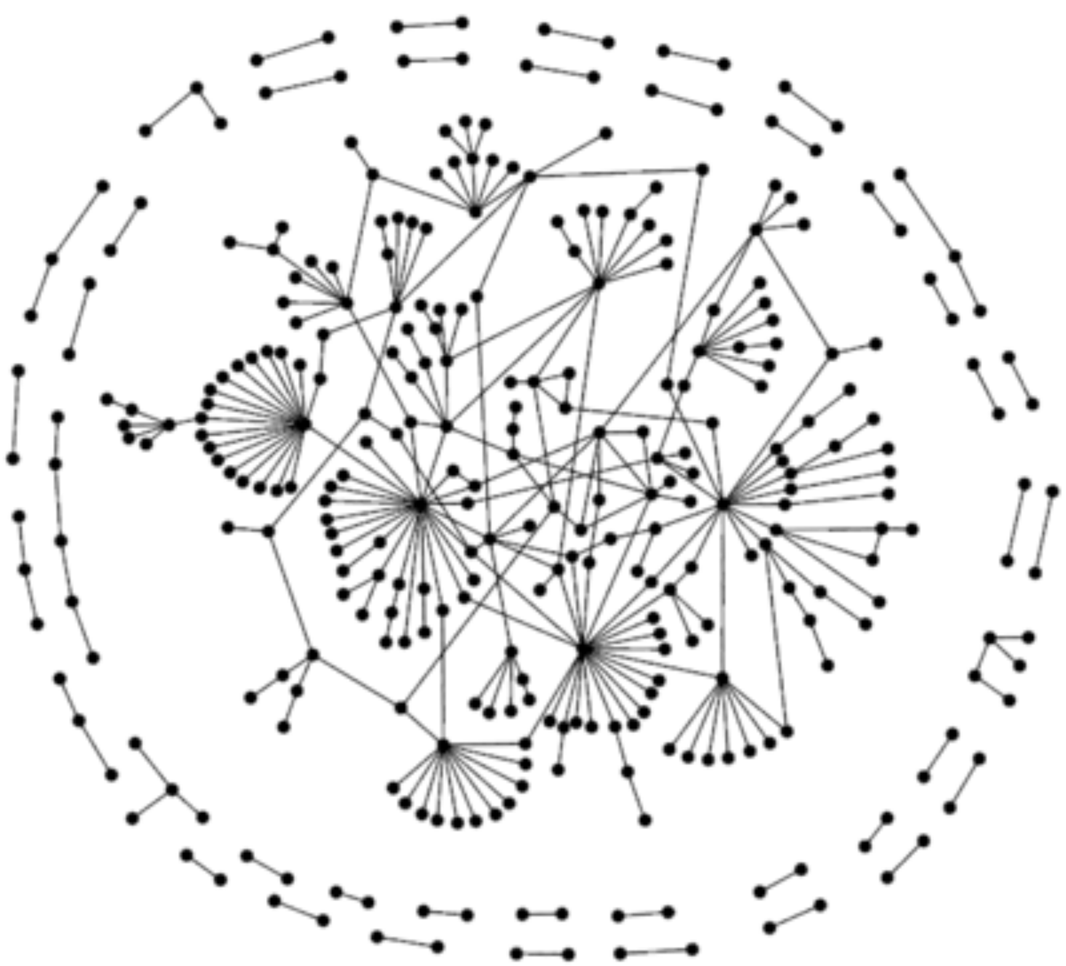
Composite Self-concordant Minimization

Volkan Cevher

*Laboratory for Information and Inference Systems-**LIONS***

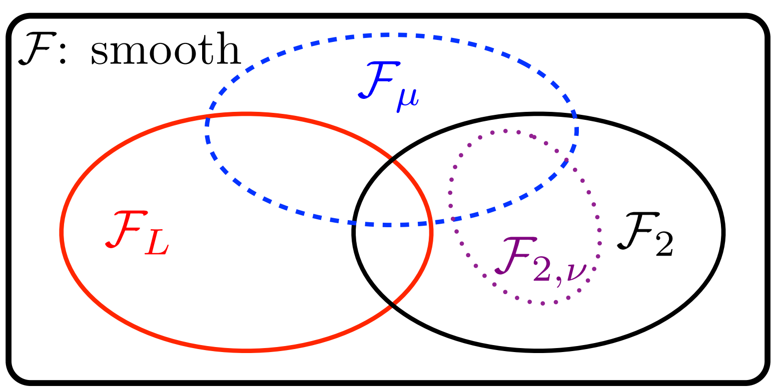
Ecole Polytechnique Federale de Lausanne (EPFL)

volkan.cevher@epfl.ch



UCL-Gatsby

June 12, 2014



joint work with

Quoc Tran Dinh

Anastasios Kyrillidis

Yen-Huan Li

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

lions@epfl

MARIE CURIE ACTIONS

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

erc

Composite minimization

$$\text{(P)} \qquad \min_{\mathbf{x} \in \mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$

f
convex and **smooth**

g
convex and possibly **nonsmooth**

Motivation

Problem (P) *covers many practical problems*:

- Unconstrained basic LASSO / logistic regression
- Graphical model selection / latent variable graphical model selection
- Poisson imaging reconstruction / LASSO problem with unknown variance
- Low-rank recovery / clustering
- Atomic norm regularization / off-the-grid array processing

g: ℓ_1-norm, nuclear norm or indicator functions

Composite minimization

(P) $$\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$

f
convex and **smooth**

g
convex and possibly **nonsmooth**

Motivation

Problem (P) *covers many practical* **LARGE SCALE** *problems*:

- Unconstrained basic LASSO / logistic regression
- Graphical model selection / latent variable graphical model selection
- Poisson imaging reconstruction / LASSO problem with unknown variance
- Low-rank recovery / clustering
- Atomic norm regularization / off-the-grid array processing

need scalable algorithms

g: ℓ_1-norm, nuclear norm or indicator functions

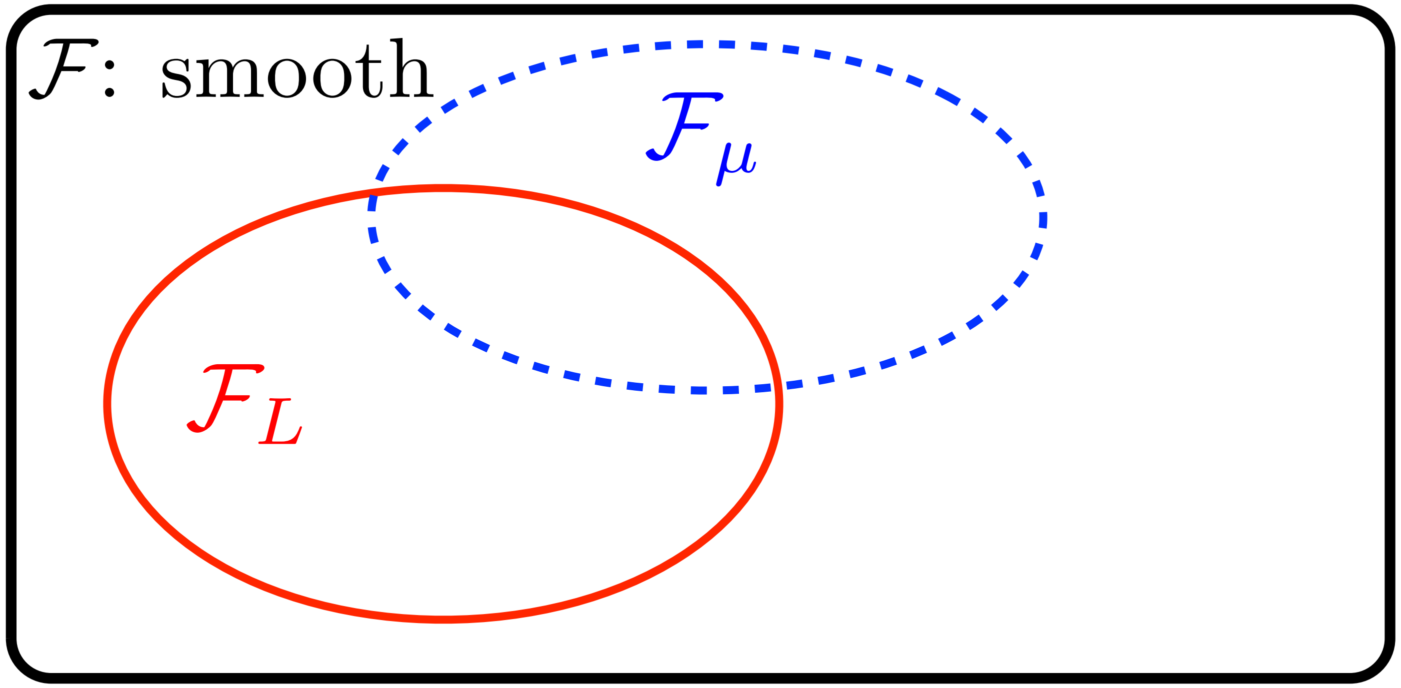
Composite minimization: modus operandi

(P) $$\min_{\mathbf{x} \in \mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$

f
convex and **smooth**

g
convex and possibly **nonsmooth**

Classes of smooth functions (f)



- $\mathcal{F}_L$ - L-Lipschitz gradient
- $\mathcal{F}_\mu$ - μ-strongly convex

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\|$$

$$\mu \mathbb{I} \preceq \nabla^2 f(x) \preceq L \mathbb{I}$$

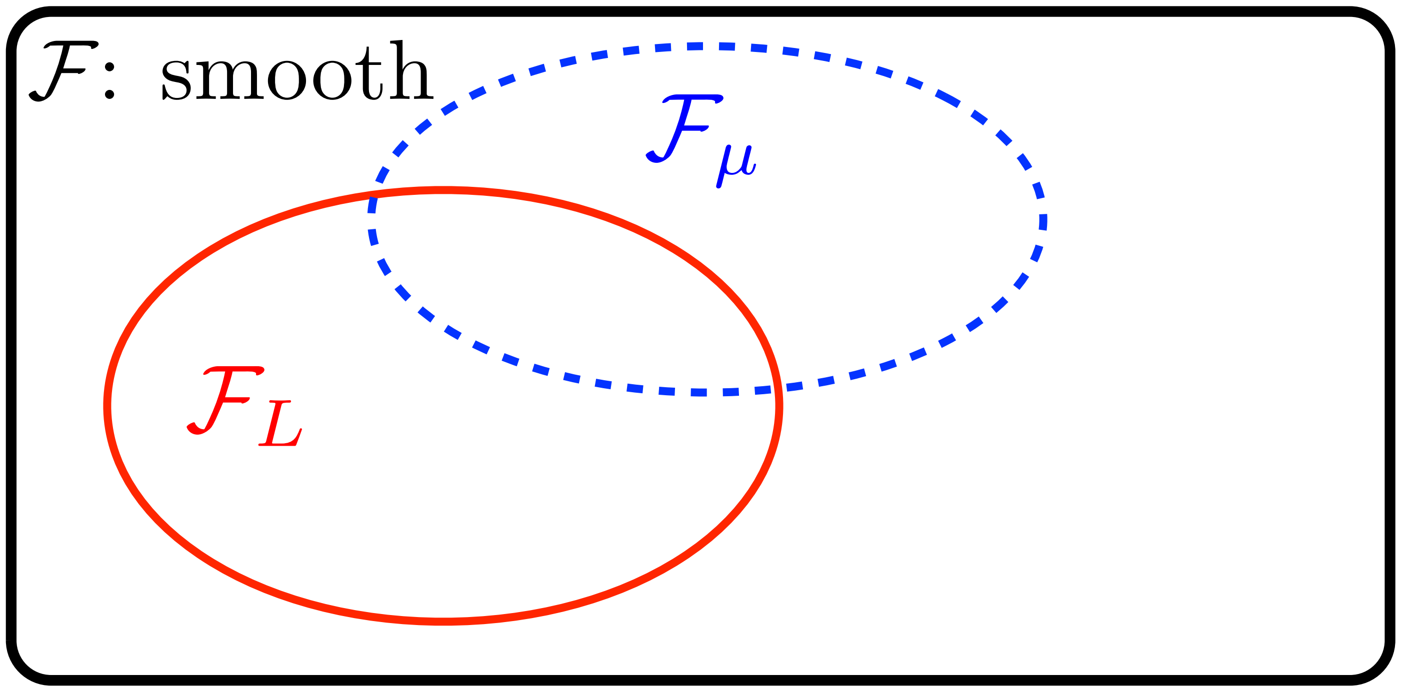
Composite minimization: modus operandi

(P) $$\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$

f
convex and **smooth**

g
convex and possibly **nonsmooth**

Classes of smooth functions (f)



- $\mathcal{F}_L$ - L-Lipschitz gradient
- $\mathcal{F}_\mu$ - μ-strongly convex

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\|$$

$$\mu\mathbb{I} \preceq \nabla^2 f(x) \preceq L\mathbb{I}$$

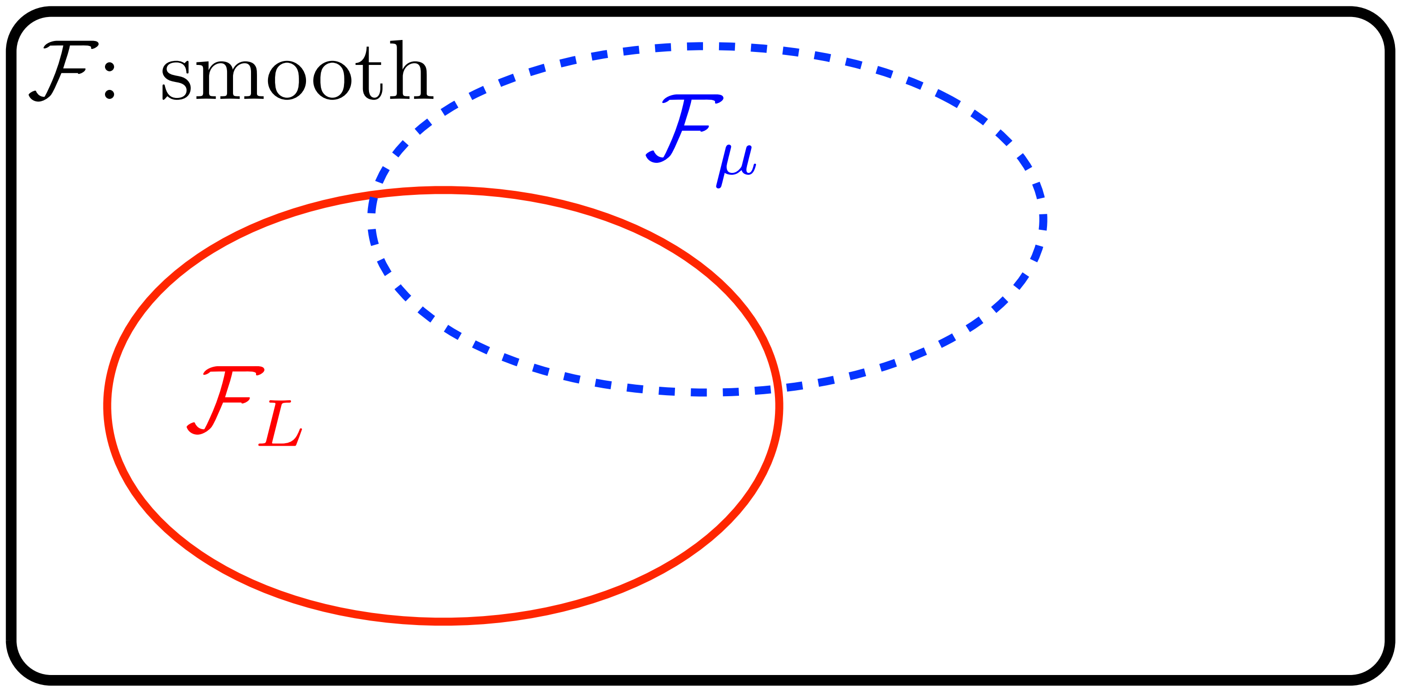
Composite minimization: modus operandi

(P) $$\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$

f
convex and **smooth**

g
convex and possibly **nonsmooth**

Classes of smooth functions (f)



- $\mathcal{F}_L$ - L-Lipschitz gradient
- $\mathcal{F}_\mu$ - μ-strongly convex

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\|$$

$$\mu\mathbb{I} \preceq \nabla^2 f(x) \preceq L\mathbb{I}$$

Following **prox** computation is ***tractable:***

$$\mathrm{prox}_{\gamma g}(\mathbf{s}) := \arg\min_{\mathbf{x}} \left\{ g(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{s}\|_2^2 \right\}$$

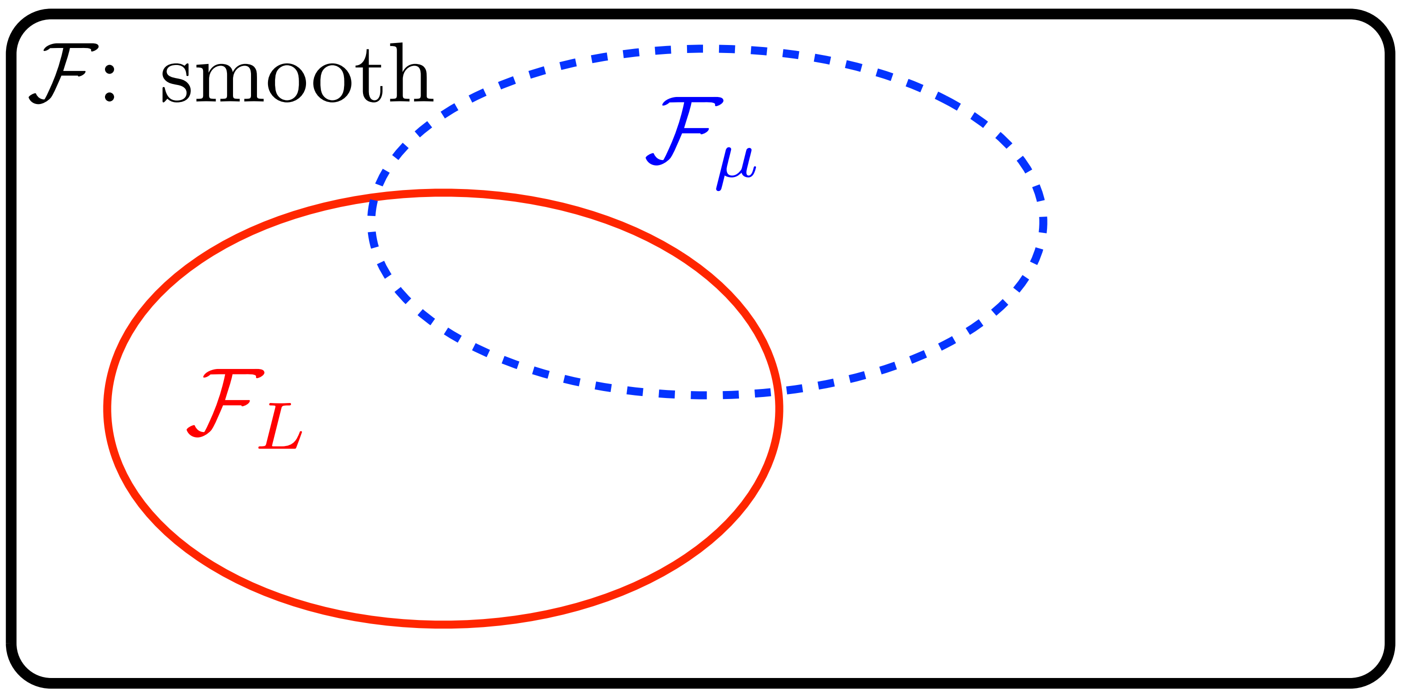
Composite minimization: modus operandi

(P) $$\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$

f
convex and **smooth**

g
convex and possibly **nonsmooth**

Classes of smooth functions (f)



- $\mathcal{F}_L$ - L-Lipschitz gradient
- $\mathcal{F}_\mu$ - μ-strongly convex

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\|$$

$$\mu\mathbb{I} \preceq \nabla^2 f(x) \preceq L\mathbb{I}$$

Following **prox** computation is ***tractable:***

$$\mathrm{prox}_{\gamma g}(\mathbf{s}) := \arg\min_{\mathbf{x}} \left\{ g(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{s}\|_2^2 \right\}$$

Example:

if $g(\mathbf{x}) = \|\mathbf{x}\|_1$, then
$\mathrm{prox}_{\gamma g}(\mathbf{s}) = \mathrm{SoftThresh}(\mathbf{s}, \gamma)$

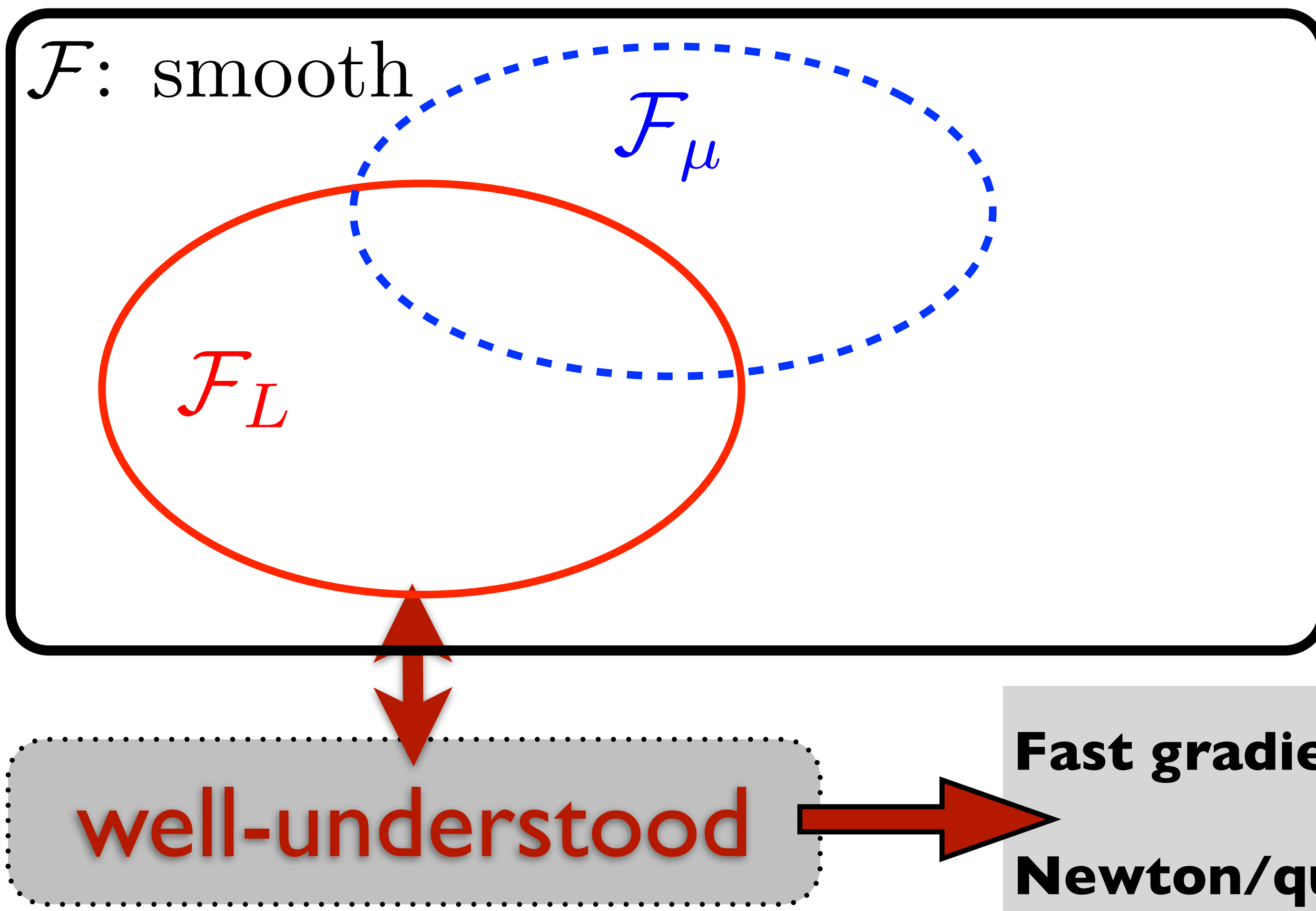
Composite minimization: modus operandi

$$\text{(P)} \qquad \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}) \right\}$$

f
convex and **smooth**

g
convex and possibly **nonsmooth**
with **"tractable"** prox

Classes of smooth functions (f)



- $\mathcal{F}_L$ - L-Lipschitz gradient
- $\mathcal{F}_\mu$ - μ-strongly convex

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\|$$

$$\mu\mathbb{I} \preceq \nabla^2 f(x) \preceq L\mathbb{I}$$

Fast gradient schemes (Nesterov's methods)

Newton/quasi Newton schemes

Composite minimization: an uncharted region

(P) $$\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$

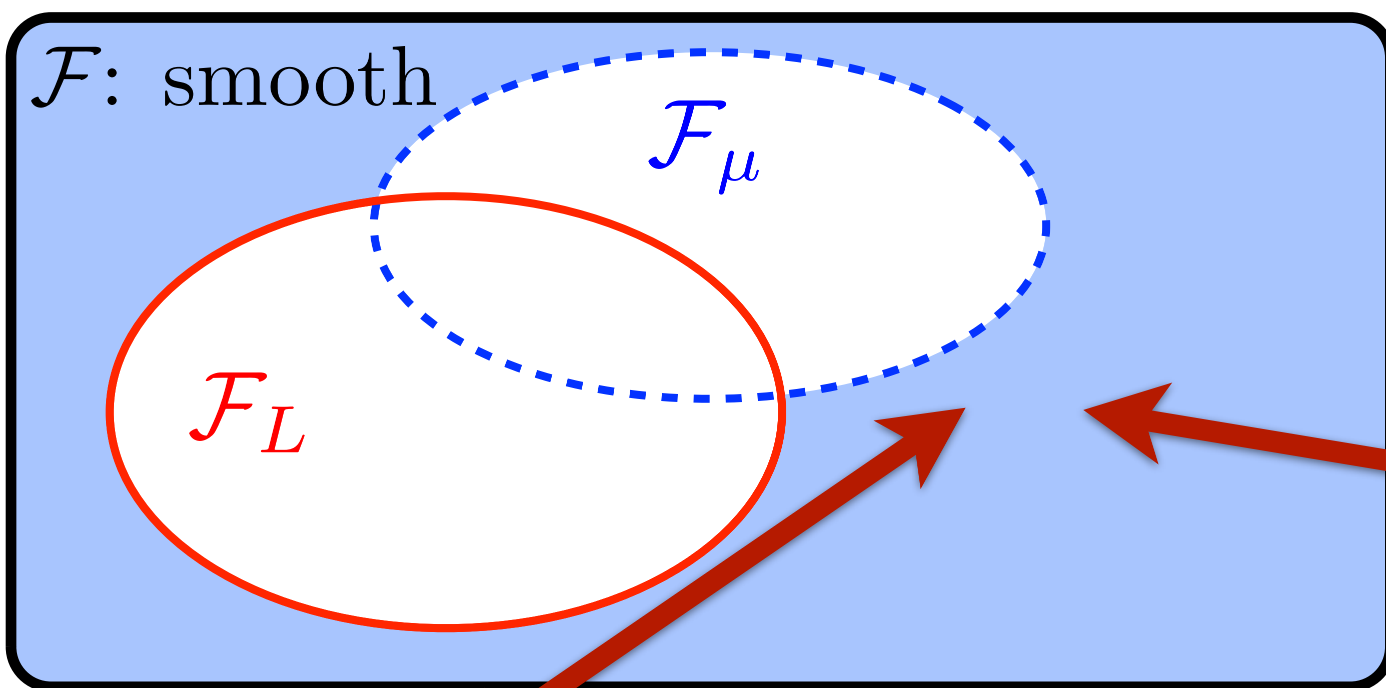
f

convex and **smooth**

g

convex and possibly **nonsmooth**
with **“tractable”** prox

Classes of smooth functions (f)



- $\mathcal{F}_L$ - L-Lipschitz gradient
- $\mathcal{F}_\mu$ - μ-strongly convex

$$\cancel{\|\nabla f(x) - \nabla f(y)\| \le L\|y - x\|}$$

$$\cancel{\mu\mathbb{I} \preceq \nabla^2 f(x) \preceq L\mathbb{I}}$$

Scalability is NOT great

~~**Fast gradient schemes (Nesterov’s methods)**~~

~~**Newton/quasi Newton schemes**~~

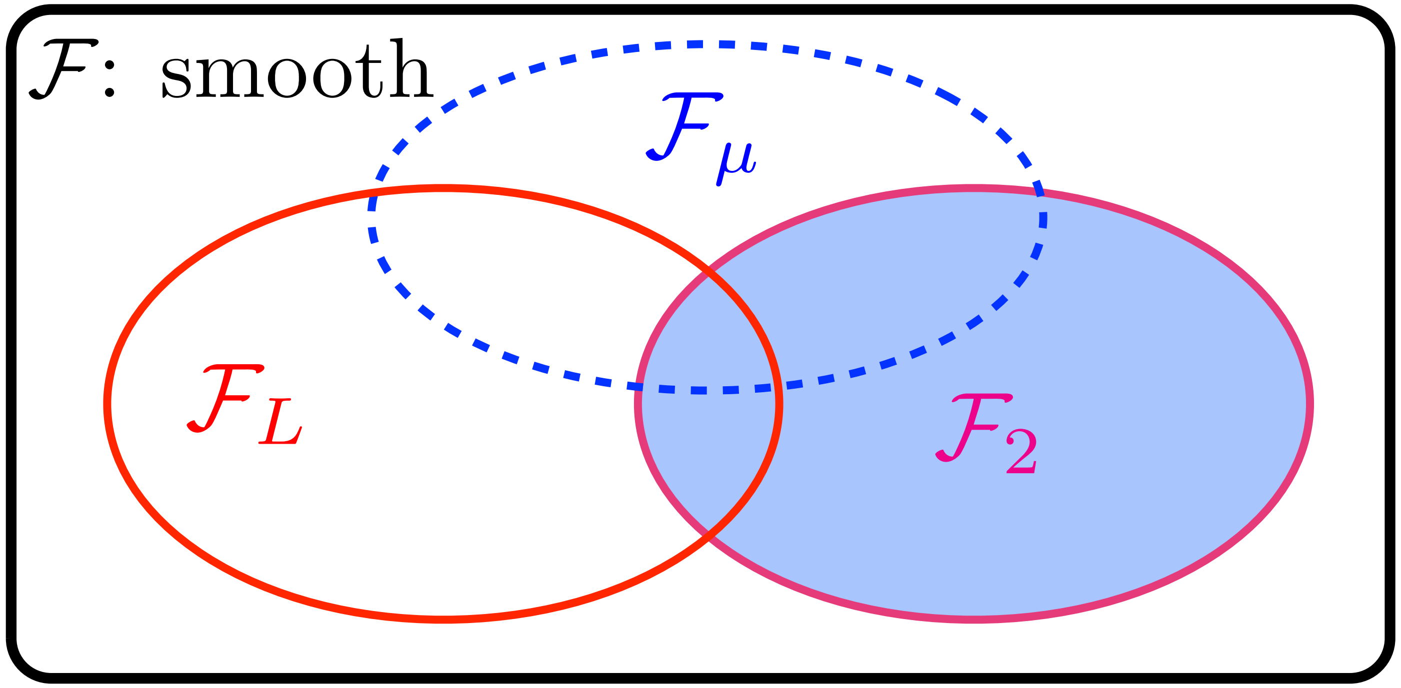
Composite self-concordant minimization

(P) $$\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$

f
convex and **self-concordant**

g
convex and possibly **nonsmooth**
with **"tractable"** prox

Classes of smooth functions (f)



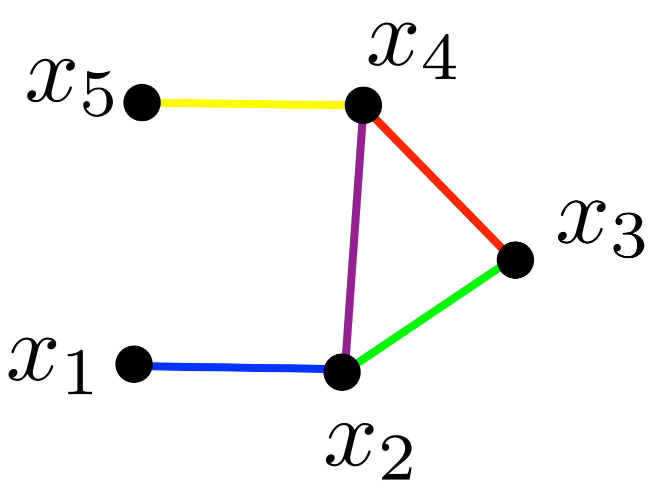
- $\mathcal{F}_L$ - L-Lipschitz gradient
- $\mathcal{F}_\mu$ - μ-strongly convex
- $\mathcal{F}_2$ - self-concordant

Key structure for the interior point method

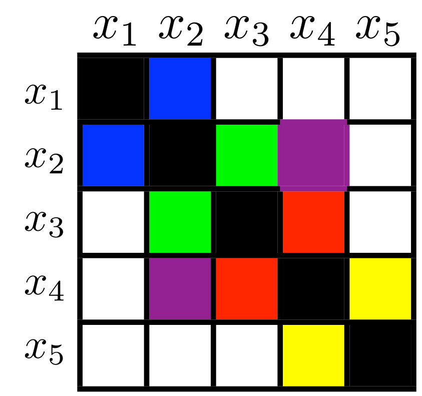
f is self-concordant if $\varphi(t) := f(\mathbf{x} + t\mathbf{d})$ satisfies $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$

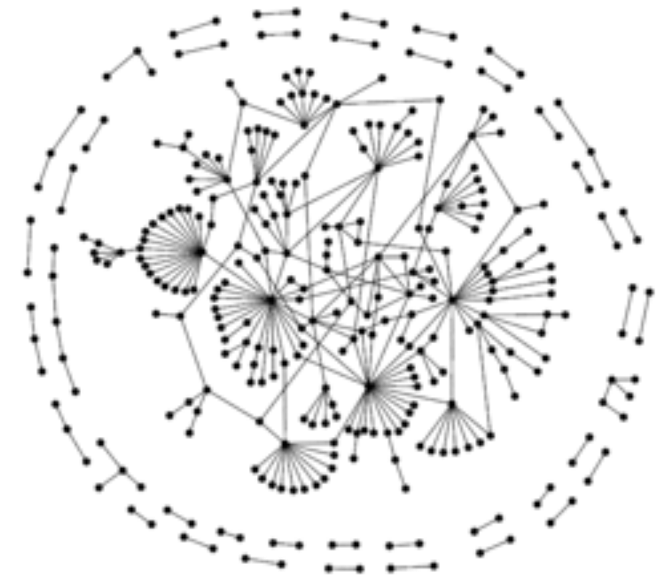
Example: Log-determinant for LMIs

- ***Application***: Graphical model selection



$\Theta =$





Given a data set $\mathcal{D} := \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_i$ is a Gaussian random variable. Let $\boldsymbol{\Sigma}$ be the **covariance matrix** corresponding to the **graphical model** of the Gausian Markov random field. The aim is to learn a **sparse matrix** Θ that approximates the inverse $\boldsymbol{\Sigma}^{-1}$.

Optimization problem
$$\min_{\boldsymbol{\Theta} \succ 0} \left\{ \underbrace{-\log\det(\boldsymbol{\Theta}) + \mathrm{trace}(\boldsymbol{\Sigma}\boldsymbol{\Theta})}_{f(\mathbf{x})} + \underbrace{\rho\|\mathrm{vec}(\boldsymbol{\Theta})\|_1}_{g(\mathbf{x})} \right\}$$

Log-barrier for linear/quadratic inequalities

- Poisson imaging reconstruction via TV regularization

$$x^* \in \operatorname*{argmin}_x \left\{ \underbrace{\sum_{i=1}^m a_i^T x - \sum_{i=1}^m y_i \log(a_i^T x + b_i)}_{f(x)} + g(x) \right\}$$

- Basic pursuit denoising problem (BPDP): Barrier formulation

$$\mathbf{x}_t^* = \operatorname*{argmin}_{\mathbf{x}} \left\{ \underbrace{-t\log\left(\sigma^2 - \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2\right)}_{=: f(\mathbf{x})} + g(\mathbf{x}) \right\}$$

- LASSO problem with unknown variance

$$\mathbf{x}^* \equiv (\phi^*, \gamma^*) = \operatorname*{argmin}_{\phi,\gamma} \left\{ \underbrace{-\log(\gamma) + \frac{1}{2n}\|\gamma\mathbf{y} - \mathbf{X}\phi\|_2^2}_{=: f(\mathbf{x})} + \underbrace{\lambda\|\phi\|_1}_{=: g(\mathbf{x})} \right\}$$

- Quantum tomography ML estimator (*another presentation!*)

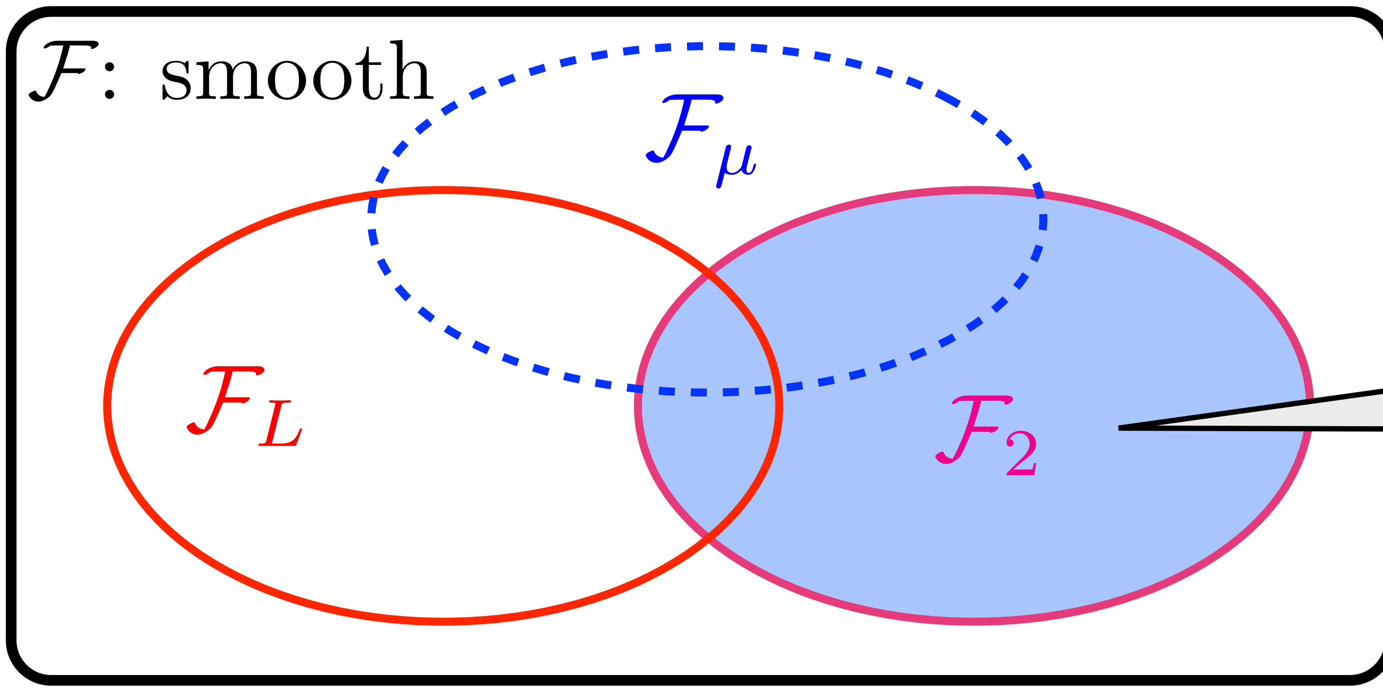
Composite self-concordant minimization

$$\text{(P)} \qquad \min_{\mathbf{x}\in\mathbb{R}^n} \left\{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\right\}$$

f
convex and **self-concordant**

g
convex and possibly **nonsmooth**
with **"tractable"** prox

Classes of smooth functions (f)



Our contributions:

i) a ***variable metric*** (path following) forward-backward framework

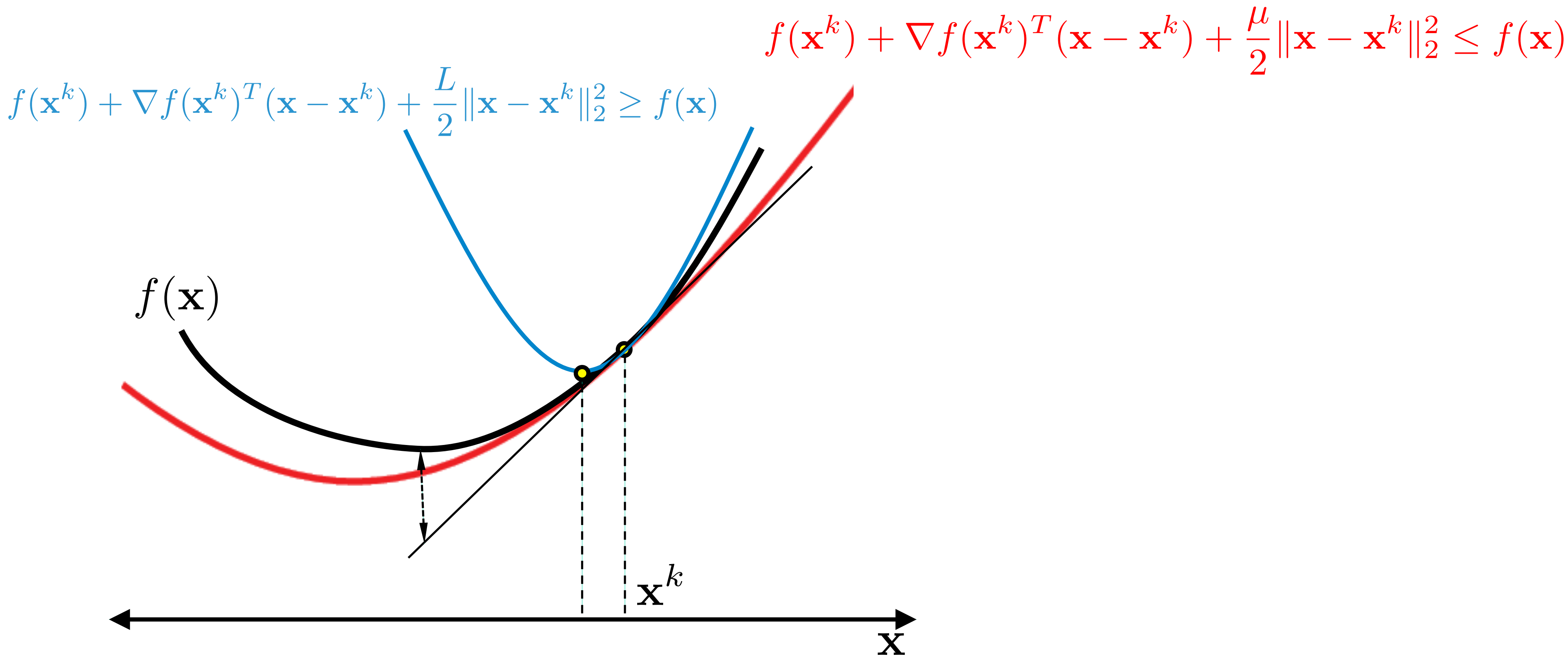
ii) convergence theory ***without the Lipschitz gradient assumption***

iii) novel variants and extensions for several applications & SCOPT

f is self-concordant if $\varphi(t) := f(\mathbf{x} + t\mathbf{d})$ satisfies $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$

Basic algorithmic framework

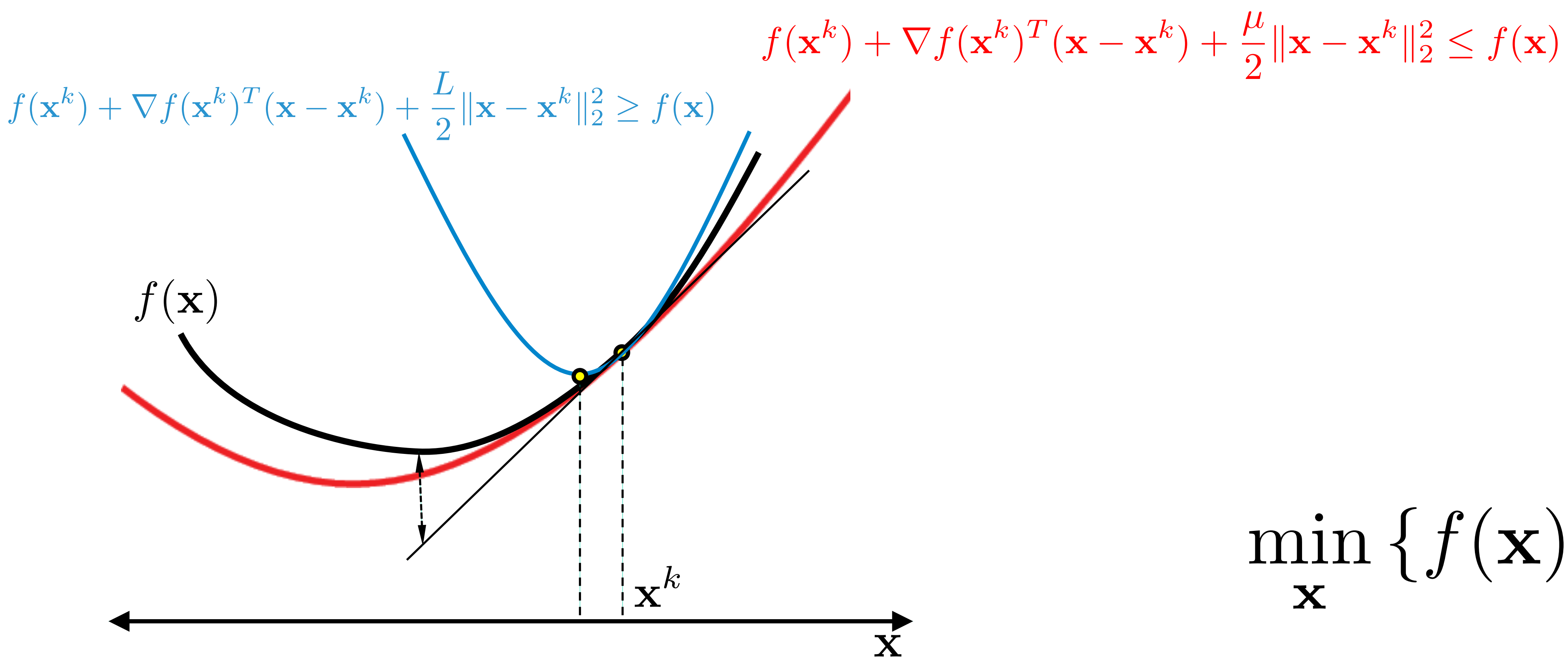
A basic composite minimization framework



- **Main properties of $\mathcal{F}_{\mu,L}$**

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	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

A basic composite minimization framework

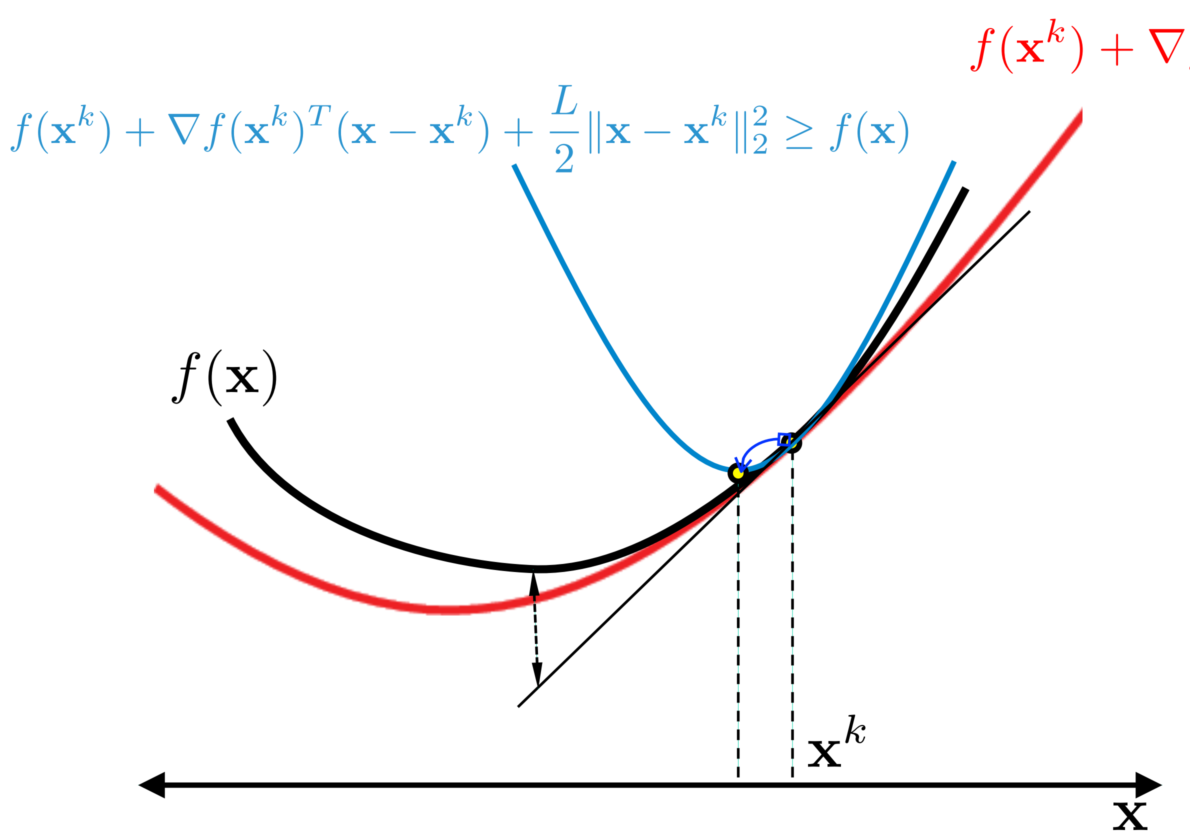


$$\min_{\mathbf{x}} \{f(\mathbf{x}) + g(\mathbf{x})\}$$

- **Main properties of $\mathcal{F}_{\mu,L}$**

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	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

A basic composite minimization framework



$$\boxed{\mathbf{x}^{k+1} = \mathrm{prox}_g\left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\right)}$$

$$\min_{\mathbf{x}} \{f(\mathbf{x}) + g(\mathbf{x})\}$$

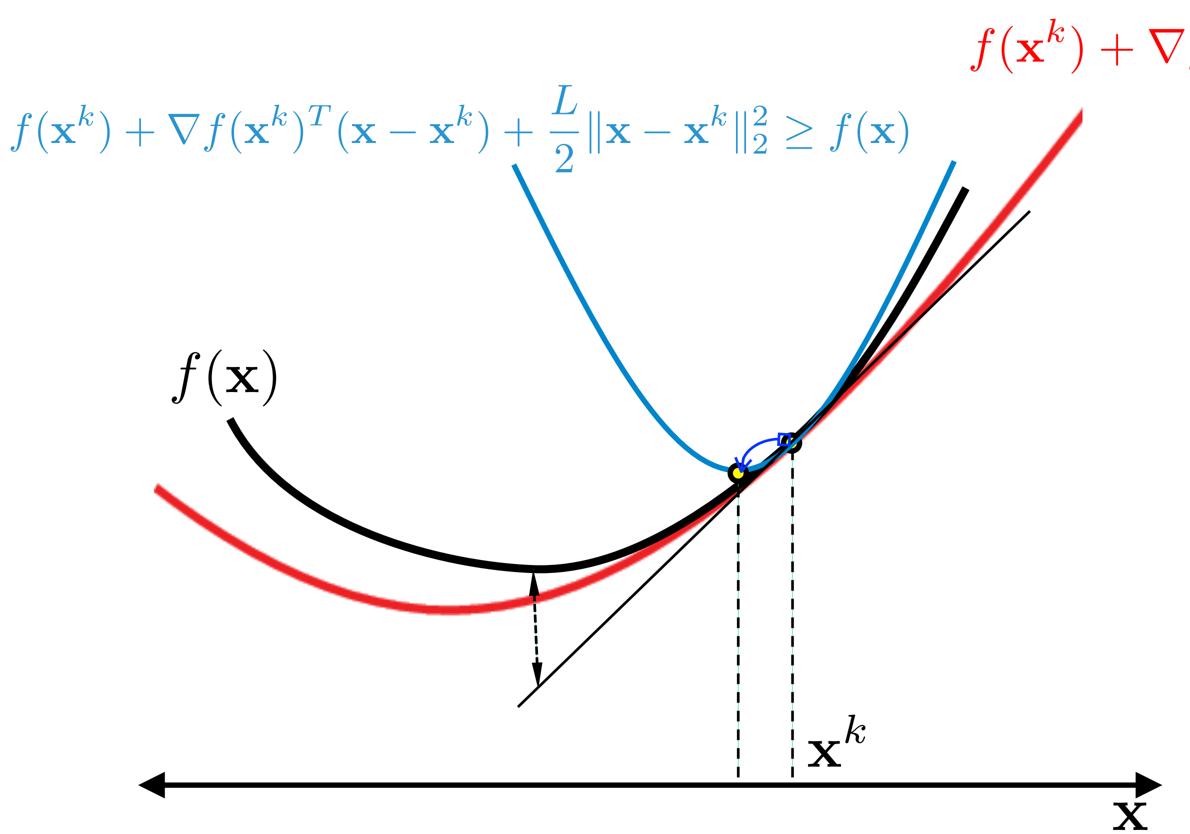
$$\mathbf{x}^{k+1} := \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2 + g(\mathbf{x}) \right\}$$

- **Main properties of $\mathcal{F}_{\mu,L}$**

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	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

$\mathcal{F}_L$

A basic composite minimization framework



ISTA

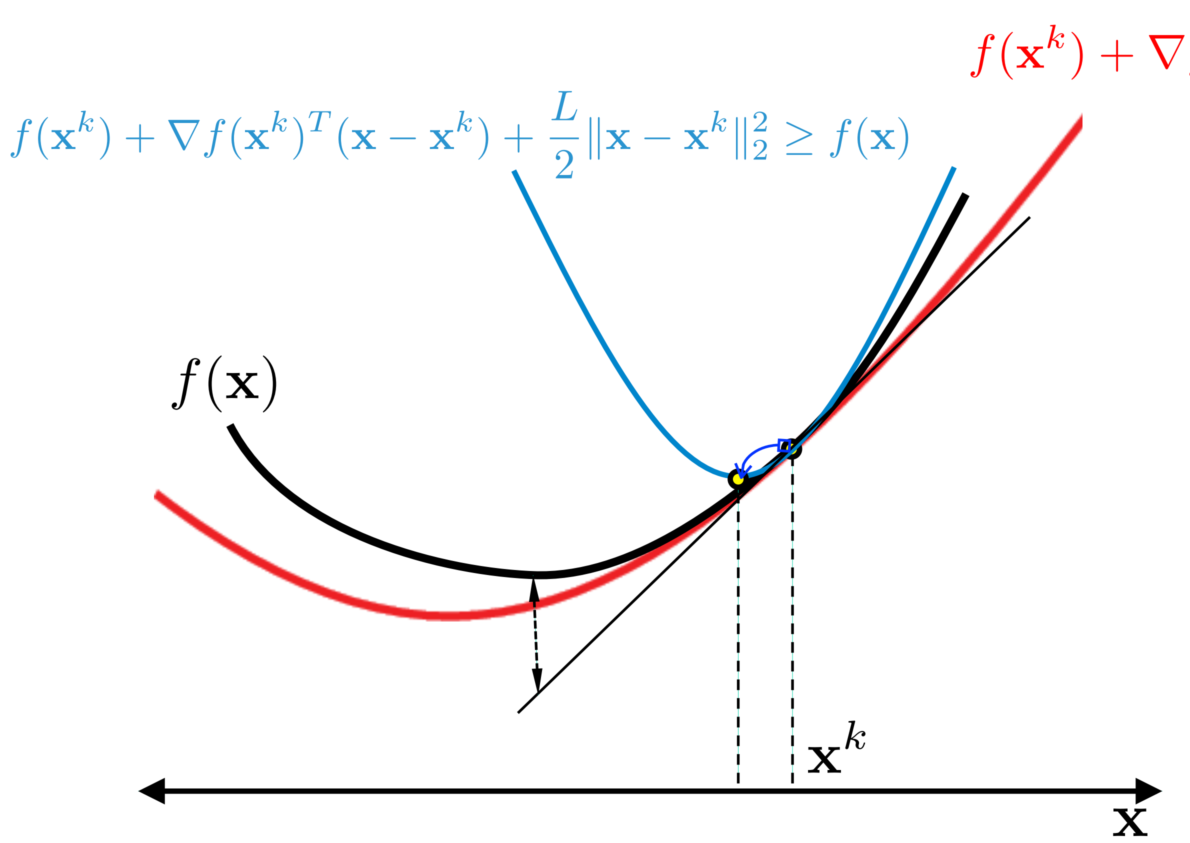
$$\mathbf{x}^{k+1} = \mathrm{prox}_g\left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\right)$$

- **Main properties of $\mathcal{F}_{\mu,L}$**

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	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

$\mathcal{F}_L$

A basic composite minimization framework



ISTA

$$\boxed{\mathbf{x}^{k+1} = \mathrm{prox}_g\left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\right)}$$

$$f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \frac{L\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2k} \Rightarrow \text{iterations} = \mathcal{O}(\epsilon^{-1})$$

- **Main properties of $\mathcal{F}_{\mu,L}$**

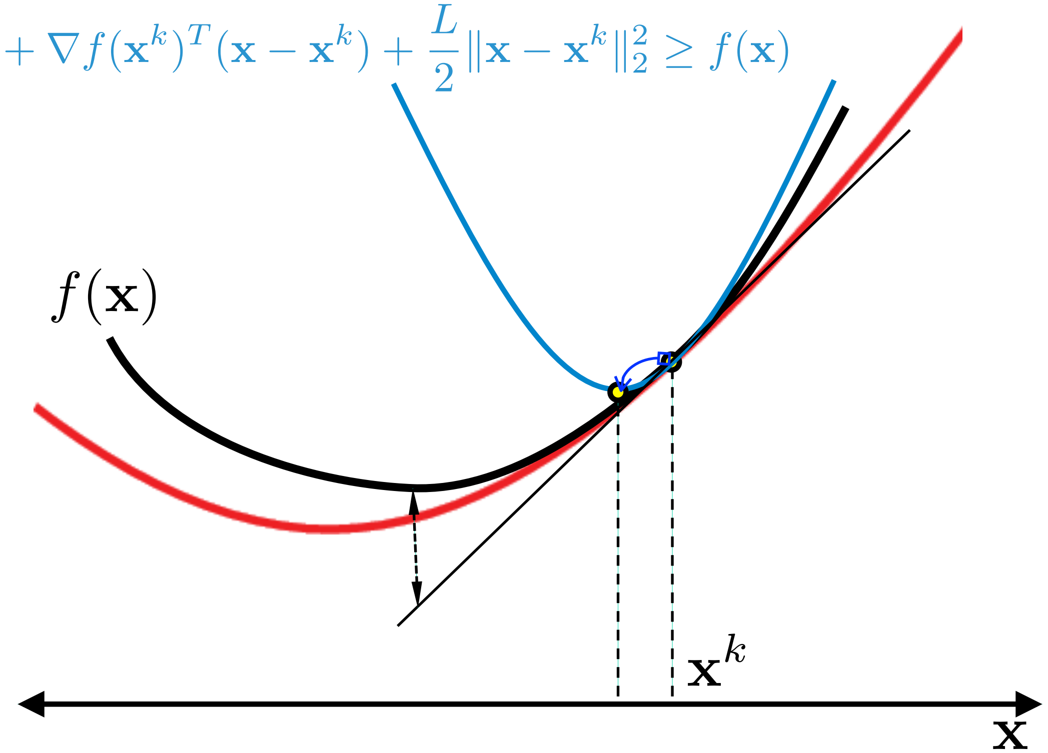
Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y}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

$\mathcal{F}_L$

A basic composite minimization framework

$$f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2 \leq f(\mathbf{x})$$

$$f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2 \geq f(\mathbf{x})$$



ISTA

$$\boxed{\mathbf{x}^{k+1} = \mathrm{prox}_g\left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\right)}$$

$$f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \frac{L\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2k} \Rightarrow \text{iterations} = \mathcal{O}(\epsilon^{-1})$$

acceleration is possible

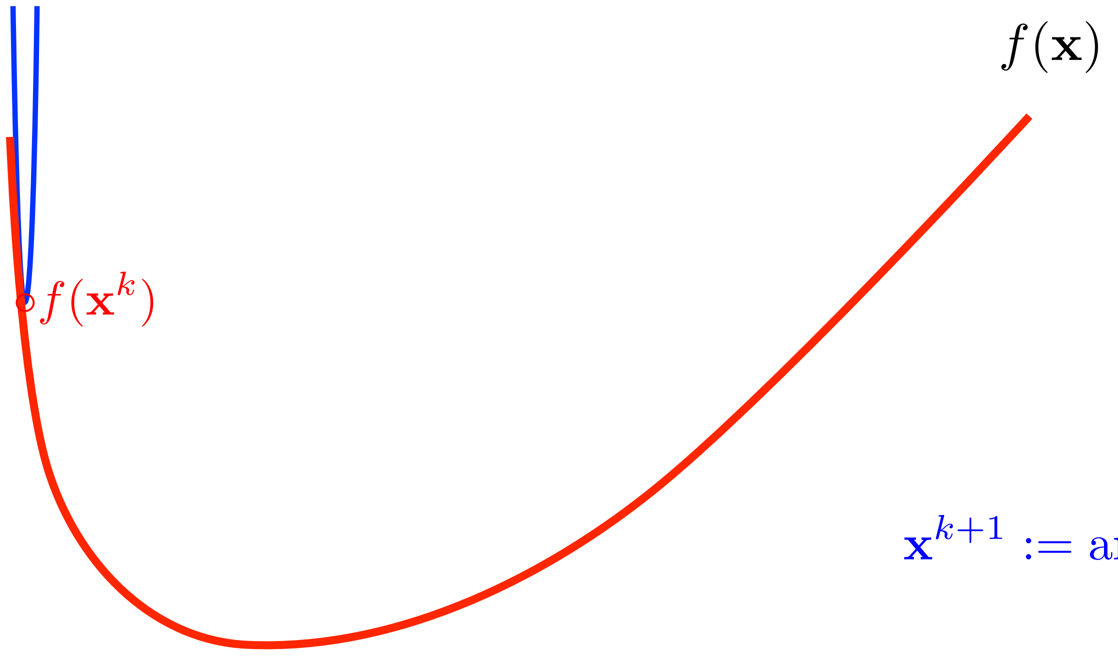
FISTA

- **Main properties of $\mathcal{F}_{\mu,L}$**

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	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

$\mathcal{F}_L$

To adapt or not to adapt?

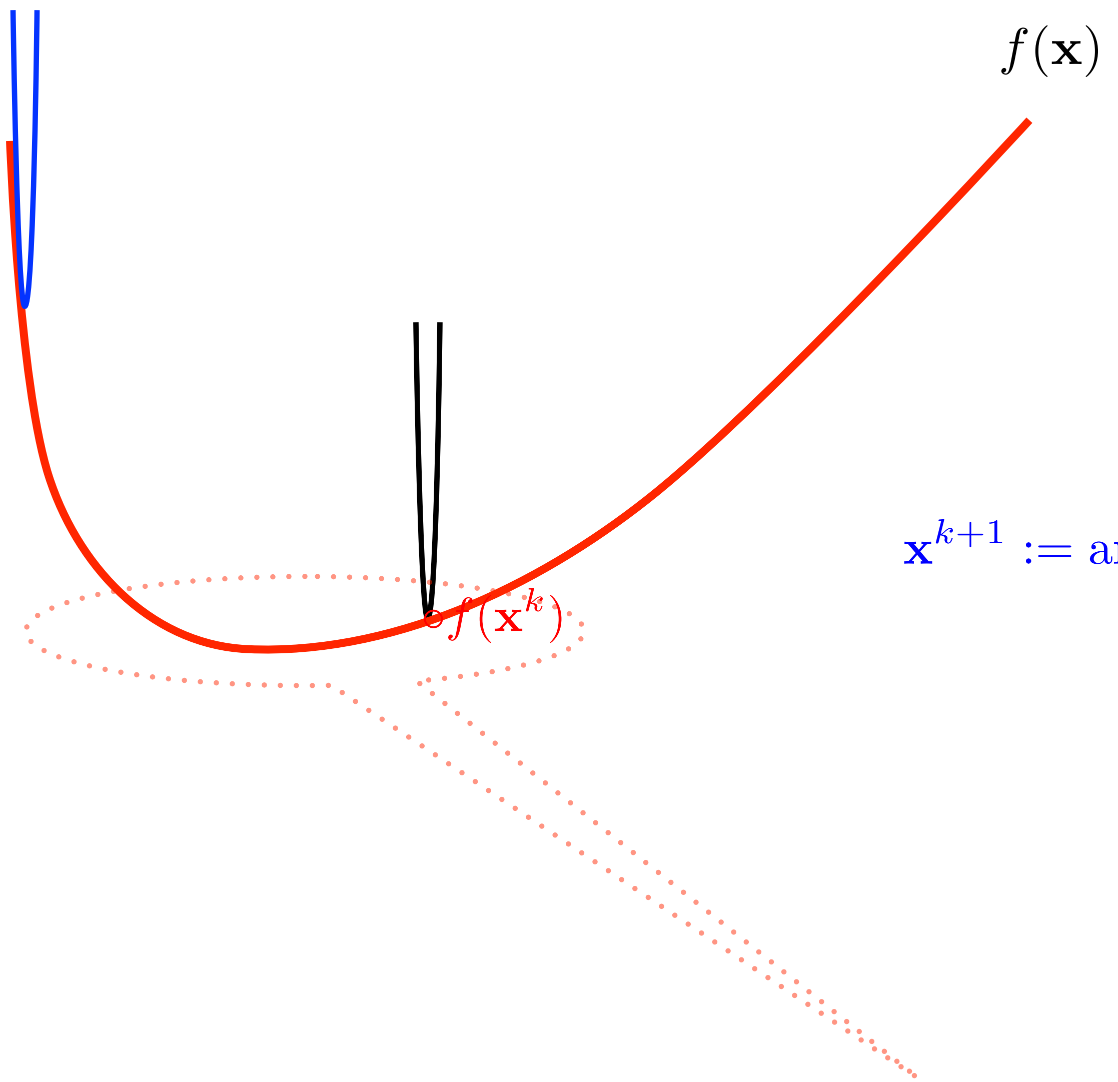


$$\mathbf{x}^{k+1} := \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2 + g(\mathbf{x}) \right\}$$

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\|$$

L is a global worst-case constant

To adapt or not to adapt?

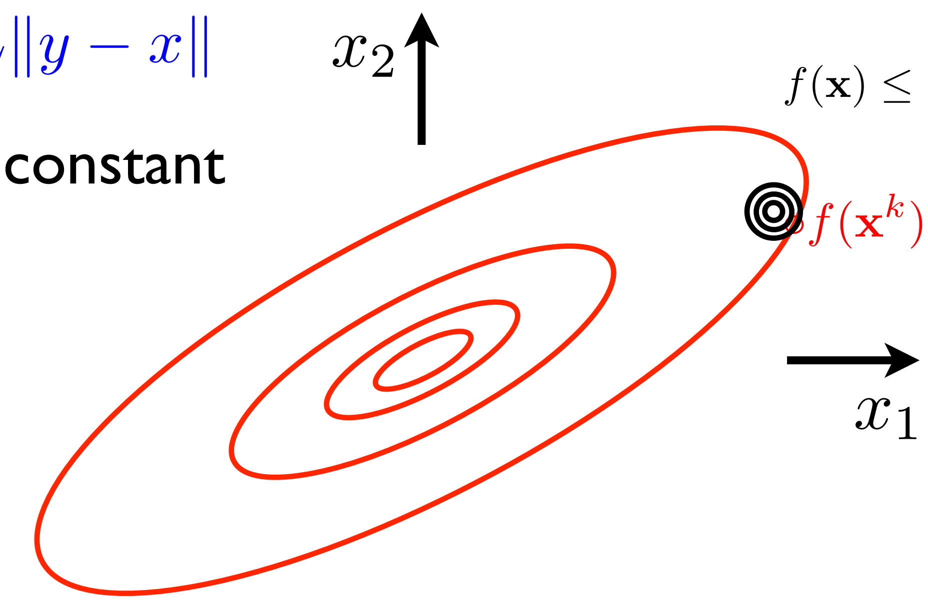


$$\mathbf{x}^{k+1} := \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2 + g(\mathbf{x}) \right\}$$

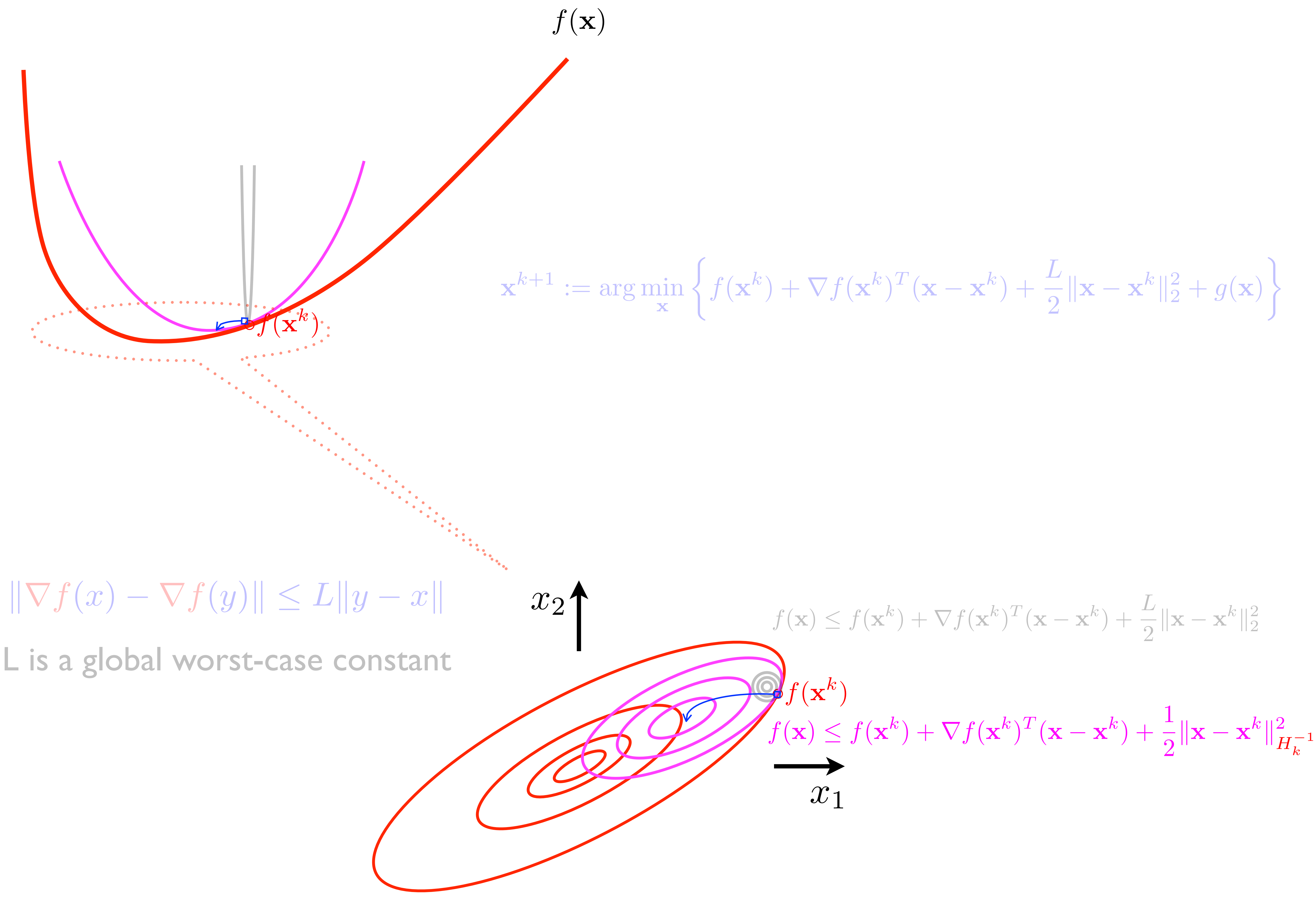
$$\|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\|$$

L is a global worst-case constant

$$f(\mathbf{x}) \leq f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2$$



To adapt or not to adapt?

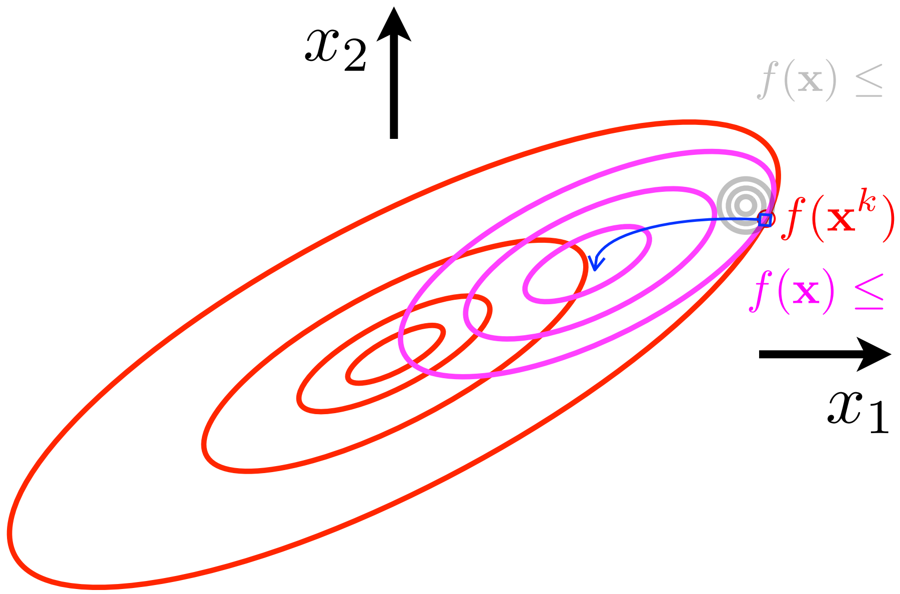


To adapt or not to adapt?

$$\mathrm{prox}_{\gamma g}(\mathbf{s}) := \arg\min_{\mathbf{x}} \left\{ g(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{s}\|_2^2 \right\}$$

Variable metric proximal point operator

$$\boxed{\mathrm{prox}_{H^{-1}g}(s) := \arg\min_{x} \left\{ g(x) + \frac{1}{2}\|x-s\|_{H^{-1}}^2 \right\}}$$



$$f(\mathbf{x}) \leq f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x}-\mathbf{x}^k) + \frac{L}{2}\|\mathbf{x}-\mathbf{x}^k\|_2^2$$

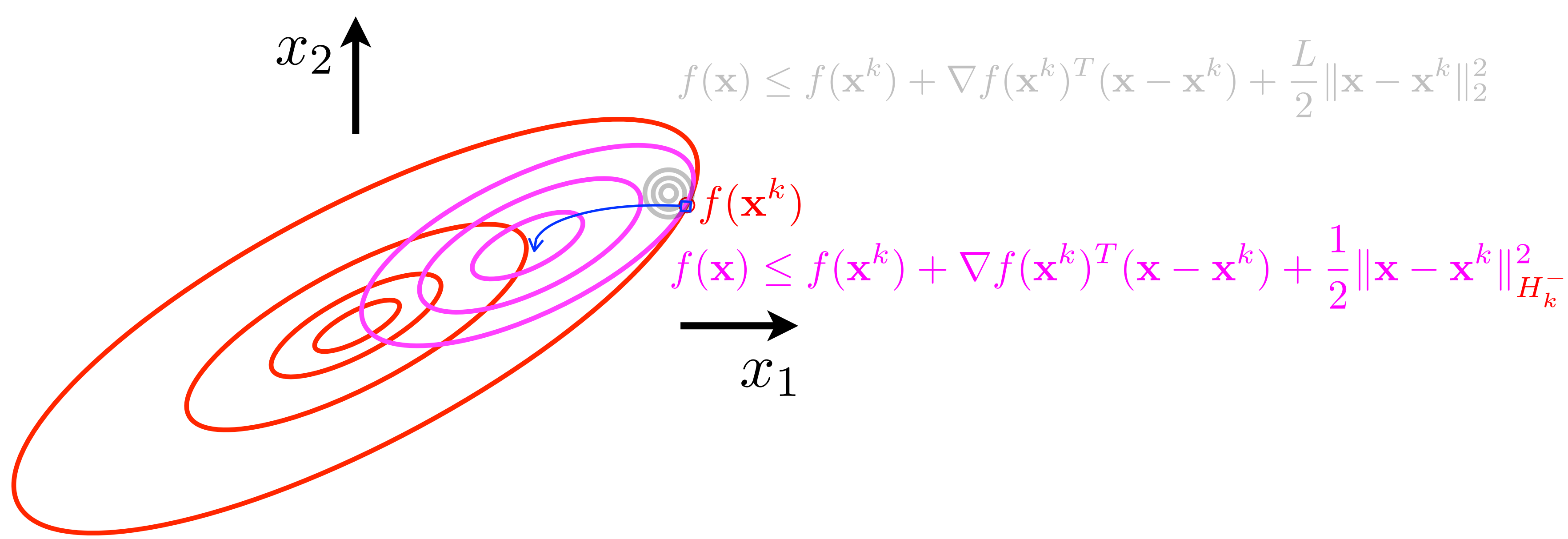
$$f(\mathbf{x}) \leq f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x}-\mathbf{x}^k) + \frac{1}{2}\|\mathbf{x}-\mathbf{x}^k\|_{H_k^{-1}}^2$$

To adapt or not to adapt?

$$\mathrm{prox}_{\gamma g}(\mathbf{s}) := \arg\min_{\mathbf{x}} \left\{ g(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{s}\|_2^2 \right\}$$

Variable metric proximal point operator

$$\boxed{\mathrm{prox}_{H^{-1}g}(s) := \arg\min_{x} \left\{ g(x) + \frac{1}{2}\|x-s\|_{H^{-1}}^2 \right\}}$$



if $g(\mathbf{x}) = \|\mathbf{x}\|_1$, then
$\mathrm{prox}_{\gamma g}(\mathbf{s}) = \mathrm{SoftThresh}(\mathbf{s}, \gamma)$
$\mathrm{prox}_{H^{-1}g}(s) = \mathrm{LASSO}$

A basic variable metric minimization framework

- Proximal point scheme with variable metric [Bonnans, 1993]

Proximal point scheme with variable metric

Given x^0, generate a sequence $\{x^k\}_{k\geq 0}$ such that

$$x^{k+1} = \mathrm{prox}_{H_k^{-1}}\left(x^k - H_k{}^{-1}\nabla f(x^k)\right)$$

where H_k is symmetric positive definite

Variable metric proximal point operator

$$\mathrm{prox}_{H^{-1}g}(s) := \arg\min_x \left\{ g(x) + \frac{1}{2}\|x - s\|^2_{H^{-1}} \right\}$$

- Additional accuracy vs. computation trade-offs

Order	Example	Components	k
1-st	Accelerated gradient	$\nabla f, \mathrm{prox}_{1/L\mathbf{I}_n}$	$\mathcal{O}(\epsilon^{-1/2})$
1^+-th	BFGS	$H_k, \nabla f, \mathrm{prox}_{H_k^{-1}}$	$\mathcal{O}(\log \epsilon^{-1})$ or faster
2-nd	Proximal Newton, IPM	$\nabla^2 f, \nabla f, \mathrm{prox}_{\nabla^2 f^{-1}}$	$\mathcal{O}(\log\log \epsilon^{-1})$

Self-concordance vs. Lipschitz gradient + SC

- **Main properties of $\mathcal{F}_{\mu,L}$**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\lVert\mathbf{y}-\mathbf{x}\rVert_2^2$	$\mathbf{x},\mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\lVert\mathbf{y}-\mathbf{x}\rVert_2^2$	$\mathbf{x},\mathbf{y} \in \mathrm{dom}(f)$
Hessian surrogates	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

Self-concordance vs. Lipschitz gradient + SC

- **Main properties of $\mathcal{F}_{\mu,L}$**

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\lVert\mathbf{y}-\mathbf{x}\rVert_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\lVert\mathbf{y}-\mathbf{x}\rVert_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Hessian surrogates	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

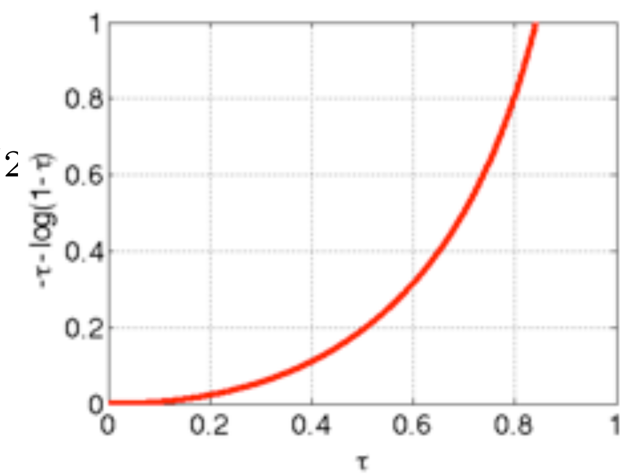
Global

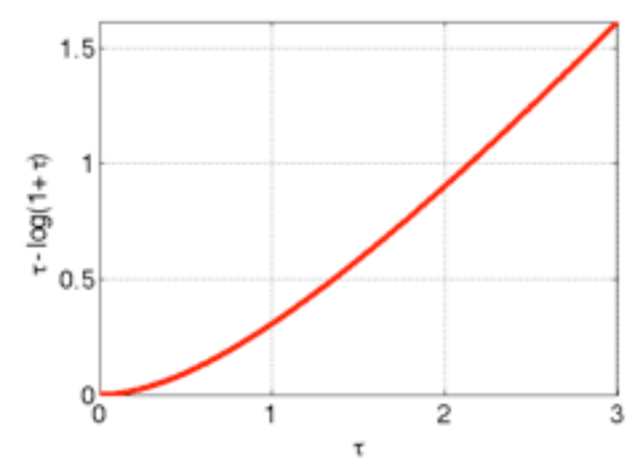
Self-concordance vs. Lipschitz gradient + SC

- **Main properties of $\mathcal{F}_2$**

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	$(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2\nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq (1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^{-2}\nabla^2 f(\mathbf{x})$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$

Utility functions: $\omega_*(\tau) = -\tau - \ln(1-\tau),\ \tau \in [0,1)$ $\qquad \omega(\tau) = \tau - \ln(1+\tau),\ \tau \geq 0$





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

$$\frac{\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2}{1+\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}} \leq (\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^T(\mathbf{y}-\mathbf{x}) \leq \underbrace{\frac{\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2}{1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}}}_{\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}<1}, \quad \forall \mathbf{x},\mathbf{y} \in \mathrm{dom}(f)$$

f is self-concordant if $\varphi(t) := f(\mathbf{x}+t\mathbf{d})$ satisfies $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$

Self-concordance vs. Lipschitz gradient + SC

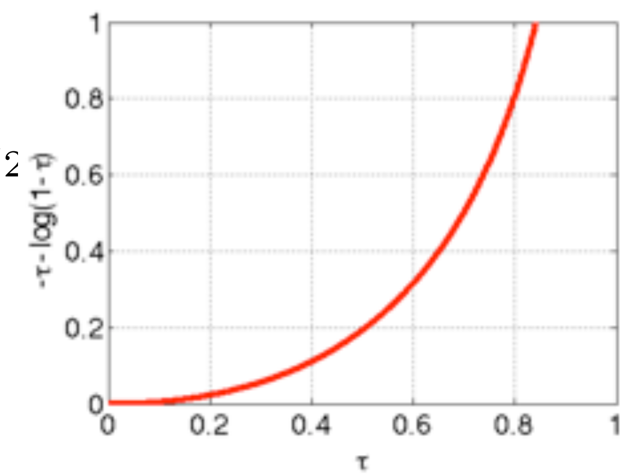
- **Main properties of $\mathcal{F}_2$**

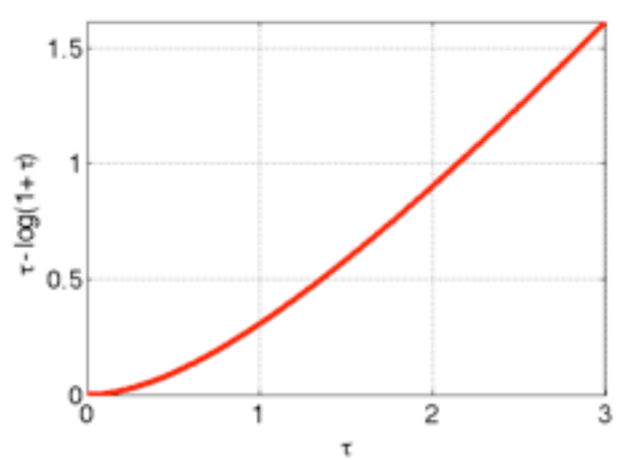
Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x})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

Utility functions: $\omega_*(\tau) = -\tau - \ln(1-\tau),\ \tau \in [0,1)$ $\qquad \omega(\tau) = \tau - \ln(1+\tau),\ \tau \geq 0$

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





$$\frac{\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2}{1+\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}} \leq (\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^T(\mathbf{y}-\mathbf{x}) \leq \underbrace{\frac{\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2}{1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}}}_{\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}<1}, \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$$

f is self-concordant if $\varphi(t) := f(\mathbf{x} + t\mathbf{d})$ satisfies $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$

Self-concordance: A mathematical tool

- **Main properties of $\mathcal{F}_2$**

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	$(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2\nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq (1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^{-2}\nabla^2 f(\mathbf{x})$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$

- New variable metric framework with rigorous convergence guarantees

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\right\}$$

Includes several algorithms: Newton, quasi-Newton, and gradient methods...

f is self-concordant if $\varphi(t) := f(\mathbf{x}+t\mathbf{d})$ satisfies $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$

Our composite self-concordant minimization framework

- Proximal Newton scheme $\mathbf{H}_k = \nabla^2 f(\mathbf{x}^k)$

Given $\mathbf{x}^0$, generate a sequence $\{\mathbf{x}^k\}_{k\geq 0}$ such that

$$\boxed{\mathbf{x}^{k+1} := \mathbf{x}^k + \alpha_k \mathbf{d}^k_{\mathbf{H}_k}}$$

where $\alpha_k \in (0, 1]$ is step-size, $\mathbf{d}^k_{\mathbf{H}_k}$ is a search direction

Our composite self-concordant minimization framework

- Proximal Newton scheme $\mathbf{H}_k = \nabla^2 f(\mathbf{x}^k)$

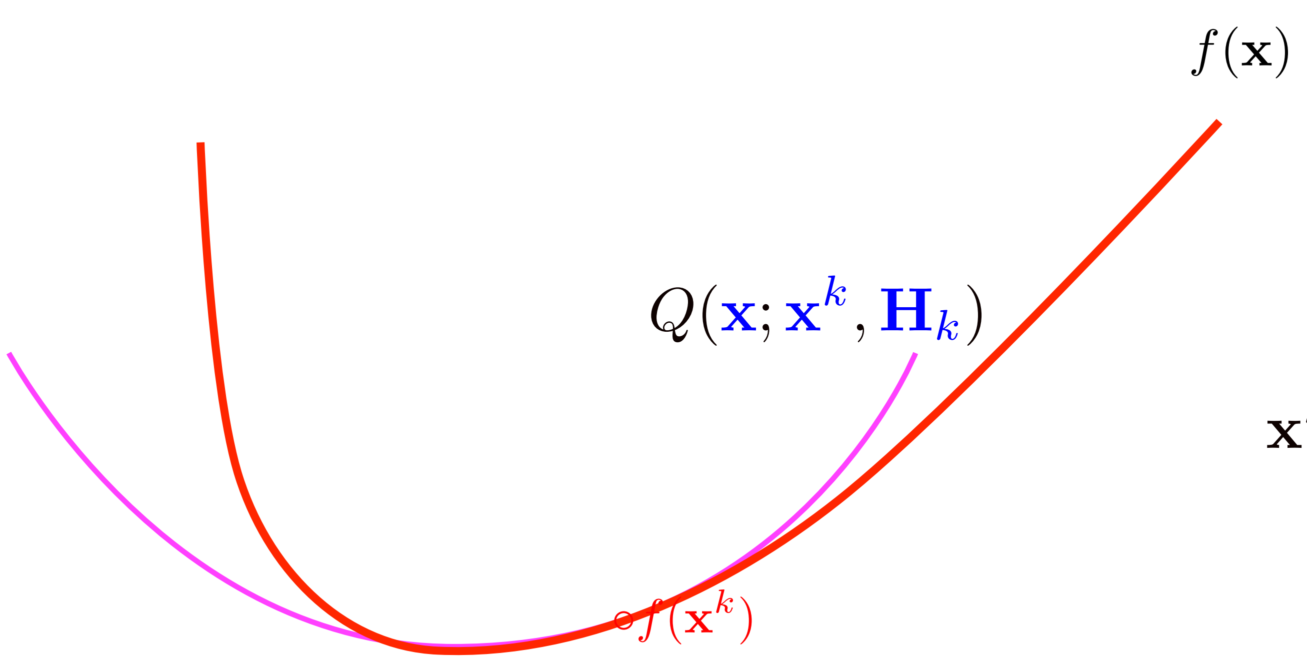
Given $\mathbf{x}^0$, generate a sequence $\{\mathbf{x}^k\}_{k\geq 0}$ such that

$$\boxed{\mathbf{x}^{k+1} := \mathbf{x}^k + \alpha_k \mathbf{d}^k_{\mathbf{H}_k}}$$

where $\alpha_k \in (0,1]$ is step-size, $\mathbf{d}^k_{\mathbf{H}_k}$ is a search direction

- How to compute the Proximal Newton direction?

$$\mathbf{d}^k_{\mathbf{H}_k} := \arg\min_{\mathbf{d}} \left\{ f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T \mathbf{d} + \frac{1}{2}\mathbf{d}^T \mathbf{H}_k \mathbf{d} + g(\mathbf{x}^k + \mathbf{d}) \right\}, \quad \mathbf{H}_k = \nabla^2 f(\mathbf{x}^k)$$



$$\mathbf{s}^k_g := \arg\min_{\mathbf{x}\in\mathrm{dom}(F)} \left\{ Q(\mathbf{x}; \mathbf{x}^k, \mathbf{H}_k) + g(\mathbf{x}) \right\}$$

$$\mathbf{x}^{k+1} := \mathbf{x}^k + \alpha_k \left(\mathbf{s}^k_g - \mathbf{x}^k \right)$$

Our composite self-concordant minimization framework

- Proximal Newton scheme $\mathbf{H}_k = \nabla^2 f(\mathbf{x}^k)$

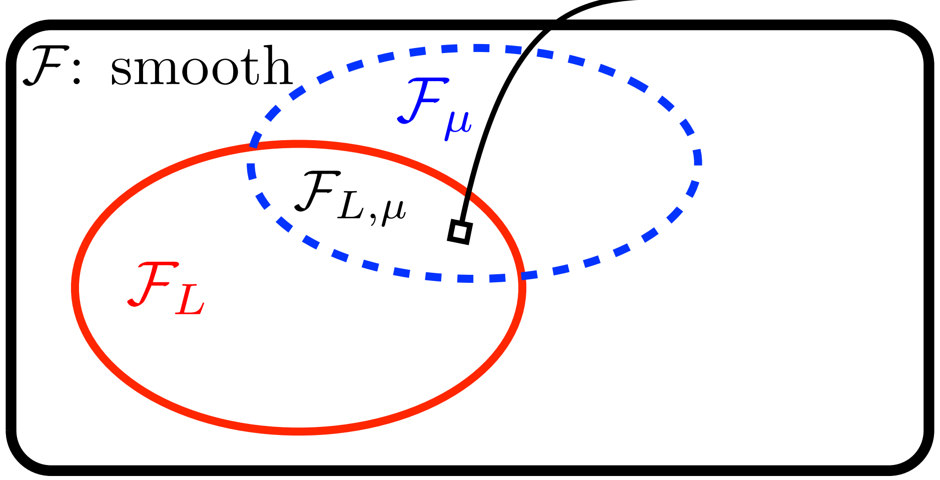
Given $\mathbf{x}^0$, generate a sequence $\{\mathbf{x}^k\}_{k\geq 0}$ such that

$$\boxed{\mathbf{x}^{k+1} := \mathbf{x}^k + \alpha_k \mathbf{d}^k_{\mathbf{H}_k}}$$

where $\alpha_k \in (0, 1]$ is step-size, $\mathbf{d}^k_{\mathbf{H}_k}$ is a search direction

- How to compute the Proximal Newton direction?

$$\mathbf{d}^k_{\mathbf{H}_k} := \arg\min_{\mathbf{d}} \left\{ f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T \mathbf{d} + \frac{1}{2}\mathbf{d}^T \mathbf{H}_k \mathbf{d} + g(\mathbf{x}^k + \mathbf{d}) \right\}, \quad \mathbf{H}_k = \nabla^2 f(\mathbf{x}^k)$$



$$\|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\|$$

$$\mu\mathbb{I} \preceq \nabla^2 f(x) \preceq L\mathbb{I}$$

Fast gradient schemes (Nesterov's methods)

Newton/quasi Newton schemes

Our composite self-concordant minimization framework

- Proximal Newton scheme

Key contribution:
step size selection procedure

$\mathbf{H}_k = \nabla^2 f(\mathbf{x}^k)$

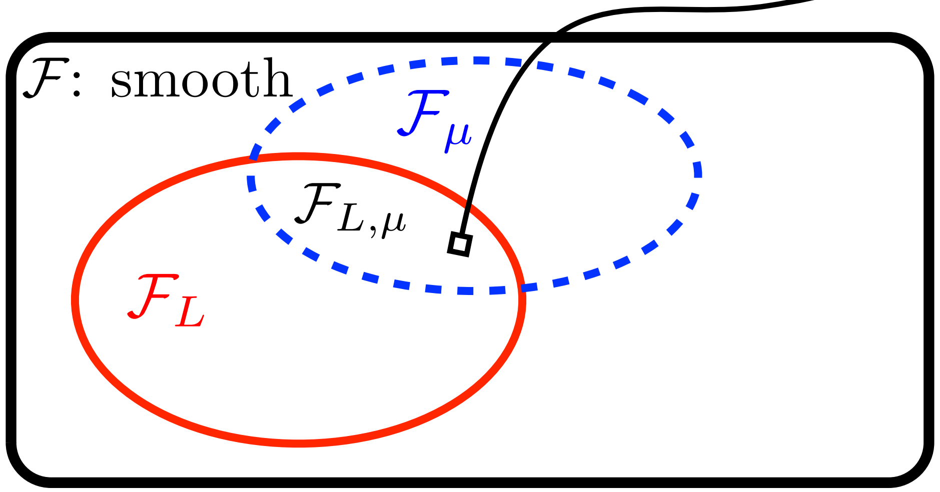
Given $\mathbf{x}^0$, generate a sequence $\{\mathbf{x}^k\}_{k\geq 0}$ such that

$$\boxed{\mathbf{x}^{k+1} := \mathbf{x}^k + \alpha_k \mathbf{d}^k_{\mathbf{H}_k}}$$

where $\alpha_k \in (0, 1]$ is step-size, $\mathbf{d}^k_{\mathbf{H}_k}$ is a search direction

- How to compute the Proximal Newton direction?

$$\mathbf{d}^k_{\mathbf{H}_k} := \arg\min_{\mathbf{d}} \left\{ f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T \mathbf{d} + \frac{1}{2}\mathbf{d}^T \mathbf{H}_k \mathbf{d} + g(\mathbf{x}^k + \mathbf{d}) \right\}, \quad \mathbf{H}_k = \nabla^2 f(\mathbf{x}^k)$$



$$\|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\|$$

$$\mu\mathbb{I} \preceq \nabla^2 f(x) \preceq L\mathbb{I}$$

Fast gradient schemes (Nesterov's methods)

Newton/quasi Newton schemes

How do we compute the step-size?

- Upper surrogate of f

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x}^{k+1} - \mathbf{x}^k) + \omega^*(\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_{\mathbf{x}^k}), \quad \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_{\mathbf{x}^k} < 1$$

- Convexity of g and optimality condition of the subproblem

$$g(\mathbf{x}^{k+1}) - g(\mathbf{x}^k) \leq -\alpha_k \nabla f(\mathbf{x}^k)^T \mathbf{d}^k_{\mathbf{H_k}} - \alpha_k \|\mathbf{d}^k_{\mathbf{H_k}}\|^2_{\mathbf{H}_k}.$$

How do we compute the step-size?

- Upper surrogate of f

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x}^{k+1} - \mathbf{x}^k) + \omega^*(\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_{\mathbf{x}^k}), \quad \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_{\mathbf{x}^k} < 1$$

- Convexity of g and optimality condition of the subproblem

$$g(\mathbf{x}^{k+1}) - g(\mathbf{x}^k) \leq -\alpha_k \nabla f(\mathbf{x}^k)^T \mathbf{d}^k_{\mathbf{H_k}} - \alpha_k \|\mathbf{d}^k_{\mathbf{H_k}}\|^2_{\mathbf{H}_k}.$$

$$\boxed{\phi(\mathbf{x}^{k+1}) \leq \phi(\mathbf{x}^k) - \alpha_k \|\mathbf{d}^k_{\mathbf{H}_k}\|^2_{\mathbf{H}_k} + \omega_*\left(\alpha_k \|\mathbf{d}^k_{\mathbf{H}_k}\|_{\mathbf{x}^k}\right)}$$

How do we compute the step-size?

- Upper surrogate of f

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x}^{k+1} - \mathbf{x}^k) + \omega^*(\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_{\mathbf{x}^k}), \quad \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_{\mathbf{x}^k} < 1$$

- Convexity of g and optimality condition of the subproblem

$$g(\mathbf{x}^{k+1}) - g(\mathbf{x}^k) \leq -\alpha_k \nabla f(\mathbf{x}^k)^T \mathbf{d}^k_{\mathbf{H_k}} - \alpha_k \|\mathbf{d}^k_{\mathbf{H_k}}\|^2_{\mathbf{H}_k}.$$

$$\boxed{\phi(\mathbf{x}^{k+1}) \leq \phi(\mathbf{x}^k) - \alpha_k \|\mathbf{d}^k_{\mathbf{H}_k}\|^2_{\mathbf{H}_k} + \omega_* \left(\alpha_k \|\mathbf{d}^k_{\mathbf{H}_k}\|_{\mathbf{x}^k}\right)}$$

- When $\mathbf{H}_k \equiv \nabla^2 f(\mathbf{x}^k)$, $\lambda_k := \|\mathbf{d}^k_{\mathbf{H}_k}\|_{\mathbf{x}^k}$

$$\boxed{\phi(\mathbf{x}^{k+1}) \leq \phi(\mathbf{x}^k) - \underbrace{[\alpha_k \lambda_k - \omega_*(\alpha_k \lambda_k)]}_{\psi(\alpha_k)}}$$

maximize $\psi(\alpha_k)$ to get optimal α_k^*

$$\boxed{\alpha_k^* = \frac{1}{\lambda_k + 1} \in (0, 1]}$$

Analytic complexity

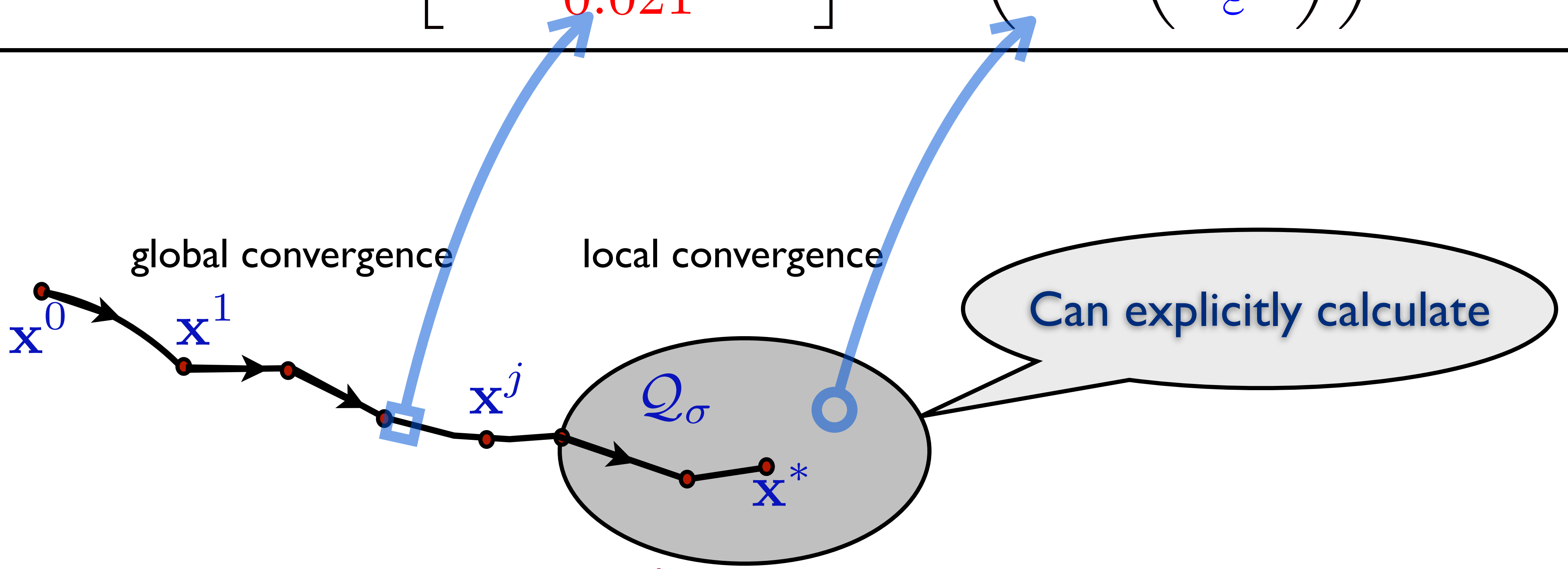
- Worst-case complexity to obtain an ε-approximate solution

$$\#\text{iterations} = \left\lfloor \frac{\phi(\mathbf{x}^0) - \phi(\mathbf{x}^*)}{0.021} \right\rfloor + O\left(\ln\ln\left(\frac{4.56}{\varepsilon}\right)\right)$$

Analytic complexity

- Worst-case complexity to obtain an ε-approximate solution

$$\#\text{iterations} = \left\lfloor \frac{\phi(\mathbf{x}^0) - \phi(\mathbf{x}^*)}{0.021} \right\rfloor + O\left(\ln\ln\left(\frac{4.56}{\varepsilon}\right)\right)$$

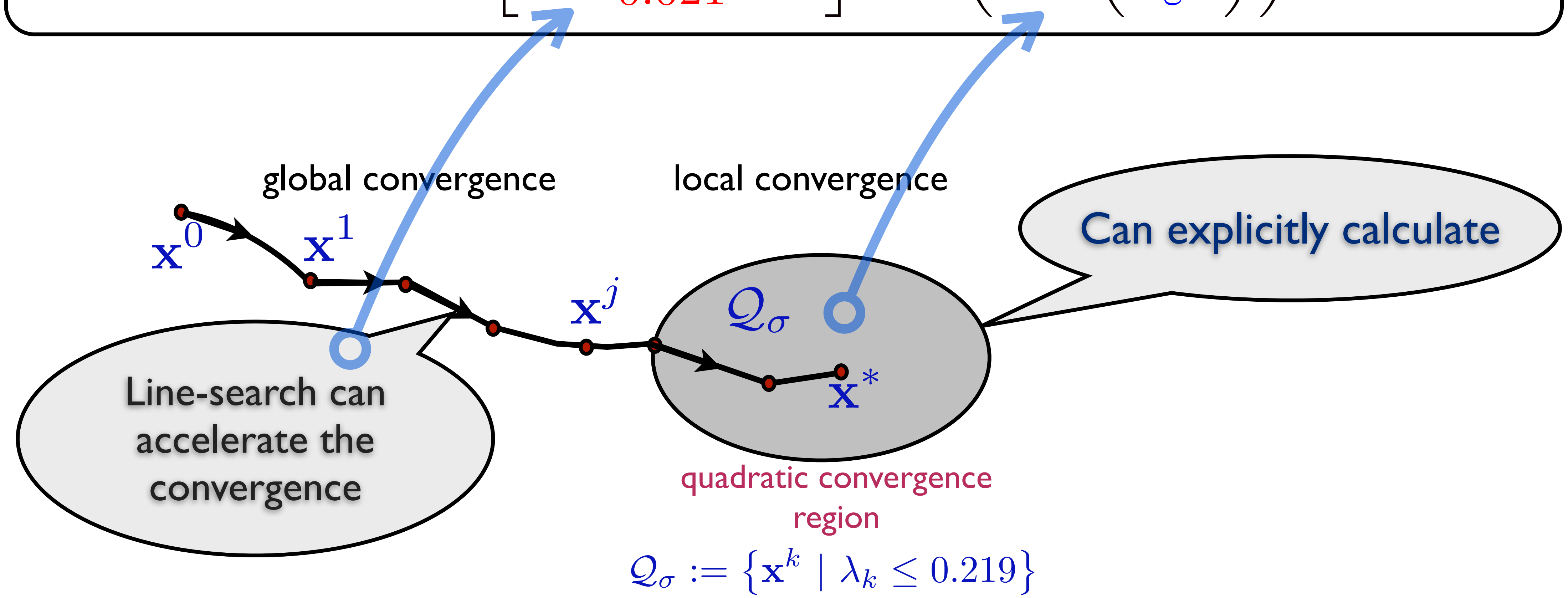


$$\mathcal{Q}_\sigma := \{\mathbf{x}^k \mid \lambda_k \le 0.219\}$$

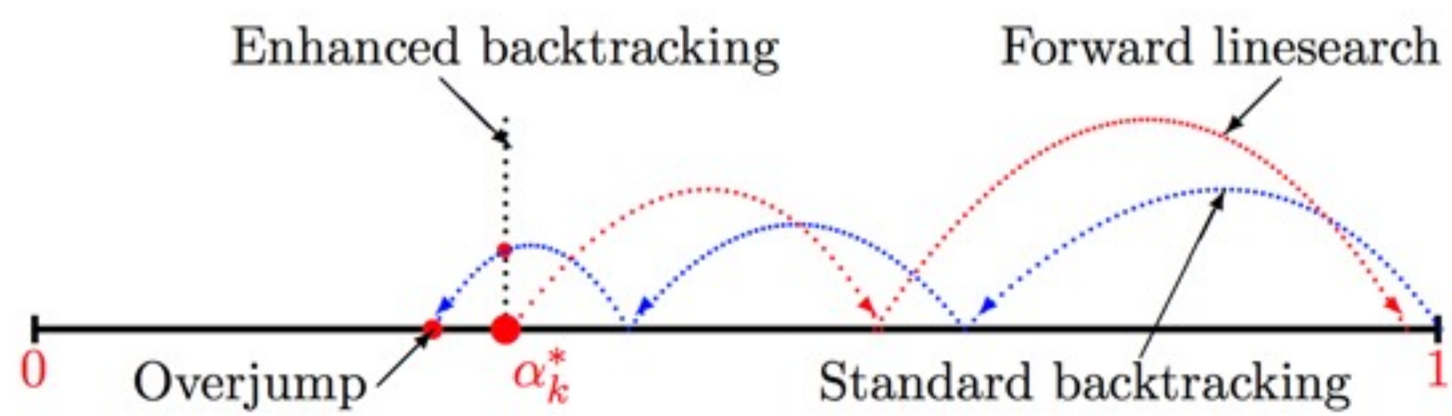
Analytic complexity

- Worst-case complexity to obtain an ε-approximate solution

$$\#\text{iterations} = \left\lfloor \frac{\phi(\mathbf{x}^0) - \phi(\mathbf{x}^*)}{0.021} \right\rfloor + O\left(\ln\ln\left(\frac{4.56}{\varepsilon}\right)\right)$$



- Line-search enhancement



Graphical model selection

- Objective:

$$\min_{\boldsymbol{\Theta} \succ 0} \left\{ \underbrace{-\log\det(\boldsymbol{\Theta}) + \mathrm{trace}(\boldsymbol{\Sigma\Theta})}_{f(\mathbf{x})} + \underbrace{\rho \|\mathrm{vec}(\boldsymbol{\Theta})\|_1}_{g(\mathbf{x})} \right\}$$

Graphical model selection

- Objective:

$$\min_{\boldsymbol{\Theta} \succ 0} \left\{ \underbrace{-\log\det(\boldsymbol{\Theta}) + \mathrm{trace}(\boldsymbol{\Sigma}\boldsymbol{\Theta})}_{f(\mathbf{x})} + \underbrace{\rho \|\mathrm{vec}(\boldsymbol{\Theta})\|_1}_{g(\mathbf{x})} \right\}$$

- Gradient and Hessian (large-scale, special structure)
  - Gradient of f: $\nabla f(\mathbf{x}) = \mathbf{vec}(\boldsymbol{\Sigma} - \boldsymbol{\Theta}^{-1})$.
  - Hessian of f: $\nabla^2 f(\mathbf{x}) = \boldsymbol{\Theta}^{-1} \otimes \boldsymbol{\Theta}^{-1}$

Graphical model selection

- Objective:

$$\boxed{\min_{\boldsymbol{\Theta} \succ 0} \left\{ \underbrace{-\log\det(\boldsymbol{\Theta}) + \mathrm{trace}(\boldsymbol{\Sigma}\boldsymbol{\Theta})}_{f(\mathbf{x})} + \underbrace{\rho \|\mathrm{vec}(\boldsymbol{\Theta})\|_1}_{g(\mathbf{x})} \right\}}$$

- Gradient and Hessian (large-scale, special structure)
  - Gradient of f: $\nabla f(\mathbf{x}) = \mathbf{vec}(\boldsymbol{\Sigma} - \boldsymbol{\Theta}^{-1})$.
  - Hessian of f: $\nabla^2 f(\mathbf{x}) = \boldsymbol{\Theta}^{-1} \otimes \boldsymbol{\Theta}^{-1}$

- How to compute the Proximal Newton direction?

$$\mathbf{d}^k_{\mathbf{H}_k} := \arg\min_{\mathbf{d}} \left\{ f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T \mathbf{d} + \frac{1}{2}\mathbf{d}^T \mathbf{H}_k \mathbf{d} + g(\mathbf{x}^k + \mathbf{d}) \right\}, \quad \mathbf{H}_k = \nabla^2 f(\mathbf{x}^k)$$

Graphical model selection

- Objective:

$$\min_{\Theta \succ 0} \left\{ \underbrace{-\log\det(\Theta) + \mathrm{trace}(\Sigma\Theta)}_{f(\mathbf{x})} + \underbrace{\rho\|\mathrm{vec}(\Theta)\|_1}_{g(\mathbf{x})} \right\}$$

- Gradient and Hessian (large-scale, special structure)
  - Gradient of f: $\nabla f(\mathbf{x}) = \mathbf{vec}(\Sigma - \Theta^{-1})$.
  - Hessian of f: $\nabla^2 f(\mathbf{x}) = \Theta^{-1} \otimes \Theta^{-1}$
- Dual approach for solving subproblem (SP)

Primal subproblem	Dual subproblem (SPGL)
$\min_{\Delta} \left\{ \frac{1}{2}\mathrm{trace}((\Theta_i^{-1}\Delta)^2) + \mathrm{trace}(\mathbf{R}_i\Delta) + \rho\|\mathrm{vec}(\Delta)\|_1 \right\}$	$\min_{\|\mathrm{vec}(\mathbf{U})\|_\infty \leq 1} \left\{ \frac{1}{2}\mathrm{trace}((\Theta_i\mathbf{U})^2) + \mathrm{trace}(\mathbf{Q}_i\mathbf{U}) \right\}$
$\mathbf{R}_i := \Sigma - 2\Theta_i^{-1}$	$\mathbf{Q}_i := \rho^{-1}[\Theta_i\Sigma\Theta_i - 2\Theta_i]$

Unconstrained LASSO problem

Graphical model selection

- Objective:

$$\min_{\Theta \succ 0} \left\{ \underbrace{-\log\det(\Theta) + \mathrm{trace}(\Sigma\Theta)}_{f(\mathbf{x})} + \underbrace{\rho\|\mathrm{vec}(\Theta)\|_1}_{g(\mathbf{x})} \right\}$$

- Gradient and Hessian (large-scale, special structure)
  - Gradient of f: $\nabla f(\mathbf{x}) = \mathbf{vec}(\Sigma - \Theta^{-1})$.
  - Hessian of f: $\nabla^2 f(\mathbf{x}) = \Theta^{-1} \otimes \Theta^{-1}$

No Cholesky decomposition and matrix inversion

- Dual approach for solving subproblem (SP)

Primal subproblem	Dual subproblem (SPGL)
$\min_{\Delta} \left\{ \frac{1}{2}\mathrm{trace}((\Theta_i^{-1}\Delta)^2) + \mathrm{trace}(\mathbf{R}_i\Delta) + \rho\|\mathrm{vec}(\Delta)\|_1 \right\}$	$\min_{\|\mathrm{vec}(\mathbf{U})\|_\infty \le 1} \left\{ \frac{1}{2}\mathrm{trace}((\Theta_i\mathbf{U})^2) + \mathrm{trace}(\mathbf{Q}_i\mathbf{U}) \right\}$
$\mathbf{R}_i := \Sigma - 2\Theta_i^{-1}$	$\mathbf{Q}_i := \rho^{-1}[\Theta_i\Sigma\Theta_i - 2\Theta_i]$

Unconstrained LASSO problem

Graphical model selection

- Objective:

$$\min_{\boldsymbol{\Theta} \succ 0} \left\{ \underbrace{-\log\det(\boldsymbol{\Theta}) + \mathrm{trace}(\boldsymbol{\Sigma}\boldsymbol{\Theta})}_{f(\mathbf{x})} + \underbrace{\rho\|\mathrm{vec}(\boldsymbol{\Theta})\|_1}_{g(\mathbf{x})} \right\}$$

- Gradient and Hessian (large-scale, special structure)
  - Gradient of f: $\nabla f(\mathbf{x}) = \mathbf{vec}(\boldsymbol{\Sigma} - \boldsymbol{\Theta}^{-1})$.
  - Hessian of f: $\nabla^2 f(\mathbf{x}) = \boldsymbol{\Theta}^{-1} \otimes \boldsymbol{\Theta}^{-1}$

No Cholesky decomposition and matrix inversion

- Dual approach for solving subproblem (SP)

Primal subproblem	Dual subproblem (SPGL)
Unconstrained LASSO problem	$\min_{\|\mathrm{vec}(\mathbf{U})\|_\infty \le 1} \left\{ \frac{1}{2}\mathrm{trace}((\boldsymbol{\Theta}_i\mathbf{U})^2) + \mathrm{trace}(\mathbf{Q}_i\mathbf{U}) \right\}$
$\mathbf{R}_i := \boldsymbol{\Sigma} - 2\boldsymbol{\Theta}_i^{-1}$	$\mathbf{Q}_i := \rho^{-1}[\boldsymbol{\Theta}_i\boldsymbol{\Sigma}\boldsymbol{\Theta}_i - 2\boldsymbol{\Theta}_i]$

- How to compute proximal Newton decrement $\lambda_i := \|\mathbf{d}^i\|_{\mathbf{x}^i}$?

$$\lambda_i := [p - 2\mathrm{trace}(\mathbf{W}_i) + \mathrm{trace}(\mathbf{W}_i^2)]^{1/2}, \quad \mathbf{W}_i = \boldsymbol{\Theta}_i(\boldsymbol{\Sigma} - \rho\mathbf{U}^*)$$

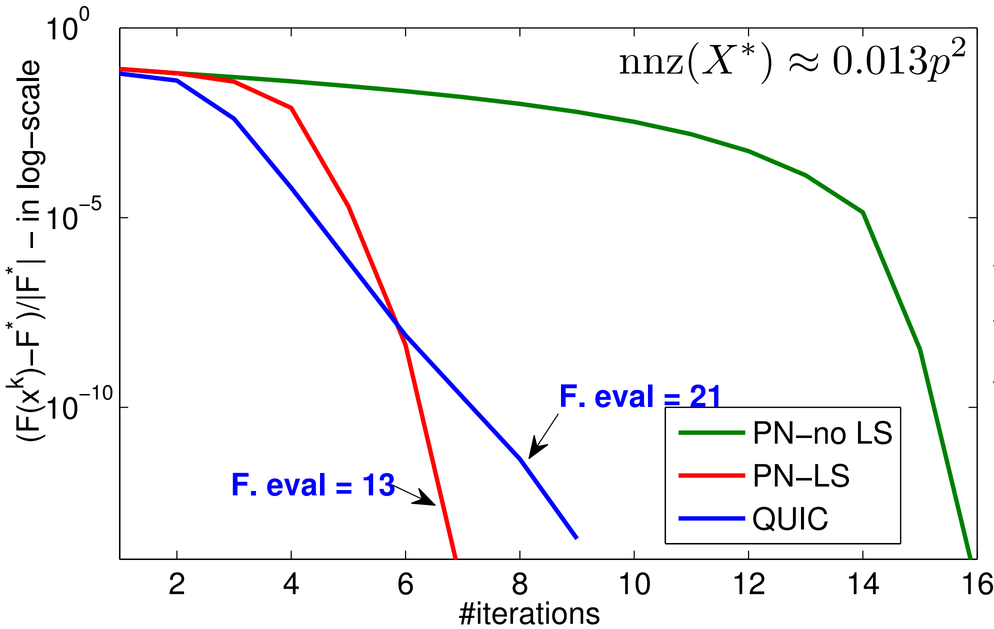
Graphical model selection: numerical examples

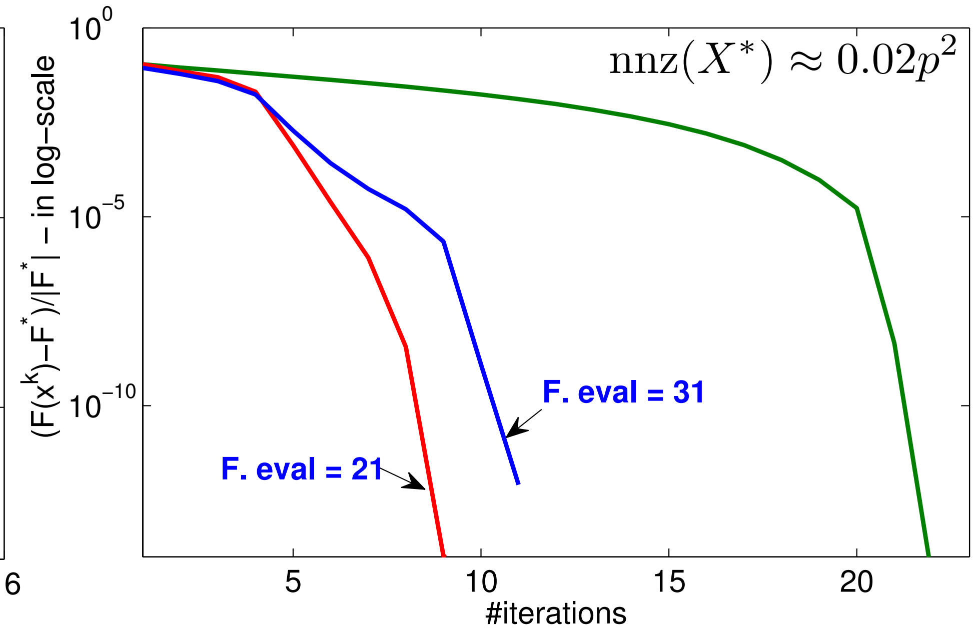
Our method vs QUIC [Hseih2011]

- QUIC subproblem solver: special block-coordinate descent

- Our subproblem solver: general proximal algorithms

Convergence behaviour [rho = 0.5]: Lymph [p = 587] (left), Leukemia [p = 1255] (right)





Step-size selection strategies: Arabidopsis [p = 834], Leukemia [p = 1255], Hereditary [p = 1869]

	Synthetic ($\rho = 0.01$)			Arabidopsis ($\rho = 0.5$)			Leukemia ($\rho = 0.1$)			Hereditary ($\rho = 0.1$)		
LS SCHEME	#iter	#chol	#Mm	#iter	#chol	#Mm	#iter	#chol	#Mm	#iter	#chol	#Mm
NoLS	25.4	-	3400	18	-	1810	44	-	9842	72	-	20960
BtkLS	25.5	37.0	2436	11	25	718	15	50	1282	19	63	2006
E-BtkLS	25.5	36.2	2436	11	24	718	15	49	1282	15	51	1282
FwLS	18.1	26.2	1632	10	17	612	12	34	844	14	44	1126

*Details: "A proximal Newton framework for composite minimization: Graph learning without Cholesky decompositions and matrix inversions," ICML'13 and lions.epfl.ch/publications.

Graphical model selection: numerical examples

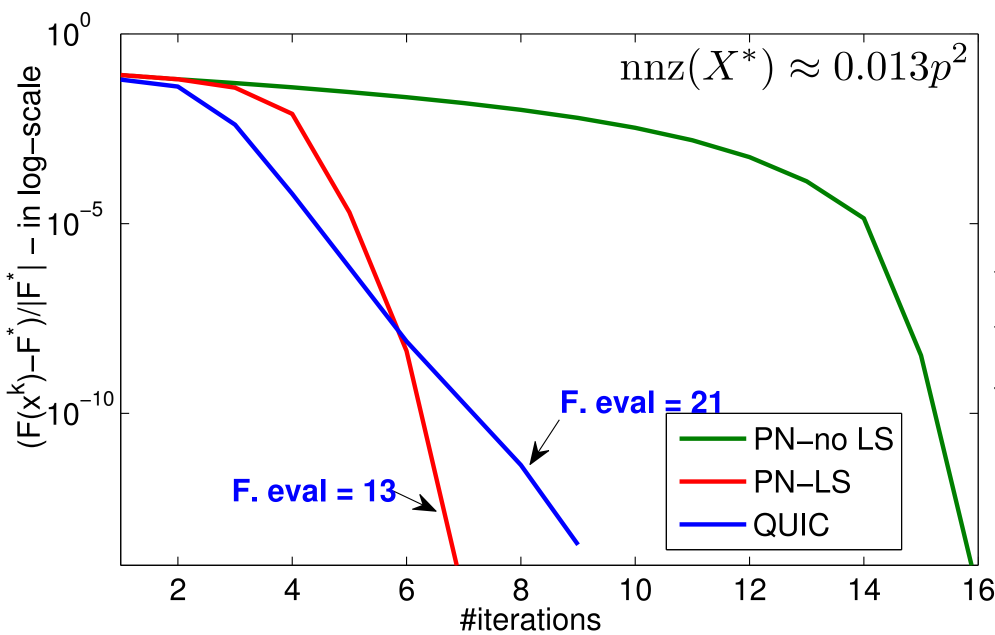
Our method vs QUIC [Hseih2011]

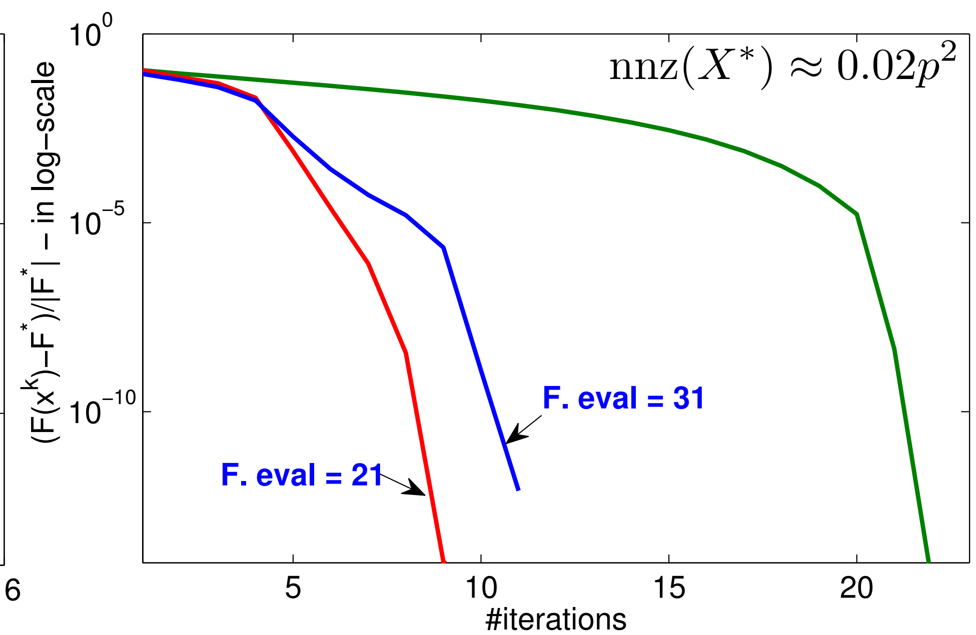
- QUIC subproblem solver: special block-coordinate descent

On the average x5 acceleration (up to x15) over Matlab QUIC

- Our subproblem solver: general proximal algorithms

Convergence behaviour [rho = 0.5]: Lymph [p = 587] (left), Leukemia [p = 1255] (right)





Step-size selection strategies: Arabidopsis [p = 834], Leukemia [p = 1255], Hereditary [p = 1869]

	Synthetic ($\rho = 0.01$)			Arabidopsis ($\rho = 0.5$)			Leukemia ($\rho = 0.1$)			Hereditary ($\rho = 0.1$)		
LS SCHEME	#iter	#chol	#Mm	#iter	#chol	#Mm	#iter	#chol	#Mm	#iter	#chol	#Mm
NoLS	25.4	-	3400	18	-	1810	44	-	9842	72	-	20960
BtkLS	25.5	37.0	2436	11	25	718	15	50	1282	19	63	2006
E-BtkLS	25.5	36.2	2436	11	24	718	15	49	1282	15	51	1282
FwLS	18.1	26.2	1632	10	17	612	12	34	844	14	44	1126

*Details: "A proximal Newton framework for composite minimization: Graph learning without Cholesky decompositions and matrix inversions," ICML'13 and lions.epfl.ch/publications.

Composite minimization: alternatives?

(P) $$\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$

f
convex and **smooth**

g
convex and possibly **nonsmooth**

Existing numerical approaches

- Splitting methods

  - **Forward-backward**: *applicable if **f** has Lipschitz gradient*

  - **Douglas-Rachford** decomposition: ***f*** *and* ***g*** *have "tractable" proximity operators*

- Augmented Lagrangian methods (e.g., D-R again)

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x},\mathbf{y}) := f(\mathbf{x}) + g(\mathbf{y})\}$$
$$\text{s.t.} \quad \boxed{\mathbf{x} - \mathbf{y} = 0}$$

$-\log\det$: full eigen decomposition

$-\log$ (with a linear operator): non-linear system

Prox operator of self-concordant functions are costly!

- New theory for AL/ADMM/decomposition (*another presentation!*)

Our “cheaper” variable metric strategies

- Proximal **gradient** scheme*

Given $\mathbf{x}^0$, generate a sequence $\{\mathbf{x}^k\}_{k\geq 0}$ such that

$$\boxed{\mathbf{x}^{k+1} := \mathbf{x}^k + \alpha_k \mathbf{d}^k_{\mathbf{H}_k}}$$

where $\alpha_k \in (0,1]$ is step-size, $\mathbf{d}^k_{\mathbf{H}_k}$ is a search direction

- How to compute the search direction?

$$\mathbf{d}^k_{\mathbf{H}_k} := \arg\min_{\mathbf{d}} \left\{ f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T \mathbf{d} + \frac{1}{2}\mathbf{d}^T \mathbf{H}_k \mathbf{d} + g(\mathbf{x}^k + \mathbf{d}) \right\}, \quad \mathbf{H}_k = \mathbf{D}_k : \text{ diagonal}$$

- No Lipschitz assumption

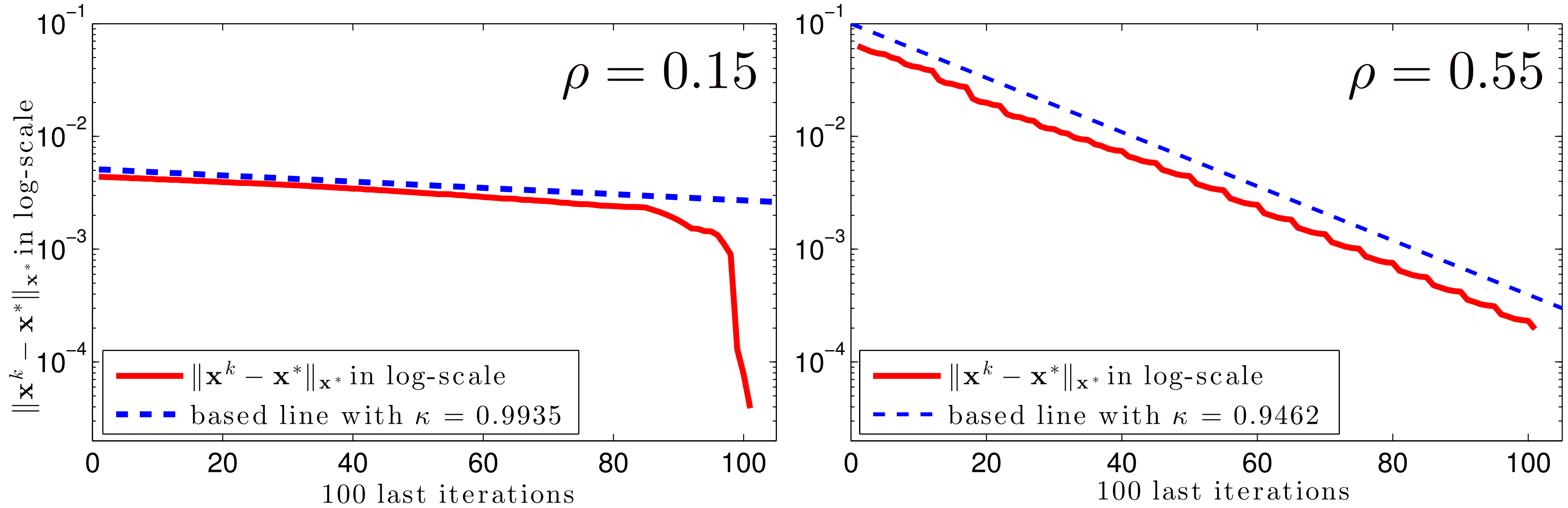
A new predictor corrector scheme (with local linear convergence*)

- Proximal **quasi**-Newton scheme (BFGS updates)*

*Details: “Composite Self-Concordant Minimization,” arxiv and lions.epfl.ch/publications.

New theory: Local linear convergence of the PG

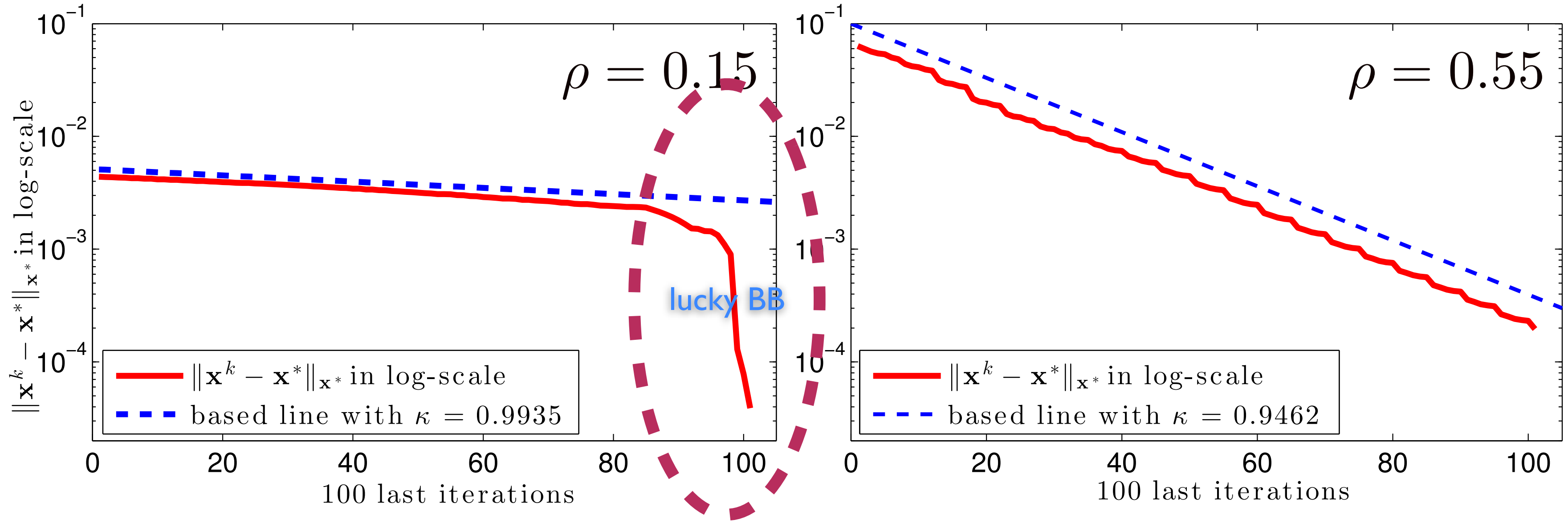
Graph learning: Lymph [p = 587]



$$\min_{\boldsymbol{\Theta} \succ 0} \left\{ \underbrace{-\log\det(\boldsymbol{\Theta}) + \mathrm{trace}(\boldsymbol{\Sigma}\boldsymbol{\Theta})}_{f(\mathbf{x})} + \underbrace{\rho\|\mathrm{vec}(\boldsymbol{\Theta})\|_1}_{g(\mathbf{x})} \right\}$$

New theory: Local linear convergence of the PG

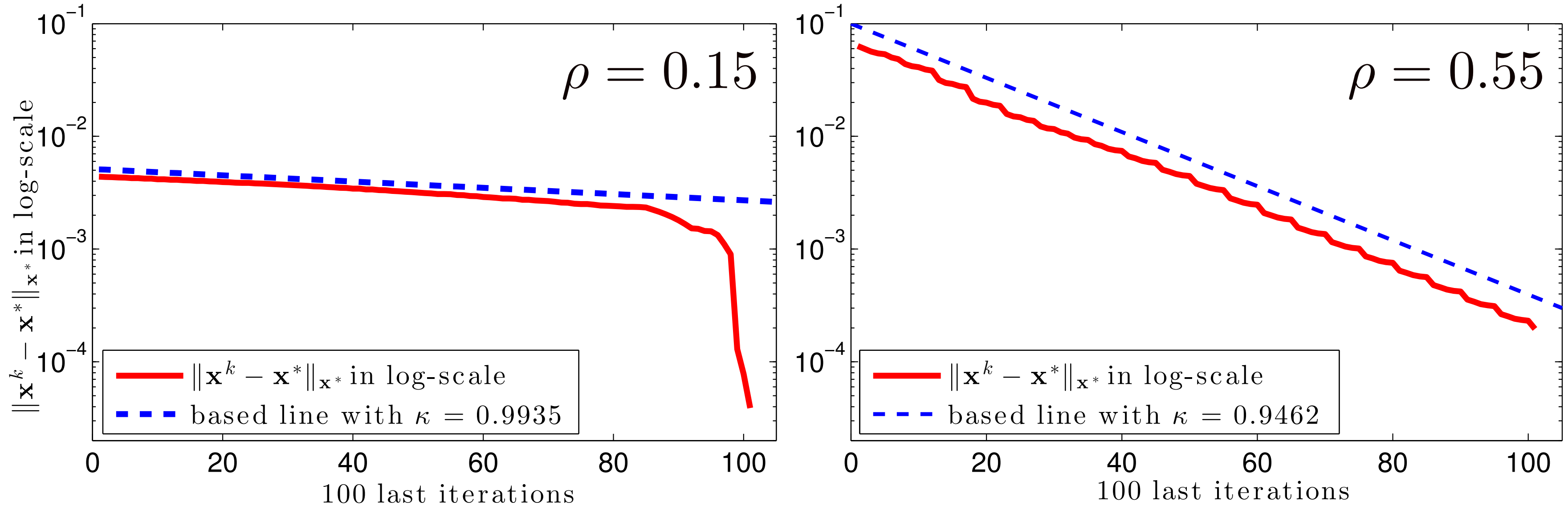
Graph learning: Lymph [p = 587]



$$\min_{\Theta \succ 0} \left\{ \underbrace{-\log\det(\Theta) + \mathrm{trace}(\Sigma\Theta)}_{f(\mathbf{x})} + \underbrace{\rho\|\mathrm{vec}(\Theta)\|_1}_{g(\mathbf{x})} \right\}$$

New theory: Local linear convergence of the PG

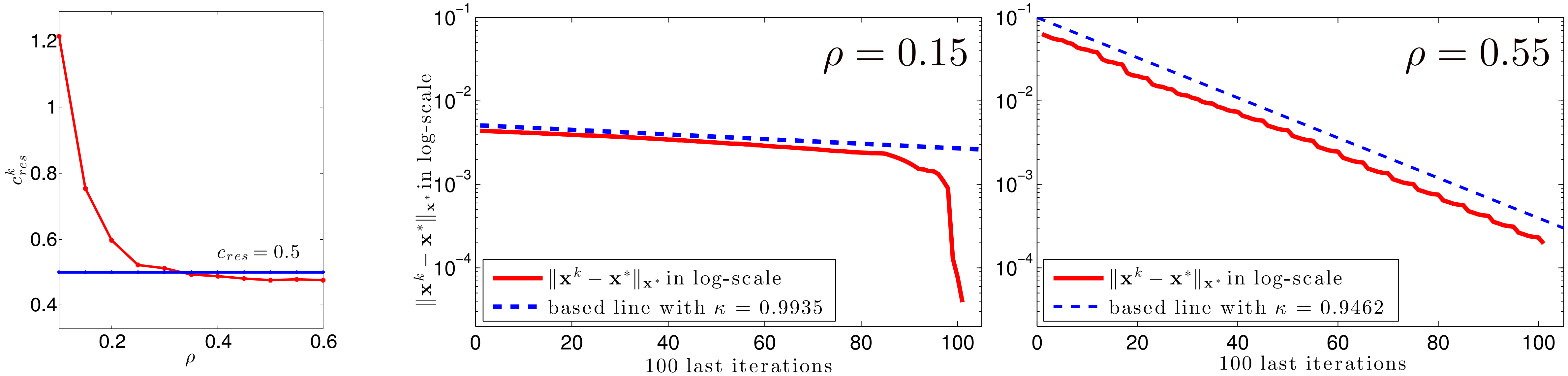
Graph learning: Lymph [p = 587]



$$\min_{\mathbf{\Theta} \succ 0} \left\{ \underbrace{-\log\det(\mathbf{\Theta}) + \mathrm{trace}(\mathbf{\Sigma}\mathbf{\Theta})}_{f(\mathbf{x})} + \underbrace{\rho\|\mathrm{vec}(\mathbf{\Theta})\|_1}_{g(\mathbf{x})} \right\}$$

New theory: Local linear convergence of the PG

Graph learning: Lymph [p = 587]

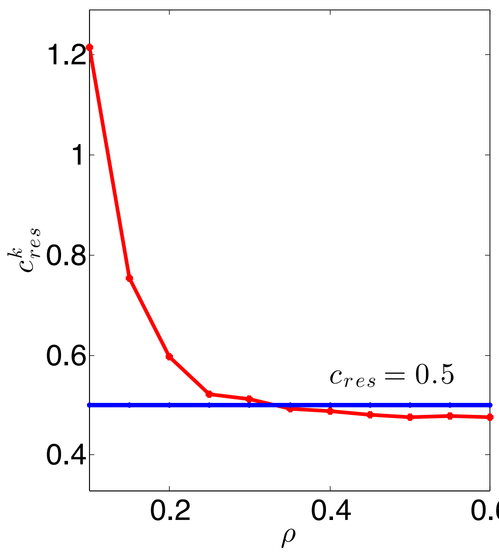


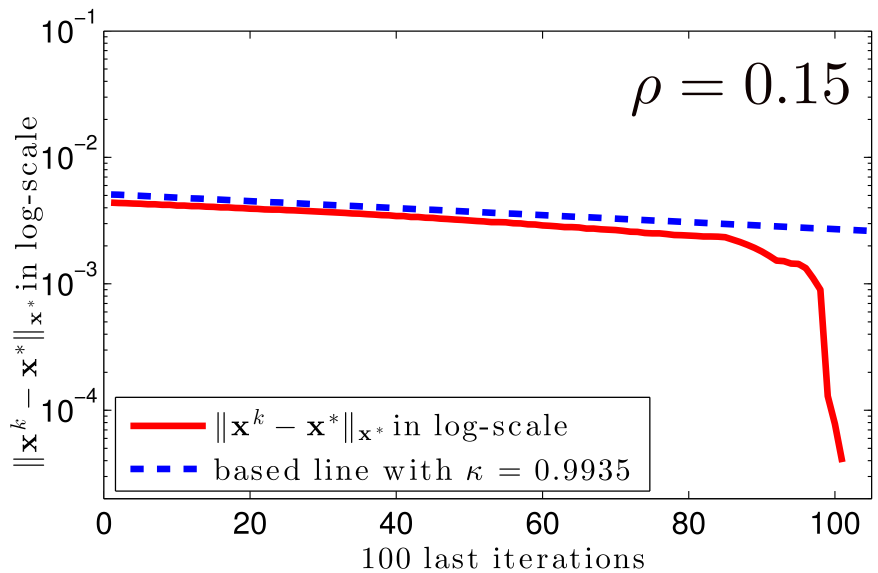
Theory is based on the notion of "restricted strong convexity"

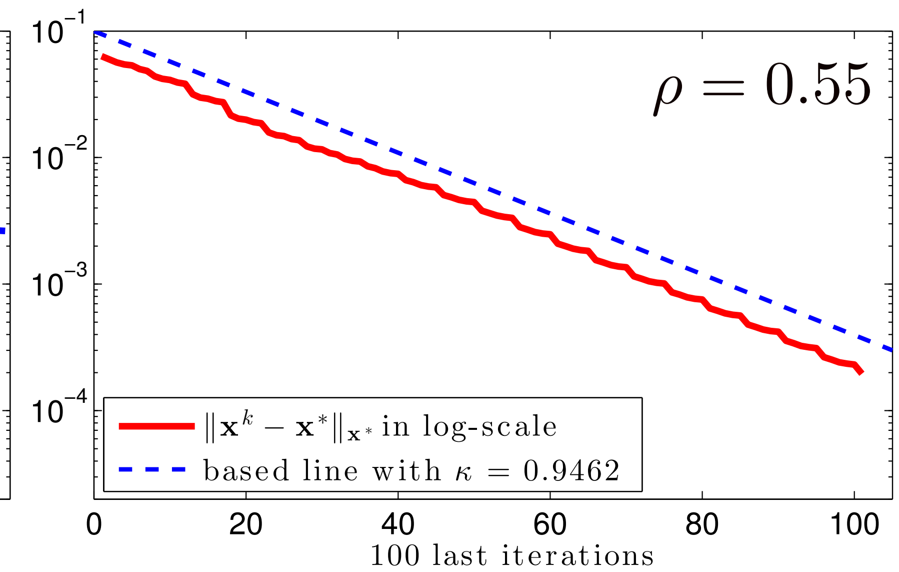
$$\min_{\boldsymbol{\Theta}\succ 0}\left\{\underbrace{-\log\det(\boldsymbol{\Theta}) + \mathrm{trace}(\boldsymbol{\Sigma}\boldsymbol{\Theta})}_{f(\mathbf{x})} + \underbrace{\rho\|\mathrm{vec}(\boldsymbol{\Theta})\|_1}_{g(\mathbf{x})}\right\}$$

New theory: Local linear convergence of the PG

Graph learning: Lymph [p = 587]

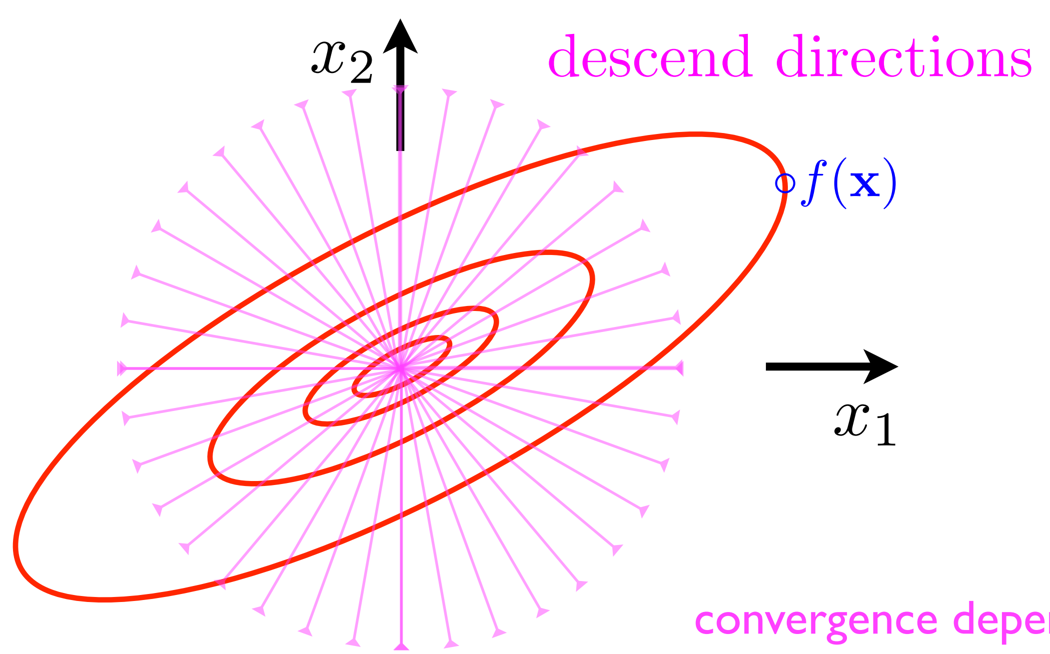






Theory is based on the notion of "restricted strong convexity"

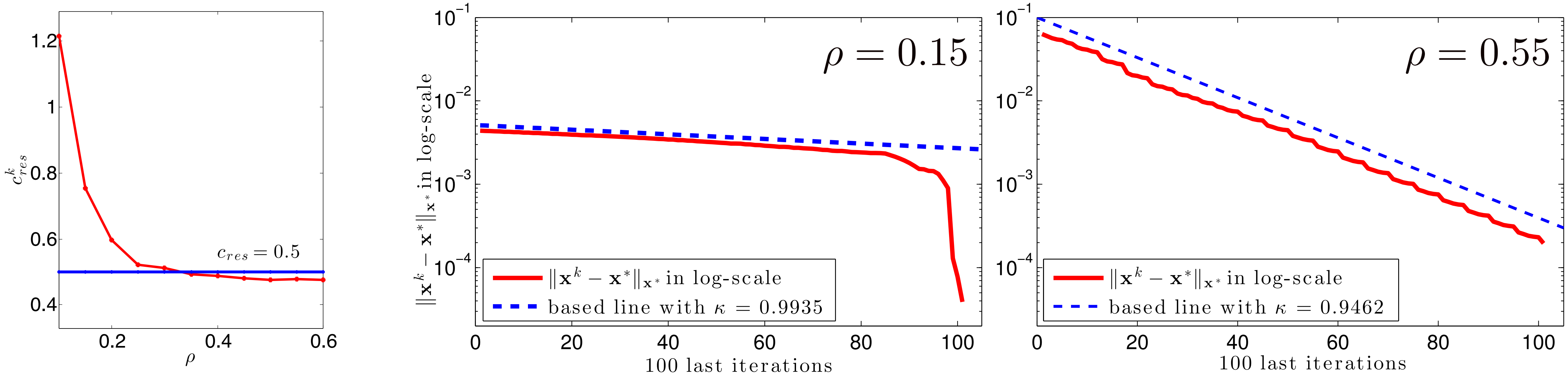
$$\min_{\Theta \succ 0} \left\{ \underbrace{-\log\det(\Theta) + \mathrm{trace}(\Sigma\Theta)}_{f(\mathbf{x})} \right\}$$



convergence depends on the full condition number

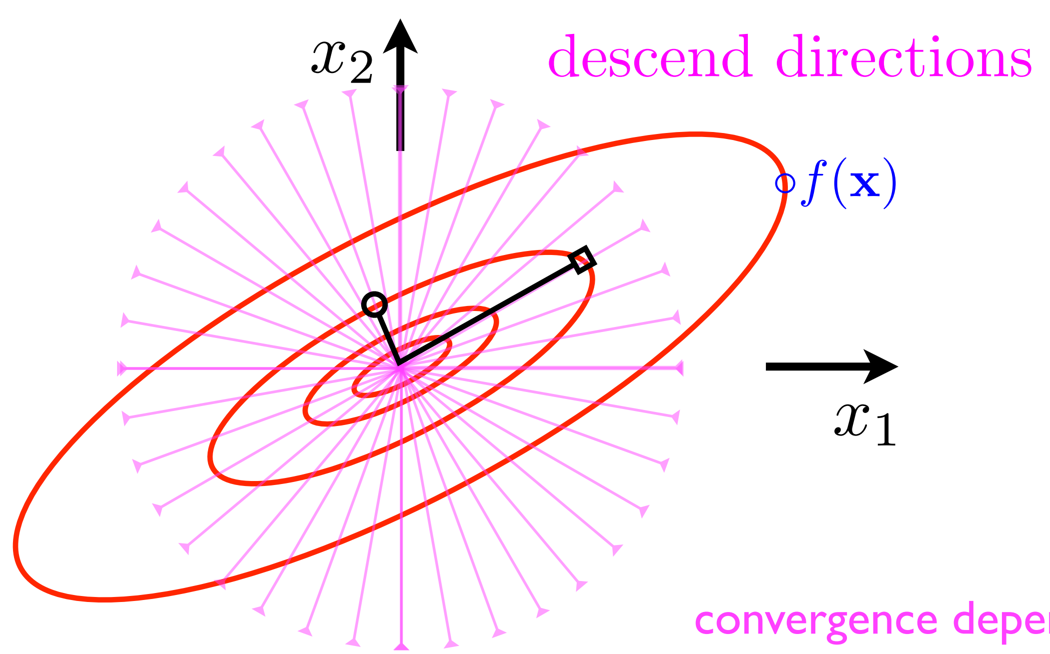
New theory: Local linear convergence of the PG

Graph learning: Lymph [p = 587]



Theory is based on the notion of "restricted strong convexity"

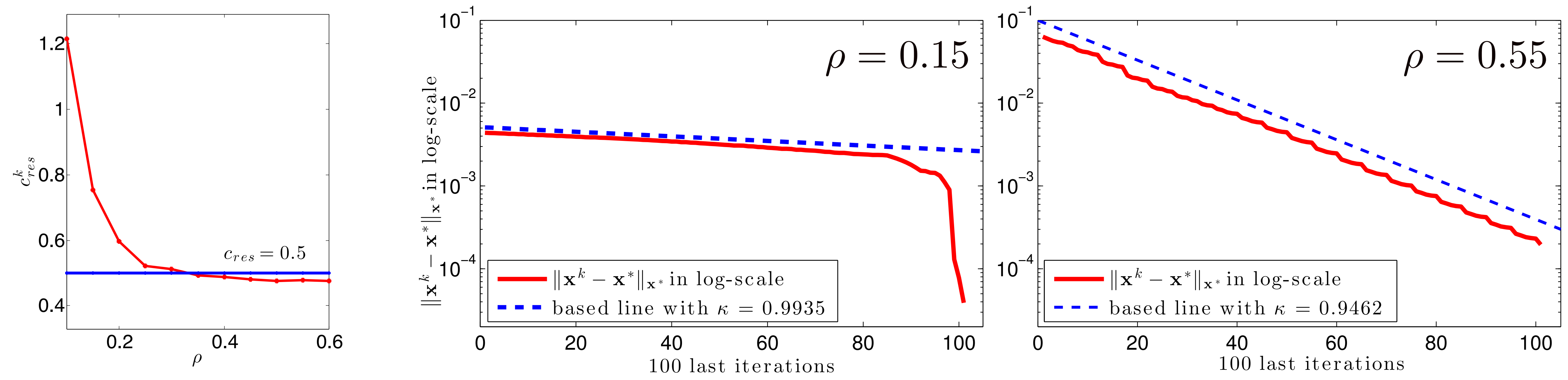
$$\min_{\boldsymbol{\Theta} \succ 0}\left\{\underbrace{-\log\det(\boldsymbol{\Theta}) + \mathrm{trace}(\boldsymbol{\Sigma}\boldsymbol{\Theta})}_{f(\mathbf{x})}\right\}$$



convergence depends on the full condition number

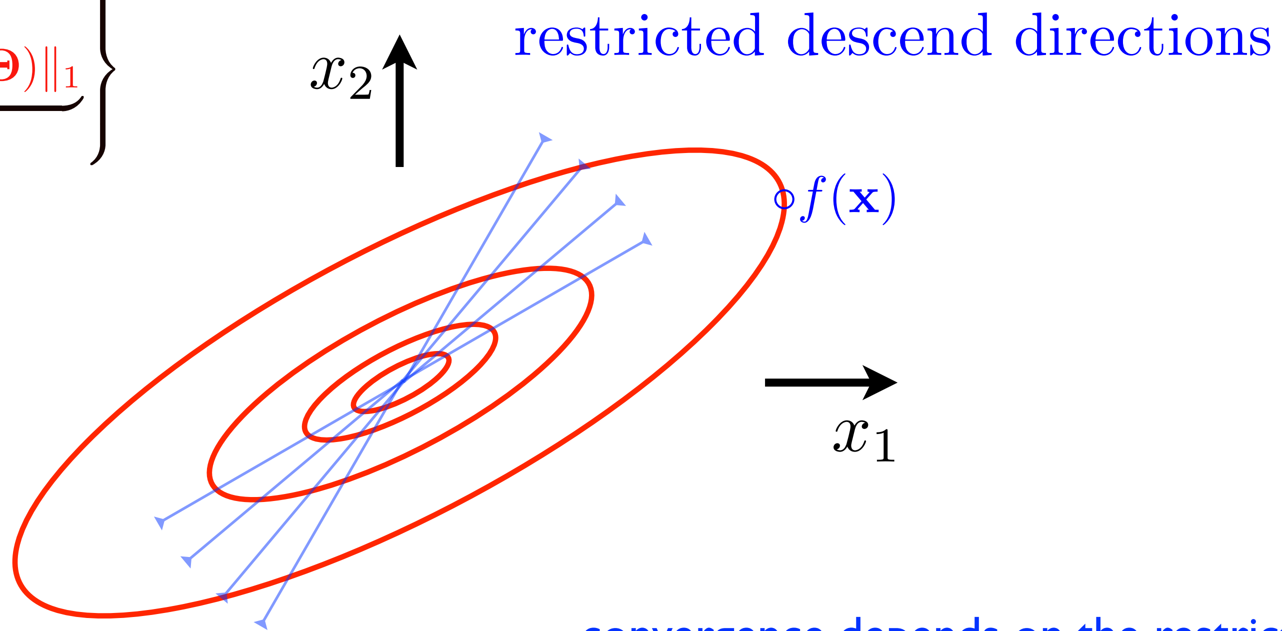
New theory: Local linear convergence of the PG

Graph learning: Lymph [p = 587]



Theory is based on the notion of "restricted strong convexity"

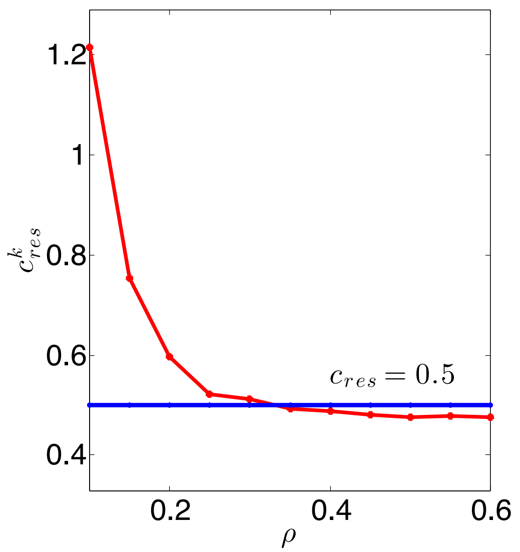
$$\min_{\Theta \succ 0} \left\{ \underbrace{-\log\det(\Theta) + \mathrm{trace}(\Sigma\Theta)}_{f(\mathbf{x})} + \underbrace{\rho\|\mathrm{vec}(\Theta)\|_1}_{g(\mathbf{x})} \right\}$$

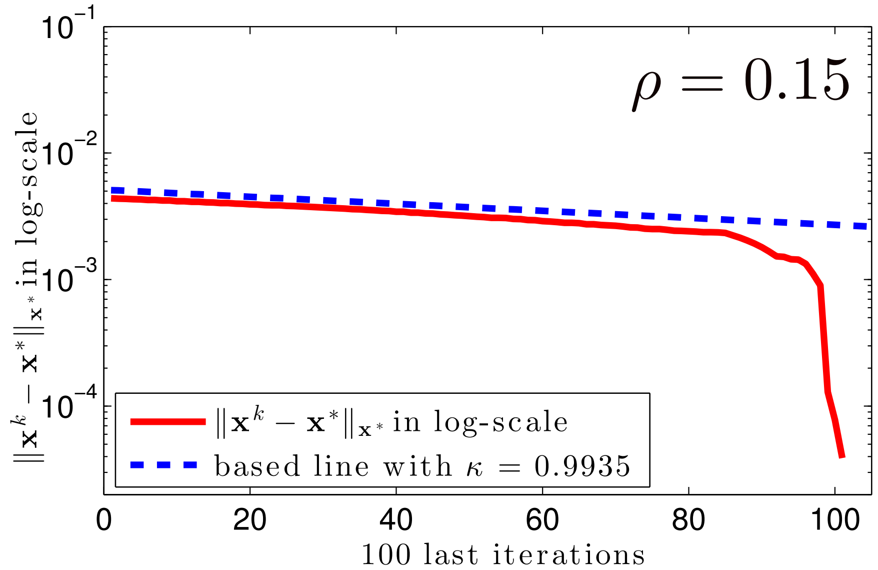


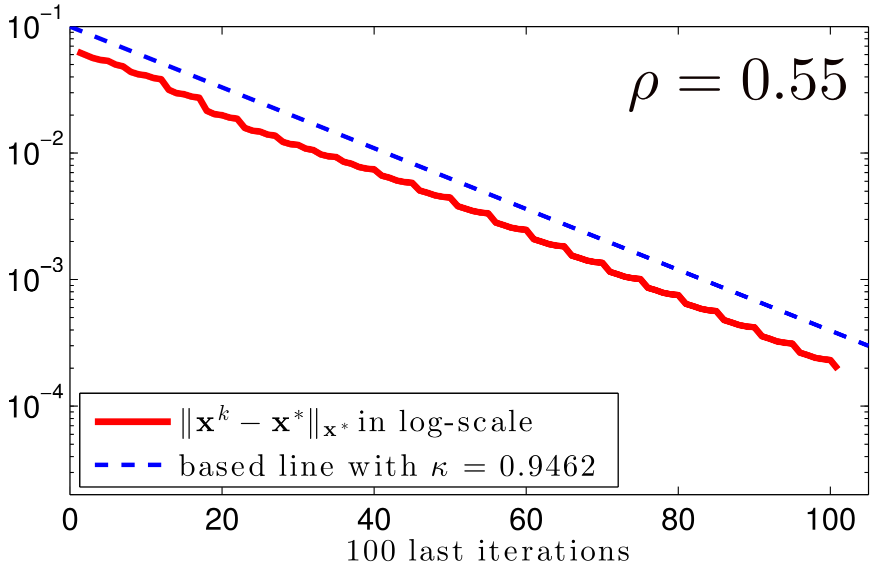
convergence depends on the restricted condition number

New theory: Local linear convergence of the PG

Graph learning: Lymph [p = 587]

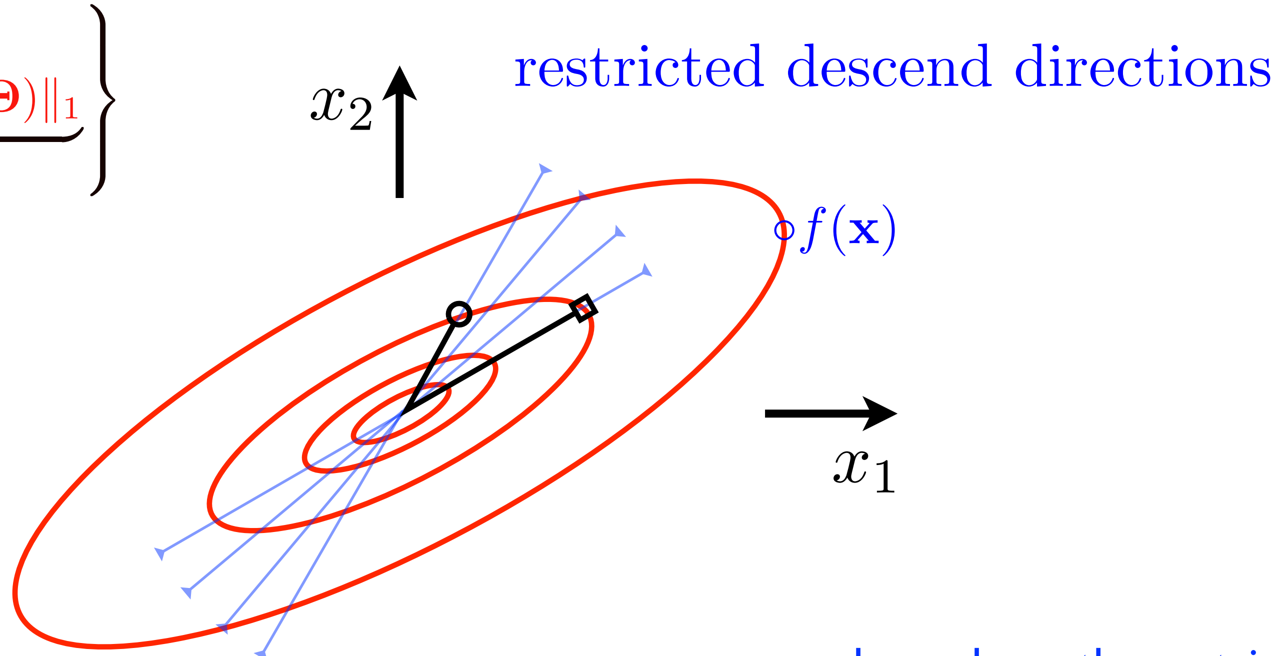






Theory is based on the notion of "restricted strong convexity"

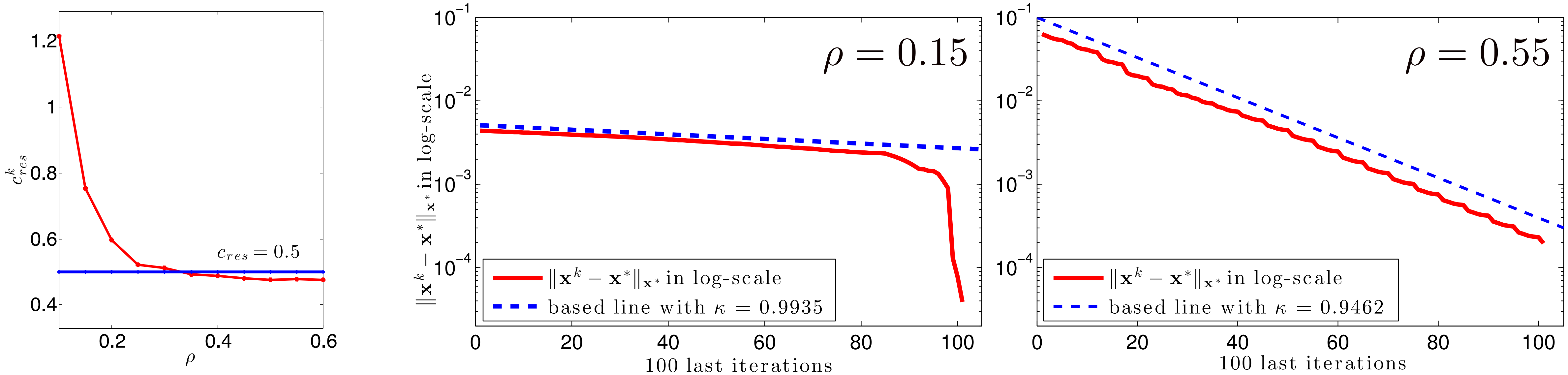
$$\min_{\Theta \succ 0} \left\{ \underbrace{-\log\det(\Theta) + \mathrm{trace}(\Sigma\Theta)}_{f(\mathbf{x})} + \underbrace{\rho\|\mathrm{vec}(\Theta)\|_1}_{g(\mathbf{x})} \right\}$$



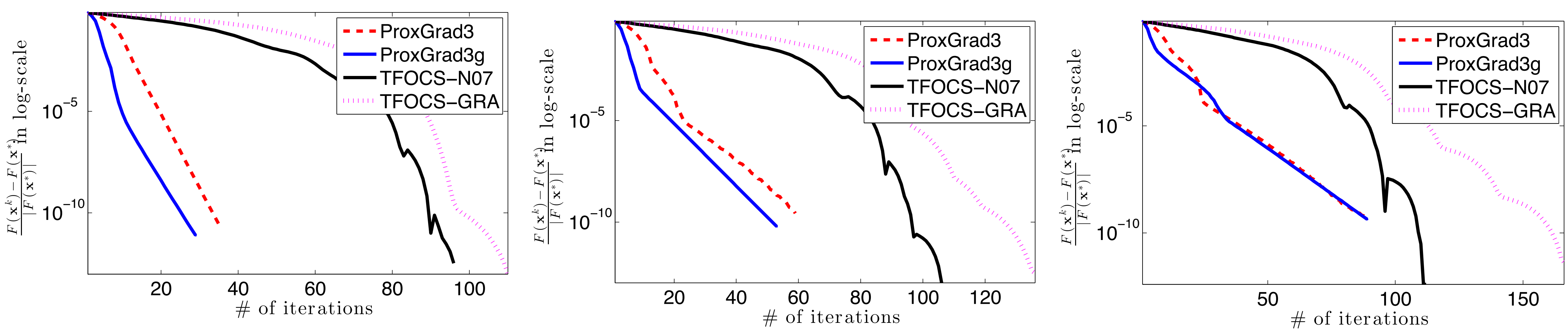
convergence depends on the restricted condition number

New theory: Local linear convergence of the PG

Graph learning: Lymph [p = 587]



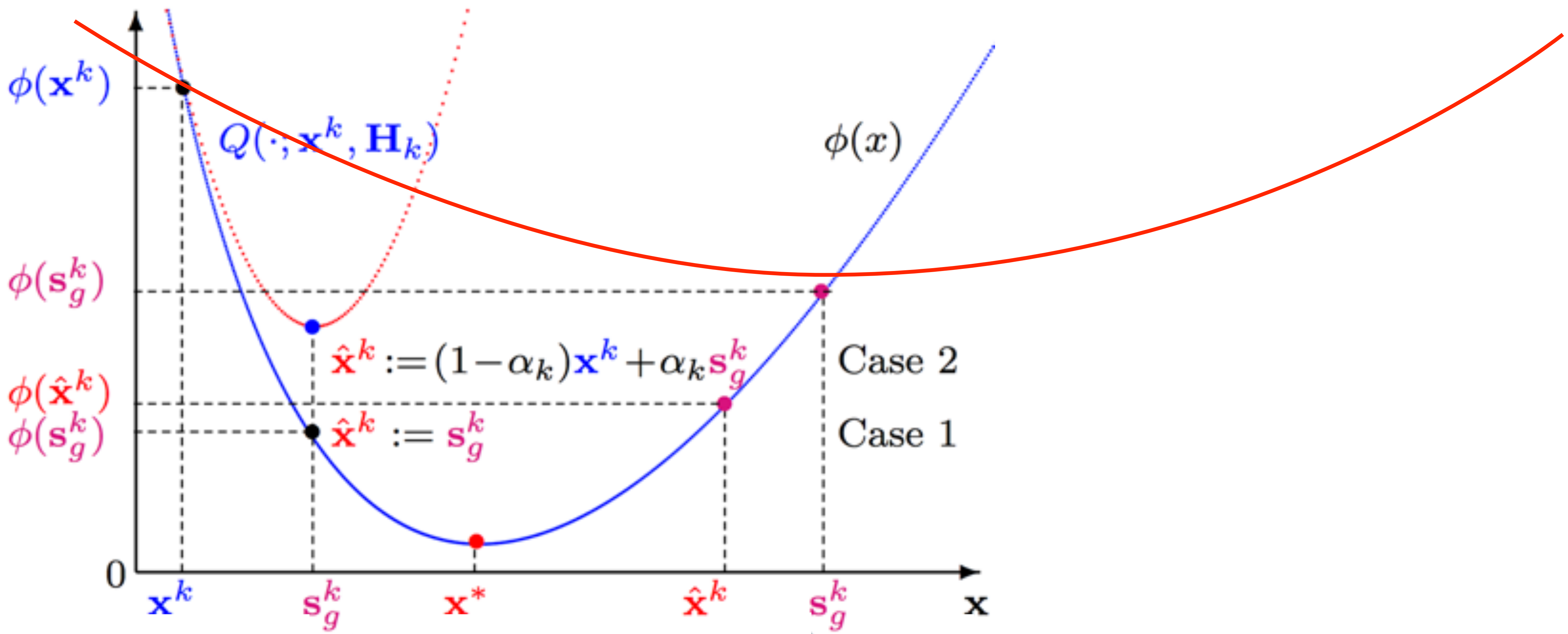
Heteroschedastic LASSO [rho decreases from left to right]



$$\mathbf{x}^* \equiv (\phi^*, \gamma^*) = \underset{\phi,\gamma}{\operatorname{argmin}} \Big\{ \underbrace{-\log(\gamma) + \frac{1}{2n}\|\gamma\mathbf{y} - \mathbf{X}\phi\|_2^2}_{=: f(\mathbf{x})} + \underbrace{\rho\|\phi\|_1}_{=: g(\mathbf{x})} \Big\}$$

Proximal gradient scheme: new engineering

- A ***greedy*** enhancement $\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$

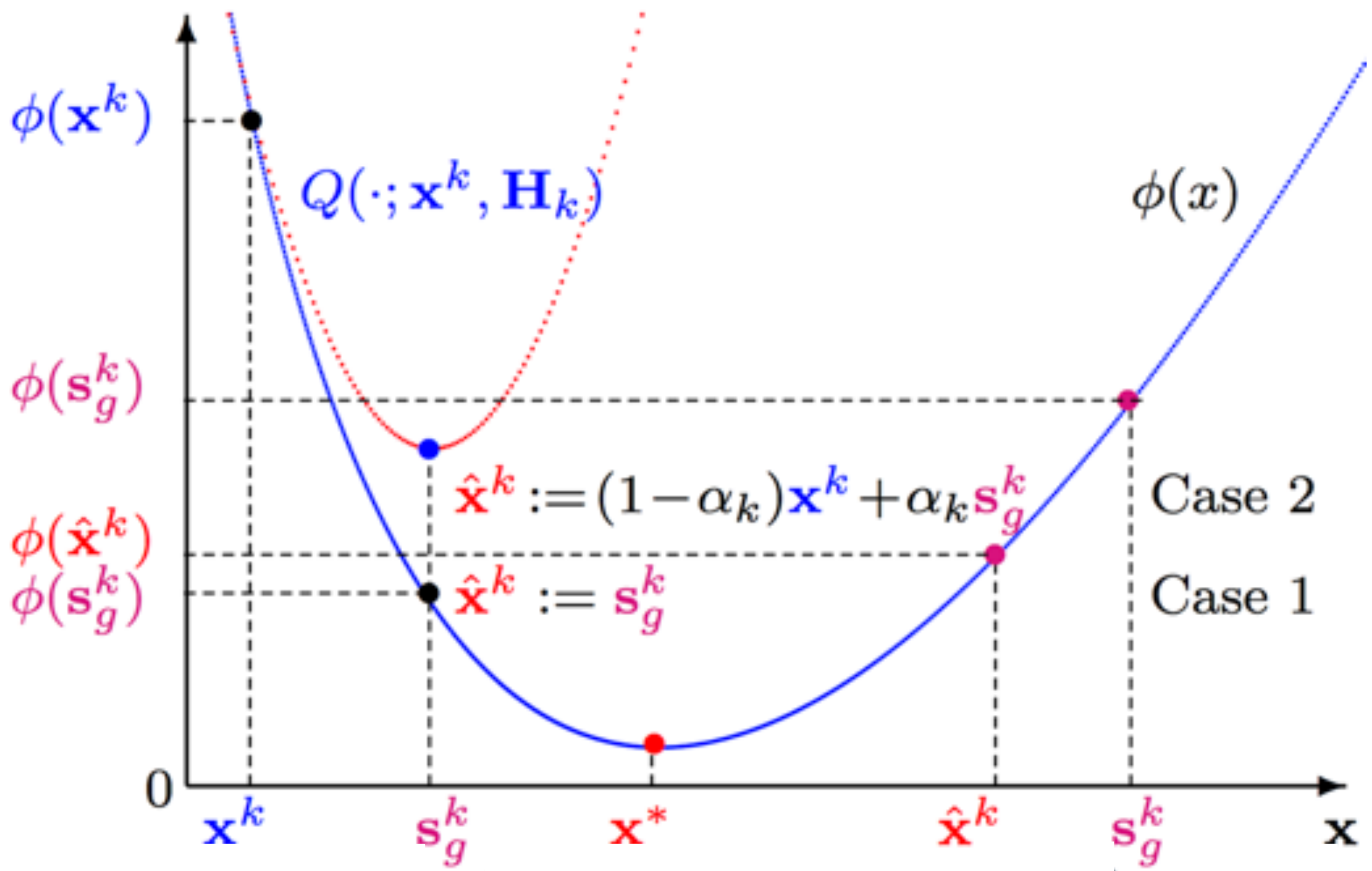


$$\hat{\mathbf{x}}^k := (1-\alpha_k)\mathbf{x}^k + \alpha_k\mathbf{s}_g^k$$

prox: $\mathbf{s}_g^k := \arg\min_{\mathbf{x}\in\mathrm{dom}(F)} \{Q(\mathbf{x};\mathbf{x}^k,\mathbf{D}_k) + g(\mathbf{x})\}$

Proximal gradient scheme: new engineering

- A ***greedy*** enhancement $\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$

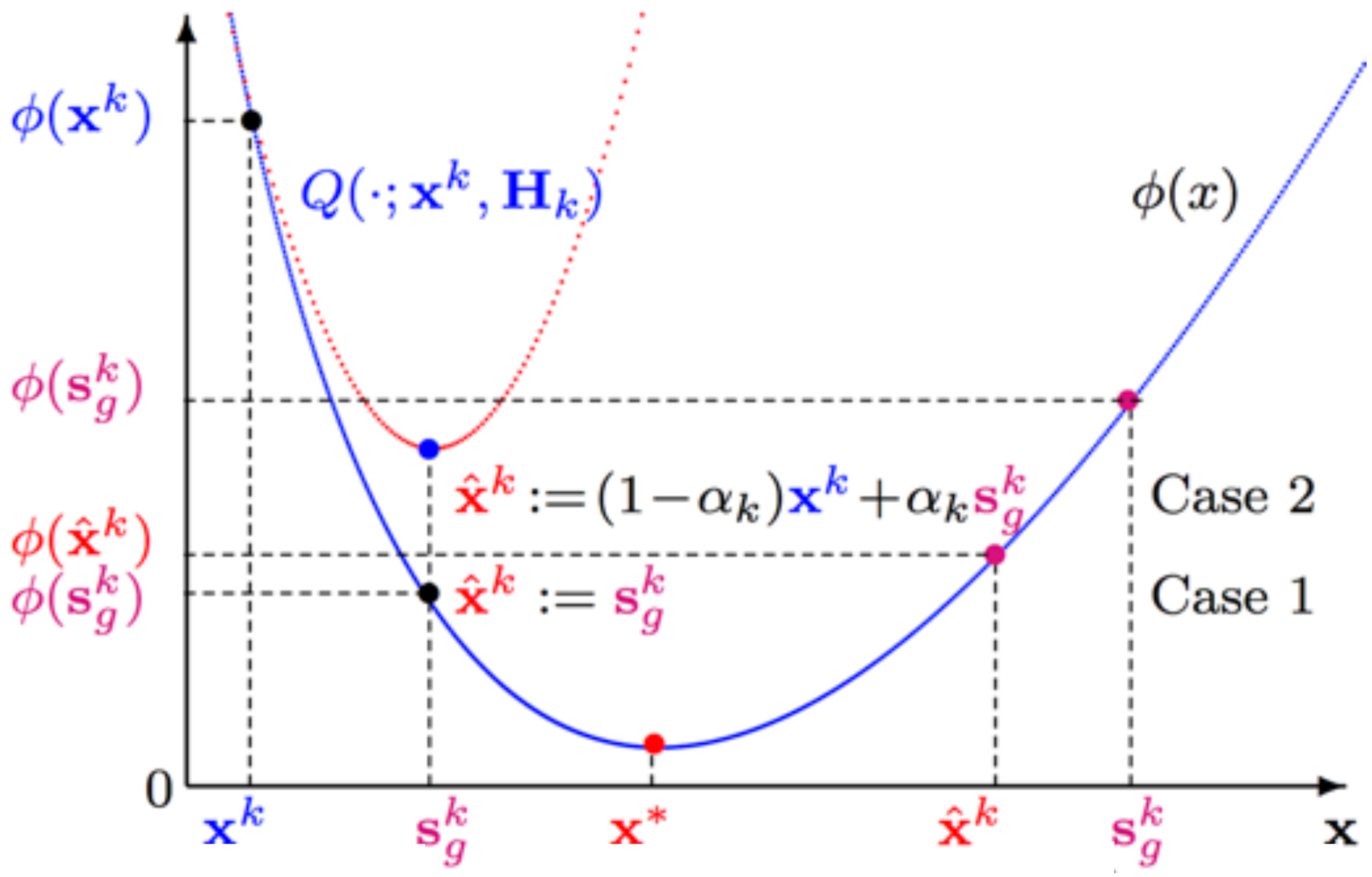


$$\hat{\mathbf{x}}^k := (1-\alpha_k)\mathbf{x}^k + \alpha_k\mathbf{s}_g^k$$

prox: $\mathbf{s}_g^k := \arg\min_{\mathbf{x}\in\mathrm{dom}(F)} \{Q(\mathbf{x};\mathbf{x}^k,\mathbf{D}_k) + g(\mathbf{x})\}$

Proximal gradient scheme: new engineering

- A ***greedy*** enhancement $\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$



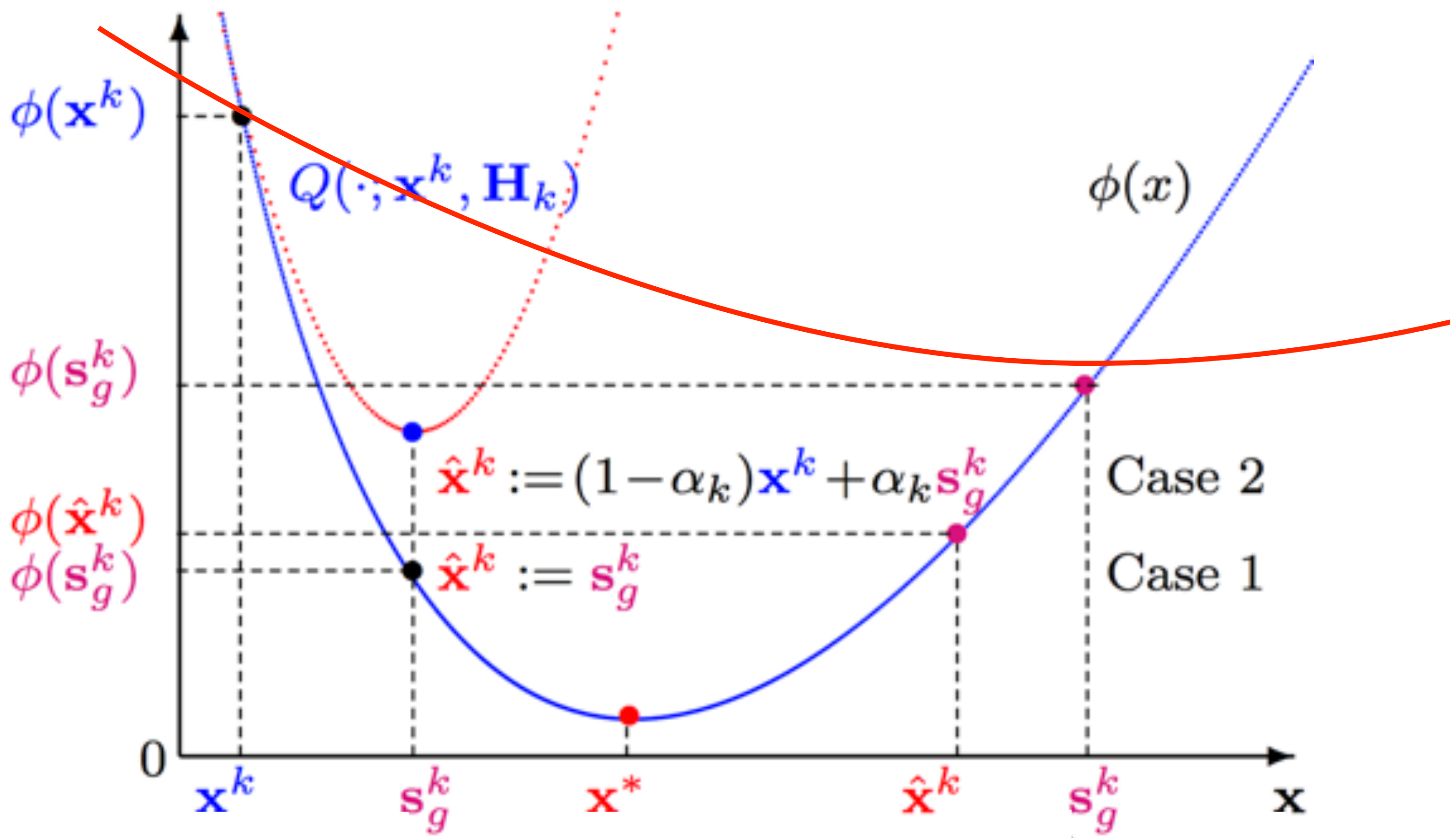
slows down convergence!

$$\hat{\mathbf{x}}^k := (1-\alpha_k)\mathbf{x}^k + \alpha_k\mathbf{s}_g^k$$

prox: $\mathbf{s}_g^k := \arg\min_{\mathbf{x}\in\mathrm{dom}(F)} \{Q(\mathbf{x};\mathbf{x}^k,\mathbf{D}_k) + g(\mathbf{x})\}$

Proximal gradient scheme: new engineering

- A ***greedy*** enhancement $\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$



simple decision:

based on the function values

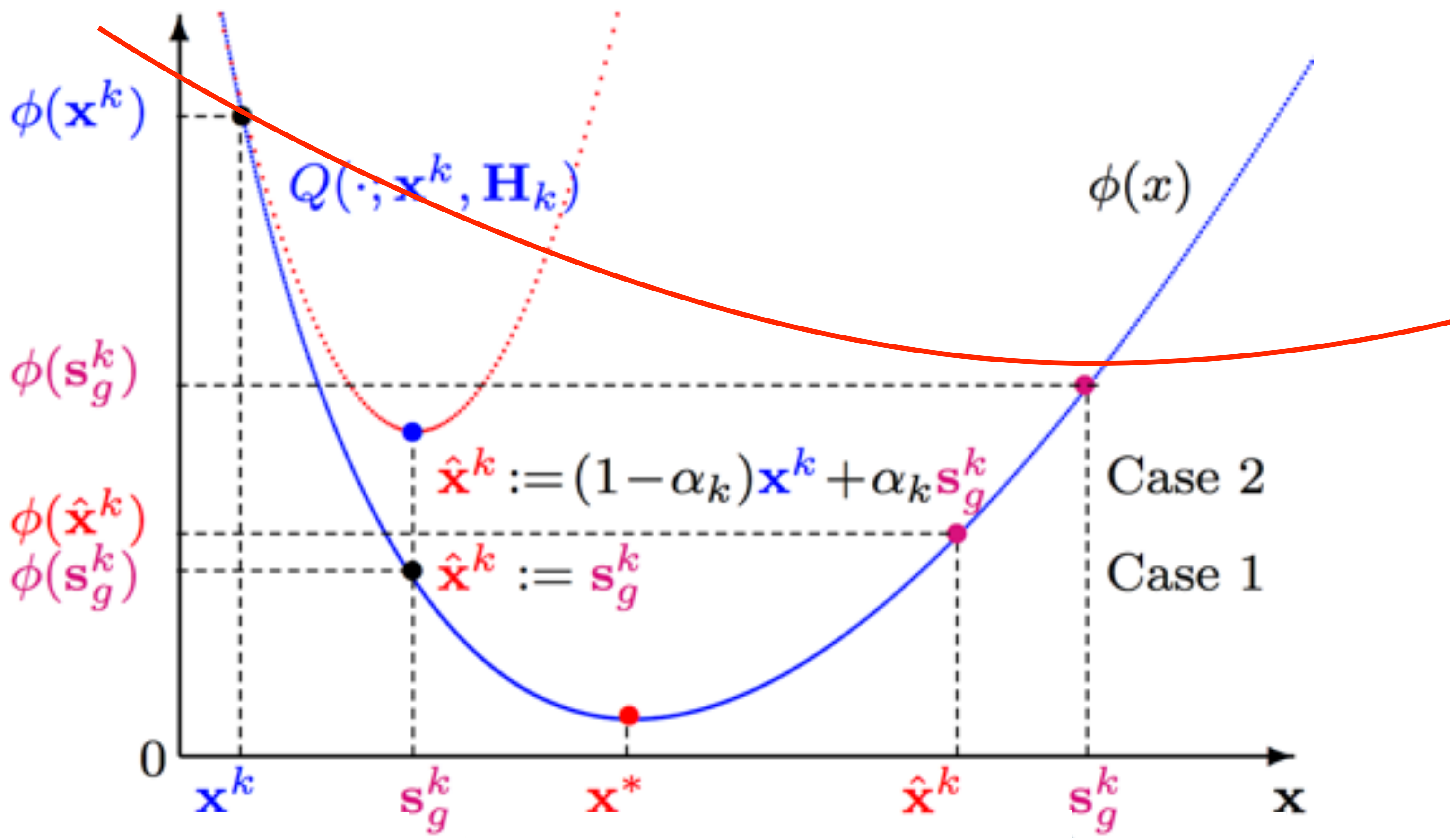
$\phi(\mathbf{s}_g^k)$, $\phi(\hat{\mathbf{x}}^k)$ and $\phi(\mathbf{x}^k)$

$$\hat{\mathbf{x}}^k := (1-\alpha_k)\mathbf{x}^k + \alpha_k\mathbf{s}_g^k$$

prox: $\mathbf{s}_g^k := \arg\min_{\mathbf{x}\in\mathrm{dom}(F)} \{Q(\mathbf{x};\mathbf{x}^k,\mathbf{D}_k) + g(\mathbf{x})\}$

Proximal gradient scheme: new engineering

- A ***greedy*** enhancement $\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$



simple decision:

based on the function values $\phi(\mathbf{s}_g^k)$, $\phi(\hat{\mathbf{x}}^k)$ and $\phi(\mathbf{x}^k)$

cost:

practically none if implemented carefully

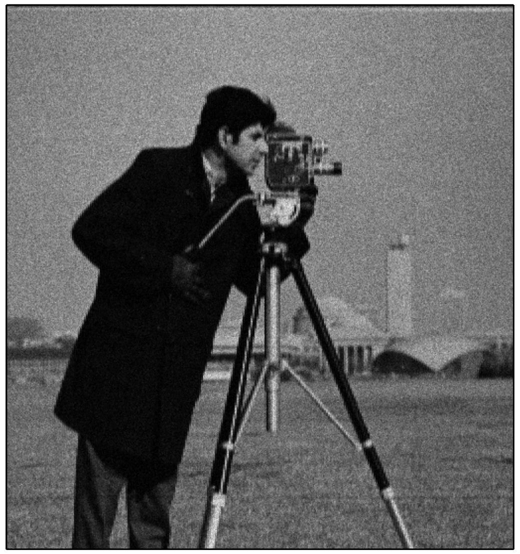
$$\hat{\mathbf{x}}^k := (1-\alpha_k)\mathbf{x}^k + \alpha_k\mathbf{s}_g^k$$

prox: $$\mathbf{s}_g^k := \arg\min_{\mathbf{x}\in\mathrm{dom}(F)} \{Q(\mathbf{x}; \mathbf{x}^k, \mathbf{D}_k) + g(\mathbf{x})\}$$

Poisson imaging reconstruction via TV

Our method vs SPIRAL-TAP [Harmany2012]

Original image



Poisson noise image

Reconstructed image (ProxGrad)

Reconstructed image (ProxGradNewton)

Reconstructed image (SPIRAL–TAP

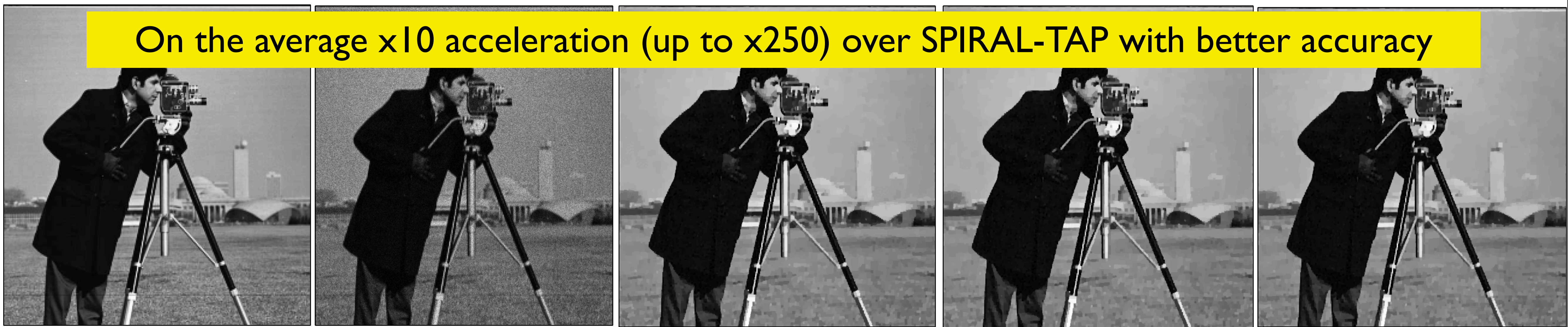
- Poisson imaging reconstruction via TV regularization

$$x^* \in \operatorname*{argmin}_x \Big\{ \underbrace{\sum_{i=1}^m a_i^T x - \sum_{i=1}^m y_i \log(a_i^T x + b_i)}_{f(x)} + g(x) \Big\}$$

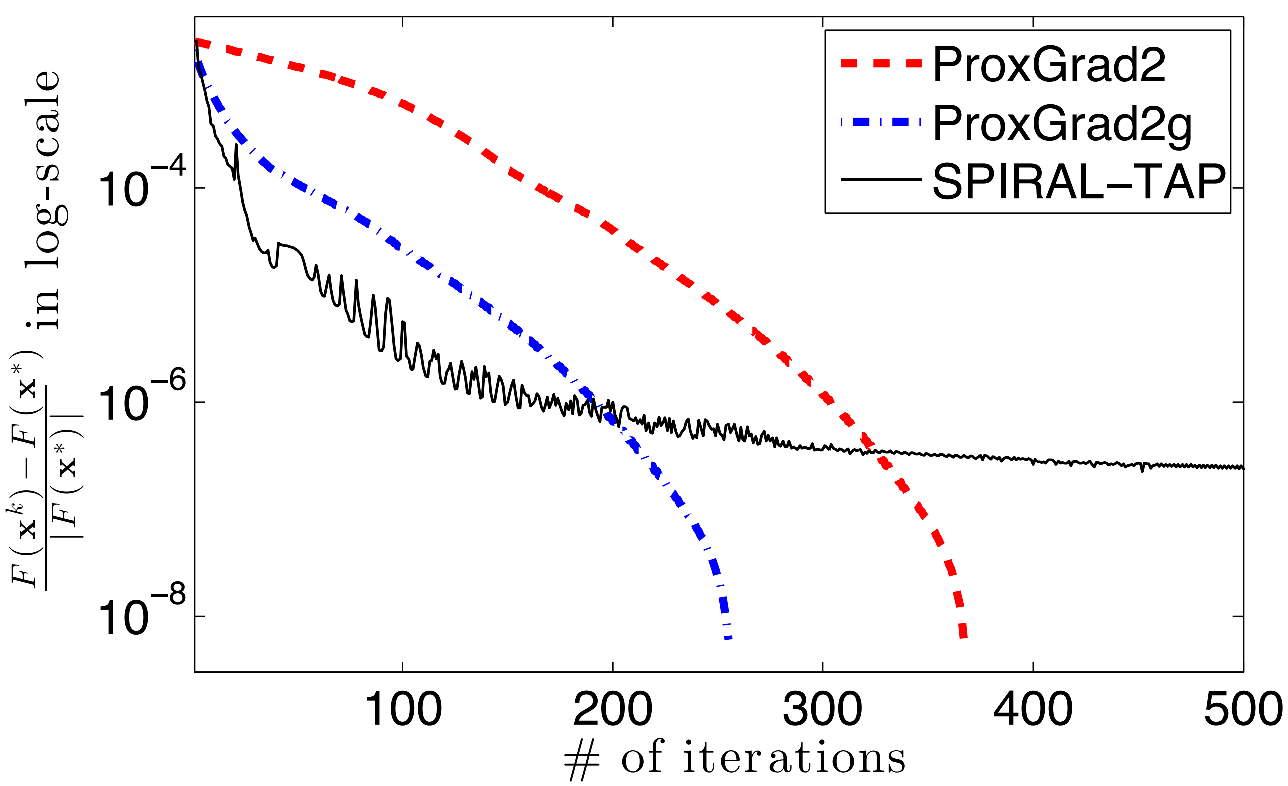
Poisson imaging reconstruction via TV

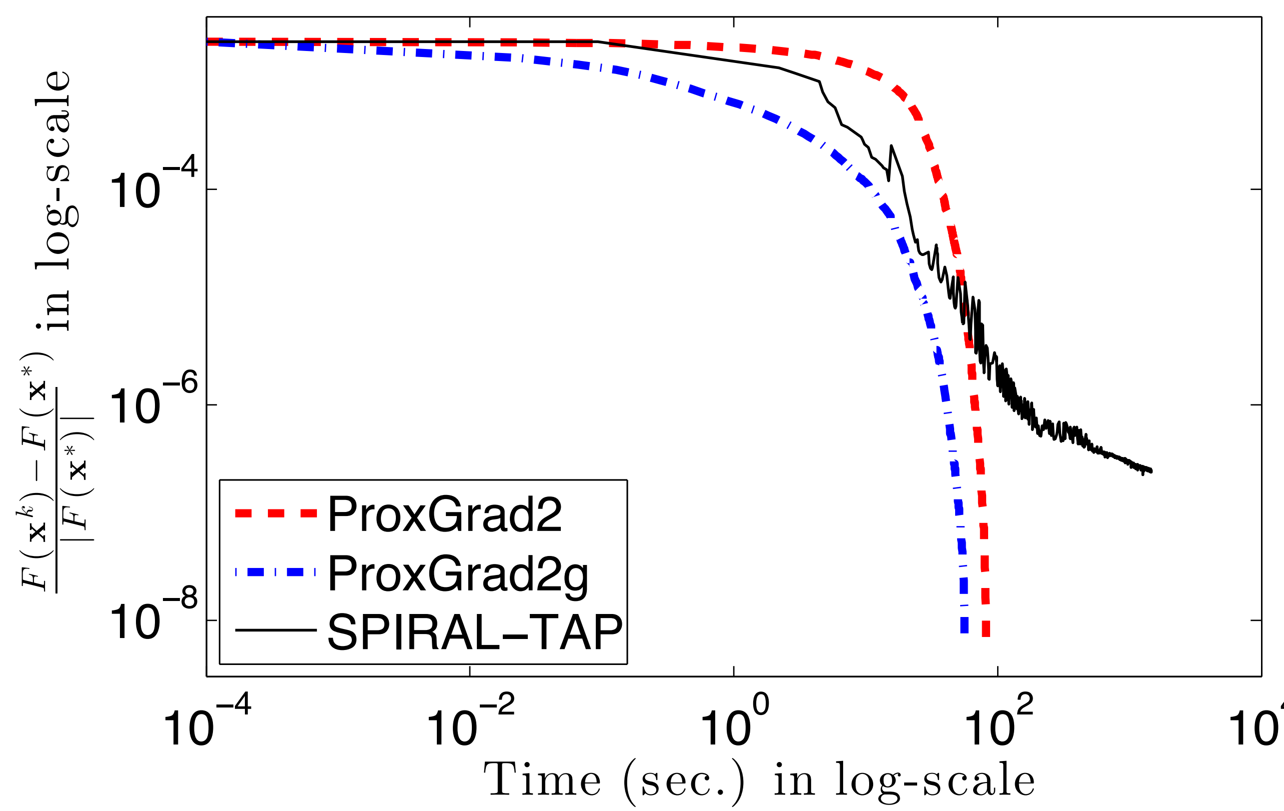
Our method vs SPIRAL-TAP [Harmany2012]

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



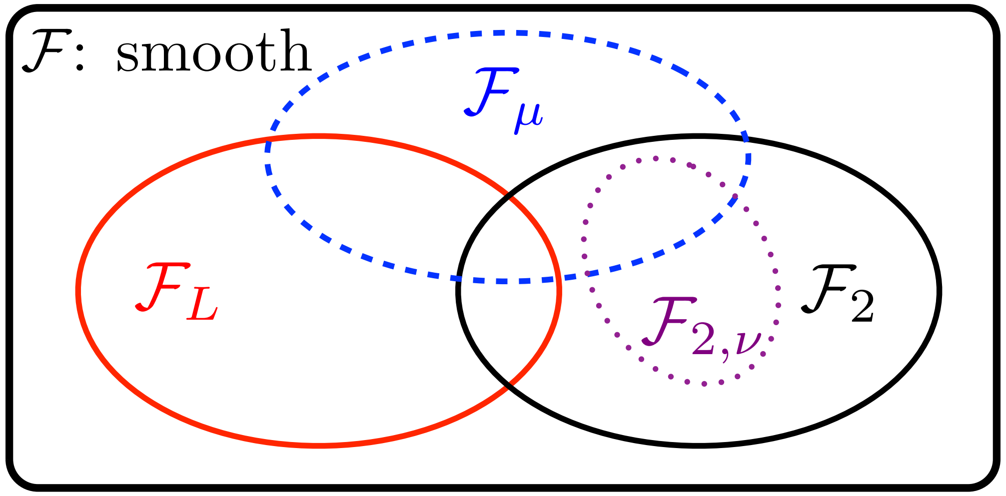
Original image | Poisson noise image | Reconstructed image (ProxGrad) | Reconstructed image (ProxGradNewton) | Reconstructed image (SPIRAL-TAP





Barrier extensions

Constrained convex problems



$$g^* := \min_{\mathbf{x}\in\Omega} g(\mathbf{x})$$

- Ω is endowed with a self-concordant barrier $f(x)$;

Self-concordant barrier

- Given **f** is smooth and convex on its domain. We define:

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

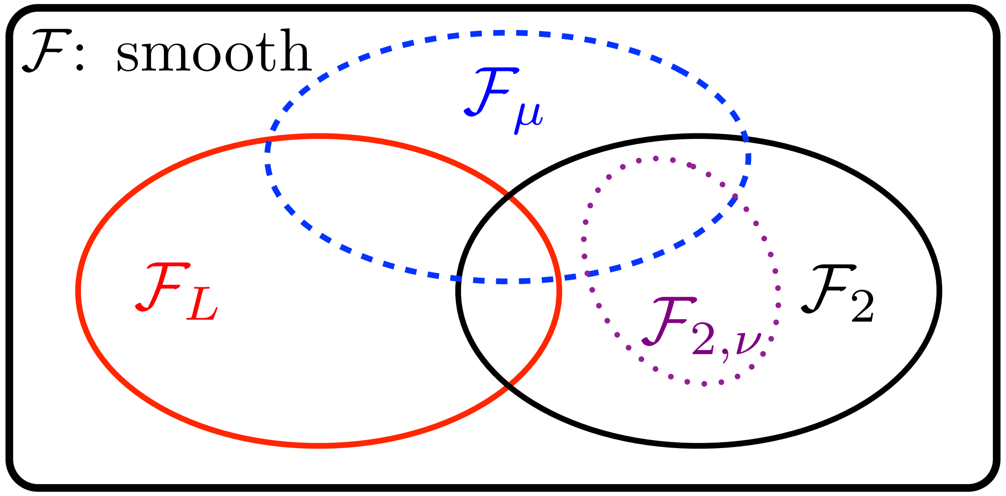
- ***Self-concordance***: **f** is standard self-concordant if

$$|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$$

- ***Self-concordant barrier***: **f** is a *self-concordant barrier* if there exists $\nu > 0$ such that:

$$\begin{cases} f \text{ is standard self-concordant,} \\ |\varphi'(t)| \leq \sqrt{\nu}\varphi''(t)^{1/2}, \\ f(\mathbf{x}) \to +\infty \text{ as } \mathbf{x} \to \partial\Omega \end{cases}$$

Constrained convex problems



$$g^* := \min_{\mathbf{x} \in \Omega} g(\mathbf{x})$$

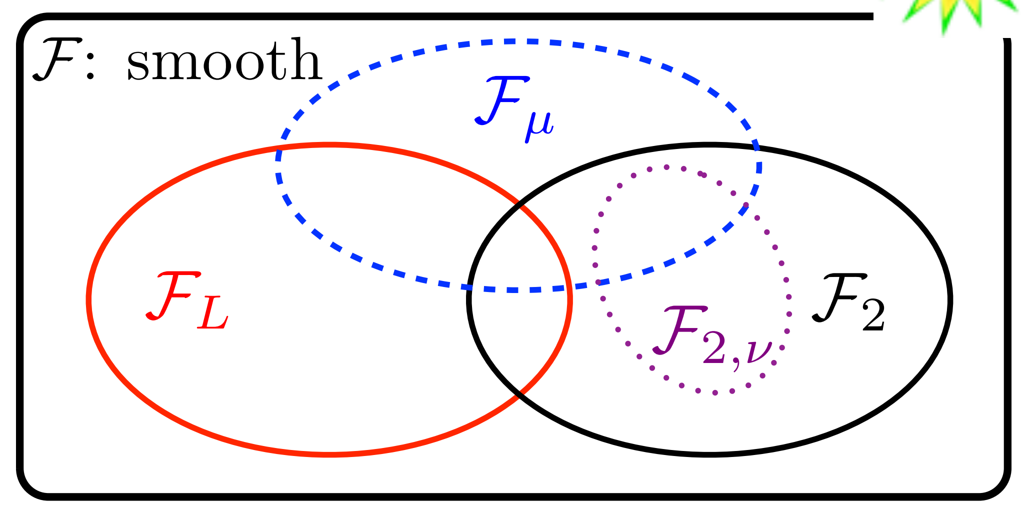
- Ω is endowed with a self-concordant barrier $f(x)$;

- Examples:

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

f is a ν-self-concordant barrier if $\varphi(t) := f(\mathbf{x} + t\mathbf{d})$ satisfies $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$ and $|\varphi'(t)| \leq \sqrt{\nu}\varphi''(t)^{1/2}$

Constrained convex problems



$$g^* := \min_{\mathbf{x}\in\Omega} g(\mathbf{x})$$

- Ω is endowed with a self-concordant barrier $f(x)$;

- Examples:

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

- Main idea: ***solve a sequence of composite self-concordant problems***

$$\min_{\mathbf{x}\in\mathrm{int}(\Omega)} \left\{ F(\mathbf{x}, \tau) := f(\mathbf{x}) + \tau^{-1} g(\mathbf{x}) \right\}$$

Proximal Point + Interior Point

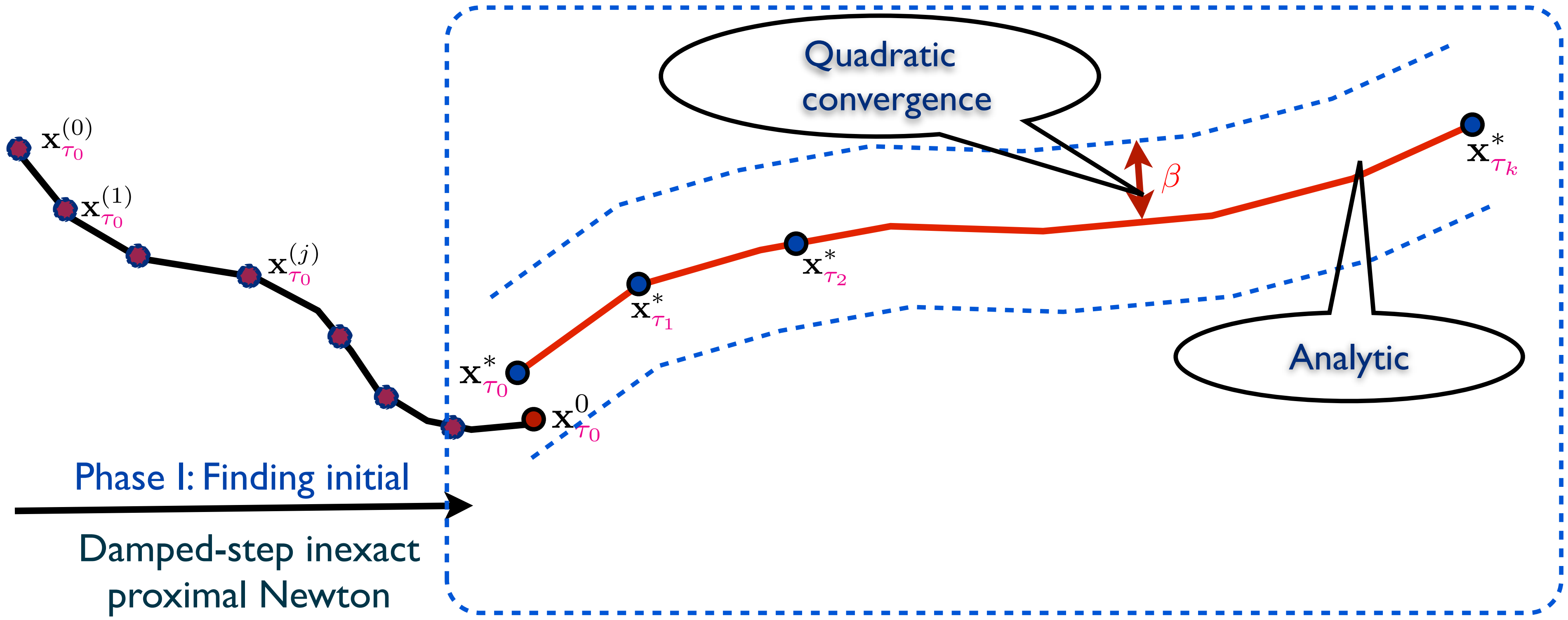
combine them together!!!

f is a ν-self-concordant barrier if $\varphi(t) := f(\mathbf{x} + t\mathbf{d})$ satisfies $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$ and $|\varphi'(t)| \leq \sqrt{\nu}\varphi''(t)^{1/2}$

How does it work?

Main idea: solve a sequence of composite self-concordant problems as opposed to DCO

$$\min_{\mathbf{x}\in\mathrm{int}(\Omega)} \left\{ F(\mathbf{x},\tau) := f(\mathbf{x}) + \tau^{-1} g(\mathbf{x}) \right\}$$



One iteration **k** requires two updates **simultaneously**:

- Update the penalty parameter:

$$t_{k+1} := (1-\sigma_k)t_k, \quad \sigma_k \in [\underline{\sigma}, 1) \quad (\text{e.g.}, \underline{\sigma} = 0.0337/\sqrt{\nu}).$$

- Update the iterative vector (*can be solved approximately*):

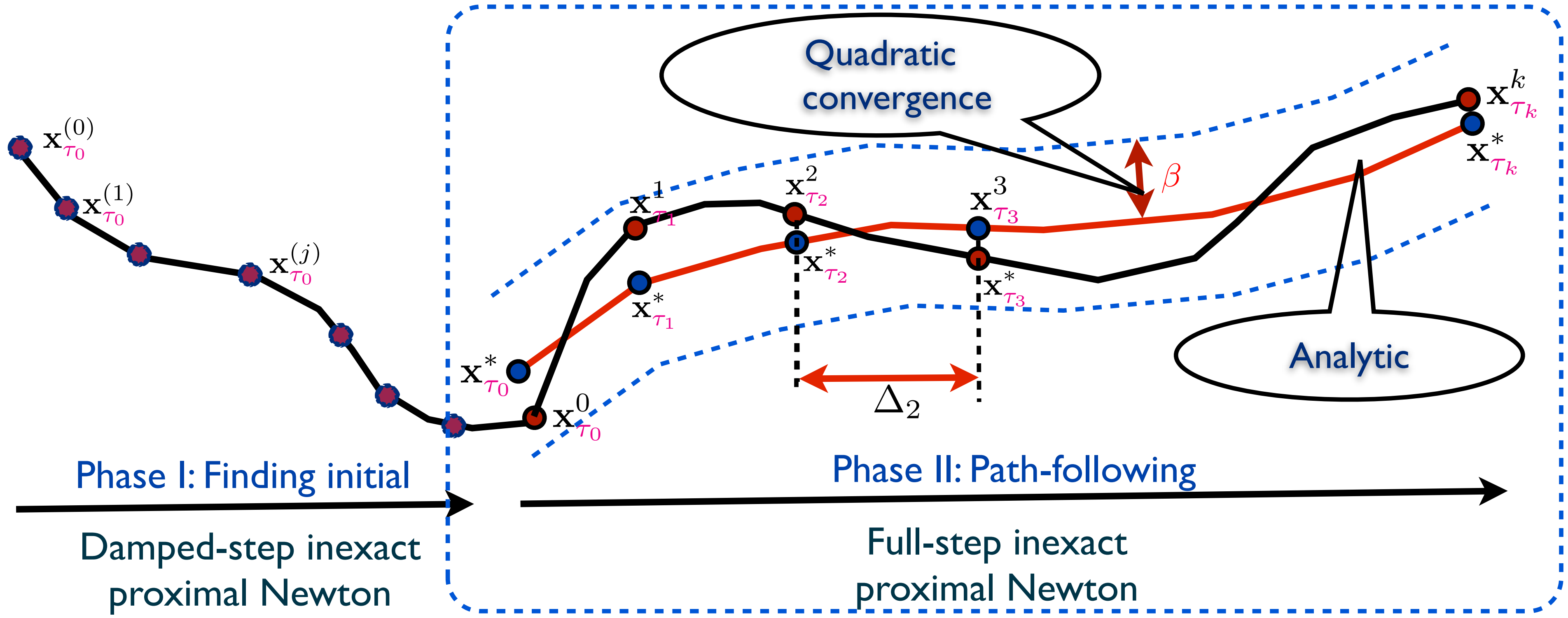
$$\mathbf{x}^{k+1} := \operatorname*{argmin}_{\mathbf{x}} \left\{ t_{k+1}\nabla f(\mathbf{x}^k)^T(\mathbf{x}-\mathbf{x}^k) + \frac{t_{k+1}}{2}(\mathbf{x}-\mathbf{x}^k)^T\nabla^2 f(\mathbf{x}^k)(\mathbf{x}-\mathbf{x}^k) + g(\mathbf{x}) \right\}$$

***Details:** "An Inexact Proximal Path-Following Algorithm for Constrained Convex Minimization," *optimization-online and lions.epfl.ch/publications.*

How does it work?

Main idea: solve a sequence of composite self-concordant problems as opposed to DCO

$$\min_{\mathbf{x}\in\mathrm{int}(\Omega)} \left\{ F(\mathbf{x},\tau) := f(\mathbf{x}) + \tau^{-1} g(\mathbf{x}) \right\}$$



One iteration **k** requires two updates **simultaneously**:

- Update the penalty parameter:

$$t_{k+1} := (1-\sigma_k)t_k, \quad \sigma_k \in [\underline{\sigma}, 1) \quad (\text{e.g.}, \underline{\sigma} = 0.0337/\sqrt{\nu}).$$

- Update the iterative vector (*can be solved approximately*):

$$\mathbf{x}^{k+1} := \operatorname*{argmin}_{\mathbf{x}} \left\{ t_{k+1}\nabla f(\mathbf{x}^k)^T(\mathbf{x}-\mathbf{x}^k) + \frac{t_{k+1}}{2}(\mathbf{x}-\mathbf{x}^k)^T\nabla^2 f(\mathbf{x}^k)(\mathbf{x}-\mathbf{x}^k) + g(\mathbf{x}) \right\}$$

Details:** "An Inexact Proximal Path-Following Algorithm for Constrained Convex Minimization," ***optimization-online and lions.epfl.ch/publications.

How does it converge?

The algorithm consists of **two** PHASES:

- **Phase I**: Find a **starting point** $\mathbf{x}^0_{p_2}$ such that: $\|\mathbf{x}^0_{p_2} - \mathbf{x}^*(t_0)\|_{\mathbf{x}^*(t_0)} \leq 0.05$

- **Phase II**: Perform the ***path-following iterations***

Tracking properties on the penalty parameter and the iterative sequence in **Phase II**

- The penalty parameter t_k **decreases** at least with a factor $\tau := 1 - \frac{0.0337}{\sqrt{\nu}}$ at each iteration **k**

$$t_{k+1} = \tau t_k$$

- **Tracking error** of the iterative sequence:

$$\text{if } \boxed{\|\mathbf{x}^k - \mathbf{x}^*(t_k)\|_{\mathbf{x}^*(t_k)} \leq 0.05} \text{ then } \boxed{\|\mathbf{x}^{k+1} - \mathbf{x}^*(t_{k+1})\|_{\mathbf{x}^*(t_{k+1})} \leq 0.05}$$

Worst-case complexity:

- **Phase I**: Finding a starting point for **Phase II** requires at most

$$j_{\max} := \left\lfloor \frac{F(\mathbf{x}^0; t_0) - F(\mathbf{x}^*(t_0); t_0)}{0.0012} \right\rfloor + 1$$

- **Phase II**: The **worst-case complexity** to reach an ε**- solution** is at most:

$$\mathcal{O}\left(\sqrt{\nu}\log\left(\frac{Ct_0}{\varepsilon}\right)\right)$$

- **Note:** This worst-case complexity is as the **same** as in *standard path-following methods* [see Nesterov2004]

Proximal path-following **Upshot: no-heavy lifting!**

Proximal path following for conic programming with rigorous guarantees

$$g^* := \min_{\mathbf{x}\in\Omega} g(\mathbf{x})$$

- Ω is endowed with a self-concordant barrier $f(x)$;

Example: Low-rank SDP matrix approximation ...

$$\begin{array}{ll}\min\limits_{\mathbf{X}} & \rho\|\mathrm{vec}(\mathbf{X}-\mathbf{M})\|_1 + (1-\rho)\mathrm{tr}(\mathbf{X}) \\ \text{s.t.} & \mathbf{X}\succeq 0,\ \mathbf{L}_{ij}\leq \mathbf{X}_{ij}\leq \mathbf{U}_{ij},\ i,j=1,\ldots,n.\end{array}$$

ρ is a regularization parameter in (0, 1), $\mathbf{M}$ is the given input matrix.

	Solver\ n	80	100	120	140	160
Size	$[n_v; n_c]$	[16,200; 9,720]	[25,250; 15,150]	[36,300; 21,780]	[49,350; 29,610]	[64,400; 38,640]
Time (sec)	New Algorithm	15.738	24.046	24.817	25.326	36.531
	SDPT3	156.340	508.418	881.398	1742.502	2948.441
	SeDuMi	231.530	970.390	3820.828	9258.429	17096.580
Objective value	New Algorithm	306.9159	497.6706	635.4304	842.4626	1096.6516
	SDPT3	306.9153	497.6754	635.4306	842.4644	1096.6540
	SeDuMi	306.9176	497.6821	635.4384	842.4776	1096.6695

Proximal path-following **Upshot: no-heavy lifting!**

Proximal path following for conic programming with rigorous guarantees

$$g^* := \min_{\mathbf{x} \in \Omega} g(\mathbf{x})$$

- Ω is endowed with a self-concordant barrier $f(x)$;

Example: Low-rank SDP matrix approximation ...

$$\begin{array}{ll} \min_{\mathbf{X}} & \rho \|\mathrm{vec}(\mathbf{X} - \mathbf{M})\|_1 + (1-\rho)\mathrm{tr}(\mathbf{X}) \\ \text{s.t.} & \mathbf{X} \succeq 0, \ \mathbf{L}_{ij} \leq \mathbf{X}_{ij} \leq \mathbf{U}_{ij}, \ i, j = 1, \dots, n. \end{array}$$

#variables
#constraints

ρ is a regularization parameter in (0, 1), **M** is the given input matrix.

	Solver\\n	80	100	120	140	160
Size	$[n_v; n_c]$	[16,200; 9,720]	[25,250; 15,150]	[36,300; 21,780]	[49,350; 29,610]	[64,400; 38,640]
Time (sec)	New Algorithm	15.738	24.046	24.817	25.326	36.531
	SDPT3	156.340	508.418	881.398	1742.502	2948.441
	SeDuMi	231.530	970.390	3820.828	9258.429	17096.580
Objective value	New Algorithm	306.9159	497.6706	635.4304	842.4626	1096.6516
	SDPT3	306.9153	497.6754	635.4306	842.4644	1096.6540
	SeDuMi	306.9176	497.6821	635.4384	842.4776	1096.6695

+ **Lift** the problem dimension (small to large-scale)

+ **Destroy** the sparsity due to **scaling-factors** in SDP (e.g., Nesterov-Todd scaling factor)

Proximal path-following **Upshot: desired scaling!**

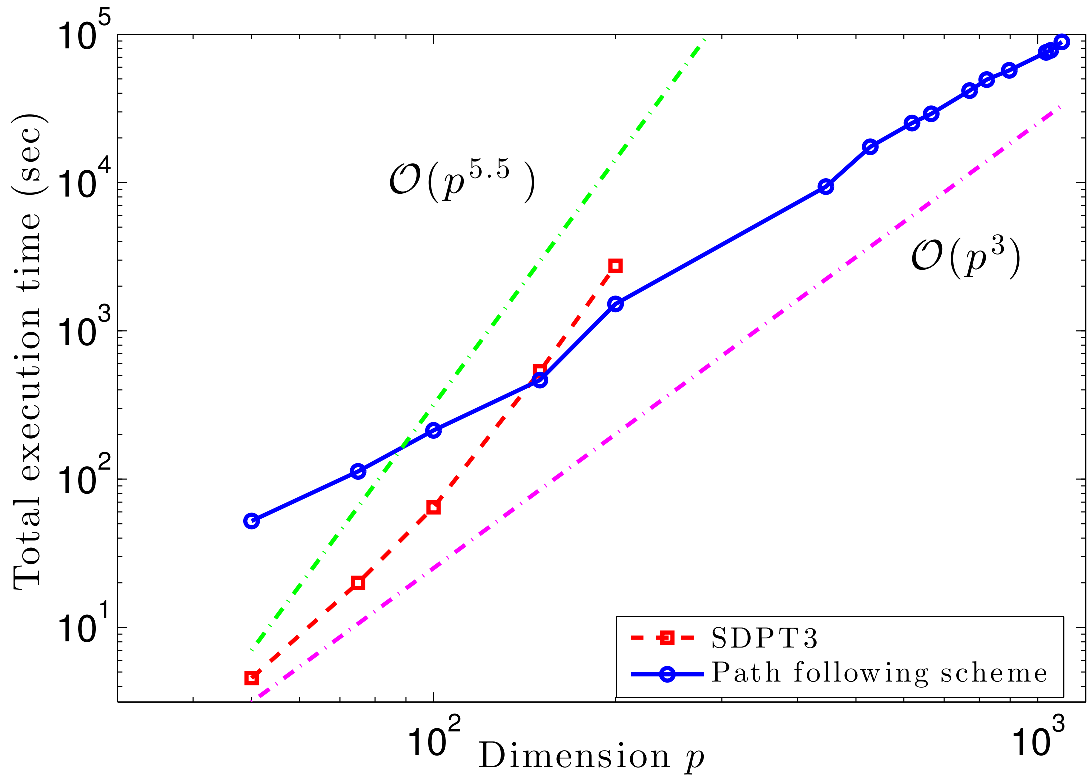
Proximal path following for conic programming with rigorous guarantees

$$g^* := \min_{\mathbf{x} \in \Omega} g(\mathbf{x})$$

- Ω is endowed with a self-concordant barrier $f(x)$;

Example: 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

	p	50	75	100	150	200
Time (sec)	PF	62.450	109.426	202.600	416.044	1573.881
	SDPT3	4.396	21.282	64.939	522.021	2588.721
	splitting	102.217	236.366	354.444	778.904	1420.844
$g(\mathbf{K}^*)$	PF	549.1567	1293.6727	2232.5897	5396.0485	9809.6066
	SDPT3	549.1860	1293.7890	2233.0747	5396.7305	9809.6934
	splitting	597.8825	1387.1379	2496.6535	5583.8605	9958.0974

Proximal path-following **Upshot: auto. regularization!**

Proximal path following for conic programming with rigorous guarantees

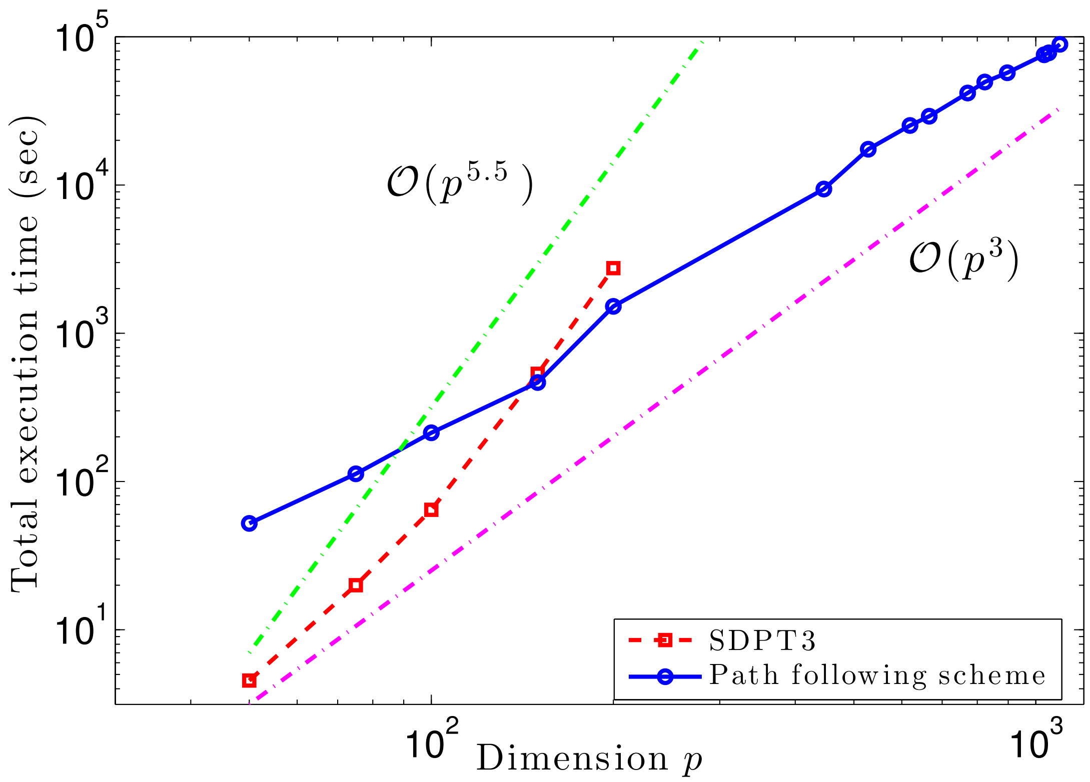
$$g^* := \min_{\mathbf{x} \in \Omega} g(\mathbf{x})$$

- Ω is endowed with a self-concordant barrier $f(x)$;

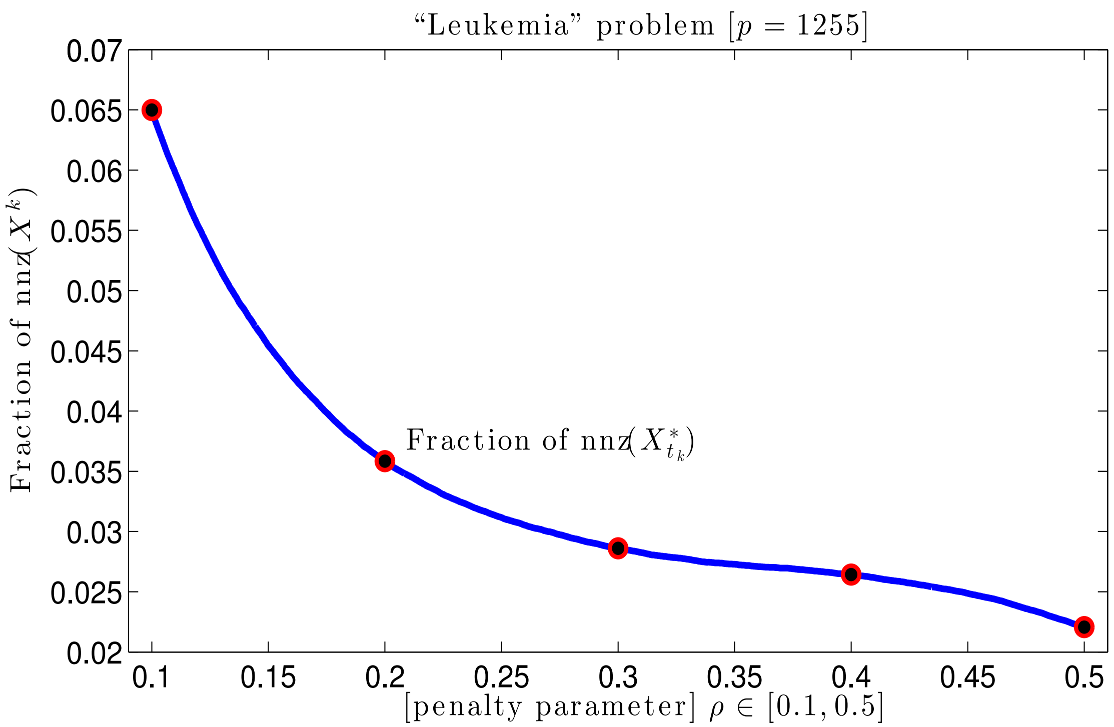
- Main idea: solve a sequence of composite self-concordant problems as opposed to DCO

$$\min_{\mathbf{x} \in \mathrm{int}(\Omega)} \{F(\mathbf{x}; t) := g(\mathbf{x}) + t f(\mathbf{x})\}$$

Example: Max-norm clustering



Example: Graph selection



Smoothing via self-concordant barrier

Composite convex minimization with nonsmooth-max structure **f**:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$

- Nonsmooth function **f** is defined as:

$$f(\mathbf{x}) := \max_{\mathbf{u}\in\mathcal{U}} \{\langle \mathbf{A}^T\mathbf{x}, \mathbf{u}\rangle - G(\mathbf{u})\}$$

- $\mathcal{U}$: endowed with a **self-concordant barrier** $b_{\mathcal{U}}$, **G**: **convex**

Motivating examples

- *Nonsmooth convex function* induced by a **norm**:

$$f(\mathbf{x}) := \|\mathbf{D}\mathbf{x}\| = \max_{\mathbf{u}} \{\langle \mathbf{D}\mathbf{x}, \mathbf{u}\rangle \mid \mathbf{u} \in \mathcal{U} := \{\mathbf{v} \mid \|\mathbf{v}\|^* \leq 1\}\}$$

- Fenchel's ***dual function*** with compact domain:

$$f(\mathbf{x}) := \max_{\mathbf{u}\in\mathrm{dom} f^*} \{\langle \mathbf{x}, \mathbf{u}\rangle - f^*(\mathbf{u})\}$$

- ***Lagrange dual problem*** in *constrained convex optimization problem* $\max_{\mathbf{u}\in\mathcal{U}} \{-G(\mathbf{u}) \mid \mathbf{A}\mathbf{u} = b\}$

$$\boxed{\min_{\mathbf{x}} \left\{ \underbrace{\max_{\mathbf{u}\in\mathcal{U}} \{\mathbf{x}^T\mathbf{A}\mathbf{u} - G(\mathbf{u})\}}_{f(\mathbf{x})} \underbrace{-b^T\mathbf{x}}_{g(\mathbf{x})} \right\}}$$

Smoothing via self-concordant barrier

Composite convex minimization with nonsmooth-max structure **f**:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$

- Nonsmooth function **f** is defined as:

$$f(\mathbf{x}) := \max_{\mathbf{u}\in\mathcal{U}} \left\{\langle \mathbf{A}^T\mathbf{x}, \mathbf{u}\rangle - G(\mathbf{u})\right\}$$

- $\mathcal{U}$: endowed with a **self-concordant barrier** $b_{\mathcal{U}}$, **G**: **convex**

Self-concordant barrier smoother of f*

$$f_\sigma(\mathbf{x}) := \max_{\mathbf{u}\in\mathrm{int}(\mathcal{U})} \left\{\langle \mathbf{A}^T\mathbf{x}, \mathbf{u}\rangle - G(\mathbf{u}) - \sigma b_{\mathcal{U}}(\mathbf{u})\right\}$$

- $\sigma > 0$ is a *smoothness parameter*

- The solution $\mathbf{u}^*{}_\sigma(\mathrm{x})$ of the maximization problem satisfies the following optimality condition:

$$0 \in \mathbf{A}^T\mathbf{x} - \partial G(\mathbf{u}^*{}_\sigma(\mathbf{x})) - \sigma\nabla b_{\mathcal{U}}(\mathbf{u}^*{}_\sigma(\mathbf{x}))$$

- This is a **generalized equation**. *If **G** is **smooth** then it reduces to a system of nonlinear equations*

Often easier to solve than general convex problems

The **gradient** of f_σ is given as

$$\nabla f_\sigma(\mathbf{x}) = \mathbf{A}\mathbf{u}^*{}_\sigma(\mathbf{x})$$

Details: Manuscript under review for ICASSP'14.

Three key properties of the barrier smoother

1. Approximation property

$$f_\sigma(\mathbf{x}) \leq f(\mathbf{x}) \leq f_\sigma(\mathbf{x}) + \sigma C\nu.$$

- $\sigma > 0$ is the *smoothness parameter, C is a given constant and* ν *is the barrier parameter.*

2. Local Lipschitz-like property

$$\|\nabla f_\sigma(\mathbf{x}) - \nabla f_\sigma(\mathbf{y})\|_2 \leq \frac{c_A^2\|\mathbf{x}-\mathbf{y}\|_2}{\sigma - c_A\|\mathbf{x}-\mathbf{y}\|_2}, \quad \forall \mathbf{x},\mathbf{y} \;\; \|\mathbf{x}-\mathbf{y}\|_2 < \frac{\sigma}{c_A}$$

- Here: $c_A \equiv c_A(\mathbf{x}) = \left[\|\mathbf{A}\nabla^2 b_{\mathcal{U}}(\mathbf{u^*}_\sigma(\mathbf{x}))\mathbf{A}^T\|\right]_2^{1/2} \leq \bar{c}_A \;\; (\text{constant})$

3. Existence of the second order derivative

- If **G** is ***twice differentiable*** and A is ***full-row rank***, then

$$\nabla^2 f_\sigma(\mathbf{x}) = \mathbf{A}\left(\nabla^2 G(\mathbf{u^*}_\sigma(\mathbf{x})) + \sigma\nabla^2 b_{\mathcal{U}}(\mathbf{u^*}_\sigma(\mathbf{x}))\right)^{-1}\mathbf{A}^T \succ 0$$

- If, in addition, **G** is **self-concordant** with the parameter M_G , then f_σ is also **self-concordant** with

$$M_{f_\sigma} = \max\left\{M_G, \frac{2}{\sqrt{\sigma}}\right\}$$

First-order method for barrier smoothing

Barrier smoothed problem

$$\phi_\sigma^* = \min_{\mathbf{x}\in\mathbb{R}^n} \{\phi_\sigma(\mathbf{x}) := f_\sigma(\mathbf{x}) + g(\mathbf{x})\}$$

Here, f_σ is **smooth** and **g** is ***"tractably" proximal***

Proximal - gradient scheme

$$\begin{cases} \mathbf{s}^k & := \mathrm{prox}^g{}_{\gamma_k}\left(\mathbf{x}^k - \gamma_k \nabla f_\sigma(\mathbf{x}^k)\right), \\ \mathbf{d}^k & := \mathbf{s}^k - \mathbf{x}^k, \\ \mathbf{x}^{k+1} & := \mathbf{x}^k + \alpha_k \mathbf{d}^k. \end{cases}$$

- **Step-sizes:** $\gamma_k \geq \frac{\sigma}{\bar{c}_A^2}$ and $\alpha_k = \frac{\sigma}{c_A^k\left(c_A^k\gamma_k + \lambda_k\right)} \in (0, 1]$ with $\lambda_k = \|\mathbf{d}^k\|_2,\ c_A^k = c_A(\mathbf{x}^k)$

Convergence and convergence rate

- **Descent** property: $\boxed{\phi_\sigma(\mathbf{x}^{k+1}) \leq \phi_\sigma(\mathbf{x}^k) - \sigma\omega\left(\frac{\lambda_k}{c_A^k}\right),}$ $\omega(\tau) = \tau - \log(1+t).$

- **Convergence rate** (in the ergodic sense): ***If*** $\lambda_0 \leq M\bar{c}_A$ ***and*** $\bar{\mathbf{x}}^k = \left(\sum_{j=0}^k \alpha_j\right)^{-1} \sum_{j=0}^k \alpha_j \mathbf{x}^j$ ***then***

$$\boxed{\phi_\sigma(\bar{\mathbf{x}}^k) - \phi_\sigma(\mathbf{x}_\sigma^*) \leq \frac{(1+M)\bar{c}_A^2}{2\sigma k}\|\mathbf{x}^0 - \mathbf{x}_\sigma^*\|_2^2} \longleftarrow \boxed{\mathcal{O}\left(\frac{\bar{c}_A^2}{\sigma k}\|\mathbf{x}^0 - \mathbf{x}_\sigma{}^*\|_2^2\right)}$$

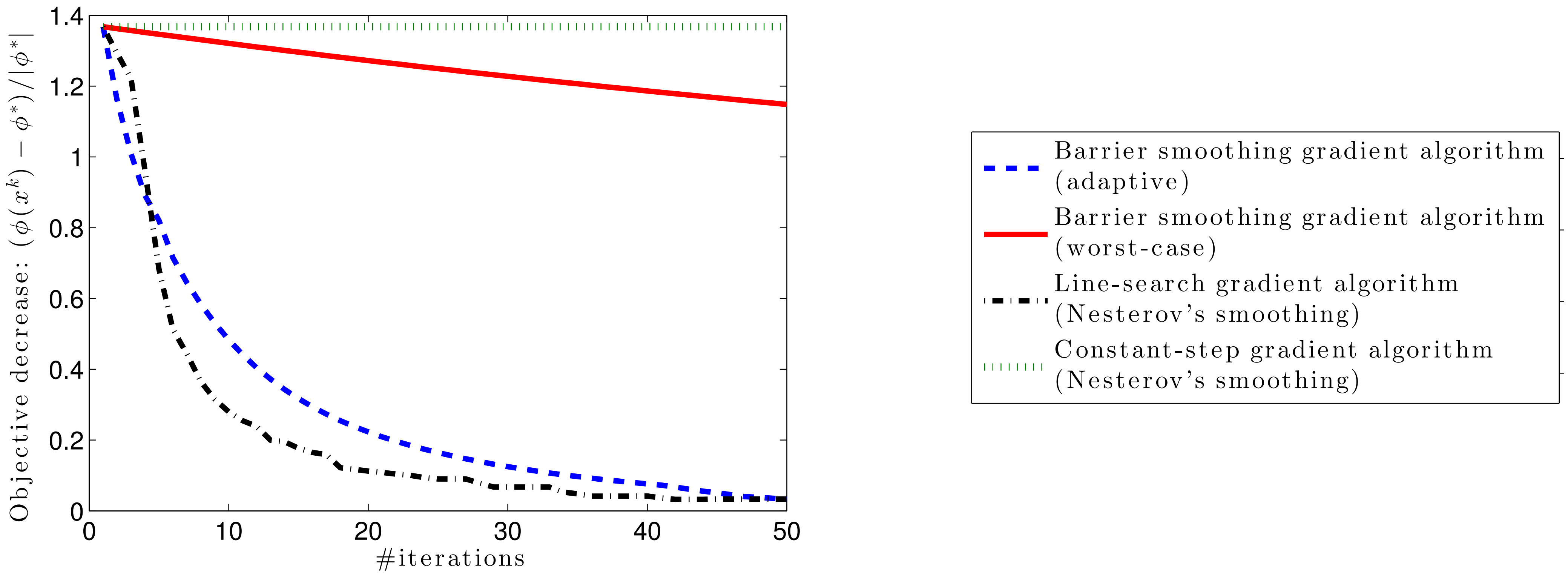
A stylized example

A simple quadratically constrained quadratic program

$$\phi^* = \min_{\mathbf{x}} \Big\{ \phi(\mathbf{x}) := \underbrace{\max_{\mathbf{u}^T\mathbf{B}\mathbf{u}\le 1} \Big\{ (\mathbf{A}\mathbf{x} - b)^T\mathbf{u} - \frac{1}{2}\mathbf{x}^T Q\mathbf{u} \Big\}}_{f(\mathbf{x})} + \underbrace{c^T\mathbf{x}}_{g(\mathbf{x})} \Big\}$$

- *Self-concordant barrier* for $\mathcal{U} := \{\mathbf{u} \in \mathbb{R}^m \mid \mathbf{u}^T\mathbf{B}\mathbf{u} \le 1\}$ is $b_{\mathcal{U}}(\mathbf{u}) := -\log(1 - \mathbf{u}^T\mathbf{B}\mathbf{u})$ with $\nu = 2$

Convergence of Nesterov's smoothing vs. Barrier smoothing (m = 200, n = 60) after **50** iterations



- Line-search ***gradient method*** *based on Nesterov's smoothing with proximity function* $p(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{B}\mathbf{x}$

- ***Gradient method*** *based on barrier smoothing with the* ***worst-case step size*** *using* $\bar{c}_A$
- ***Gradient method*** *based on barrier smoothing with the* ***adaptive step size*** *using* c_A^k

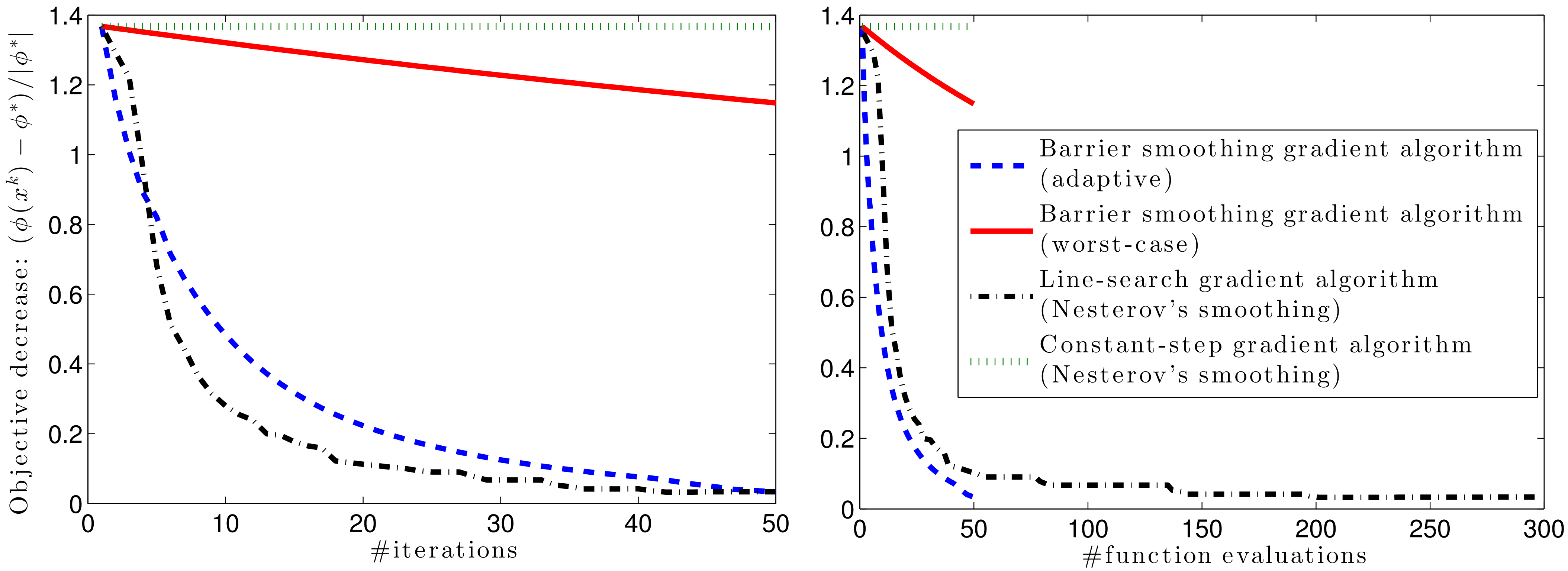
A stylized example

A simple quadratically constrained quadratic program

$$\phi^* = \min_{\mathbf{x}} \left\{ \phi(\mathbf{x}) := \underbrace{\max_{\mathbf{u}^T\mathbf{B}\mathbf{u}\leq 1} \left\{ (\mathbf{A}\mathbf{x} - b)^T\mathbf{u} - \frac{1}{2}\mathbf{x}^T Q\mathbf{u} \right\}}_{f(\mathbf{x})} + \underbrace{c^T\mathbf{x}}_{g(\mathbf{x})} \right\}$$

- *Self-concordant barrier* for $\mathcal{U} := \left\{\mathbf{u} \in \mathbb{R}^m \mid \mathbf{u}^T\mathbf{B}\mathbf{u} \leq 1\right\}$ is $b_{\mathcal{U}}(\mathbf{u}) := -\log(1 - \mathbf{u}^T\mathbf{B}\mathbf{u})$ with $\nu = 2$

Convergence of Nesterov's smoothing vs. Barrier smoothing (m = 200, n = 60) after **50** iterations



- Line-search ***gradient method*** *based on Nesterov's smoothing with proximity function* $p(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{B}\mathbf{x}$

  Line-search *requires additional function evaluations. The number of function evaluations is* ***296*** *after 50 iterations*

- ***Gradient method*** *based on barrier smoothing with the* ***worst-case step size*** *using* $\bar{c}_A$
- ***Gradient method*** *based on barrier smoothing with the* ***adaptive step size*** *using* c_A^k

Second order method for barrier smoothing

When to apply proximal-Newton?

G is **self-concordant** and **A** is **full-row rank**

Proximal-Newton scheme

$$\begin{cases} \mathbf{s}^k & := \mathrm{argmin}_{\mathbf{x}} \left\{ Q(\mathbf{x}; \mathbf{x}^k) + g(\mathbf{x}) \right\}, \\ \mathbf{d}^k & := \mathbf{s}^k - \mathbf{x}^k, \\ \alpha_k & := (1 + \lambda_k)^{-1}, \ \ \lambda_k := \sqrt{\sigma^{-1}} \|\mathbf{d}^k\|_{\mathbf{x}^k}, \\ \mathbf{x}^{k+1} & := \mathbf{x}^k + \alpha_k \mathbf{d}^k. \end{cases}$$

- **Q** is the **quadratic surrogate** of f_σ around $\mathbf{x}^k$:

$$Q(\mathbf{x}; \mathbf{x}^k) := \nabla f_\sigma(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}^k)^T \nabla^2 f_\sigma(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k)$$

Convergence

- Maintain ***local quadratic convergence*** as in *proximal-Newton method* for *self-concordant minimization*

- The worst-case complexity depends on the smoothness parameter:

$$\#\text{iterations} = \left\lfloor \frac{\phi_\sigma(\mathbf{x}^0) - \phi_\sigma{}^*}{0.017\sigma} \right\rfloor + \mathcal{O}\left(1.5 \ln\ln\left(\frac{0.28}{\sigma\varepsilon}\right)\right) + 2$$

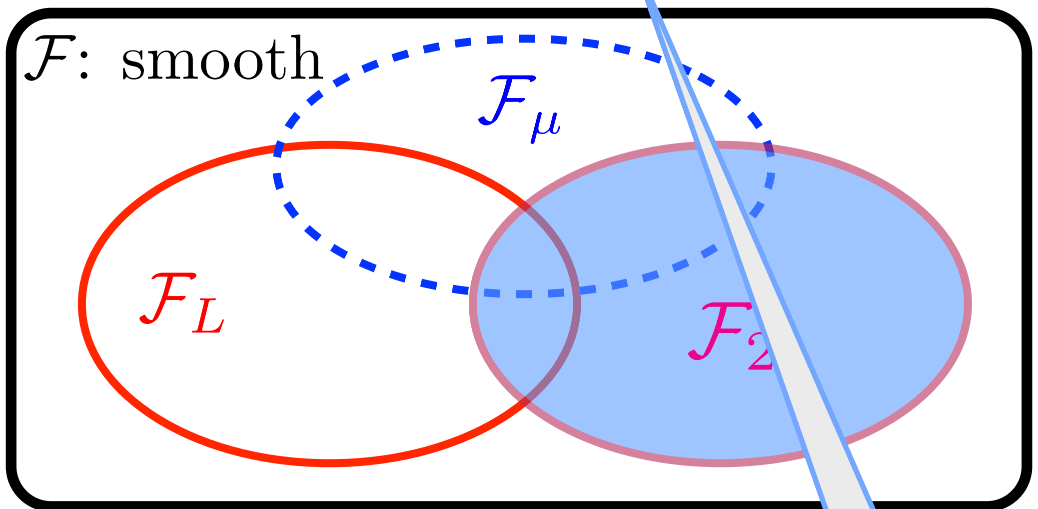
- Possible to apply the **path-following proximal Newton scheme** for tuning $\sigma \downarrow 0^+$

Simultaneously update both σ and $\mathbf{x}^k$ at each iteration **k**

Conclusions

Conclusions

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\right\}$$

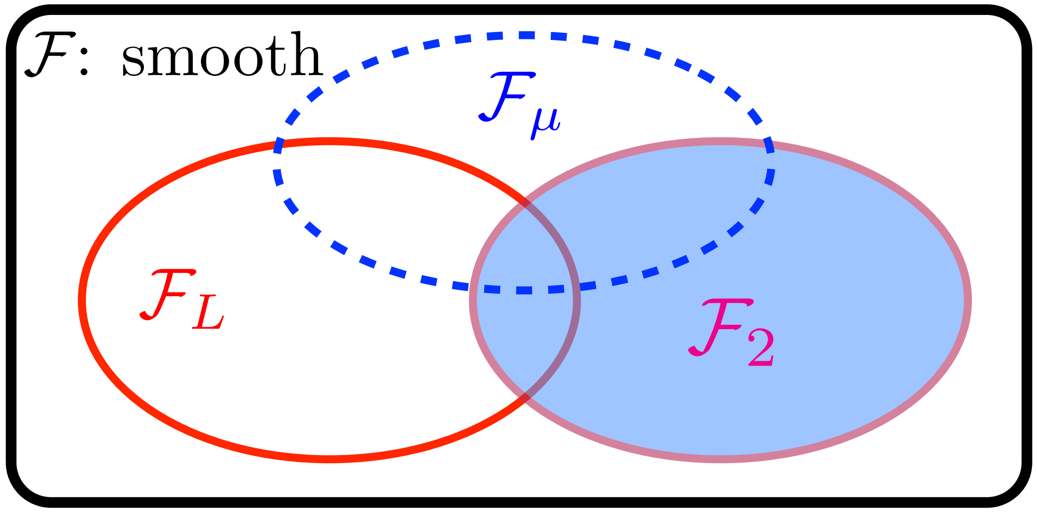


- $\mathcal{F}_\mu$ - μ-strongly convex
- $\mathcal{F}_L$ - L-Lipschitz gradient
- $\mathcal{F}_2$ - self-concordant

A new variable metric proximal-point framework for composite self-concordant minimization
+
Extensions

Conclusions

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$



- $\mathcal{F}_\mu$ - μ-strongly convex
- $\mathcal{F}_L$ - L-Lipschitz gradient
- $\mathcal{F}_2$ - self-concordant

- Highlights

- **Globalization**: a new strategy for finding step-size **explicitly**

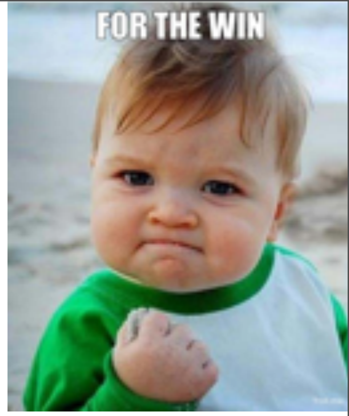
  motivate "*forward-looking*" line-search strategy

- **Search direction**: efficient (strongly convex program)

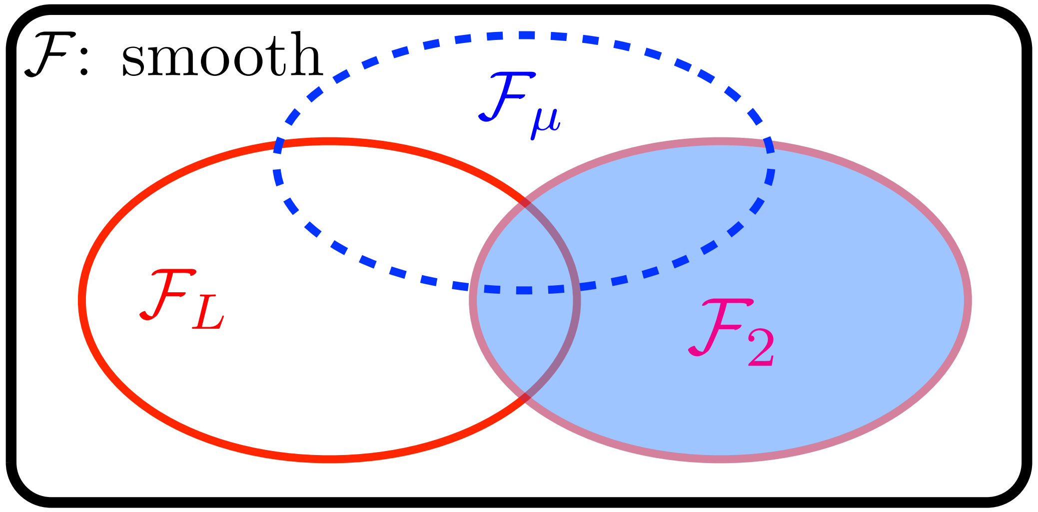
- **Local convergence**: quadratic convergence *without boundedness of the Hessian*

  analytic quadratic convergence region

Conclusions



$$\min_{\mathbf{x}\in\mathbb{R}^n} \{\phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}$$



- $\mathcal{F}_\mu$ - μ-strongly convex
- $\mathcal{F}_L$ - L-Lipschitz gradient
- $\mathcal{F}_2$ - self-concordant

- Highlights

- **Globalization**: a new strategy for finding step-size **explicitly**
  motivate "*forward-looking*" line-search strategy
- **Search direction**: efficient (strongly convex program)
- **Local convergence**: quadratic convergence *without boundedness of the Hessian*
  analytic quadratic convergence region

- Practical contributions (this talk)
  - SCOPT package has quasi-Newton / first & second order methods @lions.epfl.ch/software
  - leverage fast proximal solvers for g(x) (structured norms etc.)
  - robust to subproblem solver accuracy

SCOPT FTW

- Postdoc & PhD positions @ LIONS / EPFL

contact: volkan.cevher@epfl.ch