Analyzing Optimization Algorithms using Integral Quadratic Constraints

Laurent Lessard[1], Benjamin Recht[2], and Andrew Packard[2]

[1]University of Wisconsin–Madison [2]University of California, Berkeley

Abstract

Iterative optimization algorithms are a main engine behind large-scale data processing applications such as computer vision and machine learning. However, the design and use of such algorithms is currently more art than science. We present a new analysis method for optimization algorithms that is based on robust control theory. This framework allows one to easily compute robust performance bounds for a variety of algorithms by solving small convex programs. Rather than testing different algorithms to see which ones perform best, we can now prescribe desired properties e.g. "robust to 5% noise" and then *design the best algorithm* that meets the specification.

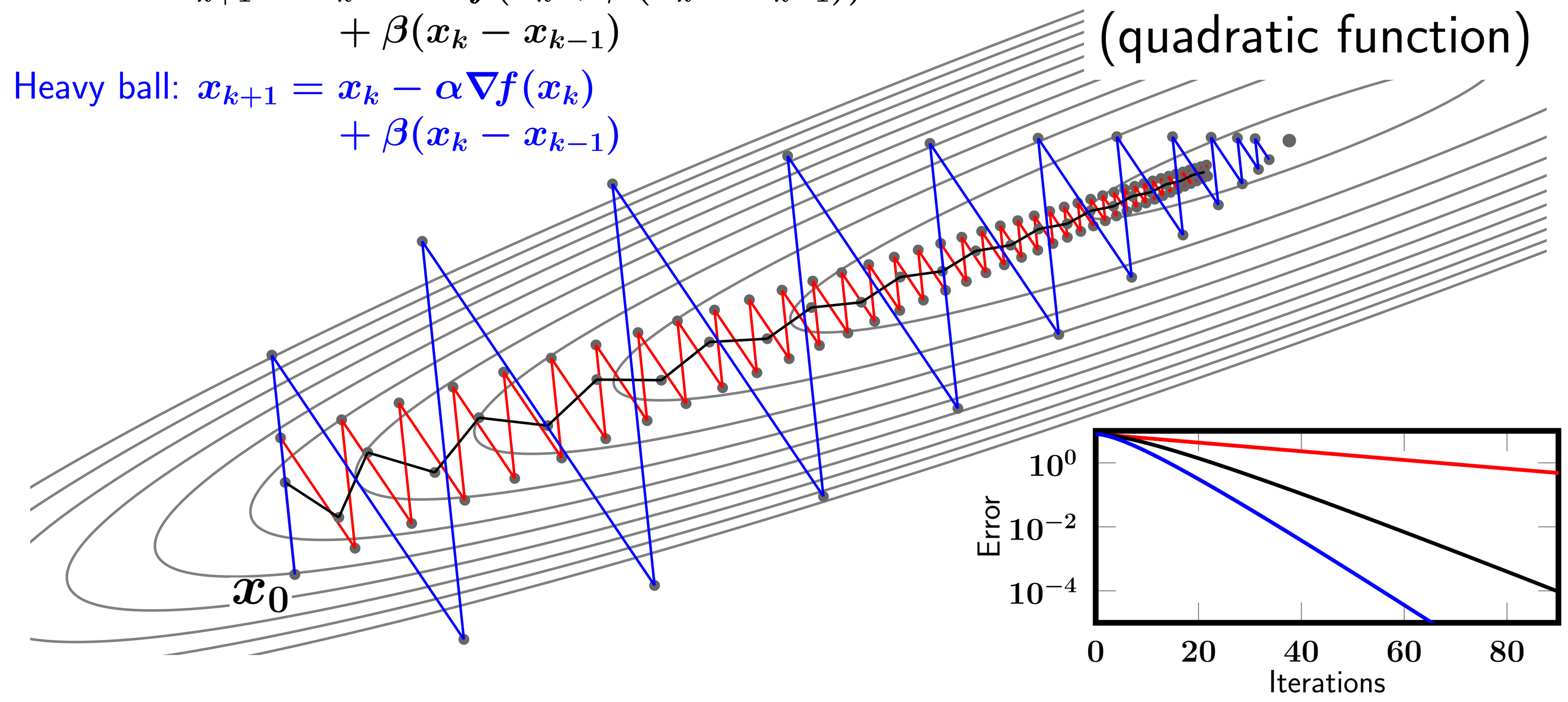
Iterative optimization algorithms

Gradient: $x_{k+1} = x_k - \alpha\nabla f(x_k)$

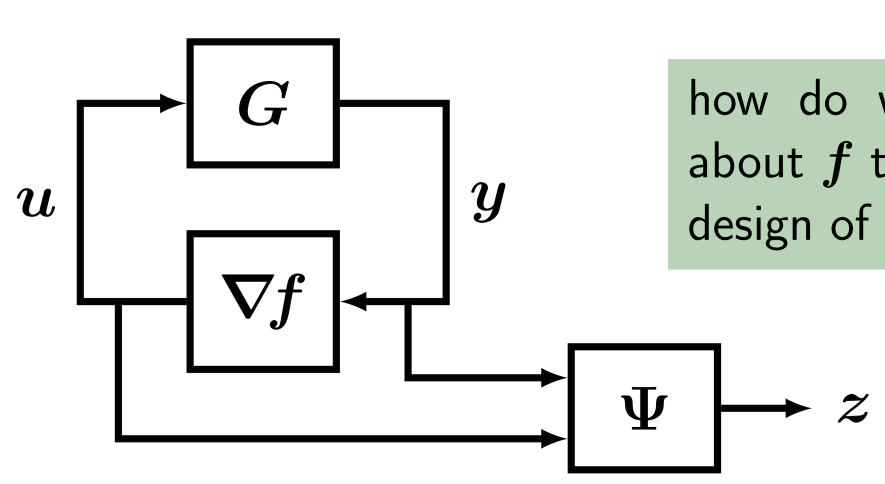
Nesterov: $x_{k+1} = x_k - \alpha\nabla f(x_k + \beta(x_k - x_{k-1})) + \beta(x_k - x_{k-1})$

Heavy ball: $x_{k+1} = x_k - \alpha\nabla f(x_k) + \beta(x_k - x_{k-1})$

contours of $f(x)$ (quadratic function)



Robust control formulation



how do we leverage knowledge about f to inform our analysis or design of the algorithm G?

- G : discrete-time linear dynamical system (the iterative algorithm)
- f : uncertain function that we will be minimizing
- Ψ : filter that characterizes the input-ouput properties of f

IQCs for characterizing nonlinearities

If f is strongly convex: $mI \preceq \nabla^2 f \preceq LI$ and is minimized at $\nabla f(x_\star) = 0$, then for any $\{y_0, y_1, \dots\}$ with $u_k := \nabla f(y_k)$,

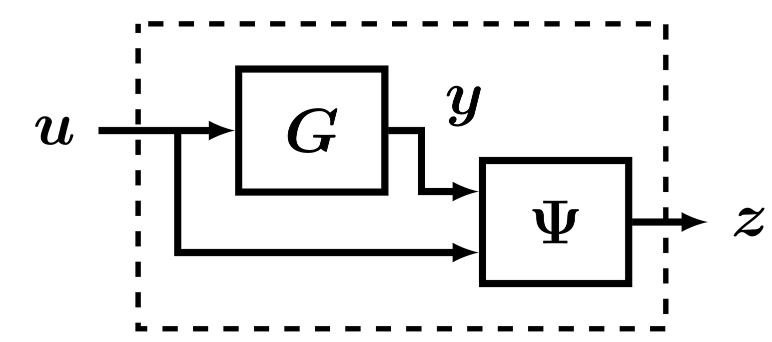
$$\sum_{t=0}^{N} \rho^{-2k}\,(z_k - z_\star)^\mathsf{T} M (z_k - z_\star) \geq 0 \quad \forall N \text{ and } 0 \leq \rho \leq 1$$

$$\Psi := \left[\begin{array}{c|cc} 0 & -L & 1 \\ \hline \rho^2 & L & -1 \\ 0 & -m & 1 \end{array}\right] \quad \text{and} \quad M := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

This is an example of a Zames-Falb IQC.

Main result [SIOPT'16]

Remove ∇f from the block diagram, obtain



$$z = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] u$$

Let x be the combined state of (G, Ψ).
If the following SDP is feasible for $P \succ 0$ and $\lambda \geq 0$

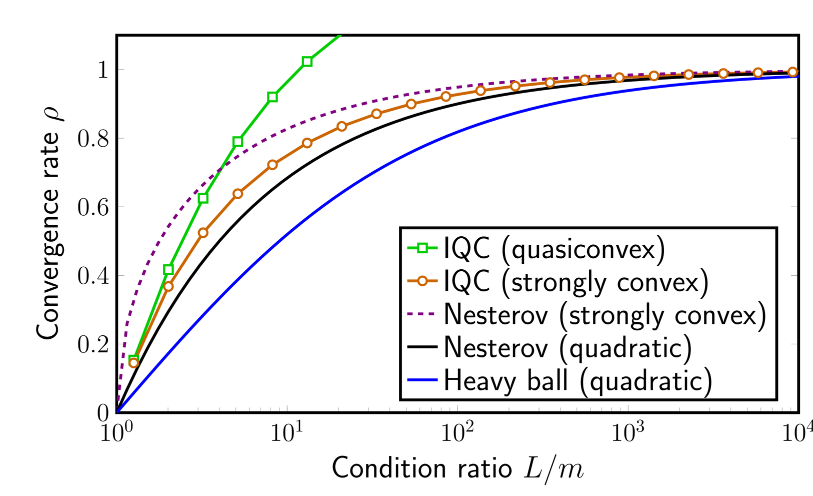
$$\begin{bmatrix} A^\mathsf{T} P A - \rho^2 P & A^\mathsf{T} P B \\ B^\mathsf{T} P A & B^\mathsf{T} P B \end{bmatrix} + \lambda \begin{bmatrix} C & D \end{bmatrix}^\mathsf{T} M \begin{bmatrix} C & D \end{bmatrix} \preceq 0$$

Then we have exponential convergence:

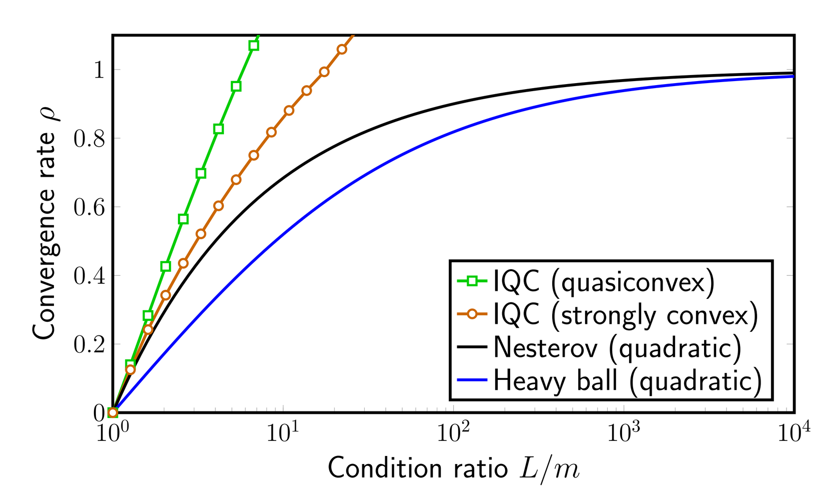
$$\|x_k - x_\star\| \leq \sqrt{\operatorname{cond}(P)}\,\rho^k\,\|x_0 - x_\star\|$$

Case study: Nesterov and Heavy ball

What is the best bound on the rate of these algorithms if we assume f is quadratic, has sector-bounded gradients, or is strongly convex?



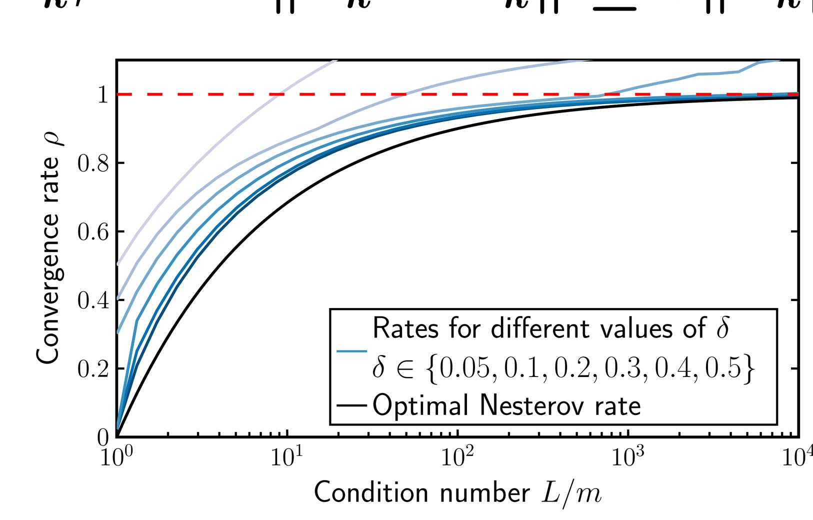
Nesterov: IQC upper bound is strictly tighter than the best-known bound (dashed purple).

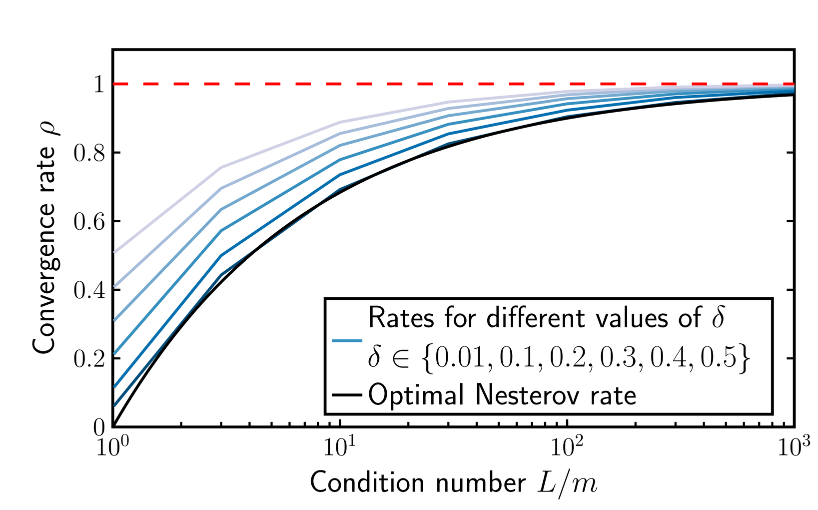


Heavy ball: IQC upper bound suggests Heavy ball is not stable for strongly convex f (verified!)

Case study: noise robustness

How robust is an algorithm to gradient noise? If we use $\bar{u}_k$ instead of u_k, where $\|u_k - \bar{u}_k\| \leq \delta\|u_k\|$, Nesterov's method is not robust.





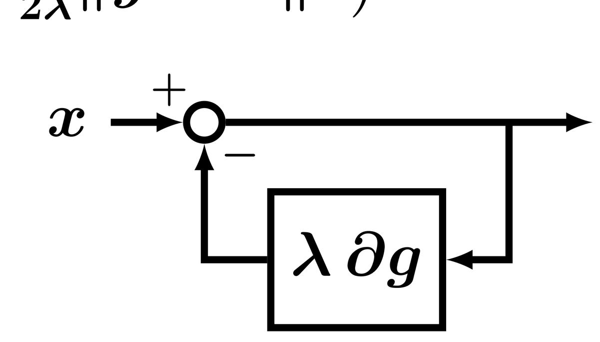
Robustness recovered by designing $(\alpha, \beta_1, \beta_2)$ in a new algorithm:
$x_{k+1} = x_k - \alpha\nabla f(x_k + \beta_1(x_k - x_{k-1})) + \beta_2(x_k - x_{k-1})$

Generalizations

The proximal operator defined below can be represented as a block

$$\operatorname{prox}_{\lambda g}(x) := \arg\min_y \bigl(g(y) + \tfrac{1}{2\lambda}\|y - x\|^2 \bigr)$$

diagram and ∂g can be characterized using IQCs. This allows analysis of constrained optimization, proximal point methods, operator-splitting methods (e.g. ADMM), ...



Laurent Lessard (laurent.lessard@wisc.edu): http://www.laurentlessard.com