ONLINE NEWTON LEARNING FOR MULTI-OUTPUT NONLINEAR MODELS UNDER NON-I.I.D. DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

We study online learning for a class of multi-output nonlinear stochastic regression models under non-i.i.d. data and general convex losses. This problem arises in several representative settings, including multi-output linear regression, learning from component-wise saturated observations, and multiclass softmax classification under the KL-divergence loss. We develop an adaptive Newton-type estimation algorithm that combines observation-driven gradients, recursively accumulated second-order information, and local curvature of the nonlinear function, allowing its updates to adapt to the evolving geometry of sequential data. We then establish parameter convergence without imposing i.i.d. data assumptions, requiring only an information-growth condition that matches the minimal requirement for least-squares estimation in linear stochastic regression. Moreover, we show that the average prediction regret decays at a logarithmic-over-linear rate without requiring additional excitation on the data. Experiments on synthetic and real-world datasets corroborate the theoretical results and demonstrate the effectiveness of the proposed method.

1 INTRODUCTION

Rapid advances in data-driven artificial intelligence have made machine learning central to prediction and decision-making. The statistical structure of the data is therefore fundamental to both practical performance and theoretical guarantees. In much of machine learning and statistics, learning is formulated either as deterministic optimization over a fixed dataset or as statistical risk minimization, where one seeks to minimize an expected loss under i.i.d. sampling; extensions beyond independence often retain stationarity and mixing data assumptions (Mohri and Rostamizadeh, 2007). By contrast, in many modern applications, including time-series forecasting, streaming regression, and feedback control, data arrive sequentially, and current observations may depend on past observations, decisions, or predictions, making ideal data assumptions difficult to justify (Anava et al., 2013). Online learning provides a natural framework for processing such data streams, but obtaining theoretical guarantees under weaker data assumptions remains challenging (Agarwal and Duchi, 2013).

Research on learning from dependent data has progressively relaxed classical data assumptions while retaining theoretical guarantees. Stability-based analyses establish generalization guarantees for online algorithms under mixing processes (Agarwal and Duchi, 2013), while stochastic optimization studies characterize the effects of mixing or Markovian sampling on the convergence and sample complexity of SGD (Ma et al., 2022; Even, 2023). Beyond mixing assumptions, Kuznetsov and Mohri (2015) derive prediction bounds for non-stationary, non-mixing processes in terms of sequential complexity and discrepancy measures. Excitation constitutes another important data requirement in online estimation and prediction. Lai and Zhu (1991) established asymptotic prediction guarantees for nonlinear autoregressive models under persistent excitation (PE) conditions. These developments substantially extend learning theory beyond the i.i.d. data setting, but guarantees for broad classes of nonlinear models still typically require nontrivial assumptions on either data dependence or persistent excitation. Establishing online learning guarantees for broader nonlinear models under weaker data conditions remains theoretically important and technically challenging.

First-order gradient methods are attractive in streaming data settings because of their low computational cost. AdaGrad adapts updates to coordinate-wise or full-matrix gradient statistics (Duchi

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2011). MetaGrad combines multiple learning rates and provides full-matrix, sketched, and coordinate-wise variants (van Erven et al., 2021). Recent gradient-based adaptive prediction theory establishes guarantees for scalar-output nonlinear models (Liu et al., 2026). These results demonstrate the feasibility of first-order recursive learning for nonlinear prediction with dependent sequential data. Nevertheless, the corresponding recursion relies on a scalar adaptive gain and may converge slowly on ill-conditioned problems (Godichon-Baggioni and Werge, 2025). Improving convergence without sacrificing stability thus calls for updates that adapt to the local curvature.

Second-order methods offer a natural alternative. Hazan et al. (2007) introduced the Online Newton Step algorithm and established logarithmic prediction error bounds for exp-concave losses. Subsequent work develops sketched and kernelized variants and computationally efficient quasi-Newton updates (Luo et al., 2016; Calandriello et al., 2017; Mhammedi and Gatmiry, 2023). Online Newton-type methods have also been studied in structured settings, including generalized linear bandits and logistic regression (Jun et al., 2017; Bercu et al., 2020). Recent stochastic quasi-Newton methods also address ill-conditioned streaming optimization, although their analyses require the i.i.d. data condition (Godichon-Baggioni and Werge, 2025).

Classical foundations also come from the system identification field. A substantial body of identification work studies linear stochastic systems (Åström and Wittenmark, 1973; Ljung, 1976; Moore, 1978; Christopeit and Helmes, 1980; Chen, 1982; Lai and Wei, 1986; Kumar, 1990; Guo and Chen, 1991; Guo, 1995; Chen and Guo, 1991). In particular, Lai and Wei (1982) used stochastic Lyapunov functions and martingale convergence theorems to establish strong consistency of least-squares estimates under an essentially minimal information-growth condition weaker than the PE condition. Beyond linear systems, Newton-type identification has recently been developed for stochastic systems with saturated observations without periodic or independent signal assumptions (Zhang and Guo, 2025). More directly related to our setting, Zhang and Zhang (2025) established online learning guarantees for nonlinear dynamical systems under squared prediction loss.

These developments motivate online methods that exploit curvature information to improve the convergence rate while extending learning guarantees beyond squared prediction loss under weak data assumptions. Accordingly, this paper studies online Newton learning for multi-output nonlinear stochastic regression models with a broad class of prediction losses under weaker data assumptions. The main contributions are summarized as follows:

- We develop an online learning framework for multi-output nonlinear stochastic regression models with general prediction losses and propose an online Newton-type estimation algorithm. The algorithm combines observation-driven gradients, recursively accumulated second-order information, and local curvature information for recursive parameter updates. Without explicitly forming or inverting the full Hessian matrix, the proposed method applies to a broad class of settings, including multi-output linear regression, component-wise saturated observations, and multiclass softmax classification with KL-divergence loss.
- We establish almost-sure performance guarantees for prediction and parameter estimation. Without additional excitation conditions, the average prediction error converges at a logarithmic-over-linear rate, improving upon existing guarantees for stochastic-gradient methods, e.g., Liu et al. (2026). Moreover, the parameter estimates are strongly consistent when the smallest eigenvalue of the adaptive information matrix dominates its log-determinant, requiring an information-growth condition of the same order as the minimal requirement for least-squares estimation in linear stochastic regression (Lai and Wei, 1982).
- We perform extensive experiments from both theoretical and practical perspectives. Synthetic experiments verify the prediction and parameter estimation guarantees under weak excitation conditions, while real-world experiments on PAMAP2 demonstrate the effectiveness and stability of the proposed method for online multiclass classification under both chronological and randomly permuted data orders.

The remainder of the paper is organized as follows: Section 2 introduces the multi-output nonlinear model, the prediction loss, and the assumptions, together with representative examples. Section 3 presents the projected Newton-type online estimation algorithm. Section 4 states the main parameter-estimation and prediction guarantees, whose proofs are given in the Appendix. Section 5 reports some experimental results, and Section 6 concludes the paper.

2 PROBLEM SETUP

2.1 MODELS AND ASSUMPTIONS

Consider the following multi-output nonlinear stochastic dynamical model:

$$y_{k+1} = g\left(\Theta^{*\top}\phi_k, e_{k+1}\right), \quad k \geq 0, \tag{1}$$

where $y_{k+1} \in \mathcal{Y} \subset \mathbb{R}^m$ is the observation vector, $\phi_k \in \mathbb{R}^d$ is the regression vector, $e_{k+1} \in \mathbb{R}^m$ is the noise vector, $\Theta^* \in \mathbb{R}^{d\times m}$ is the unknown parameter matrix, and $g : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ is the nonlinear system function.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, and let $\{\mathcal{F}_k\}_{k\geq 0}$ be a filtration representing the information available up to time k, and $\mathcal{F}_k = \sigma\big(\phi_0, e_1, \phi_1, \ldots, e_k, \phi_k\big)$. To evaluate the prediction performance, let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a known loss function, where $L(y_k, \hat{y}_k)$ quantifies the discrepancy between y_k and its prediction $\hat{y}_k$. We make the following assumption on L.

Assumption 1 *For every $k \geq 0$ and $x \in \mathbb{R}^m$, the conditional risk $\mathbb{E}\left[L\big(g(x, e_{k+1}), z\big) \mid \mathcal{F}_k\right]$ admits a unique minimizer over $z \in \mathcal{Y}$.*

Remark 1 *Assumption 1 is a joint condition on the loss function L, the system mapping g, and the distribution of the noise e_{k+1}. It covers a broad class of models and loss functions, including linear regression with the least-squares loss, saturated observation models with the least-squares loss, and softmax models with the KL-divergence loss. See Examples 1–3 below for details.*

Under Assumption 1, the unique minimizer of the conditional risk defines a function $f : \mathbb{R}^m \to \mathbb{R}^m$ given by $f(x) = \arg\min_{z\in\mathcal{Y}} \mathbb{E}[L(g(x, e_{k+1}), z) \mid \mathcal{F}_k]$. If Θ^* were known, the unique optimal predictor of y_{k+1} under the loss L would therefore be given by

$$f(\Theta^{*\top}\phi_k) = \arg\min_{z\in\mathcal{Y}} \mathbb{E}[L(y_{k+1}, z) \mid \mathcal{F}_k]. \tag{2}$$

When Θ^* is unknown, our objective is to develop an online algorithm that, at each time $k \geq 0$, estimates Θ^* by $\hat{\Theta}_k$ and predicts y_{k+1} by $f(\hat{\Theta}_k^\top \phi_k)$, using the historical data $\{(y_{\tau+1}, \phi_\tau)\}_{\tau=0}^{k-1}$ and the current regressor ϕ_k.

At each time $k \geq 0$, for a candidate parameter Θ, we define the prediction loss relative to the true-parameter prediction by

$$J_k(\Theta) = L\left(f(\Theta^{*\top}\phi_k), f(\Theta^\top\phi_k)\right). \tag{3}$$

In particular, for the online estimate $\hat{\Theta}_k$ at time k, $J_k(\hat{\Theta}_k)$ measures the prediction loss between $f(\Theta^{*\top}\phi_k)$ and $f(\hat{\Theta}_k^\top\phi_k)$. To evaluate the cumulative predictive performance of the online algorithm, we define the accumulated prediction error up to time k as

$$R_k = \sum_{\tau=0}^{k} J_\tau(\hat{\Theta}_\tau). \tag{4}$$

Accordingly, the objective is to design an online algorithm such that the resulting accumulated prediction regret R_k admits a sublinear growth rate almost surely, i.e., $R_k = o(k)$, a.s.

Now we introduce basic assumptions on the input data, parameter domain, prediction function, and loss function. Before stating these assumptions, we first define some notations.

Throughout the sequel, let $\|\cdot\|$ denote the Euclidean norm for vectors and the spectral norm for matrices, let $\|\cdot\|_F$ denote the Frobenius norm for matrices, and let $|P|$ denote the determinant of a square matrix P. The maximum and minimum eigenvalues of a square matrix X are denoted by $\lambda_{\max}\{X\}$ and $\lambda_{\min}\{X\}$, respectively. Moreover, for two real sequences $\{a_n\}, \{b_n\}$ with $b_n > 0$, $a_n = o(b_n)$ means that $a_n/b_n \to 0$ as $n \to \infty$. Also, $a_n = O(b_n)$ means that there exists a positive constant C such that $|a_n|/b_n \leq C$ for all $n > 0$.

For the theoretical analysis, we denote the gradient noise sequence as

$$w_{k+1} = \nabla_x L(y_{k+1}, f(\hat{\Theta}_k^T\phi_k)) - \nabla_x L(f(\Theta^{*T}\phi_k), f(\hat{\Theta}_k^T\phi_k)), \tag{5}$$

where $\nabla_x L(y, x) = \frac{\partial L(y,x)}{\partial x}$, $\hat{\Theta}_k$ is the estimate at time k obtained by our algorithm below. Consequently, we introduce the following assumptions:

Assumption 2 *The gradient noise process $\{w_k, \mathcal{F}_k\}$ is a martingale difference sequence adapted to $\{\mathcal{F}_k\}$. Moreover, there exist a finite constant $\sigma^2 > 0$ and a constant $\alpha > 2$ such that*

$$\sup_{k\geq 0} \mathbb{E}\left[\|w_{k+1}\|^{\alpha} \mid \mathcal{F}_k\right] < \sigma^2. \tag{6}$$

Assumption 3 *There exist constants $\delta > 0$, $c_1 \geq 0$, and $c_2 \geq 0$ such that for each $k \geq 0$*

1. *The gradient of the instantaneous loss satisfies*

$$\operatorname{tr}\left(\nabla_{\Theta} J_k(\Theta)^T(\Theta - \Theta^*)\right) \geq \delta J_k(\Theta); \tag{7}$$

2. *The squared norm of the loss gradient with respect to the prediction variable is bounded by*

$$\left\|\nabla_x L\left(f(\Theta^{*T}\phi_k), f(\Theta^T\phi_k)\right)\right\|^2 \leq c_1 J_k(\Theta) + c_2. \tag{8}$$

Remark 2 *The condition (7) is the δ-weak quasi-convexity condition relative to Θ^* (Hardt et al., 2018; Liu et al., 2026). It does not require J_k to be globally convex in Θ. In particular, if J_k is differentiable and convex in Θ, then (7) holds with $\delta = 1$, whereas the converse need not hold.*

Assumption 4 *For any $k \geq 0$, ϕ_k is $\mathcal{F}_k$-measurable and satisfies*

$$\sup_{0\leq k<\infty} \|\phi_k\| \leq N, \quad a.s. \tag{9}$$

Moreover, the unknown parameter matrix belongs to a known bounded set, i.e., Θ^* satisfies

$$\|\Theta^*\|_F \leq L, \tag{10}$$

where $N, L > 0$ are constants.

Assumption 5 *The prediction function f is continuously differentiable, and satisfies*

$$0 < m_f = \inf_{\|x\|\leq NL} \lambda_{\min}\{J_f(x)\} \leq \sup_{\|x\|\leq NL} \lambda_{\max}\{J_f(x)\} = M_f < \infty, \tag{11}$$

where N and L are defined in Assumption 4, $J_f(x)$ is the symmetric Jacobian matrix of $f(x)$.

Assumption 6 *The loss function $L(y, x)$ is twice continuously differentiable in x and satisfies $L(y, x) \geq 0$ for all admissible y and x, and $L(y, x) = 0$ if and only if $x = y$. Moreover, there exist constants $0 < m_L \leq M_L < \infty$ such that, for any two feasible prediction vectors $s = f(u)$ and $q = f(v)$ with $\|u\| \leq NL, \|v\| \leq NL$, and for every $\xi = (1-t)s + tq$ with $t \in [0, 1]$,*

$$m_L I_m \preceq \nabla^2_{xx} L(s, \xi) \preceq M_L I_m. \tag{12}$$

2.2 Examples

To illustrate the applicability of the proposed framework, we list several practical multi-output examples. Throughout this subsection, the parameter estimate Θ is restricted to the compact set $\|\Theta\|_F \leq L$, which will be ensured by the projection step in the proposed algorithm introduced later. Verification of the above assumptions is provided in the appendix.

Example 1 (Multi-output linear regression). Consider the linear model

$$y_{k+1} = \Theta^{*T}\phi_k + e_{k+1}, \tag{13}$$

where $y_{k+1} \in \mathcal{Y} = \mathbb{R}^m$, $\{e_{k+1}, \mathcal{F}_{k+1}\}$ is a martingale difference sequence satisfying the conditional moment bound in Assumption 2. We use the quadratic loss $L(s, x) = \frac{1}{2}\|s - x\|^2$, and the instantaneous loss is defined as $J_k(\Theta) = \frac{1}{2}\|(\Theta - \Theta^*)^T\phi_k\|^2$.

Example 2 (Multi-output saturated observations). Consider the saturation model as proposed in Wang et al. (2022). We extend the saturation to a component-wise multi-output model. Let

$$y_{k+1} = S(\Theta^{*T}\phi_k + e_{k+1}), \tag{14}$$

where $y_{k+1} \in \mathcal{Y} = \mathbb{R}^m$, $S : \mathbb{R}^m \to \mathbb{R}^m$ is defined component-wise by

$$S_i(z) = L_i \mathbf{1}_{\{z_i \leq L_i\}} + z_i \mathbf{1}_{\{L_i < z_i < U_i\}} + U_i \mathbf{1}_{\{z_i \geq U_i\}}, \qquad -\infty < L_i < U_i < \infty. \tag{15}$$

Assume that e_{k+1} is $\mathcal{F}_{k+1}$-measurable, independent of $\mathcal{F}_k$, and distributed as $\mathcal{N}(0, I_m)$. Let p_i and F_i denote the density and distribution function of its i-th component, respectively. The optimal mean-square predictor is the conditional expectation $f(x) = \mathbb{E}\left[S(x + e_{k+1}) \mid \mathcal{F}_k\right] \triangleq G(x)$, where $G_i(x_i) = \int_{-\infty}^{\infty} S_i(x_i + z)p_i(z)\,dz$. Here we also use the quadratic loss $L(s, x) = \frac{1}{2}\|s - x\|^2$. Consequently, the instantaneous loss is $J_k(\Theta) = \frac{1}{2}\|G(\Theta^{*\top}\phi_k) - G(\Theta^{\top}\phi_k)\|^2$.

Example 3 (Multi-output softmax classification with KL-divergence loss). Consider the stochastic model

$$y_{k+1} = f(\Theta^{*T}\phi_k) + e_{k+1}, \tag{16}$$

where $y_{k+1} \in \mathcal{Y} = [0,1]^m$, $\{e_{k+1}, \mathcal{F}_{k+1}\}$ is a martingale difference sequence satisfying the conditional moment bound in Assumption 2. Here $f : \mathbb{R}^m \to \mathbb{R}^m$ is the reduced softmax predictor for $(m+1)$ classes, defined by $f^{(i)}(x) = e^{x_i}/(1 + \sum_{j=1}^m e^{x_j})$, $i = 1, \ldots, m$, with reference-class probability $f^{(m+1)}(x) = 1 - \sum_{i=1}^m f^{(i)}(x) = 1/(1 + \sum_{j=1}^m e^{x_j})$. For probability vectors $s, x \in \mathbb{R}^m$, set $s_{m+1} = 1 - \sum_{i=1}^m s_i$ and $x_{m+1} = 1 - \sum_{i=1}^m x_i$. We use the KL-divergence loss $L(s,x) = \sum_{i=1}^{m+1} s_i \log \frac{s_i}{x_i}$, where $s_i \geq 0$ and $x_i > 0$ for $i = 1, \ldots, m+1$, and $L(s,x) := \infty$ when $x = 0$. The instantaneous prediction loss is then $J_k(\Theta) = \sum_{i=1}^{m+1} p_{k,i} \log \frac{p_{k,i}}{q_{k,i}}$, where $p_{k,i} = f^{(i)}(\Theta^{*T}\phi_k)$ and $q_{k,i} = f^{(i)}(\Theta^T\phi_k)$ are the optimal and candidate class probabilities, respectively, and $p_{k,m+1} = 1 - \sum_{i=1}^m p_{k,i}$, $q_{k,m+1} = 1 - \sum_{i=1}^m q_{k,i}$.

Since the cross-entropy loss $L_{\mathrm{CE}}(p_k, q_k) = J_k(\Theta) + H(p_k)$ and $H(p_k) = -\sum_{i=1}^{m+1} p_{k,i} \log p_{k,i}$ is independent of Θ, minimizing cross-entropy is equivalent to minimizing $J_k(\Theta)$.

3 Online Newton Learning Method

We next develop an online Newton-type learning algorithm for estimating the unknown parameter Θ^*, which integrates observation-driven gradient information, recursively accumulated second-order information, and local curvature information. Here we consider the observation-based loss $\ell_k(\Theta) := L\left(y_{k+1}, f(\Theta^\top\phi_k)\right)$, whose gradient at the current estimate is

$$\widehat{G}_k := \nabla_\Theta \ell_k(\hat{\Theta}_k) = \phi_k \left[J_f(\hat{\Theta}_k^\top \phi_k)^\top \nabla_x L\left(y_{k+1}, f(\hat{\Theta}_k^\top \phi_k)\right)\right]^\top. \tag{17}$$

To incorporate local curvature, define

$$H_k := \nabla_{xx}^2 L\left(f(\hat{\Theta}_k^\top \phi_k), f(\hat{\Theta}_k^\top \phi_k)\right), \tag{18}$$

$$h_k := \lambda_{\min}\left(J_f(\hat{\Theta}_k^\top \phi_k)^\top H_k J_f(\hat{\Theta}_k^\top \phi_k)\right). \tag{19}$$

Both quantities are computable from the current estimate and regressor, where the scalar h_k provides a lower bound on the local curvature matrix $J_f(\hat{\Theta}_k^\top \phi_k)^\top H_k J_f(\hat{\Theta}_k^\top \phi_k)$.

Let $\beta_1 > \beta_2 > 0$ and initialize $P_0 = I_d$. Given $\hat{\Theta}_k$ and P_k, define

$$a_k := \frac{1}{1 + \beta_1 h_k \phi_k^\top P_k \phi_k}. \tag{20}$$

The matrix gain is updated recursively as

$$P_{k+1} = P_k - \beta_2 a_k h_k P_k \phi_k \phi_k^\top P_k. \tag{21}$$

Applying the matrix inversion formula gives

$$P_{k+1}^{-1} = P_k^{-1} + \frac{\beta_2 h_k}{1 + (\beta_1 - \beta_2) h_k \phi_k^\top P_k \phi_k} \phi_k \phi_k^\top. \tag{22}$$

Since $h_k > 0$ under Assumptions 6 and 5, and $P_0^{-1} = I_d \succ 0$, it follows recursively that $P_k^{-1} \succ 0$ for all $k \geq 0$.

To keep the parameter estimates within the prescribed feasible set, we introduce the following weighted projection onto the Frobenius-norm ball $\{\Theta : \|\Theta\|_F \leq L\}$:

$$\Pi_{P_k^{-1}}\{X\} := \underset{\|\Omega\|_F \leq L}{\arg\min} \|X - \Omega\|_{P_k^{-1}}, \qquad X \in \mathbb{R}^{d \times m}, \tag{23}$$

where $\|Y\|_{P_k^{-1}} := \sqrt{\mathrm{tr}\left(Y^\top P_k^{-1} Y\right)}, Y \in \mathbb{R}^{d \times m}$. Since the feasible set is closed and convex, the weighted projection is nonexpansive: $\|\Pi_{P_k^{-1}}\{X\} - \Pi_{P_k^{-1}}\{Y\}\|_{P_k^{-1}} \leq \|X - Y\|_{P_k^{-1}}, X, Y \in \mathbb{R}^{d \times m}$.

Algorithm 1 Projected Online Newton Learning for Nonlinear Multi-Output Models

Require: Constraint radius L; parameters $\beta_1 > \beta_2 > 0$ and $\gamma > 0$; initial estimate $\hat{\Theta}_0$ with $\|\hat{\Theta}_0\|_F \le L$

Initialize $P_0 \leftarrow I_d$
for $k = 0, 1, 2, \ldots$ **do**
  Observe the current regressor ϕ_k
  Predict $\hat{y}_{k+1} \leftarrow f(\hat{\Theta}_k^\top \phi_k)$
  Compute H_k and h_k using (18)–(19)
  Compute a_k using (20)
  Update P_{k+1} using (21)
  Receive the observation y_{k+1}
  Compute $\widehat{G}_k$ using (17)
  Update $\hat{\Theta}_{k+1}$ using (24)
end for

The parameter estimate is then updated by preconditioning the observation-driven gradient and projecting the result onto the feasible set:

$$\hat{\Theta}_{k+1} = \Pi_{P_{k+1}^{-1}} \left\{ \hat{\Theta}_k - \gamma a_k P_k \widehat{G}_k \right\}, \tag{24}$$

where the initial estimate $\hat{\Theta}_0$ can be any matrix satisfying $\|\hat{\Theta}_0\|_F \le L$, and the positive constants β_1, β_2, and γ are chosen such that $\beta_1 > \beta_2$ and

$$C := \frac{2\gamma\delta}{1 + (\beta_1 - \beta_2) M_L M_f^2 N^2} - \frac{2c_1\gamma^2 M_f^2}{\beta_1 m_f^2 m_L} \left(\frac{1}{4} + \frac{\beta_2}{\beta_1} \right) - \frac{2\beta_2 M_f^2 M_L}{m_L m_f^2} > 0, \tag{25}$$

with the constants being defined in Assumptions 1–6. (25) can always be satisfied: for any fixed $\beta_1 > 0$, setting $\beta_2 = \gamma^2$ and choosing $\gamma > 0$ suitably small ensures both $\beta_1 > \beta_2$ and $C > 0$.

The complete recursive procedure is summarized in Algorithm 1. The matrix gain $\gamma a_k P_k$ incorporates accumulated regressor geometry through P_k and local curvature information through a_k to precondition the observation-based gradient $\hat{G}_k$. This yields a Newton-type update without forming or inverting the full parameter Hessian.

For the subsequent analysis, define

$$\tilde{\Theta}_k = \Theta^* - \hat{\Theta}_k, \tag{26}$$

$$\psi_k = \nabla_x L \left(f(\Theta^{*\top} \phi_k), f(\hat{\Theta}_k^\top \phi_k) \right), \tag{27}$$

$$\mu_k = \frac{\beta_2 h_k}{1 + (\beta_1 - \beta_2) h_k \phi_k^\top P_k \phi_k}. \tag{28}$$

4 Theoretical Guarantees

In this section, we first establish prediction guarantees and then study parameter estimation. The full proofs are provided in Appendix B. Actually, accurate prediction does not require full parameter identifiability. The following theorem establishes almost-sure convergence of the average prediction error to zero without requiring additional excitation conditions.

Theorem 1 (Cumulative prediction error) *Under Assumptions 1–6, the estimates generated by the proposed online Newton algorithm satisfy, as $n \to \infty$,*

$$\sum_{k=0}^{n} J_k(\hat{\Theta}_k) = O\big(1 + \log\det(P_{n+1}^{-1})\big), \qquad \textit{a.s.} \tag{29}$$

Corollary 1 (Average prediction error) *Under Assumptions 1–6, for fixed model dimensions, the proposed online Newton algorithm satisfies, as $n \to \infty$,*

$$\frac{1}{n+1} \sum_{k=0}^{n} J_k(\hat{\Theta}_k) = O\left(\frac{\log n}{n} \right), \qquad \textit{a.s.} \tag{30}$$

Theorem 2 (Parameter estimation) *Under Assumptions 1–6, the parameter estimates generated by the proposed online Newton algorithm satisfy, as $n \to \infty$,*

$$\|\tilde{\Theta}_n\|^2 = O\left(\frac{1+\log\det(P_n^{-1})}{\lambda_{\min}(P_n^{-1})}\right), \qquad a.s. \tag{31}$$

Corollary 2 (Strong consistency) *Suppose that Assumptions 1–6 hold and, in addition,*

$$\lim_{n\to\infty} \frac{1+\log\det(P_n^{-1})}{\lambda_{\min}(P_n^{-1})} = 0, \qquad a.s. \tag{32}$$

Then the parameter estimates are strongly consistent almost surely, i.e., $\lim_{n\to\infty} \hat{\Theta}_n = \Theta^*, \quad a.s.$

Remark 3 *For parameter estimation, condition (32) requires the accumulated information to dominate the logarithmic growth of the information matrix. This condition coincides with the classical excitation condition of Lai and Wei (1982), which is known to be essentially minimal for the strong consistency of least-squares estimates in stochastic linear regression. In particular, it is substantially weaker than the strong persistent excitation condition imposed in Lai and Zhu (1991), which requires the relevant information matrix to grow linearly with time, i.e., $n = O(\lambda_{\min}(P_n^{-1}))$, a.s. For output prediction, the average prediction error converges to zero at a logarithmic-over-linear rate, improving upon existing stochastic-gradient-type convergence guarantees, e.g., Liu et al. (2026).*

5 Experiments

5.1 Synthetic Experiments: Verification of Theoretical Results

Using synthetic data from Example 1, we validate Theorems 1–2 and illustrate the different data requirements for prediction and parameter estimation.

Experimental setup. The observation noise is generated as $e_{k+1} \sim \mathcal{N}(0, \sigma^2 I_m)$ with $\sigma = 0.1$. The total number of online observations is set to $T = 50000$, and all results are averaged over 30 independent Monte Carlo repetitions. Algorithm 1 is initialized with $\widehat{\Theta}_0 = 0$ and $P_0 = I_d$. The projection radius and algorithmic parameters are fixed as $L = 4$, $\beta_1 = 6$, $\beta_2 = 0.01$, and $\gamma = 0.1$. The unknown parameter matrix Θ^* is generated by projecting a random matrix $\Theta_0 \sim 0.5\mathcal{N}(0, I_{d\times m})$ onto the Frobenius-norm ball $\{\Theta : \|\Theta\|_F \le L\}$, where $d = 10$ and $m = 3$.

Here we consider two types of input data sequences. (1). *Insufficient excitation*: the first setting is a rank-deficient case with $\phi_k = e_1$, where only the first parameter direction is excited, and the minimum eigenvalue of the information matrix P_n^{-1} remains bounded. (2). *Non-persistent excitation*: the second setting considers a weak non-PE sequence. We set $\phi_{k,1} = 1/\sqrt{2}$ and $\phi_{k,j} = \xi_{k,j}(k+1)^{-\alpha}/\sqrt{2(d-1)}$ for $j = 2, \ldots, d$, where $\xi_{k,j}$ are independent Rademacher random variables satisfying $\mathbb{P}(\xi_{k,j} = 1) = \mathbb{P}(\xi_{k,j} = -1) = 1/2$, and $\alpha = 0.1$. It can be verified that the PE condition is not satisfied, while the information-growth condition in (32) is satisfied.

Experimental results. We first evaluate the average prediction error $\frac{1}{n}\sum_{k=0}^{n-1} J_k(\widehat{\Theta}_k)$. Figure 1(a) shows the prediction performance under two data excitation conditions, where solid curves show the Monte Carlo mean over 30 repetitions, shaded regions indicate approximate 95% confidence intervals based on the Monte Carlo standard error, and the dashed curve corresponds to the theoretical reference rate $\log n/n$. It can be seen that the average prediction error decreases to zero in both cases. In particular, insufficient excitation does not prevent accurate prediction, even though the complete parameter matrix is not identifiable. This finding is consistent with Theorem 1 and Corollary 1, which establish prediction guarantees without additional excitation conditions. Figure 1(b)-(d) illustrate parameter estimation through the estimation error $\|\widehat{\Theta}_n - \Theta^*\|_F^2$ and the associated information quantities $\lambda_{\min}(P_n^{-1})$ and $(1+\log\det(P_n^{-1}))/\lambda_{\min}(P_n^{-1})$. Under insufficient excitation, the parameter estimates cannot converge to the true values because $\lambda_{\min}(P_n^{-1})$ remains bounded and the information-growth condition in (32) is violated. In contrast, under the non-persistent excitation, $\lambda_{\min}(P_n^{-1})$ increases with n, while $(1+\log\det(P_n^{-1}))/\lambda_{\min}(P_n^{-1})$ converges to zero, leading to consistent parameter estimation as established in Corollary 2.

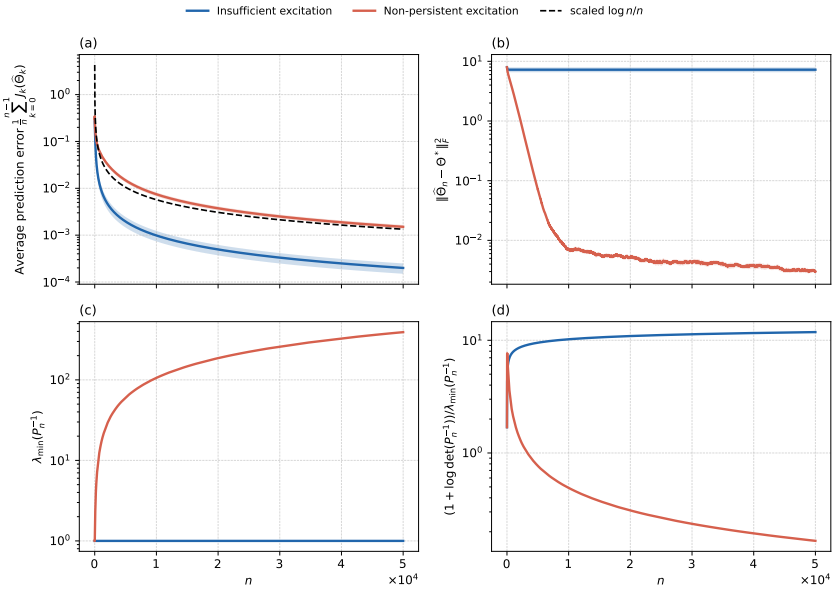


Figure 1: Average prediction and parameter estimation error under different data conditions.

5.2 Real-World Experiments: PAMAP2 Human Activity Classification

We next evaluate Algorithm 1 in the multiclass setting of Example 3 using the PAMAP2 Physical Activity Monitoring dataset (Reiss and Stricker, 2012; Reiss, 2012).

Data and preprocessing. The PAMAP2 dataset contains recordings from nine subjects performing 12 prescribed activities under the official Protocol. We segment each recording into non-overlapping 5-second windows (500 samples) after handling timestamp discontinuities. Missing values are forward-filled using at most 10 preceding samples, and invalid windows are discarded, resulting in 3,768 windows. For each window, we extract 84 statistical features, including channel-wise mean, standard deviation, and root-mean-square values, together with pairwise correlations and vector-magnitude statistics of the six three-axis sensor blocks. Features are standardized online using only past windows and normalized to control input scale. An intercept term is appended, yielding an input dimension of $d = 85$.

Online evaluation protocols. At instant k, each method receives ϕ_k and outputs the probability vector $\widehat{p}_k = (f(\widehat{\Theta}_k^\top \phi_k)^\top, 1 - \mathbf{1}^\top f(\widehat{\Theta}_k^\top \phi_k))^\top$, with prediction $\widehat{c}_{k+1} = \arg\max_{i \in \{1,\dots,12\}} \widehat{p}_{k,i}$. We evaluate two data-ordering protocols: *chronological order*, which preserves the temporal dependence within each subject, and *random order*, where all windows are randomly permuted using 10 fixed seeds. The chronological protocol evaluates learning from the original dependent stream, while the random-order protocol provides a complementary test of order sensitivity. Besides, we report the average cross-entropy loss $\mathrm{CE} = -\frac{1}{n+1}\sum_{k=0}^{n} \log \widehat{p}_{k,c_{k+1}}$, and the cumulative prequential accuracy $\mathrm{Acc} = \frac{1}{n+1}\sum_{k=0}^{n} \mathbf{1}_{\{\widehat{c}_{k+1}=c_{k+1}\}}$. Final accuracy, macro-averaged precision, recall, and F1 score are reported over all 12 classes.

Compared methods and experimental setup. In addition to Algorithm 1, we compare with four online baselines. *Adaptive SG* is the stochastic-gradient algorithm of Liu et al. (2026). *Adam* is the standard bias-corrected adaptive-moment optimization method (Kingma and Ba, 2015). *Diagonal MetaGrad* and *Full-Matrix MetaGrad* are two variants of MetaGrad that aggregate multiple learning-rate experts while maintaining diagonal and full-matrix second-order information, respectively (van Erven et al., 2021). The four methods differ in their parameter updates while sharing the same predictor, features, labels, and initialization $\hat{\Theta}_0 = 0$. Each method processes every time window once in the online stream. Algorithm 1 uses $P_0 = I_{85}$, $\gamma = 2$, $\beta_1 = 30$, $\beta_2 = 20$, and $L = 50$.

Table 1: Prequential performance on all PAMAP2 windows in chronological order.

Method	Acc. (%) ↑	Macro-P (%) ↑	Macro-R (%) ↑	Macro-F1 (%) ↑	CE (nats) ↓
Algorithm 1 (ours)	**92.57**	**93.28**	**90.92**	**91.90**	**0.335**
Adaptive SG	74.07	69.35	68.64	68.59	1.189
Adam	86.54	86.76	84.49	85.27	1.248
Diagonal MetaGrad	91.48	91.93	89.88	90.72	0.386
Full-Matrix MetaGrad	91.48	92.00	89.89	90.76	0.385

Table 2: Prequential performance on all PAMAP2 windows in random order. Entries are mean ± sample standard deviation.

Method	Acc. (%) ↑	Macro-P (%) ↑	Macro-R (%) ↑	Macro-F1 (%) ↑	CE (nats) ↓
Algorithm 1 (ours)	$\mathbf{90.75 \pm 0.46}$	$\mathbf{91.98 \pm 0.41}$	$\mathbf{89.47 \pm 0.42}$	$\mathbf{90.54 \pm 0.39}$	$\mathbf{0.397 \pm 0.020}$
Adaptive SG	78.18 ± 7.23	74.35 ± 5.39	72.32 ± 5.74	72.44 ± 5.63	1.330 ± 0.715
Adam	88.37 ± 0.38	89.15 ± 0.35	87.69 ± 0.37	88.34 ± 0.35	0.462 ± 0.012
Diagonal MetaGrad	85.44 ± 0.88	87.42 ± 0.75	82.16 ± 1.23	83.61 ± 1.36	0.695 ± 0.031
Full-Matrix MetaGrad	85.73 ± 0.48	87.66 ± 0.48	82.52 ± 0.58	84.00 ± 0.60	0.686 ± 0.014

Adaptive SG uses $\mu = 0.99$, $\beta_1^{\mathrm{SG}} = 0.55$, $\beta_2^{\mathrm{SG}} = 0$, and $\beta_3^{\mathrm{SG}} = 1.1$. Adam uses $\eta = 0.1$, $\rho_1 = 0.9$, $\rho_2 = 0.999$, and $\epsilon = 10^{-8}$. For both diagonal and full-matrix MetaGrad, we set the domain radius $R = 24$, the diameter parameter $D = 48$, the gradient-scale parameter $G = 20$, and the horizon $T = 3768$. For each method, we sampled 10 hyperparameter configurations within commonly used ranges and fixed the best-performing configuration for all evaluations.

Experimental results. Table 1 reports the results under the original temporal order. Algorithm 1 achieves the best final performance across all five metrics, attaining 92.57% accuracy and 91.90% macro-F1. Diagonal and Full-Matrix MetaGrad achieve similar performance, both attaining 91.48% accuracy and approximately 90.7% macro-F1. Compared with Adam, Algorithm 1 improves accuracy by 6.03% and macro-F1 by 6.63%, while reducing cross-entropy from 1.248 to 0.335. It also achieves higher classification performance and lower final cross-entropy than Adaptive SG. Table 2 reports the mean and sample standard deviation over 10 independent random permutations. Algorithm 1 achieves the best mean performance across all five metrics, reaching $90.75\% \pm 0.46\%$ accuracy and $90.54\% \pm 0.39\%$ macro-F1. Compared with the chronological setting, the mean accuracies of Diagonal and Full-Matrix MetaGrad decrease by 6.04% and 5.75%, respectively, whereas that of Algorithm 1 changes by only 1.82%, indicating greater stability across different data presentation orders. Overall, Algorithm 1 achieves better performance across all five metrics under both chronological and random-order protocols, demonstrating its effectiveness and stability across different data presentation orders. Additional learning curves and comparisons of computational costs and complexities are provided in Appendix A (Figure 2 and Table 3).

6 Conclusion

In this paper, we develop an online Newton-type learning framework for multi-output nonlinear stochastic regression under non-i.i.d. data and general prediction losses. The results reveal a fundamental distinction between prediction and parameter estimation: prediction consistency requires neither additional data excitation conditions nor parameter identifiability, while strong consistency requires only an essentially minimal information-growth condition. Experiments on synthetic and real-world tasks validate the theoretical results and the effectiveness of the proposed framework. Several directions remain open, including relaxing the current curvature conditions and extending the framework to closed-loop settings, such as adaptive control and embodied learning.

AI use statement

During the preparation of this manuscript, the authors used AI-based tools solely for language editing and polishing. The authors take full responsibility for the content, accuracy, and originality of the manuscript.

REFERENCES

Alekh Agarwal and John C. Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2013. doi: 10.1109/TIT.2012.2212414.

Oren Anava, Elad Hazan, Shie Mannor, and Ohad Shamir. Online learning for time series prediction. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 172–184, 2013.

Karl Johan Åström and Björn Wittenmark. On self tuning regulators. *Automatica*, 9(2):185–199, 1973. doi: 10.1016/0005-1098(73)90073-3.

Bernard Bercu, Antoine Godichon, and Bruno Portier. An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization*, 58(1): 348–367, 2020. doi: 10.1137/19M1261717.

Daniele Calandriello, Alessandro Lazaric, and Michal Valko. Second-order kernel online convex optimization with adaptive sketching. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 645–653, 2017.

Han-Fu Chen. Strong consistency and convergence rate of least squares identification. *Scientia Sinica, Series A*, 25(7):771–784, 1982.

Han-Fu Chen and Lei Guo. *Identification and Stochastic Adaptive Control*. Systems & Control: Foundations & Applications. Birkhäuser, Boston, MA, 1991. doi: 10.1007/978-1-4612-0429-9.

Norbert Christopeit and Kurt Helmes. Strong consistency of least squares estimators in linear regression models. *The Annals of Statistics*, 8(4):778–788, 1980. doi: 10.1214/aos/1176345070.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

Mathieu Even. Stochastic gradient descent under Markovian sampling schemes. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 9412–9439, 2023.

Antoine Godichon-Baggioni and Nicklas Werge. On adaptive stochastic optimization for streaming data: A Newton's method with $O(dN)$ operations. *Journal of Machine Learning Research*, 26 (59):1–49, 2025.

Lei Guo. Convergence and logarithm laws of self-tuning regulators. *Automatica*, 31(3):435–450, 1995. doi: 10.1016/0005-1098(94)00127-5.

Lei Guo and Han-Fu Chen. The Åström-Wittenmark self-tuning regulator revisited and ELS-based adaptive trackers. *IEEE Transactions on Automatic Control*, 36(7):802–812, 1991. doi: 10.1109/9.85060.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018. URL `https://jmlr.org/papers/v19/16-465.html`.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2–3):169–192, 2007. doi: 10.1007/s10994-007-5016-8.

Kwang-Sung Jun, Aniruddha Bhargava, Robert D. Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems*, volume 30, pages 99–109, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/28dd2c7955ce926456240b2ff0100bde-Abstract.html`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

P. R. Kumar. Convergence of adaptive control schemes using least-squares parameter estimates. *IEEE Transactions on Automatic Control*, 35(4):416–424, 1990. doi: 10.1109/9.52293.

Vitaly Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1): 154–166, 1982. doi: 10.1214/aos/1176345697.

Tze Leung Lai and Ching Zong Wei. Extended least squares and their applications to adaptive control and prediction in linear systems. *IEEE Transactions on Automatic Control*, 31(10):898–906, 1986. doi: 10.1109/TAC.1986.1104138.

Tze Leung Lai and Guangrui Zhu. Adaptive prediction in non-linear autoregressive models and control systems. *Statistica Sinica*, 1(2):309–334, 1991. URL https://www3.stat.sinica.edu.tw/statistica/j1n2/j1n21/j1n21.htm.

Yujing Liu, Xin Zheng, Zhixin Liu, and Lei Guo. Gradient-based adaptive prediction and control for nonlinear dynamical systems. *arXiv preprint arXiv:2602.11899*, 2026. doi: 10.48550/arXiv.2602.11899.

Lennart Ljung. Consistency of the least-squares identification method. *IEEE Transactions on Automatic Control*, 21(5):779–781, 1976. doi: 10.1109/TAC.1976.1101344.

Haipeng Luo, Alekh Agarwal, Nicolò Cesa-Bianchi, and John Langford. Efficient second order online learning by sketching. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Shaocong Ma, Ziyi Chen, Yi Zhou, Kaiyi Ji, and Yingbin Liang. Data sampling affects the complexity of online SGD over dependent data. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1296–1305, 2022.

Zakaria Mhammedi and Khashayar Gatmiry. Quasi-Newton steps for efficient online exp-concave optimization. In *Proceedings of the 36th Annual Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4473–4503, 2023.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for non-i.i.d. processes. In *Advances in Neural Information Processing Systems*, volume 20, pages 1025–1032. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/hash/85d8ce590ad8981ca2c8286f79f59954-Abstract.html.

John B. Moore. On strong consistency of least squares identification algorithms. *Automatica*, 14(5): 505–509, 1978. doi: 10.1016/0005-1098(78)90010-9.

Attila Reiss. PAMAP2 physical activity monitoring. UCI Machine Learning Repository, 2012.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*, pages 108–109, 2012. doi: 10.1109/ISWC.2012.13.

Tim van Erven, Wouter M. Koolen, and Dirk van der Hoeven. MetaGrad: Adaptation using multiple learning rates in online learning. *Journal of Machine Learning Research*, 22(161):1–61, 2021.

Fang Wang, Lantian Zhang, and Lei Guo. Applications of nonlinear recursive identification theory in sentencing data analyses. *Scientia Sinica Informationis*, 52(10):1837–1852, 2022. doi: 10.1360/SSI-2022-0325. In Chinese.

Lantian Zhang and Lei Guo. Asymptotically efficient adaptive identification under saturated output observations. *SIAM Journal on Control and Optimization*, 63(4):2338–2368, 2025. doi: 10.1137/24M1665647.

Lantian Zhang and Silun Zhang. Online learning for nonlinear dynamical systems without the I.I.D. condition. *arXiv preprint arXiv:2504.02995*, 2025. doi: 10.48550/arXiv.2504.02995. URL https://arxiv.org/abs/2504.02995.

A Additional PAMAP2 Experimental Results and Computational Analysis

This appendix provides additional results for the PAMAP2 experiments. Figure 2 presents the cumulative prediction performance under chronological and random-order protocols, while Table 3 compares the empirical computational costs and theoretical complexities of the evaluated methods.

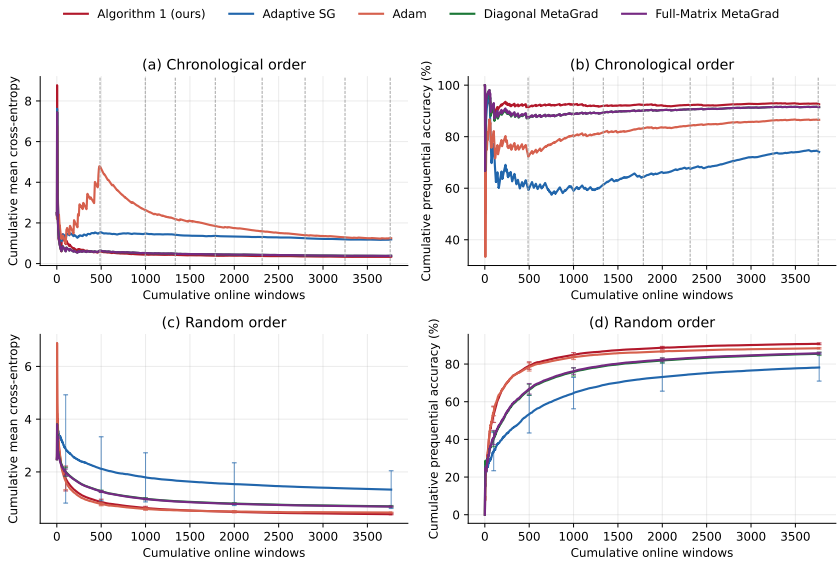


Figure 2: Learning curves of cumulative mean cross-entropy and cumulative prequential accuracy under chronological and random-order evaluation. Random-order results are averaged over 10 random permutations, with error bars indicating one standard deviation at selected window counts.

Table 3: Computational cost and complexity comparison on the PAMAP2 dataset. Time denotes the measured mean update time per window in milliseconds (ms), and State denotes the memory occupied by the stored algorithmic state in mebibytes (MiB). The update complexity reports the leading cost of the regular optimizer update, excluding the common prediction and gradient computation and any conditional weighted projection. The storage complexity accounts for model parameters and persistent optimizer states. Here, d is the feature dimension, m is the number of non-reference softmax outputs, $p = dm$ is the parameter dimension, and $M = 7$ is the number of MetaGrad learning-rate experts.

Method	Time (ms)	State (MiB)	Update complexity	Storage complexity
Algorithm 1 (ours)	1.539	0.062	$O(d^2 + m^3)$	$O(d^2 + p)$
Adaptive SG	0.269	0.007	$O(p + m^2)$	$O(p)$
Adam	0.202	0.021	$O(p)$	$O(p)$
Diagonal MetaGrad	0.652	0.064	$O(Mp)$	$O(Mp)$
Full-Matrix MetaGrad	187.574	46.746	$O(Mp^2)$	$O(Mp^2)$

B Proofs of the Main Results

To establish the performance guarantees presented in Section 4, we first introduce the following lemma and provide its proof.

Lemma 1 *Let Assumptions 1-6 be satisfied for the model (1). Then Algorithm 1 satisfies the following property:*

$$\operatorname{tr}\left\{\tilde{\Theta}_{n+1}^T P_{n+1}^{-1}\tilde{\Theta}_{n+1}\right\}+\sum_{k=0}^{n} J_k(\hat{\Theta}_k)=O(1+\log|P_{n+1}^{-1}|),\quad a.s. \tag{33}$$

Proof. By (20) and (22), we acquire

$$a_k P_{k+1}^{-1}P_k\phi_k=\frac{1}{1+(\beta_1-\beta_2)h_k\phi_k^T P_k\phi_k}\phi_k=:\zeta_k\phi_k. \tag{34}$$

For any given $k\geq 0$, we consider the stochastic Lyapunov function $V_k=\operatorname{tr}\{\tilde{\Theta}_k^T P_k^{-1}\tilde{\Theta}_k\}, k\geq 0$. By Algorithm 1, (22), (34) and the nonexpansive property of projection, we have

$$\begin{aligned}
V_{k+1}=&\|\tilde{\Theta}_{k+1}\|^2_{P_{k+1}^{-1}}\leq\|\tilde{\Theta}_k+\gamma a_k P_k\phi_k(\psi_k+w_{k+1})^T J_f(\hat{\Theta}_k^T\phi_k)\|^2_{P_{k+1}^{-1}}\\
=&V_k+\mu_k\|\tilde{\Theta}_k^T\phi_k\|^2+2\gamma a_k\operatorname{tr}\left\{\tilde{\Theta}_k^T P_{k+1}^{-1}P_k\phi_k[w_{k+1}+\psi_k]^T J_f(\hat{\Theta}_k^T\phi_k)\right\}\\
&+\gamma^2 a_k^2\phi_k^T P_k P_{k+1}^{-1}P_k\phi_k\left\|J_f(\hat{\Theta}_k^T\phi_k)^T(w_{k+1}+\psi_k)\right\|^2\\
=&V_k+\mu_k\|\tilde{\Theta}_k^T\phi_k\|^2+2\gamma\zeta_k\operatorname{tr}\left\{\tilde{\Theta}_k^T\phi_k[w_{k+1}+\psi_k]^T J_f(\hat{\Theta}_k^T\phi_k)\right\}\\
&+\gamma^2 a_k\zeta_k\phi_k^T P_k\phi_k\left\|J_f(\hat{\Theta}_k^T\phi_k)^T(w_{k+1}+\psi_k)\right\|^2\\
=&V_k+\mu_k\|\tilde{\Theta}_k^T\phi_k\|^2+2\gamma\zeta_k\phi_k^T\tilde{\Theta}_k J_f(\hat{\Theta}_k^T\phi_k)^T[w_{k+1}+\psi_k]\\
&+\gamma^2\zeta_k a_k\phi_k^T P_k\phi_k\|J_f(\hat{\Theta}_k^T\phi_k)^T[w_{k+1}+\psi_k]\|^2
\end{aligned} \tag{35}$$

Note that $\nabla_\Theta J_k(\hat{\Theta}_k)=\phi_k(J_f(\hat{\Theta}_k^T\phi_k)^T\psi_k)^T$, by (7) in Assumption 3 we have

$$2\gamma\zeta_k\phi_k^T\tilde{\Theta}_k J_f(\hat{\Theta}_k^T\phi_k)^T\psi_k\leq-\frac{2\gamma\delta}{1+(\beta_1-\beta_2)M_L M_f^2N^2}J_k(\hat{\Theta}_k), \tag{36}$$

where the above inequality holds since $h_k\leq\lambda^2_{\max}\{J_f(\hat{\Theta}_k^T\phi_k)\}\lambda_{\max}\{H_k\}\leq M_f^2M_L$ and $\phi_k^T P_k\phi_k\leq N^2$.

By the fact that $a_k\phi_k^T P_k\phi_k=\frac{a_k^2[\phi_k^T P_k\phi_k+\mu_k(\phi_k^T P_k\phi_k)^2]}{\zeta_k}$, and by (8) in Assumption 3, we also have

$$\begin{aligned}
&2\gamma^2\zeta_k a_k\phi_k^T P_k\phi_k\|J_f(\hat{\Theta}_k^T\phi_k)^T\psi_k\|^2\\
=&2\gamma^2 a_k^2[\phi_k^T P_k\phi_k+\mu_k(\phi_k^T P_k\phi_k)^2]\|J_f(\hat{\Theta}_k^T\phi_k)^T\psi_k\|^2\\
\leq&\frac{2\gamma^2M_f^2c_1}{\beta_1 m_f^2 m_L}\left(\frac{1}{4}+\frac{\beta_2}{\beta_1}\right)J_k(\hat{\Theta}_k)+2c_2\gamma^2\zeta_k M_f^2 a_k\phi_k^T P_k\phi_k,
\end{aligned} \tag{37}$$

where the last inequality holds since $a_k^2[\phi_k^T P_k\phi_k+\mu_k(\phi_k^T P_k\phi_k)^2]\leq\frac{1}{\beta_1 m_f^2 m_L}\left(\frac{1}{4}+\frac{\beta_2}{\beta_1}\right)$.

Moreover, by the mean value theorem we have

$$f(\Theta^{*T}\phi_k)-f(\hat{\Theta}_k^T\phi_k)=\Big(\int_0^1 J_f\big(\hat{\Theta}_k^T\phi_k+t(\Theta^{*T}\phi_k-\hat{\Theta}_k^T\phi_k)\big)\,dt\Big)\tilde{\Theta}_k^T\phi_k:=A_k\tilde{\Theta}_k^T\phi_k, \tag{38}$$

where A_k is symmetric and satisfies $m_f I\preceq A_k$ by Assumption 5. Applying the Taylor expansion, there exists ξ_k between $f(\hat{\Theta}_k^T\phi_k)$ and $f(\Theta^{*T}\phi_k)$ such that

$$\begin{aligned}
J_k(\hat{\Theta}_k)&=\frac{1}{2}[f(\hat{\Theta}_k^T\phi_k)-f(\Theta^{*T}\phi_k)]^T\nabla^2_{xx}L(f(\Theta^{*T}\phi_k),\xi_k)[f(\hat{\Theta}_k^T\phi_k)-f(\Theta^{*T}\phi_k)]\\
&=\frac{1}{2}(A_k\tilde{\Theta}_k^T\phi_k)^T\nabla^2_{xx}L(f(\Theta^{*T}\phi_k),\xi_k)(A_k\tilde{\Theta}_k^T\phi_k)\\
&\geq\frac{m_L m_f^2}{2}\|\tilde{\Theta}_k^T\phi_k\|^2,
\end{aligned} \tag{39}$$

from this and note that $h_k \leq M_f^2 M_L$, we have

$$\mu_k \|\tilde{\Theta}_k^T \phi_k\|^2 \leq \frac{2\beta_2 M_L M_f^2}{m_L m_f^2} J_k(\hat{\Theta}_k). \tag{40}$$

Note that $0 \leq \zeta_k \leq 1, k \geq 0$, substituting (36)-(40) into (35) and summing up both sides of (35) from $k = 0$ to n we know that

$$\begin{aligned}
&V_{n+1} + \sum_{k=0}^{n} \left[\frac{2\gamma\delta}{1 + (\beta_1 - \beta_2) M_L M_f^2 N^2} - \frac{2c_1 \gamma^2 M_f^2}{\beta_1 m_f^2 m_L} \left(\frac{1}{4} + \frac{\beta_2}{\beta_1} \right) - \frac{2\beta_2 M_f^2 M_L}{m_L m_f^2} \right] J_k(\hat{\Theta}_k) \\
\leq & V_0 + 2\gamma \sum_{k=0}^{n} \zeta_k \phi_k^T \tilde{\Theta}_k J_f(\hat{\Theta}_k^T \phi_k)^T w_{k+1} + 2c_2 \gamma^2 M_f^2 \sum_{k=0}^{n} a_k \phi_k^T P_k \phi_k \\
&+ 2\gamma^2 \sum_{k=0}^{n} a_k \phi_k^T P_k \phi_k \|J_f(\hat{\Theta}_k^T \phi_k)^T w_{k+1}\|^2.
\end{aligned} \tag{41}$$

Then, we turn to analyze the right-hand-side (RHS) of (41) term by term. By Lemma 2 in Appendix C, (40) and (39), we know that

$$\begin{aligned}
\sum_{k=0}^{n} \zeta_k \phi_k^T \tilde{\Theta}_k J_f(\hat{\Theta}_k^T \phi_k)^T w_{k+1} &= o\left(\sum_{k=0}^{n} \|\phi_k^T \tilde{\Theta}_k\|^2 \right) + O(1) \\
&= o\left(\sum_{k=0}^{n} J_k(\hat{\Theta}_k) \right) + O(1), \quad a.s.
\end{aligned} \tag{42}$$

For the third term of RHS in (41), by the definition of a_k and μ_k, we have

$$a_k \phi_k^T P_k \phi_k = \frac{\mu_k \phi_k^T P_k \phi_k}{\beta_2 h_k (1 + \mu_k \phi_k^T P_k \phi_k)}, \tag{43}$$

from this, the fact that $\frac{x}{1+x} \leq \log(1+x), x \geq 0$ and $P_{k+1}^{-1} = P_k^{-1} + \mu_k \phi_k \phi_k^T$, we know that

$$\begin{aligned}
\sum_{k=0}^{n} a_k \phi_k^T P_k \phi_k &\leq \frac{1}{\beta_2 m_f^2 m_L} \sum_{k=0}^{n} \log(1 + \mu_k \phi_k^T P_k \phi_k) \\
&= \frac{1}{\beta_2 m_f^2 m_L} \sum_{k=0}^{n} \left\{ \log |P_{k+1}^{-1}| - \log |P_k^{-1}| \right\} \\
&= O\big(\log |P_{n+1}^{-1}|\big),
\end{aligned} \tag{44}$$

where the first inequality holds since $h_k \geq \lambda_{\min}^2 \{J_f(\hat{\Theta}_k^T \phi_k)\} \lambda_{\min}\{H_k\} \geq m_f^2 m_L$.

As for the last term of RHS in (41), by Lemma 2 in Appendix C applied to the centered squared-noise sequence and from (44), we obtain

$$\begin{aligned}
&\sum_{k=0}^{n} a_k \phi_k^T P_k \phi_k \|J_f(\hat{\Theta}_k^T \phi_k)^T w_{k+1}\|^2 \\
&= \sum_{k=0}^{n} a_k \phi_k^T P_k \phi_k \mathbb{E}\Big[\|J_f(\hat{\Theta}_k^T \phi_k)^T w_{k+1}\|^2 \mid \mathcal{F}_k \Big] \\
&\quad + \sum_{k=0}^{n} a_k \phi_k^T P_k \phi_k \left\{ \|J_f(\hat{\Theta}_k^T \phi_k)^T w_{k+1}\|^2 - \mathbb{E}\Big[\|J_f(\hat{\Theta}_k^T \phi_k)^T w_{k+1}\|^2 \mid \mathcal{F}_k \Big] \right\} \\
&= O\left(\sum_{k=0}^{n} a_k \phi_k^T P_k \phi_k \right) + o\left(\sum_{k=0}^{n} a_k \phi_k^T P_k \phi_k \right) + O(1) \\
&= O\big(\log |P_{n+1}^{-1}|\big), \qquad \text{a.s.}
\end{aligned} \tag{45}$$

Combining (42)-(45) with (41) proves Lemma 1.

Proof of Theorems 1-2. Theorems 1-2 can be derived directly by Lemma 1.

C Auxiliary Result and Verification of Examples

Lemma 2 *Chen and Guo (1991) Let $\{X_n, \mathcal{F}_n\}$ be a matrix martingale difference sequence and $\{M_n, \mathcal{F}_n\}$ an adapted sequence of random matrices $\|M_n\| < \infty \quad a.s., \quad \forall n \geq 0$. If*

$$\sup_n E\left[\|X_{n+1}\|^\alpha \mid \mathcal{F}_n\right] < \infty \quad a.s. \tag{46}$$

for some $\alpha \in (0, 2]$, then as $n \to \infty$

$$\sum_{i=0}^{n} M_i X_{i+1} = O\left(s_n(\alpha) \log^{\frac{1}{\alpha}+\eta}\left(s_n^\alpha(\alpha) + e\right)\right) \ a.s., \forall \eta > 0, \tag{47}$$

where

$$s_n(\alpha) = \left(\sum_{i=0}^{n} \|M_i\|^\alpha\right)^{\frac{1}{\alpha}} \tag{48}$$

Condition Verification. We now verify the required assumptions for three examples in Section 2.2.

Example 1. We first verify Assumption 1. For every $z \in \mathbb{R}^m$, since $\mathbb{E}[e_{k+1} \mid \mathcal{F}_k] = 0$,

$$\mathbb{E}\left[\frac{1}{2}\|x + e_{k+1} - z\|^2 \mid \mathcal{F}_k\right] = \frac{1}{2}\|x - z\|^2 + \frac{1}{2}\mathbb{E}\left[\|e_{k+1}\|^2 \mid \mathcal{F}_k\right].$$

Hence the conditional risk has the unique deterministic minimizer $f(x) = x$. For the quadratic loss, $\nabla_x L(y, x) = x - y$. Therefore, by the definition of the gradient noise,

$$w_{k+1} = \nabla_x L(y_{k+1}, \widehat{\Theta}_k^T \phi_k) - \nabla_x L(\Theta^{*T} \phi_k, \widehat{\Theta}_k^T \phi_k) = -e_{k+1}.$$

Thus Assumption 2 holds. Since $\nabla_\Theta J_k(\Theta) = \phi_k\big((\Theta - \Theta^*)^T \phi_k\big)^T$, we have

$$\operatorname{tr}\left(\nabla_\Theta J_k(\Theta)^T (\Theta - \Theta^*)\right) = \|(\Theta - \Theta^*)^T \phi_k\|^2 = 2J_k(\Theta),$$

which verifies Assumption 3 with $\delta = 2$. Moreover,

$$\nabla_x L\left(f(\Theta^{*T}\phi_k), f(\Theta^T \phi_k)\right) = (\Theta - \Theta^*)^T \phi_k,$$

and hence

$$\left\|\nabla_x L\left(f(\Theta^{*T}\phi_k), f(\Theta^T \phi_k)\right)\right\|^2 = 2J_k(\Theta).$$

Thus Assumption 3 holds with $c_1 = 2$ and $c_2 = 0$.

The Hessian of the quadratic loss with respect to its second argument is

$$\nabla_{xx}^2 L(s, x) = I_m.$$

Hence Assumption 6 holds. Finally, $f(x) = x$ has Jacobian

$$J_f(x) = I_m,$$

and thus Assumption 5 holds.

Example 2. We first verify Assumption 1. Recall that $G(x) = \mathbb{E}[S(x + e_{k+1}) \mid \mathcal{F}_k]$, for every $z \in \mathbb{R}^m$,

$$\mathbb{E}\left[\frac{1}{2}\|S(x + e_{k+1}) - z\|^2 \mid \mathcal{F}_k\right] = \frac{1}{2}\|G(x) - z\|^2 + \frac{1}{2}\mathbb{E}\left[\|S(x + e_{k+1}) - G(x)\|^2 \mid \mathcal{F}_k\right].$$

Therefore, the unique deterministic minimizer is $f(x) = G(x)$, and Assumption 1 holds.

For the quadratic loss,

$$w_{k+1} = G(\Theta^{*T}\phi_k) - y_{k+1}.$$

Since e_{k+1} is independent of $\mathcal{F}_k$ and $\Theta^{*T}\phi_k$ is $\mathcal{F}_k$-measurable, $G(\Theta^{*T}\phi_k) = \mathbb{E}[y_{k+1} \mid \mathcal{F}_k]$. It follows that Assumption 2 holds. For every $t \in \mathbb{R}$,

$$G_i'(t) = F_i(U_i - t) - F_i(L_i - t) > 0.$$

Since G_i' is continuous, there exist constants $0 < m_f \le M_f < \infty$ such that

$$m_f \le G_i'(t) \le M_f, \qquad \forall |t| \le NL, \quad i = 1, \dots, m.$$

The Jacobian of G is

$$J_G(x) = \operatorname{diag}\left(G_1'(x_1), \dots, G_m'(x_m)\right),$$

which is symmetric and satisfies

$$m_f I_m \preceq J_G(x) \preceq M_f I_m, \qquad \|x\| \le NL.$$

Thus Assumption 5 holds.

Next, set

$$x_k = \Theta^T \phi_k, \qquad x_k^* = \Theta^{*T}\phi_k, \qquad u_k = x_k - x_k^* = (\Theta - \Theta^*)^T \phi_k.$$

The mean-value theorem gives

$$G(x_k) - G(x_k^*) = D_k u_k,$$

where

$$D_k = \operatorname{diag}\left(G_1'(\eta_{k,1}), \dots, G_m'(\eta_{k,m})\right),$$

and each $\eta_{k,i}$ lies between $x_{k,i}$ and $x_{k,i}^*$. Hence

$$m_f I_m \preceq D_k \preceq M_f I_m.$$

Note that

$$\nabla_\Theta J_k(\Theta) = \phi_k \left[J_G(x_k)^T \big(G(x_k) - G(x_k^*)\big)\right]^T,$$

so we have

$$\begin{aligned} \operatorname{tr}\left(\nabla_\Theta J_k(\Theta)^T (\Theta - \Theta^*)\right) &= \big(G(x_k) - G(x_k^*)\big)^T J_G(x_k) u_k \\ &= u_k^T D_k J_G(x_k) u_k \\ &\ge m_f^2 \|u_k\|^2. \end{aligned}$$

On the other hand,

$$\begin{aligned} 2J_k(\Theta) &= \|G(x_k) - G(x_k^*)\|^2 \\ &= \|D_k u_k\|^2 \le M_f^2 \|u_k\|^2. \end{aligned}$$

Therefore,

$$\operatorname{tr}\left(\nabla_\Theta J_k(\Theta)^T (\Theta - \Theta^*)\right) \ge \frac{2m_f^2}{M_f^2} J_k(\Theta).$$

Thus Assumption 3 holds with $\delta = \frac{2m_f^2}{M_f^2}$.

Furthermore,

$$\nabla_x L\left(f(\Theta^{*T}\phi_k), f(\Theta^T\phi_k)\right) = G(\Theta^T\phi_k) - G(\Theta^{*T}\phi_k),$$

and hence

$$\left\|\nabla_x L\left(f(\Theta^{*T}\phi_k), f(\Theta^T\phi_k)\right)\right\|^2 = 2J_k(\Theta).$$

Thus Assumption 3 holds with $c_1 = 2$ and $c_2 = 0$.

Finally,

$$\nabla_{xx}^2 L(s, x) = I_m.$$

Hence Assumption 6 holds.

Example 3. Let $p = f(x)$ and define $p_{m+1} = 1 - \sum_{i=1}^{m} p_i$. For every $z \in \mathcal{Z}$, set $z_{m+1} = 1 - \sum_{i=1}^{m} z_i$, so we have

$$\mathbb{E}[L(y_{k+1}, z) \mid \mathcal{F}_k] = -\sum_{i=1}^{m+1} p_i \log z_i = L(p, z) + H(p).$$

Since $H(p)$ is independent of z, Gibbs' inequality shows that the conditional risk is uniquely minimized at $z = p = f(x)$, hence Assumption 1 holds.

On the compact set $\|x\| \le NL$, there exists

$$\varepsilon = \frac{e^{-NL}}{1 + me^{NL}} > 0$$

such that

$$f^{(i)}(x) \ge \varepsilon, \qquad i = 1, \dots, m+1.$$

Let $\widehat{q}_k = f(\widehat{\Theta}_k^T \phi_k)$ and set $\widehat{q}_{k,m+1} = 1 - \sum_{i=1}^{m} \widehat{q}_{k,i}$ and $y_{k+1,m+1} = 1 - \sum_{i=1}^{m} y_{k+1,i}$. For the reduced KL loss, the i-th component of the gradient noise is

$$\begin{aligned} w_{k+1,i} &= -\frac{y_{k+1,i} - p_{k,i}}{\widehat{q}_{k,i}} + \frac{y_{k+1,m+1} - p_{k,m+1}}{\widehat{q}_{k,m+1}} \\ &= -\frac{e_{k+1,i}}{\widehat{q}_{k,i}} - \frac{\sum_{j=1}^{m} e_{k+1,j}}{\widehat{q}_{k,m+1}}, \qquad i = 1, \dots, m. \end{aligned}$$

Since $\mathbb{E}[e_{k+1} \mid \mathcal{F}_k] = 0$, we have $\mathbb{E}[w_{k+1} \mid \mathcal{F}_k] = 0$. Moreover, e_{k+1} is bounded and all components of $\widehat{q}_k$, including $\widehat{q}_{k,m+1}$, are bounded below by ε. Hence $\|w_{k+1}\|$ is uniformly bounded, thus Assumption 2 holds. By Gibbs' inequality, the KL divergence satisfies

$$J_k(\Theta) \ge 0, \qquad J_k(\Theta) = 0 \Longleftrightarrow p_k = q_k.$$

For $t \in [0, 1]$, define

$$\xi_k(t) = (1 - t)p_k + tq_k, \qquad \xi_{k,m+1}(t) = 1 - \sum_{i=1}^{m} \xi_{k,i}(t).$$

Then every component of $\xi_k(t)$, including $\xi_{k,m+1}(t)$, is bounded below by ε. Moreover,

$$\nabla^2_{\xi_k \xi_k} L(p_k, \xi_k(t)) = \operatorname{diag}\left(\frac{p_{k,1}}{\xi_{k,1}(t)^2}, \dots, \frac{p_{k,m}}{\xi_{k,m}(t)^2}\right) + \frac{p_{k,m+1}}{\xi_{k,m+1}(t)^2} \mathbf{1}\mathbf{1}^T.$$

Consequently, there exist constants $0 < m_L \le M_L < \infty$ such that

$$m_L I_m \preceq \nabla^2_{\xi_k \xi_k} L(p_k, \xi_k(t)) \preceq M_L I_m, \qquad t \in [0, 1].$$

Thus Assumption 6 holds.

Next, the Jacobian of the reduced softmax predictor is

$$J_f(x) = \operatorname{diag}(f(x)) - f(x) f^T(x).$$

For any $z \in \mathbb{R}^m$,

$$\begin{aligned} z^T J_f(x) z &= \sum_{i=1}^{m} f^{(i)}(x) z_i^2 - \left(\sum_{i=1}^{m} f^{(i)}(x) z_i\right)^2 \\ &\ge f^{(m+1)}(x) \sum_{i=1}^{m} f^{(i)}(x) z_i^2 \ge \varepsilon^2 \|z\|^2. \end{aligned}$$

Also,

$$z^T J_f(x) z \le \sum_{i=1}^{m} f^{(i)}(x) z_i^2 \le \|z\|^2.$$

Hence there exist constants $0 < m_f \le M_f$ such that

$$m_f I_m \preceq J_f(x) \preceq M_f I_m, \qquad \|x\| \le NL.$$

Thus Assumption 5 holds.

It remains to verify Assumption 3. First, for the KL loss,

$$\nabla_{q_k} L(p_k, q_k) = \left(-\frac{p_{k,1}}{q_{k,1}} + \frac{p_{k,m+1}}{q_{k,m+1}}, \dots, -\frac{p_{k,m}}{q_{k,m}} + \frac{p_{k,m+1}}{q_{k,m+1}} \right)^T.$$

For the reduced softmax map, a direct calculation gives

$$J_f(x_k) \nabla_{q_k} L(p_k, q_k) = q_k - p_k.$$

Therefore, by the chain rule,

$$\nabla_\Theta J_k(\Theta) = \phi_k (q_k - p_k)^T.$$

Let $x_k = \Theta^T \phi_k, x_k^* = \Theta^{*T} \phi_k$, and let

$$u_k = x_k - x_k^* = (\Theta - \Theta^*)^T \phi_k.$$

By the integral mean-value formula,

$$q_k - p_k = \overline{J}_k u_k, \qquad \overline{J}_k = \int_0^1 J_f\big(x_k^* + t(x_k - x_k^*)\big)\, dt.$$

Since $m_f I_m \preceq \overline{J}_k \preceq M_f I_m$, we have

$$\begin{aligned} \operatorname{tr}\left(\nabla_\Theta J_k(\Theta)^T (\Theta - \Theta^*)\right) &= (q_k - p_k)^T u_k \\ &= (q_k - p_k)^T \overline{J}_k^{-1} (q_k - p_k) \\ &\geq \frac{1}{M_f} \|q_k - p_k\|^2. \end{aligned}$$

On the other hand, since $L(p_k, q)$ is uniformly smooth with respect to q on the feasible probability region and $L(p_k, p_k) = 0$, $\nabla_q L(p_k, p_k) = 0$, we have

$$J_k(\Theta) = L(p_k, q_k) \leq \frac{M_L}{2} \|q_k - p_k\|^2.$$

Consequently,

$$\operatorname{tr}\left(\nabla_\Theta J_k(\Theta)^T (\Theta - \Theta^*)\right) \geq \frac{2}{M_f M_L} J_k(\Theta),$$

which verifies Assumption 3 with $\delta = \frac{2}{M_f M_L}$.

Finally, by the uniform smoothness of $L(p_k, q)$ with respect to q,

$$\|\nabla_{q_k} L(p_k, q_k)\| \leq M_L \|q_k - p_k\|.$$

By the uniform strong convexity of $L(p_k, q)$ with respect to q,

$$J_k(\Theta) = L(p_k, q_k) \geq \frac{m_L}{2} \|q_k - p_k\|^2.$$

Thus

$$\|\nabla_{q_k} L(p_k, q_k)\|^2 \leq \frac{2M_L^2}{m_L} J_k(\Theta).$$

Therefore Assumption 3 holds with $c_1 = \frac{2M_L^2}{m_L}$ and $c_2 = 0$.