# Efficient computation and data representation

Susanne Still
University of Hawai‘i at Mānoa

# I

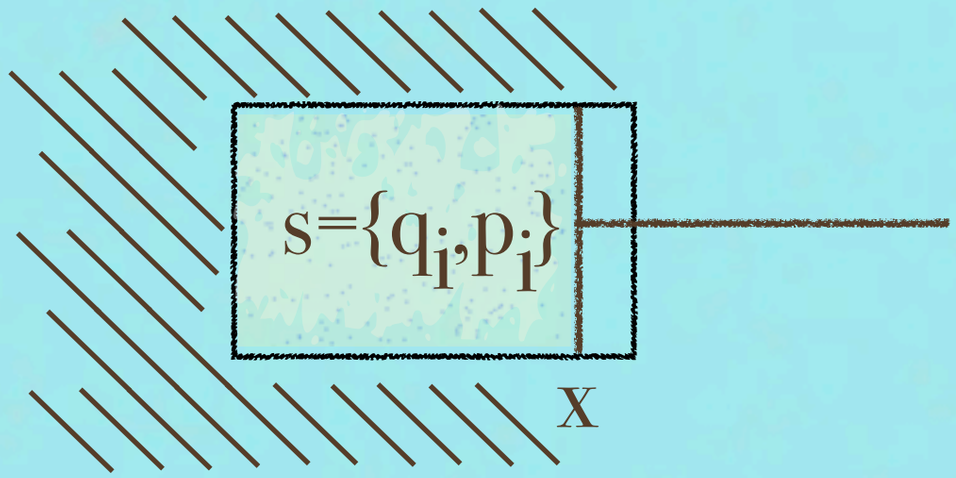
# Second law of thermodynamics and Landauer's bound

In closed system, entropy does not decrease.
Dissipated work is (on average) non-negative.
Gives lower bound on heat generated when information is erased.

# Equilibrium Thermodynamics (quick review)

- The state (variable s) of a system in thermodynamic equilibrium, given an environmental parameter, x.
  Best described by the Boltzmann distribution:

$$p_{\mathrm{eq}}(s|x) = e^{-\beta[E(s,x)-F(x)]} \qquad \beta = \frac{1}{k_B T}$$




Example: Gas in box with piston. x = piston position, s = positions and momenta of molecules. Temperature T.

- This is the maximum entropy distribution, given the known quantities [here: the average energy $E = \langle E(s,x) \rangle$]. (Jaynes 1957)
- Entropy: $S = k_B H[p_{\mathrm{eq}}] = -k_B \left\langle \ln[p_{\mathrm{eq}}] \right\rangle_{p_{\mathrm{eq}}}$
- (Equilibrium) Free Energy: $F = E - TS$

1. Start system in thermodynamic equilibrium t=0

$$p(s_0|x_0) = e^{-\beta[E(s_0,x_0)-F(x_0)]}$$

2. Drive system, performing work, W.

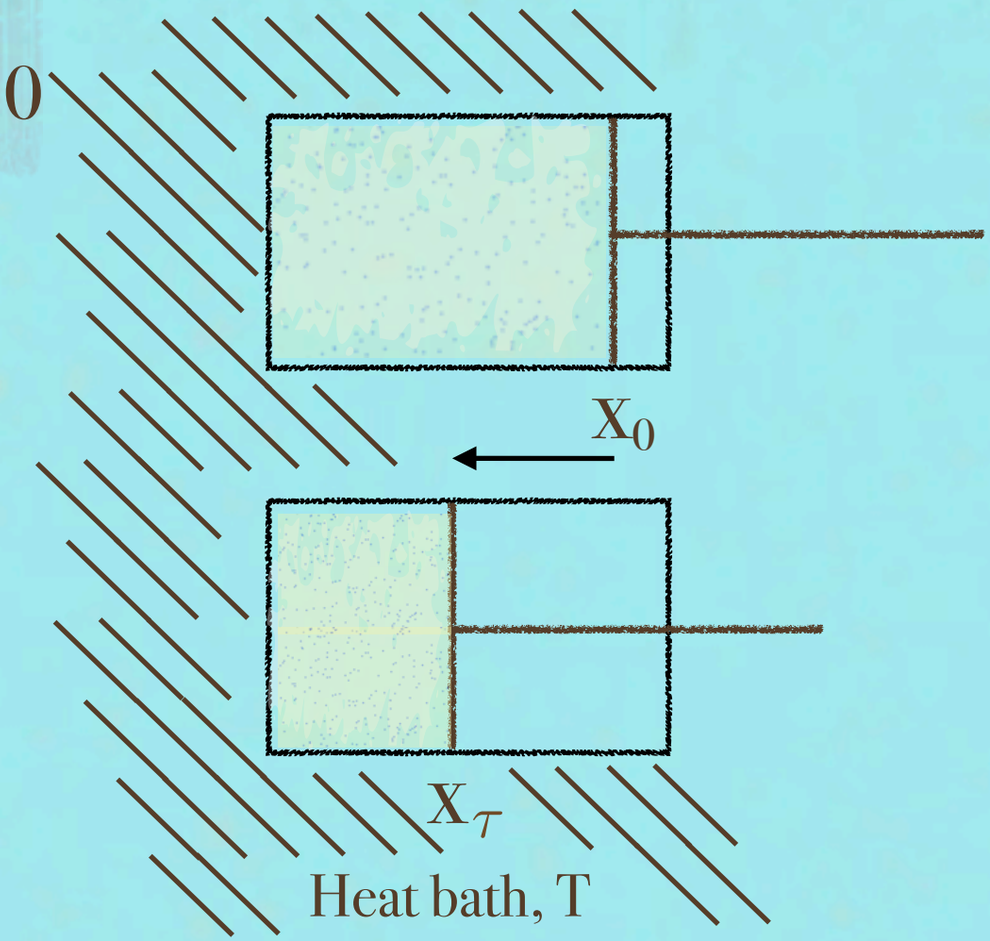
   Contact to heat reservoir at temperature T.

   Heat flowing into gas = Q.

3. Let system relax back to equilibrium. At t=$\tau$:

$$p(s_\tau|x_\tau) = e^{-\beta[E(s_\tau,x_\tau)-F(x_\tau)]}$$

   Energy is conserved: $\Delta E = W + Q$




$$\Delta E := E_\tau - E_0$$
$$\Delta F := F_\tau - F_0$$
$$\Delta S := S_\tau - S_0$$

- Free energy change: $\Delta F = \Delta E - T\Delta S$
- Average work done in excess of free energy change (dissipated work):

$$\langle W_{\rm ex}\rangle := \langle W\rangle - \Delta F = -\langle Q\rangle + T\ \Delta S\ \geq 0$$

# Landauer's principle
(1961)

- Erasing information: reset system to zero entropy state: $H_\tau = 0$

$$-\langle Q \rangle + k_B T \left( H_\tau - H_0 \right) \geq 0 \quad \Rightarrow \quad -\langle Q \rangle \geq k_B T H_0$$

- When one bit of information is erased, heat is produced, in the amount of *at least* kT ln(2).
- Direct consequence of the second law of thermodynamics.
- Transformation between thermodynamic equilibrium states.
- What if there is no time to stay in equilibrium?

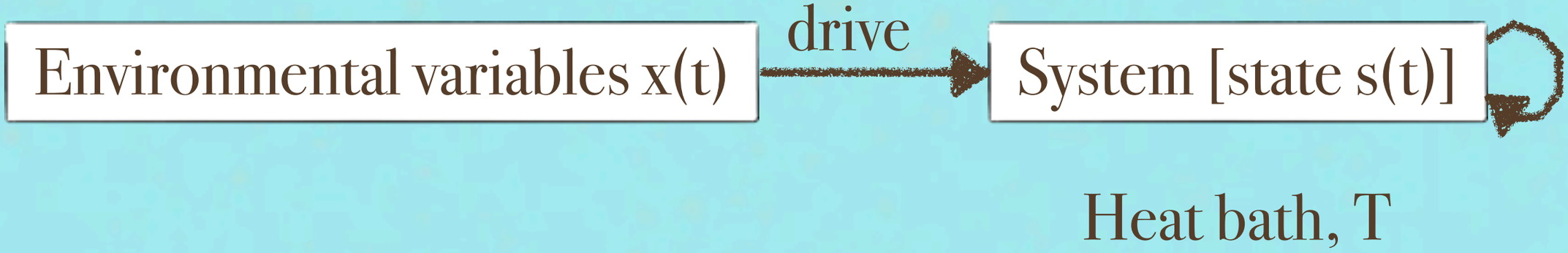
# II

# Nonpredictive information limits smallest achievable dissipation

In system driven (arbitrarily far) away from equilibrium, the dynamics matter.

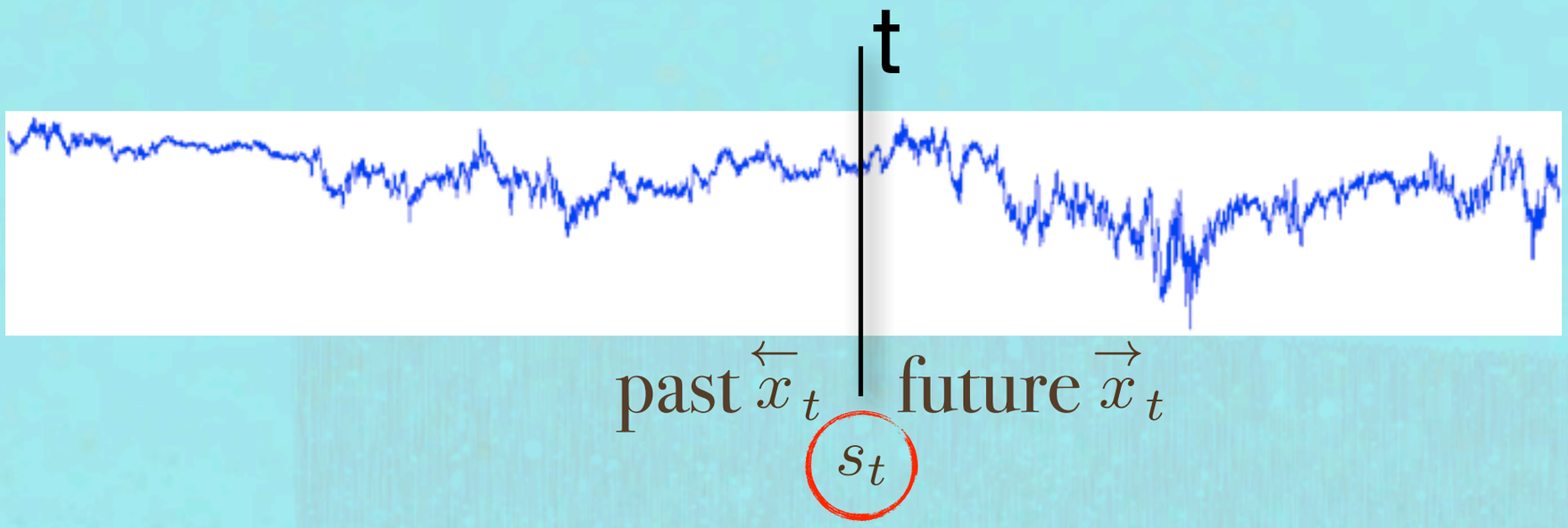
That fraction of a system's instantaneous memory about environmental signals that does not contain predictive power sets a lower bound on dissipation.

- Driven system, arbitrarily far from equilibrium.




- Environment can be stochastic.

- This is like a learning machine:



# Stochastic thermodynamics of learning machines

Still, Sivak, Bell & Crooks PRL (2012)

Signal: $x_0 \to x_1 \to \cdots \to x_{t-1} \longrightarrow x_t \to x_{t+1} \longrightarrow \cdots \to x_\tau$

System: $s_0 \to s_1 \to \cdots \longrightarrow s_{t-1} \to s_t \longrightarrow s_{t+1} \to \cdots \to s_\tau$

$$I[S_t; X_t] - I[S_t; X_{t+1}] = \beta \left\langle W_{\text{diss}}(x_t \to x_{t+1}) \right\rangle$$

Work done (on average) in excess of nonequilibrium free energy change due to change in environment is proportional to instantaneous nonpredictive information.

- This is a fundamental result that also holds for quantum systems (Grimsmo, 2013):

  - Provides a new interpretation of quantum discord as: "the thermodynamic inefficiency of the most energetically efficient classical approximation of a quantum memory".

  - There is a possible quantum advantage in terms of dissipation (when environmental signal is non-classical).

- New result: there is also a quantum predictive advantage! (Grimsmo & Still, in preperation)

# Second law and Landauer refined

Still, Sivak, Bell & Crooks PRL (2012)

- Lower bound on dissipation: $\beta \langle W_{\rm ex} \rangle \geq I_{\rm nonpred}$

( $I_{\rm nonpred}$ is the *total* instataneous nonpredictive information, summed over time.)

- Therefore: $-\beta Q \geq \mathcal{I}_{\rm erasure} + I_{\rm nonpred}$

Landauer's bound is augmented by nonpredictive information (which is a signature of the *dynamics* of the driven system)

# Motivation for efficient computation

- View system's state s as a *summary* of past environmental state(s)
- Interpret information shared with past as bit-cost
- Interpret information shared with future environmental state as predictive power
- Then: to minimize the lower bound on dissipation at fixed predictive power, bit-cost should be small.

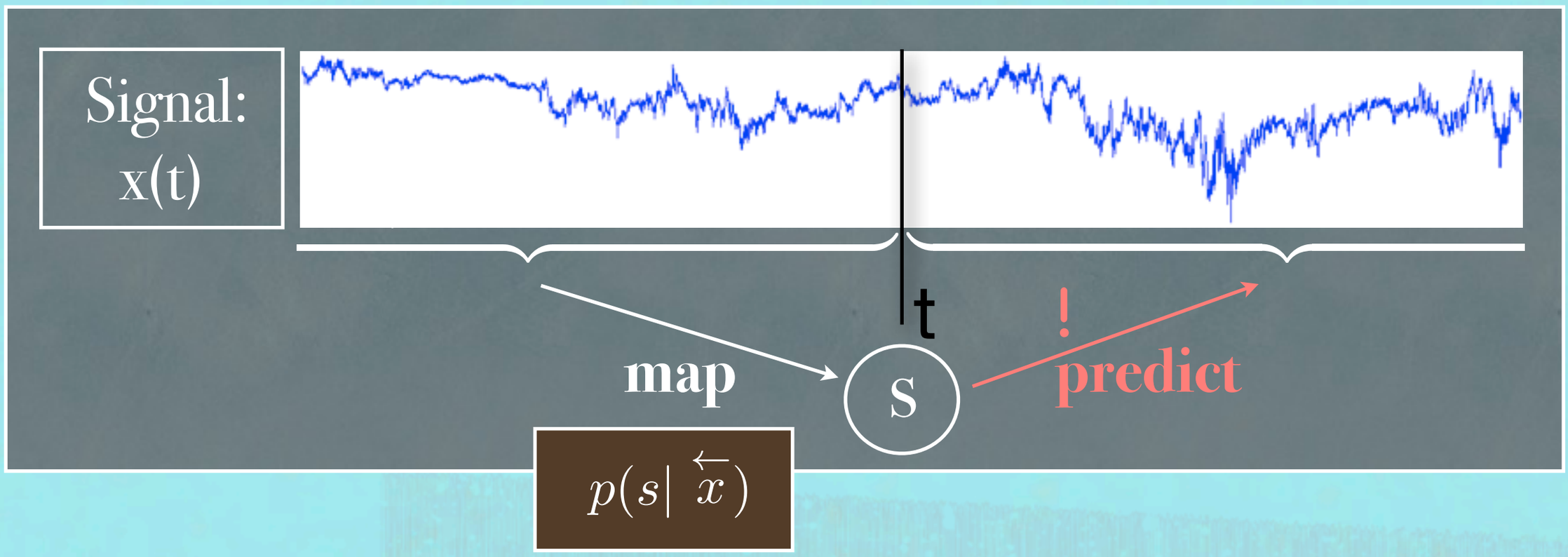
# III

# Efficient data representation

Basic challenge in communication and in data modeling:
represent continuous signal by discrete variable
in an efficient  and meaningful way
(coarse graining, clustering, abstraction).

# Optimal predictive inference

- Follow our train of thought: summarize past data by s;
  achieve smallest bit-cost at fixed predictive power.



- Summarize past data by s. Achieve smallest bit-cost at fixed predictive power:

$$\min_{p(s|\overleftarrow{x})} I[s;\overleftarrow{x}] \quad \text{s.t. } I[s;\overrightarrow{x}] \geq I$$

$$\Leftrightarrow \quad \min_{p(s|\overleftarrow{x})} \Big( I[s;\overleftarrow{x}] - \alpha I[s;\overrightarrow{x}] \Big)$$

- Maximally predictive models (at fixed memory):

$$\Rightarrow \quad p_{\rm opt}(s|\overleftarrow{x}) = \frac{p_{\rm opt}(s)}{Z(\overleftarrow{x})} e^{-\alpha D[p(\overrightarrow{x}|\overleftarrow{x})\|p(\overrightarrow{x}|s)]}$$

*Relative entropy:*

$$D[p\|q] := \left\langle \log \left[\frac{p}{q}\right] \right\rangle_p$$

- Can be used to model dynamical systems. Finds "causal states" (sufficient statistics) in the limit $\alpha \to 0$ (Still, Crutchfield and Ellison, 2010)

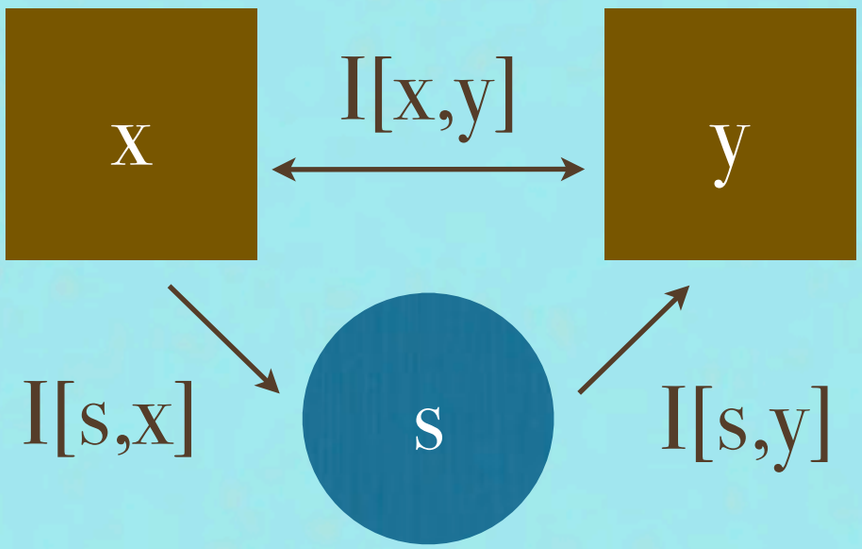
(equivalence relation: $\overleftarrow{x} \sim \overleftarrow{x}'$ if $p(\overrightarrow{x} \mid \overleftarrow{x}) = p(\overrightarrow{x} \mid \overleftarrow{x}')$ )

(Crutchfield&Young 1989, Milner "probabilistic bipartition"1986/89)

# Information Bottleneck method

(Tishby, Pereira, Bialek, 1999)

- Given a relevant quantity, y, and co-occurance statistics p(x,y)
- Achieve smallest bit-cost while retaining relevant information:




$$\min_{p(s|x)} \left( I[s,x] - \alpha I[s,y] \right) \Rightarrow \quad p_{\mathrm{opt}}(s|x) = \frac{p_{\mathrm{opt}}(s)}{Z(x)} e^{-\alpha D[p(y|x)\|p(y|s)]}$$

# Information Bottleneck method (IB)

- Solutions found numerically by iterative algorithm.
- Useful method for clustering, e.g. document classification.
- Learning: estimate of p(x,y) is subject to finite sample errors!
- Correct resulting overestimate in relevant information.
  Complexity control, e.g. estimate number of clusters
  (Still&Bialek 2004).

# How does this relate to…?

- Rate distortion theory - can be mapped onto it (next slides)
- Thermodynamic work - simple relationship (slides)
- MaxEnt - (talk to me at lunch)
- Cluster analysis - (talk to me at lunch)
- Generalization: dynamical and interactive learning (slides, self-study and/or talk to me)

# Rate-distortion theory (RDT)
(Shannon 1948)

- Continuous signal has infinite information rate.
- But infinite resolution is irrelevant for most applications, some level of distortion is tolerable.
- What is the rate of a continuous information source, if transmitted to finite resolution, i.e. for fixed average distortion?

# Rate-distortion curve

- Represent original signal, x, by encoded signal, s.
- Given: distortion function d(s,x); information source p(x).
- Achievable rate at fixed average distortion:

$$R(D) := \min_{p(s|x)} I[s,x]$$
$$\text{s.t.} \langle d(s,x) \rangle_{p(s,x)} = D$$

- (units: convert between information, in bits per symbol, and rate by multiplication with a constant: symbols per second)

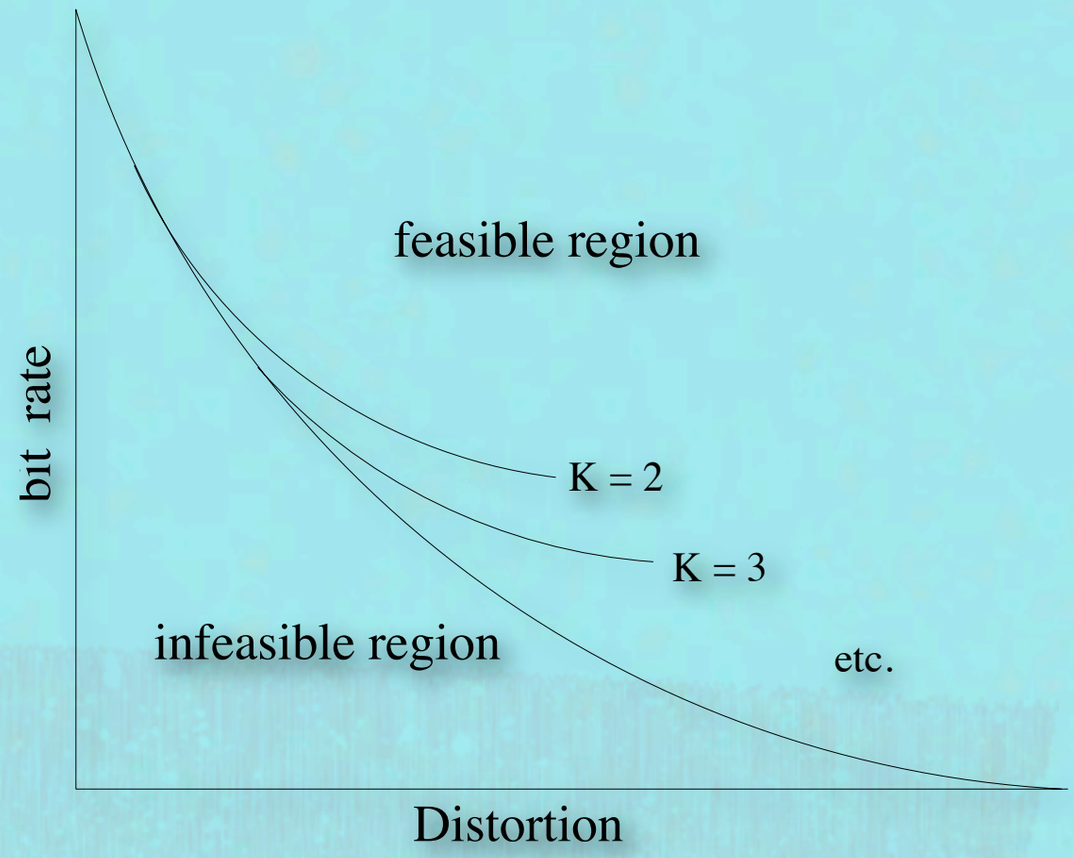
$$R(D) := \min_{p(s|x)} I[s,x]$$
$$\text{s.t.} \langle d(s,x) \rangle_{p(s,x)} = D$$

$$\Leftrightarrow \quad \min_{p(s|x)} \left( I[s,x] - \alpha \langle d(s,x) \rangle_{p(s,x)} \right)$$

$$\Rightarrow \quad p_{\text{opt}}(s|x) = \frac{p_{\text{opt}}(s)}{Z(x)} e^{-\alpha d(s,x)}$$

$$p_{\text{opt}}(s) = \langle p_{\text{opt}}(s|x) \rangle_{p(x)}$$
$$Z(x) = \langle e^{-\alpha d(s,x)} \rangle_{p_{\text{opt}}(s)}$$

- Optimal models are parameterized by $\alpha$ (controls trade-off)
- They lie on the rate-distortion curve, which delineates feasible from infeasible region

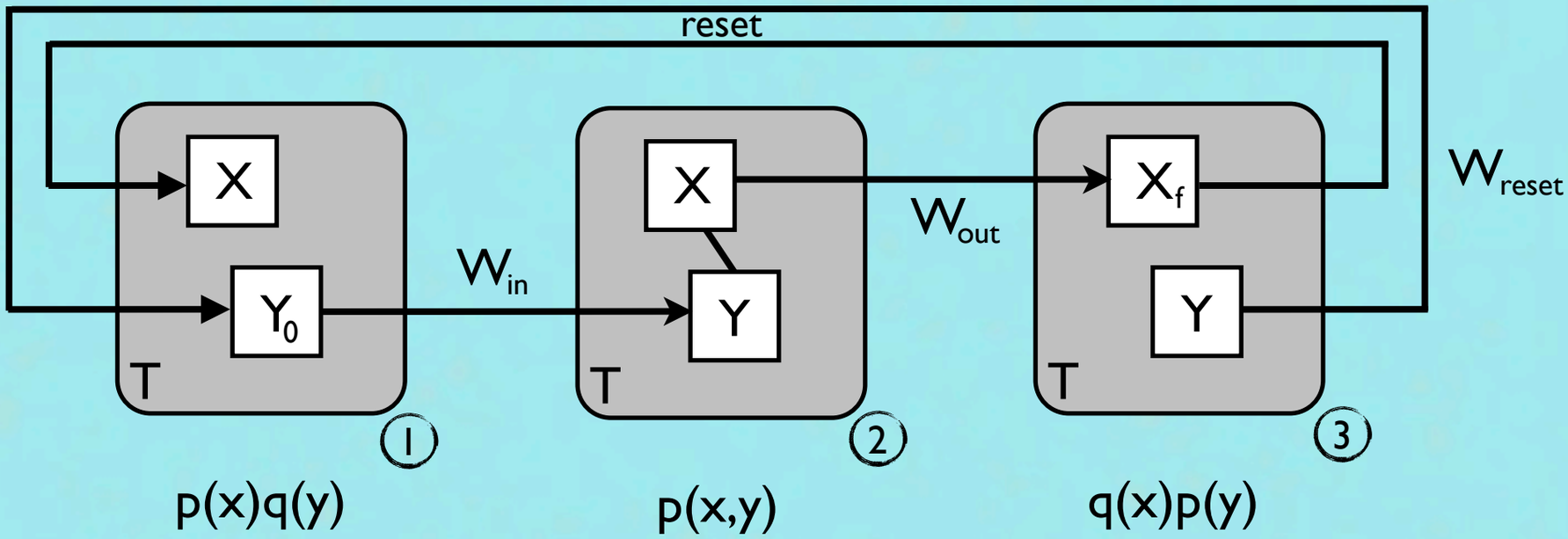


# IB is a special case of RDT

- Recall: given relevant variable, y, and p(x,y).
- Get IB by choosing the distortion function:

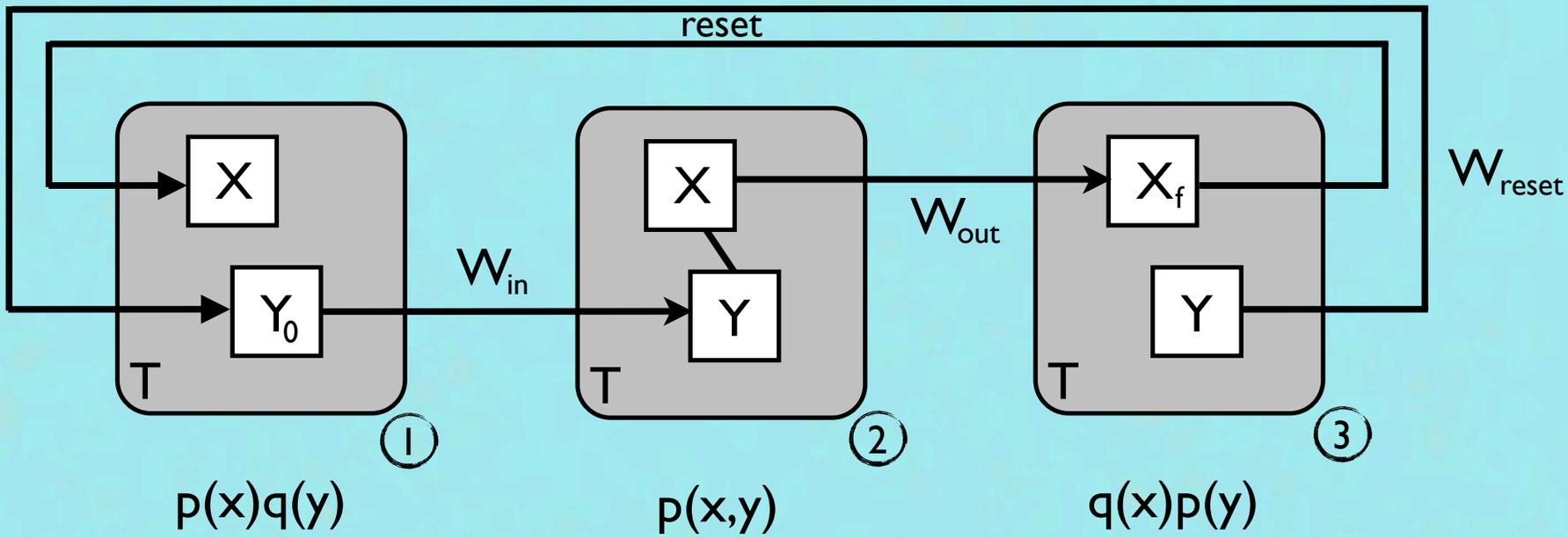
$$d(s,x) = D[p(y|x)\|p(y|s)]$$

# Thermodynamic foundations




- Thermodynamic cost: work it takes to run the memory Y
- Thermodynamic gain: extracting work from X

# Thermodynamic foundations




- Thermodynamic cost of data representation: Second Law =>
  Minimum effort required is the (nonequilibrium) free energy change:

$$W_Y \leq \Delta F_{1\to 2} + \Delta F_{\text{reset}} = \langle F[p(y|x)]\rangle_{p(x)} - F[p(y)]$$

- Maximum work extraction potential is free energy change due to inference: $-W_X \geq -(\Delta F_{2\to 3} + \Delta F_{\text{reset}}) = \langle F[p(x|y)]\rangle_{p(y)} - F[p(x)]$

# Thermodynamic foundations

- Free energy change = ave. energy change + kT entropy change
- Assume no change in overall average energy, then:
- minimum effort required for data representation *equals* maximum work extraction potential…

$$\langle F[p(y|x)]\rangle_{p(x)} - F[p(y)] = \langle F[p(x|y)]\rangle_{p(y)} - F[p(x)] =: \Delta F$$

- … and is proportional to mutual information: $\Delta F = k_B T \; I[x, y]$

  (Still 2014; Parrondo, Horowitz, Sagawa 2015)

- Lends a thermodynamic foundation to Shannon's approach!

# Thermodynamic foundations of RDT

(Still 2014)

- Given information source statistics p(x)
- Least effort principle: for a given quality of reproduction, minimize effort!

$$L(D) := \min_{p(s|x)} \Delta F$$
$$\text{s.t.} \langle d(s,x) \rangle_{p(s,x)} = D$$

- Recall: $\Delta F = k_B T \; I[s,x]$

  and: $R(D) := \min_{p(s|x)} I[s,x];\ \text{s.t.} \langle d(s,x) \rangle_{p(s,x)} = D$
- Encodings that achieve minimum effort lie on rate-distortion curve (kT adjusts the units).
- **Bit-cost is proportional to physical effort!**

# Thermodynamics of channel capacity

(Still 2014)

- Communication: given encoding and channel noise statistics, p(y | x). Information source can be more or less "matched" to channel.

- **Maximum extractable work <=> channel capacity** defined by Shannon as the maximally transmittable (rate of) information:

$$\max_{p(x)} \Delta F \quad \Leftrightarrow \quad \max_{p(x)} \; I[x, y]$$

(units adjusted by constants).

# Thermodynamic foundations of IB

(Still 2014)

- Given two correlated systems, X and Y, with p(x,y)
- Represent X by system S (at temperature T). Least effort: $k_B T\ I[s, x]$
- Can use this representation to extract work from system Y (at T')

  Work extraction potential: $k_B T'\ I[s, y]$
- Want: least effort data representation at maximum work potential. Opt. sol. fulfills: $\min_{p(s|x)} (k_B T\ I[s, x] - k_B T'\ I[s, y])$
- Recognize IB! Interpret trade-off parameter as ratio of temperatures.

# IV

# Dynamical and interactive learning

Animals and embodied systems (robots) interact with their environment and are able to change (to some degree) the world they are learning about.
Learning is dynamical, decisions are made under uncertainty.

# Interactive learning
(Still 2009)

- Learner summarizes history, and acts on environment.
- Summary: s, action: a, history: h (has to contain at least one past environmental data point; may contain past trajectories and/or past actions/states. Future signal to predict: z.
- *Behavior:* action-state pair (a,s)
- Based on the same motivation as before: **achieve behavior with maximal predictive power at fixed bit-cost**.

# Interactive learning (IAL)

- *Model:* probabilistic map from (past) experiences, h, onto summary s, p(s | h).

- *Action policy:* probabilistic map from (past) experiences onto actions a, $\pi$(a | h)

$$\max_{\substack{p(s|h)\\ \pi(a|h)}} \Big( I[\{s,a\};z] - \lambda I[s;h] - \mu I[a;h] \Big)$$

- Predictive power of behavior
- Cost of model
- Cost of policy

# Optimal action policies

- Optimal action policies have to balance between control and exploration!

$$\pi^*(a|h) \sim e^{-\frac{1}{\mu}E_{\pi^*}(a,h)}$$

$$E_{\pi^*}(a,h) = D\left[p(z|h,a)\|p_{\pi^*}(z|s^*,a)\right] - D\left[p(z|h,a)\|p_{\pi^*}(z)\right]$$

modeling accuracy (should be small)

information gain (should be large)

- Include rewards => curiosity driven reinforcement learning...

**Include rewards, r, and compare to "Boltzmann-Exploration"** (Sutton & Barto 98).

- Reinforcement learning: World state, x. Value of action: expected future reward.

  Objective: max expected return

$$Q_\pi = E_\pi \left[ \sum_{i=0}^{\tau} \gamma^i r_{t+i} \middle| x, a \right]$$

(A) **Max. entropy** => Boltzmann-Exploration:
(comp. Jaynes 57)

$$\pi^*(a|x) = \frac{1}{Z} e^{\frac{1}{\lambda} Q(x,a)}$$

**Min. information rate** => the policy that optimally trades return for bit-cost is

$$\pi^*(a|x) = \frac{p_{\pi^*}(a)}{Z} e^{\frac{1}{\lambda} Q(x,a)}$$

Exploration happens only due to randomness in decision!

(B) **Require optimal prediction** (as before) => optimal policy contains exploration-exploitation trade-off, even in absence of randomness in the decision!

$$a^*(x) = \arg\max_a \Big( D_{KL} \left[ p(x'|x,a) \| p_{\pi^*}(x') \right] + \alpha Q(x,a) \Big)$$

Exploration Exploitation

(Still&Precup 2012)

# Optimal models

- Models that achieve maximal predictive power at given coding cost fulfill

$$p^*(s|h) \sim e^{-\frac{1}{\lambda}\langle D_{KL}[p(z|h,a)\|p_{\pi^*}(z|s,a)]\rangle_{\pi^*(a|h)}}$$

- Compute solution with iterative algorithm (as before).

- This suggests definition of **interactive causal state partition**: $\mathcal{S}_A$ (take limit $\{\lambda, \mu\} \to 0$)

  - Two histories are equivalent under action policy A(h) when

$$p_A(z|h) = p_{A'}(z|h)$$

  - Two *action policies* A and A' are *causally equivalent* when they induce the same partition, i.e. $\mathcal{S}_A = \mathcal{S}_{A'}$

# Special case: no actions

1. *Recursive Information Bottleneck:*

$$\max_{p(s_{t+1}|s_t,x_t)} \left( I[\{s_{t+1}; \overrightarrow{x}_t^{(\tau)}] - \lambda I[s_{t+1}; \{s_t, x_t\}] \right)$$

3. Limit $\lambda \to 0; \{t, \tau\} \to \infty$ => $\epsilon$*-machine.*

4. History includes pasts of arbitrary length but no states => *Information Bottleneck / OCI.*

5. Time local + gaussian statistics => *PFIB*

✳ (more details next slide...)

# - Some theorems and algorithms for optimal predictive machines -

- **Dynamical learning system** finds asymptotically (Still, 2014) the $\epsilon$-**machine** (Crutchfield and Young,1989), a deterministic hidden Markov model that is maximally predictive. All characteristics of the underlying process can be computed from the $\epsilon$-machine, in many cases analytically: entropy rate, predictive/stored information,... (Crutchfield and colleagues,1989 onward).

- Used in **batch mode** (Still and Crutchfield , 2007) it is an instantiation of **Information Bottleneck Method** (Tishby, Pereira and Bialek,1999) which, in turn, is a special case of **rate-distortion theory** (Shannon, 1948). Family of optimal solutions: maximally predictive models at fixed memory. More efficient models are infeasible. Iterative algorithm (compare to Blahut/Arimoto ,1972).

- As the constraint on model complexity is relaxed, the batch method finds (Still, Crutchfield and Ellison, 2010) the "**causal state partition**" (Crutchfield and Young,1989), **minimal and unique sufficient statistics** (Shalizi and Crutchfield, 2001).

- **Gaussian assumptions** -> method is related (Creutzig and Sprekeler, 2008) to **slow feature analysis** (Wiskott and Sejnowski, 2002) in batch, and (Creutzig and Globerson and Tishby, 2009) to **canonical correlation analysis** in dynamic learning mode.

# Collaborators

(noneq. thermodynamics)

Gavin E. Crooks
(LBL)

David A. Sivak
(Simon Fraser)

Anthony J. Bell

(RCTW, Berkeley)

(learning theory)

Bill Bialek
(Princeton)

Doina Precup (McGill)



Jim Crutchfield
(UC Davis)

✳ (quantum system)

Arne Grimsmo
(Sherbroke)

# Our papers relevant to this talk


S.Still. Lossy is lazy. Proc. 7th Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-2014).

S.Still. Information Bottleneck approach to predictive inference. Entropy 2013.

S. Still, D. A. Sivak, A. J. Bell and G. E. Crooks. **The thermodynamics of prediction. Physical Review Letters 109, 120604 (2012)**

S. Still and D. Precup. An information-theoretic approach to curiosity-driven reinforcement learning. Theory in Biosciences 131 (3) pp. 139-148 (2012)

S. Still, J. P. Crutchfield, and C. Ellison. Optimal causal inference: estimating stored information and approximating causal architecture. Chaos 20, 037111 (2010)

S. Still. **Information-theoretic approach to interactive learning. EPL 85, 28005 (2009)**

S. Still and J. P. Crutchfield. Structure or Noise? arxiv:0708.0654 (2007)

S. Still and W. Bialek. How many clusters? An information theoretic perspective. Neural Computation16, pp. 2483-2506 (2004)


www2.hawaii.edu/~sstill