A Proof of Theorem 3

We start our proof with the following auxiliary claims.

Claim 1. *Suppose that* $\max_i \|W_i - A_i\| \leq \delta$ *and* $\|W_i\| = 1$. *We have:*

1. $\langle W_i, A_i \rangle \geq 1 - \delta^2/2$ *for any* $i \in [m]$;
2. $|\langle W_i, A_j \rangle| \leq \mu/\sqrt{n} + \delta$, *for any* $j \neq i \in [m]$;
3. $\sum_{j \in S \backslash \{i\}} \langle W_i, A_j \rangle^2 \leq O(\mu^2 k/n + \delta^2)$ *for any* $S \subset [m]$ *of size at most* k.

Proof. The claims (i) and (ii) clearly follow from the δ-closeness and μ-incoherence properties as shown below.

$$\langle W_i, A_i \rangle = 1 - (1/2)\|W_i - A_i\|^2 \geq 1 - \delta^2/2,$$

and

$$|\langle W_i, A_j \rangle| = |\langle A_i, A_j \rangle + \langle W_i - A_i, A_j \rangle| \leq \mu/\sqrt{n} + \delta.$$

For (iii), we apply Cauchy-Schwarz to bound each term inside the summation. Precisely, for any $j \neq i$,

$$\langle W_i, A_j \rangle^2 \leq 2\big(\langle A_i, A_j \rangle^2 + \langle W_i - A_i, A_j \rangle^2\big) \leq 2\mu^2/n + 2\langle W_i - A_i, A_j \rangle^2.$$

Together with $\|A\| = O(\sqrt{m/n}) = O(1)$, we finish proving (iii) by noting that

$$\sum_{j \in S \backslash \{i\}} \langle W_i, A_j \rangle^2 \leq 2\mu^2 k/n + 2\|A_S^T(W_i - A_i)\|_F^2 \leq 2\mu^2 k/n + 2\|A_S\|^2\|W_i - A_i\|^2 \leq O(\mu^2 k/n + \delta^2).$$

■

Claim 2. *Suppose* $\|W_i\| = 1$, *then* $\max_i |\langle W_i, \eta \rangle| \leq \sigma_\eta \log n$ *holds with high probability.*

Proof. Since η is a spherical Gaussian random vector and $\|W_i\| = 1$, $\langle W_i, \eta \rangle$ is Gaussian with mean 0 and variance σ_η^2. Using the Gaussian tail bound for $\langle W_i, \eta \rangle$ and taking the union bound over $i = 1, 2, \dots, m$, we have that $\max_i |\langle W_i, \eta \rangle| \leq \sigma_\eta \log n$ holds with high probability. ■

Proof of Theorem 3. Denote $z = W^T y + b$ and let $i \in [m]$ be fixed for a moment. (Later we use a union bound argument for account for all i). Denote $S = \mathrm{supp}(x^*)$ and $R = S \backslash \{i\}$. Notice that $x_i = 0$ if $i \notin S$ by definition. One can write the i^{th} entry z_i of the weighted sum z as

$$\begin{aligned} z_i &= W_i^T(A_S x_S^* + \eta) + b_i \\ &= \langle W_i, A_i \rangle x_i^* + \sum_{j \in R} \langle W_i, A_j \rangle x_j^* + \langle W_i, \eta \rangle + b_i \\ &= \langle W_i, A_i \rangle x_i^* + Z_i + \langle W_i, \eta \rangle + b_i, \end{aligned}$$

where we write $Z_i = \sum_{j \in R} \langle W_i, A_j \rangle x_j^*$. Roughly speaking, since $\langle W_i, A_i \rangle$ is close to 1, z_i approximately equals x_i^* if we can control the remaining terms. This will be made precise below separately for different generative models.

A.1 Case (i): Sparse coding model

For this setting, the hidden code x^* is k-sparse and is not restricted to non-negative values. The nonzero entries are mutually independent sub-Gaussian with mean $\kappa_1 = 0$ and variance $\kappa_2 = 1$. Note further that $a_1 \in (0, 1]$ and $a_2 = \infty$ and the dictionary is incoherent and over-complete.

Since the true code takes both positive and negative values as well as sparse, it is natural to consider the hard thresholding activation. The consistency is studied in [16] for the case of sparse coding (see Appendix C and also work [28], Lemma 8 for a treatment of the noise.)

A.2 Case (ii) and (iii): Non-negative k-sparse model

Recall that $S = \mathrm{supp}(x^*)$ and that $x^*_j \in [a_1, a_2]$ for $j \in S$. Cauchy-Schwarz inequality implies

$$|Z_i| = \Big|\sum_{j \in R} \langle W_i, A_j\rangle x^*_j\Big| \le \sqrt{\sum_{j \in R} \langle W_i, A_j\rangle^2} \|x^*\| \le a_2 \sqrt{\frac{\mu^2 k^2}{n} + k\delta^2},$$

where we use bound (ii) in Claim 1 and $\|x^*\| \le a_2\sqrt{k}$.

If $i \in S$, then w.h.p.

$$\begin{aligned} z_i &= \langle W_i, A_i\rangle x^*_i + Z_i + \langle W_i, \eta\rangle \\ &\ge (1 - \delta^2/2)a_1 - a_2\sqrt{\frac{\mu^2 k^2}{n} + k\delta^2} - \sigma_\eta \log n + b_i > 0 \end{aligned}$$

for $b_i \ge -(1-\delta)a_1 + a_2\delta\sqrt{k}$ and $a_2\delta\sqrt{k} \ll (1-\delta)a_1$, $k = O(1/\delta^2) = O(\log^2 n)$, $\mu \le \delta\sqrt{n}/k$, and $\sigma_\eta = O(1/\sqrt{n})$. On the other hand, when $i \notin S$ then w.h.p.

$$\begin{aligned} z_i &= Z_i + \langle W_i, \eta\rangle + b_i \\ &\le a_2\sqrt{\frac{\mu^2 k^2}{n} + k\delta^2} + \sigma_\eta \log n + b_i \\ &\le 0 \end{aligned}$$

for $b_i \le -a_2\sqrt{\frac{\mu^2 k^2}{n} + k\delta^2} - \sigma_\eta \log n \approx -a_2\delta\sqrt{k}$.

Due to the use of Claim 2, these results hold w.h.p. uniformly for all i and so $\mathrm{supp}(x) = S$ for $x = \mathrm{ReLU}(W^T y + b)$ w.h.p. by We re-use the tail bound $\mathbb{P}[Z_i \ge \epsilon]$ given in [11], Theorem 3.1.

Moreover, one can also see that with high probability $z_i > a_1/2$ if $i \in S$ and $z_i < a_2\delta\sqrt{k} < a_1/4$ otherwise. This results hold w.h.p. uniformly for all i and so $x = \mathrm{threshold}_{1/2}(z)$ has the same support as x^* w.h.p. ■

B Proof of Theorem 4

B.1 Case (i): Mixture of Gaussians

We start with simplifying the form of g_i using the generative model 3 and Theorem 3. First, from the model we can have $p_i = \mathbb{P}[x^*_i \ne 0] = \Theta(1/m)$ and $\mathbb{E}[\eta] = 0$ and $\mathbb{E}[\eta\eta^T] = \sigma_\eta^2 I$. Second, by Theorem 3 in (i), $\mathbf{1}_{x_i \ne 0} = x^*_i = 1$ with high probability. As such, under the event we have $x_i = \sigma(W_i^T y + b_i) = (W_i^T y + b_i)\mathbf{1}_{x^*_i \ne 0}$ for both choices of σ (Theorem 3).

To analyze g_i, we observe that

$$\gamma = \mathbb{E}[(W_i^T y I + b_i I + y W_i^T)(y - Wx)(\mathbf{1}_{x^*_i \ne 0} - \mathbf{1}_{x_i \ne 0})]$$

has norm of order $O(n^{-w(1)})$ since the failure probability of the support consistency event is sufficiently small for large n, and the remaining term has bounded moments. One can write:

$$\begin{aligned} g_i &= -\mathbb{E}[\mathbf{1}_{x^*_i \ne 0}(W_i^T y I + b_i I + y W_i^T)(y - Wx)] + \gamma \\ &= -\mathbb{E}[\mathbf{1}_{x^*_i \ne 0}(W_i^T y I + y W_i^T + b_i I)(y - W_i W_i^T y - b_i W_i)] + \gamma \\ &= -\mathbb{E}[\mathbf{1}_{x^*_i \ne 0}(W_i^T y I + y W_i^T)(I - W_i W_i^T)y] + b_i \mathbb{E}[\mathbf{1}_{x^*_i \ne 0}(W_i^T y I + y W_i^T)]W_i \\ &\qquad - b_i \mathbb{E}[\mathbf{1}_{x^*_i \ne 0}(I - W_i W_i^T)y] + b_i^2 W_i \mathbb{E}[\mathbf{1}_{x^*_i \ne 0}] + \gamma \\ &= g_i^{(1)} + g_i^{(2)} + g_i^{(3)} + p_i b_i^2 W_i + \gamma, \end{aligned}$$

Next, we study each of $g_i^{(t)}$, $t = 1, 2, 3$, by using the fact that $y = A_i + \eta$ as $x_i^* = 1$. To simplify the notation, denote $\lambda_i = \langle W_i, A_i \rangle$. Then

$$\begin{aligned} g_i^{(1)} &= -\mathbb{E}[(W_i^T(A_i+\eta)I + (A_i+\eta)W_i^T)(I - W_iW_i^T)(A_i+\eta)\mathbf{1}_{x_i^* \neq 0}] \\ &= -\mathbb{E}[(\lambda_i I + A_iW_i^T + \langle W_i, \eta\rangle I + \eta W_i^T)(I - W_iW_i^T)(A_i+\eta)\mathbf{1}_{x_i^* \neq 0}] \\ &= -(\lambda_i I + A_iW_i^T)(A_i - \lambda_i W_i)\mathbb{P}[x_i^* \neq 0] - \mathbb{E}[(\langle W_i, \eta\rangle I + \eta W_i^T)(I - W_iW_i^T)\eta\mathbf{1}_{x_i^* \neq 0}] \\ &= -p_i\lambda_i A_i + p_i\lambda_i^2 W_i - \mathbb{E}[(\langle W_i, \eta\rangle I + \eta W_i^T)(I - W_iW_i^T)\eta\mathbf{1}_{x_i^* \neq 0}], \end{aligned}$$

where we use $p_i = \mathbb{P}[x_i^* \neq 0]$ and denote $\|W_i\| = 1$. Also, since η is spherical Gaussian-distributed, we have:

$$\begin{aligned} \mathbb{E}[(\langle W_i, \eta\rangle I + \eta W_i^T)(I - W_iW_i^T)\eta\mathbf{1}_{x_i^* \neq 0}] &= p_i\mathbb{E}[\langle W_i, \eta\rangle\eta - \langle W_i, \eta\rangle^2 W_i] \\ &= p_i\sigma_\eta^2(1 - \|W_i\|^2)W_i = 0, \end{aligned}$$

To sum up, we have

$$g_i^{(1)} = -p_i\lambda_i A_i + p_i\lambda_i^2 W_i \tag{6}$$

For the second term,

$$\begin{aligned} g_i^{(2)} = b_i\mathbb{E}[\mathbf{1}_{x_i^* \neq 0}(W_i^T y I + yW_i^T)]W_i &= b_i\mathbb{E}[\mathbf{1}_{x_i^* \neq 0}(W_i^T(A_i+\eta)I + (A_i+\eta)W_i^T)]W_i \\ &= b_i\mathbb{E}[(\lambda_i W_i + \|W_i\|^2 A_i)\mathbf{1}_{x_i^* \neq 0}] \\ &= p_ib_i\lambda_i W_i + p_ib_iA_i. \end{aligned} \tag{7}$$

In the second step, we use the independence of spherical η and x. Similarly, we can compute the third term:

$$\begin{aligned} g_i^{(3)} = -b_i(I - W_iW_i^T)\mathbb{E}[y\mathbf{1}_{x_i^* \neq 0}] &= -b_i(I - W_iW_i^T)\mathbb{E}[(A_i+\eta)\mathbf{1}_{x_i^* \neq 0}] \\ &= -p_ib_i(I - W_iW_i^T)A_i \\ &= -p_ib_iA_i + p_ib_i\lambda_i W_i \end{aligned} \tag{8}$$

Putting (6), (7) and (8) together, we have

$$g_i = -p_i\lambda_i A_i + p_i(\lambda_i^2 + 2b_i\lambda_i + b_i^2)W_i + \gamma$$

Having established the closed-form for g_i, one can observe that when b_i such that $\lambda_i^2 + 2b_i\lambda_i + b_i^2 \approx \lambda_i$, g_i roughly points in the same desired direction to A^* and suggests the correlation of g_i with $W_i - A_i$. Now, we prove this result.

Proof of Lemma 1. Denote $v = p_i(\lambda_i^2 + 2b_i\lambda_i + b_i^2 - \lambda_i)W_i + \gamma$. Then

$$\begin{aligned} g_i &= -p_i\lambda_i A_i + p_i(\lambda_i^2 + 2b_i\lambda_i + b_i^2)W_i + \gamma \\ &= p_i\lambda_i(W_i - A_i) + v, \end{aligned} \tag{9}$$

By expanding (9), we have

$$2\langle v, W_i - A_i\rangle = \frac{1}{p_i\lambda_i}\|g_i\|^2 - p_i\lambda_i\|W_i - A_i\|^2 - \frac{1}{p_i\lambda_i}\|v\|^2.$$

Using this equality and taking inner product with $W_i - A_i$ to both sides of (9), we get

$$2\langle g_i, W_i - A_i\rangle = p_i\lambda_i\|W_i - A_i\|^2 + \frac{1}{p_i\lambda_i}\|g_i\|^2 - \frac{1}{p_i\lambda_i}\|v\|^2.$$

We need an upper bound for $\|v\|^2$. Since

$$|(b_i + \lambda_i)^2 - \lambda_i| \leq 2(1 - \lambda_i)$$

and

$$2(1-\lambda_i) = \|W_i - A_i\|^2,$$

we have:

$$|(b_i+\lambda_i)^2 - \lambda_i| \leq \|W_i - A_i\|^2 \leq \delta\|W_i - A_i\|$$

Notice that

$$\begin{aligned}\|v\|^2 &= \|p_i(\lambda_i^2 + 2b_i\lambda_i + b_i^2 - \lambda_i)W_i + \gamma\|^2 \\ &\leq 2p_i^2\delta^2\|W_i - A_i\|^2 + 2\|\gamma\|^2.\end{aligned}$$

Now one can easily show that

$$2\langle g_i, W_i - A_i\rangle \geq p_i(\lambda_i - 2\delta^2)\|W_i - A_i\|^2 + \frac{1}{p_i\lambda_i}\|g_i\|^2 - \frac{2}{p_i\lambda_i}\|\gamma\|^2.$$

■

B.2 Case (ii): General k-Sparse Coding

For this case, we adopt the same analysis as used in Case 1. The difference lies in the distributional assumption of x^*, where nonzero entries are independent sub-Gaussian. Specifically, given the support S of size at most k with $p_i = \mathbb{P}[i \in S] = \Theta(k/m)$ and $p_{ij} = \mathbb{P}[i,j \in S] = \Theta(k^2/m^2)$, we suppose $\mathbb{E}[x_i^*|S] = 0$ and $\mathbb{E}[x_S^* x_S^{*T}|S] = I$. For simplicity, we choose to skip the noise, i.e., $y = Ax^*$ for this case. Our analysis is robust to iid additive Gaussian noise in the data; see [28] for a similar treatment. Also, according to Theorem 3, we set $b_i = 0$ to obtain support consistency. With *zero* bias, the expected update rule g_i becomes

$$g_i = -\mathbb{E}[(W_i^T yI + yW_i^T)(y - Wx)\mathbf{1}_{x_i \neq 0}].$$

For $S = \text{supp}(x^*)$, then $y = A_S x_S^*$. Theorem 3 in (ii) shows that $\text{supp}(x) = S$ w.h.p., so under that event we can write $Wx = W_S x_S = W_S(W_S^T y)$. Similar to the previous cases, γ denotes a general quantity whose norm is of order $n^{-w(1)}$ due to the converging probability of the support consistency. Now, we substitute the forms of y and x into g_i:

$$\begin{aligned}g_i &= -\mathbb{E}[(W_i^T yI + yW_i^T)(y - Wx)\mathbf{1}_{x_i\neq 0}] \\ &= -\mathbb{E}[(W_i^T yI + yW_i^T)(y - W_S W_S^T y)\mathbf{1}_{x_i^*\neq 0}] + \gamma \\ &= -\mathbb{E}[(I - W_S W_S^T)(W_i^T A_S x_S^*)A_S x_S^* \mathbf{1}_{x_i^*\neq 0}] - \mathbb{E}[(A_S x_S^*)W_i^T(I - W_S W_S^T)A_S x_S^* \mathbf{1}_{x_i^*\neq 0}] + \gamma \\ &= g_i^{(1)} + g_i^{(2)} + \gamma.\end{aligned}$$

Write

$$g_{i,S}^{(1)} = -\mathbb{E}[(I - W_S W_S^T)(W_i^T A_S x_S^*)A_S x_S^* \mathbf{1}_{x_i^*\neq 0}|S],$$

and

$$g_{i,S}^{(2)} = -\mathbb{E}[(A_S x_S^*)W_i^T(I - W_S W_S^T)A_S x_S^* \mathbf{1}_{x_i^*\neq 0}|S],$$

so that $g_i^{(1)} = \mathbb{E}(g_{i,S}^{(1)})$ and $g_i^{(2)} = \mathbb{E}(g_{i,S}^{(2)})$. It is easy to see that $\mathbb{E}[x_j^* x_l^* \mathbf{1}_{x_i^*\neq 0}|S] = 1$ if $i = j = l \in S$ and $\mathbb{E}[x_i^* x_l^* \mathbf{1}_{x_i^*\neq 0}|S] = 0$ otherwise. Therefore, $g_{i,S}^{(1)}$ becomes

$$\begin{aligned}g_{i,S}^{(1)} &= -\mathbb{E}[(I - W_S W_S^T)(W_i^T A_S x_S^*)A_S x_S^* \mathbf{1}_{x_i^*\neq 0}|S] \qquad (10) \\ &= -\sum_{j,l\in S}\mathbb{E}[(I - W_S^T W_S)(W_i^T A_j)A_l x_j^* x_l^* \mathbf{1}_{x_i^*\neq 0}|S] \\ &= -\lambda_i(I - W_S W_S^T)A_i, \qquad (11)\end{aligned}$$

where we use the earlier notation $\lambda_i = W_i^T A_i$. Similar calculation of the second term results in

$$\begin{aligned}
g_{i,S}^{(2)} &= -\mathbb{E}[(A_S x_S^*) W_i^T (I - W_S W_S^T) A_S x_S^* \mathbf{1}_{x_i^* \neq 0} | S] && (12)\\
&= -\mathbb{E}[\sum_{j \in S} x_j^* A_j W_i^T (I - W_S W_S^T) \sum_{l \in S} x_l^* A_l \mathbf{1}_{x_i^* \neq 0} | S]\\
&= -\sum_{j,l \in S} \mathbb{E}[A_j W_i^T (I - W_S W_S^T) A_l x_j^* x_l^* \mathrm{sgn}(x_i^*) | S]\\
&= -A_i W_i^T (I - W_S W_S^T) A_i && (13)
\end{aligned}$$

Now we combine the results in (10) and (12) to compute the expectation over S.

$$\begin{aligned}
g_i &= \mathbb{E}[g_{i,S}^{(1)} + g_{i,S}^{(2)}] + \gamma && (14)\\
&= -\mathbb{E}[\lambda_i (I - W_S W_S^T) A_i + A_i W_i^T (I - W_S W_S^T) A_i] + \gamma\\
&= -\mathbb{E}[2\lambda_i A_i - \lambda_i \sum_{j \in S} W_j W_j^T A_i - A_i W_i^T \sum_{j \in S} W_j W_j^T A_i] + \gamma\\
&= -2p_i \lambda_i A_i + \mathbb{E}[\lambda_i \sum_{j \in S} W_j W_j^T A_i + \sum_{j \in S} \langle W_i, W_j \rangle \langle A_i, W_j \rangle A_i] + \gamma\\
&= -2p_i \lambda_i A_i + \mathbb{E}[\lambda_i^2 W_i + \sum_{j \in R} \langle A_i, W_j \rangle W_j + \lambda_i \|W_i\|^2 A_i + \sum_{j \in R} \langle W_i, W_j \rangle \langle A_i, W_j \rangle A_i] + \gamma,
\end{aligned}$$

where $p_i = \mathbb{P}[i \in S]$ and $R = S \backslash \{i\}$. Moreover, $\|W_i\| = 1$, hence

$$\begin{aligned}
g_i &= -p_i \lambda_i A_i + p_i \lambda_i^2 W_i + \sum_{j \in [m] \backslash \{i\}} p_{ij} \lambda_i \langle A_i, W_j \rangle W_j + p_{ij} \langle W_i, W_j \rangle \langle A_i, W_j \rangle A_i) + \gamma\\
&= -p_i \lambda_i A_i + p_i \lambda_i^2 W_i + \lambda_i W_{-i} \mathrm{diag}(p_{ij}) W_{-i}^T A_i + (W_i^T W_{-i} \mathrm{diag}(p_{ij}) W_{-i}^T A_i) A_i + \gamma, && (15)
\end{aligned}$$

for $W_{-i} = (W_1, \dots, W_{i-1}, W_{i+1}, \dots, W_m)$ with the i^{th} column being removed, and $\mathrm{diag}(p_{ij})$ denotes the diagonal matrix formed by p_{ij} with $j \in [m] \backslash \{i\}$.

Observe that ignoring lower order terms, g_i can be written as $p_i \lambda_i (W_i - A_i) + p_i \lambda_i (\lambda_i - 1) W_i$, which roughly points in the same desired direction to A. Rigorously, we argue the following:

Lemma 2. *Suppose W is $(\delta, 2)$-near to A. Then*

$$2\langle g_i, W_i - A_i \rangle \geq p_i \lambda_i \|W_i - A_i\|^2 + \frac{1}{p_i \lambda_i} \|g_i\|^2 - O(p_i k^2 / n^2 \lambda_i)$$

Proof. We proceed with similar steps as in the proof of Lemma 1. By nearness,

$$\|W\| \leq \|W - A\| + \|A\| \leq 3\|A\| \leq O(\sqrt{m/n}).$$

Also, $p_i = \Theta(k/m)$ and $p_{ij} = \Theta(k^2/m^2)$. Then

$$\begin{aligned}
\|W_{-i} \mathrm{diag}(p_{ij}) W_{-i}^T A_i\| &\leq p_i \|W_{-i} \mathrm{diag}(p_{ij}/p_i) W_{-i}^T\|\\
&\leq p_i \|W_{-i}\|^2 \max_{j \neq i} (p_{ij}/p_i) = O(p_i k/n).
\end{aligned}$$

Similarly,

$$\|W_i^T W_{-i} \mathrm{diag}(p_{ij}) W_{-i}^T A_i) A_i\| \leq O(p_i k/n).$$

Now we denote

$$v = p_i \lambda_i (\lambda_i - 1) W_i + \lambda_i W_{-i} \mathrm{diag}(p_{ij}) W_{-i}^T A_i + (W_i^T W_{-i} \mathrm{diag}(p_{ij}) W_{-i}^T A_i) A_i + \gamma.$$

Then

$$g_i = p_i \lambda_i (W_i - A_i) + v$$

where $\|v\| \leq p_i\lambda_i(\delta/2)\|W_i - A_i\| + O(p_ik/n) + \|\gamma\|$. Therefore, we obtain

$$2\langle g_i, W_i - A_i\rangle \geq p_i\lambda_i(1 - \frac{\delta^2}{2})\|W_i - A_i\|^2 + \frac{1}{p_i\lambda_i}\|g_i\|^2 - O(p_ik^2/n^2\lambda_i).$$

where we assume that $\|\gamma\|$ is negligible when compared with $O(p_ik/n)$. ■

Adopting the same arguments in the proof of Case (i), we are able to get the descent property column-wise for the normalized gradient update with the step size $\zeta = \max_i(1/p_i\lambda_i)$ such that there is some $\tau \in (0,1)$:

$$\|W_i^{s+1} - A_i\|^2 \leq (1-\tau)\|W_i^s - A_i\|^2 + O(p_ik^2/n^2\lambda_i).$$

Since $p_i = \Theta(k/m)$, Consequently, we will obtain the descent in Frobenius norm stated in Theorem 4, item (ii).

Lemma 3 (Maintaining the nearness)**.** $\|W - A\| \leq 2\|A\|$.

Proof. The proof follows from [16] (Lemma 24 and Lemma 32).

B.3 Case (iii): Non-negative k-Sparse Coding

We proceed with the proof similarly to the above case of general k-sparse code. Additional effort is required due to the positive mean of nonzero coefficients in x^*. For $x = \sigma(W^Ty + b)$, we have the support recovery for both choices of σ a shown in (ii) and (iii) of Theorem 3. Hence we re-use the expansion in [11] to compute the expected approximate gradient. Note that we standardize W_i such that $\|W_i\| = 1$ and ignore the noise η.

Let i be fixed and consider the approximate gradient for the i^{th} column of W. The expected approximate gradient has the following form:

$$g_i = -\mathbb{E}[\mathbf{1}_{x_i \neq 0}(W_i^T yI + b_iI + yW_i^T)(y - Wx)] = \alpha_iW_i - \beta_iA_i + e_i,$$

where

$$\begin{aligned}\alpha_i = \kappa_2p_i\lambda_i^2 &+ \kappa_2\sum_{j\neq i}p_{ij}\langle W_i, A_j\rangle^2 + 2\kappa_1^2\sum_{j\neq i}p_{ij}\lambda_i\langle W_i, A_j\rangle + \kappa_1^2\sum_{j\neq l\neq i}p_{ijl}\langle W_i, A_j\rangle\langle W_i, A_l\rangle \\ &+ 2\kappa_1p_ib_i\lambda_i + 2\kappa_1\sum_{j\neq i}p_{ij}b_i\langle W_i, A_j\rangle + p_ib_i^2;\end{aligned}$$

$$\begin{aligned}\beta_i = \kappa_2p_i\lambda_i &- \kappa_2\sum_{j\neq i}p_{ij}\langle W_i, W_j\rangle\langle A_i, W_j\rangle + \kappa_1^2\sum_{j\neq i}p_{ij}\langle W_i, A_j\rangle - \kappa_1^2\sum_{j\neq i}p_{ij}\langle W_i, W_j\rangle\langle W_j, A_j\rangle \\ &- \kappa_1^2\sum_{j\neq l\neq i}p_{ijl}\langle W_i, W_j\rangle\langle W_j, A_l\rangle - \kappa_1\sum_{j\neq i}p_{ij}b_i\langle W_i, W_j\rangle;\end{aligned}$$

and e_i is a term with norm $\|e_i\| \leq O(\max(\kappa_1^2, \kappa_2^2)p_ik/m)$ – a rough bound obtained in [11] (see the proof of Lemma 5.2 in pages 26 and 35 of [11].) As a sanity check, by plugging in the parameters of the mixture of Gaussians to α_i, β_i and e_i, we get the same expression for g_i in Case 1. We will show that only the first term in α_i is dominant except ones involving the bias b_i. The argument for β_i follows similarly.

Claim 3.

$$\begin{aligned}\alpha_i = \kappa_2p_i\lambda_i^2 &+ \kappa_2O(p_ik/m) + 2\kappa_1^2p_i\lambda_iO(k/\sqrt{m}) + \kappa_1^2O(p_ik^2/m) \\ &+ 2\kappa_1p_ib_i\lambda_i + 2\kappa_1p_ib_iO(k/\sqrt{m}) + p_ib_i^2.\end{aligned}$$

Proof. We bound the corresponding terms in α_i one by one. We start with the second term:

$$\begin{aligned}\sum_{j\neq i}^{m}p_{ij}\langle W_i, A_j\rangle^2 &\leq \max_{j\neq i}p_{ij}\sum_{j\neq i}^{m}\langle W_i, A_j\rangle^2 \\ &\leq \max_{j\neq i}p_{ij}\|A_{-i}^TW_i\|_F^2 \\ &\leq O(p_ik/m),\end{aligned}$$

since $p_{ij} = \Theta(k^2/m^2) = \Theta(p_i k/m)$. Similarly, we have

$$\begin{aligned} |\sum_{j\neq i}^{m} p_{ij}\langle W_i, A_j\rangle| &= |W_i^T \sum_{j\neq i}^{m} p_{ij} A_j| \\ &\leq \|W_i\|\|A\|\sqrt{\sum_{j\neq i} p_{ij}^2} \\ &\leq O(p_i k/\sqrt{m}), \end{aligned}$$

which leads to a bound on the third and the sixth terms. Note that this bound will be re-used to bound the corresponding term in β_i.

The next term is bounded as follows:

$$\begin{aligned} \sum_{\substack{j\neq l\\ j,l\neq i}} p_{ijl}\langle W_i, A_j\rangle\langle W_i, A_l\rangle &= W_i^T \sum_{\substack{j\neq l\\ j,l\neq i}} p_{ijl} A_j A_l^T W_i \\ &\leq \Big\| \sum_{\substack{j\neq l\\ j,l\neq i}} p_{ijl} A_j A_l^T \Big\| \|W_i\|^2 \\ &\leq O(p_i k^2/m), \end{aligned}$$

where $M = \sum_{\substack{j\neq l\\ j,l\neq i}} p_{ijl} A_j A_l^T = A_{-i} Q A_{-i}^T$ for $Q_{jl} = p_{ijl}$ for $j \neq l$ and $Q_{jl} = 0$ otherwise. Again, A_{-i} denotes the matrix W with its i^{th} column removed. We have $p_{ijl} = \Theta(k^3/m^3) \leq O(q_i k^2/m)$; therefore, $\|M\| \leq \|Q\|_F \|A\|^2 \leq O(q_i k^2/m)$. ■

Claim 4.

$$\begin{aligned} \beta_i = \kappa_2 p_i \lambda_i - \kappa_2 O(p_i k/m) + \kappa_1^2 O(p_i k/\sqrt{m}) - \kappa_1^2 O(p_i k/\sqrt{m}) \\ + \kappa_1^2 O(p_i k^2/m) - \kappa_1 b_i O(p_i k/\sqrt{m}). \end{aligned}$$

Proof. We proceed similarly to the proof of Claim 3. Due to nearness and the fact that $\|A^*\| = O(\sqrt{m/n}) = O(1)$, we can conclude that $\|W\| \leq O(1)$. For the second term, we have

$$\begin{aligned} \|\sum_{j\neq i} p_{ij}\langle W_i, W_j\rangle\langle A_i, W_j\rangle\| &= \|W_i^T \sum_{j\neq i} p_{ij} W_j W_j^T A_i\| \\ &\leq \max_{j\neq i} p_{ij} \|W_{-i} W_{-i}^T\|\|W_i\|\|A_i\| \\ &\leq O(p_i k/m), \end{aligned}$$

where $W_j W_j^T$ are p.s.d and so $0 \preceq \sum_{j\neq i} p_{ij} W_j W_j^T \preceq (\max_{j\neq i} p_{ij})(\sum_{j\neq i} W_j W_j^T) \preceq \max_{j\neq i} p_{ij} W_{-i} W_{-i}^T$. To bound the third one, we use the fact that $|\lambda_j| = |\langle W_j, A_j\rangle| \leq 1$. Hence from the proof of Claim 3,

$$\begin{aligned} \|\sum_{j\neq i} p_{ij}\langle W_i, W_j\rangle\langle W_j, A_j\rangle\| &= \|\sum_{j\neq i} p_{ij}\lambda_j\langle W_i, W_j\rangle\| \\ &\leq \|W_i\|\|W\|\sqrt{\sum_{j\neq i}(p_{ij}\lambda_j)^2} \\ &\leq O(p_i k/\sqrt{m}), \end{aligned}$$

which is also the bound for the last term. The remaining term can be bounded as follows:

$$\begin{aligned}\|\sum_{j\neq l\neq i} p_{ijl}\langle W_i, W_j\rangle\langle W_j, A_l\rangle\| &\leq \|\sum_{j\neq l\neq i} p_{ijl} W_j W_j^T A_l\| \\ &\leq \sum_{l\neq i}\|p_{ijl}W_{-i}W_{-i}^T\| \\ &\leq \sum_{l\neq i}\max_{j\neq l\neq i} p_{ijl}\|W_{-i}\|^2 \\ &\leq O(p_i k^2/m).\end{aligned}$$

■

When $b_i = 0$, from (3) and (4) and $b_i \in (-1, 0)$, we have:

$$\alpha_i = p_i(\kappa_2\lambda_i^2 + 2\kappa_1 p_i b_i\lambda_i + b_i^2) + O(\max(\kappa_1^2, \kappa_2)k/\sqrt{m})$$

and

$$\beta_i = \kappa_2 p_i\lambda_i + O(\max(\kappa_1^2, \kappa_2)k/\sqrt{m}),$$

where we implicitly require that $k \leq O(\sqrt{n})$, which is even weaker than the condition $k = O(1/\delta^2)$ stated in Theorem 3. Now we recall the form of g_i:

$$g_i = -\kappa_2 p_i\lambda_i A_i + p_i(\kappa_2\lambda_i^2 + 2\kappa_1 p_i b_i\lambda_i + b_i^2)W_i + v \tag{16}$$

where $v = O(\max(\kappa_1^2, \kappa_2)k/\sqrt{m})A_i + O(\max(\kappa_1^2, \kappa_2)k/\sqrt{m})W_i + e_i$. Therefore $\|v\| \leq O(\max(\kappa_1^2, \kappa_2)k/\sqrt{m})$.

Lemma 4. *Suppose A is δ-close to A^* and the bias satisfies $|\kappa_2\lambda_i^2 + 2\kappa_1 p_i b_i\lambda_i + b_i^2 - \kappa_2\lambda_i| \leq 2\kappa_2(1-\lambda_i)$, then*

$$2\langle g_i, W_i - A_i\rangle \geq \kappa_2 p_i(\lambda_i - 2\delta^2)\|W_i - A_i\|^2 + \frac{1}{\kappa_2 p_i\lambda_i}\|g_i\|^2 - O(\max(1, \kappa_2/\kappa_1^2)\frac{k^2}{p_i m})$$

The proof of this lemma and the descent is the same as that of Lemma 1 for the case of Gaussian mixture. Again, the condition for bias holds when $b_i = 0$ and the thresholding activation is used; but breaks down when the nonzero bias is set fixed across iterations.

Now, we give an analysis for a bias update. Similarly to the mixture of Gaussian case, the bias is updated as

$$b^{s+1} = b^s/C,$$

for some $C > 1$. The proof remains the same to guarantee the consistency and also the descent.

The last step is to maintain the nearness for the new update. Since it is tedious to argue that for the complicated form of g_i, we can instead perform a projection on convex set $\mathcal{B} = \{W | \text{W is } \delta\text{-close to } A^* \text{ and } \|W\| \leq 2\|A\|\}$ to guarantee the nearness. The details can be found in [16].

B.4 Auxiliary Lemma

In our descent analysis, we assume a normalization for W's columns after each descent update. The descent property is achieved for the unnormalized version and does not directly imply the δ-closeness for that current estimate. In fact, this is shown by the following lemma:

Lemma 5. *Suppose that $\|W_i^s\| = \|A_i\| = 1$ and $\|W_i^s - A_i\| \leq \delta_s$. The gradient update $\widetilde{W}_i^{s+1}$ satisfies $\|\widetilde{W}_i^{s+1} - A_i\| \leq (1-\tau)\|W_i^s - A_i\| + o(\delta_s)$. Then, for $\frac{1-\delta_s}{2-\delta_s} \leq \tau < 1$, we have*

$$\|W_i^{s+1} - A_i\| \leq (1 + o(1))\delta_s,$$

where $W_i^{s+1} = \frac{\widetilde{W}_i^{s+1}}{\|\widetilde{W}_i^{s+1}\|}$.

Proof. Denote $w = \|\widetilde{W}_i^{s+1}\|$. Using a triangle inequality and the descent property, we have

$$\begin{aligned}
\|\widetilde{W}_i^{s+1} - wA_i\| &= \|\widetilde{W}_i^{s+1} - A_i + (1-w)A_i\| \\
&\leq \|\widetilde{W}_i^{s+1} - A_i\| + \|(1-w)A_i\| \quad (\|A_i\| = 1) \\
&\leq (1-\tau)\|W_i^s - A_i\| + (1-\tau)\|W_i^s - A_i\| + o(\delta_s) \\
&\leq 2(1-\tau)\|W_i^s - A_i\| + o(\delta_s).
\end{aligned}$$

At the third step, we use $|1-w| \leq \|\widetilde{W}_i^{s+1} - A_i\| \leq (1-\tau)\|W_i^s - A_i\| + o(\delta_s)$. This also implies $w \geq 1-(1-\tau-o(1))\delta_s$. Therefore,

$$\begin{aligned}
\|W_i^{s+1} - A_i\| &\leq \frac{2(1-\tau)}{w}\|W_i^s - A_i\| + o(\delta_s) \\
&\leq \frac{2(1-\tau)}{(1+(1-\tau-o(1))\delta_s)}\|W_i^s - A_i\| + o(\delta_s).
\end{aligned}$$

This implies that when the condition $\frac{1+\delta_s}{2+\delta_s} \leq \tau < 1$ holds, we get:

$$\|W_i^{s+1} - A_i\| \leq (1+o(1))\delta_s.$$

■