Metabolic cost as an organizing principle for cooperative learning

David Balduzzi, Pedro A. Ortega, Michel Besserve
Max Planck Institute for Intelligent Systems, Tübingen, Germany.
{balduzzi, ortega, besserve}@tuebingen.mpg.de

Abstract

This paper investigates how a population of neuron-like agents can use metabolic cost to communicate the importance of their actions. Although decision-making by individual agents has been extensively studied, questions regarding how agents should behave to cooperate effectively remain largely unaddressed. Under assumptions that capture a few basic features of cortical neurons, we show that constraining reward maximization by metabolic cost aligns the information content of actions with their expected reward. Thus, metabolic cost provides a mechanism whereby agents encode expected reward into their outputs. Further, aside from reducing energy expenditures, imposing a tight metabolic constraint also increases the accuracy of empirical estimates of rewards, increasing the robustness of distributed learning. Finally, we present two implementations of metabolically constrained learning that confirm our theoretical finding. These results suggest that metabolic cost may be an organizing principle underlying the neural code, and may also provide a useful guide to the design and analysis of other cooperating populations.

1 Introduction

Rational decision making by individual autonomous agents is typically formalized as maximizing a reward function via reinforcement learning [1, 2]. Although questions remain concerning the choice of reward function, choice of optimization method, and how this approach relates (if at all) to human decision making, the reinforcement learning framework has been successfully applied to an enormous variety scientific and engineering problems.

In the closely related setting of a population of cooperating agents we immediately encounter a basic question. What can one agent infer from the actions of another? Since some actions are invariably more important than others, communicating the relative importance of actions is crucial to effective cooperation. This paper shows, under fairly general assumptions, that taking metabolic costs into account causes agents to signal which of their actions are important and which are not.

The cortex as a guided self-organizing system. The scenario we have in mind is the mammalian cortex, with agents corresponding to neurons. The cortex is a self-organized system of between about $10^7$ and $10^{11}$ neurons (depending on the species) that are guided by neuromodulators that signal pleasure, pain and other globally salient events. Neurons communicate with each other via spiketrains – sequences of silences and spikes they receive from and transmit to $10^3$ to $10^4$ other neurons through connections called synapses. Neurons learn by increasing and decreasing the efficacy of their synapses according to the timing of pre-synaptic (input) and post-synaptic (output) spikes, as well as the presence or absence of neuromodulators [3, 4].
The goal of the brain as a whole is to choose favorable actions for its organism in a great variety of situations. Ultimately, responsibility for choosing the right action falls onto the neurons in the central nervous system. Most neurons, however, interact with the environment extraordinarily indirectly – through millions or billions of other neurons – so the consequences for the organism of an individual neuron’s actions are difficult to pin down. As a result, it is unclear when neurons should spike or what their spikes mean.

Without further assumptions, there is no way – even assuming neurons act optimally according to a known reward function – for another neuron to know which of its spiketrains are important and which are not since, for example, neurons cannot access each other’s inputs. In other words, without further assumptions there is no way for neurons to cooperate.

**Overview of the paper.** The key to distinguishing important from unimportant actions is that spikes and silences are not abstract, interchangeable symbols. Spiking and responding to spikes carries a much higher metabolic cost than not spiking [5]. This cost is significant since the nervous system consumes a disproportionate share of an organism’s total energy budget [6].

In this paper we attempt to distill an essential feature of cortical computation – metabolic cost – and explore its implications for cooperative learning. Formally, we consider agents that maximize rewards under metabolic constraints according to a free-energy principle following [7–9]. Our first result, §3, is that the effective information generated by an agent’s actions, assuming it acts optimally, quantifies the expected subsequent reward. Thus, it is not necessary to know an agent’s reward function to compute the importance of its actions; all that is needed is an accurate model of its policy.

In fact, if an agent obeys a strong metabolic constraint, so that the relative frequency of its actions is tightly controlled, then both the effective information and expected reward of its actions couple tightly to metabolic cost. Thus, if all agents have the same metabolic cost profile and impose similarly strong constraints, it is possible to determine the relative importance of actions from their metabolic cost. This shows that metabolically constrained agents encode expected rewards into their outputs. Intriguingly, this result may generalize to neurons in cortex, suggesting a novel approach to the neural code [10].

Controlling the frequency of expensive actions (spikes) also improves an agent’s ability to estimate expected reward, §4. In the asymptotic limit, where an agent samples from its environment exhaustively, there is no problem. However, over short time scales (which are the norm in a rapidly reorganizing environment such as the brain), the expected reward does not necessarily coincide with the empirical reward estimate computed from a finite sample. It turns out that the higher the effective information generated by spikes, the more closely the empirical reward approximates the expected reward.

Finally, §5 considers some implications of these results for cooperative learning. If expensive actions encode expected rewards, then agents that wish to cooperate should take advantage of this fact by privileging expensive actions in their reward functions. We introduce two minimal models of metabolically constrained learning agents whose reward functions privilege spikes.

Metabolic cost thus provides an organizing principle that tightly binds information, reward and approximation errors together. Future work will investigate the interplay between metabolism and cooperation in more biologically realistic settings.

**Acknowledgements.** We thank Yevgeny Seldin and Giulio Tononi for useful discussions.

2 **A minimal model of neurons in cortex**

**Unique features of neurons in cortex.** The brain differs from many other populations of interacting agents in a number of key respects, two of which we mention here. First, by and large neurons do not die, nor are new neurons created; the population is essentially stable.
Second, by and large neurons do not compete for resources, and “successful” neurons do not receive a significantly larger share of metabolic resources than “unsuccessful” neurons.

We also wish to emphasize an important difference between our agents, which abstract certain features of cortical neurons, and “agents in the wild”. Individual neurons have a negligible impact on what happens next in the brain: a single neuron has little effect the neurons it targets – each receiving inputs from thousands of other neurons – and even less effect on the rest of the brain. The inability of neurons to effectively manipulate their environment tremendously simplifies any optimization problems they may choose to solve (by stripping out the recursive Bellman aspect). Thus, although the brain is in many ways extraordinarily complex, the problems faced by an individual neuron are simpler than those faced by autonomous agents, precisely because their ability to project power is extremely limited.

**Neurons as agents.** The brain is modeled as a population of $K$ homogeneous agents, where each agent $k$ corresponds to a neuron. Each agent follows its policy $\pi_k$, where $\pi_k(a|s)$ is the probability the agent picks action $a \in \mathcal{A}$ upon encountering situation $s \in \mathcal{S}$. A situation is a vector $s = (a_1, \ldots, a_K)$ whose entries are the actions of all the agents in the brain at the previous time step. Since each neuron is only exposed to a small fraction of the brain, it ignores most entries in the vector, and so effectively the policy of agent $k$ is $\pi_k(a|s) = \pi_k(a|\Sigma_k)$ for some subset $\Sigma_k \subset \{a_1, \ldots, a_K\}$. Let $P(s)$ denote the probability over the situations $\mathcal{S}$, which we assume to be i.i.d.

The population is exposed to a global neuromodulation signal $n \in \mathcal{N}$ that is used to communicate the performance of the population as a whole. The neuromodulation signal is drawn with probabilities $P(n|s)$, that is, depending on the current situation. Agents are rewarded individually for their behavior using a reward function

$$R : \mathcal{S} \times \mathcal{A} \times \mathcal{N} \to \mathbb{R}$$

$$R(s,a,n)$$

which specifies the reward an agent obtains when it issues action $a$ being in situation $s$ and subsequently receives the neuromodulation signal $n$. Since individual neurons have little influence over their environment, we assume that the future situations experienced by the agent are unaffected by its output:

$$p\left(s(t+1,\ldots)\vert s(t),a(t)\right) = p\left(s(t+1,\ldots)\vert \Sigma_k\right)$$

where $t$ refers to time. Clearly, this assumption fails at the population level – if populations of neurons had no effect on the brain state (and therefore its actions) there would be no point in them learning. Nevertheless since, for example, destroying a single neuron makes essentially no difference to a brain’s functioning, we argue that Eq. (2) is a reasonable assumption at the individual neuron level.

The goal of any agent is to issue actions that maximize the time-averaged reward

$$\frac{1}{T} \sum_{t=1}^{T} R_t.$$ Define shorthand

$$R(a,s) := \mathbb{E}[R(a, s, N) \vert p(N|a, s)],$$

for the expected reward conditional on a situation-action pair.

**Spikes are metabolically expensive.** Since spikes are more expensive than silences. We assume there is a function $c : \mathcal{A} \to [0, \infty)$ assigning a cost to each neuronal action, and that the reference distribution takes the form $P(a) = \frac{1}{\lambda c(a)}$ where $\lambda$ is a positive, monotonically decreasing function chosen such that $\sum_a P(a) = 1$.

Agents take metabolic cost into account via a cost function penalizing deviations from the pre-specified relative action frequencies $P(A)$. If we interpret information content as resource costs, then

$$C(a,s) = \pi(a|s) \log \frac{\pi(a|s)}{P(a)}$$

Footnote: Under these conditions, agents are contextual bandit players, i.e. every neuron plays against a slot machine having odds over the rewards that depend only on the current situation [11].
measures the average extra cost the neuron pays for producing action \(a\) with frequency \(\pi(a|s)\) instead of frequency \(P(a)\) in situation \(s\).

We will often consider the case where neurons have two outputs: silence and a single spike. In this case we fix metabolic prior

\[
P(a) := \begin{cases} 
1 - f & \text{if } a = 0 \\
f & \text{if } a = 1, \text{ for some } f \ll \frac{1}{2}.
\end{cases}
\]

We will also often use the notation \(a_0\) for silence and \(a_1\) for a spike.

The effective information generated by an action. To understand what agents can and do say to one another using their actions, it is useful to consider an agent as a communication channel mapping situations to actions following [12]. The information transferred by an agent on average is the mutual information \(I_\pi(S, A)\).

However, since we are interested in the information transferred by specific actions, we introduce effective information [13,14].

**Definition 1 (effective information).**

Given an agent with policy \(\pi(a|s)\) and prior \(P(S)\) on situations, the effective information generated by action \(a \in A\) is

\[
ei(\pi,a) := D\left[\pi(S|a) \parallel P(S)\right] \quad \text{where} \quad \pi(s|a) := \frac{\pi(a|s)}{P(a)} \cdot P(s),
\]

and \(D[\bullet||\bullet]\) is the Kullback-Leibler divergence \(D[p||q] := \sum p_i \log \frac{p_i}{q_i}\).

**Remark 1.** The expected effective information generated by an agent is mutual information:

\[
E[ei(\pi,a) | P(a)] = I_\pi(S, A).
\]

Effective information quantifies the information gained about \(S\) by updating prior \(P(S)\) to posterior \(\pi(S|a)\) via Bayes’ rule. Given policy \(\pi\), effective information associates a non-negative scalar to each action available to the agent:

\[
ei_\pi : A \rightarrow \mathbb{R} : a \mapsto ei(\pi,a).
\]  

An alternate approach to quantifying information transfer is local transfer entropy [15], which (roughly) computes the information transferred by a situation-action pair. Since in our setup an agent only communicates actions to other agents – and not the situations causing the actions – we average over situations to obtain effective information.

### 3 Constrained reward maximization

Inspired by ideas from thermodynamics a number of authors have recently investigated the effects of information-theoretic constraints on optimization problems [7–9]. This section adapts these ideas to investigate how metabolically constraining reward-maximization impacts the information content of actions.

The goal of each agent is to maximize expected reward, as quantified by \(R: S \times A \times \mathcal{N} \rightarrow \mathbb{R}\). Since agents take metabolic cost into account, we introduce the following optimization problem:

**Definition 2 (constrained reward maximization).**

The optimal policy solves the constrained optimization problem

\[
\pi^* := \arg \min_{\pi} \mathbb{E}\left[C_\pi(S, A) - \beta \cdot R_\pi(S, A) \mid P(s) \cdot \pi(a|s)\right]
\]

Parameter \(\beta\) controls the relative weight given to the reward and metabolic cost.

\[\text{We extend the definition of effective information in [13] to allow arbitrary priors instead of restricting to the uniform distribution.}\]
Theorem 1 (solutions to free-energy optimization).

The optimal policy \( \pi^* \), minimizing (5), satisfies

\[
\pi^*(a|s) = P(a) \cdot \frac{e^{\beta \cdot R(s,a)}}{Z(s,\beta)}
\]

where \( Z(s,\beta) = \sum_{a \in A} P(a) \cdot e^{\beta \cdot R(s,a)} \) is the normalization constant.

Proof. See appendix.

Theorem 1 is important because it explicitly describes how the optimal policy encodes rewards into decisions. However, it is not reward but rather relative reward that determines an agent's actions:

Definition 3 (relative reward).

For a given \( \beta \), let the relative reward of action \( a \) in situation \( s \) be

\[
\bar{R}(a,s) := R(a,s) - \log Z(s,\beta).
\]

The definition is motivated by the following lemma:

Lemma 2.

• Minimizing free energy with respect to rewards \( R(s,a) \) and relative rewards \( \bar{R}(s,a) \) yields the same optimal policy

\[
\arg \min_{\pi} C_\pi(S,A) - \beta \cdot R(S,A) = \arg \min_{\pi} C_\pi(S,A) - \beta \cdot \bar{R}(S,A).
\]

• The optimal policy has functional form

\[
\pi^*(a|s) = P(a) \cdot e^{\bar{R}(s,a)}
\]

Proof. If an agent finds itself in situation \( s \), it is the difference \( R(s,a) - R(s,a') \) that determines which action it chooses. Adding constant \( R(s,a) \mapsto R(s,a) + \lambda(s) \) and \( R(s,a') \mapsto R(s,a') + \lambda(s) \) makes no difference to the actions of an optimal agent since, by the assumption in Eq. (2), individual agents cannot affect the probability of situations arising in the future.

The lemma says (optimal) agents only care about the relative reward of actions, rather than their absolute reward.

Definition 4 (the importance of an action).

The importance of action \( a \) is the expected relative reward:

\[
\mathbb{E}\left[\bar{R}_\pi(S,a) \mid p(S|a)\right] = \sum_{s \in S} \pi(s|a) \cdot \bar{R}(s,a).
\]

We introduce Definition 4 to avoid repeatedly using the cumbersome phrase “expected relative reward”.

Theorem 3 (\( ei \) quantifies importance up to a metabolic discrepancy).

For an optimal policy \( \pi^* \), effective information quantifies the importance of an action up to a metabolic discrepancy term:

\[
ei(\pi^*, a) = \underbrace{\mathbb{E}\left[\bar{R}_\pi(S,a)\right]}_{\text{expected relative reward}} + \underbrace{\log P(a)}_{\text{metabolic discrepancy}} \cdot \pi^*(a|s).
\]

Proof. Writing out the definition yields

\[
ei(\pi^*, a) = \sum_{s \in S} \pi(s|a) \log \frac{\pi^*(s|a)}{P(s)} = \sum_{s \in S} \pi^*(s|a) \log \frac{\pi^*(a|s)}{\pi^*(a)}
\]
By Theorem 1, the optimal policy satisfies $\pi^*(a|s) = P(\cdot) \cdot e^{\beta \tilde{R}(s,a)}$, so

$$
ei(\pi^*, a) = \sum_s \pi^*(s|a) \cdot \tilde{R}(s,a) - \log \pi^*(a) \frac{P(\cdot)}{P(a)} = \mathbb{E} \left[ \tilde{R}_{\pi^*}(S,a) \middle| \pi^*(S|a) \right] + \log \frac{P(a)}{\pi^*(a)}.
$$

Theorem 3 is our first main result. It says that, if the frequency of actions under policy $\pi$ coincides with the metabolic prior $P(\cdot)$, then the importance of an (optimal) agent’s actions coincides with the effective information it generates.

**A strong metabolic constraint.** Agents “in the wild”, such as organisms interacting with potentially hostile environments, are forced to take or not take specific actions to avoid death. In such a scenario, there is no reason to expect a tight relationship between effective information and importance, and the metabolic discrepancy is likely to be quite large. Unlike agents “in the wild”, neuron-like agents in cortex-like environments do not face life-or-death situations. Their actions are a means of communication, and only indirect influence global outcomes. There is therefore more freedom to constrain the policies of agents since costs apply directly and neuromodulatory signals are highly indirect.

Let us consider the consequences of taking metabolic costs into account strongly

$$P(a) = \pi^*(a) \quad \text{for all } a \in A,$$

where $P(a) := \frac{1}{\lambda_{\circ c(a)}}$ and $\pi^*(a) := \sum_{s \in S} \pi^*(a|s) \cdot P(s)$ for optimal policy $\pi^*$. Section §5 presents two learning algorithms whose policies satisfy assumption $\text{MP}$.

By Theorem 3 effective information quantifies the importance of actions by an optimal agent, to the extent that the frequency of its actions coincides with the metabolically determined reference distribution $P(\cdot)$. To compute effective information, it is necessary to observe how an agent responds to different situations. Neurons in the brain do not have access to each other’s inputs, so computing effective information is difficult. Fortunately, the next theorem shows that neurons do not need to compute the effective information generated by other neurons:

**Theorem 4 (High $\beta$ couples $\text{ei}$ to metabolic cost and importance).** Suppose $\pi^*$ is an optimal policy for an agent satisfying Assumption $\text{MP}$. In the deterministic limit $\lim_{\beta \to \infty}$, for all $a \in A$, we have

$$
ei(\pi^*, a) = \mathbb{E} \left[ \tilde{R}_{\pi^*}(S,a) \middle| \pi^*(S|a) \right] = \log \left( \lambda \circ c(a) \right).
$$

**Proof.** For a deterministic policy, $\text{ei}(\pi, a) = -\log \pi(a)$, which equals $\lambda \circ c(a)$, see paragraph after Eq. (3). Effective information equals importance by Theorem 3.

The theorem says that, so long as agents share a common metabolic cost profile, effective information and importance are aligned throughout the population. As a corollary we obtain that, so long as spikes are infrequent, relative reward is almost entirely concentrated on spikes and so silences can be ignored:

**Corollary 5 (Reward after spiking approximates total reward).** Suppose an agent has two actions (silence $a_0$ and spike $a_1$) and satisfies assumption $\text{MP}$. Then the relative reward after all actions by the agent is approximately the relative reward after spiking alone:

$$
\mathbb{E} \left[ \tilde{R}_{\pi^*}(S,A) \right] = \pi(a_1) \cdot \mathbb{E} \left[ \tilde{R}_{\pi^*}(S,a_1) \right] + O(\pi(a_1)^2).
$$
Proof. Theorem 8 in Appendix A3 shows that

\[ I_\pi(S, \mathcal{A}) = \pi(a_1) \cdot ei(\pi, a_1) + O(\pi(a_1)^2) . \]

The corollary follows after observing that \( I_\pi(S, \mathcal{A}) = \sum_{a \in \mathcal{A}} \pi^*(a) \cdot ei(\pi^*, a) \) and, similarly, \( E[\bar{R}_\pi(S, \mathcal{A})] = \sum_{a \in \mathcal{A}} \pi^*(a) \cdot E[\bar{R}_\pi(S, a)] \). \hfill \square

Theorem 4 is our second main result. It says that when an agent strongly controls the frequency of its actions via the metabolic prior, assumption MP, the information generated by its actions coincides exactly with their importance, which also couples tightly with their metabolic cost. Moreover, by Corollary 5 agents do not lose much by only attending to the spikes produced by other agents and ignoring the silences. This suggests agents in a population sharing a common metabolic cost profile can use metabolic cost to distinguish between more and less important actions by their neighbors.

4 Empirical reward estimates

We have referred to a policy’s expected reward throughout this paper. However, the expected distribution of rewards is not known; agents can only access empirical estimates based on finite samples. In a dynamically changing environment like the cortex, it is important that neurons revise their estimates over short time frames—i.e., accurately estimate expected reward from small samples. Guaranteeing the accuracy of an agent’s estimate of its expected reward is critical, since the estimate determines the agent’s policy.

This section uses effective information to bound the difference between expected reward and its empirical estimate. In short: the tighter the metabolic constraint—i.e., the fewer the set of situations for which an agent spikes—the better the quality of its empirical reward estimate.

We consider expected and empirical rewards after spiking since we are interested in the expected relative reward encoded in spikes. Many models of neuronal learning use correlations between spikes or the relative timing of spikes to control synaptic (de)potentation, so that it is rewards after (or just before) output spikes that determine a neuron’s policy [16]. Restricting attention to rewards after spiking also provides a mechanism that forces neurons to specialize by focusing on a narrow band in their input stream.

The expected and empirical rewards per spike are

\[ R_\pi := E[R(s, a_1, N)] := \sum_{s, a} \pi(s|a_1) \cdot P(n|s) \cdot R(s, a_1, n) \]

and

\[ \bar{R}_\pi := \frac{1}{T_1} \sum_{(s, a_1)(\pi(s_1)=1)} R(s_1, a_1, n_t) \]

respectively, where \( T_1 \) counts spikes produced by the agent during \([1, T]\.\)

Let us introduce notation for computations with respect to the uniform prior \( U(s) = \frac{1}{|S|} \), where \(|S|\) is the total number of possible situations. Let \( \pi_u(a) = \sum_s \pi(a|s) \cdot U(s) \), \( \pi_u(s|a) = \pi(a|s) \cdot \frac{U(s)}{\pi_u(a)} \), and \( ei_u(\pi, a) = \sum_s \pi_u(s|a) \log \frac{\pi(\pi_u(s|a))}{\pi_u(s|a)} \). Note that \( ei_u \) recovers the original definition of effective information in [13], which Definition 1 generalizes.

The following theorem is proved using a version of Occam’s razor due to Seldin and Tishby [17]. Without loss of generality we assume that rewards lie in \([0, b]\. Since only relative rewards affect the choice of optimal policy, it follows that negative rewards can be stripped out of the optimization problem by introducing an additive constant.

Theorem 6 (Error bound for empirical reward).

Suppose the agent chooses a deterministic policy \( \pi \) under the constraint that it spikes for a fixed fraction of situations: \( \pi_u(a_1) = \text{const} \). Further, suppose that situations are sampled i.i.d. Then with probability at least \( 1 - \delta \),

\[ |R_\pi - \bar{R}_\pi| \leq b \cdot \sqrt{|S| \cdot \frac{ei_u(\pi, a_1)}{2T_1} \cdot \frac{1}{e^{\pi_u(\pi, a_1)}} + \frac{\log \frac{2}{\delta}}{2T_1}} \]
Theorem 6 and the experiments in Fig. 1a are our third main result. They show that the higher the effective information generated by spiking, the better an agent’s empirical estimate of its expected reward after spiking. The metabolic prior, Assumption MP, thus controls the quality of an agent’s empirical estimates of its expected reward.

Remark 2. The assumption that inputs are i.i.d. is not realistic for a system of interacting agents such as cortex. We make two remarks. First, if rewards are only non-zero in the presence of neuromodulatory signals, then the assumption is more reasonable since it states that situations directly preceding neuromodulator release are i.i.d. Second, similar results have recently been obtained for non-i.i.d. scenarios using PAC-Bayes methods \[18\] and these may be applicable to our setting.

A related result in \[19\] is that the effective information generated by empirical risk minimization coincides (essentially) with empirical VC-entropy, which appears in bounds on expected risk.
5 Implications for cooperative learning

The previous sections showed that metabolically constrained agents spike in the few situations that, on the basis of their historical experience, most frequently lead to high rewards. This suggests cooperating agents should treat each other’s spikes as intrinsically valuable.

**Example 1** (Minimal cooperative reward function).
Consider reward function
\[ R(s, a, n) = |s| \cdot |a| \cdot n, \]  
where \( |s| \) is the total number of spikes in the input, \( |a| \) is the number of spikes in the output, and \( n \) is the neuromodulatory signal.

The motivation for \( S \) is as follows.

1. Neuromodulatory signal \( n \) controls the sign of the reward. Positive signals are reinforced and negative avoided.
2. Agents reinforce or punish their decisions only if they actively contribute to the global outcome. Active contributions are distinguished by their higher metabolic cost. Term \( |a| \), which is either 0 or 1, therefore gates the reward function.
3. An individual agent spiking in isolation cannot impact global outcomes. A simple estimate of the impact of a population of agents is the number that spike. The amplitude \( |s| \) of the reward therefore reflects local spiking activity: the more spikes there are, the more likely it is that the agent belongs to a population that impacted global behavior. Term \( |s| \) therefore amplifies local rewards according to their likely global impact.

Note that learning under reward function \( S \) is a population-level phenomenon due to the \( |s| \) term. Thus, although individual agents are unable to affect their environment, reward function \( S \) encourages them to behave as if they are responsible for global outcomes, by taking responsibility for neuromodulatory signals proportionally to the size of the local spiking population they belong to.

An impractical – but nonetheless instructive – learning algorithm that implements reward function \( S \) is the following.

**Example 2** (Brute force optimization).

Given \( M \subset \mathbb{R} \), let \( M_x := \{ \text{the } x \cdot |M| \text{ elements with highest magnitude} \} \).

**Algorithm:**

1. Samples reward estimates \( M = \{ \hat{R}(s, a_1) | s \in \mathcal{S} \} \).
2. Rank elements of \( M \) and compute \( M_{P(a_1)} \).
3. The optimal policy is
   \[ \pi_{BF}(a_1 | s) = \begin{cases} 1 & s \in M_{P(a_1)} \\ 0 & \text{else.} \end{cases} \]

Brute-force optimization can be visualized as moving a window of fixed size and variable shape over the input space, searching for the configuration that maximizes reward. It implements the optimization
\[
\arg \max_x \left[ \sum_{s,a} P(s) \cdot \pi(a | s) \cdot \hat{R}(s, a) \right]
\]
subject to: \( \sum_s \pi(a_1 | s) \cdot P(s) = P(a_1) \).

Reward function \( S \) ensures the agent only has to search over rewards after spiking, i.e. those of the form \( R(s, a_1), s \in \mathcal{S} \).
An important question, motivated by the result following from assumption [MP], is how a strong metabolic constraint affects the encoding of rewards into spikes. The expected reward (not importance) after spiking is

\[ R(a_1) = \mathbb{E}\left[ |s| \cdot n \left| p(n, s | s \in M_{P(a_1)} ) \right. \right]. \]

The expected reward compresses a large amount of data – specifically, the structure of distributions \( p(n|s) \) and \( p(s|a_1) \) – into a single number. Figure 10 shows an example, where situations are ranked according to expected reward and the policy spikes for the 15% of situations with highest reward. By construction the expected reward after spiking is much higher than after silence, so spikes encode reward. Tightening the metabolic constraint, so that the policy spikes for \( \ll 15\% \) of situations, increases the reward after spiking. Furthermore, the variance in reward after spiking is much lower than after not spiking and, once again, tightening the metabolic constraint will tend to further decrease the variance after spiking.

Thus, under tight metabolic constraints, the spikes produced by brute-force optimized policies are robust signals of high expected reward.

More generally, consider two scenarios:

- If the rewards are distributed evenly over \( M_{P(a_1)} \), i.e. \( R(s, a_1) \approx R(s', a_1) \) for all \( s, s' \in M_{P(a_1)} \), then spikes are robust predictors of rewards.
- If high rewards concentrate on a small subset \( H \) of \( M_{P(a_1)} \), i.e. \( R(s', a_1) \ll R(s, a_1) \) for \( s \in H \) and \( s' \not\in H \), then rewards after spiking are more variable.

The first scenario is preferable when agents use spikes as local estimates of expected reward: variance should be as low as possible. A simple way to ensure evenly distributed rewards after spiking is to keep \( M_{P(a_1)} \) small by forcing low \( \pi(a_1) \), which ties in with the results on empirical estimates in [4].

Brute-force optimization separates exploration and exploitation into two distinct steps. A “softer” algorithm that becomes more certain as it acquires more data is a modified version of Q-learning [20].

**Example 3** (Metabolically constrained Q-learning).

If the agent chooses action \( a \) in situation \( s \) and subsequently receives neuromodulator \( n \), then it updates the Q-matrix by

\[ Q(s, a) \leftarrow Q(s, a) + \alpha \cdot \left[ R(s, a, n) - Q(s, a) \right]. \]

where \( \alpha \) controls the rate. After updating \( Q \), the agent constructs new policy

\[ \pi(a|s) = \mathcal{M}^n \left( e^{Q(s,a)} \right). \]

Operation \( \mathcal{M}(\bullet) \) renormalizes the policy twice: first by \( Z(a) \) chosen such that \( \sum_{a \in S} \pi(a|s)P(s) = P(a) \) for all \( a \), and then by \( Z(s) \) chosen such that \( \sum_{a \in A} \pi(a|s) = 1 \) for all \( s \). Setting \( n = 3 \) yields a policy that approximately implements the metabolic constraint.

Figure 2 considers how effective information and expected reward covary as agents Q-learn. We initialized 5000 Q-learning agents with \( \pi(a, s) = P(s) \cdot P(a) \) for \( P(a_1) \in \{0.1, 0.3, 0.5\} \). As the agents adapt, their policies tend to become more both more deterministic and more likely to spike in situations yielding higher rewards, so as agents adapt they both generate more effective information and more rewards after spiking. The tighter the metabolic constraint (i.e. the lower \( P(a_1) \)), the higher the expected reward after spiking. Thus, effective information provides a reliable guide to expected reward, even in the presence of a strong metabolic constraint.

**Remark 3.** Under metabolic Q-learning, term \( |s| \) in reward function [5] uses the degree of local consensus to control the speed with which an agent learns.
Figure 2: **Metabolically constrained Q-learning.** As agents learn, effective information and expected reward increase in qualitatively the same way. Tighter metabolic constraints yield both higher effective information and greater expected rewards.

An important point is that both brute-force optimization and Q-learning are not practical learning rules; they simply construct lookup tables. In contrast, cortical neurons do not control the probability of spiking for individual inputs – but rather for individual synapses. When a neuron potentiates a synapse, its response to many inputs is altered. Neurons thus generalize in a manner that depends on the structure of their dendritic tree.

This section presented two simple implementations of metabolically constrained learning that privileges spikes. Future work will investigate how populations of such learners behave.

### 6 Discussion

The space of possible policies that the cortex as a whole could implement is vast: it consists of millions or billions of agents each receiving inputs from thousands or tens of thousands of other agents. Cutting down the size of the search space is crucial. This paper has shown that maximizing reward under strong metabolic constraints leads to a tight coupling between the metabolic cost, information content and importance of actions. This suggests that it may be useful to focus the attention of agents by privileging spikes during learning.

In summary:

1. Agents attempt to maximize their expected reward, constrained by the metabolic cost of their actions.
   [ Definition 2, Assumption MP ]
2. If an agent behaves optimally then the information generated by its actions aligns with both their importance and their metabolic cost.
   [ Theorem 3, Theorem 4 ]
3. The less frequently an agent uses expensive actions, the better its empirical reward aligns with importance.
   [ Theorem 6, Figure 1a ]
4. The total contribution of metabolically inexpensive actions is negligible.
   [ Corollary 5 ]
5. If agents share a common metabolic cost profile, then it is obvious which actions predict high reward
   [ Spikes, with high confidence ]
6. Therefore, agents in a cooperating population should treat costly inputs and outputs (spikes) as intrinsically valuable.
An intriguing analogy is that spikes form a cortical currency which neurons use to communicate estimates of future rewards. If neurons spike too frequently, then the reward encoded in spikes drops, analogously to how inflation degrades the worth of a currency. Money provides a measure of importance that agents in a market economy use to guide their behavior. If all goes well, agents with consistently contribute to the global good by performing strictly local optimizations.

Spikes, and more generally metabolically expensive actions, thus provide local guidelines that a populations of cooperating agents can exploit during learning.

Rewarding spikes introduces a tendency towards runaway spiking (e.g. epileptic seizures) which needs to be controlled. Lateral inhibition, which we have not discussed in this paper, is a mechanism to impose sparse activity in the brain that may play a useful role in other cooperating populations.

Imposing the metabolic constraint requires active effort. In the case of biological neurons it has been proposed that controlling synaptic efficacies, and so the tendency of neurons to spike, is a function of slow-wave sleep. Recent experiments have found evidence in line with this hypothesis, see.

The results in this paper depend on specific assumptions and model choices. However, we believe these can be relaxed without substantially changing the main message. Future work will investigate the computational role of spike-timing dependent learning in cortex from this perspective, to better understand how neurons encode information and rewards into spikes.

References

[1] Sutton RS, Barto AG: Reinforcement Learning: An Introduction. MIT Press 1998.
[2] Russell S, Norvig P: Artificial Intelligence: A Modern Approach. Prentice Hall, 3rd edition 2009.
[3] Dan Y, Poo MM: Spike timing-dependent plasticity: from synapse to perception. Physiol Rev 2006, 86(3):1033–1048.
[4] Pawlak V, Wickens JR, Kirkwood A, Kerr JND: Timing is not everything: neuromodulation opens the STDP gate. Front. Syn. Neurosci 2010, 2(146).
[5] Hasenstaub A, Otte S, Callaway E, Sejnowski TJ: Metabolic cost as a unifying principle governing neuronal biophysics. Proc Natl Acad Sci U S A 2010, 107(27):12329–34.
[6] Attwell D, Laughlin SB: An energy budget for signaling in the grey matter of the brain. J Cereb Blood Flow Metab 2001, 21(10):1133–45.
[7] Tishby N, Pereira F, Bialek W: The information bottleneck method. In Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing. Edited by Hajek B, Sreenivas R, University of Illinois Press 1999.
[8] Tishby N, Polani D: Information theory of decisions and actions. In Perception-Reason-Action. Edited by Vassilis, Hussain, Taylor, Springer 2011.
[9] Ortega PA, Braun DA: Information, utility and bounded rationality. The fourth conference on artificial general intelligence 2011, :269–274.
[10] Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W: Spikes: exploring the neural code. MIT Press 1997.
[11] Langford J, Zhang T: The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits. In Advances in neural information processing systems 2007.
[12] Klyubin AS, Polani D, Nehaniv CL: Representations of Space and Time in the Maximization of Information Flow in the Perception-Action Loop. Neural Comp 2007, 19(9):2387–2432.
[13] Balduzzi D, Tononi G: Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. PLoS Comput Biol 2008, 4(6):e1000091.

This perspective dates back to Selfridge’s Pandemonium, a learning system comprised of “shrieking demons” analogous to our spiking agents.
Theorem 1
The optimal policy satisfies
\[ \pi(a|s) = P(a) \cdot \frac{e^{\beta R(s,a)}}{Z(s,\beta)} \]
where \( Z(s,\beta) = \sum_{a \in A} P(a) \cdot e^{\beta R(s,a)} \) is the normalization constant.

Proof. The proof is adapted from the information bottleneck method in [7]. Let \( \lambda(s) \) be Lagrange multipliers, and consider
\[ F_\pi = \sum_{s,a} P(s)\pi(a|s) \log \frac{\pi(a|s)}{P(a)} - \beta \sum_{a,s,r} P(s)\pi(a|s)P(r|a,s)R(s,a,r) - \sum_{s,a} \lambda(s)\pi(a|s). \]
The derivative with respect to \( \pi(a|s) \) is
\[ \frac{\delta F}{\delta \pi(a|s)} = P(s) \left[ 1 + \log \frac{\pi(a|s)}{P(a)} \right] - \beta \sum_{r} P(s)P(r|a,s)R(s,a,r) - \beta \lambda(s). \]
We have that \( \frac{\delta P(s)}{\delta \pi(a|s)} = 0 \) because \( P(s) \) does not depend on the choice of policy by the assumption in Eq. [7].
Theorem follows since \[ \pi \]

**Proof.**

Let \( h \) denote the complexity of hypothesis \( \mathcal{H} \). Suppose we have a set of hypotheses \( \mathcal{H} \). \( \text{Occam’s razor} \) can be paraphrased to say that the simplest hypothesis should be preferred. Let \( L : \mathcal{X} \times \mathcal{H} \rightarrow [0, \delta] \) be a loss function. Then

**Theorem 7** (Occam’s razor).

For any data generating distribution on \( \mathcal{X} \) and any prior distribution \( P(h) \) over \( \mathcal{H} \), with a probability greater than \( 1 - \delta \) over drawing an i.i.d. sample from \( \mathcal{X} \) of size \( T \), for all \( h \in \mathcal{H} \):

\[ \left| L(h) - \hat{L}(h) \right| \leq b \sqrt{\frac{-\log P(h) + \log \frac{2}{\delta}}{2T}}. \]

**Proof.** See \[17\].

Let \( \mathcal{H} = \{ \pi : \mathcal{S} \rightarrow \mathcal{A} \} \) denote the set of deterministic policies, where \( \mathcal{A} = \{a_0, a_1\} = \{0, 1\} \).

Define loss function \( L : (\mathcal{S} \times \mathcal{N}) \times \mathcal{H} \rightarrow \mathbb{R} : (s, n) \times \pi \rightarrow R(s, \pi(s), n) \).

Further, set probability distribution \( P(s, n) = P(s) \cdot P(n|s) \) on \( \mathcal{S} \times \mathcal{N} \). Theorem 7 holds for any sampling distribution. In particular we may use the policy \( \pi \) to restrict samples to situations that cause the agent to spike to obtain \( P(s, n|\pi(s) = a_1) = \pi(s|a_1) \cdot P(n|s) \). It follows that

\[ L(\pi) = R_\pi = \mathbb{E} \left[ R(s, a_1, n) \mid \pi(s|a_1) \cdot P(n|s) \right] \] is the expected reward and

\[ \hat{L}(\pi) = \hat{R}_\pi = \frac{1}{T_1} \sum_{\{s, n, \pi(s)=a_1\}} R(s_t, a_1, n_t) \] is the empirical reward, where \( T_1 \) is the number of spike produced by the agent during \([1, T]\).

**Theorem 6**

Let \( \mathcal{H}_k = \{ \pi : \mathcal{S} \rightarrow \mathcal{A} \mid |\pi^{-1}(a_1)| = k \} \) denote policies that spike for exactly \( k \) situations. Given the setup above, with probability at least \( 1 - \delta \),

\[ \left| R_\pi - \hat{R}_\pi \right| \leq b \sqrt{\frac{|\mathcal{S}|}{2T_1} \cdot e^{i_u(\pi, a_1)} + 1} \cdot \frac{2}{2T_1} \log \frac{2}{\delta}. \]

**Proof.** Let \( N = |\mathcal{S}| \) denote the number of possible situations. We put the uniform prior on \( \mathcal{H}_k \), so \( P(\pi) = \frac{1}{k^k} \). By Stirling’s approximation, \( \log \binom{N}{k} \leq k \log \left( \frac{Ne}{k} \right) \), and it follows that

\[ -\log P(\pi) = \log \binom{N}{k} \leq k \log \frac{N \cdot e}{k} = N \cdot \frac{k}{N} \cdot \log \left( \frac{N}{k} + 1 \right) \]

The theorem follows since \( \pi_u(a_1) = \frac{k}{N} \) and \( e^{i_u(\pi, a_1)} = \log \frac{N}{k} \).
Effective and mutual information

For an agent operating under high metabolic constraints, it turns out that effective information provides a useful approximation to mutual information:

**Theorem 8** (effective information approximates mutual information).

Suppose an agent has two actions (silence \( a_0 \) and spike \( a_1 \)) and produces spikes very infrequently: \( \pi(a_1) \ll 1 \). Then the total information transferred by the agent is approximately the information it transfers using spikes alone:

\[
I(S; A) = \pi(a_1) \cdot ei(\pi, a_1) + O(\pi(a_1)^2).
\]

**Proof.** Observe that

\[
P(s|a_0) = \frac{P(s) - \pi(s|a_1) \cdot \pi(a_1)}{1 - \pi(a_1)}
\]

\[
= P(s) + \pi(a_1) \left( P(s) - \pi(s|a_1) \right) + O(\pi(a_1)^2).
\]

Thus to first order in \( \pi(a_1) \), \( \pi(s|a_0) \) is of the form \( p + \delta p \) where \( \int \delta p = 0 \). We can then compute

\[
D[p + \delta p \parallel p] = \int (p + \delta p) \log_2 \frac{p + \delta p}{p} \delta p
\]

\[
= \alpha \int (p + \delta p) \frac{\delta p}{p} \left( 1 - \frac{\delta p}{2p} \right) + O((\delta p)^3) = \alpha \int \frac{(\delta p)^2}{2p} + O((\delta p)^3),
\]

where \( \alpha = \frac{1}{\ln 2} \).

Substituting \( p = P(s) \) and \( \delta p = \pi(a_1)(P(s) - \pi(s|a_1)) \) gives

\[
D[\pi(S|a_0) \parallel P(S)] = D[p + \delta p \parallel p] + O(\pi(a_1)^2)
\]

\[
= \alpha \cdot \pi(a_1)^2 \int \frac{(P(s) - \pi(s|a_1))}{2P(s)} + O(\pi(a_1)^2).
\]

Thus \( I(S; A) = \pi(a_1)D[\pi(S|a_1) \parallel P(S)] + O(\pi(a_1)^2). \) \( \square \)