Reducing Contextual Stochastic Bilevel Optimization via Structured Function Approximation

Maxime Bouscary Jiawei Zhang Saurabh Amin
Massachusetts Institute of Technology, Cambridge, MA, USA
{mbscry,jwzhang,amins}@mit.edu

Abstract

Contextual Stochastic Bilevel Optimization (CSBO) extends standard stochastic bilevel optimization (SBO) by incorporating context-dependent lower-level problems. CSBO problems are generally intractable since existing methods require solving a distinct lower-level problem for each sampled context, resulting in prohibitive sample and computational complexity, in addition to relying on impractical conditional sampling oracles. We propose a reduction framework that approximates the lower-level solutions using expressive basis functions, thereby decoupling the lower-level dependence on context and transforming CSBO into a standard SBO problem solvable using only joint samples from the context and noise distribution. First, we show that this reduction preserves hypergradient accuracy and yields an ϵ-stationary solution to CSBO. Then, we relate the sample complexity of the reduced problem to simple metrics of the basis. This establishes sufficient criteria for a basis to yield ϵ-stationary solutions with a near-optimal complexity of $\widetilde{\mathcal{O}}(\epsilon^{-3})$, matching the best-known rate for standard SBO up to logarithmic factors. Moreover, we show that Chebyshev polynomials provide a concrete and efficient choice of basis that satisfies these criteria for a broad class of problems. Empirical results on inverse and hyperparameter optimization demonstrate that our approach outperforms CSBO baselines in convergence, sample efficiency, and memory usage.

1 Introduction

Many real-world optimization tasks involve solving a bilevel problem where the lower-level (LL) solution depends on the upper-level (UL) uncertainty (or “context"). We formalize this structure as the following Contextual Stochastic Bilevel Optimization problem (CSBO):

$$\begin{aligned} \min_{x\in\mathbb{R}^{d_x}} \quad & F(x) \triangleq \mathbb{E}_{(\xi,\eta)\sim\mathbb{P}_{(\xi,\eta)}}\left[f(x, y^\star(x,\xi), \xi, \eta)\right] && \text{(CSBO)} \\ \text{s.t.} \quad & y^\star(x,\xi) = \operatorname*{arg\,min}_{y\in\mathbb{R}^{d_y}} \mathbb{E}_{\eta\sim\mathbb{P}_{\eta|\xi}}\left[g(x,y,\xi,\eta)\right], \quad \forall x\in\mathbb{R}^{d_x}, \xi\in\Xi \end{aligned}$$

where ξ is the context, η the LL uncertainty, and g is strongly convex w.r.t. y for any given x, ξ to ensure that $y^\star(x,\xi)$ is uniquely defined. Compared to standard Stochastic Bilevel Optimization (SBO), CSBO is significantly more challenging since the LL must be solved for every realization of ξ. SBO methods achieve ϵ-accurate solutions with a sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-4})$ [1, 2] and up to $\mathcal{O}(\epsilon^{-3})$ with variance reduction methods [3, 4, 5, 6, 3]. CSBO can be solved naively by solving the LL problem from scratch at each UL iteration, achieving an ϵ-stationary point with $\mathcal{O}(\epsilon^{-6})$ samples, or by partitioning the context space to approximate the problem with an SBO with multiple subproblems. Existing CSBO methods based on multilevel Monte Carlo reduce the sample complexity to $\mathcal{O}(\epsilon^{-4})$ [7] but typically assume access to a conditional sampling oracle $\mathbb{P}_{\eta|\xi}$, and their

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ML x OR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

practical performance depends on the existence of a reference point $y_0(x)$ that is independent of ξ and close to $y^\star(x,\xi)$.

We propose a reduction that parametrizes the LL solution with $y_\Phi\left(W(x),\xi\right) \triangleq W(x)\Phi(\xi)$ where $\Phi : \Xi \to \mathbb{R}^N$ is a feature map and $W(x) \in \mathbb{R}^{d_y \times N}$ are context-independent coefficients. This decouples the LL from ξ and yields a single LL problem whose uncertainty is the joint (ξ,η) distribution, resulting in the following SBO problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^{d_x}} \quad & F_\Phi(x) \triangleq \mathbb{E}_{(\xi,\eta)\sim\mathbb{P}_{(\xi,\eta)}}\left[f(x, y_\Phi\left(W^\star(x),\xi\right),\xi,\eta)\right] \\ \text{s.t.} \quad & W^\star(x) = \operatorname*{arg\,min}_{W\in\mathbb{R}^{d_y\times N}} \mathbb{E}_{(\xi,\eta)\sim\mathbb{P}_{(\xi,\eta)}}\left[g(x,y_\Phi\left(W,\xi\right),\xi,\eta)\right], \quad \forall x\in\mathbb{R}^{d_x} \end{aligned} \tag{SBO$_\Phi$}$$

As a result, (i) only i.i.d. joint samples are required (no conditional oracle), (ii) with a suitable choice of Φ, the strong convexity of the LL is preserved, and (iii) off-the-shelf SBO solvers can be used to solve the reduced problem. While recent approaches propose to solve a surrogate for the UL or LL objective [8, 9, 10, 11, 12, 13], their complex structure does not enable deriving tight error guarantees.

Our work makes the following contributions:

1. **Valid Reduction**. We show that an ϵ-stationary solution to the original CSBO can be constructed from an $\frac{\epsilon}{\sqrt{2}}$-stationary solution to the reduced SBO.
2. **Basis-dependent complexity**. We express the regularity constants of the reduced problem in terms of two simple basis metrics: the magnitude $M_\Phi(\epsilon)$ and the non-degeneracy $m_\Phi(\epsilon)$. We then show that existing SBO solvers yield a sample complexity $\widetilde{\mathcal{O}}\left(\epsilon^{-3}\text{poly}\left(M_\Phi(\epsilon), 1/m_\Phi(\epsilon)\right)\right)$.
3. **Near-optimality with Chebyshev basis**. Bases satisfying $M_\Phi(\epsilon) = \widetilde{\mathcal{O}}(1)$ and $1/m_\Phi(\epsilon) = \widetilde{\mathcal{O}}(1)$ yield a near-optimal sample complexity $\widetilde{\mathcal{O}}(\epsilon^{-3})$. Under mild assumptions on g and $\mathbb{P}_\xi$, we show that the Chebyshev basis is such a basis - matching the sample complexity lower bound for nonconvex stochastic optimization up to logarithmic factors.

2 Main Results

Throughout the paper, we make the following assumptions about problem (CSBO).

Assumption 2.1.

(i) The functions f, ∇f, ∇g, and $\nabla^2 g$, are $L_{f,0}$, $L_{f,1}$, $L_{g,1}$, and $L_{g,2}$-Lipschitz continuous with respect to (x,y) for any fixed (ξ,η), respectively.
(ii) The function g is μ-strongly convex with respect to y for all x, ξ, and η.
(iii) If $\eta \sim \mathbb{P}_{\eta|\xi}$, then the gradients $\nabla f(x,y,\xi,\eta)$, $\nabla g(x,y,\xi,\eta)$, and $\nabla^2 g(x,y,\xi,\eta)$, are unbiased and have variance bounded by σ_f^2, $\sigma_{g,1}^2$, and $\sigma_{g,2}^2$ uniformly across all x, y, and ξ, respectively.

For a given countable basis Φ, we let $\Phi^\epsilon : \Xi \to \mathbb{R}^N$ denote the smallest mapping whose components are the first N elements of Φ and that satisfies for some $W^\dagger : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y\times N}$:

$$\mathbb{E}_\xi \left\|y_{\Phi^\epsilon}(W^\dagger(x),\xi) - y^\star(x,\xi)\right\|^2 \leq \frac{\epsilon^2}{4K^2}\frac{\mu}{L_{g,1}}, \quad \forall x\in\mathbb{R}^{d_x}, \tag{1}$$

where K is a constant depending only on the regularities of f and g.

Our first theorem states that by working in the Φ^ϵ-parametrized space, any off-the-shelf SBO solver can be used to solve the CSBO instance.

Theorem 2.2. *Suppose that assumption 2.1 holds, and let Φ be an expressive basis. If $(x^\star, W^\star(x^\star))$ is an $\frac{\epsilon}{\sqrt{2}}$-stationary solution to* (SBO$_{\Phi^\epsilon}$)*, then $(x^\star, \xi \mapsto y_{\Phi^\epsilon}(W^\star(x^\star),\xi))$ is an ϵ-stationary solution to* (CSBO).

Since the upper and lower level expectations of (SBO$_{\Phi^\epsilon}$) are both taken over $\mathbb{P}_{(\xi,\eta)}$, one can obtain a solution to CSBO without a conditional sampling oracle. Moreover, if Φ induces a smooth map

$\xi \mapsto y_{\Phi^\epsilon}(x, \xi)$, then a gradient step at a given ξ generalizes to its neighbors, yielding computational efficiency and guarding against overfitting to context-specific noise.

Next, we show that the reformulation (SBO$_{\Phi^\epsilon}$) can be solved efficiently as it satisfies the assumptions of [3] with Lipschitz constants and variance bounds that depend on $M_\Phi(\epsilon)$ and $m_\Phi(\epsilon)$. Substituting these regularity coefficients into Theorem 1 of [3], we obtain the following result.

Theorem 2.3. *Suppose that assumption 2.1 holds. Then, an ϵ-stationary solution to* (SBO$_{\Phi^\epsilon}$) *can be achieved with a sample complexity in* $\widetilde{\mathcal{O}}\left(\frac{1}{\epsilon^3} \cdot poly\left(M_\Phi(\epsilon), \frac{1}{m_\Phi(\epsilon)}\right)\right)$.

Fundamentally, this Theorem shows that the sample complexity of (SBO$_{\Phi^\epsilon}$) scales polynomially in M_Φ and $1/m_\Phi$. Specifically, the complexity remains well-controlled when:

1. the magnitude of Φ^ϵ grows slowly, ensuring that the Lipschitz and smoothness constants of the reformulation remain comparable to those of the original problem, and avoiding the need for significantly smaller step-sizes.
2. the conditioning of Φ^ϵ decreases slowly, guaranteeing that the reformulation's lower level retains a pronounced strong convexity.

Under these two favorable regimes, the number of iterations and samples required to obtain an ϵ-accurate solution to (SBO$_{\Phi^\epsilon}$) is tightly controlled. Theorem 2.3 also reveals that a basis satisfying $M_\Phi(\epsilon) = \widetilde{\mathcal{O}}(1)$ and $1/m_\Phi(\epsilon) = \widetilde{\mathcal{O}}(1)$ yield a near-optimal sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-3})$.

Under mild additional assumptions on the regularity of g and the joint distribution $\mathbb{P}_{(\xi,\eta)}$ given below, we show that the basis of Chebyshev polynomials is such that the growth of $M_\Phi(\epsilon)$ and $m_\Phi^{-1}(\epsilon)$ is sub-polynomial in ϵ^{-1}. These assumptions are formalized as follows.

Theorem 2.4. *Suppose that Assumption 2.1 holds, along with the following conditions:*

(c.1) The cardinality of Ξ is finite, or there exists $\underline{c} > 0$ such that the density of $\mathbb{P}_\xi$ is lower bounded by $\underline{c}$ on Ξ.

(c.2) The support Ξ is bounded.

(c.3) The function $G(x, y, \xi) \triangleq \mathbb{E}_{\eta|\xi}\left[g(x, y, \xi, \eta)\right]$ is analytic in (y, ξ) for any fixed x.

Then the Chebyshev polynomial basis Φ satisfies $M_\Phi(\epsilon) = \widetilde{\mathcal{O}}\left(1\right)$ and $m_\Phi^{-1}(\epsilon) = \widetilde{\mathcal{O}}\left(1\right)$.

This Theorem uses the exponentially decreasing uniform error of approximation of Chebyshev series [14, 15, 16, 17, 18, 19]. We extend these convergence results to multivariate analytic functions, and, combined with the fact that real-analytic functions are analytically continuable in an open complex set, show that the number of basis functions required for Φ^ϵ to satisfy (1) grows sub-polynomially in ϵ^{-1}. Along with Theorems 2.2 and 2.3, it follows that one can reduce (CSBO) to (SBO$_{\Phi^\epsilon}$) using the Chebyshev basis and solve the latter with $\widetilde{\mathcal{O}}(\epsilon^{-3})$ samples as the terms involving M_Φ and $1/m_\Phi$ in Theorem 2.3 collapse into polylogarithmic factors.

Corollary 2.5. *Under the conditions of Theorem 2.4, an ϵ-stationary solution to* (CSBO) *can be achieved with a near-optimal sample complexity of* $\widetilde{\mathcal{O}}\left(\epsilon^{-3}\right)$.

3 Numerical Experiments

Although standard SBO covers a broad range of settings, it becomes computationally impractical when the LL problem is context-dependent or when a large (potentially infinite) number of LL problems must be solved. CSBO addresses these limitations by making the LL approximation more expressive and scalable, thereby enabling us to solve complex problems ranging from hyperparameter optimization [20, 21], inverse optimization [22], meta-learning [23], reinforcement learning from human feedback [24], and personalized federated learning [25].

We compare our proposed reduction framework against STOCBIO [1] applied to discretized SBO approximations of CSBO. For STOCBIO$[N]$, the context space Ξ is partitioned uniformly into N intervals, each treated as a subproblem, yielding a SBO formulation with N subproblems.

Our method, denoted $\mathcal{R}_\Phi[N]$, uses the first N elements of the basis Φ and STOCBIO[1] as a backbone to solve the reduced SBO problem. We experiment with monomial, Chebyshev, and Fourier bases; we report only Chebyshev and monomial results, as Fourier and Chebyshev perform nearly identically in that setting. Importantly, for any fixed N, $\textsc{stocBiO}[N]$ and $\mathcal{R}_\Phi[N]$ have identical memory usage: $d_x + Nd_y$, enabling a fair comparison under fixed resource constraints.

We consider a temperature-dependent data-cleaning task inspired by [26], a challenging hyperparameter optimization problem. In this experiment, we aim to train a linear classifier for MNIST [27] where a fixed fraction $p = 0.3$ of the training labels have been randomly corrupted. Here, the UL seeks to weight the training points to minimize the expected validation loss across a distribution of model temperatures $\mathbb{P}_\xi$, while the LL computes, for each ξ, the model parameters y by minimizing a weighted and regularized training loss. Formally:

$$\begin{aligned} \min_x \quad & \mathbb{E}_{\xi\sim\mathbb{P}_\xi}\mathbb{E}_{D\sim\mathcal{D}_{\text{val}}}\left[\mathcal{L}_\xi(y^\star(x,\xi);D)\right] \\ \text{s.t.} \quad & y^\star(x,\xi) = \arg\min_y \mathbb{E}_{D\sim\mathcal{D}_{\text{train}}}\left[\sigma(x)\mathcal{L}_\xi(y;D) + \lambda\|y\|^2\right] \end{aligned} \tag{2}$$

where $\mathcal{L}_\xi$ is a temperature-specific convex loss, and $\mathbb{P}_\xi$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{train}}$, denote the distributions over temperatures, validation data, and training data, respectively. This CSBO formulation encourages robustness to label noise while generalizing across temperatures.

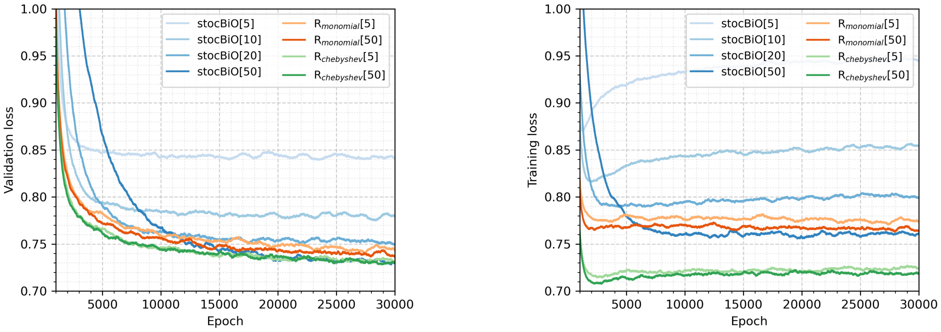

Figure 1: Moving average of the validation (left) and training (right) losses over epochs of stocBiO as well as monomial and Chebyshev bases on hyperparameter optimization.

Figure 1 compares our method $\mathcal{R}_\Phi[N]$ against $\textsc{stocBiO}[N]$ baselines. By discretizing the context space into N subproblems, larger N with $\textsc{stocBiO}[N]$ yields lower final validation loss at the cost of slower convergence. The high training loss of $\textsc{stocBiO}[N]$ indicates its inability to accurately estimate $y^\star$ across all contexts and adapt to new training weights. In contrast, our method cuts computational overhead by learning global coefficients across all contexts. As a result, $\mathcal{R}_{\text{Chebyshev}}[N]$ delivers both the fastest convergence and the lowest final training and validation losses. In comparison, the monomial basis lacks the expressiveness to substantially improve the accuracy of the lower-level solution approximation as N grows.

Due to space limits, we provide a second experiment on inverse optimization in Appendix A.2, further supporting that, by leveraging the continuity of $y^\star$, the parametrization can achieve similar or better accuracy using an order of magnitude less memory or, with the same memory budget, an order of magnitude lower optimality gap.

4 Conclusion

We presented a framework that reduces any Contextual Stochastic Bilevel Optimization (CSBO) problem to a standard Stochastic Bilevel Optimization (SBO) problem by parameterizing the lower-level solution with expressive feature maps. This decouples the context from the lower-level decision and removes the need for conditional sampling oracles. Under smoothness and strong-convexity assumptions, we showed that choosing Chebyshev polynomials as a basis yields a sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-3})$, an order-of-magnitude improvement over existing CSBO methods and matching the lower bound for nonconvex stochastic optimization up to logarithmic factors. Experiments on hyperparameter tuning and inverse optimization confirm faster convergence, lower final loss, and reduced memory compared to baselines. More broadly, this work helps bridge the gap between CSBO and SBO, offering a principled and efficient method that advances our understanding of bilevel optimization in context-rich learning environments.

References

[1] Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4882–4892. PMLR, 18–24 Jul 2021.

[2] Tianyi Chen, Yuejiao Sun, Quan Xiao, and Wotao Yin. A single-timescale method for stochastic bilevel optimization, 2022.

[3] Zhishuai Guo, Quanqi Hu, Lijun Zhang, and Tianbao Yang. Randomized stochastic variance-reduced methods for multi-task stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

[4] 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. In *Advances in Neural Information Processing Systems*, volume 34, pages 30271–30283. Curran Associates, Inc., 2021.

[5] Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Achieving $\mathcal{O}(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. *arXiv preprint arXiv:2312.03807*, 2023.

[6] Tianshu Chu, Dachuan Xu, Wei Yao, and Jin Zhang. Spaba: A single-loop and probabilistic stochastic bilevel algorithm achieving optimal sample complexity. *arXiv preprint arXiv:2405.18777*, 2024.

[7] Yifan Hu, Jie Wang, Yao Xie, Andreas Krause, and Daniel Kuhn. Contextual stochastic bilevel optimization. In *Advances in Neural Information Processing Systems*, volume 36, pages 78412–78434. Curran Associates, Inc., 2023.

[8] Ieva Petrulionyte, Julien Mairal, and Michael Arbel. Functional bilevel optimization for machine learning. *arXiv preprint arXiv:2403.20233*, 2024.

[9] Yibing Lv, Tiesong Hu, Guangmin Wang, and Zhongping Wan. A neural network approach for solving nonlinear bilevel programming problem. *Computers & Mathematics with Applications*, 55(12):2823–2829, 2008.

[10] Rahul Mihir Patel, Justin Dumouchelle, Elias Khalil, and Merve Bodur. Neur2sp: Neural two-stage stochastic programming. *Advances in neural information processing systems*, 35: 23992–24005, 2022.

[11] Jan Kronqvist, Boda Li, Jan Rolfes, and Shudian Zhao. Alternating mixed-integer programming and neural network training for approximating stochastic two-stage problems. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 124–139. Springer, 2023.

[12] Justin Dumouchelle, Esther Julien, Jannis Kurtz, and Elias Boutros Khalil. Neur2ro: Neural two-stage robust optimization. In *The Twelfth International Conference on Learning Representations*, 2023.

[13] Justin Dumouchelle, Esther Julien, Jannis Kurtz, and Elias B Khalil. Neur2bilo: Neural bilevel optimization. *arXiv preprint arXiv:2402.02552*, 2024.

[14] Walter Gautschi. *Orthogonal polynomials: computation and approximation*. OUP Oxford, 2004.

[15] Theodore S Chihara. *An introduction to orthogonal polynomials*. Courier Corporation, 2011.

[16] John P Boyd. *Chebyshev and Fourier spectral methods*. Courier Corporation, 2001.

[17] Serge Bernstein. *Sur l'ordre de la meilleure approximation des fonctions continues par des polynômes de degré donné*, volume 4. Hayez, imprimeur des académies royales, 1912.

[18] Lloyd N Trefethen. *Approximation theory and approximation practice, extended edition*. SIAM, 2019.

[19] Ben Adcock and Daan Huybrechs. On the resolution power of fourier extensions for oscillatory functions. *Journal of Computational and Applied Mathematics*, 260:312–336, 2014. ISSN 0377-0427.

[20] Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1723–1732. PMLR, 2019.

[21] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pages 1568–1577. PMLR, 2018.

[22] Ravindra K Ahuja and James B Orlin. Inverse optimization. *Operations research*, 49(5): 771–783, 2001.

[23] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

[24] Souradip Chakraborty, Amrit Bedi, Alec Koppel, Huazheng Wang, Dinesh Manocha, Mengdi Wang, and Furong Huang. Parl: A unified framework for policy alignment in reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

[25] Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, pages 9489–9502. PMLR, 2021.

[26] Quanqi Hu, Zi-Hao Qiu, Zhishuai Guo, Lijun Zhang, and Tianbao Yang. Blockwise stochastic variance-reduced methods with parallel speedup for multi-block bilevel optimization, 2023.

[27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[28] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1): 2489–2512, 2014.

[29] Martin Beckmann, Charles B McGuire, and Christopher B Winsten. Studies in the economics of transportation. Technical report, 1956.

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

[31] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

A Further Specifications on Numerical Experiments

Experiments are run in parallel on an AMD EPYC 9734 processor. Each individual run uses 16MB for a runtime ranging from 10 to 90 minutes, depending on the application and parameters used.

We omit the DL-SGD and RT-MLMC schemes [7] from our benchmarks because of their inner-loop strategy: both schemes use start from an arbitrary y_0 and use the EPOCH-GD algorithm of [28] without the projection step. The convergence of these schemes thus relies on a choice of lower-level step-size $\beta_0 \leq 1/L_{g,1}$, potentially preventing the recommended choice of $\beta_0 = 1/(4\mu)$. In practice, those schemes perform best when either:

1. $\{y^\star(x,\xi) :\ \xi \in \Xi\}$ is contained in a small ball around some $y_0(x)$, where $y_0(x)$ is known or estimable from previous iterations.
2. The lower-level problem is sufficiently well-conditioned that only a few inner-loop steps suffice for convergence.

In our experiments, such a $y_0(x)$ does not exist since the lower level solutions $y^\star(x,\cdot)$ vary substantially in ξ. Furthermore, depending on the inner-loop step size, the solution either diverges or requires a large number of steps to provide a reasonable estimate of $y^\star$ (and thus ∇F), resulting in an excessive runtime.

A.1 Hyper-parameters Optimization

We compare the performance of our proposed algorithm with stocBio, a reference algorithm that outperforms other baseline algorithms BSA, TTSA, HOAG on the MNIST Data Hyper-Cleaning task, a hyperparameter optimization problem. The dataset consists in 19,000 for training and 1,000 images for validation. The objective of Data hyper-cleaning involves training classifiers on a dataset where each label has been randomly and independently corrupted with probability p; that is, each label is replaced by a random class with chance p. The classifiers have losses with different temperatures. Formally, the objective function is:

$$\begin{aligned} \min_x \quad & \mathbb{E}_{\xi\sim\mathbb{P}_\xi}\left[\frac{1}{|\mathcal{D}_{\text{val}}|}\sum_{(X_i,Y_i)\in\mathcal{D}_{\text{val}}} L(y^\star(x,\xi)X_i/\xi),Y_i)\right] \\ \text{s.t.} \quad & y^\star(x,\xi) = \arg\min_y \frac{1}{|\mathcal{D}_{\text{train}}|}\sum_{(X_i,Y_i)\in\mathcal{D}_{\text{train}}} \sigma(x)L(yX_i/\xi),Y_i) + \lambda\|y\|^2 \end{aligned}$$

where L is the cross-entropy loss and $\sigma(\cdot)$ is the sigmoid function. Here $\mathbb{P}_\xi = \mathcal{U}(0.1, 10)$ and we choose the regularization parameter $\lambda = 10^{-3}$. The results are averaged over 20 randomized trials. We use a batch size of 512 and use grid search to choose: the inner-loop stepsize β from $\{0.001, 0.01, 0.1, 1, 10\}$, the outer-loop stepsize α from $\{10^k\ :\ k \in [\![-5, 5]\!]\}$, and the number of inner-loop steps T_{inner} from $\{1, 10, 100\}$. We choose a number of samples $K = 10$ and a scaling $s = 10^{-2}$ to approximate $\left(\mathbb{E}_{\eta\sim\mathbb{P}_{\eta|\xi}}\nabla^2_{22}g(x,y,\xi,\eta)\right)^{-1}$ as $s\cdot\prod_{k=1}^{K}\left(I - s\cdot\nabla^2_{22}g(x,y,\xi,\eta_n)\right)$.

A.2 Inverse Optimization

We consider the Static Traffic Assignment (STA) problem [29], which models network equilibrium flows under fixed origin-destination (OD) demand. Using Beckmann's convex-potential formulation and introducing a penalty for constraint violations, we cast inverse capacity estimation as a CSBO problem:

$$\begin{aligned} \min_x \quad & \mathbb{E}_{(\xi,\eta)\sim\mathbb{P}_{(\xi,\eta)}}\|y^\star(x,\xi)-\eta\| \\ \text{s.t.} \quad & y^\star(x,\xi) = \arg\min_y \sum_{e\in\mathcal{E}}\int_0^{y_e} t_e(z;x)dz + \lambda_\xi(y) \end{aligned} \tag{3}$$

where $y^\star(x,\xi)$ is the equilibrium flow given edge capacities x and OD demand ξ, η is the corresponding noisy observation of that flow, t_e is the edge performance function relating the flow to the travel time, and $\lambda_\xi(y)$ penalizes infeasible flows and the violation of OD pairs demand constraints.

We simulate a two-edge network with a single OD pair. We use the edge performance function $t_e(y;x) \triangleq t_{0,e} \cdot \left(1 + \alpha \left(\frac{y_e}{x_e}\right)^{\beta}\right)$ with $\alpha = 1$ and $\beta = 4$. To penalize negative flows and violation of the travel demand ξ, we define

$$\lambda_{\xi}(y) \triangleq \lambda_{\text{demand}} \cdot \left(\left(\xi - \sum_{e\in\mathcal{E}} y_e\right)^{+}\right)^{2} + \lambda_{+} \cdot \sum_{e\in\mathcal{E}} \left(y_e^{-}\right)^{2}$$

where z^{+} (resp. z^{-}) denotes the positive (resp. negative) part of z, $\lambda_{\text{demand}} = 100$, and $\lambda_{+} = 50$. Note that one only needs to penalize insufficient total flow since the edge performance function is strictly increasing with respect to the flow y. The results are averaged over 50 runs, with 95% confidence intervals assuming normally distributed errors across runs. Each run uses a synthetic dataset generated with the following procedure. The free flow travel times t_0 and ground truth capacities $x^{\star}$ are sampled from the uniform distribution supported on $[1,2]^{|\mathcal{E}|}$ and $[0.2, 0.8]^{|\mathcal{E}|}$, respectively. We generate $n_{\text{train}} = 1000$ training samples by drawing each context ξ independently from a standard uniform distribution. For each ξ, we compute the corresponding optimal flow $y^{\star}(x^{\star}, \xi)$ using gradient descent, and set $\eta = y^{\star}(x^{\star}, \xi) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma_0)$. We resample the noise ϵ until η is non-negative. We follow the same procedure for the $n_{\text{test}} = 1000$ samples forming the test set. We follow the same steps as in the hyperparameter optimization experiment to select α, β, T_{inner}, and set $K = 10$ and $s = 10^{-3}$.

For every trial, we sample a ground truth capacity vector $x^{\star}$ and generate training and test sets with 10^3 i.i.d. samples (ξ, η) each. The final solution $(\bar{x}, \bar{y}(\cdot))$ is averaged over the last 10% epochs. Performance is evaluated via the test loss $F(\bar{x})$, evaluated over the test set after computing the exact lower-level solutions $y^{\star}(\bar{x}, \xi)$. We also report the expected lower-level error $\Delta_y \triangleq \mathbb{E}_{\xi}\|y^{\star}(\bar{x}, \xi) - \bar{y}(\xi)\|^2$ and the the upper-level parameter error $\Delta_x \triangleq \|\bar{x} - x^{\star}\|^2$.

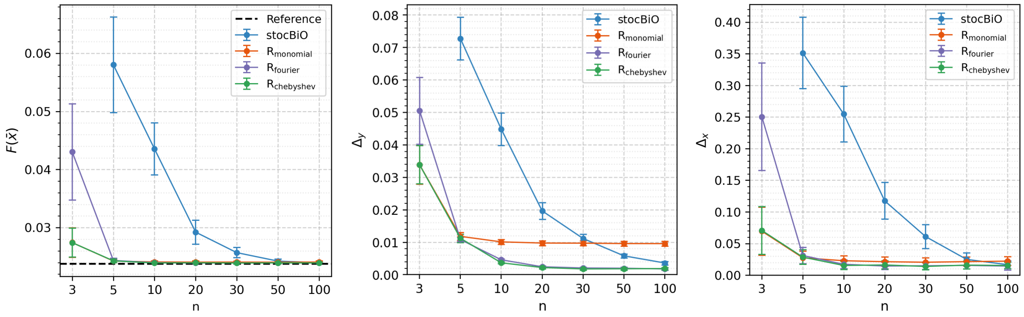

Figure 2: Loss $F(\bar{x})$ (left), lower level solution error Δ_y (center), and upper level solution error Δ_x (right) of stocBiO and our reduction framework using monomial, Fourier, and Chebyshev bases. The reference loss is $F_{\text{ref}} = F(x^{\star})$.

Figure 2 demonstrates the advantage of our CSBO reduction. As few as $N = 3$ basis functions suffice to produce reasonable solutions, whereas stocBiO struggles with instability or excessive sample requirements. By $N = 5$, all three bases close the optimality gap to within 3%, while stocBiO achieves a comparable accuracy at $N = 50$. Beyond $N = 5$, the lower-level error stops decreasing significantly for the monomial basis, while the more expressive Chebyshev and Fourier bases continue to reduce the error and consistently outperform stocBiO at every N.

At $N = 10$, $\mathcal{R}_{\text{Chebyshev}}$ and $\mathcal{R}_{\text{Fourier}}$ achieve the best performance, whereas stocBiO requires $N = 100$, and still fails to match training loss. These results validate our theoretical claims: expressive bases can compactly approximate context-sensitive solutions in CSBO. Compared to partition-based SBO approximations, the parametrization leverages the continuity of $y^{\star}$ to achieve similar or better accuracy using an order of magnitude less memory or, with the same memory budget, an order of magnitude lower optimality gap.

B Similarity of Hypergradients

Proposition B.1. *Under assumption 2.1, the following holds for any $x \in \mathbb{R}^{d_x}$:*

$$\|\nabla F(x) - \nabla F_{\Phi^\epsilon}(x)\| \leq K \cdot \mathbb{E}_\xi \|y_{\Phi^\epsilon}(W^\star(x), \xi) - y^\star(x, \xi)\| \tag{4}$$

where $K = L_{f,1} + \frac{L_{g,2}L_{f,0}}{\mu} + \frac{L_{g,2}L_{g,1}L_{f,0}}{\mu^2} + \frac{L_{f,1}L_{g,1}}{\mu}$.

Proof. Let $x \in \mathbb{R}^{d_x}$. Recall that $\nabla F(x) = \mathbb{E}_{(\xi,\eta)}\left[\nabla F(x,\xi,\eta)\right]$ and $\nabla F_{\Phi^\epsilon}(x) = \mathbb{E}_{(\xi,\eta)}\left[\nabla F_{\Phi^\epsilon}(x,\xi,\eta)\right]$, where:

$$\begin{aligned}
\nabla F(x,\xi,\eta) &\triangleq \nabla_1 f(x, y^\star(x,\xi), \xi, \eta) \\
&\quad - \nabla_{12} g(x, y^\star(x,\xi), \xi, \eta)\left[\nabla^2_{22} g(x, y^\star(x,\xi), \xi, \eta)\right]^{-1} \nabla_2 f(x, y^\star(x,\xi), \xi, \eta) \\
\nabla F_{\Phi^\epsilon}(x,\xi,\eta) &\triangleq \nabla_1 f(x, y_\Phi(W^\star(x),\xi), \xi, \eta) \\
&\quad - \nabla_{12} g(x, y_\Phi(W^\star(x),\xi), \xi, \eta)\left[\nabla^2_{22} g(x, y_\Phi(W^\star(x),\xi), \xi, \eta)\right]^{-1} \nabla_2 f(x, y_\Phi(W^\star(x),\xi), \xi, \eta)
\end{aligned}$$

Consider a sample (ξ, η). For readability, we define $\Delta_y(x,\xi) \triangleq \|y_\Phi(W^\star(x),\xi) - y^\star(x,\xi)\|$ and:

$$\begin{aligned}
A(y_\Phi) &= \nabla_{12} g(x, y_\Phi(W^\star(x),\xi), \xi, \eta) & A(y^\star) &= \nabla_{12} g(x, y^\star(x,\xi), \xi, \eta) \\
B(y_\Phi) &= \nabla^2_{22} g(x, y_\Phi(W^\star(x),\xi), \xi, \eta) & B(y^\star) &= \nabla^2_{22} g(x, y^\star(x,\xi), \xi, \eta) \\
C(y_\Phi) &= \nabla_2 f(x, y_\Phi(W^\star(x),\xi), \xi, \eta) & C(y^\star) &= \nabla_2 f(x, y^\star(x,\xi), \xi, \eta)
\end{aligned}$$

We can then write and bound $\|\nabla F_{\Phi^\epsilon}(x,\xi,\eta) - \nabla F(x,\xi,\eta)\|$ using the triangular inequality as:

$$\begin{aligned}
\|\nabla F_{\Phi^\epsilon}(x,\xi,\eta) - \nabla F(x,\xi,\eta)\| &\leq \|\nabla_1 f(x, y_\Phi(W^\star(x),\xi), \xi, \eta) - \nabla_1 f(x, y^\star(x,\xi), \xi, \eta)\| \\
&\quad + \|A(y_\Phi)B(y_\Phi)^{-1}C(y_\Phi) - A(y^\star)B(y^\star)^{-1}C(y^\star)\| \\
&\leq L_{f,1}\Delta_y(x,\xi) + \|A(y_\Phi) - A(y^\star)\| \cdot \|B(y^\star)^{-1}\| \cdot \|C(y^\star)\| \\
&\quad + \|B(y_\Phi)^{-1} - B(y^\star)^{-1}\| \cdot \|A(y_\Phi)\| \cdot \|C(y_\Phi)\| \\
&\quad + \|C(y_\Phi) - C(y^\star)\| \cdot \|A(y_\Phi)\| \cdot \|B(y^\star)^{-1}\|
\end{aligned}$$

From the regularity of f, g, and their gradient given in Assumption 2.1, we obtain:

$$\begin{aligned}
\|A(y_\Phi) - A(y^\star)\| &\leq L_{g,2}\Delta_y(x,\xi) \\
\|B(y_\Phi)^{-1} - B(y^\star)^{-1}\| &\leq \|B(y_\Phi)^{-1}\| \cdot \|B(y_\Phi) - B(y^\star)\| \cdot \|B(y^\star)^{-1}\| \leq \frac{L_{g,2}}{\mu^2}\Delta_y(x,\xi) \\
\|C(y_\Phi) - C(y^\star)\| &\leq L_{f,1}\Delta_y(x,\xi) \\
\|A(y_\Phi)\| &\leq L_{g,1}, \\
\|A(y^\star)\| &\leq L_{g,1}, \\
\|C(y_\Phi)\| &\leq L_{f,0}, \\
\|C(y^\star)\| &\leq L_{f,0},
\end{aligned}$$

where the second inequality uses the identity $\|X^{-1} - Y^{-1}\| \leq \|X^{-1}\| \cdot \|X - Y\| \cdot \|Y^{-1}\|$ and the fact that $\|B(y_\Phi)^{-1}\|$ and $\|B(y^\star)^{-1}\|$ are upper bounded by $1/\mu$ from the strong convexity of g.

It follows that:

$$\begin{aligned}
\|\nabla F_{\Phi^\epsilon}(x,\xi,\eta) - \nabla F(x,\xi,\eta)\| &\leq L_{f,1}\Delta_y(x,\xi) + \frac{L_{g,2}L_{f,0}}{\mu}\Delta_y(x,\xi) \\
&\quad + \frac{L_{g,2}L_{g,1}L_{f,0}}{\mu^2}\Delta_y(x,\xi) + \frac{L_{f,1}L_{g,1}L_{g,2}}{\mu}\Delta_y(x,\xi) \\
&= K\Delta_y(x,\xi)
\end{aligned}$$

where $K = L_{f,1} + \frac{L_{g,2}L_{f,0}}{\mu} + \frac{L_{g,2}L_{g,1}L_{f,0}}{\mu^2} + \frac{L_{f,1}L_{g,1}}{\mu}$.

We then bound the difference between the hypergradients using Jensen's inequality:

$$\begin{aligned}\|\nabla F_{\Phi^\epsilon}(x) - \nabla F(x)\| &\leq \mathbb{E}_{(\xi,\eta)}\|\nabla F_{\Phi^\epsilon}(x,\xi,\eta) - \nabla F(x,\xi,\eta)\| \\ &\leq K \cdot \mathbb{E}_{(\xi,\eta)}\left[\Delta_y(x,\xi)\right] \\ &= K \cdot \mathbb{E}_\xi \|y_\Phi(W^\star(x),\xi) - y^\star(x,\xi)\|\end{aligned}$$

□

Proposition B.2. *Under assumption 2.1, we have for any $x \in \mathbb{R}^{d_x}$ and $W \in \mathbb{R}^{d_y \times N_\Phi(\epsilon)}$:*

$$\mathbb{E}_\xi \|y_{\Phi^\epsilon}(W^\star(x),\xi) - y^\star(x,\xi)\|^2 \leq \frac{2L_{g,1}}{\mu} \cdot \mathbb{E}_\xi \|y_{\Phi^\epsilon}(W,\xi) - y^\star(x,\xi)\|^2. \tag{5}$$

Proof. Let $x \in \mathbb{R}^{d_x}$, $W \in \mathbb{R}^{d_y \times N}$, and $G(x,y,\xi) \triangleq \mathbb{E}_{\eta \sim \mathbb{P}_{\eta|\xi}}\left[g(x,y,\xi,\eta)\right]$. As g is $L_{g,1}$-smooth in (x,y) for any (ξ,η), and μ-strongly convex in y for any fixed (x,ξ,η), so is G. Since $W^\star(x)$ minimizes $\mathbb{E}_\xi\left[G(x,y_\Phi(\cdot,\xi),\xi)\right]$, we have in particular:

$$\mathbb{E}_\xi\left[G(x,y_\Phi(W^\star(x),\xi),\xi)\right] \leq \mathbb{E}_\xi\left[G(x,y_\Phi(W,\xi),\xi)\right] \tag{6}$$

Additionally, since $y^\star(x,\xi)$ minimizes $G(x,\cdot,\xi)$, it also holds that:

$$G(x,y^\star(x,\xi),\xi) \leq G(x,y_\Phi(W^\star(x),\xi),\xi)$$

Using the strong convexity of G at $W^\star(x)$ we thus obtain:

$$G(x,y_\Phi(W^\star(x),\xi),\xi) - G(x,y^\star(x,\xi),\xi) \geq \frac{\mu}{2}\|y_\Phi(W^\star(x),\xi) - y^\star(x,\xi)\|^2, \quad \forall \xi \in \Xi \tag{7}$$

On the other hand, the smoothness of G yields:

$$|G(x,y_\Phi(W,\xi),\xi) - G(x,y^\star(x,\xi),\xi)| \leq L_{g,1}\|y_\Phi(W,\xi) - y^\star(x,\xi)\|^2, \quad \forall \xi \in \Xi \tag{8}$$

Taking the expectation over ξ on both sides in (7) and (8), and using the inequality (6) we have:

$$\begin{aligned}\mathbb{E}_\xi \|y_\Phi(W^\star(x),\xi) - y^\star(x,\xi)\|^2 &\leq \frac{2}{\mu}\mathbb{E}_\xi\left[G(x,y_\Phi(W^\star(x),\xi),\xi) - G(x,y^\star(x,\xi),\xi)\right] \\ &\leq \frac{2}{\mu}\mathbb{E}_\xi\left[G(x,y_\Phi(W,\xi),\xi) - G(x,y^\star(x,\xi),\xi)\right] \\ &\leq \frac{2L_{g,1}}{\mu}\mathbb{E}_\xi\|y_\Phi(W,\xi) - y^\star(x,\xi)\|^2\end{aligned}$$

As this holds for any $x \in \mathbb{R}^{d_x}$, we obtain the desired result. □

We then proceed with the proof of Theorem 2.2.

Proof. Combining Propositions B.1 and B.2, we obtain:

$$\begin{aligned}\mathbb{E}\|\nabla F(x^\star)\|^2 &\leq \mathbb{E}\|\nabla F_{\Phi^\epsilon}(x^\star)\|^2 + \mathbb{E}\|\nabla F(x^\star) - \nabla F_{\Phi^\epsilon}(x^\star)\|^2 \\ &\overset{(1)}{\leq} \frac{\epsilon^2}{2} + \mathbb{E}\left[K^2\left(\mathbb{E}_\xi\|y_{\Phi^\epsilon}(W^\star(x^\star),\xi) - y^\star(x^\star,\xi)\|\right)^2\right] \\ &\overset{(2)}{\leq} \frac{\epsilon^2}{2} + \mathbb{E}\left[\frac{2K^2 L_{g,1}}{\mu}\mathbb{E}_\xi\|y_{\Phi^\epsilon}(W^\dagger(x^\star),\xi) - y^\star(x^\star,\xi)\|^2\right] \\ &\overset{(3)}{\leq} \frac{\epsilon^2}{2} + \frac{2K^2 L_{g,1}}{\mu} \cdot \frac{\epsilon^2 \mu}{4K^2 L_{g,1}} \\ &\leq \epsilon^2\end{aligned}$$

where (1) uses Proposition B.1 and the fact that $(x^\star, W^\star(x^\star))$ is an $\frac{\epsilon}{\sqrt{2}}$-stationary solution to (SBO$_{\Phi^\epsilon}$), (2) results from Proposition B.2 and Jensen's inequality, and (3) holds since Φ^ϵ satisfies (1). □

C Regularity of (SBO$_\Phi$)

For readability, we refer in this section to $M_\Phi(\epsilon)$ as M_Φ, and to $m_\Phi(\epsilon)$ as m_Φ.

Lemma C.1. *If $A \succeq \mu I_{d_A}$ and $B \succeq 0$ with $\mu \geq 0$, then $A \otimes B \succeq \mu I_{d_A} \otimes B$ and $B \otimes A \succeq \mu B \otimes I_{d_A}$.*

Proof. By the bilinearity of the Kronecker product we have:

$$A \otimes B - \mu I_{d_A} \otimes B = (A - \mu I_{d_A}) \otimes B$$

Since the Kronecker product of two positive definite matrix is positive definite, it holds that $(A - \mu I_{d_A}) \otimes B \succeq 0$ and thus $A \otimes B - \mu A \otimes I_{d_B} \succeq 0$. With a symmetrical argument, we obtain $B \otimes A \succeq \mu B \otimes I_{d_A}$. □

C.1 Strong convexity of $G_{\Phi^\epsilon}(x, W) \triangleq \mathbb{E}_{(\xi,\eta)}\left[g(x, W\Phi(\xi), \xi, \eta)\right]$

Lemma C.2. *Under assumptions* 2.1*-(ii) and if Φ is well-conditioned, G_{Φ^ϵ} is $\mu m_\Phi(\epsilon)$-strongly convex in W for any fixed $x \in \mathbb{R}^{d_x}$.*

Proof. We have for any fixed $x \in \mathbb{R}^{d_x}$ that G_{Φ^ϵ} is twice differentiable with respect to W as the expectation of compositions of twice differentiable and linear mapping. Additionally:

$$\nabla^2_{22} G_{\Phi^\epsilon}(x, W) = \mathbb{E}_\xi\left[\nabla^2_{22} G(x, W\Phi(\xi), \xi) \otimes \Phi(\xi)\Phi(\xi)^\top\right]$$

Since G is μ-strongly convexity with respect to y with $\mu > 0$ and $\mathbb{E}_\xi\left[\Phi(\xi)\Phi(\xi)^\top\right] \succeq m_\Phi(\epsilon) I_{N_\Phi(\epsilon)} \succeq 0$, we obtain using Lemma C.1 twice:

$$\begin{aligned}
\nabla^2_{22} G_{\Phi^\epsilon}(x, W) &\succeq \mu \mathbb{E}_\xi\left[I_{d_y} \otimes \Phi(\xi)\Phi(\xi)^\top\right] \\
&= \mu I_{d_y} \otimes \mathbb{E}_\xi\left[\Phi(\xi)\Phi(\xi)^\top\right] \\
&\succeq \mu m_\Phi(\epsilon) I_{d_y} \otimes I_{N_\Phi(\epsilon)} \\
&= \mu m_\Phi(\epsilon) I_{d_y \cdot N_\Phi(\epsilon)}
\end{aligned}$$

and we conclude that $G_{\Phi^\epsilon}(x, W)$ is $\mu m_\Phi(\epsilon)$-strongly convex with respect to W for any $x \in \mathbb{R}_{d_x}$. □

C.2 Lipschitz continuity

Lemma C.3. *Let $h(x, y, \xi, \eta) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \Xi \times \mathbb{R}^{d_\eta} \to \mathbb{R}^n$ and $l(\xi) : \Xi \to \mathbb{R}^N$. Suppose the following conditions hold:*

1. *h is L-Lipschitz continuous with respect to (x, y) for any fixed (ξ, η).*
2. *There exists a constant M such that:*
$$\|l(\xi)\| \leq M, \quad \forall \xi \in \Xi$$

Then the mapping $h_\Phi : (x, W, \xi, \eta) \mapsto h(x, W\Phi(\xi), \xi, \eta) l(\xi)$ is $LM_\Phi M$-Lipschitz continuous with respect to (x, y) for all fixed (ξ, η).

Proof. Since $h(x, y, \xi, \eta)$ is L-Lipschitz continuous in (x, y) and $\sup_{\xi \in \Xi} ||\Phi(\xi)|| \leq M_\Phi$, we have for any (x, W) and (x', W'):

$$\begin{aligned}
\|h_\Phi(x, W, \xi, \eta) - h_\Phi(x', W', \xi, \eta)\| &= \|h(x, W\Phi(\xi), \xi, \eta) l(\xi) - h(x', W'\Phi(\xi), \xi, \eta) l(\xi)\| \\
&\overset{(1)}{\leq} \|h(x, W\Phi(\xi), \xi, \eta) - h(x', W'\Phi(\xi), \xi, \eta)\| \cdot \|l(\xi)\| \\
&\overset{(2)}{\leq} ML\left(\|x - x'\| + \|W\Phi(\xi) - W'\Phi(\xi)\|\right) \\
&\overset{(3)}{\leq} ML\left(\|x - x'\| + M_\Phi \|W - W'\|\right) \\
&\overset{(4)}{\leq} LM_\Phi M\left(\|x - x'\| + \|W - W'\|\right)
\end{aligned}$$

where (1) uses the sub-multiplicativity of $\|\cdot\|$, (2) follows from the Lipschitz continuity of h and this inequality $\|l(\xi)\| \leq M$, (3) uses again the sub-multiplicativity of $\|\cdot\|$ and the inequality $\|\Phi(\xi)\| \leq M_\Phi(\epsilon)$, and (4) holds since $1 \leq M_\Phi$. Therefore h_Φ is $LM_\Phi M$-Lipschitz continuous in (x, W). □

Lemma C.4. *The following hold under assumption* 2.1*:*

1. *The functions $f_{\Phi^\epsilon}(x, W, \xi, \eta)$ is $L_{f,0}M_\Phi$-Lipschitz with respect to x and W.*
2. *The gradients $\nabla f_{\Phi^\epsilon}(x, W, \xi, \eta)$ and $\nabla g_{\Phi^\epsilon}(x, W, \xi, \eta)$ are $L_{f,1}M_\Phi^2$ and $L_{g,1}M_\Phi^2$-Lipschitz, respectively, with respect to x and W.*
3. *The second order gradients $\nabla^2_{12} g_{\Phi^\epsilon}(x, W, \xi, \eta)$ and $\nabla^2_{22} g_{\Phi^\epsilon}(x, W, \xi, \eta)$ are $L_{g,2}M_\Phi^3$-Lipschitz with respect to x and W.*

Proof. Under assumption 2.1-(i), the mappings f, ∇f, ∇g, and $\nabla^2 g$, are $L_{f,0}$, $L_{f,1}$, $L_{g,1}$, and $L_{g,2}$-Lipschitz continuous with respect to (x, y) for any fixed (ξ, η), respectively. Additionally, the chain rule gives:

$$\begin{aligned}
\nabla_1 f_{\Phi^\epsilon}(x, W, \xi, \eta) &= \nabla_1 f(x, W\Phi(\xi), \xi, \eta) \\
\nabla_2 f_{\Phi^\epsilon}(x, W, \xi, \eta) &= \nabla_2 f(x, W\Phi(\xi), \xi, \eta)\Phi(\xi)^\top \\
\nabla_1 g_{\Phi^\epsilon}(x, W, \xi, \eta) &= \nabla_1 g(x, W\Phi(\xi), \xi, \eta) \\
\nabla_2 g_{\Phi^\epsilon}(x, W, \xi, \eta) &= \nabla_2 g(x, W\Phi(\xi), \xi, \eta)\Phi(\xi)^\top \\
\nabla^2_{12} g_{\Phi^\epsilon}(x, W, \xi, \eta) &= \nabla^2_{12} g(x, W\Phi(\xi), \xi, \eta)\Phi(\xi)^\top \\
\nabla^2_{22} g_{\Phi^\epsilon}(x, W, \xi, \eta) &= \nabla^2_{22} g(x, W\Phi(\xi), \xi, \eta) \otimes \Phi(\xi)\Phi(\xi)^\top
\end{aligned}$$

We can then use Lemma C.3 to obtain the following:

h	l	L	M	Lipschitz coefficient of $h(x, W\Phi(\xi), \xi, \eta)$ w.r.t. (x, W)
f_{Φ^ϵ}	1	$L_{f,0}$	1	$L_{f,0}M_\Phi$
$\nabla_1 f_{\Phi^\epsilon}$	I_{d_x}	$L_{f,1}$	1	$L_{f,1}M_\Phi$
$\nabla_2 f_{\Phi^\epsilon}$	$\Phi^\top$	$L_{f,1}$	M_Φ	$L_{f,1}M_\Phi^2$
$\nabla_1 g_{\Phi^\epsilon}$	I_{d_x}	$L_{g,1}$	1	$L_{g,1}M_\Phi$
$\nabla_2 g_{\Phi^\epsilon}$	$\Phi^\top$	$L_{g,1}$	M_Φ	$L_{g,1}M_\Phi^2$
$\nabla^2_{12} g_{\Phi^\epsilon}$	$\Phi^\top$	$L_{g,2}$	M_Φ	$L_{g,2}M_\Phi^2$
$\nabla^2_{22} g_{\Phi^\epsilon}$	$\Phi\Phi^\top$	$L_{g,2}$	M_Φ^2	$L_{g,2}M_\Phi^3$

and we conclude using the inequality $1 \leq M_\Phi$. □

C.3 Bounded variance

Lemma C.5. *Under assumption 2.1, then for all ξ, $y^\star(\cdot, \xi)$ is L_y-Lipschitz continuous with:*

$$L_y = \frac{L_{g,1}}{\mu}$$

Proof. Let $g(x, y, \xi) = \mathbb{E}_{\eta|\xi} g(x, W, \xi, \eta)$. By definition of $y^\star$, we have $\nabla_2 g(x, y^\star(x, \xi), \xi) = 0$. Then, by taking the derivative on both sides w.r.t. x, using the chain rule, and the implicit function theorem, we obtain:

$$\nabla^2_{12} g(x, y^\star(x), \xi) + \nabla^2_{22} g(x, y^\star(x, \xi), \xi)\nabla_1 y^\star(x, \xi) = 0$$

It follows that $\|\nabla_1 y^\star(x, \xi)\| \leq \frac{L_{g,1}}{\mu}$. □

Lemma C.6. *Let $h(x, y, \xi, \eta) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \Xi \times \mathbb{R}^{d_\eta} \to \mathbb{R}^n$ and $l(\xi) : \Xi \to \mathbb{R}^N$. Suppose the following conditions hold:*

1. *$h(x, y, \xi, \eta)$ has, conditioned on ξ, a variance bounded by σ^2 i.e.*

$$\mathbb{E}_{\eta \sim \mathbb{P}_{\eta|\xi}} \left\| h(x, y, \xi, \eta) - \mathbb{E}_{\eta' \sim \mathbb{P}_{\eta|\xi}} \left[h(x, y, \xi, \eta') \right] \right\|^2 \leq \sigma^2, \quad \forall x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}, \xi \in \mathbb{R}^{d_\xi}.$$

2. *There exists a constant C such that:*

$$\|h(x,y,\xi,\eta)\| \leq C, \quad \forall x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}, \xi \in \mathbb{R}^{d_\xi}$$

3. *There exists a constant M such that:*

$$\|l(\xi)\| \leq M, \quad \forall \xi \in \Xi$$

Then the mapping $h_\Phi : (x,W,\xi,\eta) \mapsto h(x,W\Phi(\xi),\xi,\eta)l(\xi)$ has a variance bounded by $M^2\left(\sigma^2 + C^2\right)$.

Proof. Let $Z = h_\Phi(x,W,\xi,\eta)$. Using the law of total variance we have:

$$\mathrm{Var}_{(\xi,\eta)}\left[Z\right] = \mathbb{E}_\xi\left[\mathrm{Var}_{\eta|\xi}\left[Z \mid \xi\right]\right] + \mathrm{Var}_\xi\left[\mathbb{E}_{\eta|\xi}\left[Z \mid \xi\right]\right]$$

For all $\xi \in \Xi$ it holds that:

$$\begin{aligned}
\mathrm{Var}_{\eta|\xi}\left[Z \mid \xi\right] &= \mathbb{E}_{\eta|\xi}\left[\left\|h(x,W\Phi(\xi),\xi,\eta)l(\xi) - \mathbb{E}_{\eta'|\xi}\left[h(x,W\Phi(\xi),\xi,\eta')l(\xi)\right]\right\|^2 \middle| \xi\right] \\
&\leq \|l(\xi)\|^2 \cdot \mathbb{E}_{\eta|\xi}\left[\left\|h(x,W\Phi(\xi),\xi,\eta) - \mathbb{E}_{\eta'|\xi}\left[h(x,W\Phi(\xi),\xi,\eta')\right]\right\|^2 \middle| \xi\right] \\
&\leq \|l(\xi)\|^2 \cdot \sigma^2
\end{aligned}$$

where the second inequality follows from the sub-multiplicativity of $\|\cdot\|$ and the fact that $l(\xi)$ is deterministic when conditioned on ξ, and the last inequality holds under condition 1.

Similarly, we have that for all $\xi \in \Xi$ and under condition 2:

$$\begin{aligned}
\left\|\mathbb{E}_{\eta|\xi}\left[Z \mid \xi\right]\right\| &\leq \left\|\mathbb{E}_{\eta|\xi}\left[h(x,W\Phi(\xi),\xi,\eta)l(\xi) \mid \xi\right]\right\| \\
&\leq \|l(\xi)\| \cdot \left\|\mathbb{E}_{\eta|\xi}\left[h(x,W,\xi,\eta) \mid \xi\right]\right\| \\
&\leq \|l(\xi)\| \cdot C
\end{aligned}$$

Combining the above results, we obtain:

$$\begin{aligned}
\mathrm{Var}_{(\xi,\eta)}\left[Z\right] &\leq \mathbb{E}_\xi\left[\|l(\xi)\|^2\sigma^2\right] + \mathbb{E}_\xi\left[\left\|\mathbb{E}_{\eta|\xi}\left[Z \mid \xi\right]\right\|^2\right] \\
&\leq M^2\sigma^2 + \mathbb{E}_\xi\left[\|l(\xi)\|^2 \cdot C^2\right] \\
&\leq M^2\left(\sigma^2 + C^2\right)
\end{aligned}$$

where the last inequality uses condition 3. □

Lemma C.7. *Under assumption 2.1, the mappings $\nabla_1 f_{\Phi^\epsilon}(x,W,\xi,\eta)$, $\nabla_2 f_{\Phi^\epsilon}(x,W,\xi,\eta)$, $\nabla_2 g_{\Phi^\epsilon}(x,W,\xi,\eta)$, $\nabla_{12}^2 g_{\Phi^\epsilon}(x,W,\xi,\eta)$, and $\nabla_{22}^2 g_{\Phi^\epsilon}(x,W,\xi,\eta)$ have a bounded variance for all x and W such that $\|W - W^\star(x)\| \leq \Delta_0$.*

Proof. Under assumption 2.1-(iii) we have for any (x,y,ξ) that the variances of $\nabla_1 f(x,y,\xi,\eta)$ and $\nabla_2 f(x,y,\xi,\eta)$ are upper bounded by σ_f^2, and the variance of $\nabla_{12}^2 g(x,y,\xi,\eta)$ and $\nabla_{22}^2 g(x,y,\xi,\eta)$ by $\sigma_{g,2}^2$. Thus these 4 functions satisfy the first condition of Lemma C.7. From the Lipschitz continuity of f and ∇g (assumption 2.1-(i)) we have for any (x,y,ξ,η):

$$\begin{aligned}
\|\nabla_1 f(x,y,\xi,\eta)\| &\leq L_{f,0} \\
\|\nabla_2 f(x,y,\xi,\eta)\| &\leq L_{f,0} \\
\left\|\nabla_{12}^2 g(x,y,\xi,\eta)\right\| &\leq L_{g,1} \\
\left\|\nabla_{22}^2 g(x,y,\xi,\eta)\right\| &\leq L_{g,1}
\end{aligned}$$

Taking the expectation over $\eta \sim \mathbb{P}_{\eta|\xi}$, the second condition of Lemma C.7 holds for theses 4 functions. We then use the chain rule to get:

$$\begin{aligned}
\nabla_1 f_{\Phi^\epsilon}(x,W,\xi,\eta) &= \nabla_1 f(x,W\Phi(\xi),\xi,\eta) \\
\nabla_2 f_{\Phi^\epsilon}(x,W,\xi,\eta) &= \nabla_2 f(x,W\Phi(\xi),\xi,\eta)\Phi(\xi)^\top \\
\nabla_{12}^2 g_{\Phi^\epsilon}(x,W,\xi,\eta) &= \nabla_{12}^2 g(x,W\Phi(\xi),\xi,\eta)\Phi(\xi)^\top \\
\nabla_{22}^2 g_{\Phi^\epsilon}(x,W,\xi,\eta) &= \nabla_{22}^2 g(x,W\Phi(\xi),\xi,\eta) \otimes \Phi(\xi)\Phi(\xi)^\top
\end{aligned}$$

Since $\|\Phi(\xi)\| \leq M_\Phi$, Lemma C.6 applies and we obtain the following bounds:

h	σ^2	l	C	M	Bound on $\text{Var}_{(\xi,\eta)}\left[h(x,W,\xi,\eta)\right]$
$\nabla_1 f_{\Phi^\epsilon}$	σ_f^2	I_{d_x}	$L_{f,0}$	1	$\sigma_f^2 + L_{f,0}^2$
$\nabla_2 f_{\Phi^\epsilon}$	σ_f^2	$\Phi^\top$	$L_{f,0}$	M_Φ	$M_\Phi^2\left(\sigma_f^2 + L_{f,0}^2\right)$
$\nabla_{12}^2 g_{\Phi^\epsilon}$	$\sigma_{g,2}^2$	$\Phi^\top$	$L_{g,1}$	M_Φ	$M_\Phi^2\left(\sigma_{g,2}^2 + L_{g,1}^2\right)$
$\nabla_{22}^2 g_{\Phi^\epsilon}$	$\sigma_{g,2}^2$	$\Phi\Phi^\top$	$L_{g,1}$	M_Φ^2	$M_\Phi^4\left(\sigma_{g,2}^2 + L_{g,1}^2\right)$

We conclude that these 4 mappings have a bounded variance.

The case of $\nabla_2 g_{\Phi^\epsilon}$ requires extra care since it is not uniformly bounded. We have:

$$\begin{aligned}\|\nabla_2 g(x, W\Phi(\xi), \xi, \eta)\| &\leq \|\nabla_2 g(x, W\Phi(\xi), \xi, \eta) - \nabla_2 g(x, y^\star(x,\xi), \xi, \eta)\| + \|\nabla_2 g(x, y^\star(x,\xi), \xi, \eta)\| \\ &\leq L_{g,1}\|W\Phi(\xi) - y^\star(x,\xi)\| + \|\nabla_2 g(x, y^\star(x,\xi), \xi, \eta)\|\end{aligned}$$

Using the identity $(a+b)^2 \leq 2a^2 + 2b^2$, we have:

$$\|\nabla_2 g(x, W\Phi(\xi), \xi, \eta)\|^2 \leq 2L_{g,1}^2\|W\Phi(\xi) - y^\star(x,\xi)\|^2 + 2\|\nabla_2 g(x, y^\star(x,\xi), \xi, \eta)\|^2$$

We use the triangular inequality, the identity $(a+b)^2 \leq 2a^2 + 2b^2$, and $\|\Phi(\xi)\| \leq M_\Phi$ to bound:

$$\begin{aligned}\|W\Phi(\xi) - y^\star(x,\xi)\|^2 &\leq \left(\|W\Phi(\xi) - W^\star(x)\Phi(\xi)\| + \|W^\star(x)\Phi(\xi) - y^\star(x,\xi)\|\right)^2 \\ &\leq 2\|W\Phi(\xi) - W^\star(x)\Phi(\xi)\|^2 + 2\|W^\star(x)\Phi(\xi) - y^\star(x,\xi)\|^2 \\ &\leq 2M_\Phi^2\|W - W^\star(x)\|^2 + 2\|W^\star(x)\Phi(\xi) - y^\star(x,\xi)\|^2\end{aligned}$$

Finally, combining $\|W - W^\star(x)\| \leq \Delta_0$, (5), and (1) yields:

$$\begin{aligned}\mathbb{E}_\xi\|W\Phi(\xi) - y^\star(x,\xi)\|^2 &\leq 2M_\Phi^2\Delta_0^2 + 2\|W^\star(x)\Phi(\xi) - y^\star(x,\xi)\|^2 \\ &\leq 2M_\Phi^2\Delta_0^2 + \frac{4L_{g,1}}{\mu}\mathbb{E}_\xi\|W^\dagger(x)\Phi(\xi) - y^\star(x,\xi)\|^2 \\ &\leq 2M_\Phi^2\Delta_0^2 + \frac{\epsilon^2}{K^2}\end{aligned}$$

From Assumption 2.1, for any fixed ξ, $\nabla_2 g(x, y, \xi, \eta)$ is unbiased with variance bounded by $\sigma_{g,1}^2$. By definition of $y^\star(x,\xi)$ we have $\mathbb{E}_{\eta|\xi}\nabla_2 g(x, y^\star(x,\xi), \xi, \eta) = 0$ for all ξ and it follows that:

$$\begin{aligned}\mathbb{E}_{\eta|\xi}\|\nabla_2 g(x, y^\star(x,\xi), \xi, \eta)\|^2 &= \text{Var}_{\eta|\xi}\left[\nabla_2 g(x, y^\star(x,\xi), \xi, \eta)\right] \\ &\leq \sigma_{g,1}^2\end{aligned}$$

Taking the expectation over ξ yields:

$$\mathbb{E}_{(\xi,\eta)}\|\nabla_2 g(x, y^\star(x,\xi), \xi, \eta)\|^2 \leq \sigma_{g,1}^2$$

Combining the above results we have:

$$\begin{aligned}\mathbb{E}_{(\xi,\eta)}\|\nabla_2 g(x, W\Phi(\xi), \xi, \eta)\|^2 &\leq 2L_{g,1}^2\mathbb{E}_{(\xi,\eta)}\|W\Phi(\xi) - y^\star(x,\xi)\|^2 + 2\mathbb{E}_{(\xi,\eta)}\|\nabla_2 g(x, y^\star(x,\xi), \xi, \eta)\|^2 \\ &\leq 2L_{g,1}^2\left(2M_\Phi^2\Delta_0^2 + \frac{\epsilon^2}{K^2}\right) + 2\sigma_{g,1}^2\end{aligned}$$

Since $\nabla_2 g_{\Phi^\epsilon}(x, W, \xi, \eta) = \nabla_2 g(x, W\Phi(\xi), \xi, \eta)\Phi(\xi)^\top$ and $\|\Phi(\xi)\| \leq M_\Phi$ we obtain:

$$\begin{aligned}\text{Var}_{(\xi,\eta)}\left[\nabla_2 g_{\Phi^\epsilon}(x, W, \xi, \eta)\right] &\leq \mathbb{E}_{(\xi,\eta)}\|\nabla_2 g_{\Phi^\epsilon}(x, W, \xi, \eta)\|^2 \\ &\leq \mathbb{E}_{(\xi,\eta)}\left[\|\nabla_2 g(x, W\Phi(\xi), \xi, \eta)\|^2\|\Phi(\xi)\|^2\right] \\ &\leq 4L_{g,1}^2M_\Phi^4\Delta_0^2 + 2L_{g,1}^2M_\Phi^2\frac{\epsilon^2}{K^2} + 2\sigma_{g,1}^2M_\Phi^2 \\ &\leq 4M_\Phi^4\left(L_{g,1}^2\Delta_0^2 + \frac{L_{g,1}^2\epsilon^2}{K^2} + \sigma_{g,1}^2\right)\end{aligned}\tag{9}$$

which concludes the proof. □

D Proof of Theorem 2.3

In this proof, we use the notations $\lesssim$, $\simeq$, and $\gtrsim$ to denote relations up to a constant. We will show that under assumptions 2.1 and if Φ is well-conditioned, the assumptions 1 and 2 of [3] hold for f_Φ and g_Φ. Namely, we want to show:

Assumption D.1. For any $x \in \mathbb{R}^{d_x}$, $\mathbb{E}_{(\xi,\eta)}\left[g_\Phi\left(x, W, \xi, \eta\right)\right]$ is λ-strongly convex and L-smooth.

and

Assumption D.2. The following hold:

(i) $\nabla_1 f_\Phi$ is L_{fx}-Lipschitz continuous, $\nabla_2 f_\Phi$ is L_{fy}-Lipschitz continuous, $\nabla_2 g_\Phi$ is L_{gy}-Lipschitz continuous, $\nabla_{12}^2 g_\Phi$ is L_{gxy}-Lipschitz continuous, $\nabla_{22}^2 g_\Phi$ is L_{gyy}-Lipschitz continuous, all with respect to (x, W).

(ii) $\nabla_1 f_\Phi$, $\nabla_2 f_\Phi$, $\nabla_2 g_\Phi$, $\nabla_{12}^2 f_\Phi$, and $\nabla_{22}^2 f_\Phi$ have a variance bounded by σ^2.

(iii) $\|\nabla_2 \mathbb{E}_{(\xi,\eta)}\left[f(x, W, \xi, \eta)\right]\|^2 \leq C_{fy}^2$, $\|\nabla_{12}^2 \mathbb{E}_{(\xi,\eta)}\left[g(x, W, \xi, \eta)\right]\|^2 \leq C_{gxy}^2$.

Under assumption 2.1 and if Φ is well-conditioned, Lemma C.2 gives $\mathbb{E}_{(\xi,\eta)}\left[g_\Phi\left(x, W, \xi, \eta\right)\right]$ is $\mu m_\Phi(\epsilon)$-strongly convex with respect to W for any $x \in \mathbb{R}^{d_x}$. Further, we have from Lemma C.4 that $\nabla_2 g_\Phi\left(x, W, \xi, \eta\right)$ is $L_{g,1}M_\Phi^2$-Lipchitz continuous in W for all fixed (x, ξ, η). Taking the expectation over (ξ, η) we obtain:

$$\begin{aligned}&\left\|\nabla_2 \mathbb{E}_{(\xi,\eta)}\left[g_\Phi\left(x, W, \xi, \eta\right)\right] - \nabla_2 \mathbb{E}_{(\xi,\eta)}\left[g_\Phi\left(x', W', \xi, \eta\right)\right]\right\| \\ &\quad= \left\|\mathbb{E}_{(\xi,\eta)}\left[\nabla_2 g_\Phi\left(x, W, \xi, \eta\right) - \nabla_2 g_\Phi\left(x', W', \xi, \eta\right)\right]\right\| \\ &\quad\leq \mathbb{E}_{(\xi,\eta)}\left[\left\|\nabla_2 g_\Phi\left(x, W, \xi, \eta\right) - \nabla_2 g_\Phi\left(x', W', \xi, \eta\right)\right\|\right] \\ &\quad\leq L_{g,1}M_\Phi^2\left(\|x - x'\| + \|W - W'\|\right)\end{aligned}$$

where the first inequality holds since assumption 2.1 implies that $\nabla_2 g_\Phi\left(x, W, \xi, \eta\right)$ is unbiased and the first inequality uses Jensen's inequality. Therefore $\mathbb{E}_{(\xi,\eta)}\left[g_\Phi\left(x, W, \xi, \eta\right)\right]$ is $L_{g,1}M_\Phi^2$-smooth for any fixed $x \in \mathbb{R}^{d_x}$ and assumption D.1 holds with $\lambda = \mu m_\Phi(\epsilon)$ and $L = L_{g,1}M_\Phi^2$.

Using the implicit-function theorem, the accuracy of the estimate of ∇F_Φ degrades linearly with the gap $\|W_t - W^\star(x_t)\|$. Hence $\|W_t - W^\star(x_t)\|$ cannot diverge and there exists $\Delta_0 < \infty$ such that the iterates W_t satisfy $\|W_t - W^\star(x_t)\| \leq \Delta_0$. In particular, this is holds for RSVRB as per Lemma 6 of [3]. Henceforth, we restrict our analysis to $\left\{(x, W) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y \times N_\Phi(\epsilon)} :\ \|W - W^\star(x)\| \leq \Delta_0\right\}$ for the rest of the proof.

Lemma C.4 and Lemma C.7 show that assumption $D.2-(i)$ and $D.2-(ii)$ hold f_Φ and g_Φ for with Lipchitz constant L and uniform variance bound σ^2 given in the table below.

h	L	σ^2
$\nabla_1 f_{\Phi^\epsilon}$	$L_{f,1}M_\Phi$	$\sigma_f^2 + L_{f,0}^2$
$\nabla_2 f_{\Phi^\epsilon}$	$L_{f,1}M_\Phi^2$	$M_\Phi^2\left(\sigma_f^2 + L_{f,0}^2\right)$
$\nabla_2 g_{\Phi^\epsilon}$	$L_{g,1}M_\Phi^2$	$4M_\Phi^4\left(L_{g,1}^2\Delta_0^2 + \frac{\epsilon}{K^2} + \sigma_{g,1}^2\right)$
$\nabla_{12}^2 g_{\Phi^\epsilon}$	$L_{g,2}M_\Phi^2$	$M_\Phi^2\left(\sigma_{g,2}^2 + L_{g,1}^2\right)$
$\nabla_{22}^2 g_{\Phi^\epsilon}$	$L_{g,2}M_\Phi^3$	$M_\Phi^4\left(\sigma_{g,2}^2 + L_{g,1}^2\right)$

Further, Lemma C.4 also gives that f_Φ and g_ϕ are $L_{f,0}M_\Phi$ and $L_{g,0}M_\Phi$-Lipschitz continuous in (x, W), respectively. It follows that that $\|\nabla_1 \mathbb{E}_{(\xi,\eta)}\left[f_\Phi\left(x, W, \xi, \eta\right)\right]\| \leq L_{f,0}M_\Phi$ and $\|\nabla_2 \mathbb{E}_{(\xi,\eta)}\left[g_\Phi\left(x, W, \xi, \eta\right)\right]\| \leq L_{g,0}M_\Phi$. Hence condition $D.2-(iii)$ hold for f_Φ and g_Φ with $C_{fy} = L_{f,0}M_\Phi$ and $C_{gxy} = L_{g,1}M_\Phi^2$.

Therefore Theorem 1 of [3] holds. Before stating it we first bound some quantities in term of M_Φ and m_Φ.

Since Φ is well conditioned, we have $\Sigma_\Phi \succeq m_\phi I_{N_\Phi(\epsilon)}$. Therefore:

$$\begin{aligned}
\mathbb{E}_\xi\left[\|W^\star(x_0)\Phi(\xi)\|^2\right] &= \mathbb{E}_\xi\left[(W^\star(x_0)\Phi(\xi))^\top W^\star(x_0)\Phi(\xi)\right] \\
&= \mathbb{E}_\xi\left[\text{tr}\left(W^\star(x_0)\Phi(\xi)(W^\star(x_0)\Phi(\xi))^\top\right)\right] \\
&= \text{tr}\left(W^\star(x_0)\mathbb{E}_\xi\left[\Phi(\xi)\Phi(\xi)^\top\right]W^\star(x_0)^\top\right) \\
&\geq m_\phi \text{tr}\left(W^\star(x_0)W^\star(x_0)^\top\right) \\
&= m_\phi\|W^\star(x_0)\|_F^2 \\
&\geq m_\phi\|W^\star(x_0)\|^2
\end{aligned}$$

For any (x_0, W_0) and by the μ strong convexity of g we have:

$$\mathbb{E}_{(\xi,\eta)}\left[g_\Phi(x_0, W_0, \xi, \eta)\right] \geq \mathbb{E}_{(\xi,\eta)}\left[g_\Phi(x_0, W^\star(x_0), \xi, \eta)\right] + \frac{\mu}{2}\mathbb{E}_{(\xi,\eta)}\left[\|W^\star(x_0)\Phi(\xi)\|^2\right]$$

and thus, taking $W_0 = 0$:

$$\begin{aligned}
\mathbb{E}_\xi\left[\|W^\star(x_0)\Phi(\xi)\|^2\right] &\leq \frac{2}{\mu}\mathbb{E}_{(\xi,\eta)}\left[g_\Phi(x_0, W_0, \xi, \eta) - g_\Phi(x_0, W^\star(x_0), \xi, \eta)\right] \\
&\leq \frac{2}{\mu}\underbrace{\left(\mathbb{E}_{(\xi,\eta)}\left[g(x_0, 0, \xi, \eta) - \min_y g(x_0, y, \xi, \eta)\right]\right)}_{\Delta_{g,0}}
\end{aligned}$$

Combining the above results we obtain:

$$\begin{aligned}
\|W^\star(x_0)\|^2 &\leq \frac{\mathbb{E}_\xi\left[\|W^\star(x_0)\Phi(\xi)\|^2\right]}{m_\Phi} \\
&\leq \frac{2\Delta_{g,0}}{\mu \cdot m_\Phi}
\end{aligned}$$

From Lemma 2.2 in [30], $W^\star(x)$ is Lipschitz continuous in x with constant $L_W = \frac{L_{gy}}{\lambda} = \frac{L_{g,1}M_\Phi^2}{\mu m_\Phi}$. Additionally, ∇F is Lipschitz continuous in x with constant

$$\begin{aligned}
L_F &= L_{fy} + \frac{L_{fy}L_{gy}}{\lambda} + \frac{L_{gy}}{\lambda}\left(L_{fy} + \frac{L_{fy}L_{gy}}{\lambda} + L_f\left[\frac{L_{gxy}}{\lambda} + \frac{L_{gyy}L_{gy}}{\lambda^2}\right]\right) + L_f\left[\frac{L_{gxy}L_f}{\lambda} + \frac{L_{gyy}L_{gy}}{\lambda^2}\right] \\
&= L_{fy} + \frac{2L_{fy}L_{gy} + L_f^2L_{gxy}}{\lambda} + \frac{L_{fy}L_{gy}^2 + L_fL_{gy}L_{gxy} + L_fL_{gyy}L_{gy}}{\lambda^2} + \frac{L_fL_{gy}^2L_{gyy}}{\lambda^3} \\
&\lesssim M_\Phi^2 + \frac{M_\Phi^5}{m_\Phi} + \frac{M_\Phi^6}{m_\Phi^2} + \frac{M_\Phi^8}{m_\Phi^3} \\
&\lesssim \frac{M_\Phi^8}{m_\Phi^3}
\end{aligned}$$

Using the notations of [3], we have $\delta_{W,0} \triangleq \|W_1 - W^\star(x_0)\|^2$ with $W_1 = W_0 - \tau_0\tau w_1$ and:

$$\begin{aligned}
\mathbb{E}\left[\delta_{fx,0}\right] &\lesssim \|\nabla_1 f_{\Phi^\epsilon}(x_0, W_0))\|^2 \lesssim M_\Phi^2 \\
\mathbb{E}\left[\delta_{fy,0}\right] &\lesssim \|\nabla_2 f_{\Phi^\epsilon}(x_0, W_0))\|^2 \lesssim M_\Phi^4 \\
\mathbb{E}\left[\delta_{gxy,0}\right] &\lesssim \|\nabla_{12} g_{\Phi^\epsilon}(x_0, W_0))\|^2 \lesssim M_\Phi^4 \\
\mathbb{E}\left[\delta_{gyy,0}\right] &\lesssim \|\nabla_{22} g_{\Phi^\epsilon}(x_0, W_0))\|^2 \lesssim M_\Phi^6 \\
\mathbb{E}\left[\delta_{gy,0}\right] &\lesssim \mathbb{E}\|\nabla_2 g_{\Phi^\epsilon}(x_0, W_0))\|^2 \lesssim M_\Phi^4
\end{aligned}$$

where the first 4 inequalities follows from the Lipschitz constants given in Lemma C.4, and the last holds given (9) .

Theorem 1 of [3] finally give:

$$\frac{1}{2(T+1)}\mathbb{E}\left[\sum_{t=0}^{T}\|\nabla F_{\Phi^\epsilon}(x_t)\|^2\right] \leq \frac{F_\Phi(x_0)-F_\Phi(x^\star)}{\gamma\eta_T T} + \frac{C\mathbb{E}\left[\delta_{W,0}\right]}{\eta_T T} \\ + \frac{\mathbb{E}\left[\delta_{gy,0}+\delta_{fx,0}+\delta_{fy,0}+\delta_{gxy,0}+\delta_{gyy,0}\right]}{\gamma\eta_0\eta_T T} + \frac{\mathcal{O}\left(\ln(T+2)\right)}{\gamma\eta_T T} \tag{10}$$

where for $c=1$:

$$\begin{aligned}
C_0 &= \left(2L_{fx}^2 + \frac{6C_{fy}^2 L_{gxy}^2}{\lambda^2} + \frac{6C_{fy}^2 C_{gxy}^2 L_{gyy}^2}{\lambda^4} + \frac{6L_{fy}^2 C_{gxy}^2}{\lambda^2}\right) \lesssim \frac{M_\Phi^{12}}{m_\Phi^4} \\
C_1 &= 2 \\
C_2 &= \frac{6C_{fy}^2}{\lambda^2} \simeq \frac{M_\Phi^2}{m_\Phi^2} \\
C_3 &= \frac{6C_{fy}^2 C_{gxy}^2}{\lambda^4} \simeq \frac{M_\Phi^6}{m_\Phi^4} \\
C_4 &= \frac{6C_{gxy}^2}{\lambda^2} \simeq \frac{M_\Phi^4}{m_\Phi^2} \\
\tau &= \frac{1}{3L_g} \simeq \frac{1}{M_\Phi^2} \\
C &= \max\left\{\frac{4C_0}{\tau\lambda}, \frac{4\left(L_{gy}^2+L_{fx}^2+L_{fy}^2+L_{gxy}^2+L_{gyy}^2\right)}{\gamma}\right\} \lesssim \frac{M_\Phi^{18}}{m_\Phi^5} \\
\gamma &= \min\left\{\frac{\sqrt{\tau\lambda}}{8\sqrt{C}L_W}, \frac{1}{16\left(L_{gy}^2+L_{fx}^2+L_{fy}^2+L_{gxy}^2+L_{gyy}^2\right)}\right\} \gtrsim \frac{m_\Phi^4}{M_\Phi^{12}} \\
c_0 &= \max\left\{2, 64L_F^3, \left(\frac{\lambda}{16C\gamma\tau}\right)^{3/2}, \left(\frac{2}{7L_F}\right)^{3/2}, \left(2(C_1+C_2+C_3+C_4)\gamma\right)^{3/2}\right\} \simeq \frac{M_\Phi^{24}}{m_\Phi^9} \\
\eta_t &= \tau_t = \frac{1}{(c_0+t)^{1/3}}
\end{aligned}$$

Here the bounds on C and γ are obtain after considering all 4 possible cases. We can also bound $\|w_1\| \leq 2L_{g,0}M_\Phi$, thus $\|W_1\| \lesssim \frac{m_\Phi^3}{M_\Phi^7}$, and $\delta_{W,0} \leq \left(\|W_1\| + \|W^\star(x_0)\|\right)^2 \lesssim \frac{1}{m_\Phi}$.

Substituting in (10) we finally obtain:

$$\begin{aligned}
\frac{1}{2(T+1)}\mathbb{E}\left[\sum_{t=0}^{T}\|\nabla F_{\Phi^\epsilon}(x_t)\|^2\right] &= \frac{1}{\eta_T T}\Bigg[\frac{F_\Phi(x_0)-F_\Phi(x^\star)}{\gamma} + C\mathbb{E}\left[\delta_{W,0}\right] \\
&\quad + \frac{\mathbb{E}\left[\delta_{gy,0}+\delta_{fx,0}+\delta_{fy,0}+\delta_{gxy,0}+\delta_{gyy,0}\right]}{\gamma\eta_0} + \frac{\mathcal{O}\left(\ln(T+2)\right)}{\gamma}\Bigg] \\
&\lesssim \frac{1}{\eta_T T}\left[\frac{M_\Phi^{12}}{m_\Phi^4} + \frac{M_\Phi^{18}}{m_\Phi^5}\frac{1}{m_\Phi} + \frac{M_\Phi^{26}}{m_\Phi^7} + \frac{M_\Phi^{12}}{m_\Phi^4}\mathcal{O}\left(\ln(T+2)\right)\right]
\end{aligned}$$

In particular, following the analysis in [31], the sample complexity is $\widetilde{\mathcal{O}}\left(\frac{1}{\epsilon^3}\frac{\text{Poly}(M_\Phi(\epsilon))}{\text{Poly}(m_\Phi(\epsilon))}\right)$.

E Chebyshev Series: Uniform Convergence and Conditioning

Lemma E.1.
Under assumptions 2.1-(ii) and condition (c.3), $y^\star(x,\xi)$ is real-analytic over Ξ for any fixed $x \in \mathbb{R}^{d_x}$.

Proof. Recall that $G(x, y, \xi) \triangleq \mathbb{E}_{\eta \sim \mathbb{P}_{\eta|\xi}} [g(x, y, \xi, \eta)]$. Under assumption 2.1-(ii), g is μ-strongly convex in y for any fixed (x, ξ, η). Thus, G is μ-strongly convex in y for any fixed (x, ξ) and $y^\star(x, \xi)$ is the unique solution to $\nabla_2 G(x, y' \xi) = 0$. Additionally, since G is real-analytic in (y, ξ) for all $x \in \mathbb{R}^{d_x}$, so is $\nabla_2 G$. Define $H(x, y, \xi) = \nabla_2 G(x, y, \xi)$. Then $H(x, y^\star(x, \xi), \xi) = 0$ for all $x \in \mathbb{R}^{d_x}$. Further, the strong convexity of G with respect to y gives that $\nabla_2 H = \nabla_{22}^2 G$ is invertible. We can then use the analytic implicit function theorem to obtain that, for any fixed $x \in \mathbb{R}^{d_x}$, there exists a unique function $y^\dagger(x, \xi)$ real-analytic over Ξ and solution to $H(x, y, \xi) = 0$ for all $\xi \in \Xi$. By unicity, we must have $y^\star = y^\dagger$ and it follows that $y^\star$ is real-analytic over Ξ for any fixed $x \in \mathbb{R}^{d_x}$. □

Lemma E.2.
Let f be an analytic function in $[-1, 1]^m$ that is analytically continuable to the open region E_ρ delimited by the Bernstein ellipse with parameter $\rho \in (1, e^{1/2}]$, where it satisfies $|f(x)| \leq M$ for all $x \in \mathcal{R}(E_\rho)$. Then for each $k \geq 0$ its Chebyshev coefficients satisfy

$$|a_k| \leq 2M\rho^{-k}.$$

Proof. Let $F : \left\{ \begin{array}{ccc} E_\rho & \to & \mathbb{R} \\ z & \mapsto & f\left(\frac{z+z^{-1}}{2}\right) \end{array} \right.$. Since f is analytic in $\mathbb{E}_\rho$, F is also analytic in E_ρ as the composition of the two analytic functions $z \mapsto (z + z^{-1})/2$ and $x \mapsto f(x)$. From Theorem 3.1 of [18], the Chebyshev coefficients are given by

$$a_0 = \frac{2}{\pi i} \int_{|z|=1} z^{-1} F(z) dz$$

and

$$a_k = \frac{1}{\pi i} \int_{|z|=1} z^{-(1+k)} F(z) dz, \quad \forall k \geq 1.$$

If F is analytic in the closure of E_ρ, we can expand the contour to $|z| = \rho$ without changing the value of these integrals. Since $|F(z)| \leq M$ for all $z \in E_\rho$ and $\rho \in (1, e^{1/2}]$ we obtain for $k = 0$:

$$\begin{aligned} |a_0| &= \frac{2}{\pi} \left| \int_{|z|=\rho} z^{-(} F(z) dz \right| \\ &\leq \frac{M}{\pi} \int_{|z|=\rho} |z|^{-1} dz \\ &= 4M \ln(\rho) \\ &\leq 2M \end{aligned}$$

and similarly for all $k \geq 1$:

$$\begin{aligned} |a_k| &= \frac{1}{\pi} \left| \int_{|z|=\rho} z^{-(1+k)} F(z) dz \right| \\ &\leq \frac{M}{\pi} \int_{|z|=\rho} |z|^{-(1+k)} dz \\ &= 2M\rho^{-k} \end{aligned}$$

Otherwise, we can expand the contour to $|z| = s$ for any $s < \rho$, giving the same bound for all $s < \rho$ and thus also for $s = \rho$.
Therefore $|a_k| \leq 2M\rho^{-k}$ holds for any $k \geq 0$. □

Lemma E.3.
Let f be an analytic function in $[-1, 1]^m$ that is analytically continuable to the open region E_ρ^m delimited by m-dimensional Bernstein space $\prod_{i=1}^m E_\rho$ with $\rho \in (1, e^{1/2}]$ and $|f(x)| \leq M$ for all $x \in \mathcal{R}\left(E_\rho^m\right)$.
Then the coefficients of the m-dimensional Chebychev expansion:

$$f(x) = \sum_{k_1=0}^{\infty} \cdots \sum_{k_m=0}^{\infty} a_{k_1,\ldots,k_m} \prod_{i=1}^{m} T_{k_i}(x_i)$$

are such that, for any $k_1, \dots, k_m \geq 0$*,*

$$|a_{k_1,\dots,k_m}| \leq M \prod_{i=1}^{m} \left(2\rho^{-k_i}\right).$$

Proof. Let $R_\rho^l \triangleq \mathcal{R}\left(E_\rho^l\right)$ be the projection of E_ρ^l onto $\mathbb{R}^l$. We begin by defining two classes of functions:

$$H^{(l)}(M) = \left\{ f : [-1,1]^l \to \mathbb{R} \mid f \text{ analytically continuable to } E_\rho^l, \quad \sup_{x \in R_\rho^l} |f(x)| \leq M \right\}$$

$$P^{(l)}(M) = \left\{ f : [-1,1]^l \to \mathbb{R} \mid \text{the Chebychev coefficients of } f \text{ satisfy } |a_{k_1,\dots,k_l}| \leq M \prod_{i=1}^{l} \left(2\rho^{-k_i}\right) \right\}$$

Let the hypothesis of induction be that the statement holds for a dimension $l - 1 \leq m$. Namely, $H^{(l-1)}(M) \subseteq P^{(l-1)}(M)$. We want to show that $H^{(l)}(M) \subseteq P^{(l)}(M)$.

Let $f^{(l)} \in H^{(l)}(M)$. For any fixed $(x_1, \dots, x_{l-1}) \in [-1,1]^{l-1}$, define the single-variable function:

$$f_l^{(l)}(x_1, \dots, x_{l-1})[x_l] \triangleq f^{(l)}(x_1, \dots, x_{l-1}, x_l), \quad \forall x_l \in [-1,1].$$

Note that since $f^{(l)} \in H^{(l)}$ we have in particular for any fixed $(x_1, \dots, x_{l-1}) \in R_\rho^{l-1}$ that $\sup_{x_l \in R_\rho} \left| f_l^{(l)}(x_1, \dots, x_{l-1})[x_l] \right| \leq \sup_{x \in R_\rho^l} \left| f^{(l)}(x) \right| \leq M$. Since $f^{(l)}$ is jointly real-analytic in all its variables and can be continued analytically to E_ρ^l, the mapping $x_l \mapsto f_l^{(l)}(z_1, \dots, z_{l-1})[x_l]$ is also real-analytic and can be continued analytically to E_ρ for any $(z_1, \dots, z_{l-1}) \in E_\rho^{l-1}$.

Applying Lemma E.2, we have that the coefficients of the Chebyshev expansion:

$$f_l^{(l)}(x_1, \dots, x_{l-1})[x_l] = \sum_{k_l=1}^{\infty} c_{k_l}(x_1, \dots, x_{l-1}) T_{k_l}(x_l).$$

satisfy

$$|c_{k_l}(x_1, \dots, x_{l-1})| \leq 2M\rho^{-k_l}, \quad \forall k_l \geq 0.$$

Since this hold for any fixed $(x_1, \dots, x_{l-1}) \in R_\rho^{l-1}$, we have $\sup_{x \in R_\rho^{l-1}} |c_{k_l}(x)| \leq 2M\rho^{-k_l}$.

Further, because $f^{(l)}$ is jointly analytic in all its variables and can be continued analytically to E_ρ^l, for any fixed x_l, the mapping $(x_1, \dots, x_{l-1}) \mapsto f^{(l)}(x_1, \dots, x_{l-1}, x_l)$ is also analytic and can be continued analytically to E_ρ^{l-1}. By definition,

$$\begin{aligned} c_{k_l}(x_1, \dots, x_{l-1}) &\triangleq \frac{2}{\pi} \int_0^\pi f_l^{(l)}(x_1, \dots, x_{l-1})[\cos(\theta)] T_{k_l}(\cos\theta) d\theta \\ &= \frac{2}{\pi} \int_0^\pi f^{(l)}(x_1, \dots, x_{l-1}, \cos(\theta)) T_{k_l}(\cos\theta) d\theta \end{aligned}$$

so c_{k_l} is analytic on $[-1,1]^{l-1}$ and can be continued analytically on E_ρ^{l-1} as integration preserves analyticity. Therefore, $c_{k_l} \in H^{(l-1)}(2M\rho^{-k_l})$. By the induction hypothesis, it follows that $c_{k_l} \in P^{(l-1)}(2M\rho^{-k_l})$. Thus, each c_{k_l} can be expanded as:

$$c_{k_l}(x_1, \dots, x_{l-1}) = \sum_{k_1=1}^{\infty} \cdots \sum_{k_{l-1}=1}^{\infty} a_{k_1,\dots,k_{l-1},k_l} \prod_{i=1}^{l-1} T_{k_i}(x_i)$$

with

$$\begin{aligned} |a_{k_1,\dots,k_{l-1},k_l}| &\leq \left(2M\rho^{-k_l}\right) \prod_{i=1}^{l-1} \left(2\rho^{-k_i}\right) \\ &= M \prod_{i=1}^{l} \left(2\rho^{-k_i}\right) \end{aligned}$$

It follows that for any $(x_1, \dots, x_l) \in [-1, 1]^l$,

$$\begin{aligned} f^{(l)}(x_1, \dots, x_l) &= f_l^{(l)}(x_1, \dots, x_{l-1})[x_l] \\ &= \sum_{k_l=0}^{\infty} c_{k_l}(x_1, \dots, x_{l-1}) T_{k_l}(x_l) \\ &= \sum_{k_l=0}^{\infty} \left[\sum_{k_1=0}^{\infty} \cdots \sum_{k_{l-1}=0}^{\infty} a_{k_1,\dots,k_{l-1},k_l} \prod_{i=1}^{l-1} T_{k_i}(x_i) \right] T_{k_l}(x_l) \\ &= \sum_{k_1=0}^{\infty} \cdots \sum_{k_l=0}^{\infty} a_{k_1,\dots,k_{l-1},k_l} \prod_{i=1}^{l} T_{k_i}(x_i) \end{aligned}$$

with $|a_{k_1,\dots,k_{l-1},k_l}| \leq M \prod_{i=1}^{l} \left(2\rho^{-k_i}\right)$. This shows that $f^{(l)} \in P^{(l)}(M)$, completing the induction step.

The initialization ($l = 1$) of the induction reduces to Lemma E.2. We conclude by induction that:

$$H^{(m)}(M) \subseteq P^{(m)}(M)$$

□

Lemma E.4.
Let $f = \lim_{n\to\infty} f_n$ *where*

$$f_n : \left\{ \begin{array}{ccc} [-1,1]^m & \to & \mathbb{R} \\ x & \mapsto & \sum_{k_1=0}^{n} \cdots \sum_{k_m=0}^{n} a_{k_1,\dots,k_m} \prod_{i=1}^{m} T_{k_i}(x_i) \end{array} \right. , \quad \forall n \geq 1$$

and a *satisfies for* $\rho > 1$*:*

$$|a_{k_1,\dots,k_m}| \leq M \prod_{i=1}^{m} \left(2\rho^{-k_i}\right) \quad \forall k_1, \dots, k_m \geq 0.$$

Then the residual $r_n = \sup_{x\in[-1,1]^m} |f(x) - f_n(x)|$ *is bounded by*

$$r_n \leq M \left(\frac{2}{\rho - 1}\right)^m \left[1 - (1 - (1/\rho)^n)^m\right].$$

Proof. By definition, the remainder after truncation is

$$f(x) - f_n(x) = \sum_{\substack{k_1,\dots,k_m \geq 0 \\ \exists j \text{ s.t. } k_j > n}} a_{k_1,\dots,k_m} \prod_{i=1}^{m} T_{k_i}(x_i).$$

Since $|T_{k_i}(x_i)| \leq 1$ for all $x_i \in [-1, 1]$, we have

$$\begin{aligned} r_n &\leq \sup_{x\in[-1,1]^m} \sum_{\substack{k_1,\dots,k_m \geq 0 \\ \exists j \text{ s.t. } k_j > n}} |a_{k_1,\dots,k_m}| \prod_{i=1}^{m} |T_{k_i}(x_i)| \\ &\leq \sum_{\substack{k_1,\dots,k_m \geq 0 \\ \exists j \text{ s.t. } k_j > n}} |a_{k_1,\dots,k_m}|. \end{aligned}$$

Since $|a_{k_1,\dots,k_m}| \leq M \prod_{i=1}^{m}(2\rho^{-k_i})$, it follows that

$$\begin{aligned}
r_n &\leq M \sum_{\substack{k_1,\dots,k_m \geq 0 \\ \exists j \text{ s.t. } k_j > n}} \prod_{i=1}^{m}(2\rho^{-k_i}) \\
&= M \sum_{k_1=0}^{\infty} \cdots \sum_{k_m=0}^{\infty} \prod_{i=1}^{m}(2\rho^{-k_i}) - M \sum_{k_1=0}^{n} \cdots \sum_{k_m=0}^{n} \prod_{i=1}^{m}(2\rho^{-k_i}) \\
&= M \left(\sum_{k=0}^{\infty} 2\rho^{-k}\right)^{m} - M \left(\sum_{k=0}^{n} 2\rho^{-k}\right)^{m}
\end{aligned}$$

Since $\rho > 1$ we have:

$$\sum_{k=0}^{\infty} 2\rho^{-k} = \frac{2}{\rho - 1}$$

for the full sum, and

$$\sum_{k=0}^{n} 2\rho^{-k} = 2 \cdot \frac{1 - (1/\rho)^n}{\rho - 1}$$

for the truncated sum.

Substituting this back into our bound, we get:

$$\begin{aligned}
r_n &\leq M \left[\left(\frac{2}{\rho - 1}\right)^{m} - \left(\frac{2(1 - (1/\rho)^n)}{\rho - 1}\right)^{m} \right] \\
&= M \left(\frac{2}{\rho - 1}\right)^{m} \left[1 - (1 - (1/\rho)^n)^m\right]
\end{aligned}$$

which gives the desired result. □

Proposition E.5. *Let Φ be the basis of d_ξ-dimensional Chebyshev polynomials and*

$$\underline{N}(\tilde{\epsilon}) = \mathcal{O}\left(\ln^{d_\xi}\left(\tilde{\epsilon}^{-1}\right)\right).$$

If assumptions 2.1 and the conditions of Theorem 2.4 hold, then N_Φ is expressive with $N_\Phi(\epsilon) = \underline{N}\left(\frac{\epsilon}{2K}\sqrt{\frac{\mu}{L_{g,1}}}\right)$.

Proof. We give a proof for $\Xi = [-1, 1]^{d_\xi}$. This is without loss of generality, as discussed in the main text. Under assumptions 2.1-(ii) and condition (c.3), we have from Lemma E.1 that $y^\star(x, \cdot)$ is real-analytic over Ξ for any fixed $x \in \mathbb{R}^{d_x}$. Hence there exists $\rho \in (1, e^{1/2})$ such that $y^\star(x, \cdot)$ is analytically continuable to the closure of E_ρ, the open region delimited by the d_ξ-dimensional Bernstein space $\prod_{i=1}^{m} E_\rho$. Since $y^\star(x, \cdot)$ is analytic on the compact set $\overline{E_\rho}$, it is in particular continuous and $y^\star(x, \cdot)$ is bounded on E_ρ by some constant M. For any $j \in [d_y]$ Using Lemma E.3, $y_j^\star(x, \cdot)$ admits the d_ξ-dimensional Chebyshev expansion:

$$y_j^\star(x, \xi) = \sum_{k_1=0}^{\infty} \cdots \sum_{k_{d_\xi}=0}^{\infty} a^{(j)}_{k_1,\dots,k_{d_\xi}} \prod_{i=1}^{d_\xi} T_{k_i}(\xi_i)$$

where for any $k_1, \dots, k_m \geq 0$,

$$|a^{(j)}_{k_1,\dots,k_{d_\xi}}| \leq M \prod_{i=1}^{d_\xi} \left(2\rho^{-k_i}\right).$$

Let $W^\dagger$ be such that its j-th row contains the elements of $a^{(j)}$. Then

$$\begin{aligned} y_{\Phi,j}(W^\dagger(x),\xi) &= W_j^\dagger(x)\Phi(\xi) \\ &= \sum_{k_1=0}^{n} \cdots \sum_{k_{d_\xi}=0}^{n} a_{k_1,\ldots,k_{d_\xi}} \prod_{i=1}^{d_\xi} T_{k_i}(\xi_i) \end{aligned}$$

Lemma E.4 then yields for any $n \geq 1$:

$$\sup_{\xi\in\Xi} \left| y_{\Phi,j}(W^\dagger(x),\xi) - y_j^\star(x,\xi) \right| \leq M \left(\frac{2}{\rho-1}\right)^{d_\xi} \left[1 - (1-(1/\rho)^n)^{d_\xi}\right] \tag{11}$$

Define:

$$\underline{\mathrm{n}}(\tilde{\epsilon}) \triangleq \left\lceil -\ln\left(1 - \left(1 - \frac{\tilde{\epsilon}}{M\sqrt{d_y}}\left(\frac{\rho-1}{2}\right)^{d_\xi}\right)^{1/d_\xi}\right) / \ln(\rho) \right\rceil \quad \text{and} \quad \underline{\mathrm{N}}(\tilde{\epsilon}) \triangleq \underline{\mathrm{n}}(\tilde{\epsilon})^{d_\xi} \tag{12}$$

Note that $\underline{\mathrm{N}}(\tilde{\epsilon}) = \Theta\left(\ln^{d_\xi}(1/\tilde{\epsilon})\right)$ as $\tilde{\epsilon} \to 0$. Furthermore, for any number of basis functions $N \geq \underline{\mathrm{N}}(\tilde{\epsilon})$, we have at least $n = N^{1/d_\xi} \geq \underline{\mathrm{n}}(\tilde{\epsilon})$ elements per dimension. As the right hand side of (11) decreases in n, we have for any $n \geq \underline{\mathrm{n}}(\tilde{\epsilon})$:

$$\sup_{\xi\in\Xi} \left| y_{\Phi,j}(W^\dagger(x),\xi) - y_j^\star(x,\xi) \right| \leq \frac{\tilde{\epsilon}}{\sqrt{d_y}}, \quad \forall j \in [d_y]$$

Since this holds for any $j \in [d_y]$, we obtain:

$$\sup_{\xi\in\Xi} \left\| y_\Phi(W^\dagger(x),\xi) - y^\star(x,\xi) \right\|^2 \leq \tilde{\epsilon}^2.$$

□

Lemma E.6. *Let $A^{(N)} \in \mathbb{R}^{N\times N}$ be a zero-indexed matrix containing the unweighted scalar product of d-dimensional Chebyshev polynomials. Then $\lambda_{\mathit{min}}\left(A^{(N)}\right) = \Omega\left(N^{-1}\right)$.*

Proof. We first consider the 1-dimensional case and define $B \in \mathbb{R}^{n\times n}$ satisfying:

$$B_{i,j} = \frac{1}{2}\int_{-1}^{1} T_i(x)T_j(x)dx$$

For any zero-indexed vector $v \in \mathbb{R}^n$, we have:

$$\begin{aligned} v^\top Bv &= \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} \frac{1}{2} v_i v_j \int_{-1}^{1} T_i(x)T_j(x)dx \\ &= \frac{1}{2}\int_{-1}^{1}\left(\sum_{i=0}^{n-1} v_i T_i(x)\right)^2 dx \\ &= \frac{1}{2}\int_{-1}^{1}\left(\sum_{i=0}^{n-1} v_i \cos(i\arccos(x))\right)^2 dx \end{aligned}$$

After substituting $x = \cos(\theta)$ and $dx = -sin(\theta)d\theta$ we obtain for $\delta = \frac{1}{4n}$:

$$\begin{aligned} v^\top Bv &= \frac{1}{2}\int_{0}^{\pi}\left(\sum_{i=0}^{n-1} v_i \cos(i\theta)\right)^2 \sin(\theta)d\theta \\ &\geq \frac{1}{2}\int_{\delta}^{\pi-\delta}\left(\sum_{i=0}^{n-1} v_i \cos(i\theta)\right)^2 \sin(\theta)d\theta \\ &\geq \frac{\sin(\delta)}{2}\int_{\delta}^{\pi-\delta}\left(\sum_{i=0}^{n-1} v_i \cos(i\theta)\right)^2 d\theta \end{aligned}$$

On one hand, we know from Fourier theory that

$$\int_0^\pi \cos(i\theta)\cos(j\theta) = \begin{cases} 0 & \text{if } i \neq j \\ \pi & \text{if } i = j = 0 \\ \frac{\pi}{2} & \text{otherwise.} \end{cases}$$

and thus

$$\begin{aligned} \int_0^\pi \left(\sum_{i=0}^{n-1} v_i \cos(i\theta) \right)^2 d\theta &= \pi v_0^2 + \frac{\pi}{2} \sum_{i=1}^{n-1} v_i^2 \\ &\geq \frac{\pi}{2} \|v\|^2 \end{aligned}$$

On the other hand, we have using Cauchy-Schwartz inequality:

$$\begin{aligned} \int_0^\pi \left(\sum_{i=0}^{n-1} v_i \cos(i\theta) \right)^2 d\theta &= \int_\delta^{\pi-\delta} \left(\sum_{i=0}^{n-1} v_i \cos(i\theta) \right)^2 d\theta + \int_{[0,\delta]\cup[\pi-\delta,\pi]} \left(\sum_{i=0}^{n-1} v_i \cos(i\theta) \right)^2 d\theta \\ &\leq \int_\delta^{\pi-\delta} \left(\sum_{i=0}^{n-1} v_i \cos(i\theta) \right)^2 d\theta + \|v\|^2 \int_{[0,\delta]\cup[\pi-\delta,\pi]} \left(\sum_{i=0}^{n-1} \cos^2(i\theta) \right) d\theta \\ &\leq \int_\delta^{\pi-\delta} \left(\sum_{i=0}^{n-1} v_i \cos(i\theta) \right)^2 d\theta + 2\delta \|v\|^2 n \end{aligned}$$

Hence we obtain:

$$\int_\delta^{\pi-\delta} \left(\sum_{i=0}^{n} v_i \cos(i\theta) \right)^2 d\theta \geq \left(\frac{\pi}{2} - 2\delta n \right) \|v\|^2$$

and we conclude using $\sin(\delta) \geq \frac{\delta}{2}$ for any $\delta \in (0,1)$ that:

$$\begin{aligned} v^\top B v &\geq \frac{\sin(\delta)}{2} \int_\delta^{\pi-\delta} \left(\sum_{i=0}^{n} v_i \cos(i\theta) \right)^2 d\theta \\ &\geq \frac{\delta}{4} \left(\frac{\pi}{2} - 2\delta n \right) \|v\|^2 \\ &= \frac{1}{16n} \left(\frac{\pi}{2} - \frac{1}{2} \right) \|v\|^2 \end{aligned}$$

Therefore we obtain for all $n \in \mathbb{N}^*$ that $\lambda_{\text{min}}(B) \geq \frac{c_B}{n} I_n$ with $c_B = \frac{\pi-1}{32}$.

Suppose first that $N = n^d$ for some $n \in \mathbb{N}^*$. We can decompose $A^{(N)} = \bigotimes_{i=1}^{d} B$. Using Lemma C.1 we obtain:

$$\begin{aligned} A^{(N)} &\succeq \left(\frac{c_B}{n} \right)^d I_N \\ &= \frac{c_B^d}{N} I_N \end{aligned}$$

Consider now an arbitrary $N \in \mathbb{N}^*$, and $n \in \mathbb{N}^*$ such that $N \in \left[(n-1)^d + 1, n^d \right]$. Since $A^{(N)}$ is a principal submatrix of $A^{(n^d)}$, the Eigenvalue Interlacing Theorem gives:

$$\begin{aligned} \lambda_{\text{min}} \left(A^{(N)} \right) &\geq \lambda_{\text{min}} \left(A^{(n^d)} \right) \\ &\geq \frac{c_B^d}{n^d} \\ &= \frac{c_B^d}{\lceil N^{1/d} \rceil^d} \end{aligned}$$

where the second inequality uses the result of the case $N = n^d$, and the last equality hold since $N \in \left[(n-1)^d + 1, n^d\right]$.

We conclude that $\lambda_{\text{min}}\left(A^{(N)}\right) = \Omega\left(N^{-1}\right)$.

□

Proposition E.7. *Let Φ be the basis of d_ξ-dimensional Chebyshev polynomials. If condition (c.1) holds, then $m_\Phi(\epsilon) = \Omega\left(\frac{1}{N_\Phi(\epsilon)}\right)$.*

Proof. Under condition (c.1), we have that Ξ is finite or there exists $\underline{c} > 0$ such that the density of $\mathbb{P}_\xi$ is lower bounded by $\underline{c}$ on Ξ.

Suppose first that $|\Xi|$ is finite. Then for any $N \leq |\Xi|$ we have that $\{\Phi(\xi)\}_{\xi\in\Xi}$ spans $\mathbb{R}^N$. Thus there exists $c_N > 0$ such that $\Sigma_\Phi \succeq c_N I_N$. Since $n-1$ polynomials of distinct order can interpolate n point, we have $N_\Phi(\epsilon) \leq |\Xi|$. Therefore, for any $N \in \mathbb{N}^*$ we have $\Sigma_\Phi \succeq \min_{n=1}^{|\Xi|} c_n I_N$ and in particular $m_\Phi(\epsilon) \geq c \geq \frac{c}{N_\Phi(\epsilon)}$ with $c = \min_{n=1}^{|\Xi|} c_n$.

Suppose now that the density of $\mathbb{P}_\xi$ is lower bounded by $\underline{c}$ on Ξ. Then:

$$\begin{aligned}\Sigma_\Phi &= \mathbb{E}_\xi\left[\Phi(\xi)\Phi(\xi)^\top\right] \\ &= \underline{c}\int_{-1}^{1} \Phi(\xi)\Phi(\xi)^\top d\xi + \int_{-1}^{1} \Phi(\xi)\Phi(\xi)^\top (d\mathbb{P}(\xi) - \underline{c}d_\xi)\end{aligned}$$

From Lemma E.6 we have $\lambda_{\text{min}}\left(\int_{-1}^{1} \Phi(\xi)\Phi(\xi)^\top d\xi\right) = \Omega\left(1/N_\Phi(\epsilon)\right)$. Additionally, $(d\mathbb{P}(\xi) - \underline{c}d_\xi)\Phi(\xi)\Phi(\xi)^\top \succeq 0$ for any $\xi \in \Xi$ as the product of a positive term and a rank 1 matrix. Therefore we have:

$$\lambda_{\text{min}}\left(\Sigma_\Phi\right) = \Omega\left(1/N_\Phi(\epsilon)\right)$$

and we conclude that $m_\Phi(\epsilon) = \Omega(1/N_\Phi(\epsilon))$. □

Combining the above results, we obtain Theorem 2.4.

Proof. Suppose that assumptions 2.1 and the conditions of Theorem 2.4 hold. The first statement of the theorem follows from Proposition E.5. Indeed we have that Φ is expressive with $N_\Phi(\epsilon) = \underline{N}\left(\frac{\epsilon}{2K}\sqrt{\frac{\mu}{L_{g,0}}}\right)$ and since $\underline{N}(\tilde{\epsilon}) = O\left(\ln^{d_\xi}(\tilde{\epsilon}^{-1})\right)$, we obtain $N_\Phi(\epsilon) = O\left(\ln^{d_\xi}(\epsilon^{-1})\right)$. Since Φ encodes multivariate Chebyshev polynomials, we have $|\Phi_i| \leq 1$ for any $i \in [N_\Phi(\epsilon)]$ and thus $M_\Phi(\epsilon) \leq \sqrt{N_\Phi(\epsilon)} = O\left(\ln^{d_\xi/2}(\epsilon^{-1})\right)$. The second statement directly follows from Proposition E.7, where substituting $N_\Phi(\epsilon) = O\left(\ln^{d_\xi}(\epsilon^{-1})\right)$ gives into $m_\Phi(\epsilon) = \Omega\left(\frac{1}{N_\Phi(\epsilon)}\right)$ gives $m_\Phi(\epsilon) = \Omega\left(\ln^{-d_\xi}(\epsilon^{-1})\right)$. □