Understanding Gradient Regularization in Deep Learning: Efficient Finite-Difference Computation and Implicit Bias

Ryo Karakida*, Tomoumi Takase*, Tomohiro Hayase† & Kazuki Osawa♦

*AIST (Japan), †Cluster (Japan), ♦ ETH Zurich (Switzerland)

ICML 2023

Gradient Regularization (GR)

[Barrett+, Smith+, ICLR ‘21], [Zhao+ ICML ‘22]

$$\tilde{\mathcal{L}}(\theta) = \mathcal{L}(\theta) + \frac{\gamma}{2} R(\theta), \quad R(\theta) = \|\nabla_\theta \mathcal{L}(\theta)\|^2$$

- Explicitly decreasing GR enhances a convergence to flat minima and generalization performance: **Explicit GR**

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \tilde{\mathcal{L}}\left(\theta_t\right)$$

- Discrete update of (S)GD implicitly decreases GR: **Implicit GR**

Backward error analysis
(Approx. by continuous time)

$$\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}\left(\theta_t\right)$$

$$\dot{\theta} = -\nabla \tilde{\mathcal{L}}(\theta) + \mathcal{O}\left(\eta^2\right) \quad (\gamma = \eta/4)$$

Gradient Regularization (GR)

Algorithms for explicit GR

Requires computation of "gradient of gradient"

- **Double Backpropagation (DB)** $\nabla \|\nabla \mathcal{L}(\theta_t)\|^2$ Auto-grad. ×2

  [Drucker & LeCun 1992]

- **Finite Difference** $(\varepsilon > 0)$

  - Forward GR (F-GR) $$\Delta R_F(\varepsilon) = \frac{\nabla \mathcal{L}(\theta_t + \varepsilon \nabla \mathcal{L}(\theta_t)) - \nabla \mathcal{L}(\theta_t)}{\varepsilon}$$

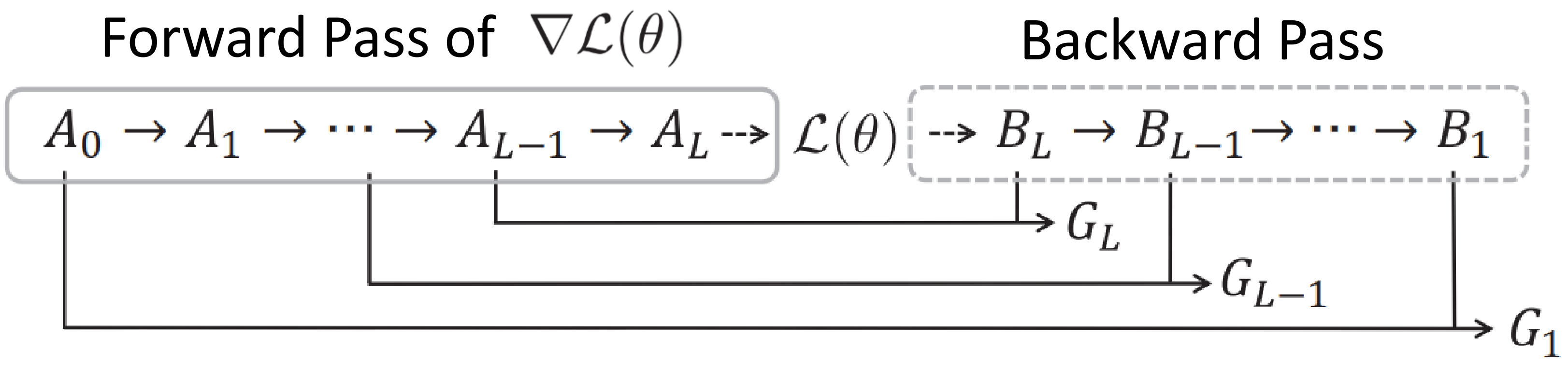
  - Backward GR (B-GR) $$\Delta R_B(\varepsilon) = \frac{\nabla \mathcal{L}(\theta_t) - \nabla \mathcal{L}(\theta_t - \varepsilon \nabla \mathcal{L}(\theta_t))}{\varepsilon}$$

Note: Centered or high-order finite differences are not used here because they require more gradient computation (backpropagation)

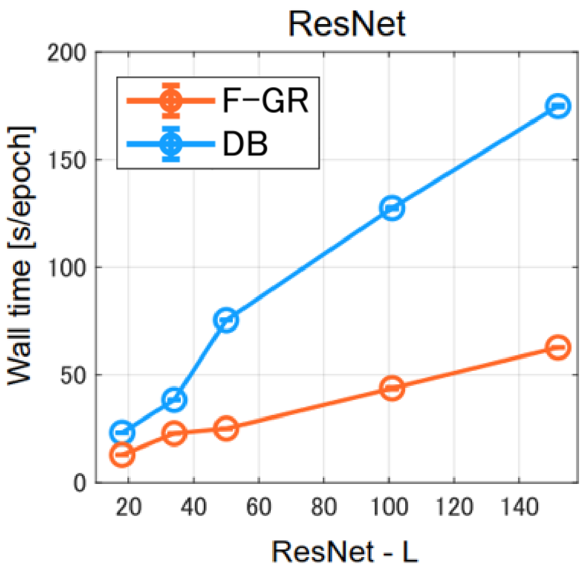
Result 1: Efficiency of GR algorithms

- Computational cost (measured by # of matrix multiplication)

For L-layered MLP, **$6L$ (Finite difference) < $9L$ (DB)**



Computational graph of DB: Each node with an incoming arrow requires one matrix multiplication for the forward pass.



Result 2: Dependence on GR Algorithms

- Double Backpropagation (DB)
- Finite Difference

$$\Delta R(\varepsilon) = \frac{\nabla \mathcal{L}\left(\theta_t + \varepsilon \nabla \mathcal{L}\left(\theta_t\right)\right) - \nabla \mathcal{L}\left(\theta_t\right)}{\varepsilon}$$

Forward GR (F-GR): $\varepsilon > 0$
Backward GR (B-GR): $\varepsilon < 0$

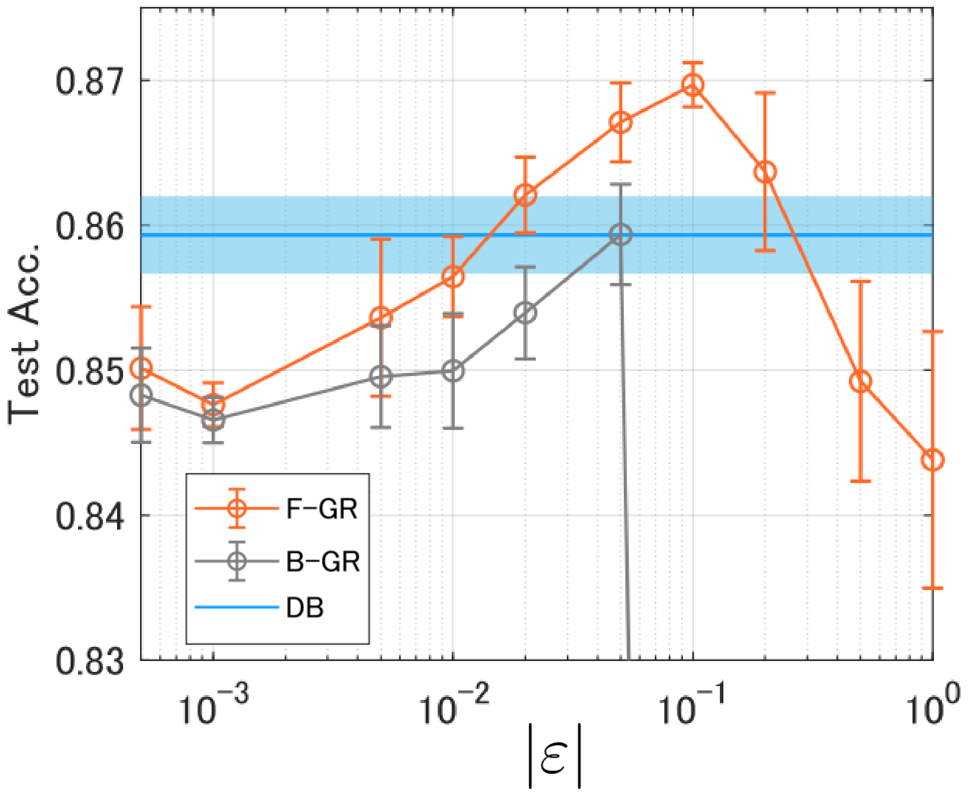


Fig: ResNet-18 trained by SGD w/ GR. On CIFAR-10.

Generalization performance highly depends on the choice of algorithms.
F-GR achieves better performance than B-GR and DB. A relatively large ascent step ($\varepsilon \sim 0.1$) is the best.

Implicit Bias in Diagonal Linear Network (DLN)

[Woodworth, … & Srebro, COLT '20]

$$f(x) = \sum_{i=1}^{D} \underbrace{(w_{+,i}^2 - w_{-,i}^2)}_{=:\,\beta_i} x_i$$

Settings: MSE Loss $L(w) = \sum_{\mu=1}^{N} \|y^\mu - f(x^\mu)\|_2^2$

Gradient dynamics $dw/dt = -\nabla \mathcal{L}$

Initialization scale $\alpha = w_{\pm,i}(t=0)$

Evaluate interpolation solutions: $X\beta = y$

Implicit Bias in Diagonal Linear Network (DLN)

[Woodworth, … & Srebro, COLT '20]

If gradient dynamics converges to the interpolation solution β^∞, it depends on α and satisfies

$$\beta^\infty(\alpha) = \underset{\beta \in \mathbb{R}^D \text{ s.t. } X\beta = y}{\arg\min} \; \phi_\alpha(\beta)$$

$$\phi_\alpha(\beta) = \sum_{i=1}^{D} \alpha^2 q\left(\beta_i/\alpha^2\right), \quad q(z) = 2 - \sqrt{4 + z^2} + z \operatorname{arcsinh}(z/2)$$

- Initialization scale (α) changes the minima

$\alpha \gg 1$: Lazy regime

$$\phi_\alpha(\beta) \sim \|\beta\|_2^2$$

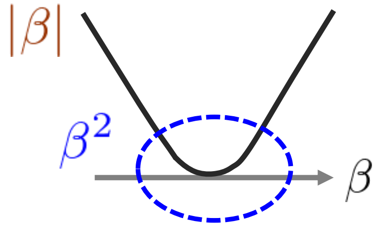
L2 norm regularization

$\alpha \ll 1$: Rich regime

$$\phi_\alpha(\beta) \sim \|\beta\|_1$$

L1 norm regularization

Shape of ϕ



Result 3: Analysis of GR in DLN

$$\frac{dw}{dt} = -\nabla\mathcal{L}(w) - \gamma\Delta R(\varepsilon) \qquad \Delta R(\varepsilon) = \frac{\nabla\mathcal{L}\left(\theta_t + \varepsilon\nabla\mathcal{L}\left(\theta_t\right)\right) - \nabla\mathcal{L}\left(\theta_t\right)}{\varepsilon}$$

Remind: F-GR ($\varepsilon > 0$), B-GR ($\varepsilon < 0$), DB ($\varepsilon \to 0$)

If the gradient dynamics with GR converges to the interpolation solution,

$$\beta^\infty(\alpha_{GR}) = \underset{\beta\in\mathbb{R}^D \text{ s.t. } X\beta=y}{\arg\min} \phi_{\alpha_{GR}}(\beta), \quad \alpha_{GR} = \alpha \circ \exp\left(-\gamma\left(c_0 + \varepsilon c_1 + \varepsilon^2 c_2\right) + \mathcal{O}\left(\gamma^2\right)\right)$$

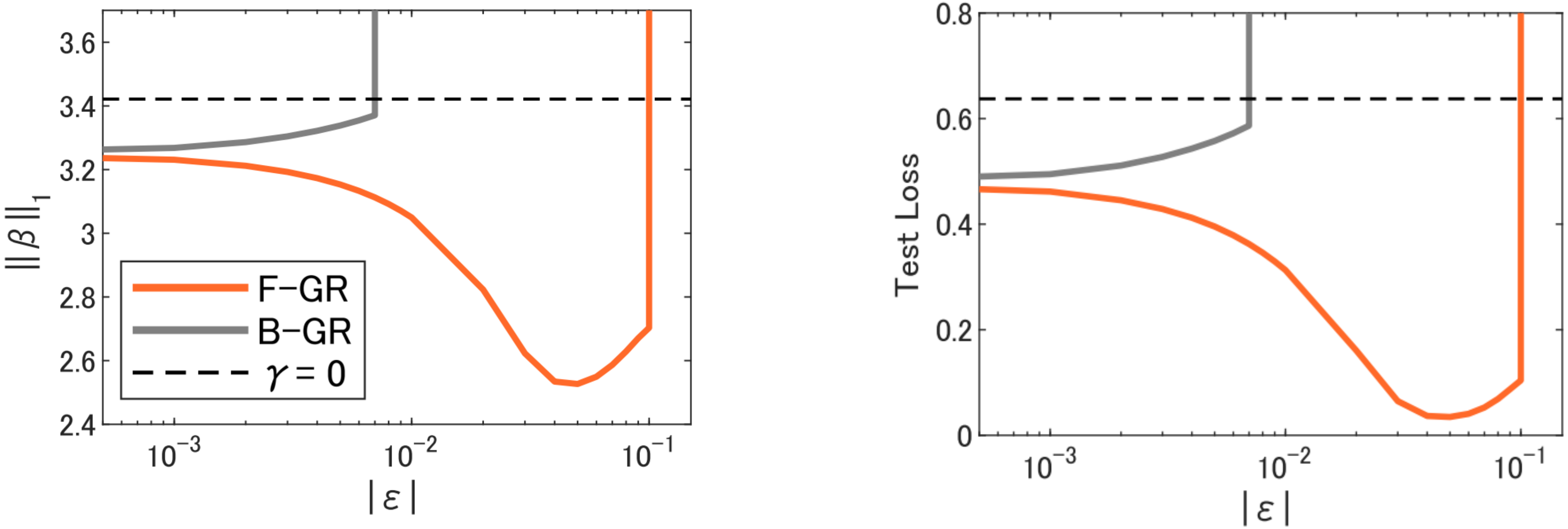
$$c_0 = \int_0^\infty (X^\top(X\beta(s) - y))^2 ds/n^2, \quad c_1 = (X^\top(X\beta(t=0) - y))^2/2n^2$$

- For F-GR, $\alpha_{GR,i} \lesssim \alpha_i \exp\left(-\gamma\varepsilon c_{1,i}/2\right)$ As ε increases, biased towards L1 (Rich regime)
- For B-GR, $\alpha_{GR,i} \gtrsim \alpha_i D^\gamma \exp\left(\gamma|\varepsilon|c_{1,i}\right)$ As |ε| increases, biased towards L2 (Lazy regime)

Experiments: DLN

Artificial data: input $x \in \mathbb{R}^D$ given by i.i.d Gaussian, $y \sim \mathcal{N}\left(\langle \beta^*, x \rangle, 0.01\right)$

(D=100, sample size $N = 50, \beta^*$: 5 non-zero entries)

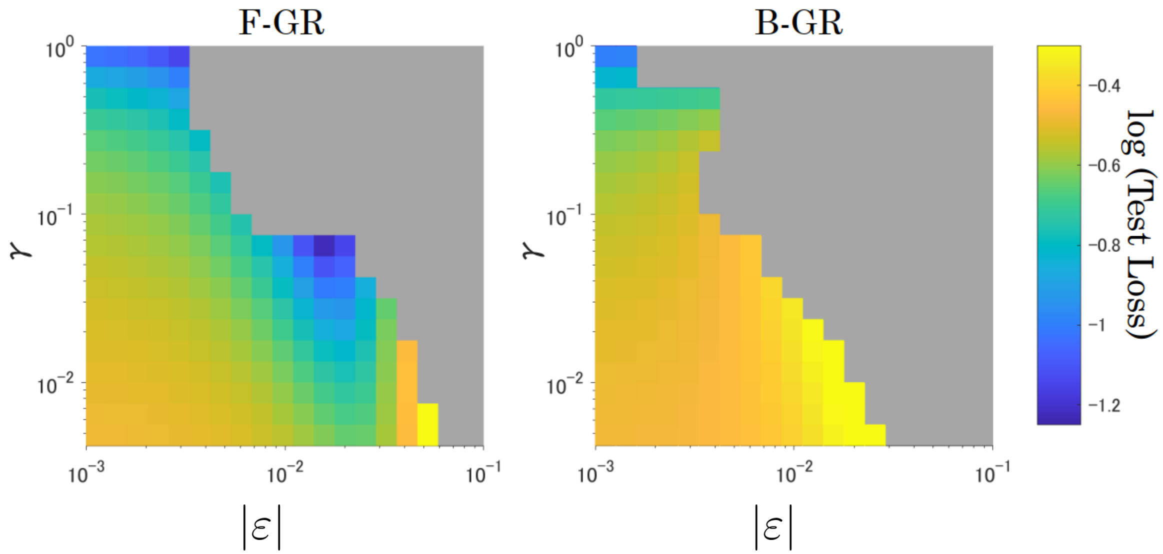


- Non-monotonicity of F-GR comes from c_2 (which is empirically negative)

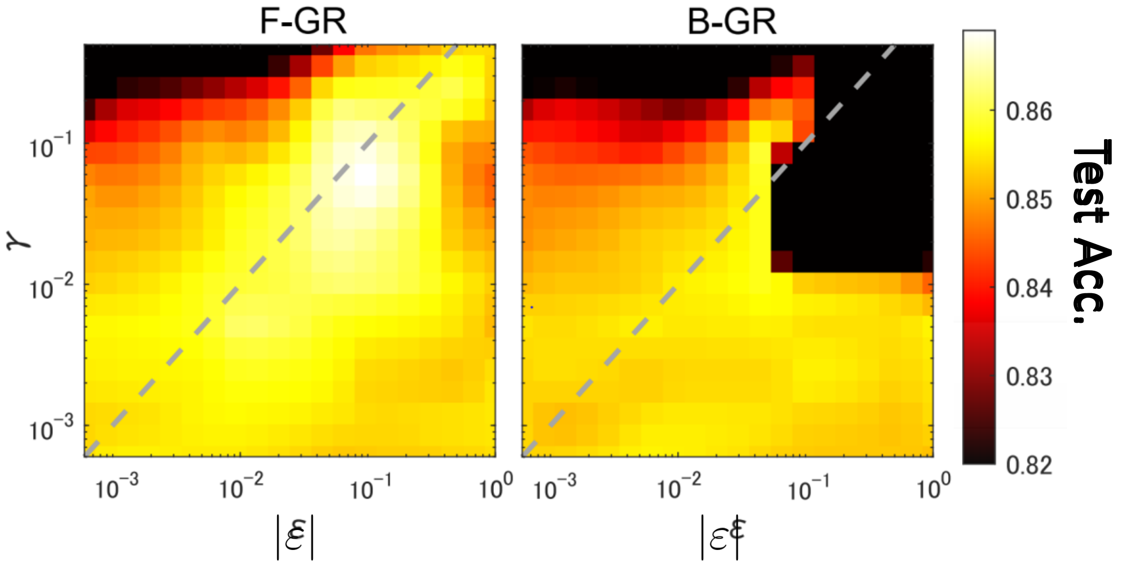
$$\alpha_{GR} = \alpha_0 \circ \exp(-\gamma(c_0 + \varepsilon c_1 + \varepsilon^2 c_2) + \mathcal{O}(\gamma^2))$$

Experiments: Grid Search on (ε, γ)

- DLN on artificial data



- ResNet-18 on CIFAR-10



Both cases are consistent in

- F-GR achieves the highest accuracy on large ascent steps
- B-GR is worse in the accuracy and is likely to explode

Result 4: Relation to other methods

Sharpness-aware minimalization (SAM) [Foret+ ICLR '21]

$$\nabla\mathcal{L}(\theta) + \frac{\gamma}{\varepsilon}\left(\nabla\mathcal{L}(\theta') - \nabla\mathcal{L}(\theta)\right) = \nabla\mathcal{L}(\theta') \quad \xrightarrow[\gamma=\varepsilon]{} \quad \theta' = \theta_t + \rho\nabla\mathcal{L}(\theta_t)$$

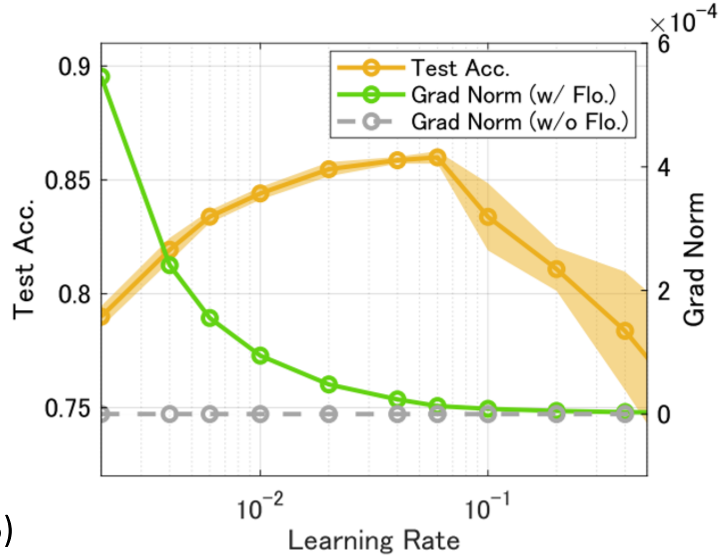
Flooding [Ishida+ ICML '21]

$$\tilde{\mathcal{L}}(\theta) = |\mathcal{L}(\theta) - b| + b \quad (b > 0), \quad \theta_{t+1} = \theta_t - \eta\,\mathrm{Sgn}(\mathcal{L} - b)\nabla\mathcal{L}$$

- Floating up from "water surface" at time t, (*i.e.*, $\mathcal{L}(\theta_t) < b, \quad \mathcal{L}(\theta_{t+1}) > b$),

$$\theta_{t+2} = \theta_t - \eta^2 \frac{\nabla\mathcal{L}(\theta_t + \eta\nabla\mathcal{L}(\theta_t)) - \nabla\mathcal{L}(\theta_t)}{\eta}$$

Equivalent to F-GR w/ $\eta = \gamma = \varepsilon$



ResNet-18 on CIFAR-10 trained with SGD + flooding (b=0.05)