ORF522 – Linear and Nonlinear Optimization

17. Operator theory

Bartolomeo Stellato — Fall 2021

Ed Forum

- What is the advantage of (1-alpha)*I + alpha* R over simply the alpha * R? it seems alpha*R also transforms the original nonexpansive function R into a contractive one, so why bother adding another term I?
- What do the graphs mean for the averaged operator? Does the domain of R(x) get adjusted by a factor of alpha around the fixed point, and then shifted by a factor of (1-alpha)*I so that it encompasses both x and the fixed point, and this becomes T(x)?
- Slide 28 uses the phrase, "component-wise soft-thresholding;" what does that mean, as opposed to not component-wise?
- Throughout the lecture, I think it was mentioned that some things are "hard but cheap" or "expensive." In this context, such as slide 26 and 27, what is the difference between hard and expensive?

Recap

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., $g(x) = \sum_{i=1}^{N} g_i(x_i)$

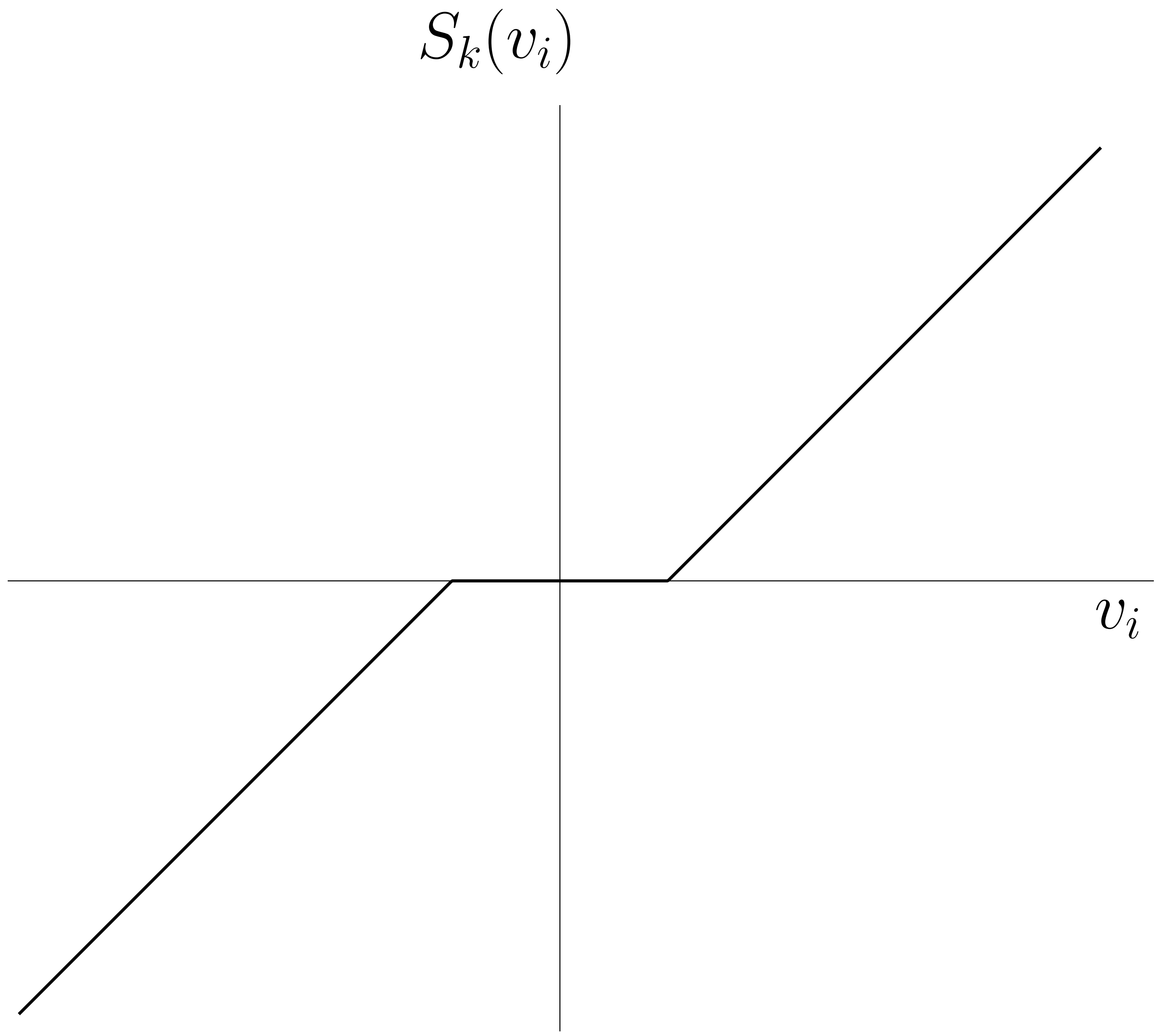
then, $(\mathbf{prox}_g(v))_i = \mathbf{prox}_{g_i}(v_i), \quad i = 1, \ldots, N$

(key to parallel/distributed proximal algorithms)

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

soft-thresholding

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

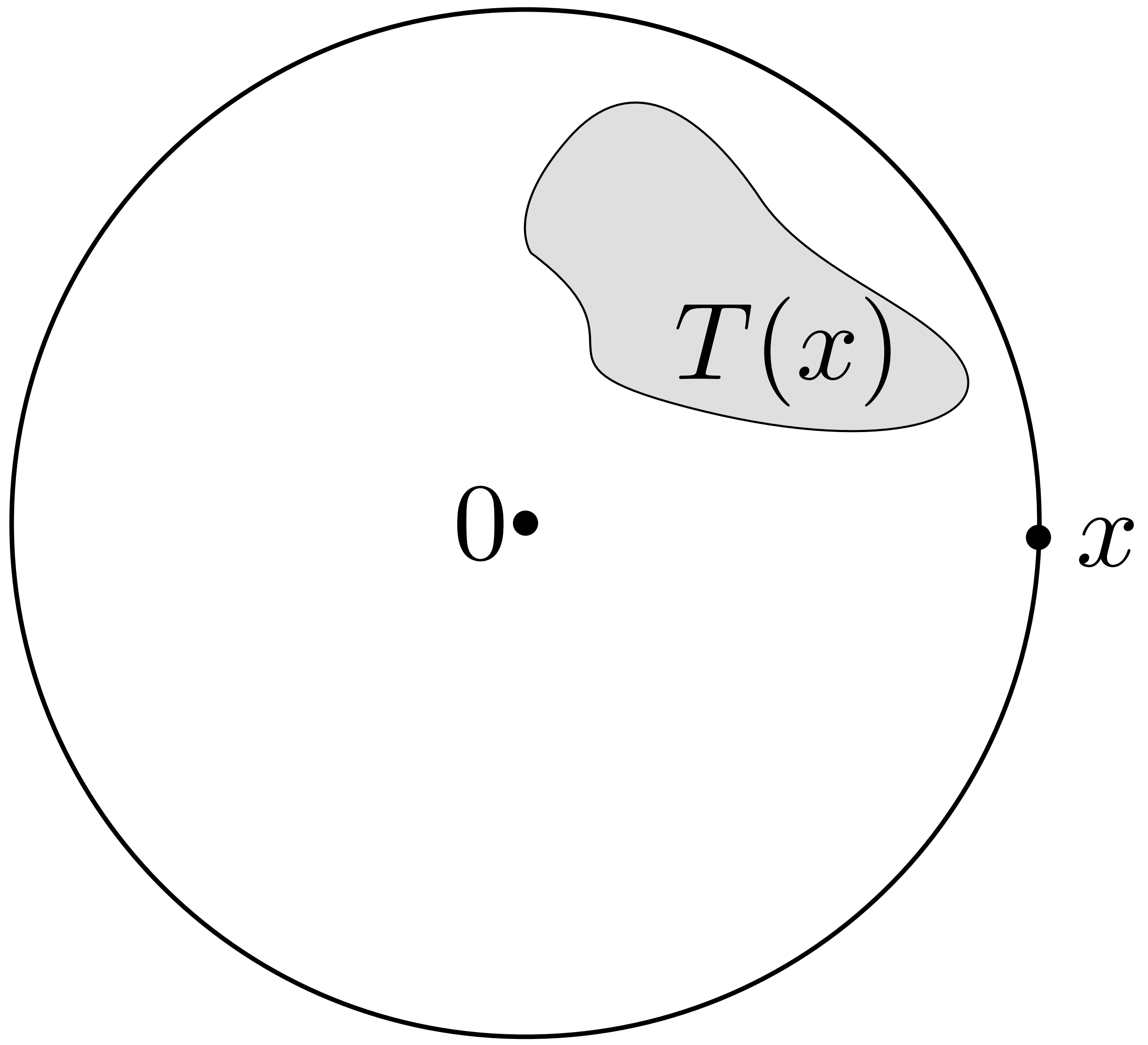


Operators

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

Zeros

Zero

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

Zero set

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

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

Fixed points

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

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

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

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

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

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

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

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

Lipschitz operators and fixed points

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

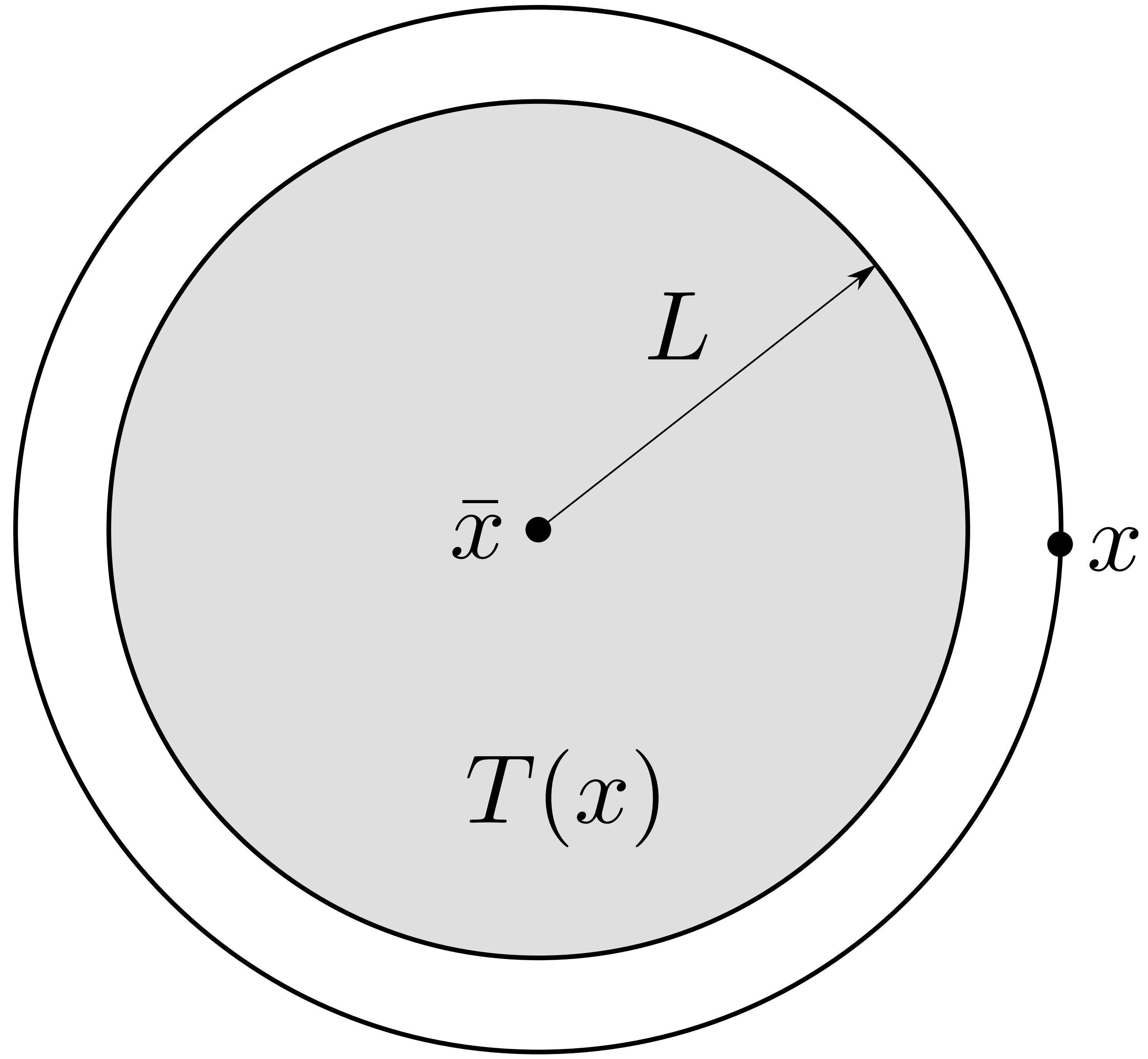
$$\|Tx - \bar{x}\| = \|Tx - T\bar{x}\| \leq L\|x - \bar{x}\|$$

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

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



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

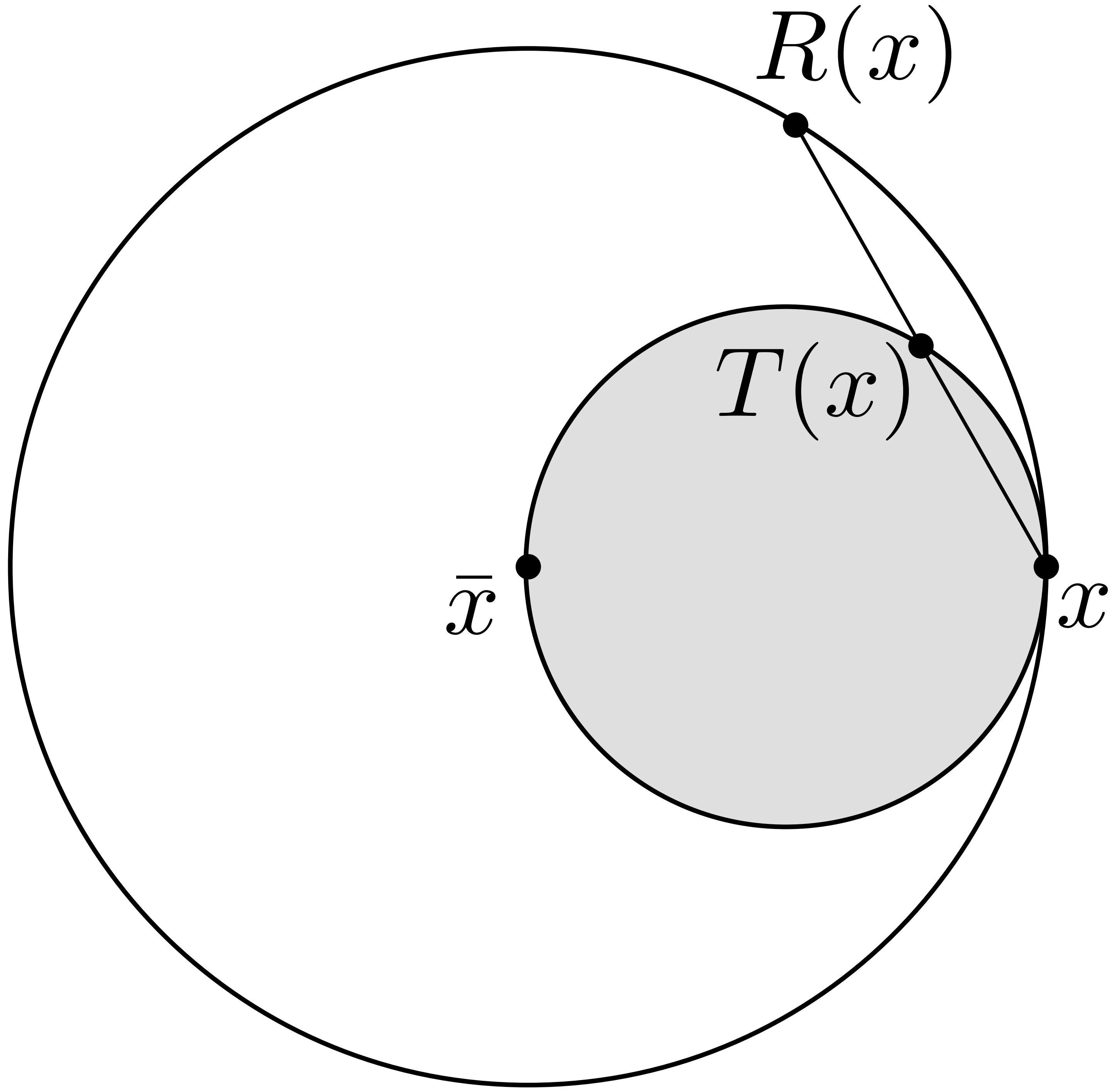
Example $T(x) = x + 2$

Averaged operators

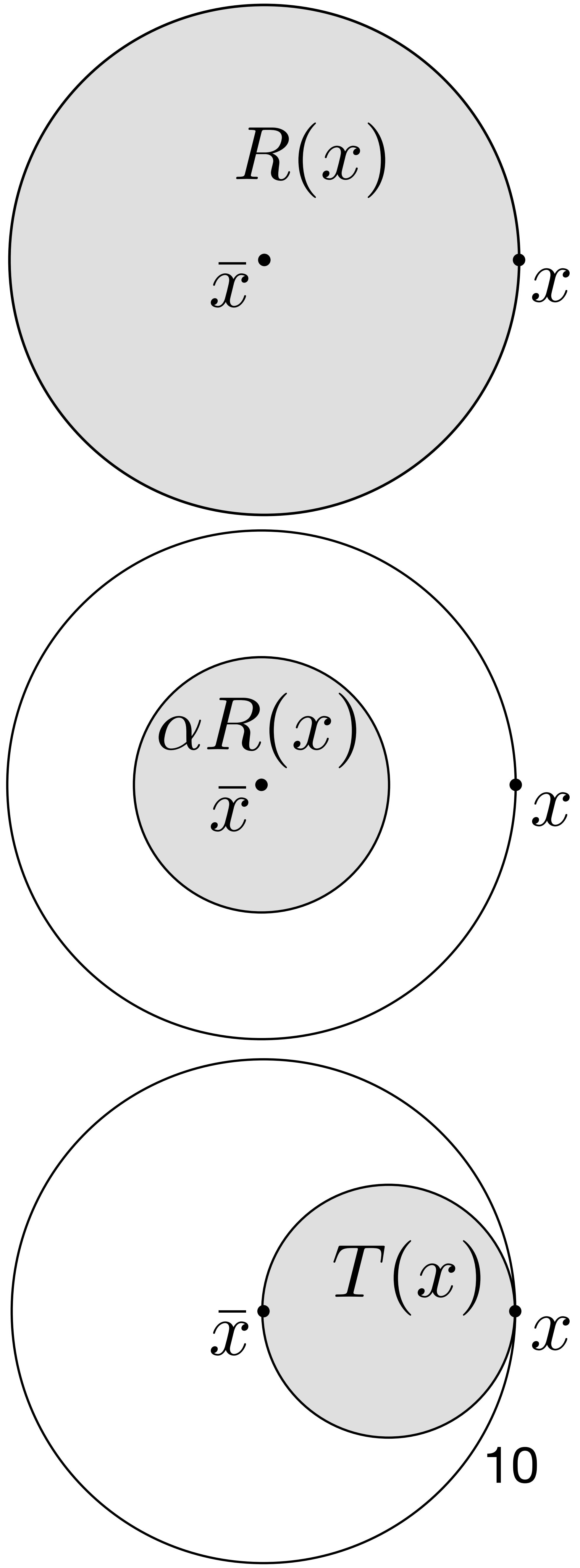
We say that an operator T is $\alpha-$**averaged** with $\alpha \in (0,1)$ if

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

and R is nonexpansive.



Example $\alpha = 1/2, \bar{x} = 0$



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 $\implies$ **linear convergence**
If T is averaged $\implies$ **sublinear convergence**

Most first order algorithms can be constructed in this way

Today's lecture

[Chapter 4, FMO][PA][PMO][LSMO]

Monotone operators

- Conjugate functions and duality
- Monotone and cocoercive operators
- Subdifferential operator and monotonicity
- Operators in optimization problems
- Operators in algorithms
- Building contractions

Conjugate functions and duality

Convex closed proper functions

A function $f : \mathbf{R}^n \to \mathbf{R}$ is called **CCP** if it is

closed $\quad \mathbf{epi}\, f$ is a closed set

convex $\quad f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \quad \alpha \in [0,1]$

proper $\quad \mathbf{dom}\, f$ is nonempty

If not otherwise stated, we assume functions to be **CCP**

Conjugate function

Given a function $f : \mathbf{R}^n \to \mathbf{R}$ we define its **conjugate** $f^* : \mathbf{R}^n \to \mathbf{R}$ as

$$f^*(y) = \max_x \; y^T x - f(x)$$

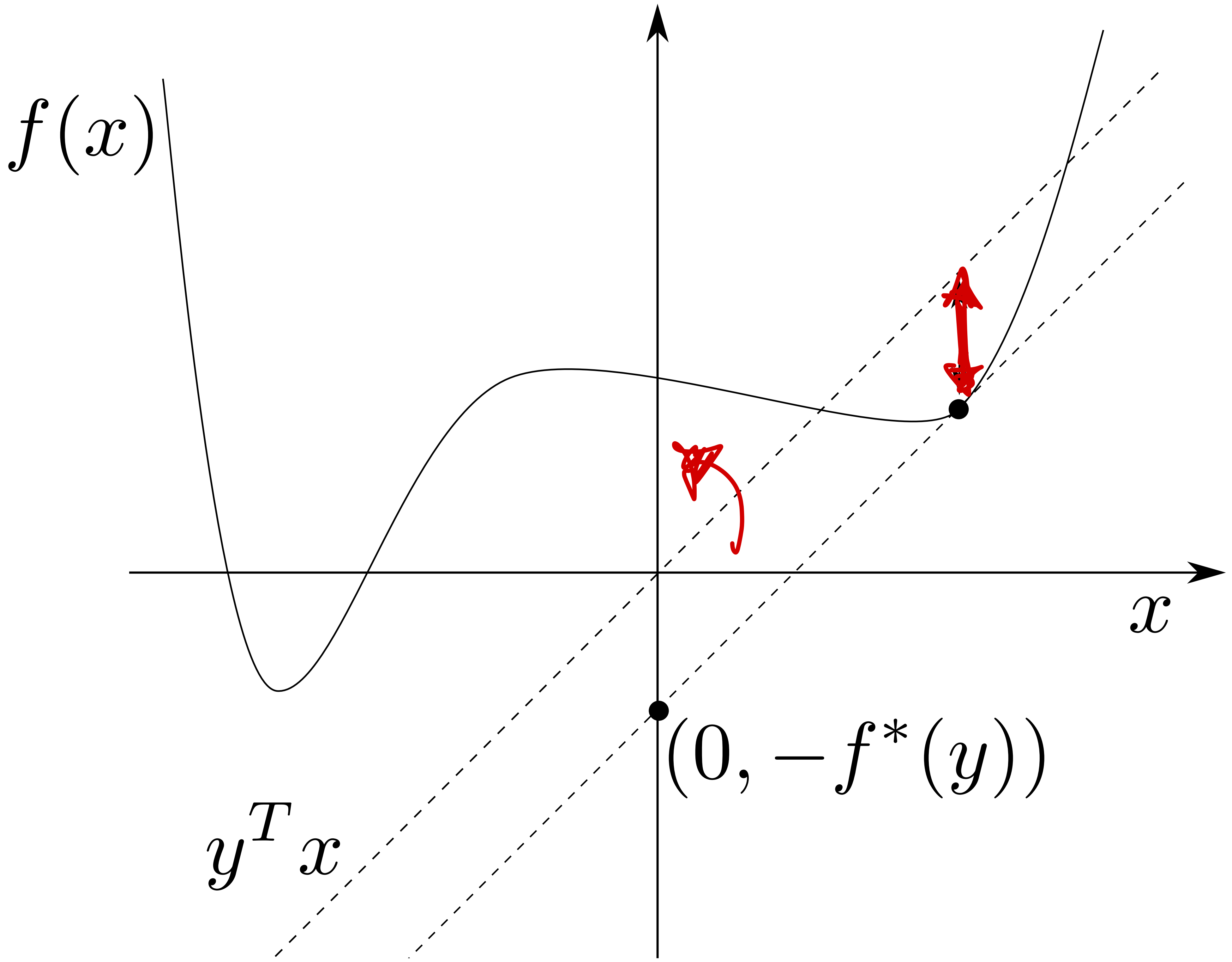
Note f^* is always convex (pointwise maximum of affine functions in y)

Conjugate function

Given a function $f : \mathbf{R}^n \to \mathbf{R}$ we define its **conjugate** $f^* : \mathbf{R}^n \to \mathbf{R}$ as

$$f^*(y) = \max_x \; y^T x - f(x)$$

Note f^* is always convex (pointwise maximum of affine functions in y)



f^* is the *maximum gap* between $y^T x$ and $f(x)$

$f^*(y) = \max_x \, y^T x - f(x)$

Conjugate function properties and examples

Properties

Fenchel's inequality $\quad f(x) + f^*(y) \geq y^T x \quad$ (from $\max$ inside conjugate)

Conjugate function properties and examples

Properties

Fenchel's inequality $f(x) + f^*(y) \geq y^T x$ (from $\max$ inside conjugate)

Biconjugate $f^{**}(y) = \max_x \; y^T x - f^*(x) \quad \Longrightarrow \quad f(x) \geq f^{**}(x)$

Conjugate function properties and examples

Properties

Fenchel's inequality $f(x) + f^*(y) \geq y^T x$ (from $\max$ inside conjugate)

Biconjugate $f^{**}(y) = \max_x \; y^T x - f^*(x) \implies f(x) \geq f^{**}(x)$

Biconjugate for CCP functions If f CCP, then $f^{**} = f$

Conjugate function properties and examples

Properties

Fenchel's inequality $f(x) + f^*(y) \geq y^T x$ (from $\max$ inside conjugate)

Biconjugate $f^{**}(y) = \max_x \; y^T x - f^*(x) \implies f(x) \geq f^{**}(x)$

Biconjugate for CCP functions If f CCP, then $f^{**} = f$

Examples

$\frac{1}{p} + 1/q = 1$

Norm $f(x) = \|x\|_p$; $f^*(y) = \mathcal{I}_{\|y\|_q \leq 1}(y)$ **indicator function of dual norm set**

Conjugate function properties and examples

Properties

Fenchel's inequality $f(x) + f^*(y) \geq y^T x$ (from $\max$ inside conjugate)

Biconjugate $f^{**}(y) = \max_x \; y^T x - f^*(x) \quad \Longrightarrow \quad f(x) \geq f^{**}(x)$

Biconjugate for CCP functions If f CCP, then $f^{**} = f$

Examples

Norm $f(x) = \|x\|$: $f^*(y) = \mathcal{I}_{\|y\|_* \leq 1}(y)$ **indicator function of dual norm set**

Indicator function $f(x) = \mathcal{I}_C(x)$: $f^*(y) = \mathcal{I}_C^*(y) = \max_{x \in C} \; y^T x = \sigma_C(y)$ **support function**

More examples of conjugate functions [Page 101, FMO]

Fenchel dual

Dual using conjugate functions

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

Equivalent form (variables split)

minimize $f(x) + g(z)$
subject to $x = z$

Fenchel dual

Dual using conjugate functions

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

Equivalent form (variables split)

$$\begin{array}{ll} \text{minimize} & f(x) + g(z) \\ \text{subject to} & x = z \end{array}$$

Lagrangian

$$L(x, z, y) = f(x) + g(z) + y^T(z - x) = -(y^T x - f(x)) - (-y^T z - g(z))$$

Fenchel dual

Dual using conjugate functions

Equivalent form (variables split)

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

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

subject to $x = z$

Lagrangian

$$L(x, z, y) = f(x) + g(z) + y^T(z - x) = -(y^T x - f(x)) - (-y^T z - g(z))$$

Dual function

$$\min_{x,z} L(x, z, y) = -f^*(y) - g^*(-y)$$

Fenchel dual

Dual using conjugate functions

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

Equivalent form (variables split)

minimize $\quad f(x) + g(z)$
subject to $\quad x = z$

Lagrangian

$$L(x, z, y) = f(x) + g(z) + y^T(z - x) = -(y^T x - f(x)) - (-y^T z - g(z))$$

Dual function

$$\min_{x,z} L(x, z, y) = -f^*(y) - g^*(-y)$$

Dual problem

maximize $\; -f^*(y) - g^*(-y)$

Fenchel dual example

Constrained optimization

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

$\longrightarrow$

Dual problem

maximize $-f^*(y) - \sigma_C(-y)$

Fenchel dual example

Constrained optimization

minimize $f(x) + \mathcal{I}_C(x)$ ⟶

Dual problem

maximize $-f^*(y) - \sigma_C(-y)$

Norm penalization

minimize $f(x) + \|x\|$ ⟶

Dual problem

maximize $-f^*(y)$

subject to $\|y\|_* \leq 1$

Fenchel dual example

Constrained optimization

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

Dual problem

maximize $-f^*(y) - \sigma_C(-y)$

Norm penalization

minimize $f(x) + \|x\|$ $\longrightarrow$

Dual problem

maximize $-f^*(y)$

subject to $\|y\|_* \leq 1$

Remarks

- Fenchel duality can simplify derivations
- Useful when conjugates are known
- Very common in operator splitting algorithms

Monotone cocoercive operators

Monotone operators

An operator T on $\mathbf{R}^n$ is **monotone** if

$$(u-v)^T(x-y) \geq 0, \quad \forall (x,u),(y,v) \in \mathbf{gph} T$$

Monotone operators

An operator T on $\mathbf{R}^n$ is **monotone** if

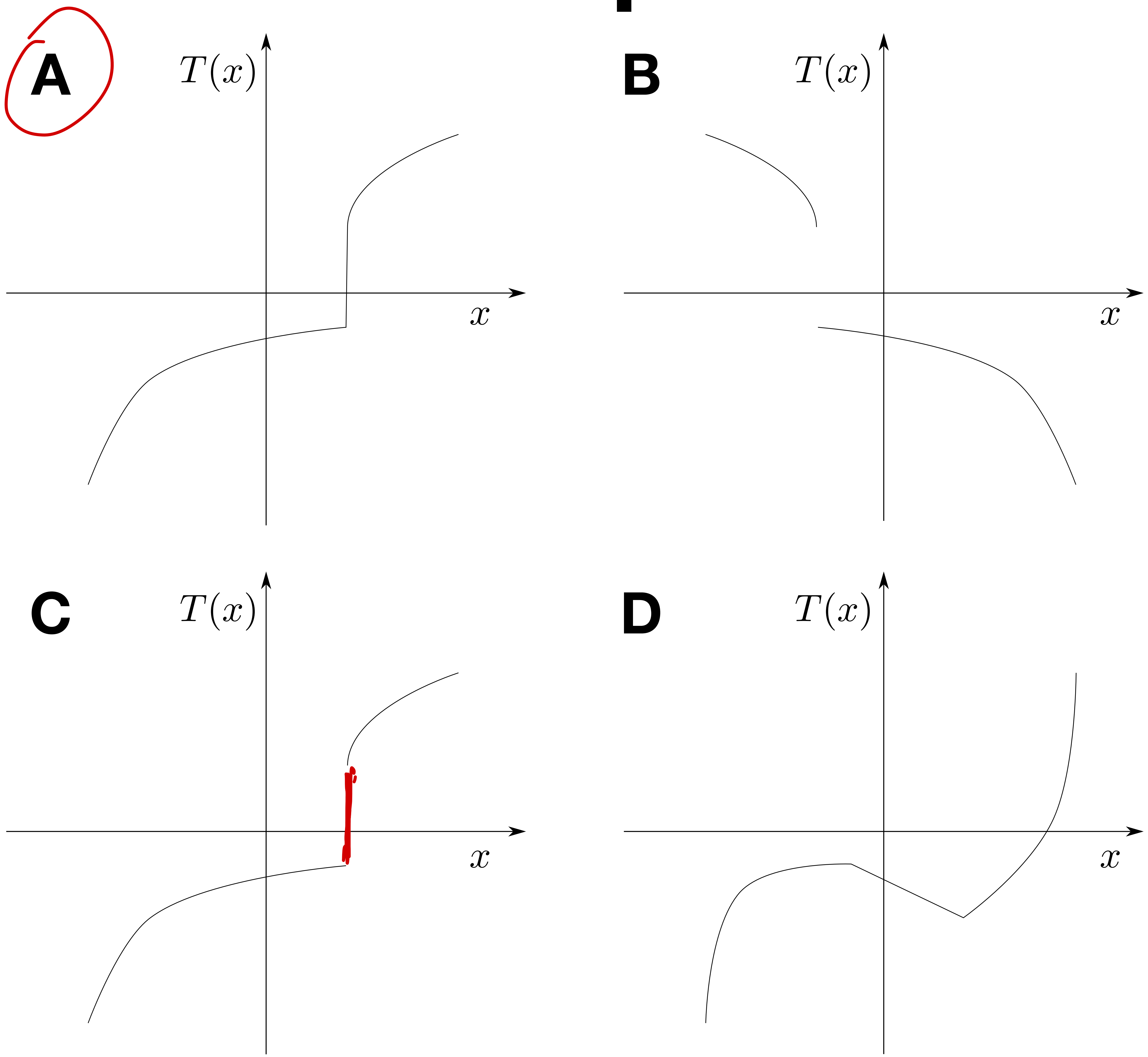
$$(u-v)^T(x-y) \geq 0, \quad \forall (x,u),(y,v) \in \mathbf{gph} T$$

T is **maximal monotone** if
$\nexists (\bar{x},\bar{u}) \notin \mathbf{gph} T$ such that

$$(\bar{u}-u)^T(\bar{x}-x) \geq 0$$

Equivalently: $\nexists$ monotone R
such that $\mathbf{gph} T \subset \mathbf{gph} R$

Monotone operators in 1D



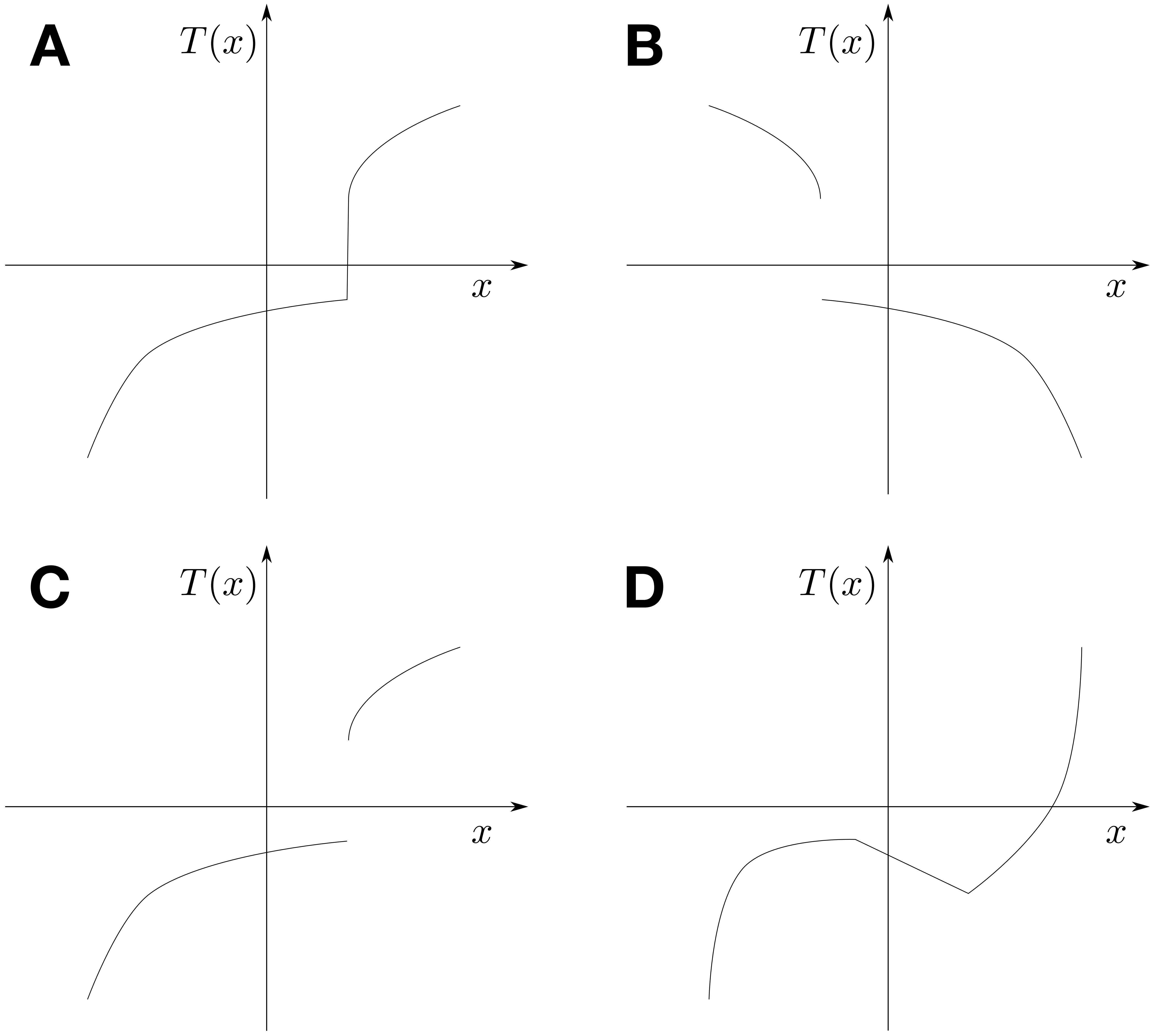
Let's fill the table

	Monotone	Max Monotone
A	✓	✓
B	✗	✗
C	✓	✗
D	✗	✗

Monotonicity

$$y > x \quad \Rightarrow \quad T(y) \geq T(x)$$

Monotone operators in 1D



Let's fill the table

	Monotone	Max Monotone
A		
B		
C		
D		

Monotonicity

$$y > x \quad \Rightarrow \quad T(y) \geq T(x)$$

Continuity

If T single-valued, continuous and monotone, then it's maximal monotone

Monotone operator properties

- **sum** $T + R$ is monotone
- **nonnegative scaling** αT with $\alpha \geq 0$ is monotone
- **inverse** T^{-1} is monotone
- **congruence** for $M \in \mathbf{R}^{n \times m}$, then $M^T T(Mz)$ is monotone on $\mathbf{R}^m$

Affine function $T(x) = Ax + b$ is maximal monotone

$$\iff \quad A + A^T \succeq 0$$

$f(x) = 1/2\, x^T P x + q^T x$

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

$P + P^T \succeq 0$

Strongly monotone operators

An operator T on $\mathbf{R}^n$ is μ-**strongly monotone** if

$$(u-v)^T(x-y) \geq \mu\|x-y\|^2, \quad \mu > 0$$

(also called μ-**coercive**)

$\forall (x,u),(y,v) \in \mathbf{gph} T$

Strongly monotone operators

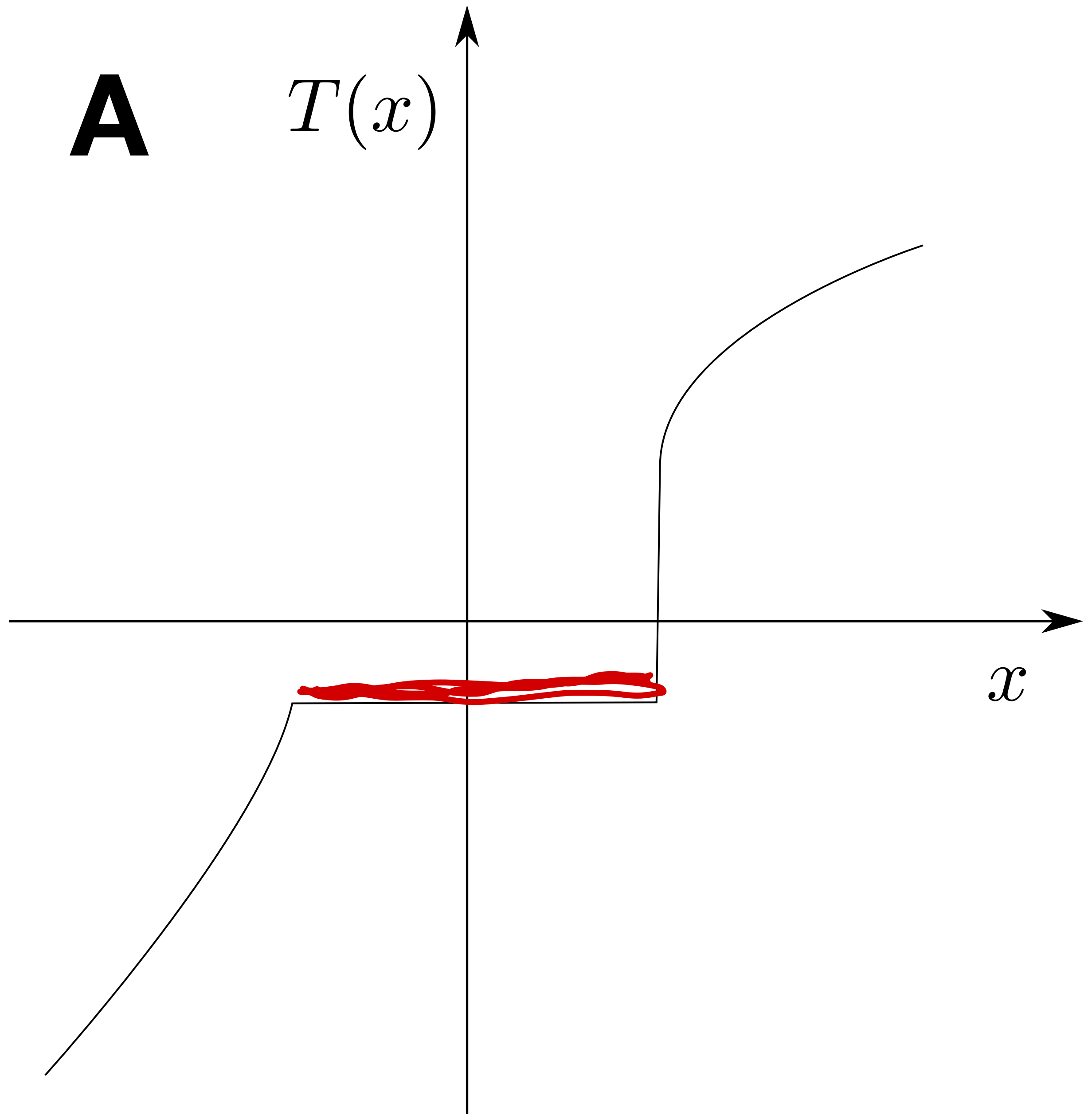
An operator T on $\mathbf{R}^n$ is μ-**strongly monotone** if

$$(u-v)^T(x-y) \geq \mu\|x-y\|^2, \quad \mu > 0$$

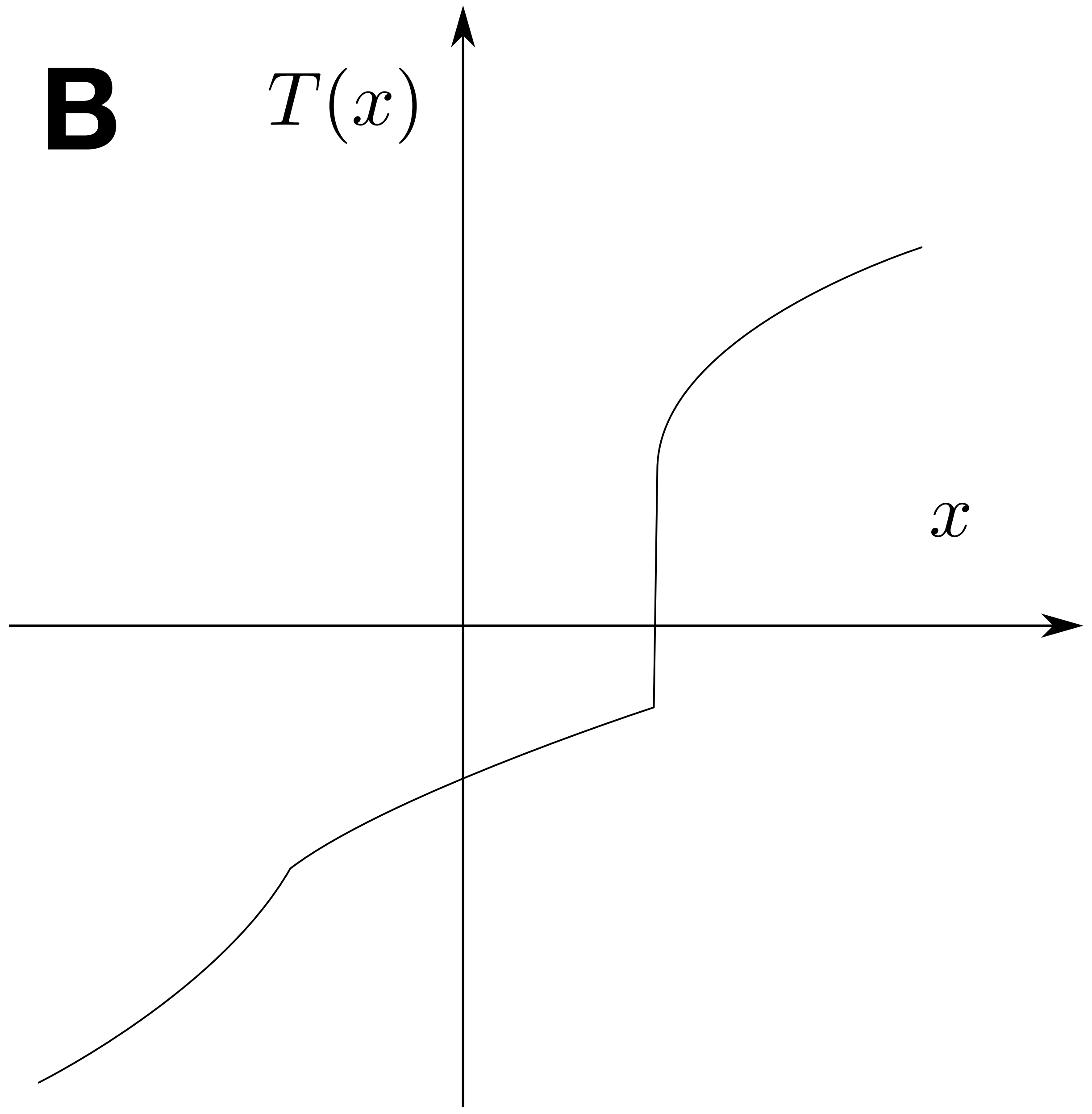
(also called μ**-coercive**)

$\forall (x,u),(y,v) \in \mathbf{gph} T$

A



B



Let's fill the table

	Monotone	Strongly Monotone
A	✓	✗
B	✓	✓

Strongly monotone operators

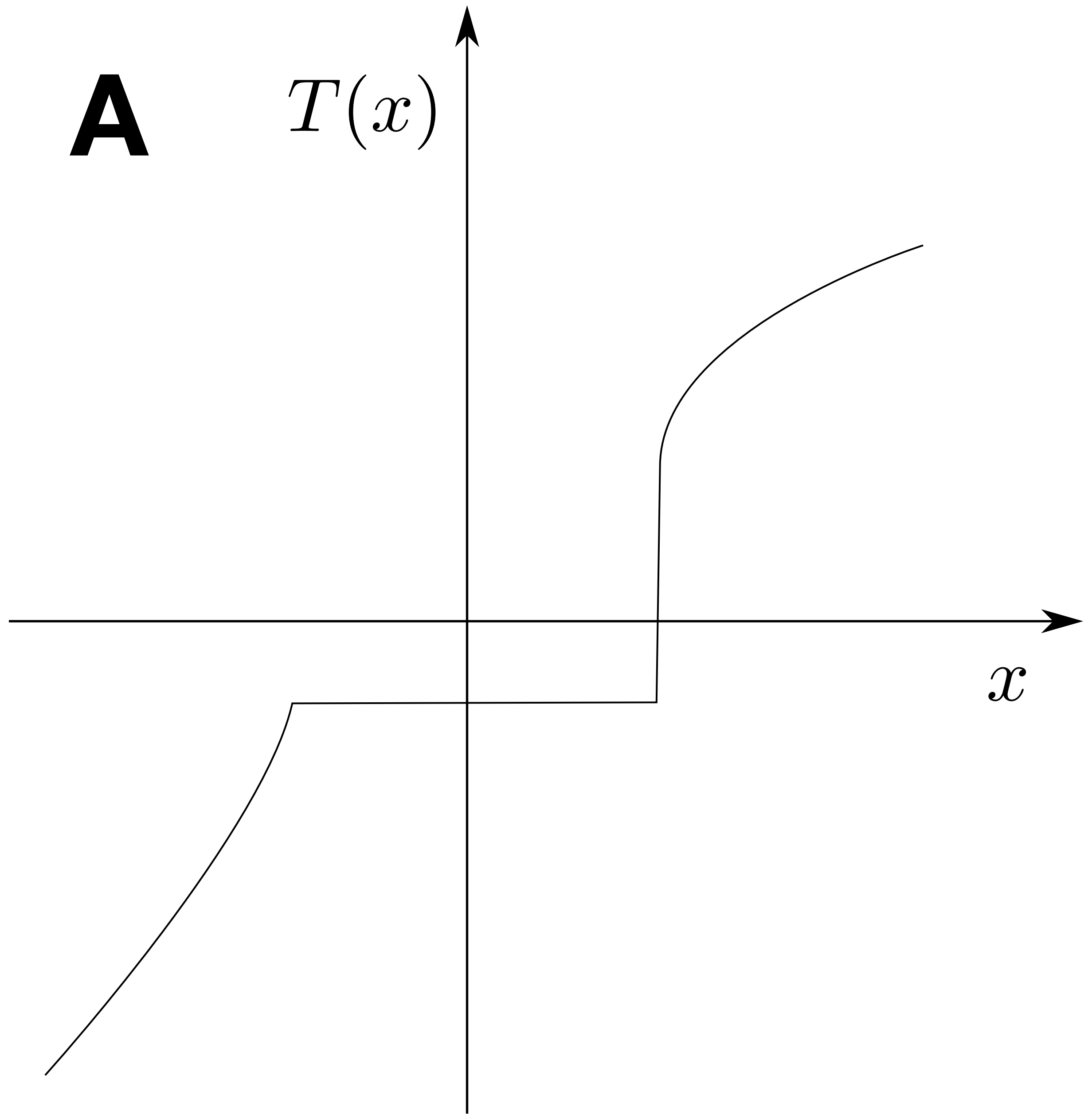
An operator T on $\mathbf{R}^n$ is μ-**strongly monotone** if

$$(u - v)^T(x - y) \geq \mu \|x - y\|^2, \quad \mu > 0$$

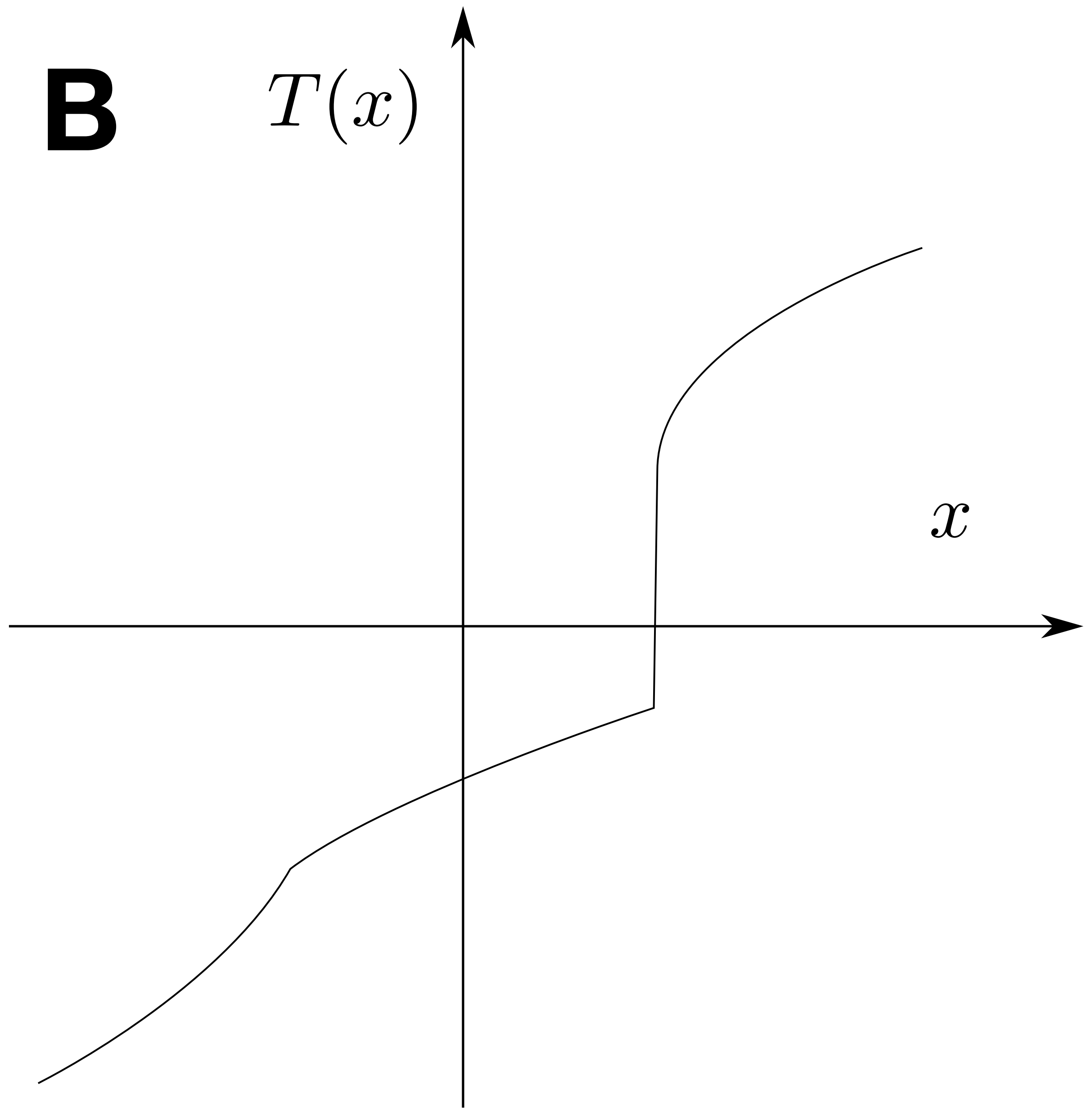
(also called μ-**coercive**)

$\forall (x, u), (y, v) \in \mathbf{gph} T$

A



B



Let's fill the table

	Monotone	Strongly Monotone
A		
B		

The slope is at least μ

Cocoercive operators

An operator T is β-**cocoercive**, $\beta > 0$, if

$$(T(x) - T(y))^T(x - y) \geq \beta \|T(x) - T(y)\|^2$$

Cocoercive operators

An operator T is β**-cocoercive**, $\beta > 0$, if

$$(T(x) - T(y))^T(x - y) \geq \beta\|T(x) - T(y)\|^2$$

If T is β**-cocoercive**, then T is $(1/\beta)$**-Lipschitz**

Proof $\beta\|T(x) - T(y)\|^2 \leq (T(x) - T(y))^T(x - y) \leq \|T(x) - T(y)\|\|x - y\|$

$$\implies \|T(x) - T(y)\| \leq (1/\beta)\|x - y\| \qquad \blacksquare$$

Cocoercive operators

An operator T is β**-cocoercive**, $\beta > 0$, if

$$(T(x) - T(y))^T (x - y) \geq \beta \|T(x) - T(y)\|^2$$

If T is β**-cocoercive**, then T is $(1/\beta)$**-Lipschitz**

Proof $\beta \|T(x) - T(y)\|^2 \leq (T(x) - T(y))^T (x - y) \leq \|T(x) - T(y)\| \|x - y\|$

$$\implies \|T(x) - T(y)\| \leq (1/\beta) \|x - y\| \qquad \blacksquare$$

If T is μ**-strongly monotone** if and only if T^{-1} is μ**-cocoercive**

Proof $(T(x) - T(x))^T (x - y) \geq \mu \|x - y\|^2$

Inverse: $u = T(x)$ and $v = T(y)$ if and only if $x \in T^{-1}(u)$ and $y \in T^{-1}(v)$

$$(u - v)^T (T^{-1}(u) - T^{-1}(v)) \geq \mu \|T^{-1}(u) - T^{-1}(v)\|^2 \qquad \blacksquare$$

Cocoercive and nonexpansive operators

If T is β-**cocoercive** if and only if $I - 2\beta T$ is **nonexpansive**

Cocoercive and nonexpansive operators

If T is β-**cocoercive** if and only if $I - 2\beta T$ is **nonexpansive**

Proof $\|(I{-}2\beta T)(y) - (I - 2\beta T)(x)\|^2 =$

$$= \|y - 2\beta T(y) - x - 2\beta T(x)\|^2$$

$$= \|y - x\|^2 - 4\beta(T(y) - T(x))^T(y - x) + 4\beta^2\|T(y) - T(x)\|^2$$

$$= |y - x\|^2 - 4\beta\left((T(y) - T(x))^T(y - x) - \beta\|T(y) - T(x)\|^2\right)$$

$$\leq \|y - x\|^2 \quad \blacksquare$$

Cococercive and nonexpansive operators

If T is β**-cocoercive** if and only if $I - 2\beta T$ is **nonexpansive**

Proof $\|(I{-}2\beta T)(y) - (I - 2\beta T)(x)\|^2 =$

$$= \|y - 2\beta T(y) - x - 2\beta T(x)\|^2$$

$$= \|y - x\|^2 - 4\beta (T(y) - T(x))^T (y - x) + 4\beta^2 \|T(y) - T(x)\|^2$$

$$= |y - x\|^2 - 4\beta \left((T(y) - T(x))^T (y - x) - \beta \|T(y) - T(x)\|^2\right)$$

$$\leq \|y - x\|^2 \quad \blacksquare$$

(cocoercive)

Summary of monotone and cocoercive operators

Monotone

$$(T(x) - T(y))^T (x - y) \geq 0$$

↑ $\mu = 0$

Strongly monotone

$$(T(x) - T(y))^T (x - y) \geq \mu \|x - y\|^2$$

⟷ $F = T^{-1}$

Lipschitz

$$\|F(x) - F(y)\| \leq L \|x - y\|$$

↑ $L = 1/\mu$

Cococercive

$$(F(x) - F(y))^T (x - y) \geq \mu \|F(x) - F(y)\|^2$$

↕ $G = I - 2\mu F$

Nonexpansive

$$\|G(x) - G(y)\| \leq \|x - y\|$$

Subdifferential operator and monotonicity

Subdifferential operator monotonicity

$$\partial f(x) = \{g \mid f(y) \geq f(x) + g^T(y - x)\}$$

$\partial f(x)$ is **monotone** (also for nonconvex functions)

Subdifferential operator monotonicity

$$\partial f(x) = \{g \mid f(y) \geq f(x) + g^T(y - x)\}$$

$\partial f(x)$ is **monotone** (also for nonconvex functions)

Proof Suppose $u \in \partial f(x)$ and $v \in \partial f(y)$ then

$$f(y) \geq f(x) + u^T(y - x), \qquad f(x) \geq f(y) + v^T(x - y)$$

By adding them, we can write $(u - v)^T(x - y) \geq 0$ ■

Subdifferential operator monotonicity

$$\partial f(x) = \{g \mid f(y) \geq f(x) + g^T(y - x)\}$$

$\partial f(x)$ is **monotone** (also for nonconvex functions)

Proof Suppose $u \in \partial f(x)$ and $v \in \partial f(y)$ then

$$f(y) \geq f(x) + u^T(y - x), \qquad f(x) \geq f(y) + v^T(x - y)$$

By adding them, we can write $(u - v)^T(x - y) \geq 0$ ■

Maximal monotonicity

If f is convex, closed and proper (CCP), then $\partial f(x)$ is maximal monotone

Strongly monotone and cocoercive subdifferential

f is μ **-strongly convex** $\quad \Longleftrightarrow \quad$ ∂f μ**-strongly monotone**

$$(\partial f(x) - \partial f(y))^T (x - y) \geq \mu \|x - y\|^2$$

Strongly monotone and cocoercive subdifferential

f is μ **-strongly convex** $\iff$ ∂f μ**-strongly monotone**

$$(\partial f(x) - \partial f(y))^T(x-y) \geq \mu \|x-y\|^2$$

f is L**-smooth**

$\iff$ ∂f L**-Lipschitz** and $\partial f = \nabla f$: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$

$\iff$ ∂f $(1/L)$**-cocoercive**: $(\nabla f(x) - \nabla f(y))^T(x-y) \geq (1/L)\|\nabla f(x) - \nabla f(y)\|^2$

Strongly monotone and cocoercive subdifferential

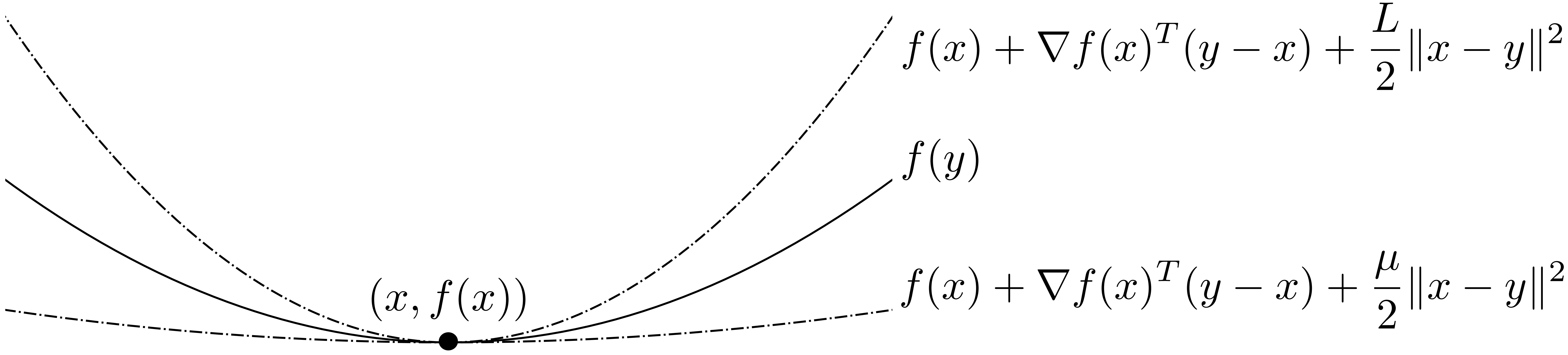
f is μ **-strongly convex** $\iff$ ∂f μ**-strongly monotone**

$$(\partial f(x) - \partial f(y))^T(x - y) \geq \mu\|x - y\|^2$$

f is L**-smooth**

$\iff$ ∂f L**-Lipschitz** and $\partial f = \nabla f$: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$

$\iff$ ∂f $(1/L)$**-cocoercive**: $(\nabla f(x) - \nabla f(y))^T(x - y) \geq (1/L)\|\nabla f(x) - \nabla f(y)\|^2$



Inverse of subdifferential

If f is CCP, then, $\qquad (\partial f)^{-1} = \partial f^*$

Inverse of subdifferential

If f is CCP, then, $\quad (\partial f)^{-1} = \partial f^*$

Proof

$$
\begin{aligned}
(u, v) \in \mathbf{gph}(\partial f)^{-1} &\iff (v, u) \in \mathbf{gph}\partial f \\
&\iff u \in \partial f(v) \\
&\iff 0 \in \partial f(v) - u \\
&\iff v \in \operatorname*{argmin}_x f(x) - u^T x \\
&\iff f^*(u) = u^T v - f(v)
\end{aligned}
$$

Inverse of subdifferential

If f is CCP, then, $\quad (\partial f)^{-1} = \partial f^*$

Proof

$$
\begin{aligned}
(u, v) \in \mathbf{gph}(\partial f)^{-1} &\iff (v, u) \in \mathbf{gph}\partial f \\
&\iff u \in \partial f(v) \\
&\iff 0 \in \partial f(v) - u \\
&\iff v \in \underset{x}{\operatorname{argmin}}\, f(x) - u^T x \\
&\iff f^*(u) = u^T v - f(v)
\end{aligned}
$$

Therefore, $f(v) + f^*(u) = u^T v$. If f is CCP, then $f^{**} = f$ and we can write

$$f^{**}(v) + f^*(u) = u^T v \quad \iff \quad (u, v) \in \mathbf{gph}\partial f^* \quad \blacksquare$$

Strong convexity is the dual of smoothness

$$f \text{ is } \mu\text{-strongly convex} \quad \Longleftrightarrow \quad f^* \text{ is } (1/\mu)\text{-smooth}$$

Strong convexity is the dual of smoothness

f is μ-strongly convex $\quad \iff \quad$ f^* is $(1/\mu)$-smooth

Proof

$$
\begin{aligned}
f \quad \mu\text{-strongly convex} \quad &\iff \quad \partial f \quad \mu\text{-strongly monotone} \\
&\iff \quad (\partial f)^{-1} = \partial f^* \quad \mu\text{-cocoercive} \\
&\iff \quad f^* \quad (1/\mu)\text{-smooth} \qquad \blacksquare
\end{aligned}
$$

Strong convexity is the dual of smoothness

$$f \text{ is } \mu\text{-strongly convex} \quad \Longleftrightarrow \quad f^* \text{ is } (1/\mu)\text{-smooth}$$

Proof

$$\begin{aligned} f \quad \mu\text{-strongly convex} \quad &\Longleftrightarrow \quad \partial f \quad \mu\text{-strongly monotone} \\ &\Longleftrightarrow \quad (\partial f)^{-1} = \partial f^* \quad \mu\text{-cocoercive} \\ &\Longleftrightarrow \quad f^* \quad (1/\mu)\text{-smooth} \qquad \blacksquare \end{aligned}$$

Remark: strong convexity and (strong) smoothness are **dual**

Operators in optimization problems

KKT operator

minimize $f(x)$

subject to $Ax = b$

$\longrightarrow$

Lagrangian

$$L(x, y) = f(x) + y^T(Ax - b)$$

KKT operator

minimize $f(x)$
subject to $Ax = b$

→

Lagrangian

$$L(x, y) = f(x) + y^T(Ax - b)$$

KKT operator

$$T(x, y) = \begin{bmatrix} \partial_x L(x, y) \\ -\partial_y L(x, y) \end{bmatrix} = \begin{bmatrix} \partial f(x) + A^T y \\ b - Ax \end{bmatrix} = \begin{bmatrix} r^{\text{dual}} \\ -r^{\text{prim}} \end{bmatrix}$$

zero set $\{(x, y) \mid 0 \in T(x, y)\}$ is the set of **primal-dual optimal points**

KKT operator

minimize $f(x)$
subject to $Ax = b$

→

Lagrangian

$$L(x, y) = f(x) + y^T(Ax - b)$$

KKT operator

$$T(x, y) = \begin{bmatrix} \partial_x L(x, y) \\ -\partial_y L(x, y) \end{bmatrix} = \begin{bmatrix} \partial f(x) + A^T y \\ b - Ax \end{bmatrix} = \begin{bmatrix} r^{\text{dual}} \\ -r^{\text{prim}} \end{bmatrix}$$

zero set $\{(x, y) \mid 0 \in T(x, y)\}$ is the set of **primal-dual optimal points**

Monotonicity

$$T(x, y) = \begin{bmatrix} \partial f(x) \\ b \end{bmatrix} + \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

KKT operator

minimize $\quad f(x)$

subject to $\quad Ax = b$

→

Lagrangian

$$L(x, y) = f(x) + y^T(Ax - b)$$

KKT operator

$$T(x, y) = \begin{bmatrix} \partial_x L(x, y) \\ -\partial_y L(x, y) \end{bmatrix} = \begin{bmatrix} \partial f(x) + A^T y \\ b - Ax \end{bmatrix} = \begin{bmatrix} r^{\text{dual}} \\ -r^{\text{prim}} \end{bmatrix}$$

zero set $\{(x, y) \mid 0 \in T(x, y)\}$ is the set of **primal-dual optimal points**

Monotonicity M

$M + M^T = 0 \succeq 0$

$$T(x, y) = \begin{bmatrix} \partial f(x) \\ b \end{bmatrix} + \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

sum of monotone operators

skew-symmetric

“multiplier to residual” mapping

$$\begin{array}{ll}\text{minimize} & f(x) \\ \text{subject to} & Ax = b\end{array}$$

"multiplier to residual" mapping

Lagrangian

minimize	$f(x)$
subject to	$Ax = b$

$\longrightarrow$

$$L(x, y) = f(x) + y^T(Ax - b)$$

“multiplier to residual” mapping

Lagrangian

minimize $f(x)$
subject to $Ax = b$

$\longrightarrow$

$$L(x, y) = f(x) + y^T(Ax - b)$$

Dual problem

maximize $g(y) = \min_x L(x, y) = -\max_x -L(x, y) = -(f^*(-A^T y) + y^T b)$

“multiplier to residual” mapping

minimize $f(x)$

subject to $Ax = b$

⟶

Lagrangian

$$L(x, y) = f(x) + y^T(Ax - b)$$

Dual problem

maximize $g(y) = \min_x L(x, y) = -\max_x -L(x, y) = -(f^*(-A^T y) + y^T b)$

Operator

$T(y) = b - Ax$, where $x = \mathrm{argmin}_z\, L(z, y)$

⟶

Monotonicity

If f CCP, then T is monotone

"multiplier to residual" mapping

minimize $f(x)$

subject to $Ax = b$

Lagrangian

$$L(x, y) = f(x) + y^T(Ax - b)$$

Dual problem

maximize $g(y) = \min_x L(x, y) = -\max_x -L(x, y) = -(f^*(-A^T y) + y^T b)$

Operator

$T(y) = b - Ax$, where $x = \mathrm{argmin}_z\, L(z, y)$

Monotonicity

If f CCP, then T is monotone

Proof

$0 \in \partial_x L(x, y) = \partial f(x) + A^T y \iff x = (\partial f)^{-1}(-A^T y)$

Therefore, $T(y) = b - A(\partial f)^{-1}(-A^T y) = \partial_y \left(b^T y + f^*(-A^T y)\right) = \partial(-g)$ ∎

“multiplier to residual” mapping

minimize $f(x)$
subject to $Ax = b$ $\longrightarrow$

Lagrangian

$$L(x, y) = f(x) + y^T(Ax - b)$$

Dual problem

maximize $g(y) = \min_x L(x, y) = -\max_x -L(x, y) = -(f^*(-A^T y) + y^T b)$

Operator

$T(y) = b - Ax$, where $x = \operatorname{argmin}_z L(z, y)$ $\longrightarrow$

Monotonicity

If f CCP, then T is monotone

Proof

$$0 \in \partial_x L(x, y) = \partial f(x) + A^T y \quad \Longleftrightarrow \quad x = (\partial f)^{-1}(-A^T y)$$

Therefore, $T(y) = b - A(\partial f)^{-1}(-A^T y) = \partial_y \left(b^T y + f^*(-A^T y)\right) = \partial(-g)$ ■

monotone

Operators in algorithms

Forward step operator

The **forward step operator** of T is defined as

$$I - \gamma T$$

In general **monotonicity of T** is not enough for convergence

Forward step operator

The **forward step operator** of T is defined as

$$I - \gamma T$$

In general **monotonicity of T** is not enough for convergence

Example

minimize $\quad x$

subject to $\quad x = 0$

KKT operator

$$T(x, y) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Monotone (skew-symmetric)

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad A + A^T = 0 \succeq 0$$

Forward step operator

The **forward step operator** of T is defined as

$$I - \gamma T$$

In general **monotonicity of T** is not enough for convergence

Example

minimize $\quad x$

subject to $\quad x = 0$

KKT operator

$$T(x, y) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Monotone (skew-symmetric)

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad A + A^T = 0 \succeq 0$$

Forward step

$$(I - \gamma T) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & -\gamma \\ \gamma & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \longrightarrow$$

Expansive

$$\left\| \begin{bmatrix} 1 & -\gamma \\ \gamma & 1 \end{bmatrix} \right\|_2 > 1, \quad \forall \gamma$$

Gradient step: special case of forward step

f L-smooth $\iff$ ∇f $(1/L)$-cocoercive $\iff$ $I - (2/L)\nabla f$ nonexpansive

Gradient step: special case of forward step

f L-smooth $\iff$ ∇f $(1/L)$-cocoercive $\iff$ $I - (2/L)\nabla f$ nonexpansive

Construct averaged iterations

$$I - \gamma \nabla f = (1-\alpha)I + \alpha(I - (2/L)\nabla f)$$

where $\alpha = \gamma L/2 \in (0,1)$ $\iff$ $\gamma \in (0, 2/L)$

Gradient step: special case of forward step

f L-smooth $\iff$ ∇f $(1/L)$-cocoercive $\iff$ $I - (2/L)\nabla f$ nonexpansive

Construct averaged iterations

$$I - \gamma \nabla f = (1-\alpha)I + \alpha(I - (2/L)\nabla f)$$

where $\alpha = \gamma L/2 \in (0,1)$ $\iff$ $\gamma \in (0, 2/L)$

(to be averaged)

Gradient step: special case of forward step

f L-smooth $\iff$ ∇f $(1/L)$-cocoercive $\iff$ $I - (2/L)\nabla f$ nonexpansive

Construct averaged iterations

$$I - \gamma \nabla f = (1-\alpha)I + \alpha(I - (2/L)\nabla f)$$

where $\alpha = \gamma L/2 \in (0,1)$ $\iff$ $\gamma \in (0, 2/L)$

(to be averaged)

Remark

- Only smoothness assumption gives **sublinear convergence**
- Similar result we obtained in gradient descent lecture

Resolvent and Cayley operators

The **resolvent** of operator A is defined as

$$R_A = (I + A)^{-1}$$

The **Cayley (reflection) operator** of A is defined as

$$C_A = 2R_A - I = 2(I + A)^{-1} - I$$

Properties

- If A is maximal monotone, $\mathbf{dom}\, R_A = \mathbf{dom}\, C_A = \mathbf{R}^n$ (Minty's theorem)
- If A is **monotone**, R_A and C_A are **nonexpansive** (thus functions)
- **Zeros** of A are **fixed points** of R_A and C_A

Resolvent and Cayley operators

The **resolvent** of operator A is defined as

$$R_A = (I + A)^{-1}$$

The **Cayley (reflection) operator** of A is defined as

$$C_A = 2R_A - I = 2(I + A)^{-1} - I$$

Properties

- If A is maximal monotone, $\mathbf{dom}\, R_A = \mathbf{dom}\, C_A = \mathbf{R}^n$ (Minty's theorem)
- If A is **monotone**, R_A and C_A are **nonexpansive** (thus functions)
- **Zeros** of A are **fixed points** of R_A and C_A

Key result we can solve $0 \in A(x)$ by finding fixed points of C_A or R_A

Fixed points of R_A and C_A are zeros of A

Proof

$$R_A = (I + A)^{-1}$$

$x \in \mathbf{fix}\, R_A$

$$\begin{aligned} 0 \in A(x) &\iff x \in (I + A)(x) \\ &\iff (I + A)^{-1}(x) = x \\ &\iff x = R_A(x) \end{aligned}$$

Fixed points of R_A and C_A are zeros of A

Proof

$$R_A = (I + A)^{-1}$$

$x \in \mathbf{fix}\, R_A$

$$\begin{aligned} 0 \in A(x) &\iff x \in (I + A)(x) \\ &\iff (I + A)^{-1}(x) = x \\ &\iff x = R_A(x) \end{aligned}$$

$x \in \mathbf{fix}\, C_A$

$$C_A(x) = 2R_A(x) - I(x) = 2x - x = x \qquad \blacksquare$$

If A is monotone, then R_A is nonexpansive

Proof

If $(x, u) \in \mathbf{gph} R_A$ and $(y, v) \in \mathbf{gph} R_A$, then

$$u + A(u) \ni x, \qquad v + A(v) \ni y$$

If A is monotone, then R_A is nonexpansive

Proof

If $(x, u) \in \mathbf{gph} R_A$ and $(y, v) \in \mathbf{gph} R_A$, then

$$u + A(u) \ni x, \qquad v + A(v) \ni y$$

Subtract to get $u - v + (A(u) - A(v)) \ni x - y$

If A is monotone, then R_A is nonexpansive

Proof

If $(x, u) \in \mathbf{gph} R_A$ and $(y, v) \in \mathbf{gph} R_A$, then

$$u + A(u) \ni x, \qquad v + A(v) \ni y$$

Subtract to get $u - v + (A(u) - A(v)) \ni x - y$

Multiply by $(u - v)^T$ and use monotonicity of A (being also a function: $\in \to =$),

$$\|u - v\|^2 \leq (x - y)^T (u - v)$$

If A is monotone, then R_A is nonexpansive

Proof

If $(x, u) \in \mathbf{gph} R_A$ and $(y, v) \in \mathbf{gph} R_A$, then

$$u + A(u) \ni x, \qquad v + A(v) \ni y$$

Subtract to get $u - v + (A(u) - A(v)) \ni x - y$

Multiply by $(u-v)^T$ and use monotonicity of A (being also a function: $\in \to =$),

$$\|u - v\|^2 \leq (x - y)^T (u - v)$$

Apply Cauchy-Schwarz and divide by $\|u - v\|$ to get

$$\|u - v\| \leq \|x - y\| \qquad \blacksquare$$

If A is monotone, then C_A is nonexpansive

Proof

Given $u = R_A(x)$ and $v = R_A(y)$ (R_A is a function)

$$\begin{aligned} \|C(x) - C(y)\|^2 &= \|(2u - x) - (2v - y)\|^2 \\ &= \|2(u - v) - (x - y)\|^2 \\ &= 4\|u - v\|^2 - 4(u - v)^T(x - y) + \|x - y\|^2 \\ &\leq \|x - y\|^2 \end{aligned}$$

Note R_A monotonicity (prev slide): $\|u - v\|^2 \leq (u - v)^T(x - y)$ ■

If A is monotone, then C_A is nonexpansive

Proof

Given $u = R_A(x)$ and $v = R_A(y)$ (R_A is a function)

$$\begin{aligned}\|C(x) - C(y)\|^2 &= \|(2u - x) - (2v - y)\|^2 \\ &= \|2(u - v) - (x - y)\|^2 \\ &= 4\|u - v\|^2 - 4(u - v)^T(x - y) + \|x - y\|^2 \\ &\leq \|x - y\|^2\end{aligned}$$

Note R_A monotonicity (prev slide): $\|u - v\|^2 \leq (u - v)^T(x - y)$ ■

Remark

R_A is nonexpansive since it is the average of I and C_A:

$$R_A = (1/2)I + (1/2)C_A = (1/2)I + (1/2)(2R_A - 1)$$

Role of maximality

We mostly consider **maximal** operators A because of

Theory: A, R_A and C_A do not bring iterates outside their domains

Practice: hard to compute R_A and C_A for non-maximal monotone operators, e.g., when $A = \partial f(x)$ where f nonconvex.

Resolvent of subdifferential: proximal operator

$$\mathbf{prox}_f = R_{\partial f} = (I + \partial f)^{-1}$$

Proof

Let $z = \mathbf{prox}_f(x)$, then

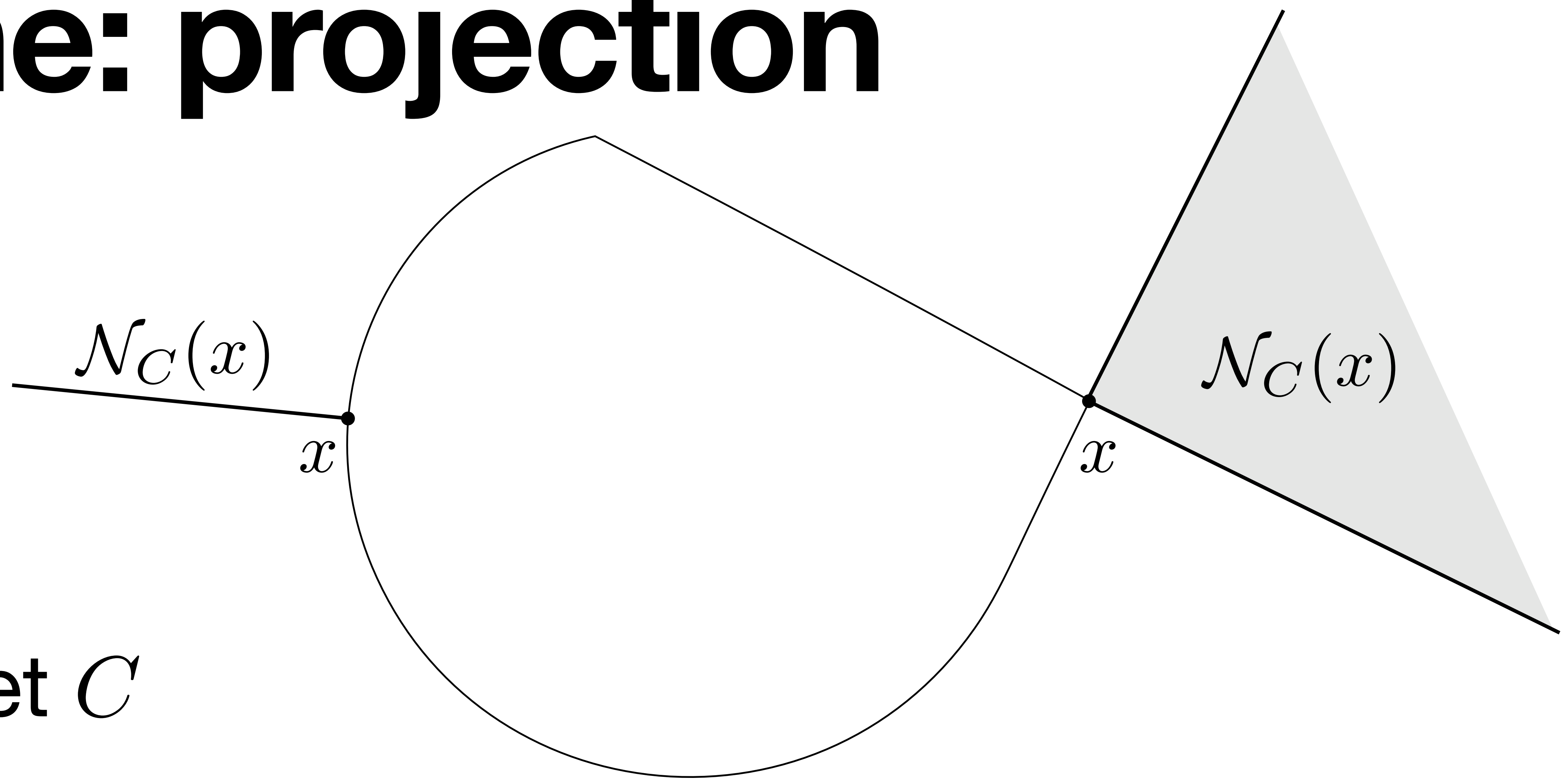
$$z = \underset{u}{\operatorname{argmin}}\, f(u) + \frac{1}{2}\|u - x\|^2$$

$$\iff \; 0 \in \partial f(z) + z - x \qquad \text{(optimality conditions)}$$

$$\iff \; x \in (I + \partial f)(z)$$

$$\iff \; z = (I + \partial f)^{-1}(x) \qquad \blacksquare$$

Resolvent of normal cone: projection



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

Proof

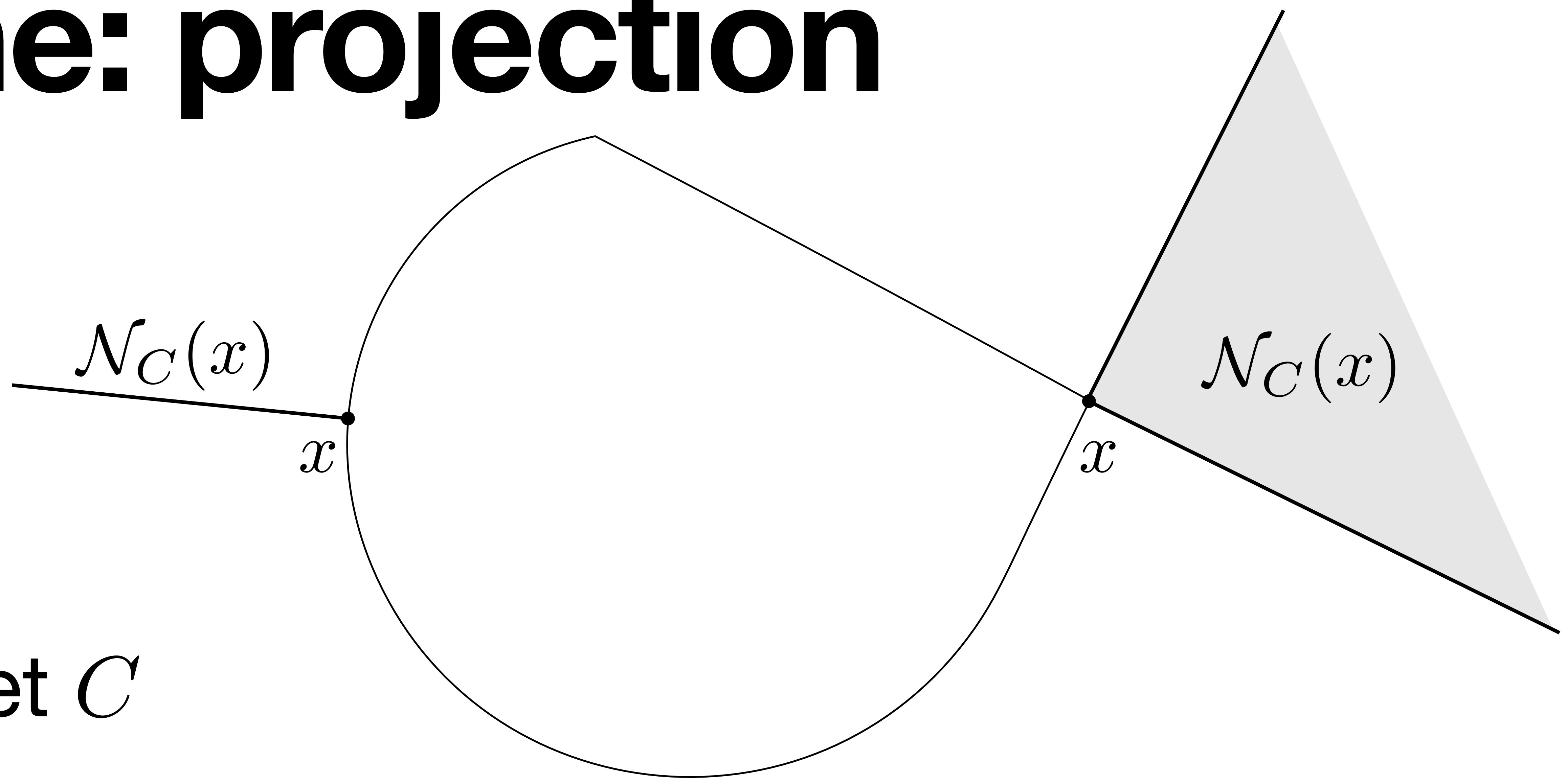
Let $f = \mathcal{I}_C$, the indicator function of a convex set C

Recall: $\partial \mathcal{I}_C(x) = \mathcal{N}_C(x)$ **normal cone operator**

$$u = (I + \partial \mathcal{I}_C)^{-1}(x) \quad \iff \quad u = \operatorname*{argmin}_z \ \mathcal{I}_C(u) + (1/2)\|z - x\|^2 = \Pi_C(x)$$

■

Resolvent of normal cone: projection



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

Proof

Let $f = \mathcal{I}_C$, the indicator function of a convex set C

Recall: $\partial \mathcal{I}_C(x) = \mathcal{N}_C(x)$ **normal cone operator**

$$u = (I + \partial \mathcal{I}_C)^{-1}(x) \quad \Longleftrightarrow \quad u = \underset{z}{\operatorname{argmin}} \ \mathcal{I}_C(u) + (1/2)\|z - x\|^2 = \Pi_C(x) \qquad \blacksquare$$

$\mathcal{N}_C$ monotone $\quad \Longrightarrow \quad$ Π_C nonexpansive

Proof of monotonicity

$$u \in \mathcal{N}_C(x) \ \Rightarrow \ u^T(z - x) \le 0, \ \forall z \in C \ \Rightarrow \ u^T(y - x) \le 0$$

$$v \in \mathcal{N}_C(y) \ \Rightarrow \ v^T(z - y) \le 0, \ \forall z \in C \ \Rightarrow \ v^T(x - y) \le 0$$

$\longrightarrow$ add to obtain monotonicity $\blacksquare$

Building contractions

Forward step contractions

Given T L-Lipschitz and μ-strongly monotone, then $I - \gamma T$ converges linearly at rate $\sqrt{1 - 2\gamma\mu + \gamma^2 L^2}$, with optimal step $\gamma = \mu / L^2$.

Forward step contractions

Given T L-Lipschitz and μ-strongly monotone, then $I - \gamma T$ converges linearly at rate $\sqrt{1 - 2\gamma\mu + \gamma^2 L^2}$, with optimal step $\gamma = \mu / L^2$.

Proof

$$\begin{aligned}
\|(I - \gamma T)(x) - (I - \gamma T)(y)\|^2 &= \|x - y + \gamma T(x) - \gamma T(y)\|^2 \\
&= \|x - y\|^2 - 2\gamma(T(x) - T(y))^T(x - y) + \gamma^2\|T(x) - T(y)\|^2 \\
&\leq (1 - 2\gamma\mu + \gamma^2 L^2)\|x - y\|^2
\end{aligned}$$

■

Forward step contractions

Given T L-Lipschitz and μ-strongly monotone, then $I - \gamma T$ converges linearly at rate $\sqrt{1 - 2\gamma\mu + \gamma^2 L^2}$, with optimal step $\gamma = \mu / L^2$.

Proof

$$\begin{aligned}
\|(I-\gamma T)(x) - (I-\gamma T)(y)\|^2 &= \|x - y + \gamma T(x) - \gamma T(y)\|^2 \\
&= \|x-y\|^2 - 2\gamma (T(x) - T(y))^T (x-y) + \gamma^2 \|T(x) - T(y)\|^2 \\
&\leq (1 - 2\gamma\mu + \gamma^2 L^2)\|x-y\|^2
\end{aligned}$$

strongly monotone

Lipschitz

■

Forward step contractions

Given T L-Lipschitz and μ-strongly monotone, then $I - \gamma T$ converges linearly at rate $\sqrt{1 - 2\gamma\mu + \gamma^2 L^2}$, with optimal step $\gamma = \mu / L^2$.

Proof

$$\begin{aligned}
\|(I - \gamma T)(x) - (I - \gamma T)(y)\|^2 &= \|x - y + \gamma T(x) - \gamma T(y)\|^2 \\
&= \|x - y\|^2 - 2\gamma (T(x) - T(y))^T (x - y) + \gamma^2 \|T(x) - T(y)\|^2 \\
&\leq (1 - 2\gamma\mu + \gamma^2 L^2)\|x - y\|^2
\end{aligned}$$

strongly monotone

Lipschitz

■

Remarks

- It applies to **gradient descent** with L-smooth and μ-strongly convex f
- Better rate in gradient descent lecture. We can get it by bounding derivative: $\|D(I - \gamma \nabla^2 f(x))\|_2 \leq \max\{|1 - \gamma L|, |1 - \gamma\mu|\}$. Optimal step $\gamma = 2/(\mu + L)$ and factor $(\mu/L - 1)(\mu/L + 1)$.

Resolvent contractions

If A is μ-strongly monotone, then

$$R_A = (I + A)^{-1}$$

is a contraction with Lipschitz parameter $1/(1+\mu)$

Resolvent contractions

If A is μ-strongly monotone, then

$$R_A = (I + A)^{-1}$$

is a contraction with Lipschitz parameter $1/(1+\mu)$

Proof

$$\begin{aligned} A\ \mu\text{-strongly monotone} \quad &\implies \quad (I+A) \qquad (1+\mu)\text{-strongly monotone} \\ &\implies \quad R_A = (I+A)^{-1} \quad (1+\mu)\text{-cocoercive} \\ &\implies \quad R_A \qquad (1/(1+\mu))\text{-Lipschitz} \qquad \blacksquare \end{aligned}$$

Cayley contractions

If A is μ-strongly monotone and L-Lipschitz, then

$$C_{\gamma A} = 2R_{\gamma A} - I = 2(I + \gamma A)^{-1} - I$$

is a contraction with factor $\sqrt{1 - 4\gamma\mu/(1+\gamma L)^2}$

Remark need also Lipschitz condition

Proof [Page 20, PMO]

Cayley contractions

If A is μ-strongly monotone and L-Lipschitz, then

$$C_{\gamma A} = 2R_{\gamma A} - I = 2(I + \gamma A)^{-1} - I$$

is a contraction with factor $\sqrt{1 - 4\gamma\mu/(1+\gamma L)^2}$

Remark need also Lipschitz condition

Proof [Page 20, PMO]

If, in addition, $A = \partial f$ where f is CCP, then $C_{\gamma A}$ converges with factor $(\sqrt{\mu/L} - 1)/(\sqrt{\mu/L} + 1)$ and optimal step $\gamma = 1/\sqrt{\mu L}$

Proof
[Linear Convergence and Metric Selection for Douglas-Rachford Splitting and ADMM, Giselsson and Boyd]

Requirements for contractions

	Operator A	**Function** f ($A = \partial f$)
Forward step $I - \gamma A$	μ-strongly monotone	μ-strongly convex L-smooth
Resolvent $R_A = (I + A)^{-1}$	μ-strongly monotone	μ-strongly convex L-smooth
Cayley $C_A = 2(I + A)^{-1} - I$	μ-strongly monotone L-Lipschitz	μ-strongly convex L-smooth

faster convergence

Key to contractions: strong monotonicity/convexity

Operator theory

Today, we learned to:

- **Use** conjugate functions to define duality
- **Define** monotone and cocoercive operators and their relations
- **Relate** subdifferential operator and monotonicity
- **Recognize** monotone operators in optimization problems
- **Apply** operators in algorithms: forward step, resolvent, Cayley
- **Understand requirements** for building contractions

Next lecture

- Operator splitting algorithms