PSL Week 2024 - *Statistical Physics and Machine Learning*
Two lectures on stochastic gradient descent

Bruno Loureiro[1]

[1]Département d'Informatique, École Normale Supérieure (ENS) - PSL & CNRS, F-75230 Paris cedex 05, France

March 04-08, 2024

Get in touch at: brloureiro@gmail.com

1 Introduction and motivation

Consider a supervised learning regression setting where we are given n i.i.d. samples $(x^\nu, y^\nu)_{\nu\in[n]} \in \mathbb{R}^d \times \mathbb{R}$ from a probability density ρ defined on $\mathbb{R}^d \times \mathbb{R}$. The goal of supervised learning is to find a (typically) parametric function $f_\theta : \mathbb{R}^d \to \mathbb{R}$ such that, given a new sample $(x_{\text{new}}, y_{\text{new}}) \sim \rho$, the prediction $\hat{y} = f_\theta(x_{\text{new}})$ is as close as possible to the true label y_{new}. The typical questions we are interested in answering are:

- How can we find a good f_θ (or equivalently a good $\theta \in \mathbb{R}^m$)? Or in other words, what algorithm should we use?
- How much information from ρ do we need to find a good f_θ? Or in other words, how much data n do we need to see?
- How rich does the function class f_θ needs to be? Or in other words, what in a good choice of model or architecture and how large m should we take?

1.1 Empirical risk minimisation

A natural idea is to choose a **loss function** $\ell : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}_+$ that quantifies the error made by a given choice of f_θ. Therefore, in this metric the goal becomes to minimise the so-called **population risk**:

$$\mathcal{R}(\theta) = \mathbb{E}_{(x,y)\sim\rho}\left[\ell(y, f_\theta(x))\right]. \qquad \text{(Population risk)}$$

There are two problems with this. The first one is that the statistician typically has no access to $\mathcal{R}$, since she typically only has access to the samples $(x^\nu, y^\nu)_{\nu\in[n]}$ and not the full distribution ρ. Therefore, the best she can aim to do is to minimise instead **empirical risk**:

$$\hat{\mathcal{R}}_n(\theta) = \frac{1}{n}\sum_{\nu=1}^{n} \ell(y^\nu, f_\theta(x^\nu)). \qquad \text{(Empirical risk)}$$

This is known as **empirical risk minimisation** (ERM), and it is one of the most popular ways of learning f_θ. However, there is no general guarantee that a global minimiser $\theta \in \operatorname{argmin} \hat{\mathcal{R}}_n$ of the empirical risk will have a low population risk $\mathcal{R}$. Indeed, for neural networks it is possible to find

global minima of the empirical risk which have very bad generalisation (high population risk) (Liu et al., 2020).

The second problem is that even if we had full access to $\mathcal{R}$, for many choices of function class $\{f_\theta : \mathbb{R}^d \to \mathbb{R} : \theta \in \mathbb{R}^m\}$ and loss function ℓ, the population $\mathcal{R}$ risk is a non-convex function of the parameter $\theta \in \mathbb{R}^m$. Therefore, finding the global minima is not easy, specially when the number of parameters m is large. In particular, the choice of optimisation algorithm and initialisation are very important when minimising non-convex objectives.

1.2 Descent algorithms

One of the most natural algorithms for optimisation is **gradient descent** (GD):

$$\theta^k = \theta^k - \gamma_k \nabla_\theta \hat{\mathcal{R}}_n(\theta^k) \tag{GD}$$

which consists of simply updating the weights in the steepest descent direction, with step sized gauged by $\gamma_k > 0$. Note that GD naturally stops at a point in which $\nabla_\theta \hat{\mathcal{R}}_n = 0$, which can be both a local or global minima. Defining a continuous function $\theta(\gamma_k k) = \theta_k$ by piecewise affine interpolation, when the step size is small $\gamma_k \to 0^+$, GD is well approximated by a continous **gradient flow**:

$$\dot{\theta}(t) = -\nabla_\theta \hat{\mathcal{R}}_n(\theta(t)). \tag{1.1}$$

where $\dot{\theta} := {}^{\mathrm{d}\theta}/_{\mathrm{d}t}$. Or, seeing things in the opposite way, GD can be seen as the Euler discretisation of gradient flow with $t = k\gamma_k$.

A drawback of GD is that at every step k, one needs to compute the full gradient over the empirical risk. This means running over the full training set at every time - which can be slow if n is large. A simple way to avoid this computational bottleneck is to estimate the gradient at each step k only on a subset $b_k \subset [n]$ (known as a *mini-batch*) of the training data, which gives **stochastic gradient descent** (SGD):

$$\theta^{k+1} = \theta^k - \gamma_k \frac{1}{|b_k|} \sum_{\nu \in b_k} \nabla_\theta \ell(y^\nu, f_\theta(x^\nu)). \tag{SGD}$$

Together with its variants, SGD is one of the most used algorithm in modern machine learning. Besides being more efficient than GD, one advantage of SGD is that it can be seen as an approximation for gradient flow on the population risk:

$$\dot{\theta} = -\nabla_\theta \mathcal{R}. \tag{1.2}$$

Indeed, note that although $\nabla_{\theta_k} \hat{\mathcal{R}}_n(\theta_k)$ is an unbiased estimate of $\nabla_\theta \mathcal{R}$ at initialisation $k = 0$, since in GD we use the same gradient at every step k, the gradient at time $k > 0$ will be a biased estimation of the true population gradient at this time (Robbins and Monro, 1951). Instead, if each mini-batch is chosen independently and without replacement (which is possible if a lot of data is available $n \gg |b_k|$), then at each $k > 0$ SGD will make a step on a direction which is an unbiased estimation of the population loss gradient. This limit is known as **one-pass SGD**[1], and if mostly often studied in the particular case of $|b_k| = 1$. Although the one-pass setting might seen unrealistic on a first sight, it is worth noting that it is a good approximation to certain scenarios, such as Large Language Models (LLM) like ChatGPT-3 which are trained of billions of tokens, see e.g. Table 2.2 in Brown et al. (2020). Note that in the one-pass case, each step corresponds to seeing one sample, and therefore the amount of data required to achieve a given error is equal to the number of SGD steps - or in other words, **convergence rates are equivalent to the sample complexity**.

[1] Sometimes also refereed to online SGD, specially in the Statistical Physics literature.

With the observation above in mind, an useful way of thinking about one-pass SGD is as a noisy version of gradient descent on the population risk:

$$\theta^{\nu+1} = \theta^{\nu} - \gamma_\nu \nabla_\theta \mathcal{R}(\theta^\nu) + \gamma_\nu \varepsilon^\nu \tag{1.3}$$

where we have switched notation $k \to \nu$ to stress that each step we take a fresh data, and we defined the effective noise:

$$\begin{aligned} \varepsilon^\nu &:= \nabla_\theta \ell\left(y^\nu, f_{\theta^\nu}(x^\nu)\right) - \nabla_\theta \mathcal{R}(\theta^\nu) \\ &= \nabla_\theta \ell\left(y^\nu, f_{\theta^\nu}(x^\nu)\right) - \mathbb{E}\left[\nabla_\theta \ell(y, f_{\theta^\nu}(x))\right] \end{aligned} \tag{1.4}$$

which has zero mean since the estimation is unbiased[2]. This observation can be surprising at first sight: on average one-pass SGD optimises the true population risk even though this is an unknown function! Therefore, understanding one-pass SGD essentially amounts to understanding two things: (1) gradient descent on the population risk, and (2) the properties of the effective noise ε^ν. It is important to stress, however, that ε^ν is a not a simple Gaussian noise, and therefore as a stochastic process **SGD can be very different from Brownian motion**.

Our goal in the following will be precisely to study one-pass SGD for a particular class of supervised learning problems.

2 Asymptotic limits of one-pass SGD

Four key ingredients define the one-pass SGD algorithm in eq. (1.3): the loss function ℓ, the parametric family f_θ (a.k.a. architecture), the data distribution ρ and the learning rate γ_k. Classical analysis of SGD consists of making strong assumptions on the objective function, for example $\theta \mapsto \ell(y, f_\theta(x))$ is a convex function[3], and deriving convergence rates which are fairly general on the data distribution ρ, see (Moulines and Bach, 2011) for an example.

Instead, here we will take a different approach. We will make strong assumptions on the data distribution ρ, with the benefit of being able to derive sharper results for a richer class of architectures f_θ. More precisely, we will consider the following setting.

Architecture: We are interested in the simplest architecture leading to a non-convex learning problem, the two-layer neural network:

$$f_\theta(x) = \frac{1}{p} \sum_{i=1}^{p} a_i \sigma(w_i \cdot x). \tag{2.1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function, and the overall normalisation is chosen for convenience[4]. Note that here we have $\theta = (a, W) \in \mathbb{R}^p \times \mathbb{R}^{p \times d}$, and therefore $m = p(d+1)$ parameters, where we have defined the weight matrix $W \in \mathbb{R}^{p \times d}$ by stacking $w_i \in \mathbb{R}^p$ row-wise. Some commonly employed activation functions are the ReLU $\sigma(x) = \max(0, x)$, the hyperbolic tangent $\sigma(x) = \tanh(x)$ and the error function $\sigma(x) = \mathrm{erf}(x)$.

Loss function: Since we are dealing with regression, we will also focus on the square loss, defined as:

$$\ell(y, f_\theta(x)) = \frac{1}{2}\left(y - f_\theta(x)\right)^2. \tag{2.2}$$

Abusing the notation, we will often omit the dependence on the data and denote $\ell(\theta) := \ell(y, f_\theta(x))$.

[2] Note that since a fresh sample is drawn at every step ν, θ^ν is independent of x^ν.

[3] Note this strongly constraints the architecture f_θ. For instance, for the square loss this implies $f_\theta(x) = \langle \theta, x \rangle$ is a linear function.

[4] Indeed, the normalisation is chosen to match the mean-field literature (Chizat and Bach, 2018; Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2022; Sirignano and Spiliopoulos, 2020).

Data: We assume i.i.d. Gaussian covariates $x^\nu \sim \mathcal{N}(0, 1/d I_d)$ and labels generated from a target function with additive noise:

$$y^\nu = f_\star(x^\nu) + \xi^\nu \tag{2.3}$$

where $\mathbb{E}[\xi^\nu] = 0$ and $\mathbb{E}[(\xi^\nu)^2] = \Delta < \infty$. Moreover, we assume the target function $f_\star : \mathbb{R}^d \to \mathbb{R}$ belongs to the class of **multi-index models**:

$$f_\star(x) = g\left(\langle w_1^\star, x\rangle, \cdots, \langle w_r^\star, x\rangle\right) \tag{2.4}$$

for a given set of (fixed) weights $w_1^\star, \dots, w_r^\star \in \mathbb{R}^d$ and link function $g : \mathbb{R}^k \to \mathbb{R}$. Note that in this setting the covariates x^ν are unstructured: they are drawn isotropically in $\mathbb{R}^d$, and all structure in the data is actually encoded in the target $f_\star$, which depends on the covariates only through the projections in the subspace spanned by the directions $(w_k^\star)_{k\in[r]}$. In the following, we will consider the asymptotic limit where $d \to \infty$ while $r = \Theta_d(1)$, with our multi-index function modelling a scenario where most information on our high-dimensional data distribution is captured by a low-dimensional subspace. For convenience, we define the matrix $W^\star \in \mathbb{R}^{r\times d}$ by stacking $w_k^\star$ row-wise. Learning the target function $f_\star$ therefore can be thought as jointly approximating the link function g and retrieving the subspace $W^\star$. Note that an important particular example of a multi-index model is given by a two-layer neural network with r hidden-units:

$$f_\star(x) = \sum_{k=1}^{r} a_k^\star \sigma_\star(\langle w_k^\star, x\rangle). \tag{2.5}$$

where $(W^\star, a^\star)$ are the first and second layer weights and $\sigma_\star : \mathbb{R} \to \mathbb{R}$ are the activation functions.

Remarks — A few remarks about the setting above are in place.

- In the context of statistical physics of learning, the setting above where the generative model for the data distribution is given by a parametric hypothesis class which we learn with another (not necessarily the same) model class is known as a **teacher-student model**[5]. Within this terminology, we refer to $f_\star$ as the **teacher** and f_θ as the student, with $\theta^\star = (W^\star, a^\star)$ the **teacher weights** and $\theta = (W, a)$ the **student weights**.
- Here we have assumed the covariates to be unstructured for simplicity. The discussion that will follow can be generalised to the correlated case $x^\nu \sim \mathcal{N}(0, \Sigma)$, see (Goldt et al., 2020).
- In the particular case where the target is a two-layer neural network eq. (2.5) with $p = k$ and $\sigma_\star = \sigma$, the global minimum of the population risk is given, up to permutation symmetry, by the teacher parameters θ^*. This is known as the **well-specified** setting, and in this case learning simplifies to estimating the parameters $\theta^\star$.
- Analogously, the case where the problem is not well-specified (e.g. $\sigma_\star \neq \sigma$) but one can still achieve perfect estimation (i.e. the lowest population risk $\mathcal{R} = \Delta$) is known as the **realisable setting**. Note it requires at least $p \geq r$.

2.1 Sufficient statistics

Consider the population risk for the problem we defined above:

$$\begin{aligned}\mathcal{R}(\theta) &= \mathbb{E}\left[\ell(y, f_\theta(x))\right] = 1/2\ \mathbb{E}\left[(y - f_\theta(x))^2\right] \\ &= 1/2\ \mathbb{E}_{x\sim\mathcal{N}(0,1/dI_d)}\left[\left(g\left(\langle w_1^\star, x\rangle, \dots, \langle w_r^\star, x\rangle\right) - \frac{1}{p}\sum_{i=1}^{p} a_i \sigma\left(\langle w_i, x\rangle\right)\right)^2\right] + \Delta/2\end{aligned} \tag{2.6}$$

[5]This terminology for the generative model for data was introduced by Gardner and Derrida (1989).

where we used the fact that the noise ξ is independent from x. Note that in order to compute the expectation over x, all we need to know is the joint distribution between the pre-activations:

$$\lambda_k^\star = \langle w_k^\star, x \rangle, \qquad \lambda_i = \langle w_i, x \rangle. \tag{2.7}$$

In the context of statistical physics, these are also known as **local fields**. Conditionally on $W, W^\star$, these are jointly Gaussian variables:

$$(\lambda^\star, \lambda) \sim \mathcal{N}\left(0_{r+p}, \Sigma\right), \qquad \Sigma := \begin{bmatrix} P & M \\ M^\top & Q \end{bmatrix} \in \mathbb{R}^{(r+p)\times(r+p)} \tag{2.8}$$

where:

$$P := \frac{1}{d} W^\star W^{\star\top} \in \mathbb{R}^{r\times r}, \qquad M^\nu := \frac{1}{d} W^\star W^\top \in \mathbb{R}^{r\times p}, \qquad Q := \frac{1}{d} W W^\top \in \mathbb{R}^{p\times p}. \tag{2.9}$$

Note that by construction Σ, P and Q are symmetric matrices. Therefore, if we are only interested in tracking the evolution of the population error throughout the dynamics, it is sufficient to track the matrix Σ^ν, or equivalently (M^ν, Q^ν) since P is fixed. In the statistical physics parlour, these covariance matrices are our **order parameters**, also known as **sufficient statistics** in statistics. These are $p(p+2r-1)/2$ parameters, in comparison with the original pd parameters for the SGD weights $W^\nu \in \mathbb{R}^{p\times d}$. Therefore, if we are only interested in understanding the risk in the regime where $d \gg p, r$, it might be advantageous to track Σ^ν instead of W^ν.

2.2 Evolution of the sufficient statistics

Motivated by the observation in section 2.1, can we derive equations for the evolution of the sufficient statistics Σ^ν from the original SGD for the weights eq. (1.3)?

$$\mathcal{E}^\nu = \mathcal{E}(a^\nu, \lambda^{\star\nu}, \lambda^\nu) := y^\nu - \frac{1}{p}\sum_{j=1}^{p} a_j^\nu \sigma\left(w_j^\nu, x^\nu\right) \tag{2.10}$$

such that $\ell(a^\nu, W^\nu) = 1/2\mathcal{E}(\lambda^\star, a^\nu, \lambda^\nu)^2$. With that notation, the gradient of the loss with respect to the weights are given by:

$$\begin{aligned} \nabla_{a_i} \ell(a^\nu, W^\nu) &= -\frac{1}{p}\mathcal{E}^\nu \sigma\left(\langle w_i^\nu, x^\nu \rangle\right) \\ \nabla_{w_i} \ell(a^\nu, W^\nu) &= -\frac{1}{p}\mathcal{E}^\nu a_i^\nu \sigma'\left(\langle w_i^\nu, x^\nu \rangle\right) x^\nu. \end{aligned} \tag{2.11}$$

Therefore, with these notations we can write the evolution of the weights under the one-pass SGD dynamics as:

$$\begin{aligned} a_i^{\nu+1} &= a_i^\nu + \frac{\gamma_a}{p}\mathcal{E}(\lambda^{\star\nu}, a^\nu, \lambda^\nu)\sigma\left(\lambda_i^\nu\right) \\ w_i^{\nu+1} &= w_i^\nu + \frac{\gamma}{p}\mathcal{E}(\lambda^{\star\nu}, a^\nu, \lambda^\nu) a_i^\nu \sigma'\left(\lambda_i^\nu\right) x^\nu \end{aligned} \tag{2.12}$$

As we will see later, it is important to keep separate learning rates for each layer (γ, γ_a) due to the difference in the scaling of the gradient. Note that the right-hand side of the above only depend on the second-layer weights a^ν and the local-fields $\lambda^{\star\nu}_k = \langle w_k^\star, x^\nu \rangle$ and $\lambda_i^\nu = \langle w_i^\nu, x^\nu \rangle$.

Equation for M^ν: Taking the dot product with respect to $w_k^\star$ at both sides and dividing by $1/d$, we can write closed form equations for the evolution of M^ν:

$$M_{ki}^{\nu+1} = M_{ki}^\nu + \frac{\gamma}{pd}\mathcal{E}(\lambda^{\star\nu}, a^\nu, \lambda^\nu) a_i^\nu \sigma'\left(\lambda_i^\nu\right) \lambda^{\star\nu}_k \tag{2.13}$$

Note that this is a stochastic differential equation (SDE) for the $r+p$ random variables $M_{\rho i}$.

Equation for Q^ν: The equation for Q is just a bit more involved. First, we dot product both sides of eq. (2.12) with respect to $w_j^{\nu+1}$:

$$w_i^{\nu+1} \cdot w_j^{\nu+1} = \left(w^\nu + \frac{\gamma_\nu}{p} \mathcal{E}^\nu a_i^\nu \sigma'\left(\lambda_i^\nu\right) x^\nu \right) \cdot w_j^{\nu+1} \tag{2.14}$$

and now we re-apply eq. (2.12), but now for $w_j^{\nu+1}$:

$$w_i^{\nu+1} \cdot w_j^{\nu+1} = \left(w^\nu + \frac{\gamma_\nu}{p} \mathcal{E}^\nu a_i^\nu \sigma'\left(\lambda_i^\nu\right) x^\nu \right) \left(w_j^\nu + \frac{\gamma_\nu}{p} \mathcal{E}^\nu a_j^\nu \sigma'\left(\lambda_j^\nu\right) x^\nu \right) \tag{2.15}$$

Dividing by $1/d$, this closes in Q:

$$Q_{ij}^{\nu+1} = Q_{ij}^\nu + \frac{\gamma_\nu}{pd} \left(\mathcal{E}^\nu a_i^\nu \sigma'\left(\lambda_i^\nu\right) \lambda_j^\nu + \mathcal{E}^\nu a_j^\nu \sigma'\left(\lambda_j^\nu\right) \lambda_i^\nu \right) + \frac{\gamma_\nu^2}{dp^2} (\mathcal{E}^\nu)^2 \sigma'\left(\lambda_i^\nu\right) \sigma'\left(\lambda_j^\nu\right) \ ||x^\nu||_2^2 \tag{2.16}$$

Summary — Putting the three equations together, we have a system of stochastic processes for the joint evolution of the local fields $(\lambda_k^{\star\nu}, \lambda_i^\nu)$ and the second layer weights a_i^ν:

$$\begin{aligned}
a_i^{\nu+1} &= a_i^\nu + \frac{\gamma_a}{p} \mathcal{E}^\nu \sigma\left(\lambda_i^\nu\right) \\
M_{ki}^{\nu+1} &= M_{ki}^\nu + \frac{\gamma}{pd} \mathcal{E}^\nu a_i^\nu \sigma'\left(\lambda_i^\nu\right) \lambda^{\star\nu}_k \\
Q_{ij}^{\nu+1} &= Q_{ij}^\nu + \frac{\gamma}{pd} \mathcal{E}^\nu \left(a_i^\nu \sigma'\left(\lambda_i^\nu\right) \lambda_j^\nu + a_j^\nu \sigma'\left(\lambda_j^\nu\right) \lambda_i^\nu \right) + \frac{\gamma_\nu^2}{dp^2} (\mathcal{E}^\nu)^2 a_i^\nu a_j^\nu \sigma'\left(\lambda_i^\nu\right) \sigma'\left(\lambda_j^\nu\right) ||x^\nu||_2^2
\end{aligned} \tag{2.17}$$

where we recall the reader the definition of the displacement vector $\mathcal{E}^\nu = y^\nu - 1/p \sum_i a_i^\nu \sigma(\langle w_i^\nu, x^\nu \rangle)$. Note, however, that these equations are not closed on the local fields and their moments. Not only they depend explicitly on $||x^\nu||^2$, but they can depend on higher moments. For notational convenience, we now define the following potential functions:

$$\begin{cases} \Psi_a(a, \lambda_\star, \lambda) &= \mathcal{E}^\nu \sigma\left(\lambda_i^\nu\right) \\ \Psi_M(a, \lambda_\star, \lambda) &= \mathcal{E}^\nu a_i^\nu \sigma'\left(\lambda_i^\nu\right) \lambda^{\star\nu}_k \end{cases} \quad \begin{cases} \Psi_Q^{(\mathrm{gf})}(a^\nu, \lambda_\star^\nu, \lambda^\nu) &= \mathcal{E}^\nu \left(a_i^\nu \sigma'\left(\lambda_i^\nu\right) \lambda_j^\nu + a_j^\nu \sigma'\left(\lambda_j^\nu\right) \lambda_i^\nu \right), \\ \Psi_Q^{(\mathrm{var})}(a^\nu, \lambda_\star^\nu, \lambda^\nu) &= (\mathcal{E}^\nu)^2 a_i^\nu a_j^\nu \sigma'\left(\lambda_i^\nu\right) \sigma'\left(\lambda_j^\nu\right) ||x^\nu||_2^2 \end{cases} \tag{2.18}$$

which allow us to write the equations as:

$$\begin{aligned}
a_i^{\nu+1} - a_i^\nu &= \Psi_a(a, \lambda_\star, \lambda) \delta t_a \\
M_{ki}^{\nu+1} - M_{ki}^\nu &= \Psi_M(a, \lambda_\star, \lambda) \delta t \\
Q_{ij}^{\nu+1} - Q_{ij}^\nu &= \left[\Psi_Q^{(\mathrm{gf})}(a^\nu, \lambda_\star^\nu, \lambda^\nu) + \frac{\gamma}{p} \Psi_Q^{(\mathrm{var})}(a^\nu, \lambda_\star^\nu, \lambda^\nu) \right] \delta t
\end{aligned} \tag{2.19}$$

where we also defined the step-sizes $\delta t = \gamma/pd$ and $\delta t_a = \gamma_a/p$. Note that this makes clear the difference in scale between the step-size of the first layer δt and the second layer δt_a with respect to the covariate dimension d. In order to obtain a homogeneous scaling with respect to dimension, from now on we consider $\gamma_a := \gamma/d$ such that $\delta t_a = \delta t$.

2.3 Deterministic limits

So far, our computations are exact: eq. (2.19) are just an alternative rewriting of eq. (1.3). Although this can provide some insight in the evolution of the correlation functions, this rewriting is not very useful since the eq. (2.19) are not autonomous. Fortunately, it can be shown they considerable simplify

in the limit of vanishing step-size $\delta t \to 0^+$, where as we will see the stochastic evolution of the summary statistics concentrate towards a deterministic limit.

Before stating this result, it will be convenient to define a short-hand notation. First, we extend the local-fields covariance matrix from eq. (2.8) by stacking the second-layer weights a^ν block-wise:

$$\Omega^\nu \coloneqq \begin{bmatrix} a^\nu & 0 & 0 \\ 0 & P & M^\nu \\ 0 & M^{\nu\top} & Q^\nu \end{bmatrix} \in \mathbb{R}^{(r+p+1)\times(r+p+1)} \tag{2.20}$$

Second, we define the expectation of the summary statistics with respect to the random draw of the sequence of data $(x^\nu, y^\nu)_{\nu\in[n]}$:

$$\bar{\Omega}^\nu = \mathbb{E}[\Omega^\nu] \tag{2.21}$$

As well as the expectation of the random potential functions eq. (2.18):

$$\bar{\Psi}_a(\bar{\Omega}) = \mathbb{E}[\Psi_a(a, \lambda_\star, \lambda)], \qquad \bar{\Psi}_M(\bar{\Omega}) = \mathbb{E}[\Psi_M(a, \lambda_\star, \lambda)], \tag{2.22}$$

$$\bar{\Psi}_Q^{(\mathrm{gf})}(\bar{\Omega}) = \mathbb{E}[\Psi_Q^{(\mathrm{gf})}(a, \lambda_\star, \lambda)], \qquad \bar{\Psi}_Q^{(\mathrm{var})}(\bar{\Omega}) = \mathbb{E}[\Psi_Q^{(\mathrm{var})}(a, \lambda_\star, \lambda)] \tag{2.23}$$

Note that these are deterministic function that depend only on the summary statistics $\bar{\Omega}$. Third, we define a continuous time variable $t \in \mathbb{R}_+$ and an extension of the summary statistics to continuous time $\bar{\Omega}(t)$ as follows: at times $t = \nu\delta t$ we have:

$$\bar{\Omega}(\nu\delta t) \coloneqq \bar{\Omega}^\nu, \qquad \nu \in [n] \tag{2.24}$$

At all other times $t \geq 0$, we define $\Omega(t)$ via linear interpolation. We are now ready to state our result.

Theorem 1 (Veiga et al. (2022), informal)**.** *Consider the stochastic process defined by eq. (2.19) with step-size $\delta t \coloneqq \gamma/dp$ from the initial condition $\bar{\Omega}(0) \coloneqq \Omega^0$ and $r = \Theta(1)$. Then, there exists a constant $C > 0$ such that for any $\nu \in [n]$:*

$$||\Omega^\nu - \bar{\Omega}(\nu\delta t)||_\infty \leq e^{C\nu\delta t}\sqrt{\frac{\gamma}{dp}} \tag{2.25}$$

Where for $\gamma/p = \Theta(1)$, $\bar{\Omega}(t)$ satisfies the following ordinary differential equation:

$$\begin{aligned} \frac{d\bar{a}}{dt} &= \bar{\Psi}_a(\bar{a}(t), \bar{M}(t), \bar{Q}(t)), \\ \frac{d\bar{M}_{ki}}{dt} &= \bar{\Psi}_M(\bar{a}(t), \bar{M}(t), \bar{Q}(t)), \\ \frac{d\bar{Q}_{ij}}{dt} &= \bar{\Psi}_Q^{(\mathrm{gf})}(\bar{a}(t), \bar{M}(t), \bar{Q}(t)) + \frac{\gamma}{p}\bar{\Psi}_Q^{(\mathrm{var})}(\bar{a}(t), \bar{M}(t), \bar{Q}(t)) \end{aligned} \tag{2.26}$$

Otherwise, if $\gamma/p = o(1)$, $\bar{\Omega}(t)$ satisfies instead the following set of simplified equations:

$$\frac{d\bar{a}}{dt} = \bar{\Psi}_a(\bar{a}(t), \bar{M}(t), \bar{Q}(t)), \tag{2.27}$$

$$\frac{d\bar{M}_{ki}}{dt} = \bar{\Psi}_M(\bar{a}(t), \bar{M}(t), \bar{Q}(t)), \tag{2.28}$$

$$\frac{d\bar{Q}_{ij}}{dt} = \bar{\Psi}_Q^{(\mathrm{gf})}(\bar{a}(t), \bar{M}(t), \bar{Q}(t)) \tag{2.29}$$

Intuitively, theorem 1 states that the stochastic process for the summary statistics stays close to its mean during a finite time horizon $n = \Theta(1)$, which is given by the solution of a deterministic ordinary differential equation. A few comments are in place.

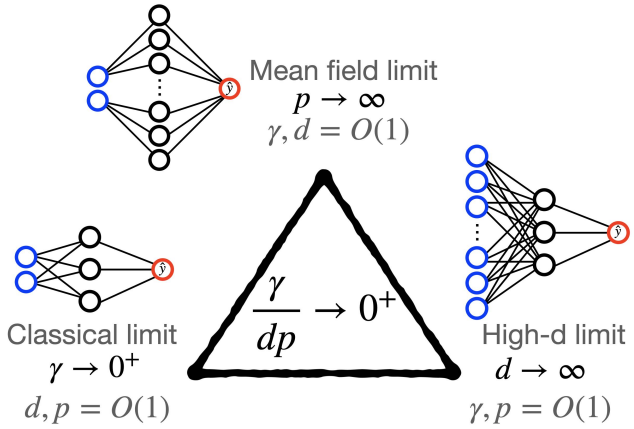

Figure 1: The different scaling limits of one-pass SGD.

Remarks —

- The result above was first derived in a series of seminal works by Saad and Solla (1996, 1995a,b) in the particular limit $d \to \infty$ with $\gamma, p = \Theta_d(1)$. Although partial mathematical results were known, e.g. (Reents and Urbanczik, 1998), a full rigorous proof only appeared in Goldt et al. (2019), which built on a proof scheme from Wang et al. (2018, 2019). This was later generalised to the $\gamma/p = o(1)$ regime by Arnaboldi et al. (2023); Veiga et al. (2022) with the motivation of studying wide neural networks $p \gg 1$.

- As we have previously stressed, one-pass SGD sees a single data sample (x^ν, y^ν) at every update step ν. Therefore, the quantity of data seen by the algorithm is exactly equal the total number of SGD steps. In continuous time, we have $t = \nu\delta t$, meaning that one continuous step corresponds to $\delta t^{-1} = dp/\gamma$ samples.

- Mathematically, the fact that the process can be asymptotically characterised by a simple ODE is crucially due to the Markovian nature of one-pass SGD: the only source of randomness in the parameters $(W^{\nu+1}, a^{\nu+1})$ at step $\nu + 1$ are, conditionally on the parameters on step ν, given by the independent draw of the sample (x^ν, y^ν). The proof of this result is out of the scope of these lectures, but it precisely builds on this observation by decomposing the process into a deterministic part and a martingale correction that can be controlled. I refer the motivated reader to Appendix A of Veiga et al. (2022).

- Note that the bound in eq. (2.25) has an exponential dependence on time $t = \nu\delta t$. This means that theorem 1 guarantees the stochastic process is close to its expectation for any fixed time horizon $T = n\delta t = \Theta(1)$, effectively implying we require n can be as large as $\delta t^{-1} = dp/\gamma$. However, this is only a bound, and it does not imply that the variance of the process necessarily blows up with time. As we will see later, for particular problems the variance stays bounded for longer time scales (see (Arnaboldi et al., 2024) for an example). We don't expect this to be always true, and understanding under which conditions we can prove results which are uniform in time is an important open problem.

- Theorem 1 was derived for standard one-pass SGD. However, similar asymptotic limits can also be derived for some of its variants, for instance spherical SGD.

2.4 The different regimes

Theorem 1 hold under the condition that the step-size is small $\delta t = \gamma/dp$. In this section we have a look at the different ways this limit can be achieved, as illustrated in fig. 1, and what this implies

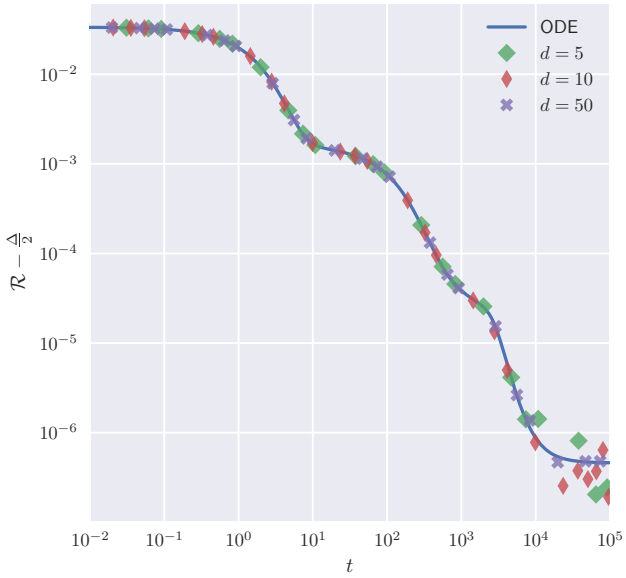

Figure 2: Evolution of the excess error $\mathcal{R} - \Delta/2$ as a function of time for a two-layer network with width $p = 10$ trained under one-pass SGD on data generated from a two-layer target function with $r = 5$, label noise $\Delta = 10^{-3}$ and activation functions $\sigma(x) = \sigma_\star(x) = \mathrm{erf}(x)$. Here, we kept the first layer fixed $a_i = a_r^\star = 1$. Dots correspond to finite size simulations at different covariate dimensions $d \in \{5, 10, 50\}$, and the solid curve corresponds to the theoretical prediction. Interestingly, starting from a dimension independent initial condition the trajectories are independent of the covariate dimension.

to the dynamics. For notational convenience, in the following the drop the distinction between the random Ω and averaged process $\bar{\Omega}$, denoting all quantities without bars.

Classical regime — One of the most studied regime in the classical machine learning literature is the gradient flow limit, corresponding to a vanishing learning rate $\gamma \to 0^+$ at fixed dimensions $p, d = \Theta(1)$. As previously discussed in section 1.2, in this limit it was shown by Robbins and Monro (1951) that one-pass SGD converges to the gradient flow on the population risk eq. (1.2). In terms of the summary statistics, the equations simplify to:

$$\begin{aligned}
\dot{a}_i(t) &= \bar{\Psi}_a(a, M, Q) = \mathbb{E}_{\Omega(\lambda_\star,\lambda)\sim\mathcal{N}(0,\Sigma(t))}[\mathcal{E}(a, \lambda_\star, \lambda)\sigma(\lambda_i)] \\
\dot{M}_{ki}(t) &= \bar{\Psi}_M(a, M, Q) = \mathbb{E}_{\Omega(\lambda_\star,\lambda)\sim\mathcal{N}(0,\Sigma(t))}[\mathcal{E}(a, \lambda_\star, \lambda)a_i(t)\sigma'(\lambda_i)\lambda_k^\star] \\
\dot{Q}_{ij}(t) &= \bar{\Psi}_Q^{(\mathrm{gf})}(a, M, Q) = \mathbb{E}_{\Omega(\lambda_\star,\lambda)\sim\mathcal{N}(0,\Sigma(t))}[\mathcal{E}(a, \lambda_\star, \lambda)\left(a_i(t)\sigma'(\lambda_i)\lambda_j + a_j(t)\sigma'(\lambda_j)\lambda_i\right)]
\end{aligned} \tag{2.30}$$

These equations correspond precisely to what one would have obtained starting directly from the gradient flow eq. (1.2). Very importantly, the equation for the covariance of the first layer weights $Q = {}^1/_d WW^\top$. This justifies our choice of notation for the potential functions $\bar{\Psi}^{(\mathrm{gf})}, \bar{\Psi}^{(\mathrm{var})}$: indeed, it is easy to show that the potential $\bar{\Psi}^{(\mathrm{gf})}$ comes from the gradient of the population risk, while $\bar{\Psi}^{(\mathrm{var})}$ is given by the variance of the effective SGD noise ε^ν (1.4). Since, as shown by Robbins and Monro (1951) the effective SGD noise is subleading in the learning rate γ, it is natural that $\bar{\Psi}^{(\mathrm{var})}$ does not contribute to the gradient flow limit. An illustration of the evolution of the population risk under the one-pass SGD dynamics in the gradient flow regime is given in fig. 2.

One important remark is that if $p, d = \Theta(1)$, the gradient flow eq. (1.2) only depend on dp parameters, while the above depend of $p(p-1)/2 + p(r+1)$ variables. Therefore, the summary statistics description might not really be convenient in this regime.

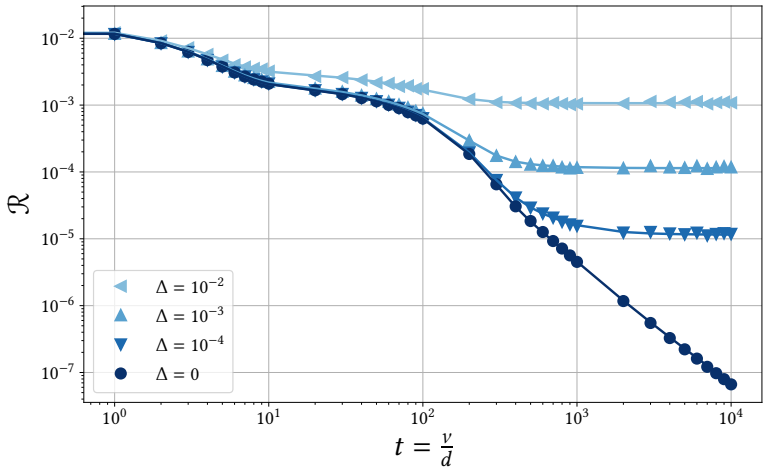

Figure 3: Evolution of the error $\mathcal{R}$ as a function of time for a two-layer network with width $p = 8$ trained under one-pass SGD on data generated from a two-layer of width $r = 4$ and activation functions $\sigma(x) = \sigma_\star(x) = \mathrm{erf}(x)$ for different label noise values $\Delta \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$. Here, we kept the first layer fixed $a_i = a_r^\star = 1$. Dots correspond to finite size simulations at dimension $d = 10^3$, and the solid curve corresponds to the theoretical prediction.

High-dimensional regime — Consider the limit of high-dimensional covariates $d \to \infty$ at fixed learning rate and width $\gamma, p = \Theta(1)$, which we will refer to as the *high-dimensional regime*. This was the limit that motivated the pioneering work of Saad and Solla (1996, 1995a,b), since in this limit the SGD updates on the parameters are intractable (one needs to track pd parameters), and hence the summary statistics description (which is independent of d) provide a real advantage.

The key difference in the high-dimensional regime is that the noise potential $\bar{\Psi}^{(\mathrm{var})}$ is of the same order as the gradient flow potential $\bar{\Psi}^{(\mathrm{gf})}$. While the equations for a_i, M_{ki} stays the same, the equation for the weights covariance is now given by:

$$\begin{aligned}\dot{Q}_{ij}(t) &= \bar{\Psi}_Q^{(\mathrm{gf})}(a, M, Q) + \frac{\gamma}{p}\bar{\Psi}_Q^{(\mathrm{var})}(a, M, Q) \\ &= \mathbb{E}[\mathcal{E}(a, \lambda_\star, \lambda)\left(a_i(t)\sigma'(\lambda_i)\lambda_j + a_j(t)\sigma'(\lambda_j)\lambda_i\right)] + \frac{\gamma}{p}\mathbb{E}[\mathcal{E}(a, \lambda_\star, \lambda)^2 a_i(t)a_j(t)\sigma'(\lambda_i)\sigma'(\lambda_j)]\end{aligned} \tag{2.31}$$

An illustration of the evolution of the risk in this regime is given in fig. 3. As we are going to see in some simple examples later, the presence contribution of the noise variance can have a non-trivial effect of the SGD dynamics.

Mean-field regime — We now consider the limit of infinite hidden-layer width $p \to \infty$ at fixed $d, \gamma = \Theta(1)$. But before discussing the evolution of the summary statistics, let's do a quick review of the main results on the mean-field limit for neural networks.

The key idea, which draws back from early works in approximation theory of neural networks (Barron, 1993; Kurkova and Sanguineti, 2001) is to see two-layer neural networks as discretisation over functions defined over measures. More precisely, define:

$$f_\mu(x) = \int \mathrm{d}\mu(a, w) a\sigma(\langle w, x\rangle) \tag{2.32}$$

for a measure μ over $\mathbb{R}^{d+1}$. Then, it is clear that our two-layer neural network is obtained as a

particular case $f_\theta(x) = f_{\hat{\mu}_p}(x)$ where we have defined the empirical density of weights:

$$\hat{\mu}_p(a, w) = \frac{1}{p}\sum_{i=1}^{p} \delta(a - a_i)\delta(w - w_i) \tag{2.33}$$

Following the same construction, we can define a population risk function in the space of measures:

$$\begin{aligned}\mathcal{R}(\mu) &= \frac{1}{2}\mathbb{E}\left[(y - f_\mu(x))^2\right] \\ &= \frac{1}{2}\mathbb{E}\left[\left(y - \int \mathrm{d}\mu(a, w) a\sigma(\langle w, x\rangle)\right)^2\right]\end{aligned} \tag{2.34}$$

A key observation in this construction is that, differently from $\mathcal{R}(a, w)$ which is a non-convex function over w, $\mathcal{R}(\mu)$ is a convex function over μ (Bach, 2017; Bengio et al., 2005; Rosset et al., 2007). Again, we can follow the same logic and define a gradient flow over the space of measures:

$$\partial_t \mu_t = \nabla_{(a,w)} \cdot (\mu_t \psi(\cdot, \mu_t)) \tag{2.35}$$

where $\psi(a, w, \mu) = (\psi_a, \psi_w)$ is the continuous equivalent of the gradient flow equations:

$$\psi_a(a, w; \mu) = \mathbb{E}[\sigma(\langle w, x\rangle)(y - f_\mu(x))] \tag{2.36}$$

$$\psi_w(a, w; \mu) = \mathbb{E}[a\sigma'(\langle w, x\rangle)x(y - f_\mu(x))] \tag{2.37}$$

A key result in the mean-field line of work (Chizat and Bach, 2018; Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2022; Sirignano and Spiliopoulos, 2020) is to show that as $p \to \infty$, under suitable conditions, one-pass SGD over the parameters (W, a) stay close to the Wasserstein gradient flow eq. (2.35). Moreover, due to convexity of $\mathcal{R}(\mu)$, under some conditions over the activation function σ and the initial conditions (W^0, a^0), it can be shown that if the flow in eq. (2.35), it must converge to a global minimum Chizat and Bach (2018). See (Bach and Chizat, 2021) for a great detailed review on these results.

The mean-field results discussed above assume little on the data distribution. Can we derive similar consequences for our setting from section 2? Note that from a first sight, from the point of view of the structure of the ordinary differential equations the $p \to \infty$ limit is exactly equal to the gradient flow limit $\gamma \to 0^+$. An important difference, however, is that in this limit the dimension of the sufficient statistics grow - recall that $a \in \mathbb{R}^p$, $M \in \mathbb{R}^{r\times p}$ and $Q \in \mathbb{R}^{p\times p}$. A natural idea would be to define, similarly to eq. (2.33) a density of the summary statistics. However, while the dimension of a, M is linear in p, Q has $p(p-1)/2$ entries. However, for $p > d$, Q is at most a rank d matrix, and the most relevant part of the dynamics should happen in the subspace spanned by the target directions $V = \mathrm{span}(w_1^\star, \cdots, w_r^\star)$. Defining the orthogonal decomposition of the weights:

$$W = P_V W + W^\perp = MP^{-1}W^\star + W^\perp \tag{2.38}$$

a similar decomposition on the local-fields and summary statistics:

$$\lambda^\perp = W^\perp x = \lambda - MP^{-1}\lambda_\star, \qquad Q^\perp = Q - MP^{-1}M^\top. \tag{2.39}$$

From the ODE for Q, it is easy to derive the asymptotic evolution of the perpendicular component $Q^\perp$:

$$\dot{Q}^\perp_{ij} = \Psi^\perp_Q(a, M, Q) = \mathbb{E}\left[\mathcal{E}(a, \lambda_\star, \lambda)\left(a_i(t)\sigma'(\lambda_i)\lambda^\perp_j + a_j(t)\sigma'(\lambda_j)\lambda^\perp_i\right)\right] \tag{2.40}$$

Note that this decomposition is exact. Surprisingly, from these equations, it can be shown that in the large-width limit $p \to \infty$ the most important component in the evolution of Q is its diagonal $q_i = Q_{ii}$.

Indeed, under a mild assumption over $W^\star$ (see Arnaboldi et al. (2023) for details), the component $W^\perp$ perpendicular to $V = \text{span}(W^\star)$ can be approximated by an independent random vector:

$$w_i^\perp \approx \sqrt{q_i} g_i, \qquad g_i \sim \text{Unif}(\mathbb{S}^{d-r-1}). \tag{2.41}$$

This means that the important correlations between the hidden-units lie in the subspace spanned by the target directions, with the correlation between the different perpendicular components of the hidden-units $Q_{ij}^\perp = {}^1\!/_d \langle w_i^\perp, w_j^\perp \rangle$ behaving as random forces in the dynamics. To make this precise, define a random version of the hidden-layer covariances Q:

$$\tilde{Q} = MP^{-1}M^\top - \text{diag}(\sqrt{q_i})\Xi\text{diag}(\sqrt{q_i}) \tag{2.42}$$

with $\Xi \in \mathbb{R}^{p\times p}$ a random matrix satisfying:

$$\Xi_{ii} = 1, \qquad \Xi_{ij} = \langle g, g' \rangle, \qquad g, g' \sim \text{Unif}(\mathcal{S}^{d-r-1}). \tag{2.43}$$

Consider $\mathcal{R}(t) \coloneqq \mathcal{R}(a(t), M(t), Q(t))$ the population risk written in terms of the continuous limit of the summary statistics. Remember this is a deterministic function. We can analogously define a random version of this function by evaluating it on $\tilde{Q}$ instead of Q. Defining its expectation:

$$\mathcal{R}(a(t), M(t), q(t)) = \mathbb{E}_\Xi[\mathcal{R}(a(t), M(t), \tilde{Q}(t))] \tag{2.44}$$

One can show the following result:

Theorem 2 (Arnaboldi et al. (2023))**.** *With probability at least* $1 - e^{-z^2}$ *on the initialisation:*

$$\sup_{t\in[0,T]} |\mathcal{R}(a(t), M(t), Q(t)) - \mathcal{R}(a(t), M(t), q(t))| \leq Ce^{CT} \frac{\sqrt{\log pT} + z}{\sqrt{p}} \tag{2.45}$$

where the reduced summary statistics (a, M, q) *solve the following averaged ODEs:*

$$\dot{a}_i = \mathbb{E}_\Xi[\bar{\Psi}_a(a, M, \tilde{Q})], \qquad \dot{M}_{ki} = \mathbb{E}_\Xi[\bar{\Psi}_M(a, M, \tilde{Q})], \qquad \dot{q}_i = \mathbb{E}_\Xi[\bar{\Psi}_Q^{(\text{gf})}(a, M, \tilde{Q})] \tag{2.46}$$

One can go even further and show that the contribution of the random matrix Ξ to the risk $\mathcal{R}(a, M, \tilde{Q})$ is $\Theta(d^{-1/2})$, and therefore in the joint limit where $p, d \to \infty$ we can completely ignore this term in the dynamics, see theorem 3.5 in (Arnaboldi et al., 2023) for a detailed discussion.

Theorem 2 considerably simplifies the sufficient statistics description, reducing the number of parameters in Q from ${}^{p(p-1)}\!/_2$ to p. This allow us to follow an analogous construction to the mean-field limit discussed above. Defining an empirical density over the summary statistics:

$$\hat{\pi}_p(a, m, q) = \frac{1}{p}\sum_{i=1}^{p} \delta(a - a_i(t))\delta(m - m_i(k))\delta(q - q_i(t)) \tag{2.47}$$

where $m_i \in \mathbb{R}^r$ are the columns of $M \in \mathbb{R}^{r\times p}$. Note this is a density over $\mathbb{R}^{r+2}$. By similar arguments to the mean-field limit, when $p \to \infty$ this can be shown to converge weakly to a density $\tilde{\mu}_t$ satisfying a Wasserstein gradient flow:

$$\partial_t \pi_t = \nabla_{(a,m,q)} \cdot (\pi_t \psi(\cdot; \pi_t)) \tag{2.48}$$

where $\psi(a, m, q; \mu_t) = (\psi_a, \psi_m, \psi_q)$ are the continuous version of the potentials in the right-hand side of eq. (2.46).

3 Simple case studies

Up to now, we have only discussed the structural form of the limiting summary statistics evolution. In this section, we look at some concrete cases where the dynamics can be solved, starting from the simplest example: linear regression.

3.1 Least-squares regression

Least-square regression is a particular case of the setting discussed in section 2 with $r = p = 1$ and $g(\langle w^\star, x\rangle) = \langle w^\star, x\rangle$. In other words, data $(x^\nu, y^\nu)_{\nu\in[n]}$ is generated from:

$$y^\nu = \langle w^\star, x^\nu\rangle + z^\nu, \qquad x^\nu \sim \mathcal{N}(0, {}^1\!/_d I_d), \qquad z^\nu \sim \mathcal{N}(0, \Delta). \tag{3.1}$$

which we seek to learn by doing one-pass SGD on the linear model $f_\theta(x) = \langle w, x\rangle$:

$$w^{\nu+1} = w^\nu - \gamma \nabla_w \ell(y^\nu, \langle w^\nu, x^\nu\rangle). \tag{3.2}$$

Note that this is a convex problem over the parameters w. Since this problem is simple, a lot can be said by directly looking at the weights. Therefore, it is a nice case study where we can compare the dynamics in parameter and summary statistics space. Indeed, the gradient of the loss in eq. (3.2) is simply given by:

$$\begin{aligned} g \coloneqq \nabla_w \ell(y, \langle w, x\rangle) &= -(y - \langle w, x\rangle)x \\ &= -\left(w^\star - w\right)^\top x x^\top - zx \end{aligned} \tag{3.3}$$

hence, its expectation is given by:

$$\mathbb{E}[g] = -{}^1\!/_d \left(w^\star - w\right) \tag{3.4}$$

Note that even if we initialise $w^0 = 0_d$, the initial gradient will negatively point towards the signal $w^\star$. As a starting point, let's look at the gradient flow equations. Following the convention from section 2, we define a step-size $\delta t = \gamma/d$, and take the limit $\gamma \to 0^+$ to get:

$$\dot{w}(t) = -d\mathbb{E}[g(t)] = w^\star - w(t) \tag{3.5}$$

This is a simple linear ODE, which admits a closed-form solution:

$$w(t) = w^\star + e^{-t}(w^0 - w^\star) \tag{3.6}$$

where $w^0 \coloneqq w(0)$ is the initial condition. This is a simple exponential relaxation from the initial condition w^0 to the global minimum $w^\star$, with a typical time scale $\tau = 1$. Recalling that $t = \nu\delta t = {}^{\nu\gamma}\!/_d$, we can see that the smaller γ/d, the more steps are needed to keep $\tau = \Theta(1)$. Therefore, GF has a sample complexity $n = \Theta(d)$. It will also be useful to write the explicit solution for the mean-squared error:

$$\text{mse}(t) = {}^1\!/_d ||w(t) - w^\star||^2 = e^{-2t} {}^1\!/_d ||w^0 - w^\star||_2^2 = e^{-2t}\text{mse}_0 \tag{3.7}$$

As we discussed in eq. (1.3), the original one-pass SGD can be seen as a noisy version of GF:

$$\begin{aligned} w^{\nu+1} - w^\nu &= \gamma(w^\star - w^\nu)^\top x^\nu x^{\nu\top} + \gamma z^\nu x^\nu \\ &= (w^\star - w^\nu) + \gamma\varepsilon^\nu \end{aligned} \tag{3.8}$$

where the effective noise is explicitly given by:

$$\varepsilon^\nu = \gamma^{-1}({}^1\!/_d I_d - \gamma x^\nu x^{\nu\top})(w^\star - w^\nu) + z^\nu x^\nu \tag{3.9}$$

Note that this is very different from unstructured Gaussian noise. In particular, it is composed by a first factor which is proportional to a projector in the direction of the samples x^ν - which at every step is just a random direction in $\mathbb{R}^d$. Note this random term is self-annealing: the closer w^ν is to the global minimum $w^\star$, the smaller it gets. The second random term is independent from the iterates, and also lives in the spam of x^ν. However, despite the fact that z^ν and x^ν are Gaussian, the product

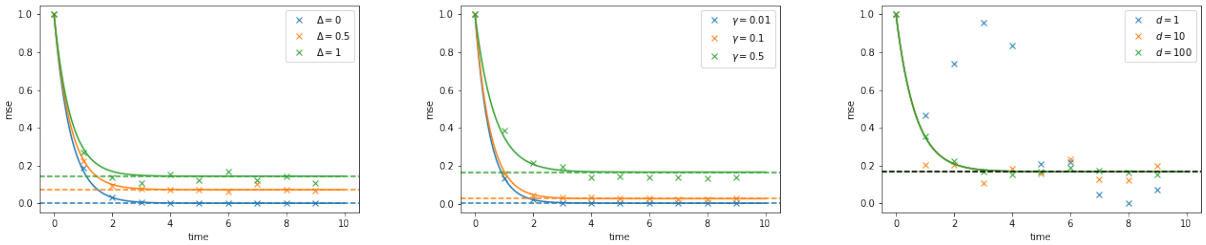

Figure 4: Mean-squared error as a function of time for a least squares problem from initial condition $w^0 = 0_d$ and target weights $w^\star \sim \text{Uni}(\mathbb{S}^{d-1})$. (**Left**) Fixed learning rate $\gamma = 0.25$ and dimension $d = 100$, for varying levels of label noise $\Delta \in \{0, 0.5, 1\}$. (**Centre**) Fixed dimension $d = 100$ and label noise level $\Delta = 0.5$, for varying levels of learning rates $\gamma \in \{10^{-2}, 10^{-1}, 0.5\}$. (**Right**) Fixed learning rate $\gamma = 0.5$ and label noise $\Delta = 0.5$ dimension $d = 100$, for varying problem dimensions $d \in \{1, 10, 100\}$. In all plots, solid lines correspond to the theoretical curves, and crosses to finite-size simulations.

$z^\nu x^\nu$ is not a Gaussian variable. With this observation in mind, one could proceed by studying the properties of the stochastic process eq. (3.8).

An alternative approach is to look at the evolution of the summary statistics in the limit where $\delta t = \gamma/d \to 0^+$. Taking $r = p = 1$ and $\sigma(x) = \sigma_\star(x) = 1$ in eq. (2.26), we get the following ODEs for $m = {}^1\!/_d\langle w, w^\star\rangle$ and $q = {}^1\!/_d||w||_2^2$:

$$\dot{m}(t) = \mathbb{E}[(\lambda_\star - \lambda + z)\lambda_\star] \tag{3.10}$$

$$\dot{q}(t) = \mathbb{E}[2(\lambda_\star - \lambda + z)\lambda + \gamma(\lambda_\star - \lambda + z)^2] \tag{3.11}$$

where the expectation is taken over:

$$(\lambda, \lambda^\star) \sim \mathcal{N}\left(0, \begin{bmatrix} \rho & m \\ m & q \end{bmatrix}\right), \qquad z \sim \mathcal{N}(0, \Delta) \tag{3.12}$$

where $\rho = {}^1\!/_d||w^\star||_2^2$. Taking the expectation explicitly is straightforward, and give us:

$$\dot{m}(t) = \rho - m(t) \tag{3.13}$$

$$\dot{q}(t) = 2(m(t) - q(t)) + \gamma(\Delta + \rho + q(t) - 2m(t)) \tag{3.14}$$

Interestingly, the equation for $m(t)$ is autonomous from $q(t)$ - this is a particular feature of the least-squares problem, and will not hold in more general cases. Again, these ODEs admit a closed form solution:

$$m(t) = (1 - e^{-t})\rho + m_0 e^{-t} \tag{3.15}$$

$$q(t) = \frac{2(\rho + q_0 - 2m_0) - \gamma(\rho + \Delta + q_0 - 2m_0)}{2 - \gamma} e^{-(2-\gamma)t} - 2e^{-t}(\rho - m_0) + \frac{\Delta\gamma}{2 - \gamma} \tag{3.16}$$

where $(m_0, q_0) = ({}^1\!/_d\langle w^0, w^\star\rangle, {}^1\!/_d||w^0||_2^2)$ are the initial values of the summary statistics. A few comments are in place.

Remarks:

- Note that for $\gamma > 2$, the norm of the predictor grows unboundedly as $t \to \infty$. This defines a critical learning rate $\gamma_\star = 2$ above which SGD over shoots the minima.

- For $\gamma \in (0, 2)$, the large time limit of the equations are given by:

$$\lim_{t\to\infty} m(t) = \rho \tag{3.17}$$

$$\lim_{t\to\infty} q(t) = \rho + \frac{\gamma\Delta}{2-\gamma} \tag{3.18}$$

  Which means we achieve perfect correlation with the global minimum. Nevertheless, for $\gamma \in (0, 2)$ we don't perfectly recover the weights $w^\star$ (which would imply $q = \rho$). Instead, we stay in a circle of radius $\gamma\Delta/(2-\gamma)$.

- We can also get a closed-form equation directly from mean-squared error:

$$\begin{aligned} \text{mse}(t) &= 1/d||w(t) - w^\star||_2^2 = \rho + q(t) - 2m(t) \\ &= e^{-(2-\gamma)t}\left(\text{mse}_0 - \frac{\gamma\Delta}{2-\gamma}\right) + \frac{\gamma\Delta}{2-\gamma} \end{aligned} \tag{3.19}$$

  which is related to the population loss as $\mathcal{R}(t) = 1/2(\text{mse} + \Delta)$. An illustration of the behaviour of the mean-squared error with varying parameters in the problem is given in fig. 4.

- In the gradient flow limit $\gamma \to 0^+$ we recover precisely the expected result from the exact solution eq. (3.6).

3.2 Phase retrieval

Least squares is in many ways a very special case: it is a convex optimisation problem with an exact closed-form solution for the summary statistics description of the dynamics. As we will see now, this is a luxury we don't have in more complicated setting.

A natural next step with respect least-squares is the real phase retrieval problem. This corresponds a problem with $p = r = 1$ and square activation function $\sigma(x) = x^2$.[6] In other words, data is generated from:

$$y^\nu = \langle w^\star, x^\nu\rangle^2 + z^\nu, \qquad x^\nu \sim \mathcal{N}(0, 1/dI_d), \qquad z^\nu \sim \mathcal{N}(0, \Delta). \tag{3.20}$$

which we seek to with a quadratic model itself $f_\theta(x) = \langle w, x\rangle^2$. Note that although the quadratic activation might seem unnatural from the perspective of neural networks and learning, the phase retrieval problem is a classic inverse problem in signal processing. It naturally appears in problems where one seeks to measure a signal with a detector which can only capture the amplitude, but not the phase, such as in X-ray crystallography and astronomical imaging. See (Dong et al., 2023; Jaganathan et al., 2016) for some examples of works in this context. Here, we study it as a proxy for the simplest non-convex optimisation problem beyond least-squares, which as we will see contain already some of the more complicated features of more general problems.

Let's start by computing the gradient:

$$g^\nu := \nabla_w \ell(w^\nu) = -\left(\langle w^\star, x^\nu\rangle^2 - \langle w^\nu, x^\nu\rangle^2 + z^\nu\right) 2\langle w^\nu, x^\nu\rangle x^\nu \tag{3.21}$$

As before, a good starting point is to look at the gradient flow limit. As in the linear case, the expected gradient can be computed exactly, with the difference that it now involves fourth moments of a Gaussian variable:

$$\mathbb{E}[g^\nu] = -2/d\left(||w^\star||_2^2 - 3||w||_2^2\right) w - 2/d\langle w^\star, w\rangle w^\star \tag{3.22}$$

Note that's very different from the linear case eq. (3.4): the zero initialisation $w^0 = 0_d$ is a fixed point of gradient flow. Moreover, the component pointing towards to the signal $w^\star$ is proportional to the

[6] Note one could also have considered $\sigma(x) = |x|$, but since this is not differentiable, it will complicate the algebra.

overlap $m = {}^1/_d\langle w^\star, w\rangle$, meaning that with a random initialisation $w^0, w^\star \sim \text{Uni}(\mathbb{S}^{d-1}(\sqrt{d}))$, we have $m \sim {}^1/\sqrt{d}$ and therefore the step towards the signal is vanishingly small in high-dimensions. Indeed, since the signal component is proportional to the correlation, for large but finite d one might require several steps to converge to the global minima. From a physical point of view, this is an entropic phenomena: finding a single direction $w^\star$ in $\mathbb{R}^d$ is like finding a needle in the haystack.

As before, analysing the summary statistics evolution eq. (2.26) involves tracking a coupled system of ODEs, which in this case read:

$$\dot{m}(t) = 6\ m(t)(\rho - q(t)) \tag{3.23}$$

$$\begin{aligned}\dot{q}(t) = 4\left(q(t)(\rho - 3q(t)) + 2m(t)^2\right) \\ + 12\gamma\left(q(t)(\rho^2 + 5q(t)^2 - 2\rho q(t)) + 4m^2(\rho - 2q(t))\right)\end{aligned} \tag{3.24}$$

This time, the evolution of m is coupled non-trivially with q.

Since most of the interesting phenomenology happens on the evolution of m, to simplify our life we will consider a spherical variant of SGD:

$$w^{\nu+1} = \sqrt{d}\frac{w^\nu - \gamma\nabla_{\mathbb{S}^{d-1}(\sqrt{d})}\ell(w^\nu)}{||w^\nu - \gamma\nabla_{\mathcal{S}^{d-1}(\sqrt{d})}\ell(w^\nu)||} \tag{3.25}$$

where $\nabla_{\mathcal{S}^{d-1}(\sqrt{d})}$ denotes de spherical gradient:

$$\begin{aligned}v = \nabla_{\mathbb{S}^{d-1}(\sqrt{d})}\ell(w) &:= \left(I_d - \frac{ww^\top}{d}\right)\nabla_w\ell(w) \\ &= g - {}^1/_d\langle g, w\rangle w \\ &= ((\lambda^\star)^2 - (\lambda^\nu)^2 + z^\nu)2\lambda^\nu(x^\nu - \lambda^\nu w^\nu)\end{aligned} \tag{3.26}$$

In particular, note that the spherical gradient is orthogonal to $w \in \mathbb{S}^{d-1}(\sqrt{d})$:

$$\langle v, w\rangle = \langle g, w\rangle - \langle g, w\rangle = 0 \tag{3.27}$$

This algorithm ensures that if we initialise $w^0 \sim \text{Uni}(\mathbb{S}^{d-1}(\sqrt{d}))$, the norm $q^\nu = {}^{||w^\nu||_2^2}/_d = 1$ stays constant for any $\nu \geq 0$. In order to derive ODEs for this spherical gradient, we need to have a closer look at the denominator:

$$\begin{aligned}\frac{\sqrt{d}}{||w^\nu - \gamma\nabla_{\mathbb{S}^{d-1}(\sqrt{d})}\ell(w^\nu)||} &= \sqrt{d}\left(||w^\nu - \gamma v^\nu||^2\right)^{-1/2} \\ &= \sqrt{d}\left(||w^\nu||^2 - 2\gamma\langle w^\nu, v^\nu\rangle + \gamma^2||v^\nu||^2\right)^{-1/2} \\ &= \sqrt{d}\left(d - \gamma^2||v^\nu||^2\right)^{-1/2} \\ &= \left(1 + \frac{\gamma^2}{d}||v^\nu||^2\right)^{-1/2}\end{aligned} \tag{3.28}$$

where we used the orthogonality of the spherical gradient $\langle v, w\rangle = 0$. One can check that $||v^\nu||^2$ is a $\Theta(1)$ quantity:

$$\begin{aligned}||v^\nu||_2^2 &= ((\lambda^{\star\nu})^2 - (\lambda^\nu)^2 + z^\nu)^2 4(\lambda^\nu)^2||x^\nu - \lambda^\nu w^\nu||_2^2 \\ &= ((\lambda^{\star\nu})^2 - (\lambda^\nu)^2 + z^\nu)^2 4(\lambda^\nu)^2\left(||x^\nu||^2 - (\lambda^\nu)^2\right)\end{aligned} \tag{3.29}$$

Therefore, when $d \gg 1$ we can expand $(1+x)^{-1/2} \approx 1 - {}^x/_2$ to get:

$$\frac{\sqrt{d}}{||w^\nu - \gamma v^\nu||} = 1 - \frac{\gamma}{2d}||v^\nu||^2 + o(d^{-1}) \tag{3.30}$$

Putting together, we have:

$$\begin{aligned} m^{\nu+1} &= \left(m^\nu - \frac{\gamma}{d}\langle w^\star, v^\nu\rangle\right)\left(1 - \frac{\gamma^2}{2d}||v^\nu||^2 + o(d^{-1})\right) \\ &= m^\nu - \frac{\gamma}{d}\langle w^\star, v\rangle - \frac{m^\nu\gamma^2}{2d}||v^\nu||_2^2 + o(d^{-1}) \\ &= m^\nu + \frac{2\gamma}{d}\left(\mathcal{E}^\nu\lambda^\nu\left(\lambda^\star - m^\nu\lambda^\nu\right) - \gamma m^\nu(\mathcal{E}^\nu)^2\left(||x^\nu||^2 - (\lambda^\nu)^2\right)\right) \end{aligned} \tag{3.31}$$

Taking the limit $d \to \infty$ with step-size $\delta t = \gamma/d$ yields the evolution for m:

$$\begin{aligned} \dot{m}(t) &= 2\,\mathbb{E}\left[\mathcal{E}\lambda(\lambda^\star - m(t)\lambda) - \gamma m(t)\mathcal{E}^2(1-\lambda^2)\right] \\ &= m(t)\left[4(1-6\gamma)(1-m(t)^2) - 2\gamma\Delta\right] \end{aligned} \tag{3.32}$$

Remarks:

- As anticipated from the discussion of the population gradient, $m = 0$ is a fixed point of the equation above. How long does it take to escape it? Considering for simplicity the gradient flow limit $\gamma \to 0^+$ and letting $m = \epsilon \ll 1$, we can keep only the higher-order terms in the right-hand side of the ODEs:

  $$\dot{\epsilon} \approx 4\epsilon \tag{3.33}$$

  which has solution $\epsilon(t) = e^{4t}\epsilon_0$. So the time it takes for to develop order one correlation $\epsilon(t) \approx 1$ from $\epsilon_0 \sim 1/d$ is given by

  $$t \approx \frac{1}{8}\log d \tag{3.34}$$

  Or in terms of sample complexity, $n = \Theta(d\log d)$. This scaling has been derived by different authors in the literature (Chen et al., 2019; Sun et al., 2016; Tan and Vershynin, 2018), including the exact escape time from the leading order stochastic correction to eq. (3.32) (Arnaboldi et al., 2024). Variants of the phase retrieval problem have been studied in the $r = 1, p > 1$ case by Arnaboldi et al. (2024) and $r, p > 1$ by Martin et al. (2023). In both cases, it was shown that overparametrisation does can only improve this sample complexity by a constant factor.

- Requiring that the first term (drift) to be positive implies we must have $\gamma \in [0, 1/6]$. Note that the critical learning rate in this case is smaller than in the least squares case.

- There are two more fixed points, given by:

  $$m_\pm = \pm\sqrt{1 - \frac{\gamma\Delta}{2(1-6\gamma)}} \tag{3.35}$$

  In the gradient flow limit $\gamma \to 0^+$, this gives $m_\pm = \pm 1$ which correspond to convergence to the global minima $\pm w^\star$.

3.3 Generalised linear models

A natural extension of the phase retrieval case is generalised linear estimation, where instead of the square activation, we allow for a generic non-linear activation σ. In other words, data $(x^\nu, y^\nu)_{\nu\in[n]}$ is generated from:

$$y^\nu = \sigma(\langle w^\star, x^\nu\rangle) + z^\nu, \qquad x^\nu \sim \mathcal{N}(0, 1/dI_d), \qquad z^\nu \sim \mathcal{N}(0, \Delta). \tag{3.36}$$

which we seek to learn by doing one-pass SGD on a generalised linear model $f_\theta(x) = \sigma(\langle w, x\rangle)$. Note that unless $\sigma(x) = x$, the loss is a non-convex function of the parameters w. This problem is also known as the *teacher-student perceptron* problem in the statistical physics of learning literature, and sometimes also called a *single-index model* in the context of learning.[7]

In the following, we will closely follow the argument by (Damian et al., 2023) for recovering the result of (Ben Arous et al., 2021) from the ODEs. As for the phase retrieval problem, to simplify the algebra we will again focus on spherical SGD:

$$w^{\nu+1} = \sqrt{d}\frac{w^\nu - \gamma v^\nu}{||w^\nu - \gamma v^\nu||} \tag{3.37}$$

with $w^\star, w \in \mathbb{S}^{d-1}(\sqrt{d})$ and v^ν the spherical gradient. given by:

$$v^\nu := \nabla_{\mathbb{S}^{d-1}(\sqrt{d})}\ell(w^\nu) = -(\sigma(\lambda^{\star\nu}) - \sigma(\lambda^\nu) + z^\nu)\sigma'(\lambda^\nu)(x - \lambda^\nu w) \tag{3.38}$$

As in the phase retrieval case, to get the ODEs, we expand the denominator and write an stochastic process for m^ν:

$$m^{\nu+1} = \left(m^\nu - \frac{\gamma}{d}\langle w^\star, v^\nu\rangle\right)\left(1 - \frac{\gamma}{2d}||g^\nu||^2 + o(d^{-1})\right) \tag{3.39}$$

$$\begin{aligned} m^{\nu+1} &= m^\nu + \frac{\gamma}{d}\langle w^\star, v^\nu\rangle - \frac{\gamma}{2d}m^\nu||v^\nu||_2^2 + o(d^{-1}) \\ &= m^\nu + \frac{\gamma}{d}\left(\mathcal{E}^\nu\sigma'(\lambda^\nu)\left(\lambda^{\star\nu} - m^\nu\lambda^\nu\right) - \frac{\gamma}{2}m^\nu(\mathcal{E}^\nu)^2\sigma'(\lambda^\nu)^2(||x^\nu||_2^2 - \lambda^{\nu 2})\right) + o(d^{-1}) \end{aligned} \tag{3.40}$$

In the $d \to \infty$ limit with vanishing step-size $\delta t = 1/d$, this gives the following ODE:

$$\dot{m} = \gamma\mathbb{E}\left[\mathcal{E}\sigma'(\lambda)\left(\lambda^\star - m\lambda\right) - \frac{\gamma}{2}m\mathcal{E}^2\sigma'(\lambda)^2(1 - \lambda^2)\right] \tag{3.41}$$

Note that in order for this ODE to be contracting, the first term must dominate over the second:

$$\gamma \leq \frac{2}{m(t)}\frac{\mathbb{E}[\langle w^\star, v(t)\rangle]}{\mathbb{E}[||v(t)||^2]} = \frac{2}{m(t)}\frac{\mathbb{E}[\mathcal{E}\sigma'(\lambda)\left(\lambda^\star - m\lambda\right)]}{\mathbb{E}[\mathcal{E}^2\sigma'(\lambda)^2(1 - \lambda^2)]} \tag{3.42}$$

In fact, this ratio can be interpreted as a signal to noise ratio in the problem, the signal being given by the correlation of the gradient with the global minimum $w^\star$ and the noise given by the norm of the gradient. To simplify things, we can take an optimal learning rate schedule, i.e. the largest learning rate such that the problem is still contracting. This simplifies the ODE to:

$$\dot{m}(t) = \frac{1}{2m(t)}\frac{\mathbb{E}[\langle w^\star, v(t)\rangle]^2}{\mathbb{E}[||v(t)||^2]} = \frac{1}{2m(t)}\frac{\mathbb{E}[\mathcal{E}\sigma'(\lambda)\left(\lambda^\star - m\lambda\right)]^2}{\mathbb{E}[\mathcal{E}^2\sigma'(\lambda)^2(1 - \lambda^2)]} \tag{3.43}$$

Therefore, we now need to deal with the expectations above. The key idea to deal with a general non-linear function σ is to expand it in a suitable basis. Since the argument of σ (i.e. the pre-activation $\lambda_\star, \lambda$) are jointly Gaussian variables, a natural option to simplify the expectations involved in the problem is to choose an orthogonal basis with respect to the Gaussian measure - the Hermite polynomials, which we now review. The Hermite polynomials $(\mathrm{He}_k(x))_{k\geq 0}$ are a family of polynomials of increasing degrees, with the first few given by:[8]

$$\mathrm{He}_0(x) = 1, \qquad \mathrm{He}_1(x) = x, \qquad \mathrm{He}_2(x) = x^2 - 1, \qquad \mathrm{He}_3(x) = x^3 - 3x \tag{3.44}$$

The normalised Hermite polynomials satisfy the following useful properties.

[7] Although some people would reserve the *single-index* terminology exclusively to misspecified problems where one needs to learn the so-called link function of the target $\sigma_\star$.

[8] Note that are different normalisation conventions for the Hermite polynomials. Here we adopt what is commonly known as the probabilist Hermite polynomials

Completness: Let μ denote the Gaussian measure, and recall:

$$L_2(\mu) = \left\{ f : \mathbb{R} \to \mathbb{R}, \int \mathrm{d}\mu(x) f(x)^2 < \infty \right\} \tag{3.45}$$

Then, any $f \in L_2(\mu)$ admits a decomposition in terms of $\mathrm{He}_k(x)$:

$$f(x) = \sum_{k \geq 0} \frac{a_k}{k!} \mathrm{He}_k(x) \tag{3.46}$$

where:

$$a_k = \mathbb{E}_{x \sim \mathcal{N}(0,1)} \left[f(x) \mathrm{He}_k(x) \right] \tag{3.47}$$

Orthogonality:

$$\mathbb{E}_{x \sim \mathcal{N}(0,1)} [\mathrm{He}_k(x) \mathrm{He}_l(x)] = k! \delta_{jl} \tag{3.48}$$

Derivative:

$$\mathrm{He}_k'(x) = k \mathrm{He}_{k-1}(x) \tag{3.49}$$

In particular, this implies that for $f \in L_2(\mu)$:

$$f'(x) = \sum_{k \geq 1} \frac{f_k}{k!} k h_{k-1}(x) = \sum_{k \geq 0} \frac{f_{k+1}}{k!} h_k(x) \tag{3.50}$$

Correlation: Consider $x, x' \sim \mathcal{N}(0,1)$ with correlation $\mathbb{E}[xx'] = \rho$. Then:

$$\mathbb{E}[\mathrm{He}_k(x) \mathrm{He}_l(x')] = k! \rho^k \delta_{kl} \tag{3.51}$$

We refer the reader to Chapter 11 of (O'Donnell, 2021) for a deeper dive into Hermite polynomials.

Completeness allow us to expand the GLM in the Hermite basis:

$$\sigma(\langle w, x \rangle) = \sum_{k \geq 0} \frac{\sigma_k}{k!} \mathrm{He}_k(\langle w, x \rangle), \qquad \sigma(\langle w^\star, x \rangle) = \sum_{k \geq 0} \frac{\sigma_k}{k!} \mathrm{He}_k(\langle w^\star, x \rangle) \tag{3.52}$$

As a warm-up, let's compute how the population risk depends on the summary statistic m:

$$\begin{aligned} \mathcal{R}(w) - \Delta/2 &= 1/2 \mathbb{E}[(\sigma(\langle w^\star, x \rangle) - \sigma(\langle w, x \rangle))^2] \\ &= 1/2 \left(\mathbb{E}[\sigma(\lambda^\star)^2] + \mathbb{E}[\sigma(\lambda)^2] \right) - \mathbb{E}[\sigma(\lambda_\star)\sigma(\lambda)] \end{aligned} \tag{3.53}$$

where we recall $\lambda_\star = \langle w^\star, x \rangle$ and $\lambda = \langle w, x \rangle$. Note that the first term are simple constants:

$$\mathbb{E}[\sigma(\lambda^\star)^2] = \mathbb{E}[\sigma(\lambda)^2] = \sum_{k,l \geq 0} \frac{\sigma_k \sigma_l}{k! l!} \mathbb{E}_{z \sim \mathcal{N}(0,1)} [\mathrm{He}_k(z) \mathrm{He}_l(z)] = \sum_{k,l \geq 0} \frac{\sigma_k \sigma_l}{k! l!} k! \delta_{kl} = \sum_{k \geq 0} \frac{\sigma_k^2}{k!} \tag{3.54}$$

$$\tag{3.55}$$

While the cross term is given by:

$$\mathbb{E}[\sigma(\lambda_\star)\sigma(\lambda)] = \sum_{k,l \geq 0} \frac{\sigma_k \sigma_l}{k! l!} \mathbb{E}[\mathrm{He}_k(\lambda^\star) \mathrm{He}_l(\lambda)] = \sum_{k,l \geq 0} \frac{\sigma_k \sigma_l}{k! l!} k! \delta_{kl} m^k = \sum_{k \geq 0} \frac{\sigma_k^2}{k!} m^k \tag{3.56}$$

Putting together:

$$\mathcal{R}(w) - \Delta/2 = \sum_{k\geq 0} \frac{\sigma_k^2}{k!}(1 - m^k) \tag{3.57}$$

Note that since $\nabla_{\mathbb{S}^{d-1}(\sqrt{d})} = 1/d\ (w^\star - mw)$, the population (spherical) gradient is given by:

$$\mathbb{E}[v^\nu] = \nabla_{\mathbb{S}^{d-1}(\sqrt{d})}\mathcal{R}(w) = -\sum_{k\geq 0} \frac{\sigma_k^2}{k!d} k m^{k-1}(w^\star - mw) \tag{3.58}$$

which implies that:

$$\mathbb{E}[\langle v^\nu, w^\star\rangle] = -\sum_{k\geq 0} \frac{\sigma_k^2}{(k-1)!} m^{k-1}(1 - m^2) \tag{3.59}$$

Comparing to the phase retrieval case eq. (3.21), we see that again, although the gradient points towards the global minimum $w^\star$, it involves a factor m^{k-1}. Since at initialisation $m \sim 1/\sqrt{d}$, for large d one needs to overcome a vanishing gradient which is dominated by the first non-vanishing Hermite coefficient $k^\star$, also known as the information exponent (Ben Arous et al., 2021):

$$\mathbb{E}[\langle v^\nu, w^\star\rangle] = -\frac{\sigma_{k^\star}^2}{(k^\star - 1)!} m^{k^\star - 1} + O(m^k) \tag{3.60}$$

It remains to deal with $\mathbb{E}[||v||^2]$. Since v is a random vector in dimension d with components $\Theta(d^{-1/2})$, we have $\mathbb{E}[||v||_2^2] = \Theta(1)$. Putting together, expanding the right-hand side of eq. (3.43) around $m = 0$ give the following ODE:

$$\dot{m} = C m^{2k^\star - 3} \tag{3.61}$$

for some constant $C > 0$. For $k^\star > 2$, this has solution a solution which corresponds to a escape time $T = O(d^{k^\star - 2})$. Noting that $T = n/d$, this gives a sample complexity $n = \Theta(d^{k^\star - 1})$. A few comments on this result:

- Note this sample complexity is far from optimal. Indeed, other algorithms are known to achieve perfect recovery with sample complexity $n = \Theta(d)$. An example is approximate message passing, see (Barbier et al., 2019).
- It can be shown that by smoothening the loss function the sample complexity can be improved to $n = \Theta(n^{k^\star/2 - 1})$, which corresponds to the optimal sample complexity in the class of Correlational Statistical Query (CSQ) algorithms (which includes one-pass SGD), see (Damian et al., 2023). This draws back to landscape smoothening ideas from statistical physics (Biroli et al., 2020).
- We have considered a well-specified setting where $p = r = 1$ and $\sigma_\star(x) = \sigma(x)$. For $\sigma \neq \sigma_\star$, one must have $p > 1$ (often $p \to \infty$) in order for the problem to be realisable, even if $r = 1$. Berthier et al. (2023) has shown that a two-layer neural network with p large enough can learn a single-index target function with $k^\star > 1$ with sample complexity $n = \Theta(d)$. Going beyond $k^\star > 1$ and generally understanding the how p must scale with d in the high-dimensional limit is an important open problem.
- The notion of information exponent can be extended to multi-index target functions $r > 1$, where it is known as leap index (Abbe et al., 2023). However, since the directions can be coupled in different ways, this makes the characterisation of which functions are "hard" to learn richer.

References

Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2552–2623. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/abbe23a.html.

Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless odes: A unifying approach to sgd in two-layers networks. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1199–1227. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/arnaboldi23a.html.

Luca Arnaboldi, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Escaping mediocrity: how two-layer networks learn hard generalized linear models with sgd, 2024.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. URL http://jmlr.org/papers/v18/14-546.html.

Francis Bach and Lenaic Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization, 2021.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019. doi: 10.1073/pnas.1802705116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1802705116.

A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993. doi: 10.1109/18.256500.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106): 1–51, 2021. URL http://jmlr.org/papers/v22/20-1288.html.

Yoshua Bengio, Nicolas Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. URL https://proceedings.neurips.cc/paper_files/paper/2005/file/0fc170ecbb8ff1afb2c6de48ea5343e7-Paper.pdf.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks, 2023.

Giulio Biroli, Chiara Cammarota, and Federico Ricci-Tersenghi. How to iron out rough landscapes and get optimal performances: averaged gradient descent and its application to tensor pca. *Journal of Physics A: Mathematical and Theoretical*, 53(17):174003, apr 2020. doi: 10.1088/1751-8121/ab7b1f. URL https://dx.doi.org/10.1088/1751-8121/ab7b1f.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33,

pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1):5–37, Jul 2019. ISSN 1436-4646. doi: 10.1007/s10107-019-01363-6.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 752–784. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/02763667a5761ff92bb15d8751bcd223-Paper-Conference.pdf.

Jonathan Dong, Lorenzo Valzania, Antoine Maillard, Thanh-an Pham, Sylvain Gigan, and Michael Unser. Phase retrieval: From computational imaging to machine learning: A tutorial. *IEEE Signal Processing Magazine*, 40(1):45–57, 2023. doi: 10.1109/MSP.2022.3219240.

E Gardner and B Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, jun 1989. doi: 10.1088/0305-4470/22/12/004. URL https://dx.doi.org/10.1088/0305-4470/22/12/004.

Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/cab070d53bd0d200746fb852a922064a-Paper.pdf.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Phys. Rev. X*, 10: 041044, Dec 2020. doi: 10.1103/PhysRevX.10.041044. URL https://link.aps.org/doi/10.1103/PhysRevX.10.041044.

Kishore Jaganathan, Yonina C. Eldar, and Babak Hassibi. Phase Retrieval: An Overview of Recent Developments. CRC Press, 2016. ISBN 9781315371474.

V. Kurkova and M. Sanguineti. Bounds on rates of variable-basis and neural-network approximation. *IEEE Transactions on Information Theory*, 47(6):2659–2665, 2001. doi: 10.1109/18.945285.

Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. Bad global minima exist and sgd can reach them. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8543–8552. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/618491e20a9b686b79e158c293ab4f91-Paper.pdf.

Simon Martin, Francis Bach, and Giulio Biroli. On the impact of overparameterization on the training of a shallow neural network in high dimensions, 2023.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi: 10.1073/pnas.1806579115. URL https://www.pnas.org/doi/abs/10.1073/pnas.1806579115.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/40008b9a5380fcacce3976bf7c08af5b-Paper.pdf.

Ryan O'Donnell. Analysis of boolean functions, 2021.

G. Reents and R. Urbanczik. Self-averaging and on-line learning. *Phys. Rev. Lett.*, 80:5445–5448, Jun 1998. doi: 10.1103/PhysRevLett.80.5445.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Saharon Rosset, Grzegorz Swirszcz, Nathan Srebro, and Ji Zhu. ℓ_1 regularization in infinite dimensional feature spaces. In Nader H. Bshouty and Claudio Gentile, editors, *Learning Theory*, pages 544–558, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-72927-3.

Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9): 1889–1935, 2022. doi: https://doi.org/10.1002/cpa.22074. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22074.

David Saad and Sara Solla. Dynamics of on-line gradient descent learning for multilayer neural networks. In D. Touretzky, M. C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1996.

David Saad and Sara A. Solla. On-line learning in soft committee machines. *Phys. Rev. E*, 52:4225–4243, Oct 1995a. doi: 10.1103/PhysRevE.52.4225. URL https://link.aps.org/doi/10.1103/PhysRevE.52.4225.

David Saad and Sara A. Solla. Exact solution for on-line learning in multilayer neural networks. *Phys. Rev. Lett.*, 74:4337–4340, May 1995b. doi: 10.1103/PhysRevLett.74.4337.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2019.06.003. URL https://www.sciencedirect.com/science/article/pii/S0304414918306197.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 2379–2383, 2016. doi: 10.1109/ISIT.2016.7541725.

Yan Shuo Tan and Roman Vershynin. Phase retrieval via randomized Kaczmarz: theoretical guarantees. *Information and Inference: A Journal of the IMA*, 8(1):97–123, 04 2018. ISSN 2049-8772. doi: 10.1093/imaiai/iay005.

Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks.

In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23244–23255. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/939bb847ebfd14c6e4d3b5705e562054-Paper-Conference.pdf.

Chuang Wang, Yonina C. Eldar, and Yue M. Lu. Subspace estimation from incomplete observations: A high-dimensional analysis. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1240–1252, 2018. doi: 10.1109/JSTSP.2018.2877405.

Chuang Wang, Hong Hu, and Yue Lu. A solvable high-dimensional model of gan. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/6b3c49bdba5be0d322334e30c459f8bd-Paper.pdf.