Oregon State University

On Convergence of Projected Gradient Descent for Minimizing a Large-Scale Quadratic over the Unit Sphere

Trung Vu, Raviv Raich, and Xiao Fu

School of EECS, Oregon State University, Corvallis, OR 97331-5501 USA

{vutru, raich, xiao.fu}@oregonstate.edu

Motivation

- Eigen-decomposition problem
  - repeatedly solving a sequence of problems of form

$$\max_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \quad \text{s.t. } \boldsymbol{x}^T\boldsymbol{x} = 1$$

- Trust-region subproblem
  - using the quadratic model to approximate the original objective function

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T\boldsymbol{x} + \boldsymbol{x}^T\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{x}$$
$$\text{s.t. } \|\boldsymbol{x}\| \le \Delta_k$$

Algorithm Power Method

1: Initialize $\boldsymbol{x}^{(0)}$ such that $[\boldsymbol{V}\boldsymbol{x}]_1 \neq 0$
2: **for** $t = 0, 1, \dots$ **do**
3: $\quad \boldsymbol{z}^{(t+1)} = \boldsymbol{A}\boldsymbol{x}^{(t)}$
4: $\quad \boldsymbol{x}^{(t+1)} = \dfrac{\boldsymbol{z}^{(t+1)}}{\left\|\boldsymbol{z}^{(t+1)}\right\|}$

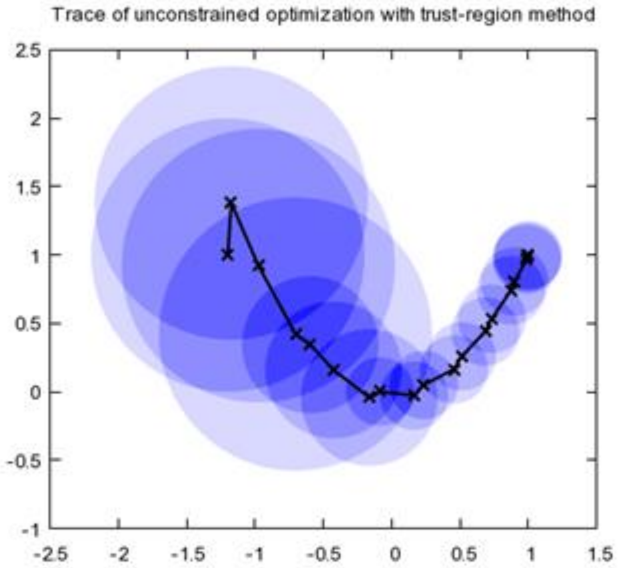
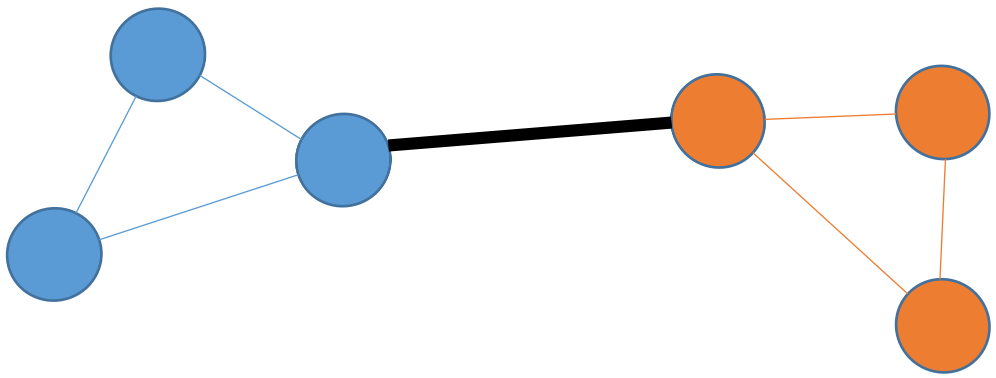


Image from https://optimization.mccormick.northwestern.edu/index.php/Trust-region_methods

Graph Partitioning as Constrained Quadratic Optimization



- *Bipartition*: cut a weighted, undirected graph into two subgraphs
  - roughly equals in size
  - the total weight of the cut edges is smallest
- Express the weight of a cut as a quadratic function of binary variables

$$\min_{\boldsymbol{x}\in\{-1,+1\}^n} \boldsymbol{x}^T\boldsymbol{L}\boldsymbol{x} \quad \text{s.t. } \mathbf{1}^T\boldsymbol{x}=0$$

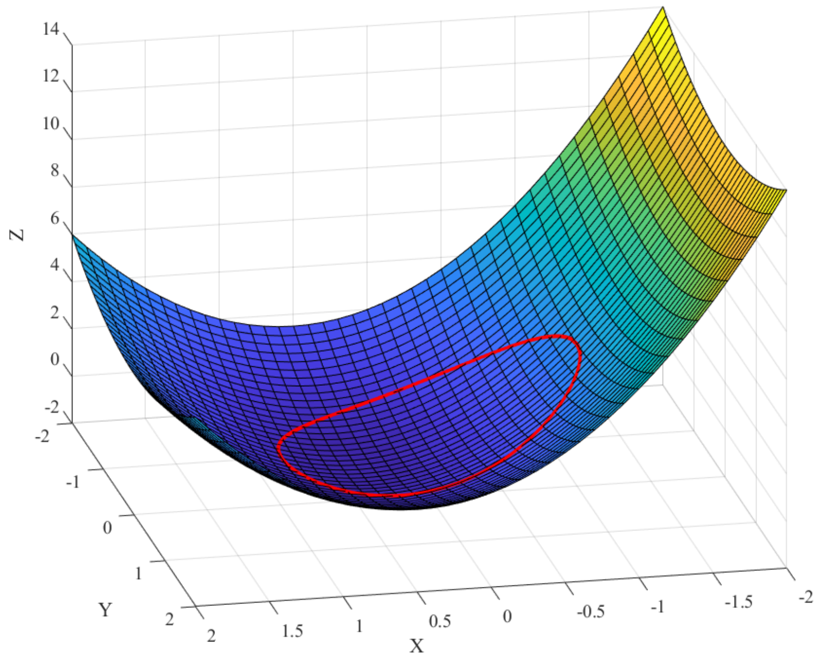
$\xrightarrow{\text{Relaxed}}$

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{x}^T\boldsymbol{L}\boldsymbol{x} \quad \text{s.t. } \mathbf{1}^T\boldsymbol{x}=0 \text{ and } \boldsymbol{x}^T\boldsymbol{x}=1$$

$\xrightarrow{\text{Regularized}}$

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{x}^T\boldsymbol{L}\boldsymbol{x}+\lambda\mathbf{1}^T\boldsymbol{x} \quad \text{s.t. } \|\boldsymbol{x}\|_2=1$$

$$\min_{x_1,x_2} \frac{1}{2}(4x_1^2 + x_2^2) - 2x_1 \quad s.t.\ x_1^2 + x_2^2 = 1$$



Problem Formulation

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} \qquad \text{subject to } \|\boldsymbol{x}\|^2 = 1$$

- $\|\cdot\|$ is the Euclidean norm
- $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric but not necessarily positive semidefinite
- Non-convex objective function with a non-convex constraint

Solution Properties

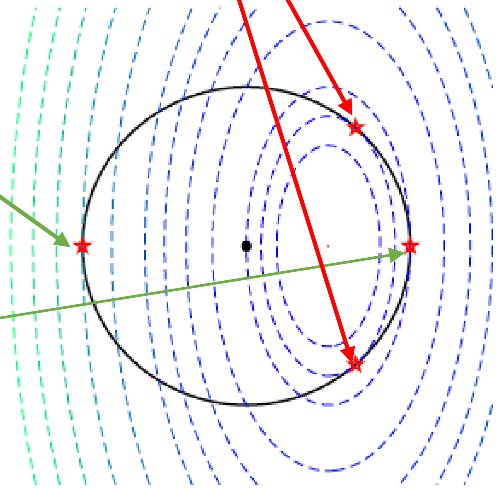
$$\mathcal{L}(\boldsymbol{x},\gamma) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} - \frac{1}{2}\gamma(\|\boldsymbol{x}\|^2 - 1)$$

- Stationary points $\begin{cases} \boldsymbol{x}_* \in \mathcal{S}^{n-1} \\ \boldsymbol{A}\boldsymbol{x}_* - \boldsymbol{b} = \gamma(\boldsymbol{x}_*)\cdot\boldsymbol{x}_* \end{cases}$
- Local minimum $\begin{cases} \boldsymbol{P}_{\boldsymbol{x}_*}^{\perp} = \boldsymbol{I} - \boldsymbol{x}_*\boldsymbol{x}_*^T \\ \gamma(\boldsymbol{x}_*) \le \lambda_{n-1}(\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}) \end{cases}$
- Global minimum $\gamma(\boldsymbol{x}_\star) \le \lambda_n(\boldsymbol{A}) \le \lambda_{n-1}(\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp})$

Local/Global minima
$\gamma = 1, \lambda_{n-1} = 8/3$

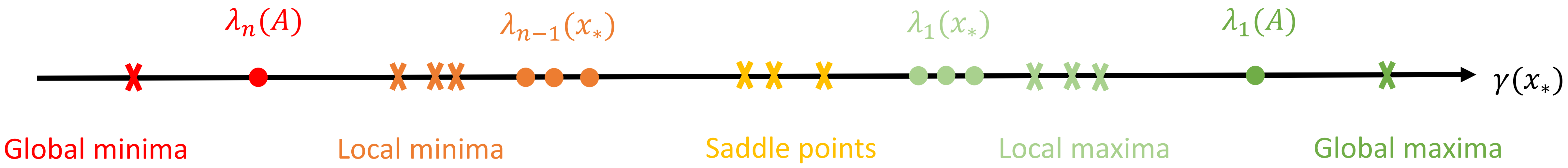
Global maximum
$\gamma = 6, \lambda_{n-1} = 1$

$$\min_{x_1,x_2} \frac{1}{2}(4x_1^2 + x_2^2) - 2x_1 \quad s.t.\ x_1^2 + x_2^2 = 1$$

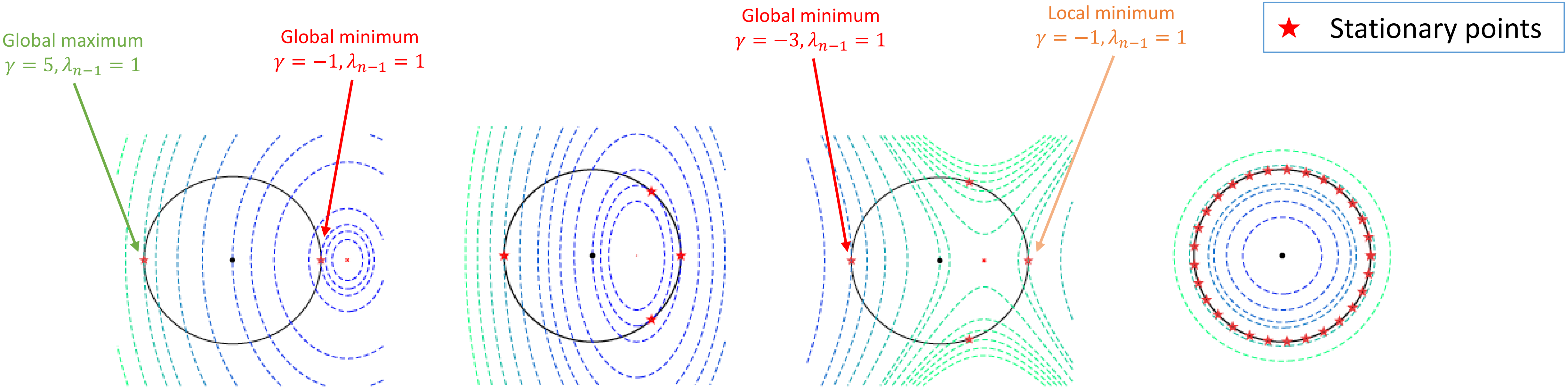


Local maximum
$\gamma = 2, \lambda_{n-1} = 1$

$$\boldsymbol{A} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$



2D-Examples



(a) $\boldsymbol{A} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$ (b) $\boldsymbol{A} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ (c) $\boldsymbol{A} = \begin{bmatrix} -2 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$ (d) $\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

Related Work

- Quadratic Constrained Quadratic Program with only one constraint (QCQP-1)
  - semidefinite relaxation (SDR)
  - Lagrangian relaxation

  → $O(n^2)$!

- Trust-region subproblem
  - solvable when n is small (by matrix factorizations)
  - for large n, iterative methods are considered
    - Parameterized eigenvalue problem [Sorensen'97]
    - Sequential Subspace Method (SSM) [Hager'01]

Projected Gradient Descent

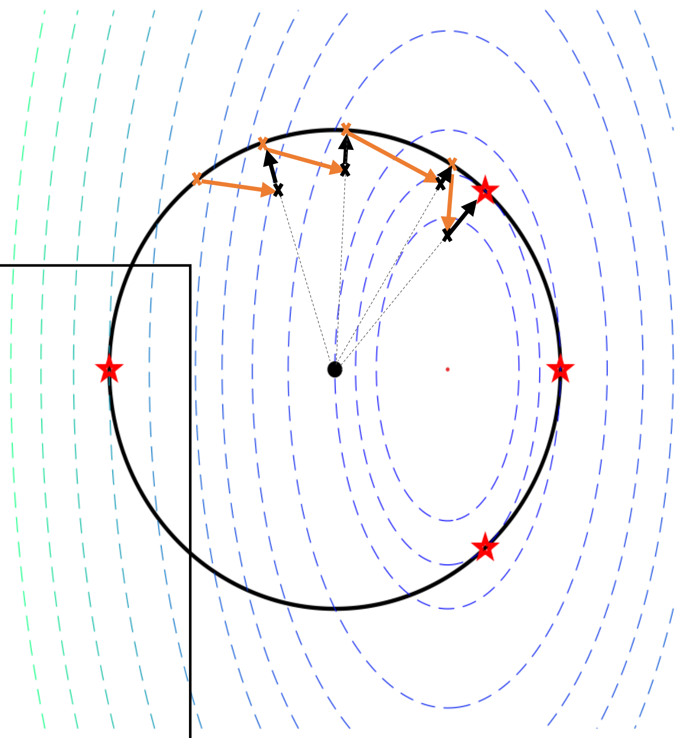
Algorithm Projected Gradient Descent (PGD)

1: Initialize $\boldsymbol{x}^{(0)} \in \mathcal{S}^{n-1}$
2: **for** $t = 0, 1, \ldots$ **do**
3: $\quad \boldsymbol{z}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha(\boldsymbol{A}\boldsymbol{x}^{(t)} - \boldsymbol{b})$
4: $\quad \boldsymbol{x}^{(t+1)} = \mathcal{P}_{\mathcal{S}^{n-1}}(\boldsymbol{z}^{(t+1)})$

$$\max_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \quad \text{s.t. } \boldsymbol{x}^T\boldsymbol{x} = 1$$

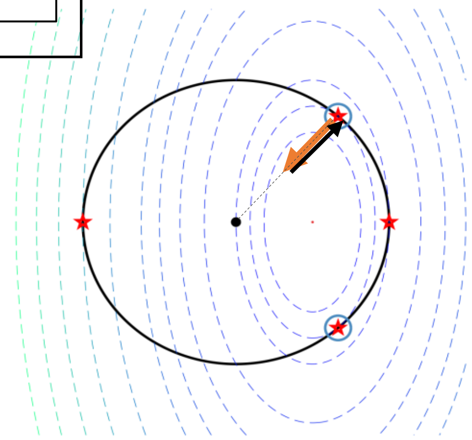
Power method $\quad \boldsymbol{\alpha} \to \infty$

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{x}^{(t)} - \infty(\boldsymbol{A}\boldsymbol{x}^{(t)}) \sim \boldsymbol{A}\boldsymbol{x}^{(t)}$$



- Fixed point

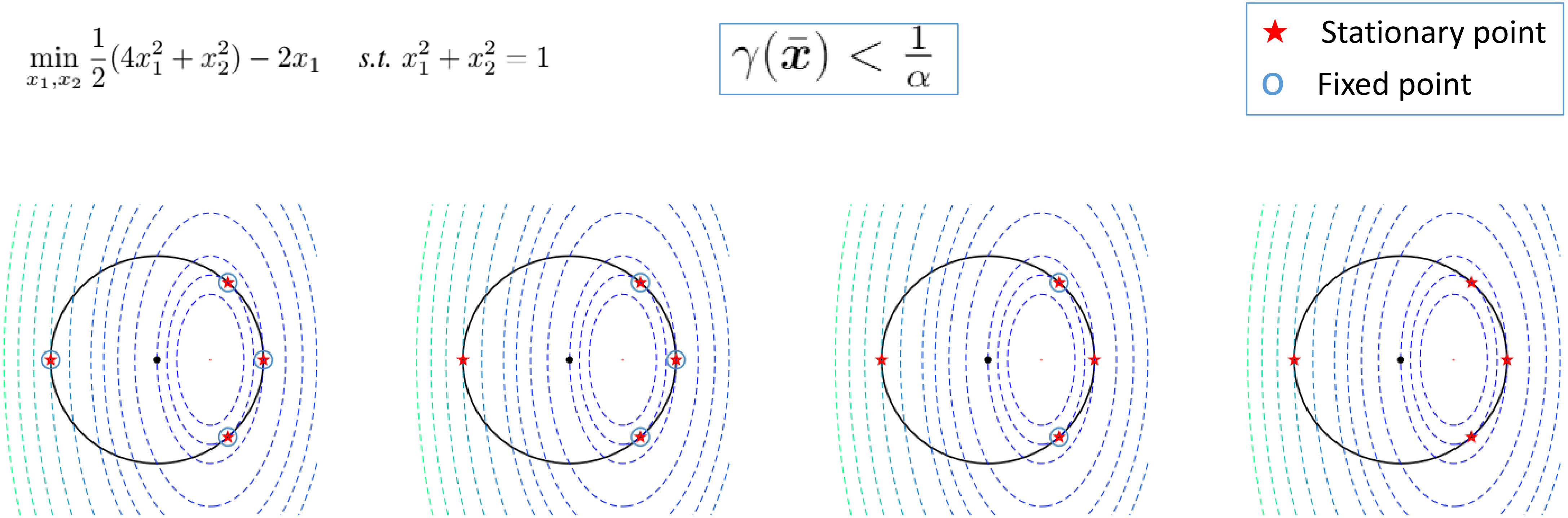
$$\mathcal{P}_{\mathcal{S}^{n-1}}(\bar{\boldsymbol{x}} - \alpha(\boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b})) = \bar{\boldsymbol{x}} \iff \begin{cases} \bar{\boldsymbol{x}} \in \mathcal{S}^{n-1} \\ \gamma(\bar{\boldsymbol{x}}) < \frac{1}{\alpha} \\ \boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b} = \gamma(\bar{\boldsymbol{x}}) \cdot \bar{\boldsymbol{x}} \end{cases}$$



Stationary Point versus Fixed Point

$$\min_{x_1,x_2} \frac{1}{2}(4x_1^2 + x_2^2) - 2x_1 \quad s.t.\ x_1^2 + x_2^2 = 1$$

$$\boxed{\gamma(\bar{\boldsymbol{x}}) < \frac{1}{\alpha}}$$



(a) $0 < \alpha < 1/6$ (b) $1/6 < \alpha < 1/2$ (c) $1/2 < \alpha < 1$ (d) $\alpha > 1$

Convergence Analysis

- Taylor Series Expansion of the Projection

$$\mathcal{P}_{\mathcal{S}^{n-1}}(\boldsymbol{x}+\boldsymbol{\delta}) = \frac{\boldsymbol{x}+\boldsymbol{\delta}}{\|\boldsymbol{x}+\boldsymbol{\delta}\|} = \mathcal{P}_{\mathcal{S}^{n-1}}(\boldsymbol{x}) + \frac{1}{\|\boldsymbol{x}\|}\left(\boldsymbol{I} - \frac{\boldsymbol{x}\boldsymbol{x}^T}{\|\boldsymbol{x}\|^2}\right)\boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^2)$$

- Recursion on the error vector

$$\begin{aligned}
\boldsymbol{\delta}^{(t+1)} &= \boldsymbol{x}^{(t+1)} - \boldsymbol{x}_* = \mathcal{P}_{\mathcal{S}^{n-1}}\left(\boldsymbol{x}^{(t)} - \alpha(\boldsymbol{A}\boldsymbol{x}^{(t)} - \boldsymbol{b})\right) - \boldsymbol{x}_* \\
&= \mathcal{P}_{\mathcal{S}^{n-1}}\left(\boldsymbol{x}_* - \alpha(\boldsymbol{A}\boldsymbol{x}_* - \boldsymbol{b}) + (\boldsymbol{I} - \alpha\boldsymbol{A})\boldsymbol{\delta}^{(t)}\right) - \boldsymbol{x}_* \\
&= \underbrace{\frac{1}{1-\alpha\gamma(\boldsymbol{x}_*)}(\boldsymbol{I} - \boldsymbol{x}_*\boldsymbol{x}_*^T)(\boldsymbol{I} - \alpha\boldsymbol{A})}_{\rho_\alpha}\boldsymbol{\delta}^{(t)} + O(\|\boldsymbol{\delta}^{(t)}\|^2)
\end{aligned}$$

If $\rho_\alpha < 1$ and $\delta^{(0)}$ is sufficiently small, the error series behaves similar to a geometric series

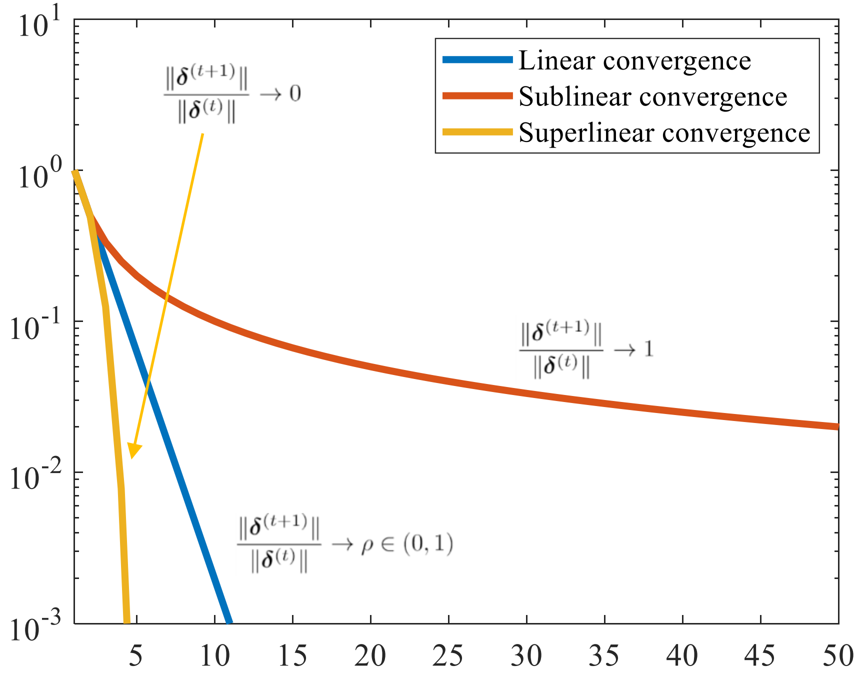
Rate of Convergence

- **Our results:**
  - PGD converges linearly locally to any strict local minimum with appropriate choice of α
  - The asymptotic rate of convergence is given by

$$\rho_\alpha(\boldsymbol{x}_*) = \max_{1\le i\le n-1} \frac{\left|1-\alpha\lambda_i(\boldsymbol{P}_{\boldsymbol{x}_*}^\perp \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^\perp)\right|}{1-\alpha\gamma(\boldsymbol{x}_*)} < 1$$

  - Optimizing over the step size yields faster convergence

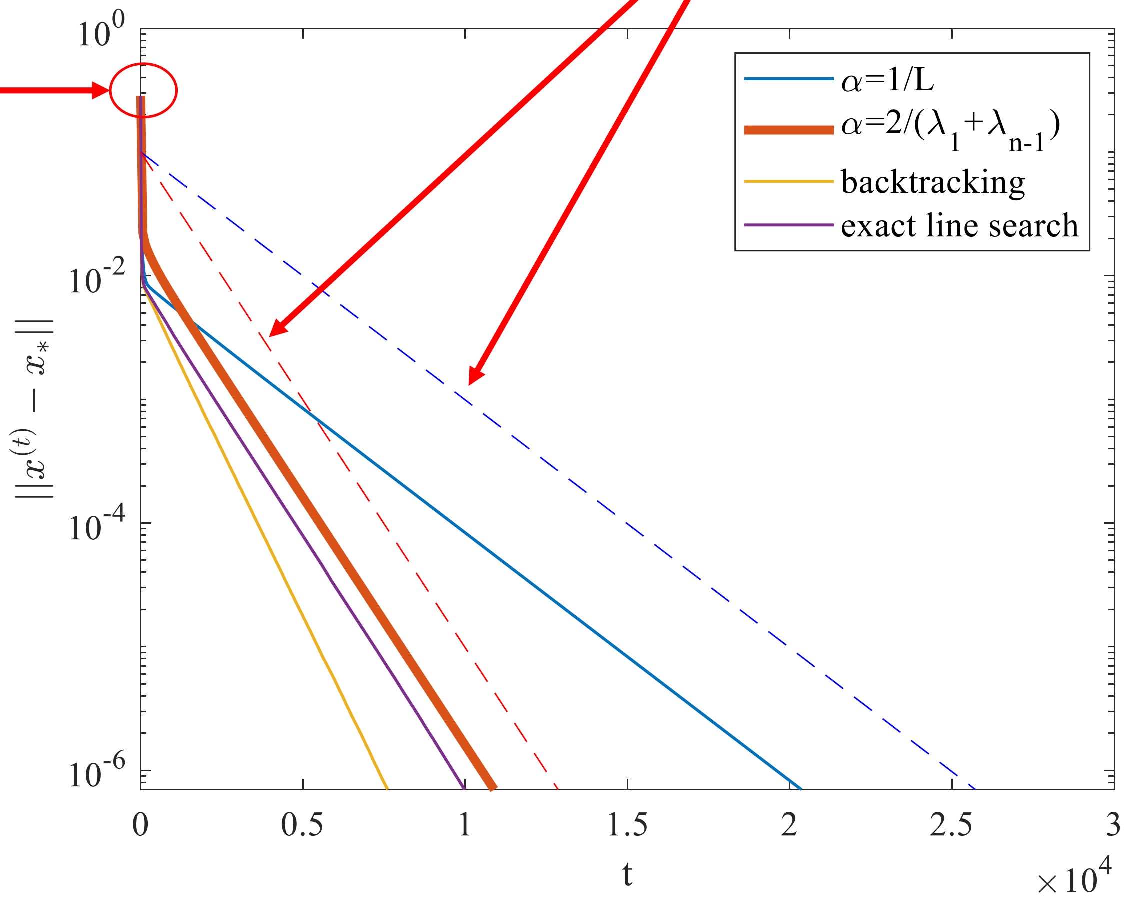
$$\alpha_* = \operatorname*{argmin}_{\alpha>0} \max_{1\le i\le n-1} \frac{\left|1-\alpha\lambda_i(\boldsymbol{P}_{\boldsymbol{x}_*}^\perp \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^\perp)\right|}{1-\alpha\gamma(\boldsymbol{x}_*)} = \frac{2}{\lambda_1(\boldsymbol{P}_{\boldsymbol{x}_*}^\perp \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^\perp)+\lambda_{n-1}(\boldsymbol{P}_{\boldsymbol{x}_*}^\perp \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^\perp)}$$



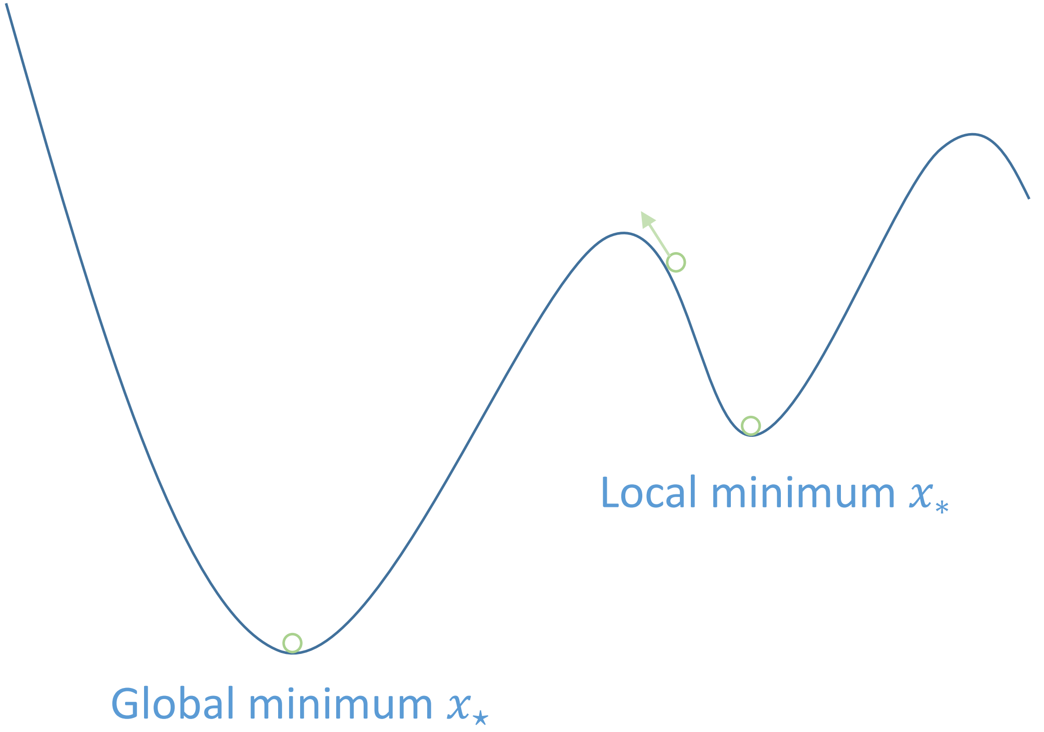
Numerical Evaluation

Theoretical analysis of $\rho_\alpha(x_*)^t$

Initialize $x^{(0)}$ near the local minimum x_*



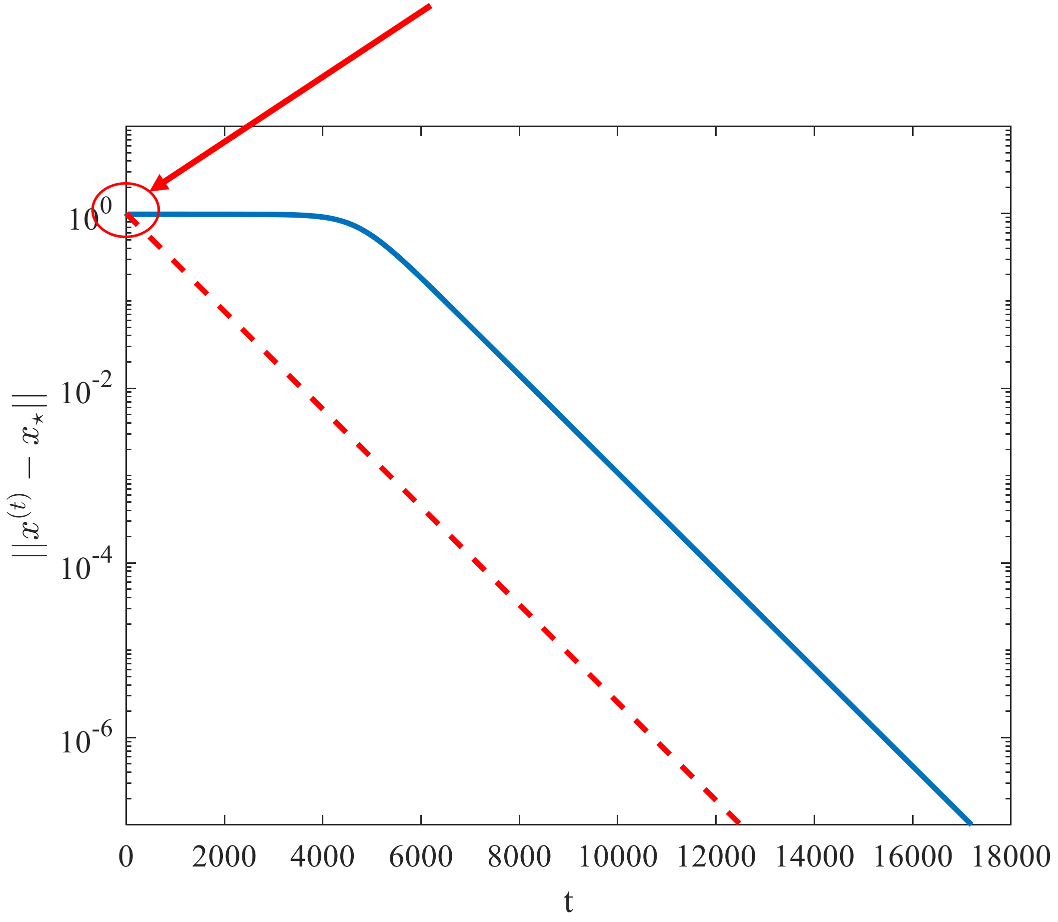
Escape from Local Minima

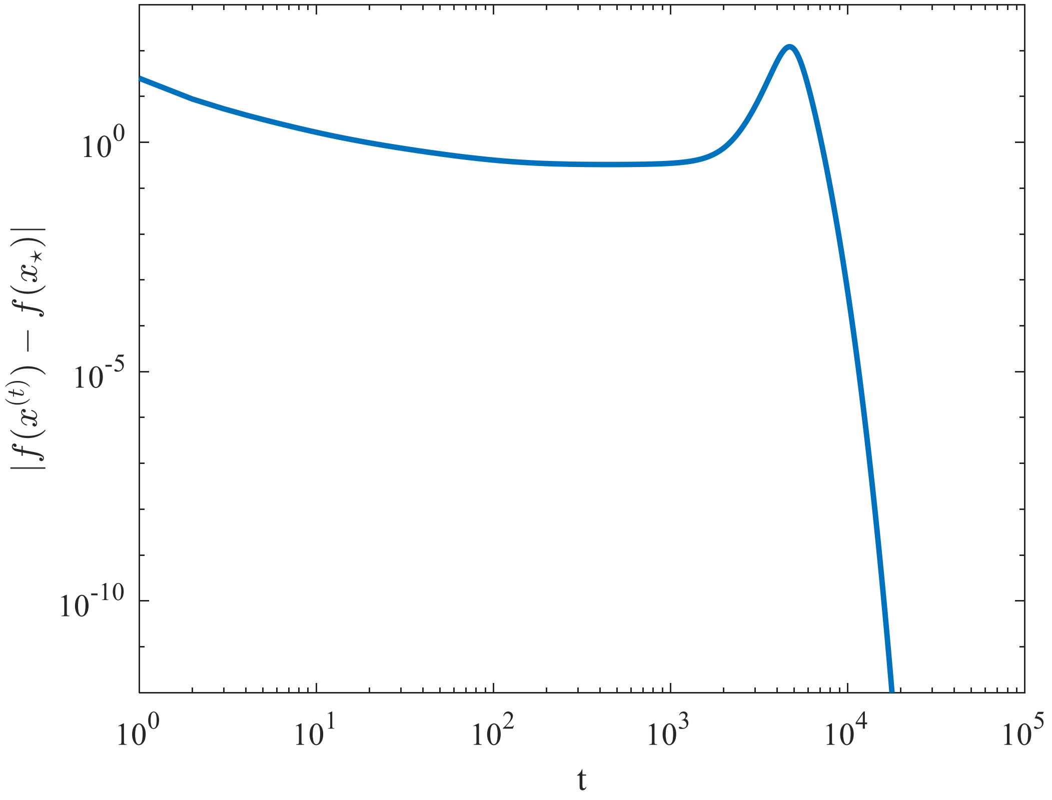


- For $\alpha > \frac{2}{\lambda_1(x_*)+\gamma(x_*)}$, the error series *w.r.t* x_* tends to diverge since $\rho_\alpha > 1$
- Define $g(\alpha, x_*) = \alpha(\lambda_1(x_*) + \gamma(x_*))$
- *Conjecture*:

Assume there exists sufficiently large α satisfying $g(\alpha, x_\star) < 2$ for any global minimum $x_\star$ and $g(\alpha, x_*) \geq 2$ for any strict local minimum x_*. Then PGD with step size α converges to one of the optimal solutions $x_\star$ at an asymptotic geometric rate of $\rho\_\alpha\ (x_\star)$.

Initialize $x^{(0)}$ near the local minimum x_*





Conclusion and Future Works

- Conclusion
  - showed PGD converges linearly to a strict local minimum in its neighborhood
  - provided the closed-form expression for asymptotic convergence rate
  - identified ways of achieving optimal rate of convergence near the optimum

- Future works
  - minimizing a quadratic over an ellipsoid
  - acceleration of gradient projection using momentum
  - analysis of convergence to a continuum of optima

THANK YOU!

References

1. Danny C Sorensen, “Minimization of a large-scale quadratic function subject to a spherical constraint,” *SIAM Journal on Optimization*, vol. 7, no. 1, pp. 141–161, 1997.
2. William W Hager, “Minimizing a quadratic over a sphere,” *SIAM Journal on Optimization*, vol. 12, no.1, pp. 188–208, 2001.