Feature learning in terms of spectral flow processes

Cyril Furtlehner

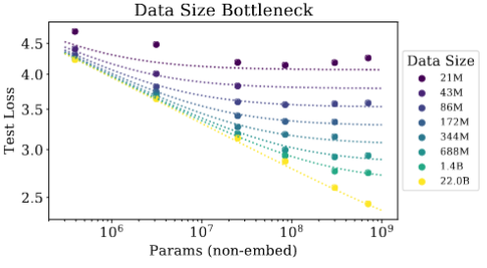
(Inria Saclay,TAU team – UPSaclay,LISN)

Un quart de siècle pour un quart de plan

Generalities on learning theory

Basic questions are

- how NN are able to generalize prediction on unseen data
- how the performance are affected by the choice of architecture (implicit biases), nature of the data, method of optimization
- Universal behaviour
- Learning dynamics

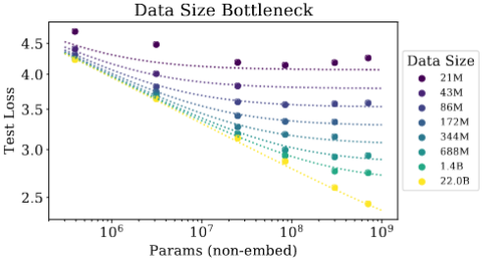


Neural scaling laws w.r.t. N (#samples) and P (#parameters) $\alpha \approx 1/d$ (Kaplan et al. '20)

Generalities on learning theory

Basic questions are

- how NN are able to generalize prediction on unseen data
- how the performance are affected by the choice of architecture (implicit biases), nature of the data, method of optimization
- Universal behaviour
- Learning dynamics

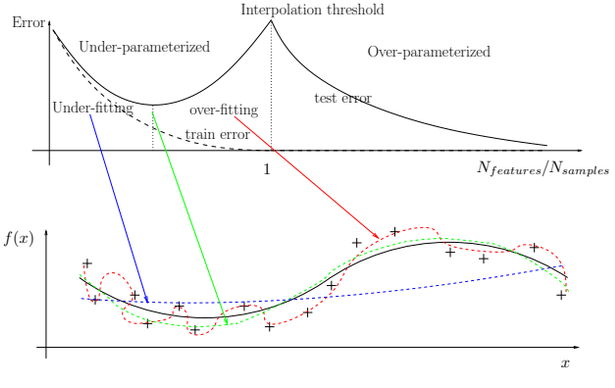


Neural scaling laws w.r.t. N (#samples) and P (#parameters) $\alpha \approx 1/d$ (Kaplan et al. '20)

2 regimes :

- Lazy training regime
- feature learning regime

(non) overfitting paradox :
mechanism of implicit regularization (early stopping,SGD, benign overfitting. . .)



(Belkin et al. '18)

Regression problem and Feature learning

dataset : $\{(\mathbf{x}_s, y_s), s = 1, \dots N\}$, $(\mathbf{x}_s, y_s) \in \mathbb{R}^d \times \mathbb{R}$ from noisy observations

$$y_s = f^\star(\mathbf{x}_s) + \epsilon_s \qquad \text{with} \begin{cases} f^\star \text{ unknown target function} \\ \epsilon \qquad \text{noise} \end{cases}$$

Regression model : $$f_{w,\theta}(\mathbf{x}) = \sum_{k=1}^{M} w_k \phi_k(\mathbf{x}|\theta) \qquad \text{with} \qquad (w, \theta) \in \mathbb{R}^M \times \mathbb{R}^P$$

$\{\phi_k(x|\theta), k = 1, \dots M\}$: features to be learned by gradient descent of the loss

$$\mathcal{L}(w, \theta) = \frac{\|w\|^2}{2} + \frac{\alpha}{N} \sum_{s=1}^{N} \big(f_{w,\theta}(\mathbf{x}_s) - y_s\big)^2.$$

α : coupling constant $=$ inverse ridge penalty.

Regression problem and Feature learning

dataset : $\{(\mathbf{x}_s, y_s), s = 1, \dots N\}$, $(\mathbf{x}_s, y_s) \in \mathbb{R}^d \times \mathbb{R}$ from noisy observations

$$y_s = f^\star(\mathbf{x}_s) + \epsilon_s \qquad \text{with} \begin{cases} f^\star \text{ unknown target function} \\ \epsilon \qquad \text{noise} \end{cases}$$

Regression model : $$f_{w,\theta}(\mathbf{x}) = \sum_{k=1}^{M} w_k \phi_k(\mathbf{x}|\theta) \qquad \text{with} \qquad (w, \theta) \in \mathbb{R}^M \times \mathbb{R}^P$$

$\{\phi_k(x|\theta), k = 1, \dots M\}$: features to be learned by gradient descent of the loss

$$\mathcal{L}(w, \theta) = \frac{\|w\|^2}{2} + \frac{\alpha}{N} \sum_{s=1}^{N} \big(f_{w,\theta}(\mathbf{x}_s) - y_s\big)^2.$$

α : coupling constant $=$ inverse ridge penalty.

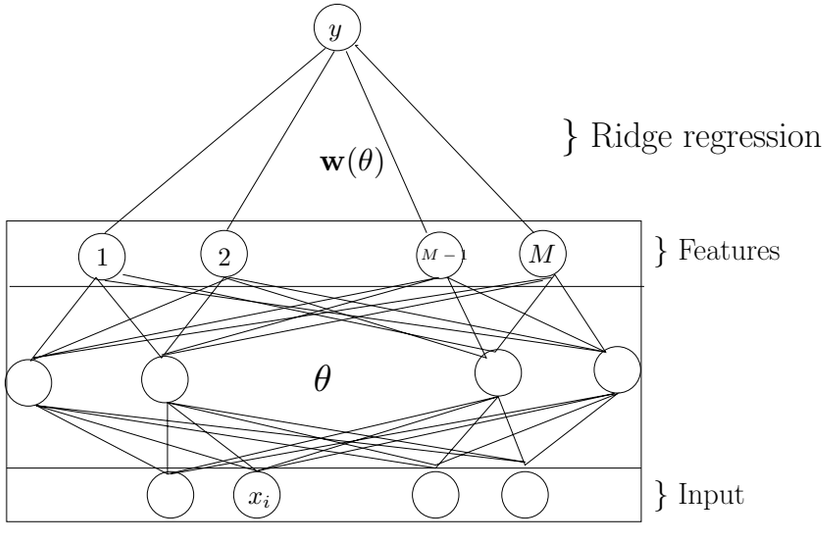
What can be said about the evolution of the features $\phi(x|\theta_t)$?

Online learning with time scales separation

Online learning : independent batch of N data at each time step.

Specific setting considered :

- assumming fast convergence of the last layer take $w_t = w(\theta_t)$ solution of the ridge regression
- one gradient step is performed on θ_t, with new batch of N data



Consider asymptotic limit : $N, M, P \to \infty$ with fixed ratios $\rho = N/M$ and P/N with RMT.

Part I : Ridge regression and feature alignment

Ridge regression problem

Given the signal $\mathbf{f}$ observed through

$$y = f(\mathbf{x}) + \mathcal{N}(0, \sigma^2), \qquad \mathbf{x} \in \mathbb{R}^d,\ y \in \mathbb{R}$$

a vector of feature functions $\phi(\mathbf{x}) \in \mathbb{R}^M$ find $\mathbf{w} \in \mathbb{R}^M$ minimizing

$$\mathcal{L}(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2} + \alpha E_{\text{train}} \qquad \text{with} \qquad E_{\text{train}} = \frac{1}{2N} \sum_{s=1}^{N} \left| y_s - \mathbf{w}^\top \phi(\mathbf{x}_s) \right|^2$$

α^{-1} : ridge penalty

Ridge regression problem

Given the signal $\mathbf{f}$ observed through

$$y = f(\mathbf{x}) + \mathcal{N}(0, \sigma^2), \qquad \mathbf{x} \in \mathbb{R}^d,\ y \in \mathbb{R}$$

a vector of feature functions $\phi(\mathbf{x}) \in \mathbb{R}^M$ find $\mathbf{w} \in \mathbb{R}^M$ minimizing

$$\mathcal{L}(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2} + \alpha E_{\text{train}} \qquad \text{with} \qquad E_{\text{train}} = \frac{1}{2N} \sum_{s=1}^{N} \left| y_s - \mathbf{w}^\top \phi(\mathbf{x}_s) \right|^2$$

α^{-1} : ridge penalty

Optimal solution : $\hat{\mathbf{w}} = \alpha \hat{G} \hat{Z}$ with

$$\begin{cases} \hat{C} & \overset{\text{def}}{=} \frac{1}{N} \sum_{s=1}^{N} \phi(\mathbf{x}_s) \phi(\mathbf{x}_s)^\top, \\ \hat{Z} & \overset{\text{def}}{=} \frac{1}{N} \sum_{s=1}^{N} \phi(\mathbf{x}_s) y_s, \\ \hat{G} & \overset{\text{def}}{=} \left(\mathbb{I} + \alpha \hat{C} \right)^{-1}. \qquad \text{(resolvent)} \end{cases}$$

Asymptotics of the generalization error

Train and test errors :

$$E_{\text{train}} = \frac{1}{\alpha}\,\mathsf{Tr}\Big[\hat{G}(\mathbb{I}-\hat{G})\phi\mathbf{f}\mathbf{f}^\top\phi^{\dagger\top}\Big] + \sigma_{\text{eff}}^2\Big(1-\rho^{-1}+\frac{1}{N}\,\mathsf{Tr}\Big[\hat{G}^2\Big]\Big),$$

$$E_{\text{test}} = \underbrace{\mathsf{Tr}\Big[\hat{G}C\hat{G}\phi\mathbf{f}\mathbf{f}^\top\phi^{\dagger\top}\Big]}_{\text{Bias}} + \underbrace{\sigma_{\text{eff}}^2\Big(1+\frac{\alpha}{N}\,\mathsf{Tr}\Big[\hat{G}(\mathbb{I}-\hat{G})C\Big]\Big)}_{\text{Variance}}$$

with

$$\rho = \frac{N}{M} \qquad \text{and} \qquad \sigma_{\text{eff}}^2 \overset{\text{def}}{=} \sigma^2 + \|\mathbf{f}^\perp\|^2.$$

and

$$C \overset{\text{def}}{=} \mathbb{E}\big(\phi(\mathbf{x})\phi(\mathbf{x})^\top\big) \qquad \textbf{population matrix}$$

Asymptotics of the generalization error

Train and test errors :

$$E_{\text{train}} = \frac{1}{\alpha}\,\mathsf{Tr}\Big[\hat{G}(\mathbb{I}-\hat{G})\phi\mathbf{f}\mathbf{f}^\top\phi^{\dagger\top}\Big] + \sigma_{\text{eff}}^2\Big(1-\rho^{-1}+\frac{1}{N}\,\mathsf{Tr}\Big[\hat{G}^2\Big]\Big),$$

$$E_{\text{test}} = \underbrace{\mathsf{Tr}\Big[\hat{G}C\hat{G}\phi\mathbf{f}\mathbf{f}^\top\phi^{\dagger\top}\Big]}_{\text{Bias}} + \underbrace{\sigma_{\text{eff}}^2\Big(1+\frac{\alpha}{N}\,\mathsf{Tr}\Big[\hat{G}(\mathbb{I}-\hat{G})C\Big]\Big)}_{\text{Variance}}$$

with

$$\rho = \frac{N}{M} \qquad \text{and} \qquad \sigma_{\text{eff}}^2 \overset{\text{def}}{=} \sigma^2 + \|\mathbf{f}^\perp\|^2.$$

and

$$C \overset{\text{def}}{=} \mathbb{E}\big(\phi(\mathbf{x})\phi(\mathbf{x})^\top\big) \qquad \textbf{population matrix}$$

RMT : asymptotics when $N, M \to \infty$ with fixed ρ heavily investigated

- Dobriban-Wager (2018), average on isotropic signal ($C = \mathbb{I}$)
- Wu-Xu (2020), Richards et al. (2021), average on non-isotropic with restrictions
- Hastié et. al (2022) finite size bounds

Leave-one out argument

Thanks to Sherman-Morrison formula we have

$$\hat{G} = \hat{G}_{\setminus s} + \gamma_s \hat{G}_{\setminus s} \phi(\mathbf{x}_s) \phi(\mathbf{x}_s)^\top \hat{G}_{\setminus s}, \tag{1}$$

with

$$\gamma_s = \frac{\alpha}{N + K_{\setminus s}(\mathbf{x}_s, \mathbf{x}_s)}$$

and

$$\hat{K}_{\setminus s}(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \hat{G}_{\setminus s} \phi(\mathbf{x}')$$

Leave-one out argument

Thanks to Sherman-Morrison formula we have

$$\hat{G} = \hat{G}_{\setminus s} + \gamma_s \hat{G}_{\setminus s} \phi(\mathbf{x}_s) \phi(\mathbf{x}_s)^\top \hat{G}_{\setminus s}, \tag{1}$$

with

$$\gamma_s = \frac{\alpha}{N + K_{\setminus s}(\mathbf{x}_s, \mathbf{x}_s)}$$

and

$$\hat{K}_{\setminus s}(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \hat{G}_{\setminus s} \phi(\mathbf{x}')$$

yields the leave-one out relation :

$$E_{\text{test}}(\mathbf{x}_s) = \left[1 + \frac{1}{N} \hat{K}(\mathbf{x}_s, \mathbf{x}_s)\right]^2 E_{\text{train}}(\mathbf{x}_s)$$

Some RMT formulas

Marchenko-Pastur (1967)

Let $\rho = \dfrac{N}{M}$ and $\begin{cases} \nu(dx) & \text{spectral density of population matrix } C \\ \sigma(d\tau) & \text{distribution of } \tau_s = \|\mathbf{x}_s\| \end{cases}$

Some RMT formulas

Marchenko-Pastur (1967)

Let $\rho = \dfrac{N}{M}$ and $\begin{cases} \nu(dx) & \text{spectral density of population matrix } C \\ \sigma(d\tau) & \text{distribution of } \tau_s = \|\mathbf{x}_s\| \end{cases}$

Mean field equations : $\begin{cases} \Gamma = \dfrac{\alpha}{\rho} \displaystyle\int \nu(dx) \frac{x}{1+\Lambda x} & \left(\overset{\text{def}}{=} \lim\limits_{N,M\to\infty} \dfrac{\alpha}{N} \mathsf{Tr}[\hat{G}C]\right) \\ \Lambda = \alpha \displaystyle\int \sigma(d\tau) \frac{\tau}{1+\Gamma\tau} & \text{(Effective coupling)} \end{cases}$

Ledoit-Péchet (2011) $\displaystyle\lim_{N,M\to\infty} \frac{1}{M} \mathsf{Tr}\left[\hat{G} h(C)\right] = \int \frac{\nu(dx) h(x)}{1+\Lambda x}$

Deterministic equivalent : $\hat{G} \sim \dfrac{1}{\mathbb{I} + \Lambda C}$

Spectral decomposition of the signal w.r.t. population matrix

Spectral decomposition of hidden signal along C + transverse modes as

$$f = f^{\parallel} + f^{\perp} \stackrel{\text{def}}{=} \sum_{a=1}^{M} f_a u_a + \sum_{b>M} f_b u_b$$

power spectrum : $\mu_a = M f_a^2 \qquad \mu^{\parallel} \stackrel{\text{def}}{=} \sum_{a=1}^{M} f_a^2 \qquad \mu^{\perp} \stackrel{\text{def}}{=} \sum_{b>M} f_b^2$

spectral moments

$$\begin{cases} g_n \stackrel{\text{def}}{=} \int \dfrac{\nu(dx)}{(1+\Lambda x)^n} & \text{specific to } C \\ \bar{\mu}_n \stackrel{\text{def}}{=} \int \dfrac{\mu(dx)}{(1+\Lambda x)^n} & \text{alignment measures} \end{cases}$$

with

$$\mu(x) = \lim_{M\to\infty} \frac{1}{M} \sum_{a=1}^{M} \mu_a \delta(x - c_a)$$

Summary of asymptotic expressions

(simplified version $\nu(\tau) = \delta(\tau - 1)$)

Train and test errors :

$$E_{\text{train}} = \frac{\Lambda^2}{\alpha^2} E_{\text{test}}$$

$$E_{\text{test}} = \frac{\rho\big(1 - \bar{\mu}_0 + \bar{\mu}_2\big)}{\big(\rho - \underbrace{(1 - 2g_1 + g_2)}_{\text{spectral parameter}}\big)}$$

The Loss : $$\mathcal{L} = \frac{\Lambda}{2}\big[1 - \bar{\mu}_0 + \bar{\mu}_1\big].$$

Fixed-point equation :

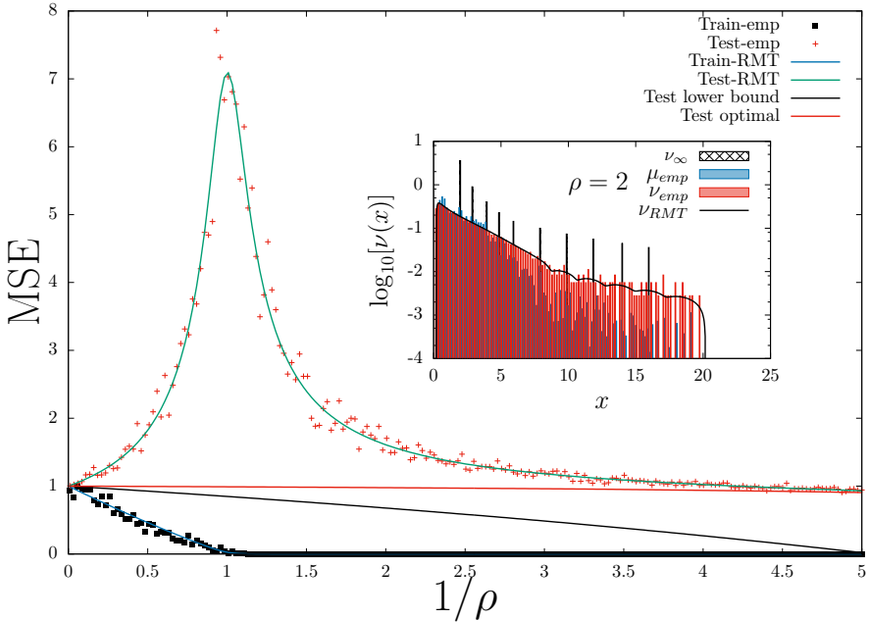
$$\Lambda = \alpha\Big(1 - \frac{1}{\rho}\Big) + \frac{\alpha}{\rho} g_1$$

signal to noise :

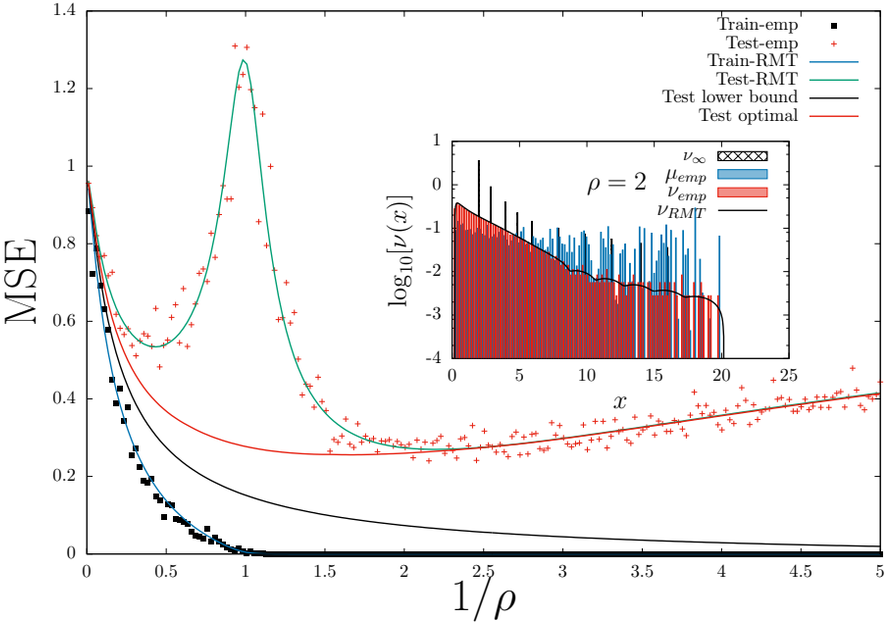
$$\bar{\mu}_0 = \frac{SNR}{1 + SNR}$$

Numerical examples

Number of prior levels (population matrix) : 10
Numbers of training samples : $N = 200$



anti-aligned



aligned

Part II : Dynamical systems for feature learning processes

Gradient of the asymptotic loss

Recall

(i) the model $\quad f_\theta(\mathbf{x}|\theta) = \sum_k w_k(\theta)\phi_k(\mathbf{x}|\theta)$

(ii) feature map frame (SVD)

$$\begin{cases} \phi_k(\mathbf{x}|\theta) = \sum_a \sqrt{c_a} v_{ka} u_a(\mathbf{x}|\theta) & \text{(feature map)} \\ C(\theta) = \int \underbrace{\omega(d\mathbf{x})}_{\text{population density}} \phi(\mathbf{x}|\theta)\phi(\mathbf{x}|\theta)^t & \text{(population matrix)} \\ f^\star(\mathbf{x}) = \sum_a f_a u_a(\mathbf{x}|\theta) + f^{\star\perp}(\mathbf{x}) & \text{(target function)} \end{cases}$$

(iii) asymptotic loss : $\quad \mathcal{L} = \frac{\Lambda}{2}\big[\mu_0 + \bar{\mu}_1\big] = \mathcal{L}\big(\{(c_a, \mu_a), a = 1, \ldots\}\big)$

Arbitrary feature deformations $d\Phi_{ab}$ yield

$$d\mathcal{L} = \sum_{a,b} \frac{\partial \mathcal{L}}{\partial \Phi_{ab}}(\{(c_a, f_a), a = 1, \ldots\})\ d\Phi_{ab},$$

Free dynamics

$$\theta \Longleftrightarrow \{\Phi_{ab}\} : \qquad \frac{d\Phi_{ab}}{dt} = -\frac{\partial \mathcal{L}}{\partial \Phi_{ab}},$$

yields the **autonomous systems** for $a = 1, \dots M$:

$$\dot{\mu}_a = \Lambda^2 \Big[\underbrace{\frac{\mu_a \mu_0^{\perp}}{1 + \Lambda c_a}}_{J_{0\to a}} + \underbrace{\frac{1}{M} \sum_{b=1, b\neq a}^{M} \frac{c_a + c_b}{c_a - c_b} \frac{\mu_a \mu_b}{\big[1 + \Lambda c_a\big]\big[1 + \Lambda c_b\big]}}_{\sum_{b\neq a} J_{b\to a}} \Big],$$

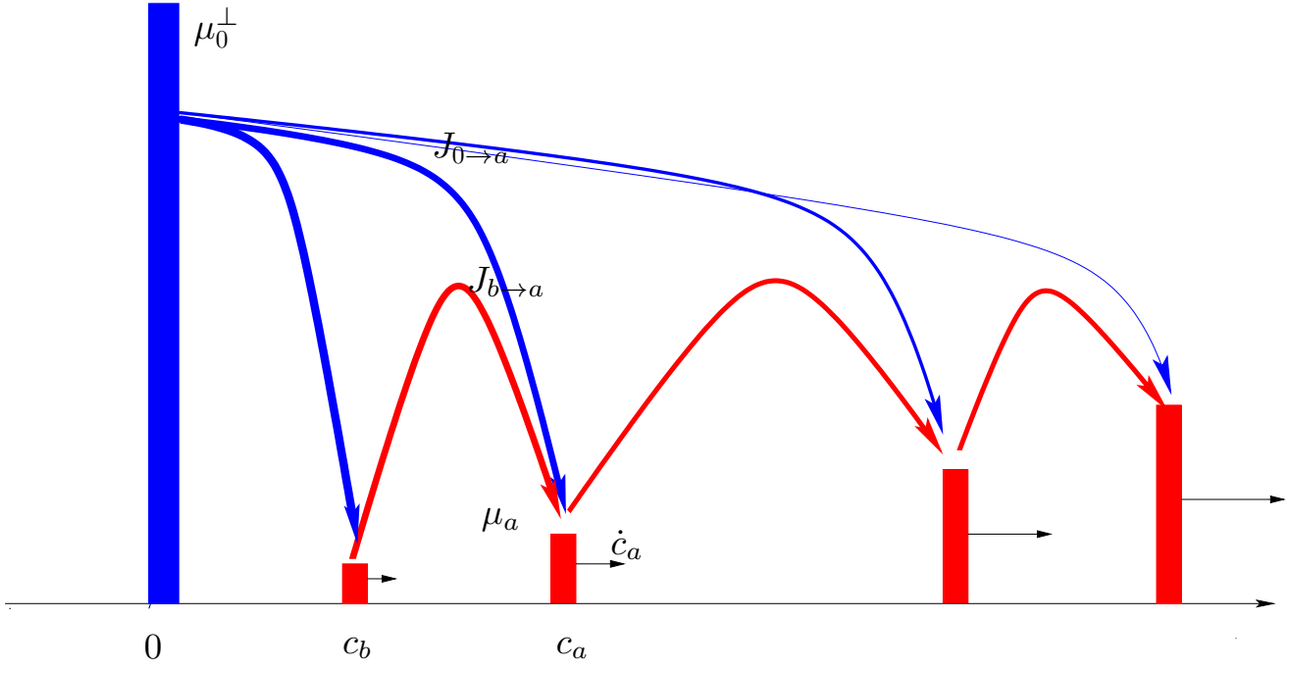
$$\dot{c}_a = \frac{1}{M} \frac{2\Lambda^2 c_a}{\big[1 + \Lambda c_a\big]^2} \Big(\mu_a + \frac{E_{\rm test}}{\rho} \Big).$$

$E_{\rm test}$, Λ and $\mu_0^{\perp}$ **macroscopic variables** functions of the $\{c_a, \mu_a\} = 2M$ **microscopic degrees of freedom**.

$\mu_a = \mathcal{O}\Big(1/\sqrt{M}\Big)$: fast variables,

$c_a = \mathcal{O}(1)$: slow variables.

Spectral Flow



$$\dot{c}_a = \frac{1}{M}\frac{2\Lambda^2 c_a}{\left[1+\Lambda c_a\right]^2}\left(\mu_a + \frac{E_{\text{test}}}{\rho}\right)$$

$$\begin{cases} J_{0\to a} = \Lambda^2 \dfrac{\mu_a \mu_0^\perp}{1+\Lambda c_a} \\ J_{b\to a} = \dfrac{\Lambda^2}{M}\dfrac{c_a + c_b}{c_a - c_b}\dfrac{\mu_a \mu_b}{\left[1+\Lambda c_a\right]\left[1+\Lambda c_b\right]} \end{cases}$$

Macroscopic equations

No closed set of equations, **hierarchy of equations** instead.
for instance :

$$\dot{\Lambda} = \frac{2\Lambda^3 \bar{\mu}_0}{M}\Big(\frac{\bar{\mu}_3 - \bar{\mu}_4}{\rho - \mathcal{Q}} + \bar{\mu}_2 \frac{g_3 - g_4}{(\rho - \mathcal{Q})^2}\Big),$$

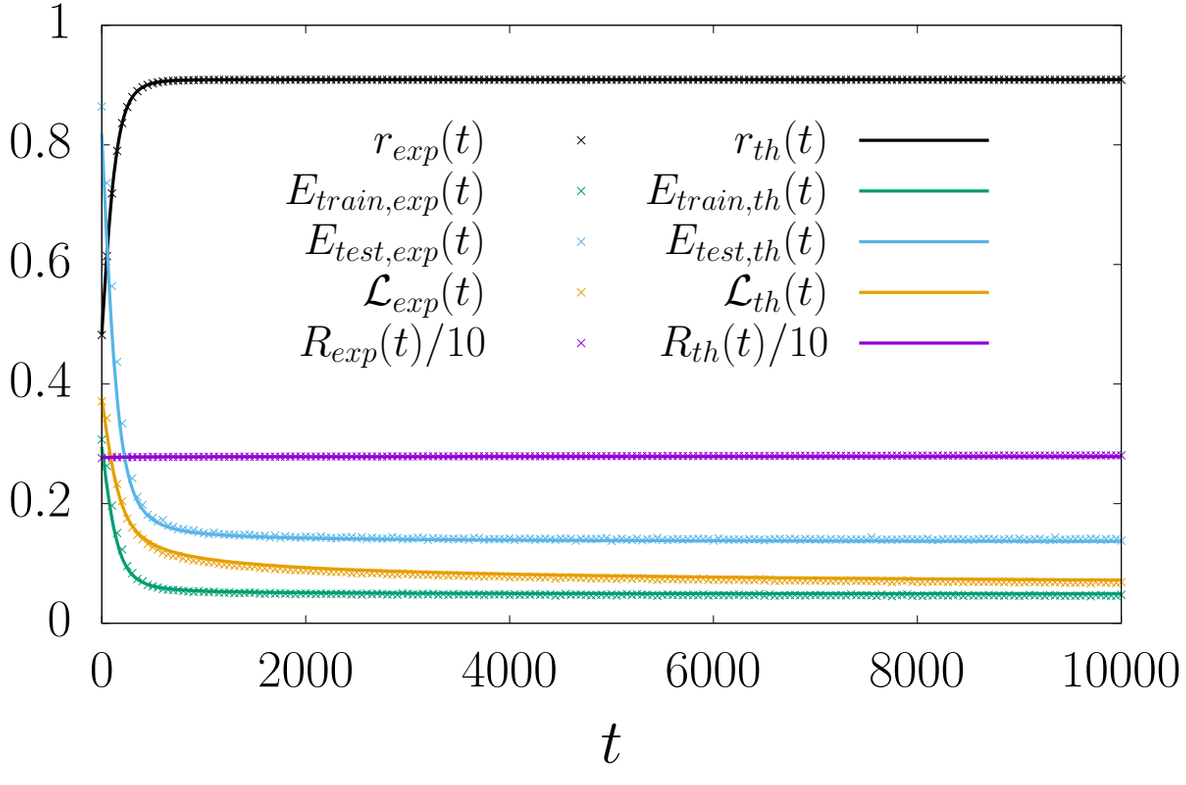
$$\dot{\mu}_0^{\perp} = -2\Lambda^2 \mu_0^{\perp}(1 - \mu_0^{\perp} - \sigma^2)\mu_1,$$

(with $\mathcal{Q} = 1 - 2g_1 + g_2$)
Sigmoid like behavior for transverse spectral weight :

$$\mu_0^{\perp}(t) = \frac{1 - \sigma^2}{1 + A \exp\Big(2 \int_0^t d\tau \Lambda^2(\tau)\bar{\mu}_1(\tau)\Big)},$$

Numerical experiments : under-parameterized regime

$$\rho = 2$$

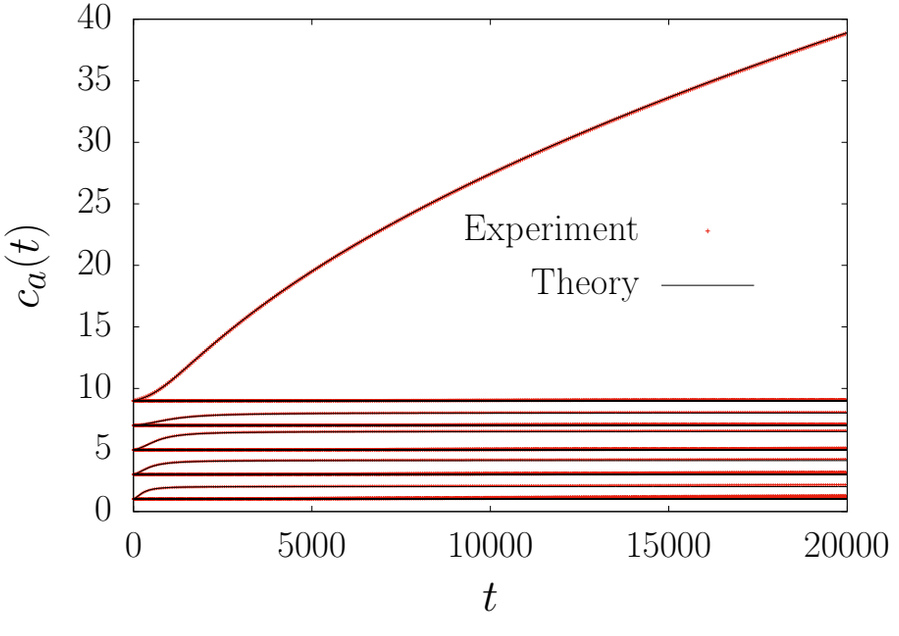


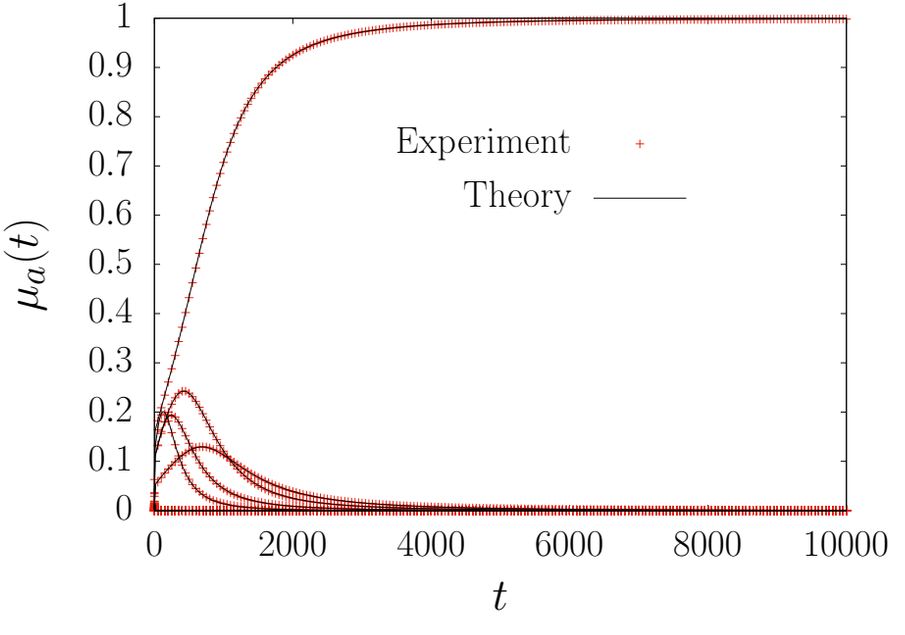
Number of prior levels (population matrix) : 5
Numbers of training samples : $N = 200$

Numerical experiments : under-parameterized regime

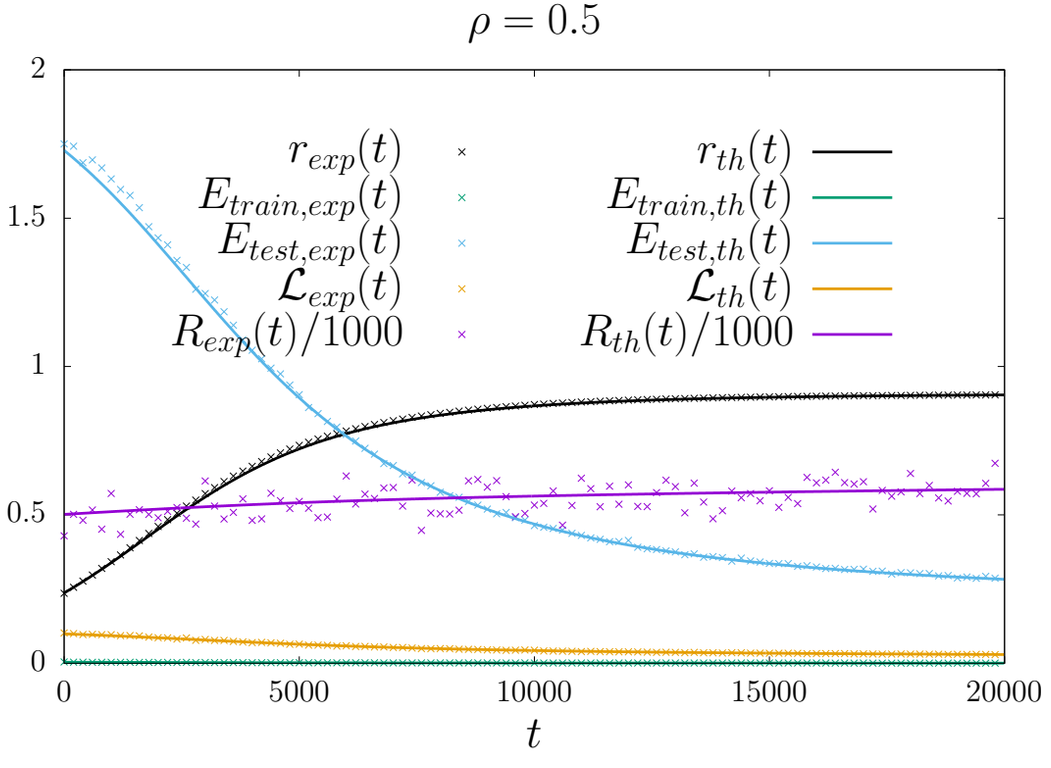
$$(\rho = 2)$$

Number of prior levels (population matrix) : 5
Numbers of training samples : $N = 200$





Numerical experiments : over-parameterized regime

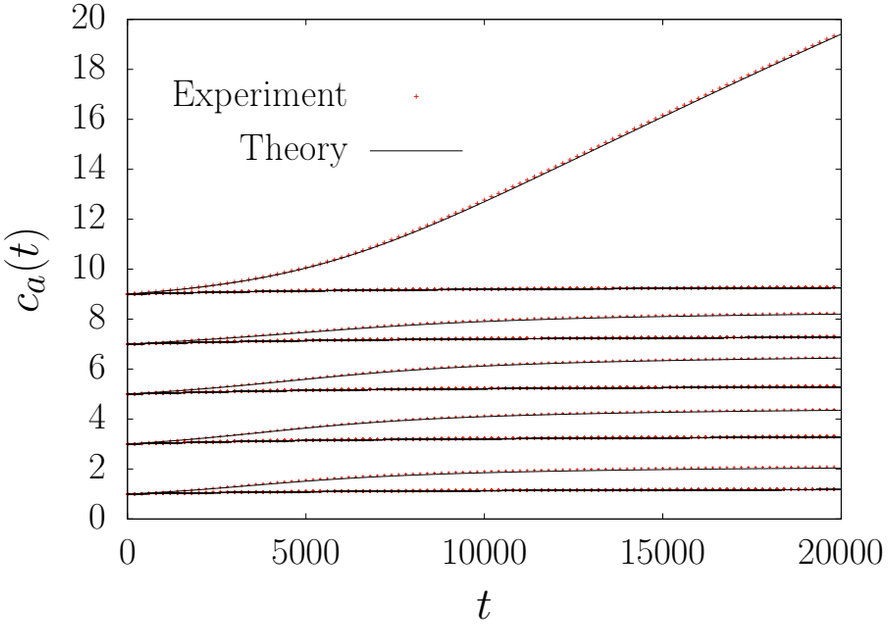


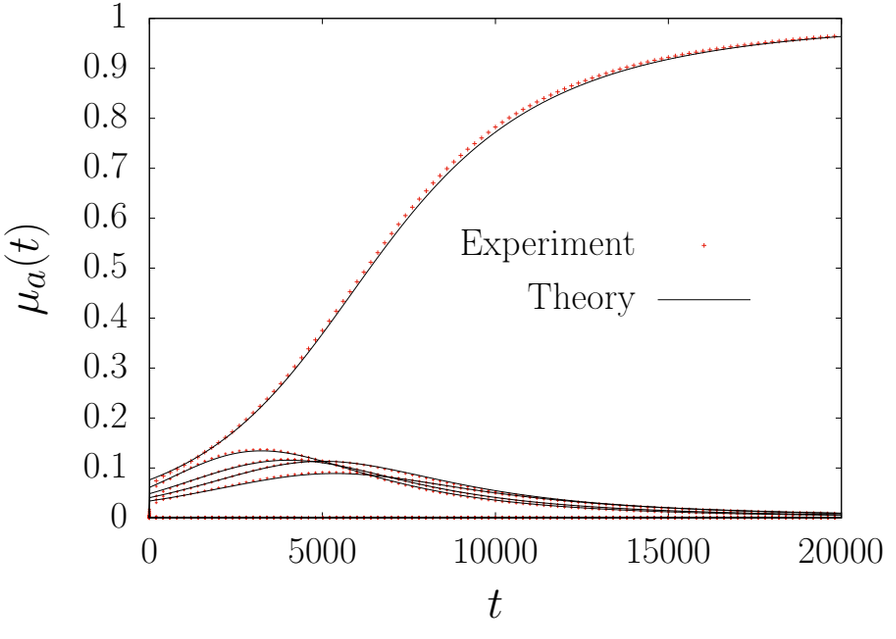
Number of prior levels (population matrix) : 5
Numbers of training samples : $N = 50$

Numerical experiments : over-parameterized regime

$$(\rho = 0.5)$$

Number of prior levels (population matrix) : 5
Numbers of training samples : $N = 50$





General case

feature dynamics depend on architecture dependent kernel :

$$\frac{d\phi_k(\mathbf{x}|\theta_t)}{dt} = -\delta_{k\ell} \int \hat{\omega}(d\mathbf{y}) K_{\theta_t}^{(k\ell)}(\mathbf{x},\mathbf{y}) \frac{\partial \mathcal{L}}{\partial \phi_\ell(\mathbf{x})}$$

Free case corresponds to no shared parameters + features and signal $\in \mathcal{H}_{K_{\theta_t}}, \forall t$+natural gradient :

$$K_{\theta_t}^{(k\ell)}(\mathbf{x},\mathbf{y}) = \delta_{k\ell} K_{\theta_t}(\mathbf{x},\mathbf{y}) \qquad \text{and} \qquad \begin{cases} u_a(\mathbf{x}|\theta_t) = \int \omega(d\mathbf{y}) K_{\theta_t}(\mathbf{x},\mathbf{y}) u_a(\mathbf{y}|\theta_t) \\ f^\star(\mathbf{x}) = \int \omega(d\mathbf{y}) K_{\theta_t}(\mathbf{x},\mathbf{y}) f^\star(\mathbf{y}) \end{cases}$$

Otherwise we obtain a non-diagonal kernel in feature frame :

$$K_{ab}^{cd} = \sum_{n=1}^{P} \frac{\partial \Phi_{ab}}{\partial \theta_n} \frac{\partial \Phi_{cd}}{\partial \theta_n} \neq \delta_{ac}\delta_{cd}$$

Case of NN without implicit bias

Way out : D is dimension of functional space, $P\#$ parameters with $\gamma = \frac{P}{D}$ fixed

$$K = \frac{d\Phi}{d\theta^t}\frac{d\Phi}{d\theta}^t \qquad (DM) \times (DM) \text{ random matrix of rank } P$$

Result depends only on : $\mathrm{Tr}[K] = MP$, $\mathrm{Tr}[K^2] = (1 + \sigma_{\mathrm{K}}^2)M^2P$.

Case of NN without implicit bias

Way out : D is dimension of functional space, $P\#$ parameters with $\gamma = \frac{P}{D}$ fixed

$$K = \frac{d\Phi}{d\theta^t}\frac{d\Phi}{d\theta}^t \qquad (DM) \times (DM) \text{ random matrix of rank } P$$

Result depends only on : $\mathrm{Tr}[K] = MP$, $\mathrm{Tr}[K^2] = (1 + \sigma_{\mathrm{K}}^2)M^2P$.

Spectral dynamics

$$\begin{cases} \dot{c}_a = \gamma \bar{v}_a + \sigma_a \, \eta_{a,t} & (\bar{v} = \text{ free velocity}) \\ \dot{\mu}_a = \gamma \sum_{b \neq a} \left(\bar{J}_{b \to a} + \sigma_{ab} \, \eta_{b \to a, t} \right) & (\bar{J} = \text{ free current}) \end{cases}$$

(with $\gamma = \frac{P}{D}$ noise $\eta_{\cdot,t} = \mathcal{N}(0,1)$, $\eta_{a \to b,t} = -\eta_{b \to a,t} = \mathcal{N}(0,1)$, $\sigma_{ab} = \sigma_{ba}$)

Case of NN without implicit bias

Way out : D is dimension of functional space, $P\#$ parameters with $\gamma = \frac{P}{D}$ fixed

$$K = \frac{d\Phi}{d\theta^t}\frac{d\Phi}{d\theta}^t \qquad (DM) \times (DM) \text{ random matrix of rank } P$$

Result depends only on : $\mathrm{Tr}[K] = MP$, $\mathrm{Tr}[K^2] = (1+\sigma_{\mathrm{K}}^2)M^2P$.

Spectral dynamics

$$\begin{cases} \dot{c}_a = \gamma \bar{v}_a + \sigma_a \, \eta_{a,t} & (\bar{v} = \text{ free velocity}) \\ \dot{\mu}_a = \gamma \sum_{b \neq a} \left(\bar{J}_{b\to a} + \sigma_{ab} \, \eta_{b\to a,t} \right) & (\bar{J} = \text{ free current}) \end{cases}$$

(with $\gamma = \frac{P}{D}$ noise $\eta_{\cdot,t} = \mathcal{N}(0,1)$, $\eta_{a\to b,t} = -\eta_{b\to a,t} = \mathcal{N}(0,1)$, $\sigma_{ab} = \sigma_{ba}$)

Main points :

Autonomous stochastic system (if $\sigma_{\mathrm{K}} = cte$) and $\sigma\left(\{c_a, \mu_a\}, \sigma_{\mathrm{K}}\right) = \mathcal{O}\left(\frac{M}{\sqrt{P}}\right) \times (|\bar{J}|\,,\, \bar{v})$

Discussion

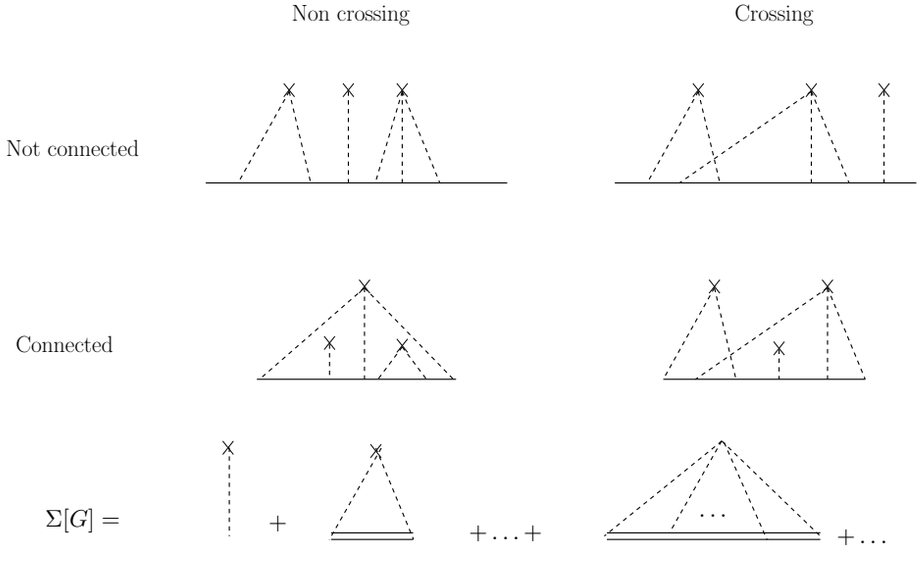
- generalization of leave-one out argument to other contexts
- incerting implicit bias into the dynamical system
- Monitoring the dynamics in real training
- Low dimensional ML systems like PINNs

Physics viewpoint on free probabilities

To compute trace of random matrices like

$$G_\rho^{(N)}(z) \stackrel{\text{def}}{=} \big(z + C_N\big)^{-1},$$

with $$C_N = C_0 + \sum_{n=1}^{N} \omega_n \omega_n^t.$$



Expansion : $G_\rho^{(N)}(z) = G_0(z)$

$$+ \sum_{p=1}^{\infty} (-1)^p G_0(z) \Big[\sum_{n=1}^{N} \omega_n \omega_n^t\Big]^p G_0(z).$$

Dyson equation : $$G_\rho(z) = G_0(z) - \rho G_0(z) \Sigma_\rho G_\rho(z).$$

Free probabilities : $$\Sigma_\rho[G] = M \sum_{n=1}^{\infty} (-1)^{n-1} \mathbb{E}\Big[\omega \big(\omega^t G \omega\big)^{n-1} \omega^t\Big] = M \mathbb{E}\Big[\frac{\omega \omega^t}{1 + \omega^t G \omega}\Big].$$

Spectral decomposition of errors

$$E_{\text{test}}(\rho,\alpha) = \mathcal{R}(\rho,\alpha)^2 E_{\text{train}}(\rho,\alpha).$$

with

$$\mathcal{R}(\rho,\alpha) = 1 + \frac{\alpha}{\rho}\int dx \nu_\infty(x) x g(x,\rho,\alpha),$$

$$E_{\text{train}}(\rho,\alpha) = -\int dx \mu^{\parallel}(x)\frac{\partial}{\partial\alpha} g(x,\rho,\alpha) + (\sigma^2+\mu^{\perp})\Big(1-\rho^{-1}+\frac{\partial}{\partial\alpha}\big[\alpha g(\rho,\alpha)\big]\Big).$$

with

$$g(\rho,\alpha) = \int dx \nu_\infty(x) g(x,\rho,\alpha),$$

$$g(x,\rho,\alpha) = \left(1+\frac{\alpha x}{\mathcal{R}(\rho,\alpha)}\right)^{-1},$$

and

- ν_∞ : spectral density of C
- $(\mu^{\parallel},\mu^{\perp})$: spectral power of signal w.r.t. C modes.

Finite size effects

To understand this deviation consider the ratio

$\rho_D \overset{\text{def}}{=} \rho \dfrac{M(D+1)}{D}$: number of training data points per direction in the embedding space

$\rho_D = 50, 200$ for $\rho = 0.5, 2$ in the experiments.

Empirical identity operator : $\hat{\mathbb{I}} = \dfrac{1}{N_{\text{train}}} \sum_{s=1}^{N_{\text{train}}} \mathbf{x}_s \mathbf{x}_s^t \underset{N_{\text{train}} \to \infty}{\longrightarrow} \mathbb{I}$

Signal as viewed from the training set : $\tilde{f} \overset{\text{def}}{=} \hat{\mathbb{I}} f$

RMT yields : $\|\mathbf{f} - \tilde{\mathbf{f}}\|^2 \approx \dfrac{\|\mathbf{f}\|^2}{\rho_D},$

training process is pointing to $\tilde{\mathbf{f}}$ instead of $\mathbf{f}$.
$\rho_D \propto M$: finite size effect, vanishing in the thermodynamics limits.

Special case : 1-level

Single degenerate level needs to be maintained artificially. Spectral parameters take simple form :

$$g_k = \frac{1}{(1+\Lambda c)^k} = \mu_k = g^k$$

Macroscopic dynamical system :

$$\dot{g} = -\frac{2\alpha^2 r}{\rho^2 M}\frac{(g+\rho-1)^3}{\rho-(1-g)^2}g^3(1-g)\Big[1+\frac{g^2}{\rho-(1-g)^2}\Big]$$

$$\dot{r} = \frac{2\alpha^2}{\rho^2}r(1-r-\sigma^2)g(\rho-1+g)^2,$$

$$E_{\text{test}} = \frac{\rho}{\rho-(1-g)^2}\big[1+r(g^2-1)\big].$$

converges to $g = g_{\min} = \max(0, 1-\rho)$.

Special case : 2-levels

Spectral parameters : $g_k = (1-\nu)y_1^k + \nu y_2^k,$

$\mu_k = (1-\mu)y_1^k + \mu y_2^k,$ with $y_{1,2} \stackrel{\text{def}}{=} \dfrac{1}{1+\Lambda c_{1,2}}$ and ν fixed.

$r \to 1-\sigma^2$ and $\mu \to 1$ fast variables

Macroscopic dynamical system

$$\dot{y}_1 = -\frac{2\Lambda^2 r}{M}\frac{y_1(1-y_1)y_2^2}{\rho-\mathcal{Q}}\Big(y_1^2 - y_2(1-y_2) - \frac{g_3-g_4}{\rho-\mathcal{Q}}\Big),$$

$$\dot{y}_2 = -\frac{2\Lambda^2 r}{M}y_2^3(1-y_2)\Big(\frac{1}{\nu} + \frac{y_2(2y_2-1)}{\rho-\mathcal{Q}} - \frac{g_3-g_4}{(\rho-\mathcal{Q})^2}\Big),$$

with

$$E_{\text{test}}(y_1,y_2) = \frac{\rho\big[1+r(y_2^2-1)\big]}{\rho-(1-\nu)(1-y_1)^2-\nu(1-y_2)^2}.$$

Phase portraits