Finding Simple Proofs for First-Order Optimization

Anonymous Authors[1]

Abstract

Large language models (LLMs) are increasingly used in research across mathematics and machine learning, and their impact is shaped by the workflows researchers build around them. Several challenges arise, from identifying tasks where LLM assistance is genuinely useful, to designing reproducible tool-augmented procedures for those tasks, and structuring the underlying problems to exploit these procedures. We believe first-order optimization offers a natural setting for advancing those challenges. Indeed, many convergence proofs admit certificate representations that can be searched, simplified, and interpreted, making the use of agentic workflows possibly very powerful.

Recent work on performance estimation problems has shown that proofs of convergence for first-order optimization methods can be discovered by searching over a structured space of Lagrangian dual certificates. This paper studies the simplification of such proofs as an optimization problem. Starting from dual certificates, we develop post-processing procedures using tools from sparse optimization and statistical learning. We measure complexity through features such as active hypotheses and residual structure, and introduce methods based on (weighted) ℓ_1-heuristics, and formulations for discovering simple proofs using intermediate lemmas.

Examples on gradient descent, proximal methods, and fast-gradient methods show that these procedures can autonomously prune redundant inequalities, reveal structured proof patterns, and, in the proximal setting, recover standard intermediate lemmas that lead to streamlined proofs. By distilling dense machine-generated certificates into compact proof structures, this workflow acts as a pre-processing step for the final proof, reducing the complexity that must be managed during human interpretation, reuse, and formalization.

Workshop fit. Rather than focusing on proposing prompting techniques, this paper contributes a solver-guided workflow for AI-assisted proof development. It casts proof simplification as a structured optimization problem: candidate inequalities, residual decompositions, and intermediate lemmas can be searched, tested, and reused by systems that combine language models with numerical solvers. This connects the work to AI for research by making part of mathematical discovery more systematic and verifiable.

In other words, *the advances on proof discovery proposed in the paper can directly contribute to enhancing agentic workflows' ability to advance research.*

1. Introduction

Proof simplification is a central part of mathematical practice. Once a result has been proved, a simpler proof can change what the result offers: it can make the argument easier to verify, reveal the mechanism behind the statement, and turn a one-off derivation into a building block for extensions. This role becomes especially important in computer-aided mathematics, where proofs and certificates may be produced, manipulated, or checked by numerical solvers, symbolic algebra systems, formal proof assistants such as Lean, and large language models. In such settings, obtaining a valid certificate is often not enough: one also wants proof structures that are sparse, modular, and reusable.

Performance estimation problems (PEPs) provide a concrete setting in which this task can be posed precisely. PEPs have made it possible to certify tight worst-case bounds for first-order optimization methods by solving numerical optimization problems (Drori & Teboulle, 2014; Taylor et al., 2017c). In the standard semidefinite programming (SDP) formulations used for interpolation-based PEPs, Lagrangian dual solutions are algebraic certificates of the corresponding convergence bounds, and hence machine-searchable proof objects. These certificates, however, are often numerical, parameter-specific, dense, and far from unique: even after the method, function class, PEP formulation, and target bound have been fixed, many different certificates may

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

prove the same inequality. In this form, a certificate may establish the rate without making clear which inequalities, residual terms, or intermediate statements actually drive the proof.

This article asks whether the same machinery that certifies a worst-case guarantee can also help identify a simple proof of it. We treat simplification as a second-stage optimization problem over Lagrangian dual certificates. Here "simple" is an operational notion: a proof may use fewer active hypotheses, organize the remaining nonnegative residual terms more transparently, expose recognizable multiplier patterns, or isolate an intermediate lemma that turns a dense identity into a reusable proof step. Starting from this viewpoint, we develop post-processing procedures based on exhaustive sparsification, weighted ℓ_1-type heuristics, and SDP-based searches for candidate intermediate lemmas. The workflow is deliberately not told in advance which proof pattern or lemma should appear; after a bound has been certified, it attempts to distill the certificate into proof ingredients suitable for human interpretation, reuse, and translation into symbolic/formal proof systems (e.g., Naldi et al., 2025).

Concretely, we consider first-order methods for problems of the form

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ \ f(x),$$

where $f:\mathbb{R}^d\to\mathbb{R}$ belongs to a function class $\mathcal{F}$; the same framework can also accommodate composite terms, constraints, and proximal structure. For a method $\mathcal{A}$ applied over k steps (e.g., Gradient Descent), let $\mathcal{T}_{\mathcal{A}}(f,k,x_0)$ denote the first $k+1$ iterates of the trajectory starting from x_0, generated by the method $\mathcal{A}$ on the function $f\in\mathcal{F}$. Moreover, let $\mathcal{P}_f$ denote the performance metric of interest (e.g., $\|x_k-x_\star\|^2$), and $\mathcal{I}_f$ denote the initialization measure (e.g., $\|x_0-x_\star\|^2$). The associated worst-case problem corresponds to finding a worst-case dimension d, function f with optimal point $x_\star$, initialization x_0, and trajectory (x_t), maximizing the effective rate ρ for those metrics, i.e.,

$$\begin{aligned}
\text{maximize}\quad & \rho\\
\text{over}\quad & \rho\in\mathbb{R},\ f\in\mathcal{F},\ d\geqslant 1,\\
& x_\star,x_0\in\mathbb{R}^d,\\
& (x_t)_{t\in[\![0,k]\!]}\in\mathcal{T}_{\mathcal{A}}(f,k,x_0)\\
\text{subject to}\quad & \mathcal{P}_f\Big((x_t)_{t\in[\![0,k]\!]},x_\star\Big)\\
& \quad=\rho\mathcal{I}_f\Big((x_t)_{t\in[\![0,k]\!]},x_\star\Big),\\
& x_\star\in\underset{x\in\mathbb{R}^d}{\operatorname{argmin}}\, f(x).
\end{aligned}\tag{PEP}$$

This problem is infinite-dimensional because the function itself is an optimization variable, and the trajectory constraints may include oracle outputs and auxiliary variables used to describe the run. When an exact finite-dimensional SDP reformulation is available, we work with the resulting finite problem and its Lagrangian dual certificates.

Given such an exact SDP formulation, a verified feasible solution of the Lagrangian dual whose objective is no larger than the target upper bound gives a numerical certificate of that bound. These certificates are typically not unique: the same guarantee can be witnessed by different choices of Lagrangian dual multipliers and slack variables. Thus, once the PEP formulation and target guarantee are fixed, simplification can be viewed as a second-stage search for an alternative certificate that still proves the chosen bound. Such simplified certificates can turn dense numerical multipliers into explicit patterns and, in favorable cases, reveal proof templates that can be adapted to related function classes or algorithms.

Contributions. **(i)** Building on standard SDP formulations for interpolation-based PEPs, we introduce certificate-complexity criteria: active inequalities and residual terms left after a chosen decomposition. **(ii)** We develop exact and heuristic sparsification procedures, from exhaustive search in small instances to weighted ℓ_1-type surrogates for larger ones. **(iii)** We propose an SDP search for deriving valid inequalities from those already available in a PEP formulation, thereby generating candidate intermediate lemmas. **(iv)** We provide examples on gradient descent (GD), fast-gradient methods (FGM), and proximal methods in Appendices A and B: the workflow recovers weaker fitted interpolation inequalities, a three-hypothesis GD proof, compact FGM multiplier patterns, and compact proofs using Lyapunov functions for the proximal point residual bound and accelerated proximal point saddle-gap estimate.

Related work and scope. This work builds on the Performance Estimation Problem (PEP) framework, originally introduced by Drori & Teboulle (2014), as formalized by Taylor (2017); Taylor et al. (2017a). In essence, PEPs facilitate the search for performance certificates, comprising both complexity bounds and their corresponding proofs, by leveraging the Lagrangian dual of the PEP. This approach has led to the discovery of numerous tight certificates (e.g., De Klerk et al., 2017; Taylor & Bach, 2019; Barré et al., 2020; Dragomir et al., 2022; Gorbunov et al., 2022; Goujaud et al., 2022; Berg Thomsen et al., 2025; Rotaru et al., 2024; Abbaszadehpeivasti et al., 2024).

More broadly, PEPs provide a rigorous framework for characterizing the proof structures inherent to first-order optimization in classical setups (Goujaud et al., 2023). This has enabled the discovery of proofs that were arguably beyond the reach of traditional analytical patterns (Kim & Fessler, 2016; 2018b;a; Ryu et al., 2020; Lieder, 2021), as well as the systematic design of new algorithms (Kim & Fessler, 2016; Drori & Taylor, 2020; Taylor & Drori, 2023; Altschuler &

Parrilo, 2025; Jang et al., 2025). These advancements are supported by dedicated software tools (Taylor et al., 2017b; Goujaud et al., 2024).

However, such certificates are rarely unique. The resulting proofs are often highly complex, making them difficult to understand, replicate, or adapt to new algorithms. This paper demonstrates that tools from sparse optimization and statistical learning can be used to make the passage from certificates to proofs more systematic: they provide a way to search for *simpler* certificates that still recover the desired performance guarantees.

Organization. The rest of the paper is organized as follows. First, in Section 2, we formalize proof structures and the certificate-complexity measures used to compare them. We then describe search procedures for sparsifying certificates and generating candidate lemmas in Section 3. The appendices contain the worked examples and technical details: Appendix A shows how candidate lemmas recover fitted interpolation inequalities for gradient descent, while sparsification exposes compact multiplier patterns for fast-gradient methods; Appendix B states proximal certificates whose closed-form proofs are given in Appendix H. Additional technical material covers the general PEP framework, sparse minimization details, the candidate-lemma SDP formulation, the concrete PEPs used in the experiments, branch-wise GD residuals, and FGM active-multiplier patterns in Appendices C to G and I.

Notation. We write $a \triangleq b$ when a is defined as b. For integers $n, m \in \mathbb{Z}$, write $[\![n, m]\!] = \{i \in \mathbb{Z} : n \leqslant i \leqslant m\}$. For finite sets S, $|S|$ denotes cardinality. We use $\langle \cdot, \cdot \rangle$ for the ambient inner product and $\|\cdot\|$ for its induced norm; for self-adjoint positive-definite B, write $\|x\|_B^2 \triangleq \langle Bx, x\rangle$. The trace is denoted by Tr. Let $\mathbb{S}^r$ be the space of $r \times r$ real symmetric matrices and $\mathbb{S}^r_+$ its positive semidefinite cone; $A \succcurlyeq 0$ means $A \in \mathbb{S}^r_+$. We write $\mathbb{R}^p_+$ for the nonnegative orthant. If $x_\star$ is an optimizer, write $f_\star = f(x_\star)$. For convex functions, ∂f denotes the convex subdifferential; for concave scalar penalties, $\partial \phi$ denotes the superdifferential.

2. Simple proof structures

We first recall classical elements on proof structures in first-order optimization.

2.1. Proof structures

As described in (PEP), a convergence proof typically certifies, for each $k \in \mathbb{N}$, a bound of the form

$$\mathcal{P}_f\Big((x_t)_{t\in[\![0,k]\!]}, x_\star\Big) \leqslant \rho(k)\ \mathcal{I}_f\Big((x_t)_{t\in[\![0,k]\!]}, x_\star\Big), \quad (1)$$

uniformly over all problem dimensions, functions $f \in \mathcal{F}$, minimizers $x_\star$, initial points x_0, and trajectories $(x_t)_t$ generated by the method $\mathcal{A}$, where $\mathcal{P}_f$ is the chosen performance metric, $\mathcal{I}_f$ is an initialization budget, and $\rho(k)$ is the certified rate at iteration k. The *performance estimation* literature (see, e.g., Drori & Teboulle, 2014; Taylor, 2017; Taylor et al., 2017c; Goujaud et al., 2023) has studied the structure of first-order convergence proofs. In particular, one key takeaway is that many first-order convergence proofs can be written as

$$\begin{aligned}\rho(k)\mathcal{I}_f\Big((x_t)_{t\in[\![0,k]\!]}, x_\star\Big) - \mathcal{P}_f\Big((x_t)_{t\in[\![0,k]\!]}, x_\star\Big) \\ = \sum_{i,j\in\mathcal{S}} \lambda_{ij} H(x_i, x_j) + \sum_{i=1}^{r} c_i R_i,\end{aligned} \quad \text{(Proof)}$$

where $\mathcal{S} = \{\star, 0, \dots, k\}$, $r \in \mathbb{N}$, Lagrangian multipliers λ_{ij}, and c_i are all non-negative, and:

1. $(H(x_i, x_j))_{i,j\in\mathcal{S}}$ are expressions provided by the definition of the class and the algorithm, which are enforced to be non-negative. For first-order function classes considered here, those expressions encompass the *interpolation inequalities* (Taylor et al., 2017c).
2. Each R_i is an explicit square or sum of squares (so $R_i \geqslant 0$).

Thus the right-hand side is non-negative term by term, eq. (Proof) proves eq. (1). Classes of problems known to admit such decompositions are detailed in Appendix C.

2.2. Proof complexity

Identity (Proof) gives a direct way to discuss the complexity of a convergence proof. Once the iteration k, the point set $\mathcal{S}$, and the allowed inequality and residual terms are fixed, the complexity measures below track which valid inequalities appear with positive multipliers and which residual terms have positive coefficients.

For a certificate of the form (Proof), when there is one valid inequality for each ordered pair, define the active inequality and residual sets by

$$\begin{aligned}\mathcal{A}_H &\triangleq \big\{(i,j) \in \mathcal{S}^2 : \lambda_{ij} > 0,\ H(x_i, x_j) \not\equiv 0\big\},\\ \mathcal{A}_R &\triangleq \{i \in [\![1, r]\!] : c_i > 0\}.\end{aligned}$$

Thus $(i,j) \in \mathcal{A}_H$ means that the proof uses the interpolation inequality $H(x_i, x_j) \geqslant 0$, while $i \in \mathcal{A}_R$ means that the residual term R_i appears in the residual decomposition. These active sets induce the two complexity measures used below:

1. the **hypothesis complexity** of a proof identity is the number of active inequalities $|\mathcal{A}_H|$,
2. the **residual complexity** of a proof identity is the number of active residual terms $|\mathcal{A}_R|$.

For numerical experiments, the active sets are thresholded: $\lambda_{ij} > 0$ and $c_i > 0$ are replaced by $\lambda_{ij} > \varepsilon_{\mathrm{act}}$ and $c_i > \varepsilon_{\mathrm{act}}$, with the tolerance specified in the corresponding experiment or table.

3. Heuristic search for simple proofs

We now turn the proof-complexity measures above into concrete search procedures. Throughout this section, a search instance consists of a fixed convergence guarantee, a finite list of available valid hypotheses, and a chosen residual decomposition. The goal is to find another certificate for essentially the same guarantee that uses fewer active hypotheses, simpler residual terms, or both.

There are two useful choices to make before applying these heuristics:

1. **Expand the hypotheses set.** When available, add redundant but valid inequalities (e.g., Descent Lemma) before sparsifying. This larger finite list fixes what counts as an *available hypothesis* and can eventually lead to simpler certificates, even though the search starts with more possible terms.

2. **Relax the target.** Let $\rho_\star > 0$ denote the best rate certified by the chosen formulation. For a target relative suboptimality tolerance $\varepsilon_{\mathrm{rel}} \geqslant 0$, set $\bar{\rho} \triangleq (1 + \varepsilon_{\mathrm{rel}})\rho_\star$.

3.1. Exhaustive sparsification

Finding a proof with minimal hypothesis complexity is a best-subset-type problem over the available hypotheses, analogous to classical best subset selection in sparse regression (Bertsimas et al., 2016). Closely related minimum-cardinality feasibility and sparse-approximation problems are already NP-hard (Amaldi & Kann, 1998; Natarajan, 1995): for each candidate active pattern, one must check whether there exists a certificate using no hypotheses outside that pattern. Exhaustive sparsification is the direct combinatorial baseline. It is useful as a ground-truth benchmark on small instances, but its cost still scales exponentially with the number of possible active hypothesis patterns.

Index the available hypotheses by $[\![1, m]\!]$. For an active pattern $\mathcal{A} \subseteq [\![1, m]\!]$, let $\rho(\mathcal{A})$ be the smallest bound certifiable using only $\mathcal{A}$, with $\rho(\mathcal{A}) = +\infty$ if no such certificate exists; call $\mathcal{A}$ admissible if $\rho(\mathcal{A}) \leqslant \bar{\rho}$. Since restricting the active set cannot improve the optimum, $\rho(\mathcal{A}) \geqslant \rho_\star$. With this notation, exhaustive sparsification computes $h_\star \triangleq \min\{|\mathcal{A}| : \mathcal{A} \subseteq [\![1, m]\!],\ \rho(\mathcal{A}) \leqslant \bar{\rho}\}$, and returns all admissible patterns of size $h_\star$.

3.2. Sparse minimization heuristics

The exhaustive baseline is exact but quickly becomes too expensive. For larger instances, we instead minimize smooth or piecewise-linear surrogates of hypothesis complexity over certificates that already prove the relaxed target, using standard ideas from sparse optimization and statistical learning such as lasso relaxations, reweighted ℓ_1 penalties, log-sum penalties, and capped ℓ_1 penalties (Tibshirani, 1996; Candès et al., 2008; Zhang, 2010). The concrete objectives and majorization scheme used in the experiments are given in Appendix D.

3.3. Discovering proof structures through intermediate lemmas

The previous heuristics simplify a certificate after the list of available hypotheses has been fixed. We can also enlarge this list by searching for additional valid inequalities that may serve as reusable steps in a shorter proof. Each candidate lemma is an auxiliary inequality whose validity is itself certified by a proof identity of the form (Proof); it is therefore not a new assumption, but a derived inequality that can be reused as a separately checkable step. The SDP formulation, extraction procedure, and final proof reconstruction are given in Appendix E.

4. Conclusion

This paper treats PEP dual solutions not only as tight worst-case certificates, but as objects for searching over simpler proof structures; active multipliers and residual terms provide tangible complexity measures for deciding when a computer-generated certificate can be made readable. Examples in the appendix showcase three uses of this idea: exhaustive sparsification exposes closed-form multiplier patterns for gradient descent, candidate lemmas recover proximal Lyapunov decrements for residual/value and saddle-gap bounds, and normalized sparsity heuristics reduce the active multiplier set for FGM, suggesting a workflow in which computer-aided worst-case analysis is followed by certificate simplification to guide proof design.

References

Abbaszadehpeivasti, H., de Klerk, E., and Zamani, M. On the rate of convergence of the difference-of-convex algorithm (dca). *Journal of Optimization Theory and Applications*, 202(1):475–496, 2024.

Altschuler, J. M. and Parrilo, P. A. Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule. *Journal of the ACM*, 72(2):1–38, 2025.

Amaldi, E. and Kann, V. On the approximability of minimizing nonzero variables or unsatisfied relations in linear

systems. *Theoretical Computer Science*, 209(1–2):237–260, 1998. doi: 10.1016/S0304-3975(97)00115-1.

Barré, M., Taylor, A., and Bach, F. Principled analyses and design of first-order methods with inexact proximal operators. *arXiv preprint arXiv:2006.06041*, 2020.

Berg Thomsen, D., Taylor, A., and Dieuleveut, A. Tight analyses of first-order methods with error feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Bertsimas, D., King, A., and Mazumder, R. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016. doi: 10.1214/15-AOS1388.

Candès, E. J., Wakin, M. B., and Boyd, S. P. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008. doi: 10.1007/s00041-008-9045-x.

De Klerk, E., Glineur, F., and Taylor, A. B. On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *Optimization Letters*, 11(7):1185–1199, 2017.

Dragomir, R.-A., Taylor, A. B., d'Aspremont, A., and Bolte, J. Optimal complexity and certification of bregman first-order methods. *Mathematical Programming*, 194(1):41–83, 2022.

Drori, Y. and Taylor, A. Efficient first-order methods for convex minimization: a constructive approach. *Mathematical Programming, Series A*, 184:183–220, 2020.

Drori, Y. and Teboulle, M. Performance of first-order methods for smooth convex minimization: A novel approach. *Mathematical Programming*, 145(1):451–482, 2014. doi: 10.1007/s10107-013-0653-0.

Fazel, M., Hindi, H., and Boyd, S. P. Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. In *Proceedings of the 2003 American Control Conference (ACC), 2003*, volume 3, pp. 2156–2162. IEEE, 2003. doi: 10.1109/ACC.2003.1243393.

Gorbunov, E., Taylor, A., and Gidel, G. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. *Advances in neural information processing systems*, 35:21858–21870, 2022.

Goujaud, B., Taylor, A., and Dieuleveut, A. Optimal first-order methods for convex functions with a quadratic upper bound. *arXiv preprint arXiv:2205.15033*, 2022.

Goujaud, B., Dieuleveut, A., and Taylor, A. On fundamental proof structures in first-order optimization. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 3023–3030. IEEE, 2023. doi: 10.1109/CDC49753.2023.10383282.

Goujaud, B., Moucer, C., Glineur, F., Hendrickx, J. M., Taylor, A. B., and Dieuleveut, A. PEPit: Computer-assisted worst-case analyses of first-order optimization methods in Python. *Mathematical Programming Computation*, 16 (3):337–367, 2024. doi: 10.1007/s12532-024-00259-7.

Gu, G. and Yang, J. Tight convergence rate in subgradient norm of the proximal point algorithm. *Optimization*, 2025. doi: 10.1080/02331934.2025.2602877.

Jang, U., Gupta, S. D., and Ryu, E. K. Computer-assisted design of accelerated composite optimization methods: Optista: U. jang et al. *Mathematical Programming*, pp. 1–109, 2025.

Kim, D. Accelerated proximal point method for maximally monotone operators. *Mathematical Programming*, 190: 57–87, 2021. doi: 10.1007/s10107-021-01643-0.

Kim, D. and Fessler, J. A. Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159(1):81–107, 2016.

Kim, D. and Fessler, J. A. Another look at the fast iterative shrinkage/thresholding algorithm (fista). *SIAM Journal on Optimization*, 28(1):223–250, 2018a.

Kim, D. and Fessler, J. A. Generalizing the optimized gradient method for smooth convex minimization. *SIAM Journal on Optimization*, 28(2):1920–1950, 2018b.

Lieder, F. On the convergence rate of the halpern-iteration. *Optimization letters*, 15(2):405–418, 2021.

Naldi, S., Safey El Din, M., Taylor, A., and Wang, W. Solving generic parametric linear matrix inequalities. In *Proceedings of the 2025 International Symposium on Symbolic and Algebraic Computation*, pp. 267–276, 2025.

Natarajan, B. K. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995. doi: 10.1137/S0097539792240406.

Rotaru, T. *Exact Performance Analysis of Fundamental First-Order Optimization Methods*. PhD thesis, KU Leuven, 2026. URL https://lirias.kuleuven.be/handle/20.500.12942/781389.

Rotaru, T., Glineur, F., and Patrinos, P. Exact worst-case convergence rates of gradient descent: a complete analysis for all constant stepsizes over nonconvex and convex functions. *arXiv preprint arXiv:2406.17506*, 2024.

Ryu, E. K., Taylor, A. B., Bergeling, C., and Giselsson, P. Operator splitting performance estimation: Tight contraction factors and optimal parameter selection. *SIAM Journal on Optimization*, 30(3):2251–2271, 2020.

Taylor, A. *Convex interpolation and performance estimation of first-order methods for convex optimization*. PhD thesis, Université Catholique de Louvain, 2017.

Taylor, A. and Bach, F. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Conference on Learning Theory*, pp. 2934–2992. PMLR, 2019.

Taylor, A. and Drori, Y. An optimal gradient method for smooth strongly convex minimization. *Mathematical Programming*, 199(1):557–594, 2023.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27 (3):1283–1313, 2017a. doi: 10.1137/16M108104X.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Performance estimation toolbox (pesto): Automated worst-case analysis of first-order optimization methods. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 1278–1283. IEEE, 2017b.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017c.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. *Journal of Optimization Theory and Applications*, 178(2):455–476, 2018. doi: 10.1007/s10957-018-1298-1.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Uschmajew, A. and Vandereycken, B. A note on the optimal convergence rate of descent methods with fixed step sizes for smooth strongly convex functions. *Journal of Optimization Theory and Applications*, 194(1):364–373, 2022. doi: 10.1007/s10957-022-02032-z.

Zhang, T. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11(35):1081–1107, 2010.

A. Experimental examples

The notebooks and code implementing the procedures used in these examples are available in the public source-code repository https://github.com/icmlworkshop2026simple/proofs. All reported experiments were run on a MacBook Pro with an Apple M4 Max chip, 14 CPU cores, and 36 GB of memory. The examples are small deterministic SDP and enumeration computations; except for the largest exhaustive active-set checks, they are not computationally intensive.

A.1. Gradient descent

Consider one step of gradient descent,

$$x_1 = x_0 - \gamma \nabla f(x_0), \tag{GD}$$

where $\gamma > 0$ is the stepsize. We are interested in worst-case guarantees on the functional residual after one step on the class $\mathcal{F}_{L,\mu}$, $0 \leqslant \mu < L$, under the initial normalization $f(x_0) - f_\star \leqslant 1$. The corresponding numerical formulations are collected in Appendix F.1.

For the function f under discussion, points $u, v \in \mathbb{R}^d$, and $0 \leqslant \mu < L$, define

$$\begin{aligned} H_{L,\mu}(u,v) \triangleq\ & f(u) - f(v) \\ & - \langle \nabla f(v), u - v \rangle - \frac{\mu}{2} \|u - v\|^2 \\ & - \frac{1}{2(L-\mu)} \left\| \begin{matrix} \nabla f(u) - \nabla f(v) \\ -\mu(u - v) \end{matrix} \right\|^2 \end{aligned} \tag{2}$$

For any $f \in \mathcal{F}_{L,\mu}$, co-coercivity gives $H_{L,\mu}(u,v) \geqslant 0$ for all points u, v.

A.1.1. Exhaustive sparsification

For the one-step functional-residual experiment, exhaustive enumeration is still practical because the points $x_\star$, x_0, and x_1 give only six nontrivial two-point co-coercivity inequalities. Figure 1 compares the raw certificate weights with the weights retained after exhaustive sparsification. The retained weights multiply $H_{L,\mu}(x_\star, x_0)$, $H_{L,\mu}(x_\star, x_1)$, and $H_{L,\mu}(x_0, x_1)$; the weights on the other three two-point inequalities vanish. They follow the branchwise closed forms below for $0 < \gamma \leqslant 2/L$, with threshold $\gamma = \frac{2}{L+\mu}$:

$$\begin{aligned} \lambda_{\star 0} &= \begin{cases} \gamma\mu(1-\gamma\mu), & \gamma \leq \frac{2}{L+\mu}, \\ (2-\gamma L)(\gamma L - 1), & \gamma \geq \frac{2}{L+\mu}, \end{cases} \\ \lambda_{\star 1} &= \min\{\gamma\mu,\ 2-\gamma L\}, \\ \lambda_{01} &= \max\{1-\gamma\mu,\ \gamma L - 1\}. \end{aligned}$$

A.1.2. Fitted curvatures for the classical rate of GD

Applying the candidate-lemma search of Section 3.3 with one singleton candidate for each two-point inequality recovers a tight one-step GD proof for every stepsize $0 < \gamma \leqslant 2/L$. The rate is the classical fixed-step function-value contraction; here, the search rediscovers it as a three-hypothesis certificate. Concretely, the singleton candidate lemmas can be recognized as co-coercivity inequalities for the fitted larger class

$$\begin{aligned} \mathcal{F}_{L,\mu} &\subseteq \mathcal{F}_{L_\gamma,\mu_\gamma}, \\ (L_\gamma, \mu_\gamma) &\triangleq \begin{cases} \left(\frac{2}{\gamma} - \mu,\ \mu\right), & 0 < \gamma \leqslant \frac{2}{L+\mu}, \\ \left(L,\ \frac{2}{\gamma} - L\right), & \frac{2}{L+\mu} \leqslant \gamma \leqslant \frac{2}{L}. \end{cases} \end{aligned} \tag{3}$$

Figure 2 shows the fitted curvatures extracted by a candidate-lemma SDP after fixing the sparse certificate weights on a representative grid; the measured points lie on the closed-form branches in (3). For $0 < \gamma \leqslant 2/L$, these constants satisfy $L_\gamma \geqslant L$, $0 \leqslant \mu_\gamma \leqslant \mu$, and $\gamma = 2/(L_\gamma + \mu_\gamma)$; hence the fitted inequalities are valid for every original $\mathcal{F}_{L,\mu}$ instance, but impose only the weaker requirements of the larger class.

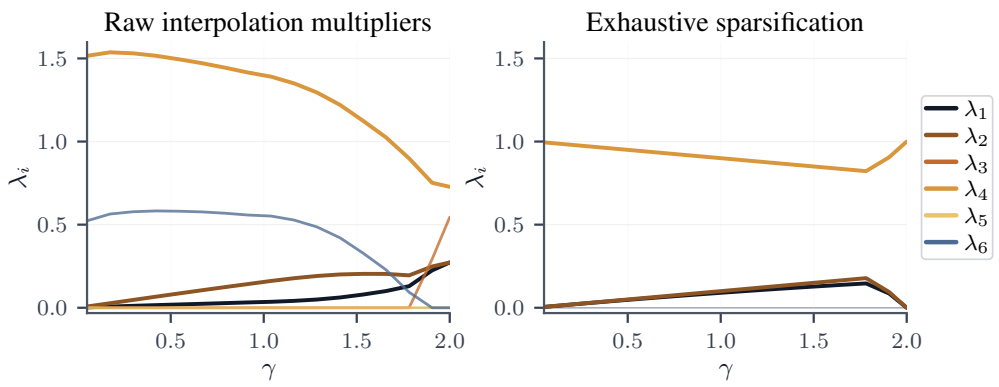

Figure 1. One-step gradient descent with functional-residual normalization. Left: raw certificate weights on the step-size grid. Right: the retained weights after exhaustive sparsification.

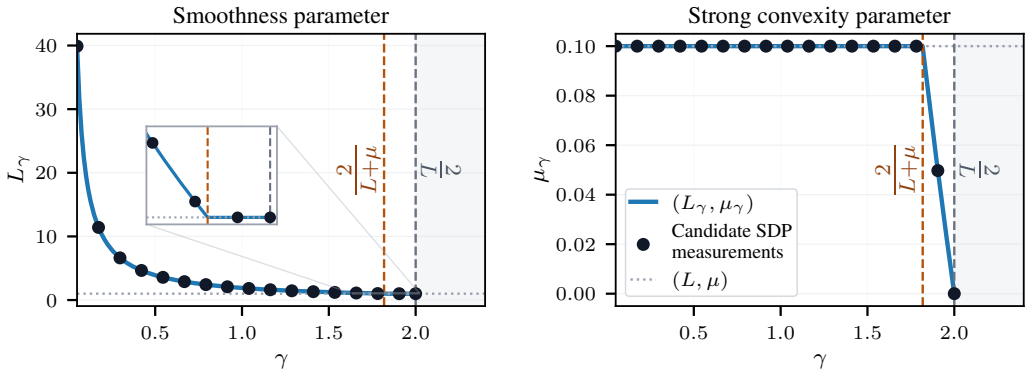

Figure 2. Fitted interpolation curvatures identified from singleton candidate lemmas for the one-step gradient descent certificate, shown for $L = 1$ and $\mu = 0.1$. The SDP measurements follow the closed-form curves: before the threshold $2/(L+\mu)$, the fitted class keeps $\mu_\gamma = \mu$ and increases L_γ; after the threshold, it keeps $L_\gamma = L$ and weakens μ_γ. Dotted horizontal lines mark the original constants.

Define

$$\begin{aligned}
\rho_\gamma &\triangleq \max\left\{(1-\gamma\mu)^2, (\gamma L - 1)^2\right\} \\
&= (1-\gamma\mu_\gamma)^2 = (\gamma L_\gamma - 1)^2, \\
\lambda_{\star 0} &\triangleq \gamma\mu_\gamma\sqrt{\rho_\gamma}, \\
\lambda_{\star 1} &\triangleq \gamma\mu_\gamma, \\
\lambda_{01} &\triangleq \sqrt{\rho_\gamma}.
\end{aligned} \tag{4}$$

The three λ's are nonnegative for $0 < \gamma \leqslant 2/L$. On the large-step branch, the fitted value $\mu_\gamma = 2/\gamma - L$ is exactly the weaker strong-convexity parameter used by Uschmajew & Vandereycken (2022, after Eq. (1.3)); the proof below uses the same weakening trick but covers the full range $0 < \gamma \leqslant 2/L$.

Theorem 1 (Worst-case performance of one-step GD)**.** *Let $f \in \mathcal{F}_{L,\mu}$ with $0 \leqslant \mu < L$, and let $0 < \gamma \leqslant 2/L$. With the quantities in (3)–(4), one step of gradient descent satisfies $f(x_1) - f_\star \leqslant \rho_\gamma(f(x_0) - f_\star)$.*

Proof. The inclusion $\mathcal{F}_{L,\mu} \subseteq \mathcal{F}_{L_\gamma,\mu_\gamma}$ makes the three H-terms below nonnegative. Substituting $x_1 - x_\star = x_0 - x_\star - \gamma\nabla f(x_0)$ and $\gamma = 2/(L_\gamma + \mu_\gamma)$ gives the certificate identity

$$\begin{aligned}
&\rho_\gamma(f(x_0) - f_\star) - (f(x_1) - f_\star) \\
&\quad = \lambda_{\star 0} H_{L_\gamma,\mu_\gamma}(x_\star, x_0) \\
&\qquad + \lambda_{\star 1} H_{L_\gamma,\mu_\gamma}(x_\star, x_1) \\
&\qquad + \lambda_{01} H_{L_\gamma,\mu_\gamma}(x_0, x_1) + R_\gamma.
\end{aligned}$$

Here

$$R_\gamma \triangleq \frac{\gamma}{4\sqrt{\rho_\gamma}} \left\| \nabla f(x_0) + \nabla f(x_1) - \gamma\mu_\gamma L_\gamma (x_0 - x_\star) \right\|^2 .$$

Since the multipliers and R_γ are also nonnegative, this proves the claim. □

Remark 1 (Comparison with original curvatures). For $0 < \mu < L$, setting the nonsmooth term h to zero in the proof of Taylor et al. (2018, Theorem 3.3) recovers this statement. Using the same three active multipliers with the original $\mathcal{F}_{L,\mu}$ co-coercivity inequalities gives instead

$$\begin{aligned} &\rho_\gamma(f(x_0) - f_\star) - (f(x_1) - f_\star) \\ &\quad = \lambda_{\star 0} H_{L,\mu}(x_\star, x_0) \\ &\qquad + \lambda_{\star 1} H_{L,\mu}(x_\star, x_1) \\ &\qquad + \lambda_{01} H_{L,\mu}(x_0, x_1) + R'_\gamma. \end{aligned}$$

Specializing their branchwise slack decomposition to $h \equiv 0$ gives a nonnegative but more complicated residual; its closed forms on the two stepsize branches are given in Appendix G.

A.2. Fast gradient method

For a horizon N, we consider the following parameterization of the smooth-convex fast gradient method, with the k-indexed updates taken over $k \in [\![0, N-2]\!]$:

$$\begin{aligned} y_0 &= x_0, \\ x_N &= y_{N-1} - \frac{1}{L}\nabla f(y_{N-1}), \\ x_{k+1} &= y_k - \frac{1}{L}\nabla f(y_k), \\ y_{k+1} &= x_{k+1} + \frac{k}{k+3}(x_{k+1} - x_k). \end{aligned} \tag{FGM}$$

In contrast with the GD example, this one is posed on the smooth-convex class $\mathcal{F}_L$, with distance normalization $\|x_0 - x_\star\|^2 \leqslant 1$ and terminal performance metric $f(x_N) - f_\star$. The corresponding numerical formulations are collected in Appendix F.2.

Let $H_L(u, v) \triangleq H_{L,0}(u, v)$; for any $f \in \mathcal{F}_L$, one has $H_L(u, v) \geqslant 0$ for all points u, v.

Target rate. Under this normalization, the conjectured optimal worst-case rate for the smooth-convex FGM endpoint is $\rho_\star(N, L) = \frac{2L}{N^2+5N+6}$ (Taylor et al., 2017a, Table 1). The same table gives the relaxed FGM target $\rho_{\mathrm{rel}}(N, L) = \frac{2L}{N^2+5N+2}$. The relative suboptimality criterion is set so that this relaxed rate can be recovered, with the addition of tolerance for noise due to numerics.

A.2.1. ℓ_1-HEURISTICS

For the computer-aided sparsity comparison we use only the base smooth-convex inequalities H_L at $N = 3$. The points $\{x_\star, y_0, y_1, y_2, x_3\}$ generate the ordered two-point inequalities used in the search. At this size, exhaustive enumeration is still practical and provides a useful calibration for the penalty-based searches. Plain ℓ_1 and log-sum are scale-sensitive on this instance and retain denser patterns, while the normalized log-sum heuristic from Appendix D moves the returned pattern closer to the exhaustive one. The active-inequality patterns for the $N = 3$ comparison are detailed in Appendix I. The same comparison across longer horizons is shown in Figure 3.

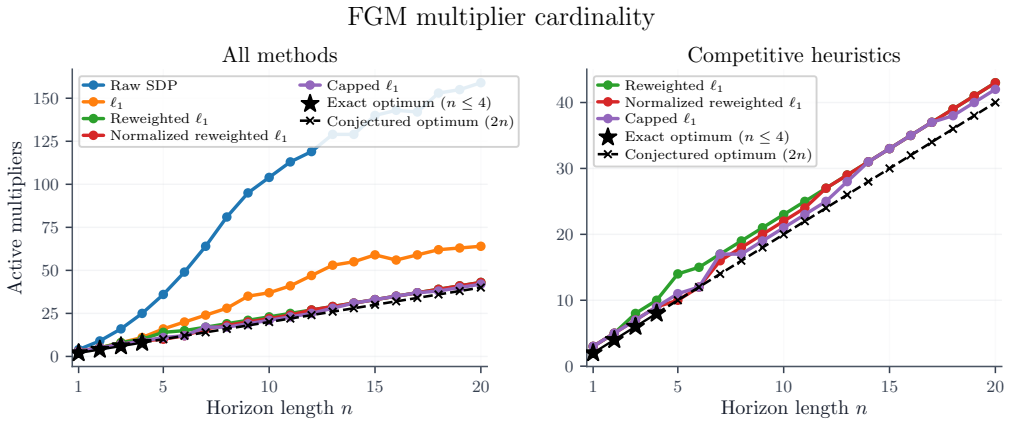

Figure 3. FGM hypothesis complexity across horizon lengths: all methods on the left, and the competitive continuous sparsification heuristics on the right, which track the exact or conjectured sparse active-inequality pattern more closely.

B. Proximal certificates from candidate lemmas

The candidate-lemma search can return a local Lyapunov inequality rather than only a shorter terminal certificate (Taylor & Bach, 2019; Rotaru, 2026). We record two proximal examples in which the SDP search identifies the potential and its one-step decrement; the closed-form proofs are given in Appendix H.

B.1. Proximal point residuals

Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be closed, proper, and convex, and let $x_\star \in \operatorname{argmin} F$ with $F_\star = F(x_\star)$. Fix a self-adjoint positive-definite $B \in \mathbb{S}^d$. For positive stepsizes $\alpha_1, \dots, \alpha_N$, the proximal point method is

$$\begin{aligned} x_k &\in \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ F(x) + \frac{1}{2\alpha_k} \|x - x_{k-1}\|_B^2 \right\}, \\ g_k &\triangleq \frac{B(x_{k-1} - x_k)}{\alpha_k} \in \partial F(x_k). \end{aligned} \tag{PPA}$$

Writing $A_N = \sum_{k=1}^N \alpha_k$, the recovered certificate gives a considerably simpler Lyapunov-function proof of the sharp last-residual estimate conjectured by Taylor et al. (2017a) and proved by Gu & Yang (2025); the same Lyapunov function yields the value bound.

Theorem 2 (PPA residual and value bounds)**.** *If $\|x_0 - x_\star\|_B \leqslant R$, then for every $N \geqslant 1$,*

$$\|g_N\|_{B^{-1}} \leqslant \frac{R}{A_N}, \qquad F(x_N) - F_\star \leqslant \frac{R^2}{4A_N}.$$

B.2. Accelerated proximal point saddle gaps

The second example recovers an operator potential for accelerated PPA and then applies it to a saddle subdifferential. Let $\mathcal{H}_1, \mathcal{H}_2$ be real Hilbert spaces, and let ϕ be a saddle function with $\phi(\cdot, v)$ and $-\phi(u, \cdot)$ closed proper convex. Assume that the saddle subdifferential $\mathcal{M}(u, v) = (\partial\phi(\cdot, v)(u), \partial(-\phi(u, \cdot))(v))$ is maximally monotone on $\mathcal{H}_1 \times \mathcal{H}_2$ and that $0 \in \mathcal{M}(u_\star, v_\star)$. For $\beta > 0$, initialize $x_0 = y_0 = y_{-1} = (u_0, v_0)$ and run the accelerated proximal point method of Kim (2021):

$$\begin{aligned} x_{i+1} &= (\mathrm{Id} + \beta\mathcal{M})^{-1}(y_i), \\ y_{i+1} &= x_{i+1} + \frac{i}{i+2}(x_{i+1} - x_i) \\ &\quad - \frac{i}{i+2}(x_i - y_{i-1}). \end{aligned} \tag{APPM}$$

The following saddle-gap bound was conjectured in Kim (2021).

Theorem 3 (APPM saddle-gap bound)**.** *Writing $x_i = (u_i, v_i)$, assume the saddle-gap quantities below are finite along the generated sequence. Then, for every $N \geqslant 1$,*

$$\phi(u_N, v_\star) - \phi(u_\star, v_N) \leqslant \frac{\|u_0 - u_\star\|^2 + \|v_0 - v_\star\|^2}{4\beta N}.$$

C. General PEP Framework

This section gives the sampled PEP derivation behind the proof structures used throughout the paper. In particular, it explains why first-order PEP dual certificates naturally take the form of identities built from valid inequalities evaluated on finitely many sampled points.

C.1. Sampled-data PEP

Fix a horizon N and let $\mathcal{S}$ be the finite set of point labels used in the proof identity, as in Section 2.1. The sampled data are

$$\mathcal{D} = \{(x_i, f_i, g_i)\}_{i\in\mathcal{S}}, \qquad f_i = f(x_i), \quad g_i \in \partial f(x_i).$$

When a formulation needs additional method-state variables or auxiliary points, we append them to $\mathcal{D}$. Such variables are included in $\mathcal{S}$ only when oracle data are sampled there and interpolation inequalities are imposed on them. Equations used only to define added variables, such as algorithm recurrences, are substituted away in the sampled expressions.

We use the same initialization and performance functionals as above and write them as $\mathcal{I}_f(\mathcal{D}, x_\star)$ and $\mathcal{P}_f(\mathcal{D}, x_\star)$. For the first-order function classes considered in this paper, we use the following class-side interpolation form.

Assumption 1 (Finite interpolation form)**.** The sampled data $\mathcal{D}$ are $\mathcal{F}$-interpolable if and only if

$$H(x_i, x_j) \geqslant 0, \qquad \forall i, j \in \mathcal{S}.$$

Here $\mathcal{F}$-interpolable means that some $f \in \mathcal{F}$ realizes the sampled triples $\{(x_i, f_i, g_i)\}_{i\in\mathcal{S}}$. For product classes, the ordered-pair family is the union of the corresponding component interpolation families.

This is the same hypothesis notation as in Section 2.1. Adding valid redundant hypotheses can change the certificate representation but not the exact PEP; omitting class-defining interpolation inequalities gives a relaxation. Under Assumption 1, the constraints $H(x_i, x_j) \geqslant 0$ in the finite PEP below are exactly the requirement that the sampled oracle data come from some function in $\mathcal{F}$.

The finite sampled-data PEP associated with the introductory worst-case problem can then be written as:

$$\begin{aligned} \underset{\mathcal{D},\ x_\star\in\mathbb{R}^d,\ d\geqslant 1}{\text{maximize}} \quad & \mathcal{P}_f(\mathcal{D}, x_\star) \\ \text{subject to} \quad & (x_t)_{t\in[\![0,N]\!]} \in \mathcal{T}_{\mathcal{A}}(f, N, x_0), \\ & D - \mathcal{I}_f(\mathcal{D}, x_\star) \geqslant 0, \\ & H(x_i, x_j) \geqslant 0, \quad \forall i, j \in \mathcal{S}. \end{aligned} \tag{f-PEP}$$

C.2. Lifted SDP and Lagrangian dual

The sampled PEP above is finite, but it is still written in terms of vectors in an arbitrary dimension. In the examples considered here, every scalar quantity depends on those vectors only through their inner products, together with the sampled function values. We therefore use the standard SDP lift: function values are collected as free scalar coordinates, and inner products are collected in a positive semidefinite Gram matrix. The following assumption records the finite lifted form used below.

Assumption 2 (Gram representability)**.** For the fixed horizon and label set $\mathcal{S}$, the sampled quantities can be encoded by scalar coordinates $s \in \mathbb{R}^p$ and a Gram matrix $G \succcurlyeq 0$ such that the lifted expressions

$$P(s, G) \triangleq \langle s, q_P\rangle + \operatorname{Tr}(GQ_P), \qquad I(s, G) \triangleq \langle s, q_I\rangle + \operatorname{Tr}(GQ_I),$$

represent $\mathcal{P}_f(\mathcal{D}, x_\star)$ and $\mathcal{I}_f(\mathcal{D}, x_\star)$, respectively, and, for every $i, j \in \mathcal{S}$, the lifted expression

$$H_{ij}(s, G) \triangleq \langle s, q_{ij}\rangle + \operatorname{Tr}(GQ_{ij})$$

represents the sampled interpolation inequality $H(x_i, x_j)$.

Under Assumption 2, the lifted SDP is

$$\begin{aligned} \underset{s\in\mathbb{R}^p,\ G\succcurlyeq 0}{\text{maximize}} \quad & P(s,G) \\ \text{subject to} \quad & D - I(s,G) \geqslant 0, \\ & H_{ij}(s,G) \geqslant 0, \qquad \forall i,j \in \mathcal{S}. \end{aligned} \tag{SDP-PEP}$$

The Lagrangian dual of this reduced SDP is

$$\begin{aligned} \underset{\tau,\lambda}{\text{minimize}} \quad & \tau D \\ \text{subject to} \quad & \tau q_I - q_P - \sum_{i,j\in\mathcal{S}} \lambda_{ij} q_{ij} = 0, \\ & \tau Q_I - Q_P - \sum_{i,j\in\mathcal{S}} \lambda_{ij} Q_{ij} \succcurlyeq 0, \\ & \tau \geqslant 0, \qquad \lambda_{ij} \geqslant 0, \quad \forall i,j \in \mathcal{S}. \end{aligned} \tag{D-SDP-PEP}$$

Feasibility of this dual gives a proof identity of the form used in (Proof). Indeed, if

$$R = \tau Q_I - Q_P - \sum_{i,j\in\mathcal{S}} \lambda_{ij} Q_{ij} \succcurlyeq 0,$$

then the scalar equality gives

$$\tau I(s,G) - P(s,G) = \sum_{i,j\in\mathcal{S}} \lambda_{ij} H_{ij}(s,G) + \operatorname{Tr}(GR).$$

Evaluating this lifted identity on sampled data recovers the corresponding identity with $\mathcal{I}_f$, $\mathcal{P}_f$, and $H(x_i, x_j)$. The PSD slack decomposition of $\operatorname{Tr}(GR)$ gives the residual terms R_i. The active hypothesis set is therefore

$$\mathcal{A}_H = \big\{(i,j) \in \mathcal{S}^2 : \lambda_{ij} > 0,\ H(x_i, x_j) \not\equiv 0\big\}.$$

D. Sparse Minimization Heuristics

This appendix gives the details behind Section 3.2. For exposition clarity, we focus on *hypothesis complexity*; *residual complexity* can be handled analogously by penalizing residual-slack structure in the SDP representation.

We now formalize the sparsity-surrogate objectives used in the experiments. Let $\mathcal{C}_{\bar{\rho}}$ denote the certificates in the fixed search space that certify a bound no larger than $\bar{\rho}$. Each such certificate $C \in \mathcal{C}_{\bar{\rho}}$ is defined by pair-indexed nonnegative hypothesis multipliers $(\lambda_{ij})_{(i,j)\in\mathcal{S}^2}$ and active hypothesis set

$$\mathcal{A}_H^C = \big\{(i,j) \in \mathcal{S}^2 : \lambda_{ij} > 0,\ H(x_i, x_j) \not\equiv 0\big\}.$$

We consider separable penalties $\phi_{ij} : \mathbb{R}_+ \to \mathbb{R}$ applied to these multipliers:

$$\underset{C\in\mathcal{C}_{\bar{\rho}}}{\text{minimize}} \quad \sum_{i,j\in\mathcal{S}} \phi_{ij}(\lambda_{ij}). \tag{5}$$

The choices below approximate the number of active hypothesis multipliers. A natural first choice is plain ℓ_1:

$$\phi_{ij}(x) = x,$$

which is the standard lasso-type convex relaxation of sparsity (Tibshirani, 1996). Second, a common nonconvex surrogate for the number of nonzero coordinates is the log-sum penalty:

$$\phi_{ij}(x) = \log(x + \delta),$$

where $\delta > 0$ stabilizes the logarithm near zero. Its first-order majorization gives reweighted ℓ_1 iterations (Candès et al., 2008).

Because the attainable ranges of the multipliers can differ by orders of magnitude, an unnormalized log-sum surrogate can impose much larger effective shrinkage on some multipliers than on others. To remedy this, one can introduce multiplier-specific normalizations, giving the normalized log-sum penalty

$$\phi_{ij}(x) = \log\left(\frac{x}{M_{ij}} + \delta\right),$$

where $M_{ij} > 0$ is a reference scale for the multiplier of $H(x_i, x_j) \geqslant 0$, estimated from its feasible range and clipped to a positive bounded interval in the experiments. To distinguish active multipliers rather than their magnitudes, one can use a capped ℓ_1 penalty:

$$\phi_{ij}(x) = \min\left\{\theta, \frac{x}{M_{ij}}\right\}.$$

The cap is linear for $x/M_{ij} < \theta$ and then saturates, so θ is the normalized activation threshold. Thus the objective behaves like a scaled active-multiplier count rather than a magnitude penalty. This is the standard capped-ℓ_1 sparsity surrogate handled by multi-stage convex relaxation (Zhang, 2010).

In practice, nonlinear choices in (5) are optimized by iterative majorization: at each step, we replace the penalty by a linear upper bound at the current multipliers. Given weights w, define

$$\textsc{Weighted-Cert}(w, \bar{\rho}) \in \underset{C\in\mathcal{C}_{\bar{\rho}}}{\operatorname{argmin}} \sum_{i,j\in\mathcal{S}} w_{ij}\lambda_{ij}. \tag{6}$$

returning a minimizing certificate and its hypothesis multipliers. The optimization problem in (6) is convex and thus solvable efficiently. Along iterations, the relaxed target $\bar{\rho}$ stays fixed; only the selection criterion changes. We update weights using superdifferentials at these multipliers: $w_{ij}^{(t+1)} \in \partial\phi_{ij}(\lambda_{ij}^{(t)})$. For normalized log-sum, $w_{ij}^{(t+1)} = (\lambda_{ij}^{(t)} + \delta M_{ij})^{-1}$, while capped ℓ_1 gives $w_{ij}^{(t+1)} = M_{ij}^{-1}$ below the cap and $w_{ij}^{(t+1)} = 0$ above it, with any supergradient between these values at the threshold. The resulting procedure is summarized in Algorithm 1.

Algorithm 1 Majorized sparsification heuristic

Require: Relaxed target $\bar{\rho}$, coordinate penalties ϕ_{ij}, initial weights $w^{(0)}$, number of iterations T

1: **for** $t \in [\![0, T-1]\!]$ **do**
2: $\quad C^{(t)} \leftarrow$ WEIGHTED-CERT$(w^{(t)}, \bar{\rho})$. We denote $\lambda^{(t)}$ the corresponding multipliers.
3: $\quad$ Choose $w_{ij}^{(t+1)} \in \partial\phi_{ij}(\lambda_{ij}^{(t)}), \forall (i,j) \in \mathcal{S}^2$
4:
5: **return** $C^{(T-1)}$.

The output certificate proves the relaxed target, and $\mathcal{A}_H^{C^{(T-1)}}$ is the selected sparsification pattern.

E. An SDP Formulation for Candidate Lemma Search

This appendix derives the SDP formulation used to search for candidate lemmas. The formulation is obtained from (D-SDP-PEP).

We use the lifted notation $H_{ij}(s,G)$, $I(s,G)$, and $P(s,G)$ from Assumption 2. Let $\mathcal{L}_{\text{cand}}$ index the candidate lemma slots included in the search. For each $\ell \in \mathcal{L}_{\text{cand}}$, fix a set $\mathcal{I}_\ell \subseteq \mathcal{S}^2$ of ordered hypothesis pairs allowed to appear in that candidate lemma. For fixed coefficients and residual matrix, a candidate lemma has the lifted analogue of the proof identities in Section 2.1:

$$\Psi_\ell(s,G) \triangleq \sum_{(i,j)\in\mathcal{I}_\ell} \alpha_{ij}^{(\ell)} H_{ij}(s,G) + \operatorname{Tr}(GR_\ell) \geqslant 0,$$

where

$$\alpha_{ij}^{(\ell)} \geqslant 0, \qquad R_\ell \succcurlyeq 0.$$

Thus $\Psi_\ell \geqslant 0$ is not a new assumption: it is certified by the original hypotheses together with the nonnegative residual term $\operatorname{Tr}(GR_\ell)$. The PSD matrix R_ℓ is the lifted representation of the residual part; a PSD decomposition of $\operatorname{Tr}(GR_\ell)$ gives residual squares of the kind denoted by R_i in (Proof). Equivalently, if $\Psi_\ell(s,G) = \langle s, q_{\Psi_\ell}\rangle + \operatorname{Tr}(GQ_{\Psi_\ell})$, then

$$q_{\Psi_\ell} = \sum_{(i,j)\in\mathcal{I}_\ell} \alpha_{ij}^{(\ell)} q_{ij}, \qquad Q_{\Psi_\ell} = \sum_{(i,j)\in\mathcal{I}_\ell} \alpha_{ij}^{(\ell)} Q_{ij} + R_\ell.$$

If the candidate lemmas $\Psi_\ell \geqslant 0$ are fixed and added to (SDP-PEP), the augmented Lagrangian dual has multipliers $\tau \geqslant 0$, $\lambda_{ij} \geqslant 0$, and $\eta_\ell \geqslant 0$, with constraints

$$\tau q_I - q_P - \sum_{i,j\in\mathcal{S}} \lambda_{ij} q_{ij} - \sum_{\ell\in\mathcal{L}_{\text{cand}}} \eta_\ell q_{\Psi_\ell} = 0,$$

$$R \triangleq \tau Q_I - Q_P - \sum_{i,j\in\mathcal{S}} \lambda_{ij} Q_{ij} - \sum_{\ell\in\mathcal{L}_{\text{cand}}} \eta_\ell Q_{\Psi_\ell} \succcurlyeq 0.$$

Here R denotes the aggregate certificate slack matrix, distinct from the per-slot residual matrices R_ℓ and $\widehat{R}_\ell$. Searching over candidate lemmas would introduce products $\eta_\ell \alpha_{ij}^{(\ell)}$ and $\eta_\ell R_\ell$. Absorb them by setting

$$\zeta_{ij}^{(\ell)} = \eta_\ell \alpha_{ij}^{(\ell)}, \qquad \widehat{R}_\ell = \eta_\ell R_\ell.$$

The absorbed variables satisfy

$$\zeta_{ij}^{(\ell)} \in \mathbb{R}_+, \qquad \widehat{R}_\ell \succcurlyeq 0.$$

Thus the target-restricted SDP feasible set used for candidate-lemma search, with the aggregate slack written after absorption, is

$$\begin{aligned}
\text{find} \quad & \tau, \quad \{\lambda_{ij}\}, \quad \{\zeta_{ij}^{(\ell)}\}, \quad \{\widehat{R}_\ell\} \\
\text{subject to} \quad & 0 \leqslant \tau \leqslant \bar{\rho}, \\
& \lambda_{ij} \geqslant 0, \quad \forall i,j \in \mathcal{S}, \\
& \zeta_{ij}^{(\ell)} \in \mathbb{R}_+, \quad \forall \ell \in \mathcal{L}_{\text{cand}}, \quad (i,j) \in \mathcal{I}_\ell, \\
& \widehat{R}_\ell \succcurlyeq 0, \quad \forall \ell \in \mathcal{L}_{\text{cand}}, \\
& 0 = \tau q_I - q_P - \sum_{i,j\in\mathcal{S}} \lambda_{ij} q_{ij} - \sum_{\ell\in\mathcal{L}_{\text{cand}}} \sum_{(i,j)\in\mathcal{I}_\ell} \zeta_{ij}^{(\ell)} q_{ij}, \\
& R \triangleq \tau Q_I - Q_P - \sum_{i,j\in\mathcal{S}} \lambda_{ij} Q_{ij} \\
& \qquad - \sum_{\ell\in\mathcal{L}_{\text{cand}}} \left(\sum_{(i,j)\in\mathcal{I}_\ell} \zeta_{ij}^{(\ell)} Q_{ij} + \widehat{R}_\ell \right) \succcurlyeq 0.
\end{aligned} \tag{CL-SDP}$$

One can encourage low residual complexity by adding rank-minimization heuristics on the slack R, such as log-det objectives (Fazel et al., 2003), while keeping the relaxed target $\bar{\rho}$ fixed.

After solving (CL-SDP), recover the candidate-lemma multiplier in slot ℓ by setting

$$\eta_\ell \triangleq \sum_{(i,j)\in\mathcal{I}_\ell} \zeta_{ij}^{(\ell)}.$$

If $\eta_\ell > 0$, use the normalization

$$\alpha_{ij}^{(\ell)} = \eta_\ell^{-1}\zeta_{ij}^{(\ell)}, \qquad R_\ell = \eta_\ell^{-1}\widehat{R}_\ell.$$

With these recovered coefficients, $\Psi_\ell(s,G)$ has the form displayed above and appears in the proof identity with multiplier η_ℓ. If $\eta_\ell = 0$, then $\zeta_{ij}^{(\ell)} = 0$ for every $(i,j) \in \mathcal{I}_\ell$, so that slot carries only residual slack. Define

$$\bar{R} \triangleq R + \sum_{\ell\in\mathcal{L}_{\text{cand}}:\ \eta_\ell=0} \widehat{R}_\ell \succcurlyeq 0.$$

Final proof reconstruction. The final proof can be written with selected intermediate lemmas as explicit steps, then expanded back into the original hypothesis language if desired. Every feasible point of (CL-SDP) yields the proof identity

$$\tau I(s,G) - P(s,G) = \sum_{i,j\in\mathcal{S}} \lambda_{ij} H_{ij}(s,G) + \sum_{\ell\in\mathcal{L}_{\text{cand}}:\ \eta_\ell>0} \eta_\ell \Psi_\ell(s,G) + \operatorname{Tr}\left(G\bar{R}\right) \geqslant 0.$$

Expanding each selected Ψ_ℓ recovers a certificate using only the original hypotheses and PSD residual terms, while keeping the selected candidate lemmas visible produces the intermediate-lemma proof identity used in Section 3.3.

F. Detailed PEP Formulations for Experimental Examples

This section instantiates the sampled-data notation of Appendix C for the two experimental examples below and writes the concrete lifted SDP formulations used in the certificate searches. Each finite PEP fixes a label set $\mathcal{S}$, sampled data $\mathcal{D} = \{(x_i, f_i, g_i)\}_{i\in\mathcal{S}}$, and concrete interpolation hypotheses $H(x_i, x_j) \geqslant 0$ on ordered label pairs.

F.1. One-step gradient descent

Let $N = 1$ and $x_1 = x_0 - \gamma\nabla f(x_0)$. The class is $\mathcal{F}_{L,\mu}$ with $0 \leqslant \mu < L$, the initialization constraint is on $\mathcal{I}_f = f(x_0) - f_\star$, and the performance metric is $\mathcal{P}_f = f(x_1) - f_\star$.

Functional-residual PEP. Set $\mathcal{S} = \{0, 1, \star\}$ and $\mathcal{D} = \{(x_i, f_i, g_i)\}_{i\in\mathcal{S}}$, with $g_\star = 0$. The sampled version of (f-PEP) is

$$\begin{aligned} \underset{\substack{x_0,x_1,x_\star,g_0,g_1\in\mathbb{R}^d,\\ f_0,f_1,f_\star\in\mathbb{R},\ d\geqslant 1}}{\text{maximize}} \quad & f_1 - f_\star \\ \text{subject to} \quad & f_0 - f_\star \leqslant 1, \\ & x_1 = x_0 - \gamma g_0, \\ & H_{L,\mu}(x_i, x_j) \geqslant 0, \qquad i, j \in \{0, 1, \star\},\ i \neq j. \end{aligned} \tag{GD-f-PEP}$$

The omitted self-pairs are only the trivial constraints $H_{L,\mu}(x_i, x_i) \equiv 0$. For $i, j \in \mathcal{S}$,

$$H_{L,\mu}(x_i, x_j) \triangleq f_i - f_j - \langle g_j, x_i - x_j\rangle - \frac{1}{2(L-\mu)}\left\|g_i - g_j - \mu(x_i - x_j)\right\|^2 - \frac{\mu}{2}\left\|x_i - x_j\right\|^2.$$

Lifted SDP. Introduce

$$s = (s_0, s_1) = (f_0 - f_\star,\ f_1 - f_\star).$$

The Gram matrix is

$$G = \begin{bmatrix} \|x_0 - x_\star\|^2 & \langle g_0, x_0 - x_\star\rangle & \langle g_1, x_0 - x_\star\rangle \\ \langle g_0, x_0 - x_\star\rangle & \|g_0\|^2 & \langle g_0, g_1\rangle \\ \langle g_1, x_0 - x_\star\rangle & \langle g_1, g_0\rangle & \|g_1\|^2 \end{bmatrix} \in \mathbb{S}^3_+.$$

The update has been substituted through $x_1 - x_\star = (x_0 - x_\star) - \gamma g_0$. Let e_0, e_1 be the canonical basis of $\mathbb{R}^2$, set $q_I = e_0$, $q_P = e_1$, $e_\star = 0$, and $q_{ij} = e_i - e_j$. Define $Q_{ij} \in \mathbb{S}^3$ by

$$\operatorname{Tr}(GQ_{ij}) = -\langle g_j, x_i - x_j\rangle - \frac{1}{2(L-\mu)}\left\|g_i - g_j - \mu(x_i - x_j)\right\|^2 - \frac{\mu}{2}\left\|x_i - x_j\right\|^2.$$

Equivalently, the interpolation residuals are represented as

$$H_{L,\mu}(x_i, x_j) = \langle s, q_{ij}\rangle + \operatorname{Tr}(GQ_{ij}).$$

Thus the reduced lifted problem is

$$\begin{aligned} \underset{s\in\mathbb{R}^2,\ G\in\mathbb{S}^3_+}{\text{maximize}} \quad & \langle s, q_P\rangle \\ \text{subject to} \quad & \langle s, q_I\rangle \leqslant 1, \\ & \langle s, q_{ij}\rangle + \operatorname{Tr}(GQ_{ij}) \geqslant 0, \qquad i, j \in \{0, 1, \star\},\ i \neq j. \end{aligned} \tag{GD-SDP-PEP}$$

Lagrangian dual certificate. The reduced Lagrangian dual is

$$\begin{aligned} \underset{\substack{\tau\geqslant 0,\\ \lambda_{ij}\geqslant 0,\ i,j\in\{0,1,\star\},\ i\neq j}}{\text{minimize}} \quad & \tau \\ \text{subject to} \quad & \tau q_I - q_P - \sum_{\substack{i,j\in\{0,1,\star\}\\ i\neq j}} \lambda_{ij} q_{ij} = 0, \\ & -\sum_{\substack{i,j\in\{0,1,\star\}\\ i\neq j}} \lambda_{ij} Q_{ij} \succcurlyeq 0. \end{aligned} \tag{GD-D-SDP-PEP}$$

F.2. Smooth-convex fast gradient method

For a horizon N, consider the smooth-convex FGM recurrence

$$y_0 = x_0, \qquad x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k), \qquad y_{k+1} = x_{k+1} + \frac{k}{k+3}(x_{k+1} - x_k),$$

with the x-update imposed for $k \in [\![0, N-1]\!]$ and the y-update imposed for $k \in [\![0, N-2]\!]$. The class is $\mathcal{F}_L$, the initialization constraint is $\mathcal{I}_f = \|x_0 - x_\star\|^2$, and the performance measure is $\mathcal{P}_f = f(x_N) - f_\star$.

Function-value PEP. Set

$$\mathcal{S}_N = [\![0, N]\!] \cup \{\star\}, \qquad \bar{x}_i = y_i \quad i \in [\![0, N-1]\!], \qquad \bar{x}_N = x_N, \qquad \bar{x}_\star = x_\star.$$

The sampled oracle data are

$$\mathcal{D}_N = \{(\bar{x}_i, f_i, g_i)\}_{i \in \mathcal{S}_N}, \qquad f_i = f(\bar{x}_i), \quad g_i = \nabla f(\bar{x}_i), \quad g_\star = 0.$$

Then

$$\begin{aligned}
\underset{\substack{\mathcal{D}_N,\, x_0,\dots,x_N,\, y_0,\dots,y_{N-1},\, x_\star,\\ d \geqslant 1}}{\text{maximize}} \quad & f_N - f_\star \\
\text{subject to} \quad & \|x_0 - x_\star\|^2 \leqslant 1, \\
& y_0 = x_0, \\
& x_{k+1} = y_k - \frac{1}{L} g_k, \\
& \qquad \forall k \in [\![0, N-1]\!], \\
& y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}), \\
& \qquad \forall k \in [\![1, N-1]\!], \\
& H_L(\bar{x}_i, \bar{x}_j) \geqslant 0, \\
& \qquad \forall i, j \in \mathcal{S}_N : \ i \neq j.
\end{aligned} \tag{FGM-f-PEP}$$

Here the concrete smooth-convex interpolation hypothesis is the smooth-convex interpolation inequality

$$H_L(\bar{x}_i, \bar{x}_j) \triangleq f_i - f_j - \langle g_j, \bar{x}_i - \bar{x}_j \rangle - \frac{1}{2L}\|g_i - g_j\|^2.$$

Lifted SDP. Use label-based scalar coordinates

$$s = (f_0 - f_\star, \dots, f_N - f_\star) \in \mathbb{R}^{N+1},$$

and let $\mathrm{Gram}\,(v_1, \dots, v_k)$ denote the matrix with entries $\langle v_i, v_j \rangle$. Set

$$G = \mathrm{Gram}\,(x_0 - x_\star, \dots, x_N - x_\star, y_0 - x_\star, \dots, y_{N-1} - x_\star,\ g_0, \dots, g_N) \in \mathbb{S}_+^{3N+2}.$$

Let $e_0, \dots, e_N$ be the canonical basis of $\mathbb{R}^{N+1}$, set $e_\star = 0$, $q_P = e_N$, $q_I = 0$, $Q_I = E_0$, and $q_{i \to j} = e_i - e_j$. We write the smooth-convex interpolation inequalities as the concrete linear forms

$$H_L(\bar{x}_i, \bar{x}_j) = \langle s, q_{i \to j} \rangle + \mathrm{Tr}\,(G Q_{i \to j}) \geqslant 0, \qquad i, j \in \mathcal{S}_N,\ i \neq j,$$

where $Q_{i \to j} \in \mathbb{S}^{3N+2}$ is chosen so that

$$\mathrm{Tr}\,(G Q_{i \to j}) = -\langle g_j, \bar{x}_i - \bar{x}_j \rangle - \frac{1}{2L}\|g_i - g_j\|^2.$$

The lifted problem is

$$\begin{aligned}
\underset{s\in\mathbb{R}^{N+1},\ G\in\mathbb{S}_+^{3N+2}}{\text{maximize}} \quad & \langle s, q_P\rangle \\
\text{subject to} \quad & \operatorname{Tr}(GQ_I) \leqslant 1, \\
& \operatorname{Tr}(M_k G) = 0, \qquad \forall k \in [\![0, N-1]\!], \\
& \operatorname{Tr}(N_0 G) = 0, \\
& \operatorname{Tr}(N_k G) = 0, \qquad \forall k \in [\![1, N-1]\!], \\
& H_L(\bar{x}_i, \bar{x}_j) \geqslant 0, \qquad \forall i, j \in \mathcal{S}_N : \ i \neq j.
\end{aligned} \tag{FGM-SDP-PEP}$$

The coefficient matrices encode

$$\operatorname{Tr}(GQ_I) = \|x_0 - x_\star\|^2, \qquad \operatorname{Tr}(M_k G) = \left\| x_{k+1} - y_k + \frac{1}{L} g_k \right\|^2,$$

$$\operatorname{Tr}(N_0 G) = \|y_0 - x_0\|^2, \qquad \operatorname{Tr}(N_k G) = \left\| y_k - x_k - \frac{k-1}{k+2}(x_k - x_{k-1}) \right\|^2,$$

while the matrices $Q_{i\to j}$ encode the Gram part of the interpolation hypotheses.

Lagrangian dual certificate. The Lagrangian dual, with free multipliers for the method equalities, is

$$\begin{aligned}
\underset{\substack{\tau\geqslant 0,\ \eta_k\in\mathbb{R},\ k\in[\![0,N-1]\!],\\ \zeta_k\in\mathbb{R},\ k\in[\![0,N-1]\!],\\ \lambda_{i\to j}\geqslant 0,\ i,j\in\mathcal{S}_N,\ i\neq j}}{\text{minimize}} \quad & \tau \\
\text{subject to} \quad & \tau q_I - q_P - \sum_{\substack{i,j\in\mathcal{S}_N\\ i\neq j}} \lambda_{i\to j} q_{i\to j} = 0, \\
& \tau Q_I + \sum_{k\in[\![0,N-1]\!]} \eta_k M_k + \sum_{k\in[\![0,N-1]\!]} \zeta_k N_k \\
& \qquad - \sum_{\substack{i,j\in\mathcal{S}_N\\ i\neq j}} \lambda_{i\to j} Q_{i\to j} \succcurlyeq 0.
\end{aligned} \tag{FGM-D-SDP-PEP}$$

The accompanying FGM example code eliminates the recurrence equalities before solving; this produces an equivalent reduced Lagrangian dual with the same interpolation multiplier labels $\lambda_{i\to j}$.

G. Original-Curvature Residuals for Gradient Descent

The comparison in Remark 1 uses the residual obtained by specializing the branchwise slack decomposition of Taylor et al. (2018, Theorem 3.3) to $h \equiv 0$. Let

$$\alpha_\gamma^- \triangleq 2L(2-\gamma\mu) - \mu(2-\gamma\mu)^2 - \gamma^2 L^2 \mu, \qquad \beta_\gamma^- \triangleq 2 - \gamma(L+\mu),$$
$$\alpha_\gamma^+ \triangleq \gamma L\big(L^2+\mu^2\big) - 2\big(L^2 - L\mu + \mu^2\big), \qquad \beta_\gamma^+ \triangleq \gamma(L+\mu) - 2.$$

Then

$$R'_\gamma \triangleq \begin{cases} R_\gamma^-, & 0 < \gamma \leqslant \frac{2}{L+\mu}, \\ R_\gamma^+, & \frac{2}{L+\mu} \leqslant \gamma \leqslant \frac{2}{L}, \end{cases} \tag{7}$$

where

$$\begin{aligned} R_\gamma^- \triangleq\ & \frac{(2-\gamma\mu)\beta_\gamma^-}{2\alpha_\gamma^-} \left\|(1-\gamma\mu)\nabla f(x_0) - \nabla f(x_1)\right\|^2 \\ & + \frac{\gamma L \mu^2 (2-\gamma\mu)}{2(L-\mu)} \left\| x_0 - x_\star - \frac{\nabla f(x_0) + \nabla f(x_1)}{\mu(2-\gamma\mu)} \right\|^2 \\ & + \frac{\gamma\mu\alpha_\gamma^-}{2L(L-\mu)(2-\gamma\mu)} \left\| \frac{(\gamma\mu-1)L\beta_\gamma^-}{\alpha_\gamma^-} \nabla f(x_0) + \frac{L\beta_\gamma^-}{\alpha_\gamma^-} \nabla f(x_1) \right\|^2, \\ R_\gamma^+ \triangleq\ & \frac{(2-\gamma L)\beta_\gamma^+}{2\gamma\alpha_\gamma^+} \left\|(1-\gamma L)\nabla f(x_0) - \nabla f(x_1)\right\|^2 \\ & + \frac{\gamma L^2 \mu (2-\gamma L)}{2(L-\mu)} \left\| x_0 - x_\star + \frac{1-\gamma L-\gamma\mu}{\gamma L\mu} \nabla f(x_0) - \frac{1}{\gamma L\mu} \nabla f(x_1) \right\|^2 \\ & + \frac{\gamma\alpha_\gamma^+}{2\mu(L-\mu)} \left\| \frac{(\gamma L-1)L\beta_\gamma^+}{\gamma\alpha_\gamma^+} \nabla f(x_0) + \frac{L\beta_\gamma^+}{\gamma\alpha_\gamma^+} \nabla f(x_1) \right\|^2. \end{aligned}$$

H. Proximal Certificate Proofs

H.1. Proximal point residual certificate

Theorem 2 (PPA residual and value bounds)**.** *If $\|x_0 - x_\star\|_B \leqslant R$, then for every $N \geqslant 1$,*

$$\|g_N\|_{B^{-1}} \leqslant \frac{R}{A_N}, \qquad F(x_N) - F_\star \leqslant \frac{R^2}{4A_N}.$$

Proof of Theorem 2. Set $A_0 = 0$ and $A_k = \sum_{i=1}^k \alpha_i$. For sampled points define the convexity gaps

$$C_{i,j} \triangleq F(x_i) - F(x_j) - \langle g_j, x_i - x_j\rangle, \qquad C_{\star,j} \triangleq F_\star - F(x_j) - \langle g_j, x_\star - x_j\rangle.$$

Since $g_j \in \partial F(x_j)$, all these quantities are nonnegative. Define

$$\mathcal{V}_k \triangleq \|x_k - x_\star\|_B^2 + 2A_k(F(x_k) - F_\star) + A_k^2 \|g_k\|_{B^{-1}}^2 .$$

Using $x_0 - x_1 = \alpha_1 B^{-1} g_1$,

$$\|x_0 - x_\star\|_B^2 - \mathcal{V}_1 = 2\alpha_1(\langle g_1, x_1 - x_\star\rangle - F(x_1) + F_\star) = 2\alpha_1 C_{\star,1} \geqslant 0.$$

Thus $\mathcal{V}_1 \leqslant \|x_0 - x_\star\|_B^2$.

For the decrement, fix $k \in [\![1, N-1]\!]$ and set $A = A_k$, $a = \alpha_{k+1}$, and $h = g_{k+1}$. From $x_k - x_{k+1} = aB^{-1}h$,

$$\|x_k - x_\star\|_B^2 - \|x_{k+1} - x_\star\|_B^2 = 2a\langle h, x_{k+1} - x_\star\rangle + a^2 \|h\|_{B^{-1}}^2$$

and

$$F(x_k) - F(x_{k+1}) = C_{k,k+1} + a\|h\|_{B^{-1}}^2 .$$

Substitution into $\mathcal{V}_k - \mathcal{V}_{k+1}$ gives

$$\mathcal{V}_k - \mathcal{V}_{k+1} = 2aC_{\star,k+1} + 2AC_{k,k+1} + A^2\Big(\|g_k\|_{B^{-1}}^2 - \|g_{k+1}\|_{B^{-1}}^2\Big).$$

The adjacent gaps satisfy

$$C_{k,k+1} + C_{k+1,k} = \langle g_k - g_{k+1}, x_k - x_{k+1}\rangle = a\langle g_k - g_{k+1}, B^{-1}g_{k+1}\rangle,$$

so

$$\|g_k\|_{B^{-1}}^2 - \|g_{k+1}\|_{B^{-1}}^2 = \frac{2}{a}(C_{k,k+1} + C_{k+1,k}) + \|g_k - g_{k+1}\|_{B^{-1}}^2 .$$

Therefore

$$\mathcal{V}_k - \mathcal{V}_{k+1} = 2aC_{\star,k+1} + \left(2A + \frac{2A^2}{a}\right)C_{k,k+1} + \frac{2A^2}{a}C_{k+1,k} + A^2\|g_k - g_{k+1}\|_{B^{-1}}^2 \geqslant 0.$$

Consequently $\mathcal{V}_N \leqslant \mathcal{V}_1 \leqslant \|x_0 - x_\star\|_B^2 \leqslant R^2$. Since $F(x_N) - F_\star \geqslant 0$, this implies

$$A_N^2 \|g_N\|_{B^{-1}}^2 \leqslant \mathcal{V}_N \leqslant R^2,$$

and hence $\|g_N\|_{B^{-1}} \leqslant R/A_N$. Finally, convexity gives $F(x_N) - F_\star \leqslant \langle g_N, x_N - x_\star\rangle$, and Young's inequality in the B/B^{-1} pair yields

$$2A_N(F(x_N) - F_\star) \leqslant \|x_N - x_\star\|_B^2 + A_N^2 \|g_N\|_{B^{-1}}^2 .$$

Adding another copy of $2A_N(F(x_N) - F_\star)$ to both sides gives $4A_N(F(x_N) - F_\star) \leqslant \mathcal{V}_N \leqslant R^2$, proving the value bound. □

H.2. Accelerated proximal point saddle-gap certificate

Theorem 3 (APPM saddle-gap bound)**.** *Writing $x_i = (u_i, v_i)$, assume the saddle-gap quantities below are finite along the generated sequence. Then, for every $N \geqslant 1$,*

$$\phi(u_N, v_\star) - \phi(u_\star, v_N) \leqslant \frac{\|u_0 - u_\star\|^2 + \|v_0 - v_\star\|^2}{4\beta N}.$$

We first prove an operator estimate. Let $\mathcal{H}$ be a real Hilbert space, let $\mathcal{M} : \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone, let $x_\star \in \operatorname{zer} \mathcal{M}$, and fix $\beta > 0$. Starting from $x_0 = y_0 = y_{-1}$, define

$$x_{i+1} = (\mathrm{Id} + \beta\mathcal{M})^{-1}(y_i), \qquad y_{i+1} = x_{i+1} + \frac{i}{i+2}(x_{i+1} - x_i) - \frac{i}{i+2}(x_i - y_{i-1}).$$

For $i \geqslant 1$, set

$$r_i \triangleq y_{i-1} - x_i, \qquad q_i \triangleq \beta^{-1} r_i.$$

The resolvent equation gives $q_i \in \mathcal{M}x_i$.

Lemma 1 (APPM operator estimate)**.** *For every $N \geqslant 1$,*

$$\langle x_N - x_\star, y_{N-1} - x_N \rangle \leqslant \frac{\|y_0 - x_\star\|^2}{4N}.$$

Proof. The recurrence implies, by induction, the trajectory identities

$$y_i = y_0 - \frac{2}{i+1}\sum_{k=1}^{i} k r_k, \qquad x_i = y_0 - \frac{2}{i}\sum_{k=1}^{i-1} k r_k - r_i, \qquad i \geqslant 1.$$

Set $z = y_0 - x_\star$, $A_i = \sum_{k=1}^{i} k r_k$ with $A_0 = 0$, and $d_i = x_i - x_\star$. Define

$$\mathcal{V}_i \triangleq i\langle d_i, r_i \rangle + \left\| i r_i - \frac{1}{2} z \right\|^2.$$

Using the trajectory formula, $d_i = z - \frac{2}{i} A_{i-1} - r_i$, and hence

$$\mathcal{V}_i = \frac{1}{4}\|z\|^2 - 2\langle A_{i-1}, r_i \rangle + i(i-1)\|r_i\|^2.$$

A direct subtraction, using $A_{i-1} = A_{i-2} + (i-1) r_{i-1}$ and the same trajectory formula for $d_{i-1} - d_i$, gives, for $i \geqslant 2$,

$$i(i-1)(d_{i-1} - d_i) = i(i-1)(r_{i-1} + r_i) - 2A_{i-1}.$$

Taking the inner product with $r_{i-1} - r_i$ gives

$$i(i-1)\langle d_{i-1} - d_i, r_{i-1} - r_i \rangle = F_i - F_{i-1}, \qquad F_i \triangleq 2\langle A_{i-1}, r_i \rangle - i(i-1)\|r_i\|^2.$$

Since $\mathcal{V}_i = \frac{1}{4}\|z\|^2 - F_i$, we obtain

$$\mathcal{V}_i - \mathcal{V}_{i-1} = -i(i-1)\langle x_{i-1} - x_i, r_{i-1} - r_i \rangle.$$

Since $r_j = \beta q_j$ with $q_j \in \mathcal{M}x_j$,

$$\langle x_{i-1} - x_i, r_{i-1} - r_i \rangle = \beta\langle x_{i-1} - x_i, q_{i-1} - q_i \rangle \geqslant 0$$

by monotonicity. Thus $\mathcal{V}_i \leqslant \mathcal{V}_{i-1}$. At $i = 1$, $x_1 = y_0 - r_1$, so

$$\mathcal{V}_1 = \langle y_0 - x_\star - r_1, r_1 \rangle + \left\| r_1 - \frac{1}{2}(y_0 - x_\star) \right\|^2 = \frac{1}{4}\|y_0 - x_\star\|^2.$$

Therefore $\mathcal{V}_N \leqslant \|y_0 - x_\star\|^2/4$. Dropping the square term in $\mathcal{V}_N$ gives

$$N\langle x_N - x_\star, r_N \rangle \leqslant \mathcal{V}_N \leqslant \frac{1}{4}\|y_0 - x_\star\|^2.$$

Since $r_N = y_{N-1} - x_N$, the claim follows. □

Proof of Theorem 3. Apply Lemma 1 on the product Hilbert space $\mathcal{H}_1 \times \mathcal{H}_2$ with $x_\star = (u_\star, v_\star)$. Write $q_N = (p_N, s_N) \in \mathcal{M}(u_N, v_N)$. Then $p_N \in \partial\phi(\cdot, v_N)(u_N)$ and $s_N \in \partial(-\phi(u_N, \cdot))(v_N)$. Convexity in u and convexity of $-\phi(u_N, \cdot)$ give

$$\phi(u_N, v_N) - \phi(u_\star, v_N) \leqslant \langle p_N, u_N - u_\star \rangle, \qquad \phi(u_N, v_\star) - \phi(u_N, v_N) \leqslant \langle s_N, v_N - v_\star \rangle.$$

Adding and using $r_N = \beta q_N$ yields

$$\phi(u_N, v_\star) - \phi(u_\star, v_N) \leqslant \langle x_N - x_\star, q_N \rangle = \beta^{-1} \langle x_N - x_\star, r_N \rangle.$$

The operator estimate gives

$$\phi(u_N, v_\star) - \phi(u_\star, v_N) \leqslant \frac{\|y_0 - x_\star\|^2}{4\beta N} = \frac{\|u_0 - u_\star\|^2 + \|v_0 - v_\star\|^2}{4\beta N},$$

as claimed. □

I. FGM Active Multiplier Patterns

For the FGM sparsity comparison, the conjectured $2N$ active-multiplier pattern for the interpolation constraints associated with the function class $\mathcal{F}_L$ is

$$\mathcal{S}_{2N} \triangleq \{x_\star \to y_k : \ k \in [\![0, N-1]\!]\} \cup \{y_k \to y_{k+1} : \ k \in [\![0, N-2]\!]\} \cup \{y_{N-1} \to x_N\}.$$

Here $u \to v$ denotes the $\mathcal{F}_L$ interpolation constraint with source sample u and target sample v. In the FGM label convention of Appendix F.2, these three blocks correspond respectively to the active Lagrangian dual multipliers $\lambda_{\star\to k}$, $\lambda_{k\to k+1}$, and $\lambda_{N-1\to N}$. Thus the middle block pairs consecutive momentum points y_k and y_{k+1}, while the only edge involving the final endpoint is the terminal pair $y_{N-1} \to x_N$. This pattern has size $N + (N-1) + 1 = 2N$. For instance, when $N = 3$ the pattern is

$$\{x_\star \to y_0,\ x_\star \to y_1,\ x_\star \to y_2, y_0 \to y_1,\ y_1 \to y_2,\ y_2 \to x_3\}.$$

The following spy-style table records the active interpolation-multiplier patterns used in the $N = 3$ FGM sparsity comparison. Rows are source points and columns are target points: a bullet in row u and column v means that the corresponding multiplier $\lambda_{u\to v}$ is above the numerical active-set threshold, while an empty cell means that it is inactive. The final panel is the conjectured $2N$ pattern above, which matches the exhaustive active pattern for this instance. The table is generated from the reduced FGM formulation equivalent to (FGM-D-SDP-PEP).

Table 1. Active interpolation-multiplier patterns for the FGM $N = 3$ example. Rows are sources and columns are targets; a bullet marks $|\lambda_{i\to j}| > 10^{-7}$ and dashes mark self-pairs.

Raw (total 16)

	$x_\star$	y_0	y_1	y_2	x_3
$x_\star$	–	•	•	•	•
y_0		–	•	•	•
y_1		•	–	•	•
y_2		•	•	–	•
x_3		•	•	•	–

Plain ℓ_1 (total 8)

	$x_\star$	y_0	y_1	y_2	x_3
$x_\star$	–	•	•	•	•
y_0		–		•	•
y_1			–		•
y_2				–	•
x_3					–

Log-sum (total 8)

	$x_\star$	y_0	y_1	y_2	x_3
$x_\star$	–	•	•	•	•
y_0		–		•	•
y_1			–		•
y_2				–	•
x_3					–

Norm. log-sum (total 7)

	$x_\star$	y_0	y_1	y_2	x_3
$x_\star$	–	•	•	•	•
y_0		–	•		
y_1			–	•	
y_2				–	•
x_3					–

Capped ℓ_1 (total 7)

	$x_\star$	y_0	y_1	y_2	x_3
$x_\star$	–	•	•	•	•
y_0		–	•		
y_1			–	•	
y_2				–	•
x_3					–

Conjecture (total 6)

	$x_\star$	y_0	y_1	y_2	x_3
$x_\star$	–	•	•	•	
y_0		–	•		
y_1			–	•	
y_2				–	•
x_3					–