AN ANALYTICAL THEORY OF AUXILIARY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Auxiliary learning is an optimization paradigm in which a neural network's performance on a target task is improved by jointly training it on additional tasks. However, the mechanisms behind this improvement remain poorly understood. We study this problem using a teacher-student framework and derive a closed system of differential equations describing the dynamics of online stochastic gradient descent in the large-input limit. For linear networks, we obtain a closed-form expression for the generalization error to leading order in the learning rate, quantifying how task correlations and label noise determine the benefit of auxiliary learning. For non-linear activation functions, we develop a fluctuation-dissipation analytical theory that establishes a general relation linking the main and auxiliary errors to the corresponding single-task error. Numerical experiments support the theoretical predictions and show how auxiliary tasks improve generalization by balancing the forcing dynamics towards the optimal solution with gradient noise.

1 INTRODUCTION

Modern neural networks are rarely trained to optimize a single task in isolation. In practice, practitioners have introduced different multi-task paradigms to improve performance on the tasks these models are designed for (Caruana, 1997; Ruder, 2017; Liebel & Körner, 2018; Crawshaw, 2020; Zhang & Yang, 2022). These paradigms use additional objectives, called *auxiliary* tasks, on which the model is trained either in parallel with its main objective (auxiliary learning) or by alternating tasks (knowledge transfer and continual learning). The empirical effectiveness of these methods clashes with the intuition that multiple competing objectives might worsen the overall performance.

Modern neural networks are prone to overfitting, but do not have a capacity problem, since they can approximate any function to any precision. Baxter (2000) suggested that multi-task learning can work as an *inductive bias* in the space of possible functions. Thus, multi-task learning might lead the network to learn a better-generalizing representation of the data, but our understanding of this phenomenon is still limited.

To begin developing a theory of auxiliary learning, we focus on a specific tractable setting. We study a variant of the classic teacher-student setup in which one or more auxiliary tasks, available only during training, can improve performance on a target task. In this framework, the target function is exactly learnable, and we analyze how auxiliary tasks affect generalization through training dynamics, while leaving the target-task optimal solution unchanged. To do so, we expand on the framework proposed for the single-output teacher-student setup by Saad & Solla (1995a;b); Biehl & Schwarze (1995), who showed that a set of ordinary differential equations can describe the dynamics of large-width networks trained with online learning. We generalize these mean-field ODEs to our auxiliary problem and, drawing upon the literature on adaptive filters, we derive a closed-form leading-order approximation of the generalization error for a multi-layer linear model.

We then focus on other nonlinear activations and find that, at convergence, the dynamics are described, to first order, by a Lyapunov equation. By solving it numerically, we can recover the generalization error, as well as the optimal weighting of the auxiliary task that minimizes it, for any choice of activation and hyperparameters. We find that the generalization error between the main and auxiliary task is described by a simple relationship involving the generalization error of the corresponding one-output teacher-student network, which our theory predicts for any choice of activation function; a novel result, since the existing literature recovers it only for some specific activations (Goldt et al., 2019).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.1 PREVIOUS LITERATURE

Auxiliary learning. Auxiliary learning uses parallel tasks during training as an inductive bias to improve either feature learning, generalization, or optimization. Caruana (1997) introduced the idea that related tasks improve generalization via learning shared representations. Baxter (2000) later expanded on this relationship. Its effectiveness has been extensively discussed; Caruana (1997); Ruder (2017); Liebel & Körner (2018); Crawshaw (2020); Zhang & Yang (2022) provide relevant literature reviews

Adaptive filters. Adaptive filters are linear functions, explicitly designed as real-time, sample-by-sample algorithms for processing streaming signals, where parameters are updated after each input signal is processed. In the adaptive filters literature, the steady-state error of the least mean squares algorithm (which is exactly online stochastic gradient descent, SGD, on linear regression with quadratic loss) has been studied extensively by Widrow & Walach (1983); Haykin (1986); Sayed (2008).

Teacher-student. A series of works on the teacher-student setup (Saad & Solla, 1995a;b; Biehl & Schwarze, 1995; Vicente & Caticha, 1997; Saad, 1999; Inoue et al., 2003) showed that in the large-input limit, when the size of the input data is much larger than the number of hidden neurons ($N \to \infty$), the dynamics of the student network trained with online stochastic gradient descent (SGD) can be modeled by a set of mean-field ordinary differential equations (ODEs) for some order parameters, whose number scale intensively as K^2, where K is the number of neurons in the hidden layer, and do not depend on the (extensive) size of the input N. This statistical mechanics framework, one of the first used to study the properties of multi-layer neural networks, was later formalized by Goldt et al. (2019) and adopted to understand specific settings related to multi-task learning; Straat & Biehl (2019) expanded it to ReLU activations, Asanuma et al. (2021); Lee et al. (2021); Saglietti et al. (2022) to continual learning, and Mori & Mignacco (2025) to dropout regularization. Citton et al. (2024; 2025b;a) have used Hermite expansions to simulate the dynamics of arbitrary activation functions.

Fluctuation-dissipation. Mandt et al. (2017) expanded the model loss around a local minimum, so that SGD reduces to a multivariate Ornstein-Uhlenbeck process whose Gaussian stationary distribution has a covariance matrix fixed by a Lyapunov equation involving the Hessian at the optimum. Several works have expanded this first result; for example, Jastrzębski et al. (2018) solved the stationary covariance in closed form under additional assumptions. Yaida (2019) gives a more general discussion of fluctuation-dissipation relations without the quadratic approximation.

2 AUXILIARY LEARNING IN THE TEACHER-STUDENT FRAMEWORK

2.1 PROBLEM SETUP

In the teacher-student setup, a *student* network is trained on inputs $\boldsymbol{x} \in \mathbb{R}^N$ drawn from the normal distribution $\mathcal{N}(0, I_N)$ with labels generated by a *teacher* network. We consider a variant of the classic teacher-student scenario in which a task (here called the main task) must be learned, with the help of an auxiliary output. The setting is thus that of a multi-output network with K hidden neurons, parametrized by weights $W \in \mathbb{R}^{K\times N}$, $\boldsymbol{v} \in \mathbb{R}^K$, $\boldsymbol{u} \in \mathbb{R}^K$ and activation function $g : \mathbb{R} \to \mathbb{R}$. The network outputs are

$$
\begin{aligned}
y_{\text{main}}(\boldsymbol{x} \,|\, \theta) &= \sum_{j=1}^{K} v_j \, g\left(h_j\right) = \sum_{j=1}^{K} v_j \, g\left(\frac{\boldsymbol{w}_j \cdot \boldsymbol{x}}{\sqrt{N}}\right), \\
y_{\text{aux}}(\boldsymbol{x} \,|\, \theta) &= \sum_{j=1}^{K} u_j \, g\left(h_j\right) = \sum_{j=1}^{K} u_j \, g\left(\frac{\boldsymbol{w}_j \cdot \boldsymbol{x}}{\sqrt{N}}\right),
\end{aligned}
\tag{1}
$$

where $\boldsymbol{w}_j$ is the j-th row of W. We call θ the set of weights $\{W, \boldsymbol{v}, \boldsymbol{u}\}$. The teacher and student networks have the same number of neurons. The main target is y_{main}, and y_{aux} is the auxiliary target.

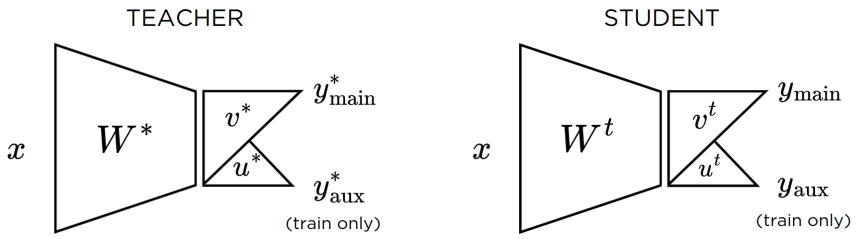


Figure 1: **Schema of the multi-index teacher-student setup.**

Given a set of teacher weights $\theta^* = \{W^*, \boldsymbol{v}^*, \boldsymbol{u}^*\}$, the labels to be learned are $y^*(\boldsymbol{x}) = y(\boldsymbol{x} \,|\, \theta^*)$, and we are interested in the generalization error on the target task

$$\epsilon_{\text{main}}(\theta, \theta^*) = \mathop{\mathbb{E}}_{\boldsymbol{x} \in \mathcal{N}} \left[\left(y^*_{\text{main}}(\boldsymbol{x}) - y_{\text{main}}(\boldsymbol{x}) \right)^2 \right]. \tag{2}$$

We initialize the student model weights from a standard Gaussian (but they can be initialized differently without loss of generality). The student is trained with online SGD: namely, at each epoch, we sample a point from the standard normal $\boldsymbol{x} \sim \mathcal{N}(0, I_N)$, and the weights are updated by

$$\begin{aligned} W^{t+1} &= W^t - \mu \nabla_W \mathcal{L}(\boldsymbol{x} \,|\, \theta^t, \theta^*), \\ v^{t+1} &= v^t - \frac{\mu}{N} \nabla_v \mathcal{L}(\boldsymbol{x} \,|\, \theta^t, \theta^*), \\ u^{t+1} &= u^t - \frac{\mu}{N} \nabla_u \mathcal{L}(\boldsymbol{x} \,|\, \theta^t, \theta^*), \end{aligned} \tag{3}$$

with learning rate $\mu > 0$ and training loss

$$\begin{aligned} &\boxed{\mathcal{L}(\boldsymbol{x}) = \alpha \mathcal{L}_{\text{main}}(\boldsymbol{x}) + \beta \mathcal{L}_{\text{aux}}(\boldsymbol{x}),} \\ &\mathcal{L}_i(\boldsymbol{x}) = \frac{1}{2} \left(y^*_i(\boldsymbol{x}) - y_i(\boldsymbol{x}) + \sigma z_i \right)^2, \quad i \in \{\text{main}, \text{aux}\} \end{aligned} \tag{4}$$

with $\sigma \in \mathbb{R}$, and $z_i \sim \mathcal{N}(0, 1)$ representing Gaussian noise added to the labels during training. Notice how the auxiliary loss does not change the minimizer, since the two tasks share the same teacher. As we will see, it plays a crucial dynamical role in trading effective label noise against effective learning rate.

A central question is how much to weight the auxiliary loss term to minimize the generalization error on the target task. In other words, we want to find the optimal values for the hyperparameters

$$(\alpha, \beta) = \mathop{\text{argmin}}_{\alpha, \beta \in S} \epsilon_{\text{main}}, \tag{5}$$

where S is some set of conditions on (α, β). We focus mainly on finding the optimal β with $\alpha = 1$, without rescaling the main generalization loss. We also explore the case $\alpha = 1 - \beta$ in Appendix E, which produces different results because of the dynamical effects governing our online learning setup.

2.2 The large-input limit

We train the student network on $i.i.d.$ Gaussian samples drawn from a normal distribution. Thus, both the learning dynamics and the generalization error depend on the student and teacher weights only through the statistics of the network pre-activations (local fields) h_i and h^*_n. Since each

pre-activation is a projection of the Gaussian sample onto the corresponding weight $\boldsymbol{w}$, the pre-activations are jointly Gaussian with zero mean, and their covariance is fully determined by

$$\begin{aligned} Q_{ik} &= \mathbb{E}\left[h_i h_k\right] = \frac{\boldsymbol{w}_i^T \boldsymbol{w}_k}{N}, \\ R_{in} &= \mathbb{E}\left[h_i h_n^*\right] = \frac{\boldsymbol{w}_i^T \boldsymbol{w}_n^*}{N}, \\ T_{mn} &= \mathbb{E}\left[h_m^* h_n^*\right] = \frac{\boldsymbol{w}_m^{*T} \boldsymbol{w}_n^*}{N}. \end{aligned} \tag{6}$$

The entries of these matrices, the student-student overlaps Q_{ik}, the student-teacher overlaps R_{in}, and the teacher-teacher overlaps T_{mn}, are the (slightly improperly named) *order parameters* of our problem, and are the only quantities through which the weights enter both the generalization error and the dynamics.

One can then compute the generalization error analytically (Saad & Solla, 1995a;b; Biehl & Schwarze, 1995; Goldt et al., 2019), obtaining

$$\epsilon_{\text{main}} = \sum_{i,j} v_i v_j I_2(i,j) + \sum_{n,m} v_n^* v_m^* I_2(n,m) - 2 \sum_{i,n} v_i v_n^* I_2(i,n), \tag{7}$$

where the notation $I_2(i,j)$ indicates $\mathbb{E}[g(h_i)\, g(h_j)]$, with i, j being indices of units in the student network and n, m teacher ones. Because I_2 only depends on the order parameters, it reduces to computing a two-dimensional Gaussian average of a nonlinear function, which, depending on the activation, can be solved analytically or approximated numerically.

The order parameters, together with the second-layer weights $\boldsymbol{v}$ and $\boldsymbol{u}$, can be evolved through the exact population dynamics of SGD (namely, in the $N \to \infty$ limit) according to a closed set of ordinary differential equations that depend on I_2 and other integrals of the activation function. This description becomes deterministic and asymptotically exact in the large-N limit, since the order parameters are self-averaging (Goldt et al., 2019). We derive the full set of equations for the two-output case in Appendix A.

3 Results

3.1 Linear model

Before studying the general case of networks with nonlinear activation functions, we develop a theory for the simpler deep linear model, defined by the choice $g(x) = x$, with fixed second-layer weights $\boldsymbol{v} = \boldsymbol{v}^*$, and $\boldsymbol{u} = \boldsymbol{u}^*$, leaving W as the only trainable parameters. Despite its simplicity, this model still provides insight into the dynamics of one-hidden-layer models. It is well understood that a linear model trained with online SGD on an MSE loss converges to a generalization error governed by the dynamical effects of training (Widrow & Walach, 1983; Haykin, 1986; Sayed, 2008). These linear models were introduced in the machine learning literature as *adaptive filters*. For normal Gaussian data, the loss converges to $\epsilon_{a.f.} \approx \mu\, \sigma^2/2$ when we keep only the leading term in μ. When either μ or σ goes to zero, we recover the original function. The dependence on μ explicitly highlights the dynamical nature of this result.

In Appendix B, we show that, for $\alpha > 0$, $\beta \geq 0$, a small learning rate μ, and with the same norms of the second layer weights (namely, $|\boldsymbol{u}|^2 = |\boldsymbol{v}|^2$), the two-layer linear model converges to a generalization loss with a leading term in μ

$$\begin{aligned} \epsilon_{\text{main}} &= \frac{\mu\sigma^2}{2} |\boldsymbol{v}|^2 \frac{\alpha^2 + \alpha\beta(1-\rho^2) + \rho^2\beta^2}{\alpha+\beta} = \\ &= \underbrace{\epsilon_{a.f.} |\boldsymbol{v}|^2}_{\epsilon_0} \frac{\alpha^2 + \alpha\beta(1-\rho^2) + \rho^2\beta^2}{\alpha+\beta}, \end{aligned} \tag{8}$$

where $\rho = \boldsymbol{v} \cdot \boldsymbol{u}/|\boldsymbol{v}||\boldsymbol{u}|$ is the correlation between the two output weights and $\epsilon_{a.f.}$ is the generalization error of the corresponding linear model trained without an auxiliary task (also truncated at the first order in μ). In Figure 2A, we plot the evolution of generalization error ϵ_{main} obtained both by training models with SGD and by numerically solving the ODEs, showing how their dynamics match and converge to the loss predicted by the theory for small learning rates μ. For SGD, the generalization error fluctuates around the steady state, so we show an average over 10000 epochs after convergence.

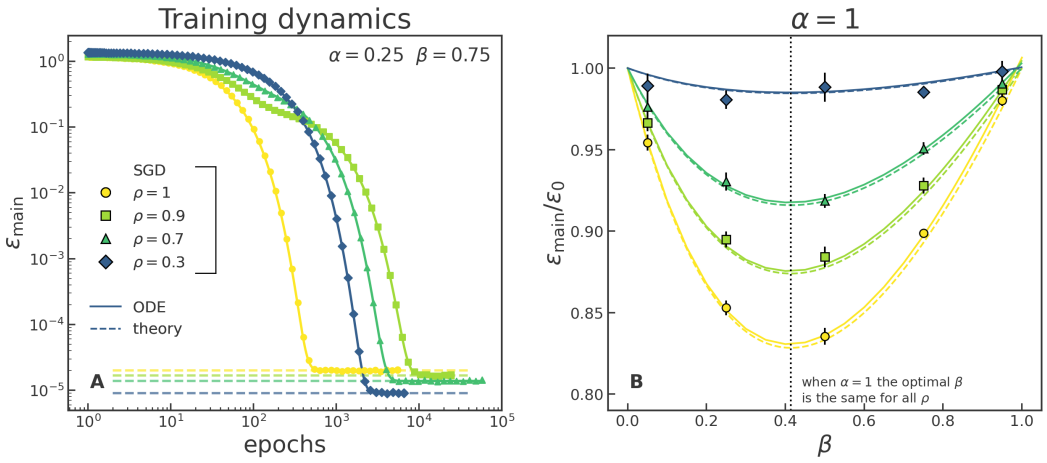


Figure 2: **Performance of linear models.** A) Sample learning trajectory obtained with ODEs (continuous line) and averages over ten runs of SGD (dots) when $\alpha = 0.25$ and $\beta = 0.75$. The dashed line is the theoretical loss at convergence, and different colors/shapes correspond to different correlations ρ. B) Relative error $\epsilon_{\text{main}}/\epsilon_0$ for a linear model with four different ρ trained with online gradient descent, $\alpha = 1$, and different values of β. The dashed lines show the theoretical solutions, the continuous lines show the results of numerically solving the mean-field ODEs, the dots show simulation results for models trained with online SGD, averaged over 10000 epochs after convergence across 10 simulations, with the vertical bars denoting the corresponding confidence intervals. Parameters for all simulations are $\mu = 0.01$, $\sigma = 0.1$, $K = 4$, $N = 784$.

When setting $\alpha = 1$, we get

$$\epsilon_{\text{main}}\Big|_{\alpha=1} = \epsilon_0 \left(1 - \rho^2 \frac{\beta(1-\beta)}{1+\beta}\right), \tag{9}$$

which is minimized by

$$\beta^* = \sqrt{2} - 1. \tag{10}$$

Interestingly, the optimal value does not depend on ρ, as shown in Figure 2B. When $\rho = 0$, the generalization error is $\epsilon_{\text{main}} = \epsilon_{a.f.}|\boldsymbol{v}|^2$ and does not depend on the choice of β: when the main and auxiliary tasks are completely decoupled, rescaling β does not rescale the main task learning rate and does not affect its generalization error. When the two tasks coincide (namely, when $\rho = 1$), the optimal value β^* is not $1/2$ as one naively expects. As shown in Appendix C, the case of $\rho = 1$ corresponds to an effective learning rate $\mu_{\text{eff}} = \mu(\alpha + \beta)$ and an effective noise $\sigma^2_{\text{eff}} = \sigma^2(\alpha^2 + \beta^2)/(\alpha + \beta)^2$. Thus, rescaling the auxiliary task also increases the effective learning rate, creating a trade-off between the learning rate and the noise variance.

A different phenomenology can be obtained with the constraint $\alpha = 1 - \beta$, when $\mu_{\text{eff}} = \mu$, without a dependence on β. We discuss this setting in Appendix E, and show how this additional constraint modifies the results.

Quite unexpectedly, from Equation 8 it follows that

$$\alpha\epsilon_{\text{main}} + \beta\epsilon_{\text{aux}} = \epsilon_{a.f.}|\boldsymbol{v}|^2(\alpha^2 + \beta^2), \tag{11}$$

where the quantities on the left depend on ρ, while those on the right do not; we return to this relationship when discussing nonlinear models.

In Appendix B, we extend these results to a teacher with different noise variance on the auxiliary task, showing it is beneficial to weight the auxiliary task more when its noise variance is smaller than the main task one, as well as a teacher with multiple auxiliary tasks, showing that they improve generalization compared to only having one.

3.2 Nonlinear activation functions

Goldt et al. (2019) previously calculated the generalization error for Soft Committee Machines (SCMs), networks for which $v_i = v_i^* = 1\ \forall i$, showing it to be $\epsilon_{\text{erf}} = \mu\sigma^2 K/(\pi\sqrt{3}) + \mathcal{O}(\mu^2)$ when the activation function is $g(x) = \text{erf}(x/\sqrt{2})$. This function is convenient because it is one of the few for which all ODE integrals are analytically computable in a simple form. In general, they also observe that $\epsilon_0 \propto \mu\sigma^2$ when $\mu \to 0$ with ReLU activation, although they did not derive the proportionality constant analytically.

We now extend these results to networks trained with an auxiliary output. We fix $\boldsymbol{v} = \boldsymbol{v}^*$ and $\boldsymbol{u} = \boldsymbol{u}^*$, with correlation ρ, and same norm $|\boldsymbol{u}| = |\boldsymbol{v}|$, and different nonlinear activations, namely erf and ReLU. The behavior is more complicated and does not depend only on ρ. Figure 3A shows simulation results for erf, and Figure 3B for ReLU.

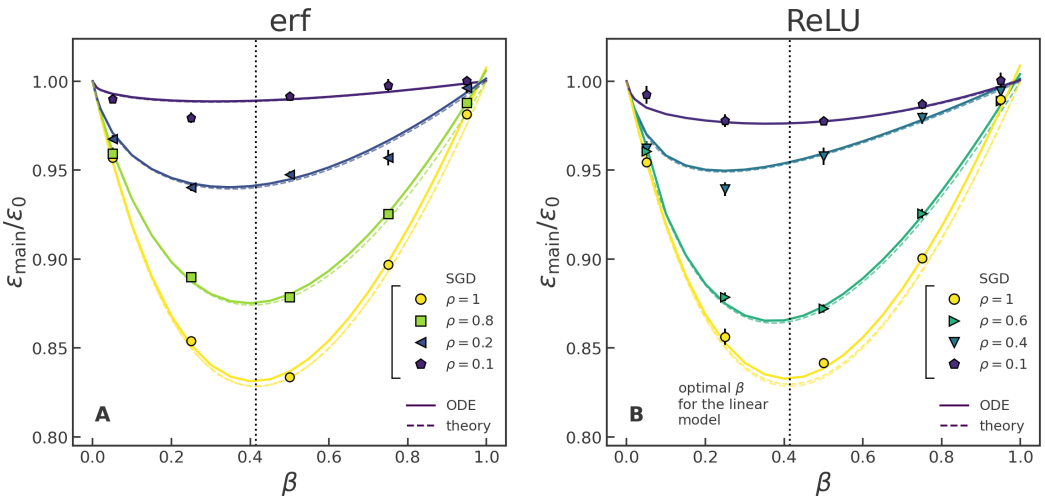


Figure 3: **Performance of models with different activations.** Relative error $\epsilon_{\text{main}}/\epsilon_0$ for a model with A) erf activation ($K = 4$) and B) ReLU activation ($K = 2$), with different ρ trained with online gradient descent, $\alpha = 1$, and different βs. The dashed lines show the theoretical solutions, the continuous lines show the results of numerically solving the mean-field ODEs, the dots show simulation results for models trained with online SGD, averaged over 10000 epochs after convergence across 5 simulations for erf and 10 for ReLU, with the vertical bars denoting the corresponding confidence intervals. Results for $\mu = 0.01$, $\sigma = 0.1$, $N = 784$.

A note on initialization. Throughout the rest of this paper, we are interested in studying the behavior of the well-converging initializations W_0 that reach a neighbourhood of the realizable optimum W^* without getting stuck in non-optimal minima or saddle points. The theoretical construction of the mean-field ODEs does not depend on the choice of initialization for W, v, and u, justifying the choice of specific initializations. We empirically observe that bad initializations converge to a non-diagonal R matrix (possibly transiently). We also see that some of these bad initializations do not converge for $\beta = 0$, namely the single-output case, but do reach a neighbourhood of the optimum in the presence of auxiliary tasks, when $\beta \neq 0$. This result suggests that auxiliary learning can also

help optimization, beyond the convergence results which are the focus of our theory. Appendix D.6 presents a preliminary computational analysis of this effect, but a complete study of how auxiliary learning might help escape bad minima in the teacher-student is beyond the scope of this work.

We can make a few empirical observations from Figure 3.

Observation 1. *The generalization error is $\epsilon_{main}|_{\alpha=\beta=1} = \epsilon_0$, exactly as in the linear case.*

Observation 2. *The relation $\alpha\epsilon_{main} + \beta\epsilon_{aux} \approx \epsilon_0(\alpha^2 + \beta^2)$ still approximately holds. While both ϵ_{main} and ϵ_{aux} depend on the specific choice of $\boldsymbol{v}$, $\boldsymbol{u}$ and their correlation ρ, the quantity on the right side of the equation does not.*

Observation 3. *Empirically, the optimal value of β is $\beta^* < 1/2$ when $\alpha = 1$, and typically $\beta^* \leq \sqrt{2} - 1$ for most sampled configurations.*

3.3 A GENERAL ANALYTIC THEORY

SGD dynamics, after converging, fluctuate around the steady state given by the realizable optimum W^*, namely the teacher weights. We now discuss how the statistics of this stationary distribution around the minimum allow us to compute the generalization error and explain the observations in the previous section. In Appendix D.1, we show how the distribution induced by online SGD is described at first order in μ by the Lyapunov equation

$$HC + CH = \mu\,\Sigma_{\text{grad}}, \tag{12}$$

with

$$\begin{aligned} C = \mathbb{E}\big[\delta\boldsymbol{W}\delta\boldsymbol{W}^T\big], \quad H = \mathbb{E}\big[\nabla^2_{\boldsymbol{W}}\mathcal{L}\big]\big|_{\boldsymbol{W}^*}, \\ \Sigma_{\text{grad}} = \mathbb{E}\big[\nabla_{\boldsymbol{W}}\mathcal{L}\,\nabla_{\boldsymbol{W}}\mathcal{L}^T\big]\big|_{\boldsymbol{W}^*}, \end{aligned} \tag{13}$$

where the weight fluctuations are $\delta\boldsymbol{W} := \boldsymbol{W} - \boldsymbol{W}^*$, and we adopted the notation $\boldsymbol{W} = \text{vec}(W)$. Also, $\mathbb{E}$ is the population average over new samples once the dynamics converge around the minimum. Notice that H and Σ_{grad} are fixed by the choice of network architecture and teacher weights.

One can numerically solve Equation 12 for C and recover the theoretical generalization error, but C is still a $KN \times KN$ matrix. In Appendix D.2, we show that explicitly computing H and Σ_{grad}, due to the tensor product structure, reduces the Lyapunov equation to a $K \times K$ equation

$$(\alpha P_{\text{main}} + \beta P_{\text{aux}})\tilde{C} + \tilde{C}(\alpha P_{\text{main}} + \beta P_{\text{aux}}) = \mu\sigma^2(\alpha^2 P_{\text{main}} + \beta^2 P_{\text{aux}}). \tag{14}$$

with

$$\begin{aligned} (P_{\text{main}})_{ij} = v_i v_j\,\mathbb{E}\big[g'(h_i)\,g'(h_j)\big]\Big|_{W^*} = v_i v_j\,J_2(ij)\big|_{W^*}, \\ (P_{\text{aux}})_{ij} = u_i u_j\,\mathbb{E}\big[g'(h_i)\,g'(h_j)\big]\Big|_{W^*} = u_i u_j\,J_2(ij)\big|_{W^*}, \end{aligned} \tag{15}$$

This new Lyapunov equation can now be solved numerically for $\tilde{C} \in \mathbb{R}^{K\times K}$. This allows for a very efficient way to compute the theoretical generalization error for a given choice of $\boldsymbol{v}$, $\boldsymbol{u}$, and hyperparameters:

$$\epsilon_i = \text{tr}(P_i\tilde{C}). \tag{16}$$

As seen from Figure 3, the theory (dotted line) matches the simulations and ODEs very well. In Appendix D.5, we discuss the regimes of μ for which this approximation holds. Let us now return to our empirical observations.

Explaining Observation 1. When $\alpha = \beta$, a diagonal $\tilde{C}$ solve our Lyapunov equation exactly. It immediately follows that

$$\alpha = \beta \implies \epsilon_{\text{main}} = \epsilon_{\text{aux}} \tag{17}$$

for any activation and any correlation ρ. Substituting this result into Observation 2, we get

$$\epsilon_{\text{main}}|_{\alpha=\beta} = \alpha\epsilon_0. \tag{18}$$

Explaining Observation 2. Taking the trace of Equation 14, we get

$$\alpha\epsilon_{\text{main}} + \beta\epsilon_{\text{aux}} = \frac{\mu\sigma^2}{2}\left(G_{\text{main}}\alpha^2 + G_{\text{aux}}\beta^2\right) \tag{19}$$

where $G_{\text{main}} = \operatorname{tr} P_{\text{main}} = \sum_i v_i^2 \, J_2(i,i)|_{W^*}$, and similarly for G_{aux}.

At the fixed point W^*, $Q_{ii} \to T_{ii} = 1$ (the student aligns with the teacher), so one can compute directly $J_2(i,i)|_{W^*} = \mathbb{E}[g'(h)^2] = J_2^*$. Then it immediately follows, since the readouts have the same norm, that $G_{\text{main}} = G_{\text{aux}} = G = J_2^*|v|^2$, and

$$\alpha\epsilon_{\text{main}} + \beta\epsilon_{\text{aux}} = \frac{\mu\sigma^2 G}{2}\left(\alpha^2 + \beta^2\right) = \epsilon_0\left(\alpha^2 + \beta^2\right), \tag{20}$$

with the auxiliary-free floor ϵ_0 obtained by setting $\alpha = 1, \beta = 0$. This result holds to leading order in μ. In Appendix D.5, we characterize the impact of higher-order corrections, showing agreement for $\mu \leq 10^{-2}$ and increasing deviations for $\mu \to 1$.

We can compute G analytically from J_2, and we get

$$G^{\text{linear}} = |\boldsymbol{v}|^2, \qquad G^{\text{erf}} = \frac{2}{\pi\sqrt{3}}|\boldsymbol{v}|^2, \qquad G^{\text{ReLU}} = \frac{1}{2}|\boldsymbol{v}|^2. \tag{21}$$

From this, we can recover the generalization error of the online single-output teacher-student for any activation function, namely, from

$$\epsilon_0 = \frac{\mu\sigma^2 G}{2}, \tag{22}$$

which aligns with our previous results for linear models, and generalizes previously known results from Goldt et al. (2019), which proposed an approach that requires the computation of all integrals in Equation 30 of Appendix A to obtain the leading term in σ^2 of the generalization error. Our approach requires only J_2, which is much easier to compute, but is limited to the leading term in μ.

Explaining Observation 3. In Figure 4, we randomly sample $\boldsymbol{v}$ and $\boldsymbol{u}$ with the same norm and a given correlation ρ, and we numerically solve the corresponding Lyapunov equation for different βs. From this, we compute the optimal β^* and the relative loss $\epsilon_{\text{main}}^*/\epsilon_0 \equiv \epsilon_{\text{main}}(\beta^*)/\epsilon_0$, and we observe that all of the β^*s lie below $\sqrt{2} - 1$ for models with erf activation, but not for ReLU.

In Appendix D.4, we discuss the case where P_{main} and P_{aux} commute, and we show that, under this strong assumption, which does not hold generally, $\beta^* < 1/2$ when $\alpha = 1$.

4 Discussion and Future Work

We showed that auxiliary learning improves performance for the teacher-student setup in the large-input and small-learning-rate limit. We can write the linear model's generalization error analytically, and it exhibits an interesting phenomenology. Namely, the optimal β^* weighting the auxiliary task is always $\sqrt{2} - 1$ regardless of the correlation between tasks, and the generalization errors of the

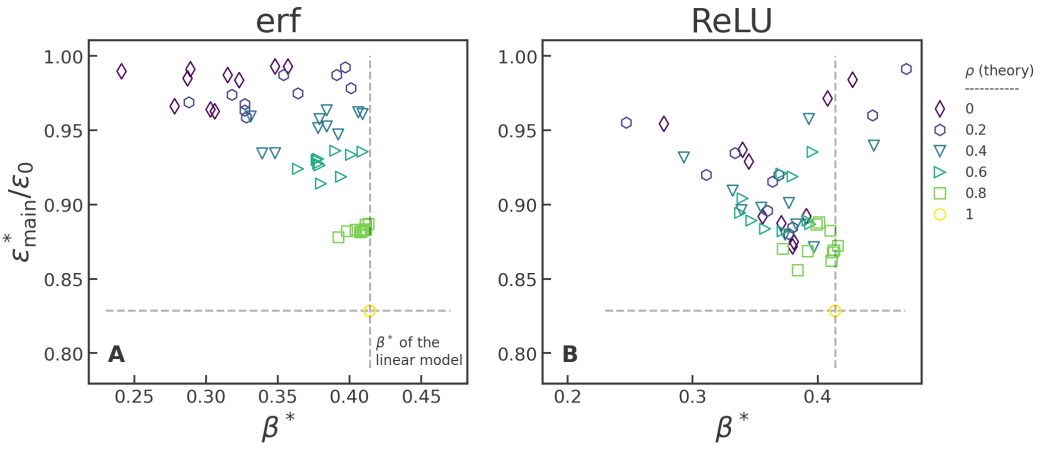


Figure 4: **Best improvement (theory).** Best relative loss against optimal β^* computed for the model with A) erf and B) ReLU activation, $\alpha = 1$. We uniformly sample ten $\boldsymbol{v}$, $\boldsymbol{u}$ with norm $|\boldsymbol{v}| = |\boldsymbol{u}| = \sqrt{K}$ and given correlation ρ. Results obtained by numerically solving the corresponding Lyapunov equation for $\mu = 0.01$, $\sigma = 0.1$, $K = 4$.

main and auxiliary tasks are tied to the generalization of the single-output model by an elegant relationship (Equation 11).

We then made similar empirical observations on models with nonlinear activations. The optimal β^* is not always the same as the linear model; the advantage curves remain similar, and the loss relationship stays empirically valid. This led us to develop an analytic theory by expanding online SGD around the teacher optimum W^*. We obtain a Lyapunov equation that we can solve numerically to predict the generalization error for any choice of second-layer weights, activation function, and hyperparameters. This theory also explains our empirical observations and could be trivially expanded to more complicated setups, such as multiple auxiliary outputs or tasks with different noise scales, even though we explored these setups only for the linear model. One of the most relevant consequences is that we can now predict the first-order term in μ of the generalization error for a one-output teacher-student with a fixed second layer and any nonlinear activation, a novel result in the teacher-student literature.

A future direction left unexplored is student models with a different number of neurons from the teacher. Goldt et al. (2019) highlighted an empirical difference in behavior between sigmoids and ReLUs, which our analytical theory could explain. Another interesting direction is to extend our results to networks with learnable second-layer weights. Once again, the model convergence can be studied using our fluctuation-dissipation argument, although the application would not be trivial since the learning rate effectively varies between student weights. Lastly, it should be possible to generalize our results to non-Gaussian inputs using Gaussian equivalence theory (Goldt et al., 2022).

AI USE STATEMENT

In this work, LLMs provided an important link to existing literature, namely with dissipation-fluctuations theory and adaptive filters, which were critical ingredients in developing our mathematical framework. After an initial formulation of generic mathematical claims, their application in the context of our work was human-made. We did not use generative AI tools to propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, support qualitative and thematic data analysis, or interpret results. Dataset-related uses do not apply to this work. Additionally, we used generative AI tools as an aid in writing software code, while most of our code is human-written, and to review the manuscript. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of LLMs.

Reproducibility statement

Code available to reproduce the numerical simulations at:

https://anonymous.4open.science/r/teacher-student-iclr.

References

Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, 1964.

Haruka Asanuma, Shiro Takagi, Yoshihiro Nagano, Yuki Yoshida, Yasuhiko Igarashi, and Masato Okada. Statistical mechanical analysis of catastrophic forgetting in continual learning with teacher and student networks. *J. Phys. Soc. Jpn.*, 90(10):104001, 2021. doi: 10.7566/jpsj.90.104001.

Jonathan Baxter. A model of inductive bias learning. *J. Artif. Intell. Res.*, 12(1):149–198, 2000. doi: 10.1613/jair.731.

Michael Biehl and Holm Schwarze. Learning by on-line gradient descent. *J. Phys. A*, 28(3):643, 1995. doi: 10.1088/0305-4470/28/3/018.

Rich Caruana. Multitask learning. *Mach. Learn.*, 28:41–75, 1997. doi: 10.1023/A:1007379606734.

Otávio Citton, Frederieke Richert, and Michael Biehl. On-line learning dynamics in layered neural networks with arbitrary activation functions. In *Proc. ESANN 2024*, pp. 437–442. i6doc.com publ., 2024. doi: 10.14428/esann/2024.ES2024-116.

Otávio Citton, Frederieke Richert, and Michael Biehl. The role of the learning rate in layered neural networks with ReLU activation function. In *Proc. ESANN 2025*, pp. 437–442. i6doc.com publ., 2025a. doi: 10.14428/esann/2025.ES2025-94.

Otávio Citton, Frederieke Richert, Michael Biehl, and Michiel Straat. Hermite polynomials facilitating on-line learning analysis of layered neural networks with arbitrary activation function. *Neurocomputing*, 655:131328, 2025b. doi: 10.1016/j.neucom.2025.131328.

Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv*, 2020. doi: 10.48550/arXiv.2009.09796.

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 *NeurIPS*, volume 32. Curran Associates, Inc., 2019. URL `proceedings.neurips.cc/paper/2019/hash/cab070d53bd0d200746fb852a922064a-Abstract.html`.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. The gaussian equivalence of generative models for learning with shallow neural networks. In *Proc. MSML*, volume 145, pp. 426–471. PMLR, 2022. URL `proceedings.mlr.press/v145/goldt22a.html`.

Simon Haykin. *Adaptive Filter Theory*. Inf. Syst. Sci. Ser. Prentice-Hall, 1986.

Masato Inoue, Hyeyoung Park, and Masato Okada. On-line learning theory of soft committee machines with correlated hidden units -Steepest gradient descent and natural gradient descent-. *J. Phys. Soc. Jpn.*, 72(4):805–810, 2003. doi: 10.1143/jpsj.72.805.

Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in SGD. *arXiv*, 2018. doi: 10.48550/arXiv.1711.04623.

Raymond Kan and Cesare Robotti. On moments of folded and truncated multivariate normal distributions. *J. Comput. Graph. Stat*, 26(4):930–934, 2017. doi: 10.1080/10618600.2017.1322092.

Sebastian Lee, Sebastian Goldt, and Andrew M. Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *Proc. ICML*, volume 139, pp. 6109–6119. PMLR, 2021. URL `proceedings.mlr.press/v139/lee21e.html`.

Lukas Liebel and Marco Körner. Auxiliary tasks in multi-task learning. *arXiv*, 2018. doi: 10.48550/arXiv.1805.06334.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *J. Mach. Learn. Res.*, 18(134):1–35, 2017. URL jmlr.org/papers/v18/17-214.html.

Francesco Mori and Francesca Mignacco. Analytic theory of dropout regularization. *Phys. Rev. E*, 112:045301, 2025. doi: 10.1103/jmdx-x3gr.

Bengt Muthén. Moments of the censored and truncated bivariate normal distribution. *Br. J. Math. Stat. Psychol.*, 43(1):131–143, 1990. doi: 10.1111/j.2044-8317.1990.tb00930.x.

Donald B. Owen. Tables for computing bivariate normal probabilities. *Ann. Math. Stat.*, 27(4): 1075–1090, 1956. doi: 10.1214/aoms/1177728074.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv*, 2017. doi: 10.48550/arXiv.1706.05098.

David Saad. *On-Line Learning in Neural Networks*. Publ. Newton Inst. Cambridge University Press, 1999. doi: 10.1017/CBO9780511569920.

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

David Saad and Sara A. Solla. On-line learning in soft committee machines. *Phys. Rev. E*, 52: 4225–4243, 1995b. doi: 10.1103/PhysRevE.52.4225.

Luca Saglietti, Stefano Mannelli, and Andrew M. Saxe. An analytical theory of curriculum learning in teacher-student networks. In *NeurIPS*, volume 35, pp. 21113–21127. Curran Associates, Inc., 2022. doi: 10.52202/068431-1535.

Ali H. Sayed. *Adaptive Filters*. John Wiley & Sons, Ltd, 2008. doi: 10.1002/9780470374122.

Michiel Straat and Michael Biehl. On-line learning dynamics of ReLU neural networks using statistical physics techniques. In *Proc. ESANN 2019*. i6doc.com publ., 2019. URL www.esann.org/sites/default/files/proceedings/legacy/es2019-173.pdf.

Renato Vicente and Nestor Caticha. Functional optimization of online algorithms in multilayer neural networks. *J. Phys. A*, 30(17):L599, 1997. doi: 10.1088/0305-4470/30/17/002.

Bernard Widrow and Eugene Walach. Adaptive signal processing for adaptive control. *IFAC Proc. Vol.*, 16(9):7–12, 1983. doi: 10.1016/S1474-6670(17)62348-6.

Sho Yaida. Fluctuation-dissipation relations for stochastic gradient descent. In *ICLR*. OpenReview, 2019. URL openreview.net/forum?id=SkNksoRctQ.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.*, 34 (12):5586–5609, 2022. doi: 10.1109/TKDE.2021.3070203.

APPENDIX

A DERIVATION OF THE AUXILIARY DYNAMICS DIFFERENTIAL EQUATIONS

The derivation of the set of differential equations for the order parameters closely follows the derivation for the one-output case by Saad & Solla (1995a;b); Biehl & Schwarze (1995); Vicente & Caticha (1997); Saad (1999); Inoue et al. (2003); Goldt et al. (2019). Let us consider the generalization error

$$\epsilon_{\text{main}}(\theta, \theta^*) = \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{N}}\left[\left(y^*_{\text{main}}(\boldsymbol{x}) - y_{\text{main}}(\boldsymbol{x})\right)^2\right]. \tag{23}$$

Expanding the square, we immediately get

$$\begin{aligned}\epsilon_{\text{main}} &= \sum_{i,j} v_i v_j \mathbb{E}\left[g(h_i)g(h_j)\right] + \sum_{n,m} v^*_n v^*_m \mathbb{E}\left[g(h^*_n)g(h^*_m)\right] - 2\sum_{i,n} v_i v^*_n \mathbb{E}\left[g(h_i)g(h^*_n)\right] \\ &= \sum_{i,j} v_i v_j I_2(i,j) + \sum_{n,m} v^*_n v^*_m I_2(n,m) - 2\sum_{i,n} v_i v^*_n I_2(i,n),\end{aligned} \tag{24}$$

with the notation $I_2(i,j) = \mathbb{E}[g(h_i)\, g(h_j)]$, i, j being indices of units in the student network and n, m teacher ones. Since the inputs are taken to be $i.i.d.$ Gaussian, the pre-activations

$$h_i = \frac{\boldsymbol{w}_i \cdot \boldsymbol{x}}{\sqrt{N}} \tag{25}$$

are jointly Gaussian together with the teacher pre-activations $\boldsymbol{h}^*$, having zero-mean and a $2K \times 2K$ covariance matrix with block structure

$$\text{cov} = \begin{pmatrix} Q & R \\ R^T & T \end{pmatrix}, \tag{26}$$

with

$$\begin{aligned} Q_{ik} &= \mathbb{E}\left[h_i h_k\right] = \frac{\boldsymbol{w}_i^T}{\sqrt{N}} \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right] \frac{\boldsymbol{w}_k}{\sqrt{N}} = \frac{\boldsymbol{w}_i^T \boldsymbol{w}_k}{N}, \\ R_{in} &= \mathbb{E}\left[h_i h^*_n\right] = \frac{\boldsymbol{w}_i^T \boldsymbol{w}^*_n}{N}, \\ T_{mn} &= \mathbb{E}\left[h^*_m h^*_n\right] = \frac{\boldsymbol{w}^{*T}_m \boldsymbol{w}^*_n}{N} \xrightarrow[N\to\infty]{} \delta_{mn}. \end{aligned} \tag{27}$$

Notice that Q and R evolve during training, while T does not.

To compute the generalization error, we only need to compute either analytically or numerically the integral I_2 (namely, a two-dimensional Gaussian average of a nonlinear function). Notice the convention, valid from now on, that i, k (and j, l) denote student indices, and n, m teacher indices.

Let us now consider the dynamics. Starting from Equation 3, for the two-output case, they are described by the set of equations

$$
\begin{aligned}
W_{ik}^{t+1} &= W_{ik}^{t} - \mu\Bigg[\alpha\, v_i\Bigg(\sum_m v_m^*\, g\left(\frac{\boldsymbol{w}_m^*\cdot\boldsymbol{x}}{\sqrt{N}}\right) - \sum_j v_j\, g\left(\frac{\boldsymbol{w}_j\cdot\boldsymbol{x}}{\sqrt{N}}\right) + \sigma z_{\text{main},t}\Bigg) \\
&\quad + \beta\, u_i\Bigg(\sum_m u_m^*\, g\left(\frac{\boldsymbol{w}_m^*\cdot\boldsymbol{x}}{\sqrt{N}}\right) - \sum_j u_j\, g\left(\frac{\boldsymbol{w}_j\cdot\boldsymbol{x}}{\sqrt{N}}\right) + \sigma z_{\text{aux},t}\Bigg)\Bigg] g'\left(\frac{\boldsymbol{w}_i\cdot\boldsymbol{x}}{\sqrt{N}}\right)\frac{x_k}{\sqrt{N}}, \\
v_i^{t+1} &= v_i^t - \frac{\mu}{N}\alpha\Bigg(\sum_m v_m^*\, g\left(\frac{\boldsymbol{w}_m^*\cdot\boldsymbol{x}}{\sqrt{N}}\right) - \sum_j v_j\, g\left(\frac{\boldsymbol{w}_j\cdot\boldsymbol{x}}{\sqrt{N}}\right) + \sigma z_{\text{main},t}\Bigg)\, g\left(\frac{\boldsymbol{w}_i\cdot\boldsymbol{x}}{\sqrt{N}}\right), \\
u_i^{t+1} &= u_i^t - \frac{\mu}{N}\beta\Bigg(\sum_m u_m^*\, g\left(\frac{\boldsymbol{w}_m^*\cdot\boldsymbol{x}}{\sqrt{N}}\right) - \sum_j u_j\, g\left(\frac{\boldsymbol{w}_j\cdot\boldsymbol{x}}{\sqrt{N}}\right) + \sigma z_{\text{aux},t}\Bigg)\, g\left(\frac{\boldsymbol{w}_i\cdot\boldsymbol{x}}{\sqrt{N}}\right),
\end{aligned}
\tag{28}
$$

where the second-layer weights have a rescaled learning rate μ/N. Taking the appropriate expectation of these equations, and rescaling the time variable $\tau = t/N$, we obtain

$$
\frac{\mathrm{d}v_k}{\mathrm{d}\tau} = \alpha\mu\Bigg(\sum_m v_m^* I_2(k,m) - \sum_j v_j I_2(k,j)\Bigg), \tag{29a}
$$

$$
\frac{\mathrm{d}u_k}{\mathrm{d}\tau} = \beta\mu\Bigg(\sum_m u_m^* I_2(k,m) - \sum_j u_j I_2(k,j)\Bigg), \tag{29b}
$$

$$
\begin{aligned}
\frac{\mathrm{d}Q^{ik}}{\mathrm{d}\tau} &= \mu\sum_m\Big[\left(\alpha v_m^* v_k + \beta u_m^* u_k\right) I_3(k,i,m) + \left(\alpha v_m^* v_i + \beta u_m^* u_i\right) I_3(i,k,m)\Big] \\
&\quad - \mu\sum_j\Big[\left(\alpha v_j v_k + \beta u_j u_k\right) I_3(k,i,j) + \left(\alpha v_j v_i + \beta u_j u_i\right) I_3(i,k,j)\Big] \\
&\quad + \mu^2\sum_{n,m}\Big[\left(\alpha^2 v_i v_k v_m^* v_n^* + \alpha\beta(v_k u_i + v_i u_k) u_n^* v_m^* + \beta^2 u_i u_k u_n^* u_m^*\right) I_4(i,k,n,m)\Big] \\
&\quad + \mu^2\sum_{j,l}\Big[\left(\alpha^2 v_i v_k v_j v_l + \alpha\beta(v_k u_i + v_i u_k) u_j v_l + \beta^2 u_i u_k u_j u_l\right) I_4(i,k,j,l)\Big] \\
&\quad - \mu^2\sum_{j,n}\Big[\left(2\alpha^2 v_i v_k v_j v_n^* + \alpha\beta(v_k u_i + v_i u_k)(u_j v_n^* + v_j u_n^*) + 2\beta^2 u_i u_k u_j u_n^*\right) I_4(i,k,j,n)\Big] \\
&\quad + \mu^2\sigma^2(\alpha^2 v_k v_i + \beta^2 u_k u_i) J_2(i,k),
\end{aligned}
\tag{29c}
$$

$$
\frac{\mathrm{d}R^{in}}{\mathrm{d}\tau} = \mu\sum_m\left(\alpha v_m^* v_i + \beta u_m^* u_i\right) I_3(i,n,m) - \mu\sum_j\left(\alpha v_i v_j + \beta u_i u_j\right) I_3(i,n,j), \tag{29d}
$$

depending on the integrals

$$
\begin{aligned}
I_2 &= \mathbb{E}[g(x)\, g(y)], \\
I_3 &= \mathbb{E}[g'(x)\, y\, g(z)], \\
I_4 &= \mathbb{E}[g'(x)\, g'(y)\, g(z)\, g(h)], \\
J_2 &= \mathbb{E}[g'(x)\, g'(y)].
\end{aligned}
\tag{30}
$$

Thus, the system dynamics are described by a non-extensive set of mean-field ODEs, which can be numerically integrated as long as we can compute these integrals. We will now discuss two common choices of activation functions.

A.1 Error function

If $g(x) = \mathrm{erf}(x/\sqrt{2})$, then all integrals can be computed analytically (Saad & Solla, 1995a;b; Biehl & Schwarze, 1995; Goldt et al., 2019) and we get

$$
\begin{aligned}
I_2(x,y) &= \mathbb{E}\left[g(x)g(y)\right] = \frac{2}{\pi}\arcsin\left(\frac{c_{xy}}{\sqrt{(1+c_{xx})(1+c_{yy})}}\right),\\
J_2(x,y) &= \mathbb{E}\left[g'(x)g'(y)\right] = \frac{2}{\pi\sqrt{\Lambda_4}},\\
I_3(x,y,z) &= \mathbb{E}\left[g'(x)\,y\,g(z)\right] = \frac{2}{\pi}\frac{(1+c_{xx})c_{yz}-c_{yx}c_{xz}}{(1+c_{xx})\sqrt{\Lambda_4}},\\
I_4(x,y,z,h) &= \mathbb{E}\left[g'(x)g'(y)g(z)g(h)\right] = \frac{4}{\pi^2}\frac{1}{\sqrt{\Lambda_4}}\arcsin\Big(\frac{\Lambda_0}{\sqrt{\Lambda_1\Lambda_2}}\Big),
\end{aligned}
\tag{31}
$$

with

$$
\begin{aligned}
\Lambda_4 &= (1+c_{xx})(1+c_{yy})-c_{xy}^2,\\
\Lambda_0 &= \Lambda_4 c_{zh} - c_{yz}c_{yh}(1+c_{xx}) - c_{xz}c_{xh}(1+c_{yy}) + c_{xy}(c_{xz}c_{yh}+c_{yz}c_{xh}),\\
\Lambda_1 &= \Lambda_4(1+c_{zz}) - c_{yz}^2(1+c_{xx}) - c_{xz}^2(1+c_{yy}) + 2c_{xy}c_{xz}c_{yz},\\
\Lambda_2 &= \Lambda_4(1+c_{hh}) - c_{yh}^2(1+c_{xx}) - c_{xh}^2(1+c_{yy}) + 2c_{xy}c_{xh}c_{yh}.
\end{aligned}
\tag{32}
$$

A.2 Rectified linear unit

If $g(x) = \text{ReLU}(x) = \max(x,0)$, then some of the integrals can be computed analytically (Straat & Biehl, 2019), namely

$$
\begin{aligned}
I_2(x,y) &= \mathbb{E}\left[g(x)g(y)\right] = \frac{c_{xy}}{4} + \frac{1}{2\pi}\left(c_{xy}\arcsin\left(\frac{c_{xy}}{\sqrt{c_{xx}c_{yy}}}\right) + \sqrt{c_{xx}c_{yy}-c_{xy}^2}\right),\\
J_2(x,y) &= \mathbb{E}\left[g'(x)g'(y)\right] = \frac{1}{4} + \frac{1}{2\pi}\arcsin\left(\frac{c_{xy}}{\sqrt{c_{xx}c_{yy}}}\right),\\
I_3(x,y,z) &= \mathbb{E}\left[g'(x)\,y\,g(z)\right] = \frac{c_{xy}}{2\pi c_{xx}}\sqrt{c_{xx}c_{zz}-c_{xz}^2} + \frac{1}{2\pi}c_{yz}\arcsin\left(\frac{c_{xz}}{\sqrt{c_{xx}c_{zz}}}\right) + \frac{c_{yz}}{4}.
\end{aligned}
\tag{33}
$$

The integral $I_4 = \mathbb{E}\left[g'(x)g'(y)g(z)g(h)\right] = \mathbb{E}\left[\mathbf{1}_{x>0}\,\mathbf{1}_{y>0}\max(z,0)\max(h,0)\right]$ cannot be computed analytically. Citton et al. (2024; 2025b;a) proposed a Hermite expansion to approximate it numerically, but we found that it suffers from convergence issues for ill-conditioned matrices. We instead used a conditional Gaussian decomposition with Gauss–Hermite quadrature to approximate it.

Partition C as

$$
C = \begin{pmatrix} A & B^T \\ B & D \end{pmatrix}, \quad A = \text{Cov}(x,y), \quad D = \text{Cov}(z,h), \quad B = \text{Cov}\big((z,h),(x,y)\big). \tag{34}
$$

By the Gaussian conditioning formula, (z,h) conditioned on $(x,y) = \mathbf{u}$ is Gaussian:

$$
(z,h) \mid (x,y) = \mathbf{u} \;\sim\; \mathcal{N}\left(\underbrace{BA^{-1}\mathbf{u}}_{=\mathbf{m}(\mathbf{u})},\; \underbrace{D - BA^{-1}B^T}_{=S}\right). \tag{35}
$$

S is the Schur complement of A in C, and is positive semi-definite. Then our integral becomes

$$
I_4 = \int_{\mathbb{R}^2_+} \underbrace{\mathbb{E}\Big[\max(z,0)\,\max(h,0)\;\Big|\;(x,y)=\mathbf{u}\Big]}_{=K(\mathbf{m}(\mathbf{u}),S)}\, p(\mathbf{u})\,d\mathbf{u}, \tag{36}
$$

Writing $s_1 = \sqrt{S_{11}}$, $s_2 = \sqrt{S_{22}}$, $\rho = S_{12}/(s_1 s_2)$, $a = m_1/s_1$, $b = m_2/s_2$, $k = \sqrt{1-\rho^2}$, the closed form for K is given by Muthén (1990); Kan & Robotti (2017).

$$K(\mathbf{m}, S) = s_1 s_2 \Big[(ab+\rho)\, \Phi_2(a, b; \rho) + a\, \varphi(b)\, \Phi\Big(\tfrac{a-\rho b}{k}\Big) + b\, \varphi(a)\, \Phi\Big(\tfrac{b-\rho a}{k}\Big) + k^2\, \varphi_2(a, b; \rho) \Big], \quad (37)$$

where φ, Φ are the univariate standard normal density and CDF, and φ_2, Φ_2 their bivariate analogues at correlation ρ. The bottleneck in this expression is Φ_2, which has no elementary closed form. We evaluate it via Owen's identity (Owen, 1956),

$$\frac{\partial \Phi_2}{\partial \rho}(a, b; \rho) = \varphi_2(a, b; \rho) \quad \Longrightarrow \quad \Phi_2(a, b; \rho) = \Phi(a)\Phi(b) + \int_0^\rho \varphi_2(a, b; t)\, dt, \quad (38)$$

which is a smooth, one-dimensional integral over a compact interval $[0, \rho] \subset [-1, 1]$ and converges rapidly under low-order Gauss–Legendre quadrature. Then I_4 can be approximated by Gauss–Hermite quadrature (Abramowitz & Stegun, 1964). When C is singular, then I_4 always reduces to one of two integrals that are known analytically. Citton et al. (2025a) described them in their additional material.

B Generalization error of the two-output linear model

Given matrix $W^* \in \mathbb{R}^{K \times N}$ and two vectors $\boldsymbol{v},\ \boldsymbol{u} \in \mathbb{R}^K$, we want to compute

$$\epsilon_{\text{main}} = \lim_{t \to \infty} \epsilon_{\text{main}}(t) = \lim_{t \to \infty} \mathop{\mathbb{E}}_{x \sim \mathcal{N}} \left[\left(\frac{\boldsymbol{v}^T W^* \boldsymbol{x}}{\sqrt{N}} - \frac{\boldsymbol{v}^T W_t \boldsymbol{x}}{\sqrt{N}} \right)^2 \right], \quad (39)$$

where

$$W_{t+1} = W_t - \mu \nabla_W \mathcal{L}(t), \quad (40)$$

and

$$\begin{aligned} \mathcal{L}(t) = \frac{1}{2} \Bigg[& \alpha \left(\frac{\boldsymbol{v}^T W^* \boldsymbol{x}_t}{\sqrt{N}} - \frac{\boldsymbol{v}^T W_t \boldsymbol{x}_t}{\sqrt{N}} + \sigma z_{\text{main},t} \right)^2 \\ & + \beta \left(\frac{\boldsymbol{u}^T W^* \boldsymbol{x}_t}{\sqrt{N}} - \frac{\boldsymbol{u}^T W_t \boldsymbol{x}_t}{\sqrt{N}} + \sigma z_{\text{aux},t} \right)^2 \Bigg], \end{aligned} \quad (41)$$

with $\alpha, \beta \in \mathbb{R}^+$, $\alpha \neq 0$, and $\boldsymbol{x}_t \sim \mathcal{N}(0, I_N)$ and $z_{\text{main}}, z_{\text{aux}} \in \mathcal{N}(0, 1)$.

Then,

$$\begin{aligned} W_{t+1} = W_t + \mu \Bigg[& \alpha \left(\frac{\boldsymbol{v}^T W^* \boldsymbol{x}_t}{\sqrt{N}} - \frac{\boldsymbol{v}^T W_t \boldsymbol{x}_t}{\sqrt{N}} + \sigma z_{\text{main},t} \right) \boldsymbol{v} \boldsymbol{x}_t^T \\ & + \beta \left(\frac{\boldsymbol{u}^T W^* \boldsymbol{x}_t}{\sqrt{N}} - \frac{\boldsymbol{u}^T W_t \boldsymbol{x}_t}{\sqrt{N}} + \sigma z_{\text{aux},t} \right) \boldsymbol{u} \boldsymbol{x}_t^T \Bigg]. \end{aligned} \quad (42)$$

We can now project the weight matrix onto the task-relevant directions, calling $\boldsymbol{p}_t := \boldsymbol{v}_t^T W/\sqrt{N}$ and $\boldsymbol{q}_t := \boldsymbol{u}_t^T W/\sqrt{N}$, then

$$\begin{aligned}\boldsymbol{p}_{t+1} &= \boldsymbol{p}_t + \frac{\mu}{N}\Big[\alpha|\boldsymbol{v}|^2(\boldsymbol{p}^{*T}\boldsymbol{x}_t - \boldsymbol{p}_t^T\boldsymbol{x}_t + \sigma z_{\text{main},t})\boldsymbol{x}_t \\ &\qquad +\beta\rho|\boldsymbol{v}||\boldsymbol{u}|(\boldsymbol{q}^{*T}\boldsymbol{x}_t - \boldsymbol{q}_t^T\boldsymbol{x}_t + \sigma z_{\text{aux},t})\boldsymbol{x}_t\Big], \\ \boldsymbol{q}_{t+1} &= \boldsymbol{q}_t + \frac{\mu}{N}\Big[\alpha\rho|\boldsymbol{v}||\boldsymbol{u}|(\boldsymbol{p}^{*T}\boldsymbol{x}_t - \boldsymbol{p}_t^T\boldsymbol{x}_t + \sigma z_{\text{main},t})\boldsymbol{x}_t \\ &\qquad +\beta|\boldsymbol{u}|^2(\boldsymbol{q}^{*T}\boldsymbol{x}_t - \boldsymbol{q}_t^T\boldsymbol{x}_t + \sigma z_{\text{aux},t})\boldsymbol{x}_t\Big],\end{aligned} \tag{43}$$

where $\rho = \boldsymbol{u}^T\boldsymbol{v}/|\boldsymbol{v}||\boldsymbol{u}|$.

Introducing the deviations from the teacher $\delta\boldsymbol{p}_t = \boldsymbol{p}^* - \boldsymbol{p}_t$ and $\delta\boldsymbol{q}_t = \boldsymbol{q}^* - \boldsymbol{q}_t$, we get

$$\begin{aligned}\delta\boldsymbol{p}_{t+1} &= \delta\boldsymbol{p}_t - \frac{\mu}{N}\left[\alpha|\boldsymbol{v}|^2(\delta\boldsymbol{p}_t^T\boldsymbol{x}_t + \sigma z_{\text{main},t})\boldsymbol{x}_t + \beta\rho|\boldsymbol{v}||\boldsymbol{u}|(\delta\boldsymbol{q}_t^T\boldsymbol{x}_t + \sigma z_{\text{aux},t})\boldsymbol{x}_t\right], \\ \delta\boldsymbol{q}_{t+1} &= \delta\boldsymbol{q}_t - \frac{\mu}{N}\left[\alpha\rho|\boldsymbol{v}||\boldsymbol{u}|(\delta\boldsymbol{p}_t^T\boldsymbol{x}_t + \sigma z_{\text{main},t})\boldsymbol{x}_t + \beta|\boldsymbol{u}|^2(\delta\boldsymbol{q}_t^T\boldsymbol{x}_t + \sigma z_{\text{aux},t})\boldsymbol{x}_t\right].\end{aligned} \tag{44}$$

Then, the generalization error at time t can be computed as

$$\epsilon_{\text{main}}(t) = \mathbb{E}_x\left[(\delta\boldsymbol{p}_t^T\boldsymbol{x})^2\right] = \delta\boldsymbol{p}_t^T\delta\boldsymbol{p}_t. \tag{45}$$

We introduce the variable $\boldsymbol{\phi}_t^T = (\delta\boldsymbol{p}_t^T, \delta\boldsymbol{q}_t^T)$, and we choose, for simplicity, our second layer weights to have the same norm, namely $|\boldsymbol{v}| = |\boldsymbol{u}|$. Then,

$$\boldsymbol{\phi}_{t+1} = \left(I - \eta A \otimes (\boldsymbol{x}_t\boldsymbol{x}_t^T)\right)\boldsymbol{\phi}_t - \eta\sigma\,\boldsymbol{b}_t \otimes \boldsymbol{x}_t, \tag{46}$$

where we defined a new learning rate $\eta := \dfrac{\mu|\boldsymbol{v}|^2}{N}$, as well as the quantities

$$A = \begin{pmatrix}\alpha & \beta\rho \\ \alpha\rho & \beta\end{pmatrix}, \tag{47}$$

and

$$\boldsymbol{b}_t = \begin{pmatrix}\alpha\, z_{\text{main},t} + \beta\rho\, z_{\text{aux},t} \\ \alpha\rho\, z_{\text{main},t} + \beta\, z_{\text{aux},t}\end{pmatrix}. \tag{48}$$

Given that the weights at time t are independent of the current input $\boldsymbol{x}_t$, namely $\boldsymbol{x}_t \perp \boldsymbol{\phi}_t$, then it follows, taking the expectation over the whole dynamics

$$\mathbb{E}[\boldsymbol{\phi}_{t+1}] = (I - \eta A \otimes I)\,\mathbb{E}[\boldsymbol{\phi}_t], \tag{49}$$

which converges if

$$0 < \mu < \frac{2N}{|\boldsymbol{v}|^2\,\lambda_{max}(A)} = \frac{4N}{|\boldsymbol{v}|^2\left(\alpha + \beta + \sqrt{(\alpha-\beta)^2 + 4\rho^2\alpha\beta}\right)}. \tag{50}$$

We are then interested in the quantity $C_t = \boldsymbol{\phi}_t\boldsymbol{\phi}_t^T$, from which we can derive the generalization error.

Applying the Gaussian moment theorem results in the identity

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T M \boldsymbol{x}\boldsymbol{x}^T] = \text{tr}(M)I + M + M^T, \tag{51}$$

we get

$$C_{t+1} = C_t - \eta(A \otimes I)C_t - \eta C_t(A^T \otimes I) + \eta^2\sigma^2\Sigma \otimes I + \mathcal{O}(\eta^3), \tag{52}$$

where

$$\Sigma = \mathbb{E}_t\left[\boldsymbol{b}_t\boldsymbol{b}_t^T\right] = \begin{pmatrix} \alpha^2 + \beta^2\rho^2 & \rho(\alpha^2 + \beta^2) \\ \rho(\alpha^2 + \beta^2) & \alpha^2\rho^2 + \beta^2 \end{pmatrix} = AA^T. \tag{53}$$

Our matrix C inherits the Kronecker structure from the other terms, namely $C = \tilde{C} \otimes I$. This reduces the equation to

$$\tilde{C}_{t+1} = \tilde{C}_t - \eta A\tilde{C}_t - \eta\,\tilde{C}_t A^T + \eta^2\sigma^2\Sigma + \mathcal{O}(\eta^3), \tag{54}$$

Now, we impose stationary, setting $\tilde{C}_{t+1} = \tilde{C}_t = \tilde{C}_\infty$, and we can solve for $\tilde{C}_\infty$, obtaining

$$\begin{aligned} \epsilon_{\text{main}} = \lim_{t\to\infty}\left[\epsilon_{\text{main}}(t)\right] = \delta\boldsymbol{p}_\infty^T\delta\boldsymbol{p}_\infty = \sum_{i=1}^{N}(C_\infty)_{ii} = N(\tilde{C}_\infty)_{11} = \\ = \frac{\mu\sigma^2|\boldsymbol{v}|^2}{2}\frac{\alpha^2 + \alpha\beta(1-\rho^2) + \beta^2\rho^2}{\alpha+\beta} + \mathcal{O}(\mu^2). \end{aligned} \tag{55}$$

Different noise variance in auxiliary labels. Let us now consider the same linear model as before, with the addition that the auxiliary labels y_{aux} may have noise with a different variance σ^2_{aux}, while the main task noise has variance σ^2. With an analogous calculation, the leading order in μ of the generalization error is

$$\begin{aligned} \epsilon_{\text{main}} = \frac{\mu|\boldsymbol{v}|^2\sigma^2}{2}\frac{\left(\alpha^2 + \alpha\beta(1-\rho^2)\right) + \frac{\sigma^2_{\text{aux}}}{\sigma^2}\rho^2\beta^2}{\alpha+\beta} = \\ = \epsilon_0\frac{\alpha^2 + \alpha\beta(1-\rho^2) + \delta^2\rho^2\beta^2}{\alpha+\beta}, \end{aligned} \tag{56}$$

where we have defined the noise scale $\delta = \sigma_{\text{aux}}/\sigma$.

When $\alpha = 1$

$$\epsilon_{\text{main}}\Big|_{\alpha=1} = \epsilon_0\left(1 - \rho^2\frac{\beta(1-\beta\delta^2)}{1+\beta}\right) \tag{57}$$

minimized by

$$\beta^* = \sqrt{1 + \frac{1}{\delta^2}} - 1. \tag{58}$$

Notice that β^* can be larger than 1 and does not depend on ρ. For every choice of σ^2_{aux}, including when $\sigma^2_{\text{aux}} > \sigma^2$, there is always a $\beta^* \neq 0$ that improves the corresponding generalization error $\epsilon_{\text{main}}|_{\beta^*} = \epsilon^*_{\text{main}}$ compared to ϵ_0, as shown in Figure S1. Of course, when $\delta > 1$, the optimal β^* is small with only marginal improvements in generalization, and setting $\beta > \beta^*$ can degrade performance.

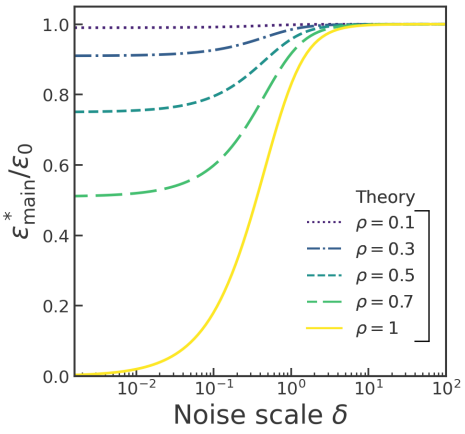


Figure S1: **Optimal generalization for different noise scales.** Leading term of the generalization error on the main task of the two-output linear model computed at the optimal β^* as a function of noise scale $\delta = \sigma_{\text{aux}}/\sigma$ and different ρ, with $\alpha = 1$.

Multiple auxiliary labels. Consider a model with two auxiliary labels instead $y_{\text{aux},1}$ and $y_{\text{aux},2}$, each having the same label noise σ. Namely, the training loss will be

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}) &= \alpha \mathcal{L}_{\text{main}}(\boldsymbol{x}) + \beta \mathcal{L}_{\text{aux},1}(\boldsymbol{x}) + \gamma \mathcal{L}_{\text{aux},2}(\boldsymbol{x}), \\ \mathcal{L}_i(\boldsymbol{x}) &= \frac{1}{2}\Big(y_i^*(\boldsymbol{x}) - y_i(\boldsymbol{x}) + \sigma z_i\Big)^2 . \end{aligned} \tag{59}$$

From this, the generalization error on the main task is, at the first order in μ,

$$\epsilon_{\text{main}} = \frac{\mu\, \sigma^2\, |\boldsymbol{v}|^2}{2}\, \frac{\mathcal{N}}{\mathcal{D}}, \tag{60}$$

with numerator

$$\begin{aligned} \mathcal{N} = \alpha^3 &\Big(\beta(1-\rho_1^2) + \gamma(1-\rho_2^2)\Big) \\ &+ \alpha^2 \Big(\big[\beta(1-\rho_1^2) + \gamma(1-\rho_2^2)\big]^2 + \beta\gamma\big[1 - \rho_{12}^2 - \Delta\big]\Big) \\ &+ \alpha\Big(\beta^3\rho_1^2(1-\rho_{12}^2) + \gamma^3\rho_2^2(1-\rho_2^2) \\ &\qquad + \beta\gamma\big[(\beta+\gamma)\Delta + \beta\rho_1^2(1-\rho_2^2) + \gamma\rho_2^2(1-\rho_{12}^2)\big]\Big) \\ &+ \beta\gamma\Big((\beta^2\rho_1^2 + \gamma^2\rho_2^2)(1-\rho_{12}^2) + \beta\gamma\big[(\rho_1 - \rho_2\rho_{12})^2 + (\rho_2 - \rho_1\rho_{12})^2\big]\Big), \end{aligned} \tag{61}$$

and denominator

$$\mathcal{D} = (\alpha+\beta)(\alpha+\gamma)(\beta+\gamma) - (\alpha+\beta+\gamma)\big(\alpha\beta\rho_1^2 + \alpha\gamma\rho_2^2 + \beta\gamma\rho_{12}^2\big) + \alpha\beta\gamma(1-\Delta), \tag{62}$$

where ρ_1 and ρ_2 are the correlations between the main task and the auxiliary tasks, ρ_{12} is the correlation between the two auxiliary tasks, and $\Delta = 1 + 2\rho_1\rho_2\rho_{12} - \rho_1^2 - \rho_2^2 - \rho_{12}^2$ is the Gram determinant of the correlation matrix between second-layer weights

$$G = \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_{12} \\ \rho_2 & \rho_{12} & 1 \end{pmatrix}. \tag{63}$$

One could focus on many different correlation patterns between tasks. For example, we choose two outputs uncorrelated with each other, each having the same correlation with the main task, namely $\rho_{12} = 0$, $\rho_1 = \rho_2 = \rho$. We also choose $\gamma = \beta$, $\alpha = 1$. Then

$$\epsilon_{\text{main}} = \epsilon_0 \left(1 - 2\rho^2 \frac{(1-\beta)\beta}{\beta+1}\right). \tag{64}$$

It is straightforward to see that $\beta^* = \sqrt{2} - 1$, analogous to the two-output setup, but corresponding to a better generalization error than the minimum of Equation 9.

Once again, regardless of the choices of correlation between outputs,

$$\alpha\epsilon_{\text{main}} + \beta\epsilon_{\text{aux},1} + \gamma\epsilon_{\text{aux},2} = \epsilon_0\big(\alpha^2 + \beta^2 + \gamma^2\big). \tag{65}$$

C Effective dynamics with perfectly correlated output

Let us consider the case where the second layer weights are the same, $\boldsymbol{v} = \boldsymbol{v}^* = \boldsymbol{u} = \boldsymbol{u}^*$,

$$\mathcal{L}(t) = \frac{1}{2}\left[\alpha(y_1^* - y_1 + \sigma z_{\text{main},t})^2 + \beta(y_1^* - y_1 + \sigma z_{\text{aux},t})^2\right], \tag{66}$$

and, regardless of the choice of activation g,

$$\begin{aligned} W^{t+1} &= W^t + \mu\left[\alpha(y_1^* - y_1 + \sigma z_{\text{main},t}) + \beta(y_1^* - y_1 + \sigma z_{\text{aux},t})\right]\left(\boldsymbol{v} \odot \boldsymbol{g}'(h)\right)\frac{\boldsymbol{x}^T}{\sqrt{N}} = \\ &= W^t + (\alpha+\beta)\mu\left[y_1^* - y_1 + \sigma\frac{\alpha z_{\text{main},t} + \beta z_{\text{aux},t}}{\alpha+\beta}\right]\left(\boldsymbol{v} \odot \boldsymbol{g}'(h)\right)\frac{\boldsymbol{x}^T}{\sqrt{N}}, \end{aligned} \tag{67}$$

where $\boldsymbol{g}'(h)_i \equiv g'(h_i)$ and $\odot$ is the Hadamard (element-wise) product. This is equivalent to training a teacher-student network with one output, with an effective learning rate and noise:

$$\begin{aligned} \mu_{\text{eff}} &= (\alpha+\beta)\mu, \\ \sigma^2_{\text{eff}} &= \frac{\alpha^2+\beta^2}{(\alpha+\beta)^2}\sigma^2. \end{aligned} \tag{68}$$

D Universality of loss

D.1 Derivation of the Lyapunov equation

Empirically, as discussed previously, for good initializations the dynamics converge near the realizable fixed point W^* (where the student recovers the teacher weights). Let us assume the second-layer weights are fixed, and write the weight difference $\delta\boldsymbol{W} = \boldsymbol{W} - \boldsymbol{W}^*$, where $\boldsymbol{W} = \text{vec}(W)$ for ease of notation. The online SGD update with a new sample is

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \mu\,\nabla_{\boldsymbol{W}}\mathcal{L}(W_t; \boldsymbol{x}_t, \boldsymbol{z}_t) \tag{69}$$

The sample gradient close to the optimum $\boldsymbol{W}^*$ can be linearized as

$$\begin{aligned} \nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}_t; \boldsymbol{x}_t, \boldsymbol{z}_t) &= \nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^*; \boldsymbol{x}_t, \boldsymbol{z}_t) + H_t\,\delta\boldsymbol{W}_t + o(\|\delta\boldsymbol{W}\|), \\ \text{with } H_t &:= \nabla^2_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^*; \boldsymbol{x}_t, \boldsymbol{z}_t). \end{aligned} \tag{70}$$

Substituting, we get

$$\delta \boldsymbol{W}_{t+1} = (I - \mu H_t)\, \delta \boldsymbol{W}_t - \mu\, \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}^*; \boldsymbol{x}_t, \boldsymbol{z}_t) + o(\|\delta \boldsymbol{W}\|)\mu. \tag{71}$$

We now take the outer product of the recursion and apply the average over $(\boldsymbol{x}_t,\, \boldsymbol{z}_t)$. Since $\boldsymbol{W}^*$ minimizes the population loss, $\mathbb{E}[\nabla_{\boldsymbol{W}} \mathcal{L}(W^*; \boldsymbol{x}_t, \boldsymbol{z}_t)]$ vanishes, while $\mathbb{E}[H_t] = H$ and $\mathbb{E}[\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}^*; \boldsymbol{x}_t, \boldsymbol{z}_t) \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}^*; \boldsymbol{x}_t, \boldsymbol{z}_t)^T] = \Sigma_{\text{grad}}$ define the population Hessian and gradient covariance at $\boldsymbol{W}^*$, and do not depend on t.

We work to leading order in the learning rate: we posit the self-consistent scaling $\mathbb{E}[\delta \boldsymbol{W} \delta \boldsymbol{W}^T] \sim \mathcal{O}(\mu)$, which is observed empirically and verified a posteriori. This is what allows us to drop the terms $o(\|\delta \boldsymbol{W}\|)$. The term $\mu^2\, \mathbb{E}[H_t\, \delta \boldsymbol{W}_t \delta \boldsymbol{W}_t^T\, H_t]$ is $\mathcal{O}(\mu^3)$ and therefore dropped, removing any dependence on Hessian fluctuations at this order. The linear cross-term $\delta \boldsymbol{W}_t\, \mathbb{E}[\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}^*)^T]$ vanishes identically, but the term $\mu^2\, \mathbb{E}[H_t\, \delta \boldsymbol{W}_t\, \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}^*)^T]$ does not, since H_t and $\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}^*)$ are correlated through a shared sample.

We get a deterministic recursion for the matrix $C_t := \mathbb{E}[\delta \boldsymbol{W}_t \delta \boldsymbol{W}_t^T]$:

$$C_{t+1} = C_t - \mu(HC_t + C_t H^T) + \mu^2\, \Sigma_{\text{grad}} + O(\mu^2)\, \mathbb{E}[\delta \boldsymbol{W}_t] + \mathcal{O}(\mu^3). \tag{72}$$

The term $\mathbb{E}[\delta \boldsymbol{W}_t]$ decays as $\mathbb{E}[\delta \boldsymbol{W}_{t+1}] = (I - \mu H)\, \mathbb{E}[\delta \boldsymbol{W}_t] \to 0$ under $\rho(I - \mu H) < 1$. We now impose stationarity, namely for $t \to \infty$, $C_{t+1} = C_t := C$. We get

$$HC + CH = \mu\, \Sigma_{\text{grad}}, \tag{73}$$

which is the corresponding Lyapunov equation of our process, where we used $H = H^T$. Notice that H and Σ_{grad} are fixed by the choice of network architecture and teacher weights.

D.2 Simplifying the Lyapunov equation

Computing them at $\boldsymbol{W}^*$, our equation simplifies further, and we get

$$\begin{aligned} H &= \alpha M_{\text{main}} + \beta M_{\text{aux}}. \\ \Sigma_{\text{grad}} &= \sigma^2 \big(\alpha^2 M_{\text{main}} + \beta^2 M_{\text{aux}}\big). \end{aligned} \tag{74}$$

where $M_i := \mathbb{E}[\nabla_{\boldsymbol{W}} y_i \nabla_{\boldsymbol{W}} y_i^T]|_{\boldsymbol{W}^*}$.

Both H and Σ_{grad} are built from the same two matrices. From these, it immediately follows that the Lyapunov equation can be written as

$$(\alpha M_{\text{main}} + \beta M_{\text{aux}})\, C + C\, (\alpha M_{\text{main}} + \beta M_{\text{aux}}) = \mu \sigma^2 (\alpha^2 M_{\text{main}} + \beta^2 M_{\text{aux}}). \tag{75}$$

The generalization errors can also be expanded around $\boldsymbol{W}^*$, obtaining

$$\epsilon_i = \mathbb{E}[\delta \boldsymbol{W}^T M_i \delta \boldsymbol{W}] = \operatorname{tr}(M_i C). \tag{76}$$

One can numerically solve Equation 75 for C and recover the theoretical generalization error, but C is still a $KN \times KN$ matrix.

Let us focus on M_{main}. Taking the explicit derivative, it can be written as

$$(M_{\text{main}})_{ik,jl} = \mathbb{E}\Big[v_i v_j\, g'(h_i) g'(h_j)\, \frac{x_k x_l}{N}\Big]. \tag{77}$$

At large N the input points decorrelates from the activations, $\mathbb{E}[x_k x_l / N] \to \delta_{kl}/N$, so

$$\begin{aligned} M_{\text{main}} &= \frac{1}{N}\, P_{\text{main}} \otimes I_N, \\ (P_{\text{main}})_{ij} &= v_i v_j\, \mathbb{E}[g'(h_i) g'(h_j)] = v_i v_j\, J_2(ij). \end{aligned} \tag{78}$$

In general, because $M_j = P_j \otimes I_N/N$, the Lyapunov equation inherits the Kronecker structure: $C = \tilde{C} \otimes I_N$ and it reduces to the $K \times K$ equation

$$(\alpha P_{\text{main}} + \beta P_{\text{aux}})\tilde{C} + \tilde{C}(\alpha P_{\text{main}} + \beta P_{\text{aux}}) = \mu\sigma^2(\alpha^2 P_{\text{main}} + \beta^2 P_{\text{aux}}). \tag{79}$$

The generalization error is then

$$\epsilon_i = \text{tr}(P_i \tilde{C}). \tag{80}$$

D.3 A CONSIDERATION ON THE STRUCTURE OF P

Now evaluate J_2 at the fixed point W^*. There, $Q = T = I$ (the student aligns with the teacher), so the h_i are $i.i.d.$ standard normal and

$$J_2(ij) = \mathbb{E}[g'(h_i)g'(h_j)] = \begin{cases} m^2 & i \neq j \\ m^2 + \kappa & i = j \end{cases} \tag{81}$$

where $\kappa := \text{Var}[g'(h)]$ and $m = \mathbb{E}[g'(h)]$. We can write $J_2 = m^2 \mathbf{1}\mathbf{1}^T + \kappa I$, and

$$P_{\text{main}} = m^2\, \boldsymbol{v}\boldsymbol{v}^T + \kappa\, D_v\,, \qquad D_v = \text{diag}(v_i^2), \tag{82}$$

with P_{aux} identical under $\boldsymbol{v} \to \boldsymbol{u}$. The linear model has $\kappa = 0$, from which P is rank-one, and it reduces to a 2×2 problem in $\text{span}(\boldsymbol{v}, \boldsymbol{u})$.

D.4 A DISCUSSION OF THE OPTIMAL BETA

In general, the matrices P do not commute, and there is no mode decomposition, but this analysis shows where β^* lies in this simplified setting. If P_{main} and P_{aux} commute, they are jointly diagonalizable. The equation decouples completely into scalars,

$$\begin{aligned} &2(\alpha a_i + \beta b_i)\,\tilde{C}_{ii} = \mu\sigma^2(\alpha^2 a_i + \beta^2 b_i), \\ &\tilde{C}_{ii} = \frac{\mu\sigma^2}{2}\frac{\alpha^2 a_i + \beta^2 b_i}{\alpha a_i + \beta b_i}, \end{aligned} \tag{83}$$

where a_i and b_i are the eigenvalues of the two matrices P.

With $\alpha = 1$ and $\epsilon_{\text{main}} = \text{tr}(P_1\tilde{C}) = \sum_i a_i \tilde{C}_{ii}$, we get:

$$\frac{\epsilon_{\text{main}}}{\epsilon_0} = \frac{\sum_i a_i \dfrac{a_i + \beta^2 b_i}{a_i + \beta b_i}}{\sum_i a_i} \tag{84}$$

with $\epsilon_0 = \mu\sigma^2 \sum_i a_i/2 = \mu\sigma^2\, \text{tr}\, P_{\text{main}}/2$, as discussed in Observation 2.

Define per-mode $f_i(\beta) = (a_i + \beta^2 b_i)/(a_i + \beta b_i)$ and $r_i = b_i/a_i$. If there is only one mode, then to compute the optimum one can impose $f_i' = 0 \Rightarrow r_i\beta^2 + 2\beta - 1 = 0$, so

$$\beta^\star(r) = \frac{\sqrt{1+r} - 1}{r} \in \big(0, \frac{1}{2}\big), \tag{85}$$

being $\sqrt{2} - 1$ when $r = 1$. The collective optimum solves $\sum_i a_i f_i'(\beta) = 0$. We get $\beta^* < 1/2$, which aligns with our empirical observation in Figure 4.

For the linear model, the two rank-one matrices P generally do not commute, but the problem is exactly solvable, and one can obtain the same results as in Appendix B.

D.5 Effects of higher-order terms

In Figure S2, we show the impact of higher-order terms on the theory developed above. We choose a set of parameters, namely the activation, $\boldsymbol{v}$, $\boldsymbol{u}$, β, and σ. We compare the theoretical improvement ϵ_{main} predicted by the theory to the one obtained by numerically integrating the ODEs for different choices of μ. As we see, they agree for small μ, while for learning rates approaching $\mu = 1$ the higher-order effects modify the results.

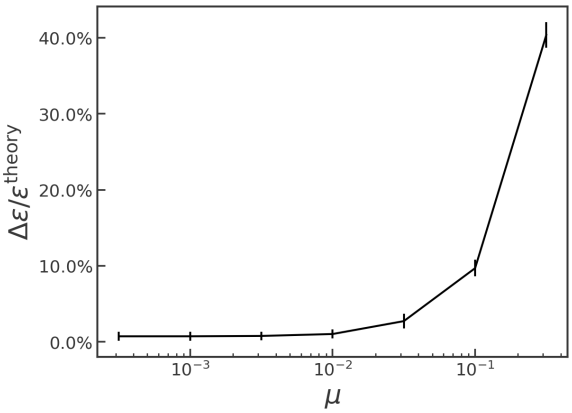


Figure S2: **Higher-order effects in μ.** Relative error $|\epsilon^{\text{ode}} - \epsilon^{\text{theory}}|/\epsilon^{\text{theory}} = \Delta\epsilon/\epsilon^{\text{theory}}$ for different μ, with ϵ^{theory} obtained by numerically solving the corresponding Lyapunov equation, and ϵ^{ode} by integrating the ODEs. Average over 10 random choices of $\boldsymbol{u}$, with erf activation, $\alpha = 1$, $\beta = 0.5$, $\sigma = 0.1$, $K = 4$, $\boldsymbol{v} = [1, 1, 1, 1]$, and correlation between tasks $\rho = 0.6$.

D.6 Convergence of different initializations

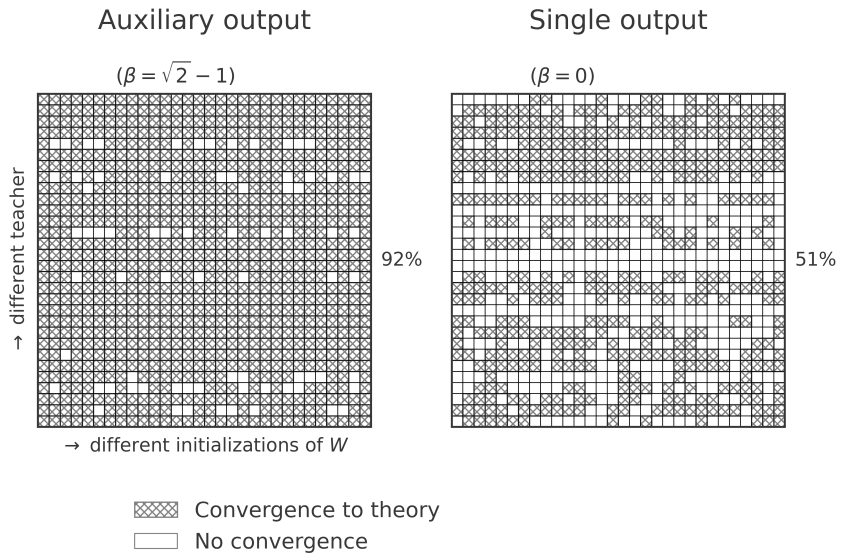


Figure S3: **Convergence of different initializations.** Comparison of how often ODE integration of different initializations of W, W^*, $\boldsymbol{v}$, $\boldsymbol{u}$ converges to the theoretical ϵ^{theory}, when $\beta = \sqrt{2} - 1$ and when $\beta = 0$. We sample entries independently from a standard normal distribution. Every row is the same teacher initialization of $\boldsymbol{v}$, $\boldsymbol{u}$, and W^*; every column is the same initialization of W. We plot 30 random initializations of W for each of 30 teachers, with erf activation, $\alpha = 1$, $\mu = 0.01$, $\sigma = 0.1$, $K = 4$. Convergence when the relative error is less than 5%.

		$\beta = 0$	
		converges	does not converge
$\beta = \sqrt{2} - 1$	converges	451	374
	does not converge	6	69

Table S1: **Convergence of different initializations.** Paired counts for the runs in Figure S3. McNemar exact test, $p < 10^{-10}$.

In Section 3.2, we discussed how not all initializations of W converge to a neighbourhood of W^* at the end of the dynamics. In this section, we sketch a preliminary exploration of this phenomenon. We sample W^*_{ij}, W_{ij}, v_i, $u_i \sim \mathcal{N}(0, 1)$, and check how often the ODE dynamics error matches the theoretical generalization obtained by numerically solving the corresponding Lyapunov equation.

As seen in Figure S3, more initializations converge to the global minimum W^* when $\beta = \sqrt{2} - 1$ than when there is no auxiliary output, namely when $\beta = 0$. Table S1 shows the counts. Applying McNemar's exact test, we reject the null hypothesis that the marginal convergence probabilities with the two β values are the same, with $p < 10^{-10}$. This suggests that auxiliary learning can help escape bad minima in the teacher-student setup.

The choice of $\beta = \sqrt{2} - 1$ for the auxiliary output is independent of $\boldsymbol{v}$, $\boldsymbol{u}$ and the related optimal β^*. This analysis could be expanded to more β values, but we do not expect major differences.

E A DIFFERENT SCALING OF THE MAIN TASK

In this section, we study an alternative weighting of the main task, namely setting $\alpha = 1 - \beta$. Although this may seem similar to choosing $\alpha = 1$, because online SGD results are intrinsically dynamical, this choice leads to completely different phenomenology.

Equation 68 shows that, for perfectly correlated outputs, the effective learning rate $\mu_{\text{eff}} = \mu(\alpha + \beta)$. From our choice of α then it follows that $\mu_{\text{eff}} = \mu$. But this choice of α implies that, when the tasks are not perfectly correlated, changing β also diminishes the learning rate for the main task. Thus, larger β values are beneficial because they reduce the learning rate on the main task. As we will see, β^* is always at least $1/2$ with this weighting (and, unlike with $\alpha = 1$, β has to be smaller than 1). In practice, for the linear case, from Equation 8, the generalization error on the main task is

$$\epsilon_{\text{main}}\Big|_{\alpha=1-\beta} = \epsilon_0\bigg(1 - \beta - \rho^2\beta(1 - 2\beta)\bigg), \tag{86}$$

minimized by

$$\beta^* = \frac{1 + \rho^2}{4\rho^2}. \tag{87}$$

Notice that if $\rho = 1$, the main and auxiliary tasks are the same, so $\beta^* = 1/2$: both tasks should have equal weights in the training loss, but generalization still improves because the noise on each task is independent. When $\rho = 0$, then the two tasks are completely decoupled, and rescaling α is effectively just rescaling the learning rate μ. This follows from Equation 44 of Appendix B when setting $\rho = 0$. Since we must have $\beta < 1$ (otherwise α would be zero or negative), from Equation 87 we see that for all $\rho < 1/\sqrt{3}$ the optimal training setup is the limit $\beta \to 1$. We show results for the linear model in Figure S4A.

When moving from linear to nonlinear models, namely with erf and ReLU activations in Figure S4B–C, we can make some empirical observations. Once again, the behavior does not depend solely on the correlation ρ between $\boldsymbol{v}$ and $\boldsymbol{u}$. We can make a few empirical observations:

Observation 1. *When $\alpha = \beta = 1/2$, then $\epsilon_{main} = \epsilon_{aux} = \epsilon_0/2$ for each ρ.*

Observation 2. $\alpha\epsilon_{\text{main}} + \beta\epsilon_{\text{aux}} = \epsilon_0(\alpha^2 + \beta^2)$.

Observation 3. *The optimal weight* $1/2 \leq \beta^* < 1$.

These observations match those made with $\alpha = 1$ and have a corresponding theoretical explanation.

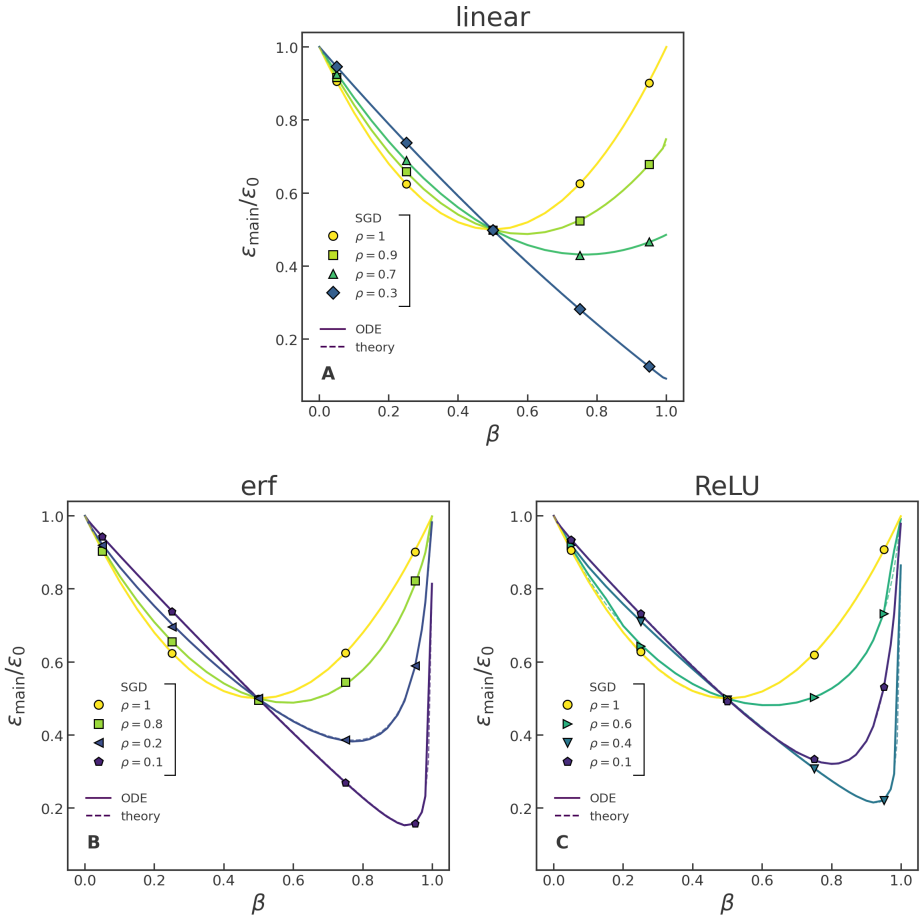


Figure S4: **Performance of models with $\alpha = 1 - \beta$.** Relative error $\epsilon_{\text{main}}/\epsilon_0$ for a model with A) linear, B) erf activation, and C) ReLU activation, and different ρ values trained with online gradient descent, $\alpha = 1 - \beta$, and different β values. The dashed lines show the theoretical solutions, the continuous lines show the results of numerically solving the mean-field ODEs, the dots show simulation results for models trained with online SGD, averaged over 10000 epochs after convergence across 5 simulations for linear and erf and 10 for ReLU, with the vertical bars denoting the corresponding confidence intervals. $\mu = 0.01$, $\sigma = 0.1$, $N = 784$, $K = 2$ for ReLU and $K = 4$ for linear and erf.