Notes on one-pass SGD for ACDL 2022

Bruno Loureiro[1]

[1]Information, Learning and Physics lab., EPFL

August 26, 2022

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 $(\boldsymbol{x}^{\nu}, y^{\nu})_{\nu\in[n]} \in \mathbb{R}^d \times \mathbb{R}$ 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}. But how do we search for f_θ in practice (or equivalently, for $\theta \in \mathbb{R}^m$)?

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**:

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

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**:

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

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) [1].

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 p 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**:

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

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

An issue with 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) [2]:

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

Together with its variants, SGD is one of the most used algorithm in modern machine learning. Besides being more efficient from GD, one advantage of SGD is that it can be seen as an approximation for gradient flow on the population risk $\dot{\theta} = -\nabla_\theta \mathcal{R}$. 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. 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 $|B_k| = 1$. Note that in this case, the 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 steps needed. Therefore, one-pass SGD can be seen as a noisy version of gradient descent on the population risk:

$$\theta^{\nu+1} = \theta^\nu - \gamma_\nu \nabla_\theta \mathcal{R}(\theta^\nu) + \gamma_\nu \varepsilon_\nu \tag{1.1}$$

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

$$\varepsilon_\nu \equiv \nabla_\theta \ell\left(y^\nu, f_{\theta^\nu}(x^\nu)\right) - \nabla_{\theta^\nu} \mathcal{R}(\theta^\nu) = \nabla_\theta \ell\left(y^\nu, f_{\theta^\nu}(x^\nu)\right) - \mathbb{E}\left[\nabla_{\theta^\nu} \ell(y, f_{\theta^\nu}(x))\right] \tag{1.2}$$

which has zero mean since the estimation is unbiased[2]. This observation is striking: up to noise, one-pass SGD optimises the true population risk even though this is an unknown function!

This discussion leads to our main question in these notes. Can we characterise the one-pass SGD dynamics? For instance, can we understand the role of the effective noise? What is the interplay between the architecture f_θ and the optimisation algorithm?

2 Sharp analysis of one-pass SGD

There are five key ingredients that define the one-pass SGD dynamics: the loss function ℓ, the parametric family f_θ (a.k.a. architecture), the data distribution ρ and the learning rate γ_k. To have a sharp description of the dynamics induced by the algorithm, we need to specify these quantities.

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

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

Architecture: In these notes, we will be interested in the simplest architecture leading to a non-convex 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[3]. 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 $w_i \in \mathbb{R}^p$ in the columns.

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

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

Data: While much can be said about the dynamics without specifying the data model, if we aim at a sharp characterisation we need to compute the averages over the data exactly (for instance in the definition of $\mathcal{R}$). Therefore, we need a model for the data distribution ρ. Here, we will consider a **teacher-student model**[4], where the labels are generated, up to an additive noise, by a similar **teacher** architecture, i.e. a two-layer neural network:

$$y^\nu = \underbrace{\frac{1}{k}\sum_{\rho=1}^{k} a^\star_\rho \sigma_\star(w^\star_\rho \cdot x^\nu)}_{f_\star(x^\nu)} + \sqrt{\Delta}\xi^\nu, \qquad\qquad \xi^\nu \sim \mathcal{N}(0,1), \quad x^\nu \sim \mathcal{N}(0, {}^1\!/_d I_d) \quad \text{i.i.d.}. \tag{2.3}$$

We refer to $f_\star$ as the **teacher** (or target function), and to $\theta^\star = (a^\star, W^\star) \in \mathbb{R}^k \times \mathbb{R}^{k\times d}$ as the **teacher parameters**. Here we have assumed the data is Gaussian i.i.d. for simplicity, but the discussion that will follow is valid for more general distributions of inputs [9]. Note that if $p = k$, the global minimum of the population risk is given, up to permutation symmetry, by the teacher parameters θ^*. Therefore, in the teacher-student model learning boils down to a statistical estimation problem. The same discussion holds in the case where $p > k$, where one can also achieve perfect estimation, corresponding to $\mathcal{R} = \Delta$ due to the intrinsic noise (however, the exact global minimum can be more complicated). These cases are sometimes referred to as **realisable setting**.

In what follows we will focus on the realisable setting with $p \geq k$ and $\sigma_\star = \sigma$. Moreover, to simplify the discussion we will fix the second layer weights $a^\star_\rho = a_i = 1$ for all $\rho \in [k]$ and $i \in [p]$. Everything we will see can be generalised to the case in which these weights are trained, and we refer the interested reader to [10] for the details.

2.1 Sufficient statistics and order parameters

For our teacher-student model introduced above, the population risk can be written very explicitly as:

$$\begin{aligned}\mathcal{R}(w) &= \mathbb{E}_{(x,y)\sim\rho}\left[\frac{1}{2}\left(y - f_\theta(x)\right)^2\right] = \mathbb{E}_{x,\xi}\left[\frac{1}{2}\left(\frac{1}{k}\sum_{\rho=1}^{k}\sigma(w^\star_\rho \cdot x) + \sqrt{\Delta}\xi - \frac{1}{p}\sum_{i=1}^{p} a_i\sigma(w_i \cdot x)\right)^2\right] \\ &= \mathbb{E}_{x\sim\mathcal{N}(0,{}^1\!/_d I_d)}\left[\frac{1}{2}\left(\frac{1}{k}\sum_{\rho=1}^{k}\sigma(w^\star_\rho \cdot x) - \frac{1}{p}\sum_{i=1}^{p}\sigma(w_i \cdot x)\right)^2\right] + \Delta \end{aligned} \tag{2.4}$$

[3] Indeed, the normalisation is chosen to match the mean-field literature [5–8].

[4] This terminology for the generative model for data was introduced by the physicist Elisabeth Gardner in [4], and it is standard in Statistical Physics of learning.

where in the third equality we used the fact that the noise is independent from the inputs. Note that in order to compute the average, all we need to know is the joint distribution of $\lambda^\star_\rho = w^\star_\rho \cdot x$ and $\lambda_i = w_i \cdot x$:

$$\mathcal{R}(w) = \mathbb{E}_{(\lambda^\star,\lambda)}\left[\frac{1}{2}\left(\frac{1}{k}\sum_{\rho=1}^{k}\sigma(\lambda^\star_\rho) - \frac{1}{p}\sum_{i=1}^{p}\sigma(\lambda_i)\right)^2\right] + \Delta \tag{2.5}$$

Although this might seen a simple observation, it is not trivial, since it tell us that in order to know the error it is sufficient to track only $k+p$ quantities instead of $(k+p)d$. The sufficient statistics $(\lambda^\star, \lambda) \in \mathbb{R}^{k+p}$ are known in the statistical physics literature as **local fields**. Note that since $x \sim \mathcal{N}(0, 1/dI_d)$, at every step $\nu \in [n]$ of SGD the local fields are jointly Gaussian variables:

$$(\lambda^\star, \lambda^\nu) = \mathcal{N}\left(0_{k+p}, \Omega^\nu\right), \qquad \Omega^\nu \equiv \begin{bmatrix} P & M^\nu \\ M^{\nu\top} & Q^\nu \end{bmatrix} \in \mathbb{R}^{(k+p)\times(k+p)} \tag{2.6}$$

where:

$$P = \frac{1}{d}W^\star W^{\star\top} \in \mathbb{R}^{k\times k}, \qquad M^\nu = \frac{1}{d}W^\star W^{\nu\top} \in \mathbb{R}^{k\times p}, \qquad Q^\nu = \frac{1}{d}W^\nu W^{\nu\top} \in \mathbb{R}^{p\times p}. \tag{2.7}$$

Note that by construction Ω^ν, P and Q^ν 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 Ω^ν, or equivalently (M^ν, Q^ν) since P is fixed. In the statistical physics parlour, these are our **order parameters**.

2.2 Low-dimensional dynamics

Since we only need to track Ω^ν to track the evolution of the error, can we derive closed-form evolution equations for Ω^ν? First, we need to compute the gradient of the loss. Defining the displacement vector at step ν:

$$\mathcal{E}^\nu = y^\nu - \frac{1}{p}\sum_{j=1}^{p}\sigma(w^\nu_j \cdot x^\nu) \tag{2.8}$$

such that $\ell(y, f_\theta(x)) = 1/2(\mathcal{E}^\nu)^2$, this is simply given by:

$$\nabla_{w_i}\ell(y, f_\theta(x)) = -\frac{1}{p}\underbrace{\mathcal{E}^\nu\sigma'(w^\nu_i \cdot x^\nu)}_{\equiv\mathcal{E}^\nu_i}x^\nu. \tag{2.9}$$

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

$$w^{\nu+1}_i = w^\nu_i + \frac{\gamma_\nu}{p}\mathcal{E}^\nu_i x^\nu \tag{2.10}$$

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

$$M^{\nu+1}_{\rho i} = M^\nu_{\rho i} + \frac{\gamma_\nu}{pd}\mathcal{E}^\nu_i\lambda^\star_\rho \tag{2.11}$$

Note that this is a stochastic differential equation (SDE) for the $k+p$ 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. (2.10) with respect to $w_j^{\nu+1}$:

$$w_i^{\nu+1} \cdot w_j^{\nu+1} = \left(w^{\nu} + \frac{\gamma_\nu}{p} \mathcal{E}_i^{\nu} x^{\nu} \right) \cdot w_j^{\nu+1} \tag{2.12}$$

and now we re-apply eq. (2.10), but now for $w_j^{\nu+1}$:

$$w_i^{\nu+1} \cdot w_j^{\nu+1} = \left(w^{\nu} + \frac{\gamma_\nu}{p} \mathcal{E}_i^{\nu} x^{\nu} \right) \left(w_j^{\nu} + \frac{\gamma_\nu}{p} \mathcal{E}_j^{\nu} x^{\nu} \right) \tag{2.13}$$

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

$$Q_{ij}^{\nu+1} = Q_{ij}^{\nu} + \frac{\gamma_\nu}{pd} \left(\mathcal{E}_i^{\nu} \lambda_j^{\nu} + \mathcal{E}_j^{\nu} \lambda_i^{\nu} \right) + \frac{\gamma_\nu^2}{dp^2} \mathcal{E}_i^{\nu} \mathcal{E}_j^{\nu} \; ||x^{\nu}||_2^2 \tag{2.14}$$

Putting these two equations together, up to the norm $||x^{\nu}||_2^2$[5] we have obtained closed-form, low-dimensional SDEs for the evolution of the order parameters:

$$\Omega^{\nu+1} = \Psi(\Omega^{\nu}) \tag{2.15}$$

Note that evolution equations eqs. (2.15) are stochastic because they depend explicitly on the random variables $\{x^{\nu}\}_{\nu \in [n]}$.

Interpretation: Looking at the evolution equations, three types of terms appear:

$$\Psi_M \equiv \frac{\gamma_\nu}{pd} \mathcal{E}_i^{\nu} \lambda_\rho^{\star}, \qquad \Psi_Q^{(1)} \equiv \frac{\gamma_\nu}{pd} \left(\mathcal{E}_i^{\nu} \lambda_j^{\nu} + \mathcal{E}_j^{\nu} \lambda_i^{\nu} \right), \qquad \Psi_Q^{(2)} \equiv \frac{\gamma_\nu^2}{p^2 d} \mathcal{E}_i^{\nu} \mathcal{E}_j^{\nu} \; ||x^{\nu}||_2^2 \tag{2.16}$$

Note is that the prefactor of Ψ_M and $\Psi_Q^{(1)}$ are the same, and different from $\Psi_Q^{(2)}$. Since, $(\mathcal{E}_i^{\nu} \lambda_j^{\star}, \mathcal{E}_i^{\nu} \lambda_j^{\nu}, \mathcal{E}_i^{\nu} \mathcal{E}_j^{\nu})$ are $O(1)$ quantities, this means these terms have different scaling with respect to γ_ν/d. Moreover, it can be shown that the deterministic component of $(\Psi_M, \Psi_Q^{(1)})$ act as an attractive drift encouraging the alignement of W to $W^{\star}$, while $\Psi_Q^{(2)}$ is a repulsion on the weights W.

Indeed, the averaged terms $\bar{\Psi}_M \equiv \mathbb{E}\Psi_M$, $\bar{\Psi}_Q \equiv \mathbb{E}\Psi_Q^{(1)}$ have a natural interpretation: they correspond exactly to the terms coming from the gradient of the population risk in eq. (1.1):

$$\nabla_{w_i} \mathcal{R}(W) = \frac{1}{p} \nabla_{w_i} \mathbb{E}_x \left[\mathcal{E}_i x \right] \tag{2.17}$$

Taking the dot product of the above with respect to the weights $(w_\rho^{\star}, w_i)$ give precisely $\bar{\Psi}_M, \bar{\Psi}_Q^{(1)}$.

2.3 Deterministic limit(s)

So far, we have done no approximation: the SDEs above are exact and valid for any (γ_ν, p, k, d, n). Now we discuss different limits where the evolution above simplifies.

[5] Which concentrates fast for d large.

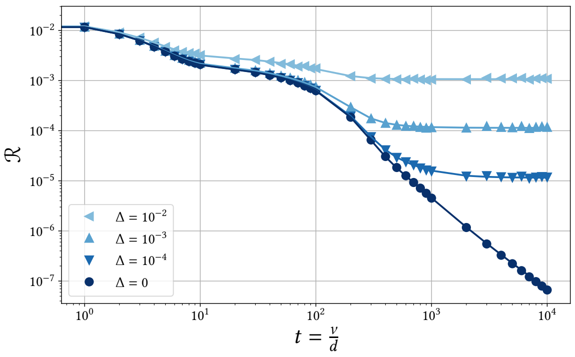


Figure 1: Population risk dynamics for the Saad & Solla scaling: $p = 8$, $k = 4$ with orthogonal initialisation. Activation: $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$. Dots represent simulations ($d = 1000$), while solid lines are obtained by integration of the ODEs given by eqs. (2.18).

Saad & Solla regime: Consider the case in which $\gamma_\nu \equiv \gamma$, p, k, n are fixed, $O(1)$ numbers, and we take the input dimension to be large $d \to \infty$. Defining $t = \nu\delta t$ with time-step $\delta t = {}^{1}\!/_{d}$, we can show that the order parameters concentrate, and the stochastic part of the SDE above is sub-leading in d. This leads to the following deterministic ODE for the evolution of the order parameters:

$$\dot{M}_{\rho i}(t) = \bar{\Psi}_M(P, M(t), Q(t)) = \frac{\gamma}{p}\mathbb{E}_{(\lambda^\star,\lambda)\sim\mathcal{N}(0,\Omega)}\left[\mathcal{E}_i\lambda^\star_\rho\right] \tag{2.18}$$

$$\begin{aligned}\dot{Q}_{ij}(t) &= \bar{\Psi}^{(1)}_Q\left(P, M(t), Q(t)\right) + \bar{\Psi}^{(2)}_Q\left(P, M(t), Q(t)\right)\\ &= \frac{\gamma}{p}\mathbb{E}_{(\lambda^\star,\lambda)\sim\mathcal{N}(0,\Omega)}\left[\mathcal{E}_i\lambda_j + \mathcal{E}_j\lambda_i\right] + \frac{\gamma^2}{p}\mathbb{E}_{(\lambda^\star,\lambda)\sim\mathcal{N}(0,\Omega)}\left[\mathcal{E}_i\mathcal{E}_j\right]\end{aligned} \tag{2.19}$$

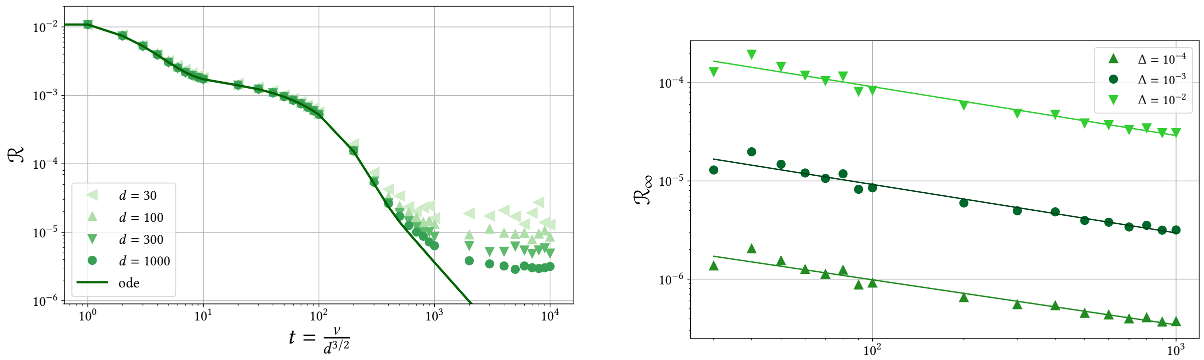
These equations have been first derived in 1995 by physicists David Saad and Sara Solla in [3], and have been rigorously proven in [10, 11]. Crucially, in this regime we have a competition between an attractive and a repulsive term, which can trap the dynamics for times $t \sim O(1)$. Indeed, in [10] it has been shown analytically that even in a realisable setting, for different architectures such as $\sigma \in \{\mathrm{relu}, \mathrm{erf}\}$ the asymptotic might not recover $W^\star$ in $t \sim O(1)$, and one can get trapped in a plateau $\mathcal{R}_\infty - \Delta \propto \Delta\gamma$, see Fig. 1. In particular, this plateau is due to the repulsion term $\bar{\Psi}^{(2)}_Q$, which is precisely the term not appearing in the population gradient flow dynamics. Indeed, this drift term is an intrinsic high-dimensional correction for the population gradient flow dynamics.

Perfect learning regime: As we have seen above, the SGD dynamics can get trapped even if, a priori, perfect prediction is possible. Can we actually improve on this? Recall our observation above: the repulsive term $\bar{\Psi}^{(2)}_Q$ which hinders perfect learning has a different scaling with respect to the attractive terms. Therefore, we might be able to get rid of this term by either: (a) overparametrising $p d d$ or (b) by choosing a fast decaying learning rate $\gamma \ll d$. Therefore, consider the case in which these quantities scale with d:

$$p \sim d^\kappa, \qquad\qquad \gamma \sim d^{-\delta} \tag{2.20}$$

with k and n fixed $O(1)$ numbers. Defining $t = \nu\delta t$ with time-step:

$$\delta t = \frac{\gamma}{pd} \sim \frac{1}{d^{1+\kappa+\delta}} \tag{2.21}$$



(a) Population risk dynamics for $\kappa = 0$ and $\delta = 1/2$. Fixed noise $\Delta = 10^{-3}$ and varying d. Dots represent simulations, while the solid line is obtained by integration of the ODEs given by eqs. (2.22). The data are compatible with the claim that as $d \to \infty$ the curve converges to zero population risk.

(b) Asymptotic population risk $\mathcal{R}_\infty - \Delta$ from simulations (dots) as a function of d for different noise levels under the scaling $\kappa = 0$ and $\delta = 1/2$. The fitted straight lines have slopes -0.458, -0.494, -0.497, for $\Delta = 10^{-4}, 10^{-3}, 10^{-2}$, respectively.

Figure 2: Network parameters: $p = 8$, $k = 4$ and orthogonal initialisation. Activation function: $\sigma(x) = \text{erf}(x/\sqrt{2})$.

we can show that for $\kappa > -\delta$, we the order parameters also concentrate, and the stochastic term vanishes. The resulting continuous limit dynamics is given in this case by:

$$\dot{M}_{\rho i}(t) = \bar{\Psi}_M(P, M(t), Q(t)) = \mathbb{E}_{(\lambda^\star,\lambda)\sim\mathcal{N}(0,\Omega)}\left[\mathcal{E}_i\lambda^\star_\rho\right] \tag{2.22}$$

$$\dot{Q}_{ij}(t) = \bar{\Psi}^{(1)}_Q\left(P, M(t), Q(t)\right) = \mathbb{E}_{(\lambda^\star,\lambda)\sim\mathcal{N}(0,\Omega)}\left[\mathcal{E}_i\lambda_j + \mathcal{E}_j\lambda_i\right]. \tag{2.23}$$

This concentration result has been recently shown in [12]. Note that we have completely got rid of $\bar{\Psi}^{(2)}_Q$! Therefore, in this regime one can achieve perfect, see Fig. 2 for a illustration. However, note that there is no free lunch: since the time scale $t = \nu\delta t$, by decreasing the learning rate γ we effective require more samples for a given time-step, so indeed it makes sense that by seeing more samples one can reach better learning on a finite time horizon. The same is true for the overparametrisation. Note that this result is consistent to the global convergence of overparametrised two-layer neural networks, and provides a bridge between the classical results for narrow networks [3] and the mean-field works [5–8].

Note that a consequence of this result is that for $\kappa = -\delta$ we get exactly the same equations as in the Saad and Solla regime where $\kappa = \delta = 0$. Therefore, in this case the phenomenology is exactly as in [3].

Bad learning regime: Although this is less interesting, we can also derive a deterministic limit in a regime where the $\Psi^{(2)}_Q$ term dominates the dynamics. This is given by choosing $t = \nu\delta t$ with:

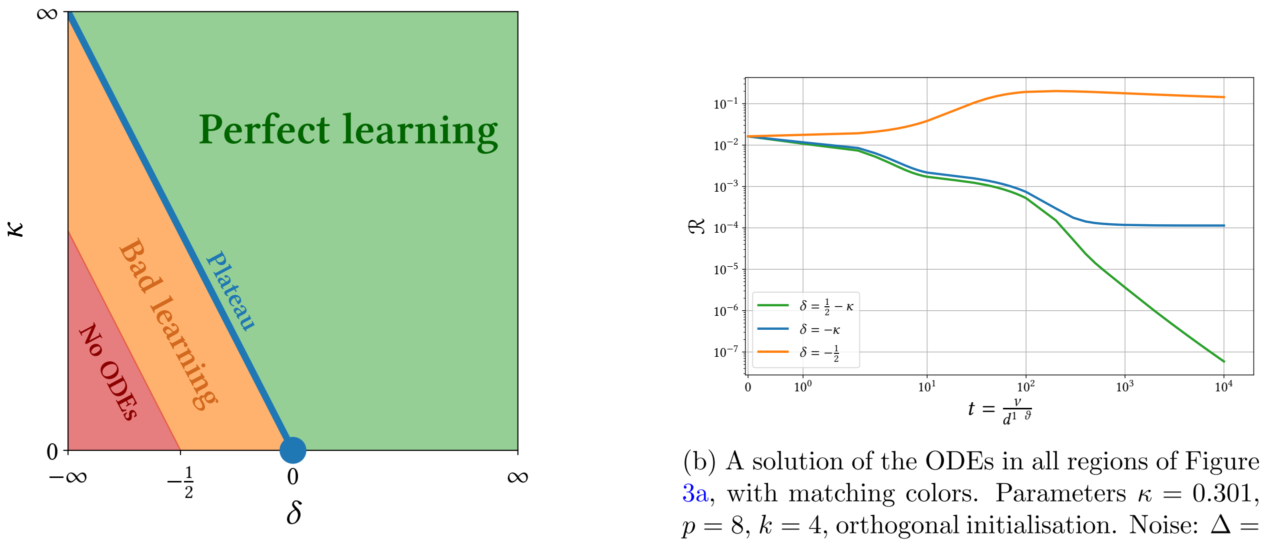
$$\delta t = \frac{\gamma^2}{p^2 d} \sim \frac{1}{d^{1+2(\kappa+\delta)}} \tag{2.24}$$

and taking $d \to \infty$ at fixed n. In this case, the resulting ODEs are simply:

$$\dot{M}_{\rho i}(t) = 0 \tag{2.25}$$

$$\dot{Q}_{ij}(t) = \bar{\Psi}^{(2)}_Q\left(P, M(t), Q(t)\right) = \mathbb{E}_{(\lambda^\star,\lambda)\sim\mathcal{N}(0,\Omega)}\left[\mathcal{E}_i\mathcal{E}_j\right]. \tag{2.26}$$

Since $\dot{M} = 0$, there is no learning at all in this regime.



(a) The phase diagram of SGD learning regimes for two-layer neural networks in the high-dimensional input layer limit $d \to \infty$. Eqs. (2.20) define proper time scalings for each of the regions. Perfect learning region: $\kappa + \delta > 0$. Plateau line: $\kappa + \delta = 0$. Bad learning region: $-1/2 < \kappa + \delta < 0$. No ODEs region: $\kappa + \delta < -1/2$.

(b) A solution of the ODEs in all regions of Figure 3a, with matching colors. Parameters $\kappa = 0.301$, $p = 8$, $k = 4$, orthogonal initialisation. Noise: $\Delta = 10^{-3}$. Activation function: $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$. The time scaling is not uniform through the phase diagram: $\vartheta = \kappa + \delta$ on green and blue regimes and $\vartheta = 2(\kappa + \delta)$ on the orange region. The green curve decays as a power law to zero excess error.

Figure 3: Phase diagram (left) and typical behavior of the ODE in each regions (right).

2.4 Summary

Since physicists love phase diagrams, we can summarise our discussion (and a bit more) in a phase diagram for the deterministic limit of the one-pass SGD dynamics, Fig. 3a.

References

[1] S. Liu, D. Papailiopoulos, D. Achlioptas, Bad Global Minima Exist and SGD Can Reach Them, Part of Advances in Neural Information Processing Systems 33 (NeurIPS 2020).

[2] H. Robbins, S. Monro. A Stochastic Approximation Method. The Annals of Mathematical Statistics, 22 (3): 400 - 407, September, 1951.

[3] Saad and S. A. Solla, On-line learning in soft committee machines, Phys. Rev. E, vol. 52, pp. 4225-4243, Oct 1995.

[4] E. Gardner, B. Derrida. Three unfinished works on the optimal storage capacity of networks. Journal of Physics A: Mathematical and General, 22(12):1983, 1989.

[5] S. Mei, A. Montanari, P.M. Nguyen, A mean field view of the landscape of two-layer neural networks, Proceedings of the National Academy of Sciences, vol. 115, no. 33, pp. E7665-E7671, 2018.

[6] Lénaïc Chizat, Francis Bach, On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport, Part of Advances in Neural Information Processing Systems 31 (NeurIPS 2018).

[7] G.M. Rotskoff, E. Vanden-Eijnden, Trainability and Accuracy of Neural Networks: An Interacting Particle System Approach, arXiv: 1805.00915 [stat.ML].

[8] J. Sirignano, K. Spiliopoulos, Mean field analysis of neural networks: A central limit theorem, Stochastic Processes and their Applications, vol. 130, no. 3, pp. 1820-1852, 2020.

[9] S. Goldt, M. Mézard, F. Krzakala, L. Zdeborová, Modeling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model, Phys. Rev. X 10, 041044 - Published 3 December 2020.

[10] S. Goldt, M. Advani, A.M. Saxe, F. Krzakala, L. Zdeborová, Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup, Part of Advances in Neural Information Processing Systems 32 (NeurIPS 2019).

[11] G. Reents, R. Urbanczik, Self-Averaging and On-Line Learning, Phys. Rev. Lett. 80, 5445 - Published 15 June 1998.

[12] R. Veiga, L. Stephan, B. Loureiro, F. Krzakala, L. Zdeborová, Phase diagram of Stochastic Gradient Descent in high-dimensional two-layer neural networks, arXiv: 2202.00293 [stat.ML].