ADABIO: STOCHASTIC BILEVEL OPTIMIZATION FOR NONCONVEX PERFORMATIVE RISK

Anonymous authors
Paper under double-blind review

ABSTRACT

While stochastic bilevel optimization provides a framework for nested learning problems, most existing methods assume that the data distributions at both levels remain fixed throughout the optimization process. This assumption can fail in performative settings, where the deployed decisions change the environment and thereby affect future data distributions. We study stochastic bilevel optimization for nonconvex performative risk when both the upper-level (UL) and lower-level (LL) distributions are decision-dependent and unknown. Unlike existing bilevel performative methods that primarily target performatively stable points, our goal is to directly optimize the performative risk and find its stationary point. We propose AdaBiO, which iteratively estimates the two unknown distribution maps using explored samples. We prove that AdaBiO converges to an ϵ-stationary point with $\mathcal{O}(\epsilon^{-2})$ computational complexity and $\mathcal{O}(\epsilon^{-1})$ sample complexity. Experiments on strategic credit classification demonstrate that AdaBiO achieves lower performative risk than both non-performative stocBiO and a bilevel performatively stable point baseline. Moreover, its performance approaches that of AdaBiO with known distribution maps.

1 INTRODUCTION

Bilevel optimization provides a framework for learning problems with nested objectives, where an upper-level (UL) decision is optimized through the solution of a lower-level (LL) learning problem. Such formulations arise in hyperparameter optimization, meta-learning, data reweighting, and many other machine learning applications. In stochastic bilevel optimization, the objectives at both levels are typically formulated as expectations over random data, and considerable progress has been made in developing efficient stochastic hypergradient methods with provable convergence guarantees (Ghadimi & Wang, 2018; Hong et al., 2020; Ji et al., 2021). A common assumption underlying these methods is that the sampling distributions remain fixed throughout the optimization process.

A motivating application where this assumption can fail is strategic credit classification. A lender uses a credit scoring model to assess whether a borrower is likely to become seriously delinquent. Once a scoring policy is deployed, borrowers may respond by adjusting actionable financial characteristics, such as credit utilization, debt ratio, or other attributes that affect their credit profiles. Consequently, the deployed model can influence borrowers' behavior and thereby change the distribution of future data observed by the learning system. In a bilevel setting, such feedback can occur at both levels. For example, the lender updates the scoring model at the UL, while borrowers respond at the LL by changing actionable financial features. These responses then change the data used to evaluate the scoring model. Similar bilevel feedback can also arise in online advertising, financial markets, and traffic management, where deployed models affect user, market, or driver behavior and thereby change future data distributions.

This deployment induced feedback is captured by performative prediction, in which the data distribution depends on the deployed decision (Perdomo et al., 2020; Mendler-Dünner et al., 2020). In such settings, optimizing a loss under a fixed distribution may not minimize the risk actually encountered after deployment. Existing performative prediction studies have therefore investigated how learning dynamics interact with decision-dependent distributions. An important direction of work focuses on computing performatively stable points, which characterize fixed points of repeated deployment and retraining (Perdomo et al., 2020; Lu, 2023). However, stability is a different objective from directly

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.

minimizing the performative risk. In the single-level setting, direct performative risk optimization has been studied under structural conditions that can make the resulting objective tractable (Miller et al., 2021). Limited work has studied when the performative risk has a bilevel structure, the UL objective is potentially nonconvex, and the decision-dependent distributions are unknown.

Motivated by this gap, we study the following stochastic bilevel performative risk minimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^{d_x}} \Phi(\boldsymbol{x}) = \mathop{\mathbb{E}}_{\boldsymbol{\xi}\sim\mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}))} f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\xi}) \tag{1a}$$

$$\text{s.t.} \quad \boldsymbol{y}^*(\boldsymbol{x}) \in \arg\min_{\boldsymbol{y}\in\mathbb{R}^{d_y}} \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x})} g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\zeta}) \tag{1b}$$

where $\boldsymbol{x}$ denotes UL model parameters, and $\boldsymbol{y}$ represents LL parameters. For a given $\boldsymbol{x}$, the LL problem returns the optimal response $\boldsymbol{y}^*(\boldsymbol{x})$. $\mathcal{C}$ and $\mathcal{D}$ are UL and LL data distribution mappings, respectively. $f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\xi})$ and $g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\zeta})$ are the UL and LL loss functions, respectively. Let $\boldsymbol{\zeta} \in \mathbb{R}^{d_\zeta}$ and $\boldsymbol{\xi} \in \mathbb{R}^{d_\xi}$ be the UL and LL data samples. We assume that the LL function is strongly convex, a common assumption in bilevel optimization (Ji et al., 2021; Lu, 2023), while the UL objective function is nonconvex. Our goal is to directly optimize the nonconvex performative risk $\Phi(\boldsymbol{x})$ and find an ϵ-stationary point of it. In contrast, prior bilevel performative prediction work (Lu, 2023) mainly targets performatively stable points. As shown in Miller et al. (2021), a stable point only reflects the outcome of repeated retraining and does not guarantee low performative risk. Directly optimizing the performative risk is therefore more closely aligned with minimizing the loss incurred after deployment.

This problem is more challenging than existing works for the following reasons. First, unlike standard bilevel optimization with fixed data distributions (Ghadimi & Wang, 2018; Hong et al., 2020; Ji et al., 2021), both levels in our problem are performative. The LL distribution map depends on $\boldsymbol{x}$, while the UL distribution map depends on the implicit response $\boldsymbol{y}^*(\boldsymbol{x})$. Therefore, differentiating the bilevel objective requires additional distributional derivative terms at both levels, beyond the usual implicit differentiation through the LL problem. Second, unlike single-level performative risk optimization (Miller et al., 2021; Jin et al., 2024), the two distributional effects are coupled through the LL solution. Changes in the LL distribution affect $\boldsymbol{y}^*(\boldsymbol{x})$, which in turn changes the UL distribution, making the performative hypergradient depend on both levels simultaneously. Third, unlike methods that target performatively stable points (Lu, 2023; Jin et al., 2024), which can work with the currently deployed distributions without estimating their mappings, finding a stationary point of the performative risk requires knowing how these distributions change with the models. Since both distribution maps are unknown in our setting, they must be estimated while the bilevel variables are optimized. Finally, these map estimation errors affect both the LL updates and the hypergradient, and must be analyzed together with other approximation errors.

1.1 Contributions

This paper develops an online stochastic bilevel framework for directly optimizing nonconvex performative risk under unknown decision-dependent distributions. By combining online distribution map estimation with performative hypergradient estimation, our method extends stochastic bilevel optimization to a setting where model deployment affects the data distributions at both levels. Our main contributions are summarized as follows.

- **AdaBiO algorithm.** We formulate stochastic bilevel performative risk minimization in which the LL distribution depends on the UL decision and the UL distribution depends on the implicit LL response. Under location-family distribution maps, we derive the corresponding performative hypergradient, which contains additional distributional derivative terms absent from standard bilevel optimization. We then propose AdaBiO, which estimates both unknown distribution maps from collected deployment samples and combines these estimates with warm-started stochastic LL updates and a stochastic Neumann hypergradient estimator.
- **Convergence and complexity analysis.** We establish error bounds for the online estimates of both distribution maps and show how these estimation errors affect the stochastic performative hypergradient and the recursion for the LL tracking error. Building on these results, we prove that AdaBiO converges to an ϵ-stationary point of the nonconvex performative risk. The resulting stochastic gradient and Hessian/Jacobian-vector product complexities retain the same ϵ^{-2} dependence as

Table 1: Comparison of stochastic bilevel optimization algorithms.

Algorithm	$\mathrm{Gc}(f,\epsilon)$	$\mathrm{Gc}(g,\epsilon)$	$\mathrm{JV}(g,\epsilon)$	$\mathrm{HV}(g,\epsilon)$	Bilevel PP
TTSA (Hong et al., 2020)	$\mathcal{O}(\mathrm{poly}(\kappa)\epsilon^{-5/2})$	$\mathcal{O}(\mathrm{poly}(\kappa)\epsilon^{-5/2})$	$\mathcal{O}(\mathrm{poly}(\kappa)\epsilon^{-5/2})$	$\mathcal{O}(\mathrm{poly}(\kappa)\epsilon^{-5/2})$	$\times$
BSA (Ghadimi & Wang, 2018)	$\mathcal{O}(\kappa^6\epsilon^{-2})$	$\mathcal{O}(\kappa^9\epsilon^{-3})$	$\mathcal{O}(\kappa^6\epsilon^{-2})$	$\widetilde{\mathcal{O}}(\kappa^6\epsilon^{-2})$	$\times$
stocBiO (Ji et al., 2021)	$\mathcal{O}(\kappa^5\epsilon^{-2})$	$\mathcal{O}(\kappa^9\epsilon^{-2})$	$\mathcal{O}(\kappa^5\epsilon^{-2})$	$\widetilde{\mathcal{O}}(\kappa^6\epsilon^{-2})$	$\times$
AdaBiO (this paper)	$\mathcal{O}(\kappa^5\epsilon^{-2})$	$\mathcal{O}(\kappa^9\epsilon^{-2})$	$\mathcal{O}(\kappa^5\epsilon^{-2})$	$\widetilde{\mathcal{O}}(\kappa^7\epsilon^{-2})$	✓

$\mathrm{Gc}(f,\epsilon)$ and $\mathrm{Gc}(g,\epsilon)$ denote the numbers of stochastic gradient evaluations with respect to f and g, respectively. $\mathrm{JV}(g,\epsilon)$ denotes the number of Jacobian-vector products associated with the LL cross derivative, and $\mathrm{HV}(g,\epsilon)$ denotes the number of Hessian-vector products involving $\nabla^2_{yy}g$. Here, $\kappa = L/\gamma_y$ denotes the condition number, and $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors in $1/\epsilon$. For TTSA, $\mathrm{poly}(\kappa)$ is used since the explicit dependence on κ is not provided in (Hong et al., 2020). The last column indicates whether the method is designed for bilevel performative prediction.

stocBiO, while the additional sample complexity for learning the unknown distribution maps is explicitly characterized in terms of the optimization trajectory.

- **Application to strategic learning.** We evaluate AdaBiO on strategic credit classification with decision-dependent LL and UL feature distributions. The simulation results show that AdaBiO achieves lower training and test performative losses than standard stocBiO and Bi-SGD baselines, while attaining performance close to that of AdaBiO with known distribution maps.

2 Proposed AdaBiO Algorithm

In this section, we propose the AdaBiO algorithm for bilevel performative risk minimization. The algorithm has two main components. Algorithm 1 estimates the unknown location-family distribution maps from collected deployment samples. Algorithm 2 then combines these estimated distribution maps with warm-started LL updates, a stochastic Neumann inverse-Hessian estimator, and a UL stochastic gradient update to optimize the nonconvex performative risk.

2.1 Location-Family Distribution Maps and Estimation

Since the deployment distributions are unknown, directly optimizing problem (1) requires information about how $\mathcal{D}(\boldsymbol{x})$ and $\mathcal{C}(\boldsymbol{y})$ vary with the deployed decision variables. Without additional structure, estimating a general map to probability distributions from finite deployment samples is not tractable. To address this difficulty, we adopt the same parametric location-family model in Miller et al. (2021); Yan & Cao (2023), which can be viewed as a first-order approximation of the performative effect: the distribution shift depends linearly on the deployed decision and higher-order effects are ignored for simplicity. The LL and UL distribution maps $\mathcal{D}(\cdot)$ and $\mathcal{C}(\cdot)$ are parameterized as $\boldsymbol{\zeta} \sim \mathcal{D}(\boldsymbol{x}) : \boldsymbol{\zeta} = \boldsymbol{\zeta}^{(0)} + \boldsymbol{A}^{(d)}\boldsymbol{x}$, $\boldsymbol{\xi} \sim \mathcal{C}(\boldsymbol{y}) : \boldsymbol{\xi} = \boldsymbol{\xi}^{(0)} + \boldsymbol{A}^{(c)}\boldsymbol{y}$, where $\boldsymbol{\zeta}^{(0)} \sim \mathcal{D}^{(0)}$, $\boldsymbol{\xi}^{(0)} \sim \mathcal{C}^{(0)}$ are unknown zero-mean subgaussian random vectors and $\boldsymbol{A}^{(d)}$, $\boldsymbol{A}^{(c)}$ are unknown matrices.

We estimate the two location matrices using stochastic least squares. Let $S_{d,r}$ and $S_{c,r}$ denote the cumulative numbers of LL and UL map estimation updates completed by the end of outer iteration r, with $S_{d,-1} = S_{c,-1} = 0$ and $S_{d,r} \geq S_{d,r-1} + 1$, $S_{c,r} \geq S_{c,r-1} + 1$. Thus, previous map estimates are preserved across outer iterations, and only the additional updates required to reach the current cumulative targets are performed. The choices of $S_{d,r}$ and $S_{c,r}$ are specified in Appendix G.

At outer iteration r, initialize $\widehat{\boldsymbol{A}}^{(d)}_{r,0} = \widehat{\boldsymbol{A}}^{(d)}_{r-1}$. For $j = 1, \ldots, S_{d,r} - S_{d,r-1}$, draw $\boldsymbol{x}^{\mathrm{exp}}_{r,j} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_x})$ and independently observe $\boldsymbol{\zeta}^{(1)}_{r,j} \sim \mathcal{D}(\boldsymbol{x}_r)$, $\boldsymbol{\zeta}^{(2)}_{r,j} \sim \mathcal{D}(\boldsymbol{x}_r + \boldsymbol{x}^{\mathrm{exp}}_{r,j})$. Since $\boldsymbol{\zeta}^{(2)}_{r,j} - \boldsymbol{\zeta}^{(1)}_{r,j} = \boldsymbol{A}^{(d)}\boldsymbol{x}^{\mathrm{exp}}_{r,j}$, the LL map estimate is updated as

$$\widehat{\boldsymbol{A}}^{(d)}_{r,j} = \widehat{\boldsymbol{A}}^{(d)}_{r,j-1} + \frac{2}{S_{d,r-1} + j + 8d_x}\Big(\boldsymbol{\zeta}^{(2)}_{r,j} - \boldsymbol{\zeta}^{(1)}_{r,j} - \widehat{\boldsymbol{A}}^{(d)}_{r,j-1}\boldsymbol{x}^{\mathrm{exp}}_{r,j}\Big)(\boldsymbol{x}^{\mathrm{exp}}_{r,j})^{\intercal}. \tag{2}$$

After updates, let $\widehat{\boldsymbol{A}}^{(d)}_{r} = \widehat{\boldsymbol{A}}^{(d)}_{r,S_{d,r}-S_{d,r-1}}$, which defines $\widehat{\mathcal{D}}(\boldsymbol{x}_r)$ through $\boldsymbol{\zeta}_r = \boldsymbol{\zeta}^{(0)}_r + \widehat{\boldsymbol{A}}^{(d)}_r\boldsymbol{x}_r$. Similarly, after obtaining $\boldsymbol{y}_{r,K}$, initialize $\widehat{\boldsymbol{A}}^{(c)}_{r,0} = \widehat{\boldsymbol{A}}^{(c)}_{r-1}$ and, for $j = 1, \ldots, S_{c,r} - S_{c,r-1}$, draw

Algorithm 1 Online Distribution Map Estimation for Location Families

Require: $\widehat{\boldsymbol{A}}^{(d)}_{r-1},\ \boldsymbol{x}_r,\ (S_{d,r-1}, S_{d,r})$ or $\widehat{\boldsymbol{A}}^{(c)}_{r-1},\ \boldsymbol{y}_{r,K},\ (S_{c,r-1}, S_{c,r})$

1: Initialize $\widehat{\boldsymbol{A}}^{(d)}_{r,0} = \widehat{\boldsymbol{A}}^{(d)}_{r-1}$ or $\widehat{\boldsymbol{A}}^{(c)}_{r,0} = \widehat{\boldsymbol{A}}^{(c)}_{r-1}$

2: **for** $j = 1, \ldots,$ $S_{d,r} - S_{d,r-1}$ or $S_{c,r} - S_{c,r-1}$ **do**

3: Draw $\boldsymbol{x}^{\exp}_{r,j} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_x})$ or $\boldsymbol{y}^{\exp}_{r,j} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_y})$, and observe

$\boldsymbol{\zeta}^{(1)}_{r,j} \sim \mathcal{D}(\boldsymbol{x}_r),\ \boldsymbol{\zeta}^{(2)}_{r,j} \sim \mathcal{D}(\boldsymbol{x}_r + \boldsymbol{x}^{\exp}_{r,j})$ or $\boldsymbol{\xi}^{(1)}_{r,j} \sim \mathcal{C}(\boldsymbol{y}_{r,K}),\ \boldsymbol{\xi}^{(2)}_{r,j} \sim \mathcal{C}(\boldsymbol{y}_{r,K} + \boldsymbol{y}^{\exp}_{r,j})$

4: Update $\widehat{\boldsymbol{A}}^{(d)}_{r,j}$ via Eq. (2) or $\widehat{\boldsymbol{A}}^{(c)}_{r,j}$ via Eq. (3)

5: **end for**

6: Update $\widehat{\boldsymbol{A}}^{(d)}_{r} = \widehat{\boldsymbol{A}}^{(d)}_{r,S_{d,r}-S_{d,r-1}}$ or $\widehat{\boldsymbol{A}}^{(c)}_{r} = \widehat{\boldsymbol{A}}^{(c)}_{r,S_{c,r}-S_{c,r-1}}$

$\boldsymbol{y}^{\exp}_{r,j} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_y})$ and observe $\boldsymbol{\xi}^{(1)}_{r,j} \sim \mathcal{C}(\boldsymbol{y}_{r,K})$ and $\boldsymbol{\xi}^{(2)}_{r,j} \sim \mathcal{C}(\boldsymbol{y}_{r,K} + \boldsymbol{y}^{\exp}_{r,j})$. The corresponding UL update is

$$\widehat{\boldsymbol{A}}^{(c)}_{r,j} = \widehat{\boldsymbol{A}}^{(c)}_{r,j-1} + \frac{2}{S_{c,r-1} + j + 8d_y}\left(\boldsymbol{\xi}^{(2)}_{r,j} - \boldsymbol{\xi}^{(1)}_{r,j} - \widehat{\boldsymbol{A}}^{(c)}_{r,j-1}\boldsymbol{y}^{\exp}_{r,j}\right)(\boldsymbol{y}^{\exp}_{r,j})^{\intercal}. \tag{3}$$

We then set $\widehat{\boldsymbol{A}}^{(c)}_{r} = \widehat{\boldsymbol{A}}^{(c)}_{r,S_{c,r}-S_{c,r-1}}$, which defines $\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})$ through $\boldsymbol{\xi}_r = \boldsymbol{\xi}^{(0)}_r + \widehat{\boldsymbol{A}}^{(c)}_{r}\boldsymbol{y}_{r,K}$.

Each stochastic least-squares update uses two newly collected deployment observations. The map estimation procedure is summarized in Algorithm 1.

2.2 Performative Hypergradient Estimation

The key difference from the standard stocBiO hypergradient is that both levels contain distributional feedback. In standard stochastic bilevel optimization, the hypergradient involves the cross derivative $\nabla^2_{xy} g$ and the inverse-Hessian vector $(\nabla^2_{yy} g)^{-1}\nabla_y f$. In our performative setting, the LL distribution $\mathcal{D}(\boldsymbol{x})$ depends on the UL variable. Therefore, differentiating the LL optimality condition with respect to $\boldsymbol{x}$ introduces the additional distributional derivative term $(\boldsymbol{A}^{(d)})^{\intercal}\nabla^2_{\zeta y} g$. Meanwhile, since the UL distribution $\mathcal{C}(\boldsymbol{y})$ depends on the LL response, differentiating the UL objective with respect to $\boldsymbol{y}$ introduces the additional term $(\boldsymbol{A}^{(c)})^{\intercal}\nabla_{\xi} f$. Consequently, both the cross derivative term and the inverse-Hessian vector in the standard bilevel hypergradient must be modified to account for performative distribution shifts. The true hypergradient of $\Phi(\boldsymbol{x})$ can be written as

$$\begin{aligned}\nabla\Phi(\boldsymbol{x}) = &\mathop{\mathbb{E}}_{\boldsymbol{\xi}\sim\mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}))} \nabla_x f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\xi}) \\ &- \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x})}\left[\nabla^2_{xy} g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\zeta}) + (\boldsymbol{A}^{(d)})^{\intercal}\nabla^2_{\zeta y} g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\zeta})\right]\boldsymbol{v}^*(\boldsymbol{x}),\end{aligned} \tag{4}$$

where

$$\begin{aligned}&\boldsymbol{v}^*(\boldsymbol{x}) \\ &= \left[\mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x})} \nabla^2_{yy} g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\zeta})\right]^{-1} \mathop{\mathbb{E}}_{\boldsymbol{\xi}\sim\mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}))}\left[\nabla_y f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\xi}) + (\boldsymbol{A}^{(c)})^{\intercal}\nabla_{\xi} f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\xi})\right].\end{aligned} \tag{5}$$

Since the exact hypergradient in Eq. (4) requires the LL optimum, the true distribution maps, and an inverse Hessian, it cannot be evaluated directly. At outer iteration r, we replace $\boldsymbol{y}^*(\boldsymbol{x}_r)$ by the current LL iterate $\boldsymbol{y}_{r,K}$ and use the estimated distribution maps. We first construct

$$\boldsymbol{v}_{r,0} = \nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r, \mathcal{B}^{(f)}_r) + (\widehat{\boldsymbol{A}}^{(c)}_r)^{\intercal}\nabla_{\xi} f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r, \mathcal{B}^{(f)}_r), \tag{6}$$

which is the stochastic counterpart of the vector on the right-hand side of the inverse Hessian in Eq. (5). To avoid explicitly computing or inverting the LL Hessian, we approximate the corresponding

Algorithm 2 AdaBiO for Bilevel Performative Risk

Require: initial points $(\boldsymbol{x}_0, \boldsymbol{y}_0)$, R, K, Q, α, β, η, $\widehat{\boldsymbol{A}}_{-1}^{(d)} = \widehat{\boldsymbol{A}}_{-1}^{(c)} = \mathbf{0}$, and $S_{d,-1} = S_{c,-1} = 0$

for $r = 0, \ldots, R-1$ **do**
  Choose the cumulative LL map estimation target $S_{d,r} \geq S_{d,r-1} + 1$
  Update $\widehat{\boldsymbol{A}}_r^{(d)}$ via Algorithm 1 using $\boldsymbol{x}_r$, $S_{d,r-1}$, and $S_{d,r}$
  Set $\boldsymbol{y}_{r,0} \leftarrow \boldsymbol{y}_{r-1,K}$ if $r \geq 1$, otherwise $\boldsymbol{y}_{0,0} \leftarrow \boldsymbol{y}_0$
  for $k = 1, \ldots, K$ **do**
    Draw an independent mini-batch $\mathcal{B}_{r,k-1}^{(gy)}$ from $\widehat{\mathcal{D}}(\boldsymbol{x}_r)$
    Update $\boldsymbol{y}_{r,k} \leftarrow \boldsymbol{y}_{r,k-1} - \beta \nabla_y g\left(\boldsymbol{x}_r, \boldsymbol{y}_{r,k-1}; \boldsymbol{\zeta}_r, \mathcal{B}_{r,k-1}^{(gy)}\right)$
  end for
  Set $\boldsymbol{y}_r \leftarrow \boldsymbol{y}_{r,K}$
  Choose the cumulative UL map estimation target $S_{c,r} \geq S_{c,r-1} + 1$
  Update $\widehat{\boldsymbol{A}}_r^{(c)}$ via Algorithm 1 using $\boldsymbol{y}_{r,K}$, $S_{c,r-1}$, and $S_{c,r}$
  Draw independent mini-batches from $\widehat{\mathcal{C}}(\boldsymbol{y}_r)$ and $\widehat{\mathcal{D}}(\boldsymbol{x}_r)$ for $\mathcal{B}_r^{(f)}$, $\mathcal{B}_r^{(v)}$, and $\mathcal{B}_r^{(gxy)}$
  Compute $\boldsymbol{v}_{r,0}$ via Eq. (6) and $\boldsymbol{v}_{r,Q}$ via Eq. (7)
  Compute $\widehat{\nabla}\Phi(\boldsymbol{x}_r)$ via Eq. (8)
  Update $\boldsymbol{x}_{r+1} = \boldsymbol{x}_r - \alpha \widehat{\nabla}\Phi(\boldsymbol{x}_r)$
end for

inverse-Hessian-vector product using a stochastic truncated Neumann series:

$$\boldsymbol{v}_{r,Q} = \eta \sum_{q=0}^{Q} \prod_{j=Q+1-q}^{Q} \left(\boldsymbol{I} - \eta \nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_{r,j}, \mathcal{B}_{r,j}^{(v)})\right) \boldsymbol{v}_{r,0}, \tag{7}$$

where $\prod_{j=Q+1}^{Q}(\cdot) = \boldsymbol{I}$. Each Hessian mini-batch in Eq. (7) is drawn independently from $\widehat{\mathcal{D}}(\boldsymbol{x}_r)$. Therefore, $\boldsymbol{v}_{r,Q}$ provides a stochastic approximation to the inverse-Hessian vector $\boldsymbol{v}^*(\boldsymbol{x}_r)$ using the current LL iterate and the estimated distribution maps. Finally, substituting $\boldsymbol{y}_{r,K}$, the estimated distribution maps, and the stochastic inverse-Hessian vector $\boldsymbol{v}_{r,Q}$ into the performative hypergradient in Eq. (4) yields the stochastic hypergradient estimator

$$\begin{aligned}\widehat{\nabla}\Phi(\boldsymbol{x}_r) = {} & \nabla_x f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r, \mathcal{B}_r^{(f)}) \\ & - \left(\nabla_{xy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_r^{(gxy)}) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal \nabla_{\zeta y}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_r^{(gxy)})\right) \boldsymbol{v}_{r,Q}. \end{aligned} \tag{8}$$

Compared with stocBiO, the Neumann-series approximation retains the same basic role, whereas the vector $\boldsymbol{v}_{r,0}$ and the cross derivative term in the final hypergradient estimator are modified to incorporate the UL and LL distributional feedback, respectively.

2.3 AdaBiO for Bilevel Performative Risk

The overall AdaBiO procedure is summarized in Algorithm 2. The cumulative update targets $S_{d,r}$ and $S_{c,r}$ determine how many additional stochastic least-squares updates are performed at outer iteration r. Their concrete choices for achieving a desired stationarity accuracy are specified in Theorem 1 and its proof. At each outer iteration r, we first recursively update the LL location map estimate using Algorithm 1 and the estimate from the previous iteration. The updated map $\widehat{\mathcal{D}}(\boldsymbol{x}_r)$ is then used for the K warm-started stochastic LL updates. After obtaining $\boldsymbol{y}_{r,K}$, we similarly update the UL location map estimate and construct $\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})$. Based on the two updated distribution maps, we compute $\boldsymbol{v}_{r,0}$, its stochastic Neumann approximation $\boldsymbol{v}_{r,Q}$, and the performative hypergradient estimator in Eq. (8). Finally, the UL variable is updated using this stochastic hypergradient.

We next clarify the sampling procedure in Algorithm 2. Every call to $\widehat{\mathcal{D}}(\boldsymbol{x}_r)$ or $\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})$ uses newly independent base-noise samples. Specifically, a sample $\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)$ is generated as $\boldsymbol{\zeta}_r = \boldsymbol{\zeta}_r^{(0)} + \widehat{\boldsymbol{A}}_r^{(d)} \boldsymbol{x}_r$ with $\boldsymbol{\zeta}_r^{(0)} \sim \mathcal{D}^{(0)}$, and a sample $\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})$ is generated as $\boldsymbol{\xi}_r = \boldsymbol{\xi}_r^{(0)} + \widehat{\boldsymbol{A}}_r^{(c)} \boldsymbol{y}_{r,K}$ with $\boldsymbol{\xi}_r^{(0)} \sim \mathcal{C}^{(0)}$. Thus, all mini-batch quantities in Algorithm 2 are averages over samples from

the corresponding estimated deployment distributions. In particular, $\mathcal{B}_r^{(v)} = \{\mathcal{B}_{r,j}^{(v)}\}_{j=1}^{Q}$ consists of mutually independent mini-batches. Each Hessian factor in Eq. (7) therefore uses a newly independent sample set from $\widehat{\mathcal{D}}(\boldsymbol{x}_r)$. The mini-batches $\mathcal{B}_r^{(f)}$, $\mathcal{B}_r^{(gxy)}$, and $\{\mathcal{B}_{r,j}^{(v)}\}_{j=1}^{Q}$ are also sampled independently.

3 Main Theoretical Results

The following assumptions adapt standard regularity conditions commonly used in stochastic bilevel optimization and performative prediction. In particular, the strong convexity, smoothness, and bounded stochastic error conditions are consistent with those used in nonconvex-strongly-convex stochastic bilevel optimization (Ghadimi & Wang, 2018; Ji et al., 2021; Chen et al., 2022), while the regularity of decision-dependent distributions follows the performative prediction literature (Perdomo et al., 2020; Miller et al., 2021; Lu, 2023). The coupled dependence between the UL and LL distributions is also consistent with the bilevel performative setting studied in (Lu, 2023). We further specialize the distribution maps to location families, which makes their online estimation tractable from collected deployment samples.

Assumption 1. *Distribution maps of $\mathcal{C}(\cdot)$ and $\mathcal{D}(\cdot)$ form location families.*

Assumption 2. *For every $\boldsymbol{x}$ and $\boldsymbol{\zeta}$, the loss function $g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\zeta})$ is γ_y-strongly convex in $\boldsymbol{y}$.*

Assumption 3. *Assume that $\nabla_\xi f$ is L_f^ξ-Lipschitz in $\boldsymbol{\xi}$, $L_{f\xi}^x$-Lipschitz in $\boldsymbol{x}$, and $L_{f\xi}^y$-Lipschitz in $\boldsymbol{y}$. The LL gradient $\nabla_y g$ is L_{gy}^ζ-Lipschitz in $\boldsymbol{\zeta}$. The UL gradients satisfy that $\nabla_x f$ is L_f^x-Lipschitz in $\boldsymbol{x}$, $\bar{L}_f^x$-Lipschitz in $\boldsymbol{y}$, and L_{fx}^ξ-Lipschitz in $\boldsymbol{\xi}$, while $\nabla_y f$ is L_f^y-Lipschitz in $\boldsymbol{x}$, $\bar{L}_f^y$-Lipschitz in $\boldsymbol{y}$, and L_{fy}^ξ-Lipschitz in $\boldsymbol{\xi}$. The LL loss g is L_g^y-smooth in $\boldsymbol{y}$. Moreover, g is twice continuously differentiable, and $\nabla_{xy}^2 g$, $\nabla_{yy}^2 g$, and $\nabla_{\zeta y}^2 g$ are Lipschitz with constants L_{gxy}^ζ, L_{gyy}^ζ, $L_{g\zeta y}^\zeta$ in $\boldsymbol{\zeta}$, L_{gxy}^x, L_{gyy}^x, $L_{g\zeta y}^x$ in $\boldsymbol{x}$, and L_{gxy}^y, L_{gyy}^y, $L_{g\zeta y}^y$ in $\boldsymbol{y}$.*

Assumption 4. *Assume that $\|\nabla_x f(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\xi})\| \le C_f^x, \|\nabla_y f(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\xi})\| \le C_f^y, \|\nabla_\xi f(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\xi})\| \le C_f^\xi$, and $\|\nabla_{xy}^2 g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\zeta})\| \le C_g^{xy}, \|\nabla_{yy}^2 g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\zeta})\| \le C_g^{yy}, \|\nabla_{\zeta y}^2 g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\zeta})\| \le C_g^{\zeta y}$ for every sample $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$. In addition, every stochastic LL gradient mini-batch is unbiased and satisfies $\mathbb{E}\|\nabla_y g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\zeta}, \mathcal{B}) - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{P}} \nabla_y g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\zeta})\|^2 \le \frac{4(C_g^y)^2}{|\mathcal{B}|}$, where each element of $\mathcal{B}$ is drawn independently from $\mathcal{P}$.*

We first analyze the estimation errors of the two distribution maps. The following result characterizes the estimation errors of the two distribution maps in terms of the cumulative numbers of stochastic least-squares updates.

Proposition 1 (Distribution map estimation error)**.** *Suppose Assumption 1 holds and $\widehat{\boldsymbol{A}}_{-1}^{(d)} = \widehat{\boldsymbol{A}}_{-1}^{(c)} = \boldsymbol{0}$. Let $S_{d,r}$ and $S_{c,r}$ be cumulative update counts satisfying $S_{d,r} \ge S_{d,r-1} + 1, S_{c,r} \ge S_{c,r-1} + 1$, and then, for every $r \ge 0$,*

$$\mathbb{E}\left[\frac{S_{d,r} + 8d_x}{[1 + \log(S_{d,r} + 8d_x)]^2} \left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|_F^2\right] \le C_\zeta, \tag{9}$$

$$\mathbb{E}\left[\frac{S_{c,r} + 8d_y}{[1 + \log(S_{c,r} + 8d_y)]^2} \left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\|_F^2\right] \le C_\xi, \tag{10}$$

where C_ζ and C_ξ are positive constants defined in Appendix C. Furthermore, for any $\boldsymbol{x}$ and $\boldsymbol{y}$, the corresponding estimated distribution maps satisfy

$$W_1\left(\widehat{\mathcal{D}}(\boldsymbol{x}), \mathcal{D}(\boldsymbol{x})\right) \le \left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\| \|\boldsymbol{x}\|, W_1\left(\widehat{\mathcal{C}}(\boldsymbol{y}), \mathcal{C}(\boldsymbol{y})\right) \le \left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\| \|\boldsymbol{y}\|. \tag{11}$$

Proposition 1 shows that the distribution map estimation errors decrease as the cumulative numbers of stochastic least-squares updates increase. In particular, the mean-square errors decay as $\widetilde{\mathcal{O}}(1/r)$ under the minimum requirement $S_{d,r}, S_{c,r} \ge r + 1$. More importantly, the weighted bounds allow the cumulative update targets to be chosen adaptively according to the current deployment points without separating the map estimation errors from the optimization trajectory. The Wasserstein bounds characterize how the matrix estimation errors perturb the corresponding deployment distributions.

We next analyze the estimation error of the stochastic performative hypergradient in Eq. (8). In addition to the LL tracking and stochastic approximation errors that arise in standard bilevel optimization, our estimator is affected by the estimation errors of both decision-dependent distribution maps.

Lemma 1 (Hypergradient estimation error)**.** *Under Assumptions 1–4 and the sampling scheme in Algorithm 2, for every $r = 0, \dots, R-1$, we have*

$$\begin{aligned}
&\mathbb{E}\left\|\widehat{\nabla}\Phi(\boldsymbol{x}_r) - \nabla\Phi(\boldsymbol{x}_r)\right\|^2 \\
&\le \mathcal{H}_y \mathbb{E}\left\|\boldsymbol{y}_{r,K} - \boldsymbol{y}^*(\boldsymbol{x}_r)\right\|^2 + \mathcal{O}\left((1-\eta\gamma_y)^{2Q+2} + \frac{1}{B} + \frac{1}{|\mathcal{B}_r^{(f)}|} + \frac{1}{|\mathcal{B}_r^{(gxy)}|}\right) \\
&\quad + \mathcal{O}\Bigg(\mathbb{E}\left[(1+\|\boldsymbol{y}_{r,K}\|^2)\left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\|^2\right] + \mathbb{E}\left[(1+\|\boldsymbol{x}_r\|^2)\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^2\right] \\
&\quad + \sqrt{\mathbb{E}\left[(1+\|\boldsymbol{x}_r\|^2)^2\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^4\right]}\sqrt{\mathbb{E}\left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\|^4}\Bigg).
\end{aligned} \tag{12}$$

Here, $\mathcal{H}_y$ is a positive constant whose explicit expression is provided in Appendix D.

Lemma 1 separates the hypergradient estimation error into four components: the LL tracking error, the truncation error of the stochastic Neumann approximation, the stochastic mini-batch errors, and the errors caused by estimating the two decision-dependent distribution maps. The last component contains trajectory-weighted second- and fourth-moment errors of the online distribution map estimates and is specific to the performative setting.

For notational simplicity, let $\mathbb{E}[\mathcal{V}_r^{\Phi}]$ denote other terms except $\|\boldsymbol{y}_{r,K} - \boldsymbol{y}^*(\boldsymbol{x}_r)\|^2$ in the hypergradient error bound of Lemma 1. Then, $\mathbb{E}\left\|\widehat{\nabla}\Phi(\boldsymbol{x}_r) - \nabla\Phi(\boldsymbol{x}_r)\right\|^2 \le \mathcal{H}_y \mathbb{E}\left\|\boldsymbol{y}_{r,K} - \boldsymbol{y}^*(\boldsymbol{x}_r)\right\|^2 + \mathbb{E}\left[\mathcal{V}_r^{\Phi}\right]$.

Lemma 2. *Under Assumptions 1–4 and the sampling scheme in Algorithm 2, for $r \ge 1$, we have*

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{y}_{r,K} - \boldsymbol{y}^*(\boldsymbol{x}_r)\right\|^2 \\
&\le \lambda_y \mathbb{E}\left\|\boldsymbol{y}_{r-1,K} - \boldsymbol{y}^*(\boldsymbol{x}_{r-1})\right\|^2 + 12\alpha^2(1-\beta\gamma_y)^K \bar{L}_y^2 \mathbb{E}\left\|\nabla\Phi(\boldsymbol{x}_{r-1})\right\|^2 + 12\alpha^2(1-\beta\gamma_y)^K \bar{L}_y^2 \\
&\quad \cdot \mathbb{E}\left[\mathcal{V}_{r-1}^{\Phi}\right] + \frac{8\beta(C_g^y)^2}{\gamma_y|\mathcal{B}_r^{(gy)}|} + \left(6(1-\beta\gamma_y)^K + 2\right)\frac{(L_{gy}^{\zeta})^2}{\gamma_y^2}\mathbb{E}\left[\|\boldsymbol{x}_r\|^2\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^2\right],
\end{aligned} \tag{13}$$

where $\lambda_y := 6(1-\beta\gamma_y)^K + 12\alpha^2(1-\beta\gamma_y)^K \bar{L}_y^2 \mathcal{H}_y$, and $\bar{L}_y$ denotes the Lipschitz constant of the true LL response mapping $\boldsymbol{y}^*(\boldsymbol{x})$.

Lemma 2 shows that, when $\lambda_y < 1$, the warm-started LL iterates contract toward the moving LL solution $\boldsymbol{y}^*(\boldsymbol{x}_r)$ across outer iterations. Unlike the standard stochastic bilevel setting, the tracking error is additionally affected by the estimation error of the LL distribution map. Increasing K strengthens the LL contraction, while increasing the cumulative number of LL map estimation updates reduces the additional tracking error induced by the unknown performative distribution, as characterized in Proposition 1. We use $\mathrm{SC}_d(\epsilon)$ and $\mathrm{SC}_c(\epsilon)$ to denote the numbers of samples for estimating the LL and UL distribution maps, respectively. The following theorem summarizes the convergence and complexity guarantees.

Theorem 1 (Convergence complexity)**.** *Suppose Assumptions 1–4 hold and the performative objective is lower bounded, i.e., $\Phi_{\inf} := \inf_{\boldsymbol{x}\in\mathbb{R}^{d_x}} \Phi(\boldsymbol{x}) > -\infty$. Let L_Φ denote the Lipschitz constant of $\nabla\Phi$, whose expression is given in Appendix F. Choose $0 < \alpha \le 1/(2L_\Phi)$, $0 < \beta \le c_\beta/C_g^{yy}$, and $0 < \eta \le c_\eta/C_g^{yy}$, where $0 < c_\beta, c_\eta < 1$. Let $\bar{\boldsymbol{x}}$ be sampled uniformly from $\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{R-1}\}$. With the parameter choices specified in Appendix G, the following results hold:*

- ***Convergence:*** *Algorithm 2 returns an ϵ-accurate stationary point satisfying $\mathbb{E}\|\nabla\Phi(\bar{\boldsymbol{x}})\|^2 \le \mathcal{O}(\epsilon)$, with $R = \mathcal{O}\left(L_\Phi \epsilon^{-1}\right)$ outer iterations.*
- ***Computation complexity:*** *The corresponding computation complexities are $\mathrm{Gc}(f,\epsilon) = \mathcal{O}\left(L_\Phi\epsilon^{-2}\right)$, $\mathrm{Gc}(g,\epsilon) = \mathcal{O}\left(L_\Phi K\epsilon^{-2}\right)$, $\mathrm{JV}(g,\epsilon) = \mathcal{O}\left(L_\Phi\epsilon^{-2}\right)$, and $\mathrm{HV}(g,\epsilon) = \widetilde{\mathcal{O}}\left(L_\Phi(\eta\gamma_y)^{-2}\epsilon^{-2}\right)$.*

- ***Distribution map sample complexity:*** *The LL and UL distribution map estimation complexities satisfy* $\mathrm{SC}_d(\epsilon) = \widetilde{\mathcal{O}}\left((L_\Phi + \max_{0\le r<R}\|\boldsymbol{x}_r\|^2)\epsilon^{-1}\right)$, $\mathrm{SC}_c(\epsilon) = \widetilde{\mathcal{O}}\left((L_\Phi + \max_{0\le r<R}\|\boldsymbol{y}_{r,K}\|^2)\epsilon^{-1}\right)$. *If* $\max_{0\le r<R}\|\boldsymbol{x}_r\|$ *and* $\max_{0\le r<R}\|\boldsymbol{y}_{r,K}\|$ *are treated as constants, then* $\mathrm{SC}_d(\epsilon) = \mathrm{SC}_c(\epsilon) = \widetilde{\mathcal{O}}(L_\Phi\epsilon^{-1})$.

Remark 1. For comparison with Ji et al. (2021), we adopt the same condition number: all smoothness constants, bounded derivative constants, and fixed constants induced by the distribution maps are absorbed into a single constant L, and define $\kappa := L/\gamma_y$. In this case, $\mathrm{Gc}(f,\epsilon) = \mathcal{O}\left(\kappa^5\epsilon^{-2}\right)$, $\mathrm{Gc}(g,\epsilon) = \mathcal{O}\left(\kappa^9\epsilon^{-2}\right)$, $\mathrm{JV}(g,\epsilon) = \mathcal{O}\left(\kappa^5\epsilon^{-2}\right)$, $\mathrm{HV}(g,\epsilon) = \widetilde{\mathcal{O}}\left(\kappa^7\epsilon^{-2}\right)$, $\mathrm{SC}_d(\epsilon) = \mathrm{SC}_c(\epsilon) = \widetilde{\mathcal{O}}\left(\kappa^3\epsilon^{-1}\right)$.

Theorem 1 establishes convergence to a stationary point without requiring convexity of the UL performative objective. Under the condition number defined above, the proposed method requires $\mathcal{O}(\kappa^3\epsilon^{-1})$ outer iterations. Its gradient and Jacobian-vector product complexities have the same orders as stocBiO, while the current analysis incurs an additional factor of κ in the Hessian-vector product complexity. The distribution map estimation reuses all previous updates, so only additional samples required by the current cumulative targets are collected. Therefore, the sample complexity has $\widetilde{\mathcal{O}}(\epsilon^{-1})$, with the dependence on $\max_{0\le r<R}\|\boldsymbol{x}_r\|$ and $\max_{0\le r<R}\|\boldsymbol{y}_{r,K}\|$.

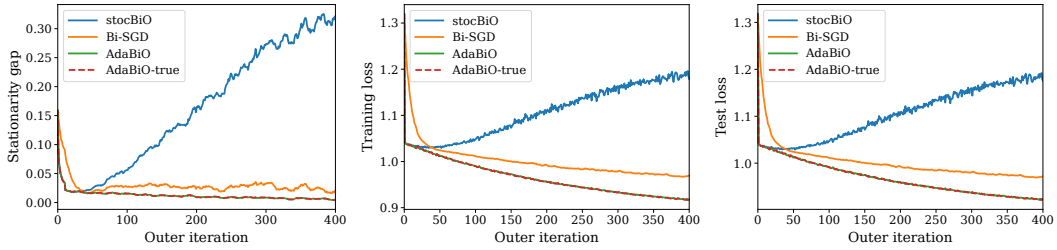
Table 1 compares the computational costs under the common nonconvex-strongly-convex bilevel setting. AdaBiO preserves the ϵ^{-2} dependence of stocBiO for all four types of computations and matches its κ-dependence for gradient and Jacobian-vector product evaluations. The current Hessian-vector product bound contains one additional factor of κ. Unlike the methods listed above, however, AdaBiO additionally handles unknown decision-dependent distributions at both levels, whose estimation costs are separately characterized by $\mathrm{SC}_d(\epsilon)$ and $\mathrm{SC}_c(\epsilon)$ in Theorem 1.

4 Applications to Strategic Learning

We conduct experiments on a strategic credit classification task to evaluate the proposed AdaBiO algorithm. We consider binary credit risk prediction on the Give Me Some Credit (GMSC) dataset (Kaggle, 2011). The LL and UL models are represented by parameter vectors $\boldsymbol{y}$ and $\boldsymbol{x}$, respectively. The distribution shifts are applied in the preprocessed credit feature space before a frozen nonlinear feature encoder. Only a subset of actionable credit attributes is allowed to change under deployment. For a frozen encoder $E_\psi(\cdot)$, we consider the LL objective $g(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\zeta}) = \ell\big(\boldsymbol{y}^\intercal E_\psi(\boldsymbol{\zeta})\big) + \frac{\lambda_y}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2$, and the UL objective $f(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\xi}) = \ell\big(\boldsymbol{y}^\intercal E_\psi(\boldsymbol{\xi})\big) + \frac{\lambda_x}{2}\|\boldsymbol{x}\|^2$, where ℓ denotes the class-weighted binary cross-entropy loss. The proximal term in the LL objective ensures strong convexity with respect to $\boldsymbol{y}$. We compare our method with the following three baselines:
stocBiO (Ji et al., 2021): Conventional stochastic bilevel optimizer trained on the base distributions without performative training. It ignores both decision-dependent distribution maps.
Bi-SGD (Lu, 2023): Bilevel performatively stable point (BPS) baseline. It uses samples from the currently deployed distributions but treats these distributions as fixed when constructing the bilevel gradient, and hence does not include the distributional derivative terms.
AdaBiO-true: A variant of AdaBiO that has access to the true matrices $\boldsymbol{A}^{(d)}$ and $\boldsymbol{A}^{(c)}$. This variant isolates the performance loss caused by online distribution map estimation.

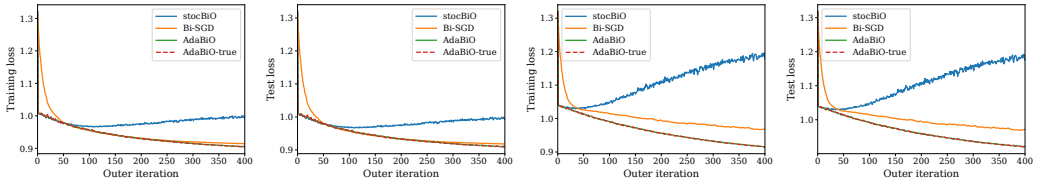
All methods are evaluated under the same true performative environment. For a controlled comparison, the bilevel methods use the same stochastic Neumann-series construction whenever an inverse-Hessian-vector product is required, while retaining their original LL update structures. In particular, stocBiO and AdaBiO use multiple stochastic LL updates, whereas Bi-SGD uses its original single-step LL update. We control the LL and UL distribution sensitivities through $\|\boldsymbol{A}^{(d)}\|_2 = \gamma_d$ and $\|\boldsymbol{A}^{(c)}\|_2 = \gamma_c$. Unless otherwise specified, we use $\gamma_d = \gamma_c = 1$. Detailed simulation settings are provided in Appendix B.

Fig. 1(a) shows that AdaBiO and AdaBiO-true achieve small stationarity gaps, whereas stocBiO deviates increasingly from performative stationarity and Bi-SGD remains at a larger gap. This difference is expected because stocBiO ignores the decision-dependent distributions, while Bi-SGD targets a performatively stable point rather than a stationary point of the performative risk. The result is also consistent with Theorem 1, which establishes the convergence of AdaBiO to a stationary



(a) Stationarity gap. (b) Training loss. (c) Test loss.

Figure 1: Performance comparisons of AdaBiO and baselines with $\gamma_c = \gamma_d = 1$.



(a) Training loss with $\gamma_c = 0.6, \gamma_d = 1$. (b) Test loss with $\gamma_c = 0.6, \gamma_d = 1$. (c) Training loss with $\gamma_c = 1, \gamma_d = 0.6$. (d) Test loss with $\gamma_c = 1, \gamma_d = 0.6$.

Figure 2: Performance comparisons of AdaBiO and baselines under different (γ_c, γ_d) pairs.

point of the performative risk. Figs. 1(b) and 1(c) further show that AdaBiO achieves lower training and test losses than stocBiO and Bi-SGD, while closely matching AdaBiO-true. The consistent improvement on both training and test data shows that the gain is not limited to the training samples. Moreover, the small gap between AdaBiO and AdaBiO-true indicates that the estimated distribution maps are sufficiently accurate for performative optimization. These results demonstrate the benefit of including the distributional derivatives in the hypergradient while learning the unknown distribution maps online.

Fig. 2 compares two settings with different UL and LL sensitivities: $(\gamma_c, \gamma_d) = (0.6, 1)$ and $(1, 0.6)$. Since the two settings simply exchange the same sensitivity values, this comparison separates the effects of stronger UL and LL distribution shifts. The performance gap between Bi-SGD and AdaBiO is larger when the UL shift is stronger, and stocBiO also shows a larger degradation. Meanwhile, AdaBiO remains close to AdaBiO-true in both settings, showing that the online map estimation remains effective under different shift strengths. These results indicate that the UL and LL distribution shifts have different effects on the optimization, and that the considered credit-classification task is more sensitive to the UL distribution shift.

5 Conclusion

This paper studied stochastic bilevel optimization for nonconvex performative risk when the data distributions at both levels depend on deployed decisions and are unknown to the learner. Unlike bilevel performative approaches that target performatively stable points, we directly considered a potentially nonconvex UL performative objective and developed AdaBiO to find an approximate stationary point of the performative risk. The proposed method combines online estimation of the two decision-dependent distribution maps with warm-started stochastic LL updates and a performative stochastic hypergradient estimator. We established error bounds for the distribution map estimation, characterized their effect on LL updates and hypergradient estimation, and proved convergence to an ϵ-stationary point. Experiments on strategic credit classification showed that AdaBiO consistently improves the performative risk over both standard stocBiO and Bi-SGD baselines, while achieving performance close to that of the variant with known distribution maps. Improvement becomes more significant as the strength of the performative feedback increases. Future work includes extending the framework beyond location-family distribution maps.

ACKNOWLEDGMENTS

No acknowledgments.

AI USE STATEMENT

Generative AI tools were used for language polishing and improving the clarity of the manuscript. The authors take full responsibility for the content of the paper.

ETHICS STATEMENT

This work studies a methodological framework for stochastic bilevel optimization and evaluates it using a publicly available credit-risk dataset. No human subjects were directly involved in our experiments.

REPRODUCIBILITY STATEMENT

Detailed experimental settings and parameter choices are provided in the appendix, together with complete proofs of the theoretical results.

REFERENCES

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *International conference on artificial intelligence and statistics*, pp. 6045–6061. PMLR, 2022.

Tianyi Chen, Yuejiao Sun, Quan Xiao, and Wotao Yin. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2466–2488. PMLR, 2022.

Xuxing Chen, Tesi Xiao, and Krishnakumar Balasubramanian. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *Journal of Machine Learning Research*, 25(151):1–51, 2024.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *Advances in Neural Information Processing Systems*, 35:26698–26710, 2022.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70, 2018.

Valia Efthymiou, Chara Podimata, Diptangshu Sen, and Juba Ziani. Incentivizing desirable effort profiles in strategic classification: The role of causality and uncertainty. *Advances in Neural Information Processing Systems*, 38:146550–146586, 2026.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In *International Conference on Machine Learning*, pp. 3672–3681. PMLR, 2021.

Xiaochuan Gong, Jie Hao, and Mingrui Liu. An accelerated algorithm for stochastic bilevel optimization under unbounded smoothness. *Advances in Neural Information Processing Systems*, 37: 78201–78243, 2024.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp. 111–122, 2016.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic, dec. 20. *arXiv preprint arXiv:2007.05170*, 2020.

Yifan Hu, Jie Wang, Yao Xie, Andreas Krause, and Daniel Kuhn. Contextual stochastic bilevel optimization. *Advances in Neural Information Processing Systems*, 36:78412–78434, 2023.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pp. 4882–4892. PMLR, 2021.

Kun Jin, Tongxin Yin, Zhongzhu Chen, Zeyu Sun, Xueru Zhang, Yang Liu, and Mingyan Liu. Performative federated learning: A solution to model-dependent and heterogeneous distribution shifts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12938–12946, 2024.

Kaggle. Give me some credit. `https://www.kaggle.com/competitions/GiveMeSomeCredit`, 2011.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in neural information processing systems*, 34:30271–30283, 2021.

Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *International Conference on Machine Learning*, pp. 6243–6253. PMLR, 2021.

Songtao Lu. Bilevel optimization with coupled decision-dependent distributions. In *International Conference on Machine Learning*, pp. 22758–22789. PMLR, 2023.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.

John P. Miller, Juan C. Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7710–7720. PMLR, 2021.

Mehrnaz Mofakhami, Ioannis Mitliagkas, and Gauthier Gidel. Performative prediction with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 11079–11093. PMLR, 2023.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International conference on machine learning*, pp. 7599–7609. PMLR, 2020.

Han Shao, Avrim Blum, and Omar Montasser. Strategic classification under unknown personalized manipulation. *Advances in Neural Information Processing Systems*, 36:26452–26484, 2023.

Songkai Xue and Yuekai Sun. Distributionally robust performative prediction. *Advances in Neural Information Processing Systems*, 37:55030–55052, 2024.

Wenjing Yan and Xuanyu Cao. Zero-regret performative prediction under inequality constraints. *Advances in Neural Information Processing Systems*, 36:1298–1308, 2023.

Tijana Zrnic, Eric Mazumdar, Shankar Sastry, and Michael Jordan. Who leads and who follows in strategic classification? *Advances in neural information processing systems*, 34:15257–15269, 2021.

A Related Work

Stochastic bilevel optimization. Stochastic bilevel optimization has been extensively studied for learning problems in which the upper-level objective depends implicitly on the solution of a stochastic lower-level problem. Ghadimi & Wang (2018) developed an early stochastic approximation framework and established convergence guarantees for bilevel problems with a strongly convex lower-level objective. Hong et al. (2020) subsequently studied a two-timescale stochastic approximation scheme for bilevel optimization. To improve the efficiency of hypergradient computation, Ji et al. (2021) proposed stocBiO, which combines stochastic lower-level updates with a Neumann-series approximation of the inverse-Hessian-vector product and provides refined complexity guarantees. In the same period, Khanduri et al. (2021) developed a single-timescale double-momentum method with near-optimal stochastic complexity. Chen et al. (2022) further proposed a single-timescale bilevel method with fixed mini-batches, while Dagréou et al. (2022) introduced a stochastic bilevel framework supporting global variance reduction. More recent studies have considered contextual bilevel problems (Hu et al., 2023), relaxed smoothness assumptions (Chen et al., 2024), and accelerated methods under unbounded smoothness (Gong et al., 2024). These approaches, however, are developed for bilevel objectives whose sampling distributions are fixed. In contrast, our setting allows the data distributions at both levels to change with the deployed decisions, which introduces additional distributional derivatives into the stochastic hypergradient and requires the distribution maps themselves to be learned online.

Performative prediction. Performative prediction studies learning systems in which deploying a model changes the distribution of future data. Perdomo et al. (2020) formalized this feedback loop, introduced performative risk and performative stability, and analyzed repeated retraining under distribution sensitivity. Mendler-Dünner et al. (2020) then studied stochastic optimization under performative feedback and characterized the role of deployment frequency. Rather than targeting only stable retraining points, Miller et al. (2021) investigated direct performative-risk minimization and identified structural conditions under which the single-level performative risk can be optimized. Brown et al. (2022) extended the framework to stateful populations, where the induced distribution depends on an evolving environmental state. Moving beyond single-level formulations, Lu (2023) introduced bilevel optimization with coupled decision-dependent distributions and studied bilevel performatively stable solutions. Other extensions have considered neural-network predictors (Mofakhami et al., 2023) and robustness against uncertainty in the performative distribution map (Xue & Sun, 2024). Different from these works, we directly optimize a potentially nonconvex bilevel performative risk and target first-order stationarity rather than a performatively stable fixed point. Moreover, both decision-dependent distribution maps are unknown in our setting and must be estimated online while the bilevel variables are being optimized.

Strategic learning. Strategic learning provides an important class of applications in which predictions alter the behavior of the population being predicted. Hardt et al. (2016) initiated the study of strategic classification, where individuals modify observable features in response to a deployed classifier. Dong et al. (2018) considered learning strategic classifiers from revealed preferences when the agents' responses are not directly known. Subsequent work examined richer interaction models, including the order of adaptation between the learner and strategic agents (Zrnic et al., 2021), learning with limited knowledge of strategic responses (Ghalme et al., 2021), and practical learning procedures for strategic classification (Levanon & Rosenfeld, 2021). More recently, Shao et al. (2023) studied unknown and personalized manipulation, while Efthymiou et al. (2026) investigated how causal structure and uncertainty affect the incentives for desirable effort. Our formulation includes such behavioral responses through decision-dependent distributions and allows performative feedback to arise at both levels of a bilevel learning problem, rather than assuming a fixed strategic-response model known to the learner.

B Simulation Details

We use the Give Me Some Credit (GMSC) dataset, which contains numerical credit attributes and a binary label indicating whether an individual experiences serious delinquency within two years. We divide the data into four disjoint subsets containing 30,000 samples for representation pretraining, 60,000 samples for the LL problem, 20,000 samples for the UL problem, and 30,000 samples for

testing, respectively. The splits are stratified according to the binary label. Since the positive class is substantially less frequent than the negative class, we use weighted binary cross entropy throughout the experiments, with the positive class weight determined from the corresponding training split.

We first train a small nonlinear feature encoder on the disjoint representation pretraining subset. The encoder consists of a fully connected layer with 32 hidden units and ReLU activation, followed by a 16-dimensional layer with $\tanh$ activation. The encoder is trained for 15 epochs using Adam with learning rate 10^{-3} and mini-batch size 512. After pretraining, the encoder parameters are frozen for all bilevel optimization experiments.

The UL and LL decision variables $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{16}$ parameterize linear classifiers on top of the frozen representation. All compared algorithms use the same frozen encoder, data splits, and initial values of $\boldsymbol{x}$ and $\boldsymbol{y}$. The LL loss is $g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\zeta}) = \ell\left(\boldsymbol{y}^\intercal E_\psi(\boldsymbol{\zeta})\right) + \frac{\lambda_y}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2$, where we set $\lambda_y = 1$. The UL loss is $f(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\xi}) = \ell\left(\boldsymbol{y}^\intercal E_\psi(\boldsymbol{\xi})\right) + \frac{\lambda_x}{2}\|\boldsymbol{x}\|^2$, with $\lambda_x = 10^{-3}$.

B.1 Performative Distribution Construction

The performative responses are introduced in the preprocessed raw feature space before the frozen encoder. For a base LL feature vector $\boldsymbol{\zeta}^{(0)}$ and a base UL feature vector $\boldsymbol{\xi}^{(0)}$, the deployed features are generated according to $\boldsymbol{\zeta} = \boldsymbol{\zeta}^{(0)} + \boldsymbol{A}^{(d)}\boldsymbol{x}$ and $\boldsymbol{\xi} = \boldsymbol{\xi}^{(0)} + \boldsymbol{A}^{(c)}\boldsymbol{y}$. We restrict strategic responses to the following five potentially actionable credit attributes: `RevolvingUtilizationOfUnsecuredLines`, `DebtRatio`, `MonthlyIncome`, `NumberOfOpenCreditLinesAndLoans`, and `NumberRealEstateLoansOrLines`. All rows of $\boldsymbol{A}^{(d)}$ and $\boldsymbol{A}^{(c)}$ corresponding to the remaining features are set to zero. The response matrix is spectrally normalized such that $\|\boldsymbol{A}^{(d)}\|_2 = \gamma_d$ and $\|\boldsymbol{A}^{(c)}\|_2 = \gamma_c$. The default experiment uses $\gamma_d = \gamma_c = 1$.

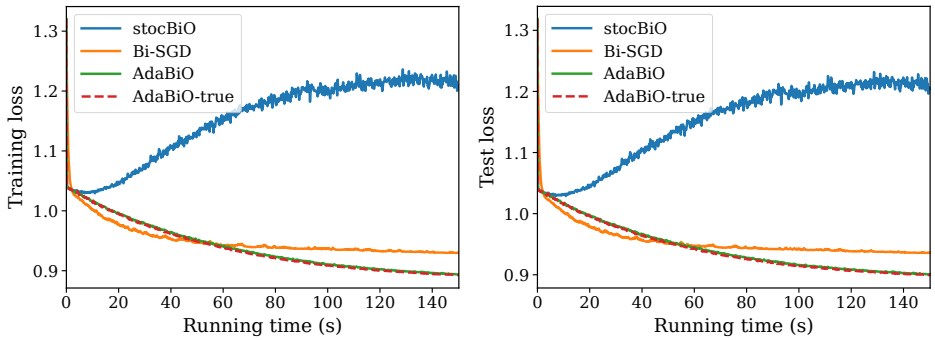
B.2 Optimization Settings

All compared methods use the same constant UL stepsize $\alpha = 0.03$ and LL stepsize $\beta = 0.03$. The Neumann-series stepsize is set to $\eta = 0.1$. For stocBiO, we use 100 stochastic LL updates per outer iteration. For AdaBiO and AdaBiO-true, we use $K = 100$ stochastic LL updates per outer iteration. Bi-SGD retains its original single-step LL recursion and therefore performs one LL stochastic update per outer iteration. We use a Neumann truncation depth of $Q = 24$ for all methods requiring an inverse-Hessian-vector estimate. For AdaBiO, we use $|\mathcal{B}^{(gy)}| = |\mathcal{B}^{(f)}| = |\mathcal{B}^{(gxy)}| = 640$, while the base Hessian batch parameter is $B = 160$. For distribution map estimation, we use the stochastic least-squares procedure in Algorithm 1. We set constants $c_d = c_c = 4$. At outer iteration r, the LL and UL map estimation targets $S_{d,r}$ and $S_{c,r}$ are selected adaptively according to Eqs. (163)–(164), based on the current norms of $\boldsymbol{x}_r$ and $\boldsymbol{y}_{r,K}$, respectively. Only the additional $S_{d,r} - S_{d,r-1}$ and $S_{c,r} - S_{c,r-1}$ stochastic least-squares updates are performed at each outer iteration, and the previous map estimates are preserved. Each stochastic map estimation update uses two independently collected deployment observations.

B.3 Additional Numerical Results

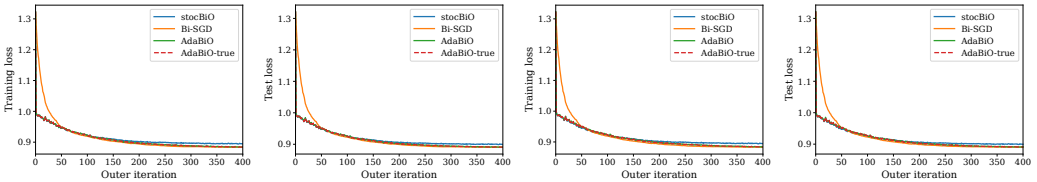
Running-time comparison. Fig. 3(a) and Fig. 3(b) report the training and test losses with respect to running time. Although Bi-SGD decreases the loss relatively quickly due to its simpler single-step LL update, AdaBiO and its true-map variant continue to improve as the optimization proceeds and achieve lower performative losses. In contrast, stocBiO gradually deteriorates. These results show that AdaBiO maintains its performance advantage when running time is used for comparison.

Additional results under different performative shifts. Fig. 4 provides additional comparisons under different combinations of the UL and LL performative shifts. When the UL shift is relatively weak ($\gamma_c = 0.2$), the four methods present similar performance, indicating that the effect of modeling the performative feedback is limited in this case. As γ_c increases, however, the difference between the methods becomes obvious: stocBiO shows a clear degradation, while Bi-SGD generally converges to a higher loss than AdaBiO and AdaBiO-true. In particular, the advantage of the two AdaBiO variants is most evident when $\gamma_c = 1.0$. Across different values of γ_d, AdaBiO remains close to

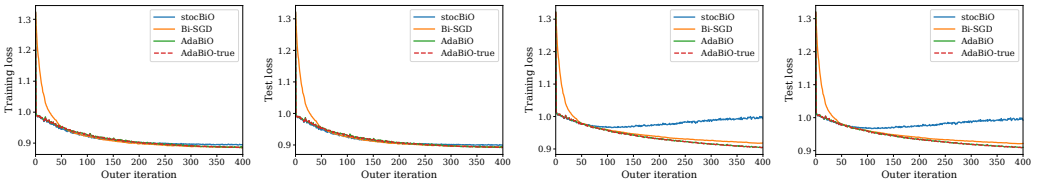


(a) Training loss. (b) Test loss.

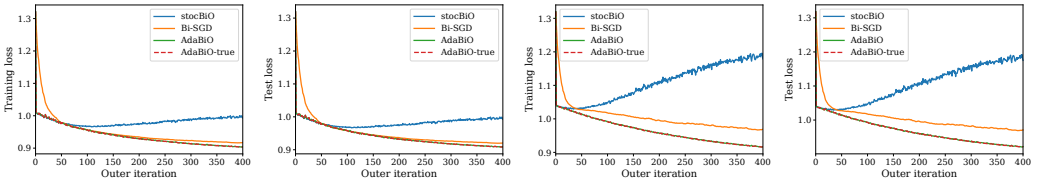
Figure 3: Convergence performance versus running time with $\gamma_c = \gamma_d = 1$.



(a) Training loss ($\gamma_c = 0.2, \gamma_d = 0.2$). (b) Test loss ($\gamma_c = 0.2, \gamma_d = 0.2$). (c) Training loss ($\gamma_c = 0.2, \gamma_d = 0.6$). (d) Test loss ($\gamma_c = 0.2, \gamma_d = 0.6$).



(e) Training loss ($\gamma_c = 0.2, \gamma_d = 1$). (f) Test loss ($\gamma_c = 0.2, \gamma_d = 1$). (g) Training loss ($\gamma_c = 0.6, \gamma_d = 0.2$). (h) Test loss ($\gamma_c = 0.6, \gamma_d = 0.2$).



(i) Training loss ($\gamma_c = 0.6, \gamma_d = 0.6$). (j) Test loss ($\gamma_c = 0.6, \gamma_d = 0.6$). (k) Training loss ($\gamma_c = 1, \gamma_d = 0.2$). (l) Test loss ($\gamma_c = 1, \gamma_d = 0.2$).

Figure 4: Performance comparison of AdaBiO and baselines with different γ_c and γ_d.

its true-map version, demonstrating that the proposed online map estimation is effective over a broad range of performative shift strengths. Overall, these results further confirm that explicitly incorporating distributional derivatives becomes increasingly important as the performative feedback grows stronger, especially for the UL distribution shift.

C Proof of Proposition 1

Based on Assumption 1, the LL distribution map $\mathcal{D}(\cdot)$ and the UL distribution map $\mathcal{C}(\cdot)$ have the representations

$$\boldsymbol{\zeta} \sim \mathcal{D}(\boldsymbol{x}) \Longleftrightarrow \boldsymbol{\zeta} = \boldsymbol{\zeta}^{(0)} + \boldsymbol{A}^{(d)}\boldsymbol{x}, \tag{14}$$

and

$$\boldsymbol{\xi} \sim \mathcal{C}(\boldsymbol{y}) \Longleftrightarrow \boldsymbol{\xi} = \boldsymbol{\xi}^{(0)} + \boldsymbol{A}^{(c)}\boldsymbol{y}. \tag{15}$$

We first describe the cumulative stochastic least-squares updates used to estimate the two location matrices. Let $S_{d,r}$ denote the total number of LL distribution map SGD updates completed by the end of outer iteration r, with $S_{d,-1} = 0$. We require

$$S_{d,r} \geq S_{d,r-1} + 1. \tag{16}$$

Thus, at outer iteration r, the LL map estimator performs $S_{d,r} - S_{d,r-1}$ new stochastic least-squares updates. The choice of $S_{d,r}$ will be specified later in Appendix G. Set $\widehat{\boldsymbol{A}}_{r,0}^{(d)} = \widehat{\boldsymbol{A}}_{r-1}^{(d)}, \widehat{\boldsymbol{A}}_{-1}^{(d)} = \mathbf{0}$. For $j = 1, \ldots, S_{d,r} - S_{d,r-1}$, independently draw $\boldsymbol{x}_{r,j}^{\exp} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_x})$, and observe $\boldsymbol{\zeta}_{r,j}^{(1)} \sim \mathcal{D}(\boldsymbol{x}_r), \boldsymbol{\zeta}_{r,j}^{(2)} \sim \mathcal{D}\left(\boldsymbol{x}_r + \boldsymbol{x}_{r,j}^{\exp}\right)$. Under the location-family model, $\boldsymbol{\zeta}_{r,j}^{(1)} = \boldsymbol{\zeta}_{r,j}^{(0,1)} + \boldsymbol{A}^{(d)}\boldsymbol{x}_r, \boldsymbol{\zeta}_{r,j}^{(2)} = \boldsymbol{\zeta}_{r,j}^{(0,2)} + \boldsymbol{A}^{(d)}\left(\boldsymbol{x}_r + \boldsymbol{x}_{r,j}^{\exp}\right)$, where $\boldsymbol{\zeta}_{r,j}^{(0,1)}$ and $\boldsymbol{\zeta}_{r,j}^{(0,2)}$ are independent copies of $\boldsymbol{\zeta}^{(0)}$. Hence,

$$\boldsymbol{\zeta}_{r,j}^{(2)} - \boldsymbol{\zeta}_{r,j}^{(1)} = \boldsymbol{A}^{(d)}\boldsymbol{x}_{r,j}^{\exp} + \boldsymbol{\zeta}_{r,j}^{(0,2)} - \boldsymbol{\zeta}_{r,j}^{(0,1)}. \tag{17}$$

The j-th LL map estimation update at outer iteration r is

$$\widehat{\boldsymbol{A}}_{r,j}^{(d)} = \widehat{\boldsymbol{A}}_{r,j-1}^{(d)} + \frac{2}{S_{d,r-1} + j + 8d_x}\left(\boldsymbol{\zeta}_{r,j}^{(2)} - \boldsymbol{\zeta}_{r,j}^{(1)} - \widehat{\boldsymbol{A}}_{r,j-1}^{(d)}\boldsymbol{x}_{r,j}^{\exp}\right)(\boldsymbol{x}_{r,j}^{\exp})^{\intercal}. \tag{18}$$

After all LL map estimation updates at outer iteration r, set $\widehat{\boldsymbol{A}}_r^{(d)} = \widehat{\boldsymbol{A}}_{r,S_{d,r}-S_{d,r-1}}^{(d)}$. The estimated LL distribution map is

$$\widehat{\boldsymbol{\zeta}} \sim \widehat{\mathcal{D}}(\boldsymbol{x}) \Longleftrightarrow \widehat{\boldsymbol{\zeta}} = \boldsymbol{\zeta}^{(0)} + \widehat{\boldsymbol{A}}_r^{(d)}\boldsymbol{x}. \tag{19}$$

Similarly, let $S_{c,r}$ denote the total number of UL distribution map SGD updates completed by the end of outer iteration r, with $S_{c,-1} = 0$, and require

$$S_{c,r} \geq S_{c,r-1} + 1. \tag{20}$$

After obtaining $\boldsymbol{y}_{r,K}$, set $\widehat{\boldsymbol{A}}_{r,0}^{(c)} = \widehat{\boldsymbol{A}}_{r-1}^{(c)}, \widehat{\boldsymbol{A}}_{-1}^{(c)} = \mathbf{0}$. For $j = 1, \ldots, S_{c,r} - S_{c,r-1}$, independently draw $\boldsymbol{y}_{r,j}^{\exp} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_y})$, and observe $\boldsymbol{\xi}_{r,j}^{(1)} \sim \mathcal{C}(\boldsymbol{y}_{r,K}), \boldsymbol{\xi}_{r,j}^{(2)} \sim \mathcal{C}\left(\boldsymbol{y}_{r,K} + \boldsymbol{y}_{r,j}^{\exp}\right)$. Under the location-family model,

$$\boldsymbol{\xi}_{r,j}^{(2)} - \boldsymbol{\xi}_{r,j}^{(1)} = \boldsymbol{A}^{(c)}\boldsymbol{y}_{r,j}^{\exp} + \boldsymbol{\xi}_{r,j}^{(0,2)} - \boldsymbol{\xi}_{r,j}^{(0,1)}, \tag{21}$$

where $\boldsymbol{\xi}_{r,j}^{(0,1)}$ and $\boldsymbol{\xi}_{r,j}^{(0,2)}$ are independent copies of $\boldsymbol{\xi}^{(0)}$. The j-th UL map estimation update at outer iteration r is

$$\widehat{\boldsymbol{A}}_{r,j}^{(c)} = \widehat{\boldsymbol{A}}_{r,j-1}^{(c)} + \frac{2}{S_{c,r-1} + j + 8d_y}\left(\boldsymbol{\xi}_{r,j}^{(2)} - \boldsymbol{\xi}_{r,j}^{(1)} - \widehat{\boldsymbol{A}}_{r,j-1}^{(c)}\boldsymbol{y}_{r,j}^{\exp}\right)(\boldsymbol{y}_{r,j}^{\exp})^{\intercal}. \tag{22}$$

After all UL map estimation updates at outer iteration r, set $\widehat{\boldsymbol{A}}_r^{(c)} = \widehat{\boldsymbol{A}}_{r,S_{c,r}-S_{c,r-1}}^{(c)}$. The estimated UL distribution map is $\widehat{\boldsymbol{\xi}} \sim \widehat{\mathcal{C}}(\boldsymbol{y}) \Longleftrightarrow \widehat{\boldsymbol{\xi}} = \boldsymbol{\xi}^{(0)} + \widehat{\boldsymbol{A}}_r^{(c)}\boldsymbol{y}$.

The following lemma characterizes the estimation errors in terms of the cumulative numbers of stochastic least-squares updates.

Lemma 3 (Estimation error of distribution maps)**.** *Suppose Assumption 1 holds. Let $S_{d,r}$ and $S_{c,r}$ be finite cumulative update indices satisfying Eqs. (16) and (20), respectively, and determined before the*

map estimation samples corresponding to the additional updates are drawn. Then there exist positive constants C_ζ and C_ξ, such that

$$\mathbb{E}\left[\frac{S_{d,r}+8d_x}{\left[1+\log(S_{d,r}+8d_x)\right]^2}\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|_F^2\right]\le C_\zeta, \tag{23}$$

$$\mathbb{E}\left[\frac{S_{c,r}+8d_y}{\left[1+\log(S_{c,r}+8d_y)\right]^2}\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|_F^2\right]\le C_\xi. \tag{24}$$

Moreover,

$$\mathbb{E}\left[\frac{(S_{d,r}+8d_x)^2}{\left[1+\log(S_{d,r}+8d_x)\right]^4}\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|_F^4\right]\le C_\zeta, \tag{25}$$

$$\mathbb{E}\left[\frac{(S_{c,r}+8d_y)^2}{\left[1+\log(S_{c,r}+8d_y)\right]^4}\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|_F^4\right]\le C_\xi. \tag{26}$$

Since $S_{d,r}\ge r+1$ and $S_{c,r}\ge r+1$, it follows that

$$\mathbb{E}\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|_F^2\le C_\zeta\frac{\left[1+\log(r+1+8d_x)\right]^2}{r+1+8d_x}, \tag{27}$$

$$\mathbb{E}\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|_F^2\le C_\xi\frac{\left[1+\log(r+1+8d_y)\right]^2}{r+1+8d_y}, \tag{28}$$

and

$$\mathbb{E}\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|_F^4\le C_\zeta\frac{\left[1+\log(r+1+8d_x)\right]^4}{(r+1+8d_x)^2}, \tag{29}$$

$$\mathbb{E}\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|_F^4\le C_\xi\frac{\left[1+\log(r+1+8d_y)\right]^4}{(r+1+8d_y)^2}. \tag{30}$$

Proof. We first prove the result for the LL distribution map estimator. We index all LL map estimation updates in the order they are performed across outer iterations and let $n=1,2,\dots$ denote their cumulative update index. Let $\mathcal{F}_n^{(d)}$ denote the algorithmic history available before the $(n+1)$-th LL map estimation update. For the proof only, let $\boldsymbol{E}_n^{(d)}$ denote the LL matrix estimation error after n cumulative updates.

By Eq. (17) and Eq. (18),

$$\boldsymbol{E}_n^{(d)}=\boldsymbol{E}_{n-1}^{(d)}\left(\boldsymbol{I}_{d_x}-\frac{2}{n+8d_x}\boldsymbol{x}_n^{\exp}(\boldsymbol{x}_n^{\exp})^\intercal\right)+\frac{2}{n+8d_x}\left(\boldsymbol{\zeta}_n^{(0,2)}-\boldsymbol{\zeta}_n^{(0,1)}\right)(\boldsymbol{x}_n^{\exp})^\intercal. \tag{31}$$

Here, $\boldsymbol{x}_n^{\exp}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_{d_x})$, and $\boldsymbol{x}_n^{\exp}$, $\boldsymbol{\zeta}_n^{(0,1)}$, and $\boldsymbol{\zeta}_n^{(0,2)}$ are mutually independent and independent of $\mathcal{F}_{n-1}^{(d)}$. Moreover, $\boldsymbol{E}_{n-1}^{(d)}$ is $\mathcal{F}_{n-1}^{(d)}$-measurable. The Gaussian identities

$$\mathbb{E}\left[\boldsymbol{x}_n^{\exp}(\boldsymbol{x}_n^{\exp})^\intercal\right]=\boldsymbol{I}_{d_x}, \tag{32}$$

$$\mathbb{E}\left[\|\boldsymbol{x}_n^{\exp}\|^2\boldsymbol{x}_n^{\exp}(\boldsymbol{x}_n^{\exp})^\intercal\right]=(d_x+2)\boldsymbol{I}_{d_x} \tag{33}$$

and the independence and centeredness of the base noises imply

$$\mathbb{E}\left[\|\boldsymbol{E}_n^{(d)}\|_F^2\mid\mathcal{F}_{n-1}^{(d)}\right]=\left[1-\frac{4}{n+8d_x}+\frac{4(d_x+2)}{(n+8d_x)^2}\right]\|\boldsymbol{E}_{n-1}^{(d)}\|_F^2+\frac{8d_x\mathbb{E}\|\boldsymbol{\zeta}^{(0)}\|^2}{(n+8d_x)^2}. \tag{34}$$

Since $\frac{2(d_x+2)}{n+8d_x}\le 1$, we obtain

$$\mathbb{E}\left[\|\boldsymbol{E}_n^{(d)}\|_F^2\mid\mathcal{F}_{n-1}^{(d)}\right]\le\left(1-\frac{2}{n+8d_x}\right)\|\boldsymbol{E}_{n-1}^{(d)}\|_F^2+\frac{8d_x\mathbb{E}\|\boldsymbol{\zeta}^{(0)}\|^2}{(n+8d_x)^2}. \tag{35}$$

Multiplying Eq. (35) by $\frac{n+8d_x}{[1+\log(n+8d_x)]^2}$ gives

$$\mathbb{E}\left[\frac{n+8d_x}{[1+\log(n+8d_x)]^2}\|\boldsymbol{E}_n^{(d)}\|_F^2 \mid \mathcal{F}_{n-1}^{(d)}\right] \leq \frac{n+8d_x-2}{[1+\log(n+8d_x)]^2}\|\boldsymbol{E}_{n-1}^{(d)}\|_F^2 + \frac{8d_x\mathbb{E}\|\boldsymbol{\zeta}^{(0)}\|^2}{(n+8d_x)\,[1+\log(n+8d_x)]^2}. \tag{36}$$

Since

$$\frac{n+8d_x-2}{[1+\log(n+8d_x)]^2} \leq \frac{n-1+8d_x}{[1+\log(n-1+8d_x)]^2}, \tag{37}$$

and

$$\sum_{n=1}^{\infty}\frac{1}{(n+8d_x)\,[1+\log(n+8d_x)]^2} < \infty, \tag{38}$$

Since the last term in Eq. (36) is summable by Eq. (38), repeatedly applying the recursion gives a bounded weighted LL map estimation error. Setting the number of updates to $S_{d,r}$ then gives Eq. (23).

We next establish the fourth-moment bound. By the location-family model introduced in Section 2.1, the centered sub-Gaussian vector $\boldsymbol{\zeta}^{(0)}$ has a finite fourth moment. For any fixed matrix $\boldsymbol{E}$, any $\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_{d_x})$, and any $\nu>0$, Gaussian moment identities give

$$\mathbb{E}\left[\|\boldsymbol{E}\,(\boldsymbol{I}_{d_x}-\nu\boldsymbol{u}\boldsymbol{u}^\intercal)\|_F^4\right] = \left[1-4\nu+2(d_x+2)\nu^2\right]\|\boldsymbol{E}\|_F^4 + \left[4\nu^2-4(d_x+4)\nu^3+(d_x+4)(d_x+6)\nu^4\right]\left(\|\boldsymbol{E}\|_F^4+2\|\boldsymbol{E}^\intercal\boldsymbol{E}\|_F^2\right). \tag{39}$$

Using $\|\boldsymbol{E}^\intercal\boldsymbol{E}\|_F^2\leq\|\boldsymbol{E}\|_F^4$ and setting $\nu=\frac{2}{n+8d_x}$, we obtain

$$\mathbb{E}\left[\left\|\boldsymbol{E}\left(\boldsymbol{I}_{d_x}-\frac{2}{n+8d_x}\boldsymbol{u}\boldsymbol{u}^\intercal\right)\right\|_F^4\right] \leq \left(1-\frac{3}{n+8d_x}\right)\|\boldsymbol{E}\|_F^4. \tag{40}$$

Taking the fourth power in Eq. (31), applying Cauchy–Schwarz and Young's inequalities to the remaining mixed terms, we obtain that

$$\mathbb{E}\left[\|\boldsymbol{E}_n^{(d)}\|_F^4 \mid \mathcal{F}_{n-1}^{(d)}\right] \leq \left(1-\frac{5}{2(n+8d_x)}\right)\|\boldsymbol{E}_{n-1}^{(d)}\|_F^4 + \frac{C}{(n+8d_x)^2}\|\boldsymbol{E}_{n-1}^{(d)}\|_F^2 + \frac{C}{(n+8d_x)^3}, \tag{41}$$

where C is a positive constant. Multiplying Eq. (41) by $\frac{(n+8d_x)^2}{[1+\log(n+8d_x)]^4}$ and using $\frac{(n+8d_x)^2\left(1-\frac{5}{2(n+8d_x)}\right)}{[1+\log(n+8d_x)]^4} \leq \frac{(n-1+8d_x)^2}{[1+\log(n-1+8d_x)]^4}$, we obtain

$$\mathbb{E}\left[\frac{(n+8d_x)^2}{[1+\log(n+8d_x)]^4}\|\boldsymbol{E}_n^{(d)}\|_F^4 \mid \mathcal{F}_{n-1}^{(d)}\right] \leq \frac{(n-1+8d_x)^2}{[1+\log(n-1+8d_x)]^4}\|\boldsymbol{E}_{n-1}^{(d)}\|_F^4 + \frac{C}{[1+\log(n+8d_x)]^4}\|\boldsymbol{E}_{n-1}^{(d)}\|_F^2 + \frac{C}{(n+8d_x)\,[1+\log(n+8d_x)]^4}. \tag{42}$$

Moreover,

$$\frac{C}{[1+\log(n+8d_x)]^4}\|\boldsymbol{E}_{n-1}^{(d)}\|_F^2 = \frac{C\,[1+\log(n-1+8d_x)]^2}{(n-1+8d_x)\,[1+\log(n+8d_x)]^4}\cdot\frac{n-1+8d_x}{[1+\log(n-1+8d_x)]^2}\|\boldsymbol{E}_{n-1}^{(d)}\|_F^2. \tag{43}$$

Since

$$\sum_{n=1}^{\infty} \frac{[1+\log(n-1+8d_x)]^2}{(n-1+8d_x)\,[1+\log(n+8d_x)]^4} < \infty, \tag{44}$$

$$\sum_{n=1}^{\infty} \frac{1}{(n+8d_x)\,[1+\log(n+8d_x)]^4} < \infty, \tag{45}$$

combining these results with the weighted second-moment bound and Eq. (42) gives Eq. (25).

For the UL estimator, we repeat the above steps for the cumulative UL map estimation updates. Replacing $d_x, \boldsymbol{\zeta}^{(0)}, \boldsymbol{A}^{(d)}, S_{d,r}$ with $d_y, \boldsymbol{\xi}^{(0)}, \boldsymbol{A}^{(c)}, S_{c,r}$, respectively, gives Eqs. (24) and (26). Finally, since the function $\frac{s+8d}{[1+\log(s+8d)]^2}$ is nondecreasing for $s \geq 0$, and $S_{d,r} \geq r+1$ and $S_{c,r} \geq r+1$, Eqs. (27)–(30) follow. □

We next relate the matrix estimation errors to the corresponding distribution estimation errors. For any fixed of $\widehat{\boldsymbol{A}}_r^{(d)}$, couple $\widehat{\mathcal{D}}(\boldsymbol{x})$ and $\mathcal{D}(\boldsymbol{x})$ using the same base sample $\boldsymbol{\zeta}^{(0)}$. Then,

$$\begin{aligned} W_1\left(\widehat{\mathcal{D}}(\boldsymbol{x}), \mathcal{D}(\boldsymbol{x})\right) &\leq \mathbb{E}\left\|\boldsymbol{\zeta}^{(0)} + \widehat{\boldsymbol{A}}_r^{(d)}\boldsymbol{x} - \boldsymbol{\zeta}^{(0)} - \boldsymbol{A}^{(d)}\boldsymbol{x}\right\| \\ &= \left\|\left(\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right)\boldsymbol{x}\right\| \\ &\leq \left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\| \|\boldsymbol{x}\|. \end{aligned} \tag{46}$$

Similarly,

$$W_1\left(\widehat{\mathcal{C}}(\boldsymbol{y}), \mathcal{C}(\boldsymbol{y})\right) \leq \left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\| \|\boldsymbol{y}\|. \tag{47}$$

Furthermore, if $\sigma_{\max}(\boldsymbol{A}^{(d)}) := \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}^{(d)}\boldsymbol{x}\|, \sigma_{\max}(\boldsymbol{A}^{(c)}) := \max_{\|\boldsymbol{y}\|=1} \|\boldsymbol{A}^{(c)}\boldsymbol{y}\|$, then

$$W_1\left(\mathcal{D}(\boldsymbol{x}), \mathcal{D}(\boldsymbol{x}')\right) \leq \sigma_{\max}(\boldsymbol{A}^{(d)})\|\boldsymbol{x} - \boldsymbol{x}'\|, \tag{48}$$

$$W_1\left(\mathcal{C}(\boldsymbol{y}), \mathcal{C}(\boldsymbol{y}')\right) \leq \sigma_{\max}(\boldsymbol{A}^{(c)})\|\boldsymbol{y} - \boldsymbol{y}'\|. \tag{49}$$

D Proof of Lemma 1

We first present the following lemmas on stochastic Neumann approximation error and first moment error of the hypergradient estimator to facilitate our proof of Lemma 1.

D.1 Stochastic Neumann Approximation Error

Lemma 4. *With Assumptions 1-4, we have*

$$\begin{aligned} \mathbb{E}\,\|\boldsymbol{v}_{r,Q} - \boldsymbol{v}_r^*\|^2 \leq\; & 2\left(\frac{16\eta^2(C_g^{yy})^2}{(1-\eta\gamma_y)\gamma_y^2 B} + \frac{4}{\gamma_y^2|\mathcal{B}_r^{(f)}|} + \frac{2(1-\eta\gamma_y)^{2Q+2}}{\gamma_y^2}\right) \\ & \cdot \left(C_f^y + \|\boldsymbol{A}^{(c)}\|C_f^\xi + \|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\|C_f^\xi\right)^2 \\ & + 4\left(\frac{L_f^\xi\|\boldsymbol{y}_{r,K}\| + \|\boldsymbol{A}^{(c)}\|L_f^\xi\|\boldsymbol{y}_{r,K}\| + C_f^\xi}{\gamma_y}\right)^2 \left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\|^2 \\ & + 4\left(\frac{\left(C_f^y + \|\boldsymbol{A}^{(c)}\|C_f^\xi\right)L_{gyy}^\zeta\|\boldsymbol{x}_r\|}{\gamma_y^2}\right)^2 \left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^2. \end{aligned} \tag{50}$$

Proof. Based on the construction of $\boldsymbol{v}_{r,Q}$, we have

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{v}_{r,Q}] \\
&= \mathbb{E}\Bigg[\eta \sum_{q=0}^{Q} \prod_{j=Q+1-q}^{Q} (I - \eta \nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_{r,j}^{(v)})) \Big(\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r, \mathcal{B}_r^{(f)}) \\
&\quad + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r, \mathcal{B}_r^{(f)}) \Big) \Bigg] \\
&= \eta \sum_{q=0}^{Q} \prod_{j=Q+1-q}^{Q} \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} [I - \eta \nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r)] \\
&\quad \cdot \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})} \Big[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) \Big] \\
&= \eta \sum_{q=0}^{Q} \left(I - \eta \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) \right)^q \\
&\quad \cdot \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})} \Big[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) \Big] .
\end{aligned} \tag{51}$$

We denote $\widehat{\boldsymbol{v}}_r^*$ under estimated distribution map as

$$\begin{aligned}
&\widehat{\boldsymbol{v}}_r^* \\
&= \left[\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) \right]^{-1} \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})} \Big[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) \Big] .
\end{aligned} \tag{52}$$

Let the Neumann stepsize satisfy $0 < \eta \leq c_\eta / C_g^{yy}$ for a fixed constant $0 < c_\eta < 1$. The conditional Neumann truncation bias between $\boldsymbol{v}_{r,Q}$ and $\widehat{\boldsymbol{v}}_r^*$ is

$$\begin{aligned}
&\|\mathbb{E}[\boldsymbol{v}_{r,Q}] - \widehat{\boldsymbol{v}}_r^*\| \\
&= \Bigg\| \mathbb{E}[\boldsymbol{v}_{r,Q}] - \left[\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) \right]^{-1} \\
&\qquad \cdot \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})} \Big[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) \Big] \Bigg\| \\
&\overset{(a)}{=} \Bigg\| \mathbb{E}[\boldsymbol{v}_{r,Q}] - \eta \sum_{q=0}^{\infty} \left(I - \eta \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) \right)^q \\
&\qquad \cdot \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})} \Big[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) \Big] \Bigg\| \\
&= \Bigg\| \eta \sum_{q=Q+1}^{\infty} \left(I - \eta \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) \right)^q \\
&\qquad \cdot \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})} \Big[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) \Big] \Bigg\| \\
&\overset{(b)}{\leq} \frac{(1 - \eta\gamma_y)^{Q+1}}{\gamma_y} \left(C_f^y + \|\widehat{\boldsymbol{A}}_r^{(c)}\| C_f^\xi \right),
\end{aligned} \tag{53}$$

where (a) is true based on the Neumann series and (b) holds due to Assumptions 2 and 4. We denote $\boldsymbol{v}_r^*$ under the true distribution map as

$$\boldsymbol{v}_r^* = \left[\mathop{\mathbb{E}}_{\boldsymbol{\zeta} \sim \mathcal{D}(\boldsymbol{x}_r)} \nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}) \right]^{-1} \mathop{\mathbb{E}}_{\boldsymbol{\xi} \sim \mathcal{C}(\boldsymbol{y}_{r,K})} \Big[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}) + (\boldsymbol{A}^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}) \Big] . \tag{54}$$

The distribution map perturbation between $\widehat{\boldsymbol{v}}_r^*$ and $\boldsymbol{v}_r^*$ is given as

$$
\begin{aligned}
&\|\widehat{\boldsymbol{v}}_r^* - \boldsymbol{v}_r^*\| \\
&= \left\| \left[\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) \right]^{-1} \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})} \left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) \right] \right. \\
&\quad \left. - \left[\mathop{\mathbb{E}}_{\boldsymbol{\zeta} \sim \mathcal{D}(\boldsymbol{x}_r)} \nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}) \right]^{-1} \mathop{\mathbb{E}}_{\boldsymbol{\xi} \sim \mathcal{C}(\boldsymbol{y}_{r,K})} \left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}) + (\boldsymbol{A}^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}) \right] \right\| \\
&\leq \left\| \left[\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) \right]^{-1} \right\| \left\| \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})} \left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) \right] \right. \\
&\quad \left. - \mathop{\mathbb{E}}_{\boldsymbol{\xi} \sim \mathcal{C}(\boldsymbol{y}_{r,K})} \left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}) + (\boldsymbol{A}^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}) \right] \right\| \\
&\quad + \left\| \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \left[\nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) \right]^{-1} - \mathop{\mathbb{E}}_{\boldsymbol{\zeta} \sim \mathcal{D}(\boldsymbol{x}_r)} \left[\nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}) \right]^{-1} \right\| \\
&\quad \cdot \mathop{\mathbb{E}}_{\boldsymbol{\xi} \sim \mathcal{C}(\boldsymbol{y}_{r,K})} \left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}) + (\boldsymbol{A}^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}) \right] \\
&\leq \frac{L_{fy}^\xi \|\boldsymbol{y}_{r,K}\| + \|\boldsymbol{A}^{(c)}\| L_f^\xi \|\boldsymbol{y}_{r,K}\| + C_f^\xi}{\gamma_y} \|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\| + \frac{\left(C_f^y + \|\boldsymbol{A}^{(c)}\| C_f^\xi \right) L_{gyy}^\zeta \|\boldsymbol{x}_r\|}{\gamma_y^2} \\
&\quad \cdot \|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\|. \qquad (55)
\end{aligned}
$$

Therefore, the bias between the Neumann estimator mean and the true inverse-Hessian target is

$$
\begin{aligned}
&\|\mathbb{E}[\boldsymbol{v}_{r,Q}] - \boldsymbol{v}_r^*\| \\
&= \|\mathbb{E}[\boldsymbol{v}_{r,Q}] - \widehat{\boldsymbol{v}}_r^* + \widehat{\boldsymbol{v}}_r^* - \boldsymbol{v}_r^*\| \\
&\leq \|\mathbb{E}[\boldsymbol{v}_{r,Q}] - \widehat{\boldsymbol{v}}_r^*\| + \|\widehat{\boldsymbol{v}}_r^* - \boldsymbol{v}_r^*\| \\
&\leq \frac{(1 - \eta\gamma_y)^{Q+1}}{\gamma_y} \left(C_f^y + \|\boldsymbol{A}^{(c)}\| C_f^\xi + \|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\| C_f^\xi \right) \\
&\quad + \frac{L_{fy}^\xi \|\boldsymbol{y}_{r,K}\| + \|\boldsymbol{A}^{(c)}\| L_f^\xi \|\boldsymbol{y}_{r,K}\| + C_f^\xi}{\gamma_y} \|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\| \\
&\quad + \frac{\left(C_f^y + \|\boldsymbol{A}^{(c)}\| C_f^\xi \right) L_{gyy}^\zeta \|\boldsymbol{x}_r\|}{\gamma_y^2} \|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\|. \qquad (56)
\end{aligned}
$$

We next bound the variance of $\boldsymbol{v}_{r,Q}$:

$$
\begin{aligned}
&\mathbb{E} \|\boldsymbol{v}_{r,Q} - \widehat{\boldsymbol{v}}_r^*\|^2 \\
&= \mathbb{E} \|\boldsymbol{v}_{r,Q} - \mathbb{E}[\boldsymbol{v}_{r,Q}] + \mathbb{E}[\boldsymbol{v}_{r,Q}] - \widehat{\boldsymbol{v}}_r^*\|^2 \\
&\leq 2\mathbb{E} \|\boldsymbol{v}_{r,Q} - \mathbb{E}[\boldsymbol{v}_{r,Q}]\|^2 + 2 \|\mathbb{E}[\boldsymbol{v}_{r,Q}] - \widehat{\boldsymbol{v}}_r^*\|^2 \\
&\leq 2\mathbb{E} \|\boldsymbol{v}_{r,Q} - \mathbb{E}[\boldsymbol{v}_{r,Q}]\|^2 + \frac{2(1 - \eta\gamma_y)^{2Q+2}}{\gamma_y^2} \left(C_f^y + \|\boldsymbol{A}^{(c)}\| C_f^\xi + \|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\| C_f^\xi \right)^2. \qquad (57)
\end{aligned}
$$

We next bound $\mathbb{E}\|\boldsymbol{v}_{r,Q} - \mathbb{E}[\boldsymbol{v}_{r,Q}]\|^2$:

$$
\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{v}_{r,Q} - \mathbb{E}[\boldsymbol{v}_{r,Q}]\right\|^2 \\
&= \mathbb{E}\left\|\eta\sum_{q=0}^{Q}\prod_{j=Q+1-q}^{Q}\left(I - \eta\nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_{r,j}^{(v)})\right)\boldsymbol{v}_{r,0} - \eta\sum_{q=0}^{Q}\left(I - \eta\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r)\right)^q\right. \\
&\qquad \left.\cdot \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})}\left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal\nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r)\right]\right\|^2 \\
&\le 2\eta^2\mathbb{E}\left\|\sum_{q=0}^{Q}\left[\prod_{j=Q+1-q}^{Q}\left(I - \eta\nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_{r,j}^{(v)})\right) - \left(I - \eta\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r)\right)^q\right]\right. \\
&\qquad \left.\cdot \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})}\left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal\nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r)\right]\right\|^2 \\
&\quad + 2\eta^2\mathbb{E}\left\|\sum_{q=0}^{Q}\prod_{j=Q+1-q}^{Q}\left(I - \eta\nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_{r,j}^{(v)})\right)\right. \\
&\qquad \left.\cdot\left(\boldsymbol{v}_{r,0} - \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})}\left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal\nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r)\right]\right)\right\|^2. \qquad (58)
\end{aligned}
$$

We first bound the second term in Eq. (58):

$$
\begin{aligned}
&2\eta^2\mathbb{E}\left\|\sum_{q=0}^{Q}\prod_{j=Q+1-q}^{Q}\left(I - \eta\nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_{r,j}^{(v)})\right)\right. \\
&\qquad \left.\cdot\left(\boldsymbol{v}_{r,0} - \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})}\left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal\nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r)\right]\right)\right\|^2 \\
&\le 2\eta^2\left(\sum_{q=0}^{Q}(1-\eta\gamma_y)^q\right)^2\mathbb{E}\left\|\boldsymbol{v}_{r,0} - \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})}\left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal\nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r)\right]\right\|^2 \\
&\le \frac{2}{\gamma_y^2}\mathbb{E}\left\|\boldsymbol{v}_{r,0} - \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})}\left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal\nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r)\right]\right\|^2 \\
&\le \frac{2\left(C_f^y + \|\boldsymbol{A}^{(c)}\|C_f^\xi + \|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\|C_f^\xi\right)^2}{\gamma_y^2|\mathcal{B}_r^{(f)}|}. \qquad (59)
\end{aligned}
$$

We next bound the first term in Eq. (58). For $q = 0, \ldots, Q$, define the vector second moment

$$
\begin{aligned}
M_q := \mathbb{E}&\left\|\left[\prod_{j=Q+1-q}^{Q}\left(I - \eta\nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_{r,j}, \mathcal{B}_{r,j}^{(v)})\right) - \left(I - \eta\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r)\right)^q\right]\right. \\
&\left.\cdot \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})}\left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}_r^{(c)})^\intercal\nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r)\right]\right\|^2. \qquad (60)
\end{aligned}
$$

Obviously, $M_0 = 0$. For $q \ge 1$, the mini-batches $\mathcal{B}_{r,j}^{(v)}$ are independently chosen, and

$$
\mathbb{E}\left[\nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_{r,j}, \mathcal{B}_{r,j}^{(v)}) - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\nabla_{yy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r)\right] = 0. \qquad (61)
$$

Therefore, expanding the squared Euclidean norm gives

$$\begin{aligned}
M_q \le & \left\| I - \eta \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \nabla^2_{yy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) \right\|^2 M_{q-1} \\
& + \eta^2 \mathbb{E} \Bigg\| \left(\nabla^2_{yy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_{r,Q+1-q}, \mathcal{B}^{(v)}_{r,Q+1-q}) - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \nabla^2_{yy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) \right) \\
& \cdot \prod_{j=Q+2-q}^{Q} \left(I - \eta \nabla^2_{yy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_{r,j}, \mathcal{B}^{(v)}_{r,j}) \right) \\
& \cdot \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})} \left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}^{(c)}_r)^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) \right] \Bigg\|^2 \\
\le & (1 - \eta\gamma_y)^2 M_{q-1} + \eta^2 (1 - \eta\gamma_y)^{2q-2} \frac{4 (C^{yy}_g)^2}{|\mathcal{B}^{(v)}_{r,Q+1-q}|} \left(C^y_f + \|\boldsymbol{A}^{(c)}\| C^\xi_f + \|\widehat{\boldsymbol{A}}^{(c)}_r - \boldsymbol{A}^{(c)}\| C^\xi_f \right)^2 .
\end{aligned} \tag{62}$$

Here we use Assumptions 2 and 4, which imply

$$\left\| I - \eta \nabla^2_{yy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_{r,j}, \mathcal{B}^{(v)}_{r,j}) \right\| \le 1 - \eta\gamma_y \tag{63}$$

for $0 < \eta \le 1/C^{yy}_g$, and

$$\begin{aligned}
& \left\| \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})} \left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\widehat{\boldsymbol{A}}^{(c)}_r)^\intercal \nabla_\xi f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) \right] \right\| \\
& \qquad \le C^y_f + \|\boldsymbol{A}^{(c)}\| C^\xi_f + \|\widehat{\boldsymbol{A}}^{(c)}_r - \boldsymbol{A}^{(c)}\| C^\xi_f .
\end{aligned} \tag{64}$$

By recursively applying Eq. (62) and using $M_0 = 0$, we obtain

$$M_q \le 4\eta^2 (C^{yy}_g)^2 (1 - \eta\gamma_y)^{2q-2} \sum_{i=1}^{q} \frac{\left(C^y_f + \|\boldsymbol{A}^{(c)}\| C^\xi_f + \|\widehat{\boldsymbol{A}}^{(c)}_r - \boldsymbol{A}^{(c)}\| C^\xi_f \right)^2}{|\mathcal{B}^{(v)}_{r,Q+1-i}|} . \tag{65}$$

Choose the integer batch sizes $|\mathcal{B}^{(v)}_{r,Q+1-j}| = \max\{1, \lceil BQ(1-\eta\gamma_y)^{j-1} \rceil\}$ for $j = 1, \ldots, Q$. Then $|\mathcal{B}^{(v)}_{r,Q+1-j}| \ge BQ(1-\eta\gamma_y)^{j-1}$, and for $q = 1, \ldots, Q$, we have

$$\begin{aligned}
M_q & \le 4\eta^2 (C^{yy}_g)^2 (1 - \eta\gamma_y)^{2q-2} \sum_{j=1}^{q} \frac{\left(C^y_f + \|\boldsymbol{A}^{(c)}\| C^\xi_f + \|\widehat{\boldsymbol{A}}^{(c)}_r - \boldsymbol{A}^{(c)}\| C^\xi_f \right)^2}{BQ(1-\eta\gamma_y)^{j-1}} \\
& = \frac{4\eta^2 (C^{yy}_g)^2}{BQ} (1 - \eta\gamma_y)^{2q-2} \left(C^y_f + \|\boldsymbol{A}^{(c)}\| C^\xi_f + \|\widehat{\boldsymbol{A}}^{(c)}_r - \boldsymbol{A}^{(c)}\| C^\xi_f \right)^2 \sum_{j=1}^{q} \left(\frac{1}{1-\eta\gamma_y} \right)^{j-1} \\
& \overset{(a)}{\le} \frac{4\eta^2 (C^{yy}_g)^2}{BQ} (1 - \eta\gamma_y)^{2q-2} \left(C^y_f + \|\boldsymbol{A}^{(c)}\| C^\xi_f + \|\widehat{\boldsymbol{A}}^{(c)}_r - \boldsymbol{A}^{(c)}\| C^\xi_f \right)^2 \frac{\left(\frac{1}{1-\eta\gamma_y} \right)^q}{\frac{1}{1-\eta\gamma_y} - 1} \\
& = \frac{4\eta (C^{yy}_g)^2}{(1-\eta\gamma_y)\gamma_y BQ} (1 - \eta\gamma_y)^q \left(C^y_f + \|\boldsymbol{A}^{(c)}\| C^\xi_f + \|\widehat{\boldsymbol{A}}^{(c)}_r - \boldsymbol{A}^{(c)}\| C^\xi_f \right)^2 ,
\end{aligned} \tag{66}$$

where we use the fact that $\sum_{j=1}^{q} \left(\frac{1}{1-\eta\gamma_y} \right)^{j-1} \le \frac{\left(\frac{1}{1-\eta\gamma_y} \right)^q}{\frac{1}{1-\eta\gamma_y} - 1}$ in (a), and based on the Cauchy-Schwarz inequality, we know

$$\begin{aligned}
\left(\sum_{q=1}^{Q} M_q^{1/2} \right)^2 & \le Q \sum_{q=1}^{Q} \frac{4\eta (C^{yy}_g)^2}{(1-\eta\gamma_y)\gamma_y BQ} (1 - \eta\gamma_y)^q \left(C^y_f + \|\boldsymbol{A}^{(c)}\| C^\xi_f + \|\widehat{\boldsymbol{A}}^{(c)}_r - \boldsymbol{A}^{(c)}\| C^\xi_f \right)^2 \\
& \le \frac{4 (C^{yy}_g)^2}{(1-\eta\gamma_y)\gamma_y^2 B} \left(C^y_f + \|\boldsymbol{A}^{(c)}\| C^\xi_f + \|\widehat{\boldsymbol{A}}^{(c)}_r - \boldsymbol{A}^{(c)}\| C^\xi_f \right)^2 .
\end{aligned} \tag{67}$$

Therefore, for the first term in Eq. (58), we obtain

$$2\eta^2\mathbb{E}\left\|\sum_{q=0}^{Q}\left[\prod_{j=Q+1-q}^{Q}\left(I-\eta\nabla^2_{yy}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r,\mathcal{B}^{(v)}_{r,j})\right)-\left(I-\eta\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\nabla^2_{yy}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)\right)^q\right]\right.$$
$$\left.\cdot\mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})}\left[\nabla_y f(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\xi}_r)+(\widehat{\boldsymbol{A}}^{(c)}_r)^\intercal\nabla_\xi f(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\xi}_r)\right]\right\|^2$$
$$\leq\frac{8\eta^2(C^{yy}_g)^2}{(1-\eta\gamma_y)\gamma_y^2B}\left(C^y_f+\|\boldsymbol{A}^{(c)}\|C^\xi_f+\|\widehat{\boldsymbol{A}}^{(c)}_r-\boldsymbol{A}^{(c)}\|C^\xi_f\right)^2. \tag{68}$$

Combining Eqs. (57), (59), and (68), we know

$$\mathbb{E}\left\|\boldsymbol{v}_{r,Q}-\widehat{\boldsymbol{v}}^*_r\right\|^2$$
$$\leq\left(\frac{16\eta^2(C^{yy}_g)^2}{(1-\eta\gamma_y)\gamma_y^2B}+\frac{4}{\gamma_y^2|\mathcal{B}^{(f)}_r|}+\frac{2(1-\eta\gamma_y)^{2Q+2}}{\gamma_y^2}\right)\left(C^y_f+\|\boldsymbol{A}^{(c)}\|C^\xi_f+\|\widehat{\boldsymbol{A}}^{(c)}_r-\boldsymbol{A}^{(c)}\|C^\xi_f\right)^2. \tag{69}$$

We further bound the error with respect to the true target $\boldsymbol{v}^*_r$. By the triangle inequality and the basic inequality $\|a+b\|^2\leq 2\|a\|^2+2\|b\|^2$,

$$\mathbb{E}\left\|\boldsymbol{v}_{r,Q}-\boldsymbol{v}^*_r\right\|^2=\mathbb{E}\left\|\boldsymbol{v}_{r,Q}-\widehat{\boldsymbol{v}}^*_r+\widehat{\boldsymbol{v}}^*_r-\boldsymbol{v}^*_r\right\|^2\leq 2\mathbb{E}\left\|\boldsymbol{v}_{r,Q}-\widehat{\boldsymbol{v}}^*_r\right\|^2+2\left\|\widehat{\boldsymbol{v}}^*_r-\boldsymbol{v}^*_r\right\|^2. \tag{70}$$

Combining Eq. (69) with Eq. (70) gives

$$\mathbb{E}\left\|\boldsymbol{v}_{r,Q}-\boldsymbol{v}^*_r\right\|^2\leq 2\left(\frac{16\eta^2(C^{yy}_g)^2}{(1-\eta\gamma_y)\gamma_y^2B}+\frac{4}{\gamma_y^2|\mathcal{B}^{(f)}_r|}+\frac{2(1-\eta\gamma_y)^{2Q+2}}{\gamma_y^2}\right)$$
$$\cdot\left(C^y_f+\|\boldsymbol{A}^{(c)}\|C^\xi_f+\|\widehat{\boldsymbol{A}}^{(c)}_r-\boldsymbol{A}^{(c)}\|C^\xi_f\right)^2+2\left\|\widehat{\boldsymbol{v}}^*_r-\boldsymbol{v}^*_r\right\|^2. \tag{71}$$

Moreover, from the estimation bias between $\widehat{\boldsymbol{v}}^*_r$ and $\boldsymbol{v}^*_r$, we have

$$\left\|\widehat{\boldsymbol{v}}^*_r-\boldsymbol{v}^*_r\right\|\leq\frac{L^\xi_f\|\boldsymbol{y}_{r,K}\|+\|\boldsymbol{A}^{(c)}\|L^\xi_f\|\boldsymbol{y}_{r,K}\|+C^\xi_f}{\gamma_y}\left\|\widehat{\boldsymbol{A}}^{(c)}_r-\boldsymbol{A}^{(c)}\right\|$$
$$+\frac{\left(C^y_f+\|\boldsymbol{A}^{(c)}\|C^\xi_f\right)L^\zeta_{gyy}\|\boldsymbol{x}_r\|}{\gamma_y^2}\left\|\widehat{\boldsymbol{A}}^{(d)}_r-\boldsymbol{A}^{(d)}\right\|. \tag{72}$$

Therefore,

$$\mathbb{E}\left\|\boldsymbol{v}_{r,Q}-\boldsymbol{v}^*_r\right\|^2\leq 2\left(\frac{16\eta^2(C^{yy}_g)^2}{(1-\eta\gamma_y)\gamma_y^2B}+\frac{4}{\gamma_y^2|\mathcal{B}^{(f)}_r|}+\frac{2(1-\eta\gamma_y)^{2Q+2}}{\gamma_y^2}\right)$$
$$\cdot\left(C^y_f+\|\boldsymbol{A}^{(c)}\|C^\xi_f+\|\widehat{\boldsymbol{A}}^{(c)}_r-\boldsymbol{A}^{(c)}\|C^\xi_f\right)^2$$
$$+4\left(\frac{L^\xi_f\|\boldsymbol{y}_{r,K}\|+\|\boldsymbol{A}^{(c)}\|L^\xi_f\|\boldsymbol{y}_{r,K}\|+C^\xi_f}{\gamma_y}\right)^2\left\|\widehat{\boldsymbol{A}}^{(c)}_r-\boldsymbol{A}^{(c)}\right\|^2$$
$$+4\left(\frac{\left(C^y_f+\|\boldsymbol{A}^{(c)}\|C^\xi_f\right)L^\zeta_{gyy}\|\boldsymbol{x}_r\|}{\gamma_y^2}\right)^2\left\|\widehat{\boldsymbol{A}}^{(d)}_r-\boldsymbol{A}^{(d)}\right\|^2. \tag{73}$$

□

D.2 First-Moment Error of the Hypergradient Estimator

We next bound the bias of the hypergradient estimator.

Lemma 5. *With Assumptions 1-4, we have*

$$
\begin{aligned}
&\left\|\mathbb{E}\widehat{\nabla}\Phi(\boldsymbol{x}_r)-\nabla\Phi(\boldsymbol{x}_r)\right\|^2 \\
&\le \frac{3(1-\eta\gamma_y)^{2Q+2}}{\gamma_y^2}\left(C_g^{xy}+\|\widehat{\boldsymbol{A}}_r^{(d)}\|C_g^{\zeta y}\right)^2\left(C_f^{y}+\|\widehat{\boldsymbol{A}}_r^{(c)}\|C_f^{\xi}\right)^2 \\
&\quad + 9(L_{fx}^{\xi})^2\|\boldsymbol{y}_{r,K}\|^2\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|^2 \\
&\quad + \frac{9}{\gamma_y^2}\left(L_{gxy}^{\zeta}\|\boldsymbol{x}_r\|+C_g^{\zeta y}+\|\boldsymbol{A}^{(d)}\|L_{g\zeta y}^{\zeta}\|\boldsymbol{x}_r\|\right)^2\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|^2\left(C_f^{y}+\|\widehat{\boldsymbol{A}}_r^{(c)}\|C_f^{\xi}\right)^2 \\
&\quad + 9\left(C_g^{xy}+\|\boldsymbol{A}^{(d)}\|C_g^{\zeta y}\right)^2\|\widehat{\boldsymbol{v}}_r^*-\boldsymbol{v}_r^*\|^2+\Bigg[9\left(\bar{L}_f^{x}+\|\boldsymbol{A}^{(c)}\|L_{fx}^{\xi}\right)^2 \\
&\quad + \frac{9\left(L_{gxy}^{y}+\|\boldsymbol{A}^{(d)}\|L_{g\zeta y}^{y}\right)^2}{\gamma_y^2}\left(C_f^{y}+\|\boldsymbol{A}^{(c)}\|C_f^{\xi}\right)^2+9\left(C_g^{xy}+\|\boldsymbol{A}^{(d)}\|C_g^{\zeta y}\right)^2 \\
&\quad \cdot\left(\frac{\bar{L}_f^{y}+\|\boldsymbol{A}^{(c)}\|L_{fy}^{\xi}+\|\boldsymbol{A}^{(c)}\|L_{f\xi}^{y}+\|\boldsymbol{A}^{(c)}\|^2L_f^{\xi}}{\gamma_y}+\frac{L_{gyy}^{y}\left(C_f^{y}+\|\boldsymbol{A}^{(c)}\|C_f^{\xi}\right)}{\gamma_y^2}\right)^2\Bigg] \\
&\quad \cdot\|\boldsymbol{y}_{r,K}-\boldsymbol{y}^*(\boldsymbol{x}_r)\|^2. \qquad (74)
\end{aligned}
$$

Proof. Since the LL distribution depends on $\boldsymbol{x}$, every cross derivative with respect to the UL variable is written as the chain rule derivative. We have

$$
\begin{aligned}
\widehat{\nabla}\Phi(\boldsymbol{x}_r) &= \nabla_x f(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\xi}_r,\mathcal{B}_r^{(f)}) \\
&\quad -\left(\nabla_{xy}^2 g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r,\mathcal{B}_r^{(gxy)})+(\widehat{\boldsymbol{A}}_r^{(d)})^{\intercal}\nabla_{\zeta y}^2 g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r,\mathcal{B}_r^{(gxy)})\right)\boldsymbol{v}_{r,Q}. \qquad (75)
\end{aligned}
$$

Define the estimated map target evaluated at $(\boldsymbol{x}_r,\boldsymbol{y}_{r,K})$ as

$$
\begin{aligned}
\widehat{\nabla}\Phi^*(\boldsymbol{x}_r,\boldsymbol{y}_{r,K}) &:= \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})}\nabla_x f(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\xi}_r) \\
&\quad -\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\left[\nabla_{xy}^2 g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)+(\widehat{\boldsymbol{A}}_r^{(d)})^{\intercal}\nabla_{\zeta y}^2 g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)\right]\widehat{\boldsymbol{v}}_r^*. \qquad (76)
\end{aligned}
$$

Similarly, define the true-map target evaluated at $(\boldsymbol{x}_r,\boldsymbol{y}_{r,K})$ as

$$
\begin{aligned}
\nabla\Phi^*(\boldsymbol{x}_r,\boldsymbol{y}_{r,K}) &:= \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\mathcal{C}(\boldsymbol{y}_{r,K})}\nabla_x f(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\xi}_r) \\
&\quad -\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\mathcal{D}(\boldsymbol{x}_r)}\left[\nabla_{xy}^2 g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)+(\boldsymbol{A}^{(d)})^{\intercal}\nabla_{\zeta y}^2 g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)\right]\boldsymbol{v}_r^*. \qquad (77)
\end{aligned}
$$

Taking conditional expectation in Eq. (75) gives

$$
\begin{aligned}
\mathbb{E}\widehat{\nabla}\Phi(\boldsymbol{x}_r) &= \widehat{\nabla}\Phi^*(\boldsymbol{x}_r,\boldsymbol{y}_{r,K}) \\
&\quad -\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\left[\nabla_{xy}^2 g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)+(\widehat{\boldsymbol{A}}_r^{(d)})^{\intercal}\nabla_{\zeta y}^2 g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)\right](\mathbb{E}[\boldsymbol{v}_{r,Q}]-\widehat{\boldsymbol{v}}_r^*). \qquad (78)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}\widehat{\nabla}\Phi(\boldsymbol{x}_r)-\nabla\Phi(\boldsymbol{x}_r) \\
&= -\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\left[\nabla_{xy}^2 g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)+(\widehat{\boldsymbol{A}}_r^{(d)})^{\intercal}\nabla_{\zeta y}^2 g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)\right](\mathbb{E}[\boldsymbol{v}_{r,Q}]-\widehat{\boldsymbol{v}}_r^*) \\
&\quad +\left[\widehat{\nabla}\Phi^*(\boldsymbol{x}_r,\boldsymbol{y}_{r,K})-\nabla\Phi^*(\boldsymbol{x}_r,\boldsymbol{y}_{r,K})\right]+\left[\nabla\Phi^*(\boldsymbol{x}_r,\boldsymbol{y}_{r,K})-\nabla\Phi(\boldsymbol{x}_r)\right]. \qquad (79)
\end{aligned}
$$

Thus,

$$\begin{aligned}
&\left\|\mathbb{E}\widehat{\nabla}\Phi(\boldsymbol{x}_r) - \nabla\Phi(\boldsymbol{x}_r)\right\|^2 \\
&\le 3\left\|\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\left[\nabla^2_{xy}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal\nabla^2_{\zeta y}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)\right](\mathbb{E}[\boldsymbol{v}_{r,Q}] - \widehat{\boldsymbol{v}}_r^*)\right\|^2 \\
&\quad + 3\left\|\widehat{\nabla}\Phi^*(\boldsymbol{x}_r,\boldsymbol{y}_{r,K}) - \nabla\Phi^*(\boldsymbol{x}_r,\boldsymbol{y}_{r,K})\right\|^2 + 3\left\|\nabla\Phi^*(\boldsymbol{x}_r,\boldsymbol{y}_{r,K}) - \nabla\Phi(\boldsymbol{x}_r)\right\|^2. \qquad (80)
\end{aligned}$$

We first bound the Neumann-truncation part. By Assumption 4,

$$\left\|\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\left[\nabla^2_{xy}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal\nabla^2_{\zeta y}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)\right]\right\| \le C_g^{xy} + \|\widehat{\boldsymbol{A}}_r^{(d)}\|C_g^{\zeta y}. \qquad (81)$$

Combining Eq. (81) with the conditional Neumann bias bound gives

$$\begin{aligned}
&3\left\|\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\left[\nabla^2_{xy}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal\nabla^2_{\zeta y}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)\right](\mathbb{E}[\boldsymbol{v}_{r,Q}] - \widehat{\boldsymbol{v}}_r^*)\right\|^2 \\
&\le \frac{3(1-\eta\gamma_y)^{2Q+2}}{\gamma_y^2}\left(C_g^{xy} + \|\widehat{\boldsymbol{A}}_r^{(d)}\|C_g^{\zeta y}\right)^2\left(C_f^{y} + \|\widehat{\boldsymbol{A}}_r^{(c)}\|C_f^{\xi}\right)^2. \qquad (82)
\end{aligned}$$

We then bound the distribution map perturbation part. From Eqs. (76)–(77),

$$\begin{aligned}
&\widehat{\nabla}\Phi^*(\boldsymbol{x}_r,\boldsymbol{y}_{r,K}) - \nabla\Phi^*(\boldsymbol{x}_r,\boldsymbol{y}_{r,K}) \\
&= \left[\mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})}\nabla_x f(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\xi}_r) - \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\mathcal{C}(\boldsymbol{y}_{r,K})}\nabla_x f(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\xi}_r)\right] \\
&\quad - \left[\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\left(\nabla^2_{xy}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal\nabla^2_{\zeta y}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)\right)\right. \\
&\quad \left. - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\mathcal{D}(\boldsymbol{x}_r)}\left(\nabla^2_{xy}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r) + (\boldsymbol{A}^{(d)})^\intercal\nabla^2_{\zeta y}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)\right)\right]\widehat{\boldsymbol{v}}_r^* \\
&\quad - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\mathcal{D}(\boldsymbol{x}_r)}\left[\nabla^2_{xy}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r) + (\boldsymbol{A}^{(d)})^\intercal\nabla^2_{\zeta y}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)\right](\widehat{\boldsymbol{v}}_r^* - \boldsymbol{v}_r^*). \qquad (83)
\end{aligned}$$

For the first term,

$$\left\|\mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})}\nabla_x f(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\xi}_r) - \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\mathcal{C}(\boldsymbol{y}_{r,K})}\nabla_x f(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\xi}_r)\right\| \le L_{fx}^{\xi}\|\boldsymbol{y}_{r,K}\|\left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\|. \qquad (84)$$

Using the same operation for the two LL distributions, the total cross derivative perturbation satisfies

$$\begin{aligned}
&\left\|\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\left[\nabla^2_{xy}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal\nabla^2_{\zeta y}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)\right]\right. \\
&\quad \left. - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\mathcal{D}(\boldsymbol{x}_r)}\left[\nabla^2_{xy}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r) + (\boldsymbol{A}^{(d)})^\intercal\nabla^2_{\zeta y}g(\boldsymbol{x}_r,\boldsymbol{y}_{r,K};\boldsymbol{\zeta}_r)\right]\right\| \\
&\le \left(L_{gxy}^{\zeta}\|\boldsymbol{x}_r\| + C_g^{\zeta y} + \|\boldsymbol{A}^{(d)}\|L_{g\zeta y}^{\zeta}\|\boldsymbol{x}_r\|\right)\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|. \qquad (85)
\end{aligned}$$

Moreover,

$$\|\widehat{\boldsymbol{v}}_r^*\| \le \frac{1}{\gamma_y}\left(C_f^{y} + \|\widehat{\boldsymbol{A}}_r^{(c)}\|C_f^{\xi}\right), \qquad (86)$$

$$\left\| \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \mathcal{D}(\boldsymbol{x}_r)} \left[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) + (\boldsymbol{A}^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) \right] \right\| \leq C_g^{xy} + \|\boldsymbol{A}^{(d)}\| C_g^{\zeta y}. \tag{87}$$

Therefore,

$$\begin{aligned}
&3 \left\| \widehat{\nabla} \Phi^*(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}) - \nabla \Phi^*(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}) \right\|^2 \\
&\leq 9 (L_{fx}^{\xi})^2 \|\boldsymbol{y}_{r,K}\|^2 \left\| \widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)} \right\|^2 \\
&\quad + \frac{9}{\gamma_y^2} \left(L_{gxy}^{\zeta} \|\boldsymbol{x}_r\| + C_g^{\zeta y} + \|\boldsymbol{A}^{(d)}\| L_{g\zeta y}^{\zeta} \|\boldsymbol{x}_r\| \right)^2 \left\| \widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)} \right\|^2 \left(C_f^y + \|\widehat{\boldsymbol{A}}_r^{(c)}\| C_f^{\xi} \right)^2 \\
&\quad + 9 \left(C_g^{xy} + \|\boldsymbol{A}^{(d)}\| C_g^{\zeta y} \right)^2 \|\widehat{\boldsymbol{v}}_r^* - \boldsymbol{v}_r^*\|^2 .
\end{aligned} \tag{88}$$

We now bound the LL tracking part. Since $\nabla \Phi(\boldsymbol{x}_r)$ is evaluated at $\boldsymbol{y}^*(\boldsymbol{x}_r)$, we have

$$\begin{aligned}
&\nabla \Phi^*(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}) - \nabla \Phi(\boldsymbol{x}_r) \\
&= \left[\mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \mathcal{C}(\boldsymbol{y}_{r,K})} \nabla_x f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) - \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}_r))} \nabla_x f(\boldsymbol{x}_r, \boldsymbol{y}^*(\boldsymbol{x}_r); \boldsymbol{\xi}_r) \right] \\
&\quad - \Bigg[\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \mathcal{D}(\boldsymbol{x}_r)} \left(\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) + (\boldsymbol{A}^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) \right) \\
&\quad - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \mathcal{D}(\boldsymbol{x}_r)} \left(\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}^*(\boldsymbol{x}_r); \boldsymbol{\zeta}_r) + (\boldsymbol{A}^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}^*(\boldsymbol{x}_r); \boldsymbol{\zeta}_r) \right) \Bigg] \boldsymbol{v}_r^* \\
&\quad - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \mathcal{D}(\boldsymbol{x}_r)} \left[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}^*(\boldsymbol{x}_r); \boldsymbol{\zeta}_r) + (\boldsymbol{A}^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}^*(\boldsymbol{x}_r); \boldsymbol{\zeta}_r) \right] (\boldsymbol{v}_r^* - \boldsymbol{v}^*(\boldsymbol{x}_r)) .
\end{aligned} \tag{89}$$

The first term is bounded by

$$\begin{aligned}
&\left\| \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \mathcal{C}(\boldsymbol{y}_{r,K})} \nabla_x f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) - \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}_r))} \nabla_x f(\boldsymbol{x}_r, \boldsymbol{y}^*(\boldsymbol{x}_r); \boldsymbol{\xi}_r) \right\| \\
&\leq \left(\bar{L}_f^x + \|\boldsymbol{A}^{(c)}\| L_{fx}^{\xi} \right) \|\boldsymbol{y}_{r,K} - \boldsymbol{y}^*(\boldsymbol{x}_r)\| .
\end{aligned} \tag{90}$$

For the total cross derivative, we have

$$\begin{aligned}
&\Bigg\| \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \mathcal{D}(\boldsymbol{x}_r)} \left[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) + (\boldsymbol{A}^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) \right] \\
&\quad - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \mathcal{D}(\boldsymbol{x}_r)} \left[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}^*(\boldsymbol{x}_r); \boldsymbol{\zeta}_r) + (\boldsymbol{A}^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}^*(\boldsymbol{x}_r); \boldsymbol{\zeta}_r) \right] \Bigg\| \\
&\leq \left(L_{gxy}^y + \|\boldsymbol{A}^{(d)}\| L_{g\zeta y}^y \right) \|\boldsymbol{y}_{r,K} - \boldsymbol{y}^*(\boldsymbol{x}_r)\| .
\end{aligned} \tag{91}$$

Also, we know that

$$\|\boldsymbol{v}_r^*\| \leq \frac{1}{\gamma_y} \left(C_f^y + \|\boldsymbol{A}^{(c)}\| C_f^{\xi} \right), \tag{92}$$

$$\begin{aligned}
&\Bigg\| \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \mathcal{C}(\boldsymbol{y}_{r,K})} \left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) + (\boldsymbol{A}^{(c)})^\intercal \nabla_{\xi} f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) \right] \\
&\quad - \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}_r))} \left[\nabla_y f(\boldsymbol{x}_r, \boldsymbol{y}^*(\boldsymbol{x}_r); \boldsymbol{\xi}_r) + (\boldsymbol{A}^{(c)})^\intercal \nabla_{\xi} f(\boldsymbol{x}_r, \boldsymbol{y}^*(\boldsymbol{x}_r); \boldsymbol{\xi}_r) \right] \Bigg\| \\
&\leq \left(\bar{L}_f^y + \|\boldsymbol{A}^{(c)}\| L_{fy}^{\xi} + \|\boldsymbol{A}^{(c)}\| L_{f\xi}^y + \|\boldsymbol{A}^{(c)}\|^2 L_f^{\xi} \right) \|\boldsymbol{y}_{r,K} - \boldsymbol{y}^*(\boldsymbol{x}_r)\| .
\end{aligned} \tag{93}$$

Using $H_1^{-1} - H_2^{-1} = H_1^{-1}(H_2 - H_1)H_2^{-1}$, we obtain

$$\|\boldsymbol{v}_r^* - \boldsymbol{v}^*(\boldsymbol{x}_r)\| \le \left[\frac{\bar{L}_f^y + \|\boldsymbol{A}^{(c)}\|L_{fy}^{\xi} + \|\boldsymbol{A}^{(c)}\|L_{f\xi}^{y} + \|\boldsymbol{A}^{(c)}\|^2 L_f^{\xi}}{\gamma_y} + \frac{L_{gyy}^{y}\left(C_f^y + \|\boldsymbol{A}^{(c)}\|C_f^{\xi}\right)}{\gamma_y^2}\right] \cdot \|\boldsymbol{y}_{r,K} - \boldsymbol{y}^*(\boldsymbol{x}_r)\|. \tag{94}$$

Combining the three lower-level terms yields

$$\begin{aligned} &3\left\|\nabla\Phi^*(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}) - \nabla\Phi(\boldsymbol{x}_r)\right\|^2 \\ &\le \left[9\left(\bar{L}_f^x + \|\boldsymbol{A}^{(c)}\|L_{fx}^{\xi}\right)^2 + \frac{9\left(L_{gxy}^{y} + \|\boldsymbol{A}^{(d)}\|L_{g\zeta y}^{y}\right)^2}{\gamma_y^2}\left(C_f^y + \|\boldsymbol{A}^{(c)}\|C_f^{\xi}\right)^2 + 9\left(C_g^{xy} + \|\boldsymbol{A}^{(d)}\|C_g^{\zeta y}\right)^2\right. \\ &\quad \left.\cdot\left(\frac{\bar{L}_f^y + \|\boldsymbol{A}^{(c)}\|L_{fy}^{\xi} + \|\boldsymbol{A}^{(c)}\|L_{f\xi}^{y} + \|\boldsymbol{A}^{(c)}\|^2 L_f^{\xi}}{\gamma_y} + \frac{L_{gyy}^{y}\left(C_f^y + \|\boldsymbol{A}^{(c)}\|C_f^{\xi}\right)}{\gamma_y^2}\right)^2\right] \|\boldsymbol{y}_{r,K} - \boldsymbol{y}^*(\boldsymbol{x}_r)\|^2. \end{aligned} \tag{95}$$

Finally, combining the preceding bounds with the bound on $\|\widehat{\boldsymbol{v}}_r^* - \boldsymbol{v}_r^*\|$, we have

$$\begin{aligned} &\left\|\mathbb{E}\widehat{\nabla}\Phi(\boldsymbol{x}_r) - \nabla\Phi(\boldsymbol{x}_r)\right\|^2 \\ &\le \frac{3(1-\eta\gamma_y)^{2Q+2}}{\gamma_y^2}\left(C_g^{xy} + \|\widehat{\boldsymbol{A}}_r^{(d)}\|C_g^{\zeta y}\right)^2\left(C_f^y + \|\widehat{\boldsymbol{A}}_r^{(c)}\|C_f^{\xi}\right)^2 \\ &\quad + 9(L_{fx}^{\xi})^2\|\boldsymbol{y}_{r,K}\|^2\left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\|^2 \\ &\quad + \frac{9}{\gamma_y^2}\left(L_{gxy}^{\zeta}\|\boldsymbol{x}_r\| + C_g^{\zeta y} + \|\boldsymbol{A}^{(d)}\|L_{g\zeta y}^{\zeta}\|\boldsymbol{x}_r\|\right)^2\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^2\left(C_f^y + \|\widehat{\boldsymbol{A}}_r^{(c)}\|C_f^{\xi}\right)^2 \\ &\quad + 9\left(C_g^{xy} + \|\boldsymbol{A}^{(d)}\|C_g^{\zeta y}\right)^2\|\widehat{\boldsymbol{v}}_r^* - \boldsymbol{v}_r^*\|^2 + \left[9\left(\bar{L}_f^x + \|\boldsymbol{A}^{(c)}\|L_{fx}^{\xi}\right)^2\right. \\ &\quad + \frac{9\left(L_{gxy}^{y} + \|\boldsymbol{A}^{(d)}\|L_{g\zeta y}^{y}\right)^2}{\gamma_y^2}\left(C_f^y + \|\boldsymbol{A}^{(c)}\|C_f^{\xi}\right)^2 + 9\left(C_g^{xy} + \|\boldsymbol{A}^{(d)}\|C_g^{\zeta y}\right)^2 \\ &\quad \left.\cdot\left(\frac{\bar{L}_f^y + \|\boldsymbol{A}^{(c)}\|L_{fy}^{\xi} + \|\boldsymbol{A}^{(c)}\|L_{f\xi}^{y} + \|\boldsymbol{A}^{(c)}\|^2 L_f^{\xi}}{\gamma_y} + \frac{L_{gyy}^{y}\left(C_f^y + \|\boldsymbol{A}^{(c)}\|C_f^{\xi}\right)}{\gamma_y^2}\right)^2\right] \\ &\quad \cdot\|\boldsymbol{y}_{r,K} - \boldsymbol{y}^*(\boldsymbol{x}_r)\|^2. \end{aligned} \tag{96}$$

□

D.3 Proof of Lemma 1

We next bound $\mathbb{E}\|\widehat{\nabla}\Phi(\boldsymbol{x}_r) - \nabla\Phi(\boldsymbol{x}_r)\|^2$. We have

$$\mathbb{E}\left\|\widehat{\nabla}\Phi(\boldsymbol{x}_r) - \nabla\Phi(\boldsymbol{x}_r)\right\|^2 \le 2\mathbb{E}\left\|\widehat{\nabla}\Phi(\boldsymbol{x}_r) - \mathbb{E}\widehat{\nabla}\Phi(\boldsymbol{x}_r)\right\|^2 + 2\left\|\mathbb{E}\widehat{\nabla}\Phi(\boldsymbol{x}_r) - \nabla\Phi(\boldsymbol{x}_r)\right\|^2. \tag{97}$$

The second term is controlled by Eq. (96) and it remains to bound the variance part. From Eq. (75),

$$\begin{aligned} \mathbb{E}\widehat{\nabla}\Phi(\boldsymbol{x}_r) = &\mathop{\mathbb{E}}_{\boldsymbol{\xi}_r\sim\widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})}\nabla_x f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r) \\ &- \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}\left[\nabla_{xy}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) + (\widehat{\boldsymbol{A}}_r^{(d)})^{\intercal}\nabla_{\zeta y}^2 g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r)\right]\mathbb{E}[\boldsymbol{v}_{r,Q}]. \end{aligned} \tag{98}$$

Therefore,

$$
\begin{aligned}
&\widehat{\nabla}\Phi(\boldsymbol{x}_r) - \mathbb{E}\widehat{\nabla}\Phi(\boldsymbol{x}_r) \\
&= \left[\nabla_x f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r, \mathcal{B}_r^{(f)}) - \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})} \nabla_x f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r)\right] \\
&\quad - \Bigg[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_r^{(gxy)}) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_r^{(gxy)}) \\
&\quad - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \left[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r)\right]\Bigg] \boldsymbol{v}_{r,Q} \\
&\quad - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \left[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r)\right] (\boldsymbol{v}_{r,Q} - \mathbb{E}[\boldsymbol{v}_{r,Q}])\,. \qquad (99)
\end{aligned}
$$

By $\|\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}\|^2 \le 3\|\boldsymbol{a}\|^2 + 3\|\boldsymbol{b}\|^2 + 3\|\boldsymbol{c}\|^2$, we have

$$
\begin{aligned}
&2\mathbb{E}\left\|\widehat{\nabla}\Phi(\boldsymbol{x}_r) - \mathbb{E}\widehat{\nabla}\Phi(\boldsymbol{x}_r)\right\|^2 \\
&\le 6\mathbb{E}\left\|\nabla_x f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r, \mathcal{B}_r^{(f)}) - \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})} \nabla_x f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r)\right\|^2 \\
&\quad + 6\mathbb{E}\Bigg\|\Bigg[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_r^{(gxy)}) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_r^{(gxy)}) \\
&\quad - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \left[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r)\right]\Bigg] \boldsymbol{v}_{r,Q}\Bigg\|^2 \\
&\quad + 6\mathbb{E}\left\|\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \left[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r)\right] (\boldsymbol{v}_{r,Q} - \mathbb{E}[\boldsymbol{v}_{r,Q}])\right\|^2. \qquad (100)
\end{aligned}
$$

The first term is

$$
6\mathbb{E}\left\|\nabla_x f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r, \mathcal{B}_r^{(f)}) - \mathop{\mathbb{E}}_{\boldsymbol{\xi}_r \sim \widehat{\mathcal{C}}(\boldsymbol{y}_{r,K})} \nabla_x f(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\xi}_r)\right\|^2 \le \frac{24(C_f^x)^2}{|\mathcal{B}_r^{(f)}|}. \qquad (101)
$$

For the second term, since $\mathcal{B}_r^{(gxy)}$ is independent of the mini-batches used to construct $\boldsymbol{v}_{r,Q}$,

$$
\begin{aligned}
&\mathbb{E}\Bigg\|\Bigg[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_r^{(gxy)}) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_r^{(gxy)}) \\
&\quad - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \left[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r)\right]\Bigg] \boldsymbol{v}_{r,Q}\Bigg\|^2 \\
&\le \frac{4\left(C_g^{xy} + \|\widehat{\boldsymbol{A}}_r^{(d)}\| C_g^{\zeta y}\right)^2}{|\mathcal{B}_r^{(gxy)}|} \mathbb{E}\|\boldsymbol{v}_{r,Q}\|^2. \qquad (102)
\end{aligned}
$$

Moreover,

$$
\|\boldsymbol{v}_{r,Q}\| \le \frac{1}{\gamma_y}\left(C_f^y + \|\widehat{\boldsymbol{A}}_r^{(c)}\| C_f^{\xi}\right). \qquad (103)
$$

Thus,

$$
\begin{aligned}
&6\mathbb{E}\Bigg\| \Big[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_r^{(gxy)}) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r, \mathcal{B}_r^{(gxy)}) \\
&- \mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \Big[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}_r)\Big] \Big] \boldsymbol{v}_{r,Q} \Bigg\|^2 \\
&\le \frac{24\left(C_g^{xy} + \|\widehat{\boldsymbol{A}}_r^{(d)}\| C_g^{\zeta y}\right)^2}{\gamma_y^2 |\mathcal{B}_r^{(gxy)}|} \left(C_f^y + \|\widehat{\boldsymbol{A}}_r^{(c)}\| C_f^{\xi}\right)^2 .
\end{aligned} \tag{104}
$$

For the third term, Eq. (81) and the variance bound for $\boldsymbol{v}_{r,Q}$ imply

$$
\begin{aligned}
&6\mathbb{E}\left\| \mathop{\mathbb{E}}_{\boldsymbol{\zeta} \sim \widehat{\mathcal{D}}(\boldsymbol{x}_r)} \Big[\nabla^2_{xy} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta}) + (\widehat{\boldsymbol{A}}_r^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}_r, \boldsymbol{y}_{r,K}; \boldsymbol{\zeta})\Big] (\boldsymbol{v}_{r,Q} - \mathbb{E}[\boldsymbol{v}_{r,Q}]) \right\|^2 \\
&\le \left(\frac{48\eta^2 \left(C_g^{xy} + \|\widehat{\boldsymbol{A}}_r^{(d)}\| C_g^{\zeta y}\right)^2 (C_g^{yy})^2}{(1-\eta\gamma_y)\gamma_y^2 B} + \frac{12\left(C_g^{xy} + \|\widehat{\boldsymbol{A}}_r^{(d)}\| C_g^{\zeta y}\right)^2}{\gamma_y^2 |\mathcal{B}_r^{(f)}|} \right) \left(C_f^y + \|\widehat{\boldsymbol{A}}_r^{(c)}\| C_f^{\xi}\right)^2 .
\end{aligned} \tag{105}
$$

Combining Eqs. (101), (104), and (105) gives the variance bound. Combining it with Eq. (96) yields

$$
\begin{aligned}
&\mathbb{E}\left\| \widehat{\nabla}\Phi(\boldsymbol{x}_r) - \nabla\Phi(\boldsymbol{x}_r) \right\|^2 \\
&\le \frac{24 (C_f^x)^2}{|\mathcal{B}_r^{(f)}|} + \frac{24\left(C_g^{xy} + \|\widehat{\boldsymbol{A}}_r^{(d)}\| C_g^{\zeta y}\right)^2}{\gamma_y^2 |\mathcal{B}_r^{(gxy)}|} \left(C_f^y + \|\widehat{\boldsymbol{A}}_r^{(c)}\| C_f^{\xi}\right)^2 \\
&\quad + \left(\frac{48\eta^2 \left(C_g^{xy} + \|\widehat{\boldsymbol{A}}_r^{(d)}\| C_g^{\zeta y}\right)^2 (C_g^{yy})^2}{(1-\eta\gamma_y)\gamma_y^2 B} + \frac{12\left(C_g^{xy} + \|\widehat{\boldsymbol{A}}_r^{(d)}\| C_g^{\zeta y}\right)^2}{\gamma_y^2 |\mathcal{B}_r^{(f)}|} \right) \left(C_f^y + \|\widehat{\boldsymbol{A}}_r^{(c)}\| C_f^{\xi}\right)^2 \\
&\quad + 2\left\| \mathbb{E}\widehat{\nabla}\Phi(\boldsymbol{x}_r) - \nabla\Phi(\boldsymbol{x}_r) \right\|^2 .
\end{aligned} \tag{106}
$$

We now take the expectation in Eq. (106). By Lemma 3 and $\|\boldsymbol{M}\| \le \|\boldsymbol{M}\|_F$, the second and fourth moments of the two distribution map estimation errors are uniformly bounded. Therefore,

$$
\mathbb{E}\Big[\left(C_g^{xy} + \|\widehat{\boldsymbol{A}}_r^{(d)}\| C_g^{\zeta y}\right)^2 \left(C_f^y + \|\widehat{\boldsymbol{A}}_r^{(c)}\| C_f^{\xi}\right)^2 \Big] = \mathcal{O}(1). \tag{107}
$$

Using

$$
\|\widehat{\boldsymbol{A}}_r^{(d)}\|^2 \le 2\|\boldsymbol{A}^{(d)}\|^2 + 2\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\|^2
$$

and that for $\widehat{\boldsymbol{A}}_r^{(c)}$, the mixed term satisfies

$$
\mathbb{E}\left[\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^2 \left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\|^2 \right] \le \sqrt{\mathbb{E}\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^4} \sqrt{\mathbb{E}\left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\|^4} = \mathcal{O}(1). \tag{108}
$$

We next consider the distribution map estimation terms in Eq. (96). The UL distribution map perturbation is bounded by $\mathcal{O}\left(\mathbb{E}\left[(1+\|\boldsymbol{y}_{r,K}\|^2)\left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\|^2\right]\right)$. For the LL distribution map perturbation, using $\left(C_f^y + \|\widehat{\boldsymbol{A}}_r^{(c)}\| C_f^{\xi}\right)^2 \le C\left(1 + \left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\|^2\right)$ for a positive constant C, we

have

$$\mathbb{E}\Big[(1+\|\boldsymbol{x}_r\|^2)\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|^2\left(1+\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|^2\right)\Big]$$
$$\leq \mathbb{E}\left[(1+\|\boldsymbol{x}_r\|^2)\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|^2\right]+\sqrt{\mathbb{E}\left[(1+\|\boldsymbol{x}_r\|^2)^2\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|^4\right]}\sqrt{\mathbb{E}\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|^4}. \tag{109}$$

Moreover, from the distribution map perturbation bound on $\widehat{\boldsymbol{v}}_r^*-\boldsymbol{v}_r^*$ derived in Appendix D.1,

$$\|\widehat{\boldsymbol{v}}_r^*-\boldsymbol{v}_r^*\|^2\leq\mathcal{O}\Big((1+\|\boldsymbol{y}_{r,K}\|^2)\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|^2+\|\boldsymbol{x}_r\|^2\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|^2\Big). \tag{110}$$

Therefore, taking the expectation in Eq. (96) gives

$$\begin{aligned}
&2\mathbb{E}\left\|\mathbb{E}\widehat{\nabla}\Phi(\boldsymbol{x}_r)-\nabla\Phi(\boldsymbol{x}_r)\right\|^2\\
&\leq\mathcal{H}_y\mathbb{E}\left\|\boldsymbol{y}_{r,K}-\boldsymbol{y}^*(\boldsymbol{x}_r)\right\|^2+\mathcal{O}\left((1-\eta\gamma_y)^{2Q+2}\right)\\
&\quad+\mathcal{O}\Bigg(\mathbb{E}\left[(1+\|\boldsymbol{y}_{r,K}\|^2)\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|^2\right]+\mathbb{E}\left[(1+\|\boldsymbol{x}_r\|^2)\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|^2\right]\\
&\qquad+\sqrt{\mathbb{E}\left[(1+\|\boldsymbol{x}_r\|^2)^2\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|^4\right]}\sqrt{\mathbb{E}\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|^4}\Bigg).
\end{aligned} \tag{111}$$

Similarly, taking the expectation of the variance terms in Eq. (106) and using Eq. (107) gives

$$2\mathbb{E}\left\|\widehat{\nabla}\Phi(\boldsymbol{x}_r)-\mathbb{E}\widehat{\nabla}\Phi(\boldsymbol{x}_r)\right\|^2=\mathcal{O}\left(\frac{1}{B}+\frac{1}{|\mathcal{B}_r^{(f)}|}+\frac{1}{|\mathcal{B}_r^{(gxy)}|}\right). \tag{112}$$

Combining Eqs. (97), (111), and (112) gives

$$\begin{aligned}
&\mathbb{E}\left\|\widehat{\nabla}\Phi(\boldsymbol{x}_r)-\nabla\Phi(\boldsymbol{x}_r)\right\|^2\\
&\leq\mathcal{H}_y\mathbb{E}\left\|\boldsymbol{y}_{r,K}-\boldsymbol{y}^*(\boldsymbol{x}_r)\right\|^2+\mathcal{O}\left((1-\eta\gamma_y)^{2Q+2}+\frac{1}{B}+\frac{1}{|\mathcal{B}_r^{(f)}|}+\frac{1}{|\mathcal{B}_r^{(gxy)}|}\right)\\
&\quad+\mathcal{O}\Bigg(\mathbb{E}\left[(1+\|\boldsymbol{y}_{r,K}\|^2)\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|^2\right]+\mathbb{E}\left[(1+\|\boldsymbol{x}_r\|^2)\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|^2\right]\\
&\qquad+\sqrt{\mathbb{E}\left[(1+\|\boldsymbol{x}_r\|^2)^2\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|^4\right]}\sqrt{\mathbb{E}\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|^4}\Bigg).
\end{aligned} \tag{113}$$

Equivalently, Eq. (128) holds with $\mathcal{V}_r^{\Phi}$ denoting other terms except $\mathcal{H}_y\mathbb{E}\left\|\boldsymbol{y}_{r,K}-\boldsymbol{y}^*(\boldsymbol{x}_r)\right\|^2$.

E Proof of Lemma 2

We next bound the LL tracking error $\mathbb{E}\|\boldsymbol{y}_{r,K}-\boldsymbol{y}^*(\boldsymbol{x}_r)\|^2$. Different from the standard stocBiO analysis, the LL update is performed under the estimated distribution map $\widehat{\mathcal{D}}$, while $\boldsymbol{y}^*(\boldsymbol{x}_r)$ is defined under the true distribution map $\mathcal{D}$. Let

$$\widehat{\boldsymbol{y}}^*(\boldsymbol{x}_r)\in\arg\min_{\boldsymbol{y}}\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)}g(\boldsymbol{x}_r,\boldsymbol{y};\boldsymbol{\zeta}_r). \tag{114}$$

By the triangle inequality,

$$\mathbb{E}\left\|\boldsymbol{y}_{r,K}-\boldsymbol{y}^*(\boldsymbol{x}_r)\right\|^2\leq 2\mathbb{E}\left\|\boldsymbol{y}_{r,K}-\widehat{\boldsymbol{y}}^*(\boldsymbol{x}_r)\right\|^2+2\mathbb{E}\left\|\widehat{\boldsymbol{y}}^*(\boldsymbol{x}_r)-\boldsymbol{y}^*(\boldsymbol{x}_r)\right\|^2. \tag{115}$$

For the second term in Eq. (115), the optimality conditions give

$$\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\widehat{\mathcal{D}}(\boldsymbol{x}_r)} \nabla_y g(\boldsymbol{x}_r, \widehat{\boldsymbol{y}}^*(\boldsymbol{x}_r); \boldsymbol{\zeta}_r) = 0, \tag{116}$$

$$\mathop{\mathbb{E}}_{\boldsymbol{\zeta}_r\sim\mathcal{D}(\boldsymbol{x}_r)} \nabla_y g(\boldsymbol{x}_r, \boldsymbol{y}^*(\boldsymbol{x}_r); \boldsymbol{\zeta}_r) = 0. \tag{117}$$

With the strong convexity of $g(\cdot)$, Assumption 3, and Proposition 1, we have

$$\|\widehat{\boldsymbol{y}}^*(\boldsymbol{x}_r) - \boldsymbol{y}^*(\boldsymbol{x}_r)\| \le \frac{L_{gy}^{\zeta}}{\gamma_y} W_1\left(\widehat{\mathcal{D}}(\boldsymbol{x}_r), \mathcal{D}(\boldsymbol{x}_r)\right) \le \frac{L_{gy}^{\zeta}\|\boldsymbol{x}_r\|}{\gamma_y}\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|. \tag{118}$$

Therefore,

$$\mathbb{E}\,\|\widehat{\boldsymbol{y}}^*(\boldsymbol{x}_r) - \boldsymbol{y}^*(\boldsymbol{x}_r)\|^2 \le \frac{(L_{gy}^{\zeta})^2}{\gamma_y^2}\mathbb{E}\left[\|\boldsymbol{x}_r\|^2\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^2\right]. \tag{119}$$

Next, we analyze the first term in Eq. (115). The LL update is

$$\boldsymbol{y}_{r,k+1} = \boldsymbol{y}_{r,k} - \beta\nabla_y g(\boldsymbol{x}_r, \boldsymbol{y}_{r,k}; \boldsymbol{\zeta}_r, \mathcal{B}_{r,k}^{(gy)}). \tag{120}$$

Therefore, by the standard SGD recursion under Assumptions 2 and 4, if $0 < \beta \le c_\beta / C_g^{yy}$ for a fixed $0 < c_\beta < 1$, then

$$\mathbb{E}\,\|\boldsymbol{y}_{r,k+1} - \widehat{\boldsymbol{y}}^*(\boldsymbol{x}_r)\|^2 \le (1-\beta\gamma_y)\mathbb{E}\,\|\boldsymbol{y}_{r,k} - \widehat{\boldsymbol{y}}^*(\boldsymbol{x}_r)\|^2 + \frac{4\beta^2(C_g^y)^2}{|\mathcal{B}_{r,k}^{(gy)}|}. \tag{121}$$

Telescoping Eq. (121) over k and choosing the same batch size at each inner step gives

$$\mathbb{E}\,\|\boldsymbol{y}_{r,K} - \widehat{\boldsymbol{y}}^*(\boldsymbol{x}_r)\|^2 \le (1-\beta\gamma_y)^K\mathbb{E}\,\|\boldsymbol{y}_{r,0} - \widehat{\boldsymbol{y}}^*(\boldsymbol{x}_r)\|^2 + \frac{4\beta(C_g^y)^2}{\gamma_y|\mathcal{B}_r^{(gy)}|}. \tag{122}$$

With the use of warm start, we have $\boldsymbol{y}_{r,0} = \boldsymbol{y}_{r-1,K}$. We compare the warm-start point with the current estimated optimizer through the true LL response map:

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{y}_{r,0} - \widehat{\boldsymbol{y}}^*(\boldsymbol{x}_r)\|^2 \le{}& 3\mathbb{E}\,\|\boldsymbol{y}_{r-1,K} - \boldsymbol{y}^*(\boldsymbol{x}_{r-1})\|^2 + 3\mathbb{E}\,\|\boldsymbol{y}^*(\boldsymbol{x}_{r-1}) - \boldsymbol{y}^*(\boldsymbol{x}_r)\|^2 \\ &+ 3\mathbb{E}\,\|\boldsymbol{y}^*(\boldsymbol{x}_r) - \widehat{\boldsymbol{y}}^*(\boldsymbol{x}_r)\|^2.\end{aligned} \tag{123}$$

By the Lipschitz continuity of the true LL response map,

$$\|\boldsymbol{y}^*(\boldsymbol{x}_{r-1}) - \boldsymbol{y}^*(\boldsymbol{x}_r)\| \le \bar{L}_y\,\|\boldsymbol{x}_{r-1} - \boldsymbol{x}_r\|, \tag{124}$$

where $\bar{L}_y$ is defined in Eq. (135). Moreover, the UL update gives

$$\boldsymbol{x}_r - \boldsymbol{x}_{r-1} = -\alpha\widehat{\nabla}\Phi(\boldsymbol{x}_{r-1}). \tag{125}$$

Therefore,

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{x}_r - \boldsymbol{x}_{r-1}\|^2 &= \alpha^2\mathbb{E}\left\|\widehat{\nabla}\Phi(\boldsymbol{x}_{r-1})\right\|^2 \\ &\le 2\alpha^2\mathbb{E}\,\|\nabla\Phi(\boldsymbol{x}_{r-1})\|^2 + 2\alpha^2\mathbb{E}\left\|\widehat{\nabla}\Phi(\boldsymbol{x}_{r-1}) - \nabla\Phi(\boldsymbol{x}_{r-1})\right\|^2.\end{aligned} \tag{126}$$

Define

$$\begin{aligned}\mathcal{H}_y :={}& 18\left(\bar{L}_f^x + \|\boldsymbol{A}^{(c)}\|L_{fx}^{\xi}\right)^2 + \frac{18\left(L_{gxy}^y + \|\boldsymbol{A}^{(d)}\|L_{g\zeta y}^y\right)^2}{\gamma_y^2}\left(C_f^y + \|\boldsymbol{A}^{(c)}\|C_f^{\xi}\right)^2 \\ &+ 18\left(C_g^{xy} + \|\boldsymbol{A}^{(d)}\|C_g^{\zeta y}\right)^2 \\ &\cdot\left(\frac{\bar{L}_f^y + \|\boldsymbol{A}^{(c)}\|L_{fy}^{\xi} + \|\boldsymbol{A}^{(c)}\|L_{f\xi}^y + \|\boldsymbol{A}^{(c)}\|^2 L_f^{\xi}}{\gamma_y} + \frac{L_{gyy}^y\left(C_f^y + \|\boldsymbol{A}^{(c)}\|C_f^{\xi}\right)}{\gamma_y^2}\right)^2.\end{aligned} \tag{127}$$

With $\mathcal{V}_r^{\Phi}$ defined after Eq. (113), we know

$$\mathbb{E}\left\|\widehat{\nabla}\Phi(\boldsymbol{x}_r)-\nabla\Phi(\boldsymbol{x}_r)\right\|^2 \le \mathbb{E}\left[\mathcal{V}_r^{\Phi}\right]+\mathcal{H}_y\mathbb{E}\left\|\boldsymbol{y}_{r,K}-\boldsymbol{y}^*(\boldsymbol{x}_r)\right\|^2. \tag{128}$$

Combining Eqs. (123), (119), (126), and (128) at iteration $r-1$, we obtain

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{y}_{r,0}-\widehat{\boldsymbol{y}}^*(\boldsymbol{x}_r)\right\|^2\\
&\le \left(3+6\alpha^2\bar{L}_y^2\mathcal{H}_y\right)\mathbb{E}\left\|\boldsymbol{y}_{r-1,K}-\boldsymbol{y}^*(\boldsymbol{x}_{r-1})\right\|^2+6\alpha^2\bar{L}_y^2\mathbb{E}\left\|\nabla\Phi(\boldsymbol{x}_{r-1})\right\|^2+6\alpha^2\bar{L}_y^2\mathbb{E}\left[\mathcal{V}_{r-1}^{\Phi}\right]\\
&\quad+3\frac{(L_{gy}^{\zeta})^2}{\gamma_y^2}\mathbb{E}\left[\|\boldsymbol{x}_r\|^2\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|^2\right].
\end{aligned} \tag{129}$$

Substituting Eq. (129) into Eq. (122) and then using Eq. (115) yields

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{y}_{r,K}-\boldsymbol{y}^*(\boldsymbol{x}_r)\right\|^2\\
&\le \left[6(1-\beta\gamma_y)^K+12\alpha^2(1-\beta\gamma_y)^K\bar{L}_y^2\mathcal{H}_y\right]\mathbb{E}\left\|\boldsymbol{y}_{r-1,K}-\boldsymbol{y}^*(\boldsymbol{x}_{r-1})\right\|^2\\
&\quad+12\alpha^2(1-\beta\gamma_y)^K\bar{L}_y^2\mathbb{E}\left\|\nabla\Phi(\boldsymbol{x}_{r-1})\right\|^2+12\alpha^2(1-\beta\gamma_y)^K\bar{L}_y^2\mathbb{E}\left[\mathcal{V}_{r-1}^{\Phi}\right]+\frac{8\beta(C_g^y)^2}{\gamma_y|\mathcal{B}_r^{(gy)}|}\\
&\quad+\left[6(1-\beta\gamma_y)^K+2\right]\frac{(L_{gy}^{\zeta})^2}{\gamma_y^2}\mathbb{E}\left[\|\boldsymbol{x}_r\|^2\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|^2\right].
\end{aligned} \tag{130}$$

Let

$$\lambda_y := 6(1-\beta\gamma_y)^K+12\alpha^2(1-\beta\gamma_y)^K\bar{L}_y^2\mathcal{H}_y. \tag{131}$$

Choosing K sufficiently large and α sufficiently small so that $\lambda_y<1$ ensures contraction of the LL tracking error across outer iterations.

Recursively applying Eq. (130) from iteration 1 to iteration r gives

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{y}_{r,K}-\boldsymbol{y}^*(\boldsymbol{x}_r)\right\|^2\\
&\le \lambda_y^r\mathbb{E}\left\|\boldsymbol{y}_{0,K}-\boldsymbol{y}^*(\boldsymbol{x}_0)\right\|^2+12\alpha^2(1-\beta\gamma_y)^K\bar{L}_y^2\sum_{s=1}^{r}\lambda_y^{r-s}\mathbb{E}\left\|\nabla\Phi(\boldsymbol{x}_{s-1})\right\|^2\\
&\quad+12\alpha^2(1-\beta\gamma_y)^K\bar{L}_y^2\sum_{s=1}^{r}\lambda_y^{r-s}\mathbb{E}\left[\mathcal{V}_{s-1}^{\Phi}\right]+\sum_{s=1}^{r}\lambda_y^{r-s}\frac{8\beta(C_g^y)^2}{\gamma_y|\mathcal{B}_s^{(gy)}|}\\
&\quad+\left[6(1-\beta\gamma_y)^K+2\right]\frac{(L_{gy}^{\zeta})^2}{\gamma_y^2}\sum_{s=1}^{r}\lambda_y^{r-s}\mathbb{E}\left[\|\boldsymbol{x}_s\|^2\left\|\widehat{\boldsymbol{A}}_s^{(d)}-\boldsymbol{A}^{(d)}\right\|^2\right].
\end{aligned} \tag{132}$$

F Smoothness of the Performative Objective

We next give the lemma of the smoothness property of the UL performative objective.

Lemma 6 (Smoothness of the performative bilevel objective)**.** *Under Assumptions 1–4, the objective $\Phi(\boldsymbol{x})$ has Lipschitz-continuous gradient on the region specified in Assumptions 3–4. That is, for any $\boldsymbol{x},\boldsymbol{x}'$ in this region, there exists a constant $L_\Phi>0$ such that*

$$\left\|\nabla\Phi(\boldsymbol{x})-\nabla\Phi(\boldsymbol{x}')\right\| \le L_\Phi\left\|\boldsymbol{x}-\boldsymbol{x}'\right\|, \tag{133}$$

where L_Φ is

$$
\begin{aligned}
&L_\Phi \\
&= L_f^x + \left(\bar{L}_f^x + L_{fx}^\xi \sigma_{\max}(\boldsymbol{A}^{(c)})\right) \bar{L}_y + \frac{C_f^y + \sigma_{\max}(\boldsymbol{A}^{(c)}) C_f^\xi}{\gamma_y} \Big(L_{gxy}^x + \sigma_{\max}(\boldsymbol{A}^{(d)}) L_{g\zeta y}^x \\
&\quad + \left(L_{gxy}^\zeta + \sigma_{\max}(\boldsymbol{A}^{(d)}) L_{g\zeta y}^\zeta\right) \sigma_{\max}(\boldsymbol{A}^{(d)}) + \left(L_{gxy}^y + \sigma_{\max}(\boldsymbol{A}^{(d)}) L_{g\zeta y}^y\right) \bar{L}_y \Big) \\
&\quad + \left(C_g^{xy} + \sigma_{\max}(\boldsymbol{A}^{(d)}) C_g^{\zeta y}\right) \Bigg(\frac{L_f^y + \sigma_{\max}(\boldsymbol{A}^{(c)}) L_{f\xi}^x}{\gamma_y} \\
&\quad + \frac{\left(\bar{L}_f^y + L_{fy}^\xi \sigma_{\max}(\boldsymbol{A}^{(c)}) + \sigma_{\max}(\boldsymbol{A}^{(c)}) L_{f\xi}^y + \sigma_{\max}^2(\boldsymbol{A}^{(c)}) L_f^\xi\right) \bar{L}_y}{\gamma_y} \\
&\quad + \frac{(C_f^y + \sigma_{\max}(\boldsymbol{A}^{(c)}) C_f^\xi) \left(L_{gyy}^x + L_{gyy}^\zeta \sigma_{\max}(\boldsymbol{A}^{(d)}) + L_{gyy}^y \bar{L}_y\right)}{\gamma_y^2} \Bigg),
\end{aligned} \tag{134}
$$

and

$$
\bar{L}_y := \frac{C_g^{xy} + \sigma_{\max}(\boldsymbol{A}^{(d)}) C_g^{\zeta y}}{\gamma_y}. \tag{135}
$$

Proof. For any $\boldsymbol{x}, \boldsymbol{x}'$, by the optimality conditions of $\boldsymbol{y}^*(\boldsymbol{x})$ and $\boldsymbol{y}^*(\boldsymbol{x}')$, we have

$$
\mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x})} \nabla_y g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\zeta}) = 0, \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x}')} \nabla_y g(\boldsymbol{x}', \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\zeta}) = 0. \tag{136}
$$

From the strong convexity of $g(\cdot)$, we know

$$
\begin{aligned}
\gamma_y \left\|\boldsymbol{y}^*(\boldsymbol{x}) - \boldsymbol{y}^*(\boldsymbol{x}')\right\| &\le \left\| \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x})} \nabla_y g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\zeta}) - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x})} \nabla_y g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\zeta}) \right\| \\
&= \left\| \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x})} \nabla_y g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\zeta}) \right\| \\
&= \left\| \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x})} \nabla_y g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\zeta}) - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x}')} \nabla_y g(\boldsymbol{x}', \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\zeta}) \right\| \\
&\le \left(C_g^{xy} + \sigma_{\max}(\boldsymbol{A}^{(d)}) C_g^{\zeta y}\right) \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|.
\end{aligned} \tag{137}
$$

Therefore, it's easy to know

$$
\left\|\boldsymbol{y}^*(\boldsymbol{x}) - \boldsymbol{y}^*(\boldsymbol{x}')\right\| \le \frac{C_g^{xy} + \sigma_{\max}(\boldsymbol{A}^{(d)}) C_g^{\zeta y}}{\gamma_y} \left\|\boldsymbol{x} - \boldsymbol{x}'\right\| = \bar{L}_y \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|. \tag{138}
$$

By implicit differentiation, the true hypergradient can be written as

$$
\begin{aligned}
&\nabla\Phi(\boldsymbol{x}) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\xi}\sim\mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}))} \nabla_x f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\xi}) - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x})} \left[\nabla_{xy}^2 g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\zeta}) + (\boldsymbol{A}^{(d)})^\intercal \nabla_{\zeta y}^2 g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\zeta})\right] \boldsymbol{v}^*(\boldsymbol{x}),
\end{aligned} \tag{139}
$$

where

$$
\boldsymbol{v}^*(\boldsymbol{x}) = \left[\mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x})} \nabla_{yy}^2 g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\zeta})\right]^{-1} \mathop{\mathbb{E}}_{\boldsymbol{\xi}\sim\mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}))} \left[\nabla_y f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\xi}) + (\boldsymbol{A}^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\xi})\right]. \tag{140}
$$

For any $\boldsymbol{x}, \boldsymbol{x}'$, by the triangle inequality,

$$\begin{aligned}
&\|\nabla\Phi(\boldsymbol{x}) - \nabla\Phi(\boldsymbol{x}')\| \\
&\leq \left\| \mathop{\mathbb{E}}_{\boldsymbol{\xi}\sim\mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}))} \nabla_x f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\xi}) - \mathop{\mathbb{E}}_{\boldsymbol{\xi}\sim\mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}'))} \nabla_x f(\boldsymbol{x}', \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\xi}) \right\| \\
&\quad + \Bigg\| \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x})} \left[\nabla^2_{xy} g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\zeta}) + (\boldsymbol{A}^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\zeta}) \right] \\
&\quad - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x}')} \left[\nabla^2_{xy} g(\boldsymbol{x}', \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\zeta}) + (\boldsymbol{A}^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}', \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\zeta}) \right] \Bigg\| \|\boldsymbol{v}^*(\boldsymbol{x})\| \\
&\quad + \left(C_g^{xy} + \sigma_{\max}(\boldsymbol{A}^{(d)}) C_g^{\zeta y} \right) \|\boldsymbol{v}^*(\boldsymbol{x}) - \boldsymbol{v}^*(\boldsymbol{x}')\| .
\end{aligned} \tag{141}$$

The first term in Eq. (141) satisfies

$$\begin{aligned}
&\left\| \mathop{\mathbb{E}}_{\boldsymbol{\xi}\sim\mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}))} \nabla_x f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\xi}) - \mathop{\mathbb{E}}_{\boldsymbol{\xi}\sim\mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}'))} \nabla_x f(\boldsymbol{x}', \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\xi}) \right\| \\
&\leq L_f^x \|\boldsymbol{x} - \boldsymbol{x}'\| + \bar{L}_f^x \|\boldsymbol{y}^*(\boldsymbol{x}) - \boldsymbol{y}^*(\boldsymbol{x}')\| + L_{fx}^{\xi} W_1 \left(\mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x})), \mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}')) \right) \\
&\leq \left[L_f^x + \left(\bar{L}_f^x + L_{fx}^{\xi} \sigma_{\max}(\boldsymbol{A}^{(c)}) \right) \bar{L}_y \right] \|\boldsymbol{x} - \boldsymbol{x}'\|.
\end{aligned} \tag{142}$$

Similarly, the second term satisfies

$$\begin{aligned}
&\Bigg\| \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x})} \left[\nabla^2_{xy} g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\zeta}) + (\boldsymbol{A}^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\zeta}) \right] \\
&\quad - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x}')} \left[\nabla^2_{xy} g(\boldsymbol{x}', \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\zeta}) + (\boldsymbol{A}^{(d)})^\intercal \nabla^2_{\zeta y} g(\boldsymbol{x}', \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\zeta}) \right] \Bigg\| \\
&\leq L_{gxy}^x \|\boldsymbol{x} - \boldsymbol{x}'\| + L_{gxy}^y \|\boldsymbol{y}^*(\boldsymbol{x}) - \boldsymbol{y}^*(\boldsymbol{x}')\| + L_{gxy}^{\zeta} W_1 \left(\mathcal{D}(\boldsymbol{x}), \mathcal{D}(\boldsymbol{x}') \right) \\
&\leq \Big(L_{gxy}^x + \sigma_{\max}(\boldsymbol{A}^{(d)}) L_{g\zeta y}^x + \left(L_{gxy}^{\zeta} + \sigma_{\max}(\boldsymbol{A}^{(d)}) L_{g\zeta y}^{\zeta} \right) \sigma_{\max}(\boldsymbol{A}^{(d)}) \\
&\quad + \left(L_{gxy}^y + \sigma_{\max}(\boldsymbol{A}^{(d)}) L_{g\zeta y}^y \right) \bar{L}_y \Big) \|\boldsymbol{x} - \boldsymbol{x}'\|.
\end{aligned} \tag{143}$$

Moreover,

$$\|\boldsymbol{v}^*(\boldsymbol{x})\| \leq \frac{C_f^y + \sigma_{\max}(\boldsymbol{A}^{(c)}) C_f^{\xi}}{\gamma_y}. \tag{144}$$

For the vector term in $\boldsymbol{v}^*(\boldsymbol{x})$, we have

$$\begin{aligned}
&\Bigg\| \mathop{\mathbb{E}}_{\boldsymbol{\xi}\sim\mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}))} \left[\nabla_y f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\xi}) + (\boldsymbol{A}^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\xi}) \right] \\
&\quad - \mathop{\mathbb{E}}_{\boldsymbol{\xi}\sim\mathcal{C}(\boldsymbol{y}^*(\boldsymbol{x}'))} \left[\nabla_y f(\boldsymbol{x}', \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\xi}) + (\boldsymbol{A}^{(c)})^\intercal \nabla_\xi f(\boldsymbol{x}', \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\xi}) \right] \Bigg\| \\
&\leq \left[L_f^y + \sigma_{\max}(\boldsymbol{A}^{(c)}) L_{f\xi}^x + \left(\bar{L}_f^y + L_{fy}^{\xi} \sigma_{\max}(\boldsymbol{A}^{(c)}) + \sigma_{\max}(\boldsymbol{A}^{(c)}) L_{f\xi}^y + \sigma^2_{\max}(\boldsymbol{A}^{(c)}) L_f^{\xi} \right) \bar{L}_y \right] \\
&\quad \cdot \|\boldsymbol{x} - \boldsymbol{x}'\|.
\end{aligned} \tag{145}$$

For the Hessian term, we have

$$\begin{aligned}
&\left\| \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x})} \nabla^2_{yy} g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \boldsymbol{\zeta}) - \mathop{\mathbb{E}}_{\boldsymbol{\zeta}\sim\mathcal{D}(\boldsymbol{x}')} \nabla^2_{yy} g(\boldsymbol{x}', \boldsymbol{y}^*(\boldsymbol{x}'); \boldsymbol{\zeta}) \right\| \\
&\leq \left(L_{gyy}^x + L_{gyy}^{\zeta} \sigma_{\max}(\boldsymbol{A}^{(d)}) + L_{gyy}^y \bar{L}_y \right) \|\boldsymbol{x} - \boldsymbol{x}'\|.
\end{aligned} \tag{146}$$

Combining Eqs. (145) and (146), and using $\|H_1^{-1} - H_2^{-1}\| \le \|H_1^{-1}\|\|H_1 - H_2\|\|H_2^{-1}\|$, we obtain

$$\begin{aligned}
&\|\boldsymbol{v}^*(\boldsymbol{x}) - \boldsymbol{v}^*(\boldsymbol{x}')\| \\
&\le \left[\frac{L_f^y + \sigma_{\max}(\boldsymbol{A}^{(c)}) L_{f\xi}^x + \left(\bar{L}_f^y + L_{fy}^{\xi} \sigma_{\max}(\boldsymbol{A}^{(c)}) + \sigma_{\max}(\boldsymbol{A}^{(c)}) L_{f\xi}^y + \sigma_{\max}^2(\boldsymbol{A}^{(c)}) L_f^{\xi} \right) \bar{L}_y}{\gamma_y} \right. \\
&\quad \left. + \frac{\left(C_f^y + \sigma_{\max}(\boldsymbol{A}^{(c)}) C_f^{\xi}\right) \left(L_{gyy}^x + L_{gyy}^{\zeta} \sigma_{\max}(\boldsymbol{A}^{(d)}) + L_{gyy}^y \bar{L}_y \right)}{\gamma_y^2} \right] \|\boldsymbol{x} - \boldsymbol{x}'\|. \qquad (147)
\end{aligned}$$

Combining Eqs. (141), (142), (143), (144), and (147), we obtain

$$\|\nabla \Phi(\boldsymbol{x}) - \nabla \Phi(\boldsymbol{x}')\| \le L_{\Phi} \|\boldsymbol{x} - \boldsymbol{x}'\|. \qquad (148)$$

□

G Proof of Theorem 1

For the descent theorem, we assume that the performative objective is lower bounded, i.e., $\Phi_{\inf} := \inf_{\boldsymbol{x} \in \mathbb{R}^{d_x}} \Phi(\boldsymbol{x}) > -\infty$. This is a standard requirement for converting the one-step descent inequality into an averaged stationarity guarantee.

The recursion in Eq. (132) is valid for $r \ge 1$, with the initial tracking term given by $\mathbb{E}\|\boldsymbol{y}_{0,K} - \boldsymbol{y}^*(\boldsymbol{x}_0)\|^2$, i.e., the error after the first inner loop rather than the input point $\boldsymbol{y}_{0,0}$. Summing Eq. (132) over $r = 1, \ldots, R-1$, adding the term at $r = 0$, and using $\lambda_y < 1$, we obtain

$$\begin{aligned}
&\sum_{r=0}^{R-1} \mathbb{E} \left\| \boldsymbol{y}_{r,K} - \boldsymbol{y}^*(\boldsymbol{x}_r) \right\|^2 \\
&\le \frac{1}{1-\lambda_y} \mathbb{E} \left\| \boldsymbol{y}_{0,K} - \boldsymbol{y}^*(\boldsymbol{x}_0) \right\|^2 + \frac{12\alpha^2 (1-\beta\gamma_y)^K \bar{L}_y^2}{1-\lambda_y} \sum_{r=0}^{R-1} \mathbb{E} \left\| \nabla \Phi(\boldsymbol{x}_r) \right\|^2 \\
&\quad + \frac{12\alpha^2 (1-\beta\gamma_y)^K \bar{L}_y^2}{1-\lambda_y} \sum_{r=0}^{R-1} \mathbb{E} \left[\mathcal{V}_r^{\Phi} \right] + \frac{1}{1-\lambda_y} \sum_{r=0}^{R-1} \frac{8\beta (C_g^y)^2}{\gamma_y |\mathcal{B}_r^{(gy)}|} \\
&\quad + \frac{6(1-\beta\gamma_y)^K + 2}{1-\lambda_y} \frac{(L_{gy}^{\zeta})^2}{\gamma_y^2} \sum_{r=0}^{R-1} \mathbb{E} \left[\|\boldsymbol{x}_r\|^2 \left\| \widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)} \right\|^2 \right]. \qquad (149)
\end{aligned}$$

By Lemma 6, Φ is L_{Φ}-smooth. For the outer update, $\boldsymbol{x}_{r+1} - \boldsymbol{x}_r = -\alpha \widehat{\nabla} \Phi(\boldsymbol{x}_r)$. If $0 < \alpha \le 1/(2L_{\Phi})$, then for every $r = 0, \ldots, R-1$,

$$\mathbb{E} \left[\Phi(\boldsymbol{x}_{r+1}) \right] \le \mathbb{E} \left[\Phi(\boldsymbol{x}_r) \right] - \frac{\alpha}{4} \mathbb{E} \left\| \nabla \Phi(\boldsymbol{x}_r) \right\|^2 + \alpha \mathbb{E} \left\| \widehat{\nabla} \Phi(\boldsymbol{x}_r) - \nabla \Phi(\boldsymbol{x}_r) \right\|^2. \qquad (150)$$

With Eq. (128), this gives

$$\mathbb{E} \left[\Phi(\boldsymbol{x}_{r+1}) \right] \le \mathbb{E} \left[\Phi(\boldsymbol{x}_r) \right] - \frac{\alpha}{4} \mathbb{E} \left\| \nabla \Phi(\boldsymbol{x}_r) \right\|^2 + \alpha \mathbb{E} \left[\mathcal{V}_r^{\Phi} \right] + \alpha \mathcal{H}_y \mathbb{E} \left\| \boldsymbol{y}_{r,K} - \boldsymbol{y}^*(\boldsymbol{x}_r) \right\|^2. \qquad (151)$$

Telescoping Eq. (151) over $r = 0, \ldots, R-1$, using the lower boundedness of $\Phi(\cdot)$, and combining it with Eq. (149), we obtain

$$\begin{aligned}
&\left[\frac{\alpha}{4} - \alpha\mathcal{H}_y \frac{12\alpha^2(1-\beta\gamma_y)^K \bar{L}_y^2}{1-\lambda_y}\right] \sum_{r=0}^{R-1} \mathbb{E}\left\|\nabla\Phi(\boldsymbol{x}_r)\right\|^2 \\
&\leq \Phi(\boldsymbol{x}_0) - \Phi_{\inf} + \alpha\left[1 + \mathcal{H}_y \frac{12\alpha^2(1-\beta\gamma_y)^K \bar{L}_y^2}{1-\lambda_y}\right] \sum_{r=0}^{R-1} \mathbb{E}\left[\mathcal{V}_r^{\Phi}\right] \\
&\quad + \alpha\mathcal{H}_y \frac{1}{1-\lambda_y} \mathbb{E}\left\|\boldsymbol{y}_{0,K} - \boldsymbol{y}^*(\boldsymbol{x}_0)\right\|^2 + \alpha\mathcal{H}_y \frac{1}{1-\lambda_y} \sum_{r=0}^{R-1} \frac{8\beta(C_g^y)^2}{\gamma_y |\mathcal{B}_r^{(gy)}|} \\
&\quad + \alpha\mathcal{H}_y \frac{6(1-\beta\gamma_y)^K + 2}{1-\lambda_y} \frac{(L_{gy}^{\zeta})^2}{\gamma_y^2} \sum_{r=0}^{R-1} \mathbb{E}\left[\|\boldsymbol{x}_r\|^2 \left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^2\right].
\end{aligned} \tag{152}$$

Choose K sufficiently large such that

$$\mathcal{H}_y \frac{12\alpha^2(1-\beta\gamma_y)^K \bar{L}_y^2}{1-\lambda_y} \leq \frac{1}{16}. \tag{153}$$

Then the coefficient on the left-hand side of Eq. (152) is lower bounded by $3\alpha/16$. Let $\bar{\boldsymbol{x}}$ be sampled uniformly from $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{R-1}\}$. Then

$$\begin{aligned}
\mathbb{E}\left\|\nabla\Phi(\bar{\boldsymbol{x}})\right\|^2 \leq \mathcal{O}\Bigg(&\frac{1}{\alpha R} + \frac{1}{R}\mathbb{E}\left\|\boldsymbol{y}_{0,K} - \boldsymbol{y}^*(\boldsymbol{x}_0)\right\|^2 + \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\mathcal{V}_r^{\Phi}\right] + \frac{1}{R}\sum_{r=0}^{R-1}\frac{1}{|\mathcal{B}_r^{(gy)}|} \\
&+ \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\|\boldsymbol{x}_r\|^2 \left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^2\right]\Bigg).
\end{aligned} \tag{154}$$

It remains to control the distribution map estimation terms inside $\mathcal{V}_r^{\Phi}$. By Lemma 1, Eq. (154) can be written as

$$\begin{aligned}
\mathbb{E}\left\|\nabla\Phi(\bar{\boldsymbol{x}})\right\|^2 = \mathcal{O}\Bigg(&\frac{1}{\alpha R} + \frac{1}{R}\mathbb{E}\left\|\boldsymbol{y}_{0,K} - \boldsymbol{y}^*(\boldsymbol{x}_0)\right\|^2 + \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\mathcal{V}_r^{\Phi,\mathrm{part}}\right] + \frac{1}{R}\sum_{r=0}^{R-1}\frac{1}{|\mathcal{B}_r^{(gy)}|} \\
&+ \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[(1+\|\boldsymbol{x}_r\|^2)\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^2\right] \\
&+ \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[(1+\|\boldsymbol{y}_{r,K}\|^2)\left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\|^2\right] \\
&+ \frac{1}{R}\sum_{r=0}^{R-1}\sqrt{\mathbb{E}\left[(1+\|\boldsymbol{x}_r\|^2)^2\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^4\right]}\sqrt{\mathbb{E}\left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\|^4}\Bigg),
\end{aligned} \tag{155}$$

where $\mathcal{V}_r^{\Phi,\mathrm{part}}$ contains only the mini-batch variance and Neumann truncation terms. For any target accuracy $0 < \epsilon \leq 1$, we next specify the parameters so that every term in Eq. (155) is bounded by a constant multiple of ϵ. Choose $\alpha = \Theta\left(\frac{1}{L_\Phi}\right), 0 < \alpha \leq \frac{1}{2L_\Phi}$, $\beta = \Theta\left(\frac{1}{C_g^{yy}}\right), 0 < \beta \leq \frac{c_\beta}{C_g^{yy}}$, and $0 < \eta \leq \frac{c_\eta}{C_g^{yy}}$, where $0 < c_\beta, c_\eta < 1$. Choose

$$R = \left\lceil \frac{c_R}{\alpha\epsilon} \right\rceil, \tag{156}$$

for a sufficiently large constant $c_R > 0$. Recall that $\lambda_y = (1-\beta\gamma_y)^K\left(6 + 12\alpha^2\bar{L}_y^2\mathcal{H}_y\right)$. A sufficient explicit choice of K that guarantees both $\lambda_y \leq 1/2$ and Eq. (153) is

$$K \geq \left\lceil \frac{\log\left(\max\left\{2\left(6 + 12\alpha^2\bar{L}_y^2\mathcal{H}_y\right),\, 384\alpha^2\bar{L}_y^2\mathcal{H}_y,\, 1\right\}\right)}{-\log(1-\beta\gamma_y)} \right\rceil. \tag{157}$$

The first term inside the maximum guarantees $\lambda_y \le 1/2$, while the second guarantees the absorption condition because $(1-\beta\gamma_y)^K \le \frac{1}{384\alpha^2\bar{L}_y^2\mathcal{H}_y}$ whenever this term is active. Consequently,

$$K = \Theta\left(\frac{1}{\beta\gamma_y}\log\left(\max\left\{6 + 12\alpha^2\bar{L}_y^2\mathcal{H}_y,\, 384\alpha^2\bar{L}_y^2\mathcal{H}_y,\, 1\right\}\right)\right). \tag{158}$$

If $\beta = \Theta(1/C_g^{yy})$, then

$$K = \Theta\left(\frac{C_g^{yy}}{\gamma_y}\log\left(\max\left\{6 + 12\alpha^2\bar{L}_y^2\mathcal{H}_y,\, 384\alpha^2\bar{L}_y^2\mathcal{H}_y,\, 1\right\}\right)\right). \tag{159}$$

For the Neumann-series estimator, choose

$$Q = \left\lceil \frac{c_Q}{\eta\gamma_y}\log\frac{1}{\epsilon}\right\rceil, \tag{160}$$

for a sufficiently large constant $c_Q > 0$. If $\eta = \Theta(1/C_g^{yy})$, then

$$Q = \Theta\left(\frac{C_g^{yy}}{\gamma_y}\log\frac{1}{\epsilon}\right). \tag{161}$$

Choose the mini-batch sizes as

$$|\mathcal{B}_r^{(f)}| = \left\lceil c_f\epsilon^{-1}\right\rceil, |\mathcal{B}_r^{(gxy)}| = \left\lceil c_{gxy}\epsilon^{-1}\right\rceil, B = \left\lceil c_v\epsilon^{-1}\right\rceil, |\mathcal{B}_r^{(gy)}| = \left\lceil c_{gy}\epsilon^{-1}\right\rceil, \tag{162}$$

where the constants are chosen sufficiently large. We now choose the cumulative numbers of distribution map updates. Let $S_{d,-1} = S_{c,-1} = 0$. At outer iteration r, choose

$$S_{d,r} := \min\left\{s \in \mathbb{N} : s \ge S_{d,r-1} + 1,\, \frac{s + 8d_x}{\left[1 + \log(s + 8d_x)\right]^2} \ge \frac{c_d\left(1 + \|\boldsymbol{x}_r\|^2\right)}{\epsilon}\right\}, \tag{163}$$

where $c_d > 0$ is a sufficiently large constant. After $\boldsymbol{y}_{r,K}$ is obtained, choose

$$S_{c,r} := \min\left\{s \in \mathbb{N} : s \ge S_{c,r-1} + 1,\, \frac{s + 8d_y}{\left[1 + \log(s + 8d_y)\right]^2} \ge \frac{c_c\left(1 + \|\boldsymbol{y}_{r,K}\|^2\right)}{\epsilon}\right\}, \tag{164}$$

where $c_c > 0$ is sufficiently large. Therefore,

$$1 + \|\boldsymbol{x}_r\|^2 \le \frac{\epsilon}{c_d}\frac{S_{d,r} + 8d_x}{\left[1 + \log(S_{d,r} + 8d_x)\right]^2}, \tag{165}$$

$$1 + \|\boldsymbol{y}_{r,K}\|^2 \le \frac{\epsilon}{c_c}\frac{S_{c,r} + 8d_y}{\left[1 + \log(S_{c,r} + 8d_y)\right]^2}. \tag{166}$$

Since $\|\boldsymbol{M}\| \le \|\boldsymbol{M}\|_F$ for every matrix $\boldsymbol{M}$, Lemma 3 and Eq. (166) imply

$$\mathbb{E}\left[\left(1 + \|\boldsymbol{x}_r\|^2\right)\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^2\right] \le \frac{\epsilon}{c_d}\mathbb{E}\left[\frac{S_{d,r} + 8d_x}{\left[1 + \log(S_{d,r} + 8d_x)\right]^2}\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|_F^2\right] \le \frac{C_\zeta}{c_d}\epsilon, \tag{167}$$

and similarly,

$$\mathbb{E}\left[\left(1 + \|\boldsymbol{y}_{r,K}\|^2\right)\left\|\widehat{\boldsymbol{A}}_r^{(c)} - \boldsymbol{A}^{(c)}\right\|^2\right] \le \frac{C_\xi}{c_c}\epsilon. \tag{168}$$

Moreover, using the fourth-moment bounds in Lemma 3, we have

$$\mathbb{E}\left[\left(1 + \|\boldsymbol{x}_r\|^2\right)^2\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|^4\right] \le \frac{\epsilon^2}{c_d^2}\mathbb{E}\left[\frac{(S_{d,r} + 8d_x)^2}{\left[1 + \log(S_{d,r} + 8d_x)\right]^4}\left\|\widehat{\boldsymbol{A}}_r^{(d)} - \boldsymbol{A}^{(d)}\right\|_F^4\right] \le \frac{C_\zeta}{c_d^2}\epsilon^2, \tag{169}$$

and

$$\mathbb{E}\left[(1+\|\boldsymbol{y}_{r,K}\|^2)^2\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|^4\right]\le\frac{C_\xi}{c_c^2}\epsilon^2. \tag{170}$$

Since $1+\|\boldsymbol{y}_{r,K}\|^2\ge 1$, Eq. (170) also gives

$$\mathbb{E}\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|^4\le\frac{C_\xi}{c_c^2}\epsilon^2. \tag{171}$$

Consequently,

$$\sqrt{\mathbb{E}\left[(1+\|\boldsymbol{x}_r\|^2)^2\left\|\widehat{\boldsymbol{A}}_r^{(d)}-\boldsymbol{A}^{(d)}\right\|^4\right]}\sqrt{\mathbb{E}\left\|\widehat{\boldsymbol{A}}_r^{(c)}-\boldsymbol{A}^{(c)}\right\|^4}\le\frac{\sqrt{C_\zeta C_\xi}}{c_d c_c}\epsilon^2\le\frac{\sqrt{C_\zeta C_\xi}}{c_d c_c}\epsilon, \tag{172}$$

where the last inequality follows from $0<\epsilon\le 1$. The remaining part of the hypergradient error satisfies

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\mathcal{V}_r^{\Phi,\mathrm{part}}\right]=\mathcal{O}\left((1-\eta\gamma_y)^{2Q+2}+\frac{1}{B}+\frac{1}{|\mathcal{B}_r^{(f)}|}+\frac{1}{|\mathcal{B}_r^{(gxy)}|}\right). \tag{173}$$

Under Eqs. (160) and (162), the right-hand side of Eq. (173) is $\mathcal{O}(\epsilon)$. Moreover, $\frac{1}{R}\sum_{r=0}^{R-1}\frac{1}{|\mathcal{B}_r^{(gy)}|}=\mathcal{O}(\epsilon)$, and for a fixed initial tracking error, $\frac{1}{R}\mathbb{E}\left\|\boldsymbol{y}_{0,K}-\boldsymbol{y}^*(\boldsymbol{x}_0)\right\|^2=\mathcal{O}(\epsilon)$. Finally, Eq. (156) gives $\frac{1}{\alpha R}=\mathcal{O}(\epsilon)$.

Combining these relations with Eqs. (167)–(172) in Eq. (155) yields

$$\mathbb{E}\left\|\nabla\Phi(\bar{\boldsymbol{x}})\right\|^2\le\mathcal{O}(\epsilon). \tag{174}$$

Hence, the uniformly sampled output $\bar{\boldsymbol{x}}$ is an ϵ-accurate stationary point.

We next characterize the computation complexity. In each outer iteration, there are K lower-level stochastic gradient steps, each using $|\mathcal{B}_r^{(gy)}|$ samples, one mini-batch evaluation of the f-gradients using $|\mathcal{B}_r^{(f)}|$ samples, one Jacobian-vector product batch $\left(\nabla_{xy}^2 g(\cdot)+(\widehat{\boldsymbol{A}}_r^{(d)})^\intercal\nabla_{\zeta y}^2 g(\cdot)\right)\boldsymbol{v}_Q$ using $|\mathcal{B}_r^{(gxy)}|$ samples, and Q Hessian-vector product batches for the Neumann-series estimator. For the Hessian-vector products, under the integer exponentially decaying batch choice

$$|\mathcal{B}_{r,Q+1-j}^{(v)}|=\max\left\{1,\left\lceil BQ(1-\eta\gamma_y)^{j-1}\right\rceil\right\}, \tag{175}$$

the total number of Hessian-vector products per outer iteration satisfies

$$\sum_{j=1}^{Q}|\mathcal{B}_{r,j}^{(v)}|\le\mathcal{O}\left(\frac{BQ}{\eta\gamma_y}\right). \tag{176}$$

Therefore, the gradient, Jacobian-vector product, and Hessian-vector product complexities are

$$\mathrm{Gc}(f,\epsilon)=\mathcal{O}\left(R|\mathcal{B}_r^{(f)}|\right)=\mathcal{O}\left(L_\Phi\epsilon^{-2}\right), \tag{177}$$

$$\mathrm{Gc}(g,\epsilon)=\mathcal{O}\left(RK|\mathcal{B}_r^{(gy)}|\right)=\mathcal{O}\left(L_\Phi K\epsilon^{-2}\right), \tag{178}$$

$$\mathrm{JV}(g,\epsilon)=\mathcal{O}\left(R|\mathcal{B}_r^{(gxy)}|\right)=\mathcal{O}\left(L_\Phi\epsilon^{-2}\right), \tag{179}$$

$$\mathrm{HV}(g,\epsilon)=\mathcal{O}\left(R\frac{BQ}{\eta\gamma_y}\right)=\mathcal{O}\left(\frac{L_\Phi}{\eta\gamma_y}\epsilon^{-2}Q\right). \tag{180}$$

With

$$Q=\mathcal{O}\left((\eta\gamma_y)^{-1}\log\frac{1}{\epsilon}\right), \tag{181}$$

we have

$$\mathrm{HV}(g,\epsilon) = \widetilde{\mathcal{O}}\left(\frac{L_\Phi}{(\eta\gamma_y)^2}\epsilon^{-2}\right). \tag{182}$$

For comparison with Ji et al. (2021), we adopt the same condition number definition. Specifically, all smoothness constants, bounded derivative constants, and fixed constants induced by the distribution maps are absorbed into a single constant L, and $\kappa := L/\gamma_y$. In this case, $\bar{L}_y = \mathcal{O}(\kappa), L_\Phi = \mathcal{O}(\kappa^3), \mathcal{H}_y = \mathcal{O}(\kappa^4), \alpha = \mathcal{O}(\kappa^{-3}), R = \mathcal{O}\left(\kappa^3\epsilon^{-1}\right), K = \mathcal{O}(\kappa)$, and $Q = \mathcal{O}\left(\kappa \log \frac{1}{\epsilon}\right)$. The mini-batch sizes scale as $|\mathcal{B}_r^{(f)}| = \mathcal{O}\left(\kappa^2\epsilon^{-1}\right), |\mathcal{B}_r^{(gxy)}| = \mathcal{O}\left(\kappa^2\epsilon^{-1}\right), B = \mathcal{O}\left(\kappa^2\epsilon^{-1}\right), |\mathcal{B}_r^{(gy)}| = \mathcal{O}\left(\kappa^5\epsilon^{-1}\right)$. Therefore, we have

$$\mathrm{Gc}(f,\epsilon) = \mathcal{O}\left(\kappa^5\epsilon^{-2}\right), \tag{183}$$

$$\mathrm{Gc}(g,\epsilon) = \mathcal{O}\left(\kappa^9\epsilon^{-2}\right), \tag{184}$$

$$\mathrm{JV}(g,\epsilon) = \mathcal{O}\left(\kappa^5\epsilon^{-2}\right), \tag{185}$$

$$\mathrm{HV}(g,\epsilon) = \widetilde{\mathcal{O}}\left(\kappa^7\epsilon^{-2}\right). \tag{186}$$

Finally, we characterize the additional sample complexity required for estimating the two unknown distribution maps. Each stochastic least-squares update requires two deployment observations. Therefore,

$$\mathrm{SC}_d(\epsilon) = 2S_{d,R-1}, \tag{187}$$

$$\mathrm{SC}_c(\epsilon) = 2S_{c,R-1}. \tag{188}$$

From Eqs. (163) and (164), the cumulative update indices increase by at least one at each outer iteration, while additional updates are performed only when the current accuracy requirement exceeds the accuracy provided by the previous cumulative updates. Consequently,

$$S_{d,R-1} \le R + \max_{0\le r<R} \min\left\{ s\in\mathbb{N} : \frac{s+8d_x}{[1+\log(s+8d_x)]^2} \ge \frac{c_d(1+\|\boldsymbol{x}_r\|^2)}{\epsilon} \right\}, \tag{189}$$

$$S_{c,R-1} \le R + \max_{0\le r<R} \min\left\{ s\in\mathbb{N} : \frac{s+8d_y}{[1+\log(s+8d_y)]^2} \ge \frac{c_c(1+\|\boldsymbol{y}_{r,K}\|^2)}{\epsilon} \right\}. \tag{190}$$

Since

$$\frac{s+8d}{[1+\log(s+8d)]^2} = \widetilde{\Theta}(s),$$

we obtain

$$\mathrm{SC}_d(\epsilon) = \widetilde{\mathcal{O}}\left(R + \frac{1+\max_{0\le r<R}\|\boldsymbol{x}_r\|^2}{\epsilon}\right), \tag{191}$$

$$\mathrm{SC}_c(\epsilon) = \widetilde{\mathcal{O}}\left(R + \frac{1+\max_{0\le r<R}\|\boldsymbol{y}_{r,K}\|^2}{\epsilon}\right). \tag{192}$$

Since $R = \mathcal{O}\left(L_\Phi\epsilon^{-1}\right)$, the exploratory sample complexities have $\widetilde{\mathcal{O}}(\epsilon^{-1})$ dependence on the target accuracy, up to the trajectory-dependent maximum norms. In particular, when these trajectory-dependent factors are treated as constants, we know

$$\mathrm{SC}_d(\epsilon) = \widetilde{\mathcal{O}}\left(L_\Phi\epsilon^{-1}\right), \tag{193}$$

$$\mathrm{SC}_c(\epsilon) = \widetilde{\mathcal{O}}\left(L_\Phi\epsilon^{-1}\right). \tag{194}$$

With the defined condition number κ, we have

$$\mathrm{SC}_d(\epsilon) = \widetilde{\mathcal{O}}\left(\kappa^3\epsilon^{-1}\right), \tag{195}$$

$$\mathrm{SC}_c(\epsilon) = \widetilde{\mathcal{O}}\left(\kappa^3\epsilon^{-1}\right). \tag{196}$$