Non-Euclidean Gradient Descent Operates at the Edge of Stability

Rustem Islamov
University of Basel

Joint work with:

Michael Crawshaw

Jeremy Cohen

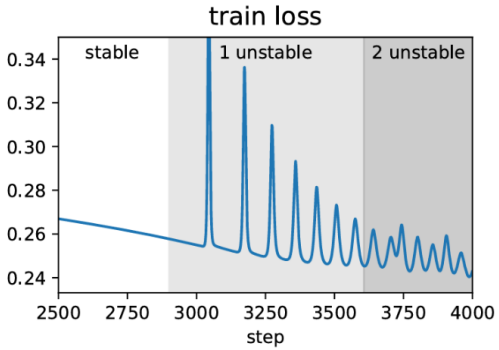
Robert Gower

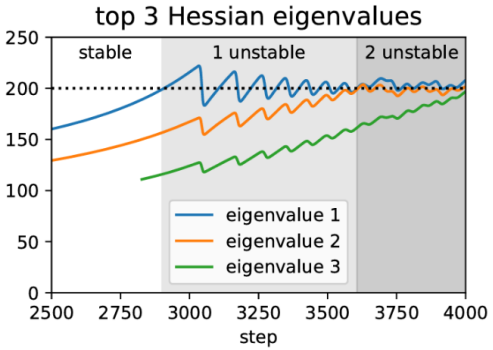
Background

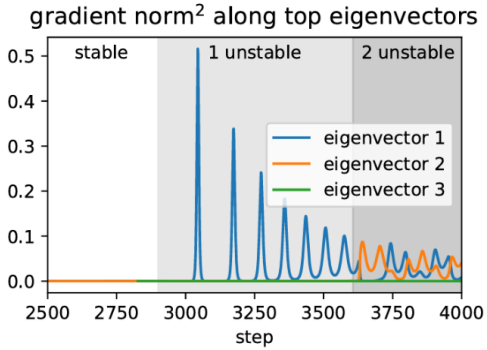
Train loss

Vanilla Gradient Descent

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \nabla f(\mathbf{w}_k)$$







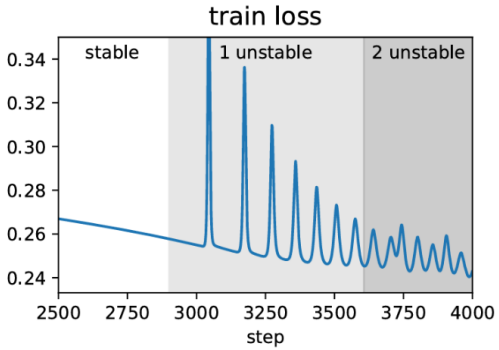
Cohen et al., Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability, ICLR 2021

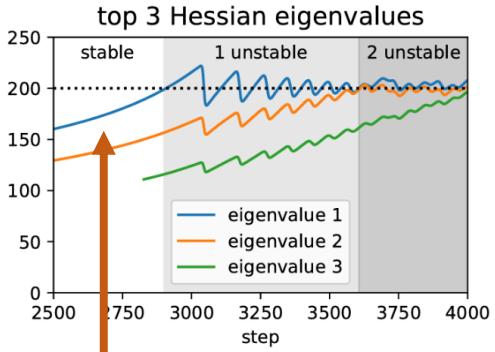
Background

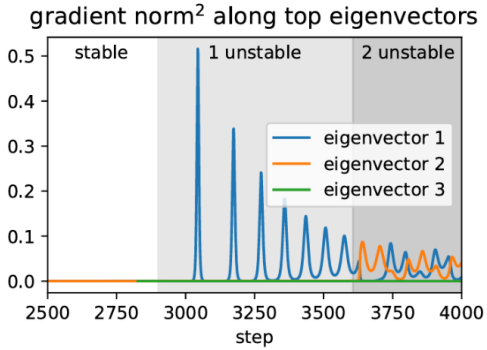
Train loss

Vanilla Gradient Descent

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \nabla f(\mathbf{w}_k)$$







Top eigenvalue of the Hessian (sharpness) grows to $2/\eta$

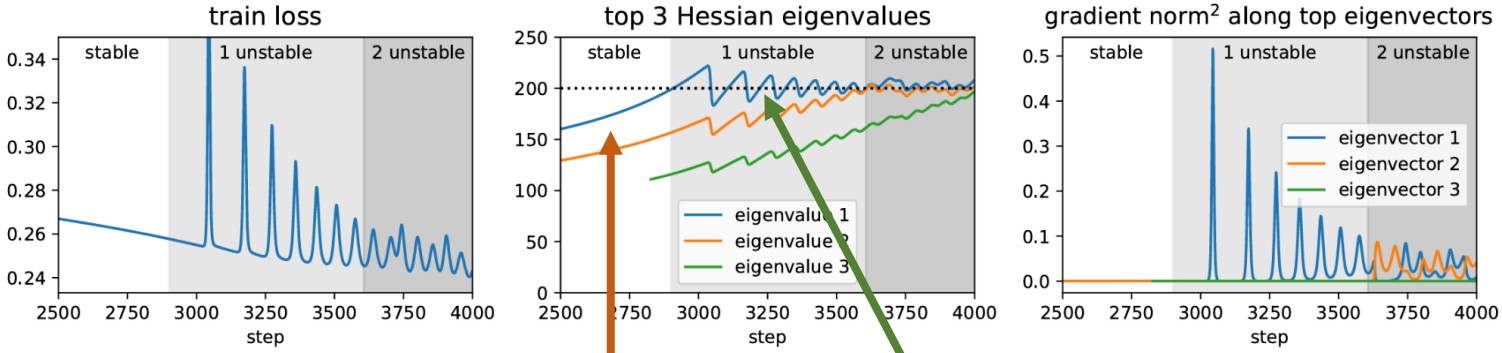
Cohen et al., Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability, ICLR 2021

Background

Train loss

Vanilla Gradient Descent

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \nabla f(\mathbf{w}_k)$$



Top eigenvalue of the Hessian (sharpness) grows to $2/\eta$

Top eigenvalue oscillates around $2/\eta$

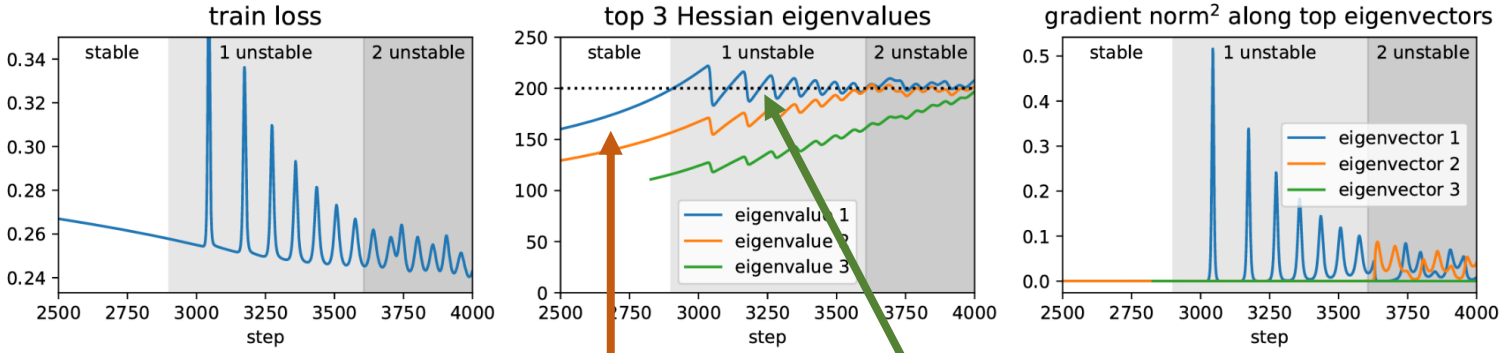
Cohen et al., Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability, ICLR 2021

Background

Vanilla Gradient Descent

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \nabla f(\mathbf{w}_k)$$

Train loss



Top eigenvalue of the Hessian (sharpness) grows to $2/\eta$

Top eigenvalue oscillates around $2/\eta$

Progressive sharpening

Edge of Stability

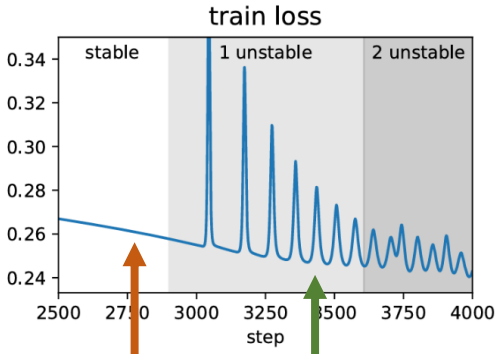
Cohen et al., Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability, ICLR 2021

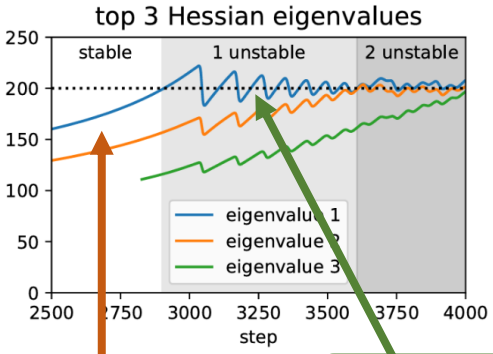
Background

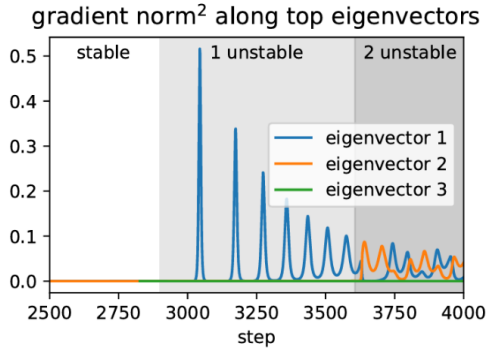
Train loss

Vanilla Gradient Descent

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \nabla f(\mathbf{w}_k)$$







Monotonically decreases

Spiking, but decreasing over time

Top eigenvalue of the Hessian (sharpness) grows to $2/\eta$

Progressive sharpening

Top eigenvalue oscillates around $2/\eta$

Edge of Stability

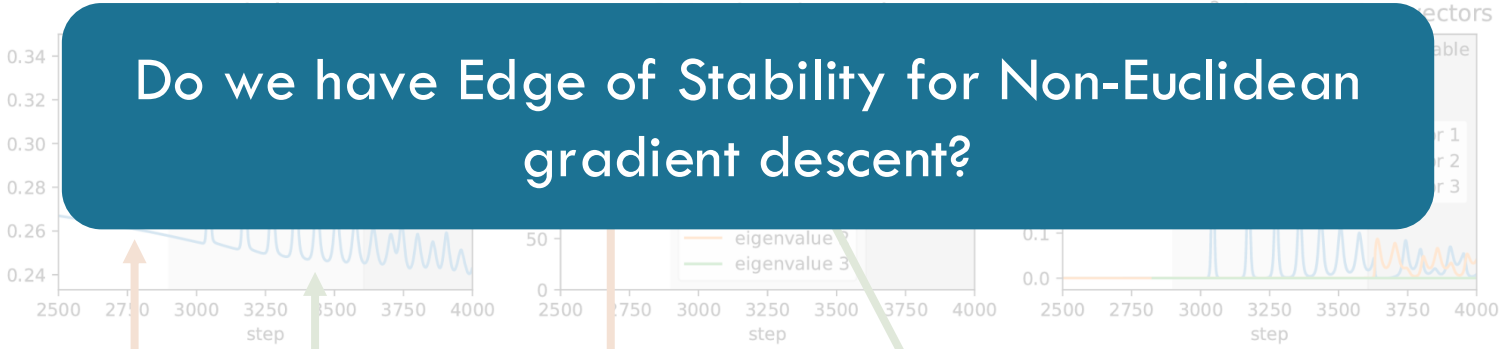
Cohen et al., Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability, ICLR 2021

Background

Train loss

Vanilla Gradient Descent $\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \nabla f(\mathbf{w}_k)$

Do we have Edge of Stability for Non-Euclidean gradient descent?



Monotonically decreases

Spiking, but decreasing over time

Top eigenvalue of the Hessian (sharpness) grows to $2/\eta$

Progressive sharpening

Top eigenvalue oscillates around $2/\eta$

Edge of Stability

Cohen et al., Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability, ICLR 2021

Directional Smoothness: A Tautology

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

Arbitrary norm

Directional Smoothness: A Tautology

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

$$D(\mathbf{x}, \mathbf{y}) = \frac{f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle}{\frac{1}{2} \|\mathbf{x} - \mathbf{y}\|^2}$$

Arbitrary norm

Directional Smoothness

Directional Smoothness: A Tautology

Non-Euclidean Gradient Descent

$$\mathbf{w}_{k+1} = \arg\min_{\mathbf{w}} f(\mathbf{w}_k) + \langle \nabla f(\mathbf{w}_k), \mathbf{w} - \mathbf{w}_k \rangle + \frac{1}{2\eta} \|\mathbf{w} - \mathbf{w}_k\|^2$$

$$= \mathbf{w}_k - \eta \|\nabla f(\mathbf{w}_k)\|_* \texttt{LMO}(\nabla f(\mathbf{w}_k))$$

$$\texttt{LMO}(\mathbf{m}) = \arg\max_{\|\mathbf{d}\|=1} \langle \mathbf{m}, \mathbf{d} \rangle$$

Directional Smoothness: A Tautology

Non-Euclidean Gradient Descent

$$\mathbf{w}_{k+1} = \arg\min_{\mathbf{w}} f(\mathbf{w}_k) + \langle \nabla f(\mathbf{w}_k), \mathbf{w} - \mathbf{w}_k \rangle + \frac{1}{2\eta} \|\mathbf{w} - \mathbf{w}_k\|^2$$

$$= \mathbf{w}_k - \eta \|\nabla f(\mathbf{w}_k)\|_* \texttt{LMO}(\nabla f(\mathbf{w}_k))$$

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

Directional Smoothness: A Tautology

Non-Euclidean Gradient Descent

$$\mathbf{w}_{k+1} = \arg\min_{\mathbf{w}} f(\mathbf{w}_k) + \langle \nabla f(\mathbf{w}_k), \mathbf{w} - \mathbf{w}_k \rangle + \frac{1}{2\eta} \|\mathbf{w} - \mathbf{w}_k\|^2$$

$$= \mathbf{w}_k - \eta \|\nabla f(\mathbf{w}_k)\|_* \texttt{LMO}(\nabla f(\mathbf{w}_k))$$

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

$$f(\mathbf{w}_{k+1}) = f(\mathbf{w}_k) - \eta \left(1 - \frac{\eta D(\mathbf{w}_{k+1}, \mathbf{w}_k)}{2}\right) \|\nabla f(\mathbf{w}_k)\|_*^2$$

Directional Smoothness: A Tautology

$$f(\mathbf{w}_{k+1}) = f(\mathbf{w}_k) - \eta \left(1 - \frac{\eta D(\mathbf{w}_{k+1}, \mathbf{w}_k)}{2}\right) \|\nabla f(\mathbf{w}_k)\|_*^2$$

Non-zero in practice

Directional Smoothness: A Tautology

$$f(\mathbf{w}_{k+1}) = f(\mathbf{w}_k) - \eta \left(1 - \frac{\eta D(\mathbf{w}_{k+1}, \mathbf{w}_k)}{2}\right) \|\nabla f(\mathbf{w}_k)\|_*^2$$

Non-zero in practice

$$f(\mathbf{w}_{k+1}) < f(\mathbf{w}_k)$$

Directional Smoothness: A Tautology

$$f(\mathbf{w}_{k+1}) = f(\mathbf{w}_k) - \eta \left(1 - \frac{\eta D(\mathbf{w}_{k+1}, \mathbf{w}_k)}{2}\right) \|\nabla f(\mathbf{w}_k)\|_*^2$$

Non-zero in practice

$$f(\mathbf{w}_{k+1}) < f(\mathbf{w}_k) \quad \Rightarrow \quad D(\mathbf{w}_{k+1}, \mathbf{w}_k) < \frac{2}{\eta}$$

Directional Smoothness: A Tautology

$$f(\mathbf{w}_{k+1}) = f(\mathbf{w}_k) - \eta \left(1 - \frac{\eta D(\mathbf{w}_{k+1}, \mathbf{w}_k)}{2}\right) \|\nabla f(\mathbf{w}_k)\|_*^2$$

Non-zero in practice

$$f(\mathbf{w}_{k+1}) < f(\mathbf{w}_k) \quad \Rightarrow \quad D(\mathbf{w}_{k+1}, \mathbf{w}_k) < \frac{2}{\eta}$$

$$f(\mathbf{w}_{k+1}) \approx f(\mathbf{w}_k)$$

Directional Smoothness: A Tautology

$$f(\mathbf{w}_{k+1}) = f(\mathbf{w}_k) - \eta \left(1 - \frac{\eta D(\mathbf{w}_{k+1}, \mathbf{w}_k)}{2}\right) \|\nabla f(\mathbf{w}_k)\|_*^2$$

Non-zero in practice

$$f(\mathbf{w}_{k+1}) < f(\mathbf{w}_k) \quad \rightarrow \quad D(\mathbf{w}_{k+1}, \mathbf{w}_k) < \frac{2}{\eta}$$

$$f(\mathbf{w}_{k+1}) \approx f(\mathbf{w}_k) \quad \rightarrow \quad D(\mathbf{w}_{k+1}, \mathbf{w}_k) \approx \frac{2}{\eta}$$

Generalized Sharpness

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \mathbf{d}_k$$

$$\mathbf{d}_k = \|\nabla f(\mathbf{w}_k)\|_* \texttt{LMO}(\nabla f(\mathbf{w}_k))$$

$$D(\mathbf{w}_{k+1}, \mathbf{w}_k) := \frac{f(\mathbf{w}_{k+1}) - f(\mathbf{w}_k) - \langle \nabla f(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle}{\frac{1}{2}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2}$$

Generalized Sharpness

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \mathbf{d}_k$$

$$\mathbf{d}_k = \|\nabla f(\mathbf{w}_k)\|_* \texttt{LMO}(\nabla f(\mathbf{w}_k))$$

$$D(\mathbf{w}_{k+1}, \mathbf{w}_k) := \frac{f(\mathbf{w}_{k+1}) - f(\mathbf{w}_k) - \langle \nabla f(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle}{\frac{1}{2}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2}$$

$$= \frac{\mathbf{d}_k^\top \nabla^2 f(\mathbf{w}_k - \xi_k \eta \mathbf{d}_k)\mathbf{d}_k}{\|\mathbf{d}_k\|^2}, \xi_k \in (0, 1)$$

Generalized Sharpness

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \mathbf{d}_k$$

$\mathbf{d}_k = \|\nabla f(\mathbf{w}_k)\|_* \texttt{LMO}(\nabla f(\mathbf{w}_k))$

$$D(\mathbf{w}_{k+1}, \mathbf{w}_k) := \frac{f(\mathbf{w}_{k+1}) - f(\mathbf{w}_k) - \langle \nabla f(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle}{\frac{1}{2}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2}$$

$$= \frac{\mathbf{d}_k^\top \nabla^2 f(\mathbf{w}_k - \xi_k \eta \mathbf{d}_k)\mathbf{d}_k}{\|\mathbf{d}_k\|^2}, \xi_k \in (0,1)$$

Hessian does not change significantly

$$\lesssim \max_{\|\mathbf{d}\|=1} \mathbf{d}^\top \nabla^2 f(\mathbf{w}_k)\mathbf{d}$$

Generalized Sharpness

Frank-Wolfe Approximation

$$S^{\|\cdot\|}(\mathbf{w}) = \max_{\|\mathbf{d}\|=1} \mathbf{d}^\top \nabla^2 f(\mathbf{w}) \mathbf{d}$$

NP-hard problem in general

Frank-Wolfe Approximation

$$S^{\|\cdot\|}(\mathbf{w}) = \max_{\|\mathbf{d}\|=1} \mathbf{d}^\top \nabla^2 f(\mathbf{w}) \mathbf{d}$$

NP-hard problem in general

$$= \max_{\|d\|\leq 1} \mathbf{d}^\top \nabla^2 f(\mathbf{w}) \mathbf{d}$$

Holds if the Hessian is not negative definite

Can be approximated by Frank-Wolfe

Frank-Wolfe Approximation

$$\max_{\|\mathbf{d}\| \leq 1} \mathbf{d}^\top \nabla^2 f(\mathbf{w}) \mathbf{d} \qquad (\star)$$

Algorithm 1 Frank-Wolfe to approximate $(\star)$

Input: norm $\|\cdot\|$, $\gamma_k = \frac{2}{2+k}$, $S_0 = 0$
for restart $m = 1, \ldots, M$ **do**
  $\mathbf{d}_0 \sim \mathcal{N}(0, \boldsymbol{I})$, $\mathbf{d}_0 = \Pi_{\|\cdot\|=1}(\mathbf{d}_0)$
  for $k = 0, 1, \ldots, K-1$ **do**
    $\mathbf{v}_k = \operatorname{argmax}_{\|\mathbf{v}\| \leq 1} \langle \nabla^2 \mathcal{L}(\mathbf{w}_t)\mathbf{d}_k, \mathbf{v} \rangle$
    $\mathbf{d}_{k+1} = (1-\gamma_k)\mathbf{d}_k + \gamma_k \mathbf{v}_k$
  end for
  $\mathbf{u}_K = \Pi_{\|\cdot\|=1}(\mathbf{d}_K)$
  $\hat{S}_m = \mathbf{d}_K^\top \nabla^2 \mathcal{L}(\mathbf{w}_t)\mathbf{d}_K$
  $S_m = \max\{S_{m-1}, \hat{S}_m\}$
end for
Return: S_M

Frank-Wolfe Approximation

$$\max_{\|\mathbf{d}\| \leq 1} \mathbf{d}^\top \nabla^2 f(\mathbf{w}) \mathbf{d} \qquad (\star)$$

Algorithm 1 Frank-Wolfe to approximate $(\star)$

Input: norm $\|\cdot\|$, $\gamma_k = \frac{2}{2+k}$, $S_0 = 0$
for restart $m = 1, \ldots, M$ **do**
  $\mathbf{d}_0 \sim \mathcal{N}(0, \boldsymbol{I})$, $\mathbf{d}_0 = \Pi_{\|\cdot\|=1}(\mathbf{d}_0)$
  for $k = 0, 1, \ldots, K-1$ **do**
    $\mathbf{v}_k = \operatorname{argmax}_{\|\mathbf{v}\| \leq 1} \langle \nabla^2 \mathcal{L}(\mathbf{w}_t) \mathbf{d}_k, \mathbf{v} \rangle$
    $\mathbf{d}_{k+1} = (1 - \gamma_k)\mathbf{d}_k + \gamma_k \mathbf{v}_k$
  end for
  $\mathbf{u}_K = \Pi_{\|\cdot\|=1}(\mathbf{d}_K)$
  $\hat{S}_m = \mathbf{d}_K^\top \nabla^2 \mathcal{L}(\mathbf{w}_t) \mathbf{d}_K$
  $S_m = \max\{S_{m-1}, \hat{S}_m\}$
end for
Return: S_M

One run of FW

Frank-Wolfe Approximation

$$\max_{\|\mathbf{d}\| \leq 1} \mathbf{d}^\top \nabla^2 f(\mathbf{w}) \mathbf{d} \qquad (\star)$$

Algorithm 1 Frank-Wolfe to approximate $(\star)$

Input: norm $\|\cdot\|$, $\gamma_k = \frac{2}{2+k}$, $S_0 = 0$
for restart $m = 1, \ldots, M$ **do**
  $\mathbf{d}_0 \sim \mathcal{N}(0, \boldsymbol{I})$, $\mathbf{d}_0 = \Pi_{\|\cdot\|=1}(\mathbf{d}_0)$
  for $k = 0, 1, \ldots, K-1$ **do**
    $\mathbf{v}_k = \operatorname{argmax}_{\|\mathbf{v}\| \leq 1} \langle \nabla^2 \mathcal{L}(\mathbf{w}_t) \mathbf{d}_k, \mathbf{v} \rangle$
    $\mathbf{d}_{k+1} = (1 - \gamma_k)\mathbf{d}_k + \gamma_k \mathbf{v}_k$
  end for
  $\mathbf{u}_K = \Pi_{\|\cdot\|=1}(\mathbf{d}_K)$
  $\hat{S}_m = \mathbf{d}_K^\top \nabla^2 \mathcal{L}(\mathbf{w}_t) \mathbf{d}_K$
  $S_m = \max\{S_{m-1}, \hat{S}_m\}$
end for
Return: S_M

We do restarts to find a better solution

One run of FW

Special Cases

$$S^{\|\cdot\|}(\mathbf{w}) = \max_{\|\mathbf{d}\|=1} \mathbf{d}^\top \nabla^2 f(\mathbf{w}) \mathbf{d}$$

Euclidean Norm:

$$S^{\|\cdot\|_2}(\mathbf{w}) = \lambda_{\max}(\nabla^2 f(\mathbf{w}))$$

Matches the definition of [Cohen et. al, 2021]

Special Cases

$$S^{\|\cdot\|}(\mathbf{w}) = \max_{\|\mathbf{d}\|=1} \mathbf{d}^\top \nabla^2 f(\mathbf{w}) \mathbf{d}$$

Euclidean Norm:

$$S^{\|\cdot\|_2}(\mathbf{w}) = \lambda_{\max}(\nabla^2 f(\mathbf{w}))$$

Matches the definition of [Cohen et. al, 2021]

Preconditioned Euclidean Norm:

$$S^{\|\cdot\|_{\mathbf{P}_k}}(\mathbf{w}) = \lambda_{\max}(\mathbf{P}_k^{-1/2} \nabla^2 f(\mathbf{w}) \mathbf{P}_k^{-1/2})$$

Preconditioner

Matches the definition of [Cohen et. al, 2025]

Special Cases: Infinity Norm

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

No closed-form solution; NP-hard in general

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \|\nabla f(\mathbf{w}_k)\|_1 \mathtt{sign}(\nabla f(\mathbf{w}_k))$$

ℓ_∞ -descent

Special Cases: Infinity Norm

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \|\nabla f(\mathbf{w}_k)\|_1 \mathtt{sign}(\nabla f(\mathbf{w}_k))$$

ℓ_∞ -descent

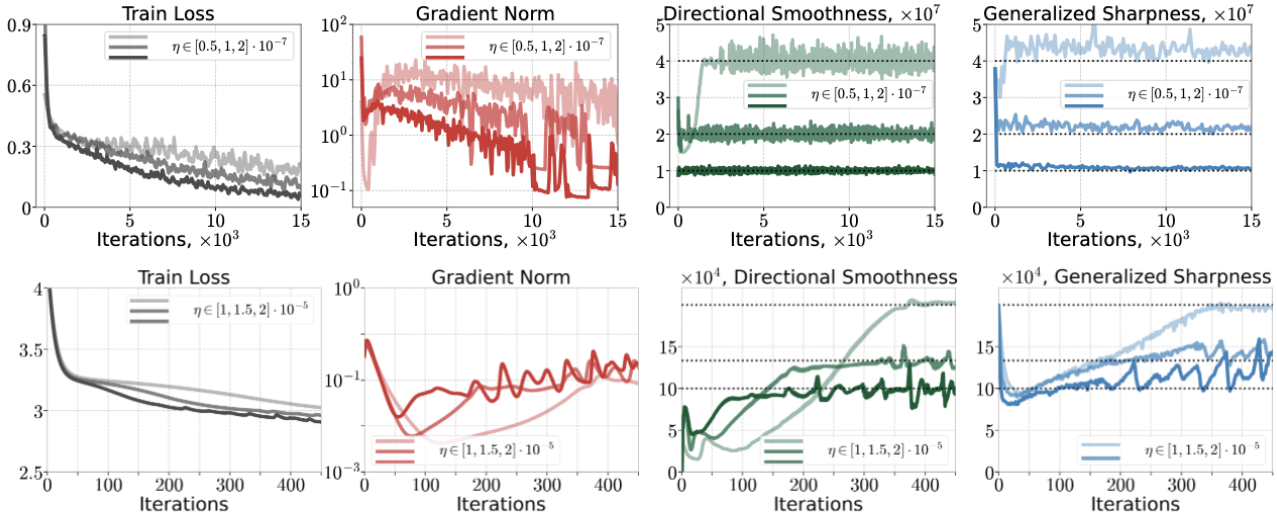


Figure 2: ℓ_∞-descent on **Top**: MLP CIFAR-10-5k; **bottom**: Transformer on Tiny Shakespeare. Columns show train loss, gradient norm, directional smoothness, and generalized sharpness (14). Dashed lines mark the stability threshold $2/\eta$.

Special Cases: Product Spectral Norm

$$\mathbf{W} = (\mathbf{W}^1, \ldots, \mathbf{W}^L) \in \mathbb{R}^{d_1^{\text{out}} \times d_1^{\text{in}}} \oplus \cdots \oplus \mathbb{R}^{d_L^{\text{out}} \times d_L^{\text{in}}}$$

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

Special Cases: Product Spectral Norm

$$\mathbf{W} = (\mathbf{W}^1, \dots, \mathbf{W}^L) \in \mathbb{R}^{d_1^{\text{out}} \times d_1^{\text{in}}} \oplus \cdots \oplus \mathbb{R}^{d_L^{\text{out}} \times d_L^{\text{in}}}$$

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

$$\mathbf{W}_{k+1}^l = \mathbf{W}_k^l - \eta \left(\sum_{j=1}^{L} \operatorname{Tr}(\mathbf{\Sigma}_k^j) \right) \mathbf{U}_k^l \mathbf{V}_k^l, \quad \nabla_l f(\mathbf{W}_k) = \mathbf{U}_k^l \mathbf{\Sigma}_k^j \mathbf{V}_k^l$$

Spectral GD

Polar factor of the gradient

SVD

Special Cases: Product Spectral Norm

$$\mathbf{W}_{k+1}^l = \mathbf{W}_k^l - \eta \left(\sum_{j=1}^{L} \mathrm{Tr}(\mathbf{\Sigma}_k^j) \right) \mathbf{U}_k^l \mathbf{V}_k^l, \quad \nabla_l f(\mathbf{W}_k) = \mathbf{U}_k^l \mathbf{\Sigma}_k^j \mathbf{V}_k^l$$

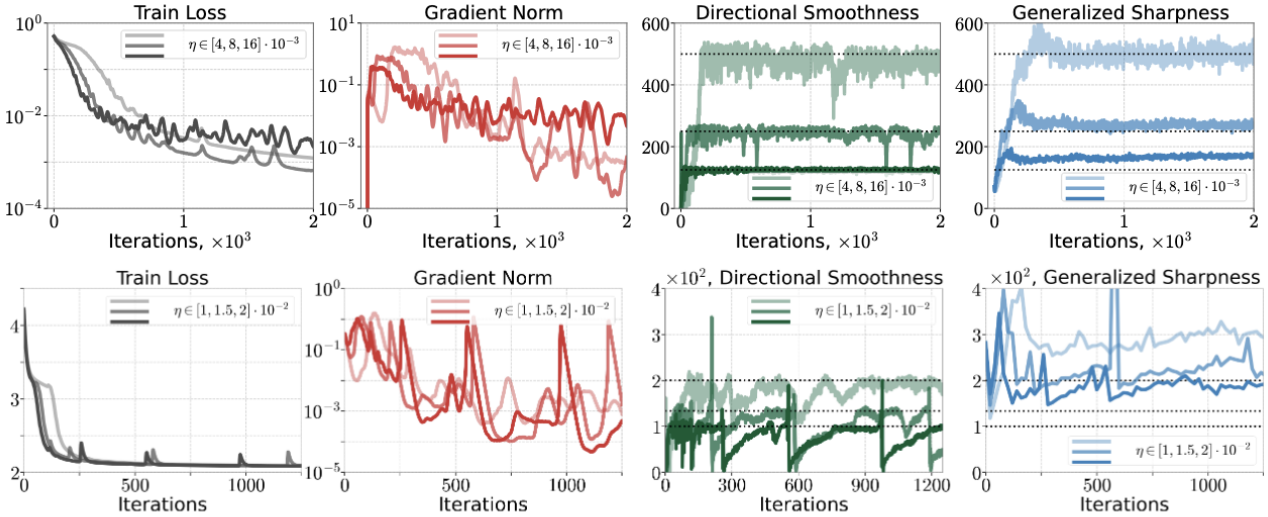


Figure 4: `Spectral GD` on **Top**: MLP, CIFAR-10, **bottom**: Transformer, Tiny Shakespeare. Columns show train loss, gradient norm, directional smoothness, and generalized sharpness (19). Dashed lines mark the stability threshold $2/\eta$.

Normalized Non-Euclidean GD

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \texttt{LMO}(\mathbf{w}_k)$$

$$\frac{2\|\nabla f(\mathbf{w}_k)\|_*}{\eta} \approx D(\mathbf{w}_{k+1}, \mathbf{w}_k), \qquad \frac{2\|\nabla f(\mathbf{w}_k)\|_*}{\eta} \lesssim S^{\|\cdot\|}(\mathbf{w}_k)$$

Normalized Non-Euclidean GD

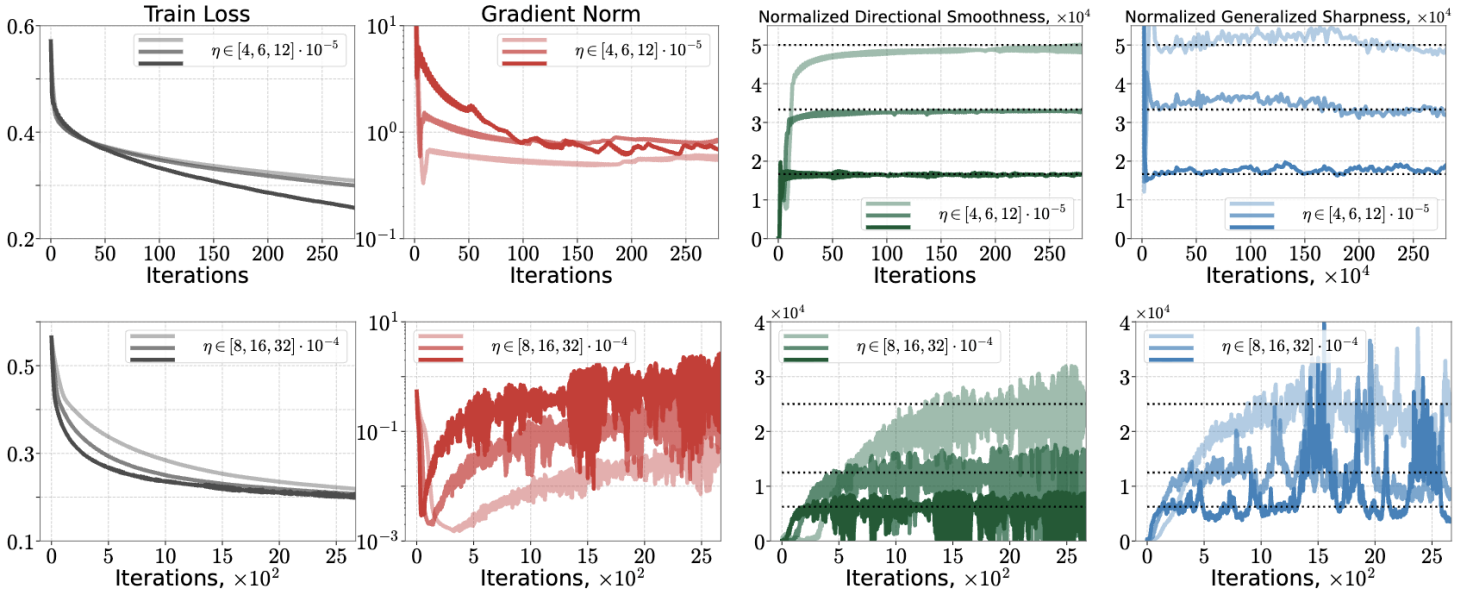


Figure 5: Normalized non-Euclidean GD. **Top**: `SignGD` on CIFAR-10-5k; **bottom**: normalized `Spectral GD` on CIFAR-10. Columns show train loss, gradient norm, directional smoothness divided by $\|\nabla \mathcal{L}(\mathbf{w}_t)\|_*$, and generalized sharpness divided by $\|\nabla \mathcal{L}(\mathbf{w}_t)\|_*$. Dashed lines mark the stability threshold $2/\eta$.

Convergence on Quadratics

$$f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbf{H}\mathbf{w}, \mathbf{H} = \mathbf{H}^\top, \mathbf{H} \succ 0$$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* = \|\mathbf{H}(\mathbf{x} - \mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\|$$

$= S^{\|\cdot\|}(\mathbf{w})$

Convergence on Quadratics

$$f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbf{H}\mathbf{w}, \mathbf{H} = \mathbf{H}^\top, \mathbf{H} \succ 0$$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* = \|\mathbf{H}(\mathbf{x} - \mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\|$$

$= S^{\|\cdot\|}(\mathbf{w})$

Euclidean GD converges iff $\eta < \frac{2}{\lambda_{\max}(\mathbf{H})} = \frac{2}{S^{\|\cdot\|_2}(\mathbf{w})}$

Convergence on Quadratics

$$f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbf{H}\mathbf{w}, \mathbf{H} = \mathbf{H}^\top, \mathbf{H} \succ 0$$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* = \|\mathbf{H}(\mathbf{x}-\mathbf{y})\|_* \leq L\|\mathbf{x}-\mathbf{y}\|$$

$= S^{\|\cdot\|}(\mathbf{w})$

Euclidean GD converges iff $\eta < \frac{2}{\lambda_{\max}(\mathbf{H})} = \frac{2}{S^{\|\cdot\|_2}(\mathbf{w})}$

Non-Euclidean GD converges if $\eta < \frac{2}{S^{\|\cdot\|}(\mathbf{w})}$

Convergence on Quadratics

$$f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbf{H}\mathbf{w}, \mathbf{H} = \mathbf{H}^\top, \mathbf{H} \succ 0$$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* = \|\mathbf{H}(\mathbf{x}-\mathbf{y})\|_* \leq L\|\mathbf{x}-\mathbf{y}\|$$

$= S^{\|\cdot\|}(\mathbf{w})$

Euclidean GD converges iff $\eta < \frac{2}{\lambda_{\max}(\mathbf{H})} = \frac{2}{S^{\|\cdot\|_2}(\mathbf{w})}$

Non-Euclidean GD converges if $\eta < \frac{2}{S^{\|\cdot\|}(\mathbf{w})}$

It is possible to construct a convergent example for ℓ_∞-descent with initialization set of positive measure and step size $\eta > 2/S^{\|\cdot\|_\infty}(\mathbf{w})$

Convergence on Quadratics

$$f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbf{H}\mathbf{w}, \mathbf{H} = \mathbf{H}^\top, \mathbf{H} \succ 0$$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* = \|\mathbf{H}(\mathbf{x}-\mathbf{y})\|_* \leq L\|\mathbf{x}-\mathbf{y}\|$$

$= S^{\|\cdot\|}(\mathbf{w})$

Euclidean GD converges iff $\eta < \frac{2}{\lambda_{\max}(\mathbf{H})} = \frac{2}{S^{\|\cdot\|_2}(\mathbf{w})}$

Non-Euclidean GD converges if $\eta < \frac{2}{S^{\|\cdot\|}(\mathbf{w})}$

Let $\hat{\mathbf{d}} = \text{argmax}_{\|\mathbf{d}\|=1} \mathbf{d}^\top \mathbf{H}\mathbf{d}$ **and** $\mathbf{w}_0 \in \text{span}(\hat{\mathbf{d}})$

then Non-Euclidean GD converges iff $\eta < \frac{2}{S^{\|\cdot\|}(\mathbf{w})}$

It is possible to construct a convergent example for ℓ_∞ -descent with initialization set of positive measure and step size $\eta > 2/S^{\|\cdot\|_\infty}(\mathbf{w})$

Thank you! Questions?

arXiv: 2603.05002

I am on a job market,

looking for an intern and research scientist positions

rustem.islamov.github.io

Minimizing Quadratic Model

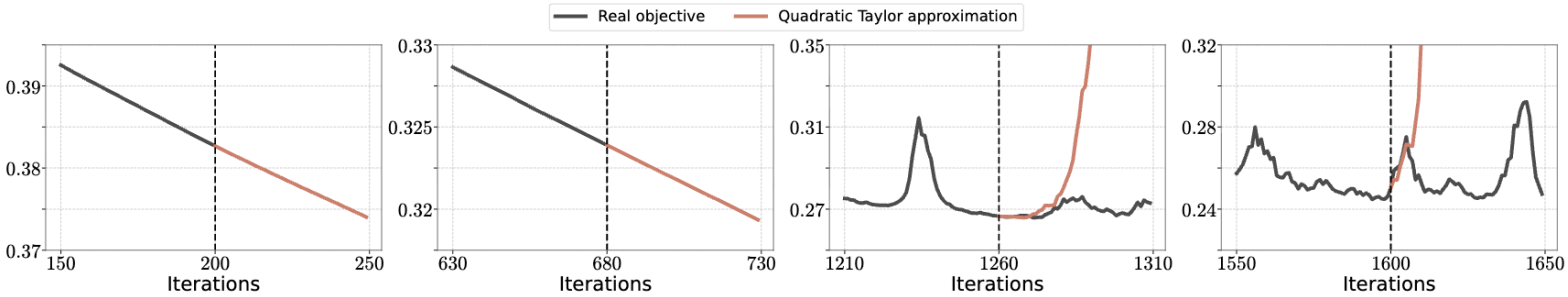


Figure 6: MSE loss for `Spectral GD` on a CNN trained on CIFAR-10 with $\eta = 0.002$. At four marked iterations, we switch from the true objective to the quadratic Taylor approximation at the current iterate. In the two left panels, before `EoS`, the quadratic approximation closely tracks the true loss; in the two right panels, during `EoS`, it quickly diverges.

Special Cases: Block L12 Norm

$$\mathbf{w} = (\mathbf{w}^1, \ldots, \mathbf{w}^L) \in \mathbb{R}^{d_1} \oplus \cdots \oplus \mathbb{R}^{d_L}$$

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

$$\mathbf{w}_{k+1}^l = \mathbf{w}_k^l - \eta \nabla_l f(\mathbf{w}_k), \text{ if } l = \text{argmax}_{s \in [L]} \|\nabla_s f(\mathbf{w}_k)\|_2$$

$$\mathbf{w}_{k+1}^l = \mathbf{w}_k^l, \text{ otherwise}$$

Greedy Block CD

$$S^{\|\cdot\|_{1,2}}(\mathbf{w}) = \max_{l \in [L]} \lambda_{\max}(\nabla_l^2 f(\mathbf{w}))$$

If the Hessian is PSD

Special Cases: Block L12 Norm

$$\mathbf{w}_{k+1}^l = \mathbf{w}_k^l - \eta \nabla_l f(\mathbf{w}_k), \text{ if } l = \text{argmax}_{s\in[L]} \|\nabla_s f(\mathbf{w}_k)\|_2, \quad \mathbf{w}_{k+1}^l = \mathbf{w}_k^l, \text{ otherwise}$$

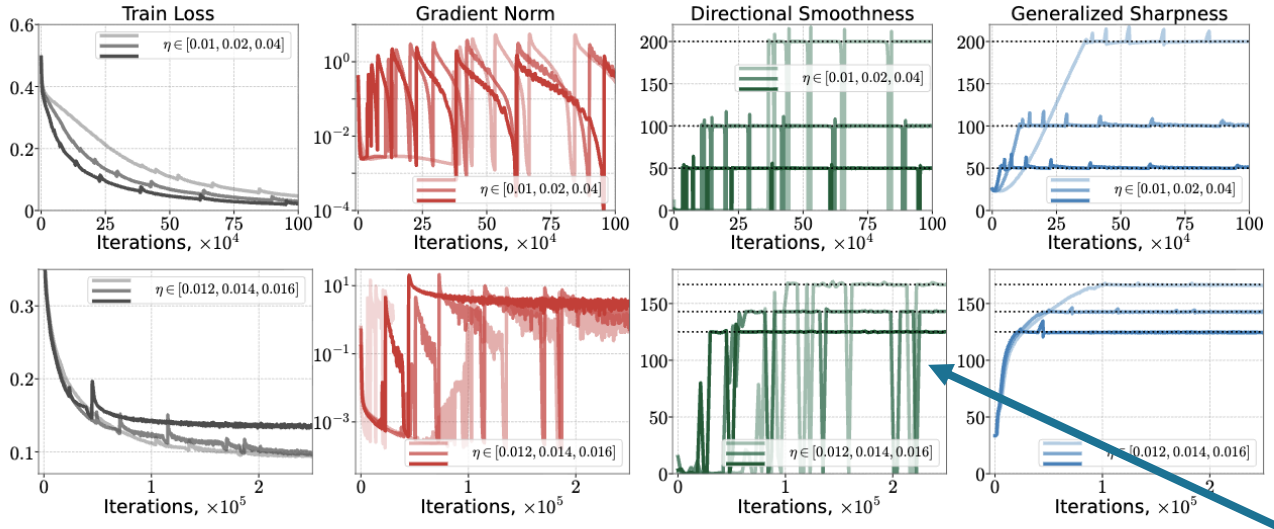


Figure 3: `Block CD` on CIFAR-10-5k. **Top**: MLP, **bottom**: CNN. Columns show train loss, gradient norm, directional smoothness, and generalized sharpness (16). Dashed lines mark the stability threshold $2/\eta$.

Drops down when another block gets updated which is not at EoS yet

Quality of FW Approximation

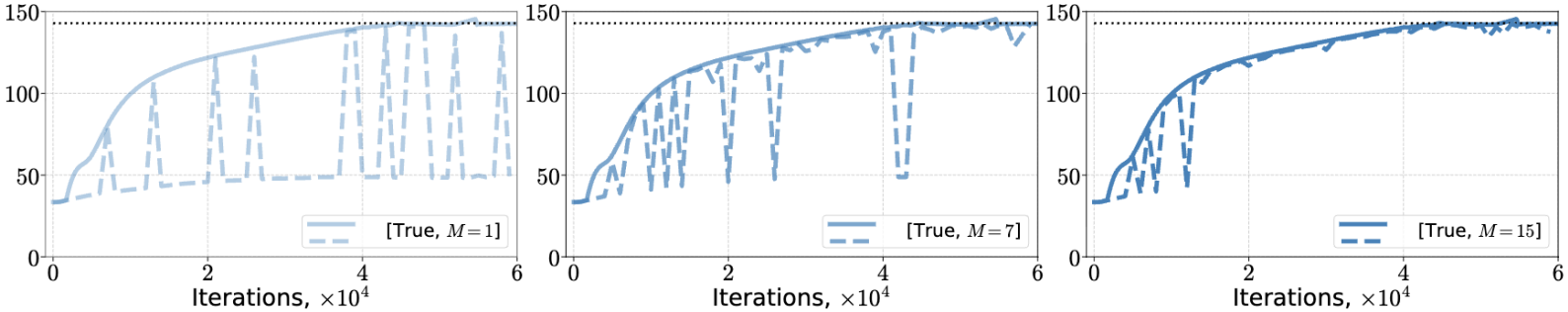


Figure F.1: The maximum block-wise Hessian eigenvalue (solid line), which is the generalized sharpness of `Block CD`, and its approximation by the Frank-Wolfe algorithm varying the number of initialization points in $\{1, 7, 15\}$ for the Frank-Wolfe algorithm. Here, M is the number of restarts of the Frank-Wolfe algorithm, varying the initialization point.

L2 Sharpness Does Not Capture EoS of Non-Euclidean GD

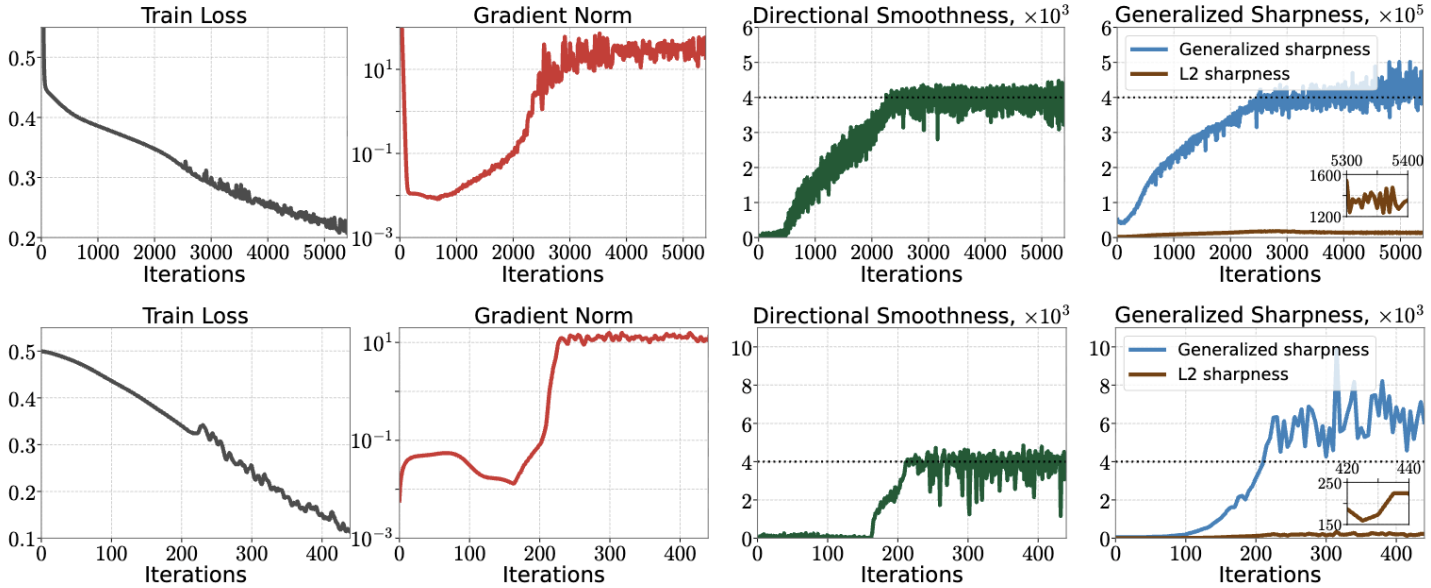


Figure G.5: (`Spectral GD`) Train loss, gradient norm, directional smoothness, generalized sharpness (19), and L2 sharpness ($\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{w}_t))$) during training ResNet20 (top, $\eta = 5 \cdot 10^{-5}$) and VGG11 (bottom, $\eta = 5 \cdot 10^{-4}$) on CIFAR-10 with `Spectral GD`. Horizontal dashed lines correspond to the value $2/\eta$.

SignGD is $\beta_2 \to \infty$ Limit of RMSprop

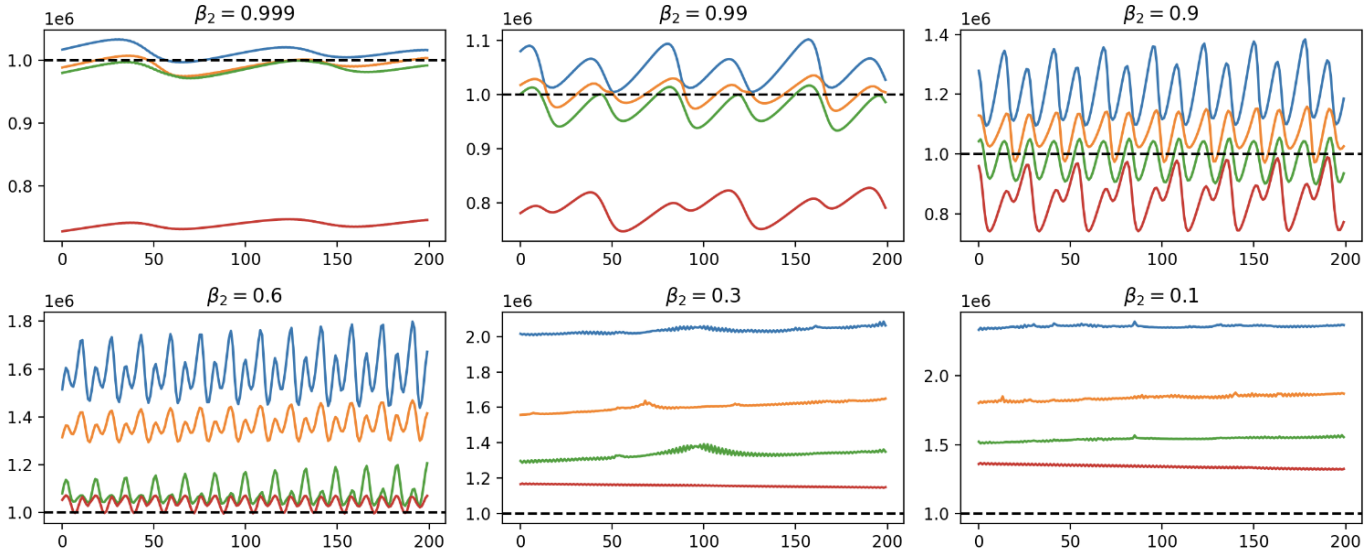


Figure H.1: Sharpness of `RMSprop` when training MLP model on a subset of CIFAR-10 dataset, varying β_2 hyperparameter. Here, colored lines correspond to the evolution of the top-4 largest eigenvalues of the preconditioned Hessian, while the dashed line is $2/\eta$ threshold. We observe that `RMSprop` reaches `AEoS` only for realistic (close to 1) values of β_2, while for small β_2 the preconditioned sharpness is not at $2/\eta$, but significantly higher.