Editors' Suggestion

Statistical Mechanics of Transfer Learning in Fully Connected Networks in the Proportional Limit

Alessandro Ingrosso,[1,*] Rosalba Pacelli,[2] Pietro Rotondo,[3] and Federica Gerace[4,†]
[1]Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6525, Nijmegen, The Netherlands
[2]I.N.F.N., sezione di Padova, Via Marzolo 8, 35131, Padova, Italy
[3]Dipartimento di Scienze Matematiche, Fisiche e Informatiche, Universitá degli Studi di Parma, Parco Area delle Scienze, 7/A 43124 Parma, Italy
[4]Dipartimento di Matematica, Universitá di Bologna, Piazza di Porta San Donato 5, 40126, Bologna (BO), Italy

(Received 23 July 2024; accepted 7 March 2025; published 30 April 2025)

Transfer learning (TL) is a well-established machine learning technique to boost the generalization performance on a specific (target) task using information gained from a related (source) task, and it crucially depends on the ability of a network to learn useful features. Leveraging recent analytical progress in the proportional regime of deep learning theory (i.e., the limit where the size of the training set P and the size of the hidden layers N are taken to infinity keeping their ratio $\alpha = P/N$ finite), in this Letter we develop a novel single-instance Franz-Parisi formalism that yields an effective theory for TL in fully connected neural networks. Unlike the (lazy-training) infinite-width limit, where TL is ineffective, we demonstrate that in the proportional limit TL occurs due to a renormalized source-target kernel that quantifies their relatedness and determines whether TL is beneficial for generalization.

DOI: 10.1103/PhysRevLett.134.177301

Introduction—Modern deep learning relies on foundation models that are pretrained on tasks closely related to the one of interest but much richer in training examples. In this way, the generalization performance of a neural network trained on a data-scarce "target task" can consistently improve by leveraging the knowledge that the pretrained model has previously acquired on a close but data-abundant "source task." This "transfer learning" (TL) practice has been amply demonstrated to enhance the generalization performance of deep learning models, especially in those settings where data is scarce or labeling is demanding [1,2].

Despite being among the dominating paradigms in deep learning applications, TL remains poorly understood from a theoretical perspective, with several fundamental questions still open. For instance, (i) how does the source-target similarity affect TL efficiency? (ii) How does the width of the transferred layers impact generalization performance?

Most theoretical results in this direction hold for a parallel form of TL in the framework of classical learning theory, and rely on proofs of worst-case bounds, based on the Vapnik-Chervonenkis dimension [3], covering number, stability, and Rademacher complexity [4] (see [5–7] for review). A recent line of research has approached TL using statistical mechanics [8–11]. However its applicability is limited, since its focus is on linear neural networks (NNs) and source-target data models are overly simplistic. In [12], the authors went one step further by proposing a theoretical framework to study TL in one-hidden layer (1HL) and nonlinear networks. Here, pretraining is purely numerical, the interaction between the source and target is encoded implicitly in the empirical covariance of the hidden units, and the first layer weights are always kept fixed to the source configurations.

The analytically tractable lazy-training infinite-width limit [13–17], one of the recent milestones in deep learning theory, is also not a viable option to theoretically investigate pretraining and transfer stages: since the statistics of the weights remains unchanged during training, no features can be transferred from the source to the target task [18]. One possibility to overcome this issue is to consider the feature-learning phase of infinite-width networks [19–25], where TL is still possible [26]. To investigate more realistic settings than the infinite-width limit, one could analyze TL in the recently explored proportional regime of deep NNs, formally defined as the limit where both the size of the training set P and the width of the hidden layers N_ℓ scale to infinity while keeping the ratios $\alpha_\ell = P/N_\ell$ fixed. This limit was first studied in linear networks [27–30] and then extended to nonlinear models [31–33]. One major advantage of such a setting is the possibility of corroborating its analytical predictions with the outcome of Bayesian learning experiments in finite-width networks, as recently done for fully connected 1HL models [34].

In this Letter, we leverage the results established in the proportional limit, combining this approach with a novel *single-instance Franz-Parisi* formalism [35] to investigate

*Contact author: alessandro.ingrosso@donders.ru.nl
†Contact author: federica.gerace@unibo.it

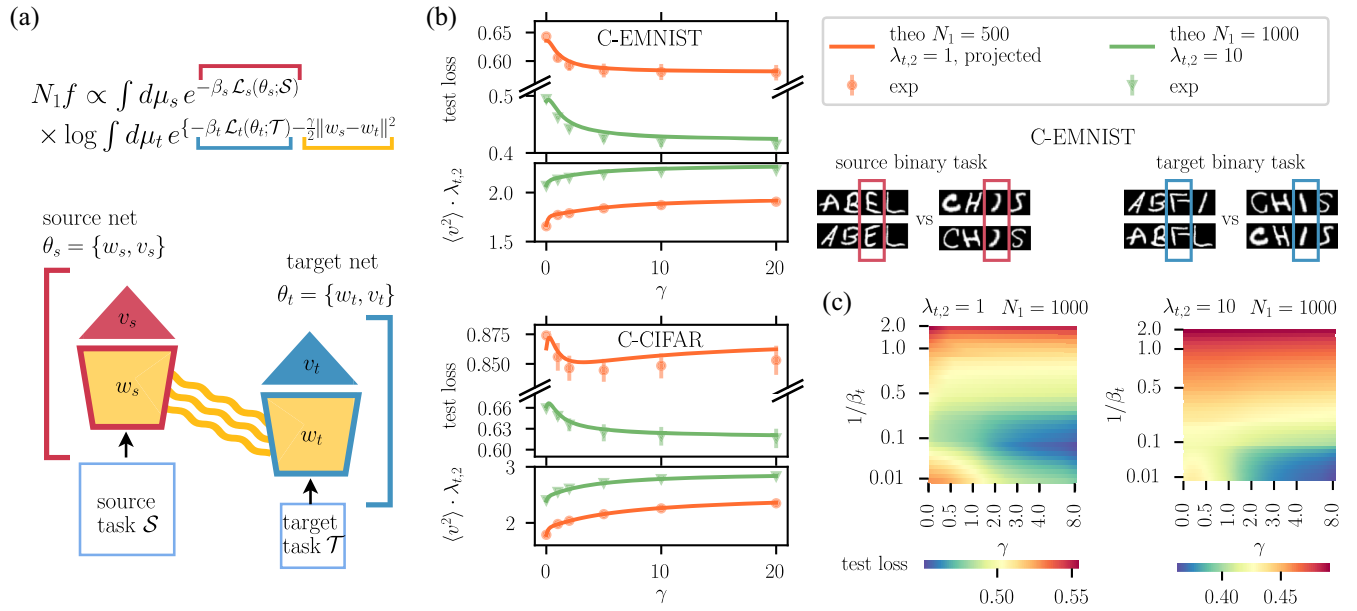

FIG. 1. TL in 1HL networks through the lens of a single-instance Franz-Parisi framework. (a) Sketch of TL with 1HL networks. The first-layer weights of the source (left) and target (right) networks are coupled, while the readout layers are optimized independently. The coupling strength γ is described by wavy yellow lines. (b) Learning curves of target 1HL networks. Experimental test loss and rescaled squared norm of the last layer are shown as a function of the coupling γ for different sizes N_1 and second-layer regularization $\lambda_{t,2}$. Markers are shown in comparison to solid lines, which represent theoretical predictions from the theory. The networks are trained on $P_s = 800$, $P_t = 100$ examples from C-EMNIST and C-CIFAR tasks [see Supplemental Material (SM) [38] for more details on the tasks]. The error bars denote one standard deviation from the mean, based on $k = 5$ independent extraction of weight configurations from the source posterior. The curves at $N_1 = 1000$ were obtained with source networks pretrained on images of $N_0 = 784$ pixels. The curves with $N_1 = 500$ refer to images projected in a lower-dimensional space of dimension $D = 300$, as described in SM. (c) The effect of temperature on transfer and fine-tuning with C-EMNIST. Predicted test loss of target network as a function of γ and $1/\beta_t$.

TL effectiveness in Bayesian NNs. In particular, we argue that the posterior over the weights of the source task plays the role of the quenched disorder in spin-glass theory [36], which can be thus integrated out using the well-known replica method. This leads to an explicit formula for the free energy in the proportional limit, describing the learning scenario of a neural network trained on the target task while coupled to a quenched copy, pretrained on the source task. The effect of the coupling is highlighted by the emergence of a renormalized source-target kernel, describing how the source-target similarity affects TL efficacy.

Single-instance Franz-Parisi formalism for TL—In the standard TL pipeline, a neural network is trained on a P_t-dimensional target set $\mathcal{T} = (X_t, y_t) \in (\mathbb{R}^{N_0} \times \mathbb{R})^{P_t}$ while keeping some of its layers frozen to the ones transferred from the P_s-dimensional source set $\mathcal{S} = (X_s, y_s) \in (\mathbb{R}^{N_0} \times \mathbb{R})^{P_s}$. Deep learning practitioners can later add a fine-tuning stage, where the transferred layers are unfrozen and the whole network is trained on the target set using a smaller learning rate.

To rationalize the effectiveness of TL in the proportional limit of fully connected networks, we introduce a novel approach based on statistical mechanics of learning [37]. Specifically, we consider a setting involving a 1HL neural network ϕ_t whose first-layer weights adapt to the target task while coupled to those learned by another network ϕ_s on the source task. A sketch of the learning setting is shown in Fig. 1(a). In the framework of statistical mechanics, this learning paradigm can be effectively described by the following free-energy density:

$$f = \frac{1}{N_1} \mathbb{E}_{\theta_s} \left[\log \int d\mu(\theta_t) e^{-\beta_t \mathcal{L}_t(\theta_t; \mathcal{T}) - \frac{\gamma}{2} \|w_s - w_t\|_2^2} \right], \quad (1)$$

where N_1 is the number of hidden units, $\theta_{s/t}$ are the collections of the first and second-layer weights $w_{s/t}$ and $v_{s/t}$ of $\phi_{s/t}$, respectively, $\mu(\theta_t) \propto \exp\{-\lambda_t \|\theta_t\|_2^2\}$ is the Gaussian prior over the target weights, describing an L_2 regularization on the first and second-layer weights with strength controlled by the parameter $\lambda_t = (\lambda_{t,1}, \lambda_{t,2})$, $\mathcal{L}_t$ is the target training loss, and β_t is the target inverse temperature. The limit $\beta_t \to \infty$ can eventually enforce perfect interpolation of the target training set. The coupling between the source and the target network is controlled by a parameter γ. This allows for continuous interpolation between two regimes: one where the network is trained from scratch on the target set with no knowledge transfer from the source task ($\gamma = 0$), and another one where the first-layer weights of the target network are kept frozen to the source weights, while the second-layer weights adapt to

the target set ($\gamma \to \infty$). The intermediate values of γ describe the fine-tuning stage, where the first-layer weights w_t are first initialized to the source configuration w_s and then trained on the target set, together with the second-layer ones v_t. The integral over θ_t in Eq. (1) defines the partition function Z of the target network when coupled to the source one.

The expectation over the source configurations θ_s ensures that we describe the typical TL behavior. To guarantee that the source configurations effectively solve the source task, we take this expectation over the *posterior* distribution of the source weights. This corresponds to a Boltzmann-Gibbs measure whose partition function only involves the source task,

$$Z_s(\beta_s) = \int d\mu(\theta_s) e^{-\beta_s \mathcal{L}_s(\theta_s, \mathcal{S})}, \tag{2}$$

where $\mu(\theta_s) \propto \exp\{-\lambda_s \|\theta_s\|_2^2\}$ is the Gaussian prior over the source weights, with $\lambda_s = (\lambda_{s,1}, \lambda_{s,2})$, and β_s is the source inverse temperature. It is important to emphasize that, since the source and target training are not performed simultaneously, the expectation over the source configurations is quenched. This is crucial to ensure that the source posterior is not affected by the source-target coupling: in a TL pipeline, the source network is indeed trained on the source task without relying on any information on the target data.

As anticipated, the free-energy density in Eq. (1) closely resembles the Franz-Parisi potential, originally introduced to analyze metastable states in spin-glass systems [35] and later used to characterize the properties of energy landscapes in machine learning [40–42] and constraint satisfaction problems [43], knowledge distillation [44], curriculum learning [45], and continual learning [11] in single-layer neural networks. At variance with previous literature [46,47], we refrain from averaging the free-energy density over the input data distribution, in the same spirit of Refs. [24,27,31–33] (note that a similar approach has been put forward to investigate Markov proximal learning [48] in an attempt to connect the neural tangent and the neural network Gaussian process kernels).

For this reason, we name our new theoretical framework *single-instance* Franz-Parisi. The quenched expectation over the source weights is tackled using the replica method, while the integral over the replicated target configurations in Eq. (1) is performed via the standard kernel renormalization approach, which is exact for deep linear networks [27,33], and can be justified using a Gaussian equivalence for nonlinear activation function [49–51]. In the following, we provide a sketch of the derivation [full details can be found in SM [38]], which is valid for source and target quadratic loss function (mean squared error).

Free energy in the proportional limit—To evaluate the quenched free energy in Eq. (1) in the proportional thermodynamic limit described in the introduction, we make use of the replica trick

$$f = \frac{1}{N_1} \lim_{n \to 0} \partial_n \mathbb{E}_{\theta_s}[Z^n]. \tag{3}$$

The calculation yields a compact expression for the replicated partition function in terms of an effective finite-n action S_n, which, for the sake of simplicity and lighter notation, we here report in the case $\beta_t = \beta_s = \beta$ and $\lambda_{s,2} = \lambda_{t,2} = \lambda$,

$$\mathbb{E}_{\theta_s}[Z^n] \sim \exp\{N_1 S_n(\mathcal{Q}, \bar{\mathcal{Q}}; n)\} \tag{4}$$

$$S_n = \mathrm{Tr}(\lambda \mathcal{Q} \bar{\mathcal{Q}}) - \log\det(\mathbb{1} + \bar{\mathcal{Q}}) - \frac{1}{N_1} \log\det(\mathbb{1} + \beta \mathcal{K}) - \frac{\beta}{N_1} y^T (\mathbb{1} + \beta \mathcal{K})^{-1} y, \tag{5}$$

where $\mathcal{Q}$ and its conjugate $\bar{\mathcal{Q}}$ are $(n+1) \times (n+1)$ order parameters matrices (see SM for the most general expression). At this level, $\mathcal{K}$ is a $(P_s + nP_t) \times (P_s + nP_t)$ replicated renormalized kernel matrix,

$$\mathcal{K}^{ab}_{\mu\nu} = \mathcal{Q}^{ab} K^{ab}_{\mu\nu}, \qquad K^{ab}_{\mu\nu} = \langle \sigma(h^a_\mu) \sigma(h^b_\nu) \rangle, \tag{6}$$

where σ is the activation function and the expectations are taken with respect to an effective distribution of the first-layer hidden representations h. The zeroth replica identifies the source network, and correspondingly y is the concatenation $y = (y_s, y_t, \ldots, y_t)$ of the output vectors. The last two terms in the effective action are $\mathcal{O}(1)$ in the proportional regime $P_s = \alpha_s N_1$ and $P_t = \alpha_t N_1$, with $\alpha_{s/t} = \mathcal{O}(1)$. Given the symmetric source-target coupling among replicas, it is reasonable to assume a replica symmetric ansatz for the matrix $\mathcal{Q}$. This implies that there are only four distinct order parameters $\mathcal{Q} = \{Q_s, Q_t, Q_{st}, Q_{tt}\}$, along with their conjugates $\bar{\mathcal{Q}}$, mediating the interaction between the source and target posterior distributions via their coupling with four distinct kernels K_s, K_t, K_{st}, K_{tt}. Of these, the crucial one is the effective source-target kernel K_{st}, measuring the similarity between the source and target hidden representations, as a direct function of the source-target input covariance $C_{st} = X_s X_t^T / N_0$. K_{st} is thus the blueprint of TL efficacy in the proportional limit.

The $\mathcal{O}(1)$ terms of S_n rebuild the effective action of the source task alone [31], while the $\mathcal{O}(n)$ terms in S_n in the $n \to 0$ yield the genuine source-target action, incorporating the effect of transfer. Through appropriate derivatives of such "transfer action," we can compute relevant observables in the equilibrium ensemble, e.g., the test loss or the norm of the last-layer weights (see SM).

TL on benchmark tasks—Figure 1(b) illustrates the good agreement between theoretical predictions (solid lines) and numerical simulations (dots) for both the test loss (top) and the norm of the last-layer weights (bottom), as a function of

the coupling parameter γ, for two different values of N_1 and last-layer regularization strength $\lambda_{t,2}$. The two rows refer to TL scenarios with two different source-target pairs of binary tasks: C-EMNIST (top) and C-CIFAR (bottom), built from the well-known benchmark datasets EMNIST [52] and CIFAR10 [53] as in [12]. An example of C-EMNIST is shown in Fig. 1: to construct the source task we group some EMNIST letters in two macro classes; the target task is then generated by replacing one letter in each source macro class. C-CIFAR follows the same idea but with different classes. These cases are meant to describe settings where the source and target tasks differ due to geometric structure of the input data (see also End Matter).

As Fig. 1(b) shows, the effectiveness of TL and fine-tuning strongly depends on data structure and thus on the source-target relation, but also on the network size. For instance, while there exists an optimal γ minimizing the C-CIFAR test loss (bottom orange curve), the same does not occur with C-EMNIST (top orange curve), where the generalization performance steadily improves with γ, eventually reaching a plateau for $\gamma \to \infty$. At larger network sizes (green curves), the optimum in the C-CIFAR setup is attained at very large γ: fine-tuning is thus less beneficial than just freezing the features when converging toward the infinite-width limit.

In Fig. 1(c) we show the phase diagram in the $\beta_t^{-1} - \gamma$ plane for two different second-layer regularizations. As we can notice, carefully tuning the temperature can improve generalization. More interestingly, we see that both TL and stronger regularization help to sample optimal solutions at lower temperatures.

Fine-tuning and source-target similarity—The level of source-target similarity is a crucial aspect in TL settings. If two tasks are poorly related, it is not unusual to observe negative transfer effects [54]. To analyze these aspects, we use the correlated hidden manifold model (CHMM), a synthetic data model for pairs of source-target tasks [12]. In the CHMM, each N_0-dimensional source data point lies on a latent D_s-dimensional manifold and is built as a nonlinear combination of some generative features $F_s \in \mathbb{R}^{N_0 \times D_s}$. The labels are instead provided by a teacher network $\theta_s \in \mathbb{R}^{D_s}$, directly acting on the data manifold [55]. To mimic realistic TL scenarios, the target task is then constructed from the source by perturbing one or more among the generative features, the teacher vector and the dimensionality of the data manifold. For instance, the source-target datasets may differ because of structure and style, as in the example in Fig. 1. In the model, this is described by the fraction ρ of rows in F_s replaced by new ones in the target task. The detailed description of the CHMM can be found in End Matter.

Figure 2(b) shows the test loss as a function of γ when training on a CHMM source-target pair with two distinct values of ρ. When a larger number of features are common to both tasks (green curve), the test loss decreases with the strength of the source-target coupling, showing that it is always convenient to constrain the first-layer weights to the source ones ($\gamma \to \infty$) rather than training from scratch ($\gamma = 0$) or slightly fine-tuning the network on the target set ($\gamma \to 0$). Instead, when the two tasks share only half of the features (orange curve), one clearly sees that fine-tuning the network on the target set at $\gamma \simeq 1$ leads to better generalization than freezing the first-layer weights or training from scratch. Figure 2(a) shows the phase diagram in the $\rho - \gamma$ plane, which highlights a continuous interpolation between these two regimes.

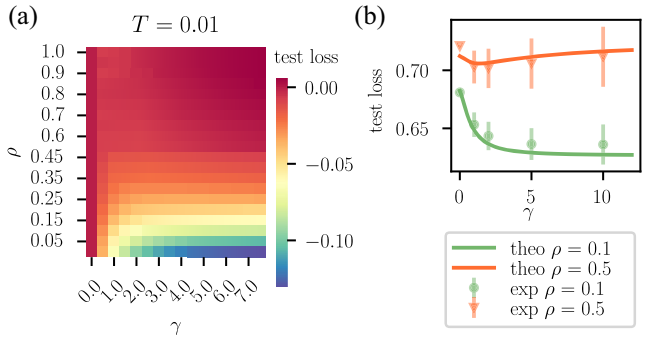

FIG. 2. Effectiveness of fine-tuning depends on source-target similarity. (a) Phase diagram of fine-tuning on the CHMM. The heat map shows the difference between the test loss at a given γ minus the same at $\gamma = 0$ for all values of ρ in the $\gamma - \rho$ plane. (b) Test loss of 1HL target network as a function of γ for $\rho = 0.1$, 0.5, corresponding to two horizontal slices of the phase diagram in panel (a). The source task has $P_s = 800$ examples, with $N_0 = 500$ and source and target latent dimensions $D_s = D_t = 300$. The $P_t = 200$ target examples share only a fraction $1 - \rho$ of features with the source task. Error bars represent one standard deviation from the average across $k = 5$ different posterior source weight extractions.

TL at proportional width vs infinite width—Figure 3(a) shows a comparison between the single-instance Franz-Parisi in the proportional regime and the infinite-width one. Specifically, we show test losses (left) and last-layer weight norms (right) of target networks for different values of N_1. Consistently, the theory reproduces the infinite-width limit behavior in the regime where $N_1 \gg P_s, P_t$. Already at $N_1 \sim 10 P_s$, there is less than 2% gain in terms of generalization performance between uncoupled model to the best (completely transferred) model, and the last-layer weights have the same average norm as before the training ($1/\lambda_{t,2}$). This behavior is compatible with the lazy-training infinite-width regime, where the statistics of the weights of the source and target networks does not change during training, forbidding TL from being beneficial. More interestingly, already at $\gamma \sim 1$, smaller architectures outperform the infinite-width performance, which is instead the best achievable performance when no knowledge transfer occurs ($\gamma = 0$). The fact that finite coupled networks outperform their best uncoupled predictor signals that

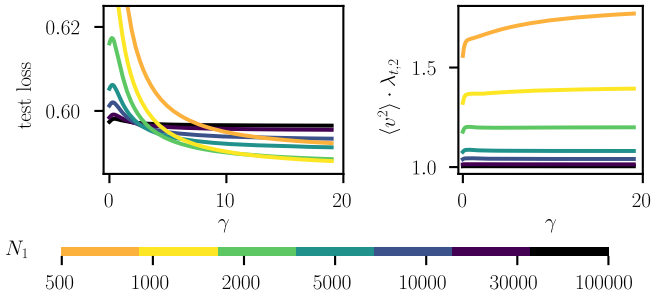

FIG. 3. TL is ineffective in the lazy-training infinite-width limit. Target test loss (left) and norm of last-layer weights (right) are shown as a function of γ for different values of N_1. The tasks are C-EMNIST with $P_s = 800$, $P_t = 100$. As the infinite-width limit is approached ($N_1 \gg P_s, P_t$), TL becomes ineffective, with no gain in terms of performance between the uncoupled model ($\gamma = 0$) and the TL model ($\gamma > 0$). Coherently, as N_1 grows, the norm converges to its prior value $1/\lambda_{t,1}$, corresponding to the infinite-width lazy-training solution $\mathcal{Q} = \mathbb{1}$ ($Q_s = Q_t = 1$, $Q_{st} = Q_{tt} = 0$).

the performance improvement is genuinely due to transfer and is not an artifact of effective regularization.

Discussion and conclusions—In this Letter, we introduced a new theoretical approach to study TL in the proportional limit that leverages techniques used in the theory of spin glasses and kernel methods, and showcased it for 1HL fully connected networks. Our theory predicts the emergence of a source-target kernel K_{st}, whose renormalization in the proportional limit captures the improvement on generalization performance of the target network, depending on source-target similarity.

A transfer pipeline can also be formulated in convex problems where there is a strong imbalance between available data for two different learning tasks: linear regression is arguably the simplest example of such a case where analytical expressions for TL can be derived without resorting to the replica method (see Sec. III of SM and Fig. S1 for a simple example in a teacher-student setting). The generalization to deep nonlinear networks and more complex layer structures, where kernel renormalization has been shown to depend on a local mechanism [32], are interesting avenues for future work we are currently pursuing. In particular, we sketch in the End Matter a tentative derivation for multilayer architectures, which we conjecture to be exact in the case of deep linear networks, and show in the SM how kernels would change for a convolutional layer.

Finally, it is reasonable to expect that the proposed effective theory breaks down in the large P regime [46,56,57], i.e., when the number of training patterns is proportional to the total number of parameters of the network, e.g., $P \propto N_1 N_0$ in a one-hidden-layer network. Addressing this regime can be considered a major challenge for scientists working in deep learning theory.

Acknowledgments—F. G., R. P. and P. R. are supported by NEXTGEN- ERATIONEU (NGEU). F. G. is partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan. P. R. is funded by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), project MNESYS (PE0000006) "A Multiscale integrated approach to the study of the nervous system in health and disease" (DN. 1553 11.10.2022). R. P. is funded by MUR project PRIN 2022HSKLK9.

[1] M. Bernhardt, D. C. Castro, R. Tanno, A. Schwaighofer, K. C. Tezcan, M. Monteiro, S. Bannur, M. P. Lungren, A. Nori, B. Glocker *et al.*, Active label cleaning for improved dataset quality under resource constraints, Nat. Commun. **13**, 1161 (2022).
[2] D. Mincu and S. Roy, Developing robust benchmarks for driving forward AI innovation in healthcare, Nat. Mach. Intell. **4**, 916 (2022).
[3] V. Vapnik, *The Nature of Statistical Learning Theory* (Springer Science & Business Media, New York, NY, 2013).
[4] P. L. Bartlett and S. Mendelson, Rademacher and Gaussian complexities: Risk bounds and structural results, J. Mach. Learn. Res. **3**, 463 (2002).
[5] L. Zhang and X. Gao, Transfer adaptation learning: A decade survey, IEEE Trans. Neural Networks Learn. Syst. **35**, 23 (2022).
[6] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, Generalizing from a few examples: A survey on few-shot learning, ACM Comput. Surv. **53**, 1 (2020).
[7] Y. Zhang and Q. Yang, A survey on multi-task learning, IEEE Trans. Knowledge Data Eng. **34**, 5586 (2021).
[8] A. K. Lampinen and S. Ganguli, An analytic theory of generalization dynamics and transfer learning in deep linear networks, arXiv:1809.10374.
[9] Y. Dar and R. G. Baraniuk, Double double descent: On generalization errors in transfer learning between linear regression tasks, SIAM J. Math. Data Sci. **4**, 1447 (2022).
[10] O. Dhifallah and Y. M. Lu, Phase transitions in transfer learning for high-dimensional perceptrons, Entropy **23**, 400 (2021).
[11] C. Li, Z. Huang, W. Zou, and H. Huang, Statistical mechanics of continual learning: Variational principle and mean-field potential, Phys. Rev. E **108**, 014309 (2023).
[12] F. Gerace, L. Saglietti, S. S. Mannelli, A. Saxe, and L. Zdeborová, Probing transfer learning with a model of synthetic correlated datasets, Mach. Learn. **3**, 015030 (2022).
[13] R. M. Neal, Priors for infinite networks, in *Bayesian Learning for Neural Networks* (Springer, New York, NY, 1996), pp. 29–53.
[14] J. Lee, J. Sohl-dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri, Deep neural networks as Gaussian processes, in *International Conference on Learning Representations* (2018).

[15] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), Vol. 31.

[16] A. G. de G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani, Gaussian process behaviour in wide deep neural networks, in *International Conference on Learning Representations* (2018), https://openreview.net/forum?id=H1-nGgWC-.

[17] B. Hanin, Random neural networks in the infinite width limit as Gaussian processes, Ann. Appl. Probab. **33**, 4798 (2023).

[18] G. Yang and J. E. Hu, Tensor programs IV: Feature learning in infinite-width neural networks, in *International Conference on Machine Learning* (2021).

[19] S. Mei, A. Montanari, and P.-M. Nguyen, A mean field view of the landscape of two-layer neural networks, Proc. Natl. Acad. Sci. U.S.A. **115**, E7665 (2018).

[20] B. Bordelon and C. Pehlevan, Self-consistent dynamical field theory of kernel evolution in wide neural networks, in *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., 2022), Vol. 35, pp. 32240–32256.

[21] G. Rotskoff and E. Vanden-Eijnden, Trainability and accuracy of artificial neural networks: An interacting particle system approach, Commun. Pure Appl. Math. **75**, 1889 (2022).

[22] J. Sirignano and K. Spiliopoulos, Mean field analysis of neural networks: A law of large numbers, SIAM J. Appl. Math. **80**, 725 (2020).

[23] L. Chizat and F. Bach, On the global convergence of gradient descent for over-parameterized models using optimal transport, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), Vol. 31.

[24] I. Seroussi, G. Naveh, and Z. Ringel, Separation of scales and a thermodynamic description of feature learning in some cnns, Nat. Commun. **14**, 908 (2023).

[25] G. Naveh and Z. Ringel, A self consistent theory of Gaussian processes captures feature learning effects in finite CNNs, in *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021), Vol. 34, pp. 21352–21364.

[26] G. Yang and E. J. Hu, Feature learning in infinite-width neural networks, arXiv:2011.14522.

[27] Q. Li and H. Sompolinsky, Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization, Phys. Rev. X **11**, 031059 (2021).

[28] B. Hanin and A. Zlokapa, Bayesian interpolation with deep linear networks, Proc. Natl. Acad. Sci. U.S.A. **120**, e2301345120 (2023).

[29] F. Bassetti, M. Gherardi, A. Ingrosso, M. Pastore, and P. Rotondo, Feature learning in finite-width Bayesian deep linear networks with multiple outputs and convolutional layers, arXiv:2406.03260.

[30] L. Tiberi, F. Mignacco, K. Irie, and H. Sompolinsky, Dissecting the interplay of attention paths in a statistical mechanics theory of transformers, arXiv:2405.15926.

[31] R. Pacelli, S. Ariosto, M. Pastore, F. Ginelli, M. Gherardi, and P. Rotondo, A statistical mechanics framework for Bayesian deep neural networks beyond the infinite-width limit, Nat. Mach. Intell. **5**, 1497 (2023).

[32] R. Aiudi, R. Pacelli, A. Vezzani, R. Burioni, and P. Rotondo, Local kernel renormalization as a mechanism for feature learning in overparametrized convolutional neural networks, arXiv:2307.11807.

[33] Q. Li and H. Sompolinsky, Globally gated deep linear networks, Adv. Neural Inf. Process. Syst. **35**, 34789 (2022).

[34] P. Baglioni, R. Pacelli, R. Aiudi, F. D. Renzo, A. Vezzani, R. Burioni, and P. Rotondo, Predictive power of a Bayesian effective action for fully-connected one hidden layer neural networks in the proportional limit, Phys. Rev. Lett. **133**, 027301 (2024).

[35] S. Franz and G. Parisi, Recipes for metastable states in spin glasses, J. Phys. I (France) **5**, 1401 (1995).

[36] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* (World Scientific Publishing Company, Singapore, 1987), Vol. 9.

[37] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, England, 2001).

[38] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.134.177301, which includes Refs. [31,32,39], for a detailed derivation of the analytical results and additional numerical experiments.

[39] C. Williams, Computing with infinite networks, in *Advances in Neural Information Processing Systems*, edited by M. Mozer, M. Jordan, and T. Petsche (MIT Press, Cambridge, MA, 1996), Vol. 9.

[40] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses, Phys. Rev. Lett. **115**, 128101 (2015).

[41] C. Baldassi, F. Gerace, C. Lucibello, L. Saglietti, and R. Zecchina, Learning may need only a few bits of synaptic precision, Phys. Rev. E **93**, 052313 (2016).

[42] G. Catania, A. Decelle, and B. Seoane, Copycat perceptron: Smashing barriers through collective learning, Phys. Rev. E **109**, 065313 (2024).

[43] M. C. Angelini and F. Ricci-Tersenghi, Limits and performances of algorithms based on simulated annealing in solving sparse hard inference problems, Phys. Rev. X **13**, 021011 (2023).

[44] L. Saglietti and L. Zdeborová, Solvable model for inheriting the regularization through knowledge distillation, in *Mathematical and Scientific Machine Learning* (PMLR, 2022), pp. 809–846.

[45] L. Saglietti, S. Mannelli, and A. Saxe, An analytical theory of curriculum learning in teacher-student networks, Adv. Neural Inf. Process. Syst. **35**, 21113 (2022).

[46] H. Cui, F. Krzakala, and L. Zdeborova, Bayes-optimal learning of deep random networks of extensive-width,

in *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 202, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (PMLR, 2023), pp. 6468–6521.
[47] F. Camilli, D. Tieplova, and J. Barbier, Fundamental limits of overparametrized shallow neural networks for supervised learning, arXiv:2307.05635.
[48] Y. Avidan, Q. Li, and H. Sompolinsky, Connecting NTK and NNGP: A unified theoretical framework for neural network learning dynamics in the kernel regime, arXiv:2309.04522 [Phys. Rev. E (to be published)].
[49] J.-M. Bardet and D. Surgailis, Moment bounds and central limit theorems for Gaussian subordinated arrays, J. Multivariate Anal. **114**, 457 (2013).
[50] I. Nourdin, G. Peccati, and M. Podolskij, Quantitative Breuer–Major theorems, Stoch. Proc. Appl. **121**, 793 (2011).
[51] P. Breuer and P. Major, Central limit theorems for non-linear functionals of Gaussian fields, J. Multivariate Anal. **13**, 425 (1983).
[52] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, Emnist: Extending mnist to handwritten letters, in *2017 international Joint Conference on Neural Networks (IJCNN)* (IEEE, New York, 2017), pp. 2921–2926.
[53] A. Krizhevsky, G. Hinton *et al.*, Learning multiple layers of features from tiny images (2009).
[54] F. Gerace, D. Doimo, S. S. Mannelli, L. Saglietti, and A. Laio, Optimal transfer protocol by incremental layer defrosting, arXiv:2303.01429.
[55] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová, Generalisation error in learning with random features and the hidden manifold model, in *International Conference on Machine Learning* (PMLR, 2020), pp. 3452–3462.
[56] Q. Li, B. Sorscher, and H. Sompolinsky, Representations and generalization in artificial and brain neural networks, Proc. Natl. Acad. Sci. U.S.A. **121**, e2311805121 (2024).
[57] F. Aguirre-López, S. Franz, and M. Pastore, Random features and polynomial rules, SciPost Phys. **18**, 039 (2025).

End Matter

Appendix A: Details of numerical experiments—In this section, we provide some details about the training algorithm and tasks used for numerical validation of our theoretical results. All experiments are performed on pairs of source-target architectures with one hidden layer and Erf (error function) activation. To ensure sampling from the posterior Gibbs ensemble of weights, which is essential to validate our theory, we train our networks using a discretized Langevin dynamics, similarly to what is done in [24,27,31,34]. At each training step t, the parameters $\theta = \{w, v\}$ are updated according to

$$\theta(t+1) = \theta(t) - \eta \nabla_\theta \tilde{\mathcal{L}}[\theta(t)] + \sqrt{2T\eta}\epsilon(t), \tag{A1}$$

where $T = 1/\beta$ is the temperature, η is the learning rate, $\epsilon(t)$ is a white Gaussian noise vector with entries drawn from a standard normal distribution, and the regularized loss function $\tilde{\mathcal{L}}$ comprises the prior rescaled by the temperature,

$$\tilde{\mathcal{L}}_{s/t} = \mathcal{L}_{s/t} + \frac{T\lambda_{s/t,1}}{2}\|w_{s/t}\|^2 + \frac{T\lambda_{s/t,2}}{2}\|v_{s/t}\|^2. \tag{A2}$$

The learning rate is fixed to $\eta = 10^{-3}$ throughout the experiments for both source and target networks.

Experimental setup: We use pairs of correlated source-task classification tasks. Two pairs involve real-world computer vision datasets (C-EMNIST and C-CIFAR), and one is a synthetic task (the CHMM). We train the source model on the source task first, then extract k equilibrium configurations of the source weights. For each of the k sets of features, we train a target network for different values of the parameter γ controlling the coupling to source weights, and average results over the k source configurations. In Fig. 1 of the main text $k = 5$; in Fig. 2 we used $k = 10$.

C-EMNIST and C-CIFAR: Similarly to [12], we build a binary source task by dividing a subset of the EMNIST letters into two distinct groups: letters $\{A, B, E, L\}$ for the first one and $\{C, H, J, S\}$ for the second one. We assign the label to each image according to the group membership. The target task is then built from the source task by replacing one letter per group (letter E with F and J with I). We call this pair of source-target tasks C-EMNIST (correlated EMNIST). C-CIFAR is constructed in a similar way from the CIFAR10 dataset: the source task includes the first eight classes, specifically {1,2,3,4} in the first group and {5,6,7,8} in the second one. The target task is obtained by replacing class 1 with 10 and class 5 with 9.

In all experiments, images from CIFAR10 and EMNIST are gray-scaled and down-sized. In Fig. 2 and Fig. 3 of the main text, the input size is set to $N_0 = 784$ pixels. In Fig. 1(b), the curves at $N_1 = 500$ have $N_0 = 784$, while those at $N_1 = 1000$ are obtained by preprocessing the input data points x with random features,

$$\hat{x} = \sigma\left(\frac{Fx}{N_0}\right), \tag{A3}$$

where $F \in \mathbb{R}^{D\times N_0}$ is the random feature matrix, whose entries are sampled independent identically distributed from a standard Gaussian. This effectively projects the input data points in a new space of dimension D. In the experiments in Fig. 1(b) we set $D = 400$.

CHMM: To analyze the extent to which the correlation between source-target tasks is essential for TL to be

beneficial, we use the correlated hidden manifold model (CHMM), a synthetic data model where the source-target similarity can be explicitly tuned via three different set of parameters meant to mimic different and realistic TL scenarios [12]. According to the CHMM, the source task is built using a hidden manifold model [55]. This model is based on the idea that real data do not span the entire input space uniformly, but rather are confined in a low-dimensional manifold. Along the lines of the hidden manifold model, in CHMM, each source input X_s^μ is built as a nonlinear combination of some generative features $F_s \in \mathbb{R}^{N_0 \times D_s}$ with some coefficients $c_s^\mu \in \mathbb{R}_s^D$,

$$X_s^\mu = f_x\left(\frac{F_s c_s^\mu}{\sqrt{D_s}}\right), \tag{A4}$$

where f_x is a nonlinear function acting pointwise, D_s is the dimension of the low-dimensional manifold or intrinsic dimension, and σ is a nonlinear function acting pointwise. In the experiments of the main manuscript, we sample both the generative features as well as the coefficients of the nonlinear combination from a normal Gaussian measure, that is $F_{ij}, c_s^\mu \sim \mathcal{N}(0,1)$. The labels y_s^μ are instead provided by a teacher network $\theta_s \in \mathbb{R}_s^D$, acting directly in the low-dimensional space,

$$y_s^\mu = f_y\left(\frac{c_s^\mu \cdot \theta_s}{\sqrt{D_s}}\right), \tag{A5}$$

where f_y is a nonlinear function acting pointwise. The target task is constructed from the source task by perturbing one or more among the generative features, the teacher vector, and the intrinsic dimension. In particular, we can distinguish among three different families of perturbations.

Feature perturbation and substitution. This type of perturbation is meant to describe those settings where the source and the target task differ for the input data structure. In the CHMM this is modeled through two distinct parameters: ρ, which determines the number of rows of F_s that are replaced in the target set, and η, which determines the amount of noise injected in each feature,

$$(\boldsymbol{F}_t)_{ij} = \begin{cases} (\tilde{\boldsymbol{F}})_{ij} & i \in [1, \rho D_s] \\ \eta(\boldsymbol{F}_s)_{ij} + \sqrt{1-\eta^2}(\tilde{\boldsymbol{F}})_{ij} & i \in [\rho D_s + 1, D_s] \end{cases}, \tag{A6}$$

with $(\tilde{\boldsymbol{F}})_{ij} \sim \mathcal{N}(0,1)$.

Teacher network perturbation. This type of perturbation accounts for those settings in which the source and the target task share the same data structure but they are labeled according to a different labeling rule. In the model, the misalignment between the source and the target labeling rule is described by the parameter q,

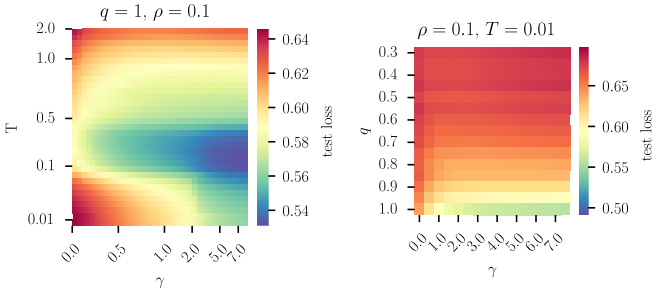

FIG. 4. Predicted target test loss for a NN trained on $P_t = 200$ examples from the CHMM, as a function of T and γ (left), and q and γ (right). Each point in the right plot has been obtained as the average over $k = 20$ realizations of the source perturbation $\rho = 1$.

$$(\boldsymbol{\theta}_t)_i = q(\boldsymbol{\theta}_s)_i + \sqrt{1-q^2}(\tilde{\boldsymbol{\theta}})_i, \tag{A7}$$

with $(\tilde{\boldsymbol{\theta}})_i \sim \mathcal{N}(0,1)$.

Feature addition or deletion. This perturbation describes the situation where the source and the target tasks live in data manifolds of different dimensionality. Increasing the dimensionality makes a task more complex. In the model, this situation is described by tuning the intrinsic dimensions of the two datasets as

$$(\boldsymbol{F}_t)_{ij} = \begin{cases} (\boldsymbol{F}_s)_{ij} & i \in [1, \min(L_s, L_t)] \\ (\tilde{\boldsymbol{F}})_{ij} & i \in [\min(L_s, L_t) + 1, L_t] \end{cases}, \tag{A8}$$

with $(\tilde{\boldsymbol{F}})_{ij} \sim \mathcal{N}(0,1)$. As a consequence, the target teacher vector will also have a different number of components,

$$(\boldsymbol{\theta}_t)_i = \begin{cases} (\boldsymbol{\theta}_s)_i & i \in [1, \min(L_s, L_t)] \\ (\tilde{\boldsymbol{\theta}})_i & \text{for } i \in [\min(L_s, L_t) + 1, L_t] \end{cases}, \tag{A9}$$

with $(\tilde{\boldsymbol{\theta}})_i \sim \mathcal{N}(0,1)$.

In Fig. 4 we show two phase diagrams reporting the predicted target test loss as a function of T, q, and γ.

Appendix B: TL in deep fully connected networks—The kernel renormalization approach can be extended to deep nonlinear networks [31] via a recursive computation of the distribution of the preactivations h_μ^ℓ across layers $\ell = 1, \ldots, L$. This approach has been recently shown to be exact for deep linear networks [28,29]. Building on such results, we can combine the replica method with deep kernel renormalization to describe TL in deep neural networks. A first attempt of the calculation is sketched in the SM, leading to the following conjecture for the deep action:

$$S_n = \sum_l^L [\mathrm{Tr}(\lambda \mathcal{Q}_l \bar{\mathcal{Q}}_l) - \log\det(\mathbb{1} + \bar{\mathcal{Q}}_l)] - \frac{1}{N_L}\log\det(\mathbb{1} + \beta\mathcal{K}_{L+1}) - \frac{\beta}{N_L} y^T(\mathbb{1} + \beta\mathcal{K}_{L+1})^{-1} y, \tag{B1}$$

where we take $\lambda_\ell \equiv \lambda$ and $N_\ell = N$ equal for all layers, and the renormalized replicated kernel of the last layer $\mathcal{K}_{L+1}$ is computed recursively,

$$[\mathcal{K}_{\ell+1}]^{ab}_{\mu\nu} = \mathcal{Q}^{ab}_\ell [K_\ell]^{ab}_{\mu\nu} \tag{B2}$$

$$[K_\ell]^{ab}_{\mu\nu} = \langle \sigma(h^a_\mu)\sigma(h^b_\nu)\rangle_{\mathcal{N}(0,K_{\ell-1})}, \qquad K_0 \equiv C. \tag{B3}$$

We expect our derivation to be exact for deep linear networks, as long as replica symmetry is not broken. Corroborating Eq. (B1) via numerical experiments is left for future work.