Information theoretic approach to interactive learning

Susanne Still

University of Hawaii at Manoa, ICS Department, Honolulu, HI 96822, USA. sstill@hawaii.edu

Abstract. - The principles of statistical mechanics and information theory play an important role in learning and have inspired both theory and the design of numerous machine learning algorithms. The new aspect in this paper is a focus on integrating feedback from the learner. A quantitative approach to interactive learning and adaptive behavior is proposed, integrating model- and decision-making into one theoretical framework. This paper follows simple principles by requiring that the observer's world model and action policy should result in maximal predictive power at minimal complexity. Classes of optimal action policies and of optimal models are derived from an objective function that reflects this trade-off between prediction and complexity. The resulting optimal models then summarize, at different levels of abstraction, the process's causal organization in the presence of the learner's actions. A fundamental consequence of the proposed principle is that the learner's optimal action policies balance exploration and control as an *emerging* property. Interestingly, the explorative component is present in the absence of policy randomness, i.e. in the optimal *deterministic* behavior. This is a direct result of requiring maximal predictive power in the presence of feedback.

Introduction. – The problem of learning a model, or model parameters, from observations obtained in experiments, appears throughout physics and the natural sciences as a whole. The statistical mechanics of learning have been discussed in many contexts [1, 2], such as neural networks, support vector machines [3, 4], and unsupervised learning via compression [5]. The latter, information theoretic approach essentially views learning as lossy compression – data are summarized with respect to some relevant quantity [6]. This can be an average variance [5], or any other measure of either distortion [7] or relevance [6]. Applied to time series data, one can show [8] that if prediction is relevant, then representations are found by this approach that constitute unique sufficient statistics [9] and which can be interpreted as underlying *causal states* [10] of the observed system.

However, the role of the observer is not always a passive one, as is assumed in the large majority of work on learning theory (see e.g. [11, 12]). In many problems ranging from quantum mechanics, to neuroscience, to animal behavior, the interactive coupling between the observer and the system that is being observed is crucial and has to be taken into account.

In this paper, an *information-theoretic approach to integrated model and decision making* is proposed. As a first step towards a general theory of adaptive behavior, let us ask a simple question: If the goal of a learner is to have as much predictive power as possible, then what is the least complex action policy, and what is the least complex world model that achieve this goal?

The ability to predict improves the performance of a learner across a large variety of specific behaviors, and is hence quite fundamental, increasing the survival chance of an autonomous agent, or an animal, and the success rate on tasks, independent of the specific nature of the task. Furthermore, a good model of the world must generalize well (see, e.g., [12])—in other words, the quality of the learner's world model can be judged by how well it predicts as-yet unseen data. For those reasons prediction is in general crucial for any adaptively behaving entity. Therefore, as a first step, we focus on prediction. To model animal behavior, other constraints, such as energy consumption, are clearly also relevant.

The approach taken here is related to, but different from *active learning* (e.g., [13–16]) and *optimal experiment design*, which has found countless applications in physics, chemistry, biology and medicine ([17]; for more recent reviews see, e.g., [18, 19]). These approaches do not usually take feedback from the learner into account. Feedback is modeled more explicitly in *reinforcement learning* (RL) [20], but this approach is limited to specific inputs, assuming that the learner receives a reward signal. In contrast to RL, we step back and ask about behavior that is optimal with respect to learning about the environment rather than with respect to fulfilling a specific task. Our approach does not require rewards.

Much of the RL literature assumes that the learner's explorative behavior is achieved by some level of randomness of the behavioral policy [21]. Here we show, in contrast, that if learning and optimal model-making are the goal, then explorative behavior emerges as one component of the optimal policy – even in the absence of stochasticity: any policy which is optimal with respect to learning maximally predictive models must balance exploration with control, including the optimal *deterministic* policy (see Eq. (21)).

Conceptually, our approach could perhaps be thought of as "rewarding" information gain and, hence, curiosity. In that sense, it is related to *curiosity driven RL* [22], where internal rewards are given that correlate with some measure of prediction error.[1] However, an important difference of the approach discussed here is that the learner's goal is not to predict future rewards, but rather to behave such that the time series it observes as a consequence of its own actions is rich in causal structure. This, in turn, then allows the learner to construct a maximally predictive model of its environment.

Optimally predictive model and decision making. – Let there be a physical system to be learned, and call it the learner's "world". A learner in parallel (i) builds a model of the world and (ii) engages in an interaction with the world. The learner's inputs are observations, $x(t)$, of (some aspects of) the world. Observations result in actions, $a(t)$, through a decision process. Actions affect the world and so change future observations.

Let us assume that the learner interacts with the environment between consecutive observations.[2] Let one decision epoch consist in mapping the current "history", h (specified below), available to the learner at time t, onto an action (sequence) a that starts at time t and takes the time Δ to be executed. The next datum is sensed at time $t + \Delta$. (We assume for simplicity that the times it takes to react and to sense are both negligible.)

The decision function, or *action policy* [20], is given by the conditional probability distribution $P(a|h)$.[3] Let the model summarize historical information, using internal states s, via the probabilistic map $P(s|h)$. The model and the policy depend upon each other, but histories are mapped *independently* onto (i) internal states (using the model $P(s|h)$), and (ii) action sequences (using the policy $P(a|h)$). Hence, actions and internal states are conditionally independent, if the history h is given:

$$P(s, a|h) = P(s|h)P(a|h). \tag{1}$$

The "internal state" does not change the statistics of the environment, but rather serves as an internal observer. The feedback due to the actions, however, changes the statistics of the environment. The action policy contains a model in the sense that if a large group of histories share the same optimal action, then the action can be viewed as a compressed representation of this "history-cluster".

The learner uses the current state, $s(t)$, together with knowledge of the action, $a(t)$, to make probabilistic predictions of future observations, $z(t)$, of length τ_f:[4]

$$P(z|s, a) = \frac{1}{P(s, a)} \langle P(z|h, a)P(a|h)P(s|h)\rangle_{P(h)} . \tag{2}$$

$P(z|h, a)$ and $P(h)$ are (for the moment) assumed to be known. A history always includes the current observation, $x(t)$. Beyond this, it may include a record of prior observations reaching some length τ_p into the past, and also previous internal state and action(s). Lengths of the internal records of past observations and past actions are assumed given by the learner's storage capacity.

The problem of interactive learning then is to choose a model and an action policy, which are optimal in that they maximize the learner's ability to predict the world, while being minimally complex.

We measure the learner's predictive ability by the mutual information [7] that the internal state, *in the presence of the action*, contains about the future:

$$I[\{s, a\}; z] = \left\langle \log\left[\frac{P(z|s, a)}{P(z)}\right]\right\rangle_{P(z,s,a)} . \tag{3}$$

The quantity $I[\{s, a\}; z] = H[z] - H[z|s, a]$ measures the reduction in the uncertainty about the future (entropy H), when state and action are known. It is zero if the future is independent of s and a. It is maximal if the knowledge of s and a eliminates all uncertainty about the future ($H[z|s, a] = 0$).

Simple models and simple action policies come at a lower coding cost, quantified by the coding rates $I[s; h]$ and $I[a; h]$, respectively. The notion that the simplest possible model is preferable is deeply rooted in our culture. William of Ockham is frequently cited on this matter, which is known as "Ockham's razor". In the same vein, out of two action policies which yield the same value of the objective, Eq. (3), one would choose the simpler policy, as there is no reason to implement a more complex policy which takes more memory.

The interactive learning problem is solved by maximizing $I[\{s, a\}; z]$ over $P(s|h)$ and $P(a|h)$, under constraints

[1] Our approach is fairly general, and to compare one has to adopt the specific RL setting, which we explore in [23].

[2] This sequential setup is useful for the sake of simplicity. However, a real agent continuously acts and senses, and an extension to this more involved case would be interesting.

[3] Short hand notation: the argument t is dropped. Actions a, internal states s, futures z, and histories h are (possibly multi-valued) random variables with values $A \in \mathcal{A}$, $S \in \mathcal{S}$, $Z \in \mathcal{Z}$, and $H \in \mathcal{H}$, respectively.

[4] Future observations, $z(t)$, are given by the signal $x(t')$ on the interval $t' \in [t + \Delta, t + \Delta + \tau_f]$, where Δ is the duration of the intervention given by the action, or the sequence of actions, initiated at time t, $a(t)$. The learner is interested in understanding how one intervention changes the future. The action choice does depend on past actions, if they are included in the learner's history, $h(t)$. However, planning of consecutive future actions is not discussed here, but an extension would be desirable. The notation $\langle\cdot\rangle_P$ denotes the average taken over P.

that select for the simplest possible model and the most efficient policy, respectively, in terms of smallest complexity measured by the coding rate. Less complex models and policies result in less predictive power. This trade-off can be implemented using Lagrange multipliers, λ and μ. Following the spirit of *rate distortion theory* [7], and, more closely related, the *information bottleneck method* (IB) [6], one can then calculate the best possible solution at each value of the Lagrange multipliers. The optimization problem for interactive learning is given by:

$$\max_{\substack{P(s|h)\\P(a|h)}} \big(I[\{s,a\};z] - \lambda I[s;h] - \mu I[a;h]\big) \tag{4}$$

The two constraints are taken into account individually, rather than as a sum,[5] so that their relative importance can be adjusted. Think, for example, about a robotic multi-agent system in which robots communicate their internal states to each other. Limited communication channel capacity may force them to produce compact internal representations, but the complexity of the action policy that each individual can implement does not have to be equally constrained.

The trade-off parameters λ and μ parameterize families of optimal models and policies, respectively, constituting those models and policies that have maximal predictive power at fixed complexity. An analogy to statistical mechanics is useful to guide intuition [5], and relates λ and μ to temperature – they control the "fuzziness" of the maps that assign histories to states and actions, respectively. This approach also relates the distortion function to the energy function of a corresponding physical system and the normalization constant to the partition function.

Optimal action policies. The action policies that solve optimization problem, Eq. (4), are given by

$$P_{\rm opt}(a|h) = \frac{P(a)}{Z_{\rm A}(h,\mu)} e^{-\frac{1}{\mu}E_{\rm A}(a,h)} \tag{5}$$

with the energy function

$$\begin{aligned} E_{\rm A}(a,h) &= \langle \mathcal{D}[P(z|h,a)\|P(z|s,a)]\rangle_{P(s|h)} \\ &\quad -\mathcal{D}[P(z|h,a)\|P(z)]\,, \end{aligned} \tag{6}$$

and the partition function

$$Z_{\rm A}(h,\mu) = \langle e^{-\frac{1}{\mu}E_{\rm A}(a,h)}\rangle_{P(a)}. \tag{7}$$

$\mathcal{D}[p\|q] = \langle \log[p/q]\rangle_p$ denotes the *relative entropy*, or *Kullback–Leibler divergence* between distributions p and q. Equations (5)-(7) must be solved self-consistently, together with Eq. (2) and

$$P(a) = \langle P(a|h)\rangle_{P(h)}\,, \tag{8}$$

$$P(z) = \Big\langle \langle P(z|h,a)\rangle_{P(a|h)}\Big\rangle_{P(h)}. \tag{9}$$

To derive this result (Eqs. (5)-(7)), one calculates $I[\{s,a\};z]$, using Eq. (2), and the functional derivative of Eq. (4) w.r.t. $P(a|h)$. Individual nonzero contributions are given by:[6]

$$\begin{aligned} \frac{\delta I[\{s,a\};z]}{\delta P(a|h)} &= P(h)\left\langle\left\langle \log\left[\frac{P(z|s,a)}{P(z)}\right]\right\rangle_{P(z|h,a)}\right\rangle_{P(s|h)} \\ &= P(h)\mathcal{D}[P(z|h,a)\|P(z)] \\ &\quad -P(h)\,\langle\mathcal{D}[P(z|h,a)\|P(z|s,a)]\rangle_{P(s|h)} \end{aligned} \tag{10}$$

$$\frac{\delta I[a;h]}{\delta P(a|h)} = P(h)\log\left[\frac{P(a|h)}{P(a)}\right]. \tag{11}$$

Observe that the most likely action is that of minimum energy (see Eq. (5)). The first term in the energy function, Eq. (6),

$$\langle \mathcal{D}[P(z|h,a)\|P(z|s,a)]\rangle_{P(s|h)} \tag{12}$$

is smaller for actions that will, on average, make the conditional future distribution $P(z|h,a)$ as *similar* as possible to the distribution that is *predicted* by the learner's internal state, $P(z|s,a)$. The average is taken over the model $P(s|h)$. This term selects for actions that bias the future towards what the learner predicts – it is therefore related to the *control* that the learner can exert on the world.

The second (negative) term in Eq. (6)

$$-\mathcal{D}[P(z|h,a)\|P(z)] \tag{13}$$

selects for actions that will make the conditional future distribution $P(z|h,a)$ as *different* as possible from the average $P(z)$. The term embodies a preference for actions that bias towards an uncommon future distribution – it is related to *exploration* and causes the learner to perturb the world away from the average.

This shows that at the root of interactive learning there is a competition between exploration and control, which arises as a fundamental consequence of the proposed optimization principle: Exploration and control have to be *balanced* in the optimal action policy to result in maximal predictive power.

Optimally predictive models. The family of models that solve optimization problem Eq. (4), is given by [7]

$$P_{\rm opt}(s|h) = \frac{P(s)}{Z_{\rm S}(h,\lambda)} e^{-\frac{1}{\lambda}E_{\rm S}(s,h)} \tag{14}$$

[5] $I[\{s,a\};h] + I[s;a] = I[a;h] + I[s;h]$, because of Eq. 1. $I[\{s,a\};h]$ is the coding rate of the learner's full behavior – consisting of both the internal state, s, and the action sequence, a. $I[s;a]$ measures the redundancy, which should be minimized together with the coding rate.

[6] Terms constant in a are omitted, because in the solution they are absorbed into $Z_{\rm A}$.

[7] The derivation is similar to that for Eq. (5) and follows [6]. Individual contributions to the functional derivative w.r.t. $P(s|h)$ are (ignoring constant terms): $\frac{\delta I[\{s,a\};z]}{\delta P(s|h)} = -P(h)\,\langle\mathcal{D}[P(z|h,a)\|P(z|s,a)]\rangle_{P(a|h)}$ and $\frac{\delta I[s;h]}{\delta P(s|h)} = P(h)\log\left[\frac{P(s|h)}{P(s)}\right]$.

with

$$E_{\mathrm{S}}(s,h) = \langle \mathcal{D}[P(z|h,a)\|P(z|s,a)]\rangle_{P(a|h)} \qquad (15)$$

and

$$Z_{\mathrm{S}}(h,\lambda) = \left\langle e^{-\frac{1}{\lambda}E_{\mathrm{S}}(s,h)}\right\rangle_{P(s)} . \qquad (16)$$

These equations must be solved self-consistently, together with Eq. (2) and

$$P(s) = \langle P(s|h)\rangle_{P(h)} . \qquad (17)$$

The most likely state minimizes the relative entropy between the actual, $P(z|h,a)$, and the predicted, $P(z|s,a)$, conditional future distribution (see Eqs. (14) and (15)), averaged over the action policy $P(a|h)$. The internal states thus capture the effect that the history has on the probability distribution over futures, under a given action policy. In that sense, the optimal model reflects the causal structure of the underlying process.

Altogether, Eqs. (5) and (14), must be solved self consistently (together with Eqs. (2), (6)-(9), and (15)-(17)) to yield the model that is optimally predictive under the optimal policy (and vice versa). This can be done iteratively, resulting in an algorithm that is similar to the IB algorithm [6]. This new algorithm, however, includes a feedback loop, due to actions.[8]

With increasing λ, the level of abstraction of the model increases, as less detail is kept. In the high temperature limit, $\lambda \to \infty$, all possible histories are effectively represented by the same internal state.[9]

Deterministic models and decisions. In the low temperature limit ($T \to 0$; $T \in \{\lambda, \mu\}$), the distributions in Eqs. (5) and (14) become deterministic mappings. To see this, let us use the discrete random variable $y \in \{a, s\}$, and let $E(y,h)$ denote the value of the energy function E_{A}, if $y = a$, and E_{S}, if $y = s$. Furthermore, define the functions $y^*(h) := \arg\min_y E(y,h)$ and $\mathcal{E}(y,h) := E(y,h) - E(y^*(h),h) \geq 0$. Now, we can write the conditional distribution for the most likely value $y^*(h)$ as

$$\begin{aligned} P(y = y^*(h)|h) &= \frac{P(y = y^*(h))}{Z(h,T)} e^{-\frac{1}{T}E(y^*(h),h)} \\ &= \left(1 + \sum_{y \neq y^*(h)} \frac{P(y)}{P(y^*(h))} e^{-\frac{1}{T}\mathcal{E}(y,h)}\right)^{-1} \end{aligned} \qquad (18)$$

Since $\mathcal{E}(y,h)$ is positive, the sum goes to zero as $T \to 0$ (assuming that $P(y^*(h)) > 0$). As a consequence, we have $P(y = y^*(h)|h) = 1$ and, due to normalization, the optimal mapping becomes deterministic: $P_{T\to 0}(y|h) = \delta_{yy^*(h)}$, where δ denotes the Kronecker-Delta.

For a *deterministic model*, specified by $P_{\lambda\to 0}(s|h) = \delta_{ss^*(h)}$, this means that a history h is assigned with probability one to the state $s = s^*(h)$ which minimizes the energy function $E_{\mathrm{S}}(s,h)$, Eq. (15):

$$s^*(h) = \arg\min_s \langle \mathcal{D}[P(z|h,a)\|P(z|s,a)]\rangle_{P(a|h)} . \qquad (19)$$

Note that without constraints on the cardinality of the state space, one can always ensure that this minimum is zero: $E_{\mathrm{S}}(s^*(h),h) = 0$. This fact then implies that the predicted information, $I[\{s,a\};z]$, reaches its maximum at the optimal deterministic model, $I[\{s^*,a\};z]$.

The maximum is given by the predictive information of the time series, in the presence of the learner's actions: $I[\{s^*,a\};z] = I[\{h,a\};z]$.[10] The optimal policy now maximizes this quantity, at fixed $I[a,h]$. This illustrates that the optimal policy makes as much information as possible available to be summarized by the model, at fixed policy complexity.

Action policies become increasingly random with increasing μ – the learner's reactions become less specific responses to the history. In the other limit, as the complexity constraint is relaxed by letting the parameter μ approach zero, one finds the optimal *deterministic* policy $a^*(h)$[11] which maximizes the predictive information of the time series, in the presence of the actions.

The special case is of particular interest in which the learner produces *deterministic* maps $s^*(h)$ and $a^*(h)$, which maximize the predictive power, Eq 3. The optimal deterministic model maps a history h to the internal state

$$s^*(h) := \arg\min_s \mathcal{D}[P(z|h,a^*(h))\|P(z|s,a^*(h))] . \qquad (20)$$

Assuming that there are no constraints on the cardinality of the state space, this map partitions the space of histories in a way that is similar to the *causal state partition* of [10]. One can show [8] that if actions are not considered (passive time series modeling), then the passive equivalent of Eq. (20) exactly recovers the causal state partition of [10]. Causal states are unique and minimal sufficient statistics – constituting a meaningful representation of the underlying process [9].

The partition specified by Eq. (20) allows for an extension of the causal state concept to *interactive* time series modeling: here the space of histories is partitioned such that all histories, $h \in \mathcal{H}_s \subset \mathcal{H}$, that are mapped to the same causal state, s, are *causally equivalent under the optimal action policy*, $a^*(h)$; meaning that their conditional future distributions $P(z|h,a^*(h))$ are the same.

[8] Details about the algorithm are given in [24], where examples are also discussed. An extension will be published elsewhere.

[9] As $\lambda \to \infty$, $P(z|s,a)$ is the same for all states s: As $\lambda \to \infty$, $P_{\mathrm{opt}}(s|h) \to P(s)$, see Eq. (14), and with that $P(s,a) = \langle P(s,a|h)\rangle_{P(h)} = \langle P(s|h)P(a|h)\rangle_{P(h)} \to P(s)P(a)$, and $P(z|s,a) \to \frac{1}{P(s)P(a)}\langle P(z|h,a)P(a|h)P(s)\rangle_{P(h)} = P(z|a)$, $\forall s$; see Eqs. (1) and (2).

[10] $I[\{s,a\};z] = I[\{h,a\};z] - \langle E_{\mathrm{S}}(s,h)\rangle_{p(s,h)}$. The second term vanishes for the optimal deterministic model. It becomes $\langle E_{\mathrm{S}}(s^*(h),h)\rangle_{p(h)} = 0$.

[11] $a^*(h)$ is given by Eq. 21.

This grouping of histories results in an equivalence class that is controlled by the action policy: under any action policy, $A(h)$ (where the map $A : h \mapsto a$ is a deterministic policy), two histories h and h' are equivalent with respect to their effect on the future, z, if $P(z|h, A(h)) = P(z|h', A(h'))$. The resulting partition, $\mathcal{S}_A$, of the history space into causal states depends on the action policy, A. The choice of the policy determines the nature of the time series which is produced by the system *coupled* to the observer through the actions. Note that there could be different action policies $A' \neq A$, which result in coupled systems with the same underlying causal state partition $\mathcal{S}_A = \mathcal{S}_{A'}$. The policy $A = a^*$ is the deterministic policy that creates the coupled world-observer system that can be predicted most effectively by a causal model.

Optimal deterministic decisions for actions are made according to the rule

$$\begin{aligned} a^*(h) := \arg\min_a \big[\quad & \langle \mathcal{D}[P(z|h,a) \| P(z|s,a)] \rangle_{P(s|h)} \\ & - \mathcal{D}[P(z|h,a) \| P(z)] \big]. \end{aligned} \tag{21}$$

It is important to note that the term related to exploration (second term) persists in the optimal deterministic action policy, Eq. (21). This is in direct contrast to "Boltzmann exploration", commonly used in RL [20]. There, exploration is implemented as policy randomization by softening of the optimal, deterministic policy (optimal in an RL sense by maximizing expected future reward). We have shown here, however, that to create data which allows for optimally predictive modeling, an exploratory component must be present even in the optimal *deterministic* policy. In our framework, exploration is hence an emerging behavior, and it is *not* the same as policy randomization.

Probability estimates and finite sampling errors. So far, we have assumed $P(z|h,a)$ and $P(h)$ to be known. However, in practice, they may have to be estimated from the observed time series. Hence there could be a bias towards overestimating $I[\{s,a\};z]$ due to finite sampling errors in the probability estimates. This may result in over-fitting. The accuracy of the estimates depends on the data set size, N. One can counteract finite sampling errors, using an approximate error correction method, such as discussed in [25]. This method has already been applied successfully to predictive inference in the absence of actions [8] and it can also be applied in the presence of actions.

Time dependent on-line learning procedure. In [25], we calculated bounds on the smallest temperature, $T^*(N)$, allowable before over-fitting occurs. This value depends on the data set size N. In the interactive learning setup, the data set size grows linearly with time. One can implement an algorithmic annealing procedure, similar to the one in [5], but different in that the temperature is kept fixed at each time step and then changes over time with growing data set size. This captures the intuition that a learner may allow itself to model an increasing amount of detail the longer it has observed the world. The temperatures in each time step are set to (upper bounds on) the values $\lambda^*(t)$ and $\mu^*(t)$, below which over-fitting would occur. Since these can be calculated, an annealing *rate*, as used in deterministic annealing [5], is not necessary. The work in [25] directly provides a bound on $\lambda^*(t)$ and could be extended to calculate a bound on $\mu^*(t)$. Tighter bounds or an exact calculation of T^* would also be desirable.

Possible extension to multi-agent systems. When multiple agents observe and interact with an environment, they often exhibit emerging co-operative behavior. Understanding the emergence of such co-operative strategies is an active field of research. In order to utilize our approach for the study of this phenomenon, we have to distinguish (i) the agents' available sensory input and (ii) whether there is communication between agents. In the simplest case, each of the agents has access only to data from the environment. Then each agent can be modeled exactly as we have outlined here, and all coupling happens implicitly, through the environment. Communication of internal states and/or the observation (or communication) of each others actions, however, means that the *other agents'* internal states and/or actions, respectively, must be included in each agent's input (history h). Furthermore, if agents try to learn about each others behavior, then we need to include the other agents' future actions into the data which ought to be predicted (future z). A detailed exploration of multi-agent learning has to be left for future research.

Summary. – This paper has proposed an information-theoretic approach to a quantitative understanding of interactive learning and adaptive behavior by means of optimal predictive modeling and decision making. A simple optimization principle was stated: use the least complex model and action policy which together provide the learner with the largest predictive ability. A fundamental consequence of this principle is that the optimal action policy finds a balance between exploration and control. This is a direct consequence of optimal prediction in the presence of feedback due to the learner's actions. The theory developed here is general in that it makes no assumptions about the detailed structure of the underlying process that generates the data, and thus is not restricted to specific model classes.

Acknowledgments. – I am deeply grateful to W. Bialek who contributed significantly to the ideas expressed in this paper. I thank J. P. Crutchfield and D. Precup for many helpful discussions and comments on the manuscript, and M. Berciu, L. Bottou, I. Nemenman, B. A. Pearlmutter and C. Watkins for very useful discussions.

REFERENCES

[1] T. L. H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Rev. Mod. Phys.*, 65(2):499–

556, 1993.

[2] D. Mahlzahn and M. Opper. Statistical Mechanics of Learning: A Variational Approach for Real Data. *Phys. Rev. Lett.*, 89:108302, 2002.

[3] R. Dietrich, M. Opper, and H. Sompolinsky. Statistical Mechanics of Support Vector Networks. *Phys. Rev. Lett.*, 82(14):2975–2978, 1999.

[4] M. Opper and R. Urbanczik. Universal learning curves of support vector machines. *Phys. Rev. Lett.*, 86(19):4410–4413, 2001.

[5] K. Rose, E. Gurewitz, and G. C. Fox. Statistical Mechanics and Phase Transitions in Clustering. *Phys. Rev. Lett.*, 65(8):945–948, 1990.

[6] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In B. Hajek and R. S. Sreenivas, editors, *Proc. 37th Annual Allerton Conference*, pages 368–377. University of Illinois, 1999.

[7] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27, 1948.

[8] S. Still, J. P. Crutchfield, and C. Ellison. Optimal Predictive Inference. 2007. Available at: http://lanl.arxiv.org/abs/0708.1580.

[9] J. P. Crutchfield and C. R. Shalizi. Thermodynamic depth of causal states: Objective complexity via minimal representations. *Phys. Rev. E*, 59(1):275–283, 1999.

[10] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Lett.*, 63:105–108, 1989.

[11] A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning.* Cambridge University Press, 2001.

[12] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, New York, 1995.

[13] D. MacKay. Information based objective functions for active data selection. *Neural Comp.*, 4(4):589–603, 1992.

[14] H. S. Seung, M. Opper, and H. Sompolinsky. Querry by Commitee. In *Proceedings of the Fifth Workshop on Computational Learning Theory*, pages 287 – 294. New York, ACM, 1992.

[15] S. Dasgupta. Coarse sample complexity bounds for active learning. In B. Schölkopf Y. Weiss and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 235–242. MIT Press, Cambridge, MA, 2006.

[16] M.-F. Balcan, A. Beygelzimmer, and J. Langford. Agnostic Active Learning. In *Proceedings of ICML 2006*, 2006.

[17] V. V. Fedorov. *Theory of optimal experiments.* Academic Press, 1972.

[18] A. C. Atkinson, B. Bogacka, and A. A. Zhiglkilavskify, editors. *Optimum Design 2000.* Springer, 2001.

[19] G. Box, J. Hunter, and W. Hunter. *Statistics for Experimenters (2nd Ed.).* Wiley, 2005.

[20] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An introduction.* MIT Press, 1998.

[21] L. Pack-Kaelbling, M. Littman, and A. Moore. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[22] J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. . In J. A. Meyer and S. W. Wilson, editors, *Proc. Int. Conf. Simulation of Adaptive Behavior: From Animals to Animats*, pages 222–227. MIT Press/Bradford Books, 1991.

[23] S. Still and D. Precup. An information theoretic approach to curiosity driven reinforcement learning. 2008. In preparation.

[24] S. Still and W. Bialek. Technical Report UH-ICS-MLL-06-06, University of Hawaii, Honolulu, USA, 2006.

[25] S. Still and W. Bialek. How many clusters? An information theoretic perspective. *Neural Comp.*, 16(12):2483–2506, 2004.