

Leveraging Continuous Time to Understand Momentum When Training Diagonal Linear Networks

Hristo Papazov* Scott Pesme* Nicolas Flammarion (TML @ EPFL)

Discrete and Continuous Momentum

Setup: $F : \mathbb{R}^D \rightarrow \mathbb{R}_{\geq 0}$ – differentiable loss function
 • $\gamma > 0$ – step size • $\beta \in [0, 1]$ – momentum parameter
 • $\lambda > 0$ – key quantity • $\varepsilon > 0$ – discretization step

MGD(γ, β): $w_{k+1} = w_k - \gamma \nabla F(w_k) + \beta(w_k - w_{k-1})$

MGF(λ): $\lambda \ddot{w}_t + \dot{w}_t + \nabla F(w_t) = 0$ ∇F locally Lipschitz Unique solution on $[0, +\infty)$

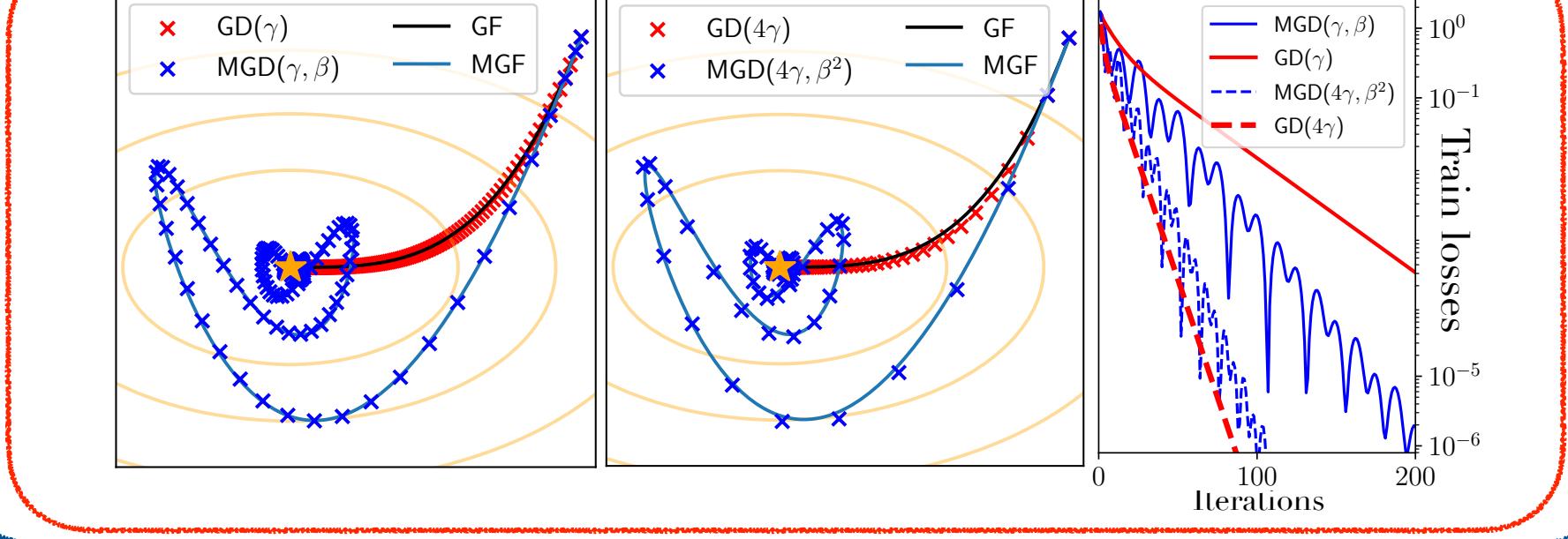
Discretization Scheme:

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

Proposition. For $(w_0, w_1) \in \mathbb{R}^{2D}$, consider MGF(λ) with $\lambda = \frac{\gamma}{(1-\beta)^2}$ initialized at $w_{t=0} = w_0$, $\dot{w}_{t=0} = (w_1 - w_0)/\varepsilon$ where $\varepsilon = \gamma/(1-\beta)$. Then, the above discretization scheme with d.s. ε leads to MGD(γ, β) initialized at (w_0, w_1) .

Error Bounds: Unnecessarily pessimistic $\max_{k \in [K]} |w_k - w(k\varepsilon)| \leq \exp(CK) \cdot \varepsilon$

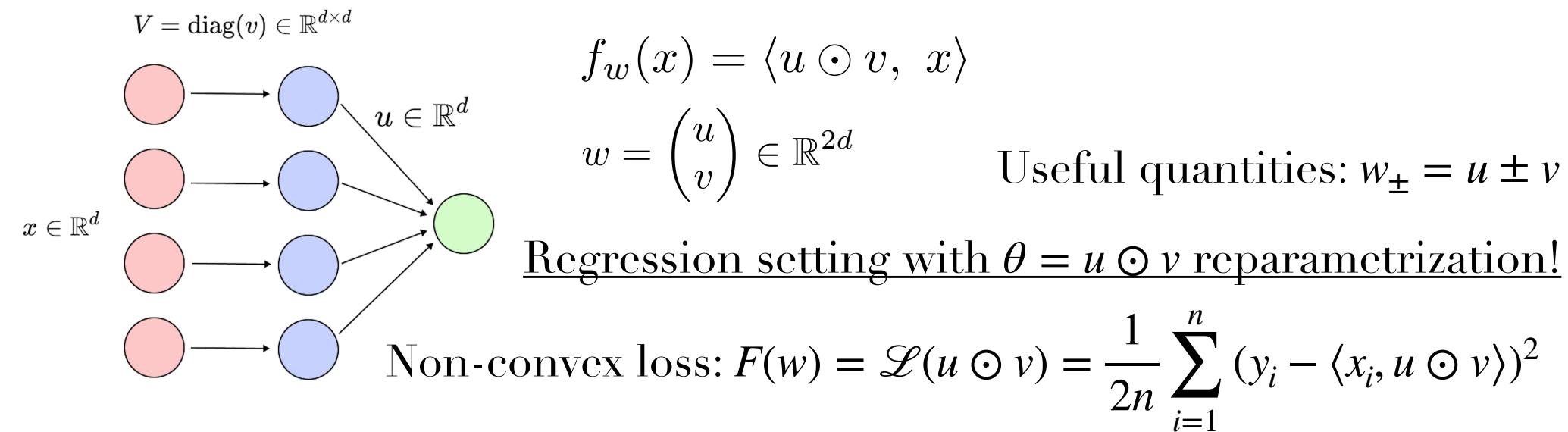
(M)GD vs. (M)GF over a 2D Quadratic



Momentum on Diagonal Linear Networks

Data: Sparse and underdetermined sample $(x_i, y_i)_{i=1}^n$
 $x_i \in \mathbb{R}^d$, $y_i = \langle \theta_s^*, x_i \rangle$, $n < d$, θ_s^* – s -sparse
 $\mathcal{S} = \theta_s^* + \text{span}(x_1, \dots, x_n)^\perp$ – set of interpolators

Architecture: 2-layer diagonal linear neural network



MGF and Stochastic MGD:

(Neurons)	$\lambda \ddot{u}_t + \dot{u}_t + \nabla \mathcal{L}(\theta_t) \odot v_t = 0$	(C1) $\Delta_0 \neq 0 \iff 2d$ degrees of freedom
	$\lambda \ddot{v}_t + \dot{v}_t + \nabla \mathcal{L}(\theta_t) \odot u_t = 0$	(C2) Zero initial speed: $\dot{u}_0 = \dot{v}_0 = 0$
(Predictors)	$u_{k+1} = u_k - \gamma \nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k) \odot v_k + \beta(u_k - u_{k-1})$	(Initialization scale) $\alpha := \max(\ u_0\ , \ v_0\)$
	$v_{k+1} = v_k - \gamma \nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k) \odot u_k + \beta(v_k - v_{k-1})$	Assumptions:
(Balancedness)	$\theta_t = u_t \odot v_t$ and $\theta_k = u_k \odot v_k$	1. (Boundedness) The optimization trajectory of MGF and SMGD is bounded. 2. (Balancedness) The asymptotic balancedness Δ_∞ of MGF and SMGD has nonzero coordinates.

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

Initialization:

(C1) $\Delta_0 \neq 0 \iff 2d$ degrees of freedom

(C2) Zero initial speed: $\dot{u}_0 = \dot{v}_0 = 0$

(Initialization scale) $\alpha := \max(\|u_0\|, \|v_0\|)$

Assumptions:

1. (Boundedness) The optimization trajectory of MGF and SMGD is bounded.

2. (Balancedness) The asymptotic balancedness Δ_∞ of MGF and SMGD has nonzero coordinates.

Recall†

Mirror Map: $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ – C^2 -smooth, strictly convex, coercive gradient

Bregman Divergence: $D_\Phi(\theta_1, \theta_2) = \Phi(\theta_1) - \Phi(\theta_2) - \langle \nabla \Phi(\theta_2), \theta_1 - \theta_2 \rangle > 0$, $\forall \theta_1 \neq \theta_2$

Hyperbolic Entropy: For $\Delta \in \mathbb{R}_{>0}^d$, $\psi_\Delta(\theta) = \frac{1}{4} \sum_{i=1}^d \left(2\theta_i \text{arcsinh} \left(\frac{2\theta_i}{\Delta_i} \right) - \sqrt{4\theta_i^2 + \Delta_i^2} + \Delta_i \right)$: $\mathbb{R}^d \rightarrow \mathbb{R}$.

Importantly, $\psi_\Delta \sim_{\Delta \rightarrow 0} \frac{\log(4/\Delta)}{2} \|\cdot\|_1$ and $\psi_\Delta \sim_{\Delta \rightarrow +\infty} \frac{1}{2\Delta} \|\cdot\|_2$.

Implicit Bias of Gradient Flow ($\lambda = 0$): $\theta^{\text{GF}} = \arg\min_{\theta \in \mathcal{S}} D_{\psi_{\Delta_0}}(\theta^*, \theta)$

Small Initialization: $\Delta_0, \theta_0 = O(\alpha^2) \ll \theta^*$, so $D_{\psi_{\Delta_0}}(\theta^*, \theta_0) \sim_{\alpha \rightarrow 0} \psi_{\Delta_0}(\theta^*) \propto_{\alpha \rightarrow 0} \|\theta^*\|_1$

Recovery of sparse interpolators!

Time-Varying Mirror Flow

Proof Strategy: Show that \exists time $T > 0$, after which the predictors θ_t follow a momentum mirror flow with time-varying potentials Φ_t

$$\lambda \frac{d^2 \nabla \Phi_t(\theta_t)}{dt^2} + \frac{d \nabla \Phi_t(\theta_t)}{dt} + \nabla \mathcal{L}(\theta_t) = 0, \text{ so } \nabla \Phi_\infty(\theta_\infty) \in \text{span}(x_1, \dots, x_n).$$

Find $\tilde{\theta}_0$ such that $\nabla \Phi_\infty(\tilde{\theta}_0) \in \text{span}(x_1, \dots, x_n)$.

Use the Bregman Cosine Theorem + $\theta^* - \theta^{\text{MGF}} \in \text{span}(x_1, \dots, x_n)^\perp$ to conclude:

$$D_{\Phi_\infty}(\theta^*, \tilde{\theta}_0) = D_{\Phi_\infty}(\theta^*, \theta^{\text{MGF}}) + D_{\Phi_\infty}(\theta^{\text{MGF}}, \tilde{\theta}_0).$$

Intertwined Roles of γ and β

Acceleration Rule: $\gamma \rightarrow \rho^2 \cdot \gamma$ $\beta \rightarrow 1 - \rho(1 - \beta)$ ρ speed-up

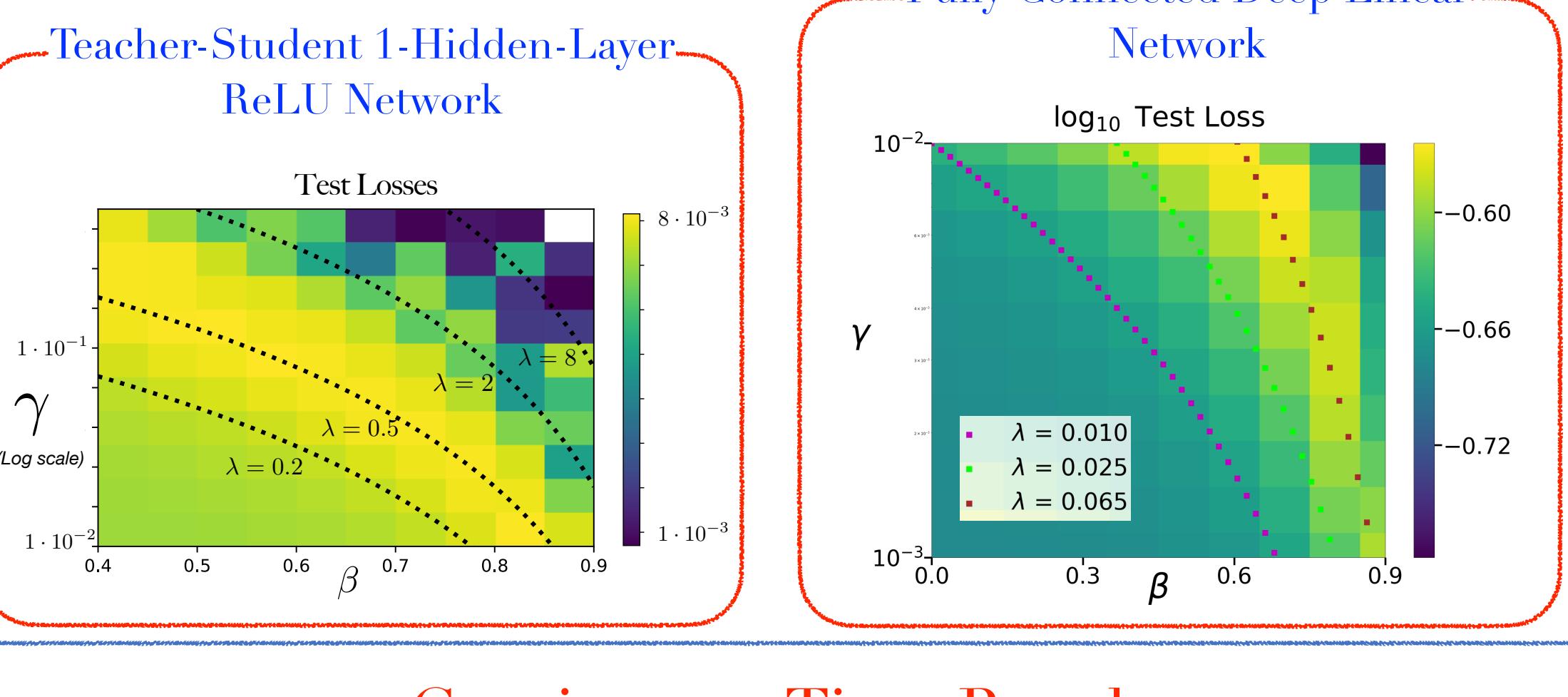
Corollary. Let MGD(γ, β) initialized at $w_0 = w_1 \in \mathbb{R}^D$ correspond to the discretization of MGF(λ) with d.s. ε . For $\rho > 0$, consider the parameter couple $\hat{\gamma} = \rho^2 \gamma$ and $\hat{\beta} = 1 - \rho(1 - \beta) \approx_{\rho \rightarrow 1} \beta^\rho$. Then, MGD($\hat{\gamma}, \hat{\beta}$) initialized at (w_0, w_1) again becomes the discretization of MGF(λ) but with discretization step $\hat{\varepsilon} = \rho \cdot \varepsilon$.

Optimization Regimes: The link between λ , γ , and β

Large β : $\lambda \gg 1$: Heavy oscillations + arbitrary slow convergence

Small γ : $\lambda \ll 1$: Gradient flow trajectory

Non-Degenerate λ : λ The momentum regime



Continuous-Time Results

Implicit Bias of MGF(λ): $\theta^{\text{MGF}} = \arg\min_{\theta^* \in \mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^*, \tilde{\theta}_0)$

Asymptotic Balancedness: $\Delta_\infty = \Delta_0 \odot \exp(-I_+ + I_-)$

Perturbed Initialization: $\tilde{\theta}_0 = \frac{1}{4} (w_{+,0}^2 \odot \exp(-2I_+) - w_{-,0}^2 \odot \exp(-2I_-))$

Experimentally: $\tilde{\theta}_0$ is negligible and $\theta^{\text{MGF}} \approx \arg\min_{\theta^* \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^*)$.

Gradient flow initialized at $(u_0 = \sqrt{\Delta_\infty}, v_0 = 0)$ and $(\dot{u}_0 = 0, \dot{v}_0 = 0)$ converges to $\theta^{\text{MGF}} = \arg\min_{\theta^* \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^*)$.

Theoretically: If $\forall t \in [0, +\infty]$, $\Delta_t > 0$, then $I_\pm = -\lambda \int_0^\infty \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}} \right)^2 dt > 0$.

$\Delta_\infty, \tilde{\theta}_0 = O(\alpha^2)$. Hence, for small initializations, $\Delta_\infty, \tilde{\theta}_0 \ll \theta^*$, $\forall \theta^* \in \mathcal{S}$.

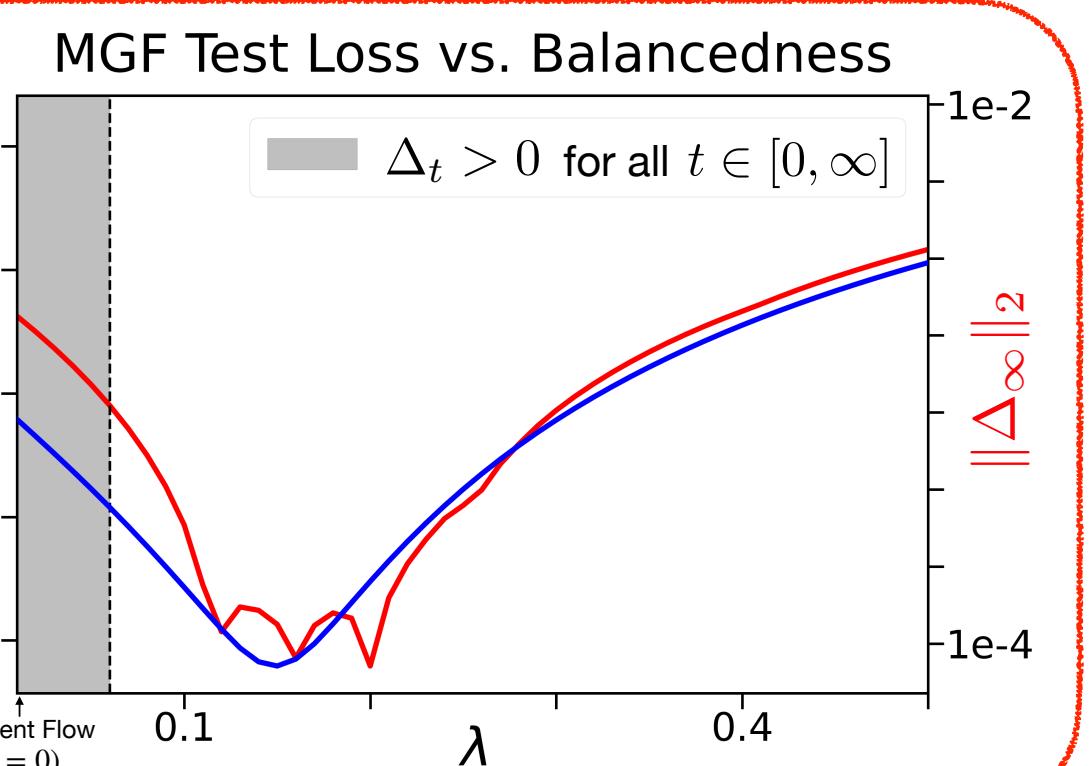
$D_{\psi_{\Delta_\infty}}(\theta^*, \theta_0) \sim_{\alpha \rightarrow 0} \psi_{\Delta_\infty}(\theta^*) \propto_{\alpha \rightarrow 0} \|\theta^*\|_1$ Sparse solutions!

Since $\Delta_\infty < \Delta_0$, we expect θ^{MGF} to be sparser than θ^{GF} !

Small λ Regime:

For $\lambda \leq \frac{n}{\|y\|_2^2} \cdot (\min_{i \in [d]} \Delta_{0,i})$, the balancedness never vanishes!

Also, $\Delta_\infty^2 \sim_{\lambda \rightarrow 0} \Delta_0^2 \exp\left(-2\lambda \int_0^\infty \nabla \mathcal{L}(\theta_t) dt\right)$.



Discrete-Time Results

Implicit Bias of SMGD(γ, β): $\theta^{\text{SMGD}} = \arg\min_{\theta^* \in \mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^*, \tilde{\theta}_0)$

Asymptotic Balancedness: $\Delta_\infty = \Delta_0 \odot \exp(-(S_+ + S_-))$

Perturbed Initialization: $\tilde{\theta}_0 = \frac{1}{4} (w_{+,0}^2 \odot \exp(-2S_+) - w_{-,0}^2 \odot \exp(-2S_-))$

$S_\pm = \frac{1}{1-\beta} \sum_{k=1}^\infty \left[r\left(\frac{w_{\pm,k+1}}{w_{\pm,k}}\right) + \beta r\left(\frac{w_{\pm,k}}{w_{\pm,k+1}}\right) \right]$, where $r(z) = (z-1) - \ln(|z|)$ for $z \neq 0$

Experimentally: Again $\tilde{\theta}_0$ is negligible and $\theta^{\text{SMGD}} \approx \arg\min_{\theta^* \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^*)$.

Theoretically: If the iterates $w_{\pm,k}$ do not change signs, then $S_\pm > 0$ and $\Delta_\infty < \Delta_0$.

$\Delta_\infty, \tilde{\theta}_0 = O(\alpha^2)$ $\Delta_\infty, \tilde{\theta}_0 \ll \theta^*$, $\forall \theta^* \in \mathcal{S}$ Sparse solutions for small α : sparser than θ^{GF}

