Softmax as Linear Attention in the Large-Prompt Regime: a Measure-based Perspective

with C. Boyer

Summary:

- Measure-based view of self attention unifying finite and infinite prompt as inputs
- Concentration of the finite prompt regime to the infinite one as $L \to \infty$
- Derive new convergence guarantees for in-context linear regression with softmax attention

I - Measure based view of attention

Classic view

An attention layer takes as inputs a prompt of d-dimensional tokens $(z_1, \dots, z_L) \in \mathbb{R}^{d \times L}$ and a query token $z \in \mathbb{R}^d$ and outputs

$$\frac{\sum_{i=1}^{L} \exp(\langle Kz_i, Qz \rangle) V z_i}{\sum_{i=1}^{L} \exp(\langle Kz_i, Qz \rangle)}$$

this is a softmax operator combined with a matrix multiplication

where K, Q, V are the model parameters (key, query and value matrices)

Measure based view

Output is independent of token order, so that we can see the prompt as a measure $\mu \in \mathcal{M}(\mathbb{R}^d)$ and an attention layer parametrized by K, Q, V is then a function:

$$T^{K,Q,V} : \begin{array}{ccc} \mathcal{M}(\mathbb{R}^d) \times \mathbb{R}^d & \longrightarrow & \mathbb{R}^d \\ (\mu, z) & \longmapsto & T^{K,Q,V}[\mu](z) = \dfrac{\int \exp(\langle Kz', Qz \rangle) V z' \, d\mu(z')}{\int \exp(\langle Kz', Qz \rangle) \, d\mu(z')} \end{array}$$

→ corresponds to classic definition when $\mu = \underbrace{\frac{1}{L}\sum_{\ell=1}^{L} \delta_{z_\ell}}_{:= \hat{\mu}_L}$

If $z_i \overset{iid}{\sim} \mu$. Then $T^{K,Q,V}[\hat{\mu}_L](z) \xrightarrow[L\to\infty]{a.s.} T^{K,Q,V}[\mu](z)$

infinite prompt regime is often easier to analyze.

Example. Gaussian inputs.

Lemma (Castin, et al, 2025)

If $\mu = \mathcal{N}(m, \Gamma)$ then $T^{K,Q,V}[\mu](z) = Vm + V\Gamma K^T Q z$

Proof $T^{K,Q,V}[\mu](z) = V \frac{\int \exp(\langle Kz', Qz\rangle)\, z'\, d\mu(z')}{\int \exp(\langle Kz', Qz\rangle)\, d\mu(z')} = V \mathbb{E}_{z'\sim\nu}[z']$

with ν skewed distribution defined as: $d\nu(z') = \frac{\exp(\langle Kz', Qz\rangle)}{\mathbb{E}_{z\sim\mu}[\exp(\langle Kz, Qz\rangle)]}\, d\mu(z')$

ν is a probability distribution and (assuming Γ non-degenerate),

$$\frac{d\nu(z')}{dz'} \propto \exp\left(z'^T K^T Q z - \frac{1}{2}(z'-m)^T \Gamma^{-1}(z'-m)\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(z'-m-\Gamma K^T Q z\right)^T \Gamma^{-1}\left(z'-m-\Gamma K^T Q z\right)\right) \cdot C(z)$$

$$\Rightarrow \quad \nu = \mathcal{N}(m + \Gamma K^T Q z,\ \Gamma)$$

So $T^{K,Q,V}[\mu](z) = V\mathbb{E}_{z\sim\nu}[z] = Vm + V\Gamma K^T Q z$ □

With Gaussian inputs, softmax attention coincide with (normalized) linear attention in the limit $L \to \infty$ where

$$T^{K,Q,V}_{lin}[\mu](z) = V \int \langle Kz', Qz \rangle z' d\mu(z')$$

$\longrightarrow$ much easier to analyse (eg optimization dynamics for ICL)

But how do they compare for finite L?

II - Concentration of softmax attention

Prop

If μ and ν are resp. σ and 1 sub-Gaussian measures with $\sigma \geq 1$, then there exists constants $c_1, c_2 > 0$ depending solely on d, V, K, Q s.t.

$$\mathbb{E}_{z_1,\dots,z_L \sim \mu^{\otimes L}} \left[\| T^{K,Q,V}[\hat{\mu}_L] - T^{K,Q,V}[\mu] \|^2_{L^2(\nu)} \right] \leq c_1 \sigma^6 \frac{\ln L}{L^{c_2/\sigma^2}}$$

$c_2 = \frac{1}{192 \|QK\|_{op}^2}$

Proof Idea: concentration on both $\frac{1}{L} \sum_{i=1}^{L} \exp(\langle Kz_i, Qz \rangle) Vz_i$

and $\frac{1}{L} \sum_{i=1}^{L} \exp(\langle Kz_i, Qz \rangle)$

but these are heavy tail random variables.

when $\max_\rho \exp(\langle K z_\rho, Q z_\rho \rangle) \leq M$ $\longrightarrow$ exponential tail concentration

when $\max_\rho \exp(\langle K z_\rho, Q z_\rho \rangle) > M$ (low proba) $\longrightarrow$ use that

$$\| T^{K,Q,V}[\hat{\mu}_L](z) \| \leq \max_\rho \|V z_\rho\| \quad \square$$

Our goal: transfer optimization results derived for linear case ($L = \infty$) to finite L and softmax attention

For that, we need concentration on the gradients. Use $U = K^T Q$ parametrization

so that $T^{U,V} = T^{K,Q,V}$ with

$$T^{U,V}[\mu](z) = \frac{\int \exp(z'^T U z) V z' \, d\mu(z')}{\int \exp(z'^T U z) \, d\mu(z')}$$

Under same assumptions + bounded moments of shared distribution:

$$\mathbb{E}_{z_1,\dots,z_L \sim \mu^{\otimes L}} \left[\| \nabla_{U,V} T^{U,V}[\hat{\mu}_L] - \nabla_{U,V} T^{U,V}[\mu] \|_{(U,V)}^2 \right] \leq c_1 \sigma^{42} \frac{\ell_n(L)^2}{L^{2/5}}$$

$\longrightarrow$ training dynamics should be the same for large values of L!

III - In-context linear regression

In-Context Learning (ICL): ability to "learn" task structure directly from the prompt, at inference time

Setting (ICL linear regression)

for each new prompt: draw parameter $w \sim \mathcal{N}(0, I_d)$

and prompt of $(z_1, \dots, z_L)$ where $z_i = (x_i, y_i)$ and $\begin{cases} x_i \sim \mathcal{N}(0, \Sigma) \\ y_i = w^T x_i \end{cases}$

Goal: given a masked query input $(x_q, 0)$, predict $y_q = w^T x_q$.

Note that given w, $z_\rho \sim \mathcal{N}(0, \Gamma_w)$ with $\Gamma_w = \begin{pmatrix} \Sigma & \Sigma w \\ (\Sigma w)^\top & \|w\|_\Sigma^2 \end{pmatrix}$

ICL Risk for $L = \infty$:

$$\mathcal{R}_\infty^{ICL}(U,V) = \frac{1}{2}\,\mathbb{E}_{w\sim\mathcal{N}(0,I_d)}\left[\mathbb{E}_{x_q\sim\mathcal{N}(0,\Sigma)}\left[\left(T^{U,V}[\mu](x_q,0)_{d+1} - w^\top x_q\right)^2\right]\right]$$

$$= \frac{1}{2}\,\mathbb{E}_{w\sim\mathcal{N}(0,I_d)}\left[\mathbb{E}_{x_q\sim\mathcal{N}(0,\Sigma)}\left[\left(v_{d+1}^\top \Gamma_w U(x_q,0)^\top - w^\top x_q\right)^2\right]\right]$$

ICL Risk for $L < \infty$!

$$\mathcal{R}_L^{ICL}(U,V) = \frac{1}{2}\,\mathbb{E}_w\left[\mathbb{E}_{\substack{x_q \\ z_1,\dots,z_L}}\left[\left(T^{U,V}[\hat{\mu}_L](x_q,0)_{d+1} - w^\top x_q\right)^2\right]\right]$$

$$= \frac{1}{2}\,\mathbb{E}_w\,\mathbb{E}_{\substack{x_q \\ z_1,\dots,z_L\sim\mathcal{N}(0,\Gamma_w)}}\left[\left(\frac{\sum_{i=1}^L \exp(z_i^\top U(x_q,0)^\top)\, v_{d+1}^\top z_i}{\sum_{i=1}^L \exp(z_i^\top U(x_q,0)^\top)} - w^\top x_q\right)^2\right]$$

Thm

For a proper init. scheme, then for any $\varepsilon > 0$, $\exists L(\varepsilon)$ s.t. $\forall L \geq L(\varepsilon)$,

$$\lim_{t\to\infty} \mathcal{R}_L^{ICL}(U_L(t), V_L(t)) \leq \mathcal{R}^{ICL,*} + \varepsilon$$

$U_L(t), V_L(t)$: params obtained by gradient flow over $\mathcal{R}_L^{ICL}$

$\mathcal{R}^{ICL,*}$: Bayes risk ($=0$ in absence of label noise)

Idea of proof:

use similarity with the $L = \infty$ case + the result by Zhang et al (2024)

that provides convergence guarantees to Bayes optimum for linear attention.