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
Composite minimization

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

$f$ 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

$g$: $\ell_1$-norm, nuclear norm or indicator functions
Composite minimization

\[(P) \quad \min_{x \in \mathbb{R}^n} \{ \phi(x) := f(x) + g(x) \}\]

\[f\] convex and smooth \hspace{1cm} \[g\] 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**

\[g: \ell_1\text{-norm, nuclear norm or indicator functions}\]
Composite minimization: modus operandi

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

\( f \) convex and smooth
\( g \) convex and possibly nonsmooth

Classes of smooth functions (\( f \))

\( \mathcal{F} \): smooth
\( \mathcal{F}_L \) - L-Lipschitz gradient
\( \mathcal{F}_\mu \) - \( \mu \)-strongly convex

\[
\| \nabla f(x) - \nabla f(y) \| \leq L \| y - x \|
\]

\[
\mu \mathbb{I} \leq \nabla^2 f(x) \leq L \mathbb{I}
\]
Composite minimization: modus operandi

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

\( f \) convex and smooth

\( g \) convex and possibly nonsmooth

Classes of smooth functions (\( f \))

- \( \mathcal{F}_L \) - L-Lipschitz gradient
- \( \mathcal{F}_\mu \) - \( \mu \)-strongly convex

\[ ||\nabla f(x) - \nabla f(y)|| \leq L ||y - x|| \]

\[ \mu \| x \| \leq \nabla^2 f(x) \leq L \| x \| \]
Composite minimization: modus operandi

\[(P) \quad \min_{x \in \mathbb{R}^n} \{ \phi(x) := f(x) + g(x) \} \]

\[\text{\textbf{f}}\] convex and smooth

\[\text{\textbf{g}}\] convex and possibly nonsmooth

Classes of smooth functions (\textbf{f})

\[\mathcal{F}: \text{smooth} \qquad \mathcal{F}_\mu \qquad \mathcal{F}_L\]

- \(\mathcal{F}_L\) - \(L\)-Lipschitz gradient
- \(\mathcal{F}_\mu\) - \(\mu\)-strongly convex

\[\|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\|\]

\[\mu I \leq \nabla^2 f(x) \leq L I\]

Following \textbf{prox} computation is \textit{tractable}:

\[\text{prox}_{\gamma g}(s) := \arg \min_x \left\{ g(x) + \frac{1}{2\gamma} \|x - s\|^2 \right\}\]
Composite minimization: modus operandi

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

\( f \) convex and smooth

\( g \) convex and possibly nonsmooth

**Classes of smooth functions (f)**

- \( \mathcal{F}_L \) - L-Lipschitz gradient
- \( \mathcal{F}_\mu \) - \( \mu \)-strongly convex

\[ \| \nabla f(x) - \nabla f(y) \| \leq L \| y - x \| \]

\[ \mu \| \| \leq \nabla^2 f(x) \leq L \| \]

Following \( \text{prox} \) computation is **tractable**: 

\[ \text{prox}_{\gamma g}(s) := \arg \min_x \left\{ g(x) + \frac{1}{2\gamma} \| x - s \|^2 \right\} \]

**Example:**

If \( g(x) = \| x \|_1 \), then \( \text{prox}_{\gamma g}(s) = \text{SoftThresh}(s, \gamma) \)
Composite minimization: modus operandi

\[(P) \quad \min_{x \in \mathbb{R}^n} \{ \phi(x) := f(x) + g(x) \} \]

\[f \text{ convex and smooth} \quad \quad g \quad \text{convex and possibly nonsmooth} \quad \text{with “tractable” prox}\]

Classes of smooth functions \((f)\)

\[\mathcal{F}: \text{smooth} \]

\[\mathcal{F}_L - L\text{-Lipschitz gradient} \quad \quad \mathcal{F}_\mu - \mu\text{-strongly convex} \]

\[\| \nabla f(x) - \nabla f(y) \| \leq L \| y - x \| \]

\[\mu \mathbb{I} \leq \nabla^2 f(x) \leq L \mathbb{I} \]

Fast gradient schemes (Nesterov’s methods)

Newton/quasi Newton schemes
Composite minimization: an uncharted region

\[(P) \quad \min_{x \in \mathbb{R}^n} \{ \phi(x) := f(x) + g(x) \}\]

\(f\) convex and smooth

\(g\) convex and possibly nonsmooth with “tractable” prox

Classes of smooth functions \((f)\)

\(\mathcal{F}:\) smooth

- \(\mathcal{F}_L\) - \(L\)-Lipschitz gradient
- \(\mathcal{F}_\mu\) - \(\mu\)-strongly convex

\[\|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\|\]

\[\mu \|x\| \leq \nabla^2 f(x) \leq L\|x\|\]

Scalability is NOT great

Fast gradient schemes (Nesterov’s methods)

Newton/quasi-Newton schemes
Composite **self-concordant** minimization

\[(P) \quad \min_{x \in \mathbb{R}^n} \{ \phi(x) := f(x) + g(x) \}\]

- \(f\) convex and **self-concordant**
- \(g\) convex and possibly **nonsmooth** with "tractable" prox

**Classes of smooth functions** (**f**)

\[ \mathcal{F}: \text{smooth} \]

\[ \mathcal{F}_L \]

\[ \mathcal{F}_\mu \]

\[ \mathcal{F}_2 \]

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

**Key structure for the interior point method**

\(f\) is self-concordant if \(\phi(t) := f(x + td)\) satisfies \(|\phi'''(t)| \leq 2\phi''(t)^{3/2}\)
Example: Log-determinant for LMIs

**Application:** Graphical model selection

Given a data set \( \mathcal{D} := \{x_1, \ldots, x_N\} \), where \( x_i \) is a Gaussian random variable. Let \( \Sigma \) be the **covariance matrix** corresponding to the **graphical model** of the Gaussian Markov random field. The aim is to learn a **sparse matrix** \( \Theta \) that approximates the inverse \( \Sigma^{-1} \).

**Optimization problem**

\[
\min_{\Theta > 0} \left\{ - \log \det(\Theta) + \text{trace}(\Sigma \Theta) + \rho \|\text{vec}(\Theta)\|_1 \right\}
\]

\(f(x)\)

\(g(x)\)
**Log-barrier for linear/quadratic inequalities**

- Poisson imaging reconstruction via TV regularization

\[ x^* \in \arg \min_x \left\{ \sum_{i=1}^{m} a_i^T x - \sum_{i=1}^{m} y_i \log(a_i^T x + b_i) + g(x) \right\} \]

- Basic pursuit denoising problem (BPDP): Barrier formulation

\[ x_t^* = \arg \min_x \left\{ -t \log\left( \sigma^2 - \| Ax - y \|^2_2 \right) + g(x) \right\} =: f(x) \]

- LASSO problem with unknown variance

\[ x^* \equiv (\phi^*, \gamma^*) = \arg \min_{\phi, \gamma} \left\{ -\log(\gamma) + \frac{1}{2n} \| \gamma y - X\phi \|^2_2 + \lambda \| \phi \|_1 \right\} =: f(x) \]

- Quantum tomography ML estimator *(another presentation!)*
Composite **self-concordant** minimization

\[
(P) \quad \min_{x \in \mathbb{R}^n} \{ \phi(x) := f(x) + g(x) \}
\]

- **f** convex and **self-concordant**
- **g** convex and possibly **nonsmooth** with "tractable" prox

**Classes of smooth functions (f)**

\[ \mathcal{F}: \text{smooth} \quad \mathcal{F}_L \quad \mathcal{F}_2 \quad \mathcal{F}_\mu \]

**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

\[ f \text{ is self-concordant if } \varphi(t) := f(x + td) \text{ satisfies } |\varphi'''(t)| \leq 2\varphi''(t)^{3/2} \]
Basic algorithmic framework
A basic composite minimization framework

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

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

**Main properties of** \( \mathcal{F}_{\mu,L} \)

<table>
<thead>
<tr>
<th>Property</th>
<th>Inequality</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower surrogate</td>
<td>( f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} |y - x|_2^2 )</td>
<td>( x, y \in \text{dom}(f) )</td>
</tr>
<tr>
<td>Upper surrogate</td>
<td>( f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} |y - x|_2^2 )</td>
<td>( x, y \in \text{dom}(f) )</td>
</tr>
<tr>
<td>Hessian surrogates</td>
<td>( \mu I \preceq \nabla^2 f(x) \preceq L I )</td>
<td>( x \in \text{dom}(f) )</td>
</tr>
</tbody>
</table>
A basic composite minimization framework

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

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

\[ \min_{x} \{ f(x) + g(x) \} \]

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

<table>
<thead>
<tr>
<th></th>
<th>( f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} |y - x|_2^2 )</th>
<th>( x, y \in \text{dom}(f) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower surrogate</td>
<td>( f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} |y - x|_2^2 )</td>
<td>( x, y \in \text{dom}(f) )</td>
</tr>
<tr>
<td>Upper surrogate</td>
<td>( \mu \mathbb{I} \leq \nabla^2 f(x) \leq L \mathbb{I} )</td>
<td>( x \in \text{dom}(f) )</td>
</tr>
</tbody>
</table>
A basic composite minimization framework

\[ f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{\mu}{2} ||x - x^k||^2 \leq f(x) \]

\[ f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{L}{2} ||x - x^k||^2 \geq f(x) \]

\[ x^{k+1} = \text{prox}_g \left( x^k - \frac{1}{L} \nabla f(x^k) \right) \]

\[ \min_x \{ f(x) + g(x) \} \]

\[ x^{k+1} := \arg \min_x \left\{ f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{L}{2} ||x - x^k||^2 + g(x) \right\} \]

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

|            | \( f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} ||y - x||^2 \) | \( x, y \in \text{dom}(f) \) |
|------------|--------------------------------------------------------------------------------|-----------------------------|
| **Lower surrogate** | \( f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} ||y - x||^2 \) | \( x, y \in \text{dom}(f) \) |
| **Upper surrogate**  |                                                                                     |                             |
| **Hessian surrogates** | \( \mu I \leq \nabla^2 f(x) \leq L I \) | \( x \in \text{dom}(f) \) |
### A basic composite minimization framework

A basic composite minimization framework aims to minimize a function $f(x)$ subject to certain constraints. The framework is given by the following equation:

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

This equation represents a composite function $F_{\mu,L}$, which is a combination of lower and upper surrogates and Hessian surrogates. The properties of $F_{\mu,L}$ are summarized in the table below:

<table>
<thead>
<tr>
<th>Property</th>
<th>Equation</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower surrogate</td>
<td>$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} |y - x|_2^2$</td>
<td>$x,y \in \text{dom}(f)$</td>
</tr>
<tr>
<td>Upper surrogate</td>
<td>$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} |y - x|_2^2$</td>
<td>$x,y \in \text{dom}(f)$</td>
</tr>
<tr>
<td>Hessian surrogates</td>
<td>$\mu \mathbb{I} \preceq \nabla^2 f(x) \preceq L \mathbb{I}$</td>
<td>$x \in \text{dom}(f)$</td>
</tr>
</tbody>
</table>

### Main properties of $F_{\mu,L}$

- **ISTA**

$$x^{k+1} = \text{prox}_g \left( x^k - \frac{1}{L} \nabla f(x^k) \right)$$
A basic composite minimization framework

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

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

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

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

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

| Lower surrogate       | \( f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}||y - x||_2^2 \) | \( x, y \in \text{dom}(f) \) |
|------------------------|-------------------------------------------------------------------|---------------------------------|
| Upper surrogate        | \( f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}||y - x||_2^2 \) | \( x, y \in \text{dom}(f) \) |
| Hessian surrogates     | \( \mu\mathbb{I} \leq \nabla^2 f(x) \leq L\mathbb{I} \)          | \( x \in \text{dom}(f) \)   |
A basic composite minimization framework

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

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

- **Main properties of** $F_{\mu, L}$

<table>
<thead>
<tr>
<th>Surrogate</th>
<th>Expression</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower surrogate</td>
<td>$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}|y - x|_2^2$</td>
<td>$x, y \in \text{dom}(f)$</td>
</tr>
<tr>
<td>Upper surrogate</td>
<td>$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}|y - x|_2^2$</td>
<td>$x, y \in \text{dom}(f)$</td>
</tr>
<tr>
<td>Hessian surrogates</td>
<td>$\mu I \preceq \nabla^2 f(x) \preceq L I$</td>
<td>$x \in \text{dom}(f)$</td>
</tr>
</tbody>
</table>

\[ x^{k+1} = \text{prox}_g\left(x^k - \frac{1}{L}\nabla f(x^k)\right) \]

\[ f(x^k) - f(x^\star) \leq \frac{L\|x^0 - x^\star\|_2^2}{2k} \Rightarrow \text{iterations} = \mathcal{O}(\varepsilon^{-1}) \]

**acceleration is possible**

**ISTA**

**FISTA**
To adapt or not to adapt?

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

\[ \| \nabla f(x) - \nabla f(y) \| \leq L \| y - x \| \]

L is a global worst-case constant
To adapt or not to adapt?

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

\[ \|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\| \]

L is a global worst-case constant
To adapt or not to adapt?

$L$ is a global worst-case constant

\[ f(x) \leq f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{L}{2}\|x - x^k\|_2^2 + g(x) \]
To adapt or not to adapt?

Variable metric proximal point operator

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

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

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

$$f(x) \leq f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2} \|x - x^k\|^2_{H_k^{-1}}$$
To adapt or not to adapt?

\[ \text{prox}_{\gamma g}(s) := \arg \min_x \left\{ g(x) + \frac{1}{2\gamma} \|x - s\|^2 \right\} \]

Variable metric proximal point operator

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

if \( g(x) = \|x\|_1 \), then
\( \text{prox}_{\gamma g}(s) = \text{SoftThresh}(s, \gamma) \)
\( \text{prox}_{H^{-1}g}(s) = \text{LASSO} \)
A basic variable metric minimization framework

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

<table>
<thead>
<tr>
<th>Order</th>
<th>Example</th>
<th>Components</th>
<th>$k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-st</td>
<td>Accelerated gradient</td>
<td>$\nabla f$, $\text{prox}_{1/L}I_n$</td>
<td>$O(\epsilon^{-1/2})$</td>
</tr>
<tr>
<td>1+-th</td>
<td>BFGS</td>
<td>$H_k$, $\nabla f$, $\text{prox}_{H_k^{-1}}$</td>
<td>$O(\log \epsilon^{-1})$ or faster</td>
</tr>
<tr>
<td>2-nd</td>
<td>Proximal Newton, IPM</td>
<td>$\nabla^2 f$, $\nabla f$, $\text{prox}_{\nabla^2 f^{-1}}$</td>
<td>$O(\log \log \epsilon^{-1})$</td>
</tr>
</tbody>
</table>
Self-concordance vs. Lipschitz gradient + SC

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

<table>
<thead>
<tr>
<th>Surrogate Type</th>
<th>Expression</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower surrogate</td>
<td>$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} |y - x|^2$</td>
<td>$x, y \in \text{dom}(f)$</td>
</tr>
<tr>
<td>Upper surrogate</td>
<td>$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} |y - x|^2$</td>
<td>$x, y \in \text{dom}(f)$</td>
</tr>
<tr>
<td>Hessian surrogates</td>
<td>$\mu \mathbb{I} \preceq \nabla^2 f(x) \preceq L \mathbb{I}$</td>
<td>$x \in \text{dom}(f)$</td>
</tr>
</tbody>
</table>
**Self-concordance vs. Lipschitz gradient + SC**

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

<table>
<thead>
<tr>
<th></th>
<th>$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} |y - x|^2$</th>
<th>$x, y \in \text{dom}(f)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lower surrogate</strong></td>
<td>$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} |y - x|^2$</td>
<td>$x, y \in \text{dom}(f)$</td>
</tr>
<tr>
<td><strong>Upper surrogate</strong></td>
<td>$\mu \mathbb{I} \preceq \nabla^2 f(x) \preceq L \mathbb{I}$</td>
<td>$x \in \text{dom}(f)$</td>
</tr>
<tr>
<td><strong>Hessian surrogates</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Self-concordance vs. Lipschitz gradient + SC

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

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

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

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

\[
\frac{||y-x||_x^2}{1 + ||y-x||_x} \leq (\nabla f(y) - \nabla f(x))^T(y-x) \leq \frac{||y-x||_x^2}{1 - ||y-x||_x}, \quad \forall x, y \in \text{dom}(f)
\]

\[
f \text{ is self-concordant if } \varphi(t) := f(x + td) \text{ satisfies } |\varphi''(t)| \leq 2\varphi''(t)^{3/2}
\]
Self-concordance vs. Lipschitz gradient + SC

- **Main properties of $F_2$**

| Lower surrogate | $f(y) \geq f(x) + \nabla f(x)^T(y - x) + \omega(||y - x||_x)$ | $x, y \in \text{dom}(f)$ |
|-----------------|------------------------------------------------|--------------------------|
| Upper surrogate | $f(y) \leq f(x) + \nabla f(x)^T(y - x) + \omega_*(||y - x||_x)$ | $||y - x||_x < 1$ |
| Hessian surrogates | $(1 - ||y - x||_x)^2 \nabla^2 f(x) \leq \nabla^2 f(y) \leq (1 - ||y - x||_x)^{-2} \nabla^2 f(x)$ | $||y - x||_x < 1$ |

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

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

$$\frac{||y - x||_x^2}{1 + ||y - x||_x} \leq (\nabla f(y) - \nabla f(x))^T(y - x) \leq \frac{||y - x||_x^2}{1 - ||y - x||_x}, \ \forall x, y \in \text{dom}(f)$$

$f$ is self-concordant if $\varphi(t) := f(x + td)$ satisfies $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$
**Self-concordance: A mathematical tool**

- **Main properties of \( \mathcal{F}_2 \)**

<table>
<thead>
<tr>
<th></th>
<th>( f(y) \geq f(x) + \nabla f(x)^T(y - x) + \omega (|y - x|_x) )</th>
<th>( x, y \in \text{dom}(f) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower surrogate</td>
<td>( f(y) \leq f(x) + \nabla f(x)^T(y - x) + \omega_x (|y - x|_x) )</td>
<td>( |y - x|_x &lt; 1 )</td>
</tr>
<tr>
<td>Upper surrogate</td>
<td>( (1 - |y - x|_x)^2 \nabla^2 f(x) \leq \nabla^2 f(y) \leq (1 - |y - x|_x)^{-2} \nabla^2 f(x) )</td>
<td>( |y - x|_x &lt; 1 )</td>
</tr>
</tbody>
</table>

- **New variable metric framework with rigorous convergence guarantees**

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

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

\( f \) is self-concordant if \( \varphi(t) := f(x + td) \) satisfies \( |\varphi'''(t)| \leq 2\varphi''(t)^{3/2} \)
Our composite self-concordant minimization framework

- Proximal Newton scheme

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

$$x^{k+1} := x^k + \alpha_k d^k_{H_k}$$

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

$$H_k = \nabla^2 f(x^k)$$
Our composite self-concordant minimization framework

- **Proximal Newton scheme**

  \[ H_k = \nabla^2 f(x^k) \]

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

  \[ x^{k+1} := x^k + \alpha_k d^k_{H_k} \]

  where \( \alpha_k \in (0, 1] \) is step-size, \( d^k_{H_k} \) is a search direction

- **How to compute the Proximal Newton direction?**

  \[ d^k_{H_k} := \arg \min_d \left\{ f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d + g(x^k + d) \right\}, \quad H_k = \nabla^2 f(x^k) \]

  \[ f(x) \]

  \[ s^k_g := \arg \min_{x \in \text{dom}(F)} \{ Q(x; x^k, H_k) + g(x) \} \]

  \[ x^{k+1} := x^k + \alpha_k \left( s^k_g - x^k \right) \]
Our composite self-concordant minimization framework

- **Proximal Newton scheme**  
  \[ H_k = \nabla^2 f(x^k) \]
  
  Given \( x^0 \), generate a sequence \( \{x^k\}_{k \geq 0} \) such that
  \[ x^{k+1} := x^k + \alpha_k d^k_{H_k} \]
  where \( \alpha_k \in (0, 1] \) is step-size, \( d^k_{H_k} \) is a search direction

- **How to compute the Proximal Newton direction?**

  \[
  d^k_{H_k} := \arg \min_d \left\{ f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d + g(x^k + d) \right\}, \quad H_k = \nabla^2 f(x^k)
  \]

\[\]
Our composite self-concordant minimization framework

• Proximal Newton scheme

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

\[ x^{k+1} := x^k + \alpha_k d^k_{H_k} \]

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

• How to compute the Proximal Newton direction?

\[ d^k_{H_k} := \arg \min_d \left\{ f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d + g(x^k + d) \right\}, \quad H_k = \nabla^2 f(x^k) \]
How do we compute the step-size?

- Upper surrogate of $f$

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

- Convexity of $g$ and optimality condition of the subproblem

$$g(x^{k+1}) - g(x^k) \leq -\alpha_k \nabla f(x^k)^T d^k_{H_k} - \alpha_k \|d^k_{H_k}\|_{H_k}^2.$$
How do we compute the step-size?

- **Upper surrogate of** $f$
  \[ f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^T (x^{k+1} - x^k) + \omega^* (\|x^{k+1} - x^k\|_{x^k}) \], \quad \|x^{k+1} - x^k\|_{x^k} < 1

- **Convexity of** $g$ **and optimality condition of the subproblem**
  \[ g(x^{k+1}) - g(x^k) \leq -\alpha_k \nabla f(x^k)^T d^{k}_{H_k} - \alpha_k \|d^{k}_{H_k}\|_{H_k}^2. \]
  \[ \phi(x^{k+1}) \leq \phi(x^k) - \alpha_k \|d^{k}_{H_k}\|_{H_k}^2 + \omega^* (\alpha_k \|d^{k}_{H_k}\|_{x^k}) \]
How do we compute the step-size?

- **Upper surrogate of** \( f \)

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

- **Convexity of** \( g \) and **optimality condition** of the subproblem

\[
g(x^{k+1}) - g(x^k) \leq -\alpha_k \nabla f(x^k)^T d_{H_k}^k - \alpha_k \|d_{H_k}^k\|_{H_k}^2.
\]

\[
\phi(x^{k+1}) \leq \phi(x^k) - \alpha_k \|d_{H_k}^k\|_{H_k}^2 + \omega^* (\alpha_k \|d_{H_k}^k\|_{x^k})
\]

- **When** \( H_k \equiv \nabla^2 f(x^k) \), \( \lambda_k := \|d_{H_k}^k\|_{x^k} \)

\[
\phi(x^{k+1}) \leq \phi(x^k) - \left[ \alpha_k \lambda_k - \omega^* (\alpha_k \lambda_k) \right]
\]

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

\[
\alpha_k^* = \frac{1}{\lambda_k + 1} \in (0, 1]
\]
Analytic complexity

- Worst-case complexity to obtain an $\varepsilon$-approximate solution

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

- Worst-case complexity to obtain an $\varepsilon$-approximate solution

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

Can explicitly calculate $Q_\sigma := \{x^k \mid \lambda_k \leq 0.219\}$
Analytic complexity

- Worst-case complexity to obtain an $\varepsilon$-approximate solution

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

- Can explicitly calculate

- Quadratic convergence region

$$Q_\sigma := \{x^k \mid \lambda_k \leq 0.219\}$$

- Line-search can accelerate the convergence

- Global convergence

- Local convergence

- Line-search enhancement

- Enhanced backtracking

- Standard backtracking

- Forward linesearch

- Overjump
Graphical model selection

- **Objective:**

\[
\min_{\Theta > 0} \left\{ -\log \det(\Theta) + \text{trace}(\Sigma\Theta) + \rho \|\text{vec}(\Theta)\|_1 \right\}
\]

\(f(x)\)

\(g(x)\)
Graphical model selection

- Objective:

\[
\min_{\Theta > 0} \left\{ -\log \det(\Theta) + \text{trace}(\Sigma \Theta) + \rho \|\vec(\Theta)\|_1 \right\}
\]

- Gradient and Hessian (large-scale, special structure)

  - Gradient of \( f \): \( \nabla f(x) = \vec(\Sigma - \Theta^{-1}) \).

  - Hessian of \( f \): \( \nabla^2 f(x) = \Theta^{-1} \otimes \Theta^{-1} \)
**Graphical model selection**

- **Objective:**
  \[
  \min_{\Theta > 0} \begin{cases} 
  - \log \det(\Theta) + \text{trace}(\Sigma \Theta) + \rho \| \text{vec}(\Theta) \|_1 \\
  f(x) & g(x)
  \end{cases}
  \]

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

- **How to compute the Proximal Newton direction?**
  \[
  d^k_{H_k} := \arg \min_d \left\{ f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d + g(x^k + d) \right\}, \quad H_k = \nabla^2 f(x^k)
  \]
Graphical model selection

- **Objective:**

\[
\min_{\Theta > 0} \left\{ -\log \det(\Theta) + \text{trace}(\Sigma \Theta) + \rho \|\text{vec}(\Theta)\|_1 \right\}
\]

\[
f(x) = \begin{cases} 
- \log \det(\Theta) + \text{trace}(\Sigma \Theta) + \rho \|\text{vec}(\Theta)\|_1 
\end{cases}
\]

\[
g(x) = \begin{cases} 
\end{cases}
\]

- **Gradient and Hessian: (large-scale, special structure)**

  - Gradient of \( f \): \( \nabla f(x) = \text{vec}(\Sigma - \Theta^{-1}) \).

  - Hessian of \( f \): \( \nabla^2 f(x) = \Theta^{-1} \otimes \Theta^{-1} \)

- **Dual approach for solving subproblem (SP)**

<table>
<thead>
<tr>
<th>Primal subproblem</th>
<th>Dual subproblem (SPGL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \min_{\Delta} \left{ \frac{1}{2} \text{trace}(\Theta_i^{-1} \Delta)^2 + \text{trace}(R_i \Delta) + \rho |\text{vec}(\Delta)|_1 \right} )</td>
<td>( \min_{|\text{vec}(U)|_{\infty} \leq 1} \left{ \frac{1}{2} \text{trace}(\Theta_i U)^2 + \text{trace}(Q_i U) \right} )</td>
</tr>
<tr>
<td>( R_i := \Sigma - 2\Theta_i^{-1} )</td>
<td>( Q_i := \rho^{-1}[\Theta_i \Sigma \Theta_i - 2\Theta_i] )</td>
</tr>
</tbody>
</table>

Unconstrained LASSO problem
Graphical model selection

- **Objective:**

\[
\min_{\Theta > 0} \left\{ -\log \det(\Theta) + \text{trace}(\Sigma \Theta) + \rho \| \text{vec}(\Theta) \|_1 \right\} \\
 f(x) \quad g(x)
\]

- **Gradient and Hessian (large-scale, special structure)**

  - Gradient of \( f \): \( \nabla f(x) = \text{vec}(\Sigma - \Theta^{-1}) \).
  - Hessian of \( f \): \( \nabla^2 f(x) = \Theta^{-1} \otimes \Theta^{-1} \).

- **Dual approach for solving subproblem (SP)**

<table>
<thead>
<tr>
<th>Primal subproblem</th>
<th>Dual subproblem (SPGL)</th>
</tr>
</thead>
</table>
| \[
\min_{\Delta} \left\{ \frac{1}{2} \text{trace}(\Theta_i^{-1} \Delta^2) + \text{trace}(R_i \Delta) + \rho \| \text{vec}(\Delta) \|_1 \right\}
\] |
| \( R_i := \Sigma - 2\Theta_i^{-1} \)                                              | \[
\min_{\|\text{vec}(U)\|_\infty \leq 1} \left\{ \frac{1}{2} \text{trace}(\Theta_i U)^2 + \text{trace}(Q_i U) \right\}
\] |
| \( Q_i := \rho^{-1}[\Theta_i \Sigma \Theta_i - 2\Theta_i] \)                      |                                                                                      |

**Unconstrained LASSO problem**

No Cholesky decomposition and matrix inversion
**Graphical model selection**

- **Objective:**

\[
\min_{\Theta > 0} \left\{ -\log \det(\Theta) + \text{trace}(\Sigma \Theta) + \rho \|\text{vec}(\Theta)\|_1 \right\}
\]

- **Gradient and Hessian** *(large-scale, special structure)*

  - Gradient of \( f(x) \): \( \nabla f(x) = \text{vec}(\Sigma - \Theta^{-1}) \).
  
  - Hessian of \( f(x) \): \( \nabla^2 f(x) = \Theta^{-1} \otimes \Theta^{-1} \).

- **Dual approach for solving subproblem (SP)**

<table>
<thead>
<tr>
<th>Primal subproblem</th>
<th>Dual subproblem (SPGL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_i := \Sigma - 2\Theta_i^{-1} )</td>
<td>( \text{Unconstrained LASSO problem} )</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
Q_i & := \rho^{-1}[\Theta_i \Sigma \Theta_i - 2\Theta_i] \\
\text{min} & \quad \left\{ \frac{1}{2} \text{trace}((\Theta_i U)^2) + \text{trace}(Q_i U) \right\}
\end{align*}
\]

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

\[
\lambda_i := [p - 2\text{trace}(W_i) + \text{trace}(W_i^2)]^{1/2}, \quad W_i = \Theta_i(\Sigma - \rho U^*)
\]
Graphical model selection: numerical examples

Our method vs QUIC [Hsieh2011]

- 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)

![Graphs showing convergence behaviour](image)

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

<table>
<thead>
<tr>
<th>LS Scheme</th>
<th>Synthetic ((\rho = 0.01))</th>
<th>Arabidopsis ((\rho = 0.5))</th>
<th>Leukemia ((\rho = 0.1))</th>
<th>Hereditary ((\rho = 0.1))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#iter</td>
<td>#chol</td>
<td>#Mm</td>
<td>#iter</td>
</tr>
<tr>
<td>NoLS</td>
<td>25.4</td>
<td>-</td>
<td>3400</td>
<td>18</td>
</tr>
<tr>
<td>BtkLS</td>
<td>25.5</td>
<td>37.0</td>
<td>2436</td>
<td>11</td>
</tr>
<tr>
<td>E-BtkLS</td>
<td>25.5</td>
<td>36.2</td>
<td>2436</td>
<td>11</td>
</tr>
<tr>
<td>FwLS</td>
<td>18.1</td>
<td>26.2</td>
<td>1632</td>
<td>10</td>
</tr>
</tbody>
</table>

*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 [\(p = 1869\)]

*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?

\[(P) \quad \min_{x \in \mathbb{R}^n} \{ \phi(x) := f(x) + g(x) \} \]

- Splitting methods
  - Forward-backward: \( f \) applicable if \( f \) has Lipschitz gradient
  - Douglas-Rachford decomposition: \( f \) and \( g \) have "tractable" proximity operators

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

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

- \( \log \det \) : 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 $x^0$, generate a sequence $\{x^k\}_{k \geq 0}$ such that
  \[
  x^{k+1} := x^k + \alpha_k d^k_{H_k}
  \]
  where $\alpha_k \in (0, 1]$ is step-size, $d^k_{H_k}$ is a search direction

- **How to compute the search direction?**
  \[
  d^k_{H_k} := \arg \min_d \left\{ f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d + g(x^k + d) \right\}, \quad H_k = D_k : \text{diagonal}
  \]

- **No Lipschitz assumption**
  A new predictor corrector scheme (with local linear convergence*)

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

New theory: Local linear convergence of the PG

Graph learning: Lymph [p = 587]

\[
\min_{\Theta > 0} \begin{cases} 
-f(x) = -\log \det(\Theta) + \text{trace}(\Sigma \Theta) + \rho \| \text{vec}(\Theta) \|_1 \\
-g(x) = \| x^k - x^* \|_x \text{ in log-scale} 
\end{cases}
\]

\[
\rho = 0.15 \quad \text{based line with } \kappa = 0.9935 \\
\rho = 0.55 \quad \text{based line with } \kappa = 0.9462
\]
New theory: Local linear convergence of the PG

Graph learning: Lymph [p = 587]

\[
\min_{\Theta > 0} \begin{cases} 
  - \log \det(\Theta) + \text{trace}(\Sigma \Theta) + \rho \|\text{vec}(\Theta)\|_1 \\
  f(x) \\
  g(x)
\end{cases}
\]
New theory: Local linear convergence of the PG

Graph learning: Lymph $[p = 587]$

\[
\begin{align*}
\min_{\Theta > 0} & \quad \left\{ \begin{array}{l}
- \log \det(\Theta) + \text{trace}(\Sigma \Theta) + \rho \| \text{vec}(\Theta) \|_1 \\
\end{array} \right\} \\
& \quad f(x) \quad g(x)
\end{align*}
\]
New theory: Local linear convergence of the PG

Graph learning: Lymph \[ p = 587 \]

Theory is based on the notion of “restricted strong convexity”

\[
\min_{\Theta > 0} \begin{cases} 
- \log \det(\Theta) + \text{trace}(\Sigma \Theta) + \rho \| \text{vec}(\Theta) \|_1 \end{cases}
\]

\[ f(x) \]

\[ g(x) \]
**New theory: Local linear convergence of the PG**

Graph learning: Lymph \([p = 587]\)

The theory is based on the notion of “restricted strong convexity”

\[
\min_{\Theta > 0} \left\{ -\log \det(\Theta) + \text{trace}(\Sigma \Theta) \right\}
\]

where \(f(x)\) is the objective function. The descent directions are given by the gradient of \(f(x)\), and the 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”

\[
\min_{\Theta > 0} \left\{ - \log \det(\Theta) + \text{trace}(\Sigma \Theta) \right\}
\]

\(\|x^k - x^*\|_x\) in log-scale

\(\rho = 0.15\)

\(\rho = 0.55\)

\(\|x^k - x^*\|_x\) in log-scale

Based line with \(\kappa = 0.9935\)

Based line with \(\kappa = 0.9462\)

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”

\[
\min_{\Theta \succ 0} \left\{ -\log \det(\Theta) + \text{trace}(\Sigma \Theta) + \rho \|\text{vec}(\Theta)\|_1 \right\}
\]

\(f(x)\)

\(g(x)\)

\(x_2\)

\(x_1\)

convergence depends on the restricted condition number

\(\rho = 0.15\)

\(\rho = 0.55\)
New theory: Local linear convergence of the PG

Graph learning: Lymph \([p = 587]\)

Theory is based on the notion of “restricted strong convexity”

\[
\min_{\Theta > 0} \left\{ -\log \det(\Theta) + \text{trace}(\Sigma \Theta) + \rho \|\text{vec}(\Theta)\|_1 \right\}
\]

restricted descend directions

convergence depends on the restricted condition number
**New theory: Local linear convergence of the PG**

**Graph learning: Lymph** \([p = 587]\)

Heteroscedastic LASSO \([\rho \text{ decreases from left to right}]\)

\[ x^* \equiv (\phi^*, \gamma^*) = \arg\min_{\phi, \gamma} \left\{ -\log(\gamma) + \frac{1}{2n} \|\gamma y - X\phi\|_2^2 + \rho \|\phi\|_1 \right\} \]

\[ = f(x) \]

\[ = g(x) \]
Proximal gradient scheme: new engineering

- A **greedy** enhancement

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

\[
\hat{x}^k := (1 - \alpha_k)x^k + \alpha_k s_g^k
\]

\[
\text{prox}: \ s_g^k := \arg \min_{x \in \text{dom}(F)} \left\{ Q(x; x^k, D_k) + g(x) \right\}
\]
Proximal gradient scheme: new engineering

- A greedy enhancement

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

\[
\hat{x}^k := (1 - \alpha_k)x^k + \alpha_k s^k_g
\]

prox: \[ s^k_g := \arg \min_{x \in \text{dom}(F)} \{ Q(x; x^k, D_k) + g(x) \} \]
Proximal gradient scheme: new engineering

- A **greedy** enhancement

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

\[
\hat{x}^k := (1 - \alpha_k)x^k + \alpha_k s_g^k
\]

**prox:**

\[
s_g^k := \arg \min_{x \in \text{dom}(F)} \{ Q(x; x^k, D_k) + g(x) \}
\]

slows down convergence!
Proximal gradient scheme: new engineering

- A greedy enhancement

\[
\begin{align*}
\min_{x \in \mathbb{R}^n} \{ \phi(x) := f(x) + g(x) \}
\end{align*}
\]

**simple decision:**

based on the function values \( \phi(s_g^k), \phi(\hat{x}^k) \) and \( \phi(x^k) \)

**Case 1:**

\[
\hat{x}^k := (1 - \alpha_k) x^k + \alpha_k s_g^k
\]

**prox:**

\[
s_g^k := \arg \min_{x \in \text{dom}(F)} \left\{ Q(x; x^k, D_k) + g(x) \right\}
\]
Proximal gradient scheme: new engineering

- A **greedy** enhancement

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

**Simple decision:**

Based on the function values $\phi(s_g^k), \phi(\hat{x}^k)$ and $\phi(x^k)$

**Cost:**

*practically none if implemented carefully*

$$\hat{x}^k := (1 - \alpha_k)x^k + \alpha ks_g^k$$

**prox:**

$$s_g^k := \arg \min_{x \in \text{dom}(F)} \{ Q(x; x^k, D_k) + g(x) \}$$
Our method vs SPIRAL-TAP [Harmany2012]

Poisson imaging reconstruction via TV regularisation

\[ x^* \in \arg\min_x \left\{ \sum_{i=1}^m a_i^T x - \sum_{i=1}^m y_i \log(a_i^T x + b_i) + g(x) \right\} \]

\[ f(x) \]
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)

\[ \frac{F(x^k) - F(x^*)}{|F(x^*)|} \text{ in log-scale} \]

# of iterations

\[ \frac{F(x^k) - F(x^*)}{|F(x^*)|} \text{ in log-scale} \]

Time (sec.) in log-scale

ProxGrad2
ProxGrad2g
SPIRAL−TAP
Barrier extensions
Constrained convex problems

\[
g^* := \min_{x \in \Omega} g(x)
\]

- \( \Omega \) is endowed with a self-concordant barrier \( f(x) \);

Self-concordant barrier

- Given \( f \) is smooth and convex on its domain. We define:

\[
\varphi(t) := f(x + tv), \text{ where } x, x + tv \in \text{dom}(f), v \in \mathbb{R}^n, t \in \mathbb{R}
\]

- **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:

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

\[ g^* := \min_{x \in \Omega} g(x) \]

- \( \Omega \) is endowed with a self-concordant barrier \( f(x) \);

**Examples:**

\[
\begin{align*}
\Omega : \mathbf{X} \succeq 0 & \quad \Rightarrow \quad f_{\Omega}(X) = -\log \det(X) \\
\Omega : \mathbf{a}^T \mathbf{x} \geq 0 & \quad \Rightarrow \quad f_{\Omega}(x) = -\log(\mathbf{a}^T \mathbf{x}) \\
\Omega : \| \mathbf{A} \mathbf{x} - \mathbf{b} \| \leq \sigma & \quad \Rightarrow \quad f_{\Omega}(x) = -\log(\sigma^2 - \| \mathbf{A} \mathbf{x} - \mathbf{b} \|^2)
\end{align*}
\]

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

\[ g^* := \min_{x \in \Omega} g(x) \]

- \( \Omega \) is endowed with a self-concordant barrier \( f(x) \);

**Examples:**

\[
\begin{align*}
\Omega : X \succeq 0 & \quad \Rightarrow \quad f_\Omega(X) = -\log \det(X) \\
\Omega : a^T x \succeq 0 & \quad \Rightarrow \quad f_\Omega(x) = -\log(a^T x) \\
\Omega : \|Ax - b\| \leq \sigma & \quad \Rightarrow \quad f_\Omega(x) = -\log(\sigma^2 - \|Ax - b\|^2)
\end{align*}
\]

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

\[
\min_{x \in \text{int}(\Omega)} \left\{ F(x, \tau) := f(x) + \tau^{-1} g(x) \right\}
\]

Proximal Point + Interior Point

combine them together!!!

\( f \) is a \( \nu \)-self-concordant barrier if \( \varphi(t) := f(x + td) \) 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

\[
\min_{x \in \text{int}(\Omega)} \left\{ F(x, \tau) := f(x) + \tau^{-1} g(x) \right\}
\]

One iteration \( k \) requires two updates **simultaneously**:

- Update the penalty parameter:
  \[
  t_{k+1} := (1 - \sigma_k) t_k, \quad \sigma_k \in [\sigma, 1) \quad \text{(e.g., } \sigma = 0.0337/\sqrt{\nu})
  \]

- Update the iterative vector (**can be solved approximately**):
  \[
  x^{k+1} := \arg\min_x \left\{ t_{k+1} \nabla f(x^k)^T (x - x^k) + \frac{t_{k+1}}{2} (x - x^k)^T \nabla^2 f(x^k)(x - x^k) + g(x) \right\}
  \]

*Details: “An Inexact Proximal Path-Following Algorithm for Constrained Convex Minimization,” [optimization-online](https://optimization-online.org/) and [lions.epfl.ch/publications].
How does it work?

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

\[
\min_{x \in \text{int}(\Omega)} \left\{ F(x, \tau) := f(x) + \tau^{-1} g(x) \right\}
\]

One iteration \( k \) requires two updates simultaneously:

- Update the penalty parameter:

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

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

\[
x^{k+1} := \arg\min_x \left\{ t_{k+1} \nabla f(x^k)^T(x - x^k) + \frac{t_{k+1}}{2} (x - x^k)^T \nabla^2 f(x^k)(x - x^k) + g(x) \right\}
\]

*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** $x^0_{p_2}$ such that: $\|x^0_{p_2} - x^*(t_0)\|_{x^*(t_0)} \leq 0.05$

- **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:
  \[
  \text{if } \left\| x^k - x^*(t_k) \right\|_{x^*(t_k)} \leq 0.05 \text{ then } \left\| x^{k+1} - x^*(t_{k+1}) \right\|_{x^*(t_{k+1})} \leq 0.05
  \]

Worst-case complexity:

- **Phase I:** Finding a starting point for **Phase II** requires at most
  \[
  j_{\text{max}} := \left\lceil \frac{F(x^0; t_0) - F(x^*(t_0); t_0)}{0.0012} \right\rceil + 1
  \]

- **Phase II:** The **worst-case complexity** to reach a $\varepsilon$-solution is at most:
  \[
  O \left( \sqrt{\nu} \log \left( \frac{C t_0}{\varepsilon} \right) \right)
  \]

- **Note:** This worst-case complexity is as the same as in **standard path-following methods** [see Nesterov2004]
Proximal path-following for conic programming with rigorous guarantees

\[ g^* := \min_{x \in \Omega} g(x) \]

- \( \Omega \) is endowed with a self-concordant barrier \( f(x) \);

**Example:** Low-rank SDP matrix approximation ...

\[ \min_{x} \rho \| \text{vec}(X - M) \|_1 + (1 - \rho) \text{tr}(X) \]

s.t. \( X \succeq 0, \quad L_{ij} \leq X_{ij} \leq U_{ij}, \quad i, j = 1, \ldots, n. \)

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

<table>
<thead>
<tr>
<th>Solver ( \backslash n )</th>
<th>80</th>
<th>100</th>
<th>120</th>
<th>140</th>
<th>160</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size ( [n_v; n_c] )</td>
<td>([16,200; 9,720] )</td>
<td>([25,250; 15,150] )</td>
<td>([36,300; 21,780] )</td>
<td>([49,350; 29,610] )</td>
<td>([64,400; 38,640] )</td>
</tr>
<tr>
<td>Time (sec)</td>
<td>New Algorithm</td>
<td>15.738</td>
<td>24.046</td>
<td>24.817</td>
<td>25.326</td>
</tr>
<tr>
<td></td>
<td>SDPT3</td>
<td>156.340</td>
<td>508.418</td>
<td>881.398</td>
<td>1742.502</td>
</tr>
<tr>
<td></td>
<td>SeDuMi</td>
<td>231.530</td>
<td>970.390</td>
<td>3820.828</td>
<td>9258.429</td>
</tr>
<tr>
<td>Objective value</td>
<td>New Algorithm</td>
<td>306.9159</td>
<td>497.6706</td>
<td>635.4304</td>
<td>842.4626</td>
</tr>
<tr>
<td></td>
<td>SDPT3</td>
<td>306.9153</td>
<td>497.6754</td>
<td>635.4306</td>
<td>842.4644</td>
</tr>
<tr>
<td></td>
<td>SeDuMi</td>
<td>306.9176</td>
<td>497.6821</td>
<td>635.4384</td>
<td>842.4776</td>
</tr>
</tbody>
</table>
Proximal path-following for conic programming with rigorous guarantees

\[ g^* := \min_{x \in \Omega} g(x) \]

- \( \Omega \) is endowed with a self-concordant barrier \( f(x) \);

**Example:** Low-rank SDP matrix approximation...

\[
\begin{align*}
\min_x & \quad \rho \| \text{vec}(X - M) \|_1 + (1 - \rho) \text{tr}(X) \\
\text{s.t.} & \quad X \succeq 0, \quad L_{ij} \leq X_{ij} \leq U_{ij}, \quad i, j = 1, \ldots, n.
\end{align*}
\]

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

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

+ **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**

Proximal path following for conic programming with rigorous guarantees

\[ g^* := \min_{x \in \Omega} g(x) \]

- \( \Omega \) is endowed with a self-concordant barrier \( f(x) \);

Example: Max-norm clustering

\[
\begin{align*}
\min_{L, R, K} & \quad \| \text{vec}(K - A) \|_1 \\
\text{s.t.} & \quad \begin{bmatrix} L & K \\ K^T & R \end{bmatrix} \succeq 0, \quad L_{ii} \leq 1, \quad R_{ii} \leq 1, \quad i = 1, \ldots, p.
\end{align*}
\]

**Upshot: desired scaling!**

<table>
<thead>
<tr>
<th>Dimension ( p )</th>
<th>Total execution time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 10^2 )</td>
<td>( O(p^3) )</td>
</tr>
<tr>
<td>( 10^3 )</td>
<td>( O(p^{5.5}) )</td>
</tr>
</tbody>
</table>

**DCO:**

<table>
<thead>
<tr>
<th>( p )</th>
<th>( \text{variables} )</th>
<th>( \text{constraints} )</th>
<th>( \text{PF scheme} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>15.1</td>
<td>2.6</td>
<td>10</td>
</tr>
<tr>
<td>75</td>
<td>33.9</td>
<td>5.8</td>
<td>22.5</td>
</tr>
<tr>
<td>100</td>
<td>60.2</td>
<td>10.2</td>
<td>40</td>
</tr>
<tr>
<td>150</td>
<td>135.3</td>
<td>22.8</td>
<td>90</td>
</tr>
<tr>
<td>200</td>
<td>240.4</td>
<td>40.4</td>
<td>160</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( p )</th>
<th>( 50 )</th>
<th>( 75 )</th>
<th>( 100 )</th>
<th>( 150 )</th>
<th>( 200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (sec)</td>
<td>PF</td>
<td>SDPT3</td>
<td>splitting</td>
<td>PF</td>
<td>SDPT3</td>
</tr>
<tr>
<td>50</td>
<td>62.450</td>
<td>4.396</td>
<td>102.217</td>
<td>549.1567</td>
<td>549.1860</td>
</tr>
<tr>
<td>100</td>
<td>202.600</td>
<td>64.939</td>
<td>354.444</td>
<td>2232.5897</td>
<td>2233.0747</td>
</tr>
<tr>
<td>150</td>
<td>416.044</td>
<td>522.021</td>
<td>778.904</td>
<td>5396.0485</td>
<td>5396.7305</td>
</tr>
<tr>
<td>200</td>
<td>1573.881</td>
<td>2588.721</td>
<td>1420.844</td>
<td>9809.6066</td>
<td>9809.6934</td>
</tr>
</tbody>
</table>

\[ g(K^*) \]
Proximal path-following  Upshot: auto. regularization!

Proximal path following for conic programming with rigorous guarantees

\[ g^* := \min_{x \in \Omega} g(x) \]

• \( \Omega \) is endowed with a self-concordant barrier \( f(x) \);

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

\[
\min_{x \in \text{int}(\Omega)} \{ F(x; t) := g(x) + tf(x) \}
\]

Example: Max-norm clustering

Example: Graph selection

\[ O(p^{5.5}) \]

\[ O(p^3) \]

<table>
<thead>
<tr>
<th>Total execution time (sec)</th>
<th>10^6</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^1</td>
<td></td>
</tr>
<tr>
<td>10^2</td>
<td></td>
</tr>
<tr>
<td>10^3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension p</th>
<th>10^2</th>
<th>10^3</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDPT3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Path following scheme</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“Leukemia” problem \( |p = 1255| \)
Smoothing via self-concordant barrier

Composite convex minimization with nonsmooth-max structure \( f \):

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

- Nonsmooth function \( f \) is defined as:

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

- \( U \): endowed with a self-concordant barrier \( b_U \), \( G \): convex

Motivating examples

- Nonsmooth convex function induced by a norm:

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

- Fenchel’s dual function with compact domain:

\[
f(x) := \max_{u \in \text{dom} f^*} \{ \langle x, u \rangle - f^*(u) \}
\]

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

\[
\min_{x} \left\{ \max_{u \in U} \{ x^T A u - G(u) \} - b^T x \right\}
\]
**Smoothing via self-concordant barrier**

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

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

- Nonsmooth function $f$ is defined as:
  $$f(x) := \max_{u \in \mathcal{U}} \{ \langle A^T x, u \rangle - G(u) \}$$

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

**Self-concordant barrier smoother** of $f^*$

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

- $\sigma > 0$ is a smoothness parameter

- The solution $u^*_{\sigma}(x)$ of the maximization problem satisfies the following optimality condition:
  $$0 \in A^T x - \partial G(u^*_{\sigma}(x)) - \sigma \nabla b_{\mathcal{U}}(u^*_{\sigma}(x))$$

- 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 $f^\sigma$ is given as

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

*Details: Manuscript under review for ICASSP'14.*
Three key properties of the barrier smoother

1. **Approximation property**

\[ f_\sigma(x) \leq f(x) \leq f_\sigma(x) + \sigma C \nu. \]

- \( \sigma > 0 \) is the smoothness parameter, \( C \) is a given constant and \( \nu \) is the barrier parameter.

2. **Local Lipschitz-like property**

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

- Here: \( c_A \equiv c_A(x) = \left( \| A \nabla^2 b_U(u^*_{\sigma}(x))A^T \|_2 \right)^{1/2} \leq \bar{c}_A \) (constant)

3. **Existence of the second order derivative**

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

\[ \nabla^2 f_\sigma(x) = A \left( \nabla^2 G(u^*_{\sigma}(x)) \right) + \sigma \nabla^2 b_U(u^*_{\sigma}(x))^{-1} A^T \succ 0 \]

- If, in addition, \( G \) is **self-concordant** with the parameter \( M_G \), then \( f_\sigma \) 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

\[ \phi^*_\sigma = \min_{x \in \mathbb{R}^n} \{ \phi_\sigma(x) := f_\sigma(x) + g(x) \} \]

Here, \( f_\sigma \) is smooth and \( g \) is “tractably” proximal

Proximal - gradient scheme

\[
\begin{aligned}
    s^k &:= \text{prox}^{g_{\gamma_k}}(x^k - \gamma_k \nabla f_\sigma(x^k)), \\
    d^k &:= s^k - x^k, \\
    x^{k+1} &:= x^k + \alpha_k d^k.
\end{aligned}
\]

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

Convergence and convergence rate

- Descent property:

\[ \phi_\sigma(x^{k+1}) \leq \phi_\sigma(x^k) - \sigma \omega \left( \frac{\lambda_k}{c_A^k} \right), \quad \omega(\tau) = \tau - \log(1 + t). \]

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

\[ \phi_\sigma(\bar{x}^k) - \phi_\sigma(x^*_\sigma) \leq \frac{(1 + M)\bar{c}_A^2}{2\sigma k} \|x^0 - x^*_\sigma\|_2^2 \]

\[ O \left( \frac{\bar{c}_A^2}{\sigma k} \|x^0 - x^*_\sigma\|_2^2 \right) \]
A stylized example

A simple quadratically constrained quadratic program

\[ \phi^* = \min_{x} \left\{ \phi(x) := \max_{u^T Bu \leq 1} \left\{ (Ax - b)^T u - \frac{1}{2} x^T Q u \right\} + c^T x \right\} \]

- **Self-concordant barrier** for \( U := \{ u \in \mathbb{R}^m \mid u^T Bu \leq 1 \} \) is \( b_U(u) := -\log(1 - u^T Bu) \) 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 \( p(x) = \frac{1}{2} x^T B x \)

- **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 \( c^k_A \)
A stylized example

A simple quadratically constrained quadratic program

\[ \phi^* = \min_x \left\{ \phi(x) := \max_{u^T B u \leq 1} \left\{ (A x - b)^T u - \frac{1}{2} x^T Q u \right\} + c^T x \right\}, \]

- Self-concordant barrier for \( \mathcal{U} := \{ u \in \mathbb{R}^m \mid u^T B u \leq 1 \} \) is \( b_{\mathcal{U}}(u) := -\log(1 - u^T B u) \) 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 \( p(x) = \frac{1}{2} x^T B x \)
  
  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

\[
\begin{align*}
    s^k & := \text{argmin}_x \{ Q(x; x^k) + g(x) \}, \\
    d^k & := s^k - x^k, \\
    \alpha_k & := (1 + \lambda_k)^{-1}, \quad \lambda_k := \sqrt{\sigma^{-1} \| d^k \|_{x^k}}, \\
    x^{k+1} & := x^k + \alpha_k d^k.
\end{align*}
\]

- \( Q \) is the quadratic surrogate of \( f_\sigma \) around \( x^k \):

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

Convergence

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

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

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

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

Simultaneously update both \( \sigma \) and \( x^k \) at each iteration \( k \).
Conclusions
Conclusions

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

- \( \mathcal{F}_\mu \) - \( \mu \)-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

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

- \( F_\mu \) - \( \mu \)-strongly convex
- \( F_L \) - \( L \)-Lipschitz gradient
- \( F_2 \) - self-concordant

- **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

- **Highlights**
Conclusions

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

- **F**: smooth
- **F\_\mu**: \( \mu \)-strongly convex
- **F\_L**: \( L \)-Lipschitz gradient
- **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](http://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