# Escaping saddle points without Lipschitz smoothness
## The power of nonlinear preconditioning

Alexander Bodard, Panagiotis Patrinos

KU Leuven & Leuven.AI

## Problem formulation

We study the *nonlinearly preconditioned gradient method* [3, 4]

$$x^{k+1} = T_{\gamma,\lambda}(x^k) := x^k - \gamma\nabla\phi^*(\lambda\nabla f(x^k)). \qquad \text{(P-GD)}$$

for minimizing a possibly nonconvex function $f \in \mathcal{C}^2$.

- If $\phi(x) = \frac{1}{2}\|x\|^2$, then (P-GD) reduces to vanilla gradient descent (GD).
- Focus: *isotropic* reference functions $\phi = h(\|\cdot\|)$ with kernel function $h : \mathbb{R} \to \overline{\mathbb{R}}$.

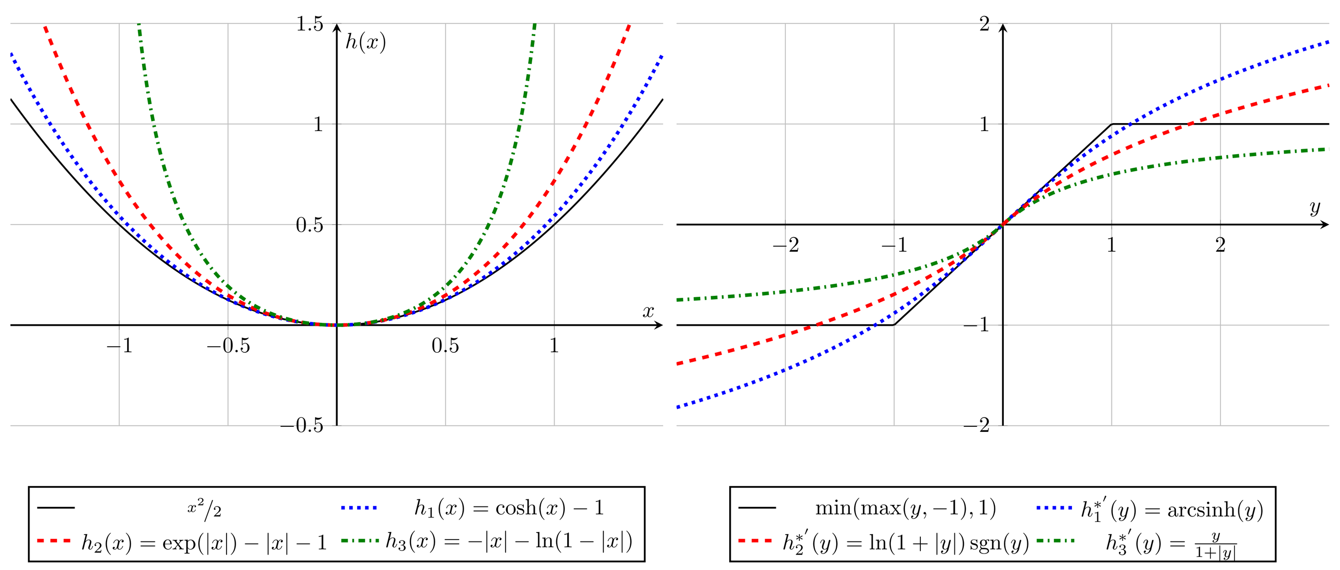


Figure 1: (a) kernel functions and (b) their corresponding nonlinear preconditioners.

- (P-GD) is naturally analyzed under *anisotropic smoothness*.
- *Gradient clipping* (cf. Fig 1(b)) often analyzed under $(L_0, L_1)$-smoothness.

### Research questions

(i) Can we formally establish anisotropic smoothness and $(L_0, L_1)$-smoothness of practical problems where traditional assumptions fail?

(ii) Does nonlinear preconditioning preserve the saddle point avoidance properties of GD under (milder) generalized smoothness assumptions?

Example: If $\phi = \cosh(\|\cdot\|) - 1$, and $\lambda = 1$, then (P-GD) becomes

$$x^{k+1} = x^k - \gamma \operatorname{arcsinh}(\|\nabla f(x^k)\|)\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}.$$

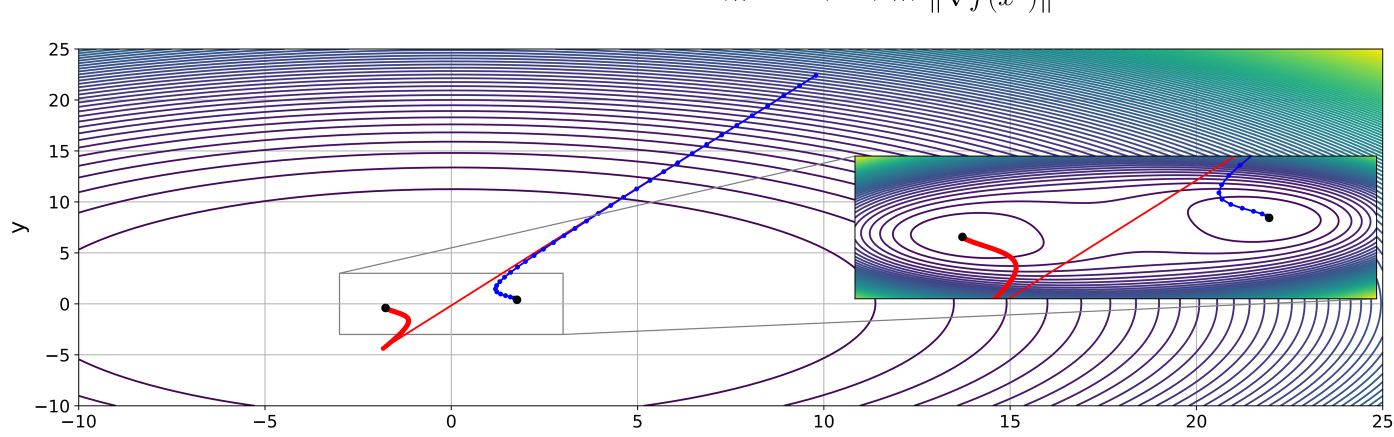


Figure 2: Iterates of GD (red) and (P-GD) (blue) on a symmetric matrix factorization problem. Because $f$ is a quartic polynomial, $\|\nabla f\|$ grows rapidly. For this reason, GD often requires a small $\gamma$, which results in tiny steps when close to a stationary point. In contrast, (P-GD) damps large gradients so that a larger $\gamma$ can be used, yielding larger steps around stationary points.

## Generalized smoothness

### Anisotropic smoothness [4]

We say that $f$ is anisotropically smooth [4] relative to $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ with constants $L, \bar{L} > 0$ if for all $x, \bar{x} \in \mathbb{R}^n$:

$$f(x) \leq f(\bar{x}) + \bar{L}L^{-1}\phi(L(x - T_{L^{-1},\bar{L}^{-1}}(\bar{x}))) - \bar{L}L^{-1}\phi(L(\bar{x} - T_{L^{-1},\bar{L}^{-1}}(\bar{x}))). \quad \text{(AS)}$$

- Under mild requirements, the condition [4, Propositions 2.6 & 2.9]
$$\nabla^2 f(x) \prec L\bar{L}[\nabla^2\phi^*(\bar{L}^{-1}\nabla f(x))]^{-1} \qquad \forall x \in \mathbb{R}^n, \qquad \text{(AS-SC)}$$
implies that (AS) holds with constants $\delta L, \bar{L}$ for any $\delta \in (0, 1)$.
- If $f \in \mathcal{C}^2$ is $(L_0, L_1)$-smooth, i.e.,
$$\|\nabla^2 f(x)\| \leq L_0 + L_1\|\nabla f(x)\| \qquad \forall x \in \mathbb{R}^n,$$
then (AS-SC) holds with $\phi(x) = -\|x\| - \ln(1 - \|x\|)$ and $(L, \bar{L}) = (L_1, L_0/L_1)$.
- Anisotropic smoothness is *more general* than $(L_0, L_1)$-smoothness, e.g.,
$$f(x) = \exp(\|x\|^2) - 2\|x\|^2.$$

### A novel sufficient condition

### A sufficient condition for anisotropic and $(L_0, L_1)$-smoothness

There exists an $R \in \mathbb{N}$ such that for all $x \in \mathbb{R}^n$

$$\|\nabla^2 f(x)\|_F \leq p_R(\|x\|), \quad \text{and} \quad \|\nabla f(x)\| \geq q_{R+1}(\|x\|),$$

where $p_R(\alpha) = \sum_{i=0}^{R} a_i\alpha^i$ and $q_{R+1}(\alpha) = \sum_{i=0}^{R+1} b_i\alpha^i$ are polynomials of degree $R$ and $R+1$, such that $b_{R+1} > 0$.

- A polynomial upper bound to $\|\nabla^2 f(x)\|_F$ was used for Bregman relative smoothness [2].
- Extra polynomial lower bound to $\|\nabla f(x)\|$ is crucial for $(L_0, L_1)$-smoothness.

### Key applications where Lipschitz smoothness fails

(i) *Phase retrieval* (assuming that measurements span $\mathbb{R}^n$)

(ii) *Symmetric matrix factorization (MF)*

(iii) *Regularized asymmetric MF*

(iv) *Burer-Monteiro factorization of MaxCut SDPs*

### Generalized smoothness holds for these applications

For the objective $f$ of Applications (i)-(iv), the following statements hold:

- For any $L_1 > 0$ there exists $L_0 > 0$ such that $f$ is $(L_0, L_1)$-smooth.
- Under mild requirements on $\phi$ and for any $\bar{L} > 0$, there exists an $L > 0$ such that $f$ satisfies (AS-SC) with constants $(L, \bar{L})$.

## Saddle point avoidance results

We generalize some classical saddle point avoidance results to hold under *anisotropic smoothness*, rather than Lipschitz smoothness.

### Asymptotic saddle avoidance

Let $\phi^* \in \mathcal{C}^2$, and $x^\star$ a strict saddle point of $f$. If (AS-SC) holds, then the iterates of (P-GD) with uniformly random initialization and $\gamma < \frac{1}{L}$ and $\lambda = \bar{L}^{-1}$ satisfy $\mathbb{P}\left(\lim_{k\to\infty} x^k = x^\star\right) = 0$.

- (P-GD) with $\phi = h(\|\cdot\|)$ and $h = \frac{1}{2}\|\cdot\|^2 + \delta_{[-1,1]}$ reduces to *gradient clipping*
$$x^{k+1} = T_{\gamma,\bar{L}^{-1}}(x^k) = x^k - \gamma\min(1/\|\nabla f(x^k)\|, \bar{L}^{-1})\nabla f(x^k).$$
- Although $\phi^* \notin \mathcal{C}^2$, a similar result can be established.

### Efficient saddle avoidance of perturbed (P-GD)

For any $\epsilon > 0$ sufficiently small, and for any $\delta \in (0, 1)$, a *perturbed* (P-GD) variant visits an $\epsilon$-second-order stationary point after at most $T/2$ iterations with probability at least $1 - \delta$, where

$$T = \tilde{O}\left(\frac{L(f(x^0) - \inf f)}{L\epsilon^2}\right).$$

- Efficient saddle point avoidance of GD is preserved by (P-GD).
- Theory holds under *milder* smoothness conditions.

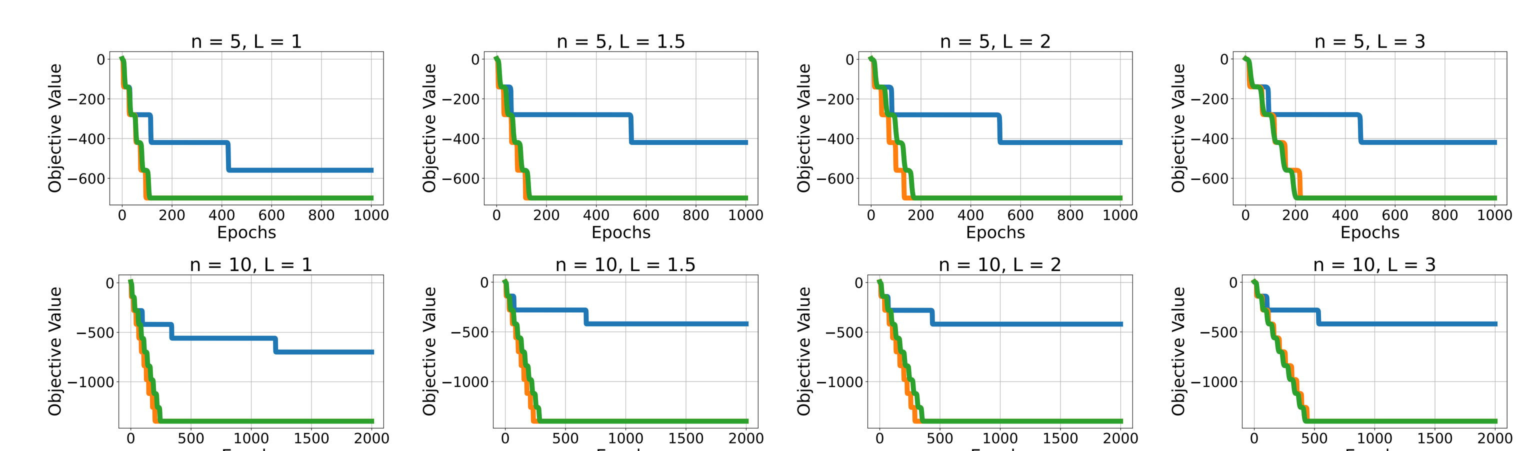


Figure 3: Performance of vanilla GD (blue), perturbed vanilla GD [1, Alg 1] (orange), and perturbed (P-GD) (green) on the 'octopus' function [1].

## References


[1] Simon S Du et al. "Gradient Descent Can Take Exponential Time to Escape Saddle Points". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[2] Haihao Lu et al. "Relatively Smooth Convex Optimization by First-Order Methods, and Applications". In: *SIAM Journal on Optimization* 28.1 (Jan. 2018), pp. 333–354.

[3] Chris J. Maddison et al. "Dual Space Preconditioning for Gradient Descent". In: *SIAM Journal on Optimization* 31.1 (Jan. 2021), pp. 991–1016.

[4] Konstantinos Oikonomidis et al. "Nonlinearly Preconditioned Gradient Methods under Generalized Smoothness". In: *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, July 2025, pp. 47132–47154.