Erciyes University Journal of Institute Of Science and Technology

Research Article
10.65520/erciyesfen.1957213

Imprint:
Volume: 42(3)
Year: 2026
Page: 523-549

Zafer DOĞAN[a*]

[a] Asst. Prof., Koç University, zdogan@ku.edu.tr

[*] Corresponding Author

Received: 5/22/2026
Accepted: 10/22/2026

Citation:
Zafer DOĞAN (2026). One-Sample Tilted Empirical Risk Minimization: Finite-Sample Tracking, Denominator Order, and Robustness. *Erciyes University Journal of Institute Of Science and Technology*, 42(3), 523-549. https://doi.org/10.65520/erciyesfen.*1957213*

Screened by iThenticate for Authors & Researchers

Except where otherwise noted, content in this article is licensed under a Creative Commons 4.0 International license. Icons by Font Awesome.

One-Sample Tilted Empirical Risk Minimization: Finite-Sample Tracking, Denominator Order, and Robustness

Abstract

Background/Aim: Classical tilted empirical risk minimization (TERM) assigns loss-dependent weights through a global normalizer. Direct evaluation on one observation removes relative weighting. We ask how closely a strict one-current-sample approximation tracks finite-sample TERM and how its finite-step behavior depends on denominator order, normalizer rate, and tilt.

Methods: A scalar moving-average normalizer carries cross-sample information between updates. We compared a previous-state denominator with a matched current-sample-first order on a toy problem using full-batch oracles, seven normalizer rates, and 20 shared streams. Scale-calibrated tilts were then evaluated with Polyak–Ruppert-averaged stochastic gradient descent (SGD) under four Breast Cancer Wisconsin Diagnostic (WDBC) label-noise conditions and seven Abalone response-contamination conditions.

Results: Both orders remained close to the toy finite-sample references. They were exactly identical at zero tilt; at nonzero tilt, their local low-rate discrepancy was approximately proportional to the product of learning and normalizer rates and changed direction with tilt. No order was uniformly superior. Negative tilt improved on empirical risk minimization (ERM) in specified corrupted conditions. Representative task-specific baselines were stronger in the most corrupted conditions, while the finite-sample TERM tracking findings were unchanged.
Conclusion: One persistent scalar can preserve tilt-dependent one-sample updates and closely track classical TERM in the tested finite-dimensional settings. The evidence is empirical and finite-horizon.

Keywords: Tilted Empirical Risk Minimization, Stochastic Approximation, Online Learning, Label Noise, Robust Regression

Tek Örnekli Eğilmiş Ampirik Risk Minimizasyonu: Sonlu Örneklem İzlemesi, Payda Sırası ve Sağlamlık

Öz

Arka Plan/Amaç: Klasik eğilmiş ampirik risk minimizasyonu (TERM), eğitim gözlemlerinin tümü üzerinden hesaplanan bir normalleştiriciyle kayba bağlı ağırlıklar atar. Tek gözlem üzerinde doğrudan değerlendirme bu göreli ağırlıklandırmayı ortadan kaldırır. Katı, tek-güncel-örnekli ve durum tutan bir yaklaşımın sonlu örneklem TERM'ini ne ölçüde izlediğini; sonlu-adım davranışının payda sırasına, normalleştirici hızına ve eğime nasıl bağlı olduğunu inceliyoruz.

Yöntemler: Skaler üstel hareketli-ortalama normalleştirici, stokastik güncellemeler arasında örnekler-arası bilgiyi taşır. Önceki-durum paydasını, tam-yığın referanslar, yedi normalleştirici hızı ve 20 ortak örnek akışıyla doğrusal-regresyon oyuncak probleminde eşleştirilmiş güncel-örnek-öncelikli sırayla karşılaştırdık. Ölçek-kalibreli negatif, sıfır ve pozitif eğimler daha sonra

dört Breast Cancer Wisconsin Diagnostic (WDBC) etiket-gürültüsü ve yedi Abalone yanıt-kirlenmesi koşulunda Polyak–Ruppert ortalamalı stokastik gradyan inişi (SGD) ile değerlendirildi.

Bulgular: İki sıra da oyuncak sonlu-örneklem referanslarına yakın kaldı. Sıfır eğimde tam olarak özdeşti; sıfır-dışı eğimde yerel düşük-hız farkı yaklaşık olarak öğrenme ve normalleştirici hızlarının çarpımıyla orantılıydı ve yönü eğime göre değişti. Hiçbir sıra her durumda üstün değildi. Negatif eğim, belirtilen kirlenmiş koşullarda ampirik risk minimizasyonunu (ERM) iyileştirdi. Göreve özgü temsili yöntemler en yüksek kirlenme koşullarında daha güçlüydü; sonlu-örneklem TERM izleme bulguları ise değişmedi.

Sonuç: Kalıcı tek bir skaler, eğime bağlı tek-örnekli güncellemeleri koruyabilir ve incelenen sonlu boyutlu ortamlarda klasik TERM'i yakından izleyebilir. Kanıtlar ampirik ve sonlu ufukludur.

Anahtar kelimeler: Eğilmiş Ampirik Risk Minimizasyonu, Stokastik Yaklaşım, Çevrim İçi Öğrenme, Etiket Gürültüsü, Sağlam Regresyon

1. INTRODUCTION

Large data sets and sequentially accessed observations motivate stochastic rules that update a model from one current example while retaining limited state. This accounting is routine for separable empirical risk minimization (ERM), because an individual sample loss provides a valid stochastic gradient of the empirical mean. It is less direct for objectives that couple observations through a global statistic. Stochastic compositional optimization addresses related compositions of expectations by recursively estimating inner functions, and finite-sum coupled variants develop single-loop methods for this broader structure [1], [2], [3]. The present study concerns a narrower operational question: how a globally normalized finite-sample tilted objective behaves when each model update may inspect only one current training example.

Tilted empirical risk minimization (TERM) aggregates sample losses through a log-sum-exp transformation [4]. Its tilt controls relative emphasis: positive tilt emphasizes larger losses, negative tilt emphasizes smaller losses, and the zero-tilt limit is empirical risk minimization (ERM). This family connects to entropic risk, tail-sensitive criteria, fairness objectives, and distributionally robust optimization (DRO), although these interpretations impose different assumptions and answer different statistical questions [5], [6]. Recent theory gives generalization and robustness guarantees for tilted empirical risk, including negative-tilt results under weak moment conditions and contaminated or shifted data [7]. Such statistical results motivate negative tilt, but they do not determine how a strict one-current-sample implementation should track a fixed finite-sample TERM objective.

The algorithmic obstruction is TERM's partition function. Its exact gradient is an exponentially weighted average in which every sample weight is normalized by all training losses. A current observation therefore does not determine its classical TERM weight in isolation. Indeed, evaluating the log-sum-exp objective on a batch of one returns the ordinary sample loss, so the tilt effect disappears. A scalar moving estimate of the normalizer can retain cross-sample information without storing losses or per-example weights. This resembles maintained inner-function estimates in compositional optimization, including recent work tailored to compositional entropic risk [8]. The resemblance does not make these algorithms identical: their assumptions, estimators, geometry, and resource models differ from the denominator recursion isolated here.

Stateful stochastic TERM is established prior work. Li et al. develop stochastic algorithms for tilted losses, analyze convergence under stated conditions, and include practical updates that maintain an objective or equivalent normalizer state [4]. Consequently, this paper does not claim the first stochastic, stateful, or one-sample TERM solver. The unresolved empirical issue is more specific. When the same observation supplies the exponential numerator and updates the running denominator, one may divide by the state available before that observation or by the state after incorporating it. Those orders agree at zero tilt but differ at finite step sizes away from zero. Their absolute tracking error,

CC: BY-NC-ND

paired finite-step discrepancy, and joint dependence on learning and normalizer rates have not been isolated here before under matched streams, tuning budgets, and state accounting.

A distinct online alternative removes the logarithm and global normalization, producing a directly computable exponential surrogate. Yıldırım et al. study this normalization-free objective for online outlier detection and robust regression [9]. That construction is state-free with respect to the classical partition function, but it optimizes a different objective: adding a common constant to every loss leaves normalized classical TERM weights unchanged while rescaling the surrogate gradient. Here the target remains the classical normalized finite-sample risk, and one scalar state approximates its partition function. The two approaches are therefore complementary rather than interchangeable.

Table 1 separates these direct methodological comparators from the task-specific robustness controls used later. Stateful stochastic TERM and compositional optimization are prior frameworks; the empirical focus here is strict one-current-sample tracking of a fixed finite-sample TERM target, including denominator order and three distinct tracking errors under matched resources.

Table 1. *Relationship among classical and stochastic tilted objectives. "One sample" describes the model update; full-data initialization and offline evaluation are separate resource choices.*

Method	Target	One sample	TERM scalar	Main role or distinction
Full-batch TERM	Exact finite-sample TERM	No	No	Exact normalized gradient uses all losses.
Li et al. stochastic TERM	Tilted loss/TERM formulations	Yes	Yes	Established stateful stochastic TERM; practical update incorporates the current sample before scaling.
Previous-state recursion	Finite-sample TERM tracking	Yes	Yes	Uses $\hat{Z}_{k-1}$ for the gradient, then updates the scalar state.
Current-sample-first	Same tracking target	Yes	Yes	Updates the scalar state before scaling; matched order comparator.
Exponential surrogate	Normalization-free exponential risk	Yes	No	Different, common-loss-shift-sensitive objective.
Compositional entropic risk	Compositional risk formulations	Varies	Inner estimate	Related coupled-state methods with different assumptions and state.

We compare two strict one-sample denominator orders. In the previous-state convention, the current exponential-weighted gradient is divided by the normalizer available at the start of the iteration; the model is updated; and the current exponential loss then enters the scalar state. In the current-sample-first convention, the same observation first refreshes the denominator and the refreshed state scales the model step. Both consume exactly one current sample, maintain one scalar TERM-specific state, use identical streams, and admit the same optimizers. A local reciprocal expansion predicts that their immediate stochastic gradient descent (SGD) difference is proportional to the product of learning and normalizer rates when the remaining factors are controlled. This is a finite-step heuristic, not a two-timescale convergence theorem; stochastic-approximation theory supplies the general coupled-state framework but not the empirical conclusion tested here [10], [11], [12].

Scale is a second practical issue. Exponential weighting depends on the dimensionless product of tilt and loss, so an identical raw tilt can produce different regimes for logistic loss, squared loss, or responses measured in different units. We map a dimensionless normalized tilt to a task-specific tilt using a fixed initialization-loss scale. This calibration is not asserted to be optimal. It makes signs and

CC: BY-NC-ND

magnitudes more comparable across tasks, reduces unit-driven saturation, and separates the scientific tilt choice from arbitrary loss units.

The robustness motivation is sign-specific but not unique to TERM. Classical M-estimation modifies the influence of residuals [13]; Catoni-type procedures robustify empirical means [14]; heavy-tailed ERM uses robust risk estimators [15]; conditional value-at-risk (CVaR) targets an upper loss tail [5]; and DRO protects against distributions in an ambiguity set [6]. Negative TERM instead changes normalized exponential weights within the observed finite sample. It neither applies a fixed robust loss nor explicitly rejects outliers, estimates a robust mean, or solves a worst-case distributional problem. If corrupted observations remain high-loss, negative tilt can reduce their influence, but clean hard examples can be reduced as well. We therefore test corruption-specific improvement and do not claim universal superiority to task-specific robust methods.

Noisy-label learning likewise includes specialized corrections and robust objectives. Unbiased loss correction can use known class-conditional transition rates [16]; transition-matrix methods estimate or apply forward and backward corrections [17]; normalized losses alter robustness properties [18]; and recent work studies both noise-ignorant ERM under identifiable signal conditions and conditional DRO for uncertain label posteriors [19], [20]. Negative TERM is corruption-agnostic loss reweighting: it uses neither a noise model nor clean-label side information. It is therefore a complementary generic control, not a replacement for methods designed around a known corruption mechanism.

The study asks two questions. The principal question is how accurately a strict one-current-sample stateful approximation tracks classical finite-sample TERM, and how its finite-step behavior depends on denominator order, normalizer rate, and tilt. A controlled linear-regression toy problem supplies high-accuracy full-batch oracles, seven normalizer rates, five tilts, matched tuning roles, and 20 shared sample streams. Objective, parameter, and normalizer errors establish absolute scale; paired order differences and a fine low-rate analysis test the local learning-rate–normalizer-rate heuristic. The secondary question is whether the same approximation preserves negative-tilt robustness. Breast Cancer Wisconsin Diagnostic (WDBC) [21] supplies four independent training/validation label-noise conditions evaluated on clean balanced test labels, and Abalone [22] supplies seven independent response-contamination conditions evaluated by clean-test regression error. Both use validation-only selection, scale-calibrated tilt, scheduled SGD, matched seeds, and Polyak–Ruppert averaging [23].

This paper makes three evidence-specific contributions. First, it quantifies strict one-sample tracking against full-batch TERM through simultaneous objective, parameter, and normalizer diagnostics. Second, it characterizes denominator order under matched resources: exact ERM identity, small differences at low normalizer rates, an approximately learning-rate–normalizer-rate proportional local regime, and a tilt-dependent direction with no uniformly preferred order. Third, it reports paired application controls across four WDBC label-noise and seven Abalone response-contamination conditions, including representative forward-correction and Huber baselines. These task-specific baselines are secondary to the tracking question. Scale calibration, augmented-state bookkeeping, optimizer compatibility, and positive tilt are supporting controls. The scope is finite-dimensional and empirical: the study establishes neither convergence nor a universal robustness law, and it does not claim an optimal scale estimator. Within that scope, it provides a reproducible finite-sample bridge between classical TERM and strict one-current-sample optimization.

2. MATERIALS AND METHODS

2.1. Preliminaries and problem setup

Let a generic observation be $Z = (X, Y) \sim P$, where P denotes the data-generating distribution, X is a feature vector, and Y is a response or class label. A model is indexed by a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$, where d is the parameter dimension. For a per-sample loss $\ell(\boldsymbol{\theta}; Z)$, the population risk is $R(\boldsymbol{\theta}) = \mathbb{E}_P[\ell(\boldsymbol{\theta}; Z)]$, with $\mathbb{E}_P$ denoting expectation under P. Population risk describes average loss under the data-generating distribution and is conceptually distinct from the empirical objectives evaluated on an

CC: BY-NC-ND

observed data set [15].

The analysis in this paper is finite-sample; population risk is introduced only to place ERM and TERM in context. Fix a training sample $\mathcal{D} = \{z_i\}_{i=1}^N$ of N observations. The index $i \in \{1, \dots, N\}$ identifies observation $z_i = (\mathbf{x}_i, y_i)$, where $\mathbf{x}_i$ is its feature vector and y_i is its response in regression or label in classification. The empirical loss of observation i is $\ell_i(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; z_i)$. ERM uses the uniformly averaged empirical objective

$$R_0(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}).$$

Thus, R_0 is determined by the fixed training sample rather than by an expectation over unobserved data. In the experiments, $\mathcal{D}$ is the training split after preprocessing and any training corruption specified by the protocol; validation and test observations are excluded from these empirical objectives. ERM averages observed sample losses uniformly. This is often appropriate, but it does not distinguish observations according to their current loss level. TERM modifies this aggregation through a continuous scalar tilt parameter. All objectives, gradients, normalizers, and full-batch oracles below refer to the same fixed training sample unless stated otherwise. The resulting notation separates the empirical optimization problem from subsequent validation and test evaluation.

2.2. Classical finite-sample TERM

Classical TERM acts on the finite collection of training losses $\{\ell_i(\boldsymbol{\theta})\}_{i=1}^N$. The finite-sample tilted empirical risk is denoted by $R_\tau(\boldsymbol{\theta})$, where the scalar τ is the tilt parameter: $\tau > 0$ emphasizes higher-loss observations, $\tau < 0$ emphasizes lower-loss observations, and the limit $\tau \to 0$ recovers the ERM objective R_0. For $\tau \neq 0$ [4],

$$R_\tau(\boldsymbol{\theta}) = \frac{1}{\tau}\log\left\{\frac{1}{N}\sum_{i=1}^{N} \exp\left[\tau \ell_i(\boldsymbol{\theta})\right]\right\}. \tag{1}$$

The scalar finite-sample exponential normalizer is

$$Z_\tau(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \exp\left[\tau \ell_i(\boldsymbol{\theta})\right].$$

$Z_\tau(\boldsymbol{\theta})$ depends on all N sample losses and provides the denominator required to form normalized exponential weights. Define

$$w_i(\boldsymbol{\theta}, \tau) = \frac{\exp[\tau \ell_i(\boldsymbol{\theta})]}{\sum_{j=1}^{N} \exp\left[\tau \ell_j(\boldsymbol{\theta})\right]}, \qquad \sum_{i=1}^{N} w_i(\boldsymbol{\theta}, \tau) = 1,$$

where j is a summation index over the same training observations. The exact finite-sample gradient is

$$\nabla R_\tau(\boldsymbol{\theta}) = \sum_{i=1}^{N} w_i(\boldsymbol{\theta}, \tau) \nabla \ell_i(\boldsymbol{\theta}),$$

where ∇ denotes differentiation with respect to $\boldsymbol{\theta}$. Each w_i is relative to all N training losses, so the exact gradient has global finite-sample dependence. Evaluating (1) on a single observation yields that observation's ordinary loss; hence the tilt effect disappears if no information from other samples is retained. To preserve cross-sample information while consuming only one current sample per update, we maintain a running scalar estimate of the exponential normalizer.

CC: BY-NC-ND

2.3. Stateful one-sample approximation

Initialize $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ and a scalar normalizer estimate $\hat{Z}_0 > 0$. The integer $k \in \{1, \dots, K\}$ is the stochastic-update index, where K is the total number of model updates, and i_k is the training index sampled uniformly with replacement at update k. The vectors $\boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_k \in \mathbb{R}^d$ are the model parameters before and after that update, respectively, and $\hat{\mathbf{g}}_k \in \mathbb{R}^d$ is its one-sample TERM gradient approximation. The exponential-loss factor a_k, the pre-update and post-update normalizer states $\hat{Z}_{k-1}$ and $\hat{Z}_k$, the learning rate η_k, the exponential moving average (EMA) rate ρ_k, and the positive numerical stabilizer ε_Z are scalars. With the loss and gradient evaluated at $\boldsymbol{\theta}_{k-1}$, the recursion is

$$a_k = \exp[\tau \ell_{i_k}(\boldsymbol{\theta}_{k-1})], \tag{2}$$

$$\widehat{\boldsymbol{g}}_k = \frac{a_k}{\hat{Z}_{k-1} + \varepsilon_Z} \nabla \ell_{i_k}(\boldsymbol{\theta}_{k-1}), \tag{3}$$

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \eta_k \widehat{\boldsymbol{g}}_k, \tag{4}$$

$$\hat{Z}_k = (1 - \rho_k)\hat{Z}_{k-1} + \rho_k a_k. \tag{5}$$

Equation (2) maps the current sample loss to its scalar exponential-loss factor. Equation (3) scales the current sample gradient by that factor and the previous running normalizer to approximate the normalized TERM gradient. Equation (4) uses η_k to control the model step, while (5) uses ρ_k to control how quickly the scalar normalizer incorporates the current sample. Thus the previous normalizer forms the current gradient, the model is updated, and only then is the scalar state refreshed. Using the pre-update normalizer prevents the current observation from entering both the numerator and the newly updated denominator of the same gradient estimate; it is not an unbiasedness correction.

Algorithm 1. Previous-state strict one-sample TERM with SGD

1. initialize $\boldsymbol{\theta}_0$ and $\hat{Z}_0 > 0$.
2. **for** $k = 1, \dots, K$, **do**
3. draw one i_k uniformly with replacement.
4. Compute a_k and $\widehat{\boldsymbol{g}}_k$ from (2) and (3).
5. Update $\boldsymbol{\theta}_k$ by (4).
6. Update $\hat{Z}_k$ by (5).
7. **end for**

Strict one-sample refers to model updates after initialization. The loss scale described below is computed once from the training split before updates begin. No full-data loss or gradient is computed during an online model update; full-data quantities are used only for offline evaluation and oracle construction. Numerically, normalized factors and the EMA are evaluated in the log domain, classical objectives use shifted log-sum-exp, and $\tau = 0$ has an exact ERM branch.

For SGD, $(\boldsymbol{\theta}_k, \hat{Z}_k)$ is the finite-dimensional stochastic state. Under independent sampling, standard state augmentation gives a time-inhomogeneous Markov description when schedules vary; coupled and two-timescale recursions are standard stochastic-approximation structures [10], [11], [12]. This is bookkeeping for the coupled recursion, not a convergence result.

2.4. Current-sample-first comparator

Once the normalizer is stateful, the current sample can be incorporated before or after its own model step. The matched comparator uses the same i_k, a_k, initialization, schedules, and sample gradient as the previous-state protocol, but first forms $\hat{Z}_k^+$, the scalar normalizer after incorporating the current sample and before the model update:

CC: BY-NC-ND

$$\hat{Z}_k^+ = (1-\rho_k)\hat{Z}_{k-1} + \rho_k a_k = \hat{Z}_{k-1} + \rho_k(a_k - \hat{Z}_{k-1}), \tag{6}$$

then uses $a_k/(\hat{Z}_k^+ + \varepsilon_Z)$ in the model update and carries $\hat{Z}_k = \hat{Z}_k^+$ forward. The update in (6) is the closest practical current-sample-first convention in Li et al. [4]. The protocols differ only in whether the current sample contributes to the denominator used for its own gradient.

Ignoring ε_Z for a local calculation, let $w_k^{\text{previous}} = a_k/\hat{Z}_{k-1}$ and $w_k^{\text{current}} = a_k/\hat{Z}_k^+$. These are the scalar sample weights used by the previous-state and current-sample-first conventions, respectively, and both multiply the same current-sample gradient $\nabla\ell_{i_k}(\boldsymbol{\theta}_{k-1})$. A first-order reciprocal expansion gives

$$w_k^{\text{current}} - w_k^{\text{previous}} = -\rho_k \frac{a_k(a_k - \hat{Z}_{k-1})}{\hat{Z}_{k-1}^2} + O(\rho_k^2). \tag{7}$$

Here w_k^{previous} and w_k^{current} are scalar multipliers, and both act on the same gradient $\nabla\ell_{i_k}(\boldsymbol{\theta}_{k-1})$. The remainder $O(\rho_k^2)$ collects terms whose magnitude is of second or higher order as ρ_k approaches zero. Thus (7) is local in ρ_k. With an SGD step η_k, the immediate model-update difference is locally of order $\eta_k\rho_k$ when the remaining factors are controlled. This is a finite-step heuristic, not a theorem, convergence rate, or asymptotic-equivalence statement.

2.5. Scale calibration

Because τ appears through the product $\tau\ell$, the loss units affect the effective exponential weighting. Scale calibration is therefore a numerical comparability device, not a theoretically optimal estimator. Let $\boldsymbol{\theta}_0$ be the model initialization, let $\mathcal{L}_0 = \{\ell_i(\boldsymbol{\theta}_0)\}_{i=1}^N$ collect all training losses at that initialization, and let $Q_p(\mathcal{L}_0)$ be the empirical p-quantile of $\mathcal{L}_0$. Define

$$s_{\text{IQR}} = Q_{0.95}(\mathcal{L}_0) - Q_{0.05}(\mathcal{L}_0), \qquad \tau = \frac{\bar{\tau}}{\max(s_{\text{IQR}}, \varepsilon_s)}. \tag{8}$$

Here s_{IQR} is the difference between the empirical 0.95 and 0.05 quantiles, $\bar{\tau}$ is the dimensionless normalized tilt, τ is the task-specific physical tilt used in the exponential, and $\varepsilon_s > 0$ is a positive floor preventing division by zero. For zero-initialized logistic regression, $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}_0 = \mathbf{0}$ and hence $l_i(\boldsymbol{\theta}_0) = \log[1 + \exp(0)] = \log 2$ for every observation, so the dispersion can vanish. The classification implementation therefore uses

$$s_{\text{class}} = \max\{s_{\text{IQR}}, Q_{0.50}(\mathcal{L}_0), \varepsilon_s\}$$

in the denominator of (8), where s_{class} is the classification-specific fallback scale. This fallback is needed because zero initialization can make every logistic loss equal and hence make $s_{\text{IQR}} = 0$. The scale is a fixed, reproducible default rather than a tuned optimum.

2.6. Experimental design

For regression, $\boldsymbol{x}_i \in \mathbb{R}^d$ is the p-dimensional raw feature vector and $y_i \in \mathbb{R}$ is the observed response. When an intercept is fitted, $\boldsymbol{x}_i$ is augmented by a leading one in the expression $\boldsymbol{x}_i^T\boldsymbol{\theta}$, so d = p + 1; otherwise d = p. The empirical half-squared loss is $\ell_i(\boldsymbol{\theta}) = \frac{1}{2}\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\theta}\right)^2$. For classification, $\boldsymbol{x}_i$ is the feature vector, $y_i \in \{-1,+1\}$ is the binary class label, $\boldsymbol{x}_i^T\boldsymbol{\theta}$ is the linear decision score, and the empirical logistic loss is $\ell_i(\boldsymbol{\theta}) = \log[1 + \exp(-y_i\boldsymbol{x}_i^T\boldsymbol{\theta})]$. SGD is the main optimizer. Adam [24] and high-order Adam (HAdam), the fourth-moment ablation, are retained only as numerical controls in Appendix E. Polyak–Ruppert averaging begins at the learning-rate switch in the application experiments [23]. The application study tests only normalized tilts $\bar{\tau} \in \{-0.5,0,0.5\}$, so its conclusions concern this tested magnitude rather than the full tilt continuum.

CC: BY-NC-ND

Regression performance is reported by root mean squared error (RMSE) and mean absolute error (MAE). For classification, sensitivity (recall) is the true-positive rate, specificity is the true-negative rate, precision is the fraction of predicted positives that are correct, and balanced accuracy is the average of sensitivity and specificity. We also report the F1 score, the harmonic mean of precision and recall; area under the receiver operating characteristic curve (AUROC); and area under the precision–recall curve (PR-AUC).

Table 2. *Primary experimental protocols. All reported aggregate results use 20 seeds.*

Study	Conditions	Horizon and state	Selection and primary metrics
Toy	Raw $\tau = -1, -0.5, 0, 0.5, 1$; seven ρ values from 0.001 to 0.1	50,000 selected updates; one scalar EMA; zero initialization	10,000-update validation tuning; exact objective gap, parameter error, normalizer error
WDBC	Label noise $r =$ 0,0.1,0.2,0.3; $\bar{\tau} = -0.5, 0, 0.5$	80 epochs; $\rho = 0.01$; two-stage SGD; averaged estimate	20-epoch validation tuning; clean balanced-test balanced accuracy and F1
Abalone	Contamination $c =$ 0,0.05, … ,0.30; $\bar{\tau} = -0.5, 0, 0.5$	300 epochs; $\rho = 0.01$; two-stage SGD; averaged estimate	50-epoch validation tuning; clean-test RMSE and MAE

The toy experiment isolates finite-sample tracking and denominator-order effects. Its data contain 290 inliers and 10 outliers with parameter order [slope,intercept]. A deterministic stratified 80/20 training/validation split is applied to the seed-0 data. The two denominator orders share every sampled index. High-accuracy full-batch solutions define the offline oracle; the full-batch optimizer is not treated as a stochastic trajectory.

The WDBC experiment tests negative-tilt robustness under binary label corruption. WDBC contains 569 observations and 30 predictors [21]. The symbol $r \in [0,1]$ denotes the label-noise rate. Stratified train/validation/test splitting precedes corruption, standardization is fitted on training features only, and test data are balanced to 43 malignant and 43 benign observations. Training and validation labels are independently flipped without replacement; test labels remain clean. Validation balanced accuracy minus failure fraction selects the schedule. For its two-stage SGD schedule, η_0 is the initial learning rate, and the switch is reported as the equivalent training epoch.

The Abalone experiment tests negative-tilt robustness under continuous-response contamination. Abalone contains 4,177 observations [22]. Each seed gives 3,341 training, 418 validation, and 418 test observations. The symbol $c \in [0,1]$ denotes the contamination fraction and n denotes the size of the split being corrupted. Training-fitted standardization and an intercept yield nine parameters. Independently within each split, $\lfloor cn \rfloor$ responses are replaced by draws from $\mathcal{N}(5, 5^2)$, where $\lfloor \cdot \rfloor$ is the floor operator and $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2. Original clean test responses are retained. Validation selection minimizes contaminated-validation RMSE plus 10^4 times the failure fraction. WDBC and Abalone corruption sets are matched by seed across tilt but independently generated and nonnested across levels.

Reported uncertainty is descriptive. Figures show mean $\pm$ one standard deviation for absolute metrics. Paired negative-tilt contrasts use 10,000 seed-pair bootstrap resamples with seed 161803 and are reported as percentile intervals. Appendix A gives complete grids, seeds, failure rules, metrics, and numerical details.

2.7. Representative robustness comparators

The robustness baselines are task-specific secondary controls rather than direct substitutes for the stochastic TERM recursion. For Abalone, Huber regression uses residual $u = \mathbf{x}^\top \boldsymbol{\theta} - y$ and the loss [13]

CC: BY-NC-ND

$$h_\delta(u) = \begin{cases} 1/2\, u^2, & |u| \le \delta, \\ \delta(|u| - 1/2\, \delta), & |u| > \delta. \end{cases}$$

It uses the same linear model, split and corruption generator, strict one-sample SGD updates, 300-epoch selected horizon, and Polyak–Ruppert convention as the TERM experiment. Six candidates cross $\delta \in \{1,2,4\}$ with $\eta_0 \in \{0.001,0.003\}$; all switch at epoch 10 and decay with exponent 1. Contaminated-validation RMSE plus the same failure penalty selects the candidate. Huber receives no contamination ratio.

For WDBC, forward loss correction is supplied the declared symmetric label-flip rate r as oracle-known information [16], [17]. Let $p_i = \sigma(\mathbf{x}_i^\top \boldsymbol{\theta})$ be the clean positive-class probability and $p_{y_i} = \sigma(y_i \mathbf{x}_i^\top \boldsymbol{\theta})$ the clean probability assigned to the observed label. The corresponding corrupted-label probability is

$$q_{y_i} = r + (1 - 2r) p_{y_i},$$

and the corrected sample loss is $-\log q_{y_i}$. At $r = 0$, this loss and its gradient reduce exactly to ordinary logistic cross-entropy. The comparator uses the same logistic model, split and corruption generator, one-sample SGD updates, 80-epoch selected horizon, Polyak–Ruppert convention, and learning-rate grid {0.001,0.003,0.01,0.03}. Corrupted-validation balanced accuracy minus failure fraction selects the rate; test data are not used. Because the true declared transition rate is supplied, this is an information-advantaged comparator.

The cited Catoni study concerns robust mean and variance estimation and does not uniquely determine a supervised linear-regression loss and tuning protocol for this experiment [14]; an ad hoc "Catoni-like" baseline was therefore not introduced.

2.8. Tracking quantities and evaluation timing

For each toy tilt, let $\boldsymbol{\theta}_\tau^\star$ denote the high-accuracy full-batch finite-sample solution. The signed objective gap is $R_\tau(\boldsymbol{\theta}_k) - R_\tau(\boldsymbol{\theta}_\tau^\star)$; it is not replaced by its absolute value. Parameter error is the absolute, unnormalized distance $\| \boldsymbol{\theta}_k - \boldsymbol{\theta}_\tau^\star \|_2$. The relative normalizer error is $\left|\hat{Z}_k / Z_\tau(\boldsymbol{\theta}_k) - 1\right|$, equivalently $\left|\hat{Z}_k - Z_\tau(\boldsymbol{\theta}_k)\right| / Z_\tau(\boldsymbol{\theta}_k)$. Thus the exact reference normalizer is evaluated at the same current parameter as the running estimate.

Paired denominator-order contrasts are current-sample-first minus previous-state, matched by seed, stream, tilt, normalizer rate, and analysis role. Positive objective- or parameter-error differences therefore favor previous-state. For the local rate diagnostic, $x_j = \eta \rho_j$ and y_j is the mean signed endpoint contrast at $\rho_j \le 0.01$ for a fixed nonzero tilt. The through-origin coefficient is $\sum_j x_j\, y_j / \sum_j x_j^2$, with uncentered $R^2 = 1 - \sum_j \left(y_j - \hat{\beta} x_j\right)^2 / \sum_j y_j^2$. The log–log slope is an ordinary least-squares slope with intercept for $\log\left|y_j\right|$ against $\log x_j$; only this slope discards the sign.

Toy trajectory measurements are post-update at iteration 1, every 250 updates, and iteration 50,000. Endpoint quantities use the current $\boldsymbol{\theta}_{50000}$ and $\hat{Z}_{50000}$ and do not use parameter averaging. Polyak–Ruppert evaluation is used only for the WDBC and Abalone application endpoints. Absolute summaries are means and sample standard deviations over 20 streams; paired intervals resample the 20 seedwise differences. Table 3 consolidates the definitions and timing.

CC: BY-NC-ND

Table 3. *Tracking quantities, references, and evaluation timing. Absolute summaries use means ± sample standard deviations; paired intervals resample matched seed differences.*

Quantity	Definition	Reference	Evaluation and aggregation
Objective gap	$R_\tau(\boldsymbol{\theta}_k) - R_\tau(\boldsymbol{\theta}_\tau^\star)$	Full-batch finite-sample optimum	Current trajectory/endpoint iterate; mean $\pm$ SD over 20 seeds
Parameter error	$\| \boldsymbol{\theta}_k - \boldsymbol{\theta}_\tau^\star \|_2$	Full-batch optimizer parameter	Absolute, unnormalized distance at the current iterate; mean $\pm$ SD
Normalizer error	$\left\|\hat{Z}_k / Z_\tau(\boldsymbol{\theta}_k) - 1\right\|$	Exact normalizer at the same iterate	Current iterate and scalar state; saved checkpoint or endpoint
Order contrast	Error_current – Error_previous	Matched previous-state run	Endpoint after 50,000 updates; seedwise difference and descriptive paired interval
Low-rate diagnostic	Through-origin fit of order contrast on $\eta\rho$	Cells with $\rho \leq 0.01$	Signed contrast for uncentered R^2; magnitude for log–log slope

2.9. Ethical approval

The study uses synthetic data and public, de-identified benchmark data. It involves no recruitment, identifiable personal data, or animal experiments; ethical approval was therefore not required.

3. RESULTS

3.1. Finite-sample tracking and denominator order

Both strict one-sample protocols remained close to their full-batch references over the tested toy horizon (Figure 1; Table 4). The exact error definitions, signs, checkpoints, and endpoint roles are given in Table 3. At the reference normalizer rate $\rho = 0.01$, mean objective gaps ranged from 2.27×10^{-5} to 6.93×10^{-4} across the five tilts and two orders; mean parameter errors ranged from 0.0103 to 0.0298. Mean relative normalizer errors were at most 0.0121. These absolute values provide the scale for the paired order contrasts.

CC: BY-NC-ND

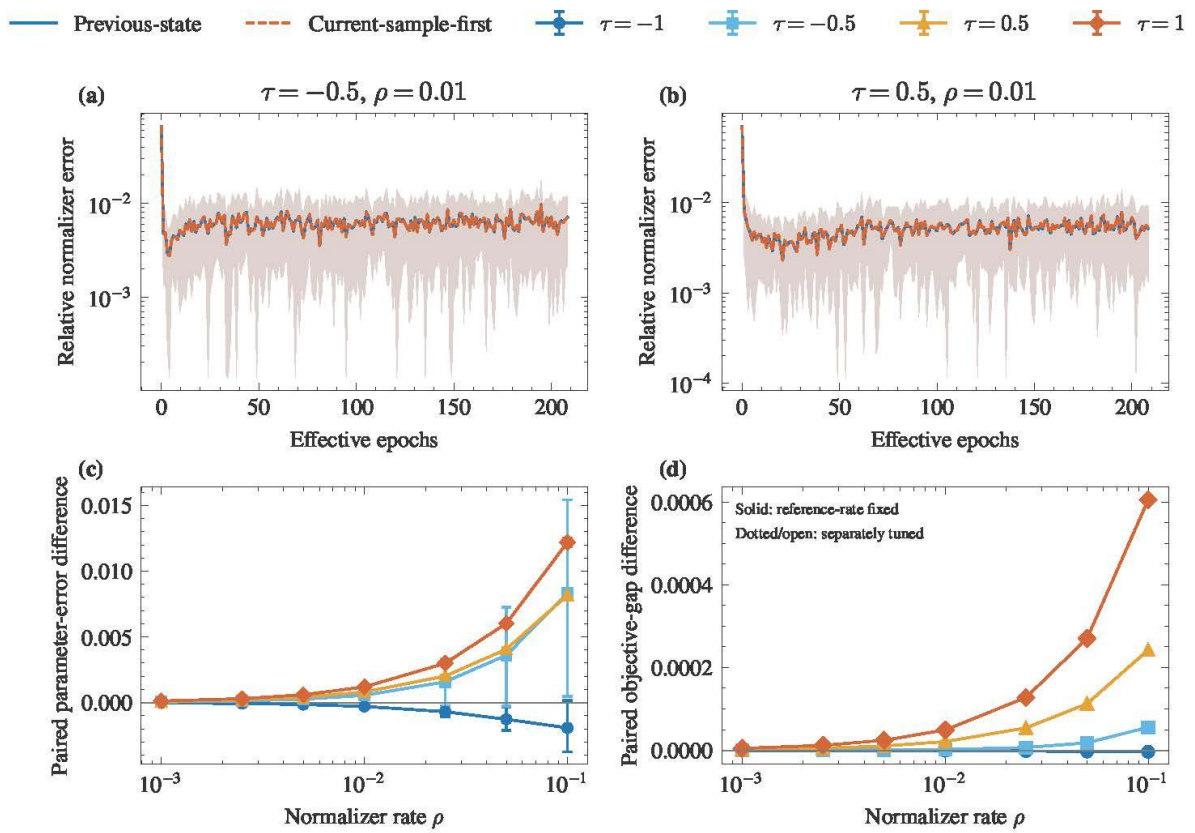

Figure 1. *Finite-step denominator-order comparison in the toy problem. Objective-gap and parameter-error contrasts are current-sample-first minus previous-state; positive values favor previous-state. Absolute errors are included to establish scale. Points and intervals summarize 20 matched streams. At zero tilt the protocols are exactly identical and the contrast is zero.*

Table 4. *Toy endpoint errors at the reference rate $\rho = 0.01$. Values are means over 20 streams; C–P is current-sample-first minus previous-state.*

τ	Previous obj.	Current obj.	Previous param.	C–P param.
−1.0	2.34×10^{-5}	2.27×10^{-5}	0.01058	-2.95×10^{-4}
−0.5	1.31×10^{-4}	1.33×10^{-4}	0.02799	5.48×10^{-4}
0.0	1.62×10^{-4}	1.62×10^{-4}	0.02291	0
0.5	4.24×10^{-4}	4.45×10^{-4}	0.02845	8.04×10^{-4}
1.0	6.43×10^{-4}	6.93×10^{-4}	0.02865	1.19×10^{-3}

At $\tau = 0$, the two protocols were exactly identical for every tested normalizer rate: objective-gap and parameter-error differences were zero, and both reduced to the same ordinary SGD updates. At $\rho = 0.01$, current-sample-first minus previous-state parameter error was -2.95×10^{-4} at $\tau = -1$, 5.48×10^{-4} at $\tau = -0.5$, 8.04×10^{-4} at $\tau = 0.5$, and 1.19×10^{-3} at $\tau = 1$. Thus current-sample-first was slightly better at $\tau = -1$, whereas previous-state was favored at $\tau = -0.5$ and the positive tilts.

The direction and magnitude changed with ρ. For positive tilt, current-sample-first parameter and objective errors separated further from previous-state as ρ increased. The same direction appeared at $\tau = -0.5$, even though current-sample-first had slightly smaller normalizer error; at $\tau = -1$, current-sample-first instead had lower mean parameter and objective error, with a broad paired interval crossing zero at $\rho = 0.1$. Normalizer error alone therefore did not determine parameter error, and neither order was uniformly superior.

CC: BY-NC-ND

In the local range $\rho \leq 0.01$, through-origin regressions of paired parameter-error differences on $\eta\rho$ gave uncentered coefficient of determination $R^2 = 0.99928–1.00000$ across the four nonzero tilts; objective-gap differences gave $R^2 = 0.99845–0.99998$. Corresponding log–log slopes were 0.994–1.043 and 0.985–1.065. This empirically supports the local $\eta\rho$ heuristic in (7) for the tested toy problem, not a global rate. Reference-ρ fixed, globally pooled fixed, and separately tuned roles selected exactly the same learning rate at every tilt: 0.003 at $\tau = -1, -0.5$ and 0.0003 at $\tau = 0,0.5,1$. Their selected metrics were consequently identical. Appendix B provides the full absolute objective, parameter, and normalizer errors across all tested normalizer rates, together with the complete learning-rate tuning surfaces and selection controls.

3.2. Robustness under label noise

Figure 2 reports WDBC performance on clean balanced test labels after independent training/validation label corruption. Negative tilt had the largest displayed mean balanced accuracy and F1 at each of the four discrete conditions. At noise 0, the paired negative-tilt-minus-ERM mean was 0.0029 for balanced accuracy, with interval $[-0.0017,0.0081]$, and 0.0029 for F1, with interval $[-0.0021,0.0083]$ (Table 5). Both intervals included zero.

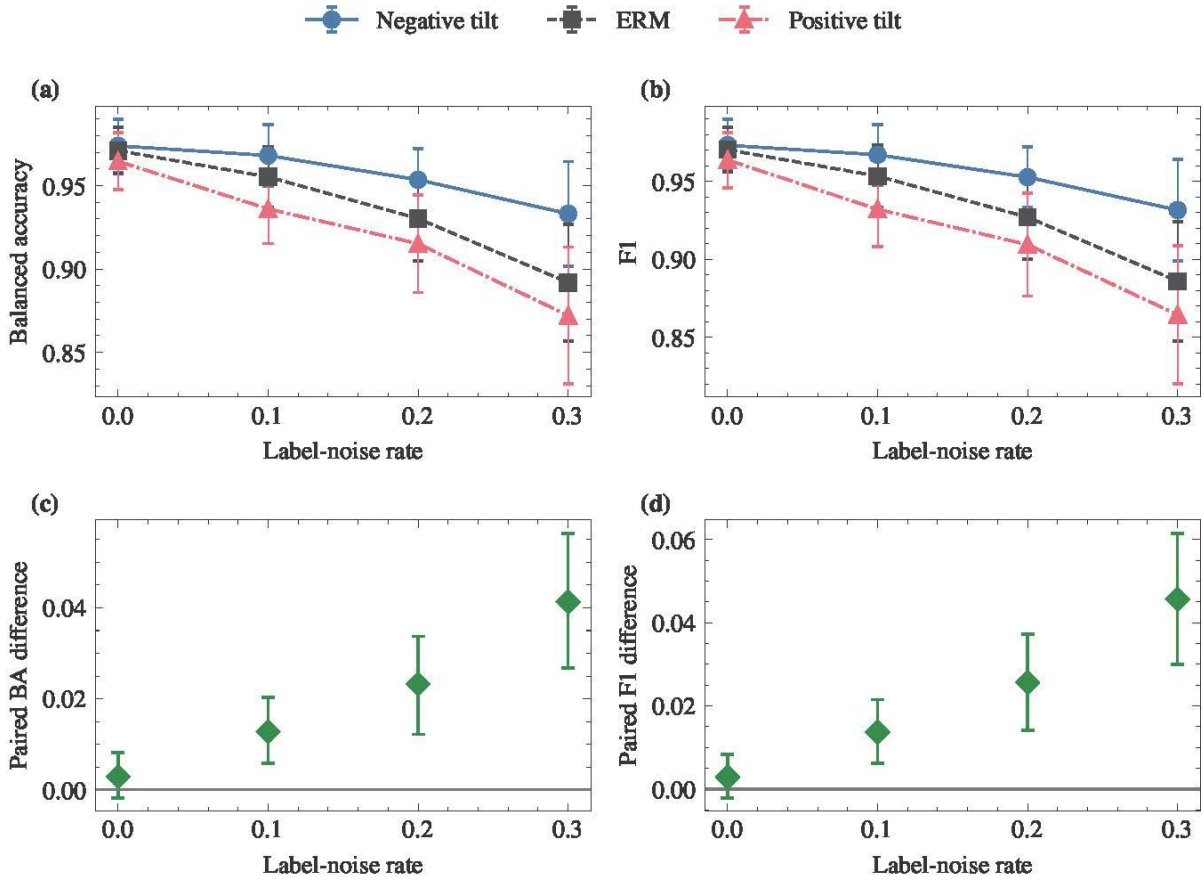

Figure 2. *WDBC label-noise results using Polyak–Ruppert-averaged SGD and clean balanced test labels. Top panels show mean ± one standard deviation over 20 seeds. Bottom panels show paired negative-tilt-minus-ERM differences with descriptive 95% bootstrap intervals. Conditions are discrete and nonnested.*

CC: BY-NC-ND

Table 5. *Paired WDBC negative-tilt-minus-ERM differences. Positive values favor negative tilt; intervals use 10,000 matched-seed bootstrap resamples.*

Noise	Balanced accuracy	F1
0%	0.003 [−0.002, 0.008]	0.003 [−0.002, 0.008]
10%	0.013 [0.006, 0.020]	0.014 [0.006, 0.022]
20%	0.023 [0.012, 0.034]	0.026 [0.014, 0.037]
30%	0.041 [0.027, 0.056]	0.046 [0.030, 0.062]

At noise 0.1, 0.2, and 0.3, the paired intervals were positive for both primary metrics. Balanced-accuracy improvements were 0.0128 [0.0058,0.0203], 0.0233 [0.0122,0.0337], and 0.0413 [0.0267,0.0564], respectively. F1 improvements were 0.0137 [0.0062,0.0215], 0.0257 [0.0141,0.0373], and 0.0457 [0.0300,0.0616]. At 30% noise, mean balanced accuracy was 0.933 for negative tilt and 0.892 for ERM; mean F1 was 0.932 and 0.886. All primary rows comprised 20 finite matched seeds. Although the displayed paired improvements increase numerically across the four conditions, the grid is coarse and nonnested and no predeclared monotonicity analysis was conducted; the conditions are interpreted as discrete comparisons. AUROC and PR-AUC appear in Appendix C.

Against oracle-rate forward loss correction, negative TERM and the comparator had descriptively close primary performance through noise 0.20: negative-TERM-minus-forward balanced-accuracy differences were 0.0029 [−0.0017,0.0081], 0.0052 [−0.0012,0.0116], and 0.0006 [−0.0070,0.0093] at noise 0, 0.1, and 0.2 (Table 7). Corresponding F1 differences were 0.0029 [−0.0021,0.0083], 0.0055 [−0.0012,0.0125], and 0.0011 [−0.0069,0.0103]. At noise 0.30, forward correction was favored: balanced accuracy was 0.9448 versus 0.9331, with paired difference −0.0116 [−0.0192, −0.0052], and the F1 difference was −0.0119 [−0.0196, −0.0051]. Forward correction received the true declared symmetric noise rate; negative TERM did not use a corruption model. Full metrics are in Appendix D.

3.3. Robustness under response contamination

Across the seven discrete Abalone conditions, negative tilt had the lowest displayed mean clean-test RMSE and MAE (Figure 3). Relative to ERM, the paired RMSE interval included zero at contamination 0 but excluded zero from 0.05 through 0.30 (Table 6). The paired MAE interval excluded zero at all seven conditions. At contamination 0.30, mean clean RMSE was 2.628 for negative tilt and 2.712 for ERM, a paired reduction of 0.0841 [0.0791,0.0891]; clean MAE was 1.794 and 1.867, a reduction of 0.0727 [0.0665,0.0793].

CC: BY-NC-ND

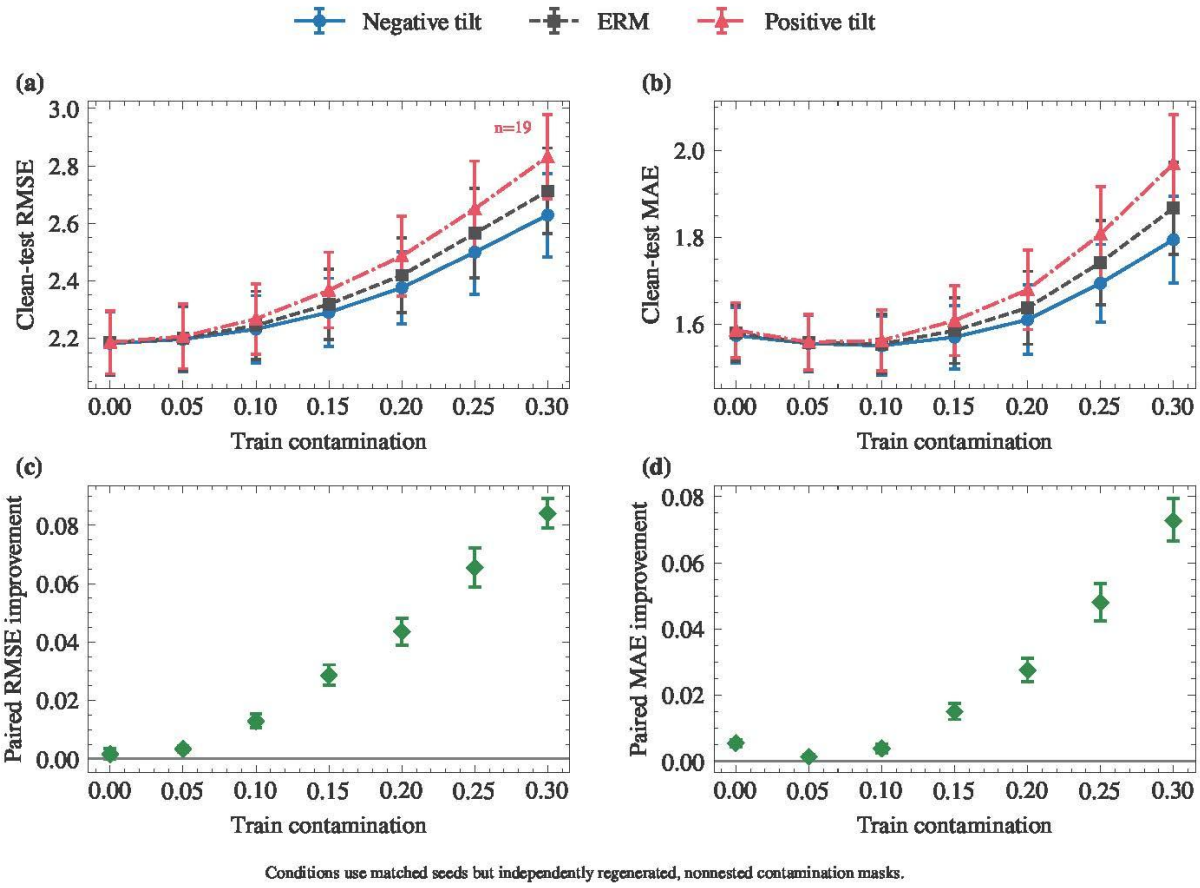

Figure 3. *Abalone response-contamination results using Polyak–Ruppert-averaged SGD. Top panels show clean-test mean ± one standard deviation; bottom panels show paired ERM-minus-negative-tilt error reductions with descriptive 95% bootstrap intervals. Each condition has 20 matched primary pairs. Positive tilt at contamination 0.30 excludes one failed selected run ($n = 19$). Conditions are discrete and nonnested.*

Table 6. *Paired Abalone ERM-minus-negative-tilt clean-error reductions. Positive values favor negative tilt; intervals use 10,000 matched-seed bootstrap resamples.*

c	Clean RMSE reduction	Clean MAE reduction
0.00	0.0016 [−0.0002, 0.0035]	0.0055 [0.0044, 0.0066]
0.05	0.0034 [0.0022, 0.0046]	0.0014 [0.0006, 0.0022]
0.10	0.0129 [0.0106, 0.0154]	0.0040 [0.0027, 0.0053]
0.15	0.0286 [0.0253, 0.0322]	0.0151 [0.0127, 0.0176]
0.20	0.0436 [0.0389, 0.0482]	0.0276 [0.0242, 0.0311]
0.25	0.0655 [0.0588, 0.0723]	0.0480 [0.0425, 0.0538]
0.30	0.0841 [0.0791, 0.0891]	0.0727 [0.0665, 0.0793]

The predeclared monotonic-trend gate failed: displayed MAE improvement decreased from 0.0055 at contamination 0 to 0.0014 at 0.05, even though other predeclared slope and balance checks passed. The evidence therefore supports differences across seven discrete contamination conditions, not a dose response or monotonic trend. Conditions were matched by seed across tilt but nonnested across contamination levels.

Negative tilt and ERM selected the same two-stage SGD schedule at six of seven conditions. At contamination 0.25, negative tilt used $\eta_0 = 0.003$ with switch epoch 10, whereas ERM used $\eta_0 = 0.001$ with switch epoch 25; that condition is schedule-confounded. A candidate-level failure is a seed–schedule run that triggers the invalid-update or divergence rule; failed rows remain in validation and receive the declared penalty rather than being discarded. Positive tilt selected the smaller initial rate 0.001 throughout and incurred all 383 candidate-level tuning failures. Its selected contamination-0.30 cell contains one failed seed (seed 7), so the finite descriptive mean uses $n = 19$. Positive tilt had higher mean clean RMSE and MAE than ERM in every valid comparison. Secondary contaminated-test and stability diagnostics are in Appendix E.

Huber regression was essentially tied with negative TERM in clean RMSE at contamination 0 and 0.05: Huber-minus-negative differences were -0.0011 $[-0.0079, 0.0052]$ and -0.0016 $[-0.0065, 0.0032]$. From contamination 0.10 through 0.30, the RMSE intervals favored Huber; at 0.30, clean RMSE was 2.4766 for Huber and 2.6284 for negative TERM, a difference of -0.1518 $[-0.1585, -0.1459]$. Huber also had lower MAE at every level; at 0.30, the means were 1.6697 and 1.7945, with difference -0.1248 $[-0.1338, -0.1167]$. Huber selected $\delta = 4$, the maximum tested threshold, and $\eta_0 = 0.003$ at every contamination level. It optimizes a different robust loss, so these results show that negative TERM's within-family advantage over ERM does not establish superiority to a task-specific robust regression method.

Table 7. *Representative task-specific robustness comparators under Polyak–Ruppert evaluation. WDBC differences are negative TERM minus oracle-rate, information-advantaged forward correction; Abalone differences are Huber minus negative TERM, and Huber optimizes a different robust loss. Intervals are descriptive paired bootstrap intervals over 20 matched seeds.*

Condition	Negative TERM	Comparator	Paired difference [95% interval]
(a) WDBC balanced accuracy; negative TERM − oracle-rate forward correction			
0%	0.9738	0.9709	0.0029 [−0.0017, 0.0081]
10%	0.9680	0.9628	0.0052 [−0.0012, 0.0116]
20%	0.9535	0.9529	0.0006 [−0.0070, 0.0093]
30%	0.9331	0.9448	−0.0116 [−0.0192, −0.0052]
(b) Abalone clean RMSE; Huber − negative TERM			
0%	2.1823	2.1811	−0.0011 [−0.0079, 0.0052]
5%	2.1963	2.1946	−0.0016 [−0.0065, 0.0032]
10%	2.2317	2.2195	−0.0122 [−0.0157, −0.0090]
15%	2.2901	2.2531	−0.0370 [−0.0419, −0.0323]
20%	2.3756	2.3086	−0.0670 [−0.0728, −0.0617]
25%	2.4996	2.3868	−0.1128 [−0.1199, −0.1061]
30%	2.6284	2.4766	−0.1518 [−0.1585, −0.1459]

4. DISCUSSION

4.1. One-sample tracking and prior TERM theory

The toy study establishes a finite-sample property of the evaluated recursion: a model coupled to one scalar running normalizer can remain close to high-accuracy classical TERM references while each post-initialization update consumes one current observation. Objective gaps, parameter distances, and normalizer errors jointly locate the stochastic endpoints relative to the batch solutions. The result is more informative than numerical survival alone, yet it is deliberately narrower than convergence. The target $Z_\tau(\boldsymbol{\theta}_k)$ moves with the parameter, a constant-rate exponential moving average has lag and sampling variation, and the numerator is instantaneous.

CC: BY-NC-ND

This evidence complements rather than duplicates existing TERM results. Li et al. provide the objective's principal theory and stochastic algorithms [4]; Aminian et al. analyze generalization and robustness, including negative tilt [7]. Our experiments instead hold the finite sample fixed and measure how two strict one-current-sample recursions track its numerical optimum. A small terminal objective gap over 50,000 updates is a finite-horizon fact about this data set, grid, initialization, and stream collection. It does not imply unbiased gradients or convergence on other problems.

The scalar state is necessary because the classical weight of one observation depends on other losses. Its persistence supplies cross-sample information that a batch-size-one evaluation discards. Maintaining an auxiliary state is standard in stochastic approximation and compositional optimization [1], [11]; the TERM-specific contribution here is not that state augmentation exists, but the matched examination of two denominator orders and their errors. Recent compositional entropic-risk methods further confirm the relevance of this broader algorithmic viewpoint [8], while differing in formulation and guarantees.

4.2. Denominator order and rate interaction

Previous-state and current-sample-first updates have identical sample and state budgets but different within-step couplings. The former uses information available before the current observation enters the normalizer. The latter partly self-normalizes that observation by moving the denominator toward its exponential loss before the model step. Neither is intrinsically more faithful at finite rates. By (7), the scalar-weight discrepancy is locally proportional to ρ, and the immediate SGD-step discrepancy is locally proportional to $\eta\rho$ when the other factors remain controlled.

The fine-rate evidence supports that local description in the tested regime. Parameter- and objective-error contrasts had near-unit log–log slopes and high through-origin R^2 against $\eta\rho$. This agreement is not a two-timescale theorem. Classical two-timescale stochastic approximation studies coupled recursions under explicit schedule, regularity, and stability assumptions [10], [12]; those assumptions are not verified here, and both rates need not approach zero in the reported experiments. The result should therefore be read as an empirical mechanism check for small finite rates, not a global asymptotic statement.

The tilt-dependent sign reversal prevents a ranking claim. Current-sample-first had slightly lower mean parameter and objective error at $\tau = -1$, whereas previous-state was favored at $\tau = -0.5$ and positive tilt as ρ increased. At the largest rate, uncertainty under strong negative tilt widened. We retain previous-state as the primary convention because its temporal accounting is explicit: the current observation does not affect the denominator used for its own gradient. The comparator remains valid and may perform better in a specific regime.

The tuning controls reduce an important confound. Reference-rate fixed, globally pooled fixed, and separately tuned policies selected the same learning rate at every tilt, so the reported order pattern was not created by different tuning budgets or selected rates. This equality is grid-specific. It neither establishes optimizer independence nor implies that a different horizon, loss distribution, or normalizer schedule would preserve the same ordering.

4.3. Scale-aware tilt

The normalized tilt addresses a comparability problem rather than a statistical optimality problem. Raw τ multiplies the loss inside an exponential; changing response units or loss conventions changes the operative weighting even when the raw number is unchanged. Dividing a dimensionless $\bar{\tau}$ by an initialization-loss scale makes the exponent regime reproducible across the studied settings and reduces numerical saturation. The classification fallback is needed because zero-initialized logistic losses can be identical, making the percentile spread zero.

This convention remains data- and initialization-dependent. A percentile spread can be unstable on a small sample, can underrepresent rare extreme losses, and changes if the initial model changes. The present paper does not prove that it is optimal. More principled scale estimators, robust online recalibration, and joint adaptation of ρ and τ are natural extensions. Any adaptive scheme must also

CC: BY-NC-ND

preserve the strict resource accounting claimed for model updates.

4.4. Robustness under two corruption mechanisms

The WDBC and Abalone findings are consistent with the loss-weighting mechanism of negative TERM and with recent theoretical interest in negative tilted risk [7]. If corrupted observations remain relatively high-loss, negative tilt gives them smaller normalized exponential weights. This mechanism is corruption-agnostic and conditional: it can also reduce clean hard examples, and corrupted examples that are fitted well need not remain downweighted.

Within the TERM/ERM family, negative tilt improved the specified corrupted conditions. On WDBC, the paired balanced-accuracy and F1 intervals for negative tilt minus ERM were positive at noise 0.1, 0.2, and 0.3; at 30%, balanced accuracy increased from 0.892 to 0.933 and F1 from 0.886 to 0.932. On Abalone, negative tilt had lower displayed mean clean RMSE and MAE than ERM at all seven discrete conditions; at 30%, clean RMSE decreased from 2.712 to 2.628. These comparisons show that the one-sample approximation preserves useful tilt behavior in the tested applications, but ERM alone is not a sufficient robustness benchmark.

The representative task-specific controls narrow that interpretation. Huber regression had lower Abalone MAE at every tested contamination and lower RMSE from 0.10 upward; its high-contamination advantage was materially larger than the negative-TERM-versus-ERM reduction. Huber changes the residual influence through a different robust loss and selected the maximum tested threshold, so this is not an objective-equivalent comparison. On WDBC, oracle-rate forward correction was favored at 30% noise, whereas the primary paired intervals included zero through 20%. Forward correction uses the true declared transition rate, information that TERM does not receive. Thus neither result supplies a universal ranking: it shows that a task-specific method can be stronger when its loss or corruption information matches the problem.

Negative TERM is also distinct from Catoni estimation, robust heavy-tailed empirical risk, conditional value-at-risk, and DRO [5], [6], [14], [15]. These approaches robustify mean estimation, replace the empirical risk estimate, target an upper tail, or optimize over an ambiguity set. The cited Catoni construction does not uniquely specify a supervised linear-regression objective for this protocol, so no ad hoc variant was compared. Robustness is therefore application-level evidence here; the primary contribution remains finite-sample tracking of classical TERM.

The contamination levels must remain discrete. Masks were independently generated and nonnested, and the predeclared joint monotonic gate failed because the displayed MAE improvement at 0.05 was smaller than at zero contamination. The 0.25 Abalone comparison is schedule-confounded, although negative tilt and ERM selected the same schedule at the other six levels. Positive tilt selected smaller rates, produced all candidate-level tuning failures, and retained one failed selected run at contamination 0.30. These facts support caution about high-loss emphasis under response corruption, but not a general instability claim.

4.5. Practical implications, limitations, and future work

The accounting is compact: one current sample is used per model update, TERM adds one persistent scalar, and the update cost remains $O(d)$ for a d-parameter model. Scheduled SGD was the nonadaptive reference, and Polyak–Ruppert averaging reduced dependence on a noisy final iterate at the cost of a parameter accumulator. Adam and the HAdam ablation establish implementation compatibility only; they are not central contributions. The normalization-free online surrogate [9] avoids the scalar state but targets a different risk, so choosing between the two approaches is also a choice of objective.

Several limitations bound interpretation. There is no convergence theorem for the moving-normalizer recursion. Oracle tracking is demonstrated on one two-parameter toy problem, and the applications use linear or logistic models on two public data sets, with WDBC particularly small. Test balancing changes the WDBC evaluation distribution. Corruption conditions are nonnested, horizons can leave optimization floors, and scale calibration and normalizer schedules are heuristic. Bootstrap

CC: BY-NC-ND

intervals summarize 20 matched seeds and are not population-level inference. Huber and oracle-rate forward correction are representative controls, not a comprehensive robustness benchmark. Additional details on the initialization-loss scale, the halted sensitivity analysis, and analyses excluded from the focused evidence hierarchy are provided in Appendix F. The study does not cover deep networks, nonconvex risks, genuine streaming shift, estimated label-transition matrices, or broad robust-risk families.

Future work should analyze the coupled model–normalizer dynamics with moving parameters, including conditions under which separate rate schedules yield tracking or convergence. Adaptive normalizer rates and principled scale estimators could reduce manual calibration. Larger regression and classification benchmarks could study well-defined supervised Catoni-type formulations, broader DRO and CVaR controls, and estimated rather than oracle transition matrices. Nonlinear and deep models, heavier noise, distribution shift, and streaming environments would test whether the finite-sample patterns survive beyond the controlled setting.

5. CONCLUSION

Classical TERM requires cross-sample normalization, so directly applying its log-sum-exp expression to one observation removes the tilt effect. In the evaluated strict one-current-sample protocol, one persistent scalar estimate of the exponential normalizer was sufficient to retain nontrivial tilt-dependent updates. On the controlled toy problem, both previous-state and current-sample-first orders tracked high-accuracy finite-sample oracles, were exactly identical at zero tilt, and differed locally in an approximately $\eta\rho$-scaled, tilt-dependent manner; neither order was uniformly superior.

With scale-calibrated tilt, scheduled SGD, and matched validation selection, negative tilt improved on ERM in the specified corrupted WDBC and Abalone conditions. Representative task-specific controls also showed the boundary of that result: Huber regression was stronger on Abalone at moderate and high contamination, and oracle-rate forward correction was stronger at 30% WDBC label noise. Robustness should therefore be read as application-level behavior rather than a general ranking. The primary result is unchanged: one scalar state provides a reproducible finite-sample bridge between classical normalized TERM and strict one-current-sample optimization, with a local, tilt-dependent $\eta\rho$ order effect and no uniformly superior denominator order. The evidence is finite-horizon and limited to linear and logistic models; theory for coupled dynamics, adaptive rates, principled scale estimators, broader benchmarks, and nonlinear models remains future work.

Peer-review: External, Independent.

Acknowledgements:
This work was supported partially by TÜBİTAK under project 124E063 in the ARDEB 1001 program and partly by KUIS AI Research Center.

Declarations:

1. Statement of Originality:
This work is original.

2. Author Contributions:
Concept: ZD; ***Conceptualization:*** ZD; ***Literature Search:*** ZD; ***Data Collection:*** ZD; ***Data Processing:*** ZD; ***Analysis:*** ZD; ***Writing – original draft:*** ZD; ***Writing – review & editing:*** ZD.

3. Ethics approval:
Not applicable.

4. Funding/Support:
This work was supported partially by TÜBİTAK under project 124E063 in the ARDEB 1001 program and partly by KUIS AI Research Center.

5. Competing Interests:

CC: BY-NC-ND

The author declares no competing interests.

6. GenAI Usage Statement:

GenAI tools were used to assist with language editing, code development and verification, and equation and notation checks. All code, analyses, numerical results, and manuscript text were reviewed and verified by the author, who takes full responsibility for the content.

7. Sustainable Development Goals:

REFERENCES

[1] Wang, M., Fang, E. X., and Liu, H. Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. Mathematical Programming, 161(1–2), 419–449, 2017.

[2] Wang, B. and Yang, T. Finite-sum coupled compositional stochastic optimization: Theory and applications. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 23292–23317. PMLR, 2022.

[3] Wang, B. and Yang, T. A near-optimal single-loop stochastic algorithm for convex finite-sum coupled compositional optimization. In Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, pages 65091–65121. PMLR, 2025.

[4] Li, T., Beirami, A., Sanjabi, M., and Smith, V. On tilted losses in machine learning: Theory and applications. Journal of Machine Learning Research, 24(142), 1–79, 2023.

[5] Rockafellar, R. T. and Uryasev, S. Optimization of conditional value-at-risk. The Journal of Risk, 2(3), 21–41, 2000.

[6] Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. The Annals of Statistics, 49(3), 1378–1406, 2021.

[7] Aminian, G., Asadi, A. R., Li, T., Beirami, A., Reinert, G., and Cohen, S. N. Generalization and robustness of the tilted empirical risk. In Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, pages 1419–1461. PMLR, 2025.

[8] Wei, X., Zhou, L., Wang, B., Lin, C.-J., and Yang, T. A geometry-aware efficient algorithm for compositional entropic risk minimization, 2026. arXiv preprint; accepted to the 43rd International Conference on Machine Learning, final proceedings metadata not yet available.

[9] Yıldırım, Y. E., Demir, S., and Doğan, Z. Benefits of online tilted empirical risk minimization: A case study of outlier detection and robust regression. In 2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP), pages 1–6. IEEE, 2025.

[10] Borkar, V. S. Stochastic approximation with two time scales. Systems & Control Letters, 29(5),

291–294, 1997.

[11] Borkar, V. S. Stochastic Approximation: A Dynamical Systems Viewpoint, volume 48 of Texts and Readings in Mathematics. Springer Nature Singapore, Singapore, 2nd edition, 2023.

[12] Kushner, H. J. and Yin, G. G. Stochastic Approximation and Recursive Algorithms and Applications, volume 35 of Stochastic Modelling and Applied Probability. Springer New York, New York, 2nd edition, 2003.

[13] Huber, P. J. Robust estimation of a location parameter. The Annals of Mathematical Statistics, 35(1), 73–101, 1964.

[14] Catoni, O. Challenging the empirical mean and empirical variance: A deviation study. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, 48(4), 1148–1185, 2012.

[15] Brownlees, C., Joly, E., and Lugosi, G. Empirical risk minimization for heavy-tailed losses. The Annals of Statistics, 43(6), 2507–2536, 2015.

[16] Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In Advances in Neural Information Processing Systems, volume 26, pages 1196–1204, 2013.

[17] Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1944–1952. IEEE, 2017.

[18] Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., and Bailey, J. Normalized loss functions for deep learning with noisy labels. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 6543–6553. PMLR, 2020.

[19] Zhu, Y., Zhang, J., Gangrade, A., and Scott, C. Label noise: Ignorance is bliss. In Advances in Neural Information Processing Systems, volume 37, 2024.

[20] Guo, H., Yi, G. Y., and Wang, B. Learning from noisy labels via conditional distributionally robust optimization. In Advances in Neural Information Processing Systems, volume 37, 2024.

[21] Wolberg, W., Mangasarian, O., Street, N., and Street, W. Breast cancer wisconsin (diagnostic). UCI Machine Learning Repository, 1993. Dataset.

[22] Nash, W., Sellers, T., Talbot, S., Cawthorn, A., and Ford, W. Abalone. UCI Machine Learning Repository, 1994. Dataset.

[23] Polyak, B. T. and Juditsky, A. B. Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization, 30(4), 838–855, 1992.

[24] Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015.

APPENDIXES

A. COMPLETE PROTOCOLS AND REPRODUCIBILITY

CC: BY-NC-ND

Table 8. ***Detailed protocol specification.***

Component	Specification
Toy data	Seed 0; 290 inliers with $x \sim \mathcal{N}(0, 0.5^2)$, $y = -x + 0.2 + \epsilon$, $\epsilon \sim \mathcal{N}(0, 0.1^2)$; 10 outliers with $x \sim \mathcal{N}(-2, 0.5^2)$, $y \sim \mathcal{N}(0, 0.5^2)$. Stratified 80/20 train/validation split, seed 1729.
Toy tuning	Protocols previous-state/current-sample-first; $\tau \in \{-1, -0.5, 0, 0.5, 1\}$; seven ρ values in Table 2; $\eta \in \{10^{-4}, 3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$; 10,000 tuning and 50,000 selected updates; stream seeds 0–19. Validation classical TERM objective selects η without oracle access.
WDBC	Stratified 60/20/20 split within class. Training-fitted standardization; clean test balanced to 43 observations per class. Label flips without replacement in training and validation only. $\bar{\tau} \in \{-0.5, 0, 0.5\}$; 20 tuning and 80 selected epochs; $\rho = 0.01$; stream seed $30{,}000 + s$.
Abalone	Seeded 80/10/10 split; 3,341/418/418 observations. Response replacements $\mathcal{N}(5, 5^2)$ independently within split. $\bar{\tau} \in \{-0.5, 0, 0.5\}$; 50 tuning and 300 selected epochs; $\rho = 0.01$; stream seed $20{,}000 + s$.
SGD schedules	WDBC: $\eta_0 \in \{10^{-3}, 3 \times 10^{-3}, 10^{-2}, 3 \times 10^{-2}\}$, switch epoch 10, decay exponent 1. Abalone: $\eta_0 \in \{10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$, switch epoch 10 or 25, exponent 1.
Comparator grids	WDBC forward correction: oracle-known symmetric noise rate and the four WDBC SGD rates. Abalone Huber: $\delta \in \{1, 2, 4\}$ crossed with $\eta_0 \in \{10^{-3}, 3 \times 10^{-3}\}$, switch epoch 10, exponent 1.
Numerics	$\hat{Z}_0 = 1$; $\varepsilon_Z = \varepsilon_s = 10^{-12}$; double-precision model updates; shifted log-sum-exp; log-domain EMA; stable logistic loss; exact ERM branch; no arbitrary exponent clipping.
Failure rules	Failure retained if an update is invalid or $\| \boldsymbol{\theta} \|_2 > 10^4$, where $\| \boldsymbol{\theta} \|_2$ denotes the Euclidean norm of the parameter vector; nonfinite required quantities are separately flagged. Abalone validation adds 10^4 times failure fraction.
Evaluation	Full-batch loss/gradient only offline. WDBC selects validation balanced accuracy minus failure fraction. Abalone selects contaminated-validation RMSE plus failure penalty. Test data never enter selection.

Full-batch toy references use analytical gradients and the high-accuracy Broyden–Fletcher–Goldfarb–Shanno (BFGS) method with gradient tolerance 10^{-10} and multiple starts. Application oracles use its limited-memory, bounded variant (L-BFGS-B) with analytical gradients, gradient tolerance 10^{-10}, function tolerance 10^{-14}, and at most 2,000 regression or 1,500 classification iterations. They define offline diagnostic references only.

Regression reports RMSE and MAE as defined in Methods. Classification uses malignant as the positive class and reports balanced accuracy, F1, AUROC, and PR-AUC; these metrics and their positive-class components are also defined in Methods. Paired intervals use 10,000 seed-pair resamples. WDBC and Abalone use bootstrap seed 161803; the toy order comparison uses 271828.

CC: BY-NC-ND

B. TOY ABSOLUTE ERRORS AND SELECTION CONTROLS

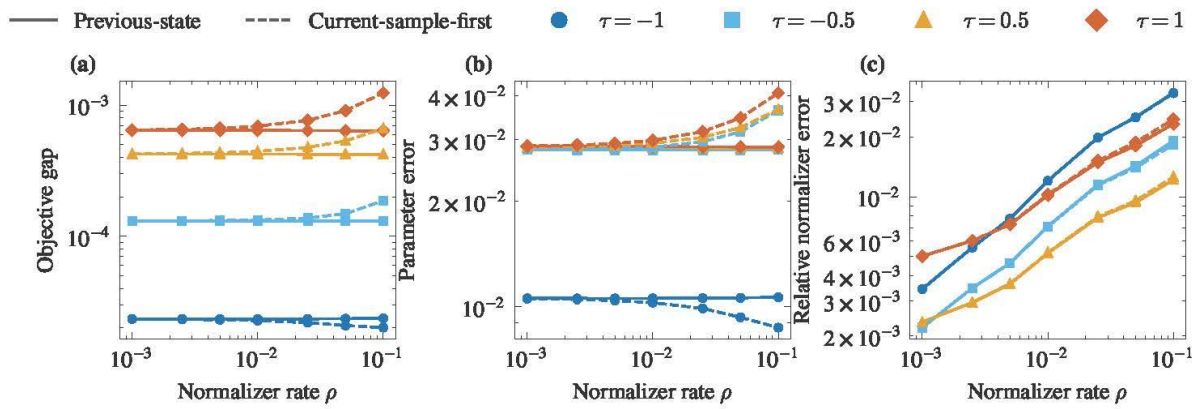

Figure 4. *Absolute toy objective, parameter, and normalizer errors over seven normalizer rates. Points summarize 20 matched streams. Both denominator orders remain close to the full-batch references; normalizer error generally increases with ρ for nonzero tilt.*

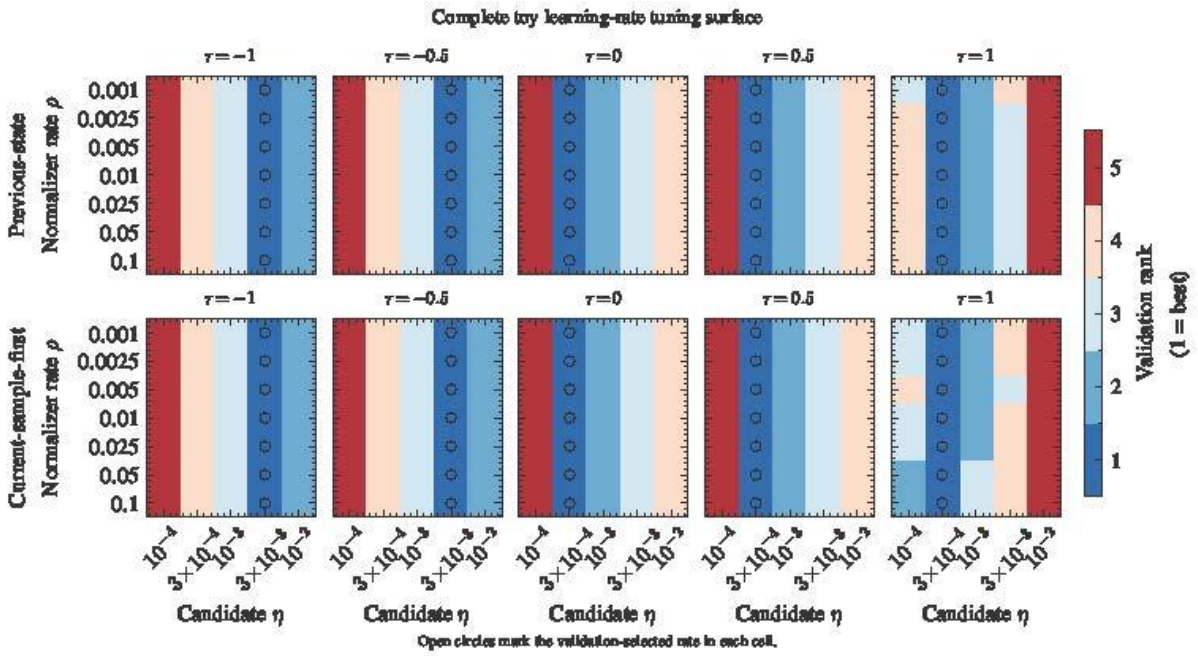

Figure 5. *Complete toy learning-rate tuning surfaces. Colors give within-cell mean validation-objective rank over the five evaluated η candidates; open circles mark the selected rate for each protocol, tilt, and normalizer rate. Validation selects* $\eta = 0.003$ *for* $\tau = -1, -0.5$ *and* $\eta = 0.0003$ *for* $\tau = 0, 0.5, 1$ *throughout. Reference- ρ fixed, globally pooled fixed, and separately tuned roles therefore retain the same tilt-specific selections.*

Table 9. *Toy learning-rate selections for all three analysis roles.*

τ	Reference-ρ η	Global η	Separate min	Separate max
-1	0.003	0.003	0.003	0.003
-0.5	0.003	0.003	0.003	0.003
0	0.0003	0.0003	0.0003	0.0003
0.5	0.0003	0.0003	0.0003	0.0003
1	0.0003	0.0003	0.0003	0.0003

All toy selected runs were finite. At $\rho = 0.1$ and $\tau = -1$, the current-minus-previous parameter-error interval was $[-0.00375, 0.00015]$, illustrating the broad uncertainty behind the strong-negative mean reversal.

C. SECONDARY APPLICATION METRICS

Table 10. *Paired WDBC negative-tilt-minus-ERM secondary metrics. Values are mean differences with descriptive 95% bootstrap intervals.*

Noise	AUROC	PR-AUC
0%	0.0007 [0.0002, 0.0013]	0.0006 [0.0002, 0.0011]
10%	0.0030 [0.0008, 0.0057]	0.0023 [0.0006, 0.0042]
20%	0.0054 [0.0018, 0.0103]	0.0044 [0.0015, 0.0082]
30%	0.0197 [0.0110, 0.0299]	0.0182 [0.0098, 0.0280]

Table 11. *Abalone contaminated-test metrics for Polyak–Ruppert SGD. Entries are mean RMSE/MAE over valid runs.*

c	Negative tilt	ERM	Positive tilt
0.00	2.182/1.573	2.184/1.579	2.186/1.586
0.05	2.657/1.764	2.658/1.765	2.660/1.765
0.10	3.096/1.968	3.095/1.968	3.097/1.970
0.15	3.337/2.142	3.337/2.147	3.343/2.156
0.20	3.598/2.336	3.595/2.346	3.596/2.362
0.25	3.857/2.562	3.856/2.581	3.860/2.609
0.30	4.128/2.810	4.125/2.838	4.143/2.887 (n=19)

RMSE/MAE; the positive-tilt contamination-0.30 cell uses n = 19 because one selected run failed.

The Abalone table reports contaminated-test errors and oracle diagnostics in addition to the clean-test results in the main article. Contaminated-test separation is smaller than clean-test separation.

CC: BY-NC-ND

Normalizer error is evaluated at the final iterate even when performance uses the Polyak–Ruppert parameter.

D. REPRESENTATIVE ROBUSTNESS COMPARATORS

Table 12. *Full WDBC comparison of negative TERM with oracle-rate forward loss correction. Differences are negative TERM minus forward correction; positive values favor negative TERM. Every row has 20 finite matched seeds and zero failures.*

Noise	Metric	Negative TERM	Forward correction	Difference [95% interval]
0%	BA	0.9738	0.9709	0.0029 [−0.0017, 0.0081]
	F1	0.9732	0.9703	0.0029 [−0.0021, 0.0083]
	AUROC	0.9962	0.9955	0.0007 [0.0002, 0.0013]
	PR-AUC	0.9967	0.9961	0.0006 [0.0002, 0.0011]
10%	BA	0.9680	0.9628	0.0052 [−0.0012, 0.0116]
	F1	0.9671	0.9616	0.0055 [−0.0012, 0.0125]
	AUROC	0.9951	0.9946	0.0005 [−0.0003, 0.0013]
	PR-AUC	0.9959	0.9954	0.0005 [−0.0002, 0.0011]
20%	BA	0.9535	0.9529	0.0006 [−0.0070, 0.0093]
	F1	0.9527	0.9516	0.0011 [−0.0069, 0.0103]
	AUROC	0.9909	0.9908	0.0001 [−0.0014, 0.0018]
	PR-AUC	0.9922	0.9921	0.0002 [−0.0012, 0.0018]
30%	BA	0.9331	0.9448	−0.0116 [−0.0192, −0.0052]
	F1	0.9317	0.9436	−0.0119 [−0.0196, −0.0051]
	AUROC	0.9817	0.9853	−0.0036 [−0.0060, −0.0015]
	PR-AUC	0.9843	0.9874	−0.0031 [−0.0052, −0.0012]

Forward correction used the declared symmetric noise rate, so it had corruption information unavailable to TERM. Validation selected $\eta_0 = 0.01$, 0.003, 0.01, and 0.01 at noise 0, 0.10, 0.20, and 0.30, respectively; every schedule switched at epoch 10 and decayed with exponent 1. Selection used corrupted-validation balanced accuracy only.

Table 13. *Full Abalone comparison of negative TERM with Huber regression. Differences are Huber error minus negative-TERM error; negative values favor Huber. Every row has 20 finite matched seeds and zero failures.*

Contam.	Metric	Negative TERM	Huber	Huber − negative [95% interval]
0%	RMSE	2.1823	2.1811	−0.0011 [−0.0079, 0.0052]
	MAE	1.5734	1.5592	−0.0142 [−0.0181, −0.0103]
5%	RMSE	2.1963	2.1946	−0.0016 [−0.0065, 0.0032]
	MAE	1.5549	1.5498	−0.0051 [−0.0083, −0.0021]
10%	RMSE	2.2317	2.2195	−0.0122 [−0.0157, −0.0090]
	MAE	1.5497	1.5449	−0.0048 [−0.0072, −0.0025]
15%	RMSE	2.2901	2.2531	−0.0370 [−0.0419, −0.0323]
	MAE	1.5696	1.5499	−0.0196 [−0.0230, −0.0163]

CC: BY-NC-ND

Contam.	Metric	Negative TERM	Huber	Huber − negative [95% interval]
20%	RMSE	2.3756	2.3086	−0.0670 [−0.0728, −0.0617]
	MAE	1.6097	1.5686	−0.0411 [−0.0456, −0.0366]
25%	RMSE	2.4996	2.3868	−0.1128 [−0.1199, −0.1061]
	MAE	1.6938	1.6113	−0.0825 [−0.0897, −0.0759]
30%	RMSE	2.6284	2.4766	−0.1518 [−0.1585, −0.1459]
	MAE	1.7945	1.6697	−0.1248 [−0.1338, −0.1167]

Huber selected $\delta = 4$ and $\eta_0 = 0.003$ at every contamination level; $\delta = 4$ is the maximum tested threshold. The six-candidate grid crossed $\delta \in \{1,2,4\}$ with $\eta_0 \in \{0.001,0.003\}$, with switch epoch 10 and decay exponent 1. Huber received no contamination ratio, and selection used contaminated-validation RMSE plus the same failure penalty. The boundary threshold is reported as a tuning outcome and is not interpreted as an optimal population value.

E. NUMERICAL AND COMPARATOR CONTROLS

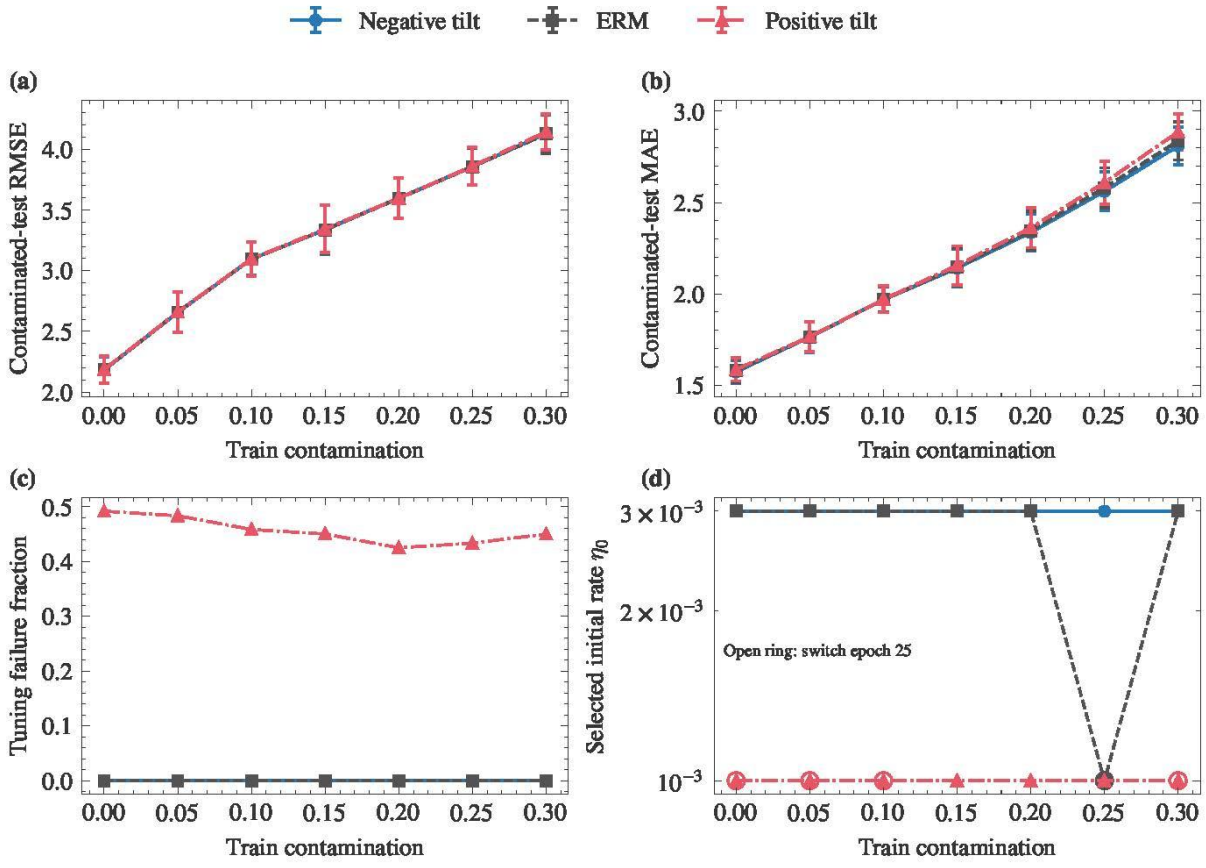

Figure 6. *Abalone secondary diagnostics: contaminated-test metrics, candidate failure fractions, and selected learning-rate schedules. Positive tilt accounts for all candidate-level failures and one selected failure at contamination 0.30.*

CC: BY-NC-ND

Table 14. *Representative application denominator-order control at the 30% corruption conditions. Mean ± standard deviation uses Polyak–Ruppert evaluation.*

Application, $\bar{\tau}$	Order	Primary metric	Objective gap
Abalone, -0.5	Previous	2.6383 ± 0.1470	0.0124 ± 0.0051
Abalone, -0.5	Current	2.6389 ± 0.1471	0.0124 ± 0.0051
Abalone, 0	Previous/current	2.7201 ± 0.1518	0.0117 ± 0.0051
WDBC, -0.5	Previous	0.9331 ± 0.0313	0.0748 ± 0.0153
WDBC, -0.5	Current	0.9459 ± 0.0273	0.0892 ± 0.0166
WDBC, 0	Previous/current	0.8919 ± 0.0350	0.0363 ± 0.0112

Application protocols received equal tuning budgets but were selected separately, so the WDBC negative-tilt rows also differ in selected learning rate. They are controls on practical finite-step implementations, not isolated denominator effects.

Table 15. *Compact optimizer control retained from the broader classification analysis. All methods also keep one scalar TERM normalizer.*

a) Pooled performance

Optimizer	Mean objective gap	Mean balanced accuracy
Adam	0.020	0.882
HAdam	0.023	0.882
SGD	0.027	0.880

b) Implementation cost

Optimizer	Runtime (s)	Relative runtime	Vector states
Adam	0.120	1.00	2
HAdam	0.150	1.25	3
SGD	0.102	0.84	0

Polyak results pooled over both datasets, both classification settings, every saved level, all three normalized tilts, and 20 seeds per condition. All optimizers also retain one scalar TERM normalizer; Adam stores first- and second-moment vectors, and HAdam additionally stores a fourth-moment vector.

SGD is the nonadaptive reference, Adam the standard adaptive baseline [24], and HAdam an auxiliary fourth-moment ablation. This pooled control includes analyses outside the focused main evidence; it establishes implementation compatibility only. HAdam retains one additional vector and has no consistent aggregate advantage in these saved controls.

F. SCALE AND EXCLUDED ANALYSES

The initialization-loss scale is retained as a transparent default. A planned sensitivity study was stopped at its numerical gate: some positive-tilt Abalone scale/learning-rate combinations were nonfinite, while zero-initialized WDBC logistic loss required the median fallback because dispersion scales vanished. The halted study is not treated as a completed result and supports no scale-estimator ranking.

CC: BY-NC-ND

Class-imbalance experiments, extensive optimizer/HAdam diagnostics, the unnormalized exponential surrogate, and prospective high-dimensional theory remain available in the repository history but are excluded from the focused evidence hierarchy. Positive tilt in particular emphasizes high-loss observations rather than minority labels directly. No deep-network, streaming-distribution-shift, or state-evolution result is claimed.

Erciyes University Journal of Institute Of Science and Technology 42(3), 2026
CC: BY-NC-ND