Scaled Relative Graph (SRG)

Wotao Yin (UCLA Math Department)

Based on: Ernest Ryu, Robert R. Hannah, W. Yin. Scaled Relative Graph: Nonexpansive operators via 2D Euclidean Geometry. arXiv:1902.09788

One World Optimization Seminar
May 11th, 2020

Many optimization methods are fixed-point iterations: $x^{k+1} = T(x^k)$.

They are analyzed with inequalities, which are rigorous but often unintuitive.

Today, an alternative 2D geometric tool

- visual and intuitive
- serve as rigorous proofs
- give tight constants.

A sample result

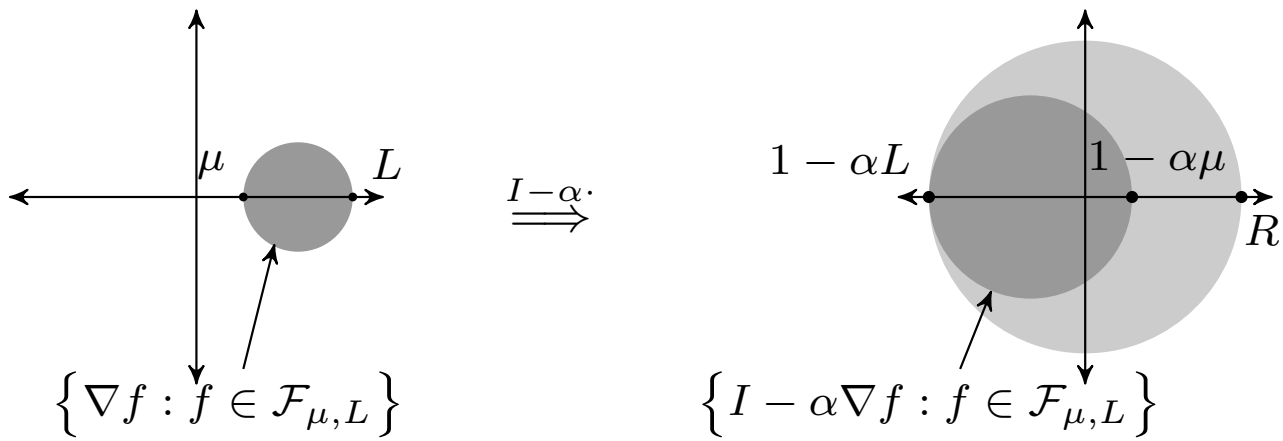
Fact: If f is μ-strongly convex and L-Lipschitz differentiable, then

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

converges linearly at sharp rate:

$$R = \max\{|1 - \alpha\mu|, |1 - \alpha L|\}.$$

Diagrams:



(We will make them a rigorous proof.)

Prior work that includes geometric illustrations

(Eckstein, 1989) and (Eckstein and Bertsekas, 1992) use disks to illustrate firm-nonexpansiveness and Lipschitz continuity

(Giselsson and Boyd, 2017; Banjac and Goulart, 2018) have illustrations on tight linear convergence rates. Lecture notes (Giselsson, 2015) used them more thoroughly.

Many have used geometric illustrations to build initial intuitions though wrote actual proofs with algebraic inequalities.

Fixed-point iterations

Find $T : \mathcal{H} \to \mathcal{H}$ such that $x^* = T(x^*)$ is a solution.

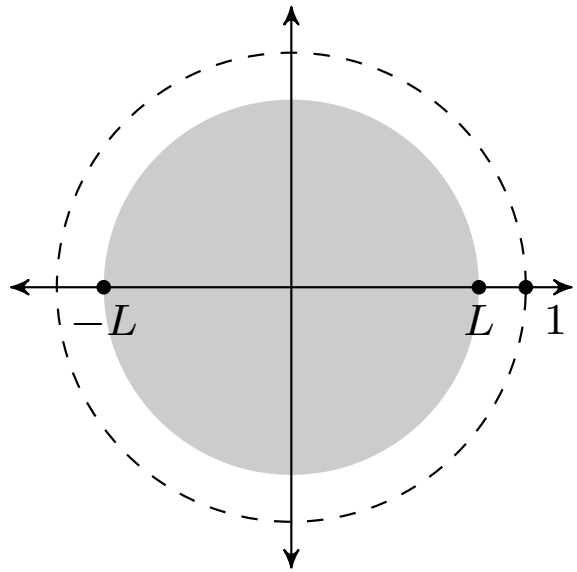
Example: under proper conditions,

1. $\min f(x) \;\Leftrightarrow\; x^* = (I - \alpha\nabla f)x^*$
2. $\min f(x) + g(x) \;\Leftrightarrow\; x^* = \mathbf{prox}_{\alpha f}(I - \alpha\nabla g)x^*$
3. $\min_{Ax+By=b} f(x) + g(y) \;\Leftrightarrow\; z^* = \frac{1}{2}(I + R_{\alpha A\partial f^*(A^T\cdot)}R_{\alpha B\partial g^*(B^T\cdot)-b})z^*$

To show $x^{k+1} = Tx^k$ converge, a standard approach takes 2 steps

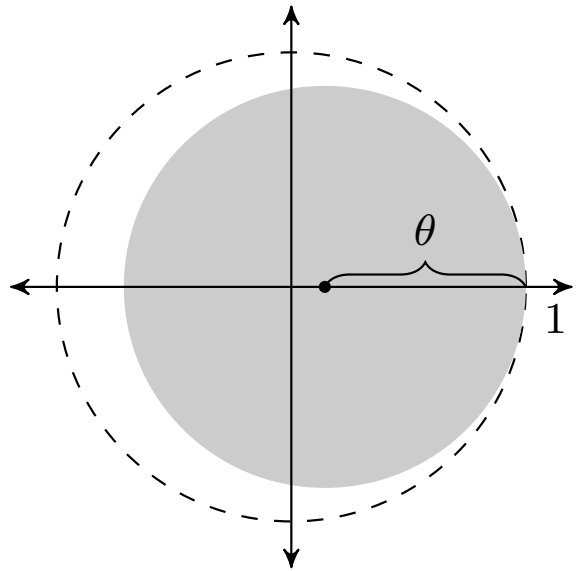
1. proving T is contractive or averaged
2. applying standard arguments.

Contractive operator



Banach fixed-point theorem: If T is contractive (L-Lipschitz with $L < 1$), then $x^{k+1} = Tx^k$ converges linearly to $x^\star = Tx^\star$.

Averaged operator



Krasnosel'skiĭ–Mann theorem:

If T with θ-averaged, $\theta \in (0, 1)$, and T has a fixed point, then $x^{k+1} = Tx^k$ converges to a fixed point with $\|x^k - Tx^k\|^2 = o(1/k)$.

Baillon, Bruck, and Reich (1978)

How to tell if T is contractive or averaged?

T is built from the scaling, addition, and inversion of identity, matrices, gradients, and subdifferentials.

Example: $T = \underbrace{\mathbf{prox}_{\alpha f}}_{(I+\alpha\partial f)^{-1}}(I - \alpha\nabla g)$

	original	transform $\mathcal{T}$	contractive / averaged
operator:	A, B	$T = \mathcal{T}(A, B)$	$\subseteq \mathcal{L}_L$ or $\mathcal{N}_\theta$
geometry:	$\mathcal{G}(A), \mathcal{G}(B)$ 2D shapes	$\mathcal{G}(T)$ new shape	$\subseteq \mathcal{G}(\mathcal{L}_L)$ or $\mathcal{G}(\mathcal{N}_\theta)$ enclosed in shape of $\mathcal{L}_L$ or $\mathcal{N}_\theta$

SRG of a (single/multi-valued) operator A

Pick $x \neq y$, $u \in Ax$ and $v \in Ay$. Plot a complex $z = re^{\phi \mathrm{i}}$ with

$$\text{size change:} \quad r := \frac{\|u - v\|}{\|x - y\|}$$

$$\text{rotation:} \quad \phi := \pm\angle(u - v, x - y).$$

For example, if $A = I$, $z \equiv (1, 0)$.

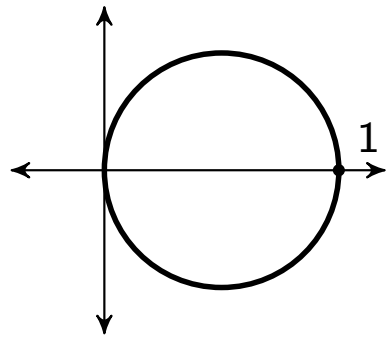
SRG consists of all such z:

$$\mathcal{G}(A) := \{z : x \neq y, u \in Ax, v \in Ay\} \left(\cup \{\infty\} \text{ if } A \text{ is multi-valued} \right)$$

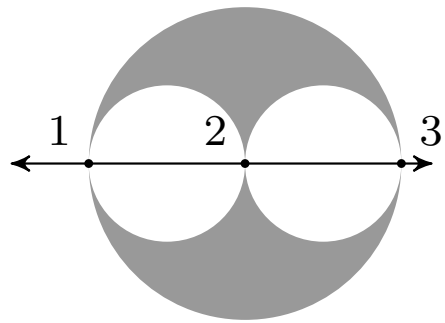
For operator class, $\mathcal{G}(\mathcal{A}) := \bigcup_{A \in \mathcal{A}} \mathcal{G}(A)$.

Examples of SRGs

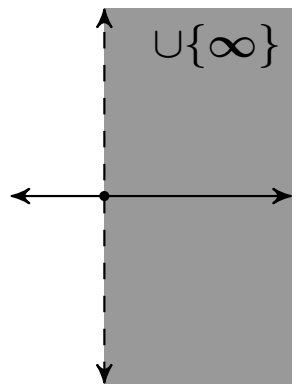
In $\mathbb{R}^2$, projection to any line:



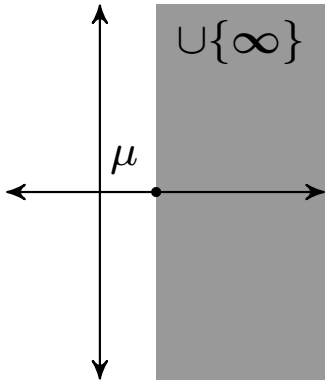
$$A \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \alpha \\ 2\beta \\ 3\gamma \end{bmatrix}$$



subdifferential of $\|\cdot\|_2$ in $\mathbb{R}^2$:

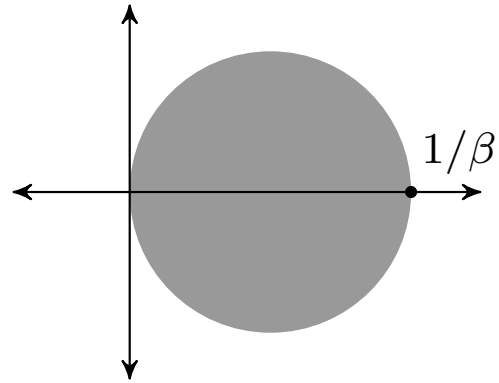


SRG of operator classes



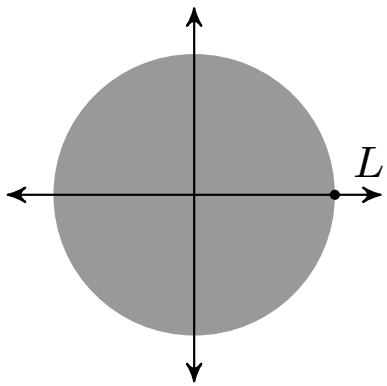
$\mathcal{M}_\mu$: μ-strongly monotone operator

$\partial\mathcal{F}_{\mu,\infty}$: subdiff'l of μ-strgly-cvx function

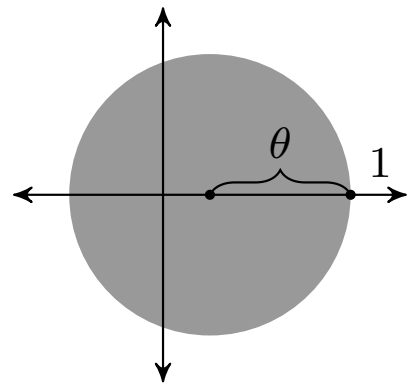


$\mathcal{C}_\beta$: β-cocoercive operator

$\partial\mathcal{F}_{0,\frac{1}{\beta}}$: gradient of $\frac{1}{\beta}$-Lip.diff.cvx function



$\mathcal{L}_L$: L-Lipschitz operator



$\mathcal{N}_\theta$: θ-averaged operator

Operator inclusion $\overset{?}{\Leftrightarrow}$ SRG inclusion

$T \in \mathcal{L}_L$ or $\mathcal{N}_\theta \quad \overset{?}{\Longleftrightarrow} \quad \mathcal{G}(T) \subseteq \mathcal{G}(\mathcal{L}_L)$ or $\mathcal{G}(\mathcal{N}_\theta)$

For any operator class $\mathcal{A}$, "$T \in \mathcal{A} \quad \Rightarrow \quad \mathcal{G}(T) \subseteq \mathcal{G}(\mathcal{A})$" follows from the definition.

The converse does not hold in general.

But fortunately, it does hold for $\mathcal{L}_L$ and $\mathcal{N}_\theta$.

SRG-full classes of operators

Definition: an operator class $\mathcal{A}$ is **SRG-full** if

$$T \in \mathcal{A} \quad \Leftrightarrow \quad \mathcal{G}(T) \subseteq \mathcal{G}(\mathcal{A}).$$

Theorem: An operator class defined by 1-homogeneous equations of $\|u-v\|^2, \|x-y\|^2, \langle u-v, x-y\rangle$ is SRG-full.

Therefore, classes $\mathcal{M}_\mu$, $\mathcal{C}_\beta$, $\mathcal{L}_L$, and $\mathcal{N}_\theta$ are SRG-full.

Transformation

Drawing $\mathcal{G}(\mathcal{T}(\mathcal{A}))$ is simplified by the following tools:

- $\mathcal{G}(\beta I + \alpha\mathcal{A}) = \beta + \alpha\mathcal{G}(\mathcal{A})$, for $\alpha, \beta \in \mathbb{R}$
- $\mathcal{G}(\mathcal{A}^{-1}) = (\mathcal{G}(\mathcal{A}))^{-1}$

and, under suitable conditions,

- $\mathcal{G}(\mathcal{A} \cap \mathcal{B}) = \mathcal{G}(\mathcal{A}) \cap \mathcal{G}(\mathcal{B})$
- $\mathcal{G}(\mathcal{A} + \mathcal{B}) = \mathcal{G}(\mathcal{A}) + \mathcal{G}(\mathcal{B})$
- $\mathcal{G}(\mathcal{A}\mathcal{B}) = \mathcal{G}(\mathcal{A}) \cdot \mathcal{G}(\mathcal{B})$

On left are operations $\alpha\cdot, +, \ ^{-1}, \cap, +, \circ$ in the space of operators.

On right are Minkowski-type operations in the complex plane.

Scaling and translation

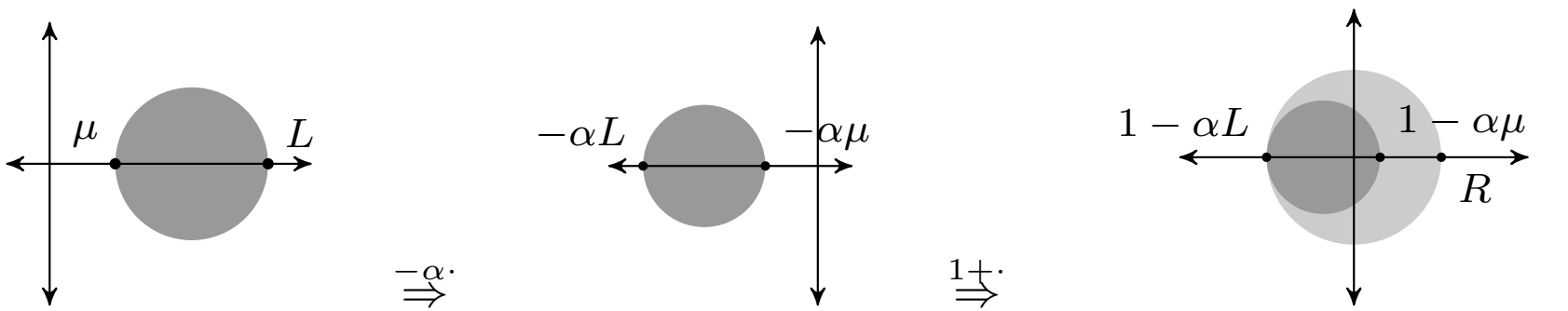
Tool: for $\alpha, \beta \in \mathbb{R}$,

$$\mathcal{G}(\beta I + \alpha \mathcal{A}) = \beta + \alpha \mathcal{G}(\mathcal{A}).$$

Fact: If f is μ-strongly convex and L-Lipschitz differentiable, then $x^{k+1} = x^k - \alpha \nabla f(x^k)$ converges linearly at sharp rate:

$$R = \max\{|1 - \alpha\mu|, |1 - \alpha L|\} < 1.$$

Proof by diagrams:



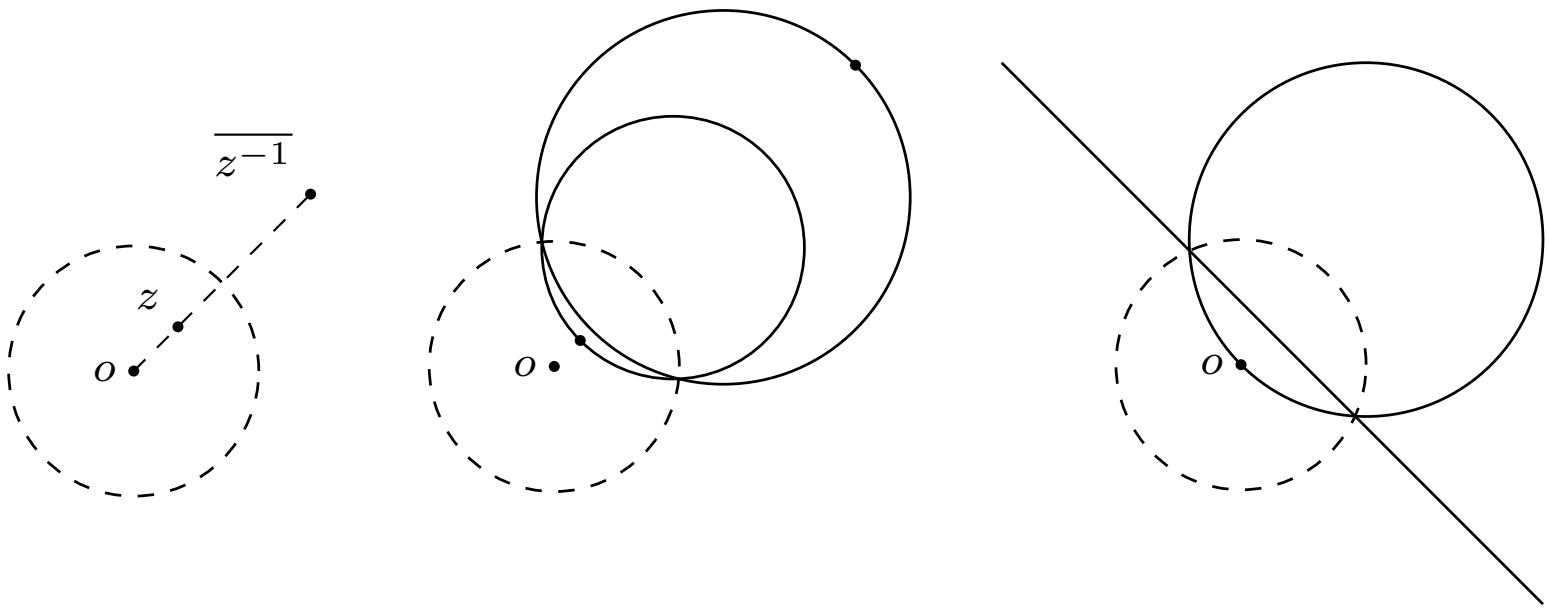
(uses scaling, translation)

Since $\mathcal{L}_R$ is SRG-full, the last diagram implies $I - \alpha \partial \mathcal{F}_{\mu,L} \subseteq \mathcal{L}_R$.

Transformation: inversion

Tool: $\mathcal{G}(\mathcal{A}^{-1}) = \big(\mathcal{G}(\mathcal{A})\big)^{-1}$ (operator inversion = geometric inversion).

Geometric inversion is known as *reflection in the unit circle*: $z \mapsto \overline{z^{-1}}$.



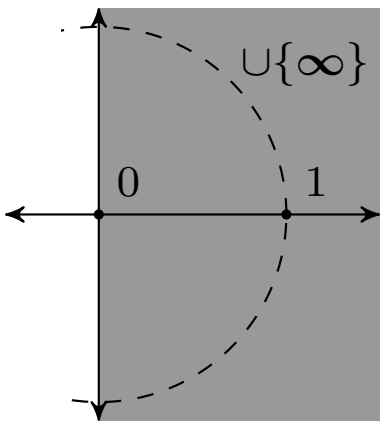
A line is a generalized circle with infinite radius.

Including this generalization, the inversion of a circle is a circle.

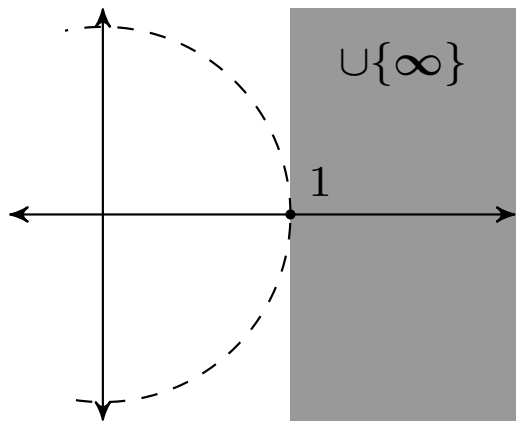
Transformation: inversion

Fact: if A is monotone and $\alpha > 0$, $J_{\alpha A} := (I + \alpha A)^{-1}$ is $1/2$-averaged (firmly nonexpansive). Iteration $x^{k+1} = J_{\alpha A}(x^k)$ has sublinear convergence.

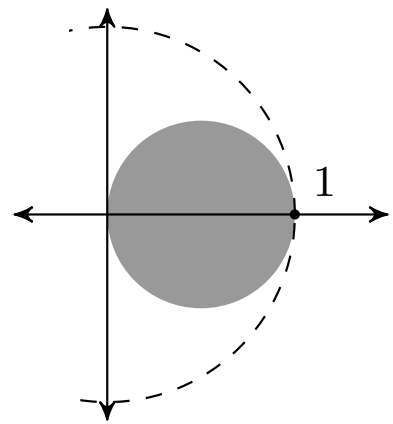
Proof by diagrams:



$\mathcal{G}(\alpha\mathcal{M})$



$\mathcal{G}(I + \alpha\mathcal{M})$

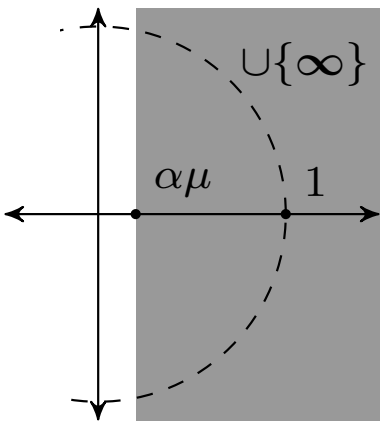


$\big(\mathcal{G}(I + \alpha\mathcal{M})\big)^{-1}$
$= \mathcal{G}\big((I + \alpha\mathcal{M})^{-1}\big)$

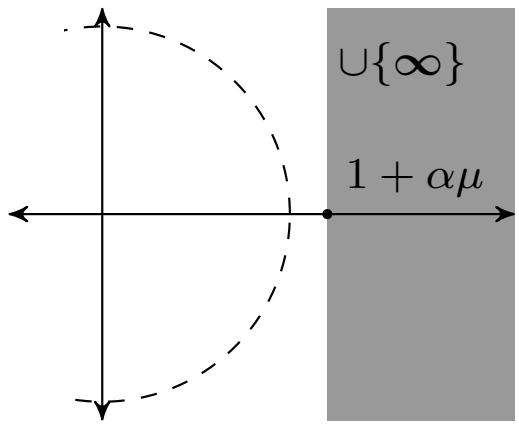
Since $\mathcal{G}\big((I + \alpha\mathcal{M})^{-1}\big) = \mathcal{G}(\mathcal{N}_{1/2})$ and $\mathcal{N}_{1/2}$ is SRG-full, $(I + \alpha\mathcal{M})^{-1} \subseteq \mathcal{N}_{1/2}$.

Fact: if A is μ-strongly monotone and $\alpha > 0$, $J_{\alpha A}$ is $1/(1+\alpha\mu)$-Lipschitz. Iteration $x^{k+1} = J_{\alpha A}(x^k)$ has linear convergence.

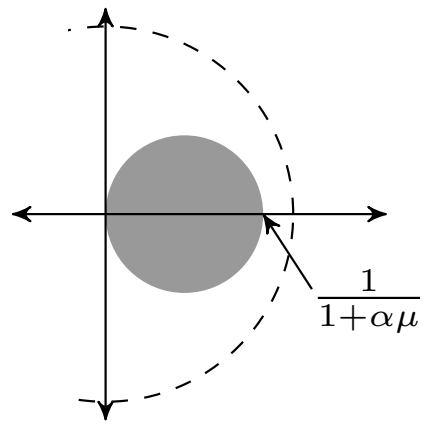
Proof by diagrams:



$\mathcal{G}(\alpha\mathcal{M})$



$\mathcal{G}(I+\alpha\mathcal{M})$



$\Big(\mathcal{G}(I+\alpha\mathcal{M})\Big)^{-1}$
$=\mathcal{G}\Big((I+\alpha\mathcal{M})^{-1}\Big)$

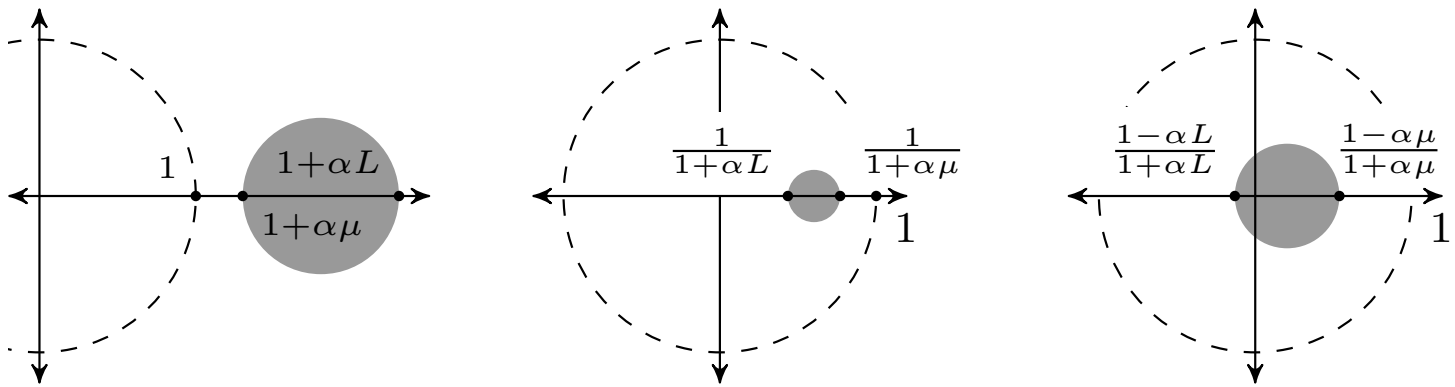
Since $\mathcal{G}\Big((I+\alpha\mathcal{M})^{-1}\Big) \subset \mathcal{G}\Big(\mathcal{L}_{\frac{1}{1+\alpha\mu}}\Big)$ and $\mathcal{L}_{\frac{1}{1+\alpha\mu}}$ is SRG-full,

$$(I+\alpha\mathcal{M})^{-1} \subset \mathcal{L}_{\frac{1}{1+\alpha\mu}}.$$

Fact[1]**:** If f is a μ-strongly convex L-Lipschitz differentiable function and $\alpha > 0$, we have

- $\mathbf{prox}_{\alpha f}$ is $\frac{1}{1+\alpha\mu}$-Lipschitz;
- $2\mathbf{prox}_{\alpha f} - I$ is R-Lipschitz for $R = \max\left\{\left|\frac{1-\alpha\mu}{1+\alpha\mu}\right|, \left|\frac{1-\alpha L}{1+\alpha L}\right|\right\}$, tight.

Proof by diagrams:



$\mathcal{G}(I + \alpha\partial\mathcal{F}_{\mu,L})$ $\mathcal{G}\left(\mathbf{prox}_{\alpha\mathcal{F}_{\mu,L}}\right)$ $\mathcal{G}\left(2\mathbf{prox}_{\alpha\mathcal{F}_{\mu,L}} - I\right)$

Middle implies $\mathbf{prox}_{\alpha\mathcal{F}_{\mu,L}} \subset \mathcal{L}_{\frac{1}{1+\alpha\mu}}$. Right implies $2\mathbf{prox}_{\alpha\mathcal{F}_{\mu,L}} - I \subseteq \mathcal{L}_R$.

[1] Giselsson and Boyd (2017, Thm 1)

Composition of operators

Theorem: If $\mathcal{A}, \mathcal{B}$ are SRG-full, then excluding $\infty \cdot \emptyset$ cases

$$\mathcal{G}(\mathcal{AB}) \supseteq \mathcal{G}(\mathcal{A})\mathcal{G}(\mathcal{B}).$$

In addition, if $\mathcal{A}$ or $\mathcal{B}$ satisfies the *arc property* then

$$\mathcal{G}(\mathcal{AB}) = \mathcal{G}(\mathcal{BA}) = \mathcal{G}(\mathcal{A})\mathcal{G}(\mathcal{B}).$$

Definition: An operator (class) $\mathcal{A}$ satisfies the *arc property* if

$$z \in \mathcal{G}(\mathcal{A}) \;\Rightarrow\; \text{either left arc or right arc } (z, \bar{z}) \subseteq \mathcal{G}(\mathcal{A}).$$

Convergence: alternating projections

Fact: For two closed convex sets $C, D \subset \mathbb{R}^n$ and $C \cap D \neq \emptyset$, iteration

$$x^{k+1} = \mathrm{Proj}_C \mathrm{Proj}_D x^k$$

converges to some $x^* \in C \cap D$.

Since projection to a closed convex set is $\frac{1}{2}$-averaged, this follows from the following result regarding $\mathcal{N}_{\theta_1}\mathcal{N}_{\theta_2}$.

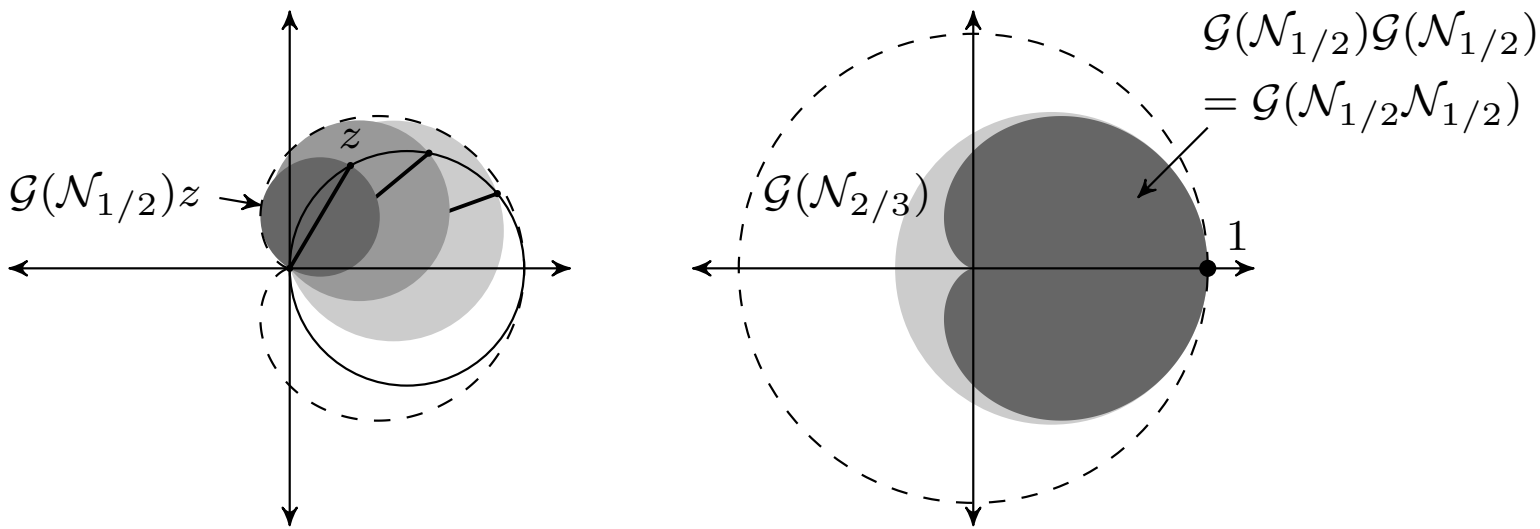
Composition of averaged operators

Fact[2]**:** Let $\mathcal{N}_\theta$ be the class of θ-averaged operators. Then,

$$\mathcal{N}_{\theta_1}\mathcal{N}_{\theta_2} \subseteq \mathcal{N}_\theta, \quad \theta = \frac{\theta_1 + \theta_2 - 2\theta_1\theta_2}{1 - \theta_1\theta_2}.$$

In particular, $\mathcal{N}_{1/2}\mathcal{N}_{1/2} \subseteq \mathcal{N}_{2/3}$.

Diagrams for $\mathcal{N}_{1/2}\mathcal{N}_{1/2}$:



[2]Ogura and Yamada (2002)

Application: Tight characterization of $\mathcal{N}_{\theta_1}\mathcal{N}_{\theta_2}$

Using geometric arguments, we can show:

Theorem: For $0 < \theta_1, \theta_2 < 1$, $\mathcal{G}(\mathcal{N}_{\theta_1}\mathcal{N}_{\theta_2})$ is the region enclosed by the curve (r, ψ) in polar coordinate:

$$r^2(\psi) - 2r(\psi)\left(\cos(\psi)(1-\theta_1)(1-\theta_2) + \theta_1\theta_2\right) + (1-2\theta_1)(1-2\theta_2) = 0.$$

Corollary: Formula of θ on last slide for $\mathcal{N}_{\theta_1}\mathcal{N}_{\theta_2} \subseteq \mathcal{N}_{\theta}$ is tight.

Huang, Ryu, and Yin (2020)

Application: Plug-n-play (PnP)

PnP: replace an operator (e.g., $\mathbf{prox}_{\mathrm{TV}}$) in classic optimization methods (e.g., forward-backward, ADMM) by a better denoising operator (e.g., BM3D, neural network)

Why? Use pre-trained denoisers when there is not sufficient data or time for end-to-end training.

Example: Forward-backward PnP denoising: let

$$H : \text{noisy image} \mapsto \text{less noisy image}$$

be a denoising operator (BM3D, DnCNN), and f be a data-fidelity function.

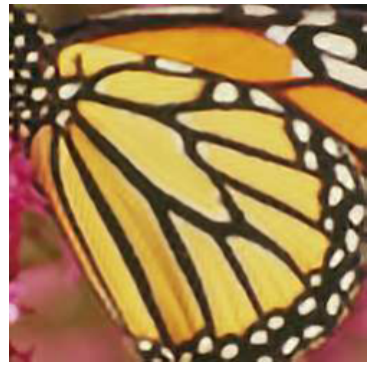
$$\textbf{PnP-FBS:} \quad x^{k+1} = H(x^k - \alpha \nabla f(x^k)).$$

Venkatakrishnan, Bouman, and Wohlberg (2013)

Experiment: Super resolution

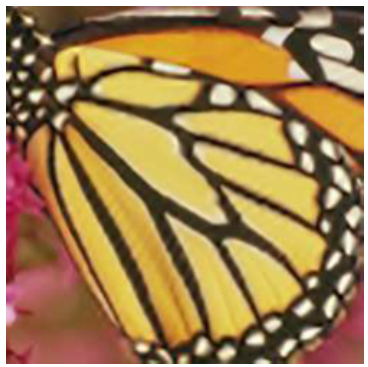
Low-res input

Other method

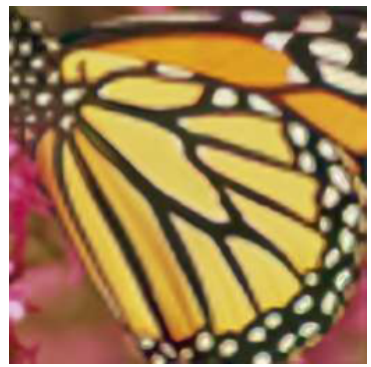


Other method

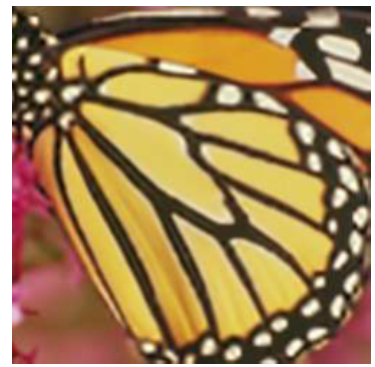
Other method



Other method



Other method



PnP-ADMM BM3D

Chan, Wang, and Elgendy (2017)

Experiment: Single photon imaging

Binary input

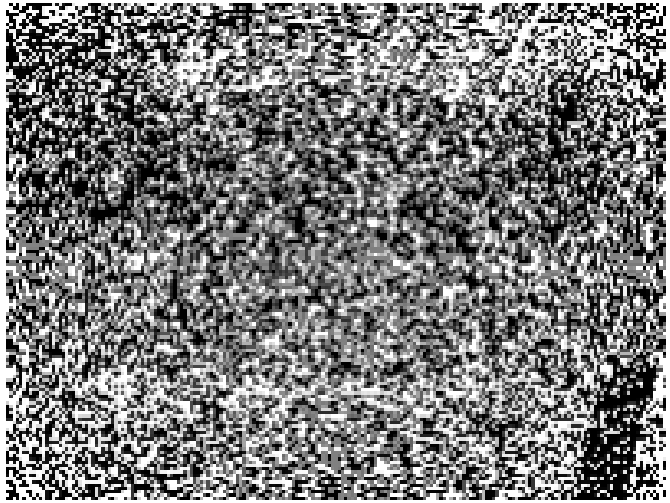
Other method

Other method

PnP-ADMM BM3D

Chan, Wang, and Elgendy (2017)

Zoom in



Binary input

Other method

Other method

PnP-ADMM BM3D

Chan, Wang, and Elgendy (2017)

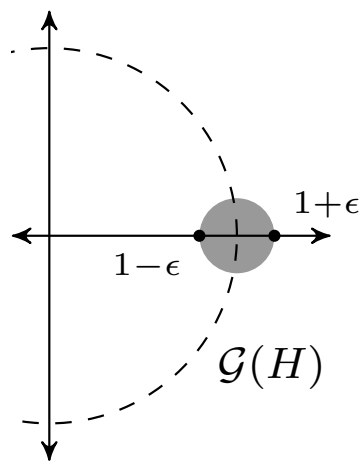
Convergence theory for PnP

Assume denoising operator

$$H : \text{noisy image} \mapsto \text{less noisy image}$$

is close to I in the following sense

$$\|(H-I)x-(H-I)y\|^2 \leq \epsilon^2\|x-y\|^2, \quad \forall x, y.$$

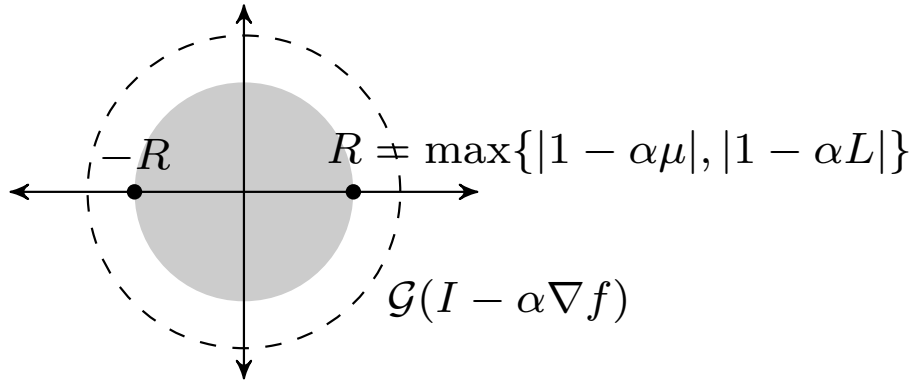


We can enforce this assumption in training using Real Spectrum Normalization

---

Ryu, Liu, Wang, Chen, Wang, and Yin (2019)

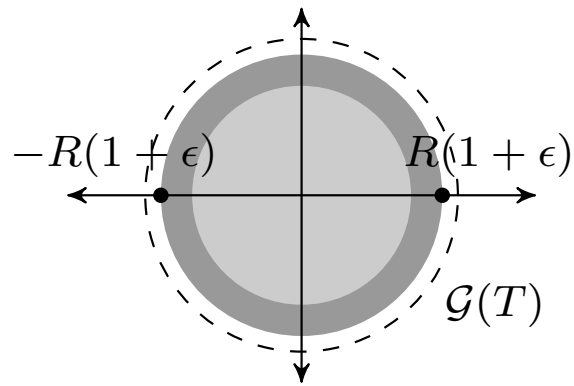
Assume f is μ-strongly convex and L-Lipschitz differentiable.



Theorem: The PnP forward-backward operator

$$T = H(I - \alpha\nabla f)$$

is contractive (thus, $x^{k+1} = Tx^k$ converges linearly) for $\epsilon < \frac{2\mu}{L-\mu}$ and $\frac{1}{\mu(1+\frac{1}{\epsilon})} < \alpha < \frac{2}{L} - \frac{1}{L(1+\frac{1}{\epsilon})}$.



Ryu, Liu, Wang, Chen, Wang, and Yin (2019)

Theorem: PnP-ADMM operator

$$T = \frac{1}{2}I + \frac{1}{2}(2H - I)(2\mathbf{prox}_f - I)$$

is contractive (thus, $x^{k+1} = Tx^k$ converges linearly) if $\epsilon < 1$ and $\alpha > \frac{\epsilon}{(1+\epsilon-2\epsilon^2)\mu}$.

Compare PnP-ADMM and PnP-FBS:

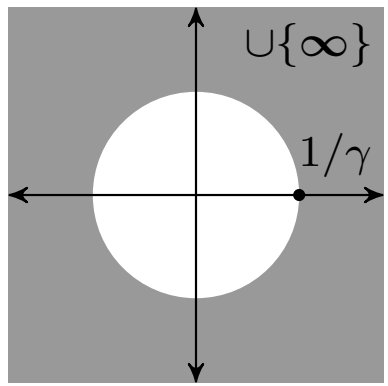
- With the same parameters, they have the same fixed points
- PnP-FBS is easier to implement.
- PnP-ADMM has better results (due to its wider allowed rangers of parameters)

Ryu, Liu, Wang, Chen, Wang, and Yin (2019)

Application: impossible results regarding "inverse Lipschitz continuity"

Let $\mathcal{L}_\gamma^{-1}$ be set of γ-inversely Lipschitz continuous operators, that is,

$$u \in Ax,\ v \in Ay \ \Rightarrow\ \gamma\|u - v\| \geq \|x - y\|.$$



$\mathcal{M}_{1/\gamma} \subset \mathcal{L}_\gamma^{-1}$ strictly: there are more inversely Lipschitz operators than strongly monotone operators.

Metric subregularity

$\mathcal{L}_\gamma^{-1}$ implies γ-metric subregularity; so, if a result fails to hold with $\mathcal{L}_\gamma^{-1}$, it will also fails to hold with γ-metric subregularity.

Definition: A is γ-metric subregular at x_0 for $y_0 \in Ax_0$ if $\mathrm{dist}(x, A^{-1}y_0) \leq \gamma \mathrm{dist}(y_0, Ax)$ in some neighborhood of x_0.

For convex function subdifferentials, equivalent to "error bound conditions[3]"

[3] Luo and Tseng (1993)

Relaxing assumptions for linear convergence :

For gradient descent, proximal-point, forward-backward, and certain ADMM algorithms[4], **relaxing strong monotonicity to metric subregularity (implied by inverse Lipschitz) maintains linear convergence**.

[4] Leventhal (2009); Bauschke, Noll, and Phan (2015); Liang, Fadili, and Peyré (2016); Latafat and Patrinos (2017); Karimi, Nutini, and Schmidt (2016); Bolte, Nguyen, Peypouquet, and Suter (2017); Drusvyatskiy and Lewis (2018); Necoara, Nesterov, and Glineur (2018); Ye, Yuan, Zeng, and Zhang (2018); Yuan, Zeng, and Zhang (2018); Zhang (2019)

This relaxation is not always possible

Define a parametrized class of Douglas-Rachford operators:

$$\mathcal{D}_{\alpha,\theta}(\mathcal{A},\mathcal{B}) = \{(1-\theta)I + \theta(2J_{\alpha A} - I)(2J_{\alpha B} - I) : A \in \mathcal{A}, B \in \mathcal{B}\}\,.$$

Theorem: Let $0 < 1/\gamma \leq L < \infty$ and $0 < \alpha, \theta < \infty$.
For $\mathcal{A} = \mathcal{M} \cap \mathcal{L}_L \cap \mathcal{M}_{1/\gamma}$ and $\mathcal{B} = \mathcal{M}$, we have, for proper ϵ,

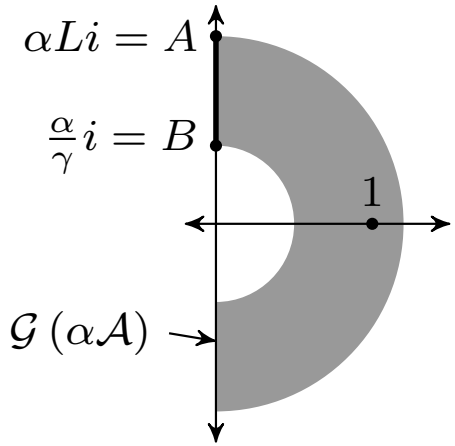
$$\mathcal{D}_{\alpha,\theta}(\mathcal{A},\mathcal{B}) \subseteq \mathcal{L}_{1-\epsilon} \quad \text{and} \quad \mathcal{D}_{\alpha,\theta}(\mathcal{B},\mathcal{A}) \subseteq \mathcal{L}_{1-\epsilon}.$$

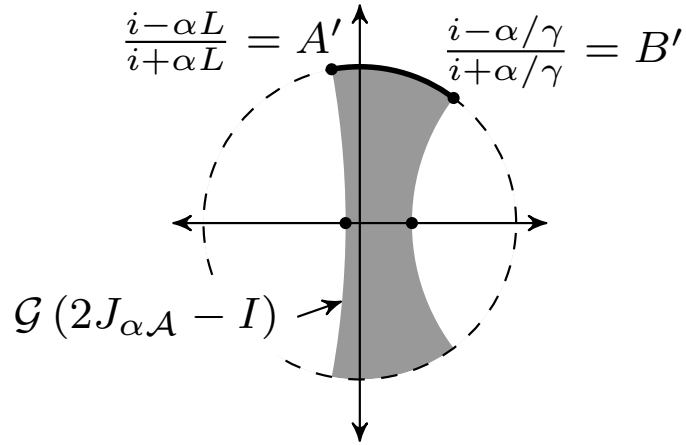
For $\mathcal{A} = \mathcal{M} \cap \mathcal{L}_L \cap \mathcal{L}_\gamma^{-1}$ and $\mathcal{B} = \mathcal{M}$, we have, for any $\epsilon \in (0,\infty)$,

$$\mathcal{D}_{\alpha,\theta}(\mathcal{A},\mathcal{B}) \not\subseteq \mathcal{L}_{1-\epsilon} \quad \text{and} \quad \mathcal{D}_{\alpha,\theta}(\mathcal{B},\mathcal{A}) \not\subseteq \mathcal{L}_{1-\epsilon}.$$

(Lions and Mercier, 1979, Proposition 4) for $\mathcal{D}_{\alpha,\theta}(\mathcal{A},\mathcal{B})$ and (Davis and Yin, 2017, Theorem 6) for $\mathcal{D}_{\alpha,\theta}(\mathcal{B},\mathcal{A})$, proved for subdifferentials of convex functions.

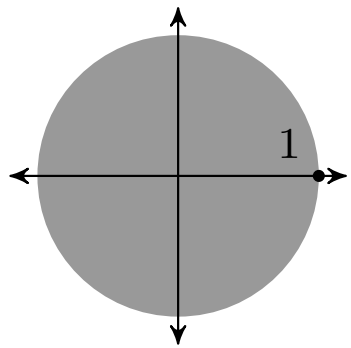
Proof by diagrams for "$\not\subseteq$":



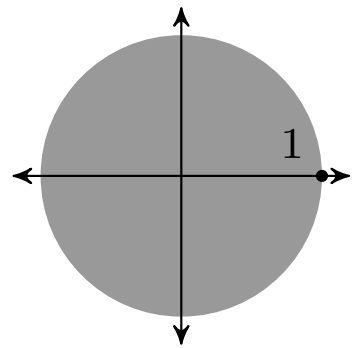


Line AB is mapped to arc $A'B'$.

Last graph $\times$



$=$



"$=$" relies on the arc property of $2J_{\alpha\mathcal{B}} - I$.

θ-averaging maintains the point 1, so $\not\subseteq \mathcal{L}_{1-\epsilon}$.

Summary

- SRG is a signature of an operator class.
- A few diagrams capture key ideas and can serve a rigorous proof.

For more results, such as addition of operators, tight bounds, and impossibilities, see:

Ernest Ryu, Robert Hannah, Wotao Yin. Scaled Relative Graph: Nonexpansive Operators via 2D Euclidean Geometry, arXiv:1902.09788.

Thank you!

References:

J. Eckstein. *Splitting methods for monotone operators with applications to parallel optimization*. PhD thesis, MIT, 1989.

J. Eckstein and D. P. Bertsekas. On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1–3), 1992.

P. Giselsson and S. Boyd. Linear convergence and metric selection for Douglas–Rachford splitting and ADMM. *IEEE Transactions on Automatic Control*, 62(2):532–544, 2017.

G. Banjac and P. J. Goulart. Tight global linear convergence rate bounds for operator splitting methods. *IEEE Transactions on Automatic Control*, 63(12):4126–4139, 2018.

P. Giselsson. Lunds universitet, lecture notes: Large-scale convex optimization, 2015. URL: `http://www.control.lth.se/education/doctorate-program/large-scale-convex-optimization/`. Last visited on 2018/12/01.

J.-B. Baillon, R. E. Bruck, and S. Reich. On the asymptotic behavior of nonexpansive mappings and semigroups in Banach spaces. *Houston Journal of Mathematics*, 4 (1):1–9, 1978.

N. Ogura and I. Yamada. Non-strictly convex minimization over the fixed point set of an asymptotically shrinking nonexpansive mapping. *Numerical Functional Analysis and Optimization*, 23(1–2):113–137, 2002.

X. Huang, E. K. Ryu, and W. Yin. Tight coefficients of averaged operators via scaled relative graph. *arXiv:1912.01593*, to appear in Journal of Mathematical Analysis and Applications, 2020.

S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg. Plug-and-Play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948, Austin, TX, USA, 2013. IEEE.

S. H. Chan, X. Wang, and O. A. Elgendy. Plug-and-Play ADMM for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2017.

E. K. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin. Plug-and-play methods provably converge with properly trained denoisers. In *International Conference on Machine Learning (ICML)*, Long Beach, CA, 2019.

Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: A general approach. *Annals of Operations Research*, 46-47(1):157–178, 1993.

D. Leventhal. Metric subregularity and the proximal point method. *Journal of Mathematical Analysis and Applications*, 360(2):681–688, 2009.

H. H. Bauschke, D. Noll, and H. M. Phan. Linear and strong convergence of algorithms involving averaged nonexpansive operators. *Journal of Mathematical Analysis and Applications*, 421(1):1–20, 2015.

J. Liang, J. Fadili, and G. Peyré. Convergence rates with inexact non-expansive operators. *Mathematical Programming*, 159(1):403–434, 2016.

P. Latafat and P. Patrinos. Asymmetric forward–backward–adjoint splitting for solving monotone inclusions involving three operators. *Comput. Optim. Appl.*, 68(1):57–93, 2017.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases (KDD)*, pages 795–811. Springer International Publishing, 2016.

J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3): 919–948, 2018.

I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 2018.

J. Ye, X. Yuan, S. Zeng, and J. Zhang. Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems. *Optimization Online Preprint*, 2018.

X. Yuan, S. Zeng, and J. Zhang. Discerning the linear convergence of admm for structured convex optimization through the lens of variational analysis. *Optimization Online Preprint*, 2018.

H. Zhang. New analysis of linear convergence of gradient-type methods via unifying error bound conditions. *Mathematical Programming*, Jan 2019.

P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979. doi: $10.1137/0716071$.

D. Davis and W. Yin. Faster convergence rates of relaxed Peaceman–Rachford and ADMM under regularity assumptions. *Mathematics of Operations Research*, 42(3): 783–805, 2017.