Inexact reduced gradient methods in smooth nonconvex optimization

Dat Tran
he9180@wayne.edu
Department of Mathematics

the talk is given at the 24th Midwest Optimization Meeting
based on a joint work with Pham Duy Khanh (HCMC University of Education, Vietnam, khanhpd@hcmue.edu.vn) and Boris S. Mordukhovich (Wayne State University, USA, aa1086@wayne.edu)

October 29th, 2022

Targeted problems

Consider the optimization problem of the form

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad x \in \mathbb{R}^n \tag{1}$$

with a continuously differentiable ($\mathcal{C}^1$-smooth) objective function $f : \mathbb{R}^n \to \mathbb{R}$.

Gradient descent finds stationary points

A necessary condition for $\bar{x} \in \mathbb{R}^n$ to be a minimizer of f is

$$\nabla f(\bar{x}) = 0. \tag{2}$$

The point $\bar{x}$ satisfies (2) is called a stationary point. To find such a point the *gradient descent* methods construct the iterative procedure

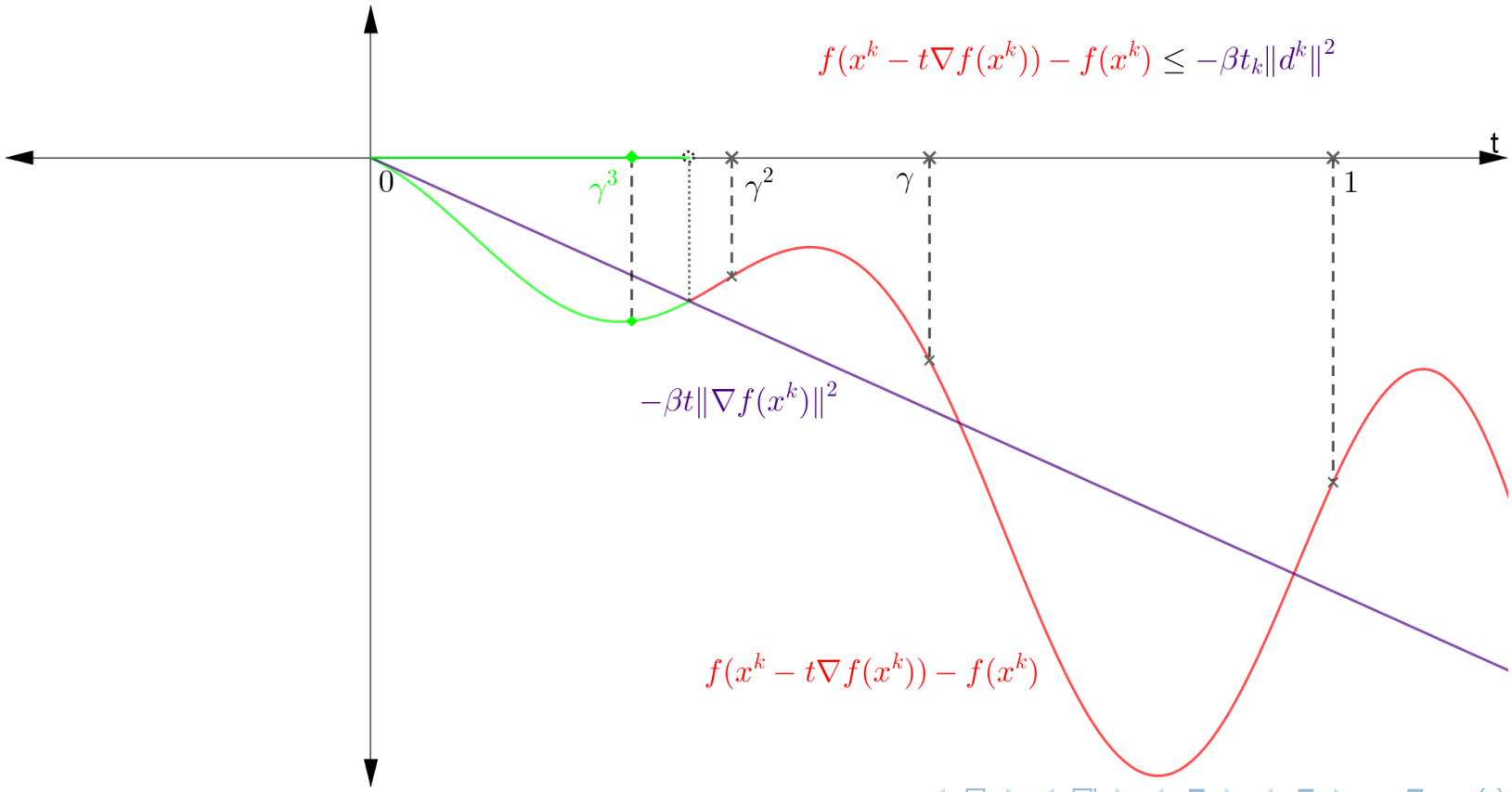
$$x^{k+1} := x^k - t_k \nabla f(x^k) \text{ for all } k \in \mathbb{N}, \tag{3}$$

where $t_k \geq 0$ is a stepsize at the k^{th} iteration, and where $\nabla f(x^k)$ is the gradient of f at x^k.

Stepsize selections - Backtracking stepsize

The stepsize sequence $\{t_k\}$ satisfies the **Armijo rule** if there exist a scalar $\beta \in (0,1)$ and a reduction factor $\gamma \in (0,1)$ such that for all $k \in \mathbb{N}$ we have the representation

$$t_k = \max_{t \in \{1,\gamma,\gamma^2,\ldots\}} \left\{ t \middle| \ f(x^k - t\nabla f(x^k)) - f(x^k) \leq -\beta t \left\|\nabla f(x^k)\right\|^2 \right\}. \quad (4)$$



Stepsize selections - Constant and Diminishing

$L-$Lipschitz continuity

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \text{ for all } x, y \in \mathbb{R}^n.$$

Constant stepsize: $t_k = \tau \in \left(0, \frac{2}{L}\right)$ for all $k \in \mathbb{N}$. A smaller modulus L gives a broader range of selections for τ.

Diminishing stepsize: $t_k \downarrow 0$ and $\sum_{k=1}^{\infty} t_k = \infty$, e.g., $t_k = \frac{1}{k}$.

For more discussions on gradient descent methods and its variants, see [Ber16, IS14, Nes18, NW16, Pol87, Rus06].

Advantages and disadvantages of stepsize selections

	Class of functions	Compute L	Speed	fval
Backtracking	$\mathcal{C}^1$	No	Moderate	Yes
Constant	$\mathcal{C}^1$+∇f $L-$Lipschitz	Yes	Fast	No
Diminishing	$\mathcal{C}^1$+∇f $L-$Lipschitz	No	Slow	No

Table: Comparison between stepsize selections

A relaxation of $L-$Lipschitz continuity of ∇f

The $L-$Lipschitz continuity of ∇f yields the $L-$descent condition

$$f(y) \leq \underbrace{f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2}_{\text{Quadratic function } f_{x,L}(y) \text{ with amplitude } \frac{L}{2}} \quad \text{for all} \quad x, y \in \mathbb{R}^n. \tag{5}$$

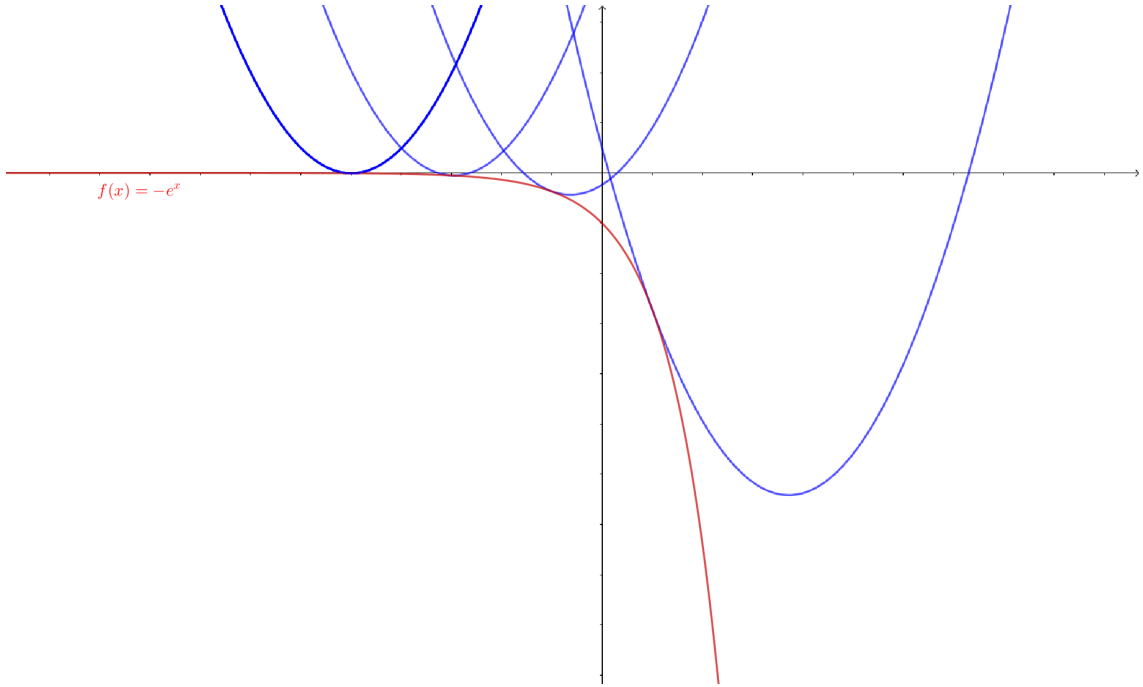


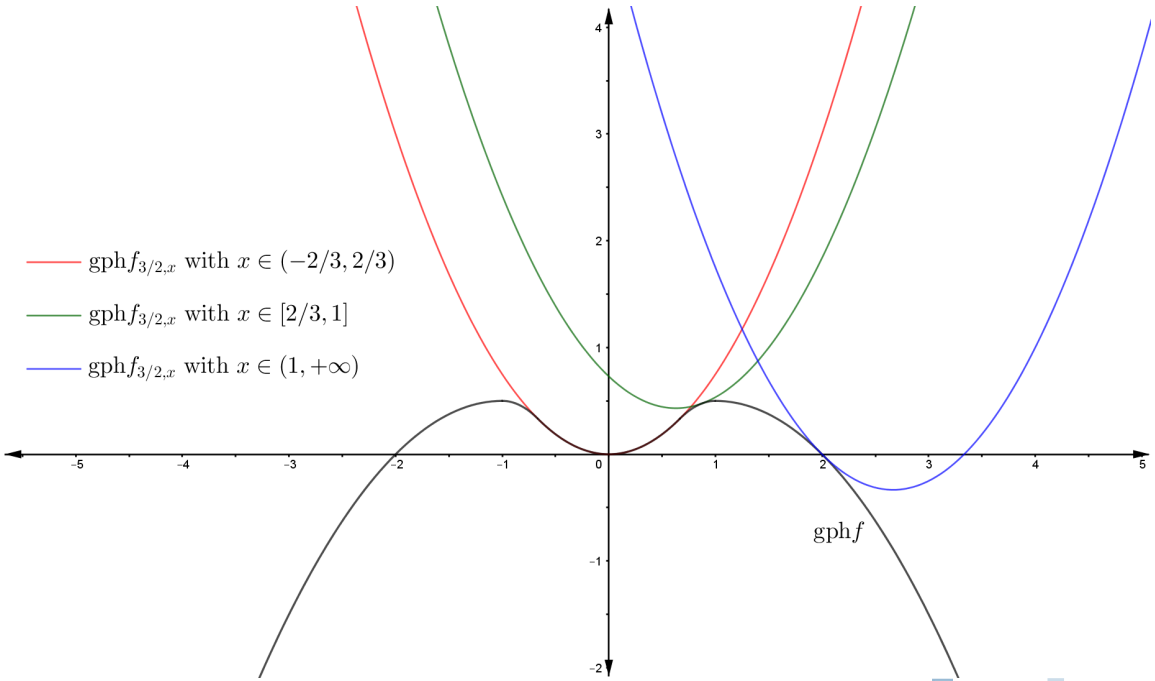
Figure: An $L-$descent function that does not have Lipschitz gradient

A relaxation of $L-$Lipschitz continuity of ∇f

Consider the univariate function f, where

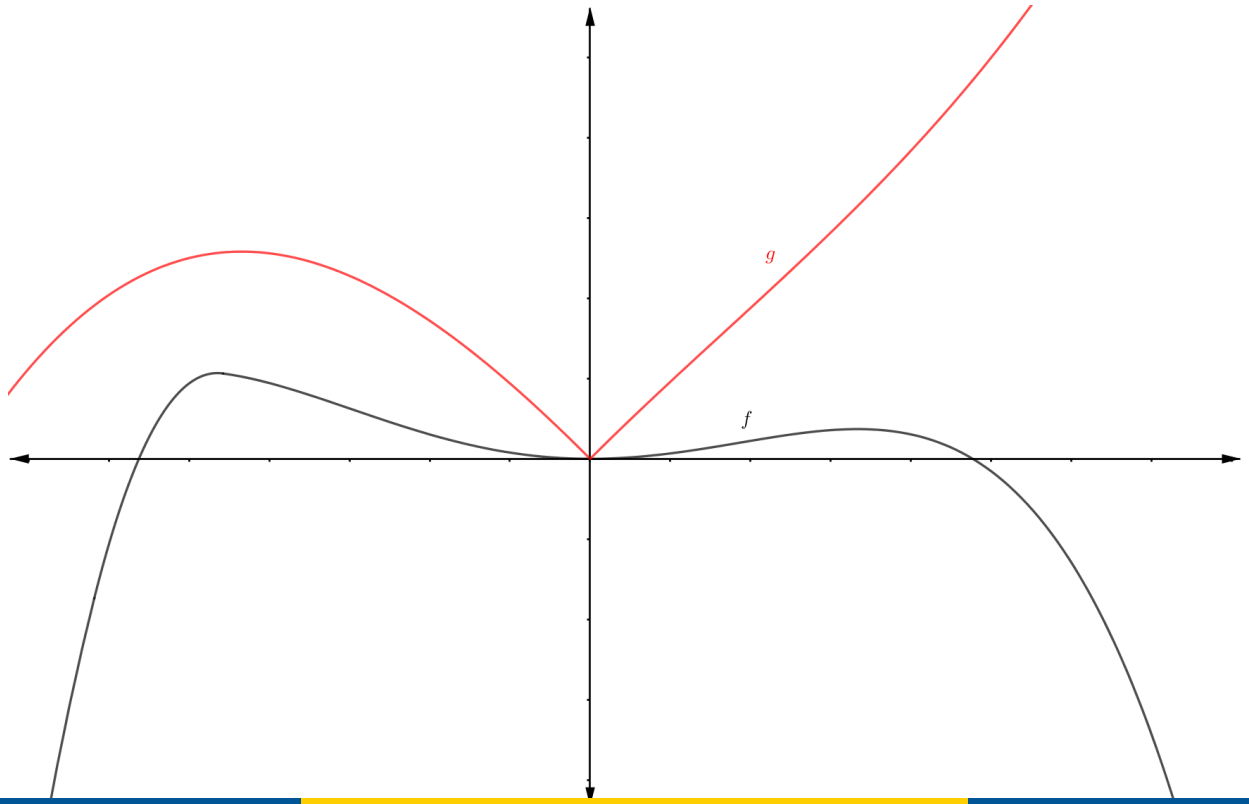
$$f(x) = \begin{cases} \frac{3}{4}x^2 & \text{if } |x| < \frac{2}{3}, \\ -\frac{3}{2}x^2 + 3x - 1 & \text{if } \frac{2}{3} \leq x \leq 1, \\ -\frac{3}{2}x^2 - 3x - 1 & \text{if } -1 \leq x \leq -\frac{2}{3}, \\ |x| - \frac{x^2}{2} & \text{if } |x| > 1. \end{cases}$$

Then ∇f is $3-$Lipschitz while f satisfies the $3/2$-descent proper.



Gradient descent methods with errors

- When the function f is a black-box function, i.e., does not have an analytic form [ACL11] .
- When the function f is noisy, [GK95].
- When the function f is a smoothing version (Moreau envelope/Forward-backward envelope) of another nonsmooth function g [RW98,STP17,THP20].



Main concerns

Question 1

Does the convergence results hold when errors appear in the calculation $\nabla f(x^k)$?

Main concerns

Question 1

Does the convergence results hold when errors appear in the calculation $\nabla f(x^k)$?

Answer 1 (Ber16, Section 1.2)

When the $\mathcal{C}^1-$*smooth function* $f : \mathbb{R}^n \to \mathbb{R}$ *has a Lipschitz gradient, the gradient descent method with diminishing step-size has stationary accumulation points.*

Main concerns

Question 1

Does the convergence results hold when errors appear in the calculation $\nabla f(x^k)$?

Answer 1 (Ber16, Section 1.2)

When the $\mathcal{C}^1-$smooth function $f : \mathbb{R}^n \to \mathbb{R}$ has a Lipschitz gradient, the gradient descent method with diminishing step-size has stationary accumulation points.

Question 2

How about the other types of step-size and the general class of $\mathcal{C}^1$-smooth functions?

Main contributions

Design the novel **Inexact reduced gradient** (IRG) methods with 3 stepsize selections:

1. $\mathcal{C}^1-$smooth functions: backtracking stepsizes.
2. $\mathcal{C}^1$-smooth functions that satisfy $L-$descent condition: constant and diminishing step stepsizes.

What we have achieved?

- Stationary accumulation points for all methods.
- *Global convergence* for all methods under the *Kurdyka-Łojasiewicz* (*KL*) *property* of the objective functions.
- *Linear convergence rates* for methods using backtracking stepsizes and constant stepsize.

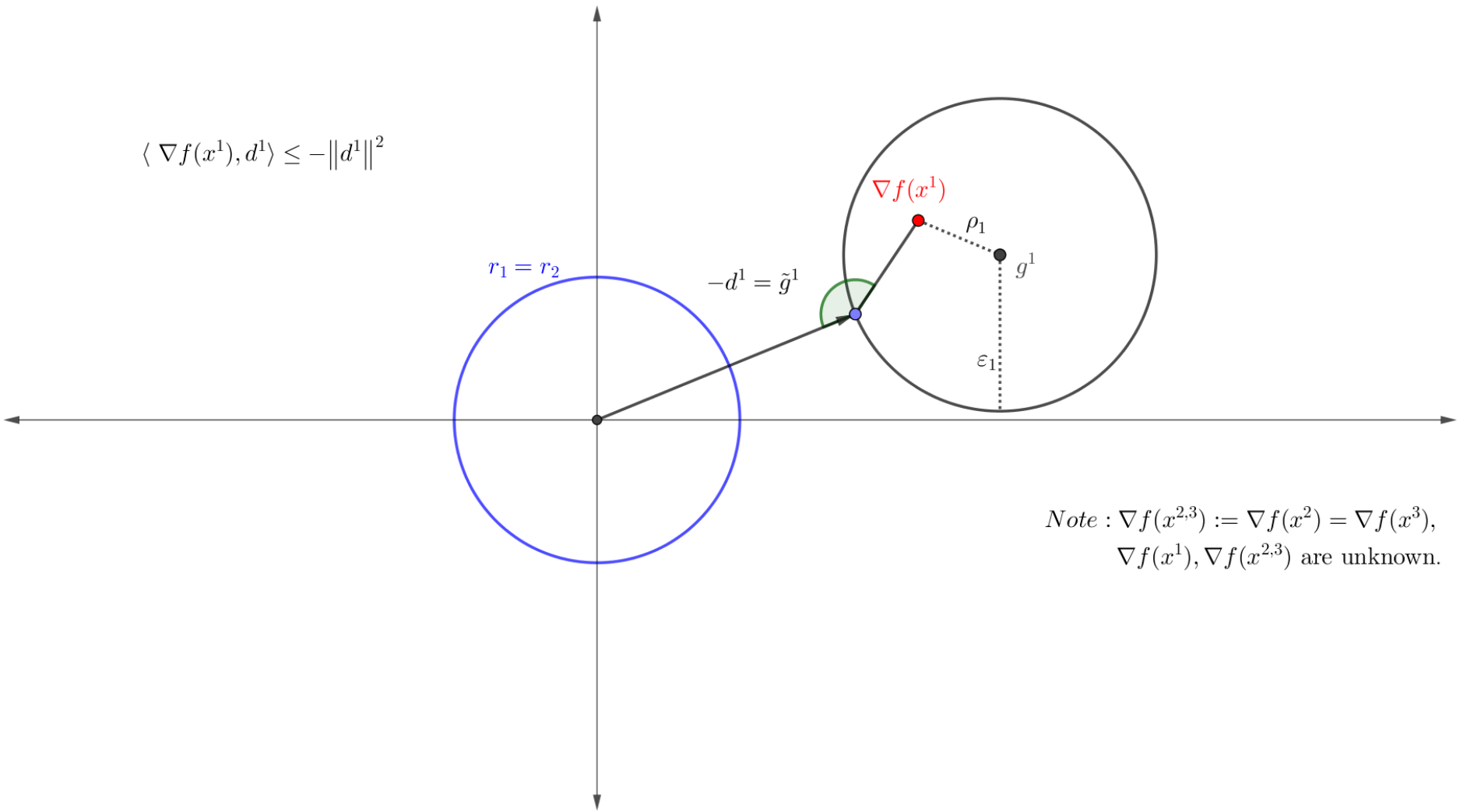
Our ideas

1. Calculate an arbitrary inexact gradient g^k satisfying

$$\left\|\nabla f(x^k) - g^k\right\| \leq \delta_k.$$

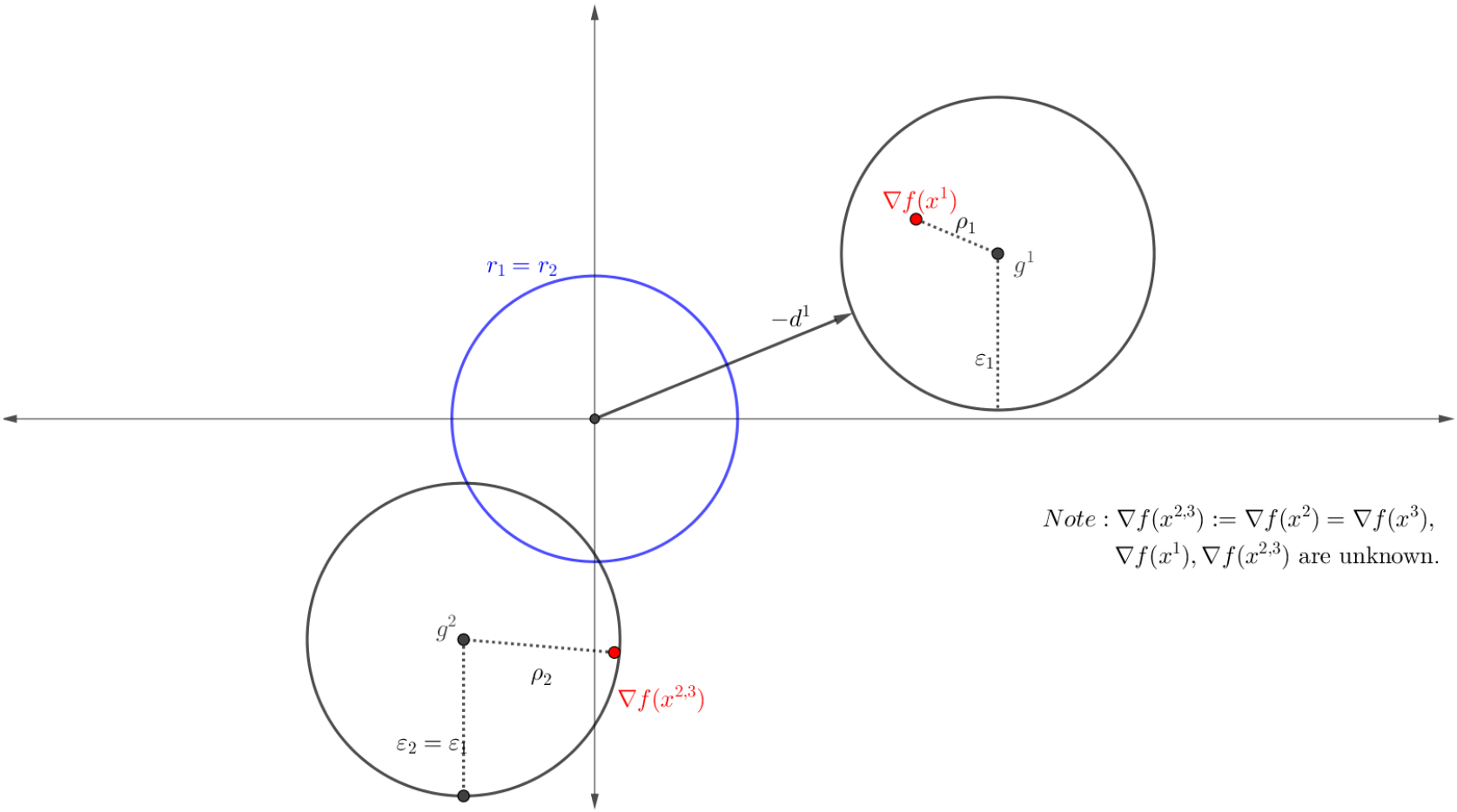
2. Choose $\widetilde{g}^k$ near g^k that have *a better property* than g^k.
3. Choose $d^k = -\widetilde{g}^k$ as a descent direction.

Geometric representation



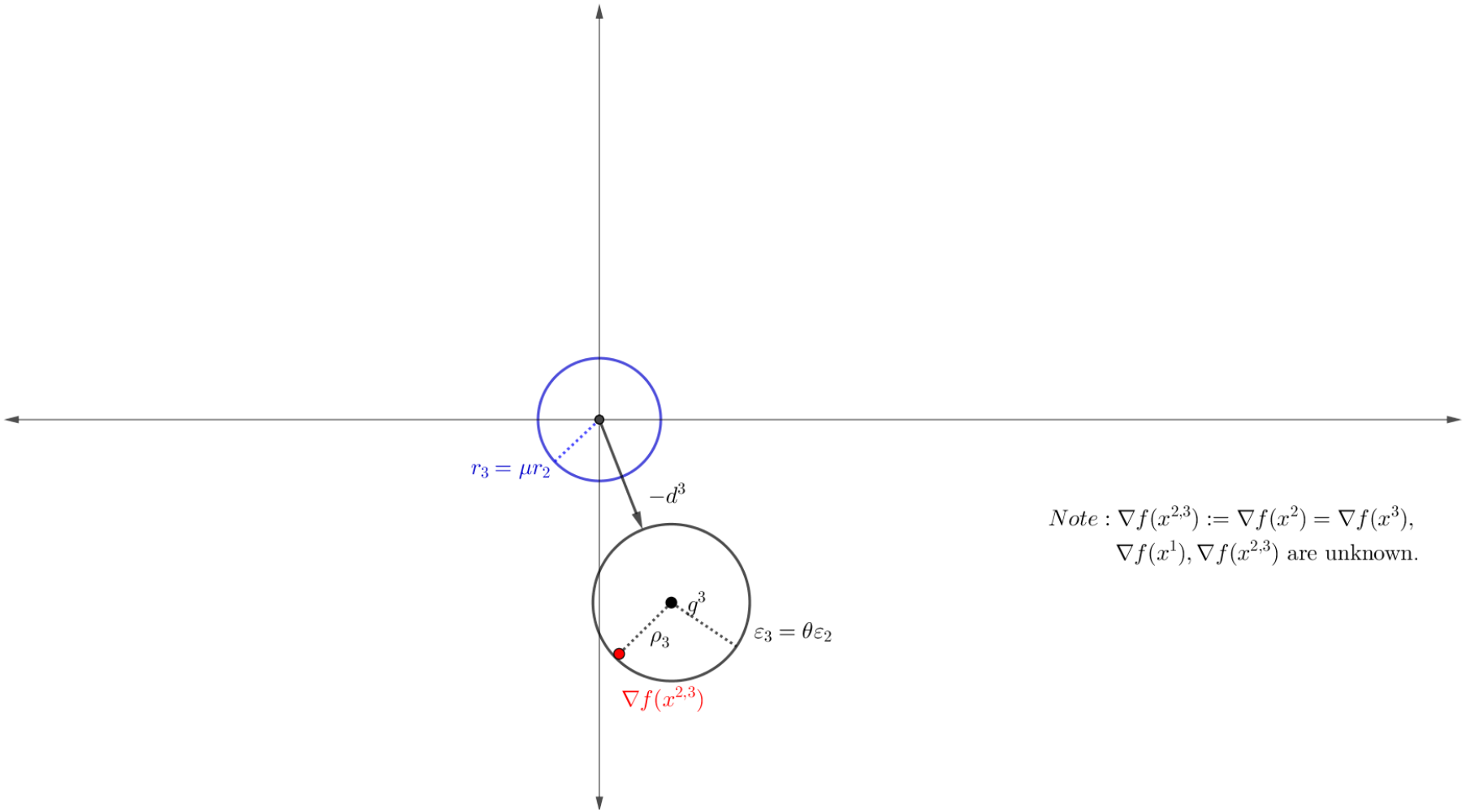
Note : $\nabla f(x^{2,3}) := \nabla f(x^2) = \nabla f(x^3)$,
$\nabla f(x^1), \nabla f(x^{2,3})$ are unknown.

Geometric representation



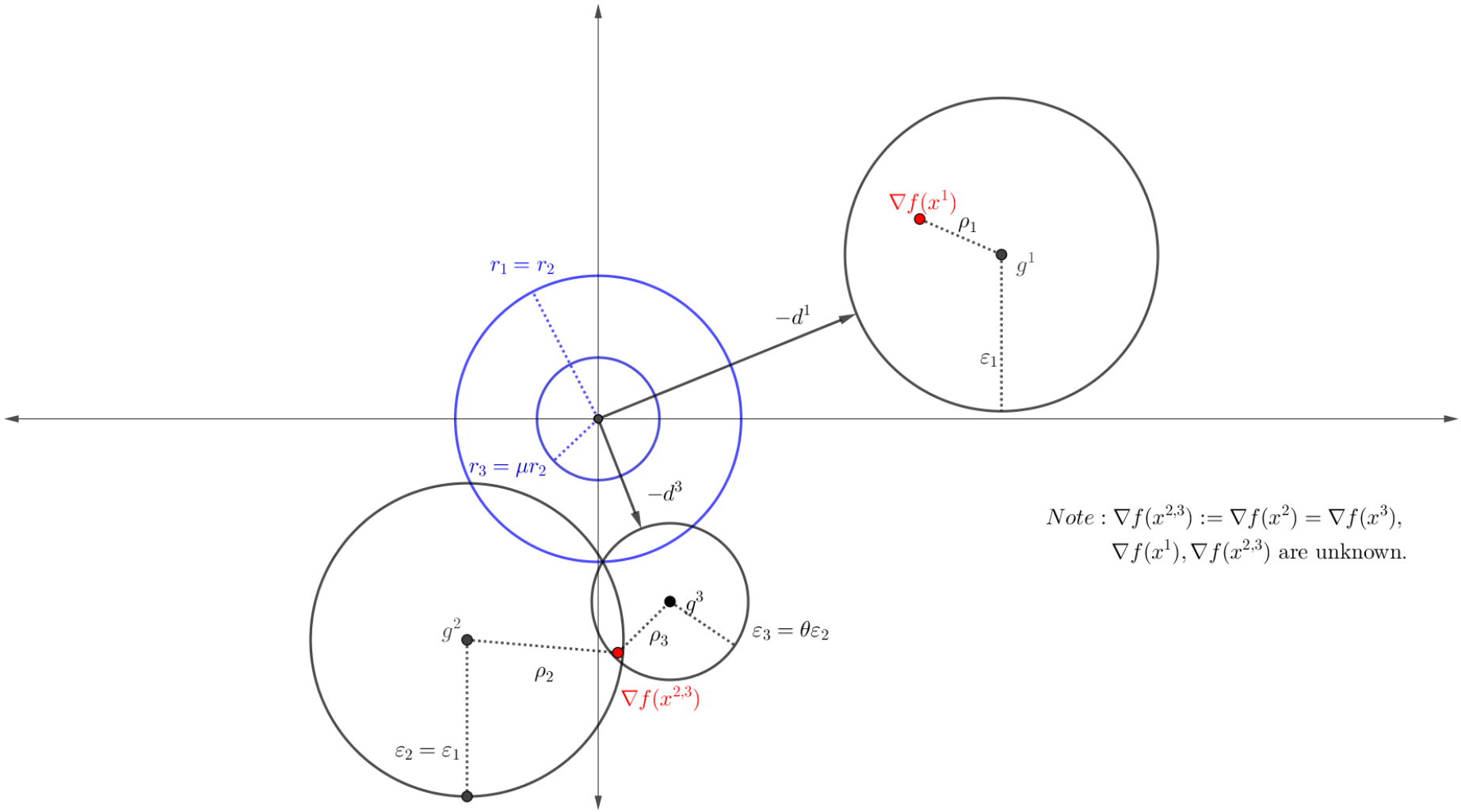
$Note : \nabla f(x^{2,3}) := \nabla f(x^2) = \nabla f(x^3),$
$\nabla f(x^1), \nabla f(x^{2,3})$ are unknown.

Geometric representation



$Note: \nabla f(x^{2,3}) := \nabla f(x^2) = \nabla f(x^3)$,
$\nabla f(x^1), \nabla f(x^{2,3})$ are unknown.

Geometric representation



$Note: \nabla f(x^{2,3}) := \nabla f(x^2) = \nabla f(x^3)$,
$\nabla f(x^1), \nabla f(x^{2,3})$ are unknown.

General framework

Step 0. (initialization) Select an initial point $x^1 \in \mathbb{R}^n$, initial radii $\varepsilon_1, r_1 > 0$, radius reduction factors $\mu, \theta \in (0,1)$, and a sequence of gradient errors $\{\rho_k\} \downarrow 0$.

Step 1. (inexact gradient and stopping criterion) Choose g^k such that

$$\left\| g^k - \nabla f(x^k) \right\| \leq \min\{\rho_k, \varepsilon_k\}. \tag{6}$$

If $\|g^k\| = \rho_k = 0$, then stop.

Step 2. (radius update) If $\|g^k\| \leq r_k + \varepsilon_k$, then set $r_{k+1} := \mu r_k$, $\varepsilon_{k+1} := \theta \varepsilon_k$, $d^k := 0$, and go to Step 3. Otherwise, set $r_{k+1} := r_k$, $\varepsilon_{k+1} := \varepsilon_k$, and

$$d^k := -\mathrm{Proj}(0, \mathbb{B}(g^k, \varepsilon_k)) = -\frac{\|g^k\| - \varepsilon_k}{\|g^k\|} g^k. \tag{7}$$

Step 3. (stepsize) Choose $t_k > 0$ by a specific rule.

Step 4. (iteration update) Set $x^{k+1} := x^k + t_k d^k$. Increase k by 1 and go back to Step 1.

Our goals

Deriving the following assertions:

1. Every accumulation point of $\left\{x^k\right\}$ is a stationary point of f.
2. The sequence $\left\{x^k\right\}$ is convergent.

Backtracking step-size for inexact gradient

Choose a line search scalar $\beta \in (0,1)$, a reduction factor $\gamma \in (0,1)$, and an artificial stepsize at stagnant iterations $\tau \in (0,1)$.

Backtracking stepsize

If $d^k = 0$, then put $t_k := \tau$. Otherwise, we set

$$t_k := \max\left\{t \;\middle|\; f(x^k + td^k) \leq f(x^k) - \beta t\left\|d^k\right\|^2,\ t = 1,\ \gamma,\ \gamma^2, \dots\right\}. \quad (8)$$

Key lemma

Lemma 1 (KMT22)

Let $\{x^k\}$ *and* $\{d^k\}$ *be sequences satisfying*

$$\sum_{k=1}^{\infty} \left\|x^{k+1} - x^k\right\| \cdot \left\|d^k\right\| < \infty. \tag{9}$$

If $\bar{x}$ *is an accumulation point of* $\{x^k\}$ *and if* 0 *is an accumulation point of* $\{d^k\}$*, then there exists an infinite set* $J \subset \mathbb{N}$ *such that*

$$x^k \xrightarrow{J} \bar{x} \quad \text{and} \quad d^k \xrightarrow{J} 0. \tag{10}$$

Stationary accumulation points

When f is $\mathcal{C}^1-$smooth and step-size is backtracking, or when f satisfies $L-$descent condition and step-size is constant or diminishing.

Theorem 2 (KMT22)

(i) *Every accumulation point of* $\{x^k\}$ *is a stationary point of* f.

(ii) *If the sequence* $\{x^k\}$ *is bounded, then the set of accumulation points of* $\{x^k\}$ *is nonempty, compact, and connected.*

(iii) *If* $\{x^k\}$ *has an isolated accumulation point, then the entire sequence* $\{x^k\}$ *converges to this point.*

Kurdyka-Łojasiewicz property

Definition 3 (AMA6)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. We say that f satisfies the *KL property* at $\bar{x} \in \mathbb{R}^n$ if there exist a number $\eta > 0$, a neighborhood U of $\bar{x}$, and a nondecreasing function $\psi : (0, \eta) \to (0, \infty)$ such that the function $1/\psi$ is integrable over $(0, \eta)$ and we have

$$\|\nabla f(x)\| \geq \psi\big(f(x) - f(\bar{x})\big) \tag{11}$$

for all $x \in U$ with $f(\bar{x}) < f(x) < f(\bar{x}) + \eta$.

Remark 1

The KL property is satisfied at every point $x \in \mathbb{R}^n$ when f is either

- analytic
- semi-algebraic (graph is built up by polynomial inequalities)
- definable in o-minimal structures

Global convergence

From now, assume that the radius reduction factors satisfy $\theta < \mu$. When f is $\mathcal{C}^1-$smooth and step-size is backtracking, or when f satisfies $L-$descent condition and step-size is constant or diminishing.

Theorem 4 (KMT22)

Assume that f satisfies the KL property at some accumulation point $\bar{x}$ of $\{x^k\}$. Then $x^k \to \bar{x}$.

Rate of convergence

Definition 5

The k^{th} iteration of Algorithm 1 is called a *null iteration* if $x^{k+1} = x^k$. The set of all null iterations is denoted by

$$\mathcal{N} := \left\{ k \in \mathbb{N} \mid x^{k+1} = x^k \right\}.$$

Remark 2

If the set of non-null iterations is finite, $\{x^k\}$ converges finitely to some stationary point $\bar{x}$. So we consider the case this set is infinite.

Assumption 1

The non-null iterations sequence $\{z^k\}$ has an accumulation point $\bar{z}$, that f satisfies the KL property at $\bar{z}$ with $\psi(t) = Mt^q$ for some $M > 0$ and $q \in [0, 1)$.

Rate of convergence

Theorem 6 (KMT22)

Suppose that the step-size considered is backtracking. Assume further that f is bounded from below, and ∇f is locally Lipschitzian around $\bar{z}$. Then

$$z^k \to \bar{x} \text{ R-linearly or Q-linearly.}$$

Rate of convergence

Theorem 6 (KMT22)

Suppose that the step-size considered is backtracking. Assume further that f is bounded from below, and ∇f is locally Lipschitzian around $\bar{z}$. Then

$$z^k \to \bar{x} \text{ R-linearly or Q-linearly.}$$

Theorem 7 (KMT22)

The same rate of convergence holds when f satisfies the $L-$descent condition and step-size is constant.

Numerical experiments

Compare the *efficiency* of our new *IRG methods* with the *reduced gradient* (RG) methods and the *gradient descent* (GD) method, and then check the *sensitivity* of the IRG methods with respect to *error accumulations* in the following settings:

1. The *accuracy* of inexact gradient g^k is *low*, i.e., $\left\|g^k - \nabla f(x^k)\right\| \leq \delta_k$, where δ_k is not too small relative to $\left\|\nabla f(x^k)\right\|$.
2. The *accuracy* $\left\|\nabla f(x^{last})\right\| \leq \nu$ required for the solution is *increasing*,
3. The *dimension* of the objective function is *increasing*.

Testing functions

Test number	Problem name	Dimension	Accuracy
1	Beale	2	0.01
2	Branin	2	0.01
3	Camel	2	0.01
4	Gol	2	0.01
5	Himmel	2	0.01
6	Beale	2	0.001
7	Branin	2	0.001
8	Camel	2	0.001
9	Gol	2	0.001
10	Himmel	2	0.001

Testing functions

Test number	Problem name	Dimension	ν
11	Dixon 20	20	0.01
12	Dixon 500	500	0.01
13	Dixon 2000	2000	0.01
14	Rosen 20	20	0.01
15	Rosen 500	500	0.01
16	Rosen 2000	2000	0.01
17	Dixon 20	20	0.001
18	Dixon 500	500	0.001
19	Dixon 2000	2000	0.001
20	Rosen 20	20	0.001
21	Rosen 500	500	0.001
22	Rosen 2000	2000	0.001

Results

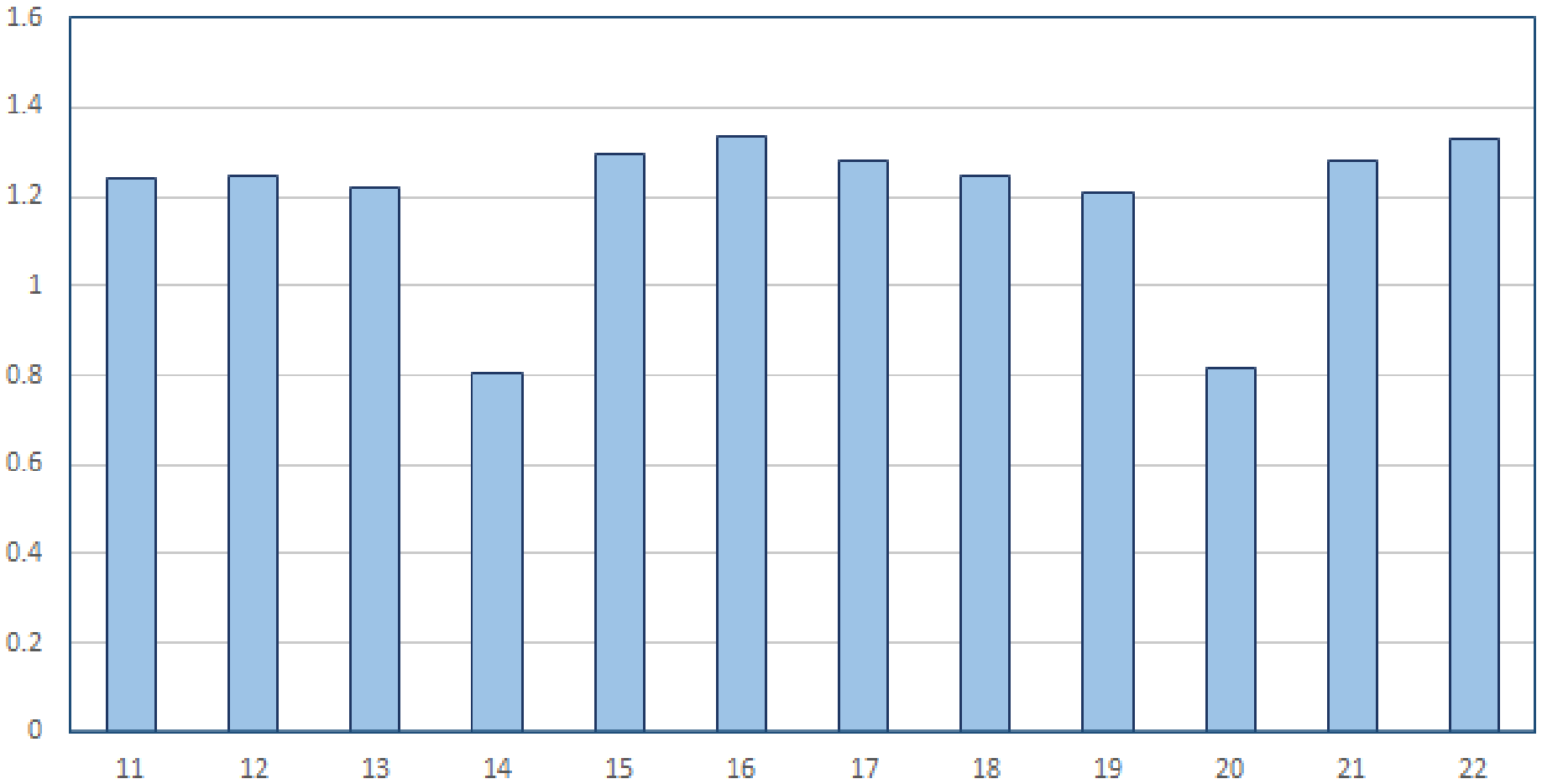


Figure: The ratio between number of iterations of IRGB and GD by tests

Results

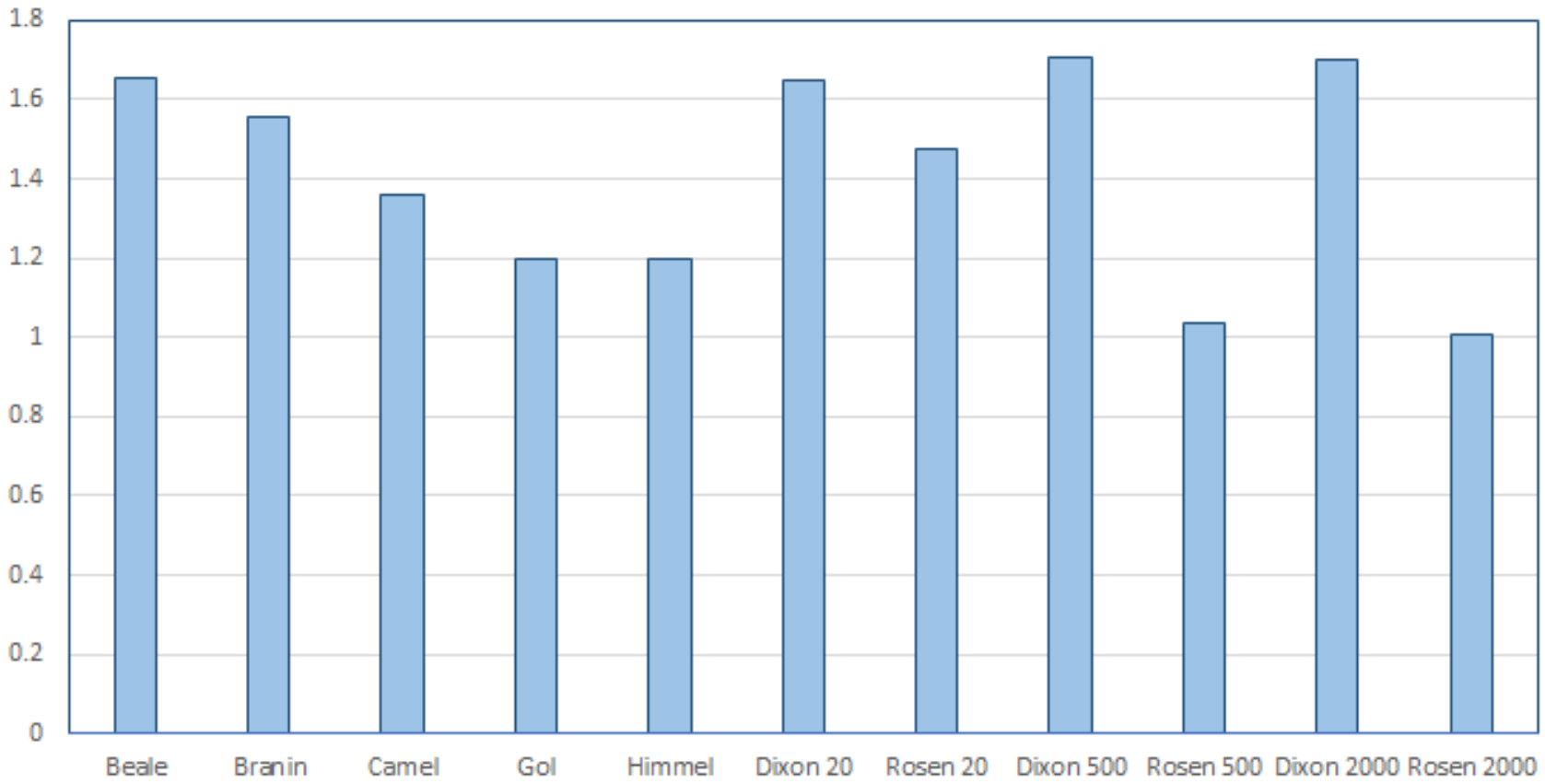


Figure: The ratio between number of iterations of IRGB when $\nu = 0.001$ and when $\nu = 0.01$ by functions

References

[AMA06] P.-A. Absil, R. Mahony and B. Andrews, Convergence of the iterates of descent methods for analytic cost functions, SIAM J. Optim. 16 (2005), 531–547.

[ACL11] A. Addis, A. Cassioli, M. Locatelli and F. Schoen, A global optimization method for the design of space trajectories, Comput. Optim. Appl. 48 (2011), 635–652.

[Ber16] D. P. Bertsekas, Nonlinear Programming, 3rd edition, Athena Scientific, Belmont, MA, 2016.

[FP03] F. Facchinei and J.-S. Pang, Finite-Dimensional Variational Inequalities and Complementarity Problems, Vol. II, Springer, New York, 2003.

[GK95] P. Gilmore and C. T. Kelley, An implicit filtering algorithm for optimization of functions with many local minima, SIAM J. Optim. 5 (1995), 269–285.

References

[KMT22] P. D. Khanh, B. S. Mordukhovich, D. B. Tran, Inexact reduced gradient methods in smooth nonconvex optimization, arxiv.org/abs/2204.01806

[Ost66] A. Ostrowski, Solution of Equations and Systems of Equations, 2nd edition, Academic Press, New York, 1966.

[IS14] A. F. Izmailov and M. V. Solodov, Newton-Type Methods for Optimization and Variational Problems, Springer, New York, 2014.

[Nes18] Yu. Nesterov, Lectures on Convex Optimization, 2nd edition, Springer, Cham, Switzerland, 2018

[NW16] J. Nocedal and S. J. Wright, Numerical Optimization, 2nd edition. New York, 2016.

[Pol87] B. T. Polyak, Introduction to Optimization, Optimization Software, New York, 1987.

References

[RW98] Rockafellar, R.T., Wets R.J-B.: Variational Analysis. Springer, Berlin, 1998
[Rus06] A. Ruszczyński, Nonlinear Optimization, Princeton University Press, Princeton, NJ, 2006.
[STP17] L. Stella, A. Themelis and P. Patrinos, Forward–backward quasi-Newton methods for nonsmooth optimization problems, Comput. Optim. Appl. 67 (2017), 443–487
[THP20] A. Themelis, B. Hermans and P. Patrinos, A new envelope function for nonsmooth DC optimization,