Position: Adopt Constraints Over Penalties in Deep Learning

Item Type Preprint

Author Juan Ramirez

Author Meraj Hashemizadeh

Author Simon Lacoste-Julien

Abstract Recent efforts to develop trustworthy AI systems with accountability guarantees have led to widespread use of machine learning formulations incorporating external requirements, or constraints. These requirements are often enforced via penalization--adding fixed-weight terms to the task loss. We argue this approach is fundamentally ill-suited since there may be no penalty coefficient that simultaneously ensures constraint satisfaction and optimal constrained performance, i.e., that truly solves the constrained problem. Moreover, tuning these coefficients requires costly trial-and-error, incurring significant time and computational overhead. We, therefore, advocate for broader adoption of tailored constrained optimization methods--such as the Lagrangian approach, which jointly optimizes the penalization "coefficients" (the Lagrange multipliers) and the model parameters. Such methods (i) truly solve the constrained problem and do so accountably, by clearly defining feasibility and verifying when it is achieved, (ii) eliminate the need for extensive penalty tuning, and (iii) integrate seamlessly with modern deep learning pipelines.

Date 2025-07-28

Short Title Position

Library Catalog arXiv.org

URL http://arxiv.org/abs/2505.20628

Accessed 9/18/2025, 11:23:16 AM

Extra arXiv:2505.20628 [cs]

DOI 10.48550/arXiv.2505.20628

Repository arXiv

Archive ID arXiv:2505.20628

Date Added 9/18/2025, 11:23:16 AM

Modified 9/18/2025, 11:23:16 AM

Tags:

Computer Science - Machine Learning, Mathematics - Optimization and Control

Notes:

Comment: Code available at https://github.com/merajhashemi/constraints-vs-penalties

Attachments

- Full Text PDF
- Snapshot

Enforcing Hard Linear Constraints in Deep Learning Models with Decision Rules

Item Type	Preprint
Author	Gonzalo E. Constante-Flores
Author	Hao Chen
Author	Can Li
Abstract	Deep learning models are increasingly deployed in safety-critical tasks where predictions must satisfy hard constraints, such as physical laws, fairness requirements, or safety limits. However, standard architectures lack built-in mechanisms to enforce such constraints, and existing approaches based on regularization or projection are often limited to simple constraints, computationally expensive, or lack feasibility guarantees. This paper proposes a model-agnostic framework for enforcing input-dependent linear equality and inequality constraints on neural network outputs. The architecture combines a task network trained for prediction accuracy with a safe network trained using decision rules from the stochastic and robust optimization literature to ensure feasibility across the entire input space. The final prediction is a convex combination of the two subnetworks, guaranteeing constraint satisfaction during both training and inference without iterative procedures or runtime optimization. We prove that the architecture is a universal approximator of constrained functions and derive computationally tractable formulations based on linear decision rules. Empirical results on benchmark regression tasks show that our method consistently satisfies constraints while maintaining competitive accuracy and low inference latency.
Date	2025-05-20
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2505.13858
Accessed	9/19/2025, 10:26:52 AM
Extra	arXiv:2505.13858 [cs]
DOI	10.48550/arXiv.2505.13858
Repository	arXiv
Archive ID	arXiv:2505.13858
Date Added	9/19/2025, 10:26:52 AM
Modified	9/19/2025, 10:26:52 AM

Tags:

Computer Science - Machine Learning

Notes:

Comment: 1 figure

This paper is very unsatisfying. It essentially just adds the output of a safe network to the output of an optimized network. This moves in the direction of safety, but not necessarily in the direction of the solution. It also seems incredibly inefficient.

Attachments

- Full Text PDF
- Snapshot

On Surjectivity of Neural Networks: Can you elicit any behavior from your model?

Item Type	Preprint
Author	Haozhe Jiang
Author	Nika Haghtalab
Abstract	Given a trained neural network, can any specified output be generated by some input? Equivalently, does the network correspond to a function that is surjective? In generative models, surjectivity implies that any output, including harmful or undesirable content, can in principle be generated by the networks, raising concerns about model safety and jailbreak vulnerabilities. In this paper, we prove that many fundamental building blocks of modern neural architectures, such as networks with pre-layer normalization and linear-attention modules, are almost always surjective. As corollaries, widely used generative frameworks, including GPT-style transformers and diffusion models with deterministic ODE solvers, admit inverse mappings for arbitrary outputs. By studying surjectivity of these modern and commonly used neural architectures, we contribute a formalism that sheds light on their unavoidable vulnerability to a broad class of adversarial attacks.
Date	2025-08-26
Short Title	On Surjectivity of Neural Networks
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2508.19445
Accessed	10/7/2025, 4:17:13 PM
Extra	arXiv:2508.19445 [cs]
DOI	10.48550/arXiv.2508.19445
Repository	arXiv
Archive ID	arXiv:2508.19445
Date Added	10/7/2025, 4:17:13 PM
Modified	10/7/2025, 4:17:13 PM

Tags:

Computer Science - Machine Learning, Statistics - Machine Learning

Attachments

- Full Text PDF
- Snapshot

HardNet: Hard-Constrained Neural Networks with Universal Approximation Guarantees

Item Type	Preprint
Author	Youngjae Min
Author	Navid Azizan
Abstract	Incorporating prior knowledge or specifications of input-output relationships into machine learning models has attracted significant attention, as it enhances generalization from limited data and leads to conforming outputs. However, most existing approaches use soft constraints by penalizing violations through regularization, which offers no guarantee of constraint satisfaction, especially on inputs far from the training distribution -- an essential requirement in safety-critical applications. On the other hand, imposing hard constraints on neural networks may hinder their representational power, adversely affecting performance. To address this, we propose HardNet, a practical framework for constructing neural networks that inherently satisfy hard constraints without sacrificing model capacity. Unlike approaches that modify outputs only at inference time, HardNet enables end-to-end training with hard constraint guarantees, leading to improved performance. To the best of our knowledge, HardNet is the first method with an efficient forward pass to enforce more than one input-dependent inequality constraint. It allows unconstrained optimization of the network parameters using standard algorithms by appending a differentiable closed-form enforcement layer to the network's output. Furthermore, we show that HardNet is expressive and retains the universal approximation capabilities of neural networks. We demonstrate the versatility and effectiveness of HardNet across various applications: learning with piecewise constraints, learning optimization solvers with guaranteed feasibility, and optimizing control policies in safety-critical systems.
Date	2025-06-03
Short Title	HardNet
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2410.10807
Accessed	10/9/2025, 11:01:02 AM
Extra	arXiv:2410.10807 [cs]
DOI	10.48550/arXiv.2410.10807
Repository	arXiv
Archive ID	arXiv:2410.10807
Date Added	10/9/2025, 11:01:02 AM

Modified 10/9/2025, 11:01:02 AM

Tags:

Computer Science - Machine Learning, Statistics - Machine Learning, Computer Science - Artificial Intelligence

Attachments

- Preprint PDF
- Snapshot

OptiMind: Teaching LLMs to Think Like Optimization Experts

Item Type	Document
Author	Zeyi Chen
Author	Xinzhi Zhang
Author	Humishka Zope
Author	Hugo Barbalho
Author	Konstantina Mellou
Author	Marco Molinaro
Author	Janardhan (Jana) Kulkarni
Author	Ishai Menache
Author	Sirui Li
Abstract	Mathematical programming - the task of expressing operations and decision-making problems in precise mathematical language - is fundamental across domains, yet remains a skill-intensive process requiring operations research expertise. Recent advances in large language models for complex reasoning have spurred interest in automating this task, translating natural language into executable optimization models. Current approaches, however, achieve limited accuracy, hindered by scarce and noisy training data without leveraging domain knowledge. In this work, we systematically integrate optimization expertise to improve formulation accuracy for mixed-integer linear programming, a key family of mathematical programs. Our approach first cleans training data through class-based error analysis to explicitly prevent common mistakes within each optimization class. We then develop multi-turn inference strategies that guide LLMs with class-specific error summaries and solver feedback, enabling iterative refinement. Experiments across multiple base LLMs demonstrate that combining cleaned data with domain-informed prompting and feedback improves formulation accuracy by 14 percentage points on average, enabling further progress toward robust LLM-assisted optimization formulation.
Date	2025-09
URL	https://www.microsoft.com/en-us/research/publication/optimind-teaching-llms-to-think-like-optimization-experts/

Date Added 10/24/2025, 4:05:43 PM
Modified 10/24/2025, 4:05:43 PM

Distributionally Robust Constrained Reinforcement Learning under Strong Duality

Item Type Preprint
Author Zhengfei Zhang
Author Kishan Panaganti
Author Laixi Shi
Author Yanan Sui
Author Adam Wierman
Author Yisong Yue
Abstract We study the problem of Distributionally Robust Constrained RL (DRC-RL), where the goal is to maximize the expected reward subject to environmental distribution shifts and constraints. This setting captures situations where training and testing environments differ, and policies must satisfy constraints motivated by safety or limited budgets. Despite significant progress toward algorithm design for the separate problems of distributionally robust RL and constrained RL, there do not yet exist algorithms with end-to-end convergence guarantees for DRC-RL. We develop an algorithmic framework based on strong duality that enables the first efficient and provable solution in a class of environmental uncertainties. Further, our framework exposes an inherent structure of DRC-RL that arises from the combination of distributional robustness and constraints, which prevents a popular class of iterative methods from tractably solving DRC-RL, despite such frameworks being applicable for each of distributionally robust RL and constrained RL individually. Finally, we conduct experiments on a car racing benchmark to evaluate the effectiveness of the proposed algorithm.
Date 2024-06-22
Library Catalog arXiv.org
URL http://arxiv.org/abs/2406.15788
Accessed 12/15/2025, 3:48:30 PM
Extra arXiv:2406.15788 [cs]
DOI 10.48550/arXiv.2406.15788
Repository arXiv
Archive ID arXiv:2406.15788
Date Added 12/15/2025, 3:48:30 PM
Modified 12/15/2025, 3:48:30 PM

Tags:

Computer Science - Machine Learning

Notes:

Comment: Accepted at the Reinforcement Learning Conference (RLC) 2024; 28 pages, 4 figures

Attachments

- Preprint PDF
- Snapshot

Lagrangian Duality for Constrained Deep Learning

Item Type	Preprint
Author	Ferdinando Fioretto
Author	Pascal Van Hentenryck
Author	Terrence WK Mak
Author	Cuong Tran
Author	Federico Baldo
Author	Michele Lombardi
Abstract	This paper explores the potential of Lagrangian duality for learning applications that feature complex constraints. Such constraints arise in many science and engineering domains, where the task amounts to learning optimization problems which must be solved repeatedly and include hard physical and operational constraints. The paper also considers applications where the learning task must enforce constraints on the predictor itself, either because they are natural properties of the function to learn or because it is desirable from a societal standpoint to impose them. This paper demonstrates experimentally that Lagrangian duality brings significant benefits for these applications. In energy domains, the combination of Lagrangian duality and deep learning can be used to obtain state-of-the-art results to predict optimal power flows, in energy systems, and optimal compressor settings, in gas networks. In transprecision computing, Lagrangian duality can complement deep learning to impose monotonicity constraints on the predictor without sacrificing accuracy. Finally, Lagrangian duality can be used to enforce fairness constraints on a predictor and obtain state-of-the-art results when minimizing disparate treatments.
Date	2020-04-06
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2001.09394
Accessed	12/15/2025, 4:03:51 PM
Extra	arXiv:2001.09394 [cs]
DOI	10.48550/arXiv.2001.09394
Repository	arXiv
Archive ID	arXiv:2001.09394
Date Added	12/15/2025, 4:03:52 PM

Modified 12/15/2025, 4:03:52 PM

Tags:

Computer Science - Machine Learning, Statistics - Machine Learning

Attachments

- Full Text PDF
- Snapshot

OptiMUS-0.3: Using Large Language Models to Model and Solve Optimization Problems at Scale

Item Type Preprint
Author Ali AhmadiTeshnizi
Author Wenzhi Gao
Author Herman Brunborg
Author Shayan Talaei
Author Connor Lawless
Author Madeleine Udell
Abstract Optimization problems are pervasive in sectors from manufacturing and distribution to healthcare. However, most such problems are still solved heuristically by hand rather than optimally by state-of-the-art solvers because the expertise required to formulate and solve these problems limits the widespread adoption of optimization tools and techniques. We introduce a Large Language Model (LLM)-based system designed to formulate and solve (mixed integer) linear programming problems from their natural language descriptions. Our system is capable of developing mathematical models, writing and debugging solver code, evaluating the generated solutions, and improving efficiency and correctness of its model and code based on these evaluations. OptiMUS-0.3 utilizes a modular structure to process problems, allowing it to handle problems with long descriptions and complex data without long prompts. Experiments demonstrate that OptiMUS-0.3 outperforms existing state-of-the-art methods on easy datasets by more than 22% and on hard datasets (including a new dataset, NLP4LP, released with this paper that features long and complex problems) by more than 24%.
Date 2025-08-27
Short Title OptiMUS-0.3
Library Catalog arXiv.org
URL http://arxiv.org/abs/2407.19633
Accessed 12/15/2025, 4:12:22 PM
Extra arXiv:2407.19633 [cs]
DOI 10.48550/arXiv.2407.19633
Repository arXiv

Archive ID	arXiv:2407.19633
Date Added	12/15/2025, 4:12:22 PM
Modified	12/15/2025, 4:12:22 PM

Tags:

Computer Science - Artificial Intelligence

Notes:

Comment: This paper documents OptiMUS-0.3, improving on OptiMUS-0.1 (arXiv:2310.06116) and OptiMUS-0.2 (arXiv:2402.10172). arXiv admin note: text overlap with arXiv:2402.10172

Attachments

- Preprint PDF
- Snapshot

OptiChat: Bridging Optimization Models and Practitioners with Large Language Models

Item Type	Preprint
Author	Hao Chen
Author	Gonzalo Esteban Constante-Flores
Author	Krishna Sri Ipsit Mantri
Author	Sai Madhukiran Kompalli
Author	Akshdeep Singh Ahluwalia
Author	Can Li
Abstract	Optimization models have been applied to solve a wide variety of decision-making problems. These models are usually developed by optimization experts but are used by practitioners without optimization expertise in various application domains. As a result, practitioners often struggle to interact with and draw useful conclusions from optimization models independently. To fill this gap, we introduce OptiChat, a natural language dialogue system designed to help practitioners interpret model formulation, diagnose infeasibility, analyze sensitivity, retrieve information, evaluate modifications, and provide counterfactual explanations. By augmenting large language models (LLMs) with functional calls and code generation tailored for optimization models, we enable seamless interaction and minimize the risk of hallucinations in OptiChat. We develop a new dataset to evaluate OptiChat's performance in explaining optimization models. Experiments demonstrate that OptiChat effectively bridges the gap between optimization models and practitioners, delivering autonomous, accurate, and instant responses.
Date	2025-09-21

Short Title OptiChat
Library Catalog arXiv.org
URL http://arxiv.org/abs/2501.08406
Accessed 12/15/2025, 4:12:42 PM
Extra arXiv:2501.08406 [cs]
DOI 10.48550/arXiv.2501.08406
Repository arXiv
Archive ID arXiv:2501.08406
Date Added 12/15/2025, 4:12:42 PM
Modified 12/15/2025, 4:12:42 PM

Tags:

Computer Science - Computation and Language, Computer Science - Human-Computer Interaction, Computer Science - Machine Learning, Mathematics - Optimization and Control

Attachments

- Preprint PDF
- Snapshot

Large Language Models for Supply Chain Decisions

Item Type Preprint
Author David Simchi-Levi
Author Konstantina Mellou
Author Ishai Menache
Author Jeevan Pathuri
Abstract Supply Chain Management requires addressing a variety of complex decision-making challenges, from sourcing strategies to planning and execution. Over the last few decades, advances in computation and information technologies have enabled the transition from manual, intuition and experience-based decision-making, into more automated and data-driven decisions using a variety of tools that apply optimization techniques. These techniques use mathematical methods to improve decision-making. Unfortunately, business planners and executives still need to spend considerable time and effort to (i) understand and explain the recommendations coming out of these technologies; (ii) analyze various scenarios and answer what-if questions; and (iii) update the mathematical models used in these tools to reflect current business environments. Addressing these challenges requires involving data science teams and/or the technology providers to explain results or make the necessary changes in the technology and hence significantly slows down decision making. Motivated by the recent advances in Large Language Models (LLMs), we report how this disruptive technology can democratize supply chain technology - namely, facilitate the

understanding of tools' outcomes, as well as the interaction with supply chain tools without human-in-the-loop. Specifically, we report how we apply LLMs to address the three challenges described above, thus substantially reducing the time to decision from days and weeks to minutes and hours as well as dramatically increasing planners' and executives' productivity and impact.

Date	2025-07-29
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2507.21502
Accessed	12/15/2025, 4:13:33 PM
Extra	arXiv:2507.21502 [cs]
DOI	10.48550/arXiv.2507.21502
Repository	arXiv
Archive ID	arXiv:2507.21502
Date Added	12/15/2025, 4:13:33 PM
Modified	12/15/2025, 4:13:33 PM

Tags:

Computer Science - Artificial Intelligence

Notes:

Comment: Forthcoming chapter in AI in Supply Chains: Perspectives from Global Thought Leaders, edited by Maxime C. Cohen and Tinglong Dai, and part of the Springer Series in Supply Chain Management (edited by Prof. Chris Tang)

Attachments

- Full Text PDF
- Snapshot

Hilbert: Recursively Building Formal Proofs with Informal Reasoning

Item Type	Preprint
Author	Sumanth Varambally
Author	Thomas Voice
Author	Yanchao Sun
Author	Zhifeng Chen
Author	Rose Yu
Author	Ke Ye
Abstract	Large Language Models (LLMs) demonstrate impressive mathematical reasoning abilities, but their solutions frequently contain errors that cannot be automatically

verified. Formal theorem proving systems such as Lean 4 offer automated verification with complete accuracy, motivating recent efforts to build specialized prover LLMs that generate verifiable proofs in formal languages. However, a significant gap remains: current prover LLMs solve substantially fewer problems than general-purpose LLMs operating in natural language. We introduce Hilbert, an agentic framework that bridges this gap by combining the complementary strengths of informal reasoning and formal verification. Our system orchestrates four components: an informal LLM that excels at mathematical reasoning, a specialized prover LLM optimized for Lean 4 tactics, a formal verifier, and a semantic theorem retriever. Given a problem that the prover is unable to solve, Hilbert employs recursive decomposition to split the problem into subgoals that it solves with the prover or reasoner LLM. It leverages verifier feedback to refine incorrect proofs as necessary. Experimental results demonstrate that Hilbert substantially outperforms existing approaches on key benchmarks, achieving 99.2% on miniF2F, 6.6% points above the best publicly available method. Hilbert achieves the best known result on PutnamBench. It solves 462/660 problems (70.0%), outperforming proprietary approaches like SeedProver (50.4%) and achieving a 422% improvement over the best publicly available baseline. Thus, Hilbert effectively narrows the gap between informal reasoning and formal proof generation.

Date	2025-09-26
Short Title	Hilbert
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2509.22819
Accessed	12/15/2025, 4:15:21 PM
Extra	arXiv:2509.22819 [cs]
DOI	10.48550/arXiv.2509.22819
Repository	arXiv
Archive ID	arXiv:2509.22819
Date Added	12/15/2025, 4:15:21 PM
Modified	12/15/2025, 4:15:50 PM

Tags:

Computer Science - Machine Learning, Computer Science - Artificial Intelligence, Computer Science - Formal Languages and Automata Theory, verifier

Attachments

- Preprint PDF
- Snapshot

Persona Vectors: Monitoring and Controlling Character Traits in Language Models

Item Type	Preprint
Author	Runjin Chen
Author	Andy Arditi
Author	Henry Sleight
Author	Owain Evans
Author	Jack Lindsey
Abstract	Large language models interact with users through a simulated 'Assistant' persona. While the Assistant is typically trained to be helpful, harmless, and honest, it sometimes deviates from these ideals. In this paper, we identify directions in the model's activation space-persona vectors-underlying several traits, such as evil, sycophancy, and propensity to hallucinate. We confirm that these vectors can be used to monitor fluctuations in the Assistant's personality at deployment time. We then apply persona vectors to predict and control personality shifts that occur during training. We find that both intended and unintended personality changes after finetuning are strongly correlated with shifts along the relevant persona vectors. These shifts can be mitigated through post-hoc intervention, or avoided in the first place with a new preventative steering method. Moreover, persona vectors can be used to flag training data that will produce undesirable personality changes, both at the dataset level and the individual sample level. Our method for extracting persona vectors is automated and can be applied to any personality trait of interest, given only a natural-language description.
Date	2025-09-05
Short Title	Persona Vectors
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2507.21509
Accessed	12/15/2025, 4:19:00 PM
Extra	arXiv:2507.21509 [cs]
DOI	10.48550/arXiv.2507.21509
Repository	arXiv
Archive ID	arXiv:2507.21509
Date Added	12/15/2025, 4:19:00 PM
Modified	12/15/2025, 4:19:06 PM

Tags:

Computer Science - Machine Learning, Computer Science - Computation and Language

Attachments

- Preprint PDF
- Snapshot

AlphaEdit: Null-Space Constrained Knowledge Editing for Language Models

Item Type	Preprint
Author	Junfeng Fang
Author	Houcheng Jiang
Author	Kun Wang
Author	Yunshan Ma
Author	Shi Jie
Author	Xiang Wang
Author	Xiangnan He
Author	Tat-seng Chua
Abstract	Large language models (LLMs) often exhibit hallucinations due to incorrect or outdated knowledge. Hence, model editing methods have emerged to enable targeted knowledge updates. To achieve this, a prevailing paradigm is the locating-then-editing approach, which first locates influential parameters and then edits them by introducing a perturbation. While effective, current studies have demonstrated that this perturbation inevitably disrupt the originally preserved knowledge within LLMs, especially in sequential editing scenarios. To address this, we introduce AlphaEdit, a novel solution that projects perturbation onto the null space of the preserved knowledge before applying it to the parameters. We theoretically prove that this projection ensures the output of post-edited LLMs remains unchanged when queried about the preserved knowledge, thereby mitigating the issue of disruption. Extensive experiments on various LLMs, including LLaMA3, GPT2-XL, and GPT-J, show that AlphaEdit boosts the performance of most locating-then-editing methods by an average of 36.7% with a single line of additional code for projection solely. Our code is available at: https://github.com/jianghoucheng/AlphaEdit.
Date	2025-04-22
Short Title	AlphaEdit
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2410.02355
Accessed	12/15/2025, 4:19:45 PM
Extra	arXiv:2410.02355 [cs]
DOI	10.48550/arXiv.2410.02355
Repository	arXiv
Archive ID	arXiv:2410.02355
Date Added	12/15/2025, 4:19:45 PM
Modified	12/15/2025, 4:19:51 PM

Tags:

Computer Science - Artificial Intelligence, Computer Science - Computation and Language

Attachments

- Preprint PDF
- Snapshot

SWE-bench: Can Language Models Resolve Real-World GitHub Issues?

Item Type	Preprint
Author	Carlos E. Jimenez
Author	John Yang
Author	Alexander Wettig
Author	Shunyu Yao
Author	Kexin Pei
Author	Ofir Press
Author	Karthik Narasimhan
Abstract	Language models have outpaced our ability to evaluate them effectively, but for their future development it is essential to study the frontier of their capabilities. We find real-world software engineering to be a rich, sustainable, and challenging testbed for evaluating the next generation of language models. To this end, we introduce SWE-bench, an evaluation framework consisting of \$2,294\$ software engineering problems drawn from real GitHub issues and corresponding pull requests across \$12\$ popular Python repositories. Given a codebase along with a description of an issue to be resolved, a language model is tasked with editing the codebase to address the issue. Resolving issues in SWE-bench frequently requires understanding and coordinating changes across multiple functions, classes, and even files simultaneously, calling for models to interact with execution environments, process extremely long contexts and perform complex reasoning that goes far beyond traditional code generation tasks. Our evaluations show that both state-of-the-art proprietary models and our fine-tuned model SWE-Llama can resolve only the simplest issues. The best-performing model, Claude 2, is able to solve a mere \$1.96\$% of the issues. Advances on SWE-bench represent steps towards LMs that are more practical, intelligent, and autonomous.
Date	2024-11-11
Short Title	SWE-bench
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2310.06770
Accessed	12/15/2025, 4:21:25 PM
Extra	arXiv:2310.06770 [cs]
DOI	10.48550/arXiv.2310.06770
Repository	arXiv
Archive ID	arXiv:2310.06770
Date Added	12/15/2025, 4:21:25 PM
Modified	12/15/2025, 4:21:37 PM

Tags:

Computer Science - Artificial Intelligence, Computer Science - Computation and Language, Computer Science - Software Engineering

Notes:

Comment: Data, code, and leaderboard are available at https://www.swebench.com ICLR 2024, https://openreview.net/forum?id=VTF8yNQM66

Attachments

- Preprint PDF

GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks

Item Type	Preprint
Author	Tejal Patwardhan
Author	Rachel Dias
Author	Elizabeth Proehl
Author	Grace Kim
Author	Michele Wang
Author	Olivia Watkins
Author	Simón Posada Fishman
Author	Marwan Aljubeh
Author	Phoebe Thacker
Author	Laurance Fauconnet
Author	Natalie S. Kim
Author	Patrick Chao
Author	Samuel Miserendino
Author	Gildas Chabot
Author	David Li
Author	Michael Sharman
Author	Alexandra Barr
Author	Amelia Glaese
Author	Jerry Tworek
Abstract	We introduce GDPval, a benchmark evaluating AI model capabilities on real-world economically valuable tasks. GDPval covers the majority of U.S. Bureau of Labor Statistics Work Activities for 44 occupations across the top 9 sectors contributing to U.S. GDP (Gross Domestic Product). Tasks are constructed from the representative work of industry professionals with an average of 14 years of experience. We find that frontier model performance on GDPval is improving roughly linearly over time, and that the current best frontier models are approaching industry experts in deliverable quality. We analyze the potential for frontier models, when paired with human oversight, to perform GDPval tasks cheaper and faster than unaided experts. We also demonstrate that increased reasoning effort, increased task context, and increased scaffolding improves model performance on GDPval. Finally, we open-source a gold

subset of 220 tasks and provide a public automated grading service at evals.openai.com to facilitate future research in understanding real-world model capabilities.

Date 2025-10-05

Short Title GDPval

Library Catalog arXiv.org

URL http://arxiv.org/abs/2510.04374

Accessed 12/15/2025, 4:22:06 PM

Extra arXiv:2510.04374 [cs]

DOI 10.48550/arXiv.2510.04374

Repository arXiv

Archive ID arXiv:2510.04374

Date Added 12/15/2025, 4:22:06 PM

Modified 12/15/2025, 4:22:06 PM

Tags:

Computer Science - Machine Learning, Computer Science - Artificial Intelligence, Computer Science - Computers and Society

Attachments

- Preprint PDF

Democratizing Optimization with Generative AI

Item Type Journal Article

Author David Simchi-Levi

Author Tinglong Dai

Author Ishai Menache

Author Michelle Xiao Wu

Abstract Recent breakthroughs in generative artificial intelligence (GenAI) have captured public imagination and interest, while mathematical optimization remains largely underappreciated outside expert circles. In this article, we argue that GenAI can finally bridge the persistent gap between optimization's potent capabilities and its limited real-world uptake. We present the 4I framework—Insight, Interpretability, Interactivity, Improvisation—as a set of design principles for combining GenAI with mathematical optimization. Insight establishes a trusted, up-to-date view of the state; Interpretability explains model logic and trade-offs; Interactivity enables conversational what-if analysis; and Improvisation supports event-driven reoptimization. By making optimization tools more intuitive, explainable, and adaptable, we envision a future where frontline decision-makers are empowered to engage in rigorous decision-making. We discuss how GenAI complements, rather than replaces, optimization: GenAI lowers barriers to modeling and interpretation, while mathematical optimization reliably enforces business goals, rules, and hard constraints. We also address emerging concerns, from hallucinations to the risk of over-reliance, and outline research directions to ensure robust, ethical integration of GenAI and optimization. Ultimately, the GenAI boom gives the optimization community a historic opportunity to expand its impact, making decision-intelligence science more accessible and trustworthy to a wider audience while elevating human capabilities.

Language en
Library Catalog Zotero
Date Added 12/16/2025, 3:30:52 PM
Modified 12/16/2025, 3:31:04 PM

Attachments

- PDF

Guiding LLMs The Right Way: Fast, Non-Invasive Constrained Generation

Item Type Web Page
URL https://arxiv.org/html/2403.06988v1
Accessed 12/29/2025, 2:25:58 PM
Date Added 12/29/2025, 2:25:58 PM
Modified 12/29/2025, 2:26:09 PM

Attachments

- Guiding LLMs The Right Way: Fast, Non-Invasive Constrained Generation

---

HardNet: Hard-Constrained Neural Networks with Universal Approximation Guarantees

Item Type Preprint
Author Youngjae Min
Author Navid Azizan
Abstract Incorporating prior knowledge or specifications of input-output relationships into machine learning models has attracted significant attention, as it enhances generalization from limited data and yields conforming outputs. However, most existing approaches use soft constraints by penalizing violations through regularization, which offers no guarantee of constraint satisfaction, especially on inputs far from the training distribution--an essential requirement in safety-critical applications. On the other hand, imposing hard constraints on neural networks may hinder their representational power, adversely affecting performance. To address this,

we propose HardNet, a practical framework for constructing neural networks that inherently satisfy hard constraints without sacrificing model capacity. Unlike approaches that modify outputs only at inference time, HardNet enables end-to-end training with hard constraint guarantees, leading to improved performance. To the best of our knowledge, HardNet is the first method that enables efficient and differentiable enforcement of more than one input-dependent inequality constraint. It allows unconstrained optimization of the network parameters using standard algorithms by appending a differentiable closed-form enforcement layer to the network's output. Furthermore, we show that HardNet retains neural networks' universal approximation capabilities. We demonstrate its versatility and effectiveness across various applications: learning with piecewise constraints, learning optimization solvers with guaranteed feasibility, and optimizing control policies in safety-critical systems.

Date	2025-10-19
Short Title	HardNet
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2410.10807
Accessed	12/31/2025, 9:00:12 AM
Extra	arXiv:2410.10807 [cs]
DOI	10.48550/arXiv.2410.10807
Repository	arXiv
Archive ID	arXiv:2410.10807
Date Added	12/31/2025, 9:00:12 AM
Modified	12/31/2025, 9:00:22 AM

Tags:

Computer Science - Artificial Intelligence, Computer Science - Machine Learning, Statistics - Machine Learning

Attachments

- Full Text PDF
- Snapshot

Lagrangian Duality for Constrained Deep Learning

Item Type	Journal Article
Author	Ferdinando Fioretto
Author	Pascal Van Hentenryck
Author	Terrence W K Mak
Abstract	This paper explores the potential of Lagrangian duality for learning applications that feature complex constraints. Such constraints arise in many science and engineering domains, where the task amounts to learning optimization problems which must be

solved repeatedly and include hard physical and operational constraints. The paper also considers applications where the learning task must enforce constraints on the predictor itself, either because they are natural properties of the function to learn or because it is desirable from a societal standpoint to impose them.

Language	en
Library Catalog	Zotero
Date Added	12/31/2025, 9:01:21 AM
Modified	12/31/2025, 9:01:21 AM

Attachments

- PDF

Advancing LLM Safe Alignment with Safety Representation Ranking

Item Type	Conference Paper
Abstract	The rapid advancement of large language models (LLMs) has demonstrated milestone success in a variety of tasks, yet their potential for generating harmful content remains a significant safety concern. Existing safety guardrail approaches typically operate directly on textual responses, overlooking the rich information embedded in the model representations. In this paper, going beyond existing defenses that focus on a single safe response, we explore the potential of ranking hidden states across diverse responses to achieve safe generation. To this end, we propose Safety Representation Ranking (SRR), a listwise ranking framework that selects safe responses using hidden states from the LLM itself. SRR encodes both instructions and candidate completions using intermediate transformer representations and ranks candidates via a lightweight similarity-based scorer. Building on this framework, our approach directly leverages internal model states and supervision at the list level to capture subtle safety signals. Experiments across multiple benchmarks show that SRR significantly improves robustness to adversarial prompts, contributing a novel paradigm for LLM safety. Our code will be available upon publication.
Date	2025/10/08
Language	en
Library Catalog	openreview.net
URL	https://openreview.net/forum?id=A3DELRfrKO
Accessed	12/31/2025, 10:58:28 AM
Conference Name	The Fourteenth International Conference on Learning Representations
Date Added	12/31/2025, 10:58:28 AM
Modified	12/31/2025, 10:58:39 AM

Attachments

- Full Text PDF

Neal Parikh Department of Computer Science Stanford University

Item Type	Journal Article
Author	Stephen Boyd
Language	en
Library Catalog	Zotero
Date Added	12/31/2025, 11:50:25 AM
Modified	12/31/2025, 11:50:27 AM

Attachments

- PDF

web.stanford.edu/class/ee364a/lectures/chance_constr.pdf#page=4.00

Item Type	Attachment
URL	https://web.stanford.edu/class/ee364a/lectures/chance_constr.pdf#page=4.00
Accessed	12/31/2025, 12:12:36 PM
Date Added	12/31/2025, 12:12:36 PM
Modified	12/31/2025, 12:12:53 PM

Tags:

Chance constraint

Optimization Learning

Item Type	Preprint
Author	Pascal Van Hentenryck
Abstract	This article introduces the concept of optimization learning, a methodology to design optimization proxies that learn the input/output mapping of parametric optimization problems. These optimization proxies are trustworthy by design: they compute feasible solutions to the underlying optimization problems, provide quality guarantees on the returned solutions, and scale to large instances. Optimization proxies are differentiable programs that combine traditional deep learning technology with repair or completion layers to produce feasible solutions. The article shows that optimization proxies can be trained end-to-end in a self-supervised way. It presents methodologies to provide performance guarantees and to scale optimization proxies to large-scale optimization problems. The potential of optimization proxies is highlighted through applications in power systems and, in particular, real-time risk assessment and security-constrained optimal power flow.
Date	2025-01-07

Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2501.03443
Accessed	12/31/2025, 12:24:52 PM
Extra	arXiv:2501.03443 [math]
DOI	10.48550/arXiv.2501.03443
Repository	arXiv
Archive ID	arXiv:2501.03443
Date Added	12/31/2025, 12:24:52 PM
Modified	12/31/2025, 12:24:52 PM

Tags:

Computer Science - Artificial Intelligence, Mathematics - Optimization and Control

Attachments

- Full Text PDF
- Snapshot

Decoding the Configuration of AI Coding Agents: Insights from Claude Code Projects

Item Type	Preprint
Author	Helio Victor F. Santos
Author	Vitor Costa
Author	Joao Eduardo Montandon
Author	Marco Tulio Valente
Abstract	Agentic code assistants are a new generation of AI systems capable of performing end-to-end software engineering tasks. While these systems promise unprecedented productivity gains, their behavior and effectiveness depend heavily on configuration files that define architectural constraints, coding practices, and tool usage policies. However, little is known about the structure and content of these configuration artifacts. This paper presents an empirical study of the configuration ecosystem of Claude Code, one of the most widely used agentic coding systems. We collected and analyzed 328 configuration files from public Claude Code projects to identify (i) the software engineering concerns and practices they specify and (ii) how these concerns co-occur within individual files. The results highlight the importance of defining a wide range of concerns and practices in agent configuration files, with particular emphasis on specifying the architecture the agent should follow.
Date	2025-11-12
Short Title	Decoding the Configuration of AI Coding Agents
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2511.09268

Accessed	12/31/2025, 12:52:53 PM
Extra	arXiv:2511.09268 [cs]
DOI	10.48550/arXiv.2511.09268
Repository	arXiv
Archive ID	arXiv:2511.09268
Date Added	12/31/2025, 12:52:53 PM
Modified	12/31/2025, 12:52:56 PM

Tags:

Computer Science - Software Engineering

Attachments

- Preprint PDF
- Snapshot

Claude Code Best Practices

Item Type	Web Page
Abstract	A blog post covering tips and tricks that have proven effective for using Claude Code across various codebases, languages, and environments.
Language	en
URL	https://www.anthropic.com/engineering/claude-code-best-practices
Accessed	12/31/2025, 12:53:46 PM
Date Added	12/31/2025, 12:53:46 PM
Modified	12/31/2025, 12:53:46 PM

Projection-Based Constrained Policy Optimization

Item Type	Preprint
Author	Tsung-Yen Yang
Author	Justinian Rosca
Author	Karthik Narasimhan
Author	Peter J. Ramadge
Abstract	We consider the problem of learning control policies that optimize a reward function while satisfying constraints due to considerations of safety, fairness, or other costs. We propose a new algorithm, Projection-Based Constrained Policy Optimization (PCPO). This is an iterative method for optimizing policies in a two-step process: the first step performs a local reward improvement update, while the second step reconciles any constraint violation by projecting the policy back onto the constraint set. We

theoretically analyze PCPO and provide a lower bound on reward improvement, and an upper bound on constraint violation, for each policy update. We further characterize the convergence of PCPO based on two different metrics: $\normltwo$ norm and Kullback-Leibler divergence. Our empirical results over several control tasks demonstrate that PCPO achieves superior performance, averaging more than 3.5 times less constraint violation and around 15\% higher reward compared to state-of-the-art methods.

Date	2020-10-07
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2010.03152
Accessed	1/2/2026, 11:55:35 AM
Extra	arXiv:2010.03152 [cs]
DOI	10.48550/arXiv.2010.03152
Repository	arXiv
Archive ID	arXiv:2010.03152
Date Added	1/2/2026, 11:55:35 AM
Modified	1/2/2026, 11:55:41 AM

Tags:

Computer Science - Artificial Intelligence, Computer Science - Machine Learning, Computer Science - Robotics

Notes:

Comment: International Conference on Learning Representations (ICLR) 2020

Attachments

- Full Text PDF
- Snapshot

First Order Constrained Optimization in Policy Space

Item Type	Preprint
Author	Yiming Zhang
Author	Quan Vuong
Author	Keith W. Ross
Abstract	In reinforcement learning, an agent attempts to learn high-performing behaviors through interacting with the environment, such behaviors are often quantified in the form of a reward function. However some aspects of behavior-such as ones which are deemed unsafe and to be avoided-are best captured through constraints. We propose a

novel approach called First Order Constrained Optimization in Policy Space (FOCOPS) which maximizes an agent's overall reward while ensuring the agent satisfies a set of cost constraints. Using data generated from the current policy, FOCOPS first finds the optimal update policy by solving a constrained optimization problem in the nonparameterized policy space. FOCOPS then projects the update policy back into the parametric policy space. Our approach has an approximate upper bound for worst-case constraint violation throughout training and is first-order in nature therefore simple to implement. We provide empirical evidence that our simple approach achieves better performance on a set of constrained robotics locomotive tasks.

Date	2020-10-25
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2002.06506
Accessed	1/2/2026, 11:56:14 AM
Extra	arXiv:2002.06506 [cs]
DOI	10.48550/arXiv.2002.06506
Repository	arXiv
Archive ID	arXiv:2002.06506
Date Added	1/2/2026, 11:56:14 AM
Modified	1/2/2026, 11:56:14 AM

Tags:

Computer Science - Artificial Intelligence, Computer Science - Machine Learning, Statistics - Machine Learning

Attachments

- Preprint PDF
- Snapshot

OptNet: Differentiable Optimization as a Layer in Neural Networks

Item Type	Preprint
Author	Brandon Amos
Author	J. Zico Kolter
Abstract	This paper presents OptNet, a network architecture that integrates optimization problems (here, specifically in the form of quadratic programs) as individual layers in larger end-to-end trainable deep networks. These layers encode constraints and complex dependencies between the hidden states that traditional convolutional and fully-connected layers often cannot capture. We explore the foundations for such an architecture: we show how techniques from sensitivity analysis, bilevel optimization, and implicit differentiation can be used to exactly differentiate through these layers and

with respect to layer parameters; we develop a highly efficient solver for these layers that exploits fast GPU-based batch solves within a primal-dual interior point method, and which provides backpropagation gradients with virtually no additional cost on top of the solve; and we highlight the application of these approaches in several problems. In one notable example, the method is learns to play mini-Sudoku (4x4) given just input and output games, with no a-priori information about the rules of the game; this highlights the ability of OptNet to learn hard constraints better than other neural architectures.

Date	2021-12-02
Short Title	OptNet
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/1703.00443
Accessed	1/2/2026, 12:35:08 PM
Extra	arXiv:1703.00443 [cs]
DOI	10.48550/arXiv.1703.00443
Repository	arXiv
Archive ID	arXiv:1703.00443
Date Added	1/2/2026, 12:35:08 PM
Modified	1/2/2026, 12:35:12 PM

Tags:

Computer Science - Artificial Intelligence, Computer Science - Machine Learning, Mathematics - Optimization and Control, Statistics - Machine Learning

Notes:

Comment: ICML 2017

Attachments

- Full Text PDF
- Snapshot

Constrained Policy Optimization

Item Type	Preprint
Author	Joshua Achiam
Author	David Held
Author	Aviv Tamar
Author	Pieter Abbeel

Abstract For many applications of reinforcement learning it can be more convenient to specify both a reward function and constraints, rather than trying to design behavior through the reward function. For example, systems that physically interact with or around humans should satisfy safety constraints. Recent advances in policy search algorithms (Mnih et al., 2016, Schulman et al., 2015, Lillicrap et al., 2016, Levine et al., 2016) have enabled new capabilities in high-dimensional control, but do not consider the constrained setting. We propose Constrained Policy Optimization (CPO), the first general-purpose policy search algorithm for constrained reinforcement learning with guarantees for near-constraint satisfaction at each iteration. Our method allows us to train neural network policies for high-dimensional control while making guarantees about policy behavior all throughout training. Our guarantees are based on a new theoretical result, which is of independent interest: we prove a bound relating the expected returns of two policies to an average divergence between them. We demonstrate the effectiveness of our approach on simulated robot locomotion tasks where the agent must satisfy constraints motivated by safety.

Date	2017-05-30
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/1705.10528
Accessed	1/4/2026, 9:59:40 AM
Extra	arXiv:1705.10528 [cs]
DOI	10.48550/arXiv.1705.10528
Repository	arXiv
Archive ID	arXiv:1705.10528
Date Added	1/4/2026, 9:59:40 AM
Modified	1/4/2026, 9:59:46 AM

Tags:

Computer Science - Machine Learning

Notes:

Comment: Accepted to ICML 2017

Attachments

- Preprint PDF
- Snapshot