A Random Matrix Viewpoint of Learning with Gradient Descent

DIMACS Workshop on Randomized Numerical Linear Algebra, Stats, and Optim

Zhenyu Liao, Romain Couillet

CentraleSupélec, Université Paris-Saclay, France
G-STATS IDEX DataScience Chair, GIPSA-lab, Université Grenoble-Alpes, France.

Outline

Motivation: the pitfalls of large dimensional statistics

- Big data era:
  large dimensional and massive amount of data, with huge learning systems;
- # of data instances n, their dimension p and # of system parameters N *all* large;
  - high resolution images $n \leq 10p$: MNIST with $n = 6\,000, p = 784$ and ImageNet with $n = 500\,000, p = 65\,536$ per class;
  - highly over-parameterized deep neural networks $N \gg 10n$: "shallow" LeNet-5 with $N = 60\,000$ and "deep" ResNet-152 with $N = 60\,200\,000$;
- $\boxed{N \gg n \sim p}$.

Figure: Samples from the MNIST database.

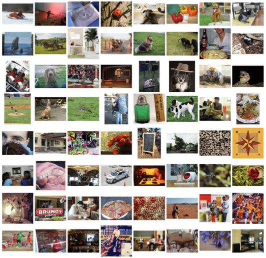


Figure: Samples from the ImageNet database.

Sample covariance matrix in the large n, p regime

- For $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, estimate the covariance matrix from n data samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$.
- Classical maximum likelihood sample covariance matrix:

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T} = \frac{1}{n} \mathbf{X}\mathbf{X}^\mathsf{T} \in \mathbb{R}^{p \times p}$$

  of rank at most n, "optimal" if $n \gg p$.
- In the regime where $n \sim p$, conventional wisdom breaks down, for $\mathbf{C} = \mathbf{I}_p$ with $n < p$, SCM will never be consistent:

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty}$$

  with at least $p - n$ zero eigenvalues (eigenvalue mismatch)!
- Typically what happens in deep learning: try to fit an enormous statistical model (60.2 M of ResNet-152) with insufficient, but still numerous data (total 14.2 M images of ImageNet dataset).

When is one under the random matrix regime?

What about $n = 100p$? Recall $n \sim 10p$ for MINST and ImageNet.
For $\mathbf{C} = \mathbf{I}_p$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$: the Marčenko–Pastur law

$$\mu(dx) = (1 + c^{-1})^{+}\delta(x) + \frac{1}{2\pi c x}\sqrt{(x - \lambda_-)^{+}(\lambda_+ - x)^{+}}dx$$

where $\lambda_- = (1 - \sqrt{c})^2$, $\lambda_+ = (1 + \sqrt{c})^2$ and $(x)^+ \equiv \max(x, 0)$.

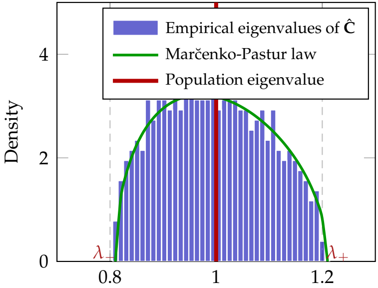


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

- eigenvalues span on $[(1-\sqrt{\mathbf{c}})^2, (1+\sqrt{\mathbf{c}})^2]$.
- for $\mathbf{n} = \mathbf{100p}$, spread on a range of $\mathbf{4}\sqrt{\mathbf{c}} = \mathbf{0.4}$ around the *population* eigenvalue $\mathbf{1}$.

Motivation: about deep learning

Some known facts:

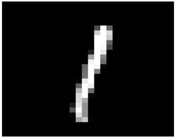
- trained with backpropagation (*gradient decent*);
- achieved *superhuman* performance in many applications;
- "generalization mystery": highly *over-parameterized* ($N \gg n \sim p$), some still generalize remarkably well;

In this work:

- Why over-parameterization does not harm generalization?
- What is the role played by gradient descent?
- $\Rightarrow$ A *general* RMT framework for **gradient descent dynamics** of simple nets!
- **Conclusion**:

  both over-parameterization and gradient descent are important for generalization!

Objective: predict the performance of simple neural nets



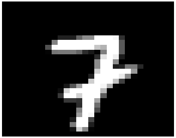


Figure: Example of MNIST images

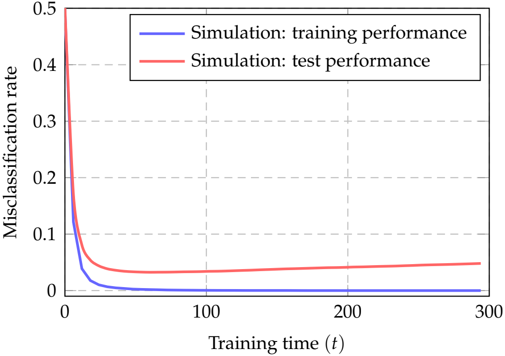


Figure: Training and test performance for MNIST data with a learning rate $\alpha = 0.01$. Results averaged over 100 runs.

A toy model of binary classification

Gaussian mixture data

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

$$\mathbf{x}_i \in \mathcal{C}_a \Leftrightarrow \mathbf{x}_i = \boldsymbol{\mu}_a + \mathbf{C}_a^{\frac{1}{2}} \mathbf{z}_i$$

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

Gradient descent dynamics

Gradient descent to minimize $\ell(\mathbf{w}) = \frac{1}{2n} \|\mathbf{y}^\mathsf{T} - \mathbf{w}^\mathsf{T} \mathbf{X}\|^2$ with $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$.
For small learning rate α, gradient flow given by

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

of explicit solution

$$\mathbf{w}(t) = e^{-\frac{\alpha t}{n} \mathbf{X}\mathbf{X}^\mathsf{T}} \mathbf{w}_0 + \left(\mathbf{I}_p - e^{-\frac{\alpha t}{n} \mathbf{X}\mathbf{X}^\mathsf{T}} \right) \underbrace{\boxed{(\mathbf{X}\mathbf{X}^\mathsf{T})^{-1} \mathbf{X}\mathbf{y}}}_{\equiv \mathbf{w}_{LS}}$$

if $\mathbf{X}\mathbf{X}^\mathsf{T}$ invertible and $\mathbf{w}_0$ the initialization.

Key object:

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

For symmetric $\mathbf{A} \in \mathbb{R}^{p\times p}$ with spectral decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}_\mathbf{A}\mathbf{U}^\mathsf{T} = \sum_{i=1}^p \lambda_i(\mathbf{A})\mathbf{u}_i\mathbf{u}_i^\mathsf{T}$,

$$e^{\mathbf{A}} = \mathbf{U}e^{\mathbf{\Lambda}_\mathbf{A}}\mathbf{U}^\mathsf{T} = \sum_{i=1}^p e^{\lambda_i(\mathbf{A})}\mathbf{u}_i\mathbf{u}_i^\mathsf{T}.$$

- projection of eigenvector weighted by $\exp(-\alpha t\lambda)$ of eigenvalue λ;
- functional of the sample covariance-type matrix $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$;
- Random Matrix Theory (RMT) provides an answer!

RMT for gradient descent dynamics

Objective: Test performance

Test performance for a new $\hat{\mathbf{x}} \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$:

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

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

$$\mathbf{w}(t)^\mathsf{T}\hat{\mathbf{x}} \mid \mathbf{w}(t) \sim \mathcal{N}(\mathbf{w}(t)^\mathsf{T}\boldsymbol{\mu}_a, \mathbf{w}(t)^\mathsf{T}\mathbf{C}_a\mathbf{w}(t))$$

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

With RMT:

- **Cauchy's integral formula** to express the functional $e^{(\cdot)}$ via contour integration;
- for random $\mathbf{X}$, both $\mathbf{w}(t)^\mathsf{T}\boldsymbol{\mu}_a$ and $\mathbf{w}(t)^\mathsf{T}\mathbf{C}_a\mathbf{w}(t)$ have tractable **asymptotically deterministic**[1] behavior: **deterministic equivalent** technique;
- $\Rightarrow$ Performance at any time is asymptotically deterministic and predictable!

[1] that only depends on data statistics and the problem dimension.

Proposed RMT analysis framework: Cauchy's integral formula

Consider $\mu_a^\mathsf{T} \mathbf{w}(t) = \mu_a^\mathsf{T} e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}} \mathbf{w}_0 + \mu_a^\mathsf{T} \left(\mathbf{I}_p - e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}} \right) \mathbf{w}_{LS}$.

Cauchy's integral formula

For $\Gamma \in \mathbb{C}$ a positively (i.e., counterclockwise) oriented simple closed curve and a complex function $f(z)$ analytic in a region containing Γ and its interior, then $-\frac{1}{2\pi\imath} \oint_\Gamma \frac{f(z)}{z_0 - z} dz = f(z_0)$ if $z_0 \in \mathbb{C}$ is enclosed by Γ and 0 otherwise.

$$f(\mathbf{A}) = \mathbf{a}^\mathsf{T} e^{\mathbf{A}} \mathbf{b} = \sum_{i=1}^p \mathbf{a}^\mathsf{T} \left(e^{\lambda_i(\mathbf{A})} \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \right) \mathbf{b} \qquad \text{(spectral decomposition of } \mathbf{A})$$

$$= \sum_{i=1}^p \mathbf{a}^\mathsf{T} \left(-\frac{1}{2\pi\imath} \oint_\Gamma \frac{\exp(z)\, dz}{\lambda_i(\mathbf{A}) - z} \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \right) \mathbf{b} \qquad \text{(Cauchy's integral formula for } \exp(\cdot))$$

$$= -\frac{1}{2\pi\imath} \oint_\Gamma \exp(z) \mathbf{a}^\mathsf{T} \mathbf{Q}_\mathbf{A}(z) \mathbf{b}\, dz \qquad \left(\mathbf{Q}_\mathbf{A}(z) \equiv (\mathbf{A} - z\mathbf{I}_p)^{-1} = \sum_{i=1}^p \frac{\mathbf{u}_i \mathbf{u}_i^\mathsf{T}}{\lambda_i(\mathbf{A}) - z} \right)$$

with $\mathbf{Q}_\mathbf{A}(z)$ the resolvent of $\mathbf{A}$ for $z \in \mathbb{C}$ not eigenvalue of $\mathbf{A}$ and Γ positively enclosed all eigenvalues of $\mathbf{A}$.

[2] Technical remark: no worries about *branch cut* with the exponential function $e^{(\cdot)}$, attention with other functions such as the complex $\log(\cdot)$.

Proposed RMT analysis framework: deterministic equivalent technique

$$f(\mathbf{A}) = \mathbf{a}^\mathsf{T} e^{\mathbf{A}} \mathbf{b} = -\frac{1}{2\pi\imath} \oint_\Gamma \exp(z) \mathbf{a}^\mathsf{T} \mathbf{Q}_\mathbf{A}(z) \mathbf{b} \, dz.$$

Resolvent and deterministic equivalents

For symmetric random matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$, define its resolvent $\mathbf{Q}_\mathbf{A}(z)$, for $z \in \mathbb{C}$ not eigenvalue of $\mathbf{A}$, as

$$\mathbf{Q}_\mathbf{A}(z) = \left(\mathbf{A} - z\mathbf{I}_p\right)^{-1}.$$

For a large family of random $\mathbf{A}$, we note $\bar{\mathbf{Q}}_\mathbf{A} \leftrightarrow \mathbf{Q}_\mathbf{A}$ and say the deterministic matrix $\bar{\mathbf{Q}}_\mathbf{A}$ is a deterministic equivalent of $\mathbf{Q}_\mathbf{A}$ if

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

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

$\Rightarrow$ To treat $\bar{\mathbf{Q}}_\mathbf{A}$ instead of the random $\mathbf{Q}_\mathbf{A}$ for n, p large!

In particular, $f(\mathbf{A}) = -\frac{1}{2\pi\imath} \oint_\Gamma \exp(z) \mathbf{a}^\mathsf{T} \mathbf{Q}_\mathbf{A}(z) \mathbf{b} \, dz \simeq -\frac{1}{2\pi\imath} \oint_\Gamma \exp(z) \mathbf{a}^\mathsf{T} \bar{\mathbf{Q}}_\mathbf{A}(z) \mathbf{b} \, dz.$

Intuition behind deterministic equivalent: concentration phenomena

Example: Gaussian concentration inequality

For Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and α-Lipschitz function $f : \mathbb{R}^p \mapsto \mathbb{R}$, then

$$P\left(|f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})]| \geq t\right) \leq 2e^{-t^2/(2\alpha^2)}$$

- dimension free in the case of single Gaussian random vector
- add a factor $n \sim p$ for (the joint behavior of cols of) random matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$
- for a random matrix $\mathbf{A} = \frac{1}{n}\mathbf{X}\mathbf{X}$ and its resolvent $\mathbf{Q}_\mathbf{A}$, the Lipschitz function $\mathbf{a}^\mathsf{T}\mathbf{Q}_\mathbf{A}\mathbf{b}$ concentrate around its expectation $\mathbf{a}^\mathsf{T}\mathbb{E}[\mathbf{Q}_\mathbf{A}]\mathbf{b}$ with high probability
- often $\mathbb{E}[\mathbf{Q}_\mathbf{A}]$ not easily accessible (high dimensional integral), but admits an asymptotic equivalent $\|\mathbb{E}[\mathbf{Q}_\mathbf{A}] - \bar{\mathbf{Q}}_\mathbf{A}\| \to 0$ as $n, p \to \infty$
- $\Rightarrow \mathbf{a}^\mathsf{T}\left(\mathbf{Q}_\mathbf{A} - \bar{\mathbf{Q}}_\mathbf{A}\right)\mathbf{b} \to 0$ almost surely as $n, p \to \infty$

A Central Limit Theorem

To evaluate test performance: $\boxed{\mathbf{w}(t)^\mathsf{T}\hat{\mathbf{x}} \mid \mathbf{w}(t) \sim \mathcal{N}(\mathbf{w}(t)^\mathsf{T}\boldsymbol{\mu}_a, \mathbf{w}(t)^\mathsf{T}\mathbf{C}_a\mathbf{w}(t))}$, with $\mathbf{w}(t) = e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}\mathbf{w}_0 + \left(\mathbf{I}_p - e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}\right)\mathbf{w}_{LS}$. For $\mathbf{w}(t)^\mathsf{T}\boldsymbol{\mu}_a$:

1. With **Cauchy's integral formula**

$$
\begin{aligned}
\boldsymbol{\mu}_a^\mathsf{T}\mathbf{w}(t) &= -\frac{1}{2\pi\imath}\oint_\Gamma \boldsymbol{\mu}_a^\mathsf{T}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\left(\exp(-\alpha t z)\mathbf{w}_0 + \frac{1-\exp(-\alpha t z)}{z}\frac{1}{n}\mathbf{X}\mathbf{y}\right)dz \\
&= -\frac{1}{2\pi\imath}\oint_\Gamma \boldsymbol{\mu}_a^\mathsf{T}\mathbf{Q}_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}(z)\left(\mathbf{w}_0\exp(-\alpha t z) + \frac{1-\exp(-\alpha t z)}{z}\frac{1}{n}\mathbf{X}\mathbf{y}\right)dz
\end{aligned}
$$

2. "replace" the random resolvent $\mathbf{Q}_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}(z)$ with its **deterministic equivalent** $\bar{\mathbf{Q}}(z)$.

To reach a CLT for $\mathbf{w}(t)^\mathsf{T}\hat{\mathbf{x}}$ of type

Generic result: asymptotic Gaussianity for $\mathbf{w}(t)^\mathsf{T}\hat{\mathbf{x}}$

For an independent test datum $\hat{\mathbf{x}} \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, the soft output $\mathbf{w}(t)^\mathsf{T}\hat{\mathbf{x}} - h_a(t) \to 0$ in distribution as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$,

$$h_a(t) \sim \mathcal{N}(E_a(t), V_a(t))$$

where $E_a(t)$ and $V_a(t)$ are given by contour integrals and depend on data statistics $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{C}_1, \mathbf{C}_2)$, gradient descent initialization $\mathbf{w}_0$ and the problem dimension n, p.

A more interpretable case: $\mathbf{C}_a = \mathbf{I}_p$

For *generic* covariance $\mathbf{C}_a$, the deterministic equivalent $\bar{\mathbf{Q}}(z)$ has no-closed form and is characterized via a system of fixed point equations[2], e.g., for centered $\mathbf{X}$,

$$\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1} \equiv \mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z) = \left(\sum_{a=1}^{2} \frac{\pi_a \mathbf{C}_a}{1+g_a(z)} - z\mathbf{I}_p\right)^{-1}, g_a(z) = \frac{1}{n}\operatorname{tr} \mathbf{C}_a \bar{\mathbf{Q}}_(z)$$

with π_a the prior probability of class $\mathcal{C}_a$ and $g_a(z)$ the unique solution of the equation.

Marčenko–Pastur equation

In the special case of $\mathbf{C}_a = \mathbf{I}_p$, closed-form solution:

$$\bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p$$

with $m(z)$ (also known as the Stieltjes transform of $\mu_{\mathbf{X}\mathbf{X}^\mathsf{T}/n}$, the spectral measure of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$) given by the Marčenko–Pastur equation such that $\Im[z] \cdot \Im[m(z)] > 0$,

$$zcm^2(z) - (1-c-z)m(z) + 1 = 0 \Rightarrow m(z) = \frac{1-c-z}{2cz} \pm \frac{\sqrt{(1-c-z)^2 - 4cz}}{2cz}$$

with c denotes (the limit of) the ratio p/n.

[2] Florent Benaych-Georges and Romain Couillet. "Spectral analysis of the Gram matrix of mixture models". In: *ESAIM: Probability and Statistics* 20 (2016), pp. 217–237, p. 3.

Special case: $\mathbf{C}_a = \mathbf{I}_p$

Theorem: asymptotic Gaussianity for $\mathbf{w}(t)^\mathsf{T}\hat{\mathbf{x}}$

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, for an independent test datum $\hat{\mathbf{x}} \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$, the soft output $\mathbf{w}(t)^\mathsf{T}\hat{\mathbf{x}} - \pm h(t) \to 0$ in distribution as $n, p \to \infty$, with

$$h(t) \sim \mathcal{N}(E(t), V(t))$$

where

$$E(t) = -\frac{1}{2\pi\imath} \oint_\Gamma \frac{1 - e^{-\alpha z t}}{z} \frac{\|\boldsymbol{\mu}\|^2 m(z)\, dz}{(\|\boldsymbol{\mu}\|^2 + c)\, m(z) + 1}$$

$$V(t) = \frac{1}{2\pi\imath} \oint_\Gamma \left[\frac{\frac{1}{z^2}\left(1 - e^{-\alpha z t}\right)^2}{(\|\boldsymbol{\mu}\|^2 + c)\, m(z) + 1} - \sigma^2 e^{-2\alpha z t} m(z) \right] dz$$

for Γ a positively oriented contour that encloses $\cup_{n=1}^{\infty} \operatorname{supp}(\mu_{\mathbf{X}\mathbf{X}^\mathsf{T}/n})$, the spectral measure of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ (know to be almost surely compact).

Practical consequence

Corollary: asymptotic test performance

For a decision threshold $\xi = 0$, we have

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

with standard Gaussian tail function $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp(-u^2/2)du$ and

$$E(t) = -\frac{1}{2\pi\imath}\oint_\Gamma \frac{1-e^{-\alpha z t}}{z}\frac{\|\boldsymbol{\mu}\|^2 m(z)\,dz}{(\|\boldsymbol{\mu}\|^2+c)\,m(z)+1}$$

$$V(t) = \frac{1}{2\pi\imath}\oint_\Gamma \left[\frac{\frac{1}{z^2}\left(1-e^{-\alpha z t}\right)^2}{(\|\boldsymbol{\mu}\|^2+c)\,m(z)+1} - \sigma^2 e^{-2\alpha z t} m(z)\right] dz$$

Not really understandable, nor interpretable…

"Break" the contour integration

- we know (almost surely) the "location" of the eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$;
- and we are "free" to choose the contour Γ!

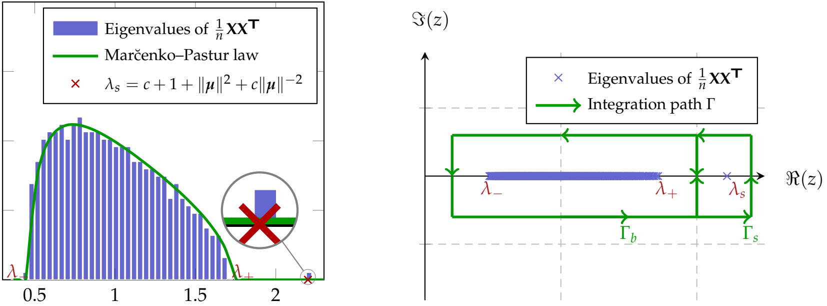


Figure: Eigenvalue distribution of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ for $\boldsymbol{\mu} = [1; \mathbf{0}_{p-1}]$, $p = 500$, $n = 5\,000$ and $\pi_1 = \pi_2 = \frac{1}{2}$.

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

Asymptotic test performance in more compact form

Corollary: (simplified) asymptotic test performance

For a decision threshold $\xi = 0$, we have

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

where

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

with $\mu(dx)$ the Marčenko–Pastur law $\mu(dx) \equiv \frac{\sqrt{(x-\lambda_-)^+(\lambda_+ - x)^+}}{2\pi c x} dx + (1-c^{-1})^+\delta(x)$, $\lambda_- = (1-\sqrt{c})^2$, $\lambda_+ = (1+\sqrt{c})^2$ and

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

for $\lambda_s = c + 1 + \|\boldsymbol{\mu}\|^2 + c\|\boldsymbol{\mu}\|^{-2}$.

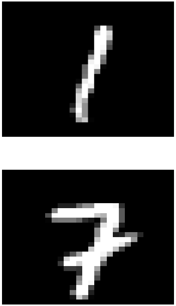


Figure: Example of MNIST images

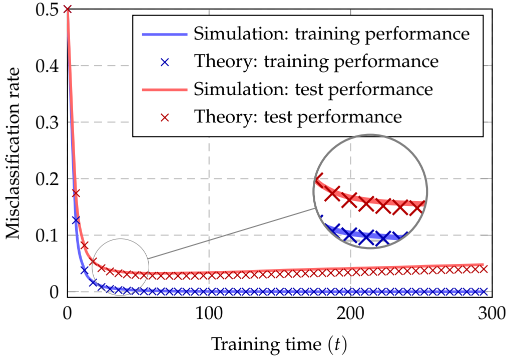


Figure: Training and test performance on MNIST data (number 1 and 7) with $n = p = 784$, $c_1 = c_2 = 1/2$, $\alpha = 0.01$ and $\sigma^2 = 1$. Results averaged over 100 runs.

Discussions on overfitting

$$E(t) = \int \frac{1-\exp(-\alpha x t)}{x} \nu(dx)$$

$$V(t) = \frac{\|\boldsymbol{\mu}\|^2 + c}{\|\boldsymbol{\mu}\|^2} \int \frac{(1-\exp(-\alpha x t))^2 \nu(dx)}{x^2} + \sigma^2 \int \exp(-2\alpha x t) \mu(dx)$$

Optimal performance

With $\int \nu(dx) = \|\boldsymbol{\mu}\|^2$ and Cauchy–Schwarz inequality:

$$\frac{E(t)}{\sqrt{V(t)}} \leq \sqrt{\frac{\int \frac{(1-\exp(-\alpha x t))^2}{x^2} \nu(dx) \cdot \int \nu(dx)}{V(t)}} \leq \frac{\|\boldsymbol{\mu}\|^2}{\sqrt{\|\boldsymbol{\mu}\|^2 + c}}$$

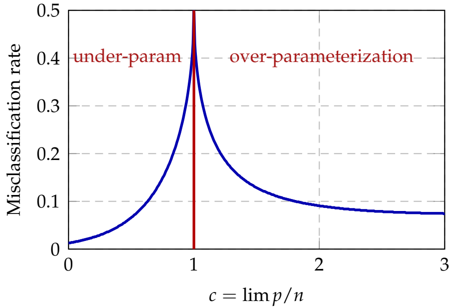
Overfitting and generalization

As $t \to \infty$, we obtain the *least squares solution* ($\mathbf{w}_{LS}$) and

$$\frac{E(\infty)}{\sqrt{V(\infty)}} = \frac{\|\boldsymbol{\mu}\|^2}{\sqrt{\|\boldsymbol{\mu}\|^2 + c}} \boxed{\sqrt{1-\min(c, c^{-1})}}$$

with $p/n \to c \in (0, \infty)$, i.e., the performance drop by a factor of $\sqrt{1-\min(c, c^{-1})}$.

The benefit of over-parametrization



For least squares solution $\mathbf{w}_{LS}$:

$$\frac{E(\infty)}{\sqrt{V(\infty)}} = \frac{\|\boldsymbol{\mu}\|^2}{\sqrt{\|\boldsymbol{\mu}\|^2 + c}} \boxed{\sqrt{1 - \min(c, c^{-1})}}$$

Figure: Classification error rate as a function of c, $\|\boldsymbol{\mu}\|^2 = 5$.

- performance contains a singularity at $p = n$!
- in this case the number of system parameters $N = p$
- for a given training set of size n, performance increase when the model gets over-parameterized ($N \uparrow$)
- similar phenomena are proved/observed for model involved models
- an argument to explain why highly over-parametrized neural nets generalize well

Take-away message: the benefit of learning with gradient descent

- for $\mathbf{w}_{LS}$: $\frac{E(\infty)}{\sqrt{V(\infty)}} = \frac{\|\boldsymbol{\mu}\|^2}{\sqrt{\|\boldsymbol{\mu}\|^2+c}} \boxed{\sqrt{1-\min(c,c^{-1})}}$, singularity at $c = 1$
- in this work, we show for any $t < \infty$, $\frac{E(t)}{\sqrt{V(t)}}$ is a smooth function of c
- $\Rightarrow$ no performance drop at $c = 1$ with "early" stopping!
- an argument to explain why gradient-based deep neural nets generalize well
  - ✔ holds for the misclassification rate in classification of Gaussian mixtures
  - ✔ holds for prediction risk in a (ridge) regression context
  - ✔ extends to nonlinear systems, e.g., nonlinear random feature-based models
  - ? convex optimization problems with no closed-form solution, e.g., logistic regression
  - ? non-convex models are mode involved, but of more practical interest

Related works

Some references and related works:

- Zhenyu Liao and Romain Couillet. "The Dynamics of Learning: A Random Matrix Approach". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. PMLR, 2018, pp. 3072–3081
- Madhu S Advani and Andrew M Saxe. "High-dimensional dynamics of generalization error in neural networks". In: *arXiv preprint arXiv:1710.03667* (2017)
- Stefano Spigler et al. "A jamming transition from under-to over-parametrization affects loss landscape and generalization". In: *arXiv preprint arXiv:1810.09665* (2018)
- Tengyuan Liang and Alexander Rakhlin. "Just interpolate: Kernel" ridgeless" regression can generalize". In: *arXiv preprint arXiv:1808.00387* (2018)
- Mikhail Belkin, Daniel J Hsu, and Partha Mitra. "Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2300–2311
- Mikhail Belkin et al. "Reconciling modern machine learning and the bias-variance trade-off". In: *arXiv preprint arXiv:1812.11118* (2018)
- Trevor Hastie et al. "Surprises in High-Dimensional Ridgeless Least Squares Interpolation". In: *arXiv preprint arXiv:1903.08560* (2019)

and many many more …

Thank you!