Convergence of Fixed-Point Iterations

Chenglong Bao

Acknowledgement: this material is from Prof. Wotao Yin(UCLA)'s slide.

Why study fixed-point iterations?

- Abstract many existing algorithms in optimization, numerical linear algebra, and differential equations
- Often require only minimal conditions
- Simplify complicated convergence proofs

Google Scholar

Search Scholar

- English
  - Business, Economics & Management
  - Chemical & Material Sciences
  - Engineering & Computer Science
  - Health & Medical Sciences
  - Humanities, Literature & Arts
  - Life Sciences & Earth Sciences
  - Physics & Mathematics
    - Mathematical Optimization
  - Social Sciences
- Chinese
- Portuguese
- Spanish
- German

Top publications - Mathematical Optimization

Learn more

	Publication	h5-index	h5-median
1.	arXiv Optimization and Control (math.OC)	66	102
2.	Mathematical Programming	50	78
3.	SIAM Journal on Optimization	43	63
4.	Fixed Point Theory and Applications	42	60
5.	SIAM Journal on Control and Optimization	36	52
6.	Structural and Multidisciplinary Optimization	35	53
7.	Journal of Optimization Theory and Applications	32	45
8.	Mathematics of Operations Research	30	45
9.	Computational Optimization and Applications	29	42
10.	Journal of Global Optimization	29	39
11.	Engineering Optimization	25	32
12.	Optimization Letters	25	32
13.	ESAIM: Control, Optimisation and Calculus of Variations	24	33
14.	Optimization Methods and Software	23	29

Notation

- **space:** Hilbert space $\mathcal{H}$ equipped with $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$
- Fine to think in $\mathbb{R}^2$ (though not always)
- An *operator* $T : \mathcal{H} \to \mathcal{H}$ (or $C \to C$ where C is closed subset of $\mathcal{H}$)
- **our focus:**
  - when $\mathrm{Fix}T := \{x \in \mathcal{H} : x = T(x)\}$ is nonempty
  - the convergence of $x^{k+1} \leftarrow T(x^k)$
- **simplification:** $T(x)$ is often written as Tx

Examples

unconstrained C^1 minimization:

$$\text{minimize}\, f(x)$$

- x^* is a **stationary point** if $\nabla f(x^*) = 0$
- **gradient descent operator:** for $\gamma > 0$

$$T := I - \gamma \nabla f$$

- the gradient descent iteration

$$x^{k+1} \leftarrow T x^k$$

- **lemma:** x^* is a stationary point if, and only if, $x^* \in \mathrm{Fix} T$

Examples

constrained C^1 minimization:

$$\text{minimize}\, f(x) \quad \text{subject to } x \in C$$

- **assume:** f is proper closed convex, C is nonempty closed convex
- **projected-gradient operator:** for $\gamma > 0$

$$T := \mathbf{proj}_C(I - \gamma \nabla f)$$

- $x^{k+1} \leftarrow Tx^k$ is the projected-gradient iteration

$$x^{k+1} \leftarrow \mathbf{proj}_C\big(x^k - \gamma \nabla f(x^k)\big)$$

- x^* is optimal if

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0 \quad \forall x \in C$$

- **lemma:** x^* is optimal if, and only if, $x^* \in \mathrm{Fix} T$

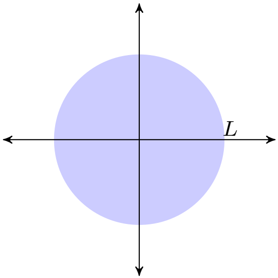
Lipschitz operator

- **definition:** an operator T is L-Lipschitz, $L \in [0, \infty)$, if

$$\|Tx - Ty\| \leq L\|x - y\|, \quad \forall x, y \in \mathcal{H}$$

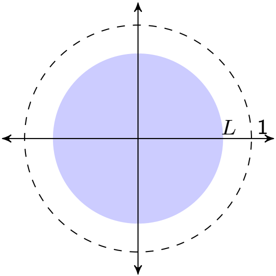
- **definition:** an operator T is L-**quasi**-Lipschitz, $L \in [0, \infty)$, if for any $x^* \in \mathrm{Fix} T$ (assumed to exist),

$$\|Tx - x^*\| \leq L\|x - x^*\|, \quad \forall x \in \mathcal{H}$$



Contractive operator

- **definition:** T is contractive if it is L-Lipschitz for $L \in [0, 1)$
- **definition:** T is **quasi**-contractive if it is L-**quasi**-Lipschitz for $L \in [0, 1)$



Banach fixed-point theorem

- **Theorem:** If T is contractive, then
  - T admits a unique fixed-point x^* (existence and uniqueness)
  - $x^k \to x^*$ (convergence)
  - $\|x^k - x^*\| \leq L^k \|x^0 - x^*\|$ (speed)
- Holds in a Banach space
- Also known as the Picard-Lindelöf Theorem

Examples

minimize a Lipschitz-differentiable strongly-convex function:

$$\text{minimize}\, f(x)$$

- **definition:** a convex f is L-Lipschitz-differentiable if

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq L\langle x - y, \nabla f(x) - \nabla f(y)\rangle \quad \forall x, y \in \mathrm{dom} f$$

- **definition:** a convex f is μ-strongly convex if, element wise,

$$\langle \partial f(x) - \partial f(y), x - y\rangle \geq \mu\|x - y\|^2 \quad \forall x, y \in \mathrm{dom} f$$

- **lemma:** Gradient descent operator $T := I - \gamma\nabla f$ is C-contractive for all γ in a certain interval.
  exercise: find the interval of γ and the formula of C in γ, L, μ

- Also true for a projected-gradient operator if C is closed convex and $C \cap \mathrm{dom} f \neq \emptyset$

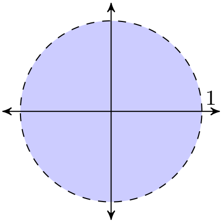
Nonexpansive operator

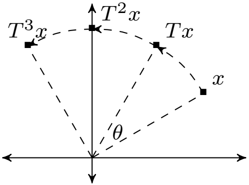
- **definition**: an operator is nonexpansive if it is 1-Lipschitz, i.e.,

$$\|Tx - Ty\| \le \|x - y\|, \quad \forall x, y \in \mathcal{H}$$

- **properties**:
  - T may not have a fixed point x^*
  - if x^* exists, $x^{k+1} = Tx^k$ is bounded
  - may diverge

- **examples:** rotation, alt. reflection





Between $L = 1$ and $L < 1$

- $L < 1$: linear (or geometric) convergence
- $L = 1$: bounded, may diverge
- A vast set of algorithms (often with sublinear convergence) **cannot** be characterized by L
  - Alternative projection (von Neumann)
  - Gradient descent without strong convexity
  - Proximal-point algorithm without strong convexity
  - Operator splitting algorithms

Averaged operator

- **fixed-point residual operator:** $R := I - T$
- $Rx^* = 0 \;\Leftrightarrow\; x^* = Tx^*$
- **averaged operator:** from some $\eta > 0$,

$$\|Tx - Ty\|^2 \leq \|x - y\|^2 - \eta\|Rx - Ry\|^2, \quad \forall x, y \in \mathcal{H}.$$

- **quasi-averaged operator:** from some $\eta > 0$,

$$\|Tx - x^*\|^2 \leq \|x - x^*\|^2 - \eta\|Rx\|^2, \quad \forall x \in \mathcal{H}.$$

- **interpretation:** improve by the amount of fixed-point violation
- **speed:** may become slower as x^k gets closer to the minimizer

- **convention:** use α instead of η following

$$\eta := \frac{1-\alpha}{\alpha}$$

- $\eta > 0 \;\Leftrightarrow\; \alpha \in (0,1)$

- α-**averaged operator:** from some $\eta > 0$,

$$\|Tx - Ty\|^2 \leq \|x - y\|^2 - \frac{1-\alpha}{\alpha}\|Rx - Ry\|^2, \quad \forall x, y \in \mathcal{H}$$

- **special case:**
    - $\alpha = \frac{1}{2}$: T is called *firmly nonexpansive*
    - $\alpha = 1$ (violating $\alpha \in (0,1)$): T is called *nonexpansive*

Why called "averaged"?

Lemma

T is α-averaged if, and only if, there exists a nonexpansive map T' so that

$$T = (1-\alpha)I + \alpha T'.$$

or equivalently,

$$T' := \big((1-\frac{1}{\alpha})I + \frac{1}{\alpha}T\big)$$

is nonexpansive.

Proof. From $T' := (I - \frac{1}{\alpha})I + \frac{1}{\alpha}T = I - \frac{1}{\alpha}R$, basic algebraic manipulation gives us: for any x and y,

$$\alpha(\|x-y\|^2 - \|T'x - T'y\|^2) = \|x-y\|^2 - \|Tx - Ty\|^2 - \frac{1-\alpha}{\alpha}\|Rx - Ry\|^2.$$

Therefore, T' is nonexpansive $\Leftrightarrow$ T is α-averaged. $\square$

Properties

- **assume:**
  - T is α-averaged
  - T has a fixed point x^*
- **iteration:** $x^{k+1} \leftarrow Tx^k$
- **claims about the iteration:** step-by-step,

  (a) $\|x^{k+1} - x^*\|^2 \le \|x^k - x^*\|^2 - \frac{1-\alpha}{\alpha}\|Rx^k - \underbrace{Rx^*}_{=0}\|^2$

  (b) by telescopic sum on (a),
  $\|x^{k+1} - x^*\|^2 \le \|x^0 - x^*\|^2 - \frac{1-\alpha}{\alpha}\sum_{j=0}^k \|Rx^j\|^2$.

  (c) $\{\|Rx^k\|^2\}$ is summable and $\|Rx^k\| \to 0$

- **claims (cont.):**
  (d) by (a), $\{\|x^k - x^*\|^2\}$ is monotonically decreasing until $x^k \in \mathrm{Fix}T$

  (e) by (d), $\lim_k \|x^k - x^*\|^2$ exists (but not necessarily zero)

  (f) by (d), $\{x^k\}$ is bounded and thus has a weak cluster point $\bar{x}$
  (note: $\mathcal{H}$ is weakly sequentially closed)

  Next: we will show that $\bar{x} \in \mathrm{Fix}T$ and then $x^k \rightharpoonup \bar{x}$.

- **claims (cont.):**

  (h) **demiclosedness principle:** Let T be nonexpansive and $R := I - T$. If $x^j \rightharpoonup x'$ and $\lim \|Rx^j\| = 0$, then $Rx' = 0$.

  Proof. Goal is to expand $\|Rx'\|^2$ into convergent terms as $j \to \infty$.

$$\begin{aligned}\|Rx'\|^2 &= \|Rx^j\|^2 + 2\langle Rx^j, Tx^j - Tx'\rangle + \|Tx^j - Tx'\|^2 \\ &\quad - \|x^j - x'\|^2 - 2\langle Rx', x^j - x'\rangle \\ &\le \|Rx^j\|^2 + 2\langle Rx^j, Tx^j - Tx'\rangle - 2\langle Rx', x^j - x'\rangle.\end{aligned}$$

  Each term on the RHS $\to 0$ as $j \to \infty$. Therefore, $\|Rx'\|^2 = 0$. $\square$

  (i) by applying (h) to any converging subsequence, each cluster point $\bar{x}$ of $\{x^k\}$ is a fixed point.

- **claims (cont.):**
  (j) By (e) and (i), $\bar{x}$ is the **unique cluster point**.

  Proof. Let $\bar{y}$ also be a cluster point.

  - $\bar{y} \in \mathrm{Fix}T$, just like $\bar{x}$.
  - by (e), both $\lim_k \|x^k - \bar{x}\|^2$ and $\lim_k \|x^k - \bar{y}\|^2$ exist.
  - algebraically,

$$2\langle x^k, \bar{x} - \bar{y} \rangle = \|x^k - \bar{x}\|^2 - \|x^k - \bar{y}\|^2 + \|\bar{x}\|^2 - \|\bar{y}\|^2,$$

  whose RHS converges to a constant, say C.

  - passing the limits of the two subsequence, to $\bar{x}$ and to $\bar{y}$,

$$2\langle \bar{x}, \bar{x} - \bar{y} \rangle = 2\langle \bar{y}, \bar{x} - \bar{y} \rangle = c.$$

  - hence, $\|\bar{x} - \bar{y}\|^2 = 0$. □

Theorem (Krasnosel'skiĭ)

Let T be an averaged operator with a fixed point. Then, the iteration

$$x^{k+1} \leftarrow Tx^k$$

converges weakly to a fixed point of T.

Mann's version

- Let T be a nonexpansive operator with a fixed point. Then, the iteration

$$\boxed{x^{k+1} \leftarrow (1-\lambda_k)x^k + \lambda_k T x^k}$$

  (known as the KM iteration) converges weakly to a fixed point of T as long as

$$\lambda_k > 0, \quad \sum_k \lambda_k(1-\lambda_k) = \infty.$$

- The λ_k condition is ensured if

$$\lambda_k \in [\epsilon, 1-\epsilon]$$

  (bounded away from 0 and 1)

Remarks

- Can be relaxed to quasi-averagedness
- Summable errors can be added to the iteration
- In finite dimension, demiclosedness principle is not needed
- This fundamental result is largely ignore, yet often reproved in $\mathbb{R}^n$
- Browder-Göhde-Kirk fixed-point theorem: If T has no fixed point and λ_k is bounded away from 0 and 1, the sequence $\{x^k\}$ is unbounded.
- Speed: $\|Rx^k\|^2 = o(1/k)$, no rate for $x^k \rightharpoonup x^*$
- Much more applications than Banach's fixed-point theorem

Special cases

proximal-point algorithm

- **problem:**

$$\operatorname{minimize} f(x)$$

- **proximal operator:** let $\lambda > 0$,

$$T := \mathbf{prox}_{\lambda f}$$

- Since T **is firmly-nonexpansive**,

$$x^{k+1} \leftarrow \mathbf{prox}_{\lambda f}(x^k)$$

  converges weakly to a minimizer of f, if it exists

Special cases

gradient descent:

- Define the **gradient-descent operator:**

$$T := I - \lambda \nabla f$$

- **iteration:**

$$x^{k+1} \leftarrow Tx^k = x^k - \gamma \nabla f(x^k)$$

- **Baillion-Haddad theorem:** if f is convex and ∇f is L-Lipschitz, then

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

- If f has a minimizer x^*, then

$$\frac{2}{L\gamma}\|\gamma \nabla f(x^k)\|^2 \leq 2\langle x^k - x^*, \gamma \nabla f(x^k)\rangle$$

- Directly expand $\|x^{k+1} - x^*\|^2$:

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

  Therefore, T is quasi-averaged if

$$\lambda \in \Big(0, \frac{2}{L}\Big).$$

- In fact, it is easy to show that T is averaged.

- The convergence result applies to gradient descent.

Composition of operators

- If $T_1, \ldots, T_m : \mathcal{H} \to \mathcal{H}$ are nonexpansive, then $T_1 \circ \cdots \circ T_m$ is nonexpansive.
- If $T_1, \ldots, T_m : \mathcal{H} \to \mathcal{H}$ are averaged, then $T_1 \circ \cdots \circ T_m$ is averaged.
- The averagedness constants get worse: let T_i be α_i-averaged (allowing $\alpha_i = 1$), then $T = T_1 \circ \cdots \circ T_m$ is α-averaged where

$$\alpha = \frac{m}{m - 1 + \frac{1}{\max_i \alpha_i}}$$

- In addition, if any T_i is contractive, $T_1 \circ \cdots \circ T_m$ is contractive.

Special cases

projected-gradient method:

- **convex problem:**

$$\underset{x}{\text{minimize}}\, f(x) \quad \text{subject to } x \in C.$$

- assume sufficient intersection between $\mathrm{dom} f$ and C
- **define:**

$$T := \mathbf{proj}_C \circ (I - \lambda \nabla f)$$

- assume ∇f is L-Lipschitz, let $\lambda \in (0, 2/L)$
- since both $\mathbf{proj}_C$ and $(I - \lambda \nabla f)$ are averaged, T is averaged
- therefore, the following sequence weakly converges to a minimizer, if exists:

$$x^{k+1} \leftarrow T x^k = \mathbf{proj}_C\big(x^k - \lambda \nabla f(x^k)\big)$$

Special cases

prox-gradient method:

- **convex problem:**

$$\underset{x}{\text{minimize}}\ f(x) + h(x)$$

- assume sufficient intersection between $\text{dom} f$ and $\text{dom} h$
- **define:**

$$T := \mathbf{prox}_{\lambda h} \circ (I - \lambda \nabla f)$$

- assume ∇f is L-Lipschitz, let $\lambda \in (0, 2/L)$
- since both $\mathbf{prox}_{\lambda h}$ and $(I - \lambda \nabla f)$ are averaged, T is averaged
- therefore, the following sequence weakly converges to a minimizer, if exists:

$$x^{k+1} \leftarrow Tx^k = \mathbf{proj}_{\lambda h}\big(x^k - \lambda \nabla f(x^k)\big)$$

Special cases

Later this course, we will see more special cases

- forward-backward iteration
- Douglas-Rachford and Peaceman-Rachford iteration
- ADMM
- Tseng's forward-backward-forward iteration
- Davis-Yin iteration
- primal-dual iteration
- $\cdots$

Summary

- Fixed-point iteration and analysis are powerful tools
- Contractive T: fixed-point exists, is unique, iteration strongly converges
- Nonexpansive T: bounded, if fixed-point exists
- Averaged T: weakly converges, if fixed-point exists
- More power: closedness under composition