On Lipschitzness, monotonicity, and cocoercivity

Zach Stoebner
Electrical & Computer Engineering, UT Austin
zstoebner@austin.utexas.edu

February 26, 2026

Abstract

In iterative optimization, an algorithm's behavior often depends on properties of the operators involved in computing the iterates. Take gradient descent for example: its convergence rate is bounded by the Lipschitzness of the objective's gradient, as shown below. In addition to Lipschitzness, the properties of monotonocity and cocoercivity are also involved in convergence behavior, but their relationship to each other and what they intuitively mean about an operator is typically overlooked and complicated in the literature. To develop pithy intuition, this note explores these three operator properties, their relationships to each other, and what they mean regarding operator geometry. The key takeaways: Lipschitzness prevents explosion, monotonicity prevents pointing backwards, and cocoercivity prevents destabilization.

Notation Unless denoted otherwise, $||\cdot|| = ||\cdot||_2$. $\mathbb{F}$ denotes the field of numbers; the results hold for $\mathbb{R}$ and $\mathbb{C}$. Regular lowercase letters, e.g., x, denote scalars. Bolded lowercase letters, e.g., $\mathbf{x}$, denote vectors. Uppercase regular letters, e.g., X, denote matrices. Caligraphic uppercase letters, e.g., $\mathcal{X}$, denote tensors.

1 Lipschitzness

An operator F is L-Lipschitz continuous if it satisfies the property:

$$||F(x) - F(y)|| \leq L[F]||x - y|| \quad \forall x, y \in dom\{F\}$$

For a differentiable convex function $f : \mathcal{D}_f \to \mathbb{F}$ with an L-Lipschitz gradient (aka f is L-smooth), the function has the first-order property:

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}||x - y||^2$$

Note that these inequalities hold for scalar- and vector-valued functions where $\nabla f(x) = J_f(x)$ in the vector-valued case. In the sequel, I will use f to denote a scalar-valued function and F to denote a vector-valued one – the facts generalize between the two.

1.1 Convergence behavior of gradient descent

WLOG assume a scalar-valued $f \in \mathcal{F}_L^1$ where $\mathcal{F}_L^1$ denotes the space of convex functions with L-smooth 1st-order derivatives, i.e., $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$. Given the gradient descent step $x_{k+1} = x_k - \eta \nabla f(x_k)$ and using the property $\frac{1}{L}||\nabla f(x) - \nabla f(y)||^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle$ (wrt to Euclidean norm), the convergence is given by:

$$\begin{aligned}
||x_{k+1} - x^*||^2 &= ||x_k - \eta \nabla f(x_k) - (x^* - \eta \nabla f(x^*))||^2 \\
&= ||x_k - x^*||^2 + \eta^2 ||\nabla f(x_k) - \nabla f(x^*)||^2 - 2\eta \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle \\
&\leq (1 + \eta^2 L^2)||x_k - x^*||^2 - 2\eta \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle \\
&\leq (1 + \eta^2 L^2 - \frac{2\eta}{L})||x_k - x^*||^2.
\end{aligned}$$

For uniform convergence (aka contraction, see below), $1 + \eta^2 L^2 - \frac{2\eta}{L} < 1 \implies \eta < \frac{1}{L}$. The proof for $\frac{1}{L}||\nabla f(x) - \nabla f(y)||_*^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle$ can be found in [[1], Thm 2.1.5] [1]. The sketch is:

[1] This property of convex functions with Lipschitz continuous gradients is a soft-intro to cocoercivity, i.e., the gradient is $\frac{1}{L}$-cocoercive.

1. Consider $\phi(y) = f(y) - \langle \nabla f(x), y \rangle \implies \nabla\phi(y) = \nabla f(y) - \nabla f(x), y^* = x$.
2. $\phi(y^*) = \min_z \phi(z) \leq \min_z \phi(y) + \langle \nabla\phi(y), z - y \rangle + \frac{L}{2}||z - y||^2$.
3. Using the general Cauchy-Schwarz inequality with the dual norm: $\langle \nabla\phi(y)z - y \rangle \leq ||\nabla\phi(y)||_*||z - y||$.
4. For $r = ||z - y|| \geq 0$, the min of the rhs in (2) equals $\min_{r\geq 0} \phi(y) - r||\nabla\phi(y)||_* + \frac{L}{2}r^2$.
5. Differentiate and solve for r^* and substitute to obtain $\frac{1}{2L}||\nabla f(y) - \nabla f(x)||_* \leq f(y) - f(x) + \langle \nabla f(x), y - x \rangle$.
6. The result in (5) also applies for $f(x) - f(y) + \langle \nabla f(y), x - y \rangle$ so summing together yields the desired result.

1.2 Intuition

With all that example out of the way, what does Lipschitzness actually tell us? Yes, it can indicate how quickly an iteration will converge but what does it tell us about the geometry of the function? Rearranging the inequality:

$$\frac{||F(x) - F(y)||}{||x - y||} \leq L[F]$$

which should look familiar. Essentially, Lipschitzness bounds the magnitude of the slope of operator F, which is very useful intuition. To build on that, we can visualize the effect of decreasing $L[F]$ for $F(x) = \cos(x)$. Using the mean-value theorem (MVT), we know that for continuous differentiable F on $[a, b]$ with $a \leq c \leq b$:

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Therefore:

$$|\cos(x) - \cos(y)| = |(-\sin(c))|\,|(x - y)| \leq |x - y|$$

where $L[\cos]$ scales with the amplitude. It follows that $\lim_{A\to 0} L[A\cos] = 0$, i.e., the function converges to 0 everywhere aka it has no variance.

Another common classification pertaining to a Lipschitz continuous operator when it comes to iterative methods is whether it's contractive, expansive, or non-expansive. Typically, we want the error to reduce between subsequent iterates. Besides GD, others iterations will have convergence bounds akin to $||x_k - x^*|| \leq L^k||x_0 - x^*||$ so, for uniform convergence, we want contraction of the errors. However, as in the sinusoidal example, a Lipschitz operator that is too contractive reduces to a line. For a black-box methods, enforcing this restricts the expressivity of the function because there is little to no variation between disparate inputs. **Takeaway: Lipschitzness prevents explosion.**

1.3 Significance of a symmetric Jacobian

Geometrically, the Jacobian $J_F(x)$ is the best linear approximation of F near x. In fact, there is an MVT for vector-valued functions. Let $c \in [0, 1]$

$$F(x) - F(y) = J_F((1 - c)x + cy) \cdot (x - y).$$

where the RHS is a matvec product. Taking the norm of both sides and applying Cauchy-Schwarz on the right gives us the familiar inequality:

$$||F(x) - F(y)|| \leq ||J_F((1 - c)x + cy)||\,||x - y|| \leq (\max_x |\sigma_{\max}(J_F(x))|)\,||x - y||$$

where $\sigma_{\max}(A) = ||A||$ is the max singular value of matrix A. If $J_F \in S^n$, then $L[F] = ||J_F|| = \max_x |\lambda_{\max}(J_F)|$.

Suppose $F = \nabla f$ for some scalar-valued $f : \mathbb{F}^n \to \mathbb{F}$. Then, $J_F = \nabla^2 f$ is the Hessian of f and if $J_F \in S^n$ then f is a potential and F is a conservative vector field, which introduces many nice properties for solving all sorts of problems, e.g., denoising [2].

2 Monotonicity

An operator F is m-monotone ($m \geq 0$) if it satisfies the property:

$$\langle F(x) - F(y), x - y \rangle \geq m||x - y||^2$$

As the name might suggest, this property relates to monotonicity in real analysis and linear functional analysis. Here are those definitions to draw intuition [3]:

1. A function $f : D \to \mathbb{R}$ is increasing $\iff \forall x, y \in D \quad x < y \implies f(x) \leq f(y)$, where D is a nonempty convex subset of the Hilbert space H. f is monotonically increasing when the inequality is strict.
2. A bounded linear operator $F : H \to H$ is monotone $\iff \forall x \in H \quad \langle x, Fx \rangle \geq 0$.

2.1 Intuition

To gain intuition, we can reverse-engineer the monotone property by one step:

$$F(x) - F(y) \geq m(x - y) \implies \langle F(x) - F(y), x - y \rangle \geq m\langle x - y, x - y \rangle$$

where the consequent is the monotone property since $||x||^2 = \langle x, x \rangle$. Looking at the antecedent, we can say $\frac{||F(x)-F(y)||}{||x-y||} \geq m$, which lower bounds the magnitude of the "slope" of the operator. For a convex function f that is m-strongly convex and L-smooth, this comes nicely into play with Lipschitzness to bound the first-order property for convexity:

$$\frac{m}{2}||y - x||^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}||y - x||^2$$

giving us a tighter band for the first-order approximation of f. From a differential view, monotonicity lower bounds variation positively s.t. steps with the operator will not point backwards ,i.e., the negative direction. **Takeaway: monotonicity prevents pointing backwards.**

2.2 Significance of a symmetric Hessian

Following the scenario in 1.3 where we have $F = \nabla f$ for some $f : \mathbb{F}^n \to \mathbb{F}$, if F is monotone then $||J_F|| \geq m$. If $J_F \in S^n$, then $J_F \succeq mI$, meaning that J_F is positive semi-definite (PSD) for $m \geq 0$ and, hence, f is a convex potential. This implies that f has positive curvature which offers up a host of nice properties, from finding unique fixed points to accelerating iterative algorithms [4].

3 Cocoercivity

An operator A is β-cocoercive ($\beta > 0$) if it satisfies:

$$\langle A(x) - A(y), x - y \rangle \geq \beta||A(x) - A(y)||^2 \quad \forall x, y \in \mathbb{F}^n.$$

Facts:

- β-cocoercive $\implies$ $\frac{1}{\beta}$-Lipschitz, monotone.
  Proof: For Lipschitzness, Cauchy-Schwarz on LHS above. For monotone, $\beta||A(x) - A(y)||^2 \geq 0$.
- $f \in \mathcal{F}_\tau^1 \implies f$ is $\frac{1}{\tau}$-cocoercive.
  Proof: Baillon-Haddad Theorem [5]. Also [[1], Thm 2.1.5].
- A is L-Lipschitz, m-monotone ($m > 0$) $\implies$ A is $\frac{m}{L^2}$-cocoercive.
  Proof: $\langle A(x) - A(y), x - y \rangle \geq m||x - y||^2 \geq \frac{m}{L^2}||A(x) - A(y)||^2$.

3.1 Intuition

Paraphrasing and expanding on the facts above:

1. "If an operator is cocoercive, it is immediately Lipschitz and monotone."
   If we take the same scenario as 1.3 where $F = \nabla f$ and $f : \mathbb{F}^n \to \mathbb{F}$ and F is $\frac{1}{L}$-cocoercive, then this implies that $0 \leq ||J_F|| \leq L$. In a dynamical systems sense, this nice bound on variation implies a stable system that will converge under mild conditions.

2. "If a function is convex with a Lipschitz gradient, then the function is cocoercive."
   This one is self-explanatory. If f is convex, it implies that its gradient is monotone since $J_F \succeq mI$. However, this property also holds for $m = 0$ whereas the third fact only holds for $m > 0$.

3. "If an operator is Lipschitz and strongly monotone, it's cocoercive."
   This fact is particularly useful in applied settings. Unlike the second fact, the operator need not be convex or have a symmetric Jacobian, but if we can enforce Lipschitzness and monotonicity, it gains many nice benefits enjoyed by convexity, such as stability and guaranteed convergence. A great practical example where cocoercivity plays a central role in an accelerated algorithm is variable-metric forward-backward splitting [6].

Takeaway: cocoercivity prevents destabilization.

References

[1] Y. Nesterov *et al.*, *Lectures on convex optimization*. Springer, 2018, vol. 137.

[2] P. Milanfar and M. Delbracio, "Denoising: a powerful building block for imaging, inverse problems and machine learning," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 383, no. 2299, 2025.

[3] P. L. Combettes, "Monotone operator theory in convex optimization," *Mathematical Programming*, vol. 170, no. 1, pp. 177–206, 2018.

[4] E. K. Ryu and S. Boyd, "Primer on monotone operator methods," *Appl. comput. math*, vol. 15, no. 1, pp. 3–43, 2016.

[5] D. Wachsmuth and G. Wachsmuth, "A simple proof of the baillon-haddad theorem on open subsets of hilbert spaces," *arXiv preprint arXiv:2204.00282*, 2022.

[6] E. K. Ryu and W. Yin, *Large-scale convex optimization: algorithms & analyses via monotone operators*. Cambridge University Press, 2022.