Autoencoders explained

Lorenzo Bozzoni

March 2024

Contents

1 Autoencoders

Autoencoders are simple learning circuits which aim to transform inputs into outputs with the least possible amount of distortion. To derive a fairly general framework, an $n/p/n$ autoencoder is defined by a t-uple $n, p, m, \mathbb{F}, \mathbb{G}, \mathcal{A}, \mathcal{B}, \mathcal{X}, \mathcal{Y}, \Delta$, where:

- $\mathbb{F}$ and $\mathbb{G}$ are sets
- n and p are positive integers
- $\mathcal{A}$ is a class of functions from $\mathbb{G}^p$ to $\mathbb{F}^n$
- $\mathcal{B}$ is a class of functions from $\mathbb{F}^n$ to $\mathbb{G}^p$
- $\mathcal{X} = x_1, \ldots, x_m$ is a set of m (training) vectors in $\mathbb{F}^n$. When the external targets are present, we let $\mathcal{Y} = y_1, \ldots, y_m$ denote the corresponding set of m target vectors in $\mathbb{F}^n$
- Δ is a dissimilarity or distortion function defined over $\mathbb{F}^n$

For any $A \in \mathcal{A}$ and $B \in \mathcal{B}$, the autoencoder transforms an input vector $x \in \mathbb{F}^n$ into an output vector $A \circ B(x) \in \mathbb{F}^n$. The corresponding **autoencoder problem** is to find $A \in \mathcal{A}$ and $B \in \mathcal{B}$ that minimize the overall distortion function:

$$\min E(A, B) = \min_{A,B} \sum_{t=1}^{m} E(A, B) = \min_{A,B} \sum_{t=1}^{m} \Delta(x_t, A \circ B(x_t)) \tag{1}$$

In the non auto-associative case, when external targets y_t are provided, the minimization problem becomes:

$$\min E(A, B) = \min_{A,B} \sum_{t=1}^{m} E(A, B) = \min_{A,B} \sum_{t=1}^{m} \Delta(y_t, A \circ B(x_t)) \tag{2}$$

It is important to notice that if $p < n$ corresponds to a compression or feature extraction, while $p > n$ corresponds to a decompression.

2 Linear Autoencoders

We consider the problem of learning from examples in layered linear feed-forward neural networks using optimization methods, such as back propagation, with respect to the usual quadratic error function E of the connection weights.

We assume to have N samples and N lables so for each x_n input vector corresponds the y_n label. The classical quadratic error function is defined as:

$$E = \sum_n \|y_n - F(x_n)\|^2$$

where F is the current function implemented by the network. We defined also the **covariance matrices**:

$$\Sigma_{XX} = \sum_n x_n x_n^\intercal$$

$$\Sigma_{XY} = \sum_n x_n y_n^\intercal$$

$$\Sigma_{YY} = \sum_n y_n y_n^\intercal$$

$$\Sigma_{YX} = \sum_n y_n x_n^\intercal$$

Where these quantities are defined.

2.1 Useful mathematical concepts

For any matrices P, Q, R we have $tr(PQR) = tr(RPQ) = tr(QRP)$ provided that these quantities are defined. Thus in particolar if P is **idempotent**, that is, $P^2 = P$, then:

$$tr(PQP) = tr(PPQ) = tr(P^2Q) = tr(PQ) \tag{a}$$

If U is orthogonal, that is $U^\intercal U = I$, then:

$$tr(UQU^\intercal) = tr(U^\intercal UQ) = tr(Q) \tag{b}$$

The **Kronecker product** $P \otimes Q$ of any two matrices P and Q is the matrix obtained from the matrix P by replacing each entry p_{ij} of P with the matrix $p_{ij}Q$. Which means that:

$$P : m \times n \text{ and } Q : r \times q \implies P \otimes Q = \begin{bmatrix} p_{11}Q & \dots & a_{1n}Q \\ \vdots & \ddots & \vdots \\ p_{m1}Q & \dots & p_{mn}Q \end{bmatrix} \text{ of shape } rm \times qn$$

The **vec operation** transforms a matrix into a column vector by stacking the columns of the matrix one underneath the other. Indeed, if P is any $m \times n$ matrix and p_j is the j-th column, then $vec(P)$ is the $mn \times 1$ vector $vec(P) = [p_1^\intercal, \dots, p_n^\intercal]^\intercal$.

We have that:

$$tr(PQ^\intercal) = (vec(P))^\intercal vec(Q) \tag{c}$$

$$vec(PQR^\intercal) = (R \otimes P)vec(Q) \tag{d}$$

$$(P \otimes Q)(R \otimes S) = PR \otimes QS \tag{e}$$

$$(P \otimes Q)^{-1} = P^{-1} \otimes Q^{-1} \tag{f}$$

$$(P \otimes Q)^\intercal = P^\intercal \otimes Q^\intercal \tag{g}$$

whenever these quantities are defined. Also: if P and Q are symmetric and positive semidefinite (resp. definite) then $P \otimes Q$ is symmetric and positive semidefinite (resp. positive definite) (h).

Finally, let us introduce the input data matrix $X = [x_1, \dots, x_N]$ and the output data matrix $Y = [y_1, \dots, y_N]$. It is easily seen that $XX^\intercal = \Sigma_{XX}$, $XY^\intercal = \Sigma_{XY}$, $YY^\intercal = \Sigma_{YY}$, $YX^\intercal = \Sigma_{YX}$ and $E(A, B) = \|vec(Y - ABX)\|^2$. In the proof of facts 1 and 2, we shall use the following well known lemma.

Lemma: the quadratic function:

$$F(z) = \|c - Mz\|^2 = c^\intercal c - 2c^\intercal Mz + z^\intercal M^\intercal Mz$$

is convex. A point z corresponds to a global minimum of F if and only if it satisfies the equation $\nabla F = 0$, or equivalently $M^\intercal Mz = M^\intercal c$. If in addition $M^\intercal M$ is positive definite, then F is strictly convex and the unique minimum of F is attained for $z = (M^\intercal M)^{-1}M^\intercal c$.

2.2 Fact 1

For any fixed $n \times p$ matrix A the function $E(A, B)$ is convex in the coefficients of B and attains its minimum for any B satisfying the equation

$$A^{\intercal} A B \Sigma_{XX} = A^{\intercal} \Sigma_{YX} \tag{3}$$

If Σ_{XX} is invertible and A is full rank p, then E is strictly convex and has unique minimum reached when:

$$B = \hat{B}(A) = (A^{\intercal} A)^{-1} A^{\intercal} \Sigma_{YX} \Sigma_{XX}^{-1} \tag{3}$$

In the auto-associative case, (3) becomes

$$B = \hat{B}(A) = (A^{\intercal} A)^{-1} A^{\intercal} \tag{3'}$$

Since $Y = X$ so $\Sigma_{YX} \Sigma_{XX}^{-1} = I$.

2.2.1 Proof of fact 1

Proof. For fixed A, use (d) to write:

$$vec(Y - ABX) = vec(Y) - vec(ABX) = vec(Y) - (X^{\intercal} \otimes A) vec(B)$$

and thus:

$$E(A, B) = \|vec(Y) - (X^{\intercal} \otimes A) vec(B)\|^2$$

By the above lemma, E is convex in the coefficients of B and B corresponds to a global minimum if and only if

$$(X^{\intercal} \otimes A)^{\intercal} (X^{\intercal} \otimes A) vec(B) = (X^{\intercal} \otimes A) vec(Y)$$

Now on one hand:

$$\begin{aligned} (X^{\intercal} \otimes A)^{\intercal} (X^{\intercal} \otimes A) vec(B) &= (X^{\intercal} \otimes A) vec(B) \\ &= (X X^{\intercal} \otimes A^{\intercal} A) vec(B) \\ &= (\Sigma_{XX} \otimes A^{\intercal} A) vec(B) \\ &= vec(A^{\intercal} A B \Sigma_{XX}) \end{aligned}$$

On the other hand:

$$\begin{aligned} (X^{\intercal} \otimes A)^{\intercal} vec(Y) &= (X \otimes A^{\intercal}) vec(Y) \\ &= vec(A^{\intercal} Y X^{\intercal}) \\ &= vec(A^{\intercal} \Sigma_{YX}) \end{aligned}$$

Therefore:

$$A^{\intercal} A B \Sigma_{XX} = A^{\intercal} \Sigma_{YX}$$

which is (2). If A is full rank, $A^{\intercal} A$ is symmetric and positive definite. As a covariance matrix, Σ_{XX} is symmetric and positive semidefinite; if, in addition,

Σ_{XX} is invertible, then Σ_{XX} is also positive definite. Because of (h), $(X^\intercal \otimes A)^\intercal (X^\intercal \otimes A) = \Sigma_{XX} \otimes A^\intercal A$ is also symmetric and positive definite. Applying the above lemma, we conclude that if Σ_{XX} is invertible and A is a fixed full rank matrix, then E is strictly convex in the coefficients of B and attains its unique minimum at the unique solution $B = \hat{B}(A) = (A^\intercal A)^{-1} A^\intercal \Sigma_{YX} \Sigma_{XX}^{-1}$ of (2), which is (3). In the auto-associative case, $x_n = y_n$. Therefore $\Sigma_{XX} = \Sigma_{YX} = \Sigma_{YY} = \Sigma_{XY}$ and the above expression simplifies to (3'). □

2.3 Fact 3

Assume that Σ_{XX} is invertible. If two matrices A and B define a critical point of E (i.e. a point where $\frac{\partial E}{\partial a_{ij}} = \frac{\partial E}{\partial b_{ij}} = 0$) then the global map $W = AB$ is of the form:

$$W = P_A \Sigma_{YX} \Sigma_{XX}^{-1} \tag{6}$$

with A satisfying

$$P_A \Sigma = P_A \Sigma P_A = \Sigma P_A \tag{7}$$

Where $\Sigma = \Sigma_{YX} = \Sigma_{XX}^{-1} = \Sigma_{XY}$. Recall also, that the matrix P_A is the matrix of the orthogonal projection onto the subspace spanned by the columns of A. In the auto-associative case, $\Sigma = \Sigma_{XX}$ and (6) and (7) become:

$$W = AB = P_A \tag{6'}$$

$$P_A \Sigma_{XX} = P_A \Sigma_{XX} P_A = \Sigma_{XX} P_A \tag{7'}$$

If A is full rank p, then A and B define a critical point of E if and only if A satisfies (7) and $B = \hat{B}(A)$, or equivalently if and only if A and W satisfy (6) and (7).

2.3.1 Proof of fact 3

Proof. Assume first that A and B define a critical point of E, with A full rank. Then from fact 1 we get $B = \hat{B}(A)$ and thus

$$W = AB = A(A^\intercal A)^{-1} A \Sigma_{YX} \Sigma_{XX}^{-1} = P_A \Sigma_{YX} \Sigma_{XX}^{-1}$$

Which is (6). Multiplication of (4) by $A^\intercal$ on the right yields

$$W \Sigma_{XX} W^\intercal = AB \Sigma_{XX} B^\intercal A^\intercal = \Sigma_{YX} B^\intercal A^\intercal = \Sigma_{YX} W^\intercal$$

Or

$$P_A \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XX} \Sigma_{XX}^{-1} \Sigma_{XY} P_A = \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} P_A$$

Or, equivalently $P_A \Sigma P_A = \Sigma P_A$. Since both Σ and P_A are symmetric, $P_A \Sigma P_A = \Sigma P_A$ is also symmetric and therefore $\Sigma P_A = (\Sigma P_A)^\intercal = P_A^\intercal \Sigma^\intercal = P_A \Sigma$, which is (7). Hence if A and B correspond to a critical point and A is full rank then (6) and (7) must hold and $B = \hat{B}(A)$.

Conversely, assume that A and W satisfy (6) and (7), with A full rank. Multiplying (6) by $(A^\intercal A)^{-1}A^\intercal$ on the left yields

$$B = (A^\intercal A)^{-1}A\Sigma_{YX}\Sigma_{XX}^{-1} = \hat{B}(A)$$

and (2) is satisfied. From $P_A\Sigma P_A = \Sigma P_A$ and using (6) we immediately get

$$AB\Sigma_{XX}B^\intercal A^\intercal = \Sigma_{YX}B^\intercal A^\intercal$$

and multiplication of both sides by $A(A^\intercal A)^{-1}$ on the right yields

$$AB\Sigma_{XX}B^\intercal = \Sigma_{YX}B^\intercal$$

which is (4). Thus A and B satisfy (2) and (4) and therefore they define a critical point of E. □

2.4 Fact 4

Assume that Σ is full-rank with n distinct eigenvalues $\lambda_1 > \cdots > \lambda_n$. If $\mathcal{I} = i_1, \ldots, i_p$ $(1 \leq i_1 < \cdots < i_p \leq n)$ is any ordered p-index set, let $U_\mathcal{I} = [u_{i_1}, \ldots, u_{i_p}]$ denote the matrix formed by the orthonormal eigenvectors of Σ associated with the eigenvalues $\lambda_{i_1}, \ldots, \lambda_{i_p}$. Then two full rank matrices A and B define a critical point of E if and only if there exist an ordered p-index set $\mathcal{I}$ and an invertible $p \times p$ matrix C such that:

$$A = U_\mathcal{I}C \tag{8}$$

$$B = C^{-1}U_\mathcal{I}^\intercal\Sigma_{YX}\Sigma_{XX}^{-1} \tag{9}$$

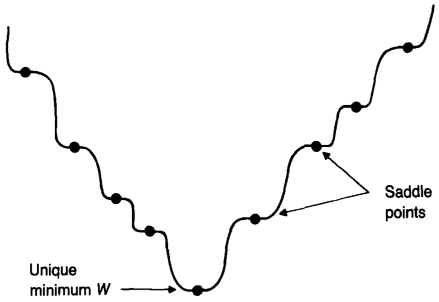
For such a critical point we have:

$$W = P_{U_\mathcal{I}}\Sigma_{YX}\Sigma_{XX}^{-1} \tag{10}$$

$$E(A,B) = tr(\Sigma_{YY}) - \sum_{i\in\mathcal{I}}\lambda_i \tag{11}$$

Therefore a critical W of rank p is always the product of the ordinary least squares regression matrix followed by an orthogonal projection onto the subspace spanned by p eigenvectors of Σ. The critical map W associated with the index set $1, 2, \ldots, p$ is the unique local and global minimum of E. The remaining $\binom{n}{p} - 1$ p-index sets correspond to saddle points. All additional critical points defined by matrices A and B which are not full rank are also saddle points and can be characterized in terms of orthogonal projections onto subspaces spanned by q eigenvectors of Σ with $q < p$ (see Figure 1).

Figure 1: Landscape of E



In the auto-associative case, $\Sigma = \Sigma_{XX}$ and (8), (9) and (10) become:

$$A = U_{\mathcal{I}} C \tag{8'}$$

$$B = C^{-1} U_{\mathcal{I}}^{\intercal} \tag{9'}$$

$$W = P_{U_{\mathcal{I}}} \tag{10'}$$

and therefore the unique locally and globally optimal map W is the orthogonal projection onto the subspace spanned by the first p eigenvectors of Σ_{XX} associated with the p largest eigenvalues.

Remark: at the global minimum, if C is the identity I_p then the activities of the units in the hidden layer are given by:

$$u_1^{\intercal} \hat{y}_n, \ldots, u_p^{\intercal} \hat{y}_n$$

the so called **principal components** of the output data $\hat{y}$. In the auto-associative case, these activities are given by $u_1^{\intercal} x_n, \ldots, u_p^{\intercal} x_n$, the principal components of the input data x. They are the coordinates of the vector x_n along the first p eigenvectors of Σ_{XX}.

2.4.1 Proof of fact 4

First notice that since Σ is a real symmetric covariance matrix, it can always be written as $\Sigma = U \Lambda U^{\intercal}$ where U is an orthogonal column matrix of eigenvectors of Σ and Λ is the diagonal matrix with non-increasing eigenvalues on its diagonal. Also if Σ is full-rank, then $\Sigma_{XX}, \Sigma_{XY}, \Sigma_{YX}$ are full rank too. Now clearly if A and B satisfy (8) and (9) for some C and some $\mathcal{I}$, then A and B are full rank p and satisfy (3) and (5). Therefore they define a critical point of E.

For the converse , we have:

$$P_{U^{\intercal} A} = U^{\intercal} A (A^{\intercal} U U^{\intercal} A)^{-1} A^{\intercal} U = U^{\intercal} A (A^{\intercal} A)^{-1} A^{\intercal} U = U^{\intercal} P_A U$$

or, equivalently, $P_A = U P_{U^{\intercal} A} U^{\intercal}$. Hence (7) yields:

$$U P_{U^{\intercal} A} U^{\intercal} U \Lambda U^{\intercal} = P_A \Sigma = \Sigma P_A = U \Lambda U^{\intercal} U P_{U^{\intercal} A} U^{\intercal}$$

and so $P_{U^\intercal A}\Lambda = \Lambda P_{U^\intercal A}$. Since $\lambda_1 > \cdots > \lambda_n > 0$, it is readily seen that $P_{U^\intercal A}$ is an orthogonal projector of rank p and its eigenvalues are 1 (p times) and 0 ($n-p$ times). Therefore there exists a unique index set $\mathcal{I} = i_1, \ldots, i_p$ with $1 \leq i_1 < \cdots < i_p \leq n$ such that $P_{U^\intercal A} = I_{\mathcal{I}}$, where $I_{\mathcal{I}}$ is the diagonal matrix with entry $i = 1$ if $i \in \mathcal{I}$ and 0 otherwise. It follows that

$$P_A = UP_{U^\intercal A}U^\intercal = UI_{\mathcal{I}}U^\intercal = U_{\mathcal{I}}U_{\mathcal{I}}^\intercal$$

where $U_{\mathcal{I}} = [u_{i_1}, \ldots, u_{i_p}]$. Thus P_A is the orthogonal projection onto the subspace spanned by the columns of $U_{\mathcal{I}}$. Since the column space of A coincides with the column space of $U_{\mathcal{I}}$, there exists an invertible $p \times p$ matrix C such that $A = U_{\mathcal{I}}C$. Moreover, $B = \hat{B}(A) = C^{-1}U_{\mathcal{I}}\Sigma_{YX}\Sigma_{XX}^{-1}$ and (8) and (9) are satisfied. There are $\binom{n}{p}$ possible choices for $\mathcal{I}$ and therefore $\binom{n}{p}$ possible critical points with full rank. From (8), (9) and (10) results immediately.

To prove (11), use (c) to write:

$$\begin{aligned} E(A,B) &= (vec(Y - ABX))^\intercal(vec(Y - ABX)) \\ &= vec(Y)^\intercal vec(Y) - 2(vec(ABX))^\intercal vec(Y) + vec(ABX)^\intercal vec(ABX) \\ &= tr(YY^\intercal) - 2tr(ABXY^\intercal) + tr(ABXX^\intercal B^\intercal A^\intercal) \\ &= tr(\Sigma_{YY}) - 2tr(W\Sigma_{XY}) + tr(W\Sigma_{XX}W^\intercal) \end{aligned}$$

If A is full rank and $B = \hat{B}(A)$, then $W = AB(A) = P_A\Sigma_{YX}\Sigma XX^{-1}$ and therefore:

$$tr(W\Sigma_{XX}W^\intercal) = tr(P_A\Sigma P_A) = tr(P_A\Sigma) = tr(UP_{U^\intercal A}U^\intercal U\Lambda U^\intercal) =$$

$$= tr(P_{U^\intercal A}U^\intercal U\Lambda) = tr(P_{U^\intercal A}\Lambda)$$

and

$$tr(W\Sigma_{YX}) = tr(P_A\Sigma) = tr(P_{U^\intercal A}\Lambda)$$

So, for an arbitrary A of rank p:

$$E(A, \hat{B}(A)) = tr(\Sigma_{YY}) - tr(P_{U^\intercal A}\Lambda)$$

If A is of the form $U_{\mathcal{I}}C$, then $P_{U^\intercal A} = I_{\mathcal{I}}$, therefore:

$$E(A, \hat{B}(A)) = tr(\Sigma_{YY}) - tr(I_{\mathcal{I}}\Lambda) = tr(\Sigma_{YY}) - \sum_{i \in \mathcal{I}} \lambda_i$$

which is (11).

We shall now establish that whenever A and B satisfy (8) and (9) with $\mathcal{I} = 1, 2, \ldots, p$ there exist matrices $\bar{A}$, $\bar{B}$ arbitrarily close to A, B such that $E(\bar{A}, \bar{B}) < E(A, B)$. For this purpose it is enough to slightly perturb the column space of A in the direction of an eigenvector associated with one of the first p eigenvalues of Σ which is not contained in $\{\lambda_i, i \in \mathcal{I}\}$. More precisely, fix two indeces j and k with $j \in \mathcal{I}, k \notin \mathcal{I}$. For any ϵ, put:

$$\tilde{u}_j = (1 + \epsilon^2)^{-\frac{1}{2}}(u_j + \epsilon u_k) = \frac{1}{\sqrt{1 + \epsilon^2}}(u_j + \epsilon u_k)$$

and construct $\tilde{U}_{\mathcal{I}}$ from $U_{\mathcal{I}}$ by replacing u_i with $\tilde{u}_j$. Since $k \notin \mathcal{I}$, we still have $\tilde{U}_{\mathcal{I}}^{\intercal}\tilde{U}_{\mathcal{I}} = I_p$. Now let $\tilde{A} = \tilde{U}_{\mathcal{I}}C$ and

$$\tilde{B} = \hat{B}(\tilde{A}) = C^{-1}\tilde{U}_{\mathcal{I}}^{\intercal}\Sigma_{YX}\Sigma_{XX}^{-1}$$

A simple calculation shows that the diagonal elements of $P_{U^{\intercal}A}$ are:

$$\tilde{\delta}_i = \begin{cases} 0 & \text{if } i \notin \mathcal{I} \cup \{k\} \\ 1 & \text{if } i \in \mathcal{I} \text{ and } i \neq j \text{ and } i \neq k \\ \dfrac{1}{1+\epsilon^2} & \text{if } i = j \\ \dfrac{\epsilon^2}{1+\epsilon^2} & \text{if } i = k \end{cases}$$

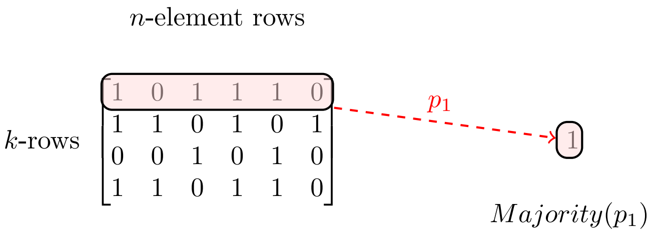
Therefore:

$$\begin{aligned} E(\tilde{A}, \tilde{B}) &= tr(\Sigma_{YY}) - tr(P_{U^{\intercal}\tilde{A}}\Lambda) \\ &= tr \end{aligned}$$

3 Boolean Autoencoders

Boolean autoencoders correspond to the case where $\mathbb{F} = \mathbb{G} = \{0,1\}$, A and B are classes of Boolean functions, and Δ is the Hamming distance. Traditionally, a Boolean function is defined as a mapping from $\{0,1\}^n$ to $\{0,1\}$. but here we use the same term more generally to refer to Boolean vector functions, that is functions from $\{0,1\}^n$ to $\{0,1\}^m$ which of course can be seen as m Boolean functions.

In the general framework, the sets $\mathcal{A}, \mathcal{B}$ contain all possible Boolean functions of the right dimensions. Given k binary column vectors $p_1, \ldots, p_k$ in the n-dimensional hypercube $\mathbb{H}^n$, we define the corresponding binary majority vector Majority(p) in $\mathbb{H}^n$ by taking in each row j the majority of the corresponding components p_{ij}. When n is even, there can be ties in which case one can flip a fair coin to assign the corresponding value.



Lemma 3.1. *The vector Majority(p) is a vector in $\mathbb{H}^n$ closest to the center of gravity of the vectors $p_1, \ldots, p_k$ and it minimizes the function $E(q) = \sum_{i=1}^{k} \Delta(p_i, q)$.*

Proof. The center of gravity is the vector c in $\mathbb{R}^n$ with coordinates

$$c_j = \frac{\left(\sum_{i=1}^{k} p_{ji}\right)}{k}$$

For any j, $(p)_j$ is the closest binary value to c_j. Furthermore:

$$\sum_{i=1}^{k} \Delta(\text{Majority}(p), p_i) = \sum_{i=1}^{k}\sum_{j=1}^{n} \Delta(\text{Majority}(p)_j, p_{ij}) = \sum_{j=1}^{n}\left(\sum_{i=1}^{k} \Delta(\text{Majority}(p)_j, p_{ij})\right)$$

and each term in the last sum is minimized by the majority vector. □

A **Voronoi partition** of $\mathbb{H}^n$ generated by the vectors $p_1, \dots, p_k$ is a partition of $\mathbb{H}^n$ into k regions $\mathcal{C}^{Vor}(p_1), \dots, \mathcal{C}^{Vor}(p_k)$ such that for each x in $\mathbb{H}^n$:

$$x \in \mathcal{C}^{Vor}(p_i) \iff \Delta(x, p_i) \leq \Delta(x, p_j) \text{ for all } j \neq i$$

And this can be visualized as:

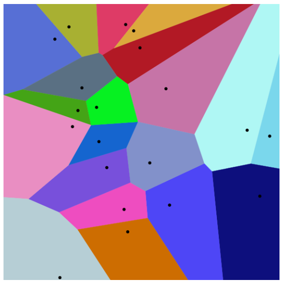


Figure 2: Voronoi diagram with Euclidean distance metric

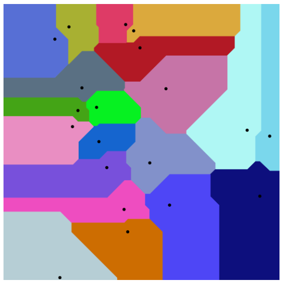


Figure 3: Voronoi diagram with Manhattan distance metric

Theorem 3.2. ***Fixed layer solution:*** *if the A mapping is fixed , then the optimal mapping B^* is given by $B^*(x) = h_i$ for any x in $\mathcal{C}_i = \mathcal{C}^{Vor}(A(h_i))$. Conversely, if B is fixed, then the optimal mapping A^* is given by $A^*(h_i) = $ Majority $\left[\mathcal{X} \cap B^{-1}(h_i)\right]$*

Proof. Assume first that A is fixed . Then for each of the 2^p possible Boolean vectors $h_1, \dots, h_{2^p}$ of the hidden layer, $A(h_1), \dots, A(h_{2^p})$ provide 2^p points (centroids) in the hypercube $\mathbb{H}^n$. One can build the corresponding Voronoi partition by assigning each point of $\mathbb{H}^n$ to its closest centroid, breaking ties arbitrarily, thus forming a partition of $\mathbb{H}^n$ into 2^p corresponding clusters $\mathcal{C}_1, \dots, \mathcal{C}_{2^p}$, with $\mathcal{C}_i = \mathcal{C}^{Vor}(A(h_i))$. The optimal mapping B^* is then given by $B^*(x) = h_i$ for any x in $\mathcal{C}_i$.

Conversely, assume that B is fixed. Then for each of the 2^p possible Boolean vectors $h_1, \ldots, h_{2^p}$ of the hidden layer, let $B^{-1}(h_i) = \{x \in \mathbb{H}^n : B(x) = h_i\}$. To minimize the reconstruction error, A^* must map h_i onto a point y of $\mathbb{H}^n$ minimizing the sum of Hamming distances to points in $\mathcal{X} \cap B^{-1}(h_i)$. By Lemma 3.1m the minimum is realized by the component-wise majority vector $A^*(h_i) = \text{Majority}[\mathcal{X} \cap B^{-1}(h_i)]$, breaking ties arbitrarily. Note that this solution minimizes the distortion on the training set. The generalization or total distortion however, is minimized by $A^*(h_i) = \text{Majority}[B^{-1}(h_i)]$. In some situations, one may have the additional constraint that the output vector must belong to the training. With this additional constraint the optimal solution is $A^*(h_i)$ should be the vector $\mathcal{X}$ that is closest to the vector $\text{Majority}[\mathcal{X} \cap B^{-1}(h_i)]$. $\square$