Leveraging Continuous Time to Understand Momentum When Training Diagonal Linear Networks

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

1. Discrete and Continuous Momentum

Setup: $F: \mathbb{R}^D \to \mathbb{R}_{\geq 0}$ — 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): $\lambda \ddot{w}_t + \dot{w}_t + \nabla F(w_t) = 0$ $\xrightarrow{\nabla F \text{ locally Lipschitz}}$ Unique solution on $[0, +\infty)$

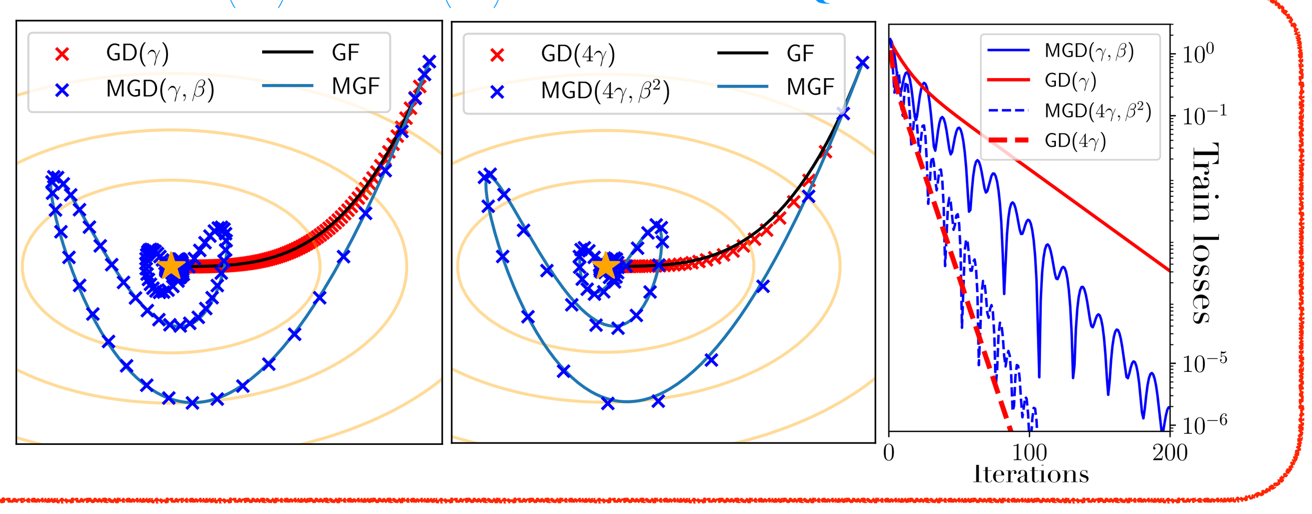
Discretization Scheme:

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

Proposition. For $(w_0, w_1) \in \mathbb{R}^{2D}$, consider MGF($\lambda$) with $\lambda = \dfrac{\gamma}{(1-\beta)^2}$ 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



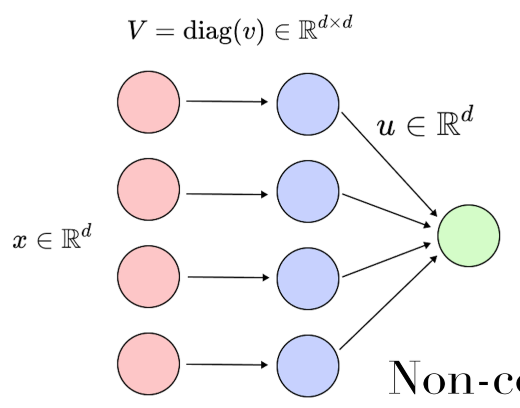
3. Momentum on Diagonal Linear Networks

Data: Sparse and underdetermined sample $(x_i, y_i)_{i=1}^n$

$x_i \in \mathbb{R}^d$, $y_i = \langle \theta_s^\star, x_i \rangle$, $n < d$, $\theta_s^\star$ — s-sparse

$\mathcal{S} = \theta_s^\star + \text{span}(x_1, \ldots, x_n)^\perp$ — set of interpolators

Architecture: 2-layer diagonal linear neural network



$f_w(x) = \langle u \odot v, x \rangle$

$w = \begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{R}^{2d}$ Useful quantities: $w_\pm = u \pm v$

Regression setting with $\theta = u \odot v$ reparametrization!

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

MGF and Stochastic MGD:

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

$u_{k+1} = u_k - \gamma \nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k) \odot v_k + \beta(u_k - u_{k-1})$
$v_{k+1} = v_k - \gamma \nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k) \odot u_k + \beta(v_k - v_{k-1})$

(Predictors) $\theta_t = u_t \odot v_t$ and $\theta_k = u_k \odot v_k$

(Balancedness) $\Delta_t := |u_t^2 - v_t^2| \in \mathbb{R}^d_{\geq 0}$
$\Delta_k := |u_k^2 - v_k^2| \in \mathbb{R}^d_{\geq 0}$

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: $\Phi: \mathbb{R}^d \to \mathbb{R}$ — 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 $\Delta \in \mathbb{R}^d_{>0}$, $\psi_\Delta(\theta) = \dfrac{1}{4} \sum_{i=1}^d \left(2\theta_i \operatorname{arcsinh}\left(\dfrac{2\theta_i}{\Delta_i}\right) - \sqrt{4\theta_i^2 + \Delta_i^2} + \Delta_i \right) : \mathbb{R}^d \to \mathbb{R}$.

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

Implicit Bias of Gradient Flow ($\lambda = 0$): $\theta^{\text{GF}} = \operatorname{argmin}_{\theta^\star \in \mathcal{S}} D_{\psi_{\Delta_0}}(\theta^\star, \theta_0)$

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

$\rightarrow$ Recovery of sparse interpolators!

5. 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.

$\rightarrow$ $\lambda \dfrac{d^2 \nabla \Phi_t(\theta_t)}{dt^2} + \dfrac{d \nabla \Phi_t(\theta_t)}{dt} + \nabla \mathcal{L}(\theta_t) = 0$, so $\nabla \Phi_\infty(\theta_\infty) - \nabla \Phi_0(\theta_0) \in \text{span}(x_1, \ldots, x_n)$.

$\rightarrow$ Find $\tilde{\theta}_0$ such that $\nabla \Phi_\infty(\tilde{\theta}_0) = \nabla \Phi_0(\theta_0)$.

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

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

†Woodworth et al., *Kernel and rich regimes in overparametrized models*. COLT 2020

2. Intertwined Roles of γ and β

Acceleration Rule: $\gamma \to \rho^2 \cdot \gamma$ + $\beta \to 1 - \rho(1-\beta)$ $\rightarrow$ 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 $\hat{\beta} = 1 - \rho(1-\beta) \approx_{\beta \to 1} \beta^\rho$. Then, MGD($\hat{\gamma}, \hat{\beta}$) initialized at (w_0, w_1) again becomes the discretization of MGF(λ) but with discretization step $\hat{\varepsilon} = \rho \cdot \varepsilon$.

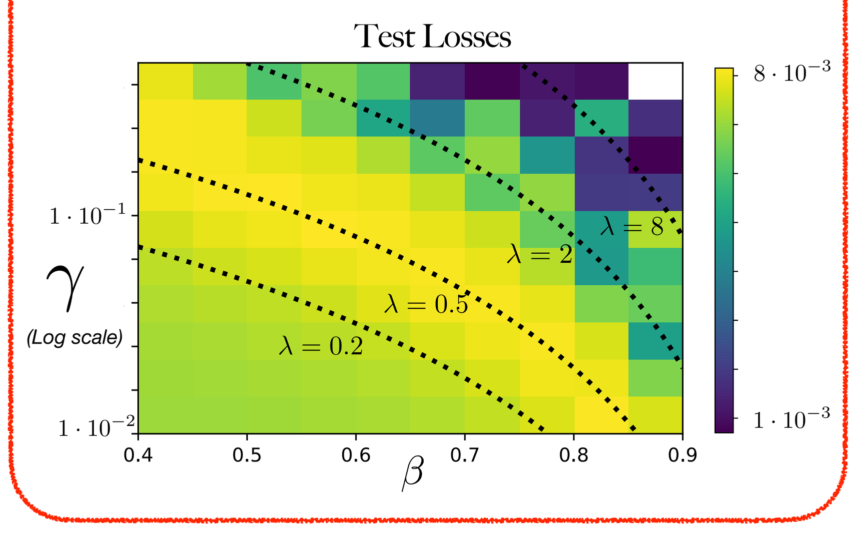
Optimization Regimes: The link between λ, γ, and β

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

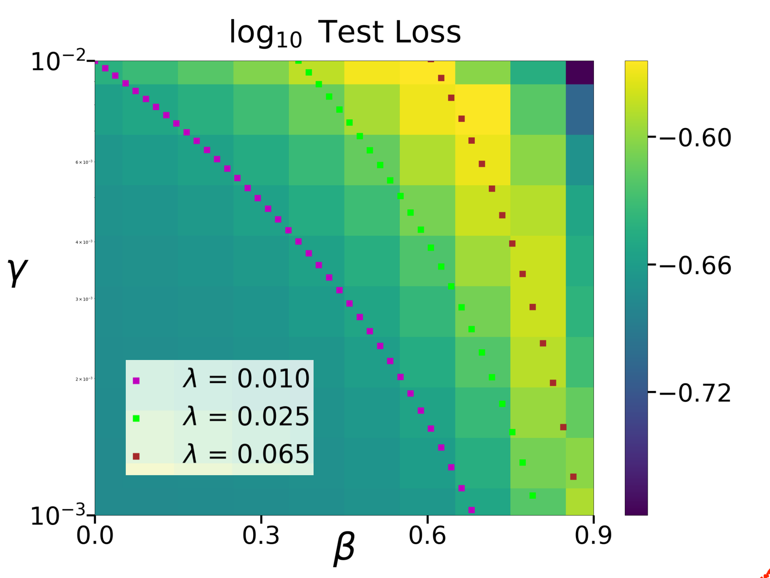
Small γ: $\rightarrow$ $\lambda \ll 1$: Gradient flow trajectory

Non-Degenerate λ: $\rightarrow$ The momentum regime

Teacher-Student 1-Hidden-Layer ReLU Network



Fully Connected Deep Linear Network



4. Continuous-Time Results

Implicit Bias of MGF(λ): $\theta^{\text{MGF}} = \operatorname{argmin}_{\theta^\star \in \mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0)$

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

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

Experimentally: $\tilde{\theta}_0$ is negligible and $\theta^{\text{MGF}} \approx \operatorname{argmin}_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$.

Gradient flow initialized at $(u_0 = \sqrt{\Delta_\infty}, v_0 = 0)$ and $(\dot{u}_0 = 0, \dot{v}_0 = 0)$ converges to $\theta^{\text{GF}}_{\Delta_\infty} = \operatorname{argmin}_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$. $\xrightarrow{\text{Simulate}}$ $\|\theta^{\text{MGF}} - \theta^{\text{GF}}_{\Delta_\infty}\|_2 / \|\theta^{\text{GF}}_{\Delta_\infty}\|_2 \ll 1$

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

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

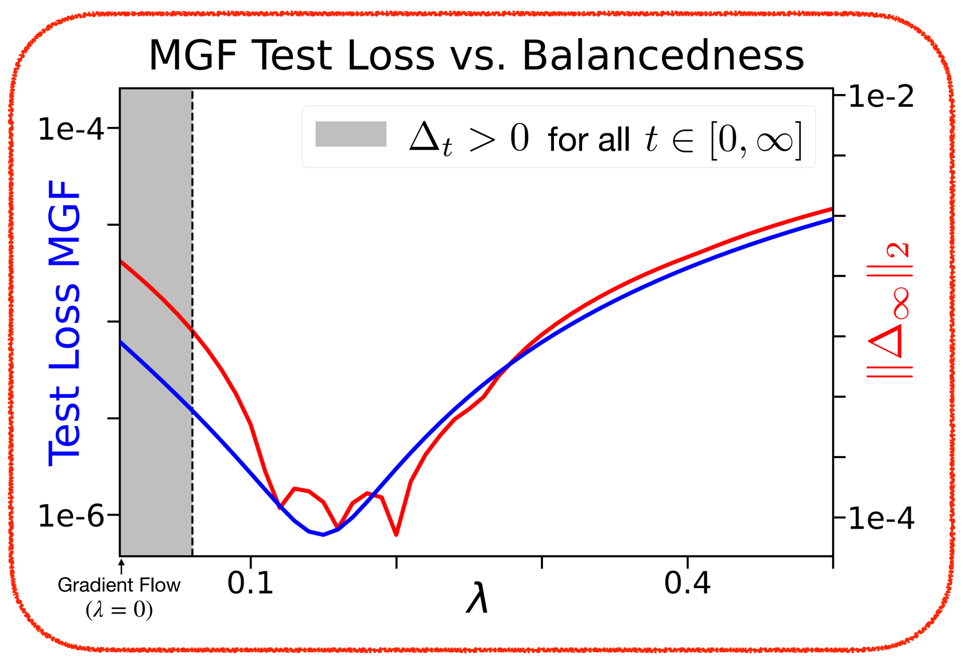
$\rightarrow$ $D_{\psi_{\Delta_\infty}}(\theta^\star, \theta_0) \sim_{\alpha \to 0} \psi_{\Delta_\infty}(\theta^\star) \propto_{\alpha \to 0} \|\theta^\star\|_1$ $\rightarrow$ Sparse solutions!

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

Small λ Regime:

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

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



6. Discrete-Time Results

Implicit Bias of SMGD(γ, β): $\theta^{\text{SMGD}} = \operatorname{argmin}_{\theta^\star \in \mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0)$

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

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

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

Experimentally: Again $\tilde{\theta}_0$ is negligible and $\theta^{\text{SMGD}} \approx \operatorname{argmin}_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$.

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

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