AN ADAPTIVE HOMEOSTATIC ALGORITHM FOR THE UNSUPERVISED LEARNING OF VISUAL FEATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

The formation of structure in the brain, that is, of the connection between cells within neural populations, is by large an unsupervised learning process: The emergence of this architecture is mostly self-organized. In the primary visual cortex of mammals, for example, one may observe during development the formation of cells selective to localized, oriented features. This leads to the development of a rough representation of contours of the retinal image in area V1. We modeled these mechanisms using sparse Hebbian learning algorithms. These algorithms alternate a coding step to encode the information with a learning step to find the proper encoder. A major difficulty faced by these algorithms is to deduce a good representation while knowing immature encoders, and to learn good encoders with a non-optimal representation. To address this problem, we propose here to introduce a new regulation process between learning and coding, called homeostasis. Our homeostasis is compatible with a neuro-mimetic architecture and allows for the fast emergence of localized filters sensitive to orientation. The key to this algorithm lies in a simple adaptation mechanism based on non-linear functions that reconciles the antagonistic processes that occur at the coding and learning time scales. We tested this unsupervised algorithm with this homeostasis rule for a range of existing unsupervised learning algorithms coupled with different neural coding algorithms. In addition, we propose a simplification of this optimal homeostasis rule by implementing a simple heuristic on the probability of activation of neurons. Compared to the optimal homeostasis rule, we show that this heuristic allows to implement a more rapid unsupervised learning algorithm while keeping a large part of its effectiveness. These results demonstrate the potential application of such a strategy in machine learning and we illustrate this with one result in a convolutional neural network.

1 INTRODUCTION

The neural architecture is a complex dynamic system that operates at different time scales. In particular, one of its properties is to succeed in representing quickly information (the coding step) while optimizing in the long term its encoding (the learning step). In the case of the mammalian primary visual cortex (V1) for instance, this rapid coding operation, of the order of 50 milliseconds in humans, is the key to the results of Hubel & Wiesel (1968), who showed that some cells of V1 have relatively localized receptive fields which are predominantly selective at different orientations. As such, one can consider the rapid coding of the retinal image as a process of transforming the raw visual information into a rough "sketch" that represents the outlines of objects in the image by using elementary edge-like features. This internal representation and the visual information share the same property of being *sparse*: for most natural images, only a relatively small number of features are necessary to describe the input. Thus, the coding step consists in choosing the right encoder that selects as few features (called atoms) as possible among a collection of them (called the *dictionary*). Amazingly, Olshausen & Field (1996) have shown that when enforcing a sparse prior on the encoding step, such edge-like filters are emerging using a simple Hebbian unsupervised learning strategy.

Additionally, recent advances in machine learning, and especially on unsupervised learning have shed new light on the functioning of the underlying biological neural processes. By definition, unsupervised learning aims at learning the best dictionary to represent the input image autonomously, that is, without using other external knowledge such as in supervised or reinforcement learning.

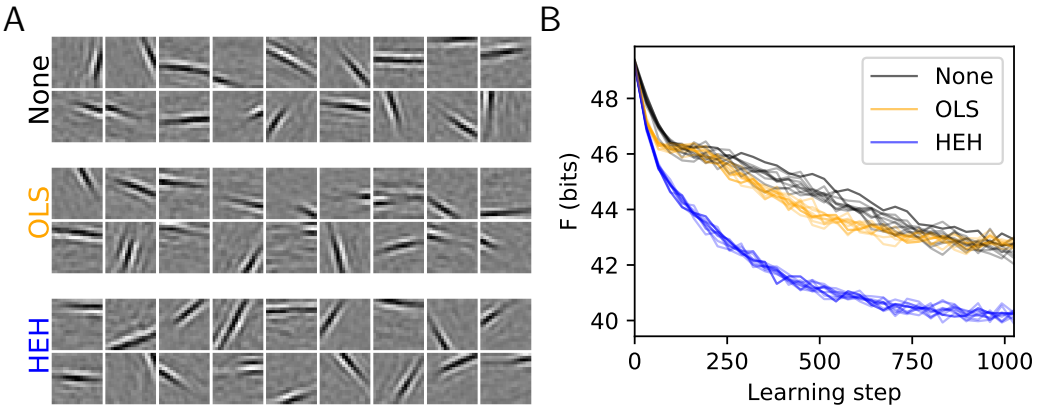

Figure 1: **Role of homeostasis in learning sparse representations**: We show the results of Sparse Hebbian Learning using different homeostasis algorithms at convergence (1024 learning epochs). The compared algorithms are : `None` (using a simple simple normalization of the atoms), `OLS` (the method of (Olshausen & Field, 1997)), `HEH` (using the optimal homeostasis described in this paper). (A) For each algorithm, we show 18 randomly drawn atoms from the 441 filters of the same size as the image patches ($M = 18 \times 18 = 324$) and presented in a matrix (separated by a white border). (B) Evolution of cost F (in bits, see Eq. 4) as a function of the number of iterations and cross-validated over 10 runs. While `OLS` provides a similar convergence than `None`, the `HEH` method provides a better final convergence.

Algorithms that combines such learning as the input to classical, supervised deep-learning show great success in tasks like image denoising (Vincent et al., 2008) or classification (Sulam et al., 2017). A variant consists in forcing the generated representation to be sparsely encoded (Makhzani & Frey, 2013), whether by adding a penalty term to the optimized cost function or by encoding each intermediate representation by a pursuit algorithm (Papyan et al., 2016). Interestingly, (Papyan et al., 2016) proposes a model of Convolutional Sparse Coding (CSC) tightly connected with Convolutional Neural Network (CNN), so much that the forward pass of the CNN is equivalent to a CSC with a thresholding pursuit algorithm. These unsupervised algorithms are equivalent to a gradient descent optimization over an informational-type coding cost (Kingma & Welling, 2013). This cost makes it then possible to *quantitatively* evaluate the joint exploration of new learning or coding strategies. As such, this remark shows us that unsupervised learning consists of two antagonistic mechanisms, a long time scale that corresponds to the learning and exploration of new components and a faster scale that corresponds to coding.

In particular, an aspect often ignored in this type of learning is the set of homeostasis mechanisms that control the average activity of neurons within a population. Indeed, there is an intrinsic complexity in unsupervised dictionary learning algorithms: how to adapt the regularization parameter of each atom to make sure no atoms are wasted because of improper regularization settings? In the original algorithms of sparse unsupervised learning (Olshausen & Field, 1997), homeostasis is implemented as a heuristic that prevents the algorithm from diverging. In most unsupervised learning algorithms it takes the form of a normalization, that is, an equalization of the energy of each atom in the dictionary (Mairal et al., 2014). However, the neural mechanisms of homeostasis are at work in many components of the neural code and are essential to the overall transduction of neural information. For example, the sub-networks of glutamate and GABA-type neurons may regulate the overall activity of neural populations (Marder & Goaillard, 2006). In particular, such mechanisms could be tuned to balance the contribution of the excitatory populations with respect to that in inhibitory populations. As a consequence, this creates a so-called balanced network which may explain many facets of the properties of the primary visual cortex (Hansel & van Vreeswijk, 2012). At the modelling level, these mechanisms are often implemented in the form of normalization rules (Schwartz & Simoncelli, 2001) which are considered as the basis of a normative theory to explain the function of the primary visual cortex (Carandini & Heeger, 2012). However, when extending such model using unsupervised learning, most modelling effort is rather intended to show that the cells' selectivity that emerges have

the same characteristics than those observed in neuro-physiology (Ringach, 2002; Rehn & Sommer, 2007; Loxley, 2017). Other algorithms use non-linearities that implicitly implement homeostatic rules in neuro-mimetic algorithms (Brito & Gerstner, 2016). These non-linearities are mainly used in the output of successive layers of deep learning networks that are nowadays widely used for image classification or artificial intelligence. However most of these non-linear normalization rules are based on heuristics mimicking neural mechanisms but are not justified as part of the global problem underlying unsupervised learning. Framing this problem in a probabilistic framework allows to consider in addition to coding and learning the intermediate time scale of homeostasis and allows us also to associate it to an adaptation mechanisms (Rao & Ballard, 1999). Our main argument is that by optimizing unsupervised learning at different time scales, we allow for the implementation of fast algorithms compatible with the performance of biological networks and in comparison with classical (Olshausen & Field, 1997) or Deep Learning approaches.

In this paper, we will first define a simple algorithm for controlling the selection of coefficients in sparse coding algorithms based on a set of non-linear functions similar to a generic neural gain normalization mechanisms. Such functions will be used to implement an homeostasis mechanism based on histogram equalization by progressively adapting these non-linear functions. In particular, this algorithm will extend an already existing algorithm of unsupervised sparse learning (Perrinet, 2010) to a more general setting. In particular, we will show quantitative results of this optimal algorithm by applying it to different pairs of coding and learning algorithms. Second, we will propose a simplification of this homeostasis algorithm based on the activation probability of each neuron and show that it yields similar quantitative results as the full homeostasis algorithm and that it converges more rapidly than classical methods (Olshausen & Field, 1997; Sandin & Martin-del Campo, 2017). All these algorithms were implemented using `Python` (version 3.6.5) with packages `NumPy` (version 1.14.3), `sklearn` (version 0.19.1) and `SciPy` (version 1.1.0) (Oliphant, 2007). Visualization was performed using `Matplotlib` (version 2.2.2) (Hunter, 2007). In particular, we focused in our architecture to be able to quantitatively cross-validate for every single hyper-parameters and all these scripts are available at `https://github.com/XXX/ZZZ`. Finally, we will conclude by showing an application of such an adaptive algorithm to CNNs and discuss on its development in real-world architectures.

2 UNSUPERVISED LEARNING AND THE OPTIMAL REPRESENTATION OF IMAGES

Visual items composing natural images are often sparse, such that knowing a model for the generation of images, the brain may use this property to reconstruct images using only a few of these items. In the context of the representation of natural images[1] $\mathbf{y} = (\mathbf{y}_k)_{k=1}^K \in \mathbb{R}^{K\times M}$ represented in a matrix as a set of K vector samples (herein, we will use a batch size of $K = 256$) as images raveled along M pixels (each $\mathbf{y}_{k,j} \in \mathbb{R}$ are the corresponding luminance values), let us assume the generic Generative Linear Model, such that for any sample k the image was generated as $\mathbf{y}_k = \Phi^T \mathbf{a}_k + \epsilon$, where by definition, the coefficients are denoted by $\mathbf{a}_k = (\mathbf{a}_{k,i})_{i=1}^N \in \mathbb{R}^N$ and the dictionary by $\Phi \in \mathbb{R}^{N\times M}$. Finally, $\epsilon \in \mathbb{R}^M$ is a Gaussian iid noise which is Normal without loss of generality by scaling the norm of the dictionary's rows. Knowing this model, unsupervised learning aims at finding the least surprising causes (the parameters $\hat{\mathbf{a}}_k$ and Φ) for the data $\mathbf{y}_k$. In particular, the cost may be formalized in a probabilistic terms as (Olshausen & Field, 1997):

$$F \approx < -\log[p(\mathbf{y}_k|\hat{\mathbf{a}}_k, \Phi)p_\Phi(\hat{\mathbf{a}}_k)] >_{k=1\dots K} = < \frac{1}{2}\|\mathbf{y}_k - \Phi\hat{\mathbf{a}}_k\|_2^2 - \log p_\Phi(\hat{\mathbf{a}}_k) >_{k=1\dots K} \quad (1)$$

Such hypothesis allows to retrieve the cost that is optimized in most of existing models of unsupervised learning. Explicitly, the representation is optimized by minimizing a cost defined on prior assumptions on representation's sparseness, that is on $\log p_\Phi(\mathbf{a}_k)$. For instance, learning is accomplished in SPARSENET (Olshausen & Field, 1997) by defining a sparse prior probability distribution function for each coefficients in the factorial form $\log p_\Phi(\mathbf{a}_k) \sim -\beta \sum_i \log(1 + \frac{a_i^2}{\sigma^2})$ where β corresponds to the steepness of the prior and σ to its scaling (see Figure 13.2 from (Olshausen, 2002)). Then, knowing this sparse solution, learning is defined as slowly changing the dictionary using Hebbian learning.

[1] We use image patches drawn from large images of outdoor scenes, as provided in the *kodakdb* database which is available in the code's repository.

Indeed, to compute the partial derivate of F with respect to Φ, we have $\forall i$:

$$\frac{\partial}{\partial \Phi_i} F =< \frac{1}{2} \frac{\partial}{\partial \Phi_i} [(\mathbf{y}_k - \Phi^T \hat{\mathbf{a}}_k)^T (\mathbf{y}_k - \Phi^T \hat{\mathbf{a}}_k)] >_{k=1 \dots K} =< \hat{\mathbf{a}}_k (\mathbf{y}_k - \Phi^T \hat{\mathbf{a}}_k) >_{k=1 \dots K} . \quad (2)$$

This allows to define unsupervised learning as the gradient descent using this equation. Similarly to Eq. 17 in (Olshausen & Field, 1997) or to Eq. 2 in (Smith & Lewicki, 2006), the relation is a linear "Hebbian" rule (Hebb, 1949) since it enhances the weight of neurons proportionally to the activity (coefficients) between pre- and post-synaptic neurons. Note that there is no learning for non-activated coefficients and also that we used a (classical) scheduling of the learning rate and a proper initialization of the weights (see Annex 2.5 & 2.6). The novelty of this formulation compared to other linear Hebbian learning rule such as (Oja, 1982) is to take advantage of the sparse (non-linear) representation, hence the name Sparse Hebbian Learning (SHL). In general, the parameterization of the prior in Eq. 1 has major impacts on results of the sparse coding and thus on the emergence of edge-like receptive fields and requires proper tuning. For instance, a L2-norm penalty term (that is, a Gaussian prior on the coefficients) corresponds to Tikhonov regularization (Tikhonov, 1977) and a L1-norm term (that is, an exponential prior for the coefficients) corresponds to the convex cost which is optimized by least-angle regression (LARS) (Efron et al., 2004) or FISTA (Beck & Teboulle, 2009).

2.1 Algorithm: Sparse Coding with a control mechanism for the selection of atoms

Concerning the choice of a proper prior distribution, the spiking nature of neural information demonstrates that the transition from an inactive to an active state is far more significant at the coding time scale than smooth changes of the firing rate. This is for instance perfectly illustrated by the binary nature of the neural code in the auditory cortex of rats (DeWeese et al., 2003). Binary codes also emerge as optimal neural codes for rapid signal transmission (Bethge et al., 2003). This is also relevant for neuromorphic systems which transmit discrete, asynchronous events (such as a network packet). With a binary event-based code, the cost is only incremented when a new neuron gets active, regardless to its (analog) value. Stating that an active neuron carries a bounded amount of information of λ bits, an upper bound for the representation cost of neural activity on the receiver end is proportional to the count of active neurons, that is, to the ℓ_0 pseudo-norm $\|\mathbf{a}_k\|_0 = |\{i, \mathbf{a}_{k,i} \neq 0\}|$:

$$F \approx < \frac{1}{2} \|\mathbf{y}_k - \Phi \mathbf{a}_k\|_2^2 + \lambda \|\mathbf{a}_k\|_0 >_{k=1 \dots K} \quad (3)$$

This cost is similar with information criteria such as the Akaike Information Criteria (Akaike, 1974) or distortion rate (Mallat, 1998, p. 571). For $\lambda = \log_2 N$, it gives the total information (in bits) to code for the residual (using entropic coding) and the list of spikes' addresses. In general, the high inter-connectivity of neurons (on average approximately 10000 synapses per neurons) justifies such an informational perspective with respect to the analog quantization of information in the point-to-point transfer of information between neurons. However, Eq. 3 defines a harder cost to optimize (in comparison to convex formulations in Equation 1 for instance) since the hard ℓ_0 pseudo-norm sparseness leads to a non-convex optimization problem which is *NP-complete* with respect to the dimension M of the dictionary (Mallat, 1998, p. 418).

Still, there are many solutions to this optimization problem and here, we will use a generalized version of the Matching Pursuit (MP) algorithm (Mallat, 1998, p. 422). A crucial aspect of this algorithm is the $\arg\max$ function as it produces at each step a competition among N neurons. For this reason, we will introduce a mechanism to tune this competition. For any signal $\mathbf{y}_k$ drawn from the database, we get the coefficients $\mathbf{a}_k = S(\mathbf{y}_k; \Psi = \{\Phi, z, N_0\})$ thanks to Algorithm 1. The parameter $N_0 > 0$ controls the amount of sparsity that we impose to the coding. The novelty of this generalization of MP lies in the scalar functions $z = \{z_i\}_{i=1 \dots N}$ which control the competition for the best match across atoms. While an identical symmetric function is chosen in the original MP algorithm (that is, $\forall i, z_i(\mathbf{a}_k) = |\mathbf{a}_k|$), we will define these at a first attempt as the rescaled non-linear rectified linear unit (ReLU) with gain γ_i: $\forall i, z_i(\mathbf{a}_{k,i}) = \gamma_i * \mathbf{a}_{k,i} * \delta(\mathbf{a}_{k,i} > 0)$ where δ is Kronecker's indicator function.

We found as in (Rehn & Sommer, 2007) that by using an algorithm like Matching Pursuit (that is using the symmetric function or setting $\forall i, \gamma_i = 1$ as in (Mairal et al., 2014) for instance), the Sparse Hebbian Learning algorithm could provide results similar to SparseNet. An advantage is

Algorithm 1 Generalized Matching Pursuit: $\mathbf{a}_k = S(\mathbf{y}_k; \Psi = \{\Phi, z, N_0\})$

1: set the sparse vector $\mathbf{a}_k$ to zero,
2: initialize $\bar{\mathbf{a}_{k}}_i =< \mathbf{y}_k, \Phi_i >$ for all i
3: **while** $\|\mathbf{a}_k\|_0 < N_0$ **do**:
4: select the best match: $i^* = \arg\max_i [z_i(\bar{\mathbf{a}_{k}}_i)]$
5: update the sparse coefficient: $\mathbf{a}_{k,i^*} = \mathbf{a}_{k,i^*} + \bar{\mathbf{a}_{k,i^*}}$,
6: update residual coefficients: $\forall i, \bar{\mathbf{a}_{k,i}} \leftarrow \bar{\mathbf{a}_{k,i}} - \mathbf{a}_{k,i^*} < \Phi_{i^*}, \Phi_i >$.

the non-parametric assumption on the prior based on this more generic ℓ_0 pseudo-norm sparseness. However, we observed that this class of algorithms could lead to solutions corresponding to a local minimum of the full objective function: Some solutions seem as efficient as others for representing the signal but do not represent edge-like features homogeneously (Figure 1-A, `None`). Moreover, using other sparse coding algorithms which are implemented in the `sklearn` library, we compared the convergence of the learning with different sparse coding algorithms. In particular, we compared the learning as implemented with matching pursuit to that with orthogonal matching pursuit (OMP) (Pati et al., 1993), LARS or FISTA (see Annex 2.4). For all these sparse coding algorithms, during the early learning step, some cells may learn "faster" than others. In particular, these cells have more peaked distributions of their activity and tend to be selected more often. There is the need for a homeostasis mechanism that will ensure convergence of learning. The goal of this work is to study the specific role of homeostasis in learning sparse representations and to propose a homeostasis mechanism based on the functions z_i which optimizes the learning of an efficient representation.

2.2 Algorithm: Histogram Equalization Homeostasis

Knowing a dictionary and a sparse coding algorithm, we may transform any data sample $\mathbf{y}_k$ into a set of sparse coefficients using the above algorithm: $\mathbf{a}_k = S(\mathbf{y}_k; \Psi = \{\Phi, z, N_0\})$ (see Algorithm 1). In particular, at any step *during* learning, dictionaries may not have been homogeneously learned and may exhibit different distributions. However, this would not be taken into account in the original cost (see Eq. 3) as we assumed as in Olshausen & Field (1997) that $p_\Phi(\hat{\mathbf{a}_k})$ is factorized, that is, that the components of the sparse vector are independent. As a consequence, we may use a deviation to this hypothesis as an additional component to the cost:

$$F \approx < \frac{1}{2}\|\mathbf{y}_k - \Phi\mathbf{a}_k\|_2^2 + \lambda\|\mathbf{a}_k\|_0 + \texttt{MI}(\mathbf{a}_k) >_{k=1\ldots K} \tag{4}$$

Where we used the mutual information `MI` as a proxy to measure the dependence between the components of the sparse vector. Indeed, as information is coded in the address of neurons, information transfer as computed through Shannon entropy, is optimized when the activity within the neural population is uniformly balanced, that is when each neuron is *a priori* selected with the same probability. In particular, a necessary (yet not sufficient) condition for minimizing this cost is that the prior probability of selecting coefficients are identical $\forall (i,j), q_\Psi(\mathbf{a}_{k,i}) = q_\Psi(\mathbf{a}_{k,j})$ to ensure the optimality of the choice of the ℓ_0 pseudo-norm and compare it to the representation in the primary visual cortex. As we have seen, we may use different transformation functions z to influence the choice of coefficients such that we may use these functions to optimize the objective cost defined by Eq. 4.

To achieve this uniformity, we may define an homeostatic gain control mechanism based on histogram equalization, that is, by transforming coefficients in terms of z-scores, that is, by setting $\forall i, z_i(\cdot) = P(\cdot > a_i)$. Such a transform is similar to the inverse transform sampling which is used to optimize representation in auto-encoders (Doersch, 2016) and can be considered as a non-parametric extension of the "re-normalization trick" used in variational auto-encoders (Kingma & Welling, 2013). Moreover, it has been found that such an adaptation mechanism is observed in the response of the retina to various contrast distributions (Laughlin, 1981). However, an important point to note is that this joint optimization problem between coding and homeostasis is circular as we can not access the true posterior $p_\Phi(\mathbf{a})$: Indeed, the coefficients depend on non-linear coefficients through $\mathbf{a}_k = S(\mathbf{y}_k; \Psi = \{\Phi, z_i, N_0\})$, while the non-linear functions depend on the (cumulative) distribution of the coefficients. We will make the assumption that such a problem can be solved iteratively by slowly learning the non-linear functions. Starting with an initial set of non-linear functions as in `None`, we will derive an approximation for the sparse coefficients. Then, the function z_i for each

coefficient of the sparse vector is calculated using an iterative moving average scheme (parameterized by time constant $1/\eta_h$) to smooth its evolution during learning. At the coding level, this z-score function is incorporated in the matching step of the matching pursuit algorithm, to modulate the choice of the most probable as that with the maximal z-score: $i^* = \arg\max_i z_i(a_i)$ (see Algorithm 1). The rest of the algorithm Sparse Hebbian Learning algorithm is left unchanged (see Algorithm 2). We will coin this algorithm variant as Histogram Equalization Homeostasis (`HEH`).

Furthermore, as we adapt the dictionaries progressively during Sparse Hebbian Learning, we may incorporate this `HEH` homeostasis during learning by choosing an appropriate learning rate η_h. To recapitulate the different choices we made from the learning to the coding and the homeostasis, the unsupervised learning can be summarized using the following steps:

Algorithm 2 Homeostatic Unsupervised Learning of Kernels: $\Phi = H(\mathbf{y}; \eta, \eta_h, N_0)$

Initialize the point non-linear gain functions z_i to similar cumulative distribution functions,
initialize atoms Φ_i to random points on the K-unit sphere,
for T epochs **do**:
  draw a new batch $\mathbf{y}$ from the database of natural images,
  for each data point $\mathbf{y}_k$ **do**:
    compute the sparse representation vector $\mathbf{a} = S(\mathbf{y}_k; \Psi = \{\Phi, z, N_0\})$ (see Algorithm 1),
    modify dictionary: $\forall i, \Phi_i \leftarrow \Phi_i + \eta \cdot \mathbf{a}_i \cdot (\mathbf{y}_k - \Phi\mathbf{a})$,
    normalize dictionary: $\forall i, \Phi_i \leftarrow \Phi_i / \|\Phi_i\|$,
    update homeostasis functions: $\forall i, z_i(\cdot) \leftarrow (1 - \eta_h) \cdot z_i(\cdot) + \eta_h \cdot \delta(a_i \leq \cdot)$.

We compared qualitatively the set Φ of receptive filters generated with different homeostasis algorithms (see Fig. 1-A). A more quantitative study of the coding is shown by comparing the decrease of the cost as a function of the iteration step (see Fig. 1-B). This demonstrate that forcing the learning activity to be uniformly spread among all receptive fields results in a faster convergence of the representation error as represented by the decrease of the cost F.

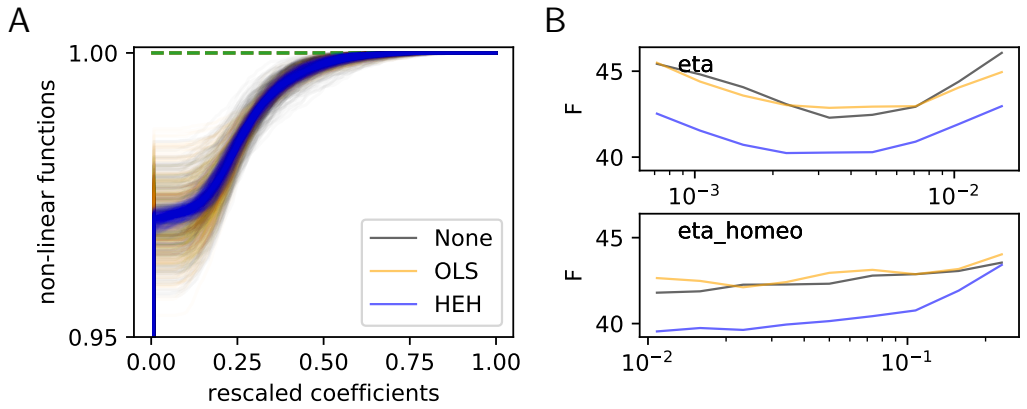

Figure 2: **Histogram Equalization Homeostasis and its role in unsupervised learning**: (A) Non-linear homeostatic functions $z_i, \forall i$ learned using Hebbian learning. These functions were computed for different homeostatic strategies (`None`, `OLS` or `HEH`) but only used in `HEH`. Note that for our choice of $N_0 = 13$, all cumulative functions start around $1 - N_0/N \approx .970$. At convergence of `HEH`, the probability of choosing any filter is equiprobable, while the distribution of coefficients is more variable for `None` and `OLS`. As a consequence, the distortion between the distributions of sparse coefficients is minimal for `HEH`, a property which is essential for the optimal representation of signals in distributed networks such as the brain. (B) Effect of learning rate η (`eta`) and homeostatic learning rate η_h (`eta_homeo`) on the final cost as computed for the same learning algorithms but with different homeostatic strategies (`None`, `OLS` or `HEH`). Parameters were explored around a default value, on a logarithmic scale and over 4 octaves. This shows that `HEH` is robust across a wide range of parameters.

2.3 RESULTS: FAST UNSUPERVISED LEARNING USING HOMEOSTASIS

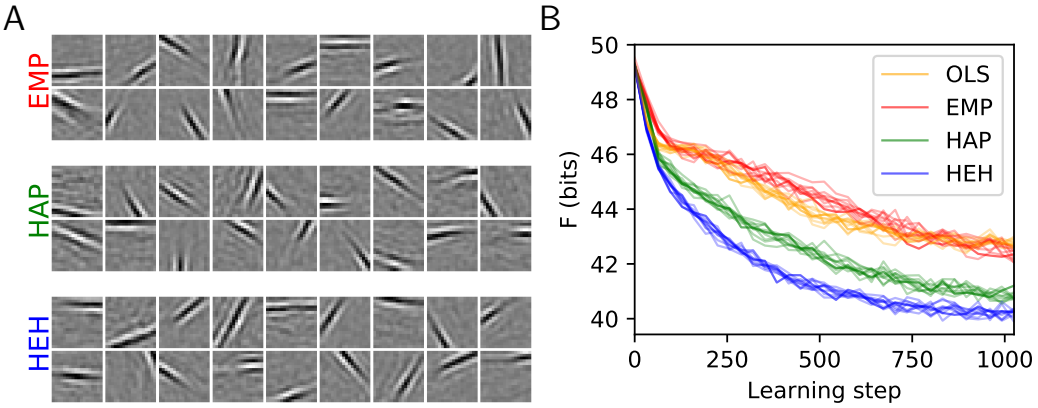

Figure 3: **Homeostasis on Activation Probability (`HAP`) and a quantitative evaluation of homeostatic strategies**: (A) 18 from the 441 dictionaries learned for the two heuristics `EMP` and `HAP` and compared to the optimal homeostasis (see Figure 1-A, `HEH`). (B) Comparison of the cost F during learning and cross-validated over 10 runs: The convergence of `OLS` is similar to `EMP`. The simpler `HAP` heuristics gets closer to the more demanding `HEH` homeostatic rule, demonstrating that this heuristic is a good compromise for fast unsupervised learning.

We have shown above that we can find an exact solution to the problem of homeostasis during Sparse Hebbian Learning. However, this solution has several drawbacks. First, it is computationally intensive on a conventional computer as it necessitates to store each z_i function to store the cumulative distribution of each coefficient. More importantly, it seems that biological neurons seem to rather use a simple gain control mechanism. This can be implemented by modifying the gain γ_i of the slope of the ReLu function to operate a gradient descent on the cost based on the distribution of each coefficients. Such strategy can be included in the SHL algorithm by replacing line 8 in Algorithm 2. For instance, the strategy of (Olshausen & Field, 1997) assumes a cost on the difference between the observed variance of coefficients V_i as computed over a set of samples compared to a desired value σ_g^2 (and assuming a multiplicative noise parameterized by α) :

$$V_i \leftarrow (1-\eta_h) \cdot V_i + \eta_h \cdot 1/K \sum_{k=1 \cdots K} a_{i,k}^2 \text{ and } \gamma_i \leftarrow \gamma_i \cdot \left(\frac{V_i}{\sigma_g^2}\right)^{\alpha} \tag{5}$$

This is similar to the mechanisms of gain normalization proposed by Schwartz & Simoncelli (2001) and which were recently shown to provide efficient coding mechanisms by Simoncelli (2017). However, compared to these methods which manipulate the gain of dictionaries based on the energy of coefficients, we propose to rather use a methodology based on the probability of activation. Indeed, the main distortion that occurs during learning is on higher statistical moments rather than variance, for instance when an atom is winning more at the earlier iterations, its pdf will typically be more kurtotic than a filter that has learned less.

Recently, such an approach was proposed by Sandin & Martin-del Campo (2017). Based on the same observations, the authors propose to optimize the coding during learning by modulating the gain of each dictionary element based on the recent activation history. They base their Equalitarian Matching Pursuit (`EMP`) algorithm on a heuristics which cancels the activation of any filter that was more often activated than a given threshold probability (parameterized by $1 + \alpha_h$). In our setting, we may compute a similar algorithm using an evaluation of probability of activation followed by binary gates:

$$p_i \leftarrow (1-\eta_h) \cdot p_i + \eta_h \cdot 1/K \sum_{k=1 \cdots K} \delta(a_{i,k} > 0) \text{ and } \gamma_i = \delta(p_i < N_0/N * (1+\alpha_h)) \tag{6}$$

Interestingly, they reported that such a simple heuristic could improve the learning, deriving a similar result as we have shown in Figure 1 and Figure 2. Again, such strategy can be included in Algorithm 2.

Similarly, we may derive an approximate homeostasis algorithm based on the current activation probability but using a gradient descent approach on gain modulation. Ideally, this corresponds to finding γ_i such that we minimize the entropy $\sum_{i=1\cdots N} p_i \log p_i$. However, the sparse coding function $S(\mathbf{y}_k; \Psi = \{\Phi, z, N_0\})$ is not derivable. One possible heuristic is then to derivate the change of modulation gain that would be necessary to achieve an equiprobable probability, that is when $\forall i, p_i = p_0 \stackrel{\text{def.}}{=} N_0/N$:

$$p_i \leftarrow (1-\eta_h) \cdot p_i + \eta_h \cdot 1/K \sum_{k=1\cdots K} \delta(a_{i,k} > 0) \text{ and } \gamma_i = \exp(-(p_i - p_0)/\alpha_h) \tag{7}$$

We will coin this variant of the algorithm Homeostasis on Activation Probability (`HAP`). Following these derivations, we quantitatively compared `OLS`, `EMP` and `HAP` to `HEH` (see Figure 3). This shows that while `EMP` slightly outperforms `OLS` (which itself is more efficient than `None`, see Figure 2-B), `HAP` proves to be closer to the optimal solution given by `HEH`. In particular, we replicated in `HAP` the result of Sandin & Martin-del Campo (2017) that while homeostasis was essential in improving unsupervised learning, the coding algorithm (MP versus OMP) mattered relatively little (see Annex 2.4). Also, we verified the dependence of this efficiency with respect to different hyperparameters (as we did in Figure 2-B). These quantitative results show that the `HEH` algorithm could be replaced by a simpler and more rapid heuristic, `HAP`, which is based on activation probability. This would generate similar efficiency for the coding of patches from natural images.

3 Discussion and conclusion

One core advantage of sparse representations is the efficient coding of complex signals using compact codes. Inputs are thus represented as combination of few elements drawn from a large dictionary of atoms. As a consequence, a common design for unsupervised learning rules relies on a gradient descent over a cost measuring representation quality with respect to sparseness. This constraint introduces a competition between atoms. In the context of the efficient processing of natural images, we proposed here that such strategies can be optimized by including a proper homeostatic regulation enforcing a fair competition between the elements of the dictionary. We implemented this rule by introducing a non-linear gain normalization similar to what is observed in biological neural networks. We validated this theoretical insight by challenging this adaptive unsupervised learning algorithm with alternate heuristics for homeostasis. Simulations show that at convergence, while the coding accuracy did not vary much, including homeostasis changed qualitatively the learned features. In particular, homeostasis results in a more homogeneous set of orientation selective filters, which is closer to what is found in the visual cortex of mammals (Ringach, 2002; Rehn & Sommer, 2007; Loxley, 2017). To further validate these results, we quantitatively compared the efficiency of the different variants of the algorithms, both at the level of homeostasis (homeostatic learning rate, parameters of the heuristics), but also to the coding (by changing M, N or N_0) and to the learning (by changing the learning rate, the scheduling or M). This demonstrated that overall, this neuro-inspired homeostatic algorithm provided with the best compromise between efficiency and computational cost.

In summary, this biologically-inspired learning rule demonstrates that principles observed in neural computations can help improve real-life machine learning algorithms. Indeed, by developing this fast learning algorithm, we hope for its rapid application in artificial intelligence algorithms. This type of architecture is economical, efficient and fast. It makes it possible to be transferred to most deep learning algorithms. Along with this, we hope that this new type of rapid unsupervised learning algorithm can provide a normative theory for the coding of information in low-level sensory processing, whether it is visual or auditory, for example. Moreover, by its nature, this algorithm can easily be extended to convolutional networks such as those used in deep learning neural networks. This extension is possible by extending the filter dictionary by the hypothesis of invariances to the translation of representations. Our results on different databases show the stable and rapid emergence of characteristic filters on these different bases (see Figure 4 and Annex 3.1). This result shows a probable prospect of extending this representation and for which we hope to obtain classification results superior to the algorithms existing in the state-of-the-art. As such, empirical evaluations of the proposed algorithms should be extended. For instance, it would be very useful to test for image classification results on standard benchmark datasets.

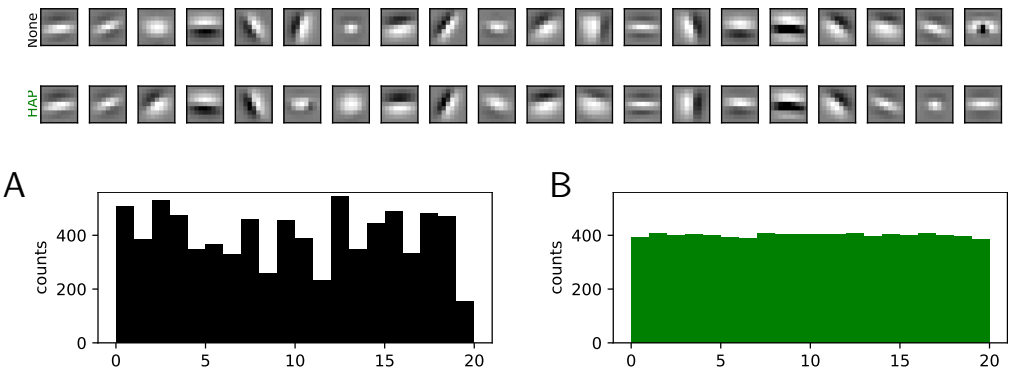

Figure 4: **Extension to Convolutional Neural Networks (CNNs)**: We extend the HAP algorithm to a single layered CNN with 20 kernels and using the ATT face database. We show here the kernels learned without (`None`, top row) and with (`HAP`, bottom row) homeostasis (note that we used the same initial conditions). As for the simpler case, we observe a heterogeneity of activation counts without homeostasis, that is, in the case which simply normalizes the energy of kernels (see (A)). With homeostasis, we observe the convergence of the activation probability for the different kernels (see (B)). This demonstrates that this heuristic extends well to a CNN architecture.

REFERENCES

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–23, 1974.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Matthias Bethge, David Rotermund, and Klaus Pawelzik. Second order phase transition in neural rate coding: Binary encoding is optimal for rapid signal transmission. *Physical Review Letters*, 90 (8):088104, 2003.

Carlos SN Brito and Wulfram Gerstner. Nonlinear hebbian learning as a unifying principle in receptive field formation. *PLoS computational biology*, 12(9):e1005070, 2016.

Matteo Carandini and David J Dj Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(November):1–12, jan 2012. doi: 10.1038/nrn3136. URL `http://discovery.ucl.ac.uk/1332718/http://www.ncbi.nlm.nih.gov/pubmed/22108672http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3273486`.

Michael R. DeWeese, Michael Wehr, and Anthony M. Zador. Binary spiking in auditory cortex. *Journal of Neuroscience*, 23(21):7940–7949, August 2003. URL `http://www.jneurosci.org/content/23/21/7940.abstract`.

Carl Doersch. Tutorial on variational autoencoders. *arXiv:1606.05908 [cs, stat]*, June 2016. URL `http://arxiv.org/abs/1606.05908`.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

David Hansel and Carl van Vreeswijk. The mechanism of orientation selectivity in primary visual cortex without a functional map. *Journal of Neuroscience*, 32(12):4049–4064, 2012.

Donald O. Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, New York, 1949.

D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195(1):215–243, March 1968. ISSN 0022-3751. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1557912/.

John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9 (3):90–95, May 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.55. URL http://dx.doi.org/10.1109/MCSE.2007.55.

D. P Kingma and M. Welling. Auto-Encoding Variational Bayes. *ArXiv e-prints*, December 2013.

S. Laughlin. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung. Section C: Biosciences*, 36(9-10):910–912, 1981. ISSN 0341-0382. URL http://view.ncbi.nlm.nih.gov/pubmed/7303823.

P. N. Loxley. The two-dimensional gabor function adapted to natural image statistics: A model of simple-cell receptive fields and sparse structure in images. *Neural Computation*, 29(10):2769–2799, oct 2017. ISSN 0899-7667. doi: 10.1162/neco_a_00997. URL http://www.ncbi.nlm.nih.gov/pubmed/28777727http://www.mitpressjournals.org/doi/abs/10.1162/neco_a_00997.

Julien Mairal, Francis Bach, Jean Ponce, et al. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

Alireza Makhzani and Brendan J. Frey. k-sparse autoencoders. *CoRR*, abs/1312.5663, 2013. URL http://arxiv.org/abs/1312.5663.

Stéphane Mallat. *A wavelet tour of signal processing*. Academic Press, second edition, 1998.

Eve Marder and Jean-Marc Goaillard. Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience*, 7(7):563, 2006.

Erkki Oja. A Simplified Neuron Model as a Principal Component Analyzer. *Journal of Mathematical biology*, 15:267–73, 1982.

T. E. Oliphant. Python for scientific computing. *Computing in Science and Engineering*, 9(3):10–20, May 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.58. URL http://dx.doi.org/10.1109/MCSE.2007.58.

Bruno A. Olshausen. Sparse codes and spikes. In Rajesh P. N. Rao, Bruno A. Olshausen, and Michael S. Lewicki (eds.), *Probabilistic Models of the Brain: Perception and Neural Function*, chapter Sparse Codes and Spikes, pp. 257–72. MIT Press, 2002.

Bruno A. Olshausen and David J. Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(6583):333–9, June 1996. ISSN 0028-0836. doi: 10.1038/381607a0. URL http://dx.doi.org/10.1038/381607a0.

Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, December 1997. ISSN 0042-6989. doi: 10.1016/S0042-6989(97)00169-7. URL http://dx.doi.org/10.1016/S0042-6989(97)00169-7.

Vardan Papyan, Yaniv Romano, and Michael Elad. Convolutional neural networks analyzed via convolutional sparse coding. *stat*, 1050:27, 2016.

Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pp. 40–44. IEEE, 1993.

Laurent U. Perrinet. Role of homeostasis in learning sparse representations. *Neural Computation*, 22(7):1812–1836, July 2010. ISSN 1530-888X. doi: 10.1162/neco.2010.05-08-795. URL http://invibe.net/LaurentPerrinet/Publications/Perrinet10shl.

Rajesh Rao and Dana Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, January 1999. ISSN 1097-6256. doi: 10.1038/4580. URL http://dx.doi.org/10.1038/4580.

Martin Rehn and Friedrich T. Sommer. A model that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *Journal of Computational Neuroscience*, 22 (2):135–46, 2007.

Dario L Ringach. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of neurophysiology*, 88(1):455–463, 2002.

Fredrik Sandin and Sergio Martin-del Campo. Dictionary learning with equiprobable matching pursuit. *arXiv preprint arXiv:1611.09333*, pp. 557–564, may 2017. doi: 10.1109/IJCNN.2017.7965902. URL http://ieeexplore.ieee.org/document/7965902/.

Odelia Schwartz and Eero P. Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–25, 2001.

Eero P Simoncelli. Optimal representations and efficient coding. *Annual Review of Vision Science*, 3 (1), 2017.

Evan C. Smith and Michael S. Lewicki. Efficient auditory coding. *Nature*, 439(7079):978–982, February 2006. ISSN 1476-4679. doi: 10.1038/nature04485. URL http://dx.doi.org/10.1038/nature04485.

Jeremias Sulam, Vardan Papyan, Yaniv Romano, and Michael Elad. Multi-layer convolutional sparse modeling: Pursuit and dictionary learning. *arXiv preprint arXiv:1708.08705*, 2017.

Andrei N. Tikhonov. *Solutions of Ill-Posed Problems*. Winston & Sons, Washington, 1977. ISBN 0470991240. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0470991240.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103. ACM, 2008.