# A Theory of Unsupervised Speech Recognition

**Anonymous ACL submission**

## Abstract

Unsupervised speech recognition (ASR-U) is the problem of learning automatic speech recognition (ASR) systems from *unpaired* speech-only and text-only corpora. While various algorithms exist to solve this problem, a theoretical framework is missing to study their properties and address such issues as sensitivity to hyperparameters and training instability. In this paper, we proposed a general theoretical framework to study the properties of ASR-U systems based on random matrix theory and the theory of neural tangent kernels. Such a framework allows us to prove various learnability conditions and sample complexity bounds of ASR-U. Extensive ASR-U experiments on synthetic languages with three classes of transition graphs provide strong empirical evidence for our theory (code available at here).

## 1 Introduction

Unsupervised speech recognition (ASR-U) is the problem of learning automatic speech recognition (ASR) systems from *unpaired* speech-only and text-only corpora. Such a system can not only significantly reduce the amount of annotation resources required for training state-of-the-art ASR system, but serve as a bridge between spoken and written language understanding tasks in the low-resource setting. Since its first proposal (Liu et al., 2018), it has seen remarkable progress and the current best system (Baevski et al., 2021) has achieved comparable performance to systems trained with paired data on various languages. However, there are several mysteries surrounding ASR-U, which potentially hinders the future development of such systems. In particular, prior experiments have found that training the current state-of-the-art ASR-U model, wav2vec-U (Baevski et al., 2021), requires careful tuning over the weights of various regularization losses to avoid converging to bad local optima and that even despite extensive regularization weight tuning, wav2vec-U may still fail to converge (Ni et al., 2022). Therefore, it remains a mystery whether or when unpaired speech and text data indeed provide sufficient information for learning an ASR system. Another mystery is whether the success of existing ASR-U models based on generative adversarial net (GAN) (Goodfellow et al., 2014) is sufficiently explained by the GAN objective function per se, or also requires other factors, such as randomness in training, quirks in the data used and careful domain-specific hyper-parameter settings, etc.

In this paper, we provide a theoretical analysis of ASR-U to investigate the mysteries surrounding ASR-U. First, we prove learnability conditions and sample complexity bounds that crucially depend on the eigenvalue spacings of the transition probability matrix of the spoken language. Random matrix theory shows that such learnability conditions are achievable with high probability. Next, we study the gradient flow of GAN-based ASR-U and provide conditions under which the generator minimizing the GAN objective converges to the true generator. Finally, to verify our theory empirically, we perform GAN-based ASR-U experiments on three classes of synthetic languages. Not only do we observe phase transition phenomena predicted by our theory, but we achieve stable training with lower test word error rate by several modifications of the existing state-of-the-art ASR-U system inspired by our theory.

## 2 Problem formulation

**General formulation** The training data comprise a set of sequences of quantized speech vectors, and a set of sequences of phoneme labels. The data are unpaired: there is no label sequence that matches any one of the speech sequences. The data are, however, matched in distribution. Let $P_{X_i}(x)$ and $P_{Y_j}(y)$ be the probability mass functions (pmfs) of the $i^{\text{th}}$ speech vector in a sequence, $x \in \mathbb{X}$, and the

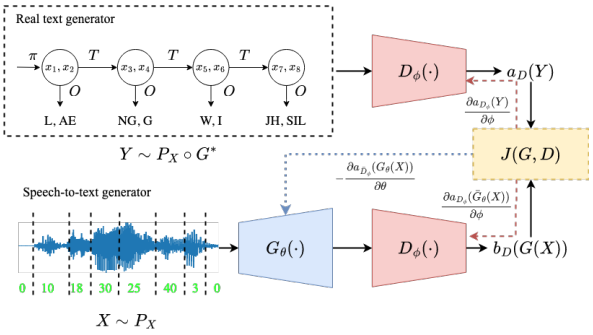



Figure 1: Overview of the ASR-U system for our analysis

$j^{\text{th}}$ phoneme in a sequence, $y \in \mathbb{Y}$, respectively: the requirement that they are *matched in distribution* is the requirement that there exists some generator function $O : (\mathbb{X}, \mathbb{Y}) \to \{0, 1\}$ such that

$$\sum_{x \in \mathbb{X}} P_{X_i}(x) O(x, y) = P_{Y_i}(y) \tag{1}$$

The problem of ASR-U is to find the generator function $O$.

**GAN-based ASR-U** Eq. (1) leverages sequence information to remove ambiguity: $O$ must be an optimal generator not only for the position-independent distributions of $X$ and $Y$, but also for their position-dependent distributions $P_{X_i}, P_{Y_i} \forall i \in \mathbb{N}^0$. In reality we cannot observe every possible sequence of speech vectors, or every possible sequence of phonemes, but instead must estimate $O$ from samples. To address this issue, a GAN can be used to reproduce the empirical distribution of the training dataset with minimum error, subject to the generator's inductive bias, e.g., subject to the constraint that the function $O$ is a matrix of the form $O \in \{0, 1\}^{|\mathbb{X}| \times |\mathbb{Y}|}$, where $|\mathbb{X}|$ and $|\mathbb{Y}|$ are the sizes of the alphabets $\mathbb{X}$ and $\mathbb{Y}$, respectively. As shown in Figure 1, a GAN achieves this goal by computing $O$ as the output of a neural network, $O = G(x, y; \theta)$, and by requiring $G$ to play a zero-sum game with another neural network called the discriminator $D$ with the following general utility function:

$$\min_G \max_D J(G, D) := \mathbb{E}_{Y \sim P_Y}[a(D(Y))] - \mathbb{E}_{X \sim P_X}[b(D(G(X)))]. \tag{2}$$

For the original GAN (Goodfellow et al., 2014), $a(D) = \log(\sigma(D))$ and $b(D) = -\log(1 - \sigma(D))$, where $\sigma$ is the sigmoid function. For the Wasserstein GAN (Arjovsky et al., 2017), $D(Y)$ is a Lipschitz-continuous scalar function, and $a(D) = b(D) = D$. A maximum mean discrepancy (MMD) GAN (Li et al., 2017) minimizes the squred norm of Eq. (2)), where $D(Y)$ is an embedding into a reproducing kernel Hilbert space (RKHS). In this paper we take the RKHS embedding to be the probability mass function of a scalar random variable $D(Y)$, and assume that the discriminator is trained well enough to maintain Eq. (2)) simplified considerably if the RKHS embedding is the probability mass function of a scalar random variable $D(Y)$ non-negative; in this situation, the MMD GAN minimizes Eq. (2) with $a(D) = b(D) = Y$. In practice, Eq. (2) is optimized by alternatively updating the parameters of the discriminator and the generator using gradient descent/ascent:

$$\phi_{i+1} = \phi_i + \eta \nabla_\phi J(G_{\theta_i}, D_{\phi_i}) \tag{3}$$

$$\theta_{i+1} = \theta_i - \nu \nabla_\theta J(G_{\theta_i}, D_{\phi_{i+1}}). \tag{4}$$

**Theoretical questions of ASR-U** The aforementioned formulation of ASR-U is ill-posed. Intuitively, the function $O$ has finite degrees of freedom ($O \in \{0, 1\}^{|\mathbb{X}| \times |\mathbb{Y}|}$), while Eq. (1) must be valid for an infinite number of distributions ($P_{X_i}$ and $P_{Y_i}$ for $i \in \mathbb{N}$), so there is no guarantee that a solution exists. On the other hand, if the sequence is unimportant ($P_{X_i} = P_{X_j} \forall i, j \in \mathbb{N}^0$), then the solution may not be unique. One important question is then: *what are the necessary and sufficient conditions for Eq. (1) to have a unique solution?* Further, it is well-known that gradient-based training of GAN can be unstable and prior works on ASR-U (Yeh et al., 2019; Baevski et al., 2021) have used various regularization losses to stabilize training. Therefore, another question of practical significance is: *what are the necessary and sufficient conditions for the alternate gradient method as described by in Eq. (3)-(4) to converge to the true generator for ASR-U?* In the subsequent sections, we set out to answer these questions.

## 3 Theoretical analysis of ASR-U

### 3.1 Learnability of ASR-U: a sufficient condition

A key assumption of our theory is that the distribution of the speech and the text units can be modeled by a single *hidden Markov model* whose hidden states are $N$-grams of speech units and whose outputs are $N$-grams of text units, as shown in Figure 1.

The parameters of this HMM are its initial probability vector, $\pi$, which specifies the distribution of the first $N$ speech vectors $X_{0:(N-1)} \in \mathbb{X}^N$, its transition probability matrix $A$, which specifies the probability of any given sequence of $N$ speech vectors given the preceding $N$ speech vectors, and its observation probability matrix, which specifies the distribution of one phone symbol given one speech vector:

$$\begin{aligned}\pi &:= P_{X_{0:N-1}} \in \Delta^{|\mathbb{X}|^N} \\ A &:= P_{X_{N:2N-1}|X_{0:N-1}} \in \Delta^{|\mathbb{X}|^N \times |\mathbb{X}|^N} \\ O &:= P_{Y|X} \in \Delta^{|\mathbb{X}| \times |\mathbb{Y}|},\end{aligned}$$

where $\Delta^k$ is the $k$-dimensional probability simplex. The first-order Markov assumption is made plausible by the use of N-gram states, $X_{0:N-1}$, rathern than unigram states; with sufficiently long $N$, natural language may be considered to be approximately first-order Markov. The connection between the $N$-gram states and the unigram observations requires the use of a selector matrix, $E = \mathbf{1}_{|\mathbb{X}|^{N-1}} \otimes I_{|\mathbb{X}|}$, where $\otimes$ denotes the Kronecker product, thus $P_{X_{kN}} = \pi^\top A^k E$, and for multiples of $N$, Eq. (1) can be written $P_{Y_{kN}} = \pi^\top A^k EO$. It turns out that a crucial feature for a spoken language to be learnable in an unsupervised fashion is that it needs to be "complex" enough such that a simple, symmetric and repetitive graph is not sufficient to generate the language. This is captured by the following assumptions on the parameters $A$ and $\pi$.

**Assumption 1.** *There exists an invertible matrix $U \in \mathbb{R}^{|\mathbb{X}|^{N-1} \times |\mathbb{X}|^{N-1}} = [U_1|U_2|\cdots|U_K]$, where the columns of each matrix $U_j = [u_{j1}|\cdots|u_{jN_j}]$ are eigenvectors with the same eigenvalue and a diagonal matrix $\Lambda = \mathrm{blkdiag}(\Lambda_1, \cdots, \Lambda_K)$, where each matrix $\Lambda_k$ is a diagonal matrix with all diagonal elements equal to the same scalar $\lambda_k$, such that $A = U\Lambda U^{-1}$ with $|\mathbb{X}|^N \geq K \geq |\mathbb{X}|$ nonzero eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_K$.*

**Assumption 2.** *For at least $|\mathbb{X}|$ values of $j$, there is at least one $k$ s.t. $\pi^\top u_{jk} \neq 0$.*

With Assumptions 1 and 2, we can consider the following algorithm: First, we construct the following matrices

$$P^X := \begin{bmatrix} P_{X_0}^\top \\ P_{X_N}^\top \\ \vdots \\ P_{X_{(L-1)N}}^\top \end{bmatrix}, P^Y := \begin{bmatrix} P_{Y_0}^\top \\ P_{Y_N}^\top \\ \vdots \\ P_{Y_{(L-1)N}}^\top \end{bmatrix}, \quad (5)$$

Then, $O$ satisfies the following matrix equation

$$P^X O = P^Y. \quad (6)$$

The binary matrix $O$ in Eq. 6) is unique if and only if $P^X$ has full column rank. The following theorem proves that this is indeed the case under our assumptions.

**Theorem 1.** *Under Assumptions 1 and 2, $P^X$ has full column rank and perfect ASR-U is possible. Further, the true phoneme assignment function is $O = P^{X+}P^Y$, where $P^{X+} = (P^{X\top}P^X)^{-1}P^{X\top}$ is the left-inverse of $P^X$.*

Further, if we measure how far the matrix $P^X$ is from being singular by its *smallest* singular value defined as

$$\sigma_{\min}(P^X) := \min_{v \in \mathbb{R}^{|\mathbb{X}|}} \frac{\|P^X v\|_2}{\|v\|_2},$$

we can see that $P^X$ becomes further and further away from being singular as the sequence length $L$ gets larger. An equivalent result for a different purpose has appeared in the Theorem 1 of (Bazán, 2000).

**Lemma 1.** *Under Assumptions 1 and 2 and for simplicity assuming the number of distinct eigenvalues $K = |\mathbb{X}|$ for $T$, then we have*

$$\sigma_{\min}(P^X) \geq \frac{\delta_{\min}^{(|\mathbb{X}|-1)/2|\mathbb{X}|} \sum_{l=0}^{L-|\mathbb{X}|-1} \lambda_{\min}^{2l}(A)}{\kappa(V_{|\mathbb{X}|}(\lambda_{1:|\mathbb{X}|}))} \min_j \|\hat{r}_j\| \quad (7)$$

*where $\delta_{\min} := \min_{i \neq j} |\lambda_i(A) - \lambda_j(A)|$, $\lambda_{\min}(A)$ is the smallest eigenvalue of square matrix $A$, $\kappa(V_{|\mathbb{X}|}(\lambda_{1:|\mathbb{X}|}))$ is the condition number of the square Vandermonde matrix created from eigenvalues $\lambda_1(A), \ldots, \lambda_{|\mathbb{X}|}(A)$, $r_j = \pi^T U_j \Omega_j^\top E$, and $\Omega_j^\top$ is the set of rows of $U^{-1}$ corresponding to eigenvalue $\lambda_j(A)$, after orthogonalizing them from every other block of rows, i.e., $U^{-1} = L[\Omega_1|\cdots|\Omega_K]^T$ such that $L$ is lower-triangular, and the blocks $\Omega_i$ and $\Omega_j$ are orthogonal.*

Next, we will show that Assumption 1 can be easily met using random matrix arguments.

### 3.2 Finite-sample learnability of ASR-U

**Matched setup** Now we show that the requirement for distinct eigenvalues is a mild one as it can easily be satisfied with *random* transition matrices. According to such a result, ASR-U is feasible

with high probability in the (empirically) *matched* setting commonly used in the ASR-U literature, where the *empirical* generated and true distributions can be matched exactly by some generator in the function class (Liu et al., 2018). Our proof relies crucially on the seminal work of (Nguyen et al., 2017) on eigenvalue gaps of symmetric random matrices with independent entries.

In the context of ASR-U, it is of particular interest to study the eigenvalue gaps of a Markov random matrix, which unlike the symmetric case, is asymmetric with correlated entries. Fortunately, by modifying the proof for Theorem 2.6 of (Nguyen et al., 2017), we can show that if the language model belongs to a special but rather broad class of Markov random matrices defined below and the states are *non-overlapping* $N$-gram instead of the more common overlapping ones, it should have at least $|\mathbb{X}|$ distinct eigenvalues with minimum spacing depending on $|\mathbb{X}|$ and the $N$ for the $N$-gram.

**Definition 1.** *(symmetric Markov random matrix) A symmetric Markov random matrix is a matrix of the form $A := D^{-1}W$, where the* adjacency matrix *$W$ is a real, symmetric random matrix with positive entries and bounded variance and $D$ is a diagonal matrix with $d_{ii} = \sum_j W_{ij} > 0$.*

Intuitively, a symmetric Markov random matrix is the transition matrix for a *reversible* Markov chain formed by normalizing edge weights of a weighted, undirected graph.

**Theorem 2.** *(simple spectrum of symmetric Markov random matrix) For any real symmetric Markov random matrix $A_n = D_n^{-1}W_n \in \mathbb{R}^{n\times n}$ with adjacency matrix $W_n = F_n + X_n$, where $F_n$ is a deterministic symmetric matrix with eigenvalues of order $n^\gamma$ and $X_n$ is a symmetric random matrix of zero-mean, unit variance sub-Gaussian random variables, we have for any $C > 0$, there exists $B > 4\gamma' C + 7\gamma' + 1$ such that*

$$\max_{1\le i\le n-1} \Pr[|\lambda_i - \lambda_{i+1}| \le n^{-B}] \le n^{-C},$$

*with probability at least $1 - O(\exp(-\alpha_0 n))$ for some $\alpha_0 > 0$ dependent on $B$ and $\gamma' = \max\{\gamma, 1/2\}$.*

**Corollary 1.** *Suppose the speech feature transition probability is a symmetric Markov random matrix $A := D^{-1}W$ with entries $W_{ij} \sim \mathrm{Uniform}(0, 2\sqrt{3})$ and $D$ is a diagonal matrix with $d_{ii} = \sum_j W_{ij}$, then for any $\epsilon > 0$, there exists $\alpha_0 > 0$ such that with probability at least $1 - O\left(|\mathbb{X}|^{-CN} + \exp\left(-\alpha_0|\mathbb{X}|^N\right)\right)$, the transition probability matrix $A$ has $|\mathbb{X}|^N$ distinct eigenvalues with minimum gap $|\mathbb{X}|^{-BN} > 0$.*

The proof of Theorem 2 and Corollary 1 are presented in detail in the Appendix .

**Unmatched setup** In the finite-sample, unmatched setup, the empirical distribution of the fake text data generated by the GAN does not necessarily match the empirical distribution of the true text data. Assuming the discriminator is perfect in the sense that it maintains Eq. (2) non-negative, and assuming $D(Y)$ is a scalar random variable, then minimizing Eq. (2) is equivalent to minimizing a *divergence measure* $d(\cdot,\cdot)$, between the empirical text distribution, $P^Y$, and the text distribution generated by $O_x(y) = \hat{P}_{Y|X}(y|x)$:

$$\min_{O\in\Delta^{|\mathbb{X}|\times|\mathbb{Y}|}} d^\gamma(P^Y, P^X O), \tag{8}$$

where $\gamma > 0$. For example, for the original GAN, $d(\cdot,\cdot)$ is the Jensen-Shannon distance and for the MMD GAN, $d(\cdot,\cdot)$ is the $L_\gamma$ distance between the expectations $\mathbb{E}[D(Y)]$ under the distributions $P^Y$ and $P^X O$. In both cases, however, Eq. (8) can be minimized using a *decomposable* discriminator defined to be:

$$\mathbb{E}_{P_Y}[a(D(Y))] = \sum_{l=0}^{L-1} \mathbb{E}_{P_{Y_l}}[a(D_l(Y_l))] \tag{9}$$

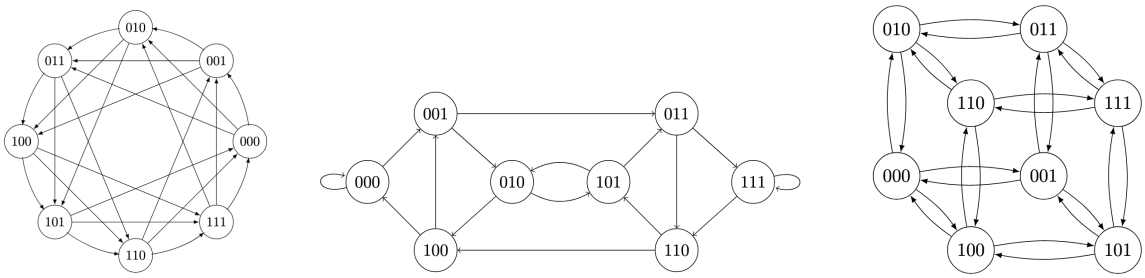
$$\mathbb{E}_{P_X}[b(D(G(X)))] = \sum_{l=0}^{L-1} \mathbb{E}_{P_{X_l}}[b(D_l(G_l(X))], \tag{10}$$

with *components* $D_l : |\mathbb{Y}| \mapsto \mathbb{R}, l = 1, \cdots, L$. Under the assumption that $D$ is decomposable and that the MMD GAN is used, we have the following sample complexity bound on perfect ASR-U.

**Theorem 3.** *The empirical risk minimizer (ERM) of Eq. (8) recovers the true assignment $O$ perfectly from $n^X$ speech frames and $n^Y$ text characters with probability $1 - 2\delta$ if*

$$\sigma_{\min}(P^X) \ge \sqrt{\frac{4L|\mathbb{Y}|(n^X + n^Y) + L|\mathbb{X}|n^X}{n^X n^Y}} + 10\sqrt{\frac{L\log\frac{1}{\delta}}{n^X \wedge n^Y}},$$

*where $n^X \wedge n^Y := \min\{n^X, n^Y\}$.*




(a) Circulant graph (b) De Bruijn graph (c) Hypercube

Figure 2: Various types of Markov transition graphs

### 3.3 Training dynamic of GAN-based ASR-U

So far, we have assumed the GAN training is able to find the optimal parameters for the discriminator and the generator. However, there is no guarantee that this is indeed the case with gradient updates such as Eq. (3). To analyze the behaviour of the GAN training dynamic for ASR-U, we follow prior works on neural tangent kernel (NTK) (Jacot et al., 2018) to focus on the *infinite-width, continuous-time* regime, or NTK regime, where the generator and the discriminator are assumed to be neural networks with an infinite number of hidden neurons trained with gradient descent at an infinitely small learning rate. Though highly idealized, studying such a regime is practically useful as results from this regime can often be converted to finite-width, discrete-time settings (See, e.g., (Du et al., 2019)).

For simplicity, denote $f_\tau := D_{\phi_\tau}$ and $g_t := G_{\theta_t}$ and define $\mathcal{L}_t(f) := J(g_t, f)$, then in the NTK regime, between each generator step, the training dynamic of the discriminator can be described by the following partial differential equation (PDE):

$$\partial_\tau \phi_\tau = \nabla_{\phi_\tau} \mathcal{L}_t(f_\tau). \tag{11}$$

Let $f^*_{P_t} = \lim_{\tau\to\infty} f_\tau$ be the limit of Eq. (11). If the limit exists and is unique, the generator loss is well-defined as $\mathcal{C}_t(g_t) := J(g_t, f^*_{P_t})$. Note that the output of the ASR-U generator is discrete, which is not a differentiable function per se, but we can instead directly parameterize the *generated text distribution* as $P_{g_t} := P_X \circ O_t$ for some softmax posterior distribution $O_t$:

$$O_{t,x}(y) := \prod_{l=1}^{L} \frac{\exp(h_{\theta,y_l}(x_l))}{\sum_{y'_l} \exp(h_{\theta,y'_l}(x_l))}, \tag{12}$$

where $h_\theta$ is a neural network, and is assumed to be one layer in our analysis, though it can be extended to multiple layers with slight modifications using techniques similar to those in (Du et al., 2019). Using such a generator, the generator dynamic can be then described by the following PDE:

$$\partial_t \theta_t = \sum_{y\in\mathbb{Y}^L} b(f^*_{g_t}(y)) \nabla_{\theta_t} P_{g_t}(y), \tag{13}$$

where the right-hand side is the term in the gradient of $\mathcal{C}_t$ with respect to $\theta_t$ ignoring the dependency of the discriminator $f^*_{g_t}$. Define the NTKs of the discriminator and the generator (distribution) as

$$K_{f_\tau}(y, y') = \mathbb{E}_{\phi_0\sim\mathcal{W}} \left[ \frac{\partial f_\tau(y)}{\partial \phi_\tau}^\top \frac{\partial f_\tau(y')}{\partial \phi_\tau} \right] \tag{14}$$

$$K_{g_t}(y, y') = \mathbb{E}_{\theta_0\sim\mathcal{W}} \left[ \frac{\partial P_{g_t}(y)}{\partial \theta_t}^\top \frac{\partial P_{g_t}(y')}{\partial \theta_t} \right], \tag{15}$$

where $\mathcal{W}$ is the initialization distribution (usually Gaussian). Note that the NTKs are $|\mathbb{Y}|^L \times |\mathbb{Y}|^L$ matrices for ASR-U due to the discrete nature of the generator. A key result in (Jacot et al., 2018) states that as the widths of the hidden layers of the discriminator and generator go to infinity, $K_{f_\tau} \to K_D, K_{g_t} \to K_G$ stay constant during gradient descent/ascent and we have

$$\begin{aligned} \partial_\tau f_\tau = K_D \big(&\mathrm{diag}(P_Y) \nabla_{f_\tau} a \\ &- \mathrm{diag}(P_{g_t}) \nabla_{f_\tau} b\big), \end{aligned} \tag{16}$$

$$\partial_t P_{g_t} = K_G \mathbf{b}_{f_{g_t}}, \tag{17}$$

where $\nabla_f \{a, b\} = \left[\frac{\partial\{a,b\}(f(y))}{\partial f(y)}\right]_{y\in\mathbb{Y}^L}$ and $\mathbf{b}_f = (b_f(y))_{y\in\mathbb{Y}^L}$. However, Eq. (16)-(17) is in general highly nonlinear and it remains an open problem as to their convergence properties. Instead, we focus on the case when the discriminator $f_t$ is decomposable with components $f_{t,l}, l = 1, \cdots, L$, and simplify Eq. (16) and Eq. (17) into PDEs involving

only samples at a particular time step:

$$\partial_\tau f_{\tau,l} = K_{D,l}\left(\mathrm{diag}(P_l^Y)\nabla_{f_{\tau,l}}\mathbf{a}_{f_{\tau,l}} - \mathrm{diag}(P_l^{g_t})\nabla_{f_{\tau,l}}\mathbf{b}_{f_{\tau,l}}\right), \quad (18)$$

$$\partial_t O_{t,x}^\top = \sum_{l=1}^{L} P_l^X(x) K_{O_{t,x}}\mathbf{b}_{f_{g_t,l}}, \quad (19)$$

for all $l = 1, \cdots, L$, $x \in \mathbb{X}$ in terms of the *stepwise* NTKs defined as:

$$K_{D,l}(y, y') := \mathbb{E}_{\phi_0\sim\mathcal{W}}\left[\frac{\partial f_\tau(y)}{\partial \phi_\tau}^\top \frac{\partial f_\tau(y')}{\partial \phi_\tau}\right]$$

$$K_{O_{t,x}}(y, y') := \mathbb{E}_{\theta_0\sim\mathcal{W}}\left[\frac{\partial O_{t,x}(y)}{\partial \theta_\tau}^\top \frac{\partial O_{t,x}(y')}{\partial \theta_\tau}\right].$$

We focus on the special case that $f_{\tau,l}$ is parameterized by a two-layer neural network with ReLU activation, though the framework can be extended to network of arbitrary depths:

$$f_{\tau,l}(y) = \lim_{m\to\infty} \frac{1}{\sqrt{m}} \sum_{r=1}^{m} v_r^l \max\{W_{ry}^l, 0\}. \quad (20)$$

In this case, under mild regularity conditions, we can show that the generator trained with the alternate gradient method minimizes Eq. (8), which under the same condition as in Section 3.2, implies ASR-U is feasible.

**Theorem 4.** *Suppose the following assumptions hold:*

1. *The discriminator is decomposable and parameterized by Eq. (20), whose parameters are all initialized by standard Gaussian variables;*
2. *The generator is linear before the softmax layer;*
3. *The GAN objective is MMD;*
4. *The linear equation $P^X O = P^Y$ has at least one solution.*

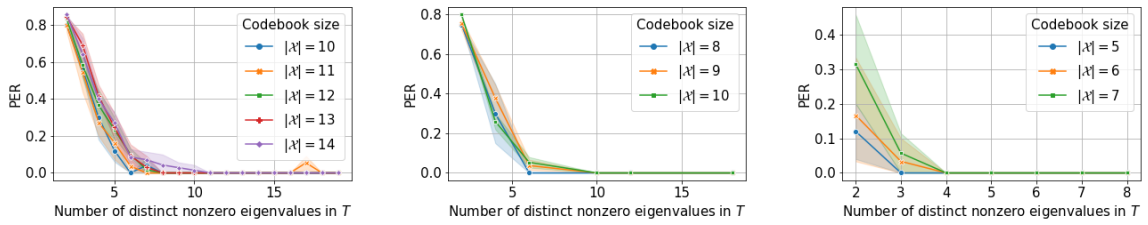
*Then we have for any solution $O_t$ of Eq. (19),* $\lim_{t\to\infty} P^X O_t = P^Y$.

## 4 Experiments

**Synthetic language dataset** To allow easy control of the eigenvalue spacings of the transition matrix $T$ and thus observe the phase transition phenomena predicted by our theory, we design six synthetic languages with HMM language models as follows. First, we create the HMM transition graph by treating non-overlapping *bigrams* as hidden states of the HMM. The hidden state of the HMM will henceforth be referred to as the "speech unit", while the observation emitted by the HMM will be referred to as the "text unit". For the asymptotic ASR-U, we control the number of eigenvalues of the Markov transition graph by varying the number of disjoint, identical subgraphs. The number of distinct eigenvalues of the whole graph will then be equal to the number of eigenvalues of each subgraph. For the finite sample setting, we instead select only Hamiltonian graphs and either gradually decrease the degrees of the original graph to its Hamiltonian cycle or interpolate between the graph adjacency matrix and that of its Hamiltonian cycle. Thus, we can increase $\sigma_{\min}(P^X)$ by increasing $w$. For both the subgraph in the former case and the Hamiltonian graph in the latter, we experiments with circulant, de Bruijn graphs (de Bruijn, 1946) and hypercubes, as illustrated in Figure 2. Next, we randomly permute the hidden state symbols to form the true generator mapping from the speech units to text units. To create matched speech-text data, we simply sample matched speech and text unit sequences using a single HMM. For unmatched datasets, we sample the speech and text data independently with two HMMs with the same parameters. Please refer to Appendix B for more details.

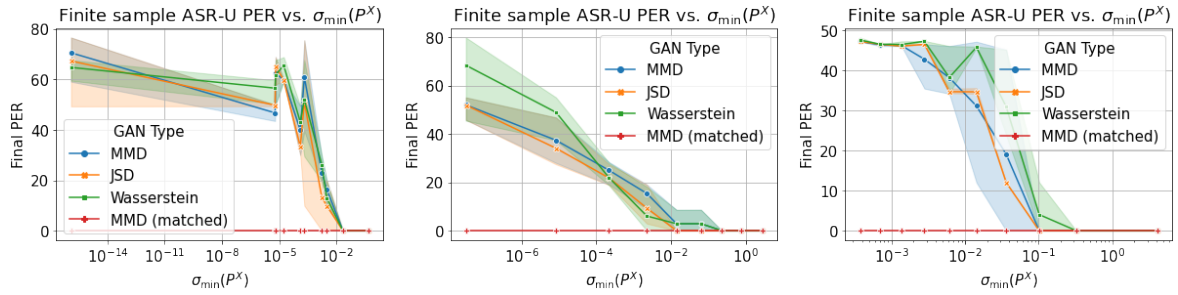
**Model architecture** For finite-sample ASR-U, we use wav2vec-U (Baevski et al., 2021) with several modifications. In particular, we experiment with various training objectives other than the Jensen-Shannon (JS) GAN used in the original wav2vec-U, including the Wasserstein GAN (Liu et al., 2018) and the MMD GAN. All additional regularization losses are *disabled*. Moreover, we experimentally manipulate two hyperparameters: (1) the averaging strategy used by the generator, and (2) whether to *reset* the discriminator weights to zero at the beginning of each discriminator training loop. More details can be found in Appendix B.

**Phase transition of PER vs. eigenvalue gaps: asymptotic case** The phoneme error rate (PER) as a function of the number of eigenvalues of $A$ for the asymptotic ASR-U on the synthetic datasets are shown in Figure 3. For all three graphs, we observe clear phase transitions as the number of eigenvalues exceeds the number of speech units, and an




(a) Circulant graph (b) De Bruijn graph (c) Hypercube

Figure 3: Asymptotic ASR-U PER vs number of distinct nonzero eigenvalues for various Markov transition graphs




(a) Circulant graph (b) De Bruijn graph (c) Hypercube

Figure 4: Finite-sample ASR-U PER vs $\sigma_{\min}(P^X)$ for various Markov transition graphs

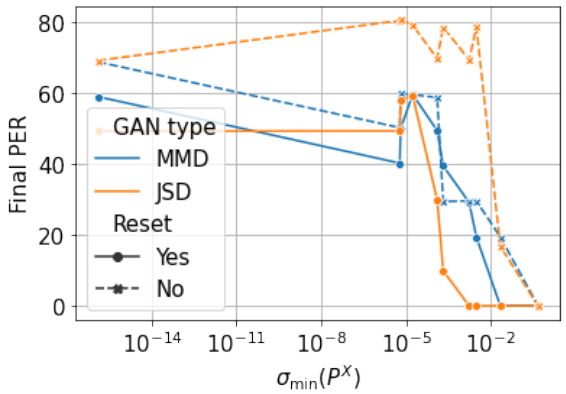



Figure 5: Effect of discriminator resetting at every update

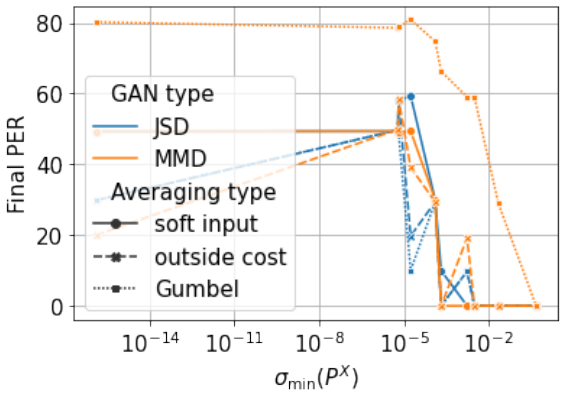



Figure 6: Effect of different type of averaging for the generator

increase of the number of distinct, nonzero eigenvalues required for perfect ASR-U as the number of speech units increases.

**Phase transition of PER vs. eigenvalue gaps: finite-sample case** The PER as a function of the least singular value $\sigma_{\min}(P^X)$ for the finite-sample ASR-U on the synthetic datasets are shown in Figure 4. As we can see, the ASR-U exhibit the phase transition phenomena in all three graphs, albeit with differences in the critical point and their rate of approaching the perfect ASR-U regime. While the PER generally decreases as $\sigma_{\min}(P^X)$ gets larger, we found a dip in PER in the circulant graph case as $\sigma_{\min}(P^X)$ moves from $10^{-31}$ to $10^{-15}$. Though unexpected, this observation is not contradictory to our theory since our theory does not make explicit predictions about the rate of phase transition for ASR-U. Across different GAN models, we found that JSD generally approaches perfect ASR-U at a faster rate than MMD in all three graphs, suggesting the use of nonlinear dynamic may be beneficial. Nevertheless, the overall trends for different GANs remain in large part homogeneous. Between Wasserstein and MMD, we observe very little difference in performance, suggesting the reg-

ularization effect of NTK is sufficient to control the Lipschitz coefficient of the network. Finally, for the MMD GAN in the matched setting, we found the network is able to achieve perfect ASR-U regardless of the spectral properties of the Markov transition graphs, which confirms our theory that a symmetric Markov random matrix tends to have simple eigenvalue spectrum suitable for ASR-U.

**Effect of discriminator reset** As pointed out by (Franceschi et al., 2021), a discriminator may suffer from residual noise from previous updates and fail to approximate the target divergence measure. We analyze such effects for MMD and JSD as shown in Figure 5. We observed consistent trends that models whose weights are reset to the initial weights every discriminator loop outperform those without resetting. The effect is more pronounced for JSD GAN than MMD GAN and for smaller $\sigma_{\min}(P^X)$.

**Effect of generator averaging strategy** The original wav2vec-U (Baevski et al., 2021) directly feeds the text posterior probabilities $O$ into the discriminator, which we refer to as the *"soft input"* approach. Alternatively, we can instead calculate a weighted average of the gradient form over the samples $y \in \mathbb{Y}^L$ as in Eq. (13), which we refer to as the *"outside cost"* approach. The comparison between the two approaches are shown in Figure 6. We observed mixed results: for MMD GANs, the soft-input approach outperforms the outside-cost approach and performs best among the models in the high-$\sigma_{\min}(P^X)$ setting; for JSD GANs, we found that the outside-cost approach performs slightly better than the soft-input approach. Such inconsistencies may be another consequence of the regularization effect predicted by the GANTK. We leave the theoretical explanation as future work.

## 5 Related works

The problem of ASR-U was first proposed by (Liu et al., 2018). In the same paper, they develop a speech-to-text generator with ground-truth phoneme boundaries and quantized speech features as inputs to match the real text distribution by optimizing the GAN objective. (Chen et al., 2019) later replaced the ground truth boundaries with unsupervised ones refined iteratively by an HMM, which also incorporates language model information into the system. (Yeh et al., 2019) explored the cross entropy loss for matching the generated and real text distribution, but it is prone to mode collapse and needs the help of additional regularization losses such as smoothness weight. More recently, (Baevski et al., 2021; Liu et al., 2022) proposed another GAN-based model using continuous features from the last hidden layer of the wav2vec 2.0 (Baevski et al., 2020) model and additional regularization losses to stabilize training. Their approach achieves ASR error rates comparable to the supervised system on multiple languages, making it the current state-of-the-art system. To better understand the learning behavior of ASR-U systems, (Lin et al., 2022) analyze the robustness of wav2vec-U against empirical distribution mismatch between the speech and text, and found that $N$-gram language model is predictive of the success of ASR-U.

Our analysis on the sufficient condition of ASR-U is based on previous works on the asymptotic behaviour of GAN objective functions (Goodfellow et al., 2014; Arjovsky et al., 2017). Our finite-sample analysis takes inspiration from later works extending the asymptotic analysis to the finite-sample regimes (Arora et al., 2017; Bai et al., 2019). Such frameworks, however, do not account for the alternate gradient optimization method of GAN and inevitably lead to various inconsistencies between the theory and empirical observations of GAN training (Franceschi et al., 2021). Building upon prior works (Mescheder et al., 2017, 2018; Domingo-Enrich et al., 2020; Mroueh and Nguyen, 2021; Balaji et al., 2021), (Franceschi et al., 2021) proposed a unified framework called GANTK based on NTK (Jacot et al., 2018) to describe the training dynamic of *any* GAN objectives and network architectures. Our analysis on the training dynamic of ASR-U adopts and extends the GANTK framework to handle *discrete, sequential* data such as natural languages.

## 6 Conclusion

In this paper, we develop a theoretical framework to study the fundamental limits of ASR-U as well as the convergence properties of GAN-based ASR-U algorithms. In doing so, our theory sheds light on the underlying causes of training instability for such algorithms, as well as several new directions for more reliable ASR-U training.

## 7 Limitations

Our theory currently assumes that input speech features are quantized into discrete units, as in (Chen

et al., 2019). However, more recent works (Baevski et al., 2021; Liu et al., 2022) have shown that continuous features, with the help of additional regularization losses, can achieve almost perfect ASR-U. Such phenomena is beyond explanations based on our current theory and require generalizing our theory to continuous speech features. Further, our model assumes that sufficiently reliable phoneme boundaries are fed to the ASR-U system, and kept fixed during training. It will be interesting to extend our framework to systems with trainable phoneme boundaries, such as the wav2vec-U systems, to better understand its effect on training stability.

## References


G. Anderson, A. Guionnet, and O. Zeitouni. 2009. *An introduction to random matrices*. Cambridge University Press.

Martin Arjovsky, Soumith Chintala, and Leon Bottou. 2017. Wasserstein GAN. In *International Conference on Machine Learning*.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. 2017. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning*.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. In *Neural Information Processing System*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Neural Information Processing System*.

Yu Bai, Tengyu Ma, and Andrej Risteski. 2019. Approximability of discriminators implies diversity in GANs. In *International Conference on Learning Representations*.

Yogesh Balaji, Mohammadmahdi Sajedi, Neha Mukund Kalibhat, Mucong Ding, Dominik Stöger, Mahdi Soltanolkotabi, and Soheil Feizi. 2021. Understanding overparameterization in generative adversarial networks. In *International Conference on Learning Representations*.

Fermán S. V. Bazán. 2000. Conditioning of rectangular vandermonde matrices with nodes in the unit disk. *SIAM Journal on Matrix Analysis and Applications*, 21(2):679–693.

Nicolaas Govert de Bruijn. 1946. A combinatorial problem. *Indagationes Mathematicae*, page 758–764.

Kuan-Yu Chen, Che-Ping Tsai, Da-Rong Liu, Hung-Yi Lee, and Lin shan Lee. 2019. Completely unsupervised speech recognition by a generative adversarial network harmonized with iteratively refined hidden markov models. In *Interspeech*.

Charles Delorme and Jean Pierre Tillich. 1998. The spectrum of de bruijn and kautz graphs. *European Journal Combinatorics*, pages 307–319.

Carles Domingo-Enrich, Samy Jelassi, Arthur Mensch, Grant M. Rotskoff, and Joan Bruna. 2020. A mean-field analysis of two-player zero-sum games. In *Neural Information Processing System*.

Simon S. Du, Xiyu Zhai, Barnabás Poczós, and Aarti Singh. 2019. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*.

P. Erdös. 1945. On a lemma of Littlewood and Offord. *Bulletin of the American Mathematical Society*, 51:898–902.

Jean-Yves Franceschi, Emmanuel de Bézenac, Ibrahim Ayed, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. 2021. A neural tangent kernel perspective of GANs. In *International Conference on Machine Learning*.

Bolin Gao and Lacra Pavel. 2017. On the properties of the softmax function with application in game theory and reinforcement learning. In *ArKiv*.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Neural Information Processing System*.

Arthur Jacot, Franck Gabriel, and Clément Hongler. 2018. Neural tangent kernel: Convergence and generalization in neural networks. In *Neural Information Processing System*.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. 2017. Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30.

Guan-Ting Lin, Chan-Jan Hsu, Da-Rong Liu, Hung-Yi Lee, and Yu Tsao. 2022. Analyzing the robustness of unsupervised speech recognition. In *ICASSP*.

Alexander H. Liu, Wei-Ning Hsu, Michael Auli, and Alexei Baevski. 2022. Towards end-to-end unsupervised speech recognition. In *ArKiv*.

Da-Rong Liu, Kuan-Yu Chen, Hung-Yi Lee, and Lin shan Lee. 2018. Completely unsupervised phoneme recognition by adversarially learning mapping relationships from audio embeddings. In *Interspeech*.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. 2018. Which training methods for GANs do actually converge? In *International Conference on Machine Learning*.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. 2017. The numerics of GANs. In *Neural Information Processing System*.

Youssef Mroueh and Truyen Nguyen. 2021. On the convergence of gradient descent in GANs: Mmd GAN as a gradient flow. In *International Conference on Artificial Intelligence and Statistics*.

Hoi Nguyen, Terence Tao, and Van Vu. 2017. Random matrices: Tail bounds for gaps between eigenvalues. *Probability Theory and Related Fields*, page 777–816.

Hoi Nguyen and Van Vu. 2011. Optimal Littlewood-Offord theorems. *Advances in Mathematics*, 226(6):5298–5319.

Junrui Ni, Liming Wang, Heting Gao, Kaizhi Qian, Yang Zhang, Shiyu Chang, and Mark Hasegawa-Johnson. 2022. Unsupervised text-to-speech synthesis by unsupervised automatic speech recognition. In *Interspeech*.

Mark Rudelson and Roman Vershynin. 2008. The Littlewood-Offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633.

Terence Tao and Van Vu. 2009. Inverse littlewood–offord theorems and the condition number of random matrices. *Annual of Mathematics*, 169(2):595–632.

Chih-Kuan Yeh, Jianshu Chen, Chengzhu Yu, and Dong Yu. 2019. Unsupervised speech recognition via segmental empirical output distribution matching. In *International Conference on Learning Representations*.

# A Proofs of theoretical results

## A.1 Learnability of ASR-U: a sufficient condition

*Proof.* (Theorem 1)

For simplicity, we assume that the eigenvalues of $A$ are real though a similar argument applies to complex eigenvalues as well. By Assumptions 1 and 2, it can be verified that

$$\begin{aligned} P_{X_{kN}} &= \pi^\top A^k E \\ &= \pi^\top U \Lambda^k U^{-1} E, \end{aligned}$$

where $E = \mathbf{1}_{|\mathbb{X}|^{N-2}} \otimes I_{|\mathbb{X}|}$, where $\otimes$ denotes the Kronecker product. Define $c_{jk} = \pi^\top u_{jk}$. Define $r_{jk}^\top$ to be the $k^{\text{th}}$ row of the $j^{\text{th}}$ block of the matrix $U^{-1}E$, i.e., $UU^{-1}E = \sum_{j=1}^{K}\sum_{k=1}^{N_j} u_{jk} r_{jk}^\top$. Define the matrix $R_K$ as $R_K = [r_1, \cdots, r_K]$, where $r_j = \sum_{k=1}^{N_j} c_{jk} r_{jk}$. Then we have:

$$\begin{aligned} P_{X_{kN}}^\top &= \sum_{j=1}^{K} \lambda_j^k r_j^\top \\ P^X &= V_L(\lambda_{1:K})^\top R_K^\top, \end{aligned}$$

where $V_L(\lambda_{1:K})$ is the Vandermonde matrix formed by nonzero eigenvalues $\lambda_1, \cdots, \lambda_K$ and with $L$ columns, $K \geq |\mathbb{X}|$ by Assumption 1. $R_K$ has full column rank of $K \geq |X|$ by Assumption 2, therefore it is possible to write $R_K = \hat{R}_K L$, where $\hat{R}_K = \hat{r}_1, \ldots, \hat{r}_K]$ is a matrix with orthogonal columns, and $L$ is lower-triangular. As a result, we have $P^X$ is full rank iff $V_L(\lambda_{1:K})$ has full row rank of at least $|\mathbb{X}|$, which holds by Assumption 1. □

*Proof.* (Lemma 1)

Use the Rayleigh-characterization of eigenvalues of the matrix $P^{X\top}P^X$, we have

$$\begin{aligned} &\sigma_{\min}(P^X) \\ &= \sqrt{\lambda_{\min}(P^{X\top}P^X)} \\ &= \sqrt{\min_{\|w\|=1} w^\top P^{X\top} P^X w} \\ &= \sqrt{\min_{\|w\|=1} w^\top R_K V_L V_L^\top R_K^\top w} \\ &\geq \sqrt{\sum_{l=0}^{L-|\mathbb{X}|-1} \lambda_{\min}^{2l} \min_{\|w\|=1} w^\top R_K V_{|\mathbb{X}|} V_{|\mathbb{X}|}^\top R_K^\top w} \\ &= \sigma_{\min}(P^X_{1:|\mathbb{X}|}) \sqrt{\sum_{l=0}^{L-|\mathbb{X}|-1} \lambda_{\min}^{2l}}, \end{aligned}$$

where $\lambda_{\min}$ is the eigenvalue of $A$ with minimum absolute value, and $P^X_{1:|\mathbb{X}|}$ is the first $|\mathbb{X}|$ rows of $P^X$. Therefore, to lower bound $\sigma_{\min}(P^X)$, it suffices to lower bound $\sigma_{\min}(P^X_{1:|\mathbb{X}|})$. But note that

$$\begin{aligned} &\sigma_{\min}(P^X_{1:|\mathbb{X}|}) \\ =& \min_{\|w\|=1} \|V_{|\mathbb{X}|}^T R_K^T w\| \\ \geq& \sigma_{\min}(V_{|\mathbb{X}|}^T) \min_{\|w\|=1} \|R_K^T w\| \\ \geq& \frac{\sigma_{\max}(V_{|\mathbb{X}|})}{\kappa(V_{|\mathbb{X}|})} \min_j \|\hat{r}_j\| \\ \geq& \frac{|\det(V_{|\mathbb{X}|})|^{1/|\mathbb{X}|}}{\kappa(V_{|\mathbb{X}|})} \min_j \|\hat{r}_j\| \\ =& \frac{|\prod_{1\leq i<j\leq|\mathbb{X}|} |\lambda_i - \lambda_j|^{1/|\mathbb{X}|}}{\kappa(V_{|\mathbb{X}|})} \min_j \|\hat{r}_j\| \\ \geq& \frac{\delta_{\min}^{(|\mathbb{X}|-1)/2|\mathbb{X}|}}{\kappa(V_{|\mathbb{X}|})} \min_j \|\hat{r}_j\| \end{aligned}$$

where the last equality uses the closed-form formula of the determinant of a square Vandermonde matrix, and where the behaviour of $\kappa(V_{|\mathbb{X}|})$, the condition number of the Vandermonde matrix, has been studied in depth in (Bazán, 2000). □

### A.2 Finite-sample learnability of ASR-U: matched setup

**Theory of small ball probability** The proof of Theorem 2 makes extensive use of the theory of small ball probability. Therefore, we briefly provide some background on the subject. First, we define the *small ball probability* of a vector $x$ as follows.

**Definition 2.** *(Small ball probability) Given a fixed vector $x = (x_1, \cdots, x_n)$, and i.i.d random variables $\xi = (\xi_i, \cdots, \xi_n)$, the small ball probability is defined as*

$$\rho_\delta(x) := \sup_{a \in \mathbb{R}} \Pr[|\xi^\top x - a| \leq \delta].$$

Intuitively, small ball probability is the amount of "*additive* structure" in $x$: for example, if the coordinates of $x$ are integer multiples of each other and $\xi_i$'s are symmetric Bernoulli variables, the product $\xi^\top x$ tends to have small magnitude as terms cancel out each other very often. Since sparser vectors tend to have less additive structure, small ball probability can also be used to measure how *sparse* the weights of $x$ are. Another way to look at this is that, if the L2 norm of $x$ is fixed and most of the weight of $x$ is gathered in a few coordinates, the product $\xi^\top x$ has higher variance and is thus less likely to settle in any fixed-length intervals. This is quantitatively captured by the celebrated Offord-Littlewood-Erdös (OLE) anti-concentration inequality (and its inverse) for general subgaussian random variables:

**Lemma 2.** *(Erdös, 1945; Rudelson and Vershynin, 2008; Tao and Vu, 2009) Let $\epsilon > 0$ be fixed, let $\delta > 0$, and let $v \in \mathbb{R}^m$ be a unit vector with*

$$\rho_\delta(v) \geq m^{-\frac{1}{2}+\epsilon}.$$

*Then all but at most $\epsilon m$ of the coefficients of $v$ have magnitude at most $\delta$.*

Note that here we use a slight generalization of the notion of sparsity called *compressibility* defined as follows.

**Definition 3.** *($(\alpha, \delta)$-compressible) A vector $v \in \mathbb{R}^n$ is $(\alpha, \delta)$-compressible if at most $\lfloor \alpha n \rfloor$ of its coefficients have magnitude above $\delta$.*

Note that a sparse vector with a support of size at most $\lfloor \alpha n \rfloor$ is $(\alpha, 0)$-compressible.

A more generally applicable anti-concentration inequality requires the following definition of generalized arithmetic progression, which is used to quantify the amount of additive structure of a vector.

**Definition 4.** *(Generalized arithmetic progression) A generalized arithmetic progression (GAP) is a set of the form*

$$Q = \{a^\top w : a \in \mathbb{Z}^r, |a_i| \leq N_i, 1 \leq i \leq r\},$$

*where $r \geq 0$ is called the* rank *of the GAP and $w_1, \cdots, w_r \in \mathbb{R}$ are called* generators of the GAP. *Further, the quantity*

$$\mathrm{vol}(Q) := \prod_{i=1}^{r} (2N_i + 1)$$

*is called the* volume *of the GAP.*

**Lemma 3.** *(Continuous inverse Littlewood-Offord theorem, Theorem 2.9 of (Nguyen and Vu, 2011)) Let $\epsilon > 0$ be fixed, let $\delta > 0$ and let $v \in \mathbb{R}^n$ be a unit vector whose small ball probability $\rho := \rho_\delta(v)$ obeys the lower bound*

$$\rho \gg n^{-O(1)}.$$

*Then there exists a generalized arithmetic progression $Q$ of volume*

$$vol(Q) \leq \max\left(O\left(\frac{1}{\rho\sqrt{\alpha n}}\right), 1\right)$$

*such that all but at most $\alpha n$ of the coefficients $v_1, \cdots, v_n$ of $v$ lie within $\delta$ of $Q$. Furthermore, if $r$ denotes the rank of $Q$, then $r = O(1)$ and all the generators $w_1, \cdots, w_r$ of $Q$ have magnitude $O(1)$.*

While applicable for any $\rho \gg n^{-\epsilon}$ rather than only those with $\rho_\delta(v) \geq n^{-1/2+\epsilon}$ as required by Lemma 2, Lemma 3 is *weaker* than Lemma 2 in the sense that rather than showing that the vector is compressible with high probability and thus covered by the set of compressible vectors, it proves that the vector is covered by a small set with high probability.

A related notion that is often more convenient for our analysis is the *segmental* small ball probability, which is simply small ball probability computed on a segment of the vector:

$$\rho_{\delta,\alpha}(x) = \inf_{I \subseteq \{1,\cdots,n\}:|I|=\lfloor \alpha n \rfloor} \rho_\delta(x_I),$$

From the definition, it is not hard to see that $\rho_{\delta,\alpha}(x) \geq \rho_\delta(x)$.

**Eigen-gaps of symmetric Markov random matrix** Armed with tools from the theory of small ball probability, we will establish guarantees of eigenvalue gaps for a symmetric Markov random matrix. First, we shall show that Theorem 2 implies Corollary 1.

*Proof.* (Proof of Corollary 1) Using Theorem 2 and union bound, the probability that a symmetric Markov random matrix has at least $|\mathbb{X}|$ distinct eigenvalues can be bounded as

$$\begin{aligned}&\Pr\left[\min_{1\le i\le|\mathbb{X}|}|\lambda_i-\lambda_{i+1}|\le|\mathbb{X}|^{-BN}\right]\le\\&|\mathbb{X}|\max_i\Pr\left[|\lambda_i-\lambda_{i+1}|\le|\mathbb{X}|^{-BN}\right]\\&=O(|\mathbb{X}|^{-CN}),\end{aligned}$$

with probability at least $1-O(\exp(-\alpha_0|\mathbb{X}|^N))$. □

It turns out that a symmetric Markov random matrix enjoys various properties analogous to a symmetric matrix. First, we can show that its eigenvalues are real. This can be proved by noting that for a symmetric Markov random matrix $A_n := D_n^{-1}W_n$ and for any of its eigenvalues $\lambda$ with eigenvector $v$,

$$\begin{aligned}&D_n^{-1}W_nv=\lambda v\\\Longleftrightarrow&D_n^{-1/2}W_nD_n^{-1/2}(D_n^{1/2}v)=\lambda D_n^{1/2}v,\end{aligned}\quad(21)$$

which implies $A_n$ has the same spectrum as $D_n^{-1/2}W_nD_n^{-1/2}$, which is symmetric and thus has a real spectrum. Further, we can prove a variant of Cauchy's interlace theorem for symmetric Markov random matrix.

**Lemma 4.** *Suppose $A_n = D_n^{-1}W_n \in \mathbb{R}^{n\times n}$ is a symmetric Markov random matrix with adjacency matrix $W_n$ and eigenvalues $\lambda_1 \ge \cdots \ge \lambda_n$ and $A_m = D_m^{-1}W_m$ with adjacency matrix $W_{m-1}$ and eigenvalues $\nu_1 \ge \cdots \ge \nu_m, m < n$ is formed by successively deleting $i$-rows and $i$-columns, then*

$$\lambda_i \le \nu_i \le \lambda_{i+n-m}.$$

*Proof.* Using the previous observation in Eq. 21, we can apply the standard Cauchy's interlacing theorem on $A'_n := D_n^{-1/2}W_nD_n^{-1/2}$ and $A'_m := D_m^{-1/2}W_mD_m^{-1/2}$, then we have

$$\begin{aligned}\lambda_i(A_n)=\lambda_i(A'_n)\le\lambda_i(A'_m)=\lambda_i(A_m)\\\le\lambda_{i+n-m}(A'_n)=\lambda_{i+n-m}(A_n).\end{aligned}$$

□

Next, we can show that the eigenvalues of a symmetric Markov random matrix and its adjacency matrix are simultaneously distributed within the bounded intervals $[-10n^{\gamma-1}, 10n^{\gamma-1}]$ and $[-10n^{\gamma}, 10n^{\gamma}]$ with high probability. For this and subsequent proofs, we will assume $\gamma' = \gamma > 1/2$.

**Lemma 5.** *Let $A_n = D_n^{-1}W_n$ be a symmetric Markov random matrix with adjacency matrix $W_n$ and properties defined in Theorem 2, then we have with probability at least $1 - O(\exp(-\alpha_0 n))$,*

$$\begin{aligned}\lambda_i(A_n)\in[-10n^{\gamma-1},10n^{\gamma-1}],\\\lambda_i(W_n)\in[-10n^{\gamma},10n^{\gamma}],\end{aligned}$$

*for any $1 \le i \le n$ and some $\alpha_0 > 0$.*

*Proof.* First, by definition, we can let $W_n = F_n + X_n$, where $F_n$ is a deterministic matrix with eigenvalues of order $n^\gamma$ and $X_n$ is a symmetric matrix whose elements are independent zero-mean unit-variance subgaussian random variables. Using standard results from random matrix theory (Anderson et al., 2009), we have

$$\{\lambda_1(X_n),\cdots,\lambda_n(X_n)\}\subset[-10n^{\gamma-1},10n^{\gamma-1}],$$

with probability at least $1-O(\exp(-\alpha_0 n))$. Therefore, Weyl's matrix perturbation inequality then ensures that

$$\{\lambda_1(W_n),\cdots,\lambda_n(W_n)\}\in[-10n^{\gamma},10n^{\gamma}],$$

with probability at least $1 - O(\exp(-\alpha_1 n))$. Suppose this event occurs and use Lemma 4 and the variational characterization of eigenvalues, we have

$$\begin{aligned}\lambda_i(A_n)&=\min_{V_{i-1}}\max_{\substack{v\in V_{i-1}^\perp\\\|v\|_2=1}}v^\top D_n^{-1/2}W_nD_n^{-1/2}v\\&=\min_{V_{i-1}}\max_{\substack{v\in V_{i-1}^\perp\\v^\top Dv=1}}v^\top W_nv\\&=\min_{V_{i-1}}\max_{v\in V_{i-1}^\perp}\frac{v^\top W_nv}{v^\top D_nv},\end{aligned}$$

where $V_{i-1}$ is a subspace of dimension $i-1$. Combining the two results, we have with probability at least $1 - O(\exp(-\alpha_1 n))$,

$$\begin{aligned}\max_{v\in V_{i-1}^\perp}\left|\frac{v^\top W_nv}{v^\top D_nv}\right|&\le\frac{\max_{v:\|v\|=1}|v^\top W_nv|}{\min_{v:\|v\|=1}|v^\top D_nv|}\\&=\frac{\lambda_1(W)}{\min_i|d_{ii}|}\end{aligned}$$

Recall that $d_{ii} = \sum_{j=1}^{n} w_{ij} = \sum_{j=1}^{n}(f_{ij} + x_{ij})$, where $w_{ij}$, $f_{ij}$, and $x_{ij}$ are the $(i,j)^{\text{th}}$ elements of $W_n, F_n$, and $X_n$ respectively. Since $A_n = D_n^{-1}W_n$ is a Markov matrix we assume that $f_{ij}$ and the distribution of $x_{ij}$ are selected to guarantee that $w_{ij} \geq 0$, e.g., it must be true that $f_{ij} \geq 0$. We also know that $x_{ij}$ is a zero-mean unit-variance sub-Gaussian random variable, therefore

$$\begin{aligned}\Pr\{w_{ij} < \delta\} &= \Pr\{x_{ij} < -f_{ij} + \delta\} \\ &\leq 2\exp\left(-\frac{1}{2}(f_{ij}-\delta)^2\right) \\ \Pr\{d_{ii} < n\delta\} &= \Pr\left\{\sum_{j=1}^{n} w_{ij} < n\delta\right\} \\ &\leq 2\exp\left(-\alpha_2 n\right)\end{aligned}$$

where $\alpha_2 = -\frac{1}{2}(\bar{f}_i - \delta)^2$, and $\bar{f}_i = \frac{1}{n}\sum_j f_{ij}$. Therefore, with probability at least $1 - O(\exp(-\alpha_0 n))$ where $\alpha_0 = \alpha_1 + \alpha_2$,

$$\lambda_i(A_n) \in [-10n^{\gamma-1}, 10n^{\gamma-1}],\ 1 \leq i \leq n \quad (22)$$

□

**Remark.** Lemma 5 ensures that for any symmetric Markov random matrix $A_n = D_n^{-1}W_n$ with properties defined in Theorem 2, we can focus our attention on any eigenvector $v$ whose eigenvalue is no greater than $O(n^{\gamma-1})$ and whose $\|W_n v\|_2$ is of order $n^{\gamma}$ with high probability. Therefore, we will assume such conditions in later proofs.

Using Lemmas 4-5, we can reduce Theorem 2 to the following statement on small ball probability of the *eigenvectors* of $X_n$, analogous to the arguments for symmetric random matrices in (Nguyen et al., 2017).

**Lemma 6.** *Let $A_n = D_n^{-1}W_n \in \mathbb{R}^{n\times n}$ be a symmetric Markov random matrix with adjacency matrix $W_n$. Let $\lambda_i(A_n)$ and $w = [u^\top, b]^\top \in \mathbb{R}^n$ be the $i$-th eigenvalue and eigenvector of the matrix $A_n$, respectively, where $u \in \mathbb{R}^{n-1}$ and $b \in \mathbb{R}$. Then we have*

$$\begin{aligned}&\Pr[|\lambda_i(A_n) - \lambda_{i+1}(A_n)| \leq \delta] \leq \\ &\quad n\Pr[\rho_{\delta n^{\gamma+1}}(v) \geq c_0 n^{\gamma+1}\delta] + c_0 n^{\gamma+2}\delta \\ &\qquad + O(\exp(-\alpha_0 n)),\end{aligned}$$

*for some $c_0, \alpha_0 > 0$.*

*Proof.* Let $W_{n-1}$ and $D_{n-1}$ be the $(n-1)$-dimensional minors of $W_n$ and $D_n$, respectively, then

$$\begin{bmatrix} W_{n-1} & w_n \\ w_n^\top & w_{nn}\end{bmatrix}\begin{bmatrix} u \\ b\end{bmatrix} = \lambda\begin{bmatrix} D_{n-1} & \mathbf{0}_n \\ \mathbf{0}_n^\top & d_{nn}\end{bmatrix}\begin{bmatrix} u \\ b\end{bmatrix},$$

where $w_n$ is the last column of $W_n$. Let $v$ be the $i$-th eigenvector of matrix $A_{n-1} := D_{n-1}^{-1}W_{n-1}$, we have

$$\begin{aligned}&v^\top W_{n-1}u + v^\top W b = \lambda_i(A_n)v^\top D_{n-1}u \\ \Longrightarrow&|(\lambda_i(X_{n-1}) - \lambda_i(X_n))| \max_{1\leq i\leq n} d_{ii} \geq \\ &|(\lambda_i(A_{n-1}) - \lambda_i(A_n))v^\top D_{n-1}u| = |v^\top w_n b|.\end{aligned}$$

Therefore,

$$\begin{aligned}&\Pr[|\lambda_i(A_n) - \lambda_i(A_{n-1})| \leq \delta] \\ \leq&\Pr\left[\frac{|v^\top w_n|}{\max_{1\leq i\leq n} d_{ii}} \leq \frac{\delta}{b}\right].\end{aligned}$$

By Lemma 4, $\lambda_{i+1}(A_n) \leq \lambda_i(A_{n-1}) \leq \lambda_i(A_n)$ and we have

$$\begin{aligned}&\Pr[|\lambda_i(A_n) - \lambda_{i+1}(A_n)| \leq \delta] \leq \\ &\quad\Pr[|\lambda_i(A_{n-1}) - \lambda_i(A_n)| \leq \delta] \leq \\ &\qquad\Pr\left[\frac{|v^\top w_n|}{\max_{1\leq i\leq n} d_{ii}} \leq \frac{\delta}{b}\right].\end{aligned}$$

$d_{ii}$ is typically $O(n)$, but we have been unable to prove that it is necessarily $O(n)$. Consider that $w_{ij} = f_{ij} + x_{ij}$, where $F_n$ is a symmetric matrix with eigenvalues $\lambda_i(F_n) = O(n^{\gamma})$, therefore

$$\begin{aligned}\sum_{j=1}^{n} f_{ij} = (F_n\mathbf{1}_n)_i &\leq \|F_n\mathbf{1}_n\|_2 = \|F_n\|_1 \\ &\leq n^{1/2}\|F_n\|_2 = O\left(n^{\gamma+\frac{1}{2}}\right).\end{aligned}$$

$W_n = F_n + X_n$, therefore

$$\begin{aligned}&\Pr\left\{d_{ii} \neq O\left(n^{\gamma+\frac{1}{2}}\right)\right\} \\ &\leq \Pr\left\{\sum_{j=1}^{n} x_{ij} > \sum_{j=1}^{n} f_{ij} - n\delta\right\} \\ &\leq O(\exp(-\alpha_2 n))\end{aligned}$$

Now, by the law of total probability,

$$\begin{aligned}
&\Pr\left[\frac{|v^\top w_n|}{\max_{1\le i\le n} d_{ii}} \le \frac{\delta}{b}\right] \\
\le &\Pr\left[\frac{|v^\top w_n|}{\max_{1\le i\le n} d_{ii}} \le \frac{\delta}{b}, \max_{1\le i\le n} d_{ii} \le O(n^{\gamma+\frac{1}{2}})\right] \\
&+ \Pr\left[\max_{1\le i\le n} d_{ii} \ne O\left(n^{\gamma+\frac{1}{2}}\right)\right] \\
\le &\Pr\left[|v^\top w_n| = O\left(\frac{\delta n^{\gamma+\frac{1}{2}}}{b}\right)\right] + O(\exp(-\alpha_2 n)).
\end{aligned}$$

By symmetry, we can choose any row and the corresponding column to split the matrix and derive inequality of the same form. Further, suppose for some $b_1 > 0$, with probability at least $1 - \exp(-c_1 n)$, there are at least $n_T$ coordinates of $w$ that are at least $b_1$ and suppose we choose the split index $J$ uniformly at random. Let the $J$-th column of $W_n$ be $W$ and the $J$-th coefficient of the eigenvector of $W_n$ be $w_J$, then we have

$$\begin{aligned}
&\Pr[|\lambda_i(A_n) - \lambda_{i+1}(A_n)| \le \delta] \\
\le &\Pr\left[|v^\top W| \ne O\left(\frac{\delta n^{\gamma+\frac{1}{2}}}{w_J}\right) | N_b \ge n_b\right] \\
&+ O(\exp(-c_1 n)) + O(\exp(-\alpha_2 n)) \\
\le &\frac{n}{n_T} \Pr\left[|v^\top W| \ne O\left(\frac{\delta n^{\gamma+\frac{1}{2}}}{b_1}\right) | N_b \ge n_b\right] \\
&+ O(\exp(-c_1 n)) + O(\exp(-\alpha_2 n)),
\end{aligned}$$

where the second inequality can be proved as follows. Define

$$\begin{aligned}
\mathcal{E} &= \{N_b \ge n_b\}, \\
\mathcal{F} &= \{w_J \ge b_1\}, \\
\mathcal{G} &= \left\{|v^\top W| \ne O\left(\frac{\delta n^{\gamma+1/2}}{w_J}\right)\right\}, \\
\mathcal{H} &= \left\{|v^\top W| \ne O\left(\frac{\delta n^{\gamma+1/2}}{b_1}\right)\right\}.
\end{aligned}$$

Then use the above definitions and the fact that $\mathcal{F}$ and $\mathcal{G}$ are conditionally independent given $N_b$, we have

$$\begin{aligned}
&\Pr\left[|v^\top W| \ne O\left(\frac{\delta n^{\gamma+\frac{1}{2}}}{b_1}\right) | N_b \ge n_b\right] \\
= &\Pr(\mathcal{H}|\mathcal{E}) \ge \Pr(\mathcal{F}\cap\mathcal{G}|\mathcal{E}) \ge \frac{n_T}{n}\Pr(\mathcal{G}|\mathcal{E}) \\
= &\frac{n_T}{n}\Pr\left[|v^\top W| \ne O\left(\frac{\delta n^{\gamma+1/2}}{w_J}\right) | N_b \ge n_b\right].
\end{aligned}$$

Further, to remove the dependency on $N_b$, notice that

$$\Pr(\mathcal{H}|\mathcal{E}) \le \frac{\Pr(\mathcal{H})}{\Pr(\mathcal{E})} = \Pr(\mathcal{H}) + O(\exp(-c_1 n)).$$

Next, by the pigeonhole principle, at least one coordinate of the unit eigenvector $w$ is at least $n^{-1/2}$, and thus we can let $c_1 = \infty$, $n_b = 1$ and $b_1 = n^{-1/2}$ and arrive at

$$\begin{aligned}
&\Pr\left[|\lambda_i(A_n) - \lambda_{i+1}(A_n)| \le \delta\right] \\
\le &n \Pr\left[|v^\top W| \ne O\left(\delta n^{\gamma+1}\right)\right] + O(e^{-\alpha_0 n}) \\
\le &n\rho_{\delta O(1) n^{\gamma+1}}(v) + O(\exp(-\alpha_0 n)), \qquad (23)
\end{aligned}$$

where $\alpha_0 = c_1 + \alpha_2$. Finally, recall the definition of small ball probability, we have

$$\begin{aligned}
\Pr\left[|v^\top W| \le \delta\right] &\le \Pr\left[|v^\top W| \le \delta | \rho_\delta(v) \le \epsilon\right] \\
&\quad + \Pr[\rho_\delta(v) > \epsilon] \\
&\le \Pr[\rho_\delta(v) > \epsilon] + \epsilon,
\end{aligned}$$

and thus applying this inequality with $\delta := c_0 \delta n^{\gamma+1}$ on Eq. (23) yields the result. □

**Remark.** We can sharpen the bound in Lemma 6 by extending the delocalization theorem for a symmetric Wigner matrix (see Theorem 4.2 of (Nguyen et al., 2017)) to a symmetric Markov random matrix and using it to choose a larger $n_b$ in the proof. This will be left as future work.

With the help of Lemma 6, we can reduce Theorem 2 to the following theorem.

**Theorem 5.** *Let $A_n \in \mathbb{R}^{n\times n}$ be a symmetric Markov random matrix matrix and $v$ be an eigenvector with eigenvalue $\lambda = O(n^{\gamma-1})$, then for any fixed $C > 0$, there exists some $B > \max\{4\gamma C + 3\gamma, 4\gamma + 1\}$ such that*

$$\rho_{n^{-B}}(v) \le n^{-C},$$

*with probability at least $1 - O(\exp(-\alpha_0 n))$ for some $\alpha_0$ depending on $B$.*

Similar to the proof for the perturbed symmetric matrices in (Nguyen et al., 2017), we reduce Theorem 5 to the following.

**Theorem 6.** *Let $v$ be the eigenvector and $B$ be the constant defined in Theorem 5. Then for any $n^{-B} \le \delta \le n^{-B/2}$, we have with probability $O(\exp(-\alpha_0 n))$,*

$$n^{-C} \le \rho_{n^\gamma \delta}(v) \le n^{0.49}\rho_\delta(v). \qquad (24)$$

To show that Theorem 6 implies Theorem 5, we prove the contrapositive of the statement, that is, if $\rho_{n^{-B}}(v) > n^{-C}$, then there exists $n^{-B} \le \delta \le n^{-B/2}$ such that Eq. 24 holds with probability at least $1 - O(\exp(-\alpha_0 n))$. To construct such $\delta$, let

$$\delta_0 := n^{-B}$$
$$\delta_{j+1} := n^{\gamma}\delta_j,$$

for $j = 0, \cdots, J-1$ with $J = \lfloor B/2\gamma \rfloor$. By construction, we have

$$n^{-B} = \delta_0 \le \delta_j \le \delta_J \le n^{-B/2}$$
$$\rho_{\delta_j}(v) \ge \rho_{\delta_0}(v) \ge n^{-C}.$$

Suppose Eq. 24 does not hold for any $\delta := \delta_j$, or otherwise the result follows, we have

$$\rho_{\delta_J}(v) \ge n^{0.49J}\rho_{n^{-B}}(v) \ge n^{0.49\lfloor B/2\gamma \rfloor - C} > 1,$$

if $B \ge 4\gamma C + 3\gamma$, which contradicts the fact that $\rho_{\delta_J}(v) \le 1$. As a result, there has to exist some $j$ such that Eq. 24 holds.

Again similar to the perturbed symmetric matrix case in (Nguyen et al., 2017), we divide the proof of Theorem 6 into the compressible case and the non-compressible case. For the compressible case, we first prove the following lemma.

**Lemma 7.** *Suppose $v$ is an eigenvector of a symmetric Markov random matrix $A_n := D_n^{-1}W_n$ with adjacency matrix $W_n$ and the same properties defined in Theorem 2, and suppose there exists $\delta \in [n^{-B}, n^{-B/2}]$ such that $\rho_{\delta,\alpha}(v) \ge (\alpha n)^{-1/2+\epsilon}$, we have with probability $O(\exp(-\alpha_0 n))$,*

$$n^{-C} \le \rho_{n^{\gamma}\delta}(v) \le n^{0.49}\rho_{\delta}(v).$$

*Proof.* Using concentration inequalities, we have with probability at least $1 - O(\exp(-\alpha_2 n))$ for some $\alpha_2 > 0$,

$$d_{ii} = O\left(n^{\gamma+\frac{1}{2}}\right),\ 1 \le i \le n \qquad (25)$$

Further, since $\rho_{\delta,\alpha}(v) \ge (\alpha n)^{-1/2+\epsilon}$, by Lemma 2, we have $v$ is $(O(\alpha), \delta)$ compressible, and thus there exists $I$ of of size $O(\alpha n)$ such that $v_i > \delta$ only if $i \in I$. Without loss of generality, let $I = \{n-k, \cdots, n\}$ for $k = O(\alpha n)$ and $\mathbb{E}[A_{ij}] = 1$. Further, split $v = [v'^{\top}, v''^{\top}]^{\top}$, then by definition of eigenvalues and eigenvectors,

$$\begin{bmatrix} W_{n-k} & F \\ F^{\top} & W_k \end{bmatrix} \begin{bmatrix} v' \\ v'' \end{bmatrix} = \lambda \begin{bmatrix} D_{n-k} & \mathbf{0} \\ \mathbf{0}^{\top} & D_k \end{bmatrix} \begin{bmatrix} v' \\ v'' \end{bmatrix}.$$

Reading off the first line of the matrix equation, we have

$$\|Fv''\|_2 = \|(W_{n-k} - \lambda D_{n-k})v'\|_2$$
$$\le \|W_{n-k}v'\|_2 + \|\lambda D_{n-k}v'\|_2.$$

Notice that assuming Eq. 25 and Eq. 22 occur, we have that all elements $v'_i$ of $v'$ have $|v'_i| < \delta$, therefore $\|v'\|_2 \le \delta n^{-1/2}$, therefore

$$\|W_{n-k}v'\|_2 \le \delta n^{1/2} \max_{v:\|v\|_2=1} \|Wv\|_2$$
$$= O(n^{-B/2+1/2+\gamma})$$

Furthermore, if we assume that Eq. (22) and Eq. (25) occur, then

$$\|\lambda D_{n-k}v'\|_2 = O(n^{\gamma-1} \cdot n^{\gamma+1/2} \cdot \delta n^{1/2})$$
$$= O(n^{-B/2+2\gamma}).$$

Thus, using the fact that $B \ge 4\gamma + 1$,

$$\|Fv''\|_2 = O(n^{-B/2+2\gamma}) = O(n^{-1/2}).$$

On the other hand, using a standard epsilon-net argument, with probability at least $1 - O(\exp(-\alpha_3 n))$,

$$\inf_{w \in \mathbb{R}^k : \|w\|=1} \|Fw\|_2 \ge n^{-1/2}.$$

Now, define the events

$$\mathcal{E} := \{v \text{ is an eigenvector of } A\}$$
$$\mathcal{E}_{\alpha,\delta} := \{v \text{ is } (O(\alpha), \delta)\text{-compressible}\}$$
$$\mathcal{E}_I := \{\|W_{I^c,I}v_I\|_2 \gg O(n^{-1/2})\},$$

then by the previous discussion, we have

$$\Pr(\mathcal{E}_I | \mathcal{E} \cap \mathcal{E}_{\alpha,\delta}) = O(\exp(-\alpha_2 n))$$
$$\Pr(\mathcal{E}_I^c | \mathcal{E}) = O(\exp(-\alpha_3 n)).$$

Note that to prove the lemma, it suffices to show that the eigenvector $v$ is not $(O(\alpha), \delta)$-compressible with high probability, or $\Pr(\mathcal{E}_{\alpha,\delta}|\mathcal{E})$ is small, since that will lead to $\rho_{\delta,\alpha}(v) < (\alpha n)^{-1/2+\epsilon}$ with high probability and thus a contradiction with high probability. Indeed, we have

$$\Pr(\mathcal{E}_{\alpha,\delta}|\mathcal{E}) \le \Pr(\mathcal{E}_{\alpha,\delta} \cap \mathcal{E}_I | \mathcal{E}) + \Pr(\mathcal{E}_{\alpha,\delta} \cap \mathcal{E}_I^c | \mathcal{E})$$
$$\le \Pr(\mathcal{E}_I | \mathcal{E} \cap \mathcal{E}_{\alpha,\delta}) + \Pr(\mathcal{E}_I^c | \mathcal{E})$$
$$= O(\exp(-\alpha_0 n))$$

for some $\alpha_0 > 0$. □

For the incompressible case, we apply the continuous inverse Offord-Littlewood theorem to discretize the set of eigenvectors, and prove the following result analogous to the symmetric case in (Nguyen and Vu, 2011).

**Lemma 8.** *Suppose $v$ is an eigenvector of a symmetric Markov random matrix $A_n := D_n^{-1}W_n$ with adjacency matrix $W_n$ and the same properties defined in Theorem 2, and suppose there exists $\delta \in [n^{-B}, n^{-B/2}]$ such that $q := \rho_{\delta,\alpha}(v) < (\alpha n)^{-1/2+\epsilon}$, we have with probability $O(\exp(-\alpha_0 n))$,*

$$n^{-C} \le \rho_{n^\gamma \delta}(v) \le n^{0.49}\rho_\delta(v).$$

To prove this result, we need the following useful lemmas.

**Lemma 9.** *For any eigenvector-eigenvalue pair $(v, \lambda)$ and $\alpha > 0$ with $|\lambda| = O(n^{\gamma-1})$, suppose $n^{-C} < \rho_{\delta,\alpha}(v) =: q \le (\alpha n)^{-1/2+\epsilon}$, then with probability at least $1 - O(\exp(-\alpha_0 n))$ there exists a subset $\mathcal{N}$ of $\mathbb{R}^n \times \mathbb{R}$ of size $O(n^{-n/2+O(\alpha n)}q^{-n+O(\alpha n)})$ such that, there exists $(\tilde{v}, \tilde{\lambda}) \in \mathcal{N}$ with the properties:*

1. *$|v_j - \tilde{v}_j| \le \delta$ for $1 \le j \le n$;*
2. *$|\lambda - \tilde{\lambda}| \le n^\gamma \delta$.*

*Proof.* Split $\{1, \cdots, n\}$ into sets of length differing by at most 1, $I_1, \cdots, I_m$, $m = \lfloor \frac{1}{\alpha} \rfloor + 1$, then we have the length of each set is greater than or equal to $\lfloor \alpha n \rfloor$, and its small ball probability is

$$\rho_\delta(v_{I_i}) \ge \rho_{\delta,\alpha}(v) = q, 1 \le i \le m.$$

Therefore, since $q \le (\alpha n)^{-\frac{1}{2}+\epsilon}$ and $n^{-C} < q$, there exists a GAP

$$Q_i = \left\{ \sum_{j=1}^{r_i} a_{ij} w_{ij} : \begin{array}{ll} a_j \in \mathbb{Z}, \\ |a_{ij}| \le N_{ij}, \\ 1 \le j \le r_i \end{array} \right\}$$

such that

$$\sup_{j \in I_i \setminus S} \inf_{\tilde{v}_j \in Q_i} |v_j - \tilde{v}_j| \le \delta,$$

with volume

$$\mathrm{vol}(Q_i) \le O((\alpha n)^{-1/2+\epsilon}/q), 1 \le i \le m,$$

for all except at most $O(\alpha^2 n)$ indices from some exceptional set $S$. Further, for each $Q_i$, we can quantize its generators $w_{i1}, \cdots, w_{ir_i}$ to the closest multiple of $q\delta$, $\tilde{w}_{i1}, \cdots, \tilde{w}_{ir_i}$. This introduces an additional approximation error of at most

$$\begin{aligned} &\left| \sum_{j=1}^{r_i} a_{ij} w_{ij} - \sum_{j=1}^{r_i} a_{ij} \tilde{w}_{ij} \right| \\ \le &\mathrm{vol}(Q_i) \cdot q\delta \le (\alpha n)^{-1/2+\epsilon}/q \cdot q\delta \\ = &(\alpha n)^{-1/2+\epsilon}\delta = O(\delta). \end{aligned}$$

Next, for the coefficients from the exceptional set $S$, we also round them to the closest multiple of $q\delta$ and let the set of such values be $R$, which ensures that

$$\sup_{j \in S} \inf_{v' \in R} |v_j - v'| = O(\delta).$$

Therefore, for fixed generators $w_{ij}$'s and a given $S$, we can construct a finite set of vectors

$$\{\tilde{v} : \tilde{v}_j \in \cup_{i=1}^m Q_i, \forall j \notin S \text{ and } v'_j \in R, \forall j \in S\}$$

of size at most

$$\begin{aligned} &\left(m \sup_i \mathrm{vol}(Q_i)\right)^{n-|S|} |R|^{|S|} \\ \le &O\left(\frac{1}{\alpha}\frac{(\alpha n)^{-1/2+\epsilon}}{q}\right)^n \cdot O((1/q\delta)^{O(\alpha n)}) \\ \le &O\left(n^{-\frac{n}{2}+\epsilon n} q^{-n+O(\alpha n)}\right) O\left(n^{B\alpha n}\right) \\ = &O(n^{-n/2+O(\alpha n)}q^{-n+O(\alpha n)}), \\ = &O(n^{-n/2+O(\alpha n)}q^{-n}), \end{aligned}$$

that approximates $v$ within $O(\delta)$ for every coefficients. The third line uses $\delta > n^{-B}$ and $\alpha = O(1)$; the fourth line assumes $\epsilon = O(\alpha)$. Further, if we allow the generators to be variable and assume $S$ to be unknown, the quantization mentioned previously and the crude bound of the number of possible $S$ by $2^n$ enlarges the set of vectors by a factor of

$$\begin{aligned} &O\left((1/q\delta)^{\sum_{i=1}^m r_i}\right) \cdot O(2^n) = O(n^{O(m)}) \cdot O(2^n) \\ &= O(n^{O(1/\alpha)}) \cdot O(2^n) = O(n^{O(\alpha n)}). \end{aligned}$$

For the eigenvalue, we also have there exists a set that covers its domain to be within $\delta n^\gamma$ with a set of size

$$O\left(\frac{n^{\gamma-1}}{n^\gamma \delta}\right) = O(n^{B-1}) \le O(n^{O(\alpha n)}).$$

with probability at least $1 - O(\exp(-\alpha_0 n))$. Composing the sets, we find the set $\mathcal{N}$ has size $O(n^{-n/2+O(\alpha n)}q^{-n+O(\alpha n)})$. □

**Lemma 10.** *For any eigenvector-eigenvalue pair $(v, \lambda)$ of an symmetric Markov random matrix $A_n = D_n^{-1} W_n$ with adjacency matrix $W_n$ and the same properties defined in Theorem 2 and let $(\tilde{v}, \tilde{\lambda}) \in \mathcal{N}$ be the tuple that well approximates it as defined in Lemma 9, we have*

$$\|A_{I^c,I}\tilde{v}_I - u\|_2 = O(\delta n^\gamma),$$

*where $A_{I,J}$ is the matrix formed by row indices from $I$ and column indices from $J$ and $u := (\tilde{\lambda} - A_{I^c,I^c})\tilde{v}_{I^c}$.*

*Proof.* By symmetry, we can let $I = \{1, \cdots, k\}$ for $k = \lfloor \alpha n \rfloor$. Notice by definition we can split $A$ as

$$\begin{bmatrix} A_k & G \\ F^\top & A_{n-k} \end{bmatrix} \begin{bmatrix} w \\ v' \end{bmatrix} = \lambda \begin{bmatrix} w \\ v' \end{bmatrix},$$

where $v = [w^\top, v'^\top]^\top$, and as a result,

$$\begin{aligned}
&\|F^\top \tilde{v}_I - (\tilde{\lambda} - A_{n-k})\tilde{v}_{I^c}\|_2 \\
\le &\|F^\top w - (\lambda - A_{n-k})v'\|_2 + \\
&\|F^\top(\tilde{v}_I - w)\|_2 + \|(\tilde{\lambda} - \lambda)\tilde{v}_{I^c}\|_2 + \\
&\|(\lambda - A_{n-k})(\tilde{v}_{I^c} - v')\|_2 \\
= &\|F^\top(\tilde{v}_I - w)\|_2 + |(\tilde{\lambda} - \lambda)\tilde{v}_{I^c}\|_2 \\
&+ \|(\lambda - A_{n-k})(\tilde{v}_{I^c} - v')\|_2 \\
= &O(n^{\gamma-1} \cdot \delta n^{1/2}) + O(n^\gamma \delta) \\
&+ O(n^{\gamma-1} \cdot \delta n^{1/2}) = O(n^\gamma \delta).
\end{aligned}$$

□

Now we are ready to prove Lemma 8.

*Proof.* Let $\mathcal{E}$ be the event that there exists some $\delta \in [n^{-B}, n^{-B/2}]$ such that

$$n^{-C} \le \rho_{n^\gamma \delta}(v) \le n^{0.49} \rho_\delta(v) =: n^{0.49} q$$

with $q := \rho_\delta(v)$ and $\mathcal{G}$ be the event that

$$\|A_{I^c,I}\tilde{v}_I - u\|_2 = O(\delta n^\gamma),$$

where $u := (\tilde{\lambda} - A_{I^c,I^c})\tilde{v}_{I^c}$ and $(\tilde{v}, \lambda)$ well approximates $(v, \lambda)$ as defined in Lemma 10. Let $k := |I| = O(\alpha n)$, from Lemma 9, we have

$$\Pr(\mathcal{G}^c) = O(\exp(-\alpha_0 n)).$$

On the other hand, if $\mathcal{E}$ occurs, define $A_{I^c,I} = [a_{k+1}, \ldots, a_n]^\top$, $u = [u_{k+1}, \ldots, u_n]^\top$, then we have

$$\begin{aligned}
\Pr(\mathcal{G}|\mathcal{E}) \le &\sum_{(w',\tilde{v},\tilde{\lambda}) \in \mathcal{N}} \\
&\Pr\left[\sum_{i=k+1}^{n} |a_i^\top w' - u_i|^2 = O(\delta^2 n^{2\gamma+1})\right] \\
&\le |\mathcal{N}|(\rho_{n^\gamma\delta}(v))^{n-k} \le |\mathcal{N}|(n^{0.49}q)^{n-k} \\
&= O(n^{-0.01n + O(\alpha n)}),
\end{aligned}$$

which is $O(\exp(-\alpha_0 n))$ if $\alpha$ is chosen small enough. As a result, we have

$$\Pr(\mathcal{E}) \le \Pr(\mathcal{G}|\mathcal{E}) + \Pr(\mathcal{G}^c) = O(\exp(-\alpha_0 n)).$$

□

### A.3 Finite-sample learnability of ASR-U: Unmatched setup

*Proof.* (Theorem 3) Under the assumptions that the discriminator is perfect and decomposable and the GAN objective is MMD with a linear kernel over the embeddings $D(Y) = \hat{P}^Y$, Eq. (8) becomes the following least squares regression problem

$$\min_{O' \in \mathbb{R}^{|\mathbb{X}| \times |\mathbb{Y}|}} \|\hat{P}^X O' - \hat{P}^Y\|_F^2. \tag{26}$$

Let $\hat{O}$ be the ERM of Eq. (26) and $O$ be the true assignment matrix, by definition and triangle inequality,

$$\begin{aligned}
&\|\hat{P}^X \hat{O} - \hat{P}^Y\|_F \\
&\le \|\hat{P}^X O - \hat{P}^Y\|_F \\
&\le \|\hat{P}^X O - P^Y\|_F + \|\hat{P}^Y - P^Y\|_F.
\end{aligned}$$

Apply the triangle inequality again, we have

$$\begin{aligned}
&\|\hat{P}^X(\hat{O} - O)\|_F \\
&\le \|\hat{P}^X \hat{O} - \hat{P}^Y\|_F + \|\hat{P}^X O - \hat{P}^Y\|_F \\
&\le 2\|\hat{P}^X O - P^Y\|_F + 2\|\hat{P}^Y - P^Y\|_F
\end{aligned}$$

Note that if we replace any $X^{(i)} \to X^{(i)'}$ and let the resulting empirical distribution be $\hat{P}^{X'}$,

$$\begin{aligned}
&\left|\|\hat{P}^X O - P^Y\|_F - \|\hat{P}^{X'} O - P^Y\|_F\right| \\
\le &\|(\hat{P}^X - \hat{P}^{X'})O\|_F \le \frac{\sqrt{2L}}{n^X},
\end{aligned}$$

and similarly for $\hat{P}^X$ and $\hat{P}^Y$,

$$\begin{aligned}
\left|\|\hat{P}^X - P^X\|_F - \|\hat{P}^{X'} - P^X\|_F\right| &\le \frac{\sqrt{2L}}{n^X} \\
\left|\|\hat{P}^Y - P^Y\|_F - \|\hat{P}^{Y'} - P^Y\|_F\right| &\le \frac{\sqrt{2L}}{n^Y}.
\end{aligned}$$

Therefore, we can apply McDiarmid's inequality to obtain

$$\Pr\left[\|\hat{P}^X - P^X\|_F \geq \frac{\sqrt{L|\mathbb{X}|}}{\sqrt{n^X}} + \epsilon\right] \leq e^{-\frac{n^X\epsilon^2}{L}}$$

$$\Pr\left[\|\hat{P}^X O - P^Y\|_F \geq \frac{\sqrt{L|\mathbb{Y}|}}{\sqrt{n^X}} + \epsilon\right] \leq e^{-\frac{n^X\epsilon^2}{L}}$$

$$\Pr\left[\|\hat{P}^Y - P^Y\|_F \geq \frac{\sqrt{L|\mathbb{Y}|}}{\sqrt{n^Y}} + \epsilon\right] \leq e^{-\frac{n^Y\epsilon^2}{L}}.$$

Moreover, let $\epsilon^{XX} := \frac{\sqrt{L|\mathbb{X}|}}{\sqrt{n^X}} + \epsilon$, $\epsilon^{YX} := \frac{\sqrt{L|\mathbb{Y}|}}{\sqrt{n^X}} + \epsilon$, $\epsilon^{YY} = \frac{\sqrt{L|\mathbb{Y}|}}{\sqrt{n^Y}} + \epsilon$, then by a union bound, we have

$$\Pr\left[\|\hat{P}^X(\hat{O} - O)\|_F \geq \epsilon^{YX} + \epsilon^{YY}\right] \leq$$
$$\Pr\left[\|\hat{P}^X\hat{O} - P^Y\|_F + \|\hat{P}^Y - P^Y\|_F \geq \frac{\epsilon^{YX} + \epsilon^{YY}}{2}\right]$$
$$\leq \Pr\left[\|\hat{P}^{YX}\hat{O} - P^{YY}\|_F \geq \frac{\epsilon^{YX}}{2}\right] +$$
$$\Pr\left[\|\hat{P}^Y - P^Y\|_F \geq \frac{\epsilon^{YY}}{2}\right] \leq e^{-\frac{n^X\epsilon^2}{4L}} + e^{-\frac{n^Y\epsilon^2}{4L}}.$$

Therefore, we have with probability at least $1 - e^{-\frac{n^X\epsilon^2}{4L}} - e^{-\frac{n^Y\epsilon^2}{4L}}$,

$$\epsilon^{YX} + \epsilon^{YY} \geq \|\hat{P}^X(\hat{O} - O)\|_F$$
$$\geq \|P^X(\hat{O} - O)\|_F - \|\hat{P}^X - P^X\|_F\|\hat{O} - O\|_F$$
$$\geq (\sigma_{\min}(P^X) - \|\hat{P}^X - P^X\|_F)\|\hat{O} - O\|_F,$$

and combined with the bound on $\|\hat{P}^X - P^X\|_F$, we obtain with probability at least $(1 - e^{-\frac{n^X\epsilon^2}{4L}} - e^{-\frac{n^Y\epsilon^2}{4L}})(1 - e^{-\frac{n^X\epsilon^2}{4L}})$,

$$\|\hat{O} - O\|_F \leq \frac{\epsilon^{YX} + \epsilon^{YY}}{\sigma_{\min}(P^X) - \epsilon^{XX}}.$$

Assume the correct mapping is deterministic, so that $O_{xy} \in \{0, 1\}$ and each row has only one nonzero element, then to achieve perfect ASR-U, we need for any $x \in \mathbb{X}$ and $y \neq G(x)$,

$$|\hat{O}_{xG(x)} - \hat{O}_{xy}| > 0$$
$$\Longleftarrow 1 - |\hat{O}_{xG(x)} - O_{xG(x)}| - |\hat{O}_{xy} - O_{xy}| > 0$$
$$\Longleftarrow 1 - 2\|\hat{O} - O\|_\infty > 0 \Longleftrightarrow \|\hat{O} - O\|_F < \frac{1}{2},$$

which occurs if

$$\sigma_{\min}(P^X) > \epsilon^{XX} + 2\epsilon^{YX} + 2\epsilon^{YY}.$$

□

### A.4 Training dynamic of ASR-U

To prove Theorem 4, we need the following lemma on the properties of the gradient of the softmax function based on (Gao and Pavel, 2017).

**Lemma 11.** *Let $H(x)$ be the Jacobian matrix of the softmax function $\sigma : \mathbb{R}^d \mapsto \mathbb{R}^d$ with $\sigma_i(x) = \frac{e^{x_i}}{\sum_{j=1}^d e^{x_j}}$, then we have $H(x) = \mathrm{diag}(\sigma(x)) - \sigma(x)\sigma(x)^\top$ and $H(x)$ is positive semi-definite (PSD) with the null space $\mathrm{span}\{\mathbf{1}_d\}$.*

*Proof.* Apply product rule of calculus, we have

$$H_{ij}(x) = \frac{\partial\sigma_i(x)}{\partial x_j}$$
$$= \delta_{ij}\sigma_i(x) - \frac{e^{x_i}e^{x_j}}{(\sum_{j=1}^d e^{x_j})^2}$$
$$= \delta_{ij}\sigma_i(x) - \sigma_i(x)\sigma_j(x),$$

and therefore $H(x) = \mathrm{diag}(\sigma(x)) - \sigma(x)\sigma(x)^\top$. To show that $H(x)$ is PSD, notice that

$$v^\top H(x)v = v^\top\mathrm{diag}(\sigma(x))v - (v^\top\sigma(x))^2$$
$$= \mathbb{E}_{I\sim\sigma(x)}[v_I^2] - \mathbb{E}^2_{I\sim\sigma(x)}[v_I]$$
$$= \mathrm{Var}(v_I) \geq 0,$$

where by Jensen's inequality, achieves "=" if and only if $v_i = \sigma^\top v = C$, $\forall i$ for some constant $C$. □

Next, we shall establish explicit formula for NTKs of the discriminator and the generator. For clarity, we will copy the formula for the discriminator and the generator used in our analysis:

$$f_{\tau,l}(y) = \lim_{m\to\infty}\frac{1}{\sqrt{m}}\sum_{r=1}^m v_r^{\tau,l}\max\{W_{ry}^{\tau,l}, 0\}, \tag{27}$$

$$P_l^{g_t}(y) = \mathbb{E}_{X\sim P_l^X}\left[O_t(y|X)\right]$$
$$:= \mathbb{E}_{X\sim P_l^X}\left[\frac{\exp(U_y^{t\top}x)}{\sum_{y'\in\mathbb{Y}}\exp(U_{y'}^{t\top}x)}\right]. \tag{28}$$

**Lemma 12.** *For the NTKs of the discriminators defined by Eq. (27), we have $K_{D,l} \equiv K_{D,1}$, $1 \leq l \leq L$ and $\mathbf{1}_{|\mathbb{Y}|}$ is an eigenvector of $K_{D,1}$.*

*Proof.* For simplicity, we ignore the dependency on $\tau$ for the terms in the proof. First, by definition, we have

$$\frac{\partial f_l(y)}{\partial W_r^l} = \lim_{m\to\infty}\frac{1}{\sqrt{m}}\sum_{r=1}^m v_r^l e_y\mathbb{1}[W_{ry}^l \geq 0],$$
$$\frac{\partial f_l(y)}{\partial v_r^l} = \lim_{m\to\infty} = \frac{1}{\sqrt{m}}\max\{W_{ry}^l, 0\}$$

and therefore

$$\mathbb{E}_{v^l,W^l\sim\mathcal{N}(0,I)}\left[\frac{\partial f_l(y)}{\partial W_r^l}^\top\frac{\partial f_l(y)}{\partial W_r^l}\right]=$$
$$\lim_{m\to\infty}\frac{1}{m}\mathbb{E}_{v^l,W^l\sim\mathcal{N}(0,I)}\sum_{r=1}^m\delta_{yy'}v_r^2\mathbb{1}[W_{ry}^l\geq 0]$$
$$=\delta_{yy'}\frac{1}{m}\sum_{r=1}^m\mathbb{E}_{W_{ry}^l\sim\mathcal{N}(0,1)}[\mathbb{1}[W_{ry}^l\geq 0]]$$
$$=\frac{1}{2}\delta_{yy'}.$$

On the other hand,

$$\mathbb{E}_{v^l,W^l\sim\mathcal{N}(0,I)}\left[\frac{\partial f_l(y)}{\partial v^l}^\top\frac{\partial f_l(y')}{\partial v^l}\right]$$
$$=\frac{1}{m}\mathbb{E}_{v^l,W^l}\left[\sum_{r=1}^m\max\{W_{ry}^l,0\}\max\{W_{ry'}^l,0\}\right]$$
$$=\begin{cases}\mathbb{E}_{v_1^1,W_1^1}\left[\max\{W_{11}^1,0\}^2\right] & \text{if } y=y',\\ \mathbb{E}_{v_1^1,W_1^1}\left[\max\{W_{11}^1,0\}\right]^2 & \text{otherwise.}\end{cases}$$

Therefore,

$$K_{D,l}(y,y')=$$
$$\begin{cases}\left(\frac{1}{2}+\mathbb{E}_{v_1^1,W_1^1}\left[\max\{W_{11}^1,0\}^2\right]\right) & \text{if y=y',}\\ \mathbb{E}_{v_1^1,W_1^1}\left[\max\{W_{11}^1,0\}\right]^2 & \text{otherwise.}\end{cases}$$

Notice that the sum of every row in $K_{D,l}$ is

$$\left(\frac{1}{2}+\mathbb{E}_{v_1^1,W_1^1}\left[\max\{W_{11}^1,0\}^2\right]\right)+$$
$$(|\mathbb{Y}|-1)\mathbb{E}_{v_1^1,W_1^1}\left[\max\{W_{11}^1,0\}\right]^2,$$

and thus $\mathbf{1}_{|\mathbb{Y}|}$ is an eigenvector of $K_{D,l}$. □

**Lemma 13.** *For the generator defined by Eq. (28), we have*

$$K_{O_{t,x}}=$$
$$\mathbb{E}_{U_{1:|\mathbb{Y}|}\sim\mathcal{N}(0,I)}\left[(\mathrm{diag}(O_x)-O_xO_x^\top)^2\right]. \quad (29)$$

*Further, the null space of $K_{O_{t,x}}$ is* $\mathrm{span}\{\mathbf{1}_{|\mathbb{Y}|}\}$.

*Proof.* For simplicity, we ignore the dependency on $t$ for the terms in the proof. By chain rule,

$$\frac{\partial O_x(y)}{\partial U_{x'y'}}=\frac{\partial h_{y'}(x)}{\partial U_{xy'}}\frac{\partial O_x(y)}{\partial h_{y'}(x)}$$
$$=\delta_{xx'}(O(y|x)\delta_{yy'}-O(y|x)O(y'|x))$$

As a result,

$$\sum_{d,y'}\frac{\partial O_x(y)}{\partial U_{dy'}}^\top\frac{\partial O_x(y'')}{\partial U_{dy'}}$$
$$=\sum_{y'}(O_x(y)\delta_{yy'}-O_x(y)O_x(y'))$$
$$(O_x(y'')\delta_{y''y'}-O_x(y'')O_x(y'))\cdot$$
$$=((\mathrm{diag}(O_x)-O_xO_x^\top)^2)_{yy''}$$

Take the expectation over $U$ and put everything in matrix form, we obtain

$$K_{O_x}=\mathbb{E}_{U\sim\mathcal{N}(0,I)}\left[(\mathrm{diag}(O_x)-O_xO_x^\top)^2\right].$$

Next we shall study the null space of $K_{O_x}$. From Lemma 11, we have $H_x:=\mathrm{diag}(O_x)-O_xO_x^\top$ is PSD with null space $\mathrm{span}\{\mathbf{1}_{|\mathbb{Y}|}\}$, and thus

$$v^\top K_{O_x}v=\mathbb{E}_{U\sim\mathcal{N}(0,I)}\left[\|H_xv\|^2\right]\geq 0,$$

with equality achieved if and only if

$$H_xv=0,\forall x\in\mathbb{X}\Leftrightarrow v\in\mathrm{span}(\mathbf{1}_{|\mathbb{Y}|}).$$

□

We are now ready to prove Theorem 4.

*Proof.* (Theorem 4) When the objective is MMD, the discriminator can be decomposed as

$$a_{f_\tau}(y)=f_\tau(y)=\sum_{l=1}^L f_{\tau,l}(y_l),$$

we have

$$\mathcal{L}_t(f)=\sum_{l=1}^L\mathbb{E}_{Y_l\sim P_l^Y}[f_l(Y_l)]-\mathbb{E}_{Y_l'\sim P_l^XO_t}[f_l(Y_l')], \quad (30)$$

and the discriminator dynamic PDE Eq. (18) becomes:

$$\partial_\tau f_{\tau,l}=K_{D,l}(P_l^Y-P_l^XO_t)^\top.$$

Without much loss of generality, suppose we initialize $f_{0,l}(y)\equiv 0$ and stop training the discriminator after $\tau_{\max}$ steps. The solution for the discriminator PDE is then simply

$$f_{g_t,l}=\tau_{\max}K_{D,l}(P_l^Y-P^XO_t)^\top. \quad (31)$$

Plug this expression into the generator loss and apply Lemma 12, we obtain

$$\begin{aligned}\mathcal{C}_t(g_t) &:= \tau_{\max} \sum_{l=1}^{L} \|P_l^Y - P_l^X O_t\|_{K_{D,l}}^2 \\ &= \tau_{\max} \|P^Y - P^X O_t\|_{K_{D,1}}^2,\end{aligned}$$

where $\|A\|_K = \sqrt{\mathrm{Tr}(AKA^\top)}$ is the kernelized norm of $A$ by kernel $K$.

Further, plug Eq. (31) into the generator PDE Eq. (19), we obtain

$$\begin{aligned}\partial_t O_{t,x}^\top &= K_{O_{t,x}} \sum_{l=1}^{L} P_l^X(x) K_{D,l} (P_l^Y - P_l^X O_t)^\top \\ &= K_{O_{t,x}} K_{D,1} (P^Y - P^X O)^\top \tilde{P}_x^X,\end{aligned}$$

where $\tilde{P}_x^X$ is the $x$-th column of $P^X$. Next, notice that

$$\begin{aligned}&\frac{\partial \mathcal{C}_t}{\partial O_{t,xy}} \\ =&2\tau_{\max} K_{D,1}(y,\cdot)(P^X O - P^Y)^\top \tilde{P}_x^X \\ \Longrightarrow &\frac{\partial \mathcal{C}_t}{\partial O_t} = P^{X\top}(P^X O - P^Y) K_{D,1}.\end{aligned}$$

Then apply the chain rule,

$$\begin{aligned}\partial_t \mathcal{C}_t &= \mathrm{Tr}\left(\frac{\partial \mathcal{C}_t}{\partial O_t}^\top \frac{\partial O_t}{\partial t}\right) \\ &= \sum_{x\in\mathbb{X}} \mathrm{Tr}\left(\frac{\partial \mathcal{C}_t}{\partial O_{t,x}} \frac{\partial O_{t,x}}{\partial t}^\top\right) = \\ -\tau_{\max} &\sum_{x\in\mathbb{X}} \|\tilde{P}_x^{X\top}(P^Y - P^X O_t)\|_{K_{D,l} K_{G,l} K_{D,l}}^2.\end{aligned}$$

Now, apply Lemma 12, we have

$$\begin{aligned}&\partial_\tau f_{\tau,l}^\top \mathbf{1}_{|\mathbb{Y}|} \\ =&(P_l^Y - P_l^X O_t) K_{D,l} \mathbf{1}_{|\mathbb{Y}|} \\ =&\lambda (P_l^Y - P_l^X O_t)\mathbf{1}_{|\mathbb{Y}|} = 1 - 1 = 0 \\ \Longrightarrow &\mathbf{1}_{|\mathbb{Y}|} \perp K_{D,l}(P_l^Y - P_l^X O_t)^\top,\end{aligned}$$

where $\lambda$ is the eigenvalue of $K_{D,l}$ associated with $\mathbf{1}_{|\mathbb{Y}|}$, and thus

$$K_{D,l}(P^Y - P^X O_t)^\top \tilde{P}_x^X \perp \mathbf{1}_{|\mathbb{Y}|}.$$

As a result, using Lemma 13, we conclude that the kernelized residual vector $\partial_\tau f_{\tau,l}$ is always perpendicular to the null space of the stepwise generator NTK $K_{O_{t,x}}$ for all $1 \le l \le L$, $x \in \mathbb{X}$, and thus

$$\begin{aligned}&\|K_{D,l}(P^Y - P^X O_t)^\top \tilde{P}_x^X\|_{K_{G,l}} \\ \ge&\lambda_G \|K_{D,l}(P^Y - P^X O_t)^\top \tilde{P}_x^X\|_2 \\ \ge&\lambda_G \lambda_D \|P^Y - P^X O_t\|_{K_{D,1}},\end{aligned}$$

where

$$\begin{aligned}\lambda_G &\ge \min_{1\le l\le L} \lambda_{|\mathbb{Y}|-2}(K_{G,l}) > 0, \\ \lambda_D &\ge \lambda_{\min}(K_{D,1}) > 0.\end{aligned}$$

Summing over $x$, we obtain

$$\partial_t \mathcal{C}_t \le -\tau_{\max} \lambda_G \lambda_D \|P^{X\top}(P^Y - P^X O_t)\|_{K_{D,1}}^2.$$

Under the assumption that $P^X O = P^Y$ has at least one solution, we have $P^Y - P^X O$ is in the range space of $P^X$, which implies

$$\begin{aligned}&\|P^{X\top}(P^Y - P^X O_t)\|_{K_{D,1}}^2 \ge \\ &\qquad \lambda_X \|P^Y - P^X O_t\|_{K_{D,1}}^2,\end{aligned}$$

for some $\lambda_X > 0$. Put together the results, we can bound the convergence rate of the generator loss by

$$\begin{aligned}&\partial_t \mathcal{C}_t \le -\tau_{\max} \lambda_G \lambda_D \lambda_X \mathcal{C}_t \\ \Longrightarrow &\mathcal{C}_t \le \mathcal{C}_0 e^{-\tau_{\max}\lambda_G\lambda_D\lambda_X t} \xrightarrow{t\to\infty} 0,\end{aligned}$$

which implies that $\lim_{t\to\infty} P^X O_t = P^Y$.

□

## B Reproducibility checklist

**Synthetic language creation** To create a synthetic HMM language, we need to specify the initial probability vector $\pi$, the transition probability matrix $T$, the generator matrix $O$ and the maximal length of the utterances $L$.

*Initial probability*: we create $\pi$ by first uniformly randomly sampling each coefficient between $[0, 1]$ and then normalizing the resulting vector by its sum.

*Transition probability*: for the asymptotic setting, for all three languages, we control the number of eigenvalues $m$ of its transition matrix using a disjoint union of identical sub-graphs with $m$ eigenvalues, with the remainder of the nodes being self-loops. The parameters and the procedure used to determine them are as follows:

- *Circulant graph*: only undirected cycles or equivalently, circulant graph with the action set $\{-1, 1\}$, are used. Since the distinct eigenvalues of an undirected $n$-cycle $C_n$ are $-\cos\left(\frac{2\pi k}{n}\right), k = 0, \cdots, \lfloor \frac{n-1}{2} \rfloor + 1$, we can create a Markov graph with $|\mathbb{X}|^N$ nodes and $n \pm 1$ eigenvalues by a disjoint union of $\lfloor \frac{|\mathbb{X}|^N}{2n-1} \rfloor$ $C_{2n-1}$ graphs. In our phase transition experiment, we fix $N = 2$ and vary $10 \leq |\mathbb{X}| \leq 14$ and $2 \leq n \leq 20$;

- *De Bruijn graph*: an undirected de Bruijn graph $\mathrm{DB}(k, m)$ is a graph with $k^m$ nodes such that node $i$ connects to any node $j$ whose $k$-ary numerals $v(i)$ and $v(j)$ satisfies $v_{2:m}(i) = v_{1:m-1}(j)$. Clearly, $m$ is the in/out-degree of the graph. The eigenvalues of $\mathrm{DB}(k, m)$ are known to be $\cos\left(\frac{i\pi}{j}\right), 0 \leq i < j \leq m + 1$ (Delorme and Tillich, 1998). Therefore, we can create a Markov graph with $|\mathbb{X}|^N$ nodes and at most $n$, $n \leq (\lfloor \log_k |\mathbb{X}|^N \rfloor + 1)^2/2$ distinct eigenvalues by a disjoint union of $\frac{|\mathbb{X}|^N}{k^{\sqrt{2m}-1}}$ $\mathrm{DB}(k, \sqrt{2n} - 1)$ graphs. For the phase transition experiment, we set the in/out-degree of the de Bruijn subgraphs to be 2 and the $N$-gram size $N = 3$, and we vary $8 \leq |\mathbb{X}| \leq 11$ and $2 \leq n \leq 32$ with a step size of 2 for the latter.

- *Hypercube*: an $n$-cube $Q_n$ is a graph with $2^n$ nodes such that node $i$ connects to any node $j$ with Hamming distance between their binary numerals $d_H(b(i), b(j)) = 1$. The eigenvalues of the adjacency matrix of $Q_n$ is $1 - \frac{2k}{n}, k = 0, \cdots, n$. Therefore, we can create a Markov graph with $|\mathbb{X}|^N$ nodes and $n \leq \lfloor N \log_2 |\mathbb{X}| \rfloor$ eigenvalues by a disjoint union of $\lfloor \frac{|\mathbb{X}|^N}{2^n} \rfloor$ $n$-cubes. For the phase transition experiment, we fix $N = 4$, and vary $5 \leq |\mathbb{X}| \leq 8$ and $2 \leq n \leq 9$.

In the finite-sample setting, we create transition matrices for phase transition experiments using two different setups:

- For the circulant graph, we vary its action set to be $\{1, \cdots, d\}$, where $d$ takes values from 2 to 81 with a step size of 8;

- For the other two graphs, we linearly interpolate between the underlying graph $T_G$ and its Hamiltonian cycle $T_C$ as

$$T = (1 - w)T_G + wT_C, \qquad (32)$$

with a weight $w \in [0, 1]$. In particular, for the de Bruijn graph, the weight for the cycle $w$ takes 10 different values equally spaced between $[0, 1]$; for the $n$-cube, the weight $w$ takes 10 different values equally spaced between $[0.98, 1]$.

*Generator matrix $O$*: set by assuming $|\mathbb{X}| = |\mathbb{Y}|$ and randomly permuting the rows of the $|\mathbb{X}| \times |\mathbb{X}|$ identity matrix.

*Sampling*: in the asymptotic case, no sampling is needed and we simply set maximal length $L = 20$ for cycle graph and 10 for the other two graphs. For the finite-sample case, the synthetic speech and text datasets are created independently by sampling from the same HMM twice. For all three graphs, we sample $n^X = n^Y = 2560$ utterances for both text and speech with $L = 40$ for the de Bruijn graph and $L = 80$ for the other two graphs.

**Model architecture** We use a one-layer linear generator with $|\mathbb{X}|$ input nodes and $|\mathbb{Y}|$ output nodes, with no bias. Next, for all experiments except the experiment on different generator averaging strategies, we use a one-layer CNN with $|\mathbb{Y}|$ input channels, 1 output channel and a $1 \times L$ kernel with no bias. For the experiment on different averaging strategies, we use instead a sequence of 2-layer MLPs with 128 hidden nodes and ReLU activation function, one at each time step, as the discriminators. For all experiments, we disable the logits for special tokens and silences during training and testing.

**Training setting** SGD with a learning rate of 1.0 is used to train the discriminator, while Adam with a learning rate of 0.005 is used to train the generator. The dataset is used as a single batch for all experiments, though we do not observe any significant drop in performance using smaller batch sizes. No weight decays or dropouts is used. Further, we alternatively train the generator and discriminator 1 epoch each, and reset the discriminator weight to 0 for the linear case and to random Gaussian weights using Xavier initialization in the nonlinear case. All experiments are conducted on a single 12GB NVIDIA GeForce GTX 1080Ti GPU.