Implicit biases in multitask and continual learning from a backward error analysis perspective

Benoit Dherin
Google
dherin@google.com

Abstract

Using backward error analysis, we compute implicit training biases in multitask and continual learning settings for neural networks trained with stochastic gradient descent. In particular, we derive modified losses that are implicitly minimized during training. They have three terms: the original loss, accounting for convergence, an implicit flatness regularization term proportional to the learning rate, and a last term, the conflict term, which can theoretically be detrimental to both convergence and implicit regularization. In multitask, the conflict term is a well-known quantity, measuring the gradient alignment between the tasks, while in continual learning the conflict term is a new quantity in deep learning optimization, although a basic tool in differential geometry: The Lie bracket between the task gradients.

1 Introduction

Overparameterized neural networks trained to interpolate are able to generalize surprisingly well in spite of the high complexity of their hypothesis space [1]. One key concept to understand this phenomenon is that of implicit regularization or implicit training biases, which are quantities that are not explicitly regularized in the loss during training but by other mechanisms, guiding the network toward simpler solutions [2, 3]. Several groups [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15] have recently used a technique called **Backward Error Analysis** (BEA) to compute implicit biases as a measure of the discrepancy between an optimizer iterates and the solutions of Gradient Flow (GF), which are the unique continuous paths of steepest descent. Because of its flexibility BEA has been used to compute optimizer implicit biases in many settings: Gradient Descent (GD) [4], Stochastic Gradient Descent (SGD) [5, 10], Momentum [16], Adam and RMSProp [14], GAN's [7, 11, 12], and diffusion processes [9], among others [6, 8, 13, 15].

Our contribution: We add to this body of work by computing implicit biases in multitask learning [17] and continual learning [18, 19] settings optimized with SGD. In both cases, the output of BEA is a modified loss implicitly minimized by the optimizer. It consists of the original loss plus additional terms, which can be split in two parts: 1) a beneficial implicit flatness regularizer proportional to the learning rate and already observed in single-task learning (in [4, 5] using BEA as well as with other approaches in [20, 21, 22]), and 2) a conflict term, due to the presence of several tasks, and which can be detrimental to both convergence and implicit flatness regularization. In multitask, the conflict term is the inner product between the task gradients, creating an implicit propensity in the learning dynamics to seek misaligned task gradients, which is known to be detrimental and needs to be mitigated [23, 24, 25]. In continual learning, the conflict term is the Lie bracket [26] between the task gradients whose non-vanishing may possibly be related to catastrophic forgetting [18, 19] where the performance of previous tasks degrades as new ones are learned. We hope to foster interest on Lie brackets in optimization, which is one of the basic tools in differential geometry [26].

2 Background on backward error analysis

To illustrate BEA, we now derive an implicit bias of SGD after a single mini-batch update by adapting the derivation for full-batch GD from [4]. Consider the loss $L_B(\theta)$ computed on a batch of data B from a dataset D. At a given step, the SGD iterate is $\theta' = \theta - h\nabla L_B(\theta)$, while the solution of GF (which exactly minimizes the batch loss) is the curve $\theta(t)$ solving the differential equation $\dot{\theta}(t) = -\nabla L_B(\theta(t))$ with $\theta(0) = \theta$. The **discretization drift** is the difference between the two, i.e., $\|\theta(h) - \theta'\|$, and it is of order $\mathcal{O}(h^2)$ for GD (see [27] for details). BEA proposes to compute a **modified equation**, in the form of GF plus corrections in terms of powers of the learning rate

$$\dot{\theta} = -\nabla L_B(\theta) + hf_1(\theta) + h^2 f_2(\theta) + \cdots \tag{1}$$

so that the solution $\tilde{\theta}(t)$ of the modified equation exactly coincides with the GD iterate: $\theta' = \tilde{\theta}(h)$. The idea of BEA is to use the continuous modified equation to analyze the discrete optimizer. Note that if we truncate the modified equation at order n (i.e. removing the terms of order h^n and higher), the discretization drift becomes of order only $\mathcal{O}(h^{n+1})$. Let us compute f_1 following [4]: First, we expand the solution of the modified equation in a Taylor series:

$$\tilde{\theta}(h) = \theta - h\nabla_\theta L_B(\theta) + h^2 \left(f_1(\theta) + \frac{1}{4}\nabla_\theta \|\nabla_\theta L_B(\theta)\|^2 \right) + \mathcal{O}(h^3) \tag{2}$$

For $\tilde{\theta}(h)$ to coincide with $\theta' = \theta - h\nabla L_B(\theta)$, we need all the terms in power of h^2 or higher to vanish. This gives us the first correction (and recursively the higher ones too as needed; see [4, 6, 13, 27]): $f_1(\theta) = -\nabla\left(\frac{1}{4}\|\nabla L_B(\theta)\|^2\right)$. Therefore the gradient update θ' follows a GF with drift only of order $\mathcal{O}(h^3)$ but for a modified loss $\widetilde{L}_B$, since the modified equation is of the form

$$\dot{\theta} = -\nabla \left(L_B(\theta) + \frac{h}{4}\|\nabla L_B(\theta)\|^2 \right) + \mathcal{O}(h^2), \quad \widetilde{L}_B(\theta) := L_B(\theta) + \frac{h}{4}\|\nabla L_B(\theta)\|^2, \tag{3}$$

The second term in this modified loss is a flatness bias, called **Implicit Gradient Regularization** (IGR) in [4], which prefers optimization paths with shallower slopes (i.e. lower gradients) guiding the trajectory toward flatter regions, very much in line with other flatness biases in SGD found by different means [21, 22, 28, 29]. We now turn to applying BEA to multitask and continual settings.

3 Modified loss and implicit biases in Multitask learning settings

Multitask learning trains a neural network jointly on several tasks hoping that knowledge gained from each task will transfer to the other tasks, helping generalization, and useful in case of data scarcity [30, 31]. However, it has been observed at times that learning multiple tasks at once can be detrimental, a circumstance attributed to the loss gradients for each task being misaligned [23, 24, 25, 32]. BEA shed some theoretical light on this since the implicit multitask dynamics of SGD given by its modified equation has a term, the conflict term, with propensity to guide the training in regions with misaligned gradients. In terms of losses, the simplest multitask setting corresponds to having two losses $L_1(\phi_1, \theta)$ for the first task and $L_2(\phi_2, \theta)$ for the second task. The parameters θ correspond to the part of the network that is shared between the two tasks, while ϕ_1 and ϕ_2 are the parameters corresponding to the two task heads. The training setup is to devise a global loss

$$L_{\alpha,\beta}(\phi_1, \phi_2, \theta) := \alpha L_1(\phi_1, \theta) + \beta L_2(\phi_2, \theta) \tag{4}$$

consisting on a weighted average of the two losses and update the network parameters with

$$\omega' = \omega - h\nabla_\omega L_{\alpha,\beta}(\phi_1, \phi_2, \theta), \quad \text{with} \quad \omega = (\phi_1, \phi_2, \theta). \tag{5}$$

Note that the update above can be considered either as a full-batch GD update as in [4], or a single-step batch update in SGD within an epoch as in Section 2.

Theorem 3.1. *At any given SGD step the two-task iterate* (5) *follows an exact GF* $\dot{\omega} = -\nabla_\omega \widetilde{L}_{\alpha,\beta}(\omega)$ *with a modified loss*

$$\widetilde{L}_{\alpha,\beta} = L_{\alpha,\beta} + \frac{h\alpha^2}{4}\|\nabla_{\omega_1} L_1\|^2 + \frac{h\beta^2}{4}\|\nabla_{\omega_2} L_2\|^2 + \frac{h\alpha\beta}{2}\langle \nabla_\theta L_1, \nabla_\theta L_2 \rangle, \tag{6}$$

with discretization drift $\|\tilde{\omega}(h) - \omega'\|$ *of order* $\mathcal{O}(h^3)$, *where* $\tilde{\omega}$ *is the solution of the modified GF starting at* ω *and where* $\omega_i := (\phi_i, \theta)$ *for* $i = 1, 2,$.

Proof. This follows from an immediate application of the supervised modified loss and modified equation in (3) to the special form of the multitask loss (4). $\qquad\square$

Interpretation: The modified loss (6) has two implicit biases: a IGR term and a conflict term

$$\text{IGR} = \frac{\alpha^2 h}{4}\|\nabla_{\omega_1} L_1\|^2 + \frac{\beta^2 h}{4}\|\nabla_{\omega_2} L_2\|^2, \qquad \text{conflict} = \frac{h\alpha\beta}{2}\langle \nabla_\theta L_1, \nabla_\theta L_2 \rangle.$$

The IGR term is beneficial: It consists of the sum of two implicit flatness regularizers for each task loss proportional to the learning rate h as in the single-task case, where it has been shown to be beneficial, guiding optimization paths toward flatter regions with greater generalization power [4, 5]. *The conflict term can be detrimental:* The implicit dynamics from the multitask modified equation encourages this term to become negative possibly at the expense of the original losses or the IGR terms. Regions where the conflict term can be negative are also regions where the gradients of the two losses w.r.t. the shared parameters are in opposite directions, creating smaller updates for the shared parameters, resulting in possibly stalled learning. It turns out that mechanisms preventing this conflict term to become negative (e.g. by projection [23, 24, 32] or direct regularization [25]) have been identified and used successfully to improve train and test performance in multitask settings.

4 Modified loss and implicit biases in continual learning settings

Continual learning is concerned with learning from a data distribution that is changing over time with tasks corresponding to locally stationary phases of the evolution [18, 19]. One of its major issues is catastrophic forgetting, when the updates from latter tasks degrade the performance on earlier tasks. While catastrophic forgetting is an issue for all modern approaches [18, 19, 33], its causes remain unclear. As we will see, the BEA modified equation for continual learning may help shed some new light on this issue. Namely, consider the continual learning setting where we perform two successive SGD updates $\theta_1 = \theta_0 - h\nabla L_1(\theta_0)$ and $\theta_2 = \theta_1 - h\nabla L_2(\theta_1)$, with the two losses computed on two successive batches from a changing data distribution. Using BEA, we want to first compute a modified loss whose continuous minimization approximates well the two successive updates. Then we want to identify possibly detrimental terms in the modified equation that may be responsible for a decreased performance on the first batch by the second update. It turns out that such a detrimental term pops up, controlled by the Lie bracket of the two batch gradients:

Definition 4.1. *Given two vector fields on $\mathbb{R}^n$, that is, two differentiable functions $F, G : \mathbb{R}^n \to \mathbb{R}^n$, their Lie Bracket is the vector field $[F, G] : \mathbb{R}^n \to \mathbb{R}^n$ defined as follows*

$$[F, G](\theta) = \nabla G(\theta)F(\theta) - \nabla F(\theta)G(\theta), \tag{7}$$

where $\nabla G(\theta)$ and $\nabla F(\theta)$ are the Jacobians of the vector fields.

Lie brackets are fundamental tools in differential geometry [26]. They help quantify how flows intertwine. For instance, if the Lie bracket between loss gradients for different tasks vanish, i.e., $[\nabla L_1, \nabla L_2] = 0$, this implies that their gradient flows *commute*: Following the gradient flow of first L_1 and then L_2 yields the same result as the reverse [26], with their flows somehow spanning "non-interacting" subspaces. The next theorem states that when this happens two consecutive SGD updates as above can be approximated by GF for a modified loss of the form:

$$\widetilde{L}_{1,2}(\theta) = L_1(\theta) + L_2(\theta) + \frac{h}{4}\|\nabla L_1(\theta)\|^2 + \frac{h}{4}\|\nabla L_2(\theta)\|^2, \tag{8}$$

where the IGR terms encourage the learning trajectory toward flatter regions for each task. Note that flatness preservation between tasks seems helpful to combat catastrophic forgetting [34, 35]. However, when $[\nabla L_1, \nabla L_2] \neq 0$, a term of order h in the modified equation (theorem below) and proportional to the Lie bracket can potentially disrupt that implicit flatness regularization induced by the modified loss above. Since it is the only term of order h that can do so, we conjecture that the non-vanishing of the Lie bracket between loss gradients pertaining to different tasks may be linked to catastrophic forgetting in continual learning. The following theorem gives an exact description of how this Lie bracket affects the implicit gradient regularization dynamics:

Theorem 4.2. *Consider two consecutive mini-batch gradient descent updates θ_1 and θ_2 as above. The solution $\tilde{\theta}(t)$ of the modified equation*

$$\dot{\theta}(t) = -\nabla\widetilde{L}_{1,2}(\theta(t)) + \frac{h}{2}[\nabla L_1, \nabla L_2](\theta(t)) + O(h^2), \tag{9}$$

where $\widetilde{L}_{1,2}$ is the modified loss in Equation (8) follows the composition iterate θ_2 with discretization drift $\|\tilde{\theta}(h) - \theta_2\|$ of order $\mathcal{O}(h^3)$.

3

Proof. To simplify the notation, let us start with two consecutive Euler updates for general vector fields F and G. First we consider an Euler update for the first vector field: $\theta_1 = \theta_0 + hF(\theta_0)$. Then, we compose this update with an Euler step in the direction of the second vector field G and expand the result into a Taylor's Series:

$$\begin{aligned}
\theta_2 &= \theta_1 + hG(\theta_1) \\
&= \theta_0 + hF(\theta_0) + hG(\theta_0 + hF(\theta_0)) \\
&= \theta_0 + h(F(\theta_0) + G(\theta_0)) + h^2 \nabla G(\theta_0)F(\theta_0) + \mathcal{O}(h^3).
\end{aligned}$$

Now, we want to find a modified equation of the form

$$\dot{\theta} = H_0(\theta) + hH_1(\theta) + h^2 H_2(\theta) + \cdots \tag{10}$$

whose solution starting at θ_0 coincides with θ_2 after time h. For that, we can compute the Taylor expansion of the modified equation solution and compare the powers in h to obtain recursive formulas for the H_i's. It is easy to verify the the first orders of the solution Taylor's Series are given by the following expression:

$$\theta(h) = \theta_0 + hH_0(\theta_0) + h^2 (H_1(\theta_0) + \frac{1}{2}\nabla H_0(\theta_0)H_0(\theta_0)) + \mathcal{O}(h^3). \tag{11}$$

To have that $\theta_2 = \theta(h)$ at first order, we obtain the following condition: $H_0(\theta_0) = F(\theta_0) + G(\theta_0)$. This yields for the second order the following condition:

$$H_1(\theta_0) + \frac{1}{2}\nabla H_0(\theta_0)H_0(\theta_0) = \nabla G(\theta_0)H(\theta_0). \tag{12}$$

Using the first order condition and expanding, we immediately obtain

$$H_1(\theta_0) = -\frac{1}{2}\left(\nabla F(\theta_0)F(\theta_0) + \nabla G(\theta_0)G(\theta_0)\right) + \frac{1}{2}[F, G](\theta_0), \tag{13}$$

where the last term is the Lie bracket between F and G evaluated at θ_0. Now if we specialized for the gradient fields $F = -\nabla L_1$ and $G = -\nabla L_2$, we obtain that

$$\begin{aligned}
H_0(\theta) &= -\nabla(L_1 + L_2) \tag{14} \\
H_1(\theta) &= -\nabla\left(\frac{1}{4}\|\nabla L_1\|^2 + \frac{1}{4}\|\nabla L_2\|^2\right) + \frac{1}{2}[\nabla L_1, \nabla L_2], \tag{15}
\end{aligned}$$

which concludes the theorem. $\qquad\square$

Remark 4.3. *Observe that when the two losses L_1 and L_2 come from batches pertaining to the same task (i.e., close to i.i.d.) their gradients are more likely to be aligned. By the anti-symmetry of the Lie bracket, $[\nabla L_1, \nabla L_2]$ is then more likely to be close to zero. However, when the data distribution changes, creating a sharp contrast between the two task loss-gradients then the Lie bracket $[\nabla L_1, \nabla L_2]$ is likely to be large. This seems to be relevant to the* stability gap *noticed in [19], when a large and sudden decrease in performance is observed after the first update for the second task.*

5 Conclusion

We computed implicit biases in multitask and continual learning optimized with SGD using backward error analysis. These biases are local, measuring the discrepancy between one step of SGD and gradient flow on the batch loss. In both cases we found a beneficial flatness bias proportional to the learning rate and preferring smaller slopes on the loss surface for each task along the learning trajectories similar to single-task supervised learning [4, 5]. We also found a detrimental implicit bias in both cases (due to the presence of several tasks and which we called conflict term) that has the potential to steer the learning dynamics away from the flatter regions with higher generalization power. For multitask learning, the detrimental implicit bias is controlled by the inner product of the task loss-gradients $\langle \nabla L_1, \nabla L_2 \rangle$, which is a known key quantity in multitask learning already (e.g., [23, 24, 25, 32]). For continual learning the detrimental bias is a new quantity, the Lie bracket $[\nabla L_1, \nabla L_2]$ between the task loss-gradients measuring how much their respective gradient flows span independent regions of the parameter space. Despite their wide use in many areas of mathematics, Lie brackets are new to deep learning optimization to the best of our knowledge. We hope this work will help foster the use of backward error analysis in deep learning, and serve as a theoretical motivation to devise methods relying on Lie brackets in continual learning.

Acknowledgments and Disclosure of Funding

References

[1] Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

[2] Benoit Dherin, Michael Munn, Mihaela Rosca, and David GT Barrett. Why neural networks find simple solutions: The many regularizers of geometric complexity. In *NeurIPS*, 2022.

[3] Benoit Dherin, Micheal Munn, and David G. T. Barrett. The geometric occam's razor implicit in deep learning. In *NeurIPS 2021, 13th Annual Workshop on Optimization for Machine Learning*, 2021.

[4] David G.T. Barrett and Benoit Dherin. Implicit gradient regularization. In *ICLR*, 2021.

[5] Samuel L Smith, Benoit Dherin, David G.T. Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *ICLR*, 2021.

[6] Mihaela Rosca, Yan Wu, Chongli Qin, and Benoit Dherin. On a continuous time model of gradient descent dynamics and instability in deep learning. In *TMLR*, 2023.

[7] Mihaela Rosca, Yan Wu, Benoit Dherin, and David G.T. Barrett. Discretization drift in two-player games. In *ICML*, 2021.

[8] Guilherme França, Michael I Jordan, and René Vidal. On dissipative symplectic integration with applications to gradient-based optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(4), 2021.

[9] Yansong Gao, Pan Zhibong, Xin Zhou, Le Kang, and Pratik Chaudhari. Fast diffusion probabilistic model sapling through the lens of backward error analysis. *arXiv:2304.11446*, 2023.

[10] Qianxiao Li, Cheng Tai, et al. Stochastic modified equations and adaptive stochastic gradient algorithms. In *ICML*, 2017.

[11] Haihao Lu. An o(sr)-resolution ode framework for understanding discrete-time algorithms and applications to the linear convergence of minimax problems. *Math. Program.*, 194(1–2), 2022.

[12] Mihaela Rosca and Marc Deisenroth. Implicit regularisation in stochastic gradient descent: from single-objective to two-player games. In *HLD 2023: 1st Workshop on High-dimensional Learning Dynamics*, 2023.

[13] Taiki Miyagawa. Toward equation of motion for deep neural networks: Continuous-time gradient descent and discretization error analysis. In *NeurIPS*, 2022.

[14] Matias Cattaneo, Jason Klusowski, and Boris Shigida. On the implicit bias of adam. *arXiv:2309.00079*, 2023.

[15] Luis Barba, Martin Jaggi, and Yatin Dandi. Implicit gradient alignment in distributed and federated learning. In *AAAI Conference on Artificial Intelligence*, AAAI'22, 2021.

[16] Avrajit Ghosh, He Lyu, Xitong Zhang, and Rongrong Wang. Implicit regularization in heavy-ball momentum accelerated stochastic gradient descent. *ICLR*, 2023.

[17] Liyuan Wang, Zingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *arXiv:2302.00487*, 2023.

[18] Albin Soutif-Cormerais, Antonio Carta, Andrea Cossu, Julio Hurtado, Hamed Hemati, Vincenzo Lomonaco, and Joost Van de Weijer. A comprehensive empirical evaluation on online continual learning, 2023.

[19] Matthias De Lange, Gido M van de Ven, and Tinne Tuytelaars. Continual evaluation for lifelong learning: Identifying the stability gap. In *ICLR*, 2023.

[20] Gal Vardi and Ohad Shamir. Implicit regularization in relu networks with the square loss. In *Conference on Learning Theory*, 2021.

[21] Alex Damian, Tengyu Ma, and Jason D. Lee. Label noise SGD provably prefers flat global minimizers. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *NeurIPS 2021*, 2021.

[22] Chao Ma and Lexing Ying. On linear stability of sgd and input-smoothness of neural networks. In *NeurIPS*, 2021.

[23] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *NeurIPS*, 2020.

[24] Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *ICLR*, 2021.

[25] Seanie Lee, Hae Beom Lee, Juho Lee, and Sung Ju Hwang. Sequential reptile: Inter-task gradient alignment for multilingual learning. In *ICLR*, 2022.

[26] John Lee. Introduction to smooth manifolds. In *Springer*, 2022.

[27] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, 2006.

[28] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Annual Conference Computational Learning Theory*, 2019.

[29] Stanislaw Kamil Jastrzebski, Devansh Arpit, Oliver Åstrand, Giancarlo Kerg, Huan Wang, Caiming Xiong, richard socher, Kyunghyun Cho, and Krzysztof J. Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *ICML*, 2021.

[30] Sebastian Ruder. An overview of multi-task learning in deep neural networks, 2017.

[31] Yu Zhang and Qiang Yang. A survey on multi-task learning, 2021.

[32] Drago Anguelov, Henrik Kretzschmar, Jiquan Ngiam, Thang Luong, Yanping Huang, Yuning Chai, and Zhao Chen. Just pick a sign: Reducing gradient conflict in deep networks with gradient sign dropout. In *NeurIPS*, 2020.

[33] ameya prabhu, Torr philip, and Dokania Puneet. Gdumb: A simple approach that questions our progress in continual learning. In *ECCV 2020*, 2020.

[34] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An empirical investigation of the role of pre-training in lifelong learning. *Journal of Machine Learning Research*, 24(214):1–50, 2023.

[35] Wenhang Shi, Yiren Chen, Zhe Zhao, Wei Lu, Kimmo Yan, and Xiaoyong Du. Create and find flatness: Building flat training spaces in advance for continual learning. *arXiv preprint arXiv:2309.11305*, 2023.