Non-Euclidean Gradient Descent Operates at the Edge of Stability

Rustem Islamov[1] Michael Crawshaw[2] Jeremy Cohen[3] Robert M. Gower[3]

[1]University of Basel [2]George Mason University [3]Flatiron Institute

The Edge of Stability (EoS)

[Cohen et al., 2021] observed that training neural networks with vanilla `GD` exhibits two regimes

- Progressive Sharpening: loss decreases monotonically, while the sharpness grows;
- Edge of Stability (`EoS`): loss oscillates, but goes down over time; the sharpness oscillates around $2/\eta$.

Does EoS extend beyond vanilla `GD` to a family of all non-Euclidean methods (e.g, ℓ_∞-descent, `Spectral GD`) and their normalized variants?

Non-Euclidean Gradient Descent

For a norm $\|\cdot\|$ with associated dual norm $\|\cdot\|_*$, the **non-Euclidean GD** method minimizes a regularized linearization of the loss $\mathcal{L}$ at $\mathbf{w}_t$:

$$\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w}} f(\mathbf{w}_t) + \langle \nabla f(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t \rangle + \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_t\|^2$$
$$= \mathbf{w}_t - \eta \underbrace{\|\nabla \mathcal{L}(\mathbf{w}_t)\|_* (\nabla \mathcal{L}(\mathbf{w}_t))_*}_{:= \mathbf{d}_t \text{ (update direction)}},$$

where $(\nabla \mathcal{L}(\mathbf{w}_t))_* := \arg\max_{\|\mathbf{y}\|=1} \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{y} \rangle$ is the linear maximization oracle. Omitting the dual-norm factor gives **normalized** non-Euclidean GD:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\, (\nabla \mathcal{L}(\mathbf{w}_t))_*, \quad \mathbf{d}_t := (\nabla \mathcal{L}(\mathbf{w}_t))_*.$$

Block $\ell_{1,2}$ Norm

Let $\mathbf{w} = (\mathbf{w}^1, \dots, \mathbf{w}^L) \in \mathbb{R}^{d_1} \oplus \cdots \oplus \mathbb{R}^{d_L}$ with the norm $\|\cdot\|_{1,2}$ defined as

$$\|\mathbf{w}\|_{1,2} = \sum_{\ell=1}^{L} \|\mathbf{w}^\ell\|_2.$$

Let $\ell_{\max} := \arg\max_{\ell \in [L]} \|\nabla_{\mathbf{w}^\ell} \mathcal{L}(\mathbf{w}_t)\|_2$, then `GD` in this norm is `Block GD`

$$\mathbf{w}_{t+1}^{\ell_{\max}} = \mathbf{w}_t^{\ell_{\max}} - \eta \nabla_{\mathbf{w}^{\ell_{\max}}} \mathcal{L}(\mathbf{w}_t)$$
$$\mathbf{w}_{t+1}^{\ell} = \mathbf{w}_t^{\ell} \quad \text{otherwise.}$$

The generalized sharpness has a closed-form solution:

$$S^{\|\cdot\|_{1,2}}(\mathbf{w}) = \max_{\ell \in [L]} \lambda_{\max}(\nabla^2_{\mathbf{w}^\ell} \mathcal{L}(\mathbf{w})).$$

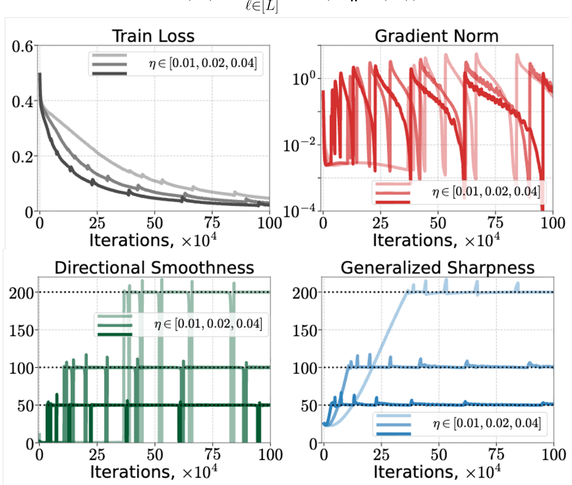

Figure 1: `Block CD` in training MLP on CIFAR10-5k.

Directional Smoothness

We consider the equality

$$f(\mathbf{x}) = f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{D^{\|\cdot\|}(\mathbf{y}, \mathbf{x})}{2}\|\mathbf{x} - \mathbf{y}\|^2,$$

where $D^{\|\cdot\|}(\mathbf{y}, \mathbf{x})$ is the **directional smoothness** given by

$$D^{\|\cdot\|}(\mathbf{y}, \mathbf{x}) := \frac{\mathcal{L}(\mathbf{x}) - \mathcal{L}(\mathbf{y}) - \langle \nabla \mathcal{L}(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle}{\frac{1}{2}\|\mathbf{y} - \mathbf{w}\|^2}$$

Substituting one step of non-Euclidean `GD` gives the **key identity**

$$\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) = -\eta\left(1 - \tfrac{\eta}{2} D^{\|\cdot\|}(\mathbf{w}_t, \mathbf{w}_{t+1})\right) \|\nabla \mathcal{L}(\mathbf{w}_t)\|_*^2,$$

so whenever $\|\nabla \mathcal{L}(\mathbf{w}_t)\|_* > 0$ and the loss decreases monotonically, we have

$$\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) \le 0 \implies D^{\|\cdot\|}(\mathbf{w}_t, \mathbf{w}_{t+1}) \le 2/\eta.$$

If the loss oscillates, then

$$\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) \approx 0 \implies D^{\|\cdot\|}(\mathbf{w}_t, \mathbf{w}_{t+1}) \approx 2/\eta$$

ℓ_∞ Norm

Here we consider $\|\mathbf{w}\|_\infty := \max_{j \in [d]} |\mathbf{w}_j|$. The resulting method is ℓ_∞-descent is given by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \|\nabla \mathcal{L}(\mathbf{w}_t)\|_1 \operatorname{sign}(\nabla \mathcal{L}(\mathbf{w}_t)).$$

The generalized sharpness under this norm is

$$S^{\|\cdot\|_\infty}(\mathbf{w}) := \max_{\|\mathbf{d}\|=1} \mathbf{d}^\top \nabla^2 \mathcal{L}(\mathbf{w}) \mathbf{d}$$

which is NP-hard problem and thus does not have a closed-form solution.

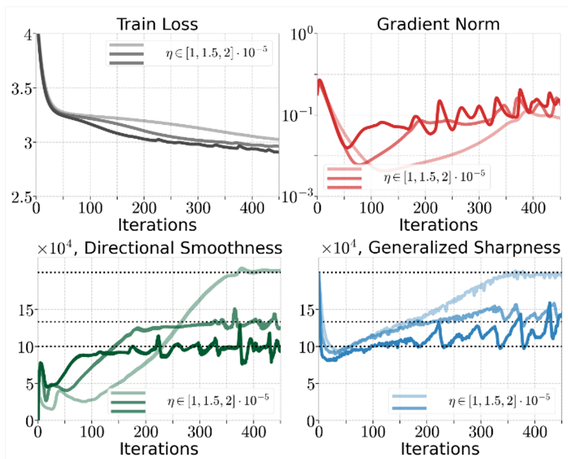

Figure 2: ℓ_∞-descent in training Transformer on Tiny Shakespeare.

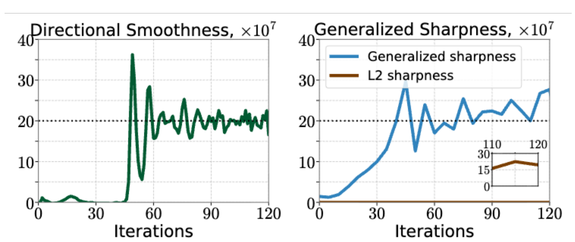

Figure 3: ℓ_∞-descnt in training VGG11 on CIFAR10.

Generalized Sharpness

Using the definition of $D^{\|\cdot\|}(\mathbf{w}_t, \mathbf{w}_{t+1})$, we have

$$D^{\|\cdot\|}(\mathbf{w}_t, \mathbf{w}_{t+1}) = \frac{\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) - \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle}{\frac{1}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2}$$
$$= \frac{\mathbf{d}_t^\top \nabla^2 \mathcal{L}(\mathbf{w}_t - \xi_t \eta \mathbf{d}_t)\mathbf{d}_t}{\|\mathbf{d}_t\|^2}, \quad \xi_t \in (0, 1)$$
$$\le \max_{\mathbf{d} \neq 0} \frac{\mathbf{d}^\top \nabla^2 \mathcal{L}(\mathbf{w}_t - \xi_t \eta \mathbf{d}_t)\mathbf{d}}{\|\mathbf{d}\|^2}.$$

Assuming $\nabla^2 \mathcal{L}(\mathbf{w}_t - \xi_t \eta \mathbf{d}_t) \approx \nabla^2 \mathcal{L}(\mathbf{w}_t)$, then we arrive at the following definition of the **generalized sharpness**

$$S^{\|\cdot\|}(\mathbf{w}) := \max_{\mathbf{d} \neq 0} \frac{\mathbf{d}^\top \nabla^2 \mathcal{L}(\mathbf{w}) \mathbf{d}}{\|\mathbf{d}\|^2} = \max_{\|\mathbf{d}\|=1} \mathbf{d}^\top \nabla^2 \mathcal{L}(\mathbf{w}) \mathbf{d}.$$

Remark: for Euclidean norm $\|\cdot\|_2$ and Mahalanobis distance $\|\cdot\|_{\mathbf{P}_t}$ the definition of generalized sharpness matches those introduced in prior work [Cohen et al., 2021; Cohen et al., 2025].

Spectral $\|\cdot\|_{\infty,2}$ Norm.

Let $\mathbf{W} = (\mathbf{W}^1, \dots, \mathbf{W}^L) \in \mathbb{R}^{d_{\text{out}}^1 \times d_{\text{in}}^1} \oplus \cdots \oplus \mathbb{R}^{d_{\text{out}}^L \times d_{\text{in}}^L}$ with the inner product $\langle \mathbf{W}, \mathbf{G} \rangle := \operatorname{Tr}(\mathbf{W}^\top \mathbf{G})$ and norm defined as

$$\|\mathbf{W}\|_{\infty,2} := \max_{\ell \in [L]} \|\mathbf{W}^\ell\|_2, \quad \|\mathbf{W}^\ell\|_2 = \max_{\|\mathbf{d}\|_2=1} \|\mathbf{W}^\ell \mathbf{d}\|_2.$$

Under this geometry, the non-Euclidean `GD` is `Spectral GD`

$$\mathbf{W}_{t+1}^\ell = \mathbf{W}_t^\ell - \eta \left(\sum_{j=1}^{L} \operatorname{Tr}(\mathbf{\Sigma}_t^j)\right) \mathbf{U}_t^\ell \mathbf{V}_t^\ell,$$

where $\mathbf{U}_t^\ell \mathbf{\Sigma}_t^\ell \mathbf{V}_t^\ell = \nabla_{\mathbf{W}^\ell} \mathcal{L}(\mathbf{W}_t)$ is SVD of the gradient of ℓ^{th} layer.

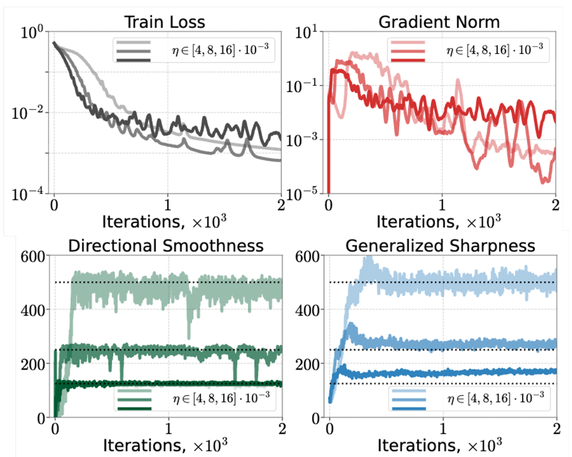

Figure 4: `Spectral GD` in training MLP on CIFAR10.

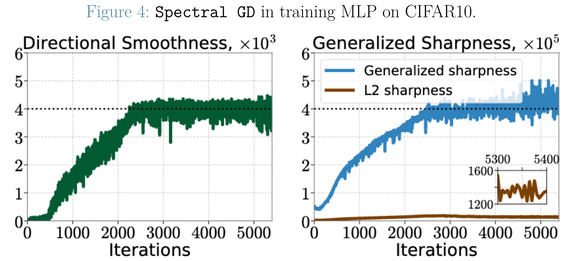

Figure 5: `Spectral GD` in training Resnet20 on CIFAR10.

Approximating Generalized Sharpness via Frank–Wolfe

To approximate the value of generalized sharpness, we use Frank-Wolfe algorithm (`FW`) that computes the approximation for the quadratic problem

$$\max \mathbf{d}^\top \nabla^2 \mathcal{L}(\mathbf{w}) \mathbf{d} \quad \text{s.t. } \|\mathbf{d}\|^2 \le 1.$$

Algorithm 1: `FW` for generalized sharpness

Input: norm $\|\cdot\|$, $\gamma_k = \frac{2}{2+k}$, $S_0 = 0$
For restart $m = 1, \dots, M$ **do**
  $\mathbf{d}_0 \sim \mathcal{N}(0, \mathbf{I})$, project $\mathbf{d}_0 \leftarrow \Pi_{\|\cdot\|=1}(\mathbf{d}_0)$ //Random initialization
  For $k = 0, \dots, K-1$ **do**
    $\mathbf{v}_k = \arg\max_{\|\mathbf{v}\| \le 1} \langle \nabla^2 \mathcal{L}(\mathbf{w}_t)\mathbf{d}_k, \mathbf{v} \rangle$ //Computing LMO
    $\mathbf{d}_{k+1} = (1 - \gamma_k)\mathbf{d}_k + \gamma_k \mathbf{v}_k$
  End For
  Project $\mathbf{u}_K = \Pi_{\|\cdot\|=1}(\mathbf{d}_K)$
  $\hat{S}_m = \mathbf{u}_K^\top \nabla^2 \mathcal{L}(\mathbf{w}_t)\mathbf{u}_K$; $S_m = \max\{S_{m-1}, \hat{S}_m\}$
End For
Return: S_M

Remark: restarts are needed to explore the high-dimensional unit ball $\{\mathbf{d} \mid \|\mathbf{d}\| \le 1\}$ better and provide a more accurate solution.

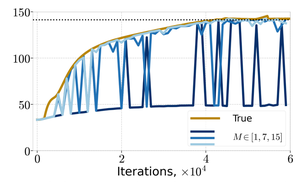

Figure 6: Comparison of the true value of the generalized sharpness against `FW` approximation varying the number of restarts $M \in [1, 7, 15]$ in training CNN model with `Block GD` on CIFAR10-5k.

Theory on Quadratics

We consider the problem

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbf{H} \mathbf{w}, \quad \mathbf{H} \succ 0, \quad \mathbf{H}^\top = \mathbf{H}.$$

It is known that for Euclidean norm, vanilla `GD` converges **iff** $\eta < \frac{2}{\lambda_{\max}(\mathbf{H})} = \frac{2}{S^{\|\cdot\|_2}(\mathbf{w})}$. With $\eta < \frac{2}{S^{\|\cdot\|}(\mathbf{w})}$ non-Euclidean `GD` does converge.

Convergence of Non-Euclidean `GD`

For some norm $\|\cdot\|$, non-Euclidean `GD` converges linearly from any initial point $\mathbf{w}_0$ with any step size $\eta < 2/S$, where $S = S^{\|\cdot\|}(\mathbf{w}) = \max_{\|\mathbf{d}\| \le 1} \mathbf{d}^\top \mathbf{H} \mathbf{d}$.

However, the converse ($\eta > \frac{2}{S}$) is not true in general, and convergence depends on the initialization. Let

$$\hat{\mathbf{d}} \in \arg\max_{\|\mathbf{d}\|=1} \mathbf{d}^\top \mathbf{H} \mathbf{d}, \quad \text{then } (\mathbf{H}\hat{\mathbf{d}})_* = \hat{\mathbf{d}}.$$

Convergence of Non-Euclidean `GD`

For some norm $\|\cdot\|$, non-Euclidean `GD` diverges from the starting point $\mathbf{w}_0 \in \operatorname{span}(\hat{\mathbf{d}})$ with any step size $\eta > 2/S$, where $S = S^{\|\cdot\|}(\mathbf{w}) = \max_{\|\mathbf{d}\| \le 1} \mathbf{d}^\top \mathbf{H} \mathbf{d}$.