The Dynamics of Learning: A Random Matrix Approach

ICML 2018, Stockholm, Sweden

Zhenyu Liao, Romain Couillet

L2S, CentraleSupélec, Université Paris-Saclay, France
GSTATS IDEX DataScience Chair, GIPSA-lab, Université Grenoble-Alpes, France.

Outline

1. Motivation
2. Problem Statement
3. Main Results
4. Summary

Motivation

About deep learning:

- Some known facts:
  - trained with backpropagation (gradient decent)
  - has achieved superhuman performance in many applications
  - highly over-parameterized, but some still generalize remarkably well in practice!
- and some (more) mysteries:
  - how do neural networks learn from training data? what features are learned?
  - why they generalize without overfitting? memorize or generalize?
  - can the network performance be guaranteed or ... even predicted?

$\Rightarrow$ The learning dynamics of neural networks!

In particular: under so-called double asymptotic regime (RMT regime):

number of network parameters and number of data instances comparably large!

In this work:

A general RMT framework for studying learning dynamics of a single-layer network!

As a consequence, more insights on:

- random initialization of training
- overfitting in neural networks
- (explicit or implicit) regularization: early stopping, l_2-penalization

Problem Setup

A toy model of binary classification:

Gaussian mixture data

Consider data $\mathbf{x}_i$ drawn from a two-class Gaussian mixture model: for $a = 1, 2$

$$\mathbf{x}_i \in \mathcal{C}_a \Leftrightarrow \mathbf{x}_i = (-1)^a \boldsymbol{\mu} + \mathbf{z}_i$$

with $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$. With label $y_i = -1$ for $\mathcal{C}_1$ and $+1$ for $\mathcal{C}_2$.

Objective: Learning dynamics

Gradient descent on loss $L(\mathbf{w}) = \frac{1}{2n}\|\mathbf{y}^\mathsf{T} - \mathbf{w}^\mathsf{T}\mathbf{X}\|^2$ with $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. For small learning rate α, with continuous-time approximation:

$$\frac{d\mathbf{w}(t)}{dt} = -\alpha \frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = \frac{\alpha}{n}\mathbf{X}\left(\mathbf{y} - \mathbf{X}^\mathsf{T}\mathbf{w}(t)\right)$$

of explicit solution $\mathbf{w}(t) = e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}\mathbf{w}_0 + \left(\mathbf{I}_p - e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}\right)(\mathbf{X}\mathbf{X}^\mathsf{T})^{-1}\mathbf{X}\mathbf{y}$ if $\mathbf{X}\mathbf{X}^\mathsf{T}$ invertible and $\mathbf{w}_0$ the initialization of gradient descent.

- projection of eigenvector weighted by $\exp(-\alpha t \lambda)$ of eigenvalue λ
- functional of sample covariance matrix $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$:

Random Matrix Theory is the answer!

Problem Setup

Objective: Test performance

Test performance for a new $\hat{\mathbf{x}}$:

$$\mathrm{P}(\mathbf{w}(t)^{\mathsf{T}}\hat{\mathbf{x}} > 0 \mid \hat{\mathbf{x}} \in \mathcal{C}_1), \ \mathrm{P}(\mathbf{w}(t)^{\mathsf{T}}\hat{\mathbf{x}} < 0 \mid \hat{\mathbf{x}} \in \mathcal{C}_2).$$

Since $\hat{\mathbf{x}}$ Gaussian and independent of $\mathbf{w}(t)$:

$$\mathbf{w}(t)^{\mathsf{T}}\hat{\mathbf{x}} \sim \mathcal{N}(\pm\mathbf{w}(t)^{\mathsf{T}}\boldsymbol{\mu}, \|\mathbf{w}(t)\|^2)$$

recall $\mathbf{w}(t) = e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}}\mathbf{w}_0 + \left(\mathbf{I}_p - e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}}\right)(\mathbf{X}\mathbf{X}^{\mathsf{T}})^{-1}\mathbf{X}\mathbf{y}$.

With RMT:

- although $\mathbf{X}$ random: $\mathbf{w}(t)^{\mathsf{T}}\boldsymbol{\mu}$ and $\|\mathbf{w}(t)\|^2$ have asymptotically deterministic behavior (only depends on data statistics and dimensions): the technique of **deterministic equivalent**
- **Cauchy's integral formula** to express the functional $\exp(\cdot)$ via contour integration

$\Rightarrow$ Network performance at any time is in fact deterministic and predictable!

Proposed analysis framework

Resolvent and deterministic equivalents

Consider an $n \times n$ Hermitian random matrix $\mathbf{M}$. Define its resolvent $\mathbf{Q}_{\mathbf{M}}(z)$, for $z \in \mathbb{C}$ not eigenvalue of $\mathbf{M}$

$$\mathbf{Q}_{\mathbf{M}}(z) = (\mathbf{M} - z\mathbf{I}_n)^{-1}.$$

For certain simple distributions of $\mathbf{M}$, define a so-called deterministic equivalent $\bar{\mathbf{Q}}_{\mathbf{M}}$ of $\mathbf{Q}_{\mathbf{M}}$: a deterministic matrix such that

- $\frac{1}{n}\operatorname{tr}(\mathbf{A}\mathbf{Q}_{\mathbf{M}}) - \frac{1}{n}\operatorname{tr}\left(\mathbf{A}\bar{\mathbf{Q}}_{\mathbf{M}}\right) \to 0$
- $\mathbf{a}^{\mathsf{T}}\left(\mathbf{Q}_{\mathbf{M}} - \bar{\mathbf{Q}}_{\mathbf{M}}\right)\mathbf{b} \to 0$

almost surely as $n \to \infty$, with $\mathbf{A}, \mathbf{a}, \mathbf{b}$ of bounded norm (operator and Euclidean).

$\Rightarrow$ Study $\bar{\mathbf{Q}}_{\mathbf{M}}$ instead of the random $\mathbf{Q}_{\mathbf{M}}$ for n large!

However, for more sophisticated functionals of $\mathbf{M}$:

Cauchy's integral formula

Example: for $f(\mathbf{M}) = \mathbf{a}^{\mathsf{T}} e^{\mathbf{M}} \mathbf{b}$,

$$f(\mathbf{M}) = -\frac{1}{2\pi i}\oint_{\gamma} \exp(z)\mathbf{a}^{\mathsf{T}}\mathbf{Q}_{\mathbf{M}}(z)\mathbf{b}\,dz \approx -\frac{1}{2\pi i}\oint_{\gamma} \exp(z)\mathbf{a}^{\mathsf{T}}\bar{\mathbf{Q}}_{\mathbf{M}}(z)\mathbf{b}\,dz.$$

with γ a positively oriented path circling around all the eigenvalues of $\mathbf{M}$.

Test performance

To evaluate test performance: $\mathbf{w}(t)^\mathsf{T}\hat{\mathbf{x}} \sim \mathcal{N}(\pm\mathbf{w}(t)^\mathsf{T}\boldsymbol{\mu}, \|\mathbf{w}(t)\|^2)$ with $\mathbf{w}(t) = e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}\mathbf{w}_0 + \left(\mathbf{I}_p - e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}\right)(\mathbf{X}\mathbf{X}^\mathsf{T})^{-1}\mathbf{X}\mathbf{y}$. For $\mathbf{w}(t)^\mathsf{T}\boldsymbol{\mu}$:

- **Cauchy's integral formula**: for $f_t(x) \equiv \exp(-\alpha t x)$,
  $\boldsymbol{\mu}^\mathsf{T}\mathbf{w}(t) = -\frac{1}{2\pi i}\oint_\gamma \boldsymbol{\mu}^\mathsf{T}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\left(f_t(z)\mathbf{w}_0 + \frac{1-f_t(z)}{z}\frac{1}{n}\mathbf{X}\mathbf{y}\right)dz.$
- "replace" the random $\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}$ by its **deterministic equivalent**.

Theorem (Test Performance)

Let $p/n \to c \in (0,\infty)$ and the initialization $\mathbf{w}_0$ be a random vector with i.i.d. entries of zero mean, variance σ^2/p. Then, as $n \to \infty$, with probability one

$$\mathrm{P}(\mathbf{w}(t)^\mathsf{T}\hat{\mathbf{x}} > 0 \mid \hat{\mathbf{x}} \in \mathcal{C}_1) - Q\left(\frac{E}{\sqrt{V}}\right) \to 0, \quad \mathrm{P}(\mathbf{w}(t)^\mathsf{T}\hat{\mathbf{x}} < 0 \mid \hat{\mathbf{x}} \in \mathcal{C}_2) - Q\left(\frac{E}{\sqrt{V}}\right) \to 0$$

for $E \equiv -\frac{1}{2\pi i}\oint_\gamma \frac{1-f_t(z)}{z}\frac{\|\boldsymbol{\mu}\|^2 m(z)\ dz}{(\|\boldsymbol{\mu}\|^2+c)m(z)+1}$, $V \equiv \frac{1}{2\pi i}\oint_\gamma \left[\frac{\frac{1}{z^2}(1-f_t(z))^2}{(\|\boldsymbol{\mu}\|^2+c)m(z)+1} - \sigma^2 f_t^2(z)m(z)\right]dz.$

γ a closed positively oriented path that contains all eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ and the origin, $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp(-u^2/2)du$ and $m(z)$ given by the popular Marčenko–Pastur equation.

Not really understandable, nor interpretable...

Simplification: "break" the contour integration

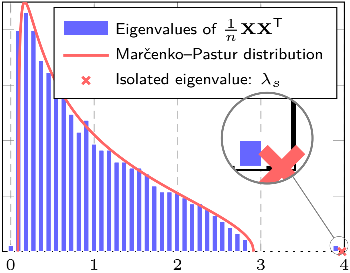

Figure: Eigenvalue distribution of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ for $\boldsymbol{\mu} = [1.5; \mathbf{0}_{p-1}]$, $p = 512$, $n = 1\,024$ and $c_1 = c_2 = 1/2$.

"Main bulk" ($[\lambda_-, \lambda_+]$): sum of real line integrals; isolated eigenvalue (λ_s): residue theorem.

(Simplified) test performance

$$E = \int \frac{1 - f_t(x)}{x} \mu(dx), \; V = \frac{\|\boldsymbol{\mu}\|^2 + c}{\|\boldsymbol{\mu}\|^2} \int \frac{(1 - f_t(x))^2 \mu(dx)}{x^2} + \sigma^2 \int f_t^2(x) \nu(dx)$$

Discussions

(Simplified) test performance

$$E = \int \frac{1 - f_t(x)}{x} \mu(dx), \; V = \frac{\|\boldsymbol{\mu}\|^2 + c}{\|\boldsymbol{\mu}\|^2} \int \frac{(1 - f_t(x))^2 \mu(dx)}{x^2} + \sigma^2 \int f_t^2(x) \nu(dx)$$

where we recall $f_t(x) \equiv \exp(-\alpha t x)$ and the popular Marčenko–Pastur distribution $\nu(dx) \equiv \frac{\sqrt{(x-\lambda_-)^+(\lambda_+-x)^+}}{2\pi c x} dx + \left(1 - \frac{1}{c}\right)^+ \delta(x)$ with $\lambda_- \equiv (1 - \sqrt{c})^2$, $\lambda_+ \equiv (1 + \sqrt{c})^2$ and

$$\mu(dx) \equiv \frac{\sqrt{(x - \lambda_-)^+(\lambda_+ - x)^+}}{2\pi(\lambda_s - x)} dx + \frac{(\|\boldsymbol{\mu}\|^4 - c)^+}{\|\boldsymbol{\mu}\|^2} \delta_{\lambda_s}(x)$$

with $\lambda_s = c + 1 + \|\boldsymbol{\mu}\|^2 + c/\|\boldsymbol{\mu}\|^2$.

Some remarks:

1. $\mu(dx)$: continuous distribution $[\lambda_-, \lambda_+]$ vs. Dirac measure at λ_s: comparable information!
2. $\int \mu(dx) = \|\boldsymbol{\mu}\|^2$ together with Cauchy–Schwarz inequality: $E^2 \leq \int \frac{(1-f_t(x))^2}{x^2} d\mu(x) \cdot \int d\mu(x) \leq \frac{\|\boldsymbol{\mu}\|^4}{\|\boldsymbol{\mu}\|^2 + c} V$, with equality if and only if the (initialization) variance $\sigma^2 = 0$. $\Rightarrow$ in fact performance drop due to large σ^2!
3. How much we over-fit? As $t \to \infty$, the performance drop by a factor $\sqrt{1 - \min(c, c^{-1})}$, with $p/n \to c \in (0, \infty)$.

Numerical validations

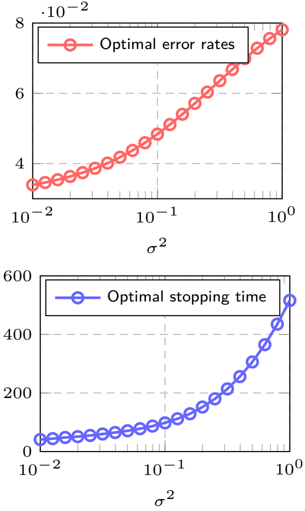

Figure: Optimal performance and stopping time as function of σ^2 with $c = 1/2$, $\|\boldsymbol{\mu}\|^2 = 4$ and $\alpha = 0.01$.

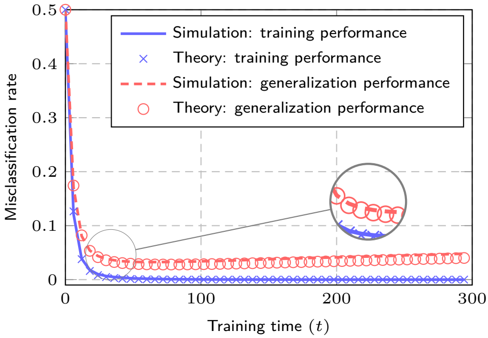

Figure: Training and generalization performance for MNIST data (number 1 and 7) with $n = p = 784$, $c_1 = c_2 = 1/2$, $\alpha = 0.01$ and $\sigma^2 = 0.1$. Results averaged over 100 runs.

Summary

Take-away messages:

- RMT framework to understand and predict learning dynamics:

  Cauchy's integral formula + technique of **deterministic equivalent**

- easily extended to more elaborate data models: e.g., Gaussian mixture model with different means and covariances
- a byproduct: choose the initialization variance σ^2 even smaller!

Thank you!

Any question? Poster # 189!