Convex quadratic maps and invariance under rescaling

Alexandra Zverovich
(joint work with Matt Hutchings and Bertrand Gauthier)

14 January 2025

Cardiff University

Table of contents

Motivation and framework

Quadratic map

Let $\mathbf{Q} \in \mathbb{R}^{N \times N}$ be a symmetric positive semi-definite (SPSD) matrix, with $N \in \mathbb{N}$ and consider $\boldsymbol{c} \in \mathrm{span}\{\mathbf{Q}\}$, $(\boldsymbol{c} \neq \mathbf{0})$.

Define the convex quadratic map

$$D(\boldsymbol{x}) = \boldsymbol{x}^T \mathbf{Q} \boldsymbol{x} - 2\boldsymbol{c}^T \boldsymbol{x} + \boldsymbol{c}^T \boldsymbol{\alpha}, \quad \boldsymbol{x} \in \mathbb{R}^N.$$

with $\boldsymbol{\alpha} = \mathbf{Q}^\dagger \boldsymbol{c}$ (so that $\min_{\boldsymbol{x} \in \mathbb{R}^N} D(\boldsymbol{x}) = D(\boldsymbol{\alpha}) = 0$).

Remark: in practice, $\boldsymbol{\alpha}$ is unknown.

Existing techniques

For large N, direct approaches are intractable.

Instead, use iterative solvers:

- Conjugate-gradient method
  - Converges in N iterations.
  - Worst-case time-complexity per iteration is $\mathcal{O}(N^2)$, so intractable for very large problems.
- Coordinate-descent methods (e.g. Gauss-Seidel)
  - Sparse solvers.
  - Slow convergence.
  - Worst-case time-complexity per iteration is $\mathcal{O}(N)$, so suitable for very large problems.

Invariance under rescaling

We define the *relaxed map*

$$R(\boldsymbol{x}) = \min_{s \geqslant 0} D(s\boldsymbol{x}) = \begin{cases} \boldsymbol{c}^T \boldsymbol{\alpha} - (\boldsymbol{c}^T \boldsymbol{x})^2 / (\boldsymbol{x}^T \mathbf{Q} \boldsymbol{x}) \text{ if } \boldsymbol{x} \in \mathscr{A}, \\ \boldsymbol{c}^T \boldsymbol{\alpha} \text{ otherwise,} \end{cases}$$

with $\mathscr{A} = \{\boldsymbol{x} \in \mathbb{R}^N \mid \boldsymbol{c}^T \boldsymbol{x} > 0\}$.

We have $R(\boldsymbol{x}) = D(s_{\boldsymbol{x}} \boldsymbol{x})$, with

$$s_{\boldsymbol{x}} = \begin{cases} (\boldsymbol{c}^T \boldsymbol{x}) / (\boldsymbol{x}^T \mathbf{Q} \boldsymbol{x}) \text{ if } \boldsymbol{x} \in \mathscr{A}, \\ 0 \text{ otherwise.} \end{cases}$$

The relaxed map R is *invariant under rescaling*, that is, $R(s\boldsymbol{x}) = R(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^N$ and $s > 0$.

Properties of the relaxed map

Directional derivative and gradient

Setting $\mathscr{Z} = \{\boldsymbol{x} \in \mathbb{R}^N \,|\, \mathbf{Q}\boldsymbol{x} = 0\}$, the directional derivative $\Lambda(\boldsymbol{x};\boldsymbol{v})$ of R at $\boldsymbol{x} \in \mathbb{R}^N$ along $\boldsymbol{v} \in \mathbb{R}^N$ is

$$\Lambda(\boldsymbol{x};\boldsymbol{v}) = \lim_{t\to 0^+} \frac{1}{t}\left[R(\boldsymbol{x}+t\boldsymbol{v}) - R(\boldsymbol{x})\right] = \begin{cases} -\infty \text{ if } \boldsymbol{x} \in \mathscr{Z} \text{ and } \boldsymbol{v} \in \mathscr{A}, \\ 2s_{\boldsymbol{x}}\boldsymbol{v}^T(s_{\boldsymbol{x}}\mathbf{Q}\boldsymbol{x} - \boldsymbol{c}) \text{ otherwise.} \end{cases}$$

The gradient of R at $\boldsymbol{x} \notin \mathscr{Z}$ is $\nabla R(\boldsymbol{x}) = 2s_{\boldsymbol{x}}(s_{\boldsymbol{x}}\mathbf{Q}\boldsymbol{x} - \boldsymbol{c})$.

Properties of R

Theorem 1 (Pseudoconvex relaxation)

The map R is ***quasiconvex*** on $\mathbb{R}^N$, and ***pseudoconvex*** on the real convex cone $\mathscr{A}$.

Quasiconvexity: For $\boldsymbol{\xi} = \boldsymbol{x} + \rho(\boldsymbol{x} - \boldsymbol{u})$, $\boldsymbol{x}, \boldsymbol{u} \in \mathbb{R}^N$, $\rho \in [0, 1]$, we have $R(\boldsymbol{\xi}) \leqslant \max\{R(\boldsymbol{x}), R(\boldsymbol{u})\}$.

Pseudoconvexity: For $\boldsymbol{x}, \boldsymbol{u} \in \mathscr{A}$, if $\Lambda(\boldsymbol{x}; \boldsymbol{u} - \boldsymbol{x}) \geqslant 0$, then $R(\boldsymbol{x}) \leqslant R(\boldsymbol{u})$.

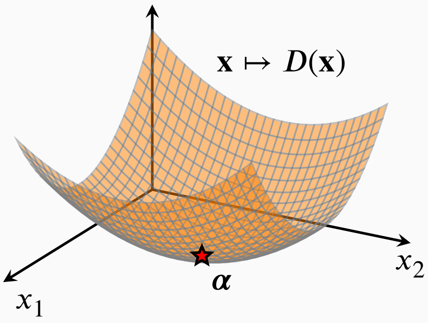

Figure 1: Graphical representation of the maps D and R over $\mathbb{R}^N_{\geqslant 0}$, $N = 2$ (illustration).

Exact line search

We can characterise the descent directions along which $\boldsymbol{R}$ can be minimised via exact line search.

Due to the invariance under scaling of $\boldsymbol{R}$, the iterate of an exact line search from $\boldsymbol{x}$ along $\boldsymbol{v}$ minimises $\boldsymbol{R}$ over $\text{span}\{\boldsymbol{x}, \boldsymbol{v}\}$.

To simplify the expression for the optimal step size, we set

$$\Upsilon(\boldsymbol{x}; \boldsymbol{v}) = (\boldsymbol{c}^T \boldsymbol{v})(\boldsymbol{x}^T \mathbf{Q} \boldsymbol{x}) - (\boldsymbol{c}^T \boldsymbol{x})(\boldsymbol{v}^T \mathbf{Q} \boldsymbol{x}).$$

Theorem 2 (Optimal step size)

Consider $\boldsymbol{x} \in \mathscr{A}$ and $\boldsymbol{v} \in \mathbb{R}^N$ and set $\boldsymbol{z}_t = \boldsymbol{x} + t\boldsymbol{v}$, $t \in \mathbb{R}$. If $\mathbf{Q}\boldsymbol{x}$ and $\mathbf{Q}\boldsymbol{v}$ are non-collinear, the following assertions hold.

(i) If $\Upsilon(\boldsymbol{v};\boldsymbol{x}) > 0$, then the function $t \mapsto R(\boldsymbol{z}_t)$, $t \in \mathbb{R}$, is minimum at $\tau = \Upsilon(\boldsymbol{x};\boldsymbol{v})/\Upsilon(\boldsymbol{v};\boldsymbol{x})$; we in this case have $\boldsymbol{z}_\tau \in \mathscr{A}$ and $R(\boldsymbol{z}_\tau) = \min_{\boldsymbol{z} \in \operatorname{span}\{\boldsymbol{x},\boldsymbol{v}\}} R(\boldsymbol{z})$.

(ii) If $\Upsilon(\boldsymbol{v};\boldsymbol{x}) \leqslant 0$, then the function $t \mapsto R(\boldsymbol{z}_t)$, $t \in \mathbb{R}$, is monotonic, and $\inf_{t\in\mathbb{R}} R(\boldsymbol{z}_t) = \min\{R(-\boldsymbol{v}), R(\boldsymbol{v})\}$.

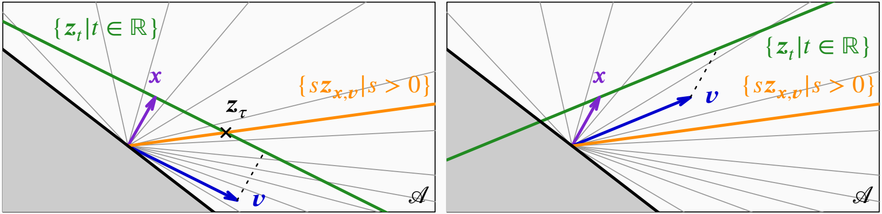

Figure 2: Schematic representation of the situations discussed in Theorem 2. The left plot corresponds to the case $\Upsilon(v; x) > 0$, and the right plot to $\Upsilon(v; x) \leqslant 0$. In each plot, the grey region indicates the set $\{x \in \mathbb{R}^N | c^T x \leqslant 0\}$, and the grey lines are level sets of the map R on $\text{span}\{x, v\}$. The direction $z_{x,v} \in \mathscr{A}$ is defined as $z_{x,v} = \Upsilon(v; x)x + \Upsilon(x; v)v$; it verifies

$$\arg \min_{z \in \text{span}\{x,v\}} R(z) = \{s z_{x,v} | s > 0\}.$$

Improvement score

Introduce $\mathcal{I}_R(\boldsymbol{x};\boldsymbol{v}) = R(\boldsymbol{x}) - \min_{\boldsymbol{z}\in\operatorname{span}\{\boldsymbol{x},\boldsymbol{v}\}} R(\boldsymbol{z}) \geqslant 0.$

Lemma 1 (Improvement score for R)

Consider $\boldsymbol{x} \in \mathscr{A}$ and $\boldsymbol{v} \in \mathbb{R}^N$, we have

$$\mathcal{I}_R(\boldsymbol{x};\boldsymbol{v}) = \big(\boldsymbol{v}^T(s_{\boldsymbol{x}}\mathbf{Q}\boldsymbol{x} - \boldsymbol{c})\big)^2 \big/ \big((\boldsymbol{v}^T\mathbf{Q}\boldsymbol{v}) - (\boldsymbol{v}^T\mathbf{Q}\boldsymbol{x})^2/(\boldsymbol{x}^T\mathbf{Q}\boldsymbol{x})\big);$$

if $\mathbf{Q}\boldsymbol{x}$ and $\mathbf{Q}\boldsymbol{v}$ are non-collinear, and $\mathcal{I}(\boldsymbol{x};\boldsymbol{v}) = 0$ otherwise.

Improvement scores

Setting $\mathcal{I}_D(\boldsymbol{x};\boldsymbol{v}) = D(\boldsymbol{x}) - \min_{t\in\mathbb{R}} D(\boldsymbol{x}+t\boldsymbol{v})$, $\boldsymbol{x}$ and $\boldsymbol{v} \in \mathbb{R}^N$, we have

$$\mathcal{I}_D(\boldsymbol{x};\boldsymbol{v}) = \begin{cases} \left(\boldsymbol{v}^T(\mathbf{Q}\boldsymbol{x}-\boldsymbol{c})\right)^2/(\boldsymbol{v}^T\mathbf{Q}\boldsymbol{v}) \text{ if } \boldsymbol{v} \notin \mathscr{Z}, \\ 0 \text{ otherwise.} \end{cases}$$

Link between improvement scores, and corrective term

For $\boldsymbol{x} \in \mathscr{A}$ and $\boldsymbol{v} \in \mathbb{R}^N$ such that $\mathbf{Q}\boldsymbol{x}$ and $\mathbf{Q}\boldsymbol{v}$ are non-collinear, we have $\mathcal{I}_R(\boldsymbol{x};\boldsymbol{v}) = \mathcal{I}_D(s_{\boldsymbol{x}}\boldsymbol{x};\boldsymbol{v})\mathcal{C}(\boldsymbol{x};\boldsymbol{v})$, with

$$\mathcal{C}(\boldsymbol{x};\boldsymbol{v}) = \left(1 - \frac{(\boldsymbol{v}^T\mathbf{Q}\boldsymbol{x})^2}{(\boldsymbol{v}^T\mathbf{Q}\boldsymbol{v})(\boldsymbol{x}^T\mathbf{Q}\boldsymbol{x})}\right)^{-1}.$$

Minimisation of the relaxed map

Coordinate descent with gradient-based rules

We minimise the relaxed map R using exact coordinate descent (i.e. iterations consist of exact line searches along directions in $\{\boldsymbol{e}_i\}_{i\in[N]}$):

- Select an initial iterate $\boldsymbol{x}^{(0)} \in \mathcal{A}$.
- Set $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \tau^{(k)}\boldsymbol{e}_{i^{(k)}}$, $k \in \mathbb{N}_0$, with $i^{(k)}$ selected using some selection rule and $\tau^{(k)}$ given by Theorem 2.

For the coordinate selection, we consider gradient-based rules.
Other rules, such as cyclic or randomised, could be considered.

Coordinate selection

A natural rule for the selection of a coordinate is

$$i_{R,\mathrm{BI},\boldsymbol{x}} \in \arg\max_{i\in[N]} \mathcal{I}_R(\boldsymbol{x};\boldsymbol{e}_i), \quad (\boldsymbol{x} \in \mathscr{A}).$$

This corresponds to the coordinate leading to the *best improvement* (BI) of R.

Another selection rule is the $\mathcal{H}$*-coordinate*,

$$i_{R,\mathcal{H},\boldsymbol{x}} \in \arg\max_{i\in[N]} \mathcal{I}_D(s_{\boldsymbol{x}}\boldsymbol{x};\boldsymbol{e}_i).$$

This corresponds to the coordinate potential $\mathbf{Q}\boldsymbol{e}_i$, $i \in [N]$, that aligns the most with $\nabla R(\boldsymbol{x})$ in the reproducing kernel Hilbert space $\mathcal{H} = \mathrm{span}\{\mathbf{Q}\}$.

Convergence properties

Define

$$\iota_{\mathbf{Q}} = \frac{\lambda_{\min}(\mathbf{Q})}{N \max_{i\in[N]} \mathbf{Q}_{i,i}} \in (0, 1].$$

Theorem 3 (Convergence)

Consider the minimisation of R over $\mathbb{R}^N$; the sequence of iterates $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}_0}$ generated by an exact coordinate descent with $\mathcal{H}$ rule verifies $\lim_{k\to\infty} R(\boldsymbol{x}^{(k)}) = 0$, with

$$R(\boldsymbol{x}^{(k)}) \leqslant (1 - \iota_{\mathbf{Q}})^k R(\boldsymbol{x}^{(0)}), \quad k \in \mathbb{N}_0.$$

The assertions of Theorem 3 also hold for the BI rule.

Experiments

Experiment 1

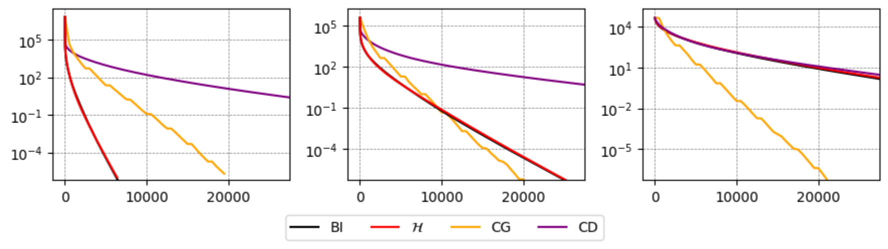

Figure 3: Decay of the map D for different ranges of the corrective terms C; 18 - 31, 5 - 7, 1.03 - 1.07 (left to right). We compare BI and $\mathcal{H}$ for the relaxed map R with popular methods (conjugate gradient and coordinate descent for D) by the number of matrix-column calls. The quadratic maps are generated using random $\mathbf{Q}$, $\mathbf{c}$ and $\boldsymbol{\alpha}$ for $N = 500$ and rank($\mathbf{Q}$) $= 250$.

Experiment 2

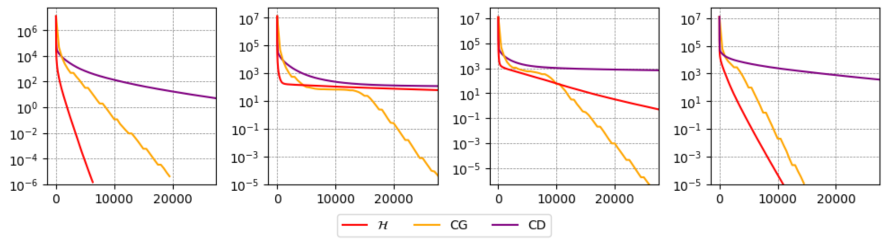

Figure 4: Decay of the map D for varying values of "nugget" added to $\mathbf{Q}$; 0, 0.2, 2, 20 (left to right). We compare BI and $\mathcal{H}$ for the relaxed map R with popular methods (conjugate gradient and coordinate descent for D) by the number of matrix-column calls. The quadratic maps are generated using random $\mathbf{Q}$, $\mathbf{c}$ and $\boldsymbol{\alpha}$ for $N = 500$ and $\text{rank}(\mathbf{Q}) = 250$.

Concluding remarks

Summary

- Study of the properties of the maps resulting from the introduction of an invariance under rescaling into convex quadratic maps.
- Investigation of behaviours of coordinate descent algorithms arising from the minimisation of such maps.

Ongoing investigations and future work

- Explore more numerical experiments and applications of the presented method.
- Gain a better understanding of the situations in which the acceleration occurs.

Thank you for your attention