Leveraging Continuous Time to Understand Momentum When Training Diagonal Linear Networks

Hristo Papazov[*]
EPFL
hristo.papazov@epfl.ch

Scott Pesme[*]
EPFL
scott.pesme@epfl.ch

Nicolas Flammarion
EPFL
nicolas.flammarion@epfl.ch

Abstract

In this work, we investigate the effect of momentum on the optimisation trajectory of gradient descent. We leverage a continuous-time approach in the analysis of momentum gradient descent with step size γ and momentum parameter β that allows us to identify an intrinsic quantity $\lambda = \frac{\gamma}{(1-\beta)^2}$ which uniquely defines the optimisation path and provides a simple acceleration rule. When training a 2-layer diagonal linear network in an overparametrised regression setting, we characterise the recovered solution through an implicit regularisation problem. We then prove that small values of λ help to recover sparse solutions. Finally, we give similar but weaker results for stochastic momentum gradient descent. We provide numerical experiments which support our claims.

1 Introduction

Momentum methods (Sutskever et al., 2013) have now become a staple of optimal neural network training due to the provided gains in both optimisation efficiency and generalisation performance. This pivotal role is underscored by the widespread use of momentum in the successful training of most state-of-the-art deep networks, including CLIP (Radford et al., 2021), Chinchilla (Hoffmann et al., 2022), GPT-3 (Brown et al., 2020), and PaLM (Chowdhery et al., 2022).

Originating in the work of Polyak (1964), momentum first featured in the heavy-ball method devised to accelerate convergence in convex optimisation. However, when applied to neural network training, momentum exhibits a distinct and complementary characteristic: a steering towards models with superior generalisation performance compared to networks trained with gradient descent. We note that while the effect of momentum on optimisation has been researched extensively (Defazio, 2020; Sun et al., 2019), the generalisation aspect of momentum has been left relatively underexplored.

The performance of gradient descent methods presents intriguing challenges from a theoretical perspective. First, establishing convergence is highly non-trivial. Second, the existence of numerous global minima for the training objective, some of which generalise poorly, adds to the puzzle (Zhang et al., 2017). To elucidate this second point, the notion of implicit regularisation has come to the forefront. It posits that the optimisation process implicitly favors solutions with strong generalisation properties, even in the absence of explicit regularisation. The canonical example is overparametrised linear regression with more trainable parameters than the number of samples. While there exist infinitely many solutions that fit the data, gradient methods navigate in a restricted parameter subspace and converge towards the solution closest in terms of the ℓ_2 distance (Lemaire, 1996).

In this work, we aim to expand our understanding of the implicit bias of momentum by analysing its impact on the optimisation trajectory in 2-layer diagonal linear networks. The 2-layer diagonal linear network has garnered significant attention recently (Woodworth et al., 2020; Vaškevičius et al., 2019; HaoChen et al., 2021; Pesme et al., 2021; Pillaud-Vivien et al., 2022). Despite its apparent simplicity, this network has surprisingly shed light on training behaviours typically associated with much more complex architectures. Some of these insights include the influence of initialisation (Woodworth et al., 2020), the impact of noise (Pesme et al., 2021), and the role of the step size (Even et al., 2023). Consequently, this architecture serves as an excellent surrogate model for gaining a deeper understanding of intricate phenomena such as the role of momentum in the generalisation performance.

[*]Equal contribution. Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1.1 Main Contributions

In this paper, we investigate the influence of momentum on the optimisation trajectory of neural networks trained with momentum gradient descent (MGD). Leveraging the continuous-time approximation of MGD – momentum gradient flow (MGF), we show that the optimisation trajectory strongly depends on the key quantity $\lambda = \frac{\gamma}{(1-\beta)^2}$, where γ and β denote the step size and momentum parameter of MGD, respectively. Surprisingly, this continuous-time framework experimentally proves to be a good approximation of the discrete trajectory even for large values of γ.

We proceed to list our main contributions.

- First, using the key quantity λ, we derive a straightforward acceleration rule that maintains the optimisation path while accelerating the optimisation speed.
- Then, focusing on MGF on 2-layer diagonal linear networks, we precisely characterise the recovered solution and prove that for suitably small values of λ, MGF recovers solutions which generalise better than the ones selected by gradient flow (GF) in a sparse regression setting.
- Finally, we provide similar but slightly weaker results for stochastic MGD.

1.2 Related Works

Momentum and Acceleration. Momentum algorithms have their roots in acceleration methods, and many studies have investigated their convergence speed when optimising both convex and non-convex functions: (Ghadimi et al., 2015; Flammarion and Bach, 2015; Kidambi et al., 2018; Can et al., 2019; Sebbouh et al., 2021; Mai and Johansson, 2020; Liu et al., 2020; Cutkosky and Mehta, 2020; Defazio, 2020; Orvieto et al., 2020; Sebbouh et al., 2021). Moreover, apart from accelerating training, heavy-ball methods come with the additional advantage of always escaping saddle points (Jin et al., 2018; Sun et al., 2019).

Momentum and Continuous-Time Models. Building upon the foundational work of Alvarez (2000); Attouch et al. (2000), researchers have analysed accelerated gradient methods using second-order differential equations. Su et al. (2014) extended the previous ODE to encompass the Nesterov accelerated method, demonstrating convergence rates similar to the discrete case. Wibisono et al. (2016) adopted a variational perspective to scrutinise the mechanics of acceleration. A significant advancement emerged with the introduction of Lyapunov analysis, undertaken by Wilson et al. (2021); Sanz Serna and Zygalakis (2021); Moucer et al. (2023). This analytical approach sheds light on the stability and convergence properties of these methods. Further refinement has been achieved by Shi et al. (2021), who developed high-resolution ODEs tailored to various momentum-based acceleration techniques and able to distinguish between Nesterov's Accelerated Gradient and Polyak's Heavy Ball methods. Finally, error bounds for the discretisation of MGF have been developed by Kovachki and Stuart (2021).

Momentum and Implicit Bias. Sutskever et al. (2013); Leclerc and Madry (2020) have empirically shown significant generalisation improvements in architectures trained with momentum on common vision tasks. Building on these empirical observations, Jelassi and Li (2022) designed a synthetic binary classification problem where a 2-layer convolutional neural network trained with MGD provably generalises better than gradient descent (GD). Recently, Ghosh et al. (2023) reveal that the MGD trajectory closely resembles the gradient flow trajectory of a regularised loss. Through the specific regularisation, the authors argue that the MGD trajectory favors flatter minima than the GD trajectory. The study's findings apply to any reasonable loss, but due to the finite time horizon restriction, cannot characterise the solution to which MGD converges. Additionally, Wang et al. (2023) show that in deep diagonal linear networks with identical weights across layers, increasing the depth biases the optimisation towards sparse solutions.

2 From Discrete to Continuous

Momentum Gradient Descent. We consider minimising a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ using *momentum gradient descent* (MGD) with step size $\gamma > 0$ and momentum parameter $\beta \in [0, 1)$. Initialised at two points $(w_0, w_1) \in \mathbb{R}^{2d}$, the iterates follow the discrete recursion for $k \geq 1$:

$$w_{k+1} = w_k - \gamma \nabla F(w_k) + \beta(w_k - w_{k-1}). \quad (\mathrm{MGD}(\gamma, \beta))$$

Momentum Gradient Flow. Directly analysing the discrete recursion $\mathrm{MGD}(\gamma, \beta)$ appears intractable in many settings. To overcome this difficulty, we follow the classical approach of considering a second order differential equation of the form

$$a\ddot{w}_t + b\dot{w}_t + \nabla F(w_t) = 0 \tag{1}$$

with leading coefficient $a \geq 0$ and damping coefficient $b > 0$. In fact, without loss of generality, the previous differential equation can be reduced to a new one which depends on a single parameter λ. Indeed, assume that w_t follows ODE (1) with initialisation $(w_{t=0}, \dot{w}_{t=0}) = (w_0, \dot{w}_0)$, then a simple chain rule shows that $\tilde{w}_t = w_{bt}$

follows

$$\frac{a}{b^2}\ddot{\tilde{w}}_t + \dot{\tilde{w}}_t + \nabla F(\tilde{w}_t) = 0,$$

with initialisation $(\tilde{w}_{t=0}, \dot{\tilde{w}}_{t=0}) = (w_0, b\dot{w}_0)$. Hence, up to a time reparametrisation, it is sufficient to consider the following differential equation which depends on a unique parameter $\lambda \geq 0$:

$$\lambda \ddot{w}_t + \dot{w}_t + \nabla F(w_t) = 0. \qquad (\mathrm{MGF}(\lambda))$$

We call the differential equation MGF(λ) *momentum gradient flow* (MGF) with parameter λ. To show the link with the MGD(γ, β) recursion, we discretise MGF(λ) with a second-order central difference, first-order backward difference, and discretisation step $\varepsilon > 0$ as carried out by Kovachki and Stuart (2021):

$$\lambda\, \frac{w_{k+1} - 2w_k + w_{k-1}}{\varepsilon^2} + \frac{w_k - w_{k-1}}{\varepsilon} + \nabla F(w_k) = 0. \qquad (2)$$

Rewriting, we obtain

$$w_{k+1} = w_k - \frac{\varepsilon^2}{\lambda}\nabla F(w_k) + (1 - \frac{\varepsilon}{\lambda})(w_k - w_{k-1}),$$

which corresponds to momentum gradient descent with parameters $\gamma = \frac{\varepsilon^2}{\lambda}$ and $\beta = 1 - \frac{\varepsilon}{\lambda}$. Solving for ε and λ leads to the following proposition:

Proposition 1. *For* $(w_0, w_1) \in \mathbb{R}^{2d}$, *consider momentum gradient flow* MGF(λ) *with*

$$\lambda = \frac{\gamma}{(1-\beta)^2}$$

and initialisation $w_{t=0} = w_0$, $\dot{w}_{t=0} = (w_1 - w_0)/\sqrt{\lambda\gamma}$. *Then, discretising as* (2) *with discretisation step* $\varepsilon = \sqrt{\lambda\gamma} = \gamma/(1-\beta)$ *leads to the momentum gradient descent recursion* MGD(γ, β) *with step size* γ*, momentum parameter* β*, and initialisation* (w_0, w_1).

Proposition 1 motivates studying MGF(λ) as a continuous proxy for MGD(γ, β) assuming that the discretisation (2) closely approximates the continuous path.

Discretisation Error Bounds. Unfortunately, applying known discretisation error bounds to our setting leads to very pessimistic bounds. Indeed, for step size γ and momentum parameter β, consider the iterates w_k from MGD(γ, β) initialised at (w_0, w_1). Now, let $w(t)$ be the solution of MGF(λ) with $\lambda = \gamma/(1-\beta)^2$ and the appropriate initialisation from Proposition 1. Then, for a finite horizon $K > 0$, classical discretisation error bounds (see Kovachki and Stuart (2021), Theorem 4) lead to a catastrophic

$$\sup_{k \leq K} \|w_k - w(k\varepsilon)\| \leq \exp(CK)\varepsilon,$$

where the constant C depends on λ and F. Such an exponential dependence in the time horizon K questions the suitability of momentum gradient flow as a good proxy for momentum gradient descent. However, empirically, the above bound appears excessively pessimistic (see Figure 1: Left and Middle). The MGF and MGD trajectories behave similarly in various settings, even with non-convex losses F and relatively large step sizes γ (see Appendix F for additional experiments).

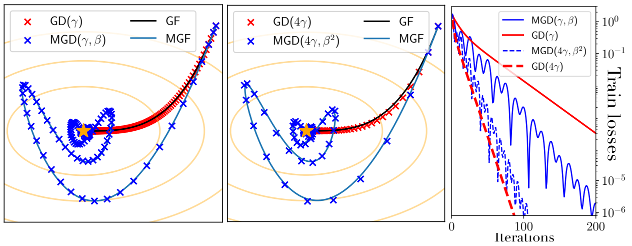

Figure 1: (M)GD over a 2D quadratic. *Left and Middle*: The (M)GD trajectories closely follow the continuous trajectories of (M)GF as suggested by Proposition 1. *Right*: MGD($4\gamma, \beta^2$) follows the same trajectory as MGD(γ, β) but twice as fast as suggested by Corollary 1. In contrast, GD(4γ) runs four times faster than GD(γ).

Intertwined Roles of γ and β. When the discretisation accurately follows the continuous path, Proposition 1 implies that the trajectory of MGD(γ, β) is solely determined by a single parameter $\lambda = \gamma/(1-\beta)^2$, intertwining step size and momentum as observed in Figures 1 and 2. **Consequently, γ and β serve interchangeable roles in influencing the trajectory of** MGD(γ, β)**.** Note that this single-parameter dependence aligns with empirical results from Leclerc and Madry (2020) where generalisation performance with large step sizes can be replicated with momentum and smaller step sizes. Though the quantity $\gamma/(1-\beta)^2$ spontaneously appears in works studying MGD (Ghosh et al., 2023), to the best of our knowledge, its natural presence was never clearly explained and motivated.

MGD Acceleration Rule. Though all couples (γ, β) with the same same value of λ yield the same trajectory, the iterates do not follow this path at the same speed.

Corollary 1 (Acceleration rule)**.** *Let* MGD(γ, β) *initialised at* $w_0 = w_1 \in \mathbb{R}^d$ *correspond to the discretisation of* MGF(λ) *with discretisation step* ε*. Now, for* $\rho \in \mathbb{R}_{>0}$*, consider the different parameter couple*

$$\hat{\gamma} = \rho^2\gamma \quad \textit{and} \quad \hat{\beta} = 1 - \rho(1-\beta) \approx_{\beta \to 1} \beta^\rho.^1$$

Then, since $\hat{\gamma}/(1-\hat{\beta})^2 = \lambda$, MGD($\hat{\gamma}, \hat{\beta}$) *initialised at* $w_0 = w_1$ *becomes the discretisation of the same* MGF(λ) *but with discretisation step* $\hat{\varepsilon} = \rho \cdot \varepsilon$.

Following the notations of the previous corollary for an integer $\rho \geq 2$ and letting w_k and $\hat{w}_k$ denote the iterates

[1] The approximation symbol abbreviates the Taylor-expansion bound $1 - \rho(1-\beta) = \beta^\rho + O((1-\beta)^2)$.

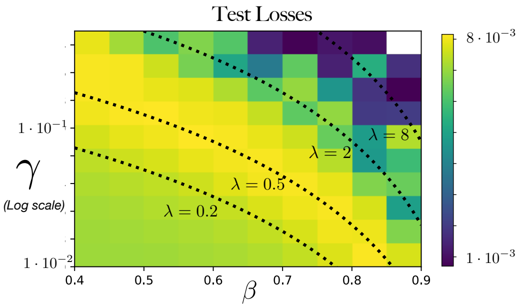

Figure 2: Teacher-student framework with a fully-connected 1-hidden layer ReLU network. The level lines of the test loss after training with MGD(γ, β) correspond to values of γ, β which have a fixed value $\lambda = \gamma/(1-\beta)^2$, as predicted by Proposition 1.

of MGD(γ, β) and MGD$(\hat{\gamma}, \hat{\beta})$, respectively, Corollary 1 implies that we expect $w_{\rho \cdot k}$ and $\hat{w}_k$ to be close. This is in contrast with gradient descent, where scaling the step size by a factor ρ^2 leads to a speedup of ρ^2. This acceleration rule is illustrated in Figure 1 with $\rho = 2$.

Optimisation Regimes. The link between λ, γ, and β highlights several regimes:

- **β large – the iterates converge arbitrarily slow.** Taking β close to 1 while keeping γ constant leads to $\lambda \gg 1$. As explained previously, a chain rule shows that $\tilde{w}_t = w_{\sqrt{\lambda}t}$ follows the ODE $\ddot{\tilde{w}}_t + \lambda^{-1/2} \cdot \dot{\tilde{w}}_t + \nabla F(\tilde{w}_t) = 0$. Consequently, the damping parameter $\lambda^{-1/2}$ goes to 0, and we expect the iterates to heavily oscillate and converge arbitrarily slowly.
- **γ small – the iterates follow GF.** Taking $\gamma \to 0$ while keeping β fixed leads to $\lambda \ll 1$, and MGF(λ) boils down to gradient flow. We expect the MGD(γ, β) iterates to be close to the discretisation of GF with discretisation step $\varepsilon = \gamma/(1-\beta)$. That is, MGD$(\gamma, \beta)$ will approximate GD with step size $\gamma/(1-\beta)$. Hence, MGD gains a speed-up of $1/(1-\beta)$ over GD without a change of trajectory.
- **The "momentum" regime.** In this regime, γ and β are such that λ is non-degenerate, and gradient flow cannot capture the trajectory of MGD(γ, β). Hence, β has an impact on the optimisation path, and the iterates can still converge in reasonable time.

3 Momentum Gradient Flow over Diagonal Linear Networks

Overparametrised Linear Regression. We consider a linear regression over n samples $(x_i, y_i)_{i=1}^n$ with inputs x_i living in $\mathbb{R}^d$ and scalar outputs $y_i \in \mathbb{R}$. We assume the dimension d to be larger than the number of samples n, in which case there exists an infinite number of vectors $\theta^\star$ which perfectly fit the dataset with $y_i = \langle \theta^\star, x_i \rangle$ for all $1 \leq i \leq n$. We call these vectors *interpolators* and we denote by $\mathcal{S}$ the set of such vectors: $\mathcal{S} = \{\theta^\star \in \mathbb{R}^d : y_i = \langle \theta^\star, x_i \rangle, \ \forall i \in [n]\}$. Note that $\mathcal{S}$ is an affine space of dimension at least $(d-n)$ equal to $\theta^\star + \mathrm{span}(x_1, \ldots, x_n)^\perp$ for any interpolator $\theta^\star$. We consider the quadratic loss:

$$\mathcal{L}(\theta) = \frac{1}{2n} \sum_{i=1}^n (y_i - \langle x_i, \theta \rangle)^2. \tag{3}$$

MGF over Least Squares. A classical result found in Lemaire (1996) and Gunasekar et al. (2018) shows that when initialised at θ_0, gradient flow over the quadratic loss (3) converges to the orthogonal projection of the initialisation on $\mathcal{S}$: $\arg\min_{\theta^\star \in \mathcal{S}} \|\theta^\star - \theta_0\|_2$. This next proposition from Alvarez (2000) shows that momentum does not fundamentally change the implicit bias.

Proposition 2 (Alvarez (2000))**.** *Initialised at θ_0 with initial speed $\dot{\theta}_0$, momentum gradient flow* MGF(λ) *over the least squares loss* (3) *converges towards*

$$\operatorname*{argmin}_{\theta^\star \in \mathcal{S}} \|\theta^\star - (\theta_0 + \lambda \dot{\theta}_0)\|_2.$$

MGF(λ) recovers the same solution as gradient flow but with an effective initialisation $\theta_0 + \lambda \dot{\theta}_0$ which takes into account the drift along $\mathrm{span}(x_1, \cdots, x_n)^\perp$ due to the initial speed $\dot{\theta}_0$. Note that in practice, $\dot{\theta}_0$ is chosen equal to 0, in which case the presence of momentum has no effect on the recovered solution.

To better understand momentum's effect on neural networks, we move beyond simple linear parametrization.

2-Layer Diagonal Linear Network. We consider a toy neural network, which corresponds to reparametrising the regression vector as $\theta = u \odot v$ for weights $(u, v) \in \mathbb{R}^{2d}$. This parametrisation can be viewed as a simple neural network $x \mapsto \langle u, \sigma(\mathrm{diag}(v)x) \rangle$, where the output weights are u, the inner weights are the diagonal matrix $\mathrm{diag}(v)$, and where the activation function σ is the identity. The loss function over the trainable weights $w = (u, v) \in \mathbb{R}^{2d}$ now writes

$$F(w) = \mathcal{L}(u \odot v) = \frac{1}{2n} \sum_{i=1}^n (y_i - \langle x_i, u \odot v \rangle)^2,$$

where $\odot$ denotes the Hadamard product. Despite the simplicity of this reparametrisation, the loss function F is non-convex and challenging to analyse.

Momentum Gradient Flow. We consider momentum gradient flow MGF(λ) with parameter $\lambda \geq 0$ over the diagonal-linear-network loss F:

$$\begin{aligned} \lambda \ddot{u}_t + \dot{u}_t + \nabla \mathcal{L}(\theta_t) \odot v_t &= 0 \\ \lambda \ddot{v}_t + \dot{v}_t + \nabla \mathcal{L}(\theta_t) \odot u_t &= 0. \end{aligned} \tag{4}$$

We initialise the flow with zero speed $\dot{u}_0 = \dot{v}_0 = 0$, and apart from requiring the quantity $|u_0^2 - v_0^2|$ to have nonzero coordinates[2], we impose no further constraints on the weight initialisations (u_0, v_0). In what follows, we often rely on the reparametrisation $(w_{+,t}, w_{-,t}) \coloneqq (u_t + v_t, u_t - v_t)$ which makes our formulas more succinct. We will also make use of the *initialisation scale* α, which we define as $\alpha \coloneqq \max(\|u_0\|_\infty, \|v_0\|_\infty)$ and consider as a small quantity.

Balancedness. In our results, the *balancedness* of the weights plays a key role. We recall its definition here.

Definition (Balancedness)**.** The *balancedness*[3] of the weights of the diagonal linear network corresponds to the quantity $\Delta_t \coloneqq |u_t^2 - v_t^2| \in \mathbb{R}^d_{\geq 0}$. We define $\Delta_\infty \coloneqq \lim_{t\to\infty} \Delta_t$ as the *asymptotic balancedness.*

Notice that with the above definition we simply adapted the classical notion of balancedness for general linear neural networks (see Du et al., 2018; Arora et al., 2019) to our toy setting. In the case of gradient flow, a simple derivation shows that balancedness is a conserved quantity: i.e., $\Delta_t = \Delta_0$ for all $t \geq 0$. However, the evolution of Δ_t becomes more complicated as soon as $\lambda > 0$, and our findings emphasise that the *asymptotic balancedness* Δ_∞ plays a crucial role in the generalisation properties of the recovered solution.

Experimental Details. In our numerical experiments, we explore the effects of momentum in the noiseless sparse regression setting with **uncentered data** as in (Nacson et al., 2022). Specifically, we choose $(x_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu\mathbf{1}, \sigma^2 I_d)$ and $y_i = \langle x_i, \theta_s^\star \rangle$ for $i \in [n]$, where $\theta_s^\star$ is s-sparse with nonzero entries equal to $1/\sqrt{s}$. The use of uncentered data is necessary in order to experimentally observe a clear impact of momentum over the training trajectory (see Figure 9 for experiments with centered data). We train a 2-layer diagonal linear network with (M)GD and (M)GF with a uniform initialisation $u_0 = \alpha \cdot \mathbf{1}$, $v_0 = 0$, where $\alpha = 0.01$. For the plots presented in the main part of our paper, we fixed $(n, d, s) = (20, 30, 5)$, $(\mu, \sigma) = (1, 1)$. We show results averaged over 5 replications. We refer the reader to Appendix F for additional experiments where we vary the parameters of the data distribution (e.g., centered data), change the architecture of the trained model, and give further details on the implementation of the (M)GF simulation.

Notations. We let $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n\times d}$ denote the feature matrix and $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ – the output vector. For a vector $z \in \mathbb{R}^d$ and a scalar function $f : \mathbb{R} \to \mathbb{R}$, the action of f on z must be understood element-wise: $f(z) \in \mathbb{R}^d$ represents the vector $(f(z_k))_{k=1}^d$. Inequalities between vectors will also be interpreted as holding coordinate-wise. Additionally, when we write $q_\pm$ for some placeholder quantity q, we mean that we refer to both q_+ and q_-. For example: $w_{\pm,t} = (u_t \pm v_t)$. Finally, for a strictly convex function $\Phi : \mathbb{R}^d \to \mathbb{R}$, which we call a *potential*, the Bregman divergence is defined as the nonnegative quantity $D_\Phi(\theta_1, \theta_2) = \Phi(\theta_1) - \Phi(\theta_2) - \langle \nabla\Phi(\theta_2), \theta_1 - \theta_2 \rangle$, $\forall \theta_1, \theta_2 \in \mathbb{R}^d$.

3.1 Implicit Bias of Gradient Flow

Before analysing the effect of momentum, we start by recalling the known results for gradient flow on diagonal linear networks, which corresponds to taking $\lambda = 0$ in eq. (4). Woodworth et al. (2020) show that the predictors $\theta_t = u_t \odot v_t$ converge towards an interpolator $\theta^{\texttt{GF}}$ uniquely defined by the following constrained minimisation problem:

$$\theta^{\texttt{GF}} = \operatorname*{argmin}_{\theta^\star \in \mathcal{S}} D_{\psi_{\Delta_0}}(\theta^\star, \theta_0), \tag{5}$$

where for $\Delta \in \mathbb{R}^d_{>0}$, $\psi_\Delta : \mathbb{R}^d \to \mathbb{R}$ denotes the hyperbolic entropy function (Ghai et al., 2020) at scale Δ:

$$\psi_\Delta(\theta) = \tfrac{1}{4}\sum_{i=1}^d \Big(2\theta_i \operatorname{arcsinh}\Big(\tfrac{2\theta_i}{\Delta_i}\Big) - \sqrt{4\theta_i^2 + \Delta_i^2} + \Delta_i\Big), \tag{6}$$

and D_{ψ_Δ} is the Bregman divergence. Note that through eq. (5), $\theta^{\texttt{GF}}$ corresponds to the Bregman-projection of the initialisation on the set of interpolators.

Effect of the Initialisation Scale. For a small initialisation scale α, $\theta_0 = O(\alpha^2)$ becomes much smaller than any interpolator $\theta^\star \in \mathcal{S}$. Hence, $D_{\psi_{\Delta_0}}(\theta^\star, \theta_0)$ roughly equals $D_{\psi_{\Delta_0}}(\theta^\star, 0)$, and eq. (5) should be thought of as

$$\theta^{\texttt{GF}} \approx \operatorname*{argmin}_{\theta^\star \in \mathcal{S}} \psi_{\Delta_0}(\theta^\star). \tag{7}$$

This last equation highlights the fact that the recovered solution simply depends on the initial balancedness Δ_0, making it a key quantity. Importantly, the hyperbolic entropy is a convex function which interpolates between the ℓ_1 and ℓ_2 norms as the magnitude of Δ_0 goes from 0 to $+\infty$ (see Woodworth et al. (2020), Theorem 2). So, as $\Delta_0 = O(\alpha^2)$ goes to 0, ψ_{Δ_0} becomes asymptotically identical to the ℓ_1-norm (see Appendix E). Hence, as seen through eq. (7), a small initialisation scale α leads to the recovery of a solution with a small ℓ_1-norm, which facilitates sparse recovery and explains why this setting is referred to as the "rich" or "feature-learning" regime. On the other hand, larger initialisation scales lead to the so-called "kernel" or "lazy" regime, where gradient flow selects small-ℓ_2-norm solutions. **Overall,**

[2] If initially $u_{i,0} = \pm v_{i,0}$ for some coordinate $i \in [d]$, then $u_{i,t} = \pm v_{i,t}$, $\forall t \geq 0$. Hence, imposing $|u_0^2 - v_0^2| \neq 0$ becomes equivalent to working with $2d$ distinct weights. See Appendix C.3.2 for the full argument from uniqueness.

[3] The absolute value in the definition must be understood coordinate-wise.

the smaller the initialisation scale, the closer the retrieved solution will be to the minimum-ℓ_1-norm solution. We refer the reader to the work of Wind et al. (2023) for precise recovery bounds. However, as noted in Even et al. (2023), the picture remains incomplete if we do not take into account the homogeneity of Δ_0. Indeed, initialisations with entries of different magnitudes can hinder the recovery of a sparse vector. However, in our case, our experiments (for uncentered data) verify that the overall magnitudes of Δ_0 and Δ_∞ are sufficient to explain the effects of momentum. We therefore put aside potential homogeneity considerations.

3.2 Implicit Bias of Momentum Gradient Flow

We now move on to describe the impact of momentum on the solution recovered by $\mathrm{MGF}(\lambda)$. Our work proceeds under the following assumption.

Assumption 1 (Boundedness)**.** The optimisation trajectory $(u_t, v_t)_{t\geq 0}$ of MGF (4) is bounded.

Unfortunately, even though Assumption 1 holds true in all our experiments, the boundedness of the trajectory of a second-order gradient flow has only been established under stronger assumption on the loss function (Alvarez, 2000; Goudou and Munier, 2009; Apidopoulos et al., 2022). We defer further details to Appendix C.1. Crucially, the boundedness assumption allows us to prove the convergence of the iterates, and we let $(u_\infty, v_\infty) \coloneqq \lim_{t\to\infty}(u_t, v_t)$. Our goal now becomes to characterise the recovered predictor which we denote with $\theta^{\texttt{MGF}} \coloneqq u_\infty \odot v_\infty$. For our proofs, we make the following additional assumption.

Assumption 2 (Balancedness)**.** The asymptotic balancedness Δ_∞ has non-zero coordinates: $\Delta_{\infty,i} > 0$ for all $i \in [d]$.

Again, Assumption 2 holds true empirically in all our experiments, and in Proposition 3, we prove that small values of λ lead to nonzero asymptotic balancedness. Positing Assumption 2 allows us to prove that the recovered solution $\theta^{\texttt{MGF}}$ interpolates the dataset.

3.2.1 General Characterisation of MGF Bias

In our main result for MGF, we prove that the iterates converge towards an interpolator characterised as the solution of a constrained minimisation problem which involves the hyperbolic entropy (6) scaled at the asymptotic balancedness Δ_∞. Moreover, we derive an insightful description of the asymptotic balancedness in terms of the full optimisation trajectory which allows us to compare the generalisation properties of MGF and GF for small values of λ. Before stating our main continuous-time theorem, we define two integral quantities which appear in our results.

Lemma 1. *The following integral quantities Ω_+ and Ω_- are well-defined and finite:*

$$\Omega_\pm \coloneqq \int_0^\infty \mathrm{m.p.v.} \int_0^t \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_{\pm,t} w_{\pm,s}) \mathrm{d}s \; \mathrm{d}t$$

where $\operatorname{sgn}(\cdot)$ *denotes the sign function,* $w_{\pm,t} = u_t \pm v_t$*, and* m.p.v. *denotes a modified Cauchy principal value defined in Appendix A.*

The fact that the weights $w_{\pm,t}$ can cross zero necessitates the use of the modified Cauchy principal value since otherwise the integrals would diverge. Now, for succinctness, let us introduce the integral quantities

$$I_\pm \coloneqq \Omega_\pm + \Lambda_\pm,$$

where the terms $\Lambda_\pm$ vanish whenever the balancedness Δ_t remains strictly positive for all $t \in [0, \infty]$. The precise form of $\Lambda_\pm$ is uninformative and can be found in Equation (19), Appendix C.3.2. We now proceed to characterise the recovered solution $\theta^{\texttt{MGF}}$.

Theorem 1. *The solution $\theta^{\texttt{MGF}}$ of MGF (4) interpolates the dataset and satisfies the following implicit regularisation:*

$$\theta^{\texttt{MGF}} = \underset{\theta^\star \in \mathcal{S}}{\operatorname{argmin}} \; D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0).$$

In the above expression, $D_{\psi_{\Delta_\infty}}$ denotes the Bregman divergence with potential ψ_{Δ_∞}, where the asymptotic balancedness equals

$$\Delta_\infty = \Delta_0 \odot \exp\big(-(I_+ + I_-)\big)$$

and $\tilde{\theta}_0 = \frac{1}{4}(w_{+,0}^2 \odot \exp(-2I_+) - w_{-,0}^2 \odot \exp(-2I_-))$ *denotes a perturbed initialisation term.*

The proof of Theorem 1 appears in Appendix C.3 as well as explicit formulas for Δ_∞ and $\tilde{\theta}_0$. We explain the significance and shed more light on the different parts of Theorem 1 below.

Perturbed Initialisation $\tilde{\theta}_0$. In all our experiments, we observe that the perturbed initialisation $\tilde{\theta}_0$ remains negligible in the sense that for any interpolator $\theta^\star \in \mathcal{S}$, $\|\tilde{\theta}_0\|_2 \ll \|\theta^\star\|_2$. Moreover, in the next section, we prove that whenever the balancedness remains nonzero during training, $\tilde{\theta}_0$ becomes smaller than α^2, where α stands for the initialisation scale. Hence, exactly for the same reasons as for gradient flow, the implicit regularisation problem from Theorem 1 should be though of as

$$\theta^{\texttt{MGF}} \approx \underset{\theta^\star \in S}{\operatorname{argmin}} \, \psi_{\Delta_\infty}(\theta^\star). \tag{8}$$

Appendix C.3.3 provides more details. Thus, the asymptotic balancedness Δ_∞ becomes the key quantity governing the properties of the recovered solution.

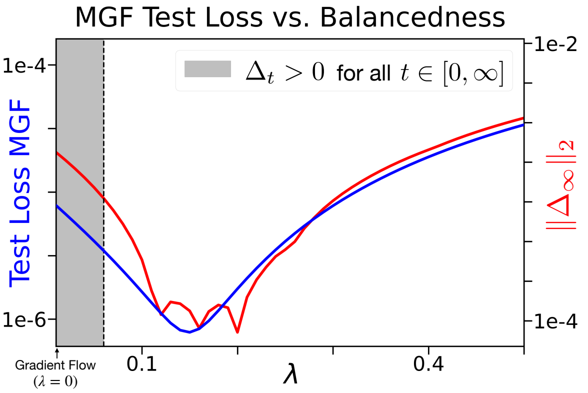

Figure 3: Test loss (in blue) and magnitude of balancedness (in red) at convergence of MGF(λ) over a diagonal linear network in a sparse regression setting with uncentered data. As predicted by Theorem 1, a more balanced solution generalises better. The shaded zone corresponds to values of λ for which the balancedness never hits zero during training and for which Corollary 2 therefore holds.

Key Role of Δ_∞. If during optimisation the weights become more balanced, i.e. $\Delta_\infty < \Delta_0$, then as discussed previously, based on the properties of ψ_{Δ_∞}, the recovered solution will enjoy better sparsity guarantees than the solution of gradient flow. Figure 3 illustrates this point: the smaller the magnitude of Δ_∞, the better the generalisation. Finally note that by eqs. (5) and (8), $\theta^{\texttt{MGF}}$ approximately equals the solution recovered from gradient flow initialised at $u_0 = \sqrt{\Delta_\infty}, v_0 = 0$, which we denote by $\theta^{\texttt{GF}}_{\Delta_\infty}$. We observe $||\theta^{\texttt{MGF}} - \theta^{\texttt{GF}}_{\Delta_\infty}||_2 / ||\theta^{\texttt{GF}}_{\Delta_\infty}||_2 < 0.01$ in all our experiments, which validates the approximation in eq. (8).

Path-Dependent Quantity. Unfortunately, the asymptotic balancedness depends on the whole optimisation trajectory in a very intricate way, and we cannot compare $\|\Delta_\infty\|$ and $\|\Delta_0\|$. Thus, in general, we cannot meaningfully compare the recovered interpolators $\theta^{\texttt{MGF}}$ and $\theta^{\texttt{GF}}$. However, in the following section we prove that with the additional assumption that the balancedness remains nonzero, we have $\Delta_\infty < \Delta_0$.

3.3 Provable Benefits of Momentum for Small Values of λ

In this subsection, we prove that for small values of the momentum flow parameter λ, the recovered solution becomes more balanced (and therefore sparser) than the solution of gradient flow. As a starting point for our argument, notice that if the balancedness $\Delta_t = |u_t^2 - v_t^2| = |w_{+,t} w_{-,t}|$ remains strictly positive throughout training, then the weights $w_{\pm,t}$ never change sign. Hence, the integral quantities $\Lambda_\pm$ become 0, and $\Omega_\pm > 0$. Thus, $I_\pm > 0$, which combined with Theorem 1 implies the following corollary.

Corollary 2. *For $\lambda > 0$, if the balancedness Δ_t remains strictly positive during training (i.e. $\Delta_t \neq 0$ for $t \in [0, +\infty]$), then the perturbed initialisation satisfies $|\tilde{\theta}_0| < \alpha^2$ and*

$$\Delta_\infty = \Delta_0 \odot \exp\Big(- \lambda \int_0^\infty \Big(\frac{\dot{w}_{+,t}}{w_{+,t}}\Big)^2 + \Big(\frac{\dot{w}_{-,t}}{w_{-,t}}\Big)^2 \mathrm{d}t \Big).$$

Importantly, $\Delta_\infty < \Delta_0$.

In words, the above corollary (proved in Appendix C.4) implies that if the balancedness Δ_t does not hit zero during training, then (i) the perturbation term $\tilde{\theta}_0$ is provably negligible, (ii) the asymptotic balancedness is coordinate-wise smaller than initial balancedness Δ_0 which translates into a solution with better sparsity properties than the gradient flow interpolator. This regime corresponds to the gray zone in Figure 3. The following proposition proved in Appendix E demonstrates that for small values of λ, the balancedness remains strictly positive.

Proposition 3. *For $\lambda \leq \frac{n}{\|y\|_2^2} \cdot (\min_{i \leq d} \Delta_{0,i})$, the balancedness Δ_t never vanishes: $\Delta_t \neq 0, \ \forall t \in [0, +\infty]$.*

Hence, through Proposition 3 and Corollary 2, we show that small values of λ lead to solutions with better sparse recovery guarantees.

Limitations of Our Analysis. In Appendix C.3.2, we prove that Δ_t can vanish at most a finite number of times. Experimentally, Δ_t never hits 0 for much larger values of λ than $\frac{n}{\|y\|_2^2} \cdot (\min_{i \leq d} \Delta_{0,i})$, making the bound from Proposition 3 relatively loose. In Figure 3, we empirically observe an interval $(0, \lambda_{max})$ in which MGF(λ) outperforms GF in terms of generalisation. Moreover, there exists an optimal value $\lambda^\star$ (roughly corresponding to the smallest Δ_∞) which brings about the most improvement compared to gradient flow. Unfortunately, as observed Figure 3, the balancedness vanishes for $\lambda = \lambda^\star$, and therefore Corollary 2 does not cover the optimal value. Also note that $(0, \lambda_{max})$ and $\lambda^\star$ depend on the data.

Behaviour of Δ_∞ for Small Values of λ. Unfortunately, determining the precise effect of λ on Δ_∞ is challenging. Nonetheless, for small λ, we informally show in Appendix C.5 that

$$\Delta_\infty^2 \underset{\lambda \to 0}{\approx} \Delta_0^2 \odot \exp\Big(- 2\lambda \int_0^\infty \nabla \mathcal{L}(\theta_s)^2 \mathrm{d}s \Big).$$

This approximate equivalence for small λ echoes the implicit bias of SGD (Even et al., 2023; Pesme et al., 2021), which involves a similar formulation for the effective initialisation where the step size γ appears instead of λ. Note that the above approximation suggests that for small values of λ, Δ_∞ monotonically decreases with λ as experimentally confirmed by Figure 3.

3.4 Sketch of Proof

Implicit Bias through a Second-Order Time-Varying Mirror Flow. A natural way of showing the implicit regularisation (5) of gradient flow on a 2-layer diagonal linear network goes through proving that the predictors $\theta_t^{\mathtt{GF}}$ follow the mirror flow $\mathrm{d}\nabla\psi_{\Delta_0}(\theta_t^{\mathtt{GF}}) = -\nabla\mathcal{L}(\theta_t^{\mathtt{GF}})\mathrm{d}t$. In our setting, we prove that the predictors $\theta_t^{\mathtt{MGF}}$ follow a second-order time-varying mirror flow. Specifically, we define a family of potentials $(\Phi_t)_{t\geq 0}$ with $\Phi_t(\theta) \coloneqq \psi_{\Delta_t}(\theta) - \langle \phi_t, \theta\rangle$ where ψ_{Δ_t} corresponds to the hyperbolic entropy (6) depending on the balancedness Δ_t and a perturbation function ϕ_t. We then prove the following proposition.

Proposition 4. *The predictors $\theta_t^{\mathtt{MGF}}$ follow a momentum mirror flow with time-varying potentials Φ_t:*

$$\lambda\frac{\mathrm{d}^2\nabla\Phi_t(\theta_t^{\mathtt{MGF}})}{\mathrm{d}t^2} + \frac{\mathrm{d}\nabla\Phi_t(\theta_t^{\mathtt{MGF}})}{\mathrm{d}t} + \nabla\mathcal{L}(\theta_t^{\mathtt{MGF}}) = 0.$$

The implicit regularisation follows from integrating the ODE: $\nabla\Phi_\infty(\theta^{\mathtt{MGF}}) = -\int_0^\infty \nabla\mathcal{L}(\theta_t^{\mathtt{MGF}})\mathrm{d}t \in \mathtt{span}(x_1,\ldots,x_n)$ which exactly corresponds to the KKT conditions of the constrained minimisation from Theorem 1. Assuming that $w_{\pm,t}$ do not change sign, the proof of Proposition 4 comes naturally and relies on the writing $w_{\pm,t} = \mathrm{sgn}(w_{\pm,0})\exp(\rho_{\pm,t})$. When the iterates cross 0, this reparametrisation does not hold anymore. The analysis can still be carried out by decomposing $\mathbb{R}_{\geq 0}$ into intervals on which the iterates have constant sign and appropriately sticking the intervals using a modified Cauchy principal value.

4 Momentum SGD over Diagonal Linear Networks

In this section, we move from continuous to discrete time and focus on the original MGD(γ,β) recursion for which we can prove similar but slightly weaker results than the ones for MGF. In fact, our results hold for stochastic momentum gradient descent (SMGD) with any batch size $B\in[n]$. For step size $\gamma>0$ and momentum parameter $\beta\in[0,1)$, the SMGD recursion writes as follows:

$$\begin{aligned} u_{k+1} &= u_k - \gamma\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)\odot v_k + \beta(u_k - u_{k-1}) \\ v_{k+1} &= v_k - \gamma\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)\odot u_k + \beta(v_k - v_{k-1}), \end{aligned} \tag{9}$$

where $L_{\mathcal{B}_k}(\theta) \coloneqq \frac{1}{2B}\sum_{i\in\mathcal{B}_k}(y_i - \langle u\odot v, x_i\rangle)^2$ corresponds to the partial loss over the batch $\mathcal{B}_k\subset[n]$ of size B. The batches could be sampled with or without replacement. As for continuous time, we let $\theta_k = u_k\odot v_k$ correspond to the regression predictor. We initialise at $u_1 = u_0$ and $v_1 = v_0$, and we again consider the balancedness of the weights $\Delta_k \coloneqq |u_k^2 - v_k^2|$ for $k\geq 0$, the reparametrised iterates $w_{\pm,k}\coloneqq u_k\pm v_k$, and the *initialisation scale* $\alpha\coloneqq \max(\|u_0\|_\infty,\|v_0\|_\infty)$. In contrast to our continuous-time prerequisites where we only assumed boundedness of the optimisation trajectory, here we assume that the iterates converge:

Assumption 3 (Convergence)**.** The iterates (u_k,v_k) converge towards the limiting weights (u_∞,v_∞). We denote by $\theta^{\mathtt{SMGD}}\coloneqq u_\infty\odot v_\infty$ the recovered predictor.

As in continuous time, we again assume that the asymptotic balancedness is nonzero.

Assumption 4 (Balancedness)**.** The asymptotic balancedness $\Delta_\infty\coloneqq|u_\infty^2 - v_\infty^2|$ has non-zero coordinates.

Similar to Lemma 1, we define two discrete infinite sums which depend on the entire trajectory and appear in our discrete-time result.

Lemma 2. *The following two sums S_+ and S_- converge to finite vectors:*

$$S_\pm = \frac{1}{1-\beta}\sum_{k=1}^\infty\left[r\Big(\frac{w_{\pm,k+1}}{w_{\pm,k}}\Big) + \beta r\Big(\frac{w_{\pm,k}}{w_{\pm,k+1}}\Big)\right],$$

where $r(z) = (z-1) - \ln(|z|)$ for $z\neq 0$.

Importantly, the function $r(z)$ from Lemma 2 is positive for $z>0$. Contrary to the continuous-time case, in discrete time, the iterates $w_{\pm,k}$ never exactly equal zero. Indeed, since ∇L is linear, we have that for all $k\geq 0$, $w_{\pm,k}(\gamma,\beta)$ is a polynomial in (γ,β). Therefore, the set of pairs (γ,β) for which there exists $k\geq 0$ such that $w_{\pm,k}(\gamma,\beta)=0$ is a negligible set in $\mathbb{R}^2$. The iterates therefore 'jump' over zero, making the sums from Lemma 2 well-defined.

4.1 General Characterisation of SMGD Bias

The following theorem represents the discrete counterpart of Theorem 1 and generalises (Even et al., 2023, Theorem 1) which considers SGD without momentum.

Theorem 2. *The solution $\theta^{\mathtt{SMGD}}$ of SMGD (9) interpolates the dataset and satisfies the following implicit regularisation:*

$$\theta^{\mathtt{SMGD}} = \underset{\theta^\star\in\mathcal{S}}{\operatorname{argmin}}\; D_{\psi_{\Delta_\infty}}(\theta^\star,\tilde\theta_0).$$

In the above expression, $D_{\psi_{\Delta_\infty}}$ denotes the Bregman divergence with potential ψ_{Δ_∞}, where the asymptotic balancedness equals

$$\Delta_\infty = \Delta_0\odot\exp\big(-(S_+ + S_-)\big)$$

and $\tilde\theta_0 = \frac{1}{4}(w_{+,0}^2\odot\exp(-2S_+)) - w_{-,0}^2\odot\exp(-2S_-))$ denotes a perturbed initialisation term.

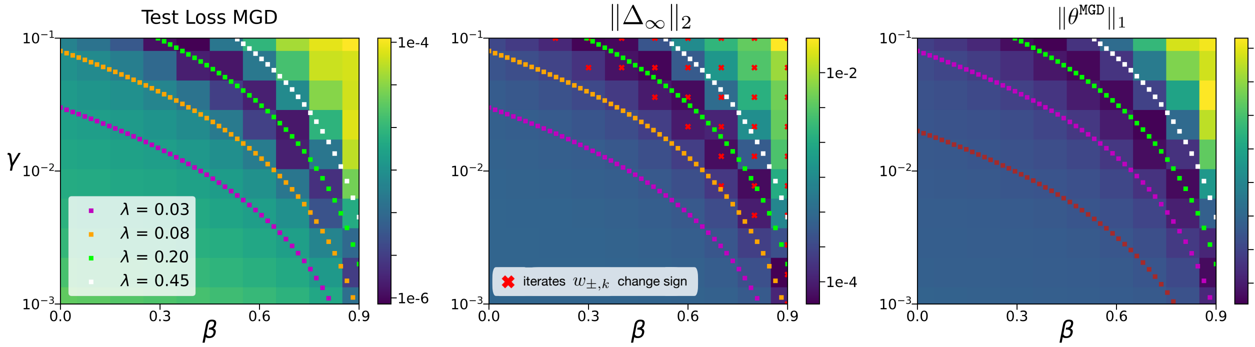

Figure 4: (Non-stochastic) MGD over a diagonal linear network in a sparse regression setting with uncentered data. As predicted by Proposition 1, the three quantities at convergence only depend on the single parameter $\lambda := \gamma/(1-\beta)^2$. As predicted by Theorem 2, a more balanced solution (*center plot*) leads to a solution with a smaller ℓ_1-norm (*right plot*), which in turn translates into better generalisation (*left plot*). Finally, as predicted by Corollary 3, the trajectories for which the iterates do not cross zero satisfy $\Delta_\infty < \Delta_0$, where Δ_0 (approximately) corresponds to the asymptotic balancedness for $\beta = 0$ and $\gamma = 10^{-3}$.

Due to the strong similarities with Theorem 1, we proceed by making similar comments. In our experiments, the norm of the perturbed initialisation $\tilde{\theta}_0$ remains much smaller than that of any interpolator $\theta^\star$. Hence, arguing as before, the implicit regularisation problem from Theorem 2 should be though of as

$$\theta^{\mathtt{SMGD}} \approx \operatorname*{argmin}_{\theta^\star \in S} \psi_{\Delta_\infty}(\theta^\star). \tag{10}$$

Again, the asymptotic balancedness Δ_∞ controls the generalisation properties of the recovered solution. Thus, if $\|\Delta_\infty(\gamma,\beta)\|_2 < \|\Delta_\infty(\gamma',\beta')\|_2$, we expect the interpolator $\theta^{\mathtt{SMGD}}(\gamma,\beta)$ to be sparser than $\theta^{\mathtt{SMGD}}(\gamma',\beta')$. Figure 4 illustrates this point: the smaller the magnitude of Δ_∞ (center plot), the better the sparsity of the interpolator (right plot), which translates into better generalisation (left plot). Unfortunately, as for MGF, the asymptotic balancedness Δ_∞ depends on the whole optimisation trajectory in an intricate way, which prevents us from extracting an insightful formula for Δ_∞ in terms of γ and β. However, Figure 4 indicates that Δ_∞ effectively depends on the single parameter $\lambda = \gamma/(1-\beta)^2$. As in Figure 2, λ again clearly appears to be the relevant quantity which governs the performance of MGD, and not γ and β considered individually. These empirical observations support the idea that even for 'practical' step sizes γ and momentum parameters β, MGD(γ,β) closely follows MGF(λ).

Figure 4 also clearly shows that the asymptotic balancedness decreases as the key quantity λ increases over an interval $[0,\lambda^\star]$ where $\lambda^\star$ denotes the parameter inducing the best generalisation performances. Then, for λ above $\lambda^\star$, the magnitude of Δ_∞ starts to grow and the sparsity of the solutions deteriorates. We expect proving this phenomenon to be very challenging. Such a proof would require a fine-grained analysis of the sums $S_\pm$, which becomes already quite involved when $\beta = 0$ as performed by Even et al. (2023).

Now, similar to the continuous-time result, the following corollary shows that if the iterates do not change sign, then the asymptotic balancedness becomes smaller than the initial balancedness.

Corollary 3. *For $\gamma, \beta > 0$, if the iterates $w_{\pm,k} = (u_k \pm v_k)$ do not change sign during training, then $|\tilde{\theta}_0| < \alpha^2$ and $\Delta_\infty < \Delta_0$.*

The above corollary implies that the recovered solution $\theta^{\mathtt{SMGD}}$ must perform at least as well as the gradient flow interpolator $\theta^{\mathtt{GF}}$. However, in contrast to the continuous case and even though we believe it to be true, we were unable to prove that the SMGD iterates do not change sign for small values of λ.

5 Conclusion

Considering an appropriate second-order differential equation which discretises into MGD, we highlight the existence of a single key quantity $\lambda = \gamma/(1-\beta)^2$ which fully determines the trajectory of MGF. This continuous-time perspective also provides a simple acceleration rule and insight into several relevant optimisation regimes. Then, focusing on 2-layer diagonal linear networks, we prove that the asymptotic balancedness Δ_∞ solely governs the generalisation performances of MGF and SMGD. We additionally prove that small values of λ aid the recovery of sparse MGF solutions. Future work should consider MGF/MGD optimisation on more complex architectures and understand precisely the non-trivial effect of λ on the asymptotic balancedness Δ_∞.

Acknowledgements. The authors would like to thank Jana Vuckovic for investigating various discretisation methods for MGF, Yoana Nakeva and Mathieu Even for providing feedback on the original version of the paper, and Aditya Varre for the helpful discussions which led to the proofs of Theorem 1 and Theorem 2. This project was supported by the Swiss National Science Foundation (grant number 212111).

References

Felipe Alvarez. On the minimizing property of a second order dissipative system in hilbert spaces. *SIAM Journal on Control and Optimization*, 38(4):1102–1119, 2000.

Vassilis Apidopoulos, Nicolò Ginatta, and Silvia Villa. Convergence rates for the heavy-ball continuous dynamics for non-convex optimization, under polyak–lojasiewicz condition. *Journal of Global Optimization*, 84(3):563–589, 2022. ISSN 1573-2916.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. *Implicit Regularization in Deep Matrix Factorization*. Curran Associates Inc., 2019.

H. Attouch, X. Goudou, and P. Redont. The heavy ball with friction method, i. the continuous dynamical system: Global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipitive dynamical system. *Communications in Contemporary Mathematics*, 2(1):1–34, 2000.

Heinz G. Bauschke and Jonathan M. Borwein. Legendre functions and the method of random bregman projections. *Journal of Convex Analysis*, 4:27–67, 1997.

Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Math. Oper. Res.*, 42:330–348, 2017.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Bugra Can, Mert Gurbuzbalaban, and Lingjiong Zhu. Accelerated linear convergence of stochastic momentum methods in wasserstein distances. In *International Conference on Machine Learning*, pages 891–901. PMLR, 2019.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020.

Aaron Defazio. Understanding the role of momentum in non-convex optimization: Practical insights from a lyapunov analysis. *ArXiv*, 2020.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (s)gd over diagonal linear networks: Implicit regularisation, large stepsizes and edge of stability. *NeurIPS 2023*, 2023.

Nicolas Flammarion and Francis Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695. PMLR, 2015.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.

Udaya Ghai, Elad Hazan, and Yoram Singer. Exponentiated gradient meets gradient descent. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 386–407. PMLR, 2020.

Avrajit Ghosh, He Lyu, Xitong Zhang, and Rongrong Wang. Implicit regularization in heavy-ball momentum accelerated stochastic gradient descent. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

X. Goudou and J. Munier. The gradient and heavy ball with friction dynamical systems: the quasiconvex case. *Mathematical Programming*, 116(1):173–191, 2009.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1827–1836. PMLR, 2018.

Jeff Z. HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2315–2357. PMLR, 15–19 Aug 2021.

A. Haraux and M.A. Jendoubi. Convergence of solutions of second-order gradient-like systems with analytic nonlinearities. *Journal of Differential Equations*, 144(2):313–320, 1998.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9965–10040. PMLR, 2022.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pages 1042–1085. PMLR, 2018.

Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

Nikola B. Kovachki and Andrew M. Stuart. Continuous time analysis of momentum methods. *J. Mach. Learn. Res.*, 22(1), 2021.

Guillaume Leclerc and Aleksander Madry. The two regimes of deep network training. *arXiv preprint arXiv:2002.10376*, 2020.

B Lemaire. An asymptotical variational principle associated with the steepest descent method for a convex function. *Journal of Convex Analysis*, 3(1):63–70, 1996.

Zhiyuan Li, Tianhao Wang, Jason D. Lee, and Sanjeev Arora. Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35 - 36th Conference on Neural Information Processing Systems, NeurIPS 2022*, Advances in Neural Information Processing Systems. Neural information processing systems foundation, 2022.

Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271, 2020.

Vien Mai and Mikael Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International conference on machine learning*, pages 6630–6639. PMLR, 2020.

Céline Moucer, Adrien Taylor, and Francis Bach. A systematic approach to lyapunov analyses of continuous-time models in convex optimization. *SIAM Journal on Optimization*, 33(3):1558–1586, 2023.

Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16270–16295. PMLR, 2022.

A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* John Wiley & Sons, New York, 1979.

Antonio Orvieto, Jonas Kohler, and Aurelien Lucchi. The role of memory in stochastic optimization. In *Uncertainty in Artificial Intelligence*, pages 356–366. PMLR, 2020.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

L. Pillaud-Vivien, J. Reygner, and N. Flammarion. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2127–2159. PMLR, 2022.

Boris Polyak and Pavel Shcherbakov. Lyapunov functions: An optimization theory perspective. *IFAC-PapersOnLine*, 50(1):7456–7461, 2017.

B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Jesús María Sanz Serna and Konstantinos C Zygalakis. The connections between lyapunov functions for some optimization algorithms and differential equations. *SIAM Journal on Numerical Analysis*, 59(3):1542–1565, 2021.

Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, pages 3935–3971. PMLR, 2021.

Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pages 1–70, 2021.

Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *Advances in neural information processing systems*, 27, 2014.

Tao Sun, Dongsheng Li, Zhe Quan, Hao Jiang, Shengguo Li, and Yong Dou. Heavy-ball algorithms always escape saddle points. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 3520–3526. AAAI Press, 2019.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 2013. PMLR.

Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Li Wang, Zhiguo Fu, Yingcong Zhou, and Zili Yan. The implicit regularization of momentum gradient descent in overparametrized models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37 (8):10149–10156, 2023.

Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

Ashia C Wilson, Ben Recht, and Michael I Jordan. A lyapunov analysis of accelerated methods in optimization. *The Journal of Machine Learning Research*, 22 (1):5040–5073, 2021.

Johan S Wind, Vegard Antun, and Anders C Hansen. Implicit regularization in ai meets generalized hardness of approximation in optimization–sharp results for diagonal linear networks. *arXiv preprint arXiv:2307.07410*, 2023.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3635–3673. PMLR, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Organisation of the Appendix.

The appendix is organised as follows:

- In Appendix A, we introduce additional notation and provide further comments on the discretisation methods.
- In Appendix B, we present a useful reparametrisation of our problem.
- In Appendix C, we offer proofs for our continuous-time results, specifically Theorem 1, Corollary 2, and Proposition 4.
- In Appendix D, we detail the proofs for our discrete-time results, namely Lemma 2, Theorem 2 and Corollary 3.
- In Appendix E, we introduce technical lemmas necessary for our main results and prove Proposition 3.
- In Appendix F, we provide more details on the main-paper experiments and showcase further experimental results.

A Additional Notations and Comments on Discretisation Methods

Vector Operations. Moving forward, all arithmetic operations and real-valued functions will be considered as being applied coordinate-wise. In other words, if a and b are vectors in $\mathbb{R}^d$ and $p, q \in \mathbb{Q}$, then $a^p b^q \in \mathbb{R}^d$ will be used as a shorthand for the vector with entries $\{a_i^p b_i^q\}_{i=1}^d$. And for any $f : \mathbb{R} \to \mathbb{R}$, $f(a)$ will represent the vector with entries $\{f(a_i)\}_{i=1}^d$. Inequalities between vectors will also be interpreted as holding coordinate-wise.

Mirror Maps. Various definitions of a *mirror map* $\Phi : \mathbb{R}^d \to (-\infty, +\infty]$ exist in the optimization literature (see Nemirovsky and Yudin, 1979; Li et al., 2022), and a common one coincides with the concept of a Legendre function (see Bauschke et al., 2017; Bauschke and Borwein, 1997). In our proofs, we do not deal with extended real-valued functions, and the term mirror map is applied to C^∞-smooth strictly convex functions with coercive gradients. In particular, our mirror maps are of Legendre type.

For such a mirror map $\Phi : \mathbb{R}^d \to \mathbb{R}$, we define the Bregman divergence $D_\Phi(\theta_1, \theta_2)$ for $\theta_1, \theta_2 \in \mathbb{R}^d$ as

$$D_\Phi(\theta_1, \theta_2) = \Phi(\theta_1) - \Phi(\theta_2) - \langle \nabla\Phi(\theta_2), \theta_1 - \theta_2 \rangle.$$

Notice that due to the strict convexity of Φ, $D_\Phi(\theta_1, \theta_2) > 0$ whenever $\theta_1 \neq \theta_2$.

Modified Cauchy Principal Value. Let $f : \mathbb{R}_{\geq 0} \to [-\infty, +\infty]$ be an extended real-valued function with a finite set of poles $\mathcal{T} = \{T_1, T_2, \dots, T_N\}$ (*i.e.* points $t \in \mathbb{R}_{\geq 0}$ at which $f(t) = \pm\infty$) such that f is continuous on $\mathbb{R}_{\geq 0} \setminus \mathcal{T}$. Let $0 < T_1 < \cdots < T_N$. Let $T \in \mathcal{T}$ and let $\varepsilon > 0$ be small enough such that $(T - \varepsilon, T + \varepsilon) \cap \mathcal{T} = \{T\}$. Recall that, provided the limit below exists, the Cauchy principal value p.v. $\int_{T-\varepsilon}^{T+\varepsilon} f(t)\mathrm{d}t$ is defined as

$$\text{p.v.} \int_{T-\varepsilon}^{T+\varepsilon} f(t)\mathrm{d}t \coloneqq \lim_{\delta \to 0} \left[\int_{T-\varepsilon}^{T-\delta} f(t)\mathrm{d}t + \int_{T+\delta}^{T+\varepsilon} f(t)\mathrm{d}t \right].$$

Now, let $\varepsilon_m > 0$ be such that $(T_m - \varepsilon_m, T_m + \varepsilon_m) \cap \mathcal{T} = \{T_m\}$ for $m \in [N]$. Moreover, let $T_0 = \varepsilon_0 = 0$ and $T_{N+1} = +\infty$. Suppose f has finite Cauchy principal values at all poles. Then, for any $\tau \geq 0$ such that $\tau \notin \mathcal{T}$, we could define p.v. $\int_0^\tau f(t)\mathrm{d}t$ as

$$\text{p.v.} \int_0^\tau f(t)\mathrm{d}t \coloneqq \sum_{m : T_{m+1} < \tau} \left[\text{p.v.} \int_{T_m - \varepsilon_m}^{T_m + \varepsilon_m} f(t)\mathrm{d}t + \int_{T_m + \varepsilon_m}^{T_{m+1} - \varepsilon_{m+1}} f(t)\mathrm{d}t \right] + \text{p.v.} \int_{T_k - \varepsilon_k}^{T_k + \varepsilon_k} f(t)\mathrm{d}t + \int_{T_k + \varepsilon_k}^{\tau} f(t)\mathrm{d}t,$$

where $T_k < \tau < T_{k+1}$.

For our proofs of Lemma 1 and Theorem 1, we require a modification to the Cauchy principal value. For the aforementioned function f with the described properties and for $T \in \mathcal{T}$, $\varepsilon > 0$ such that $(T - \varepsilon, T + \varepsilon) \cap \mathcal{T} = \{T\}$, we define the modified principal value m.p.v. $\int_{T-\varepsilon}^{T+\varepsilon} f(t)\mathrm{d}t$ as

$$\text{m.p.v.} \int_{T-\varepsilon}^{T+\varepsilon} f(t)\mathrm{d}t \coloneqq \lim_{\delta \to 0} \left[\int_{T-\varepsilon}^{T-\delta} f(t)\mathrm{d}t \cdot e^{\frac{\delta}{\lambda}} + \int_{T+\delta}^{T+\varepsilon} f(t)\mathrm{d}t \cdot e^{-\frac{\delta}{\lambda}} \right], \tag{11}$$

where λ denotes our familiar MGF parameter. We also extend the m.p.v. definition to integrals $\int_0^\tau f(t)\mathrm{d}t$ for arbitrary $\tau \geq 0$ by mimicking the Cauchy-principal-value construction:

$$\text{m.p.v.}\int_0^\tau f(t)\mathrm{d}t := \sum_{m:T_{m+1}<\tau}\left[\text{m.p.v.}\int_{T_m-\varepsilon_m}^{T_m+\varepsilon_m} f(t)\mathrm{d}t + \int_{T_m+\varepsilon_m}^{T_{m+1}-\varepsilon_{m+1}} f(t)\mathrm{d}t\right] + \text{m.p.v.}\int_{T_k-\varepsilon_k}^{T_k+\varepsilon_k} f(t)\mathrm{d}t + \int_{T_k+\varepsilon_k}^{\tau} f(t)\mathrm{d}t,$$

where $T_k < \tau < T_{k+1}$. Note that the above definition implies that whenever f has no poles on an interval $(a, b) \subset \mathbb{R}_{\geq 0}$, then

$$\text{m.p.v.}\int_a^b f(t)\mathrm{d}t = \int_a^b f(t)\mathrm{d}t.$$

Additional Comments on the Discretisation of MGF(λ)**.** Following our discussion from Section 2, we want to point out that that there are other ways of discretising

$$\lambda\ddot{w}_t + \dot{w}_t + \nabla F(w_t) = 0.$$

Indeed, instead of discretising as (2) in the main paper

$$\lambda\,\frac{w_{k+1} - 2w_k + w_{k-1}}{\varepsilon^2} + \frac{w_k - w_{k-1}}{\varepsilon} + \nabla F(w_k) = 0,$$

one could also consider a central first-order difference:

$$\lambda\,\frac{w_{k+1} - 2w_k + w_{k-1}}{\varepsilon^2} + \frac{w_{k+1} - w_{k-1}}{2\varepsilon} + \nabla F(w_k) = 0.$$

Rearranging, this leads to

$$w_{k+1} = w_k - \frac{\varepsilon^2}{\lambda(1+\frac{\varepsilon}{2\lambda})}\nabla F(w_k) + \frac{1-\frac{\varepsilon}{2\lambda}}{1+\frac{\varepsilon}{2\lambda}}(w_k - w_{k-1}),$$

which corresponds to momentum with $\gamma = \frac{\varepsilon^2}{\lambda(1+\frac{\varepsilon}{2\lambda})}$ and $\beta = \frac{1-\frac{\varepsilon}{2\lambda}}{1+\frac{\varepsilon}{2\lambda}}$. Solving for ε and λ, we get

$$\lambda = \frac{(1+\beta)\gamma}{2(1-\beta)^2} \qquad \text{and} \qquad \varepsilon = \frac{\gamma}{1-\beta}.$$

Hence, we obtain the same discretisation step ε as in Proposition 1 and a slightly different expression for λ. However, note that the two versions of λ become indistinguishable for large values of β since $\frac{1+\beta}{2} \to_{\beta\to 1} 1$. Experimentally, running MGF(λ) with the two different values for λ leads to similar results. Thus, the discretisation scheme from the main paper was chosen due to the more concise definition of λ in this case.

B (w_+, w_-)-Reparametrisation

MGF Reparametrisation. We recall that we consider momentum gradient flow MGF(λ) with parameter $\lambda > 0$ over the diagonal-linear-network loss $F((u, v))) = \mathcal{L}(u \odot v)$:

$$\lambda\ddot{u}_t + \dot{u}_t + \nabla L(\theta_t) \odot v_t = 0;$$
$$\lambda\ddot{v}_t + \dot{v}_t + \nabla L(\theta_t) \odot u_t = 0.$$

For proof-writing convenience, we consider the simple reparametrisation outlined below.

In order to eliminate the cross-dependencies in (u, v) in the above equations, it is natural to consider the quantities $(w_{+,t}, w_{-,t})$ where $w_{\pm,t} = u_t \pm v_t$ for $t \geq 0$. Hence, we get the following reparametrised ODE:

$$\begin{cases}\lambda\ddot{w}_{\pm,t} + \dot{w}_{\pm,t} \pm \nabla\mathcal{L}(\theta_t) \odot w_{\pm,t} = 0;\\ w_{\pm,0} = u_0 \pm v_0, \;\; \dot{w}_{\pm,0} = 0.\end{cases} \tag{12}$$

Notice that with these new quantities, we have

$$\theta_t = \frac{w_{+,t}^2 - w_{-,t}^2}{4} \quad \text{and} \quad \Delta_t = |w_{+,t} w_{-,t}|.$$

MGD Reparametrisation. For the discrete-time setting, we follow the same reparametrisation from the MGD recursion:

$$\begin{aligned} u_{k+1} &= u_k - \gamma \nabla L(\theta_k) \odot v_k + \beta(u_k - u_{k-1}); \\ v_{k+1} &= v_k - \gamma \nabla L(\theta_k) \odot u_k + \beta(v_k - v_{k-1}). \end{aligned}$$

We let $w_{\pm,k} = u_k \pm v_k$ for $k \geq 0$. Then, for $k \geq 1$, the equations above transform into

$$\begin{cases} w_{\pm,k+1} = w_{\pm,k} \mp \gamma \nabla \mathcal{L}(\theta_k) \odot w_{\pm,k} + \beta(w_{\pm,k} - w_{\pm,k-1}); \\ w_{\pm,1} = w_{\pm,0} = u_0 \pm v_0. \end{cases} \tag{13}$$

Again, with the newly defined quantities, we have

$$\theta_k = \frac{w_{+,k}^2 - w_{-,k}^2}{4} \quad \text{and} \quad \Delta_k = |w_{+,k} w_{-,k}|.$$

C Continuous-Time Theorems

C.1 Convergence of Momentum Gradient Flow

Momentum gradient flow (with $\lambda > 0$),

$$\lambda \ddot{w}_t + \dot{w}_t + \nabla F(w_t) = 0,$$

also known in the optimisation literature as the heavy-ball with friction ODE or the heavy-ball dynamical system with constant damping coefficient, has been the object of extensive mathematical study over the years (Haraux and Jendoubi, 1998; Attouch et al., 2000; Alvarez, 2000; Goudou and Munier, 2009; Polyak and Shcherbakov, 2017; Apidopoulos et al., 2022). If we abstract away from the diagonal linear network setting and consider an unspecified loss $F \in C^1(\mathbb{R}^D, \mathbb{R}_{\geq 0})$ with locally Lipschitz gradient, we can still identify a useful Lyapunov function, which perhaps motivated the study of the ODE in the first place. The function in question happens to be the energy of the system

$$E_t = F(w_t) + \frac{\lambda}{2} \|\dot{w}_t\|_2^2, \tag{14}$$

whose nonpositive time-derivative $\dot{E}_t = -\|\dot{w}_t\|_2^2$ allows us to prove the global existence and uniqueness of a solution to MGF [Attouch et al. (2000), Theorem 3.1] in this more general setting. We note that by an easy inductive argument, when the function F is C^k-smooth, the MGF solution w_t is C^{k+1}-smooth. Hence, in our setting where the diagonal-neural-network loss F is C^∞-smooth, the learning trajectory w_t is also C^∞-smooth.

Convergence under Assumption 1.

Under the assumption of a bounded trajectory – $w_t \in L^\infty(0, \infty)$, one can prove the following convergences (Attouch et al., 2000):

$$\lim_{t \to \infty} \dot{w}_t = \lim_{t \to \infty} \nabla F(w_t) = 0.$$

However, even when bounded, the iterates w_t need not converge as demonstrated by the coercive function from Section 4.3 in (Attouch et al., 2000). Nevertheless, when the loss F is also analytic, as in the case of diagonal linear networks, assuming boundedness, one can further prove iterate convergence $\lim_{t \to \infty} w_t = w_\infty$ [Haraux and Jendoubi (1998)].

Unfortunately, without assuming boundedness, iterate convergence has been established only in the cases of convex loss [Alvarez (2000)], quasiconvex loss [Goudou and Munier (2009)], and loss satisfying the Polyak-Lojasiewicz inequality [Apidopoulos et al. (2022)]. Thus, the square loss for a diagonal linear network (and neural networks in general) falls out of the scope of these few favorable cases due to non-convexity and an abundance of local and global minima. For that reason, we posit Assumption 1, which holds true empirically in all our experiments on diagonal linear networks.

Convergence to 0 Loss under Assumption 2.

Let us now go back to the specific case of diagonal linear networks where the loss is given by $F(w) = \mathcal{L}(u \odot v)$ for $w = (u, v)$. Notice that from the discussion above, if we assume boundedness of the trajectory, we have

$$\lim_{t\to\infty} \nabla F(w_t) = (\nabla\mathcal{L}(\theta_\infty) \odot v_\infty, \nabla\mathcal{L}(\theta_\infty) \odot u_\infty) = 0.$$

Therefore, since $\nabla\mathcal{L}(\theta_\infty) \odot \Delta_\infty = 0$, if the balancedness at infinity Δ_∞ has nonzero coordinates, we can conclude that $\nabla\mathcal{L}(\theta_\infty) = 0$. Recalling that $\mathcal{L}$ is convex, we get that $\mathcal{L}(\theta_\infty) = 0$. Hence, θ_∞ interpolates the dataset.

C.2 Proof of Time-Varying Momentum Mirror Flow

In our discussion in Appendix C.1, we saw that assuming

1) iterate boundedness: $u_t, v_t \in L^\infty(0, \infty)$, and
2) nonzero balancedness at infinity: $\Delta_{\infty,i} \neq 0, \ \forall i \in [d]$,

we can prove that MGF over a diagonal linear network (4) converges to an interpolator θ_∞.[4] Before we jump into the proof of Proposition 4, we need to establish the following lemma.

Lemma 3. *Assuming that $u_t, v_t \in L^\infty(0, \infty)$ and $\Delta_{\infty,i} \neq 0, \ \forall i \in [d]$, the following integral limit exists:*

$$\lim_{t\to\infty} \int_0^t \nabla\mathcal{L}(\theta_s) ds = \int_0^\infty \nabla\mathcal{L}(\theta_t) \mathrm{d}t.$$

Consequently,

$$\lim_{t\to\infty} \int_0^t \nabla\mathcal{L}(\theta_s) e^{-\frac{t-s}{\lambda}} ds = 0.$$

Proof. Let us consider the (w_+, w_-)-reparametrisation of MGF (4) given by Equation (12):

$$\lambda \ddot{w}_{\pm,t} + \dot{w}_{\pm,t} \pm \nabla\mathcal{L}(\theta_t) \odot w_{\pm,t} = 0.$$

Since we assumed that Δ_∞ has nonzero coordinates, there exists $T \geq 0$ such that for all $t \geq T$, $w_{\pm,t}$ have nonzero coordinates. Hence, for $t \geq T$, we can safely divide by $w_{\pm,t}$ to obtain

$$\lambda \frac{\mathrm{d}^2 \ln |w_{\pm,t}|}{\mathrm{d}t^2} + \frac{\mathrm{d} \ln |w_{\pm,t}|}{\mathrm{d}t} \pm \nabla\mathcal{L}(\theta_t) + \lambda \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 = 0.$$

Let us notice a couple of things. First, as we discussed in Appendix C.1, the boundedness of the iterates forces $w_{\pm,t}$ to converge to some vectors $w_{\pm,\infty}$ with nonzero coordinates since we assumed the coordinates of $\Delta_\infty = w_{+,\infty} w_{-,\infty}$ are nonzero. Hence,

$$\|\frac{\dot{w}_{\pm,t}^2}{w_{\pm,t}^2}\|_\infty \leq \texttt{const} \cdot (\|\dot{u}_t^2\|_\infty + \|\dot{v}_t^2\|_\infty),$$

where the RHS is integrable as we saw in the proof of Proposition 3. Second, from the discussion in Appendix C.1, we know that $\lim_{t\to\infty} \dot{w}_{\pm,t} = 0$, so $\lim_{t\to\infty} \frac{\mathrm{d} \ln |w_{\pm,t}|}{\mathrm{d}t} = 0$.

Now, for $t \geq T$,

$$\begin{aligned}\int_T^t \nabla\mathcal{L}(\theta_s) ds &= \mp \left(\lambda \int_T^t \frac{\mathrm{d}^2 \ln |w_{\pm,s}|}{\mathrm{d}t^2} ds + \int_T^t \frac{\mathrm{d} \ln |w_{\pm,s}|}{\mathrm{d}t} ds + \int_T^t \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 ds\right) \\ &= \mp \left(\lambda \frac{\mathrm{d} \ln |w_{\pm,s}|}{\mathrm{d}t}\Big|_T^t + \ln |w_{\pm,s}|\Big|_T^t + \int_T^t \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 ds\right).\end{aligned}$$

[4] Note that we also refer to θ_∞ as $\theta^{\mathtt{MGF}}$.

So, using the above observations and letting $t \to \infty$ yields

$$\lim_{t\to\infty} \int_T^t \nabla\mathcal{L}(\theta_s)ds = \mp\left(-\lambda\frac{\mathrm{d}\ln|w_{\pm,T}|}{\mathrm{d}t} - \ln|w_{\pm,T}| + \ln|w_{\pm,\infty}| - \int_T^\infty \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 \mathrm{d}t\right).$$

Thus, we conclude that $\lim_{t\to\infty}\int_0^t \nabla\mathcal{L}(\theta_s)ds$ exists, and therefore, $\lim_{t\to\infty}\int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}ds = 0$. □

We are now well-equipped to prove Proposition 4. We note that we phrased Proposition 4 rather succinctly in the main part of the paper due to space considerations. In what follows, we restate Proposition 4 by precisely specifying the underlying assumptions.

Proposition. *Assume the solution (u_t, v_t) of MGF (4) is bounded. If we also assume that the balancedness at infinity Δ_∞ has nonzero coordinates, then there exists a time $T \geq 0$, after which the predictors $\theta_t = u_t \odot v_t$ follow a momentum mirror flow with time-varying potentials Φ_t:*

$$\lambda\frac{\mathrm{d}^2\nabla\Phi_t(\theta_t)}{\mathrm{d}t^2} + \frac{\mathrm{d}\nabla\Phi_t(\theta_t)}{\mathrm{d}t} + \nabla\mathcal{L}(\theta_t) = 0.$$

Furthermore, if we assume that the balancedness Δ_t remains nonzero for $t \in [0, +\infty]$, then the momentum mirror flow holds for every $t \geq 0$.

Proof. We will consider the (w_+, w_-)-reparametrisation of momentum gradient flow (4) introduced in Appendix B. For convenience of the reader, we recall this reparametrisation here:

$$\lambda\ddot{w}_{\pm,t} + \dot{w}_{\pm,t} \pm \nabla\mathcal{L}(\theta_t)\odot w_{\pm,t} = 0.$$

Now, let $\xi : \mathbb{R}_{\geq 0} \to \mathbb{R}^d$ be the $C^\infty(0,\infty)$ solution of the following ODE:

$$\lambda\ddot{\xi}_t + \dot{\xi}_t + \nabla\mathcal{L}(\theta_t) = 0,$$

with the constraint $\xi_0 = \dot{\xi}_0 = 0$. Hence, by Lemma 5,

$$\xi_t = -\int_0^t \nabla\mathcal{L}(\theta_s)(1 - e^{-\frac{t-s}{\lambda}})\mathrm{d}s,$$

and by Lemma 3,

$$\xi_\infty = -\int_0^\infty \nabla\mathcal{L}(\theta_t)\mathrm{d}t.$$

Thus, $\xi_t \in \mathtt{span}(x_1, \ldots, x_n)$, $\forall t \in [0, +\infty]$.

Having fixed ξ_t, we define the quantities $\alpha_{\pm,t}$ for every $t \in [0, +\infty]$ through the following relation:

$$\alpha_{\pm,t} = w_{\pm,t}\exp(\mp\xi_t).$$

So, $\Delta_t = |w_{+,t}w_{-,t}| = |\alpha_{+,t}\alpha_{-,t}|$. Furthermore,

$$\begin{aligned}\theta_t &= \frac{1}{4}(w_{+,t}^2 - w_{-,t}^2)\\ &= \frac{1}{4}(\alpha_{+,t}^2\exp(2\xi_t) - \alpha_{-,t}^2\exp(-2\xi_t))\\ &= \frac{1}{2}\Delta_t\sinh\left(2\xi_t + \ln\frac{|\alpha_{+,t}|}{|\alpha_{-,t}|}\right)\end{aligned}$$

Since we assumed that Δ_∞ has nonzero coordinates, there exists $T \geq 0$ such that for all $t \geq T$, $w_{\pm,t}$ have nonzero coordinates. Hence, for $t \geq T$, the logarithm $\ln\frac{|\alpha_{+,t}|}{|\alpha_{-,t}|}$ is well-defined. If we assume positive balancedness for $t \in [0, +\infty]$, then we can choose $T = 0$. From now until the end of the proof, whenever a time-dependent quantity features division by Δ_t, we will tacitly assume that $t \geq T$.

Let us now introduce the helper quantity ϕ_t through the following identity:

$$\phi_t = \frac{1}{2}\ln\frac{|\alpha_{+,t}|}{|\alpha_{-,t}|} = \frac{1}{2}\operatorname{arcsinh}\left(\frac{\alpha_{+,t}^2 - \alpha_{-,t}^2}{2\Delta_t}\right).$$

Then,

$$\frac{1}{2}\operatorname{arcsinh}\left(\frac{2\theta_t}{\Delta_t}\right) - \phi_t = \xi_t \in \mathtt{span}(x_1, \ldots, x_n).$$

So, if we consider the time-varying potential

$$\begin{aligned}\Phi_t(\theta) &= \frac{1}{4}\sum_{i=1}^{d}\left(2\theta_i \operatorname{arcsinh}\left(\frac{2\theta_i}{\Delta_{t,i}}\right) - \sqrt{4\theta_i^2 + \Delta_{t,i}^2} + \Delta_{t,i}\right) - \langle \phi_t, \theta\rangle \\ &= \psi_{\Delta_t}(\theta) - \langle \phi_t, \theta\rangle,\end{aligned} \tag{15}$$

where ψ_{Δ_t} is the hyperbolic entropy defined in Equation (6), then

$$\nabla\Phi_t(\theta) = \frac{1}{2}\operatorname{arcsinh}\left(\frac{2\theta}{\Delta_t}\right) - \phi_t.$$

Notice that $\nabla^2\Phi_t = \mathtt{diag}\left(1/\sqrt{4\theta^2 + \Delta_t^2}\right) \succ 0$. Hence, Φ_t is a mirror map. Furthermore, $\nabla\Phi_t(\theta_t) = \xi_t$ for $t \geq T$, so

$$\lambda\frac{\mathrm{d}^2\nabla\Phi_t(\theta_t)}{\mathrm{d}t^2} + \frac{\mathrm{d}\nabla\Phi_t(\theta_t)}{\mathrm{d}t} + \nabla\mathcal{L}(\theta_t) = 0.$$

□

C.3 Proof of Theorem 1

We are now ready to prove our main result for the implicit bias of momentum gradient flow on diagonal linear networks.

Theorem 1. *The solution θ^{MGF} of MGF* (4) *interpolates the dataset and satisfies the following implicit regularisation:*

$$\theta^{\mathit{MGF}} = \operatorname*{argmin}_{\theta^\star \in \mathcal{S}}\; D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0).$$

In the above expression, $D_{\psi_{\Delta_\infty}}$ denotes the Bregman divergence with potential ψ_{Δ_∞}, where the asymptotic balancedness equals

$$\Delta_\infty = \Delta_0 \odot \exp\big(-(I_+ + I_-)\big)$$

and $\tilde{\theta}_0 = \frac{1}{4}(w_{+,0}^2 \odot \exp(-2I_+) - w_{-,0}^2 \odot \exp(-2I_-))$ denotes a perturbed initialisation term.

We split the proof into two parts for conceptual clarity. In the first part, we utilise the time-varying mirror flow from Proposition 4 to derive the implicit regularisation $\theta^{\mathsf{MGF}} = \operatorname{argmin}_{\theta^\star\in\mathcal{S}}\; D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0)$. Then, in the second part, we prove that the integral quantities $I_\pm$ from Lemma 1 are well-defined, and we give the trajectory-dependent characterisations of the asymptotic balancedness Δ_∞ and the perturbed initialisation $\tilde{\theta}_0$.

C.3.1 Proof of Implicit Regularisation.

In Proposition 4, we proved that whenever the MGF trajectory is bounded and the coordinates of Δ_∞ are nonzero, there exists a time $T \geq 0$, after which the predictors θ_t follow a momentum mirror flow with potentials given by Equation (15). Recall that for $t \geq T$,

$$\nabla\Phi_t(\theta_t) = \frac{1}{2}\operatorname{arcsinh}\left(\frac{2\theta}{\Delta_t}\right) - \phi_t = -\xi_t \in \mathtt{span}(x_1, \ldots, x_n).$$

where $\xi_t = -\int_0^t \nabla\mathcal{L}(\theta_s)(1 - e^{-\frac{t-s}{\lambda}})ds$, $\alpha_{\pm,t} = w_{\pm,t}\exp(\mp\xi_t)$, and $\phi_t = \frac{1}{2}\operatorname{arcsinh}\left(\frac{\alpha_{+,t}^2 - \alpha_{-,t}^2}{2\Delta_t}\right)$.

Now, as $t \to \infty$, ξ_t and the MGF iterates converge, so we know that $\nabla \Phi_\infty(\theta_\infty) \in \mathtt{span}(x_1, \dots, x_n)$, where $\Phi_\infty(\theta) = \psi_{\Delta_\infty}(\theta) - \langle \phi_\infty, \theta \rangle$. Thus, we can use the familiar Bregman-Cosine-Theorem trick to characterise the interpolator θ_∞. We proceed with this characterisation.

Let $\tilde{\theta}_0$ be a perturbation term such that $\nabla \Phi_\infty(\tilde{\theta}_0) = 0$. Equivalently,

$$\frac{1}{2}\operatorname{arcsinh}\left(\frac{2\tilde{\theta}_0}{\Delta_\infty^2}\right) - \phi_\infty = 0 \iff$$

$$\operatorname{arcsinh}\left(\frac{2\tilde{\theta}_0}{\Delta_\infty^2}\right) - \operatorname{arcsinh}\left(\frac{\alpha_{+,\infty}^2 - \alpha_{-,\infty}^2}{2\Delta_\infty^2}\right) = 0 \iff$$

$$\tilde{\theta}_0 = \frac{\alpha_{+,\infty}^2 - \alpha_{-,\infty}^2}{4}.$$

Note that $\alpha_{\pm,\infty} = w_{\pm,\infty} \exp(\pm \int_0^\infty \nabla \mathcal{L}(\theta_t) \mathrm{d}t)$ by Lemma 3 and $\Delta_\infty = |\alpha_{+,\infty}\alpha_{-,\infty}|$.

Now, let $\theta^\star \in \mathcal{S}$ be an arbitrary interpolator of the dataset. Then, $\theta^\star - \theta_\infty \in \ker(X) = \mathtt{span}(x_1, \dots, x_n)^\perp$. Hence, the Bregman Cosine Theorem yields

$$\begin{aligned} D_{\Phi_\infty}(\theta^\star, \tilde{\theta}_0) &= D_{\Phi_\infty}(\theta^\star, \theta_\infty) + D_{\Phi_\infty}(\theta_\infty, \tilde{\theta}_0) + \langle \theta^\star - \theta_\infty, \nabla\Phi(\theta_\infty) - \nabla\Phi(\tilde{\theta}_0) \rangle \\ &= D_{\Phi_\infty}(\theta^\star, \theta_\infty) + D_{\Phi_\infty}(\theta_\infty, \tilde{\theta}_0), \end{aligned}$$

where we used that $\nabla\Phi(\theta_\infty) - \nabla\Phi(\tilde{\theta}_0) \in \mathtt{span}(x_1, \dots, x_n)$. Thus,

$$\begin{aligned} \theta_\infty &= \operatorname*{argmin}_{\theta^\star \in \mathcal{S}} D_{\Phi_\infty}(\theta^\star, \tilde{\theta}_0) \\ &= \operatorname*{argmin}_{\theta^\star \in \mathcal{S}} \Phi_\infty(\theta^\star). \end{aligned}$$

Finally, notice that $\nabla \psi_{\Delta_\infty}(\tilde{\theta}_0) = \frac{1}{2}\operatorname{arcsinh}\left(\frac{2\tilde{\theta}_0}{\Delta_\infty}\right) = \phi_\infty$ as we showed above. Hence,

$$D_{\psi_{\Delta_\infty}}(\theta, \tilde{\theta}_0) = \Phi_\infty(\theta) - \psi_{\Delta_\infty}(\tilde{\theta}_0) + \langle \nabla\psi_{\Delta_\infty}(\tilde{\theta}_0), \tilde{\theta}_0 \rangle.$$

Thus, we conclude that

$$\theta_\infty = \operatorname*{argmin}_{\theta^\star \in \mathcal{S}} \Phi_\infty(\theta^\star) = \operatorname*{argmin}_{\theta^\star \in \mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0).$$

□

C.3.2 Proof of Trajectory-Dependent Characterisation.

We just showed that the recovered interpolator by MGF solves the constrained minimisation problem $\theta_\infty = \operatorname{argmin}_{\theta^\star \in \mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0)$, where $\Delta_\infty = |\alpha_{+,\infty}\alpha_{-,\infty}|$, $\tilde{\theta}_0 = (\alpha_{+,\infty}^2 - \alpha_{-,\infty}^2)/4$, and $\alpha_{\pm,\infty} = w_{\pm,\infty} \exp(\pm \int_0^\infty \nabla \mathcal{L}(\theta_t)\mathrm{d}t)$. Clearly, these opaque characterisations of Δ_∞ and $\tilde{\theta}_0$ prevent us from describing how the magnitude of these quantities compares to the magnitude of the initial balancedness Δ_0 and the initialisation scale $\alpha = \max(|u_0|, |v_0|)$. Ideally, we would like to find formulas for Δ_∞ and $\tilde{\theta}_0$ which show that $\tilde{\theta}_0 \ll \theta^\star$, $\forall \theta^\star \in \mathcal{S}$ and $\Delta_\infty < \Delta_0$ so that we can conclude that $\theta^{\mathtt{MGF}} \approx \operatorname{argmin}_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$ enjoys better sparsity guarantees than $\theta^{\mathtt{GF}} \approx \operatorname{argmin}_{\theta^\star \in \mathcal{S}} \psi_{\Delta_0}(\theta^\star)$. In what follows, we derive such formulas.

In our subsequent arguments, for a vector $z \in \mathbb{R}^d$ and a coordinate $i \in [d]$, we will denote with $z(i)$ the i^{th} coordinate of z in order to reduce the index bloat. Let us consider again the (w_+, w_-)-reparametrisation of MGF discussed in Appendix B:

$$\lambda \ddot{w}_{\pm,t} + \dot{w}_{\pm,t} \pm \nabla \mathcal{L}(\theta_t) \odot w_{\pm,t} = 0. \tag{16}$$

Notice that if for some $T > 0$ and $i \in [d]$, $w_{+,T}(i) = 0$, then $\dot{w}_{+,T}(i)$ must be nonzero. Indeed, as we argued in Appendix C.1, MGF (4) admits a unique global solution. And if $w_{+,T}(i) = \dot{w}_{+,T}(i) = 0$, then we could construct another solution $(w'_{+,t}, w'_{-,t})$ of MGF such that $w'_{+,t}(i) = \dot{w}'_{+,t}(i) = 0,\ \forall t \geq 0$, and $w_{\pm,T} = w'_{\pm,T}$. By uniqueness, we get that $w_{\pm,t} = w'_{\pm,t},\ \forall t \geq 0$ However, the newly constructed solution will not be consistent with the imposed initialisation $\Delta_0 \neq 0$,. Hence, $\Delta_0 \neq 0$ prevents $w_{+,t}(i)$ and $\dot{w}_+(i)$ from hitting 0 simultaneously. Similarly, this situation cannot occur for $w_{-,t}$.

Until further notice, we fix a coordinate $i \in [d]$ and consider eq. (16) only in the i^{th} coordinate without explicit mention. If $w_{\pm,T} = 0$ for some $T > 0$, then $\dot{w}_{\pm,T} \neq 0$. Hence, for some small $\delta > 0$, $\dot{w}_{\pm,t}$ does not change sign on $[T-\delta, T+\delta]$, so $w_{\pm,t}$ either strictly increases or decreases on $[T-\delta, T+\delta]$. Therefore, $w_{\pm,t} \neq 0$ on $[T-\delta, T+\delta] \setminus \{T\}$ implying that $w_{\pm,t}$ equals 0 at most a countable number of times. Recall that by Assumption 2, there exists a time T_∞ after which $w_\pm$ does not change sign. Therefore, if we assume that $w_\pm$ vanishes on infinitely many points $T_1 < T_2 < \cdots < T_\infty$, then by compactness, the limit $\tau = \lim_{m\to\infty} T_m$ exists. Since $w_\pm$ is continuous, we infer that $w_{\pm,\tau} = 0$. Moreover, by the Mean Value Theorem, for every $m \geq 1$, there exists $T'_m \in (T_m, T_{m+1})$ such that $\dot{w}_{T'_m} = 0$. Notice that $\lim_{m\to\infty} T'_m = \tau$ as well. Hence, by continuity, $w_{\pm,\tau} = \dot{w}_{\pm,\tau} = 0$ – a contradiction.

Hence, $w_\pm$ vanishes on a finite set of points. Let us order these vanishing times as $0 < T_1 < \cdots < T_N$ and let $T_0 = 0$ and $T_{N+1} = +\infty$. Observe that for $t \notin \mathcal{T} = \{T_i : i \in [N]\}$, we can safely divide both sides of eq. (16) by $w_{\pm,t}$ to obtain

$$\lambda \frac{\ddot{w}_{\pm,t}}{w_{\pm,t}} + \frac{\dot{w}_{\pm,t}}{w_{\pm,t}} \pm \nabla\mathcal{L}(\theta_t) = 0.$$

The last expression is equivalent to

$$\lambda \left(\frac{\ddot{w}_{\pm,t}}{w_{\pm,t}} - \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 \right) + \frac{\dot{w}_{\pm,t}}{w_{\pm,t}} \pm \nabla\mathcal{L}(\theta_t) + \lambda \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 = 0,$$

which can be rewritten as

$$\lambda \frac{\mathrm{d}^2 \ln(\operatorname{sgn}(w_{\pm,t}) w_{\pm,t})}{\mathrm{d}t^2} + \frac{\mathrm{d} \ln(\operatorname{sgn}(w_{\pm,t}) w_{\pm,t})}{\mathrm{d}t} \pm \nabla\mathcal{L}(\theta_t) + \lambda \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 = 0.$$

Let us define a new function $g_\pm : \mathbb{R}_{\geq 0} \setminus \mathcal{T} \to \mathbb{R}^d$ through the relation $g_{\pm,t} = \ln(\operatorname{sgn}(w_{\pm,t}) w_{\pm,t})$. Then, $g_\pm$ is C^∞-smooth on $\mathbb{R}_{\geq 0} \setminus \mathcal{T}$ and satisfies the following ODE:

$$\lambda \ddot{g}_{\pm,t} + \dot{g}_{\pm,t} \pm \nabla\mathcal{L}(\theta_t) + \lambda \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 = 0. \tag{17}$$

Induction on Vanishing Times. Now, we proceed to prove by induction on $N - 1 \geq m \geq 0$ that for $\tau \in (T_m, T_{m+1})$ the following 3 things hold:

- The following integral quantities[5] exist and are finite:

$$\text{m.p.v.} \int_0^\tau \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 e^{-\frac{\tau-s}{\lambda}} \operatorname{sgn}(w_{\pm,\tau} w_{\pm,t}) \mathrm{d}t \quad \text{and} \quad \int_0^\tau \text{m.p.v.} \int_0^t \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_{\pm,t} w_{\pm,s}) \mathrm{d}s \; \mathrm{d}t.$$

- The following identity holds:

$$\dot{g}_{\pm,\tau} = -\text{m.p.v.} \int_0^\tau \left[\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 \pm \frac{1}{\lambda} \nabla\mathcal{L}(\theta_t) \right] e^{-\frac{\tau-t}{\lambda}} \operatorname{sgn}(w_{\pm,\tau} w_{\pm,t}) \mathrm{d}t - \frac{1}{\lambda} \sum_{k=1}^m (-1)^{m-k} e^{-\frac{\tau-T_k}{\lambda}}.$$

- The following identity holds:

$$g_{\pm,\tau} = g_{\pm,0} - \int_0^\tau \text{m.p.v.} \int_0^t \left[\left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 \pm \frac{1}{\lambda} \nabla\mathcal{L}(\theta_s) \right] e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_{\pm,t} w_{\pm,s}) \mathrm{d}s \; \mathrm{d}t$$
$$- \sum_{k=1}^m (-1)^{m-k} \left(1 - e^{-\frac{\tau-T_k}{\lambda}}\right) + 2 \sum_{1 \leq i < j \leq m} (-1)^{j-i} \left(1 - e^{-\frac{T_j - T_i}{\lambda}}\right).$$

[5] See Equation (11) for the definition of m.p.v.

Recall that $\nabla\mathcal{L}(\theta_t)$ is a bounded function, so if the modified principal value from the first bullet point exists, then the modified principal values in the above identities are also well-defined.

Base case: $m = 0$. Recall from the proof of Proposition 4 in Appendix C.2 that $\dot{w}_\pm \in L^2(0,\infty)$. Now, since $w_{\pm,t}$ does not change signs on the interval (T_0, τ), we know that $1/w_{\pm,t} = \Omega(1)$. Hence,

$$\left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}} \in L^1(0,\infty).$$

Similarly, $\left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2$ is integrable on all intervals $[T_i+\varepsilon, T_{i+1}-\varepsilon]$ for any small $\varepsilon > 0$. Consequently, the integral quantities

$$\text{m.p.v.} \int_0^\tau \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_{\pm,\tau} w_{\pm,t}) \mathrm{d}t = \int_0^\tau \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 e^{-\frac{t-s}{\lambda}} \mathrm{d}t$$

and

$$\begin{aligned}\int_0^\tau \text{m.p.v.} \int_0^t \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_{\pm,t} w_{\pm,s}) \mathrm{d}s \ \mathrm{d}t &= \int_0^\tau \int_0^t \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}} \mathrm{d}s \ \mathrm{d}t \\ &= \int_0^\tau \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 (1 - e^{-\frac{\tau-t}{\lambda}}) \mathrm{d}t\end{aligned}$$

are well-defined. Moreover, after applying Lemma 5 to eq. (17), we get

$$\begin{aligned}\dot{g}_{\pm,\tau} &= -\int_0^\tau \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 e^{-\frac{\tau-t}{\lambda}} \mathrm{d}t \ \mp \frac{1}{\lambda} \int_0^\tau \nabla\mathcal{L}(\theta_t) e^{-\frac{\tau-t}{\lambda}} \mathrm{d}t \\ g_{\pm,\tau} &= g_{\pm,0} - \lambda \int_0^\tau \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 (1 - e^{-\frac{\tau-t}{\lambda}}) \mathrm{d}t \ \mp \int_0^\tau \nabla\mathcal{L}(\theta_t)(1 - e^{-\frac{\tau-t}{\lambda}}) \mathrm{d}t,\end{aligned}$$

which concludes the proof of the base case.

Induction step: $m \to m+1$. For $m \geq 0$, assume that for every $\tau \in [0, T_{m+1}) \setminus \mathcal{T}$ the expressions

$$\dot{g}_{\pm,\tau} = -\text{m.p.v.} \int_0^\tau \left[\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_t)\right] e^{-\frac{\tau-t}{\lambda}} \operatorname{sgn}(w_{\pm,\tau} w_{\pm,t}) \mathrm{d}t - \frac{1}{\lambda}\sum_{k=1}^m (-1)^{m-k} e^{-\frac{\tau - T_k}{\lambda}}$$

and

$$\begin{aligned}g_{\pm,\tau} = g_{\pm,0} &- \int_0^\tau \text{m.p.v.} \int_0^t \left[\left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right] e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_{\pm,t} w_{\pm,s}) \mathrm{d}s \ \mathrm{d}t \\ &- \sum_{k=1}^m (-1)^{m-k} \left(1 - e^{-\frac{\tau-T_k}{\lambda}}\right) + 2 \sum_{1 \leq i < j \leq m} (-1)^{j-i} \left(1 - e^{-\frac{T_j - T_i}{\lambda}}\right).\end{aligned}$$

are true and well-defined. We now want to extend the validity of these identities to $\tau \in (T_{m+1}, T_{m+2})$. For ease of notation during the induction step, let $T_{m+1} = T$, $w_\pm = w$, and $g_\pm = g$. Let $\varepsilon > 0$ and let $T_\pm = T \pm \varepsilon$.

Now, applying Lemma 5 to eq. (17) yields

$$\begin{aligned}\dot{g}_\tau &= \dot{g}_{T_+} e^{-\frac{\tau - T_+}{\lambda}} - \int_{T_+}^\tau \left[\left(\frac{\dot{w}_t}{w_t}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_t)\right] e^{-\frac{\tau-t}{\lambda}} \mathrm{d}t \\ g_\tau &= g_{T_+} + \dot{g}_{T_+} \int_{T_+}^\tau e^{-\frac{t-T_+}{\lambda}} \mathrm{d}t - \int_{T_+}^\tau \int_{T_+}^t \left[\left(\frac{\dot{w}_s}{w_s}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right] e^{-\frac{t-s}{\lambda}} \mathrm{d}s \ \mathrm{d}t.\end{aligned}$$

For further ease of notation and with some abuse of notation, let $f_t = \left(\frac{\dot{w}_t}{w_t}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_t)$ on $\mathbb{R}_{\geq 0} \setminus \mathcal{T}$. We will shortly prove that $g_{T_+} - g_{T_-} = O(\varepsilon)$ and $\dot{g}_{T_+} + \dot{g}_{T_-} + \frac{1}{\lambda} = O(\varepsilon)$.[6] Hence, the following limits will hold:

$$\dot{g}_\tau = \lim_{\varepsilon \to 0} \left[-\frac{1}{\lambda} e^{-\frac{\tau - T_+}{\lambda}} - \dot{g}_{T_-} e^{-\frac{\tau - T_+}{\lambda}} - \int_{T_+}^{\tau} f_t e^{-\frac{\tau - t}{\lambda}} \mathrm{d}t \right]$$

$$g_\tau = \lim_{\varepsilon \to 0} \left[g_{T_-} - \frac{1}{\lambda} \int_{T_+}^{\tau} e^{-\frac{t - T_+}{\lambda}} \mathrm{d}t - \dot{g}_{T_-} \int_{T_+}^{\tau} e^{-\frac{t - T_+}{\lambda}} \mathrm{d}t - \int_{T_+}^{\tau} \int_{T_+}^{t} f_s e^{-\frac{t-s}{\lambda}} \mathrm{d}s \ \mathrm{d}t \right].$$

Induction step for $\dot{g}_\tau$. Let us begin to untangle the first limit by substituting $\dot{g}_{T_-}$ with its integral formula given by the induction hypothesis. Notice that

$$\begin{aligned}
\dot{g}_{T_-} e^{-\frac{\tau - T_+}{\lambda}} &= -\text{m.p.v.} \int_0^{T_-} f_t e^{-\frac{T_- - t}{\lambda}} \operatorname{sgn}(w_{T_-} w_t) \mathrm{d}t \cdot e^{-\frac{\tau - T_+}{\lambda}} - \frac{e^{-\frac{\tau - T_+}{\lambda}}}{\lambda} \sum_{k=1}^{m} (-1)^{m-k} e^{-\frac{T_- - T_k}{\lambda}} \\
&= \text{m.p.v.} \int_0^{T_-} f_t e^{-\frac{\tau - t}{\lambda}} \operatorname{sgn}(w_\tau w_t) \mathrm{d}t \cdot e^{\frac{2\varepsilon}{\lambda}} + \frac{e^{\frac{2\varepsilon}{\lambda}}}{\lambda} \sum_{k=1}^{m} (-1)^{(m+1)-k} e^{-\frac{\tau - T_k}{\lambda}},
\end{aligned}$$

where we used that $\operatorname{sgn}(\tau) = -\operatorname{sgn}(T_-)$ since w changes signs at T. Hence, we have that

$$\begin{aligned}
\dot{g}_\tau = &- \lim_{\varepsilon \to 0} \Big[\frac{1}{\lambda} e^{-\frac{\tau - T_+}{\lambda}} + \frac{e^{\frac{2\varepsilon}{\lambda}}}{\lambda} \sum_{k=1}^{m} (-1)^{(m+1)-k} e^{-\frac{\tau - T_k}{\lambda}} \\
&+ \text{m.p.v.} \int_0^{T_-} f_t e^{-\frac{\tau - t}{\lambda}} \operatorname{sgn}(w_\tau w_t) \mathrm{d}t \cdot e^{\frac{2\varepsilon}{\lambda}} + \int_{T_+}^{\tau} f_t e^{-\frac{\tau - t}{\lambda}} \operatorname{sgn}(w_\tau w_t) \mathrm{d}t \Big] \\
= &\mp \int_0^{\tau} \frac{1}{\lambda} \nabla \mathcal{L}(\theta_t) \operatorname{sgn}(w_\tau w_t) - \frac{1}{\lambda} \sum_{k=1}^{m+1} (-1)^{(m+1)-k} e^{-\frac{\tau - T_k}{\lambda}} \\
&- \lim_{\varepsilon \to 0} \left[\text{m.p.v.} \int_0^{T_-} \left(\frac{\dot{w}_t}{w_t} \right)^2 e^{-\frac{\tau - t}{\lambda}} \operatorname{sgn}(w_\tau w_t) \mathrm{d}t \cdot e^{\frac{2\varepsilon}{\lambda}} + \int_{T_+}^{\tau} \left(\frac{\dot{w}_t}{w_t} \right)^2 e^{-\frac{\tau - t}{\lambda}} \operatorname{sgn}(w_\tau w_t) \mathrm{d}t \right],
\end{aligned}$$

where the limit on the last line formally equals the modified principal value m.p.v. $\int_0^\tau \left(\frac{\dot{w}_t}{w_t}\right)^2 e^{-\frac{\tau - s}{\lambda}} \operatorname{sgn}(w_\tau w_t) \mathrm{d}t$ whose existence we want to prove as part of the induction step. In fact, notice that we just proved the existence of m.p.v. $\int_0^\tau \left(\frac{\dot{w}_t}{w_t}\right)^2 e^{-\frac{\tau - s}{\lambda}} \operatorname{sgn}(w_\tau w_t) \mathrm{d}t$ since both $\dot{g}_\tau$ and $\mp \int_0^\tau \frac{1}{\lambda} \nabla \mathcal{L}(\theta_t) \operatorname{sgn}(w_\tau w_t)$ are finite quantities. Hence, for $\tau \in (T_{m+1}, T_{m+2})$,

$$\dot{g}_\tau = -\text{m.p.v.} \int_0^{\tau} \left[\left(\frac{\dot{w}_t}{w_t} \right)^2 \pm \frac{1}{\lambda} \nabla \mathcal{L}(\theta_t) \right] e^{-\frac{\tau - t}{\lambda}} \operatorname{sgn}(w_\tau w_t) \mathrm{d}t - \frac{1}{\lambda} \sum_{k=1}^{m+1} (-1)^{(m+1)-k} e^{-\frac{\tau - T_k}{\lambda}}.$$

Induction step for g_τ. We move on to untangle the limit which equals g_τ. By the induction hypothesis,

$$\begin{aligned}
\dot{g}_{T_-} &= -\text{m.p.v.} \int_0^{T_-} f_t e^{-\frac{T_- - t}{\lambda}} \operatorname{sgn}(w_{T_-} w_t) \mathrm{d}t - \frac{1}{\lambda} \sum_{k=1}^{m} (-1)^{m-k} e^{-\frac{T_- - T_k}{\lambda}} \\
g_{T_-} &= g_0 - \int_0^{T_-} \text{m.p.v.} \int_0^{t} f_s e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_t w_s) \mathrm{d}s \ \mathrm{d}t \\
&- \sum_{k=1}^{m} (-1)^{m-k} \left(1 - e^{-\frac{T_- - T_k}{\lambda}} \right) + 2 \sum_{1 \le i < j \le m} (-1)^{j-i} \left(1 - e^{-\frac{T_j - T_i}{\lambda}} \right).
\end{aligned}$$

[6] Whenever we write an equation of the form $A = B + O(\varepsilon^r)$ for some $r > 0$, we mean that $A = B + C$, where $|C| = O(\varepsilon^r)$.

Again, we can substitute $\operatorname{sgn}(w_{T_-})$ with $-\operatorname{sgn}(w_\tau)$, and after performing the familiar integral and limit manipulations, we obtain

$$
\begin{aligned}
g_\tau = g_0 &\mp \int_0^\tau \int_0^t \frac{1}{\lambda} \nabla \mathcal{L}(\theta_s) e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_t w_s) \mathrm{d}s \ \mathrm{d}t - \lim_{\varepsilon \to 0} \left[A_\varepsilon + B_\varepsilon + C_\varepsilon\right] \\
&- \sum_{k=1}^{m+1} (-1)^{(m+1)-k} \left(1 - e^{-\frac{\tau - T_k}{\lambda}}\right) + 2 \sum_{1 \le i < j \le m+1} (-1)^{j-i} \left(1 - e^{-\frac{T_j - T_i}{\lambda}}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
A_\varepsilon &= \int_0^{T_-} \text{m.p.v.} \int_0^t \left(\frac{\dot{w}_s}{w_s}\right)^2 e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_t w_s) \mathrm{d}s \ \mathrm{d}t \\
B_\varepsilon &= \int_{T_+}^\tau \text{m.p.v.} \int_0^{T_-} \left(\frac{\dot{w}_s}{w_s}\right)^2 e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_t w_s) \mathrm{d}s \ \mathrm{d}t \cdot e^{\frac{2\varepsilon}{\lambda}} \\
C_\varepsilon &= \int_{T_+}^\tau \int_{T_+}^t \left(\frac{\dot{w}_s}{w_s}\right)^2 e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_t w_s) \mathrm{d}s \ \mathrm{d}t.
\end{aligned}
$$

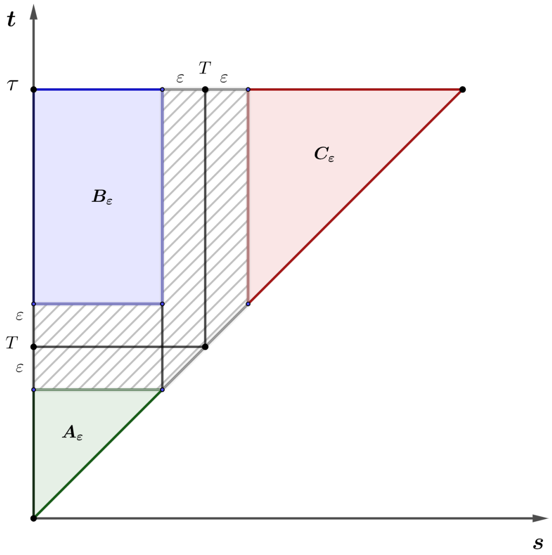

Figure 5: A visualisation of the areas over which we integrate $\left(\frac{\dot{w}_s}{w_s}\right)^2 e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_t w_s)$ in the above limit.

Notice that formally the limit $\lim_{\varepsilon \to 0}[A_\varepsilon + B_\varepsilon + C_\varepsilon]$ equals the integral quantity

$$
\int_0^\tau \text{m.p.v.} \int_0^t \left(\frac{\dot{w}_s}{w_s}\right)^2 e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_t w_s) \mathrm{d}s \ \mathrm{d}t,
$$

whose existence we just proved as a consequence of the fact that

$$
g_0 \mp \int_0^\tau \int_0^t \frac{1}{\lambda} \nabla \mathcal{L}(\theta_s) e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_t w_s) \mathrm{d}s \ \mathrm{d}t - \sum_{k=1}^{m+1} (-1)^{(m+1)-k} \left(1 - e^{-\frac{\tau - T_k}{\lambda}}\right) + 2 \sum_{1 \le i < j \le m+1} (-1)^{j-i} \left(1 - e^{-\frac{T_j - T_i}{\lambda}}\right) - g_\tau
$$

is well-defined and finite. Thus, for $\tau \in (T_{m+1}, T_{m+2})$,

$$
\begin{aligned}
g_\tau = g_0 &- \int_0^\tau \text{m.p.v.} \int_0^t \left[\left(\frac{\dot{w}_s}{w_s}\right)^2 \pm \frac{1}{\lambda} \nabla \mathcal{L}(\theta_s)\right] e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_t w_s) \mathrm{d}s \ \mathrm{d}t \\
&- \sum_{k=1}^{m+1} (-1)^{(m+1)-k} \left(1 - e^{-\frac{\tau - T_k}{\lambda}}\right) + 2 \sum_{1 \le i < j \le m+1} (-1)^{j-i} \left(1 - e^{-\frac{T_j - T_i}{\lambda}}\right).
\end{aligned}
$$

Proof of bounds. In order to conclude the induction step, we still have to prove the following bounds:

$$g_{T_+} - g_{T_-} = O(\varepsilon) \quad \text{and} \quad \dot{g}_{T_+} + \dot{g}_{T_-} + \frac{1}{\lambda} = O(\varepsilon).$$

Recall that $g_{T\pm\varepsilon} = \log|w_{T\pm\varepsilon}|$ and that $w_T = 0,\ \dot{w_T} \neq 0$. From the Taylor expansion of w_t, we know that

$$w_{T\pm\varepsilon} = \pm\varepsilon\dot{w}_T + O(\varepsilon^2).$$

Hence, $|w_{T+\varepsilon}/w_{T-\varepsilon}| = 1 + O(\varepsilon)$. Therefore, using the Taylor expansion of the logarithm around 1, we get that

$$|g_{T_+} - g_{T_-}| = |\log(1 + O(\varepsilon))| = O(\varepsilon).$$

Now, recall that $\dot{g}_{T\pm\varepsilon} = \dot{w}_{T\pm\varepsilon}/w_{T\pm\varepsilon}$ and observe that

$$\begin{aligned} w_{T\pm\varepsilon} &= \pm\varepsilon\dot{w}_T + \frac{1}{2}\varepsilon^2\ddot{w}_T + O(\varepsilon^3) \\ \dot{w}_{T\pm\varepsilon} &= \dot{w}_T \pm \varepsilon\ddot{w}_T + O(\varepsilon^2). \end{aligned}$$

Hence,

$$\frac{w_{T+\varepsilon}\dot{w}_{T-\varepsilon} + w_{T-\varepsilon}\dot{w}_{T+\varepsilon}}{w_{T+\varepsilon}w_{T-\varepsilon}} = \frac{-\varepsilon^2\dot{w}_T\ddot{w}_T + O(\varepsilon^3)}{-\varepsilon^2\dot{w}_T^2 + O(\varepsilon^3)} = \frac{\ddot{w}_T}{\dot{w}_T} + O(\varepsilon).$$

Since, $\lambda\ddot{w}_T + \dot{w}_T \pm \nabla\mathcal{L}(\theta_T) \odot w_T = 0$ and $w_T = 0$, we get that $\frac{\ddot{w}_T}{\dot{w}_T} = -\frac{1}{\lambda}$, which concludes the induction step.

Proof of Lemma 1. Thus, we proved that for $\tau \in (T_m, T_{m+1})$, $m \in \{0, 1, \ldots, N\}$,

$$\begin{aligned} \ln|w_\tau| = \ln|w_0| &- \int_0^\tau \text{m.p.v.} \int_0^t \left[\left(\frac{\dot{w}_s}{w_s}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right] e^{-\frac{t-s}{\lambda}}\,\mathrm{sgn}(w_t w_s)\mathrm{d}s\ \mathrm{d}t \\ &- \sum_{k=1}^{m}(-1)^{m-k}\left(1 - e^{-\frac{\tau - T_k}{\lambda}}\right) + 2\sum_{1\le i<j\le m}(-1)^{j-i}\left(1 - e^{-\frac{T_j - T_i}{\lambda}}\right). \end{aligned}$$

Recall that throughout our inductive proof we worked with a fixed coordinate $i \in [d]$ of $w_\pm$. Different coordinates of $w_\pm$ vanish at different points in time, so writing the sum the last line in a coordinate-agnostic way becomes impossible. Thus, deriving a simple expression for the full d-dimensional vector $w_{\pm,\tau}$ for any $\tau \in \mathbb{R}_{\geq 0}$ also becomes impossible. However, remembering that the finite nonzero limits $\lim_{\tau\to\infty}|w_{\pm,\tau}| = |w_{\pm,\infty}|$ exist and letting $\tau \to \infty$ yields an interesting result for the weights at infinity. Indeed, notice that for every vanishing time T, $\lim_{\tau\to\infty}\left(1 - e^{-\frac{\tau - T_k}{\lambda}}\right) = 1$. Hence,

$$\frac{1}{\lambda}\sum_{k=1}^{N}(-1)^{N-k}\left(1 - e^{-\frac{\tau - T_k}{\lambda}}\right) = \mathbf{1}_{\{N-\text{ odd}\}}.$$

For every $i \in [d]$, let $N_\pm(i)$ denote the number of vanishing points for the coordinate $w_\pm(i)$. Let us define the d-dimensional parity vectors $P_\pm \in \{0,1\}^d$ such that $P_\pm(i) \equiv N_\pm(i) \mod 2$. Let us also define the d-dimensional vectors $Q_\pm \in \mathbb{R}^d$ such that for each coordinate $k \in [d]$,

$$Q_\pm(k) \coloneqq -2\sum_{1\le i<j\le N_\pm(k)}(-1)^{j-i}\left(1 - e^{-\frac{T_{\pm,k}(j) - T_{\pm,k}(i)}{\lambda}}\right),$$

where $0 < T_{\pm,k}(1) < \cdots < T_{\pm,k}(N_\pm(k))$ denote the vanishing times of the weight $w_\pm(k)$. Hence, we obtain the formula

$$|w_{\pm,\infty}| = |w_{\pm,0}|e^{-(P_\pm + Q_\pm)}\exp\left(-\int_0^\infty \text{m.p.v.}\int_0^t \left[\left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right] e^{-\frac{t-s}{\lambda}}\,\mathrm{sgn}(w_{\pm,t}w_{\pm,s})\mathrm{d}s\ \mathrm{d}t\right). \quad (18)$$

Recall that in Lemma 3, we proved that the limit

$$\lim_{t\to\infty}\int_0^t \nabla\mathcal{L}(\theta_s)ds = \int_0^\infty \nabla\mathcal{L}(\theta_t)\mathrm{d}t = \frac{1}{\lambda}\int_0^\infty\int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}\mathrm{d}s\ \mathrm{d}t$$

exists and is finite. Therefore, we can decouple $\left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2$ and $\nabla\mathcal{L}(\theta_s)$ from the above integral and show that the following integral limits exist and are finite:

$$\int_0^\infty \text{m.p.v.} \int_0^t \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_{\pm,t}w_{\pm,s})\mathrm{d}s \ \mathrm{d}t \quad \text{and} \quad \int_0^\infty \int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_{\pm,t}w_{\pm,s})\mathrm{d}s \ \mathrm{d}t.$$

Hence, the integral quantities $\Omega_\pm$ from Lemma 1 are well-defined and finite. Thus, we finally proved Lemma 1.

Trajectory-Dependent Characterisation. We started this section with a promise for more insightful representations of $\Delta_\infty = |\alpha_{+,\infty}\alpha_{-,\infty}|$ and $\tilde{\theta}_0 = (\alpha_{+,\infty}^2 - \alpha_{-,\infty}^2)/4$. We now deliver on that promise.

Recall that $\alpha_{\pm,\infty} = w_{\pm,\infty}\exp\left(\pm\int_0^\infty \nabla\mathcal{L}(\theta_t)\mathrm{d}t\right)$ and notice that

$$\int_0^\infty \int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}\mathrm{d}s \ \mathrm{d}t - \int_0^\infty \int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_{\pm,t}w_{\pm,s})\mathrm{d}s \ \mathrm{d}t = 2\int_0^\infty \int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}} \mathbf{1}_{\{w_{\pm,t}w_{\pm,s}<0\}}\mathrm{d}s \ \mathrm{d}t.$$

Hence, using the formula for $w_\pm$ from Equation (18), we derive the following:

$$\begin{aligned}|\alpha_{\pm,\infty}| = |w_{\pm,0}|e^{-(P_\pm+Q_\pm)} &\odot \exp\left(-\int_0^\infty \text{m.p.v.}\int_0^t \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_{\pm,t}w_{\pm,s})\mathrm{d}s \ \mathrm{d}t\right) \\ &\odot \exp\left(\pm\frac{2}{\lambda}\int_0^\infty \int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}} \mathbf{1}_{\{w_{\pm,t}w_{\pm,s}<0\}}\mathrm{d}s \ \mathrm{d}t\right).\end{aligned}$$

Now, let

$$\Lambda_\pm \coloneqq \mp\frac{2}{\lambda}\int_0^\infty \int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}} \mathbf{1}_{\{w_{\pm,t}w_{\pm,s}<0\}}\mathrm{d}s \ \mathrm{d}t + P_\pm + Q_\pm, \tag{19}$$

where the quantities $P_\pm$ and $Q_\pm$ were defined in the previous paragraph. Notice that as we promised underneath Lemma 1, $\Lambda_\pm$ vanish whenever the balancedness Δ_t remains strictly positive. Using the abbreviation $I_\pm = \Omega_\pm + \Lambda_\pm$, we get that

$$|\alpha_{\pm,\infty}| = |w_{\pm,0}| \odot \exp\left(-I_\pm\right).$$

Multiplying $|\alpha_{+,\infty}|$ by $|\alpha_{-,\infty}|$, we derive a formula for the asymptotic balancedness:

$$\begin{aligned}\Delta_\infty = \Delta_0 e^{-(P_++P_-+Q_++Q_-)} &\odot \exp\left(-\int_0^\infty \text{m.p.v.}\int_0^t \left[\left(\frac{\dot{w}_{+,s}}{w_{+,s}}\right)^2 + \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right] e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_{+,t}w_{+,s})\mathrm{d}s \ \mathrm{d}t\right) \\ &\odot \exp\left(-\int_0^\tau \text{m.p.v.}\int_0^t \left[\left(\frac{\dot{w}_{-,s}}{w_{-,s}}\right)^2 - \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right] e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_{-,t}w_{-,s})\mathrm{d}s \ \mathrm{d}t\right) \\ &\odot \exp\left(\frac{2}{\lambda}\int_0^\infty \int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}} \left[\mathbf{1}_{\{w_{+,t}w_{+,s}<0\}} - \mathbf{1}_{\{w_{-,t}w_{-,s}<0\}}\right]\mathrm{d}s \ \mathrm{d}t\right).\end{aligned} \tag{20}$$

Now, we can write Δ_∞ and $\tilde{\theta}_0$ more succinctly as

$$\Delta_\infty = \Delta_0 \odot \exp\Big(-\left(I_+ + I_-\right)\Big)$$

and

$$\tilde{\theta}_0 = \frac{1}{4}\Big(w_{+,0}^2 \odot \exp\left(-2I_+\right) - w_{-,0}^2 \odot \exp\left(-2I_-\right)\Big),$$

which concludes the proof of Theorem 1. □

C.3.3 Consequences for Generalisation.

We just proved that whenever MGF on a diagonal linear network converges and the balancedness at infinity is nonzero, we can characterize the recovered interpolator through the implicit regularization problem

$$\begin{aligned}\theta^{\texttt{MGF}} &= \operatorname*{argmin}_{\theta^\star\in\mathcal{S}} \ D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0) \\ &= \operatorname*{argmin}_{\theta^\star\in\mathcal{S}} \left[\psi_{\Delta_\infty}(\theta^\star) - \langle\nabla\psi_{\Delta_\infty}(\tilde{\theta}_0), \theta^\star\rangle\right].\end{aligned}$$

Since

$$\psi_{\Delta_\infty}(\theta) = \frac{1}{4}\sum_{i=1}^{d}\left(2\theta_i \operatorname{arcsinh}\left(\frac{2\theta_i}{\Delta_{\infty,i}}\right) - \sqrt{4\theta_i^2 + \Delta_{\infty,i}^2} + \Delta_{\infty,i}\right)$$

and

$$\nabla\psi_{\Delta_\infty}(\theta) = \frac{1}{2}\operatorname{arcsinh}\left(\frac{2\theta}{\Delta_\infty}\right),$$

for a small asymptotic balancedness $\Delta_\infty = O(\Delta_0) = O(\alpha^2)$ and small perturbed initialisation $|\tilde{\theta}_0| = O(\alpha^2) \ll |\theta^\star|$, we would expect $\psi_{\Delta_\infty}(\theta^\star)$ to dominate $\langle \nabla\psi_{\Delta_\infty}(\tilde{\theta}_0), \theta^\star\rangle$. More formally, for a fixed $\theta^\star \in \mathcal{S}$ and small Δ_∞ and $\tilde{\theta}_0$, we have the following asymptotic equivalence:

$$\psi_{\Delta_\infty}(\theta^\star) \underset{\alpha\to 0}{\sim} D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0).$$

Hence, $\theta^{\texttt{MGF}} \approx \operatorname{argmin}_{\theta^\star\in\mathcal{S}} \psi_{\Delta_\infty}(\theta^\star) = \theta^{\texttt{GF}}_{\Delta_\infty}$ as we discussed in Section 3.1. So, if $\Delta_\infty < \Delta_0$, Lemma 4 implies that the MGF predictor will benefit from better sparsity guarantees than the GF solution.

Therefore, to recap, for a small initialisation scale α and provided that the bounds $\Delta_\infty = O(\alpha^2)$ and $\tilde{\theta}_0 = O(\alpha^2)$ hold, we conclude that the asymptotic balancedness at infinity Δ_∞ roughly controls the sparsity of the recovered interpolator. And when $\Delta_\infty < \Delta_0$, $\theta^{\texttt{MGF}}$ will be sparser than $\theta^{\texttt{GF}}_\alpha$. Unfortunately, without the assumption that the balancedness Δ_t remains strictly positive for all $t \in [0, +\infty]$, we cannot formally compare Δ_∞ and $\tilde{\theta}_0$ with α.

Note that even without the bounds $\Delta_\infty = O(\alpha^2)$ and $\tilde{\theta}_0 = O(\alpha^2)$, if $|\tilde{\theta}_0| \ll |\theta^\star|$, then $\psi_{\Delta_\infty}(\theta^\star)$ still dominates $\langle \nabla\psi_{\Delta_\infty}(\tilde{\theta}_0), \theta^\star\rangle$. Indeed, our experiments clearly show that the perturbation term $\tilde{\theta}_0$ can safely be ignored since $\theta^{\texttt{MGF}} \approx \operatorname{argmin}_{\theta^\star\in\mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$ (see the discussion around Figure 8.)

C.4 Non-Vanishing Balancedness

If we work under the assumption that the balancedness $\Delta_t = |w_{+,t}w_{-,t}|$ never vanishes, then much of the analysis from Appendix C.3.2 greatly simplifies. First, the integral quantities $P_\pm$ and $Q_\pm$ from the previous subsection become 0. Second, the multipliers $\operatorname{sgn}(w_{\pm,t}w_{\pm,s})$ become equal to 1 for all $t, s \in \mathbb{R}_{\geq 0}$. Hence, using Fubini's Theorem as in the proof of Lemma 5, we get that

$$\begin{aligned}\int_0^\tau \text{m.p.v.}\int_0^t \left[\left(\frac{\dot{w}_s}{w_s}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right] e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_t w_s)\mathrm{d}s\ \mathrm{d}t \\ = \int_0^\tau\int_0^t \left[\left(\frac{\dot{w}_s}{w_s}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right] e^{-\frac{t-s}{\lambda}}\mathrm{d}s\ \mathrm{d}t \\ = \lambda\int_0^\tau \left(\frac{\dot{w}_t}{w_t}\right)^2\left(1 - e^{-\frac{\tau-t}{\lambda}}\right)\mathrm{d}t \pm \int_0^\tau \nabla\mathcal{L}(\theta_t)\left(1 - e^{-\frac{\tau-t}{\lambda}}\right)\mathrm{d}t.\end{aligned}$$

Therefore, referencing eq. (18), we can express the evolution of the iterates as follows:

$$w_{\pm,\tau} = w_{\pm,0}\exp\left(-\lambda\int_0^\tau \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2\left(1 - e^{-\frac{\tau-t}{\lambda}}\right)\mathrm{d}t\right)\exp\left(\mp\int_0^\tau \nabla\mathcal{L}(\theta_s)\left(1 - e^{-\frac{\tau-t}{\lambda}}\right)\mathrm{d}t\right). \tag{21}$$

Thus, the balancedness evolves as

$$\Delta_t = \Delta_0\exp\left(-\lambda\int_0^\tau \left[\left(\frac{\dot{w}_{+,t}}{w_{+,t}}\right)^2 + \left(\frac{\dot{w}_{-,t}}{w_{-,t}}\right)^2\right]\left(1 - e^{-\frac{\tau-t}{\lambda}}\right)\mathrm{d}t\right). \tag{22}$$

Now, from Lemma 3, we know that $\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2$ is integrable and that

$$\lim_{\tau\to\infty}\int_0^\tau \nabla\mathcal{L}(\theta_s)\left(1 - e^{-\frac{\tau-t}{\lambda}}\right)\mathrm{d}t = \int_0^\infty \nabla\mathcal{L}(\theta_s)\mathrm{d}t$$

exists. Furthermore, from Lemma 6, we know that that

$$\lim_{\tau\to\infty}\int_0^\tau \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2\left(1 - e^{-\frac{\tau-t}{\lambda}}\right)\mathrm{d}t = \int_0^\infty \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2)\mathrm{d}t.$$

Therefore, letting $\tau \to \infty$, we obtain the formulas

$$w_{\pm,\tau} = w_{\pm,0} \exp\left(-\lambda \int_0^\infty \Big(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\Big)^2 \mathrm{d}t\right) \exp\left(\mp \int_0^\infty \nabla\mathcal{L}(\theta_s)\mathrm{d}t\right) \tag{23}$$

$$\Delta_\infty = \Delta_0 \exp\left(-\lambda \int_0^\infty \left[\Big(\frac{\dot{w}_{+,t}}{w_{+,t}}\Big)^2 + \Big(\frac{\dot{w}_{-,t}}{w_{-,t}}\Big)^2\right] \mathrm{d}t\right). \tag{24}$$

Hence, clearly, $\Delta_\infty < \Delta_0$.

Finally, let us consider how the perturbed initialisation $\tilde{\theta}_0$ looks like when Δ_t remains nonzero. Recall that $\tilde{\theta}_0 = (\alpha_+^2 - \alpha_-^2)/4$, where $\alpha_{\pm,\infty} = w_{\pm,\infty} \exp\big(\pm \int_0^\infty \nabla\mathcal{L}(\theta_t)\mathrm{d}t\big)$. Thus,

$$\alpha_{\pm,\infty} = w_{\pm,0} \exp\left(-\lambda \int_0^\infty \Big(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\Big)^2 \mathrm{d}t\right)$$

and

$$\tilde{\theta}_0 = \frac{1}{4}\left[w_{+,0}^2 \exp\left(-2\lambda \int_0^\infty \Big(\frac{\dot{w}_{+,t}}{w_{+,t}}\Big)^2 \mathrm{d}t\right) - w_{,0}^2 \exp\left(-2\lambda \int_0^\infty \Big(\frac{\dot{w}_{-,t}}{w_{-,t}}\Big)^2 \mathrm{d}t\right)\right]. \tag{25}$$

Now, $\alpha_{\pm,\infty} < w_{\pm,0} \leq 2\alpha$, where $\alpha = \max(\|u_0\|_\infty, \|v_0\|_\infty)$ stood for the initialisation scale. Hence, $|\tilde{\theta}_0| < \alpha^2$.

Therefore, we just proved

Corollary 2. *For $\lambda > 0$, if the balancedness Δ_t remains strictly positive during training (i.e. $\Delta_t \neq 0$ for $t \in [0, +\infty]$), then the perturbed initialisation satisfies $|\tilde{\theta}_0| < \alpha^2$ and*

$$\Delta_\infty = \Delta_0 \odot \exp\Big(-\lambda \int_0^\infty \Big(\frac{\dot{w}_{+,t}}{w_{+,t}}\Big)^2 + \Big(\frac{\dot{w}_{-,t}}{w_{-,t}}\Big)^2 \mathrm{d}t\Big).$$

Importantly, $\Delta_\infty < \Delta_0$.

C.5 Behaviour of Δ_∞ for Small Values of λ

Since a precise asymptotic result for small λ is technically difficult, in this section we focus on giving some qualitative results. For $\lambda > 0$, recall that our iterates follow

$$\lambda \ddot{w}_{\pm,t}^{(\lambda)} + \dot{w}_{\pm,t}^{(\lambda)} \pm \nabla\mathcal{L}(\theta_t^{(\lambda)}) \odot w_{\pm,t}^{(\lambda)} = 0,$$

where we explicitly highlight the dependency on λ. Therefore, we have

$$\frac{\dot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}^{(\lambda)}} = \mp\nabla\mathcal{L}(\theta_t^{(\lambda)}) - \lambda \frac{\ddot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}}$$

and

$$\Big(\frac{\dot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}^{(\lambda)}}\Big)^2 = \nabla\mathcal{L}(\theta_t^{(\lambda)})^2 + \lambda^2 \Big(\frac{\ddot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}}\Big)^2 \pm 2\lambda\mathcal{L}(\theta_t^{(\lambda)})\Big(\frac{\ddot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}}\Big).$$

Informally, we expect $(t \mapsto \nabla L(\theta_t^{(\lambda)}))_{0<\lambda\leq 1} \in L^2(0, +\infty)$ and $(t \mapsto \lambda \frac{\ddot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}})_\lambda \underset{\lambda\to 0}{\longrightarrow} 0$ in L^2-norm (see Theorem 5.1 in Attouch et al. (2000)). Hence, we get

$$\int_0^\infty \Big(\frac{\dot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}^{(\lambda)}}\Big)^2 \underset{\lambda\to 0}{\sim} \int_0^\infty \nabla\mathcal{L}(\theta_t^{(\lambda)})^2 \mathrm{d}t$$

and

$$\Delta_\infty \underset{\lambda\to 0}{\approx} \Delta_0 \exp\Big(-2\lambda \int_0^\infty \nabla\mathcal{L}(\theta_s^{(\lambda)})^2 \mathrm{d}s\Big).$$

D Discrete-Time Results

In this section, we cover the proofs of our discrete-time results from Section 4. We first recall the SMGD recursion (13) with the $w_\pm$-parametrisation from Appendix B. Initialised at $w_{\pm,0} = w_{\pm,1} \in \mathbb{R}^d$, for $k \geq 1$, the iterates follow

$$w_{\pm,k+1} = w_{\pm,k} \mp \gamma \nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k) \odot w_{\pm,k} + \beta(w_{\pm,k} - w_{\pm,k-1}). \tag{26}$$

In what follows, we will adapt our continuous-time proof technique to the discrete case and identify a quantity which follows a momentum mirror descent with time-varying potentials. Our proofs closely follow the proof techniques from (Even et al., 2023) which considers SGD without momentum.

D.1 Proof of Lemma 2, Theorem 2 and Corollary 3

We start by recalling the main-paper results. The first lemma introduces two convergent series which will appear in our main result.

Lemma 2. *The following two sums S_+ and S_- converge to finite vectors:*

$$S_\pm = \frac{1}{1-\beta} \sum_{k=1}^{\infty} \left[r\Big(\frac{w_{\pm,k+1}}{w_{\pm,k}}\Big) + \beta r\Big(\frac{w_{\pm,k}}{w_{\pm,k+1}}\Big) \right],$$

where $r(z) = (z-1) - \ln(|z|)$ for $z \neq 0$.

The proof of the lemma can be found in the proof of the following main theorem.

Theorem 2. *The solution $\theta^{\texttt{SMGD}}$ of SMGD (9) interpolates the dataset and satisfies the following implicit regularisation:*

$$\theta^{\texttt{SMGD}} = \underset{\theta^\star \in \mathcal{S}}{\operatorname{argmin}}\ D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0).$$

In the above expression, $D_{\psi_{\Delta_\infty}}$ denotes the Bregman divergence with potential ψ_{Δ_∞}, where the asymptotic balancedness equals

$$\Delta_\infty = \Delta_0 \odot \exp\big(-(S_+ + S_-)\big)$$

and $\tilde{\theta}_0 = \frac{1}{4}(w_{+,0}^2 \odot \exp(-2S_+)) - w_{-,0}^2 \odot \exp(-2S_-))$ denotes a perturbed initialisation term.

Proving Convergence towards an Interpolator. By Assumption 3, we have that the iterates $w_{\pm,k}$ converge towards limiting weights $w_{\pm,\infty}$ and that the predictors converge towards a vector $\theta^{\texttt{MGF}}$. Taking the limit in Equation (26), we get that $\lim_{k\to\infty} \nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k) \odot w_{\pm,k} = 0$. By Assumption 4, $w_{\pm,\infty}$ have non-zero coordinates. Therefore, $\lim_{k\to\infty} \nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k) = 0$. For any fixed batch $\mathcal{B} \subset \{1, \cdots, n\}$, the sampling with or without replacement is such that (almost surely) the set $M_k := \{k \geq 0, \mathcal{B}_k = \mathcal{B}\}$ is infinite. Hence, by continuity of $\nabla L_{\mathcal{B}}$, $\lim_{k\to\infty, k\in M_k} \nabla L_{\mathcal{B}}(\theta_k) = \nabla L_{\mathcal{B}}(\theta^{\texttt{SMGD}})$. Therefore, for all fixed batches $\mathcal{B}$, $\nabla L_{\mathcal{B}}(\theta^{\texttt{SMGD}}) = 0$ and hence $\theta^{\texttt{SMGD}}$ interpolates the dataset.

From here on now, for ease of notation, we do the proof for deterministic MGD. The proof for stochastic MGD is exactly the same after replacing $\nabla\mathcal{L}(\theta_k)$ with $\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)$.

Deriving the Momentum Mirror Descent. Recall that the set of pairs (γ, β) such that there exists k where $w_{\pm,k} = 0$ is negligible in $\mathbb{R}^2$. We can hence assume that the iterates are never exactly zero, and we consider the logarithmic reparametrisation of the iterates $w_{\pm,k}$ as

$$g_{\pm,k} = \begin{cases} \ln(w_{\pm,k}), & \text{if } w_{\pm,k} > 0, \\ \ln(|w_{\pm,k}|) + i\pi, & \text{if } w_{\pm,k} < 0. \end{cases}$$

This way we have that that $w_{\pm,k} = \exp(g_{\pm,k})$ for all k. Equation (26) then becomes

$$\exp(g_{\pm,k+1}) = \exp(g_{\pm,k}) \mp \gamma\nabla\mathcal{L}(\theta_k) \odot \exp(g_{\pm,k}) + \beta(\exp(g_{\pm,k}) - \exp(g_{\pm,k-1})).$$

Dividing by $\exp(g_{\pm,k})$ yields

$$\exp(g_{\pm,k+1} - g_{\pm,k}) = 1 \mp \gamma \nabla \mathcal{L}(\theta_k) + \beta(1 - \exp(-(g_{\pm,k} - g_{\pm,k-1})).$$

Now, for $k \geq 1$, let $\delta_{\pm,k} = g_{\pm,k} - g_{\pm,k-1}$ so that we can more compactly write the above recurrence as

$$\exp(\delta_{\pm,k+1}) = 1 \mp \gamma \nabla \mathcal{L}(\theta_k) + \beta(1 - \exp(-\delta_{\pm,k})).$$

The trick, inspired by Even et al. (2023), is to consider the function $q(z) = \exp(z) - (1 + z)$ for $z \in \mathbb{C}$. Importantly, note that $q(z) \geq 0$ for $z \in \mathbb{R}$. Using this function, we can now rewrite the recurrence as

$$\delta_{\pm,k+1} + q(\delta_{\pm,k+1}) = \mp \gamma \nabla \mathcal{L}(\theta_k) + \beta(\delta_{\pm,k} - q(-\delta_{\pm,k})).$$

Setting the residues $Q_{\pm,k} \coloneqq q(\delta_{\pm,k+1}) + \beta q(-\delta_{\pm,k})$ leads to

$$\delta_{\pm,k+1} = \beta \delta_{\pm,k} \mp \gamma \nabla \mathcal{L}(\theta_k) - Q_{\pm,k}.$$

This can be seen as a first-order recurrence relation with variable coefficients. For $\beta = 0$ we exactly recover the analysis from Even et al. (2023). For $\beta > 0$, since $\delta_{\pm,1} = 0$, for $m \geq 1$, we can expand the relation as

$$\delta_{\pm,m+1} = -\sum_{k=1}^{m} \beta^{m-k} \left[\pm \gamma \nabla \mathcal{L}(\theta_k) + Q_{\pm,k} \right].$$

Summing over m, we now get for $N \geq 1$ the following expression:

$$\begin{aligned} g_{\pm,N+1} - g_{\pm,1} &= \sum_{m=1}^{N} \delta_{\pm,m+1} \\ &= -\sum_{m=1}^{N} \sum_{k=1}^{m} \beta^{m-k} \left[\pm \gamma \nabla \mathcal{L}(\theta_k) + Q_{\pm,k} \right] \end{aligned}$$

Finally, taking the exponential for $N \geq 1$, we obtain

$$\begin{aligned} w_{\pm,N+1} &= w_{\pm,0} \exp\left(-\sum_{m=1}^{N} \sum_{k=1}^{m} \beta^{m-k} \left[\pm \gamma \nabla \mathcal{L}(\theta_k) + Q_{\pm,k} \right] \right) \\ &= w_{\pm,0} \exp\left(\pm \sum_{m=1}^{N} \sum_{k=1}^{m} \beta^{m-k} Q_{\pm,k} \right) \exp\left(\mp \gamma \sum_{m=1}^{N} \sum_{k=1}^{m} \beta^{m-k} \nabla \mathcal{L}(\theta_k) \right) \\ &= w_{\pm,0} \exp\left(-\frac{1}{1-\beta} \sum_{m=1}^{N} (1 - \beta^{N+1-m}) Q_{\pm,m} \right) \exp\left(\mp \frac{\gamma}{1-\beta} \sum_{m=1}^{N} (1 - \beta^{N+1-m}) \nabla \mathcal{L}(\theta_m) \right), \end{aligned}$$

where the last equality is obtained by changing the order of summation. Following our continuous-time approach, for $N \geq 2$, we define $\alpha_{\pm,N+1}$ as

$$\begin{aligned} \alpha_{\pm,N+1} &\coloneqq w_{\pm,0} \exp\left(\pm \sum_{m=1}^{N} \sum_{k=1}^{m} \beta^{m-k} Q_{\pm,k} \right) \\ &= w_{\pm,0} \exp\left(-\frac{1}{1-\beta} \sum_{m=1}^{N} (1 - \beta^{N+1-m}) Q_{\pm,m} \right). \end{aligned} \tag{27}$$

We can now write the iterates $w_{\pm,k}$ as

$$w_{\pm,N+1} = \alpha_{\pm,N+1} \exp\big(\mp \gamma \sum_{m=1}^{N} \sum_{k=1}^{m} \beta^{m-k} \nabla \mathcal{L}(\theta_k) \big).$$

Thus, the regression parameter θ_N becomes

$$\begin{aligned}\theta_{N+1} &= \frac{1}{4}(w^2_{+,N+1} - w^2_{-,N+1})\\ &= \frac{1}{4}\alpha^2_{+,N+1}\exp\big(-2\gamma\sum_{m=1}^{N}\sum_{k=1}^{m}\beta^{m-k}\nabla\mathcal{L}(\theta_k)\big) - \frac{1}{4}\alpha^2_{-,N+1}\exp\big(2\gamma\sum_{m=1}^{N}\sum_{k=1}^{m}\beta^{m-k}\nabla\mathcal{L}(\theta_k)\big)\\ &= \frac{1}{2}\Delta_{N+1}\ \sinh\left(-2\gamma\sum_{m=1}^{N}\sum_{k=1}^{m}\beta^{m-k}\nabla\mathcal{L}(\theta_k) + \operatorname{arcsinh}\left(\frac{\alpha^2_{+,N+1}-\alpha^2_{-,N+1}}{2\Delta_{N+1}}\right)\right),\end{aligned}$$

where we recall that $\Delta_N = |w_{+,N}w_{-,N}| = |\alpha_{+,N}\alpha_{-,N}|$. Hence, similar to the continuous case,

$$\frac{1}{2}\operatorname{arcsinh}\left(\frac{2\theta_{N+1}}{\Delta_{N+1}}\right) - \frac{1}{2}\operatorname{arcsinh}\left(\frac{\alpha^2_{+,N+1}-\alpha^2_{-,N+1}}{2\Delta_{N+1}}\right) = -\gamma\sum_{m=1}^{N}\sum_{k=1}^{m}\beta^{m-k}\nabla\mathcal{L}(\theta_k).$$

For $N \geq 1$, the above identity becomes exactly

$$\nabla\Phi_{N+1}(\theta_{N+1}) = -\gamma\sum_{m=1}^{N}\sum_{k=1}^{m}\beta^{m-k}\nabla\mathcal{L}(\theta_k), \tag{28}$$

where the time-varying potential $\Phi_N : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\begin{aligned}\Phi_N(\theta) &= \frac{1}{4}\sum_{i=1}^{d}\left(2\theta_i\operatorname{arcsinh}\left(\frac{2\theta_i}{\Delta_{N,i}}\right) - \sqrt{4\theta_i^2 + \Delta^2_{N,i}} + \Delta_{N,i}\right) + \langle\phi_N, \theta\rangle\\ &= \psi_{\Delta_N}(\theta) + \langle\phi_N, \theta\rangle,\end{aligned}$$

where $\phi_N = \frac{1}{2}\operatorname{arcsinh}\left(\frac{\alpha^2_{+,N}-\alpha^2_{-,N}}{2\Delta_N}\right)$ and ψ_{Δ_N} is the hyperbolic entropy defined in Equation (6). Notice that with this definition we arrive at the following time-varying momentum mirror descent for $N \geq 1$:

$$\nabla\Phi_{N+1}(\theta_{N+1}) = \nabla\Phi_N(\theta_N) - \gamma\nabla\mathcal{L}(\theta_N) + \beta(\nabla\Phi_N(\theta_N) - \nabla\Phi_{N-1}(\theta_{N-1})). \tag{29}$$

Convergent Quantities. From Lemma 7, we have that $\alpha_{\pm,N}$ must converge and that the limiting vectors $\alpha_{\pm,\infty}$ have non-zero coordinates. Therefore, the series $\sum_{m=1}^{\infty}\sum_{k=1}^{m}\beta^{m-k}Q_{\pm,k}$ are convergent and their terms must hence converge to zero: $\sum_{k=1}^{m}\beta^{m-k}Q_{\pm,k} \underset{m\to\infty}{\longrightarrow} 0$. Therefore,

$$\alpha_{\pm,N} \to \alpha_{\pm,\infty} = w_{\pm,0}\exp\left(-\frac{1}{1-\beta}\sum_{m=1}^{\infty}Q_{\pm,m}\right).$$

We now develop the formulas for $Q_{\pm,m}$ in order to arrive at the sums $S_\pm$ from Lemma 2. Recall that for $m \geq 1$, $Q_{\pm,m} = q(\delta_{\pm,m+1}) + \beta q(-\delta_{\pm,m})$ and $\delta_{\pm,1} = q(\delta_{\pm,1}) = 0$. Therefore,

$$\begin{aligned}\sum_{m=1}^{\infty}Q_{\pm,m} &= \sum_{m=1}^{\infty}q(\delta_{\pm,m+1}) + \beta q(-\delta_{\pm,m})\\ &= \sum_{m=1}^{\infty}q(\delta_{\pm,m+1}) + \beta q(-\delta_{\pm,m+1}).\end{aligned}$$

Since $\delta_{\pm,m+1} = g_{\pm,m+1} - g_{\pm,m}$, we have

$$\delta_{\pm,m+1} = \begin{cases}\ln\left(\frac{w_{\pm,m+1}}{w_{\pm,m}}\right) & \text{if } w_{\pm,m+1} \text{ and } w_{\pm,m} \text{ have the same sign,}\\ \ln\left(\left|\frac{w_{\pm,m+1}}{w_{\pm,m}}\right|\right) + \operatorname{sgn}(w_{\pm,m})i\pi & \text{if they have different signs.}\end{cases}$$

It remains to notice that since $q(z) = \exp(z) - (1+z)$, we get that

$$q(\ln(z)) = (z-1) - \ln(z) \qquad \text{for } z \in \mathbb{R}_{>0},$$
$$q(\ln(|z|) \pm i\pi) = (z-1) - (\ln(|z|) \pm i\pi) \qquad \text{for } z \in \mathbb{R}_{<0}.$$

Therefore letting $r(z) = (z-1) - \ln(|z|)$ as in Lemma 2, we get

$$\begin{aligned} q(\delta_{\pm,m+1}) &= r\Big(\frac{w_{\pm,m+1}}{w_{\pm,m}}\Big) - \xi_{\pm,m}\,\mathrm{sgn}(w_{\pm,m}) i\pi \\ q(-\delta_{\pm,m+1}) &= r\Big(\frac{w_{\pm,m}}{w_{\pm,m+1}}\Big) + \xi_{\pm,m}\,\mathrm{sgn}(w_{\pm,m}) i\pi, \end{aligned}$$

where $\xi_{\pm,m} = 0$ if $\mathrm{sgn}(w_{\pm,m+1}) = \mathrm{sgn}(w_{\pm,m})$ and 1 otherwise. This leads to

$$\begin{aligned} \frac{1}{1-\beta}\sum_{m=1}^{\infty} Q_{\pm,m} &= \frac{1}{1-\beta}\sum_{m=1}^{\infty}\Big[r\Big(\frac{w_{\pm,m+1}}{w_{\pm,m}}\Big) + \beta r\Big(\frac{w_{\pm,m}}{w_{\pm,m+1}}\Big)\Big] - \sum_{m=1}^{\infty}\xi_{\pm,m}\,\mathrm{sgn}(w_{\pm,m}) i\pi \\ &= S_{\pm} - \sum_{m=1}^{\infty}\xi_{\pm,m}\,\mathrm{sgn}(w_{\pm,m}) i\pi < \infty. \end{aligned}$$

The last equality is due to the definition of $S_{\pm}$ from Lemma 2, and the last inequality is due to the summability of $(Q_{\pm,m})_m$. This therefore proves lemma Lemma 2. Now notice that

$$\alpha_{\pm,\infty}^2 = w_{\pm,0}^2 \exp\left(-2S_{\pm}\right).$$

Since $\Delta_\infty = |\alpha_{+,\infty}\alpha_{-,\infty}|$, we finally get that

$$\Delta_\infty = \Delta_0 \odot \exp\Big(-(S_+ + S_-)\Big).$$

Implicit Regularisation Problem. Notice that

$$\nabla\Phi_{N+1}(\theta_{N+1}) = -\gamma\sum_{m=1}^{N}\sum_{k=1}^{m}\beta^{m-k}\nabla\mathcal{L}(\theta_k) \in \mathtt{span}(x_1, \cdots, x_n).$$

Let $\Phi_\infty(\theta) \coloneqq \psi_{\Delta_\infty}(\theta) + \langle\phi_\infty, \theta\rangle$ and consider

$$\nabla\Phi_\infty(\theta^{\mathtt{MGD}}) = (\nabla\Phi_\infty(\theta^{\mathtt{MGD}}) - \nabla\Phi_\infty(\theta_N)) + (\nabla\Phi_\infty(\theta_N) - \nabla\Phi_N(\theta_N)) + \nabla\Phi_N(\theta_N).$$

The first two terms converge to 0: the first due to the convergence $\theta_N \to \theta^{\mathtt{MGD}}$ and the second due to the uniform convergence of $\nabla\Phi_N$ to $\nabla\Phi_\infty$ on compact sets. The last term is in $\mathtt{span}(x_1, \cdots, x_n)$ for all N. Therefore, we get that $\nabla\Phi_\infty(\theta_\infty) \in \mathtt{span}(x_1, \cdots, x_n)$, and following the exact same proof as in the continuous-time framework, we finally get that

$$\theta^{\mathtt{MGD}} = \underset{\theta^\star \in \mathcal{S}}{\operatorname{argmin}}\, D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde\theta_0)$$

where

$$\begin{aligned} \tilde\theta_0 &= \frac{\alpha_{+,\infty}^2 - \alpha_{-,\infty}^2}{4} \\ &= \frac{1}{4}\Big(w_{+,0}^2 \odot \exp\left(-2S_+\right) - w_{-,0}^2 \odot \exp\left(-2S_-\right)\Big). \end{aligned}$$

□

We recall and prove the following corollary.

Corollary 3. *For $\gamma, \beta > 0$, if the iterates $w_{\pm,k} = (u_k \pm v_k)$ do not change sign during training, then $|\tilde\theta_0| < \alpha^2$ and $\Delta_\infty < \Delta_0$.*

Proof. The corollary follows from the fact that if the iterates $w_{\pm,k}$ do not change sign, then since $r(z) \geq 0$ for $z > 0$, we get that $S_\pm > 0$ and $\Delta_\infty < \Delta_0$. Furthermore, $|\tilde\theta_0| < \max(w_{+,0}^2, w_{-,0}^2)/4 \leq \alpha^2$ □

D.2 Link to the Continuous-Time Result.

In this subsection we link our continuous results with the discrete when the iterates do not cross zero. Indeed, at first sight, the discrete-time expression for Δ_∞ might seem quite different from its continuous-time counterpart:

$$\Delta_\infty^{\texttt{MGD}} = \Delta_0 \exp\left(-\frac{1}{1-\beta}\sum_{k=1}^{\infty}\left[r\left(\frac{w_{+,k+1}}{w_{+,k}}\right) + r\left(\frac{w_{-,k+1}}{w_{-,k}}\right)\right] + \beta\left[r\left(\frac{w_{+,k}}{w_{+,k+1}}\right) + r\left(\frac{w_{-,k}}{w_{-,k+1}}\right)\right]\right)$$
$$\Delta_\infty^{\texttt{MGF}} = \Delta_0 \exp\left(-\lambda\int_0^\infty \left(\frac{\dot w_{+,t}}{w_{+,t}}\right)^2 + \left(\frac{\dot w_{-,t}}{w_{-,t}}\right)^2 \mathrm{d}t\right).$$

However, upon closer inspection, by letting the discretisation step $\varepsilon = \sqrt{\lambda\gamma} = \frac{\gamma}{(1-\beta)}$ from Proposition 1 go to 0, we can recover the continuous-time result. Indeed, as $\varepsilon \to 0$, we expect successive iterates $w_{\pm,k}$ to be close and hence $w_{\pm,k+1}/w_{\pm,k} \approx 1$. Now, since $r(z) \sim_{z\to 1} (z-1)^2/2$, we roughly have

$$r\Big(\frac{w_{\pm,k+1}}{w_{\pm,k}}\Big) \approx \frac{1}{2}\Big(\frac{w_{\pm,k+1} - w_{\pm,k}}{w_{\pm,k}}\Big)^2$$

and

$$r\Big(\frac{w_{\pm,k}}{w_{\pm,k+1}}\Big) \approx \frac{1}{2}\Big(\frac{w_{\pm,k+1} - w_{\pm,k}}{w_{\pm,k+1}}\Big)^2 \approx \frac{1}{2}\Big(\frac{w_{\pm,k+1} - w_{\pm,k}}{w_{\pm,k}}\Big)^2$$

Putting the approximations together:

$$\begin{aligned}\frac{1}{1-\beta}\sum_k\left[r\Big(\frac{w_{\pm,k+1}}{w_{\pm,k}}\Big) + \beta r\Big(\frac{w_{\pm,k}}{w_{\pm,k+1}}\Big)\right] &\approx \frac{1}{2}\frac{\varepsilon(1+\beta)}{1-\beta}\sum_k\left(\frac{w_{\pm,k+1} - w_{\pm,k}}{\varepsilon}\right)^2\frac{1}{(w_{\pm,k})^2}\cdot\varepsilon \\ &\approx \frac{1+\beta}{2}\frac{\gamma}{(1-\beta)^2}\int_0^\infty\left(\frac{\dot w_{\pm,t}}{w_{\pm,t}}\right)^2\mathrm{d}t \\ &= \frac{1+\beta}{2}\lambda\int_0^\infty\left(\frac{\dot w_{\pm,t}}{w_{\pm,t}}\right)^2\mathrm{d}t.\end{aligned}$$

Notice that in order for λ to remain constant and ε to go to 0, we must both have $\gamma \to 0$ and $\beta \to 1$. Hence, $(1+\beta)/2 \to 1$, and we recover the continuous-time expression for the balancedness.

However, note that when the iterates cross zero it is unclear to the authors how the continuous formula and its discrete counterpart compare.

Another Safe-Check Computation. Recall that MGD with stepsize γ and momentum parameter β corresponds to the discretisation of MGF with $\lambda = \gamma/(1-\beta)^2$ and discretisation step $\varepsilon = \sqrt{\lambda\gamma}$. To check the consistency between the discrete time equations and continuous time equations, we look at the value of $\exp(-\frac{t-s}{\lambda})$ and times '$t = m\varepsilon$' and '$s = k\varepsilon$':

$$\begin{aligned}\exp(-\frac{t-s}{\lambda}) &= \exp(-\frac{(m-k)\varepsilon}{\lambda}) \\ &= \exp(-(m-k)(1-\beta)) \\ &= [\exp(\beta-1)]^{m-k} \\ &\sim_{\beta\to 1} \beta^{m-k}.\end{aligned}$$

This small computation serves as a safe-check, affirming the correspondence between the continuous-time analysis expression $\exp(-\frac{t-s}{\lambda})$ and its discrete-time counterpart β^{m-k}.

E Technical Lemmas

In this section we present various technical lemmas which allow us to prove our main results. For $\Delta \in \mathbb{R}^d_{>0}$, we recall the definition of the hyperbolic entropy function (Ghai et al., 2020) $\psi_\Delta : \mathbb{R}^d \to \mathbb{R}$ at scale Δ:

$$\psi_\Delta(\theta) = \frac{1}{4}\sum_{i=1}^d\left(2\theta_i \operatorname{arcsinh}\left(\frac{2\theta_i}{\Delta_i}\right) - \sqrt{4\theta_i^2 + \Delta_i^2} + \Delta_i\right).$$

The following lemma shows that the potential behaves as the ℓ_1-norm as Δ approaches 0.

Lemma 4. *For $\theta \in \mathbb{R}^d$ the following asymptotic equivalence holds:*

$$\psi_\Delta(\theta) \underset{\Delta \to 0}{\sim} \frac{1}{4} \sum_{i=1}^{d} \ln\left(\frac{1}{\Delta_i}\right) |\theta_i|.$$

Proof. The lemma easily follows from the asymptotic convergence

$$\operatorname{arcsinh}(x) \underset{|x| \to \infty}{\sim} \operatorname{sgn}(x) \ln |x|.$$

□

The following lemma is a classical result which gives a closed-form expression to the solution of a first order ODE.

Lemma 5. *Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}^d$ be a differentiable function and let $g : \mathbb{R}_{\geq 0} \to \mathbb{R}^d$ be a continuous function such that for some $\lambda \neq 0$,*

$$\lambda \dot{f} + f + g = 0, \ \ \forall t \in \mathbb{R}_{\geq 0}.$$

Then,

$$f(t) = f(0) e^{-\frac{t}{\lambda}} - \frac{1}{\lambda} \int_0^t g(s) e^{-\frac{(t-s)}{\lambda}} ds.$$

Moreover, we have the following formula for the integral of $f(t)$:

$$\int_0^T f(t) dt = \lambda f(0)(1 - e^{-\frac{T}{\lambda}}) - \int_0^T g(t)(1 - e^{-\frac{(T-t)}{\lambda}}) dt.$$

Proof. If we integrate the identity $\frac{d}{dt}\left[f(t)e^{t/\lambda}\right] = -\frac{1}{\lambda} g(t) e^{t/\lambda}$, we get that

$$f(t) = f(0) e^{-\frac{t}{\lambda}} - \frac{1}{\lambda} \int_0^t g(s) e^{-\frac{(t-s)}{\lambda}} ds.$$

As for the second part of the lemma, notice that

$$\int_0^T f(t) dt = \int_0^T \left[f(0) e^{-\frac{t}{\lambda}} - \frac{1}{\lambda} \int_0^t g(s) e^{-\frac{(t-s)}{\lambda}} ds \right] dt.$$

Hence, using Fubini, we get

$$\begin{aligned} \int_0^T \int_0^t g(s) e^{-\frac{(t-s)}{\lambda}} ds dt &= \int_0^T \int_0^T g(s) \mathbf{1}_{s \leq t}(s,t) e^{-\frac{(t-s)}{\lambda}} ds dt \\ &= \int_0^T g(s) \int_0^T \mathbf{1}_{s \leq t}(s,t) e^{-\frac{(t-s)}{\lambda}} dt ds \\ &= \int_0^T g(s) \int_s^T e^{-\frac{(t-s)}{\lambda}} dt ds \\ &= \int_0^T g(s) \lambda (1 - e^{-\frac{(T-s)}{\lambda}}) ds, \end{aligned}$$

which concludes the proof of the lemma. □

The following lemma gives various properties on integrability and convergence of the solution f of the aforementioned ODE.

Lemma 6. *Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}^d$ be a differentiable function such that $f(0) = 0$ and let $g : \mathbb{R}_{\geq 0} \to \mathbb{R}^d$ be a continuous function such that for some $\lambda \neq 0$,*

$$\lambda \dot{f} + f + g = 0, \ \ \forall t \in \mathbb{R}_{\geq 0}.$$

If $g \in L^\infty(0, +\infty)$, then $f \in L^\infty(0, +\infty)$ and $\|f\|_\infty \leq \|g\|_\infty$. Moreover, if $g \in L^1(0, +\infty)$, then the following hold:

- $f \in L^1(0, +\infty)$ *and* $\int_0^t |f(s)|ds \leq \int_0^t |g(s)|ds, \;\; \forall t \in [0, +\infty]$;
- $\lim\limits_{t\to\infty} f(t) = 0$;
- $\displaystyle\int_0^\infty f = -\int_0^\infty g.$

Proof. First, assume $g \in L^\infty(0, \infty)$. From Lemma 5, we have that $f(t) = -\frac{1}{\lambda}\int_0^t g(s)e^{-\frac{(t-s)}{\lambda}}ds$. Hence,

$$\begin{aligned}|f(t)| &\leq \frac{\|g\|_\infty}{\lambda}\int_0^t e^{-\frac{(t-s)}{\lambda}}ds \\ &= \|g\|_\infty(1 - e^{-t/\lambda}) \leq \|g\|_\infty,\end{aligned}$$

which proves the first assertion.

Second, assume $g \in L^1(0, \infty)$. Then, $|f(t)| \leq \frac{1}{\lambda}\int_0^t |g(s)|e^{-\frac{(t-s)}{\lambda}}ds$. Therefore,

$$\begin{aligned}\int_0^t |f(s)|ds &\leq \int_0^t |g(s)|(1 - e^{-\frac{(t-s)}{\lambda}})ds \\ &\leq \int_0^t |g(s)|ds \leq \|g\|_{L^1}.\end{aligned}$$

Moving on, we will show that $\lim_{t\to\infty} f(t) = 0$. Recall that $f(t) = -\frac{1}{\lambda}\int_0^t g(s)e^{-\frac{(t-s)}{\lambda}}ds$. Then,

$$\begin{aligned}\left|\int_0^t g(s)e^{-\frac{(t-s)}{\lambda}}ds\right| &= \left|\int_0^{t/2} g(s)e^{-\frac{(t-s)}{\lambda}}ds + \int_{t/2}^t g(s)e^{-\frac{(t-s)}{\lambda}}ds\right| \\ &\leq e^{-\frac{t}{2\lambda}}\int_0^{t/2}|g(s)|ds + \int_{t/2}^\infty |g(s)|ds \\ &\xrightarrow{t\to\infty} 0.\end{aligned}$$

Finally, notice that

$$\begin{aligned}&\lim_{t\to\infty}\left[\lambda\int_0^t \dot{f} + \int_0^t (f+g)\right] = 0 \iff \\ &\lambda\lim_{t\to\infty} f(t) + \int_0^\infty (f+g) = 0 \iff \\ &\int_0^\infty f + \int_0^\infty g = 0,\end{aligned}$$

where we used that $\lim_{t\to\infty} f(t) = 0$ and the linearity of the Lebesgue integral. □

With the help of Lemma 5 and Lemma 6, we can finally prove Proposition 3, which considers ODE (4) and establishes the positivity of the balancedness for small λ.

Proposition 3. *For* $\lambda \leq \frac{n}{\|y\|_2^2}\cdot(\min_{i\leq d}\Delta_{0,i})$, *the balancedness* Δ_t *never vanishes:* $\Delta_t \neq 0, \;\; \forall t \in [0, +\infty]$.

Proof. We consider MGF(λ) with the diagonal-linear-network loss $F(w) = \mathcal{L}(u \odot v)$, where $w = (u, v)$. From the energy of the system, defined in Equation (14) as $E_t = F(w_t) + \frac{\lambda}{2}\|\dot{w}_t\|_2^2$ with derivative $\dot{E}_t = -\|\dot{w}_t\|_2^2$, we get that

$$\mathcal{L}(\theta_t) = \frac{\|y\|_2^2}{2n} - \frac{\lambda}{2}\|\dot{w}_t\|_2^2 - \int_0^t \|\dot{w}_s\|_2^2 ds.$$

Hence, since the LHS of the above equation is nonnegative, we get

$$\int_0^\infty \|\dot{w}_t\|^2 dt \leq \frac{\|y\|^2}{2n}.$$

Therefore,

$$\int_0^\infty |\dot{u}_t^2 - \dot{v}_t^2| dt < \frac{\|y\|^2}{2n}\mathbf{1}.$$

Consequently, $\dot{u}_t^2 - \dot{v}_t^2 \in L^1(0, \infty)$. Now, notice that from ODE (4), we obtain

$$\lambda(\ddot{u}_t u_t - \ddot{v}_t v_t) + (\dot{u}_t u_t - \dot{v}_t v_t) = 0 \iff$$
$$\lambda \frac{d}{dt}(\dot{u}_t u_t - \dot{v}_t v_t) + (\dot{u}_t u_t - \dot{v}_t v_t) - \lambda(\dot{u}_t^2 - \dot{v}_t^2) = 0.$$

Applying Lemma 5 yields

$$\dot{u}_t u_t - \dot{v}_t v_t = \int_0^t (\dot{u}_s^2 - \dot{v}_s^2) e^{-\frac{(t-s)}{\lambda}} ds$$

and

$$u_t^2 - v_t^2 = \Delta_0 + 2\lambda \int_0^t (\dot{u}_s^2 - \dot{v}_s^2)(1 - e^{-\frac{(t-s)}{\lambda}}) ds. \tag{30}$$

Applying Lemma 6 allows us to conclude that for every $t \in [0, +\infty]$,

$$\begin{aligned}\Delta_t &\geq \Delta_0 - 2\lambda \int_0^t |\dot{u}_s^2 - \dot{v}_s^2| ds \\ &> \Delta_0 - \frac{\lambda \|y\|_2^2}{n}\mathbf{1} \geq 0,\end{aligned}$$

where the last inequality is due to the inequality assumption over λ. □

Our final technical lemma helps with the proof of Theorem 2. The definition of the quantities $Q_{\pm,m}$ can be found in the proof of this theorem.

Lemma 7. *The quantities $\alpha_{\pm,N}$ defined in eq. (27):*

$$\alpha_{\pm,N+1} = \alpha \exp\left(-\frac{1}{1-\beta}\sum_{m=1}^{N}(1-\beta^{N+1-m})Q_{\pm,m}\right),$$

converge as $N \to \infty$ to vectors $\alpha_{\pm,\infty}$ with non-zero coordinates.

Proof. From Assumption 3 and Assumption 4, we have that the iterates $w_{\pm,N}$ converge towards vectors $w_{\pm,\infty}$ such that $\Delta_\infty = |w_{+,\infty} \odot w_{-,\infty}|$ has non-zero coordinates. This means that there exists $N_0 > 0$ such that $w_{\pm,N}$ do not change sign for $N \geq N_0$. Consequently, the imaginary parts of $g_{\pm,N}$ are constant (equal to 0 or π depending on the sign of $w_{\pm,\infty}$) for $N \geq N_0$, and $\delta_{\pm,N} \in \mathbb{R}$ for $N \geq N_0$. This finally means that $Q_{\pm,N} \geq 0$ for $N \geq N_0$ and

$$\sum_{m=1}^{N}(1-\beta^{N+1-m})Q_{\pm,m} = \sum_{m=1}^{N_0}(1-\beta^{N+1-m})Q_{\pm,m} + \sum_{m=N_0+1}^{N}(1-\beta^{N+1-m})Q_{\pm,m}$$

The first term converges to $\sum_{m=1}^{N_0} Q_{\pm,m}$ as $N \to \infty$. The second term is increasing because $Q_{\pm,N}$ are positive for $N \geq N_0$ and $(1-\beta^{N+1-m})$ is increasing. Therefore, the second term also converges to a finite value since otherwise $\alpha_{\pm,\infty} = 0$, which contradicts $\Delta_\infty = |\alpha_{+,\infty}\alpha_{-,\infty}| \neq 0$. □

F Additional Experiments

In this section of the appendix, we clarify experimental details and discuss additional experiments.

F.1 MGF: A Good Continuous Surrogate

Most of our experiments deal with 2-layer diagonal networks, but before we constrain ourselves to that tractable setting, we present a couple of experiments on more general architectures. These experiments highlight our observation from Section 2 that $\mathrm{MGF}(\lambda)$ serves as a good continuous proxy for $\mathrm{MGD}(\gamma,\beta)$ even for complicated non-convex losses F and large step sizes γ. We provide evidence for that conclusion by showing that the single parameter $\lambda = \gamma/(1-\beta)^2$ controls the generalisation performance of models trained with $\mathrm{MGD}(\gamma,\beta)$.

Teacher-Student Fully Connected Network. We detail the experimental setting which leads to Figure 2. We consider a teacher-student setup where the teacher is a one-hidden-layer fully-connected ReLU network with 5 hidden neurons and the student is a one-hidden-layer fully-connected ReLU network with 20 hidden neurons. We randomly generate 15 inputs $x_i \in \mathbb{R}^2$ according to a standard multivariate normal distribution. Each y_i corresponds to the output by the teacher network on input x_i. The student is trained using momentum gradient descent with a square loss. Figure 2 corresponds to the test loss after the student reaches 10^{-5} training error. Each grid point corresponds to the same data set and initialisation of the student network. We observe that the quantity $\lambda = \frac{\gamma}{(1-\beta)^2}$ aligns well with the level lines of the test loss as expected from Proposition 1.

Deep Linear Network. The network used for Figure 6 contains 5 layers with widths (30, 60, 120, 60, 1) and was trained for 1000 epochs for each pair of momentum parameter β and step size γ. Each network weight was randomly initialised according to $\mathcal{N}(0, 0.1^2)$ with fixed randomness for each (γ,β)-trial. The training data was chosen as follows: $(x_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu\mathbf{1}, \sigma^2 I_d)$ and $y_i = \langle x_i, \theta_s^\star\rangle$ for $i \in [n]$ where $\theta_s^\star$ is s-sparse with nonzero entries equal to $1/\sqrt{s}$, where $(n, d, s) = (20, 30, 5)$ and $(\mu, \sigma) = (1, 1)$. We show results averaged over 5 replications.

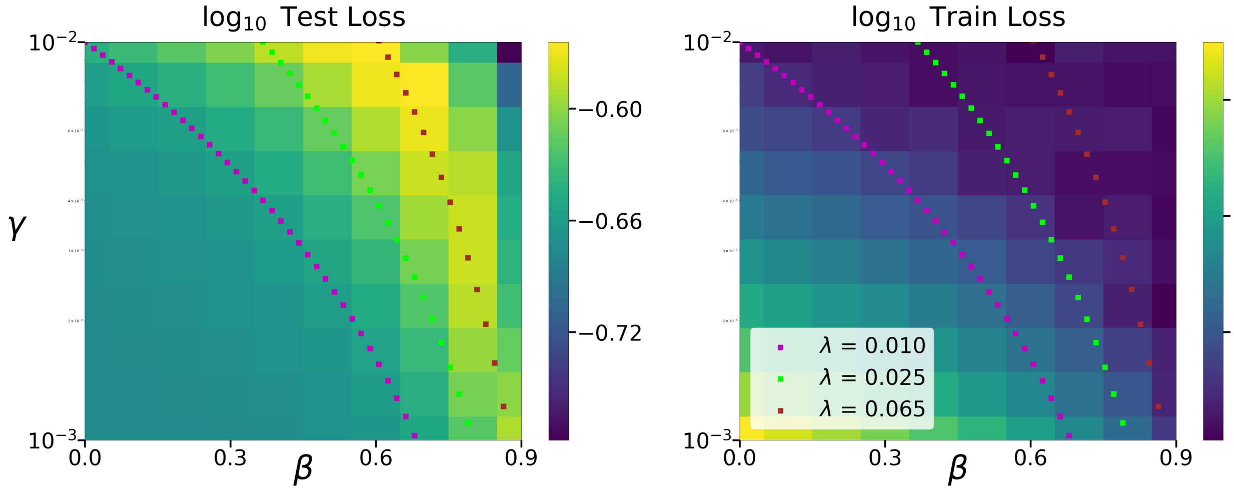

Figure 6: Test and train loss of a fully connected deep linear network trained with $\mathrm{MGD}(\gamma,\beta)$ in a noiseless sparse overparametrised regression setting. The test loss appears considerably correlated with the intrinsic parameter $\lambda = \gamma/(1-\beta)^2$, evincing that $\mathrm{MGF}(\lambda)$ approximates $\mathrm{MGD}(\gamma,\beta)$ sufficiently well even on complex architectures.

2-Layer Diagonal Linear Network. The plots from Figure 7 were obtained for a 2-layer diagonal linear network trained in the noiseless sparse overparametrised regression setting described above. The first network layer was initialised with the uniform initialisation $\alpha\mathbf{1}$, where $\alpha = 0.01$, and the weights of the second layer were set to 0. The momentum gradient flow evolution of the weights was simulated with the default version of the ODE solver `scipy.integrate.odeint`.

F.2 Experiments with Diagonal Linear Networks

Having seen empirical proof that $\mathrm{MGF}(\lambda)$ approximates well the optimisation trajectory of $\mathrm{MGD}(\gamma,\beta)$ on complicated models, we proceed with experiments that illustrate the conclusions of our results for 2-layer diagonal linear networks. In particular, we provide experimental evidence that both in the continuous and discrete-time cases, the recovered interpolators by MGD and MGF satisfy

$$\theta^{\texttt{MGF/MGD}} = \operatorname*{argmin}_{\theta^\star \in \mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0) \approx \operatorname*{argmin}_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star),$$

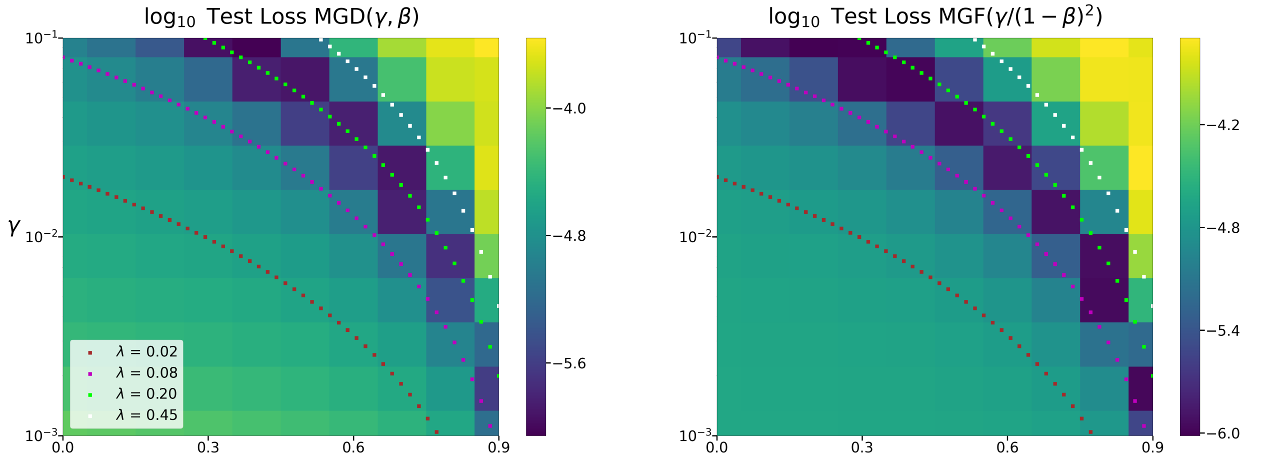

Figure 7: *Left*: Decimal logarithm of the test loss of a 2-layer diagonal linear network trained with MGD(γ, β) for 1 million epochs. *Right*: Decimal logarithm of the test loss of a 2-layer diagonal linear whose weights evolved according to MGF(λ) – where $\lambda = \gamma/(1-\beta)^2$ – and converged to an interpolator of the training dataset. We observe an almost one-to-one correspondence in terms of generalisation capacity, which demonstrates that MGF(λ) serves as a suitable continuous surrogate for MGD(γ, β) in the diagonal linear setting.

as we explain underneath Theorem 1, Theorem 2, and in Appendix C.3.3. Indeed, we observe that the perturbation term $\tilde{\theta}_0$ can be safely ignored even without the assumption of strictly positive balancedness. The asymptotic balancedness Δ_∞ then uniquely controls the properties of the recovered solution. We now specify our experimental setting.

Experimental Details. We work in the noiseless sparse overparametrised regression setting with uncentered data. More precisely, we let $(x_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu \mathbf{1}, \sigma^2 I_d)$ and $y_i = \langle x_i, \theta_s^\star \rangle$ for $i \in [n]$ where $\theta_s^\star$ is s-sparse with nonzero entries equal to $1/\sqrt{s}$. We train a 2-layer diagonal linear network with (M)GD and (M)GF with the uniform initialisation $u_0 = \alpha \mathbf{1}$, where $\alpha = 0.01$ and $v_0 = 0$. In order to simulate gradient flow or momentum gradient flow on the network weights, we use the vanilla version of the ODE solver `scipy.integrate.odeint`. For most of the incoming plots, we have fixed $(n, d, s, \sigma) = (20, 30, 5, 1)$ and we let $\mu \in \{0, 0.5, 1, 1.5\}$. In what follows, all plots show results averaged over 5 replications.

F.2.1 Continuous-Time Plots

We first present a set of 3 continuous-time plots (Figure 8) for the setting where the input data follows a Gaussian distribution $\mathcal{N}(\mu \mathbf{1}, I_d)$ with $\mu = 1$.

Experimental Setup. For a sampled dataset (X, y), we train our diagonal network with MGF(λ), $\lambda \in [0, 1]$, and initialisation $(u_0, v_0) = (\alpha \cdot \mathbf{1}, 0)$ until convergence to an interpolator [7] $\theta^{\texttt{MGF}}$. During the training of MGF(λ), we also take note of whether the balancedness Δ_t remains strictly positive at all times, thereby checking the explanatory range of Section 3.3. Having completed the MGF training, we plot the **Test Loss** of $\theta^{\texttt{MGF}}$, the **ℓ_2-Norm of** Δ_∞, and the **ℓ_1-Norm of** $\theta^{\texttt{MGF}}$ in order to visualise the gain in generalisation performance.

Insignificance of $\tilde{\theta}_0$. Now, recall from Theorem 1 that $\theta^{\texttt{MGF}} = \operatorname{argmin}_{\theta^\star \in \mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0)$ and that for $\|\tilde{\theta}_0\|_\infty \ll \|\theta^{\texttt{MGF}}\|_\infty$, $D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0) \approx \psi_{\Delta_\infty}(\theta^\star)$. We proved that for small values of λ, the balancedness remains strictly positive at all times, which allowed us to show that $\|\tilde{\theta}_0\|_\infty < \alpha^2$. We conjecture that $\theta^{\texttt{MGF}} \approx \operatorname{argmin}_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$ continues to hold for larger values of λ. We experimentally test this claim by measuring the precise distance between $\theta^{\texttt{MGF}}$ and $\theta^{\texttt{GF}}_{\Delta_\infty} = \operatorname{argmin}_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$. Indeed, we initialise a gradient flow with initial balancedness equal to Δ_∞ and such that $\theta_0 = 0$, which converges to the predictor $\theta^{\texttt{GF}}_{\Delta_\infty}$ as discussed in Section 3.1. Hence, we can calculate the **Normalised Distance between** $\theta^{\texttt{MGF}}$ **and** $\theta^{\texttt{GF}}_{\Delta_\infty}$ equal to $\|\theta^{\texttt{MGF}} - \theta^{\texttt{GF}}_{\Delta_\infty}\|_2 / \|\theta^{\texttt{GF}}_{\Delta_\infty}\|_2$, and we

[7] We know that $\theta^{\texttt{MGF}}$ interpolates the dataset (X, y) because we also record the **Train Loss** $(\theta^{\texttt{MGF}})$, which falls under 10^{-20}.

obtain that $\|\theta^{\texttt{MGF}} - \theta^{\texttt{GF}}_{\Delta_\infty}\|_2 / \|\theta^{\texttt{GF}}_{\Delta_\infty}\|_2 < 0.01$ for $\lambda \in (0,1)$.

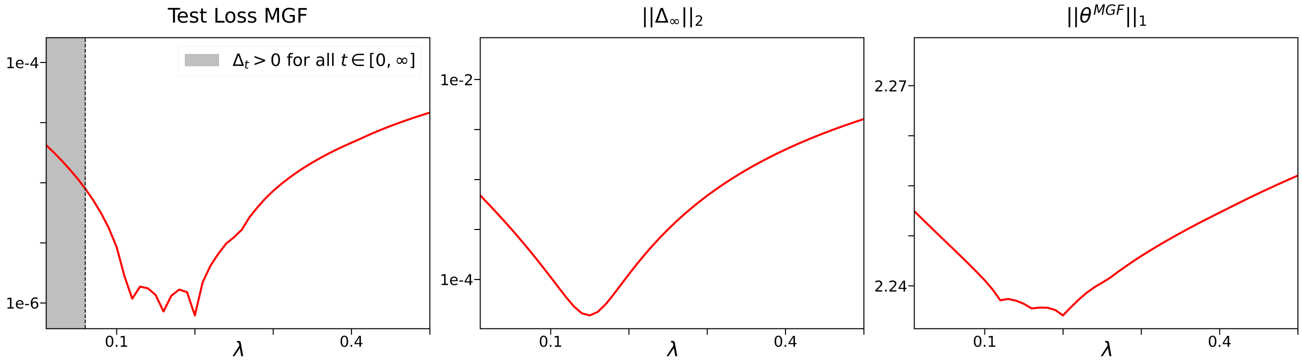

Figure 8: Continuous-time experiments on uncentered data with mean $\mu = 1$. Here, $\theta^{\texttt{MGF}}$ denotes the interpolator recovered by MGF(λ) and Δ_∞ stands for the balancedness at infinity for MGF(λ). We observe that the test loss and sparsity of $\theta^{\texttt{MGF}}$ correlate with the magnitude of Δ_∞ as predicted by Theorem 1.

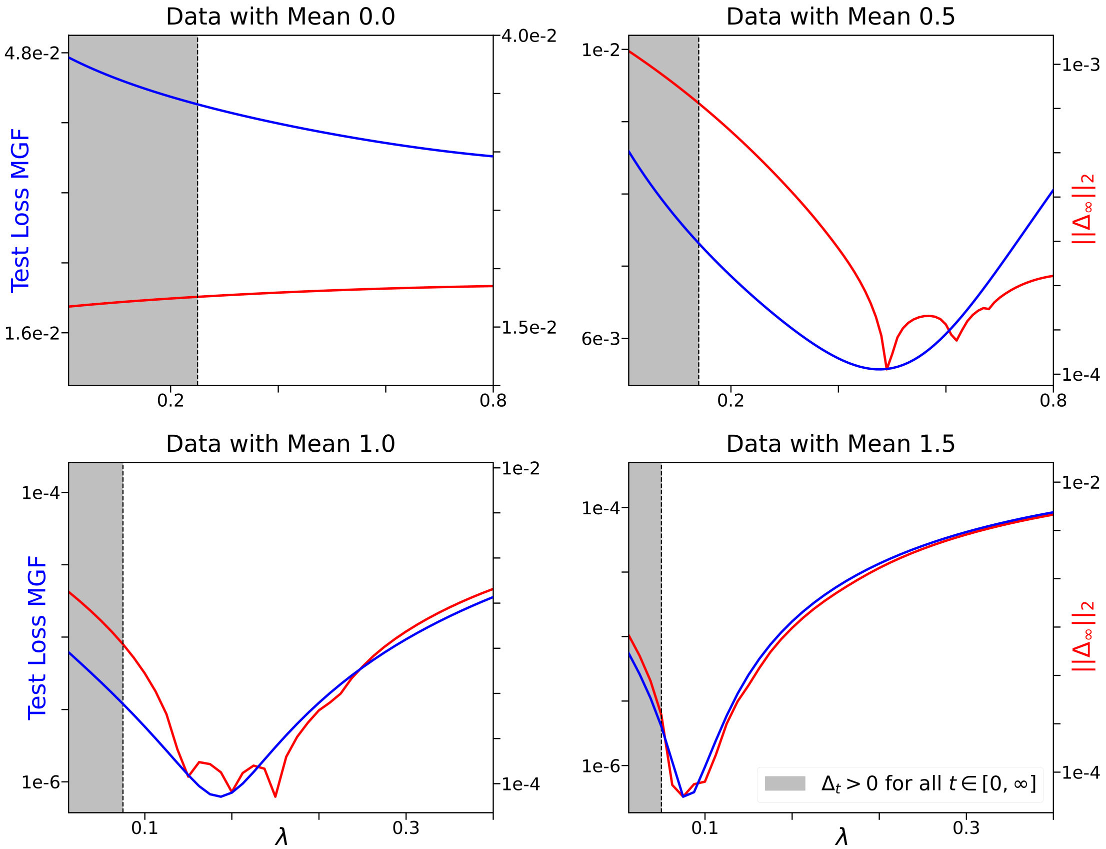

Figure 9: We observe that for uncentered data the magnitude of the balancedness at infinity Δ_∞ correlates with the test loss of the interpolator selected by MGF(λ). However, this relationship breaks for centered data.

Insights from Continuous-Time Experiments. First, we observe that no matter the mean of the data distribution[8] or the size of $\lambda \in (0,1)$, the normalised distance between $\theta^{\texttt{MGF}}$ and $\theta^{\texttt{GF}}_{\Delta_\infty}$ is always upper-bounded by 0.01. Hence, we can empirically confirm our conjecture from Theorem 1 that $\theta^{\texttt{MGF}} \approx \theta^{\texttt{GF}}_{\Delta_\infty}$ for larger λ when the balancedness changes sign. Second, we see that regardless of the mean of the dataset, the balancedness at infinity (i.e., the effective initialisation Δ_∞) controls the generalisation behavior of the recovered interpolator. We can explain this observation again through the approximate equivalence $\theta^{\texttt{MGF}} \approx \operatorname{argmin}_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$.

[8] We performed the continuous-time experiments depicted in Figure 9 for data with mean $\mu = 0, 0.5, 1, 1.5$.

The Effect of the Data Mean. In Figure 9, we summarise our empirical results for data with various means. Notice that there exists a difference between the generalisation behavior for centered and uncentered data. Indeed, for centered data (top left), the key quantity λ has little impact on the sparsity of the recovered solution. This circumstance is reminiscent of the observations from (Nacson et al., 2022) and (Even et al., 2023). However, for uncentered data, we observe an interval $\mathcal{I}_{\mathcal{D}_x} = (0, \lambda_{\max})$ (which depends on the data distribution $\mathcal{D}_x$) for which MGF with $\lambda \in \mathcal{I}_{\mathcal{D}_x}$ outperforms GF in terms of generalisation. Furthermore, there appears to exist a constant $\lambda^\star_{\mathcal{D}_x} \in \mathcal{I}_{\mathcal{D}_x}$ (roughly corresponding to the minimum magnitude of Δ_∞) which brings about the most improvement compared to gradient flow. We note that the following tendency seems to hold empirically:

$$\lim_{|\mu| \to +\infty} \lambda^\star_{\mathcal{D}_x} = 0.$$

F.2.2 Discrete-Time Plots

For the sake of brevity[9], we only present a single set of plots for the discrete-time noiseless sparse recovery given in Figure 4. Our input data follows a unit-mean Gaussian distribution $\mathcal{N}(\mathbf{1}, I_d)$.

Experimental Setup. For a sampled dataset (X, y) and hyperparameter pair (β, γ), we train our 2-layer diagonal linear network with MGD(γ, β) initialised at $(u_0, v_0) = (\alpha\mathbf{1}, 0)$ for 1 million epochs (which suffices for convergence[10]). During the MGD(γ, β) training, we also take note of whether the iterates $w_{\pm,k}$ change sign or not thereby checking the explanatory range of Corollary 3. Having completed the MGD training, we plot the **Test Loss** of $\theta^{\texttt{MGD}}$, the ℓ_2**-Norm of** Δ_∞, and the ℓ_1**-Norm of** $\theta^{\texttt{MGD}}$ in order to visualise the gain in generalisation performance.

Insignificance of $\tilde{\theta}_0$. Recall from Theorem 2 that $\theta^{\texttt{MGD}} = \operatorname{argmin}_{\theta^\star \in \mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0)$. Again, we want to characterise the recovered interpolator as $\theta^{\texttt{MGD}} \approx \operatorname{argmin}_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$. In order to verify empirically that the effect of the perturbation term is negligible, we follow the same strategy as in the continuous-time case. We initialise a gradient flow with initial balancedness equal to Δ_∞ and $\theta_0 = 0$, which converges to the predictor $\theta^{\texttt{GF}}_{\Delta_\infty}$ as discussed in Section 3.1. Hence, we can calculate the **Normalised Distance between** $\theta^{\texttt{MGD}}$ **and** $\theta^{\texttt{GF}}_{\Delta_\infty}$ equal to $\|\theta^{\texttt{MGD}}_{\gamma,\beta,\alpha} - \theta^{\texttt{GF}}_{\Delta_\infty}\|_2 / \|\theta^{\texttt{GF}}_{\Delta_\infty}\|_2$, and we find that $\|\theta^{\texttt{MGD}}_{\gamma,\beta,\alpha} - \theta^{\texttt{GF}}_{\Delta_\infty}\|_2 / \|\theta^{\texttt{GF}}_{\Delta_\infty}\|_2 < 0.01$ for all pairs (γ, β) in Figure 4. This experimentally shows that $\theta^{\texttt{MGD}} \approx \operatorname{argmin}_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$ and that the asymptotic balancedess is the key quantity which predicts the recovered solution.

Insights from Discrete-Time Experiments. As predicted by Theorem 2, a more balanced solution (center plot) leads to a solution with a lower ℓ_1-norm (right plot), which in turn translates to better generalisation (left plot). Finally, as proven in Corollary 3, the trajectories for which the iterates do not cross zero satisfy $\Delta_\infty < \Delta_0$, where Δ_0 (approximately) corresponds to the asymptotic balancedness for the pair $(\beta, \gamma) = (0, 10^{-3})$ in the bottom left corner of the center plot. Clearly, the pairs (β, γ) for which $w_{\pm,k}$ do not change sign lead to better generalisation than the pair $(0, 10^{-3})$. Again, we note that for centered data the story changes, and we lose the clear correspondence between small $\|\Delta_\infty\|_2$ and small $\|\theta^{\texttt{MGD}}\|_1$.

[9] We performed discrete-time experiments for data with means $\mu = 0, 0.5, 1, 1.5$.

[10] Again, we record the **Train Loss** $(\theta^{\texttt{MGD}}_{\gamma,\beta\alpha})$, which falls under 10^{-8}.