Supplementary materials for convolutional dictionary learning based auto-encoders for natural exponential-family distributions

1 Gradient dynamics of shallow exponential auto-encoder (SEA)

Theorem 1.1. *(informal). Given a "good" initial estimate of the dictionary from the binomial dictionary learning problem, and infinitely many examples, the binomial SEA, when trained by gradient descent through backpropagation, learns the dictionary.*

Theorem 1.2. *Suppose the generative model satisfies (A1) - (A14). Given infinitely many examples (i.e., $J \to \infty$), the binomial SEA with $\mathcal{S}_{\mathbf{b}} = ReLU_{\mathbf{b}}$ trained by approximate gradient descent followed by normalization using the learning rate of $\kappa = O(p/s)$ (i.e., $\mathbf{w}_i^{(l+1)} = normalize(\mathbf{w}_i^{(l)} - \kappa g_i)$) recovers $\mathbf{A}$. More formally, there exists $\delta \in (0,1)$ such that at every iteration l, $\forall i$ $\|\mathbf{w}_i^{(l+1)} - \mathbf{a}_i\|_2^2 \leq (1-\delta)\|\mathbf{w}_i^{(l)} - \mathbf{a}_i\|_2^2 + \kappa \cdot O(\frac{\max(s^2, s^3/p^{\frac{2}{3}+2\xi})}{p^{1+6\xi}})$.*

Theorem 1.3. *Suppose the generative model satisfies (A1) - (A13). Given infinitely many examples (i.e., $J \to \infty$), the binomial SEA with $\mathcal{S}_{\mathbf{b}} = HT_{\mathbf{b}}$ trained by approximate gradient descent followed by normalization using the learning rate of $\kappa = O(p/s)$ (i.e., $\mathbf{w}_i^{(l+1)} = normalize(\mathbf{w}_i^{(l)} - \kappa g_i)$) recovers $\mathbf{A}$. More formally, there exists $\delta \in (0,1)$ such that at every iteration l, $\forall i$ $\|\mathbf{w}_i^{(l+1)} - \mathbf{a}_i\|_2^2 \leq (1-\delta)\|\mathbf{w}_i^{(l)} - \mathbf{a}_i\|_2^2 + \kappa \cdot O(\frac{\max(s^2, s^3/p^{\frac{2}{3}+2\xi})}{p^{1+6\xi}})$.*

In proof of the above theorem, our approach is similar to [Nguyen et al., 2019].

1.1 Generative model and architecture

We have J binomial observations $\mathbf{y}^j = \sum_{m=1}^{M_j} \mathbb{1}_m^j$ where $\mathbf{y}^j$ can be seen as sum of M_j independent Bernoulli random variables (i.e., $\mathbb{1}_m^j$). We can express $\sigma^{-1}(\boldsymbol{\mu}) = \mathbf{A}\mathbf{x}^*$, where $\sigma(z) = \frac{\mathbf{e}^z}{1+\mathbf{e}^z}$ is the inverse of the corresponding link function (sigmoid), $\mathbf{A} \in \mathbb{R}^{n \times p}$ is a matrix dictionary, and $\mathbf{x}^* \in \mathbb{R}^p$ is a sparse vector. Hence, we have

$$E[\mathbf{y}^j] = \boldsymbol{\mu} = \frac{\mathbf{e}^{\mathbf{A}\mathbf{x}^*}}{1 + \mathbf{e}^{\mathbf{A}\mathbf{x}^*}} = \sigma(\mathbf{A}\mathbf{x}^*). \tag{1}$$

In this analysis, we assume that there are infinitely many examples (i.e., $J \to \infty$), hence, we use the expectation of the gradient for backpropagation at every iteration. We also assume that there are infinite number of Bernoulli observation for each binomial observation (i.e., $M_j \to \infty$). Hence, from the Law of Large Numbers, we have the following convergence in probability

$$\lim_{M_j \to \infty} \frac{1}{M_j}\mathbf{y}^j = \lim_{M_j \to \infty} \frac{1}{M_j} \sum_{m=1}^{M_j} \mathbb{1}_m^j = \boldsymbol{\mu} = \sigma(\mathbf{A}\mathbf{x}^*), \tag{2}$$

We drop j for ease of notation. Algorithm 1 shows the architecture when the code is initialized to $\mathbf{0}$. $\mathbf{W} \in \mathbb{R}^{n \times p}$ are the weights of the auto-encoder. The encoder is unfolded only once and the step size of the proximal mapping is set to 4 (i.e., assuming the maximum singular value of $\mathbf{A}$ is 1, then 4 is the largest step size to ensure convergence of the encoder as the first derivative of sigmoid is bounded by $\frac{1}{4}$. For Theorem 1.2, $\mathcal{S}_{\mathbf{b}}(\mathbf{z})$ is an element-wise operator where $\mathcal{S}_{b_i}(z_i) = \mathrm{ReLU}_{b_i}(z_i) = z_i \cdot \mathbb{1}_{|z_i| \geq b_i}$, and $\frac{1}{2} = \sigma(\mathbf{0})$ appears in the first layer as of the initial code estimate is $\mathbf{0}$. From the definition of ReLU, we can see that $\mathbf{x} = \mathcal{S}_{\mathbf{b}}(\mathbf{c}_1) = \mathbb{1}_{\mathbf{x} \neq 0}(\mathbf{c}_1 - \mathbf{b})$ where $\mathbb{1}_{\mathbf{x} \neq 0}$ is an indicator function. For Theorem 1.3, $\mathcal{S}_{\mathbf{b}}(\mathbf{z})$ is an element-wise operator where $\mathcal{S}_{b_i}(z_i) = \mathrm{HT}_{b_i}(z_i) = z_i \cdot \mathbb{1}_{|z_i| \geq b_i}$, and $\mathbf{x} = \mathcal{S}_{\mathbf{b}}(\mathbf{c}_1) = \mathbb{1}_{\mathbf{x} \neq 0}\mathbf{c}_1$.

Algorithm 1: SEA.

```
Input: $\mathbf{y}, \mathbf{W}, \mathbf{b}$
Output: $\mathbf{c}_2$
$\mathbf{c}_1 = 4\mathbf{W}^{\mathrm{T}}(\mathbf{y} - \frac{1}{2})$
$\mathbf{x} = \mathcal{S}_{\mathbf{b}}(\mathbf{c}_1)$
$\mathbf{c}_2 = \mathbf{W}\mathbf{x}$
```

1.2 Assumptions and definitions

Given the following definition and notations,

(D1) $\mathbf{W}$ is q-close to $\mathbf{A}$ if there is a permutation π and sign flip operator u such that $\forall i$ $\|u(i)\mathbf{w}_{\pi(i)} - \mathbf{a}_i\|_2 \leq q$.

(D2) $\mathbf{W}$ is (q, ε)-near to $\mathbf{A}$ if $\mathbf{W}$ is q-close to $\mathbf{A}$ and $\|\mathbf{W} - \mathbf{A}\|_2 \leq \varepsilon \|\mathbf{A}\|_2$.

(D3) A unit-norm columns matrix $\mathbf{A}$ is η-incoherent if for every pair (i, j) of columns, $|\langle \mathbf{a}_i, \mathbf{a}_j \rangle| \leq \frac{\eta}{\sqrt{n}}$.

(D4) We define column i of $\mathbf{W}$ as $\mathbf{w}_i$.

(D5) $\mathbf{w}_i$ is τ_i-correlated to $\mathbf{a}_i$ if $\tau_i = \langle \mathbf{w}_i, \mathbf{a}_i \rangle = \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_i$. Hence, $\|\mathbf{w}_i - \mathbf{a}_i\|_2^2 = 2(1-\tau_i)$.

(D6) From the binomial likelihood, the loss would be $\lim_{M\to\infty} \mathcal{L}_{\mathbf{W}}(\mathbf{y}, \mathbf{W}\mathbf{x}) = \lim_{M\to\infty} -\frac{1}{M}\big(\mathbf{W}\mathbf{x}\big)^{\mathrm{T}}\mathbf{y} + \mathbf{1}_n^{\mathrm{T}} \log\big(1 + \exp\big(\mathbf{W}\mathbf{x}\big)\big)$.

(D7) We denote the expectation of the gradient of the loss defined in (D6) with respect to $\mathbf{w}_i$ to be $g_i = E[\lim_{M\to\infty} \frac{\partial \mathcal{L}_{\mathbf{W}}}{\partial \mathbf{w}_i}]$.

(D8) $\mathbf{W}_{\setminus i}$ denotes the matrix $\mathbf{W}$ with column i removed, and $S^{\setminus i}$ denotes S excluding i.

(D9) $[\mathbf{z}]_d$ denotes z_d (i.e., the d^{th} element fo the vector $\mathbf{z}$).

(D10) $[p]$ denotes the set $\{1, \ldots, p\}$, and $[p]^{\setminus i}$ denotes $[p]$ excluding i.

(D11) For $\mathbf{A} \in \mathbb{R}^{n\times p}$, $\mathbf{A}_S \in \mathbb{R}^{n \times s}$ indicates a matrix with columns from the set S. Similarly, for $\mathbf{x}^* \in \mathbb{R}^p$, $\mathbf{x}_S^* \in \mathbb{R}^s$ indicates a vector containing only the elements with indices from S.

we assume the generative model satisfies the following assumptions:

(A1) Let the code $\mathbf{x}^*$ be s-sparse and have support S (i.e., $\mathrm{supp}(\mathbf{x}) = S$) where each element of S is chosen uniformly at random without replacement from the set $[p]$. Hence, $p_i = P(i \in S) = s/p$ and $p_{ij} = P(i, j \in S) = s(s-1)/(p(p-1))$.

(A2) Each code is bounded (i.e., $|x_i| \in [L_x, C_x]$) where $0 \leq L_x \leq C_x$ and $C_x = O(\frac{1}{p^{\frac{1}{3}+\xi}})$, and $\xi > 0$. Then $\|\mathbf{x}_S^*\|_2 \leq \sqrt{s} C_x$. For the case when $\mathcal{S}_{\mathbf{b}} = \mathrm{ReLU}_{\mathbf{b}}$, we assume the code is non-negative.

(A3) Given the support, we assume $\mathbf{x}_S^*$ is i.i.d, zero-mean, and has symmetric probability density function. Hence, $E[\mathbf{x}_i^* \mid S] = 0$ and $E[\mathbf{x}_S^* \mathbf{x}_S^{*\mathrm{T}} \mid S] = \nu \mathbf{I}$ where $\nu \leq C_x$.

(A4) From the forward pass of the encoder, $\mathrm{supp}(\mathbf{x}) = \mathrm{supp}(\mathbf{x}^*) = S$ with high probability. We call this code consistency, a similar definition from [Nguyen et al., 2019]. This code consistency enforces some conditions (i.e., based on L_x and C_x for ReLU and L_x for HT) on the value of $\mathbf{b}$ which we do not explicitly express. For when $\mathcal{S}_{\mathbf{b}} = \mathrm{ReLU}_{\mathbf{b}}$, $\mathbf{W}\mathbf{x} = \mathbf{W}_S \mathbf{x}_S = 4\mathbf{W}_S \mathbf{W}_S^{\mathrm{T}}(\mathbf{y} - \frac{1}{2}) - \mathbf{W}_S \mathbf{b}_S$, and for when $\mathcal{S}_{\mathbf{b}} = \mathrm{HT}_{\mathbf{b}}$, $\mathbf{W}\mathbf{x} = 4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\mathbf{y} - \frac{1}{2})$.

(A5) We assume $\forall i$ $\|\mathbf{a}_i\|_2 = 1$.

(A6) Given $s < n \leq p$, we have $\|\mathbf{A}\|_2 = O(\sqrt{p/n})$ and $\|\mathbf{A}_S\|_2 = O(1)$.

(A7) $\mathbf{W}$ is $(q, 2)$-near $\mathbf{A}$; thus, $\|\mathbf{W}\|_2 \leq \|\mathbf{W} - \mathbf{A}\|_2 + \|\mathbf{A}\|_2 \leq O(\sqrt{p/n})$.

(A8) $\mathbf{A}$ is η-incoherent.

(A9) $\mathbf{w}_i$ is τ_i-correlated to $\mathbf{a}_i$.

(A10) For any $i \neq j$, we have $|\langle \mathbf{w}_i, \mathbf{a}_j \rangle| = |\langle \mathbf{a}_i, \mathbf{a}_j\rangle + \langle \mathbf{w}_i - \mathbf{a}_i, \mathbf{a}_j \rangle| \leq \frac{\eta}{\sqrt{n}} + \|\mathbf{w}_i - \mathbf{a}_i\|_2 \|\mathbf{a}_j\|_2 \leq \frac{\eta}{\sqrt{n}} + q$.

(A11) We assume the network is trained by approximate gradient descent followed by normalization using the learning rate of κ. Hence, the gradient update for column i at iteration l is $\mathbf{w}_i^{(l+1)} = \mathbf{w}_i^{(l)} - \kappa g_i$. At the normalization step, $\forall i$, we enforce $\|\mathbf{w}_i\|_2 = 1$. Lemma 5 in [Nguyen et al., 2019] shows that descent property can also be achieved with the normalization step.

(A12) We use the Taylor series of $\sigma(z)$ around 0. Hence, $\sigma(z) = \frac{1}{2} + \frac{1}{4}z + \nabla^2\sigma(\bar{z})(z)^2$, where $0 \leq \bar{z} \leq z$ and ∇^2 denotes Hessian.

(A13) To simplify notation, we assume that the permutation operator $\pi(.)$ is identity and the sign flip operator $u(.)$ is $+1$.

(A14) When $\mathcal{S}_{\mathbf{b}} = \mathrm{ReLU}_{\mathbf{b}}$, at every iteration of the gradient descent, given τ_i, the bias $\mathbf{b}$ in the network satisfies $|\nu\tau_i(\tau_i - 1) + b_i^2| \leq 2\tau_i(1-\tau_i)$.

1.3 Non-negative sparse coding with $\mathcal{S}_{\mathbf{b}} = \mathrm{ReLU}_{\mathbf{b}}$

1.3.1 Gradient derivation

First, we derive g_i when $\mathcal{S}_{\mathbf{b}} = \mathrm{ReLU}_{\mathbf{b}}$. In this derivation, by dominated convergence theorem, we interchange the limit and derivative. We also compute the limit inside $\sigma(.)$ as it is a continuous function.

$$\begin{aligned}
\lim_{M\to\infty} \frac{\partial \mathcal{L}_{\mathbf{W}}}{\partial \mathbf{w}_i} &= \lim_{M\to\infty} \frac{\partial \mathbf{c}_1}{\partial \mathbf{w}_i}\frac{\partial \mathcal{L}_{\mathbf{W}}}{\partial \mathbf{c}_1} + \frac{\partial \mathbf{c}_2}{\partial \mathbf{w}_i}\frac{\partial \mathcal{L}_{\mathbf{W}}}{\partial \mathbf{c}_2} = \frac{\partial \mathbf{c}_1}{\partial \mathbf{w}_i}\frac{\partial \mathbf{x}}{\partial \mathbf{c}_1}\frac{\partial \mathbf{c}_2}{\partial \mathbf{x}}\frac{\mathcal{L}_{\mathbf{W}}}{\partial \mathbf{c}_2} + \frac{\partial \mathbf{c}_2}{\partial \mathbf{w}_i}\frac{\partial \mathcal{L}_{\mathbf{W}}}{\partial \mathbf{c}_2} \\
&= \left(\underbrace{[0, 0, \ldots, 4(\boldsymbol{\mu} - \frac{1}{2}), \ldots, 0]}_{n\times p} \underbrace{\mathrm{diag}(\mathcal{S}_{\mathbf{b}}'(\mathbf{c}_1))}_{p \times p} \mathbf{W}^{\mathrm{T}} + \mathbb{1}_{\mathbf{x}_i \neq 0}(\mathbf{w}_i^{\mathrm{T}} 4(\boldsymbol{\mu} - \frac{1}{2})\mathbf{I} - b_i \mathbf{I}) \right) \\
&\times \left(-\sigma(\mathbf{A}\mathbf{x}^*) + \sigma(4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu} - \frac{1}{2}) - \mathbf{W}_S\mathbf{b}_S) \right) \\
&= \left(\mathcal{S}_{b_i}'(\mathbf{c}_{1,i}) 4(\boldsymbol{\mu} - \frac{1}{2})\mathbf{w}_i^{\mathrm{T}} + \mathbb{1}_{\mathbf{x}_i \neq 0}(\mathbf{w}_i^{\mathrm{T}} 4(\boldsymbol{\mu} - \frac{1}{2}) - b_i)\mathbf{I} \right) \left(\sigma(4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu} - \frac{1}{2}) - \mathbf{W}_S\mathbf{b}_S) - \sigma(\mathbf{A}\mathbf{x}^*) \right).
\end{aligned} \tag{3}$$

We further expand the gradient, by replacing $\sigma(.)$ with its Taylor expansion. We have

$$\sigma(\mathbf{A}\mathbf{x}^*) = \frac{1}{2} + \frac{1}{4}\mathbf{A}\mathbf{x}^* + \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^{\mathrm{T}}$, $\epsilon_d = \nabla^2\sigma(u_d)([\mathbf{A}\mathbf{x}^*]_d)^2$, and $0 \leq u_d \leq [\mathbf{A}\mathbf{x}^*]_d$. Similarly,

$$\sigma(4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu} - \frac{1}{2}) - \mathbf{W}_S\mathbf{b}_S) = \frac{1}{2} + \mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu} - \frac{1}{2}) - \frac{1}{4}\mathbf{W}_S\mathbf{b}_S + \tilde{\boldsymbol{\epsilon}}, \tag{5}$$

where $\tilde{\boldsymbol{\epsilon}} = [\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n]^{\mathrm{T}}$, $\tilde{\epsilon}_d = \nabla^2\sigma(\tilde{u}_d)([4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu} - \frac{1}{2}) - \mathbf{W}_S\mathbf{b}_S]_d)^2$, and $0 \leq \tilde{u}_d \leq [4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu} - \frac{1}{2}) - \mathbf{W}_S\mathbf{b}_S]_d$. Again, replacing $\boldsymbol{\mu}$ with Taylor expansion of $\sigma(\mathbf{A}\mathbf{x}^*)$, we get

$$\sigma(4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu} - \frac{1}{2}) - \mathbf{W}_S\mathbf{b}_S) = \frac{1}{2} + \mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\frac{1}{4}\mathbf{A}\mathbf{x}^* + \boldsymbol{\epsilon}) - \frac{1}{4}\mathbf{W}_S\mathbf{b}_S + \tilde{\boldsymbol{\epsilon}}. \tag{6}$$

By symmetry, $E[\boldsymbol{\epsilon} \mid S] = E[\tilde{\boldsymbol{\epsilon}} \mid S] = 0$. The expectation of gradient g_i would be

$$g_i = E[\mathbb{1}_{\mathbf{x}_i \neq 0}\left((\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}} + (\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{I} - b_i\mathbf{I}\right)(\frac{1}{4}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})(\mathbf{A}\mathbf{x}^*) - \frac{1}{4}\mathbf{W}_S\mathbf{b}_S + (\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\boldsymbol{\epsilon} + \tilde{\boldsymbol{\epsilon}})]. \tag{7}$$

1.3.2 Gradient dynamics

Given the code consistency from the forward pass of the encoder, we replace $\mathbb{1}_{\mathbf{x}_i \neq 0}$ with $\mathbb{1}_{\mathbf{x}_i^* \neq 0}$ and denote the error by γ as below which is small for large p [Nguyen et al., 2019].

$$\gamma = E[(\mathbb{1}_{\mathbf{x}_i^* \neq 0} - \mathbb{1}_{\mathbf{x}_i \neq 0})\left((\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}} + \mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{I} - b_i\mathbf{I}\right)(\frac{1}{4}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})(\mathbf{A}\mathbf{x}^*) - \frac{1}{4}\mathbf{W}_S\mathbf{b}_S + (\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\boldsymbol{\epsilon} + \tilde{\boldsymbol{\epsilon}})]. \tag{8}$$

Now, we write g_i as

$$g_i = E[\mathbb{1}_{\mathbf{x}_i^* \neq 0}\left((\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}} + \mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{I} - b_i\mathbf{I}\right)(\frac{1}{4}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})(\mathbf{A}\mathbf{x}^*) - \frac{1}{4}\mathbf{W}_S\mathbf{b}_S + (\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\boldsymbol{\epsilon} + \tilde{\boldsymbol{\epsilon}})] + \gamma. \tag{9}$$

We can see that if $i \notin S$ then $\mathbb{1}_{\mathbf{x}_i^*} = 0$ hence, $g_i = 0$. Thus, in our analysis, we only consider the case $i \in S$. We decompose g_i as below.

$$g_i = g_i^{(1)} + g_i^{(2)} + g_i^{(3)} + \gamma, \tag{10}$$

where

$$g_i^{(1)} = E[\frac{1}{4}\mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^*]. \tag{11}$$

$$g_i^{(2)} = E[\frac{1}{4}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^*]. \tag{12}$$

$$g_i^{(3)} = E[\left((\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}} + \mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{I}\right)((\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\boldsymbol{\epsilon} + \tilde{\boldsymbol{\epsilon}})]. \tag{13}$$

$$g_i^{(4)} = E[(-b_i)((\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\boldsymbol{\epsilon} + \tilde{\boldsymbol{\epsilon}})]. \tag{14}$$

$$g_i^{(5)} = E[(-b_i)(\frac{1}{4}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})(\mathbf{A}\mathbf{x}^*) - \frac{1}{4}\mathbf{W}_S\mathbf{b}_S)]. \tag{15}$$

$$g_i^{(6)} = E[\left((\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}} + \mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{I}\right)(-\frac{1}{4}\mathbf{W}_S\mathbf{b}_S)]. \tag{16}$$

We define

$$g_{i,S}^{(1)} = E[\frac{1}{4}\mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^* \mid S]. \tag{17}$$

$$g_{i,S}^{(2)} = E[\frac{1}{4}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^* \mid S]. \tag{18}$$

$$g_{i,S}^{(3)} = E[\left((\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}} + \mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{I}\right)((\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\boldsymbol{\epsilon} + \tilde{\boldsymbol{\epsilon}}) \mid S]. \tag{19}$$

$$g_{i,S}^{(4)} = E[(-b_i)((\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\boldsymbol{\epsilon} + \tilde{\boldsymbol{\epsilon}}) \mid S]. \tag{20}$$

$$g_{i,S}^{(5)} = E[(-b_i)(\frac{1}{4}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})(\mathbf{A}\mathbf{x}^*) - \frac{1}{4}\mathbf{W}_S\mathbf{b}_S) \mid S]. \tag{21}$$

$$g_{i,S}^{(6)} = E[\left((\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}} + \mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{I}\right)(-\frac{1}{4}\mathbf{W}_S\mathbf{b}_S) \mid S]. \tag{22}$$

Hence, $g_i^{(k)} = E[g_{i,S}^{(k)}]$ for $k = 1, \ldots, 6$, where the expectations are with respect to the support S.

$$\begin{aligned}
g_{i,S}^{(1)} &= E[\frac{1}{4}\mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^* \mid S] \\
&= \sum_{j,l \in S} E[\frac{1}{4}\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_j\mathbf{x}_j^*(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{a}_l\mathbf{x}_l^* \mid S] + E[\mathbf{w}_i^{\mathrm{T}}\boldsymbol{\epsilon}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^* \mid S] \\
&= \frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_i(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{a}_i + \sum_{l \in S \setminus i} \frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{a}_l + e_1 = \frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_i(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{a}_i + r_1 + e_1.
\end{aligned} \tag{23}$$

We denote $r_1 = \sum_{l \in S \setminus i} \frac{1}{4} \nu \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_l (\mathbf{W}_S \mathbf{W}_S^{\mathrm{T}} - \mathbf{I}) \mathbf{a}_l$ and $e_1 = E[\mathbf{w}_i^{\mathrm{T}} \boldsymbol{\epsilon} (\mathbf{W}_S \mathbf{W}_S^{\mathrm{T}} - \mathbf{I}) \mathbf{A}\mathbf{x}^* \mid S]$. Similarly for $g_{i,S}^{(2)}$, we have

$$g_{i,S}^{(2)} = \frac{1}{4} \nu \mathbf{a}_i \mathbf{w}_i^{\mathrm{T}} (\mathbf{W}_S \mathbf{W}_S^{\mathrm{T}} - \mathbf{I}) \mathbf{a}_i + r_2 + e_2. \tag{24}$$

We denote $r_2 = \sum_{l \in S \setminus i} \frac{1}{4} \nu \mathbf{a}_l \mathbf{w}_i^{\mathrm{T}} (\mathbf{W}_S \mathbf{W}_S^{\mathrm{T}} - \mathbf{I}) \mathbf{a}_l$, $e_2 = E[\boldsymbol{\epsilon} \mathbf{w}_i^{\mathrm{T}} (\mathbf{W}_S \mathbf{W}_S^{\mathrm{T}} - \mathbf{I}) \mathbf{A}\mathbf{x}^* \mid S]$, $e_3 = g_{i,S}^{(3)}$, and $e_4 = g_{i,S}^{(4)}$. We compute $g_{i,S}^{(5)}$ and $g_{i,S}^{(6)}$ next.

$$g_{i,S}^{(5)} = E[(-b_i)(\frac{1}{4}(\mathbf{W}_S \mathbf{W}_S^{\mathrm{T}} - \mathbf{I})(\mathbf{A}\mathbf{x}^*) - \frac{1}{4} \mathbf{W}_S \mathbf{b}_S) \mid S] = \frac{1}{4} b_i^2 \mathbf{w}_i + \frac{1}{4} b_i \sum_{j \in S \setminus i} \mathbf{w}_j b_j \tag{25}$$

$$g_{i,S}^{(6)} = E[\big((\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon}) \mathbf{w}_i^{\mathrm{T}} + \mathbf{w}_i^{\mathrm{T}} (\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon}) \mathbf{I}\big) (-\frac{1}{4} \mathbf{W}_S \mathbf{b}_S) \mid S] = 0 \tag{26}$$

We denote $\beta = E[r_1 + r_2 + e_1 + e_2 + e_3 + e_4] + \gamma$. Combining the terms,

$$\begin{aligned} g_i &= E[\frac{1}{4} \nu \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_i (\mathbf{W}_S \mathbf{W}_S^{\mathrm{T}} - \mathbf{I}) \mathbf{a}_i + \frac{1}{4} \nu \mathbf{a}_i \mathbf{w}_i^{\mathrm{T}} (\mathbf{W}_S \mathbf{W}_S^{\mathrm{T}} - \mathbf{I}) \mathbf{a}_i + \frac{1}{4} b_i^2 \mathbf{w}_i + \frac{1}{4} b_i \sum_{j \in S \setminus i} \mathbf{w}_j b_j] + \beta \\ &= E[-\frac{1}{2} \nu \tau_i \mathbf{a}_i + \frac{1}{4} \nu \tau_i \sum_{j \in S} \mathbf{w}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_i + \frac{1}{4} \nu \mathbf{a}_i \mathbf{w}_i^{\mathrm{T}} \sum_{j \in S} \mathbf{w}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_i + \frac{1}{4} b_i^2 \mathbf{w}_i + \frac{1}{4} b_i \sum_{j \in S \setminus i} \mathbf{w}_j b_j] + \beta \\ &= E[-\frac{1}{2} \nu \tau_i \mathbf{a}_i + \frac{1}{4} \nu \tau_i^2 \mathbf{w}_i + \frac{1}{4} \nu \tau_i \sum_{j \in S \setminus i} \mathbf{w}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_i + \frac{1}{4} \nu \tau_i \|\mathbf{w}_i\|_2^2 \mathbf{a}_i \\ &+ \frac{1}{4} \nu \left(\mathbf{a}_i \mathbf{w}_i^{\mathrm{T}} \right) \sum_{j \in S \setminus i} \mathbf{w}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_i + \frac{1}{4} b_i^2 \mathbf{w}_i + \frac{1}{4} b_i \sum_{j \in S \setminus i} \mathbf{w}_j b_j] + \beta \\ &= -\frac{1}{4} p_i \nu \tau_i \mathbf{a}_i + p_i \frac{1}{4} (\nu \tau_i^2 + b_i^2) \mathbf{w}_i + \zeta + \beta, \end{aligned} \tag{27}$$

where $\zeta = \sum_{j \in [p] \setminus i} \frac{1}{4} p_{ij} \nu \tau_i \mathbf{w}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_i + \frac{1}{4} p_{ij} \nu \mathbf{a}_i \mathbf{w}_i^{\mathrm{T}} \mathbf{w}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_i + \frac{1}{4} p_{ij} b_i b_j \mathbf{w}_j$. We continue

$$g_i = \frac{1}{4} p_i \nu \tau_i (\mathbf{w}_i - \mathbf{a}_i) + v, \tag{28}$$

where we denote $v = \frac{1}{4} p_i (\nu \tau_i (\tau_i - 1) + b_i^2) \mathbf{w}_i + \zeta + \beta$.

Lemma 1.4. *Suppose the generative model satisfies* $(A1) - (A14)$*. Then*

$$\|v\|_2 \leq \frac{1}{4} p_i \tau_i \nu q \|\mathbf{w}_i - \mathbf{a}_i\|_2 + O(\max(C_x^3 s \sqrt{s}, C_x^4 s^2)) \tag{29}$$

Proof.

$$\begin{aligned} \|\zeta\|_2 &= \| \sum_{j \in [p] \setminus i} \frac{1}{4} p_{ij} \nu \tau_i \mathbf{w}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_i + \frac{1}{4} p_{ij} \nu \mathbf{a}_i \mathbf{w}_i^{\mathrm{T}} \mathbf{w}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_i + \frac{1}{4} p_{ij} b_i b_j \mathbf{w}_j \|_2 \\ &= \| \frac{1}{4} p_{ij} \nu \tau_i \mathbf{W}_{\setminus i} \mathbf{W}_{\setminus i}^{\mathrm{T}} \mathbf{a}_i + \frac{1}{4} p_{ij} \nu \mathbf{a}_i \mathbf{w}_i^{\mathrm{T}} \mathbf{W}_{\setminus i} \mathbf{W}_{\setminus i}^{\mathrm{T}} \mathbf{a}_i + \frac{1}{4} p_{ij} b_i \mathbf{W}_{\setminus i} b_{\setminus i} \|_2 \\ &\leq \frac{1}{4} p_{ij} \nu \tau_i \|\mathbf{W}_{\setminus i}\|_2^2 \|\mathbf{a}_i\|_2 + \frac{1}{4} p_{ij} \nu \|\mathbf{w}_i\|_2 \|\mathbf{W}_{\setminus i}\|_2^2 \|\mathbf{a}_i\|_2^2 + \frac{1}{4} p_{ij} |b_i| \|\mathbf{W}_{\setminus i}\|_2 \|b_{\setminus i}\|_2 \\ &= \frac{1}{4} O(\nu \tau_i s^2/(np)) + \frac{1}{4} O(\nu s^2/(np)) + \frac{1}{4} O(\sqrt{\frac{p}{n}} s^2/p^2) = O(s^2/(np)). \end{aligned} \tag{30}$$

$$\begin{aligned} \|E[r_1]\|_2 = \|E[\sum_{l \in S \setminus i} \frac{1}{4} \nu \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_l (\mathbf{W}_S \mathbf{W}_S^{\mathrm{T}} - \mathbf{I}) \mathbf{a}_l]\|_2 = \|E[\sum_{l \in S \setminus i} \frac{1}{4} \nu \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_l \mathbf{W}_S \mathbf{W}_S^{\mathrm{T}} \mathbf{a}_l - \sum_{l \in S \setminus i} \frac{1}{4} \nu \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_l \mathbf{a}_l]\|_2 \\ = \| \sum_{l \neq j \neq i} p_{ijl} \frac{1}{4} \nu \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_l \mathbf{w}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_l + \sum_{j \neq i} p_{ij} \frac{1}{4} \nu \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_j \mathbf{w}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_j + p_{il} \frac{1}{4} \nu \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_l \mathbf{w}_i \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_l - \sum_{l \neq i} p_{il} \frac{1}{4} \nu \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_l \mathbf{a}_l \|_2 \leq O(s^2/(np)). \end{aligned} \tag{31}$$

where each terms is bounded as below

$$\| \sum_{l \neq j \neq i} p_{ijl} \frac{1}{4} \nu \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_l \mathbf{w}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_l \|_2 = \frac{1}{4} \nu (s^3/p^3) \|\mathbf{W}_{\setminus i}\|_2 \leq O(s^3/(pn\sqrt{n})). \tag{32}$$

where $z_j = \sum_{l \neq i,j} \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_l \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_l$, hence, $\|z\|_2 \leq O(\frac{p\sqrt{p}}{n})$.

$$\| \sum_{j \neq i} p_{ij} \frac{1}{4} \nu \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_j \mathbf{w}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_j \|_2 = \| \frac{1}{4} \nu (s^2/p^2) \|\mathbf{W}_{\setminus i} z\|_2 \leq O(s^2/(np)). \tag{33}$$

where $z_j = \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_j$, hence, $\|z\|_2 \leq O(\frac{\sqrt{p}}{\sqrt{n}})$.

$$\| \sum_{l \neq j \neq i} p_{ijl} \frac{1}{4} \nu \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_l \mathbf{w}_j \mathbf{w}_j^{\mathrm{T}} \mathbf{a}_l \|_2 = \frac{1}{4} \nu (s^3/p^3) \|\mathbf{W}_{\setminus i}\|_2 \leq O(s^3/(pn\sqrt{n})). \tag{34}$$

$$\| p_{il} \frac{1}{4} \nu \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_l \mathbf{w}_i \mathbf{w}_i^{\mathrm{T}} \mathbf{a}_l \|_2 \leq O(s^2/(np^2)). \tag{35}$$

Following a similar approach for r_2, we get

$$\begin{aligned}\|E[r_2]\|_2 &= \|E[\sum_{l\in S\setminus i}\frac{1}{4}\nu\mathbf{a}_l\mathbf{w}_i^{\mathrm{T}}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}-\mathbf{I})\mathbf{a}_l]\|_2 = \|E[\sum_{l\in S\setminus i}\frac{1}{4}\nu\mathbf{a}_l\mathbf{w}_i^{\mathrm{T}}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}})\mathbf{a}_l - \sum_{l\in S\setminus i}\frac{1}{4}\nu\mathbf{a}_l\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l]\|_2\\ &= \|\sum_{l\neq j\neq i}p_{ijl}\frac{1}{4}\nu\mathbf{a}_l\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_l + \sum_{j\neq i}p_{ij}\frac{1}{4}\nu\mathbf{a}_j\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_j + p_{il}\frac{1}{4}\nu\mathbf{a}_l\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l - \sum_{l\neq i}p_{il}\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l\mathbf{a}_l\|_2\\ &= \|\sum_{l\neq j\neq i}p_{ijl}\frac{1}{4}\nu\mathbf{a}_l\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_l + \sum_{j\neq i}p_{ij}\frac{1}{4}\nu\mathbf{a}_j\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_j\|_2 \leq O(s^2/(np)).\end{aligned} \tag{36}$$

Next, we bound $\|\boldsymbol{\epsilon}\|_2$. We know that Hessian of sigmoid is bounded (i.e., $\|\nabla^2\sigma(u_t)\|_2 \leq C \approx 0.1$). We denote row t of the matrix $\mathbf{A}$ by $\tilde{\mathbf{a}}_t$.

$$\|\boldsymbol{\epsilon}\|_2 \leq \sum_{t=1}^{n}\|(\tilde{\mathbf{a}}_{t,S}^{\mathrm{T}}\mathbf{x}_S^*)^2\nabla^2\sigma(u_t)\|_2 \leq \sum_{t=1}^{n}\|\tilde{\mathbf{a}}_{t,S}^{\mathrm{T}}\mathbf{x}_S^*\|_2^2\|\nabla^2\sigma(u_t)\|_2 \leq \|\mathbf{A}_S\|_2^2\|\mathbf{x}_S^*\|_2^2\|\nabla^2\sigma(u_t)\|_2 \leq O(C_x^2 s). \tag{37}$$

Following a similar approach, we get

$$\begin{aligned}\|\tilde{\boldsymbol{\epsilon}}\|_2 &\leq \sum_{t=1}^{n}\|[4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\mu-\frac{1}{2})-\mathbf{W}_S\mathbf{b}_S]_t^2\nabla^2\sigma(u_t)\|_2 \leq (\|4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\mu-\frac{1}{2})\|_2^2+\|\mathbf{W}_S\mathbf{b}_S\|_2^2)\|\nabla^2\sigma(u_t)\|_2\\ &\leq O(C\|\mathbf{W}_S\|_2^2\|\mathbf{A}_S\mathbf{x}_S^*+\boldsymbol{\epsilon}\|_2^2) \leq O(C_x^2 s)\end{aligned} \tag{38}$$

So,

$$\|\mathbf{w}_i^{\mathrm{T}}\boldsymbol{\epsilon}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}-\mathbf{I})\mathbf{A}\mathbf{x}^*\|_2 \leq \|\mathbf{w}_i\|_2\|\boldsymbol{\epsilon}\|_2(\|\mathbf{W}_S^{\mathrm{T}}\|_2^2+1)\|\mathbf{A}_S\|_2\|\mathbf{x}_S^*\|_2 \leq O(C_x^3 s\sqrt{s}). \tag{39}$$

Hence,

$$\|E[e_1]\|_2 \leq O(C_x^3 s\sqrt{s}). \tag{40}$$

Similarly, we have $\|E[e_2]\|_2 \leq O(C_x^3 s\sqrt{s})$.

$$\begin{aligned}&\|\left((\mathbf{A}\mathbf{x}^*+\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}}+\mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^*+\boldsymbol{\epsilon})\mathbf{I}\right)((\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}-\mathbf{I})\boldsymbol{\epsilon}+\tilde{\boldsymbol{\epsilon}})\|_2\\ &\leq 2(\|\mathbf{A}_S\|_2\|\mathbf{x}_S^*\|_2+\|\boldsymbol{\epsilon}\|_2)\|\mathbf{w}_i\|_2\left((\|\mathbf{W}_S^{\mathrm{T}}\|_2^2+1)\|\boldsymbol{\epsilon}\|_2+\|\tilde{\boldsymbol{\epsilon}}\|_2\right)\\ &= O((\|\mathbf{x}_S^*\|_2+\|\boldsymbol{\epsilon}\|_2)\|\boldsymbol{\epsilon}\|_2) \leq O(\max(C_x^3 s\sqrt{s}, C_x^4 s^2)).\end{aligned} \tag{41}$$

Hence,

$$\|E[e_3]\|_2 \leq O(\max(C_x^3 s\sqrt{s}, C_x^4 s^2)). \tag{42}$$

We have

$$\|(-b_i)((\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}-\mathbf{I})\boldsymbol{\epsilon}+\tilde{\boldsymbol{\epsilon}})\|_2 \leq O(C_x^2 s) \tag{43}$$

Hence,

$$\|E[e_4]\|_2 \leq O(\max(C_x^2 s). \tag{44}$$

Using the above bounds, we have

$$\|\beta\|_2 \leq O(\max(C_x^3 s\sqrt{s}, C_x^4 s^2)). \tag{45}$$

Using (A14), we get

$$\begin{aligned}\|v\|_2 &= \|\frac{1}{4}p_i(\nu\tau_i(\tau_i-1)+b_i^2)\mathbf{w}_i+\zeta+\beta\|_2 \leq \frac{1}{4}p_i|\nu\tau_i(\tau_i-1)+b_i^2|\|\mathbf{w}_i\|_2+\|\zeta\|_2+\|\beta\|_2\\ &\leq \frac{1}{4}p_i(2\nu\tau_i(1-\tau_i))+\|\zeta\|_2+\|\beta\|_2 \leq \frac{1}{4}p_i\nu\tau_i q\|\mathbf{w}_i-\mathbf{a}_i\|_2+\|\zeta\|_2+\|\beta\|_2\\ &\leq \frac{1}{4}p_i\nu\tau_i q\|\mathbf{w}_i-\mathbf{a}_i\|_2+O(\max(C_x^3 s\sqrt{s}, C_x^4 s^2)).\end{aligned} \tag{46}$$

□

Lemma 1.5. *Suppose the generative model satisfies* $(A1)-(A14)$*. Then*

$$2\langle g_i, \mathbf{w}_i-\mathbf{a}_i\rangle \geq (\frac{1}{4}\nu\tau_i s/p)(1-2q^2)\|\mathbf{w}_i-\mathbf{a}_i\|_2^2+\frac{1}{(\frac{1}{4}\nu\tau_i s/p)}\|g_i\|_2^2-O(C_x^6 p\max(s^2/\tau_i, C_x^2 s^3/\tau_i)). \tag{47}$$

Proof. From Lemma 1.4, we have

$$\|v\|_2 \leq \frac{1}{4}p_i\tau_i\nu q\|\mathbf{w}_i-\mathbf{a}_i\|_2+O(\max(C_x^3 s\sqrt{s}, C_x^4 s^2)). \tag{48}$$

Hence,

$$\|v\|_2^2 \leq 2(\frac{1}{4}\tau_i\nu qs/p)^2\|\mathbf{w}_i-\mathbf{a}_i\|_2^2+O(\max(C_x^6 s^3, C_x^8 s^4)). \tag{49}$$

We have $g_i = \frac{1}{4}p_i\nu\tau_i(\mathbf{w}_i-\mathbf{a}_i)+v$. Taking the norm,

$$\|g_i\|_2^2 = (\frac{1}{4}p_i\nu\tau_i)^2\|\mathbf{w}_i-\mathbf{a}_i\|_2^2+\|v\|_2^2+2(\frac{1}{4}p_i\nu\tau_i)\langle v, \mathbf{w}_i-\mathbf{a}_i\rangle. \tag{50}$$

$$2\langle v, \mathbf{w}_i-\mathbf{a}_i\rangle = -(\frac{1}{4}p_i\nu\tau_i)\|\mathbf{w}_i-\mathbf{a}_i\|_2^2+\frac{1}{(\frac{1}{4}p_i\nu\tau_i)}\|g_i\|_2^2-\frac{1}{(\frac{1}{4}p_i\nu\tau_i)}\|v\|_2^2. \tag{51}$$

$$\begin{aligned} 2\langle g_i, \mathbf{w}_i - \mathbf{a}_i\rangle &= \frac{1}{4}p_i\nu\tau_i\|\mathbf{w}_i-\mathbf{a}_i\|_2^2 + \frac{1}{(\frac{1}{4}p_i\nu\tau_i)}\|g_i\|_2^2 - \frac{1}{(\frac{1}{4}p_i\nu\tau_i)}\|v\|_2^2 \\ &\geq (\frac{1}{4}\nu\tau_i s/p)(1-2q^2)\|\mathbf{w}_i-\mathbf{a}_i\|_2^2 + \frac{1}{(\frac{1}{4}\nu\tau_i s/p)}\|g_i\|_2^2 - O(C_x^6 p\max(s^2/\tau_i, C_x^2 s^3/\tau_i)). \end{aligned} \tag{52}$$

□

Intuitively, Lemma 1.5 suggests that the gradient is approximately along the same direction as $\mathbf{w}_i - \mathbf{a}_i$, so at every iteration of the gradient descent, $\mathbf{w}_i$ gets closer and closer to $\mathbf{a}_i$. Given Lemma 1.5, rigorously, from the descent property of Theorem 6 in [Arora et al., 2015], we can see that given the learning rate $\kappa = \max_i(\frac{1}{\frac{1}{4}\nu\tau_i s/p})$, letting $\delta = \kappa(\frac{1}{4}\nu\tau_i s/p)(1-2q^2) \in (0,1)$, we have the descent property as follows

$$\|\mathbf{w}_i^{(l+1)} - \mathbf{a}_i\|_2^2 \leq (1-\delta)\|\mathbf{w}_i^{(l)} - \mathbf{a}_i\|_2^2 + \kappa \cdot O(C_x^6 p\max(s^2/\tau_i, C_x^2 s^3/\tau_i)). \tag{53}$$

Lemma 1.6. *Suppose* $\|\mathbf{w}_i^{(l+1)} - \mathbf{a}_i\|_2^2 \leq (1-\delta)\|\mathbf{w}_i^{(l)} - \mathbf{a}_i\|_2^2 + \kappa \cdot O(C_x^6 p\max(s^2/\tau_i, C_x^2 s^3/\tau_i))$ *where* $\delta = \kappa(\frac{1}{4}\nu\tau_i s/p)(1-2q^2) \in (0,1)$ *and* $O(\frac{C_x^4 p^2\max(s, C_x^4 s^2)}{\tau_i^2(1-2q^2)}) < \|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2$. *Then*

$$\|\mathbf{w}_i^{(L)} - \mathbf{a}_i\|_2^2 \leq (1-\delta/2)^L\|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2. \tag{54}$$

Proof. Performing the gradient update L times,

$$\begin{aligned} \|\mathbf{w}_i^{(L)} - \mathbf{a}_i\|_2^2 &\leq (1-\delta)^L\|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2 + \frac{1}{(\frac{1}{4}\nu\tau_i s/p)(1-2q^2)}O(C_x^6 p\max(s^2/\tau_i, C_x^2 s^3/\tau_i)) \\ &\leq (1-\delta)^L\|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2 + O(\frac{C_x^6 p^2\max(s, C_x^2 s^2)}{\tau_i^2(1-2q^2)}). \end{aligned} \tag{55}$$

From Theorem 6 in [Arora et al., 2015], if $O(\frac{C_x^6 p^2\max(s, C_x^2 s^2)}{\tau_i^2(1-2q^2)}) < \|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2$, then we have

$$\|\mathbf{w}_i^{(L)} - \mathbf{a}_i\|_2^2 \leq (1-\delta/2)^L\|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2. \tag{56}$$

□

Corollary 1.6.1. *Given (A2), the condition of Lemma 1.6 is simplified to* $O(\frac{\max(s, s^2/p^{\frac{2}{3}+2\xi})}{p^{6\xi}\tau_i^2(1-2q^2)}) < \|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2$.

The intuition behind the bound on the amplitude of $\mathbf{x}^*$ in (A2) is that as C_x gets smaller, the range of $\sigma(\mathbf{A}\mathbf{x}^*)$ is concentrated around the linear region of the sigmoid function (i.e., around $\sigma(\mathbf{0})$); thus $\boldsymbol{\epsilon}$, which is the difference between $\sigma(\mathbf{A}\mathbf{x}^*)$ and the linear region of sigmoid $\frac{1}{2} + \frac{1}{4}\mathbf{A}\mathbf{x}^*$, is smaller. Hence, the upper bound on $\|v\|_2$ would be smaller and $O(\frac{\max(s, s^2/p^{\frac{2}{3}+2\xi})}{p^{6\xi}\tau_i^2(1-2q^2)})$ would get smaller.

1.4 Sparse coding with $\mathcal{S}_{\mathbf{b}} = \mathrm{HT}_{\mathbf{b}}$

1.4.1 Gradient derivation

This is section, we derive g_i for the case when $\mathcal{S}_{\mathbf{b}} = \mathrm{HT}_{\mathbf{b}}$ following a similar approach to the previous section.

$$\begin{aligned} \lim_{M\to\infty}\frac{\partial\mathcal{L}_{\mathbf{W}}}{\partial\mathbf{w}_i} &= \lim_{M\to\infty}\frac{\partial\mathbf{c}_1}{\partial\mathbf{w}_i}\frac{\partial\mathcal{L}_{\mathbf{W}}}{\partial\mathbf{c}_1} + \frac{\partial\mathbf{c}_2}{\partial\mathbf{w}_i}\frac{\partial\mathcal{L}_{\mathbf{W}}}{\partial\mathbf{c}_2} = \frac{\partial\mathbf{c}_1}{\partial\mathbf{w}_i}\frac{\partial\mathbf{x}}{\partial\mathbf{c}_1}\frac{\partial\mathbf{c}_2}{\partial\mathbf{x}}\frac{\mathcal{L}_{\mathbf{W}}}{\partial\mathbf{c}_2} + \frac{\partial\mathbf{c}_2}{\partial\mathbf{w}_i}\frac{\partial\mathcal{L}_{\mathbf{W}}}{\partial\mathbf{c}_2} \\ &= \left(\underbrace{[0,0,\dots,4(\boldsymbol{\mu}-\frac{1}{2}),\dots,0]}_{n\times p}\underbrace{\mathrm{diag}(\mathcal{S}'_{\mathbf{b}}(\mathbf{c}_1))}_{p\times p}\mathbf{W}^{\mathrm{T}} + \mathbb{1}_{\mathbf{x}_i\neq 0}\mathbf{w}_i^{\mathrm{T}}4(\boldsymbol{\mu}-\frac{1}{2})\mathbf{I}\right)\left(-\sigma(\mathbf{A}\mathbf{x}^*) + \sigma(4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu}-\frac{1}{2}))\right) \\ &= \left(\mathcal{S}'_{\mathbf{b}}(\mathbf{c}_{1,i})4(\boldsymbol{\mu}-\frac{1}{2})\mathbf{w}_i^{\mathrm{T}} + \mathbb{1}_{\mathbf{x}_i\neq 0}\mathbf{w}_i^{\mathrm{T}}4(\boldsymbol{\mu}-\frac{1}{2})\mathbf{I}\right)\left(\sigma(4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu}-\frac{1}{2})) - \sigma(\mathbf{A}\mathbf{x}^*)\right). \end{aligned} \tag{57}$$

We further expand the gradient, by replacing $\sigma(.)$ with its Taylor expansion. We have

$$\sigma(\mathbf{A}\mathbf{x}^*) = \frac{1}{2} + \frac{1}{4}\mathbf{A}\mathbf{x}^* + \boldsymbol{\epsilon}, \tag{58}$$

where $\boldsymbol{\epsilon} = [\epsilon_1,\dots,\epsilon_n]^{\mathrm{T}}$, $\epsilon_d = \nabla^2\sigma(u_d)([\mathbf{A}\mathbf{x}^*]_d)^2$, and $0 \leq u_d \leq [\mathbf{A}\mathbf{x}^*]_d$. Similarly,

$$\sigma(4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu}-\frac{1}{2})) = \frac{1}{2} + \mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu}-\frac{1}{2}) + \tilde{\boldsymbol{\epsilon}}, \tag{59}$$

where $\tilde{\boldsymbol{\epsilon}} = [\tilde{\epsilon}_1,\dots,\tilde{\epsilon}_n]^{\mathrm{T}}$, $\tilde{\epsilon}_d = \nabla^2\sigma(\tilde{u}_d)([4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu}-\frac{1}{2})]_d)^2$, and $0 \leq \tilde{u}_d \leq [4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu}-\frac{1}{2})]_d$. Again, replacing $\boldsymbol{\mu}$ with Taylor expansion of $\sigma(\mathbf{A}\mathbf{x}^*)$, we get

$$\sigma(4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\boldsymbol{\mu}-\frac{1}{2})) = \frac{1}{2} + \mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\frac{1}{4}\mathbf{A}\mathbf{x}^* + \boldsymbol{\epsilon}) + \tilde{\boldsymbol{\epsilon}}. \tag{60}$$

By symmetry, $E[\boldsymbol{\epsilon} \mid S] = E[\tilde{\boldsymbol{\epsilon}} \mid S] = 0$. The expectation of gradient g_i would be

$$g_i = E[\big(\mathbb{1}_{\mathbf{x}_i\neq 0}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}} + \mathbb{1}_{\mathbf{x}_i\neq 0}\mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{I}\big)\,(\frac{1}{4}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})(\mathbf{A}\mathbf{x}^*) + (\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\boldsymbol{\epsilon} + \tilde{\boldsymbol{\epsilon}})]. \tag{61}$$

1.4.2 Gradient dynamics

Given the code consistency from the forward pass of the encoder, we replace $\mathbb{1}_{\mathbf{x}_i \neq 0}$ with $\mathbb{1}_{\mathbf{x}_i^* \neq 0}$ and denote the error by γ as below which is small for large p [Nguyen et al., 2019].

$$\gamma = E[(\mathbb{1}_{\mathbf{x}_i^* \neq 0} - \mathbb{1}_{\mathbf{x}_i \neq 0}) \big((\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}} + \mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{I}\big) (\frac{1}{4}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})(\mathbf{A}\mathbf{x}^*) + (\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\boldsymbol{\epsilon} + \tilde{\boldsymbol{\epsilon}})]. \tag{62}$$

Now, we write g_i as

$$g_i = E[\mathbb{1}_{\mathbf{x}_i^* \neq 0} \big((\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}} + \mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{I}\big) (\frac{1}{4}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})(\mathbf{A}\mathbf{x}^*) + (\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\boldsymbol{\epsilon} + \tilde{\boldsymbol{\epsilon}})] + \gamma. \tag{63}$$

We can see that if $i \notin S$ then $\mathbb{1}_{\mathbf{x}_i^*} = 0$ hence, $g_i = 0$. Thus, in our analysis, we only consider the case $i \in S$. We decompose g_i as below.

$$g_i = g_i^{(1)} + g_i^{(2)} + g_i^{(3)} + \gamma, \tag{64}$$

where

$$g_i^{(1)} = E[\frac{1}{4}\mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^*]. \tag{65}$$

$$g_i^{(2)} = E[\frac{1}{4}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^*]. \tag{66}$$

$$g_i^{(3)} = E[\big((\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}} + \mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{I}\big) ((\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\boldsymbol{\epsilon} + \tilde{\boldsymbol{\epsilon}})]. \tag{67}$$

We define

$$g_{i,S}^{(1)} = E[\frac{1}{4}\mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^* \mid S]. \tag{68}$$

$$g_{i,S}^{(2)} = E[\frac{1}{4}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^* \mid S]. \tag{69}$$

$$g_{i,S}^{(3)} = E[\big((\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}} + \mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})\mathbf{I}\big) ((\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\boldsymbol{\epsilon} + \tilde{\boldsymbol{\epsilon}}) \mid S]. \tag{70}$$

Hence, $g_i^{(k)} = E[g_{i,S}^{(k)}]$ for $k = 1, \ldots, 3$ where the expectations are with respect to the support S.

$$\begin{aligned} g_{i,S}^{(1)} &= E[\frac{1}{4}\mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + 4\boldsymbol{\epsilon})(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^* \mid S] \\ &= \sum_{j,l \in S} E[\frac{1}{4}\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_j\mathbf{x}_j^*(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{a}_l\mathbf{x}_l^* \mid S] + E[\mathbf{w}_i^{\mathrm{T}}\boldsymbol{\epsilon}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^* \mid S] \\ &= \frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_i(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{a}_i + \sum_{l \in S^{\setminus i}} \frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{a}_l + e_1 = \frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_i(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{a}_i + r_1 + e_1. \end{aligned} \tag{71}$$

We denote $r_1 = \sum_{l \in S^{\setminus i}} \frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{a}_l$ and $e_1 = E[\mathbf{w}_i^{\mathrm{T}}\boldsymbol{\epsilon}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^* \mid S]$. Similarly for $g_{i,S}^{(2)}$, we have

$$g_{i,S}^{(2)} = \frac{1}{4}\nu\mathbf{a}_i\mathbf{w}_i^{\mathrm{T}}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{a}_i + r_2 + e_2. \tag{72}$$

We denote $r_2 = \sum_{l \in S^{\setminus i}} \frac{1}{4}\nu\mathbf{a}_l\mathbf{w}_i^{\mathrm{T}}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{a}_l$, $e_2 = E[\boldsymbol{\epsilon}\mathbf{w}_i^{\mathrm{T}}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{A}\mathbf{x}^* \mid S]$, and $e_3 = g_{i,S}^{(3)}$. We also denote $\beta = E[r_1 + r_2 + e_1 + e_2 + e_3] + \gamma$. Combining the terms,

$$\begin{aligned} g_i &= E[\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_i(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{a}_i + \frac{1}{4}\nu\mathbf{a}_i\mathbf{w}_i^{\mathrm{T}}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}} - \mathbf{I})\mathbf{a}_i] + \beta \\ &= E[-\frac{1}{2}\nu\tau_i\mathbf{a}_i + \frac{1}{4}\nu\tau_i\sum_{j \in S}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_i + \frac{1}{4}\nu\mathbf{a}_i\mathbf{w}_i^{\mathrm{T}}\sum_{j \in S}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_i] + \beta \\ &= E[-\frac{1}{2}\nu\tau_i\mathbf{a}_i + \frac{1}{4}\nu\tau_i^2\mathbf{w}_i + \frac{1}{4}\nu\tau_i\sum_{j \in S^{\setminus i}}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_i + \frac{1}{4}\nu\tau_i\|\mathbf{w}_i\|_2^2\mathbf{a}_i + \frac{1}{4}\nu\left(\mathbf{a}_i\mathbf{w}_i^{\mathrm{T}}\right)\sum_{j \in S^{\setminus i}}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_i] + \beta \\ &= -\frac{1}{4}p_i\nu\tau_i\mathbf{a}_i + p_i\frac{1}{4}\nu\tau_i^2\mathbf{w}_i + \zeta + \beta, \end{aligned} \tag{73}$$

where $\zeta = \sum_{j \in [p]^{\setminus i}} \frac{1}{4}p_{ij}\nu\tau_i\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_i + \frac{1}{4}p_{ij}\nu\mathbf{a}_i\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_i$. We continue

$$g_i = \frac{1}{4}p_i\nu\tau_i(\mathbf{w}_i - \mathbf{a}_i) + v, \tag{74}$$

where we denote $v = \frac{1}{4}p_i\nu\tau_i(\tau_i - 1)\mathbf{w}_i + \zeta + \beta$.

Lemma 1.7. *Suppose the generative model satisfies* $(A1) - (A13)$. *Then*

$$\|v\|_2 \leq \frac{1}{8}p_i\nu\tau_i q\|\mathbf{w}_i - \mathbf{a}_i\|_2 + O(\max(C_x^3 s\sqrt{s}, C_x^4 s^2)) \tag{75}$$

Proof.

$$\begin{aligned} \|\zeta\|_2 &= \|\sum_{j \in [p]^{\setminus i}} \frac{1}{4}p_{ij}\nu\tau_i\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_i + \frac{1}{4}p_{ij}\nu\mathbf{a}_i\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_i\|_2 = \|\frac{1}{4}p_{ij}\nu\tau_i\mathbf{W}_{\setminus i}\mathbf{W}_{\setminus i}^{\mathrm{T}}\mathbf{a}_i + \frac{1}{4}p_{ij}\nu\mathbf{a}_i\mathbf{w}_i^{\mathrm{T}}\mathbf{W}_{\setminus i}\mathbf{W}_{\setminus i}^{\mathrm{T}}\mathbf{a}_i\|_2 \\ &\leq \frac{1}{4}p_{ij}\nu\tau_i\|\mathbf{W}_{\setminus i}\|_2^2\|\mathbf{a}_i\|_2 + \frac{1}{4}p_{ij}\nu\|\mathbf{w}_i\|_2\|\mathbf{W}_{\setminus i}\|_2^2\|\mathbf{a}_i\|_2^2 = \frac{1}{4}O(\nu\tau_i s^2/(np)) + \frac{1}{4}O(\nu s^2/(np)) = O(s^2/(np)). \end{aligned} \tag{76}$$

$$
\begin{aligned}
\|E[r_1]\|_2 &= \|E[\sum_{l\in S\setminus i}\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}-\mathbf{I})\mathbf{a}_l]\|_2 = \|E[\sum_{l\in S\setminus i}\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}\mathbf{a}_l - \sum_{l\in S\setminus i}\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l\mathbf{a}_l]\|_2 \\
&= \|\sum_{l\neq j\neq i} p_{ijl}\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_l + \sum_{j\neq i}p_{ij}\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_j\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_j + p_{il}\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l - \sum_{l\neq i}p_{il}\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l\mathbf{a}_l\|_2 \leq O(s^2/(np)).
\end{aligned} \tag{77}
$$

where each terms is bounded as below

$$
\|\sum_{l\neq j\neq i} p_{ijl}\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_l\|_2 = \frac{1}{4}\nu(s^3/p^3)\|\mathbf{W}_{\setminus i}\|_2 \leq O(s^3/(pn\sqrt{n})). \tag{78}
$$

where $z_j = \sum_{l\neq i,j}\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_l$, hence, $\|z\|_2 \leq O(\frac{p\sqrt{p}}{n})$.

$$
\|\sum_{j\neq i}p_{ij}\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_j\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_j\|_2 = \|\frac{1}{4}\nu(s^2/p^2)\|\mathbf{W}_{\setminus i}z\|_2 \leq O(s^2/(np)). \tag{79}
$$

where $z_j = \mathbf{w}_i^{\mathrm{T}}\mathbf{a}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_j$, hence, $\|z\|_2 \leq O(\frac{\sqrt{p}}{\sqrt{n}})$.

$$
\|\sum_{l\neq j\neq i} p_{ijl}\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_l\|_2 = \frac{1}{4}\nu(s^3/p^3)\|\mathbf{W}_{\setminus i}\|_2 \leq O(s^3/(pn\sqrt{n})). \tag{80}
$$

$$
\|p_{il}\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l\|_2 \leq O(s^2/(np^2)). \tag{81}
$$

Following a similar approach for r_2, we get

$$
\begin{aligned}
\|E[r_2]\|_2 &= \|E[\sum_{l\in S\setminus i}\frac{1}{4}\nu\mathbf{a}_l\mathbf{w}_i^{\mathrm{T}}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}-\mathbf{I})\mathbf{a}_l]\|_2 = \|E[\sum_{l\in S\setminus i}\frac{1}{4}\nu\mathbf{a}_l\mathbf{w}_i^{\mathrm{T}}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}})\mathbf{a}_l - \sum_{l\in S\setminus i}\frac{1}{4}\nu\mathbf{a}_l\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l]\|_2 \\
&= \|\sum_{l\neq j\neq i} p_{ijl}\frac{1}{4}\nu\mathbf{a}_l\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_l + \sum_{j\neq i}p_{ij}\frac{1}{4}\nu\mathbf{a}_j\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_j + p_{il}\frac{1}{4}\nu\mathbf{a}_l\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l - \sum_{l\neq i}p_{il}\frac{1}{4}\nu\mathbf{w}_i^{\mathrm{T}}\mathbf{a}_l\mathbf{a}_l\|_2 \\
&= \|\sum_{l\neq j\neq i} p_{ijl}\frac{1}{4}\nu\mathbf{a}_l\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_l + \sum_{j\neq i}p_{ij}\frac{1}{4}\nu\mathbf{a}_j\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j\mathbf{w}_j^{\mathrm{T}}\mathbf{a}_j\|_2 \leq O(s^2/(np)).
\end{aligned} \tag{82}
$$

Next, we bound $\|\boldsymbol{\epsilon}\|_2$. We know that Hessian of sigmoid is bounded (i.e., $\|\nabla^2\sigma(u_t)\|_2 \leq C \approx 0.1$). We denote row t of the matrix $\mathbf{A}$ by $\tilde{\mathbf{a}}_t$.

$$
\|\boldsymbol{\epsilon}\|_2 \leq \sum_{t=1}^{n}\|(\tilde{\mathbf{a}}_{t,S}^{\mathrm{T}}\mathbf{x}_S^*)^2\nabla^2\sigma(u_t)\|_2 \leq \sum_{t=1}^{n}\|\tilde{\mathbf{a}}_{t,S}^{\mathrm{T}}\mathbf{x}_S^*\|_2^2\|\nabla^2\sigma(u_t)\|_2 \leq \|\mathbf{A}_S\|_2^2\|\mathbf{x}_S^*\|_2^2\|\nabla^2\sigma(u_t)\|_2 \leq O(C_x^2 s). \tag{83}
$$

Following a similar approach, we get

$$
\begin{aligned}
\|\tilde{\boldsymbol{\epsilon}}\|_2 &\leq \sum_{t=1}^{n}\|[4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\mu-\frac{1}{2})]_t^2\nabla^2\sigma(u_t)\|_2 \leq \|4\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}(\mu-\frac{1}{2})\|_2^2\|\nabla^2\sigma(u_t)\|_2 \\
&\leq O(C\|\mathbf{W}_S\|_2^2\|\mathbf{A}_S\mathbf{x}_S^* + \boldsymbol{\epsilon}\|_2^2) \leq O(C_x^2 s)
\end{aligned} \tag{84}
$$

So,

$$
\|\mathbf{w}_i^{\mathrm{T}}\boldsymbol{\epsilon}(\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}-\mathbf{I})\mathbf{A}\mathbf{x}^*\|_2 \leq \|\mathbf{w}_i\|_2\|\boldsymbol{\epsilon}\|_2(\|\mathbf{W}_S^{\mathrm{T}}\|_2^2+1)\|\mathbf{A}_S\|_2\|\mathbf{x}_S^*\|_2 \leq O(C_x^3 s\sqrt{s}). \tag{85}
$$

Hence,

$$
\|E[e_1]\|_2 \leq O(C_x^3 s\sqrt{s}). \tag{86}
$$

Similarly, we have $\|E[e_2]\|_2 \leq O(C_x^3 s\sqrt{s})$.

$$
\begin{aligned}
&\|\left((\mathbf{A}\mathbf{x}^* + \boldsymbol{\epsilon})\mathbf{w}_i^{\mathrm{T}} + \mathbf{w}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^* + \boldsymbol{\epsilon})\mathbf{I}\right)((\mathbf{W}_S\mathbf{W}_S^{\mathrm{T}}-\mathbf{I})\boldsymbol{\epsilon} + \tilde{\boldsymbol{\epsilon}})\|_2 \\
&\leq 2(\|\mathbf{A}_S\|_2\|\mathbf{x}_S^*\|_2 + \|\boldsymbol{\epsilon}\|_2)\|\mathbf{w}_i\|_2\left((\|\mathbf{W}_S^{\mathrm{T}}\|_2^2+1)\|\boldsymbol{\epsilon}\|_2 + \|\tilde{\boldsymbol{\epsilon}}\|_2\right) \\
&= O((\|\mathbf{x}_S^*\|_2 + \|\boldsymbol{\epsilon}\|_2)\|\boldsymbol{\epsilon}\|_2) \leq O(\max(C_x^3 s\sqrt{s}, C_x^4 s^2)).
\end{aligned} \tag{87}
$$

Hence,

$$
\|E[e_3]\|_2 \leq O(\max(C_x^3 s\sqrt{s}, C_x^4 s^2)). \tag{88}
$$

Using the above bounds, we have

$$
\|\beta\|_2 \leq O(\max(C_x^3 s\sqrt{s}, C_x^4 s^2)). \tag{89}
$$

Hence,

$$
\begin{aligned}
\|v\|_2 &= \|\frac{1}{4}p_i\nu\tau_i(\tau_i-1)\mathbf{w}_i + \zeta + \beta\|_2 \leq \frac{1}{4}p_i\nu\tau_i|(\tau_i-1)|\|\mathbf{w}_i\|_2 + \|\zeta\|_2 + \|\beta\|_2 \\
&\leq \frac{1}{4}p_i\nu\tau_i(\frac{1}{2}q\|\mathbf{w}_i-\mathbf{a}_i\|_2) + \|\zeta\|_2 + \|\beta\|_2 \leq \frac{1}{8}p_i\nu\tau_i q\|\mathbf{w}_i-\mathbf{a}_i\|_2 + O(\max(C_x^3 s\sqrt{s}, C_x^4 s^2)).
\end{aligned} \tag{90}
$$

□

Lemma 1.8. *Suppose the generative model satisfies* $(A1)-(A13)$*. Then*

$$2\langle g_i, \mathbf{w}_i - \mathbf{a}_i\rangle \geq (\frac{1}{4}\nu\tau_i s/p)(1-\frac{q^2}{2})\|\mathbf{w}_i - \mathbf{a}_i\|_2^2 + \frac{1}{(\frac{1}{4}\nu\tau_i s/p)}\|g_i\|_2^2 - O(C_x^6 p \max(s^2/\tau_i, C_x^2 s^3/\tau_i)). \tag{91}$$

Proof. From Lemma 1.7, we have

$$\|v\|_2 \leq \frac{1}{8}p_i\nu\tau_i q\|\mathbf{w}_i - \mathbf{a}_i\|_2 + O(\max(C_x^3 s\sqrt{s}, C_x^4 s^2)). \tag{92}$$

Hence,

$$\|v\|_2^2 \leq 2(\frac{1}{8}\tau_i\nu q s/p)^2\|\mathbf{w}_i - \mathbf{a}_i\|_2^2 + O(\max(C_x^6 s^3, C_x^8 s^4)). \tag{93}$$

We have $g_i = \frac{1}{4}p_i\nu\tau_i(\mathbf{w}_i - \mathbf{a}_i) + v$. Taking the norm,

$$\|g_i\|_2^2 = (\frac{1}{4}p_i\nu\tau_i)^2\|\mathbf{w}_i - \mathbf{a}_i\|_2^2 + \|v\|_2^2 + 2(\frac{1}{4}p_i\nu\tau_i)\langle v, \mathbf{w}_i - \mathbf{a}_i\rangle. \tag{94}$$

$$2\langle v, \mathbf{w}_i - \mathbf{a}_i\rangle = -(\frac{1}{4}p_i\nu\tau_i)\|\mathbf{w}_i - \mathbf{a}_i\|_2^2 + \frac{1}{(\frac{1}{4}p_i\nu\tau_i)}\|g_i\|_2^2 - \frac{1}{(\frac{1}{4}p_i\nu\tau_i)}\|v\|_2^2. \tag{95}$$

$$\begin{aligned} 2\langle g_i, \mathbf{w}_i - \mathbf{a}_i\rangle &= \frac{1}{4}p_i\nu\tau_i\|\mathbf{w}_i - \mathbf{a}_i\|_2^2 + \frac{1}{(\frac{1}{4}p_i\nu\tau_i)}\|g_i\|_2^2 - \frac{1}{(\frac{1}{4}p_i\nu\tau_i)}\|v\|_2^2 \\ &\geq (\frac{1}{4}\nu\tau_i s/p)(1-\frac{q^2}{2})\|\mathbf{w}_i - \mathbf{a}_i\|_2^2 + \frac{1}{(\frac{1}{4}\nu\tau_i s/p)}\|g_i\|_2^2 - O(C_x^6 p\max(s^2/\tau_i, C_x^2 s^3/\tau_i)). \end{aligned} \tag{96}$$

□

Lemma 1.8 suggests that the gradient is approximately along the same direction as $\mathbf{w}_i - \mathbf{a}_i$, so at every iteration of the gradient descent, $\mathbf{w}_i$ gets closer and closer to $\mathbf{a}_i$. Given Lemma 1.8 from the descent property of Theorem 6 in [Arora et al., 2015], we can see that given the learning rate $\kappa = \max_i(\frac{1}{\frac{1}{4}\nu\tau_i s/p})$, letting $\delta = \kappa(\frac{1}{4}\nu\tau_i s/p)(1-\frac{q^2}{2}) \in (0,1)$, we have the descent property as follows

$$\|\mathbf{w}_i^{(l+1)} - \mathbf{a}_i\|_2^2 \leq (1-\delta)\|\mathbf{w}_i^{(l)} - \mathbf{a}_i\|_2^2 + \kappa\cdot O(C_x^6 p\max(s^2/\tau_i, C_x^2 s^3/\tau_i)). \tag{97}$$

Lemma 1.9. *Suppose* $\|\mathbf{w}_i^{(l+1)} - \mathbf{a}_i\|_2^2 \leq (1-\delta)\|\mathbf{w}_i^{(l)} - \mathbf{a}_i\|_2^2 + \kappa\cdot O(C_x^6 p\max(s^2/\tau_i, C_x^2 s^3/\tau_i))$ *where* $\delta = \kappa(\frac{1}{4}\nu\tau_i s/p)(1-\frac{q^2}{2}) \in (0,1)$ *and* $O(\frac{C_x^4 p^2\max(s, C_x^4 s^2)}{\tau_i^2(1-\frac{q^2}{2})}) < \|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2$*. Then*

$$\|\mathbf{w}_i^{(L)} - \mathbf{a}_i\|_2^2 \leq (1-\delta/2)^L\|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2. \tag{98}$$

Proof. Performing the gradient update L times,

$$\begin{aligned} \|\mathbf{w}_i^{(L)} - \mathbf{a}_i\|_2^2 &\leq (1-\delta)^L\|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2 + \frac{1}{(\frac{1}{4}\nu\tau_i s/p)(1-\frac{q^2}{2})}O(C_x^6 p\max(s^2/\tau_i, C_x^2 s^3/\tau_i)) \\ &\leq (1-\delta)^L\|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2 + O(\frac{C_x^6 p^2\max(s, C_x^2 s^2)}{\tau_i^2(1-\frac{q^2}{2})}). \end{aligned} \tag{99}$$

From Theorem 6 in [Arora et al., 2015], if $O(\frac{C_x^6 p^2\max(s, C_x^2 s^2)}{\tau_i^2(1-\frac{q^2}{2})}) < \|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2$, then we have

$$\|\mathbf{w}_i^{(L)} - \mathbf{a}_i\|_2^2 \leq (1-\delta/2)^L\|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2. \tag{100}$$

□

Corollary 1.9.1. *Given (A2), the condition of Lemma 1.9 is simplified to* $O(\frac{\max(s, s^2/p^{\frac{2}{3}+2\xi})}{p^{6\xi}\tau_i^2(1-\frac{q^2}{2})}) < \|\mathbf{w}_i^{(0)} - \mathbf{a}_i\|_2^2$.

2 BCOMP algorithm

We implement binomial convolutional orthogonal matching pursuit (BCOMP) as a baseline for ECDL task, as mentioned in the Experiments section. BCOMP solves Eq. (2) with ℓ_0 psuedo-norm $\|\mathbf{x}^j\|_0$, instead of $\|\mathbf{x}^j\|_1$, and combines the idea of convolutional greedy pursuit [Mailhé et al., 2011] and binomial greedy pursuit [Vincent and Bengio, 2002, Lozano et al., 2011]. BCOMP is a computationally efficient algorithm for ECDL, as 1) the greedy algorithms are generally considered faster than algorithms for ℓ_1-regularized problems [Tropp and Gilbert, 2007] and 2) it exploits the localized nature of $\mathbf{h}_c$ to speed up the computation of both CSC and CDU steps.

The superscript g refers to one iteration of the the alternating-minimization procedure, for $g = 1, \cdots, G$. We assume sparsity level of T for BCOMP, which means that there are at most T non-zeros values for $\mathbf{x}^j$, set differently according to the application. The subscript t refers to a single iteration of the CSC step, where additional support for $\mathbf{x}^j$ is identified. The set $\mathcal{R}_t$ contains indices of the columns from $\mathbf{H}$ that were chosen up to iteration t. The notation $\mathbf{H}_i$ refers to the i^{th} column of $\mathbf{H}$. The index $n_{c,i}^j$ denotes the occurrence of the i^{th} event from filter c (the nonzero entries of $\mathbf{x}^j$ corresponding to filter c) in the j^{th} observation. The optimization problems in line 10 and 17 are both constrained convex optimization problems that can be

Algorithm 2: ECDL by BCOMP

```
Input: $\{\mathbf{y}^j\}_{j=1}^J \in \mathbb{R}^N$, $\{\mathbf{h}_c^{(0)}\}_{c=1}^C \in \mathbb{R}^K$
Output: $\{\mathbf{x}^{j,(G)}\}_{j=1}^J \in \mathbb{R}^{C(N-K+1)}$, $\{\mathbf{h}_c^{(G)}\}_{c=1}^C \in \mathbb{R}^K$
for $g = 1$ to $G$ do
    ($CSC$ step)
    for $j = 1$ to $J$ do
        $\mathcal{R}_0 = \emptyset$, $\mathbf{x}_1^{j,(g-1)} = \mathbf{0}$
        for $t = 1$ to $T$ do
            $\widetilde{\mathbf{y}}_t^j = \mathbf{y}_t^j - f^{-1}\big(\mathbf{H}^{(g-1)}\mathbf{x}_{t-1}^{j,(g-1)}\big)$
            $c^*, n^* = \arg\max_{c,n}\{(\mathbf{h}_c^{(g-1)} \star \widetilde{\mathbf{y}}_t^j)[n]\}_{c,n=1}^{C,N-K+1}$
            $i = c^*(N-K+1) + n^*$
            $\mathcal{R}_t = \mathcal{R}_{t-1} \cup \mathbf{H}_i^{(g-1)}$
            $\mathbf{x}_t^{j,(g)} = \arg\min_{\mathbf{x}^j} -\log p(\mathbf{y}^j|\{\mathbf{h}_c^{(g-1)}\}_{c=1}^C, \mathbf{x}^j)$, s.t. $\begin{cases} \mathbf{x}^j[n] \geq 0 \text{ for } n \in \mathcal{R}_t \\ \mathbf{x}^j[n] = 0 \text{ for } n \notin \mathcal{R}_t \end{cases}$

    ($CDU$ step)
    for $j = 1$ to $J$ do
        for $c = 1$ to $C$ do
            for $i = 1$ to $N_c^j$ do
                $\mathbf{X}_{c,i}^{j,(g)} = \begin{pmatrix} \mathbf{0}_{n_{c,i}^j \times K} & \mathbf{x}^{c,(g)}[n_{c,i}^j] \cdot \mathbf{I}_{K\times K} & \mathbf{0}_{(N-K-n_{c,i}^j)\times K} \end{pmatrix}^{\mathrm{T}}$
    $\{\mathbf{h}_c^{(g)}\}_{c=1}^C = \arg\min_{\{\mathbf{h}_c\}_{c=1}^C} -\sum_{j=1}^J \log p(\mathbf{y}^j|\{\mathbf{h}_c\}_{c=1}^C, \{\mathbf{X}_{c,i}^{j,(g)}\}_{c,i,j=1})$, s.t. $||\mathbf{h}_c||_2 = 1$
```

solved using standard convex programming packages.

We found that BCOMP converged in $G = 5$ alternating-minimization iterations in the simulations, and $G = 10$ iterations in the analyses of the real data. After convergence, the CSC step of the BCOMP can be used for inference on the test dataset, similar to using the encoder of DCEA for inference. Algorithm 3 shows the forward pass of the DCEA architecture. For notational convenience, we have dropped the superscript j indexing the J inputs.

Algorithm 3: DCEA($\mathbf{y}, \mathbf{h}, b$): Forward pass of DCEA architecture.

```
Input: $\mathbf{y}, \mathbf{h}, b, \alpha$
Output: $\mathbf{w}$
$\mathbf{x}_0 = \mathbf{0}$
for $t = 1$ to $T$ do
    $\mathbf{x}_t = \mathcal{S}_{\mathbf{b}}\left(\mathbf{x}_{t-1} + \alpha\mathbf{H}^T\big(\mathbf{y} - f^{-1}\big(\mathbf{H}\mathbf{x}_{t-1}\big)\big)\right)$
$\mathbf{w} = \mathbf{H}\mathbf{x}_T$
```

3 DCEA architecture

Implementation of the DCEA encoder We implemented the DCEA architecture in PyTorch. In the case of 1D, we accelerate the computations performed by the DCEA encoder by replacing ISTA with its faster version FISTA [Beck and Teboulle, 2009]. FISTA uses a momentum term to accelerate the converge of ISTA. The resulting encoder is similar to the one from [Tolooshams et al., 2020]. We trained it using backpropagation with the ADAM optimizer [Kingma and Ba, 2014], on an Nvidia GPU (GeForce GTX 1060).

Hyperparameters used for training the DCEA architecture in using the simulated and real neural spiking data In these experiments, we treat λ as hyperparameter where $b = \alpha\lambda$. λ is tuned by grid search in the interval of $[0.1, 1.5]$. Following the grid search, we used $\lambda = 0.38$ in the simulations and $\lambda = 0.12$ for the real data. The DCEA encoder performs $T = 250$ and $T = 5{,}000$ iterations of FISTA, respectively for the simulated and for the real data. We found that such large numbers, particularly for the real data, were necessary for the encoder to produce sparse codes. We used $\alpha = 0.2$ in the simulations and $\alpha = 0.5$ for the real data. We used batches of size 256 neurons in the simulations, and a single neuron per batch in the analyses of the real data.

Processing of the output of the DCEA encoder after training in neural spiking experiment The encoder of the DCEA architecture performs ℓ_1-regularized logistic regression using the convolutional dictionary $\mathbf{H}$, the entries of which are highly correlated because of the convolutional structure. Suppose a binomial observation $\mathbf{y}^j$ is generated according to the binomial generative model with mean of $\boldsymbol{\mu}_j = f^{-1}\big(\mathbf{H}\mathbf{x}^j\big)$, where $f^{-1}(\cdot)$ is a sigmoid function. We observed that the estimate $\mathbf{x}_T^j$ of $\mathbf{x}^j$ obtained by feeding the group of observations to the DCEA encoder is a vector whose nonzero entries are clustered around those of $\mathbf{x}^j$. This is depicted in black in Fig. 1, and is a well-known issue with ℓ_1-regularized regression with correlated dictionaries [Bhaskar et al., 2013]. Therefore, for the neural spiking data, after training the DCEA architecture, we processed the output of the encoder as follows

1. Clustering: We applied k-means clustering to $\mathbf{x}_T^j$ to identify 16 clusters.
2. Support identification: For each cluster, we identified the index of the largest entry from $\mathbf{x}_T^j$ in the cluster. This yielded a set of indices that correspond to the estimated support of $\mathbf{x}^j$.
3. Logistic regression: We performed logistic regression using the group of observations and $\mathbf{H}$ restricted to the support identified in the previous step. Note that this is a common procedure for ℓ_1-regularized problems [Tang et al., 2013, Mardani et al., 2018]. This yielded a new set of codes $\mathbf{x}^j$ that were used to re-estimate $\mathbf{H}$, similar to a single iteration of BCOMP.

The outcome of these three steps is shown in red circle in the supplementary Fig. 1.

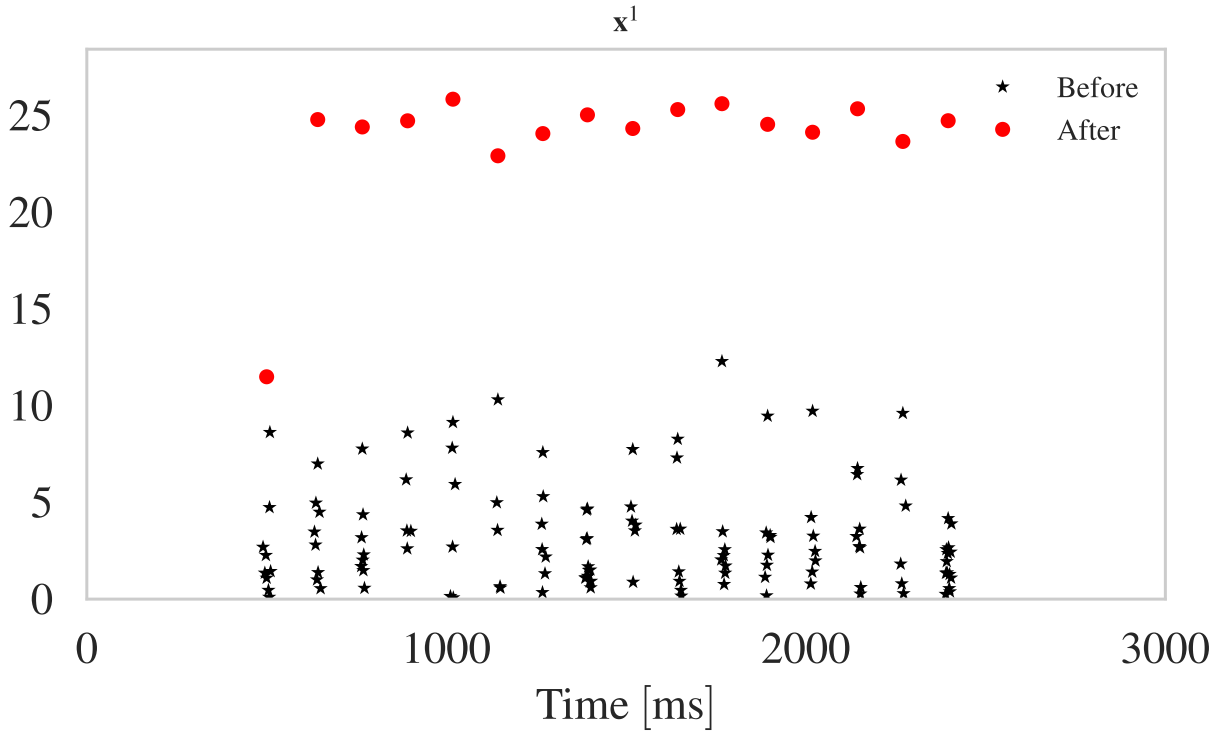

Figure 1: Output of the the DCEA encoder before and after post-processing.

4 Generalized linear model (GLM) for whisker experiment

In this section, for ease of notation, we consider the simple case of $M_j = 1$ (Bernoulli). However, the detail can be generalized to the binomial generative model.

We describe the GLM [Truccolo et al., 2005] used for analyzing the neural spiking data from the whisker experiment [Ba et al., 2014], and which we compared to BCOMP and DCEA in Fig. 4. Fig. 4(b) depicts a segment of the periodic stimulus used in the experiment to deflect the whisker. The units are in $\frac{\text{mm}}{10}$. The full stimulus lasts 3000 ms and is equal to zero (whisker at rest) during the two baseline periods from 0 to 500 ms and 2500 to 3000 ms. In the GLM analysis, we used whisker velocity as a stimulus covariate, which corresponds to the first difference of the position stimulus $\mathbf{s} \in \mathbb{R}^{3000}$. The blue curve in Fig. 4(c) represents one period of the whisker-velocity covariate. We associated a single stimulus coefficient $\boldsymbol{\beta}_{\text{stim}} \in \mathbb{R}$ to this covariate. In addition to the stimulus covariate, we used history covariates in the GLM. We denote by $\boldsymbol{\beta}_H^j \in \mathbb{R}^{L_j}$ the coefficients associated with these covariates, where $j = 1, \cdots, J$ is the neuron index. We also define a^j to be the base firing rate for neuron j. The GLM is given by

$$\begin{aligned} &\mathbf{y}^j[n] \sim \text{Bernoulli}(\mathbf{p}^j[n]) \\ &\text{s.t. } \mathbf{p}^j[n] = \Big(1 + \exp\big(-a^j - \boldsymbol{\beta}_{\text{stim}} \cdot \underbrace{(\mathbf{s}[n] - \mathbf{s}[n-1])}_{\text{whisker velocity}} - \sum_{l=1}^{L_j} \boldsymbol{\beta}_H^j[l] \cdot \mathbf{y}^j[n-l]\big)\Big)^{-1} \end{aligned} \tag{101}$$

The parameters $\{a^j\}_{j=1}^J$, $\boldsymbol{\beta}_{\text{stim}}$, and $\{\boldsymbol{\beta}_H^j\}_{j=1}^J$ are estimated by minimizing the negative likelihood of the neural spiking data $\{\mathbf{y}^j\}_{j=1}^{10}$ with $M_j = 30$ *from all neurons* using IRLS. We picked the order L_j (in ms) of the history effect for neuron j by fitting the GLM to each of the 10 neurons *separately* and finding the value of $\approx 5 \leq L_j \leq 100$ that minimizes the Akaike Information Criterion [Truccolo et al., 2005].

Interpretation of the GLM as a convolutional model Because whisker position is periodic with period 125 ms, so is whisker velocity. Letting $\mathbf{h}_1$ denote whisker velocity in the interval of length 125 ms starting at 500 ms (blue curve in Fig. 4(c)), we can interpret the GLM in terms of the convolutional model of Eq. 8. In this interpretation, $\mathbf{H}$ is the convolution matrix associated with the *fixed* filter $\mathbf{h}_1$ (blue curve in Fig. 4(c)), and $\mathbf{x}^j$ is a sparse vector with 16 equally spaced nonzero entries all equal to $\boldsymbol{\beta}_{\text{stim}}$. The first nonzero entry of $\mathbf{x}^j$ occurs at index 500. The number of indices between nonzero entries is 125. The blue dots in Fig. 4(d) reflect this interpretation.

Incorporating history dependence in the generative model GLMs of neural spiking data [Truccolo et al., 2005] include a constant term that models the baseline probability of spiking a^j, as well as a term that models the effect of spiking history. This motivates us to use the model

$$\log \frac{p(\mathbf{y}^{j,m} \mid \{\mathbf{h}_c\}_{c=1}^{C}, \mathbf{x}^j, \mathbf{x}_H^j)}{1 - p(\mathbf{y}^{j,m} \mid \{\mathbf{h}_c\}_{c=1}^{C}, \mathbf{x}^j, \mathbf{x}_H^j)} = a^j + \mathbf{H}\mathbf{x}^j + \mathbf{Y}_j \mathbf{x}_H^j, \tag{102}$$

where $\mathbf{y}^{j,m} \in \{0,1\}^N$ refers to m^{th} trial of the binomial data $\mathbf{y}^j$. The n^{th} row of $\mathbf{Y}_j \in \mathbb{R}^{N \times L_j}$ contains the spiking history of neuron j at trial m from $n - L_j$ to n, and $\mathbf{x}_H^j \in \mathbb{R}^{L_j}$ are coefficients that capture the effect of spiking history on the propensity of neuron j to spike. We use the same L_j estimated from GLM. We estimate a^j from the average firing probability during the baseline period. The addition of the history term simply results in an additional set of variables to alternate over in the alternating-minimization interpretation of ECDL. We estimate it by adding a loop around BCOMP or backpropagation through DCEA. Every iteration of this loop first assumes $\mathbf{x}_H^j$ are fixed. Then, it updates the filters and $\mathbf{x}^j$. Finally, it solves a convex optimization problem to update $\mathbf{x}_H^j$ given the filters and $\mathbf{x}^j$. In the interest of space, we do not describe this algorithm formally.

5 Kolmogorov-smirnov plots and the time-rescaling theorem

Loosely, the time-rescaling theorem states that rescaling the inter-spike intervals (ISIs) of the neuron using the (unknown) underlying conditional intensity function (CIF) will transform them into i.i.d. samples from an exponential random variable with rate 1. This implies that, if we apply the CDF of an exponential random variable with rate 1 to the rescaled ISIs, these should look like i.i.d. draws from a uniform random variable in the interval $[0, 1]$. KS plots are a visual depiction of this result. They are obtained by computing the rescaled ISIs using an estimate of the underlying CIF and applying the CDF of an exponential random variable with rate 1 to them. These are then sorted and plotted against ideal uniformly-spaced empirical quantiles from a uniform random variable in the interval $[0, 1]$. The CIF that fits the data the best is the one that yields a curve that is the closest to the 45-degree diagonal. Fig. 4(e) depicts the KS plots obtained using the CIFs estimated using DCEA, BCOMP and the GLM.

6 Image denoising

This section visualizes several test images for Poisson image denoising.

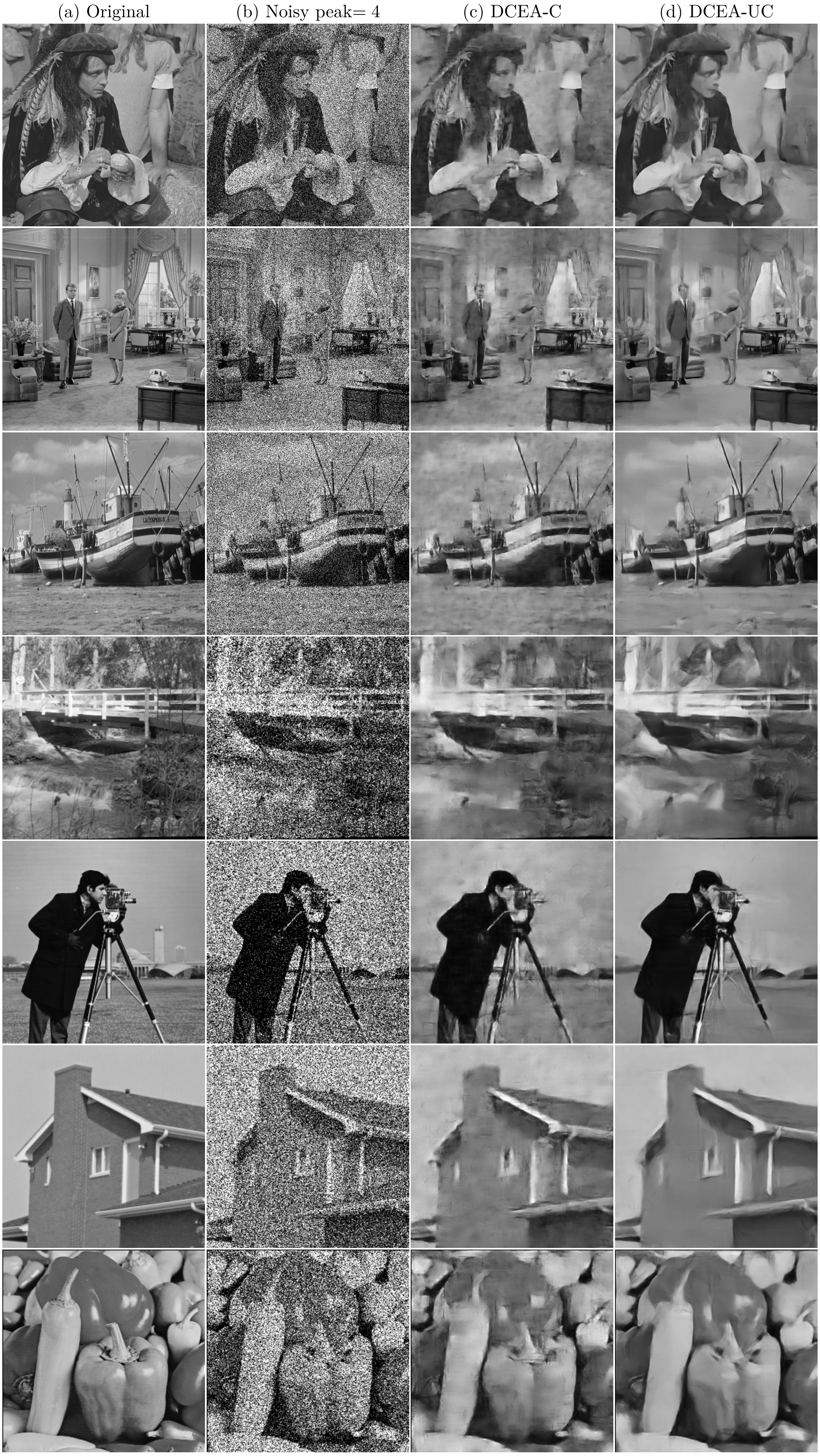

Figure 2: Denoising performance on test images with peak= 4. (a) Original, (b) noisy, (c) DCEA-C, and (d) DCEA-UC.

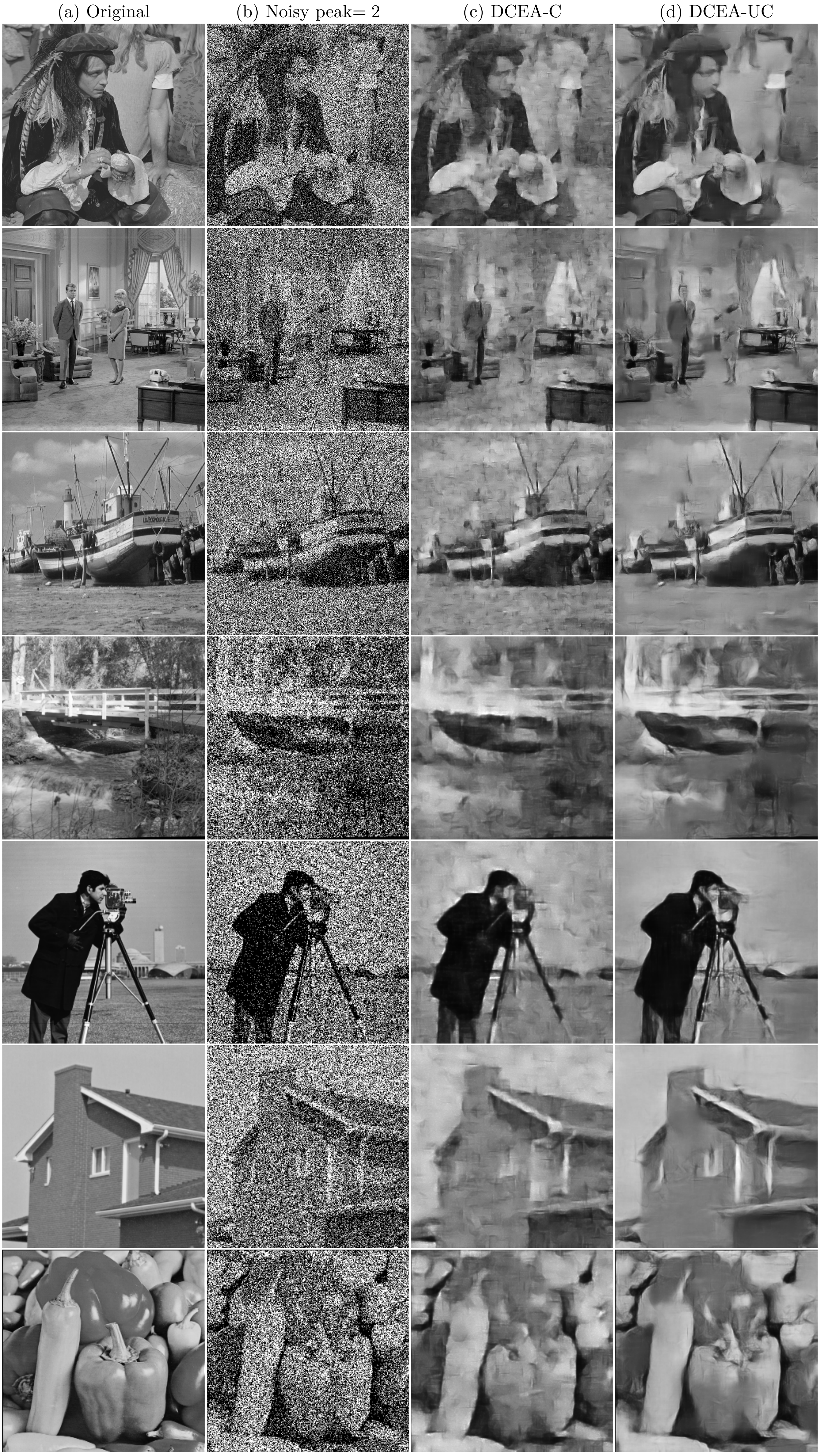

Figure 3: Denoising performance on test images with peak= 2. (a) Original, (b) noisy, (c) DCEA-C, and (d) DCEA-UC.

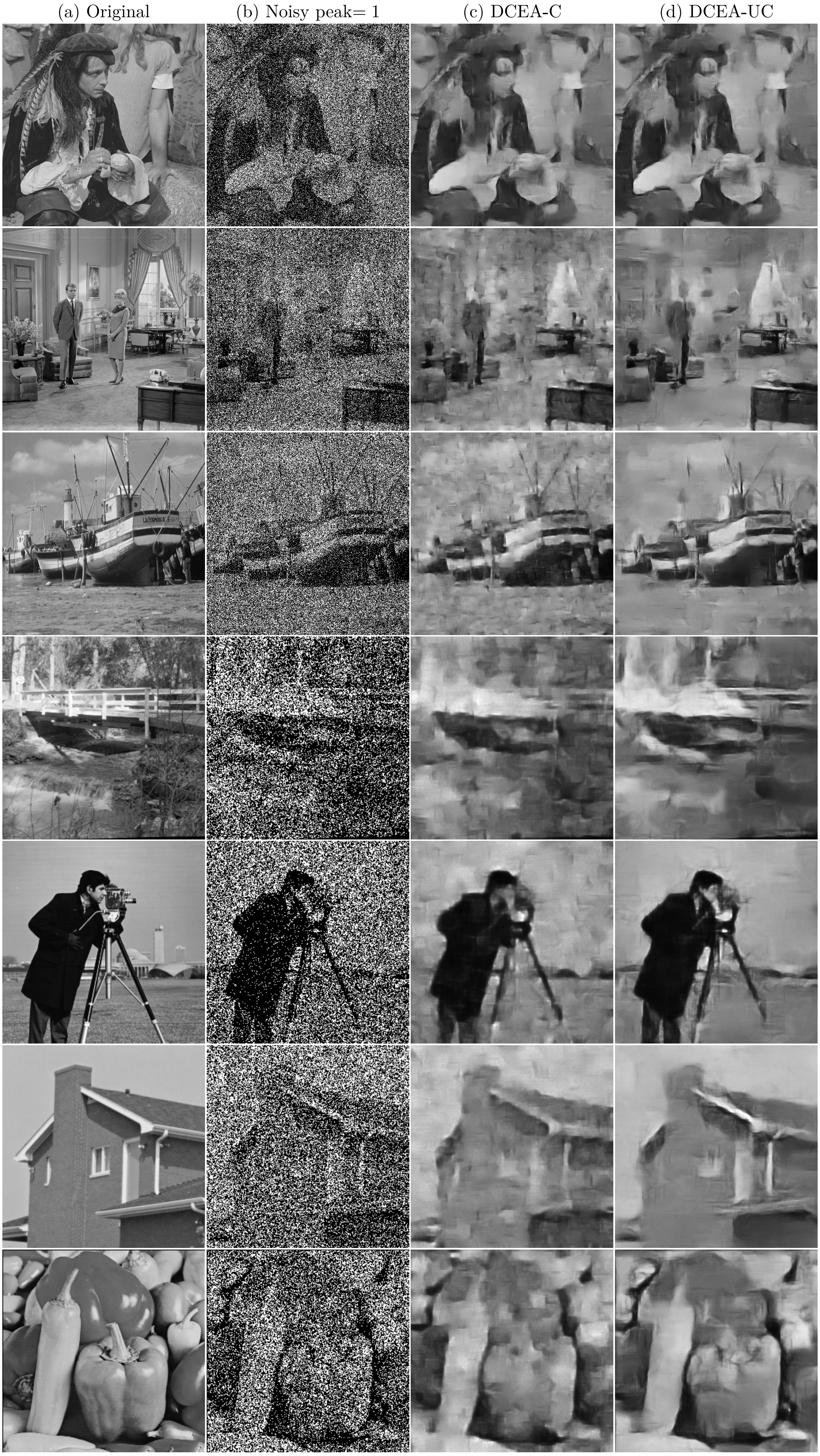

Figure 4: Denoising performance on test images with peak= 1. (a) Original, (b) noisy, (c) DCEA-C, and (d) DCEA-UC.

References

[Arora et al., 2015] Arora, S., Ge, R., Ma, T., and Moitra, A. (2015). Simple, efficient, and neural algorithms for sparse coding. In *Proc. the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 113–149, Paris, France. PMLR.

[Ba et al., 2014] Ba, D., Temereanca, S., and Brown, E. (2014). Algorithms for the analysis of ensemble neural spiking activity using simultaneous-event multivariate point-process models. *Frontiers in Computational Neuroscience*, 8:6.

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

[Bhaskar et al., 2013] Bhaskar, B. N., Tang, G., and Recht, B. (2013). Atomic norm denoising with applications to line spectral estimation. *IEEE Transactions on Signal Processing*, 61(23):5987–5999.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. In *Proc. the 3rd International Conference on Learning Representations (ICLR)*, pages 1–15.

[Lozano et al., 2011] Lozano, A., Swirszcz, G., and Abe, N. (2011). Group orthogonal matching pursuit for logistic regression. *Journal of Machine Learning Research*, 15:452–460.

[Mailhé et al., 2011] Mailhé, B., Gribonval, R., Vandergheynst, P., and Bimbot, F. (2011). Fast orthogonal sparse approximation algorithms over local dictionaries. *Signal Processing*, 91:2822–2835.

[Mardani et al., 2018] Mardani, M., Sun, Q., Vasawanala, S., Papyan, V., Monajemi, H., Pauly, J., and Donoho, D. (2018). Neural proximal gradient descent for compressive imaging. In *Proc. Advances in Neural Information Processing Systems 31*, pages 9573–9683.

[Nguyen et al., 2019] Nguyen, T. V., Wong, R. K. W., and Hegde, C. (2019). On the dynamics of gradient descent for autoencoders. In *Proc. Machine Learning Research*, volume 89, pages 2858–2867. PMLR.

[Tang et al., 2013] Tang, G., Bhaskar, B. N., and Recht, B. (2013). Sparse recovery over continuous dictionaries-just discretize. In *2013 Asilomar Conference on Signals, Systems and Computers*, pages 1043–1047.

[Tolooshams et al., 2020] Tolooshams, B., Dey, S., and Ba, D. (2020). Deep residual autoencoders for expectation maximization-inspired dictionary learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15.

[Tropp and Gilbert, 2007] Tropp, J. A. and Gilbert, A. C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666.

[Truccolo et al., 2005] Truccolo, W., Eden, U. T., Fellows, M., Donoghue, J., and Brown, E. N. (2005). A Point Process Framework for Relating Neural Spiking Activity to Spiking History, Neural Ensemble, and Extrinsic Covariate Effects. *Journal of Neurophysiology*, 93(2):1074–1089.

[Vincent and Bengio, 2002] Vincent, P. and Bengio, Y. (2002). Kernel matching pursuit. *Machine Learning*, 48(1):165–187.