(Stratonovich, 1968):

> *Is there a theory for a synthesis of optimal adaptive, self-learning and self-organising systems?*

Learning, Optimization, Self-Adaptation

(Stratonovich, 1968):

> *Is there a theory for a synthesis of optimal adaptive, self-learning and self-organising systems?*

- Is there a difference between Bayesian (statistical) estimation and control theory and theory of learning systems?

Learning, Optimization, Self-Adaptation

(Stratonovich, 1968):

Is there a theory for a synthesis of optimal adaptive, self-learning and self-organising systems?

- Is there a difference between Bayesian (statistical) estimation and control theory and theory of learning systems?
- Learning algorithms usually involve iterations $x_{t+1} = \Lambda(x_t)$.

Learning, Optimization, Self-Adaptation

(Stratonovich, 1968):

> *Is there a theory for a synthesis of optimal adaptive, self-learning and self-organising systems?*

- Is there a difference between Bayesian (statistical) estimation and control theory and theory of learning systems?
- Learning algorithms usually involve iterations $x_{t+1} = \Lambda(x_t)$.
- Learning problems are characterized by prior uncertainty.

What is learning?

Proposition

There is no need to learn, if there is nothing to optimize.

What is learning?

Proposition

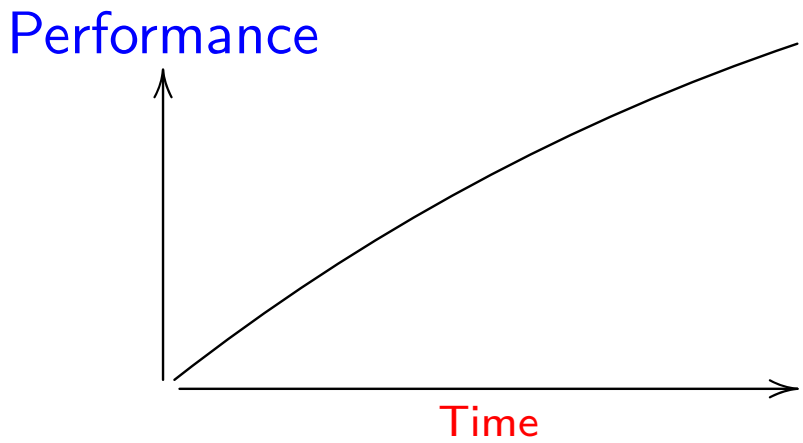
There is no need to learn, if there is nothing to optimize.

Proposition

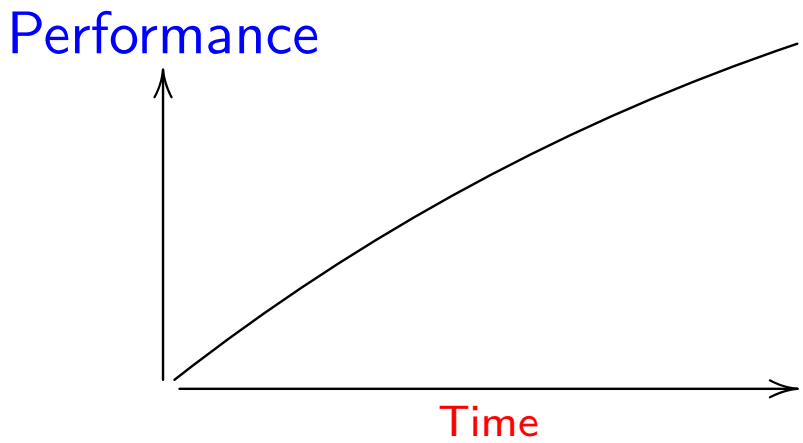
There is nothing to learn, if one has full information.

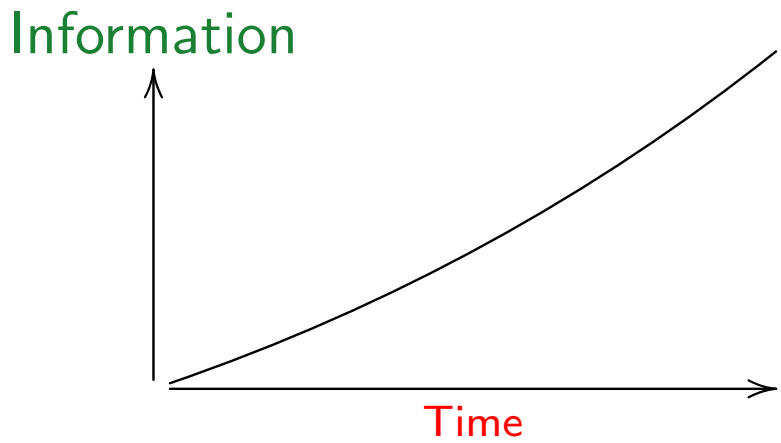
Learning Systems

Learning Systems

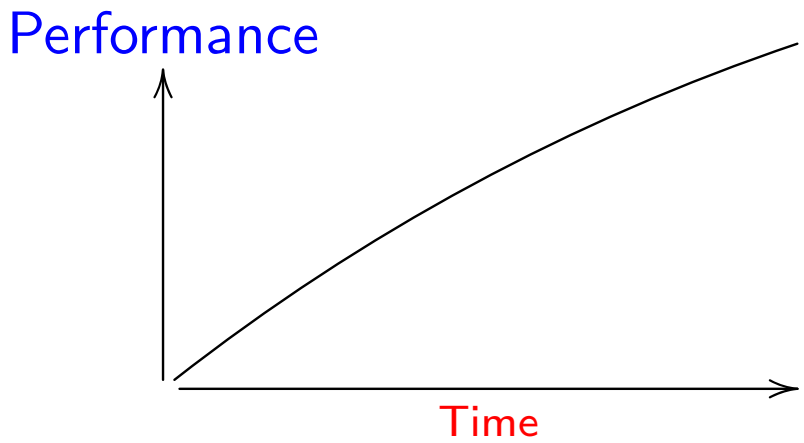


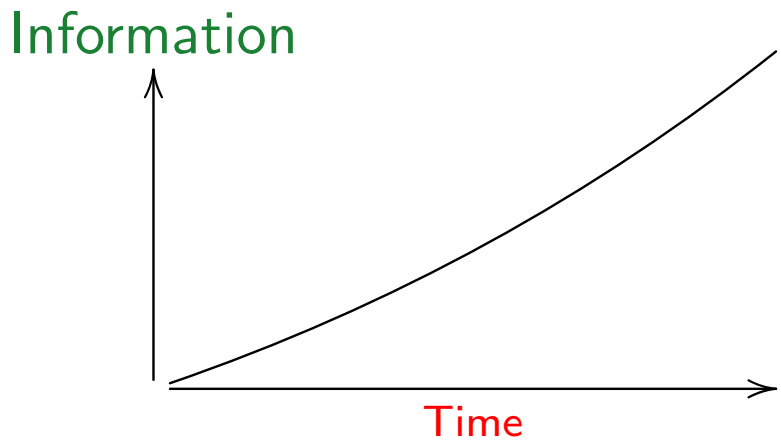
Learning Systems



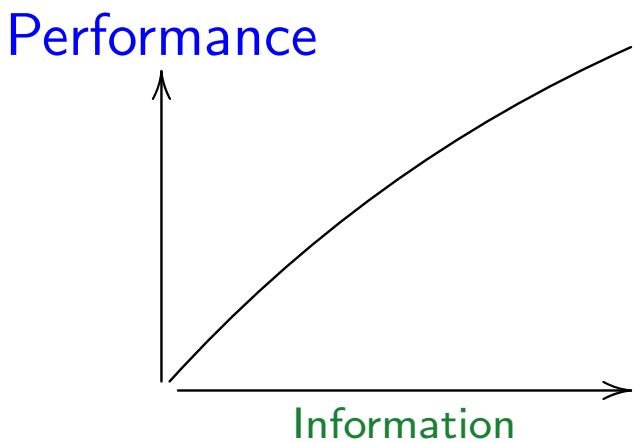


Learning Systems





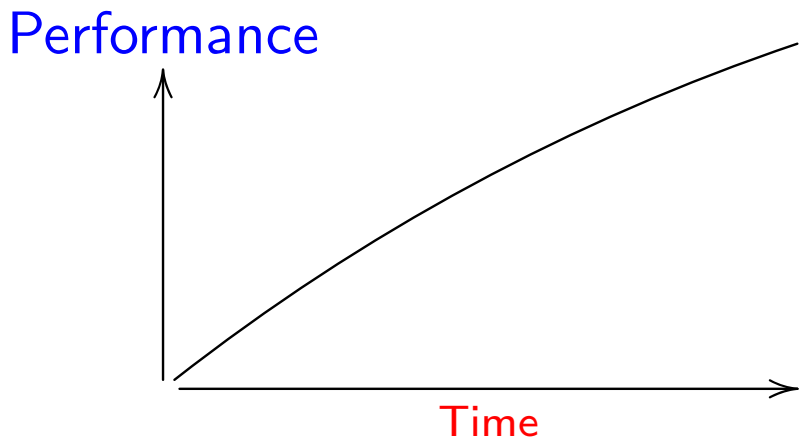
Optimal learning

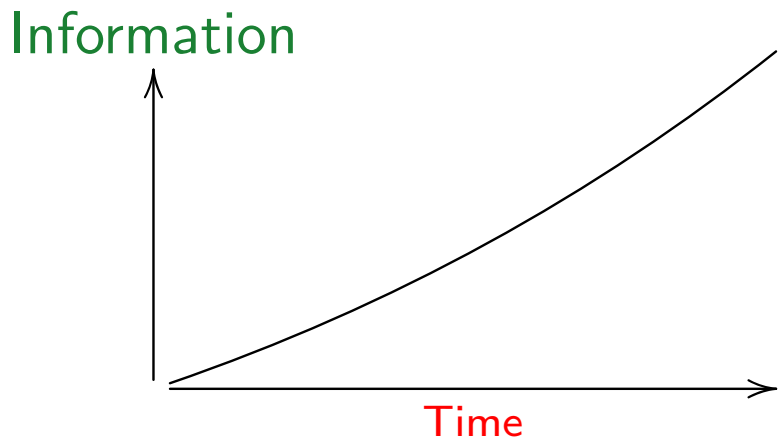


Maximize performance

s.t. information $\leq \lambda$

Learning Systems

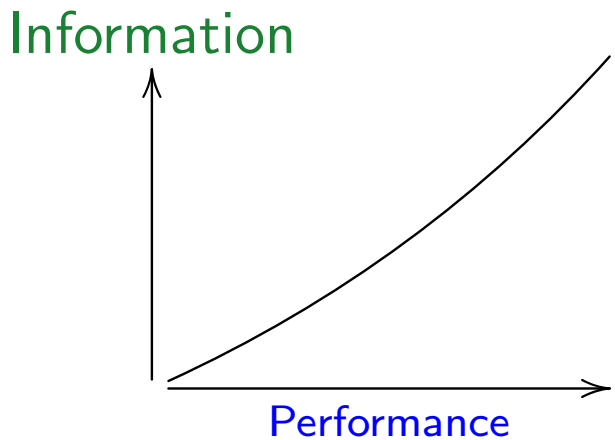




Optimal learning

Minimize information

s.t. performance $\geq v$



Optimization vs Learning

- (Ω, E, p) — probability space.

Optimization vs Learning

- (Ω, E, p) — probability space.
- $u : \Omega \to \mathbb{R}$ — utility function.

Optimization vs Learning

- (Ω, E, p) — probability space.
- $u : \Omega \to \mathbb{R}$ — utility function.
- Performance can be measured by the expected utility:

$$\mathbb{E}_p\{u\} := \sum_{\omega \in \Omega} u(\omega)\, p(\omega)$$

Optimization vs Learning

- (Ω, E, p) — probability space.
- $u : \Omega \to \mathbb{R}$ — utility function.
- Performance can be measured by the expected utility:

$$\mathbb{E}_p\{u\} := \sum_{\omega \in \Omega} u(\omega)\, p(\omega)$$

Example (Least squares estimation)

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

$$\nabla_z \mathbb{E}_p\{u(x, z)\} = \sum \nabla_z u(x, z)\, p(x) = \hat{z} - \sum x\, p(x) = 0$$

Optimization vs Learning

- (Ω, E, p) — probability space.
- $u : \Omega \to \mathbb{R}$ — utility function.
- Performance can be measured by the expected utility:

$$\mathbb{E}_p\{u\} := \sum_{\omega \in \Omega} u(\omega)\, p(\omega)$$

Example (Least squares estimation)

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

$$\nabla_z \mathbb{E}_p\{u(x,z)\} = \sum \nabla_z u(x,z)\, p(x) = \hat{z} - \sum x\, p(x) = 0$$

- $\hat{z} = \mathbb{E}_P\{x\}$

Optimization vs Learning

- (Ω, E, p) — probability space.
- $u : \Omega \to \mathbb{R}$ — utility function.
- Performance can be measured by the expected utility:

$$\mathbb{E}_{?}\{u\} := \sum_{\omega \in \Omega} u(\omega)\, ?(\omega)$$

Example (Least squares estimation)

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

$$\nabla_z \mathbb{E}_p\{u(x, z)\} = \sum \nabla_z u(x, z)\, p(x) = \hat{z} - \sum x\, p(x) = 0$$

- $\hat{z} = \mathbb{E}_P\{x\}$

Optimization vs Learning

- (Ω, E, p) — probability space.
- $u : \Omega \to \mathbb{R}$ — utility function.
- Performance can be measured by the expected utility:

$$\mathbb{E}_{\frac{n(\omega)}{n}}\{u\} := \sum_{\omega \in \Omega} u(\omega)\, \frac{n(\omega)}{n}$$

Example (Least squares estimation)

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

$$\nabla_z \mathbb{E}_p\{u(x,z)\} = \sum \nabla_z u(x,z)\, p(x) = \hat{z} - \sum x\, p(x) = 0$$

- $\hat{z} = \mathbb{E}_P\{x\}$

Dynamics of learning

- Given n i.i.d. observatons of E with probability $p(E)$:

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{n(E)}{n} - p(E)\right| > \varepsilon\right\} = 0$$

(the weak law of large numbers)

Dynamics of learning

- Given n i.i.d. observatons of E with probability $p(E)$:

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{n(E)}{n} - p(E)\right| > \varepsilon\right\} = 0$$

  (the weak law of large numbers)
- Exchangeable sequences:

$$p(E_1, E_2) = p(E_2, E_1)$$

  (e.g. independent events are exchangeable $p(E_1, E_2) = p(E_1)p(E_2)$)

Dynamics of learning

- Given n i.i.d. observatons of E with probability $p(E)$:

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{n(E)}{n} - p(E)\right| > \varepsilon\right\} = 0$$

(the weak law of large numbers)

- Exchangeable sequences:

$$p(E_1, E_2) = p(E_2, E_1)$$

(e.g. independent events are exchangeable $p(E_1, E_2) = p(E_1)p(E_2)$)

Example (i.i.d in learning)

- A cat learns a distribution of mice in a house.

Dynamics of learning

- Given n i.i.d. observatons of E with probability $p(E)$:

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{n(E)}{n} - p(E)\right| > \varepsilon\right\} = 0$$

(the weak law of large numbers)

- Exchangeable sequences:

$$p(E_1, E_2) = p(E_2, E_1)$$

(e.g. independent events are exchangeable $p(E_1, E_2) = p(E_1)p(E_2)$)

Example (i.i.d in learning)

- A cat learns a distribution of mice in a house.
- What is the limit of the cat's empirical distribution $n(E)/n$?

Dynamics of learning

- Given n i.i.d. observatons of E with probability $p(E)$:

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{n(E)}{n} - p(E)\right| > \varepsilon\right\} = 0$$

(the weak law of large numbers)

- Exchangeable sequences:

$$p(E_1, E_2) = p(E_2, E_1)$$

(e.g. independent events are exchangeable $p(E_1, E_2) = p(E_1)p(E_2)$)

Example (i.i.d in learning)

- A cat learns a distribution of mice in a house.
- What is the limit of the cat's empirical distribution $n(E)/n$?
- Does $n(E)/n$ converge?

Dynamics of learning

- Given n i.i.d. observatons of E with probability $p(E)$:

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{n(E)}{n} - p(E)\right| > \varepsilon\right\} = 0$$

(the weak law of large numbers)

- Exchangeable sequences:

$$p(E_1, E_2) = p(E_2, E_1)$$

(e.g. independent events are exchangeable $p(E_1, E_2) = p(E_1)p(E_2)$)

Example (i.i.d in learning)

- A cat learns a distribution of mice in a house.
- What is the limit of the cat's empirical distribution $n(E)/n$?
- Does $n(E)/n$ converge?
- Oh! The mice learn the distribution of the cat.

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:

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

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:

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

- Entropy as self-information:

$$I\{x,x\} = H\{x\}$$

- Information upper bound:

$$0 \le I\{x,y\} \le H\{x\}$$

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:

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

- Entropy as self-information:

$$I\{x,x\} = H\{x\}$$

- Information upper bound:

$$0 \le I\{x,y\} \le H\{x\}$$

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

Entropy and information

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

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

- 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:

$$0 \leq I\{x, y\} \leq H\{x\}$$

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

Entropy and information

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

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

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

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

- Entropy as self-information:

$$I\{x, x\} = H\{x\}$$

- Information upper bound:

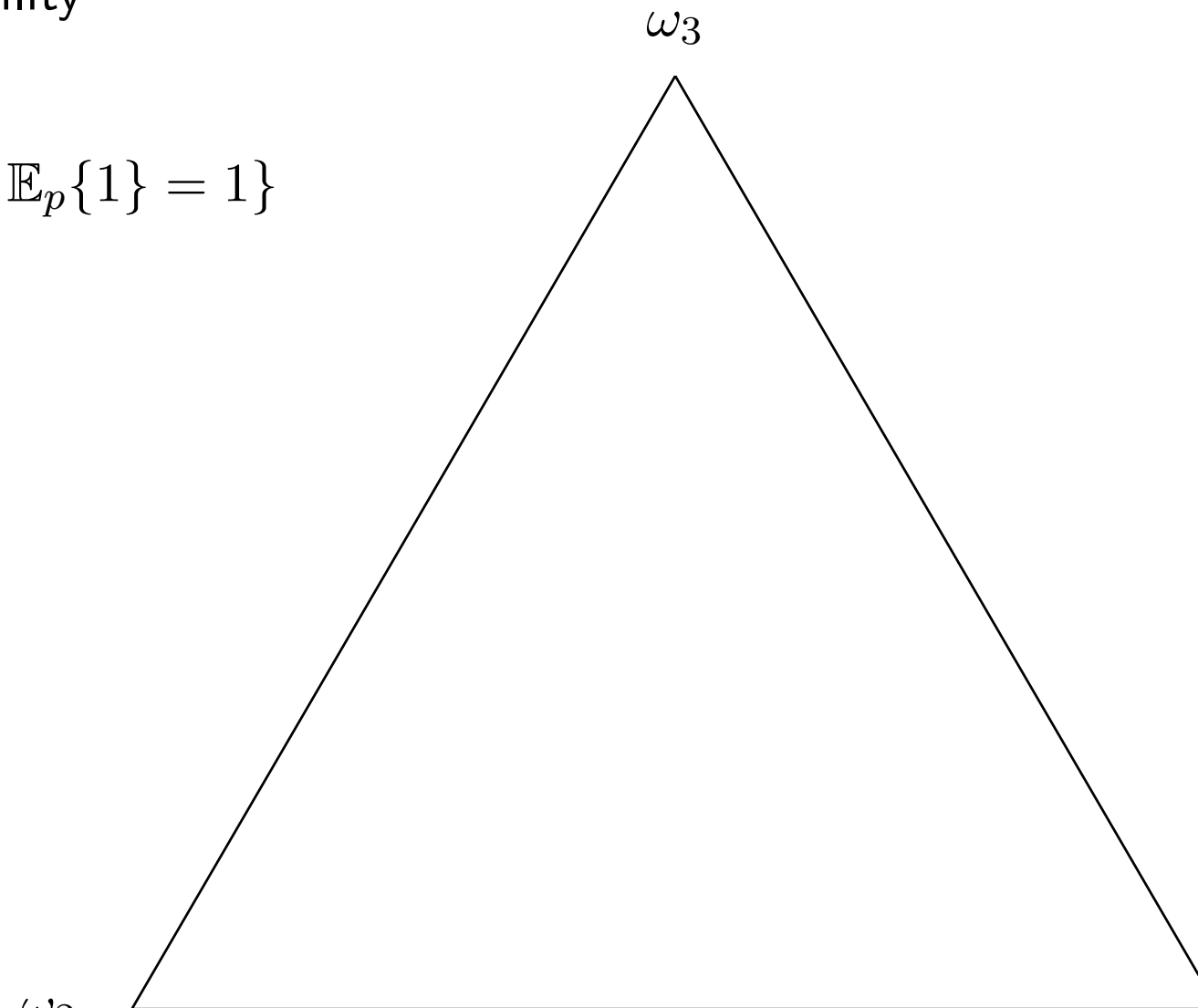
$$0 \le I\{x, y\} \le H\{x\}$$

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

Information-geometruc view

- The set of all probability measures

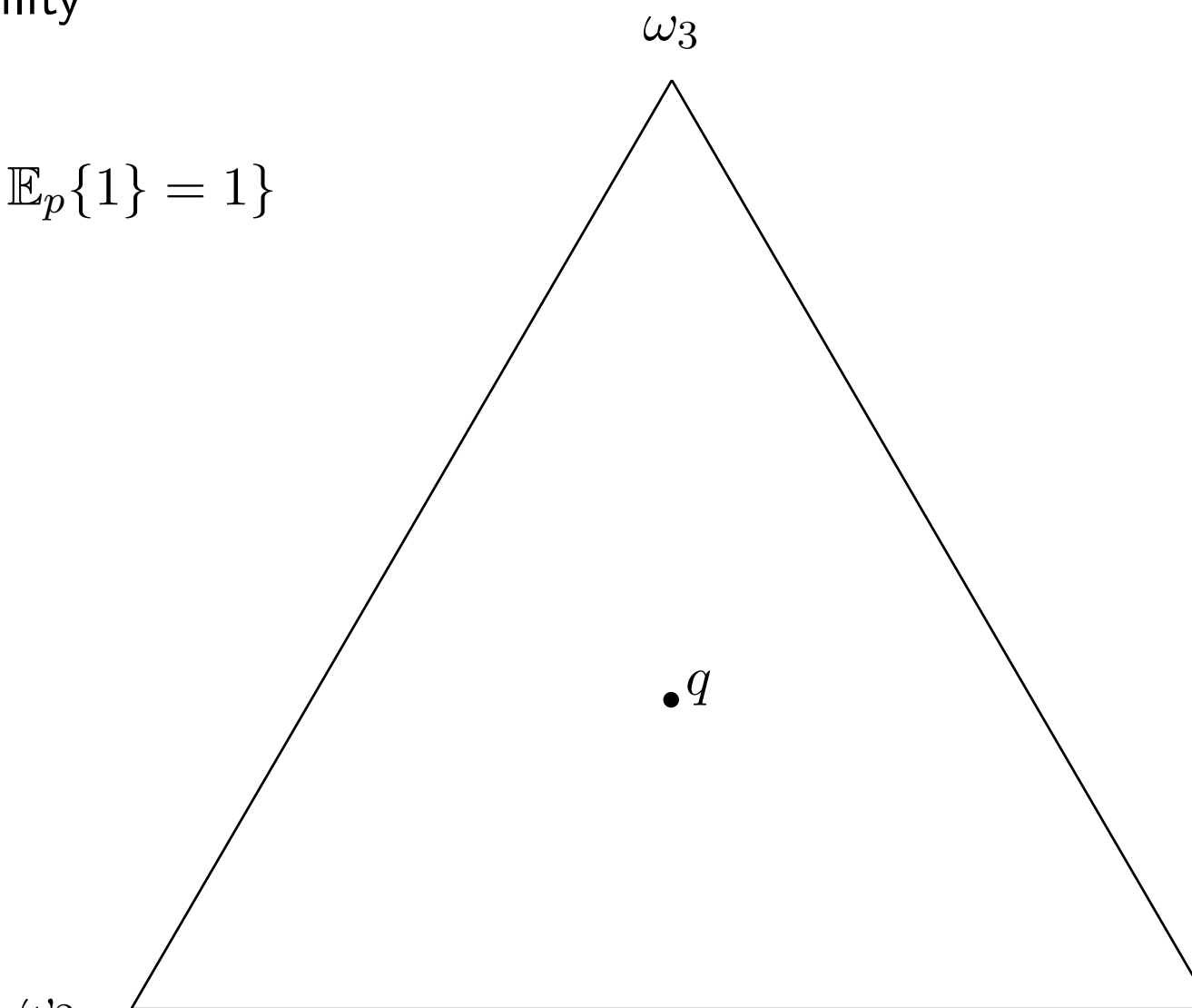
$$\mathcal{P}(\Omega) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$



Information-geometruc view

- The set of all probability measures

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

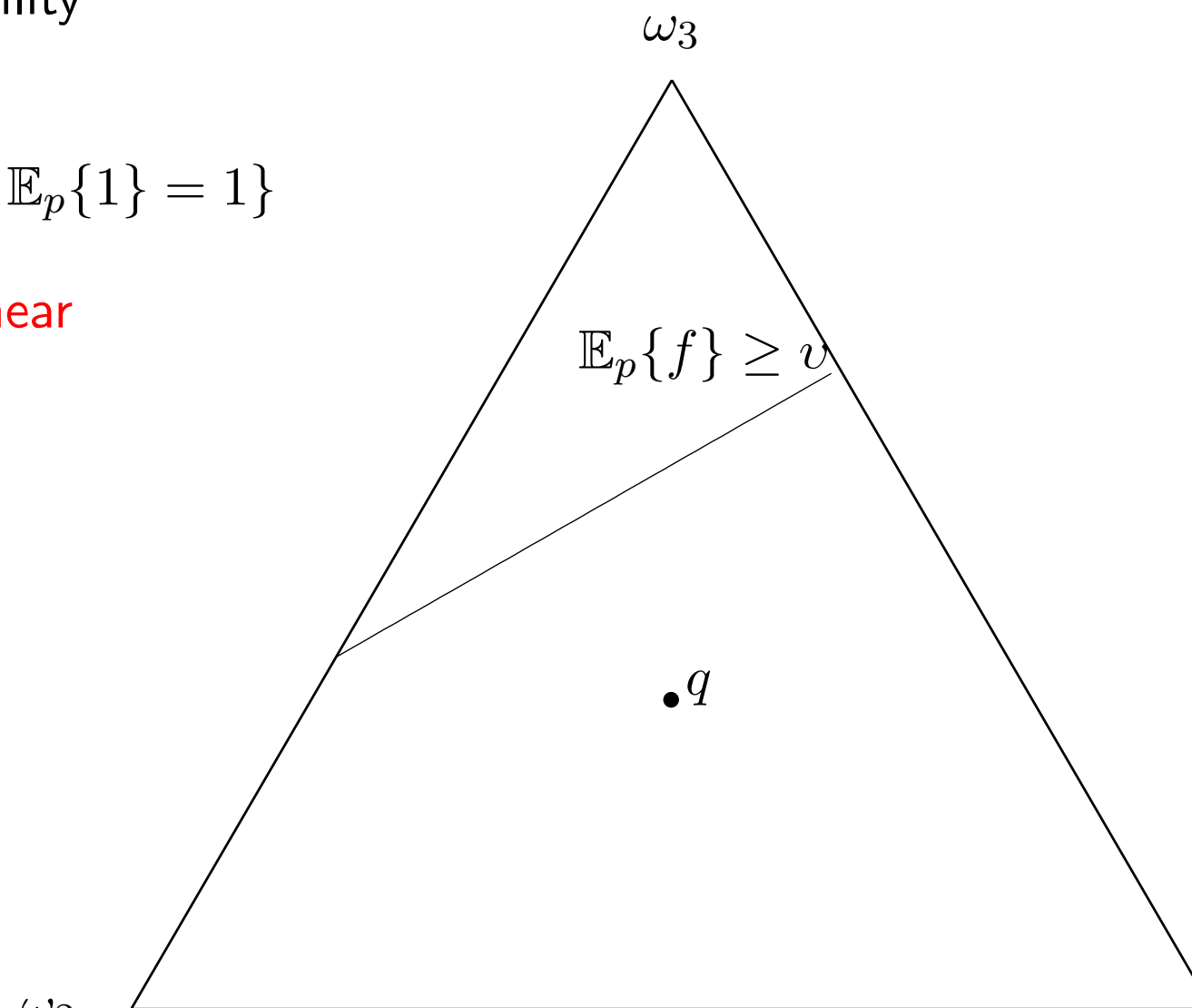


Information-geometruc view

- The set of all probability measures

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

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

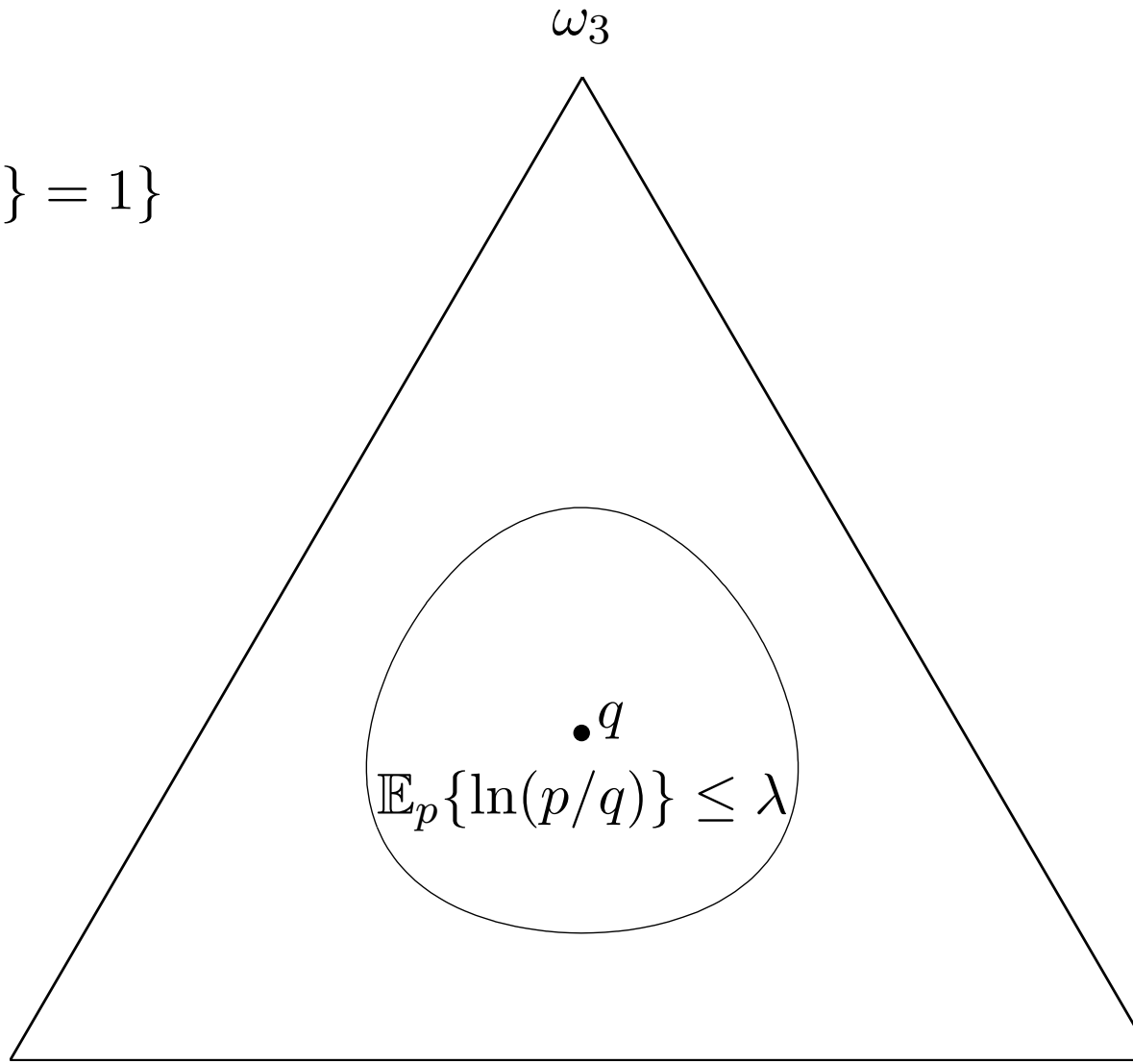


Information-geometruc view

- The set of all probability measures

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

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

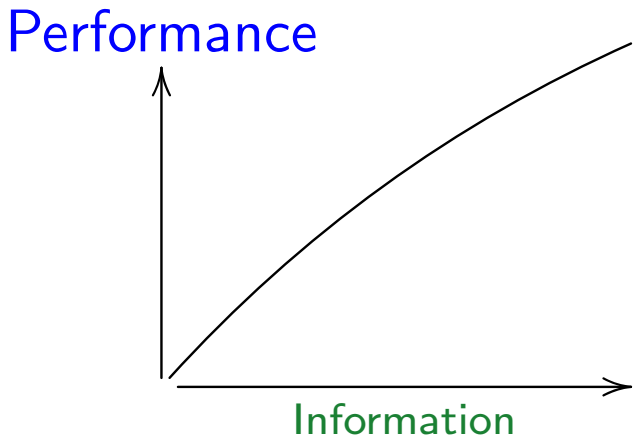


First variational problem

Problem I

- Linear programming problem:

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



Maximize performance

s.t. information $\leq \lambda$

First variational problem

Problem I

- Linear programming problem:

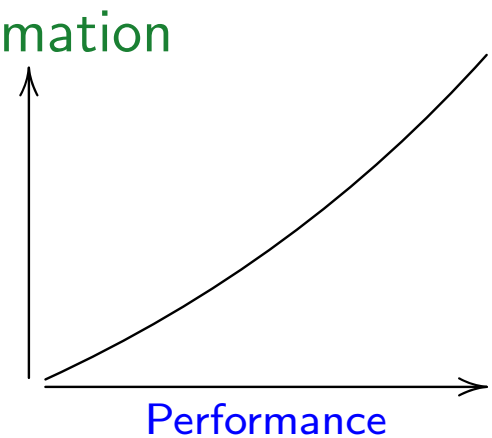
$$\text{maximize (minimize)} \quad \mathbb{E}_p\{u\} \quad \text{subject to} \quad \mathbb{E}_p\{\ln(p/q)\} \le \lambda$$

- The inverse convex programming problem:

$$\text{minimize} \quad \mathbb{E}_p\{\ln(p/q)\} \quad \text{subject to} \quad \mathbb{E}_p\{u\} \ge \upsilon \quad \left(\mathbb{E}_p\{u\} \le \upsilon\right)$$

Minimize information

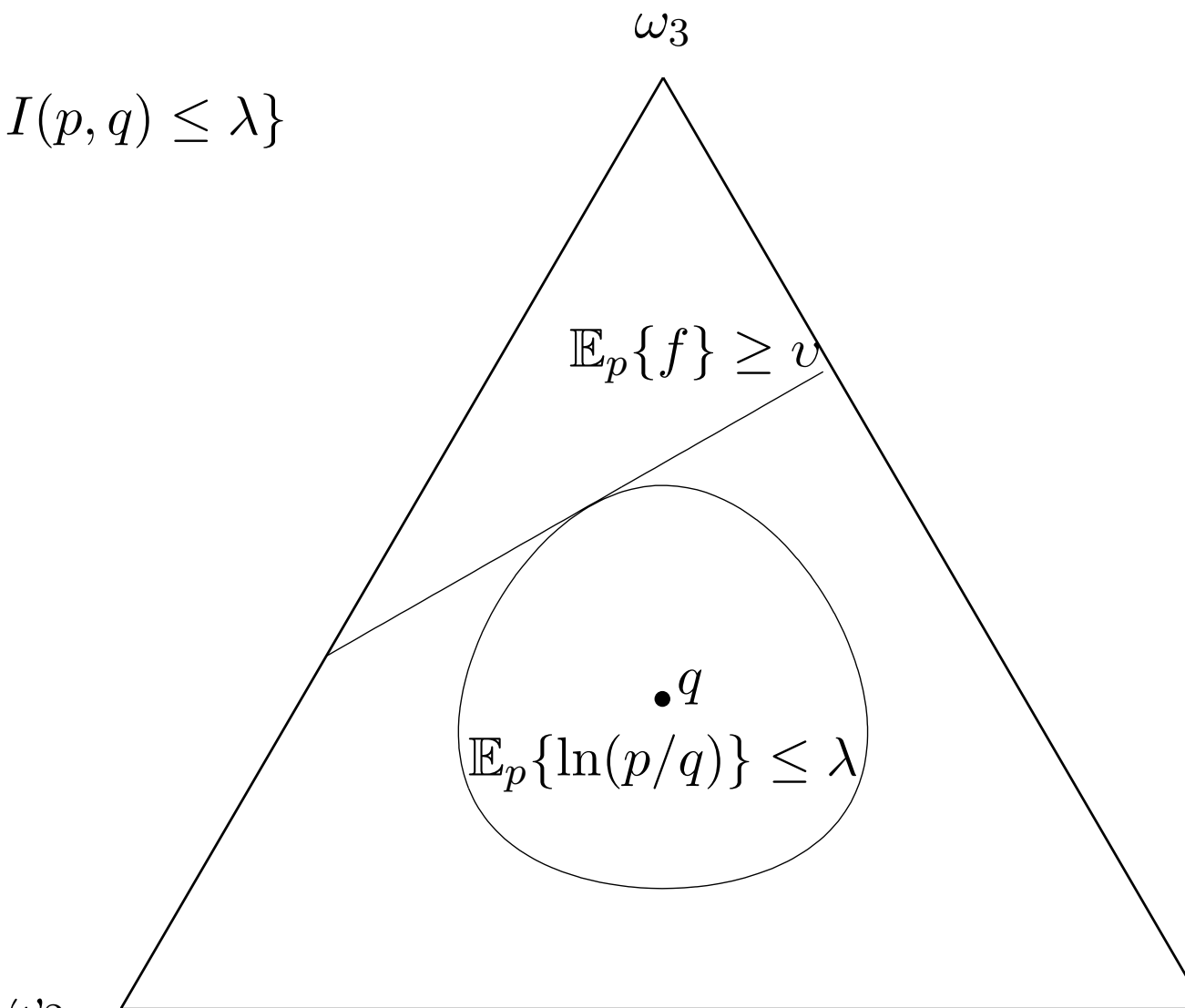
s.t. performance $\ge \upsilon$



First variational problem

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

$$v(\lambda) := \sup\{\mathbb{E}_p\{u\} : I(p,q) \le \lambda\}$$



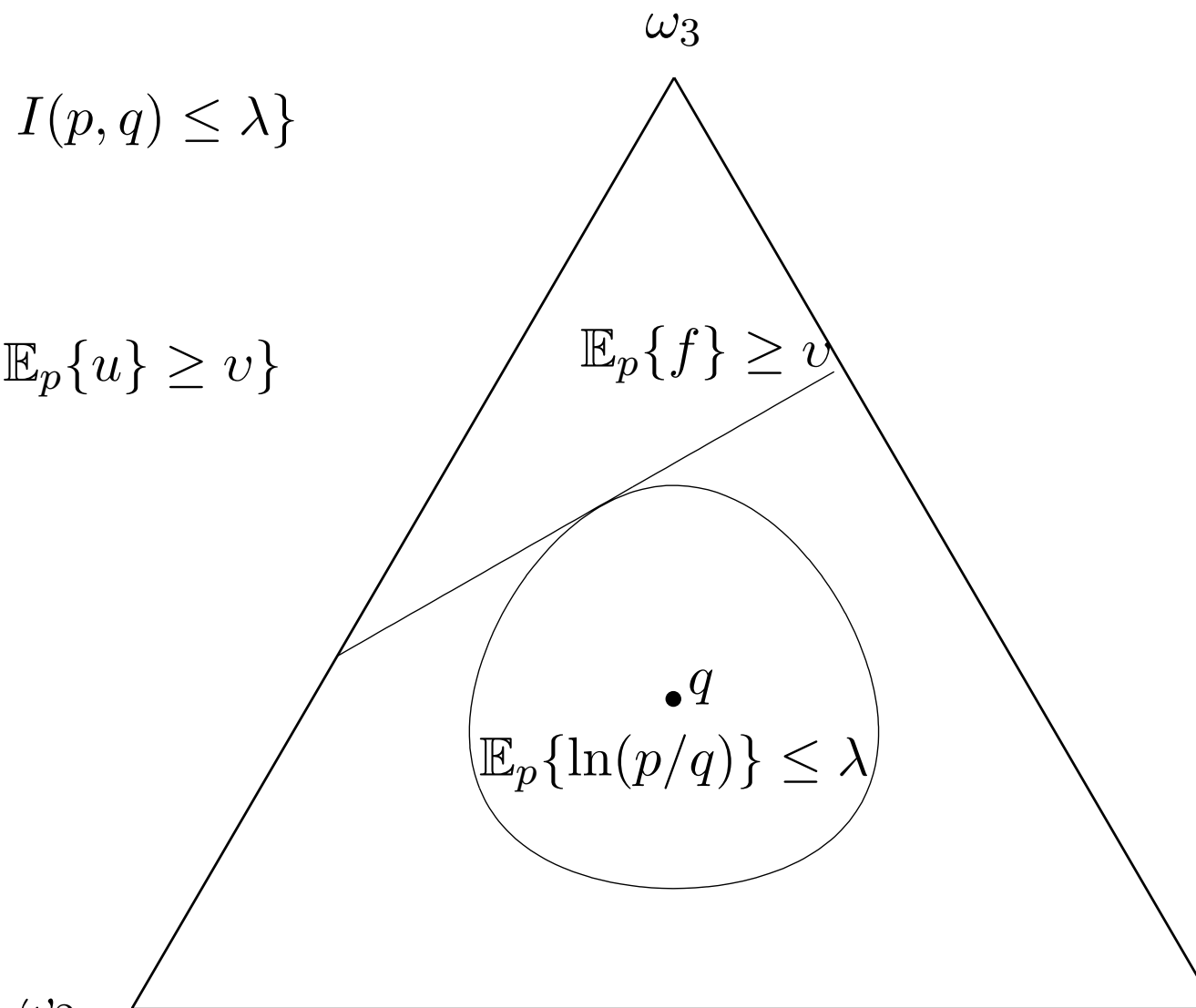
First variational problem

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

  $$v(\lambda) := \sup\{\mathbb{E}_p\{u\} : I(p,q) \le \lambda\}$$

- Minimize $I(p,q)$:

  $$\lambda(v) := \inf\{I(p,q) : \mathbb{E}_p\{u\} \ge v\}$$



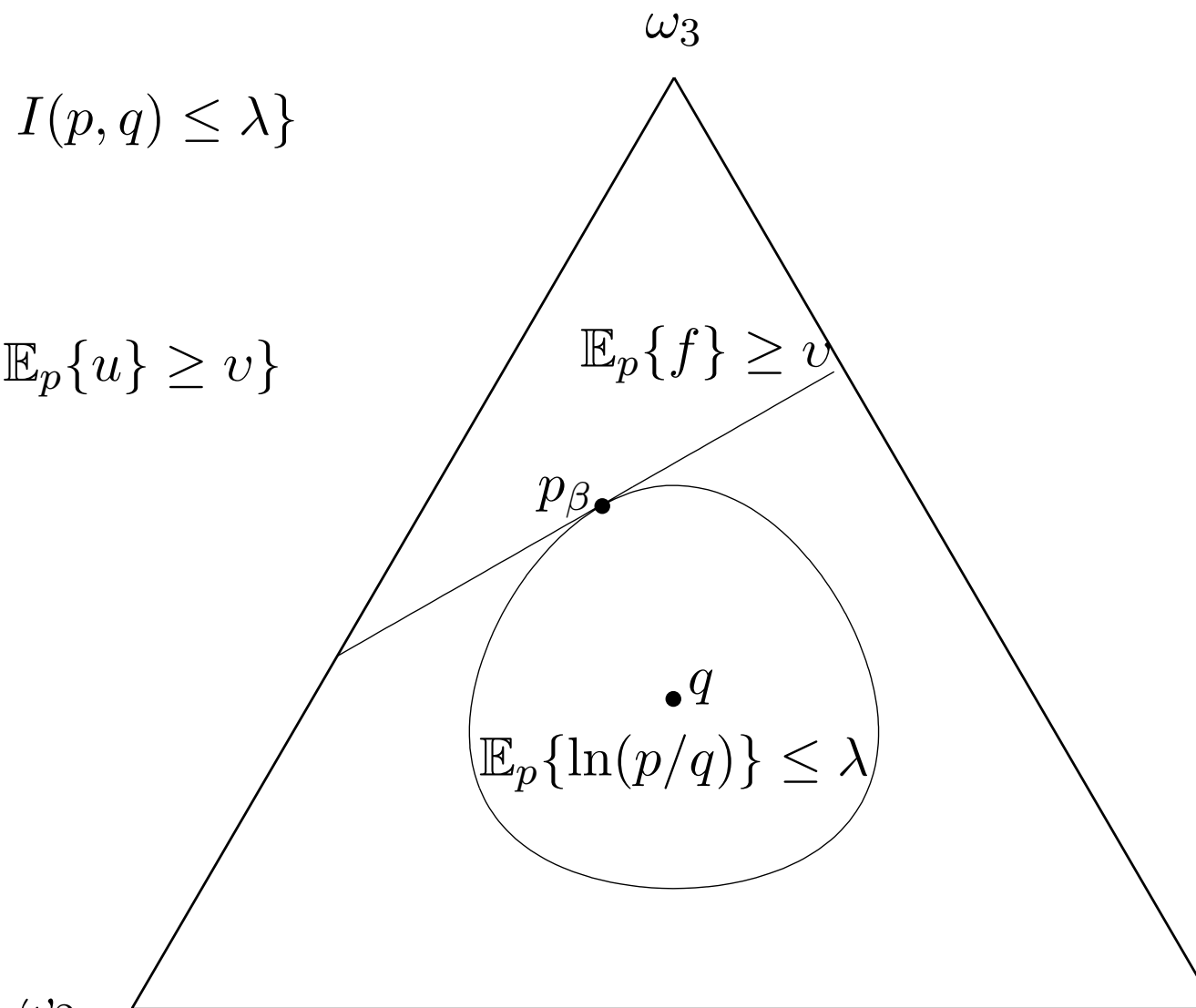
First variational problem

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

  $$\upsilon(\lambda) := \sup\{\mathbb{E}_p\{u\} : I(p,q) \le \lambda\}$$

- Minimize $I(p,q)$:

  $$\lambda(\upsilon) := \inf\{I(p,q) : \mathbb{E}_p\{u\} \ge \upsilon\}$$



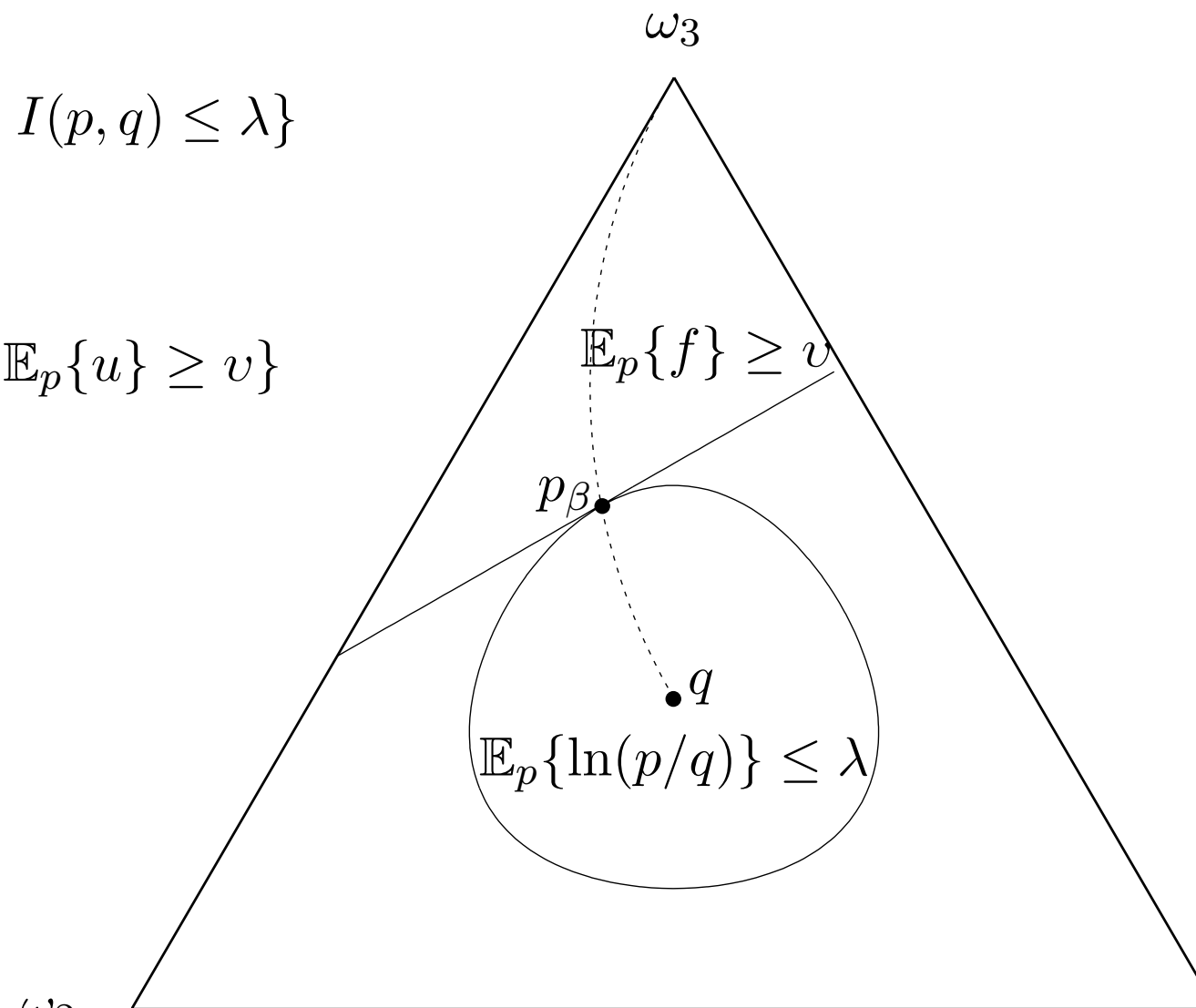
First variational problem

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

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

- Minimize $I(p,q)$:

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



Solution

- Lagrange function

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

Solution

- Lagrange function

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

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

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

Solution

- Lagrange function

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

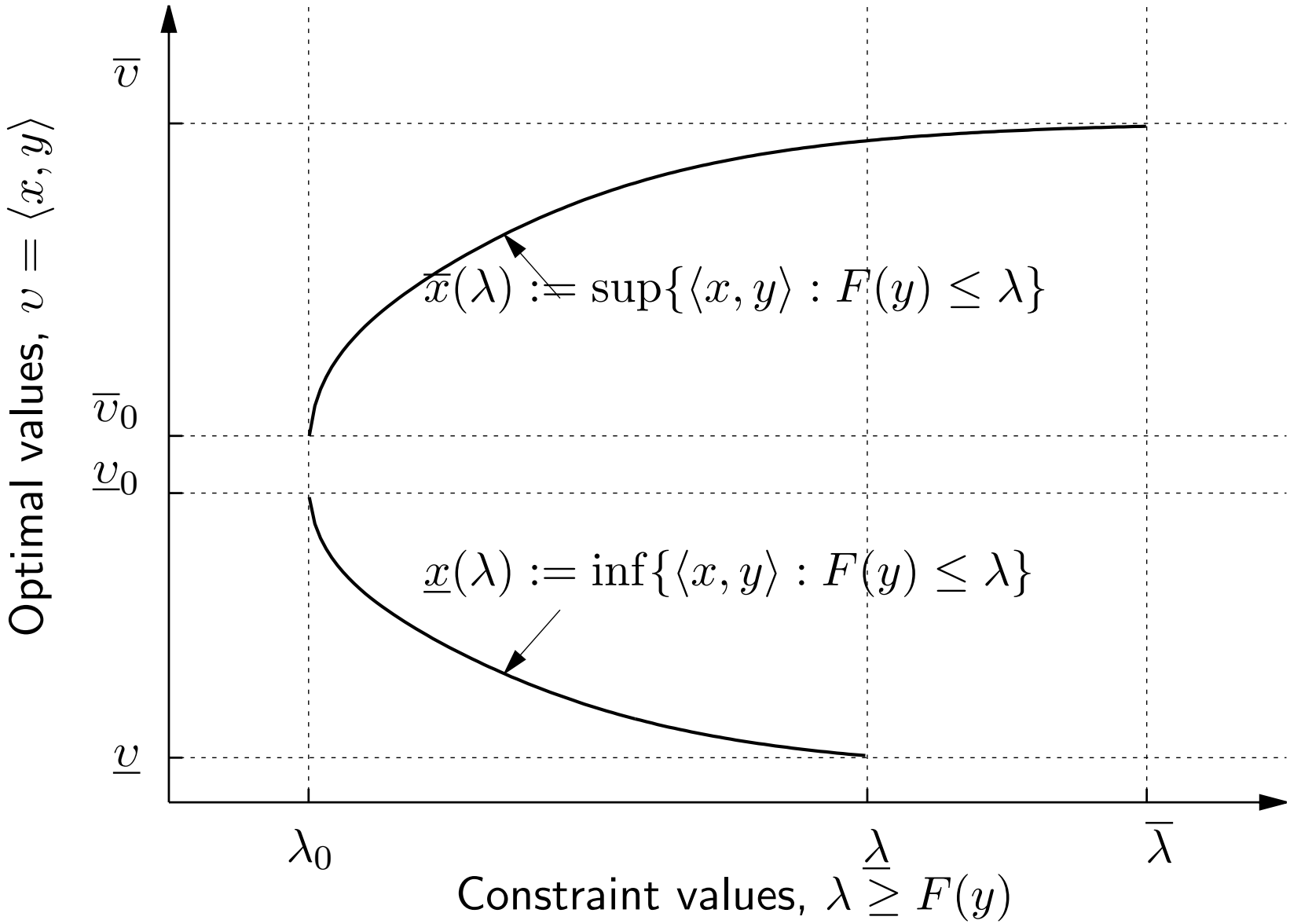
- Necessary and sufficient conditions $\nabla K(p,\beta) = 0$:

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

- Optimal solutions:

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

Concave and convex value functions



Value of information

- x — hidden, y — observed, z — control, $u(x, z)$ — utility:

Value of information

- x — hidden, y — observed, z — control, $u(x, z)$ — utility:
-

$$P(x) \qquad \mathbb{E}\{u(x, z) \mid z\}$$

Value of information

- x — hidden, y — observed, z — control, $u(x, z)$ — utility:
-

$$P(x) \qquad \sup_z \mathbb{E}\{u(x, z) \mid z\}$$

Value of information

- x — hidden, y — observed, z — control, $u(x,z)$ — utility:
-

$$P(x) \qquad \sup_z \mathbb{E}\{u(x,z) \mid z\} =: \overline{u}(0)$$

Value of information

- x — hidden, y — observed, z — control, $u(x, z)$ — utility:
-

$$x = f^{-1}(y)$$

$$P(x) \qquad \sup_z \mathbb{E}\{u(x, z) \mid z\} =: \overline{u}(0)$$

Value of information

- x — hidden, y — observed, z — control, $u(x,z)$ — utility:
-

$$x = f^{-1}(y) \qquad\qquad u(x,z)$$

$$P(x) \qquad\qquad \sup_z \mathbb{E}\{u(x,z) \mid z\} =: \overline{u}(0)$$

Value of information

- x — hidden, y — observed, z — control, $u(x, z)$ — utility:
-

$$x = f^{-1}(y) \qquad \sup_{z(x)} u(x, z)$$

$$P(x) \qquad \sup_{z} \mathbb{E}\{u(x, z) \mid z\} =: \overline{u}(0)$$

Value of information

- x — hidden, y — observed, z — control, $u(x,z)$ — utility:
-

$$x = f^{-1}(y) \qquad \sup_{z(x)} u(x,z) =: \overline{u}(\infty)$$

$$P(x) \qquad \sup_{z} \mathbb{E}\{u(x,z) \mid z\} =: \overline{u}(0)$$

Value of information

- x — hidden, y — observed, z — control, $u(x, z)$ — utility:
-

$$x = f^{-1}(y) \qquad \sup_{z(x)} u(x, z) =: \overline{u}(\infty)$$

$$P(x \mid y)$$

$$P(x) \qquad \sup_{z} \mathbb{E}\{u(x, z) \mid z\} =: \overline{u}(0)$$

Value of information

- x — hidden, y — observed, z — control, $u(x, z)$ — utility:
-

$$x = f^{-1}(y) \qquad \sup_{z(x)} u(x, z) =: \overline{u}(\infty)$$

$$P(x \mid y) \qquad \mathbb{E}\{u(x, z) \mid z, y\}$$

$$P(x) \qquad \sup_{z} \mathbb{E}\{u(x, z) \mid z\} =: \overline{u}(0)$$

Value of information

- x — hidden, y — observed, z — control, $u(x, z)$ — utility:
-

$$x = f^{-1}(y) \qquad \sup_{z(x)} u(x, z) =: \overline{u}(\infty)$$

$$P(x \mid y) \qquad \sup_{P(z|x):I\{x,z\}\leq\lambda} \mathbb{E}\{u(x, z) \mid z, y\}$$

$$P(x) \qquad \sup_{z} \mathbb{E}\{u(x, z) \mid z\} =: \overline{u}(0)$$

Value of information

- x — hidden, y — observed, z — control, $u(x, z)$ — utility:
-

$$x = f^{-1}(y) \qquad \sup_{z(x)} u(x, z) =: \overline{u}(\infty)$$

$$P(x \mid y) \qquad \sup_{P(z|x):I\{x,z\}\leq\lambda} \mathbb{E}\{u(x, z) \mid z, y\} =: \overline{u}(\lambda)$$

$$P(x) \qquad \sup_{z} \mathbb{E}\{u(x, z) \mid z\} =: \overline{u}(0)$$

Value of information

- x — hidden, y — observed, z — control, $u(x,z)$ — utility:
-

$$x = f^{-1}(y) \qquad \sup_{z(x)} u(x,z) =: \overline{u}(\infty)$$

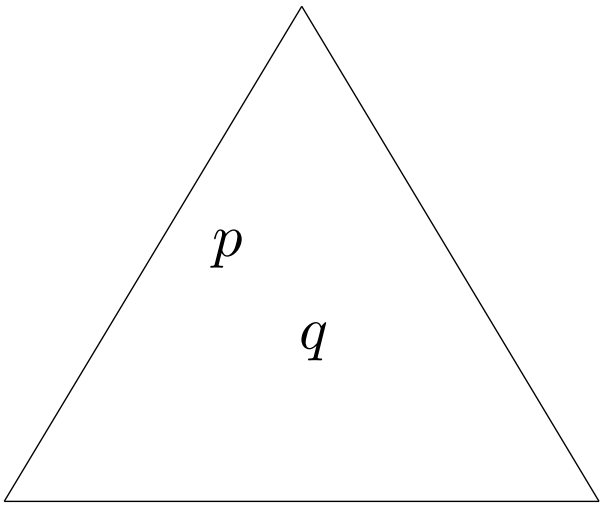
$$P(x \mid y) \qquad \sup_{P(z|x):I\{x,z\}\leq\lambda} \mathbb{E}\{u(x,z) \mid z,y\} =: \overline{u}(\lambda)$$

$$P(x) \qquad \sup_{z} \mathbb{E}\{u(x,z) \mid z\} =: \overline{u}(0)$$

Value of Information (Stratonovich, 1965)

$$V(\lambda) := \overline{u}(\lambda) - \overline{u}(0) = \sup\{\mathbb{E}_w\{u\} : I\{x,z\} \leq \lambda\} - \text{const}$$

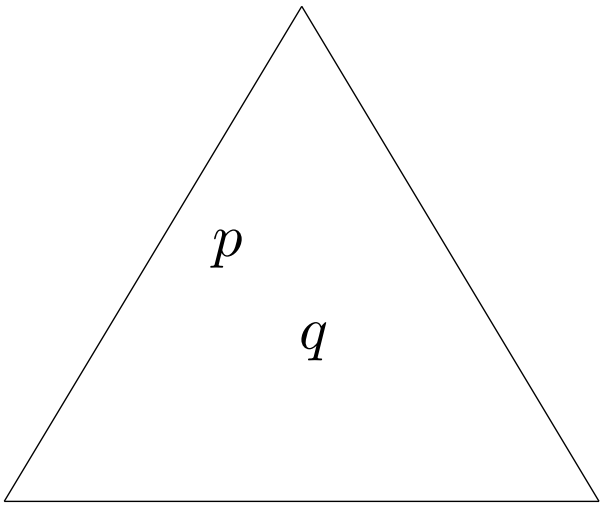
Variational problems for composite systems



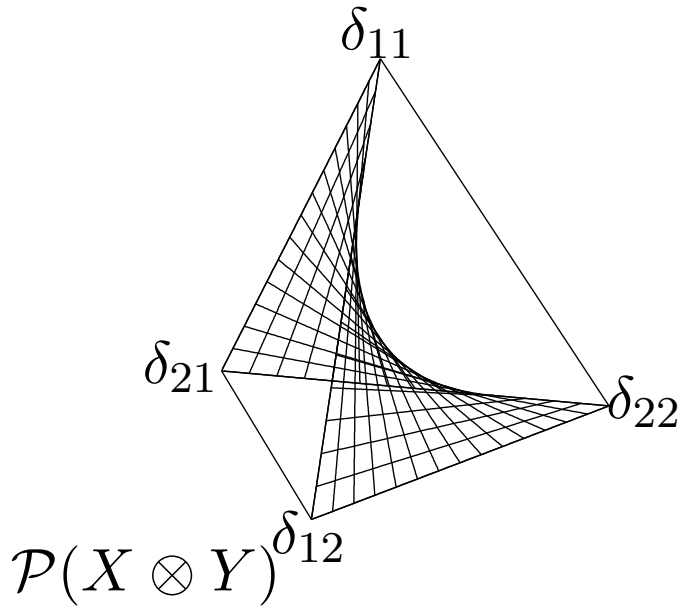
$\mathcal{P}(X)$

type I : Find optimal $p \in \mathcal{P}(X)$

Variational problems for composite systems



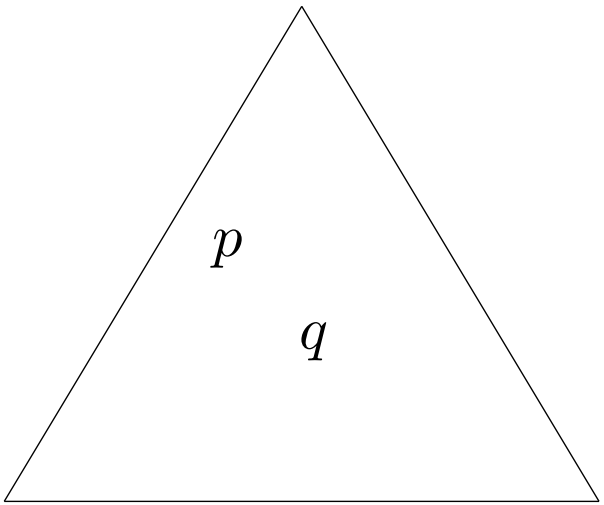
$\mathcal{P}(X)$



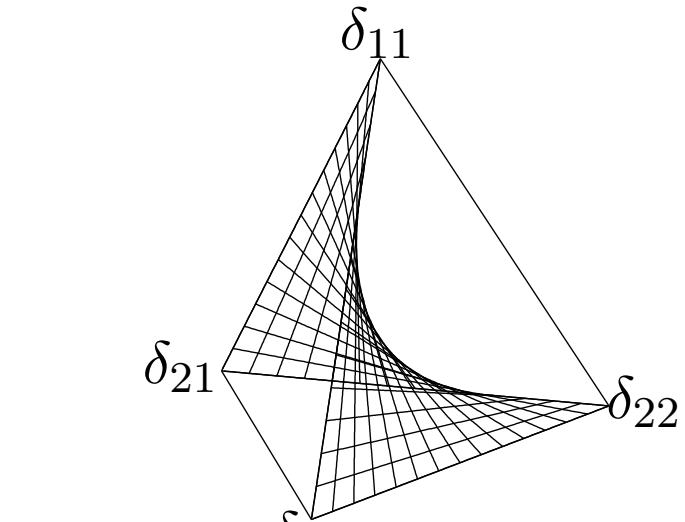
$\mathcal{P}(X \otimes Y)$

type I : Find optimal $p \in \mathcal{P}(X)$

Variational problems for composite systems



$\mathcal{P}(X)$

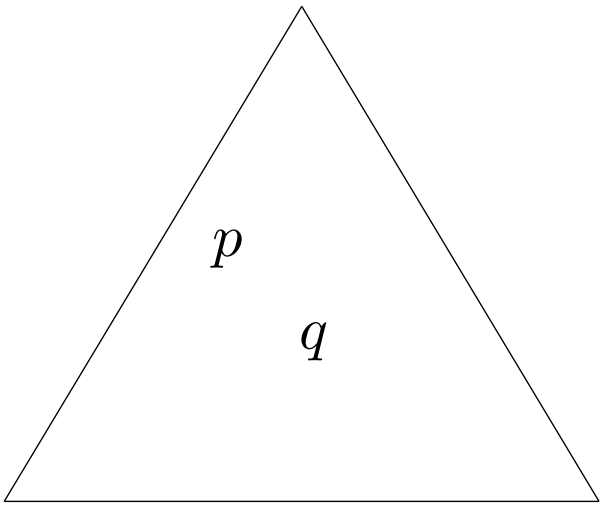


$\mathcal{P}(X \otimes Y)$

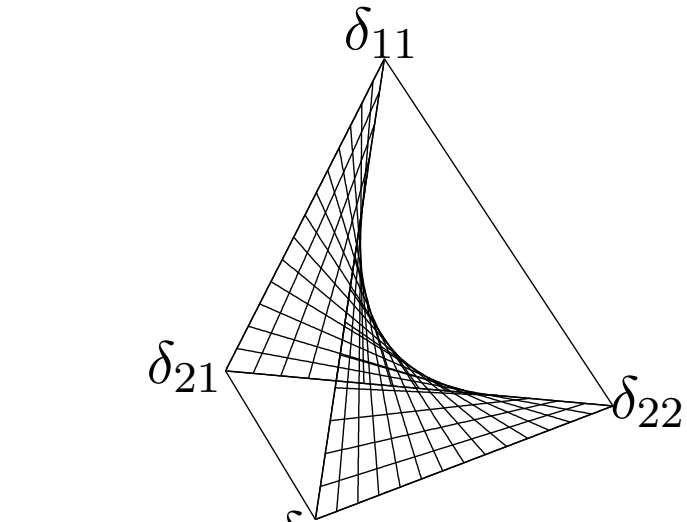
type I : Find optimal $p \in \mathcal{P}(X)$

type II : Find optimal input (marginal) $q \in \mathcal{P}(X)$ for fixed channel $\Lambda : \mathcal{P}(X) \to \mathcal{P}(Y)$.

Variational problems for composite systems



$\mathcal{P}(X)$



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

type I : Find optimal $p \in \mathcal{P}(X)$

type II : Find optimal input (marginal) $q \in \mathcal{P}(X)$ for fixed channel $\Lambda : \mathcal{P}(X) \to \mathcal{P}(Y)$.

type III : Find optimal channel $\Lambda : \mathcal{P}(X) \to \mathcal{P}(Y)$ for fixed marginal $q \in \mathcal{P}(X)$.

Solving Problems by Iterative Procedures

Optimization and Search Problems

Given a set X (a search space) and an objective function $f : X \to \mathbb{R}$, find $\overline{x}$ such that $f(\overline{x}) = \upsilon$ (e.g. $f(\overline{x}) = \sup f(x)$ or $\nabla f(\overline{x}) = 0$, $\nabla^2 f(\overline{x}) \leq 0$).

Solving Problems by Iterative Procedures

Optimization and Search Problems

Given a set X (a search space) and an objective function $f : X \to \mathbb{R}$, find $\overline{x}$ such that $f(\overline{x}) = \upsilon$ (e.g. $f(\overline{x}) = \sup f(x)$ or $\nabla f(\overline{x}) = 0$, $\nabla^2 f(\overline{x}) \leq 0$).

- The solution often involves an iteration $\Lambda : X \to X$:

$$x_1 \quad = \quad \Lambda(x_0)$$

Solving Problems by Iterative Procedures

Optimization and Search Problems

Given a set X (a search space) and an objective function $f : X \to \mathbb{R}$, find $\overline{x}$ such that $f(\overline{x}) = \upsilon$ (e.g. $f(\overline{x}) = \sup f(x)$ or $\nabla f(\overline{x}) = 0$, $\nabla^2 f(\overline{x}) \leq 0$).

- The solution often involves an iteration $\Lambda : X \to X$:

$$\begin{aligned} x_1 &= \Lambda(x_0) \\ x_2 &= \Lambda(x_1) \end{aligned}$$

Solving Problems by Iterative Procedures

Optimization and Search Problems

Given a set X (a search space) and an objective function $f : X \to \mathbb{R}$, find $\overline{x}$ such that $f(\overline{x}) = \upsilon$ (e.g. $f(\overline{x}) = \sup f(x)$ or $\nabla f(\overline{x}) = 0$, $\nabla^2 f(\overline{x}) \leq 0$).

- The solution often involves an iteration $\Lambda : X \to X$:

$$\begin{aligned} x_1 &= \Lambda(x_0) \\ x_2 &= \Lambda(x_1) \\ &\vdots \\ x_{t+1} &= \Lambda(x_t) \end{aligned}$$

Solving Problems by Iterative Procedures

Optimization and Search Problems

Given a set X (a search space) and an objective function $f : X \to \mathbb{R}$, find $\overline{x}$ such that $f(\overline{x}) = \upsilon$ (e.g. $f(\overline{x}) = \sup f(x)$ or $\nabla f(\overline{x}) = 0$, $\nabla^2 f(\overline{x}) \leq 0$).

- The solution often involves an iteration $\Lambda : X \to X$:

$$\begin{aligned} x_1 &= \Lambda(x_0) \\ x_2 &= \Lambda(x_1) \\ &\vdots \\ x_{t+1} &= \Lambda(x_t) \end{aligned}$$

Example (Gradient descent)

$$\Lambda(x_t) = x_t + \Delta x\,, \qquad \Delta x = -[\nabla^2 f(x_t)]^{-1} \nabla f(x_t) + \mathrm{rand}(X)$$

Solving Problems by Iterative Procedures

Optimization and Search Problems

Given a set X (a search space) and an objective function $f : X \to \mathbb{R}$, find $\overline{x}$ such that $f(\overline{x}) = \upsilon$ (e.g. $f(\overline{x}) = \sup f(x)$ or $\nabla f(\overline{x}) = 0$, $\nabla^2 f(\overline{x}) \leq 0$).

- The solution often involves an iteration $\Lambda : X \to X$:

$$\begin{aligned} x_1 &= \Lambda(x_0) \\ x_2 &= \Lambda(x_1) \\ &\vdots \\ x_{t+1} &= \Lambda(x_t) \end{aligned}$$

Example (Gradient descent)

$$\Lambda(x_t) = x_t + \Delta x\,, \qquad \Delta x = -[\nabla^2 f(x_t)]^{-1} \nabla f(x_t) + \mathrm{rand}(X)$$

Remark (Markov property)

Observe that $x_{t+1} = \Lambda(x_t)$ does not depend on $x_{t-1}, \ldots, x_1, x_0$.

Iterative procedures as Markov morphisms

Definition (Markov transition kernel from $(X, \mathcal{X})$ to $(Y, \mathcal{Y})$)

A conditional probability measure $p(Y_i \mid x)$ on $(Y, \mathcal{Y})$ that is $\mathcal{X}$-measurable for each $Y_i \in \mathcal{Y}$.

Iterative procedures as Markov morphisms

Definition (Markov transition kernel from $(X, \mathcal{X})$ to $(Y, \mathcal{Y})$)

A conditional probability measure $p(Y_i \mid x)$ on $(Y, \mathcal{Y})$ that is $\mathcal{X}$-measurable for each $Y_i \in \mathcal{Y}$.

- $x_{t+1} = \Lambda(x_t)$ can be represented by $p(x_{t+1} \mid x_t)$.

Iterative procedures as Markov morphisms

Definition (Markov transition kernel from $(X, \mathcal{X})$ to $(Y, \mathcal{Y})$)

A conditional probability measure $p(Y_i \mid x)$ on $(Y, \mathcal{Y})$ that is $\mathcal{X}$-measurable for each $Y_i \in \mathcal{Y}$.

- $x_{t+1} = \Lambda(x_t)$ can be represented by $p(x_{t+1} \mid x_t)$.
- Markov morphism $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$p(x_{t+1}) = \underbrace{\sum_{x_t \in X} P(x_{t+1} \mid x_t)\ q(x_t)}_{\Lambda}$$

Iterative procedures as Markov morphisms

Definition (Markov transition kernel from $(X, \mathcal{X})$ to $(Y, \mathcal{Y})$)

A conditional probability measure $p(Y_i \mid x)$ on $(Y, \mathcal{Y})$ that is $\mathcal{X}$-measurable for each $Y_i \in \mathcal{Y}$.

- $x_{t+1} = \Lambda(x_t)$ can be represented by $p(x_{t+1} \mid x_t)$.
- Markov morphism $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$p(x_{t+1}) = \underbrace{\sum_{x_t \in X} P(x_{t+1} \mid x_t)\ q(x_t)}_{\Lambda}$$

Deterministic transitions are defined by a function $\phi : X \to X$:

$$p(x_{t+1} \mid x_t) = \delta_{\phi(x_t)}(x_{t+1}) := \begin{cases} 1 & \text{if } x_{t+1} = \phi(x_t) \\ 0 & \text{otherwise} \end{cases}$$

Iterative procedures as Markov morphisms

Definition (Markov transition kernel from $(X, \mathcal{X})$ to $(Y, \mathcal{Y})$)

A conditional probability measure $p(Y_i \mid x)$ on $(Y, \mathcal{Y})$ that is $\mathcal{X}$-measurable for each $Y_i \in \mathcal{Y}$.

- $x_{t+1} = \Lambda(x_t)$ can be represented by $p(x_{t+1} \mid x_t)$.
- Markov morphism $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$p(x_{t+1}) = \underbrace{\sum_{x_t \in X} P(x_{t+1} \mid x_t)\ q(x_t)}_{\Lambda}$$

Deterministic transitions are defined by a function $\phi : X \to X$:

$$p(x_{t+1} \mid x_t) = \delta_{\phi(x_t)}(x_{t+1}) := \begin{cases} 1 & \text{if } x_{t+1} = \phi(x_t) \\ 0 & \text{otherwise} \end{cases}$$

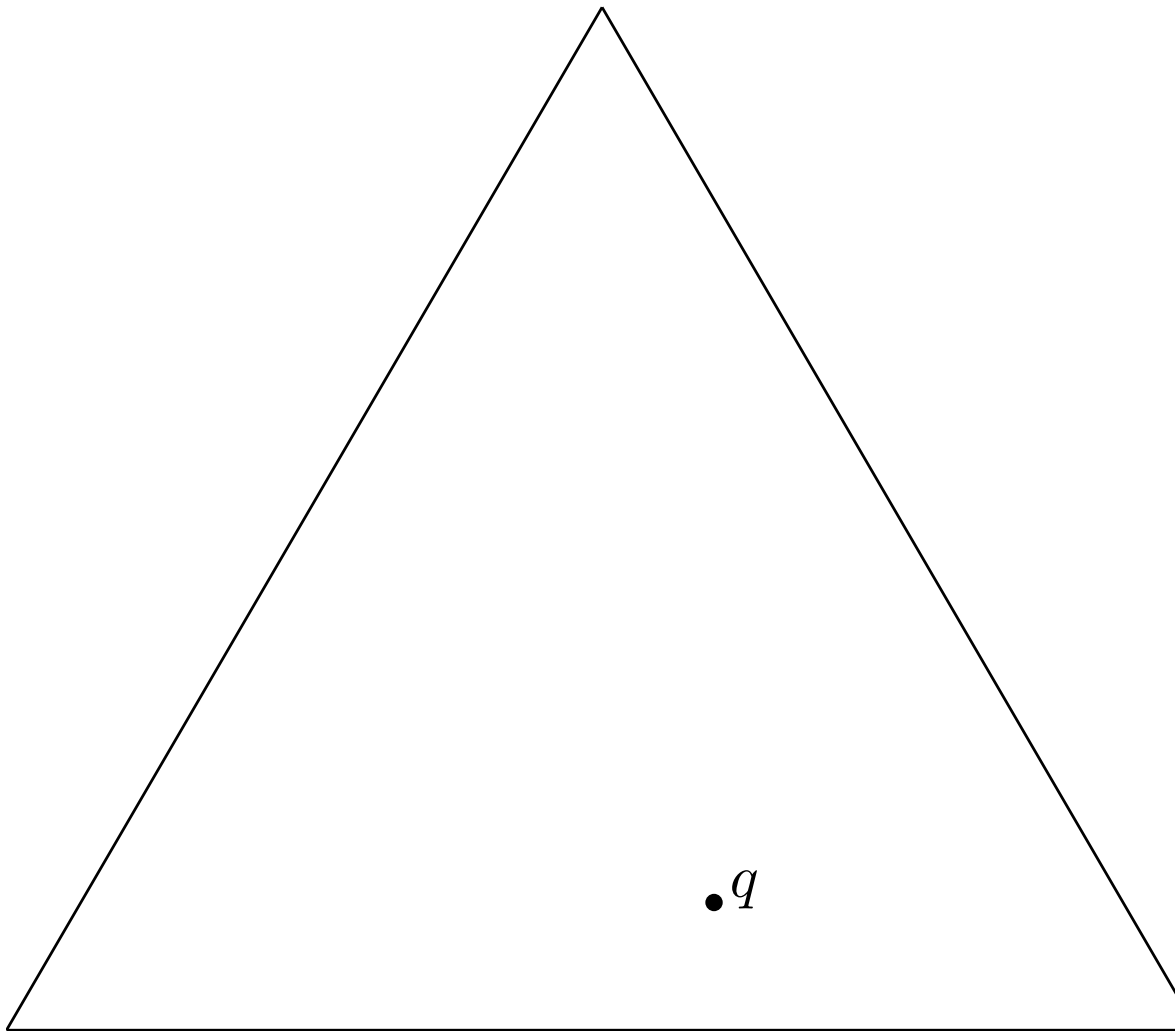
Randomized (probabilistic) transitions are such that $p(x_{t+1} \mid x_t) > 0$ for several x_{t+1} and some x_t.

Optimal learning trajectory

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

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

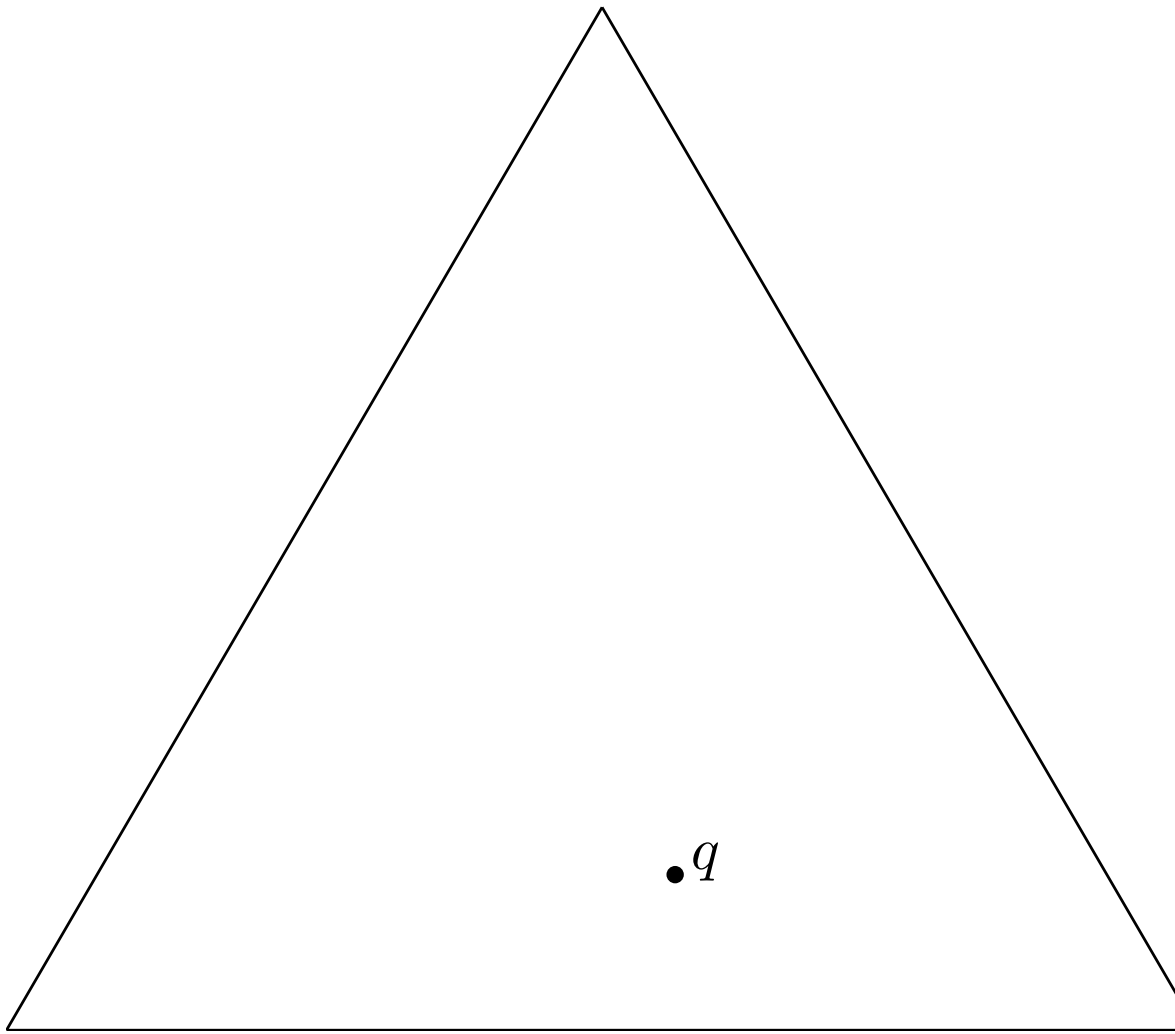


Optimal learning trajectory

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

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

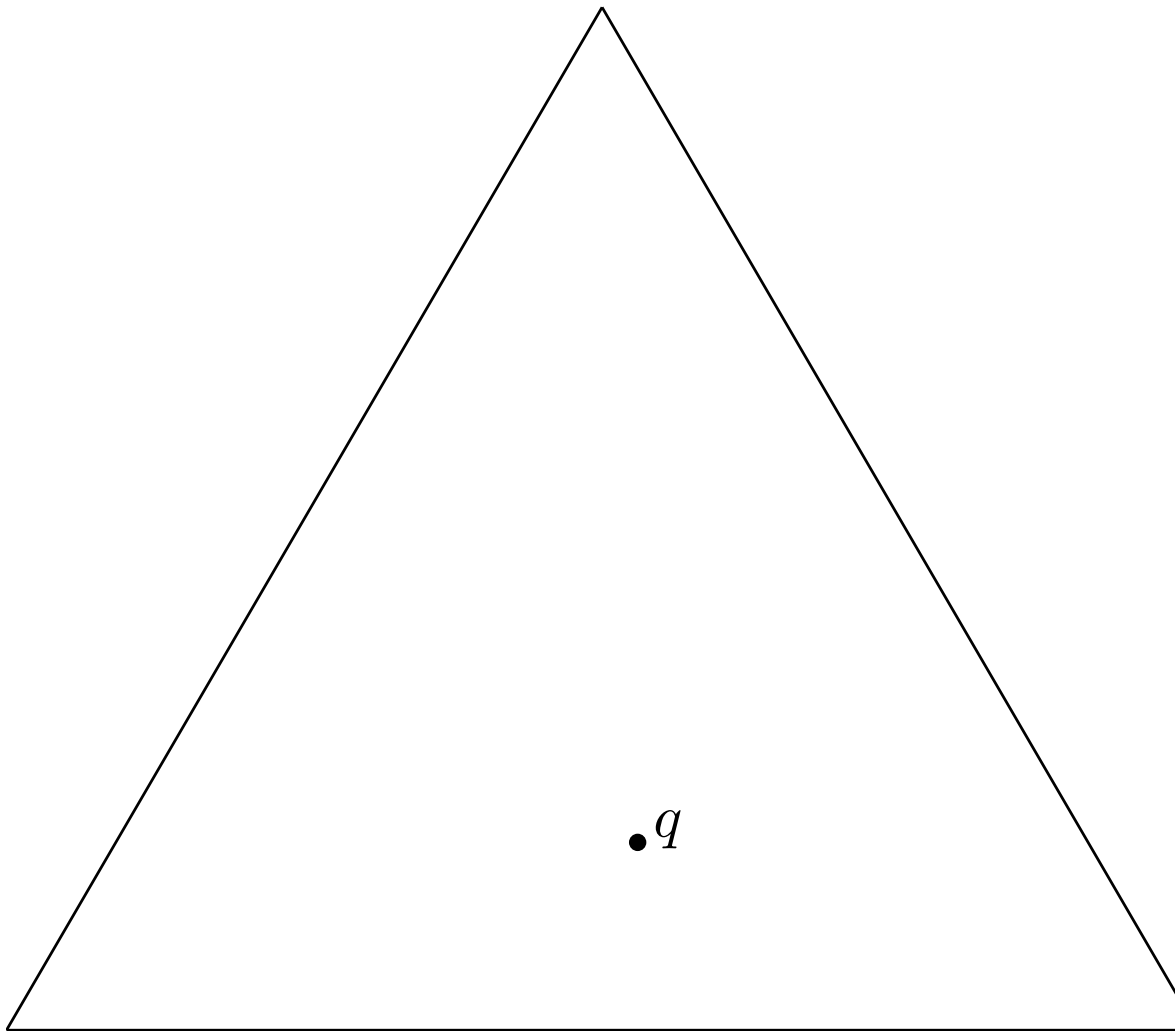


Optimal learning trajectory

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

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

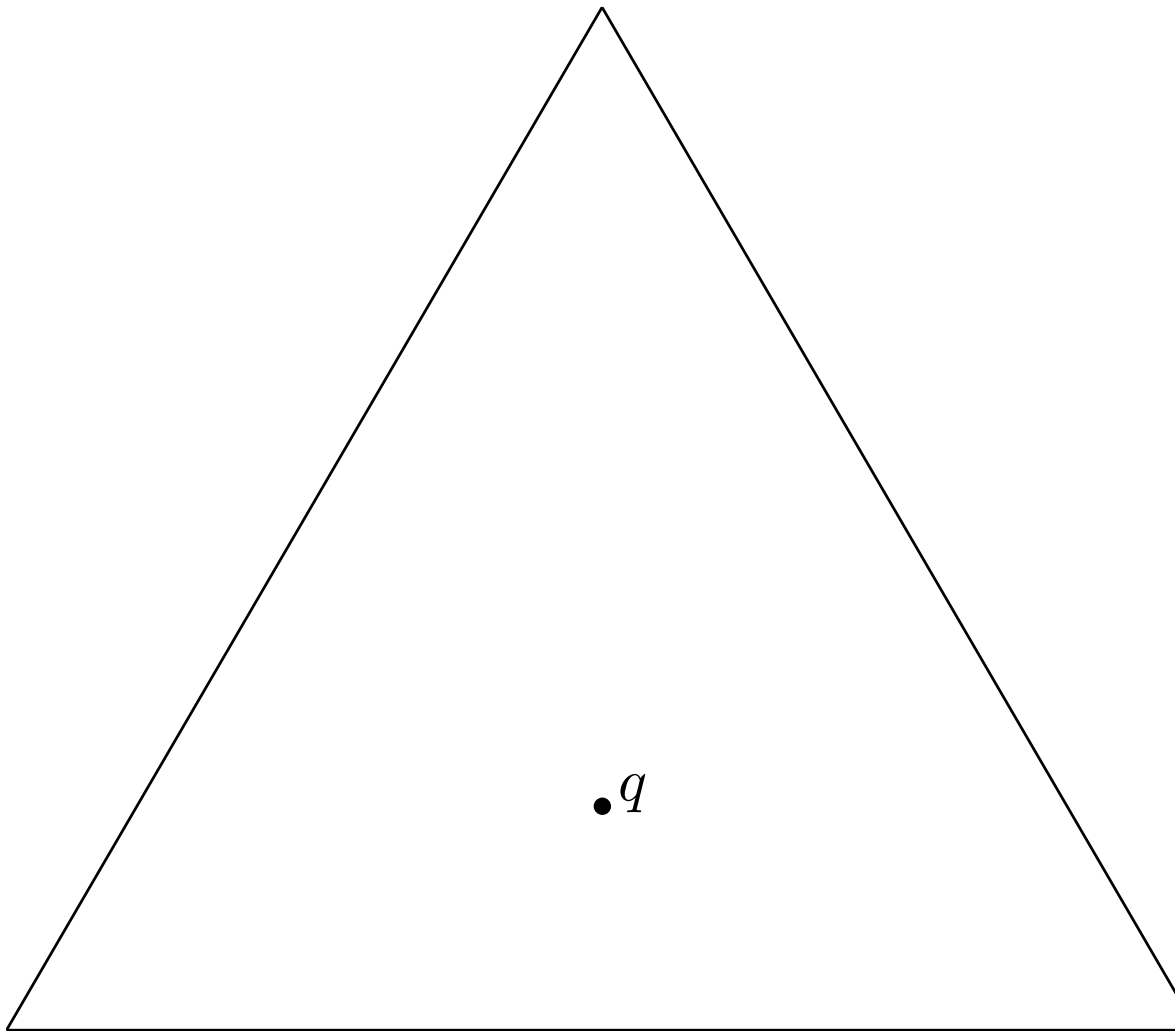


Optimal learning trajectory

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

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

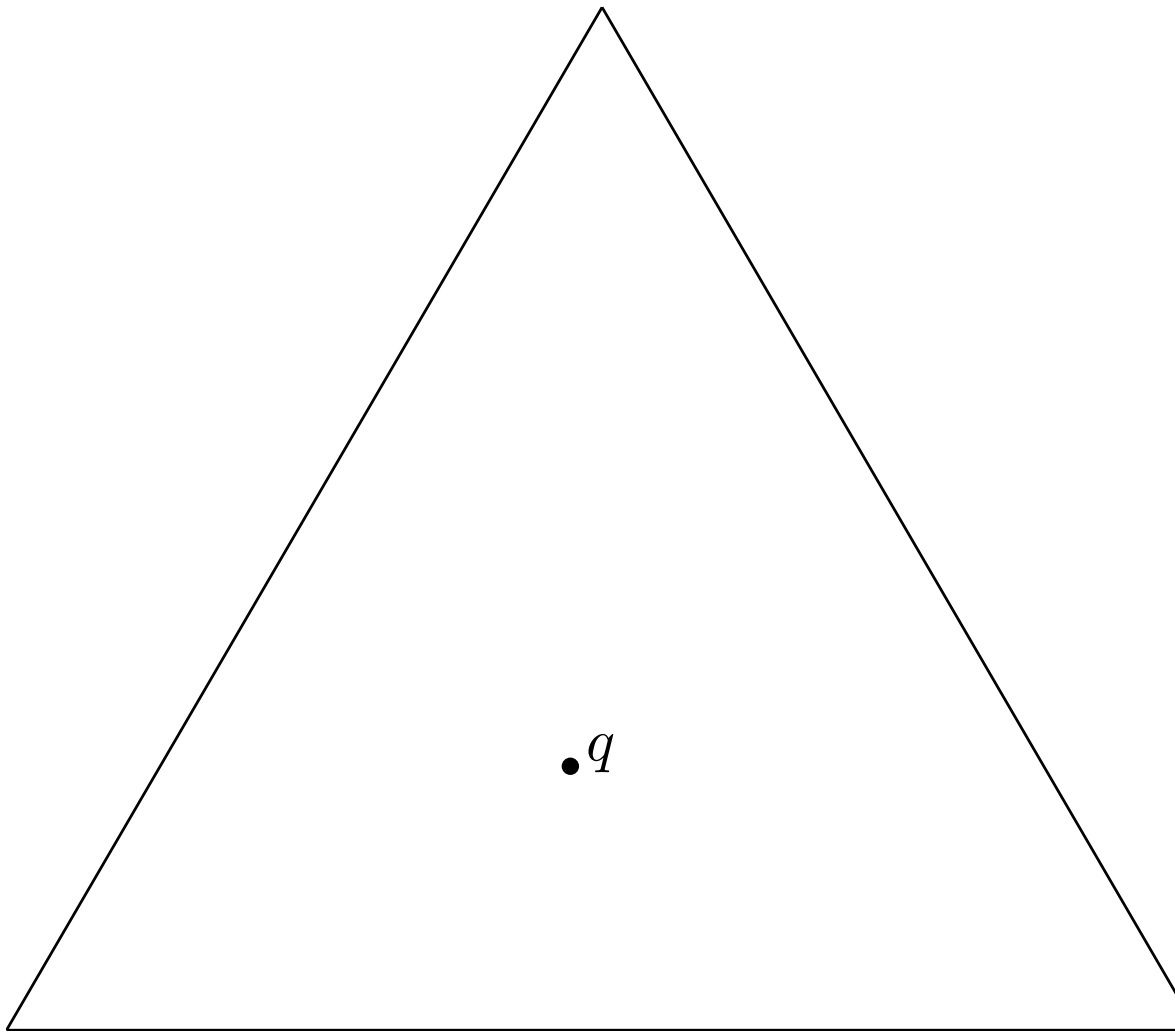


Optimal learning trajectory

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

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

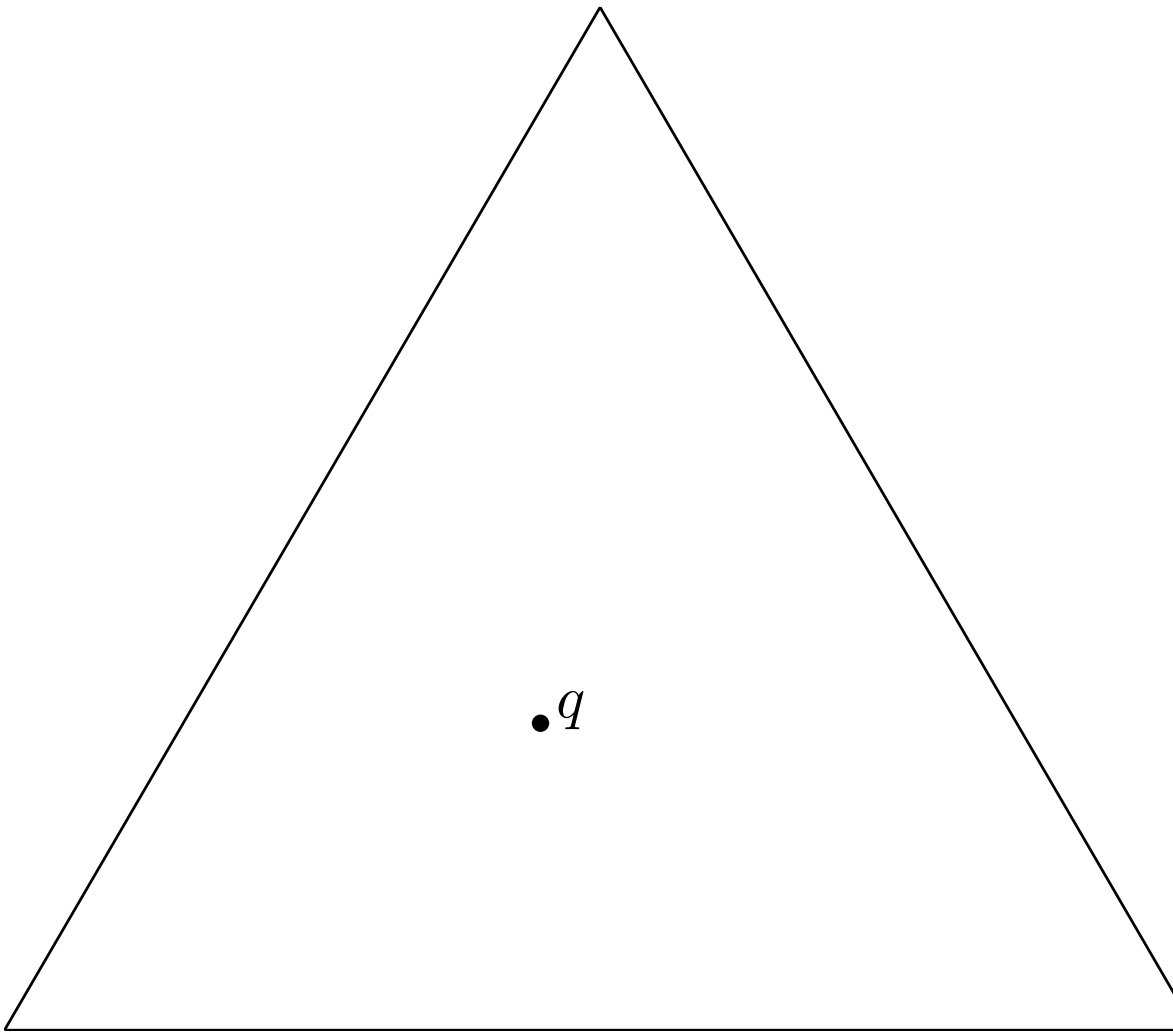


Optimal learning trajectory

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

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

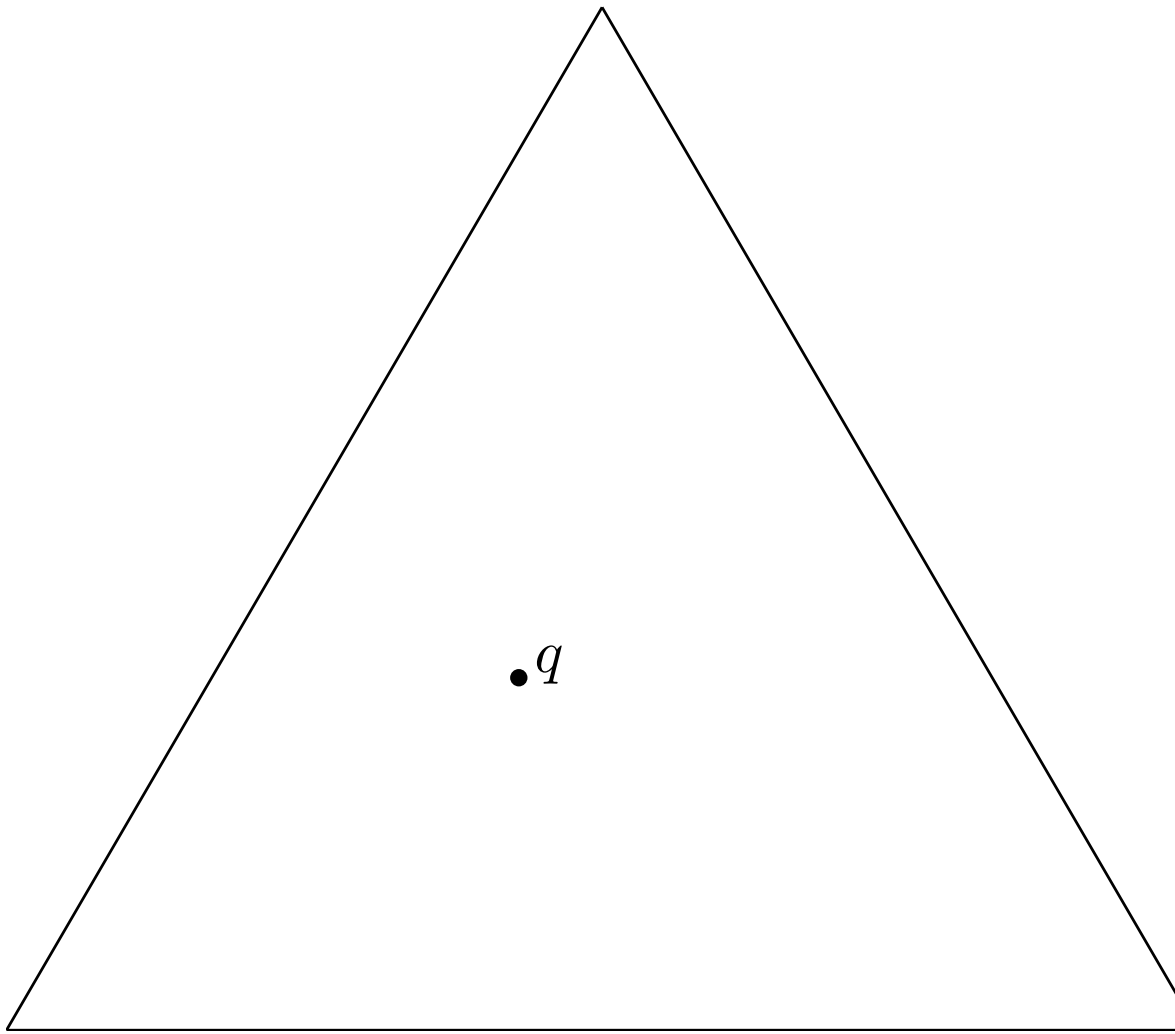


Optimal learning trajectory

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

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

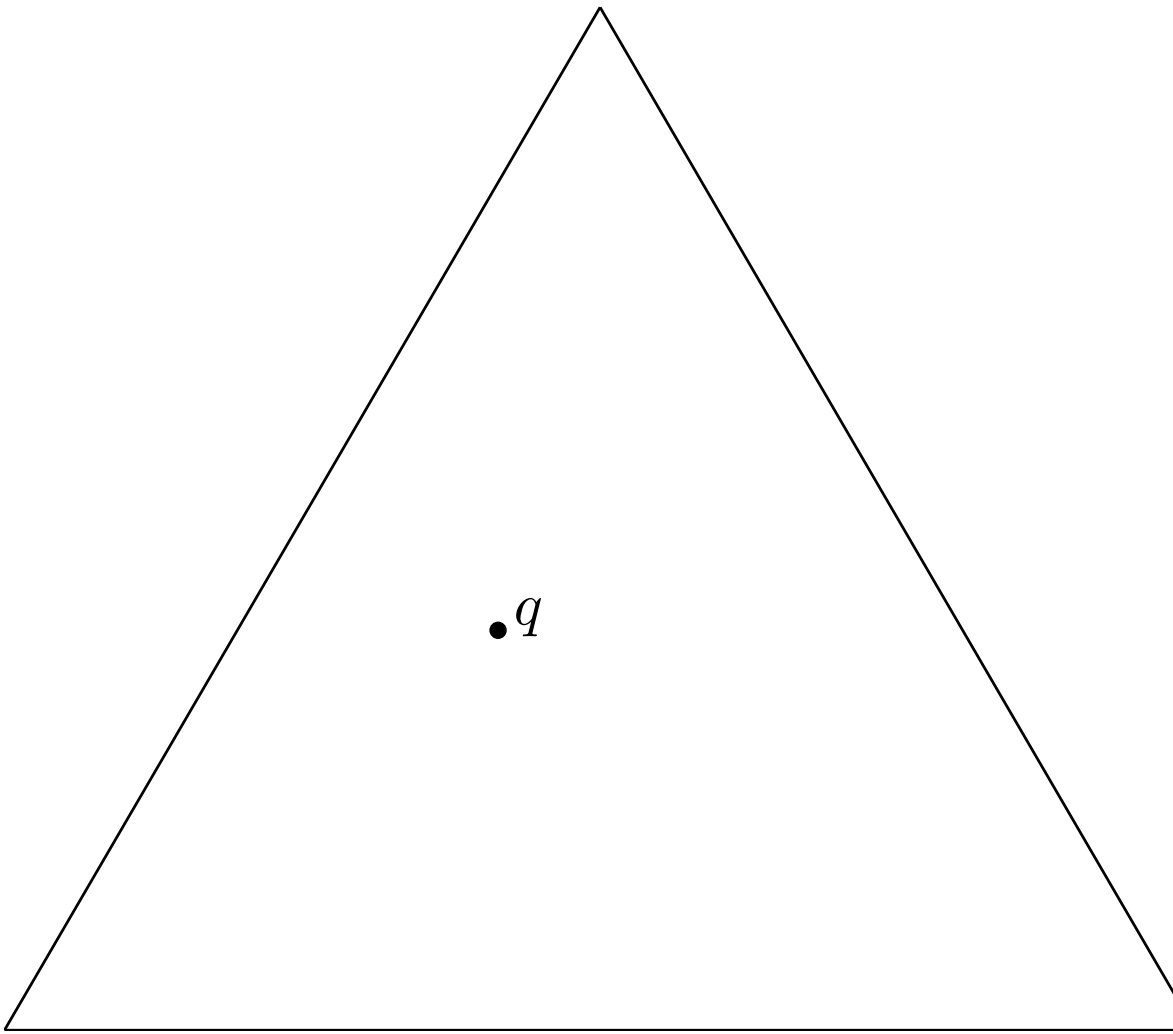


Optimal learning trajectory

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

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

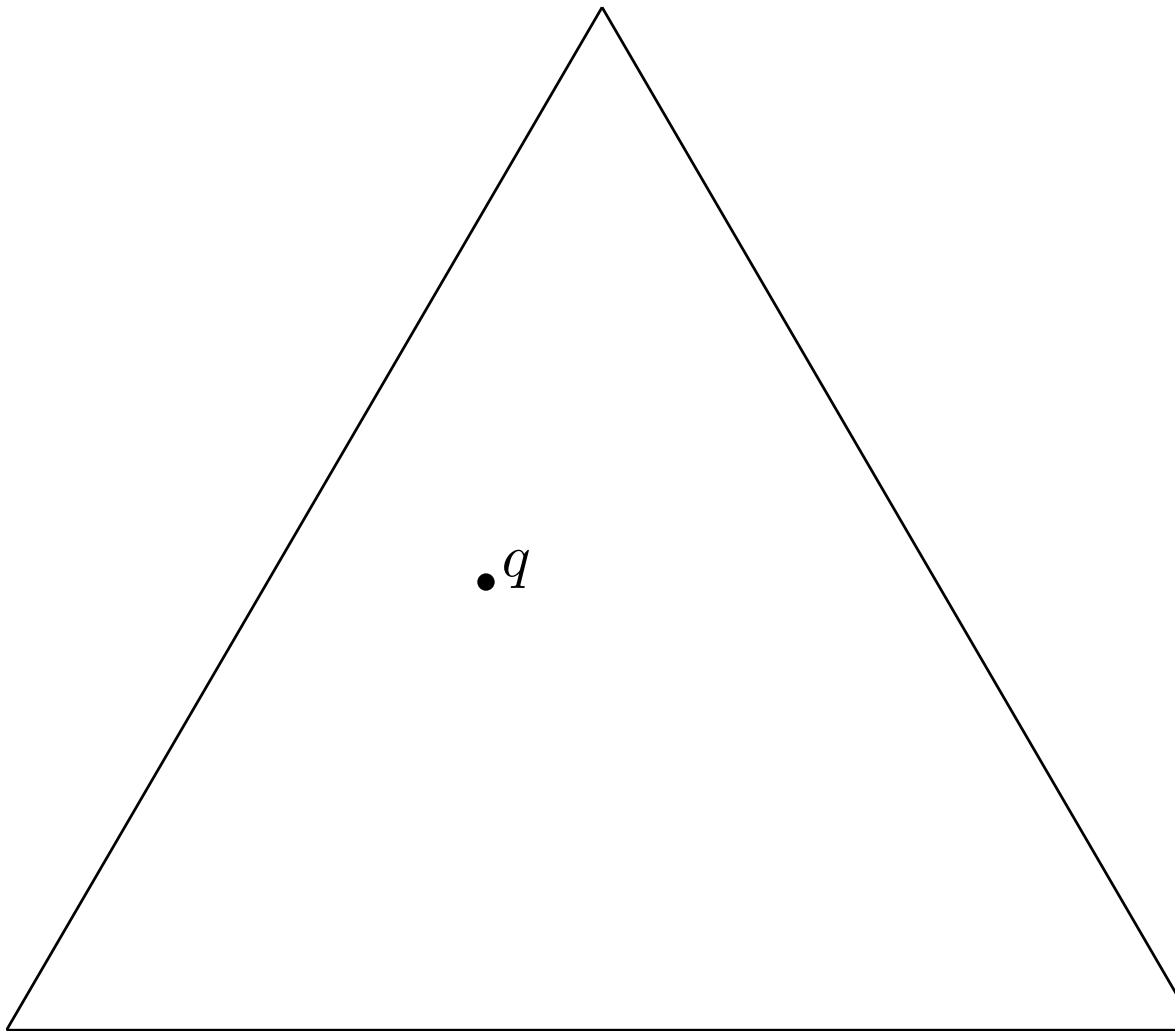


Optimal learning trajectory

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

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

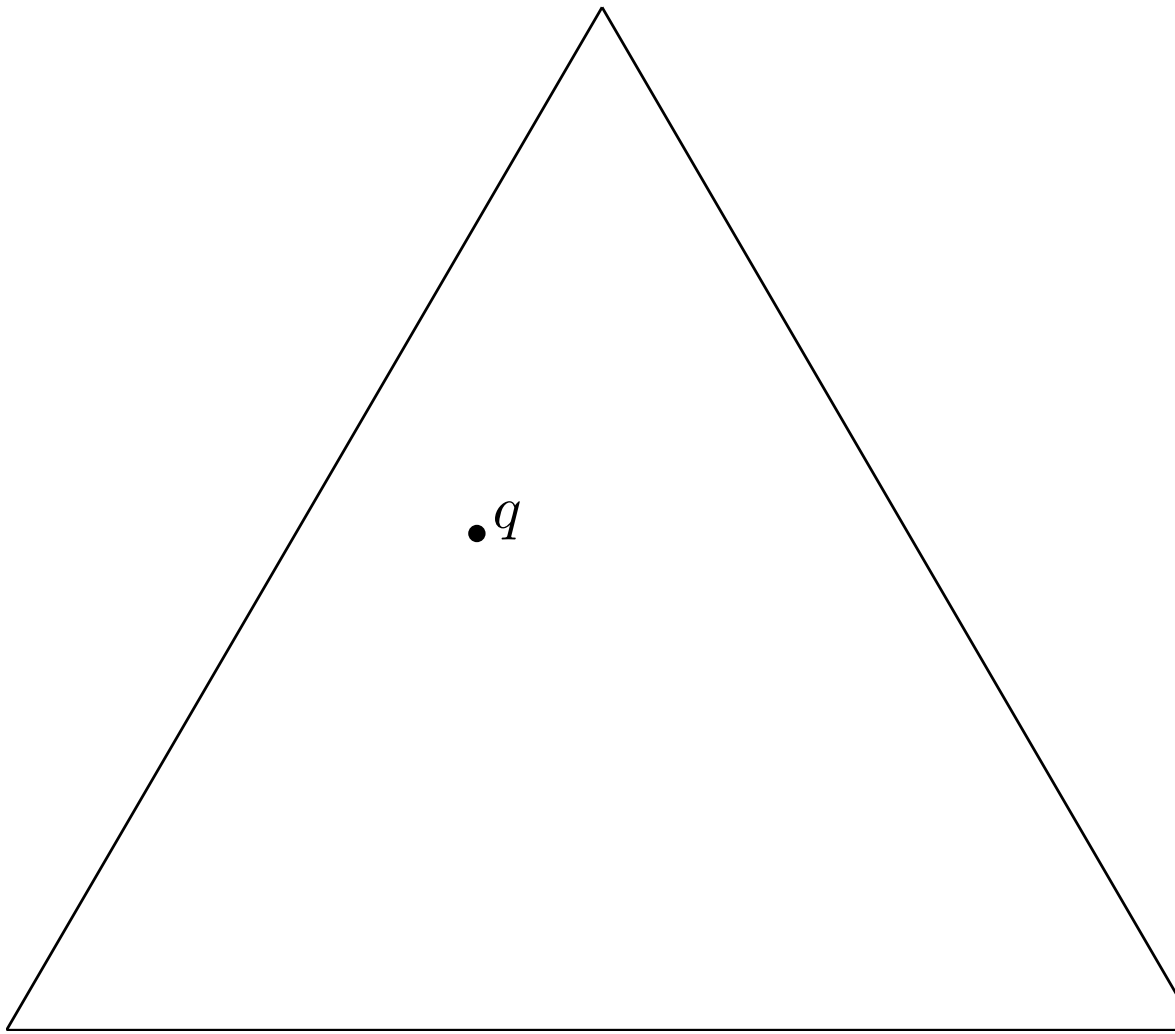


Optimal learning trajectory

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

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

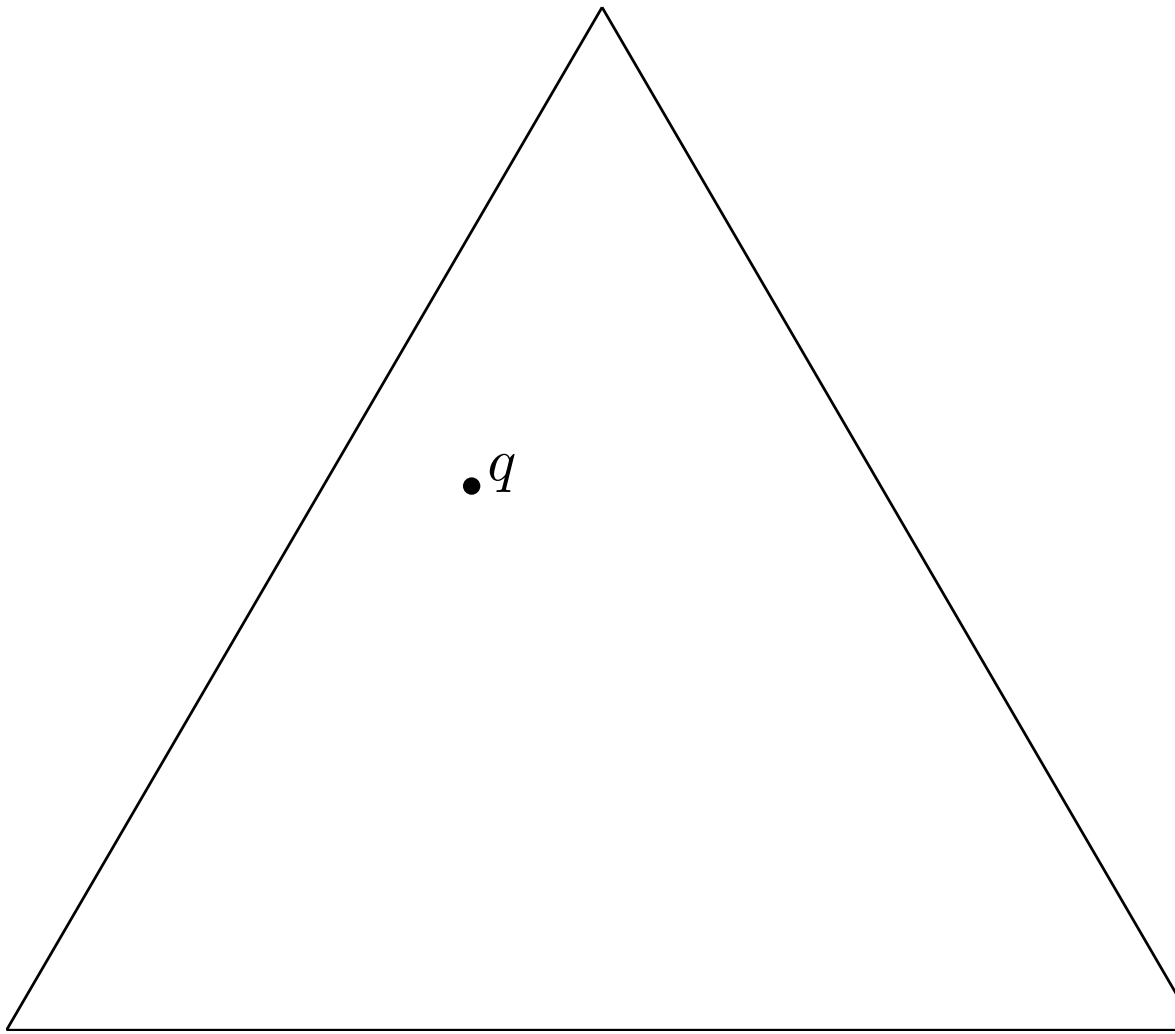


Optimal learning trajectory

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

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

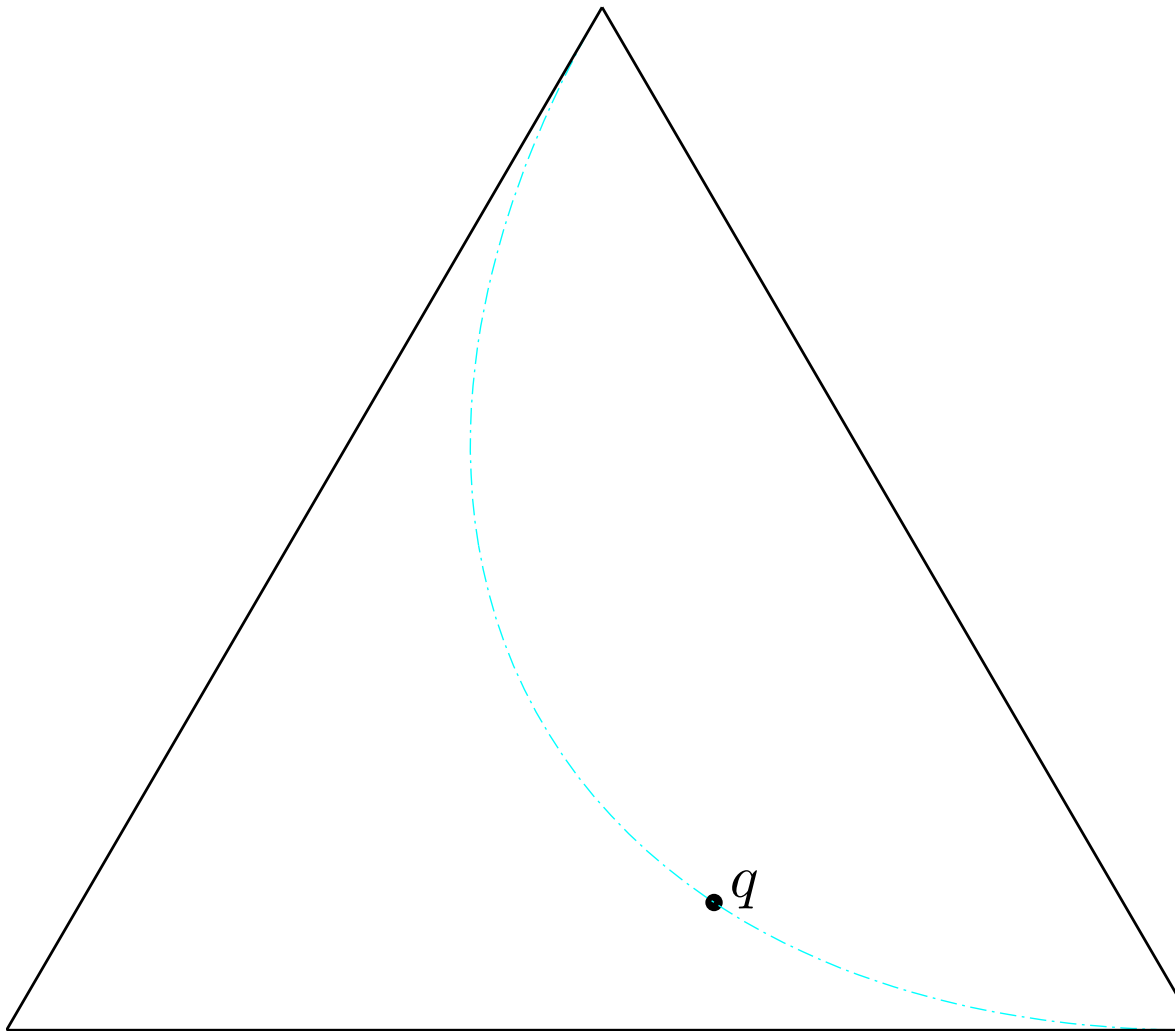


Optimal learning trajectory

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

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$



Optimal learning trajectory

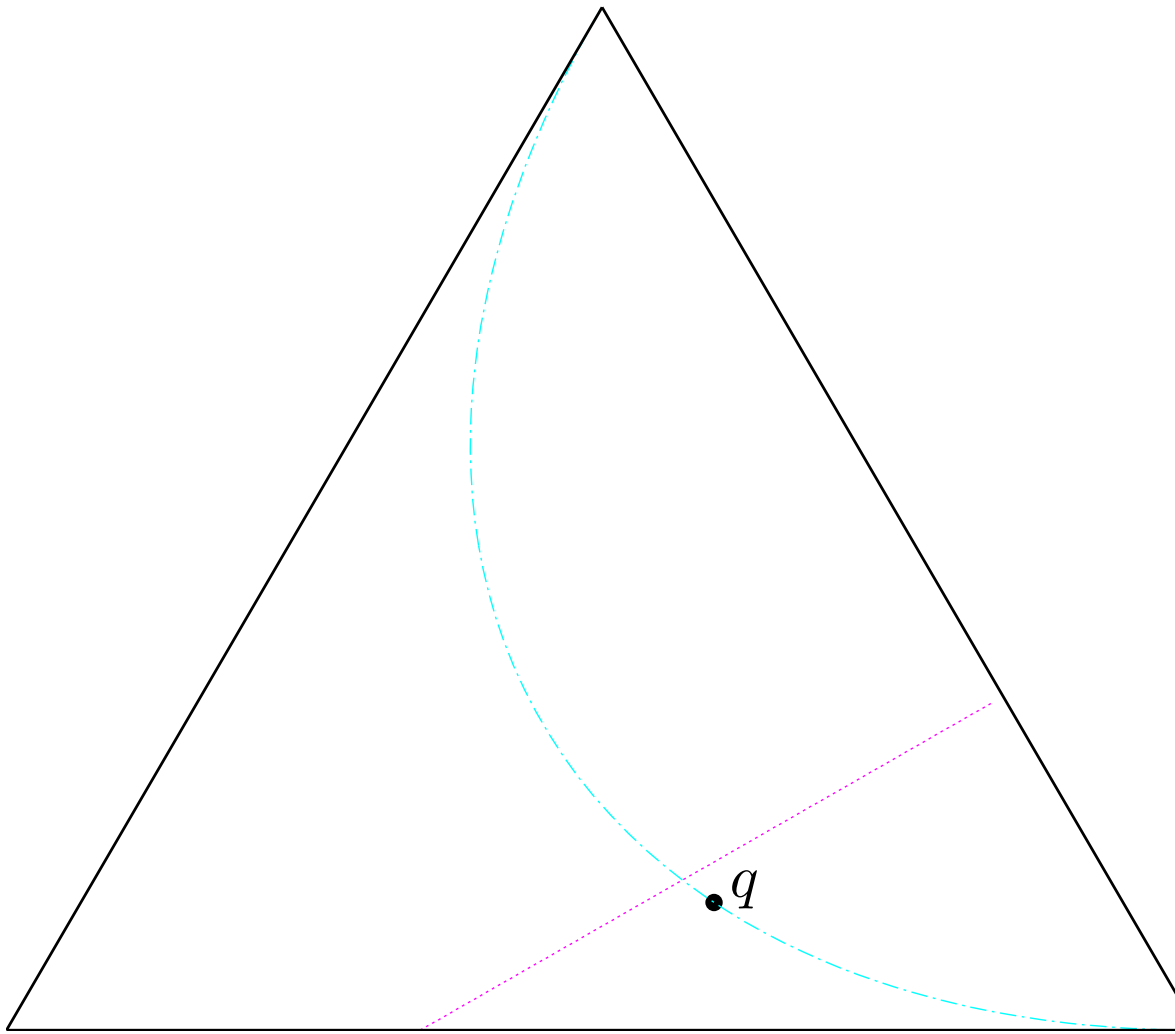
$$\mathcal{P}(X) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

- Expected utility:

$$\mathbb{E}_{q(t)}\{u\} = \sum u\, q(t)$$



Optimal learning trajectory

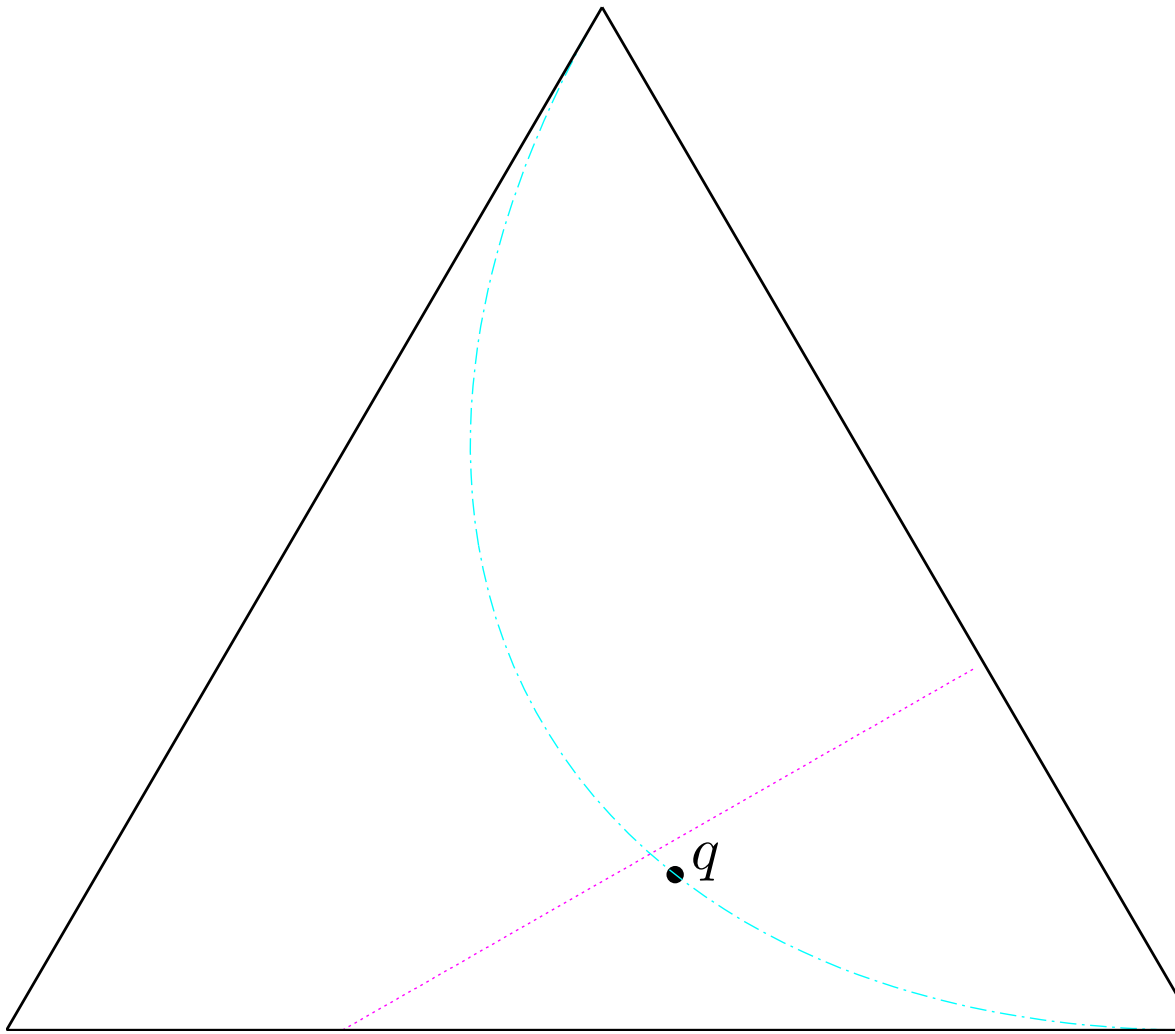
$$\mathcal{P}(X) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

- Expected utility:

$$\mathbb{E}_{q(t)}\{u\} = \sum u\, q(t)$$



Optimal learning trajectory

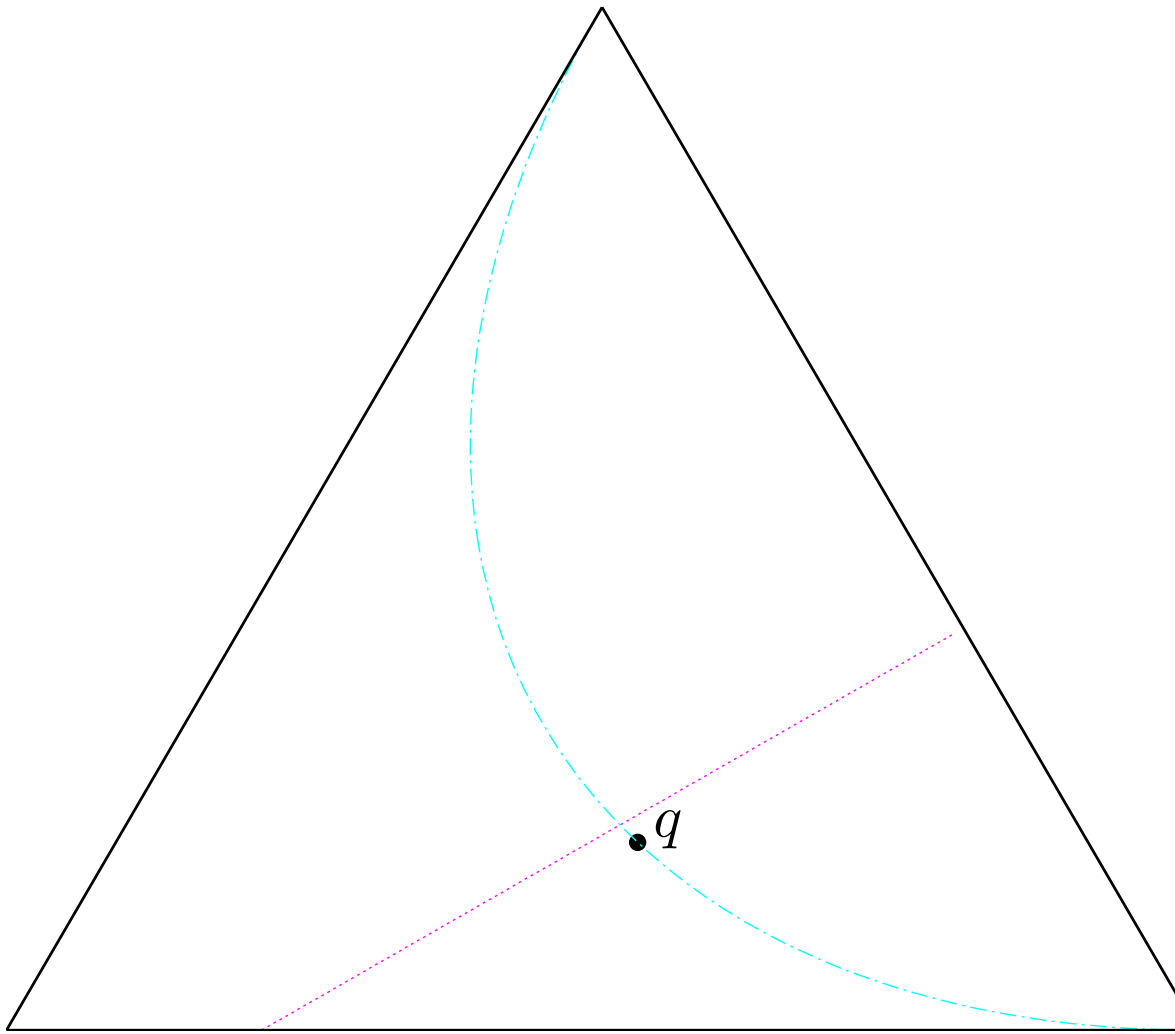
$$\mathcal{P}(X) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

- Expected utility:

$$\mathbb{E}_{q(t)}\{u\} = \sum u\, q(t)$$



Optimal learning trajectory

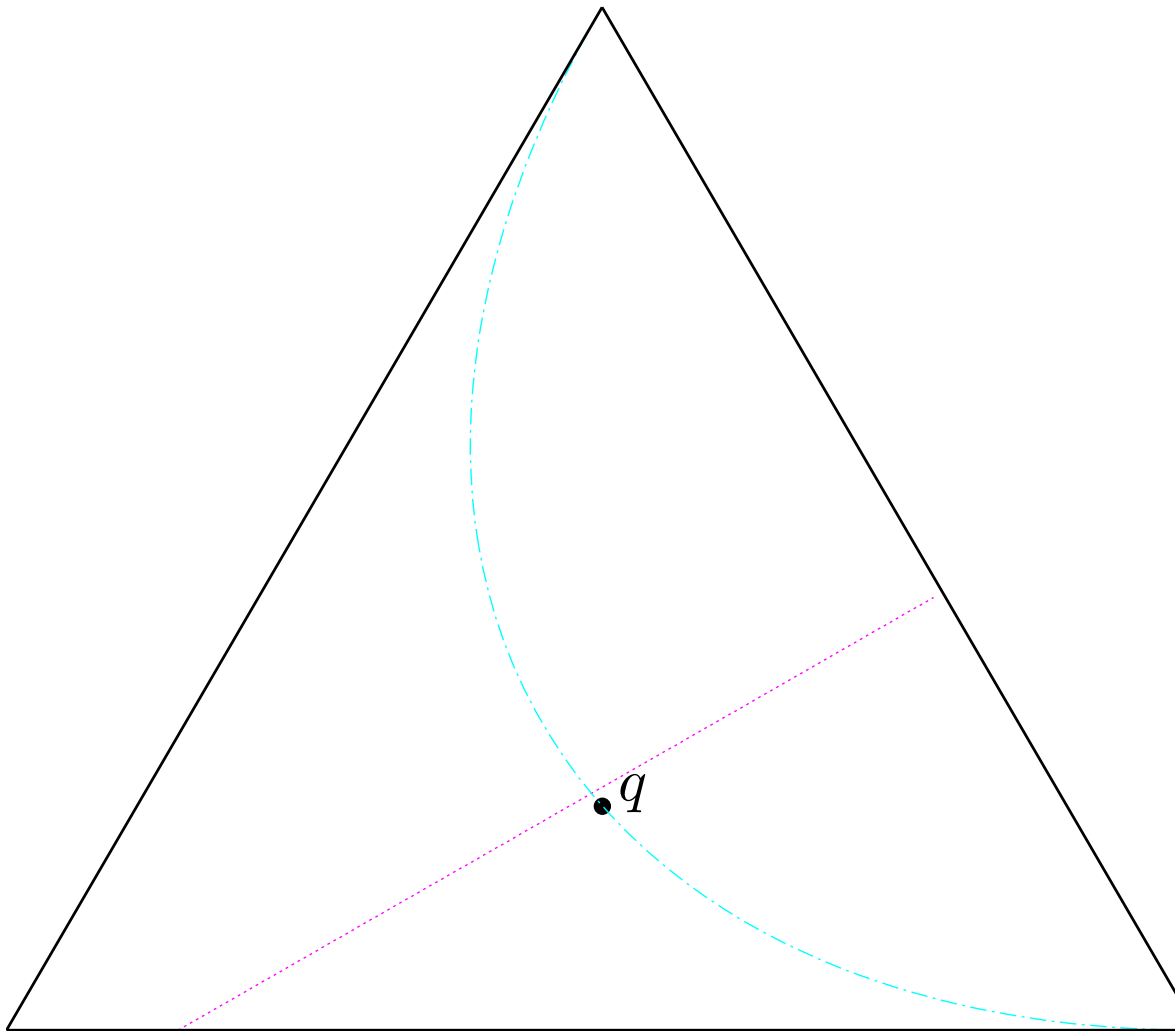
$$\mathcal{P}(X) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

- Expected utility:

$$\mathbb{E}_{q(t)}\{u\} = \sum u\, q(t)$$



Optimal learning trajectory

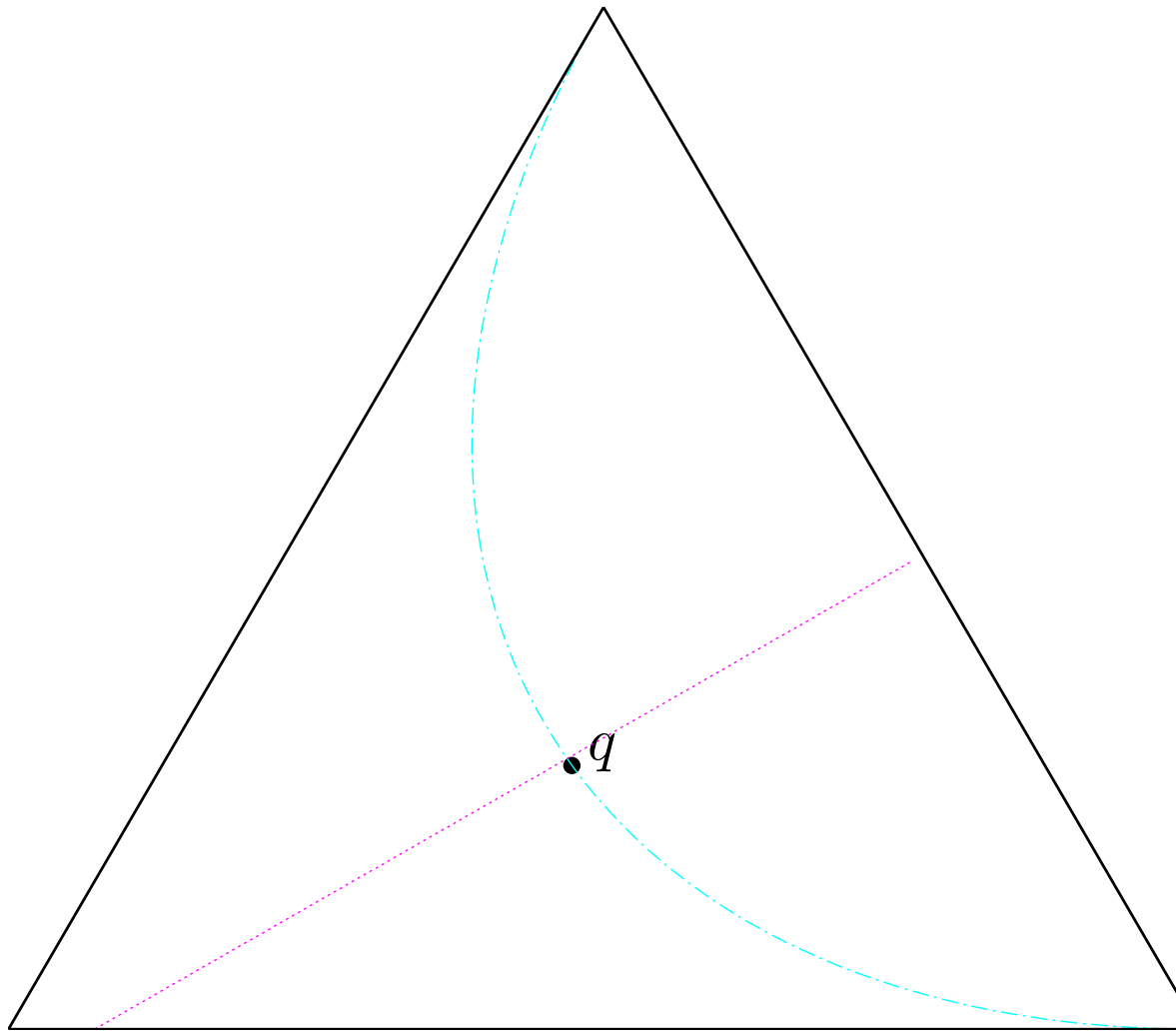
$$\mathcal{P}(X) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

- Expected utility:

$$\mathbb{E}_{q(t)}\{u\} = \sum u\, q(t)$$



Optimal learning trajectory

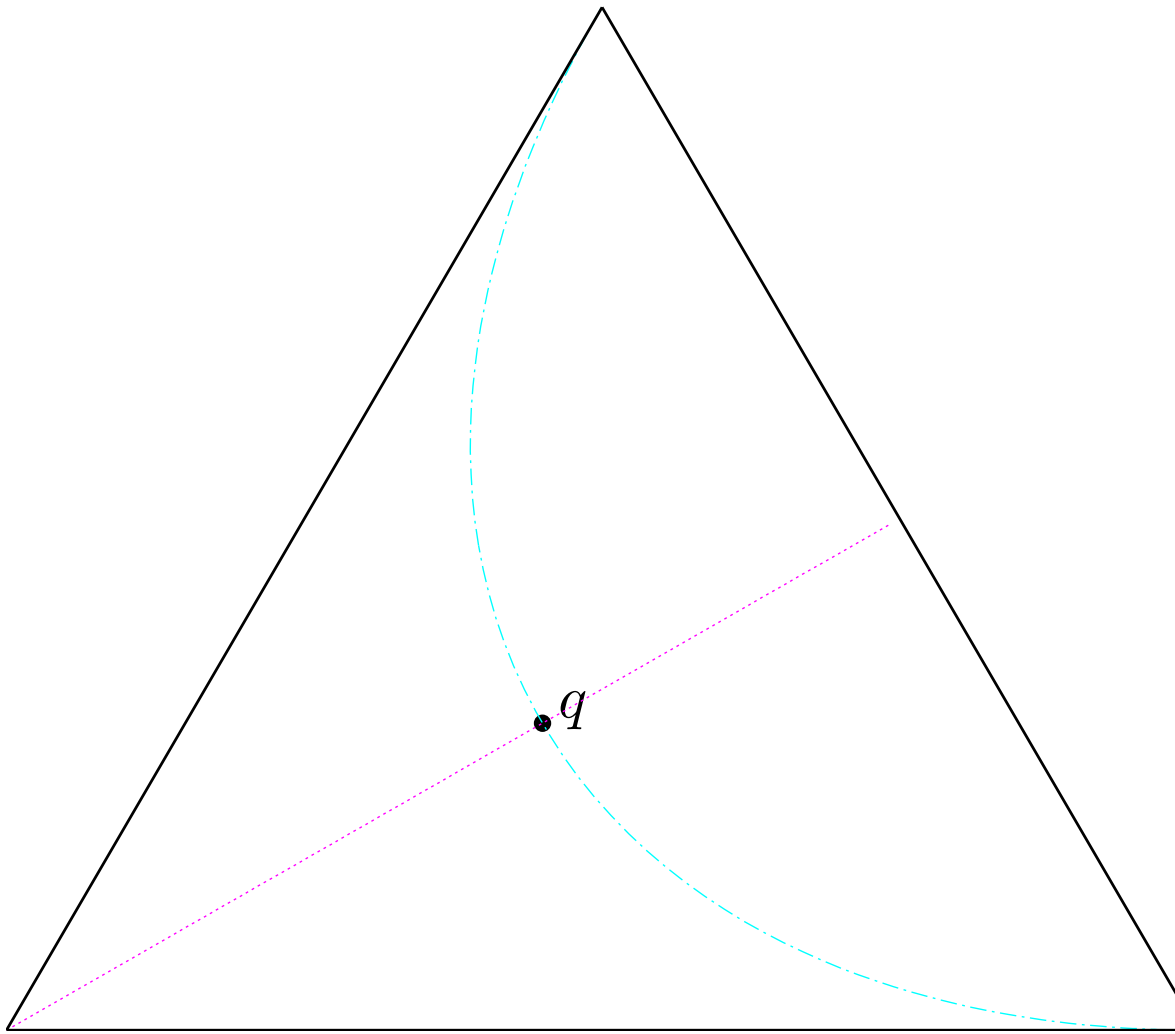
$$\mathcal{P}(X) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

- Expected utility:

$$\mathbb{E}_{q(t)}\{u\} = \sum u\, q(t)$$



Optimal learning trajectory

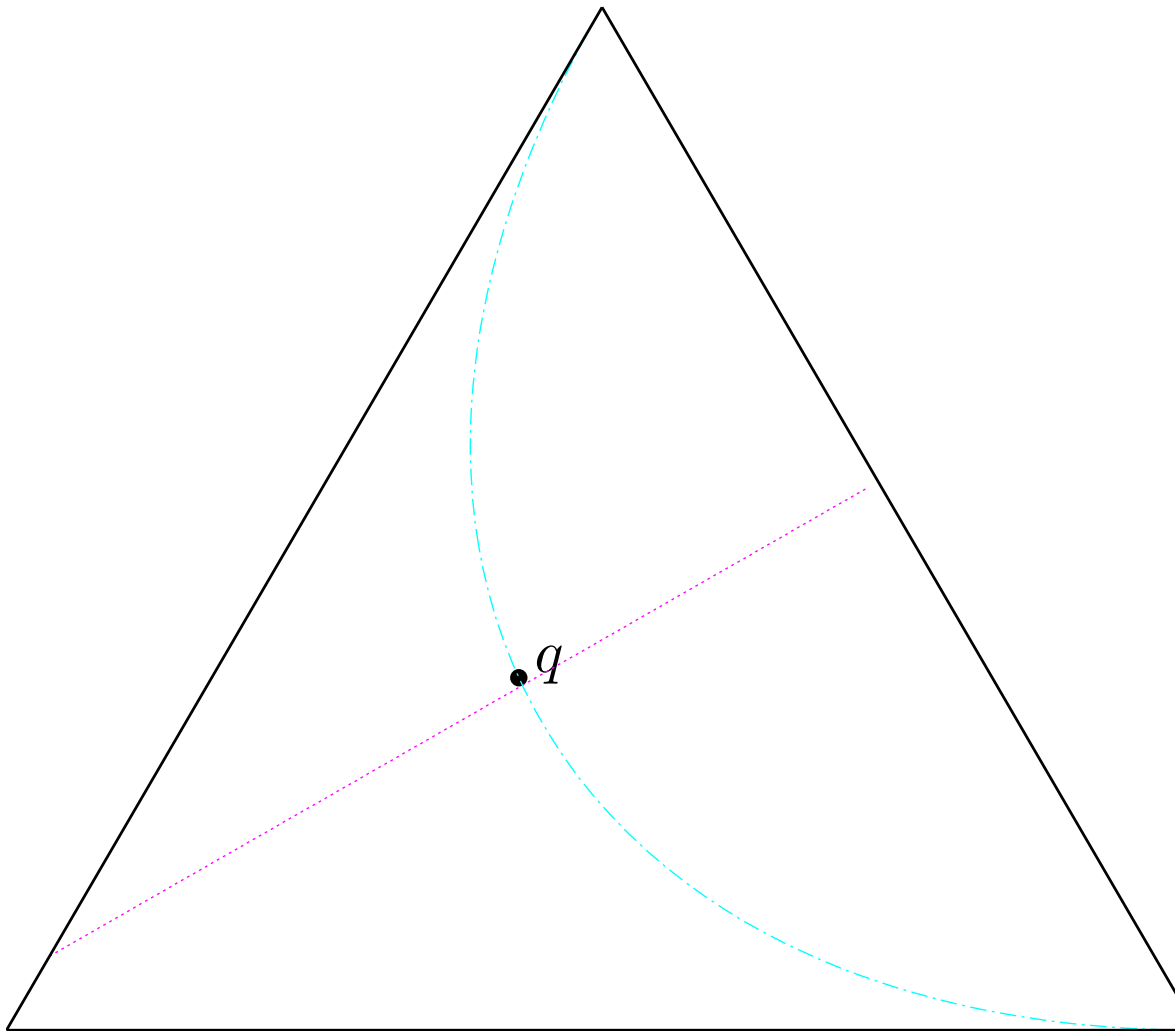
$$\mathcal{P}(X) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

- Expected utility:

$$\mathbb{E}_{q(t)}\{u\} = \sum u\, q(t)$$



Optimal learning trajectory

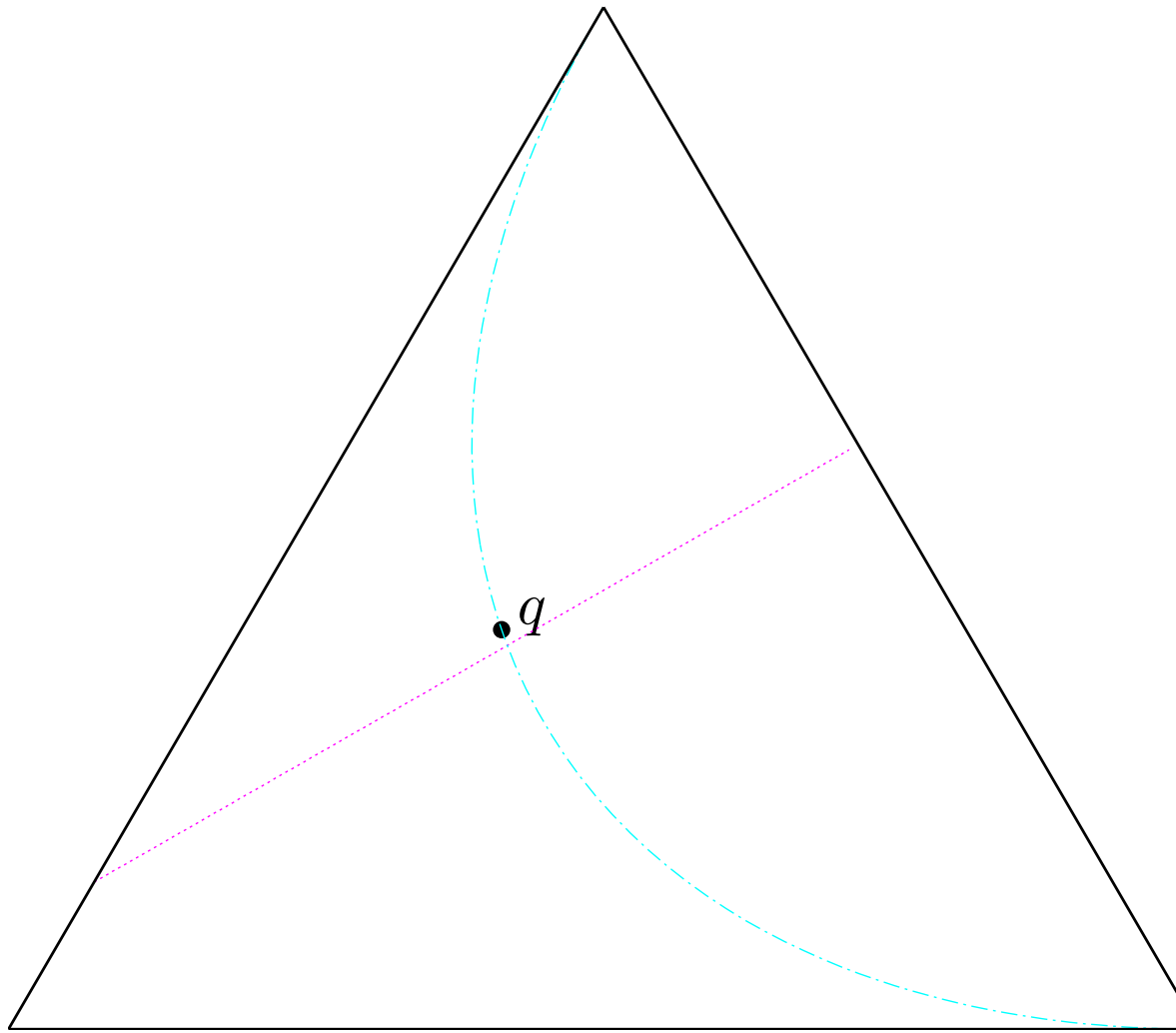
$$\mathcal{P}(X) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

- Expected utility:

$$\mathbb{E}_{q(t)}\{u\} = \sum u\, q(t)$$



Optimal learning trajectory

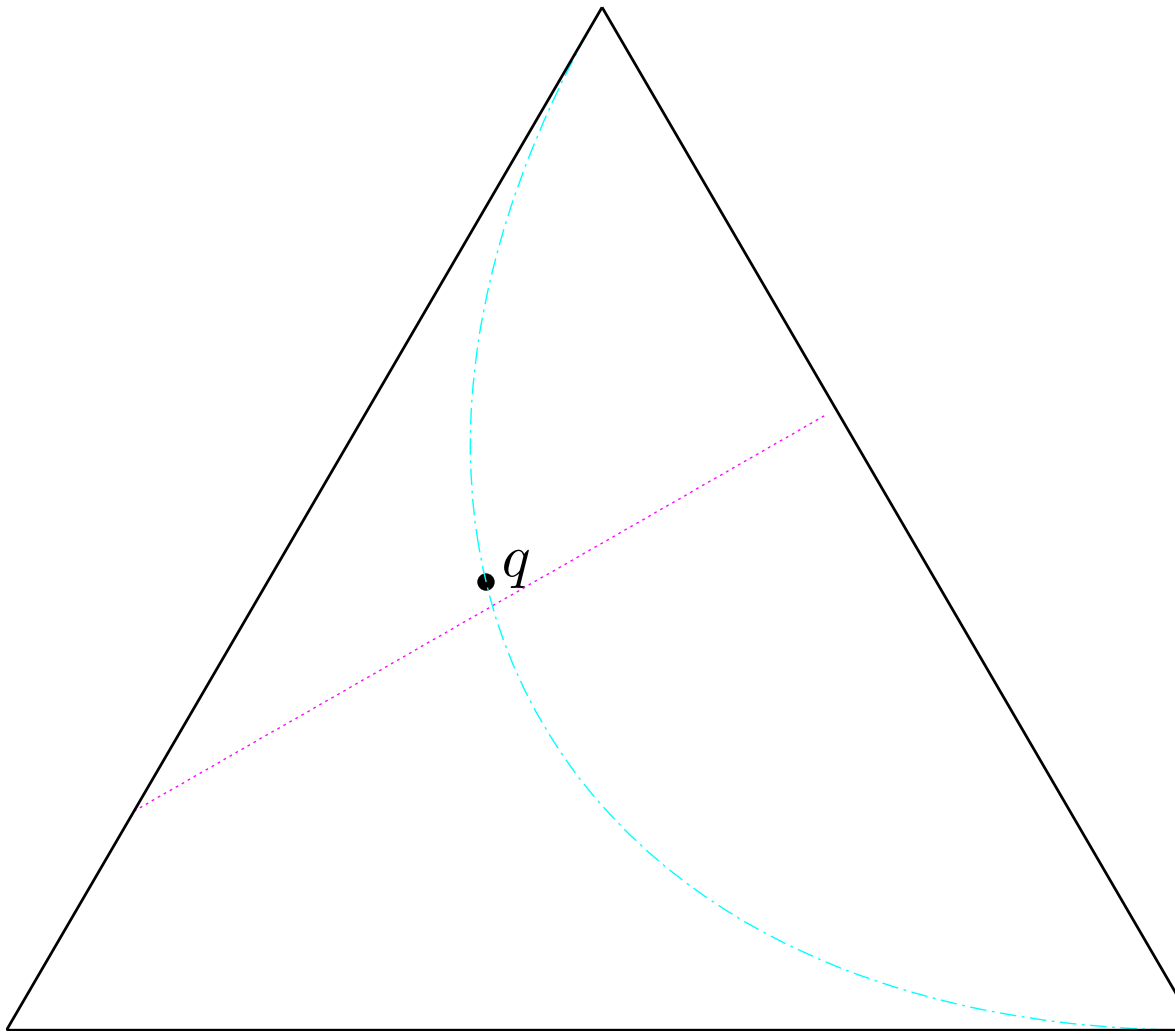
$$\mathcal{P}(X) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

- Expected utility:

$$\mathbb{E}_{q(t)}\{u\} = \sum u\, q(t)$$



Optimal learning trajectory

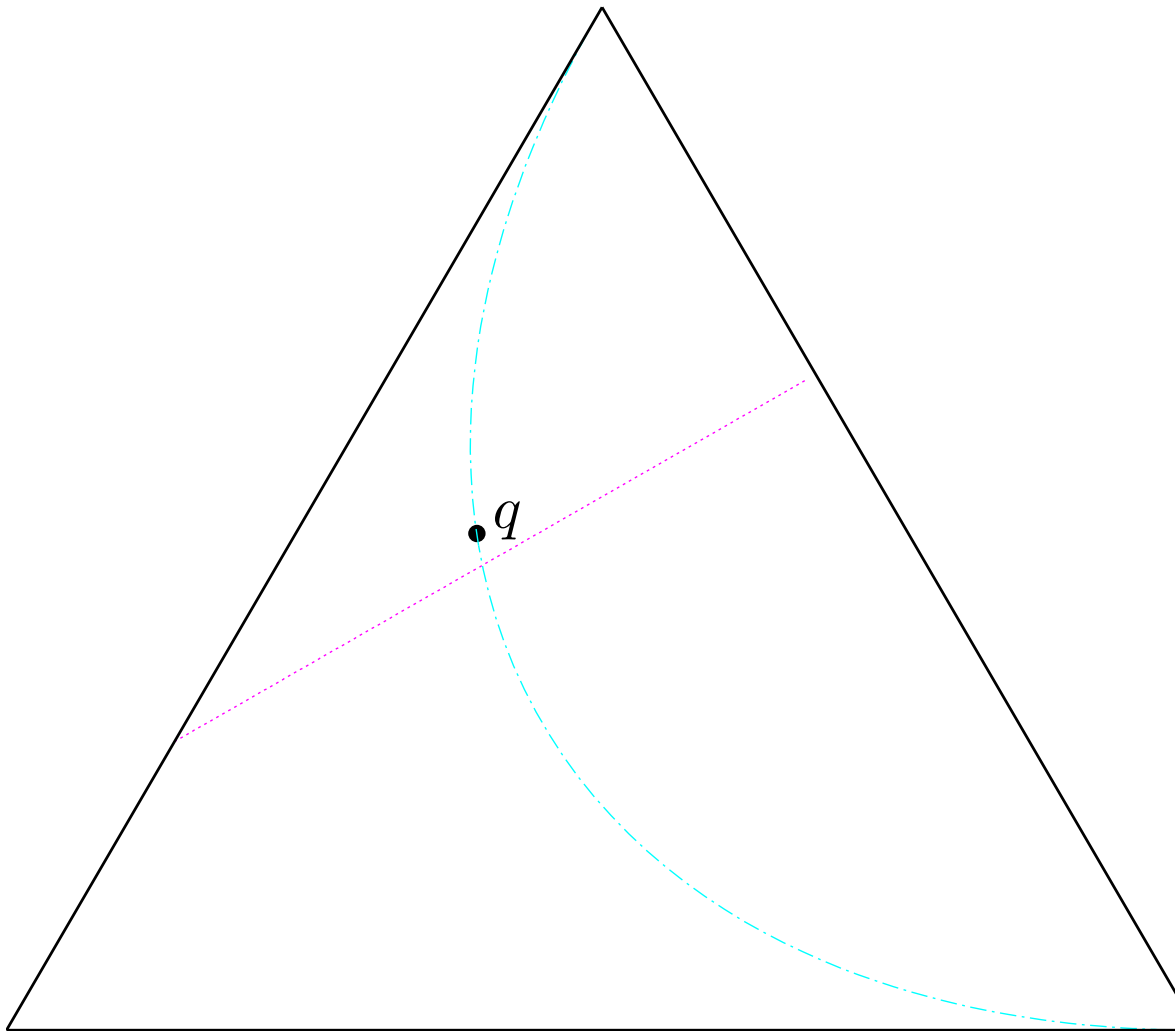
$$\mathcal{P}(X) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

- Expected utility:

$$\mathbb{E}_{q(t)}\{u\} = \sum u\, q(t)$$



Optimal learning trajectory

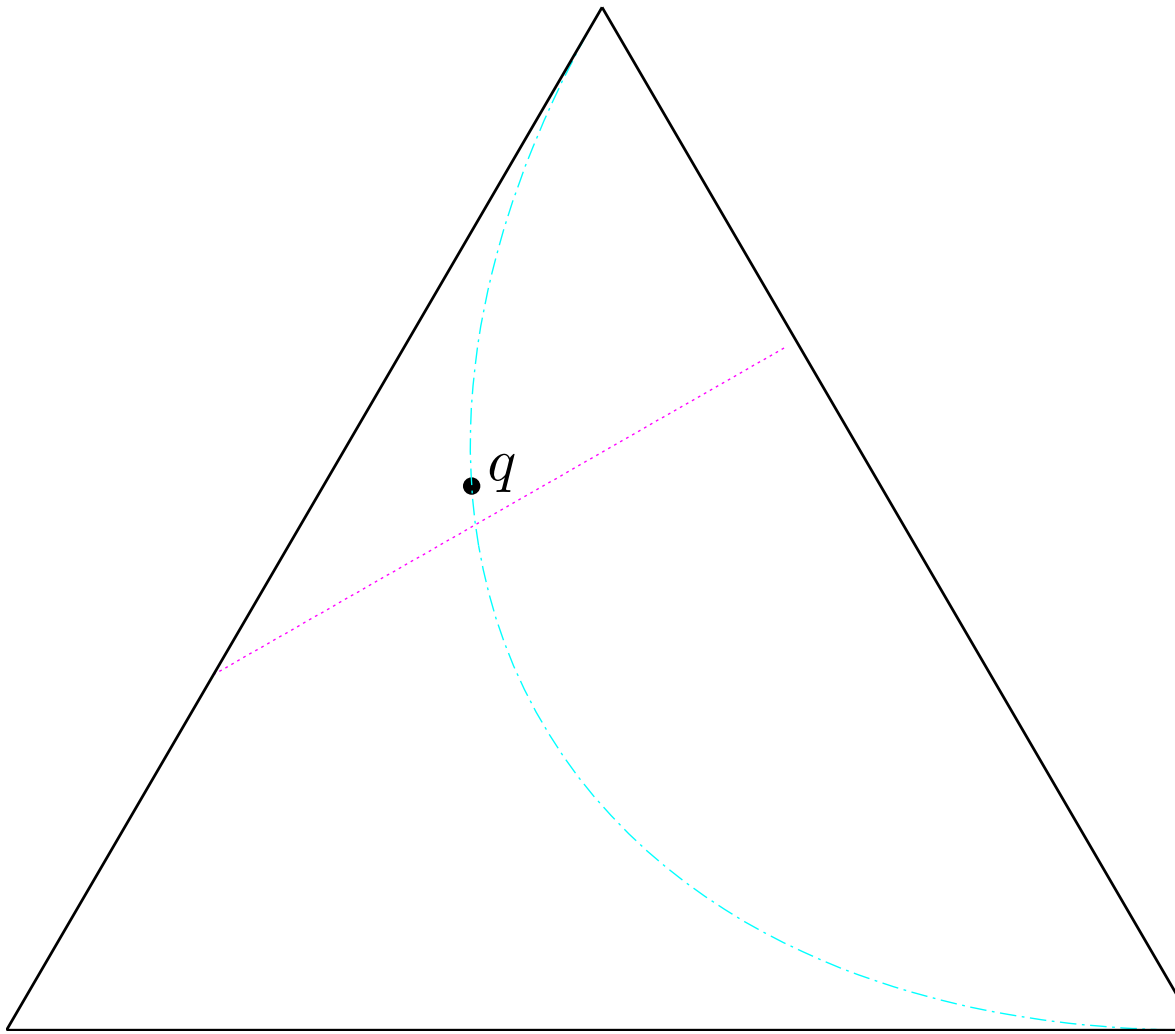
$$\mathcal{P}(X) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Evolution
  $\Lambda : \mathcal{P}(X) \to \mathcal{P}(X)$:

$$q(t+1) = \Lambda\, q(t)$$

- Expected utility:

$$\mathbb{E}_{q(t)}\{u\} = \sum u\, q(t)$$



Optimal learning trajectory

Third variational problem

- Linear programming problem:

$$\text{maximize} \quad \mathbb{E}_w\{u(x,z)\} \quad \text{subject to} \quad I\{x,z\} \leq \lambda$$

Third variational problem

- Linear programming problem:

$$\text{maximize} \quad \mathbb{E}_w\{u(x,z)\} \quad \text{subject to} \quad I\{x,z\} \leq \lambda$$

- The inverse convex programming problem:

$$\text{minimize} \quad I\{x,z\} \quad \text{subject to} \quad \mathbb{E}_w\{u(x,z)\} \geq \upsilon$$

Third variational problem

- Linear programming problem:

$$\text{maximize} \quad \mathbb{E}_w\{u(x,z)\} \quad \text{subject to} \quad I\{x,z\} \leq \lambda$$

- The inverse convex programming problem:

$$\text{minimize} \quad I\{x,z\} \quad \text{subject to} \quad \mathbb{E}_w\{u(x,z)\} \geq \upsilon$$

- Information:

$$I\{x,z\} = H\{z\} - H\{z \mid x\} \leq H\{z\} \leq \ln|Z|$$

Optimal transition kernels

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

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

Optimal transition kernels

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

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

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

$$I\{x, z\} \leq \lambda$$

Optimal transition kernels

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

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

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

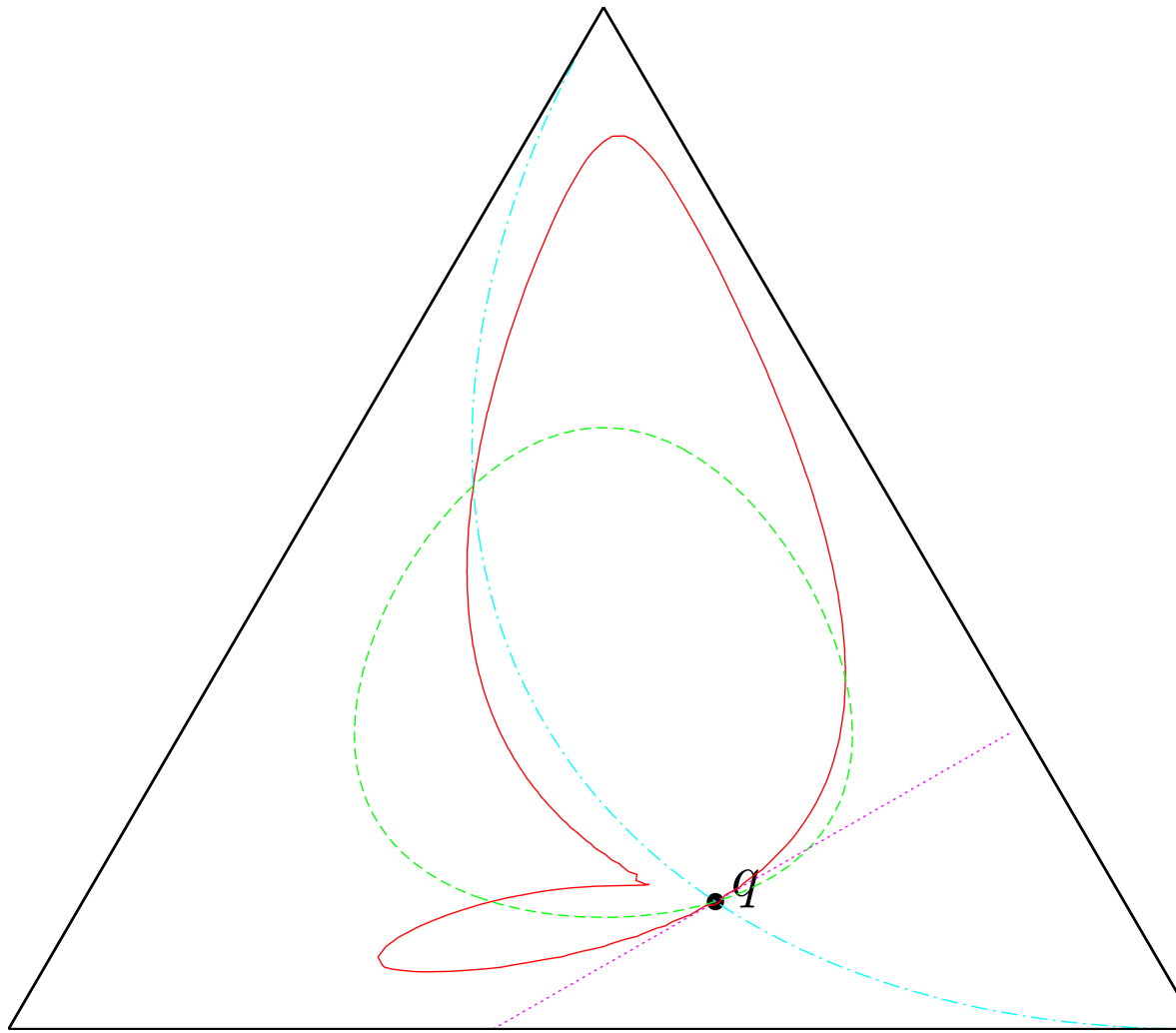
$$I\{x, z\} \leq \lambda$$

- One can show that temperature β^{-1} is the slope of $V(\lambda)$:

$$\beta^{-1} = \frac{dV(\lambda)}{d\lambda}$$

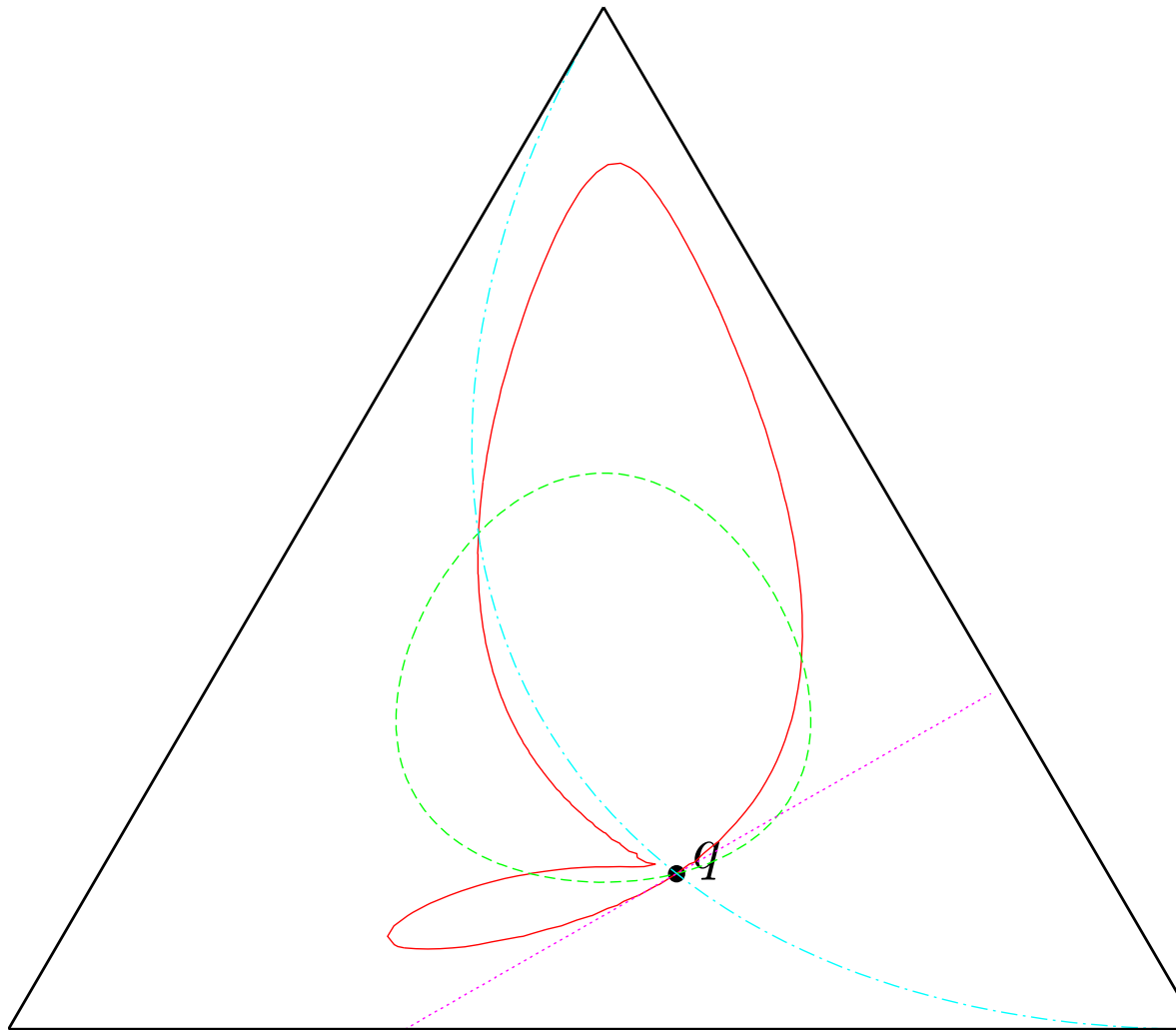
Geometry of Optimal Evolution

$$
\begin{aligned}
H(q,p) &\geq H(q) \\
H(p) &\geq H(q) \\
I(p,q) &\geq I(q,p) \\
I(w, q \otimes q) &\geq H(q) \\
I(p,q) &\geq H\{a \mid b\}
\end{aligned}
$$



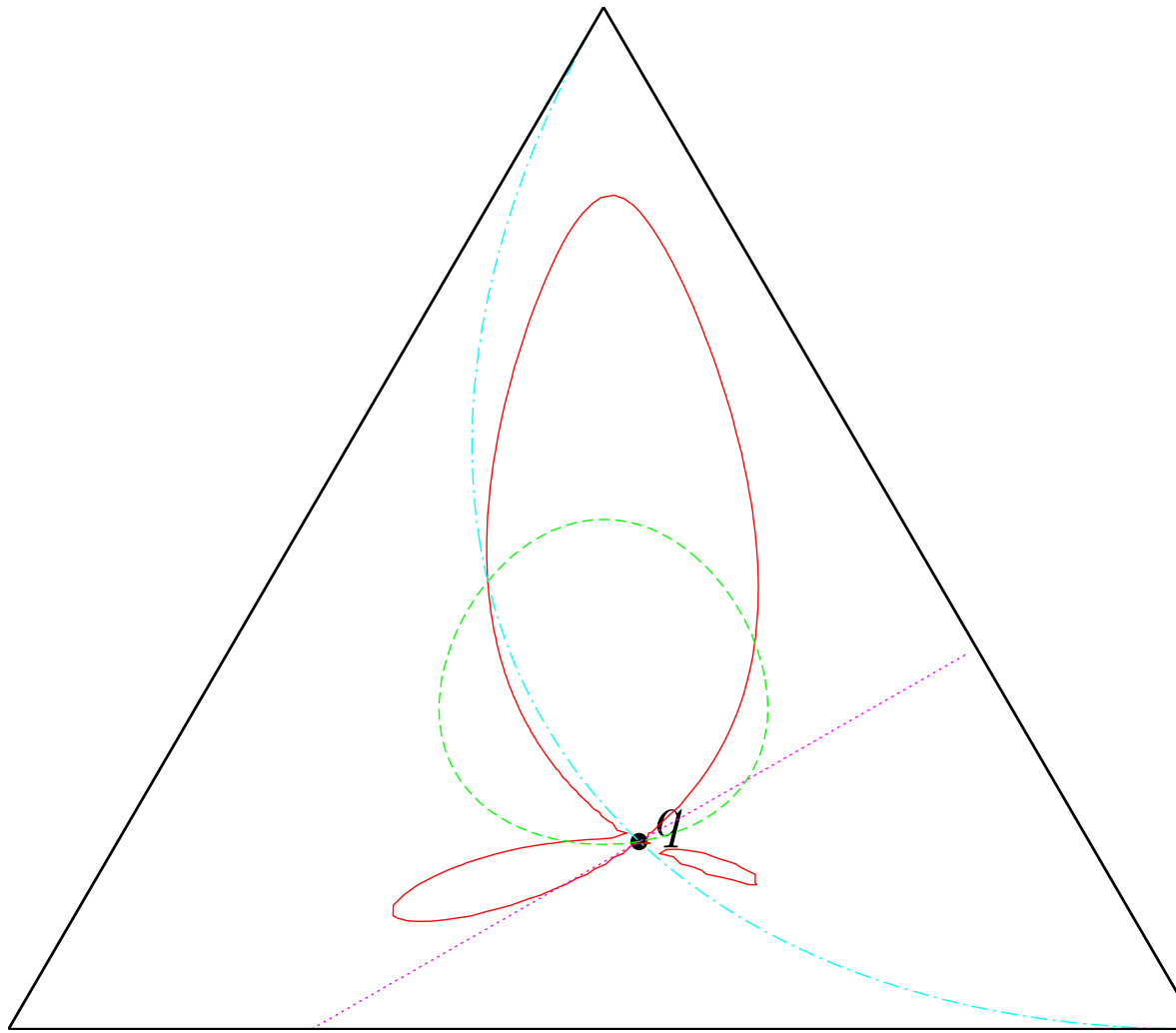
Geometry of Optimal Evolution

$$
\begin{aligned}
H(q,p) &\geq H(q)\\
H(p) &\geq H(q)\\
I(p,q) &\geq I(q,p)\\
I(w, q\otimes q) &\geq H(q)\\
I(p,q) &\geq H\{a \mid b\}
\end{aligned}
$$



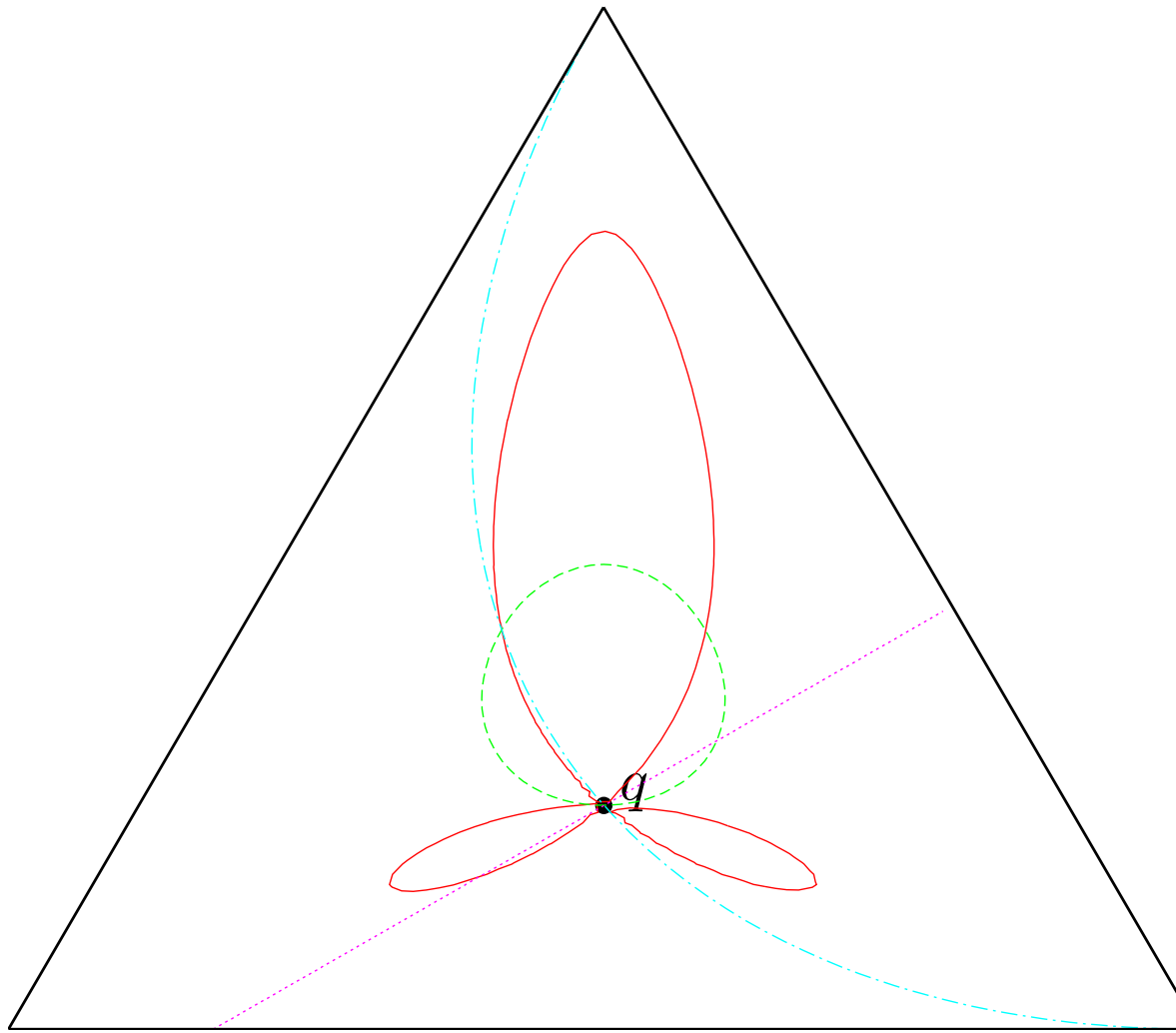
Geometry of Optimal Evolution

$$
\begin{aligned}
H(q,p) &\geq H(q) \\
H(p) &\geq H(q) \\
I(p,q) &\geq I(q,p) \\
I(w, q \otimes q) &\geq H(q) \\
I(p,q) &\geq H\{a \mid b\}
\end{aligned}
$$



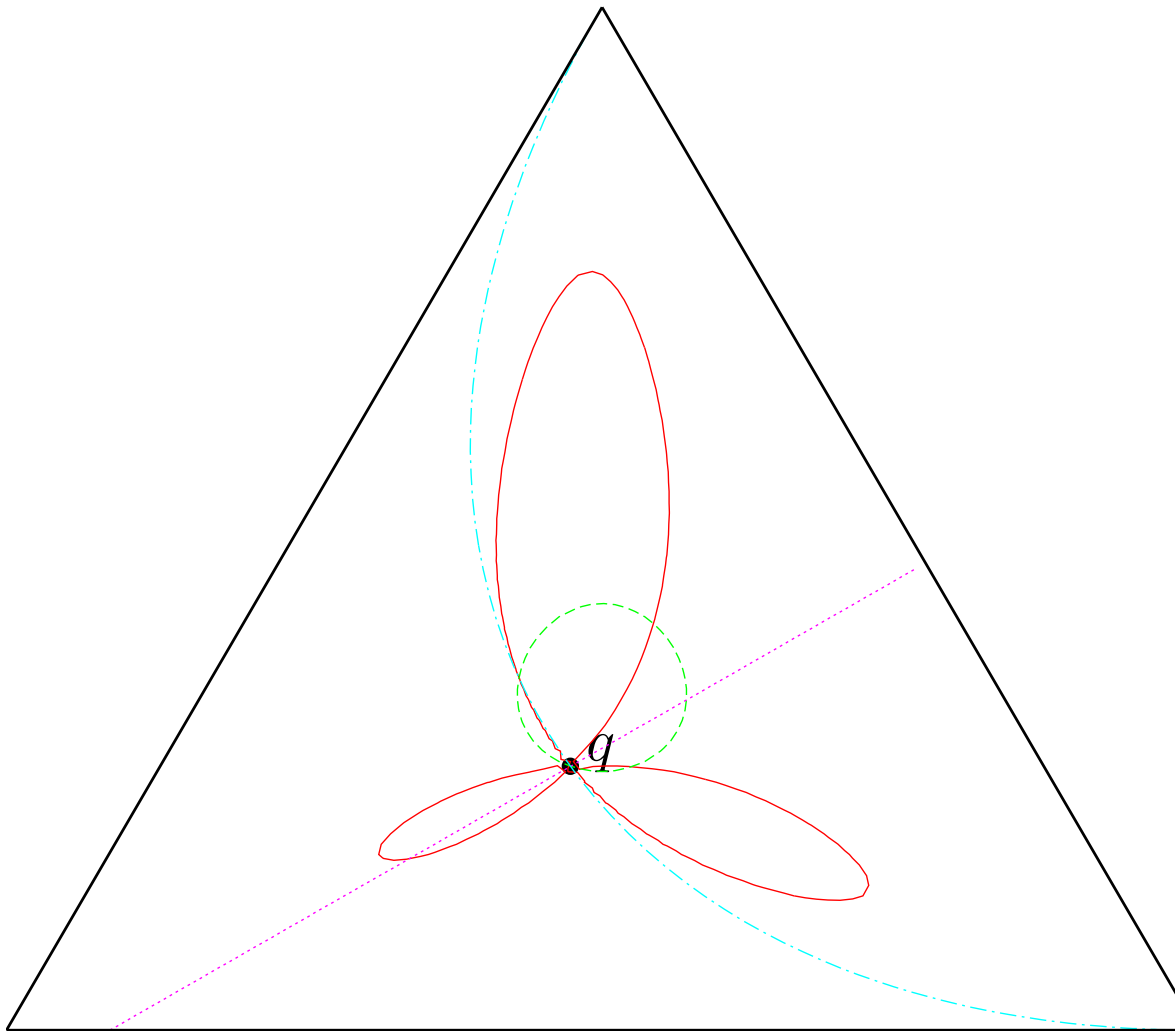
Geometry of Optimal Evolution

$$
\begin{aligned}
H(q,p) &\geq H(q) \\
H(p) &\geq H(q) \\
I(p,q) &\geq I(q,p) \\
I(w, q \otimes q) &\geq H(q) \\
I(p,q) &\geq H\{a \mid b\}
\end{aligned}
$$



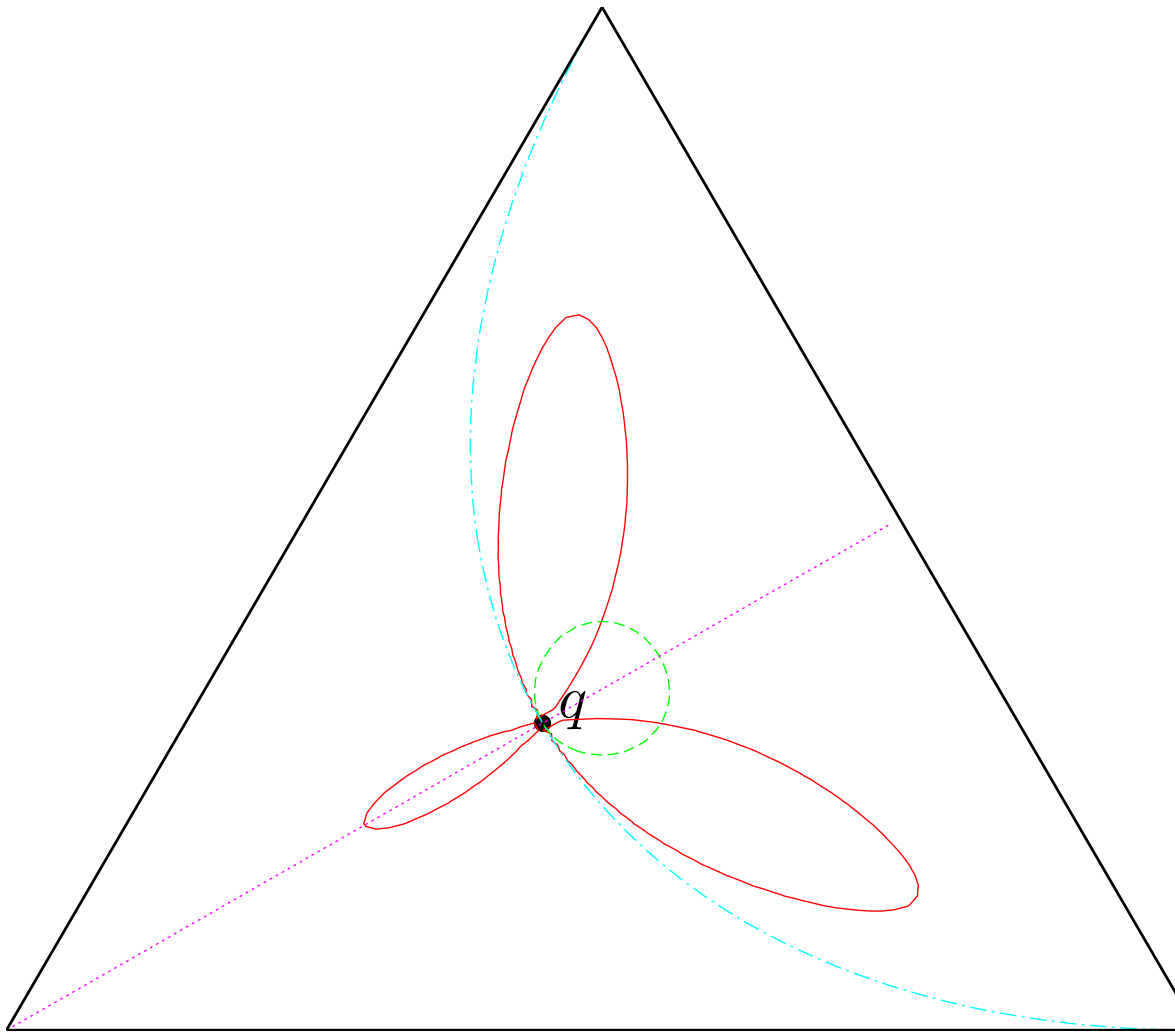
Geometry of Optimal Evolution

$$
\begin{aligned}
H(q,p) &\geq H(q) \\
H(p) &\geq H(q) \\
I(p,q) &\geq I(q,p) \\
I(w, q \otimes q) &\geq H(q) \\
I(p,q) &\geq H\{a \mid b\}
\end{aligned}
$$



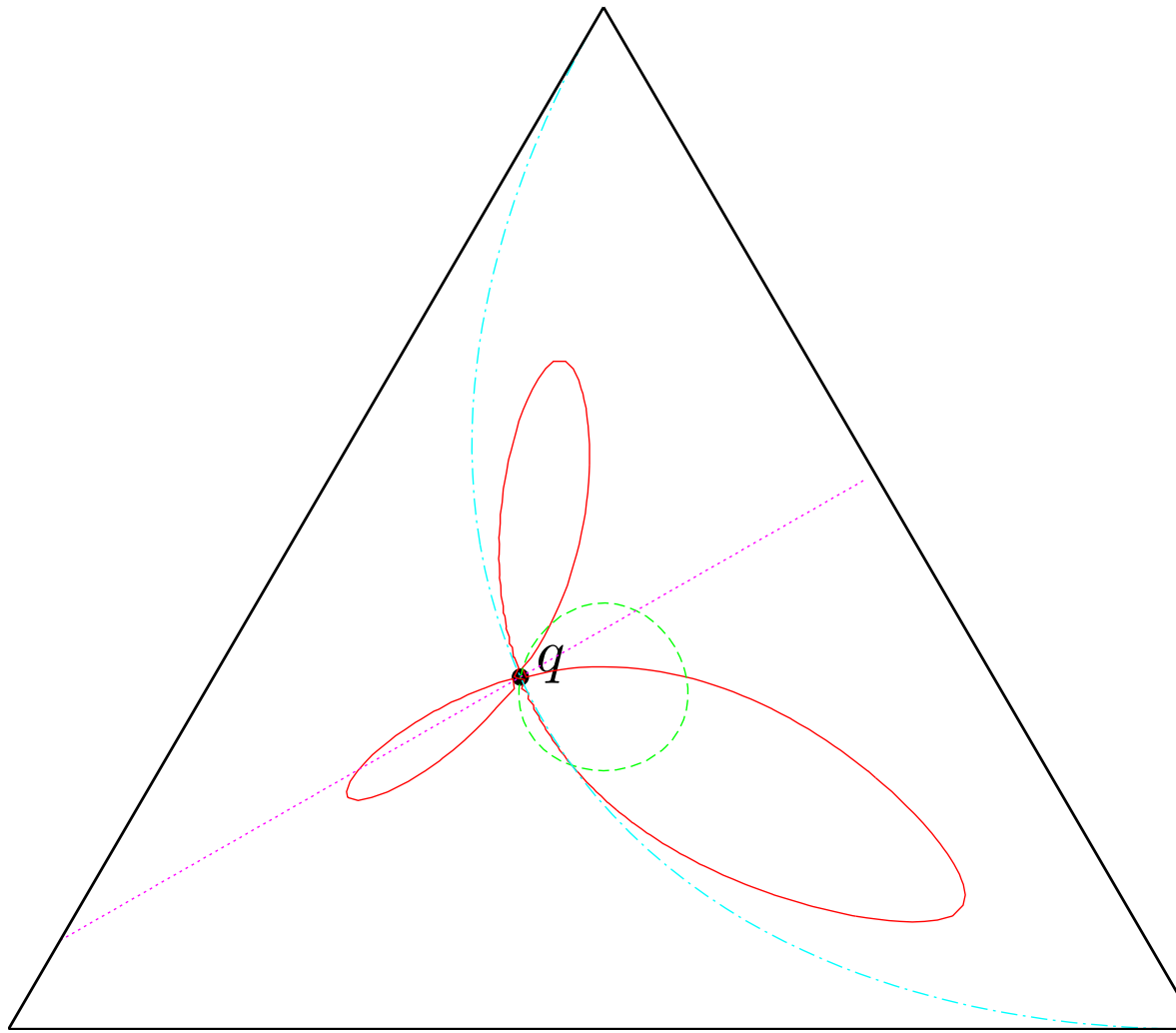
Geometry of Optimal Evolution

$$
\begin{aligned}
H(q,p) &\leq H(q) \\
H(p) &\leq H(q) \\
I(p,q) &\leq I(q,p) \\
I(w, q \otimes q) &\leq H(q) \\
I(p,q) &\leq H\{a \mid b\}
\end{aligned}
$$



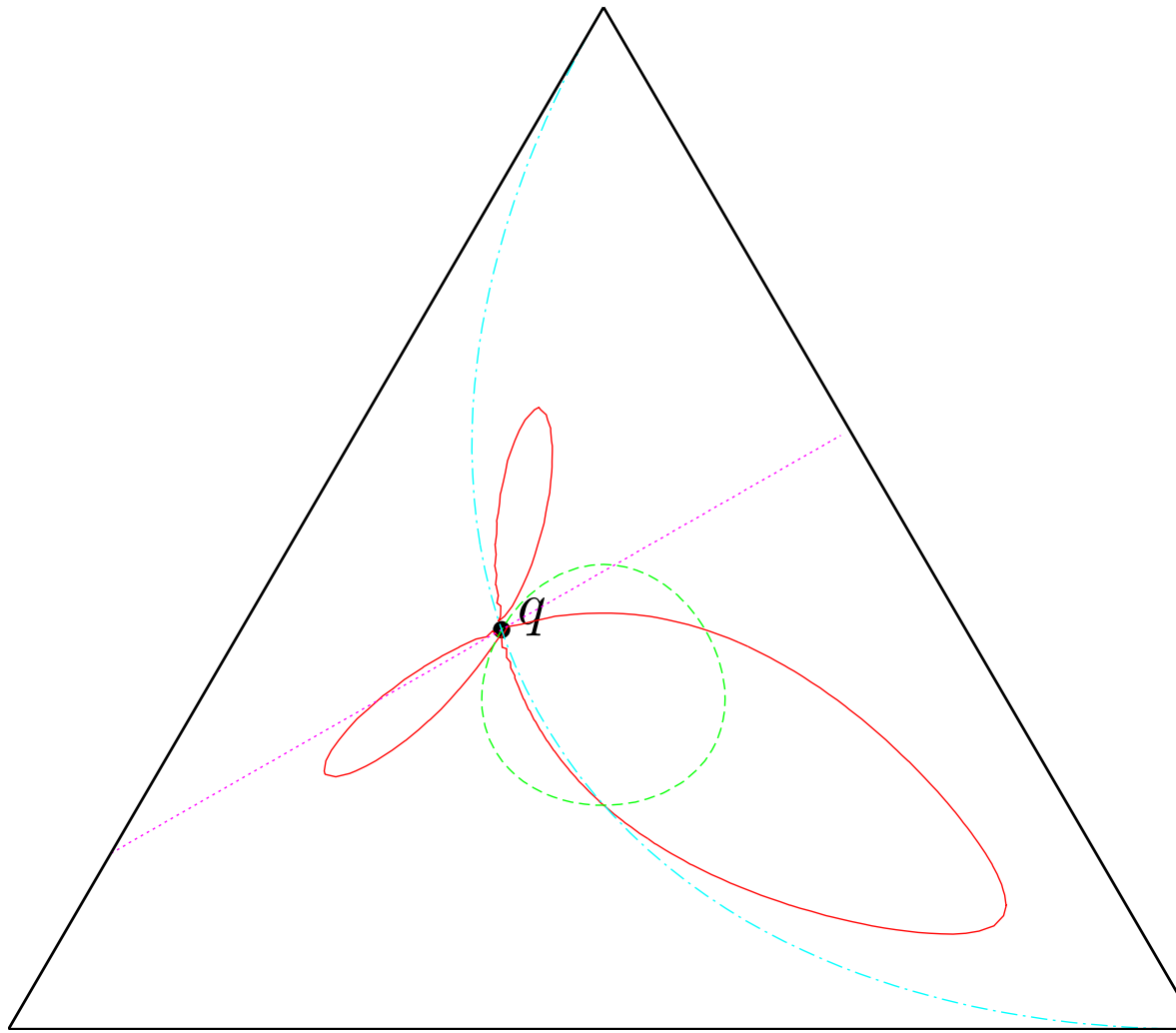
Geometry of Optimal Evolution

$$
\begin{aligned}
H(q,p) &\leq H(q) \\
H(p) &\leq H(q) \\
I(p,q) &\leq I(q,p) \\
I(w, q \otimes q) &\leq H(q) \\
I(p,q) &\leq H\{a \mid b\}
\end{aligned}
$$



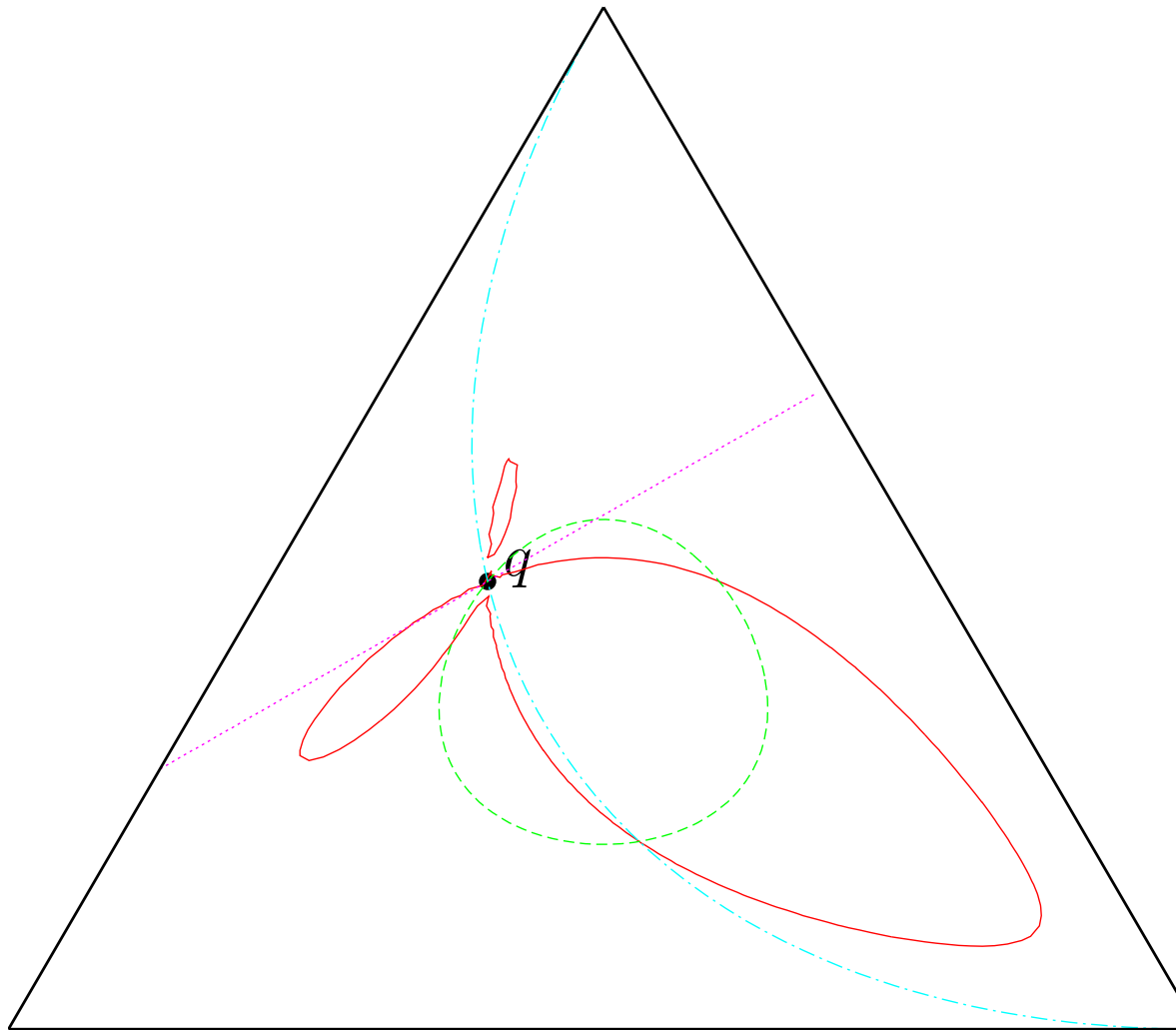
Geometry of Optimal Evolution

$$
\begin{aligned}
H(q,p) &\leq H(q) \\
H(p) &\leq H(q) \\
I(p,q) &\leq I(q,p) \\
I(w, q \otimes q) &\leq H(q) \\
I(p,q) &\leq H\{a \mid b\}
\end{aligned}
$$



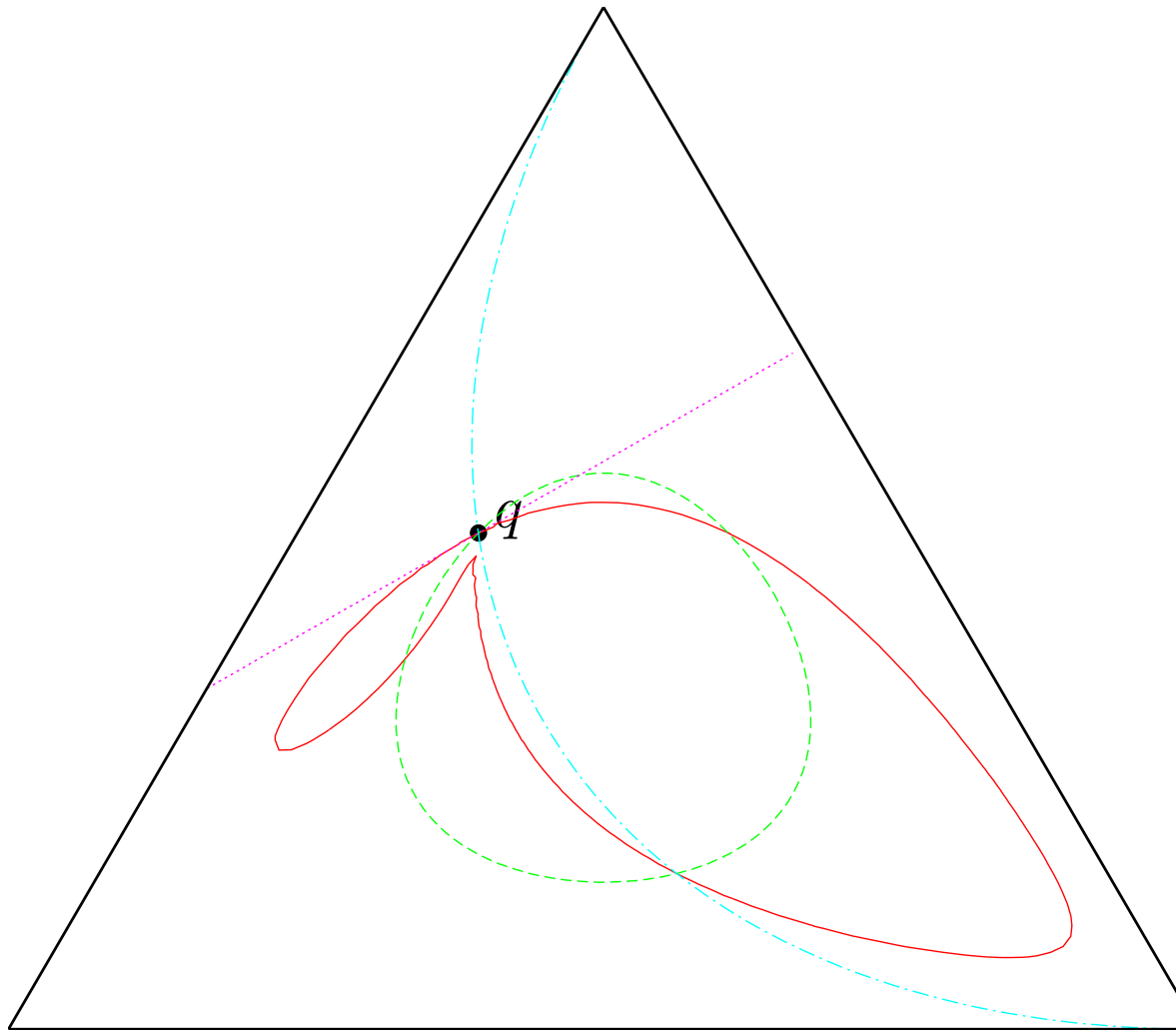
Geometry of Optimal Evolution

$$
\begin{aligned}
H(q,p) &\leq H(q) \\
H(p) &\leq H(q) \\
I(p,q) &\leq I(q,p) \\
I(w, q \otimes q) &\leq H(q) \\
I(p,q) &\leq H\{a \mid b\}
\end{aligned}
$$



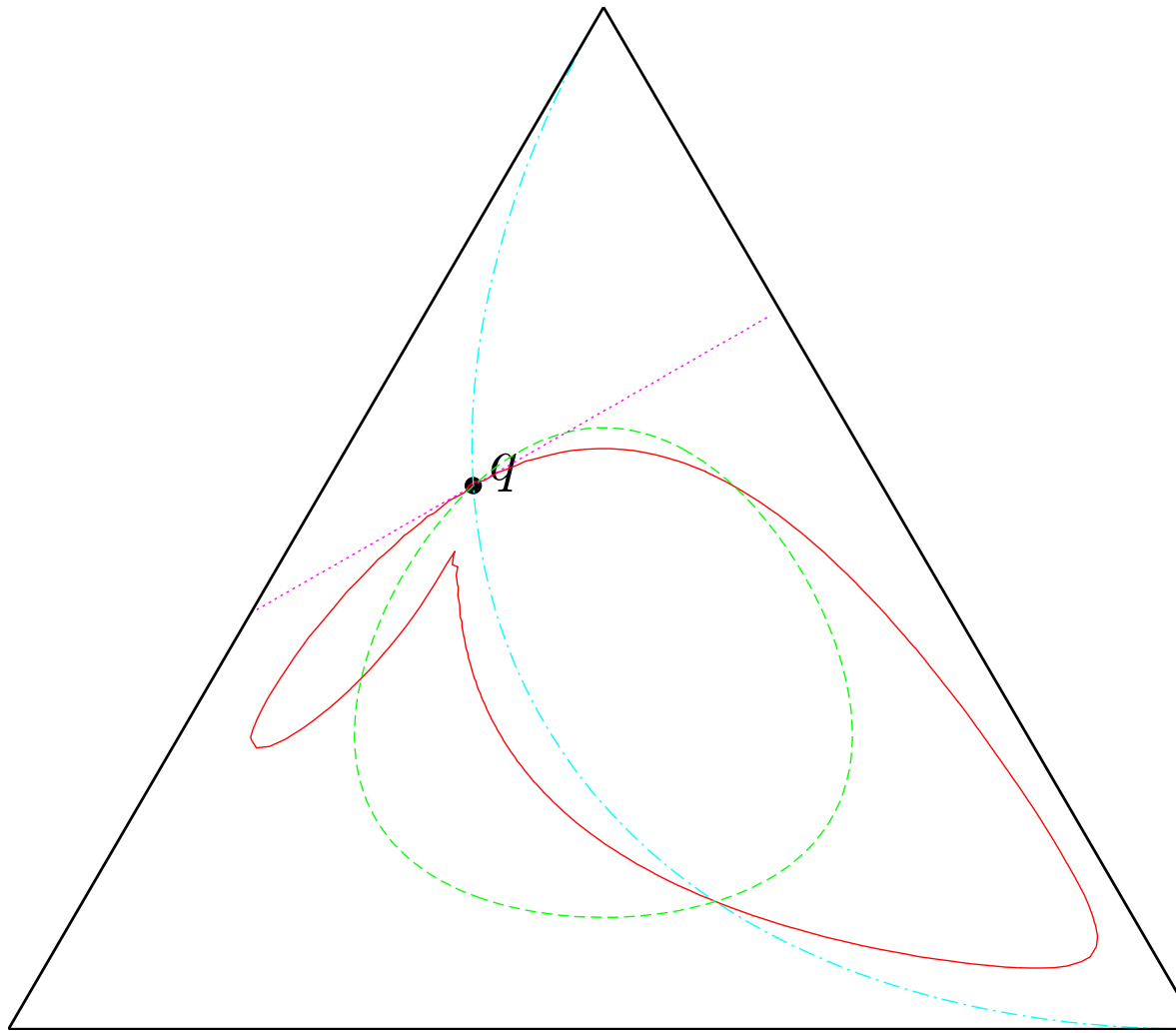
Geometry of Optimal Evolution

$$
\begin{aligned}
H(q,p) &\leq H(q) \\
H(p) &\leq H(q) \\
I(p,q) &\leq I(q,p) \\
I(w, q \otimes q) &\leq H(q) \\
I(p,q) &\leq H\{a \mid b\}
\end{aligned}
$$



Geometry of Optimal Evolution

$$
\begin{aligned}
H(q,p) &\leq H(q) \\
H(p) &\leq H(q) \\
I(p,q) &\leq I(q,p) \\
I(w, q \otimes q) &\leq H(q) \\
I(p,q) &\leq H\{a \mid b\}
\end{aligned}
$$



Geometry of optimal transition kernels

Theorem (Belavkin, 2013)

- *Let* $\{p(\beta)\}$ *be a family of* $p(\beta) \in \mathcal{P}(X \times Y)$

  maximizing $\mathbb{E}_p\{u\}$ *on sets* $\{p : I(p,q) \leq \lambda\}$, $\forall\, \lambda \in \mathbb{R} \cup \{\infty\}$

Geometry of optimal transition kernels

Theorem (Belavkin, 2013)

- *Let* $\{p(\beta)\}$ *be a family of* $p(\beta) \in \mathcal{P}(X \times Y)$

  maximizing $\mathbb{E}_p\{u\}$ *on sets* $\{p : I(p,q) \leq \lambda\}$, $\forall\, \lambda \in \mathbb{R} \cup \{\infty\}$

- *Let* $I(p,q)$ *be minimized at* $q \in \partial I^*(0,q) \subset \mathrm{Int}(\mathcal{P}(X \times Y))$.

Geometry of optimal transition kernels

Theorem (Belavkin, 2013)

- *Let* $\{p(\beta)\}$ *be a family of* $p(\beta) \in \mathcal{P}(X \times Y)$

  maximizing $\mathbb{E}_p\{u\}$ *on sets* $\{p : I(p,q) \leq \lambda\}$, $\forall\, \lambda \in \mathbb{R} \cup \{\infty\}$

- *Let* $I(p,q)$ *be minimized at* $q \in \partial I^*(0,q) \subset \mathrm{Int}(\mathcal{P}(X \times Y))$.
- *If* $I^*(u,q)$ *is strictly convex in* u, *then*

Geometry of optimal transition kernels

Theorem (Belavkin, 2013)

- *Let* $\{p(\beta)\}$ *be a family of* $p(\beta) \in \mathcal{P}(X \times Y)$

  maximizing $\mathbb{E}_p\{u\}$ *on sets* $\{p : I(p,q) \leq \lambda\}$, $\forall\, \lambda \in \mathbb{R} \cup \{\infty\}$

- *Let* $I(p,q)$ *be minimized at* $q \in \partial I^*(0,q) \subset \mathrm{Int}(\mathcal{P}(X \times Y))$.
- *If* $I^*(u,q)$ *is strictly convex in* u, *then*
  1. $p(\beta)$ *is deterministic iff there is no constraint* ($\lambda \geq \sup I$).

Geometry of optimal transition kernels

Theorem (Belavkin, 2013)

- *Let* $\{p(\beta)\}$ *be a family of* $p(\beta) \in \mathcal{P}(X \times Y)$

  maximizing $\mathbb{E}_p\{u\}$ *on sets* $\{p : I(p,q) \leq \lambda\}$, $\forall\, \lambda \in \mathbb{R} \cup \{\infty\}$

- *Let* $I(p,q)$ *be minimized at* $q \in \partial I^*(0,q) \subset \mathrm{Int}(\mathcal{P}(X \times Y))$.
- *If* $I^*(u,q)$ *is strictly convex in* u, *then*
  1. $p(\beta)$ *is deterministic iff there is no constraint* ($\lambda \geq \sup I$).
  2. *If* $p_\phi \in \mathcal{P}(X \times Y)$ *is deterministic and* $I(p_\phi, q) = I(p(\beta), q)$, *then*

$$\mathbb{E}_{p_\phi}\{u\} < \mathbb{E}_{p(\beta)}\{u\}$$

Geometry of optimal transition kernels

Theorem (Belavkin, 2013)

- *Let* $\{p(\beta)\}$ *be a family of* $p(\beta) \in \mathcal{P}(X \times Y)$

  maximizing $\mathbb{E}_p\{u\}$ *on sets* $\{p : I(p,q) \leq \lambda\}$, $\forall\, \lambda \in \mathbb{R} \cup \{\infty\}$

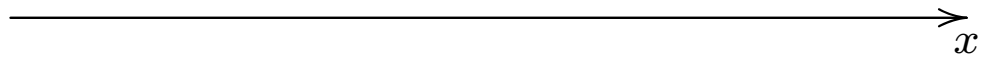
- *Let* $I(p,q)$ *be minimized at* $q \in \partial I^*(0,q) \subset \mathrm{Int}(\mathcal{P}(X \times Y))$.
- *If* $I^*(u,q)$ *is strictly convex in* u, *then*
  1. $p(\beta)$ *is deterministic iff there is no constraint* ($\lambda \geq \sup I$).
  2. *If* $p_\phi \in \mathcal{P}(X \times Y)$ *is deterministic and* $I(p_\phi, q) = I(p(\beta), q)$, *then*

     $$\mathbb{E}_{p_\phi}\{u\} < \mathbb{E}_{p(\beta)}\{u\}$$

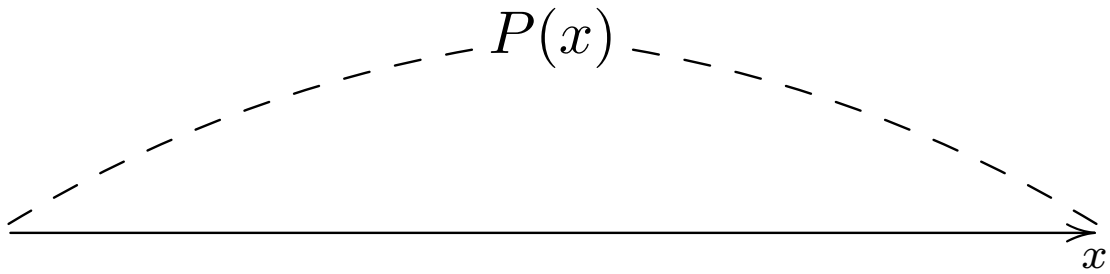
  3. *If* $p_\phi \in \mathcal{P}(X \times Y)$ *is deterministic and* $\mathbb{E}_{p_\phi}\{u\} = \mathbb{E}_{p(\beta)}\{u\}$, *then*

     $$I(p_\phi, q) > I(p(\beta), q)$$

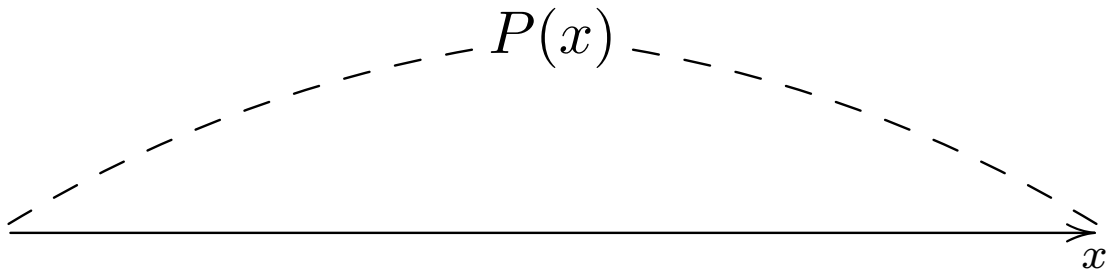
Example: Optimal Train Station



Example: Optimal Train Station

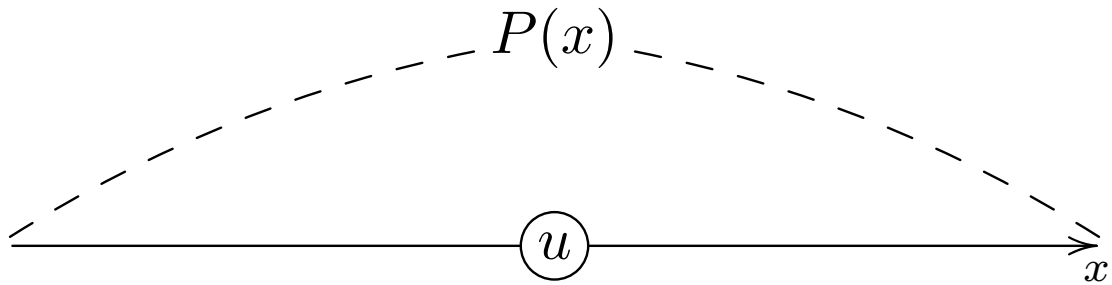


Example: Optimal Train Station



- Cost function $d(x, u) = |x - u|^2$

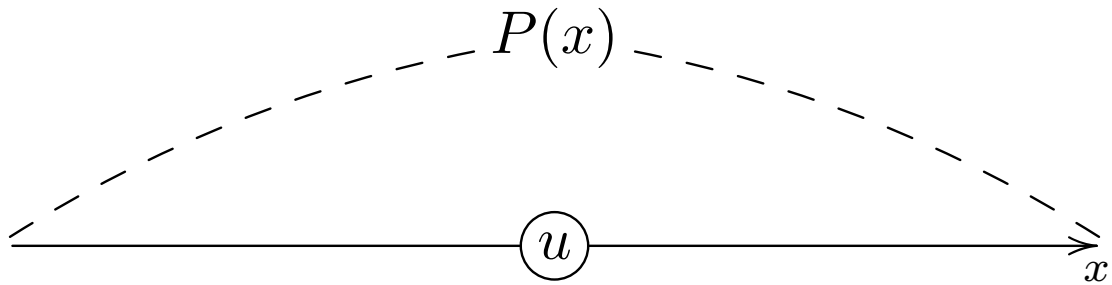
Example: Optimal Train Station



- Cost function $d(x,u) = |x-u|^2$
- Find u minimizing

$$\mathbb{E}_P\{d(x,u)\} = \int_{-\infty}^{\infty} |x-u|^2 \, dP(x)$$

Example: Optimal Train Station

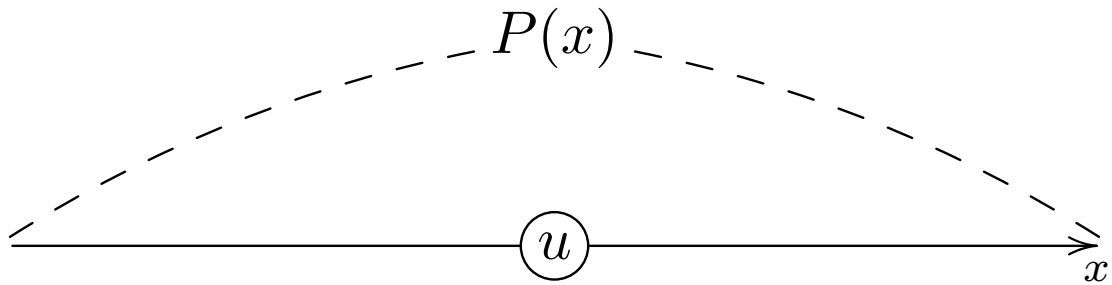


- Cost function $d(x,u) = |x-u|^2$
- Find u minimizing

$$\mathbb{E}_P\{d(x,u)\} = \int_{-\infty}^{\infty} |x-u|^2 \, dP(x)$$

- Standard solution $u = \mathbb{E}_P\{x\}$

Example: Optimal Train Station

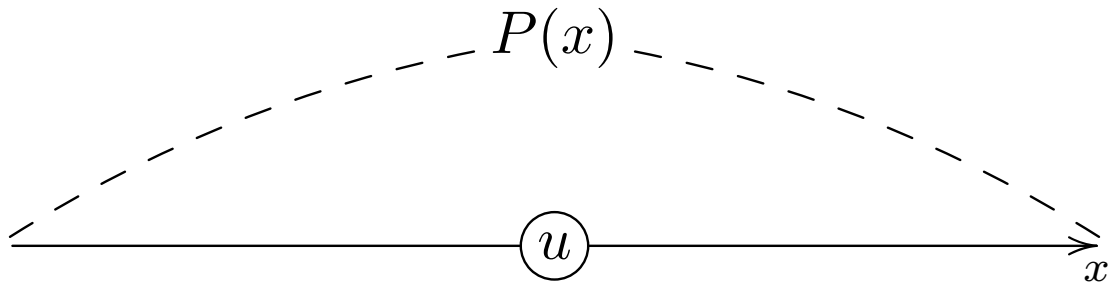


- Cost function $d(x,u) = |x-u|^2$
- Find u minimizing

$$\mathbb{E}_P\{d(x,u)\} = \int_{-\infty}^{\infty} |x-u|^2 \, dP(x)$$

- Standard solution $u = \mathbb{E}_P\{x\}$
- (or $u = \arg\max P(x)$ if expectation of x and mode coincide).

Example: Optimal Train Station



- Cost function $d(x,u) = |x-u|^2$
- Find u minimizing

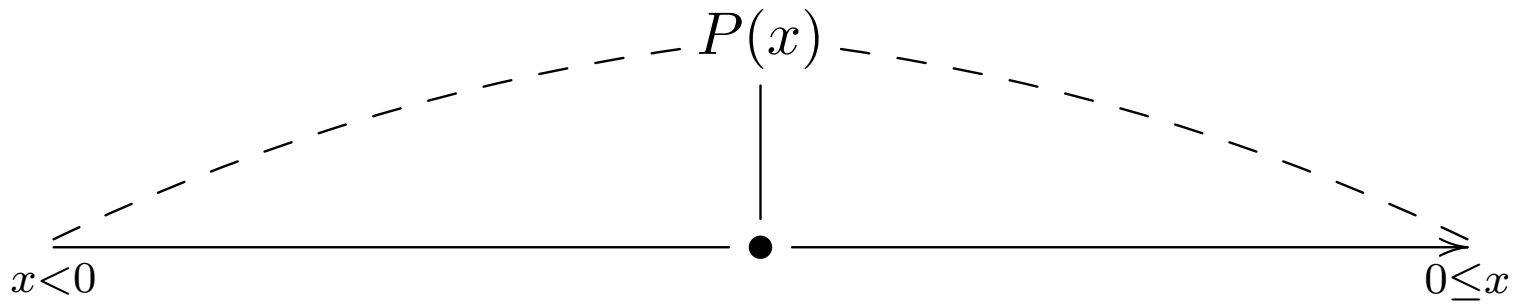
$$\mathbb{E}_P\{d(x,u)\} = \int_{-\infty}^{\infty} |x-u|^2 \, dP(x)$$

- Standard solution $u = \mathbb{E}_P\{x\}$
- (or $u = \arg\max P(x)$ if expectation of x and mode coincide).

Remark

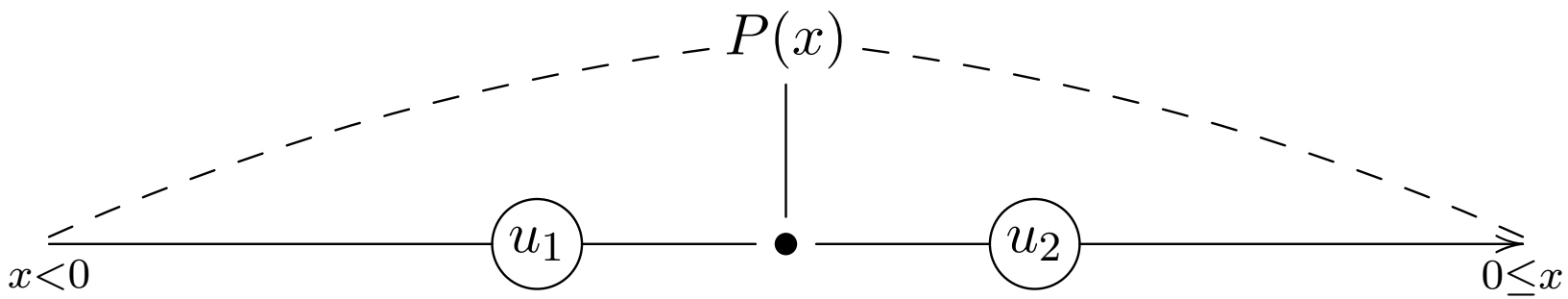
If x has Cauchy distribution $dP(x) = [\pi(x^2+1)]^{-1} \, dx$, then $\mathbb{E}_P\{d(x,u)\} = \infty$ for any u!

Example: Multiple Optimal Train Stations



- Variable $y \in \{y_1, y_2\}$ communicates information $x < 0$ or $0 \leq x$

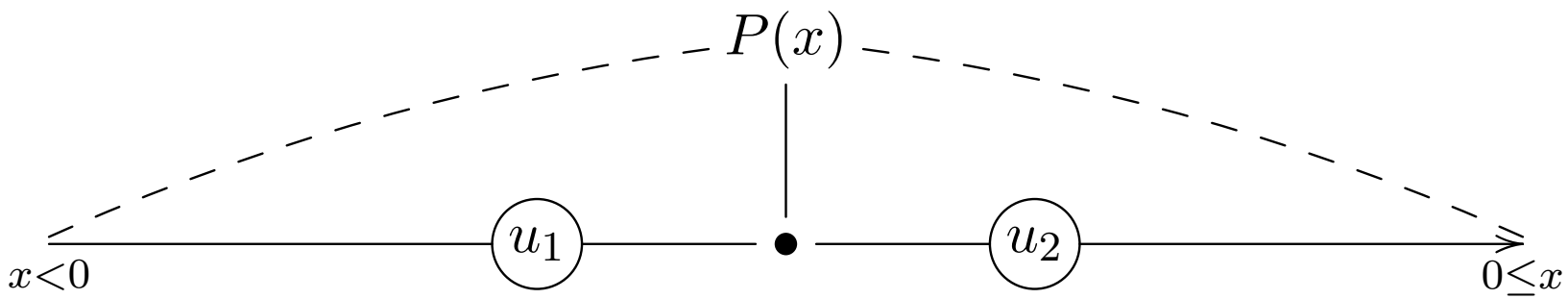
Example: Multiple Optimal Train Stations



- Variable $y \in \{y_1, y_2\}$ communicates information $x < 0$ or $0 \leq x$
- Find $u \in \{u_1, u_2\}$ minimizing

$$\mathbb{E}_P\{d(x, u_1) \mid y_1\} + \mathbb{E}_P\{d(x, u_2) \mid y_2\}$$

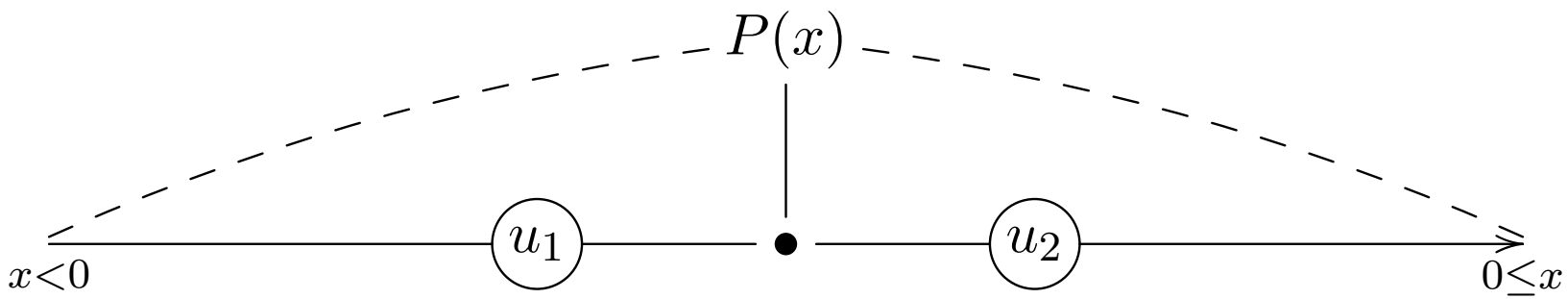
Example: Multiple Optimal Train Stations



- Variable $y \in \{y_1, y_2\}$ communicates information $x < 0$ or $0 \leq x$
- Find $u \in \{u_1, u_2\}$ minimizing

$$\mathbb{E}_P\{d(x, u_1) \mid y_1\} + \mathbb{E}_P\{d(x, u_2) \mid y_2\}$$

Example: Multiple Optimal Train Stations

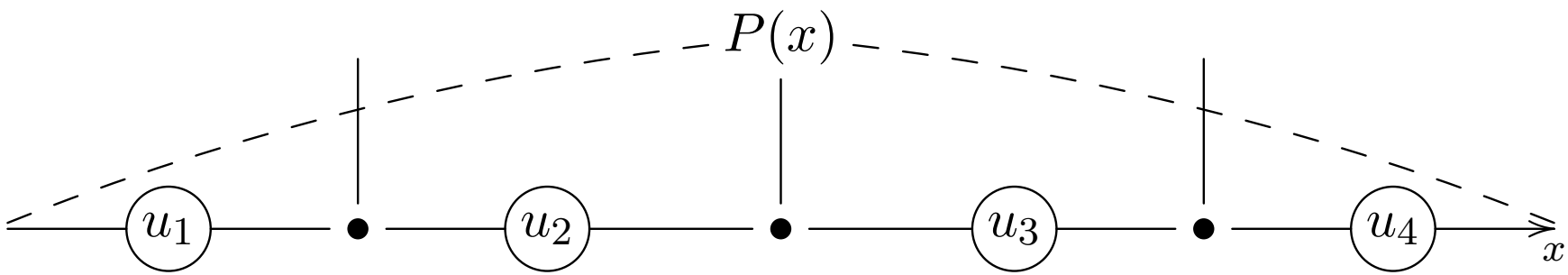


- Variable $y \in \{y_1, y_2\}$ communicates information $x < 0$ or $0 \leq x$
- Find $u \in \{u_1, u_2\}$ minimizing

$$\mathbb{E}_P\{d(x,u_1) \mid y_1\} + \mathbb{E}_P\{d(x,u_2) \mid y_2\}$$
$$= \int_{-\infty}^{0} |x-u_1|^2 \, dP(x \mid y_1) + \int_{0}^{\infty} |x-u_2|^2 \, dP(x \mid y_2)$$

- For $y \in \{y_1, \ldots, y_k\}$ find $\min_{u_1,\ldots,u_k} \sum_{i=1}^{k} \mathbb{E}_P\{d(x,u_i) \mid y_i\}$

Example: Multiple Optimal Train Stations

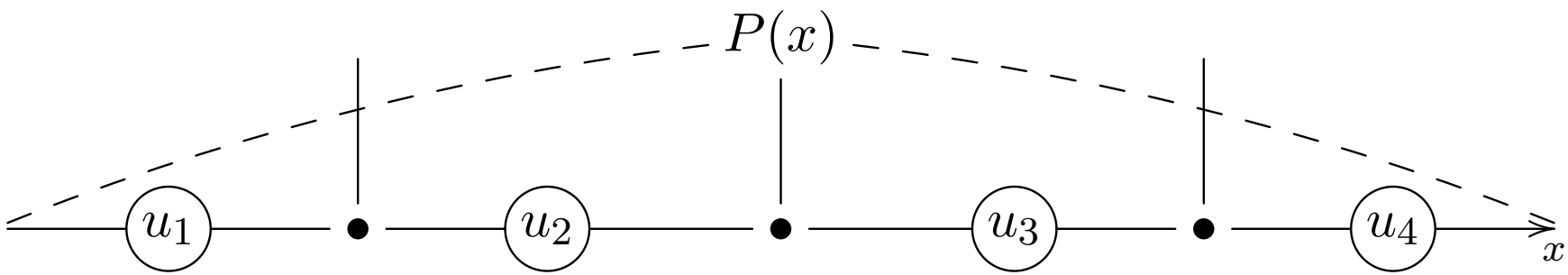


- Variable $y \in \{y_1, y_2\}$ communicates information $x < 0$ or $0 \leq x$
- Find $u \in \{u_1, u_2\}$ minimizing

$$\mathbb{E}_P\{d(x, u_1) \mid y_1\} + \mathbb{E}_P\{d(x, u_2) \mid y_2\}$$
$$= \int_{-\infty}^{0} |x - u_1|^2 \, dP(x \mid y_1) + \int_{0}^{\infty} |x - u_2|^2 \, dP(x \mid y_2)$$

- For $y \in \{y_1, \ldots, y_k\}$ find $\min_{u_1,\ldots,u_k} \sum_{i=1}^{k} \mathbb{E}_P\{d(x, u_i) \mid y_i\}$

Example: Multiple Optimal Train Stations



- Variable $y \in \{y_1, y_2\}$ communicates information $x < 0$ or $0 \leq x$
- Find $u \in \{u_1, u_2\}$ minimizing

$$\mathbb{E}_P\{d(x,u_1) \mid y_1\} + \mathbb{E}_P\{d(x,u_2) \mid y_2\}$$
$$= \int_{-\infty}^{0} |x-u_1|^2 \, dP(x \mid y_1) + \int_{0}^{\infty} |x-u_2|^2 \, dP(x \mid y_2)$$

- For $y \in \{y_1, \ldots, y_k\}$ find $\min_{u_1,\ldots,u_k} \sum_{i=1}^{k} \mathbb{E}_P\{d(x,u_i) \mid y_i\}$

Remark

If x has Cauchy distribution $dP(x) = [\pi(x^2+1)]^{-1} \, dx$, then
$\sum_{i=1}^{k} \mathbb{E}_P\{d(x,u_i) \mid y_i\} = \infty$ for any $k \in \mathbb{N}$ (i.e. any finite partition of $\mathbb{R}$)

Overwhelming an Algorithm

- Let X be a countable set of problems, and let $f(x)$ be a deterministic algorithm with time complexity $O(n^m)$.

Overwhelming an Algorithm

- Let X be a countable set of problems, and let $f(x)$ be a deterministic algorithm with time complexity $O(n^m)$.
- Distribution $P(x)$ on problems induces $P(n) = P\{x : x \text{ is of size } n\}$

Overwhelming an Algorithm

- Let X be a countable set of problems, and let $f(x)$ be a deterministic algorithm with time complexity $O(n^m)$.
- Distribution $P(x)$ on problems induces $P(n) = P\{x : x \text{ is of size } n\}$
- The expected time $\mathbb{E}\{t\}$ of a deterministic algorithm depends entirely on $P(n)$:

$$\mathbb{E}_P\{t\} \leq \sum_{n \in \mathbb{N}} n^m P(n)$$

Overwhelming an Algorithm

- Let X be a countable set of problems, and let $f(x)$ be a deterministic algorithm with time complexity $O(n^m)$.
- Distribution $P(x)$ on problems induces $P(n) = P\{x : x \text{ is of size } n\}$
- The expected time $\mathbb{E}\{t\}$ of a deterministic algorithm depends entirely on $P(n)$:

$$\mathbb{E}_P\{t\} \leq \sum_{n \in \mathbb{N}} n^m P(n)$$

Remark

If $P(n) = n^{-(m+1)}/\zeta(m+1)$, *where* $\zeta(m+1) = \sum_{n \in \mathbb{N}} n^{-(m+1)}$, *then*

$$\mathbb{E}_P\{t\} = \zeta(1)/\zeta(m+1) = \infty$$

Example: Genetic search in Hamming Space

Example (Hamming metric)

Let $X = \{1, \dots, \alpha\}$ with Hamming metric d_H:

$$d(x, y) = l - \sum_{i=1}^{l} \delta_{x_i}(y_i)$$

Example: Genetic search in Hamming Space

Example (Hamming metric)

Let $X = \{1, \dots, \alpha\}$ with Hamming metric d_H:

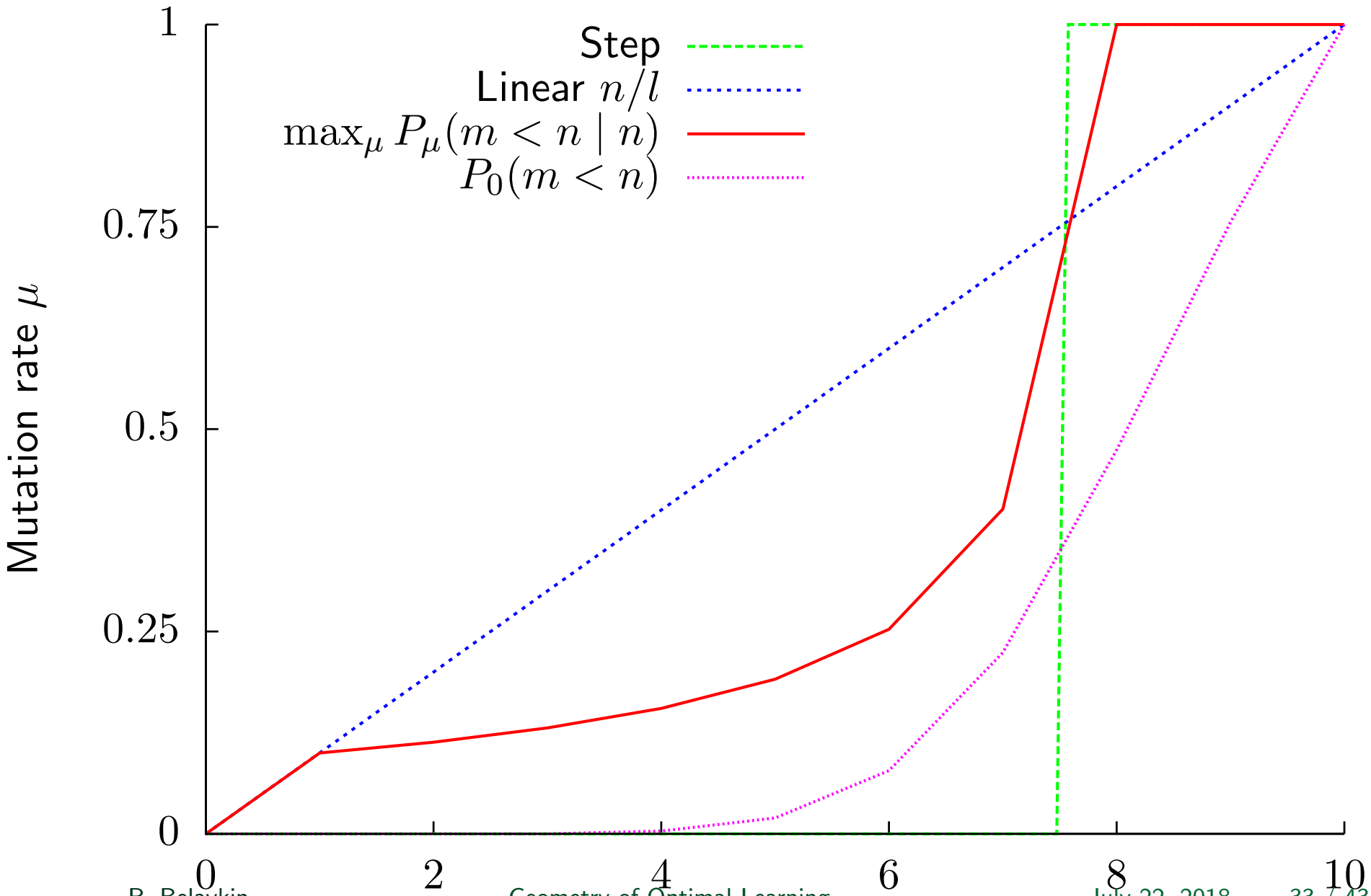
$$d(x, y) = l - \sum_{i=1}^{l} \delta_{x_i}(y_i)$$

Solution

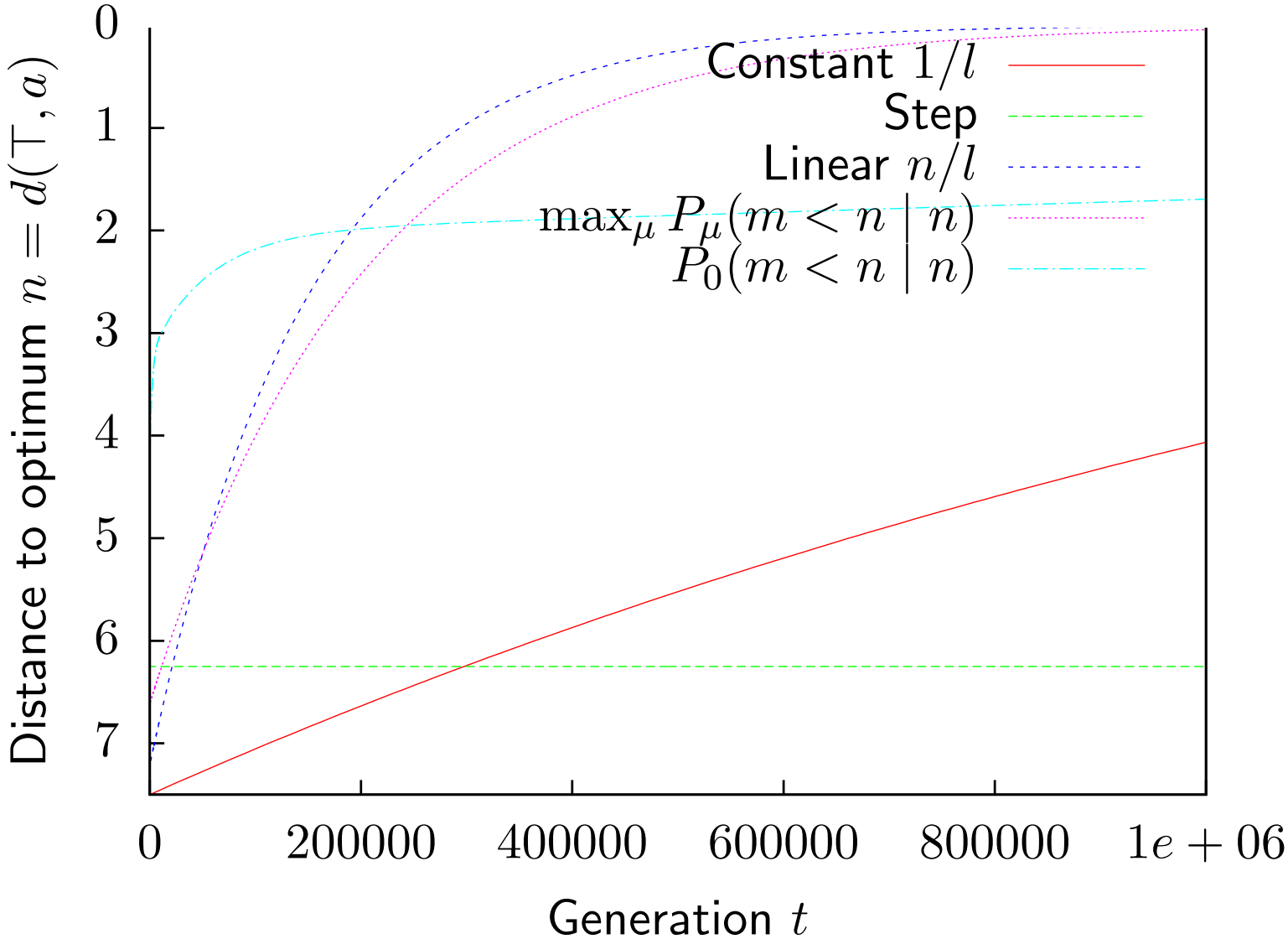
$$P_\beta(y \mid x) = \frac{e^{-\beta\, d(x,y)}}{[1 + (\alpha - 1)e^{-\beta}]^l} = \prod_{i=1}^{l} \frac{e^{-\beta\, \delta_{x_i}(y_i)}}{1 + (\alpha - 1)e^{-\beta}}$$

$\beta = \ln\left(\mu^{-1} - 1\right) + \ln(\alpha - 1)$, where $\mu = \upsilon/l$ is the mutation rate, defined by the constraint $\mathbb{E}\{d(x, y)\} \leq \upsilon$.

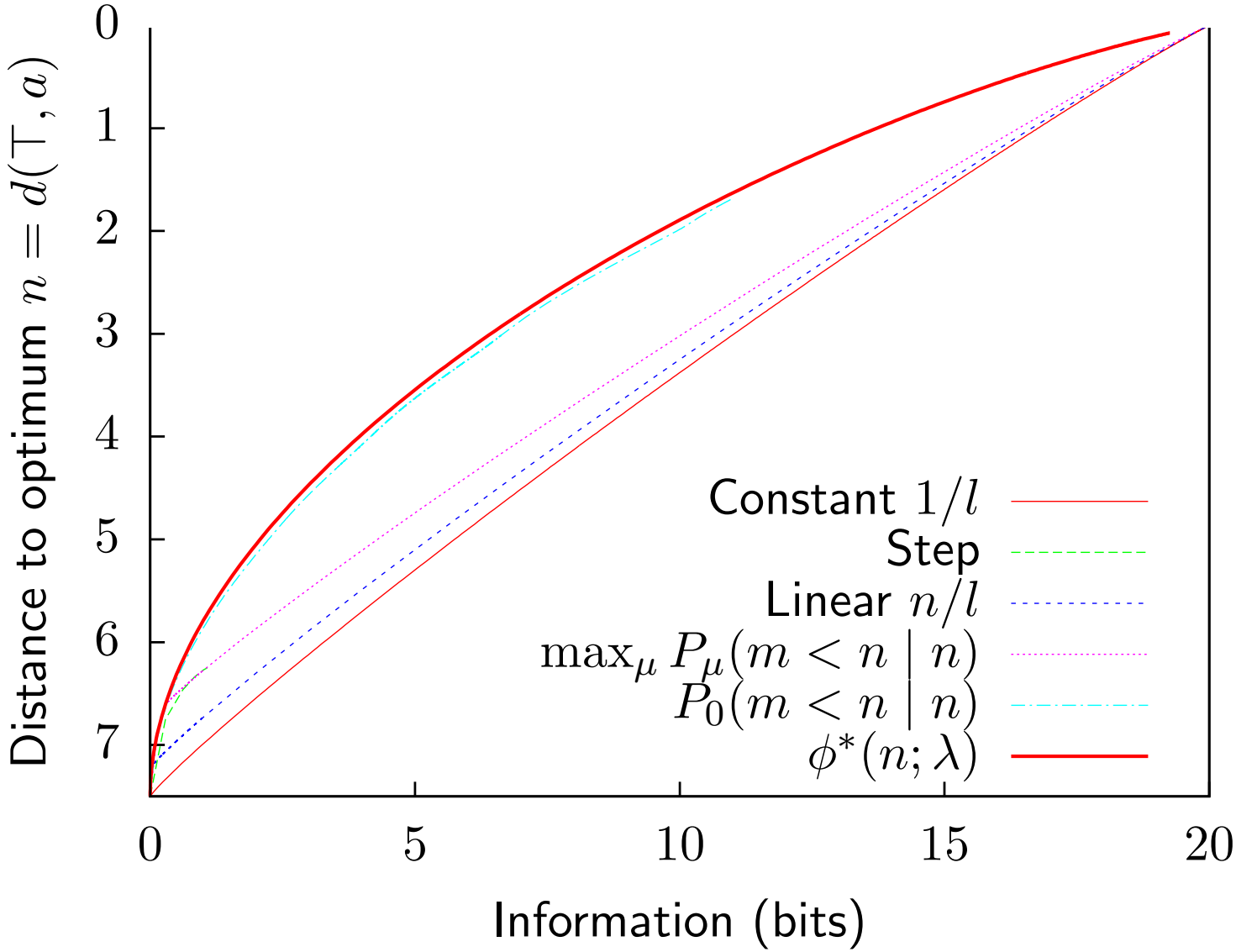
Mutation rate control functions in $\mathcal{H}_4^{10}$



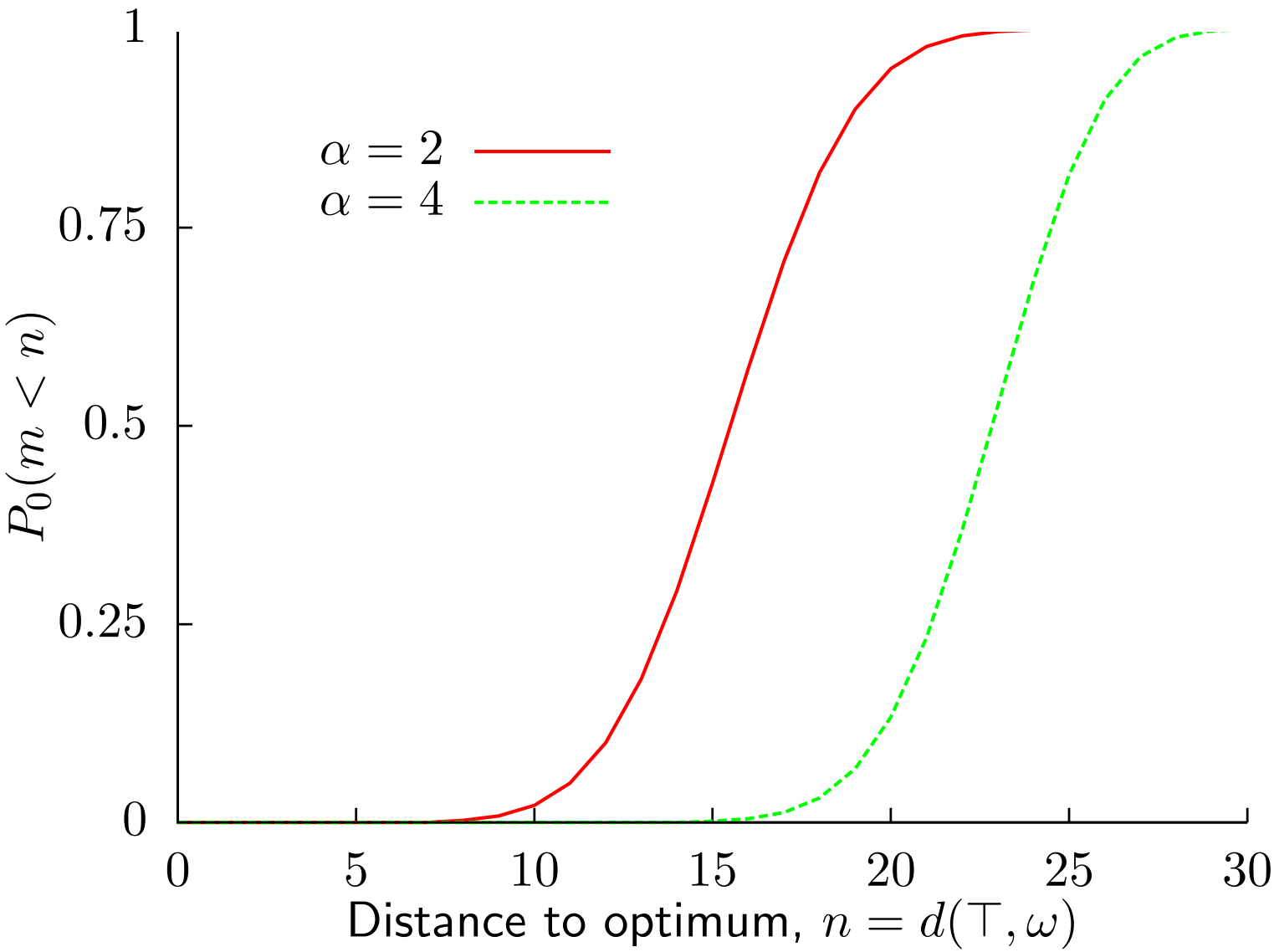
Expected Fitness in Time



Expected Fitness in Information



CDF Control


1
0.75
0.5
0.25
0
$P_0(m < n)$
$\alpha = 2$
$\alpha = 4$
0
5
10
15
20
25
30
Distance to optimum, $n = d(\top, \omega)$

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

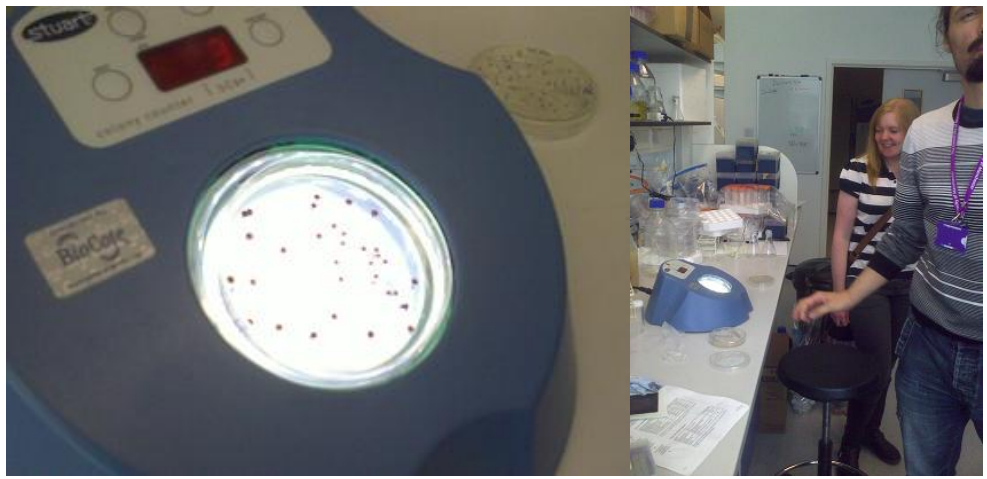
Evolution as an Information Dynamic System

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

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

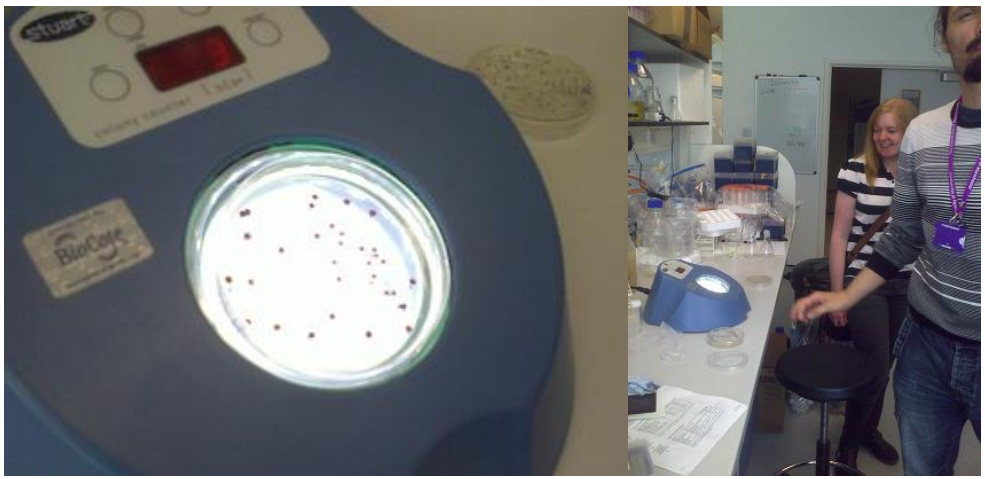
  Middlesex University : Roman Belavkin
  University of Warwick : John Aston
  University of Keele : Alastair Channon & Elizabeth Aston
  University of Manchester : Chris Knight, Rok Krašovec & Danna Gifford

Mutation Rate Control in *E. coli*



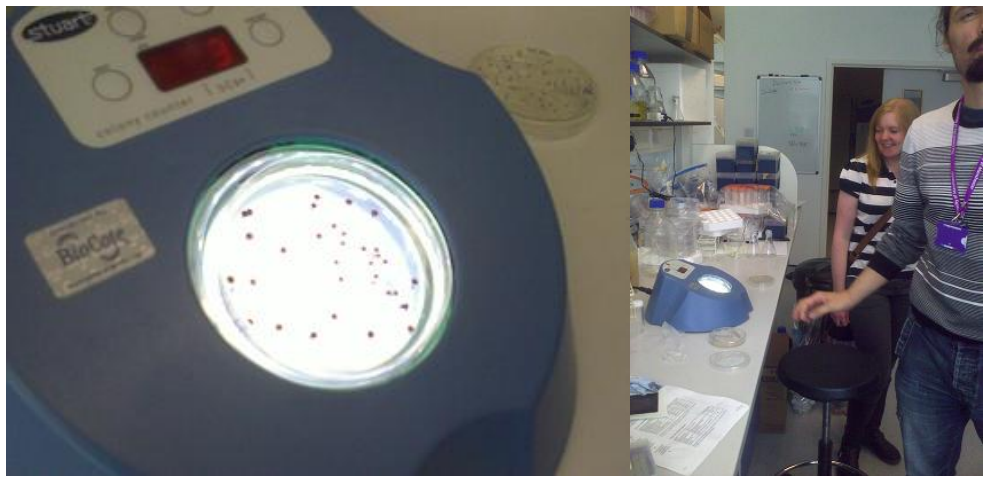
- Used strains of *Escherichia coli* K-12 MG1665

Mutation Rate Control in *E. coli*



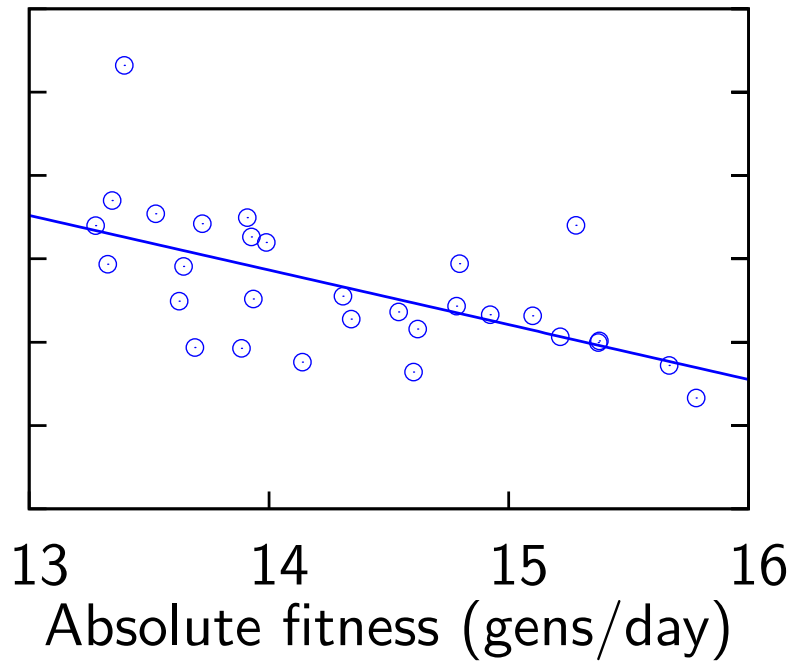
- Used strains of *Escherichia coli* K-12 MG1665
- Fluctuation test using media 50μg/ml of Rifamipicin

Mutation Rate Control in *E. coli*

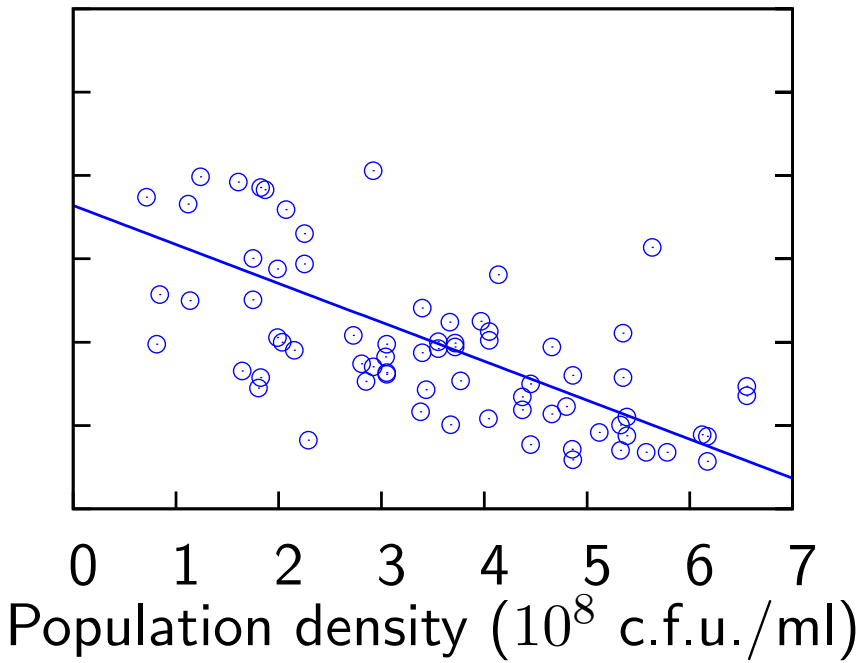


- Used strains of *Escherichia coli* K-12 MG1665
- Fluctuation test using media 50μg/ml of Rifamipicin
- Estimated mutation rates μ in *E.coli* strains grown in Davis minimal medium with different amount of glucose.

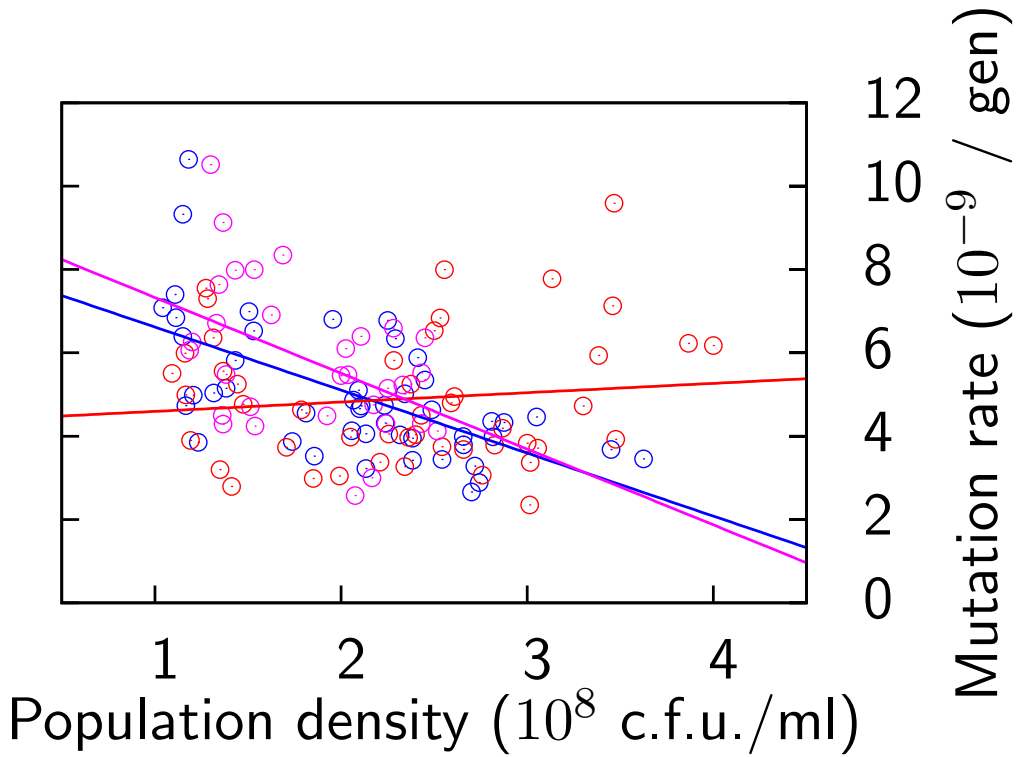
Experimental Results (Krašovec et al., 2014)



Experimental Results (Krašovec et al., 2014)

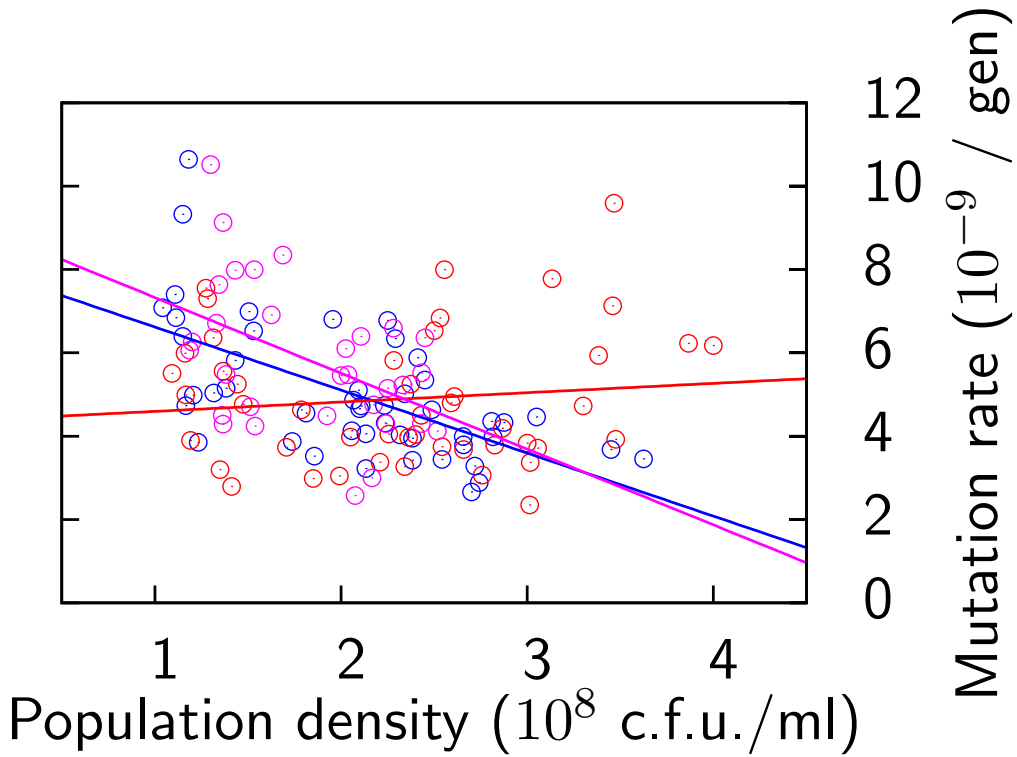


Experimental Results (Krašovec et al., 2014)



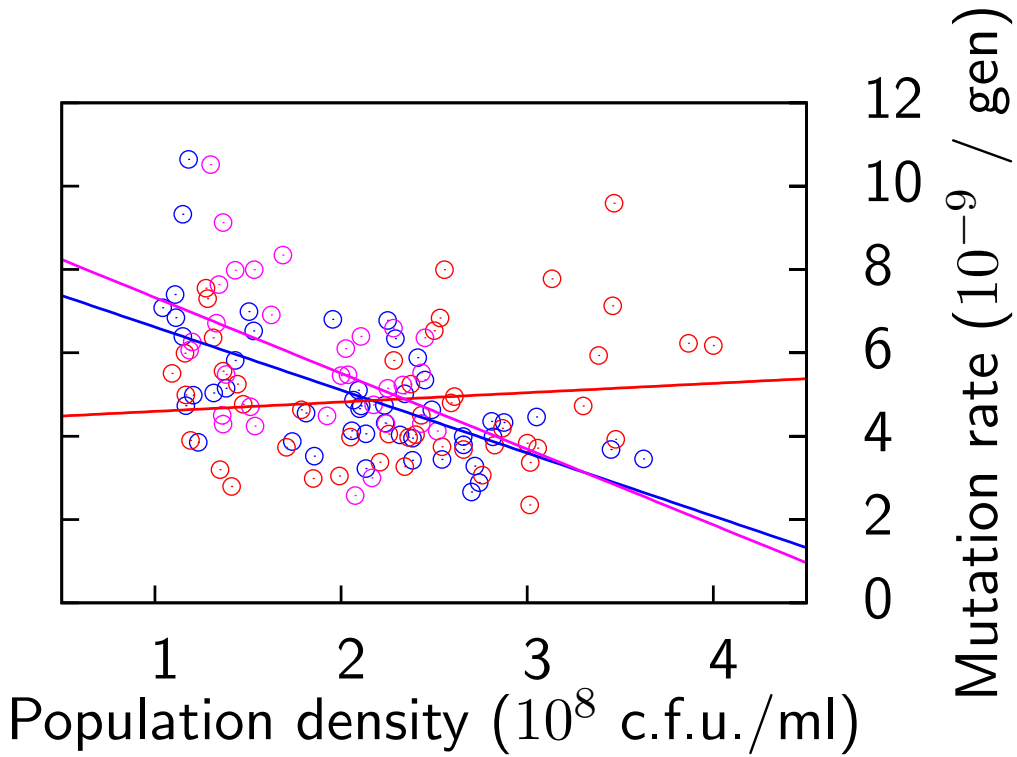
- Strong relationship between μ and density of cells ($p < .0001$).

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 ($p = .0234$).

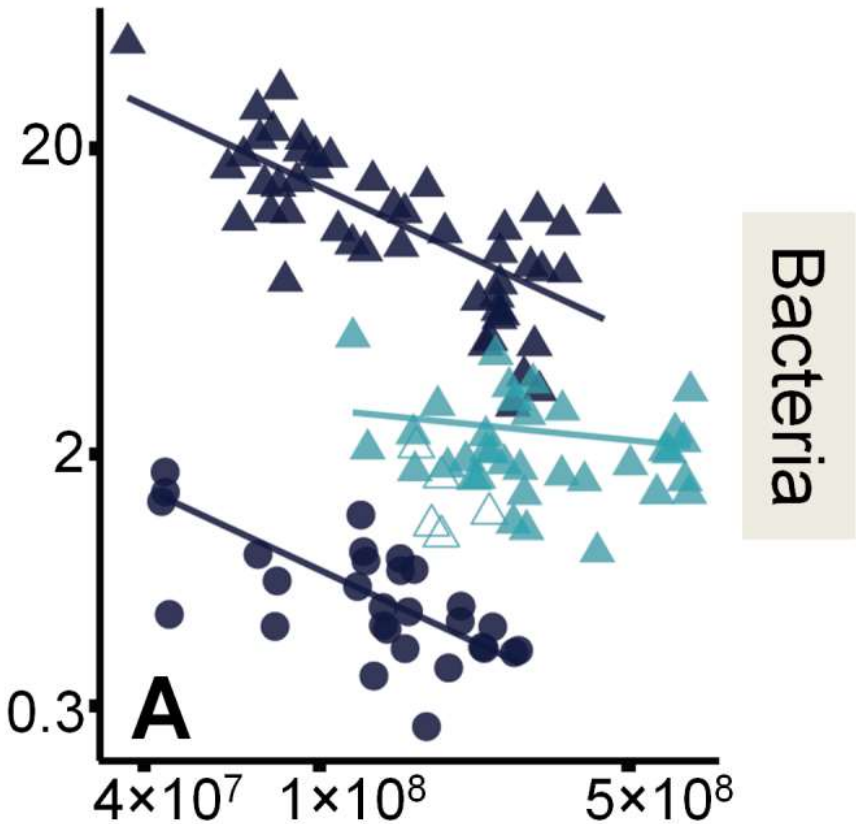
Experimental Results (Krašovec et al., 2014)



- 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).

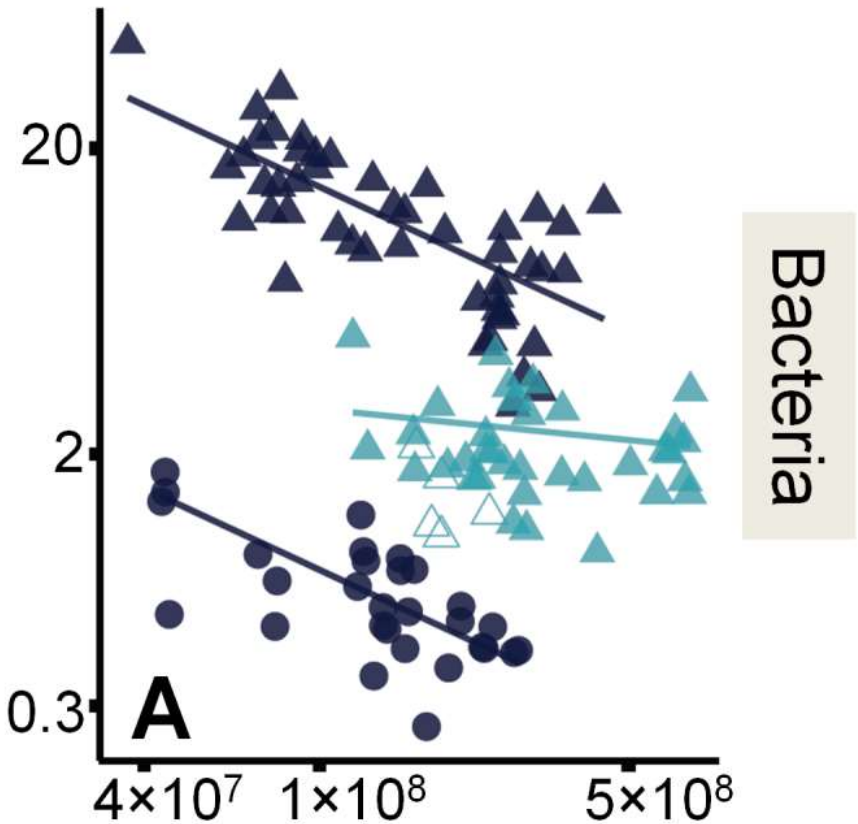
Plastic mutation rates in bacteria (Krašovec et al., 2017)



- *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.

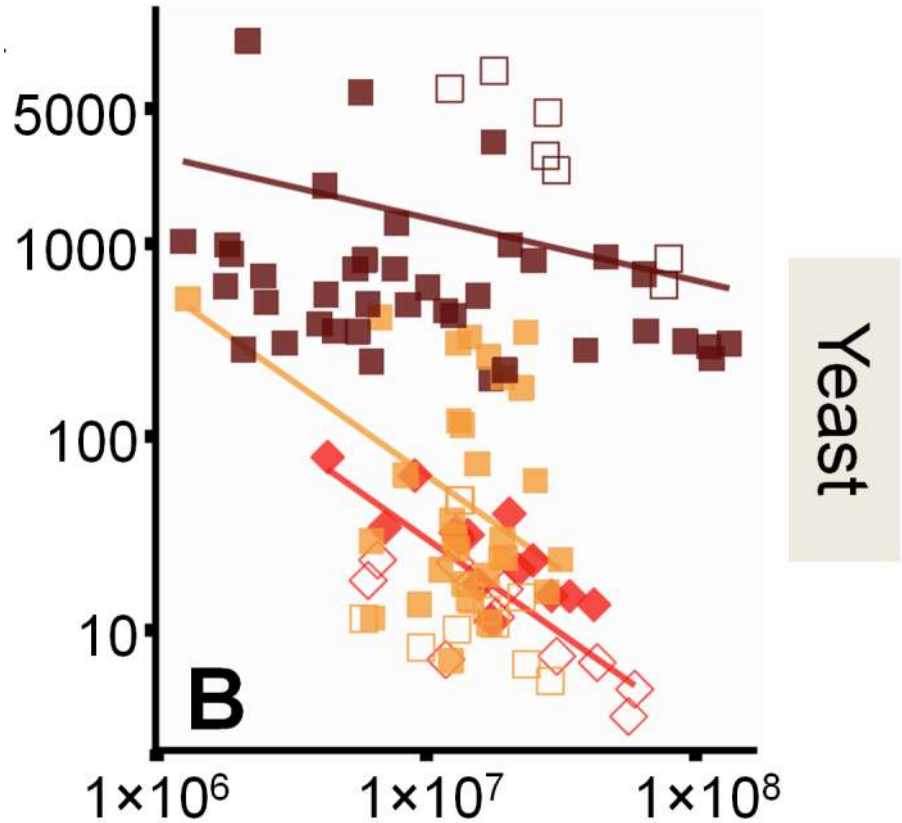
Plastic mutation rates in bacteria (Krašovec et al., 2017)



- *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.

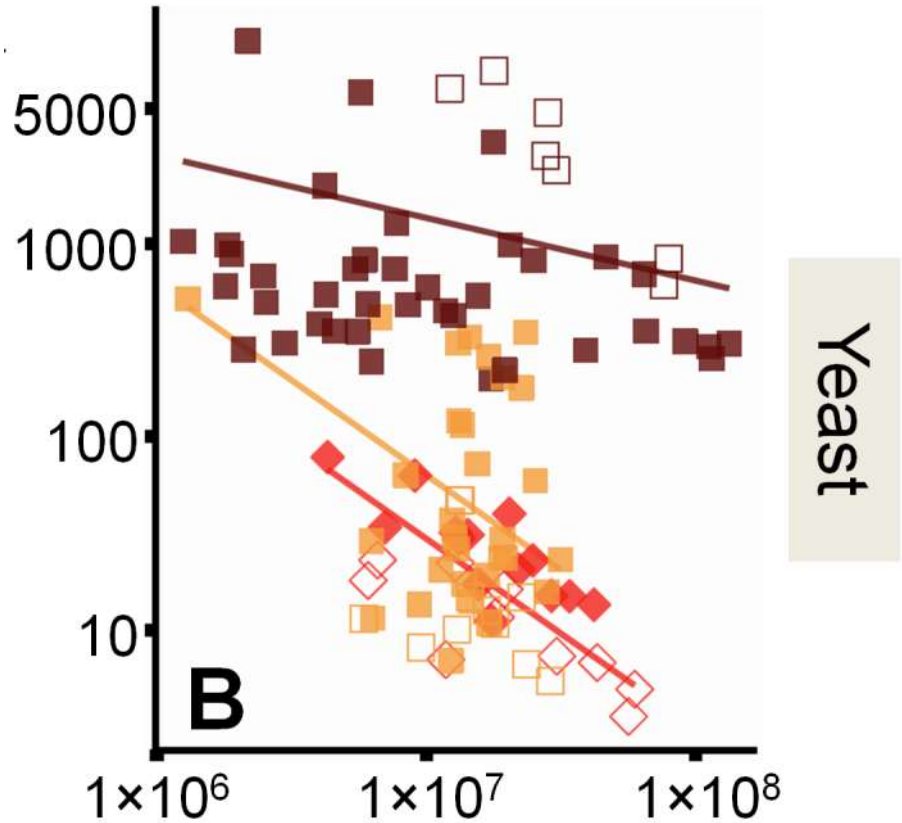
Plastic mutation rates in yeast (Krašovec et al., 2017)



- *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.

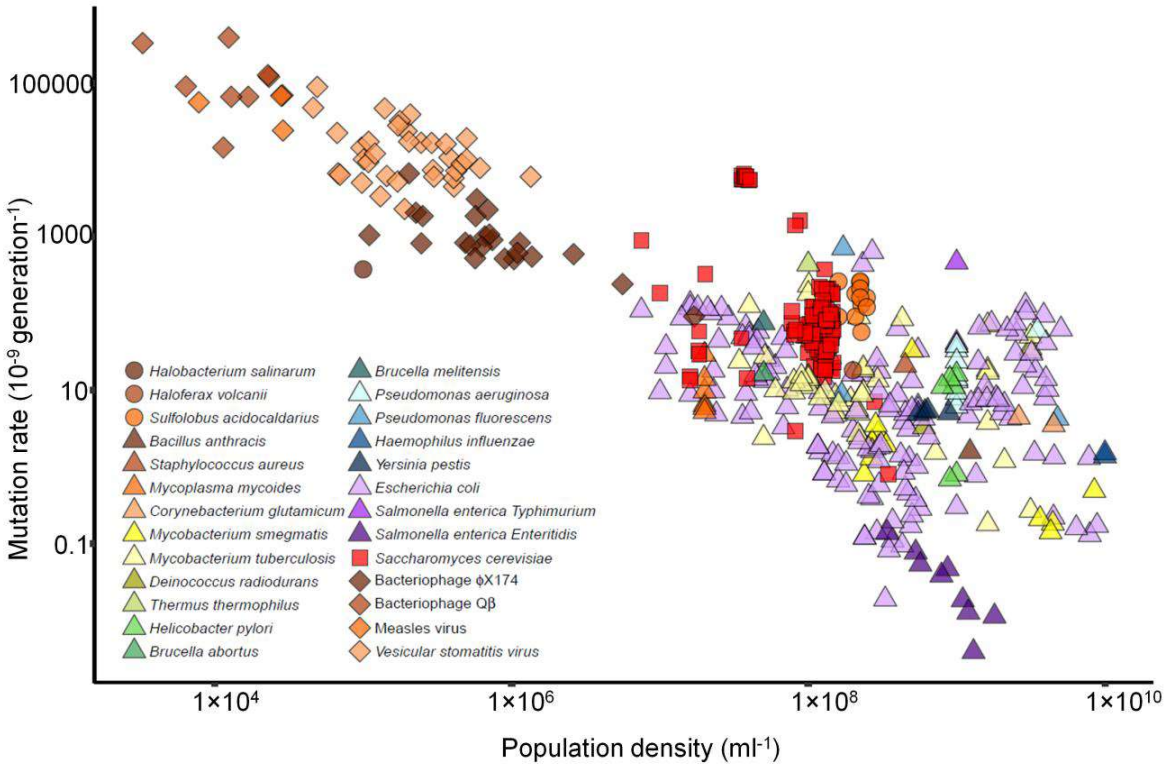
Plastic mutation rates in yeast (Krašovec et al., 2017)



- *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 (Kraševec et al., 2017)



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

Kraševec, 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.

Optimal learning

First variational problem

Value of information

Iterative algorithms as Markov morphisms

Optimal transition kernels

Optimal control of mutation rate

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

Krašovec, R., Belavkin, R. V., Aston, J. A. D., Channon, A., Aston, E., Rash, B. M., et al. (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., et al. (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. (1968). Is there a theory for a synthesis of optimal adaptive, self-learning and self-organising systems? *Automatics and Telemechanics*, *1*. (In Russian)