Composite Self-concordant Minimization

UT Simons Seminar Series
October 2nd, 2015

Anastasios Kyrillidis

Recall

$$\begin{aligned} \underset{\mathbf{X} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad & g(\mathbf{X}) \\ \text{subject to} \quad & \mathbf{X} \succeq 0 \end{aligned}$$

Recall

$$\begin{aligned} \underset{\mathbf{X}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad & g(\mathbf{X}) \\ \text{subject to} \quad & \mathbf{X} \succeq 0 \end{aligned}$$

Assumptions: $g(\mathbf{X})$ is convex, Lipschitz gradient and strongly convex.

$$\mu\mathbf{I} \preceq \nabla^2 g(\mathbf{X}) \preceq L\mathbf{I}$$

Recall

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad g(\mathbf{U}\mathbf{U}^\top)$$

$$\text{where} \qquad r \leq n$$

$$\left(\text{since} \quad \mathbf{U}\mathbf{U}^\top \succeq 0\right)$$

Assumptions: $g(\mathbf{X})$ is convex, Lipschitz gradient and strongly convex.

$$\mu\mathbf{I} \preceq \nabla^2 g(\mathbf{X}) \preceq L\mathbf{I}$$

Recall

$$\underset{\mathbf{U} \in \mathbb{R}^{n \times r}}{\text{minimize}} \quad g(\mathbf{U}\mathbf{U}^\top)$$

- Provable linear convergence rate guarantees, even under non-convexity.
- First-order oracle (i.e., at most $\nabla g(\cdot)$).

"Dropping convexity for faster semi-definite optimization", Bhojanapalli, Kyrillidis, Sanghavi, 2015

Assumptions: $g(\mathbf{X})$ is convex, Lipschitz gradient and strongly convex.

$$\mu \mathbf{I} \preceq \nabla^2 g(\mathbf{X}) \preceq L \mathbf{I}$$

In this talk

$$\underset{\mathbf{X} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad f(\mathbf{X}) + \lambda \cdot g(\mathbf{X})$$

Composite convex optimization

In this talk

$$\underset{\mathbf{X} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad f(\mathbf{X}) + \lambda \cdot g(\mathbf{X})$$

Composite convex optimization

In this talk

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad f(\mathbf{X}) + \lambda \cdot g(\mathbf{X})$$

- $g(\mathbf{X})$ is a convex function, but possibly non-smooth!

- $g(\mathbf{X})$ ``models`` a-priori knowledge.

Composite convex optimization

In this talk

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad f(\mathbf{X}) + \lambda \cdot g(\mathbf{X})$$

- $g(\mathbf{X})$ is a convex function, but possibly non-smooth!
- $g(\mathbf{X})$ ``models`` a-priori knowledge.

Examples: $g(\mathbf{X}) \equiv \|\mathbf{X}\|_1$ (for sparsity),

$g(\mathbf{X}) \equiv \|\mathbf{X}\|_*$ (for low-rankness),

$\vdots$

Composite convex optimization

In this talk

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad f(\mathbf{X}) + \lambda \cdot g(\mathbf{X})$$

- $f(\mathbf{X})$ is a convex function, but possibly non-globally Lipschitz gradient and strongly convex!
- $f(\mathbf{X})$ is usually a data fidelity term.

Composite convex optimization

In this talk

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad f(\mathbf{X}) + \lambda \cdot g(\mathbf{X})$$

- $f(\mathbf{X})$ is a convex function, but possibly <u>non-globally Lipschitz gradient and strongly convex!</u>
- $f(\mathbf{X})$ is usually a data fidelity term.

Examples: $f(\mathbf{X}) \equiv \frac{1}{2}\|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_2^2$ (least squares, MC, PT...),

$f(\mathbf{X}) = -\log\det(\mathbf{X}) + \langle \mathbf{X}, \mathbf{C}\rangle$ (GRMF)

$\vdots$

Composite convex optimization

In this talk

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad f(\mathbf{X}) + \lambda \cdot g(\mathbf{X})$$

- $f(\mathbf{X})$ is a convex function, but possibly <u>non-globally Lipschitz gradient and strongly convex!</u>
- $f(\mathbf{X})$ is usually a data fidelity term.

<u>Assumption</u>: $g(\mathbf{X})$ has a ``tractable`` proximity operator:

$$\text{prox}_{\lambda g}(\mathbf{W}) = \arg\min_{\mathbf{X}} \; g(\mathbf{X}) + \frac{1}{2\lambda}\|\mathbf{X}-\mathbf{W}\|_F^2$$

Composite convex optimization

In this talk

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad f(\mathbf{X}) + \lambda \cdot g(\mathbf{X})$$

- $f(\mathbf{X})$ is a convex function, but possibly <u>non-globally Lipschitz gradient and strongly convex!</u>
- $f(\mathbf{X})$ is usually a data fidelity term.

<u>Assumption</u>: $g(\mathbf{X})$ has a ``tractable`` proximity operator:

$$\text{prox}_{\lambda\|\cdot\|_1}(\mathbf{W}) = \text{SoftThresh}(\mathbf{W}, \lambda)$$

Note

$$\underset{\mathbf{X} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad f(\mathbf{X}) + \lambda \cdot g(\mathbf{X})$$

$$\Longleftrightarrow$$

$$\underset{\mathbf{x} \in \mathbb{R}^{n}}{\text{minimize}} \quad f(\mathbf{x}) + \lambda \cdot g(\mathbf{x})$$

What is known for

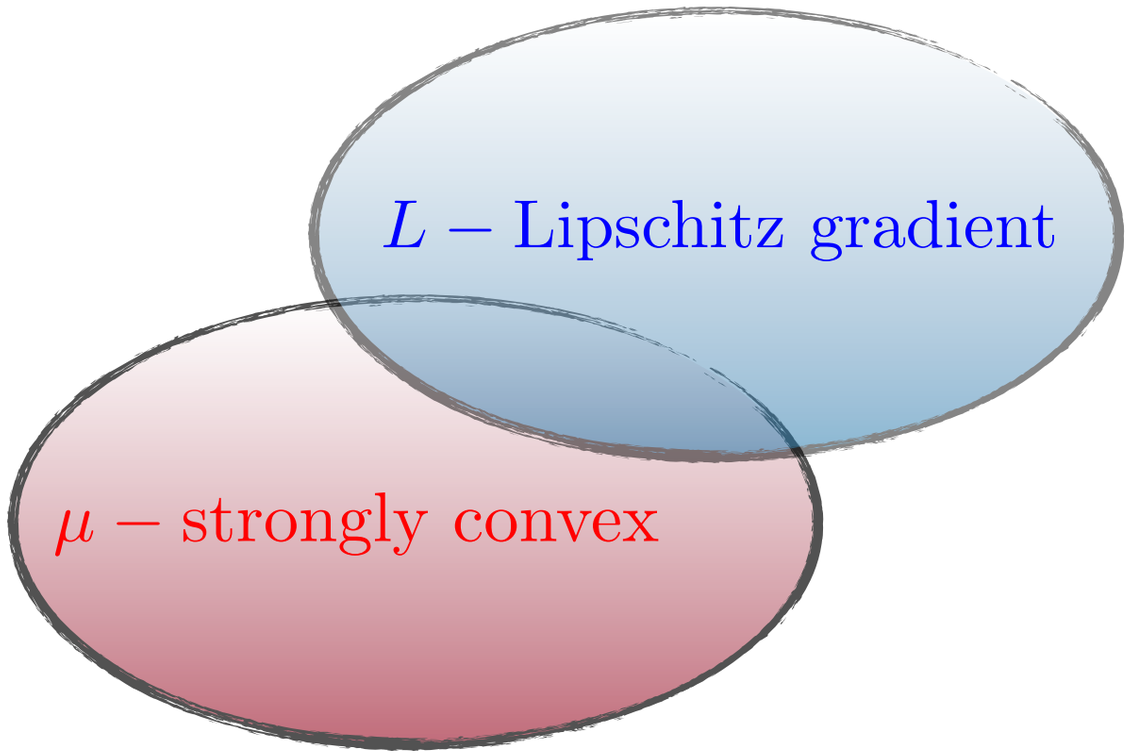
$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad f(\mathbf{x}) + \lambda \cdot g(\mathbf{x})$$

- Methods inseparable with Lipschitz gradient and strongly convex assumption:
  - Proximal first order methods [CW05, Nes07, EB92, SRB11]
  - Accelerated first order methods [BT09]
  - Smoothing techniques [Nes03]
  - Proximal Newton methods [BF12, LSS12, TKC15]
  - ADMM / ALM approaches [BPCPE12, GM12]
  - Primal-dual approaches [TC15] (and references therein)

- Methods inseparable with Lipschitz gradient and strongly convex assumption:
  - Proximal first order methods [CW05, Nes07, EB92, SRB11]
  - Accelerated first order methods [BT09]
  - Smoothing techniques [Nes03]
  - Proximal Newton methods [BF12, LSS12, TKC15]
  - ADMM / ALM approaches [BPCPE12, GM12]
  - Primal-dual approaches [TC15] (and references therein)

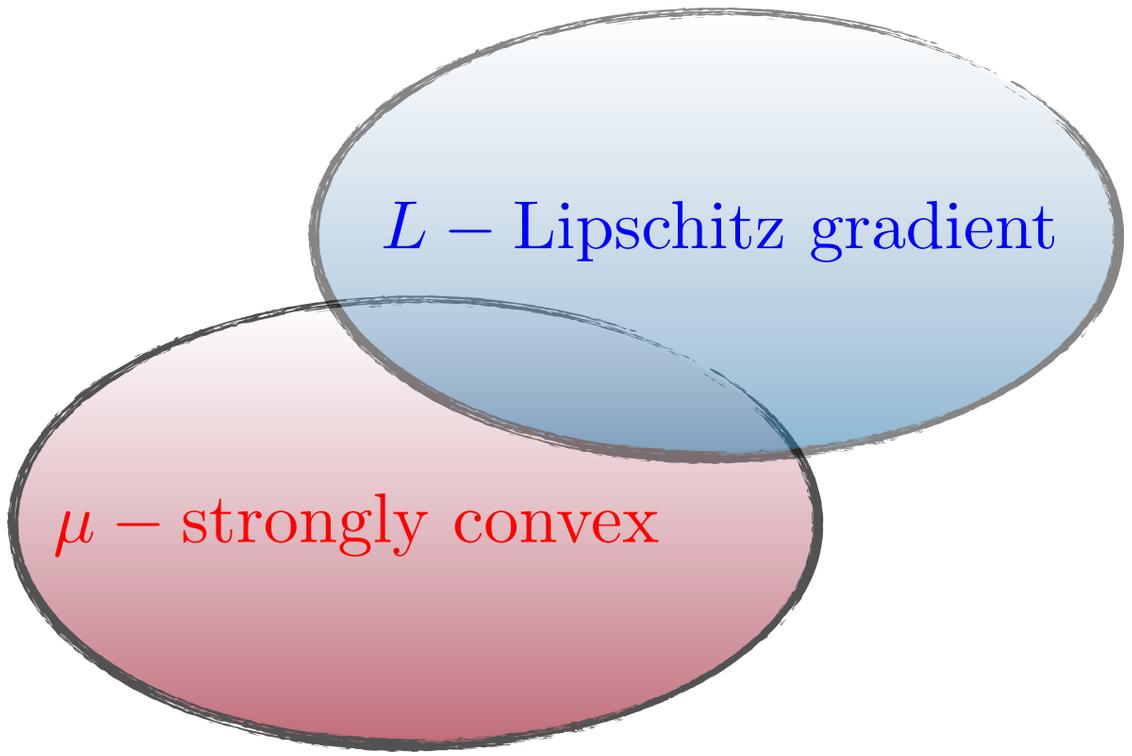
- Self-concordant functions are not new! [NN94]
  - Known results only for smooth case.

Differentiable $f(\mathbf{x})$

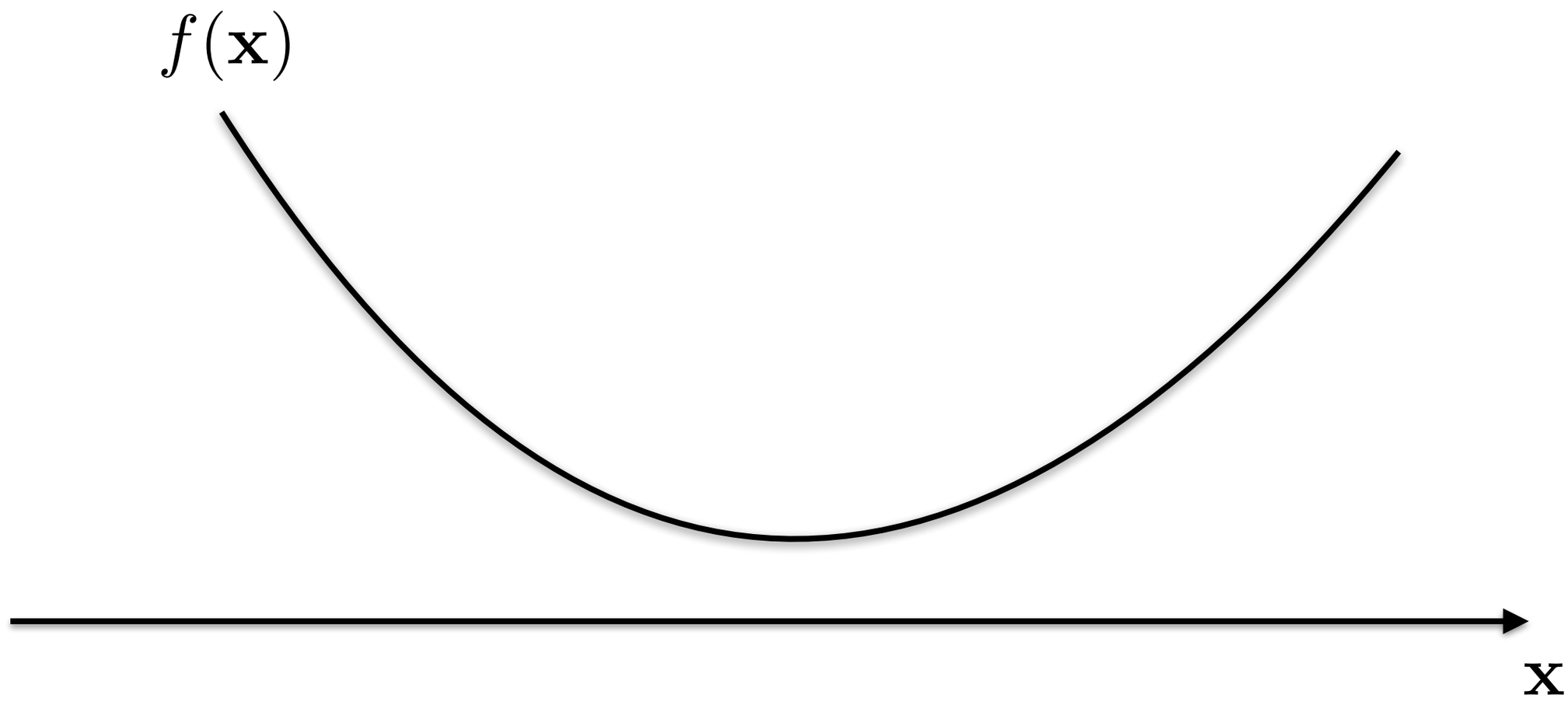


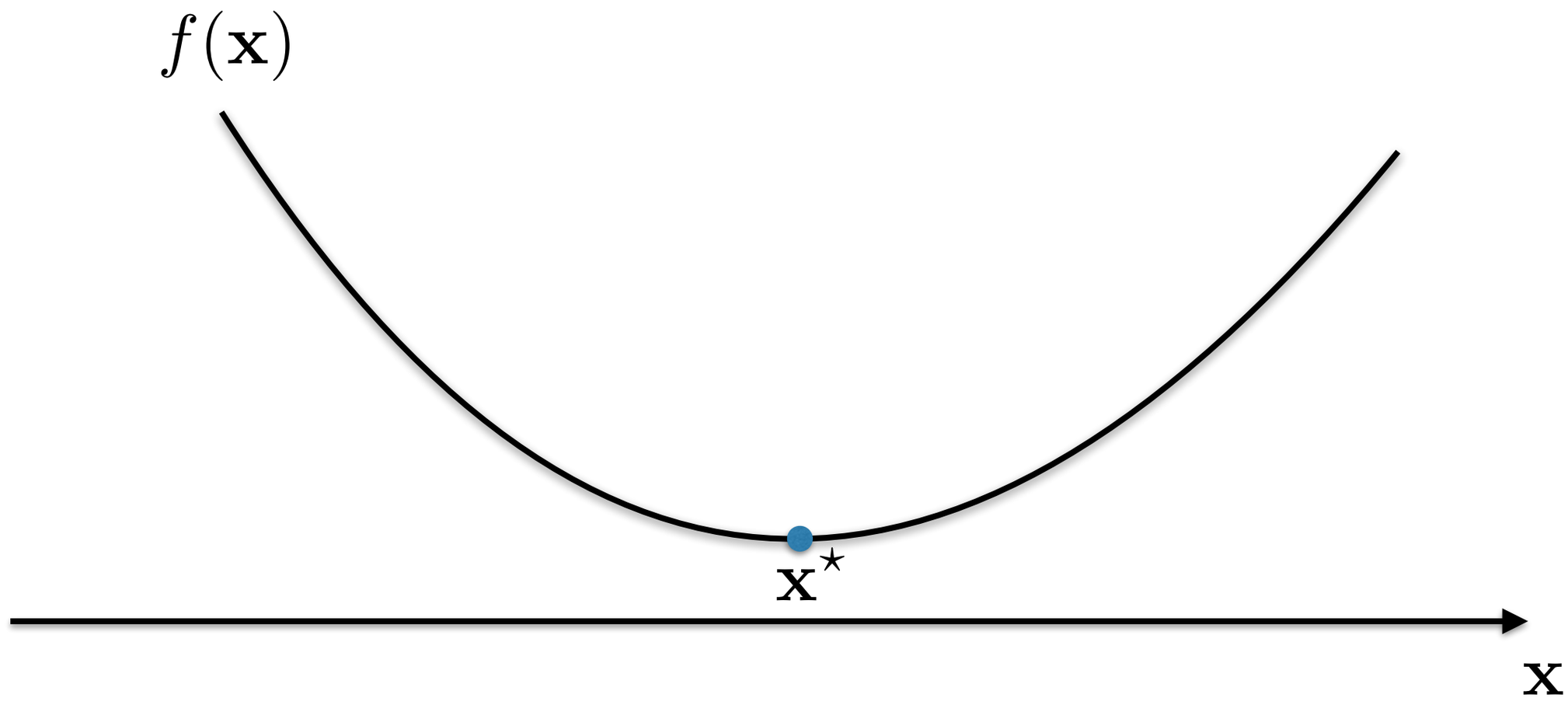
- L-Lipschitz gradient functions: $\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\| \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|$
- μ - strongly convex functions: $\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \quad \mu > 0$

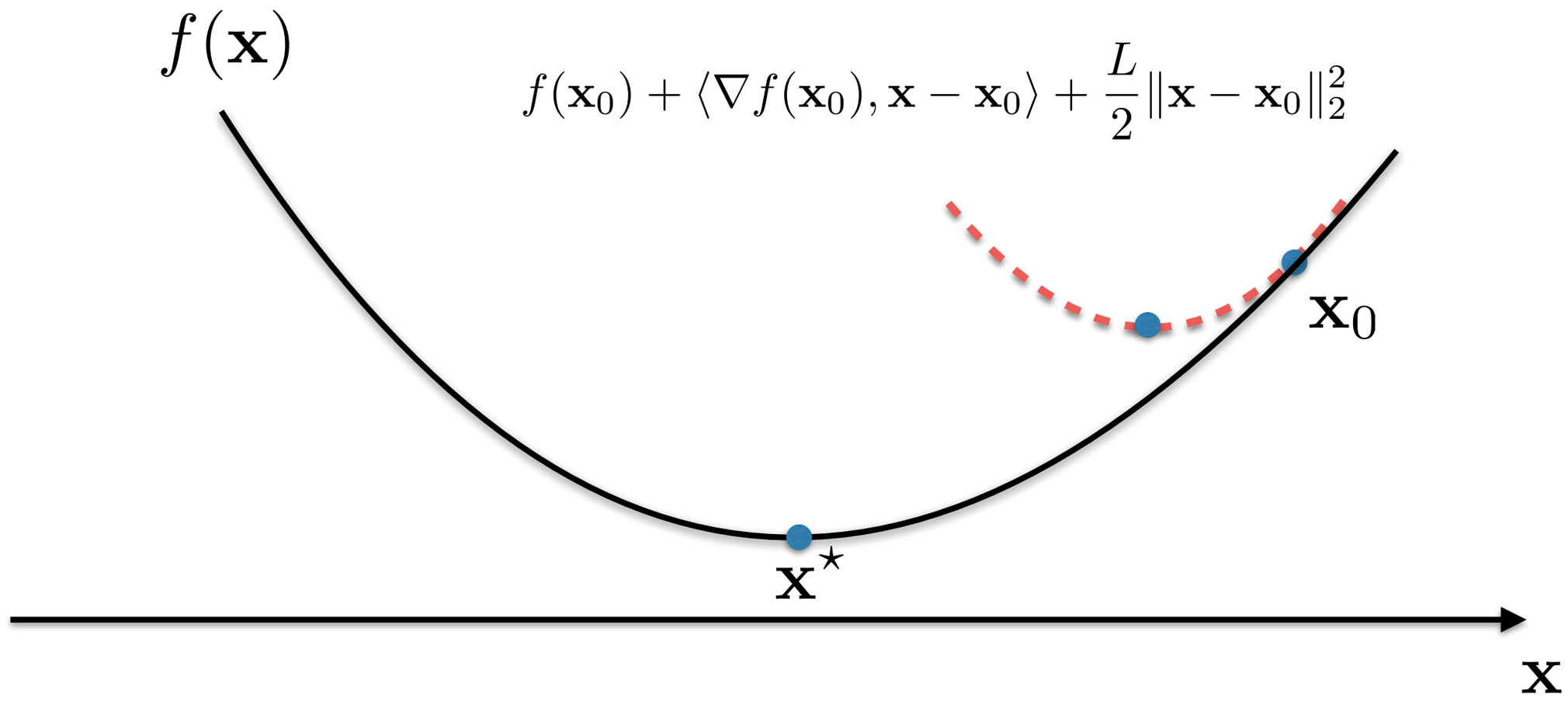
Differentiable $f(\mathbf{x})$

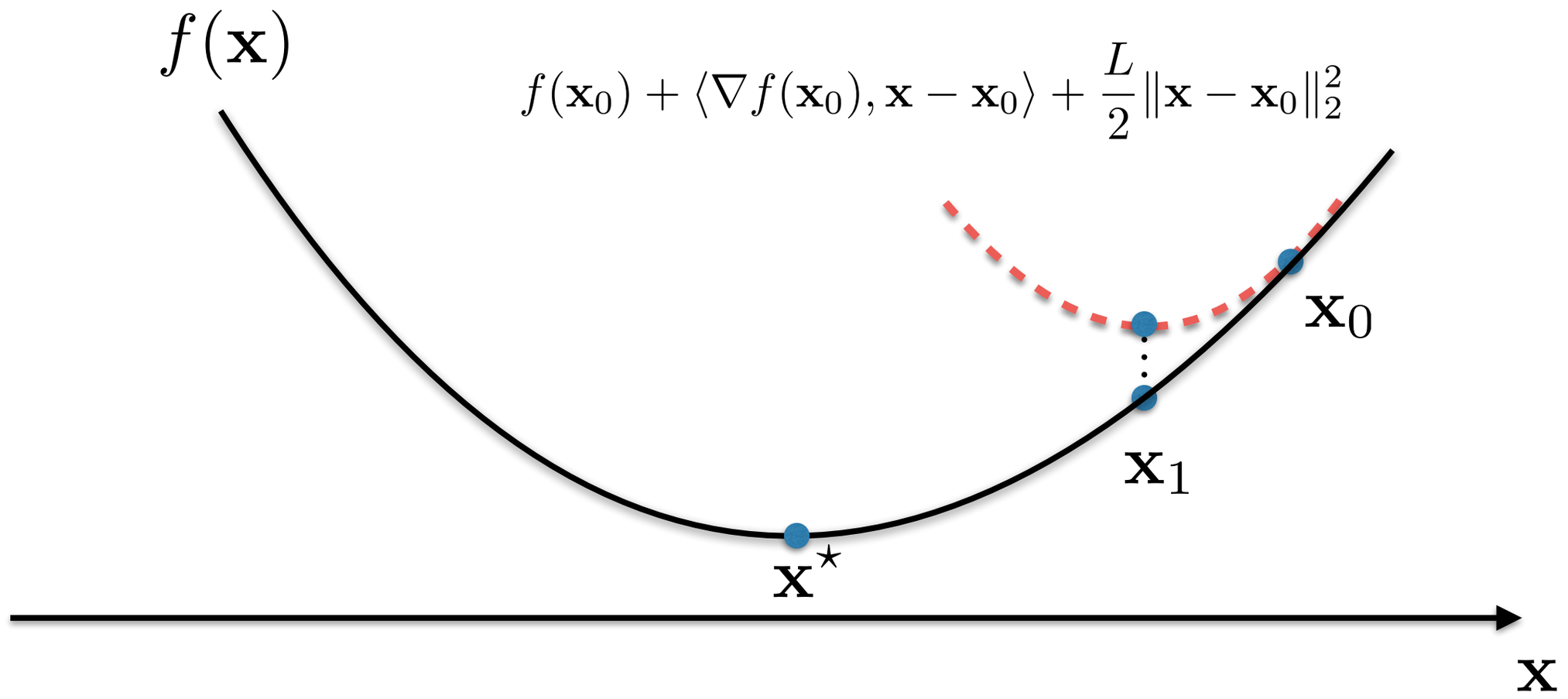


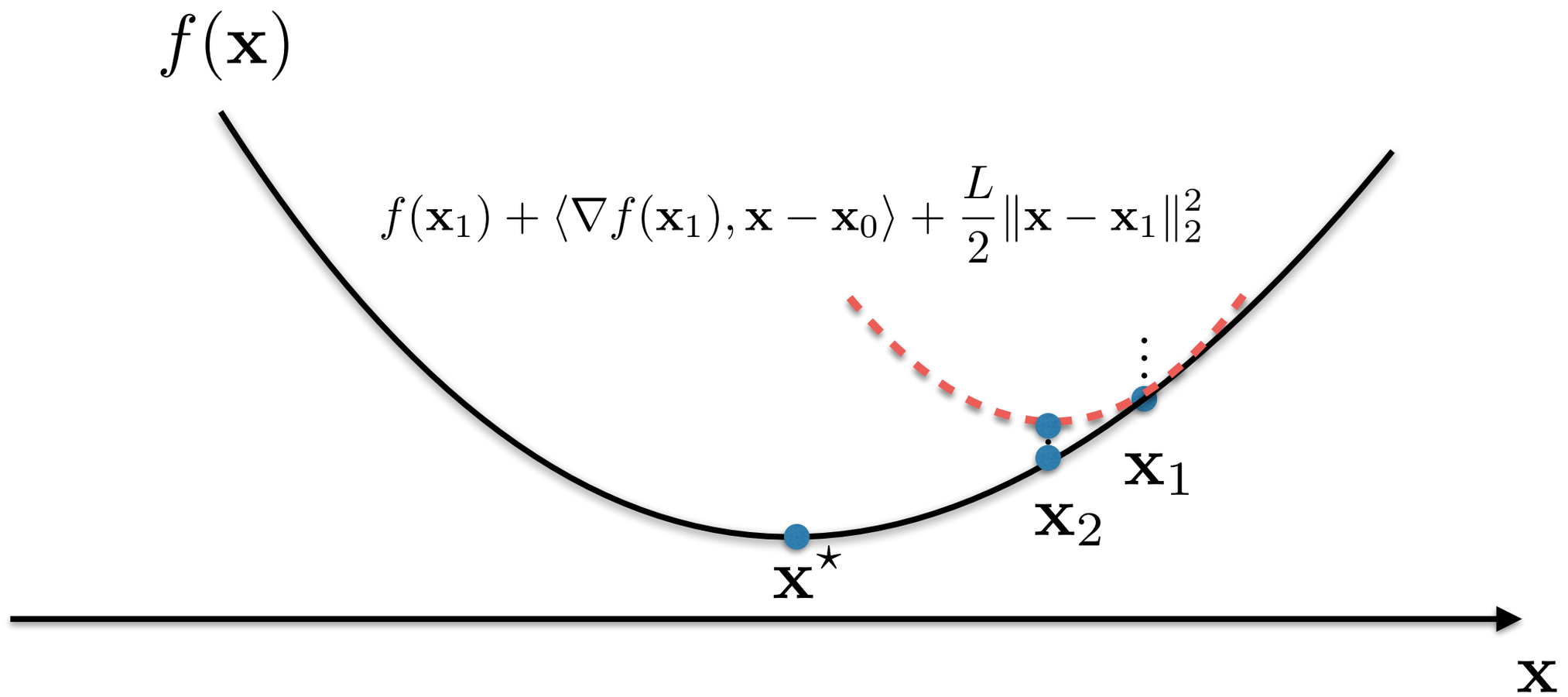
- L-Lipschitz gradient functions: $\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\| \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|$
- μ - strongly convex functions: $\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \quad \mu > 0$
- Such assumptions lead to convergence guarantees, step size selection...


$f(\mathbf{x})$
x


$f(\mathbf{x})$
$\mathbf{x}^\star$
$\mathbf{x}$

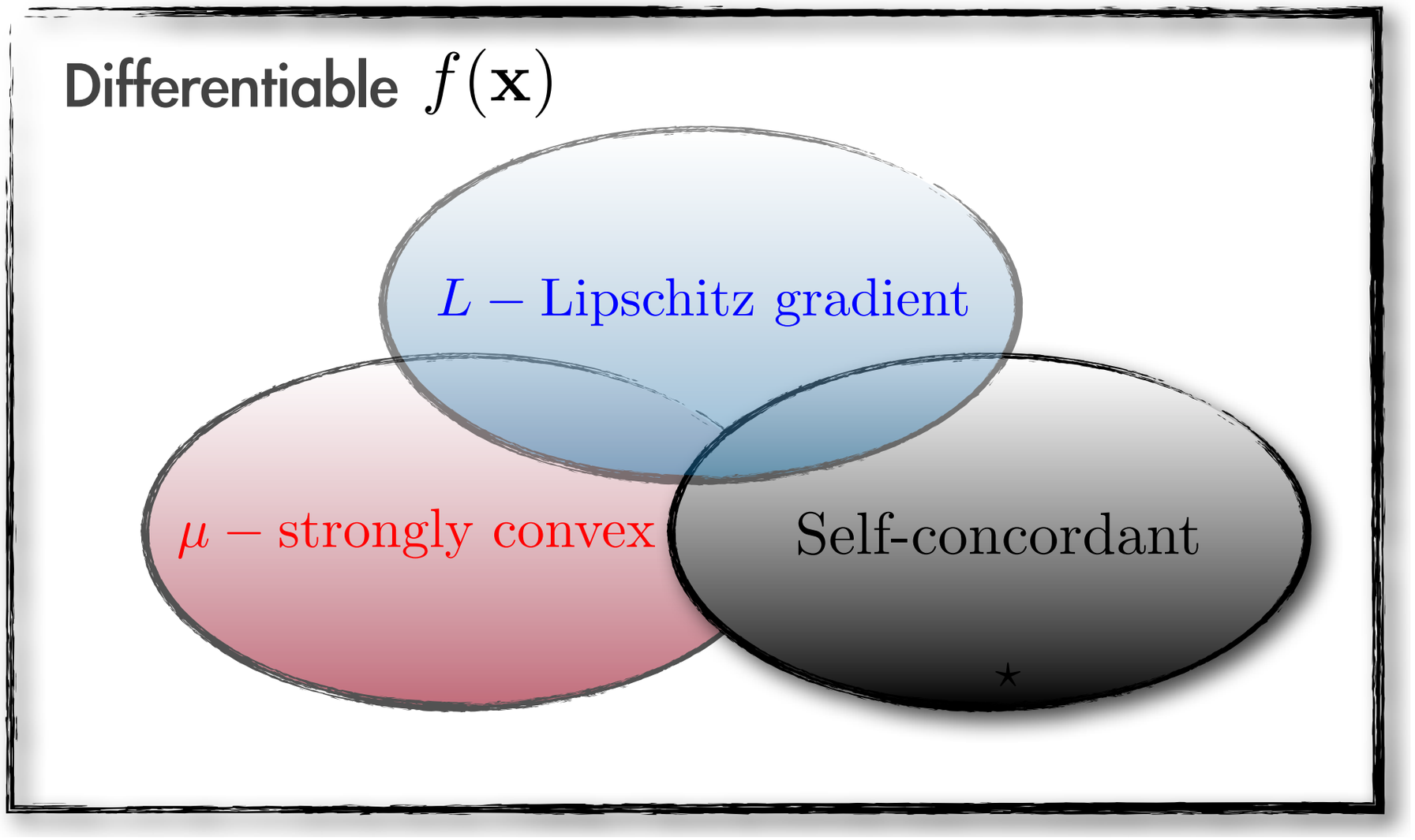

$f(\mathbf{x})$
$f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$
$\mathbf{x}_0$
$\mathbf{x}^\star$
$\mathbf{x}$


$f(\mathbf{x})$
$f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$
$\mathbf{x}_0$
$\mathbf{x}_1$
$\mathbf{x}^\star$
$\mathbf{x}$

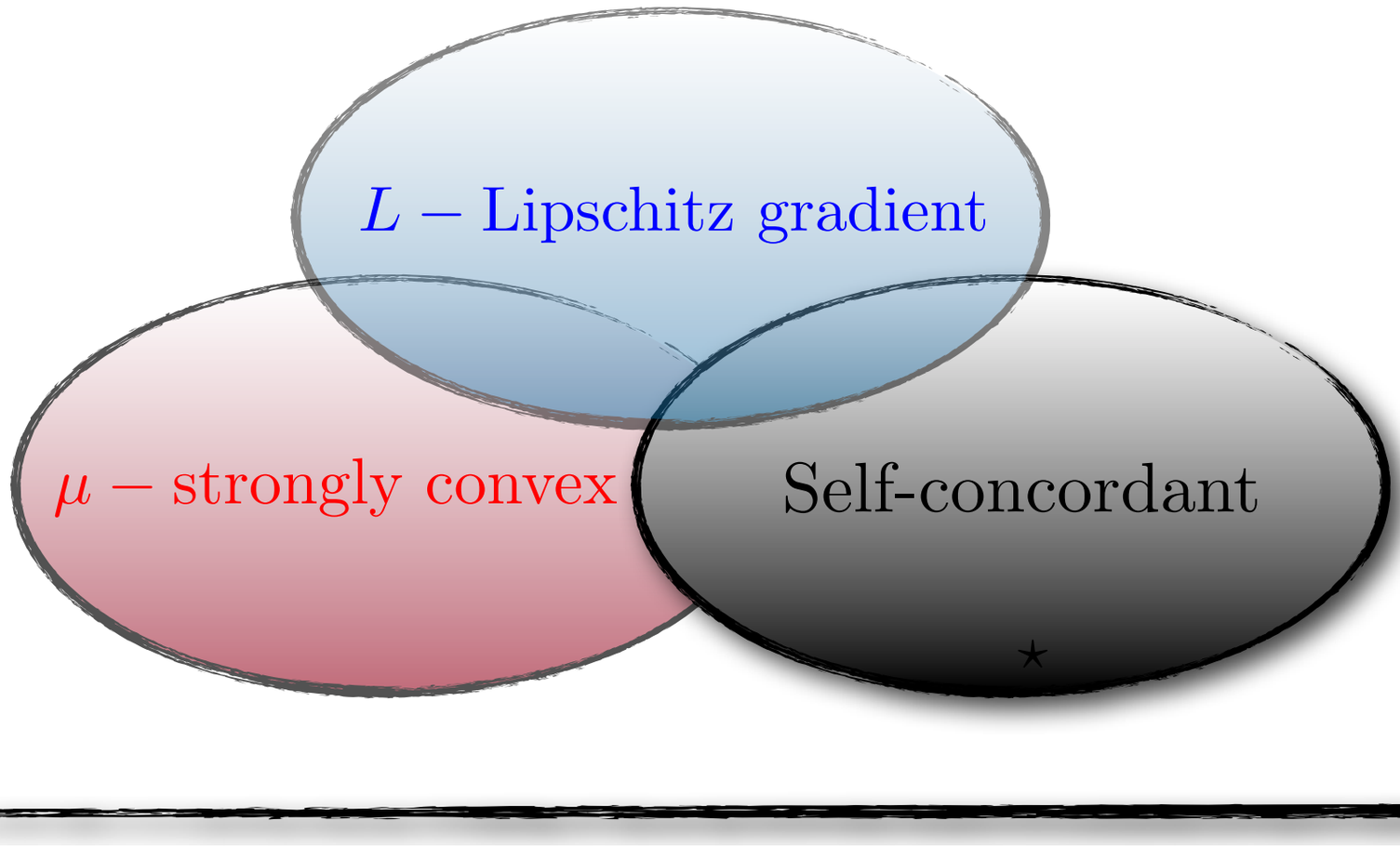

$f(\mathbf{x})$
$f(\mathbf{x}_1) + \langle \nabla f(\mathbf{x}_1), \mathbf{x} - \mathbf{x}_0 \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_1\|_2^2$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}^\star$
$\mathbf{x}$

In this talk

Class of self-concordant functions


Differentiable $f(\mathbf{x})$
$L-$ Lipschitz gradient
$\mu-$ strongly convex
Self-concordant
$\star$

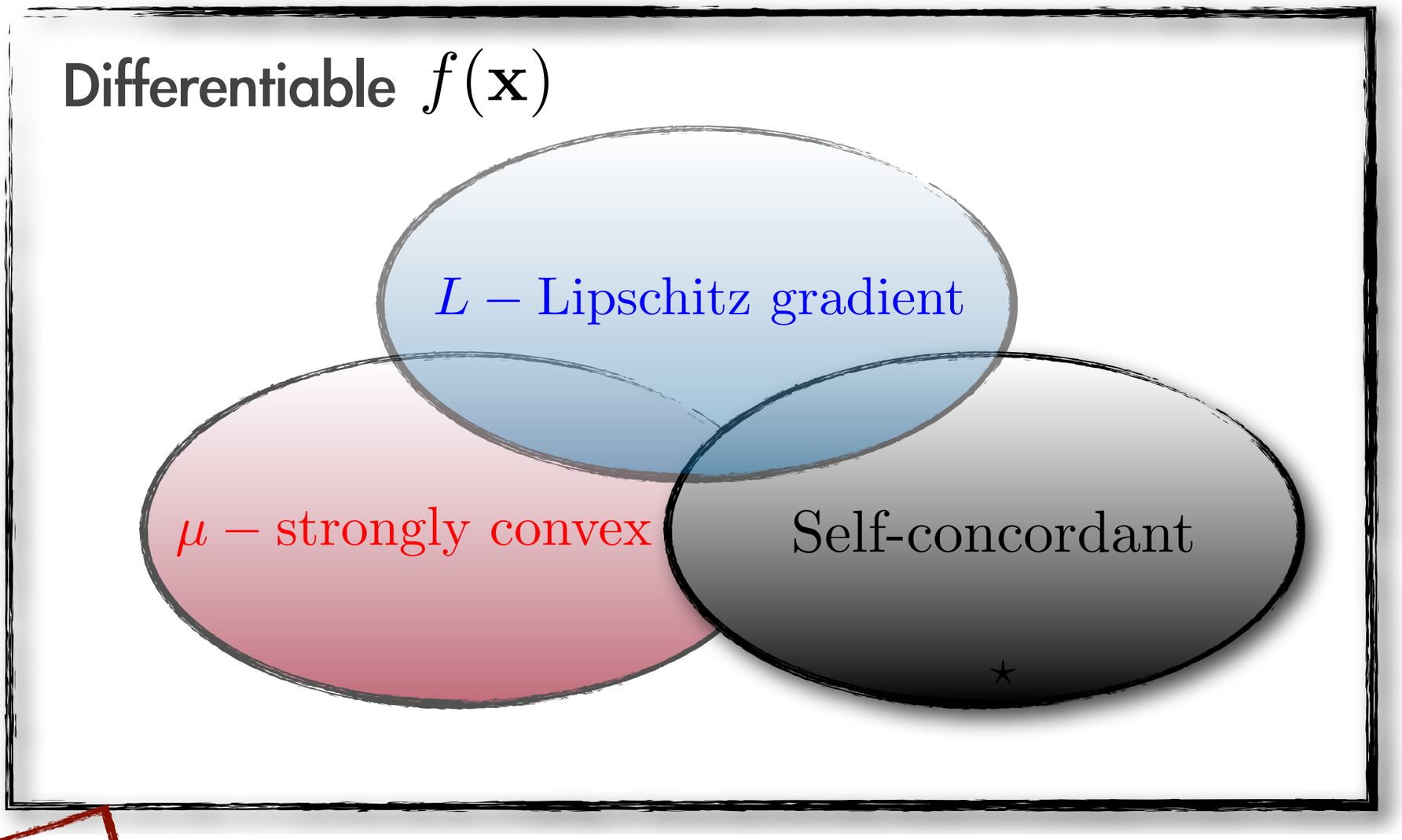
Differentiable $f(\mathbf{x})$



$f(\mathbf{x})$ is self-concordant* [NN94] $\iff$ $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$

$$\varphi(t) := f(\mathbf{x} + t\mathbf{d}), \quad \mathbf{x} \in \mathrm{dom}(f), \mathbf{d} \in \mathbb{R}^n$$

* Related to the affine invariant convergence guarantee of Newton method on self-concordant functions.

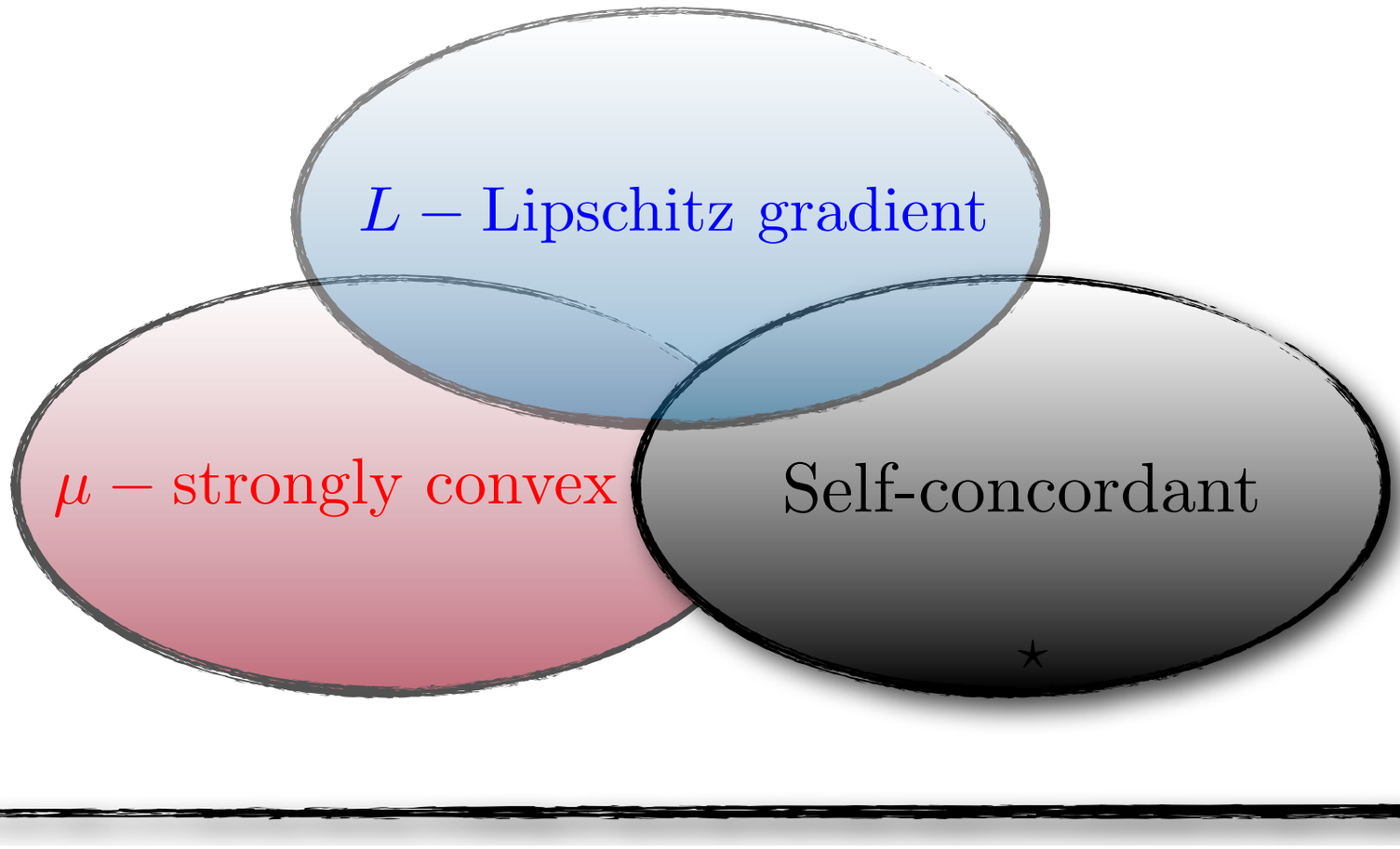


Intuition

$$\|\nabla^2 f(\mathbf{x}_1) - \nabla^2 f(\mathbf{x}_2)\| \leq M\|\mathbf{x}_1 - \mathbf{x}_2\|$$

E.g., Newton $\|\mathbf{x}_{i+1} - \mathbf{x}^\star\| \leq \dfrac{M\|\mathbf{x}_i - \mathbf{x}^\star\|^2}{2(\mu - M\|\mathbf{x}_i - \mathbf{x}^\star\|)}$

Differentiable $f(\mathbf{x})$



Intuition

$$\|\nabla^2 f(\mathbf{x}_1) - \nabla^2 f(\mathbf{x}_2)\| \leq M\|\mathbf{x}_1 - \mathbf{x}_2\|$$

E.g., Newton $\|\mathbf{x}_{i+1} - \mathbf{x}^\star\| \leq \frac{M\|\mathbf{x}_i - \mathbf{x}^\star\|^2}{2(\mu - M\|\mathbf{x}_i - \mathbf{x}^\star\|)}$

- Constants change with a new basis $f(\mathbf{Ax}$

Self-concordance does not lead to global conditions!

Self-concordance does not lead to global conditions!

- **Main properties of Lipschitz gradient + strongly convex functions**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Hessian surrogates	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

Self-concordance does not lead to global conditions!

- **Main properties of Lipschitz gradient + strongly convex functions**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Hessian surrogates	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

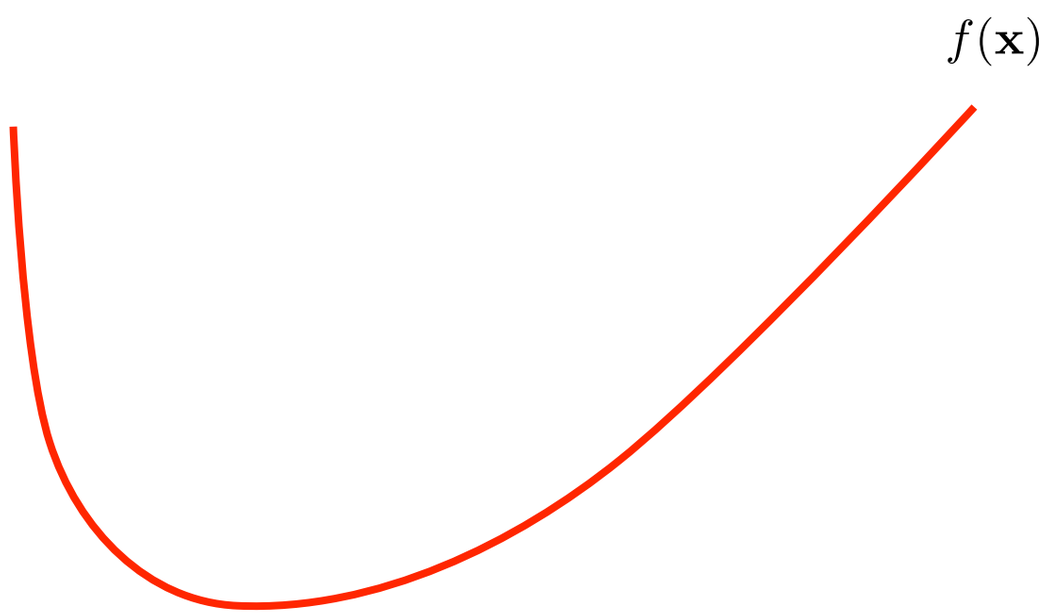
Global

Self-concordance does not lead to global conditions!

- **Main properties of Lipschitz gradient + strongly convex functions**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Hessian surrogates	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

Global



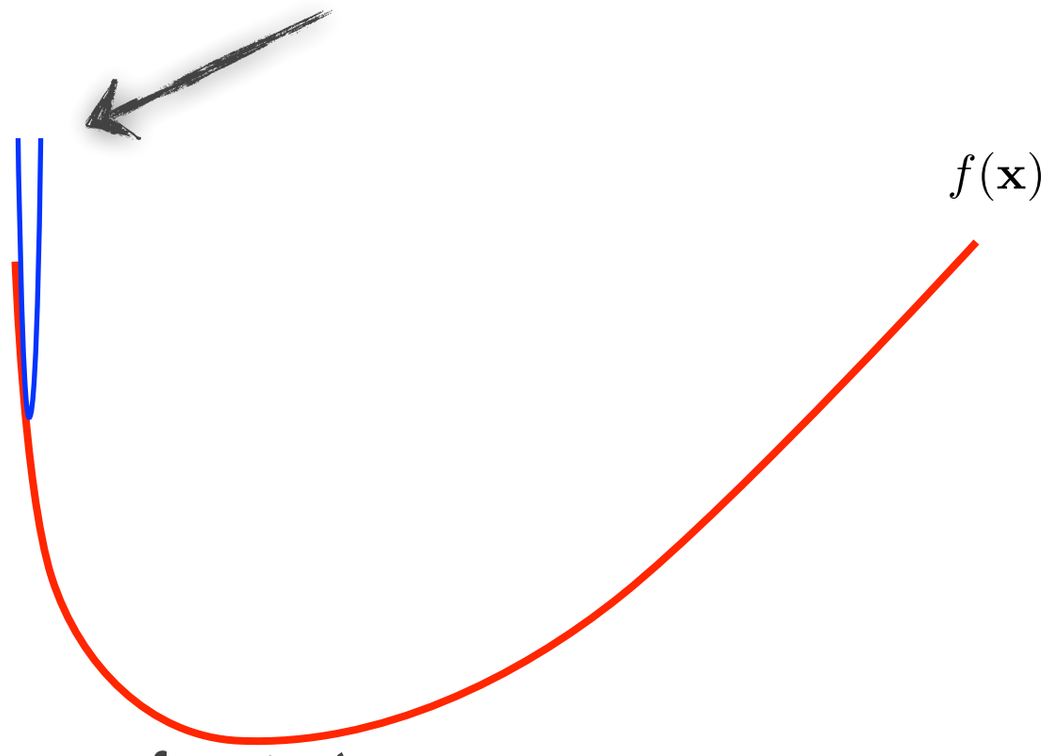
Self-concordance does not lead to global conditions!

- **Main properties of Lipschitz gradient + strongly convex functions**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Hessian surrogates	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

Global

$$\mathbf{x}^{k+1} := \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x}-\mathbf{x}^k) + \frac{L}{2}\|\mathbf{x}-\mathbf{x}^k\|_2^2 + g(\mathbf{x}) \right\}$$



L is a global constant
(defines ``width`` of surrogate function)

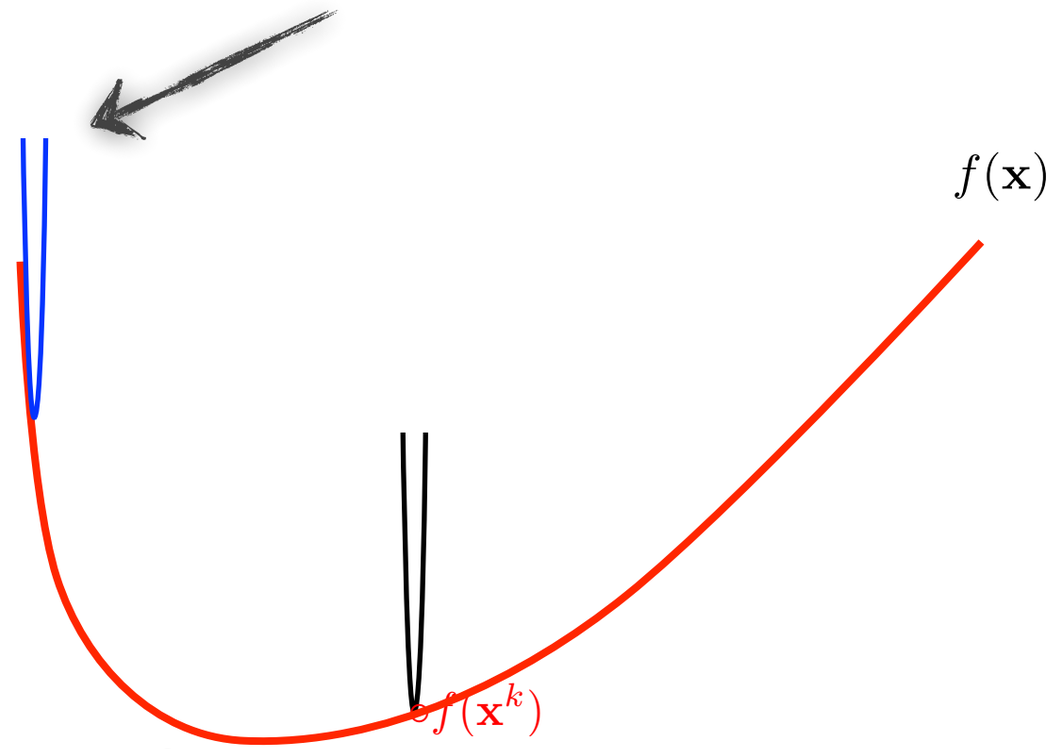
Self-concordance does not lead to global conditions!

- **Main properties of Lipschitz gradient + strongly convex functions**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Hessian surrogates	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

Global

$$\mathbf{x}^{k+1} := \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x}-\mathbf{x}^k) + \frac{L}{2}\|\mathbf{x}-\mathbf{x}^k\|_2^2 + g(\mathbf{x}) \right\}$$



L is a global constant
(defines ``width`` of surrogate function)

Self-concordance does not lead to global conditions!

- **Main properties of Lipschitz gradient + strongly convex functions**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Hessian surrogates	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

Global

- **Main properties of self-concordant functions**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega_*\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$
Hessian surrogates	$(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2\nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq (1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^{-2}\nabla^2 f(\mathbf{x})$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$

Local norm: $\|\mathbf{u}\|_{\mathbf{x}} := \left[\mathbf{u}^T\nabla^2 f(\mathbf{x})\mathbf{u}\right]^{1/2}$

Self-concordance does not lead to global conditions!

- **Main properties of Lipschitz gradient + strongly convex functions**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Hessian surrogates	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

Global

- **Main properties of self-concordant functions**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega_*\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$
Hessian surrogates	$(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2\nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq (1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^{-2}\nabla^2 f(\mathbf{x})$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$

Local

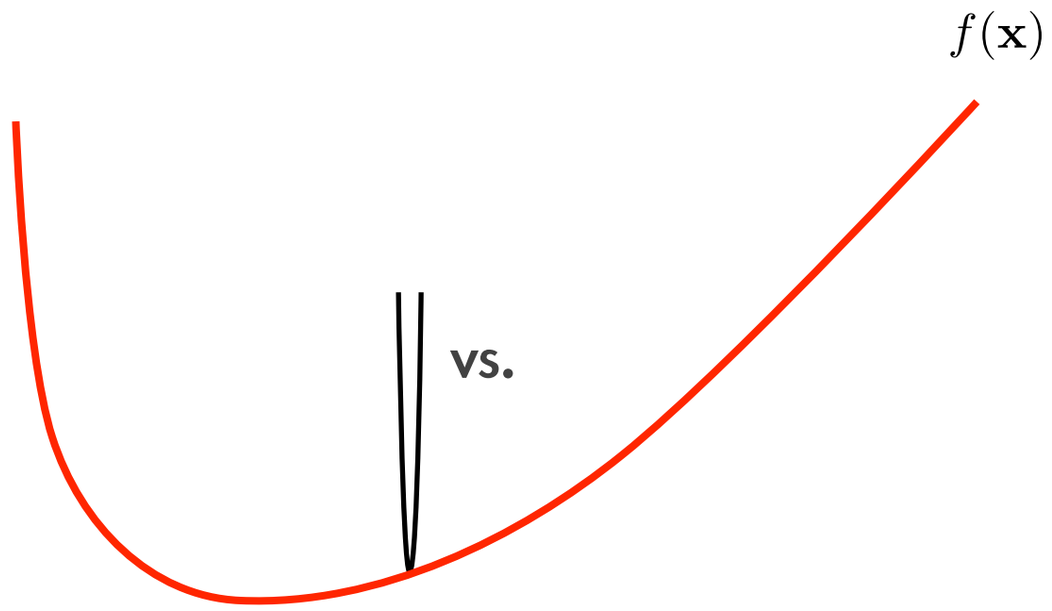
Local norm: $\|\mathbf{u}\|_{\mathbf{x}} := \left[\mathbf{u}^T\nabla^2 f(\mathbf{x})\mathbf{u}\right]^{1/2}$

Self-concordance does not lead to global conditions!

- **Main properties of self-concordant functions**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega_*\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$
Hessian surrogates	$(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2\nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq (1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^{-2}\nabla^2 f(\mathbf{x})$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$

Local

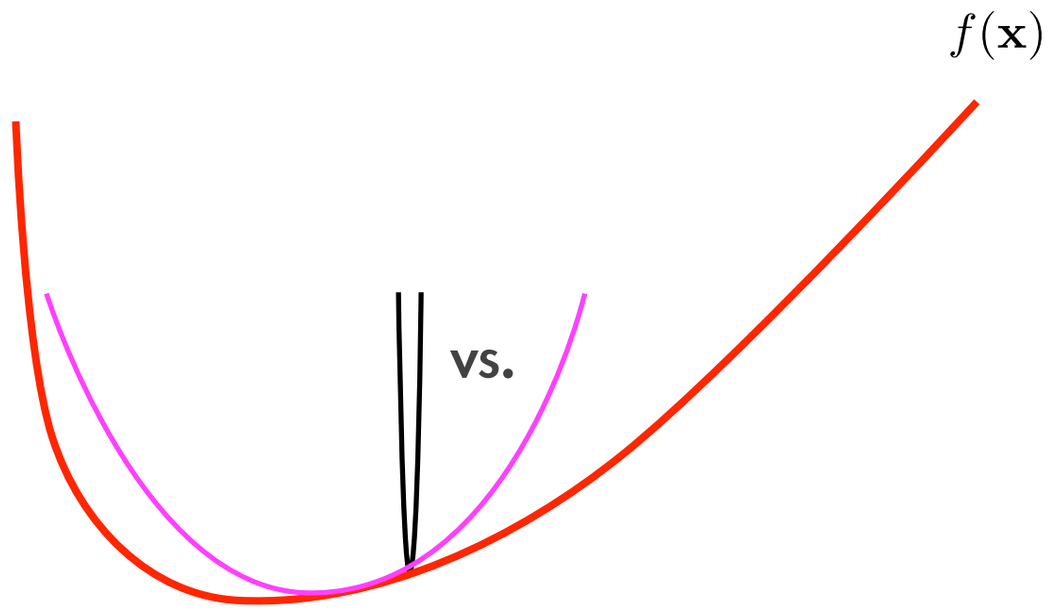


Self-concordance does not lead to global conditions!

- **Main properties of self-concordant functions**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega_*\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$
Hessian surrogates	$(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2 \nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq (1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^{-2} \nabla^2 f(\mathbf{x})$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$

Local

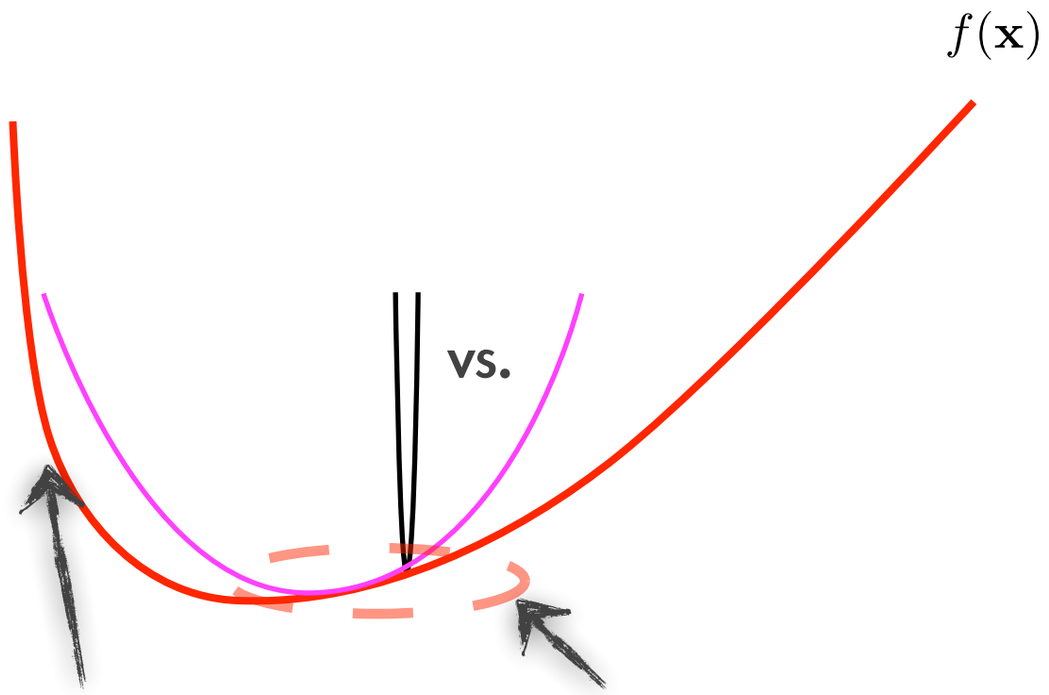


Self-concordance does not lead to global conditions!

- **Main properties of self-concordant functions**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega_*\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$
Hessian surrogates	$(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2\nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq (1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^{-2}\nabla^2 f(\mathbf{x})$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$

Local



Holds locally but not globally!

Examples

Examples:

1. $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ is L-Lipschitz gradient

Examples:

1. $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ is L-Lipschitz gradient

independent of $\mathbf{x}_1, \mathbf{x}_2$

$$\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_2 \leq \|\mathbf{A}^\top \mathbf{A}\|_{\text{sp}} \|\mathbf{x}_1 - \mathbf{x}_2\|_2$$

Examples:

1. $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ is L-Lipschitz gradient

independent of $\mathbf{x}_1, \mathbf{x}_2$

$$\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_2 \leq \|\mathbf{A}^\top \mathbf{A}\|_{\text{sp}} \|\mathbf{x}_1 - \mathbf{x}_2\|_2$$

But it is also self-concordant (any linear, quadratic function)

Examples:

1. $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ is L-Lipschitz gradient

independent of $\mathbf{x}_1, \mathbf{x}_2$

$$\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_2 \leq \|\mathbf{A}^\top \mathbf{A}\|_{\text{sp}} \|\mathbf{x}_1 - \mathbf{x}_2\|_2$$

But it is also self-concordant (any linear, quadratic function)

2. $f_{\mathbb{S}^n_{++}}(\mathbf{X}) = -\log\det(\mathbf{X})$ is L-Lipschitz gradient, but not globally!

Examples:

1. $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ is L-Lipschitz gradient

independent of $\mathbf{x}_1, \mathbf{x}_2$

$$\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_2 \leq \|\mathbf{A}^\top \mathbf{A}\|_{\text{sp}} \|\mathbf{x}_1 - \mathbf{x}_2\|_2$$

But it is also self-concordant (any linear, quadratic function)

2. $f_{\mathbb{S}^n_{++}}(\mathbf{X}) = -\log\det(\mathbf{X})$ is L-Lipschitz gradient, but not globally!

$$\|\nabla f(\mathbf{X}_1) - \nabla f(\mathbf{X}_2)\|_2 \leq \frac{1}{\alpha^2}\|\mathbf{X}_1 - \mathbf{X}_2\|_2$$

where $\alpha = \min\{\lambda_{\min}(\mathbf{X}_1),\ \lambda_{\min}(\mathbf{X}_2)\}$ depends on $\mathbf{X}_1,\ \mathbf{X}_2$

Examples:

1. $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ is L-Lipschitz gradient

independent of $\mathbf{x}_1, \mathbf{x}_2$

$$\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_2 \leq \|\mathbf{A}^\top \mathbf{A}\|_{\text{sp}} \|\mathbf{x}_1 - \mathbf{x}_2\|_2$$

But it is also self-concordant (any linear, quadratic function)

2. $f_{\mathbb{S}^n_{++}}(\mathbf{X}) = -\log\det(\mathbf{X})$ is L-Lipschitz gradient, but not globally!

$$\|\nabla f(\mathbf{X}_1) - \nabla f(\mathbf{X}_2)\|_2 \leq \frac{1}{\alpha^2}\|\mathbf{X}_1 - \mathbf{X}_2\|_2$$

where $\alpha = \min\{\lambda_{\min}(\mathbf{X}_1),\ \lambda_{\min}(\mathbf{X}_2)\}$ depends on $\mathbf{X}_1,\ \mathbf{X}_2$

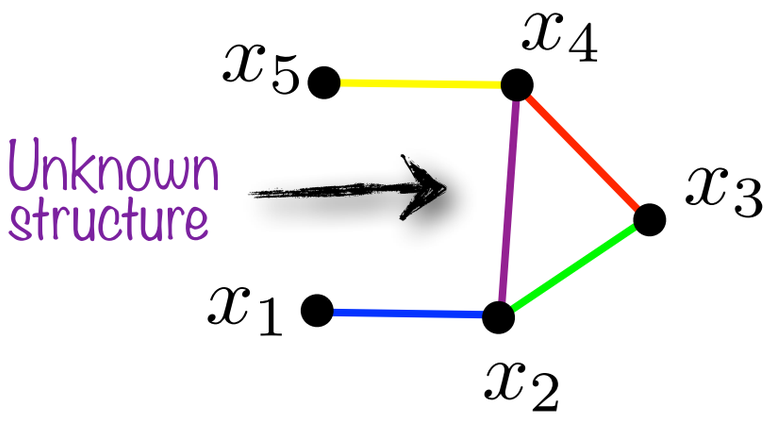
But it is self-concordant!

...but are all these useful?

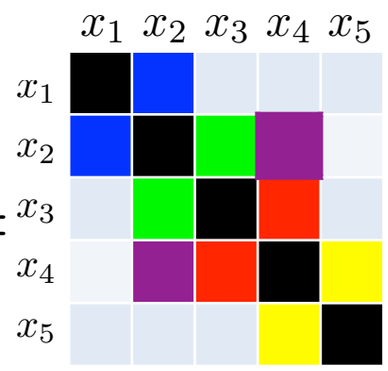
Application #1: Graphical model selection

[Dem72, BGA08, HSDR11]

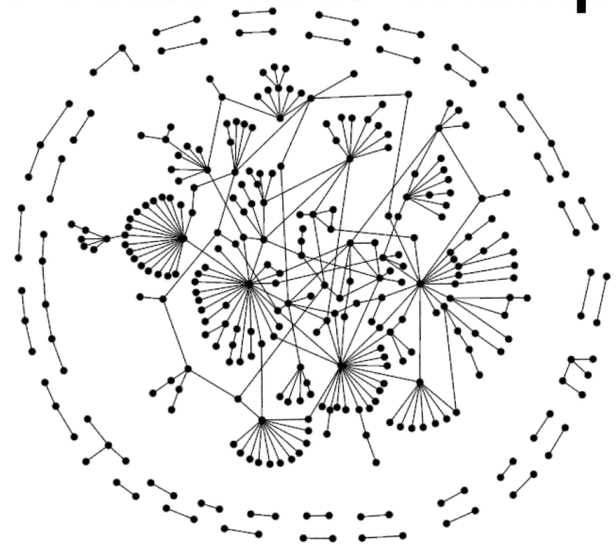
- Toy example



$$\Sigma^{-1} =$$



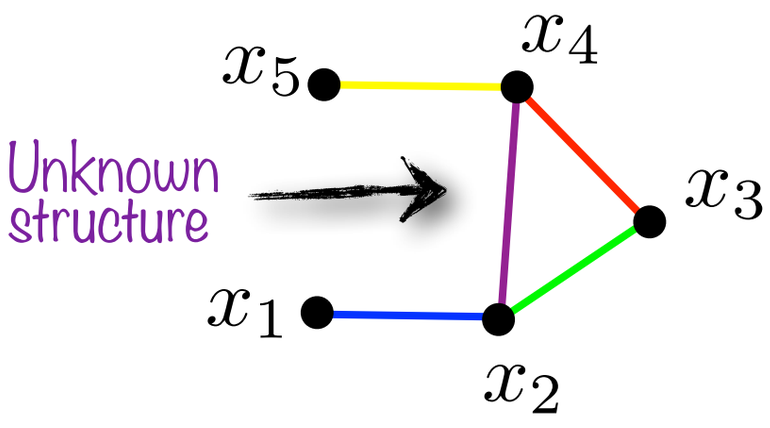
Large-scale real example



Application #1: Graphical model selection

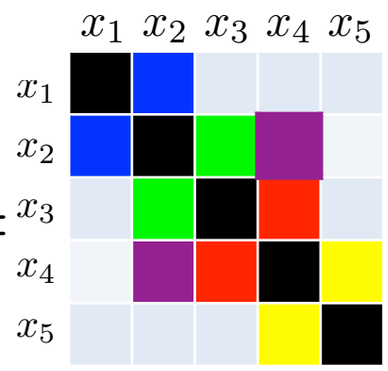
[Dem72, BGA08, HSDR11]

- Toy example

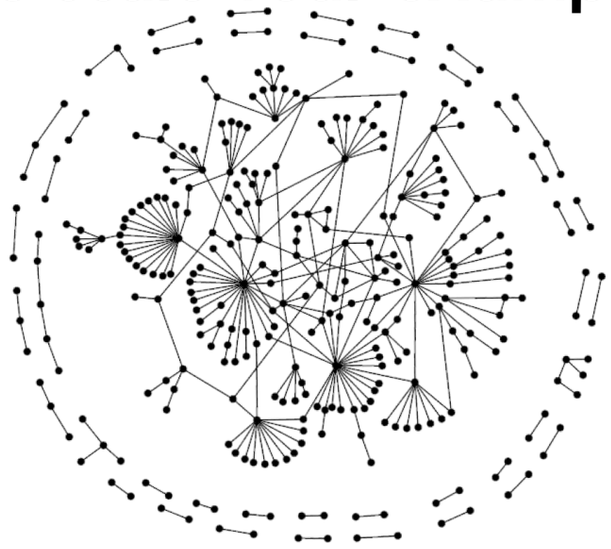


Target

$$\Sigma^{-1} =$$



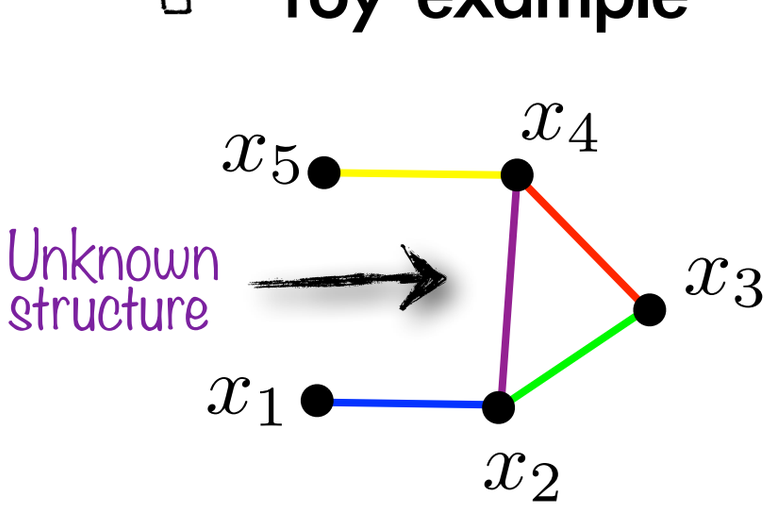
Large-scale real example



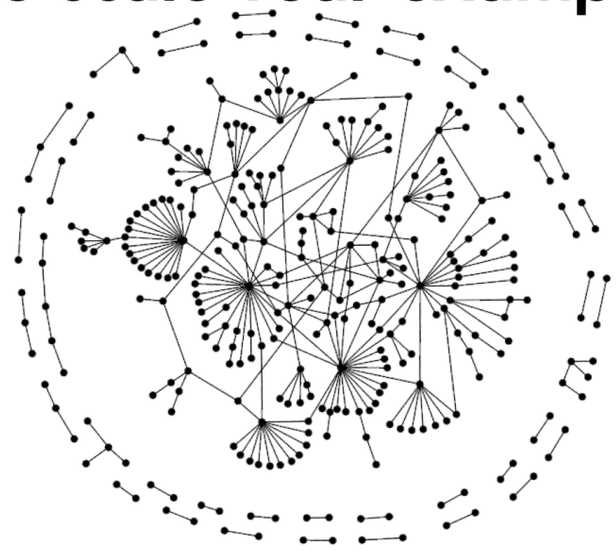
Application #1: Graphical model selection

[Dem72, BGA08, HSDR11]

- Toy example



Large-scale real example



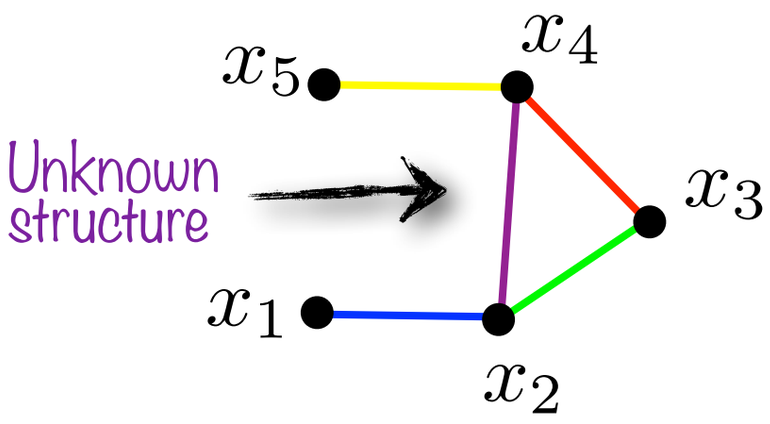
- Given a data set $\mathcal{D}$, drawn from a joint pdf with unknown covariance $\mathbf{\Sigma}$, the aim is to learn a sparse matrix $\mathbf{\Theta}$ that approximates $\mathbf{\Sigma}^{-1}$.

  Input: sample covariance $\widehat{\mathbf{\Sigma}}$ calculated usually from limited samples

Application #1: Graphical model selection

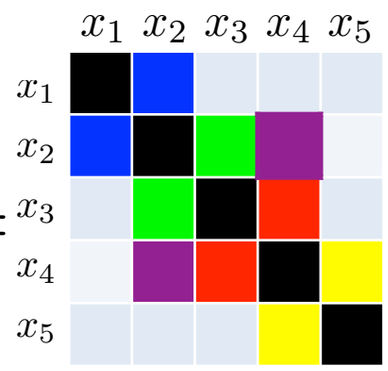
[Dem72, BGA08, HSDR11]

- Toy example

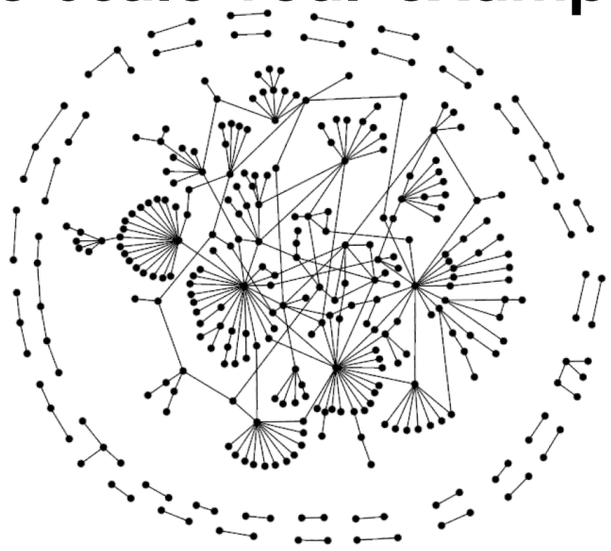


Target

$\Sigma^{-1} =$



Large-scale real example



- Given a data set $\mathcal{D}$, drawn from a joint pdf with unknown covariance $\mathbf{\Sigma}$, the aim is to learn a sparse matrix $\mathbf{\Theta}$ that approximates $\mathbf{\Sigma}^{-1}$.

Input: sample covariance $\widehat{\mathbf{\Sigma}}$ calculated usually from limited samples

Optimization problem

$$\min_{\mathbf{\Theta}} \left\{ \underbrace{-\log\det(\mathbf{\Theta}) + \mathrm{trace}(\widehat{\mathbf{\Sigma}}\mathbf{\Theta})}_{f(\cdot)} + \underbrace{\rho\|\mathbf{\Theta}\|_1}_{g(\cdot)} \right\}$$

Self-concordant function

Real-world problems include thousands of variables to optimize

Theory

"Composite self-concordant minimization", Tran-Dinh, Kyrillidis, Cevher, JMLR 2015

- Recall:

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

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Recall:

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

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$$

Direction to move

Step size

- Recall:

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

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy: $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$ ← Direction to move

Step size

- For L-Lipschitz gradient (and strongly convex) $f(\mathbf{x})$:

$$\alpha_i \propto (L, \mu)$$

- Recall:

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

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy: $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$ ← Direction to move

Step size

- For L-Lipschitz gradient (and strongly convex) $f(\mathbf{x})$:

$$\alpha_i \propto (L, \mu)$$

- For direction, proximal first-order schemes:
  - $\mathbf{d}_i = \arg\min_{\mathbf{d}} \left\{ f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{d} \rangle + \frac{L}{2}\|\mathbf{d}\|_2^2 + g(\mathbf{x}_i + \mathbf{d}) \right\}$

- Recall:

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

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy: $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$ ← Direction to move

Step size

- For L-Lipschitz gradient (and strongly convex) $f(\mathbf{x})$:

$$\alpha_i \propto (L, \mu)$$

- For direction, proximal first-order schemes:
  - $\mathbf{d}_i = \arg\min_{\mathbf{d}} \left\{ f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{d} \rangle + \frac{L}{2}\|\mathbf{d}\|_2^2 + g(\mathbf{x}_i + \mathbf{d}) \right\}$
- ...and proximal Newton schemes:
  - $\mathbf{d}_i = \arg\min_{\mathbf{d}} \left\{ f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{d} \rangle + \frac{1}{2}\langle \nabla^2 f(\mathbf{x}_i)\mathbf{d}, \mathbf{d} \rangle + g(\mathbf{x}_i + \mathbf{d}) \right\}$

- Recall:

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

In this talk

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy: $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$ ← Direction to move

Step size

- For ~~L-Lipschitz gradient (and strongly convex)~~ $f(\mathbf{x})$:

~~$\alpha_i \propto (L, \mu)$~~

- For direction, proximal first-order schemes:
  - $\mathbf{d}_i = \arg\min_{\mathbf{d}} \left\{ f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{d} \rangle + \frac{L}{2}\|\mathbf{d}\|_2^2 + g(\mathbf{x}_i + \mathbf{d}) \right\}$
- ...and proximal Newton schemes:
  - $\mathbf{d}_i = \arg\min_{\mathbf{d}} \left\{ f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{d} \rangle + \frac{1}{2}\langle \nabla^2 f(\mathbf{x}_i)\mathbf{d}, \mathbf{d} \rangle + g(\mathbf{x}_i + \mathbf{d}) \right\}$

- Recall:

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

In this talk

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy: $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$ ← Direction to move

Step size

- For ~~L-Lipschitz gradient (and strongly convex)~~ $f(\mathbf{x})$:

$\alpha_i \propto (L, \mu)$ $\qquad$ α_i ?

- For direction, proximal first-order schemes:
  - $\mathbf{d}_i = \arg\min_{\mathbf{d}} \left\{ f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{d} \rangle + \frac{L}{2}\|\mathbf{d}\|_2^2 + g(\mathbf{x}_i + \mathbf{d}) \right\}$
- ...and proximal Newton schemes:
  - $\mathbf{d}_i = \arg\min_{\mathbf{d}} \left\{ f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{d} \rangle + \frac{1}{2}\langle \nabla^2 f(\mathbf{x}_i)\mathbf{d}, \mathbf{d} \rangle + g(\mathbf{x}_i + \mathbf{d}) \right\}$

- Recall:

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

In this talk

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$$

$\mathbf{d}_i$ ← Direction to move

α_i ← Step size

Step size

- Recall:

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

In this talk

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy: $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$ ← Direction to move

Step size

Step size

- Define Newton decrement: $\lambda_i = \sqrt{\mathbf{d}_i^T \nabla^2 f(\mathbf{x}_i) \mathbf{d}_i}$
- Step size selection: $\alpha_i = \dfrac{1}{\lambda_i + 1} \in (0, 1]$

- Recall:

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

In this talk

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy: $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$ ← Direction to move

Step size

Step size

- Define Newton decrement: $\lambda_i = \sqrt{\mathbf{d}_i^T \nabla^2 f(\mathbf{x}_i) \mathbf{d}_i}$
- Step size selection: $\alpha_i = \dfrac{1}{\lambda_i + 1} \in (0, 1]$

Requires second order information

- Recall:

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

In this talk

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$$

$\mathbf{d}_i$ ← Direction to move

α_i ← Step size

- Recall:

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

In this talk

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$$

$\mathbf{d}_i$ ← Direction to move

α_i ← Step size

Algorithm (Newton case)

Damped Newton phase $(\alpha_i < 1)$

Newton phase $(\alpha_i = 1)$

- Recall:

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

In this talk

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$$

$\mathbf{d}_i$: Direction to move

α_i: Step size

Algorithm (Newton case)

Damped Newton phase $(\alpha_i < 1)$

- Compute $\mathbf{d}_i$

Newton phase $(\alpha_i = 1)$

- Recall:

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

In this talk

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$$

$\mathbf{d}_i$ ← Direction to move

α_i ← Step size

Algorithm (Newton case)

Damped Newton phase $(\alpha_i < 1)$

- Compute $\mathbf{d}_i$
- Compute λ_i and α_i

Newton phase $(\alpha_i = 1)$

- Recall:

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

In this talk

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy: $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$ ← Direction to move

Step size

Algorithm (Newton case)

Damped Newton phase $(\alpha_i < 1)$

- Compute $\mathbf{d}_i$
- Compute λ_i and α_i
- Do $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$

Newton phase $(\alpha_i = 1)$

- Recall:

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

In this talk

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy: $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$ ← Direction to move

Step size

Algorithm (Newton case)

Damped Newton phase $(\alpha_i < 1)$

- Compute $\mathbf{d}_i$
- Compute λ_i and α_i
- Do $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$

Newton phase $(\alpha_i = 1)$

- If $\lambda_i \leq 0.219$ then, $\mathbf{x}_{i+1} = \mathbf{x}_i + \mathbf{d}_i$

- Recall:

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

In this talk

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

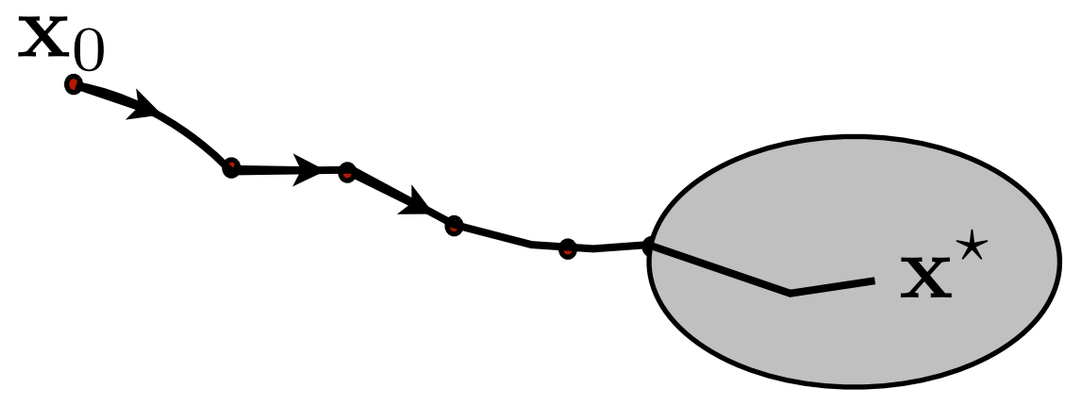
- Generic strategy: $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$ ← Direction to move

Step size

Algorithm (Newton case)

Damped Newton phase $(\alpha_i < 1)$

- Compute $\mathbf{d}_i$
- Compute λ_i and α_i
- Do $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$



Newton phase $(\alpha_i = 1)$

- If $\lambda_i \leq 0.219$ then, $\mathbf{x}_{i+1} = \mathbf{x}_i + \mathbf{d}_i$

- Recall:

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

In this talk

$$\text{prox}_{\lambda g}(\mathbf{w}) = \arg\min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{w}\|_2^2$$

- Generic strategy: $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$ ← Direction to move

Step size

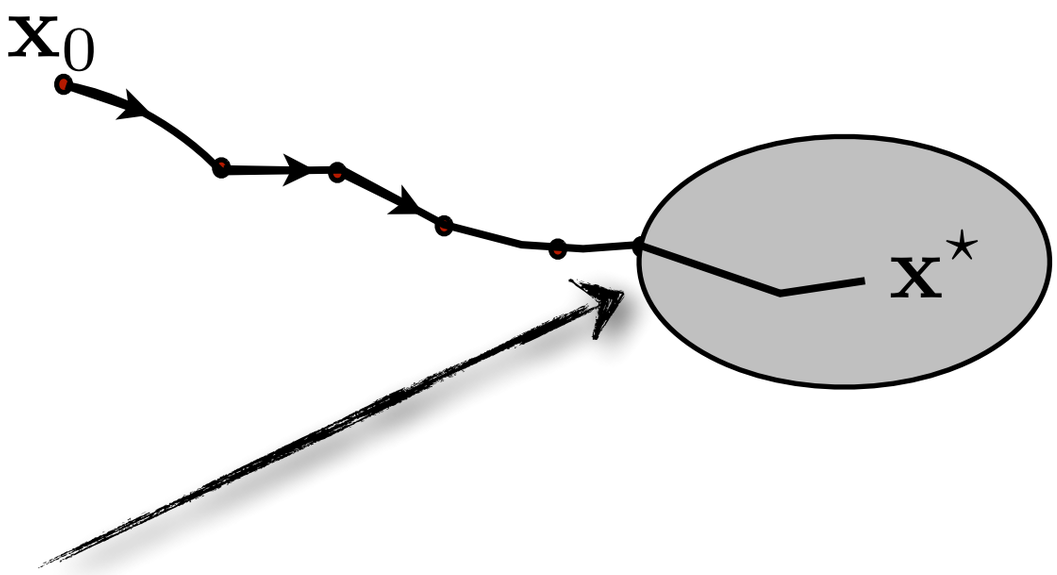
Algorithm (Newton case)

Damped Newton phase $(\alpha_i < 1)$

- Compute $\mathbf{d}_i$
- Compute λ_i and α_i
- Do $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$

Newton phase $(\alpha_i = 1)$

- If $\lambda_i \leq 0.219$ then, $\mathbf{x}_{i+1} = \mathbf{x}_i + \mathbf{d}_i$



- Theorem for Newton case:

$$\lambda_i = \sqrt{\mathbf{d}_i^T \nabla^2 f(\mathbf{x}_i) \mathbf{d}_i}$$

If $\alpha_i = \dfrac{1}{\lambda_i + 1} \in (0, 1]$, then $\mathbf{x}_{i+1}$ satisfies:

$$f(\mathbf{x}_{i+1}) + g(\mathbf{x}_{i+1}) \leq f(\mathbf{x}_i) + g(\mathbf{x}_i) - \omega(\lambda_i), \quad \text{where} \quad \omega(t) = t - \log(1 + t) \geq 0$$

Moreover, step size α_i is optimal.

- Theorem for Newton case:

$$\lambda_i = \sqrt{\mathbf{d}_i^T \nabla^2 f(\mathbf{x}_i) \mathbf{d}_i}$$

If $\alpha_i = \dfrac{1}{\lambda_i + 1} \in (0, 1]$, then $\mathbf{x}_{i+1}$ satisfies:

$$f(\mathbf{x}_{i+1}) + g(\mathbf{x}_{i+1}) \leq f(\mathbf{x}_i) + g(\mathbf{x}_i) - \omega(\lambda_i), \quad \text{where} \quad \omega(t) = t - \log(1 + t) \geq 0$$

Moreover, step size α_i is optimal.

- Proof sketch

$$\omega_*(t) = -t - \log(1 - t), \quad t \in [0, 1]$$

  - By convexity of $f(\cdot)$ and self-concordance:

$$f(\mathbf{x}_{i+1}) \leq f(\mathbf{x}_i) + \alpha_i \nabla f(\mathbf{x}_i)^T \mathbf{d}_i + \omega_* \left(\alpha_i \sqrt{\mathbf{d}_i^T \nabla^2 f(\mathbf{x}_i) \mathbf{d}_i} \right)$$

  - By convexity of $g(\cdot)$ and optimality conditions:

$$g(\mathbf{x}_{i+1}) \leq g(\mathbf{x}_i) - \alpha_i \left(\nabla f(\mathbf{x}_i)^T \mathbf{d}_i - \mathbf{d}_i^T \nabla^2 f(\mathbf{x}_i) \mathbf{d}_i \right)$$

  - Basic algebra:

$$f(\mathbf{x}_{i+1}) + g(\mathbf{x}_{i+1}) \leq f(\mathbf{x}_i) + g(\mathbf{x}_i) - (\alpha_i \lambda_i - \omega_*(\alpha_i \lambda_i))$$

- Theorem for Newton case:

$$\lambda_i = \sqrt{\mathbf{d}_i^T \nabla^2 f(\mathbf{x}_i) \mathbf{d}_i}$$

If $\alpha_i = \dfrac{1}{\lambda_i + 1} \in (0, 1]$, then $\mathbf{x}_{i+1}$ satisfies:

$$f(\mathbf{x}_{i+1}) + g(\mathbf{x}_{i+1}) \leq f(\mathbf{x}_i) + g(\mathbf{x}_i) - \omega(\lambda_i), \quad \text{where} \quad \omega(t) = t - \log(1 + t) \geq 0$$

Moreover, step size α_i is optimal.

- Guaranteed objective function decrease per iteration.
- No back-tracking line search = no function evaluations (expensive)
- Guaranteed (local) quadratic convergence rate.
- Exact analytic complexity on the number of iterations.

- Theorem for Newton case:

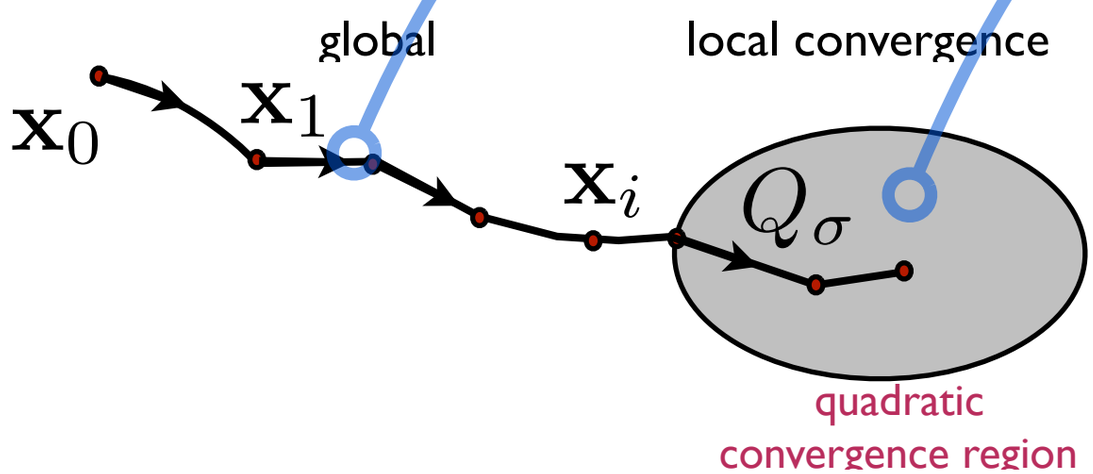
$$\lambda_i = \sqrt{\mathbf{d}_i^T \nabla^2 f(\mathbf{x}_i) \mathbf{d}_i}$$

If $\alpha_i = \dfrac{1}{\lambda_i + 1} \in (0, 1]$, then $\mathbf{x}_{i+1}$ satisfies:

$$f(\mathbf{x}_{i+1}) + g(\mathbf{x}_{i+1}) \leq f(\mathbf{x}_i) + g(\mathbf{x}_i) - \omega(\lambda_i), \quad \text{where} \quad \omega(t) = t - \log(1+t) \geq 0$$

Moreover, step size α_i is optimal.

$$\#\text{iterations} = \left\lfloor \frac{\phi(\mathbf{x}_0) - \phi(\mathbf{x}^\star)}{0.017} \right\rfloor + O\left(\log\log\left(0.3/\epsilon\right)\right)$$



$$\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})$$

$$\|\mathbf{x}_i - \mathbf{x}^\star\| \leq 2\epsilon$$

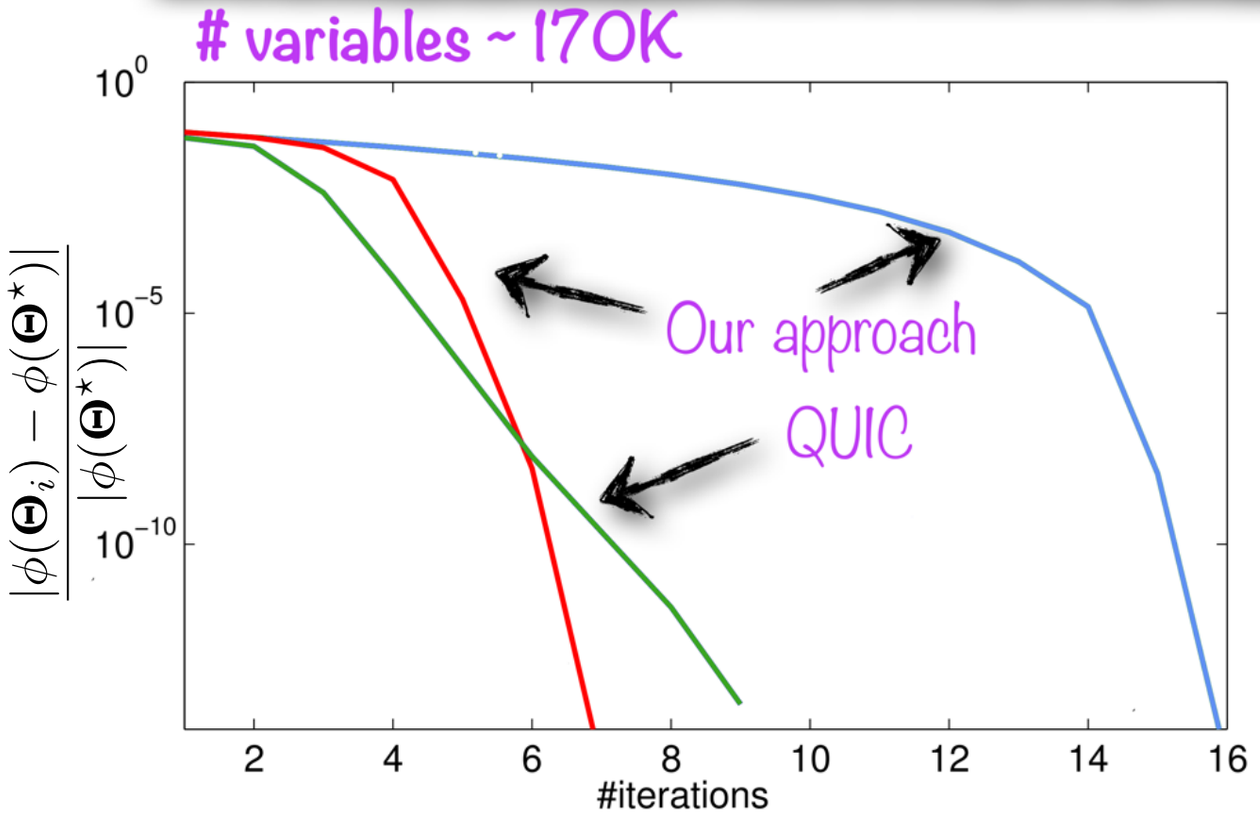
$$Q_\sigma = \{\mathbf{x}_i \mid \lambda_i \leq 0.219\}$$

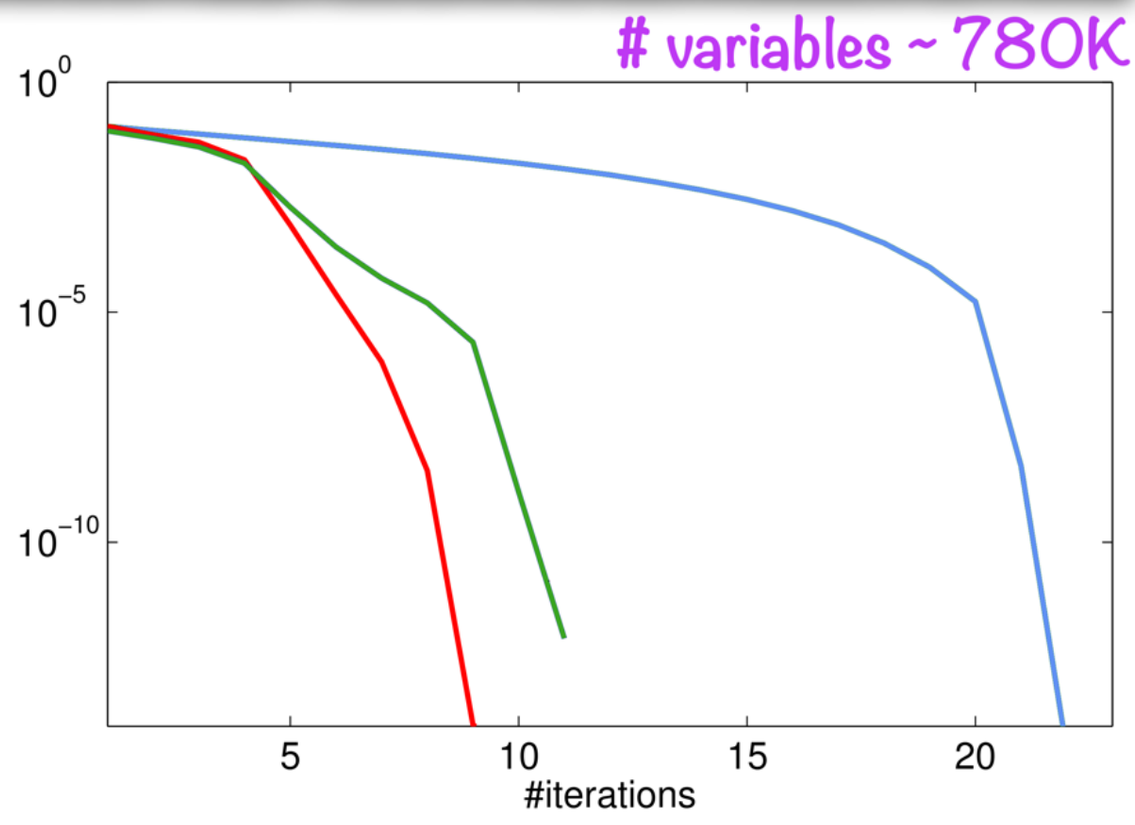
Results

$$\phi(\cdot) := f(\cdot) + g(\cdot)$$

$$\min_{\boldsymbol{\Theta}} \left\{ \underbrace{-\log\det(\boldsymbol{\Theta}) + \mathrm{trace}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta})}_{f(\cdot)} + \underbrace{\rho\|\boldsymbol{\Theta}\|_1}_{g(\cdot)} \right\}$$

Convergence behavior on gene data [Dem72, BGA08, HSDR11]



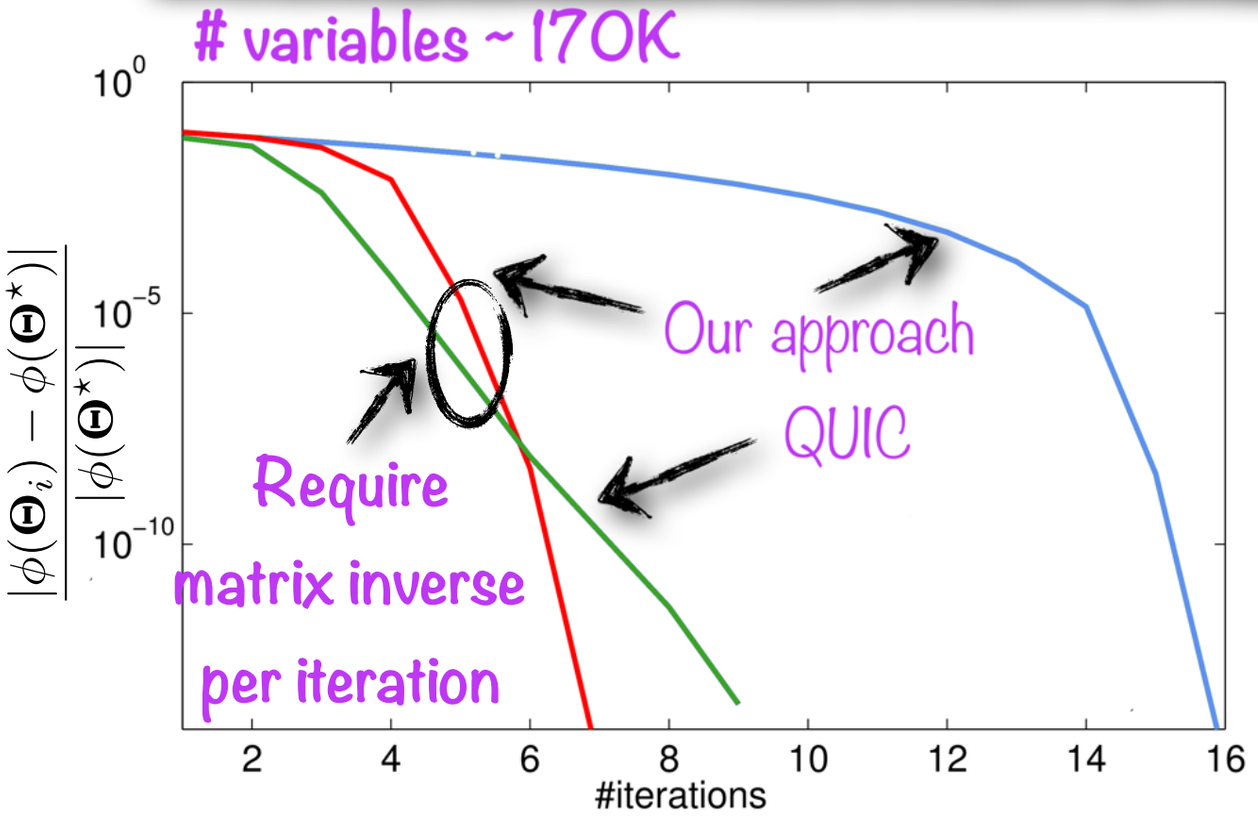


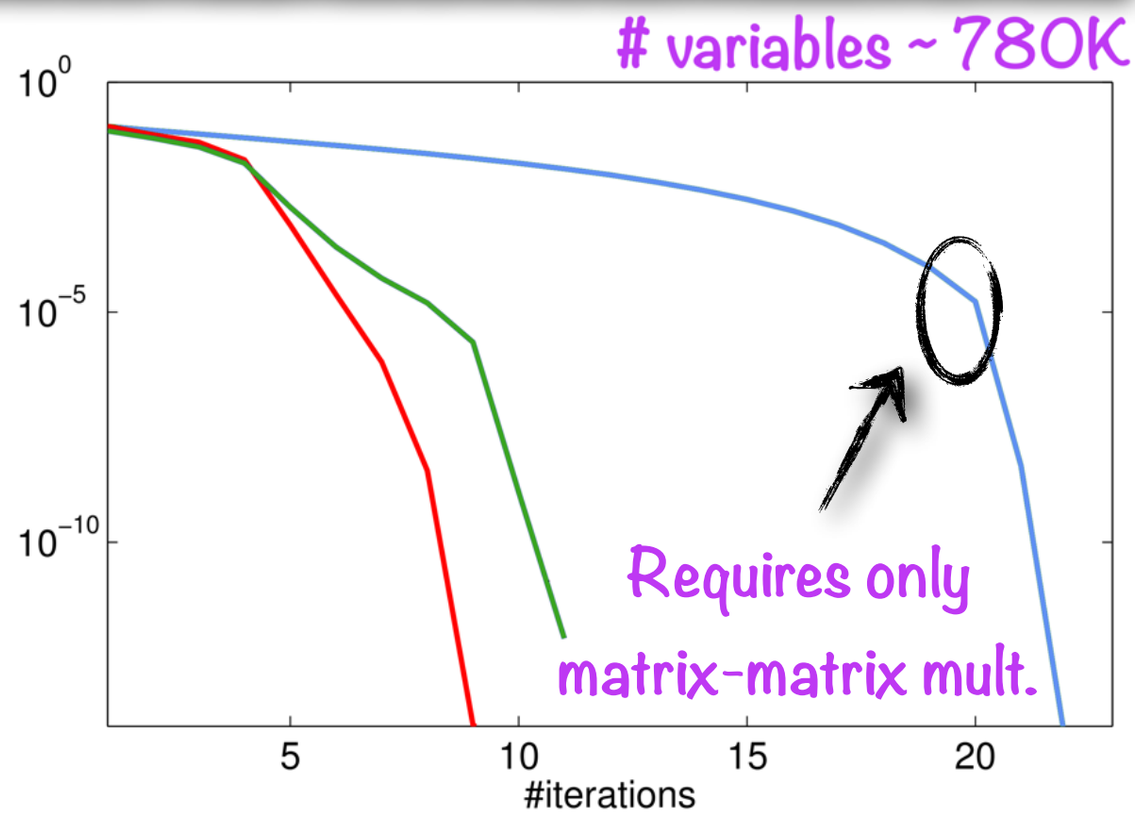
$$\nabla f(\boldsymbol{\Theta}) = \boldsymbol{\Theta}^{-1} + \widehat{\boldsymbol{\Sigma}}$$

$$\phi(\cdot) := f(\cdot) + g(\cdot)$$

$$\min_{\boldsymbol{\Theta}} \left\{ \underbrace{-\log\det(\boldsymbol{\Theta}) + \mathrm{trace}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta})}_{f(\cdot)} + \underbrace{\rho\|\boldsymbol{\Theta}\|_1}_{g(\cdot)} \right\}$$

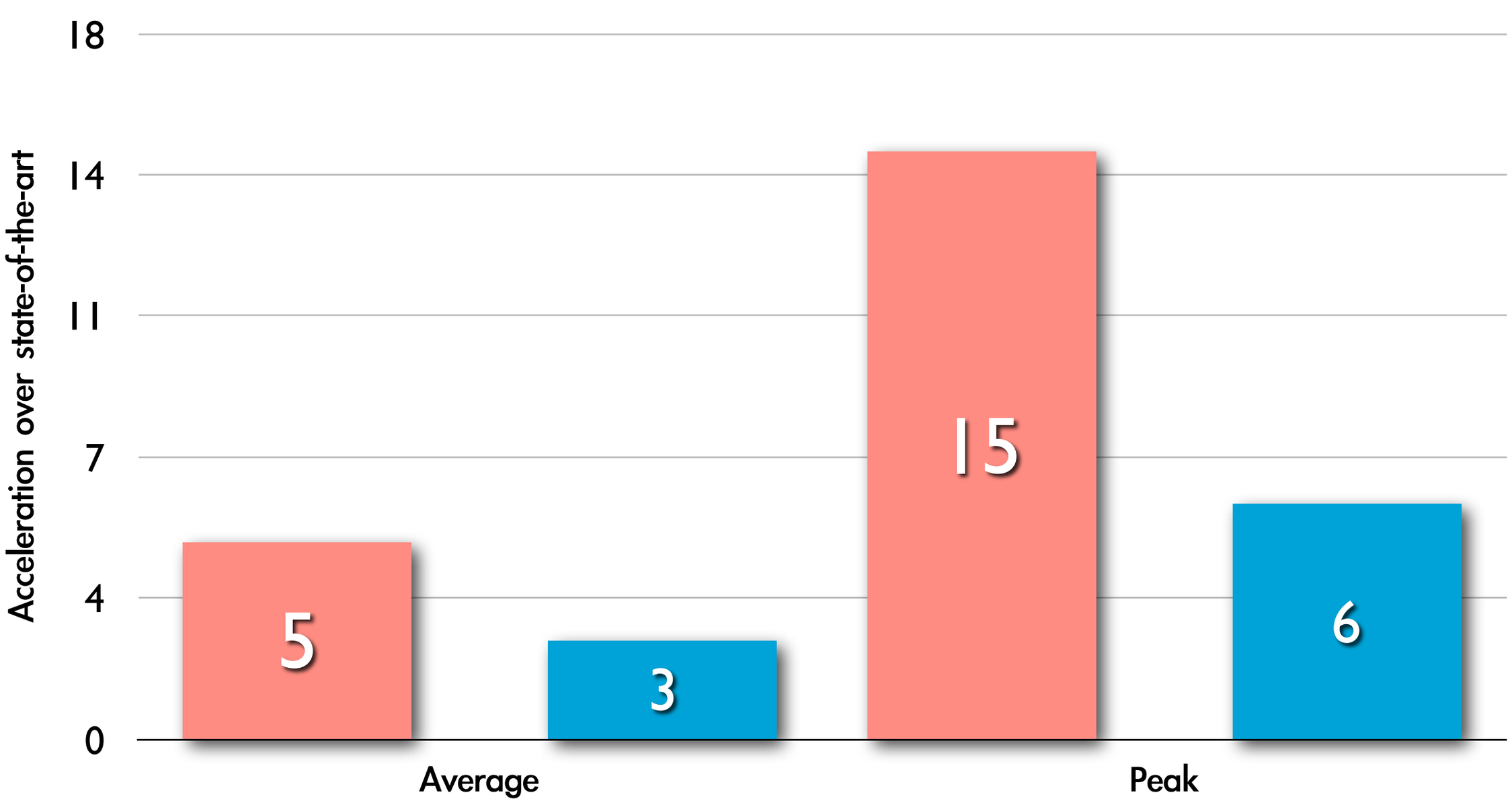
Convergence behavior on gene data [Dem72, BGA08, HSDR11]





Our approach with inversions/Cholesky`s dec.

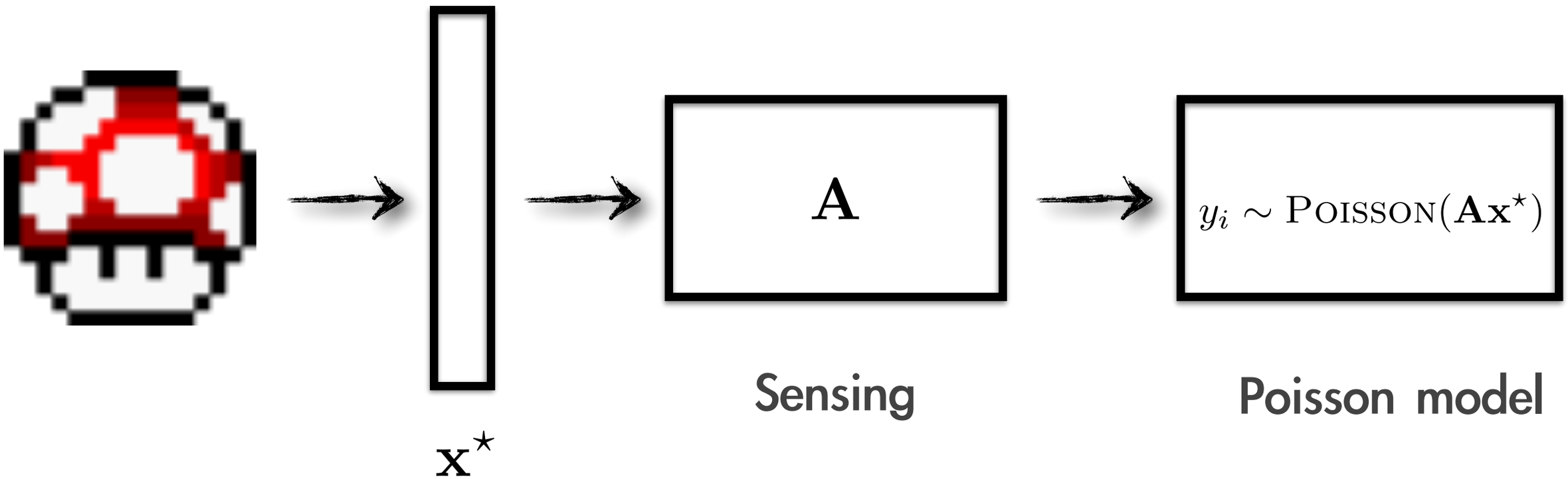
Our approach with matrix mult.



Application #2: Poisson imaging reconstruction

Optimization problem

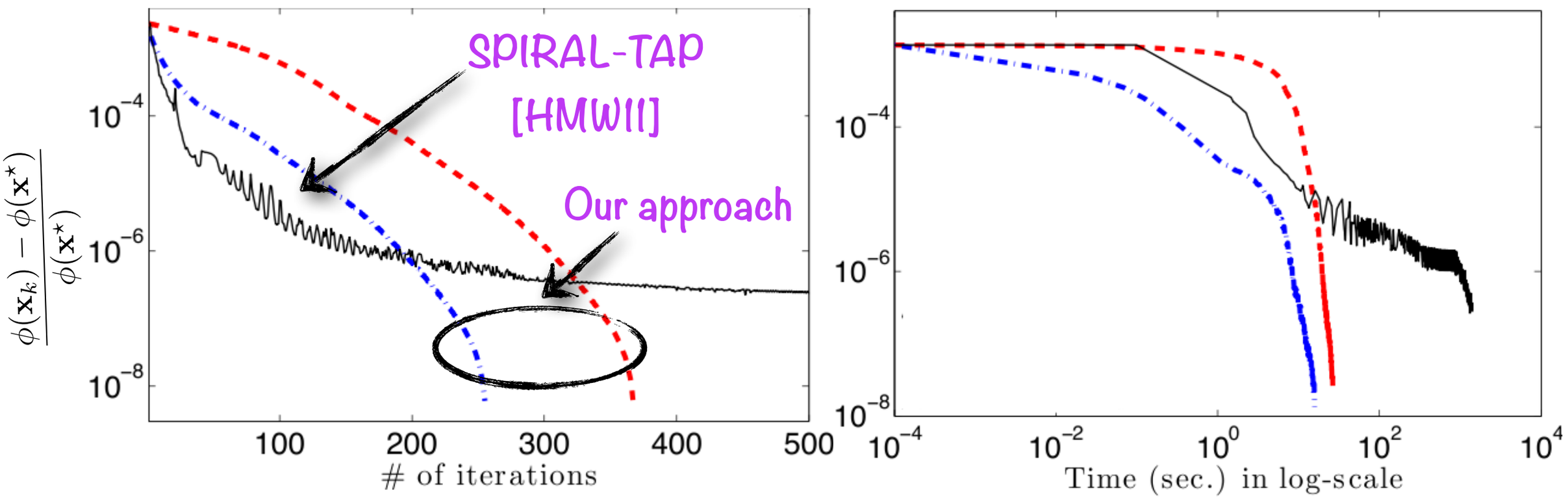
$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}} \left\{ \underbrace{\sum_{i=1}^{m} \mathbf{a}_i^T \mathbf{x} - \sum_{i=1}^{m} y_i \log\left(\mathbf{a}_i^T \mathbf{x} + \beta_i\right) + g(\mathbf{x})}_{:=\phi(\mathbf{x})} \right\}$$



Application #2: Poisson imaging reconstruction

Optimization problem

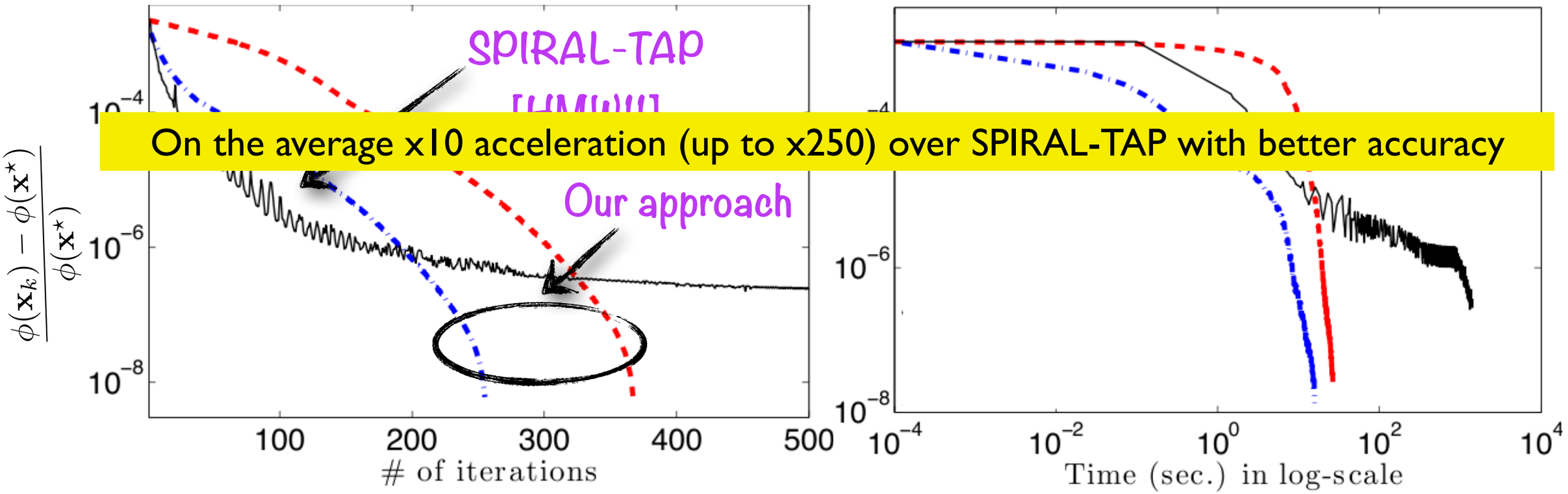
$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}} \left\{ \underbrace{\sum_{i=1}^{m} \mathbf{a}_i^T \mathbf{x} - \sum_{i=1}^{m} y_i \log\left(\mathbf{a}_i^T \mathbf{x} + \beta_i\right) + g(\mathbf{x})}_{:=\phi(\mathbf{x})} \right\}$$



Application #2: Poisson imaging reconstruction

Optimization problem

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}} \left\{ \underbrace{\sum_{i=1}^{m} \mathbf{a}_i^T \mathbf{x} - \sum_{i=1}^{m} y_i \log\left(\mathbf{a}_i^T \mathbf{x} + \beta_i\right) + g(\mathbf{x})}_{:=\phi(\mathbf{x})} \right\}$$



On the average x10 acceleration (up to x250) over SPIRAL-TAP with better accuracy

Bonus: Path following, composite self-concordant minimization

"An inexact proximal path-following algorithm for constrained convex minimization",
Tran-Dinh, Kyrillidis, Cevher, SIAM Journal on Optimization, 2015

Until this point

$$\underset{\mathbf{X} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad f(\mathbf{X}) + \lambda \cdot g(\mathbf{X})$$

- $f(\mathbf{X})$ appears naturally in the objective
- $f(\mathbf{X})$ is self-concordant

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad g(\mathbf{X})$$

$$\text{subject to} \quad \mathbf{X}\in\Omega$$

- $\Omega\subseteq\mathbb{R}^{n\times n}$ is endowed with a self-concordant barrier $f(\mathbf{X})$.

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad g(\mathbf{X})$$

$$\text{subject to} \quad \mathbf{X}\in\Omega$$

- $\Omega \subseteq \mathbb{R}^{n\times n}$ is endowed with a self-concordant barrier $f(\mathbf{X})$.

- **Examples:**

$$\begin{aligned}
\Omega : \mathbf{X} \succeq 0 \quad &\Rightarrow \quad f_\Omega(\mathbf{X}) = -\log\det(\mathbf{X}) \quad \text{(solves SDP problem)} \\
\Omega : \mathbf{a}^T\mathbf{x} \geq 0 \quad &\Rightarrow \quad f_\Omega(\mathbf{x}) = -\log(\mathbf{a}^T\mathbf{x}) \\
\Omega : \|\mathbf{A}\mathbf{x}-\mathbf{b}\| \leq \sigma \quad &\Rightarrow \quad f_\Omega(\mathbf{x}) = -\log(\sigma^2 - \|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2)
\end{aligned}$$

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad g(\mathbf{X})$$

$$\text{subject to} \quad \mathbf{X}\in\Omega$$

Approach

- Solve a sequence of composite self-concordant problems

$$\underset{\mathbf{X}\in\text{int}(\Omega)}{\text{minimize}} \quad g(\mathbf{X}) + t\cdot f(\mathbf{X})$$

as t decreases to zero

Contributions in a nutshell:

$$\underset{\mathbf{X}\in\mathrm{int}(\Omega)}{\text{minimize}} \quad g(\mathbf{X}) + t \cdot f(\mathbf{X})$$

- First (?) path following scheme in non-smooth, composite form.

Contributions in a nutshell:

$$\underset{\mathbf{X}\in\mathrm{int}(\Omega)}{\text{minimize}} \quad g(\mathbf{X}) + t \cdot f(\mathbf{X})$$

- First (?) path following scheme in non-smooth, composite form.
- Adaptive selection of t_k values such that we approximately track the solution trajectory.

Contributions in a nutshell:

$$\underset{\mathbf{X}\in\mathrm{int}(\Omega)}{\text{minimize}} \quad g(\mathbf{X}) + t \cdot f(\mathbf{X})$$

- First (?) path following scheme in non-smooth, composite form.
- Adaptive selection of t_k values such that we approximately track the solution trajectory.
- For each t_k, new estimate is computed from a one-iteration proximal Newton scheme!

Contributions in a nutshell:

$$\underset{\mathbf{X}\in\mathrm{int}(\Omega)}{\text{minimize}} \quad g(\mathbf{X}) + t \cdot f(\mathbf{X})$$

- First (?) path following scheme in non-smooth, composite form.
- Adaptive selection of t_k values such that we approximately track the solution trajectory.
- For each t_k, new estimate is computed from a one-iteration proximal Newton scheme!
- Rigorous convergence guarantees (not presented here).

Contributions in a nutshell:

$$\underset{\mathbf{X}\in\mathrm{int}(\Omega)}{\text{minimize}} \quad g(\mathbf{X}) + t \cdot f(\mathbf{X})$$

- First (?) path following scheme in non-smooth, composite form.
- Adaptive selection of t_k values such that we approximately track the solution trajectory.
  - For each t_k, new estimate is computed from a one-iteration proximal Newton scheme!
- Rigorous convergence guarantees (not presented here).

Algorithm in a nutshell:

- Adaptively set t_k (closed form solution).
- Solve $\arg\min\{g(\mathbf{x}) + t_k \cdot f(\mathbf{x})\}$ up to δ accuracy with one-iteration proximal Newton scheme.

Points on central path

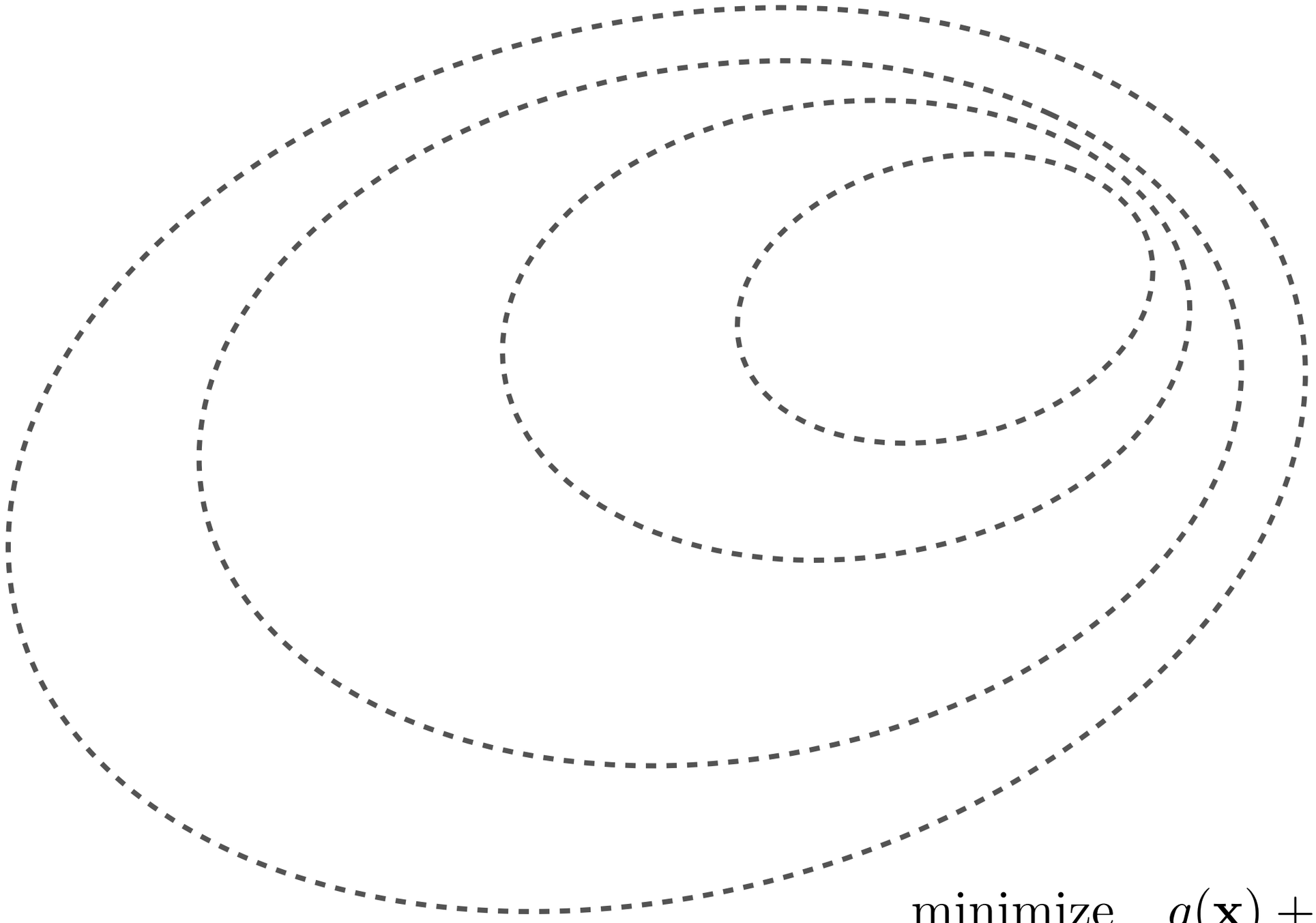
Central path trajectory

Inexact solutions

Points on central path

Central path trajectory

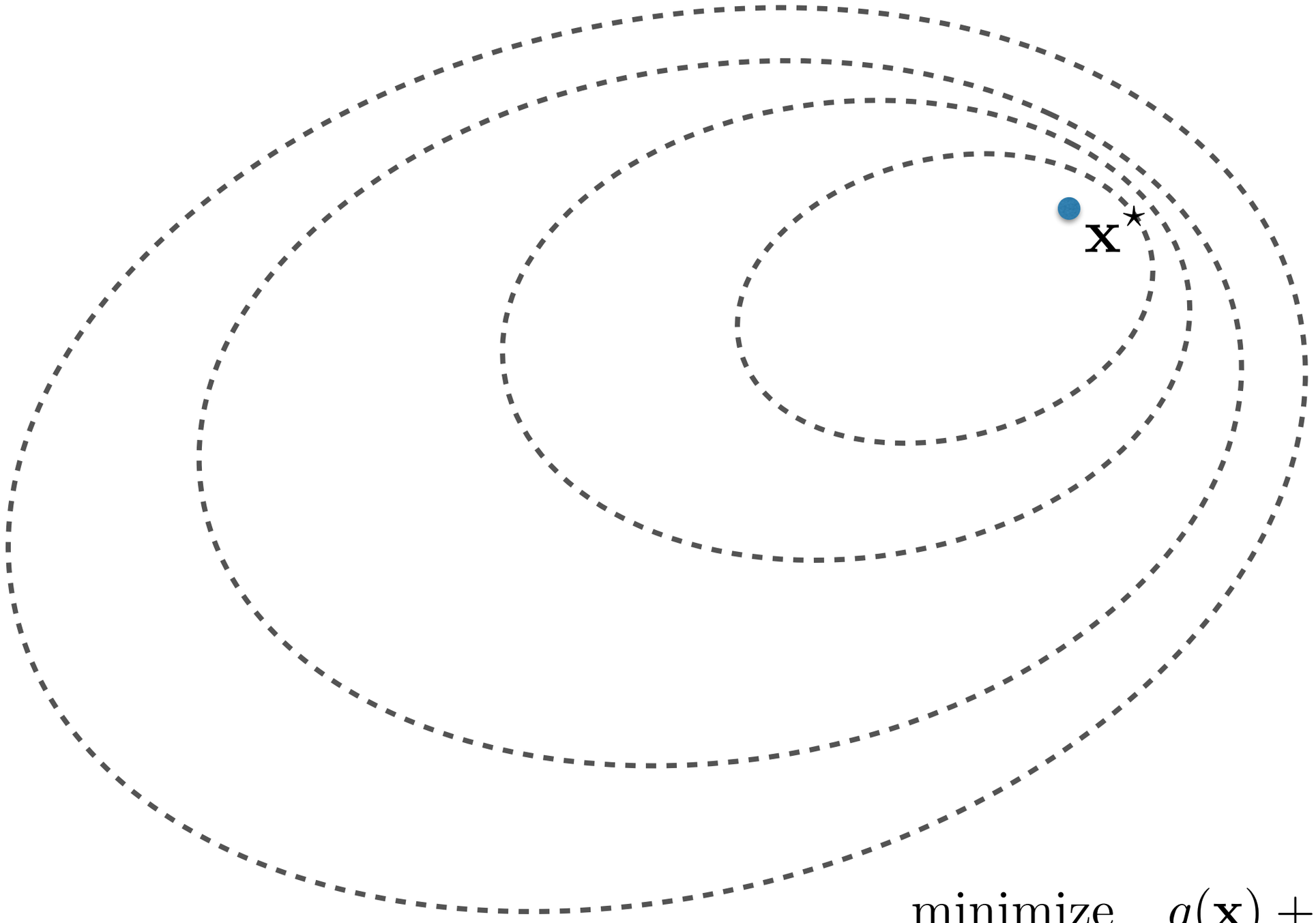
Inexact solutions



$$\underset{\mathbf{x}\in\mathrm{int}(\Omega)}{\text{minimize}} \quad g(\mathbf{x}) + t \cdot f(\mathbf{x})$$

- Points on central path
- Central path trajectory
- Inexact solutions

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\Omega} g(\mathbf{x})$$

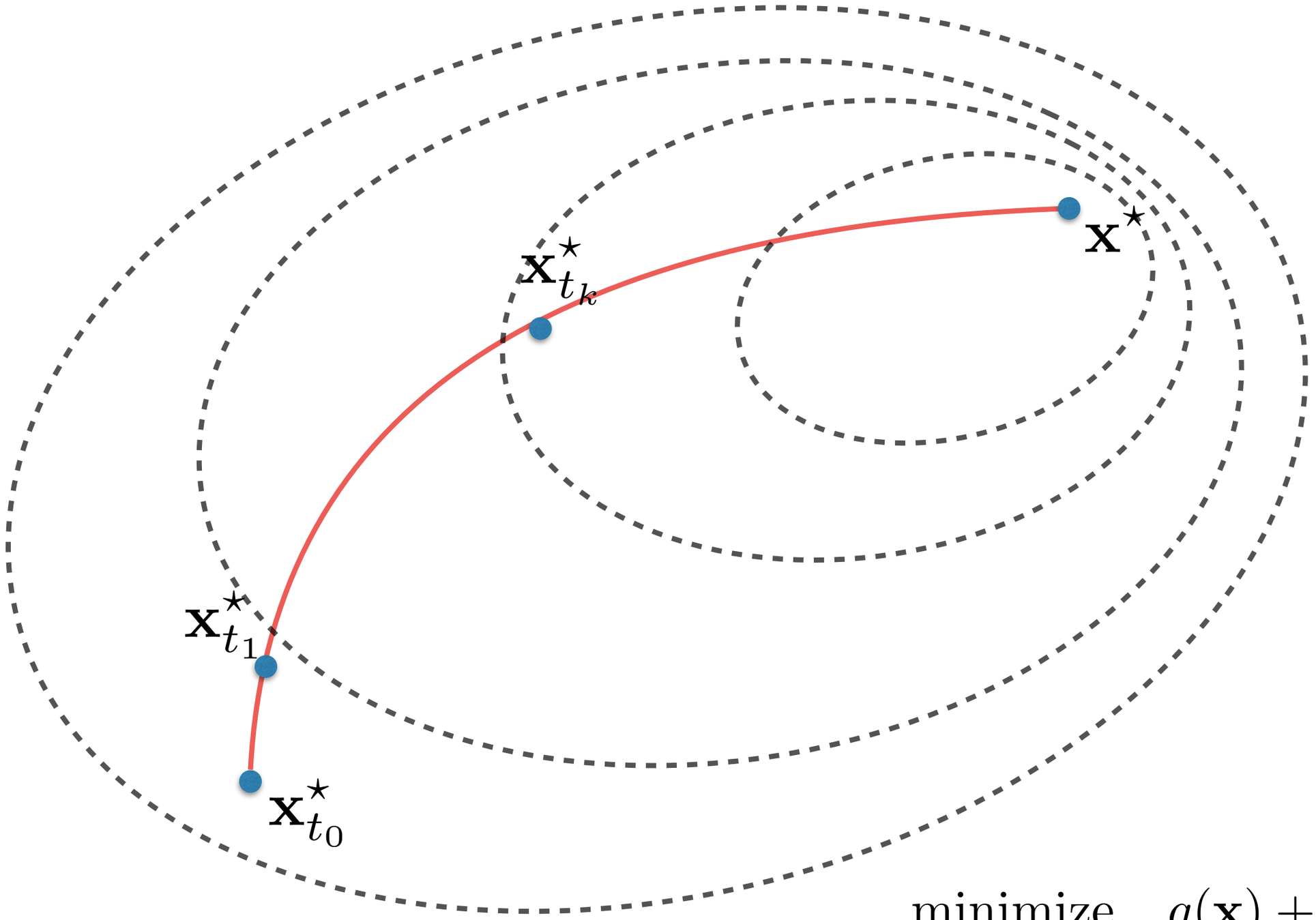


$$\underset{\mathbf{x}\in\mathrm{int}(\Omega)}{\text{minimize}} \quad g(\mathbf{x}) + t \cdot f(\mathbf{x})$$

- Points on central path
- Central path trajectory
- Inexact solutions

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\Omega} g(\mathbf{x})$$

$$\mathbf{x}^\star_{t_k} = \arg\min \{g(\mathbf{x}) + t_k \cdot f(\mathbf{x})\}$$



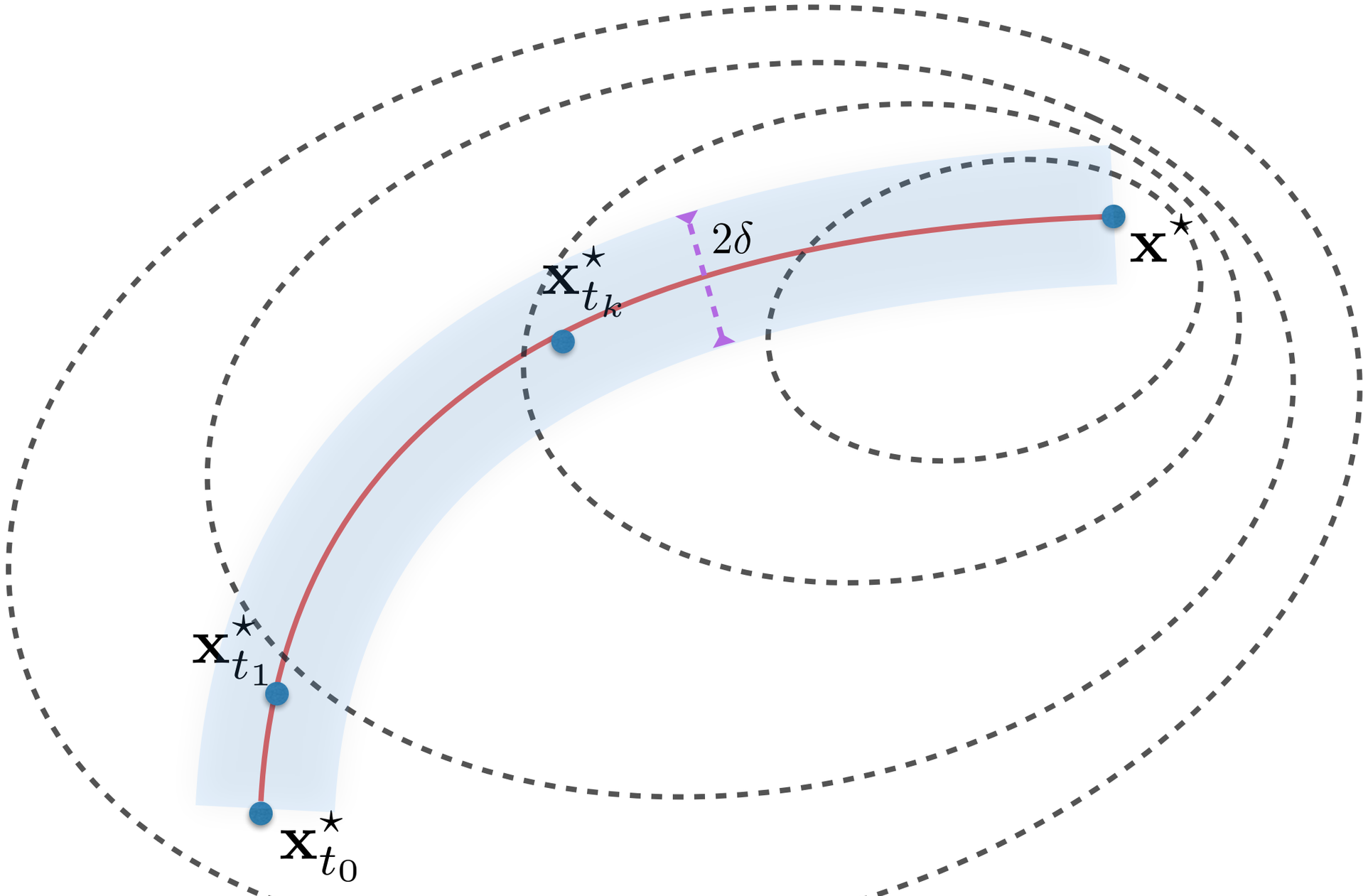
$$\underset{\mathbf{x}\in\mathrm{int}(\Omega)}{\text{minimize}} \quad g(\mathbf{x}) + t \cdot f(\mathbf{x})$$

- Points on central path
- Central path trajectory
- Inexact solutions

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\Omega} g(\mathbf{x})$$

$$\mathbf{x}^\star_{t_k} = \arg\min \{g(\mathbf{x}) + t_k \cdot f(\mathbf{x})\}$$

$\mathbf{x}^k$: δ-approximate solution with one proximal Newton step!



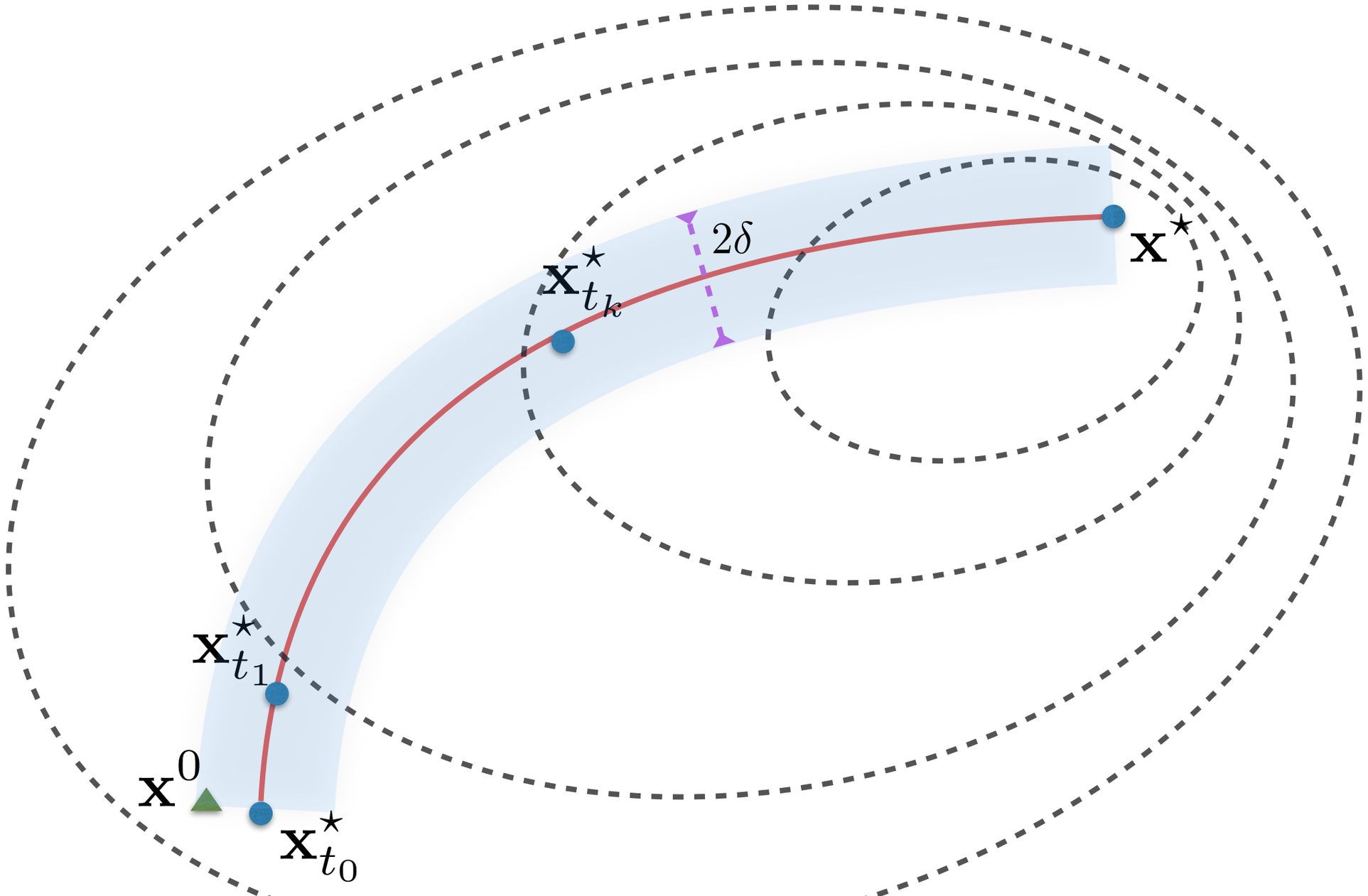
$$\underset{\mathbf{x}\in\mathrm{int}(\Omega)}{\text{minimize}} \quad g(\mathbf{x}) + t \cdot f(\mathbf{x})$$

- Points on central path
- Central path trajectory
- Inexact solutions

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\Omega} g(\mathbf{x})$$

$$\mathbf{x}^\star_{t_k} = \arg\min \{g(\mathbf{x}) + t_k \cdot f(\mathbf{x})\}$$

$\mathbf{x}^k$: δ-approximate solution with one proximal Newton step!



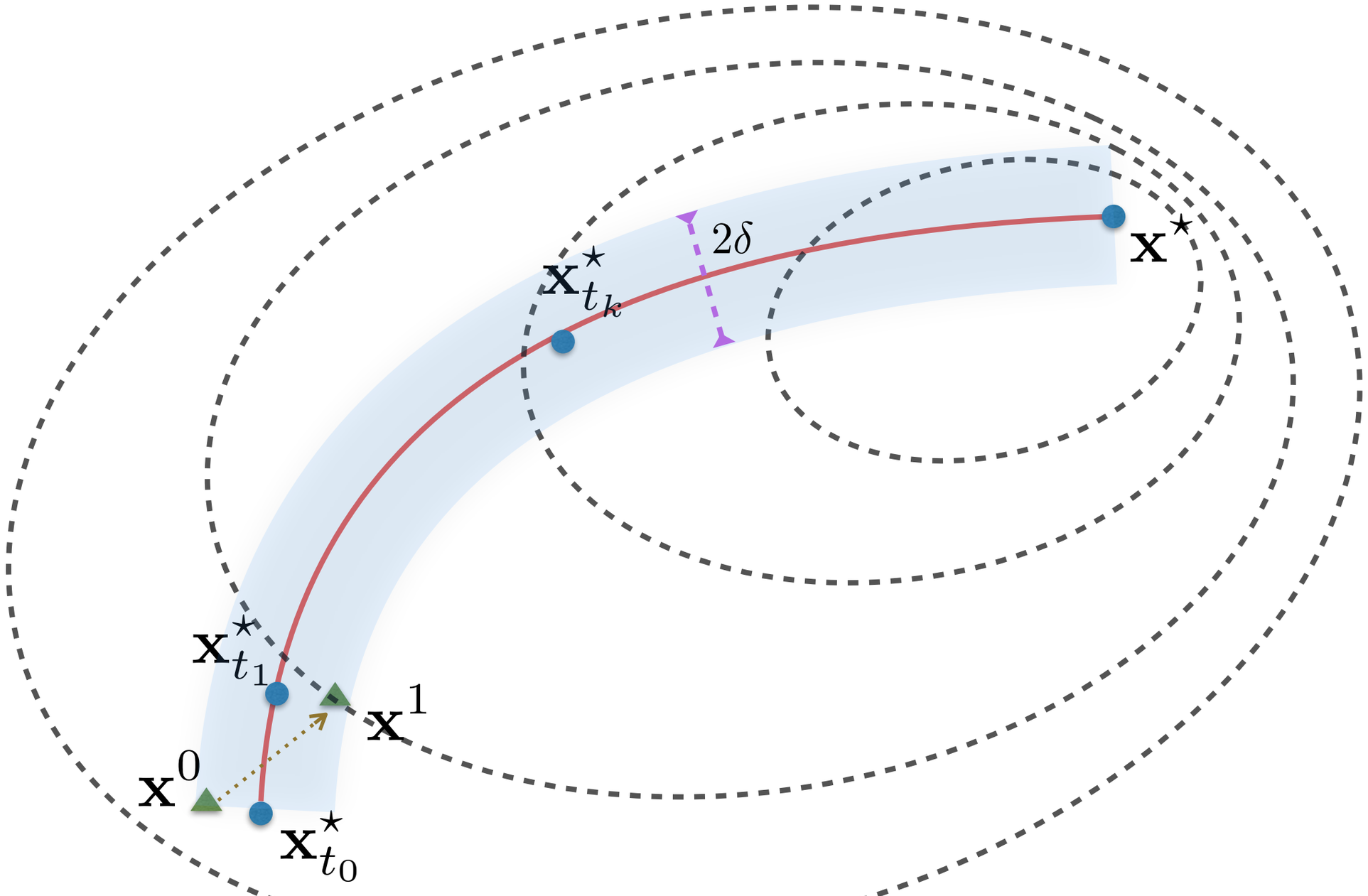
$$\underset{\mathbf{x}\in\mathrm{int}(\Omega)}{\text{minimize}} \quad g(\mathbf{x}) + t \cdot f(\mathbf{x})$$

- Points on central path
- Central path trajectory
- Inexact solutions

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\Omega} g(\mathbf{x})$$

$$\mathbf{x}^\star_{t_k} = \arg\min \{g(\mathbf{x}) + t_k \cdot f(\mathbf{x})\}$$

$\mathbf{x}^k$: δ-approximate solution with one proximal Newton step!



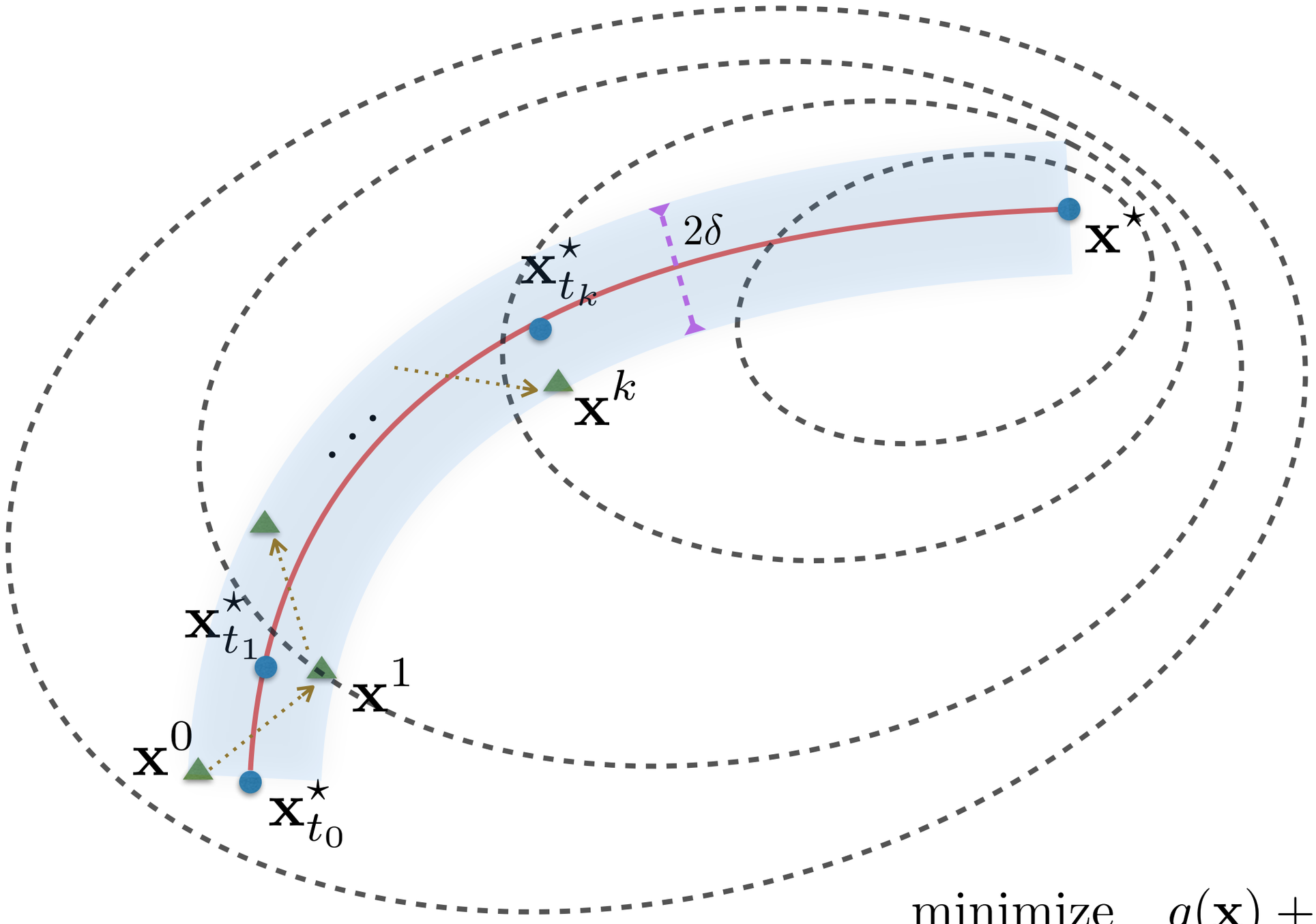
$$\underset{\mathbf{x}\in\mathrm{int}(\Omega)}{\text{minimize}} \quad g(\mathbf{x}) + t \cdot f(\mathbf{x})$$

- Points on central path
- Central path trajectory
- Inexact solutions

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\Omega} g(\mathbf{x})$$

$$\mathbf{x}^\star_{t_k} = \arg\min \{g(\mathbf{x}) + t_k \cdot f(\mathbf{x})\}$$

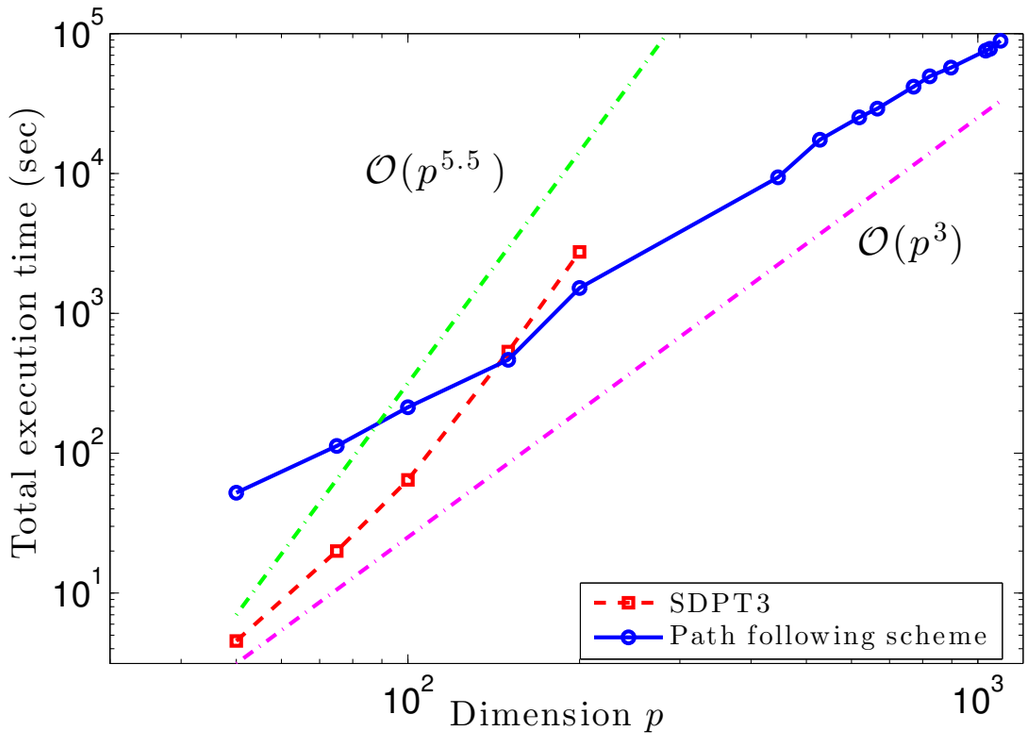
$\mathbf{x}^k$: δ-approximate solution with one proximal Newton step!



$$\underset{\mathbf{x}\in\mathrm{int}(\Omega)}{\text{minimize}} \quad g(\mathbf{x}) + t \cdot f(\mathbf{x})$$

Application #3: Max-norm clustering

$$\begin{array}{ll} \min\limits_{\mathbf{L},\mathbf{R},\mathbf{K}} & \|\mathrm{vec}\,(\mathbf{K}-\mathbf{A})\|_1 \\ \text{s.t.} & \begin{bmatrix} \mathbf{L} & \mathbf{K} \\ \mathbf{K}^T & \mathbf{R} \end{bmatrix} \succ 0,\ \mathbf{L}_{ii} \le 1,\ \mathbf{R}_{ii} \le 1,\ i = 1,\ldots,p. \end{array}$$



DCO:

	SDPT3		PF scheme
p	variables	constraints	variables
50	15.1	2.6	10
75	33.9	5.8	22.5
100	60.2	10.2	40
150	135.3	22.8	90
200	240.4	40.4	160

(in thousands)

Overview

- New results in composite convex optimization
- Non-smooth + self-concordant objective function
- Extensions to path-following schemes for constrained optimization
- Several extensions to be exploited

Thank you!