

Scan and Snap: Understanding Training Dynamics and Token Composition in 1-layer Transformer

Yuandong Tian¹ Yiping Wang^{2,4} Beidi Chen^{1,3} Simon Du²

¹Meta AI (FAIR)

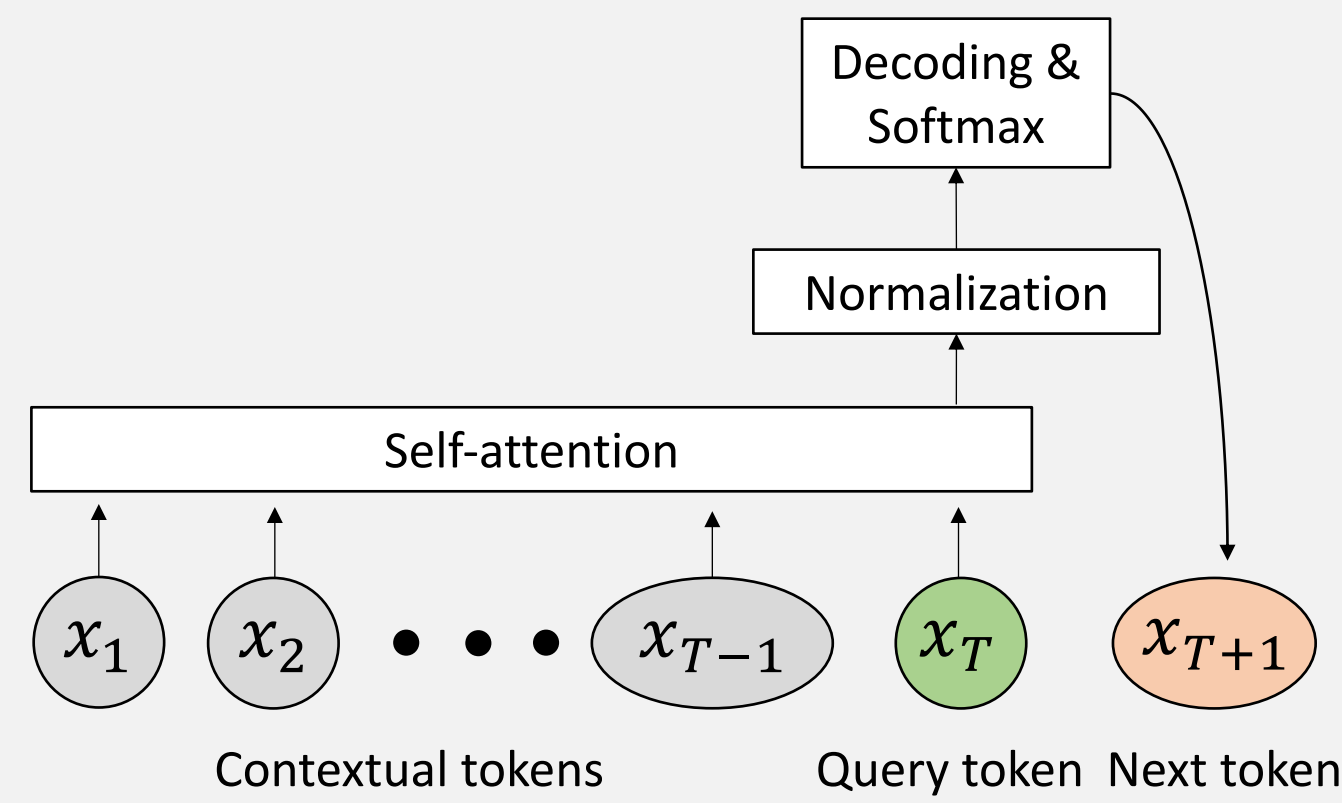
²University of Washington

³Carnegie Mellon University

⁴Zhejiang University



Problem Setting



Notation

$U = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_M]^T$: token embedding matrix

$$\hat{\mathbf{u}}_T = \sum_{t=1}^{T-1} b_{tT} \mathbf{u}_{x_t} = U^T X^T \mathbf{b}_T$$

$$b_{tT} := \frac{\exp(\mathbf{u}_{x_T}^\top W_Q W_K^\top \mathbf{u}_{x_t} / \sqrt{d})}{\sum_{t=1}^{T-1} \exp(\mathbf{u}_{x_T}^\top W_Q W_K^\top \mathbf{u}_{x_t} / \sqrt{d})}$$

Normalized version $\tilde{\mathbf{u}}_T = U^T \text{LN}(X^T \mathbf{b}_T)$

Objective:

$$\max_{W_K, W_Q, W_V, U} J = \mathbb{E}_D \left[\mathbf{u}_{x_{T+1}}^\top W_V \tilde{\mathbf{u}}_T - \log \sum_l \exp(\mathbf{u}_l^\top W_V \tilde{\mathbf{u}}_T) \right]$$

Reparameterization

Reparametrize the problem with independent variable Y and Z

- Decoder $Y = U W_V^T U^T$
- Self-attention $Z = U W_Q W_K^T U^T$

$$Z = \begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix} \mathbf{z}_m$$

$\mathbf{z}_m$: All logits of the contextual tokens when attending to last token $x_T = m$

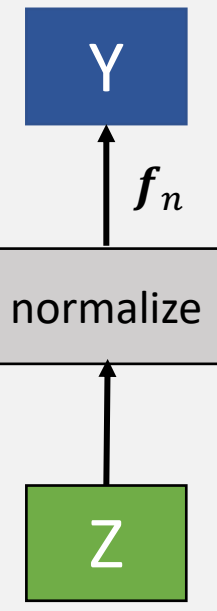
Then the SGD dynamics can be simplified:

$$\dot{Y} = \eta_Y \text{LN}(X^T \mathbf{b}_T)(x_{T+1} - \alpha)^T$$

$$\dot{Z} = \eta_Z x_T (x_{T+1} - \alpha)^T Y^T \frac{P_{X^T \mathbf{b}_T}^\perp}{\|X^T \mathbf{b}_T\|_2} X^T \text{diag}(\mathbf{b}_T) X$$

Here $Z = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M]^T$, each $\mathbf{z}_m \in \mathbb{R}^M$ is the attention score for query/last token m :

$$\dot{\mathbf{z}}_m = \eta_Z X^\top [i] \text{diag}(\mathbf{b}_T [i]) X [i] \frac{P_{X^\top [i] \mathbf{b}_T [i]}^\perp}{\|X^\top [i] \mathbf{b}_T [i]\|_2} Y(x_{T+1} [i] - \alpha [i])$$

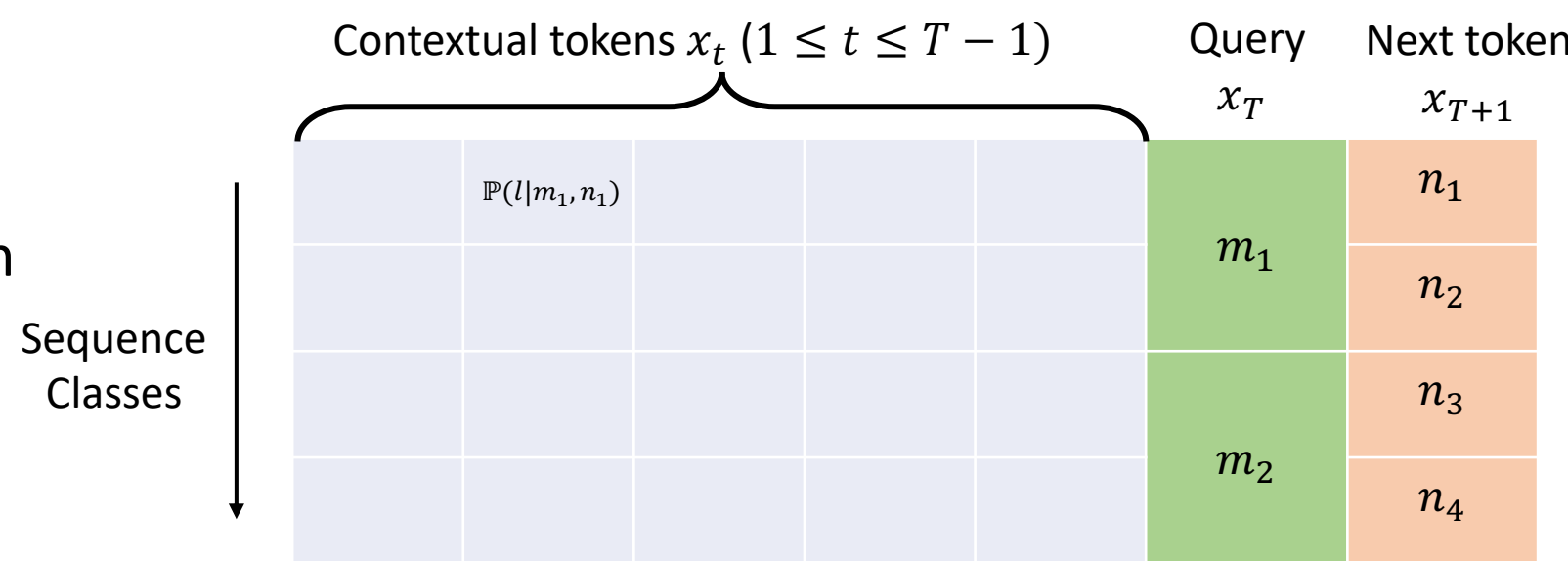


Major Assumptions

- No positional encoding
- Sequence length $T \rightarrow +\infty$
- Learning rate of decoder Y is larger than self-attention layer Z ($\eta_Y \gg \eta_Z$)

For other technical assumptions, please check the paper.

Data Distribution



Assume $m = \psi(n)$, i.e., no next token shared among different last tokens

$\mathbb{P}(l|m, n) = \mathbb{P}(l|n)$ is the conditional probability of token l given last token $x_T = m$ and $x_{T+1} = n$

The power of infinite sequence length $T \rightarrow +\infty$

$$c_{l|m,n} := \frac{T \mathbb{P}(l|m, n) \exp(z_{ml})}{\sum_{l'} T \mathbb{P}(l'|m, n) \exp(z_{ml'})} = \frac{\mathbb{P}(l|m, n) \exp(z_{ml})}{\sum_{l'} \mathbb{P}(l'|m, n) \exp(z_{ml'})} =: \frac{\tilde{c}_{l|m,n}}{\sum_{l'} \tilde{c}_{l'|m,n}}$$

Lemma 2. Given the event $\{x_T = m, x_{T+1} = n\}$, when $T \rightarrow +\infty$, we have

$$X^\top \mathbf{b}_T \rightarrow \mathbf{c}_{m,n}, \quad X^\top \text{diag}(\mathbf{b}_T) X \rightarrow \text{diag}(\mathbf{c}_{m,n})$$

where $\mathbf{c}_{m,n} = [c_{1|m,n}, c_{2|m,n}, \dots, c_{M|m,n}]^\top \in \mathbb{R}^M$. Note that $\mathbf{c}_{m,n}^\top \mathbf{1} = 1$.

Define $\mathbf{f}_n := \mathbf{f}_{m,n} := \mathbf{c}_{m,n} / \|\mathbf{c}_{m,n}\|_2$ a ℓ_2 -normalized version of $\mathbf{c}_{m,n}$.

Dynamics of Decoder Y

Since $\eta_Y \gg \eta_Z$, we analyze the dynamics of decoder Y first, treating the output of Z as constant.

$$\dot{Y} = \eta_Y \mathbf{f}_n (\mathbf{e}_n - \alpha_n)^\top, \quad \alpha_n = \frac{\exp(Y^\top \mathbf{f}_n)}{\mathbf{1}^\top \exp(Y^\top \mathbf{f}_n)}$$

K : number of possible next tokens to be predicted

Theorem 1. If Assumption 2 holds, the initial condition $Y(0) = 0$, $M \gg 100$, η_Y satisfies $M^{-0.99} \ll \eta_Y < 1$, and each sequence class appears uniformly during training, then after $t \gg K^2$ steps of batch size 1 update, given event $x_{T+1}[i] = n$, the backpropagated gradient $\mathbf{g}[i] := Y(x_{T+1}[i] - \alpha[i])$ takes the following form:

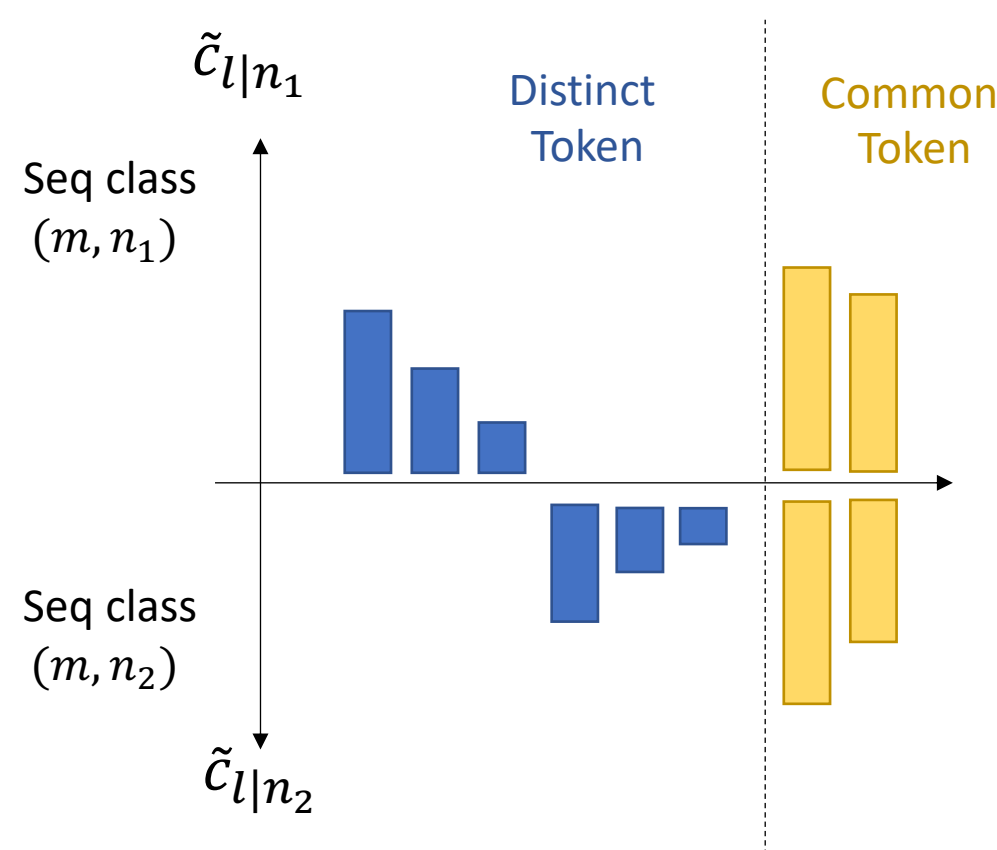
$$\mathbf{g}[i] = \gamma \left(\iota_n \mathbf{f}_n - \sum_{n' \neq n} \beta_{nn'} \mathbf{f}_{n'} \right) \quad (9)$$

Here the coefficients $\iota_n(t)$, $\beta_{nn'}(t)$ and $\gamma(t)$ are defined in Appendix with the following properties:

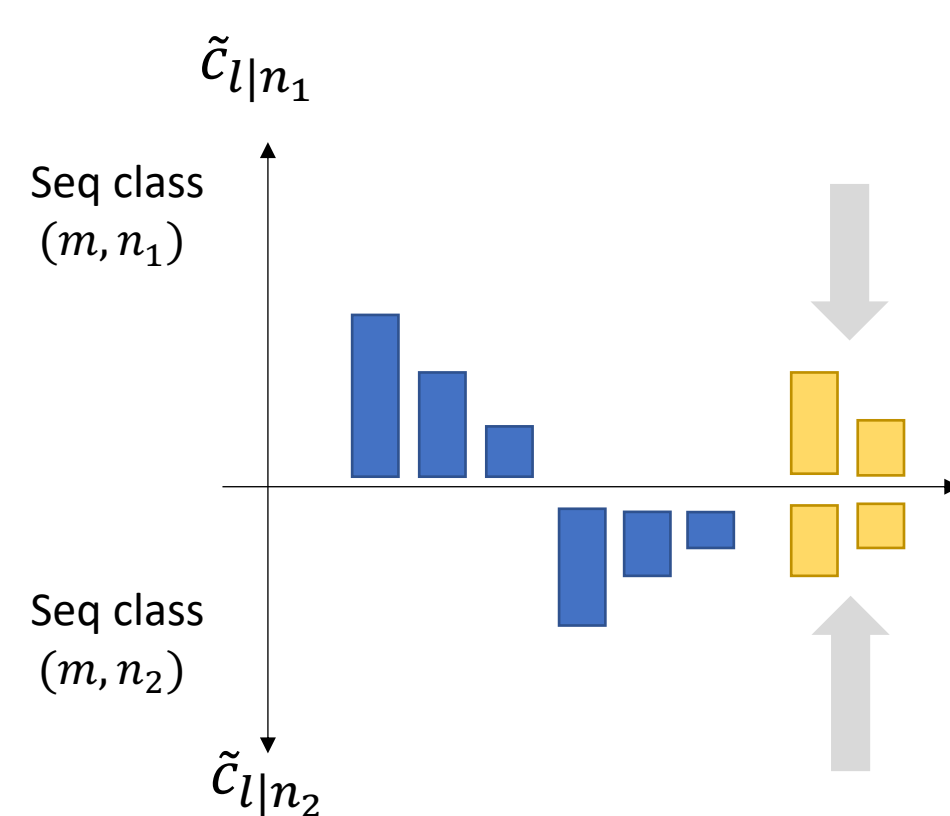
- (a) $\xi_n(t) := \gamma(t) \sum_{n' \neq n} \beta_{nn'}(t) \mathbf{f}_n^\top(t) \mathbf{f}_{n'}(t) > 0$ for any $n \in [K]$ and any t ;
- (b) The speed control coefficient $\gamma(t) > 0$ satisfies $\gamma(t) = O(\eta_Y t / K)$ when $t \leq \frac{\ln(M) \cdot K}{\eta_Y}$ and $\gamma(t) = O\left(\frac{K \ln(\eta_Y t / K)}{\eta_Y t}\right)$ when $t \geq \frac{2(1+\delta') \ln(M) \cdot K}{\eta_Y}$ with $\delta' = \Theta\left(\frac{\ln \ln M}{\ln M}\right)$.

Dynamics of Self-attention Z

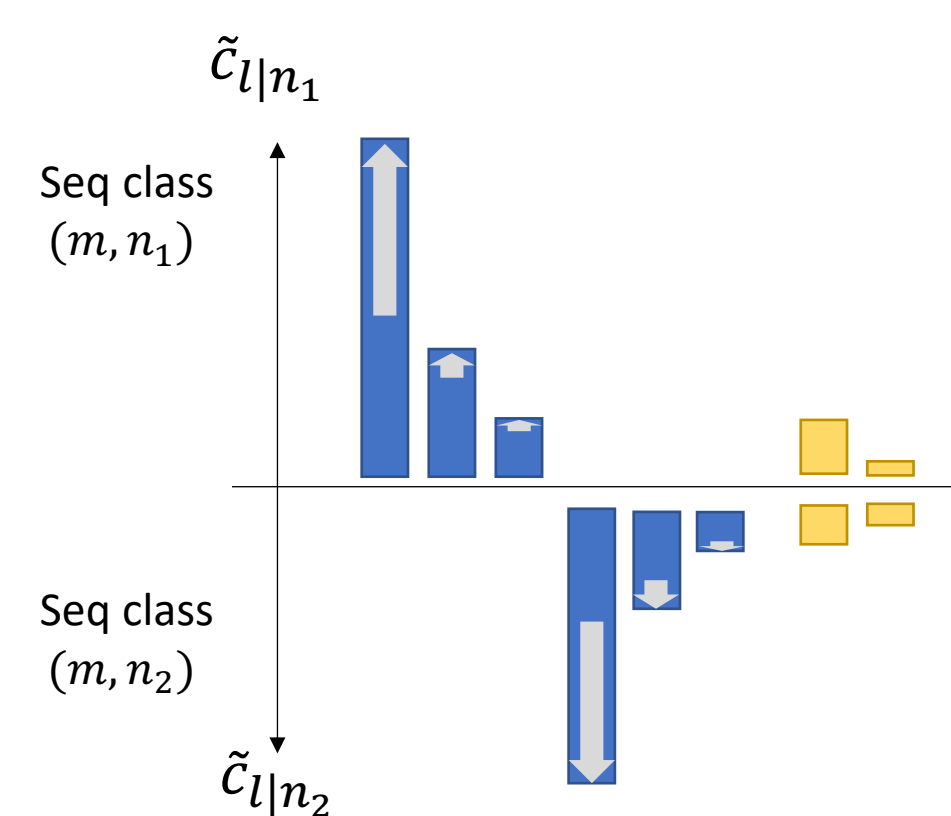
At initialization



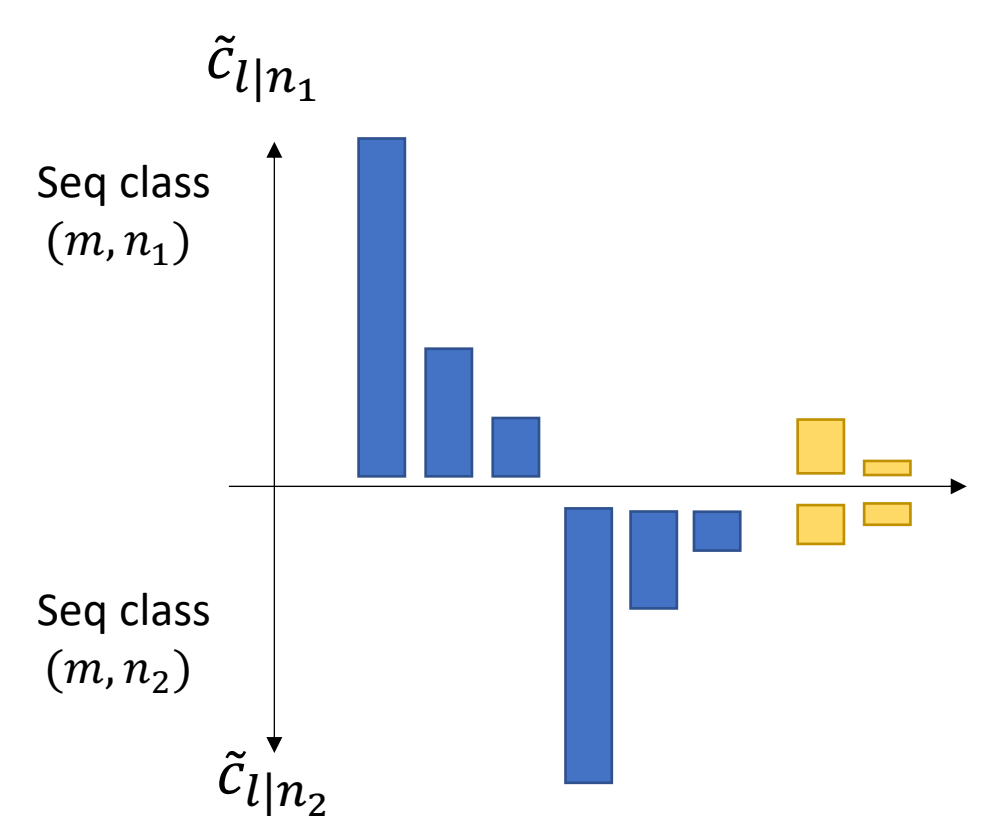
Common Token Suppression



Winners emerge



Attention frozen



- $z_{ml} < 0$, for common token l
- $z_{ml} > 0$, for distinct token l
- $z_{ml}(t)$ grows faster with larger $\mathbb{P}(l|m, n)$

Theorem 3

Relative gain $r_{l/l'|n}(t) := \frac{\tilde{c}_{l|n}^2(t)}{\tilde{c}_{l'|n}^2(t)} - 1$ has a close form:

$$r_{l/l'|n}(t) = r_{l/l'|n}(0) \chi_l(t)$$

If l_0 is the dominant token: $r_{l_0/l|n}(0) > 0$ for all $l \neq l_0$ then

$$e^{2f_{n,l_0}^2(0)B_n(t)} \leq \chi_{l_0}(t) \leq e^{2B_n(t)}$$

where $B_n(t) \geq 0$ is a monotonously increasing function with $B_n(0) = 0$.

Theorem 4. When $t \rightarrow +\infty$,

$$B_n(t) \sim \ln \left(C_0 + 2K \frac{\eta_Z}{\eta_Y} \ln^2 \left(\frac{M \eta_Y t}{K} \right) \right)$$

Attention scanning:

When training starts, $B_n(t) = O(\ln t)$

Attention snapping:

When $t \geq t_0 = O\left(\frac{2K \ln M}{\eta_Y}\right)$, $B_n(t) = O(\ln \ln t)$

- η_Z and η_Y are large, $B_n(t)$ is large and attention is sparse
- Fixing η_Z , large η_Y leads to slightly small $B_n(t)$ and denser attention

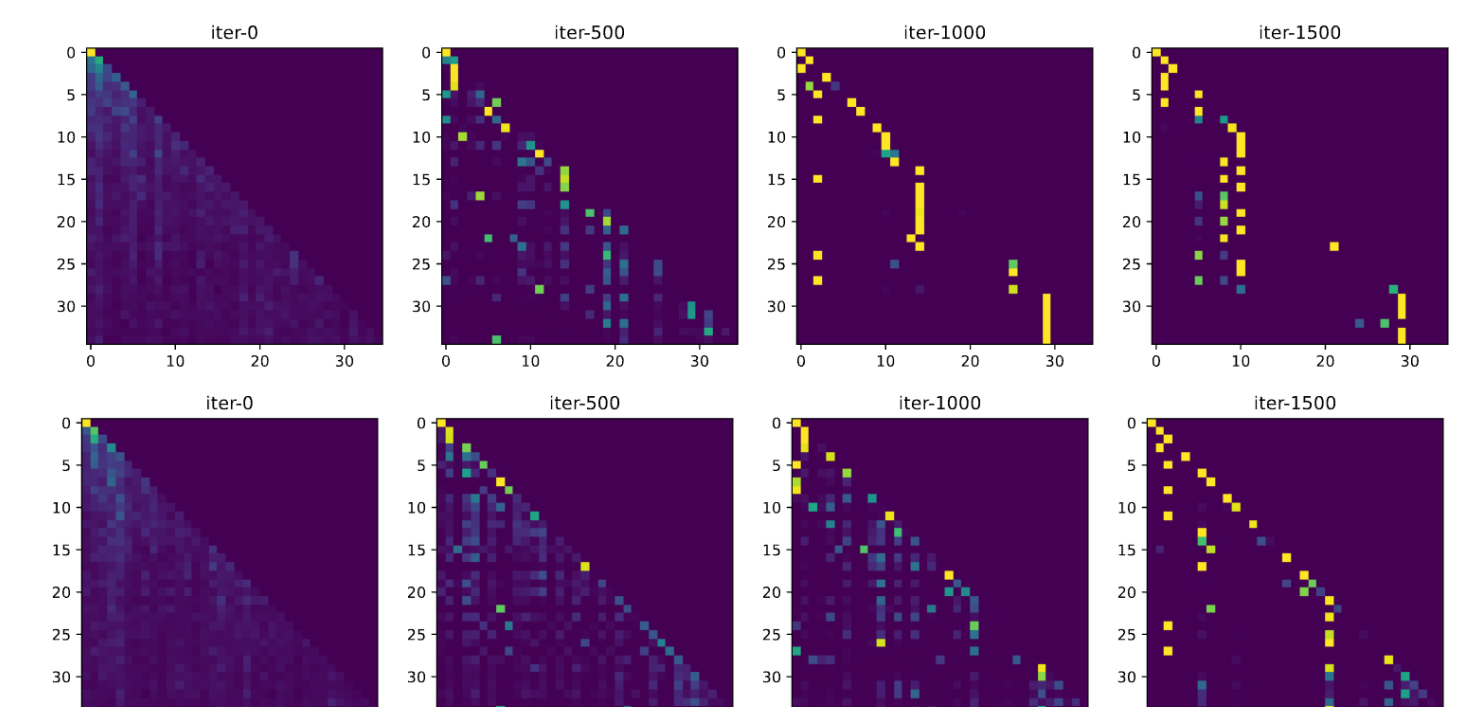
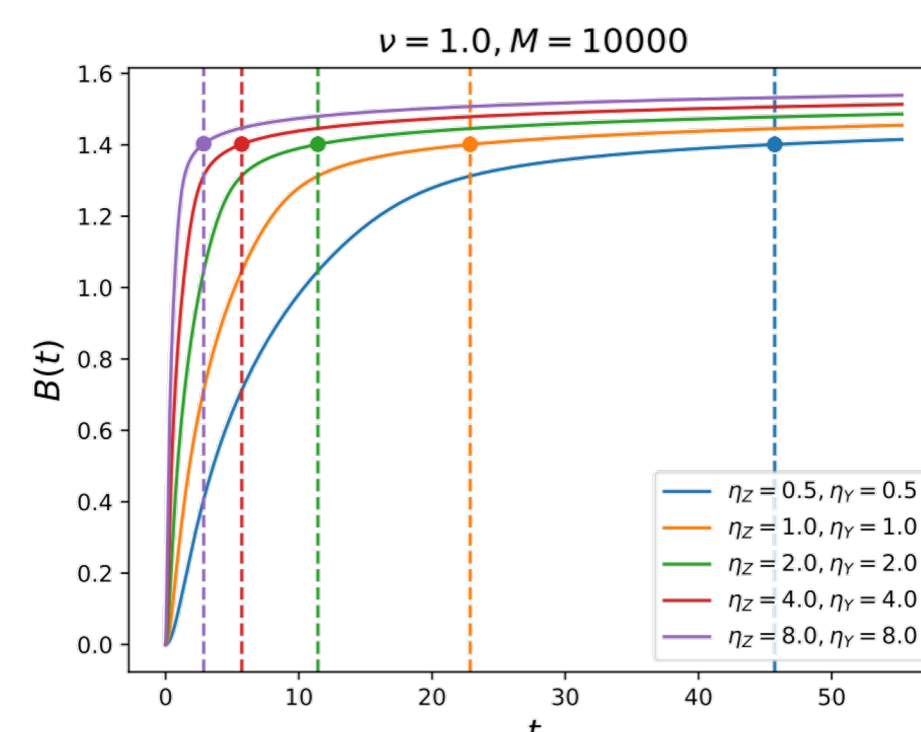


Figure 7: Attention patterns in the lowest self-attention layer for 1-layer (top) and 3-layer (bottom) Transformer trained on WikiText2 using SGD (learning rate is 5). Attention becomes sparse over training.

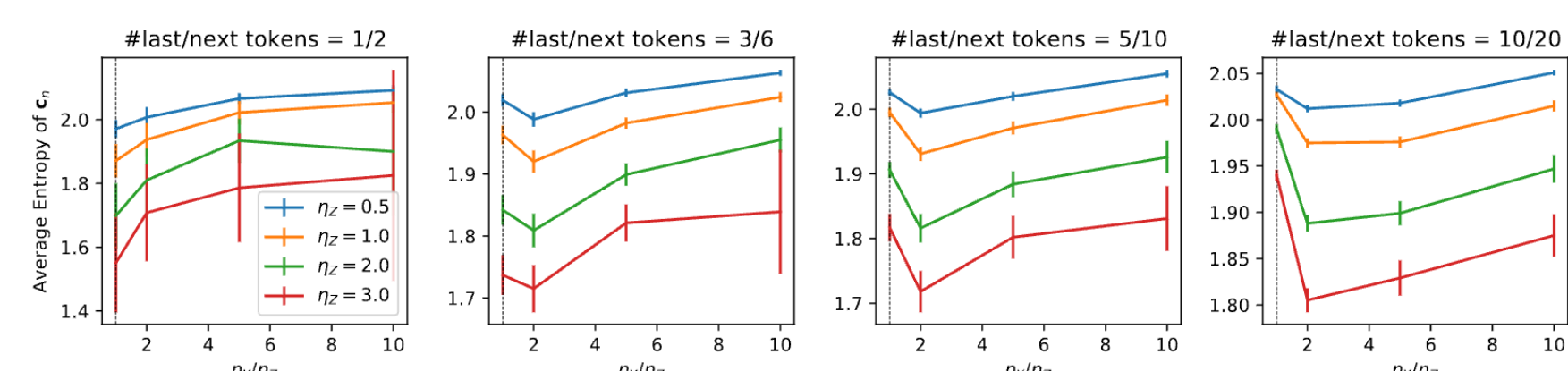


Figure 6: Average entropy of $\mathbf{c}_n$ (Eqn. 5) on distinct tokens versus learning rate ratio η_Y/η_Z with more last tokens M /next tokens K . We report mean values over 10 seeds and standard derivation of the mean.