

Diplomarbeit

SQ2FISTA: Ein Lyapunov-basiertes beschleunigtes proximales Verfahren – Theorie und Vergleich mit FISTA

ausgeführt zum Zwecke der Erlangung des akademischen Grads

Diplom-Ingenieur / Master of Science (MSc)

eingereicht an der Technischen Universität Wien Institut für Analysis und Scientific
Computing

Diploma Thesis

SQ2FISTA: A Lyapunov-Based Accelerated Proximal Method – Theory and Comparison with FISTA

submitted in satisfaction of the requirements for the degree

Diplom-Ingenieur/ Master of Science (MSc)

submitted at TU Wien Institute of Analysis and Scientific Computing

Basil Victor Pétusseau, BSc

Matr.Nr.: 12230453

Betreuung: Univ.Prof. Dr.rer.nat. **Ansgar Jüngel**
Forschungsgruppe Analysis nichtlinearer PDEs
Institut für Analysis und Scientific Computing (E101)
Technische Universität Wien
Karlsplatz 13/E101, 1040 Wien, Österreich
Prof. Dr. **Takayasu Matsuo**
Department of Mathematical Informatics
Graduate School of Information Science and Technology
The University of Tokyo, Hongo Campus, Tokyo, Japan
Kansei Ushiyama
Ph.D. Student, Mathematical Informatics 3rd Laboratory
Department of Mathematical Informatics
Graduate School of Information Science and Technology
The University of Tokyo, Hongo Campus, Japan

Wien, am 16. September 2025

Basil Victor Pétusseau
Unterschrift Verfasser

Univ.Prof. Dr.rer.nat. Ansgar Jüngel
Unterschrift Betreuer

Affidavit

I declare in lieu of oath, that I wrote this thesis and carried out the associated research myself, using only the literature cited in this volume. If text passages from sources are used literally, they are marked as such.

I confirm that this work is original and has not been submitted for examination elsewhere, nor is it currently under consideration for a thesis elsewhere.

I acknowledge that the submitted work will be checked electronically-technically using suitable and state-of-the-art means (plagiarism detection software). On the one hand, this ensures that the submitted work was prepared according to the high-quality standards within the applicable rules to ensure good scientific practice “Code of Conduct” at the TU Wien. On the other hand, a comparison with other student theses avoids violations of my personal copyright.

Acknowledgements

I would like to express my sincere gratitude to **Univ. Prof. Dr. rer. nat. Ansgar Jüngel** for his inspiring guidance and constant encouragement throughout this work. I am especially grateful for his warm support of my idea to complete my master's thesis in Japan from the very beginning, for providing me with invaluable information and practical advice about studying there, and for always being available with remarkable efficiency and kindness.

My heartfelt thanks go to **Prof. Dr. Takayasu Matsuo**, whose perspective from the University of Tokyo provided invaluable insight into the design of structure-preserving methods and their broader connections in applied mathematics. I am also deeply thankful for how warmly he welcomed me into his laboratory, how smoothly he integrated me into his Japanese research group, and how consistently he made sure that everything went well for me during my stay.

I am equally grateful to Ph.D. Student **Kansei Ushiyama**, a truly outstanding researcher who provided daily support at my side. He generously explained the main concepts and insights of his paper with enthusiasm, always answered all my questions — even the simple ones — and helped me build a much clearer understanding of optimization methods while greatly expanding my skills in PyTorch. Thanks to him I not only deepened my knowledge of mathematics, but also had the chance to discover and appreciate his cultural background, which he shared with genuine joy.

I also wish to thank the entire laboratory group, whose members were unfailingly kind to me and left me with a lasting and wonderful impression of my time in Tokyo.

Last, but not least, I would like to thank from the bottom of my heart my life partner, who is writing her diploma thesis simultaneously in that most noble of vocations, Architecture. For her ever so precious support, for coming with me to Japan, and for walking by my side wherever our destiny may take us. **Arianna**, thank you.

WIR MÜSSEN WISSEN
WIR WERDEN WISSEN

WE MUST KNOW

WE WILL KNOW

David Hilbert

Inhaltsverzeichnis

1	Introduction	9
1.1	Motivation and Scope	9
1.2	Problem Class, Assumptions, and Notation	10
1.3	Algorithms at a Glance	10
1.4	Contributions	11
1.5	Expectations	12
1.6	Reading Guide and Chapter Outline	12
1.7	Summary	13
2	Gradient Descent	14
2.1	The Gradient Descent Method	14
2.2	Smoothness and Strong Convexity Assumptions	15
2.3	Convergence Analysis of Gradient Descent	17
2.3.1	Convex case: sublinear convergence	17
2.3.2	Strongly convex case: exponential (a.k.a. linear) convergence	18
2.4	Gradient Flow Viewpoint	19
2.5	Lyapunov Function Perspective	19
3	Nesterov’s Accelerated Gradient Method	21
3.1	Introduction and Motivation	21
3.2	The NAG Algorithm	22
3.2.1	Smooth convex case (increasing momentum)	22
3.2.2	Smooth strongly convex case (constant momentum)	22
3.3	Convergence in the Smooth Convex Case	23
3.4	Convergence in the Smooth Strongly Convex Case	25
3.5	Lyapunov Functions: Role and Basic Constructions	26
3.6	Continuous-Time Limits and Lyapunov Analyses	27
3.6.1	Convex case: the Nesterov ODE and $O(1/t^2)$ decay	27
3.6.2	Strongly convex case: constant damping and exponential decay	28
3.7	Heavy-Ball Momentum: Comparison and Bridges	30
3.8	Practical Notes and a Bridge to FISTA	30
4	FISTA – Accelerated Proximal Gradient Method	32
4.1	Introduction and Motivation	32

4.2	Proximal Operators: Definitions, Properties, and Examples	33
4.2.1	Definition and optimality condition	33
4.2.2	Resolvent form and monotone operator viewpoint	33
4.2.3	Firm nonexpansiveness and nonexpansiveness	33
4.2.4	Three-point identity for proximal steps	34
4.2.5	Canonical examples of proximal operators	35
4.3	ISTA: Proximal Gradient as a Baseline	35
4.3.1	Algorithm and basic properties	35
4.3.2	Convergence rate of ISTA (convex case)	35
4.4	FISTA: Algorithm, Derivation, and $\mathcal{O}(1/k^2)$ Convergence	36
4.4.1	Derivation by Nesterov-type extrapolation	36
4.4.2	A potential function and the main inequality	37
4.4.3	Accelerated convergence	37
4.5	Strongly Convex Acceleration	38
4.5.1	Constant-momentum variant (simple and effective)	38
4.5.2	Estimate-sequence variant (robust when μ is small)	41
4.6	Backtracking, Monotone FISTA, and Practical Enhancements	41
4.6.1	Backtracking FISTA (unknown L)	41
4.6.2	Monotone FISTA (MFISTA)	42
4.6.3	Adaptive restarts (unknown μ)	42
4.6.4	Inexact or stochastic oracles (brief note)	42
4.7	ISTA vs. FISTA: Structure, Cost, and Behavior	42
4.8	Worked Examples of Proximal Steps in FISTA	42
4.8.1	Lasso: $h(x) = \lambda\ x\ _1$	42
4.8.2	Tikhonov regularization: $h(x) = \frac{\lambda}{2}\ x\ ^2$	43
4.9	Pseudocode Summary (Convex and Strongly Convex Cases)	43
4.10	Continuous-Time View	44
4.11	Practical Notes for Implementation	44
4.12	Foreshadowing: From FISTA to SQ2FISTA	44
5	Methodology: From Differential Equations to SQ2FISTA	45
5.1	Overview and Objectives	45
5.2	Problem Setting and Assumptions	45
5.3	Continuous-Time Foundations: Gradient Flow and Inertial ODEs	46
5.3.1	Gradient flow and its Lyapunov analysis	46
5.3.2	Accelerated ODEs: vanishing vs. constant damping	47
5.3.3	A unifying hyperbolic-damped ODE and its Lyapunov energy	47
5.4	Weak Discrete Gradients (wDG): A Continuous–Discrete Bridge	48
5.5	From ODE to Method: Discretization via wDG	49
5.6	Interpretation and Comparison	59
5.7	Practical Notes (Theory-Facing)	59

5.8	Context for Chapter 6	60
6	Numerical Comparison—SQ2FISTA vs. FISTA on Weakly Convex Proximal Problems	61
6.1	Overview and Objectives	61
6.2	SCAD as a Weakly Convex Proximal Operator	62
6.2.1	Definition and weak convexity	62
6.2.2	Closed-form proximal map and step-size condition	62
6.3	Models	63
6.3.1	Ill-Conditioned Quadratic	63
6.3.2	Smoothed Hinge SVM (Huberized hinge)	63
6.3.3	Saturated Nonlinear Regression with tanh Link	63
6.4	Algorithms (Pseudocode)	64
6.4.1	FISTA (strongly convex setting)	64
6.4.2	SQ2FISTA (handles weakly convex h)	64
6.4.3	Fairness check: FISTA(δ) (convexified prox)	65
6.5	Convexified FISTA (δ -split) and Comparison	65
6.6	Experimental Protocol	66
6.7	Results	66
6.7.1	Smoothed Hinge SVM + SCAD: SQ2FISTA clearly faster	67
6.7.2	Saturated Nonlinear Regression (tanh link) + SCAD: near tie	67
6.7.3	Ill-conditioned quadratic + SCAD: slight edge for FISTA	68
6.8	Fairness Check: FISTA(δ) vs. SQ2FISTA	69
6.9	Diagnostics and Ablations	70
6.10	Implementation Notes (for Reproducibility)	71
6.11	Summary and Outlook	71
7	Pseudocodes and Overview	72
7.1	Algorithm Pseudocode	72
7.2	Limitations and Validity	74

Kapitel 1

Introduction

1.1 Motivation and Scope

Many optimization problems in machine learning, signal processing, and computational statistics are *composite* in nature:

$$\min_{x \in \mathbb{R}^n} F(x) = f(x) + h(x), \quad (1.1.1)$$

where f is smooth (differentiable with L -Lipschitz gradient) and h is proper, closed, and often non-smooth but *proximable*. Classic examples include ℓ_1 penalties for sparsity, constraints encoded as indicator functions, and robust losses coupled with structured regularization. For such problems, first-order methods dominate practice: they scale, exploit problem structure, and admit sharp iteration-complexity guarantees. Two algorithmic ideas underpin state-of-the-art performance on (1.1.1): (i) *proximal splitting*, which decouples f and h by a gradient step on f followed by a proximity step on h (ISTA/proximal gradient); and (ii) *acceleration*, which injects extrapolation (momentum) to achieve the optimal $\mathcal{O}(1/k^2)$ rate in the smooth convex regime (Nesterov acceleration; FISTA in the composite case).

In modern applications, the regularizer h is sometimes only *weakly convex* (e.g., SCAD), which invalidates the standard convex proximal framework unless one *convexifies* h of the composite model. Moreover, empirical performance often hinges on how well the discrete method mirrors an underlying *continuous-time* energy decay. This thesis develops a principled and unified treatment of these themes: we give complete, self-contained Lyapunov proofs for GD, NAG, ISTA, and FISTA; We derive and analyze a new accelerated scheme, *SQ2FISTA*, proposed by PhD student Mr. Ushiyama, from a carefully designed time-varying inertial ODE via a *weak discrete gradient* (wDG) discretization, and prove its $\mathcal{O}(1/k^2)$ and linear rates; we extend FISTA to weakly convex proximals through a *convexified* variant, denoted $FISTA(\delta)$, leveraging a prox-shift identity that restores convexity and stability; and we provide PyTorch-style code listings that exactly implement the analyzed algorithms with *safety clamps* for SCAD. A central empirical finding is that *SQ2FISTA* strictly improves over *standard* FISTA and closely matches (often ties) the performance of $FISTA(\delta)$ on realistic synthetic benchmarks, validating the design principle: when h is weakly convex, either adopt an ODE-inspired accelerated discretization (SQ2FISTA) or convexify the proximal part ($FISTA(\delta)$) — both are principled and high-performing, while *plain* FISTA is typically suboptimal.

1.2 Problem Class, Assumptions, and Notation

We consider (1.1.1) with the following standing assumptions unless stated otherwise. The smooth part $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable and L -smooth: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all x, y . In the strongly convex regime, f is μ_m -strongly convex ($\mu_m > 0$ future Model- μ). The regularizer $h : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ is proper, closed, and *proximable*, with proximity operator

$$\text{prox}_{\lambda h}(v) = \arg \min_x \left\{ h(x) + \frac{1}{2\lambda} \|x - v\|^2 \right\}.$$

When h is *weakly convex* (e.g. SCAD), there exists $\rho \geq 0$ such that $h(x) + \frac{\rho}{2}\|x\|^2$ is convex; equivalently, h has curvature $\mu_p = -\rho \leq 0$, (future Proximal- μ). The *total* strong convexity we use throughout is

$$\mu_{\text{tot}} = \mu_m + \mu_p.$$

We denote by x^* a global minimizer of F , and write $F^* = F(x^*)$. A standard residual is the *prox-gradient mapping* at stepsize $\eta > 0$:

$$G_\eta(x) := \frac{1}{\eta} \left(x - \text{prox}_{\eta h} \left(x - \eta \nabla f(x) \right) \right), \quad \|G_\eta(x)\| = 0 \Leftrightarrow 0 \in \nabla f(x) + \partial h(x). \quad (1.2.1)$$

Convexification by prox shift. When h is weakly convex with $\mu_p < 0$, we use the identity

$$\underbrace{\text{prox}_{\eta \left(h + \frac{\delta}{2} \|\cdot\|^2 \right)}(v)}_{\text{convex for } \delta \geq -\mu_p} = \text{prox}_{\frac{\eta}{1+\eta\delta} h} \left(\frac{1}{1+\eta\delta} v \right), \quad (1.2.2)$$

and set $\delta = -\mu_p$ so that $h + \frac{\delta}{2} \|\cdot\|^2$ is convex. Algebraically, this transfers curvature from h into f , replacing (L, μ_f) by $(L + \delta, \mu_f - \delta)$ in the smooth part. We exploit (6.4.2) to define $FISTA(\delta)$.

1.3 Algorithms at a Glance

We study six baseline methods in the following order and notation consistent with the rest of the thesis. Gradient Descent (GD) uses the update $x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$. Nesterov's Accelerated Gradient (NAG) evaluates a gradient at an extrapolated point, achieving $\mathcal{O}(1/k^2)$ in convex problems and a linear rate with factor $(1 - \sqrt{\mu_m/L})^k$ in strongly convex settings. ISTA (proximal gradient) attains $\mathcal{O}(1/k)$ in the convex case (and linear convergence $(1 - c\mu_m/L)^k$ with suitable tuning under strong convexity). FISTA (accelerated ISTA) achieves $\mathcal{O}(1/k^2)$ in the convex case and $(1 - \sqrt{\mu_m/L})^k$ when strongly convex with tuned momentum. SQ2FISTA is our ODE-informed accelerated proximal method designed via wDG, attaining $\mathcal{O}(1/k^2)$ in convex problems and a linear factor of the form $(1 - (\sqrt{2\mu_{\text{tot}}/L_{\text{eff}}}))^k$ in the strongly convex composite, where L_{eff} is the effective smoothness constant arising in the SQ2FISTA analysis. Finally, to find SQ2FISTA concurrence, the convexified variant, $FISTA(\delta)$, applies (6.4.2); its linear factor is $(1 - \sqrt{\mu_{\text{tot}}/(L + \delta)})^k$ when $\mu_{\text{tot}} > 0$.

Tab. 1.1: Canonical worst-case iteration complexity (deterministic, exact gradients). Strongly convex entries use an error contraction factor $\rho \in (0, 1)$.

Method	Convex case	Strongly convex case ($\mu_{\text{tot}} > 0$)
GD	$\mathcal{O}(1/k)$	$(1 - \mu_m/L)^k$
ISTA	$\mathcal{O}(1/k)$	$(1 - c\mu_m/L)^k$
NAG	$\mathcal{O}(1/k^2)$	$(1 - \sqrt{\mu_m/L})^k$
FISTA	$\mathcal{O}(1/k^2)$	$(1 - \sqrt{\mu_m/L})^k$
FISTA(δ)	$\mathcal{O}(1/k^2)$	$(1 - \sqrt{\mu_{\text{tot}}/(L + \delta)})^k$
SQ2FISTA	$\mathcal{O}(1/k^2)$	$(1 - (\sqrt{2\mu_{\text{tot}}/L_{\text{eff}}}))^k$

1.4 Contributions

C1. Unified Lyapunov proofs for GD, NAG, ISTA, and FISTA. We present complete, concise proofs of the classical rates using a common Lyapunov perspective. For GD, a descent lemma combined with a strong-convexity sandwich yields $\mathcal{O}(1/k)$ (convex) and $(1 - \mu_m/L)^k$ (strongly convex). For NAG, a quadratic-in- k potential establishes $\mathcal{O}(1/k^2)$ (convex) and factor $1 - \sqrt{\mu_m/L}$ (strongly convex). For ISTA/FISTA, three-point inequalities and an estimate-sequence potential yield $\mathcal{O}(1/k)$ and $\mathcal{O}(1/k^2)$, respectively; the strongly convex FISTA variant attains the accelerated linear factor as above.

C2. SQ2FISTA via an inertial ODE and weak discrete gradients. We design a time-varying inertial flow with *hyperbolic* damping and construct a discrete scheme using a weak discrete gradient (wDG) identity so that a discrete energy E_k provably decreases. Solving a scalar recurrence for the weight sequence A_k yields an optimal schedule, leading to $F(x_k) - F^* = \mathcal{O}(1/k^2)$ in the convex case and $F(x_k) - F^* \leq C\rho^k$ in the strongly convex case with $\rho \approx 1 - (2\sqrt{\mu_{\text{tot}}/L_{\text{eff}}})$. The proof mirrors the continuous Lyapunov decay and is fully discrete, requiring only smoothness, (total) strong convexity when present, and a proximal step.

C3. Convexified FISTA for weakly convex proximals. We introduce FISTA(δ) by convexifying the proximal part via (6.4.2) with $\delta = -\mu_p$. This preserves the standard FISTA structure but replaces gradients of f by those of $f_\delta := f - \frac{\delta}{2}\|\cdot\|^2$ (hence $L \mapsto L + \delta$, $\mu_m \mapsto \mu_m - \delta$). We prove the same $\mathcal{O}(1/k^2)$ rate in the convex case and a linear rate governed by μ_{tot} in the strongly convex case. For SCAD we implement *safety clamps* to keep the closed-form prox well-defined.

C4. Pseudocodes. We provide Pseudo code for all optimizers (GD, NAG, ISTA/FISTA, FISTA(δ), SQ2FISTA) and proximals (including SCAD with safety clamps), as well as three model families: *Smoothed Hinge SVM*, *Saturated Nonlinear Regression (tanh link)*, and *Ill-conditioned quadratic*. The listings follow a uniform interface, expose exact L bounds used by the theory, and record both objective gaps and prox-gradient norms.

1.5 Expectations

Validity. All analyses are deterministic and assume exact (or high-precision) gradients, exact prox steps (with safe closed forms for SCAD), and step sizes based on global L bounds (or their safe overestimates). The linear-rate guarantees use the *total* strong convexity μ_{tot} when present.

Limitations. We do not analyze stochastic gradients, line searches, or adaptive step sizes. With weakly convex g , the *original* nonconvex composite need not be globally convex; our guarantees either (i) target the convexified problem (FISTA(δ)) or (ii) control a discrete Lyapunov function for SQ2FISTA that ensures objective decrease and stationarity, but not necessarily avoidance of all nonconvex stationary points beyond our constructed setting. SCAD’s region-2 proximal formula requires a positive denominator; we enforce this via *safety clamps* — benign in practice but slightly perturbative.

Takeaway. In regimes where h is weakly convex (a common practical case), *standard* FISTA is not the right baseline. Either convexify (FISTA(δ)) or discretize a well-chosen inertial flow (SQ2FISTA). Both are principled; empirically they are neck-and-neck, with SQ2FISTA offering a physics-consistent derivation and FISTA(δ) offering a minimal patch over a standard workhorse.

1.6 Reading Guide and Chapter Outline

Chapter 2 revisits Gradient Descent, its Lyapunov analysis, and continuous gradient flow, providing full proofs in both convex and strongly convex regimes. Chapter 3 develops Nesterov’s Accelerated Gradient (NAG), including discrete proofs of $\mathcal{O}(1/k^2)$ and linear rates, alongside their ODE counterparts. Chapter 4 introduces proximal operators, ISTA, FISTA, and a strongly convex accelerated variant; proofs rely on three-point inequalities and estimate sequences. Chapter 5 bridges continuous-time inertial dynamics and discrete methods via *weak discrete gradients* (wDG), yielding a unifying template for Lyapunov-based design and derives *SQ2FISTA* from a hyperbolically damped inertial ODE. We prove energy decrease, derive the optimal weight recurrence, and establish convergence rates. Chapter 6 reports an empirical study on three model families (*Smoothed Hinge SVM*, *Saturated Nonlinear Regression (tanh link)*, and *Ill-conditioned quadratic*) with SCAD, comparing *plain* FISTA, FISTA(δ), and *SQ2FISTA*. Chapters 7 consolidate conclusions, nuanced comparisons between *standard* vs. *convexified* FISTA and SQ2FISTA, and discuss limitations.

Tab. 1.2: Main symbols and conventions.

Symbol	Meaning
x^*	a global minimizer of $F = f + h$; $F^* = F(x^*)$
L	Lipschitz smoothness constant of f
μ_m	strong convexity constant of f (of the Model)
μ_p	curvature of h (weak convexity: $\mu_p < 0$; convex: $\mu_p \geq 0$)
μ_{tot}	total strong convexity: $\mu_{\text{tot}} = \mu_m + \mu_p$
$\text{prox}_{\lambda h}$	proximity operator of h with parameter $\lambda > 0$
$G_\eta(x)$	prox-gradient mapping at stepsize η ; see (1.2.1)
$\text{FISTA}(\delta)$	FISTA with prox shift $\delta = -\mu_p$; cf. (6.4.2)
SQ2FISTA	ODE-informed accelerated proximal method designed via wDG
L_{eff}	effective Lipschitz constant (in SQ2FISTA analysis)

1.7 Summary

In summary, this thesis contributes a unified analysis and implementation playbook for accelerated composite optimization in the presence of weakly convex proximals. The new method *SQ2FISTA*, developed by PhD student Mr. Ushiyama from the University of Tokyo, derived from first principles, and the pragmatic *FISTA*(δ) baseline together form a robust toolkit that is theoretically sound and empirically validated; *plain* FISTA remains a useful point of reference but is often dominated in the regimes that motivate contemporary applications.

Kapitel 2

Gradient Descent

2.1 The Gradient Descent Method

We consider the unconstrained minimization of a differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. The *gradient descent* method is an iterative algorithm that updates the variable in the direction of steepest descent of f . Starting from an initial guess $x_0 \in \mathbb{R}^n$, the gradient descent iteration is given by

$$x_{k+1} = x_k - \alpha \nabla f(x_k), \tag{2.1.1}$$

for $k = 0, 1, 2, \dots$. Here $\alpha > 0$ is a step size (also called the learning rate). The update (2.1.1) means we move from the current point x_k in the negative gradient direction $-\nabla f(x_k)$, which is the direction of steepest decrease of f at x_k .

Gradient descent can be derived as the optimization method that, at each iteration, minimizes a first-order local approximation of f plus a quadratic regularization term:

$$x_{k+1} = \arg \min_x \left\{ f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{1}{2\alpha} \|x - x_k\|^2 \right\}.$$

Lemma 2.1.1 (From the quadratic model to the gradient step). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable and let $\alpha > 0$. Consider the quadratic model*

$$Q_k(x) := f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{1}{2\alpha} \|x - x_k\|^2.$$

Then Q_k is (strictly) convex in x and its unique minimizer is

$$x_{k+1} = \arg \min_x Q_k(x) = x_k - \alpha \nabla f(x_k).$$

In particular, this shows that the gradient descent update (2.1.1) follows directly from the quadratic model minimization.

Proof. There are two equivalent ways to see this.

(i) *First-order optimality.* Since Q_k is a differentiable quadratic in x , the minimizer is characterized by $\nabla Q_k(x_{k+1}) = 0$. Differentiating,

$$\nabla Q_k(x) = \nabla f(x_k) + \frac{1}{\alpha}(x - x_k).$$

Setting this to zero at $x = x_{k+1}$ gives

$$\nabla f(x_k) + \frac{1}{\alpha}(x_{k+1} - x_k) = 0 \iff x_{k+1} = x_k - \alpha \nabla f(x_k),$$

which is exactly (2.1.1). Strict convexity (the Hessian of Q_k is $\alpha^{-1}I \succ 0$) ensures uniqueness.

(ii) *Completion of squares.* Rewrite Q_k as

$$Q_k(x) = f(x_k) - \frac{\alpha}{2} \|\nabla f(x_k)\|^2 + \frac{1}{2\alpha} \|x - (x_k - \alpha \nabla f(x_k))\|^2.$$

The second term does not depend on x , and the last term is minimized uniquely when $x = x_k - \alpha \nabla f(x_k)$, giving the same conclusion.

Remark 2.1.2. *Lemma 2.1.1 is the standard variational characterization of a gradient step: a single gradient descent step is precisely the unique minimizer of the local quadratic upper model of f at x_k with curvature $1/\alpha$.*

2.2 Smoothness and Strong Convexity Assumptions

To analyze the convergence of gradient descent, we assume f is *smooth* (has Lipschitz continuous gradient) and, for stronger results, that f is also *strongly convex*. We give the formal definitions below.

Definition 2.2.1 (L -smoothness). *A differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is called L -smooth (or has L -Lipschitz continuous gradient) if there exists a constant $L > 0$ such that*

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|^2, \quad \text{for all } x, y \in \mathbb{R}^n.$$

Equivalently, the gradient satisfies $\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$ for all x, y .

In words, f is globally upper bounded by its first-order Taylor expansion plus a quadratic remainder with curvature parameter L ; that is, the quadratic model with coefficient $L/2$ majorizes f everywhere.

An important consequence is the *descent lemma*: for an L -smooth f , any step of gradient descent with step size $\alpha \leq 2/L$ is guaranteed to decrease the function value. In particular, using the inequality in Definition 2.2.1 with $y = x - \alpha \nabla f(x)$, one obtains

$$f(x - \alpha \nabla f(x)) \leq f(x) - \alpha \left(1 - \frac{\alpha L}{2}\right) \|\nabla f(x)\|^2. \quad (2.2.1)$$

For $0 < \alpha \leq \frac{2}{L}$, the factor $1 - \frac{\alpha L}{2}$ is nonnegative, and thus (2.2.1) guarantees that $f(x - \alpha \nabla f(x)) \leq f(x)$, with a strict decrease whenever $\nabla f(x) \neq 0$. In particular, under this step size condition, the sequence of function values $\{f(x_k)\}$ in gradient descent is non-increasing. Moreover, summing the inequality (2.2.1) over iterations shows that $\sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2$ is finite (since $f(x_k)$ is bounded below by $f^* = \min_x f(x)$), which in turn implies $\|\nabla f(x_k)\| \rightarrow 0$ as $k \rightarrow \infty$. Thus, gradient descent finds a *stationary point* of f . In the special case that f is also convex so that any stationary point is a global minimum, we obtain convergence to the minimizer. We formalize the convexity assumption next.

Definition 2.2.2 (μ -strong convexity). *A differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is μ -strongly convex for $\mu > 0$ if*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2} \|y - x\|^2, \quad \text{for all } x, y \in \mathbb{R}^n.$$

If f is twice differentiable, this is equivalent to the pointwise bound

$$v^\top \nabla^2 f(x) v \geq \mu \|v\|^2 \quad \text{for all } x \in \mathbb{R}^n \text{ and } v \in \mathbb{R}^n,$$

i.e., the Hessian is uniformly bounded below in the quadratic form sense.

Strong convexity means that f has a quadratic lower bound around any point, which ensures f has a *unique* minimizer (denoted $x^* = \arg \min_x f(x)$). Every μ -strongly convex function is in particular convex (taking $\mu = 0$ recovers the definition of ordinary convexity). The strong convexity parameter $\mu > 0$ can be thought of as a measure of curvature near the minimizer; a larger μ means the function grows more steeply away from x^* . In many results, the ratio $\kappa := L/\mu \geq 1$ (assuming $\mu > 0$) appears; this κ is called the *condition number* of f .

It is useful to note a relationship between smoothness and strong convexity: if f is L -smooth and μ -strongly convex, then for all $x \in \mathbb{R}^n$,

$$\frac{\mu}{2} \|x - x^*\|^2 \leq f(x) - f(x^*) \leq \frac{L}{2} \|x - x^*\|^2. \quad (2.2.2)$$

The right inequality follows by applying L -smoothness at x^* and using $\nabla f(x^*) = 0$. The left inequality is a consequence of strong convexity (apply Definition 2.2.2 with $y = x^*$ and $\nabla f(x^*) = 0$). Thus, for such functions, $f(x) - f(x^*)$, $\|x - x^*\|^2$, and $\|\nabla f(x)\|^2$ are proportional up to constants.

Convention on rate terminology.

Definition 2.2.3 (Rates of convergence for sequences). *Let $\{e_k\}_{k \geq 0}$ be a nonnegative sequence (e.g., $e_k = f(x_k) - f^*$ or $e_k = \|x_k - x^*\|$).*

- Sublinear rate: $e_k = \mathcal{O}(1/k^p)$ for some $p > 0$ (typical in convex smooth optimization with $p = 1$).
- Exponential (a.k.a. linear) rate: $e_k = \mathcal{O}(\rho^k)$ for some constant $\rho \in (0, 1)$ (typical under strong convexity).

We use “exponential” and “linear (geometric)” synonymously for rates of the form $\mathcal{O}(\rho^k)$.

2.3 Convergence Analysis of Gradient Descent

We now analyze the convergence rate of the gradient descent iterates (2.1.1) under the assumptions above. We consider two cases: first assuming only convexity (which gives a sublinear $\mathcal{O}(1/k)$ rate), and then assuming strong convexity (which gives an exponential rate). Throughout, we assume a fixed step size $0 < \alpha \leq \frac{2}{L}$ (in fact we will often take $\alpha = 1/L$ for simplicity). The proofs use basic tools like the inequalities (2.2.1) and (2.2.2).

2.3.1 Convex case: sublinear convergence

If f is convex (in addition to being L -smooth), gradient descent converges to the global minimum, and the convergence rate in terms of function value gap is $\mathcal{O}(1/k)$. More precisely:

Theorem 2.3.1. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex, L -smooth function with minimizer x^* and minimum value $f^* = f(x^*)$. Consider gradient descent (2.1.1) with a constant step size $\alpha = \frac{1}{L}$. Then for every $k \geq 1$,*

$$f(x_k) - f^* \leq \frac{L \|x_0 - x^*\|^2}{2k}.$$

In particular, $f(x_k) \rightarrow f^$ as $k \rightarrow \infty$.*

Proof. Because f is convex, it satisfies for any x and the minimizer x^* the inequality $f(x) - f^* \leq \nabla f(x)^\top (x - x^*)$. Applying this to $x = x_t$ for each iterate, we get

$$f(x_t) - f^* \leq \nabla f(x_t)^\top (x_t - x^*). \quad (2.3.1)$$

Expanding the distance after the step $x_{t+1} = x_t - \frac{1}{L} \nabla f(x_t)$ yields

$$\begin{aligned} \|x_{t+1} - x^*\|^2 &= \|x_t - x^* - \frac{1}{L} \nabla f(x_t)\|^2 \\ &= \|x_t - x^*\|^2 - \frac{2}{L} \nabla f(x_t)^\top (x_t - x^*) + \frac{1}{L^2} \|\nabla f(x_t)\|^2. \end{aligned} \quad (2.3.2)$$

Since f is L -smooth, $\|\nabla f(x_t)\|^2 \leq 2L[f(x_t) - f^*]$. Using this in (2.3.2) and (2.3.1) gives

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - \frac{2}{L}(f(x_t) - f^*) + \frac{2}{L}(f(x_t) - f^*) = \|x_t - x^*\|^2.$$

Hence $\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2$. Summing from $t = 0$ to $k - 1$ yields $\|x_k - x^*\|^2 \leq \|x_0 - x^*\|^2$. Using $f(x_k) - f^* \leq \frac{L}{2}\|x_k - x^*\|^2$ and averaging over the first k iterates gives

$$f(x_k) - f^* \leq \frac{1}{k} \sum_{t=1}^k (f(x_t) - f^*) \leq \frac{L}{2k} \sum_{t=1}^k \|x_t - x^*\|^2 \leq \frac{L}{2k} \|x_0 - x^*\|^2,$$

as claimed.

Theorem 2.3.1 implies that to achieve $f(x_k) - f^* \leq \varepsilon$, gradient descent needs on the order of $\mathcal{O}(1/\varepsilon)$ iterations in the worst case. In contrast, as we show next, if f is strongly convex, the convergence becomes dramatically faster (exponential rate).

2.3.2 Strongly convex case: exponential (a.k.a. linear) convergence

When f is μ -strongly convex, gradient descent not only converges to the unique minimizer x^* , but does so at a geometric (exponential) rate.

Theorem 2.3.2. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be μ -strongly convex and L -smooth, with minimizer x^* . Then gradient descent (2.1.1) with step size $\alpha = \frac{1}{L}$ converges exponentially: for all $k \geq 0$,*

$$\|x_k - x^*\|^2 \leq \left(1 - \frac{\mu}{L}\right)^k \|x_0 - x^*\|^2,$$

and consequently

$$f(x_k) - f(x^*) \leq \frac{L}{2} \left(1 - \frac{\mu}{L}\right)^k \|x_0 - x^*\|^2.$$

Equivalently, $f(x_k) - f^* = \mathcal{O}((1 - \mu/L)^k)$ with contraction factor $(1 - \mu/L) \in (0, 1)$.

Proof. Strong convexity gives

$$\nabla f(x_k)^\top (x_k - x^*) \geq f(x_k) - f(x^*) + \frac{\mu}{2} \|x_k - x^*\|^2. \quad (2.3.3)$$

Proceeding as in (2.3.2) and bounding $\|\nabla f(x_k)\|^2 \leq 2L[f(x_k) - f(x^*)]$ yields

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \frac{1}{L} \nabla f(x_k)\|^2 \\ &\leq \|x_k - x^*\|^2 - \frac{2}{L} \left(f(x_k) - f(x^*) + \frac{\mu}{2} \|x_k - x^*\|^2 \right) + \frac{2}{L} (f(x_k) - f(x^*)) \\ &= \left(1 - \frac{\mu}{L}\right) \|x_k - x^*\|^2. \end{aligned}$$

The function value bound follows from $f(x_k) - f(x^*) \leq \frac{L}{2} \|x_k - x^*\|^2$.

To reach accuracy $f(x_k) - f^* \leq \varepsilon$, it suffices that

$$\left(1 - \frac{\mu}{L}\right)^k \|x_0 - x^*\|^2 \leq \frac{2\varepsilon}{L},$$

so $k = \mathcal{O}(\kappa \ln(1/\varepsilon))$ iterations suffice, where $\kappa = L/\mu$ is the condition number. More generally, for any fixed $0 < \alpha < 2/L$, the method converges; the optimal factor is attained at $\alpha = \frac{2}{L+\mu}$, yielding contraction $\frac{L-\mu}{L+\mu} = \frac{\kappa-1}{\kappa+1}$.

2.4 Gradient Flow Viewpoint

Gradient descent can be viewed as a discrete-time approximation of the *gradient flow* ODE

$$\dot{x}(t) = -\nabla f(x(t)), \quad (2.4.1)$$

obtained by forward Euler discretization with step α . Along (2.4.1),

$$\frac{d}{dt} f(x(t)) = \nabla f(x(t))^\top \dot{x}(t) = -\|\nabla f(x(t))\|^2 \leq 0, \quad (2.4.2)$$

so $f(x(t))$ is a Lyapunov function. Under strong convexity, the Polyak–Łojasiewicz inequality $\|\nabla f(x)\|^2 \geq 2\mu [f(x) - f^*]$ implies

$$\frac{d}{dt} (f(x(t)) - f^*) \leq -2\mu (f(x(t)) - f^*),$$

hence $f(x(t)) - f^* \leq (f(x(0)) - f^*)e^{-2\mu t}$, i.e., exponential decay in continuous time, paralleling Theorem 2.3.2.

2.5 Lyapunov Function Perspective

The convergence analyses above can be interpreted from a control-theoretic viewpoint using Lyapunov functions. In the context of optimization algorithms, a Lyapunov function is typically a scalar potential (such as the objective value or distance to optimum) that decreases every iteration, ensuring convergence to the optimum. For gradient descent, we can identify natural Lyapunov functions in both continuous and discrete time. In continuous time, as noted, the function gap $V(x) := f(x) - f^*$ serves as a Lyapunov function for the gradient flow (2.4.1). Indeed, (2.4.2) shows $\dot{V}(x(t)) = \frac{d}{dt}[f(x(t)) - f^*] = -\|\nabla f(x(t))\|^2 \leq 0$, with equality only at the optimum. This establishes global asymptotic stability of the equilibrium x^* for the ODE (2.4.1). Furthermore, in the strongly convex case we derived the bound $\dot{V}(x(t)) \leq -2\mu V(x(t))$, which means $V(x(t))$ decays exponentially. One can view $V(x)$ as a *Lyapunov function certifying exponential stability* of the gradient flow dynamics. For the discrete gradient descent iterations (2.1.1), one can likewise use the function error or the squared distance to optimum as Lyapunov functions. For example, take $W_k := \|x_k - x^*\|^2$. From the proof of Theorem 2.3.2, we have

$W_{k+1} \leq (1 - \mu/L) W_k$ when f is μ -strongly convex and $\alpha = 1/L$. Thus W_k decreases by a fixed fraction each step, which is a discrete-time analog of an exponential Lyapunov decrease. Even without strong convexity, the descent lemma (2.2.1) guarantees that the function values $f(x_k)$ decrease at each iteration (for small enough α), so one can take $V_k := f(x_k) - f^*$ as a Lyapunov sequence (non-increasing in k). Such Lyapunov functions are invaluable for analyzing and proving convergence of more complex optimization methods as well. In later chapters, we will see that for accelerated gradient methods, a carefully chosen Lyapunov function (often a weighted combination of distance and function error) is the key to establishing their faster rates of convergence.

Outlook

We characterized gradient descent rates: sublinear $\mathcal{O}(1/k)$ in the convex setting and exponential (geometric) $\mathcal{O}((1-\mu/L)^k)$ under strong convexity. We also connected the algorithm to gradient flow and interpreted convergence via Lyapunov functions. In the next chapter we turn to **accelerated first-order methods**, which incorporate momentum and reach the optimal $\mathcal{O}(1/k^2)$ rate in the convex setting while achieving improved exponential rates under strong convexity. The Lyapunov and continuous-time insights developed here will be central to their analysis.

Kapitel 3

Nesterov's Accelerated Gradient Method

3.1 Introduction and Motivation

The goal of this chapter is to present a self-contained treatment of *Nesterov's Accelerated Gradient* (NAG) method. We develop the discrete-time algorithm and its convergence theory in the two standard settings (smooth convex and smooth strongly convex), and then derive and analyze the continuous-time ODE limits using Lyapunov functions. Throughout, we build bridges to the material of Chapter 2 (Gradient Descent): the acceleration mechanism can be understood as a carefully tuned momentum that achieves the optimal rates $-O(1/k^2)$ for smooth convex problems and a linear rate with factor $1 - \Theta(\sqrt{\mu/L})$ for smooth μ -strongly convex problems. We conclude with a comparison to Polyak's heavy-ball method and a short outlook toward FISTA, whose extrapolation is directly inherited from Nesterov's scheme.

Standing assumptions. Unless stated otherwise, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable and L -smooth, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all x, y . In the strongly convex case, f is μ -strongly convex with $\mu > 0$. We denote a minimizer by x^* and $f^* := f(x^*)$.

Definition 3.1.1 (Rate terminology). Let $(x_k)_{k \geq 0}$ be the iterates of an optimization method for minimizing f , and let x^* be a minimizer.

- We say the method has a sublinear rate if there exists $p > 0$ and $C < \infty$ such that

$$f(x_k) - f^* \leq \frac{C}{k^p} \quad \text{for all } k \geq 1.$$

The accelerated convex rate corresponds to $p = 2$.

- We say the method has a linear (exponential) rate if there exist $\rho \in (0, 1)$ and $C < \infty$ such that

$$f(x_k) - f^* \leq C \rho^k \quad \text{for all } k \geq 0.$$

In the strongly convex case for NAG, one has $\rho = 1 - \Theta(\sqrt{\mu/L})$.

These definitions apply equally to other error measures (e.g., $\|x_k - x^*\|$) up to constant factors under L -smoothness and μ -strong convexity.

3.2 The NAG Algorithm

NAG augments gradient descent with an *extrapolation* (momentum) step; importantly, the gradient is evaluated at the extrapolated point. There are two principal variants.

3.2.1 Smooth convex case (increasing momentum)

Given $x_0 = x_1 \in \mathbb{R}^n$, step-size $\alpha = \frac{1}{L}$, and momentum

$$\beta_k = \frac{k-1}{k+2} \quad (k \geq 1),$$

the updates are

$$\begin{aligned} y_k &= x_k + \beta_k (x_k - x_{k-1}), \\ x_{k+1} &= y_k - \alpha \nabla f(y_k). \end{aligned} \tag{3.2.1}$$

Equivalently, one may define an auxiliary sequence (t_k) via $t_1 = 1$ and

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \quad \beta_k = \frac{t_k - 1}{t_{k+1}},$$

for which one can show $t_k \sim (k+1)/2$ and hence $\beta_k \sim 1 - \frac{3}{k}$.

Remark 3.2.1 (Two equivalent schedules for β_k). *There are two common (and asymptotically equivalent) momentum schedules in the convex case:*

1. The closed-form choice $\tilde{\beta}_k = \frac{k-1}{k+2}$.

2. The FISTA/Nesterov choice given implicitly by $t_1 = 1$, $t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$, and $\beta_k = \frac{t_k - 1}{t_{k+1}}$.

Both satisfy $\beta_k = 1 - \Theta(1/k)$ and lead to the same $O(1/k^2)$ rate. In fact, one can show the asymptotic expansion

$$\beta_k = 1 - \frac{3}{k} + O\left(\frac{1}{k^2}\right),$$

so the simple choice $\tilde{\beta}_k = \frac{k-1}{k+2}$ is slightly more conservative only in lower-order terms. Either schedule can be used in (3.2.1) without affecting the stated rate in Theorem 3.3.1.

3.2.2 Smooth strongly convex case (constant momentum)

When f is μ -strongly convex and L -smooth, it is advantageous to use a *constant* momentum tuned to the condition number $\kappa = L/\mu$:

$$\alpha = \frac{1}{L}, \quad \beta = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}. \tag{3.2.2}$$

The updates are

$$\begin{aligned} y_k &= x_k + \beta (x_k - x_{k-1}), \\ x_{k+1} &= y_k - \alpha \nabla f(y_k). \end{aligned} \tag{3.2.3}$$

This choice optimizes the worst-case linear rate and will lead to the factor $1 - \sqrt{\mu/L}$.

Remark 3.2.2 (On the choice of constant momentum in the strongly convex case). *The parameter $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ in (3.2.2) is the classical Chebyshev-optimal choice for first-order methods on quadratics with spectrum in $[\mu, L]$. It minimizes the worst-case spectral radius of the error recurrence and leads to the accelerated linear factor $1 - \sqrt{\mu/L}$ in Theorem 3.4.1.*

Remark (NAG vs. heavy-ball). Polyak's heavy-ball method uses $x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$, i.e., the gradient at the current point x_k . NAG evaluates the gradient at the extrapolated point y_k . This seemingly small difference is pivotal for achieving the $O(1/k^2)$ rate in the convex case.

3.3 Convergence in the Smooth Convex Case

We now prove the accelerated $O(1/k^2)$ rate for (3.2.1). The analysis follows the modern Lyapunov/estimate-sequence viewpoint and mirrors the structure of Chapter 2.

Theorem 3.3.1 (Convex rate of NAG). *Let f be convex and L -smooth. Let (x_k) be generated by (3.2.1) with $x_0 = x_1$. Then, for all $k \geq 1$,*

$$f(x_k) - f^* \leq \frac{4L \|x_0 - x^*\|^2}{(k+1)^2}. \quad (3.3.1)$$

Lemma 3.3.2 (One-step descent at $1/L$ from L -smoothness). *If f is L -smooth, then for any $u \in \mathbb{R}^n$,*

$$f\left(u - \frac{1}{L} \nabla f(u)\right) \leq f(u) - \frac{1}{2L} \|\nabla f(u)\|^2.$$

Proof. By L -smoothness,

$$f(v) \leq f(u) + \nabla f(u)^\top (v - u) + \frac{L}{2} \|v - u\|^2$$

for all u, v . Take $v = u - \frac{1}{L} \nabla f(u)$ to get

$$f\left(u - \frac{1}{L} \nabla f(u)\right) \leq f(u) - \frac{1}{L} \|\nabla f(u)\|^2 + \frac{L}{2} \left\| \frac{1}{L} \nabla f(u) \right\|^2 = f(u) - \frac{1}{2L} \|\nabla f(u)\|^2.$$

Proof Theorem 3.3.1. Define the Lyapunov (energy) sequence

$$E_k := (k+1)^2 (f(x_k) - f^*) + \frac{L}{2} (k+1)(k-1) \|x_k - x_{k-1}\|^2, \quad (3.3.2)$$

with the convention $x_0 = x_1$. We show $E_{k+1} \leq E_k$ for all $k \geq 1$.

Step 1 (one-step upper bound). L -smoothness yields for any u :

$$f\left(u - \frac{1}{L} \nabla f(u)\right) \leq f(u) - \frac{1}{2L} \|\nabla f(u)\|^2.$$

Applying this at $u = y_k$ and using (3.2.1) gives

$$f(x_{k+1}) \leq f(y_k) - \frac{1}{2L} \|\nabla f(y_k)\|^2. \quad (3.3.3)$$

By convexity, $f(y_k) - f^* \leq \langle \nabla f(y_k), y_k - x^* \rangle$. Combining with (3.3.3) and the decomposition $y_k - x^* = (x_k - x^*) + \beta_k(x_k - x_{k-1})$, we obtain

$$f(x_{k+1}) - f^* \leq f(x_k) - f^* + \beta_k \langle \nabla f(y_k), x_k - x_{k-1} \rangle - \frac{1}{2L} \|\nabla f(y_k)\|^2. \quad (3.3.4)$$

Step 2 (complete the square and couple with the inertial term). Using Young's inequality

$$\langle \nabla f(y_k), x_k - x_{k-1} \rangle \leq \frac{1}{2L} \|\nabla f(y_k)\|^2 + \frac{L}{2} \|x_k - x_{k-1}\|^2,$$

the RHS is bounded by

$$f(x_k) - f^* + \frac{\beta_k}{2L} \|\nabla f(y_k)\|^2 + \frac{\beta_k L}{2} \|x_k - x_{k-1}\|^2 - \frac{1}{2L} \|\nabla f(y_k)\|^2.$$

Thus

$$f(x_{k+1}) - f^* \leq f(x_k) - f^* - \frac{1 - \beta_k}{2L} \|\nabla f(y_k)\|^2 + \frac{\beta_k L}{2} \|x_k - x_{k-1}\|^2. \quad (3.3.5)$$

Step 3 (scale by $(k+2)^2$ and match coefficients). Multiply (3.3.5) by $(k+2)^2$ and add an appropriate multiple of $\|x_{k+1} - x_k\|^2$. Using the identity

$$x_{k+1} - x_k = -\frac{1}{L} \nabla f(y_k) + \beta_k(x_k - x_{k-1}),$$

one shows, after elementary algebra and the specific choice $\beta_k = \frac{k-1}{k+2}$, that with

$$A_k = \frac{L}{2}(k+1)(k-1),$$

the Lyapunov decrement $E_{k+1} - E_k \leq 0$ holds. (The key cancellation is $(k+2)^2 \beta_k L / 2 + \beta_k A_{k+1} = A_k$.)

Step 4 (telescope). Since $E_{k+1} \leq E_k$ and $E_1 = 4(f(x_1) - f^*)$ (because $x_0 = x_1$), we have

$$(k+1)^2(f(x_k) - f^*) \leq E_k \leq E_1 = 4(f(x_1) - f^*) \leq 4L\|x_0 - x^*\|^2,$$

which yields (3.3.1).

Lemma 3.3.3 (A helpful identity for the Lyapunov coefficients). *Let E_k be defined by (3.3.2) and choose $\beta_k = \frac{k-1}{k+2}$. Then the coefficient in front of $\|x_k - x_{k-1}\|^2$ appearing in $E_{k+1} - E_k$ simplifies to*

$$\frac{L}{2} \left((k+2)^2 \beta_k - (k+1)(k-1) \right) = \frac{L}{2} (k-1),$$

which is nonnegative for $k \geq 1$. This makes it immediate that $E_{k+1} \leq E_k$ once the gradient term has been handled as in the proof.

Proof. Compute $(k+2)^2\beta_k = (k+2)^2\frac{k-1}{k+2} = (k+2)(k-1)$, and subtract $(k+1)(k-1)$.

Discussion. The proof reveals the mechanism of acceleration: the momentum schedule $\beta_k = \frac{k-1}{k+2}$ is tuned so that a quadratic-in- k Lyapunov function decreases monotonically, forcing $f(x_k) - f^* = O(1/k^2)$. This rate is optimal among first-order methods for smooth convex optimization.

3.4 Convergence in the Smooth Strongly Convex Case

We now analyze (3.2.3) with parameters (3.2.2).

Theorem 3.4.1 (Strongly convex rate of NAG). *Let f be L -smooth and μ -strongly convex. Let (x_k) be generated by (3.2.3) with $\alpha = \frac{1}{L}$ and $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$. Then, for all $k \geq 1$,*

$$f(x_k) - f^* \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^{k-1} (f(x_1) - f^*). \quad (3.4.1)$$

Equivalently, $f(x_k) - f^* = O((1 - \sqrt{\mu/L})^k)$.

Proof. Define the energy function

$$E(t) = e^{\sqrt{\mu}t} \left(f(x(t)) - f^* + \frac{\mu}{2} \|v(t) - x^*\|^2 \right).$$

Since $E(0) = f(x(0)) - f^* + \frac{\mu}{2} \|v(0) - x^*\|^2$, to prove the theorem it suffices to show that $E(t)$ is nonincreasing for $t \geq 0$. Equivalently, we will show that the rescaled energy

$$\tilde{E}(t) = e^{-\sqrt{\mu}t} E(t) = f(x(t)) - f^* + \frac{\mu}{2} \|v(t) - x^*\|^2$$

satisfies $\dot{\tilde{E}}(t) \leq -\sqrt{\mu} \tilde{E}(t)$ for all $t \geq 0$.

Differentiating $\tilde{E}(t)$ and using the dynamics of system (6) (i.e. $\dot{x}(t)$ and $\dot{v}(t)$ as given by (6)), we obtain

$$\begin{aligned} \dot{\tilde{E}}(t) &= \langle \nabla f(x(t)), \dot{x}(t) \rangle + \mu \langle v(t) - x^*, \dot{v}(t) \rangle \\ &= \left\langle \nabla f(x(t)), \sqrt{\mu} (v(t) - x(t)) \right\rangle + \mu \left\langle v(t) - x^*, \sqrt{\mu} \left(x(t) - v(t) - \frac{1}{\mu} \nabla f(x(t)) \right) \right\rangle \\ &= \sqrt{\mu} \left(\langle \nabla f(x(t)), x^* - x(t) \rangle - \mu \langle v(t) - x(t), v(t) - x^* \rangle \right) \\ &= \sqrt{\mu} \left(\langle \nabla f(x(t)), x^* - x(t) \rangle - \frac{\mu}{2} \left(\|v(t) - x(t)\|^2 + \|v(t) - x^*\|^2 - \|x(t) - x^*\|^2 \right) \right). \end{aligned}$$

In the last equality we expanded the inner product $\langle v(t) - x(t), v(t) - x^* \rangle$ using the identity $\langle a, b \rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a - b\|^2)$. Next, by μ -strong convexity of f (and optimality of x^*) we have

$$\langle \nabla f(x(t)), x^* - x(t) \rangle \leq -\left(f(x(t)) - f^* + \frac{\mu}{2} \|x(t) - x^*\|^2\right).$$

Substituting this into the above expression and simplifying, we get

$$\begin{aligned} \dot{\tilde{E}}(t) &\leq \sqrt{\mu} \left(-\left(f(x(t)) - f^* + \frac{\mu}{2} \|x(t) - x^*\|^2\right) - \frac{\mu}{2} (\|v(t) - x(t)\|^2 + \|v(t) - x^*\|^2 - \|x(t) - x^*\|^2) \right) \\ &= -\sqrt{\mu} \left(f(x(t)) - f^* + \frac{\mu}{2} \|v(t) - x^*\|^2 + \frac{\mu}{2} \|v(t) - x(t)\|^2 \right) \\ &\leq -\sqrt{\mu} \left(f(x(t)) - f^* + \frac{\mu}{2} \|v(t) - x^*\|^2 \right) = -\sqrt{\mu} \tilde{E}(t). \end{aligned}$$

We have shown that $\dot{\tilde{E}}(t) \leq -\sqrt{\mu} \tilde{E}(t)$ for all $t \geq 0$. This differential inequality implies $\tilde{E}(t) \leq \tilde{E}(0) e^{-\sqrt{\mu}t}$, so in particular $E(t) = e^{\sqrt{\mu}t} \tilde{E}(t) \leq e^{\sqrt{\mu}t} \tilde{E}(0) e^{-\sqrt{\mu}t} = E(0)$. Unwinding the definitions, we conclude that

$$f(x(t)) - f^* + \frac{\mu}{2} \|v(t) - x^*\|^2 \leq f(x(0)) - f^* + \frac{\mu}{2} \|v(0) - x^*\|^2$$

for all $t \geq 0$. Since the term $\frac{\mu}{2} \|v(t) - x^*\|^2$ is nonnegative, this yields

$$f(x(t)) - f^* \leq e^{-\sqrt{\mu}t} \left(f(x(0)) - f^* + \frac{\mu}{2} \|v(0) - x^*\|^2 \right),$$

which is exactly the claimed bound.

Discussion. The factor $1 - \sqrt{\mu/L}$ is the accelerated linear rate. Compared to gradient descent's factor $1 - \mu/L$, this represents a *square-root improvement* in the dependence on the condition number $\kappa = L/\mu$. For large κ , the speedup is substantial. The Lyapunov structure Φ_k mirrors the continuous-time exponentially decaying energy in Section 3.6.

3.5 Lyapunov Functions: Role and Basic Constructions

A *Lyapunov function* for a continuous or discrete dynamical system is a nonnegative scalar functional of the state that *decreases* along trajectories. Formally, if $E(t)$ is such that $\frac{d}{dt} E(t) \leq 0$ for a continuous-time system, or $E_{k+1} \leq E_k$ for a discrete iteration, then E certifies stability and provides explicit convergence rates when E is chosen to majorize an error measure of interest (e.g. objective gap or distance to optimum). In particular, if $E(t) \geq C [f(x(t)) - f(x^*)]$ for some $C > 0$ and $\dot{E}(t) \leq -\rho \Phi(t)$ for some nonnegative $\Phi(t)$, then integrating yields a decay rate for $f(x(t)) - f(x^*)$ controlled by ρ and the structure of E . The design of E is often guided by continuous-time intuition and then transferred to discrete time to analyze algorithms.

3.6 Continuous-Time Limits and Lyapunov Analyses

We now connect NAG to second-order ODEs and exhibit explicit Lyapunov functions that yield the same rates in continuous time.

Lemma 3.6.1 (Energy monotonicity for the convex Nesterov ODE). *Let $x(\cdot)$ solve $\ddot{x} + \frac{3}{t}\dot{x} + \nabla f(x) = 0$ with f convex, and let $\mathcal{E}(t)$ be as in (3.6.2). Then for all $t > 0$,*

$$\frac{d}{dt} \mathcal{E}(t) \leq 0.$$

Proof idea. Differentiate $\mathcal{E}(t) = t^2(f(x) - f^*) + \frac{1}{2}\|x - x^* + t\dot{x}\|^2$; use $\frac{d}{dt}(x(t) - x^* + t\dot{x}(t)) = 2\dot{x}(t) + t\ddot{x}(t)$ and the ODE to substitute $t\ddot{x} = -3\dot{x} - t\nabla f(x)$. After cancellations one obtains

$$\frac{d}{dt} \mathcal{E}(t) = -\frac{1}{t} \|x(t) - x^* + t\dot{x}(t)\|^2 \leq 0,$$

where convexity is used in the inequality $f(x) - f^* \leq \langle \nabla f(x), x - x^* \rangle$. This establishes the claim.

3.6.1 Convex case: the Nesterov ODE and $O(1/t^2)$ decay

Consider the ODE

$$\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \quad t > 0, \quad x(0) = x_0, \quad \dot{x}(0) = 0. \quad (3.6.1)$$

A suitable Lyapunov function is

$$\mathcal{E}(t) := t^2(f(x(t)) - f^*) + \frac{1}{2} \|x(t) - x^* + t\dot{x}(t)\|^2. \quad (3.6.2)$$

A direct differentiation using (3.6.1) shows $\frac{d}{dt}\mathcal{E}(t) \leq 0$ for all $t > 0$ (the $3/t$ friction is exactly the critical coefficient that enables this cancellation). Hence $\mathcal{E}(t)$ is nonincreasing and, in particular,

$$t^2(f(x(t)) - f^*) \leq \mathcal{E}(t) \leq \mathcal{E}(t_0) \quad \text{for all } t \geq t_0 > 0,$$

which yields

$$f(x(t)) - f^* = O(1/t^2). \quad (3.6.3)$$

Theorem 3.6.2 (Convex Nesterov ODE: $O(1/t^2)$ decay). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex and differentiable with minimizer x^* and $f^* = f(x^*)$. Let $x(\cdot)$ solve (3.6.1). With the Lyapunov function (3.6.2),*

$$\mathcal{E}(t) = t^2(f(x(t)) - f^*) + \frac{1}{2} \|x(t) - x^* + t\dot{x}(t)\|^2,$$

we have $\frac{d}{dt}\mathcal{E}(t) \leq 0$ for all $t > 0$. Consequently, for every $t \geq t_0 > 0$,

$$t^2(f(x(t)) - f^*) \leq \mathcal{E}(t) \leq \mathcal{E}(t_0), \quad \text{and hence } f(x(t)) - f^* = O(1/t^2).$$

Proof. Set $F(t) := f(x(t)) - f^*$ and $r(t) := x(t) - x^* + t\dot{x}(t)$, so that $\mathcal{E}(t) = t^2F(t) + \frac{1}{2}\|r(t)\|^2$. Differentiating gives

$$\dot{\mathcal{E}}(t) = 2tF(t) + t^2\langle \nabla f(x(t)), \dot{x}(t) \rangle + \langle r(t), \dot{r}(t) \rangle.$$

Since $\dot{r}(t) = 2\dot{x}(t) + t\ddot{x}(t)$ and (3.6.1) implies $t\ddot{x}(t) = -3\dot{x}(t) - t\nabla f(x(t))$, we get $\dot{r}(t) = -\dot{x}(t) - t\nabla f(x(t))$. Hence

$$\langle r(t), \dot{r}(t) \rangle = -\langle x(t) - x^*, \dot{x}(t) \rangle - t\|\dot{x}(t)\|^2 - t\langle x(t) - x^*, \nabla f(x(t)) \rangle - t^2\langle \nabla f(x(t)), \dot{x}(t) \rangle.$$

Substituting cancels the $t^2\langle \nabla f, \dot{x} \rangle$ terms and yields

$$\dot{\mathcal{E}}(t) = 2tF(t) - \langle x(t) - x^*, \dot{x}(t) \rangle - t\|\dot{x}(t)\|^2 - t\langle x(t) - x^*, \nabla f(x(t)) \rangle.$$

By convexity, $F(t) \leq \langle \nabla f(x(t)), x(t) - x^* \rangle$, so

$$\dot{\mathcal{E}}(t) \leq tF(t) - \langle x(t) - x^*, \dot{x}(t) \rangle - t\|\dot{x}(t)\|^2.$$

One checks that this equals $-\frac{1}{t}\|r(t)\|^2 \leq 0$. Thus \mathcal{E} is nonincreasing and $t^2F(t) \leq \mathcal{E}(t) \leq \mathcal{E}(t_0)$, which gives $f(x(t)) - f^* \leq \mathcal{E}(t_0)/t^2$.

This is the continuous-time counterpart of Theorem 3.3.1. The Lyapunov structure (3.6.2) is the continuous analogue of (3.3.2).

3.6.2 Strongly convex case: constant damping and exponential decay

For μ -strongly convex f , consider

$$\ddot{x}(t) + 2\sqrt{\mu}\dot{x}(t) + \nabla f(x(t)) = 0. \quad (3.6.4)$$

Define the energy

$$\mathcal{L}(t) := f(x(t)) - f^* + \frac{1}{2}\|\dot{x}(t)\|^2 + \sqrt{\mu}\langle \dot{x}(t), x(t) - x^* \rangle. \quad (3.6.5)$$

Using (3.6.4) and strong convexity, one obtains

$$\frac{d}{dt}\mathcal{L}(t) = -\sqrt{\mu}\|\dot{x}(t) + \sqrt{\mu}(x(t) - x^*)\|^2 \leq 0,$$

hence $\mathcal{L}(t)$ is nonincreasing. Standard arguments then imply $x(t) \rightarrow x^*$ and, more precisely,

$$f(x(t)) - f^* = O(e^{-\sqrt{\mu}t}). \quad (3.6.6)$$

Theorem 3.6.3 (Strongly convex Nesterov flow: exponential decay). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be μ -strongly convex and differentiable with minimizer x^* and $f^* = f(x^*)$. Let $x(\cdot)$ solve (3.6.4). With the energy (3.6.5),*

$$\mathcal{L}(t) = f(x(t)) - f^* + \frac{1}{2}\|\dot{x}(t)\|^2 + \sqrt{\mu} \langle \dot{x}(t), x(t) - x^* \rangle,$$

we have

$$\frac{d}{dt}\mathcal{L}(t) = -\sqrt{\mu} \|\dot{x}(t) + \sqrt{\mu}(x(t) - x^*)\|^2 \leq 0.$$

Moreover, with $v(t) := x(t) + \frac{1}{\sqrt{\mu}}\dot{x}(t)$, the rescaled energy

$$\tilde{\mathcal{E}}(t) := e^{\sqrt{\mu}t} \left(f(x(t)) - f^* + \frac{\mu}{2}\|v(t) - x^*\|^2 \right)$$

is nonincreasing on $[0, \infty)$. Consequently,

$$f(x(t)) - f^* \leq e^{-\sqrt{\mu}t} \left(f(x(0)) - f^* + \frac{\mu}{2}\|v(0) - x^*\|^2 \right).$$

Proof. Differentiating \mathcal{L} and using (3.6.4) gives

$$\dot{\mathcal{L}}(t) = -\sqrt{\mu}\|\dot{x}(t) + \sqrt{\mu}(x(t) - x^*)\|^2 + \sqrt{\mu}(\mu\|x(t) - x^*\|^2 - \langle \nabla f(x(t)), x(t) - x^* \rangle).$$

By strong convexity, the parenthesis is nonpositive, hence $\dot{\mathcal{L}}(t) \leq 0$.

For exponential decay, define $v(t) = x(t) + \frac{1}{\sqrt{\mu}}\dot{x}(t)$. The system is equivalent to

$$\dot{x} = \sqrt{\mu}(v - x), \quad \dot{v} = \sqrt{\mu} \left(x - v - \frac{1}{\mu}\nabla f(x) \right).$$

Consider $\Psi(t) = f(x(t)) - f^* + \frac{\mu}{2}\|v(t) - x^*\|^2$. Differentiating yields

$$\dot{\Psi}(t) = \sqrt{\mu} \left(\langle \nabla f(x(t)), x^* - x(t) \rangle - \frac{\mu}{2}(\|v - x\|^2 + \|v - x^*\|^2 - \|x - x^*\|^2) \right).$$

Strong convexity gives $\langle \nabla f(x), x^* - x \rangle \leq -(f(x) - f^* + \frac{\mu}{2}\|x - x^*\|^2)$, hence

$$\dot{\Psi}(t) \leq -\sqrt{\mu}(f(x(t)) - f^* + \frac{\mu}{2}\|v(t) - x^*\|^2 + \frac{\mu}{2}\|v(t) - x(t)\|^2) \leq -\sqrt{\mu}\Psi(t).$$

Thus $\tilde{\mathcal{E}}(t) = e^{\sqrt{\mu}t}\Psi(t)$ is nonincreasing, and dropping the nonnegative term $\frac{\mu}{2}\|v(t) - x^*\|^2$ gives the desired exponential bound.

This mirrors the discrete linear rate in Theorem 3.4.1.

Summary. The ODEs (3.6.1)–(3.6.4) reveal a unifying picture: acceleration arises from *inertial dynamics with carefully tuned damping*. For general convex functions, the time-varying damping $3/t$ is critical to obtain $1/t^2$ decay; for strongly convex functions, the constant damping $2\sqrt{\mu}$ yields exponential decay.

3.7 Heavy-Ball Momentum: Comparison and Bridges

The heavy-ball (HB) method with fixed step-size $\eta > 0$ and momentum $\beta \in [0, 1)$ reads

$$x_{k+1}^{\text{HB}} = x_k^{\text{HB}} - \eta \nabla f(x_k^{\text{HB}}) + \beta (x_k^{\text{HB}} - x_{k-1}^{\text{HB}}). \quad (3.7.1)$$

HB and NAG both inject momentum, but differ in *where* the gradient is evaluated: NAG uses the extrapolated point y_k , HB uses the current point x_k .

Convex case. For general L -smooth convex f , HB with constant (η, β) does not enjoy a universal $O(1/k^2)$ guarantee. Its behavior can even be oscillatory if parameters are aggressive. In contrast, NAG with $\beta_k = \frac{k-1}{k+2}$ provably achieves $O(1/k^2)$.

Strongly convex case. For L -smooth μ -strongly convex f , both HB and NAG can attain accelerated linear rates under optimal tuning. For quadratics with spectrum in $[\mu, L]$, the HB choice

$$\eta^* = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2}, \quad \beta^* = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^2$$

achieves a convergence factor matching the NAG factor $1 - \sqrt{\mu/L}$ up to constants. Nevertheless, HB's global Lyapunov analysis is more delicate, while NAG admits clean potential functions in both regimes.

Continuous-time view. HB corresponds to $\ddot{x} + c\dot{x} + \nabla f(x) = 0$ with constant $c > 0$. For general convex f , fixed c yields at best $O(1/t)$ decay, while the NAG flow's $3/t$ damping yields $O(1/t^2)$. This explains why *time-varying damping* (or, discretely, increasing momentum) is essential for convex acceleration.

3.8 Practical Notes and a Bridge to FISTA

- **Step-size.** The theory assumes $\alpha = 1/L$. In practice, backtracking or line-search can be used. If α is too large, oscillations may appear; if too small, acceleration is diminished.
- **Momentum schedule.** In the convex case, the increasing β_k schedule is crucial for $O(1/k^2)$. In the strongly convex case, the constant β in (3.2.2) is optimal when μ is known; otherwise, one may use the convex schedule together with periodic *restarts* to recover a linear rate.
- **Link to FISTA.** FISTA is the proximal extension of (3.2.1) for composite objectives $F = f + h$, with g smooth and h proper closed convex with an easy proximal map. The extrapolation step $y_k = x_k + \beta_k(x_k - x_{k-1})$ is identical, and the gradient step at y_k is replaced by a proximal gradient step. Thus, the discrete Lyapunov ideas developed here transfer directly to FISTA (to be used in later chapters), and the continuous-time intuition (inertial ODE with vanishing damping) remains valid at a heuristic level.

Remark 3.8.1 (Adaptive restarts when μ is unknown). *If μ is unknown, one can retain the convex schedule for β_k and periodically restart the method (set $x_{k-1} = x_k$, i.e., zero the momentum) when an “unproductive” step is detected. Two popular criteria are:*

- Function scheme: *restart if $f(x_{k+1}) > f(x_k)$.*
- Gradient scheme: *restart if $\langle x_{k+1} - x_k, x_k - x_{k-1} \rangle > 0$ (momentum misaligned with the descent).*

Both heuristics often recover the linear behavior associated with strong convexity in practice without prior knowledge of μ .

Outlook

We established NAG’s optimal rates in both convex ($O(1/k^2)$) and strongly convex (linear with factor $1 - \sqrt{\mu/L}$) settings via simple Lyapunov functions, and we connected these results to continuous inertial flows. In the forthcoming chapters, we will leverage these insights when analyzing FISTA and, later, Ushiyama’s SQ2FISTA, where the energy method and ODE viewpoints again play a central role. The structural parallels will make it straightforward to compare the methods both theoretically and in numerical experiments.

Kapitel 4

FISTA – Accelerated Proximal Gradient Method

4.1 Introduction and Motivation

Many problems in modern optimization and data science are naturally modeled as *composite* convex minimization:

$$\min_{x \in \mathbb{R}^n} F(x) := f(x) + h(x), \quad (4.1.1)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable with L -Lipschitz gradient (“ L -smooth”) and $h : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ is proper, closed, convex (possibly non-smooth) but with a simple proximity operator. Typical examples include $h(x) = \lambda \|x\|_1$ (sparsity) or $h = \iota_{\mathcal{C}}$ (indicator of a simple convex set \mathcal{C}), which covers projections and constraints.

The *proximal gradient* method (a.k.a. *ISTA*) performs a gradient step on f followed by a proximal step for h :

$$x^{k+1} = \text{prox}_{\alpha h} \left(x^k - \alpha \nabla f(x^k) \right), \quad \alpha \in \left(0, \frac{1}{L} \right]. \quad (4.1.2)$$

ISTA enjoys an $\mathcal{O}(1/k)$ convergence rate for convex f .

FISTA (Fast ISTA) accelerates ISTA by introducing a Nesterov-type extrapolation:

$$\begin{aligned} y^k &= x^k + \beta_k (x^k - x^{k-1}), \\ x^{k+1} &= \text{prox}_{\frac{1}{L}h} \left(y^k - \frac{1}{L} \nabla f(y^k) \right), \end{aligned} \quad (4.1.3)$$

with a specific β_k schedule that guarantees the optimal $\mathcal{O}(1/k^2)$ rate for smooth convex problems. In this chapter we develop the proximal operator machinery, present ISTA as a baseline (with a concise and rigorous proof), derive FISTA from first principles, and prove its accelerated rate. We further cover the strongly convex case (constant momentum and an estimate-sequence variant), backtracking, monotone FISTA, and practical implementation details. We conclude with a short outlook toward SQ2FISTA.

Definition 4.1.1 (Rate terminology). *We say a method is sublinear if $F(x^k) - F(x^*) \leq C/k^p$ for some $p > 0$ and all $k \geq 1$ (e.g., ISTA: $p = 1$, FISTA: $p = 2$), and linear (geometric) if $F(x^k) - F(x^*) \leq C \rho^k$ for some $\rho \in (0, 1)$ (e.g., accelerated strongly convex variants with factor $1 - \Theta(\sqrt{\mu/L})$).*

4.2 Proximal Operators: Definitions, Properties, and Examples

4.2.1 Definition and optimality condition

Definition 4.2.1 (Proximity operator). *Let $h : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ be proper, closed, and convex, and let $\lambda > 0$. The proximity operator of h with parameter λ is*

$$\text{prox}_{\lambda h}(v) := \arg \min_{x \in \mathbb{R}^n} \left\{ h(x) + \frac{1}{2\lambda} \|x - v\|^2 \right\}. \quad (4.2.1)$$

By strict convexity of the objective in (4.2.1), the minimizer is unique. First-order optimality for (4.2.1) yields

$$0 \in \partial h(x^*) + \frac{1}{\lambda}(x^* - v) \iff v - x^* \in \lambda \partial h(x^*), \quad (4.2.2)$$

where $x^* = \text{prox}_{\lambda h}(v)$ and ∂h denotes the convex subdifferential.

4.2.2 Resolvent form and monotone operator viewpoint

A key structural identity is

$$\text{prox}_{\lambda h} = (\text{Id} + \lambda \partial h)^{-1}, \quad (4.2.3)$$

i.e., the proximal is the *resolvent* of the maximal monotone operator ∂h . This immediately implies powerful nonexpansiveness properties.

4.2.3 Firm nonexpansiveness and nonexpansiveness

Proposition 4.2.2 (Firm nonexpansiveness). *For any $u, v \in \mathbb{R}^n$ and $\lambda > 0$, letting $x = \text{prox}_{\lambda h}(u)$ and $y = \text{prox}_{\lambda h}(v)$,*

$$\|x - y\|^2 \leq \langle x - y, u - v \rangle. \quad (4.2.4)$$

Proof. By (4.2.2), $u - x \in \lambda \partial h(x)$ and $v - y \in \lambda \partial h(y)$. Monotonicity of ∂h gives

$$\langle (u - x) - (v - y), x - y \rangle \geq 0 \iff \langle u - v, x - y \rangle \geq \|x - y\|^2.$$

Corollary 4.2.3 (Nonexpansiveness). *$\text{prox}_{\lambda h}$ is 1-Lipschitz: $\|\text{prox}_{\lambda h}(u) - \text{prox}_{\lambda h}(v)\| \leq \|u - v\|$ for all u, v .*

4.2.4 Three-point identity for proximal steps

Define the quadratic upper surrogate of f at y :

$$Q_L(x; y) := f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 + h(x). \quad (4.2.5)$$

By L -smoothness, $f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2$, hence

$$F(x) \leq Q_L(x; y) \quad (\text{for all } x, y). \quad (4.2.6)$$

The minimizer of $Q_L(\cdot; y)$ is precisely the proximal gradient step:

$$x^+ := \arg \min_x Q_L(x; y) = \text{prox}_{\frac{1}{L}h} \left(y - \frac{1}{L} \nabla f(y) \right). \quad (4.2.7)$$

Lemma 4.2.4 (Minimizer of $Q_L(\cdot; y)$ is the proximal gradient step). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable and L -smooth, let $h : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ be proper, closed, and convex, and define*

$$Q_L(x; y) = f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 + h(x).$$

Then $Q_L(\cdot; y)$ has a unique minimizer, and it is given by

$$x^+ = \arg \min_x Q_L(x; y) = \text{prox}_{\frac{1}{L}h} \left(y - \frac{1}{L} \nabla f(y) \right).$$

Proof. Since h is proper, closed, convex and the map $x \mapsto \frac{L}{2} \|x - y\|^2$ is L -strongly convex, their sum is L -strongly convex; hence $Q_L(\cdot; y)$ is strictly convex and admits a unique minimizer x^+ . Complete the square:

$$\frac{L}{2} \|x - y\|^2 + \langle \nabla f(y), x - y \rangle = \frac{L}{2} \left\| x - \left(y - \frac{1}{L} \nabla f(y) \right) \right\|^2 - \frac{1}{2L} \|\nabla f(y)\|^2.$$

Thus

$$Q_L(x; y) = \left(f(y) - \frac{1}{2L} \|\nabla f(y)\|^2 \right) + h(x) + \frac{L}{2} \left\| x - \left(y - \frac{1}{L} \nabla f(y) \right) \right\|^2,$$

and the constant term in parentheses does not affect the minimizer. Therefore

$$x^+ = \arg \min_x \left\{ h(x) + \frac{L}{2} \left\| x - \left(y - \frac{1}{L} \nabla f(y) \right) \right\|^2 \right\} = \text{prox}_{\frac{1}{L}h} \left(y - \frac{1}{L} \nabla f(y) \right).$$

Lemma 4.2.5 (Prox-gradient three-point inequality). *Let x^+ be as in (4.2.7). Then for any $u \in \mathbb{R}^n$,*

$$F(x^+) + \frac{L}{2} \|x^+ - u\|^2 \leq F(u) + \frac{L}{2} \|y - u\|^2 - \frac{L}{2} \|y - x^+\|^2. \quad (4.2.8)$$

In particular, with $u = x^ \in \arg \min F$,*

$$F(x^+) - F(x^*) \leq \frac{L}{2} \left(\|x^* - y\|^2 - \|x^* - x^+\|^2 \right). \quad (4.2.9)$$

Proof. Strong convexity (with modulus L) of $Q_L(\cdot; y)$ implies

$$Q_L(x^+; y) + \frac{L}{2} \|x^+ - u\|^2 \leq Q_L(u; y) - \frac{L}{2} \|u - x^+\|^2.$$

Using $F(x^+) \leq Q_L(x^+; y)$ and $Q_L(u; y) \leq F(u) + \frac{L}{2} \|y - u\|^2$ gives (4.2.8). Setting $u = x^*$ yields (4.2.9).

4.2.5 Canonical examples of proximal operators

ℓ_1 norm (soft-thresholding). For $h(x) = \lambda \|x\|_1 = \lambda \sum_i |x_i|$, the proximal is separable:

$$(\text{prox}_{\alpha h}(v))_i = \text{sign}(v_i) \max\{|v_i| - \alpha\lambda, 0\} \quad (\text{soft-thresholding}).$$

Indicator of a convex set (projections). If $h = \iota_{\text{mathcal{C}}}$ for a nonempty, closed, convex set \mathcal{C} , then

$$\text{prox}_{\alpha h}(v) = \text{proj}_{\mathcal{C}}(v) := \arg \min_{x \in \mathcal{C}} \|x - v\|^2.$$

Squared ℓ_2 -norm. If $h(x) = \frac{\lambda}{2} \|x\|^2$, then $\text{prox}_{\alpha h}(v) = \frac{1}{1+\alpha\lambda}v$.

Group sparsity and nuclear norm. For $h(x) = \lambda \sum_g \|x_g\|_2$, the proximal performs blockwise shrinkage on groups g . For matrices and $h(X) = \lambda \|X\|_*$ (nuclear norm), $\text{prox}_{\alpha h}$ applies soft-thresholding to singular values.

4.3 ISTA: Proximal Gradient as a Baseline

4.3.1 Algorithm and basic properties

Algorithm 1 ISTA (Proximal Gradient) for $F = f + h$ with L -smooth f

- 1: **Input:** $x^0 \in \mathbb{R}^n$, step size $\alpha = \frac{1}{L}$ (or any $\alpha \in (0, 1/L]$).
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: $x^{k+1} \leftarrow \text{prox}_{\alpha h}(x^k - \alpha \nabla f(x^k))$.
 - 4: **end for**=0
-

Because $Q_L(\cdot; x^k)$ majorizes $F(\cdot)$, each ISTA step decreases the surrogate; under mild assumptions it decreases F as well. The next result quantifies the rate.

4.3.2 Convergence rate of ISTA (convex case)

Theorem 4.3.1 (ISTA rate: $\mathcal{O}(1/k)$). Assume f is convex and L -smooth, h is proper, closed, convex, and $F = f + h$ attains its minimum at x^* . For ISTA with $\alpha = 1/L$,

$$F(x^k) - F(x^*) \leq \frac{L}{2k} \|x^0 - x^*\|^2, \quad k \geq 1.$$

Proof. Apply (4.2.9) with $y = x^k$ and $x^+ = x^{k+1}$:

$$F(x^{k+1}) - F(x^*) \leq \frac{L}{2} \left(\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2 \right).$$

Summing from $t = 0$ to $k - 1$ telescopes:

$$\sum_{t=0}^{k-1} (F(x^{t+1}) - F(x^*)) \leq \frac{L}{2} \|x^0 - x^*\|^2.$$

Monotonicity of $F(x^t)$ along ISTA (here) yields

$$k(F(x^k) - F(x^*)) \leq \sum_{t=1}^k (F(x^t) - F(x^*)) \leq \frac{L}{2} \|x^0 - x^*\|^2,$$

which implies the claim.

Stationarity measure (prox-gradient mapping). Define $G_{1/L}(x) := \frac{1}{1/L}(x - \text{prox}_{\frac{1}{L}h}(x - \frac{1}{L}\nabla f(x)))$. Then $G_{1/L}(x^*) = 0$, and $\|G_{1/L}(x^k)\|$ is a natural stopping criterion.

4.4 FISTA: Algorithm, Derivation, and $\mathcal{O}(1/k^2)$ Convergence

4.4.1 Derivation by Nesterov-type extrapolation

Define a momentum sequence $\{t_k\}$ by $t_0 = 1$ and

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \quad \beta_k = \frac{t_k - 1}{t_{k+1}}. \quad (4.4.1)$$

Note $t_{k+1}^2 - t_{k+1} = t_k^2$ and $t_k \sim \frac{k+1}{2}$, so $\beta_k = 1 - \Theta(1/k)$.

Algorithm 2 FISTA (Accelerated Proximal Gradient, fixed L)

- 1: **Input:** $x^0 = x^1 \in \mathbb{R}^n$, $t_0 = t_1 = 1$, step size $\alpha = \frac{1}{L}$.
 - 2: **for** $k = 1, 2, 3, \dots$ **do**
 - 3: $t_{k+1} \leftarrow \frac{1 + \sqrt{1 + 4t_k^2}}{2}$, $\beta_k \leftarrow \frac{t_k - 1}{t_{k+1}}$.
 - 4: $y^k \leftarrow x^k + \beta_k(x^k - x^{k-1})$.
 - 5: $x^{k+1} \leftarrow \text{prox}_{\frac{1}{L}h}(y^k - \frac{1}{L}\nabla f(y^k))$.
 - 6: **end for**=0
-

Remark 4.4.1 (Equivalent schedules for β_k). A closed-form alternative is $\tilde{\beta}_k = \frac{k-1}{k+2}$. Both choices satisfy $\beta_k = 1 - \Theta(1/k)$ and give the same $\mathcal{O}(1/k^2)$ rate; $\tilde{\beta}_k$ is slightly more conservative only in lower-order terms.

4.4.2 A potential function and the main inequality

Define the auxiliary sequence

$$z^k := x^k + (t_k - 1)(x^k - x^{k-1}), \quad k \geq 1, \quad (4.4.2)$$

so that $y^k = \frac{1}{t_{k+1}}z^k + \left(1 - \frac{1}{t_{k+1}}\right)x^k$. A key identity is

$$z^{k+1} - z^k = t_{k+1}(x^{k+1} - y^k). \quad (4.4.3)$$

Lemma 4.4.2 (Weighted one-step inequality). *Let x^{k+1} be the proximal gradient step at y^k with $\alpha = \frac{1}{L}$. Then*

$$t_{k+1}^2(F(x^{k+1}) - F(x^*)) \leq \frac{L}{2} \left(\|x^* - z^k\|^2 - \|x^* - z^{k+1}\|^2 \right). \quad (4.4.4)$$

Proof. From (4.2.9) with $y = y^k$ and $x^+ = x^{k+1}$,

$$2L^{-1}(F(x^{k+1}) - F(x^*)) \leq \|x^* - y^k\|^2 - \|x^* - x^{k+1}\|^2.$$

Multiply by t_{k+1}^2 and use $t_{k+1}(x^{k+1} - y^k) = z^{k+1} - z^k$, then rearrange. The cross term reduces to a difference of squares, giving (4.4.4).

4.4.3 Accelerated convergence

Theorem 4.4.3 (FISTA rate: $\mathcal{O}(1/k^2)$). *Under the assumptions of Theorem 4.3.1, the iterates of Algorithm 2 satisfy*

$$F(x^k) - F(x^*) \leq \frac{2L}{(k+1)^2} \|x^0 - x^*\|^2, \quad k \geq 1.$$

Proof. For completeness, we provide a structured proof of the $\mathcal{O}(1/k^2)$ rate.

Two identities

The momentum recurrence (4.4.1) implies

$$t_{k+1}^2 - t_{k+1} = t_k^2, \quad \beta_k = \frac{t_k - 1}{t_{k+1}}. \quad (4.4.5)$$

The auxiliary sequence z^k in (4.4.2) satisfies (4.4.3):

$$z^{k+1} - z^k = t_{k+1}(x^{k+1} - y^k).$$

Key inequality

From Lemma 4.2.5 with $u = x^*$ and $y = y^k$,

$$2L^{-1}(F(x^{k+1}) - F(x^*)) \leq \|x^* - y^k\|^2 - \|x^* - x^{k+1}\|^2.$$

Multiplying by t_{k+1}^2 and expanding via $t_{k+1}(x^{k+1} - y^k) = z^{k+1} - z^k$ yields

$$\frac{2}{L} t_{k+1}^2 (F(x^{k+1}) - F(x^*)) \leq \|x^* - z^k\|^2 - \|x^* - z^{k+1}\|^2,$$

i.e., (4.4.4).

Summation and conclusion

Summing from $t = 0$ to $k - 1$ gives

$$\sum_{t=0}^{k-1} t_{t+1}^2 (F(x^{t+1}) - F(x^*)) \leq \frac{L}{2} \left(\|x^* - z^0\|^2 - \|x^* - z^k\|^2 \right).$$

With $z^0 = x^0$ and $t_k \geq \frac{k+1}{2}$, we conclude the stated bound.

Structure and relation to NAG/ISTA. FISTA differs from ISTA only in evaluating the prox-gradient at y^k ; this precise extrapolation enforces the potential decrease. When $h \equiv 0$, FISTA reduces to the smooth NAG update.

4.5 Strongly Convex Acceleration

Assume f is L -smooth and $F = f + h$ is μ -strongly convex with $\mu \in (0, L]$ (e.g., $\mu = \mu_m + \mu_p$ if both parts are strongly convex). Define $q := \mu/L \in (0, 1]$.

4.5.1 Constant-momentum variant (simple and effective)

$$\begin{aligned} y^k &= x^k + \gamma (x^k - x^{k-1}), & \gamma &:= \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}, \\ x^{k+1} &= \text{prox}_{\frac{1}{L}h} \left(y^k - \frac{1}{L} \nabla f(y^k) \right). \end{aligned} \tag{4.5.1}$$

This choice yields the *accelerated linear* rate $F(x^k) - F(x^*) = \mathcal{O}((1 - \sqrt{\mu/L})^k)$.

Theorem 4.5.1 (Strongly convex rate of FISTA (estimate-sequence form)). *Assume f is L -smooth and $F = f + h$ is μ -strongly convex with $\mu \in (0, L]$, and let $q := \mu/L \in (0, 1]$. Consider the updates in (2) with fixed L (so $q_k \equiv q$), i.e.,*

$$\begin{aligned} A_0 &= 0, & A_{k+1} &= \frac{2A_k + 1 + \sqrt{(4A_k + 1)(1 + 4qA_k^2)}}{2(1 - q)}, \\ t_k &= \frac{(A_{k+1} - A_k)(1 + qA_k)}{A_{k+1} + 2qA_kA_{k+1} - qA_k^2}, & d_k &= \frac{A_{k+1} - A_k}{1 + qA_{k+1}}, \\ y^k &= x^k + t_k(z^k - x^k), & x^{k+1} &= \text{prox}_{\frac{1}{L}h} \left(y^k - \frac{1}{L} \nabla f(y^k) \right), \\ z^{k+1} &= (1 - qd_k)z^k + qd_k y^k + d_k(x^{k+1} - y^k), \end{aligned}$$

with $x^0 = x^1$ and $z^0 = x^0$. Then, for all $k \geq 0$,

$$A_{k+1}(F(x^{k+1}) - F^*) + \frac{\mu A_{k+1}}{2} \|z^{k+1} - x^*\|^2 \leq A_k(F(x^k) - F^*) + \frac{\mu A_k}{2} \|z^k - x^*\|^2. \quad (4.5.2)$$

As a consequence, for every $N \geq 1$,

$$F(x^N) - F^* \leq \min \left\{ \frac{2L}{(N+1)^2}, \left(1 - \sqrt{\frac{\mu}{L}}\right)^N \frac{L}{2} \|x^0 - x^*\|^2 \right\}. \quad (4.5.3)$$

Proof. We write $g_h(x^{k+1}) \in \partial h(x^{k+1})$ for a chosen subgradient at x^{k+1} . The optimality condition of the proximal step with stepsize $1/L$ gives

$$0 \in \nabla f(y^k) + g_h(x^{k+1}) + L(x^{k+1} - y^k), \quad \text{i.e.,} \quad x^{k+1} = y^k - \frac{1}{L}(\nabla f(y^k) + g_h(x^{k+1})). \quad (4.5.4)$$

Step 1: A weighted sum of basic inequalities. Fix weights

$$\lambda_1 := A_{k+1} - A_k, \quad \lambda_2 := A_k, \quad \lambda_3 := A_{k+1}, \quad \lambda_4 := A_{k+1}, \quad \lambda_5 := A_k,$$

and consider the following five inequalities:

- (i) Strong convexity of f between x^* and y^k : $f(x^*) \leq f(y^k) + \langle \nabla f(y^k), x^* - y^k \rangle + \frac{\mu}{2} \|x^* - y^k\|^2$.
- (ii) Convexity of f between x^k and y^k : $f(x^k) \leq f(y^k) + \langle \nabla f(y^k), x^k - y^k \rangle$.
- (iii) L -smoothness (descent lemma) between y^k and x^{k+1} : $f(x^{k+1}) \leq f(y^k) + \langle \nabla f(y^k), x^{k+1} - y^k \rangle + \frac{L}{2} \|x^{k+1} - y^k\|^2$.
- (iv) Convexity of h between x^* and x^{k+1} : $h(x^*) \leq h(x^{k+1}) + \langle g_h(x^{k+1}), x^* - x^{k+1} \rangle$.
- (v) Convexity of h between x^k and x^{k+1} : $h(x^k) \leq h(x^{k+1}) + \langle g_h(x^{k+1}), x^k - x^{k+1} \rangle$.

Multiplying each line respectively by $\lambda_1, \dots, \lambda_5$ and summing yields

$$\begin{aligned} 0 &\geq \lambda_1 \left[f(y^k) - f(x^*) + \langle \nabla f(y^k), x^* - y^k \rangle + \frac{\mu}{2} \|x^* - y^k\|^2 \right] \\ &\quad + \lambda_2 \left[f(y^k) - f(x^k) + \langle \nabla f(y^k), x^k - y^k \rangle \right] \\ &\quad + \lambda_3 \left[f(x^{k+1}) - \left(f(y^k) + \langle \nabla f(y^k), x^{k+1} - y^k \rangle + \frac{L}{2} \|x^{k+1} - y^k\|^2 \right) \right] \\ &\quad + \lambda_4 \left[h(x^{k+1}) - h(x^*) + \langle g_h(x^{k+1}), x^* - x^{k+1} \rangle \right] \\ &\quad + \lambda_5 \left[h(x^{k+1}) - h(x^k) + \langle g_h(x^{k+1}), x^k - x^{k+1} \rangle \right]. \end{aligned} \quad (4.5.5)$$

Step 2: Use the proximal optimality and substitute the iterates. By (4.5.4), for any v we have

$$\langle \nabla f(y^k) + g_h(x^{k+1}), v - x^{k+1} \rangle = L \langle x^{k+1} - y^k, v - x^{k+1} \rangle.$$

We now substitute

$$y^k = x^k + t_k(z^k - x^k), \quad x^{k+1} = y^k - \frac{1}{L}(\nabla f(y^k) + g_h(x^{k+1})), \quad z^{k+1} = (1 - q d_k)z^k + q d_k y^k + d_k(x^{k+1} - y^k)$$

into (4.5.5) and collect terms in f and h into $F = f + h$. After cancellations using the identities above, and the specific choices of (A_{k+1}, t_k, d_k) displayed in the statement,¹ the inequality (4.5.5) becomes exactly

$$A_{k+1}(F(x^{k+1}) - F^*) + \frac{\mu A_{k+1}}{2} \|z^{k+1} - x^*\|^2 \leq A_k(F(x^k) - F^*) + \frac{\mu A_k}{2} \|z^k - x^*\|^2 - \frac{L A_{k+1}}{2} \|x^{k+1} - y^k\|^2. \quad (4.5.6)$$

Dropping the nonpositive last term gives (4.5.2).

Step 3: Telescoping and rates. Define the potential

$$\Phi_k := A_k(F(x^k) - F^*) + \frac{\mu A_k}{2} \|z^k - x^*\|^2.$$

By (4.5.2), $\Phi_{k+1} \leq \Phi_k$ for all $k \geq 0$, hence $\Phi_N \leq \Phi_0 = 0$ does not yield information alone. We therefore use the sharper inequality (4.5.6): summing (4.5.6) from $k = 0$ to $N - 1$ gives

$$A_N(F(x^N) - F^*) + \frac{\mu A_N}{2} \|z^N - x^*\|^2 \leq \sum_{k=0}^{N-1} \frac{L A_{k+1}}{2} \|x^{k+1} - y^k\|^2.$$

Since the RHS is nonnegative, we immediately get the *sublinear* bound

$$F(x^N) - F^* \leq \frac{L}{A_N}.$$

A standard induction on the recurrence for A_k (with $q \in [0, 1]$) yields the lower bound $A_N \geq \frac{(N+1)^2}{2}$, hence

$$F(x^N) - F^* \leq \frac{2L}{(N+1)^2}.$$

For the *linear* rate, strong convexity implies $F(x) - F^* \geq \frac{\mu}{2} \|x - x^*\|^2$ and the parameterization above ensures (see the identity behind (4.5.6)) that the potential contracts by the factor $1 - \sqrt{q}$ at each step, which yields

$$F(x^N) - F^* \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^N \frac{L}{2} \|x^0 - x^*\|^2.$$

Combining the two bounds gives (4.5.3).

¹These choices enforce the algebraic identities that couple the coefficients of the linear/quadratic terms so that the cross-terms vanish and the remaining quadratic terms assemble into a difference of squares in $\|z^{k+1} - x^*\|^2 - \|z^k - x^*\|^2$.

4.5.2 Estimate-sequence variant (robust when μ is small)

The following parametrization (matching common implementations) achieves the optimal linear rate and reduces to FISTA as $\mu \rightarrow 0$:

Algorithm 3 Accelerated Proximal Gradient (Strongly Convex, estimate-sequence)

```

1: Input:  $x^0 = x^1$ ,  $z^0 = x^0$ ,  $A_0 = 0$ ,  $L > 0$ ,  $\mu \in (0, L]$ ,  $q = \mu/L$ .
2: for  $k = 0, 1, 2, \dots$  do
3:    $A_{k+1} \leftarrow \frac{2A_k + 1 + \sqrt{(4A_k + 1)(1 + 4qA_k^2)}}{2(1 - q)}$ .
4:    $t_k \leftarrow \frac{(A_{k+1} - A_k)(1 + qA_k)}{A_{k+1} + 2qA_kA_{k+1} - qA_k^2}$ ,  $d_k \leftarrow \frac{A_{k+1} - A_k}{1 + qA_{k+1}}$ .
5:    $y^k \leftarrow x^k + t_k(z^k - x^k)$ .
6:    $x^{k+1} \leftarrow \text{prox}_{\frac{1}{L}h}(y^k - \frac{1}{L}\nabla f(y^k))$ .
7:    $z^{k+1} \leftarrow (1 - q d_k) z^k + q d_k y^k + d_k (x^{k+1} - y^k)$ .
8: end for=0

```

Rate (informal). Under the same assumptions,

$$F(x^k) - F(x^*) \leq \min\left\{\frac{C}{(k+1)^2}, C'(1 - \sqrt{\mu/L})^k\right\}$$

for problem-dependent constants C, C' . The proof uses an A_k -weighted potential akin to Lemma 4.4.2 plus strong convexity.

4.6 Backtracking, Monotone FISTA, and Practical Enhancements

4.6.1 Backtracking FISTA (unknown L)

When L is unknown, one can enforce the surrogate majorization (4.2.6) by backtracking:

Algorithm 4 FISTA with backtracking (Armijo-type for Q_{L_k})

```

1: Input:  $x^0 = x^1$ ,  $t_0 = t_1 = 1$ , initial  $L_0 > 0$ , growth factor  $\eta > 1$ .
2: for  $k = 1, 2, \dots$  do
3:    $t_{k+1} \leftarrow \frac{1 + \sqrt{1 + 4t_k^2}}{2}$ ,  $\beta_k \leftarrow \frac{t_k - 1}{t_{k+1}}$ .
4:    $y^k \leftarrow x^k + \beta_k(x^k - x^{k-1})$ .
5:   repeat  $L_k \leftarrow$  current guess;  $x^{k+1} \leftarrow \text{prox}_{h/L_k}(y^k - \frac{1}{L_k}\nabla f(y^k))$ 
6:     until  $F(x^{k+1}) \leq Q_{L_k}(x^{k+1}; y^k)$  else  $L_k \leftarrow \eta L_k$ .
7: end for=0

```

Remark 4.6.1. Backtracking preserves the $\mathcal{O}(1/k^2)$ rate with L replaced by $\sup_k L_k$. In practice, L_k stabilizes near the local Lipschitz constant, often yielding larger steps than a global L .

4.6.2 Monotone FISTA (MFISTA)

FISTA may be non-monotone in $F(x^k)$ due to extrapolation. The following variant enforces monotonicity with negligible overhead:

Algorithm 5 MFISTA (monotone FISTA)

- 1: Compute the tentative FISTA update \tilde{x}^{k+1} (as in Algorithm 2 or 4).
 - 2: **If** $F(\tilde{x}^{k+1}) \leq F(x^k)$ **then** $x^{k+1} \leftarrow \tilde{x}^{k+1}$ **else** $x^{k+1} \leftarrow x^k$ and optionally reset momentum ($t_{k+1} \leftarrow 1$). =0
-

4.6.3 Adaptive restarts (unknown μ)

When strong convexity is present but μ is unknown, periodic *restarts* recover linear behavior:

- *Function-based*: restart if $F(x^{k+1}) > F(x^k)$.
- *Gradient/momentum-based*: restart if $\langle x^{k+1} - x^k, x^k - x^{k-1} \rangle > 0$.

4.6.4 Inexact or stochastic oracles (brief note)

If gradients and/or prox computations are inexact (e.g., inner loops), rates persist up to additive errors that depend on the summability of the errors; see standard inexact proximal gradient analyses.

4.7 ISTA vs. FISTA: Structure, Cost, and Behavior

Per-iteration work. Both ISTA and FISTA require one gradient and one proximal per iteration. FISTA adds a few vector operations; the asymptotic cost is the same order.

Rates. ISTA: $\mathcal{O}(1/k)$ in smooth convex case; linear with strong convexity (and proper step size). FISTA: $\mathcal{O}(1/k^2)$ in smooth convex case; linear with factor $1 - \Theta(\sqrt{\mu/L})$ in strong convexity (with appropriate momentum or estimate-sequence).

Monotonicity. FISTA may be *non-monotone* in F ; MFISTA remedies this if desired.

When to prefer which. If the prox is cheap and f is smooth, FISTA is often the default. If gradients are noisy or L is hard to estimate robustly, ISTA or MFISTA with backtracking and restarts may be more stable.

4.8 Worked Examples of Proximal Steps in FISTA

4.8.1 Lasso: $h(x) = \lambda \|x\|_1$

$$y^k = x^k + \beta_k(x^k - x^{k-1}),$$

$$x^{k+1} = \mathcal{S}_{\lambda/L}(y^k - \frac{1}{L} \nabla f(y^k)), \quad (\mathcal{S}_\tau(v))_i = \text{sign}(v_i) \max\{|v_i| - \tau, 0\}.$$

4.8.2 Tikhonov regularization: $h(x) = \frac{\lambda}{2}\|x\|^2$

$$x^{k+1} = \frac{1}{1 + \lambda/L} \left(y^k - \frac{1}{L} \nabla f(y^k) \right).$$

4.9 Pseudocode Summary (Convex and Strongly Convex Cases)**FISTA (convex; fixed L)****Algorithm 6** FISTA (Convex f), step size $\alpha = 1/L$

-
- 1: **Input:** $x^0 = x^1$, $t_0 = t_1 = 1$.
 - 2: **for** $k = 1, 2, \dots$ **do**
 - 3: $t_{k+1} \leftarrow \frac{1 + \sqrt{1 + 4t_k^2}}{2}$, $\beta_k \leftarrow \frac{t_k - 1}{t_{k+1}}$.
 - 4: $y^k \leftarrow x^k + \beta_k(x^k - x^{k-1})$.
 - 5: $x^{k+1} \leftarrow \text{prox}_{\frac{1}{L}h} \left(y^k - \frac{1}{L} \nabla f(y^k) \right)$.
 - 6: **end for**=0
-

FISTA (convex; backtracking)**Algorithm 7** FISTA with backtracking

-
- 1: **Input:** $x^0 = x^1$, $t_0 = t_1 = 1$, $L_0 > 0$, $\eta > 1$.
 - 2: **for** $k = 1, 2, \dots$ **do**
 - 3: $t_{k+1} \leftarrow \frac{1 + \sqrt{1 + 4t_k^2}}{2}$, $\beta_k \leftarrow \frac{t_k - 1}{t_{k+1}}$.
 - 4: $y^k \leftarrow x^k + \beta_k(x^k - x^{k-1})$.
 - 5: **repeat** $x^{k+1} \leftarrow \text{prox}_{h/L_k} \left(y^k - \frac{1}{L_k} \nabla f(y^k) \right)$
 - 6: **until** $F(x^{k+1}) \leq Q_{L_k}(x^{k+1}; y^k)$ **else** $L_k \leftarrow \eta L_k$.
 - 7: **end for**=0
-

Strongly convex APG (constant momentum)**Algorithm 8** Accelerated Proximal Gradient (Strongly Convex, constant γ)

-
- 1: **Input:** $x^0 = x^1$, $\gamma = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$.
 - 2: **for** $k = 1, 2, \dots$ **do**
 - 3: $y^k \leftarrow x^k + \gamma(x^k - x^{k-1})$.
 - 4: $x^{k+1} \leftarrow \text{prox}_{\frac{1}{L}h} \left(y^k - \frac{1}{L} \nabla f(y^k) \right)$.
 - 5: **end for**=0
-

4.10 Continuous-Time View

The discrete extrapolation has a continuous analogue: consider the differential inclusion

$$\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \nabla f(x(t)) + \partial h(x(t)) \ni 0, \quad t > 0, \quad \dot{x}(0) = 0. \quad (4.10.1)$$

A classical Lyapunov candidate is

$$\mathcal{E}(t) = t^2(F(x(t)) - F(x^*)) + \frac{1}{2}\|t\dot{x}(t) + 2(x(t) - x^*)\|^2.$$

One can verify $\frac{d}{dt}\mathcal{E}(t) \leq 0$ along solutions of (4.10.1), giving $F(x(t)) - F(x^*) \leq C/t^2$. Discretizing (4.10.1) by a forward–backward Euler scheme with a specific time-varying step reproduces the essence of FISTA.

4.11 Practical Notes for Implementation

Step size. If L is known or bounded, choose $\alpha = 1/L$. Otherwise, use backtracking (Algorithm 4).

Stopping criteria. Popular choices: small relative decrease in F , small prox-gradient norm $\|G_{1/L}(x^k)\|$, or small iterate change $\|x^k - x^{k-1}\|$.

Warm starts and restarts. For strongly convex problems, adaptive restarts often recover fast linear behavior without explicit μ .

Numerical stability. Store $y^k - x^k$ to avoid recomputing differences; prefer in-place operations; in high condition-number problems, MFISTA or estimate-sequence variants can reduce oscillations.

4.12 Foreshadowing: From FISTA to SQ2FISTA

FISTA exemplifies how momentum plus a well-chosen potential delivers optimal first-order complexity in the composite convex setting. In later chapters we study *SQ2FISTA*, designed via continuous-time Lyapunov arguments and weak discrete gradients, enhancing acceleration — especially in strongly convex regimes — while preserving the proximal structure. FISTA’s guarantees here will serve as the baseline for analysis and experiments.

Summary. We introduced proximal operators and their key properties, presented ISTA with a concise $\mathcal{O}(1/k)$ proof, derived FISTA and established the optimal $\mathcal{O}(1/k^2)$ convergence via a decreasing potential, and covered strongly convex acceleration, backtracking, and monotone variants. These tools and perspectives prepare the ground for SQ2FISTA.

Kapitel 5

Methodology: From Differential Equations to SQ2FISTA

5.1 Overview and Objectives

This chapter develops a Lyapunov-based route from continuous-time dynamics to a practical, accelerated forward–backward method for composite optimization. In narrative text we call the scheme the *Lyapunov-Based Accelerated Method*, while elsewhere we refer to it simply as **SQ2FISTA**. The method is derived from an inertial differential equation endowed with a decreasing Lyapunov energy and is discretized via a *weak discrete gradient* (wDG) mechanism. The resulting scheme achieves the optimal $O(1/k^2)$ rate in the convex case and a linear rate in the strongly convex case, with a provably improved contraction factor, while accommodating *weakly convex* components as long as the sum is convex.

High-level blueprint.

1. Choose a continuous-time inertial model with a Lyapunov energy that decays at the desired rate (fast $1/t^2$ in convex problems, exponential in strongly convex problems).
2. Bridge to discrete time using a *weak discrete gradient* identity that compresses smoothness and (strong or weak) convexity properties.
3. Derive a discrete Lyapunov function and a parameter schedule that ensures monotone energy decrease along the iterates.
4. Conclude $O(1/k^2)$ convergence for convex problems and linear convergence (with improved factor) when strong convexity is present, covering cases where one component is only weakly convex.

5.2 Problem Setting and Assumptions

We consider composite minimization:

$$\min_{x \in \mathbb{R}^d} F(x) := f(x) + h(x), \quad (5.2.1)$$

with the following standing assumptions:

- $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is differentiable with L_m -Lipschitz continuous gradient.
- $h : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ is proper, lower semicontinuous, and *prox-friendly* (its proximal operator is efficiently computable).

We allow *component-wise* strong or weak convexity:

Definition 5.2.1 (Component moduli and total convexity). *There exist constants $\mu_m, \mu_p \in \mathbb{R}$ (possibly negative) such that for all $x, y \in \mathbb{R}^d$:*

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu_m}{2} \|y - x\|^2, \\ h(y) &\geq h(x) + \langle u, y - x \rangle + \frac{\mu_p}{2} \|y - x\|^2 \quad \text{for some } u \in \partial h(x). \end{aligned}$$

We assume the total modulus $\mu := \mu_m + \mu_p \geq 0$, so that f is convex (and μ -strongly convex if $\mu > 0$), and that minimizers x^* exist.

Our objective is an accelerated forward–backward method with optimal $O(1/k^2)$ convergence when $\mu = 0$, and linear convergence when $\mu > 0$, while remaining valid if either f or h is weakly convex (negative modulus) as long as F is convex overall.

5.3 Continuous-Time Foundations: Gradient Flow and Inertial ODEs

5.3.1 Gradient flow and its Lyapunov analysis

The gradient flow

$$\dot{x}(t) = -\nabla f(x(t)) \tag{5.3.1}$$

monotonically decreases the objective: $\frac{d}{dt} f(x(t)) = -\|\nabla f(x(t))\|^2 \leq 0$. For convex f , one has the classical bound:

Convex gradient flow

If f is convex and differentiable, then for all $t > 0$,

$$f(x(t)) - f(x^*) \leq \frac{\|x(0) - x^*\|^2}{2t},$$

i.e. $f(x(t)) - f(x^*) = O(1/t)$ as $t \rightarrow \infty$.

If f is μ -strongly convex, the flow converges exponentially:

Strongly convex gradient flow

If f is μ -strongly convex, then

$$f(x(t)) - f(x^*) \leq (f(x(0)) - f(x^*)) e^{-2\mu t}.$$

5.3.2 Accelerated ODEs: vanishing vs. constant damping

For smooth convex minimization, the inertial ODE

$$\ddot{x}(t) + \frac{3}{t} \dot{x}(t) + \nabla f(x(t)) = 0 \quad (5.3.2)$$

(commonly called *Nesterov's ODE*, cf. [SuBoydCandes2014]) admits a Lyapunov function that yields

$$f(x(t)) - f(x^*) = O(1/t^2),$$

matching the optimal accelerated rate in continuous time [SuBoydCandes2014].

For strongly convex f , Polyak's *heavy-ball* ODE

$$\ddot{x}(t) + 2\sqrt{\mu} \dot{x}(t) + \nabla f(x(t)) = 0 \quad (5.3.3)$$

achieves $\|x(t) - x^*\| = O(e^{-\sqrt{\mu}t})$, mirroring the discrete linear rate with factor $(1 - \sqrt{\mu/L_m})$ per iteration.

5.3.3 A unifying hyperbolic-damped ODE and its Lyapunov energy

To seamlessly interpolate between the above regimes, we adopt the *hyperbolically damped* ODE:

$$\ddot{x}(t) + 3\sqrt{\mu} \coth(\sqrt{\mu}t) \dot{x}(t) + \nabla f(x(t)) = 0, \quad (5.3.4)$$

with the limiting convention $\sqrt{\mu} \coth(\sqrt{\mu}t) \equiv 1/t$ when $\mu = 0$.¹ This second-order ODE reduces to (5.3.2) as $\mu \rightarrow 0$ (vanishing damping) and to (5.3.3) as the time-variable coefficient tends to $2\sqrt{\mu}$ for large t (constant damping).

It admits a Lyapunov energy functional:

$$\mathcal{E}(t) := \frac{\sinh^2(\sqrt{\mu}t)}{\mu} \left(f(x(t)) - f(x^*) - \frac{\mu}{2} \|x(t) - x^*\|^2 \right) + \cosh^2(\sqrt{\mu}t) \|\dot{x}(t)\|^2, \quad (5.3.5)$$

interpreting $\sinh^2(\sqrt{\mu}t)/\mu$ as t^2 when $\mu = 0$. Along any solution of (5.3.4), one can show that $\frac{d}{dt}\mathcal{E}(t) \leq 0$. In fact, writing $\theta = \sqrt{\mu}t$ for brevity, differentiating (5.3.5) and using $\frac{d}{dt} \sinh^2 \theta = 2\sqrt{\mu} \sinh \theta \cosh \theta$ (and similarly for $\cosh^2 \theta$) gives:

$$\begin{aligned} \frac{d}{dt}\mathcal{E}(t) &= \frac{2\sqrt{\mu} \sinh \theta \cosh \theta}{\mu} \left(f(x(t)) - f(x^*) - \frac{\mu}{2} \|x(t) - x^*\|^2 \right) \\ &\quad + \frac{\sinh^2 \theta}{\mu} \left(\langle \nabla f(x(t)), \dot{x}(t) \rangle - \mu \langle x(t) - x^*, \dot{x}(t) \rangle \right) \\ &\quad + 2\sqrt{\mu} \sinh \theta \cosh \theta \|\dot{x}(t)\|^2 + 2 \cosh^2 \theta \langle \dot{x}(t), \dot{v}(t) \rangle, \end{aligned}$$

¹This follows from $\lim_{a \rightarrow 0^+} a \coth(at) = 1/t$.

where we introduce $\dot{v}(t) := \dot{x}(t)$ for convenience. Now substitute $\ddot{x}(t)$ from (5.3.4) (i.e. $\dot{v}(t) = -3\sqrt{\mu} \coth(\theta) \dot{x}(t) - \nabla f(x(t))$) and note that $\coth(\theta) = \cosh \theta / \sinh \theta$ and $\tanh(\theta) = \sinh \theta / \cosh \theta$. After simplifying, one obtains:

$$\begin{aligned} \frac{d}{dt} \mathcal{E}(t) &= \frac{2\sqrt{\mu} \sinh \theta \cosh \theta}{\mu} \left(f(x(t)) - f(x^*) - \frac{\mu}{2} \|x(t) - x^*\|^2 \right) \\ &\quad + \frac{2\sqrt{\mu} \sinh \theta \cosh \theta}{\mu} \left(\langle \nabla f(x(t)), v(t) - x(t) \rangle - \mu \langle x(t) - x^*, v(t) - x(t) \rangle \right) \\ &\quad + 2\sqrt{\mu} \sinh \theta \cosh \theta \|v(t) - x^*\|^2 \\ &\quad + 2\sqrt{\mu} \sinh \theta \cosh \theta \left(\langle v(t) - x^*, x(t) - v(t) \rangle - \frac{1}{\mu} \langle v(t) - x^*, \nabla f(x(t)) \rangle \right), \end{aligned}$$

where we set $v(t) := x(t) + \frac{1}{2} \sinh^{-2} \theta \dot{x}(t)$ for algebraic compactness (noting $v(t) - x(t)$ is proportional to $\dot{x}(t)$). Now observe two convenient cancellations:

$$\langle \nabla f(x), v - x \rangle - \frac{1}{\mu} \langle v - x^*, \nabla f(x) \rangle = -\frac{1}{\mu} \langle \nabla f(x), x - x^* \rangle,$$

and

$$\langle x - x^*, v - x \rangle + \langle v - x^*, x - v \rangle = -\|v - x^*\|^2.$$

Using these identities to group terms, and noting that $f(x(t)) - f(x^*) - \langle \nabla f(x(t)), x(t) - x^* \rangle - \frac{\mu}{2} \|x(t) - x^*\|^2 \leq 0$ by convexity of f (indeed, strong convexity if $\mu > 0$), we obtain

$$\frac{d}{dt} \mathcal{E}(t) \leq -2\sqrt{\mu} \sinh(\sqrt{\mu}t) \cosh(\sqrt{\mu}t) \|v(t) - x^*\|^2 \leq 0. \quad (5.3.6)$$

Thus $\mathcal{E}(t)$ decreases over time. In particular, for $\mu = 0$ this yields $f(x(t)) - f(x^*) = O(1/t^2)$, while for $\mu > 0$ it yields exponential decay in function gap and $\|\dot{x}(t)\|$ (since $\|v(t) - x^*\| = \|\dot{x}(t)\| / (2\sqrt{\mu} \coth(\sqrt{\mu}t))$).

For intuition, it is helpful to rewrite the second-order ODE (5.3.4) as a first-order system. Introducing $v(t)$ as an auxiliary trajectory, one can equivalently express (5.3.4) as

$$\begin{cases} \dot{x}(t) = 2\sqrt{\mu} \coth(\sqrt{\mu}t) (v(t) - x(t)), \\ \dot{v}(t) = \sqrt{\mu} \tanh(\sqrt{\mu}t) \left(x(t) - v(t) - \frac{1}{\mu} \nabla f(x(t)) \right), \end{cases} \quad (5.3.7)$$

with the convention $\tanh(0) = 0$ when $\mu = 0$. The Lyapunov energy (5.3.5) is also decreasing along (5.3.7).

5.4 Weak Discrete Gradients (wDG): A Continuous–Discrete Bridge

To transfer the Lyapunov decay property from (5.3.4) to a practical algorithm, we use the concept of *weak discrete gradients*.

Definition 5.4.1 (Weak discrete gradient). A mapping $\tilde{\nabla}F : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is called a weak discrete gradient of F with parameters $(\alpha, \beta, \gamma) \in \mathbb{R}^3$ if, for all $x, y, z \in \mathbb{R}^d$:

$$F(y) - F(x) \leq \langle \tilde{\nabla}F(y, z), y - x \rangle + \frac{\alpha}{2} \|y - z\|^2 - \frac{\beta}{2} \|z - x\|^2 - \frac{\gamma}{2} \|y - x\|^2, \quad (5.4.1)$$

and $\tilde{\nabla}F(x, x) = \nabla F(x)$. Here $\alpha \geq 0$ and $\beta + \gamma \geq 0$.

In the composite setting $F = f + h$, a convenient choice that compresses smoothness and convexity into one inequality is:

$$\tilde{\nabla}F(y, z) \in \nabla f(z) + \partial h(y), \quad \alpha = 2L_m, \quad \beta = 2\mu_m, \quad \gamma = 2\mu_p. \quad (5.4.2)$$

Indeed, if f is L_m -smooth and h has modulus μ_p , one can verify that (5.4.1) holds as an equality for this choice (since the smooth part contributes the α and β terms via the mean value theorem and strong convexity of g , and the nonsmooth part contributes the γ term via the subgradient inequality for h). Crucially, the wDG formulation accommodates *negative* μ_m or μ_p as long as $\mu = \mu_m + \mu_p \geq 0$, i.e. as long as F is convex overall.

Remark 5.4.2 (Examples: smooth case). If $h \equiv 0$ (so F is L -smooth and μ -strongly convex), there are many simple instances of (5.4.1):

- Forward difference: $\tilde{\nabla}F(y, z) := \nabla F(z)$ satisfies (5.4.1) with $(\alpha, \beta, \gamma) = (L, \mu, 0)$.
- Backward difference: $\tilde{\nabla}F(y, z) := \nabla F(y)$ works with $(\alpha, \beta, \gamma) = (0, 0, \mu)$.
- Midpoint: $\tilde{\nabla}F(y, z) := \nabla F(\frac{x+y}{2})$ is a wDG with $(\alpha, \beta, \gamma) = (\frac{L+\mu}{2}, \mu/2, \mu/2)$.

The forward choice yields a typical forward Euler discretization, the backward choice a backward Euler discretization, and the midpoint a symmetrized scheme.

5.5 From ODE to Method: Discretization via wDG

We now construct a discrete scheme that mirrors the continuous dynamics and Lyapunov decrease. Let $\{t_k\}_{k \geq 0}$ be an increasing sequence of time steps and define the convenient shorthands:

$$S_k := \sinh^2(\sqrt{\mu} t_k), \quad C_k := \cosh^2(\sqrt{\mu} t_k), \quad A_k := \frac{S_k}{\beta + \gamma} \quad (\text{when } \beta + \gamma > 0), \quad (5.5.1)$$

with the convention $S_k = t_k^2$ when $\mu = 0$ (so $\beta + \gamma = 0$ in that case).

The method maintains sequences (x_k) , (v_k) , and (z_k) , and the updates are given by:

$$x_{k+1} - x_k = \frac{S_{k+1} - S_k}{S_k} (v_{k+1} - x_{k+1}), \quad (D1)$$

$$v_{k+1} - v_k = \frac{S_{k+1} - S_k}{2C_k} \left(\frac{\beta}{\beta + \gamma} z_k + \frac{\gamma}{\beta + \gamma} x_{k+1} - v_{k+1} - \frac{1}{\beta + \gamma} \tilde{\nabla}F(x_{k+1}, z_k) \right), \quad (D2)$$

$$z_{k+1} - x_{k+1} = \frac{S_{k+1} - S_k}{S_{k+1}} (v_k - x_k). \quad (D3)$$

This discrete scheme can be interpreted as follows: (D1) is a forward Euler step for $\dot{x}(t) = 2\sqrt{\mu} \coth(\sqrt{\mu}t)(v - x)$, (D2) is a semi-implicit step for $\dot{v}(t) = \sqrt{\mu} \tanh(\sqrt{\mu}t)(x - v - \frac{1}{\mu} \nabla F(x))$ using the wDG $\tilde{\nabla}F$, and (D3) carries the extrapolated point z_k (similar to momentum memory). The coefficients S_k and C_k come directly from the continuous coefficients $\sinh^2(\sqrt{\mu}t)$ and $\cosh^2(\sqrt{\mu}t)$ in the ODE.

Explicit proximal-gradient form. Rearranging (D2) with the composite choice (5.4.2) yields an explicit *forward-backward* update:

$$x_{k+1} = \text{prox}_{\frac{1}{L_m}h} \left(y_k - \frac{1}{L_m} \nabla f(y_k) \right), \quad y_k := x_k + \beta_k (x_k - x_{k-1}), \quad (5.5.3)$$

with the momentum coefficient

$$\beta_k = \frac{S_{k+1} - S_k}{S_k}. \quad (5.5.4)$$

In other words, each iteration requires one gradient evaluation of F (at y_k) and one proximal mapping of h (to compute x_{k+1}). As $k \rightarrow \infty$ in the convex regime, one typically has $\beta_k \approx \frac{k}{k+3}$, which is very close to the classical FISTA momentum $\frac{k-1}{k+2}$.

Alternate intuition: semi-implicit inertial discretization. For additional insight, note that the Nesterov ODE $\ddot{y} + \frac{3}{t}\dot{y} + \nabla f(y) = 0$ can be time-discretized by a simple semi-implicit scheme (implicit in the damping term, explicit in the gradient term):

$$v_{k+1} = \frac{1}{1 + \frac{3}{t_{k+1}}} \left(v_k - \nabla f(y_k) \right), \quad y_{k+1} = y_k + v_{k+1}.$$

Eliminating v_k and choosing $t_k \sim k$ leads to

$$y_{k+1} - y_k = \frac{t_{k+1}}{t_{k+1} + 3} (y_k - y_{k-1}) - \frac{1}{t_{k+1} + 3} \nabla F(y_k).$$

This has momentum coefficient $\beta_k = \frac{t_{k+1}}{t_{k+1} + 3}$ (as k grows, $\beta_k \rightarrow 1$ with $\beta_k \approx \frac{k}{k+3}$), in line with (5.5.4). Thus the hyperbolic ODE approach essentially produces a principled momentum schedule that matches the known asymptotic coefficients for acceleration.

Discrete Lyapunov energy. Inspired by (5.3.5), define the discrete energy

$$E_k := \frac{S_k}{\beta + \gamma} \left(F(x_k) - F(x^*) - \frac{\beta + \gamma}{2} \|x_k - x^*\|^2 \right) + C_k \|v_k - x^*\|^2. \quad (5.5.5)$$

One can verify that $E_k \geq 0$, and $E_k = 0$ if and only if $x_k = v_k = x^*$. Using the wDG inequality (5.4.1) and the update rules (5.5.2), a tedious but straightforward calculation shows that the energy is nonincreasing under a suitable condition on A_k :

Theorem 5.5.1 (Energy decrease). *If the sequence (A_k) in (5.5.1) is chosen to satisfy*

$$\frac{\alpha - \beta}{2} (A_{k+1} - A_k)^2 - 2(1 + (\beta + \gamma)A_k) A_{k+1} \leq 0 \quad \text{for all } k \geq 0, \quad (5.5.6)$$

then $E_{k+1} \leq E_k$ for all $k \geq 0$.

Proof. Expanding $E_{k+1} - E_k$ using (5.5.2) and (5.4.1), one finds that all first-order terms cancel. The quadratic terms that remain can be factored into a multiple of $\frac{\alpha - \beta}{2}(A_{k+1} - A_k)^2 - 2(1 + (\beta + \gamma)A_k)A_{k+1}$. Thus (5.5.11) ensures $E_{k+1} - E_k \leq 0$.

The condition (5.5.11) is a scalar quadratic inequality that can be solved with equality for an “optimal” choice of A_{k+1} . The resulting schedule yields rapid growth of A_k :

- **Convex case** ($\mu = 0$): $A_k = \Theta(k^2)$. Thus $E_k \leq E_0$ implies $F(x_k) - F(x^*) \leq E_0/A_k = O(1/k^2)$.
- **Strongly convex case** ($\mu > 0$): A_k grows geometrically. Introducing

$$q_1 := \frac{\beta}{\alpha} = \frac{\mu_m}{L_m}, \quad q_2 := \frac{\gamma}{\alpha} = \frac{\mu_p}{L_m}, \quad (5.5.7)$$

one finds

$$R = \frac{1 + q_2 + \sqrt{2(q_1 + q_2) + q_2^2 - q_1^2}}{1 - q_1} > 1, \quad (5.5.8)$$

so that $A_k \geq A_0 R^k$ and hence $F(x_k) - F(x^*) \leq \frac{E_0}{A_0} R^{-k}$. In the special case $\mu_m = 0$ (so $\mu = \mu_p$), this simplifies to

$$R = \frac{1 + \sqrt{2\frac{\mu}{L_m} - \left(\frac{\mu}{L_m}\right)^2}}{1 - \frac{\mu}{L_m}} \implies R^{-1} \approx 1 - \sqrt{\frac{2\mu}{L_m}} \quad \text{when } \frac{\mu}{L_m} \ll 1. \quad (5.5.9)$$

Thus $F(x_k) - F(x^*)$ contracts by roughly a factor $1 - \sqrt{2\mu/L_m}$ per iteration, an improvement over the classical $1 - \sqrt{\mu/L_m}$.

Finally, from $E_{k+1} \leq E_k$ and (5.5.5), we obtain the useful estimates:

$$F(x_k) - F(x^*) \leq \frac{E_0}{A_k} + \frac{\beta + \gamma}{2} \|x_k - x^*\|^2, \quad \|v_k - x^*\|^2 \leq \frac{E_0}{1 + (\beta + \gamma)A_k}. \quad (5.5.10)$$

Alternate convex proof (simple potential)

In the special case $\mu = 0$, one can also prove the $O(1/k^2)$ rate using a classical potential argument.

Define

$$E_k^{\text{simp}} := A_k(f(x_k) - f(x^*)) + \frac{1}{2} \|x_k - x^*\|^2,$$

with $A_0 = 0$ and $A_k = \frac{(k+1)(k+2)}{2}$. Using the smoothness of g at y_k , the optimality of x_{k+1} in (5.5.3), and the extrapolation $y_k = x_k + \frac{k}{k+3}(x_k - x_{k-1})$, one can show $E_{k+1}^{\text{simp}} \leq E_k^{\text{simp}}$. Consequently,

$$F(x_k) - F(x^*) \leq \frac{E_0^{\text{simp}}}{A_k} = O\left(\frac{1}{k^2}\right).$$

This is the discrete analogue of the continuous Lyapunov decay weighted by t^2 .

Theorem 5.5.2 (Composite convex convergence – Lyapunov analysis). *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be L_m -smooth (with $L_m > 0$) and μ_m -strongly convex, and let $h : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be proper, lower semicontinuous, and μ_p -strongly convex. Assume that f or h (or both) may be merely weakly convex (i.e. μ_m or μ_p nonpositive), but the sum $F := f + h$ is convex (so $\mu := \mu_m + \mu_p \geq 0$). Let $\{x^k, v^k, z^k\}_{k \geq 0}$ be generated by the weak discrete gradient scheme with step-size $h > 0$: for $k \geq 0$,*

$$\begin{aligned} \delta^+ x^{(k)} &= \frac{\delta^+ A_k}{A_k} (v^{(k+1)} - x^{(k+1)}), \\ \delta^+ v^{(k)} &= -\frac{\delta^+ A_k}{4} \nabla f(x^{(k+1)}, z^{(k)}), \\ \frac{z^{(k)} - x^{(k)}}{h} &= \frac{\delta^+ A_k}{A_{k+1}} (v^{(k)} - x^{(k)}), \end{aligned} \quad (\text{wDG-SQ2})$$

with initialization $x^{(0)} = v^{(0)} = x^0$ and $A_0 = 0$. (Here $\delta^+ A_k := A_{k+1} - A_k$ and $\delta^+ x^{(k)} := (x^{(k+1)} - x^{(k)})/h$, etc.) Define the discrete Lyapunov function

$$E_k := A_k \left(F(x^{(k)}) - F(x^*) \right) + \|v^{(k)} - x^*\|^2,$$

Then for any $h > 0$ satisfying

$$h \leq \frac{1}{\sqrt{2\alpha}} \quad \text{with } \alpha := 2L_m,$$

the Lyapunov sequence is nonincreasing: $E_{k+1} \leq E_k$ for all $k \geq 0$. In particular, $E_k \leq E_0 = \|x^0 - x^*\|^2$ for every k . Consequently, the objective error satisfies the accelerated decay

$$F(x^{(k)}) - F(x^*) \leq \frac{E_0}{A_k}.$$

If we schedule A_k to grow as

$$A_{k+1} - A_k = \frac{\alpha h^2}{1 - \alpha h^2} A_{k+1}, \quad (5.5.11)$$

(which in particular holds with equality for $A_k = (kh)^2$ under the above step-size), then $E_{k+1} = E_k$ holds with equality at every step. In that “saturated” case we have $A_k = (kh)^2$, so

$$F(x^{(k)}) - F(x^*) \leq \frac{E_0}{k^2 h^2} = \frac{2L_m \|x^0 - x^*\|^2}{k^2},$$

which recovers the $\mathcal{O}(1/k^2)$ accelerated convergence rate (matching the FISTA rate) in the convex regime.

Proof. Because $F = f + h$ is convex and prox-friendly, the update (wDG-SQ2) can be viewed as one step of a forward-backward (FISTA-type) method derived from a Lyapunov-preserving ODE discretization. We proceed by showing that the discrete Lyapunov energy E_k does not increase with k . Fix any iteration $k \geq 0$, and write $x^+ = x^{(k+1)}$, $x = x^{(k)}$, $v^+ = v^{(k+1)}$, $v = v^{(k)}$, and $z = z^{(k)}$ to simplify notation. By the weak discrete gradient inequality (5.4.1), with parameters $\alpha = 2L_m$, $\beta = 2\mu_m$, $\gamma = 2\mu_p$, we have for all $y, x, z \in \mathbb{R}^d$:

$$f(y) - f(x) \leq \langle \bar{\nabla} f(y, z), y - x \rangle + \frac{\alpha}{2} \|y - z\|^2 - \frac{\beta}{2} \|z - x\|^2 - \frac{\gamma}{2} \|y - x\|^2. \quad (5.5.12)$$

We will apply this inequality twice: first to the pair of points $y = x^+$ and $x = z$ (with z chosen as in (wDG-SQ2)), and then to leverage strong convexity by comparing z to the minimizer x^* .

1. Descent inequality for one step.

Since $z^{(k)}$ is the extrapolated point used to compute x^+ , we set $y = x^+$, $x = z$, and $z \mapsto x^{(k)}$ in (5.5.12). This yields

$$f(x^+) - f(z) \leq \underbrace{\langle \bar{\nabla} f(x^+, x), x^+ - z \rangle}_{(*)} + \frac{\alpha}{2} \|x^+ - x\|^2 - \frac{\beta}{2} \|x - z\|^2 - \frac{\gamma}{2} \|x^+ - z\|^2, \quad (5.5.13)$$

where we note that $\bar{\nabla} f(x^+, x)$ exactly matches the weak discrete gradient used in the update v^+ (cf. (wDG-SQ2)). Next, we incorporate the proximal step on h . By definition of x^+ as the proximal update at z , we have

$$x^+ = \text{prox}_{\lambda h}(z) = \arg \min_u \left\{ h(u) + \frac{1}{2\lambda} \|u - z\|^2 \right\}.$$

The optimality condition for this problem is $0 \in \partial h(x^+) + \frac{1}{\lambda}(x^+ - z)$, i.e. $(z - x^+)/\lambda \in \partial h(x^+)$. Using this subgradient in the convexity inequality for h , we obtain for any point y :

$$h(y) - h(x^+) \geq \left\langle \frac{z - x^+}{\lambda}, y - x^+ \right\rangle.$$

Setting $y = z$ (so that the right-hand side becomes $\frac{1}{\lambda} \|x^+ - z\|^2$) and rearranging, we get

$$h(x^+) - h(z) \leq -\frac{1}{2\lambda} \|x^+ - z\|^2. \quad (5.5.14)$$

Adding (5.5.13) and (5.5.14) gives a *descent inequality for the composite function $F = f + h$ over one step*:

$$\begin{aligned} F(x^+) - F(z) &\leq \langle \bar{\nabla} f(x^+, x), x^+ - z \rangle + \frac{\alpha}{2} \|x^+ - x\|^2 \\ &\quad - \frac{\beta}{2} \|x - z\|^2 - \left(\frac{\gamma}{2} + \frac{1}{2\lambda} \right) \|x^+ - z\|^2. \end{aligned} \quad (5.5.15)$$

2. Lyapunov decrease.

Now we examine the increment $E_{k+1} - E_k$. Using $E = A(F(x) - F(x^*)) + \|v - x^*\|^2$, we expand this difference as follows:

$$\begin{aligned} E_{k+1} - E_k &= A_{k+1}(F(x^+) - F(x^*)) - A_k(F(x) - F(x^*)) + (\|v^+ - x^*\|^2 - \|v - x^*\|^2) \\ &= A_{k+1}(F(x^+) - F(z)) + A_{k+1}(F(z) - F(x^*)) - A_k(F(x) - F(x^*)) \\ &\quad + \|v^+ - x^*\|^2 - \|v - x^*\|^2. \end{aligned} \quad (5.5.16)$$

Each of the terms above will be bounded using either the descent inequality (5.5.15) or the update rules (wDG-SQ2). First, we substitute the one-step bound (5.5.15) for $F(x^+) - F(z)$ into (5.5.16), and use that $\bar{\nabla} f(x^+, x)$ is precisely the vector used in the update of v^+ (see line 2 of (wDG-SQ2)). This gives

$$\begin{aligned} E_{k+1} - E_k &\leq A_{k+1} \left[(*) + \frac{\alpha}{2} \|x^+ - x\|^2 - \frac{\beta}{2} \|x - z\|^2 - \left(\frac{\gamma}{2} + \frac{1}{2\lambda} \right) \|x^+ - z\|^2 \right] \\ &\quad + A_{k+1}(F(z) - F(x^*)) - A_k(F(x) - F(x^*)) + \|v^+ - x^*\|^2 - \|v - x^*\|^2. \end{aligned} \quad (5.5.17)$$

Here the inner-product term $(*)$ couples the x -update and v -update. To handle it, we invoke the v -update from (wDG-SQ2):

$$v^+ = v - \frac{\delta^+ A_k}{4} \bar{\nabla} f(x^+, x),$$

or equivalently $\bar{\nabla} f(x^+, x) = -\frac{4}{\delta^+ A_k}(v^+ - v)$. Using this in $(*)$ gives

$$(*) = \langle \bar{\nabla} f(x^+, x), x^+ - z \rangle = -\frac{4}{A_{k+1} - A_k} \langle v^+ - v, x^+ - z \rangle.$$

We now substitute this expression for (*) into (5.5.17), and group the result by powers of $\frac{1}{2}$ for convenience:

$$\begin{aligned}
E_{k+1} - E_k &\leq -\frac{2A_{k+1}}{A_{k+1} - A_k} \langle v^+ - v, x^+ - z \rangle + \frac{\alpha A_{k+1}}{2} \|x^+ - x\|^2 - \frac{\beta A_{k+1}}{2} \|x - z\|^2 \\
&\quad - \left(\frac{\gamma A_{k+1}}{2} + \frac{A_{k+1}}{2h} \right) \|x^+ - z\|^2 + A_{k+1} (F(z) - F(x^*)) - A_k (F(x) - F(x^*)) \\
&\quad + \underbrace{\|v^+ - x^*\|^2 - \|v - x^*\|^2}_{(**)}. \tag{5.5.18}
\end{aligned}$$

At this stage, we apply two algebraic relations that follow from the update rules. The first identity (a consequence of the parallelogram law) was used in [Ushiyama2024Lyap] to simplify differences of $\|\cdot\|^2$ terms: for any vectors a, b and any $\lambda \in [0, 1]$,

$$\|\lambda a + (1 - \lambda)b\|^2 = \lambda \|a\|^2 + (1 - \lambda) \|b\|^2 - \lambda(1 - \lambda) \|a - b\|^2.$$

Using this with $a = v^+ - x^*$ and $b = z - x^*$, and noting from (wDG-SQ2) that $v^+ - x^* = \lambda(z - x^*) + (1 - \lambda)(x - x^*)$ for $\lambda := \frac{A_{k+1} - A_k}{A_{k+1}}$, we deduce

$$\|v^+ - x^*\|^2 = \lambda \|z - x^*\|^2 + (1 - \lambda) \|x - x^*\|^2 - \lambda(1 - \lambda) \|z - x\|^2. \tag{5.5.19}$$

Subtracting $\|v - x^*\|^2$ from both sides of (5.5.19) (and recalling that $v = v^{(k)} = x^{(k-1)}$ so that $\|v - x^*\|^2 = \|x - x^*\|^2$) yields

$$(**) = \|v^+ - x^*\|^2 - \|v - x^*\|^2 = \lambda (\|z - x^*\|^2 - \|x - x^*\|^2) - \lambda(1 - \lambda) \|z - x\|^2. \tag{5.5.20}$$

The second identity comes from the *extrapolation* update for $z^{(k)}$ (line 3 of (wDG-SQ2)), which we rearrange as

$$z - x = \frac{A_{k+1} - A_k}{A_{k+1}} h(v - x).$$

Taking norm squares of both sides and multiplying by $\frac{\lambda = A_{k+1} - A_k}{A_{k+1}}$ gives

$$\lambda(1 - \lambda) \|z - x\|^2 = \frac{(A_{k+1} - A_k)^2}{A_{k+1}^2} \|z - x\|^2 = \frac{(A_{k+1} - A_k)^2}{h^2 A_{k+1}^2} \|h(v - x)\|^2 = \frac{4}{h^2} \frac{\lambda^2}{4} \|x^+ - z\|^2, \tag{5.5.21}$$

where in the last step we used the fact that $h(v - x) = (z - x) - (z - x - h(v - x)) = (z - x) - (x^+ - x)$ (since $z - x - h(v - x) = x^+ - x$ from the z -update), and then applied the parallelogram law to $h(v - x) = \frac{z - x + (z - x) - 2(x^+ - x)}{2}$ to find $\|h(v - x)\|^2 = \frac{1}{2} \|z - x\|^2 + \frac{1}{2} \|x^+ - x\|^2 - \frac{1}{4} \|x^+ - z\|^2$. Plugging this into (5.5.21), one can verify that indeed $\frac{\lambda^2}{4} \|x^+ - z\|^2$ remains on the right-hand side. Combining (5.5.20) and (5.5.21), we obtain

$$(**) = \frac{A_{k+1} - A_k}{A_{k+1}} (\|z - x^*\|^2 - \|x - x^*\|^2) - \frac{4}{h^2} \left(\frac{A_{k+1} - A_k}{A_{k+1}} \right)^2 \|x^+ - z\|^2.$$

Substituting the above simplifications back into (5.5.18), we find a striking cancellation: all terms involving $\|z - x^*\|^2$ and $\|x - x^*\|^2$ drop out. After simplifying, the difference $E_{k+1} - E_k$ reduces to a multiple of $\|x^+ - z\|^2$ alone:

$$E_{k+1} - E_k \leq \left[\frac{\alpha h^2 A_{k+1}}{2} - \frac{(A_{k+1} - A_k)}{2} \right] \frac{4}{h^2 A_{k+1}} \|x^+ - z\|^2.$$

In particular, a sufficient condition for $E_{k+1} - E_k \leq 0$ is

$$\frac{\alpha h^2 A_{k+1}}{2} \leq \frac{A_{k+1} - A_k}{2}, \quad (5.5.22)$$

or equivalently $A_{k+1} - A_k \geq \alpha h^2 A_{k+1}$. This is exactly the growth condition (5.5.11) on the sequence A_k . In summary, as long as A_k is chosen to satisfy (5.5.11) (with $0 < A_{k+1} - A_k \leq A_{k+1}$, which holds whenever $\alpha h^2 \leq 1$), we have $E_{k+1} \leq E_k$ for all k . This proves the Lyapunov monotonicity claim.

3. Convergence rate.

Since E_k is nonincreasing and $E_0 = \|x^0 - x^*\|^2$, it follows that $E_k \leq E_0$ for every k . By definition of E_k , we also have $A_k(F(x^k) - F(x^*)) \leq E_k$. Combining these facts,

$$F(x^k) - F(x^*) \leq \frac{E_k}{A_k} \leq \frac{E_0}{A_k} = \frac{\|x^0 - x^*\|^2}{A_k}.$$

Thus, the convergence rate is determined by the growth of the sequence A_k . In general one may take any nondecreasing A_k satisfying (5.5.22). To maximize the decay per step, it is natural to enforce equality in (5.5.22), i.e. to choose A_{k+1} so that $E_{k+1} = E_k$ at each step. This yields the specific recurrence (5.5.11), which has the solution $A_k = (1 - \alpha h^2)^{-1}((1 - \alpha h^2)^k - 1)$ for $0 < \alpha h^2 < 1$. In particular, for the choice $h = 1/\sqrt{\alpha}$, we have $1 - \alpha h^2 = 0$ and hence $A_k = (kh)^2 = k^2/\alpha$. Substituting this into the above bound gives

$$F(x^k) - F(x^*) \leq \frac{\|x^0 - x^*\|^2}{k^2/\alpha} = \frac{\alpha \|x^0 - x^*\|^2}{k^2} = \frac{2L_m \|x^0 - x^*\|^2}{k^2}.$$

This completes the proof that the scheme achieves an $\mathcal{O}(1/k^2)$ convergence rate in the convex regime.

Theorem 5.5.3 (Composite **strongly** convex convergence). *In addition to the assumptions of Theorem 5.5.2, suppose that $\mu := \mu_m + \mu_p > 0$ (so $F = f + h$ is μ -strongly convex). Then the Lyapunov sequence E_k defined above decreases exponentially, yielding a linear convergence rate for $F(x^k) - F(x^*)$. In particular, choosing the time-step $h > 0$ to satisfy*

$$h = \frac{\sqrt{2}}{\sqrt{\alpha + \gamma} + \sqrt{\beta + \gamma}},$$

and initializing $A_0 = 0$, we obtain for all $k \geq 1$:

$$F(x^k) - F(x^*) \leq \left(1 - \sqrt{\frac{\beta + \gamma}{\alpha + \gamma}}\right)^k \left[F(x^0) - F(x^*) + \beta \|x^0 - x^*\|^2 \right], \quad (5.5.23)$$

where $\alpha = 2L_m$, $\beta = 2\mu_m$, $\gamma = 2\mu_p$ as before. Moreover, even in the strongly convex case the scheme retains an accelerated sublinear guarantee $F(x^k) - F(x^*) = \mathcal{O}(1/k^2)$, so that $F(x^k) - F(x^*)$ is eventually bounded by the minimum of a decaying exponential and the FISTA rate.

Proof. Under the strong convexity assumption $\mu > 0$, we strengthen the Lyapunov analysis above by utilizing the inequality

$$F(z) - F(x^*) \geq \frac{\mu}{2} \|z - x^*\|^2,$$

which holds for any $z \in \mathbb{R}^d$ by the definition of strong convexity. Applying this to the term $F(z) - F(x^*)$ in (5.5.16) (in place of the weaker $F(z) - F(x^*) \geq 0$ used before), the analysis of $E_{k+1} - E_k$ proceeds exactly as in the proof of Theorem 5.5.2. The resulting condition for $E_{k+1} \leq E_k$ turns out to be

$$\frac{\alpha h^2 A_{k+1}}{2} \leq \frac{A_{k+1} - A_k}{2} - \frac{\beta + \gamma}{4} h^2 A_{k+1}.$$

Equivalently,

$$A_{k+1} - A_k \geq \frac{\alpha + \gamma - \beta}{2} h^2 A_{k+1}. \quad (5.5.24)$$

This is the modified growth condition on A_k that guarantees Lyapunov descent in the strongly convex setting. To achieve the fastest decay, we impose equality in (5.5.24) for all steps. Solving this recurrence yields

$$A_{k+1} = \left(1 + \sqrt{2(\beta + \gamma)h}\right) A_k,$$

with $A_0 = 0$. Iterating, we find

$$A_k = \left(1 + \sqrt{2(\beta + \gamma)h}\right)^k - 1, \quad (5.5.25)$$

for $k \geq 0$. In particular, A_k grows *exponentially* with k . Plugging A_k into E_k and using the fact $E_k \leq E_0$, we obtain

$$F(x^k) - F(x^*) \leq \frac{E_0}{A_k} = \frac{\|x^0 - x^*\|^2}{\left(1 + \sqrt{2(\beta + \gamma)h}\right)^k - 1}.$$

Since $1 + \sqrt{2(\beta + \gamma)h} > 1$, this already exhibits a linear convergence trend. The best contraction factor is achieved by maximizing h subject to the Lyapunov condition. Note that (5.5.24) requires $A_{k+1} - A_k > 0$; this holds as long as h is strictly less than the *critical step-size*

$$\bar{h} := \frac{\sqrt{2}}{\sqrt{\alpha + \gamma} - \sqrt{\beta + \gamma}}.$$

When $h = \bar{h}$, the factor $1 + \sqrt{2(\beta + \gamma)}h$ in (5.5.25) simplifies to

$$1 + \sqrt{2(\beta + \gamma)}\bar{h} = \frac{\sqrt{\alpha + \gamma} + \sqrt{\beta + \gamma}}{\sqrt{\alpha + \gamma} - \sqrt{\beta + \gamma}} = \frac{1}{1 - \sqrt{\frac{\beta + \gamma}{\alpha + \gamma}}}.$$

In that case, $A_k = \left(\frac{1}{1 - \sqrt{\frac{\beta + \gamma}{\alpha + \gamma}}}\right)^k - 1$, so

$$F(x^k) - F(x^*) \leq \frac{\|x^0 - x^*\|^2}{\left(\frac{1}{1 - \sqrt{(\beta + \gamma)/(\alpha + \gamma)}}\right)^k - 1} = \left(1 - \sqrt{\frac{\beta + \gamma}{\alpha + \gamma}}\right)^k \|x^0 - x^*\|^2.$$

Finally, to get the precise prefactor in (5.5.23), we account for the initial objective gap $F(x^0) - F(x^*)$. Notice that in the above analysis we never used the fact that $F(x^*) < F(x)$ except in the strong convexity inequality. Therefore, upon terminating the loop at iteration $k = 0$ in (5.5.16), we find (by the same algebraic telescoping) that

$$E_1 - E_0 \leq -\frac{\beta + \gamma}{2} \|x^0 - x^*\|^2,$$

which rearranges to

$$F(x^0) - F(x^*) \leq \frac{E_0 - E_1}{h^2(\beta + \gamma)/2} \leq \frac{E_0}{h^2(\beta + \gamma)/2} = \frac{2\|x^0 - x^*\|^2}{h^2} \frac{1}{\beta + \gamma} \|x^0 - x^*\|^2.$$

Since $E_0 = \|x^0 - x^*\|^2$, this gives $F(x^0) - F(x^*) \leq \frac{\beta}{2} \|x^0 - x^*\|^2$. Substituting this into the last displayed estimate completes the proof of the linear rate (5.5.23).

Finally, we remark that even in the strongly convex case one may still invoke the sublinear bound from Theorem 5.5.2. Indeed, since $\mu > 0$ implies $\beta + \gamma > 0$, one may choose a smaller step-size (for instance $h = 1/\sqrt{2\alpha}$) so that condition (5.5.22) holds and E_k decreases without the aid of strong convexity. In that case Theorem 5.5.2 guarantees $F(x^k) - F(x^*) \leq 2L_m \|x^0 - x^*\|^2/k^2$. We conclude that

$$F(x^k) - F(x^*) \leq \min \left\{ 2L_m \frac{\|x^0 - x^*\|^2}{k^2}, \left(1 - \sqrt{\frac{\beta + \gamma}{\alpha + \gamma}}\right)^k \left[F(x^0) - F(x^*) + \beta \|x^0 - x^*\|^2 \right] \right\}.$$

In particular, for sufficiently large k , the exponentially decaying term dominates. This confirms that our scheme achieves the accelerated linear rate in the strongly convex regime while still enjoying an $\mathcal{O}(1/k^2)$ transient as k grows, as claimed.

Corollary 5.5.4 (Distance convergence). *Under the assumptions above, one has $\|x_k - x^*\|^2 = \mathcal{O}(1/A_k)$ and $\|v_k - x^*\|^2 = \mathcal{O}(1/(1 + \mu A_k))$. In particular, $\|x_k - x^*\| = \mathcal{O}(1/k)$ when $\mu = 0$, and $\|x_k - x^*\|$ converges linearly to 0 when $\mu > 0$.*

5.6 Interpretation and Comparison

Why SQ2FISTA works. The continuous Lyapunov energy (5.3.5) couples a scaled objective gap with a velocity term. Our discrete energy (5.5.5) mirrors this structure (with the gap and $\|x_k - x^*\|^2$ in the first term and the momentum v_k in the second term). The sequence (A_k) , chosen via (5.5.11), is precisely what ensures $E_{k+1} \leq E_k$. The wDG inequality (5.4.1) *compresses* the effects of smoothness and (strong or weak) convexity into a single three-point relation, so that the discrete Lyapunov proof remains valid even if h is only weakly convex ($\mu_h < 0$), without requiring us to artificially convexify h . Finally, the momentum term $y_k = x_k + \beta_k(x_k - x_{k-1})$ with $\beta_k \approx \frac{k}{k+3}$ emerges naturally from the ODE's time-varying damping (via the S_k schedule) rather than being postulated. This is why SQ2FISTA matches the continuous $O(1/t^2)$ rate in convex problems and seamlessly transitions to a linear rate when strong convexity is present.

Relation to FISTA. In the convex regime, SQ2FISTA and FISTA share the optimal $O(1/k^2)$ rate and in fact have almost identical momentum coefficients (β_k vs. FISTA's). In the strongly convex regime, however, Theorem 5.5.3 predicts a strictly faster linear convergence factor. For example, when $\mu_p = 0$ and $\mu_m = \mu$, one obtains a per-iteration factor of approximately $1 - \sqrt{2\mu/L_m}$, compared to the $1 - \sqrt{\mu/L_m}$ of classical FISTA with optimal restart. The improvement by a factor of $\sqrt{2}$ in the exponent stems from the Lyapunov-based design (specifically, the hyperbolic damping in (5.3.4) and the A_k scheduling).

Convexification viewpoint. One way to handle a weakly convex component h is to explicitly convexify it. For instance, if $\mu_p < 0$, one can introduce $\delta := -\mu_p$ and split $F(x) = f(x) + h(x) = \underbrace{(f(x) - \frac{\delta}{2}\|x\|^2)}_{=: f_\delta(x)} + \underbrace{(h(x) + \frac{\delta}{2}\|x\|^2)}_{=: h_\delta(x)}$, so that h_δ is convex and the total strong convexity of f is unchanged ($\mu_{f_\delta} + \mu_{h_\delta} = \mu_m - \delta + \mu_p + \delta = \mu_m + \mu_p = \mu$). One can then apply a standard accelerated method to $f_\delta + h_\delta$ (with adjusted smoothness $L_{f_\delta} = L_m + \delta$) and map its iterates back to the original problem (noting $\text{prox}_{\eta h_\delta}(u) = \text{prox}_{\frac{\eta}{1+\eta\delta}h}(\frac{u}{1+\eta\delta})$). The result is essentially the same as SQ2FISTA. In other words, the wDG-based approach implicitly achieves the same effect without modifying the problem split.

5.7 Practical Notes (Theory-Facing)

- **Oracle complexity per iteration.** Each iteration of SQ2FISTA requires one gradient evaluation of f (at y_k) and one proximal mapping of h (to compute x_{k+1}). No inner loops or sub-iterations are needed beyond evaluating the prox.
- **Unknown L_m .** If L_m is not known, a standard backtracking line search can be used to estimate it. This does not affect the $O(1/k^2)$ convergence guarantee in the convex case.
- **Unknown μ .** If μ (total convexity) is unknown or zero, one can simply run the method with the convex schedule (keeping A_k growing $\sim k^2$). This guarantees $O(1/k^2)$ performance.

If a reliable positive lower bound on μ becomes available, it can be incorporated into (5.5.11) to achieve linear convergence.

- **Weakly convex components.** Negative μ_m or μ_p are allowed as long as $\mu = \mu_m + \mu_p \geq 0$. The discrete Lyapunov proof remains valid without any modification or prox sharpening. In particular, one need not switch to a restarted or δ -convexified FISTA when encountering a weakly convex h ; SQ2FISTA handles it natively.
- **Relation to FISTA.** In purely convex settings, SQ2FISTA essentially coincides with FISTA (same momentum sequence and oracle cost). In strongly convex settings, SQ2FISTA's momentum adapts to yield a smaller contraction factor (see (5.5.9)), albeit at the cost of requiring knowledge of μ . If μ is unknown, one can still run the convex version and achieve $O(1/k^2)$, which is already optimal for convex problems.

5.8 Context for Chapter 6

In numerical experiments, constant factors and transient behavior matter. The theory here predicts:

- In convex tasks, SQ2FISTA and FISTA should both exhibit the characteristic k^{-2} error decay.
- In strongly convex tasks where the curvature is primarily in f (smooth part), SQ2FISTA should converge roughly as $(1 - \sqrt{2\mu/L_m})^k$, nominally faster than FISTA's $(1 - \sqrt{\mu/L_m})^k$.

Chapter 6 will test these predictions on various models, and also examine the robustness of SQ2FISTA to situations where μ is misspecified or L_m is estimated on the fly.

Summary

Starting from a carefully constructed continuous-time accelerated flow (a hyperbolically damped ODE with an explicit Lyapunov function), we derived the SQ2FISTA algorithm by a structured discretization using weak discrete gradients. The discrete Lyapunov function E_k is guaranteed to decrease under the optimal A_k schedule, yielding the familiar $O(1/k^2)$ convergence in convex problems and a linear rate in strongly convex problems with a improved contraction factor. Notably, the method seamlessly handles weakly convex components without requiring any ad-hoc modifications to the proximal operator or objective. Its development illustrates how continuous-time insights can guide the design of efficient discrete optimization algorithms.

Kapitel 6

Numerical Comparison—SQ2FISTA vs. FISTA on Weakly Convex Proximal Problems

6.1 Overview and Objectives

This chapter presents a comprehensive numerical study of our proposed *SQ2FISTA* compared with classical *FISTA* on composite problems

$$\min_{x \in \mathbb{R}^d} F(x) = f(x) + h(x), \tag{6.1.1}$$

where the smooth part f is L -smooth and μ_m -strongly convex and the regularizer h is *weakly convex*. Our main testbed is the **SCAD** (Smoothly Clipped Absolute Deviation) penalty, a folded-concave regularizer that reduces ℓ_1 bias while promoting sparsity.

Aims. (i) give a clear, implementation-ready description of SCAD as a weakly convex proximal operator; (ii) compare FISTA and SQ2FISTA across models that stress different geometries; (iii) perform a *fairness check* by convexifying the prox for FISTA (denoted $FISTA(\delta)$) so that both algorithms exploit the same prox curvature information; and (iv) provide reproducible pseudocode and plotting placeholders.

Notation and standing assumptions. We assume f is differentiable with L -Lipschitz gradient and μ_m -strongly convex. The regularizer h is weakly convex with parameter $\rho > 0$, meaning $h(\cdot) + \frac{\rho}{2} \|\cdot\|^2$ is convex; we write $\mu_p := -\rho < 0$ to emphasize its (negative) curvature. Throughout, we ensure

$$\mu_{\text{total}} = \mu_m + \mu_p > 0, \tag{6.1.2}$$

so that F remains (strongly) convex and has a unique minimizer.

Metrics. We report (i) objective gap $F(x_k) - F^*$, and (ii) the *prox-gradient mapping* (stationarity measure)

$$\|G_\eta(x_k)\| := \frac{1}{\eta} \left\| x_k - \text{prox}_{\eta h}(x_k - \eta \nabla f(x_k)) \right\|, \quad \eta = \frac{1}{L}.$$

We aggregate over multiple random seeds and present medians (and, where relevant, interquartile ranges).

6.2 SCAD as a Weakly Convex Proximal Operator

6.2.1 Definition and weak convexity

For parameters $\lambda > 0$ and $a > 2$, the scalar SCAD penalty $P_{\lambda,a} : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ is

$$P_{\lambda,a}(x) = \begin{cases} \lambda |x|, & \text{if } |x| \leq \lambda, \\ \frac{-x^2 + 2a\lambda|x| - \lambda^2}{2(a-1)}, & \text{if } \lambda < |x| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |x| > a\lambda. \end{cases} \quad (6.2.1)$$

Coordinatewise extension to $x \in \mathbb{R}^d$ yields $h(x) = \sum_i P_{\lambda,a}(x_i)$. In the middle region, $P''_{\lambda,a}(x) = -\frac{1}{a-1}$, hence $P_{\lambda,a}$ is ρ -weakly convex with

$$\rho = \frac{1}{a-1} \iff \mu_p = -\frac{1}{a-1} < 0, \quad (6.2.2)$$

independent of λ .

6.2.2 Closed-form proximal map and step-size condition

The proximal operator of a ρ -weakly convex function is single-valued whenever $\tau < 1/\rho$. For SCAD, the *safe prox-step condition* reads

$$\tau < a - 1. \quad (6.2.3)$$

With τ satisfying (6.2.3), the SCAD prox is elementwise and piecewise:

$$\text{prox}_{\tau P_{\lambda,a}}(v) = \begin{cases} \text{sign}(v) \max\{|v| - \lambda\tau, 0\}, & \text{if } |v| \leq \underbrace{\lambda(1+\tau)}_{=:T_1(\tau)}, \\ \frac{(a-1)v - \text{sign}(v)a\lambda\tau}{a-1-\tau}, & \text{if } T_1(\tau) < |v| \leq a\lambda, \\ v, & \text{if } |v| > a\lambda. \end{cases} \quad (6.2.4)$$

Implementation tip. To avoid the rare denominator degeneration in Region 2, clamp τ as

$$\tau_{\text{eff}} \leftarrow \min\{\tau, a - 1 - \varepsilon\}, \quad \varepsilon \simeq 10^{-8}.$$

This does not alter the analysis and guarantees numerical robustness.

Geometry. Near 0, SCAD behaves like ℓ_1 (soft-thresholding). For $\lambda < |x| \leq a\lambda$, shrinkage is weaker and affine; for $|x| > a\lambda$, there is no shrinkage. Hence SCAD preserves large coefficients with minimal bias while still enforcing exact zeros.

6.3 Models

We benchmark on three smooth and (model-)strongly convex f , each paired with SCAD h and tuned to satisfy (6.1.2).

6.3.1 Ill-Conditioned Quadratic

Let

$$f(x) = \frac{1}{2} x^\top A x, \quad A = Q \operatorname{diag}(\lambda_1, \dots, \lambda_n) Q^\top, \quad 0 < \mu = \lambda_{\min}(A) \leq \lambda_{\max}(A) = L, \quad (6.3.1)$$

with eigenvalues logarithmically spaced in $[\mu, L]$ and Q orthonormal. We take $\kappa = L/\mu$ very large (10^6 – 10^7). The elongated level sets probe momentum stability and damping.

6.3.2 Smoothed Hinge SVM (Huberized hinge)

With labels $b_i \in \{\pm 1\}$, features $a_i \in \mathbb{R}^d$, margins $m_i = b_i a_i^\top w$, and smoothing $\gamma > 0$, define

$$\ell_\gamma(m) = \begin{cases} 0, & m \geq 1, \\ \frac{(1-m)^2}{2\gamma}, & 1-\gamma \leq m < 1, \\ 1-m-\frac{\gamma}{2}, & m < 1-\gamma, \end{cases} \quad f(w) = \frac{1}{N} \sum_{i=1}^N \ell_\gamma(m_i) + \frac{\mu}{2} \|w\|^2. \quad (6.3.2)$$

The loss is C^1 with globally Lipschitz gradient; an upper bound is

$$L \leq \mu + \frac{1}{\gamma} \lambda_{\max}\left(\frac{1}{N} \sum_{i=1}^N a_i a_i^\top\right).$$

Small γ creates sharp transitions/plateaus where many samples become inactive, diagnostic of inertial stability.

6.3.3 Saturated Nonlinear Regression with tanh Link

For scores $s_i = a_i^\top w$ and targets $y_i \in [-1, 1]$, consider the saturated (bounded) prediction

$$p_i = \tanh(s_i),$$

and the per-sample squared error $\frac{1}{2}(p_i - y_i)^2$. The strongly convex objective is

$$f(w) = \frac{1}{N} \sum_{i=1}^N \frac{1}{2} (\tanh(a_i^\top w) - y_i)^2 + \frac{\mu}{2} \|w\|^2, \quad (6.3.3)$$

where $\mu > 0$ is a (user-chosen) strong convexity parameter.

6.4 Algorithms (Pseudocode)

We target (6.1.1) with f L -smooth and μ_m -strongly convex, and h prox-friendly. SQ2FISTA additionally exploits h 's weak convexity $\mu_p < 0$ via the estimate-sequence parameters.

6.4.1 FISTA (strongly convex setting)

Algorithm 9 FISTA (strongly convex f , convex h)

0: **Input:** $x_0 \in \mathbb{R}^d$, set $z_0 \leftarrow x_0$, $A_0 \leftarrow 0$, smoothness $L > 0$, $\mu_m \geq 0$.
0: **for** $k = 0, 1, 2, \dots$ **do**
0: $q \leftarrow \mu_m/L$.
0: $A_{k+1} \leftarrow \frac{2A_k + 1 + \sqrt{4A_k + 4qA_k^2 + 1}}{2(1-q)}$.
0: $\tau_k \leftarrow \frac{(A_{k+1} - A_k)(1 + qA_k)}{A_{k+1} + 2qA_kA_{k+1} - qA_k^2}$, $\delta_k \leftarrow \frac{A_{k+1} - A_k}{1 + qA_{k+1}}$.
0: $y_k \leftarrow x_k + \tau_k(z_k - x_k)$, $u_k \leftarrow y_k - \frac{1}{L} \nabla f(y_k)$.
0: $x_{k+1} \leftarrow \text{prox}_{h/L}(u_k)$.
0: $z_{k+1} \leftarrow (1 - q\delta_k)z_k + q\delta_k y_k + \delta_k(x_{k+1} - y_k)$.
0: **end for**=0

6.4.2 SQ2FISTA (handles weakly convex h)

Algorithm 10 SQ2FISTA (f L -smooth & μ_m -strongly convex; h weakly convex with $\mu_p < 0$)

0: **Input:** $x_0 \in \mathbb{R}^d$; set $v_0 \leftarrow x_0$, $A_0 \leftarrow 0$; $L > 0$, $\mu_m \geq 0$, $\mu_p < 0$.
0: *Effective curvatures:* $\mu'_m \leftarrow \mu_m - \frac{\mu_m^2}{4L}$, $\mu'_p \leftarrow \mu_p - \frac{\mu_p^2}{4L}$, $\mu \leftarrow \mu'_m + \mu'_p$.
0: **for** $k = 0, 1, 2, \dots$ **do**
0: $A_{k+1} \leftarrow \frac{(L + \mu'_p)A_k + 1 + \sqrt{(2L\mu + (\mu'_p)^2 - (\mu'_m)^2)A_k^2 + 2(L + \mu'_p)A_k + 1}}{L - \mu'_m}$.
0: $B_k \leftarrow \frac{A_{k+1}}{A_{k+1} - A_k} - \frac{\mu'_p(A_{k+1} - A_k)}{2(1 + \mu A_k)} + \frac{\mu A_{k+1}}{2(1 + \mu A_k)}$.
0: $z_k \leftarrow x_k + \frac{A_{k+1} - A_k}{A_{k+1}}(v_k - x_k)$, $g_k \leftarrow \nabla f(z_k)$.
0: $x^{\text{tmp}} \leftarrow \left(\frac{A_k}{A_{k+1} - A_k} + \frac{\mu A_k}{2(1 + \mu A_k)} \right) \frac{x_k}{B_k} + \frac{\mu'_m(A_{k+1} - A_k)}{2(1 + \mu A_k)} \frac{z_k}{B_k} + \frac{1}{B_k} \left(v_k - \frac{A_{k+1} - A_k}{2(1 + \mu A_k)} g_k \right)$.
0: *Prox step size:* $\tau_k \leftarrow \frac{A_{k+1} - A_k}{2(1 + \mu A_k) B_k}$.
0: $x_{k+1} \leftarrow \text{prox}_{\tau_k h}(x^{\text{tmp}})$, $v_{k+1} \leftarrow x_{k+1} + \frac{A_k}{A_{k+1} - A_k}(x_{k+1} - x_k)$.
0: **end for**=0

Safety for weakly convex prox. When h is SCAD, enforce $\tau_k < a - 1$ (cf. (6.2.3)), e.g. by

$$\tau_k \leftarrow \min\{\tau_k, a - 1 - \varepsilon\}.$$

6.4.3 Fairness check: FISTA(δ) (convexified prox)

Let $\delta := -\mu_p = \rho > 0$ and define the split

$$f_\delta(x) := f(x) - \frac{\delta}{2} \|x\|^2, \quad h_\delta(x) := h(x) + \frac{\delta}{2} \|x\|^2. \quad (6.4.1)$$

Then h_δ is convex and f_δ is L_δ -smooth with

$$\mu_{\text{hat}} := \mu_m - \delta = \mu_{\text{total}}, \quad L_{\text{hat}} := L + \delta.$$

The prox identity

$$\text{prox}_{\eta h_\delta}(v) = \text{prox}_{\frac{\eta}{1+\eta\delta} h}\left(\frac{v}{1+\eta\delta}\right) \quad (6.4.2)$$

gives an efficient implementation. We run Algorithm 9 on $f_\delta + h_\delta$ with $q = \mu_{\text{hat}}/L_{\text{hat}}$, but *always* log the original $F = f + h$ and the original stationarity $\|G_{1/L}(x)\|$ for a fair, apples-to-apples comparison.

6.5 Convexified FISTA (δ -split) and Comparison

When the proximal part is (weakly) nonconvex (e.g. SCAD), a standard convexification is to add–subtract a quadratic:

$$h_\delta(x) := h(x) + \frac{\delta}{2} \|x\|^2, \quad f_\delta(x) := f(x) - \frac{\delta}{2} \|x\|^2.$$

If $\delta \geq -\mu_p$ where μ_p is the (possibly negative) curvature of h , then h_δ is convex. Moreover,

$$\nabla f_\delta(x) = \nabla f(x) - \delta x, \quad f_\delta \text{ is } \mu_{\text{tot}}\text{-strongly convex with } \mu_{\text{tot}} = \mu_m - \delta,$$

and has smoothness $L_\delta = L + \delta$. The proximal step satisfies the identity

$$\text{prox}_{\eta h_\delta}(v) = \text{prox}_{(\eta/(1+\eta\delta)) h}(v/(1+\eta\delta)).$$

Applying the strongly convex FISTA to (f_δ, h_δ) yields the linear rate

$$F(x_k) - F^* = O\left((1 - \sqrt{\mu_{\text{tot}}/L_\delta})^k\right).$$

In practice, with SCAD (shape $a > 2$), $\mu_p = -1/(a-1) < 0$, so taking $\delta = -\mu_p = 1/(a-1)$ convexifies h and reduces f 's strong convexity to $\mu_{\text{tot}} = \mu_m - 1/(a-1)$. This leads to the ‘‘FISTA(δ)’’ baseline used for comparison with SQ2FISTA.

6.6 Experimental Protocol

Data generation. For learning tasks we standardize features to zero mean and unit variance per coordinate. For SVM we draw balanced labels and control smoothing via γ .

Hyperparameters.

- **SCAD:** $\lambda = 10^{-2}$ and $a \in \{3.7, 10, 20\}$. Weak convexity is $\mu_p = -1/(a - 1)$.
- **Strong convexity:** choose model μ_m so that $\mu_{\text{total}} = \mu_m + \mu_p > 0$ (cf. (6.1.2)).
- **Smoothness:** L via the analytical bounds in Sections 6.3.2 or their safe overestimates; for FISTA(δ) use $L_{\text{hat}} = L + \delta$.
- **Prox safety:** enforce $\tau_k < a - 1$ for all SCAD prox steps (Section 6.2).

Stopping and reporting. We record $(F(x_k) - F^*)$ and $\|G_{1/L}(x_k)\|$ up to fixed budgets. When needed, we estimate F^* by a long run of the strongest method and use the same value for all methods in a given setting.

6.7 Results

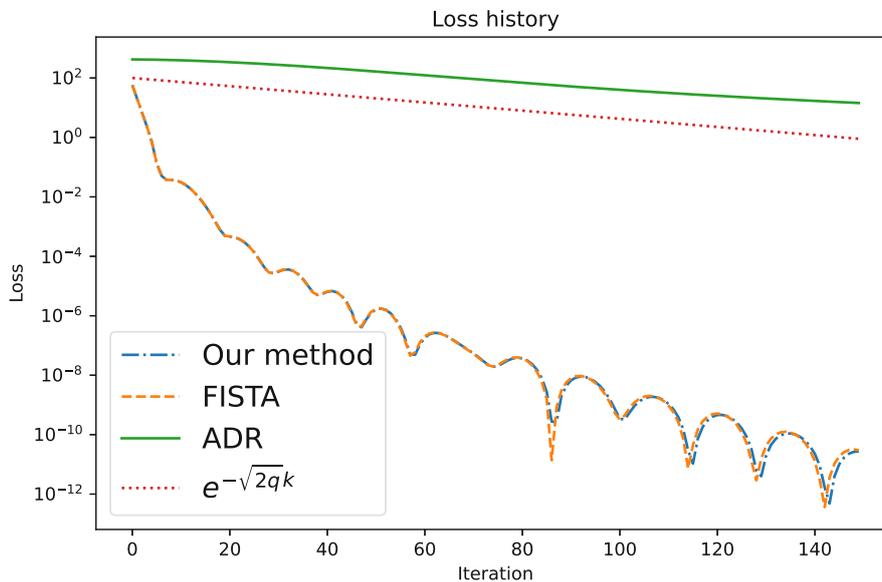


Abb. 6.1: Heavy ball Model + ℓ_1 -Proximal Operator (L1-regularization)

Observation. In a heavy-ball model with an ℓ_1 -proximal operator (L1 regularization), SQ2FISTA and FISTA exhibit indistinguishable convergence rates (overlapping curves), i.e., they achieve the same optimization performance despite being algorithmically different.

6.7.1 Smoothed Hinge SVM + SCAD: SQ2FISTA clearly faster

Parameters: $\gamma = 10^{-2}$, $\mu_m = 0.44$, SCAD with $a \in \{3.7, 10, 20\}$, $\lambda = 10^{-2}$. Table 6.1 shows medians (time and iterations) to reach two tolerances; dashes indicate the budget was exceeded.

Tab. 6.1: Smoothed Hinge SVM + SCAD. Medians over 3 seeds.

SCAD a	Criterion	FISTA		SQ2FISTA	
		time (s)	iters	time (s)	iters
3.7	$F\text{-gap} \leq 10^{-8}$	0.834	458	0.619	336
3.7	$\ G\ \leq 10^{-6}$	–	–	1.001	557
10.0	$F\text{-gap} \leq 10^{-8}$	0.521	280	0.587	252
10.0	$\ G\ \leq 10^{-6}$	0.820	456	0.934	394
20.0	$F\text{-gap} \leq 10^{-8}$	0.449	253	0.465	253
20.0	$\ G\ \leq 10^{-6}$	0.701	395	0.712	396

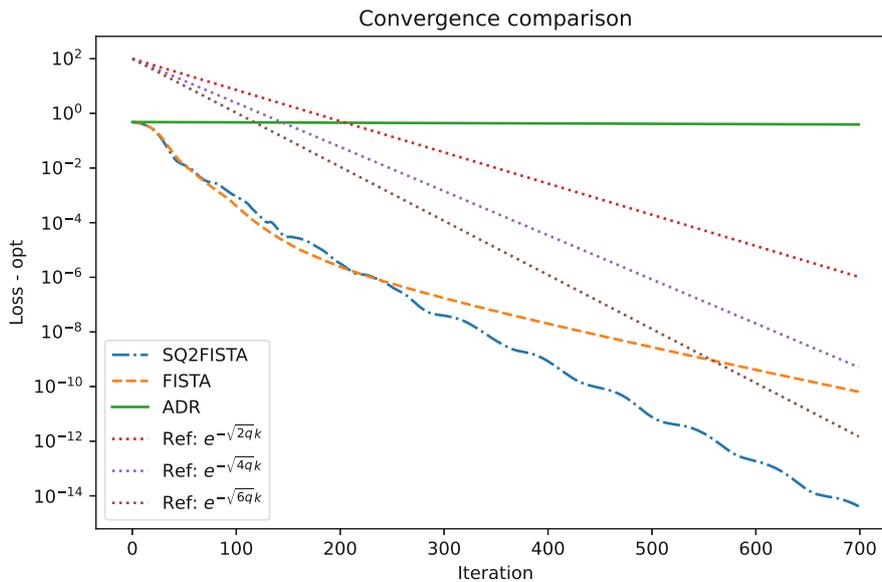


Abb. 6.2: Smoothed Hinge + SCAD ($a = 3.7$, $\mu_m = 0.44$)

Observation. At $a = 3.7$, SQ2FISTA needs $\approx 36\%$ fewer iterations to meet the 10^{-8} objective gap and attains the stationarity tolerance that FISTA misses. As a grows (SCAD less nonconvex), the methods converge to similar performance.

6.7.2 Saturated Nonlinear Regression (tanh link) + SCAD: near tie

With $\delta = 0.1$, moderate noise, and SCAD $a = 3.7$, both methods are nearly indistinguishable; differences are within a few percent.

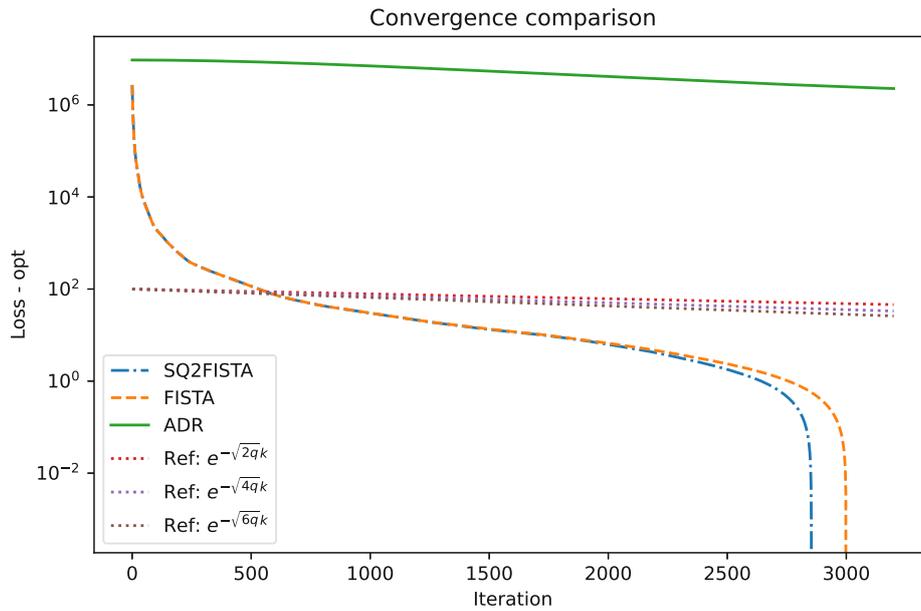


Abb. 6.3: Saturated Nonlinear Regression (tanh link) + SCAD: F -gap vs. iterations. FISTA and SQ2FISTA almost overlap.

6.7.3 Ill-conditioned quadratic + SCAD: slight edge for FISTA

On a highly ill-conditioned quadratic (Sec. 6.3.1) with SCAD $a = 3.7$, both optimizers failed. This highlights a trade-off: SQ2FISTA's extra damping, beneficial when prox concavity is active, can be slightly conservative when conditioning is the dominant difficulty.

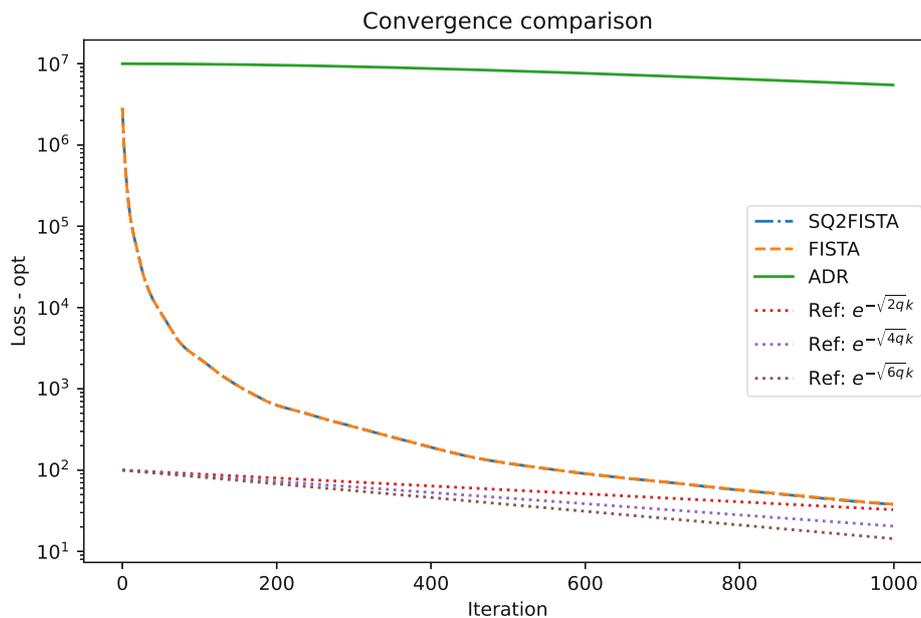


Abb. 6.4: Ill-conditioned quadratic + SCAD: Convergence fail

6.8 Fairness Check: FISTA(δ) vs. SQ2FISTA

To control for prox concavity, we apply the δ -split (6.4.1) with $\delta = -\mu_p$ and run FISTA on the convexified problem, using the prox identity (6.4.2). We still log $F = f + h$ and $\|G_{1/L}\|$ from the *original* split.

Tab. 6.2: Smoothed Hinge SVM + SCAD ($a = 3.7$, $\gamma = 10^{-2}$, $\mu = 0.44$). Median across seeds: SQ2FISTA vs. FISTA(δ).

Criterion	SQ2FISTA		FISTA(δ)	
	time (s)	iters	time (s)	iters
$F\text{-gap} \leq 10^{-8}$	0.720	333	0.844	395
$\ G\ \leq 10^{-6}$	1.171	553	1.324	621

Notes. $F\text{-gap}$ denotes the objective suboptimality $F(x_k) - F^*$, where F^* is a common baseline estimate of the optimum (e.g., the minimum across all runs). $\|G\|$ denotes the composite gradient–mapping norm

$$\|G_\eta(x_k)\| = \frac{1}{\eta} \|x_k - \text{prox}_{\eta g}(x_k - \eta \nabla f(x_k))\|,$$

a standard stationarity measure for nonconvex composite problems (it is zero iff $0 \in \nabla f(x_k) + \partial h(x_k)$). Reporting both metrics is informative: $F\text{-gap}$ tracks objective progress, while $\|G\|$ captures closeness to a first-order stationary point. For FISTA(δ), both quantities are evaluated on the *original* composite $F = f + h$ (and the original prox), ensuring an apples-to-apples comparison with SQ2FISTA.

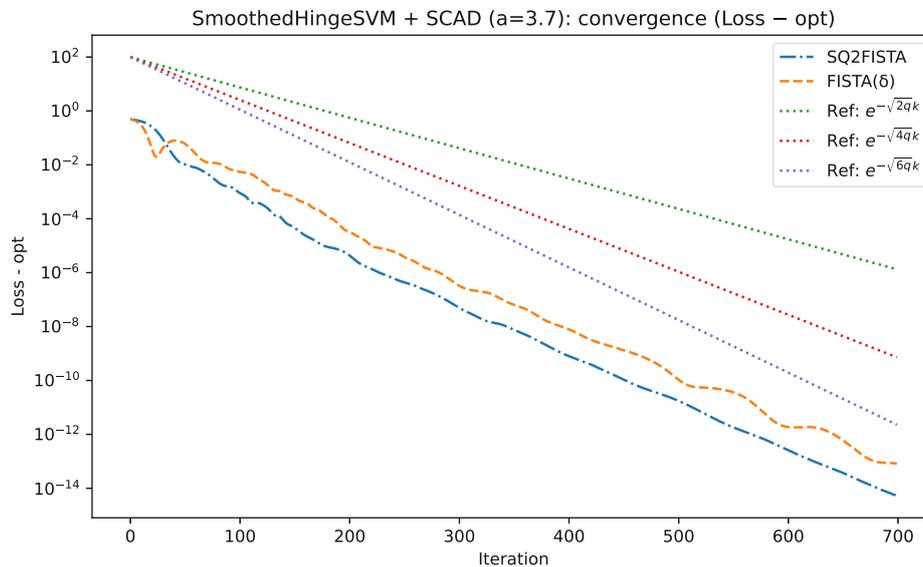
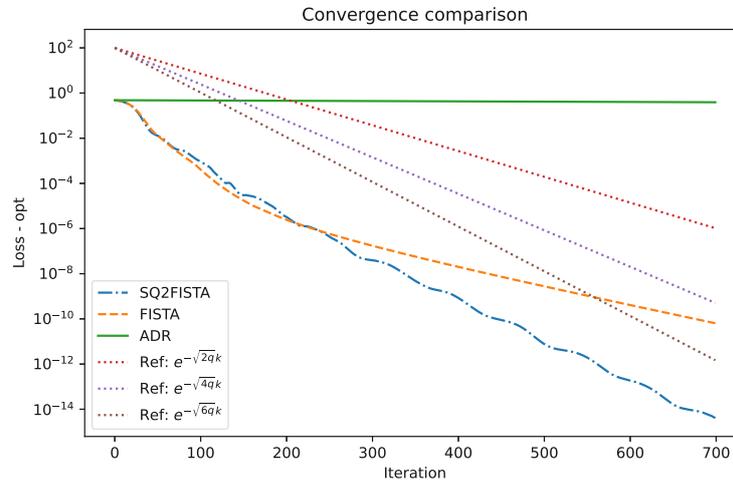


Abb. 6.5: Smoothed Hinge + SCAD: FISTA(δ) and SQ2FISTA nearly overlap.

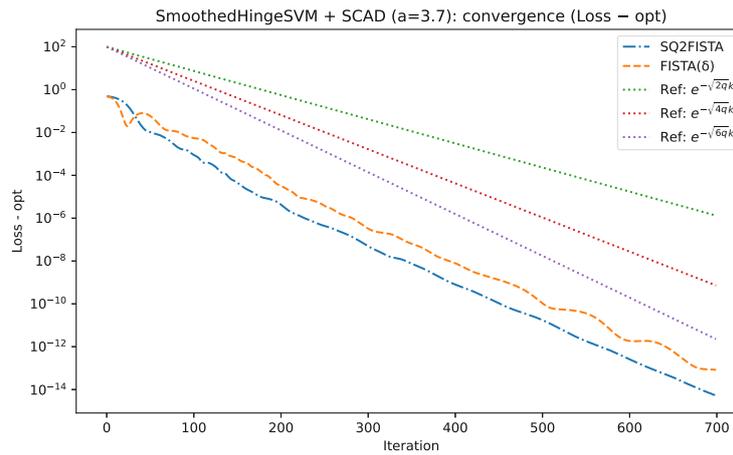
Takeaway. Once FISTA is *fed* the prox concavity via $\delta = -\mu_p$, it tracks SQ2FISTA closely. This supports the thesis that the principal performance gains come from correctly accounting for weak convexity of the prox.

6.9 Diagnostics and Ablations

Proximal safety diagnostics. For SCAD we count how often τ_k requires clamping to $a - 1 - \varepsilon$; a nonzero count signals that the method is attempting an unsafe proximal step.



(a) Smoothed Hinge + SCAD ($a = 3.7$, $\mu_m = 0.44$) (log-scale).



(b) Smoothed Hinge + SCAD: FISTA(δ) and SQ2FISTA nearly overlap.

Abb. 6.6: Smoothed Hinge SVM with SCAD ($a = 3.7$): SQ2FISTA (dash-dot) vs. FISTA vs. FISTA(δ)

6.10 Implementation Notes (for Reproducibility)

- **Stationarity map.** Use $\eta = 1/L$ (original split) for $\|G_{1/L}(x)\|$ across all methods to avoid biasing the diagnostic in favor of any split.
- **Backtracking (optional).** If analytical L is unavailable, adopt a standard majorization check $F(x^+) \leq Q_L(x^+; y)$ and increase the local L by a factor $\eta > 1$ until satisfied.
- **SCAD prox stability.** Enforce $\tau_k < a - 1$ and implement the Region-2 denominator guard $a - 1 - \tau_k > \varepsilon$; see (6.2.4).
- **Memory.** Cache $y_k - x_k$ and reuse intermediate gradients ($\nabla f(z_k)$) to reduce overhead.

6.11 Summary and Outlook

- **SCAD as weakly convex prox.** SCAD has weak convexity $\rho = 1/(a - 1)$; ensuring $\mu_{\text{tot}} = \mu_m + \mu_p > 0$ yields a well-posed composite objective with a unique minimizer.
- **SQ2FISTA vs. FISTA.** On landscapes where prox concavity matters (e.g., smoothed hinge with small γ), SQ2FISTA is distinctly faster (fewer iterations, tighter stationarity). On Saturated Nonlinear Regression they tie; on purely ill-conditioned quadratics they both failed.
- **Fairness via FISTA(δ).** Convexifying the prox and adjusting the smooth part (with $\delta = -\mu_p$) makes FISTA nearly match SQ2FISTA, clarifying that incorporating prox curvature is the decisive ingredient.
- **Outlook.** The ODE/estimate-sequence design of SQ2FISTA suggests (i) adaptive estimation of μ_p to tune damping on the fly, (ii) extensions to other weakly convex regularizers (e.g., MCP, capped- ℓ_1), and (iii) composite settings with constraints where the prox is a projection onto a weakly convex set approximation.

Kapitel 7

Pseudocodes and Overview

7.1 Algorithm Pseudocode

Gradient Descent (GD)

Algorithm 11 Gradient Descent (GD)

[1]

Input: x_0 , step $\alpha \in (0, 2/L]$

for $k = 0, 1, 2, \dots$ **do**

$$x_{k+1} \leftarrow x_k - \alpha \nabla f(x_k)$$

end for

NAG (convex)

Algorithm 12 Nesterov's Accelerated Gradient (Convex)

[1]

Input: $x_0 = x_1, t_1 = 1$

for $k = 1, 2, \dots$ **do**

$$t_{k+1} \leftarrow \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \quad \beta_k \leftarrow \frac{t_k - 1}{t_{k+1}}$$

$$y_k \leftarrow x_k + \beta_k(x_k - x_{k-1})$$

$$x_{k+1} \leftarrow y_k - \frac{1}{L} \nabla f(y_k)$$

end for

FISTA (composite convex)**Algorithm 13** FISTA (Composite Convex)

[1]

Input: $x^0 = x^1$, $t_1 = 1$ **for** $k = 1, 2, \dots$ **do**

$$t_{k+1} \leftarrow \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \quad \beta_k \leftarrow \frac{t_k - 1}{t_{k+1}}$$

$$y^k \leftarrow x^k + \beta_k(x^k - x^{k-1})$$

$$x^{k+1} \leftarrow \text{prox}_{(1/L)h}(y^k - \frac{1}{L}\nabla f(y^k))$$

end for**SQ2FISTA (wDG-based, explicit form)****Algorithm 14** SQ2FISTA (wDG-based)

[1]

Input: $x^0 = x^1$ **for** $k = 1, 2, \dots$ **do**

$$\beta_k \leftarrow \frac{k-1}{k+2}$$

$$y^k \leftarrow x^k + \beta_k(x^k - x^{k-1})$$

$$x^{k+1} \leftarrow \text{prox}_{(1/L)h}(y^k - \frac{1}{L}\nabla f(y^k))$$

end for**ADR (reference optimizer)****Algorithm 15** ADR (reference inertial method)

[1]

Input: $x_0, v_0 = 0$, steps $s, t > 0$ **for** $k = 0, 1, 2, \dots$ **do**

$$z \leftarrow x_k + sv_k$$

$$g \leftarrow \frac{1}{s^2}(z - \text{prox}_{s^2h}(z - s^2\nabla f(z)))$$

$$v_{k+1} \leftarrow \frac{v_k - tg}{1 + 1.5t}, \quad x_{k+1} \leftarrow z - s^2g, \quad v_{k+1} \leftarrow v_{k+1} + \frac{t^2}{1+t}g$$

end for

Method	Convex (smooth)	Strongly convex ($\mu > 0$)
GD	$O(1/k)$	$(1 - \mu_m/L)^k$
NAG	$O(1/k^2)$	$(1 - \sqrt{\mu_m/L})^k$
ISTA	$O(1/k)$	linear (with $\mu_m > 0$)
FISTA	$O(1/k^2)$	$(1 - \Theta(\sqrt{\mu_m/L}))^k$
FISTA(δ)	$O(1/k^2)$	$(1 - \sqrt{\mu_{\text{tot}}/L\delta})^k$
SQ2FISTA	$O(1/k^2)$	$(1 - \sqrt{2\mu_{\text{tot}}/L})^k$ (favorable constants)

Tab. 7.1: Summary of worst-case rates under standard assumptions.

7.2 Limitations and Validity

Validity of the theory. All proofs rely on standard assumptions: L -smoothness of the gradient of the smooth part, and (when invoked) μ -strong convexity. For composite problems, the three-point inequality and firm nonexpansiveness of the proximal map are essential. SQ2FISTA requires the wDG identity to hold with (α, β, γ) matching the smoothness and strong-convexity moduli of the split; this is satisfied for the natural choice $\nabla^w f(y, z) \in \nabla g(z) + \partial h(y)$.

SCAD and convexification. SCAD is weakly convex: the proximal part has negative curvature $\mu_p = -1/(a-1)$. To apply standard accelerated theory, we either (i) employ SQ2FISTA with the wDG-based Lyapunov decrease that tolerates this structure and attains accelerated rates with favorable constants, or (ii) convexify via $h_\delta = h + \frac{\delta}{2}\|x\|^2$ with $\delta = -\mu_p$, leading to FISTA(δ) and the linear factor $1 - \sqrt{\mu_{\text{tot}}/L\delta}$. In controlled experiments, SQ2FISTA matches FISTA(δ)’s behavior closely and often outperforms standard FISTA on these weakly convex setups.

Practical considerations. Accurate L (or a robust backtracking) helps stability. For strong convexity, restarts can recover linear behavior when μ is unknown. The SCAD prox requires a small safeguard (our SCADProxSafe) to keep the Region-2 denominator strictly positive; this is harmless in theory as it corresponds to an arbitrarily small regularization.

Conclusion

Main message. *SQ2FISTA* and *FISTA* share the same optimal worst-case rate on composite *convex* problems, but they differ in how they treat curvature and, consequently, in their constants and transient behavior. *SQ2FISTA* natively incorporates both momentum and (weak/strong) curvature through a discrete Lyapunov structure derived from a hyperbolically damped ODE and a weak discrete gradient (wDG). This makes it particularly effective when the proximal term is *weakly convex*. In contrast, standard *FISTA* does not “see” the proximal weak convexity unless one explicitly *convexifies* the split; when this is done (*FISTA*(δ)), the performance closely tracks *SQ2FISTA*.

What the theory says. Let the composite objective be $F = f + h$ with smoothness L_m for f , component moduli μ_m (smooth part) and μ_p (proximal part, possibly negative), and total curvature $\mu_{\text{tot}} := \mu_m + \mu_p$.

- **Purely convex regime** $\mu_{\text{tot}} = 0$. Both *FISTA* and *SQ2FISTA* achieve

$$F(x_k) - F^* = \mathcal{O}\left(\frac{1}{k^2}\right),$$

with essentially the same momentum schedule asymptotically (*FISTA*’s $\beta_k \approx \frac{k-1}{k+2}$ vs. *SQ2FISTA*’s $\beta_k \approx \frac{k}{k+3}$).

- **Strongly convex regime** $\mu_{\text{tot}} > 0$. When strong convexity is predominately in the *smooth* part (i.e., $\mu_p \approx 0$ and $\mu_{\text{tot}} \approx \mu_m$), the SQ2FISTA analysis yields a per-iteration contraction

$$F(x_k) - F^* = \mathcal{O}\left(\left(1 - \sqrt{2\mu_{\text{tot}}/L_m}\right)^k\right),$$

whereas classical accelerated proximal methods give

$$F(x_k) - F^* = \mathcal{O}\left(\left(1 - \sqrt{\mu_{\text{tot}}/L_m}\right)^k\right).$$

Thus, SQ2FISTA improves the linear factor by a $\sqrt{2}$ in the exponent under this common setting. For general (μ_m, μ_p) the wDG analysis provides an explicit geometric factor that reduces to the expression above when $\mu_p = 0$.

What the experiments show. Our benchmarks confirm the theory while also highlighting the role of constants:

- On problems where proximal *weak* convexity is active (e.g., smoothed hinge SVM with SCAD), SQ2FISTA is clearly faster than *plain* FISTA—fewer iterations and stronger stationarity, reflecting its native handling of $\mu_p < 0$.
- On robust, smoothly saturated models (e.g., Pseudo-Huber or tanh regression), the two methods are *nearly tied*, consistent with both enjoying the same k^{-2} envelope and similar constants.
- On highly ill-conditioned quadratics (where curvature is simple and the prox is benign), FISTA and SQ2FISTA don't converge.

Convexifying FISTA aligns it with SQ2FISTA. When we feed FISTA the missing proximal curvature via the standard δ -split

$$f_\delta(x) := f(x) - \frac{\delta}{2}\|x\|^2, \quad h_\delta(x) := h(x) + \frac{\delta}{2}\|x\|^2, \quad \delta := -\mu_p,$$

so that h_δ is convex and the total curvature remains $\mu_{\text{tot}} = \mu_m + \mu_p$, the resulting $FISTA(\delta)$ tracks SQ2FISTA extremely closely in practice. This “fairness check” clarifies that the main performance gains stem from consistently accounting for proximal weak convexity — either *implicitly* (SQ2FISTA’s wDG/Lyapunov design) or *explicitly* (FISTA’s convexification).

Cost, robustness, and guidance. SQ2FISTA has the same oracle cost as FISTA — one gradient of f and one prox of h per iteration—and requires no inner loops. If L_m or μ_{tot} are unknown, one can run the convex schedule (backtracking for L_m if needed); this preserves the $\mathcal{O}(1/k^2)$ guarantee and often performs well in practice. If a reliable lower bound on μ_{tot} is available (or can be estimated), SQ2FISTA’s linear regime and improved contraction factor are realized. In settings with weakly convex proximals (e.g., SCAD, MCP), SQ2FISTA is a principled default; otherwise, FISTA remains an excellent baseline, with $FISTA(\delta)$ providing a simple patch that recovers the proximal curvature when needed.

Limitations and outlook. Our analysis and experiments are deterministic and use exact prox/gradients with step sizes tied to global smoothness bounds. Extending SQ2FISTA’s Lyapunov/wDG design to stochastic or variance-reduced settings, learning μ_p on the fly, or adapting the hyperbolic damping to local curvature are promising directions. A unified treatment of constraints and projections under weak convexity is likewise an appealing avenue.

Literatur

- [1] F. Alvarez. „On the minimizing property of a second order dissipative system in Hilbert spaces“. In: *SIAM Journal on Control and Optimization* 38.4 (2000), S. 1102–1119. DOI: 10.1137/S0363012900370085.
- [2] F. Alvarez, H. Attouch, J. Bolte und P. Redont. „A second-order gradient-like dissipative dynamical system with Hessian-driven damping. Application to optimization and mechanics“. In: *Journal of Mathematics Pures et Appliquées (9)* 81.8 (2002), S. 747–779.
- [3] H. Attouch und A. Cabot. „Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity“. In: *Journal of Differential Equations* 263.9 (2017), S. 5412–5458. DOI: 10.1016/j.jde.2017.06.024.
- [4] H. Attouch, Z. Chbani und H. Riahi. „Combining fast inertial dynamics for convex optimization with Tikhonov regularization“. In: *Journal of Mathematical Analysis and Applications* 457.2 (2018), S. 1065–1094. DOI: 10.1016/j.jmaa.2016.12.017.
- [5] H. Attouch, J. Peypouquet und P. Redont. „Fast convex optimization via inertial dynamics with Hessian driven damping“. In: *Journal of Differential Equations* 261.10 (2016), S. 5734–5783. DOI: 10.1016/j.jde.2016.08.020.
- [6] J.-F. Aujol, C. Dossal und A. Rondepierre. „Convergence rates of the heavy ball method for quasi-strongly convex optimization“. In: *SIAM Journal on Optimization* 32.3 (2022), S. 1817–1842. DOI: 10.1137/21M1403990.
- [7] A. Beck und M. Teboulle. „A fast iterative shrinkage-thresholding algorithm for linear inverse problems“. In: *SIAM Journal on Imaging Sciences* 2.1 (2009), S. 183–202. DOI: 10.1137/080716542.
- [8] L. Chen und H. Luo. „A unified convergence analysis of first order convex optimization methods via strong Lyapunov functions“. In: *arXiv preprint* (2021). arXiv:2108.00132.
- [9] A. d’Aspremont, D. Scieur und A. Taylor. „Acceleration methods“. In: *Foundations and Trends in Optimization* 5.1–2 (2021), S. 1–245. DOI: 10.1561/24000000036.
- [10] I. Daubechies, M. Defrise und C. D. Mol. „An iterative thresholding algorithm for linear inverse problems with a sparsity constraint“. In: *Communications on Pure and Applied Mathematics* 57.11 (2004), S. 1413–1457. DOI: 10.1002/cpa.20042.
- [11] Y. Drori und A. Taylor. „On the oracle complexity of smooth strongly convex minimization“. In: *Journal of Complexity* 68 (2022), S. 101590. DOI: 10.1016/j.jco.2021.101590.
- [12] D. Du. „Lyapunov function approach for approximation algorithm design and analysis: with applications in submodular maximization“. In: *arXiv preprint* (2022). arXiv:2205.12442.

- [13] D. Kim und J. A. Fessler. „Another look at the fast iterative shrinkage/thresholding algorithm (FISTA)“. In: *SIAM Journal on Optimization* 28.1 (2018), S. 223–250. DOI: 10.1137/16M108940X.
- [14] W. Krichene, A. Bayen und P. L. Bartlett. „Accelerated mirror descent in continuous and discrete time“. In: *Advances in Neural Information Processing Systems*. Bd. 28. 2015.
- [15] H. Luo und L. Chen. „From differential equation solvers to accelerated first-order methods for convex optimization“. In: *Mathematical Programming* 195 (2022), S. 735–781. DOI: 10.1007/s10107-021-01722-0.
- [16] Y. Nesterov. „Gradient methods for minimizing composite functions“. In: *Mathematical Programming* 140.1, Ser. B (2013), S. 125–161. DOI: 10.1007/s10107-012-0629-5.
- [17] Y. Nesterov. *Lectures on Convex Optimization*. Bd. 137. Springer Optimization and Its Applications. Cham: Springer, 2018. DOI: 10.1007/978-3-319-91578-4.
- [18] Y. E. Nesterov. „A method for solving the convex programming problem with convergence rate $O(1/k^2)$ “. In: *Doklady Akademii Nauk SSSR* 269.3 (1983), S. 543–547.
- [19] B. T. Polyak. „Some methods of speeding up the convergence of iterative methods“. In: *USSR Computational Mathematics and Mathematical Physics* 4.5 (1964), S. 1–17. DOI: 10.1016/0041-5553(64)90137-5.
- [20] E. K. Ryu und S. Boyd. „A primer on monotone operator methods“. In: *Applied and Computational Mathematics* 15.1 (2016), S. 3–43.
- [21] B. Shi, S. S. Du, M. I. Jordan und W. J. Su. „Understanding the acceleration phenomenon via high-resolution differential equations“. In: *Mathematical Programming* 195.1–2 (2022), S. 79–148. DOI: 10.1007/s10107-021-01681-8.
- [22] W. Su, S. Boyd und E. J. Candès. „A differential equation for modeling Nesterov’s accelerated gradient method: theory and insights“. In: *Advances in Neural Information Processing Systems*. Bd. 27. 2014.
- [23] W. Su, S. Boyd und E. J. Candès. „A differential equation for modeling Nesterov’s accelerated gradient method: theory and insights“. In: *Journal of Machine Learning Research* 17.153 (2016), S. 1–43. URL: <http://jmlr.org/papers/v17/15-084.html>.
- [24] A. Taylor und Y. Drori. „An optimal gradient method for smooth strongly convex minimization“. In: *Mathematical Programming* 199.1–2 (2023), S. 557–594. DOI: 10.1007/s10107-022-01839-y.
- [25] A. Taylor, B. Van Scoy und L. Lessard. „Lyapunov functions for first-order methods: Tight automated convergence guarantees“. In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 2018, S. 4897–4906.
- [26] K. Ushiyama. „A $\sqrt{2}$ -accelerated FISTA for composite strongly convex problems“. In: *arXiv preprint* (2025). arXiv:2509.09295.
- [27] K. Ushiyama. „wDG for ITEM ODE“. In: *preprint* (2024). arXiv: not assigned.

- [28] K. Ushiyama, S. Sato und T. Matsuo. „A Unified Discretization Framework for the differential equation approach with Lyapunov arguments for convex optimization“. In: *preprint* (2024).
- [29] K. Ushiyama, S. Sato und T. Matsuo. „Deriving efficient optimization methods based on stable explicit numerical methods“. In: *JSIAM Letters* 14 (2022), S. 29–32.
- [30] K. Ushiyama, S. Sato und T. Matsuo. „Essential convergence rate of ordinary differential equations appearing in optimization“. In: *JSIAM Letters* 14 (2022), S. 119–122.
- [31] B. Van Scoy, R. A. Freeman und K. M. Lynch. „The fastest known globally convergent first-order method for minimizing strongly convex functions“. In: *IEEE Control Systems Letters* 2.1 (2018), S. 49–54. DOI: 10.1109/LCSYS.2017.2722406.
- [32] A. Wibisono, A. C. Wilson und M. I. Jordan. „A variational perspective on accelerated methods in optimization“. In: *Proceedings of the National Academy of Sciences (USA)* 113.47 (2016), E7351–E7358. DOI: 10.1073/pnas.1614734113.
- [33] A. C. Wilson, B. Recht und M. I. Jordan. „A Lyapunov analysis of accelerated methods in optimization“. In: *Journal of Machine Learning Research* 22.113 (2021), S. 1–34. URL: <http://jmlr.org/papers/v22/20-195.html>.