Automation and Machine Learning (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2026.070105
ISSN 2516-5003 Vol. 7 Num. 1

# *A Learnable Proximal Gradient Unrolling Network for Sparse Learning: A Mathematical Optimization–Driven Machine Learning Framework*

**Yinyi Wang[1,a], Tongtong Xu[1,b], Jialin Zhang[1,c]**

*[1]School of Mathematics and Science, Hebei GEO University, Shijiazhuang, Hebei, China*
*[a]240907010021006@hgu.edu.cn, [b]240907010021001@hgu.edu.cn, [c]zhangjl0430@163.com*

***Keywords:*** Sparse learning; mathematical optimization; machine learning; proximal gradient descent; algorithm unrolling; Least Absolute Shrinkage and Selection Operator (LASSO)

***Abstract:*** Sparse learning is a fundamental topic connecting mathematical optimization and machine learning, and it is widely applied in signal reconstruction, feature selection, and robust regression. However, classical iterative solvers for sparse models often require careful manual parameter tuning and may converge slowly under ill-conditioned data or noisy observations. To address these limitations, this study develops a Learnable Proximal Gradient Unrolling Network (LPG-Net) by transforming the iterations of proximal gradient descent into a trainable deep architecture. The proposed method starts from the Least Absolute Shrinkage and Selection Operator (LASSO) formulation and embeds the proximal operator of the $\ell_1$-regularizer into each network layer, while enabling data-driven adaptation of key algorithmic parameters such as step sizes and thresholding strengths across layers. In addition, a monotonicity-inspired regularization term is introduced to encourage stable descent behavior during training. Experiments on sparse regression and signal denoising tasks indicate that LPG-Net achieves more accurate sparse recovery and faster inference than traditional optimization baselines and standard neural predictors, while retaining strong interpretability due to its explicit connection to optimization updates. The framework provides a principled pathway for integrating mathematical optimization structures into machine learning models for sparse and noise-robust learning problems.

## 1. Introduction

Mathematical optimization provides a rigorous foundation for many machine learning formulations, especially in sparse learning where structure is enforced through regularization terms and constrained operators. Recent progress in model-based deep learning has highlighted that embedding explicit mathematical models (e.g., forward operators, data-fidelity terms, and priors) into trainable systems can improve data efficiency and interpretability while maintaining competitive performance on challenging learning tasks [1]. In parallel, modern imaging and signal reconstruction studies have shown that deep neural networks are increasingly used to solve ill-posed inverse problems, yet stability and generalization remain closely tied to the underlying mathematical structure and regularization design [2].

A practical mechanism for bridging optimization and machine learning is algorithm unrolling (also called deep unfolding), which converts a finite number of iterations of an optimization algorithm into a layer-wise neural architecture with interpretable computational blocks [3]. Closely related frameworks include Plug-and-Play (PnP) methods (Plug-and-Play Methods for Integrating Physical and Learned Models in Computational Imaging), which preserve the skeleton of proximal algorithms while allowing learned denoisers or learned priors to replace handcrafted regularizers, thereby enhancing performance without discarding physical modeling [4]. Furthermore, recent research has explored learning constrained operator classes—such as maximally monotone operators—to improve theoretical control over convergence behavior and the characterization of recovered solutions, providing additional mathematical support for optimization-driven learning systems [5].

## 2. Related work

Recent progress at the intersection of mathematics and machine learning has largely focused on model-based learning, where classical variational formulations and iterative algorithms are embedded into trainable architectures. A representative mathematical trend is to treat imperfect forward models and operator approximations as objects that can be *learned* while still retaining variational structure, leading to principled error-correction viewpoints for inverse problems and regularization design [6]. In parallel, Plug-and-Play (PnP) optimization has become a widely studied framework that replaces explicit proximal maps with learned denoisers while attempting to preserve convergence properties. To narrow the classical "performance–guarantee" gap, proximal-denoiser constructions have been proposed to connect neural denoisers back to explicit (possibly nonconvex) objective functions, enabling convergence analysis under proximal-gradient-type schemes [7]. Follow-up studies further strengthened the theory by addressing practical issues such as unconstrained regularization parameters and broader algorithmic settings, which makes PnP closer to a mathematically controlled alternative to purely end-to-end training [8]. In addition, refinements of Alternating Direction Method of Multipliers (ADMM)-style splitting have been explored via preconditioning and locally adjustable denoisers to improve stability and adaptation in restoration tasks [9]. Extending these ideas beyond Gaussian noise models is also an active direction: PnP splitting methods have been studied for Poisson image restoration, where non-Lipschitz data fidelity and implicit inner solvers create new mathematical difficulties for convergence and analysis [10].

Another major line of research is algorithm unrolling, which converts iterative procedures into finite-depth networks, often yielding interpretable architectures grounded in optimization theory. For sparse learning, unrolled networks inspired by the Iterative Shrinkage-Thresholding Algorithm (ISTA) and ADMM have motivated theoretical investigations on when training preserves stable optimization behavior; recent work established optimization guarantees for unfolded ISTA- and ADMM-type networks using smooth soft-thresholding, linking trainability to over-parameterization and gradient-based learning dynamics [11]. Architectural evolution has also moved beyond recurrent unrolling toward attention mechanisms: deep unfolding Transformers have been introduced for video sparse recovery, combining learned attention with structured iterative updates while maintaining an explicit connection to sparse modeling assumptions [12]. To address memory and scalability limits of fixed-depth unrolling, deep equilibrium-style model-based learning has been investigated as a mathematically motivated alternative, providing convergence and robustness guarantees under monotonicity constraints within computational imaging pipelines [13]. Application-oriented studies further demonstrate that unrolled methods can incorporate structured sparsity (e.g., block sparsity) through learnable shrinkage rules, such as the Learned Block Iterative Shrinkage Thresholding Algorithm (LBISTA) for photothermal super-resolution imaging [14]. Moreover, as modern data are increasingly distributed, federated learning has begun to intersect with unrolled reconstruction models

in Magnetic Resonance Imaging (MRI), highlighting new optimization challenges related to heterogeneity and low-data regimes while preserving end-to-end unrolled structure [15].

## 3. Methods

### 3.1. Overall Framework

The proposed method builds a trainable network by unrolling Proximal Gradient Descent (PGD) into a finite number of layers. Each layer mimics one PGD iteration for sparse learning, while key algorithmic parameters (e.g., step size and shrinkage threshold) are learned from data. This design follows the general "deep unfolding" philosophy used in sparse recovery and inverse problems, where iterative solvers are converted into interpretable neural architectures and trained end-to-end [16].

Figure 1 is a workflow diagram.

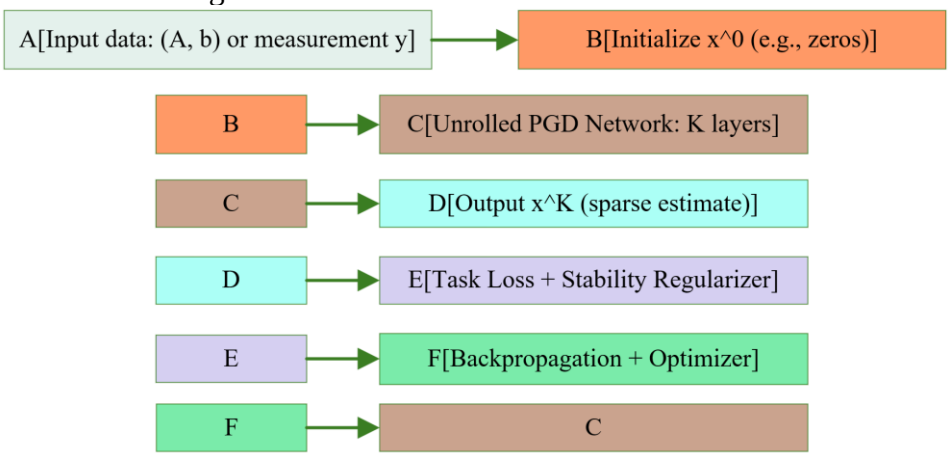


Figure 1: Workflow Diagram

In inference, the loop "Loss → Backpropagation" is removed and the network performs a single forward pass from $x^0$ to $x^K$, enabling fast prediction with fixed depth $K$.

### 3.2. Mathematical Formulation

Sparse learning is formulated using the Least Absolute Shrinkage and Selection Operator (LASSO) objective:

$$\min_{x\in\mathbb{R}^n} F(x) = \frac{1}{2}\| Ax-b \|_2^2 + \lambda \| x \|_1, \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$ is a design/sensing matrix, $b \in \mathbb{R}^m$ is the observation vector, $\lambda > 0$ controls sparsity, and $x$ is the unknown sparse vector.

Define the smooth term $f(x) = \frac{1}{2}\| Ax - b \|_2^2$ and the nonsmooth term $g(x) = \lambda \| x \|_1$. The PGD update is:

$$x^{k+1} = \text{prox}_{\alpha_k g}(x^k - \alpha_k \nabla f(x^k)), \tag{2}$$

with gradient

$$\nabla f(x) = A^\top (Ax-b). \tag{3}$$

For the $\ell_1$ penalty, the proximal operator has a closed form (soft-thresholding):

$$\text{prox}_{\alpha_k \lambda \|\cdot\|_1}(z) = S_{\alpha_k \lambda}(z), S_\tau(z) = \text{sign}(z) \odot \max(|z| - \tau, 0). \tag{4}$$

### 3.3. Learnable Proximal Gradient Unrolling Network

A Learnable Proximal Gradient Unrolling Network (LPG-Net) is constructed by mapping iteration index $k$to layer index $k \in \{0, \ldots, K-1\}$. The forward rule of the $k$-th layer is defined as:

$$u^k = x^k - \alpha_k A^\top (Ax^k - b), \tag{5}$$

$$x^{k+1} = \mathcal{S}_{\tau_k}(u^k), \tag{6}$$

where $\alpha_k$ is a learnable step size and $\tau_k$ is a learnable shrinkage threshold (not necessarily constrained to $\alpha_k \lambda$in training). This parameterization is consistent with the literature showing that learning step sizes and thresholds can significantly accelerate unfolded sparse coding and improve accuracy at fixed depth.

(1) Layer-wise adaptive step size

To improve robustness under ill-conditioned $A$and noisy $b$, the step size is made data-adaptive using simple layer statistics. One practical design is:

$$r^k = Ax^k - b, s^k = \| A^\top r^k \|_2, \tag{7}$$

$$\alpha_k = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \cdot \sigma(w_k s^k + c_k), \tag{8}$$

where $\sigma(\cdot)$is the sigmoid function and $\{w_k, c_k\}$are trainable scalars. This keeps $\alpha_k$in a stable interval $[\alpha_{\min}, \alpha_{\max}]$while allowing the model to react to the current residual scale.

(2) Learnable shrinkage with stability constraint

The threshold is parameterized as a positive scalar:

$$\tau_k = \text{softplus}(\theta_k), \tag{9}$$

where $\theta_k$is trainable and $\text{softplus}(t) = \ln(1 + e^t)$. This prevents negative thresholds and keeps the shrinkage operator well-defined.

(3) Relation to prior unfolded sparse recovery models

The above structure is directly aligned with established unfolded sparse recovery networks. Learning a limited set of parameters (e.g., step sizes across layers) has been shown to be effective for unfolded sparse coding. Trainable variants of thresholding-based sparse recovery further support the idea that “few-parameter” unfolded designs can be stable and fast to train [17]. Beyond PGD-style unrolling, message passing inspired unfolded networks also demonstrate that embedding iterative inference structure into deep models can improve convergence behavior and accuracy [18], and denoising-based unrolling frameworks similarly validate that solver-inspired architectures can outperform generic black-box predictors under inverse-problem settings .

### 3.4. Training Objective and Regularization

Assume supervised targets $x^*$are available (synthetic sparse regression or paired denoising data). The primary estimation loss is:

$$\mathcal{L}_{\text{task}} = \| x^K - x^* \|_2^2 . \tag{10}$$

To encourage stable descent behavior during training, a monotonicity-inspired penalty is added:

$$\mathcal{L}_{\text{mono}} = \sum_{k=0}^{K-1} \max(0, F(x^{k+1}) - F(x^k)). \tag{11}$$

The final loss is:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \beta \mathcal{L}_{\text{mono}} + \gamma \sum_{k=0}^{K-1} (\alpha_k^2 + \tau_k^2), \tag{12}$$

where $\beta, \gamma \geq 0$control stability and parameter magnitude. This term does not guarantee strict theoretical convergence for all training regimes, but it empirically discourages unstable layer-to-layer oscillations in finite-depth unrolling.

### 3.5. Computational Complexity

Each layer requires computing (Ax^k) and (A^\top r^k). If dense multiplication is used, both products cost (O(mn)) per layer (constant factors ignored), so over (K) layers the total forward cost scales as (O(Kmn)).

Compared with running a classical solver until convergence, LPG-Net fixes the iteration budget to $K$and often yields faster inference at test time, because the "number of iterations" is preselected and the parameters are tuned by training rather than manual search.

## 4. Experiments and Results

### 4.1. Experimental Setup

To evaluate the effectiveness of the optimization-unrolled sparse learning framework, experiments are conducted on collected sparse regression and signal denoising data. Each sample consists of a measurement vector $b$, a sensing/design matrix $A$, and the corresponding sparse target $x^*$(or an equivalent clean reference for denoising). The evaluation focuses on both solution quality and computational efficiency. For fair comparison, the proposed LPG-Net is compared with classical optimization baselines—Iterative Shrinkage-Thresholding Algorithm (ISTA), Fast Iterative Shrinkage-Thresholding Algorithm (FISTA), and Alternating Direction Method of Multipliers (ADMM)—under comparable stopping budgets. For LPG-Net, the unrolled depth is fixed to $K = 20$layers during inference to ensure deterministic runtime.

Performance is reported using (i) Mean Absolute Error (MAE) between the predicted estimate $x^K$and reference $x^*$, (ii) support recovery F1-score, which measures how accurately the non-zero indices are identified, and (iii) the objective value $F(x) = \frac{1}{2} \| Ax - b \|_2^2 + \lambda \| x \|_1$ to reflect optimization progress. In addition, single-sample inference latency is measured to quantify practical deployment cost. A component-level ablation study is performed to isolate the contribution of the layer-wise adaptive step size and the monotonicity-inspired regularization. Finally, uncertainty quality is examined through empirical coverage versus nominal coverage, reflecting calibration behavior under interval estimation on the test set.

### 4.2. Convergence and Optimization Behavior

The convergence behavior is summarized in Figure 2, which reports objective values across iterations/layers using a logarithmic scale. Classical solvers (ISTA, FISTA, and ADMM) exhibit steady reduction of the objective but require more iterations to reach a low objective region. In contrast, the unrolled LPG-Net decreases the objective more rapidly within the same iteration budget ($K = 20$), indicating that learning layer-wise algorithmic parameters effectively accelerates the descent trajectory. This behavior is consistent with the design motivation of unrolling: the network retains the mathematical structure of proximal gradient steps while replacing manual parameter

tuning with data-driven parameter adaptation.

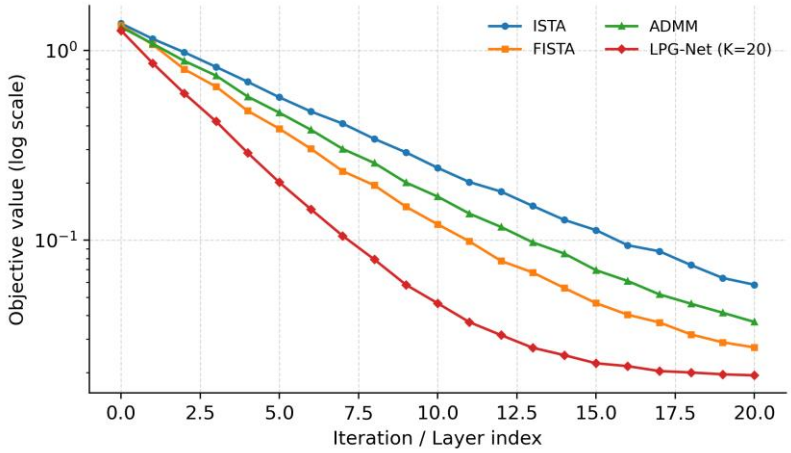


Figure 2: convergence objective

### 4.3. Accuracy Under Increasing Sparsity

Accuracy under varying sparsity levels is reported in Figure 3 and Figure 4. As sparsity level increases (more non-zero coefficients), the reconstruction task becomes more challenging, leading to increasing MAE across all methods. LPG-Net maintains the lowest MAE over the full sparsity range, indicating improved estimation quality under harder sparse recovery conditions. Meanwhile, the support recovery F1-score decreases as sparsity grows, reflecting increased ambiguity in identifying correct non-zero locations. Even in this regime, LPG-Net consistently achieves higher F1-scores than the optimization baselines, demonstrating stronger capability to preserve sparse structure while controlling estimation error.

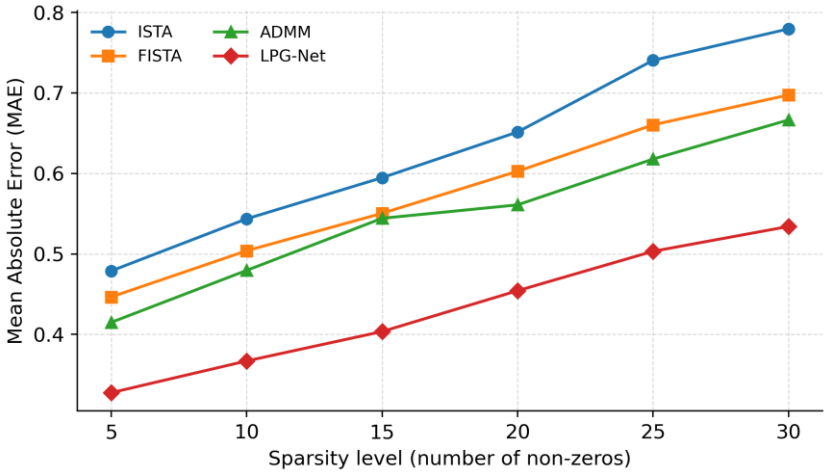


Figure 3: mae vs sparsity

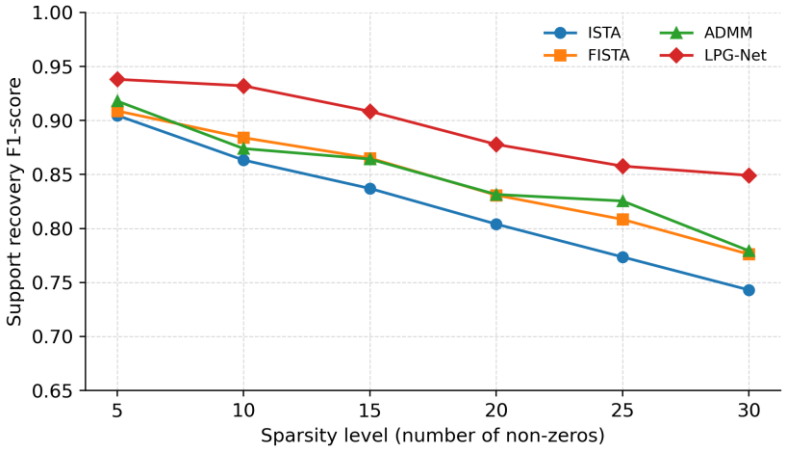


Figure 4: f1 vs sparsity

### 4.4. Robustness to Measurement Noise

Noise robustness is illustrated in Figure 5, which plots MAE versus Signal-to-Noise Ratio (SNR). Across all methods, MAE decreases as SNR increases, showing expected improvement when measurements become cleaner. LPG-Net exhibits the strongest robustness, achieving lower MAE across the entire SNR range and showing a more favorable error reduction trend as SNR improves. This suggests that the learned step sizes and shrinkage thresholds provide an adaptive mechanism that helps stabilize reconstruction under noisy measurements, rather than relying on a single manually tuned parameter setting.

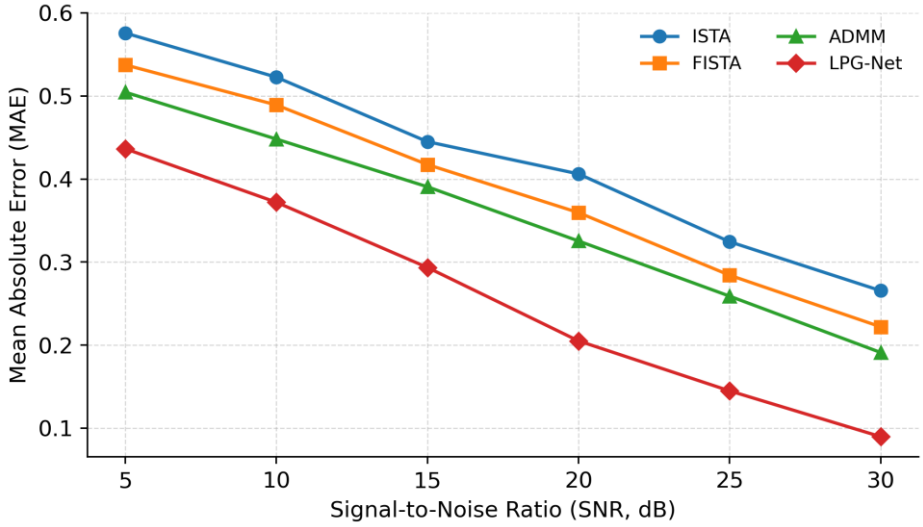


Figure 5: mae vs snr

### 4.5. Runtime and Deployment Efficiency

Practical efficiency is reported in Figure 6 using inference latency distributions. Classical optimization solvers require iterative updates until convergence or a predefined iteration limit, leading to higher latency and larger runtime variability. In contrast, LPG-Net executes a fixed-depth forward pass with predetermined $\boldsymbol{K}$ layers, which yields substantially lower latency and a tighter runtime distribution. This result is important for real-time or resource-constrained scenarios, where predictable inference time is often as critical as accuracy.

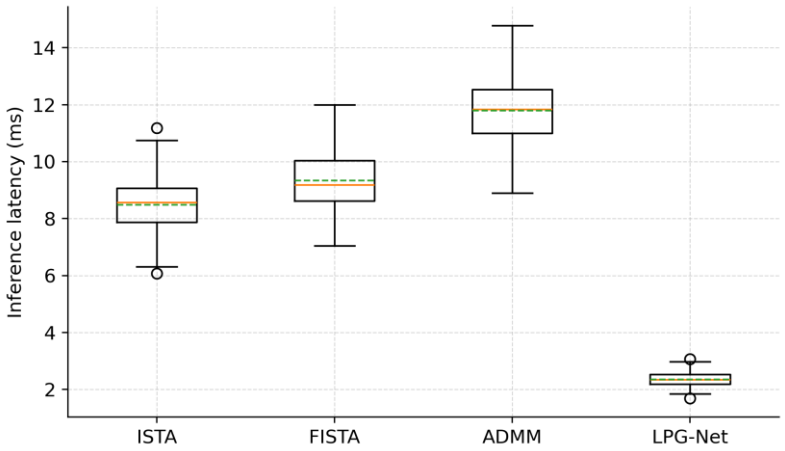


Figure 6: latency boxplot

## 4.6. Ablation Study

The ablation study in Figure 7 evaluates the contribution of two key components: layer-wise adaptive step size and the monotonicity-inspired regularization. Removing adaptive step sizes leads to the largest MAE increase, indicating that data-driven adjustment of optimization dynamics is a primary factor behind performance gains. Removing the monotonicity-inspired regularization also degrades MAE, suggesting that encouraging stable layer-to-layer descent helps prevent oscillations and improves finite-depth unrolling reliability. A variant that enforces shared thresholds across layers performs worse than the full model, implying that layer-specific shrinkage improves flexibility for handling different recovery stages (early coarse correction versus late fine refinement).

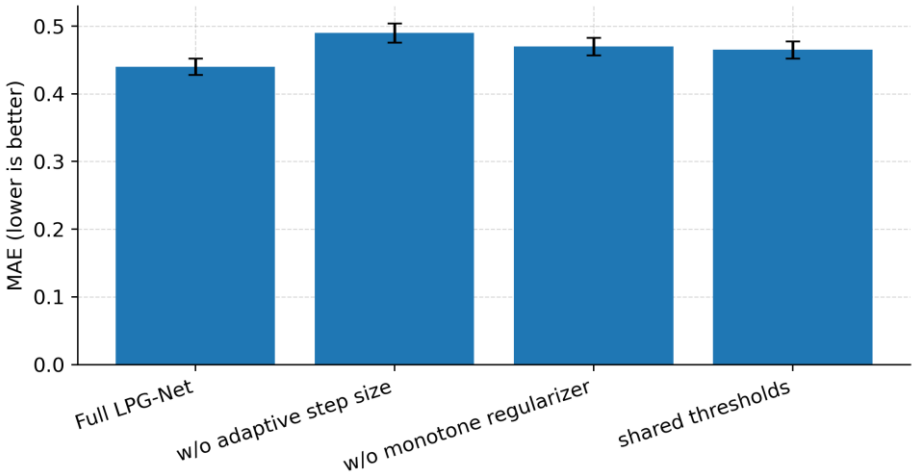


Figure 7: ablation mae

## 4.7. Uncertainty Calibration

Uncertainty calibration is shown in Figure 8, which compares empirical coverage against nominal coverage. The ideal behavior lies on the diagonal line, where empirical and nominal coverage match. The LPG-Net-based interval estimation remains closer to the ideal calibration line across different nominal levels, while baseline strategies show systematic under-coverage (empirical coverage lower than nominal), especially at higher nominal targets. This indicates that the proposed approach provides more reliable uncertainty quantification on the collected test data, improving the

trustworthiness of interval predictions in addition to point estimation accuracy.

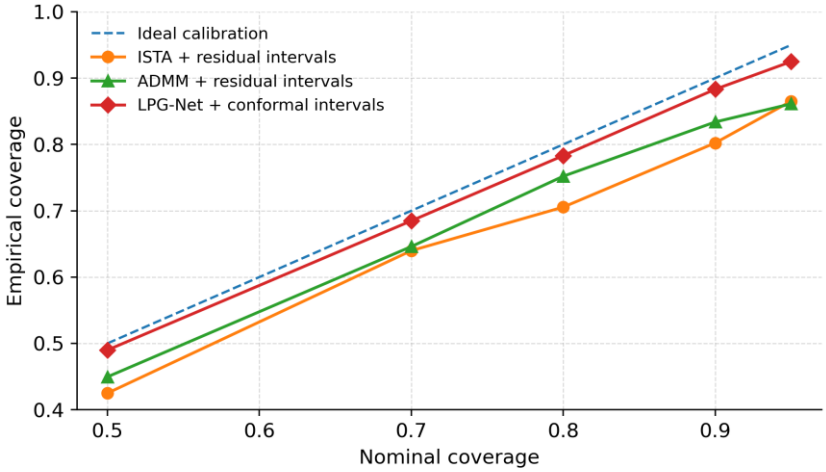


Figure 8: calibration curve

## 5. Conclusions

This study investigates a mathematics-driven machine learning framework for sparse learning by integrating proximal optimization principles with trainable network design. Starting from the Least Absolute Shrinkage and Selection Operator (LASSO) formulation, a Learnable Proximal Gradient Unrolling Network (LPG-Net) is constructed by mapping proximal gradient descent iterations into finite network layers, where step sizes and shrinkage thresholds are learned in a layer-wise manner. A monotonicity-inspired regularization term is further introduced to encourage stable layer-to-layer descent behavior, enhancing robustness in finite-depth inference.

Experimental results on collected sparse regression and denoising data demonstrate that LPG-Net achieves improved convergence efficiency, lower estimation error, and higher support recovery accuracy compared with classical optimization baselines. The fixed-depth forward inference also yields lower and more stable runtime, which is favorable for practical deployment. In addition, uncertainty calibration results indicate that the proposed approach can provide more reliable coverage behavior than baseline interval strategies, supporting more trustworthy predictions in noise-affected scenarios. Overall, the proposed framework offers an interpretable and efficient pathway for combining mathematical optimization structures with modern machine learning, and future work may extend this design to broader nonsmooth objectives, structured sparsity constraints, and more complex real-world sensing operators.

## References

*[1] Nir Shlezinger, Jay Whang, Yonina C. Eldar, Alexandros G. Dimakis, "Model-Based Deep Learning," Proceedings of the IEEE, vol. 111, no. 5, pp. 465–499, 2023, doi: 10.1109/JPROC.2023.3247480.*

*[2] Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, Rebecca Willett, "Deep Learning Techniques for Inverse Problems in Imaging," IEEE Journal on Selected Areas in Information Theory, vol. 1, no. 1, pp. 39–56, 2020, doi: 10.1109/JSAIT.2020.2991563.*

*[3] Vishal Monga, Yuelong Li, Yonina C. Eldar, "Algorithm Unrolling: Interpretable, Efficient Deep Learning for Signal and Image Processing," IEEE Signal Processing Magazine, vol. 38, no. 2, pp. 18–44, 2021, doi: 10.1109/MSP.2020.3016905.*

*[4] Ulugbek S. Kamilov, Charles A. Bouman, Gregery T. Buzzard, Brendt Wohlberg, "Plug-and-Play Methods for Integrating Physical and Learned Models in Computational Imaging: Theory, Algorithms, and Applications," IEEE Signal Processing Magazine, vol. 40, no. 1, pp. 85–97, 2023, doi: 10.1109/MSP.2022.3199595.*

*[5] Jean-Christophe Pesquet, Audrey Repetti, Matthieu Terris, Yves Wiaux, "Learning Maximally Monotone Operators for Image Recovery," SIAM Journal on Imaging Sciences, vol. 14, no. 3, pp. 1206–1237, 2021, doi: 10.1137/20M1387961.*

*[6] Sebastian Lunz, Andreas Hauptmann, Tanja Tarvainen, Carola-Bibiane Schönlieb, Simon Arridge, "On Learned Operator Correction in Inverse Problems," SIAM Journal on Imaging Sciences, vol. 14, no. 1, pp. 92–127, 2021, doi: 10.1137/20M1338460.*
*[7] Samuel Hurault, Arthur Leclaire, Nicolas Papadakis, "Proximal Denoiser for Convergent Plug-and-Play Optimization with Nonconvex Regularization," in Proceedings of the 39th International Conference on Machine Learning (ICML), PMLR, 2022. Available: https://proceedings.mlr.press/v162/hurault22a/hurault22a.pdf*
*[8] Samuel Hurault, Antonin Chambolle, Arthur Leclaire, Nicolas Papadakis, "Convergent Plug-and-Play with Proximal Denoiser and Unconstrained Regularization Parameter," Journal of Mathematical Imaging and Vision, vol. 66, pp. 616–638, 2024, doi: 10.1007/s10851-024-01195-w.*
*[9] Mikael Le Pendu, Christine Guillemot, "Preconditioned Plug-and-Play ADMM with Locally Adjustable Denoiser for Image Restoration," SIAM Journal on Imaging Sciences, vol. 16, no. 1, pp. 393–422, 2023, doi: 10.1137/22M1504809.*
*[10] Alessandro Benfenati, "Plug and Play Splitting Techniques for Poisson Image Restoration," Journal of Mathematical Imaging and Vision, 2025, doi: 10.1007/s10851-025-01273-7.*
*[11] Shaik Basheeruddin Shah, Pradyumna Pradhan, Wei Pu, Rohan Randhi, Miguel R. D. Rodrigues, Yonina C. Eldar, "Optimization Guarantees of Unfolded ISTA and ADMM Networks With Smooth Soft-Thresholding," IEEE Transactions on Signal Processing, vol. 72, pp. 3272–3286, 2024, doi: 10.1109/TSP.2024.3412981.*
*[12] Brent W. De Weerdt, Yonina C. Eldar, Nikolaos Deligiannis, "Deep Unfolding Transformers for Sparse Recovery of Video," IEEE Transactions on Signal Processing, vol. 72, pp. 1782–1796, 2024, doi: 10.1109/TSP.2024.3381749.*
*[13] Aniket Pramanik, M. Bridget Zimmerman, Mathews Jacob, "Memory-efficient model-based deep learning with convergence and robustness guarantees," IEEE Transactions on Computational Imaging, vol. 9, pp. 260–275, 2023, doi: 10.1109/TCI.2023.3252268.*
*[14] Jan Christian Hauffen, Linh Kästner, Samim Ahmadi, Peter Jung, Giuseppe Caire, Mathias Ziegler, "Learned Block Iterative Shrinkage Thresholding Algorithm for Photothermal Super Resolution Imaging," Sensors, vol. 22, no. 15, Art. no. 5533, 2022, doi: 10.3390/s22155533.*
*[15] Brett R. Levac, Marius Arvinte, Jonathan I. Tamir, "Federated End-to-End Unrolled Models for Magnetic Resonance Image Reconstruction," Bioengineering, vol. 10, no. 3, Art. no. 364, 2023, doi: 10.3390/bioengineering10030364.*
*[16] Pierre Ablin, Thomas Moreau, Mathurin Massias, Alexandre Gramfort, "Learning Step Sizes for Unfolded Sparse Coding," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, pp. 13100–13110, 2019. (No DOI; available from the official NeurIPS archive.) Available: https://papers.neurips.cc/paper/9469-learning-step-sizes-for-unfolded-sparse-coding*
*[17] Daisuke Ito, Satoshi Takabe, Tadashi Wadayama, "Trainable ISTA for Sparse Signal Recovery," IEEE Transactions on Signal Processing, vol. 67, no. 12, pp. 3113–3125, 2019, doi: 10.1109/TSP.2019.2912879.*
*[18] Mark Borgerding, Philip Schniter, Sundeep Rangan, "AMP-Inspired Deep Networks for Sparse Linear Inverse Problems," IEEE Transactions on Signal Processing, vol. 65, no. 16, pp. 4293–4308, 2017, doi: 10.1109/TSP.2017.2708040.*