Leveraging Non-uniformity in First-order Non-convex Optimization

Jincheng Mei[1,2,*], **Yue Gao**[1,*], Bo Dai[2],

Csaba Szepesvari[3,1], Dale Schuurmans[2,1]

[1]University of Alberta, [2]Google Brain, [3]DeepMind

* Equal contribution

UNIVERSITY OF ALBERTA

Main contributions

Two new properties:

non-uniform smoothness (NS), non-uniform Łojasiewicz (NŁ)

One new algorithm: geometry-aware normalized gradient descent (GNGD)

Two applications:

policy gradient optimization (PG) in reinforcement learning (RL),

generalized linear model (GLM) training in supervised learning (SL)

Non-uniform properties and algorithms

NS: $$\left| f(\theta') - f(\theta) - \left\langle \frac{df(\theta)}{d\theta}, \theta' - \theta \right\rangle \right| \leq \frac{\beta(\theta)}{2} \cdot \|\theta' - \theta\|_2^2$$ from β to $\beta(\theta)$

NŁ: $$\left\| \frac{df(\theta)}{d\theta} \right\|_2 \geq C(\theta) \cdot |f(\theta) - f(\theta^*)|^{1-\xi}$$ from C to $C(\theta)$

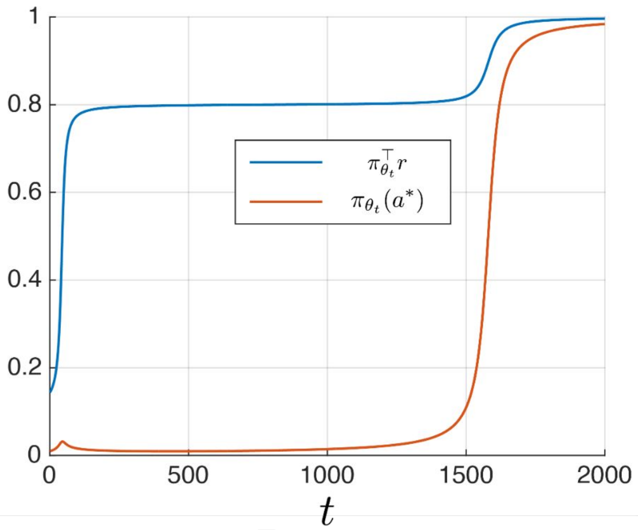
Geometry-aware normalized gradient descent (GNGD):

$$\theta_{t+1} \leftarrow \theta_t - \eta \cdot \frac{\nabla f(\theta_t)}{\beta(\theta_t)}$$

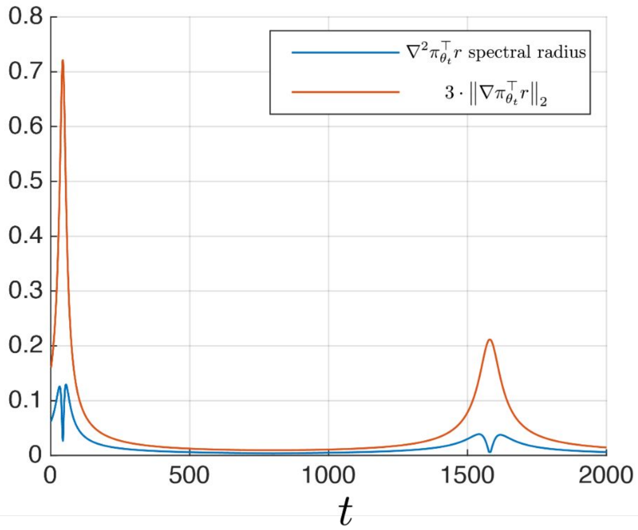
Example I: policy gradient optimization (PG)

Non-uniform smoothness (NS): standard PG

$$\left\| \frac{d^2 \pi_\theta^\top r}{d\theta^2} \right\|_2 \le 3 \cdot \left\| \frac{d \pi_\theta^\top r}{d\theta} \right\|_2$$



(a) $\pi_{\theta_t}^\top r$ and $\pi_{\theta_t}(a^*)$

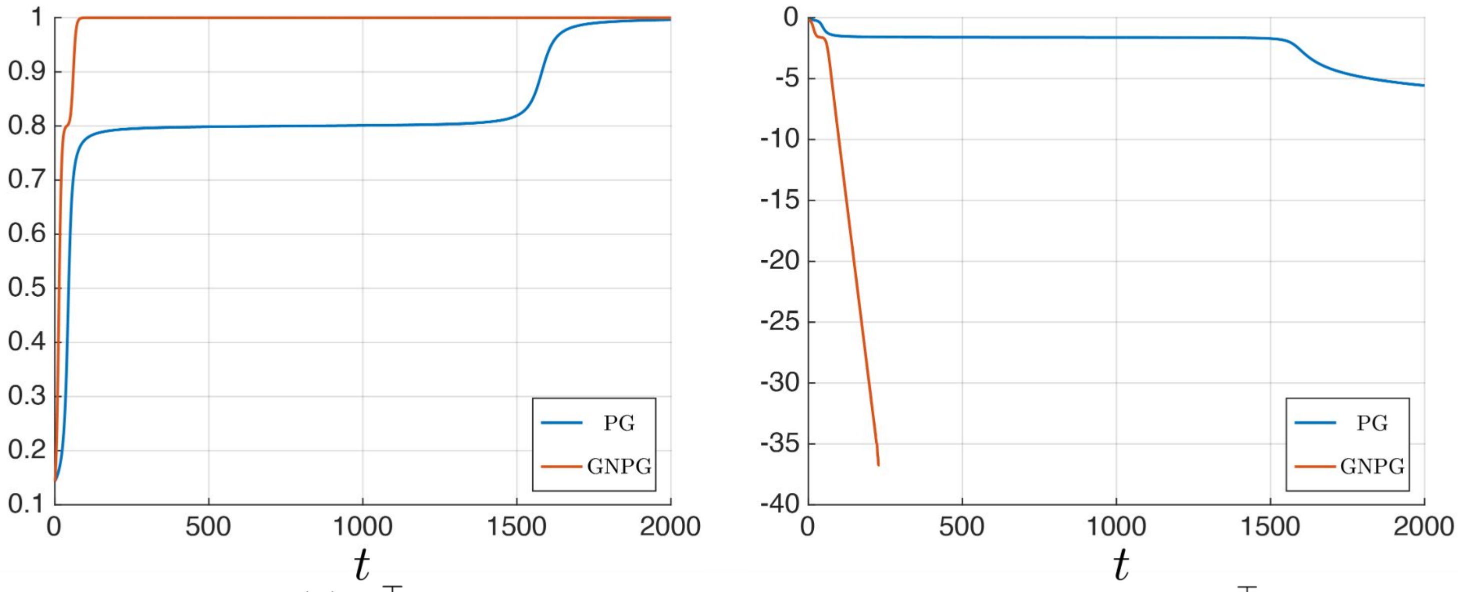


(b) Hessian spectral radius and PG norm

Key point: the smoothness of value function is non-uniform over parameters.

Geometry-aware normalized PG (GNPG)

Normalize the NS coefficient (PG norm): $$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2$$



(a) $\pi_{\theta_t}^\top r$ (b) $\log\left(\pi^* - \pi_{\theta_t}\right)^\top r$

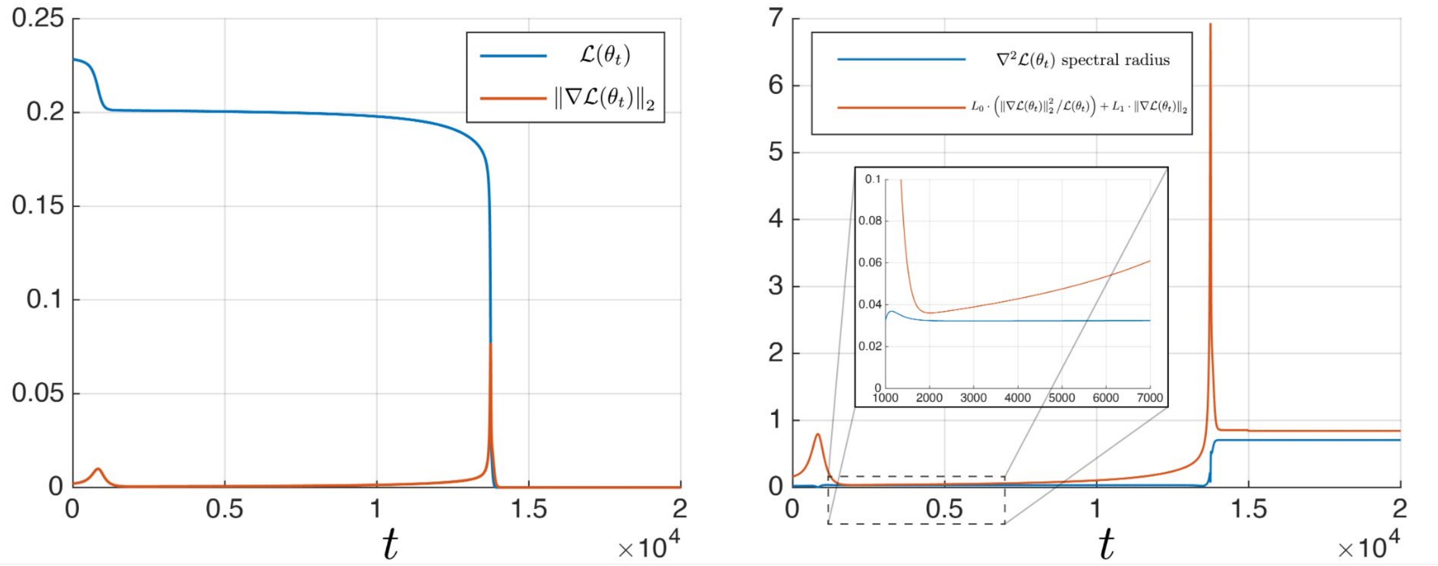
Faster rate:
from O(1/t) to
O(1/e^(c*t))

Faster escaping landscape plateau

Example II: generalized linear model (GLM)

NŁ and NS: standard gradient descent (GD)

$$\beta(\theta) = L_1 \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 + L_0 \cdot \left(\left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2^2 \Big/ \mathcal{L}(\theta) \right)$$



(a) $\mathcal{L}(\theta_t)$ and $\|\nabla \mathcal{L}(\theta_t)\|_2$ (b) Hessian spectral radius and NS

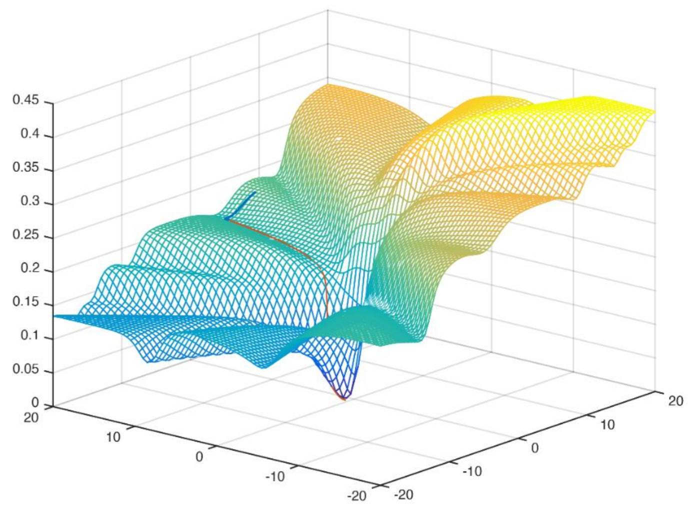
Key point:
satisfies NŁ

Key point:
different NS coefficient

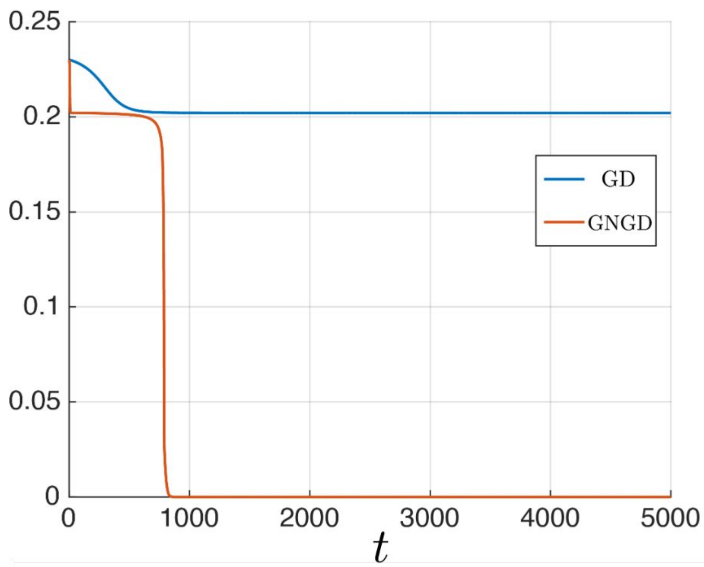
Geometry-aware normalized GD (GNGD)

Normalize the NS coefficient:

$$\theta_{t+1} \leftarrow \theta_t - \eta \cdot \frac{\nabla f(\theta_t)}{\beta(\theta_t)}$$



(a) MSE landscape in GLM



(b) Sub-optimality $\mathcal{L}(\theta_t)$

Faster rate: from O(1/e^(c^2 * t)) to O(1/e^(c * t))

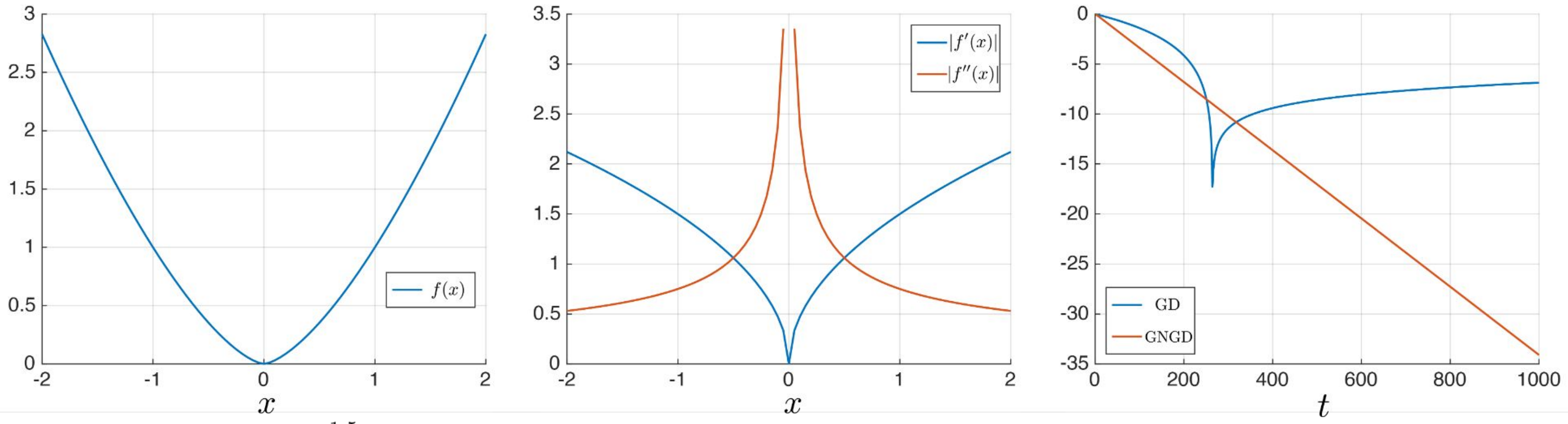
Faster escaping landscape plateau

A general non-uniform analysis

GNGD vs. GD in general optimization.

GNGD can be faster than Omega(1/t) lower bound of convex-smooth optimization.

GNGD converges when GD diverges.



(a) $f : x \mapsto |x|^{1.5}$ (b) gradient and Hessian (c) $\log \delta(x_t)$

Check our paper:

https://arxiv.org/abs/2105.06072

Thank you!