ORF522 – Linear and Nonlinear Optimization

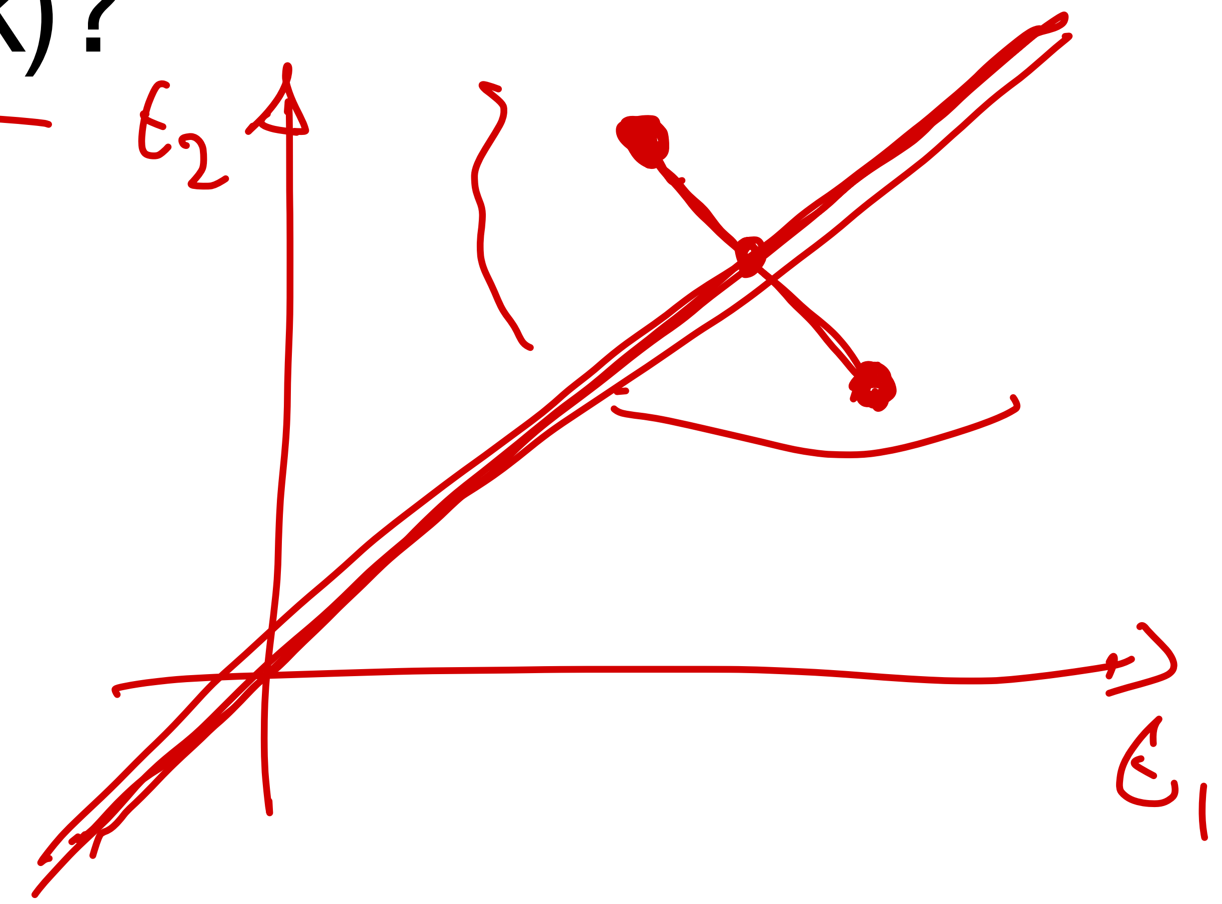
16. Proximal methods and introduction to operator theory

Bartolomeo Stellato — Fall 2021

Ed Forum

- Since there might be multiple subgradients that are very different, is there way to sometimes choose a 'best' subgradient for a given function that helps the algorithm converges faster?

- In Page 41 of Lecture 15, for the first fraction in this page, how do we conclude that it attains minimum when all t_k are equal based on the fact that the fraction is convex and symmetric in (t_1,...,t_k)?



-

Recap

Gradients and epigraphs

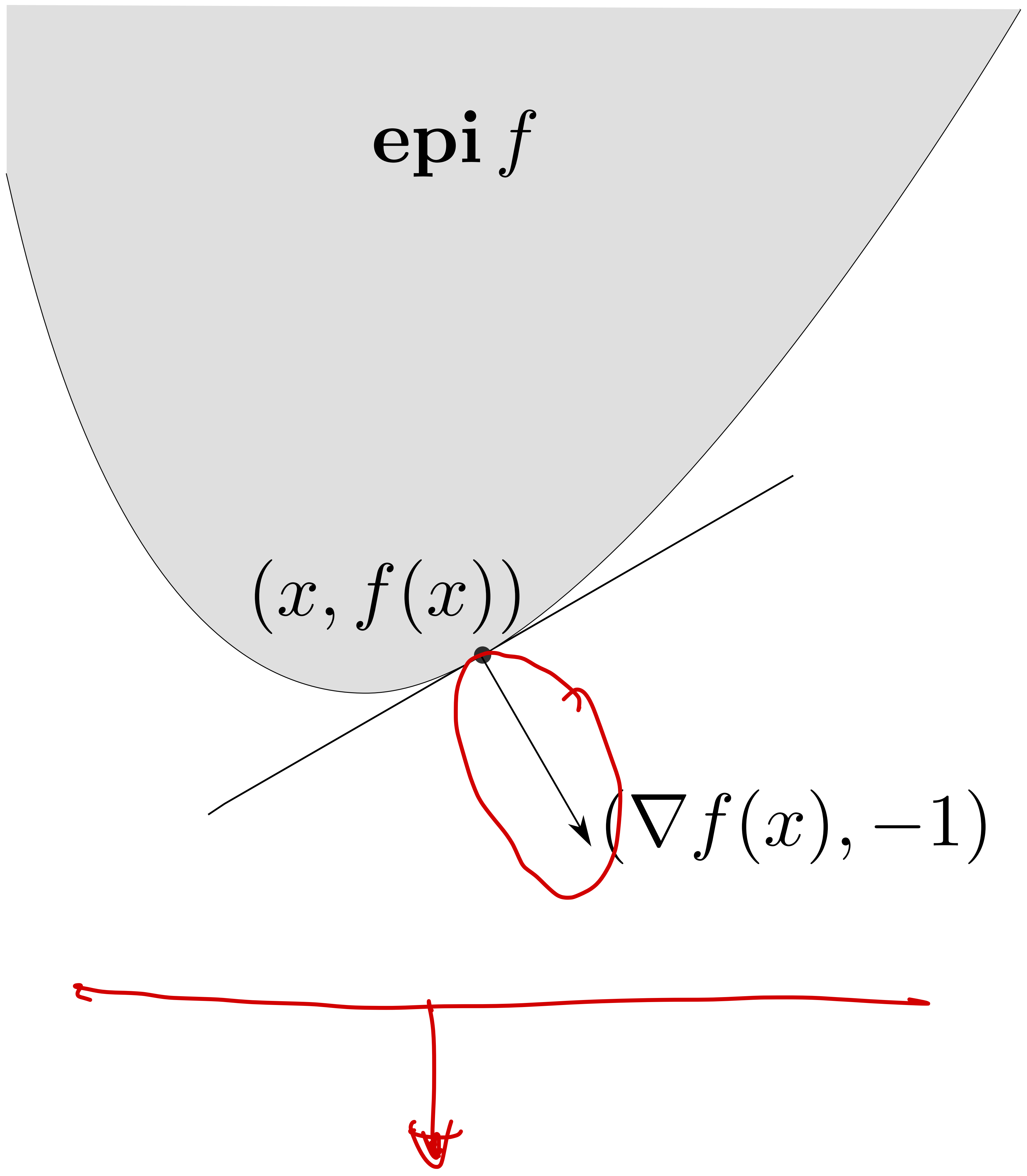
For a convex differentiable function f, i.e.

$$f(y) \geq f(x) + \nabla f(x)^T (y - x), \quad \forall y \in \mathbf{dom}\, f$$

Gradients and epigraphs

For a convex differentiable function f, i.e.

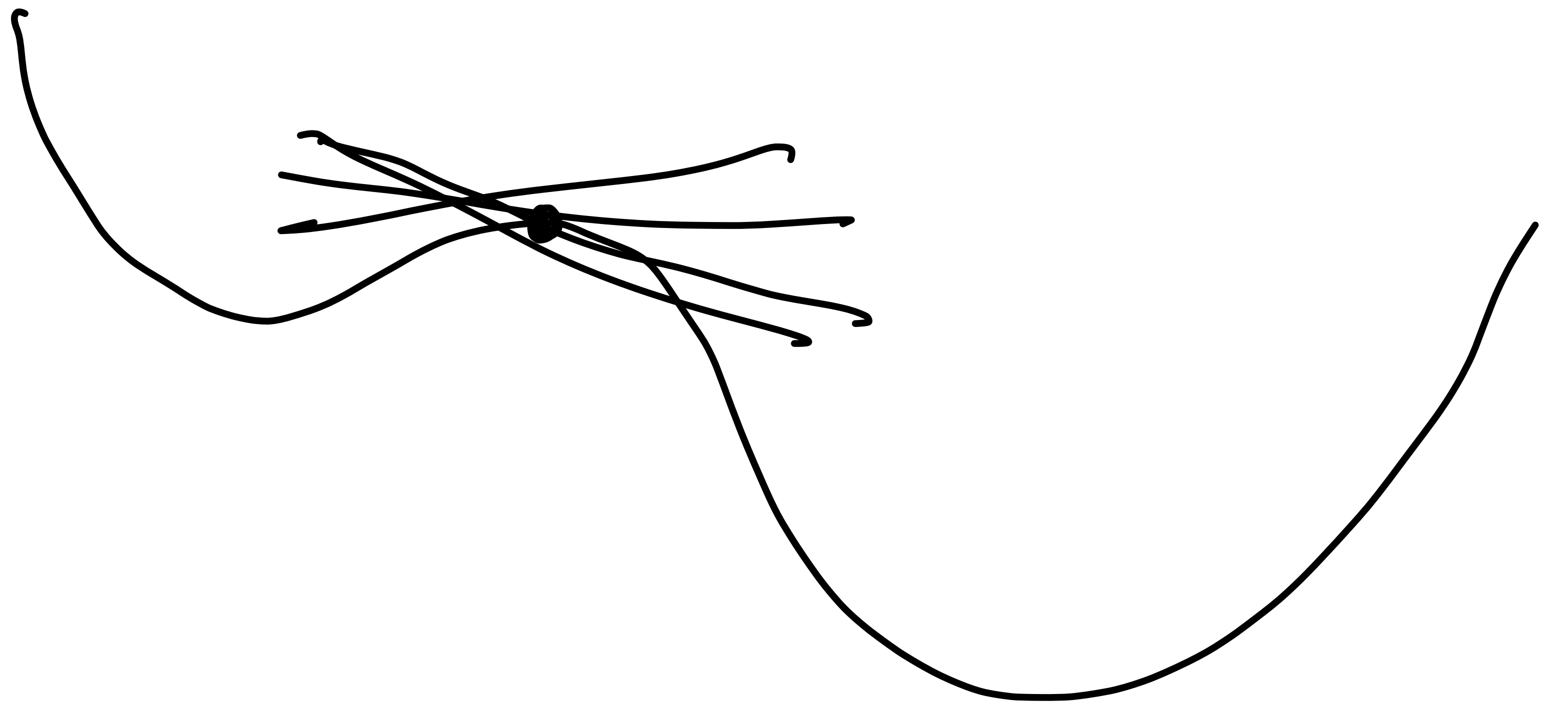
$$f(y) \geq f(x) + \nabla f(x)^T (y - x), \quad \forall y \in \textbf{dom}\, f$$



$(\nabla f(x), -1)$ defines a **supporting hyperplane** to epigraph of f at $(x, f(x))$

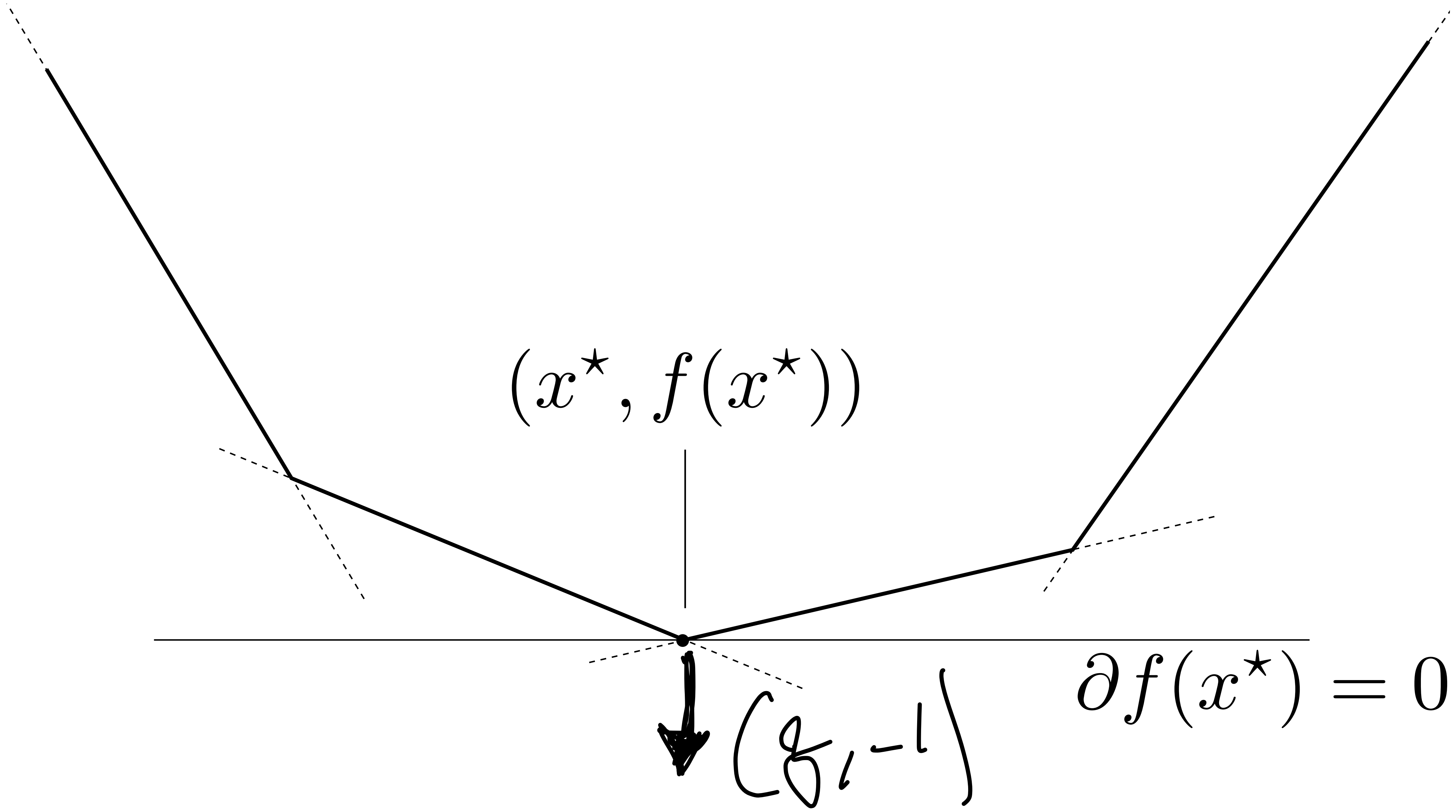
$$\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0, \quad \forall (y, t) \in \textbf{epi}\, f$$

Fermat's optimality condition



For any (not necessarily convex) function f where $\partial f(x^\star) \neq \emptyset$, $x^\star$ is a global minimizer if and only if

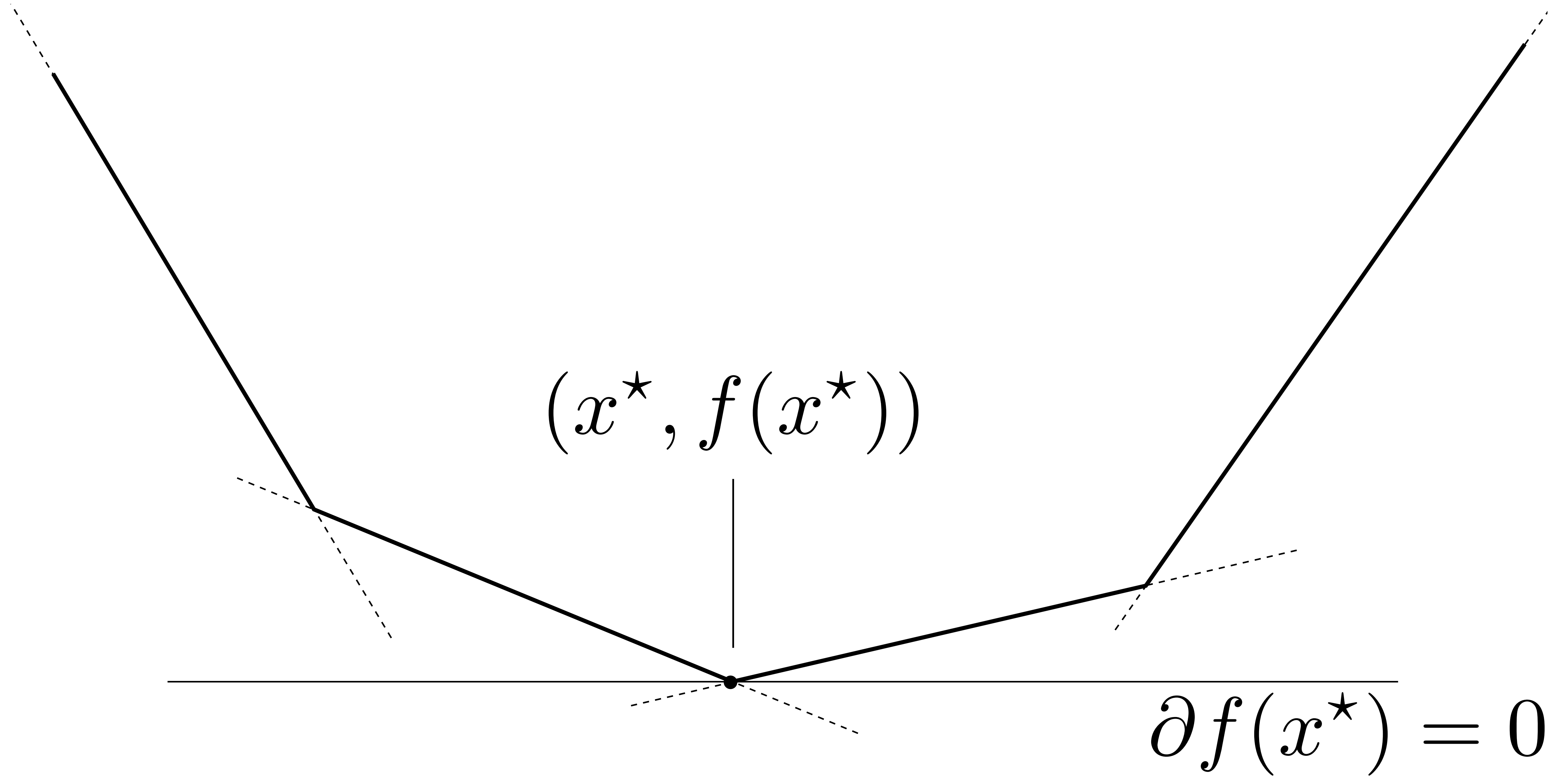
$$0 \in \partial f(x^\star)$$



Fermat's optimality condition

For any (not necessarily convex) function f where $\partial f(x^\star) \neq \emptyset$,
$x^\star$ is a global minimizer if and only if

$$0 \in \partial f(x^\star)$$



Proof

A subgradient $g = 0$ means that, for all y

$$f(y) \geq f(x^\star) + 0^T(y - x^\star) = f(x^\star) \quad \blacksquare$$

Note differentiable case with $\partial f(x) = \{\nabla f(x)\}$

Subgradient method

Convex optimization problem

minimize $\quad f(x) \qquad$ (optimal cost $f^\star$)

Iterations

$$x^{k+1} = x^k - t_k g^k, \qquad g^k \in \partial f(x^k)$$

g^k is **any subgradient** of f at x^k

Not a descent method, keep track of the best point

$$f_{\text{best}}^k = \min_{i=1,\dots,k} f(x^i)$$

Implications for step size rules

$$f_{\text{best}}^{k} - f^{\star} \leq \frac{R^2 + G^2 \sum_{i=0}^{k} t_i^2}{2 \sum_{i=0}^{k} t_i}$$

Fixed: $t_k = t$ for $k = 0, \ldots$

$$f_{\text{best}}^{k} - f^{\star} \leq \frac{R^2 + G^2 (k+1) t^2}{2(k+1)t}$$

May be suboptimal

$$\lim_{k \to \infty} f_{\text{best}}^{k} \leq f^{\star} + \frac{G^2 t}{2}$$

Diminishing: $\sum_{k=0}^{\infty} t_k^2 < \infty, \quad \sum_{k=0}^{\infty} t_k = \infty$

e.g., $t_k = \tau/(k+1)$ or $t_k = \tau/\sqrt{k+1}$

Optimal

$$\lim_{k \to \infty} f_{\text{best}}^{k} = f^{\star}$$

Summary subgradient method

- Simple
- Handles general nondifferentiable convex functions
- Very slow convergence $O(1/\epsilon^2)$
- No good stopping criterion

Summary subgradient method

- Simple
- Handles general nondifferentiable convex functions
- Very slow convergence $O(1/\epsilon^2)$
- No good stopping criterion

Can we do better?

Can we incorporate constraints?

Today's lecture
[Chapter 3 and 6, FMO] [PA] [PMO]

Proximal methods and introduction to operators

- Optimality conditions with subdifferentials
- Proximal operators
- Proximal gradient method
- Operator theory
- Fixed point iterations

Optimality conditions with subdifferentials

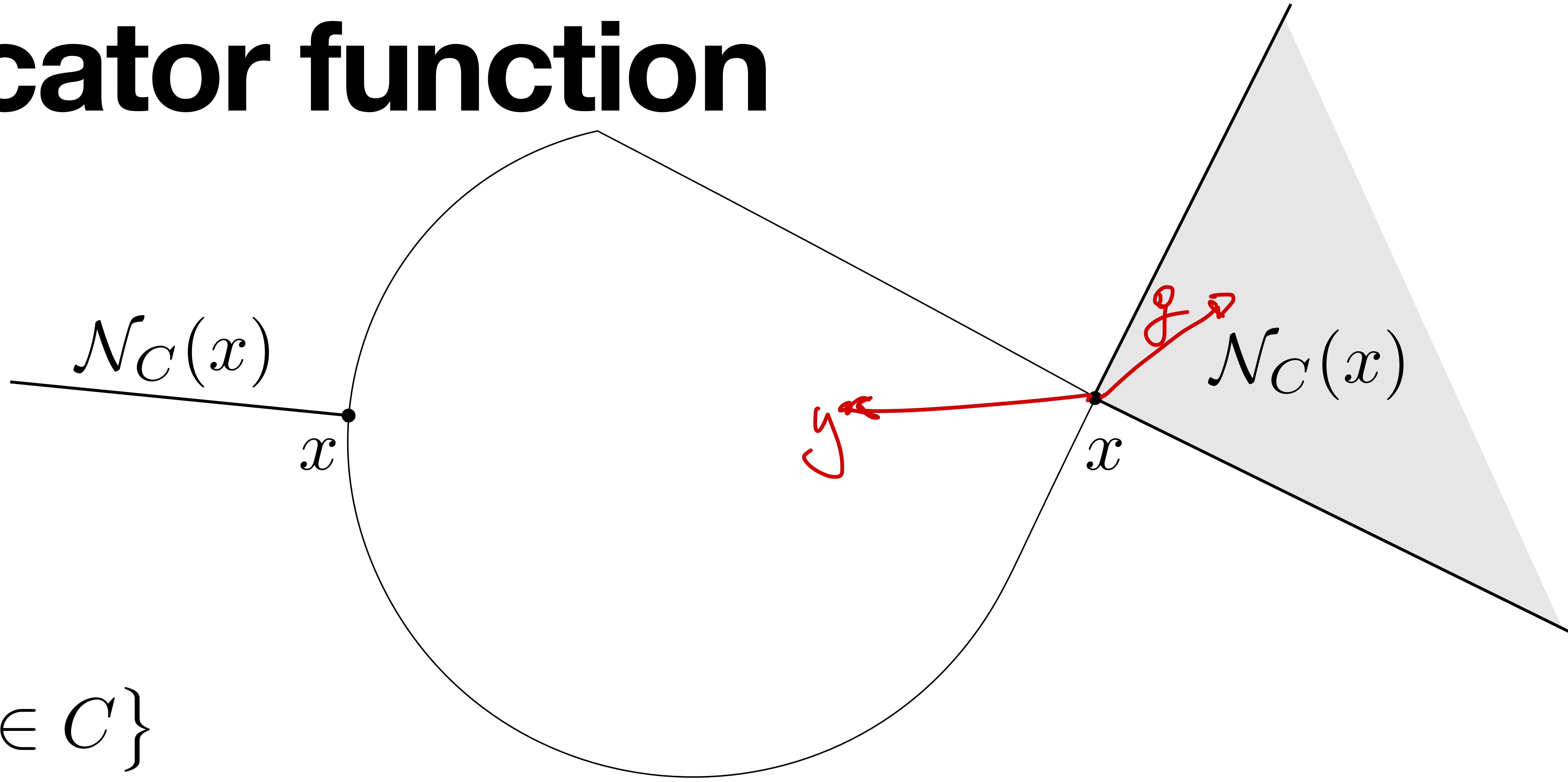
Subgradient of indicator function

The subdifferential of the **indicator functon** is the **normal cone**

$$\partial \mathcal{I}_C(x) = \mathcal{N}_C(x)$$

where,

$$\mathcal{N}_C(x) = \{g \mid g^T(y - x) \leq 0, \quad \text{for all } y \in C\}$$



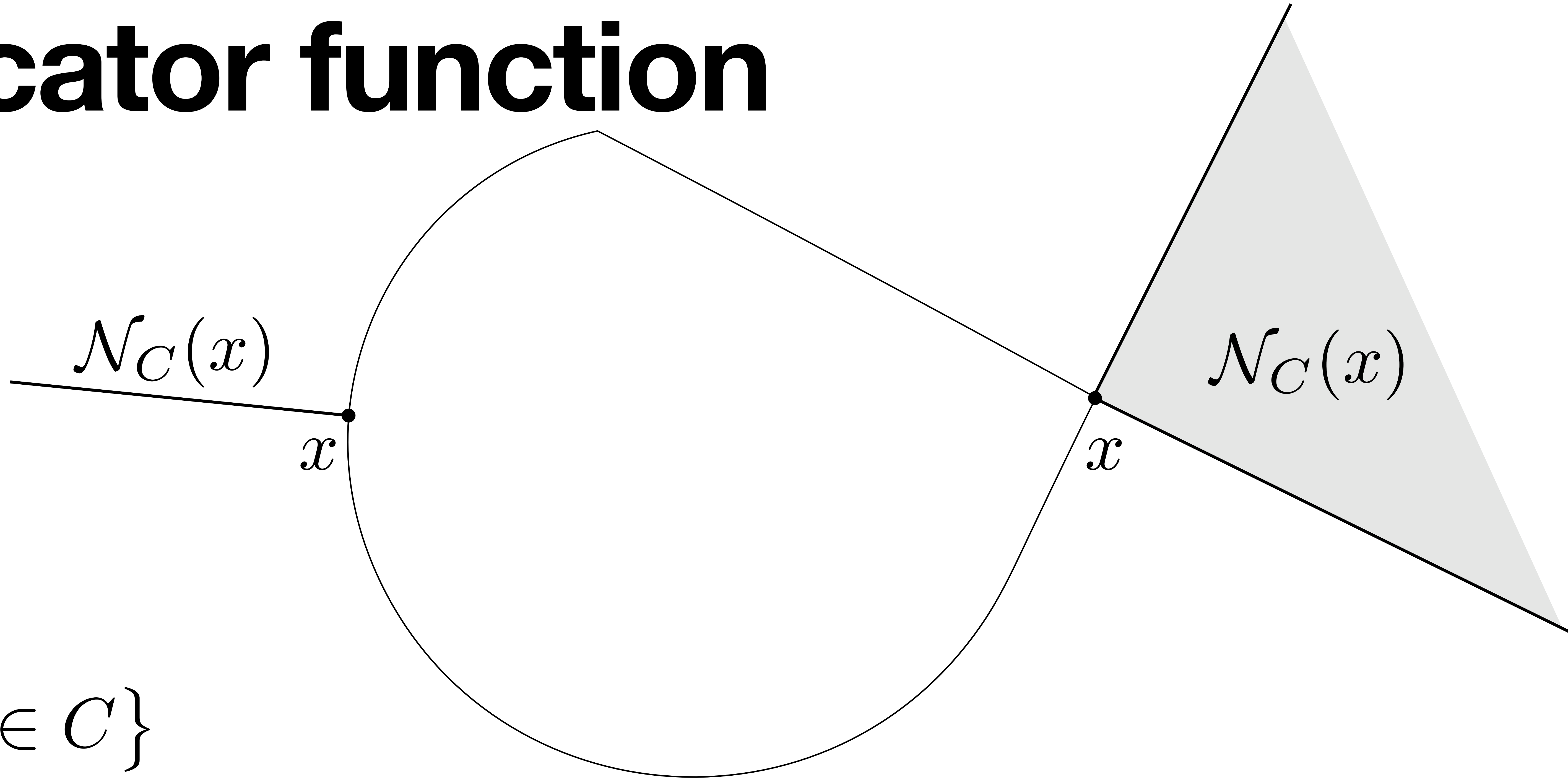
Subgradient of indicator function

The subdifferential of the **indicator functon** is the **normal cone**

$$\partial \mathcal{I}_C(x) = \mathcal{N}_C(x)$$

where,

$$\mathcal{N}_C(x) = \{g \mid g^T(y-x) \leq 0, \quad \text{for all } y \in C\}$$



Proof

By definition of subgradient g, $\mathcal{I}_C(y) \geq \mathcal{I}_C(x) + g^T(y-x), \quad \forall y$

$$y \notin C \quad \Longrightarrow \quad \mathcal{I}_C(y) = \infty$$

$$y \in C \quad \Longrightarrow \quad 0 \geq g^T(y-x)$$

■

Constrained optimization

Indicator function
of a convex set

$$\mathcal{I}_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$$

Constrained form

minimize $\quad f(x)$

subject to $\quad x \in C$

$\longleftrightarrow$

Unconstrained form

minimize $\quad f(x) + \mathcal{I}_C(x)$

First-order optimality conditions from subdifferentials

minimize $\quad f(x) + \mathcal{I}_C(x)$ $\qquad$ (f smooth, C convex)

First-order optimality conditions from subdifferentials

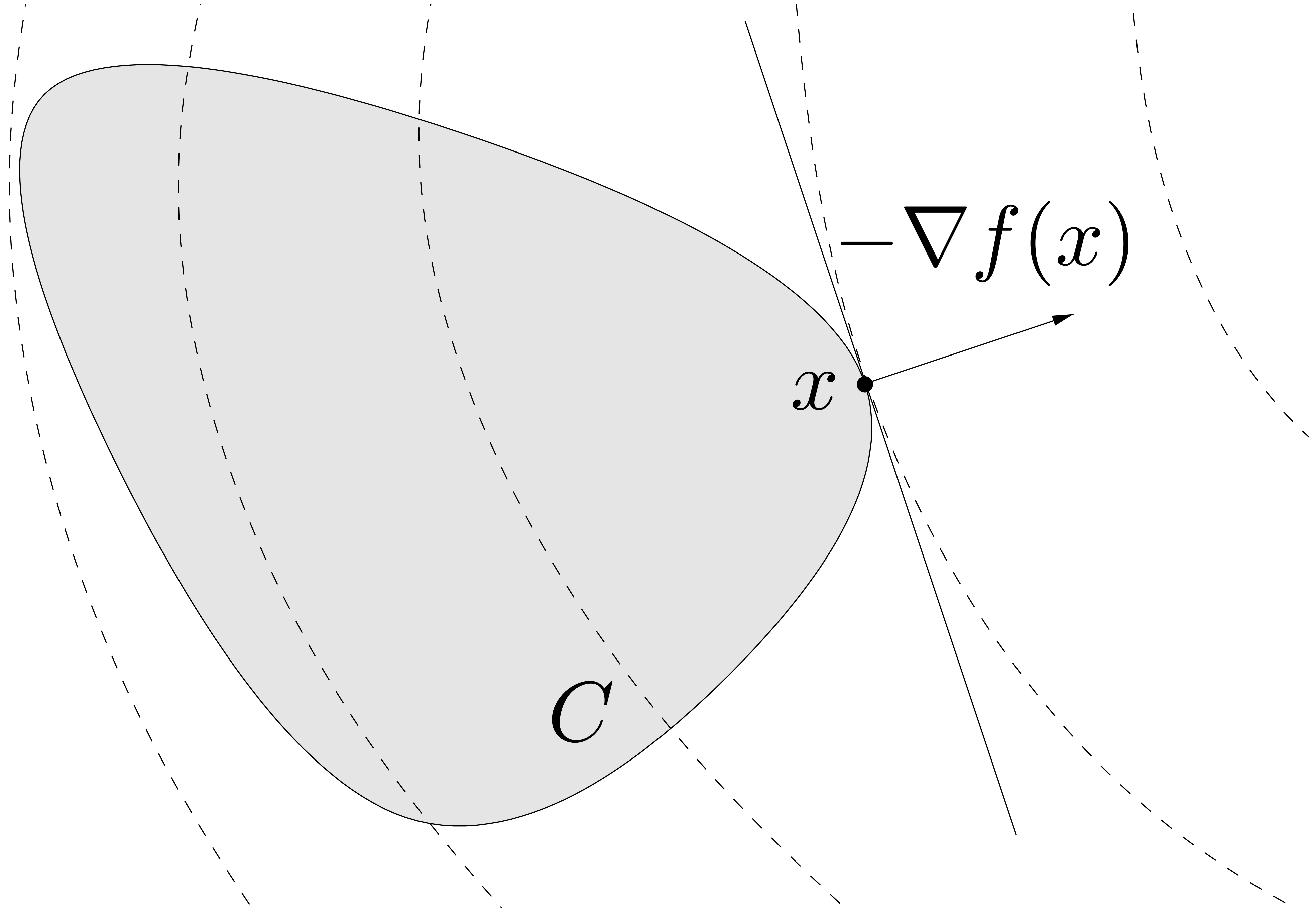
minimize $\quad f(x) + \mathcal{I}_C(x)$ $\qquad$ (f smooth, C convex)

Fermat's optimality condition

$$0 \in \partial(f(x) + \mathcal{I}_C(x))$$

$$\iff 0 \in \{\nabla f(x)\} + \mathcal{N}_C(x)$$

$$\iff -\nabla f(x) \in \mathcal{N}_C(x)$$



First-order optimality conditions from subdifferentials

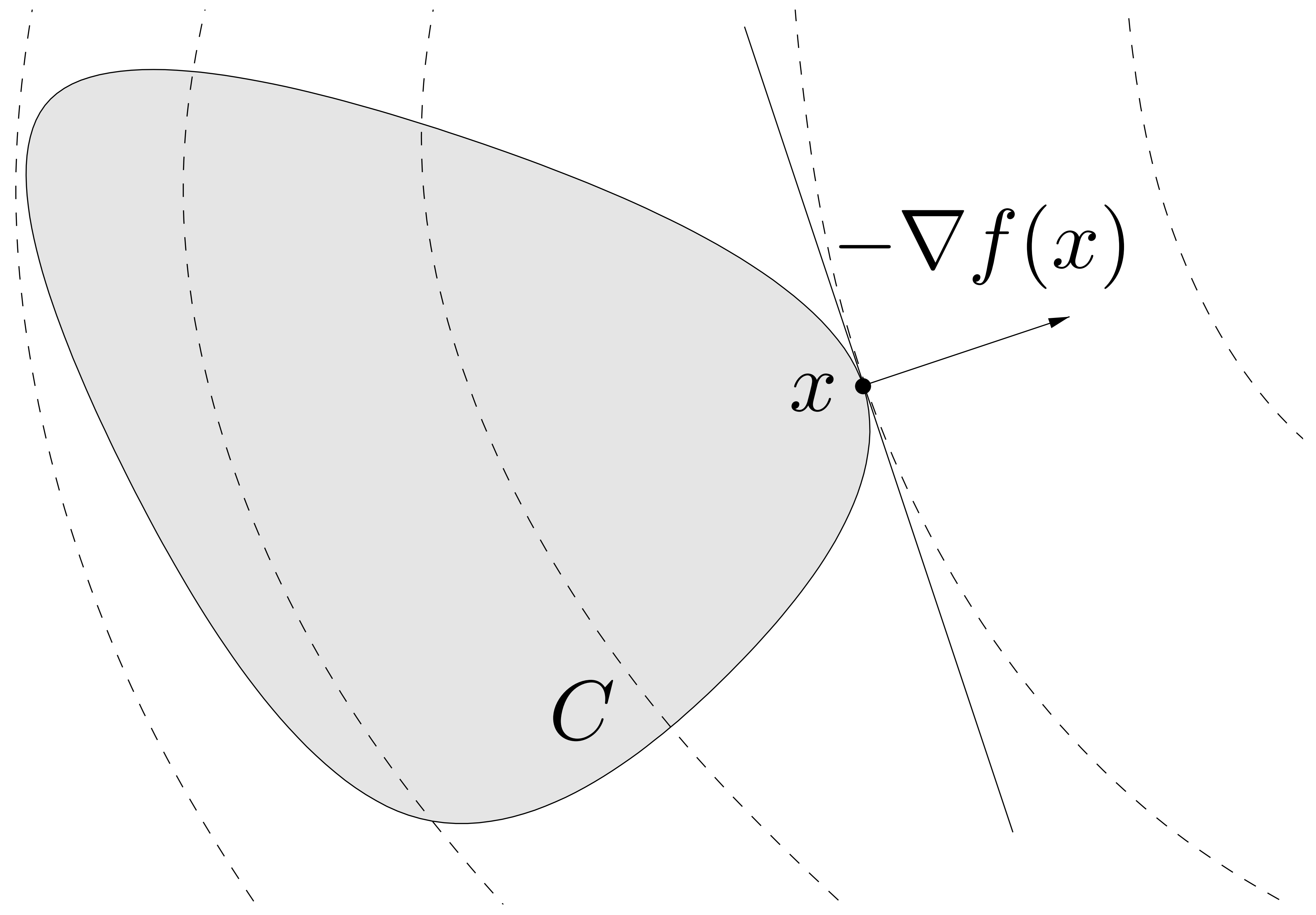
minimize $\quad f(x) + \mathcal{I}_C(x)$ $\qquad$ (f smooth, C convex)

Fermat's optimality condition

$$0 \in \partial(f(x) + \mathcal{I}_C(x))$$

$$\iff 0 \in \{\nabla f(x)\} + \mathcal{N}_C(x)$$

$$\iff -\nabla f(x) \in \mathcal{N}_C(x)$$



Equivalent to

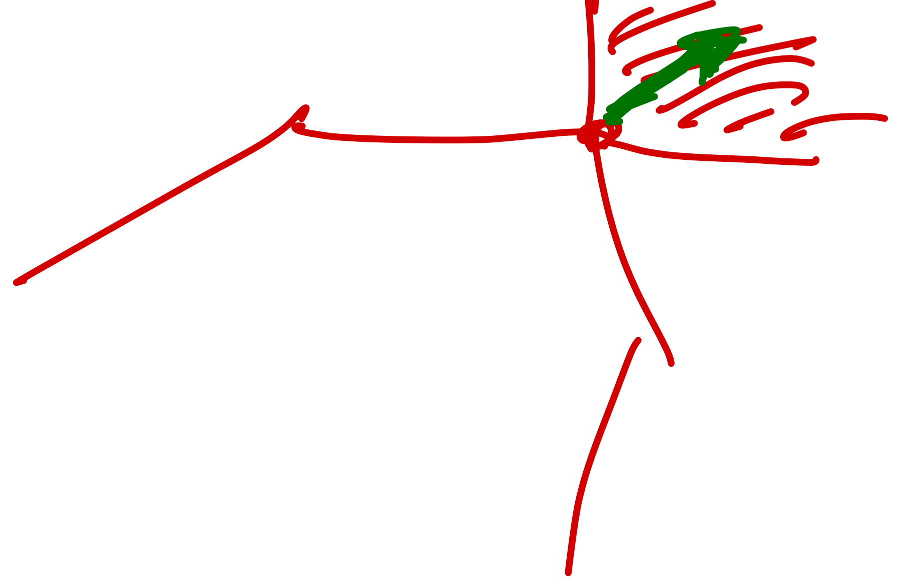
$$\nabla f(x)^T(y - x) \geq 0, \quad \forall y \in C$$

Example: KKT of a quadratic program

minimize $\quad (1/2)x^T Px + q^T x$
subject to $\quad Ax \leq b$

$\longrightarrow$ minimize $\quad (1/2)x^T Px + q^T x + \mathcal{I}_{\{Ax \leq b\}}(x)$

Example: KKT of a quadratic program



minimize	$(1/2)x^T Px + q^T x$
subject to	$Ax \leq b$

$\longrightarrow$ minimize $(1/2)x^T Px + q^T x + \mathcal{I}_{\{Ax \leq b\}}(x)$

Gradient

$$\nabla f(x) = Px + q$$

Normal cone to polyhedron

Idea: [Lecture 13].
Proof: [Theorem 6.46, Variational Analysis, Rockafellar & Wets]

$$\mathcal{N}_{\{Ax \leq b\}}(x) = \{A^T y \mid y \geq 0 \quad \text{and} \quad y_i(a_i^T x - b_i) = 0\}$$

Example: KKT of a quadratic program

$$\begin{array}{ll} \text{minimize} & (1/2)x^T P x + q^T x \\ \text{subject to} & Ax \le b \end{array} \longrightarrow \text{minimize} \quad (1/2)x^T P x + q^T x + \mathcal{I}_{\{Ax \le b\}}(x)$$

Gradient

$$\nabla f(x) = Px + q$$

Normal cone to polyhedron

Idea: [Lecture 13].
Proof: [Theorem 6.46, Variational Analysis, Rockafellar & Wets]

$$\mathcal{N}_{\{Ax \le b\}}(x) = \{A^T y \mid y \ge 0 \quad \text{and} \quad y_i(a_i^T x - b_i) = 0\}$$

First-order optimality condition

$$-\nabla f(x) \in \partial \mathcal{I}_{\{Ax \le b\}}(x) = \mathcal{N}_{\{Ax \le b\}}(x)$$

$\longleftrightarrow$

KKT Optimality conditions

$$Px + q + A^T y = 0$$ DUAL FEAS

$$y \ge 0$$ DUAL FEAS

$$Ax - b \le 0$$ PRIMAL FEAS

$$y_i(a_i^T x - b_i) = 0, \quad i = 1, \dots, m$$

COMPL SLACK

Proximal operators

Composite models

$$\text{minimize} \quad f(x) + g(x)$$

$f(x)$ convex and smooth
$g(x)$ convex (may be not differentiable)

Examples

- Regularized regression: $g(x) = \|x\|_1$
- Constrained optimization: $g(x) = \mathcal{I}_C(x)$

Proximal operator

Definition

The **proximal operator** of the function $g : \mathbf{R}^n \to \mathbf{R}$ is

$$\mathbf{prox}_g(x) = \underset{z}{\operatorname{argmin}} \left(g(z) + \frac{1}{2}\|z - x\|_2^2 \right)$$

Proximal operator

Definition

The **proximal operator** of the function $g : \mathbf{R}^n \to \mathbf{R}$ is

$$\mathbf{prox}_g(x) = \underset{z}{\text{argmin}} \left(g(z) + \frac{1}{2}\|z - x\|_2^2 \right)$$

Optimality conditions of prox

$$0 \in \partial g(z) + z - x \quad \Longrightarrow \quad x - z \in \partial g(z)$$

Proximal operator

Definition

The **proximal operator** of the function $g : \mathbf{R}^n \to \mathbf{R}$ is

$$\mathbf{prox}_g(x) = \underset{z}{\operatorname{argmin}} \left(g(z) + \frac{1}{2}\|z - x\|_2^2 \right)$$

Optimality conditions of prox

$$0 \in \partial g(z) + z - x \quad \Longrightarrow \quad x - z \in \partial g(z)$$

Properties

- It involves solving an optimization problem (not always easy!)
- Easy to evaluate for many standard functions, i.e. **proxable functions**
- Generalizes many well-known algorithms

Generalized projection

The $\mathbf{prox}$ operator of the indicator function $\mathcal{I}_C$ is the projection onto C

$$\mathbf{prox}_{\mathcal{I}_C}(v) = \underset{x \in C}{\operatorname{argmin}} \|x - v\|_2^2 = \Pi_C(v)$$

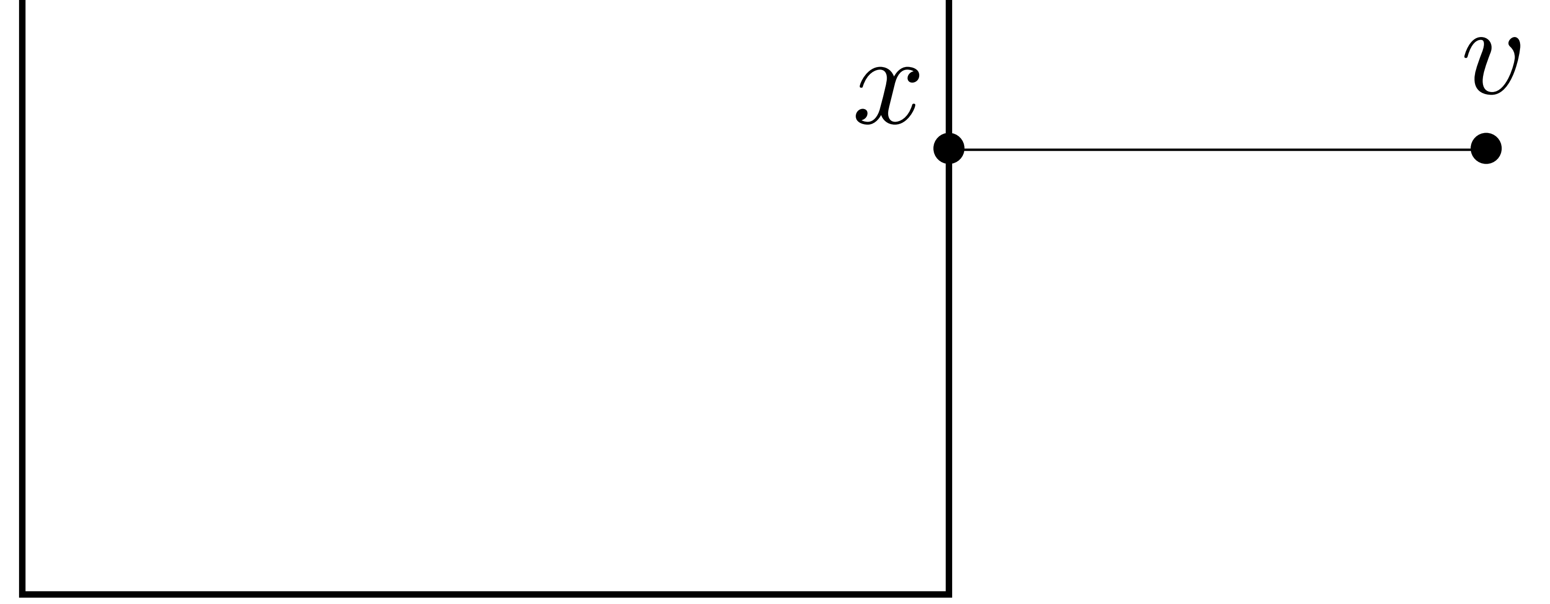
Generalized projection

The $\mathbf{prox}$ operator of the indicator function $\mathcal{I}_C$ is the projection onto C

$$\mathbf{prox}_{\mathcal{I}_C}(v) = \underset{x \in C}{\operatorname{argmin}} \|x - v\|_2^2 = \Pi_C(v)$$

Example projection onto a box $C = \{x \mid l \leq x \leq u\}$

$$\Pi_C(v)_i = \begin{cases} l_i & v_i \leq l_i \\ v_i & l_i \leq v_i \leq u_i \\ u_i & v_i \geq u_i \end{cases}$$



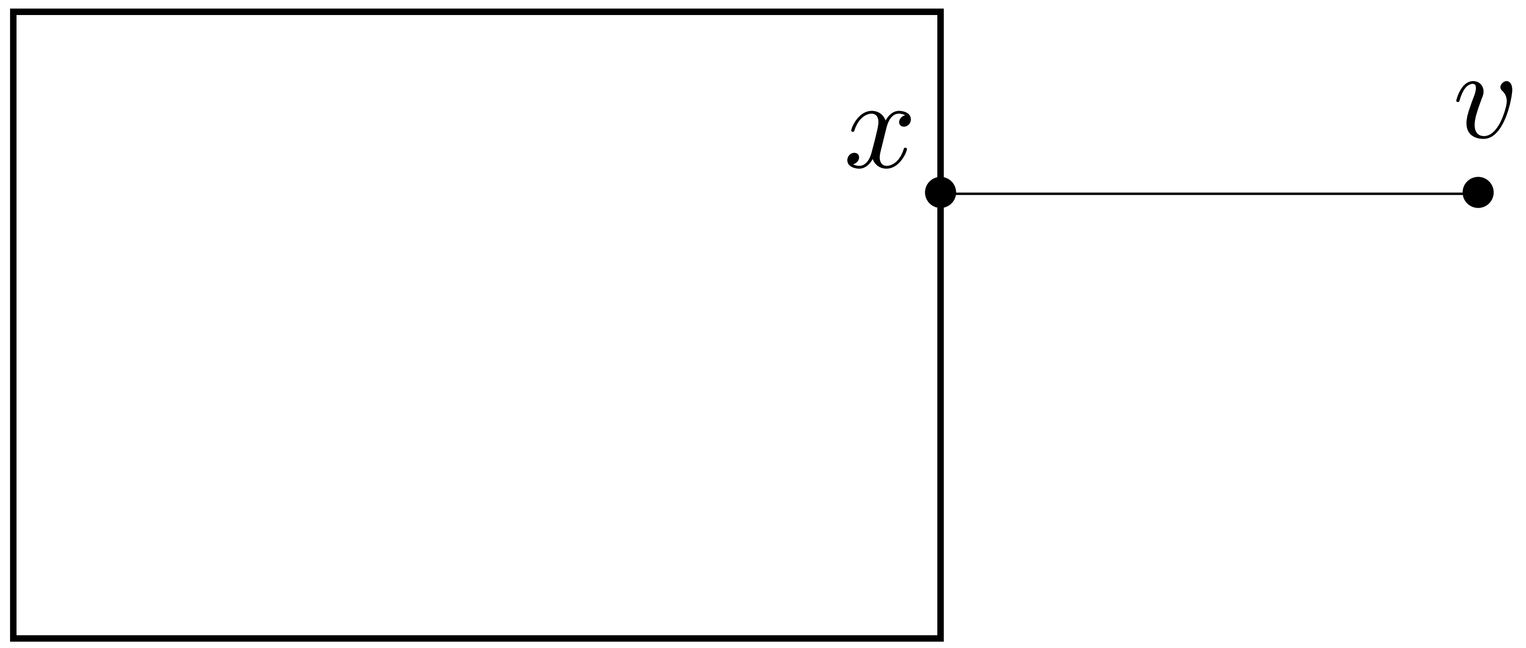
Generalized projection

The $\mathbf{prox}$ operator of the indicator function $\mathcal{I}_C$ is the projection onto C

$$\mathbf{prox}_{\mathcal{I}_C}(v) = \underset{x \in C}{\operatorname{argmin}} \|x - v\|_2^2 = \Pi_C(v)$$

Example projection onto a box $C = \{x \mid l \leq x \leq u\}$

$$\Pi_C(v)_i = \begin{cases} l_i & v_i \leq l_i \\ v_i & l_i \leq v_i \leq u_i \\ u_i & v_i \geq u_i \end{cases}$$



Remarks

- Easy for many common sets (e.g., closed form)
- Can be hard for surprisingly simple lets, e.g., $C = \{Ax \leq b\}$

Quadratic functions

$\mathrm{prox}_g(v) = \mathrm{argmin}_z \; g(z) + \frac{1}{2}\|z - v\|_2^2$

If $g(x) = (1/2)x^T Px + q^T x + r$ with $P \succeq 0$, then

$$\mathbf{prox}_g(v) = (I + P)^{-1}(v - q)$$

Remarks

- Closed-form always solvable (even with P not full rank)
- Symmetric, positive definite and usually sparse linear system
- Can prefactor $I + P$ and solve for different v

Separable sum

If $g(x)$ is block separable, i.e., $g(x) = \sum_{i=1}^{N} g_i(x_i)$

then, $(\mathbf{prox}_g(v))_i = \mathbf{prox}_{g_i}(v_i), \quad i = 1, \ldots, N$

(key to parallel/distributed proximal algorithms)

Separable sum

If $g(x)$ is block separable, i.e., $$g(x) = \sum_{i=1}^{N} g_i(x_i)$$

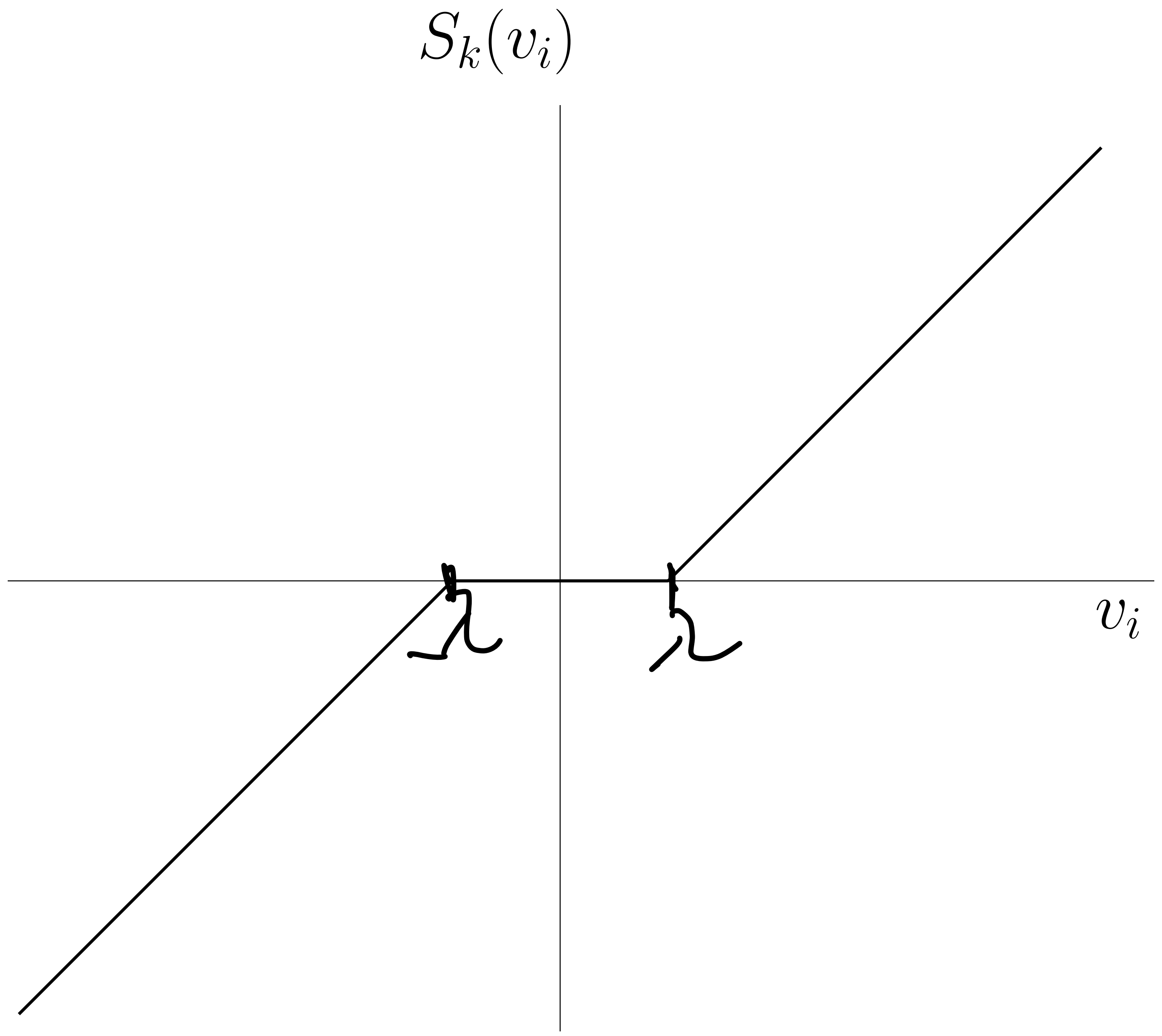
then, $$(\mathbf{prox}_g(v))_i = \mathbf{prox}_{g_i}(v_i), \quad i = 1, \ldots, N$$

(key to parallel/distributed proximal algorithms)

Example: $g(x) = \lambda \|x\|_1 = \sum_{i=1}^{n} \lambda |x_i|$

soft-thresholding

$$(\mathbf{prox}_g(v))_i = \mathbf{prox}_{\lambda|\cdot|}(v_i) = S_\lambda(v_i) = \begin{cases} v_i - \lambda & v_i > \lambda \\ 0 & |v_i| \leq \lambda \\ v_i + \lambda & v_i < -\lambda \end{cases}$$



Basic rules

Examples

- **Scaling and translation:** $g(x) = ah(x) + b$ with $a > 0$, then
$$\mathbf{prox}_g(x) = \mathbf{prox}_{ah}(x)$$
- **Affine addition:** $g(x) = h(x) + a^T x + b$, then
$$\mathbf{prox}_g(x) = \mathbf{prox}_h(x - a)$$
- **Affine transformation:** $g(x) = h(ax + b)$, with $a \neq 0, a \in \mathbf{R}$,
$$\mathbf{prox}_g(x) = \frac{1}{a}\left(\mathbf{prox}_{a^2 h}(ax + b) - b\right)$$

Proofs (exercise):

- Rearrange proximal term: $(1/2)\|z - x\|_2^2$
- Apply $\mathbf{prox}$ optimality conditions

Many more examples at [p. 156, FMO]

Proximal gradient method

Gradient descent interpretation

Problem

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

Iterations

$$x^{k+1} = x^k - t\nabla f(x^k)$$

Quadratic approximation, replacing Hessian $\nabla^2 f(x^k)$ with $\dfrac{1}{t}I$

$$x^{k+1} = \underset{z}{\text{argmin}}\ f(x^k) + \nabla f(x^k)^T(z - x^k) + \frac{1}{2t}\|z - x^k\|_2^2$$

Let's exploit the smooth part

minimize $\quad f(x) + g(x)$

$f(x)$ convex and smooth
$g(x)$ convex (may be not differentiable)

Quadratic approximation of f while keeping g

$$x^{k+1} = \underset{z}{\text{argmin}}\ g(z) + f(x^k) + \nabla f(x^k)^T(z - x^k) + \frac{1}{2t}\|z - x^k\|_2^2$$

Let's exploit the smooth part

minimize $\quad f(x) + g(x)$

$f(x)$ convex and smooth
$g(x)$ convex (may be not differentiable)

Quadratic approximation of f while keeping g

$$x^{k+1} = \underset{z}{\text{argmin}}\; g(z) + f(x^k) + \nabla f(x^k)^T(z - x^k) + \frac{1}{2t}\|z - x^k\|_2^2$$

← same as gradient descent

Let's exploit the smooth part

minimize $\quad f(x) + g(x)$

$f(x)$ convex and smooth
$g(x)$ convex (may be not differentiable)

Quadratic approximation of f while keeping g

$$x^{k+1} = \underset{z}{\text{argmin}}\ g(z) + f(x^k) + \nabla f(x^k)^T(z - x^k) + \frac{1}{2t}\|z - x^k\|_2^2$$

⟵ same as gradient descent

Equivalent to

$$x^{k+1} = \underset{z}{\text{argmin}}\ tg(z) + \frac{1}{2}\left\|z - (x^k - t\nabla f(x^k))\right\|_2^2 = \mathbf{prox}_{tg}\left(x^k - t\nabla f(x^k)\right)$$

Let's exploit the smooth part

minimize $\quad f(x) + g(x)$

$f(x)$ convex and smooth
$g(x)$ convex (may be not differentiable)

Quadratic approximation of f while keeping g

$$x^{k+1} = \operatorname*{argmin}_z \; g(z) + f(x^k) + \nabla f(x^k)^T (z - x^k) + \frac{1}{2t}\|z - x^k\|_2^2$$

⟵ same as gradient descent

Equivalent to

Proximal operator

$$x^{k+1} = \operatorname*{argmin}_z \; tg(z) + \frac{1}{2}\left\|z - (x^k - t\nabla f(x^k))\right\|_2^2 = \mathbf{prox}_{tg}\left(x^k - t\nabla f(x^k)\right)$$

↑ make g small

↑ stay close to gradient update

Proximal gradient method

minimize $\quad f(x) + g(x)$

$f(x)$ convex and smooth
$g(x)$ convex (may be not differentiable)

Iterations

$$x^{k+1} = \mathbf{prox}_{tg}\left(x^k - t\nabla f(x^k)\right)$$

Properties

- Alternates between gradient updates of f and proximal updates on g
- Useful if $\mathbf{prox}_{tg}$ is inespensive
- Can handle nonsmooth and constrained problems

Special cases

Generalized gradient descent

Problem

minimize $\quad f(x) + g(x)$

Iterations

$$x^{k+1} = \mathbf{prox}_{tg}\left(x^k - t\nabla f(x^k)\right)$$

Special cases

Generalized gradient descent

Problem

minimize $\quad f(x) + g(x)$

Iterations

$$x^{k+1} = \mathbf{prox}_{tg}\left(x^k - t\nabla f(x^k)\right)$$

Smooth

$$g(x) = 0 \quad \Longrightarrow \quad \mathbf{prox}_{tg}(x) = x$$

Gradient descent

$$\Longrightarrow \quad x^{k+1} = x^k - t\nabla f(x^k)$$

Special cases

Generalized gradient descent

Problem

minimize $f(x) + g(x)$

Iterations

$$x^{k+1} = \mathbf{prox}_{tg}\left(x^k - t\nabla f(x^k)\right)$$

Smooth

$$g(x) = 0 \quad \Longrightarrow \quad \mathbf{prox}_{tg}(x) = x$$

Gradient descent

$$\Longrightarrow \quad x^{k+1} = x^k - t\nabla f(x^k)$$

Constraints

$$g(x) = \mathcal{I}_C(x) \quad \Longrightarrow \quad \mathbf{prox}_{tg}(x) = \Pi_C(x)$$

Projected gradient descent

$$\Longrightarrow \quad x^{k+1} = \Pi_C(x^k - t\nabla f(x^k))$$

Special cases

Generalized gradient descent

Problem

minimize $\quad f(x) + g(x)$

Iterations

$$x^{k+1} = \mathbf{prox}_{tg}\left(x^k - t\nabla f(x^k)\right)$$

Smooth

$$g(x) = 0 \quad \Longrightarrow \quad \mathbf{prox}_{tg}(x) = x$$

Gradient descent

$$\Longrightarrow \quad x^{k+1} = x^k - t\nabla f(x^k)$$

Constraints

$$g(x) = \mathcal{I}_C(x) \quad \Longrightarrow \quad \mathbf{prox}_{tg}(x) = \Pi_C(x)$$

Projected gradient descent

$$\Longrightarrow \quad x^{k+1} = \Pi_C(x^k - t\nabla f(x^k))$$

Non smooth

$$f(x) = 0$$

Proximal minimization

$$\Longrightarrow \quad x^{k+1} = \mathbf{prox}_{tg}(x^k)$$

Note: useful if $\mathbf{prox}_{tg}$ is cheap

What happens if we cannot evaluate the prox?

At every iteration, it can be very expensive to evaluate

$$\mathbf{prox}_g(x) = \operatorname*{argmin}_z \left(g(z) + \frac{1}{2}\|z - x\|_2^2 \right)$$

Idea: solve it approximately!

What happens if we cannot evaluate the prox?

At every iteration, it can be very expensive to evaluate

$$\mathbf{prox}_g(x) = \underset{z}{\operatorname{argmin}} \left(g(z) + \frac{1}{2}\|z - x\|_2^2 \right)$$

Idea: solve it approximately!

If you precisely control the $\mathbf{prox}_g(x)$ evaluation errors
you can obtain the same convergence guarantees (and rates)
as the exact evaluations.

[Schmidt et al. (2011), "Convergence rates of inexact proximal-gradient methods for convex optimization"]

Example: Lasso

Iterative Soft Thresholding Algorithm (ISTA)

$$\text{minimize} \quad (1/2)\|Ax - b\|_2^2 + \lambda\|x\|_1$$

Example: Lasso

Iterative Soft Thresholding Algorithm (ISTA)

minimize $\underbrace{(1/2)\|Ax - b\|_2^2}_{f(x)} + \underbrace{\lambda\|x\|_1}_{g(x)}$

Example: Lasso

Iterative Soft Thresholding Algorithm (ISTA)

minimize $\underbrace{(1/2)\|Ax - b\|_2^2}_{f(x)} + \underbrace{\lambda\|x\|_1}_{g(x)}$

Proximal gradient descent

$$x^{k+1} = \mathbf{prox}_{tg}\left(x^k - t\nabla f(x^k)\right)$$

$$\nabla f(x) = A^T(Ax - b)$$

$$\mathbf{prox}_{tg}(x) = S_{\lambda t}(x)$$

(component wise soft-thresholding)

Example: Lasso

Iterative Soft Thresholding Algorithm (ISTA)

$$\text{minimize} \quad \underbrace{(1/2)\|Ax - b\|_2^2}_{f(x)} + \underbrace{\lambda\|x\|_1}_{g(x)}$$

Proximal gradient descent

$$x^{k+1} = \mathbf{prox}_{tg}\left(x^k - t\nabla f(x^k)\right)$$

$$\nabla f(x) = A^T(Ax - b)$$

$$\mathbf{prox}_{tg}(x) = S_{\lambda t}(x)$$

(component wise soft-thresholding)

Closed-form iterations

$$x^{k+1} = S_{\lambda t}\left(x^k - tA^T(Ax^k - b)\right)$$

Example: Lasso

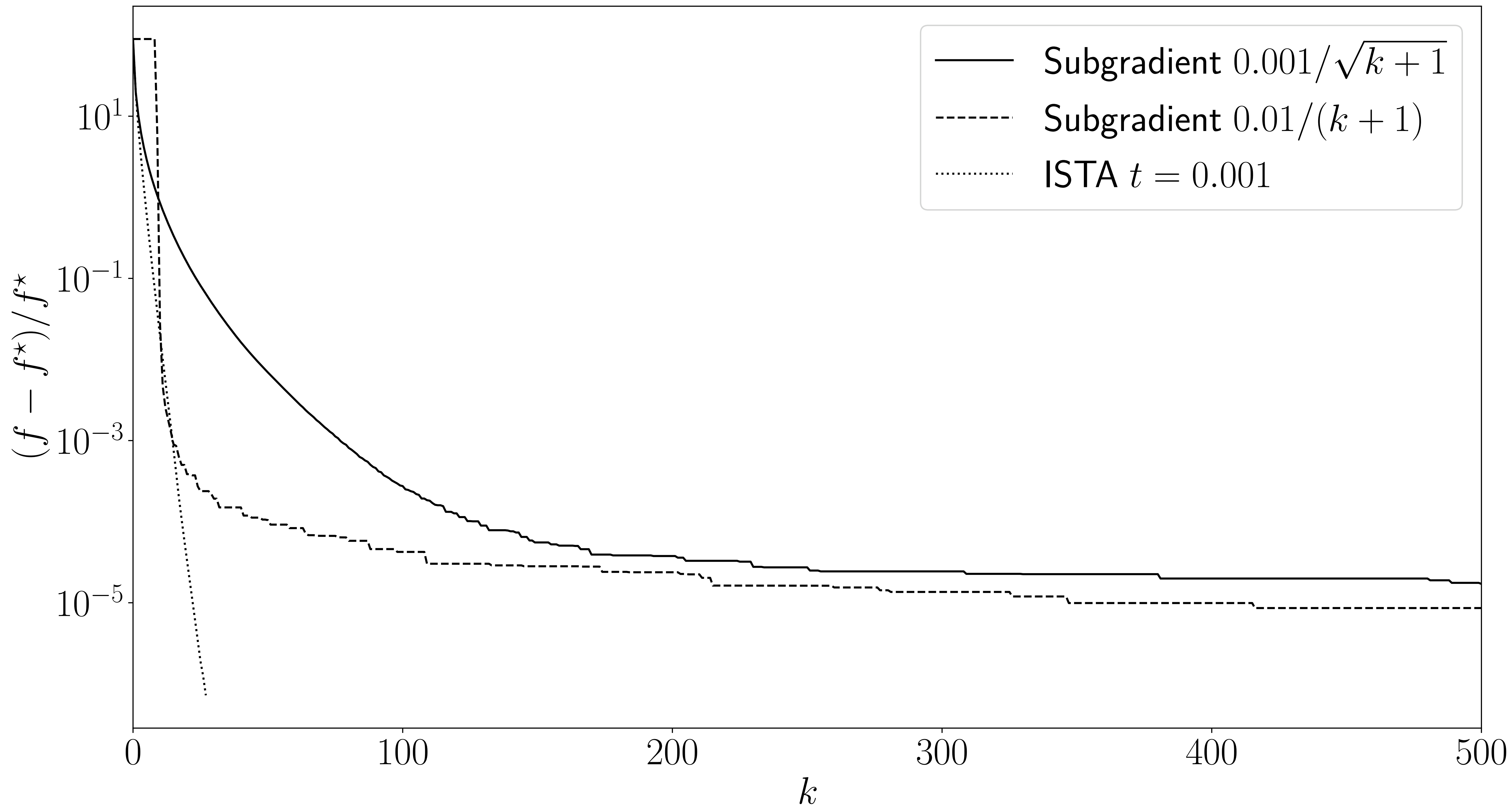
Iterative Soft Thresholding Algorithm (ISTA)

$A \in \mathbf{R}^{500 \times 100}$

$$\text{minimize} \quad (1/2)\|Ax - b\|_2^2 + \lambda \|x\|_1$$

Closed-form iterations

$$x^{k+1} = S_{\lambda t}\left(x^k - tA^T(Ax^k - b)\right)$$



Example: Lasso

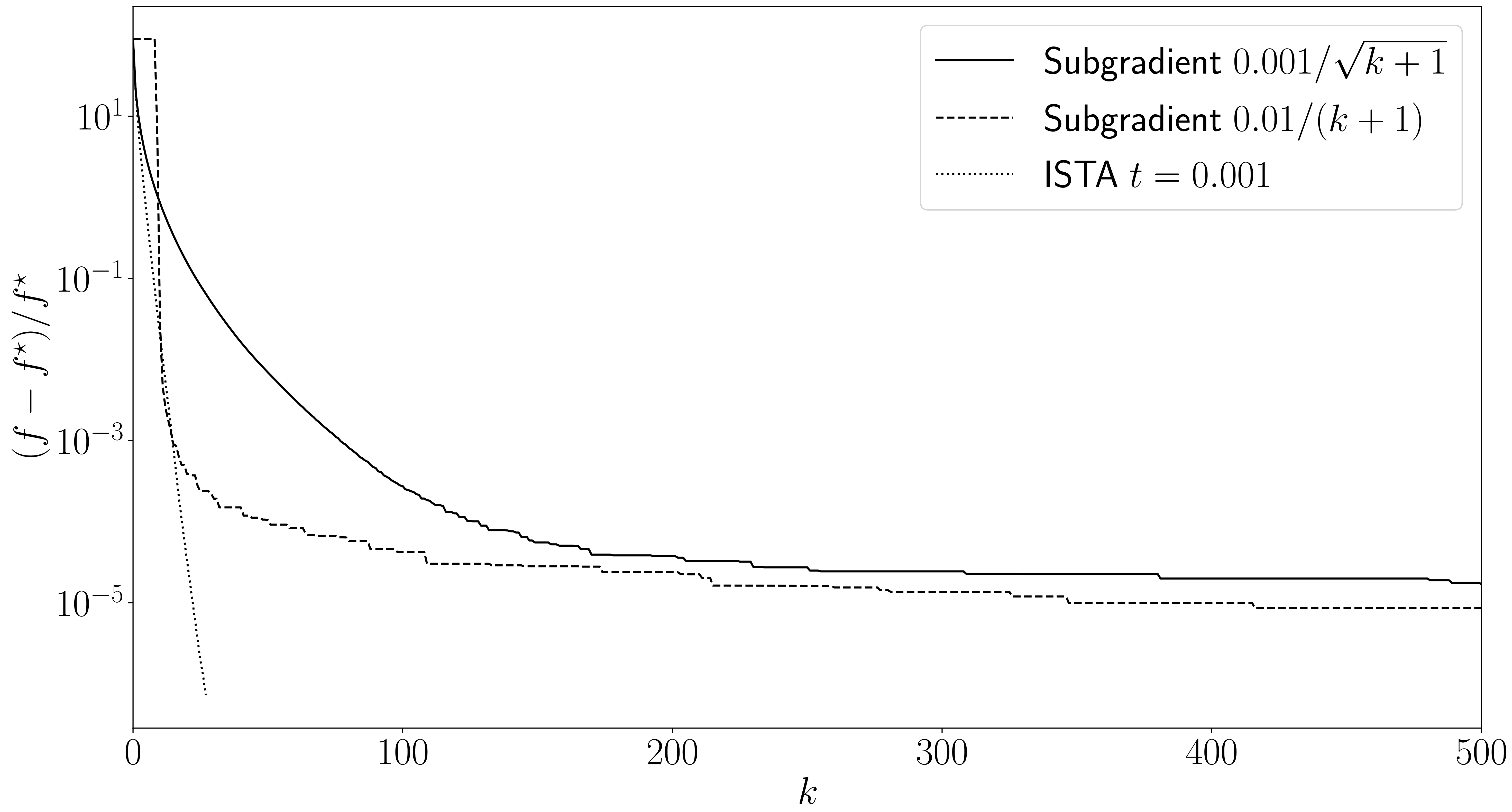
Iterative Soft Thresholding Algorithm (ISTA)

$A \in \mathbf{R}^{500 \times 100}$

minimize $\quad (1/2)\|Ax - b\|_2^2 + \lambda \|x\|_1$

Closed-form iterations

$$x^{k+1} = S_{\lambda t}\left(x^k - tA^T(Ax^k - b)\right)$$



Better convergence

Can we prove convergence generally?

Can we combine different operators?

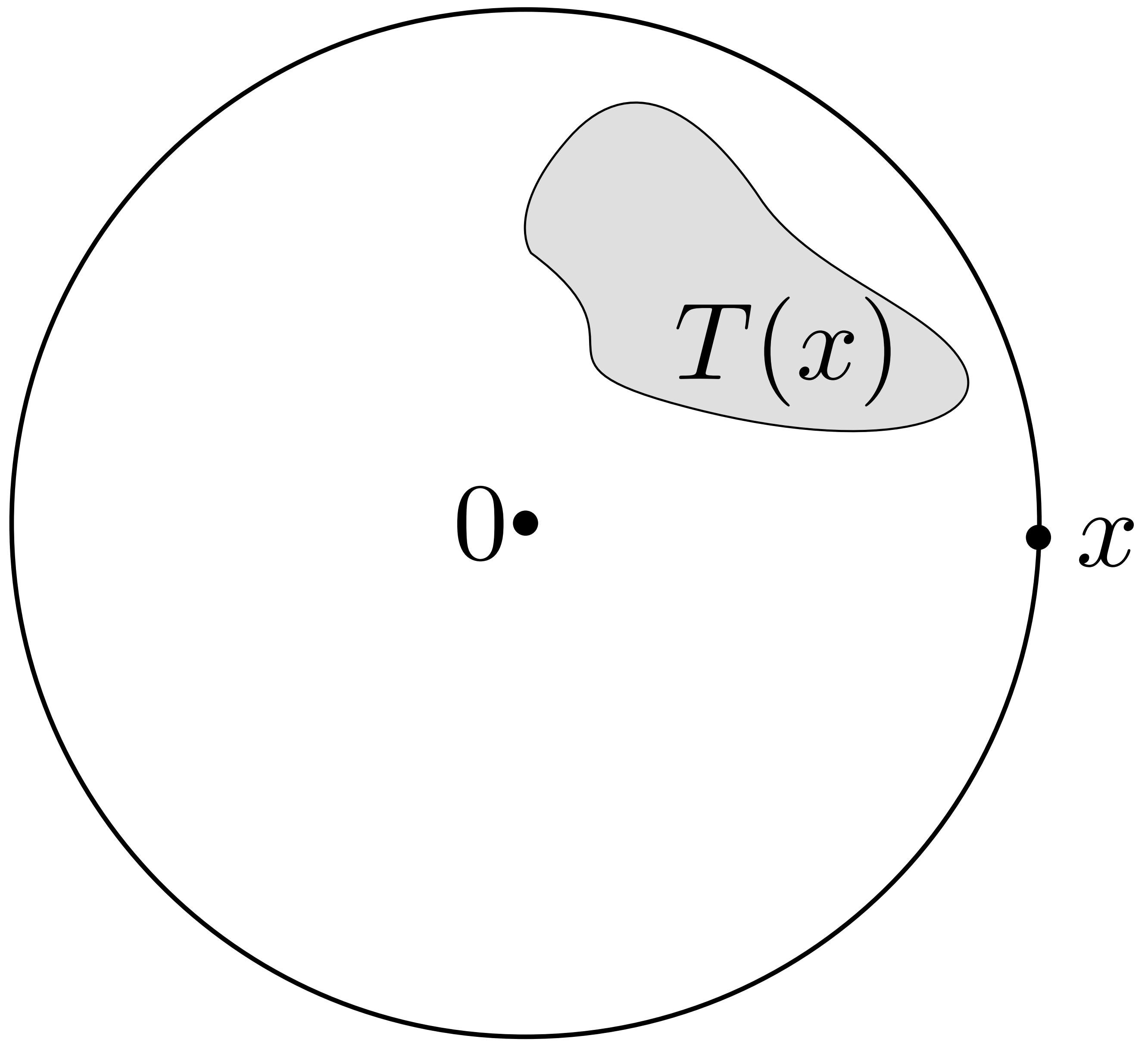
Introduction to operators

Operators

An operator T maps each point in $\mathbf{R}^n$ to a subset of $\mathbf{R}^n$

- **set valued** $T(x)$ returns a set
- **single-valued** $T(x)$ (function) returns a singleton

The **domain** of T is the set $\mathbf{dom}\, T = \{x \mid T(x) \neq \emptyset\}$

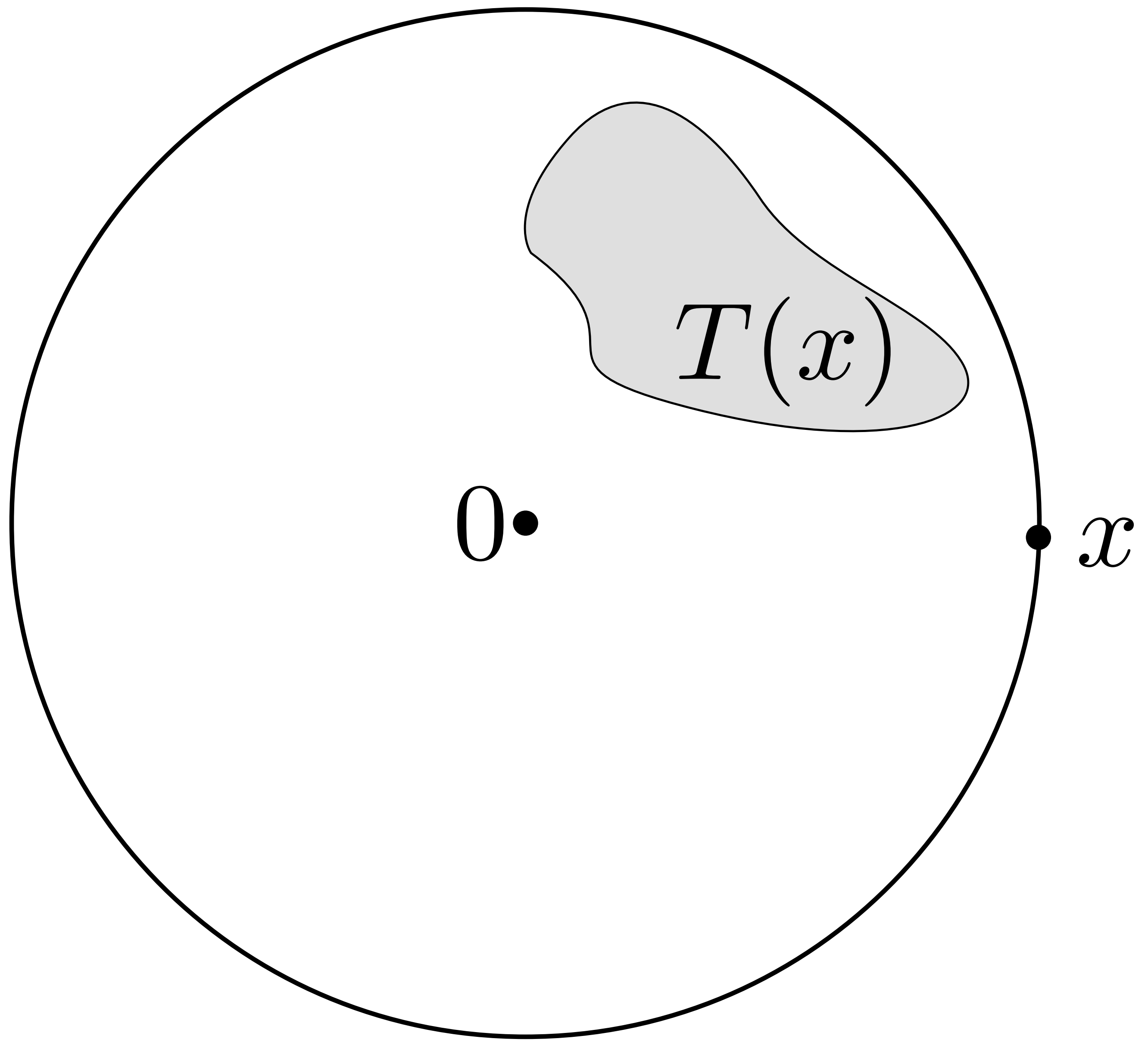


Operators

An operator T maps each point in $\mathbf{R}^n$ to a subset of $\mathbf{R}^n$

- **set valued** $T(x)$ returns a set
- **single-valued** $T(x)$ (function) returns a singleton

The **domain** of T is the set $\mathbf{dom}\, T = \{x \mid T(x) \neq \emptyset\}$



Example

- The subdifferential ∂f is a set-valued operator
- The gradient ∇f is a single-valued operator

Graph and inverse operators

Graph
The graph of an operator T is defined as

$$\mathbf{gph} T = \{(x, y) \mid y \in T(x)\}$$

In other words, all the pairs of points (x, y) such that $y \in T(x)$.

Graph and inverse operators

Graph

The graph of an operator T is defined as

$$\mathbf{gph} T = \{(x, y) \mid y \in T(x)\}$$

In other words, all the pairs of points (x, y) such that $y \in T(x)$.

Inverse

The graph of the inverse operator T^{-1} is defined as

$$\mathbf{gph} T^{-1} = \{(y, x) \mid (x, y) \in \mathbf{gph} T\}$$

Therefore, $y \in T(x)$ if and only if $x \in T^{-1}(y)$.

Zeros

Zero

x is a **zero** of T if $\quad 0 \in T(x)$

Zero set

The set of all the zeros $\quad T^{-1}(0) = \{x \mid 0 \in T(x)\}$

Zeros

Zero

x is a **zero** of T if $\qquad 0 \in T(x)$

Zero set

The set of all the zeros $\qquad T^{-1}(0) = \{x \mid 0 \in T(x)\}$

Example
If $T = \partial f$ and $f : \mathbf{R}^n \to \mathbf{R}$, then
$0 \in T(x)$ means that x minimizes f

Many problems
can be posed as finding zeros
of an operator

$I(x) = x$

Fixed points

$\bar{x}$ is a **fixed-point** of a single-valued operator T if

$$\bar{x} = T(\bar{x})$$

Set of fixed points $\quad \mathbf{fix}\, T = \{x \in \mathbf{dom}\, T \mid x = T(x)\} = (I - T)^{-1}(0)$

$(I - T)x = 0$

Examples

- **Identity** $T(x) = x$. Any point is a fixed point
- **Zero operator** $T(x) = 0$. Only 0 is a fixed point

Lipschitz operators

An operator T is L-Lipschitz if

$$\|T(x) - T(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbf{dom}\, T$$

Lipschitz operators

An operator T is L-Lipschitz if

$$\|T(x) - T(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbf{dom}\, T$$

Fact If T is Lipschitz, then it is single-valued

Proof If $y = T(x), z = T(x)$, then $\|y - z\| \leq L\|x - x\| = 0 \implies y = z$ $\blacksquare$

Lipschitz operators

An operator T is L-Lipschitz if

$$\|T(x) - T(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbf{dom}\, T$$

Fact If T is Lipschitz, then it is single-valued

Proof If $y = T(x), z = T(x)$, then $\|y - z\| \leq L\|x - x\| = 0 \implies y = z$ ■

For $L = 1$ we say T is **nonexpansive**

For $L < 1$ we say T is **contractive** (with contraction factor L)

Lipschitz operators examples

Lipschitz affine functions

$$T(x) = Ax + b \quad \longleftrightarrow \quad L = \|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$$

maximum singular value

$$\|T(y) - T(x)\| = \|A(y-x)\| \leq \|A\|_2 \|y-x\|_2$$

Lipschitz operators examples

Lipschitz affine functions

$T(x) = Ax + b$ ⟷ maximum singular value $L = \|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$

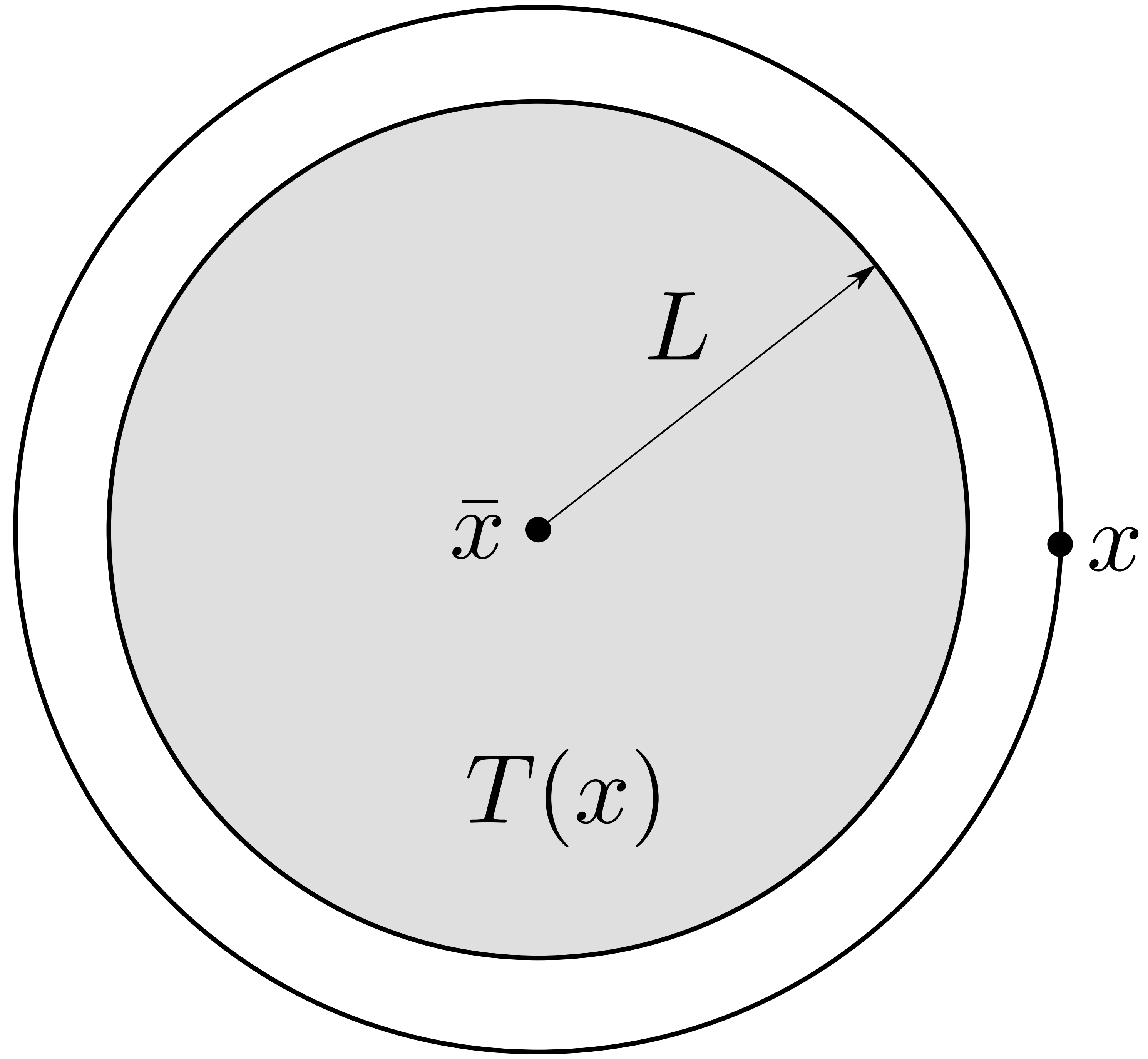
Lipschitz differentiable functions

T such that there exists derivative DT ⟷ derivative is bounded $\|DT\|_2 \leq L$

Lipschitz operators and fixed points

Given a L-Lipschitz operator T and a fixed point $\bar{x} = T\bar{x}$,

$$\|Tx - \bar{x}\| = \|Tx - T\bar{x}\| \leq L\|x - \bar{x}\|$$



Lipschitz operators and fixed points

Given a L-Lipschitz operator T and a fixed point $\bar{x} = T\bar{x}$,

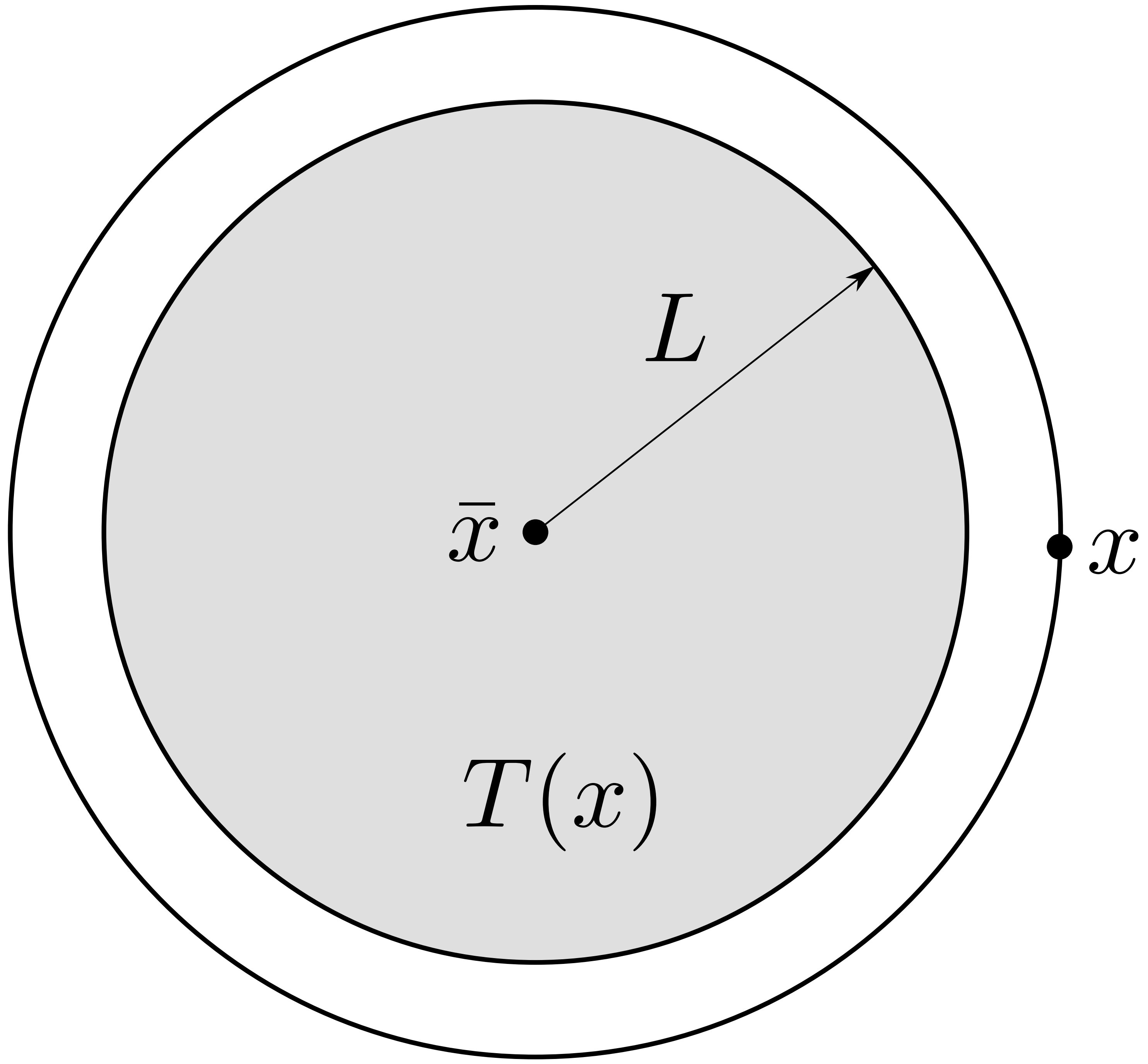
$$\|Tx - \bar{x}\| = \|Tx - T\bar{x}\| \leq L\|x - \bar{x}\|$$

A contractive operator ($L < 1$) can have at most one fixed point, i.e., $\mathbf{fix}\, T = \{\bar{x}\}$

Proof

If $\bar{x}, \bar{y} \in \mathbf{fix}\, T$ and $\bar{x} \neq \bar{y}$ then

$\|\bar{x} - \bar{y}\| = \|T(\bar{x}) - T(\bar{y})\| < \|\bar{x} - \bar{y}\|$ (contradiction) ■



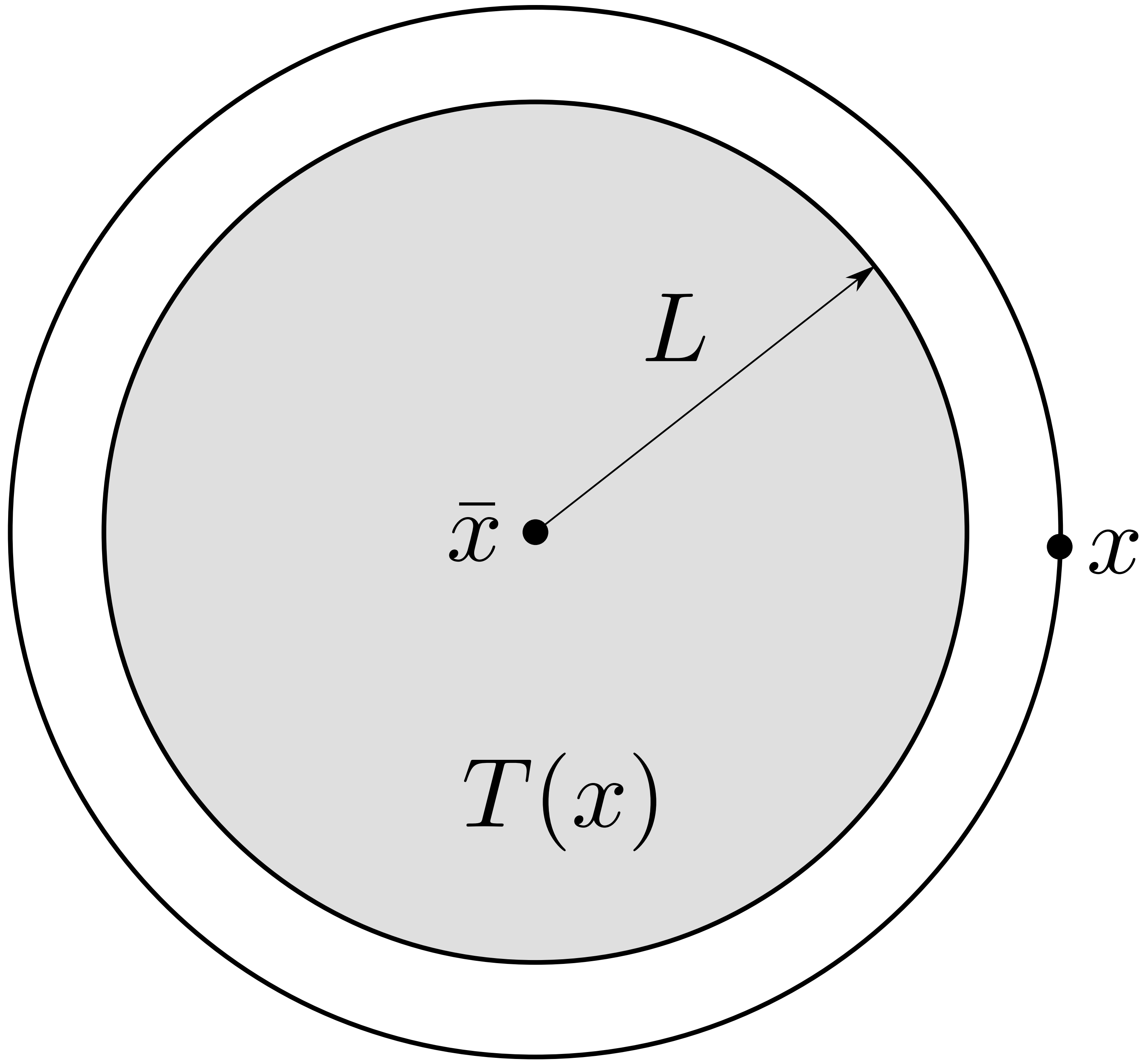
Lipschitz operators and fixed points

Given a L-Lipschitz operator T and a fixed point $\bar{x} = T\bar{x}$,

$$\|Tx - \bar{x}\| = \|Tx - T\bar{x}\| \leq L\|x - \bar{x}\|$$

A contractive operator ($L < 1$) can have at most one fixed point, i.e., $\mathbf{fix}\, T = \{\bar{x}\}$

Proof
If $\bar{x}, \bar{y} \in \mathbf{fix}\, T$ and $\bar{x} \neq \bar{y}$ then
$\|\bar{x} - \bar{y}\| = \|T(\bar{x}) - T(\bar{y})\| < \|\bar{x} - \bar{y}\|$ (contradiction) ■



A nonexpansive operator ($L = 1$) need not have a fixed point

Example $T(x) = x + 2$

Combining Lipschitz operators

T_1 is L_1-Lipschitz and T_2 is L_2-Lipschitz

Combining Lipschitz operators

T_1 is L_1-Lipschitz and T_2 is L_2-Lipschitz

The **composition** T_1T_2 is L_1L_2-Lipschitz

Proof $\|T_1T_2x - T_1T_2y\|_2 \leq L_1\|T_2x - T_2y\|_2 \leq L_1L_2\|x - y\|_2$ ■

- Composition of *nonexpansive* is nonexpansive
- Composition of *nonexpansive* and *contractive* is contractive

Combining Lipschitz operators

T_1 is L_1-Lipschitz and T_2 is L_2-Lipschitz

The **composition** $T_1 T_2$ is $L_1 L_2$-Lipschitz

Proof $\|T_1 T_2 x - T_1 T_2 y\|_2 \leq L_1 \|T_2 x - T_2 y\|_2 \leq L_1 L_2 \|x - y\|_2$ ■

- Composition of *nonexpansive* is nonexpansive
- Composition of *nonexpansive* and *contractive* is contractive

The **weighted average** $\theta T_1 + (1-\theta) T_2,\ \theta \in (0,1)$ is $(\theta L_1 + (1-\theta) L_2)$-Lipschitz

Proof (exercise)

- Weighted average of *nonexpansive* is nonexpansive
- Weighted average of *nonexpansive* and *contractive* is contractive

Fixed point iterations

Fixed point iteration

Apply operator

$$x^{k+1} = T(x^k)$$

until you reach $\bar{x} \in \mathbf{fix}\, T$

Fixed point iteration

Apply operator

$$x^{k+1} = T(x^k)$$

until you reach $\bar{x} \in \mathbf{fix}\, T$

Main approach

1. Find a suitable T such that $\bar{x} \in \mathbf{fix}\, T$ solve your problem
2. Show that the fixed point iteration converges

Fixed point iteration

Apply operator

$$x^{k+1} = T(x^k)$$

until you reach $\bar{x} \in \mathbf{fix}\, T$

Main approach

1. Find a suitable T such that $\bar{x} \in \mathbf{fix}\, T$ solve your problem
2. Show that the fixed point iteration converges

Fixed point residual to terminate

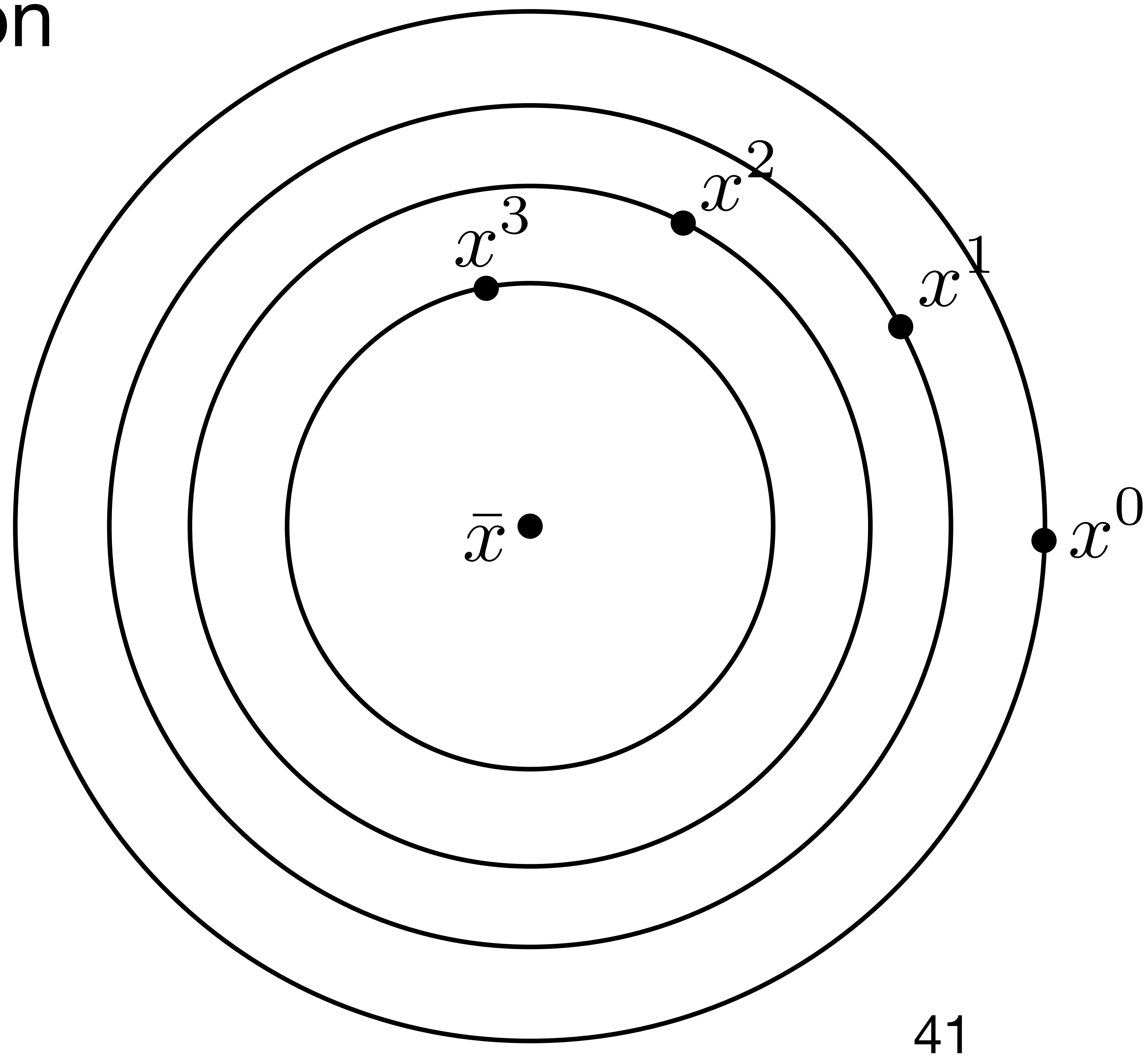
$$r^k = T(x^k) - x^k$$

Contractive fixed point iterations

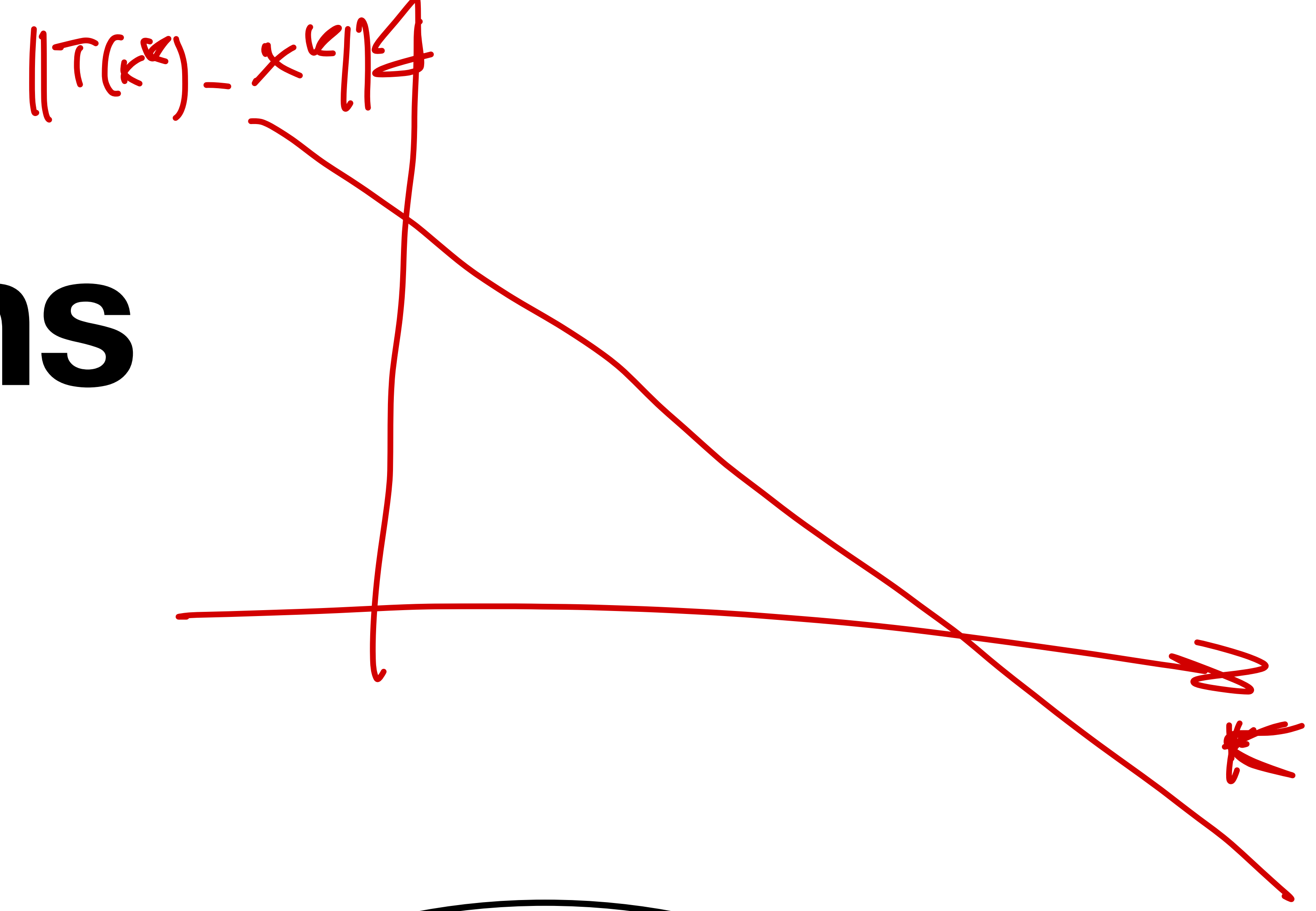
Contraction mapping theorem
If T is L-Lipschitz with $L < 1$ (contraction), the iteration

$$x^{k+1} = T(x^k)$$

converges to $\bar{x}$, the unique fixed point of T



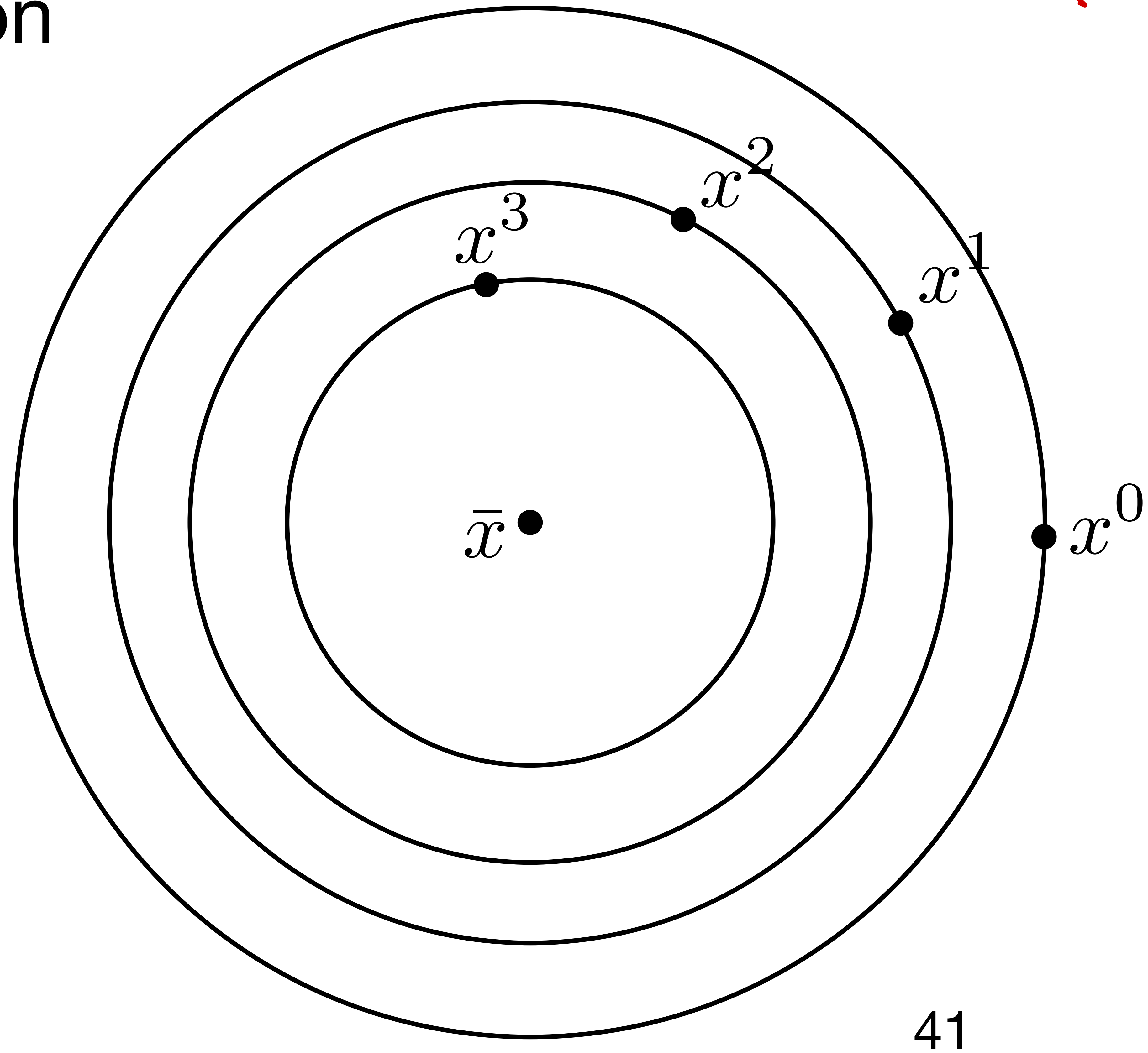
Contractive fixed point iterations



Contraction mapping theorem
If T is L-Lipschitz with $L < 1$ (contraction), the iteration

$$x^{k+1} = T(x^k)$$

converges to $\bar{x}$, the unique fixed point of T



Properties

- Distance to $\bar{x}$ decreases at each step

$$\|x^{k+1} - \bar{x}\| \leq L\|x^k - \bar{x}\|$$

  (iteration is **Fejer monotone**)

- Linear convergence rate L

Contraction mapping theorem

Proof

The sequence x^k is Cauchy

$$
\begin{aligned}
\|x^{k+\ell} - x^k\| &\le \|x^{k+\ell} - x^{k+\ell-1}\| + \cdots + \|x^{k+1} - x^k\| \\
&\le (L^{\ell-1} + \cdots + 1)\|x^{k+1} - x^k\| && \text{(Lipschitz constant)} \\
&\le \frac{1}{1-L}\|x^{k+1} - x^k\| && \text{(geometric series)} \\
&\le \frac{L^k}{1-L}\|x^1 - x^0\| && \text{(Lipschitz constant)}
\end{aligned}
$$

Contraction mapping theorem

Proof

The sequence x^k is Cauchy

$$
\begin{aligned}
\|x^{k+\ell} - x^k\| &\leq \|x^{k+\ell} - x^{k+\ell-1}\| + \cdots + \|x^{k+1} - x^k\| \\
&\leq (L^{\ell-1} + \cdots + 1)\|x^{k+1} - x^k\| && \text{(Lipschitz constant)} \\
&\leq \frac{1}{1-L}\|x^{k+1} - x^k\| && \text{(geometric series)} \\
&\leq \frac{L^k}{1-L}\|x^1 - x^0\| && \text{(Lipschitz constant)}
\end{aligned}
$$

Therefore it converges to a point $\bar{x}$ which must be the (unique) fixed point of T

Contraction mapping theorem

Proof

The sequence x^k is Cauchy

$$
\begin{aligned}
\|x^{k+\ell} - x^k\| &\leq \|x^{k+\ell} - x^{k+\ell-1}\| + \cdots + \|x^{k+1} - x^k\| \\
&\leq (L^{\ell-1} + \cdots + 1)\|x^{k+1} - x^k\| && \text{(Lipschitz constant)} \\
&\leq \frac{1}{1-L}\|x^{k+1} - x^k\| && \text{(geometric series)} \\
&\leq \frac{L^k}{1-L}\|x^1 - x^0\| && \text{(Lipschitz constant)}
\end{aligned}
$$

Therefore it converges to a point $\bar{x}$ which must be the (unique) fixed point of T

The convergence is linear (geometric) with rate L

$$\|x^k - \bar{x}\| = \|T(x^{k-1}) - T(\bar{x})\| \leq L\|x^{k-1} - \bar{x}\| \leq L^k\|x^0 - x^\star\|$$

■

Nonexpansive fixed point iterations

If T is L-Lipschitz with $L = 1$ (nonexpansive), the iteration

$$x^{k+1} = T(x^k)$$

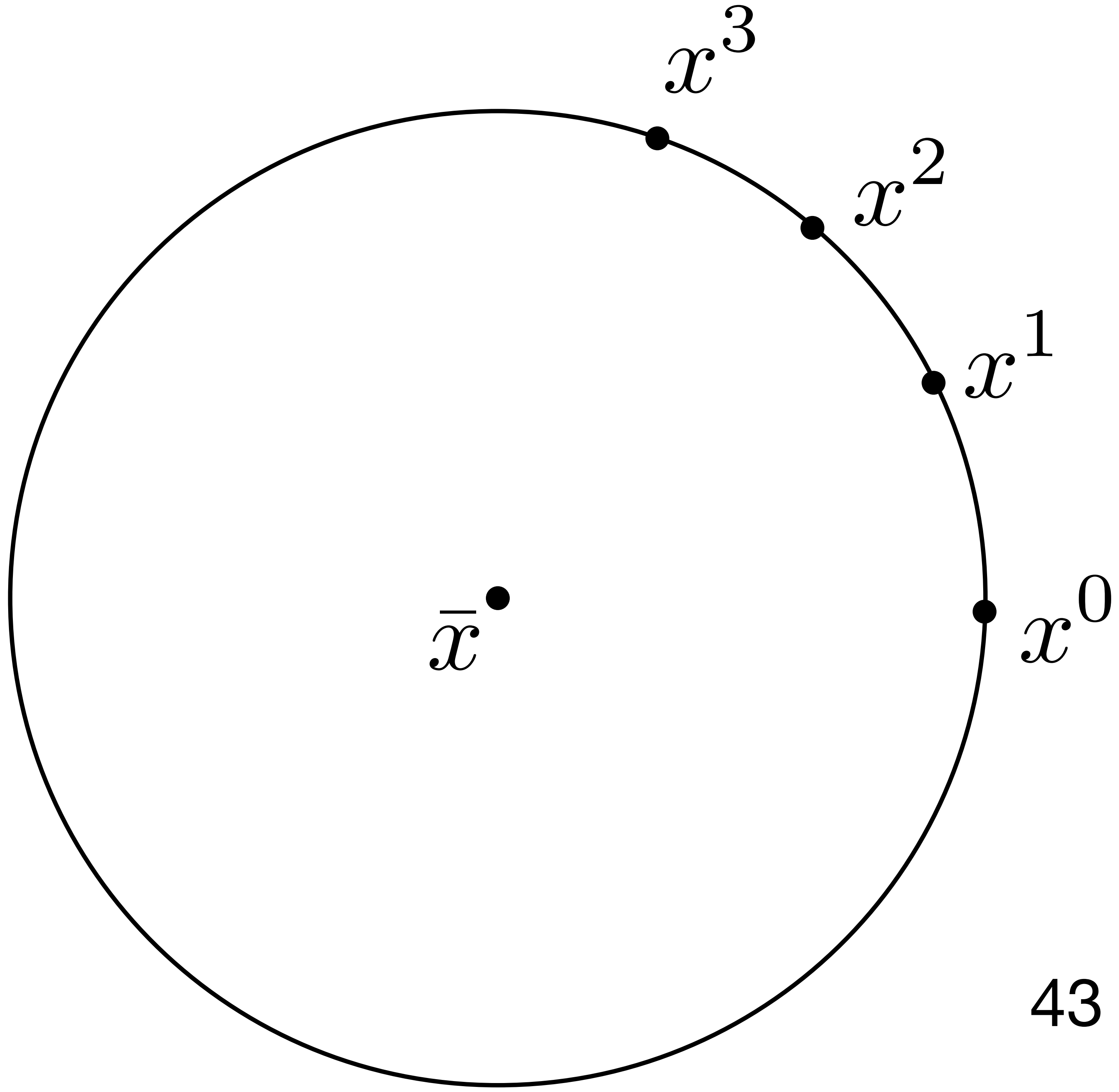
need not converge to a fixed point, even if one exists.

Nonexpansive fixed point iterations

If T is L-Lipschitz with $L = 1$ (nonexpansive), the iteration

$$x^{k+1} = T(x^k)$$

need not converge to a fixed point, even if one exists.



Example

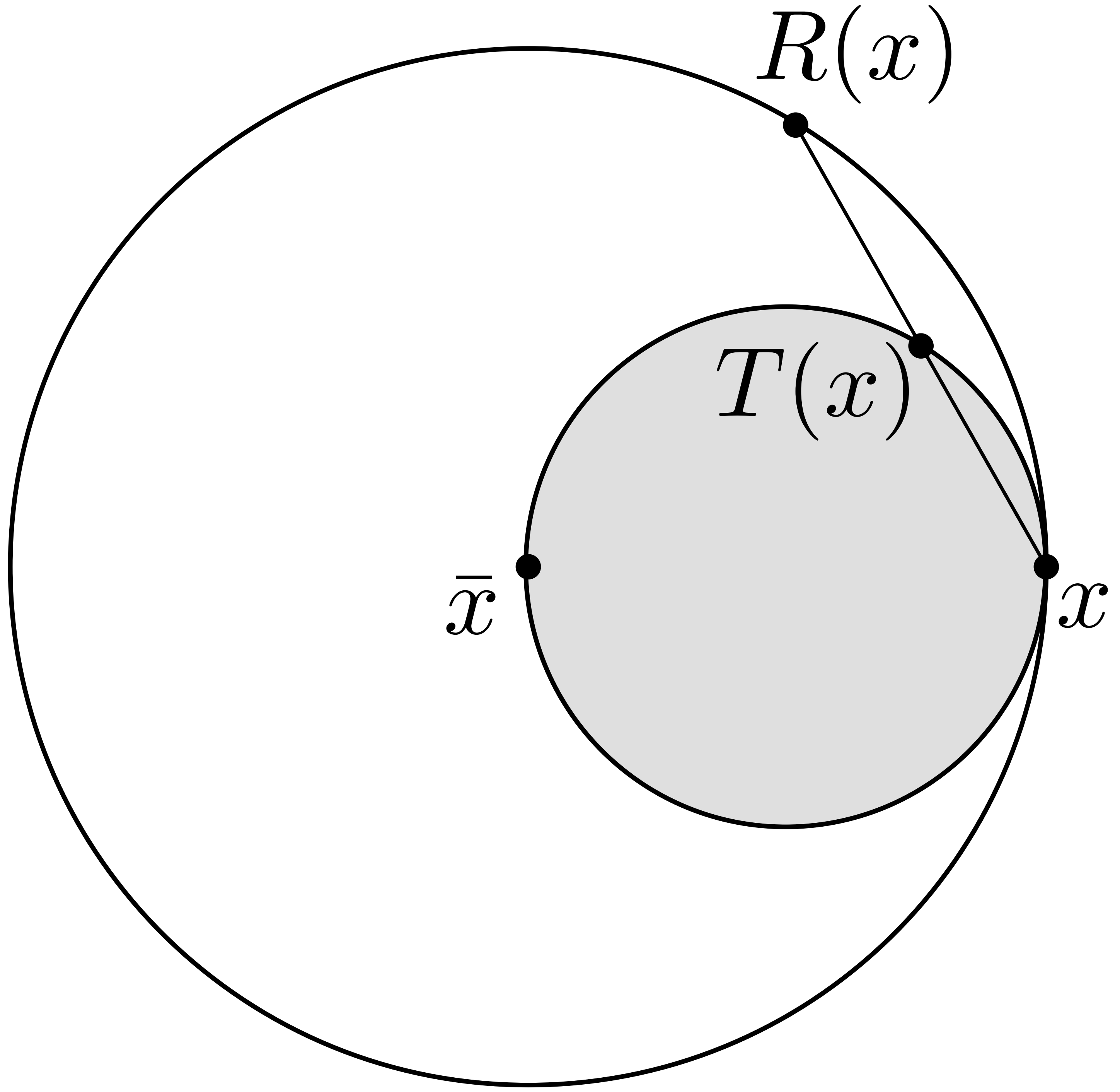
- Let T be a rotation around the origin
- T is nonexpansive and has a fixed point $\bar{x} = 0$
- $\|x^k\|$ never decreases

Averaged operators

We say that an operator T is $\alpha-$**averaged** with $\alpha \in (0,1)$ if

$$T = (1-\alpha)I + \alpha R$$

and R is nonexpansive.

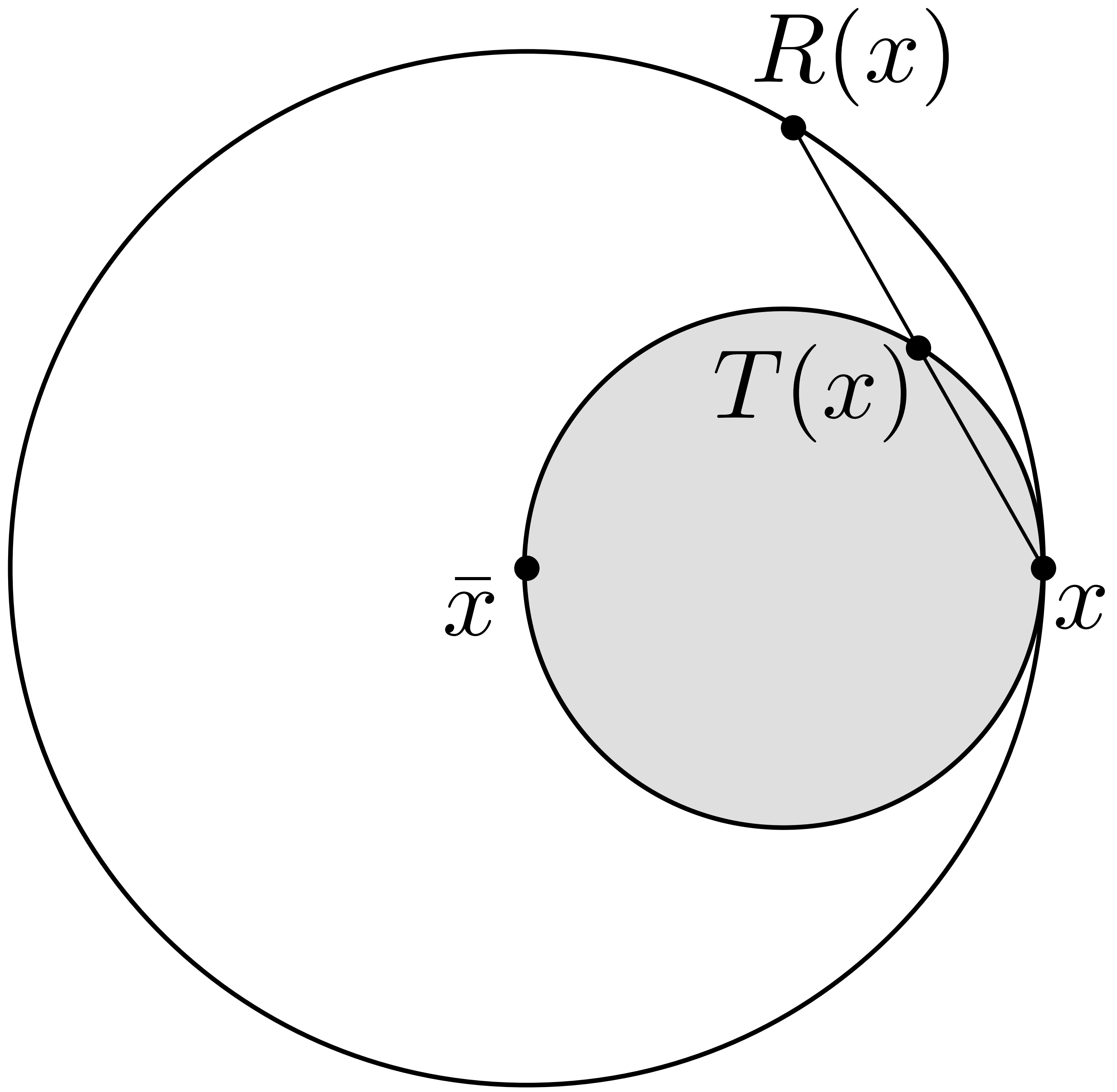


Averaged operators

We say that an operator T is $\alpha-$**averaged** with $\alpha \in (0,1)$ if

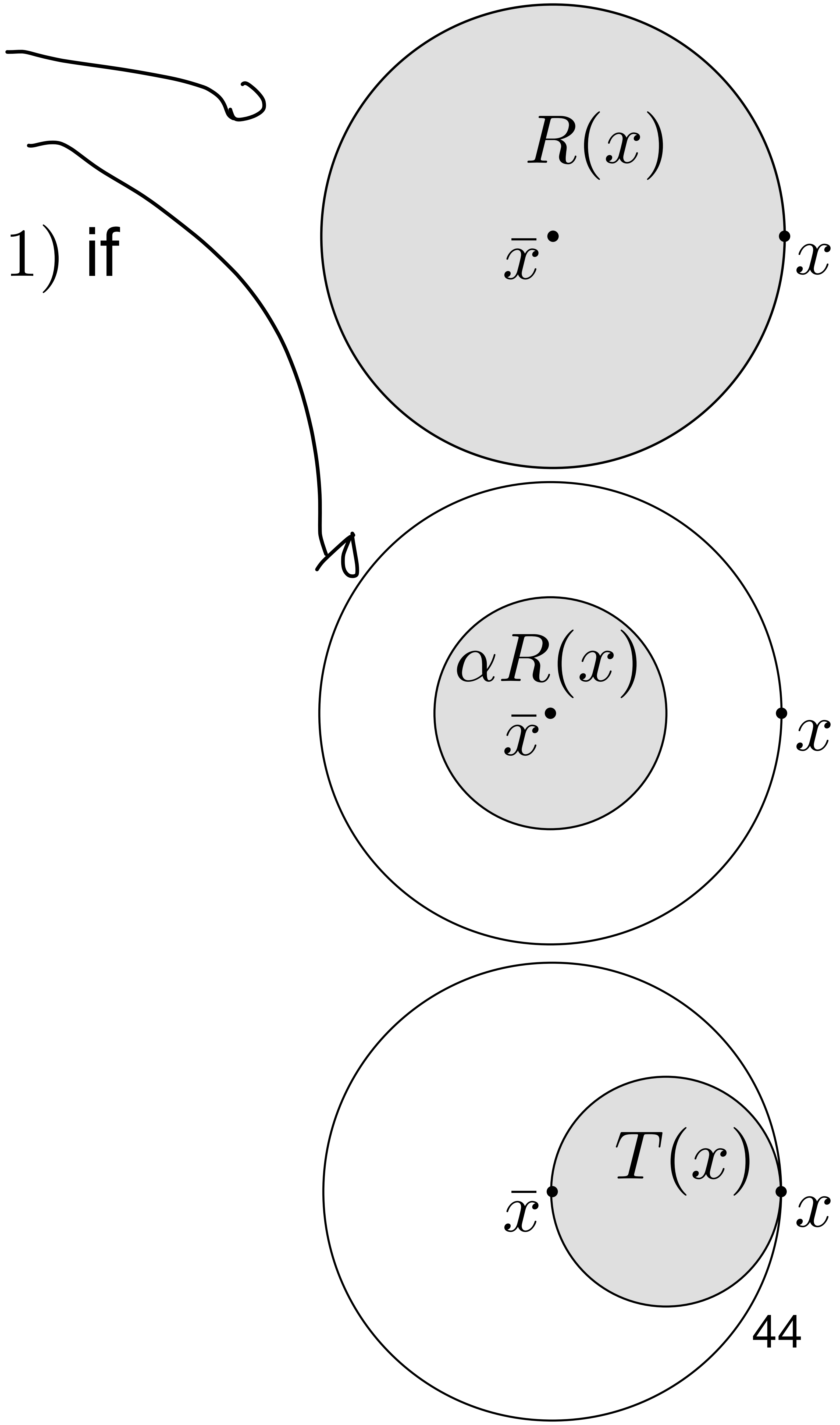
$$T = (1-\alpha)I + \alpha R$$

and R is nonexpansive.



Example $\alpha = 1/2$

SCALED RELATIVE GRAPH



Averaged operators fixed points

We say that an operator T is $\alpha-$**averaged** with $\alpha \in (0,1)$ if

$$T = (1-\alpha)I + \alpha R$$

Averaged operators fixed points

We say that an operator T is $\alpha-$**averaged** with $\alpha \in (0,1)$ if

$$T = (1-\alpha)I + \alpha R$$

Fact If T is α-averaged, then $\mathbf{fix}\, T = \mathbf{fix}\, R$

Proof
$$\begin{aligned} \bar{x} = T(\bar{x}) &= (1-\alpha)I(\bar{x}) + \alpha R(\bar{x}) \\ &= (1-\alpha)\bar{x} + \alpha R(\bar{x}) \\ \iff \quad & \alpha\bar{x} = \alpha R(\bar{x}) \\ \iff \quad & \bar{x} = R(\bar{x}) \end{aligned}$$

$\blacksquare$

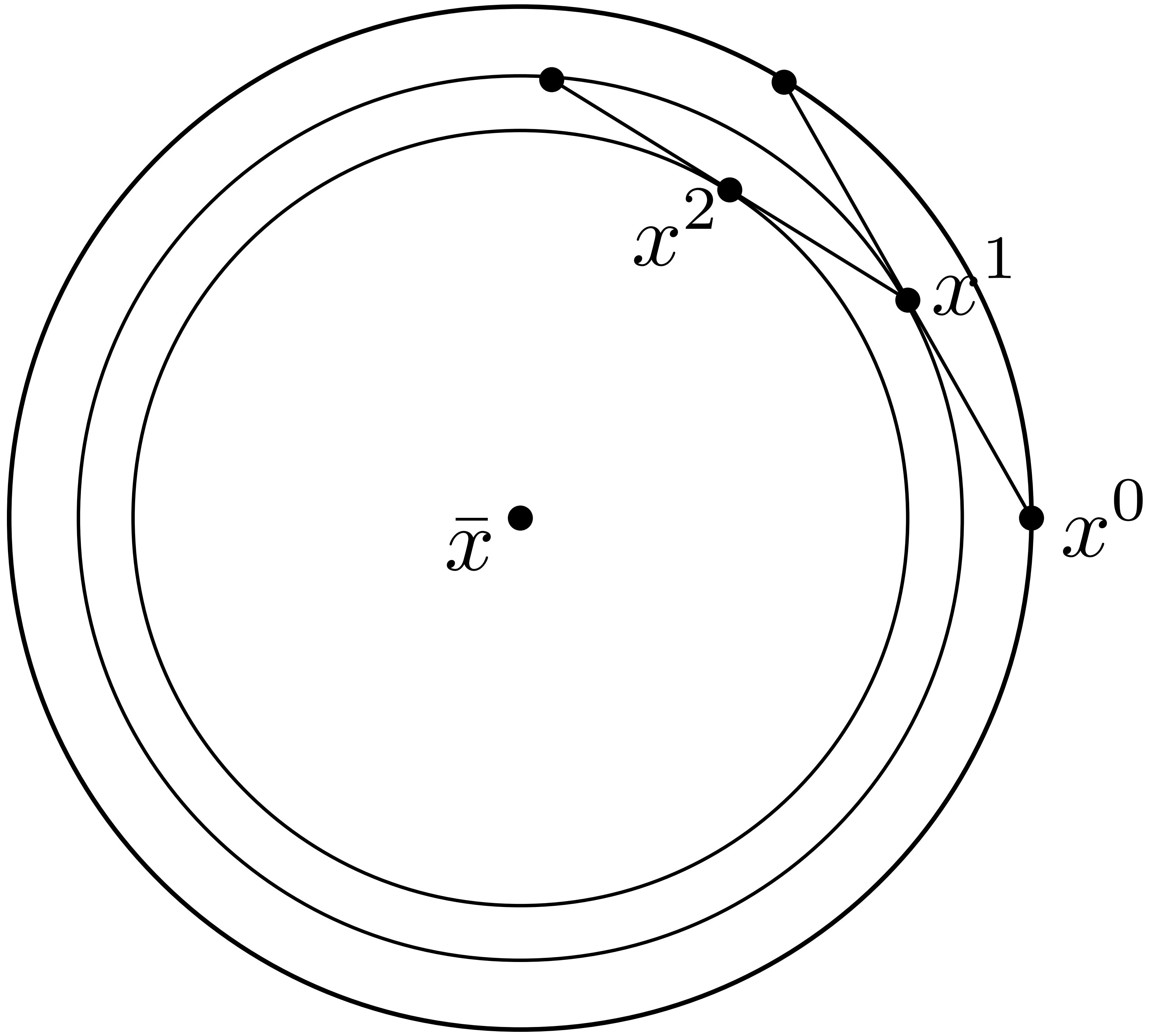
Averaged fixed point iterations

If $T = (1-\alpha)I + \alpha R$ is α-averaged
($\alpha \in (0,1)$ and R nonexpansive), the iteration

$$x^{k+1} = T(x^k)$$

converges to $\bar{x} \in \mathbf{fix}\, T$

(also called damped, averaged
or Mann-Krasnosel'skii iteration)



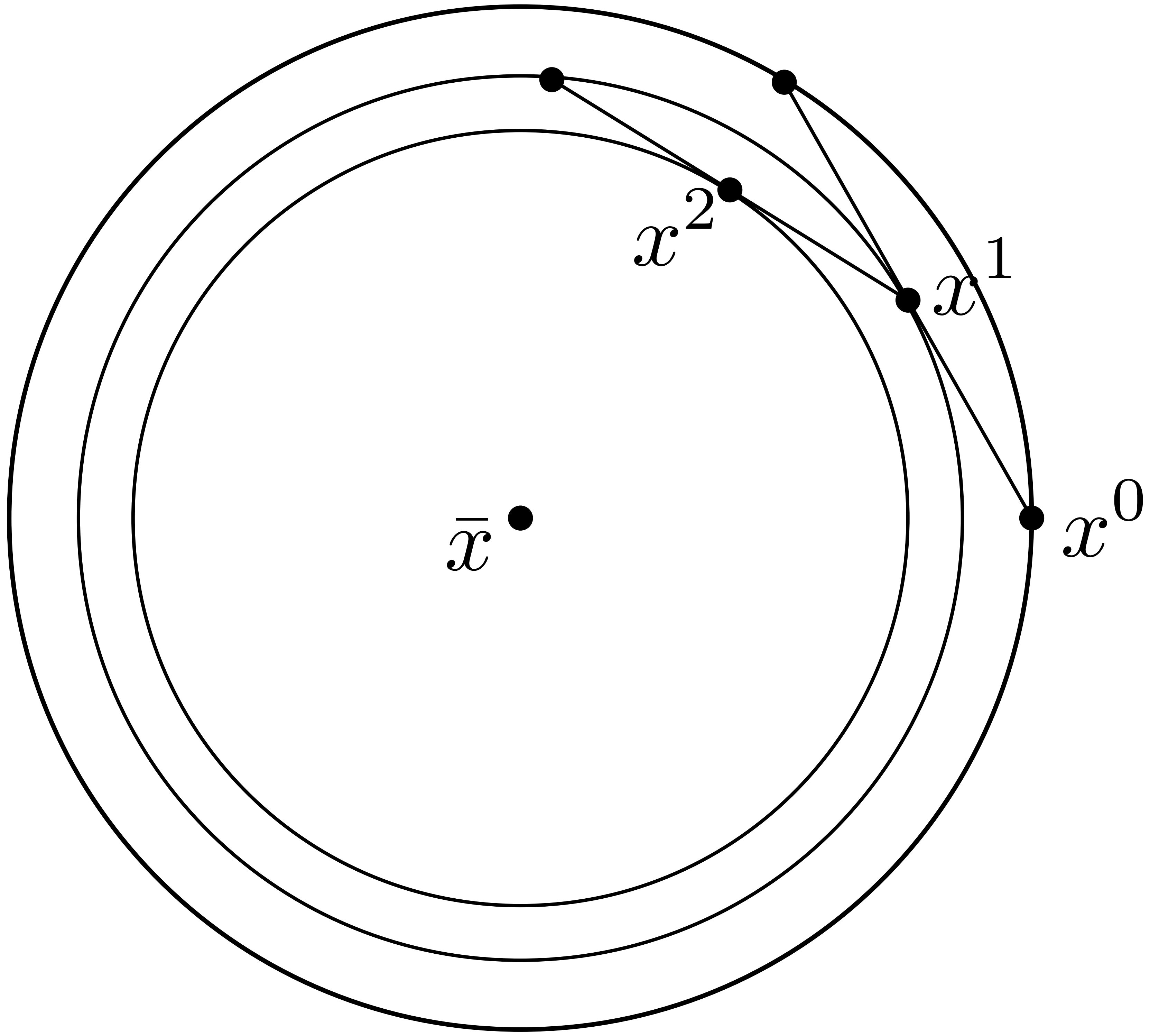
Averaged fixed point iterations

If $T = (1-\alpha)I + \alpha R$ is α-averaged
($\alpha \in (0,1)$ and R nonexpansive), the iteration

$$x^{k+1} = T(x^k)$$

converges to $\bar{x} \in \mathbf{fix}\, T$

(also called damped, averaged
or Mann-Krasnosel'skii iteration)



Properties

- Distance to $\bar{x}$ decreases at each step (**Fejer monotone**)
- Sublinear convergence to fixed-point residual

$$\|R(x^k) - x^k\| \leq \frac{1}{\sqrt{(k+1)\alpha(1-\alpha)}}\|x^0 - \bar{x}\|$$

Averaged fixed point iterations

Proof

Use the identity (proof by expanding)

$$\|(1-\alpha)a + \alpha b\|^2 = (1-\alpha)\|a\|^2 + \alpha\|b\|^2 - \alpha(1-\alpha)\|a-b\|^2$$

and apply it to

$$x^{k+1} - \bar{x} = (1-\alpha)(x^k - \bar{x}) + \alpha(R(x^k) - \bar{x})$$

Averaged fixed point iterations

Proof

Use the identity (proof by expanding)

$$\|(1-\alpha)a + \alpha b\|^2 = (1-\alpha)\|a\|^2 + \alpha\|b\|^2 - \alpha(1-\alpha)\|a-b\|^2$$

and apply it to

$$x^{k+1} - \bar{x} = (1-\alpha)(\underbrace{x^k - \bar{x}}_{a}) + \alpha(\underbrace{R(x^k) - \bar{x}}_{b})$$

Averaged fixed point iterations

Proof

Use the identity (proof by expanding)

$$\|(1-\alpha)a + \alpha b\|^2 = (1-\alpha)\|a\|^2 + \alpha\|b\|^2 - \alpha(1-\alpha)\|a-b\|^2$$

and apply it to

$$x^{k+1} - \bar{x} = (1-\alpha)(\underbrace{x^k - \bar{x}}_{a}) + \alpha(\underbrace{R(x^k) - \bar{x}}_{b})$$

obtaining

$$\begin{aligned}
\|x^{k+1} - \bar{x}\|^2 &= (1-\alpha)\|x^k - \bar{x}\|^2 + \alpha\|R(x^k) - \bar{x}\|^2 - \alpha(1-\alpha)\|x^k - R(x^k)\|^2 \\
&\le (1-\alpha)\|x^k - \bar{x}\|^2 + \alpha\|x^k - \bar{x}\|^2 - \alpha(1-\alpha)\|x^k - R(x^k)\|^2 \quad \text{(nonexpansive)} \\
&= \|x^k - \bar{x}\|^2 - \alpha(1-\alpha)\|x^k - R(x^k)\|^2
\end{aligned}$$

Averaged fixed point iterations

Proof

Use the identity (proof by expanding)

$$\|(1-\alpha)a + \alpha b\|^2 = (1-\alpha)\|a\|^2 + \alpha\|b\|^2 - \alpha(1-\alpha)\|a-b\|^2$$

and apply it to

$$x^{k+1} - \bar{x} = (1-\alpha)(\underbrace{x^k - \bar{x}}_{a}) + \alpha(\underbrace{R(x^k) - \bar{x}}_{b})$$

obtaining

$$\begin{aligned}\|x^{k+1} - \bar{x}\|^2 &= (1-\alpha)\|x^k - \bar{x}\|^2 + \alpha\|R(x^k) - \bar{x}\|^2 - \alpha(1-\alpha)\|x^k - R(x^k)\|^2 \\ &\leq (1-\alpha)\|x^k - \bar{x}\|^2 + \alpha\|x^k - \bar{x}\|^2 - \alpha(1-\alpha)\|x^k - R(x^k)\|^2 \quad \text{(nonexpansive)} \\ &= \|x^k - \bar{x}\|^2 \underbrace{- \alpha(1-\alpha)\|x^k - R(x^k)\|^2}_{\leq 0}\end{aligned}$$

Iterations are Fejer monotone

Averaged fixed point iterations

Proof (continued)

iterate righthand side over k steps

$$\|x^{k+1} - \bar{x}\|^2 \leq \|x^0 - \bar{x}\|^2 - \alpha(1-\alpha)\sum_{i=0}^{k}\|x^i - R(x^i)\|^2$$

Averaged fixed point iterations

Proof (continued)

iterate righthand side over k steps

$$\|x^{k+1} - \bar{x}\|^2 \leq \|x^0 - \bar{x}\|^2 - \alpha(1-\alpha)\sum_{i=0}^{k}\|x^i - R(x^i)\|^2$$

Since $\|x^{k+1} - \bar{x}\|^2 \geq 0$, we have

$$\sum_{i=0}^{k}\|x^i - R(x^i)\|^2 \leq \frac{1}{\alpha(1-\alpha)}\|x^0 - \bar{x}\|^2$$

Averaged fixed point iterations

Proof (continued)

iterate righthand side over k steps

$$\|x^{k+1} - \bar{x}\|^2 \leq \|x^0 - \bar{x}\|^2 - \alpha(1-\alpha)\sum_{i=0}^{k}\|x^i - R(x^i)\|^2$$

Since $\|x^{k+1} - \bar{x}\|^2 \geq 0$, we have

$$\sum_{i=0}^{k}\|x^i - R(x^i)\|^2 \leq \frac{1}{\alpha(1-\alpha)}\|x^0 - \bar{x}\|^2$$

Using $\sum_{i=0}^{k}\|x^i - R(x^i)\|^2 \geq (k+1)\min_{i=0,\dots,k}\|x^i - R(x^i)\|^2$, we obtain

$$\min_{i=0,\dots,k}\|x^i - R(x^i)\|^2 \leq \frac{1}{(k+1)\alpha(1-\alpha)}\|x^0 - \bar{x}\|^2$$

Averaged fixed point iterations

Proof (continued)

iterate righthand side over k steps

$$\|x^{k+1} - \bar{x}\|^2 \leq \|x^0 - \bar{x}\|^2 - \alpha(1-\alpha)\sum_{i=0}^{k}\|x^i - R(x^i)\|^2$$

Since $\|x^{k+1} - \bar{x}\|^2 \geq 0$, we have $\displaystyle\sum_{i=0}^{k}\|x^i - R(x^i)\|^2 \leq \frac{1}{\alpha(1-\alpha)}\|x^0 - \bar{x}\|^2$

Using $\displaystyle\sum_{i=0}^{k}\|x^i - R(x^i)\|^2 \geq (k+1)\min_{i=0,\dots,k}\|x^i - R(x^i)\|^2$, we obtain

$$\min_{i=0,\dots,k}\|x^i - R(x^i)\|^2 \leq \frac{1}{(k+1)\alpha(1-\alpha)}\|x^0 - \bar{x}\|^2$$

(R is nonexpansive $\rightarrow$ $\min$ at k) $\displaystyle\|x^k - R(x^k)\|^2 \leq \frac{1}{(k+1)\alpha(1-\alpha)}\|x^0 - \bar{x}\|^2$ ■

Average fixed point iteration convergence rates

$$\|R(x^k) - x^k\| \leq \frac{1}{\sqrt{(k+1)\alpha(1-\alpha)}}\|x^0 - \bar{x}\|$$

Average fixed point iteration convergence rates

$$\|R(x^k) - x^k\| \leq \frac{1}{\sqrt{(k+1)\alpha(1-\alpha)}}\|x^0 - \bar{x}\|$$

Righthand side minimized when $\alpha = 1/2$

$$\|R(x^k) - x^k\| \leq \frac{2}{\sqrt{k+1}}\|x^0 - \bar{x}\|$$

Iterations

$$x^{k+1} = (1/2)x^k + (1/2)R(x^k)$$

Average fixed point iteration convergence rates

$$\|R(x^k) - x^k\| \leq \frac{1}{\sqrt{(k+1)\alpha(1-\alpha)}}\|x^0 - \bar{x}\|$$

Righthand side minimized when $\alpha = 1/2$

$$\|R(x^k) - x^k\| \leq \frac{2}{\sqrt{k+1}}\|x^0 - \bar{x}\|$$

Iterations

$$x^{k+1} = (1/2)x^k + (1/2)R(x^k)$$

Remarks

- Sublinear convergence (same as subgrad method), in general not the actual rate
- $\alpha = 1/2$ is very common for averaged operators

How to design an algorithm

Problem

minimize $\quad f(x)$

Algorithm (operator) construction

1. Find a suitable T such that $\bar{x} \in \mathbf{fix}\, T$ solve your problem
2. Show that the fixed point iteration converges

How to design an algorithm

Problem

minimize $\quad f(x)$

Algorithm (operator) construction

1. Find a suitable T such that $\bar{x} \in \mathbf{fix}\, T$ solve your problem
2. Show that the fixed point iteration converges

If T is contractive $\quad \Longrightarrow \quad$ **linear convergence**
If T is averaged $\quad \Longrightarrow \quad$ **sublinear convergence**

Most first order algorithms can be constructed in this way

Proximal methods and introduction to operators

Today, we learned to:

- **Derive** optimality conditions for constrained optimization problems using subdifferentials
- **Define** and **evaluate** proximal operators for various common functions
- **Apply** proximal operators to generalize gradient descent (vanilla, projected, proximal)
- **Use operator theory** to construct general fixed-point iterations and prove their convergence

Next lecture

- Monotone operators and operator splitting algorithms