Novelty Assessment Report

Paper: Cautious Weight Decay
PDF URL: https://openreview.net/pdf?id=Gwe6gbGng5
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

We introduce Cautious Weight Decay (CWD), a one-line, optimizer-agnostic modification that applies weight decay only to parameter coordinates whose signs align with the optimizer update. Unlike standard decoupled decay, which implicitly optimizes a regularized or constrained objective, CWD preserves the original loss and admits a bilevel interpretation: it induces sliding-mode behavior upon reaching the stationary manifold, allowing it to search for locally Pareto-optimal stationary points of the unmodified objective. In practice, CWD is a drop-in change for optimizers such as AdamW, Lion, and Muon, requiring no new hyperparameters or additional tuning. For language model pre-training and ImageNet classification, CWD consistently improves final loss and accuracy at million- to billion-parameter scales.

Core Task Landscape

This paper addresses: **Selective Weight Decay Based on Parameter-Update Sign Alignment**

A total of **4 papers** were analyzed and organized into a taxonomy with **5 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Sign-Based Weight Decay Mechanisms**
- **Sign-Based Gradient Aggregation in Federated Learning**
- **Theoretical Analysis of Weight Decay in Optimization**

Complete Taxonomy Tree

- Selective Weight Decay Based on Parameter-Update Sign Alignment Survey Taxonomy
- Sign-Based Weight Decay Mechanisms
  - Selective Decay via Sign Alignment ★ (1 papers)
  - [0] Cautious Weight Decay (Anon et al., 2026) View paper
  - Sign-Based Transfer Learning Regularization (1 papers)
  - [3] Transfer learning with partial observability applied to cervical cancer screening (Kelwin Fernandes, 2017) View paper
- Sign-Based Gradient Aggregation in Federated Learning
  - Sign Election for Byzantine Robustness (1 papers)
  - [4] FedSECA: Sign Election and Coordinate-wise Aggregation of Gradients for Byzantine Tolerant Federated Learning (Joseph Geo Benjamin, 2024) View paper
  - Pruning-Aware Backdoor Defense (1 papers)
  - [2] Silencer: Pruning-aware backdoor defense for decentralized federated learning (T Huang, 2024) View paper
- Theoretical Analysis of Weight Decay in Optimization (1 papers)
  - [1] Investigating the Role of Weight Decay in Enhancing Nonconvex SGD (Tao Sun, 2025) View paper

Narrative

Core task: Selective weight decay based on parameter-update sign alignment. This emerging area explores how regularization can be made adaptive by examining the directional consistency between weight updates and decay forces. The taxonomy reveals three main branches that capture distinct facets of this idea. The first branch, Sign-Based Weight Decay Mechanisms, focuses on methods that modulate or selectively apply decay by checking whether parameter updates align or conflict with the decay direction. The second branch, Sign-Based Gradient Aggregation in Federated Learning, examines how sign information can guide communication and aggregation strategies when training is distributed across clients. The third branch, Theoretical Analysis of Weight Decay in Optimization, provides formal guarantees and convergence insights for weight decay variants in both convex and nonconvex settings. Together, these branches illustrate that sign alignment is relevant not only for single-machine training but also for distributed scenarios and for understanding the underlying optimization dynamics.

Within the first branch, a small handful of works have begun to explore selective or cautious decay strategies. Cautious Weight Decay[0] introduces a mechanism that applies regularization only when the sign of a parameter's gradient aligns with the sign of the parameter itself, aiming to prevent premature shrinkage of weights that are still being adjusted in conflicting directions. This approach contrasts with classical uniform decay and sits alongside other recent efforts such as Silencer[2], which targets specific subsets of parameters based on different criteria. Meanwhile, Weight Decay Nonconvex[1] offers theoretical perspectives on how decay interacts with nonconvex landscapes, providing a complementary lens on why selective strategies might improve generalization. By conditioning decay on sign alignment, Cautious Weight Decay[0] occupies a niche that bridges heuristic parameter selection and principled regularization, addressing scenarios where indiscriminate penalization may hinder learning.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

Both subtopics leverage parameter sign information as a signal for regularization, but apply it in fundamentally different contexts. The original leaf focuses on selective weight decay during standard training by checking alignment between parameter signs and optimizer updates. The sibling applies sign consistency as a regularization mechanism specifically for transfer learning scenarios, encouraging alignment between source and target model parameters.

Similarities: - Both use parameter sign information as a meaningful signal for regularization decisions - Both selectively apply regularization rather than uniform penalties across all parameters - Both aim to improve model training through sign-aware parameter control

Differences: - Original leaf applies decay conditionally during single-model optimization based on update-sign alignment; sibling enforces sign consistency across two models (source and target) in transfer learning - Original leaf's scope is standard training with optimizer-aware decay; sibling's scope is explicitly transfer learning regularization - Original leaf excludes transfer learning applications; sibling excludes direct weight decay methods, indicating complementary application domains

Suggested Search Directions: - Sign-based regularization methods in continual learning or domain adaptation - Hybrid approaches combining update-sign alignment with cross-model sign consistency - Theoretical analysis of why parameter sign information is effective for regularization across different learning paradigms

Sibling Subtopics

- **Sign-Based Transfer Learning Regularization** (leaves: 1, papers: 1)
- Scope: Regularization strategies encouraging sign consistency across source and target models for transfer learning.
- Exclude: Sign-based methods for federated aggregation or direct weight decay belong in their respective categories.

Contributions Analysis

Overall novelty summary. The paper proposes Cautious Weight Decay (CWD), which applies weight decay only when parameter signs align with optimizer updates. According to the taxonomy, this work resides in the 'Selective Decay via Sign Alignment' leaf under 'Sign-Based Weight Decay Mechanisms'. Notably, this leaf contains only the original paper itself—no sibling papers are listed. This positioning suggests the paper occupies a relatively sparse research direction within the broader field of sign-based regularization strategies, where most related work focuses on federated aggregation or transfer learning rather than direct sign-conditioned decay.

The taxonomy reveals three main branches: sign-based decay mechanisms, federated gradient aggregation, and theoretical weight decay analysis. The original paper's leaf sits within the first branch, which also includes a transfer learning regularization approach (paper 52a87076). Neighboring branches address Byzantine-robust aggregation and pruning-based defenses in federated settings, as well as convergence theory for standard weight decay. The scope notes clarify that CWD's sign-alignment conditioning distinguishes it from methods using sign information for aggregation or pruning, placing it in a distinct methodological niche focused on single-machine optimizer modifications.

Among the three contributions analyzed, the CWD algorithm itself examined two candidates with zero refutable prior work, while the bilevel interpretation examined three candidates with zero refutations. The Lyapunov-based convergence analysis examined ten candidates and found two that appear to provide overlapping theoretical frameworks. Given the limited search scope of fifteen total candidates, these statistics suggest the algorithmic and interpretive contributions may be more novel, whereas the convergence analysis builds on established techniques. The analysis does not claim exhaustive coverage but indicates that among top-ranked semantic matches, substantial algorithmic overlap is minimal.

Based on the limited literature search, CWD appears to introduce a relatively underexplored mechanism within sign-based regularization. The taxonomy structure and sibling-paper absence suggest this direction has received less attention than federated or transfer learning applications of sign information. However, the search examined only fifteen candidates, so broader prior work outside top semantic matches remains unassessed. The convergence analysis shows more connection to existing theory, while the core algorithm and bilevel framing appear more distinctive within the examined scope.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Cautious Weight Decay (CWD) algorithm

Description: The authors propose a simple modification to decoupled weight decay that selectively applies decay only when the optimizer update and parameter signs agree. This is implemented as a one-line change requiring no new hyperparameters and is compatible with optimizers such as AdamW, Lion, and Muon.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Sign-Entropy Regularization for Personalized Federated Learning

URL: View paper

Brief Assessment

Sign Entropy Federated[18] focuses on personalized federated learning with entropy-based regularization of gradient sign patterns across distributed clients, not on selective weight decay based on optimizer-parameter sign alignment in centralized training.

---

2. Adaptive adam-based optimizers using second-order weight decoupling and gradient-aware weight decay for vision transformer

URL: View paper

Brief Assessment

Adaptive Adam Decoupling[19] focuses on second-order weight decoupling and gradient-aware weight decay for Vision Transformers, which is a different mechanism than the sign-selective weight decay based on optimizer update and parameter alignment proposed in CWD.

Contribution 2: Bilevel interpretation and sliding-mode dynamics

Description: The authors establish that CWD optimizes the original objective without implicit regularization bias. They show it induces sliding-mode dynamics within the stationary manifold, converging to locally Pareto-optimal stationary points that minimize parameter magnitudes while remaining stationary.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Artificial Neural Networks as Surrogate Models in Multi-Objective Optimization for Chemical Reactor Design

URL: View paper

Brief Assessment

Neural Reactor Design[6] applies bilevel optimization and sliding-mode control to chemical reactor design problems, not to optimizer dynamics in deep learning. The technical domains and problem formulations are fundamentally different.

---

2. Pareto Optimal Design of a Fuzzy Adaptive Hierarchical Sliding-mode Controller for an X-Z Inverted Pendulum System

URL: View paper

Brief Assessment

Fuzzy Inverted Pendulum[5] applies sliding-mode control to a specific mechanical system (X-Z inverted pendulum) for stabilization and tracking, not bilevel optimization or Pareto-optimal stationary point convergence in general optimization contexts.

---

3. Optimal Operation of Power System Based on Artificial Intelligence Algorithm

URL: View paper

Brief Assessment

Power System AI[7] focuses on power grid optimization using genetic algorithms and fuzzy decision methods for distributed energy systems. It does not address bilevel optimization frameworks, sliding-mode dynamics, or Pareto-optimal stationary points in the context of optimizer weight decay mechanisms.

Contribution 3: Lyapunov-based convergence analysis

Description: The authors construct Lyapunov functions for several optimizers equipped with CWD and prove asymptotic stability and convergence to the stationary set of the original objective. They also provide a convergence rate for discrete-time Adam with CWD under additional assumptions.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. State space representation and phase analysis of gradient descent optimizers

URL: View paper

Brief Assessment

State Space Optimizers[12] focuses on gradient descent optimizers (SGD, SGD-M, NAG) using Lyapunov functions for stability analysis of control system models. The original paper analyzes modern adaptive optimizers (Adam, Lion-k) with cautious weight decay, a different algorithmic modification not present in the candidate.

---

2. Training dynamics aware neural network optimization with stabilization

URL: View paper

Brief Assessment

Stabilization Aware Training[13] applies Lyapunov analysis to neural network training dynamics as a switched linear system, focusing on stability of parameter updates across mini-batches. The ORIGINAL paper applies Lyapunov analysis to specific optimizers (SGD, Adam, Lion-k) with cautious weight decay to prove convergence to stationary points of the original objective. These are fundamentally different applications of Lyapunov theory—one for training stabilization via regularization, the other for proving convergence properties of a novel weight decay variant.

---

3. A lyapunov based adaptive and stable neural network weight regularization algorithm

URL: View paper

Brief Assessment

Lyapunov Weight Regularization[14] focuses on neural network weight regularization with adaptive learning rates for stability, not on optimizer convergence analysis with weight decay. The candidate addresses weight decay in neural networks through regularization signals, while the original paper analyzes convergence properties of modern optimizers (SGD, Adam, Lion-k) equipped with cautious weight decay.

---

4. Memory-efficient llm training with online subspace descent

URL: View paper

Prior Art Analysis

Online Subspace Descent[10] demonstrates prior work on Lyapunov-based convergence analysis for optimizers with weight decay. The candidate paper constructs Lyapunov functions (Hamiltonians) for multiple optimizers including SGD with momentum, Lion-k, and Adam, and proves convergence to stationary points. Specifically, the candidate shows that when applying online subspace descent to Hamiltonian+descent systems, the Hamiltonian+descent structure is preserved and yields a Lyapunov function guaranteeing convergence. The candidate also provides explicit Hamiltonian functions for these optimizers and proves that the derivative of the Hamiltonian is non-positive, establishing asymptotic stability. This work predates the original paper's claims of being the first to provide such analysis.

Evidence

Evidence 1 - **Rationale**: The candidate proves convergence to stationary points using LaSalle's invariance principle, which is a standard tool in Lyapunov stability analysis, demonstrating prior work on this type of convergence guarantee. - **Original**: we use lyapunov analysis to show that standard optimizers (sgd(m), lion-k, adam) with cautious weight decay are asymptotically stable and unbiased - **Candidate**: theorem 4.2. assume assumption 4.1 holds. let (wt, st, pt)t be a bounded solution of (7), then all the accumulation points {wt} as t → +∞ are stationary points of l(w). proof. by lasalle's invariance principle, the positive limit set of(wt, st, pt)t must be contained in i

---

5. Lyam: Robust non-convex optimization for stable learning in noisy environments

URL: View paper

Brief Assessment

Lyam[8] focuses on integrating Lyapunov stability theory with Adam optimizer for robust training in noisy environments, not on proving convergence guarantees for optimizers with cautious weight decay (CWD) as in the original paper.

---

6. A Hypersonic Target Trajectory Prediction Method Based on EGNN and Transformer

URL: View paper

Brief Assessment

Hypersonic Trajectory Prediction[15] applies Lyapunov stability theory to dynamically regulate learning rates in a whale optimization algorithm for trajectory prediction, not to analyze convergence of optimizers with weight decay. The technical focus and application domain are entirely different.

---

7. Lion secretly solves constrained optimization: As lyapunov predicts

URL: View paper

Prior Art Analysis

Lion Constrained[11] demonstrates that Lyapunov functions were previously constructed for Lion-k optimizers with weight decay, establishing asymptotic stability and convergence guarantees. The candidate paper presents a comprehensive Lyapunov analysis for Lion-k algorithms (which includes Lion, Muon, and other variants) with decoupled weight decay, proving convergence to stationary points and providing explicit convergence rates. This prior work shows that Lyapunov-based convergence analysis for optimizers with weight decay was already established before the original paper's submission.

Evidence

Evidence 1 - **Rationale**: This pair demonstrates that Lion Constrained[11] already presented Lyapunov functions for Lion-k family optimizers with weight decay (λ), establishing the theoretical framework that the original paper claims as novel. The candidate explicitly constructs Lyapunov functions for the same optimizer family with weight decay constraints. - **Original**: table 1: comparison of the continuous-time dynamics of different optimizers. sgdm represents sgd with momentum. lion-kincludes lion(k=∥·∥ 1) and muon(k=∥·∥ tr) as special cases. f:r d →ris assumed to be differentiable and lower bounded byf ⋆. optimizer continuous-time dynamics lyapunov function sgd ... - **Candidate**: the crest of this work is to show that, whenεγ ≤ 1, lion-k ode solves the following optimization: min x∈rd f(x) with f(x) := αf(x) + γ λk∗(λx), (4) where k∗(x) := supz(x⊤z -k(z)) is the conjugate function of k. because we may have k∗(x) = +∞ for some x, solving (4) requires to enforce a constraint o...

Evidence 2 - **Rationale**: This evidence pair shows that Lion Constrained[11] provided both continuous-time and discrete-time Lyapunov analysis for optimizers with weight decay (λ parameter), which directly refutes the novelty claim of constructing Lyapunov functions for optimizers with weight decay and proving convergence guarantees. - **Original**: we construct lyapunov functions for the continuous-time limits of several standard optimizers equipped with cautious weight decay. alyapunov functionis a lower bounded function with nonpositive derivative that is used to certify the stability of systems of differential equations. - **Candidate**: we now present a result on the discrete-time lion-k parallel to the continous-time results in theorem 3.1, but work for non-differentiable convex functionsk. we analyze a slight reform of (2): mt+1 = β2mt - (1 - β2)∇f(xt) ˜mt+1 = β1mt - (1 - β1)∇f(xt) xt+1 = xt + ϵ(∇k(˜mt+1) - λxt+1), (15) in which...

---

8. Implicit Bias of AdamW: Norm Constrained Optimization

URL: View paper

Brief Assessment

AdamW Implicit Bias[9] focuses on characterizing the implicit bias of AdamW as $\ell\infty$-norm constrained optimization through KKT point analysis, not on Lyapunov-based convergence analysis for optimizers with weight decay. The candidate's theoretical framework is fundamentally different from the original paper's Lyapunov function construction approach.

---

9. A Lyapunov Analysis of FISTA with Local Linear Convergence for Sparse Optimization

URL: View paper

Brief Assessment

FISTA Lyapunov[17] analyzes FISTA (a specific proximal gradient method) for sparse optimization using Lyapunov functions, while the original paper analyzes general optimizers (SGD, Adam, Lion-k) with cautious weight decay. The technical settings and algorithmic frameworks are fundamentally different.

---

10. Semi-Gradient SARSA Routing with Theoretical Guarantee on Traffic Stability and Weight Convergence

URL: View paper

Brief Assessment

SARSA Routing Stability[16] applies Lyapunov analysis to traffic routing problems with semi-gradient SARSA, not to general optimizer weight decay convergence. The technical domains and objectives differ fundamentally.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Cautious Weight Decay View paper
- [1] Investigating the Role of Weight Decay in Enhancing Nonconvex SGD View paper
- [2] Silencer: Pruning-aware backdoor defense for decentralized federated learning View paper
- [3] Transfer learning with partial observability applied to cervical cancer screening View paper
- [4] FedSECA: Sign Election and Coordinate-wise Aggregation of Gradients for Byzantine Tolerant Federated Learning View paper
- [5] Pareto Optimal Design of a Fuzzy Adaptive Hierarchical Sliding-mode Controller for an X-Z Inverted Pendulum System View paper
- [6] Artificial Neural Networks as Surrogate Models in Multi-Objective Optimization for Chemical Reactor Design View paper
- [7] Optimal Operation of Power System Based on Artificial Intelligence Algorithm View paper
- [8] Lyam: Robust non-convex optimization for stable learning in noisy environments View paper
- [9] Implicit Bias of AdamW: Norm Constrained Optimization View paper
- [10] Memory-efficient llm training with online subspace descent View paper
- [11] Lion secretly solves constrained optimization: As lyapunov predicts View paper
- [12] State space representation and phase analysis of gradient descent optimizers View paper
- [13] Training dynamics aware neural network optimization with stabilization View paper
- [14] A lyapunov based adaptive and stable neural network weight regularization algorithm View paper
- [15] A Hypersonic Target Trajectory Prediction Method Based on EGNN and Transformer View paper
- [16] Semi-Gradient SARSA Routing with Theoretical Guarantee on Traffic Stability and Weight Convergence View paper
- [17] A Lyapunov Analysis of FISTA with Local Linear Convergence for Sparse Optimization View paper
- [18] Sign-Entropy Regularization for Personalized Federated Learning View paper
- [19] Adaptive adam-based optimizers using second-order weight decoupling and gradient-aware weight decay for vision transformer View paper