Escaping mediocrity: how two-layer networks learn hard generalized linear models

Luca Arnaboldi[1], Florent Krzakala[1], Bruno Loureiro[2], Ludovic Stephan[1]

[1] IdePHICS, EPFL; [2] ENS Paris

Context & Motivation

The study of *generalized linear models*

$$y = \sigma_\star\left(w_\star^\top x\right) + \sqrt{\Delta} z,$$

with Gaussian data $x \sim \mathcal{N}(0, 1/d I_d)$, $z \sim \mathcal{N}(0,1)$ has been developed recently, leading to the following results:

- for matching activation function, the sample complexity of one-pass SGD is determined by the first non-zero Hermite coefficient of the target $\sigma_\star$, also known as the *information exponent* [2];
- wide two-layer networks can achieve the well-specified sample complexity of $n = O(d)$ under one-pass SGD, *provided that all Hermite coefficients of both $\sigma_\star, \sigma$ are non-zero* (IE$=1$)[3].

Aim: *Compute the exact convergence rate of SGD for this class of models.*

Setting

The exact model we are going to study is the following:

- Input data is generated from independent Gaussian distributions:

$$\boldsymbol{x}^\nu \sim \mathcal{N}\left(\mathbf{0}_d, \frac{1}{d}\mathrm{I}_d\right)$$

Labels are generated by

$$y = \left(w_\star^\top x\right)^2 + \sqrt{\Delta} z, \qquad w_\star \in \mathbb{S}^{d-1}(\sqrt{d})$$

where Δ is the artificial noise.

- We are training a two-layer network with **square activations**:

$$f_\Theta(x) = \frac{1}{p}\sum_{i=1}^{p} a_i (w_i^\top x)^2 \qquad w_j \in \mathbb{S}^{d-1}(\sqrt{d})$$

- We are using the **square loss function**. The population risk is given by:

$$\mathcal{R}(\Theta) \coloneqq \mathbb{E}_{(\boldsymbol{x},y)\sim\rho}\left[\frac{1}{2}(f_\Theta(\boldsymbol{x}) - y)^2\right] + \frac{\Delta}{2}$$

- We consider both **standard** & **projected online SGD**:

$$\boldsymbol{w}_j^{\nu+1} = \frac{\boldsymbol{w}_j^\nu - \gamma\nabla_{\boldsymbol{w}_j}\ell(y^\nu, f_{\Theta^\nu}(x^\nu))}{\left\|\boldsymbol{w}_j^\nu - \gamma\nabla_{\boldsymbol{w}_j}\ell(y^\nu, f_{\Theta^\nu}(x^\nu))\right\|}\sqrt{d}$$

High dimensional limit ODE description

We can introduce the following *sufficient statistics*:

$$\Omega^\nu \coloneqq \begin{pmatrix} \mathrm{Q}^\nu & m^\nu \\ m^{\nu\top} & \rho \end{pmatrix} = \frac{1}{d}\begin{pmatrix} \mathrm{W}^\nu \mathrm{W}^{\nu\top} & \mathrm{W}^\nu w^{\star\top} \\ w^\star \mathrm{W}^{\nu\top} & w^\star w^{\star\top} \end{pmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}$$

We can derive a closed set of stochastic processes

$$\begin{aligned}
a_j^{\nu+1} - a_j^\nu &= \frac{\gamma}{pd}\mathcal{E}^\nu \lambda_j^2 \\
m_j^{\nu+1} - m_j^\nu &\eqqcolon \mathcal{M}_j(a, \lambda_\star, \lambda) = 2\frac{\gamma}{pd}\mathcal{E}^\nu a_j \lambda_j \lambda_\star \\
Q_{jl}^{\nu+1} - Q_{jl}^\nu &\eqqcolon \mathcal{Q}_{jl}(a, \lambda_\star, \lambda) = 2\frac{\gamma}{pd}\mathcal{E}^\nu\left(a_j + a_l\right)\lambda_j\lambda_l \\
&\quad + 4\frac{\gamma^2}{p^2 d}\mathcal{E}^{\nu 2}||x^\nu||^2 a_j a_l \lambda_j \lambda_l
\end{aligned}$$

where the *local fields* are jointly Gaussian vectors

$$(\boldsymbol{\lambda}^\nu, \boldsymbol{\lambda}^{\star\nu}) \sim \mathcal{N}(\mathbf{0}_{p+k}, \Omega^\nu).$$

Informally, when $\frac{\gamma}{pd} \to 0^+$ there is ODEs approximation

$$\begin{aligned}
\frac{\mathrm{d}\bar{a}_j}{\mathrm{d}t} &= \mathbb{E}_{(\lambda,\lambda_\star)\sim\mathcal{N}(0_{p+1},\Omega)}\left[\mathcal{E}\lambda_j^2\right] \\
\frac{\mathrm{d}\bar{m}_j}{\mathrm{d}t} &= \mathbb{E}_{(\lambda,\lambda_\star)\sim\mathcal{N}(0_{p+1},\Omega)}\left[\mathcal{M}_j(a, \lambda_\star, \lambda)\right] \eqqcolon \Psi_j(\Omega) \\
\frac{\mathrm{d}\bar{Q}_{jl}}{\mathrm{d}t} &= \mathbb{E}_{(\lambda,\lambda_\star)\sim\mathcal{N}(0_{p+1},\Omega)}\left[\mathcal{Q}_{jl}(a, \lambda_\star, \lambda)\right] \eqqcolon \Phi_{jl}(\Omega)
\end{aligned}$$

Projected SGD The modified equations for the spherical constraint are

$$\begin{aligned}
\frac{\mathrm{d}\bar{m}_j}{\mathrm{d}t} &= \Psi_j(\Omega) - \frac{\bar{m}_j}{2}\Phi_{jj}(\Omega), \\
\frac{\mathrm{d}\bar{Q}_{jl}}{\mathrm{d}t} &= \Phi_{jl}(\Omega) - \frac{\bar{Q}_{jl}}{2}\left(\Phi_{jj}(\Omega) + \Phi_{ll}(\Omega)\right).
\end{aligned}$$

Note that $\mathrm{Q}_{jj} = 1$ is consistently fixed.

Escaping mediocrity at initialization

In the absence of knowledge of the process that generated the data, it is customary to initialize the weights randomly:

$$w_j^0 \sim \mathcal{N}(0, I_d), \qquad j = 1, \cdots, p.$$

In high-dimesnion, this means $w_j \perp w_l \perp w_\star$. In terms of the sufficient statistics, this corresponds to

$$\mathrm{Q}_{jj} \sim \text{Dirac}(1), \quad j \neq l: \ \sqrt{d}\, \mathrm{Q}_{jl}^0 \xrightarrow{d\to+\infty} \mathcal{N}(0,1)$$
$$\sqrt{d}\, m_j^0 \xrightarrow{d\to+\infty} \mathcal{N}(0,1)$$

Needle in the haystack: the proliferation of flat directions close to initialization severely slows down the SGD dynamics at high-dimensions; the starting point is a fixed point of the ODEs.

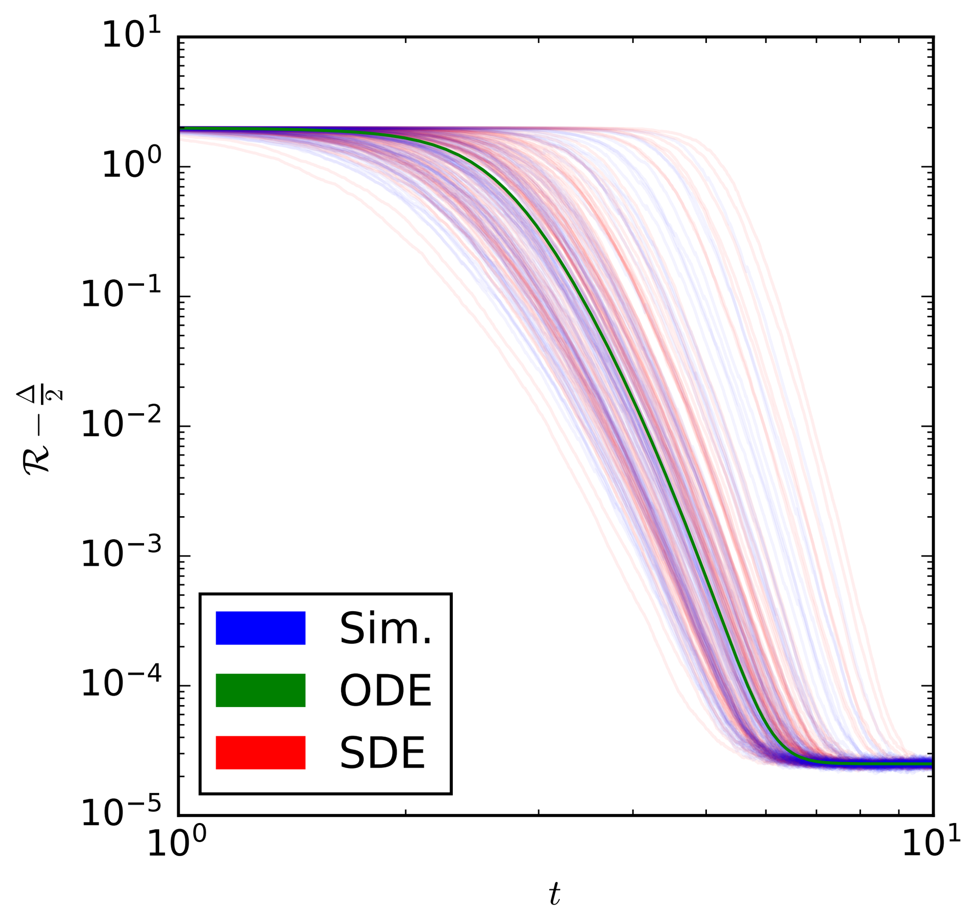
Escaping mediocrity in the well-specified scenario

Given that $p = 1$ there is one single parameter describing the system:

$$m \equiv \frac{w^\top w^\star}{d}.$$

The ODE and the risk are written as

$$\frac{\mathrm{d}\bar{m}(t)}{\mathrm{d}t} = \bar{m}(t)\left[4(1-6\gamma)(1-\bar{m}^2(t)) - 2\gamma\Delta\right]$$
$$\mathcal{R}(\bar{m}) = 2\left(1 - \bar{m}^2\right) + \frac{\Delta}{2}$$



Some *nice* side effects:

- **Bounds on the learning rate**: it must be in the range $0 < \gamma < 1/6$.
- **Minimal risk reached** Fixed point of the equation:

$$\lim_{t\to\infty}\mathcal{R}(\bar{m}(t)) - \Delta/2 = \frac{\gamma\Delta}{1 - 6\gamma}$$

Measuring the escaping time

Compute the time t_{ext} needed to reach a given threshold T

$$(1-T)\left(\mathcal{R}(\bar{m}(0)) - \frac{\Delta}{2}\right) = \left(\mathcal{R}\left(\bar{m}\left(t_{\text{ext}}\right)\right) - \frac{\Delta}{2}\right). \qquad \text{(EXT)}$$

We can average the solution over the initial condition:

- *before* solving, **annealed formula**

$$t_{\text{ext}}^{(\text{anl})} = \frac{\log\left[Td + (1-T)\right]}{8(1-6\gamma) - 4\gamma\Delta}$$

- *after* solving, **quenched formula**

$$t_{\text{ext}}^{(\text{qnc})} = \mathbb{E}_{\mu_0\sim\chi^2(1)}\left[\frac{\log\left[\frac{Td}{\mu_0} + (1-T)\right]}{8(1-6\gamma) - 4\gamma\Delta}\right]$$

We arrive at the following conclusions:

- By concavity of the logarithm function, we have $t_{\text{ext}}^{(\text{qnc})} \geq t_{\text{ext}}^{(\text{anl})}$.
- $t_{\text{ext}} = O(\log d) \implies n = O(d \log d)$ as in [2].
- There exist an **optimal learning rate**:

$$\gamma_{\text{opt}} = \frac{1}{12 + \Delta}$$

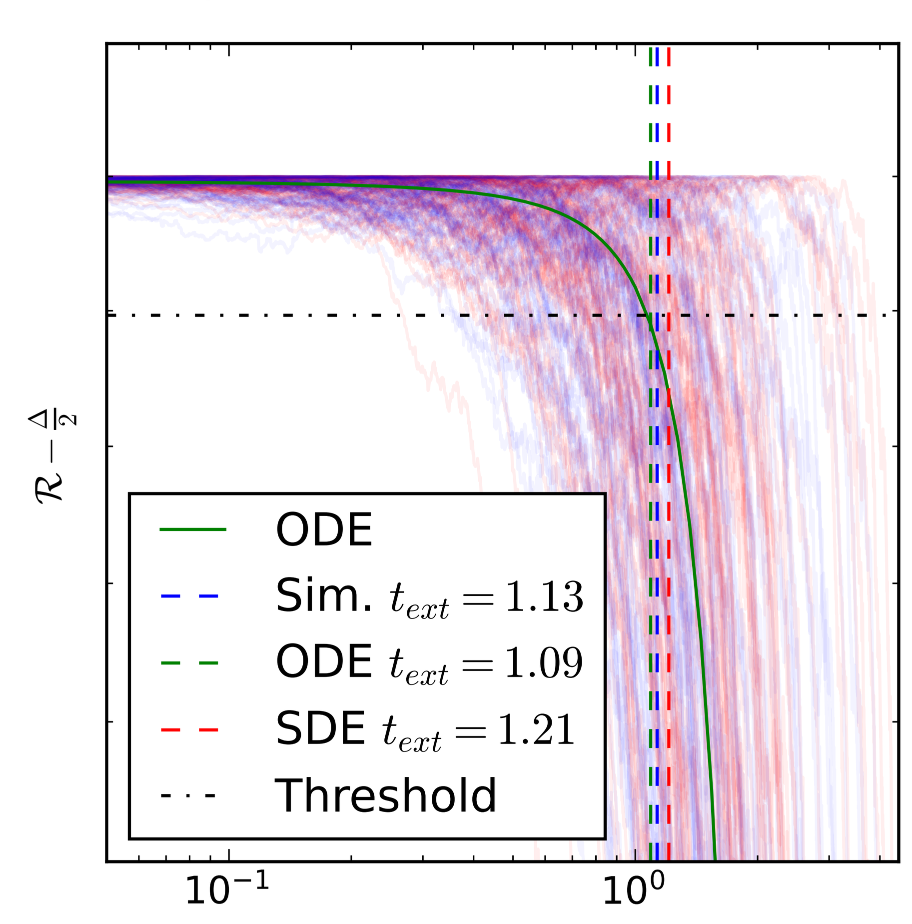
This minimizes the escaping time but not the time to learn nor the minimal risk.

Does stochasticity help?

Add the first correction to the expected value of the ODE

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \left(\Psi_1(\Omega) - \frac{m}{2}\Phi_{11}(\Omega)\right)\mathrm{d}t + \sqrt{\frac{\gamma}{d}}\left(\boldsymbol{\sigma}_m - \frac{m}{2}\boldsymbol{\sigma}_{\mathrm{Q}}\right)\cdot \mathrm{d}B_t$$

where $\boldsymbol{\sigma}_m$ and $\boldsymbol{\sigma}_{\mathrm{Q}}$ are the standard deviations of $\mathcal{M}$ and $\mathcal{Q}$.



Take-homes:

- The SDE grasps the sample stochasticity of the SGD dynamics.
- **The exit time is not affected by the stochasticity** though.

Wide networks

Eq. (EXT) is valid for any $p \geq 1$. We can derive again the two formulae for the escaping time:

- **annealed formula**

$$t_{\text{ext}}^{(\text{anl})} = \frac{\log\left[\frac{T(p+1)d + (p+1)(1-T)}{2p}\right]}{8\left[1 - \frac{\gamma}{p}\left(1 + \frac{1}{p} + \frac{4}{p^2} + \frac{\Delta}{2}\right)\right]},$$

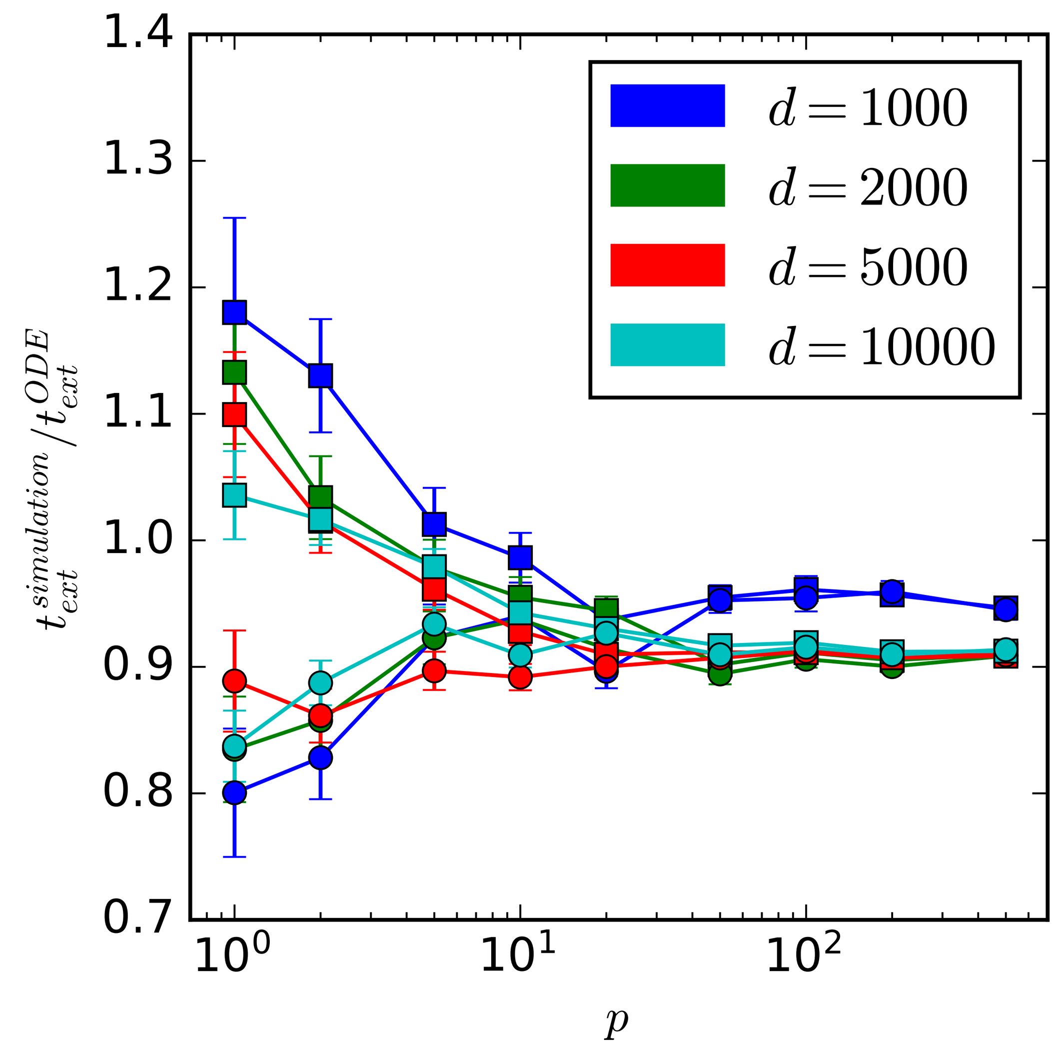
- **quenched formula**

$$t_{\text{ext}}^{(\text{qnc})} = \mathbb{E}_{\mu_0,\tau_0\sim\mathcal{P}_p^d}\left[\frac{\log\left[\frac{Tp(p+1)d + (2\mu_0 p - \tau_0)(1-T)}{2\mu_0 p}\right]}{8\left[1 - \frac{\gamma}{p}\left(1 + \frac{1}{p} + \frac{4}{p^2} + \frac{\Delta}{2}\right)\right]}\right]$$

where $\mu_0, \tau_0 \sim \mathcal{P}_p^d$ and

$$\mathcal{P}_p^d \equiv \left(d\sum_{j=1}^{p}(u_j\cdot v)^2, 2d\sum_{j=1}^{p}\sum_{l=j+1}^{p}(u_j\cdot u_l)^2\right)$$

with $v, u_j \sim \mathbb{S}^{d-1}(1)$.



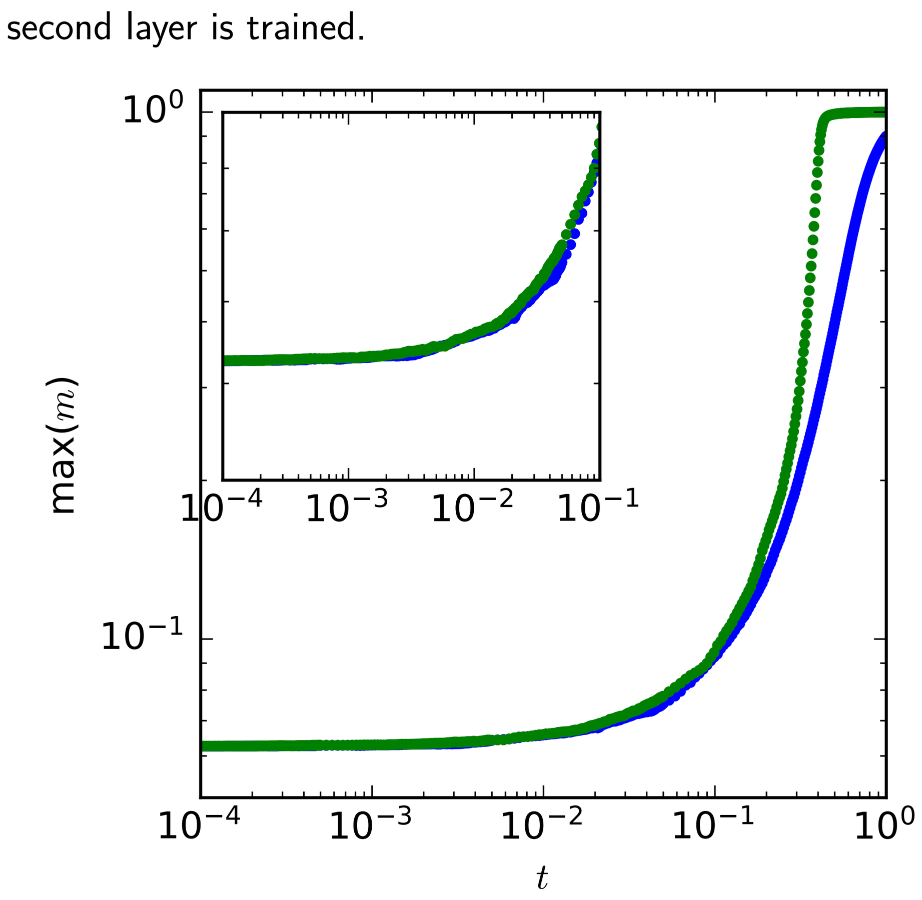
- The formulae match when $p \to \infty$.
- **The sample complexity is again $n = O(d \log d)$**
- There exits an **optimal learning rate** $\gamma_{\text{opt}}(p, \Delta)$.
- Traing with γ_{opt} at every p allow us to estimate the **gain factor of overparametrization**:

$$\frac{\text{SGD steps at } p = 1}{\text{SGD steps at } p \to +\infty} = \frac{12 + \Delta}{2 + \Delta}$$

No significant improvement over the $p = 1$ case.

Training the second layer

As of now, we fixed $a_j = 1$ for all j, but we can train them as well. We *numerically* showed that we can extend the results when the second layer is trained.



References

[1] ***Escaping mediocrity: how two-layer networks learn hard single-index models with SGD***, Luca Arnaboldi, Florent Krzakala, Bruno Loureiro, Ludovic Stephan arXiv preprint arXiv:2305.18502, 2023 [stat.ML]

[2] *On the sample complexity of learning generalized linear models with one-pass stochastic gradient descent*, Gérard Ben Arous, Reza Gheissari, Aukosh Jagannath. The Journal of Machine Learning Research, Volume 22, Issue 1, 2021.

[3] *Learning generalized linear models with two-layer neural networks*, Raphaël Berthier, Andrea Montanari, Kangjie Zhou. arXiv preprint arXiv:2303.00055, 2023