

MLPGradientFlow: going with the flow of multilayer perceptrons (and finding minima fast and accurately)

Johanni Brea, Flavio Martinelli, Berfin Şimşek, Wulfram Gerstner

January 26, 2023

Abstract

MLPGradientFlow is a software package to solve numerically the gradient flow differential equation $\dot{\theta} = -\nabla\mathcal{L}(\theta; \mathcal{D})$, where θ are the parameters of a multi-layer perceptron, \mathcal{D} is some data set, and $\nabla\mathcal{L}$ is the gradient of a loss function. We show numerically that adaptive first- or higher-order integration methods based on Runge-Kutta schemes have better accuracy and convergence speed than gradient descent with the Adam optimizer. However, we find Newton’s method and approximations like BFGS preferable to find fixed points (local and global minima of \mathcal{L}) efficiently and accurately. For small networks and data sets, gradients are usually computed faster than in pytorch and Hessian are computed at least $5\times$ faster. Additionally, the package features an integrator for a teacher-student setup with bias-free, two-layer networks trained with standard Gaussian input in the limit of infinite data. The code is accessible at <https://github.com/jbrea/MLPGradientFlow.jl>.

We consider multi-layer perceptrons

$$f_{\theta} : \mathbb{R}^{D_{\text{in}}} \rightarrow \mathbb{R}^{D_{\text{out}}} \quad (1)$$

$$x \mapsto f(x) = f_{\theta^L}^L \circ f_{\theta^{L-1}}^{L-1} \circ \dots \circ f_{\theta^1}^1(x) \quad (2)$$

$$f_{\theta^i=(W^i, b^i)}^i : \mathbb{R}^{D_{i-1}} \rightarrow \mathbb{R}^{D_i} \quad (3)$$

$$x \mapsto \begin{cases} \sigma^i(W^i x + b^i) & \text{if } \chi_i = \text{true} \\ \sigma^i(W^i x) & \text{otherwise} \end{cases}, \quad (4)$$

with weights $W^i \in \mathbb{R}^{D_i \times D_{i-1}}$, bias indicator $\chi^i \in \{\text{true}, \text{false}\}$, biases $b^i \in \mathbb{R}^{D_i}$ and element-wise applied activation function $\sigma^i : \mathbb{R} \rightarrow \mathbb{R}$, $D_0 = D_{\text{in}}, D_L = D_{\text{out}}$ and $\theta = (\theta^1, \dots, \theta^L) = (W^1, b^1, \dots, W^L, b^L)$. With data $\mathcal{D} = ((x_1, y_1), (x_2, y_2), \dots, (x_N, y_N))$ and single-sample loss $\ell(f_{\theta}(x_i), y_i) \in \mathbb{R}$, we define the loss

$$\mathcal{L}(\theta; \mathcal{D}) = \sum_i^N \ell(f_{\theta}(x_i), y_i) + R(\theta) \quad (5)$$

with barrier function

$$R(\theta) = \begin{cases} 0 & \text{if } \frac{1}{2}\|\theta\|_2^2 \leq c \\ (\frac{1}{2}\|\theta\|_2^2 - c)^2 & \text{otherwise} \end{cases} \quad (6)$$

finite data constructor	<code>Net(layers = ((D₁, σ¹, χ₁), ..., (D_L, σ^L, χ_L)), input = (x₁, ..., x_N), target = (y₁, ..., y_N))</code>
infinite data constructor	<code>NetI(θ_{student}, θ_{teacher})</code>
<hr/>	
loss and gradient	<code>loss(net, θ), gradient(net, θ)</code>
Hessian and its spectrum	<code>hessian(net, θ), hessian_spectrum(net, θ)</code>
solving Equation 7	<code>train(net, θ)</code>

Table 1: **Most important functions of MLPGradientFlow.**

where $c \in \mathbb{R}_+$ is some constant. The barrier function is included to guarantee fixed points with finite norm of the parameter values, unless c is chosen to be infinite. We define the gradient flow differential equation

$$\dot{\theta}(t) = -\eta \nabla \mathcal{L}(\theta(t); \mathcal{D}), \quad (7)$$

where η is a constant and $\nabla \mathcal{L}$ the gradient of the loss in Euclidean metric. We aim at finding solutions

$$\theta(T) = \int_0^T \dot{\theta}(t) dt \quad (8)$$

of the gradient flow differential equation (Equation 7) for initial conditions $\theta(0) = \theta_0$.

If the target values are generated by a multi-layer perceptron, i.e. $y_i = f_{\theta^*}(x_i)$ with parameters θ^* , we speak of a (noise-free) teacher-student setup. Note that the teacher does not need to have the same number of parameters as the student.

Results

MLPGradientFlow (<https://github.com/jbrea/MLPGradientFlow.jl>) is written in Julia, but it can also be used and was tested with a Python interface. The most important functions are listed in Table 1.

All results where CPU time matters were obtained on a single thread of an AMD Ryzen 7 PRO 5850U using Julia 1.8.5 on Linux. Differential equations are integrated with methods from the OrdinaryDiffEq.jl package [1]. To search fixed points of the differential equation, i.e. (local) minima of the loss, we use optimizers defined in Optim.jl [2] and NLOpt.jl [3]. Gradient descent (**Descent**) and Adam (**Adam**) are copied from Flux.jl [4] and defined directly in MLPGradientFlow.jl.

Adaptive Ordinary Differential Equation Solvers are preferable over Gradient Descent or Adam

To compare different integration methods, we solved the gradient flow differential equation Equation 7 with a reference algorithm at tolerance 10^{-6} and kept the solution as a reference trajectory $(\theta^*(0), \theta^*(t_1), \dots, \theta^*(T^*))$ at 10^6 logarithmically spaced time points ($t_0 = 0, \dots, t_{10^4} = T^*$). The reference algorithm was Rodas5 for the network with 33 parameters and CVODE_BDF for the network with 601 parameters. Next, we ran different

name	needs Hessian	ode solver	works with minibatches
Gradient Descent	no	yes (small step size)	yes
Runge-Kutta 4 (RK4)	no	yes	no
CVODE_BDF	no	yes	no
KenCarp58	yes	yes	no
Rodas5	yes	yes	no
Adam	no	no	yes
(L)-BFGS	no	no	no
SLSQP	no	no	no
Newton(TrustRegion)	yes	no	no

Table 2: **Method properties.** Differential equation solvers are listed on the upper half of the table, fixed point solvers on the lower half. There exist online (L)-BFGS methods (see e.g.[5, 6]) that we did not consider here.

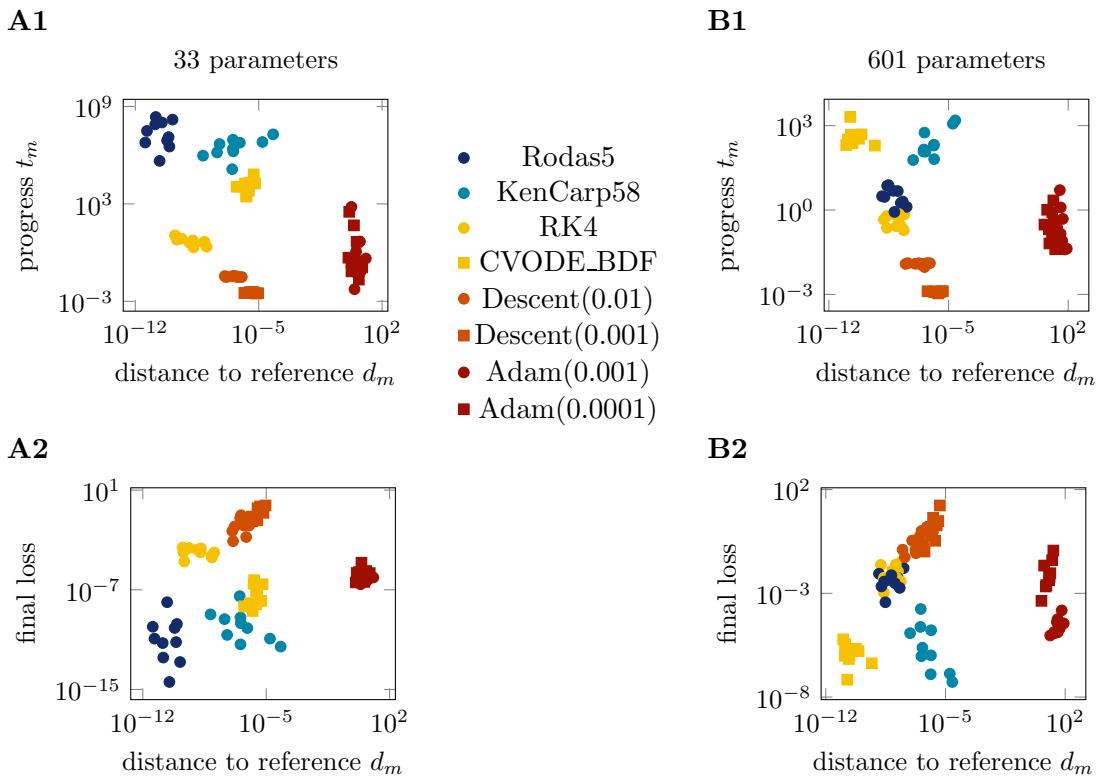


Figure 1: **Efficiency and accuracy of different approximations to the gradient flow.** For given CPU time, higher-order methods Rodas5 and KenCarp58 reach high accuracy (low distance d_m to gradient flow) and integrate the differential equation to a large t_m (A1) and a low loss (A2) for networks with 33 parameters. For larger networks (601 parameters in B1 and B2), the gap between first order and higher order methods decreases because of the long durations for computing Hessians, but the differential equation solver CVODE_BDF still excels. We used 10^4 samples in 2 input dimensions in these simulations.

integration methods m for a fixed duration on a single CPU thread and kept their trajectories $(\theta^m(0), \theta^m(t_1), \dots, \theta^m(T^m))$ at 10^3 logarithmically spaced time points. For each method m , we computed the distance to the reference target trajectory

$d_m = \frac{1}{10^3} \sum_{i=1}^{10^3} \min_j \|\theta^*(t_j) - \theta^m(t_i)\|_2^2$, the progress $t_m = \arg \min_t \|\theta^*(t) - \theta^m(T^m)\|_2^2$ and the final loss at $\theta^m(T^m)$. This was repeated for two different network sizes, and multiple random seeds for the dataset and the initial condition. We find for small networks that higher-order methods Rodas5 and KenCarp58 are usually preferable over CVODE_BDF, Runge Kutta 4 (RK4), Euler integration (Gradient Descent) or Adam and for larger networks CVODE_BDF works best (Figure 1).

Finding Fixed Points

Finding fixed points is surprisingly difficult even for small networks and datasets. If we are only interested in finding low loss values within a given time budget, (approximate) second order optimizers like (L)BFGS, SLSQP or Newton’s method are usually better than differential equation solvers, presumably because they do not spend time following accurately the trajectory of the gradient flow (Figure 2A&B). To find fixed points efficiently, we therefore suggest switching to (approximate) second-order optimization after using differential equation solvers for a given time or iteration limit.

We investigated convergence in a teacher-student setup, where two-layer teachers ($L = 2$) are generated with independently and normally distributed parameters and the students have at least as many neurons in layer 1 as the teachers. In this setting, the students can theoretically reach zero loss. However, because we use double precision (64-bits) for all floating point numbers, the round-off errors for the difference between target values y_i and predictions of the students $f_\theta(x_i)$ are on the order of 10^{-16} . Hence, the mean squared error is on the order of $(10^{-16})^2 = 10^{-32}$ for simulations that converge to a global minimum.

We observe, first, that it is indeed possible to reach loss levels on the order of 10^{-32} with KenCarp58, Newton’s method and SLSQP. BFGS usually stalls at loss levels around 10^{-14} , unless the gradients are scaled with $\eta \gg 1$. However, even with rescaling BFGS usually does not find global minima as accurately as Newton’s method or SLSQP. Second, we find that it can take very long to converge, even with efficient higher-order methods; almost 20% of the simulations reported in Figure 2 did not fully converge within 10^4 steps of KenCarp58 and 30×10^4 steps of Newton’s method. Third, it is difficult to find a reliable convergence criterion. Tiny gradient norms and smallest Hessian eigenvalues near zero do not imply that a fixed point is reached (see “non-converged” point that is probably on the path to a global minimum in Figure 2C2). And even if we call a point “probably converged”, because no decrease of the loss is found in subsequent epochs of 10^4 steps of Newton optimization, simulations may not have fully converged to a fixed point (see “converged” point that is probably on the path to a global minimum in Figure 2C2).

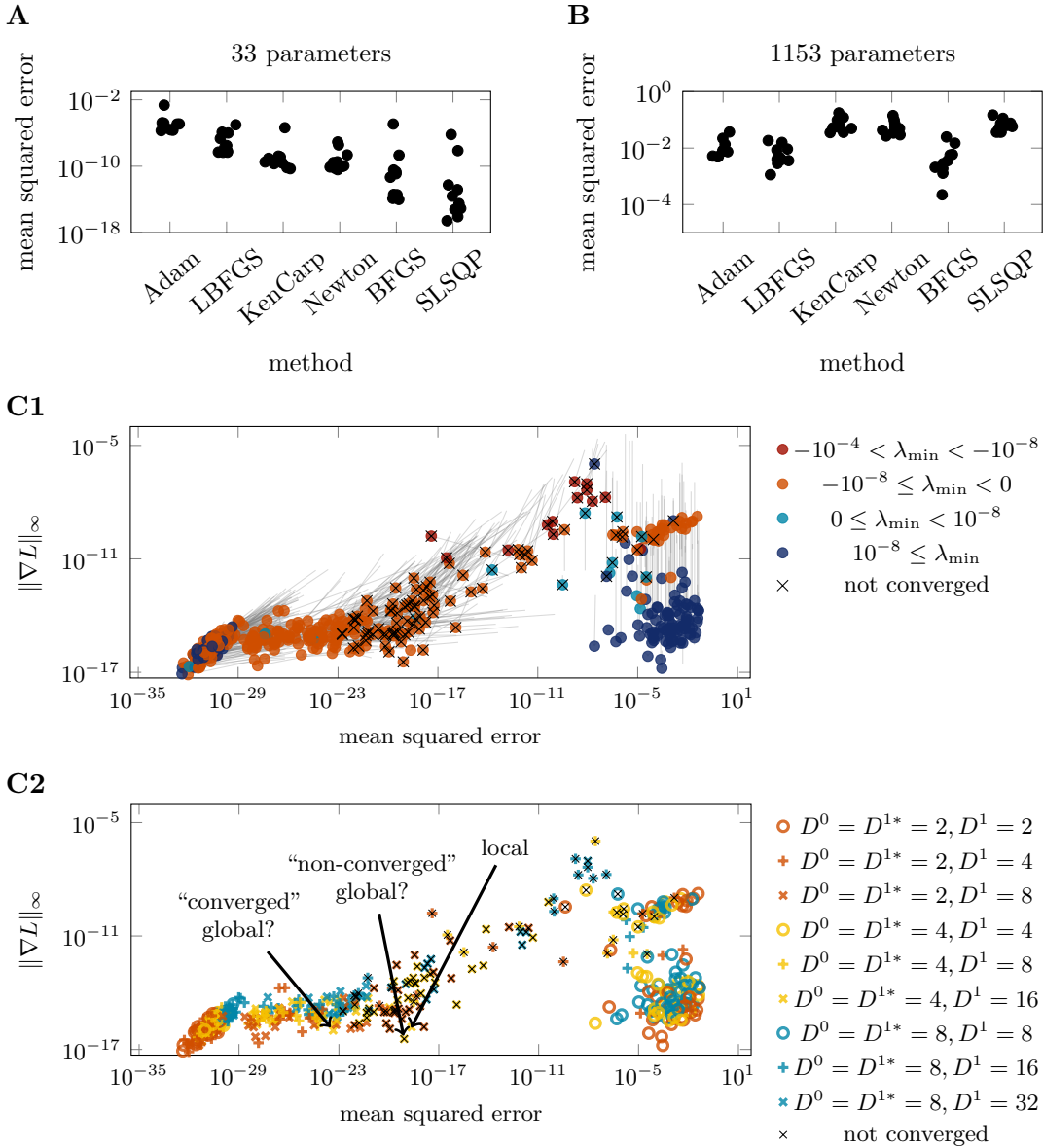


Figure 2: **Finding fixed points efficiently.** For a fixed time budget, the approximate second order method SLSQP finds typically the lowest loss value for small networks (**A**), whereas BFGS is preferable for larger networks (**B**). In **C** we show simulation results, where we know that the smallest loss is at 0. The starts of the gray lines in **C1** show the gradient norm and loss value after 10^4 steps of KenCarp58 and the points show the gradient norm, loss value and minimal eigenvalue of the Hessian (in color code) after at most 30 epochs of 10^4 additional steps of Newton’s method. We call a simulation “probably converged”, if the loss found by Newton’s method stayed the same or increased relative to the previous epoch. If this never happened within 30 epochs, we call the simulation “not converged” (marked with a black cross). In **C2** the show the same data with symbol- and color-code indicating the number of input dimensions D^0 the number of hidden neurons of the teacher D^{1*} and the number of hidden neurons of the student D^1 . Based on inspection of the parameters, we believe that the points marked “global?” will converge to a global minimum, if given more compute time, whereas the nearby point marked with “local” is probably stuck in a local minimum. Therefore, it is unclear, if the points marked as “probably converged” (no black cross) at losses between 10^{-30} and 10^{-16} reached a local minimum or if they did not fully converge to a global minimum, despite our stringent definition of convergence.

Efficiency

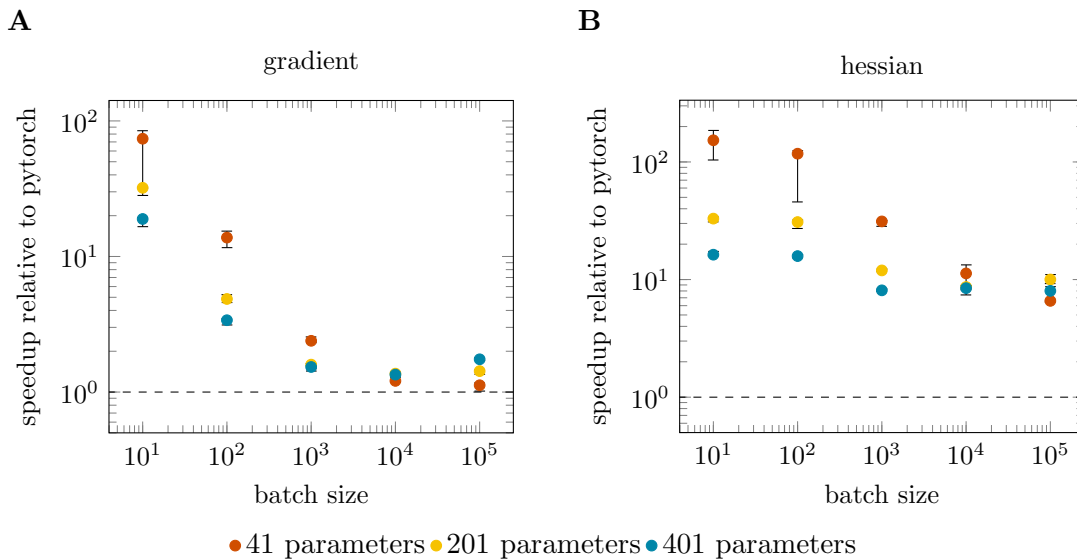


Figure 3: **Speedup relative to pytorch.** The speedup factor (duration MLPGradientFlow/duration pytorch) for different network sizes is measured on a single CPU thread with multiple repetitions of gradient and Hessian computations using the Python module “timeit” and calling MLPGradientFlow from Python.

We use a custom implementation of gradient and Hessian computations, taking inspiration from Bishop (1992) [7] and SimpleChains.jl (<https://github.com/PumasAI/SimpleChains.jl>). Relative to pytorch we observe a significant speedup, in particular for the computation of Hessians (Figure 3).

Finite versus Infinite Data

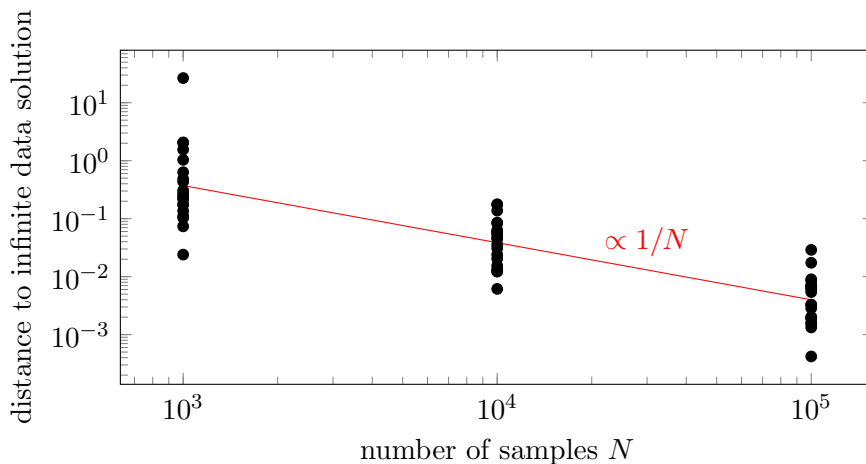


Figure 4: **Infinite, standard Gaussian input data.** The accuracy of the finite-data solution improves with rate $1/N$. These results were obtained with 9 input dimensions.

In a teacher-student setting with bias-free 2-layer networks and normally distributed input data, the loss function can be written in terms of one- to three-dimensional Gaussian integrals, independently of the input dimension [8]. For linear, relu and erf activation functions these integrals can be evaluated analytically, otherwise they can be computed numerically using cubature methods (we use Cuba.jl <https://github.com/giordano/Cuba.jl>). However, accurately computing these integrals in every step of a differential equation solver is time consuming. Therefore we fitted 3-layer perceptrons with 96 neurons in each hidden layer to datasets with approximately 10^5 samples, such that they approximate the numerical integrals with high accuracy (typically, loss and gradients are within 10^{-5} Euclidean distance from the numerical integrals on test data, and Hessians are within 10^{-3} distance). We recommend using these approximations for the differential equation solvers and switching to numerical integration, if fine-tuning of the fixed points is desired. In contrast to existing differential equation solvers for this setting (e.g. [8]), this package supports more activation functions and provides also Hessians, which enables the usage of higher-order solvers.

As an alternative to numerical integration or approximations thereof, one can sample datasets with normally distributed input. But the accuracy of the finite-data solution improves only with a rate inversely proportional to the size of the dataset (Figure 4), whereas the computational cost increases linearly with the size of the dataset. In particular in high input dimensions, the numerical integrals and its approximations are preferable over approximations with finite data sets.

Conclusions

We described the software package MLPGradientFlow.jl, a tool to integrate the gradient flow differential equation of the loss function of neural networks. We illustrated the advantages and drawbacks of different methods with toy examples. The package is mainly targeted at small networks with a few thousand parameters, at most. For larger networks, the time and space requirements of computing and inverting Hessians becomes impractical and Hessian-free methods should be considered (e.g. [9] or integrators based on Krylov methods [1]). The largest setting we considered so far is a network with 79'510 parameters and 70'000 samples in 784 input dimensions, which we integrated with CVODE_BDF (see “scripts/mnist.jl” in <https://github.com/jbrea/MLPGradientFlow.jl>).

References

- [1] Christopher Rackauckas and Qing Nie. Differentialequations.jl – a performant and feature-rich ecosystem for solving differential equations in julia. *The Journal of Open Research Software*, 5(1), 2017. Exported from <https://app.dimensions.ai> on 2019/05/05.
- [2] Patrick Kofod Mogensen and Asbjørn Nilsen Riseth. Optim: A mathematical optimization package for Julia. *Journal of Open Source Software*, 3(24):615, 2018.
- [3] Steven G. Johnson. The nlopt nonlinear-optimization package.

- [4] Michael Innes, Elliot Saba, Keno Fischer, Dhairya Gandhi, Marco Concetto Rudilosso, Neethu Mariya Joy, Tejan Karmali, Avik Pal, and Viral Shah. Fashionable modelling with flux. *CoRR*, abs/1811.01457, 2018.
- [5] Philipp Moritz, Robert Nishihara, and Michael Jordan. A linearly-convergent stochastic l-bfgs algorithm. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 249–258, Cadiz, Spain, 09–11 May 2016. PMLR.
- [6] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, Jan 2018.
- [7] Chris Bishop. Exact calculation of the hessian matrix for the multilayer perceptron. *Neural Computation*, 4(4):494–501, Jul 1992.
- [8] 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.
- [9] Barak A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Computation*, 6(1):147–160, Jan 1994.