On the Convergence, Implicit Bias and Edge of Stability of Gradient Descent in Deep Learning

Hancheng Min, *Member, IEEE*, Lachlan Ewen MacDonald and René Vidal, *Fellow, IEEE*

Despite the increasing proliferation of deep learning into everyday life, an adequate theoretical understanding of this technology remains elusive. Among the key challenges is understanding the *training* of *deep neural networks* (DNNs), which refers to the algorithmic process by which the parameters of a DNN are modified so that it learns to produce the correct outputs on training data. This training is ordinarily achieved by some variant of *gradient descent* (GD). Specifically, letting $\mathcal{L}(\boldsymbol{\theta})$ denote the *loss function* of the network, which measures the average deviation of the network outputs over the training data from their desired values at a given network parameter configuration $\boldsymbol{\theta}$, the GD update is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_t), \tag{1}$$

where η is the step size or *learning rate* and ∇ is the (sub-)gradient operator.

It is an empirical fact that DNNs trained via GD with random initialization and without any regularization enjoy good generalization performance in practice despite being highly *overparametrized*, i.e., the number of trainable parameters is much larger than the number of training samples. To theoretically understand this puzzling phenomenon, many works on convergence analysis for GD algorithms on neural networks have been developed over the last half-decade. In this article, we review these research efforts and discuss how they address three specific questions related to this puzzle. The first question is *why GD finds a global minimum efficiently*, which the literature has addressed by studying what level of overparametrization (width, depth, etc.) and what type of initialization lead to a benign optimization landscape along the training trajectory, facilitating a linear convergence rate of GD. The next question is *why the global minimum found by GD generalizes well*, which has been addressed by showing that

The first two authors contributed equally to this work. H. Min is with the Institute of Natural Sciences (INS), the School of Mathematical Sciences (SMS), and the MOE-LSC, Shanghai Jiao Tong University. L. E. MacDonald and R. Vidal are with the Center for Innovation in Data Engineering and Science (IDEAS) and the Department of Electrical and Systems Engineering (ESE), School of Engineering and Applied Science, University of Pennsylvania. R. Vidal is also with the Department of Radiology, Perelman School of Medicine, University of Pennsylvania.

overparametrization induces an implicit simplicity bias along the GD trajectory. More recently, it has been observed in practice that DNNs trained with larger learning rates than theoretically permissible converge rapidly "at the edge of stability" (EOS), wherein the loss converges to zero non-monotonically. This leads to the third question of *how training with large learning rates is possible*, which has been recently addressed by identifying a self-stabilization mechanism and implicit bias towards flat minima.

I. Convergence of GD on overparametrized networks

Classical theories of optimization focus primarily on minimizing *convex* objectives, where all first-order stationary points are global minima and can therefore be easily found using methods such as GD. DNN loss landscapes, on the other hand, are highly *non-convex*, admitting not only global minima but also many suboptimal stationary points in which training might be expected to get "stuck" without being able to proceed to a global minimum. Despite this, properly initialized GD on DNNs is generally observed to *avoid* all these suboptimal stationary points and consistently converge to global minima.

Studies concerning training neural networks with GD can be traced back several decades ago, for example, the work of Baldi [1] in 1988. Those with modernized views appeared within the past decade, among which Lee et al. [2] are among the earliest studying GD from an optimization landscape perspective: They show that GD converges to a global minimum when the loss function satisfies the following properties: 1) all of its local minima are global minima; and 2) every saddle point has a Hessian with at least one strict negative eigenvalue. Prior works suggest that shallow networks [3], and certain positively homogeneous networks [4] have such a landscape property, but unfortunately condition 2) does not hold for networks with multiple hidden layers [3]. Moreover, the landscape-based analyses generally do not characterize the convergence rate, except for a local rate around the equilibrium [2]. Those challenges in identifying a set of universal conditions on the loss landscape that admits global convergence of GD can be also attributed to the many undesired topological properties, such as non-closeness, lack of inverse stability, etc., of the function spaces realized by neural networks of fixed architecture, as investigated by Petersen et al. [5].

Nonetheless, it is natural to conjecture that while a good loss landscape does not hold globally, some benign non-convexity might do wherever visited by our optimization algorithms. Thus, *convergence analysis for GD on DNNs* has been developed, showing that with sufficient overparametrization and suitable initialization, the loss monotonically decreases at a linear rate along the GD trajectory. The key object in this research is a time-varying, positive-semidefinite operator called the *Neural Tangent Kernel* (NTK) [6], whose condition number controls the convergence rate of GD. This line of research focuses

on sufficient conditions on width, depth, initialization, step size, etc., under which the condition number of NTK can be controlled throughout training.

Recipe for linear convergence

Proving linear convergence of GD on non-convex functions typically requires the following standard assumptions. Let $\mathcal{L} : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function whose global minimum we assume for simplicity and without loss of generality to be zero. Given $L > 0$, $\mathcal{L}$ is said to be L-*smooth* if $\|\nabla\mathcal{L}(\boldsymbol{z}) - \nabla\mathcal{L}(\boldsymbol{z}')\| \leq L\|\boldsymbol{z} - \boldsymbol{z}'\|, \forall \boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^m$. Given $\mu > 0$, $\mathcal{L}$ is said to satisfy the μ-Polyak-Łojasiewicz (μ-PŁ) inequality if $\frac{1}{2}\|\nabla\mathcal{L}(\boldsymbol{z})\|^2 \geq \mu\mathcal{L}(\boldsymbol{z}), \forall \boldsymbol{z} \in \mathbb{R}^m$ (Here and for the remainder of this article, we let $\|\cdot\|$ to denote the Euclidean norm for vectors and the spectral norm for matrices). Taken together, L-smoothness and μ-PŁ suffice to guarantee convergence of GD $\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \eta\nabla\mathcal{L}(\boldsymbol{z}_t)$ at a linear rate in the following fashion [7].

The L-smoothness property is equivalent to the following inequality holding for any pair $\boldsymbol{z}, \boldsymbol{z}'$:

$$\mathcal{L}(\boldsymbol{z}) \leq \mathcal{L}(\boldsymbol{z}') + \left\langle \nabla\mathcal{L}(\boldsymbol{z}'), \boldsymbol{z} - \boldsymbol{z}' \right\rangle + \frac{L}{2}\|\boldsymbol{z} - \boldsymbol{z}'\|^2. \tag{2}$$

At iteration t, substituting $\boldsymbol{z}' = \boldsymbol{z}_t$ and $\boldsymbol{z} = \boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \eta\nabla\mathcal{L}(\boldsymbol{z}_t)$ yields the *descent lemma*,

$$\mathcal{L}(\boldsymbol{z}_{t+1}) \leq \mathcal{L}(\boldsymbol{z}_t) - \left(\eta - \frac{L\eta^2}{2}\right)\|\nabla\mathcal{L}(\boldsymbol{z}_t)\|^2. \tag{3}$$

Supposing the step size η satisfies $0 < \eta < \frac{2}{L}$, the descent lemma ensures the loss decreases monotonically. To derive a *rate* of decrease, one invokes the PL-inequality. Following the previously derived descent lemma, we have

$$\mathcal{L}(\boldsymbol{z}_{t+1}) \leq \mathcal{L}(\boldsymbol{z}_t) - \left(\eta - \frac{L\eta^2}{2}\right)\|\nabla\mathcal{L}(\boldsymbol{z}_t)\|^2 \overset{\left(\eta=\frac{1}{L}\right)}{=} \mathcal{L}(\boldsymbol{z}_t) - \frac{1}{2L}\|\nabla\mathcal{L}(\boldsymbol{z}_t)\|^2 \overset{(\mu\text{-PŁ})}{\leq} \left(1 - \frac{\mu}{L}\right)\mathcal{L}(\boldsymbol{z}_t).$$

Therefore, $\mathcal{L}(\boldsymbol{z}_{t+1}) \leq \left(1 - \frac{\mu}{L}\right)^{t+1}\mathcal{L}(\boldsymbol{z}_0)$. The loss converges to its global minimum at a linear rate $1 - \frac{\mu}{L}$.

Although DNN loss functions typically satisfy neither smoothness nor PŁ inequalities globally, they can be shown to hold *locally*, in a neighborhood of the optimization trajectory, within which convergence at a linear rate can thus be guaranteed in the above fashion. We examine this in more detail below.

GD in deep learning

Loss functions in deep learning are defined as follows. Let $g_{\boldsymbol{\theta}}(\cdot)$ be the input-output map of a neural network with parameters $\boldsymbol{\theta}$, and let the training data be $(\boldsymbol{X}, \boldsymbol{Y})$ where, $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$ contains n training data points and $\boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n]$ are the corresponding desired outputs. Then the loss $\mathcal{L}$ is a composition of two functions:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}), \quad \text{where } \boldsymbol{Z} = f_{\boldsymbol{X}}(\boldsymbol{\theta}). \tag{4}$$

Here, $f_{\boldsymbol{X}} : \boldsymbol{\theta} \mapsto \boldsymbol{Z} := [g_{\boldsymbol{\theta}}(\boldsymbol{x}_1), \cdots, g_{\boldsymbol{\theta}}(\boldsymbol{x}_n)]$ is the function that maps the network parameters $\boldsymbol{\theta}$ to the (concatenated) outputs $\boldsymbol{Z}$, and the loss $\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z})$ measures the discrepancy between the network outputs $\boldsymbol{Z}$ and the desired outputs $\boldsymbol{Y}$. We assume that $\mathcal{L}_{\boldsymbol{Y}}$ satisfies L-smoothness and μ-PŁ inequality globally; For example, for a regression problem, one typical choice for $\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z})$ is $\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{Z}\|_F^2$. We further suppose the neural network $g_{\boldsymbol{\theta}}$ is *overparametrized* such that $f_{\boldsymbol{X}}$ is surjective for any $\boldsymbol{X}$, except for those with repeated input data. Simply speaking, the neural network parametrization should be expressive enough (often requiring sufficient width, and appropriately placed bias terms) so that when given any input data with some desired output, one should be able to find some parameter under which the network produces the desired output, which is indeed often the case in modern practice [8]. In the case of $\mathcal{L}_{\boldsymbol{Y}}$ being the square loss, this overparametrization condition for $g_{\boldsymbol{\theta}}$ implies that the global minimum of $\mathcal{L}(\boldsymbol{\theta})$ is zero given any size-n training dataset $(\boldsymbol{X}, \boldsymbol{Y})$ with non-repeated input data.

Our goal is to recall the proof [7] of linear convergence of $\mathcal{L}$ towards zero under GD. Every time we update $\boldsymbol{\theta}$, it induces some update on $\boldsymbol{Z}$, resulting in a corresponding change in the loss $\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z})$. We have a full understanding of how changes in $\boldsymbol{Z}$ affect $\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z})$ since $\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z})$ has the required properties (Smoothness, PŁ inequality) discussed in the previous section. The only thing left unclear is the induced update on $\boldsymbol{Z}$.

For the sake of simplicity, assume $f_{\boldsymbol{X}}(\boldsymbol{\theta})$ is twice-differentiable w.r.t. $\boldsymbol{\theta}$. First of all, given the decomposition of the loss $\mathcal{L}$ as in (4), we use the chain rule to write $\nabla\mathcal{L}$ as

$$\nabla\mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{J}_{f_{\boldsymbol{X}}}^{\top}(\boldsymbol{\theta}) \cdot \nabla\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}) \,, \tag{5}$$

where $\nabla\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z})$ is the gradient of $\mathcal{L}_{\boldsymbol{Y}}$ w.r.t. the network outputs $\boldsymbol{Z}$, and $\boldsymbol{J}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta})$ is the Jacobian of $f_{\boldsymbol{X}}$ w.r.t. $\boldsymbol{\theta}$. Then, for every iteration t, we would like to know how $\boldsymbol{Z}_t = f_{\boldsymbol{X}}(\boldsymbol{\theta}_t)$ changes when $\boldsymbol{\theta}_t$ is updated, for which we rely on the first-order Taylor approximation of $f_{\boldsymbol{X}}$ around $\boldsymbol{\theta}_t$, evaluated at $\boldsymbol{\theta}_{t+1}$:

$$f_X(\boldsymbol{\theta}_{t+1}) = f_X(\boldsymbol{\theta}_t) + \boldsymbol{J}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)\,(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) + \mathcal{O}(\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^2) \,. \tag{6}$$

We see that the GD update from $\boldsymbol{\theta}_t$ to $\boldsymbol{\theta}_{t+1}$,

$$\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t = -\eta\nabla\mathcal{L}(\boldsymbol{\theta}_t) \overset{(5)}{=} -\eta\boldsymbol{J}_{f_{\boldsymbol{X}}}^{\top}(\boldsymbol{\theta}_t) \cdot \nabla\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t) \,,$$

induces a preconditioned GD on $\boldsymbol{Z}$ (up to an $\mathcal{O}(\eta^2)$ error term):

$$\boldsymbol{Z}_{t+1} = \boldsymbol{Z}_t - \eta \underbrace{\boldsymbol{J}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)\boldsymbol{J}_{f_{\boldsymbol{X}}}^{\top}(\boldsymbol{\theta}_t)}_{:=\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)} \cdot \nabla\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t) + \mathcal{O}(\eta^2) \,, \tag{7}$$

where the preconditioner $\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)$ is positive semi-definite and changes over GD iterations. This preconditioner $\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)$ is generally known as the *Neural Tangent Kernel (NTK)* [6], which has been studied

extensively over the past years from various perspectives. It has played a particularly important role in convergence analyses, as we now describe.

Linear convergence for GD on overparametrized networks

Let us first understand how the first-order term $-\eta \boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t) \cdot \nabla \mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t)$ in (7) affects the convergence of GD (neglecting the $\mathcal{O}(\eta^2)$ error term [1]). If $\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)$ is some scalar multiple of the identity matrix for all iterations t, then (7) is GD with a rescaled step size, and the linear convergence follows immediately from the aforementioned recipe. An issue may arise if $\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)$ is ill-conditioned: suppose $\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)$ has a zero eigenvalue, i.e., $\mathrm{Ker}\left(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)\right)$ is non-trivial; if $\mathbf{0} \neq \nabla \mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t) \in \mathrm{Ker}\left(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)\right)$, then $\boldsymbol{Z}_t$ receives zero update, and so does $\boldsymbol{\theta}_t$, which corresponds to GD being stuck at a (possibly suboptimal) critical point of the loss. Therefore, one would like $\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)$ to be well-conditioned throughout the GD trajectory for linear convergence.

Indeed, assuming $\mathcal{L}_{\boldsymbol{Y}}$ is L-smooth and satisfies the μ-PŁ inequality, the smoothness property of $\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z})$ as in (2), together with the preconditioned GD update in (7) leads to the following descent lemma:

$$\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_{t+1}) \leq \mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t) - \left(\lambda_{\min}(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t))\eta - \frac{\lambda^2_{\max}(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t))L\eta^2}{2}\right) \|\nabla \mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t)\|_F^2 + \mathcal{O}(\eta^2)\,. \tag{8}$$

By comparing the descent lemma in (8) and that in (3), one immediately realizes the main difference is due to the NTK. If we are able to show that for all $t \geq 0$ one has:

1) **Well-conditioned NTK**: $\exists \alpha > 0$ such that $\lambda_{\min}(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)) \geq \alpha$;
2) **Sufficiently small step size** η such that the sum of the higher-order terms is no larger than $\frac{\alpha\mu}{2}\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t)\eta$,

then linear convergence of $\mathcal{L}_{\boldsymbol{Y}}$ (equivalently, $\mathcal{L}$) can be derived from the descent lemma:

$$\begin{aligned}
\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_{t+1}) &\overset{\text{(Descent Lemma)}}{\leq} \mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t) - \left(\lambda_{\min}(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t))\eta - \frac{\lambda^2_{\max}(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t))L\eta^2}{2}\right) \|\nabla \mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t)\|_F^2 + \mathcal{O}(\eta^2) \\
&\overset{\text{(Well-conditioned NTK)}}{\leq} \mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t) - \alpha\|\nabla \mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t)\|_F^2\eta + \frac{\lambda^2_{\max}(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t))L}{2}\|\nabla \mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t)\|_F^2\eta^2 + \mathcal{O}(\eta^2) \\
&\overset{\text{(Sufficient small } \eta)}{\leq} \mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t) - \alpha\|\nabla \mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t)\|_F^2\eta + \frac{\alpha\mu}{2}\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t)\eta \\
&\overset{\text{(PL inequality)}}{\leq} \left(1 - \frac{\alpha\mu}{2}\eta\right)\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t)\,.
\end{aligned}$$

If one considers an "infinitesimal" step size ($\eta \to 0$), GD becomes the continuous-time *gradient flow* (GF), $\dot{\boldsymbol{\theta}} = -\nabla \mathcal{L}(\boldsymbol{\theta})$, which is also of great theoretical interest. A similar NTK-conditioning argument

[1]We note that since the error terms in (7) scale as $\mathcal{O}(\eta^2)$, it can always be neglected when η is sufficiently small; how small it should be depends on the choice of $f_{\boldsymbol{X}}$ and the iterates $\boldsymbol{\theta}_t$.

applies to show the exponential convergence of the loss $\mathcal{L}(\boldsymbol{\theta}(t))$ under GF: the time-derivative of $\mathcal{L}$ can be upper bounded as follows,

$$\dot{\mathcal{L}} = -\|\nabla_{\boldsymbol{\theta}}\mathcal{L}\|_F^2 = -\left\langle \nabla_{\boldsymbol{Z}}\mathcal{L}_{\boldsymbol{Y}}, \boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta})\nabla_{\boldsymbol{Z}}\mathcal{L}_{\boldsymbol{Y}}\right\rangle_F \leq -\lambda_{\min}\left(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta})\right)\|\nabla\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}_t)\|_F^2 . \tag{9}$$

If we can show $\lambda_{\min}(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)) \geq \alpha, \forall t \geq 0$, then we have $\dot{\mathcal{L}} \leq -\alpha\mu\mathcal{L}$, from which one conclude that $\mathcal{L}(t) \leq \exp\left(-\alpha\mu t\right)\mathcal{L}(0)$, yielding exponential convergence of the loss.

Notice that the aforementioned analysis is general in the sense that it only invokes basic properties from $\mathcal{L}_{\boldsymbol{Y}}$ (smoothness, PŁ) and $f_{\boldsymbol{X}}$ (twice-differentiability), has nothing to do with them being defined from a network training problem. In fact, $f_{\boldsymbol{X}}$ can be any parametric machine learning model that is twice-differentiable. The neural nature of the model is more associated with addressing the challenges of showing the NTK $\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta})$ is well-conditioned throughout the GD/GF trajectory, given its complicated dependence on the $f_{\boldsymbol{X}}$ and the fact that it varies as $\boldsymbol{\theta}$ gets updated during training, where specific architectural assumptions must be imposed on $f_{\boldsymbol{X}}$. In the remainder of this section, we review the past research efforts in addressing this challenge.

Convergence in the kernel regime

The most prominent line of work studies GD on DNNs [9]–[11] including fully-connected neural networks, residual networks, and convolutional neural networks. They assume an extremely large width at every layer and Gaussian random weight initialization with some properly chosen variance. In this heavily overparametrized regime (often referred to as the *kernel regime*), it can be shown that as the minimum width m among all layers grows, with high probability over random initialization, every entry of the NTK *at initialization* concentrates around the corresponding entry of a fixed kernel $\kappa_g^{\infty}(\cdot,\cdot)$ [10]:

$$[\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_0)]_{ij} = \left\langle \frac{\partial}{\partial\boldsymbol{\theta}} g_{\boldsymbol{\theta}(0)}(\boldsymbol{x}_i), \frac{\partial}{\partial\boldsymbol{\theta}} g_{\boldsymbol{\theta}(0)}(\boldsymbol{x}_j)\right\rangle \overset{m\to\infty}{\longrightarrow} \kappa_g^{\infty}(\boldsymbol{x}_i,\boldsymbol{x}_j) . \tag{10}$$

This fixed kernel $\kappa_g^{\infty}(\cdot,\cdot)$ varies depending on the network architecture, and it forms the basis of the analyses in the kernel regime.

Recall that for linear convergence, one requires well-conditioned NTK throughout GD iterations. In the kernel regime, this is guaranteed via the following arguments:

1) From (10), the NTK at initialization concentrates around some $\boldsymbol{H}_{f_{\boldsymbol{X}}}^{\infty} \succeq 0$ (often referred to as the *infinite-width NTK*), whose entries are given by $\kappa_g^{\infty}(\boldsymbol{x}_i,\boldsymbol{x}_j), 1 \leq i,j \leq n$:

$$\|\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_0) - \boldsymbol{H}_{f_{\boldsymbol{X}}}^{\infty}\| = o_m(1) ; \tag{11}$$

2) Furthermore, one can show that the weights do not deviate too much from its initialization during training: $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)\|_F$ is sufficiently small $\forall t \geq 0$. As a result, the NTK does not vary much throughout GD, i.e.,

$$\|\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t) - \boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_0)\| = o_m(1), \forall t \geq 0\,, \tag{12}$$

where the $o_m(1)$ error terms vanish as the width m grow to infinity. If additionally, one shows that $\boldsymbol{H}^{\infty}_{f_{\boldsymbol{X}}}$ is well-conditioned: $\exists \alpha > 0$ such that $\lambda_{\min}(\boldsymbol{H}^{\infty}_{f_{\boldsymbol{X}}}) \geq 2\alpha$, and that the width is sufficiently large (often exponential in the network depth and polynomial in the dataset size) so that the error terms in (11) and (12) are both smaller than $\frac{\alpha}{2}$, we conclude that

$$\lambda_{\min}(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)) \geq \lambda_{\min}(\boldsymbol{H}^{\infty}_{f_{\boldsymbol{X}}}) - \underbrace{\|\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t) - \boldsymbol{H}^{\infty}_{f_{\boldsymbol{X}}}\|}_{\leq \|\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_0) - \boldsymbol{H}^{\infty}_{f_{\boldsymbol{X}}}\| + \|\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t) - \boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_0)\| \leq \alpha} \geq 2\alpha - \alpha = \alpha\,, \forall t \geq 0\,.$$

The linear convergence is shown by the preceding analysis[2].

The fact that NTK is approximately constant during GD/GF in the kernel regime has an implication on the learning: training neural networks in this regime is almost equivalent to kernel regression with the infinite-width NTK [10]. We explain this point with GF in the kernel regime. Assume the network has a scalar output (then $\boldsymbol{Y} \in \mathbb{R}^n$) and the loss is $\mathcal{L}_{\boldsymbol{Y}}(\boldsymbol{Z}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{Z}\|^2$. Then, under GF, the evolution dynamics of network output $\boldsymbol{Z}_t = f_{\boldsymbol{X}}(\boldsymbol{\theta}(t))$ over training data $\boldsymbol{X}$, together with the network prediction $\boldsymbol{z}'_t = f_{\boldsymbol{x}'}(\boldsymbol{\theta}(t))$ over a new data point $\boldsymbol{x}'$, can be written as (by a similar argument to that for (7))

$$\frac{d}{dt}\begin{bmatrix}\boldsymbol{Z}_t \\ \boldsymbol{z}'_t\end{bmatrix} = \begin{bmatrix}\boldsymbol{J}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)\boldsymbol{J}^{\top}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t) \\ \boldsymbol{J}_{f_{\boldsymbol{x}'}}(\boldsymbol{\theta}_t)\boldsymbol{J}^{\top}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)\end{bmatrix} \underbrace{(\boldsymbol{Y} - \boldsymbol{Z}_t)}_{=\nabla \mathcal{L}_{\boldsymbol{Y}}}\,. \tag{13}$$

From (10), as width $m \to \infty$, $\boldsymbol{J}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)\boldsymbol{J}^{\top}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)$ converges to $\boldsymbol{H}^{\infty}_{f_{\boldsymbol{X}}}$, and $\boldsymbol{J}_{f_{\boldsymbol{x}'}}(\boldsymbol{\theta}_t)\boldsymbol{J}^{\top}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)$ converges to $[\kappa^{\infty}_g(\boldsymbol{x}', \boldsymbol{x}_1), \cdots, \kappa^{\infty}_g(\boldsymbol{x}', \boldsymbol{x}_n)] =: \boldsymbol{h}^{\infty}_{f_{\boldsymbol{x}'}, f_{\boldsymbol{X}}}$. If, additionally, the variance for the weight initialization is chosen such that the outputs $\boldsymbol{Z}_0, \boldsymbol{z}'_0$ vanish as the width grows, then (13) is well-approximated by

$$\frac{d}{dt}\begin{bmatrix}\boldsymbol{Z}_t \\ \boldsymbol{z}'_t\end{bmatrix} = -\begin{bmatrix}\boldsymbol{H}^{\infty}_{f_{\boldsymbol{X}}} & 0 \\ \boldsymbol{h}^{\infty}_{f_{\boldsymbol{x}'}, f_{\boldsymbol{X}}} & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{Z}_t \\ \boldsymbol{z}'_t\end{bmatrix} + \begin{bmatrix}\boldsymbol{H}^{\infty}_{f_{\boldsymbol{X}}} \\ \boldsymbol{h}^{\infty}_{f_{\boldsymbol{x}'}, f_{\boldsymbol{X}}}\end{bmatrix}\boldsymbol{Y}, \quad \boldsymbol{Z}_0 = 0, \boldsymbol{z}'_0 = 0\,. \tag{14}$$

Examining the solution to this linear dynamical system shows that

$$\lim_{t\to\infty} \boldsymbol{z}'_t = \lim_{t\to\infty} \boldsymbol{h}^{\infty}_{f_{\boldsymbol{x}'}, f_{\boldsymbol{X}}}\left(\boldsymbol{H}^{\infty}_{f_{\boldsymbol{X}}}\right)^{-1}\left(e^{-\boldsymbol{H}^{\infty}_{f_{\boldsymbol{X}}}t} - I\right)\boldsymbol{Y} = \boldsymbol{h}^{\infty}_{f_{\boldsymbol{x}'}, f_{\boldsymbol{X}}}\left(\boldsymbol{H}^{\infty}_{f_{\boldsymbol{X}}}\right)^{-1}\boldsymbol{Y}\,, \tag{15}$$

suggesting that the final predicted value $\boldsymbol{z}'_{\infty}$ for the new data $\boldsymbol{x}'$ coincides with the one obtained from kernel regression with the infinite-width kernel $\kappa^{\infty}_g(\cdot, \cdot)$; This is formally established in the work of Arora

[2]Allen-Zhu et al. [11] use a slightly different argument, but is the same in spirit: PŁ and smoothness inequalities are shown to hold over a small neighborhood of initialization, from which linear convergence of (stochastic) GD is derived

et al. [10]. Broadly speaking, these results are the consequence of a general observation [12] that training in the kernel regime is effectively training a linear model given by the linearization of the network around its weight initialization.

While these techniques and assumptions in the kernel regime enable strong convergence and learning guarantees, the fact that they ensure that training remains close to initialization is now known to be problematic. Specifically, models trained in the kernel regime fail to learn "good" features from data and achieve poor generalization performance in practice [13]. This motivates the study of the learning dynamics outside the kernel regime. For example, the work of Yang et al. [14] studies the infinite-width limit of neural networks initialized under the Maximal Update Parametrization (or μP), which has a smaller scale (i.e., the variance of the random Gaussian for weight initialization) than that of the kernel regime. This initialization scheme leads to significant changes in features learned from the input data. Moreover, for single-hidden-layer networks, such a choice of initialization scale allows for an elegant characterization of the weight evolution during training as some mean-field dynamics [15]. These studies show a more practical regime where the weights significantly change during training (the so-called *feature learning regime*), and the transition from the kernel regime to the feature learning regime can be controlled by the initialization scale [14], [16]. The training dynamics in the feature learning regime are often implicitly biased towards structurally simple models; we shall entertain this point in the second part of the article. For the remainder of this part, we discuss existing convergence analysis in this regime.

Convergence outside the kernel regime

In the feature learning regime, the weights quickly move away from their initialization following the GF/GD updates, thus so does the NTK. For this reason, convergence analyses outside the kernel regime are quite limited due to the necessity of characterizing spectral properties of the time-varying NTK along GD/GF trajectories. That is, to analyze the GF solution $\boldsymbol{\theta}(t)$ starting from some initial condition $\boldsymbol{\theta}(0)$, we would like a lower bound on the $\lambda_{\min}\left(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta}_t)\right)$ over the trajectory $\{\boldsymbol{\theta}(t), t \geq 0\}$ to show exponential convergence of the loss, which is challenging without exact knowledge of the solution. However, such lower bounds are possible under certain architectural assumptions. For instance, MacDonald et al. [17] show that a global (albeit non-uniform) lower-bound can be attained for deep networks with skip connections and appropriate weight normalization. Another example is the convergence analyses for linear networks [18], [19] (fully-connected networks with all activation functions being the identity) trained with gradient flow by exploiting conservation laws [20], namely functions $C_g : \Theta \to \mathbb{R}^{n_c}$ (that depend on the network $g_{\boldsymbol{\theta}}(\cdot)$) such that $C_g(\boldsymbol{\theta}(t)) = C_g(\boldsymbol{\theta}(0)), \forall t \geq 0$ for the GF solution $\boldsymbol{\theta}(t)$. Using conservation laws, the entire GF trajectory is restricted to an invariant subset of the parameter space

defined by the conserved quantities, and we can relax our search for a lower bound on $\lambda_{\min}\left(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta})\right)$ over the GF trajectory to over the invariant subset, as illustrated below:

$$\min_{\boldsymbol{\theta}} \lambda_{\min}\left(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta})\right) \ \ s.t.\ \boldsymbol{\theta} \in \{\boldsymbol{\theta}(t) : t \geq 0\} \qquad \text{(Examing NTK over GF trajectory)}$$

$$\text{(Relaxation)} \Downarrow$$

$$\min_{\boldsymbol{\theta}} \lambda_{\min}\left(\boldsymbol{H}_{f_{\boldsymbol{X}}}(\boldsymbol{\theta})\right) \ \ s.t.\ C_g(\boldsymbol{\theta}) = C_g(\boldsymbol{\theta}(0)) \qquad \text{(Examing NTK over invariant set)}$$

Specifically, for an L layer linear network with weights $\boldsymbol{W}_l$ at layer $l \leq L$, the quantities $\boldsymbol{W}_{l+1}^\top \boldsymbol{W}_{l+1} - \boldsymbol{W}_l \boldsymbol{W}_l^\top, l = 1, \cdots, L-1$ (called *imbalance matrices* [19]) are time-invariant under GF. This is sufficient to show convergence: Min et al. [19] show that for two-layer linear networks, as long as the (only) imbalance matrix has sufficiently large rank at initialization (a probability-one event), the least eigenvalue of the NTK is lower bounded by some constant for all time, leading to exponential convergence of GF. This result can be extended to certain deep linear networks [21]. Complementarily, the case of all imbalance matrices being zero, generally referred to as the *balanced case*, is also of particular research interest, under which the GF on network parameters $\boldsymbol{\theta}$ induces a Riemannian gradient flow on the network input-output linear map (the product of weight matrices). The global convergence of the induced Riemannian gradient flow (except for a zero-measure set of initialization) is shown by Bah et al. [22] for two-layer linear networks, and an exponential convergence rate can be derived if the initial matrix product is sufficiently close to the global optimum [18]. Lastly, analyses for GF can be translated into analyses for GD with small stepsize [18], where the imbalance matrices are shown to be approximately preserved along GD trajectories.

II. Sparse and low-rank implicit bias of GD

In many signal processing problems, one is given observations from few measurements, and seeks to recover or estimate the true signal underlying these observations from the infinitely many signals which are compatible with them [23]. The seemingly daunting task of recovering the true signal under limited measurements has been elegantly addressed by solving *explicitly regularized* optimization problems that promote simplicity in the recovered signal, motivated by the fact that the natural signals we encounter in practice often possess certain notions of simplicity (sparsity, low-rankness, etc.).

In deep learning, a related phenomenon exists. We have seen in previous sections that having a large number of trainable parameters, i.e., overparametrization, facilitates the linear convergence of GD. However, overparametrization also implies the existence of *many* global minima, which perfectly fit the training data, but differ in their ability to *generalize* to new data. Thus overparametrization brings with it the *learning problem* of ensuring that GD finds only those minima that generalize well to test data.

In the same spirit as in signal processing problems, among all the infinitely many models that fit a training set, low-complexity models should be expected to generalize better than high-complexity models. What differs in deep learning, however, is that there is often no need to explicitly regularize the model during training (although weight decay is frequently used in practice). When training DNNs, overparametrization and appropriate initialization *implicitly regularize* training via an *implicit bias* of GD towards low-complexity models that generalize well.

However, it should be noted that implicit biases towards low-complexity models are rather ubiquitous in various training regimes, and they mainly differ in the way the model complexity is measured. For example, in the kernel regime, the GF/GD generally finds, among all interpolating functions that fit the training set, the one with the least Reproducing Kernel Hilbert Space (RKHS) norm induced by the infinite-width NTK [10]. As the initialization scale decreases (moving toward the feature learning regime), these implicit regularization effects promote (certain notions of) sparsity in the learned network function [24]. Moreover, such regularization effects can be precisely characterized for certain linear networks (so that the learned function is linear) as minimizing ℓ_1 or nuclear norms that promote sparsity or low-rankness. In this section, we introduce several lines of work characterizing such a sparsity-promoting implicit bias, or implicit regularization, in the context of classic problems in signal processing, such as sparse recovery [23] and low-rank matrix sensing [25]. In Table I, we summarize the problems to be covered and list some classic regularization approaches for solving them. In addition, we list related overparametrized formulations studied in deep learning theory, for which GD and its variants have been shown to exhibit certain implicit regularization that aligns with those explicitly enforced in classic approaches.

Regularization via implicit mirror flow

First, we consider the linear regression problem and discuss how overparametrization induces an implicit regularization on the solution obtained by GF, explained through an implicit mirror descent perspective [31]. Consider a loss function $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{w}\|^2$, where $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{w} = f(\boldsymbol{\theta})$ is the image of network parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ under a differentiable map $f : \boldsymbol{\Theta} \to \mathbb{R}^d$. By minimizing $\mathcal{L}$, we seek some parameter $\boldsymbol{\theta}$ whose corresponding linear predictor $\boldsymbol{w}$ agrees with the observations $\boldsymbol{y}$ through some linear measurement $\boldsymbol{A}$. We say this problem is *underdetermined* when the number of observations n is less than the dimension d of the linear model to be estimated, thus admitting infinitely many solutions.

Notice that $\mathcal{L} = \ell \circ f$ is the composition of two functions, $\ell : \boldsymbol{w} \to \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{w}\|^2$ and f; hence, our previous discussion in the section "GD in deep learning" on how gradient update on $\boldsymbol{\theta}$ induces a

TABLE I

EXPLICIT AND IMPLICIT REGULARIZATION FOR OBTAINING LOW-COMPLEXITY MODELS.

	Signal Processing	**Deep Learning**	
	Explicit Regularization	Overparametrization	Implicit Regularization via Small Initialization
Sparse Recovery [23]	$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\beta}\|^2+\gamma\|\boldsymbol{\beta}\|_1$	$\min_{u,v\in\mathbb{R}^d} \|\boldsymbol{y}-\boldsymbol{A}(\overbrace{\boldsymbol{u}\odot\boldsymbol{u}\cdots\odot\boldsymbol{u}}^{:=\boldsymbol{u}^{\odot L}}-\boldsymbol{v}^{\odot L})\|^2$ ($:=\boldsymbol{w}$)	$\|\boldsymbol{w}\|_1$ [16], [26]
Compressed Sensing [23]	$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \|\boldsymbol{y}-\boldsymbol{\mathcal{A}}(\boldsymbol{\beta})\|^2+\gamma\|\mathcal{F}\boldsymbol{\beta}\|_1$	$\min_{\boldsymbol{w}_i\in\mathbb{R}^d} \|\boldsymbol{y}-\boldsymbol{\mathcal{A}}(\overbrace{\boldsymbol{w}_1\circledast\boldsymbol{w}_2\circledast\cdots\circledast\boldsymbol{w}_L}^{:=\boldsymbol{w}})\|^2$	$\|\mathcal{F}\boldsymbol{\beta}\|_1$ [26], [27]
Low-rank Matrix Sensing [25]	$\min_{\boldsymbol{X}\in\mathbb{R}^{d\times d}} \|\boldsymbol{y}-\boldsymbol{\mathcal{A}}(\boldsymbol{X})\|^2+\gamma\|\boldsymbol{X}\|_*$ $\min_{\boldsymbol{U},\boldsymbol{V}\in\mathbb{R}^{d\times r}} \|\boldsymbol{y}-\boldsymbol{\mathcal{A}}(\boldsymbol{U}\boldsymbol{V}^T)\|^2$	$\min_{\substack{\boldsymbol{W}_i\in\mathbb{R}^{d_i\times d_{i+1}}\\ d_1=d,d_{L+1}=d}} \|\boldsymbol{y}-\boldsymbol{\mathcal{A}}(\overbrace{\boldsymbol{W}_1\boldsymbol{W}_2\cdots\boldsymbol{W}_L}^{:=\boldsymbol{W}})\|^2$	$\|\boldsymbol{W}\|_*$ [28] rank$(\boldsymbol{W})$ [29], [30]

$\mathcal{F}$: Discrete Fourier Transform; $\odot$: Hadamard (element-wise) product; $\circledast$: circular convolution

preconditioned gradient update on $\boldsymbol{w}$ still holds. In particular, under GF $\dot{\boldsymbol{\theta}} = -\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$, we have the following induced continuous dynamics of $\boldsymbol{w}$:

$$\dot{\boldsymbol{w}} = -\underbrace{J_f(\boldsymbol{\theta})J_f^\top(\boldsymbol{\theta})}_{:=\boldsymbol{H}_f(\boldsymbol{\theta})}\nabla\ell(\boldsymbol{w}) = \boldsymbol{H}_f(\boldsymbol{\theta})\boldsymbol{A}^\top\boldsymbol{r}(\boldsymbol{w})\,, \tag{16}$$

where $\boldsymbol{H}_f(\boldsymbol{\theta})$ is derived in the same way as in (7), and $\boldsymbol{r}(\boldsymbol{w}) = \boldsymbol{y}-\boldsymbol{A}\boldsymbol{w}$ is the residual.

The solution $\boldsymbol{w}(t)$ of the ordinary differential equation in (16) is still challenging to analyze due to the dependence on the evolution of $\boldsymbol{\theta}(t)$; thus we wish to write $\boldsymbol{H}_f(\boldsymbol{\theta})$ as some function of $\boldsymbol{w}$. Given some GF trajectory $\boldsymbol{\theta}(t)$, if there exists some convex ψ over $\mathbb{R}^d$, such that

$$\boldsymbol{H}_f^{-1}(\boldsymbol{\theta}(t)) = \nabla^2\psi(\boldsymbol{w}(\boldsymbol{\theta}(t)))\,, \forall t\geq 0\,, \tag{17}$$

then under that trajectory $\boldsymbol{\theta}(t)$, we have

$$\text{(Implicit Mirror Flow)}\quad \dot{\boldsymbol{w}} = \left(\nabla^2\psi(\boldsymbol{w})\right)^{-1}\boldsymbol{A}^\top\boldsymbol{r}(\boldsymbol{w})\,, \tag{18}$$

which is the continuous-time limit of mirror descent [32] on $\ell(\boldsymbol{w})$ w.r.t. potential function ψ. This implicit mirror flow favors the interpolating solution that is the closest to the initial condition $\boldsymbol{w}(0)$ in the *Bregman divergence* w.r.t. ψ. The following theorem is a formal version of the discussion in [31]:

Theorem 1. *Consider $\boldsymbol{w}(t) = f(\boldsymbol{\theta}(t))$, where $\boldsymbol{\theta}(t)$ is the solution to GF $\dot{\boldsymbol{\theta}} = -\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$ from some initial condition $\boldsymbol{\theta}(0)$. Assume that 1) $\boldsymbol{w}(\infty) = \lim_{t\to\infty}\boldsymbol{w}(t)$ exists and is an interpolating solution,*

i.e. $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{w}(\infty)$*; and 2) there exists a convex function* $\psi : \mathbb{R}^d \to \mathbb{R}$ *such that* (17) *holds (*ψ *generally depends on* $\boldsymbol{\theta}(0)$*), then* $\boldsymbol{w}(\infty)$ *is a global optimum of*

$$\min_{\boldsymbol{w}} D_\psi(\boldsymbol{w}, \boldsymbol{w}(0)), \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{w}\,, \tag{19}$$

where $D_\psi(\boldsymbol{w}, \boldsymbol{w}') = \psi(\boldsymbol{w}) - \psi(\boldsymbol{w}') - \langle \nabla\psi(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}' \rangle$ *is the Bregman divergence w.r.t. potential* ψ*.*

Proof. We write the Lagrangian $L(\boldsymbol{w}, \boldsymbol{\nu}) = D_\psi(\boldsymbol{w}, \boldsymbol{w}(0)) + \langle \boldsymbol{\nu}, \boldsymbol{y} - \boldsymbol{A}\boldsymbol{w} \rangle$, where $\boldsymbol{\nu}$ is the Lagrangian multiplier. Since (19) has a convex objective and affine constraints, it suffices to show that $\boldsymbol{w}(\infty)$ satisfies the Karush–Kuhn–Tucker (KKT) condition of (19): the primal feasibility that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{w}(\infty)$ and the stationarity of the Lagrangian that $\exists \boldsymbol{\nu}^*$ such that

$$\frac{\partial L}{\partial \boldsymbol{w}}(\boldsymbol{w}(\infty), \boldsymbol{\nu}^*) = \nabla\psi(\boldsymbol{w}(\infty)) - \nabla\psi(\boldsymbol{w}(0)) - \boldsymbol{A}^\top \boldsymbol{\nu}^* = 0\,. \tag{20}$$

The former is given by our assumption, hence only the latter remains to be proved. From (18), we have

$$\frac{d}{dt}\left(\nabla\psi(\boldsymbol{w})\right) = \nabla^2\psi(\boldsymbol{w})\dot{\boldsymbol{w}} \overset{(18)}{=} \boldsymbol{A}^\top \boldsymbol{r}(\boldsymbol{w})\,, \tag{21}$$

and integrating both side gives exactly the stationarity of the Lagrangian:

$$\nabla\psi(\boldsymbol{w}(\infty)) - \nabla\psi(\boldsymbol{w}(0)) = \boldsymbol{A}^\top \underbrace{\int_0^\infty \boldsymbol{r}(\boldsymbol{w}(t))dt}_{:=\boldsymbol{\nu}^*}\,, \tag{22}$$

where the integral over residual $\boldsymbol{r}(\boldsymbol{w}(t))$ can be shown to exist because the left-hand side is finite. □

Theorem 1 may be viewed as a continuation of our discussion on how the preconditioner $\boldsymbol{H}_f(\boldsymbol{\theta}(t))$ affects GD/GF training. We have shown in the previous section that well-conditioned $\boldsymbol{H}_f(\boldsymbol{\theta}(t))$ guarantees convergence of the loss. Here, we take one step further to understand its implication on the implicit bias of GF: if the preconditioner has the particular form given in (17), then the implicit regularization on the obtained interpolating solution can be fully described by Theorem 1. Now, we give an example network for which this theorem applies (In the following discussions, x_i denotes the i-th entry of a vector $\boldsymbol{x}$, which then can be written as $\boldsymbol{x} = [x_i]_{i=1}^d$).

Example 2 (Depth-2 diagonal linear networks)**.** Consider $\boldsymbol{\theta} = (\boldsymbol{u}, \boldsymbol{v})$ and $\boldsymbol{w} = \boldsymbol{u} \odot \boldsymbol{u} - \boldsymbol{v} \odot \boldsymbol{v} := f(\boldsymbol{u}, \boldsymbol{v})$ ($\odot$ represents Hadamard (element-wise) product). Then the GF solution $\boldsymbol{\theta}(t)$ from $\boldsymbol{\theta}(0) = (\boldsymbol{u}(0), \boldsymbol{v}(0)) = (\varepsilon\boldsymbol{a}, \varepsilon\boldsymbol{a})$ satisfies (17) with some ψ whose corresponding Bregman divergence in (19) is given by

$$D_\psi(\boldsymbol{w}, \boldsymbol{w}(0)) = \sum_{i=1}^d \frac{k_i}{4}\left(1 + \frac{w_i}{k_i}\operatorname{arcsinh}\left(\frac{w_i}{k_i}\right) - \sqrt{1 + \left(\frac{w_i}{k_i}\right)^2}\right)\,, \tag{23}$$

where $\operatorname{arcsinh}(x) = \log(x + \sqrt{1 + x^2})$ and $k_i = 2\varepsilon^2 a_i^2$.

Proof. By definition of $f(\boldsymbol{\theta})$, its Jacobian is $2\begin{bmatrix}\mathrm{diag}(\boldsymbol{u}) & \mathrm{diag}(\boldsymbol{v})\end{bmatrix}$, thus

$$\boldsymbol{H}_f(\boldsymbol{\theta}) = 4\left(\mathrm{diag}(\boldsymbol{u}\odot\boldsymbol{u}) + \mathrm{diag}(\boldsymbol{v}\odot\boldsymbol{v})\right) = \mathrm{diag}\left(\left[4\sqrt{(u_i^2-v_i^2)^2+4(u_iv_i)^2}\right]_{i=1}^{d}\right). \tag{24}$$

We want to rewrite $\boldsymbol{H}_f(\theta)$ in terms of $\boldsymbol{w}$. Notice that the right-hand side of (24) already has $u_i^2-v_i^2=w_i$; We eliminate the u_iv_i term by the training invariance: one can show that $\frac{d}{dt}(\boldsymbol{u}\odot\boldsymbol{v})=0$, thus $u_i(t)v_i(t) = u_i(0)v_i(0) = a_i^2\varepsilon^2, \forall i, \forall t$. Therefore, we have

$$\nabla^2\psi(\boldsymbol{w}) = \boldsymbol{H}_f^{-1}(\boldsymbol{\theta}) = \mathrm{diag}\left(\left[\frac{1}{4\sqrt{w_i^2+4a_i^4\varepsilon^4}}\right]_{i=1}^{d}\right), \tag{25}$$

and the desired ψ function are of the form $\sum_{i=1}^d q_i(w_i)$ for some $q_i(z)$ whose second order derivative is $\frac{1}{4\sqrt{z^2+4a_i^4\varepsilon^4}}$, from which we obtain the Bregman divergence as in (23). □

[24] characterizes (23) when the *initialization shape* $\boldsymbol{a}$ is fixed and the *intialization scale* ε vanishes or grows to infinity. They show that as $\varepsilon\to\infty$, $D_\psi(\boldsymbol{w},\boldsymbol{w}(0)) \propto \sum_{i=1}^d \frac{w_i^2}{a_i^2}$, an ℓ_2-norm weighted by the initialization shape, whereas when $\varepsilon\to 0$, the divergence $D_\psi(\boldsymbol{w},\boldsymbol{w}(0)) \propto \sum_{i=1}^d |w_i|$ approximates the standard ℓ_1-norm regardless of the shape. This observation reveals the crucial role of the initialization scale in determining the implicit regularization of GF on overparametrized networks, showing the potential benefit of small initialization for finding interpolating solutions that are sparse.

More generally, Theorem 1 applies to diagonal linear networks with a depth larger than two. For example, Woodworth et al. [24] consider $\boldsymbol{w} = \boldsymbol{u}^{\odot L} - \boldsymbol{v}^{\odot L}$ with uniform initialization shape $\boldsymbol{u}(0) = \boldsymbol{v}(0) = \varepsilon\mathbf{1}$ and Yun et al. [26] consider $\boldsymbol{w} = \boldsymbol{w}_1\odot\boldsymbol{w}_2\odot\cdots\odot\boldsymbol{w}_L$ with non-uniform shape $\boldsymbol{w}_1(0) = \cdots = \boldsymbol{w}_{L-1}(0) = \varepsilon\boldsymbol{a}$, $\boldsymbol{w}_L(0) = \mathbf{0}$. In those cases, D_ψ has no closed form, but one can still characterize its asymptotic behavior: as the initialization scale vanishes, the $D_\psi(\boldsymbol{w},\boldsymbol{w}(0))$ in the former case approximates standard ℓ_1-norm [24], and notably, the $D_\psi(\boldsymbol{w},\boldsymbol{w}(0))$ in the latter case approximates $\sum_{i=1}^d \frac{|w_i|}{|a_i|^{L-2}}$, a weighted ℓ_1-norm by the initialization shape [26]. Moreover, the analysis in the work of Yun et al. [26] also covers certain linear convolution networks $\boldsymbol{w} : (\boldsymbol{w}_l)_{l=1}^L \mapsto \boldsymbol{w}_1 \circledast \boldsymbol{w}_2 \circledast \cdots \circledast \boldsymbol{w}_L$, where $\circledast$ represents circular convolution. The key observation here is that for Discrete Fourier Transform (DFT) $\mathcal{F}$, we have $\mathcal{F}(\boldsymbol{w}_1 \circledast \boldsymbol{w}_2) = (\mathcal{F}\boldsymbol{w}_1)\odot(\mathcal{F}\boldsymbol{w}_2)$; That is, linear circular convolution networks are equivalent to diagonal linear networks in the Fourier domain, thus training with circular convolutional networks promotes sparsity in the DFT coefficients of the resulting linear model [26]. Furthermore, this implicit mirror flow perspective applies to matrix sensing problems when all the measurement matrices commute and an implicit nuclear norm regularization exists under small initialization [28], [33]. Lastly, it is worth noting that while the mirror-flow-based analysis can cover many problem instances, its key underlying assumption of the existence of a convex potential that can be related to the NTK in the form of (17) is

limited to very specific network structures. Therefore, to expand our understanding of implicit biases to broader classes of networks, complementary analysis and tools are needed, which we discuss next.

Regularization via incremental and spectral learning

Until now, we have recalled how to characterize the implicit sparse regularization on the steady-state solution achieved by GF via small initialization. This section discusses a complementary perspective of this simplicity bias, focusing on transient incremental and spectral learning phenomena at the early training phase. To be more precise, *incremental learning* [30], [34] refers to the sequential fitting of components of the underlying ground truth function (assuming it exists) in some canonical decomposition, without learning spurious components that would lead to overfitting to the training data.. While similar phenomena can occur in modern network architectures [35], the precise characterization of incremental learning has only been studied for classic problems where the associated canonical decomposition are well understood. One important testbed is the *low-rank matrix sensing* [25] problem with the associated decomposition being the singular value decomposition, where a *spectral learning* mechanism of learning singular vectors is studied as one crucial dynamic process to achieve incremental learning.

We consider the loss function $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{\mathcal{A}}(\boldsymbol{W})\|_F^2$, where $\boldsymbol{W} = f(\boldsymbol{\theta})$, by minimizing which we seek some parameter $\boldsymbol{\theta}$ whose image under the differentiable map $f : \boldsymbol{\Theta} \mapsto \mathbb{R}^{d\times d}$ agrees with the observations $\boldsymbol{Y}$ through some linear measurement $\boldsymbol{\mathcal{A}}$. We will start by explaining the idea of incremental and spectral learning in the symmetric matrix factorization setting, where $\boldsymbol{Y}$ is positive semi-definite, $\boldsymbol{\mathcal{A}}$ is the identity map, and we let $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{U}^\top := f(\boldsymbol{U})$. Then the loss is $\mathcal{L}(\boldsymbol{U}) = \frac{1}{4}\|\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{U}^\top\|_F^2$ (we adjust the constant before the Frobenius norm to remove any extra constant in the GF equations), and the GF defined by $\dot{\boldsymbol{U}} = -\nabla_{\boldsymbol{U}}\mathcal{L}(\boldsymbol{U}) = (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U}$ induces a continuous dynamic flow of $\boldsymbol{W}$:

$$\dot{\boldsymbol{W}} = (\boldsymbol{Y} - \boldsymbol{W})\boldsymbol{W} + \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{W})\,. \tag{26}$$

Letting $\boldsymbol{Y} = \boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\boldsymbol{\Phi}^\top$ be the SVD of $\boldsymbol{Y}$ and making the change of variable $\tilde{\boldsymbol{W}} = \boldsymbol{\Phi}^\top\boldsymbol{W}\boldsymbol{\Phi}$, we immediately have

$$\dot{\tilde{\boldsymbol{W}}} = (\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \tilde{\boldsymbol{W}})\tilde{\boldsymbol{W}} + \tilde{\boldsymbol{W}}(\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \tilde{\boldsymbol{W}})\,. \tag{27}$$

Notably, initializations of the form $\boldsymbol{U}(0) = \alpha^{1/2}\boldsymbol{\Phi}\boldsymbol{\Sigma}_0^{1/2}$ (for some diagonal matrix $\boldsymbol{\Sigma}_0 \succeq 0$), called *spectral initialization* (we use $\alpha^{1/2}$ for the initialization scale and $\boldsymbol{\Sigma}_0^{1/2}$ the shape), induce a diagonal $\tilde{\boldsymbol{W}}(0)$, starting from which the solution $\tilde{\boldsymbol{W}}(t)$ to (27) remains diagonal for all time t. The following result describes how the singular values of $\tilde{\boldsymbol{W}}(t)$ evolve along the GF trajectory (We omitted its proof, which is simply solving ordinary differential equations (ODEs) in (27) when $\tilde{\boldsymbol{W}}$ is diagonal.).

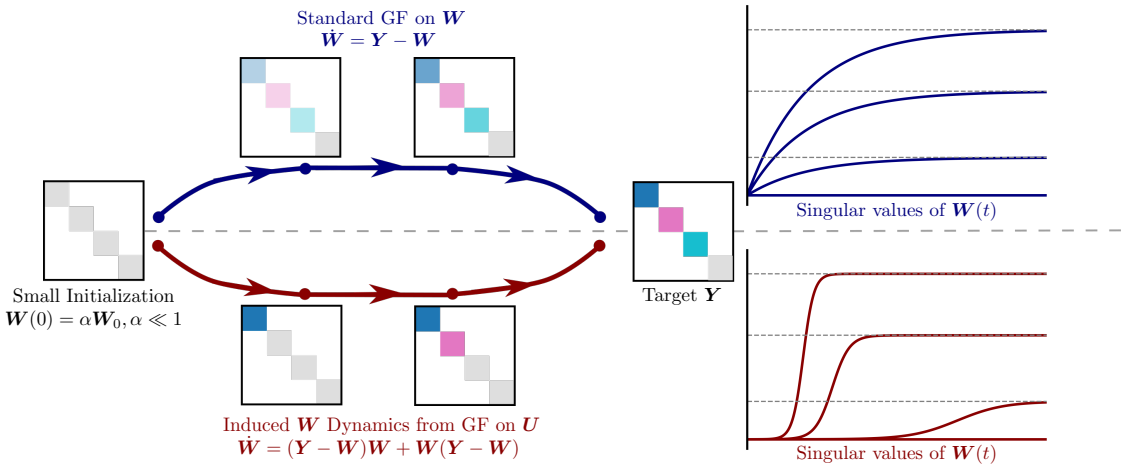

Fig. 1. Incremental learning in GF on symmetric matrix factorization problem with small initialization. Compared to standard GF on $\boldsymbol{W}$, where all the singular components of $\boldsymbol{Y}$ are being learned simultaneously at the same rate, the GF on the factor $\boldsymbol{U}$ induces nonlinear dynamics of $\boldsymbol{W}$ whose solution exhibits an incremental learning phenomenon that larger singular component of $\boldsymbol{Y}$ is learned earlier so that one can find low-rank approximations of $\boldsymbol{Y}$ along the GF trajectory.

Proposition 3. *Suppose* $\boldsymbol{\Sigma}_{\boldsymbol{Y}} = \operatorname{diag}\left([\sigma_i]_{i=1}^d\right)$. *Under some spectral initialization* $\boldsymbol{U}(0) = \alpha^{1/2}\boldsymbol{\Phi}\boldsymbol{\Sigma}_0^{1/2}$, *where* $\boldsymbol{\Sigma}_0 = \operatorname{diag}\left([\sigma_{i,0}^{1/2}]_{i=1}^d\right)$, *the GF* $\dot{\boldsymbol{U}} = -\nabla_{\boldsymbol{U}}\mathcal{L}(\boldsymbol{U})$ *induces* $\boldsymbol{W}(t) = \boldsymbol{U}(t)\boldsymbol{U}(t)^\top$ *of the form* $\boldsymbol{W}(t) = \boldsymbol{\Phi}\operatorname{diag}\left([\sigma_{i,\boldsymbol{W}}(t)]_{i=1}^d\right)\boldsymbol{\Phi}^\top$, *where*

$$\sigma_{i,\boldsymbol{W}}(t) = \frac{\alpha\sigma_i\sigma_{i,0}e^{2\sigma_i t}}{\sigma_i + \alpha\sigma_{i,0}(e^{2\sigma_i t} - 1)}\,,\ \textit{if } \sigma_i \neq 0; \qquad \sigma_{i,\boldsymbol{W}}(t) = \frac{\alpha\sigma_{i,0}}{1 + 2\alpha\sigma_{i,0}t}\,,\ \textit{if } \sigma_i = 0\,. \tag{28}$$

As suggested by this proposition, under spectral initialization, each singular component σ_i of $\boldsymbol{Y}$ is learned by one of the components $\sigma_{i,\boldsymbol{W}}(t)$ (a monotonically increasing function w.r.t. t when $\alpha\sigma_{i,0} < \sigma_i$) of $\boldsymbol{W}(t)$, through a nonlinear ODE whose solution (shown in (28)) has an interesting transient behavior when the initialization scale is small: For a fixed index i, consider $\alpha \ll 1$ and some $c, C > 0$ such that $c\sigma_{i,0} \ll \sigma_i$ and $C\sigma_{i,0} \gg \sigma_i$. Then we see that

$$\sigma_{i,\boldsymbol{W}}\left(\frac{1}{2\sigma_i}\log\frac{c}{\alpha}\right) = \frac{c\sigma_i\sigma_{i,0}}{\sigma_i - \alpha\sigma_{i,0} + c\sigma_{i,0}} \simeq c\sigma_{i,0},\ \ \sigma_{i,\boldsymbol{W}}\left(\frac{1}{2\sigma_i}\log\frac{C}{\alpha}\right) = \frac{C\sigma_i\sigma_{i,0}}{\sigma_i - \alpha\sigma_{i,0} + C\sigma_{i,0}} \simeq \sigma_i\,, \tag{29}$$

suggesting a sharp transition in $\sigma_{i,\boldsymbol{W}}(t)$ from a small value $c\sigma_{i,0}$ to one close to its target value σ_i at a time that scales as $\Theta\left(\frac{1}{\sigma_i}\log\frac{1}{\alpha}\right)$. Notably, the transition time depends inverse-proportionally on σ_i; thus, large singular values of $\boldsymbol{Y}$ get learned first in $\boldsymbol{W}(t)$, and small ones are learned later. This *incremental learning* phenomenon can be formally described as follows:

Theorem 4 (Incremental learning under small spectral initialization)**.** *Suppose the* K *non-zero singular values* $\sigma_1, \sigma_2, \cdots, \sigma_K$ *of* $\boldsymbol{Y}$ *are distinct and ordered in decreasing order. Let the spectral initialization*

$\boldsymbol{U}(0) = \alpha^{1/2}\boldsymbol{\Phi}\boldsymbol{\Sigma}_0^{1/2}$ have a uniform initialization shape $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$. Given any $0 < \varepsilon \le \sigma_K$, let $c_\varepsilon = \varepsilon$, and $C_\varepsilon = \frac{\sigma_1^2}{\varepsilon}$, Suppose α is sufficiently small such that $\alpha \le \varepsilon$ and that $\frac{-\log\alpha+\log c_\varepsilon}{-\log\alpha+\log C_\varepsilon} > \max_{1\le k\le K-1}\frac{\sigma_{k+1}}{\sigma_k}$, then the $\boldsymbol{W}(t)$ in Proposition 3 satisfies that $\forall 1 \le k \le K$,

$$\|\boldsymbol{W}(t) - \hat{\boldsymbol{Y}}_k\| \le \varepsilon, \quad \forall t \in \left[\frac{1}{2\sigma_k}\log\frac{C_\varepsilon}{\alpha}, \frac{1}{2\sigma_{k+1}}\log\frac{c_\varepsilon}{\alpha}\right] \neq \emptyset, \tag{30}$$

where $\hat{\boldsymbol{Y}}_k := \arg\min_{\mathrm{rank}(\boldsymbol{Z})=k}\|\boldsymbol{Y} - \boldsymbol{Z}\|_F$ is the best rank-k approximation of $\boldsymbol{Y}$ and we let $\frac{1}{0} := \infty$.

Proof. We have assumed small α such that $\frac{-\log\alpha+\log c_\varepsilon}{-\log\alpha+\log C_\varepsilon} > \max_{1\le k\le K-1}\frac{\sigma_{k+1}}{\sigma_k}$, which implies that for every $1 \le k \le K$, $\left[\frac{1}{2\sigma_k}\log\frac{C_\varepsilon}{\alpha}, \frac{1}{2\sigma_{k+1}}\log\frac{c_\varepsilon}{\alpha}\right] \neq \emptyset$. For a fixed k, choose any t within this interval. Since both $\boldsymbol{W}(t)$ and $\hat{\boldsymbol{Y}}_k$ are diagonal, (30) is equivalent to $|\sigma_{l,\boldsymbol{W}}(t) - \sigma_l| \le \varepsilon, \forall l \le k$ and $|\sigma_{l,\boldsymbol{W}}(t)| \le \varepsilon, \forall l > k$. Now consider the solution provided in Proposition 3, we have, $\forall 1 \le l \le k$,

$$\sigma_l \overset{(*)}{>} \sigma_{l,\boldsymbol{W}}(t) \overset{(*)}{\ge} \sigma_{l,\boldsymbol{W}}\left(\frac{1}{2\sigma_l}\log\frac{C_\varepsilon}{\alpha}\right) \overset{(29)}{=} \sigma_l - \left(\sigma_l - \frac{C_\varepsilon\sigma_l}{(\sigma_l-\alpha)+C_\varepsilon}\right) \overset{(\star)}{\ge} \sigma_l - \frac{\sigma_l^2}{C_\varepsilon} \ge \sigma_l - \varepsilon\,, \tag{31}$$

and $\forall k < l \le K$,

$$0 \overset{(*)}{<} \sigma_{l,\boldsymbol{W}}(t) \overset{(*)}{\le} \sigma_{l,\boldsymbol{W}}\left(\frac{1}{2\sigma_l}\log\frac{c_\varepsilon}{\alpha}\right) \overset{(29)}{=} \frac{c_\varepsilon\sigma_l}{(\sigma_l-\alpha)+c_\varepsilon} \overset{(\star)}{\le} \varepsilon\,, \tag{32}$$

where $(*)$ uses the monotonicity of $\sigma_{i,\boldsymbol{W}}(t)$, and $(\star)$ uses the fact that $\sigma_l \ge \sigma_K \ge \varepsilon \ge \alpha, \forall l \le K$. Lastly, since we assumed $\alpha \le \varepsilon$, we have $0 < \sigma_{l,\boldsymbol{W}}(t) = \frac{\alpha}{1+2\alpha t} \le \varepsilon, \forall l > K$. The bound in (30) is verified. □

Incremental learning opens the possibility of obtaining low-rank approximations of $\boldsymbol{Y}$ (if desired) by adopting early stopping in standard GD/GF training without regularization. This result contrasts with that of the often-called "non-overparametrized" setting: Suppose one runs GF directly on $\boldsymbol{W}$ with the $\mathcal{L}(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{W}\|_F^2$ with small initialization $\boldsymbol{W} = \alpha\boldsymbol{W}_0$, the resulting solution is

$$\boldsymbol{W}(t) = \boldsymbol{Y} + (\alpha\boldsymbol{W}_0 - \boldsymbol{Y})e^{-t} \overset{(\alpha\ll 1)}{\simeq} \boldsymbol{Y}(1 - e^{-t})\,, \tag{33}$$

suggesting that all singular values of $\boldsymbol{Y}$ are being learned simultaneously at the same rate, thus no low-rank approximation along the GF trajectory. We illustrate this difference in Figure 1. A similar analysis applies to GF training of linear networks on linear regression ($\mathcal{A} : \boldsymbol{W} \mapsto \boldsymbol{X}\boldsymbol{W}$ for some input data matrix X, and $\boldsymbol{W} : (\boldsymbol{W}_1, \boldsymbol{W}_2) \mapsto \boldsymbol{W}_1\boldsymbol{W}_2$). This study of spectral initialization for linear networks started in an influential work of Saxe et al. [34] on deep learning theory, where the nonlinear dynamics of singular value growth and discussion around incremental learning have appeared.

We have discussed the GF trajectory under spectral initialization in detail. Does the incremental learning phenomenon still happen for arbitrary initialization shapes? In the latter case, despite the fact that $\boldsymbol{U}(t)$ initially does not have the right singular vectors $\boldsymbol{\Phi}$ of $\boldsymbol{Y}$ (for example, when all its entries are initialized

randomly from a Gaussian), it can learn $\boldsymbol{\Phi}$ through a *spectral learning* mechanism [29], [36] during the early stage of GF. Consider $\boldsymbol{U} = \alpha \boldsymbol{U}_0$ with small scale α. At the early stage of the training, we have

$$\dot{\boldsymbol{U}} = (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U} \simeq \boldsymbol{Y}\boldsymbol{U}\,. \tag{34}$$

Then, $\boldsymbol{U}(t)$ evolves approximately as a linear system whose solution leads to the following approximation:

$$\boldsymbol{U}(t) \simeq \boldsymbol{\Phi}\alpha e^{\Sigma_{\boldsymbol{Y}} t}\boldsymbol{\Phi}^\top \boldsymbol{U}_0 \overset{\left(t=\frac{1}{\sigma_1}\log\frac{c}{\alpha}\right)}{=} \boldsymbol{\phi}_1 c\left(\boldsymbol{\phi}_1^\top \boldsymbol{U}_0\right) + \sum_{k=2}^{K} \boldsymbol{\phi}_k c^{\frac{\sigma_k}{\sigma_1}} \alpha^{1-\frac{\delta_k}{\delta_1}} \boldsymbol{\phi}_k^\top \boldsymbol{U}_0\,, \tag{35}$$

where $\sigma_1, \cdots, \sigma_K$ are non-zero singular values of $\boldsymbol{Y}$ in decreasing order, and $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_k$ are the corresponding singular vectors. The approximation in (35) holds up to an $o(c^2)$ error when c is small and $\alpha \ll c$, whose exact derivation is omitted due to the space constraints. We shall focus on its implications: if the initialization shape $\boldsymbol{U}_0$ has some initial alignment with the top singular vector $\boldsymbol{\phi}_1$, i.e., $\boldsymbol{\phi}_1^\top \boldsymbol{U}_0 \neq 0$ and α is sufficiently small, then at time $\Theta\left(\frac{1}{\sigma_1}\log\frac{c}{\alpha}\right)$, $\boldsymbol{U}(t)$ is approximately rank-1 and the dominant left singular vector of $\boldsymbol{U}(t)$ is almost aligned with $\boldsymbol{\phi}_1$. Such alignment allows the top singular value of $\boldsymbol{W}(t) = \boldsymbol{U}(t)\boldsymbol{U}(t)^\top$ to learn the top singular component $\boldsymbol{\phi}_1\sigma_1\boldsymbol{\phi}_1$ of $\boldsymbol{Y}$ while the rest of its components remain small, which is formally shown in the work of Li et al. [29]. Afterward, the incremental learning proceeds as follows (a rough explanation): the learned top singular component ($\boldsymbol{\phi}_1\sigma_1\boldsymbol{\phi}_1$) in $\boldsymbol{U}(t)\boldsymbol{U}(t)^\top$ stays, the rest singular component of $\boldsymbol{U}(t)$ go through another spectral learning phase similar to (34), with $\boldsymbol{Y}$ replaced by the yet-to-learn residual $\boldsymbol{Y} - \boldsymbol{\phi}_1\sigma_1\boldsymbol{\phi}_1$, during which the second singular vector of $\boldsymbol{Y}$ is learned, followed by $\boldsymbol{U}(t)\boldsymbol{U}(t)^\top$ learning $\boldsymbol{\phi}_2\sigma_2\boldsymbol{\phi}_2$. This procedure continues until all singular components of $\boldsymbol{Y}$ are learned. Li et al. [29] refer to this procedure as "greedy low-rank learning" and provide partial theoretical support by characterizing the early GF trajectory of $\boldsymbol{W}(t)$ that learns the top singular component of $\boldsymbol{Y}$ from a small initialization. Concurrently to their work [29], this spectral learning phenomenon is studied for GD on the symmetric matrix sensing problem (with the linear measurement $\boldsymbol{\mathcal{A}}$ satisfying some Restricted Isometric Property (RIP)) in the work of Stöger and Soltanolkotabi [36], where many technical contributions on quantitative analyses on GD trajectories are introduced. Those contributions [29], [36] finally led to a rigorous proof for incremental learning in GD on symmetric matrix sensing problems with RIP measurements under non-spectral initialization shape [30].

Similar spectral learning phenomena have also been studied for GF with small initialization on two-layer ReLU networks [37], primarily for binary classification problems. The previously described early training phase when the weights learn some critical singular vectors associated with the training problem is called the *alignment phase*, during which the *neurons* (rows of the first-layer weights) align their directions with one of the two class centers, through a nonlinear directional dynamics (due to ReLU nonlinearity) that move every neuron toward the centroid of the data points that has a positive inner

product of the neuron, the challenges associated with analyzing such nonlinear dynamics is primarily on tracking the activation patterns of each neuron (w.r.t. training dataset) along GF trajectory [37]. After the alignment phase, the neurons grow their norm while keeping good alignment with the classes, akin to the phase of learning singular values of $\boldsymbol{Y}$ in the matrix sensing example, resulting in a low-rank weight matrix in the trained network.

Implicit bias in classification problems

So far, we have mostly discussed the implicit bias of GF/GD for regression problems due to their close relations to many important signal processing problems. Concurrently, significant progress has also been made toward understanding such bias when training overparametrized networks for classification problems. The venture begins with Soudry et al. [38] showing GD for logistic regression (more generally, classification problems with an exponential-like loss) on linearly separable data implicitly finding a linear classifier that corresponds to the global solution to the ℓ_2-max-margin problem. Successors extend the analysis to those overparametrized models we introduced when discussing implicit mirror flow, including diagonal linear networks [27], linear convolutional networks [27] , whose results are unified in the work of Yun et al. [26]. They show that GD on separable data with an overparmeterized model finds a linear classifier that corresponds to a KKT point of a max-margin problem w.r.t. some (quasi-)norm that depends on the overparametrization function. Notably, this implicit regularization promotes certain notions of sparsity when the underlying network is diagonal linear networks or linear convolutional networks [27]. Finally, the aforementioned analyses extend to any network that is positively homogeneous w.r.t. its weights [39], [40], showing that the weights converge in direction under GF/GD to a KKT point of an ℓ_2-max-margin problem. We refer the interested readers to the survey paper from Vardi [41] for a more detailed discussion on this line of work.

III. GD AT THE EDGE OF STABILITY

Most deep learning optimization theory prior to 2021, including all aforementioned works, focused on analyzing GD in the *small learning rate regime* in which the learning rate is pointwise upper-bounded by twice the reciprocal of the largest eigenvalue of the loss Hessian (frequently now called the *sharpness* in the DNN literature [42]). Convergence guarantees in this setting typically ensure *monotonic decrease* of the loss as a function of training time. This paradigm was shaken in 2021 with the publication of the work of Cohen et al. [43], which indicated empirically that the learning rates η used for training DNNs in practice were typically orders of magnitude larger than those required for monotonic decrease of the loss by GD. Whereas training with such a high learning rate would diverge for a quadratic objective, it

was shown in their work [43] that for DNNs, *despite* the possibility of "jumps" in the loss due to too high a learning rate, training nonetheless continues to converge to a global minimum. Following [43], this regime of non-monotonic convergence is now referred to as the *edge of stability* (EOS).

The natural question that arises is the following:

What is the implicit bias of GD with a large learning rate? How does this enable convergence despite non-monotonic loss decrease?

Two lines of work have been particularly successful in answering these questions. The question of *implicit bias* is addressed in the work of Damian et al. [42], which details a *self-stabilization* mechanism via which GD with a large learning rate is biased towards *flat* (i.e. low-curvature) regions of the loss landscape. While in principle this makes convergence *possible*, the problem of providing *convergence guarantees* in this regime was not addressed until the later works of Wu et al. [44] which, however, do not make explicit use of the self-stabilization mechanism of Damian et al. [42]. More recent work by MacDonald et al. [45] has shown that in fact self-stabilization *can* be used to prove convergence theorems by synthesizing Riemannian geometry and dynamical systems theory. This latter approach will be the main focus of this this tutorial, however for ease of presentation we will consider the following significantly simplified setting.

Definition 5. Let $\mathcal{L} : \mathbb{R}^2 \to \mathbb{R}$ be a smooth loss function. We will denote by $\lambda : \mathbb{R}^2 \to \mathbb{R}$ the largest eigenvalue field (sharpness) of $\nabla^2\mathcal{L}$ and by $\boldsymbol{u} : \mathbb{R}^2 \to \mathbb{R}^2$ the (possibly ill-defined) largest eigenvector field of $\nabla^2\mathcal{L}$. We initialize GD with large learning rate η at a point $\boldsymbol{\theta}_0 \in \mathbb{R}^2$ in an open neighborhood U of a 1-dimensional subspace M of $\mathbb{R}^2$ for which the following hold:

1) The sharpness λ is non-constant and continuously differentiable over all of M.
2) One has $\nabla\mathcal{L}(\boldsymbol{\theta}), \nabla\lambda(\boldsymbol{\theta}) \in M$ and $\langle \nabla\mathcal{L}(\boldsymbol{\theta}), \nabla\lambda(\boldsymbol{\theta}) \rangle \geq 0$ for all $\boldsymbol{\theta} \in M$.
3) The largest eigenvector $\boldsymbol{u}(\boldsymbol{\theta})$ is well-defined, constant ($= \boldsymbol{u}$) and orthogonal to M for all $\boldsymbol{\theta} \in M$.
4) There is $\epsilon \geq 0$ such that, for all, $\boldsymbol{\theta} \in U$, $\nabla\mathcal{L}(\boldsymbol{\theta})$ deviates no further than ϵ from its second-order Taylor expansion about the projection $\boldsymbol{\theta}^{\|}$ of $\boldsymbol{\theta}$ to M.

These assumptions are supposed to capture the intuitive picture of "valley" along which the sharpness can vary. GD in the EOS phase corresponds in this intuitive picture to learning-rate-dependent oscillations across the valley, with an implicit drift towards flatter regions of the valley wherein the oscillations are dampened (see Figure 2, bottom-right). Quadratic objectives, having constant sharpness λ fail Assumption 1: GD with step-size $\eta > 2/\lambda$ necessarily exhibits oscillations of exponentially increasing magnitude and diverges. On the other hand, the objective $\mathcal{L}(\boldsymbol{\theta}) = \exp(-\theta_1)\theta_2^2$, having a valley along which sharpness

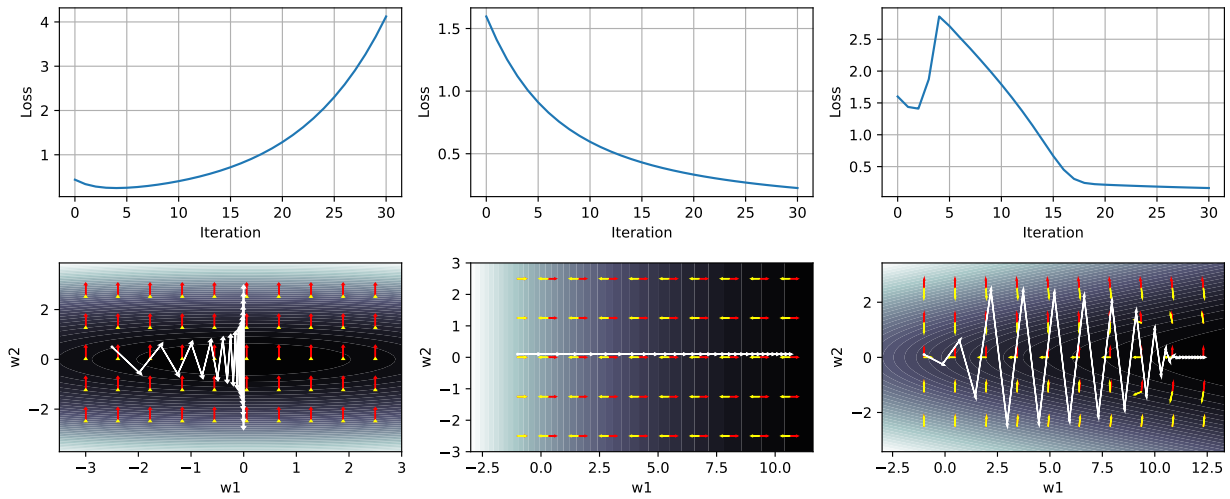

Fig. 2. Loss (top) and loss contour plot (bottom) with GD trajectory (white), largest Hessian eigenfield (red) and normalized sharpness gradient field (yellow) for a large-learning-rate-GD on a quadratic (left), and binary logistic regression with two antipodal datapoints (middle) and close-together datapoints (right). The quadratic objective has constant sharpness, so does not meet Assumption 1 of Definition 5, leading to divergence GD. The middle logistic regression, corresponding to two antipodal datapoints, fails both Assumptions 3 and 1 of Definition 5 (the largest Hessian eigenvector is everywhere parallel to the gradient of the sharpness), and GD with arbitrarily large step size converges without entering the EOS phase. The rightmost logistic regression, corresponding to two close-together datapoints, meets all assumptions (with M being the positive horizontal axis); the large learning rate causes initial increasing oscillations in the largest eigenvector direction and non-monotonic loss decrease, while the EOS mechanism pushes the trajectory towards flatter regions of the loss landscape, ultimately leading to convergence.

decreases monotonically, gives rise to the EOS phase of GD (as well as ultimate convergence).

A class of both examples and non-examples of EOS-inducing loss landscapes in machine learning is logistic regression on two datapoints $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of the same norm, with labels $+1$ and -1 respectively:

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2}\big(\ln(1+\exp(-\boldsymbol{x}_1^\top\boldsymbol{\theta})) + \ln(1+\exp(+\boldsymbol{x}_2^\top\boldsymbol{\theta})\big). \tag{36}$$

In this case, M is the max-margin line, which can be assumed without loss of generality to be the horizontal axis (Figure 2, right). The conditions of Definition 5 are all met when the angular separation ω of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ satisfies $0 < \omega < \pi/2$, leading to a widening valley along M about which EOS oscillations can occur (Figure 2, right); on the other hand, when $\boldsymbol{x}_1 = -\boldsymbol{x}_2$, there is no "valley" in which GD can oscillate, so will not enter the EOS phase (Figure 2, center).

Self-stabilization: a third-order mechanism for the EOS

The self-stabilization mechanism for the EOS phase proposed by Damian et al. [42] derives from a third-order Taylor expansion of the loss. Our own derivation of the mechanism, in line with that of

MacDonald et al. [45], is based on a slightly different intuition: while the Taylor expansions in their work [42] are taken around points of sharpness $2/\eta$, we instead Taylor expand around the bottom of the "valley" M in the loss landscape (Definition 5). This simplifies the derivation considerably, making it clear that the implicit bias towards flat regions arises from an implicit GD on the sharpness λ, as well as enabling transition to a simple convergence theorem in the case of the $\mathcal{L}(\boldsymbol{\theta}) = \exp(-\theta_1)\theta_2^2$ example.

We will illustrate the EOS mechanism for the GD iterates $\boldsymbol{\theta}_t$ in terms of their projections $\boldsymbol{\theta}_t^\perp$ and $\boldsymbol{\theta}_t^\parallel$ onto $M^\perp$ and M respectively, by Taylor expanding around $\boldsymbol{\theta}_t^\parallel$. As Theorem 6 below shows, these projections have quite different dynamics: $\boldsymbol{\theta}_t^\perp$ exhibits exponentially increasing oscillations due to too large a step size relative to the *second-order* derivative of $\mathcal{L}$; thanks to the *third-order* derivative of $\mathcal{L}$, on the other hand, $\boldsymbol{\theta}_t^\parallel$ evolves like GD on the sharpness λ, thus acting to dampen the oscillations.

Theorem 6. *Subject to the Assumptions of Definition 5, one has:*

$$\boldsymbol{\theta}_{t+1}^\perp = (1 - \eta\lambda(\boldsymbol{\theta}_t^\parallel))\boldsymbol{\theta}_t^\perp + \mathcal{O}(\eta\epsilon), \qquad \boldsymbol{\theta}_{t+1}^\parallel = \boldsymbol{\theta}_t^\parallel - \eta\left(\frac{1}{2}\|\boldsymbol{\theta}_t^\perp\|^2\nabla\lambda(\boldsymbol{\theta}_t^\parallel) + \nabla\mathcal{L}(\boldsymbol{\theta}_t^\parallel)\right) + \mathcal{O}(\eta\epsilon) \tag{37}$$

Proof. Given any $\boldsymbol{\theta}$, any $k \geq 1$ and tangent vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}$, we use $\nabla^k\mathcal{L}(\boldsymbol{\theta})[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}]$ to denote the multilinear map $\nabla^k\mathcal{L}(\boldsymbol{\theta})$ evaluated on the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}$. We invoke Assumption 4 to approximate $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla\mathcal{L}(\boldsymbol{\theta}_t)$ using the second order Taylor expansion of $\nabla\mathcal{L}(\boldsymbol{\theta}_t)$ about $\boldsymbol{\theta}_t^\parallel$:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\left(\nabla\mathcal{L}(\boldsymbol{\theta}_t^\parallel) + \nabla^2\mathcal{L}(\boldsymbol{\theta}_t^\parallel)[\boldsymbol{\theta}_t^\perp] + \frac{1}{2}\nabla^3\mathcal{L}(\boldsymbol{\theta}_t^\parallel)[\boldsymbol{\theta}_t^\perp, \boldsymbol{\theta}_t^\perp] + \mathcal{O}(\epsilon)\right), \tag{38}$$

Recalling that the parameter space is 2-dimensional, Assumption 3 guarantees that $\boldsymbol{\theta}_t^\perp$ is parallel to the largest eigenvector $\mathbf{u}$ of $\nabla^2\mathcal{L}(\boldsymbol{\theta}_t^\parallel)$, so one has

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\left(\nabla\mathcal{L}(\boldsymbol{\theta}_t^\parallel) + \lambda(\boldsymbol{\theta}_t^\parallel)\boldsymbol{\theta}_t^\perp + \frac{1}{2}\|\boldsymbol{\theta}_t^\perp\|^2\nabla^3\mathcal{L}(\boldsymbol{\theta}_t^\parallel)[\boldsymbol{u}, \boldsymbol{u}] + \mathcal{O}(\epsilon)\right). \tag{39}$$

Taking the gradient of the equation $\nabla^2\mathcal{L}[\boldsymbol{u}, \boldsymbol{u}] = \lambda$, one sees that $\nabla^3\mathcal{L}[\boldsymbol{u}, \boldsymbol{u}] = \nabla\lambda$, so that

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\left(\nabla\mathcal{L}(\boldsymbol{\theta}_t^\parallel) + \lambda(\boldsymbol{\theta}_t^\parallel)\boldsymbol{\theta}_t^\perp + \frac{1}{2}\|\boldsymbol{\theta}_t^\perp\|^2\nabla\lambda(\boldsymbol{\theta}_t^\parallel) + \mathcal{O}(\epsilon)\right). \tag{40}$$

Projecting onto $M^\perp$ and M respectively and invoking Assumption 2 then yield the claimed updates

$$\boldsymbol{\theta}_{t+1}^\perp = (1 - \eta\lambda(\boldsymbol{\theta}_t^\parallel))\boldsymbol{\theta}_t^\perp + \mathcal{O}(\eta\epsilon), \qquad \boldsymbol{\theta}_{t+1}^\parallel = \boldsymbol{\theta}_t^\parallel - \eta\left(\frac{1}{2}\|\boldsymbol{\theta}_t^\perp\|^2\nabla\lambda(\boldsymbol{\theta}_t^\parallel) + \nabla\mathcal{L}(\boldsymbol{\theta}_t^\parallel)\right) + \mathcal{O}(\eta\epsilon). \tag{41}$$

□

We now describe the self-stabilization mechanism induced by Theorem 6 in a similar manner to [42]. We assume for the sake of illustration that $\nabla\mathcal{L}(\boldsymbol{\theta}_t^\parallel) = 0$ for all t (as is the case for $\mathcal{L}(\theta_1, \theta_2) := \exp(-\theta_1)\theta_2^2$), and neglect the $\mathcal{O}(\eta\epsilon)$ terms. Then the self-stabilization mechanism can be understood in the following two phases:

1) **Instability:** Whenever $\eta > 2/\lambda(\boldsymbol{\theta}_t^{\parallel})$ one has $(1 - \eta\lambda(\boldsymbol{\theta}_t^{\parallel}))$ *strictly* less than -1. Thus if η remains larger than $2/\lambda(\boldsymbol{\theta}_t^{\parallel})$ over several time steps, $\boldsymbol{\theta}_{t+1}^{\perp} = (1 - \eta\lambda(\boldsymbol{\theta}_t^{\parallel}))\boldsymbol{\theta}_t^{\perp}$ exhibits *exponential growth* in magnitude with *oscillating sign* during this time (see Figure 3). Were $\lambda(\boldsymbol{\theta}_t^{\parallel})$ unable to change (as for a quadratic objective), these oscillations would ultimately lead to divergence.
2) **Return to stability:** The projection $\boldsymbol{\theta}_{t+1}^{\parallel} = \boldsymbol{\theta}_t^{\parallel} - \eta\frac{\|\boldsymbol{\theta}_t^{\perp}\|^2}{2}\nabla\lambda(\boldsymbol{\theta}_t^{\parallel})$ of the GD update to M is precisely GD on *the restriction of* λ *to* M with step-size proportional to $\|\boldsymbol{\theta}_t^{\perp}\|^2$. Exponential growth in $\|\boldsymbol{\theta}_t^{\perp}\|^2$ during the instability phase thus leads to increasingly *aggressive* steps in the descent direction of λ. Insofar as $\lambda(\boldsymbol{\theta}_t^{\parallel})$ decreases due to these descent steps, the factor $(1 - \eta\lambda(\boldsymbol{\theta}_t^{\parallel}))$ in the evolution of $\boldsymbol{\theta}_t^{\perp}$ decreases in magnitude, dampening the rate of increase of $\|\boldsymbol{\theta}_t^{\perp}\|$ (see again Figure 3). If eventually $\lambda(\boldsymbol{\theta}_t^{\parallel})$ decreases to below $2/\eta$, $\|\boldsymbol{\theta}_t^{\perp}\|$ will decrease, potentially enabling convergence.

Denoting $\alpha_t := \langle\boldsymbol{\theta}_t^{\perp}, \boldsymbol{u}\rangle$, the oscillation $\boldsymbol{\theta}_t^{\perp}$ relative to the largest eigenvector $\boldsymbol{u}$ of $\nabla^2\mathcal{L}$ along M, Theorem 6 can be used to predict the loss values $\mathcal{L}(\boldsymbol{\theta}_t)$ as well as α_t and $\lambda_t := \lambda(\boldsymbol{\theta}_t^{\parallel})$ in terms of the initial values α_0 and λ_0 and the values of $\mathcal{L}_t := \mathcal{L}(\boldsymbol{\theta}_t^{\parallel})$, $\nabla\mathcal{L}_t := \nabla\mathcal{L}(\boldsymbol{\theta}_t^{\parallel})$ and $\nabla\lambda_t := \nabla\lambda(\boldsymbol{\theta}_t^{\parallel})$ at M. Dropping remainder terms arising from Taylor expansions one has

$$\mathcal{L}(\boldsymbol{\theta}_t) \approx \mathcal{L}_t + \frac{\alpha_t^2}{2}\lambda_t, \qquad \alpha_{t+1} \approx (1 - \eta\lambda_t)\alpha_t, \qquad \lambda_{t+1} \approx \lambda_t - \eta\left(\frac{1}{2}\alpha_t^2\|\nabla\lambda_t\|^2 + \langle\nabla\mathcal{L}_t, \nabla\lambda_t\rangle\right) \quad (42)$$

Figure 3 plots these predictions against their true values for binary logistic regression (top) and $\mathcal{L}(\theta_1, \theta_2) = \exp(-\theta_1)\theta_2^2$ (bottom). Note the relatively poor accuracy of the approximation in the former case, due to the relative weakness of the third-order Taylor approximation of the loss. In the next section, we will see how additional assumptions on the properties of $\mathcal{L}$ and λ can be leveraged to prove a quantitative convergence theorem for GD in the EOS phase in Theorem 8 below.

Convergence theorems at the EOS

Although the work of Damian et al. [42] exhibits a mechanism for EOS, no attempt is made therein to prove a *convergence theorem* which takes EOS into account. Theorem 6 does, however, suggest how a convergence theorem might be possible: if the sharpness λ is monotonically decreasing to zero along the subspace M, then even if η is larger than the *initial* twice-reciprocal-sharpness $2/\lambda(\boldsymbol{\theta}_0^{\parallel})$, thus causing *initial* instability, the implicit GD on λ from Theorem 6 should force the algorithm into stability $(\eta < 2/\lambda(\boldsymbol{\theta}_t^{\parallel}))$ after some finite time t, thus enabling convergence.

The assumption that λ decreases monotonically along M is satisfied by the examples we have considered so far. Monotonic decrease of λ along a subspace is more generally satisfied by binary logistic regression on any linearly separable dataset, and this fact is used by Wu et al. [44] to derive EOS convergence theorems which guarantee *eventual* stability and convergence in this case. However, the

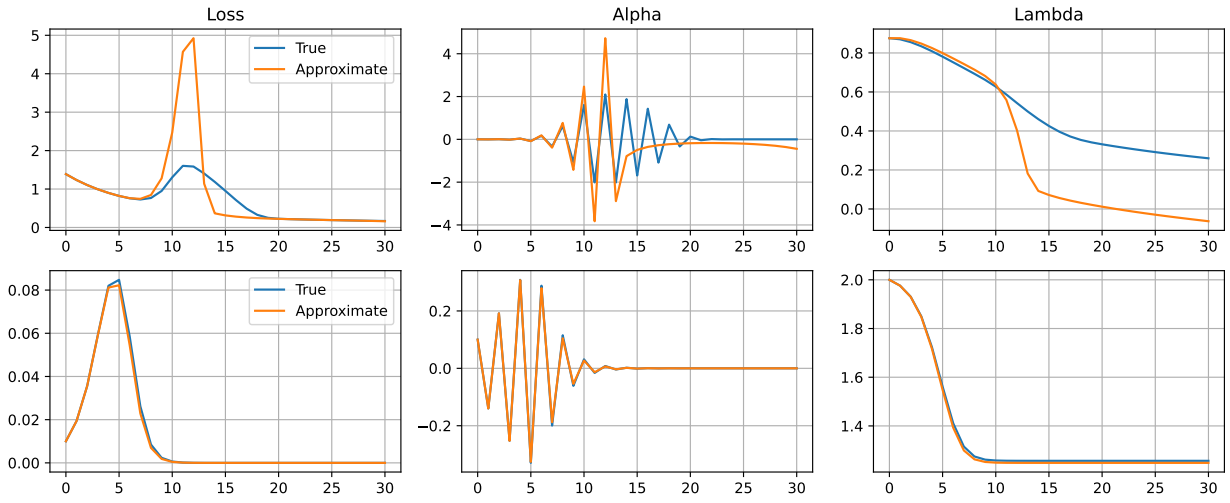

Fig. 3. Plots of approximate (orange) and true (blue) loss, sharpness λ_t and oscillation value $\alpha_t := (1-\eta\lambda_t)$ for binary logistic regression (top) and $\mathcal{L}(\theta_1,\theta_2) := \exp(-\theta_1)\theta_2^2$ (bottom). Note that the function $\mathcal{L}(\theta_1,\theta_2) = \exp(-\theta_1)\theta_2^2$ is identically quadratic in the direction of oscillation, resulting in a close match between approximate and true dynamics. The logistic regression loss, on the other hand, is *poorly* approximated by its cubic Taylor expansion in the direction of oscillation, leading to poor approximate dynamics. This mismatch was previously noted by Damian et al. [42], wherein it is fixed by including an additional term in the dynamics of α_t to account for the poor cubic approximation.

techniques used by Wu et al. [44] do not make use of the self-stabilization mechanism [42]. They are moreover specific to the (convex) logistic loss, and in particular do not extend to the (non-convex) objective $\mathcal{L}(\boldsymbol{\theta}) = \exp(-\theta_1)\theta_2^2$. In what follows, we will demonstrate that the self-stabilization equations derived in Theorem 6 can be used to derive an upper-bound on the duration of instability for the example $\mathcal{L}(\boldsymbol{\theta}) = \exp(-\theta_1)\theta_2^2$, thus enabling a convergence theorem in this *non-convex* case too. Since in this example $M = \mathrm{span}\{(1,0)\} \subset \mathbb{R}^2$, $\boldsymbol{\theta} = (\theta_1,\theta_2)$ decomposes into $\boldsymbol{\theta}^{\|} = (\theta_1, 0)$ and $\boldsymbol{\theta}^{\perp} = (0,\theta_2)$, with $\lambda(\boldsymbol{\theta}^{\|}) = 2\exp(-\theta_1)$. Motivated by this, we first prove the following lemma.

Lemma 7. *Let $(x_t)_{t=1}^{\infty}$ denote the iterates of gradient descent with step size η on the objective $\lambda(x) := 2\exp(-x)$ on $\mathbb{R}$. Then $\lambda(x_t) \leq \frac{\lambda(x_0)}{1+\lambda(x_0)\eta t}$ for all $t \in \mathbb{N}$.*

Proof. Given $t \in \mathbb{N}$, one has

$$\lambda(x_{t+1}) = 2\exp(-x_{t+1}) = 2\exp(-(x_t + 2\eta\exp(-x_t))) = \lambda(x_t)\exp(-\eta\lambda(x_t)) \leq \frac{\lambda(x_t)}{1+\eta\lambda(x_t)}. \quad (43)$$

Taking reciprocals, one has $1/\lambda(x_{t+1}) - 1/\lambda(x_t) \geq \eta$. Inductively it follows that $1/\lambda(x_t) \geq 1/\lambda(x_0) + \eta t$, from which the result follows. □

The theorem that follows proves convergence of GD in the EOS regime for $\mathcal{L}(\boldsymbol{\theta}) = \exp(-\theta_1)\theta_2^2$; it can be viewed as an easy version of the "subcritical" convergence theorem of MacDonald et al. [45].

Theorem 8. *Consider the objective $\mathcal{L}(\boldsymbol{\theta}) := \exp(-\theta_1)\theta_2^2$ on $\mathbb{R}^2$. Suppose that GD with any learning rate η is initialized at a point $\boldsymbol{\theta}_0$ at which $\eta > 2/\lambda(\boldsymbol{\theta}_0^{\|})$, and denote*

$$t_* := \left\lceil \frac{\eta\lambda(\boldsymbol{\theta}_0^{\|}) - 2}{\eta\lambda(\boldsymbol{\theta}_0^{\|})\|\boldsymbol{\theta}_0^{\perp}\|^2} \right\rceil + 1. \tag{44}$$

Then for any $s \geq t_*$, one has $\eta < 2/\lambda(\boldsymbol{\theta}_s^{\|}) < 2/\lambda(\boldsymbol{\theta}_{t_*}^{\|})$. Consequently, there is positive $\alpha \equiv \alpha(t_*) < 1$ such that $\{\boldsymbol{\theta}_t\}_{t=1}^{\infty}$ converges to a global minimum of $\mathcal{L}$ with rate $\mathcal{O}(\alpha^t)$.

Proof. Since $\mathcal{L}$ is precisely quadratic along any line orthogonal to $M = \text{span}\{(1,0)\} \subset \mathbb{R}^2$, one may globally take $\epsilon = 0$ in Assumption 4 of Definition 5, so the update equations of Theorem 6 reduce in this example to

$$\boldsymbol{\theta}_{t+1}^{\perp} = (1 - \eta\lambda(\boldsymbol{\theta}_t^{\|}))\boldsymbol{\theta}_t^{\perp}, \qquad \boldsymbol{\theta}_{t+1}^{\|} = \boldsymbol{\theta}_t^{\|} - \frac{\eta}{2}\|\boldsymbol{\theta}_t^{\perp}\|^2\nabla\lambda(\boldsymbol{\theta}_t^{\|}). \tag{45}$$

Denote by t_0 the earliest time t at which one has $\eta < 2/\lambda(\boldsymbol{\theta}_t^{\|})$.

Assume for a contradiction that $t_0 > t_*$. By the $\perp$-equation in (45) and the assumption that $\eta > 2/\lambda(\boldsymbol{\theta}_0^{\|})$, one then has $\|\boldsymbol{\theta}_t^{\perp}\| > \|\boldsymbol{\theta}_0^{\perp}\|$ for all $1 \leq t \leq t_*$. Since $\lambda : \boldsymbol{\theta}^{\|} = (\theta_1, 0) \mapsto \lambda(\boldsymbol{\theta}^{\|}) = 2\exp(-\theta_1)$ is monotonically decreasing along M, the value of λ at the t_* iterate obtained from GD on λ along M with *increasing* step size $\eta\|\boldsymbol{\theta}_t^{\perp}\|^2/2$ (as in the $\boldsymbol{\theta}^{\|}$ update in (45)) is strictly upper-bounded by its value at the t_*-iterate of GD on λ along M with the *constant* step size $(\eta/2)\|\boldsymbol{\theta}_0^{\perp}\|^2$; by Lemma 7, this yields

$$\lambda(\boldsymbol{\theta}_{t_*}^{\|}) < \frac{\lambda(\boldsymbol{\theta}_0^{\|})}{1 + \lambda(\boldsymbol{\theta}_0^{\|})\,(\eta/2)\,\|\boldsymbol{\theta}_0^{\perp}\|^2\,t_*} < \frac{2}{\eta}, \tag{46}$$

where the rightmost inequality follows from the definition of t_*. This contradicts the assumption that $t_0 > t_*$. It thus follows that $t_0 \leq t_*$ as claimed.

We now prove the convergence result. First, since the $\boldsymbol{\theta}^{\|}$ update in (45) is GD on the monotonically decreasing function λ along M, the sequence $\{\lambda(\boldsymbol{\theta}_t^{\|})\}_{t \geq t_*}$ is monotonically non-increasing. While this implies that $|1 - \eta\lambda(\boldsymbol{\theta}_t^{\|})| < 1$ for all $t \geq t_*$, the convergence theorem requires something stronger, namely $\alpha < 1$ such that $|1 - \eta\lambda(\boldsymbol{\theta}_t^{\|})| \leq \alpha$ for all $t \geq t_*$. To prove this stronger bound, we demonstrate a finite, uniform upper-bound on $\|\boldsymbol{\theta}_t^{\|}\|$ for all $t \geq t_*$. Consider the function $V : \mathbb{R}^2 \to \mathbb{R}$ defined by

$V(\boldsymbol{\theta}) := \|\boldsymbol{\theta}^{\perp}\|^2 + (4 - 2\eta\lambda(\boldsymbol{\theta}^{\|}_{t_*}))\|\boldsymbol{\theta}^{\|}\|$, noting that $\boldsymbol{\theta}^{\|}_{t_*}$ is *constant* here and not a function of the variable $\boldsymbol{\theta}$. Given $t \geq t_*$, denoting $V_t := V(\boldsymbol{\theta}_t)$ and $\lambda_t := \lambda(\boldsymbol{\theta}^{\|}_t)$, one sees using (45) that

$$\begin{aligned} V_{t+1} &= (1 - \eta\lambda_t)^2\|\boldsymbol{\theta}^{\perp}_t\|^2 + (4 - 2\eta\lambda_{t_*})\left(\|\boldsymbol{\theta}^{\|}_t\| + \frac{\eta\lambda_t\|\boldsymbol{\theta}^{\perp}_t\|^2}{2}\right) \\ &= V_t + \eta^2\lambda_t^2\|\boldsymbol{\theta}^{\perp}_t\|^2 - \eta^2\lambda_{t_*}\lambda_t\|\boldsymbol{\theta}^{\perp}_t\|^2 \\ &\leq V_t \end{aligned}$$

since λ_t is monotonically non-increasing. One deduces that $V_t \leq V_{t_*}$ for all $t \geq t_*$, hence $\|\boldsymbol{\theta}^{\|}_t\| \leq \frac{1}{4-2\eta\lambda_{t_*}}V_t \leq \frac{1}{4-2\eta\lambda_{t_*}}V_{t_*} = \|\boldsymbol{\theta}^{\|}_{t_*}\| + \frac{\|\boldsymbol{\theta}^{\perp}_{t_*}\|^2}{4-2\eta\lambda_{t_*}}$ for all $t \geq t_*$. Since λ_t is monotonically non-increasing, we deduce that

$$\tilde{\lambda}_{t_*} := \lambda\left(\boldsymbol{\theta}^{\|}_{t_*} + \frac{\|\boldsymbol{\theta}^{\perp}_{t_*}\|^2}{(4 - 2\eta\lambda_{t_*})}(1, 0)\right) \leq \lambda_t \leq \lambda_{t_*}$$

for all $t \geq t_*$. Finally, setting $\alpha := \max\{|1 - \eta\lambda_{t_*}|, |1 - \eta\tilde{\lambda}_{t_*}|\} < 1$, by the $\boldsymbol{\theta}^{\perp}$ update of (38) one has $\|\boldsymbol{\theta}^{\perp}_t\| \leq \alpha^{t-t_*}\|\boldsymbol{\theta}^{\perp}_{t_*}\|$ for all $t \geq t_*$, from which the result follows. □

We conclude this section by referring the interested reader to MacDonald et al. [45], wherein the ideas we have presented in this section are studied in substantially greater generality for certain least-squares problems. Therein, M is allowed to be a smooth manifold rather than a linear subspace, and the analogue of Theorem 6 exhibits an implicit Riemannian gradient descent on the sharpness along M. Moreover, [45] provides convergence proofs beyond the "eventually monotonic" regimes considered in Theorem 8 and by Wu et al. [44], including polynomial convergence to the optimally flat global minimum and exponential convergence to a periodic cycle centred on the optimally flat global minimum depending on the step size.

CONCLUSION

This paper presented an overview of recent progress on the convergence, implicit bias and edge of stability of gradient descent in deep learning. In Section I we focused on the convergence of gradient descent. We reviewed results showing that *overparametrization*, suitably characterized, implies that in regions where the derivative J_f of the map f from network parameters to network outputs is well-conditioned, any critical point of the loss is a global minimum and GD can be guaranteed to rapidly converge to such a minimum despite the non-convexity of the loss. Most works to date have invoked overparametrization in conjunction with an NTK style analysis that amounts to the impractical assumption that the GD dynamics are well-approximated by their linearization in a neighborhood of initialization. However, we expect that NTK analysis does not exhaust the utility of overparametrization. For instance, in the context of least squares regression, overparametrization in the sense of well-conditioned J_f implies,

by the regular value theorem, that the solutions form a smooth manifold; this fact has recently been used by MacDonald et al. [45] in EOS convergence analysis.

In Section II we focused on the implicit bias of gradient descent. Specifically, we studied recent work showing that the superior performance of neural networks in various learning problems is associated with some form of low-complexity regularization induced by the training algorithms. The success of existing works can be attributed to their identification of particular learning problems and network structures for which implicit regularization can be made explicit in the form of familiar norms known to promote sparsity (ℓ_1 norm) or rank deficiency (nuclear norm). However, it is more valuable in our view to answer these questions in reverse: given the network architectures that are known to work well in practice, how can we understand the implicit regularization therein by possibly identifying new measures for model complexity? We believe the signal processing community can bring valuable perspectives to these questions.

In Section III we focused on GD at the EOS, the regime most commonly used in practice. We studied recent works showing that higher order Taylor expansions reveal a self-stablization mechanism pushing large step size GD towards flatter regions of the loss landscape [42]. The bias of GD towards flatter regions of the loss landscape has been utilized in both logistic regression [44] and least squares problems [45] for convergence guarantees that apply even if the learning rate is larger than the threshold required for GD to monotonically decrease the objective value. While the theory of GD at the EOS is still in its very early stages, we anticipate that the self-stablization mechanism [42], [45] will play a key role in future analysis for more general problems than have so far been considered. An intriguing open question is how the implicit bias of large step size GD toward flat minima affects model complexity and generalization; we believe the perspective of the signal processing community could be valuable in addressing this question also.

REFERENCES

[1] P. Baldi, "Linear learning: Landscapes and algorithms," in *NeurIPS*, 1988.

[2] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *COLT*, 2016.

[3] K. Kawaguchi, "Deep learning without poor local minima," in *NeurIPS*, 2016.

[4] B. D. Haeffele and R. Vidal, "Global optimality in neural network training," in *CVPR*, 2017.

[5] P. Petersen, M. Raslan, and F. Voigtlaender, "Topological properties of the set of functions generated by neural networks of fixed size," *Foundations of computational mathematics*, vol. 21, no. 2, pp. 375–444, 2021.

[6] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *NeurIPS*, 2018.

[7] B. T. Polyak, *Introduction to Optimization*. Optimization Software, 1987.

[8] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *ICLR*, 2017.

[9] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, "Gradient descent finds global minima of deep neural networks," in *ICML*, 2019.

[10] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang, "On exact computation with an infinitely wide neural net," in *NeurIPS*, 2019.

[11] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *ICML*, 2019.

[12] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, "Wide neural networks of any depth evolve as linear models under gradient descent," in *NeurIPS*, 2019.

[13] L. Chizat, E. Oyallon, and F. Bach, "On lazy training in differentiable programming," in *NeurIPS*, 2019.

[14] G. Yang and E. J. Hu, "Feature learning in infinite-width neural networks," in *ICML*, 2021.

[15] S. Mei, T. Misiakiewicz, and A. Montanari, "Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit," in *COLT*, 2019.

[16] E. Moroshko, B. E. Woodworth, S. Gunasekar, J. D. Lee, N. Srebro, and D. Soudry, "Implicit bias in deep linear classification: Initialization scale vs training accuracy," in *NeurIPS*, 2020.

[17] L. MacDonald, H. Saratchandran, J. Valmadre, and S. Lucey, "On skip connections and normalisation layers in deep optimisation," in *NeurIPS*, 2023.

[18] S. Arora, N. Cohen, N. Golowich, and W. Hu, "A convergence analysis of gradient descent for deep linear neural networks," in *ICLR*, 2018.

[19] H. Min, S. Tarmoun, R. Vidal, and E. Mallada, "On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks," in *ICML*, 2021.

[20] S. Marcotte, R. Gribonval, and G. Peyré, "Abide by the law and follow the flow: conservation laws for gradient flows," in *NeurIPS*, 2023.

[21] H. Min, R. Vidal, and E. Mallada, "On the convergence of gradient flow on multi-layer linear models," in *ICML*, 2023.

[22] B. Bah, H. Rauhut, U. Terstiege, and M. Westdickenberg, "Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers," *Information and Inference: A Journal of the IMA*, vol. 11, no. 1, pp. 307–353, 2022.

[23] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[24] B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro, "Kernel and rich regimes in overparametrized models," in *COLT*, 2020.

[25] Y. Chen and Y. Chi, "Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 14–31, 2018.

[26] C. Yun, S. Krishnan, and H. Mobahi, "A unifying view on implicit bias in training linear neural networks," in *ICLR*, 2020.

[27] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro, "Implicit bias of gradient descent on linear convolutional networks," in *NeurIPS*, 2018.

[28] S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Implicit regularization in matrix factorization," in *NeurIPS*, 2017.

[29] Z. Li, Y. Luo, and K. Lyu, "Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning," in *ICLR*, 2021.

[30] J. Jin, Z. Li, K. Lyu, S. S. Du, and J. D. Lee, "Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing," in *ICML*, 2023.

[31] S. Azulay, E. Moroshko, M. S. Nacson, B. E. Woodworth, N. Srebro, A. Globerson, and D. Soudry, "On the implicit bias of initialization shape: Beyond infinitesimal mirror descent," in *ICML*, 2021.

[32] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[33] S. Arora, N. Cohen, W. Hu, and Y. Luo, "Implicit regularization in deep matrix factorization," *NeurIPS*, 2019.

[34] A. M. Saxe, J. L. Mcclelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural network," in *ICLR*, 2014.

[35] E. Abbe, S. Bengio, E. Boix-Adsera, E. Littwin, and J. Susskind, "Transformers learn through gradual rank increase," *NeurIPS*, vol. 36, 2023.

[36] D. Stöger and M. Soltanolkotabi, "Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction," in *NeurIPS*, 2021.

[37] H. Min, E. Mallada, and R. Vidal, "Early neuron alignment in two-layer ReLU networks with small initialization," in *ICLR*, 2024.

[38] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *JMLR*, vol. 19, no. 1, pp. 2822–2878, 2018.

[39] Z. Ji and M. Telgarsky, "Directional convergence and alignment in deep learning," in *NeurIPS*, 2020.

[40] K. Lyu and J. Li, "Gradient descent maximizes the margin of homogeneous neural networks," in *ICLR*, 2020.

[41] G. Vardi, "On the implicit bias in deep-learning algorithms," *Commun. ACM*, vol. 66, no. 6, p. 86–93, May 2023.

[42] A. Damian, E. Nichani, and J. Lee, "Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability," in *ICLR*, 2023.

[43] J. M. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar, "Gradient descent on neural networks typically occurs at the edge of stability," in *ICLR*, 2021.

[44] J. Wu, P. L. Bartlett, M. Telgarsky, and B. Yu, "Large Stepsize Gradient Descent for Logistic Loss: Non-Monotonicity of the Loss Improves Optimization Efficiency," in *COLT*, 2024.

[45] L. MacDonald, H. Min, L. Palma, S. Tarmoun, Z. Xu, and R. Vidal, "Convergence Rates for Gradient Descent at the Edge of Stability in Overparametrised Least Squares," in *NeurIPS*, 2025.

BIOGRAPHIES

Hancheng Min (*Member, IEEE*) received his B.S. degree in automation from Tongji University, Shanghai, China in 2016, M.S. degree in systems engineering from University of Pennsylvania in 2018, and Ph.D. degree in electrical and computer engineering from Johns Hopkins University in 2023. He was a Postdoc Researcher at the Center for Innovation in Data Engineering and Science (IDEAS), University of Pennsylvania from 2023 to 2025. He is currently an Associate Professor at the Institute of Natural Sciences (INS), the School of Mathematical Sciences (SMS), and the MOE-LSC at Shanghai Jiao Tong University. His primary research interest is deep learning theory, specifically training dynamics of gradient-based algorithms on neural networks. He is a Member of IEEE.

Lachlan MacDonald received his PhD in pure mathematics from the University of Wollongong in 2019 for his thesis on the noncommutative geometry of foliated manifolds. He has since held postdoctoral fellowships in pure mathematics at the Australian National University and the University of Adelaide, and in machine learning at the Australian Institute for Machine Learning (AIML), University of Adelaide, and the Mathematical Institute for Data Science (MINDS), Johns Hopkins University. He is now a postdoctoral fellow at the Center for Innovation in Data Engineering and Science (IDEAS), University of Pennsylvania. He has published articles in differential and noncommutative topology and geometry, as well as deep learning theory as it relates to equivariance and optimization.

René Vidal (*Fellow, IEEE*) received his B.S. degree in electrical engineering from the Pontificia Universidad Católica de Chile in 1997, and his M.S. and Ph.D. degrees in electrical engineering and computer science from the University of California at Berkeley in 2000 and 2003, respectively. He is currently the Rachleff University Professor in the Departments of Electrical and Systems Engineering and Radiology at the University of Pennsylvania, where he also directs the Center for Innovation in Data Engineering and Science (IDEAS). He has coauthored the book Generalized Principal Component Analysis (Springer, 2016) and more than 300 articles in machine learning, computer vision, biomedical image analysis, signal processing, robotics and control. He is a Fellow of IEEE.