

745 **Additional notation.** Throughout the appendices, to simplify the notation, we write

$$a_i(X) := p^\top E_{x_i}, \quad q_i(X) := \frac{\exp(a_i(X))}{\sum_{j=1}^T \exp(a_j(X))}, \quad (19)$$

746 so that  $f(X; p, \mathbf{E}) = \sum_{i=1}^T q_i(X) E_{x_i}^\top v$ . We will drop the dependence on  $X$  in  $a_i(X), q_i(X)$  when  
747 there is no confusion. We also denote

$$\gamma_i(X, y) := y E_{x_i}^\top v, \quad (20)$$

748 dropping again the dependency on  $X, y$  when there is no confusion. Finally, we define

$$g(X, y) := \frac{1}{1 + \exp(yf(X; p, \mathbf{E}))}. \quad (21)$$

749 **Properties of initialization.** By standard concentration inequalities, with probability at least  $1 - \delta$ ,  
750 at initialization we have

$$\begin{aligned} \max \left\{ \max_{s \neq s' \in \mathcal{S}} |E_s^\top E_{s'}|, \max_{s \in \mathcal{S}} |E_s^\top v|, \max_{s \in \mathcal{S}} |E_s^\top p|, |p^\top v| \right\} &\leq \frac{1}{\sqrt{d}} \sqrt{2 \log \frac{|\mathcal{S}|^2}{\delta}}, \\ \max \left\{ \max_{s \in \mathcal{S}} \|E_s\|_2, \|p\|_2 \right\} &\leq 2, \quad \min_{s \in \mathcal{S}} \|E_s\|_2 \geq \frac{1}{2}. \end{aligned} \quad (22)$$

751 For all results of the paper holding with probability at least  $1 - \delta$ , we will be implicitly conditioning  
752 on (22).

## 753 A Technical lemmas

754 **Lemma A.1.** *The gradients of the empirical loss are given by*

$$\begin{aligned} \nabla_{E_s} \mathcal{L}(\mathbf{E}, p) &= -\widehat{\mathbb{E}} \left[ yg(X, y) \left( \sum_{i=1}^T \left( \sum_{j \neq i} (\mathbb{1}_{x_i=s} - \mathbb{1}_{x_j=s}) q_i(X) q_j(X) \right) E_{x_i}^\top v p + \sum_{i=1}^T \mathbb{1}_{x_i=s} q_i v \right) \right], \\ \nabla_p \mathcal{L}(\mathbf{E}, p) &= -\widehat{\mathbb{E}} \left[ yg(X, y) \left( \sum_{i=1}^T \left( \sum_{j \neq i} q_i(X) q_j(X) (E_{x_i} - E_{x_j}) \right) E_{x_i}^\top v \right) \right], \end{aligned}$$

755 where we have defined  $g(X, y) = \frac{1}{1 + \exp(yf(X))}$ .

756 *Proof.* We start by taking the gradient of  $q_i$  as

$$\begin{aligned} \nabla_{E_s} q_i(X) &= \frac{\mathbb{1}_{x_i=s} \exp(E_{x_i}^\top p) p \left( \sum_{j=1}^T \exp(E_{x_j}^\top p) \right) - \left( \sum_{j=1}^T \mathbb{1}_{x_j=s} \exp(E_{x_j}^\top p) p \right) \exp(E_{x_i}^\top p)}{\left( \sum_{j=1}^T \exp(E_{x_j}^\top p) \right)^2} \\ &= \frac{p \sum_{j=1}^T (\mathbb{1}_{x_i=s} - \mathbb{1}_{x_j=s}) \exp(E_{x_j}^\top p) \exp(E_{x_i}^\top p)}{\left( \sum_{j=1}^T \exp(E_{x_j}^\top p) \right)^2} \\ &= p \left( \sum_{j=1}^T (\mathbb{1}_{x_i=s} - \mathbb{1}_{x_j=s}) q_i q_j \right) \\ &= p \left( \sum_{j \neq i} (\mathbb{1}_{x_i=s} - \mathbb{1}_{x_j=s}) q_i q_j \right), \end{aligned}$$

$$\begin{aligned}
\nabla_p q_i(X) &= \frac{(\exp(E_{x_i}^\top p) E_{x_i}) \left( \sum_{j=1}^T \exp(E_{x_j}^\top p) \right) - \sum_{j=1}^T \exp(E_{x_j}^\top p) E_{x_j} \exp(E_{x_i}^\top p)}{\left( \sum_{j=1}^T \exp(E_{x_j}^\top p) \right)^2} \\
&= \frac{\sum_{j=1}^T \exp(E_{x_j}^\top p) \exp(E_{x_i}^\top p) (E_{x_i} - E_{x_j})}{\left( \sum_{j=1}^T \exp(E_{x_j}^\top p) \right)^2} \\
&= \sum_{j=1}^T q_i q_j (E_{x_i} - E_{x_j}) \\
&= \sum_{j \neq i} q_i q_j (E_{x_i} - E_{x_j}).
\end{aligned}$$

758 Next, we look at the gradient of  $f(X; p, \mathbf{E})$ :

$$\begin{aligned}
\nabla_{E_s} f(X; p, \mathbf{E}) &= \sum_{i=1}^T (\nabla_{E_s} q_i) E_{x_i}^\top v + \sum_{i=1}^T \mathbb{1}_{x_i=s} q_i v \\
&= \sum_{i=1}^T \left( \sum_{j \neq i} (\mathbb{1}_{x_i=s} - \mathbb{1}_{x_j=s}) q_i q_j \right) E_{x_i}^\top v p + \sum_{i=1}^T \mathbb{1}_{x_i=s} q_i v, \\
\nabla_p f(X; p, \mathbf{E}) &= \sum_{i=1}^T \left( \sum_{j \neq i} q_i q_j (E_{x_i} - E_{x_j}) \right) E_{x_i}^\top v.
\end{aligned}$$

759 This allows us to conclude that

$$\begin{aligned}
\nabla_{E_s} \mathcal{L}(\mathbf{E}, p) &= \widehat{\mathbb{E}} \left[ \frac{-y}{1 + \exp(yf(X; p, \mathbf{E}))} \nabla_{E_s} f(X; p, \mathbf{E}) \right] \\
&= \widehat{\mathbb{E}} \left[ \frac{-y}{1 + \exp(yf(X; p, \mathbf{E}))} \left( \sum_{i=1}^T \left( \sum_{j \neq i} (\mathbb{1}_{x_i=s} - \mathbb{1}_{x_j=s}) q_i(X) q_j(X) \right) E_{x_i}^\top v p \right. \right. \\
&\quad \left. \left. + \sum_{i=1}^T \mathbb{1}_{x_i=s} q_i(X) v \right) \right], \\
\nabla_p \mathcal{L}(\mathbf{E}, p) &= \widehat{\mathbb{E}} \left[ \frac{-y}{1 + \exp(yf(X; p, \mathbf{E}))} \nabla_p f(X; p, \mathbf{E}) \right] \\
&= \widehat{\mathbb{E}} \left[ \frac{-y}{1 + \exp(yf(X; p, \mathbf{E}))} \left( \sum_{i=1}^T \left( \sum_{j \neq i} q_i(X) q_j(X) (E_{x_i} - E_{x_j}) \right) E_{x_i}^\top v \right) \right],
\end{aligned}$$

760 thus concluding the proof.  $\square$

761 **Lemma A.2.** For any vector  $\widehat{p}$ , we have

$$-\widehat{p}^\top \nabla_p \mathcal{L}(\mathbf{E}, p) = \widehat{\mathbb{E}} \left[ g(X, y) \left( \sum_{i=1}^T \sum_{j > i} (\widehat{a}_i(X) - \widehat{a}_j(X)) q_i(X) q_j(X) (\gamma_i(X, y) - \gamma_j(X, y)) \right) \right],$$

762 where  $\widehat{a}_i = \widehat{p}^\top E_{x_i}$  for all  $i \in \{1, \dots, T\}$ .

763 *Proof.* From Lemma A.1, we have

$$\begin{aligned}\nabla_p \mathcal{L}(\mathbf{E}, p) &= -\widehat{\mathbb{E}} \left[ yg(X, y) \left( \sum_{i=1}^T \left( \sum_{j \neq i} q_i(X) q_j(X) (E_{x_i} - E_{x_j}) \right) E_{x_i}^\top v \right) \right] \\ &= -\widehat{\mathbb{E}} \left[ g(X, y) \left( \sum_{i=1}^T \left( \sum_{j \neq i} q_i(X) q_j(X) (E_{x_i} - E_{x_j}) \right) \gamma_i(X, y) \right) \right] \\ &= -\widehat{\mathbb{E}} [g(X, y) \mathbf{E}_X^\top (\text{Diag}(q_X) - q_X q_X^\top) \gamma(X, y)],\end{aligned}$$

764 where  $q_X = [q_1(X), \dots, q_T(X)]^\top$ ,  $\gamma(X, y) = [\gamma_1(X, y), \dots, \gamma_T(X, y)]^\top$  and  $\text{Diag}(q_X)$  denotes the  
765 diagonal matrix with  $[\text{Diag}(q_X)]_{i,i} = q_i(X)$ .

766 Thus, letting  $\hat{a} = [\hat{a}_1, \dots, \hat{a}_T] \in \mathbb{R}^T$  with  $\hat{a}_i = \hat{p}^\top E_{x_i}$ , we have

$$\begin{aligned}-\hat{p}^\top \nabla_p \mathcal{L}(\mathbf{E}, p) &= \widehat{\mathbb{E}} [g(X, y) \hat{p}^\top \mathbf{E}_X^\top (\text{Diag}(q_X) - q_X q_X^\top) \gamma(X, y)] \\ &= \widehat{\mathbb{E}} [g(X, y) \hat{a}^\top (\text{Diag}(q_X) - q_X q_X^\top) \gamma(X, y)] \\ &= \widehat{\mathbb{E}} \left[ g(X, y) \left( \sum_{i=1}^T \hat{a}_i q_i (1 - q_i) \gamma_i - \sum_{i=1}^T \sum_{j \neq i} \hat{a}_i q_i q_j \gamma_j \right) \right] \\ &= \widehat{\mathbb{E}} \left[ g(X, y) \left( \sum_{i=1}^T \sum_{j \neq i} \hat{a}_i q_i q_j (\gamma_i - \gamma_j) \right) \right] \quad (\text{use } 1 - q_i = \sum_{j \neq i} q_j) \\ &= \widehat{\mathbb{E}} \left[ g(X, y) \left( \frac{1}{2} \sum_{i=1}^T \sum_{j \neq i} \hat{a}_i q_i q_j (\gamma_i - \gamma_j) + \frac{1}{2} \sum_{j=1}^T \sum_{i \neq j} \hat{a}_j q_i q_j (\gamma_j - \gamma_i) \right) \right] \\ &= \widehat{\mathbb{E}} \left[ g(X, y) \left( \frac{1}{2} \sum_{i=1}^T \sum_{j \neq i} (\hat{a}_i - \hat{a}_j) q_i q_j (\gamma_i - \gamma_j) \right) \right] \\ &= \widehat{\mathbb{E}} \left[ g(X, y) \left( \sum_{i=1}^T \sum_{j > i} (\hat{a}_i - \hat{a}_j) q_i q_j (\gamma_i - \gamma_j) \right) \right].\end{aligned}$$

767 □

768 **Lemma A.3** (Convergence lemma). *Let  $\|p_t\|_2 \rightarrow \infty$  and suppose there exists  $\hat{p}$  such that, for any*  
769  *$\epsilon > 0$ , there is a  $\bar{t}(\epsilon)$  ensuring*

$$-\frac{\hat{p}^\top}{\|\hat{p}\|_2} \nabla_p \mathcal{L}(\mathbf{E}, p_t) \geq -(1 - \epsilon) \frac{p_t^\top}{\|p_t\|_2} \nabla_p \mathcal{L}(\mathbf{E}, p_t), \quad \text{for all } t \geq \bar{t}(\epsilon). \quad (23)$$

770 *Then, if  $\lim_{t \rightarrow \infty} \frac{p_t}{\|p_t\|_2}$  exists, we have*

$$\lim_{t \rightarrow \infty} \frac{p_t}{\|p_t\|_2} = \frac{\hat{p}}{\|\hat{p}\|_2}.$$

771 *Proof.* By the definition of the gradient flow, (23) is equivalent to

$$\frac{\hat{p}^\top}{\|\hat{p}\|_2} \frac{dp_t}{dt} \geq (1 - \epsilon) \frac{p_t^\top}{\|p_t\|_2} \frac{dp_t}{dt}.$$

772 We note that

$$\frac{p_t^\top}{\|p_t\|_2} \frac{dp_t}{dt} = \frac{d}{dt} \|p_t\|_2.$$

773 Thus, by integrating both sides from  $[\bar{t}(\epsilon), t]$ , we have:

$$\frac{\hat{p}^\top}{\|\hat{p}\|_2} (p_t - p_{\bar{t}(\epsilon)}) \geq (1 - \epsilon) (\|p_t\|_2 - \|p_{\bar{t}(\epsilon)}\|_2),$$

774 which gives

$$\frac{\hat{p}^\top p_t}{\|\hat{p}\|_2 \|p_t\|_2} \geq (1 - \epsilon) - (1 - \epsilon) \frac{\|p_{\bar{t}(\epsilon)}\|_2}{\|p_t\|_2} + \frac{\hat{p}^\top p_{\bar{t}(\epsilon)}}{\|\hat{p}\|_2 \|p_t\|_2}.$$

775 Since  $p_{\bar{t}(\epsilon)}, \hat{p}$  have finite norm for fixed  $\epsilon$ , by taking the limit on both sides, we have

$$\liminf_{t \rightarrow \infty} \frac{\hat{p}^\top p_t}{\|\hat{p}\|_2 \|p_t\|_2} \geq 1 - \epsilon.$$

776 As we assume that  $\lim_{t \rightarrow \infty} \frac{p_t}{\|p_t\|_2}$  exist and the above argument holds for any  $\epsilon$ , we conclude

$$\lim_{t \rightarrow \infty} \frac{p_t}{\|p_t\|_2} = \frac{\hat{p}}{\|\hat{p}\|_2}.$$

777

□

778 **Lemma A.4.** Given a sequence  $X$ , model parameters  $\mathbf{E}, p, v$ , and indices  $i_*, j$  s.t.  $x_{i_*} \in \mathcal{S}_X(p), x_j \in$   
779  $X \setminus \mathcal{S}_X(p)$ , the following results hold.

780 1. We have

$$\frac{1}{T} \leq q_{i_*} \leq 1.$$

781 2. If there exist  $\tau > 0$  such that  $p^\top (E_{x_{i_*}} - E_{x_j}) \geq \tau$  for all  $x_{i_*} \in \mathcal{S}_X(p)$ , then we have

$$q_j \leq \frac{1}{1 + \exp(\tau)}.$$

782 3. If there exist  $\tau > 0$  such that  $p^\top (E_{x_{i_*}} - E_{x_j}) \leq \tau$  for all  $x_{i_*} \in \mathcal{S}_X(p)$ , then we have

$$q_j \geq \frac{1}{T \exp(\tau)}.$$

783 *Proof.* The upper bound on  $q_{i_*}$  is trivial. For the lower bound:

$$\begin{aligned} q_{i_*} &= \frac{\exp(p^\top E_{x_{i_*}})}{\exp(p^\top E_{x_{i_*}}) + \sum_{j \neq i_*} \exp(p^\top E_{x_j})} \\ &\geq \frac{\exp(p^\top E_{x_{i_*}})}{T \exp(p^\top E_{x_{i_*}})} = \frac{1}{T}. \end{aligned}$$

784 If there exists  $\tau > 0$  such that  $p^\top (E_{x_{i_*}} - E_{x_j}) \geq \tau$  for all  $x_i \in \mathcal{S}_X(p)$ , then we have

$$\begin{aligned} q_j &= \frac{1}{1 + \sum_{i \neq j} \exp(p^\top (E_{x_i} - E_{x_j}))} \\ &\leq \frac{1}{1 + \exp(p^\top (E_{x_{i_*}} - E_{x_j}))} \\ &\leq \frac{1}{1 + \exp(\tau)}. \end{aligned}$$

785 If there exists  $\tau > 0$  such that  $p^\top (E_{x_{i_*}} - E_{x_j}) \leq \tau$  for all  $x_{i_*} \in \mathcal{S}_X(p)$ , then we have

$$\begin{aligned} q_{i_*} &= \frac{1}{1 + \sum_{i \neq j} \exp(p^\top (E_{x_i} - E_{x_j}))} \\ &\geq \frac{1}{1 + (T - 1) \exp(p^\top (E_{x_{i_*}} - E_{x_j}))} \quad (\text{by definition of } \mathcal{S}_X(p)) \\ &\geq \frac{1}{T \exp(\tau)}. \end{aligned}$$

786

□

## 787 **B Properties after the first gradient step**

788 **Lemma B.1** (Boundedness of the embeddings). *For any  $\delta > 0$ , let*

$$d \geq \max \left\{ 256, \left( 2 \log \frac{|\mathcal{S}|^2}{\delta} \right)^2 \right\},$$

789 *then with probability at least  $1 - \delta$ ,*

$$\max_{s \in \mathcal{S}} \|E_s^1\|_2 \leq 2(1 + 2\eta_0), \quad \|p^1\|_2 \leq 2 + 11\eta_0 d^{-\frac{1}{4}}.$$

790 *Proof.* By using (22), we have that

$$\begin{aligned} \max_{s \in \mathcal{S}} \|E_s^1\|_2 &\leq \max_s \left( \|E_s^0\|_2 + \frac{\eta_0}{2} \|v\|_2 + \|err_s\|_2 \right) \\ &\leq \max_{s \in \mathcal{S}} \left( 2 + \frac{\eta_0}{2} + 11\eta_0 d^{-\frac{1}{4}} \right) \\ &\leq 2 + 4\eta_0, \end{aligned}$$

791 and that

$$\|p^1\|_2 \leq \|p^0\|_2 + \|err_p\|_2 \leq 2 + 11\eta_0 d^{-\frac{1}{4}}. \quad (24)$$

792 □

793 **Lemma B.2** (Upper bound on the loss). *For any  $\delta > 0$ , let*

$$d \geq \max \left\{ 256, \left( 2 \log \frac{|\mathcal{S}|^2}{\delta} \right)^2, (88\eta_0^2 + 111\eta_0 + 2)^8 \right\},$$

794 *then with probability at least  $1 - \delta$ ,*

$$\mathcal{L}(E^1, p^1) \leq \widehat{\mathbb{E}} \left[ \log \left( 1 + \exp \left( -\frac{1}{T} \sum_{i=1}^T \frac{\eta_0}{2} y \alpha_{x_i} + \frac{1}{22\eta_0} \right) \right) \right].$$

795 *Proof.* We first lower bound  $yf(X; p, E)$  for each pair  $X, y$ . After the first step, we have

$$\begin{aligned} \max_{s, s'} |(p^1)^\top (E_s^1 - E_{s'}^1)| &= \max_{s, s'} |(p^0)^\top (E_s^0 - E_{s'}^0) + \frac{\eta_0}{2} (\alpha_s - \alpha_{s'}) (p^0)^\top v \\ &\quad + err_p^\top (E_s^1 - E_{s'}^1) + (err_s - err_{s'})^\top p^1|. \end{aligned}$$

796 We bound each term separately:

$$\begin{aligned} \max_{s, s'} |(p^0)^\top (E_s^0 - E_{s'}^0)| &\leq 2 \max_s |(p^0)^\top E_s^0| \leq 2d^{-\frac{1}{4}}, \\ \frac{\eta_0}{2} (\alpha_s - \alpha_{s'}) |(p^0)^\top v| &\leq \eta_0 |(p^0)^\top v| \leq \eta_0 d^{-\frac{1}{4}}, \\ |err_p^\top (E_s^1 - E_{s'}^1)| &\leq \|err_p\|_2 \|E_s^1 - E_{s'}^1\|_2 \leq 44\eta_0 d^{-\frac{1}{4}} (1 + 2\eta_0), \\ |(err_s - err_{s'})^\top p^1| &\leq 2\|p^1\|_2 \max_s \|err_s\|_2 \leq 22\eta_0 d^{-\frac{1}{4}} (2 + 11\eta_0 d^{-\frac{1}{4}}), \end{aligned}$$

797 where we have used (22). By picking  $d \geq (88\eta_0^2 + 111\eta_0 + 2)^8$ , we get  $\max_{s, s'} |(p^1)^\top (E_s^1 - E_{s'}^1)| \leq$   
798  $d^{-\frac{1}{8}}$ , which implies that, for any  $X$  and any  $i \in \{1, \dots, T\}$ ,

$$\frac{1}{T} - \frac{2d^{-\frac{1}{8}}}{T} \leq q_i(X) \leq \frac{1}{T} + \frac{2d^{-\frac{1}{8}}}{T}.$$

799 Thus, we lower bound  $yf(X; p, \mathbf{E})$  for each pair  $(X, y)$  as

$$\begin{aligned}
yf(X; p, \mathbf{E}) &= \sum_{i=1}^T q_i(X) \gamma_i(X) \\
&\geq \frac{1}{T} \sum_{i=1}^T \frac{\eta_0}{2} y \alpha_{x_i} - \sum_{i=1}^T \frac{2d^{-\frac{1}{8}}}{T} \frac{\eta_0}{2} \alpha_{x_i} + \sum_{i=1}^T y q_i(X) v^\top (E_{x_i}^0 + \mathbf{err}_{x_i}) \\
&\geq \frac{1}{T} \sum_{i=1}^T \frac{\eta_0}{2} y \alpha_{x_i} - d^{-\frac{1}{8}} \eta_0 - (1 + 2d^{-\frac{1}{8}}) v^\top (E_{x_i}^0 + \mathbf{err}_{x_i}) \\
&\geq \frac{1}{T} \sum_{i=1}^T \frac{\eta_0}{2} y \alpha_{x_i} - d^{-\frac{1}{8}} \eta_0 - 3(1 + 11\eta_0) d^{-\frac{1}{4}} \\
&\geq \frac{1}{T} \sum_{i=1}^T \frac{\eta_0}{2} y \alpha_{x_i} - \frac{1}{22\eta_0},
\end{aligned}$$

800 which allows us to conclude that

$$\begin{aligned}
\mathcal{L}(\mathbf{E}^1, p^1) &= \widehat{\mathbb{E}} [\log(1 + \exp(-yf(X; p, \mathbf{E})))] \\
&\leq \widehat{\mathbb{E}} \left[ \log \left( 1 + \exp \left( -\frac{1}{T} \sum_{i=1}^T \frac{\eta_0}{2} y \alpha_{x_i} + \frac{1}{22\eta_0} \right) \right) \right]. \tag{25}
\end{aligned}$$

801

□

## 802 C Proofs for Section 4

### 803 C.1 Proof of Lemma 4.1

804 For simplicity, in the proof we drop the time dependency in all the variables. By picking

$$d \geq \left( 2 \log \frac{|S|^2}{\delta} \right)^2,$$

805 from (22) we have

$$\begin{aligned}
\max \left\{ \max_{s \in \mathcal{S}} |E_s^\top v|, \max_{s \in \mathcal{S}} |E_s^\top p|, |p^\top v| \right\} &\leq d^{-\frac{1}{4}}, \\
\max \left\{ \max_{s \in \mathcal{S}} \|E_s\|_2, \|p\|_2 \right\} &\leq 2.
\end{aligned}$$

806 Thus, at initialization, we have that, for all  $s$ ,

$$\exp \left( -d^{-\frac{1}{4}} \right) \leq \exp(p^\top E_s) \leq \exp \left( d^{-\frac{1}{4}} \right),$$

807 which implies that, for any sequence  $X$  and any position  $i$ ,

$$\frac{1}{T + 2T \left( d^{-\frac{1}{4}} \right)} \leq \frac{1}{1 + (T-1) \exp \left( 2d^{-\frac{1}{4}} \right)} \leq q_i(X) \leq \frac{1}{1 + (T-1) \exp \left( -2d^{-\frac{1}{4}} \right)} \leq \frac{1}{T - 2T \left( d^{-\frac{1}{4}} \right)},$$

808 where we use the fact that for  $z \in [-1, 1]$ ,  $1 - |z| \leq \exp(z) \leq 1 + |z|$ .

809 Furthermore, for  $d > 256$  and for any sequence  $(X, y)$ , we have

$$\frac{1}{T} - \frac{4d^{-\frac{1}{4}}}{T} \leq q_i(X) \leq \frac{1}{T} + \frac{4d^{-\frac{1}{4}}}{T},$$

810 and

$$-2d^{-\frac{1}{4}} \leq \frac{-Td^{-\frac{1}{4}}}{T - 2Td^{-\frac{1}{4}}} \leq yf(X; p, \mathbf{E}) \leq \frac{Td^{-\frac{1}{4}}}{T - 2Td^{-\frac{1}{4}}} \leq 2d^{-\frac{1}{4}}.$$

811 Then,

$$g(X, y) \leq \frac{1}{1 + \exp(-2d^{-\frac{1}{4}})} \leq \frac{1}{2 - 2d^{-\frac{1}{4}}} \leq \frac{1}{2} + d^{-\frac{1}{4}},$$

812 and similarly

$$g(X, y) \geq \frac{1}{2} - d^{-\frac{1}{4}}.$$

813 Now we look at the gradient update of the first step. By Lemma A.1, we have

$$\begin{aligned} -\nabla_{E_s} \mathcal{L}(\mathbf{E}, p) &= \widehat{\mathbb{E}} \left[ yg(X, y) \left( \sum_{i=1}^T \left( \sum_{j \neq i} (\mathbb{1}_{x_i=s} - \mathbb{1}_{x_j=s}) q_i q_j \right) E_{x_i}^\top v p + \sum_{i=1}^T \mathbb{1}_{x_i=s} q_i v \right) \right] \\ &= \frac{1}{2T} \widehat{\mathbb{E}} \left[ y \sum_{i=1}^T \mathbb{1}_{x_i=s} \right] v \\ &\quad + \frac{1}{2} \widehat{\mathbb{E}} \left[ y \sum_{i=1}^T \mathbb{1}_{x_i=s} \left( q_i - \frac{1}{T} \right) \right] v \\ &\quad + \widehat{\mathbb{E}} \left[ yg(X, y) \left( \sum_{i=1}^T \left( \sum_{j \neq i} (\mathbb{1}_{x_i=s} - \mathbb{1}_{x_j=s}) q_i q_j \right) E_{x_i}^\top v p \right) \right] \\ &\quad + \widehat{\mathbb{E}} \left[ y \left( g(X, y) - \frac{1}{2} \right) \sum_{i=1}^T \mathbb{1}_{x_i=s} q_i v \right], \\ -\nabla_p \mathcal{L}(\mathbf{E}, p) &= \widehat{\mathbb{E}} \left[ yg(X, y) \left( \sum_{i=1}^T \left( \sum_{j \neq i} q_i q_j (E_{x_i} - E_{x_j}) E_{x_i}^\top v \right) \right) \right]. \end{aligned}$$

814 We note that

$$\frac{1}{2T} \widehat{\mathbb{E}} \left[ y \sum_{i=1}^T \mathbb{1}_{x_i=s} \right] v = \frac{1}{2} \alpha_s v,$$

815 and we bound the remaining error terms.

816 We have that

$$\left\| \frac{1}{2} \widehat{\mathbb{E}} \left[ y \sum_{i=1}^T \mathbb{1}_{x_i=s} \left( q_i - \frac{1}{T} \right) \right] v \right\|_2 \leq d^{-\frac{1}{4}},$$

817 and

$$\begin{aligned} &\left\| \widehat{\mathbb{E}} \left[ yg(X, y) \left( \sum_{i=1}^T \left( \sum_{j \neq i} (\mathbb{1}_{x_i=s} - \mathbb{1}_{x_j=s}) q_i q_j \right) E_{x_i}^\top v p \right) \right. \right. \\ &\quad \left. \left. + y \left( g(X, y) - \frac{1}{2} \right) \sum_{i=1}^T \mathbb{1}_{x_i=s} q_i v \right] \right\|_2 \leq 10d^{-\frac{1}{4}}. \end{aligned}$$

818 Furthermore, we also have that

$$\|\nabla_p \mathcal{L}(\mathbf{E}, p)\|_2 \leq 8d^{-\frac{1}{4}}.$$

819 Thus, the desired claim follows.

## 820 C.2 Proof of Lemma 4.2

821 *Proof.* We first show that, if (12) is feasible, then the solution is unique. Indeed, assume by con-  
822 tradiction that  $p_1, p_2$  are two different solutions of (12). Clearly,  $p_1$  and  $p_2$  have the same norm, so

823  $\frac{p_1^\top p_2}{\|p_1\|_2 \|p_2\|_2} \neq 1$ . Then, any convex combination of  $p_1, p_2$  gives a feasible solution with a strictly  
 824 smaller norm, which is a contradiction.

825 Next, we show that (12) is always feasible. To see this, by definition, there exists some  $\tau$  such that

$$p_o^\top (E_s - E_{s'}) \geq \tau, \quad \forall s \in \mathcal{S}_X(p_o), \forall s' \in \overline{\mathcal{S}_X(p_o)}, \forall X \in \mathcal{X}_n.$$

826 Then,  $\frac{p_o}{\tau}$  is a feasible solution of (12) which concludes the proof of uniqueness.

827 To characterize  $p_*(p_o)$ , we first note that (12) can be equivalently written as:

$$\begin{aligned} & \arg \min_p \frac{1}{2} \|p\|_2^2 \\ & \text{s.t. } p^\top (E_s - E_{s'}) \geq 1, \quad \forall s \in \mathcal{S}_X(\mathcal{P}_{p_o}), \forall s' \in \overline{\mathcal{S}_X(\mathcal{P}_{p_o})}, \forall X \in \mathcal{X}_n. \end{aligned} \quad (26)$$

828 Now we characterize the solution of (26) explicitly. First of all, we can rewrite the constraints as

$$\mathbf{1}_N - \mathbf{M}p \leq 0.$$

829 Then we can write the Lagrangian of (26) as

$$L(p, \lambda) = \frac{1}{2} \|p\|_2^2 + \lambda^\top (\mathbf{1}_N - \mathbf{M}p),$$

830 where  $\lambda \in \mathbb{R}^N$  and  $p$  is a KKT point if

$$\begin{aligned} \nabla_p L(p, \lambda) &= p - \mathbf{M}^\top \lambda = 0, \\ \nabla_\lambda L(p, \lambda) &= \mathbf{1}_N - \mathbf{M}p = 0. \end{aligned}$$

831 Since the objective function is convex and the constraints are affine, the global optimum is achieved  
 832 at the KKT point, which satisfies  $\mathbf{M}p = \mathbf{1}_N$ . Thus, if there exists a  $p$  satisfying this condition, we  
 833 can rewrite (26) as

$$\begin{aligned} & \arg \min_p \frac{1}{2} \|p\|_2^2 \\ & \text{s.t. } \mathbf{M}p = \mathbf{1}_N, \end{aligned}$$

834 whose solution is

$$\hat{p} = \mathbf{M}^\dagger \mathbf{1}_N.$$

835 It remains to show that there exists a feasible  $p$ . Since  $d > |\mathcal{S}| + 2$ , we have that, with high-probability,  
 836  $\mathbf{E}^0$  is full rank. Furthermore,  $\mathbf{E}^1 = \mathbf{E}^0 + \mathbf{\Delta}$  and each row of  $\mathbf{\Delta}$  is in the subspace generated by  $v$   
 837 and  $p_0$ . Thus, we can pick  $\hat{p} \perp v, p_0$ , so that

$$\mathbf{E}^1 \hat{p} = \mathbf{E}^0 \hat{p}.$$

838 Then, we define  $a \in \mathbb{R}^{|\mathcal{S}|}$  such that  $a_i = 1$  for all  $i \in \bigcup_X \mathcal{S}_X(p_o)$ , and  $a_i = 0$  otherwise. Let

$$\mathbf{E}^0 \hat{p} = a.$$

839 Since  $d > |\mathcal{S}|$  and  $\mathbf{E}^0$  has full row rank, there exists a non-zero  $\hat{p}$  that solves the above equation,  
 840 which finishes the proof.

841 □

### 842 C.3 Proof of Theorem 4.3

843 We prove each part separately. We first show that  $\lim_{t \rightarrow \infty} \|p_t\|_2 = \infty$ .

844 **Lemma C.1.** *Under Assumption 1, for any  $\delta > 0$ , by picking*

$$d \geq \max \left\{ 256, \left( 2 \log \frac{|\mathcal{S}|^2}{\delta} \right)^2, |\mathcal{S}| + 3 \right\},$$

845 *with probability at least  $1 - \delta$ , we have  $\lim_{t \rightarrow \infty} \|p_t\|_2 = \infty$ .*



846 *Proof.* It is sufficient to show that there exists a non-zero finite-norm  $\hat{p}$ , such that for any finite norm  
 847  $p$ ,

$$\hat{p}^\top \nabla_p \mathcal{L}(\mathbf{E}^1, p) \neq 0.$$

848 Indeed, the above condition means that there is no stationary point for any finite-norm  $p$ . For gradient  
 849 flow, we have that

$$\lim_{t \rightarrow \infty} \nabla_p \mathcal{L}(\mathbf{E}^1, p_t) = 0,$$

850 which by contradiction implies the desired result.

851 Now we construct such  $\hat{p}$ . Since  $d > |\mathcal{S}| + 2$ , we have that with high-probability  $\mathbf{E}^0$  is full rank.  
 852 Furthermore,  $\mathbf{E}^1 = \mathbf{E}^0 + \Delta$  and each row of  $\Delta$  is in the subspace generated by  $v$  and  $p_0$ . Thus, we  
 853 can pick  $\hat{p} \perp v, p_0$ , so that

$$\mathbf{E}^1 \hat{p} = \mathbf{E}^0 \hat{p}.$$

854 Without loss of generality, let  $x_1$  be an important token in a positive sequence  $X_k$ , i.e.,  $\gamma_1(X_k) \geq \frac{\eta_0}{4nT}$ .  
 855 Then, we define  $a \in \mathbb{R}^{|\mathcal{S}|}$  such that  $a_1 = 1$  and  $a_i = 0$  for all  $i \neq 1$ . Let

$$\mathbf{E}^0 \hat{p} = a.$$

856 Since  $d > |\mathcal{S}|$  and  $\mathbf{E}^0$  has full row rank, there exists a non-zero  $\hat{p}$  that solves the above equation. By  
 857 Lemma A.2, we have that, for any  $p$ ,

$$\begin{aligned} -\hat{p}^\top \nabla_p \mathcal{L}(\mathbf{E}^1, p) &= \hat{\mathbb{E}} \left[ g(X, y) \left( \sum_{i=1}^T \sum_{j>i} (a_i - a_j) q_i q_j (\gamma_i - \gamma_j) \right) \right] \\ &= g(X_k, y_k) \sum_{j>1} q_1(X_k) q_j(X_k) \frac{\eta_0}{4nT} > 0, \end{aligned}$$

858 which concludes the proof.  $\square$

859 Next, we show that, if the directional limit exists, then it must select all completely positive/negative  
 860 tokens.

861 **Lemma C.2.** *Under Assumption 1, for any  $\delta > 0$ , by picking*

$$\eta_0 \geq 4n^2 T^2, \quad d \geq \max \left\{ 256, \left( 2 \log \frac{|\mathcal{S}|^2}{\delta} \right)^2, (88\eta_0^2 + 111\eta_0 + 2)^8 \right\},$$

862 *with probability at least  $1 - \delta$ , if  $p_\infty = \lim_{t \rightarrow \infty} \frac{p_t}{\|p_t\|_2}$  exists, then  $p_\infty$  satisfies*

$$s_*^X \in \mathcal{S}_X(p_\infty), \quad \text{for all } X \in \mathcal{X}_n,$$

863 *where  $s_*^X$  denotes the unique completely positive/negative token in the sequence  $X$ .*

864 *Proof.* We prove the lemma by contradiction. W.l.o.g., assume by contradiction that there exists  
 865  $X \in \mathcal{X}_n$  containing the important token  $x_1$  s.t.  $x_1 \notin \mathcal{S}_X(p_\infty)$ . We show that there exists  $\bar{t}$  such that,  
 866 for all  $t \geq \bar{t}$ ,

$$\mathcal{L}(\mathbf{E}^1, p^t) > \mathcal{L}(\mathbf{E}^1, p^1),$$

867 which contradicts the fact that the gradient flow always decreases the loss.

868 To see this, we first note that by the definition of  $\mathcal{S}_X(p_\infty)$ , there exists some  $\tau > 0$  independent of  $t$   
 869 such that

$$\min_{j \neq 1} p_\infty^\top (E_{x_1} - E_{x_j}) = -\tau.$$

870 W.l.o.g, we assume that  $x_2$  is the token that achieves the minimum.

871 As  $\lim_{t \rightarrow \infty} \|p_t\|_2 = \infty$  and  $\lim_{t \rightarrow \infty} \frac{p_t}{\|p_t\|_2} = p_\infty$ , we have that, for any  $\mu > 0, R > 0$ , there exists a  
 872 large enough  $\bar{t}$  such that

$$\|p_t\|_2 \geq 2R, \quad \left\| \frac{p_t}{\|p_t\|_2} - p_\infty \right\|_2 \leq \mu, \quad \text{for all } t \geq \bar{t}.$$

873 Thus, we have:

$$\begin{aligned} \frac{p_t^\top}{\|p_t\|_2} (E_{x_1} - E_{x_2}) &= p_\infty^\top (E_{x_1} - E_{x_2}) + \left( \frac{p_t}{\|p_t\|_2} - p_\infty \right)^\top (E_{x_1} - E_{x_2}) \\ &\leq -\tau + 2\mu(4\eta_0 + 2)^2, \end{aligned}$$

874 where we have used the result of Lemma B.1. Thus, by picking  $\mu = \frac{\tau}{4(4\eta_0 + 2)^2}$ , we have

$$\frac{p_t^\top}{\|p_t\|_2} (E_{x_1} - E_{x_2}) \leq -\frac{\tau}{2},$$

875 which implies that

$$p_t^\top (E_{x_1} - E_{x_2}) \leq -\tau R.$$

876 Next, we upper bound  $yf(X; p_t, \mathbf{E}^1)$ . We first note that

$$\frac{q_1}{q_2} = \exp(p_t^\top (E_{x_1} - E_{x_2})) \leq \exp(-\tau R),$$

877 which gives

$$q_1 \leq \exp(-\tau R).$$

878 Note that

$$\begin{aligned} yf(X; p_t, \mathbf{E}^1) &= \sum_{i=1}^T q_i \gamma_i \\ &\leq \exp(-\tau R) \gamma_1 + \max_{j \neq 1} \gamma_j \\ &\leq \exp(-\tau R) \left( \frac{\eta_0}{2} + (1 + 11\eta_0) d^{-\frac{1}{4}} \right) + (1 + 11\eta_0) d^{-\frac{1}{4}}. \end{aligned}$$

879 Thus, by picking  $R \geq \frac{\log d}{4\tau}$ , we have

$$yf(X; p_t, \mathbf{E}^1) \leq \left( \frac{3}{2} + \frac{23}{2} \eta_0 \right) d^{-\frac{1}{4}} \leq \frac{3}{4} d^{-\frac{1}{8}},$$

880 which implies a lower bound on the loss:

$$\mathcal{L}(\mathbf{E}^1, p_t) \geq \frac{1}{n} \log(1 + \exp(-yf(X; p_t, \mathbf{E}^1))) \geq \frac{1}{n} \log\left(1 + \exp\left(-\frac{3}{4} d^{-\frac{1}{8}}\right)\right) \geq \frac{1}{2n}, \quad (27)$$

881 where we used that  $d \geq 256$  in the last passage. Under Assumption 1, by Lemma 4.1, we have that  
 882  $y\alpha_{x_i} \geq 1/(nT)$  if  $x_i$  is either the completely positive or the completely negative token in  $X$ , and  
 883 otherwise  $y\alpha_{x_i} = 0$ . Hence, given that each sequence  $X$  contains a completely positive or negative  
 884 token, we have that

$$\frac{1}{T} \sum_{i=1}^T y\alpha_{x_i} \geq \frac{1}{nT^2}.$$

885 As  $\eta_0 > 4n^2T^2 > \sqrt{2nT^2/11}$ , by applying Lemma B.2, we obtain

$$\mathcal{L}(\mathbf{E}^1, p_1) \leq \log\left(1 + \exp\left(-\frac{\eta_0}{4nT^2}\right)\right) \leq \log(1 + \exp(-n)) \leq \exp(-n) < \frac{1}{2n},$$

886 which gives a contradiction and concludes the proof.  $\square$

887 Finally, we show that for each possible selection, if  $p_t$  converges in direction, it must converge to the  
 888 max-margin solution. In particular, we first prove the following lemma which gives an approximation  
 889 to the directional gradient of the locally optimal selection. To do so, we define the secondary selection  
 890 set and the locally optimal selection as follows:

891 **Definition C.3.** Given a vector  $p$ , for each sequence  $X$ , denote by  $\mathcal{S}_X^2(p)$  the secondary selection set  
 892 given by

$$\mathcal{S}_X^2(p) = \arg \max\{s : p^\top E_s, s \notin \mathcal{S}_X(p)\}. \quad (28)$$

893 We also denote by  $\mathcal{S}_X^<(p)$  the set of tokens that are not chosen in the first and in the second place, i.e.,

$$\mathcal{S}_X^<(p) = X \setminus (\mathcal{S}_X(p) \cup \mathcal{S}_X^2(p)). \quad (29)$$

894 **Definition C.4.** Given a vector  $p$ , we say that  $p$  is locally optimal if for every  $(X, y)$  pair, we have

$$\sum_{i \in \mathcal{S}_X(p)} (\gamma_i(X, y) - \gamma_j(X, y)) \geq \mu > 0, \quad \text{for all } j \in \mathcal{S}_X^2(p),$$

895 for some constant  $\mu$  that does not depends on  $p$ .

896 In the definition above and for the rest of this appendix, to help readability, we will abuse notation by  
897 letting indices (e.g.,  $i, j$  above) also denote the corresponding tokens (e.g.,  $x_i, x_j$  above).

898 **Lemma C.5.** Let  $\bar{p}$  be a unit-norm vector and  $p = R\bar{p}$  for some positive constant  $R$ . Suppose  $\bar{p}$  is  
899 a locally optimal direction as defined in Definition C.4 with some  $\mu$  that does not depends on  $R$ .  
900 Moreover, suppose there exists a constant  $\tau_1$  that may depend on  $\bar{p}, \eta_0, n, T, d$  but not on  $R$ , such  
901 that:

$$\begin{aligned} \min_X \{\bar{p}^\top (E_s - E_{s'}), \forall s \in \mathcal{S}_X(p), \forall s' \in \mathcal{S}_X^2(p)\} &\geq \tau_1, \\ \min_X \{\bar{p}^\top (E_s - E_{s'}), \forall s \in \mathcal{S}_X^2(p), \forall s' \in \mathcal{S}_X^<(p)\} &\geq \tau_1. \end{aligned} \quad (30)$$

902 Then, for any  $\epsilon > 0$ , for any  $\hat{p} \cong p$  such that  $\|\hat{p}\|_2$  does not depend on  $R$  and

$$\min_X \{\hat{p}^\top (E_s - E_{s'}), \forall s \in \mathcal{S}_X(\hat{p}), \forall s' \in X \setminus \mathcal{S}_X(\hat{p})\} \geq \tau_2,$$

903 there exists  $R$  large enough such that:

$$\begin{aligned} -\hat{p}^\top \nabla \mathcal{L}(\mathbf{E}^1, p) &\leq (1 + \epsilon) \widehat{\mathbb{E}} \left[ \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\hat{a}_i(X) - \hat{a}_j(X)) h_{i,j}(X, y, p) \right], \\ -\hat{p}^\top \nabla \mathcal{L}(\mathbf{E}^1, p) &\geq (1 - \epsilon) \widehat{\mathbb{E}} \left[ \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\hat{a}_i(X) - \hat{a}_j(X)) h_{i,j}(X, y, p) \right], \end{aligned}$$

904 where  $\hat{a}_i(X) = \hat{p}^\top E_{x_i}$ ,  $\hat{a}_j(X) = \hat{p}^\top E_{x_j}$  and

$$h_{i,j}(X, y, p) = g(X, y) q_i(X) q_j(X) (\gamma_i(X, y) - \gamma_j(X, y)).$$

905 *Proof.* By Lemma A.2, we can write the directional gradient as follows:

$$\begin{aligned} -\hat{p}^\top \nabla_p \mathcal{L}(\mathbf{E}^1, p) &= \widehat{\mathbb{E}} \left[ \sum_{i=1}^T \sum_{j>i} (\hat{a}_i(X) - \hat{a}_j(X)) h_{i,j}(X, y, p) \right] \\ &= \widehat{\mathbb{E}} \left[ \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\hat{a}_i(X) - \hat{a}_j(X)) h_{i,j}(X, y, p) \right] \end{aligned} \quad (\text{B0})$$

$$+ \widehat{\mathbb{E}} \left[ \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^<(p)} (\hat{a}_i(X) - \hat{a}_j(X)) h_{i,j}(X, y, p) \right] \quad (\text{B1})$$

$$+ \widehat{\mathbb{E}} \left[ \sum_{i \in X \setminus \mathcal{S}_X(p)} \sum_{j>i: j \in X \setminus \mathcal{S}_X(p)} (\hat{a}_i(X) - \hat{a}_j(X)) h_{i,j}(X, y, p) \right]. \quad (\text{B2})$$

906 The rest of the proof is to show that

$$\begin{aligned} -C_1 \exp(-\tau_1 R) (\text{B0}) &\leq (\text{B1}) \leq C_1 \exp(-\tau_1 R) (\text{B0}), \\ -C_2 \exp(-\tau_1 R) (\text{B0}) &\leq (\text{B2}) \leq C_2 \exp(-\tau_1 R) (\text{B0}), \end{aligned}$$

907 for some  $C_1, C_2 > 0$  that do not depend on  $R$ . Then, by taking  $R$  large enough, we obtain the desired  
908 result.

First, we simplify (B0). Note that, for all  $i, i_0 \in \mathcal{S}_X(p)$ , we have that  $\hat{a}_i(X) = \hat{a}_{i_0}(X)$ . Hence, by switching the order of  $i, j$ , we obtain

$$\begin{aligned} \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\hat{a}_i(X) - \hat{a}_j(X)) h_{i,j}(X, y, p) &= \sum_{j \in \mathcal{S}_X^2(p)} (\hat{a}_{i_0}(X) - \hat{a}_j(X)) \sum_{i \in \mathcal{S}_X(p)} h_{i,j}(X, y, p) \\ &= g(X, y) \sum_{j \in \mathcal{S}_X^2(p)} (\hat{a}_{i_0}(X) - \hat{a}_j(X)) q_{i_0}(X) q_j(X) \sum_{i \in \mathcal{S}_X(p)} (\gamma_i(X, y) - \gamma_j(X, y)), \end{aligned}$$

for any  $i_0 \in \mathcal{S}_X(p)$ . Since  $p$  is a locally optimal direction, we have

$$\sum_{i \in \mathcal{S}_X(p)} (\gamma_i(X, y) - \gamma_j(X, y)) \geq \mu, \quad \text{for all } j \in \mathcal{S}_X^2(p).$$

Now, we compare (B1) and (B0). By the exact same reason above, we can rewrite

$$\begin{aligned} \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\hat{a}_i(X) - \hat{a}_j(X)) h_{i,j}(X, y, p) \\ = g(X, y) \sum_{j \in \mathcal{S}_X^2(p)} (\hat{a}_{i_0}(X) - \hat{a}_j(X)) q_{i_0}(X) q_j(X) \sum_{i \in \mathcal{S}_X(p)} (\gamma_i(X, y) - \gamma_j(X, y)), \end{aligned}$$

for any  $i_0 \in \mathcal{S}_X(p)$ , and we compare to (B0) term-by-term. Namely, for any  $X, j \in \mathcal{S}_X^2(p)$  and  $k \in \mathcal{S}_X^<(p)$ , we have:

$$\frac{|\hat{a}_{i_0}(X) - \hat{a}_k(X)|}{\hat{a}_{i_0}(X) - \hat{a}_j(X)} \leq \frac{\|\hat{p}\|_2 \|E_{x_{i_0}} - E_{x_j}\|_2}{\tau_2} \leq \frac{2\|\hat{p}\|_2 \max_s \|E_s\|_2}{\tau_2} := C_3, \quad (31)$$

$$\frac{q_k(X)}{q_j(X)} = \exp(a_k(X) - a_j(X)) \leq \exp(-\tau_1 R), \quad (32)$$

$$\frac{\sum_{i \in \mathcal{S}_X(p)} |\gamma_i(X, y) - \gamma_k(X, y)|}{\sum_{i \in \mathcal{S}_X(p)} (\gamma_i(X, y) - \gamma_j(X, y))} \leq \frac{2T \max_s |\gamma_s|}{\mu} \leq \frac{2T \max_s \|E_s\|_2}{\mu} := C_4, \quad (33)$$

which implies that, for any  $X, j \in \mathcal{S}_X^2(p)$  and  $k \in \mathcal{S}_X^<(p)$ ,

$$\begin{aligned} |\hat{a}_{i_0}(X) - \hat{a}_k(X)| q_{i_0}(X) q_k(X) \sum_{i \in \mathcal{S}_X(p)} |\gamma_i(X, y) - \gamma_k(X, y)| \\ \leq \exp(-\tau_1 R) C_3 C_4 (\hat{a}_{i_0}(X) - \hat{a}_j(X)) q_{i_0}(X) q_j(X) \sum_{i \in \mathcal{S}_X(p)} (\gamma_i(X, y) - \gamma_j(X, y)). \end{aligned}$$

Thus, we get that:

$$|(\text{B1})| \leq \exp(-\tau_1 R) T C_3 C_4 |(\text{B0})|.$$

Next, we compare (B2) and (B0). Take any  $i' \in X \setminus \mathcal{S}_X(p), k > i' \in X \setminus \mathcal{S}_X(p), i_0 \in \mathcal{S}_X(p), j \in \mathcal{S}_X^2(p)$ . We compare

$$(\hat{a}_{i'}(X) - \hat{a}_k(X)) h_{i',k}(X, y, p)$$

with each term in (B1). We note that the bounds on  $\frac{\hat{a}_{i'}(X) - \hat{a}_k(X)}{\hat{a}_{i_0}(X) - \hat{a}_j(X)}$  and  $\frac{|\gamma_{i'}(X) - \gamma_k(X)|}{\sum_{i \in \mathcal{S}_X(p)} (\gamma_i(X, y) - \gamma_j(X, y))}$  are the same as those in (31) and (33). Furthermore,

$$\frac{q_{i'} q_k}{q_{i_0} q_j} \leq \exp(-\tau_1 R),$$

which gives that

$$|(\text{B2})| \leq T^2 \exp(-\tau_1 R) C_3 C_4 |(\text{B0})|,$$

thus concluding the proof.  $\square$

**Lemma C.6.** Under Assumption 1, for any  $\delta > 0$ , by picking

$$\eta_0 \geq 4n^2 T^2, \quad d \geq \max \left\{ 256, \left( 2 \log \frac{|S|^2}{\delta} \right)^2, (88\eta_0^2 + 111\eta_0 + 2)^8 \right\},$$

with probability  $\geq 1 - \delta$  over the initialization, if  $p_\infty = \lim_{t \rightarrow \infty} \frac{p_t}{\|p_t\|_2}$  exists, then  $p_\infty \in \mathcal{P}_*(p_\infty)$ .

926 *Proof.* We prove the lemma by contradiction. We first assume that there exists  $p_\infty$  such that  
 927  $p_\infty \notin \mathcal{P}_*(p_\infty)$  and  $p_\infty = \lim_{t \rightarrow \infty} \frac{p_t}{\|p_t\|_2}$ . Then, we show that there exists  $\hat{p} \in \mathcal{P}_*(p_\infty)$  such that, for  
 928 any  $\epsilon > 0$ , there is  $\bar{t}(\epsilon)$  ensuring

$$-\frac{\hat{p}^\top}{\|\hat{p}\|_2} \nabla_p \mathcal{L}(\mathbf{E}^1, p_t) \geq -(1 - \epsilon) \frac{p_t^\top}{\|p_t\|_2} \nabla_p \mathcal{L}(\mathbf{E}^1, p_t), \quad \text{for all } t \geq \bar{t}(\epsilon).$$

929 As a consequence, by Lemma A.3, we have that  $p_\infty = \frac{\hat{p}}{\|\hat{p}\|_2}$ , which gives a contradiction.

930 For the rest of the proof, we fix any  $\epsilon > 0$  and denote  $R = \|p_t\|_2$ . We define  $\bar{p}_t = \frac{p_t \|\hat{p}\|_2}{\|p_t\|_2}$ , and we  
 931 equivalently show that:

$$-\hat{p}^\top \nabla_p \mathcal{L}(\mathbf{E}^1, p_t) \geq -(1 - \epsilon) \bar{p}_t^\top \nabla_p \mathcal{L}(\mathbf{E}^1, p_t). \quad (34)$$

932 To prove this, we first note that since  $p_\infty \notin \mathcal{P}_*(p_\infty)$ , for all  $\frac{\hat{p}}{\|\hat{p}\|_2} \in \mathcal{P}_*(p_\infty)$ , there exists  $\tau_0$   
 933 independent of  $R$  such that

$$\|\hat{p} - p_\infty\|_2 \geq \tau_0.$$

934 Thus, by the definition of  $\mathcal{P}_*(p_\infty)$ , there exists  $\mathcal{X}_0 \subseteq \mathcal{X}_n$  such that for each sequence  $X \in \mathcal{X}_0$ , we  
 935 can find a pair of indices  $(i, j)$  with  $i \in \mathcal{S}_X(p_\infty)$ ,  $j \in X \setminus \mathcal{S}_X(p_\infty)$  violating the margin, i.e.,

$$(\|\hat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j}) \leq 1 - 3\tau,$$

936 for some  $\tau < \frac{1}{6}$  that does not depend on  $R$ . With a slight abuse of notation, we define  $\tau$  as

$$\begin{aligned} \tau = & \frac{1}{3} \min \left\{ \min_{X \in \mathcal{X}_0} \{1 - (\|\hat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j}), i \in \mathcal{S}_X(p_\infty), j \in \mathcal{S}_X^2(p_\infty)\}, \right. \\ & \min_{X \in \mathcal{X}_n} \{(\|\hat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j}), i \in \mathcal{S}_X(p_\infty), j \in \mathcal{S}_X^2(p_\infty)\}, \\ & \left. \min_{X \in \mathcal{X}_n} \{(\|\hat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j}), i \in \mathcal{S}_X^2(p_\infty), j \in \mathcal{S}_X^<(p_\infty)\} \right\}. \end{aligned}$$

937 This means that, for all  $X \in \mathcal{X}_n$  and for all  $(i, j)$  pairs such that  $i \in \mathcal{S}_X(p_\infty)$ ,  $j \in \mathcal{S}_X^2(p_\infty)$ , we have

$$(\|\hat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j}) \geq 3\tau;$$

938 for all pairs  $(i, j)$  such that  $i \in \mathcal{S}_X^2(p_\infty)$ ,  $j \in \mathcal{S}_X^<(p_\infty)$ , we have

$$(\|\hat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j}) \geq 3\tau;$$

939 and for all  $X \in \mathcal{X}_0$ ,  $i \in \mathcal{S}_X(p_\infty)$ ,  $j \in \mathcal{S}_X^2(p_\infty)$ , we have

$$(\|\hat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j}) \leq 1 - 3\tau,$$

940 with some  $\tau$  that does not depend on  $R$ .

941 Now, we compute the overlap with  $\bar{p}_t$ . For all  $X$  and  $(i, j)$ , we have

$$\bar{p}_t^\top (E_{x_i} - E_{x_j}) = (\|\hat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j}) + (\bar{p}_t - \|\hat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j}).$$

942 We upper bound

$$|(\bar{p}_t - \|\hat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j})| \leq \|\hat{p}\|_2 \left\| \frac{p_t}{\|p_t\|_2} - p_\infty \right\|_2 \|E_{x_i} - E_{x_j}\|_2,$$

943 and since  $\|\hat{p}\|_2, \|E_{x_1} - E_{x_2}\|_2$  are finite, we have

$$\lim_{t \rightarrow \infty} |(\bar{p}_t - \|\hat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j})| = 0.$$

944 Thus, we can pick  $t_1$ , such that for  $t \geq t_1$ , we have

$$|(\bar{p}_t - \|\hat{p}\|_2 p_\infty)^\top (E_{x_i} - E_{x_j})| \leq \tau,$$

945 which implies that, for all  $X \in \mathcal{X}_n$  and for all  $(i, j)$  pairs such that  $i \in \mathcal{S}_X(p_\infty)$ ,  $j \in \mathcal{S}_X^2(p_\infty)$ , we  
 946 have

$$\bar{p}_t^\top (E_{x_i} - E_{x_j}) \geq \tau;$$

947 for all  $(i, j)$  pairs such that  $i \in \mathcal{S}_X^2(p_\infty), j \in \mathcal{S}_X^<(p_\infty)$ , we have

$$\bar{p}_t^\top (E_{x_i} - E_{x_j}) \geq \tau;$$

948 and for all  $X \in \mathcal{X}_0, i \in \mathcal{S}_X(p_\infty), j \in \mathcal{S}_X^2(p_\infty)$ , we have:

$$\bar{p}_t^\top (E_{x_i} - E_{x_j}) \leq 1 - \tau,$$

949 for some  $\tau$  that does not depend on  $R$ .

950 Next, we show that  $\bar{p}_t$  is a locally optimal solution as per Definition C.4. By Lemma C.1,  $p_\infty$  selects  
 951 all the completely positive/negative tokens. Thus, as  $\bar{p}_t \approx p_\infty$ ,  $\bar{p}_t$  also selects such tokens, the rest  
 952 being irrelevant by Assumption 1. Hence, for any pair  $(X, y)$  and for any  $j \in X \setminus \mathcal{S}_X(p_t)$ , we have:

$$\sum_{i \in \mathcal{S}_X(p_t)} (\gamma_i(X, y) - \gamma_j(X, y)) \geq \frac{\eta_0}{4nT},$$

953 by picking  $d$  large enough (as per the hypothesis of the lemma). By construction,  $\hat{p} \cong p_t$ ,  $\|\hat{p}\|_2$  does  
 954 not depends on  $R$  and, moreover, for any  $X$ ,

$$\hat{p}^\top (E_{x_i} - E_{x_j}) \geq 1, \quad \text{for all } i \in \mathcal{S}_X(\hat{p}), j \in X \setminus \mathcal{S}_X(\hat{p}).$$

955 By applying Lemma C.5 on both  $\hat{p}$  and  $\bar{p}_t$ , we have that for any  $\epsilon_1 > 0$  there exist  $t_2$  s.t. for all  
 956  $t \geq \max\{t_1, t_2\}$ , we have

$$\begin{aligned} -\hat{p}^\top \nabla_p \mathcal{L}(\mathbf{E}^1, p_t) &\geq (1 - \epsilon_1) \widehat{\mathbb{E}} \left[ \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\hat{a}_i(X) - \hat{a}_j(X)) h_{i,j}(X, y, p_t) \right], \\ -\bar{p}_t^\top \nabla_p \mathcal{L}(\mathbf{E}^1, p_t) &\leq (1 + \epsilon_1) \widehat{\mathbb{E}} \left[ \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\bar{a}_i(X) - \bar{a}_j(X)) h_{i,j}(X, y, p_t) \right], \end{aligned}$$

957 where  $\bar{a}_i(X), \bar{a}_j(X)$  are defined analogously to  $\hat{a}_i(X), \hat{a}_j(X)$  by replacing  $\hat{p}$  with  $\bar{p}_t$ .

958 Now, we further show that, for any  $\epsilon_2 > 0$ , there exist  $t_3$  such that for all  $t \geq t_3$ ,

$$\begin{aligned} &\widehat{\mathbb{E}} \left[ \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\bar{a}_i(X) - \bar{a}_j(X)) h_{i,j}(X, y, p_t) \right] \\ &\leq (1 + \epsilon_2) \widehat{\mathbb{E}}_{X \in \mathcal{X}_0} \left[ \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\bar{a}_i(X) - \bar{a}_j(X)) h_{i,j}(X, y, p_t) \right]. \end{aligned}$$

959 To see this, we use the same idea as in the proof of Lemma C.5. We can write

$$\begin{aligned} &\widehat{\mathbb{E}} \left[ \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\bar{a}_i(X) - \bar{a}_j(X)) h_{i,j}(X, y, p_t) \right] \\ &= \widehat{\mathbb{E}}_{X \in \mathcal{X}_0} \left[ \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\bar{a}_i(X) - \bar{a}_j(X)) h_{i,j}(X, y, p_t) \right] \end{aligned} \quad (\text{A0})$$

$$+ \widehat{\mathbb{E}}_{X' \in \mathcal{X}_n \setminus \mathcal{X}_0} \left[ \sum_{i \in \mathcal{S}_{X'}(p)} \sum_{j \in \mathcal{S}_{X'}^2(p)} (\bar{a}_i(X') - \bar{a}_j(X')) h_{i,j}(X', y', p_t) \right], \quad (\text{A1})$$

960 and it is sufficient to show that

$$(\text{A1}) \leq \epsilon_2 (\text{A0}).$$

961 To prove this, we compare term-by-term. Let  $X \in \mathcal{X}_0, X' \in \mathcal{X}_n \setminus \mathcal{X}_0, j \in \mathcal{S}_X^2(p_t), j' \in \mathcal{S}_{X'}^2(p_t)$ , and  
 962 recall that:

$$\begin{aligned} &\sum_{i \in \mathcal{S}_X(p_t)} (\bar{a}_i(X) - \bar{a}_j(X)) h_{i,j}(X, y, p_t) \\ &= g(X, y) (\bar{a}_{i_0}(X) - \bar{a}_j(X)) q_{i_0}(X) q_j(X) \sum_{i \in \mathcal{S}_X(p_t)} (\gamma_i(X, y) - \gamma_j(X, y)), \end{aligned}$$

963

$$\begin{aligned} & \sum_{i \in \mathcal{S}_{X'}(p_t)} (\bar{a}_i(X') - \bar{a}_{j'}(X')) h_{i,j'}(X', y', p_t) \\ &= g(X', y') (\bar{a}_{i_1}(X') - \bar{a}_{j'}(X')) q_{i_1}(X') q_{j'}(X') \sum_{i \in \mathcal{S}_{X'}(p_t)} (\gamma_i(X', y') - \gamma_{j'}(X', y')), \end{aligned}$$

964 for any  $i_0 \in \mathcal{S}_X(p_t), i_1 \in \mathcal{S}_{X'}(p_t)$ . Note that

$$\frac{g(X', y')}{g(X, y)} \leq \frac{\max_{X,y} g(X, y)}{\min_{X,y} g(X, y)} \leq \max_{X,y} (1 + \exp(yf(X))) \leq (1 + \exp(\eta_0)) := C_5.$$

965 By using the same argument as in (31) and (33), we have

$$\begin{aligned} & \frac{\bar{a}_{i_1}(X') - \bar{a}_{j'}(X')}{\bar{a}_{i_0}(X) - \bar{a}_j(X)} \leq C_3, \\ & \frac{\sum_{i \in \mathcal{S}_{X'}(p_t)} (\gamma_i(X', y') - \gamma_{j'}(X', y'))}{\sum_{i \in \mathcal{S}_X(p_t)} (\gamma_i(X, y) - \gamma_j(X, y))} \leq C_4. \end{aligned}$$

966 Finally, we need to upper bound:

$$\frac{q_{i_1}(X') q_{j'}(X')}{q_{i_0}(X) q_j(X)}.$$

967 We note that

$$\begin{aligned} a_{i_1}(X') - a_{j'}(X') &\geq R/\|\hat{p}\|_2, \\ a_{i_0}(X) - a_j(X) &\leq (1 - \tau)R/\|\hat{p}\|_2, \end{aligned}$$

968 where  $a_i(X) = p_t^\top E_{x_i}$ . Thus by Lemma A.4, we have:

$$q_{i_0}(X) \geq \frac{1}{T}, \quad q_j(X) \geq \frac{1}{T \exp((1 - \tau)R/\|\hat{p}\|_2)}, \quad q_{i_1}(X') \leq 1, \quad q_{j'}(X') \leq \frac{1}{\exp(R/\|\hat{p}\|_2)},$$

969 which implies that

$$\frac{q_{i_1}(X') q_{j'}(X')}{q_{i_0}(X) q_j(X)} \leq T^2 \exp(-\tau R/\|\hat{p}\|_2).$$

970 Thus, for each  $X \in \mathcal{X}_0, X' \in \mathcal{X}_n \setminus \mathcal{X}_0, j \in \mathcal{S}_X^2(p_t), j' \in \mathcal{S}_{X'}^2(p_t)$ , we have

$$\sum_{i \in \mathcal{S}_{X'}(p)} (\bar{a}_i(X') - \bar{a}_{j'}(X')) h_{i,j'}(X', y', p_t) \leq C_6 \exp(-\tau R/\|\hat{p}\|_2) \sum_{i \in \mathcal{S}_X(p)} (\bar{a}_i(X) - \bar{a}_j(X)) h_{i,j}(X, y, p_t).$$

971 Thus by picking large enough  $t_3$  which gives large enough  $R$ , we have:

$$(A1) \leq \epsilon_2(A0).$$

972 This allows us to conclude that

$$\begin{aligned} -\hat{p}^\top \nabla_p \mathcal{L}(\mathbf{E}^1, p_t) &\geq (1 - \epsilon_1) \widehat{\mathbb{E}} \left[ \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\hat{a}_i(X) - \hat{a}_j(X)) h_{i,j}(X, y, p_t) \right] \\ &\geq (1 - \epsilon_1) \widehat{\mathbb{E}}_{X \in \mathcal{X}_0} \left[ \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\hat{a}_i(X) - \hat{a}_j(X)) h_{i,j}(X, y, p_t) \right], \\ -\bar{p}_t^\top \nabla_p \mathcal{L}(\mathbf{E}^1, p_t) &\leq (1 + \epsilon_1)(1 + \epsilon_2) \widehat{\mathbb{E}}_{X \in \mathcal{X}_0} \left[ \sum_{i \in \mathcal{S}_X(p)} \sum_{j \in \mathcal{S}_X^2(p)} (\bar{a}_i(X) - \bar{a}_j(X)) h_{i,j}(X, y, p_t) \right]. \end{aligned}$$

973 Note that, for each  $X \in \mathcal{X}_0$ ,

$$\hat{a}_i(X) - \hat{a}_j(X) \geq 1, \quad \bar{a}_i(X) - \bar{a}_j(X) \leq 1 - \tau,$$

974 which gives that

$$-\hat{p}^\top \nabla_p \mathcal{L}(\mathbf{E}^1, p_t) \geq -\frac{1 - \epsilon_1}{(1 + \epsilon_1)(1 + \epsilon_2)(1 - \tau)} \bar{p}_t^\top \nabla_p \mathcal{L}(\mathbf{E}^1, p_t).$$

975 Since  $\epsilon_1, \epsilon_2$  can be arbitrarily small, the proof is complete.  $\square$

976 **C.4 Proof of Lemma 4.5**

977 *Proof.* Let  $\hat{p}'$  be the max-margin solution of (12) with a different selection. By Theorem 4.3, we  
 978 have that, for all  $X, s_*^X \in \mathcal{S}_X(\hat{p}')$ . We denote by  $i_*^X$  the index of  $s_*^X$ . Assume by contradiction  
 979  $p_\infty = \frac{\|\hat{p}'\|_2}{\|\hat{p}\|_2}$ . We will now show that this implies the following statement: for any  $\epsilon > 0$ , there is a  
 980  $t(\epsilon)$  ensuring

$$-\frac{\hat{p}^\top}{\|\hat{p}\|_2} \nabla_p \mathcal{L}(\mathbf{E}, p_t) \geq -(1 - \epsilon) \frac{p_t^\top}{\|p_t\|_2} \nabla_p \mathcal{L}(\mathbf{E}, p_t), \quad \text{for all } t \geq t(\epsilon). \quad (35)$$

981 Then, by Lemma A.3, we have that  $p_\infty = \frac{\|\hat{p}\|_2}{\|\hat{p}'\|_2}$ , which gives a contradiction.

982 As in the proof of Lemma C.6, we define  $\bar{p}_t = \frac{p_t}{\|p_t\|_2} \|\hat{p}\|_2$ . Thus, (35) is equivalent to

$$-\hat{p}^\top \nabla_p \mathcal{L}(\mathbf{E}, p_t) \geq -(1 - \epsilon) \bar{p}_t^\top \nabla_p \mathcal{L}(\mathbf{E}, p_t).$$

983 First of all, since  $\hat{p}, \hat{p}'$  are two max-margin solutions, by Lemma 4.2 we have that all the constraints  
 984 are tight, that is:

$$\begin{aligned} \hat{p}^\top (E_{s_*^X} - E_s) &= 1, & \forall s \in X \setminus s_*^X, \forall X \in \mathcal{X}_n, \\ (\hat{p}')^\top (E_s - E_{s'}) &= 1, & \forall s \in \mathcal{S}_X(\hat{p}'), \forall s' \in X \setminus \mathcal{S}_X(\hat{p}'), \forall X \in \mathcal{X}_n, \end{aligned}$$

985 which implies that

$$\frac{\hat{p}'^\top \|\hat{p}\|_2}{\|\hat{p}'\|_2} (E_s - E_{s'}) = \frac{\|\hat{p}\|_2}{\|\hat{p}'\|_2} = 1 - \mu < 1, \quad \forall s \in \mathcal{S}_X(\hat{p}'), \forall s' \in X \setminus \mathcal{S}_X(\hat{p}'), \forall X \in \mathcal{X}_n.$$

986 As  $\|\bar{p}_t\|_2 = \|\hat{p}\|_2 < \|\hat{p}'\|_2$ ,  $\bar{p}_t$  violates the max-margin condition. Moreover, as  $\lim_{t \rightarrow \infty} \bar{p}_t =$   
 987  $\frac{\hat{p}'^\top \|\hat{p}\|_2}{\|\hat{p}'\|_2}$ , for any  $\epsilon_1 \in (0, \mu)$ , there exists a  $t_1$  ensuring the following for all  $t \geq t_1$ : for all  $(i, j)$  pairs  
 988 such that  $i \in \mathcal{S}_X(p_\infty), j \in \mathcal{S}_X^2(p_\infty)$ , we have

$$\bar{p}_t^\top (E_{x_i} - E_{x_j}) \leq 1 - \mu + \epsilon_1 \leq 1.$$

989 By applying Lemma C.5 to  $\bar{p}_t$ , we obtain that, for any  $\epsilon_2 > 0$ , there exists a  $t_2$  ensuring that, for all  
 990  $t \geq t_2$ ,

$$\begin{aligned} -\bar{p}_t^\top \nabla_p \mathcal{L}(\mathbf{E}^1, p_t) &\leq (1 + \epsilon_2) \widehat{\mathbb{E}} \left[ g(X, y) \sum_{i \in \mathcal{S}_X(p_t)} \sum_{j \in \mathcal{S}_X^2(p_t)} (\bar{a}_i(X) - \bar{a}_j(X) q_i(X) q_j(X) (\gamma_i(X) - \gamma_j(X))) \right] \\ &= (1 + \epsilon_2) \widehat{\mathbb{E}} \left[ g(X, y) \sum_{j \in \mathcal{S}_X^2(p_t)} (\bar{a}_{i_*^X}(X) - \bar{a}_j(X) q_{i_*^X}(X) q_j(X) (\gamma_{i_*^X}(X) - \gamma_j(X))) \right] \\ &\quad + (1 + \epsilon_2) \widehat{\mathbb{E}} \left[ g(X, y) \sum_{i \in \mathcal{S}_X(p_t), i \neq i_*^X} \sum_{j \in \mathcal{S}_X^2(p_t)} (\bar{a}_i(X) - \bar{a}_j(X) q_i(X) q_j(X) (\gamma_i(X) - \gamma_j(X))) \right]. \end{aligned}$$

991 We then compute by Lemma A.2 that

$$\begin{aligned} -\hat{p}^\top \nabla_p \mathcal{L}(\mathbf{E}^1, p_t) &= \widehat{\mathbb{E}} \left[ g(X, y) \sum_{j \in X \setminus \mathcal{S}_X(p_t)} (\widehat{a}_{i_*^X}(X) - \widehat{a}_j(X) q_{i_*^X}(X) q_j(X) (\gamma_{i_*^X}(X) - \gamma_j(X))) \right] \\ &\geq \widehat{\mathbb{E}} \left[ g(X, y) \sum_{j \in \mathcal{S}_X^2(p_t)} (\widehat{a}_{i_*^X}(X) - \widehat{a}_j(X) q_{i_*^X}(X) q_j(X) (\gamma_{i_*^X}(X) - \gamma_j(X))) \right] \\ &= (1 - \mu) \widehat{\mathbb{E}} \left[ g(X, y) \sum_{j \in \mathcal{S}_X^2(p_t)} (\widehat{a}_{i_*^X}(X) - \widehat{a}_j(X) q_{i_*^X}(X) q_j(X) (\gamma_{i_*^X}(X) - \gamma_j(X))) \right] \\ &\quad + \mu \widehat{\mathbb{E}} \left[ g(X, y) \sum_{j \in \mathcal{S}_X^2(p_t)} (\widehat{a}_{i_*^X}(X) - \widehat{a}_j(X) q_{i_*^X}(X) q_j(X) (\gamma_{i_*^X}(X) - \gamma_j(X))) \right], \end{aligned}$$



992 where in the first equality we use the fact that, for all  $j, j' \neq i_*^X$ ,  $\widehat{a_{j'}}(X) - \widehat{a_j}(X) = 0$ , and in the  
 993 second inequality we use the fact that all the terms in the summand are positive.

994 We note that:

$$\begin{aligned} & (1 + \epsilon_2) \widehat{\mathbb{E}} \left[ g(X, y) \sum_{j \in \mathcal{S}_X^2(p_t)} (\overline{a_{i_*^X}}(X) - \overline{a_j}(X)) q_{i_*^X}(X) q_j(X) (\gamma_{i_*^X}(X) - \gamma_j(X)) \right] \\ & < (1 - \mu) \widehat{\mathbb{E}} \left[ g(X, y) \sum_{j \in \mathcal{S}_X^2(p_t)} (\widehat{a_{i_*^X}}(X) - \widehat{a_j}(X)) q_{i_*^X}(X) q_j(X) (\gamma_{i_*^X}(X) - \gamma_j(X)) \right], \end{aligned}$$

995 as

$$\widehat{a_{i_*^X}}(X) - \widehat{a_j}(X) = 1, \quad \overline{a_{i_*^X}}(X) - \overline{a_j}(X) \leq 1 - \mu - \epsilon_1.$$

996 It remains to show that

$$\begin{aligned} & \mu \widehat{\mathbb{E}} \left[ g(X, y) \sum_{j \in \mathcal{S}_X^2(p_t)} (\widehat{a_{i_*^X}}(X) - \widehat{a_j}(X)) q_{i_*^X}(X) q_j(X) (\gamma_{i_*^X}(X) - \gamma_j(X)) \right] \\ & \geq (1 + \epsilon_2) \widehat{\mathbb{E}} \left[ g(X, y) \sum_{i \in \mathcal{S}_X(p_t) : i \neq i_*^X} \sum_{j \in \mathcal{S}_X^2(p_t)} (\overline{a_i}(X) - \overline{a_j}(X)) q_i(X) q_j(X) (\gamma_i(X) - \gamma_j(X)) \right]. \end{aligned} \tag{36}$$

997 We have that, for each  $i \in \mathcal{S}_X(p_t) : i \neq i_*^X, j \in \mathcal{S}_X^2(p_t)$ ,

$$|\gamma_i(X) - \gamma_j(X)| \leq 2d^{-1/4}, \quad \gamma_{i_*^X}(X) - \gamma_j(X) \geq \frac{\eta_0}{4}.$$

998 As  $d \geq \left( \frac{8T(1-\mu)}{\mu\eta_0} \right)^4$ , (36) holds and the proof is complete.  $\square$

## 999 D Details of numerical experiments

For all numerical simulations, we use the AdamW optimizer from `torch.optim`, and we reduce the learning rate in a multiplicative fashion by a factor  $\gamma = 0.1$  at epochs 100 and 200, i.e.,

$$\text{LR}_{\text{new}} = \text{LR}_{\text{old}} \cdot \gamma.$$

1000 We adhere to the batch size of 128 and fix the embedding dimension to 2048.

1001 **IMDB and Yelp datasets.** The hyperparameters *do not* differ between the two-layer model and the  
 1002 one-layer model. We set the number of training epochs to 500, the learning rate to 0.01, and the  
 1003 weight decay to  $10^{-8}$ .

1004 **Synthetic data.** We set the number of training epochs to 196, the learning rate to  $10^{-4}$ , and the  
 1005 weight decay to  $10^{-4}$ .