Analysis of Second-Order Methods via non-convex Performance Estimation

EUROPT 2024

Nizar Bousselmi

joint work with Anne Rubbens, Julien Hendrickx, and François Glineur

icteam

UCLouvain

fnrs
LA LIBERTÉ DE CHERCHER

Goal : Analysis of optimization methods

- Optimization method $\mathcal{M}$ (e.g. gradient, Newton methods,…)
- Function class $\mathcal{F}$ (e.g. convex, smooth, self-concordant,…)
- Problem : $\min_x f(x)$

Question : Worst-case performance of $\mathcal{M}$ on instance of $\mathcal{F}$?

Example: Worst-case performance of Gradient Method on L-smooth convex functions after N iterations?

$$f(x_N) - f^* \leq \frac{L}{2} \frac{||x_0 - x^*||^2}{2N+1}$$

Constructing a proof of convergence rate

Definition of function class $\mathcal{F}$ → **Step 1: Inequalities** from definition of $\mathcal{F}$ → **Step 2: Optimal combination** of inequalities via **PEP** → Guarantee on the performance of $\mathcal{M}$

$\mathcal{F}$ = {smooth convex functions} → **Step 1: Inequalities**

$$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2, \ \forall x, y.$$

$$f(x) \geq f(y) + \langle \nabla f(y), x-y \rangle$$

→ **Step 2: combination** of inequalities and $x_{k+1} = x_k - \nabla f(x_k)$ → $f(x_N) - f^* \leq \frac{L}{2}\frac{||x_0 - x^*||^2}{2N+1}$

2 sources of (possible) conservatism on the guarantee:

1) **Inequalities** are not necessary and sufficient conditions to the class and allow « undesired functions »;
2) The **combination** is not optimal;

Optimal combination of **exact** inequalities leads to exact/tight worst-case analysis

Outline

1. **Performance Estimation Problem (PEP) Framework**

2. Non-convex PEP for second-order methods

3. New convergence results

Conceptual PEP: maximizing the worst-case performance

Idea: Finding the worst-case performance as an optimization problem

$$\max_{x_0, x^*, f} \quad \text{Perf}(x_N, f)$$

$$f \in \mathcal{F}$$

$$x_N = \mathcal{M}(x_0, f)$$

$$||\nabla f(x^*)||^2 = 0$$

$$||x_0 - x^*||^2 \leq 1$$

- Maximize Perf of $\mathcal{M}$ among the set of functions $f \in F$
- $\text{Perf}(x_N, f)$ can be : $||x_N - x^*||$, $||\nabla f(x_N)||$, $\mathrm{f_N} - \mathrm{f}^*$

Issue: Untractable since optimization in function space

Solution: Discretizing function f
(w.l.o.g. by black-box property of optimization methods)

From conceptual PEP to tractable PEP (1)

Example: Worst-case performance of gradient method on L-smooth convex functions

$$\max_{\text{points } x_i, x^*, \text{function } f} \quad f(x_N) - f(x^*)$$
$$\text{s.t.} \quad f \ L\text{-smooth convex},$$
$$x_{i+1} = x_i - \frac{1}{L}\nabla f(x_i),$$
$$||x^* - x_0||^2 \leq 1,$$
$$||\nabla f(x^*)||^2 = 0.$$

$\Longleftrightarrow$

$$\max_{\text{points } x_i, x^*, f_i, f^*, g_i, g^*} \quad f_N - f^*$$
$$\text{s.t.} \quad \exists f \ L\text{-smooth convex}: \quad f(x_i) = f_i, \quad \nabla f(x_i) = g_i,$$
$$f(x^*) = f^*, \quad \nabla f(x^*) = g^*,$$
$$x_{i+1} = x_i - \frac{1}{L} g_i,$$
$$||x^* - x_0||^2 \leq 1,$$
$$||g^*||^2 = 0.$$

Key concept: necessary and sufficient **interpolation conditions**

Interpolation conditions

Step 1: Inequalities from definition of $\mathcal{F}$

Theorem 1: f is L-smooth convex if and only if for all $x, y \in R^n$

$$f(y) \le f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2, \ \forall x, y.$$

$$f(x) \ge f(y) + \langle \nabla f(y), x - y \rangle$$

Theorem 2: f is L-smooth convex if and only if for all $x, y \in R^n$

$$f(y) \ge f(x) + \nabla f(x)^T (y - x) + \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2, \ \forall x, y.$$

Proof/PEP does **not use all $x, y \in R^n$, only $x_0, \ldots, x_N, x^*$**

Given $\left\{(x_1, g_1, f_1), \ldots, (x_N, g_N, f_N)\right\}$,

$\exists$ L-smooth convex f such that $\begin{cases} f(x_i) & = f_i, \quad \forall i \\ \nabla f(x_i) & = g_i, \quad \forall i \end{cases}$ if, and only if,

$$f_i \ge f_k + g_k^T (x_i - x_k) + \frac{1}{2L} ||g_i - g_k||^2 \quad \forall (i, k).$$

From conceptual PEP to tractable PEP (2)

Example: Worst-case performance of gradient method on L-smooth convex functions

$$
\begin{aligned}
\max_{\text{points } x_i, x^*, f_i, f^*, g_i, g^*} \quad & f_N - f^* \\
\text{s.t.} \quad & \exists f \text{ } L\text{-smooth convex}: \quad f(x_i) = f_i, \quad \nabla f(x_i) = g_i, \\
& \qquad\qquad f(x^*) = f^*, \quad \nabla f(x^*) = g^*, \\
& x_{i+1} = x_i - \frac{1}{L} g_i, \\
& ||x^* - x_0||^2 \leq 1, \\
& ||g^*||^2 = 0.
\end{aligned}
$$

$\Longleftrightarrow$

$$
\begin{aligned}
\max_{\text{points } x_i, x^*, f_i, f^*, g_i, g^*} \quad & f_N - f^* \\
\text{s.t.} \quad & f_i \geq f_k + g_k^T (x_i - x_k) + \frac{1}{2L} ||g_i - g_k||^2, \\
& x_{i+1} = x_i - \frac{1}{L} g_i, \\
& ||x^* - x_0||^2 \leq 1, \\
& ||g^*||^2 = 0.
\end{aligned}
$$

- Non-convex Quadratically Constrained Quadratic Problem (QCQP)
- Linear on f_i and $x_i^T g_i$, $x_i^T x_j$, $g_i^T g_j$
- It can be formulated as convex semidefinite program efficiently solvable !
- PEP gives the exact worst-case numerically (which helps to prove it analytically) [Drori, Teboulle 14] [Taylor, Hendrickx, Glineur 17]
- It gives all the answers, but we should ask the relevant questions

Convex formulation of PEP

Convex formulation of PEP when:

Only First-Order methods

- Method analyzed is linear combination of (previous or future) gradients g_i and iterates x_i.
- Interpolation conditions are convex in f_i and $x_i^T g_i, \;\; x_i^T x_j, \;\; g_i^T g_j$

1. Gradient method : $$x_{i+1} = x_i - \frac{h}{L}\nabla f(x_i)$$

2. Fast gradient method :
$$y_{i+1} = x_i - \frac{1}{L}\nabla f(x_i)$$
$$\theta_{i+1} = \frac{1+\sqrt{4\theta_i^2+1}}{2}$$
$$x_{i+1} = y_{i+1} + \frac{\theta_i - 1}{\theta_{i+1}}(y_{i+1} - y_i)$$

OK

See more examples in « PEPit’s documentation »

3. Proximal method: $x_{i+1} = prox_{f(.)}(x_i) = \; x_i - \nabla f(x_{i+1})$

4. Chambolle-Pock method:
$$\begin{cases} x_{i+1} & = \operatorname{prox}_{\tau f}\left(x_i - \tau M^T u_i\right), \\ u_{i+1} & = \operatorname{prox}_{\sigma g^*}\left(u_i + \sigma M(2x_{i+1} - x_i)\right), \end{cases}$$

[Drori, Teboulle 14]
[Taylor, Hendrickx, Glineur 17a]
[Taylor, Hendrickx, Glineur 17b]
[B, Hendrickx, Glineur 23]

No Convex formulation of PEP

Convex formulation of PEP when:

Only First-Order methods

- Method analyzed is linear combination of (previous or future) gradients g_i and iterates x_i.
- Interpolation conditions are convex in f_i and $x_i^T g_i,\ \ x_i^T x_j,\ \ g_i^T g_j$

1. Newton method: $x_{i+1} = x_i - [\nabla^2 f(x_i)]^{-1} \nabla f(x_i)$ ([de Klerk, Glineur, Taylor 2020] did it for one step)

KO

2. Finite differences : $x_{i+1} = x_i - \frac{f_j - f_i}{x_j - x_i}$

3. Adaptive methods:

$$\lambda_k = \min\left\{\sqrt{1+\theta_{k-1}}\lambda_{k-1}, \frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|}\right\}$$
$$x^{k+1} = x^k - \lambda_k \nabla f(x^k)$$
$$\theta_k = \frac{\lambda_k}{\lambda_{k-1}}$$

[Malitsky, Mishchenko 2020]

It seems impossible to formulate these PEP in a convex way

Non-Convex formulation of PEP

Idea: Tackle the non-convex formulation of PEP [Das Gupta, Van Parys, Ryu 2022]

- Analysis of (almost) any method is possible
- Heavy computational cost (global branch and bound solver)

$$\max_{\text{points } x_i, x^*, f_i, f^*, g_i, g^*} \quad f_N - f^*$$

$$\text{s.t.} \quad f_i \geq f_k + g_k^T(x_i - x_k) + \frac{1}{2L}||g_i - g_k||^2,$$

$$x_{i+1} = x_i - \frac{1}{L} g_i,$$

$$||x^* - x_0||^2 \leq 1,$$

$$||g^*||^2 = 0.$$

- Solve the non-convex (QCQP)
- We do not avoid « **Step 1** », we still need a good description of the class considered
- Integer variables and non-quadratic constraints also possible

Idea introduced in [Das Gupta, Van Parys, Ryu 2022] to design methods and used in [Das Gupta, Freund, Sun, Taylor, 2023] to analyze nonlinear conjugate gradients methods.

Outline

1. Performance Estimation Problem (PEP) Framework

2. Non-convex PEP for second-order methods

3. New convergence results

Analysis of Second-order methods via Non-Convex PEP

Example: Analysis of Newton method

$$\max_{x_k \in \mathbb{R}^d, g_k \in \mathbb{R}^d, h_k \in \mathbb{R}^{d\times d}, p_k \in \mathbb{R}^d} \quad ||x_N - x^*||^2$$

$$\text{s.t.} \quad \exists f \in \mathcal{F} \text{ s.t. } f(x_k) = f_k, \nabla f(x_k) = g_k, \nabla^2 f(x_k) = h_k,$$

$$\text{(Newton step)} \quad x_{k+1} = x_k - p_k,$$

$$h_k p_k = g_k,$$

$$||x_0 - x^*||^2 \leq R^2,$$

$$||g^*||^2 = 0,$$

Or any other second order scheme:

- Cubic Newton method : $T_M(x) \in \operatorname{Arg\,min}_y \left[\langle f'(x), y-x\rangle + \frac{1}{2}\langle f''(x)(y-x), y-x\rangle + \frac{M}{6}\|y-x\|^3 \right], \quad (2.4)$ [Nesterov, Polyak 2008]
- Damped Newton method: $x_{k+1} = x_k - \frac{1}{1+M_f \lambda_f(x_k)}[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$
- Gradient Regularized Newton method: $\lambda_k = \sqrt{H\|\nabla f(x^k)\|}$
  [Mishchenko 2022] $x^{k+1} = x^k - (\nabla^2 f(x^k) + \lambda_k \mathbf{I})^{-1} \nabla f(x^k)$

Interpolation conditions for univariate Hessian Lipschitz functions

Step 1: Inequalities from definition of $\mathcal{F}$

We focus on univariate functions for simplicity:

$\mathcal{D}_M$: univariate functions with Lipschitz continuous Hessian.

Definition. $f \in \mathcal{D}_M$ if, and only if

$$|f''(x) - f''(y)| \leq M|x - y| \quad \forall x, y. \tag{S}$$

Theorem. *If* $f \in \mathcal{D}_M$ *then,*

Not interpolation condition

$$|f(y) - f(x) - f'(x)(y - x) - \frac{1}{2}f''(x)(y - x)^2| \leq \frac{M}{6}|y - x|^3 \quad \forall x, y. \tag{S2}$$

Theorem. $f \in \mathcal{D}_M$ *if, and only if*

Interpolation condition

$$\begin{aligned} f(y) - f(x) - f'(x)(y - x) - \frac{1}{2}f(x)''(y - x)^2 \leq & \frac{M}{6}|y - x|^3 \\ & - \frac{(f'(y) - f'(x) - f''(x)(y - x) - \frac{M}{2}(y - x)|y - x|)^2}{2(M|y - x| - (f''(y) - f''(x)))} \\ & - \frac{(M|y - x| - (f''(y) - f''(x))^3}{96M^2} \quad \forall x, y. \end{aligned} \tag{S3}$$

Curiosity: « (S2) => (S)» is an open question as far as we know

Outline

1. Performance Estimation Problem (PEP) Framework

2. Non-convex PEP for second-order methods

3. New convergence results

Global convergence rate of Cubic Newton Method

$$x_{i+1} = \arg\min_x f(x) + f'(x_i)(x - x_i) + \frac{1}{2} f''(x_i)(x - x_i)^2 + \frac{M}{6}|x - x_i|^3. \qquad \text{(CNM)}$$

Theorem. *(CNM) on Hessian M-Lipschitz univariate functions satisfy*

$$f(x_i) - f(x_{i+1}) \geq \frac{M}{12}\left(\frac{|f'(x_{i+1})|}{M}\right)^{3/2}.$$

Moreover, if the function is bounded below by f^*, then

$$\min_{i=1,\ldots,N} |f'(x_i)| \leq 4M\left(\frac{3(f(x_0) - f^*)}{2MN}\right)^{2/3}.$$

[Nesterov, Polyak 2008] (in multivariate)

Step 1: Inequalities from definition of $\mathcal{F}$

Step 2: combination of inequalities and Iteration of $\mathcal{M}$

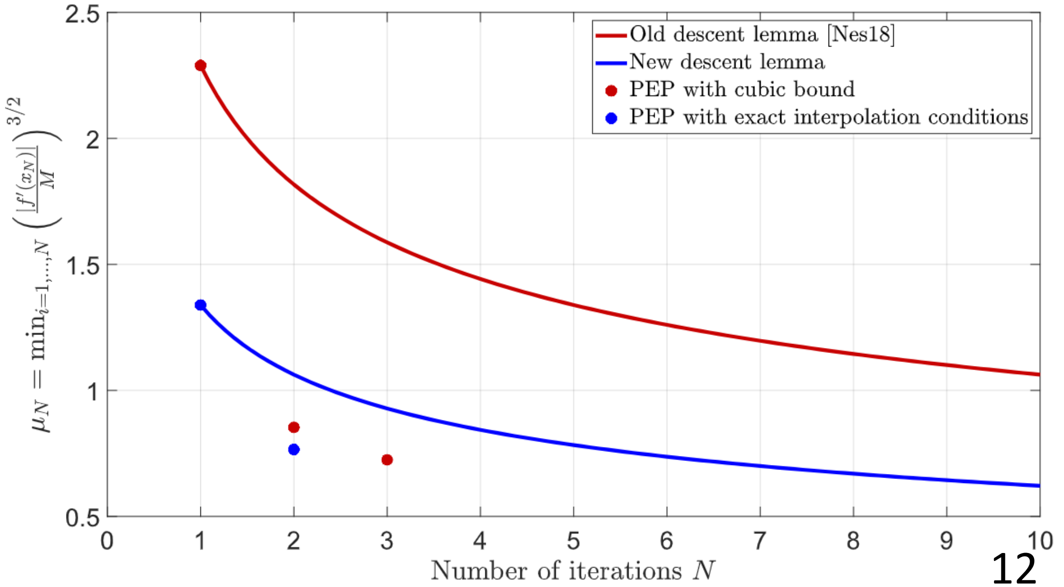
Theorem. *(CNM) on Hessian M-Lipschitz univariate functions satisfy*

$$f(x_i) - f(x_{i+1}) \geq \frac{5M}{12}\left(\frac{|f'(x_{i+1})|}{M}\right)^{3/2}.$$

Moreover, if the function is bounded below by f^*, then

$$\min_{i=1,\ldots,N} |f'(x_i)| \leq \frac{4M}{5^{2/3}}\left(\frac{3(f(x_0) - f^*)}{2MN}\right)^{2/3}.$$

[Rubbens, B, Hendrickx, Glineur 2024]



Local quadratic convergence rate of Newton Method

Theorem. *If*

- f *has a* M*-Lipschitz continuous Hessian,*
- $\exists x^*$ *such that* $\nabla f(x^*) = 0,\ \nabla^2 f(x^*) = \mu I \succ 0,$
- $\frac{M}{\mu}||x_0 - x^*|| \leq \frac{2}{3},$

then all Newton iterations $x_{k+1} = x_k - \nabla^2 f(x_k)^{-1}\nabla f(x_k)$ *satisfy*

$$||x_{k+1} - x^*|| \leq \frac{\frac{M}{\mu}||x_k - x^*||^2}{2\left(1 - \frac{M}{\mu}||x_k - x^*||\right)}$$

[Nesterov 2018]

Observation: PEP numerical results exactly match the bound

Theorem. *Theorem above is tight and attained by the following univariate cubic by parts function.*

$$f_1(x) = \begin{cases} \frac{Mx^3}{6} + \mu\frac{x^2}{2} & \text{if } x \leq 0, \\ -\frac{Mx^3}{6} + \mu\frac{x^2}{2} & \text{if } x > 0. \end{cases}$$

[Rubbens, B, Hendrickx, Glineur 2024]

Univariate case is « sufficiently rich » to attain the worst-case performance

Optimal step size of fixed damped Newton method

Fixed Damped Newton method : $x_{k+1} = x_k - \alpha \frac{f'(x_k)}{f''(x_k)}$

α that optimize the worst-case performance?

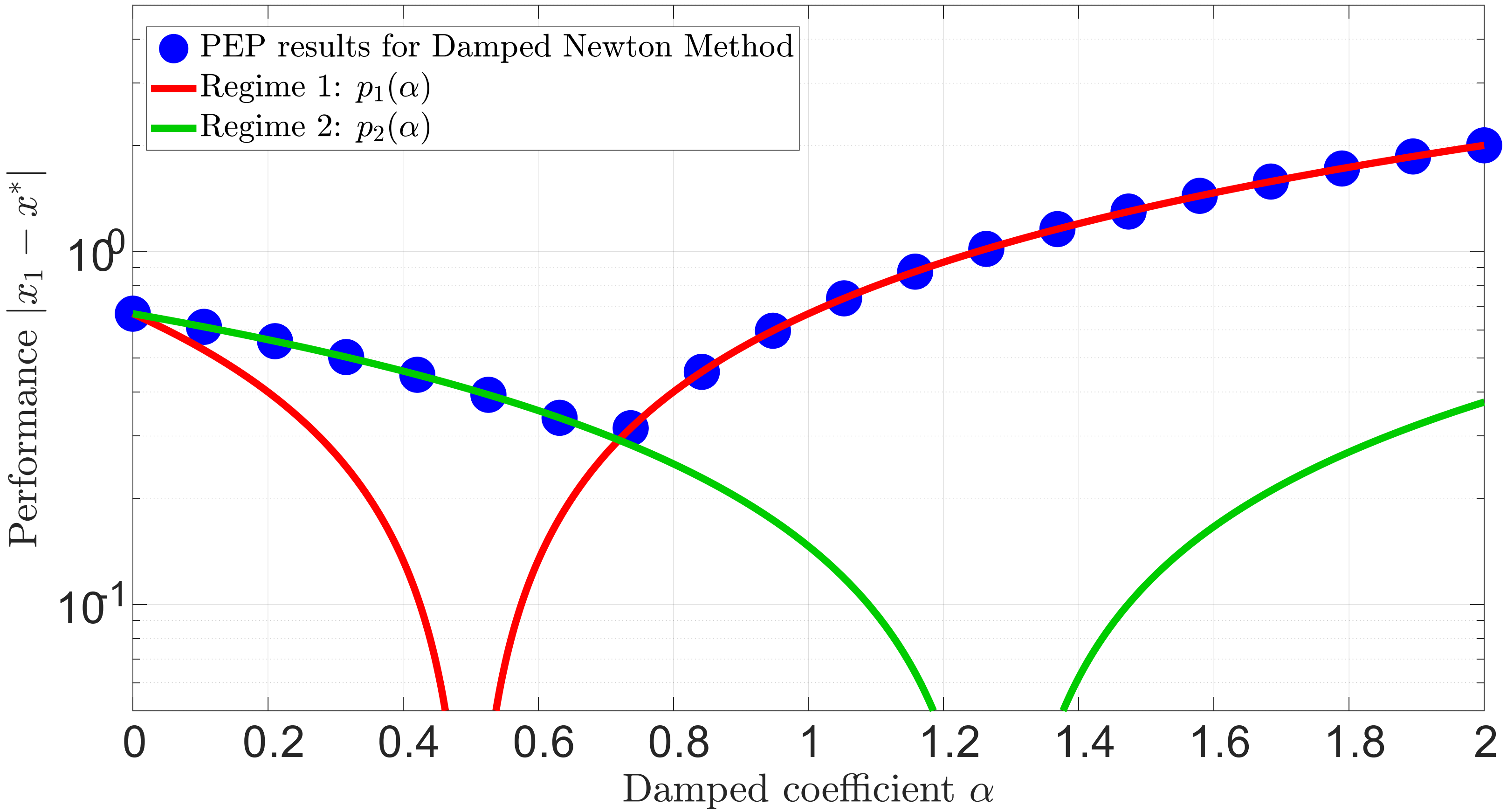


Fig. 1: $M = \mu = 1$ and $|x_0 - x^*| = \frac{2}{3}$

Summary (1/2)

State of the art

1. Tight worst-case performance requires **Step 1: Inequalities** and **Step 2: Combination of them**
2. PEP combines them **automatically and optimally**
3. Convex PEP is very efficient and useful to analyze **fixed first-order methods (see PEPit's documentation)**
4. Non-convex PEP allows to **analyze any method but is very costly**

Contributions

1. Interpolation conditions for **univariate Hessian Lipschitz functions**

Theorem. *$f \in \mathcal{D}_M$ if, and only if*

$$f(y) - f(x) - f'(x)(y-x) - \frac{1}{2} f(x)''(y-x)^2 \leq \frac{M}{6}|y-x|^3 - \frac{\left(f'(y) - f'(x) - f''(x)(y-x) - \frac{M}{2}(y-x)|y-x|\right)^2}{2(M|y-x| - (f''(y) - f''(x)))} - \frac{(M|y-x| - (f''(y) - f''(x))^3}{96M^2} \quad \forall x, y.$$

2. Applying non-convex PEP to **second-order methods**

Summary (2/2)

Contributions

3. **Improved Descent lemma of CNM** by a factor 5 (for univariate functions)

Theorem. *(CNM) on Hessian M-Lipschitz univariate functions satisfy*

$$f(x_i) - f(x_{i+1}) \geq \frac{5M}{12}\left(\frac{|f'(x_{i+1})|}{M}\right)^{3/2}.$$

4. **Exhibit a function attaining the worst** local quadratic convergece of Newton method

$$f_1(x) = \begin{cases} \frac{Mx^3}{6} + \mu\frac{x^2}{2} & \textit{if } x \leq 0, \\ -\frac{Mx^3}{6} + \mu\frac{x^2}{2} & \textit{if } x > 0. \end{cases}$$

5. **Step size selection** of damped Newton method (for univariate functions)

Future perspectives

Exploiting **non-convex PEP** to analyze new:

1. **Second-order schemes**: Gradient regularized Newton method, adaptive damped Newton method, etc
2. **Classes of functions**: self-concordant, etc
3. **Optimization methods**: zeroth order, adaptive, quasi-Newton methods, etc