Tight convergence rates for the difference-of-convex algorithm (DCA) and proximal gradient descent (PGD)

Teodor Rotaru [1,2] François Glineur [2] Panos Patrinos [1]

[1]ESAT-STADIUS, KU Leuven, Belgium

[2]ICTEAM-INMA, UCLouvain, Belgium

April 29, 2025

Equivalences of DCA with other first-order methods

- DCA is also known as the convex-concave procedure (CCCP) (Yuille and Rangarajan 2001; Lanckriet and Sriperumbudur 2009; Lipp and Boyd 2016)
- DCA is an instance of the Frank-Wolfe algorithm (Yurtsever and Sra 2022)
- Proximal gradient descent (PGD) is an instance of DCA (Le Thi and Pham Dinh 2018)

Outline

- Convergence rates for the difference-of-convex algorithm (DCA):
  - any choice of curvatures
  - second function may be nonconvex or concave
  - two different performance measures
  - exact analysis of a single iteration
- Improved "DC" splitting via curvature shifting
- Tight convergence rates for proximal gradient descent (PGD)

Motivation

- PGD iterations are equivalent to DCA ones
- No previous tight rates for PGD for nonconvex objectives with stepsizes $> \frac{1}{\text{Lipschitz ct.}}$
- Stepsizes longer than the inverse Lipschitz constant correspond to f_2 **weakly-convex**
- Standard DCA analysis only covers stepsizes $\leq \frac{1}{\text{Lipschitz ct.}}$
- Complicated derivations of previous tight rates for (standard) DCA

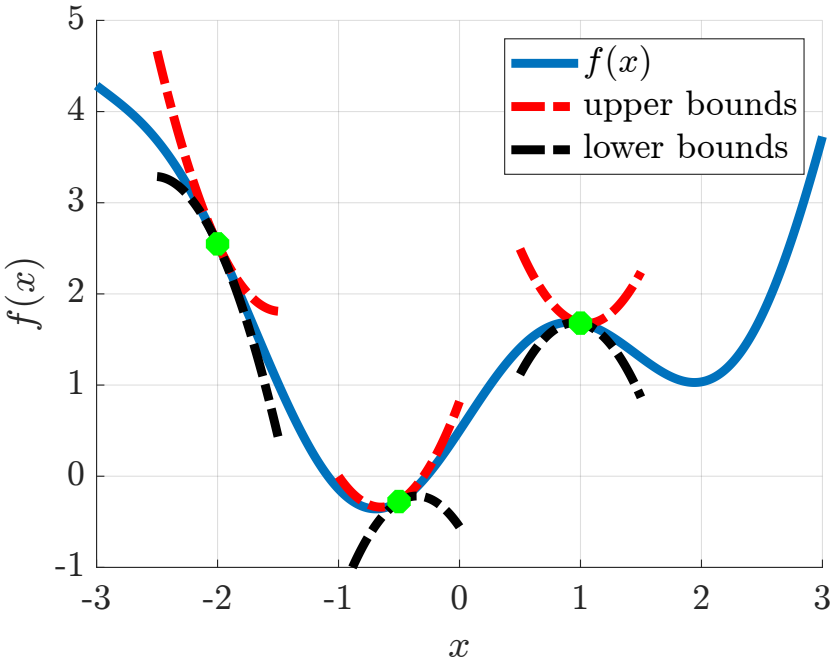
Setup

$$\boxed{\underset{x\in\mathbb{R}^d}{\text{minimize}}\, F(x) \coloneqq f_1(x) - f_2(x)}$$

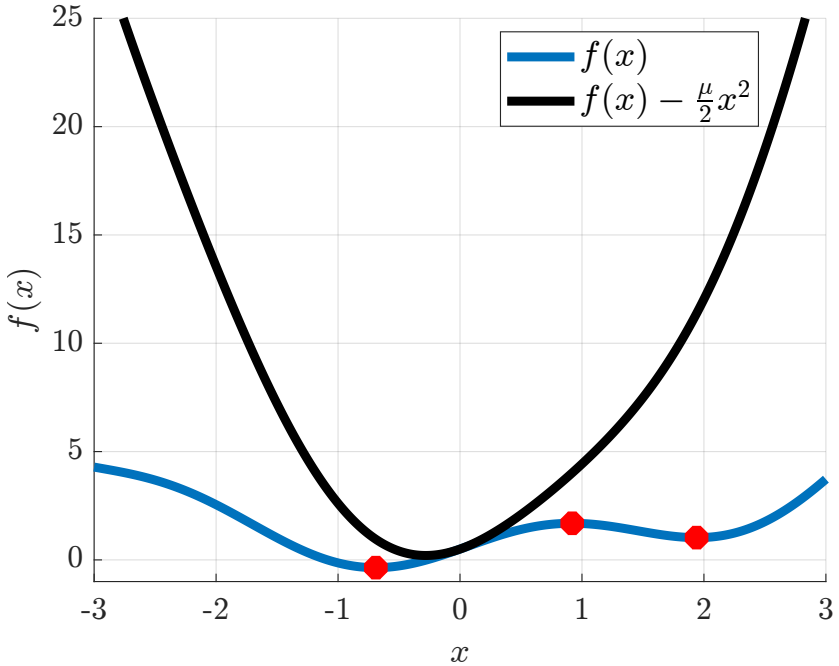
- **Function class:** $f \in \mathcal{F}_{\mu,L}$
  - **Upper** curvature $\boldsymbol{L}$: $\frac{L}{2}\|x\|^2 - f(x)$ is convex;
  - **Lower** curvature $\boldsymbol{\mu}$: $f(x) - \frac{\mu}{2}\|x\|^2$ is convex;
  - When $f \in \mathcal{C}^2$, $\mu I \preceq \nabla^2 f(x) \preceq LI$, i.e., "curvature" $\in [\mu, L]$;
- Assumptions:
  - $L_1 > 0$, $\mu_1 \in [0, L_1)$: {convex, strongly convex};
  - $\mu_2 \in (-\infty, \infty)$, $L_2 \in (\mu_2, \infty]$: {weakly-convex, (strongly) convex; concave ($L_2 < 0$) };
  - $F_{\text{lo}} = \inf_x F(x) > -\infty$;
- N iterations of **DCA**, $k = 0, \ldots, N-1$:

$$\boxed{\begin{aligned}&1.\ \text{Select } g_2^k \in \partial f_2(x^k)\\&2.\ \text{Select } x^{k+1} \in \underset{w\in\mathbb{R}^d}{\text{argmin}}\{f_1(w) - \langle g_2^k, w\rangle\}\end{aligned}} \qquad \text{(DCA)}$$

Smooth weakly-convex functions - example



(a) $f(x)$ weakly-convex



(b) $f(x) - \mu\frac{x^2}{2}$ convex ($\mu < 0$)

- **Intuition**: can be lifted up to a convex function.
- **Quadratic upper bounds** of curvature L: $x \mapsto L\frac{x^2}{2} - f(x)$ is convex
- **Quadratic lower bounds** of curvature μ: $x \mapsto f(x) - \mu\frac{x^2}{2}$ is convex

Proximal gradient descent (PGD)

- Splitting $F(x) \coloneqq \varphi(x) + h(x)$
  - $\varphi \in \mathcal{F}_{\mu_\varphi, L_\varphi}$ smooth (Lipschitz gradient)
  - $h \in \mathcal{F}_{\mu_h, L_h}$
- With stepsize $\gamma > 0$:

$$\boxed{x^+ \in \operatorname{prox}_{\gamma h}\left[x - \gamma \nabla \varphi(x)\right]}$$

- **Proximal mapping:** $\operatorname{prox}_{\gamma h} : \mathbb{R}^n \rightrightarrows \operatorname{dom} h$ with $\gamma > 0$ where $h : \mathbb{R}^n \to \bar{\mathbb{R}}$:

$$\operatorname{prox}_{\gamma h}(x) \coloneqq \operatorname*{argmin}_{w \in \mathbb{R}^n} \left\{ h(w) + \frac{1}{2\gamma} \|w - x\|^2 \right\}$$

- **Optimality condition:**

$$\frac{x - x^+}{\gamma} \in \nabla \varphi(x) + \partial h(x^+)$$

DCA and PGD are iterate equivalent (Le Thi and Pham Dinh 2018)

- **PGD** with stepsize γ applied on $F = \varphi + h \iff$
- **DCA** on $F = f_1 - f_2$ with $f_1 := h + \frac{1}{2\gamma}\|\cdot\|^2$ and $f_2 := \frac{1}{2\gamma}\|\cdot\|^2 - \varphi$.

DCA	$\leftrightarrow$	**PGD**
$F \in \mathcal{F}_{\mu_1 - L_2, L_1 - \mu_2}$	$\leftrightarrow$	$F \in \mathcal{F}_{\mu_\varphi + \mu_h, L_\varphi + L_h}$
μ_1	$\leftrightarrow$	$\gamma^{-1} + \mu_h$
L_1	$\leftrightarrow$	$\gamma^{-1} + L_h$
μ_2	$\leftrightarrow$	$\gamma^{-1} - L_\varphi$
L_2	$\leftrightarrow$	$\gamma^{-1} - \mu_\varphi$
$\mu_1 + \mu_2 > 0$	$\leftrightarrow$	$\gamma < \frac{2}{L_\varphi - \mu_h}$
$\mu_2 \geq 0$	$\leftrightarrow$	$\gamma \leq \frac{1}{L_\varphi}$

Particular cases of PGD

- PGD iteration $x^+ \in \operatorname{prox}_{\gamma h}[x - \gamma \nabla \varphi(x)]$

Gradient descent
$h = 0$

DCA	$\leftrightarrow$	PGD
μ_1	$\leftrightarrow$	γ^{-1}
L_1	$\leftrightarrow$	γ^{-1}
μ_2	$\leftrightarrow$	$\gamma^{-1} - L_\varphi$
L_2	$\leftrightarrow$	$\gamma^{-1} - \mu_\varphi$

Projected GD
$h = \delta_Q$ indicator function of a convex set Q

DCA	$\leftrightarrow$	PGD
μ_1	$\leftrightarrow$	γ^{-1}
L_1	$\leftrightarrow$	∞
μ_2	$\leftrightarrow$	$\gamma^{-1} - L_\varphi$
L_2	$\leftrightarrow$	$\gamma^{-1} - \mu_\varphi$

Proximal point (PPA)
$\varphi = 0$

DCA	$\leftrightarrow$	PGD
μ_1	$\leftrightarrow$	$\gamma^{-1} + \mu_h$
L_1	$\leftrightarrow$	$\gamma^{-1} + L_h$
μ_2	$\leftrightarrow$	γ^{-1}
L_2	$\leftrightarrow$	γ^{-1}

- Complete analysis of gradient descent in Rotaru, Glineur, Patrinos 2024
- DCA analysis only requires 4 parameters vs. PGD (4 curvature parameters + stepsize)

Theoretical background

- **Subdifferential** of proper, lower semicontinuous (l.s.c.) functions
  - *convex* functions:

$$\partial f(x) \coloneqq \{g \in \mathbb{R}^d \,|\, f(y) \geq f(x) + \langle g, y - x\rangle \;\forall y \in \mathbb{R}^d\}$$

  - *weakly-convex* functions: $f \in \mathcal{F}_{\mu,\infty}$, $\mu < 0$

$$\tilde{f}(x) \coloneqq f(x) - \mu\frac{\|x\|^2}{2} \text{ convex } \Rightarrow \partial f(x) \coloneqq \{\tilde{g} + \mu x \,|\, \tilde{g} \in \partial\tilde{f}(x)\}$$

  - If f is differentiable at x: $\partial f(x) = \{\nabla f(x)\}$

- **Range** and **Domains** definitions
  - $\operatorname{dom} f \coloneqq \{x \in \mathbb{R}^d : f(x) < \infty\}$
  - $\operatorname{range} f \coloneqq \{y \in \mathbb{R} : \exists x \in \operatorname{dom} f \text{ with } y = f(x)\}$
  - $\operatorname{dom} \partial f = \{x \in \mathbb{R}^d : \partial f(x) \neq \emptyset\}$
  - $\operatorname{range} \partial f = \cup\{\partial f(x) : x \in \operatorname{dom} \partial f\}$

- **Convex conjugate** of a l.s.c. function f is closed & convex:

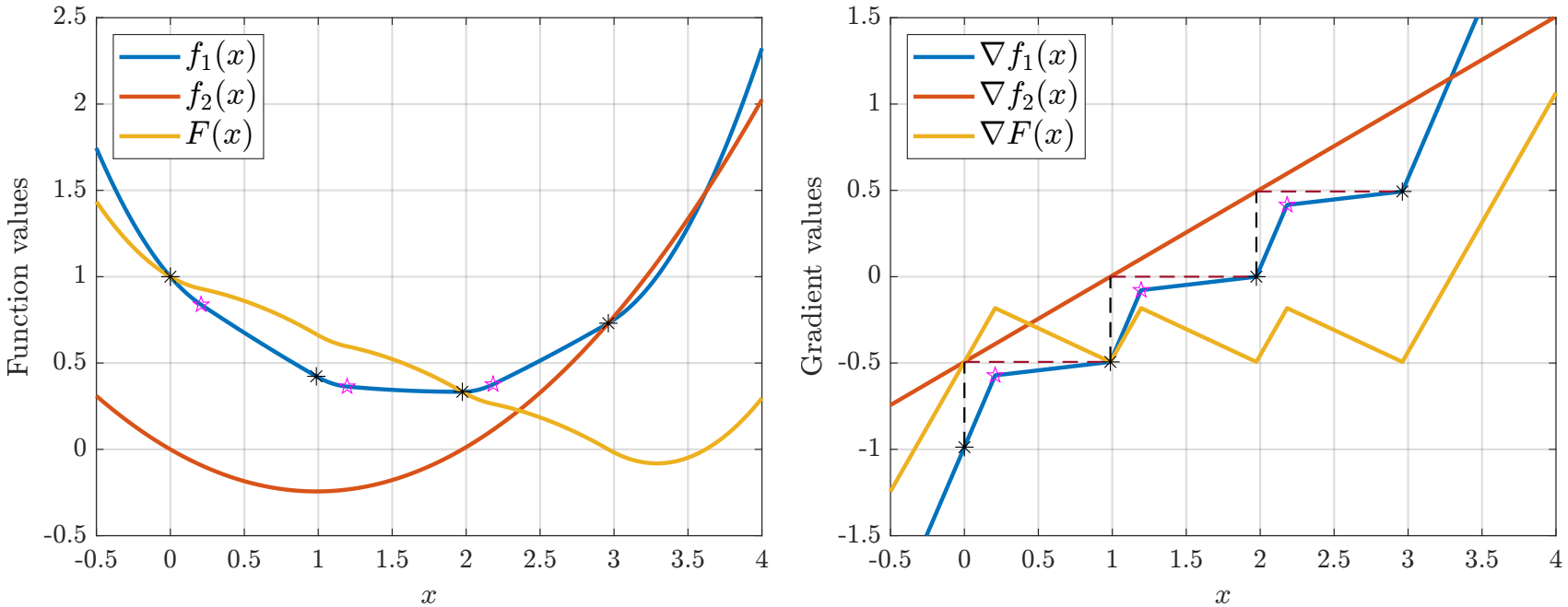
$$f^*(y) \coloneqq \sup_{x \in \operatorname{dom} f} \{\langle y, x\rangle - f(x)\}$$

Assumptions on DCA

$$\boxed{\begin{array}{l}1.\ \text{Select } g_2^k \in \partial f_2(x^k)\\ 2.\ \text{Select } x^{k+1} \in \underset{w\in\mathbb{R}^d}{\operatorname{argmin}}\{f_1(w) - \langle g_2^k, w\rangle\}\end{array}} \qquad \text{(DCA)}$$

- DCA iteration rewrites: $\boxed{x^{k+1} \in \partial f_1^*(\partial f_2(x^k))}$
- **Well-definedness**: $\varnothing \neq \operatorname{dom}\partial f_1 \subseteq \operatorname{dom}\partial f_2$ and $\operatorname{range}\partial f_2 \subseteq \operatorname{range}\partial f_1$.
- **Optimality conditions:** $(\exists)\ g_1^{k+1} \in \partial f_1(x^{k+1})$ such that $g_1^{k+1} = g_2^k$, with $g_2^k \in \partial f_2(x^k)$, $\forall k = 0, \ldots, N-1$.
- **Exact oracles** of ∂f_1 and ∂f_1^*.

Example of DCA iterations



- ***black*** stars: DCA iterates x^k
- ***magenta*** pentagrams: inflection points of f_1

Criticality and stationarity

- **DCA** finds **critical** points!
- **Critical** points $x^\star$: $\partial f_2(x^\star) \cap \partial f_1(x^\star) \neq \emptyset$
- Also **Stationary** points?
  - ✓ f_1 smooth[1] & f_2 smooth: $\nabla F(x^\star) = \nabla f_1(x^\star) - \nabla f_2(x^\star) = 0$
  - ✓ f_1 nonsmooth[2] & f_2 smooth: $0 \in \partial F(x^\star) = \partial f_1(x^\star) - \nabla f_2(x^\star)$ (Rockafellar and Wets 1998)
  - ✗ f_1 smooth & f_2 nonsmooth: $\partial F(x^\star) \subseteq \nabla f_1(x^\star) - \partial f_2(x^\star)$ (ibid.)
  - ✗ f_1 nonsmooth & f_2 nonsmooth: *necessary* $\partial f_2(x^\star) \subseteq \partial f_1(x^\star)$ (Hiriart-Urruty 1989)
- In our analysis:
  1. both f_1 and f_2 smooth
  2. case with only one of f_1 and f_2 smooth obtained to the limit

  If both f_1 and f_2 nonsmooth – different performance measure with $\mathcal{O}(\frac{1}{N})$ rate (not in this talk)

---

[1] Smooth in the sense of functions with Lipschitz gradient: $f \in \mathcal{F}_{\mu,L}$ with $L < \infty$.

[2] $L_1 = \infty$

Convergence measures

- best **residual gradient norm** vs. initial objective accuracy

$$\boxed{\frac{1}{2}\min_{0\le k\le N}\left\{\|g_1^k - g_2^k\|^2\right\} \le \frac{F(x_0) - F_{\text{lo}}}{\tau_N^{\nabla}(L_1, \mu_1, L_2, \mu_2)}}$$

  - $g_1^k \in \partial f_1(x^k)$, $g_2^k \in \partial f_2(x^k)$
  - **PGD** terms: $g_1^k - g_2^k = \nabla\varphi(x^k) + g_h^k$, where $g_h^k \in \partial h(x^k)$

Convergence measures

- best **residual gradient norm** vs. initial objective accuracy

$$\boxed{\frac{1}{2}\min_{0\le k\le N}\left\{\|g_1^k - g_2^k\|^2\right\} \le \frac{F(x_0) - F_{\mathrm{lo}}}{\tau_N^{\nabla}(L_1,\mu_1,L_2,\mu_2)}}$$

  - $g_1^k \in \partial f_1(x^k)$, $g_2^k \in \partial f_2(x^k)$
  - **PGD** terms: $g_1^k - g_2^k = \nabla\varphi(x^k) + g_h^k$, where $g_h^k \in \partial h(x^k)$

- best **gradient mapping norm** vs. initial objective accuracy

$$\boxed{\frac{1}{2}\min_{0\le k\le N}\left\{\|x^k - x^{k+1}\|^2\right\} \le \frac{F(x_0) - F_{\mathrm{lo}}}{\tau_N^{\Delta}(L_1,\mu_1,L_2,\mu_2)}}$$

  - (prox-)gradient mapping norm, where $g_h^{k+1} \in \partial h(x^{k+1})$:

$$\|\gamma^{-1}\big(x^k - \underbrace{\mathrm{prox}_{\gamma h}[x^k - \gamma\nabla\varphi(x)]}_{x^{k+1}}\big)\| = \|\nabla\varphi(x^k) + g_h^{k+1}\|$$

  - **PGD** terms: $\gamma^{-1}\left(x^k - x^{k+1}\right) = \nabla\varphi(x^k) + g_h^{k+1}$
  - **DCA** terms: measure $\|\mu_1(x^k - x^{k+1})\|$ $(= \|(\gamma^{-1} + \mu_h)(x^k - x^{k+1})\|)$

Contents

Literature review of convergence rates

Measure	Curvatures	References	Tight rates
τ_N^{Δ}	$\mu_1, \mu_2 \geq 0,\ \mu_1 + \mu_2 > 0$	Dinh and Thi 1997	✗
		Hadi Abbaszadehpeivasti et al. 2023	✓
	$\mu_1 \geq 0, L_1 > 0; \mu_2, L_2 \in \mathbb{R}$ $\mu_1 + \mu_2 > 0$	**this talk** Rotaru, Patrinos, et al. 2025	✓
τ_N^{∇}	$\mu_1, \mu_2 \geq 0$	Hadi Abbaszadehpeivasti et al. 2023	✓
	$\mu_1 \geq 0, L_1 > 0; \mu_2, L_2 \in \mathbb{R}$ $\mu_1 + \mu_2 > 0$ or $\mu_1 = \mu_2 = 0$	**this talk** Rotaru, Patrinos, et al. 2025	✓

- best **gradient mapping norm** $\tau_N^{\Delta}(L_1, \mu_1, L_2, \mu_2) = \dfrac{F(x_0) - F_{\text{lo}}}{\min\limits_{0 \leq k \leq N} \{0.5\|x^k - x^{k+1}\|^2\}}$
- best **residual gradient norm** $\tau_N^{\nabla}(L_1, \mu_1, L_2, \mu_2) = \dfrac{F(x_0) - F_{\text{lo}}}{\min\limits_{0 \leq k \leq N} \{0.5\|g_1^k - g_2^k\|^2\}}$

Sufficient condition for convergence

- Descent result (Dinh and Thi 1997, Proposition 2)

$$F(x^k) - F_{\text{lo}} \geq F(x^k) - F(x^{k+1}) \geq (\mu_1 + \mu_2)\tfrac{1}{2}\|x^k - x^{k+1}\|^2 \qquad (1)$$

- Decrease after each iteration if $\mu_1 + \mu_2 \geq 0$
- Strict decrease if $\mu_1 + \mu_2 > 0$ (unless $x^{k+1} = x^k$)

Sufficient condition for convergence

- Descent result (Dinh and Thi 1997, Proposition 2)

$$F(x^k) - F_{\text{lo}} \geq F(x^k) - F(x^{k+1}) \geq (\mu_1 + \mu_2)\tfrac{1}{2}\|x^k - x^{k+1}\|^2 \tag{1}$$

- Decrease after each iteration if $\mu_1 + \mu_2 \geq 0$
- Strict decrease if $\mu_1 + \mu_2 > 0$ (unless $x^{k+1} = x^k$)
- **PGD** terms: ($\gamma = \frac{1}{L_\varphi}$ means $\mu_2 = 0$)

$$F(x^k) - F_{\text{lo}} \geq F(x^k) - F(x^{k+1}) \geq \frac{\mu_1 + \mu_2}{\mu_1^2}\tfrac{1}{2}\|\mu_1(x^k - x^{k+1})\|^2$$

Sufficient condition for convergence

- Descent result (Dinh and Thi 1997, Proposition 2)

$$F(x^k) - F_{\text{lo}} \geq F(x^k) - F(x^{k+1}) \geq (\mu_1 + \mu_2)\tfrac{1}{2}\|x^k - x^{k+1}\|^2 \tag{1}$$

- Decrease after each iteration if $\mu_1 + \mu_2 \geq 0$
- Strict decrease if $\mu_1 + \mu_2 > 0$ (unless $x^{k+1} = x^k$)
- **PGD** terms: ($\gamma = \frac{1}{L_\varphi}$ means $\mu_2 = 0$)

$$\begin{aligned} F(x^k) - F_{\text{lo}} \geq F(x^k) - F(x^{k+1}) &\geq \frac{\mu_1 + \mu_2}{\mu_1^2}\tfrac{1}{2}\|\mu_1(x^k - x^{k+1})\|^2 \\ &\equiv \frac{2\gamma^{-1} - L_\varphi + \mu_h}{(\gamma^{-1} - \mu_h)^2}\tfrac{1}{2}\|(\gamma^{-1} - \mu_h)(x^k - x^{k+1})\|^2 \end{aligned}$$

Sufficient condition for convergence

- Descent result (Dinh and Thi 1997, Proposition 2)

$$F(x^k) - F_{\text{lo}} \geq F(x^k) - F(x^{k+1}) \geq (\mu_1 + \mu_2)\tfrac{1}{2}\|x^k - x^{k+1}\|^2 \tag{1}$$

- Decrease after each iteration if $\mu_1 + \mu_2 \geq 0$
- Strict decrease if $\mu_1 + \mu_2 > 0$ (unless $x^{k+1} = x^k$)
- **PGD** terms: ($\gamma = \frac{1}{L_\varphi}$ means $\mu_2 = 0$)

$$\begin{aligned} F(x^k) - F_{\text{lo}} \geq F(x^k) - F(x^{k+1}) &\geq \frac{\mu_1 + \mu_2}{\mu_1^2}\tfrac{1}{2}\|\mu_1(x^k - x^{k+1})\|^2 \\ &\equiv \frac{2\gamma^{-1} - L_\varphi + \mu_h}{(\gamma^{-1} - \mu_h)^2}\tfrac{1}{2}\|(\gamma^{-1} - \mu_h)(x^k - x^{k+1})\|^2 \\ &\overset{\mu_h=0}{=} \frac{2\gamma^{-1} - L_\varphi}{\gamma^{-2}}\tfrac{1}{2}\|\nabla\varphi(x^k) + \partial h(x^{k+1})\|^2 \end{aligned}$$

Sufficient condition for convergence

- Descent result (Dinh and Thi 1997, Proposition 2)

$$F(x^k) - F_{\text{lo}} \geq F(x^k) - F(x^{k+1}) \geq (\mu_1 + \mu_2)\tfrac{1}{2}\|x^k - x^{k+1}\|^2 \tag{1}$$

- Decrease after each iteration if $\mu_1 + \mu_2 \geq 0$
- Strict decrease if $\mu_1 + \mu_2 > 0$ (unless $x^{k+1} = x^k$)
- **PGD** terms: ($\gamma = \frac{1}{L_\varphi}$ means $\mu_2 = 0$)

$$\begin{aligned}
F(x^k) - F_{\text{lo}} \geq F(x^k) - F(x^{k+1}) &\geq \frac{\mu_1 + \mu_2}{\mu_1^2}\tfrac{1}{2}\|\mu_1(x^k - x^{k+1})\|^2 \\
&\equiv \frac{2\gamma^{-1} - L_\varphi + \mu_h}{(\gamma^{-1} - \mu_h)^2}\tfrac{1}{2}\|(\gamma^{-1} - \mu_h)(x^k - x^{k+1})\|^2 \\
&\overset{\mu_h = 0}{=} \frac{2\gamma^{-1} - L_\varphi}{\gamma^{-2}}\tfrac{1}{2}\|\nabla\varphi(x^k) + \partial h(x^{k+1})\|^2 \\
&\overset{\gamma^{-1} = L_\varphi}{=} \frac{1}{L_\varphi}\tfrac{1}{2}\|\nabla\varphi(x^k) + \partial h(x^{k+1})\|^2
\end{aligned}$$

Contents

Rates for gradient residual norm τ_N^{∇}

Theorem (One-step decrease)

Assume $\mu_1 \geq 0$ *and* $\mu_2 \in \mathbb{R}$ *such that* $\mu_1 + \mu_2 > 0$ *or* $\mu_1 = \mu_2 = 0$. *Then:*

$$F(x^k) - F(x^{k+1}) \geq \sigma_i \tfrac{1}{2}\|g_1^k - g_2^k\|^2 + \sigma_i^+ \tfrac{1}{2}\|g_1^{k+1} - g_2^{k+1}\|^2$$

- $p_i(L_1, \mu_1, L_2, \mu_2) = \sigma_i(L_1, \mu_1, L_2, \mu_2) + \sigma_i^+(L_1, \mu_1, L_2, \mu_2)$
- *6 distinct decrease lemmas*

Theorem (DCA sublinear rates)

Assume $\mu_1 \geq 0$ *and* $\mu_2 \in \mathbb{R}$ *such that* $\mu_1 + \mu_2 > 0$ *or* $\mu_1 = \mu_2 = 0$.
After N *iterations starting from* x^0*:*

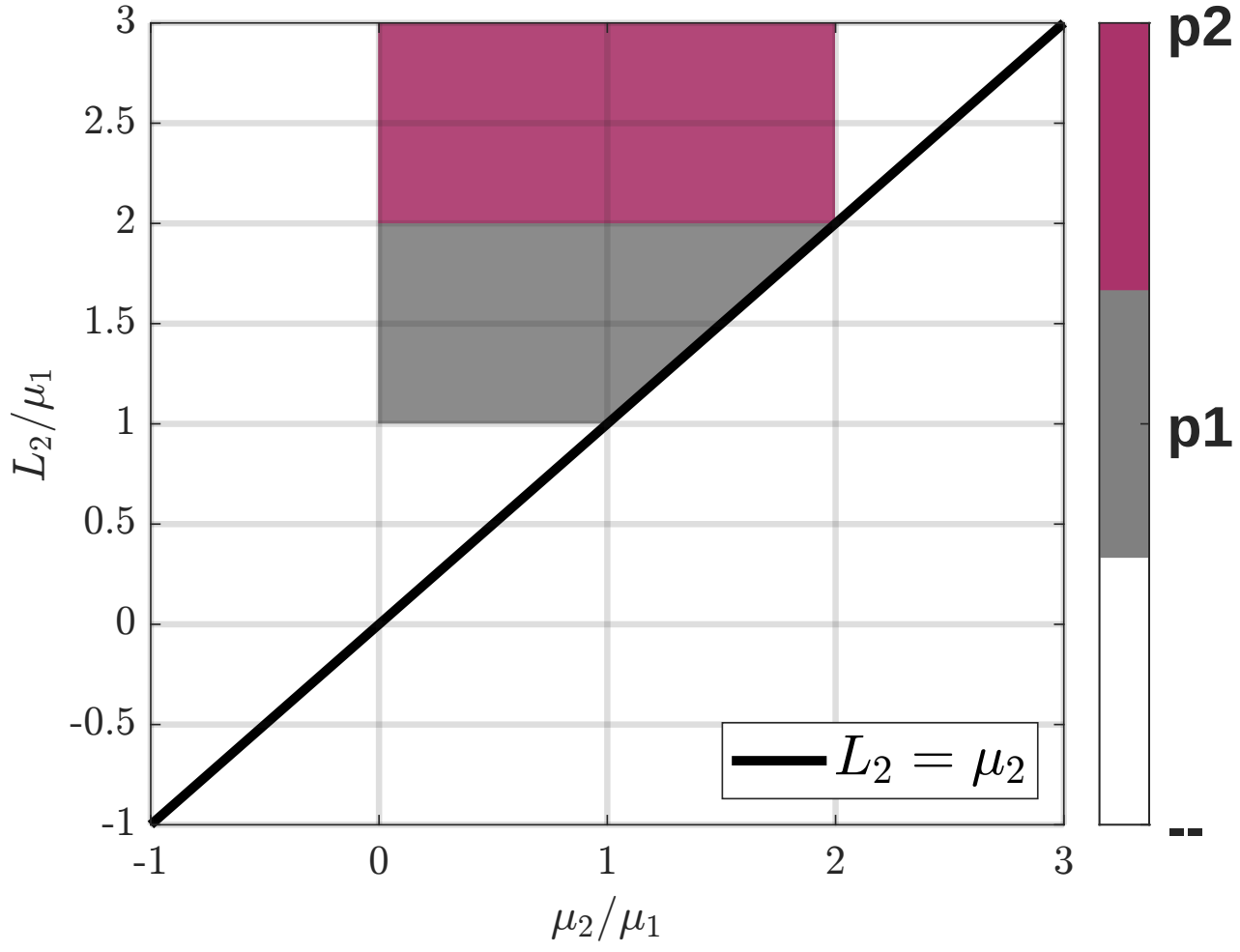
$$\tfrac{1}{2} \min_{0 \leq k \leq N} \{\|g_1^k - g_2^k\|^2\} \leq \frac{F(x^0) - F(x^N)}{p_i(L_1, \mu_1, L_2, \mu_2)N}$$

If F *is nonconcave (i.e.,* $L_1 > \mu_2$*):*

$$\tfrac{1}{2} \min_{0 \leq k \leq N} \{\|g_1^k - g_2^k\|^2\} \leq \frac{F(x^0) - F_{lo}}{p_i(L_1, \mu_1, L_2, \mu_2)N + \frac{1}{L_1 - \mu_2}}$$

- τ_N^{∇} *includes 6 regimes* p_i, $i = \{1, \ldots, 6\}$

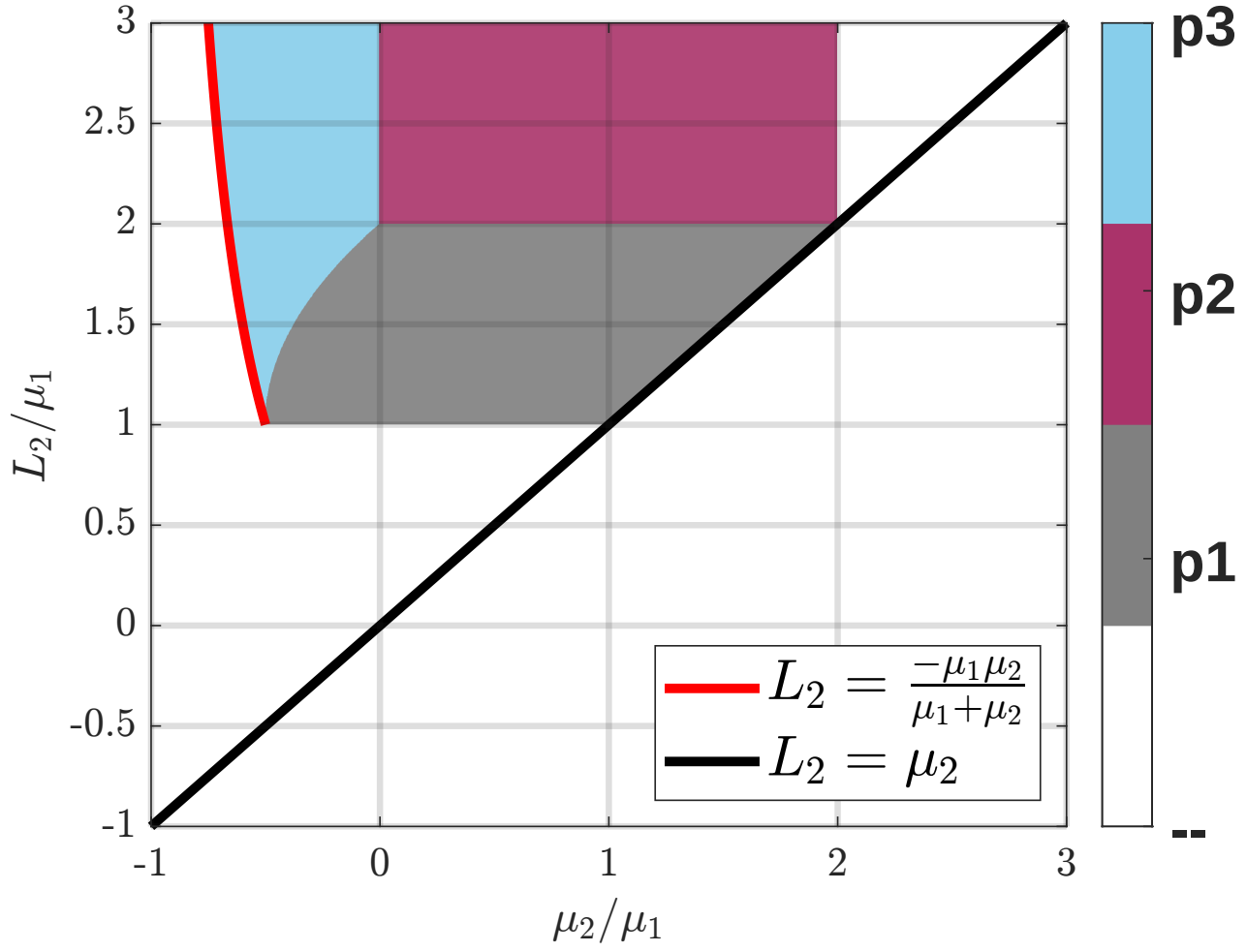
Domains of the 6 regimes (best gradient residual τ_N^{∇})



Standard DCA: f_1, f_2 convex ($\mu_1, \mu_2 \geq 0$)

$$\mu_1 > 0,\ \frac{L_1}{\mu_1} = 2,\ \frac{L_2}{\mu_1} \leq 3$$

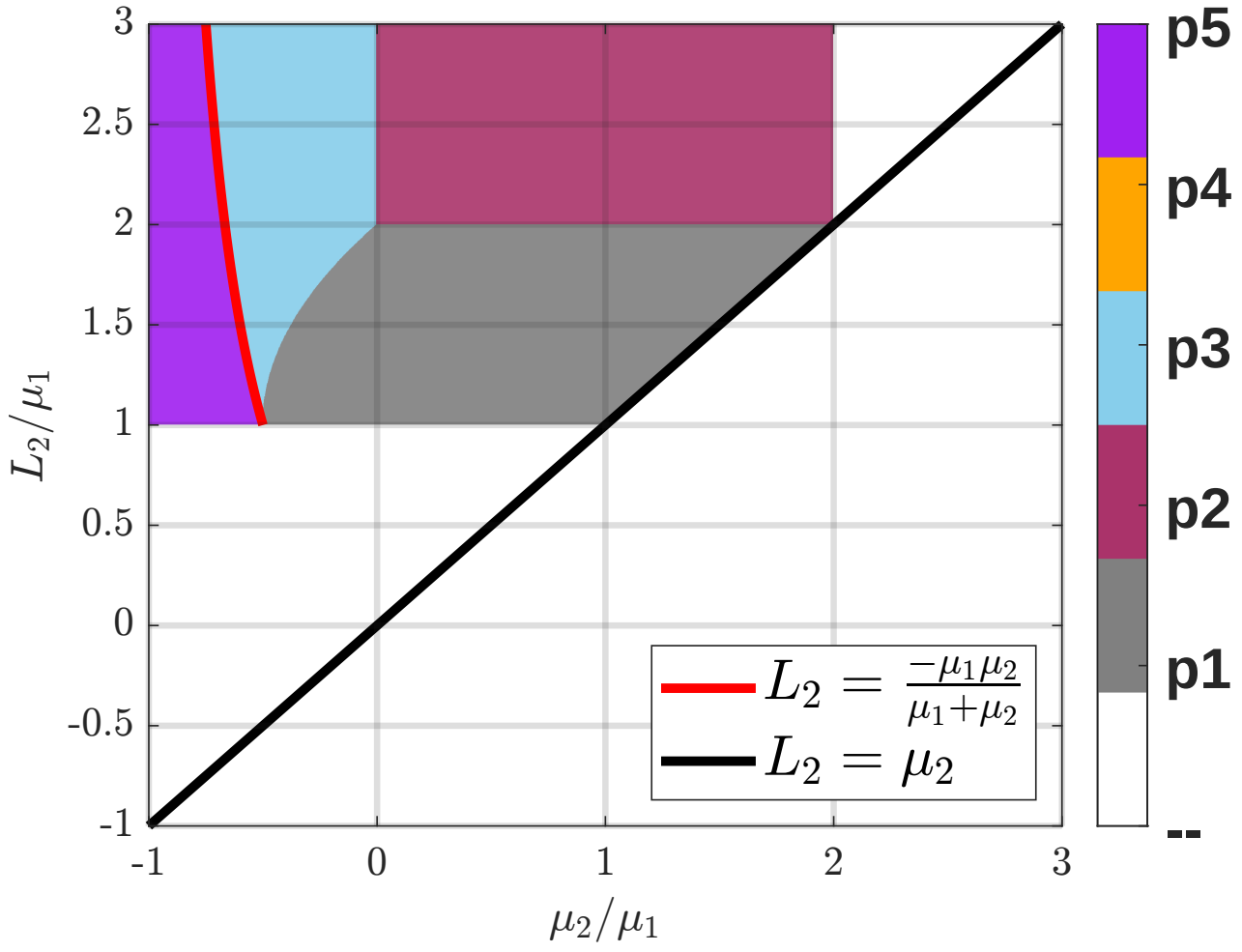
Domains of the 6 regimes (best gradient residual τ_N^{∇})



Tight regimes with f_2 weakly convex ($\mu_2 \leq 0$; $\mu_1 + \mu_2 > 0$)

$$\mu_1 > 0,\ \frac{L_1}{\mu_1} = 2,\ \frac{L_2}{\mu_1} \leq 3$$

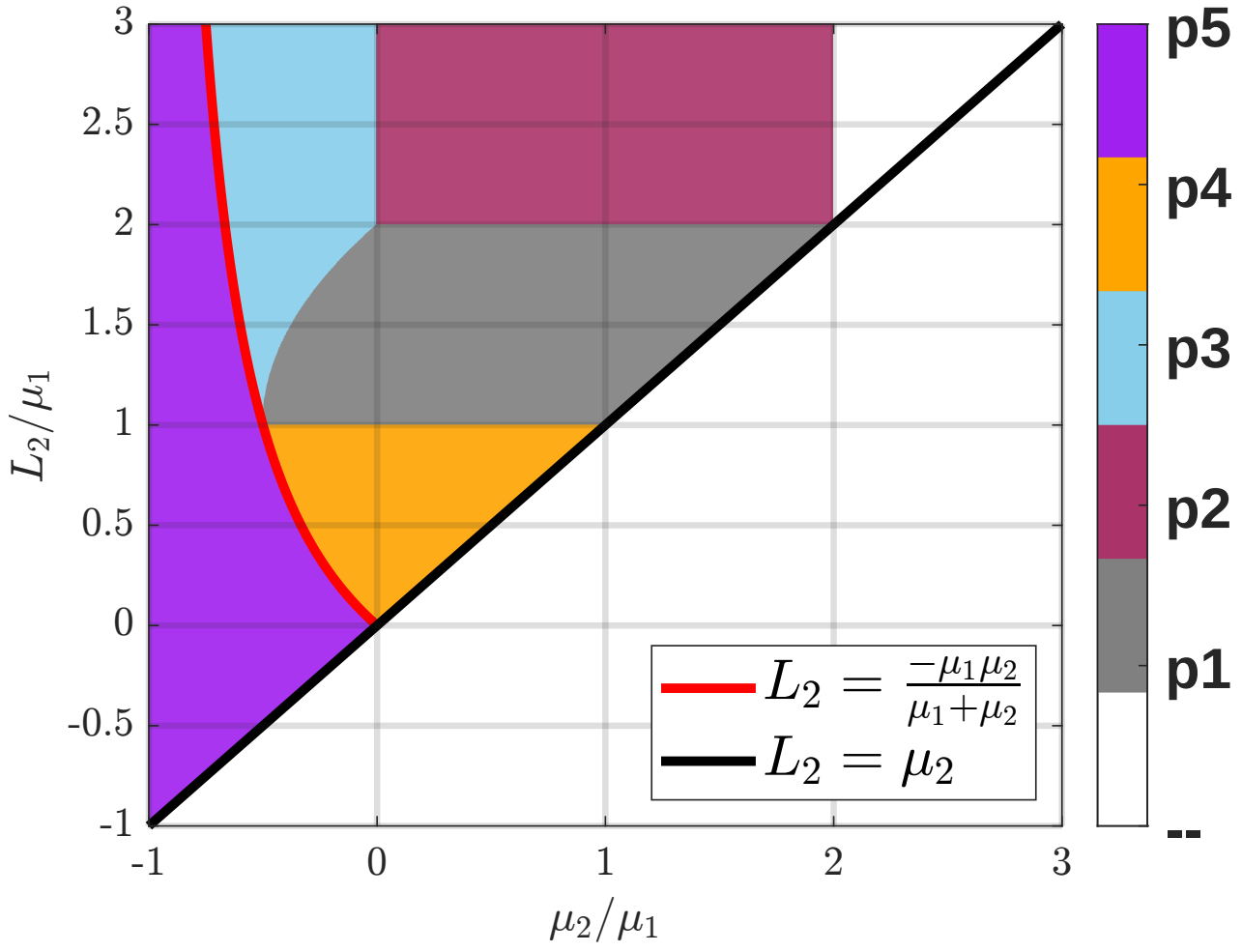
Domains of the 6 regimes (best gradient residual τ_N^{∇})



f_2 weakly convex ($\mu_2 \leq 0$; $\mu_1 + \mu_2 > 0$) & F nonconvex-nonconcave

$$\mu_1 > 0,\ \frac{L_1}{\mu_1} = 2,\ \frac{L_2}{\mu_1} \leq 3$$

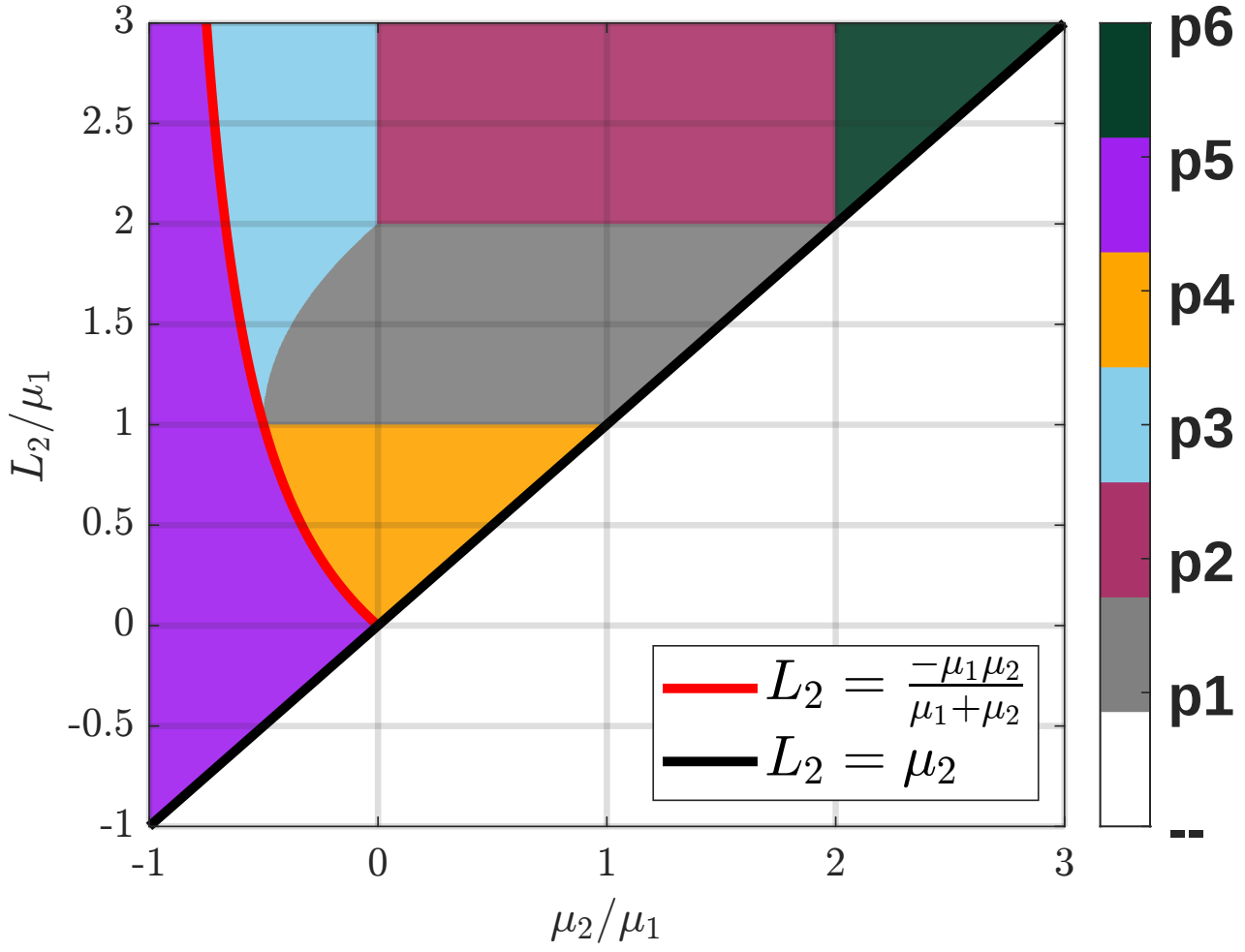
Domains of the 6 regimes (best gradient residual τ_N^{∇})



All regimes for F (strongly) convex

$$\mu_1 > 0,\ \frac{L_1}{\mu_1} = 2,\ \frac{L_2}{\mu_1} \leq 3$$

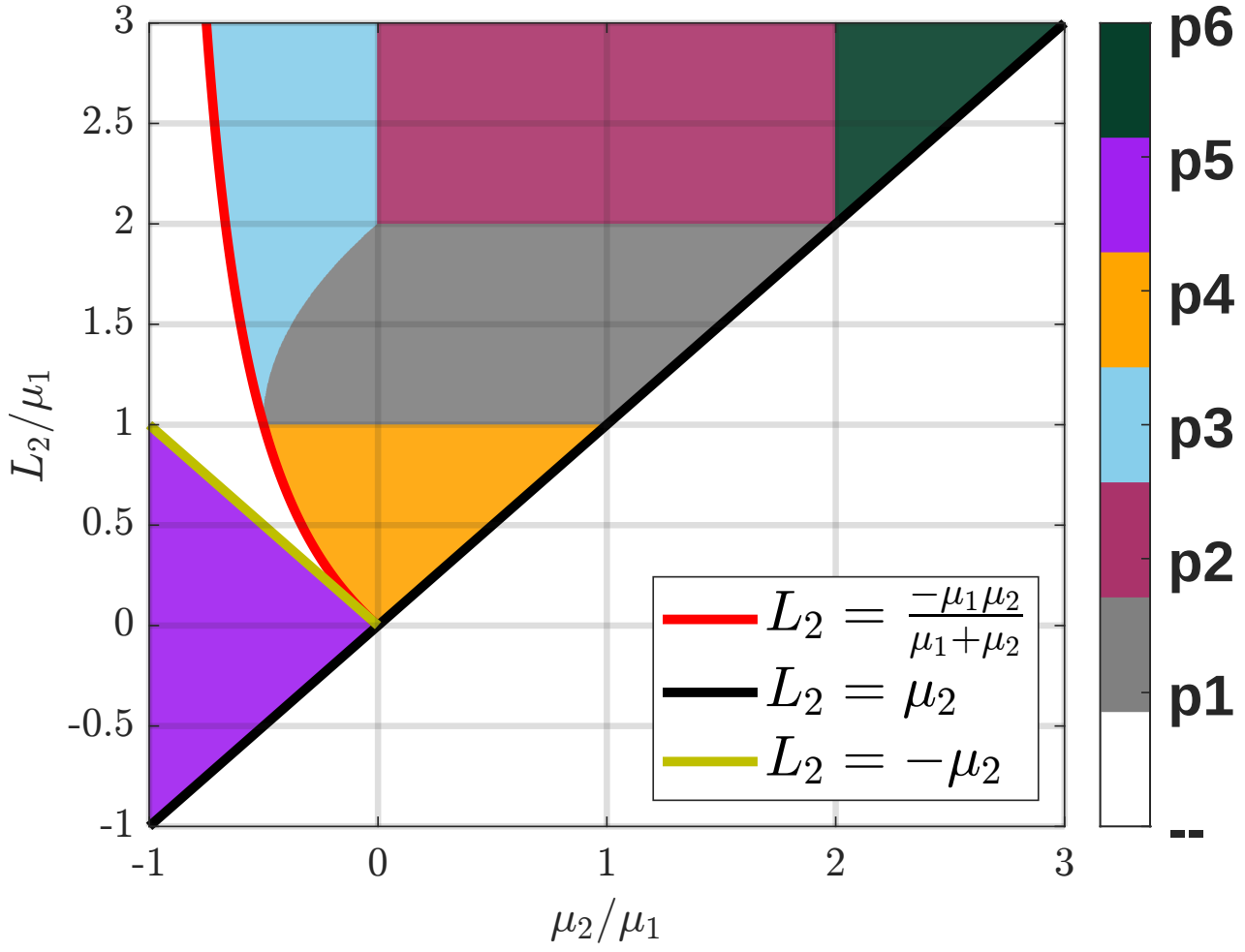
Domains of the 6 regimes (best gradient residual τ_N^{∇})



Domains for any F

$$\mu_1 > 0,\ \frac{L_1}{\mu_1} = 2,\ \frac{L_2}{\mu_1} \leq 3$$

Domains of the 6 regimes (best gradient residual τ_N^{∇})



All regimes with analytical proofs

$$\mu_1 > 0,\ \frac{L_1}{\mu_1} = 2,\ \frac{L_2}{\mu_1} \leq 3$$

Expressions of the regimes

- $F(x) - F(x^+) \geq \sigma_i \frac{1}{2}\|g_1 - g_2\|^2 + \sigma_i^+ \frac{1}{2}\|g_1^+ - g_2^+\|^2$

Regime	σ_i	σ_i^+	**Description**
p_1	$L_2^{-1}\dfrac{L_2-\mu_1}{L_1-\mu_1}$	$L_2^{-1}\left(1+\dfrac{L_2^{-1}-L_1^{-1}}{\mu_1^{-1}-L_1^{-1}}\right)$	f_1 convex f_2 convex or nonconvex F nonconvex-nonconcave
p_2	$L_1^{-1}\left(1+\dfrac{L_1^{-1}-L_2^{-1}}{\mu_2^{-1}-L_2^{-1}}\right)$	$L_1^{-1}\dfrac{L_1-\mu_2}{L_2-\mu_2}$	f_1, f_2 convex F nonconvex-nonconcave
p_3	$\dfrac{L_1^{-1}(\mu_1^{-1}+\mu_2^{-1}+L_2^{-1})}{\mu_1^{-1}+\mu_2^{-1}+L_2^{-1}-L_1^{-1}}$	$\dfrac{1}{L_2+\mu_2}$	f_1 strongly convex f_2 nonconvex F nonconvex-nonconcave
p_4	0	$\dfrac{L_2+\mu_1}{L_2^2}$	f_1 strongly convex f_2 convex or nonconvex F (strongly) convex
p_5	0	$\dfrac{\mu_1+\mu_2}{\mu_2^2}$	f_1 strongly convex f_2 nonconvex & F nonconvex-nonconcave or f_2 concave & F (strongly) convex
p_6	$\dfrac{L_1+\mu_2}{L_1^2}$	0	f_1 convex f_2 strongly convex F concave

- If $\mu_1 = \mu_2 = 0$: $F(x) - F(x^+) \geq \frac{1}{2L_1}\|g_1 - g_2\|^2 + \frac{1}{2L_2}\|g_1^+ - g_2^+\|^2$

Tightness of the regimes

- Sublinear rates for p_1, p_2, p_3 are tight for any N
- Sublinear rates for p_4, p_5, p_6 are non-tight for $N \geq 2$
- Apart from a subdomain of p_5 we can prove *linear* rates

Tightness of the regimes

- Sublinear rates for p_1, p_2, p_3 are tight for any N
- Sublinear rates for p_4, p_5, p_6 are non-tight for $N \geq 2$
- Apart from a subdomain of p_5 we can prove *linear* rates
- We have conjectured tight rates covering whole space (more difficult proofs)

Conjectured rates

- $E_k(z) := \sum_{j=1}^{2k} z^{-j} = \begin{cases} \frac{-1+z^{-2k}}{1-z}, & z \in \mathbb{R}_{>0,\neq 1} \\ 2k, & z = 1 \end{cases}$; $E_0(z) = 0$.

Conjecture (F nonconvex-nonconcave)

Let $\mu_2 < 0 < \mu_1 < L_2$, $\mu_1 + \mu_2 > 0$, $\mu_1^{-1} + \mu_2^{-1} + L_2^{-1} < 0$ *and* $N \geq 2$. *Then*

$$\frac{1}{2\mu_1} \min_{0 \leq k \leq N} \left\{ \|g_1^k - g_2^k\|^2 \right\} \leq \frac{F(x^0) - F(x^N)}{\min\left\{ P_N\left(\frac{L_2}{\mu_1}, \frac{\mu_2}{\mu_1}\right), E_N\left(\frac{\mu_2}{\mu_1}\right) \right\}},$$

where[3] $P_N(\eta, \rho) := \frac{(1+\eta)(1+\rho)}{\eta+\rho} \left[N + \frac{(1-\eta)(1-\rho)}{\eta-\rho} \sum_{k=0}^{N} \left[E_N(\eta) - E_N(\rho) \right]_+ \right]$

Conjecture (F (strongly) convex)

Let $\mu_1 > 0$, $\mu_2 \lesseqgtr 0$, $\mu_1 + \mu_2 > 0$, $L_2 \in (0, \mu_1]$, $N \geq 1$. *Then*

$$\frac{1}{2\mu_1} \min_{0 \leq k \leq N} \left\{ \|g_1^k - g_2^k\|^2 \right\} \leq \frac{F(x^0) - F(x^N)}{\min\left\{ E_N\left(\frac{L_2}{\mu_1}\right), E_N\left(\frac{\mu_2}{\mu_1}\right) \right\}}$$

[3] $[x]_+ := \max\{0, x\}$

Nonsmooth case

- Assume either $L_1 = \infty$ or $L_2 = \infty$
- $F(x) - F(x^+) \geq \sigma_i \frac{1}{2}\|g_1 - g_2\|^2 + \sigma_i^+ \frac{1}{2}\|g_1^+ - g_2^+\|^2$, $p_i = \sigma_i + \sigma_i^+$
- $p_1 = p_4$ and $p_2 = p_6$

Regime	σ_i	σ_i^+	**Domain**
$p_{1,4}$	0	$\frac{L_2 + \mu_1}{L_2^2}$	$\boldsymbol{L_1 = \infty} > L_2 \geq \mu_1 \geq 0$ $\mu_2\left(\mu_1^{-1} + \mu_2^{-1} + L_2^{-1}\right) \geq 0$
$p_{2,6}$	$\frac{L_1 + \mu_2}{L_1^2}$	0	$\boldsymbol{L_2 = \infty} > L_1 \geq \mu_2 \geq 0$ $\mu_1\left(\mu_1^{-1} + \mu_2^{-1} + L_1^{-1}\right) \geq 0$
p_3	$\frac{L_1^{-1}(\mu_1^{-1}+\mu_2^{-1})}{\mu_1^{-1}+\mu_2^{-1}-L_1^{-1}}$	0	$\boldsymbol{L_2 = \infty}$; $\mu_1 > -\mu_2 > 0$
p_5	0	$\frac{\mu_1 + \mu_2}{\mu_2^2}$	$\boldsymbol{L_1 = \infty}$; $\mu_1 > -\mu_2 > 0$ $0 < \mu_1^{-1} + \mu_2^{-1} + L_2^{-1}$

Contents

Rates for gradient mapping norm τ_N^Δ

Theorem (DCA sublinear rates)

Assume $\mu_1 \geq 0$ *and* $\mu_2 \in \mathbb{R}$ *such that* $\mu_1 + \mu_2 > 0$ *or* $\mu_1 = \mu_2 = 0$.
After $N+1 \geq 2$ *iterations starting from* x^0*:*

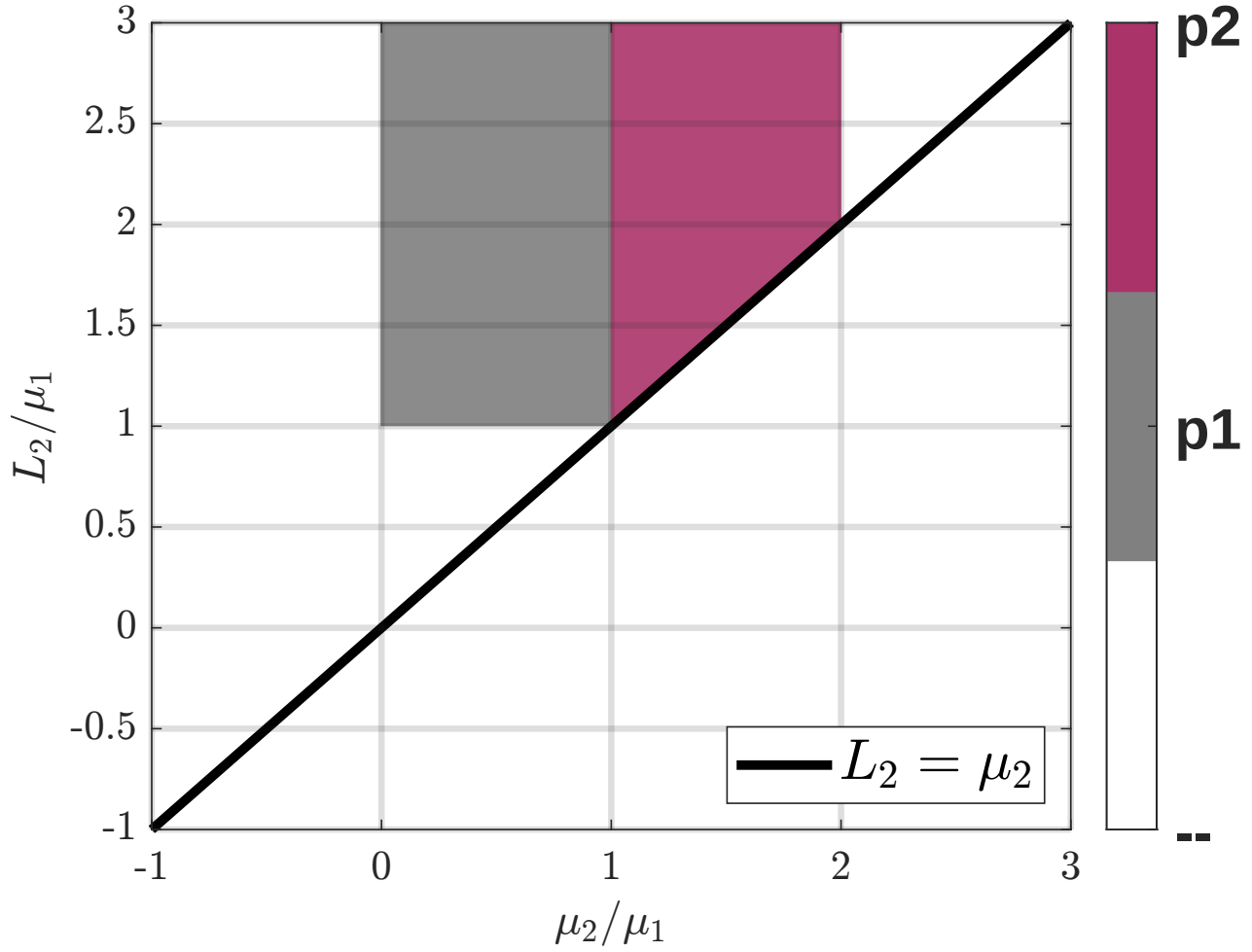
$$\frac{1}{2} \min_{0 \leq k \leq N} \{\|x^k - x^{k+1}\|^2\} \leq \frac{F(x^0) - F(x^{N+1})}{\mu_1 + \mu_2 + p_i(L_1, \mu_1, L_2, \mu_2)N}.$$

- τ_N^Δ *includes 6 regimes* p_i, $i = \{1, \ldots, 6\}$

Regime	**Domain**		**Description**
$p_1 = \mu_1 + \mu_2 + \frac{(\mu_1-\mu_2)^2}{L_2-\mu_2}$	$\mu_1 \geq \mu_2 \geq 0$	$L_2 > \mu_1$	f_1, f_2 convex F n.c.-n.ccv.
$p_2 = \mu_1 + \mu_2 + \frac{(\mu_1-\mu_2)^2}{L_1-\mu_1}$	$\mu_2 \geq \mu_1 \geq 0$	$L_1 > \mu_2$	f_1, f_2 convex F n.c.-n.ccv.
$p_3 = \frac{(\mu_1+\mu_2)(L_2+\mu_1)}{L_2+\mu_2}$	$\mu_2 \in [\frac{-L_2\mu_1}{L_2+\mu_1}, 0)$	$L_2 \geq \mu_1 > 0$	f_1 s.c., f_2 n.c. F n.c.-n.ccv.
$p_4 = \frac{\mu_1^2(L_2+\mu_1)}{L_2^2}$	$\mu_2 \in \left[\frac{-L_2\mu_1}{L_2+\mu_1}, L_2\right)$	$L_2 \in [0, \mu_1]$	f_1 s.c., f_2 n.c. F convex
$p_5 = \frac{\mu_1^2(\mu_1+\mu_2)}{\mu_2^2}$	$\mu_2 < 0$, $\mu_2 \in (-\mu_1, \frac{-L_2\mu_1}{L_2+\mu_1}]$		f_1 s.c., f_2 n.c. F n.ccv.
$p_6 = \frac{\mu_2^2(L_1+\mu_2)}{L_1^2}$	$\mu_2 \geq \mu_1 \geq 0$	$L_1 \in (0, \mu_2]$	f_1 convex, f_2 s.c. F ccv.

(s.c.: strongly convex; n.c.: nonconvex; ccv.: concave; n.ccv.: nonconcave)

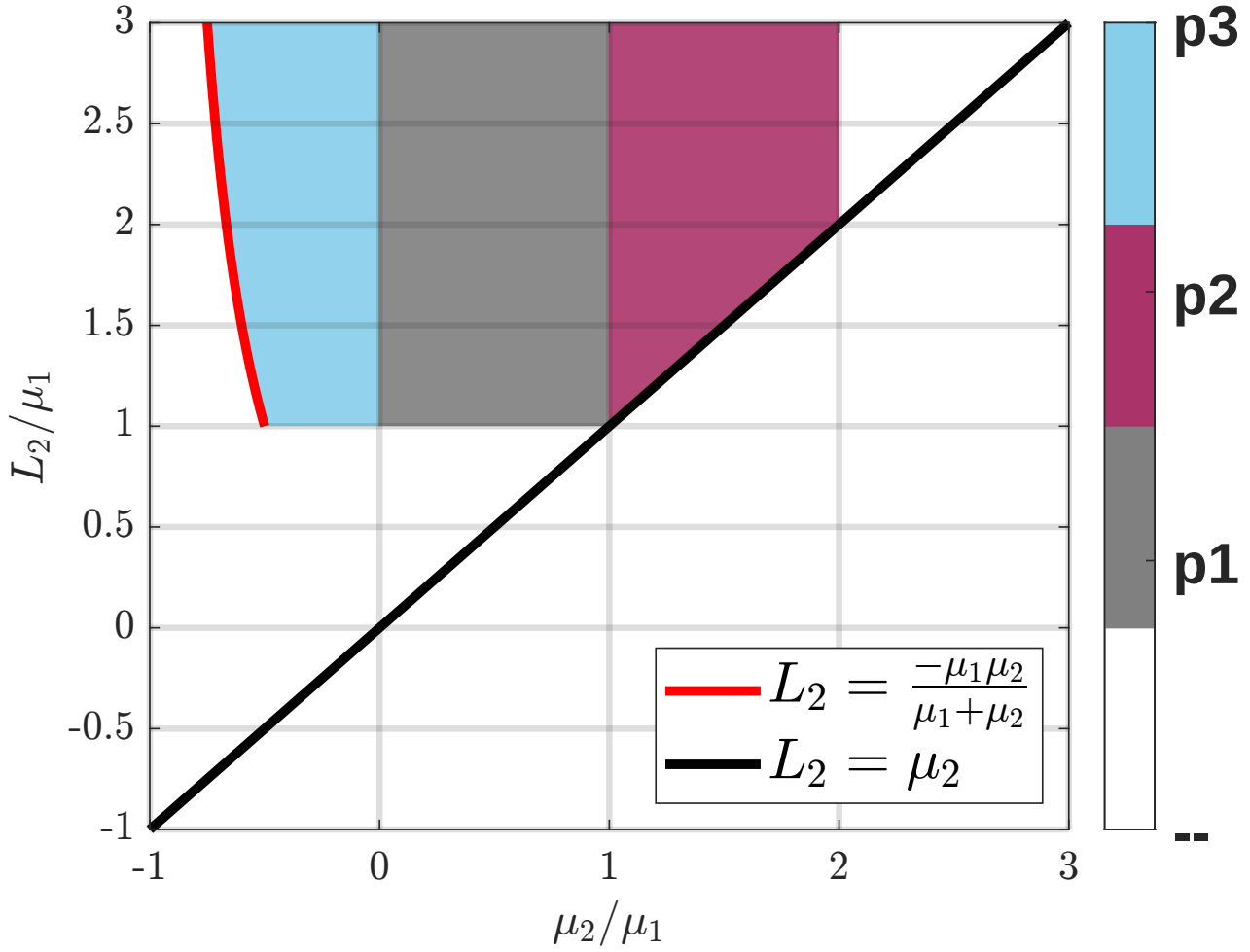
Domains of the 6 regimes (best gradient mapping τ_N^{Δ})



Standard DCA: f_1, f_2 convex $(\mu_1, \mu_2 \geq 0)$

$\mu_1 > 0$, $\frac{L_1}{\mu_1} = 2$, $\frac{L_2}{\mu_1} \leq 3$.

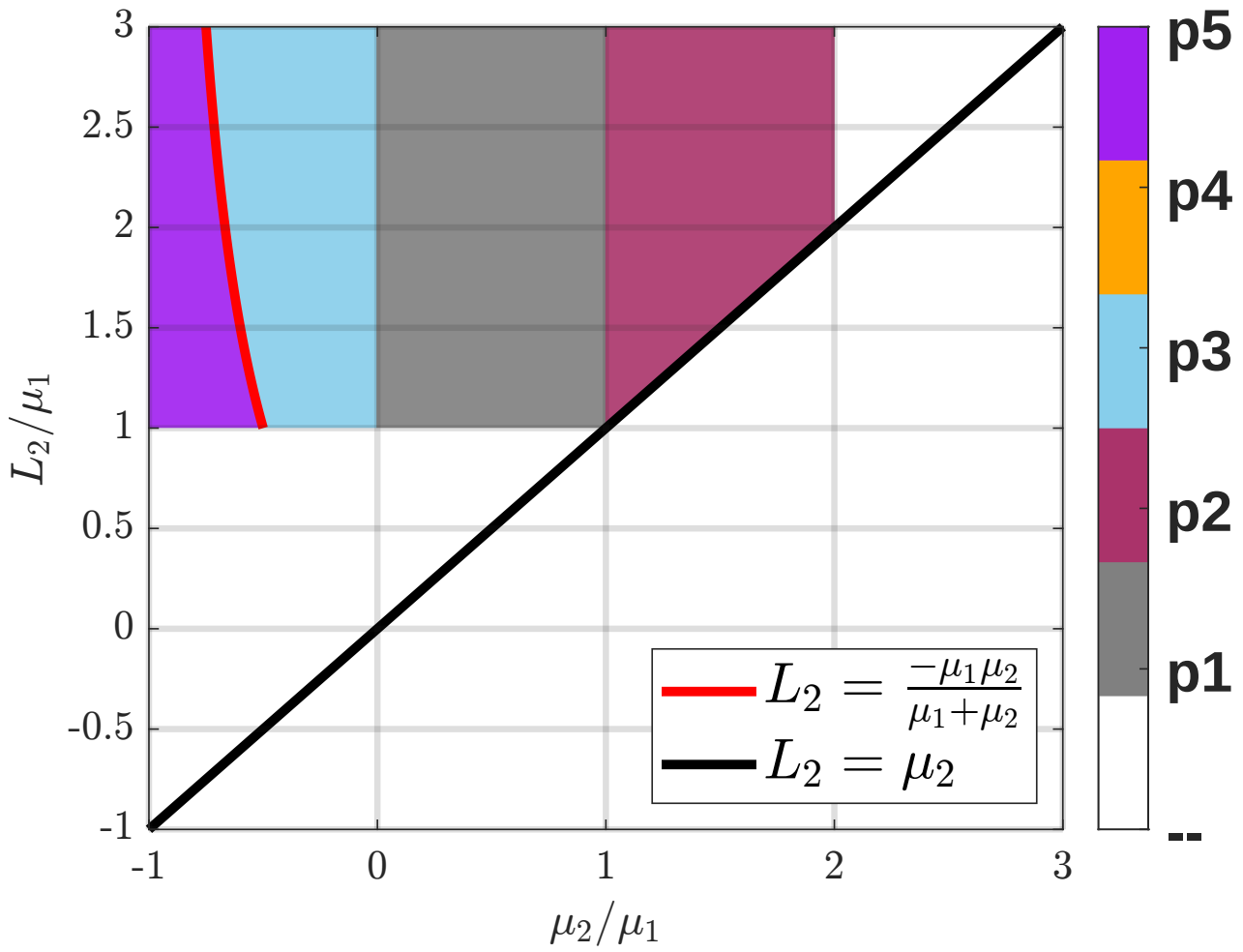
Domains of the 6 regimes (best gradient mapping τ_N^{Δ})



Tight regimes with f_2 weakly convex ($\mu_2 \leq 0$; $\mu_1 + \mu_2 > 0$)

$\mu_1 > 0$, $\frac{L_1}{\mu_1} = 2$, $\frac{L_2}{\mu_1} \leq 3$.

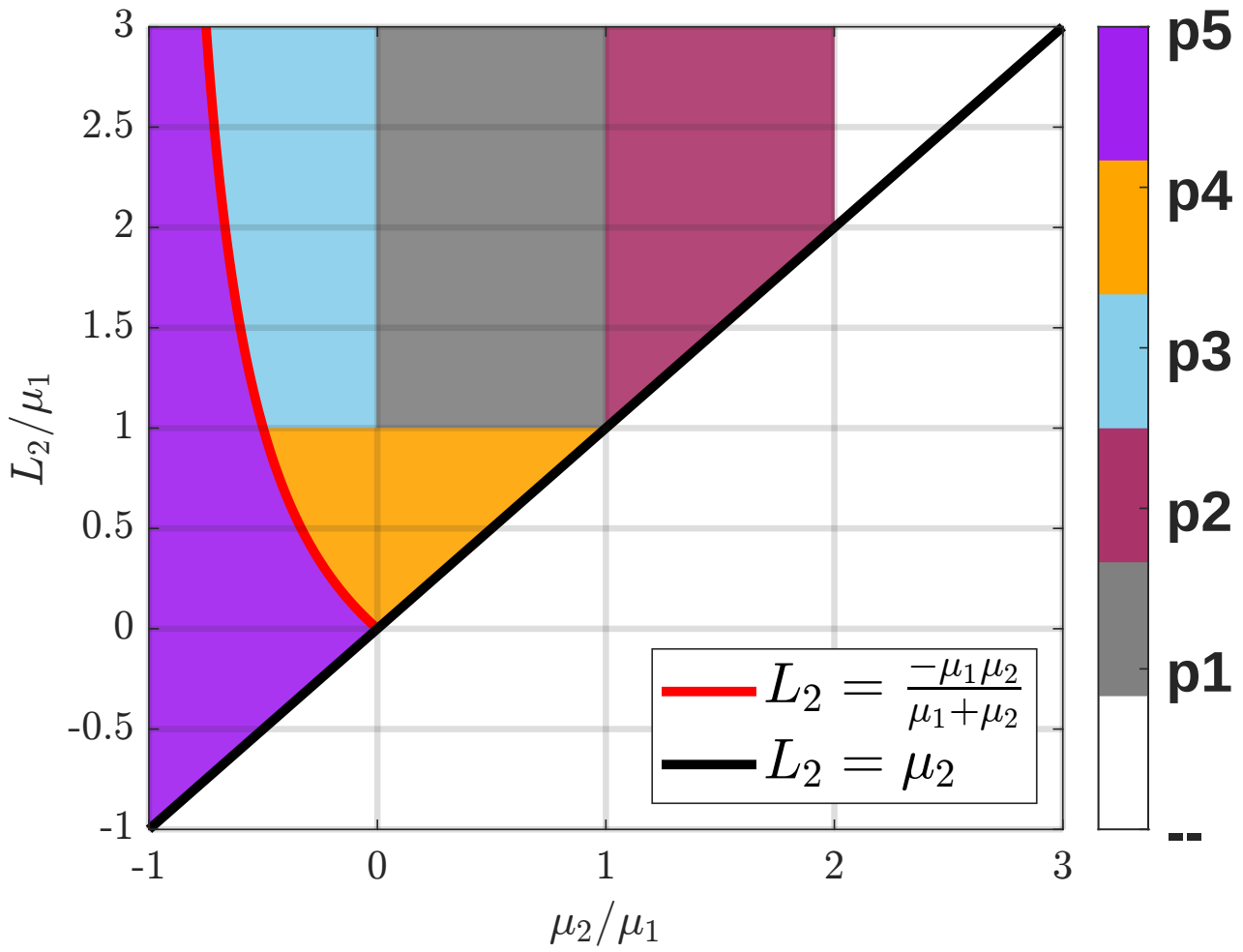
Domains of the 6 regimes (best gradient mapping τ_N^{Δ})



f_2 weakly convex ($\mu_2 \leq 0$; $\mu_1 + \mu_2 > 0$) & F nonconvex-nonconcave

$\mu_1 > 0$, $\frac{L_1}{\mu_1} = 2$, $\frac{L_2}{\mu_1} \leq 3$.

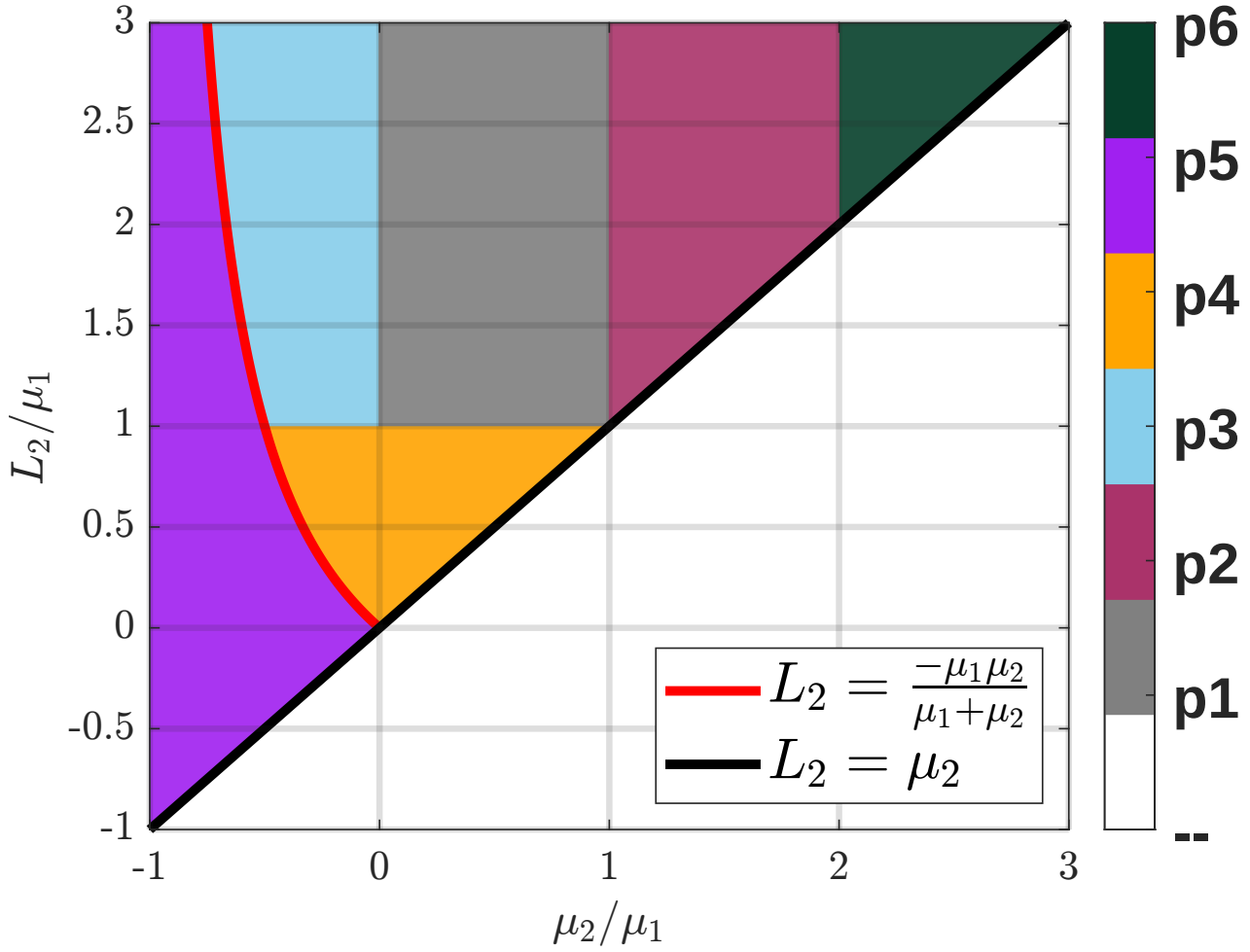
Domains of the 6 regimes (best gradient mapping τ_N^{Δ})



All regimes for F (strongly) convex

$\mu_1 > 0$, $\frac{L_1}{\mu_1} = 2$, $\frac{L_2}{\mu_1} \leq 3$.

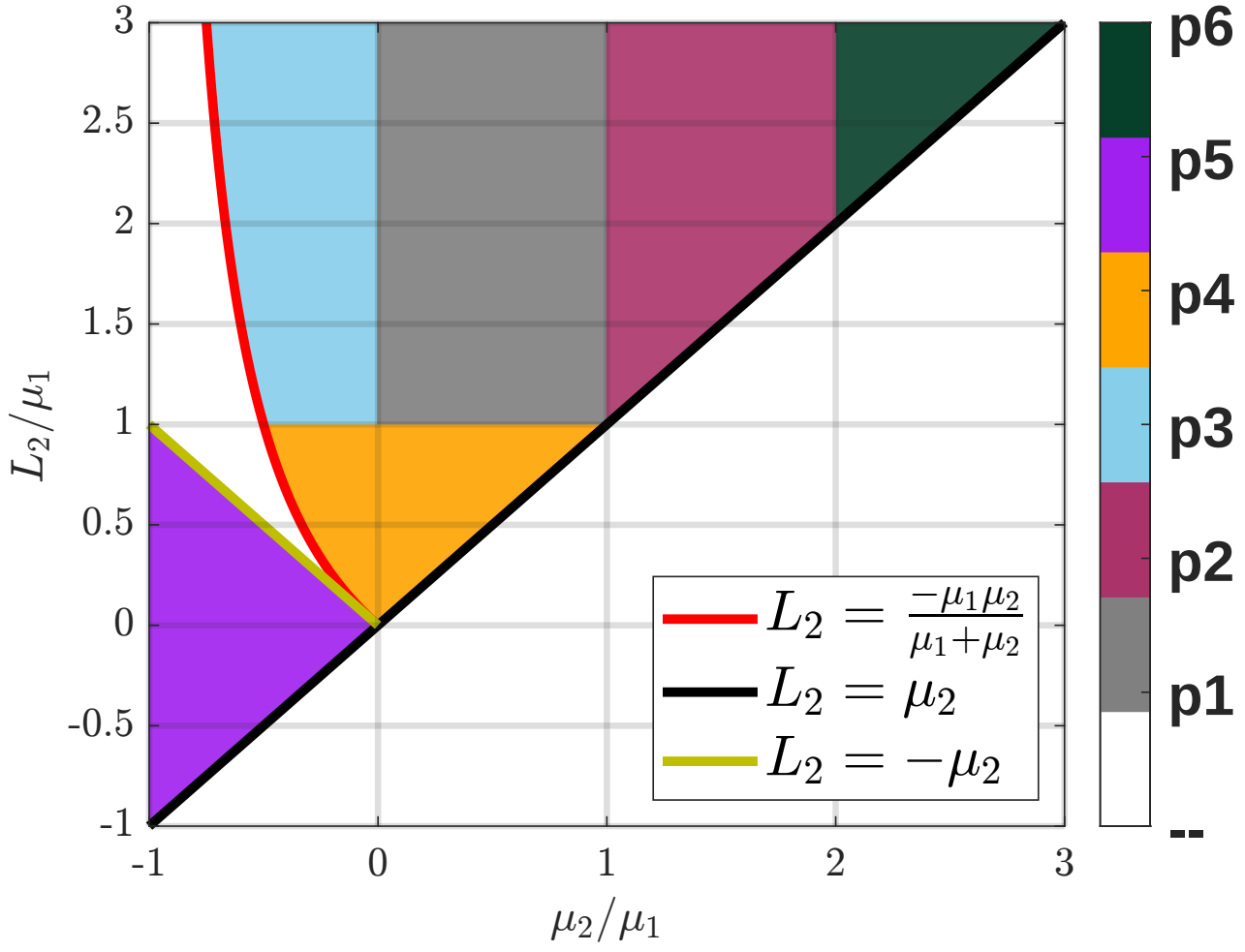
Domains of the 6 regimes (best gradient mapping τ_N^{Δ})



Domains for any F

$\mu_1 > 0$, $\frac{L_1}{\mu_1} = 2$, $\frac{L_2}{\mu_1} \leq 3$.

Domains of the 6 regimes (best gradient mapping τ_N^{Δ})



All regimes with analytical proofs

$\mu_1 > 0,\ \frac{L_1}{\mu_1} = 2,\ \frac{L_2}{\mu_1} \leq 3.$

Toland duality

- Standard DCA $\mu_1 \geq 0$, $\mu_2 \geq 0$ (regimes p_1, p_2): sublinear rates for **gradient mapping** are obtained from sublinear rates for **gradient residual**
- By Toland duality (Toland 1979):

$$\min f_1(x) - f_2(x) = \min f_2^*(x) - f_1^*(x)$$

- Result exploited by Hadi Abbaszadehpeivasti et al. 2023
- We have $x^{k+1} \in \partial f_1^*(g_2^k)$ and $x^k \in \partial f_2^*(g_2^k)$
- $f_1^* \in \mathcal{F}_{L_1^{-1},\mu_1^{-1}}$ and $f_2^* \in \mathcal{F}_{L_2^{-1},\mu_2^{-1}}$
- Rates on **gradient mapping** obtained by making the substitutions: $\mu_1 \leftrightarrow L_2^{-1}$, $L_1 \leftrightarrow \mu_2^{-1}$, $\mu_2 \leftrightarrow L_1^{-1}$, $L_2 \leftrightarrow \mu_1^{-1}$

Nonsmooth case (gradient mapping)

- f_1 nonsmooth: $L_1 = \infty$ (criticality $\Rightarrow$ stationarity!)

Regime	**Domain**		**Description**
$p_1 = \mu_1 + \mu_2 + \frac{(\mu_1-\mu_2)^2}{L_2-\mu_2}$	$\mu_1 \geq \mu_2 \geq 0$	$L_2 > \mu_1$	f_1, f_2 convex F n.c.-n.ccv.
$p_2 = \mu_1 + \mu_2 + \frac{(\mu_1-\mu_2)^2}{L_1-\mu_1}$	$\mu_2 \geq \mu_1 \geq 0$	$\mu_2 < \infty$	f_1, f_2 convex F n.c.-n.ccv.
$p_3 = \frac{(\mu_1+\mu_2)(L_2+\mu_1)}{L_2+\mu_2}$	$\mu_2 \in [\frac{-L_2\mu_1}{L_2+\mu_1}, 0)$	$L_2 \geq \mu_1 > 0$	f_1 s.c., f_2 n.c. F n.c.-n.ccv.
$p_4 = \frac{\mu_1^2(L_2+\mu_1)}{L_2^2}$	$\mu_2 \in \left[\frac{-L_2\mu_1}{L_2+\mu_1}, L_2\right)$	$L_2 \in [0, \mu_1]$	f_1 s.c., f_2 n.c. F convex
$p_5 = \frac{\mu_1^2(\mu_1+\mu_2)}{\mu_2^2}$	$\mu_2 < 0$, $\mu_2 \in \left(-\mu_1, \frac{-L_2\mu_1}{L_2+\mu_1}\right]$		f_1 s.c., f_2 n.c. F n.ccv.

(s.c.: strongly convex; n.c.: nonconvex; ccv.: concave; n.ccv.: nonconcave)

Nonsmooth case (gradient mapping)

- f_2 nonsmooth: $L_2 = \infty$ (criticality ~~$\Rightarrow$~~ stationarity!)

Regime	**Domain**		**Description**
$p_1 = \mu_1 + \mu_2 + \cancel{\frac{(\mu_1-\mu_2)^2}{L_2-\mu_2}}$	$\mu_1 \geq \mu_2 \geq 0$	$\mu_1 < \infty$	f_1, f_2 convex F n.c.-n.ccv.
$p_2 = \mu_1 + \mu_2 + \frac{(\mu_1-\mu_2)^2}{L_1-\mu_1}$	$\mu_2 \geq \mu_1 \geq 0$	$L_1 > \mu_2$	f_1, f_2 convex F n.c.-n.ccv.
$p_6 = \frac{\mu_2^2(L_1+\mu_2)}{L_1^2}$	$\mu_2 \geq \mu_1 \geq 0$	$L_1 \in (0, \mu_2]$	f_1 convex, f_2 s.c. F ccv.

(s.c.: strongly convex; n.c.: nonconvex; ccv.: concave; n.ccv.: nonconcave)

Conjectured rates

- $E_k(z) := \sum_{j=1}^{2k} z^{-j} = \begin{cases} \frac{-1+z^{-2k}}{1-z}, & z \in \mathbb{R}_{>0,\neq 1} \\ 2k, & z = 1 \end{cases}$; $E_0(z) = 0$.

Conjecture (F nonconvex-nonconcave)

Let $\mu_2 < 0 < \mu_1 < L_2$, *and* $N+1 \geq 2$. *Then*

$$\frac{1}{2\mu_1} \min_{0 \leq k \leq N} \left\{ \|\mu_1(x^k - x^{k+1})\|^2 \right\} \leq \frac{F(x^0) - F(x^{N+1})}{1 + \frac{\mu_2}{\mu_1} + \min\left\{ P_N\left(\frac{L_2}{\mu_1}, \frac{\mu_2}{\mu_1}\right), E_N\left(\frac{\mu_2}{\mu_1}\right) \right\}},$$

where[4] $P_N(\eta, \rho) := \frac{(1+\eta)(1+\rho)}{\eta+\rho} \left[N + \frac{(1-\eta)(1-\rho)}{\eta-\rho} \sum_{k=0}^{N} [E_N(\eta) - E_N(\rho)]_+ \right]$

Conjecture (F (strongly) convex)

Let $\mu_1 > 0$, $\mu_2 \lesseqgtr 0$, $\mu_1 + \mu_2 > 0$, $L_2 \in (0, \mu_1]$, $N+1 \geq 2$. *Then*

$$\frac{1}{2\mu_1} \min_{0 \leq k \leq N} \left\{ \|\mu_1(x^k - x^{k+1})\|^2 \right\} \leq \frac{F(x^0) - F(x^{N+1})}{1 + \frac{\mu_2}{\mu_1} + \min\left\{ E_N\left(\frac{L_2}{\mu_1}\right), E_N\left(\frac{\mu_2}{\mu_1}\right) \right\}}$$

[4] $[x]_+ := \max\{0, x\}$

Contents

Curvature shifting techniques

- Invariance of objective F to curvature shifting (where $\lambda \leq \mu_1$)

$$F = \underbrace{\left(f_1 - \lambda \frac{\|\cdot\|^2}{2}\right)}_{f_1^\lambda} - \underbrace{\left(f_2 - \lambda \frac{\|\cdot\|^2}{2}\right)}_{f_2^\lambda}$$

- Adjusted curvatures:
  - $\mu_1^\lambda = \mu_1 - \lambda$
  - $L_1^\lambda = L_1 - \lambda$
  - $\mu_2^\lambda = \mu_2 - \lambda$
  - $L_2^\lambda = L_2 - \lambda$
- Define $p^\lambda := p(L_1^\lambda, \mu_1^\lambda, L_2^\lambda, \mu_2^\lambda)$ be the corresponding denominator of the 6 regimes
- We compute $\lambda^* = \underset{\lambda \in (-\infty, \mu_1]}{\operatorname{argmax}} \; p^\lambda$
- Assume ∂f_1^* can be computed efficiently
- We focus on the measure of the **best gradient residual**

Example 1: Smooth DCA vs. Gradient Descent (GD)

- Assume $F \in \mathcal{F}_{\mu,L}$ is smooth
- GD iteration: $x^+ = x - \gamma \nabla F(x)$, where $\gamma \in (0, \frac{2}{L_F})$
- Compare the denominators after one iteration:
  $\frac{1}{2} \min\{\|\nabla F(x)\|^2, \|\nabla F(x^+)\|^2\} \leq \frac{F(x)-F(x^+)}{p_{\mathrm{DCA/GD}}}$

Example 1.1: F smooth-nonconvex

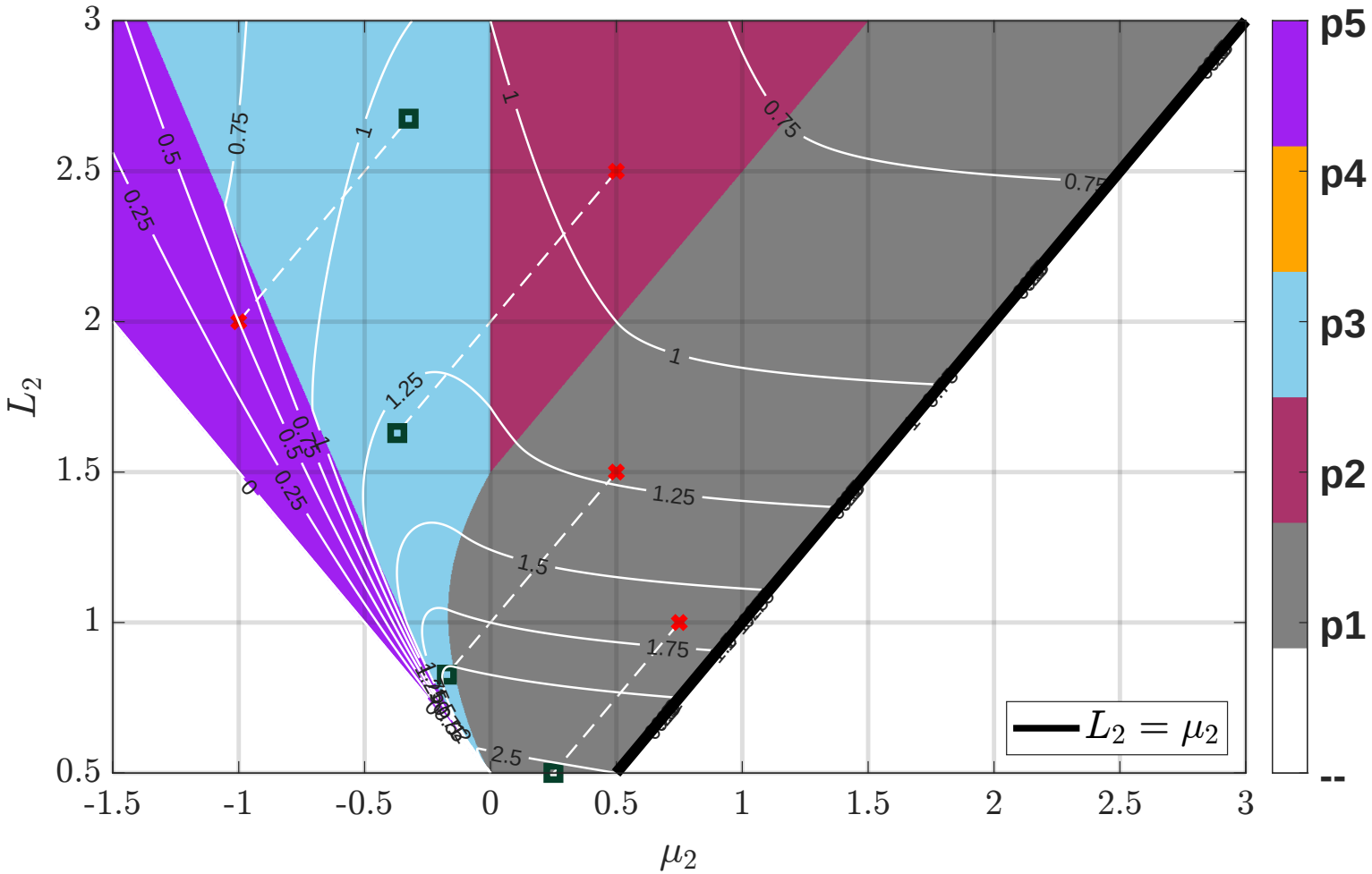
- Assume initial splitting $F = f_1 - f_2$ with $\mu_1 = 1.5$, $L_1 = 2$, $\mu_2 = 1$, $L_2 = 2.5$
- $\Rightarrow$ regime $\boxed{p_2 = 0.9167}$
- We have $\mu_F = \mu_1 - L_2 = -1$ and $L_F = L_1 - \mu_2 = 1$
- Optimal constant stepsize for GD: $\gamma^* = \frac{2}{\sqrt{3}L_F}$ (Hadi Abbaszadehpeivasti et al. 2022)
- GD rate: $\boxed{p_{\text{GD}} = 1.5396}$; 67% better than p_2
- Best DCA splitting obtained with $\lambda^* = 1.0091$: $\tilde{f}_1^{\lambda^*} \in \mathcal{F}_{0.4009,0.9009}$ and $\tilde{f}_2^{\lambda^*} \in \mathcal{F}_{-0.0991,1.4009}$
- Best in the area of regime p_3: $\boxed{p_{\text{DCA}}^{\lambda^*} = 1.724}$: 12% larger than p_{GD}

Example 1.2: F smooth-convex

- Assume initial splitting $F = f_1 - f_2$ with $\mu_1 = 1$, $L_1 = 2$, $\mu_2 = 10.$, $L_2 = 1$
- $\Rightarrow$ regime $\boxed{p_5 = 2}$
- We have $\mu_F = \mu_1 - L_2 = 0$ and $L_F = L_1 - \mu_2 = 1.9$
- Optimal constant stepsize for GD after one iteration: $\gamma^* = \frac{3}{2L_F}$ (Rotaru et al. 2024)
- GD rate: $\boxed{p_{\text{GD}} = 0.831}$
- Best DCA splitting obtained with $\lambda^* = 0.4$: $\tilde{f}_1^{\lambda^*} \in \mathcal{F}_{0.6,1.6}$ and $\tilde{f}_2^{\lambda^*} \in \mathcal{F}_{-0.3,0.6}$
- Best in the area of regime p_5: $\boxed{p_{\text{DCA}}^{\lambda^*} = 3.3333}$

Example 2: Best DCA splitting on F smooth-nonconvex

- Assume $\mu_F = \mu_1 - L_2 = -0.5$ and $L_F = L_1 - \mu_2 = 1$. What is the best splitting?



Improvements on DCA rates

- $\max\{\mu_1, \mu_2\} < \min\{L_1, L_2\}$
- Relative improvement: $Ratio = \frac{P(\lambda^*) - P(0)}{P(0)}$
  - p^0: initial denominator before splitting
  - p^{λ^*}: optimal (largest) denominator obtained through curvature shifting

Setup	μ_1	L_1	μ_2	L_2	p^0	λ^*	p^{λ^*}	**Ratio**
$\mu_1 > \mu_2$	0.2	3	0.1	4	$p_2 = 0.4535$	0.1221	$p_3 = 0.6031$	33%
	0.2	1000	0.1	3	$p_1 = 0.3255$	0.1494	$p_1 = p_3 = 0.358$	10%
	1	2	0.5	1.5	$p_1 = 0.7600$	0.6733	$p_3 = 2.0321$	167%
$\mu_1 = \mu_2$	1	4	1	3	$p_1 = 0.4583$	$\mu_1 = 1$	$p_1 = p_2 = 0.8333$	81%
$\mu_1 < \mu_2$	0.1	3	0.2	4	$p_2 = 0.4539$	$\mu_1 = 0.1$	$p_2 = 0.602$	32%
	0.001	4	0.002	3	$p_1 = 0.4531$	$\mu_1 = 0.001$	$p_1 = 0.5835$	28%
$\mu_1 > 0;\ \mu_1 + \mu_2 > 0$	2	4	-1.75	3	$p_5 = 0.4688$	-0.4855	$p_3 = 0.52$	11%
	2.99	4	-2.9	3	$p_5 = 0.4994$	-0.935	$p_5 = 0.5076$	1.6%
$\mu_1 > 0;\ \mu_1 + \mu_2 < 0$	1	2	-1.5	1.5	–	-0.6526	$p_3 = 0.8516$	–

Numerical experiments: SPCA

- Sparse Principal Component Analysis (SPCA) (DC formulations in Journée et al. 2010 and Themelis et al. 2020) with elastic-net regularization

$$\underset{x\in\bar{B}(0,1)}{\text{minimize}}\, F(x) := \kappa\|x\|_1 + \eta\frac{\|x\|^2}{2} - \frac{1}{2}x^T\Sigma x$$

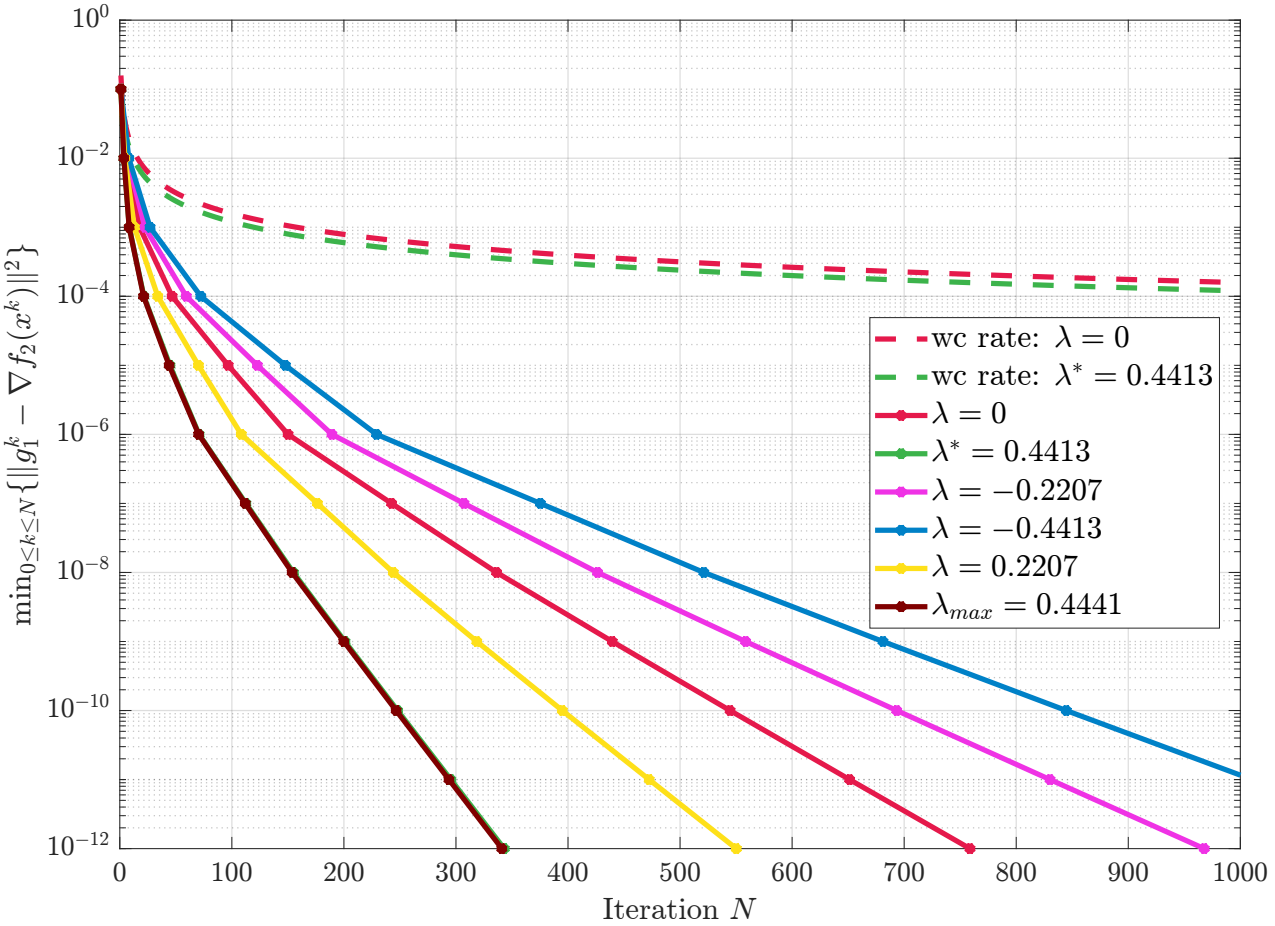
- $\bar{B}(0,1) := \{x \mid \|x\|_2 \leq 1\}$: closed Euclidean unit ball
- $\Sigma = A^T A$: sample covariance matrix
  - $A \in \mathbb{R}^{20n\times n}$ $(n = 200)$: sparse random matrix (normal distribution), normalized by maximum eigenvalue
- κ, η: l_1-(sparsity inducing) and l_2-regularization parameters
- $f_1(x) := \kappa\|x\|_1 + \eta\frac{\|x\|^2}{2} + \delta_{\bar{B}(0,1)}(x)$; thus: $\mu_1 = \eta$, $L_1 = \infty$
- $f_2(x) := \frac{1}{2}x^T\Sigma x$, with $\delta_C(x)$ as the indicator function of a set C
  - thus: $\mu_2 = \min\{\Lambda(\Sigma)\}$ and $L_2 = \max\{\Lambda(\Sigma)\}$ (extreme eigenvalues of Σ)
  - we use a sample of Σ such that: $f_2 \in \mathcal{F}_{\mu_2=0.3882, L_2=1}$
- $M = 1000$ random initial points $\in \bar{B}(0,1)$

Numerical experiments: SPCA

- Testing curvature shifting policy
  - $f_1^\lambda = f_1 - \lambda \frac{\|\cdot\|^2}{2} \in \mathcal{F}_{\eta-\lambda,\infty}$
  - $f_2^\lambda = f_2 - \lambda \frac{\|\cdot\|^2}{2} \in \mathcal{F}_{\mu_2-\lambda,L_2-\lambda}$
- $\lambda \in \{0, \pm\lambda^*, \pm 0.5\lambda^*, \lambda_{\max}\}$
  - λ^* is the optimal curvature shift
  - $\lambda_{\max} := \frac{\mu_1+\min\{\mu_1,\mu_2\}}{2}$ maximum shift guaranteeing convergence

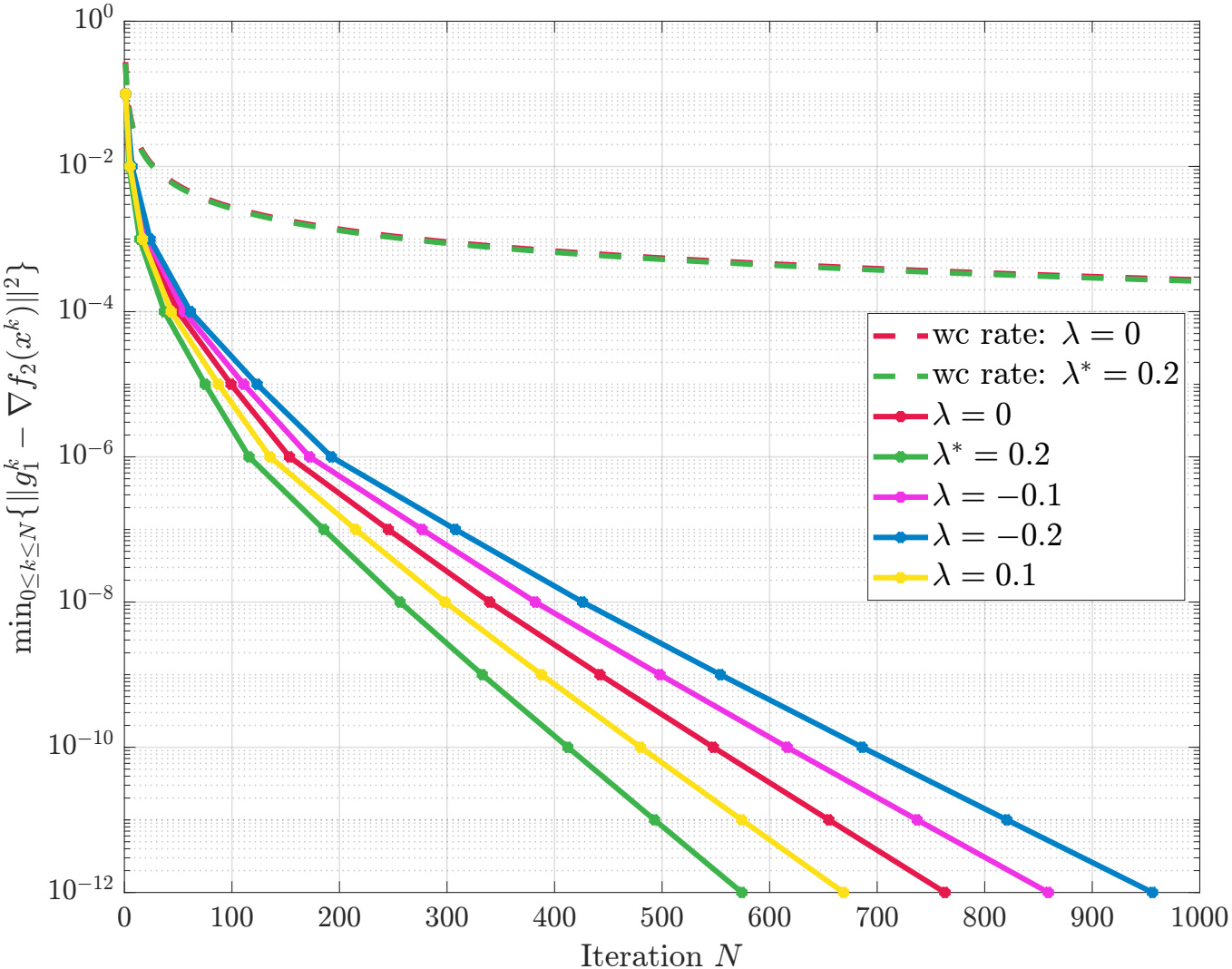
Case 1: $\mu_1 > \mu_2$

- $\eta = \mu_1 = 0.5$, $\kappa = 0.02$; $\lambda^* = 0.4413$, $\lambda_{\max} = 0.4441$
- if $\lambda > \mu_2 = 0.3882 \Rightarrow f_2^\lambda$ weakly convex
- better convergence by subtracting curvature



Case 2: $\mu_1 < \mu_2$

- $\eta = \mu_1 = 0.2$, $\kappa = 0.02$; $\lambda^* = \lambda_{\max} = \mu_1 = 0.02$
- $\approx 20\%$ improvement vs. initial splitting



Summary on curvature shifting

- If $\mu_1 \leq \mu_2$: best rates obtained by making f_1^λ convex
- If $\mu_1 > \mu_2$: best rates obtained for some $\mu_2 < 0$ (not covered by standard DCA!)
- PGD analogy: larger the stepsizes, better the performance

Curvature shifting in gradient mapping criterion

- General rates type:

$$\min_{0\le k\le N} \frac{\|x^k - x^{k+1}\|^2}{2} \le \frac{F(x^0) - F(x^{N+1})}{(\mu_1+\mu_2) + \boxed{P_N(L_1,\mu_1,L_2,\mu_2)}}$$

- No scaling of the iteration differences!
- Natural scaling through μ_1 (essentially, $\mu_1 = \gamma^{-1}$)
- Therefore, we must optimize in λ the denominator of:

$$\min_{0\le k\le N} \frac{\|\mu_1^\lambda(x^k - x^{k+1})\|^2}{2} \le \frac{F(x^0) - F(x^{N+1})}{\frac{\mu_1^\lambda+\mu_2^\lambda}{(\mu_1^\lambda)^2} + \boxed{\left(\mu_1^\lambda\right)^{-2} P_N(L_1^\lambda,\mu_1^\lambda,L_2^\lambda,\mu_2^\lambda)}}$$

Contents

Literature review

- For φ strongly convex and different performance metrics: A. B. Taylor et al. 2018
- Let $\rho = 1 - \gamma L_\varphi$

Initialization	$\|x^0 - x_\star\|^2$	$F(x^0) - F_{\text{lo}}$	$\|\nabla\varphi(x^0) + g_h^0\|^2$
$\|x^N - x_\star\|^2 \leq$	$\rho^{2N}\|x^0 - x_\star\|^2$	$\frac{2}{\mu}\rho^{2N}(F(x^0) - F_{\text{lo}})$	$\frac{1}{\mu^2}\rho^{2N}\|\nabla\varphi(x^0) + g_h^0\|^2$
$F(x^N) - F_{\text{lo}} \leq$	✗	$\rho^{2N}(F(x^0) - F_{\text{lo}})$	$\frac{1}{2\mu}\rho^{2N}\|\nabla\varphi(x^0) + g_h^0\|^2$
$\|\nabla\varphi(x^N) + g_h^N\|^2 \leq$	✗	**this work**	$\rho^{2N}\|\nabla\varphi(x^0) + g_h^0\|^2$

- For φ nonconvex, with $\mu_\varphi = -L_\varphi$ H. Abbaszadehpeivasti 2024
- General remark: We recover the same rates from the unconstrained case! (Thus the conjectures before...)

Getting PGD rates

- Obtained from DCA analysis by just replacing curvatures

DCA	$\leftrightarrow$	**PGD**
$F \in \mathcal{F}_{\mu_1 - L_2, L_1 - \mu_2}$	$\leftrightarrow$	$F \in \mathcal{F}_{\mu_\varphi + \mu_h, L_\varphi + L_h}$
μ_1	$\leftrightarrow$	$\gamma^{-1} + \mu_h$
L_1	$\leftrightarrow$	$\gamma^{-1} + L_h$
μ_2	$\leftrightarrow$	$\gamma^{-1} - L_\varphi$
L_2	$\leftrightarrow$	$\gamma^{-1} - \mu_\varphi$
$\mu_1 + \mu_2 > 0$	$\leftrightarrow$	$\gamma < \frac{2}{L_\varphi - \mu_h}$
$\mu_2 \geq 0$	$\leftrightarrow$	$\gamma \leq \frac{1}{L_\varphi}$

- Standard PGD setting: h is convex nonsmooth, hence $\mu_h = 0$, $L_h = \infty$, thus $\mu_1 = \gamma^{-1} > 0$ and $L_1 = \infty$

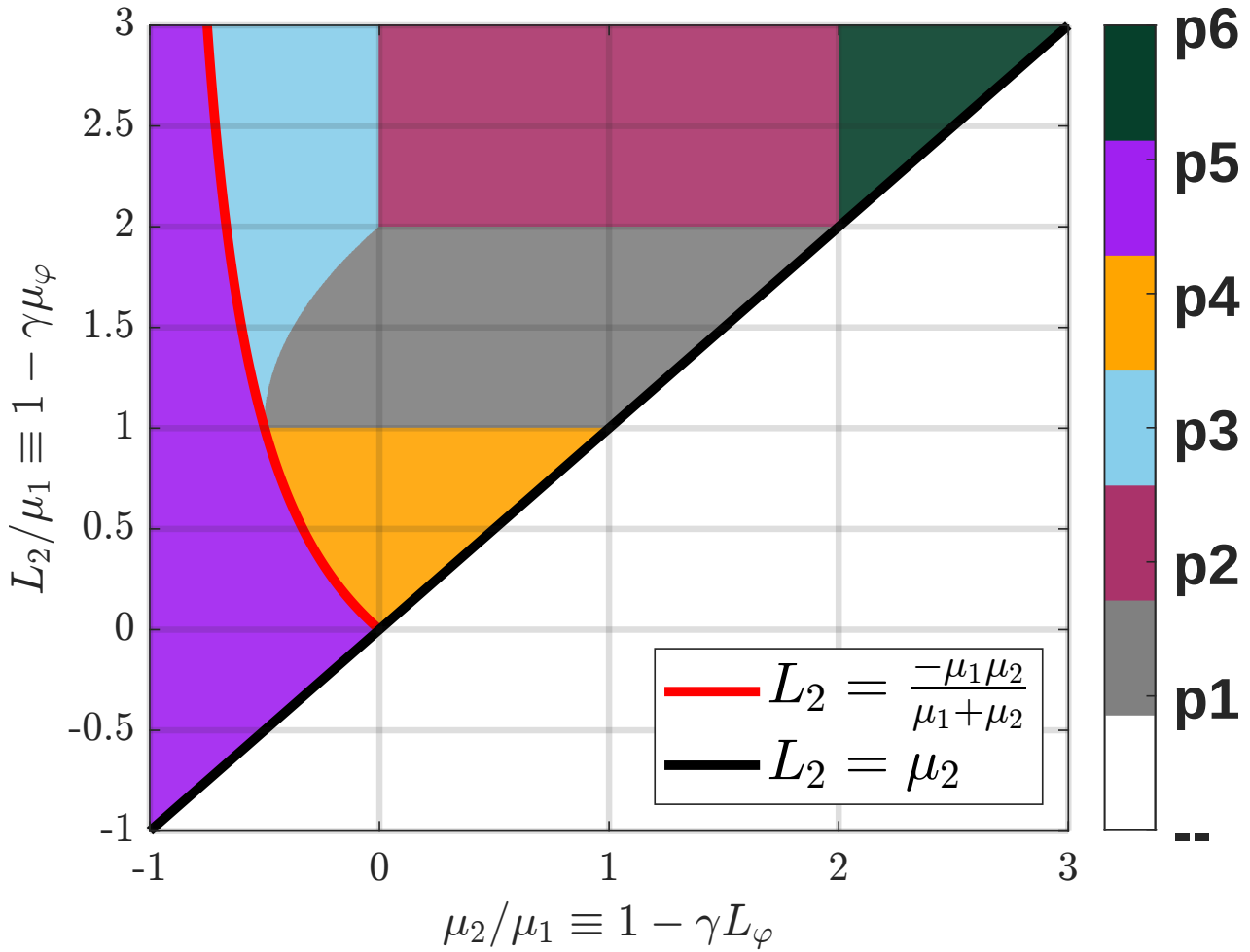
Getting PGD rates

- Obtained from DCA analysis by just replacing curvatures
- Standard PGD setting: h is convex nonsmooth, hence $\mu_h = 0$, $L_h = \infty$, thus $\mu_1 = \gamma^{-1} > 0$ and $L_1 = \infty$
- Regimes in gradient residual metric:

Convexity of φ	**Stepsize γ**	$\mu_2 = \gamma^{-1} - L_\varphi$	$L_2 = \gamma^{-1} - \mu_\varphi$	**Regime**
nonconvex $\mu_\varphi < 0$	$\gamma \in (0, \frac{1}{L_\varphi})$	$\mu_2 > 0$	$L_2 > \mu_1$	p_1
	$\gamma = \frac{1}{L_\varphi}$	$\mu_2 = 0$		p_1
	$\gamma \in (\frac{1}{L_\varphi}, \frac{2}{L_\varphi})$	$\mu_2 < 0$		p_1 or p_5
convex $\mu_\varphi = 0$	$\gamma \in (0, \frac{1}{L_\varphi})$	$\mu_2 > 0$	$L_2 = \mu_1$	$p_1 = p_4$
	$\gamma = \frac{1}{L_\varphi}$	$\mu_2 = 0$		$p_1 = p_4$
	$\gamma \in (\frac{1}{L_\varphi}, \frac{2}{L_\varphi})$	$\mu_2 < 0$		p_5
strongly convex $\mu_\varphi > 0$	$\gamma \in (0, \frac{1}{L_\varphi})$	$\mu_2 > 0$	$\lvert\mu_2\rvert < L_2 < \mu_1$	p_4
	$\gamma = \frac{1}{L_\varphi}$	$\mu_2 = 0$		p_4
	$\gamma \in (\frac{1}{L_\varphi}, \frac{2}{L_\varphi + \mu_\varphi})$	$\mu_2 < 0$		p_4 or p_5
	$\gamma \in [\frac{2}{L_\varphi + \mu_\varphi}, \frac{2}{L_\varphi})$		$L_2 \leq -\mu_2 < \mu_1$	p_5

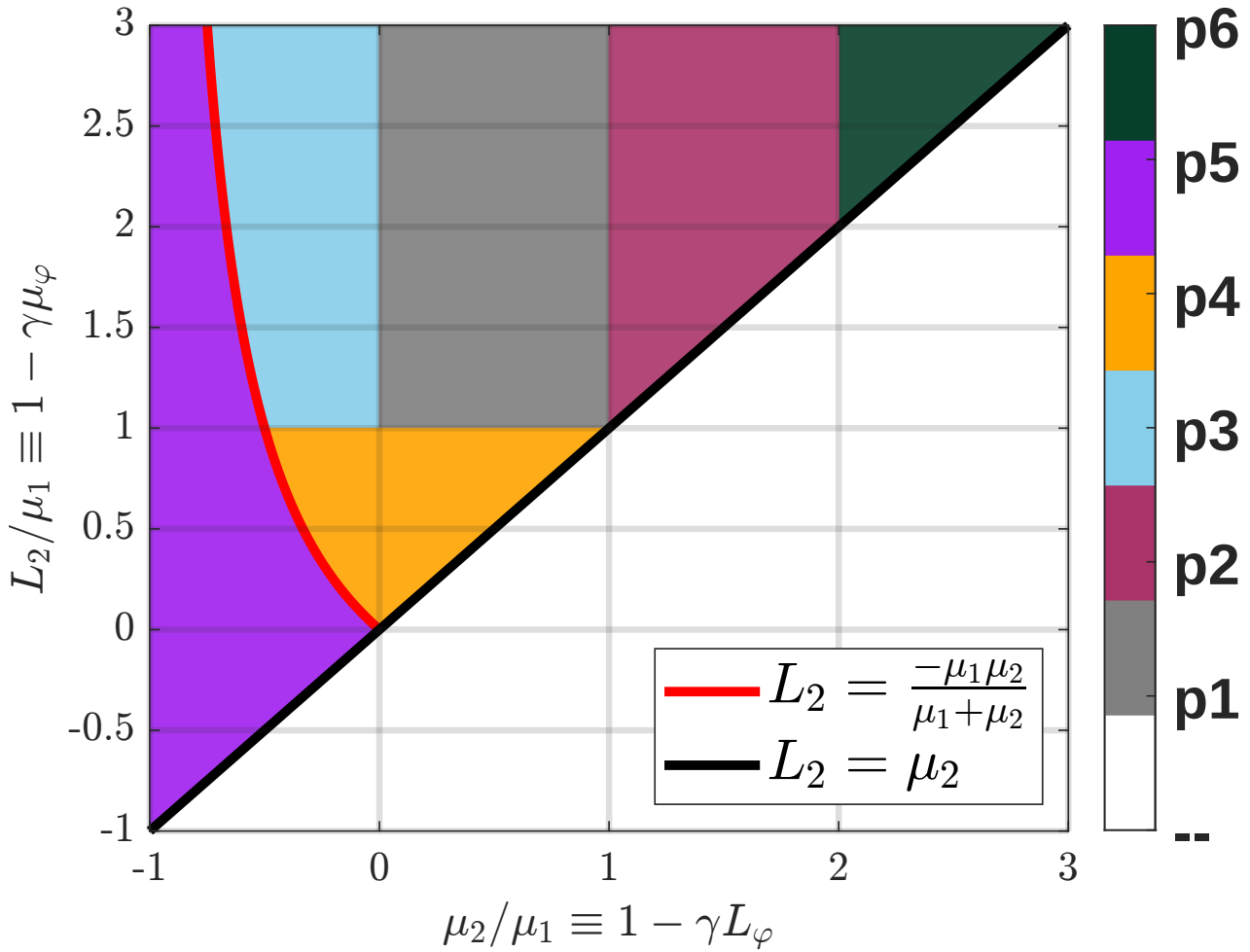
Gradient residual regimes

- $1-\gamma\mu_\varphi = \begin{cases} <1 & \text{strongly convex } \varphi \\ =1 & \text{convex } \varphi \\ >1 & \text{weakly-convex } \varphi \end{cases}$
- $1-\gamma L_\varphi = \begin{cases} \in(-1,0) & \text{long stepsizes } \gamma > L_\varphi^{-1} \\ \in[0,1) & \text{short stepsizes } 0<\gamma\leq L_\varphi^{-1} \\ >1 & \text{negative stepsizes } \gamma < L_\varphi^{-1} \end{cases}$ (here $L_1 = 2$)



Gradient mapping regimes

- $1-\gamma\mu_\varphi = \begin{cases} <1 & \text{strongly convex } \varphi \\ =1 & \text{convex } \varphi \\ >1 & \text{weakly-convex } \varphi \end{cases}$
- $1-\gamma L_\varphi = \begin{cases} \in(-1,0) & \text{long stepsizes } \gamma > L_\varphi^{-1} \\ \in[0,1) & \text{short stepsizes } 0<\gamma\leq L_\varphi^{-1} \\ >1 & \text{negative stepsizes } \gamma < L_\varphi^{-1} \end{cases}$ (here $L_1 = 2$)



Contents

Proof technique

- Combine inequalities on **consecutive** iterations.

Interpolation conditions (A. Taylor et al. 2017)

A set of triplets $\mathcal{T} = \{(x^i, g^i, f^i)\}_{i \in \mathcal{I}}$ is $\mathcal{F}_{\mu,L}$-interpolable ($-\infty < \mu < L > 0$), i.e., there exists $f \in \mathcal{F}_{\mu,L}$ such that $f(x^i) = f^i$, $g^i \in \partial f(x^i)$, **if and only if** $\forall i, j$:

$$f^i - f^j - \langle g^j, x^i - x^j \rangle \geq \frac{\mu}{2}\|x^i - x^j\|^2 + \frac{1}{2(L-\mu)}\|g^i - g^j - \mu(x^i - x^j)\|^2 \quad (Q_f^{[i,j]})$$

- **Key observation:** Any tight proof is a *linear combination* of interpolation inequalities.

Proof technique

- Combine inequalities on **consecutive** iterations.

Interpolation conditions (A. Taylor et al. 2017)

A set of triplets $\mathcal{T} = \{(x^i, g^i, f^i)\}_{i\in\mathcal{I}}$ is $\mathcal{F}_{\mu,L}$-interpolable ($-\infty < \mu < L > 0$), i.e., there exists $f \in \mathcal{F}_{\mu,L}$ such that $f(x^i) = f^i$, $g^i \in \partial f(x^i)$, **if and only if** $\forall i, j$:

$$f^i - f^j - \langle g^j, x^i - x^j\rangle \geq \frac{\mu}{2}\|x^i - x^j\|^2 + \frac{1}{2(L-\mu)}\|g^i - g^j - \mu(x^i - x^j)\|^2 \quad (Q_f^{[i,j]})$$

- **Key observation:** Any tight proof is a *linear combination* of interpolation inequalities.
- **Target inequality:** Consider nonnegative weights $\alpha^{[i,j]}$,$\beta^{[i,j]}$:

$$\sum_{0\leq i,j\leq N} \alpha^{[i,j]}Q_{f_1}^{[i,j]} + \beta^{[i,j]}Q_{f_2}^{[i,j]} \;:\; F(x^0) - F(x^N) \geq \tfrac{1}{2}\sum_{i=0}^{N}\sigma_i\|g_1^i - g_2^i\|^2 \quad (\text{☺})$$

Proof technique

Lemma

Let $f_1 \in \mathcal{F}_{\mu_1,L_1}$ and $f_2 \in \mathcal{F}_{\mu_2,L_2}$, with $L_2 > \mu_2$ and $L_1 > \mu_1 \geq 0$. Consider one DCA iteration connecting x^k and x^{k+1}. Let $\Delta x^k = x^k - x^{k+1}$, $g_1^j \in \partial f_1(x^j)$, $g_2^j \in \partial f_2(x^j)$ and $G^j = g_1^j - g_2^j$, where $j = \{k, k+1\}$. Then:

$$\Delta F(x^k) \geq \tfrac{\mu_1+\mu_2}{2}\|\Delta x^k\|^2 + \tfrac{\|G^{k+1}-\mu_2\Delta x^k\|^2}{2(L_2-\mu_2)} + \tfrac{\|G^k-\mu_1\Delta x^k\|^2}{2(L_1-\mu_1)} \qquad (B^k)$$

$$\langle G^k, \Delta x^k\rangle \geq \mu_1\|\Delta x^k\|^2 + \tfrac{\|G^k-\mu_1\Delta x^k\|^2}{L_1-\mu_1} \qquad (C_{f_1}^k)$$

$$\langle G^{k+1}, \Delta x^k\rangle \geq \mu_2\|\Delta x^k\|^2 + \tfrac{\|G^{k+1}-\mu_2\Delta x^k\|^2}{L_2-\mu_2} \qquad (C_{f_2}^k)$$

- Lemma obtained by summing up interpolation inequalities and using the property $g_1^{k+1} = g_2^k$.
- All distance-1 proofs correct inequality (B^k) via weighted combination of $(C_{f_1}^k)$ and $(C_{f_2}^k)$.

Contents

Summary

- Complete rates for DCA in terms of gradient residual and gradient mapping
- Curvature shifting improves DCA rates
- PGD rates obtained as byproduct
- **AISTATS 2025 paper available on**:

Bibliography I

Abbaszadehpeivasti, H. (2024). Performance Analysis of Optimization Methods for Machine Learning. CentER dissertation. CentER, Tilburg University. URL: https://books.google.be/books?id=YEn60AEACAAJ.

Abbaszadehpeivasti, Hadi et al. (July 2022). "The exact worst-case convergence rate of the gradient method with fixed step lengths for L-smooth functions". In: Optimization Letters 16.6, pp. 1649–1661. URL: https://doi.org/10.1007/s11590-021-01821-1.

— (2023). "On the Rate of Convergence of the Difference-of-Convex Algorithm (DCA)". In: Journal of Optimization Theory and Applications. URL: https://doi.org/10.1007/s10957-023-02199-z.

Dinh, Tao Pham and Hoai An Le Thi (1997). "Convex analysis approach to d.c. programming: Theory, Algorithm and Applications". In: Acta Matehmatica Vietnamica. Vol. 22.

Hiriart-Urruty, J.-B. (1989). "From Convex Optimization to Nonconvex Optimization. Necessary and Sufficient Conditions for Global Optimality". In: Nonsmooth Optimization and Related Topics. Boston, MA: Springer US, pp. 219–239. URL: https://doi.org/10.1007/978-1-4757-6019-4_13.

Journée, Michel et al. (2010). "Generalized Power Method for Sparse Principal Component Analysis". In: Journal of Machine Learning Research 11.15, pp. 517–553. URL: http://jmlr.org/papers/v11/journee10a.html.

Bibliography II

Lanckriet, Gert and Bharath K. Sriperumbudur (2009). "On the Convergence of the Concave-Convex Procedure". In: Advances in Neural Information Processing Systems 22. Ed. by Y. Bengio et al. URL: https://proceedings.neurips.cc/paper_files/paper/2009/file/8b5040a8a5baf3e0e67386c2e3a9b903-Paper.pdf.

Le Thi, Hoai An and Tao Pham Dinh (2018). "DC programming and DCA: thirty years of developments". In: Mathematical Programming 169.1, pp. 5–68. URL: https://doi.org/10.1007/s10107-018-1235-y.

Lipp, Thomas and Stephen Boyd (2016). "Variations and extension of the convex–concave procedure". In: Optimization and Engineering 17.2, pp. 263–287. URL: https://doi.org/10.1007/s11081-015-9294-x.

Rockafellar, R. Tyrrell and Roger J.-B. Wets (1998). Variational Analysis. Heidelberg, Berlin, New York: Springer Verlag.

Rotaru, Teodor, François Glineur, et al. (2024). "Exact worst-case convergence rates of gradient descent: a complete analysis for all constant stepsizes over nonconvex and convex functions". In: arXiv preprint arXiv:2406.17506. URL: https://arxiv.org/abs/2406.17506.

Rotaru, Teodor, Panagiotis Patrinos, et al. (2025). "Tight Analysis of Difference-of-Convex Algorithm (DCA) Improves Convergence Rates for Proximal Gradient Descent". In: arXiv preprint arXiv:2503.04486. URL: https://arxiv.org/abs/2503.04486.

Bibliography III

Taylor, Adrien et al. (Jan. 2017). "Smooth strongly convex interpolation and exact worst-case performance of first-order methods". en. In: Mathematical Programming 161.1-2, pp. 307–345. URL: http://link.springer.com/10.1007/s10107-016-1009-3 (visited on 12/20/2021).

Taylor, Adrien B. et al. (Aug. 2018). "Exact Worst-Case Convergence Rates of the Proximal Gradient Method for Composite Convex Minimization". In: Journal of Optimization Theory and Applications 178.2, pp. 455–476. URL: https://doi.org/10.1007/s10957-018-1298-1.

Themelis, Andreas et al. (2020). "A new envelope function for nonsmooth DC optimization". In: 2020 59th IEEE Conference on Decision and Control (CDC), pp. 4697–4702.

Toland, J. F. (1979). "A duality principle for non-convex optimisation and the calculus of variations". In: Archive for Rational Mechanics and Analysis 71.1, pp. 41–61. URL: https://doi.org/10.1007/BF00250669.

Yuille, Alan L and Anand Rangarajan (2001). "The Concave-Convex Procedure (CCCP)". In: Advances in Neural Information Processing Systems. Ed. by T. Dietterich et al. Vol. 14. MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/2001/file/a012869311d64a44b5a0d567cd20de04-Paper.pdf.

Yurtsever, Alp and Suvrit Sra (2022). "CCCP is Frank-Wolfe in disguise". In: Advances in Neural Information Processing Systems. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., pp. 35352–35364. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/e589022774244df75b97eae18bb3628d-Paper-Conference.pdf.