Scalable sparse covariance estimation via self-concordance

Anastasios Kyrillidis   Rabeeh Karimi Mahabadi   Quoc Tran-Dinh   Volkan Cevher

Laboratory for Information and Inference Systems (LIONS), EPFL

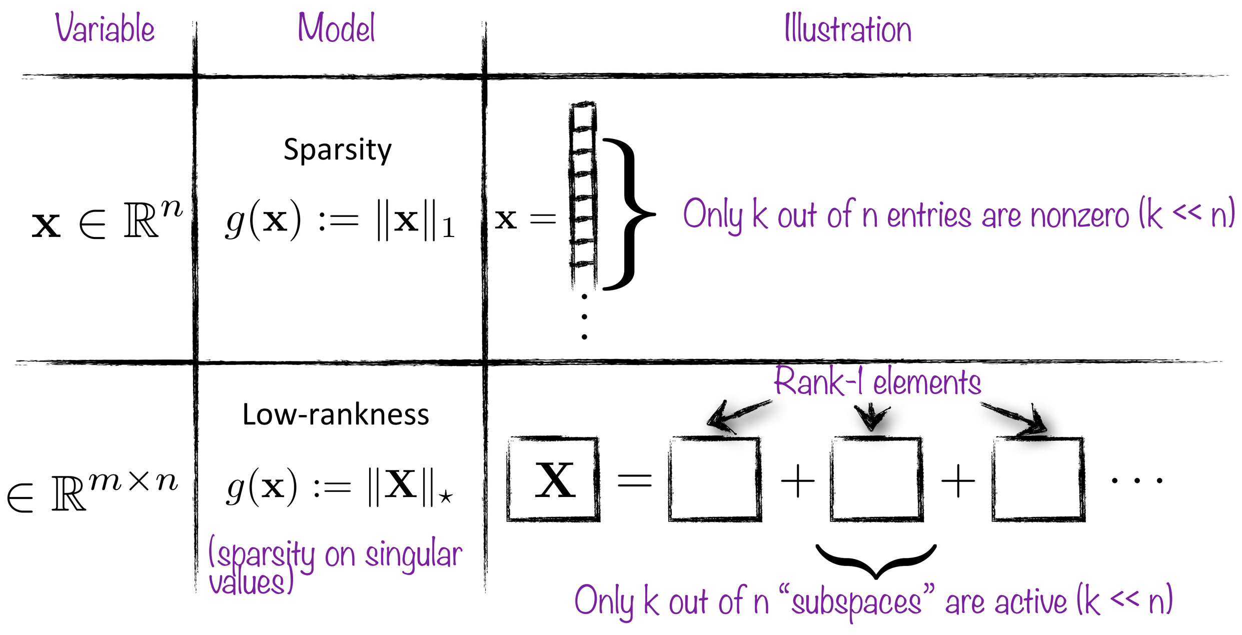
{anastasios.kyrillidis, quoc.trandinh, volkan.cevher}@epfl.ch,
rabeehk@student.ethz.ch

Big picture on convex optimization

- **Current trend in convex optimization:**

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{F(\mathbf{x}) \,:\, F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}, \quad \text{where } f \text{ is smooth convex and } g \text{ is non-smooth convex.}$$

- **"Hot" trend in optimization:** usage of low-dimensional models through g function:

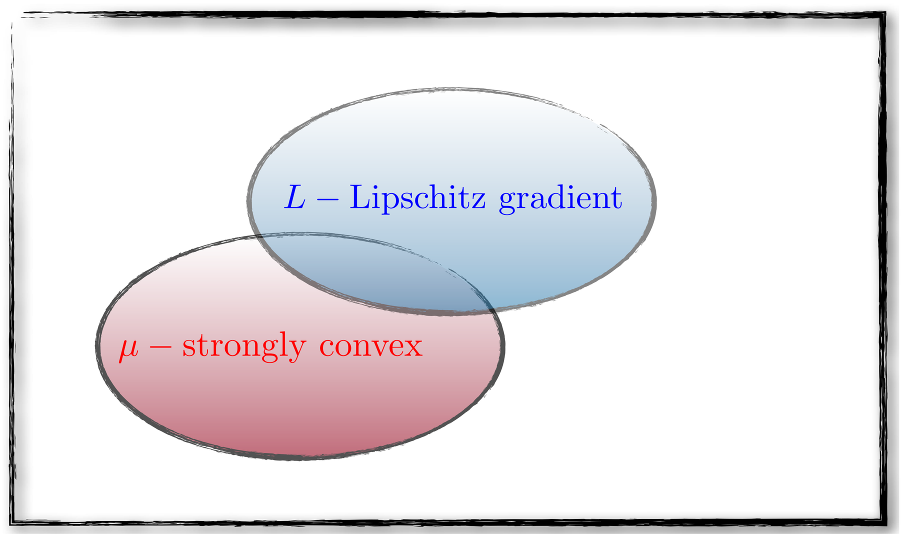


- Tractability of proximity operator for $g(\cdot)$:

$$\text{prox}_g^{\mathbf{H}}(\mathbf{y}) := \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ g(\mathbf{x}) + 1/2\|\mathbf{x}-\mathbf{y}\|_{\mathbf{H}}^2 \right\}$$

- Usually lead to harder-to-solve optimization problems...

- **Generic strategy:** $\mathbf{x}_{i+1} = \mathbf{x}_i + \tau_i \mathbf{d}_i$ where $\mathbf{d}_i$ is a direction to move and $\tau_i \in (0,1)$ is a step size.
- **How to choose $\tau_i, \mathbf{d}_i$:** By using assumptions on $f(\cdot)$:
  - $f(\cdot)$ "lives" into well-known classes of functions:



Lipschitz gradient continuity:

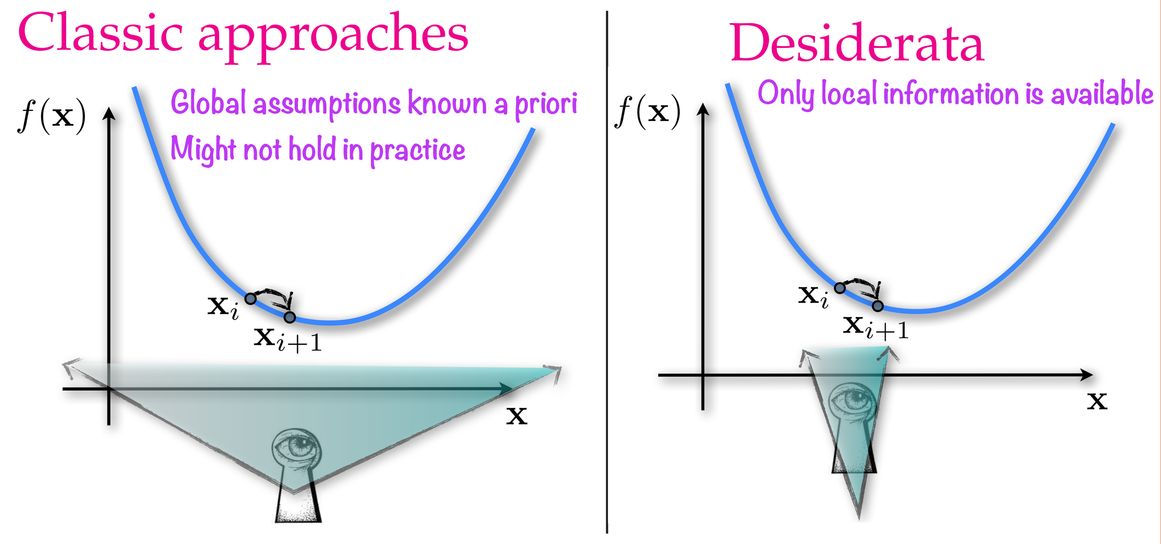
$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x}-\mathbf{y}\|_2$$

μ-strong convexity:

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

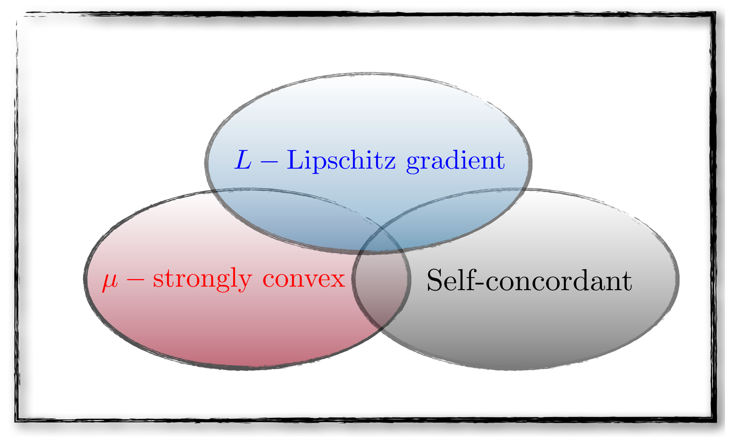
Self-concordance in optimization

- $\mathcal{F}_L$ and $\mathcal{F}_\mu$ are well-established assumptions but they might not hold in practice:



- Self-concordance: provides *affine invariance* in Newton methods – used in IP methods.

Definition 1 *A convex function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be self-concordant with parameter $M \geq 0$, if $|\varphi'''(t)| \leq M\varphi''(t)^{3/2}$, where $\varphi(t) := f(\mathbf{x}+t\mathbf{v})$ for all $t \in \mathbb{R}$, $\mathbf{x} \in dom(f)$ and $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{x}+t\mathbf{v} \in dom(f)$.*



The SCOPT framework [2]

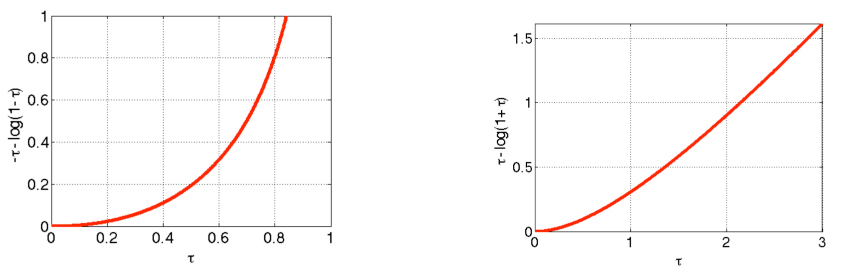
- Using self-concordant bounds:

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

Local

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

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



Algorithm 1 Inexact SCOPT for sparse cov. estimation

1: **Input:** $\mathbf{x}_0, \rho, \lambda > 0, \sigma = \frac{3}{40}, \epsilon, \gamma > 0$.
2: **while** $\varepsilon_i \leq \gamma$ or $i \leq I^{\max}$ **do**
3: Solve (4) for $\boldsymbol{\delta}_i$ with accuracy ϵ and parameters ρ, λ.
4: Compute $\varepsilon_i = \|\boldsymbol{\delta}_i - \mathbf{x}_i\|_{\mathbf{x}_i}$
5: **if** $(\varepsilon_i > \sigma)$
6: $\mathbf{x}_{i+1} = (1-\tau_i)\mathbf{x}_i + \tau_i\boldsymbol{\delta}_i$ for $\tau_i = \frac{\varepsilon_i - \sqrt{2\epsilon}}{\varepsilon_i(\varepsilon_i - \sqrt{2\epsilon}+1)}$.
7: **else** $\mathbf{x}_{i+1} = \boldsymbol{\delta}_i$
8: **end while**

Convergence guarantees

Theorem 1 (Global convergence guarantee) *Let $\tau_i := \frac{\varepsilon_i - \sqrt{2\epsilon}}{\varepsilon_i(\varepsilon_i - \sqrt{2\epsilon}+1)} \in (0,1)$ where $\varepsilon_i := \|\mathbf{d}_i - \mathbf{x}_i\|_{\mathbf{x}_i}$ is the* Newton decrement, *$\mathbf{d}_i$ is a direction to move and ϵ is the requested accuracy for finding $\mathbf{d}_i$. Assume $\varepsilon_i \geq \sqrt{2\epsilon}$, $\forall i$, and let the set $\{\mathbf{x} \in \text{dom}(F) : F(\mathbf{x}) \leq F(\mathbf{x}_0)\}$ be bounded. Then, SCOPT generates $\{\mathbf{x}_i\}_{i\geq 0}$ such that $\mathbf{x}_{i+1}$ satisfies:*

$$F(\mathbf{x}_{i+1}) \leq F(\mathbf{x}_i) - \xi(\tau_i), \quad \textit{where}$$

$\xi(\tau_i) = -\omega_*(\tau_i\varepsilon_i) - \tau_i\left(\epsilon - \frac{1}{2}\left(\varepsilon_i - \sqrt{2\epsilon}\right)^2 - \frac{1}{2}\varepsilon_i^2\right) \geq 0, \forall i$, *i.e.,* $\{F(\mathbf{x}_i)\}_{i\geq 0}$ *is a* strictly non-increasing *sequence.*

- We prove the convergence rate towards the minimizer using *local information* in norm measures: as long as $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|$ is away from 0, the algorithm has not yet converged to $\mathbf{x}^\star$. We observe:

$$\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_{\mathbf{x}_i} = \|\tau_i(\boldsymbol{\delta}_i - \mathbf{x}_i)\|_{\mathbf{x}_i} \propto \|\boldsymbol{\delta}_i - \mathbf{x}_i\|_{\mathbf{x}_i} := \varepsilon_i.$$

Theorem 2 (Local quadratic convergence rate) *Assume $\tau_i = 1$ or $\tau_i = \frac{\varepsilon_i - \sqrt{2\epsilon}}{\varepsilon_i(\varepsilon_i - \sqrt{2\epsilon}+1)} \in (0,1)$. Then, SCOPT satisfies:*

$$\varepsilon_{i+1} \leq \beta\varepsilon_i^2 + c,$$

where $\beta = \mathcal{O}\left(\frac{1}{1-\varepsilon_i}\right)$, $c = \sqrt{2\epsilon}$ and ϵ is user-defined. I.e., SCOPT has locally quadratic convergence rate *where $c > 0$ is small-valued and bounded.*

Sparse covariance estimation for portfolio optimization

- Classic Markowitz portfolio:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$$
$$\text{subject to} \quad \mathbf{w}^T\mathbf{r} = \mu, \quad \sum_i w_i = C, \quad w_i \geq 0, \ \forall i.$$

Usually, $\boldsymbol{\Sigma}$ is unknown...

- To approximate $\boldsymbol{\Sigma}$, we propose the self-concordant minimization:

$$\boldsymbol{\Theta}^\star = \arg\min_{\boldsymbol{\Theta}} \left\{ \frac{1}{2\rho}\|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Sigma}}\|_F^2 - \log\det(\boldsymbol{\Theta}) + \frac{\lambda}{\rho}\|\boldsymbol{\Theta}\|_1 \right\}$$

One cannot easily use L-Lipschitz and μ-strongly convex assumptions...

- Other applications: sparse graph selection, Poisson imaging, etc.

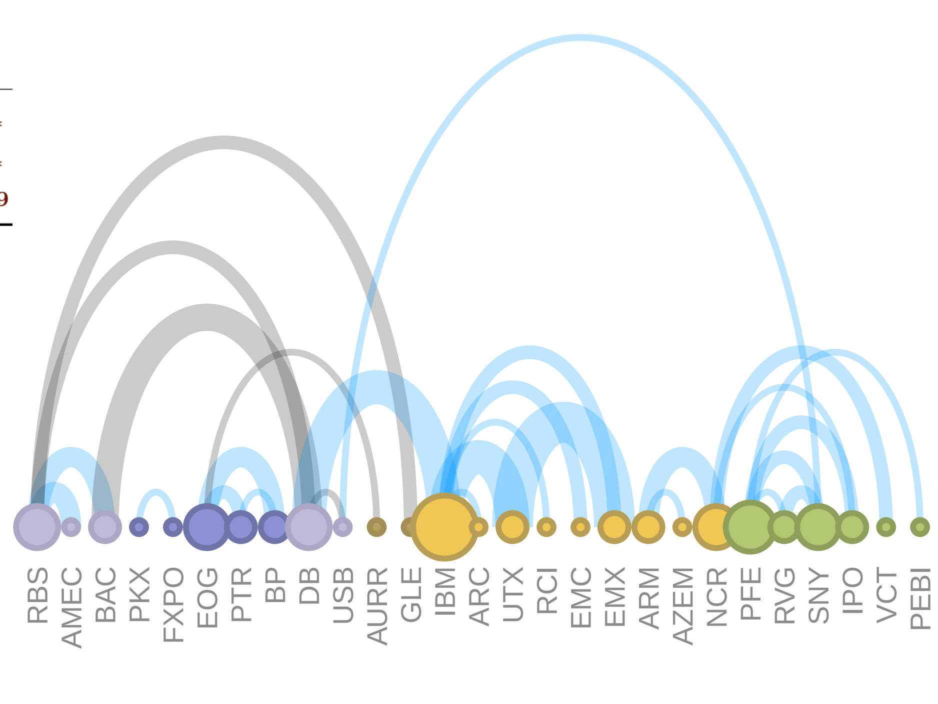
Table 2: Summary of comparison results for time efficiency.

Model				$F(\boldsymbol{\Theta}^\star)$ $(\times 10^2)$			Time (secs)		
	n		λ	[3]	iSCOPT	iSCOPT FLS	[3]	iSCOPT	iSCOPT FLS
$\boldsymbol{\Sigma}_3$	100	$\frac{k}{n^2} = 0.05$	1	32.013	**31.919**	**31.919**	8.288	9.996	**3.584**
		$\frac{k}{n^2} = 0.1$	0.5	36.190	**34.689**	**34.689**	10.470	12.761	**5.012**
		$\frac{k}{n^2} = 0.2$	0.5	62.143	**53.081**	**53.081**	18.446	14.720	**6.257**
	1000	$\frac{k}{n^2} = 0.05$	1	–	–	**2711.931**	$>$ T	$>$ T	**759.724**
		$\frac{k}{n^2} = 0.1$	1	–	–	**4734.251**	$>$ T	$>$ T	**875.344**
		$\frac{k}{n^2} = 0.2$	1	–	–	**5553.508**	$>$ T	$>$ T	**1059.709**

Table 3: Summary of comparison results for reconstruction of efficiency.

Model			$\|\boldsymbol{\Theta}^\star - \boldsymbol{\Sigma}\|_F / \|\boldsymbol{\Sigma}\|_F$			Time		
	n	N	[4]	[1]	iSCOPT FLS	[4]	[1]	iSCOPT FLS
$\boldsymbol{\Sigma}_3$	100	$n/2$	1.180	0.912	**0.908**	0.456	**0.252**	2.604
		n	0.920	0.554	**0.542**	0.494	**0.108**	0.155
		$10n$	0.396	0.192	**0.190**	0.451	0.108	**0.054**
	2000	$n/2$	–	**0.428**	**0.428**	$>$ T	350.145	**203.515**
		n	–	**0.352**	**0.352**	$>$ T	385.340	**167.688**
		$10n$	–	0.211	**0.209**	$>$ T	401.970	**122.535**

- Most correlations between assets tend to be zero in practice...



References

[1] Xue, L., Ma, S., and Zou, H.,"Positive definite ℓ_1 penalized estimation of large covariance matrices", Journal of the American Statistical Association, 2012

[2] Tran-Dinh, Q. Kyrillidis, A. and Cevher, V., "Composite self-concordant minimization", ArXiV.

[3] Rothman, A. J, "Positive definite estimators of large covariance matrices", Biometrika 99(3):733-740, 2012.

[4] Wang, H, "Two new algorithms for solving covariance graphical lasso based on coordinate descent and ECM", arXiv preprint arXiv:1205.4120, 2012.