Attention-Based PCA

Rodrigo Maulén Soto

GdT Transformers, May 2026.

Joint work with Claire Boyer

Outline

1. Introduction and problem statement
2. Softmax attention performs PCA
3. Linear attention also performs PCA
4. ICL Risk with Spiked Wishart Prior

Introduction

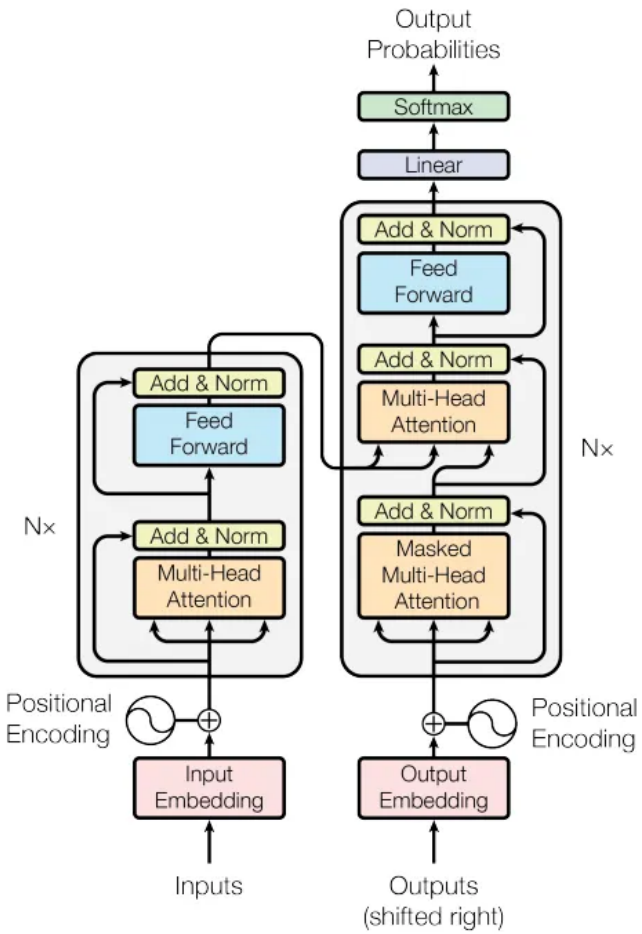


Figure 1: The Transformer - model architecture.

- Attention-based models, in particular Transformers, have achieved state of the art performance across a wide range of learning tasks.

Introduction

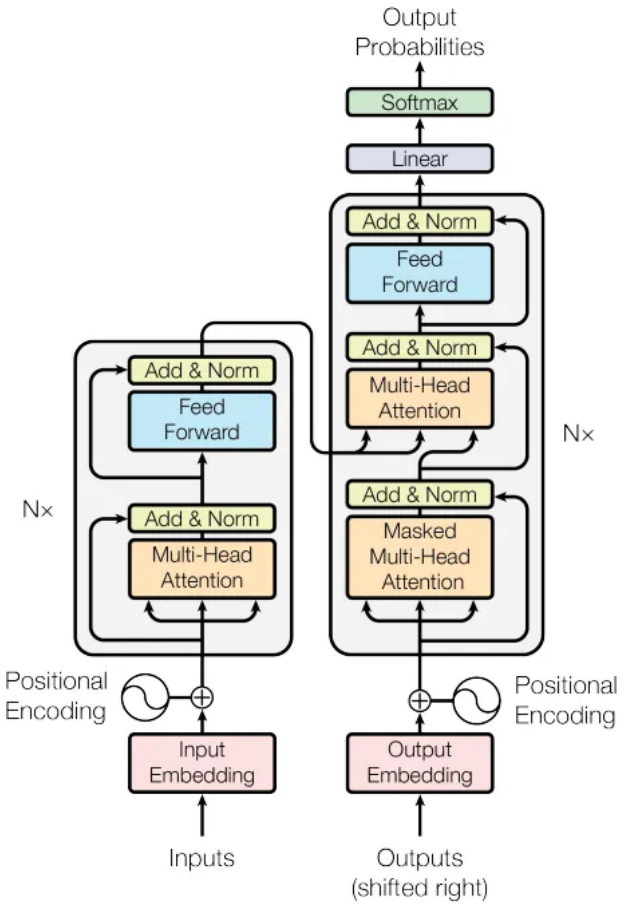


Figure 1: The Transformer - model architecture.

- Attention-based models achieve state-of-the-art performance across numerous tasks.
- Attention formula:

$$Att(Q, K, V) = \mathrm{softmax}_\lambda\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

Introduction

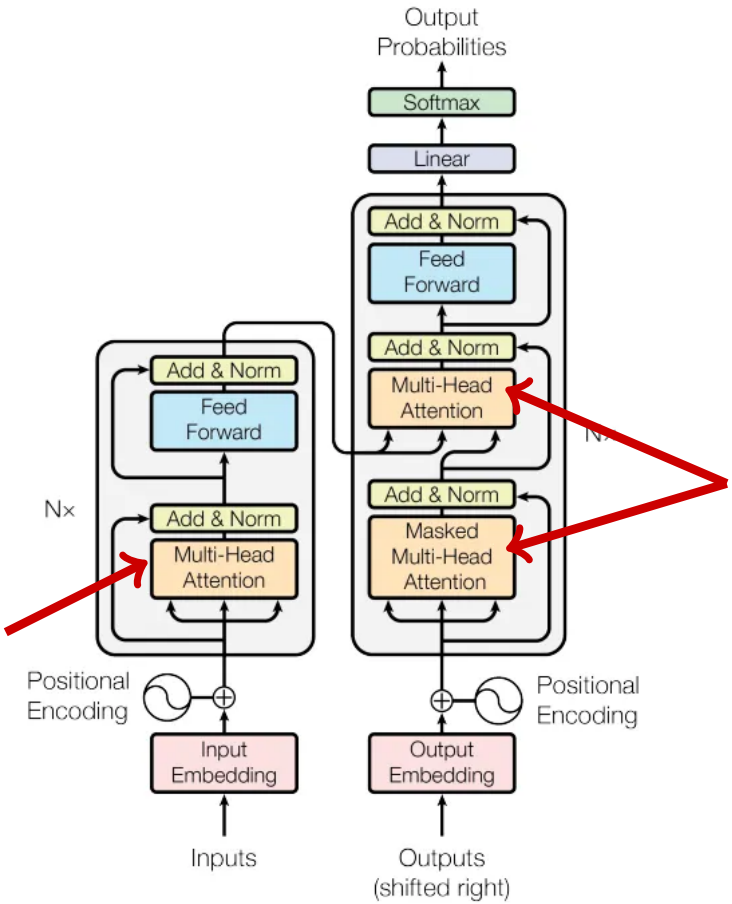


Figure 1: The Transformer - model architecture.

- Attention-based models achieve state-of-the-art performance across numerous tasks.
- Attention formula:

$$Att(Q, K, V) = \mathrm{softmax}_\lambda \left(\frac{QK^T}{\sqrt{d_k}} \right) V.$$

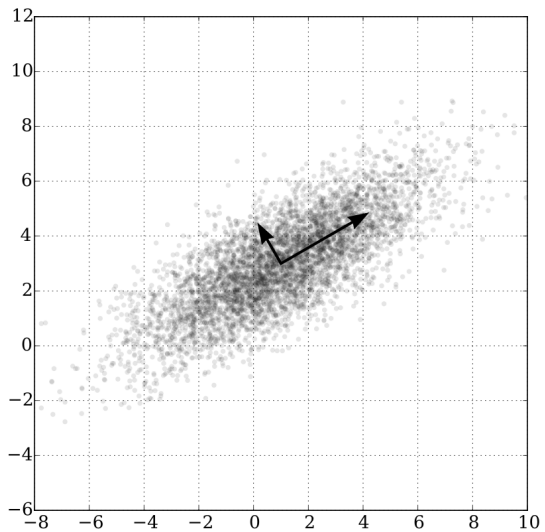
- Theoretical understanding of this mechanism is **not yet** developed.
- We focus on simpler yet representative problems.

Problem statement

- Given i.i.d. tokens drawn from

$$X_\ell \sim \mathcal{N}(0, \Sigma),$$

we consider the PCA problem, i.e., identify the leading eigenpairs (eigenvectors/eigenvalues) of the covariance matrix Σ.



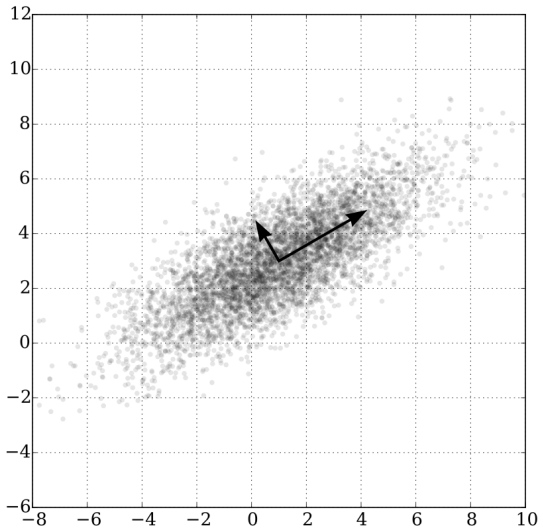
- Clasical tools: Power iteration, SVD, Oja's rule.

Problem statement

- Given i.i.d. tokens drawn from

$$X_\ell \sim \mathcal{N}(0, \Sigma),$$

we consider the PCA problem, which corresponds to identify the leading eigenpairs (eigenvectors/eigenvalues) of the covariance matrix Σ.



- Our goal is to **theoretically prove the retrieval of the leading eigenpair** using simplified attention mechanisms in an unsupervised setting.

Problem statement

- Consider the data matrix $\mathbb{X} = (X_1, \ldots, X_L)^\top \in \mathbb{R}^{L \times d}$, the attention formula becomes

$$Att(\mathbb{X}Q, \mathbb{X}K, \mathbb{X}V) = \mathrm{softmax}_\lambda \left(\mathbb{X}QK^\top \mathbb{X}^\top \right) \mathbb{X}V,$$

  where $Q, K, V \in \mathbb{R}^{d \times p}$ are called query, key and value matrices, and the $\mathrm{softmax}$ is applied row-wise.

Problem statement

- Consider the data matrix $\mathbb{X} = (X_1, \ldots, X_L)^\top \in \mathbb{R}^{L \times d}$, the attention formula becomes

$$Att(\mathbb{X}Q, \mathbb{X}K, \mathbb{X}V) = \mathrm{softmax}_\lambda \left(\mathbb{X}QK^\top \mathbb{X}^\top \right) \mathbb{X}V,$$

  where $Q, K, V \in \mathbb{R}^{d \times p}$ are called query, key and value matrices, and the $\mathrm{softmax}$ is applied row-wise.

- Our proposed simplified attention head drops the value matrix and consider $p = 1$, with $K = Q = \mu \in \mathbb{R}^d$, therefore

$$T_L^\mu(\mathbb{X})_\ell = \sum_{k=1}^{L} \mathrm{softmax}(\lambda X_\ell^\top \mu\mu^\top X_k) X_k.$$

Problem statement

- Consider the data matrix $\mathbb{X} = (X_1, \ldots, X_L)^\top \in \mathbb{R}^{L \times d}$, the attention formula becomes

$$Att(\mathbb{X}Q, \mathbb{X}K, \mathbb{X}V) = \mathrm{softmax}_\lambda \left(\mathbb{X}QK^\top \mathbb{X}^\top \right) \mathbb{X}V,$$

  where $Q, K, V \in \mathbb{R}^{d \times p}$ are called query, key and value matrices, and the $\mathrm{softmax}$ is applied row-wise.

- Our proposed simplified attention head drops the value matrix and consider $p = 1$, with $K = Q = \mu \in \mathbb{R}^d$, therefore

$$T_L^\mu(\mathbb{X})_\ell = \sum_{k=1}^{L} \mathrm{softmax}(\lambda X_\ell^\top \mu\mu^\top X_k) X_k.$$

- **Goal to keep in mind:** Define risk $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$, s.t. $\mu^\star = \min_\mu \mathcal{R}(T_L^\mu)$ aligns with the principal eigenvector of Σ.

Measure-based formalism

$$T_L^{\mu}(\mathbb{X})_\ell = \sum_{k=1}^{L} \mathrm{softmax}(\lambda X_\ell^\top \mu\mu^\top X_k) X_k.$$

Measure-based formalism

$$T_L^{\mu}(\mathbb{X})_\ell = \sum_{k=1}^{L} \mathrm{softmax}(\lambda X_\ell^\top \mu\mu^\top X_k) X_k.$$

A self-attention layer with attention parameter μ can be seen as an operator acting on measures:

$$T^{\lambda,\mu} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d,$$

$$(\nu, z) \mapsto T^{\lambda,\mu}[\nu](z) = \frac{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') z' \, d\nu(z')}{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') \, d\nu(z')}.$$

Measure-based formalism

$$T_L^{\mu}(\mathbb{X})_\ell = \sum_{k=1}^{L} \mathrm{softmax}(\lambda X_\ell^\top \mu\mu^\top X_k) X_k.$$

A self-attention layer with attention parameter μ can be seen as an operator acting on measures:

$$T^{\lambda,\mu} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d,$$

$$(\nu, z) \mapsto T^{\lambda,\mu}[\nu](z) = \frac{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') z' \, d\nu(z')}{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') \, d\nu(z')}.$$

When the prompt $\mathbb{X} = (X_1, \ldots, X_L)^\top$ is encoded by its empirical measure

$$\hat{\nu}_L = \frac{1}{L} \sum_{\ell=1}^{L} \delta_{X_\ell},$$

one retrieves the softmax attention formula, i.e.,

$$T^{\lambda,\mu}[\hat{\nu}_L](X_\ell) = T_L^{\mu}(\mathbb{X})_\ell.$$

Measure-based formalism

$$T^{\lambda,\mu} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d,$$

$$(\nu, z) \mapsto T^{\lambda,\mu}[\nu](z) = \frac{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') z' \, d\nu(z')}{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') \, d\nu(z')},$$

Measure-based formalism

$$T^{\lambda,\mu} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d,$$

$$(\nu, z) \mapsto T^{\lambda,\mu}[\nu](z) = \frac{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') z' \, d\nu(z')}{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') \, d\nu(z')},$$

Let $\hat{\nu}_L = \frac{1}{L}\sum_{\ell=1}^{L} \delta_{X_\ell}$, when the prompt length grows, the attention operator converges to its infinite-prompt counterpart, i.e.,

$$T^{\lambda,\mu}[\hat{\nu}_L](z) \xrightarrow[L\to\infty]{a.s.} T^{\lambda,\mu}[\mathcal{N}(0,\Sigma)](z),$$

Measure-based formalism

$$T^{\lambda,\mu} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d,$$

$$(\nu, z) \mapsto T^{\lambda,\mu}[\nu](z) = \frac{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') z' \, d\nu(z')}{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') \, d\nu(z')},$$

Let $\hat{\nu}_L = \frac{1}{L}\sum_{\ell=1}^{L} \delta_{X_\ell}$, when the prompt length grows, the attention operator converges to its infinite-prompt counterpart, i.e.,

$$T^{\lambda,\mu}[\hat{\nu}_L](z) \xrightarrow[L\to\infty]{a.s.} T^{\lambda,\mu}[\mathcal{N}(0,\Sigma)](z),$$

We can show that $T^{\lambda,\mu}[\mathcal{N}(0,\Sigma)](z) = \lambda\Sigma\mu\mu^\top z$.

Measure-based formalism

$$T^{\lambda,\mu} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d,$$

$$(\nu, z) \mapsto T^{\lambda,\mu}[\nu](z) = \frac{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') z' \, d\nu(z')}{\int_{\mathbb{R}^d} \exp(\lambda z^\top \mu\mu^\top z') \, d\nu(z')},$$

Let $\hat{\nu}_L = \frac{1}{L}\sum_{\ell=1}^{L} \delta_{X_\ell}$, when the prompt length grows, the attention operator converges to its infinite-prompt counterpart, i.e.,

$$T^{\lambda,\mu}[\hat{\nu}_L](z) \xrightarrow[L\to\infty]{a.s.} T^{\lambda,\mu}[\mathcal{N}(0,\Sigma)](z),$$

We can show that $T^{\lambda,\mu}[\mathcal{N}(0,\Sigma)](z) = \lambda\Sigma\mu\mu^\top z$. Thus

$$T_L^{\mu}(\mathbb{X})_\ell \xrightarrow[L\to\infty]{a.s.} T_\infty^{\mu}(X_\ell) := \lambda\Sigma\mu\mu^\top X_\ell.$$

Outline

1. Introduction and problem statement
2. Softmax attention performs PCA
3. Linear attention also performs PCA
4. ICL Risk with Spiked Wishart Prior

Risk functions

We study the minimization of the two following risk functions:

Finite-prompt risk.

$$\mathcal{R}_L(\mu) = \mathbb{E}\big[\|X_1 - T_L^\mu(\mathbb{X})_1\|^2\big],$$
$$T_L^\mu(\mathbb{X})_1 = \sum_{k=1}^{L} \mathrm{softmax}\big(\lambda X_1^\top \mu\mu^\top X_k\big)\, X_k.$$

Infinite-prompt risk.

$$\mathcal{R}_\infty(\mu) = \mathbb{E}\big[\|X_1 - T_\infty^\mu(X_1)\|^2\big],$$
$$T_\infty^\mu(X_1) = \lambda\, \Sigma\mu\mu^\top X_1.$$

Risk functions

We study the minimization of the two following risk functions:

Finite-prompt risk.

$$\mathcal{R}_L(\mu) = \mathbb{E}\big[\|X_1 - T_L^\mu(\mathbb{X})_1\|^2\big],$$
$$T_L^\mu(\mathbb{X})_1 = \sum_{k=1}^{L} \mathrm{softmax}\big(\lambda X_1^\top \mu\mu^\top X_k\big)\, X_k.$$

Infinite-prompt risk.

$$\mathcal{R}_\infty(\mu) = \mathbb{E}\big[\|X_1 - T_\infty^\mu(X_1)\|^2\big],$$
$$T_\infty^\mu(X_1) = \lambda\, \Sigma\mu\mu^\top X_1.$$

Strategy.

- $\mathcal{R}_\infty$ admits a closed-form expression, derive analysis here (μ aligns with the principal eigenvector).
- $\mathcal{R}_L$ is an empirical approximation of $\mathcal{R}_\infty$.
- Transfer results from $\mathcal{R}_\infty$ to $\mathcal{R}_L$.

Landscape of infinite prompt risk

$$\min_{\mu \in \mathbb{R}^d} \mathcal{R}_\infty(\mu) = \mathbb{E}\left[\|X_1 - \lambda \Sigma \mu \mu^\top X_1\|^2\right],$$

we have that

$$\mathbb{E}\left[\|X_1 - \lambda \Sigma \mu \mu^\top X_1\|^2\right] = \operatorname{tr}(\Sigma) - 2\lambda(\mu^\top \Sigma^2 \mu) + \lambda^2 (\mu^\top \Sigma \mu)(\mu^\top \Sigma^2 \mu).$$

Proposition

Let $(\sigma_i, u_i)_{i=1}^d$ be the eigenpairs of Σ such that $\sigma_1 > \ldots > \sigma_d$. Then

- $\operatorname{crit}(\mathcal{R}_\infty) = \{0\} \cup \{\pm \alpha_i u_i\}_{i=1,\ldots,d}$ *for some* $\alpha_i > 0$.

Landscape of infinite prompt risk

$$\min_{\mu \in \mathbb{R}^d} \mathcal{R}_\infty(\mu) = \mathbb{E}\left[\|X_1 - \lambda \Sigma \mu \mu^\top X_1\|^2\right],$$

we have that

$$\mathbb{E}\left[\|X_1 - \lambda \Sigma \mu \mu^\top X_1\|^2\right] = \operatorname{tr}(\Sigma) - 2\lambda(\mu^\top \Sigma^2 \mu) + \lambda^2 (\mu^\top \Sigma \mu)(\mu^\top \Sigma^2 \mu).$$

Proposition

Let $(\sigma_i, u_i)_{i=1}^d$ be the eigenpairs of Σ such that $\sigma_1 > \ldots > \sigma_d$. Then

- $\operatorname{crit}(\mathcal{R}_\infty) = \{0\} \cup \{\pm \alpha_i u_i\}_{i=1,\ldots,d}$ *for some* $\alpha_i > 0$.
- 0 *is a local maximum.*
- $\pm \alpha_i u_i$ *for* $i = 2, \ldots, d$ *is a strict saddle point.*
- $\pm \alpha_1 u_1$ *is a local minimum.*

Landscape of infinite prompt risk

$$\min_{\mu \in \mathbb{R}^d} \mathcal{R}_\infty(\mu) = \mathbb{E}\left[\|X_1 - \lambda \Sigma \mu \mu^\top X_1\|^2\right],$$

we have that

$$\mathbb{E}\left[\|X_1 - \lambda \Sigma \mu \mu^\top X_1\|^2\right] = \operatorname{tr}(\Sigma) - 2\lambda(\mu^\top \Sigma^2 \mu) + \lambda^2 (\mu^\top \Sigma \mu)(\mu^\top \Sigma^2 \mu).$$

Proposition

Let $(\sigma_i, u_i)_{i=1}^d$ be the eigenpairs of Σ such that $\sigma_1 > \ldots > \sigma_d$. Then

- $\operatorname{crit}(\mathcal{R}_\infty) = \{0\} \cup \{\pm \alpha_i u_i\}_{i=1,\ldots,d}$ *for some* $\alpha_i > 0$.
- 0 *is a local maximum.*
- $\pm \alpha_i u_i$ *for* $i = 2, \ldots, d$ *is a strict saddle point.*
- $\pm \alpha_1 u_1$ *is a local minimum.*
- *Coercivity implies* $\pm \alpha_1 u_1$ *is a global minimum.*

Landscape of infinite prompt risk

$$\min_{\mu \in \mathbb{R}^d} \mathcal{R}_\infty(\mu) = \mathbb{E}\left[\|X_1 - \lambda \Sigma \mu \mu^\top X_1\|^2\right],$$

we have that

$$\mathbb{E}\left[\|X_1 - \lambda \Sigma \mu \mu^\top X_1\|^2\right] = \operatorname{tr}(\Sigma) - 2\lambda(\mu^\top \Sigma^2 \mu) + \lambda^2(\mu^\top \Sigma \mu)(\mu^\top \Sigma^2 \mu).$$

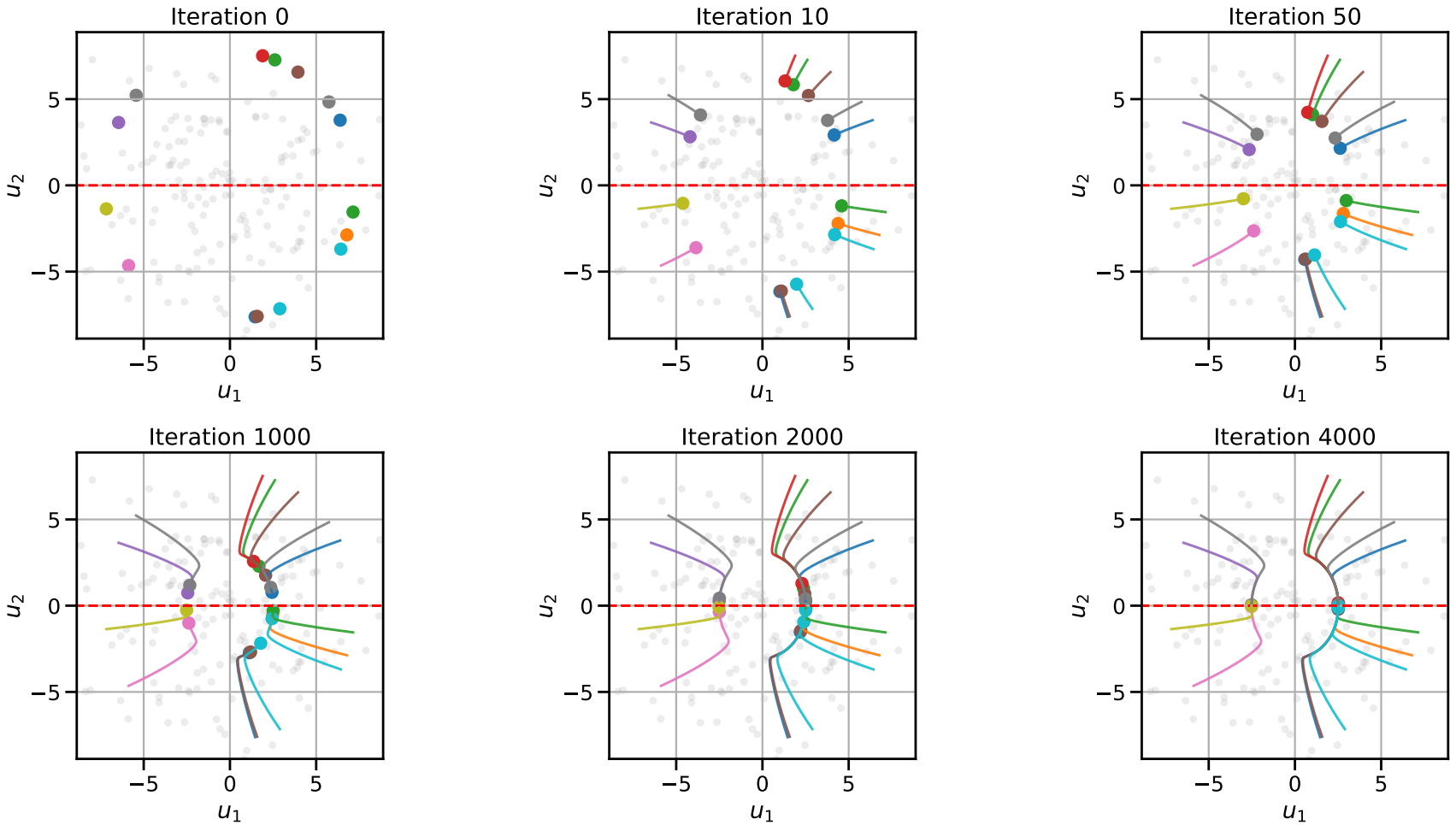
Proposition

Let $(\sigma_i, u_i)_{i=1}^d$ be the eigenpairs of Σ such that $\sigma_1 > \ldots > \sigma_d$. Then

- $\operatorname{crit}(\mathcal{R}_\infty) = \{0\} \cup \{\pm \alpha_i u_i\}_{i=1,\ldots,d}$ *for some* $\alpha_i > 0$.
- 0 *is a local maximum.*
- $\pm \alpha_i u_i$ *for* $i = 2, \ldots, d$ *is a strict saddle point.*
- $\pm \alpha_1 u_1$ *is a local minimum.*
- *Coercivity implies* $\pm \alpha_1 u_1$ *is a global minimum.*
- *Gradient flow on* $\mathcal{R}_\infty$ *with generic initalization converges to its global minimum.*

Visualization

Gradient descent dynamics projected onto the PCA plane



Brief detour: Connection to Oja's flow

Gradient flow on
$\mathcal{R}_\infty(\mu) = \mathrm{tr}(\Sigma) - 2\lambda(\mu^\top \Sigma^2 \mu) + \lambda^2(\mu^\top \Sigma \mu)(\mu^\top \Sigma^2 \mu)$, give us

$$\dot{\mu} = 4\lambda\Sigma^2\mu - 2\lambda^2[(\mu^\top \Sigma^2 \mu)\Sigma\mu + (\mu^\top \Sigma \mu)\Sigma^2\mu].$$

Brief detour: Connection to Oja's flow

Gradient flow on
$\mathcal{R}_\infty(\mu) = \mathrm{tr}(\Sigma) - 2\lambda(\mu^\top \Sigma^2 \mu) + \lambda^2(\mu^\top \Sigma \mu)(\mu^\top \Sigma^2 \mu)$, give us

$$\dot{\mu} = 4\lambda\Sigma^2\mu - 2\lambda^2[(\mu^\top \Sigma^2 \mu)\Sigma\mu + (\mu^\top \Sigma \mu)\Sigma^2\mu].$$

Introducing the change of variables $w = \Sigma^{\frac{1}{2}}\mu$, we get

$$\dot{w} = \Sigma\big[A(w)\,\Sigma w - B\,(w^\top \Sigma w)\,w\big], \quad A(w) := 2\lambda(2 - \lambda\, w^\top w),\ B := 2\lambda^2.$$

Brief detour: Connection to Oja's flow

Gradient flow on
$\mathcal{R}_\infty(\mu) = \mathrm{tr}(\Sigma) - 2\lambda(\mu^\top \Sigma^2 \mu) + \lambda^2(\mu^\top \Sigma \mu)(\mu^\top \Sigma^2 \mu)$, give us

$$\dot{\mu} = 4\lambda\Sigma^2\mu - 2\lambda^2[(\mu^\top \Sigma^2 \mu)\Sigma\mu + (\mu^\top \Sigma \mu)\Sigma^2\mu].$$

Introducing the change of variables $w = \Sigma^{\frac{1}{2}}\mu$, we get

$$\dot{w} = \Sigma\big[A(w)\,\Sigma w - B\,(w^\top \Sigma w)\,w\big], \quad A(w) := 2\lambda(2 - \lambda\, w^\top w),\ B := 2\lambda^2.$$

It can be viewed as a variant of Oja's flow,

$$\dot{w} = \Sigma w - (w^\top \Sigma w)\, w,$$

known continuous-time model for extracting the principal eigenvector of Σ. Its discrete counterpart, called Oja's rule, is a stochastic online algorithm to do so.

Concentration bounds: finite prompt vs infinite prompt

Proposition

We have that for $k \in \{0, 1, 2\}$,

$$\sup_{\mu \in B(0,\rho)} \mathbb{E}\left[\|D_\mu^k T_L^\mu(\mathbb{X})_1 - D_\mu^k T_\infty^\mu(X_1)\|_F^2\right] = \mathcal{O}(\psi(L)),$$

with $\psi(L) = L^{-\xi}(1 + \ln L)^{1-\xi}$, *for some* $0 < \xi < 1$.

Concentration bounds: finite prompt vs infinite prompt

Proposition

We have that for $k \in \{0, 1, 2\}$,

$$\sup_{\mu \in B(0,\rho)} \mathbb{E}\Big[\|D_\mu^k T_L^\mu(\mathbb{X})_1 - D_\mu^k T_\infty^\mu(X_1)\|_F^2\Big] = \mathcal{O}(\psi(L)),$$

with $\psi(L) = L^{-\xi}(1 + \ln L)^{1-\xi}$, *for some* $0 < \xi < 1$. *Then for* $k \in \{0, 1, 2\}$,

$$\sup_{\mu \in B(0,\rho)} \|\nabla^k \mathcal{R}_L(\mu) - \nabla^k \mathcal{R}_\infty(\mu)\|_F^2 = \mathcal{O}(\psi(L)).$$

In particular for $k \in \{0, 1, 2\}$,

$$\nabla^k \mathcal{R}_L \underset{L\to\infty}{\to} \nabla^k \mathcal{R}_\infty \quad \textit{uniformly on } B(0, \rho).$$

Landscape of finite prompt risk

$$\min_{\mu \in \mathbb{R}^d} \mathcal{R}_L(\mu) = \mathbb{E}\big[\|X_1 - T_L^{\mu}(\mathbb{X}_1)\|^2\big],$$

where $T_L^{\mu}(\mathbb{X}_1) = \sum_{k=1}^{L} \mathrm{softmax}(\lambda X_1^\top \mu\mu^\top X_k) X_k$.

Proposition

Let $(\sigma_i, u_i)_{i=1}^d$ be the eigenpairs of Σ such that $\sigma_1 > \ldots > \sigma_d$. Then for $\rho > 0$ and $L > 0$ large enough

- $\mathrm{crit}(\mathcal{R}_L) \cap B(0, \rho) = \{\mu^\star_{L,0}\} \cup \{\pm \mu^\star_{L,\sigma_i}\}_{i=1,\ldots,d}$.

Landscape of finite prompt risk

$$\min_{\mu \in \mathbb{R}^d} \mathcal{R}_L(\mu) = \mathbb{E}\left[\|X_1 - T_L^{\mu}(\mathbb{X}_1)\|^2\right],$$

where $T_L^{\mu}(\mathbb{X}_1) = \sum_{k=1}^{L} \mathrm{softmax}(\lambda X_1^\top \mu\mu^\top X_k) X_k$.

Proposition

Let $(\sigma_i, u_i)_{i=1}^d$ be the eigenpairs of Σ such that $\sigma_1 > \ldots > \sigma_d$. Then for $\rho > 0$ and $L > 0$ large enough

- $\mathrm{crit}(\mathcal{R}_L) \cap B(0, \rho) = \{\mu^\star_{L,0}\} \cup \{\pm\mu^\star_{L,\sigma_i}\}_{i=1,\ldots,d}$.
- *$\mu^\star_{L,0}$ is a local maximum, and $\mu^\star_{L,0} \underset{L\to\infty}{\to} 0$.*
- *$\pm\mu^\star_{L,\sigma_i}$ for $i = 2, \ldots, d$ is a strict saddle point, and $\mu^\star_{L,\sigma_i} \underset{L\to\infty}{\to} \alpha_i u_i$ for some $\alpha_i > 0$.*
- *$\pm\mu^\star_{L,\sigma_1}$ is a local minimum, and $\mu^\star_{L,\sigma_1} \underset{L\to\infty}{\to} \alpha_1 u_1$.*

Landscape of finite prompt risk

$$\min_{\mu \in \mathbb{R}^d} \mathcal{R}_L(\mu) = \mathbb{E}\big[\|X_1 - T_L^\mu(\mathbb{X}_1)\|^2\big],$$

where $T_L^\mu(\mathbb{X}_1) = \sum_{k=1}^{L} \mathrm{softmax}(\lambda X_1^\top \mu\mu^\top X_k) X_k$.

Proposition

Let $(\sigma_i, u_i)_{i=1}^d$ be the eigenpairs of Σ such that $\sigma_1 > \ldots > \sigma_d$. Then for $\rho > 0$ and $L > 0$ large enough

- $\mathrm{crit}(\mathcal{R}_L) \cap B(0,\rho) = \{\mu^\star_{L,0}\} \cup \{\pm\mu^\star_{L,\sigma_i}\}_{i=1,\ldots,d}$.
- *$\mu^\star_{L,0}$ is a local maximum, and $\mu^\star_{L,0} \underset{L\to\infty}{\to} 0$.*
- *$\pm\mu^\star_{L,\sigma_i}$ for $i = 2, \ldots, d$ is a strict saddle point, and $\mu^\star_{L,\sigma_i} \underset{L\to\infty}{\to} \alpha_i u_i$ for some $\alpha_i > 0$.*
- *$\pm\mu^\star_{L,\sigma_1}$ is a local minimum, and $\mu^\star_{L,\sigma_1} \underset{L\to\infty}{\to} \alpha_1 u_1$.*
- *Gradient flow on $\mathcal{R}_L$ with generic initalization on $B(0,\rho)$ converges to $\pm\mu^\star_{L,\sigma_1}$, which asymptotically aligns with u_1.*

Sketch of the proof

Lemma (Persistence of nondegenerate critical points)

Let $f \in C^2(\mathbb{R}^d)$ and let $x^\star$ be a nondegenerate critical point. Assume $f_n \to f$ in C^2_{loc}. Then for large enough n:

1. *There exists a unique critical point $x_n^\star$ of f_n near $x^\star$.*
2. *$x_n^\star \to x^\star$ and $x_n^\star$ is nondegenerate.*
3. *If $\mathrm{crit}(f) = \{x^{(1)}, \dots, x^{(\Lambda)}\}$ are all nondegenerate, then for ρ, n large enough,*

$$\mathrm{crit}(f_n) \cap B(0,\rho) = \{x_n^{(1)}, \dots, x_n^{(\Lambda)}\},$$

and the type (minimum / saddle) is preserved.

Sketch of the argument: Define

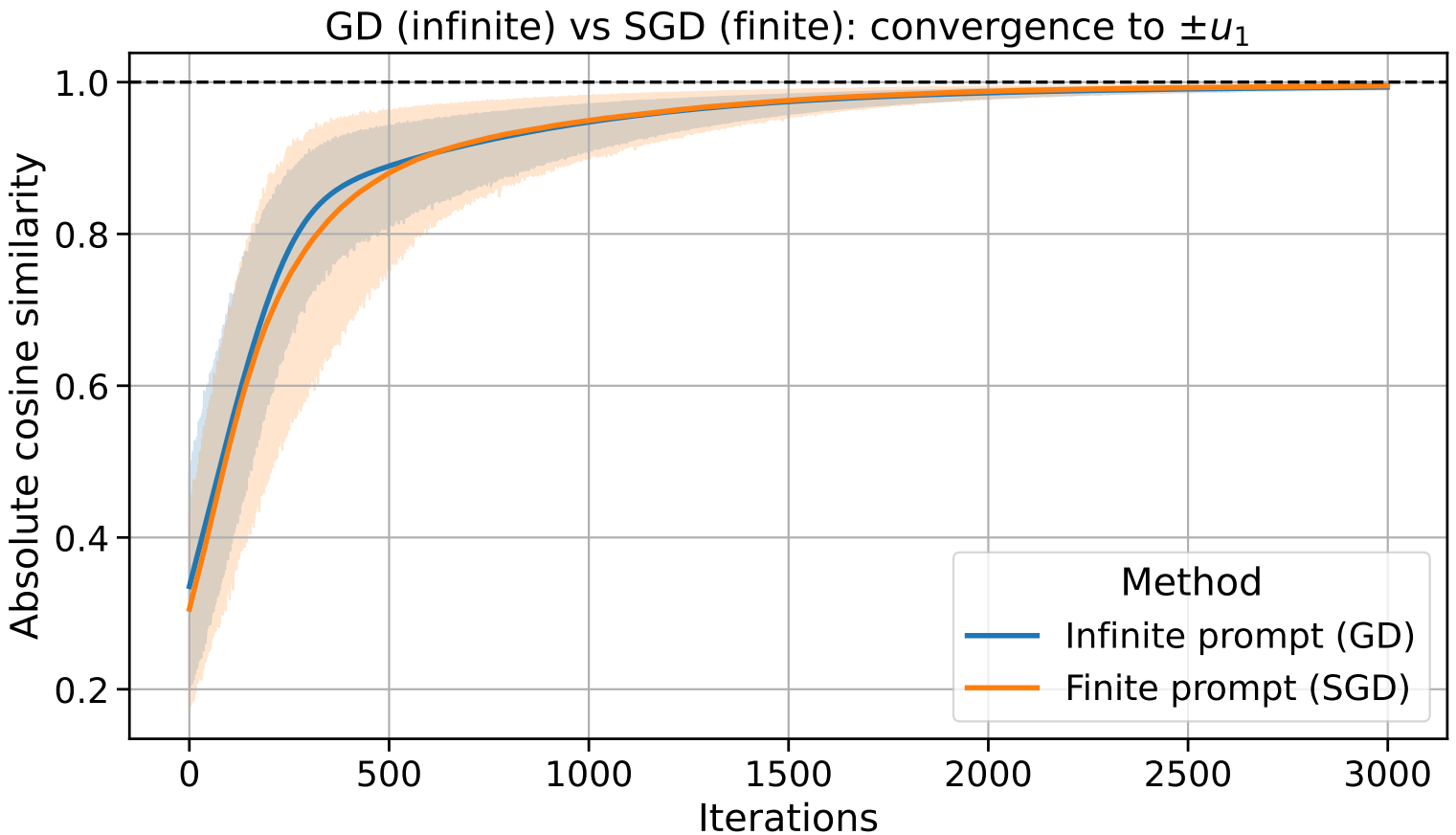
$$H_n(t,x) = (1-t)\nabla f(x) + t\nabla f_n(x).$$

Since $D_x H_n$ is unif. invertible around $x^\star$, the IFT gives a path $x_n(t)$ with $H_n(t, x_n(t)) = 0, x_n(0) = x^\star$. Set $x_n^\star := x_n(1)$.

Numerical results

We run GD on $\mathcal{R}_\infty$ and SGD on $\mathcal{R}_L$ ($L = 100$), we plot iterations vs abs. cosine similarity,

$$\left| \left\langle \frac{\mu_k}{\|\mu_k\|}, u_1 \right\rangle \right|.$$



Distribution of the encodings

Setup. Let $X_1 \sim \mathcal{N}(0, \Sigma)$ and define $T_\infty^\mu(X_1) = \lambda \Sigma \mu \mu^\top X_1$.

Distribution.

$$T_\infty^\mu(X_1) \sim \mathcal{N}(0, \Gamma(\mu)), \quad \Gamma(\mu) = \lambda^2 (\mu^\top \Sigma \mu)\, (\Sigma\mu)(\Sigma\mu)^\top.$$

Distribution of the encodings

Setup. Let $X_1 \sim \mathcal{N}(0, \Sigma)$ and define $T_\infty^\mu(X_1) = \lambda \Sigma \mu \mu^\top X_1$.

Distribution.

$$T_\infty^\mu(X_1) \sim \mathcal{N}(0, \Gamma(\mu)), \quad \Gamma(\mu) = \lambda^2 (\mu^\top \Sigma \mu)\, (\Sigma\mu)(\Sigma\mu)^\top.$$

At the optimum. Let $\mu^\star = \alpha_1 u_1$, with (σ_1, u_1) the principal eigenpair.

$$T_\infty^{\mu^\star}(X_1) \sim \mathcal{N}(0, \sigma_1 u_1 u_1^\top)$$

- Recovers the law of the PCA projection, $\langle X, u_1 \rangle u_1$, optimal rank-1 reconstruction of $X \sim \mathcal{N}(0, \Sigma)$.

Distribution of the encodings

Setup. Let $X_1 \sim \mathcal{N}(0, \Sigma)$ and define $T_\infty^\mu(X_1) = \lambda \Sigma \mu \mu^\top X_1$.

Distribution.

$$T_\infty^\mu(X_1) \sim \mathcal{N}(0, \Gamma(\mu)), \quad \Gamma(\mu) = \lambda^2 (\mu^\top \Sigma \mu)\,(\Sigma\mu)(\Sigma\mu)^\top.$$

At the optimum. Let $\mu^\star = \alpha_1 u_1$, with (σ_1, u_1) the principal eigenpair.

$$T_\infty^{\mu^\star}(X_1) \sim \mathcal{N}(0, \sigma_1 u_1 u_1^\top)$$

- Recovers the law of the PCA projection, $\langle X, u_1 \rangle u_1$, optimal rank-1 reconstruction of $X \sim \mathcal{N}(0, \Sigma)$.

Finite vs infinite prompt.

$$W_2^2\big(\mathcal{L}(T_L^\mu),\, \mathcal{L}(T_\infty^\mu)\big) = \mathcal{O}\big(L^{-\epsilon}(1 + \ln L)^{1-\epsilon}\big)\,.$$

- At the optimum $\mu^\star = \alpha_1 u_1$:
  $W_2^2\Big(\mathcal{L}(T_L^{\mu^\star}),\, \mathcal{L}(T_\infty^{\mu^\star})\Big) = \mathcal{O}\big(L^{-1/145}(1 + \ln L)^{144/145}\big)$

Outline

1. Introduction and problem statement
2. Softmax attention performs PCA
3. Linear attention also performs PCA
4. ICL Risk with Spiked Wishart Prior

Linear attention

Let $X_1, \dots, X_L$ i.i.d. $\mathcal{N}(0, \Sigma)$, we introduce the linear attention operator

$$T_L^{\mathrm{lin},\mu}(\mathbb{X})_\ell = \frac{\lambda}{L} \sum_{k=1}^{L} (X_\ell^\top \mu \mu^\top X_k) X_k.$$

By the Strong Law of Large Numbers,

$$T_L^{\mathrm{lin},\mu}(\mathbb{X})_\ell \xrightarrow[L\to\infty]{a.s.} T_\infty^{\mathrm{soft},\mu}(X_\ell) = \lambda \Sigma \mu \mu^\top X_\ell, \quad a.s..$$

Linear attention

Let $X_1, \dots, X_L$ i.i.d. $\mathcal{N}(0, \Sigma)$, we introduce the linear attention operator

$$T_L^{\mathrm{lin},\mu}(\mathbb{X})_\ell = \frac{\lambda}{L} \sum_{k=1}^{L} (X_\ell^\top \mu \mu^\top X_k) X_k.$$

By the Strong Law of Large Numbers,

$$T_L^{\mathrm{lin},\mu}(\mathbb{X})_\ell \xrightarrow[L\to\infty]{a.s.} T_\infty^{\mathrm{soft},\mu}(X_\ell) = \lambda \Sigma \mu \mu^\top X_\ell, \quad a.s..$$

In the infinite-prompt limit, minimizing the risk

$$\mathcal{R}_{\mathrm{lin},L}(\mu) = \mathbb{E}[\|X_1 - T_L^{\mathrm{lin},\mu}(\mathbb{X})_1\|^2].$$

retrieves the principal eigenvector of Σ.

Linear attention

Let $X_1, \dots, X_L$ i.i.d. $\mathcal{N}(0, \Sigma)$, we introduce the linear attention operator

$$T_L^{\mathrm{lin},\mu}(\mathbb{X})_\ell = \frac{\lambda}{L} \sum_{k=1}^{L} (X_\ell^\top \mu \mu^\top X_k) X_k.$$

By the Strong Law of Large Numbers,

$$T_L^{\mathrm{lin},\mu}(\mathbb{X})_\ell \xrightarrow[L\to\infty]{a.s.} T_\infty^{\mathrm{soft},\mu}(X_\ell) = \lambda \Sigma \mu \mu^\top X_\ell, \quad a.s..$$

In the infinite-prompt limit, minimizing the risk

$$\mathcal{R}_{\mathrm{lin},L}(\mu) = \mathbb{E}[\|X_1 - T_L^{\mathrm{lin},\mu}(\mathbb{X})_1\|^2].$$

retrieves the principal eigenvector of Σ.

In this linear setting, this is also the case for any finite prompt.

Landscape of linear attention risk

$$\min_{\mu \in \mathbb{R}^d} \mathcal{R}_{\mathrm{lin},L}(\mu) = \mathbb{E}[\|X_1 - T_L^{\mathrm{lin},\mu}(\mathbb{X})_1\|^2], \quad \text{where}$$

$$\mathbb{E}[\|X_1 - T_L^{\mathrm{lin},\mu}(\mathbb{X})_1\|^2] = tr(\Sigma) - \frac{2\lambda}{L} tr(\Sigma)(\mu^\top \Sigma \mu) - \frac{2\lambda(L+1)}{L}(\mu^\top \Sigma^2 \mu) + \frac{\lambda^2(L+2)(L+3)}{L^2}(\mu^\top \Sigma \mu)(\mu^\top \Sigma^2 \mu).$$

Proposition

Let $(\sigma_i, u_i)_{i=1}^d$ be the eigenpairs of Σ such that $\sigma_1 > \ldots > \sigma_d$. Then

- $\mathrm{crit}(\mathcal{R}_{\mathrm{lin},L}) = \{0\} \cup \{\pm \alpha_{i,L} u_i\}_{i=1,\ldots,d}$ *for some* $\alpha_{i,L} > 0$.
- 0 *is a local maximum.*
- $\pm \alpha_{i,L} u_i$ *for* $i = 2, \ldots, d$ *is a strict saddle point.*
- $\pm \alpha_{1,L} u_1$ *is a global minimum.*

Landscape of linear attention risk

$$\min_{\mu \in \mathbb{R}^d} \mathcal{R}_{\mathrm{lin},L}(\mu) = \mathbb{E}[\|X_1 - T_L^{\mathrm{lin},\mu}(\mathbb{X})_1\|^2], \quad \text{where}$$

$$\mathbb{E}[\|X_1 - T_L^{\mathrm{lin},\mu}(\mathbb{X})_1\|^2] = tr(\Sigma) - \frac{2\lambda}{L} tr(\Sigma)(\mu^\top \Sigma \mu) - \frac{2\lambda(L+1)}{L}(\mu^\top \Sigma^2 \mu) + \frac{\lambda^2(L+2)(L+3)}{L^2}(\mu^\top \Sigma \mu)(\mu^\top \Sigma^2 \mu).$$

Proposition

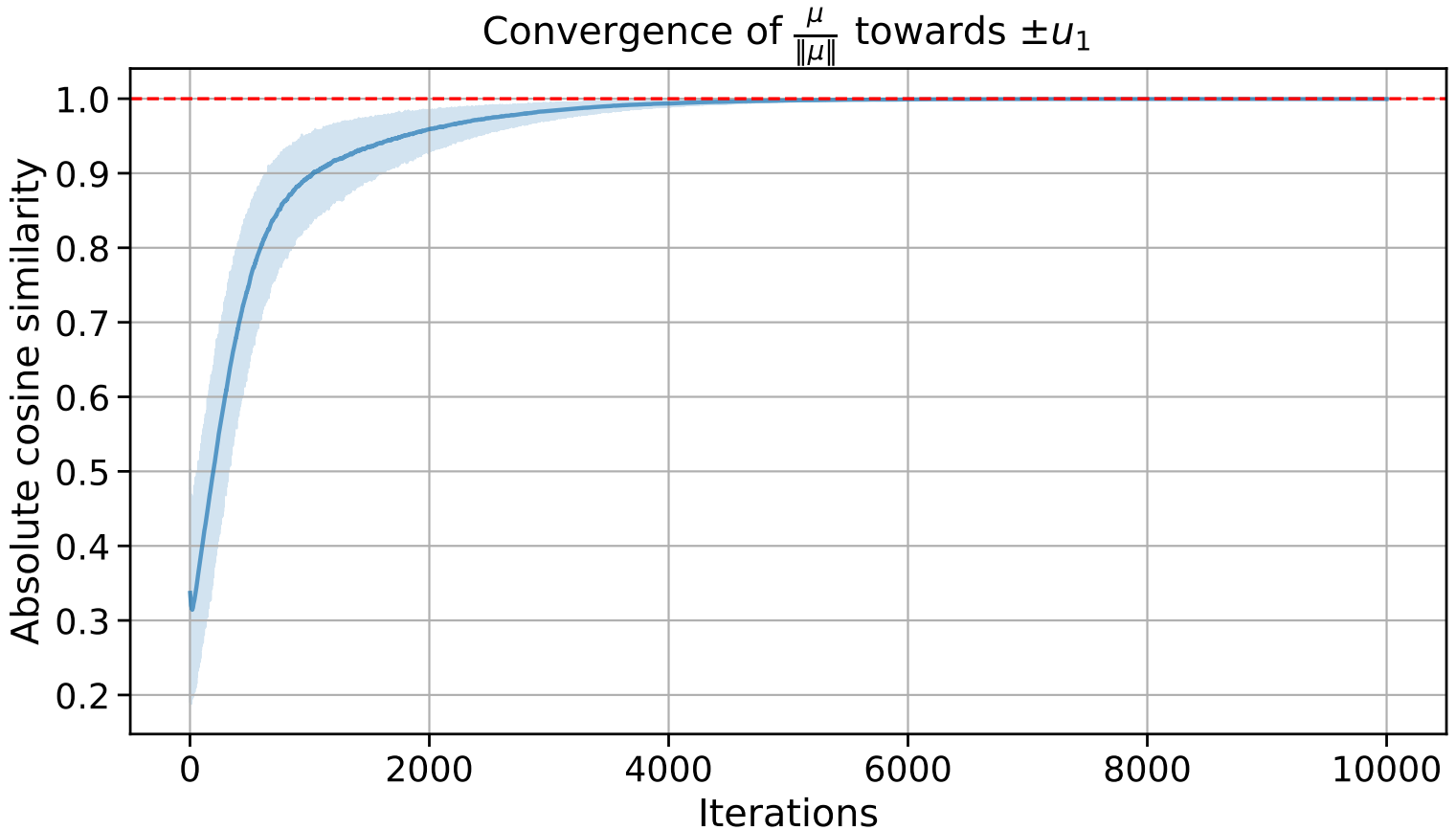
Let $(\sigma_i, u_i)_{i=1}^d$ be the eigenpairs of Σ such that $\sigma_1 > \ldots > \sigma_d$. Then

- $\mathrm{crit}(\mathcal{R}_{\mathrm{lin},L}) = \{0\} \cup \{\pm \alpha_{i,L} u_i\}_{i=1,\ldots,d}$ *for some* $\alpha_{i,L} > 0$.
- 0 *is a local maximum.*
- $\pm \alpha_{i,L} u_i$ *for* $i = 2, \ldots, d$ *is a strict saddle point.*
- $\pm \alpha_{1,L} u_1$ *is a global minimum.*
- *Gradient flow on* $\mathcal{R}_{\mathrm{lin},L}$ *with generic initalization converges to its global minimum.*

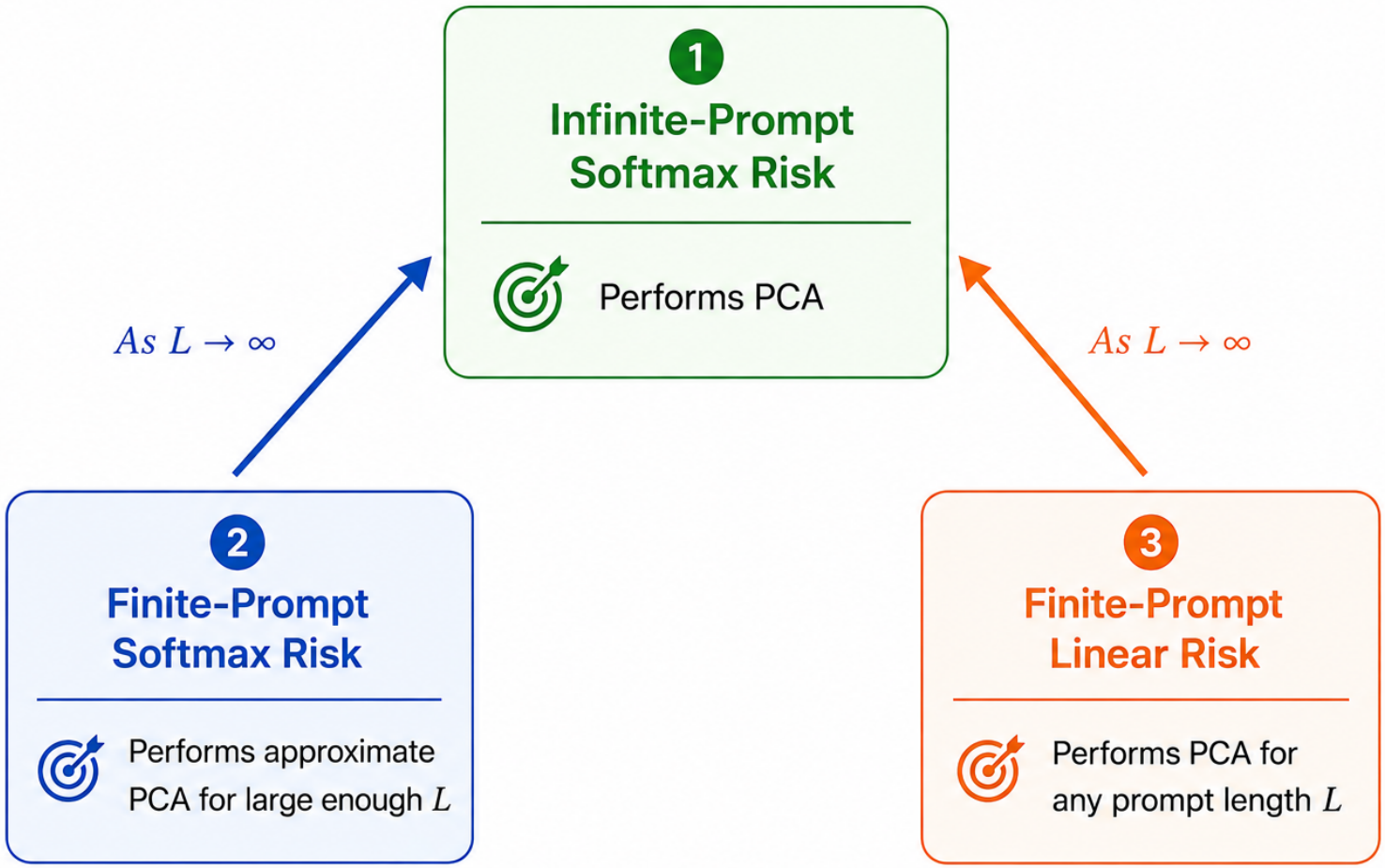
Numerical results: linear risk

We run GD on $\mathcal{R}_{\mathrm{lin},L}$, we plot iterations vs abs. cosine similarity,

$$\left| \left\langle \frac{\mu_k}{\|\mu_k\|}, u_1 \right\rangle \right|.$$



Summary of Results



Outline

1. Introduction and problem statement
2. Softmax attention performs PCA
3. Linear attention also performs PCA
4. ICL Risk with Spiked Wishart Prior

ICL Risk with Spiked Wishart Prior

Recall infinite prompt softmax risk.

$$\mathcal{R}^{(\Sigma)}_{\mathrm{soft},\infty}(\mu) = \mathrm{tr}(\Sigma) - 2\lambda\,\mu^\top\Sigma^2\mu + \lambda^2\,(\mu^\top\Sigma\mu)(\mu^\top\Sigma^2\mu)$$

ICL infinite prompt risk (random covariance).

$$\mathcal{R}^{\mathrm{ICL}}_{\infty}(\mu) = \mathbb{E}_{\Sigma\sim\mathcal{D}}\left[\mathcal{R}^{(\Sigma)}_{\mathrm{soft},\infty}(\mu).\right]$$

ICL Risk with Spiked Wishart Prior

Recall infinite prompt softmax risk.

$$\mathcal{R}^{(\Sigma)}_{\mathrm{soft},\infty}(\mu) = \mathrm{tr}(\Sigma) - 2\lambda\,\mu^\top\Sigma^2\mu + \lambda^2\,(\mu^\top\Sigma\mu)(\mu^\top\Sigma^2\mu)$$

ICL infinite prompt risk (random covariance).

$$\mathcal{R}^{\mathrm{ICL}}_{\infty}(\mu) = \mathbb{E}_{\Sigma\sim\mathcal{D}}\left[\mathcal{R}^{(\Sigma)}_{\mathrm{soft},\infty}(\mu).\right]$$

Model $\mathcal{D}$: Spiked Wishart prior

$$\Sigma \sim W_d(V, n) \Longleftrightarrow \Sigma = \sum_{i=1}^{n} X_i X_i^\top, \quad X_i \sim \mathcal{N}(0, V).$$

$$V = \xi^2 I_d + \theta\, vv^\top, \quad \|v\| = 1.$$

- Interpolates between isotropic case ($\theta = 0$) and structured signal ($\theta > 0$).
- Introduces a **latent direction** v
- **Goal:** Does gradient flow on $\mathcal{R}^{\mathrm{ICL}}_{\infty}$ recover v?

ICL Risk landscape

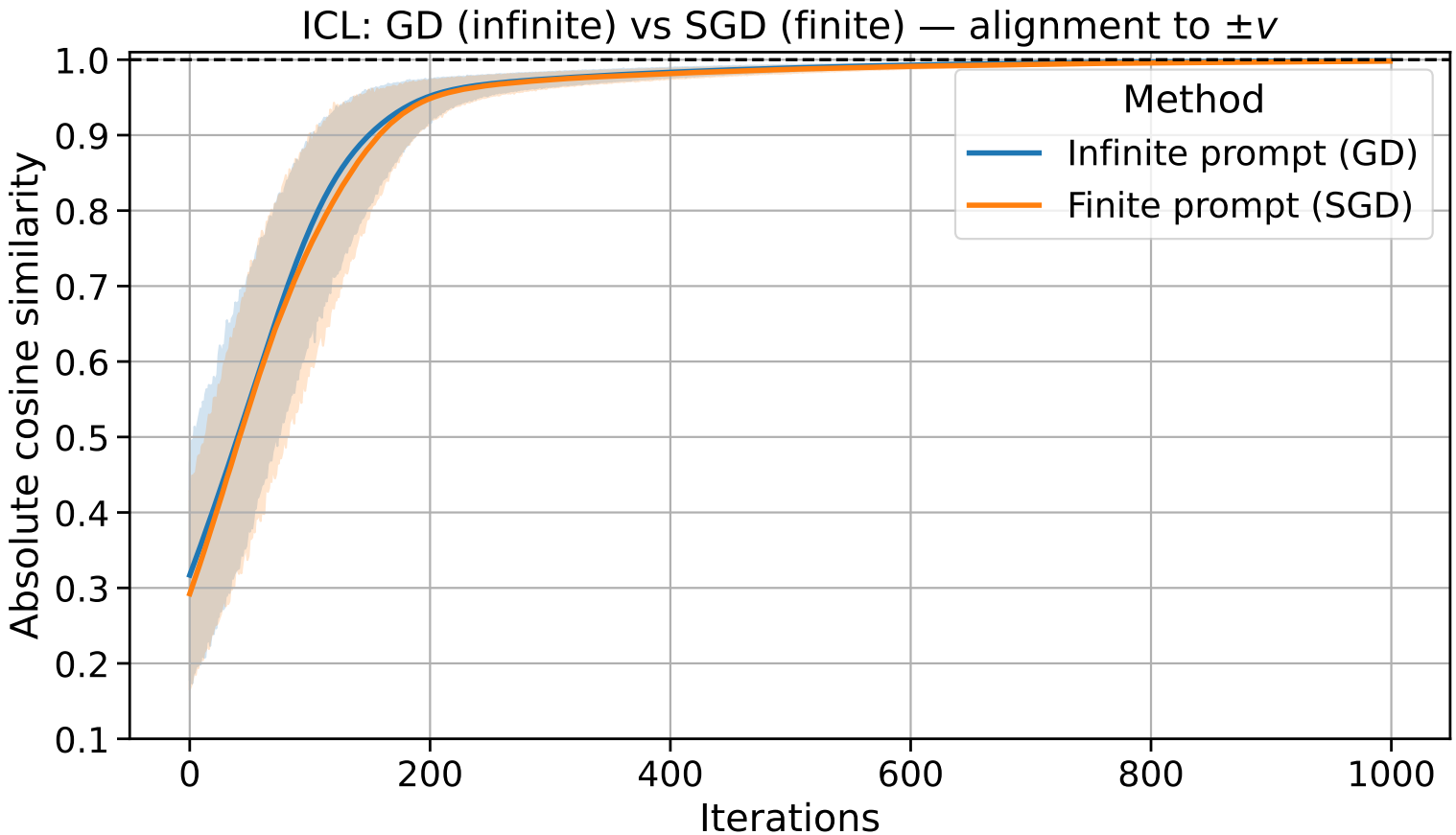
Proposition (Informal version)

The critical points of $\mathcal{R}_{\infty}^{\mathrm{ICL}}(\mu) = \mathbb{E}_{\Sigma \sim W_d(\xi^2 I_d + \theta v v^\top, n)}\left[\mathcal{R}_{\mathrm{soft},\infty}^{(\Sigma)}(\mu)\right]$ *are* 0, *which is a local maximum;* $2(d-1)$ *strict saddle points corresponding to directions orthogonal to* v*; and* $\pm\alpha^\star v$*, for some* $\alpha^\star > 0$*, which are global minimizers.*

- Gradient flow on $\mathcal{R}_{\infty}^{\mathrm{ICL}}$ with generic initialization will recover the signal direction v.
- Similar arguments as before let us obtain that gradient flow on $\mathcal{R}_{L}^{\mathrm{ICL}}(\mu) := \mathbb{E}_{\Sigma \sim W_d(\xi^2 I_d + \theta v v^\top, n)}\left[\mathcal{R}_{\mathrm{soft},L}^{(\Sigma)}(\mu)\right]$ will converge to some $\mu_L^\star$ such that $\mu_L^\star \underset{L\to\infty}{\to} \pm\alpha^\star v$.

We run GD on $\mathcal{R}_{\infty}^{\mathrm{ICL}}$ and SGD on $\mathcal{R}_{L}^{\mathrm{ICL}}$, we plot iterations vs abs. cosine similarity,

$$\left| \left\langle \frac{\mu_k}{\|\mu_k\|}, v \right\rangle \right|.$$



Conclusion

- Attention based mechanisms can recover PCA structures.
- In the infinite-prompt regime, gradient dynamics align with the principal eigenvector and connect to Oja's flow.

Conclusion

- Attention based mechanisms can recover PCA structures.
- In the infinite-prompt regime, gradient dynamics align with the principal eigenvector and connect to Oja's flow.
- Finite-prompt landscapes converge to the infinite-prompt limit as the context length grows.
- The learned encodings recover the optimal rank-one PCA projection in distribution.

Conclusion

- Attention based mechanisms can recover PCA structures.
- In the infinite-prompt regime, gradient dynamics align with the principal eigenvector and connect to Oja's flow.
- Finite-prompt landscapes converge to the infinite-prompt limit as the context length grows.
- The learned encodings recover the optimal rank-one PCA projection in distribution.
- The framework also extends to linear attention and spiked Wishart in-context learning settings.

Thanks for your attention!