FULLY FIRST-ORDER METHODS FOR CONTEXTUAL STOCHASTIC BILEVEL OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Contextual stochastic bilevel optimization (CSBO) is a new paradigm for decision making under uncertainty. It generalizes stochastic bilevel optimization (SBO) by integrating contextual information in the lower-level optimization problem and thus offers a stronger modeling capability. Nevertheless, owing to its semi-infinite nature, CSBO is extremely challenging from a computational perspective, hindering its real-world applications. Indeed, many algorithms designed for SBO are not applicable to CSBO. In this paper, we propose a double-loop fully first-order algorithm for solving CSBO and prove that both expected sample and gradient complexities of the algorithm are $\widetilde{\mathcal{O}}(\epsilon^{-6})$. To tackle the increasing number of inner-loop iterations, we further develop an accelerated version of our algorithm using the random truncated multilevel Monte Carlo technique and an adaptive stepsize for empirical stabilization. The accelerated algorithm enjoys the improved expected complexities of $\widetilde{\mathcal{O}}(\epsilon^{-4})$. These sample and gradient complexity orders match those of the corresponding Hessian-based algorithms for CSBO, while our algorithms use only first-order derivatives of g. Numerical experiments on meta-learning and a contextual electricity pricing problem demonstrate the superiority of the proposed algorithms over existing Hessian-based and reduction-based algorithms in terms of both speed and solution quality.

1 INTRODUCTION

Contextual stochastic bilevel optimization (CSBO) is a new bilevel optimization framework that accommodates contextual information or personalization in the lower level and takes the form

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^{d_x}} & F(x) \coloneqq \mathbb{E}_{\eta,\xi}[f(x, y^\star(x;\xi);\eta,\xi)] \\ \text{where} & y^\star(x;\xi) \coloneqq \arg\min\limits_{y\in\mathbb{R}^{d_y}} \mathbb{E}_{\eta|\xi}[g(x,y;\eta,\xi)] \quad \forall \xi\in\Xi\subseteq\mathbb{R}^{d_\xi},\ x\in\mathbb{R}^{d_x}, \end{array} \tag{1}$$

see (Hu et al., 2023b). We assume that $g(x, y; \xi, \eta)$ is strongly convex in y for any fixed $x \in \mathbb{R}^{d_x}$, $\xi \in \Xi$ and $\eta \in \mathbb{R}^{d_\eta}$ so that the lower-level minimizer $y^*(x; \xi)$ is unique. Further assumptions on the functions f and g are presented in Section 2. The upper-level expectation is with respect to the joint distribution of the two random vectors ξ and η, while the lower-level expectation is with respect to the random vector η conditional on ξ (throughout we fix a regular conditional distribution $\mathbb{P}_{\eta|\xi}$). We allow the support Ξ of ξ to be uncountably infinite, and we assume no knowledge of the distributions, only access to oracles generating independent samples from $\mathbb{P}_\xi$ and $\mathbb{P}_{\eta|\xi}$.

CSBO subsumes stochastic bilevel optimization (SBO) (Ghadimi & Wang, 2018; Kwon et al., 2023) as a special case. Compared with SBO, CSBO offers two distinctive modeling advantages: (i) the lower-level decision y in CSBO can be coupled not only with the upper-level decision x but also with the context ξ; (ii) the number of lower-level decision makers in CSBO can be arbitrary as Ξ is not necessarily a finite set. Unlike SBO, CSBO also generalizes contextual stochastic optimization (Bertsimas & Kallus, 2020) and conditional stochastic optimization (Hu et al., 2020a;b). Thus, CSBO serves as a versatile modeling paradigm with a wide range of applications, such as meta-learning (Rajeswaran et al., 2019), contextual inventory management (Ban & Rudin, 2019), end-to-end learning (Rychener et al., 2023), personalized federated learning (Shamsian et al., 2021), hierarchical representation learning (Yao et al., 2019), Wasserstein distributionally robust optimization with side information (Yang et al., 2026; Donti et al., 2017), instrumental variable regression (Muandet et al., 2020; Kwon et al., 2023), and dynamic pricing (Guo, 2023; Faruqui & Sergici, 2010).

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.

The increased modeling power of CSBO comes at a high computational cost, and many algorithms developed for SBO do not extend to CSBO. In particular, numerous single-loop methods that are efficient for SBO (Guo et al., 2021; 2025; Chen et al., 2021; 2022b; 2024; Hong et al., 2023; Yang et al., 2021), where each iteration simultaneously updates the upper- and lower-level variables without solving an inner subproblem to completion, cannot be directly applied to solve CSBO. Indeed, in SBO, the lower-level solution $y^\star(x)$ depends solely on the upper-level decision variable x, whereas in CSBO, the lower-level solution $y^\star(x;\xi)$ depends on both x and the context ξ. This structural difference also undermines the warm-start strategies employed in many double-loop SBO algorithms (Kwon et al., 2023; 2024b; Chen et al., 2025; Gong et al., 2024), where the inner-loop solver is initialized with the solution from the previous outer iteration, thus degrading their practical performance and invalidating their theoretical guarantees. As a remedy, Bouscary et al. (2026) recently proposed a method to reformulate CSBO as an instance of SBO, thereby making CSBO amenable to existing SBO algorithms. While this approach represents a promising step toward bridging the two problem classes, it relies on restrictive structural assumptions. Specifically, it requires the lower-level objective $\mathbb{E}_{\eta|\xi}[g(x,y;\eta,\xi)]$ to be real-analytic in (y,ξ) and the context ξ to be either a discrete random variable with finite support or a continuous random variable whose density function is uniformly bounded away from zero on a compact support. These requirements may be restrictive in practice, motivating the development of alternative approaches that rely on weaker assumptions.

For general CSBO, Hu et al. (2023b) proposed a double-loop algorithm in which the outer loop performs stochastic gradient descent (SGD) on the upper-level problem, while the inner loop approximates the lower-level minimizer $y^\star(x;\xi)$ to construct an upper-level gradient estimator. This algorithm achieves sample and gradient complexities of $\widetilde{\mathcal{O}}(\epsilon^{-6})$. To reduce the number of inner iterations, Hu et al. (2023b) further developed an accelerated variant using the random truncated multilevel Monte Carlo (RT-MLMC) technique, attaining the improved complexity $\widetilde{\mathcal{O}}(\epsilon^{-4})$. However, both algorithms require computing multiple Hessian estimates and a matrix of mixed second-order derivatives of g at each iteration, resulting in high per-iteration costs that limit their practical efficiency.

To address the above challenges, we aim to develop a fully first-order algorithm for solving general CSBO problems, while matching the sample and gradient complexity orders of Hessian-based algorithms in (Hu et al., 2023b). Our contributions can be summarized as follows.

- We propose a fully first-order double-loop algorithm (see Algorithm 1) for CSBO and prove that its sample and gradient complexities are both $\widetilde{\mathcal{O}}(\epsilon^{-6})$, matching the complexity order of the Hessian-based double-loop algorithm in (Hu et al., 2023b). Our algorithm differs fundamentally from that in (Hu et al., 2023b) and is based on a suitable penalty formulation of problem (1). To our best knowledge, this is the first fully first-order method for nonconvex-strongly-convex CSBO that handles general context distributions without requiring a second-order oracle of g.
- To reduce the number of inner iterations, we develop an accelerated variant of our algorithm using the RT-MLMC technique, which achieves expected sample and gradient complexities of $\widetilde{\mathcal{O}}(\epsilon^{-4})$. This matches the complexity order of the RT-MLMC-based algorithm of Hu et al. (2023b) while requiring only first-order information.
- Unlike in (Hu et al., 2023b), a growing penalty parameter, an artifact of using the RT-MLMC technique within our fully first-order design, inflates the variance of the MLMC gradient estimator and destabilizes the algorithm in practice. To address this challenge, we propose a novel level-dependent stepsize schedule that stabilizes the algorithm in practice without compromising the theoretical complexity. To the best of our knowledge, this is the first such strategy for empirical stabilization in RT-MLMC-based gradient algorithms, and may be of independent interest.
- We validate our algorithms on two real-data applications, a meta-learning task and a contextual dynamic pricing problem for residential electricity demand response (Guo, 2023; Faruqui & Sergici, 2010), showing that they outperform the Hessian-based algorithm of Hu et al. (2023b) and the reduction-based algorithm of Bouscary et al. (2026).

1.1 Related Work

Bilevel optimization is an active research area with a long history. Below we provide a brief review of recent developments, focusing on SBO and CSBO. Assuming access to second-order oracles for g, Ghadimi & Wang (2018) developed the first SBO algorithm with a provable non-asymptotic

complexity guarantee, in which the inner loop computes an approximate minimizer $y^\star(x)$ that is then used to estimate the gradient of F. This seminal work spurred the development of a diverse range of second-order methods, including stocBiO (Ji et al., 2021), SOBA and SABA (Dagréou et al., 2022), SPABA (Chu et al., 2024), MDBO (Gao et al., 2023), TTSA (Hong et al., 2023), SUSTAIN (Khanduri et al., 2021), ALSET (Chen et al., 2021), SVRB (Guo et al., 2021), and BSVRB (Hu et al., 2023a). Among these, SVRB and BSVRB are designed for SBO with multiple lower-level problems, which can be viewed as a special case of CSBO in which ξ takes finitely many values.

The computational burden of second-order derivatives and Hessian inverses required by the above algorithms motivates the design of fully first-order methods for SBO, pioneered by Kwon et al. (2023). Such methods fall into two classes: (i) single-loop methods based on Lagrangian or penalty reformulations, such as Kwon et al. (2024b), and (ii) double-loop methods that maintain the bilevel hierarchy and approximately solve the lower-level subproblem at each outer iteration, such as F^{2}SA-p (Chen et al., 2026), F^{2}BA and F^{2}BSA (Chen et al., 2025), and the method of Kwon et al. (2024a).[1]

A common feature of all the aforementioned algorithms is their exploitation of the fact that the lower-level minimizer $y^\star(x)$ depends only on the upper-level variable x. For instance, SPABA (Chu et al., 2024) achieves optimal sample complexity for arbitrary distributions of η and ξ, yet assumes that the follower minimizes an *un*conditional expectation. In CSBO, by contrast, the lower-level minimizer $y^\star(x;\xi)$ depends on both x and the contextual variable ξ, which prevents warm-starting the inner loop across iterations without additional control of inter-context solution variation; when the support Ξ is finite, per-context caches are possible but the complexity scales linearly with $|\Xi|$. This renders the above algorithms inapplicable to general CSBO. This distinction gives rise to substantially greater analytical and computational challenges compared to SBO.

In the CSBO setting, Hu et al. (2023b) proposed a double-loop algorithm based on second-order oracles and developed an accelerated variant by incorporating the RT-MLMC technique. Both algorithms were discussed in detail above. Bouscary et al. (2026) proposed an alternative approach that reformulates CSBO as an SBO problem, enabling the use of standard SBO algorithms, but requires real-analyticity of the lower-level objective and additional assumptions on ξ, as discussed above. Finally, several works have studied bilevel reinforcement learning as a special case of CSBO (Chen et al., 2022a; Chakraborty et al., 2024; Shen et al., 2025; Yang et al., 2025; Thoma et al., 2024), but their methods exploit problem-specific structure and do not extend to general CSBO.

1.2 Preliminaries and Notation

We write $\widetilde{\mathcal{O}}(\cdot)$ for the big-O notation with polylogarithmic factors suppressed, and $[M] := \{1,\dots,M\}$ for a positive integer M. For a function $\psi : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$, we denote its gradient and Hessian with respect to the full argument (x,y) by $\nabla\psi$ and $\nabla^2\psi$, and use subscripts to indicate partial derivatives: $\nabla_1\psi$, $\nabla_2\psi$ are the gradients with respect to x and y; $\nabla^2_{11}\psi$, $\nabla^2_{22}\psi$ are the corresponding Hessians; and $\nabla^2_{12}\psi$, $\nabla^2_{21}\psi$ are the $d_x \times d_y$ and $d_y \times d_x$ matrices of mixed second-order partial derivatives, with (i,j)-th entries $\partial^2_{x_i y_j}\psi$ and $\partial^2_{y_i x_j}\psi$, respectively. For a function $h : \mathbb{R}^{d_z} \to \mathbb{R}^{d_h}$, we say h is L-Lipschitz continuous if $\|h(z_1) - h(z_2)\| \le L\|z_1 - z_2\|$ for all $z_1, z_2 \in \mathbb{R}^{d_z}$. The Jacobian of h is the $d_h \times d_z$ matrix $Dh := [\nabla h_1, \dots, \nabla h_{d_h}]^\top$, where h_i denotes the i-th component of h. For $z = (z_1, z_2)$, the partial Jacobian of h with respect to z_i is denoted as $D_i h$. When $d_h = 1$, we say that h is S-smooth if it is differentiable with S-Lipschitz continuous gradient, and we further say that h is μ-strongly convex if $h(z) - \frac{\mu}{2}\|z\|^2$ is convex.

2 Algorithms

We first state our standing assumptions and fix the notion of approximate stationarity. Then, based on a penalty formulation (2), we develop our double-loop algorithm and its RT-MLMC variant for the CSBO problem (1).

2.1 Assumptions and Aims

Throughout, $\|\cdot\|$ denotes the Euclidean norm on vectors and the induced operator norm on matrices. Furthermore, we assume that all sample functions and derivatives appearing below are integrable whenever an expectation is taken. Our results will rely on the following regularity condition.

[1] The complexity results of Kwon et al. (2024a) are particularly relevant to ours. A detailed comparison, including a technical remark on their upper-bound analysis, is provided in Appendix A.2.

Assumption 2.1. *Problem* (1) *obeys the following regularity conditions for all* $\eta \in \mathbb{R}^{d_\eta}$ *and* $\xi \in \Xi$.

(i) $f(x, y; \eta, \xi)$ *and* $g(x, y; \eta, \xi)$ *are twice continuously differentiable in* (x, y).

(ii) $g(x, y; \eta, \xi)$ *is* μ_g*-strongly convex in* y *for all* $x \in \mathbb{R}^{d_x}$.

(iii) f, ∇f, ∇g, *and* $\nabla^2 g$ *are* $\ell_{f,0}$*-*, $\ell_{f,1}$*-*, $\ell_{g,1}$*-, and* $\ell_{g,2}$*-Lipschitz continuous in* (x, y), *respectively.*

(iv) For some $\tau_g > 0$ *and all* $x \in \mathbb{R}^{d_x}$ *and* $y \in \mathbb{R}^{d_y}$, *the conditional variance of* ∇g *satisfies*

$$\mathbb{E}\big[\|\nabla g(x, y; \eta, \xi) - \mathbb{E}[\nabla g(x, y; \eta, \xi) \mid \xi]\|^2 \mid \xi\big] \leq \tau_g^2.$$

(v) The upper-level objective is bounded below, i.e., $F^\star := \inf_{x \in \mathbb{R}^{d_x}} F(x) > -\infty$.

Note that (iii) implies $\|\nabla f\| \leq \ell_{f,0}$. Hence, $\mathbb{E}[\|\nabla f - \mathbb{E}[\nabla f|\xi]\|^2|\xi]$, and the variance of $\mathbb{E}[\nabla f|\xi]$ are both bounded by $\ell_{f,0}^2$. By Assumptions 2.1 (i) and 2.1 (iii), dominated convergence allows differentiation under the conditional expectation, i.e., $\mathbb{E}_{\eta|\xi}[\nabla f(x, y; \eta, \xi)] = \nabla \mathbb{E}_{\eta|\xi}[f(x, y; \eta, \xi)]$, and likewise for g. Moreover, Assumption 2.1 (ii) implies that $\mathbb{E}_{\eta|\xi}[g(x, \cdot; \eta, \xi)]$ is μ_g-strongly convex and coercive, and therefore the lower-level problem admits a unique minimizer $y^\star(x; \xi)$.

Remark 2.1. *Conditions of the form in Assumption 2.1 are standard in the SBO and CSBO literature Ghadimi & Wang (2018); Guo et al. (2021); Chen et al. (2021; 2022b); Hong et al. (2023); Hu et al. (2023b). Although (iii) constrains* $\nabla^2 g$*, our algorithms never query a second-order oracle.*

Given a stationarity accuracy $\epsilon > 0$, we seek algorithms whose iterates $x_1, \ldots, x_K$ are asymptotically optimal in the sense that, for $R \sim \mathrm{Unif}([K])$ independent of all oracle randomness, we have

$$\mathbb{E}[\|\nabla F(x_R)\|^2] = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|\nabla F(x_k)\|^2] \leq \epsilon^2.$$

2.2 The Basic Algorithm

Inspired by the work of Kwon et al. (2023) on SBO, our algorithm rests on the penalty formulation

$$\min_{x \in \mathbb{R}^{d_x},\, z \in \mathcal{Z}} \quad \mathbb{E}_{\eta,\xi}\big[L\big(x, z(\xi), y^\star(x; \xi), \lambda; \eta, \xi\big)\big], \tag{2}$$

where $\mathcal{Z}$ is the space of all measurable contextual decision rules $z : \Xi \to \mathbb{R}^{d_y}$, $\lambda > 0$ is a penalty parameter and $L(x, z, y, \lambda; \eta, \xi) := f(x, z; \eta, \xi) + \lambda(g(x, z; \eta, \xi) - g(x, y; \eta, \xi))$. Writing $\bar{f}(x, y; \xi) := \mathbb{E}_{\eta|\xi}[f(x, y; \eta, \xi)]$ and $\bar{g}(x, y; \xi) := \mathbb{E}_{\eta|\xi}[g(x, y; \eta, \xi)]$, the penalty term in the objective function of (2) is the expected value of the lower-level optimality gap $\bar{g}(x, z(\xi); \xi) - \bar{g}(x, y^\star(x; \xi); \xi) \geq 0$, which vanishes precisely when $z(\xi) = y^\star(x; \xi)$ almost surely. Moreover, (2) decouples across contexts. For any fixed x, minimizing the objective function of (2) over functions $z \in \mathcal{Z}$ amounts to minimizing the perturbed lower-level objective $\bar{g}(x, z; \xi) + \frac{1}{\lambda}\bar{f}(x, z; \xi)$ over vectors $z \in \mathbb{R}^{d_y}$ simultaneously for all ξ. Hence, we define $y_\gamma^\star(x; \xi) := \arg\min_y \{\bar{g}(x, y; \xi) + \gamma \bar{f}(x, y; \xi)\}$ for any sufficiently small $\gamma \geq 0$, so that $y_0^\star = y^\star$ is the lower-level minimizer in (1), and $y_{1/\lambda}^\star$ is the optimal penalized decision rule in (2). As $\nabla \bar{f}$ is $\ell_{f,1}$-Lipschitz and $\bar{g}$ is μ_g-strongly convex by Assumption 2.1, the perturbed objective is strongly convex in y, and $y_{1/\lambda}^\star$ is well-defined if $\lambda > \ell_{f,1}/\mu_g$.

Below we suppress the arguments (x, ξ) of $y_0^\star$ and $y_{1/\lambda}^\star$ to avoid clutter, and we argue informally. Computing $\mathrm{d}y_\gamma^\star/\mathrm{d}\gamma|_{\gamma=0}$ via the implicit function theorem, we obtain the Taylor expansion

$$y_{1/\lambda}^\star - y_0^\star = -\frac{1}{\lambda} \left[\nabla_{22}^2 \bar{g}(x, y_0^\star; \xi)\right]^{-1} \nabla_2 \bar{f}(x, y_0^\star; \xi) + \mathcal{O}(\lambda^{-2}). \tag{3}$$

Thus, for large λ the two minimizers are close, and their displacement captures the action of the inverse lower-level Hessian. In addition, we have

$$\begin{aligned}
\mathbb{E}_{\eta|\xi}\Big[\nabla_1 L\big(x, y_{1/\lambda}^\star, y_0^\star, \lambda; \eta, \xi\big)\Big] &= \nabla_1 \bar{f}(x, y_{1/\lambda}^\star; \xi) + \lambda\big(\nabla_1 \bar{g}(x, y_{1/\lambda}^\star; \xi) - \nabla_1 \bar{g}(x, y_0^\star; \xi)\big) \\
&= \nabla_1 \bar{f}(x, y_{1/\lambda}^\star; \xi) + \lambda \nabla_{12}^2 \bar{g}(x, y_0^\star; \xi)\big(y_{1/\lambda}^\star - y_0^\star\big) + \mathcal{O}\big(\lambda \|y_{1/\lambda}^\star - y_0^\star\|^2\big) \\
&= \nabla_1 \bar{f}(x, y_{1/\lambda}^\star; \xi) - \nabla_{12}^2 \bar{g}(x, y_0^\star; \xi) \left[\nabla_{22}^2 \bar{g}(x, y_0^\star; \xi)\right]^{-1} \nabla_2 \bar{f}(x, y_0^\star; \xi) + \mathcal{O}\big(\lambda^{-1}\big),
\end{aligned}$$

Algorithm 1 Fully first-order double-loop method for CSBO

Input: $x_1 \in \mathbb{R}^{d_x}$, $K \in \mathbb{N}$, a sufficiently small constant $c_\alpha > 0$, a sufficiently large constant $C_T > 0$

1: **for** $k = 1, \dots, K$ **do**
2: Set $\lambda_k = \frac{2\ell_{f,1}}{\mu_g}(k+1)^{1/4}$, $\alpha_k = \frac{c_\alpha}{\sqrt{k+1}}$, $T_k = \lceil C_T \lambda_k^2 \log(\lambda_k + 1) \rceil$.
3: Sample ξ_k from $\mathbb{P}_\xi$, compute an initial point y_k^0 satisfying Assumption 2.2, and set $z_k^0 = y_k^0$.
4: **for** $t = 0, 1, \dots, T_k - 1$ **do**
5: Sample η_k^t from $\mathbb{P}_{\eta|\xi_k}$, set $\beta_t = \min\{\frac{8}{\mu_g(t+1)}, \frac{\mu_g}{16\ell_{g,1}^2}\}$.
6: $y_k^{t+1} = y_k^t - \beta_t \nabla_2 g(x_k, y_k^t; \eta_k^t, \xi_k)$
7: $z_k^{t+1} = z_k^t - \frac{\beta_t}{\lambda_k} \nabla_2 L(x_k, z_k^t, y_k^t, \lambda_k; \eta_k^t, \xi_k)$.
8: **end for**
9: Set $z_{k+1} = z_k^{T_k}$, $y_{k+1} = y_k^{T_k}$
10: Sample η_k from $\mathbb{P}_{\eta|\xi_k}$
11: $x_{k+1} = x_k - \alpha_k \nabla_1 L(x_k, z_{k+1}, y_{k+1}, \lambda_k; \eta_k, \xi_k)$.
12: **end for**

Output: x_R, where $R \sim \mathrm{Unif}([K])$ is independent of all oracle randomness.

where the first and the second equality follow from the definition of L and a Taylor expansion of $\nabla_1 \bar{g}$ with respect to y, while the last equality exploits (3). Finally, one can show that

$$\nabla F(x) = \mathbb{E}_\xi \left[\nabla_1 \bar{f}(x, y_0^\star; \xi) - \nabla_{12}^2 \bar{g}(x, y_0^\star; \xi) \left[\nabla_{22}^2 \bar{g}(x, y_0^\star; \xi) \right]^{-1} \nabla_2 \bar{f}(x, y_0^\star; \xi) \right],$$

see, e.g., (Hu et al., 2023b, § 2). Comparison of the last two equations reveals that

$$\left\| \nabla F(x) - \mathbb{E}_{\eta,\xi} \left[\nabla_1 L\big(x, y_{1/\lambda}^\star, y_0^\star, \lambda; \eta, \xi\big) \right] \right\| = \mathcal{O}(\lambda^{-1}),$$

that is, $\nabla_1 L$ provides a stochastic estimator of the hypergradient ∇F with bias $\mathcal{O}(\lambda^{-1})$, computable from first-order derivatives of f and g alone. See Lemma A.8 for a rigorous derivation.

This motivates Algorithm 1. The outer loop, indexed by k, performs inexact SGD on F. Line 11 moves x_k along $\nabla_1 L(x_k, z_{k+1}, y_{k+1}, \lambda_k; \eta_k, \xi_k)$, whose bias $\mathcal{O}(\lambda_k^{-1})$ vanishes as $\lambda_k \to \infty$. Its two remaining inputs are supplied by the inner loop, indexed by t, which performs SGD on $\bar{g}(x_k, \cdot; \xi_k)$ and $\bar{g}(x_k, \cdot; \xi_k) + \frac{1}{\lambda_k} \bar{f}(x_k, \cdot; \xi_k)$ in order to compute $y_{k+1} \approx y_0^\star(x_k; \xi_k)$ and $z_{k+1} \approx y_{1/\lambda_k}^\star(x_k; \xi_k)$. Since a fresh context ξ_k is drawn at every outer iteration, the inner loop cannot be warm-started from y_k, which is precisely the obstacle that distinguishes CSBO from SBO. The inner accuracy is instead controlled by letting T_k grow with λ_k. See Algorithm 1 and Appendix A for details.

2.3 The Accelerated Algorithm

In Algorithm 1, the hypergradient bias is progressively reduced by increasing the penalty parameter λ_k and the number of inner iterations T_k. This results in a heavy computational burden for large k. A natural alternative is to retain the same biased SGD structure but maintain a fixed, accuracy-dependent bias. Specifically, to obtain an output x_R satisfying $\mathbb{E}[\|\nabla F(x_R)\|^2] \le \epsilon^2$, it suffices to control the bias at order $\mathcal{O}(\epsilon)$, rather than require it to decrease indefinitely with k. However, using the same hypergradient estimator with a deterministic inner-loop budget would still require sufficiently accurate solutions of both lower-level subproblems at every outer iteration. Under the inner-loop accuracy prescription of Algorithm 1, this entails choosing $\lambda = \Theta(\epsilon^{-1})$ and using $\widetilde{\mathcal{O}}(\epsilon^{-2})$ inner iterations per outer iteration. Thus, replacing an increasing inner-loop budget by a fixed one does not by itself resolve the computational bottleneck.

To reduce this cost, we construct a new hypergradient estimator using the RT-MLMC technique (Hu et al., 2023b; 2021). The basic idea is to retain the small bias associated with an accurate hypergradient approximation without paying the corresponding large inner-loop cost at every outer iteration. To describe this construction, we first introduce the following *base estimator*, which we define as the hypergradient estimator computed by Algorithm 1 with $T_k = 2^n$ inner iterations:

$$G_k(n, \lambda) := \nabla_1 L\big(x_k, z_k^{2^n}(\lambda), y_k^{2^n}, \lambda; \eta_k, \xi_k\big). \tag{4}$$

Here, $k \in [K]$ denotes the outer iteration count, while $n \in \mathbb{N}$ is a hyperparameter that determines how accurately the inner loop solves the subproblems. Indeed, $y_k^{2^n}$ and $z_k^{2^n}(\lambda)$ are approximate minimizers of $\bar{g}(x_k, \cdot; \xi_k)$ and its perturbed version $\bar{g}(x_k, \cdot; \xi_k) + \frac{1}{\lambda} \bar{f}(x_k, \cdot; \xi_k)$, respectively, obtained

Algorithm 2 Adaptive RT-MLMC fully first-order method for CSBO

Input: $x_1 \in \mathbb{R}^{d_x}$, $N = \mathcal{O}(1)\log(\epsilon^{-1})$, $a_1 \in (0,1)$, $\alpha_0 = \mathcal{O}(\epsilon^2/N)$ with $\alpha_0 \le a_1/(8\ell_{F,1})$, $c_0 \in (0,1]$

1: **for** $k = 1, \dots, K$ **do**
2: Sample ν_k from $\mathbb{P}_\nu$, where $\mathbb{P}_\nu := \sum_{n=1}^N p_n \delta_n$, $p_n := \frac{2^{-n}}{1-2^{-N}}$, and δ_n denotes the Dirac measure at n.
3: Sample ξ_k from $\mathbb{P}_\xi$.
4: Set $\lambda_j = \frac{2\ell_{f,1}}{\mu_g}(2^j)^{\frac{1}{2}}$ for $j \in \{0, \nu_k - 1, \nu_k\}$, compute an initial point y_k^0 satisfying Assumption 2.2, and set $z_k^0(\lambda_{\nu_k}) = z_k^0(\lambda_{\nu_k-1}) = y_k^0$.
5: **for** $t = 0, 1, \dots, 2^{\nu_k} - 1$ **do**
6: Sample η_k^t from $\mathbb{P}_{\eta|\xi_k}$, set $\beta_t = \min\{\frac{8}{\mu_g(t+1)}, \frac{\mu_g}{16\ell_{g,1}^2}\}$.
7: $y_k^{t+1} = y_k^t - \beta_t \nabla_2 g(x_k, y_k^t; \eta_k^t, \xi_k)$
8: $z_k^{t+1}(\lambda_{\nu_k}) = z_k^t(\lambda_{\nu_k}) - \frac{\beta_t}{\lambda_{\nu_k}} \nabla_2 L(x_k, z_k^t(\lambda_{\nu_k}), y_k^t, \lambda_{\nu_k}; \eta_k^t, \xi_k)$
9: $z_k^{t+1}(\lambda_{\nu_k-1}) = z_k^t(\lambda_{\nu_k-1}) - \frac{\beta_t}{\lambda_{\nu_k-1}} \nabla_2 L(x_k, z_k^t(\lambda_{\nu_k-1}), y_k^t, \lambda_{\nu_k-1}; \eta_k^t, \xi_k)$.
10: **If** $t = 0$ and $\nu_k > 1$, **then** $z_k^1(\lambda_0) = y_k^0 - \frac{\beta_0}{\lambda_0} \nabla_2 L(x_k, z_k^0(\lambda_0), y_k^0, \lambda_0; \eta_k^0, \xi_k)$.
11: **end for**
12: Sample η_k from $\mathbb{P}_{\eta|\xi_k}$.
13: **If** $\nu_k > c_0 N$, **then** $\alpha = a_1\alpha_0$; **else** $\alpha = \alpha_0$.
14: $x_{k+1} = x_k - \alpha\big(G_k(0,\lambda_0) + p_{\nu_k}^{-1}[G_k(\nu_k, \lambda_{\nu_k}) - G_k(\nu_k - 1, \lambda_{\nu_k-1})]\big)$
15: **end for**

Output: x_R, where $R \sim \mathrm{Unif}([K])$ is independent of all oracle randomness.

by running SGD for 2^n iterations. To apply the RT-MLMC technique, we let λ depend on n, so that the base estimator is parametrized by the single parameter n, which we call the *level*. With a suitable increasing sequence $\{\lambda_n\}$, together with the increasing inner-loop budget 2^n, the bias of the level-n base estimator $G_k(n, \lambda_n)$ decreases as n increases. Let N be the highest level and let ν be a random variable supported on $[N]$ with probability mass function $\mathbb{P}(\nu = n) = p_n$. Then the crux of the RT-MLMC technique lies in the following telescoping observation:

$$\begin{aligned} G_k(N, \lambda_N) &= G_k(0,\lambda_0) + \textstyle\sum_{n=1}^N p_n \frac{G_k(n,\lambda_n) - G_k(n-1,\lambda_{n-1})}{p_n} \\ &= G_k(0,\lambda_0) + \mathbb{E}_\nu\left[\frac{G_k(\nu,\lambda_\nu) - G_k(\nu-1,\lambda_{\nu-1})}{p_\nu}\right]. \end{aligned} \tag{5}$$

By sampling a random level ν, we can thus define the following *RT-MLMC hypergradient estimator*:

$$G_k(0,\lambda_0) + p_\nu^{-1}\big(G_k(\nu,\lambda_\nu) - G_k(\nu-1,\lambda_{\nu-1})\big). \tag{6}$$

As is evident from (5), the RT-MLMC estimator (6) is an unbiased estimator for the highest-level base estimator $G_k(N, \lambda_N)$ and thus has the same (small) bias as $G_k(N, \lambda_N)$ with respect to ∇F. Lemma A.13 shows that this bias is of order $\mathcal{O}(1/\lambda_N)$. However, constructing the RT-MLMC estimator only requires constructing three base estimators at levels 0, ν and $\nu - 1$. If ν is small, this can be significantly cheaper than constructing $G_k(N, \lambda_N)$. To ensure that evaluating the RT-MLMC estimator is cheap with high probability, we let ν skew towards the left. Specifically, we take ν to be a truncated geometric random variable on $[N]$ by setting $p_n = 2^{-n}/(1 - 2^{-N})$ for all $n \in [N]$. Algorithm 2 is designed to perform SGD on $F(x)$ using the RT-MLMC estimator (6).

Compared with the Hessian-based RT-MLMC method of Hu et al. (2023b), integrating the RT-MLMC technique into our penalty-based framework introduces an additional variance challenge. More precisely, in order for $\nabla_1 L$ to provide an accurate approximation of ∇F, the penalty parameter λ must increase with the level n. Since λ multiplies the difference of the stochastic gradients of g in $\nabla_1 L$, a large λ amplifies the fluctuations in the resulting base estimators and, consequently, in the RT-MLMC corrections. Moreover, high sampled levels n are associated with small probabilities p_n, so that the corresponding importance-weighted corrections can occasionally be large. As a result, although the RT-MLMC construction already yields an accelerated complexity comparing to the Algorithm 1, its numerical performance can be unstable in practice.

To tackle this issue, we develop an adaptive stepsize strategy; see step 14 in Algorithm 2. Specifically, if the sampled level ν_k exceeds a given threshold, we shrink the stepsize by a factor $a_1 \in (0,1)$. This damps the effect of the potentially large high-level RT-MLMC corrections and improves the stability of the outer updates. Moreover, the resulting algorithm, Algorithm 2, retains the improved

complexity $\widetilde{\mathcal{O}}(\epsilon^{-4})$. To the best of our knowledge, this is the first use of such a level-dependent stepsize strategy to stabilize RT-MLMC-type stochastic gradient methods. An ablation study of the adaptive parameters a_1 and c_0 is provided in Appendix B.1.4, showing that the adaptive stepsize substantially reduces the variability of the iterates relative to the non-adaptive variant while remaining competitive in mean performance.

Finally, for the purpose of the theoretical analysis, we assume that the initialization gap in the lower-level problems is bounded. This assumption is also used implicitly in (Hu et al., 2023b); see for example the proof of Lemma 3 therein.

Assumption 2.2. *There exists $b > 0$ such that for any $k \geq 1$, we have*

$$\mathbb{E}_{\eta|\xi_k}[g(x_k, y_k^0; \eta, \xi_k) - g(x_k, y^\star(x_k, \xi_k); \eta, \xi_k)] \leq b.$$

Assumption 2.2 holds for a common class of lower-level problems with quadratic regularization. Specifically, suppose that $g(x, y; \eta, \xi) = h(x, y; \eta, \xi) + \frac{\gamma}{2}\|y - r(x, \xi)\|^2$, where $\gamma > 0$, $r : \mathbb{R}^{d_x} \times \Xi \to \mathbb{R}^{d_y}$, and $h(x, \cdot; \eta, \xi)$ is convex and G-Lipschitz continuous uniformly in (x, η, ξ). One can initialize the inner loop at $y_k^0 = r(x_k, \xi_k)$. Then, for fixed (x, ξ), writing $r = r(x, \xi)$, $y^\star = y^\star(x, \xi)$, and $d = \|y^\star - r\|$ for simplicity, we have

$$\bar{g}(x, r; \xi) - \bar{g}(x, y^\star; \xi) = \bar{h}(x, r; \xi) - \bar{h}(x, y^\star; \xi) - \tfrac{\gamma}{2}d^2 \leq Gd - \tfrac{\gamma}{2}d^2 = \tfrac{G^2}{2\gamma} - \tfrac{\gamma}{2}\left(d - \tfrac{G}{\gamma}\right)^2 \leq \tfrac{G^2}{2\gamma}.$$

Thus, Assumption 2.2 holds with $b = G^2/(2\gamma)$.

In particular, the meta-learning problem (7) in Section 4.1 is of this form: when the data features are uniformly bounded, the multiclass logistic loss is uniformly Lipschitz, and the sufficient condition holds by taking $y_k^0 = x_k$. More generally, for general strongly convex g, write $g = h + \frac{\mu_g}{2}\|\cdot\|^2$ with $h := g - \frac{\mu_g}{2}\|\cdot\|^2$, then h is convex. If h is further uniformly G-Lipschitz continuous, we can apply the same argument to claim that Assumption 2.2 holds by taking $y_k^0 = 0$.

For the dynamic-pricing problem in (8), although it is not of the regularized form above, Assumption 2.2 still holds; see Appendix B.2.5 for details.

3 Complexity Analysis

Due to the bilevel structure and potential non-convexity of f, the objective function F is in general nonconvex in x and thus we aim to find an ϵ-stationarity point x_R as defined in Section 2.1.

Our first main theoretical result concerns the sample and gradient complexities of Algorithm 1. The proof can be found in Appendix A.9.

Theorem 3.1. *Suppose that Assumptions 2.1 and 2.2 hold. For the output x_R of Algorithm 1 to be an ϵ-stationarity point, it suffices to set $K = \widetilde{\mathcal{O}}(\epsilon^{-4})$. Moreover, the overall sample and gradient complexities are of order $\widetilde{\mathcal{O}}(\epsilon^{-6})$. More precisely, the sample complexity of ξ and the gradient complexities of $\nabla_1 f$, $\nabla_1 g$ are of order $\widetilde{\mathcal{O}}(\epsilon^{-4})$, while the sample complexity of η and the gradient complexities of $\nabla_2 g$, $\nabla_2 f$ are of order $\widetilde{\mathcal{O}}(\epsilon^{-6})$.*

Thanks to the RT-MLMC technique, which greatly reduces the average number of inner iterations, we next show the improved theoretical complexities. A key step is to analyze the variance of the RT-MLMC gradient estimator in (6), summarized in the following lemma.

Lemma 3.1. *Under Assumptions 2.1 and 2.2, consider Algorithm 2, we have*

$$\mathbb{E}\Big[\Big\|\mathbb{E}\big[\nabla_1 L(x_k, z_k^{2^N}(\lambda_N), y_k^{2^N}, \lambda_N; \eta_k, \xi_k) \mid \mathcal{F}_k\big] - \big(G_k(0, \lambda_0) + p_{\nu_k}^{-1}(G_k(\nu_k, \lambda_{\nu_k}) - G_k(\nu_k - 1, \lambda_{\nu_k - 1}))\big)\Big\|^2 \mid \mathcal{F}_k\Big] \leq \mathcal{O}(N),$$

where $\mathcal{F}_k$ denote the σ-algebra generated by all randomness revealed prior to outer iteration k.

Remark 3.1. *The variance of the Hessian-based RT-MLMC gradient estimator in Hu et al. (2023b) is $\mathcal{O}(\log(\epsilon^{-1}))$ (see page 16 therein). Unlike Hessian-based algorithms, our Algorithm 2 uses only first-order information. Consequently, the corresponding penalty parameter amplifies the variance and requires additional treatment in the technical analysis. Specifically, with $N = \mathcal{O}(1)\log(\epsilon^{-1})$ as defined in Algorithm 2, the displacement estimate in Lemma A.14 yields a logarithmic variance bound $\mathcal{O}(\log(\epsilon^{-1})) = \widetilde{\mathcal{O}}(1)$ which leads to the improved complexity $\widetilde{\mathcal{O}}(\epsilon^{-4})$.*

Theorem 3.2. *Suppose that Assumptions 2.1 and 2.2 hold. For the output x_R of Algorithm 2 to be an ϵ-stationarity point, it suffices to set $K = \widetilde{\mathcal{O}}(\epsilon^{-4})$, $N = C_N \log(\epsilon^{-1})$, where C_N is sufficiently large depending on c_0, and $\alpha_0 = \Theta(\epsilon^2/N)$ with $\alpha_0 \le a_1/(8\ell_{F,1})$. Moreover, the expected sample complexities of ξ and η, and the expected gradient complexities of $\nabla_1 f$, $\nabla_1 g$, $\nabla_2 g$, and $\nabla_2 f$ are of order $\widetilde{\mathcal{O}}(\epsilon^{-4})$.*

Per accelerated iteration, the expected counts are one context ξ, $\mathcal{O}(N)$ conditional draws $\eta|\xi$, $\mathcal{O}(1)$ upper- and $\mathcal{O}(N)$ lower-level gradient evaluations; exact counts are in Appendix A.1.

Remark 3.2. *Our $\widetilde{\mathcal{O}}(\epsilon^{-4})$ rate matches the best-known Hessian-based CSBO complexity order in Hu et al. (2023b) and the exponent of the SBO lower bound in Chen et al. (2026). However, that lower bound embeds the hard instance of Arjevani et al. (2023), whose rare-reveal coordinate forces the per-sample gradient norms and smoothness constants to grow as $\epsilon \to 0$; hence it violates the uniform sample-wise regularity in Assumption 2.1(iii)and does not settle whether our rate is minimax for this class. A matching lower bound under these assumptions remains open; see Appendix A.2.*

Lemma A.9 is a key technical ingredient in our RT-MLMC variance analysis and the resulting improved $\widetilde{\mathcal{O}}(\epsilon^{-4})$ complexity bound. Establishing this result is nontrivial: separately bounding the two inner-loop mean-square errors $\mathbb{E}\|z_k^t(\lambda) - y_{1/\lambda}^\star\|^2$ and $\mathbb{E}\|y_k^t - y_0^\star\|^2$ is too coarse, because the resulting bounds are amplified by the growing penalty parameter λ in the hypergradient estimator. We instead directly control their coupled difference, $\big(z_k^t(\lambda) - y_{1/\lambda}^\star\big) - \big(y_k^t - y_0^\star\big)$, which yields the sharper bound in Lemma A.9. Moreover, a counterexample shows that the analogous contraction estimate used in Kwon et al. (2024a) need not hold under the stated assumptions. Our fixed-anchor estimate in (17) *avoids this issue and is sharper. In our setting, replacing Equation* (K) *with this bound yields a tighter overall complexity bound, suggesting that the same argument may also sharpen the analysis of related penalty-based methods for classical SBO.*

4 Numerical Experiments

We evaluate our proposed first-order algorithms on two real-world CSBO applications: meta-learning (Finn et al., 2017; Rajeswaran et al., 2019) and contextual dynamic pricing for residential electricity demand response (Guo, 2023; Faruqui & Sergici, 2010). We compare our methods against the RT-MLMC Hessian-based method (Hu et al., 2023b) and the reduction framework (Bouscary et al., 2026) with the reformulated SBO problem solved by stocBiO (Ji et al., 2021) ("reduction + stocBiO"). Complete implementation details and parameter settings are provided in Appendix B.

4.1 Meta-learning

We consider learning a shared meta-parameter across tasks so that task-specific adaptation on training data yields low validation loss. Complete experimental details are given in Appendix B.1, including an ablation analysis of the adaptive stepsize parameters in Appendix B.1.4. Formally, this problem is the following special case of CSBO:

$$\begin{aligned} &\min_{x\in\mathbb{R}^{d_x}} \ \mathbb{E}_{\xi\sim\mathbb{P}_\xi}\mathbb{E}_{\eta_\xi^{\mathrm{val}}\sim\mathbb{P}_{\eta|\xi}}[l_\xi(y^\star(x;\xi),\eta_\xi^{\mathrm{val}})] \\ &\text{where } y^\star(x;\xi) = \arg\min_{y\in\mathbb{R}^{d_y}} \mathbb{E}_{\eta_\xi^{\mathrm{tr}}\sim\mathbb{P}_{\eta|\xi}}\left[l_\xi(y,\eta_\xi^{\mathrm{tr}}) + \tfrac{\gamma}{2}\|y-x\|^2\right], \forall \xi\in[M],\ x\in\mathbb{R}^{d_x}. \end{aligned} \tag{7}$$

Here, $\mathbb{P}_\xi$ is the distribution over the M tasks, $\mathbb{P}_{\eta|\xi}$ is the task-specific data distribution, η_ξ^{tr} and η_ξ^{val} are the corresponding training and validation data, x is the shared meta-parameter, $y^\star(x;\xi)$ is the adapted task parameter, l_ξ is the task loss, and $\gamma > 0$ is the regularization parameter. We report the estimated upper-level objective and validation prediction error.

Figure 1 reports measurements averaged over 10 independent runs. Algorithm 2 achieves the largest objective reduction in wall-clock time and best validation error in the sample used, while both proposed methods attain lower validation error than the baselines across the reported sample budgets. Further discussion is provided in Appendix B.1.3.

4.2 Contextual Dynamic Pricing for Residential Electricity Demand

We consider a contextual dynamic pricing problem for residential electricity demand response (Guo, 2023; Faruqui & Sergici, 2010) using the Low Carbon London (LCL) dataset (UK Power Networks,

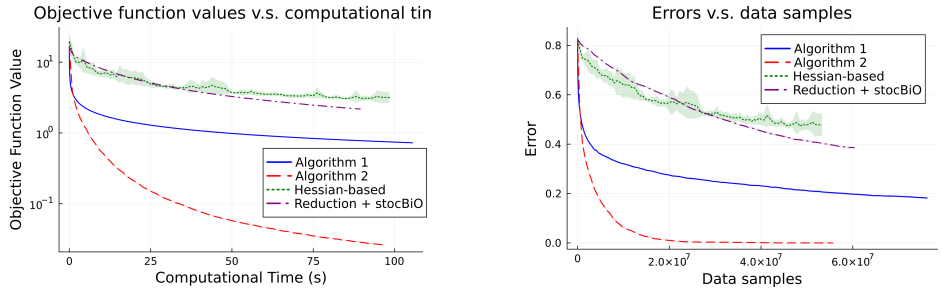


Figure 1: Meta-learning comparison of Algorithm 1, Algorithm 2, the RT-MLMC Hessian-based method (Hu et al., 2023b), and reduction + stocBiO (Bouscary et al., 2026; Ji et al., 2021). Left: objective versus computational time. Right: validation error versus number of data samples. Shaded regions show mean ± 1 standard deviation over 10 independent runs.

2015). A retailer chooses a context-dependent day-ahead tariff, while households adjust their consumption in response to the tariff and their baseline loads. This yields the following CSBO formulation:

$$f(x,z;\eta,\xi) = \lambda_{\text{peak}} \frac{\text{smax}_{\tau_p}(\ell)}{\ell_{\text{ref}}^{\text{peak}}} + \lambda_{\text{flat}} \frac{\text{Disp}_{\delta_f}(\ell)}{\ell_{\text{ref}}^{\text{flat}}} + \lambda_{\text{bill}} \frac{1}{H} \sum_{h=1}^{H} \rho_{\tau_b}\Big(\frac{p_x(\xi)^\top \eta_h - p_{\text{ref}}^\top \eta_h}{p_{\text{ref}}^\top \eta_h}\Big),$$
$$g(x,z;\eta,\xi) = \sum_{h=1}^{H} \Big[\Phi_{r(h)}(z_h) + \big(S(\eta_h)(p_x(\xi) - p_{\text{ref}})\big)^\top z_h\Big]. \tag{8}$$

Here, x parameterizes the context-dependent tariff and z represents household consumption adjustments. The lower-level objective g models the trade-off between discomfort and bill incentives, while f penalizes peak load, load dispersion, and bill deviations. We use $H = 100$, yielding $d_x = 30$ and $d_y = 4800$, and report the objective and the peak-to-average ratio (PAR), with lower values indicating better performance. Full model construction and evaluation details are provided in Appendix B.2.

Figure 2 reports measurements averaged over 10 independent runs. On this instance the proposed methods and the Hessian-based method end at nearly the same final metrics: the latter is marginally better on PAR and marginally worse on the objective. However, both proposed methods are faster than the Hessian-based method in wall-clock time. Algorithm 2 does not show acceleration compared with Algorithm 1 here because the lower-level problems are easy to solve: g has a moderate condition number and its minimizer lies close to the initialization, so the inner iterates make rapid progress in very few gradient steps and leave little inner-loop cost for the RT-MLMC estimator to spare. Conversely, the RT-MLMC estimator pays for its own variance and per-step overhead, so its theoretical gain never materializes on this instance

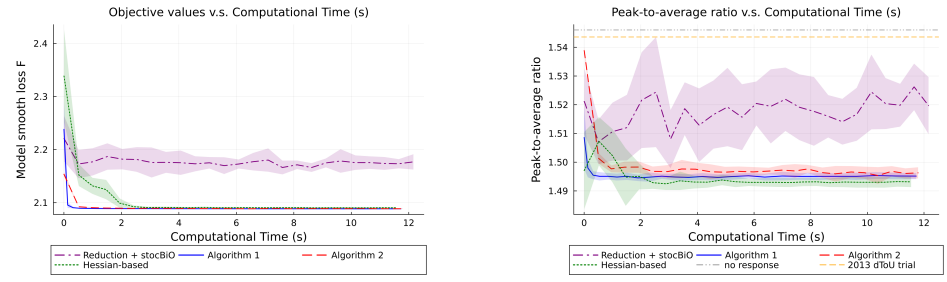


Figure 2: Contextual dynamic pricing for residential electricity demand response on LCL. Left: objective versus computational time. Right: peak-to-average ratio versus computational time. The horizontal dashed lines in the right panel mark the PAR under no response (gray, dash-dot-dot) and under the 2013 LCL dynamic time-of-use trial tariff (orange, dashed). The reduction+stocBiO uses the pruned 2-degree Chebyshev basis. Shaded regions show mean ± 1 standard deviation over 10 independent runs.

REPRODUCIBILITY STATEMENT

All theoretical claims in this paper are accompanied by full proofs, which are included in the Appendix, and are cited explicitly from the main text. The numerical experiments are fully reproducible: we provide the complete implementation (Julia code), all scripts for data preprocessing, training, and evaluation, as part of the supplementary materials. Any parameters, random seeds, hardware details, and dependencies used are documented in the supplementary material.

AI USE STATEMENT

We used LLMs during the preparation of this manuscript in limited, well-defined ways:

- We built the structure of the paper and wrote the core paragraphs ourselves. After that, we used LLMs to polish language, improve grammar, and enhance clarity and readability.
- In conducting the literature review, we used LLMs to help identify relevant papers we may originally have overlooked, to ensure thorough coverage.
- **No** theoretical results, proofs, algorithmic design, or experimental code were produced using LLMs; all substantive scientific contributions are our own.

We verified all content suggested by the LLMs. Any suggestions or drafts were carefully reviewed, edited, and corrected by us. We assume full responsibility for all content in this manuscript, including parts that were edited or polished via LLMs.

REFERENCES

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1–2): 165–214, 2023.

Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

Maxime Bouscary, Jiawei Zhang, and Saurabh Amin. Reducing contextual stochastic bilevel optimization via structured function approximation. In *International Conference on Learning Representations*, 2026.

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.

Lesi Chen, Yaohua Ma, and Jingzhao Zhang. Near-optimal nonconvex-strongly-convex bilevel optimization with fully first-order oracles. *Journal of Machine Learning Research*, 26(109):1–56, 2025.

Lesi Chen, Junru Li, El Mahdi Chayti, and Jingzhao Zhang. Faster gradient methods for highly-smooth stochastic bilevel optimization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Siyu Chen, Donglin Yang, Jiayang Li, Senmiao Wang, Zhuoran Yang, and Zhaoran Wang. Adaptive model design for Markov decision process. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3679–3700. PMLR, 2022a.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25294–25307, 2021.

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, 2022b.

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.

Tianshu Chu, Dachuan Xu, Wei Yao, and Jin Zhang. SPABA: A single-loop and probabilistic stochastic bilevel algorithm achieving optimal sample complexity. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8848–8903. PMLR, 2024.

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

Priya Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ahmad Faruqui and Sanem Sergici. Household Response to Dynamic Pricing of Electricity: A Survey of 15 Experiments. *Journal of Regulatory Economics*, 38(2):193–225, 2010.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Hongchang Gao, Bin Gu, and My T Thai. On the convergence of distributed stochastic bilevel optimization algorithms over a network. In *International conference on artificial intelligence and statistics*, pp. 9238–9281. PMLR, 2023.

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

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

Bowei Guo. The Spillover Effect of Peak Pricing. *Journal of Environmental Economics and Management*, 121:102845, 2023.

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.

Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. Unified convergence analysis for adaptive optimization with moving average estimator. *Machine Learning*, 114(4):86, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

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. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13550–13583. PMLR, 2023a.

Yifan Hu, Xin Chen, and Niao He. Sample complexity of sample average approximation for conditional stochastic optimization. *SIAM Journal on Optimization*, 30(3):2103–2133, 2020a.

Yifan Hu, Siqi Zhang, Xin Chen, and Niao He. Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2759–2770, 2020b.

Yifan Hu, Xin Chen, and Niao He. On the bias-variance-cost tradeoff of stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22119–22131, 2021.

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

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*, pp. 4882–4892. PMLR, 2021.

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, pp. 30271–30283, 2021.

András Kovács. Bilevel Programming Approach to Demand Response Management with Day-Ahead Tariff. *Journal of Modern Power Systems and Clean Energy*, 7(6):1632–1643, 2019.

András Kovács. Inverse Optimization Approach to the Identification of Electricity Consumer Models. *Central European Journal of Operations Research*, 29(2):521–537, 2021.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pp. 18083–18113. PMLR, 2023.

Jeongyeol Kwon, Dohyun Kwon, and Hanbaek Lyu. On the complexity of first-order methods in stochastic bilevel optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 25784–25811. PMLR, 2024a.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. In *International Conference on Learning Representations*, 2024b.

Zhaoyu Li, Yufan Zhang, and Qian Ai. Shape-Based Clustering for Demand Response Potential Evaluation: A Perspective of Comprehensive Evaluation Metrics. *Sustainable Energy, Grids and Networks*, 36:101213, 2023.

Vasilis Michalakopoulos, Elissaios Sarmas, Ioannis Papias, Panagiotis Skaloumpakas, Vangelis Marinakis, and Haris Doukas. A Machine Learning-Based Framework for Clustering Residential Electricity Load Profiles to Enhance Demand Response Programs. *Applied Energy*, 361:122943, 2024.

Mohammed Ali Moustafa. Tiny imagenet. Kaggle, 2017. URL https://www.kaggle.com/competitions/tiny-imagenet.

Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2710–2721, 2020.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2018.

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

Yves Rychener, Daniel Kuhn, and Tobias Sutter. End-to-end learning for stochastic optimization: A bayesian perspective. In *International Conference on Machine Learning*, pp. 29455–29472. PMLR, 2023.

Javier Saez-Gallego, Juan M. Morales, Marco Zugno, and Henrik Madsen. A Data-Driven Bidding Model for a Cluster of Price-Responsive Consumers of Electricity. *IEEE Transactions on Power Systems*, 31(6):5001–5011, 2016.

Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *International conference on machine learning*, pp. 9489–9502. PMLR, 2021.

Han Shen, Zhuoran Yang, and Tianyi Chen. Principled penalty-based methods for bilevel reinforcement learning and RLHF. *Journal of Machine Learning Research*, 26(114):1–49, 2025.

Mingyang Sun, Yi Wang, Fei Teng, Yujian Ye, Goran Strbac, and Chongqing Kang. Clustering-Based Residential Baseline Estimation: A Probabilistic Perspective. *IEEE Transactions on Smart Grid*, 10(6):6014–6028, 2019.

Vinzenz Thoma, Barna Pasztor, Andreas Krause, Giorgia Ramponi, and Yifan Hu. Contextual bilevel reinforcement learning for incentive alignment. In *Advances in Neural Information Processing Systems*, volume 37, pp. 127369–127435, 2024.

UK Power Networks. Smartmeter energy consumption data in london households. London Datastore, Greater London Authority, 2015. Available at https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households.

Yang Weng, Jiafan Yu, and Ram Rajagopal. Probabilistic Baseline Estimation Based on Load Patterns for Better Residential Customer Rewards. *International Journal of Electrical Power & Energy Systems*, 100:508–516, 2018.

Jincheng Yang, Luhao Zhang, Ningyuan Chen, Rui Gao, and Ming Hu. Decision making with side information: A causal transport robust approach. *Operations Research*, 74(4):2249–2269, 2026.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13670–13682, 2021.

Yan Yang, Bin Gao, and Ya-xiang Yuan. Bilevel reinforcement learning via the development of hyper-gradient without lower-level convexity. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4780–4788. PMLR, 2025.

Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. Hierarchically structured meta-learning. In *International conference on machine learning*, pp. 7045–7054. PMLR, 2019.

A Proofs of Main Results

A.1 Counting details

For the gradient estimator, each draw of $\eta \sim \mathbb{P}_{\eta|\xi}$ is counted as one conditional sample, while the same realized functions and partial gradients are reused to keep the count consistent with the analysis.

For the MLMC analysis, conditional on (x_k, ξ_k), define a master i.i.d. stream $\{\eta_k^t\}_{t\geq 0}$ from $\mathbb{P}_{\eta|\xi_k}$ and an independent outer sample η_k, all independent of ν_k. At every hypothetical level m, all inner loops start from the common initialization, use the first 2^m samples of this stream, and use the same outer sample for the hypergradient. The implementation evaluates only levels $0, \nu_k - 1, \nu_k$; the remaining levels are analytical coupling variables. Consequently, the telescoping identity (5) holds pathwise.

With $p_n = 2^{-n}/(1 - 2^{-N})$ on $[N]$,

$$\mathbb{E}[2^{\nu_k}] = \sum_{n=1}^{N} p_n 2^n = \frac{N}{1 - 2^{-N}}.$$

Thus one accelerated outer iteration uses one context draw and $1 + \mathbb{E}[2^{\nu_k}] = \mathcal{O}(N)$ conditional draws, with $\mathcal{O}(1)$ and $\mathcal{O}(N)$ upper- and lower-level gradient evaluations in expectation.

A.2 COMPLEXITY-COMPARISON DETAILS

All rates in this paper use $\mathbb{E}\|\nabla F(x)\|^2 \le \epsilon^2$. Rates stated under $\mathbb{E}\|\nabla F(x)\|^2 \le \epsilon$ or $\mathbb{E}\|\nabla F(x)\| \le \epsilon$ must be converted before their exponents are compared. For nonconvex–strongly-convex SBO, Chen et al. (2026) establish an $\Omega(\epsilon^{-4})$ lower bound by embedding the bounded-variance single-level hard instance of Arjevani et al. (2023) into a fully separable bilevel problem. They also analyze F^{2}SA-p with complexity $\widetilde{\mathcal{O}}(p\kappa^{9+2/p}\epsilon^{-4-2/p})$. The one-sided penalty quotient in our method has $\mathcal{O}(\lambda^{-1})$ bias by Lemma A.8, corresponding to first-order finite-difference accuracy.

Common-random-number access is not what separates the two settings: the oracle model of Arjevani et al. (2023) already permits multiple simultaneous queries sharing one random seed, and its statistical-learning extension exposes a sampled function that can be queried at several points. The relevant distinction is the uniform samplewise regularity imposed here. Writing $w = (x, y)$, Assumption 2.1(iii) implies, for every sample (η, ξ),

$$\|\nabla f(w;\eta,\xi)\| \le \ell_{f,0}, \qquad \|\nabla f(w;\eta,\xi) - \nabla f(w';\eta,\xi)\| \le \ell_{f,1}\|w - w'\|,$$

and hence also the conditional mean-squared-smoothness bound

$$\mathbb{E}\big[\|\nabla f(w;\eta,\xi) - \nabla f(w';\eta,\xi)\|^2 \mid \xi\big] \le \ell_{f,1}^2\|w - w'\|^2.$$

Here $\ell_{f,0}$ and $\ell_{f,1}$ are problem constants independent of ϵ. In contrast, the bounded-variance ϵ^{-4} hard instance of Arjevani et al. (2023) reveals a new chain coordinate with probability $p_{\mathrm{rev}} = \Theta(\epsilon^2/\sigma^2)$, where σ^2 is its variance budget, and uses a compensating $1/p_{\mathrm{rev}}$ factor to preserve unbiasedness. Consequently, its samplewise gradient bounds and samplewise smoothness constants cannot be bounded uniformly in ϵ; the statistical-learning extension retains the same amplification. Therefore, the Chen–Arjevani lower bound is an informative exponent-level benchmark, but it does not establish minimax optimality for our exact samplewise-regular CSBO class. A matching lower bound under these assumptions remains open.

A proof-level caveat for Kwon et al. (2024a) In the proof of their Lemma A.8, a pure-contraction step requires an inequality of the form

$$-2\gamma\Big\langle v_k^\star - v^{k,t}, \lambda^{-1}\nabla_y f(y^{k,t}) + \nabla_y g(y^{k,t}) - \nabla_y g(z^{k,t})\Big\rangle \le -\mu_g\gamma\|v_k^\star - v^{k,t}\|^2, \qquad \text{(K)}$$

where $y^{k,t}$ and $z^{k,t}$ are respectively the penalized and unpenalized iterates, $v^{k,t} = y^{k,t} - z^{k,t}$, and $v_k^\star = y_\lambda^\star - y^\star$. Strong convexity of g alone does not imply equation K. Indeed, consider the deterministic one-dimensional functions

$$g(y) = \tfrac{1}{2}y^2, \qquad f(y) = 1 - \cos y.$$

They satisfy the stated smoothness conditions, g is 1-strongly convex, and $|\nabla f(y)| \le 1$. For every $\lambda \ge 2$, both the unpenalized minimizer of g and the penalized minimizer of $g + \lambda^{-1}f$ equal zero. Set

$$y^{k,t} = -\frac{\pi}{2}, \qquad z^{k,t} = -\frac{\pi}{2} - \frac{1}{\lambda}.$$

Then $v^{k,t} = 1/\lambda$, $v_k^\star = 0$, while

$$\lambda^{-1}\nabla f(y^{k,t}) + \nabla g(y^{k,t}) - \nabla g(z^{k,t}) = -\frac{1}{\lambda} + \frac{1}{\lambda} = 0.$$

Consequently, the left-hand side of equation K is zero but its right-hand side is $-\gamma/\lambda^2 < 0$. Moreover, $|y^{k,t} - z^{k,t}| = 1/\lambda = \ell_{f,0}/(\mu_g\lambda)$ for $\ell_{f,0} = \mu_g = 1$, so the pair lies exactly at the stated trust-region radius rather than outside it.

The roles of the two inner variables must be reversed when translating to our notation: their penalized iterate is our $z_k^t(\lambda_m)$ and their unpenalized iterate is our y_k^t. Hence

$$\Delta_m^t = (z_k^t(\lambda_m) - z_m^\star) - (y_k^t - y_0^\star) = v^{k,t} - v_k^\star.$$

Our fixed-anchor argument in Lemma A.9 does not invoke the invalid pure contraction. It anchors the two drift fields at their respective minimizers and retains the remainder terms, yielding instead

$$\langle \Delta_m^t, \bar{A}_m^t\rangle \ge \frac{\mu_g}{4}\|\Delta_m^t\|^2 - C\lambda_m^{-2}\|e_m^t\|^2 - C\lambda_m^{-4}.$$

These anchor-error terms are the valid replacement needed to control the displacement between the two inner loops. The counterexample above concerns a key inequality in the published upper-bound proof; it does not establish that the complexity theorem itself is false.

A.3 METHODOLOGIES AND ROADMAP

The basic idea to construct a fully first-order algorithm for solving CSBO problems is to estimate ∇F using only first-order information of f and g, and then perform SGD for F.

We consider the following optimization problem with $\delta \in [0, \delta_{\max}]$, where $\delta_{\max} := \mu_g/(2\ell_{f,1})$:

$$\min_y \; Q(x, y, \delta; \xi) := \; \bar{g}(x, y; \xi) + \delta \bar{f}(x, y; \xi), \tag{9}$$

Let $y^\star(x, \delta; \xi)$ denote its solution and write $y_0^\star = y^\star(x, 0; \xi)$.

To do so, we first show in Lemma A.8 that ∇F can be approximated by $\mathbb{E}_{\eta,\xi}[\nabla_1 L(x, y^\star(x, \frac{1}{\lambda}; \xi), y^\star(x, 0; \xi), \lambda; \eta, \xi)]$:

$$\mathbb{E}_{\eta,\xi}[\nabla_1 L(x, y^\star(x, \frac{1}{\lambda}; \xi), y^\star(x, 0; \xi), \lambda; \eta, \xi)] \xrightarrow[\text{Lemma A.8}]{\text{Approximating}} \nabla F,$$

where L is defined in Section 2.2, $y^\star(x, \delta; \xi)$ is the solution to (9). Then the inner loop of our algorithms is applying a SGD-type manner to minimize $Q(x_k, y, 0; \xi)$ and $Q(x_k, y, \frac{1}{\lambda_k}; \xi)$ for y_{k+1} and z_{k+1} that approximate $y^*(x_k, 0; \xi)$ and $y^*(x_k, \frac{1}{\lambda_k}; \xi)$, respectively:

$$\text{Sample } \xi_k \sim \mathbb{P}_\xi \rightarrow \left.\begin{array}{l} \min_y \; Q(x_k, y, 0; \xi_k) \xrightarrow{\text{SGD}} y_{k+1} \\ \min_y \; Q(x_k, y, \frac{1}{\lambda_k}; \xi_k) \xrightarrow{\text{SGD}} z_{k+1} \end{array}\right\} \xrightarrow[\text{Lemmas A.6 and A.10}]{\text{Approximating}} \begin{array}{r} \mathbb{E}_{\eta,\xi}[\nabla_1 L(x, y^\star(x, \frac{1}{\lambda}; \xi), \\ y^\star(x, 0; \xi), \lambda; \eta, \xi)] \end{array}$$

Therefore, we use only the first-order information of f and g, and only SGD-type methods to approximate ∇F.

To further accelerate our Algorithm 1, we employ the multilevel Monte Carlo techniques, which, in addition to the previous framework, use extra approximations:

$$\text{Sample } \xi_k \sim \mathbb{P}_\xi, \text{Sample } \nu_k \sim \mathbb{P}_\nu \rightarrow \left.\begin{array}{l} \min_y \; Q(x_k, y, 0; \xi_k) \xrightarrow{\text{SGD}} y_k^{2^{\nu_k - 1}}, y_k^{2^{\nu_k}} \\ \min_y \; Q(x_k, y, \frac{1}{\lambda_{\nu_k}}; \xi_k) \xrightarrow{\text{SGD}} z_k^{2^{\nu_k}}(\lambda_{\nu_k}) \\ \min_y \; Q(x_k, y, \frac{1}{\lambda_{\nu_k - 1}}; \xi_k) \xrightarrow{\text{SGD}} z_k^{2^{\nu_k - 1}}(\lambda_{\nu_k - 1}) \end{array}\right\}$$

$$\xrightarrow{\text{Obtain}} G_k(0, \lambda_0) + p_{\nu_k}^{-1}[G_k(\nu_k, \lambda_{\nu_k}) - G_k(\nu_k - 1, \lambda_{\nu_k - 1})] \xrightarrow[(5)]{\text{Approximating}} \begin{array}{r} \mathbb{E}_{\eta,\xi}[\nabla_1 L(x_k, z_k^{2^N}(\lambda_N), \\ y_k^{2^N}, \lambda_N; \eta, \xi)] \end{array}$$

and

$$\mathbb{E}_{\eta,\xi}[\nabla_1 L(x_k, z_k^{2^N}(\lambda_N), y_k^{2^N}, \lambda_N; \eta, \xi) \mid \mathcal{F}_k] \xrightarrow[\text{Lemma A.11}]{\text{Approximating}}$$

$$\mathbb{E}_{\eta,\xi}[\nabla_1 L(x, y^\star(x, \frac{1}{\lambda}; \xi), y^\star(x, 0; \xi), \lambda; \eta, \xi)]$$

where $\mathbb{P}_N$ is the truncated geometric distribution whose upper bound is N defined in Algorithm 2 and $p_k \propto 2^{-\nu_k}$; λ_{ν_k}, $y_k^{2^{\nu_k - 1}}$, $y_k^{2^{\nu_k}}$, $z_k^{2^{\nu_k - 1}}(\lambda_{\nu_k - 1})$ and $z_k^{2^{\nu_k}}(\lambda_{\nu_k})$ are defined in Algorithm 2; $G_k(n, \lambda)$ is defined in (4).

A.4 USEFUL LEMMA

Lemma A.1. *(Nesterov, 2018, Lemma 1.2.3) If $g : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable on $\mathbb{R}^d$. The first derivative of g is Lipschitz continuous on $\mathbb{R}^d$ with constant $\ell_{g,1}$, then*

$$|g(y) - g(x) - \langle \nabla g(x), y - x \rangle| \leq \frac{\ell_{g,1}}{2} \|y - x\|^2.$$

Lemma A.2. *(Nesterov, 2018, Lemma 1.2.4) If $g : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable on $\mathbb{R}^d$. The second derivative of g is Lipschitz continuous on $\mathbb{R}^d$ with constant $\ell_{g,2}$, then*

$$\|\nabla g(y) - \nabla g(x) - \nabla^2 g(x)(y - x)\| \leq \frac{\ell_{g,2}}{2} \|y - x\|^2$$

$$|g(y) - g(x) - \langle \nabla g(x), y - x \rangle - \frac{1}{2} \langle \nabla^2 g(x)(y - x), y - x \rangle| \leq \frac{\ell_{g,2}}{6} \|y - x\|^3.$$

Similar to the proof of (Nesterov, 2018, Lemma 1.2.3), we have the following result:

Lemma A.3. *Suppose $G : \mathbb{R}^d \to \mathbb{R}^m$ is continuously differentiable, and $DG : \mathbb{R}^d \to \mathbb{R}^{m\times d}$ is Lipschitz continuous with modulus L in the following sense:*

$$\|DG(x) - DG(y)\|_2 \le L\|x - y\| \quad \forall x, y \in \mathbb{R}^d.$$

where $\|\cdot\|_2$ denotes the spectral norm of matrices. Then, for all $x, y \in \mathbb{R}^d$, it holds that

$$\|G(y) - G(x) - DG(x)(y - x)\| \le \frac{L}{2}\|x - y\|^2 \quad \forall x, y \in \mathbb{R}^d.$$

A.5 THE SMOOTHNESS OF F

Lemma A.4. *Under Assumption 2.1, there exists $\ell_{F,1} > 0$ such that*

$$\|\nabla F(x) - \nabla F(x')\| \le \ \ell_{F,1}\|x - x'\|.$$

Proof. Fix ξ and let $y_x = y^\star(x, 0; \xi)$. We use the operator norm for matrices. Conditional expectation preserves the uniform constants in Assumption 2.1; hence $\bar f(\cdot,\cdot;\xi)$ is $\ell_{f,0}$-Lipschitz, $\nabla\bar f(\cdot,\cdot;\xi)$ is $\ell_{f,1}$-Lipschitz, $\nabla\bar g(\cdot,\cdot;\xi)$ is $\ell_{g,1}$-Lipschitz, and $\nabla^2\bar g(\cdot,\cdot;\xi)$ is $\ell_{g,2}$-Lipschitz. In particular,

$$\|\nabla\bar f(x, y; \xi)\| \le \ell_{f,0}, \qquad \|\nabla^2\bar g(x, y; \xi)\| \le \ell_{g,1},$$

and $\nabla^2_{22}\bar g(x, y; \xi) \succeq \mu_g I$.

We first control the variation of the lower-level solution. Its first-order optimality condition is $\nabla_2\bar g(x, y_x; \xi) = 0$. Strong convexity in y, followed by the Lipschitz continuity of $\nabla_2\bar g(\cdot,\cdot;\xi)$, gives

$$\begin{aligned}\mu_g\|y_x - y_{x'}\|^2 &\le \big\langle \nabla_2\bar g(x', y_{x'}; \xi) - \nabla_2\bar g(x, y_{x'}; \xi), y_x - y_{x'}\big\rangle\\ &\le \ell_{g,1}\|x - x'\|\,\|y_x - y_{x'}\|.\end{aligned}$$

Therefore, with $c_y := \ell_{g,1}/\mu_g$,

$$\|y_x - y_{x'}\| \le c_y\|x - x'\|. \tag{10}$$

Set $G(x, y) := \nabla_2\bar g(x, y; \xi)$, so $G(x, y_x) = 0$. Since G is continuously differentiable and $D_yG(x, y_x) = \nabla^2_{22}\bar g(x, y_x; \xi) \succeq \mu_g I$ is nonsingular, the implicit function theorem and uniqueness imply that $x \mapsto y_x$ is continuously differentiable. Differentiating $G(x, y_x) = 0$ gives

$$0 = D_xG(x, y_x) + D_yG(x, y_x)D_1y_x = \nabla^2_{21}\bar g(x, y_x; \xi) + \nabla^2_{22}\bar g(x, y_x; \xi)D_1y_x.$$

Consequently,

$$D_1y_x = -\big(\nabla^2_{22}\bar g(x, y_x; \xi)\big)^{-1}\nabla^2_{21}\bar g(x, y_x; \xi).$$

Thus, for $\phi_\xi(x) := \bar f(x, y_x; \xi)$, set $w_x := (x, y_x)$ and $H_x := \nabla^2_{22}\bar g(w_x; \xi)$. The chain rule gives

$$\nabla\phi_\xi(x) = \nabla_1\bar f(w_x; \xi) - \nabla^2_{12}\bar g(w_x; \xi)H_x^{-1}\nabla_2\bar f(w_x; \xi). \tag{11}$$

It remains to show that the right-hand side of (11) is uniformly Lipschitz in x. By (10), $\|w_x - w_{x'}\| \le (1 + c_y)\|x - x'\|$. Moreover,

$$\begin{aligned}&\|\nabla_1\bar f(w_x; \xi) - \nabla_1\bar f(w_{x'}; \xi)\|,\ \|\nabla_2\bar f(w_x; \xi) - \nabla_2\bar f(w_{x'}; \xi)\| \le \ell_{f,1}\|w_x - w_{x'}\|,\\ &\|\nabla^2_{12}\bar g(w_x; \xi) - \nabla^2_{12}\bar g(w_{x'}; \xi)\|,\ \|H_x - H_{x'}\| \le \ell_{g,2}\|w_x - w_{x'}\|,\end{aligned}$$

while $\|\nabla^2_{12}\bar g(w_x; \xi)\| \le \ell_{g,1}$, $\|\nabla_2\bar f(w_x; \xi)\| \le \ell_{f,0}$, and $\|H_x^{-1}\| \le \mu_g^{-1}$. Using $H_x^{-1} - H_{x'}^{-1} = H_x^{-1}(H_{x'} - H_x)H_{x'}^{-1}$ and expanding the difference of $\nabla^2_{12}\bar g(w_x; \xi)H_x^{-1}\nabla_2\bar f(w_x; \xi)$, we obtain

$$\begin{aligned}\|\nabla\phi_\xi(x) - \nabla\phi_\xi(x')\| &\le \left(\ell_{f,1} + \frac{\ell_{f,0}\ell_{g,2}}{\mu_g} + \frac{\ell_{f,0}\ell_{g,1}\ell_{g,2}}{\mu_g^2} + \frac{\ell_{g,1}\ell_{f,1}}{\mu_g}\right)\|w_x - w_{x'}\|\\ &\le (1 + c_y)^2\left(\ell_{f,1} + \frac{\ell_{f,0}\ell_{g,2}}{\mu_g}\right)\|x - x'\|.\end{aligned}$$

The bound is independent of ξ. Also, $\|\nabla\phi_\xi(x)\| \le \ell_{f,0}(1 + c_y)$ by (11). We may therefore differentiate under the expectation and obtain $\nabla F(x) = \mathbb{E}_\xi[\nabla\phi_\xi(x)]$. The preceding estimate implies

$$\|\nabla F(x) - \nabla F(x')\| \le \mathbb{E}_\xi[\|\nabla\phi_\xi(x) - \nabla\phi_\xi(x')\|] \le \ell_{F,1}\|x - x'\|,$$

where one valid choice is $\ell_{F,1} = (1 + \ell_{g,1}/\mu_g)^2(\ell_{f,1} + \ell_{f,0}\ell_{g,2}/\mu_g)$. □

A.6 THE CONVERGENCE RATE OF INNER LOOP

The next lemma shows that $Q(x, y, \delta; \xi)$ is strongly convex provided δ is sufficiently small. This is useful to analyze the convergence rate of the inner loop.

Lemma A.5. *Under Assumption 2.1, if $\delta < \frac{\mu_g}{\ell_{f,1}}$, then for any $\xi \in \Xi$, $Q(x, y, \delta; \xi)$ in (9) is $(\mu_g - \delta\ell_{f,1})$-strongly convex in y.*

Proof. It follows from Assumption 2.1(iii) that

$$\bar{f}(x, z_2; \xi) - \bar{f}(x, z_1; \xi) \leq \langle \nabla_2 \bar{f}(x, z_2; \xi),\, z_2 - z_1 \rangle + \frac{\ell_{f,1}}{2}\|z_1 - z_2\|^2.$$

Since $\bar{g}(x, y; \xi)$ is μ_g-strongly convex in y, we have

$$\bar{g}(x, z_1; \xi) - \bar{g}(x, z_2; \xi) \geq \langle \nabla_2 \bar{g}(x, z_2; \xi),\, z_1 - z_2 \rangle + \frac{\mu_g}{2}\|z_1 - z_2\|^2.$$

Combining the above two inequalities, we get

$$\begin{aligned}
&\delta\bar{f}(x, z_1; \xi) - \delta\bar{f}(x, z_2; \xi) + \bar{g}(x, z_1; \xi) - \bar{g}(x, z_2; \xi) \\
&\geq \delta\langle \nabla_2 \bar{f}(x, z_2; \xi),\, z_1 - z_2 \rangle - \frac{\delta\ell_{f,1}}{2}\|z_1 - z_2\|^2 + \langle \nabla_2 \bar{g}(x, z_2; \xi),\, z_1 - z_2 \rangle + \frac{\mu_g}{2}\|z_1 - z_2\|^2 \\
&= \langle \nabla_2 Q(x, z_2, \delta; \xi),\, z_1 - z_2 \rangle + (\frac{\mu_g}{2} - \frac{\delta\ell_{f,1}}{2})\|z_1 - z_2\|^2.
\end{aligned}$$

This completes the proof. □

The next lemma gives a uniform sublinear convergence bound for either coupled inner loop. At iteration k, we sample ν_k (for Algorithm 2 only) and ξ_k independently of $\mathcal{F}_k$, and define

$$\mathcal{H}_k := \mathcal{F}_k \vee \sigma(\nu_k, \xi_k), \qquad \mathcal{G}_{k,t} := \mathcal{H}_k \vee \sigma(\eta_k^0, \ldots, \eta_k^{t-1}),$$

with the convention $\sigma(\nu_k, \xi_k) := \sigma(\xi_k)$ for Algorithm 1. The inner iterate y_k^t is $\mathcal{G}_{k,t}$-measurable, and η_k^t is conditionally independent of $\mathcal{G}_{k,t}$ given ξ_k, with conditional distribution $\mathbb{P}_{\eta|\xi=\xi_k}$. The outer sample η_k used in the gradient estimator is drawn independently of $\mathcal{G}_{k,T_k}$ given ξ_k, where T_k denotes the number of inner iterations at step k. For simplicity, we also define $\tau_f = \ell_{f,0}$.

Lemma A.6. *Suppose that Assumptions 2.1 and 2.2 hold. Fix the k-th outer iteration and any penalty $\lambda \geq 2\ell_{f,1}/\mu_g$ used by either algorithm. Let $z_k^t(\lambda)$ denote the corresponding penalized inner iterate (and write z_k^t in Algorithm 1). Then*

$$\mathbb{E}[\|y_k^t - y^\star(x_k, 0; \xi_k)\|^2 \mid \mathcal{H}_k] \leq \mathcal{O}\left(\frac{1}{t+1}\right),$$
$$\mathbb{E}[\|z_k^t(\lambda) - y^\star(x_k, \lambda^{-1}; \xi_k)\|^2 \mid \mathcal{H}_k] \leq \mathcal{O}\left(\frac{1}{t+1}\right).$$

Proof. Fix x_k, ξ_k, condition on $\mathcal{H}_k$, and put

$$\delta = \lambda^{-1}, \qquad \Phi_\delta(w; \eta) = g(x_k, w; \eta, \xi_k) + \delta f(x_k, w; \eta, \xi_k), \qquad \bar{\Phi}_\delta(w) = \mathbb{E}_{\eta|\xi_k}[\Phi_\delta(w; \eta)].$$

Let $w_t = z_k^t(\lambda)$ and $w^\star = y^\star(x_k, \delta; \xi_k)$. The z-update in either algorithm is equivalently

$$w_{t+1} = w_t - \beta_t \nabla\Phi_\delta(w_t; \eta_k^t),$$

because $\lambda^{-1}\nabla_z L = \nabla_2 g + \delta\nabla_2 f$. By the filtration definition above, w_t is $\mathcal{G}_{k,t}$-measurable and η_k^t, conditionally on $\mathcal{G}_{k,t}$, has law $\mathbb{P}_{\eta|\xi_k}$. Hence Assumption 2.1 gives

$$\mathbb{E}\big[\|\nabla\Phi_\delta(w_t; \eta_k^t) - \nabla\bar{\Phi}_\delta(w_t)\|^2 \mid \mathcal{G}_{k,t}\big] \leq 2\tau_g^2 + 2\delta^2\tau_f^2 =: \sigma_\delta^2.$$

Moreover, $\bar{\Phi}_\delta$ is $\mu_\delta := (\mu_g - \delta\ell_{f,1})$-strongly convex and $L_\delta := (\ell_{g,1} + \delta\ell_{f,1})$-smooth. Since $\delta\ell_{f,1} \leq \mu_g/2$ and $\mu_g \leq \ell_{g,1}$, we have $\mu_\delta \geq \mu_g/2$ and $L_\delta \leq 3\ell_{g,1}/2$. Using $\nabla\bar{\Phi}_\delta(w^\star) = 0$, strong monotonicity, and the conditional variance bound gives, for every $t \geq 0$,

$$\mathbb{E}[\|w_{t+1} - w^\star\|^2 \mid \mathcal{G}_{k,t}] \leq \big(1 - 2\mu_\delta\beta_t + 2L_\delta^2\beta_t^2\big)\|w_t - w^\star\|^2 + 2\sigma_\delta^2\beta_t^2.$$

The same inequality holds for the unpenalized y-loop with $\delta = 0$.

For either inner loop, let $(v_t, v^\star)$ denote either $(w_t, w^\star)$ for the penalized z-loop or $(y_k^t, y^\star(x_k, 0; \xi_k))$ for the unpenalized y-loop, and define its conditional mean-square error by

$$a_t := \mathbb{E}[\|v_t - v^\star\|^2 \mid \mathcal{H}_k].$$

We now treat explicitly the initial steps for which β_t need not be small. Define the fixed, problem-dependent burn-in index

$$t_{\mathrm{b}} := \left\lceil \frac{128\ell_{g,1}^2}{\mu_g^2} \right\rceil.$$

For every $t \geq t_{\mathrm{b}}$, the definition of t_{b} implies

$$\frac{8}{\mu_g(t+1)} \leq \frac{\mu_g}{16\ell_{g,1}^2}.$$

By Assumption 2.2 and strong convexity, $\|y_k^0 - y^\star(x_k, 0; \xi_k)\|^2 \leq 2b/\mu_g$. The common initialization gives $w_0 = y_k^0$, while the two first-order optimality conditions imply

$$\|w^\star - y^\star(x_k, 0; \xi_k)\| \leq \frac{\delta \ell_{f,0}}{\mu_\delta} \leq \frac{2\delta \ell_{f,0}}{\mu_g}.$$

Thus there is a deterministic constant A_0 that bounds both initial squared errors uniformly in k, ξ_k, and the admissible δ. Dropping the negative term in the preceding one-step bound and using the uniform bounds $L_\delta \leq 3\ell_{g,1}/2$ and $\sigma_\delta^2 \leq 2\tau_g^2 + 2(\mu_g/(2\ell_{f,1}))^2\tau_f^2 =: \sigma_{\max}^2$ yields

$$a_{t+1} \leq (1 + 2L_{\max}^2\beta_t^2)a_t + 2\sigma_{\max}^2\beta_t^2, \qquad 0 \leq t < t_{\mathrm{b}},$$

where $L_{\max} := 3\ell_{g,1}/2$. Iterating this bound over $0 \leq t < t_{\mathrm{b}}$ gives the explicit uniform bound

$$\max_{0 \leq t \leq t_{\mathrm{b}}} a_t \leq \left(A_0 + 2\sigma_{\max}^2 \sum_{s=0}^{t_{\mathrm{b}}-1} \beta_s^2 \right) \prod_{s=0}^{t_{\mathrm{b}}-1} (1 + 2L_{\max}^2\beta_s^2) = B < \infty.$$

We now spell out the contraction coefficient in the one-step bound. Since $\mu_\delta \geq \mu_g/2$, $L_\delta \leq 3\ell_{g,1}/2$, and $\beta_t \leq \mu_g/(16\ell_{g,1}^2)$,

$$\begin{aligned} 1 - 2\mu_\delta\beta_t + 2L_\delta^2\beta_t^2 &\leq 1 - \mu_g\beta_t + 2\left(\frac{3\ell_{g,1}}{2}\right)^2 \beta_t^2 \\ &\leq 1 - \mu_g\beta_t + 8\ell_{g,1}^2\beta_t^2 \\ &\leq 1 - \frac{\mu_g}{2}\beta_t = 1 - \frac{4}{t+1}. \end{aligned}$$

Moreover, $\sigma_\delta^2 \leq \sigma_{\max}^2$, and therefore

$$2\sigma_\delta^2\beta_t^2 \leq 2\sigma_{\max}^2 \left(\frac{8}{\mu_g(t+1)}\right)^2 = \frac{C_0}{(t+1)^2}, \qquad C_0 := \frac{128\sigma_{\max}^2}{\mu_g^2}.$$

Since $\mathcal{H}_k \subseteq \mathcal{G}_{k,t}$, taking conditional expectation of the one-step bound with respect to $\mathcal{H}_k$ and using the tower property gives

$$a_{t+1} \leq \left(1 - \frac{4}{t+1}\right) a_t + \frac{C_0}{(t+1)^2}, \qquad t \geq t_{\mathrm{b}}.$$

Thus C_0 depends only on the problem constants and is independent of k, ξ_k, λ, and t. Choose $D \geq \max\{(t_{\mathrm{b}}+1)B, C_0/3\}$.

We now use the finite-prefix estimate explicitly. Since $\max_{0 \leq t \leq t_{\mathrm{b}}} a_t \leq B$ and $D \geq (t_{\mathrm{b}}+1)B$, for every $0 \leq t \leq t_{\mathrm{b}}$,

$$a_t \leq B \leq \frac{D}{t_{\mathrm{b}}+1} \leq \frac{D}{t+1}.$$

In particular, $a_{t_\mathrm{b}} \le D/(t_\mathrm{b}+1)$, which is the base case for the tail induction. To verify the induction step, suppose that $a_t \le D/(t+1)$ for some $t \ge t_\mathrm{b}$ and set $s := t+1$. Since $\mu_g \le \ell_{g,1}$ implies $t_\mathrm{b} \ge 128$, the coefficient $1 - 4/s$ is nonnegative. The one-step bound and $D \ge C_0/3$ therefore give

$$a_{t+1} \le \left(1 - \frac{4}{s}\right)\frac{D}{s} + \frac{C_0}{s^2} = \frac{D}{s} - \frac{4D - C_0}{s^2} \le \frac{D}{s} - \frac{D}{s^2} \le \frac{D}{s+1} = \frac{D}{t+2},$$

where the penultimate inequality follows from $(s-1)/s^2 \le 1/(s+1)$. Thus induction proves $a_t \le D/(t+1)$ for every $t \ge t_\mathrm{b}$, while the preceding finite-prefix argument already proves the same estimate for $0 \le t \le t_\mathrm{b}$.

Applying this argument to both $\delta = 0$ and $\delta = \lambda^{-1}$ yields

$$\mathbb{E}[\|w_t - w^\star\|^2 \mid \mathcal{H}_k] + \mathbb{E}[\|y_k^t - y^\star(x_k, 0; \xi_k)\|^2 \mid \mathcal{H}_k] \le \frac{C}{t+1},$$

which proves both claims. □

A.7 Estimate bias

In this subsection, we shall show the bias of the gradient estimator of ∇F used in Algorithm 1 and Algorithm 2 is controllable. Specifically, we will show that $\|\nabla F(x_k) - \mathbb{E}[\nabla_1 L(x_k, z_{k+1}, y_{k+1}, \lambda_k; \eta_k, \xi_k) \mid \mathcal{F}_k]\|$ and $\|\nabla F(x_k) - \mathbb{E}_{\eta,\xi,\nu_k}[(G_k(0,\lambda_0) + p_{\nu_k}^{-1}(G_k(\nu_k, \lambda_{\nu_k}) - G_k(\nu_k - 1, \lambda_{\nu_k - 1}))) \mid \mathcal{F}_k]\|$ are upper bounded.

When δ in (9) is chosen such that $\delta < \frac{\mu_g}{\ell_{f,1}}$, by Lemma A.5, we know $Q(x, \cdot, \delta, \xi)$ is strongly convex, and hence the solution of (9) exists and is unique. The next Lemma shows that for any $\xi \in \Xi$, the solution to (9) $y^\star(x, \delta; \xi)$ is Lipschitz continuous in δ and x, respectively, provided δ is carefully selected.

Lemma A.7. *Under Assumption 2.1, let $\delta_{\max} := \mu_g/(2\ell_{f,1})$. There exists a constant $\ell_{y,0} > 0$ such that, for any $0 \le \delta' \le \delta \le \delta_{\max}$ and $\xi \in \Xi$,*

$$\begin{aligned}\|y^\star(x, \delta; \xi) - y^\star(x, \delta'; \xi)\| &\le \ell_{y,0}|\delta - \delta'|,\\ \|y^\star(x, 0; \xi) - y^\star(x', 0; \xi)\| &\le \ell_{y,0}\|x - x'\|,\end{aligned}$$

where $\ell_{y,0} := \max\{(\ell_{g,1} + \delta_{\max}\ell_{f,1})/(\mu_g - \delta_{\max}\ell_{f,1}), \ell_{f,0}/(\mu_g - \delta_{\max}\ell_{f,1})\}$.

Proof. By the definition of $y^\star(x, \delta; \xi)$ and the first-order optimality condition,

$$\nabla_2 \bar{g}(x, y^\star(x, \delta; \xi); \xi) + \delta \nabla_2 \bar{f}(x, y^\star(x, \delta; \xi); \xi) = 0,$$

Differentiating with respect to x and δ gives

$$\begin{aligned}&\nabla_{21}^2 Q(x, y^\star(x, \delta; \xi), \delta; \xi) + \nabla_{22}^2 Q(x, y^\star(x, \delta; \xi), \delta; \xi) D_1 y^\star(x, \delta; \xi) = 0,\\ &\nabla_2 \bar{f}(x, y^\star(x, \delta; \xi); \xi) + \nabla_{22}^2 Q(x, y^\star(x, \delta; \xi), \delta; \xi) D_\delta y^\star(x, \delta; \xi) = 0.\end{aligned}$$

By Lemma A.5,

$$\begin{aligned}D_1 y^\star(x, \delta; \xi) &= -\left[\nabla_{22}^2 Q(x, y^\star(x, \delta; \xi), \delta; \xi)\right]^{-1} \nabla_{21}^2 Q(x, y^\star(x, \delta; \xi), \delta; \xi),\\ D_\delta y^\star(x, \delta; \xi) &= -\left[\nabla_{22}^2 Q(x, y^\star(x, \delta; \xi), \delta; \xi)\right]^{-1} \nabla_2 \bar{f}(x, y^\star(x, \delta; \xi); \xi),\end{aligned} \tag{12}$$

where Q is defined by (9). Consequently, for every $\delta \in [0, \delta_{\max}]$,

$$\|D_1 y^\star(x, \delta; \xi)\| \le \frac{\ell_{g,1} + \delta_{\max}\ell_{f,1}}{\mu_g - \delta_{\max}\ell_{f,1}}, \qquad \|D_\delta y^\star(x, \delta; \xi)\| \le \frac{\ell_{f,0}}{\mu_g - \delta_{\max}\ell_{f,1}}.$$

The two Lipschitz bounds follow from the mean-value theorem. □

The following Lemma shows that $\nabla F(x)$ can be approximated using only first-order information of L, which plays a crucial role in our analysis.

Lemma A.8. *Suppose that Assumption 2.1 holds, and $\lambda \geq \frac{2\ell_{f,1}}{\mu_g}$. Let the solution to* (9) *be $y^\star(x, \delta; \xi)$. Then we have*

$$\|\nabla F(x) - \mathbb{E}_{\eta,\xi}[\nabla_1 L(x, y^\star(x, \frac{1}{\lambda}; \xi), y^\star(x, 0; \xi), \lambda; \eta, \xi)]\| = \mathcal{O}(\frac{1}{\lambda}).$$

Proof. By (12), we know that

$$D_1 y^\star(x, 0; \xi) = -\nabla^2_{22}\bar{g}(x, y^\star(x, 0; \xi); \xi)^{-1}\nabla^2_{21}\bar{g}(x, y^\star(x, 0; \xi); \xi)$$

The above equality and the chain rule imply

$$\begin{aligned}\nabla F(x) &= \mathbb{E}_\xi[\nabla_1 \bar{f}(x, y^\star(x, 0; \xi); \xi) + D_1 y^\star(x, 0; \xi)^\top \nabla_2 \bar{f}(x, y^\star(x, 0; \xi); \xi)] \\ &= \mathbb{E}_\xi[\nabla_1 \bar{f}(x, y^\star(x, 0; \xi); \xi) \\ &\quad - \nabla^2_{12}\bar{g}(x, y^\star(x, 0; \xi); \xi)(\nabla^2_{22}\bar{g}(x, y^\star(x, 0; \xi); \xi))^{-1}\nabla_2 \bar{f}(x, y^\star(x, 0; \xi); \xi)],\end{aligned} \tag{13}$$

It follows from the definition of L that

$$\begin{aligned}&\mathbb{E}_{\eta,\xi}[\nabla_1 L(x, y^\star(x, \frac{1}{\lambda}; \xi), y^\star(x, 0; \xi), \lambda; \eta, \xi)] \\ &= \mathbb{E}_\xi\left[\nabla_1 \bar{f}(x, y^\star(x, \frac{1}{\lambda}; \xi); \xi) + \lambda\left(\nabla_1 \bar{g}(x, y^\star(x, \frac{1}{\lambda}; \xi); \xi) - \nabla_1 \bar{g}(x, y^\star(x, 0; \xi); \xi)\right)\right]\end{aligned} \tag{14}$$

Put $\delta = 1/\lambda$, $y_\delta = y^\star(x, \delta; \xi)$, $y_0 = y^\star(x, 0; \xi)$, $H_0 = \nabla^2_{22}\bar{g}(x, y_0; \xi)$, and $b_0 = \nabla_2 \bar{f}(x, y_0; \xi)$. By Lemma A.7, $\|y_\delta - y_0\| = \mathcal{O}(\delta)$. The two first-order optimality conditions give

$$0 = \nabla_2 \bar{g}(x, y_\delta; \xi) + \delta \nabla_2 \bar{f}(x, y_\delta; \xi), \qquad 0 = \nabla_2 \bar{g}(x, y_0; \xi).$$

The Lipschitz continuity of $\nabla^2 \bar{g}$ and $\nabla \bar{f}$ therefore yields

$$H_0(y_\delta - y_0) + \delta b_0 = -(r_g + \delta r_f), \qquad \|r_g\| \leq \frac{\ell_{g,2}}{2}\|y_\delta - y_0\|^2, \quad \|r_f\| \leq \ell_{f,1}\|y_\delta - y_0\|.$$

Since $H_0 \succeq \mu_g I$, this proves, without requiring a Lipschitz assumption on $D_\delta y^\star$, that

$$y_\delta - y_0 = -\delta H_0^{-1} b_0 + r_y, \qquad \|r_y\| = \mathcal{O}(\delta^2). \tag{15}$$

All constants in this display are uniform in ξ under Assumption 2.1. Applying Lemma A.3 only to the map $y \mapsto \nabla_1 \bar{g}(x, y; \xi)$, whose Jacobian is $\nabla^2_{12}\bar{g}(x, y; \xi)$, gives

$$\nabla_1 \bar{g}(x, y_\delta; \xi) - \nabla_1 \bar{g}(x, y_0; \xi) = \nabla^2_{12}\bar{g}(x, y_0; \xi)(y_\delta - y_0) + r_1^g = -\delta \nabla^2_{12}\bar{g}(x, y_0; \xi) H_0^{-1} b_0 + r_3^g,$$

where $\|r_1^g\| \leq (\ell_{g,2}/2)\|y_\delta - y_0\|^2$ and $\|r_3^g\| = \mathcal{O}(\delta^2) = \mathcal{O}(\lambda^{-2})$. It follows from (13), (14) and the above equality that

$$\begin{aligned}&\nabla F(x) - \mathbb{E}_{\eta,\xi}[\nabla_1 L(x, y^\star(x, \frac{1}{\lambda}; \xi), y^\star(x, 0; \xi), \lambda; \eta, \xi)] \\ &= \mathbb{E}_\xi[\nabla_1 \bar{f}(x, y^\star(x, 0; \xi); \xi) - \nabla_1 \bar{f}(x, y^\star(x, \frac{1}{\lambda}; \xi); \xi)] + r_4^g,\end{aligned}$$

where $\|r_4^g\| \leq \lambda \mathbb{E}_\xi[\|r_3^g\|] = \mathcal{O}(1/\lambda)$. Combining the Lipschitz continuity of $\nabla_1 \bar{f}$ with Lemma A.7 proves the claim.

□

Lemma A.9 (Fixed-anchor displacement)**.** *Under Assumptions 2.1 and 2.2, fix the k-th outer iteration and condition on $\mathcal{H}_k$. For any $\lambda \geq 2\ell_{f,1}/\mu_g$, consider the two coupled inner loops driven by the same samples and initialized from the common point y_k^0. Both inner loops use the clipped stepsize*

$$\beta_t = \min\left\{\frac{8}{\mu_g(t+1)}, \frac{\mu_g}{16\ell_{g,1}^2}\right\},$$

employed by both Algorithm 1 and Algorithm 2. Write $z_k^t(\lambda)$ for the iterates of the penalized inner loop and define

$$\delta := \lambda^{-1}, \qquad y_0^\star := y^\star(x_k, 0; \xi_k), \qquad z_\lambda^\star := y^\star(x_k, \delta; \xi_k),$$
$$e^t := y_k^t - y_0^\star, \qquad \Delta_\lambda^t := (z_k^t(\lambda) - z_\lambda^\star) - e^t.$$

There is a deterministic constant $C > 0$, independent of k, ξ_k, λ, t, such that, for every integer $t \geq 0$,

$$\mathbb{E}[\|\Delta_\lambda^t\|^2 \mid \mathcal{H}_k] \leq C\lambda^{-4} + \frac{C\lambda^{-2}}{t+1}. \tag{16}$$

In particular, for $\lambda_m = (2\ell_{f,1}/\mu_g)(2^m)^{1/2}$, abbreviate $z_m^\star := z_{\lambda_m}^\star$, $e_m^t := e^t$, and $\Delta_m^t := \Delta_{\lambda_m}^t$. Then

$$\mathbb{E}[\|\Delta_m^{2^m}\|^2 \mid \mathcal{H}_k] = \mathcal{O}(\lambda_m^{-4}), \qquad \mathbb{E}[\|\Delta_m^{2^m}\|^2 \mid \mathcal{F}_k] = \mathcal{O}(\lambda_m^{-4}),$$

where the second conclusion follows from the tower property.

Proof. Conditionally on $\mathcal{H}_k$, the quantities x_k, ξ_k, and λ are fixed. Put

$$\Phi_0(y;\eta) := g(x_k, y; \eta, \xi_k), \qquad \Phi_\delta(z;\eta) := g(x_k, z; \eta, \xi_k) + \delta f(x_k, z; \eta, \xi_k),$$

where all gradients in this proof are taken with respect to the lower-level variable. Define

$$\begin{aligned} A_\lambda^t &:= \nabla\Phi_\delta(z_k^t(\lambda); \eta_k^t) - \nabla\Phi_\delta(z_\lambda^\star; \eta_k^t) - \nabla\Phi_0(y_k^t; \eta_k^t) + \nabla\Phi_0(y_0^\star; \eta_k^t), \\ \zeta_\lambda^t &:= \nabla\Phi_\delta(z_\lambda^\star; \eta_k^t) - \nabla\Phi_0(y_0^\star; \eta_k^t). \end{aligned}$$

The two first-order optimality conditions and the conditional sampling rule imply $\mathbb{E}[\zeta_\lambda^t \mid \mathcal{G}_{k,t}] = 0$. Subtracting the two inner updates gives

$$\Delta_\lambda^{t+1} = \Delta_\lambda^t - \beta_t(A_\lambda^t + \zeta_\lambda^t).$$

Let $\bar{\Phi}_\delta := \mathbb{E}_{\eta|\xi_k}[\Phi_\delta]$, $\bar{A}_\lambda^t := \mathbb{E}[A_\lambda^t \mid \mathcal{G}_{k,t}]$, and $d_\lambda^\star := z_\lambda^\star - y_0^\star$. The mean-value theorem gives the exact decomposition

$$\begin{aligned} \bar{A}_\lambda^t &= H_\lambda^t \Delta_\lambda^t + r_\lambda^t, \\ H_\lambda^t &:= \int_0^1 \nabla^2\bar{\Phi}_\delta(z_\lambda^\star + e^t + s\Delta_\lambda^t)\, ds, \\ r_\lambda^t &:= \nabla\bar{\Phi}_\delta(z_\lambda^\star + e^t) - \nabla\bar{\Phi}_\delta(z_\lambda^\star) - \nabla\bar{\Phi}_0(y_0^\star + e^t) + \nabla\bar{\Phi}_0(y_0^\star). \end{aligned}$$

Since $\delta\ell_{f,1} \leq \mu_g/2$, we have $H_\lambda^t \succeq (\mu_g/2)I$. Moreover,

$$r_\lambda^t = \int_0^1 \Big(\nabla^2\bar{\Phi}_0(y_0^\star + d_\lambda^\star + se^t) - \nabla^2\bar{\Phi}_0(y_0^\star + se^t)\Big)e^t\, ds + \delta\big(\nabla\bar{f}(z_\lambda^\star + e^t) - \nabla\bar{f}(z_\lambda^\star)\big).$$

By Lemma A.7, $\|d_\lambda^\star\| \leq C\lambda^{-1}$; the Lipschitz assumptions therefore give $\|r_\lambda^t\| \leq C\lambda^{-1}\|e^t\|$. Hence

$$\langle \Delta_\lambda^t, \bar{A}_\lambda^t \rangle \geq \frac{\mu_g}{4}\|\Delta_\lambda^t\|^2 - C\lambda^{-2}\|e^t\|^2 - C\lambda^{-4}. \tag{17}$$

Here the last term is a harmless weakening of the bound obtained directly from Young's inequality. This exact decomposition contains no cubic term in $\|\Delta_\lambda^t\|$.

Furthermore,

$$A_\lambda^t + \zeta_\lambda^t = \nabla\Phi_\delta(z_k^t(\lambda); \eta_k^t) - \nabla\Phi_0(y_k^t; \eta_k^t).$$

Because $z_k^t(\lambda) - y_k^t = \Delta_\lambda^t + d_\lambda^\star$ and $\|\nabla_2 f\| \leq \ell_{f,0}$, the samplewise Lipschitz bound on $\nabla_2 g$ yields the explicit estimate

$$\mathbb{E}[\|A_\lambda^t + \zeta_\lambda^t\|^2 \mid \mathcal{G}_{k,t}] \leq 4\ell_{g,1}^2\|\Delta_\lambda^t\|^2 + C\lambda^{-2}.$$

Taking squared norms in the displacement update and using the preceding two displays gives

$$\begin{aligned} D_{t+1} &\leq \Big(1 - \frac{\mu_g}{2}\beta_t + 4\ell_{g,1}^2\beta_t^2\Big) D_t + C\beta_t\lambda^{-2}E_t + C\beta_t\lambda^{-4} + C\beta_t^2\lambda^{-2}, \\ D_t &:= \mathbb{E}[\|\Delta_\lambda^t\|^2 \mid \mathcal{H}_k], \quad E_t := \mathbb{E}[\|e^t\|^2 \mid \mathcal{H}_k]. \end{aligned}$$

We next separate the clipped and harmonic regimes of the schedule. Set

$$t_{\mathrm{b}} := \left\lceil \frac{128\ell_{g,1}^2}{\mu_g^2} \right\rceil.$$

The clipping gives $\beta_t \le \mu_g/(16\ell_{g,1}^2)$ at every iteration, so $4\ell_{g,1}^2\beta_t^2 \le (\mu_g/4)\beta_t$ for every t. On the finite prefix $0 \le t < t_{\mathrm{b}}$, the schedule also satisfies $\beta_t \le 8/\mu_g$. Dropping the negative drift term in the last display, using $E_t \le C$ from Lemma A.6, and using $\lambda^{-4} \le C\lambda^{-2}$ for admissible λ gives

$$D_{t+1} \le (1 + C\beta_t^2)D_t + C(\beta_t + \beta_t^2)\lambda^{-2}, \qquad 0 \le t < t_{\mathrm{b}}.$$

The common initialization and Lemma A.7 give $D_0 = \|y_0^\star - z_\lambda^\star\|^2 \le C\lambda^{-2}$. Iteration of this bound over $0 \le t < t_{\mathrm{b}}$ thus yields a constant C_{b}, independent of k, ξ_k, λ, t, such that

$$\max_{0 \le t \le t_{\mathrm{b}}} D_t \le C_{\mathrm{b}}\lambda^{-2}.$$

For $t \ge t_{\mathrm{b}}$, the schedule reduces to $\beta_t = 8/(\mu_g(t+1))$. Therefore, by Lemma A.6, $E_t \le C/(t+1)$, and the one-step bound becomes

$$D_{t+1} \le \left(1 - \frac{2}{t+1}\right) D_t + \frac{a\lambda^{-4}}{t+1} + \frac{b\lambda^{-2}}{(t+1)^2}, \tag{18}$$

for fixed constants $a, b > 0$. To spell out the induction, choose $A \ge a/2$ and $B \ge \max\{C_{\mathrm{b}}(t_{\mathrm{b}} + 1), b\}$. If $D_t \le A\lambda^{-4} + B\lambda^{-2}/(t+1)$, then (18) gives, with $s = t+1$,

$$D_{t+1} \le A\lambda^{-4} + \frac{B\lambda^{-2}}{s} - \frac{(2B-b)\lambda^{-2}}{s^2} \le A\lambda^{-4} + \frac{B\lambda^{-2}}{s+1}.$$

The choice of B makes the estimate valid at $t = t_{\mathrm{b}}$; the finite-prefix bound covers $0 \le t < t_{\mathrm{b}}$ after enlarging C. This proves (16) for arbitrary admissible λ and every $t \ge 0$.

□

We next control the error caused by the coupled inner loop of Algorithm 1.

Lemma A.10. *Suppose that Assumptions 2.1 and 2.2 hold, and consider Algorithm 1 with* $T_k = \lceil C_T\lambda_k^2 \log(\lambda_k + 1)\rceil$ *for a sufficiently large constant* C_T. *With*

$$y_{0,k}^\star := y^\star(x_k, 0; \xi_k), \quad z_k^\star := y^\star(x_k, \lambda_k^{-1}; \xi_k),$$
$$D_k := \nabla_1 L(x_k, z_{k+1}, y_{k+1}, \lambda_k; \eta_k, \xi_k) - \nabla_1 L(x_k, z_k^\star, y_{0,k}^\star, \lambda_k; \eta_k, \xi_k),$$

we have

$$\big\|\mathbb{E}[D_k \mid \mathcal{F}_k]\big\|^2 \le \mathcal{O}(\lambda_k^{-2}).$$

Proof. For Algorithm 1, $\mathcal{H}_k = \mathcal{F}_k \vee \sigma(\xi_k)$. Condition on $\mathcal{H}_k$ and define

$$e_t := y_k^t - y_{0,k}^\star, \qquad \Delta_t := (z_k^t - z_k^\star) - e_t.$$

For clarity, the fixed-anchor estimate invoked below is based on an exact mean-value decomposition, not on a Taylor remainder containing $\|\Delta_t\|^2$. Put $\delta = \lambda_k^{-1}$, $\bar\Phi_\delta(w) = \bar g(x_k, w; \xi_k) + \delta\bar f(x_k, w; \xi_k)$, $\bar\Phi_0(w) = \bar g(x_k, w; \xi_k)$, and

$$\bar A_t := \nabla\bar\Phi_\delta(z_k^t) - \nabla\bar\Phi_\delta(z_k^\star) - \nabla\bar\Phi_0(y_k^t) + \nabla\bar\Phi_0(y_{0,k}^\star).$$

Define the exact mean-value terms

$$H_t := \int_0^1 \nabla^2\bar\Phi_\delta(z_k^\star + e_t + s\Delta_t)\, ds,$$
$$r_t := \nabla\bar\Phi_\delta(z_k^\star + e_t) - \nabla\bar\Phi_\delta(z_k^\star) - \nabla\bar\Phi_0(y_{0,k}^\star + e_t) + \nabla\bar\Phi_0(y_{0,k}^\star).$$

Then, as proved in Lemma A.9,

$$\bar A_t = H_t\Delta_t + r_t, \qquad H_t \succeq \frac{\mu_g}{2} I, \qquad \|r_t\| \le C\lambda_k^{-1}\|e_t\|.$$

Consequently,

$$\langle \Delta_t, \bar A_t\rangle \ge \frac{\mu_g}{4}\|\Delta_t\|^2 - C\lambda_k^{-2}\|e_t\|^2 - C\lambda_k^{-4},$$

so no cubic term in $\|\Delta_t\|$ is used or needs to be absorbed.

The fixed-anchor estimate in Lemma A.9, applied with $\lambda = \lambda_k$ and $t = T_k$, gives

$$\mathbb{E}[\|\Delta_{T_k}\|^2 \mid \mathcal{H}_k] \le C\lambda_k^{-4} + \frac{C\lambda_k^{-2}}{T_k + 1} = \mathcal{O}(\lambda_k^{-4}).$$

Moreover, Lemma A.6 gives

$$\mathbb{E}[\|e_{T_k}\|^2 \mid \mathcal{H}_k] = \mathcal{O}((T_k + 1)^{-1}) = \mathcal{O}(\lambda_k^{-2}).$$

Since $z_{k+1} - z_k^\star = e_{T_k} + \Delta_{T_k}$, it follows that

$$\mathbb{E}[\|z_{k+1} - z_k^\star\|^2 \mid \mathcal{H}_k] = \mathcal{O}(\lambda_k^{-2}).$$

Using the definition of L, the mean-value theorem, the samplewise Lipschitz bounds, and $\|z_k^\star - y_{0,k}^\star\| = \mathcal{O}(\lambda_k^{-1})$, we obtain

$$\mathbb{E}[\|D_k\|^2 \mid \mathcal{H}_k] \le C\mathbb{E}[\|z_{k+1} - z_k^\star\|^2 \mid \mathcal{H}_k] + C\lambda_k^2\mathbb{E}[\|\Delta_{T_k}\|^2 \mid \mathcal{H}_k] + C\mathbb{E}[\|e_{T_k}\|^2 \mid \mathcal{H}_k] = \mathcal{O}(\lambda_k^{-2}).$$

Finally, conditional Jensen's inequality and the tower property yield

$$\|\mathbb{E}[D_k \mid \mathcal{F}_k]\|^2 \le \mathbb{E}[\|D_k\|^2 \mid \mathcal{F}_k] = \mathbb{E}\big[\mathbb{E}[\|D_k\|^2 \mid \mathcal{H}_k] \mid \mathcal{F}_k\big] \le \mathcal{O}(\lambda_k^{-2}),$$

which proves the claim.

□

The same conditional argument controls the terminal RT–MLMC level.

Lemma A.11. *Suppose that Assumptions 2.1 and 2.2 hold and consider Algorithm 2. Then*

$$\Big\|\mathbb{E}\big[\nabla_1 L(x_k, z_k^{2^N}(\lambda_N), y_k^{2^N}, \lambda_N; \eta_k, \xi_k) - \nabla_1 L(x_k, y^\star(x_k, \lambda_N^{-1}; \xi_k), y^\star(x_k, 0; \xi_k), \lambda_N; \eta_k, \xi_k) \mid \mathcal{F}_k\big]\Big\|^2 \le \mathcal{O}(\lambda_N^{-2}).$$

Proof. Condition on $\mathcal{H}_k$ and use the notation of Lemma A.9 with $\lambda = \lambda_N$. That lemma and Lemma A.6 give

$$\mathbb{E}[\|\Delta_{\lambda_N}^{2^N}\|^2 \mid \mathcal{H}_k] = \mathcal{O}(\lambda_N^{-4}), \qquad \mathbb{E}[\|e^{2^N}\|^2 \mid \mathcal{H}_k] = \mathcal{O}(\lambda_N^{-2}).$$

Let $D_{k,N}$ denote the gradient difference inside the conditional expectation in the statement and put $r_{k,N} := z_k^{2^N}(\lambda_N) - z_{\lambda_N}^\star$. The same mean-value estimate as in Lemma A.10 yields

$$\mathbb{E}[\|D_{k,N}\|^2 \mid \mathcal{H}_k] \le C\mathbb{E}[\|r_{k,N}\|^2 \mid \mathcal{H}_k] + C\lambda_N^2\mathbb{E}[\|\Delta_{\lambda_N}^{2^N}\|^2 \mid \mathcal{H}_k] + C\mathbb{E}[\|e^{2^N}\|^2 \mid \mathcal{H}_k] = \mathcal{O}(\lambda_N^{-2}).$$

Conditional Jensen's inequality followed by the tower property from $\mathcal{H}_k$ to $\mathcal{F}_k$ proves the claim.

□

Now, combining all lemmas in this subsection, we can upper bound $\|\nabla F(x_k) - \mathbb{E}[\nabla_1 L(x_k, z_{k+1}, y_{k+1}, \lambda_k; \eta_k, \xi_k) \mid \mathcal{F}_k]\|$ and $\|\nabla F(x_k) - \mathbb{E}_{\eta,\xi,\nu_k}[(G_k(0, \lambda_0) + p_{\nu_k}^{-1}(G_k(\nu_k, \lambda_{\nu_k}) - G_k(\nu_k - 1, \lambda_{\nu_k - 1}))) \mid \mathcal{F}_k]\|$ using triangle inequality. Then the bias of gradient estimator is controllable. The results are summarized in the following two lemmas.

Lemma A.12. *Under Assumptions 2.1 and 2.2, consider Algorithm 1 with $T_k = \lceil C_T \lambda_k^2 \log(\lambda_k + 1)\rceil$, we have*

$$\|\nabla F(x_k) - \mathbb{E}[\nabla_1 L(x_k, z_{k+1}, y_{k+1}, \lambda_k; \eta_k, \xi_k) \mid \mathcal{F}_k]\| \le \mathcal{O}(\frac{1}{\lambda_k}).$$

Proof. Let $G_k^\star := \nabla_1 L(x_k, z_k^\star, y_{0,k}^\star, \lambda_k; \eta_k, \xi_k)$, with $z_k^\star$ and $y_{0,k}^\star$ as in Lemma A.10. By Lemma A.8, $\|\nabla F(x_k) - \mathbb{E}[G_k^\star \mid \mathcal{F}_k]\| = \mathcal{O}(\lambda_k^{-1})$. Moreover, Lemma A.10 and the definition of D_k give $\|\mathbb{E}[D_k \mid \mathcal{F}_k]\| = \mathcal{O}(\lambda_k^{-1})$. The conditional triangle inequality proves the claim. □

Lemma A.13. *Under Assumptions 2.1 and 2.2, consider Algorithm 2, we have*

$$\|\mathbb{E}[G_k(0,\lambda_0) + p_{\nu_k}^{-1}(G_k(\nu_k,\lambda_{\nu_k}) - G_k(\nu_k-1,\lambda_{\nu_k-1})) \mid \mathcal{F}_k] - \nabla F(x_k)\|^2 \leq \mathcal{O}(\frac{1}{\lambda_N^2}).$$

Proof. By (5), we obtain

$$\begin{aligned}&\mathbb{E}[G_k(0,\lambda_0) + p_{\nu_k}^{-1}(G_k(\nu_k,\lambda_{\nu_k}) - G_k(\nu_k-1,\lambda_{\nu_k-1})) \mid \mathcal{F}_k]\\ =\ &\mathbb{E}[\nabla_1 L(x_k, z_k^{2^N}(\lambda_N), y_k^{2^N}, \lambda_N; \eta_k, \xi_k) \mid \mathcal{F}_k].\end{aligned}$$

Then the desired result is due to Lemma A.8, Lemma A.11 and the above equality. □

A.8 The variance of RT-MLMC

The following lemma estimates the level difference $\mathbb{E}[\|G_k(n,\lambda_n) - G_k(n-1,\lambda_{n-1})\|^2 \mid \mathcal{F}_k]$, which will be used in the RT-MLMC variance bound.

Lemma A.14. *Under Assumptions 2.1 and 2.2, consider Algorithm 2 with $\lambda_n = \frac{2\ell_{f,1}}{\mu_g}(2^n)^{1/2}$. We have*

$$\mathbb{E}[\|G_k(n,\lambda_n) - G_k(n-1,\lambda_{n-1})\|^2 \mid \mathcal{F}_k] \leq \mathcal{O}(\frac{1}{2^{n-1}}).$$

Proof. For any λ, denote the ideal fixed-anchor hypergradient by

$$\Gamma_k(\lambda;\eta_k,\xi_k) := \nabla_1 L(x_k, y^\star(x_k,\frac{1}{\lambda};\xi_k), y^\star(x_k,0;\xi_k), \lambda;\eta_k,\xi_k).$$

We first compare each implemented level with this ideal fixed-anchor quantity. For $m \in \{n, n-1\}$, use the notation of Lemma A.9 and define

$$\begin{aligned}B_m :=\ &\nabla_1 g(x_k, z_k^{2^m}(\lambda_m);\eta_k,\xi_k) - \nabla_1 g(x_k, z_m^\star;\eta_k,\xi_k)\\ &- \nabla_1 g(x_k, y_k^{2^m};\eta_k,\xi_k) + \nabla_1 g(x_k, y_0^\star;\eta_k,\xi_k).\end{aligned}$$

Writing $d_m^\star = z_m^\star - y_0^\star$, the mean-value theorem gives

$$\begin{aligned}B_m =\ &\int_0^1 \nabla_{12}^2 g(x_k, z_m^\star + e_m^{2^m} + s\Delta_m^{2^m};\eta_k,\xi_k)\Delta_m^{2^m}\, ds\\ &+ \int_0^1 \Big(\nabla_{12}^2 g(x_k, y_0^\star + d_m^\star + se_m^{2^m};\eta_k,\xi_k) - \nabla_{12}^2 g(x_k, y_0^\star + se_m^{2^m};\eta_k,\xi_k)\Big)e_m^{2^m}\, ds.\end{aligned}$$

Therefore, the boundedness and Lipschitz continuity of $\nabla_{12}^2 g$, together with $\|d_m^\star\| = \mathcal{O}(\lambda_m^{-1})$, imply

$$\|B_m\| \leq C\|\Delta_m^{2^m}\| + C\lambda_m^{-1}\|e_m^{2^m}\|.$$

Since

$$G_k(m,\lambda_m) - \Gamma_k(\lambda_m;\eta_k,\xi_k) = \nabla_1 f(x_k, z_k^{2^m}(\lambda_m);\eta_k,\xi_k) - \nabla_1 f(x_k, z_m^\star;\eta_k,\xi_k) + \lambda_m B_m,$$

Lemma A.9 and Lemma A.6 yield, for $m \in \{n, n-1\}$,

$$\mathbb{E}[\|G_k(m,\lambda_m) - \Gamma_k(\lambda_m;\eta_k,\xi_k)\|^2 \mid \mathcal{F}_k] \leq \mathcal{O}(\lambda_m^{-2}).$$

Consequently,

$$\begin{aligned}G_k(n,\lambda_n) - G_k(n-1,\lambda_{n-1}) =\ &G_k(n,\lambda_n) - \Gamma_k(\lambda_n;\eta_k,\xi_k) + \Gamma_k(\lambda_n;\eta_k,\xi_k) - \Gamma_k(\lambda_{n-1};\eta_k,\xi_k)\\ &+ \Gamma_k(\lambda_{n-1};\eta_k,\xi_k) - G_k(n-1,\lambda_{n-1}).\end{aligned}$$

Next, put $\delta = 1/\lambda$ and define

$$\begin{aligned}&U(\delta) := \Gamma_k(1/\delta;\eta_k,\xi_k)\\ &\quad= \nabla_1 f(x_k, y^\star(x_k,\delta;\xi_k);\eta_k,\xi_k) + \delta^{-1}\Big(\nabla_1 g(x_k, y^\star(x_k,\delta;\xi_k);\eta_k,\xi_k) - \nabla_1 g(x_k, y_0^\star;\eta_k,\xi_k)\Big).\end{aligned}$$

For $\delta > 0$, the quotient can be written as

$$\int_0^1 \nabla_{12}^2 g(x_k, y_0^\star + s(y^\star(x_k, \delta; \xi_k) - y_0^\star); \eta_k, \xi_k) \cdot \frac{y^\star(x_k, \delta; \xi_k) - y_0^\star}{\delta}\, ds.$$

To justify the dependence of this quotient on δ, set $d_\delta := y^\star(x_k, \delta; \xi_k) - y_0^\star$ and

$$A_\delta := \int_0^1 \nabla_{22}^2 \bar{g}(x_k, y_0^\star + s d_\delta; \xi_k)\, ds.$$

Subtracting the two first-order optimality conditions gives

$$A_\delta d_\delta + \delta \nabla_2 \bar{f}(x_k, y^\star(x_k, \delta; \xi_k); \xi_k) = 0,$$

and hence

$$w_\delta := \frac{d_\delta}{\delta} = -A_\delta^{-1} \nabla_2 \bar{f}(x_k, y^\star(x_k, \delta; \xi_k); \xi_k).$$

By Lemma A.7 and the Lipschitz continuity of $\nabla^2 \bar{g}$, A_δ is Lipschitz in δ. Moreover, $A_\delta \succeq \mu_g I$, so the inverse-matrix identity and the Lipschitz continuity of $\nabla \bar{f}$ show that w_δ is also Lipschitz. The integral representation above and the samplewise Lipschitz bounds on ∇f and $\nabla^2 g$ therefore make U Lipschitz, uniformly in (η_k, ξ_k). Consequently,

$$\|\Gamma_k(\lambda_n; \eta_k, \xi_k) - \Gamma_k(\lambda_{n-1}; \eta_k, \xi_k)\| \le C \left| \frac{1}{\lambda_n} - \frac{1}{\lambda_{n-1}} \right| \le \mathcal{O}(\lambda_{n-1}^{-1}).$$

Combining the three estimates and using $\lambda_{n-1}^2 \asymp 2^{n-1}$, we conclude that

$$\mathbb{E}[\|G_k(n, \lambda_n) - G_k(n-1, \lambda_{n-1})\|^2 \mid \mathcal{F}_k] \le \mathcal{O}(\lambda_n^{-2}) + \mathcal{O}(\lambda_{n-1}^{-2}) + \mathcal{O}(\lambda_{n-1}^{-2}) \le \mathcal{O}(2^{-(n-1)}).$$

This completes the proof. □

Lemma A.15. *Under Assumptions 2.1 and 2.2, consider Algorithm 2 and set*

$$Z_N := \sum_{j=1}^{N} 2^{-j} = 1 - 2^{-N}, \qquad p_n := \frac{2^{-n}}{Z_N}, \quad n \in [N].$$

Define the conditional terminal mean

$$\bar{G}_{k,N} := \mathbb{E}[G_k(N, \lambda_N) \mid \mathcal{F}_k] = \mathbb{E}[\nabla_1 L(x_k, z_k^{2^N}(\lambda_N), y_k^{2^N}, \lambda_N; \eta_k, \xi_k) \mid \mathcal{F}_k].$$

Then

$$\mathbb{E}\Big[\big\| \bar{G}_{k,N} - \big(G_k(0, \lambda_0) + p_{\nu_k}^{-1} [G_k(\nu_k, \lambda_{\nu_k}) - G_k(\nu_k - 1, \lambda_{\nu_k - 1})] \big) \big\|^2 \,\Big|\, \mathcal{F}_k \Big] \qquad \le \mathcal{O}(N).$$

Proof. Write

$$v_k(n) := G_k(0, \lambda_0) + p_n^{-1} [G_k(n, \lambda_n) - G_k(n-1, \lambda_{n-1})].$$

The triangle inequality gives

$$\begin{aligned} \mathbb{E}[\|\bar{G}_{k,N} - v_k(\nu_k)\|^2 \mid \mathcal{F}_k] &\le 2\mathbb{E}[\|\bar{G}_{k,N} - G_k(0, \lambda_0)\|^2 \mid \mathcal{F}_k] \\ &\quad + 2\mathbb{E}\Big[\big\| p_{\nu_k}^{-1} [G_k(\nu_k, \lambda_{\nu_k}) - G_k(\nu_k - 1, \lambda_{\nu_k - 1})] \big\|^2 \Big| \mathcal{F}_k \Big]. \end{aligned} \tag{19}$$

By Lemmas A.6 and A.7 and $2^N \asymp \lambda_N^2$,

$$\mathbb{E}[\|z_k^{2^N}(\lambda_N) - y_k^{2^N}\|^2 \mid \mathcal{F}_k] \le \mathcal{O}(2^{-N}) + \mathcal{O}(\lambda_N^{-2}) = \mathcal{O}(\lambda_N^{-2}).$$

Consequently, Jensen's inequality, the definition of G_k, and the Lipschitz assumptions imply

$$\|\bar{G}_{k,N}\|^2 \le \mathbb{E}[\|G_k(N, \lambda_N)\|^2 \mid \mathcal{F}_k] \le 2\ell_{f,0}^2 + 2\lambda_N^2 \ell_{g,1}^2 \mathbb{E}[\|z_k^{2^N}(\lambda_N) - y_k^{2^N}\|^2 \mid \mathcal{F}_k] = \mathcal{O}(1).$$

The same one-step estimate at the fixed level 0, together with Assumption 2.2, gives $\mathbb{E}[\|G_k(0, \lambda_0)\|^2 \mid \mathcal{F}_k] = \mathcal{O}(1)$. Thus the first term on the right of (19) is $\mathcal{O}(1)$.

For the second term, averaging over the sampled level and using Lemma A.14 yields

$$\begin{aligned}\mathbb{E}\Big[\big\|p_{\nu_k}^{-1}[G_k(\nu_k,\lambda_{\nu_k})-G_k(\nu_k-1,\lambda_{\nu_k-1})]\big\|^2\Big|\mathcal{F}_k\Big]&\\ =\sum_{n=1}^{N}p_n^{-1}\mathbb{E}[\|G_k(n,\lambda_n)-G_k(n-1,\lambda_{n-1})\|^2\mid\mathcal{F}_k]&\\ \le\sum_{n=1}^{N}p_n^{-1}\mathcal{O}(2^{-(n-1)})=\mathcal{O}(N),&\end{aligned}$$

because $p_n^{-1}=Z_N2^n$. Substitution in (19) proves the result.

□

A.9 Proof of Theorem 3.1

For brevity, let

$$G_k:=\nabla_1L(x_k,z_{k+1},y_{k+1},\lambda_k;\eta_k,\xi_k),\qquad \bar{G}_k:=\mathbb{E}[G_k\mid\mathcal{F}_k].$$

By Lemmas A.6 and A.7 and $T_k\ge C_T\lambda_k^2\log(\lambda_k+1)$, there is a constant $C_{zy}>0$, independent of k, such that

$$\mathbb{E}[\|z_{k+1}-y_{k+1}\|^2\mid\mathcal{F}_k]\le\mathcal{O}(T_k^{-1})+\mathcal{O}(\lambda_k^{-2})\le C_{zy}\lambda_k^{-2}. \tag{20}$$

The Lipschitz assumptions and the definition of L imply

$$\mathbb{E}[\|G_k\|^2\mid\mathcal{F}_k]\le2\ell_{f,0}^2+2\lambda_k^2\ell_{g,1}^2\mathbb{E}[\|z_{k+1}-y_{k+1}\|^2\mid\mathcal{F}_k]\le2\ell_{f,0}^2+2\ell_{g,1}^2C_{zy}=:C_G. \tag{21}$$

Set $C_m:=\ell_{F,1}C_G/2$. The $\ell_{F,1}$-smoothness of F, followed by conditional expectation, gives

$$\begin{aligned}\mathbb{E}[F(x_{k+1})-F(x_k)\mid\mathcal{F}_k]&\le-\alpha_k\langle\nabla F(x_k),\bar{G}_k\rangle+\frac{\ell_{F,1}\alpha_k^2}{2}\mathbb{E}[\|G_k\|^2\mid\mathcal{F}_k]\\ &\le-\frac{\alpha_k}{2}\|\nabla F(x_k)\|^2+\frac{\alpha_k}{2}\|\nabla F(x_k)-\bar{G}_k\|^2+C_m\alpha_k^2,\end{aligned}$$

where we used $-\langle a,b\rangle=-\frac12\|a\|^2-\frac12\|b\|^2+\frac12\|a-b\|^2$ and dropped the nonpositive term involving $\|\bar{G}_k\|^2$. By Lemma A.12, there is a constant C_B such that independent of k, such that $\|\nabla F(x_k)-\bar{G}_k\|^2\le C_B\lambda_k^{-2}$.

Taking total expectation, rearranging, and summing the resulting inequality without first dividing by the varying stepsize gives

$$\frac12\sum_{k=1}^{K}\alpha_k\mathbb{E}[\|\nabla F(x_k)\|^2]\le\mathbb{E}[F(x_1)]-F^\star+\frac{C_B}{2}\sum_{k=1}^{K}\alpha_k\lambda_k^{-2}+C_m\sum_{k=1}^{K}\alpha_k^2. \tag{22}$$

This argument uses only the assumed lower bound $F^\star$ and does not require a uniform upper bound on the iterates' objective values.

With the prescribed choices $\alpha_k=c_\alpha(k+1)^{-1/2}$ for a sufficiently small constant $c_\alpha>0$ and $\lambda_k=c_\lambda(k+1)^{1/4}$, the stepsizes are nonincreasing and

$$\sum_{k=1}^{K}\alpha_k\lambda_k^{-2}\le\frac{c_\alpha}{c_\lambda^2}\log(K+1),\qquad\sum_{k=1}^{K}\alpha_k^2\le c_\alpha^2\log(K+1).$$

Since $\alpha_k\ge\alpha_K$, setting $C_l:=C_Bc_\alpha/c_\lambda^2+2C_{\mathrm{m}}c_\alpha^2$, (22) yields

$$\frac1K\sum_{k=1}^{K}\mathbb{E}[\|\nabla F(x_k)\|^2]\le\frac{2\{\mathbb{E}[F(x_1)]-F^\star\}+C_l\log(K+1)}{K\alpha_K}=\mathcal{O}\left(\frac{1+\log(K+1)}{\sqrt{K}}\right). \tag{23}$$

Consequently, it suffices to take $K = \widetilde{\mathcal{O}}(\epsilon^{-4})$ to make the left-hand side at most ϵ^2. Moreover, $T_K = \widetilde{\mathcal{O}}(\epsilon^{-2})$ and

$$\sum_{k=1}^{K} T_k = \widetilde{\mathcal{O}}(K^{3/2}) = \widetilde{\mathcal{O}}(\epsilon^{-6}).$$

Thus the ξ-sample complexity and the complexities of $\nabla_1 f, \nabla_1 g$ are $\widetilde{\mathcal{O}}(\epsilon^{-4})$, whereas the η-sample complexity and the complexities of $\nabla_2 f, \nabla_2 g$ are $\widetilde{\mathcal{O}}(\epsilon^{-6})$.

A.10 Proof of Theorem 3.2

It follows from Lemma A.4 that $F(x)$ is $\ell_{F,1}$-smooth, which implies

$$F(x_{k+1}) - F(x_k) \le \langle \nabla F(x_k), x_{k+1} - x_k \rangle + \frac{\ell_{F,1}}{2}\|x_{k+1} - x_k\|^2.$$

For notational simplicity, we adopt the following conventions:

$$v_k(\nu_k, \lambda_{\nu_k}) = \; G_k(0, \lambda_0) + p_{\nu_k}^{-1}(G_k(\nu_k, \lambda_{\nu_k}) - G_k(\nu_k - 1, \lambda_{\nu_k - 1})).$$

Let $M = \lfloor c_0 N \rfloor$, $q_M = \sum_{n=M+1}^{N} p_n$, and

$$s_n := \mathbf{1}_{\{n \le M\}} + a_1 \mathbf{1}_{\{n > M\}}.$$

Then $x_{k+1} = x_k - \alpha_0 s_{\nu_k} v_k(\nu_k, \lambda_{\nu_k})$. Although $\mathbb{E}[v_k(\nu_k, \lambda_{\nu_k}) \mid \mathcal{F}_k] = \mathbb{E}[G_k(N, \lambda_N) \mid \mathcal{F}_k]$, the level-dependent multiplier s_{ν_k} is correlated with the sampled level and its correction. Hence, unless $a_1 = 1$, the actual weighted update is not an unbiased terminal-level update. The MLMC identity must instead be applied with the indicators retained inside the expectation.

In particular,

$$\mathbb{E}[\mathbf{1}_{\{\nu_k \le M\}} v_k(\nu_k, \lambda_{\nu_k}) \mid \mathcal{F}_k] = \mathbb{E}[G_k(M, \lambda_M) - q_M G_k(0, \lambda_0) \mid \mathcal{F}_k],$$
$$\mathbb{E}[\mathbf{1}_{\{\nu_k > M\}} v_k(\nu_k, \lambda_{\nu_k}) \mid \mathcal{F}_k] = \mathbb{E}[G_k(N, \lambda_N) - G_k(M, \lambda_M) + q_M G_k(0, \lambda_0) \mid \mathcal{F}_k].$$

Consequently, with

$$\bar{G}_j := \mathbb{E}[G_k(j, \lambda_j) \mid \mathcal{F}_k], \qquad j \in \{0, M, N\},$$

we have

$$\mathbb{E}[s_{\nu_k} v_k(\nu_k, \lambda_{\nu_k}) \mid \mathcal{F}_k] = a_1 \bar{G}_N + (1 - a_1)(\bar{G}_M - q_M \bar{G}_0).$$

Using the smoothness of F, we obtain

$$\begin{aligned}
&\mathbb{E}[F(x_{k+1}) - F(x_k) \mid \mathcal{F}_k] \\
&\le -\alpha_0 \left\langle \nabla F(x_k), a_1 \bar{G}_N + (1 - a_1)(\bar{G}_M - q_M \bar{G}_0) \right\rangle \\
&\quad + \frac{\ell_{F,1} \alpha_0^2}{2} \mathbb{E}[s_{\nu_k}^2 \|v_k(\nu_k, \lambda_{\nu_k})\|^2 \mid \mathcal{F}_k].
\end{aligned}$$

Applying $-\langle a, b \rangle \le -\frac{1}{2}\|a\|^2 + \frac{1}{2}\|a - b\|^2$ to the two weighted inner-product terms gives

$$\begin{aligned}
&- \alpha_0 \left\langle \nabla F(x_k), a_1 \bar{G}_N + (1 - a_1)(\bar{G}_M - q_M \bar{G}_0) \right\rangle \\
&\quad \le -\frac{\alpha_0}{2}\|\nabla F(x_k)\|^2 + \frac{a_1 \alpha_0}{2}\|\nabla F(x_k) - \bar{G}_N\|^2 \\
&\qquad + \frac{(1 - a_1)\alpha_0}{2}\|\nabla F(x_k) - (\bar{G}_M - q_M \bar{G}_0)\|^2.
\end{aligned}$$

By Lemma A.13 and Lemma A.11 with N replaced by M,

$$\|\nabla F(x_k) - \bar{G}_N\|^2 = \mathcal{O}(\lambda_N^{-2}), \qquad \|\nabla F(x_k) - \bar{G}_M\|^2 = \mathcal{O}(\lambda_M^{-2}).$$

Moreover, $q_M = \mathcal{O}(2^{-M})$, and Jensen's inequality together with the fixed-level estimate used in Lemma A.15 gives $\|\bar{G}_0\|^2 = \mathcal{O}(1)$. Since $\lambda_M^{-2} \asymp 2^{-M}$,

$$\|\nabla F(x_k) - (\bar{G}_M - q_M \bar{G}_0)\|^2 \le \mathcal{O}(\lambda_M^{-2}) + \mathcal{O}(q_M^2) = \mathcal{O}(\lambda_M^{-2}).$$

For the quadratic term, since $s_{\nu_k} \le 1$, Lemma A.13 and Lemma A.15 yield

$$\begin{aligned}&\mathbb{E}[s_{\nu_k}^2 \|v_k(\nu_k, \lambda_{\nu_k})\|^2 \mid \mathcal{F}_k] \\ &\le 2\|\nabla F(x_k)\|^2 + 4\|\bar{G}_N - \nabla F(x_k)\|^2 + 4\mathbb{E}[\|\bar{G}_N - v_k(\nu_k, \lambda_{\nu_k})\|^2 \mid \mathcal{F}_k] \\ &\le 2\|\nabla F(x_k)\|^2 + \mathcal{O}(\lambda_N^{-2}) + \mathcal{O}(N).\end{aligned}$$

Combining the displays above and taking the total expectation yields

$$\begin{aligned}\mathbb{E}[F(x_{k+1}) - F(x_k)] \le \Big(-\frac{\alpha_0}{2} + \ell_{F,1}\alpha_0^2\Big)\,\mathbb{E}[\|\nabla F(x_k)\|^2] \\ + \alpha_0 \mathcal{O}(\lambda_N^{-2} + \lambda_M^{-2}) + \alpha_0^2 \mathcal{O}(\lambda_N^{-2} + N).\end{aligned}$$

Because $a_1 \le 1$ and $\alpha_0 \le a_1/(8\ell_{F,1})$, the descent coefficient satisfies

$$-\frac{\alpha_0}{2} + \ell_{F,1}\alpha_0^2 \le -\frac{\alpha_0}{2} + \frac{a_1\alpha_0}{8} \le -\frac{3\alpha_0}{8} \le -\frac{\alpha_0}{4},$$

so that

$$\mathbb{E}[F(x_{k+1}) - F(x_k)] \le -\frac{\alpha_0}{4}\mathbb{E}[\|\nabla F(x_k)\|^2] + \alpha_0 \mathcal{O}(\lambda_N^{-2} + \lambda_M^{-2}) + \alpha_0^2 \mathcal{O}(\lambda_N^{-2} + N).$$

Therefore,

$$\mathbb{E}[\|\nabla F(x_k)\|^2] \le \frac{4}{\alpha_0}\left(\mathbb{E}[F(x_k)] - \mathbb{E}[F(x_{k+1})]\right) + \mathcal{O}(\lambda_N^{-2} + \lambda_M^{-2}) + \alpha_0 \mathcal{O}(\lambda_N^{-2} + N).$$

The above inequality and the definition of λ_n in Algorithm 2 imply

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\|\nabla F(x_k)\|^2] \le \frac{4\mathbb{E}[F(x_1) - F(x_{K+1})]}{\alpha_0 K} + \mathcal{O}(\lambda_N^{-2} + \lambda_M^{-2}) + \alpha_0 \mathcal{O}(\lambda_N^{-2} + N).$$

Since $F(x_{K+1}) \ge F^\star$, the numerator in the first term is at most $\mathbb{E}[F(x_1)] - F^\star$. With $p_n = 2^{-n}/(1 - 2^{-N})$ for $n \in [N]$, the average number of conditional-sampling rounds in an inner loop is

$$\mathbb{E}[2^{\nu_k}] = \sum_{n=1}^{N} 2^n \frac{2^{-n}}{1 - 2^{-N}} = \frac{N}{1 - 2^{-N}} \le 2N.$$

Together with the one fresh outer sample, this gives $\mathbb{E}[\#\eta\text{-samples per outer iteration}] = 1 + \mathbb{E}[2^{\nu_k}] = \mathcal{O}(N)$. Each conditional-sampling round uses only a constant number of lower-level first-order oracle calls. Reusing one sampled function at several query points does not merge the corresponding gradient costs: one draw is counted as one conditional sample, while every evaluation of $\nabla_i f$ or $\nabla_i g$ is counted as a separate partial-gradient call. To ensure $\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\|\nabla F(x_k)\|^2] \le \epsilon^2$, it suffices to set $\alpha_0 = \Theta(\epsilon^2/N)$ (with $\alpha_0 \le a_1/(8\ell_{F,1})$), $K = \widetilde{\mathcal{O}}(\epsilon^{-4})$, and N such that $M = \lfloor c_0 N \rfloor \ge 1$ and $2^{-M} = \mathcal{O}(\epsilon^2)$; equivalently, it suffices to take $N = \lceil C_N \log(\epsilon^{-1}) + C_{\mathrm{add}} \rceil$ with $C_N > 2/(c_0 \log 2)$ and a sufficiently large constant C_{add} depending only on the problem constants. Over all K outer iterations, the ξ-sample count and the $\nabla_1 f, \nabla_1 g$ call counts are $\mathcal{O}(K)$, whereas the η-sample count and the $\nabla_2 f, \nabla_2 g$ call counts are $\mathcal{O}(KN)$. Hence all of these complexities have the stated $\widetilde{\mathcal{O}}(\epsilon^{-4})$ order.

B Details of Numerical Experiments

This section gives the data, evaluation, implementation, and parameter settings for the two numerical experiments in Section 4. All experiments are implemented in Julia 1.12 and run on an Apple MacBook Pro with M4 Pro (14 cores) and 48 GB memory. The gradients of f and g, and the Hessian-vector products required by the Hessian-based baseline, are evaluated in explicit closed form rather than through automatic differentiation; no Hessian matrix is explicitly formed; all operators act in place on preallocated buffers. Before the timed runs, every solver is invoked once without timing and the random seed is reset afterwards, so that the reported wall-clock times exclude compilation and reflect steady-state execution.

The implementations are tuned for performance and therefore do not follow the theoretical parameter prescriptions verbatim; the specific choices are reported together with the corresponding experiment. In particular, the outer budgets K are chosen so that the methods are compared at matched wall-clock cost rather than at matched iteration counts.

For reduction + stocBiO we use Chebyshev basis functions, following the theoretical guarantees of the reduction framework in (Bouscary et al., 2026). We select degree of the basis based on two criteria: the resulting Gram matrix of the basis must be full rank, as required by the theorems, and the basis must be able to represent a context-dependent lower-level solution; a candidate degree that is rank-deficient on the observed contexts is repaired by pruning the linearly dependent directions and orthonormalizing the remaining span. Among the admissible degrees, we use the smallest one with the best measured performance, so that the comparison is made against its strongest admissible configuration. When the ranking of the admissible degrees by expressiveness is data-dependent, it is settled by measurement and the remaining admissible degrees are reported for reference, see Table 1; when that ranking is fixed by the number of contexts, the strongest admissible degree is identified analytically and we do not sweep different degrees.

B.1 META-LEARNING

B.1.1 DATA AND PROBLEM SETTINGS

We use $M = 3$ tasks with 7 classes per task and 500 images per class. The dataset consists of Tiny ImageNet features (Moustafa, 2017) extracted by a pre-trained ResNet-18 network (He et al., 2016); each image is represented by a 512-dimensional feature vector, and we use a $90\%/10\%$ train/validation split within each class. For every task, l_ξ is the multiclass logistic loss of a linear classifier and is convex in the classifier parameters. Since the lower-level quadratic regularizer has $\gamma = 2$, the actual samplewise strong-convexity modulus certified by the model is $\mu_g = \gamma = 2$.

B.1.2 ALGORITHMIC PARAMETERS AND EVALUATIONS

For all algorithms, we use numerically tuned proxies $L_{\mathrm{tune}} = \mu_{\mathrm{tune}} = 1000$. For Algorithm 1, we use outer stepsize $\alpha_k = 5/\sqrt{k+1}$, inner stepsize $\beta_t = 50/(\mu_{\mathrm{tune}}(t+1))$, $C_T = 1$, and $K = 1500$ outer iterations (evaluation every 15 iterates). For Algorithm 2, we use $\epsilon = 10^{-4}$, truncation level $N = 12$, outer base stepsize $\alpha_0 = 0.4$, adaptive parameters $a_1 = 0.05$ and $c_0 = 0.3$, inner stepsize $\beta_t = 50/(\mu_{\mathrm{tune}}(t+1))$, and $K = 22000$ outer iterations (evaluation every 220 iterates). For the RT-MLMC Hessian-based method, we use maximum iterations 18000, RT-MLMC level parameter 12, $L_{g,1} = 10$, outer stepsizes $\alpha_t = 0.2/\sqrt{t}$, and inner stepsize coefficient $\beta = 1.0$. For reduction + stocBiO the degree of the Chebyshev basis functions is set to 2, the lowest degree whose basis has one function per task and is therefore square and full rank on the observed task contexts; lower degrees cannot represent a distinct lower-level solution per task, and higher degrees only increase the dimension of the reduced variable. Specifically, all Gram matrices for degrees 2, 10 and 50 have rank 3, equal to the number of tasks, with the nullity and dimension growing and the condition number staying below 2.2. We use maximum iterations 9000, $D = 50$ inner iterations, stepsizes $\alpha = 0.01$, $\beta = 0.1$, Neumann stepsize $\eta = 10^{-3}$, and Neumann length $Q = 15$.

Optimizer minibatches use 100 training samples and 100 validation samples per draw. For each algorithm, we evaluate 100 equally spaced outer iterates, including the initialization. At each selected iterate, assuming the full knowledge of all contexts and data, we form the exact lower-level objective and solve it by gradient descent with Armijo backtracking until the gradient residual certifies $\|\hat{y} - y^\star\| \le \|\nabla_y \bar{g}\|/\mu_g \le 5 \times 10^{-4}$ with $\mu_g = \gamma = 2$. This is only for evaluation and certifies that the reconstructed lower-level solution is within 5×10^{-4} of the exact one, so that the reported upper-level objective, evaluated at $\hat{y}$, differs from its exact value by at most $\ell_{f,0}$ times this tolerance. For each selected iterate, we report the upper-level objective along with the validation loss and classification error, averaged over all tasks and over the full validation sets. All results are further averaged over 10 independent runs with fixed seeds for data construction and sampling.

B.1.3 DISCUSSION OF THE MAIN COMPARISON

The Hessian-based method and reduction + stocBiO converge more slowly, reflecting the heavy computational overhead of Neumann-series approximations for Hessian-vector products and high-

dimensional reduced-variable optimization, respectively. The heavy costs also limit the sample efficiency of the competitors, since most of the computational time is spent on second-order oracle approximations or reduced-variable optimization rather than on processing data: both of our proposed methods achieve lower validation error than the competitors across all sample budgets. These results demonstrate the practical advantage of the proposed approach on this instance, particularly the RT-MLMC acceleration in Algorithm 2.

B.1.4 Adaptive-stepsize ablation

To study the adaptive stepsize in Algorithm 2, we rerun the main experiment for several (a_1, c_0) configurations, varying the scaling parameter a_1 and the threshold c_0. For each configuration, we keep everything else unchanged as in the main experiment, including the same tuned constants and the same fixed seeds, so that the runs differ only in these two parameters. Following the main experiment, we compare configurations by their mean objective and validation error, focusing mainly on their variances. Recall that when the sampled level satisfies $\nu_k > c_0 N$, the outer stepsize is reduced from α_0 to $a_1\alpha_0$. Therefore, intuitively, smaller a_1 gives a more aggressive reduction, while smaller c_0 triggers it more frequently.

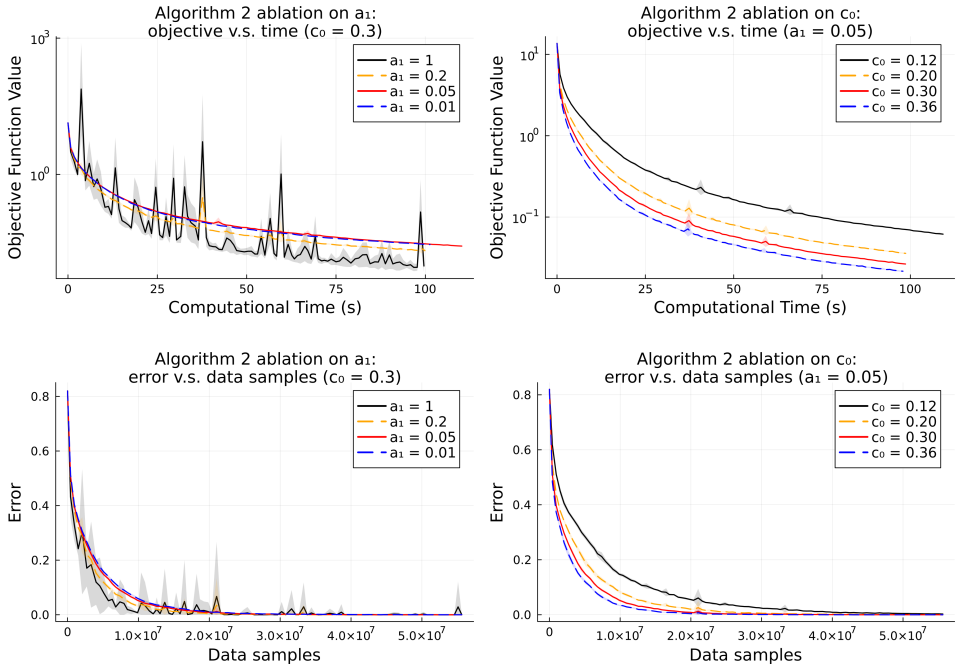


Figure 3: Ablation of the adaptive stepsize parameters a_1 and c_0 of Algorithm 2. Left pair: varying a_1 with fixed $c_0 = 0.3$ (objective versus time; validation error versus data samples). Right pair: varying c_0 with fixed $a_1 = 0.05$ (objective versus time; validation error versus data samples). Shaded regions show mean ± 1 standard deviation over 10 runs.

The left pair of Figure 3 examines the effect of a_1 with fixed $c_0 = 0.3$. The non-adaptive setting $a_1 = 1$ attains the lowest mean objective but with significantly larger variance. In contrast, adaptive settings with smaller a_1 monotonically trade mean objective for stability. All settings reach zero validation error. The right pair shows the effect of c_0 with fixed $a_1 = 0.05$: frequent triggering degrades the mean objective. On this instance the benefit of the adaptive stepsize comes from the rare deep RT-MLMC levels it is designed for: it shrinks the occasional spike in the update direction, which lowers the variance of the outer updates. Triggering it often, by contrast, only slows the outer iteration, because most such triggers fall on shallow levels, whose updates are already "stable". The default $(a_1 = 0.05, c_0 = 0.3)$ is therefore a robust choice.

To rule out the effect of a mere uniform stepsize reduction, we include a non-adaptive control in Figure 4. Specifically, we disable the adaptive stepsize by setting $a_1 = 1$ and compare the stepsizes

$\alpha_0 = 0.4, 0.2, 0.1$ against the adaptive default ($\alpha_0 = 0.4, a_1 = 0.05$). As shown in the figures, uniformly reducing α_0 does not match the adaptive scheme: the mean objective is at best comparable, and it degrades markedly once α_0 is reduced further. The variances of the non-adaptive stepsizes, by contrast, stay significantly larger than those of the adaptive scheme at every value of α_0. This therefore confirms that the lower variance comes from the level-dependent shrinkage itself, not from a smaller stepsize.

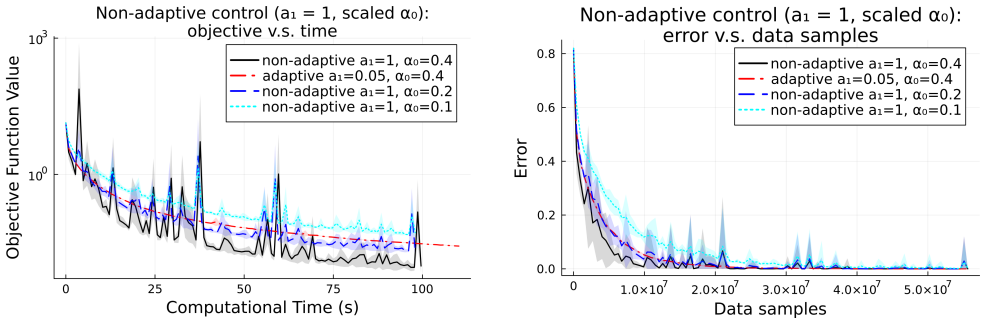


Figure 4: Non-adaptive control for the ablation of the adaptive stepsize of Algorithm 2: the adaptive stepsize is disabled ($a_1 = 1$) and the base stepsize is uniformly reduced instead, compared with the adaptive default ($a_1 = 0.05$). Left: objective versus computational time. Right: validation error versus data samples. Shaded regions show mean ± 1 standard deviation over 10 runs.

B.2 Contextual dynamic pricing for residential electricity demand response

B.2.1 Backgrounds

We consider a contextual dynamic pricing problem for residential electricity demand response (Guo, 2023) calibrated on the Low Carbon London (LCL) dataset (UK Power Networks, 2015). An electricity retailer or aggregator serves a fixed portfolio of H residential households and seeks to reduce the portfolio peak load, flatten the aggregate load profile, and keep household bills near those under a reference tariff. To this end, each day the retailer publishes a day-ahead dynamic tariff consisting of 48 half-hourly prices, tailored to the day's context ξ, which encodes calendar information (weekday, seasonality), forecasted conditions, and recent portfolio consumption. After observing the tariff, each household adjusts its consumption profile relative to its baseline (no-response) load, trading off bill savings against the discomfort of departing from baseline behavior.

This application can be naturally formulated as a CSBO problem for two reasons. First, the number of distinct day-ahead contexts is effectively unbounded due to continuously varying features such as weather forecasts and recent consumption patterns, making it infeasible to fit a separate tariff for each context; instead, a shared parametric policy pools information across contexts. Second, the decisions induced by the tariff must account for the conditional distribution of household baseline loads given the context, $\mathbb{P}_{\eta|\xi}$, which is exactly the lower-level structure in (1). CSBO captures both aspects: the upper level learns the shared tariff policy parameter x, while the lower-level models each household's optimal consumption adjustment given the tariff and the day's conditions.

The Low Carbon London project (UK Power Networks, 2015) collected half-hourly smart meter readings from 5,567 London households between November 2011 and February 2014. Approximately 1,100 households participated in a dynamic time-of-use tariff trial in 2013, while the remaining households stayed on a flat rate. The dToU trial tariff set day-ahead prices at three levels: Low at 3.99 p/kWh, Normal at 11.76 p/kWh, and High at 67.20 p/kWh, with the flat-rate tariff at 14.228 p/kWh; we take the Normal level as the reference tariff p_{ref}. We use the period from July 1, 2012, to September 30, 2013, which provides a full year of pre-treatment data for baseline clustering and a sufficient trial window for response calibration and validation. Households are selected based on data completeness: we require at least 90% coverage over both the pre-treatment window (July–December 2012) and the trial window (January–September 2013), and we exclude households whose average daily consumption is below 1.0 kWh. This yields 4,421 eligible households, of which 3,522 are on the flat rate and 899 participated in the dToU trial.

We construct the CSBO instance in four steps. We first cluster households by their pre-treatment load shapes (Appendix B.2.2), which gives the pooling index used by everything that follows. We then fit the conditional no-response law $\mathbb{P}_{\eta|\xi}$ on flat-rate households alone (Appendix B.2.3), so that it never sees a dynamic tariff. Using that law as the counterfactual baseline, we calibrate the cluster-specific household response by inverse optimization on the dToU trial (Appendix B.2.4), which yields the lower-level objective g. Finally, we specify the tariff map and the retailer objective f, completing the CSBO instance (Appendix B.2.5).

B.2.2 Baseline household clustering

Households are clustered by their pre-treatment load shapes to enable statistical pooling in both the conditional no-response model and the response calibration: households with similar timing and scale of consumption are expected to share a similar price response, so a common discomfort potential Φ_r can be identified from the dToU trial. Load-shape clustering is standard for residential demand response, including on London smart-meter panels (Michalakopoulos et al., 2024; Li et al., 2023). All features are computed from complete days in the July–December 2012 window, before any dynamic tariff is applied. The features have two blocks. The shape block is the weekday and weekend mean 48-slot profiles, each normalized to daily mean one so that clustering sees timing rather than size. The scale block consists of four scalars: log mean daily energy; the weekday load factor, measured by the mean over peak of the unnormalized weekday profile; the log ratio of morning (07:00–11:00) to evening-peak (17:00–20:00) weekday load, which separates daytime-active from evening-peaking households; and the coefficient of variation of daily energy, which measures day-to-day irregularity. Each coordinate is z-scored across households; the two blocks are then rescaled so that they contribute equally to the squared Euclidean norm used by k-means. We compare $K = 4, \ldots, 12$ using k-means, evaluating each candidate via silhouette score and bootstrap stability, i.e., adjusted Rand index over repeated 80% subsamples. The selected clustering, $K = 5$, achieves the highest combined silhouette and stability score. The five clusters exhibit distinct load patterns: Cluster 1 is evening-peaking, Cluster 2 is mid-morning peaking, Cluster 3 is evening-peaking with a lower load factor, Cluster 4 is overnight-peaking and consistent with Economy-7 night-storage heating, and Cluster 5 is evening-peaking with high evening concentration. Cluster 4 is a small cluster, but its overnight profile is sufficiently distinctive that merging it with the evening-peaking majority would mix two different response regimes. Cluster membership is fixed throughout and serves as the response heterogeneity index r_h in the lower-level model.

B.2.3 Construction of $\mathbb{P}_{\eta|\xi}$

The conditional no-response load $\eta \in \mathbb{R}^{H\times 48}$ represents what the portfolio would consume on a given day if no dynamic tariff were applied. We construct $\mathbb{P}_{\eta|\xi}$ empirically from flat-rate households using a clustering-based conditional-mean-plus-residual approach (Sun et al., 2019; Weng et al., 2018), which preserves the cross-household and cross-time correlation structure essential for aggregate system metrics. Precisely speaking, we first fit a context-dependent conditional mean for each household, then resample whole-day residuals around that mean so that a draw of $\eta \mid \xi$ is the predicted profile plus a jointly sampled shock.

We first construct the conditional mean model. For each household h, we model the expected no-response load multiplicatively: $\mathbb{E}[\eta_{h,t} \mid \xi] = s_h \cdot u_{c(h),t}(\xi)$, where s_h is household h's pre-treatment scale, i.e., averaged half-hourly consumption, and $u_{c(h),t}(\xi)$ is a cluster-level normalized load profile that depends on the context. The cluster profiles are fitted via ridge regression of scale-normalized load on day-level features: weekday indicators, two annual Fourier harmonics, previous-day same-slot load, previous-day daily mean, and a seven-day trailing mean. The model is trained on 2,818 flat-rate households over July 1, 2012–June 30, 2013, with 704 households held out for validation. On the validation window (July–September 2013), the model achieves a mean absolute error of 0.0994 kWh per half-hour and reproduces the aggregate portfolio peak to within 3.3%. The held-out group, which receives the same baseline model, exhibits a comparable error of 0.1003 kWh, confirming that the counterfactual baseline is not overfitted to the flat-rate training group.

To capture uncertainty and preserve correlation, we next construct a bank of whole-day joint residuals. For each flat-rate household-day in the training window, we compute the residual $R_{h,d} = L_{h,d} - \widehat{m}_h(\xi_d)$, where $L_{h,d}$ is the observed load vector and $\widehat{m}_h(\xi_d)$ is the conditional mean computed previously. We retain the 365 days over July 1, 2012–June 30, 2013, for which a portfolio

of 350 flat-rate households has no missing data, storing each day's full 350×48 residual matrix R_d. Crucially, we resample entire days rather than individual household or time-slot residuals. This preserves system-wide shocks (weather, holidays, calendar effects) and the cross-household covariance essential for peak and aggregate metrics.

Given a context ξ, a scenario is drawn as $\eta = \widehat{m}(\xi) + R_d$, where $\widehat{m}(\xi) = (\widehat{m}_h(\xi))_h$ stacks the household-level conditional means and d is sampled from the bank with weights determined by context similarity. We use stratified sampling: days are first grouped by weekday/weekend and season, then weighted within each stratum by a Gaussian kernel on the standardized previous-day aggregate load and peak. The bank is built to preserve dependence across households and slots, not to deliver calibrated predictive intervals. Under a leave-one-day-out check over the 365 bank days, its nominal 80% interval covers 77.0% of the daily aggregate peaks and 76.8% of the per-slot marginals, so the law is mildly underdispersed. The aggregate-peak misses are small in absolute terms, however: they average 0.014 kWh and never exceed 0.07 kWh. Every method is evaluated on scenarios drawn from the same bank, so this common underdispersion shifts them alike and does not drive the comparison. What the bank must deliver is instead the dependence: resampling residuals independently for each household or slot, rather than whole days, would discard the cross-household covariance that the aggregate metrics rely on.

B.2.4 Household response model calibration

Evaluating a candidate tariff requires a model of how households will respond to it: the retailer cannot observe the mapping from prices to adjustments, so the lower-level objective g must be recovered from data. We take g to encode each household's trade-off between bill savings and the discomfort of departing from baseline consumption, which makes day-ahead tariff design a bilevel demand-response problem (Kovács, 2019). The parameters of this model are not known a priori; we identify them by inverse optimization on the observed dToU trial response (Kovács, 2021; Saez-Gallego et al., 2016). We recover household adjustments as deviations from a counterfactual no-response baseline, then fit a cluster-specific convex potential whose minimizers match those responses.

We first construct observed responses as deviations from a counterfactual no-response baseline. For each dToU household-day during the trial (January–September 2013), we define the observed response as $y^{\mathrm{obs}}_{h,d} = L^{\mathrm{ToU}}_{h,d} - \widehat{\eta}^0_{h,d}$, where $L^{\mathrm{ToU}}_{h,d}$ is the recorded consumption and $\widehat{\eta}^0_{h,d}$ is the counterfactual no-response load predicted by the conditional mean model in Appendix B.2.3. Any systematic bias in the baseline model would appear in both the dToU households and untreated flat-rate households. We compute the mean relative response by pooling over all dToU household-days: it is -6.4% in High-price slots, $+2.0\%$ in Low-price slots, and -2.2% in Normal-price slots. The High-price figure is 93.5 times the same construction on untreated households, confirming that the response signal is well identified.

The observed response is heterogeneous across baseline clusters, so we use cluster-specific response parameters. In particular, Cluster 2 (mid-morning peaking households) exhibits the strongest High-price avoidance at -12.2%, while Cluster 3 is weaker at -4.0%. The response is also load-dependent: cutting the same 0.05 kWh is a much larger relative change in a low-load overnight slot than at the evening peak, so households are more sensitive to a given absolute adjustment when the baseline load is low. We therefore work in relative-adjustment coordinates $z_h \in \mathbb{R}^{48}$, with induced absolute adjustment $y_h = S(\eta_h) z_h$, where $S(\eta_h) = \mathrm{diag}(\eta_{h,1} + \epsilon, \ldots, \eta_{h,48} + \epsilon)$ and $\epsilon = 0.1$ kWh, equivalently $z_{h,t} = y_{h,t}/(\eta_{h,t} + \epsilon)$.

We then fit for each cluster, by inverse optimization, a structured convex potential that includes a quadratic term, a linear term, and saturating penalties along five time-of-day blocks (overnight, morning, midday, evening peak, post-peak), matching the resolution at which the dToU trial tariff varies. A purely quadratic potential cannot fit the data: the per-unit-price response differs by a factor of 2.3 between Low and High tariffs, a diminishing marginal response to the price signal (the same price change elicits a smaller adjustment at higher prices), so the saturating terms are needed. The resulting potential for household h in cluster r takes the form

$$\Phi_r(z_h) = \frac{1}{2} z_h^\top \bar{Q}_r z_h + \sum_{k=1}^{5} a_{r,k}\, \psi_\delta(v_{r,k}^\top z_h) + c_r^\top z_h,$$

where $\bar{Q}_r$ is a symmetric positive definite matrix encoding the quadratic curvature, $\psi_\delta(s) = s^2 - \delta^2 \log\cosh(s/\delta)$ is a smooth convex saturating function, $v_{r,1}, \ldots, v_{r,5}$ are orthonormal temporal modes corresponding to the five time-of-day blocks (overnight, morning, midday, evening peak, post-peak), and $a_{r,k} \geq 0$ are nonnegative weights. The linear term c_r captures the uniform reduction in consumption observed even at the Normal tariff (-2.2% relative response).

The matrix $\bar{Q}_r$ is not a free 48×48 matrix: the dToU trial tariff varies only across the five time-of-day blocks, so the inverse problem identifies curvature only at that coarser resolution. We expand $\bar{Q}_r$ in a 17-dimensional basis consisting of a scaled identity, an adjacent-slot smoothness matrix, and the 15 independent symmetric couplings among the five blocks. The five saturating modes $v_{r,k}$ are the same block indicators. This reduced parameterization is the richest quadratic-plus-nonlinear family that the tariff design can identify since a slot-wise 48-dimensional curvature would be underdetermined by the three trial price levels. The parameters $\theta_r = (\bar{Q}_r, a_{r,\cdot}, c_r)$ are fitted by inverse optimization, i.e., by solving a bilevel optimization problem. The lower level assumes that, given a tariff, households choose the relative adjustment minimizing discomfort plus the bill incentive; the upper level then searches for the potential that makes those minimizers match the observed trial responses:

$$\min_{\theta_r} \sum_{h,d} \left\| z_h^{\mathrm{obs}}(d) - z^\star(\theta_r; w_{h,d}) \right\|^2, \qquad z^\star(\theta_r; w) \in \arg\min_z \left(\Phi_r(z; \theta_r) + w^\top z \right),$$

where $z_h^{\mathrm{obs}}(d) = S(\eta_{h,d})^{-1} y_{h,d}^{\mathrm{obs}}$ and $w_{h,d} = (\eta_{h,d} + \epsilon \mathbf{1}) \odot (p_d - p_{\mathrm{ref}})$ is the effective price on trial day d. We initialize by minimizing the first-order-condition residual $\|\nabla_z \Phi_r(z^{\mathrm{obs}}) + w\|^2$, then refine by Gauss–Newton on the upper-level prediction loss. The fitted curvature is projected so that its smallest eigenvalue is at least a fixed fraction of its largest, which keeps Φ_r uniformly strongly convex. The saturation parameter $\delta = 0.05$ is selected by the profile error on held-out trial days.

Once each Φ_r is fitted, we can sum $\Phi_{r(h)}(z_h)$ and the incremental bill to get the discomfort–bill trade-off for the household h. Because $y_h = S(\eta_h) z_h$, the incremental bill under a tariff p is $\left(p - p_{\mathrm{ref}}\right)^\top y_h = \left(S(\eta_h)(p - p_{\mathrm{ref}})\right)^\top z_h$. Summing over the portfolio gives

$$g(x, z; \eta, \xi) = \sum_{h=1}^{H} \left[\Phi_{r(h)}(z_h) + \left(S(\eta_h)(p_x(\xi) - p_{\mathrm{ref}})\right)^\top z_h \right],$$

where $p_x(\xi)$ is the tariff map which is deferred to the next subsection. This g is uniformly strongly convex in z, with a Hessian that does not depend on the scenario. The second derivative $\psi_\delta''(s) = 2 - \mathrm{sech}^2(s/\delta)$ takes values in $[1, 2)$, so the Hessian of Φ_r is sandwiched between two z-independent matrices $\bar{Q}_r + \sum_k a_{r,k} v_{r,k} v_{r,k}^\top$ and $\bar{Q}_r + 2\sum_k a_{r,k} v_{r,k} v_{r,k}^\top$. The eigenvalues of those two matrices are therefore uniform strong-convexity and smoothness constants.

B.2.5 THE CSBO FORMULATION

Here, $\eta = (\eta_1, \ldots, \eta_H)$ denotes a day's no-response load scenario, where $\eta_h \in \mathbb{R}^{48}$ is the profile of household h and $r(h)$ is its load-shape cluster. The variable z_h denotes the relative adjustment of household h, which is mapped to an absolute change through $S(\eta_h)$. The quantity $p_x(\xi)$ denotes the context-dependent tariff, and $\ell \in \mathbb{R}^{48}$ is the resulting portfolio mean load averaged over all H households. The map $p_x(\xi)$ is parameterized by six context features (calendar, seasonality, and recent portfolio load) across five time-of-day blocks, hence $d_x = 30$.

To transfer the problem as a CSBO instance, the remaining piece is the tariff map as a function of the day's context. We take $\xi \in \mathbb{R}^6$ to be the policy features of that day: an intercept, a weekend flag, two annual harmonics, and the previous-day portfolio level and peak. We keep prices inside the trial's Low–High band so that the policy stays in the range on which the household model was identified, and so that every x-dependent derivative of f and g is uniformly bounded. Concretely,

$$p_x(\xi)_t = p_L + (p_H - p_L)\, \sigma\big(a_{\mathrm{ref}} + [B(\xi)x]_t\big), \qquad \sigma(z) = \tfrac{1}{1+e^{-z}},$$

where $p_L = 3.99$ and $p_H = 67.20$ are the trial's Low and High price bounds, a_{ref} is chosen so that $x = 0$ corresponds to the reference tariff $p_{\mathrm{ref}} = 11.76 \cdot \mathbf{1}$, and $B(\xi) \in \mathbb{R}^{48 \times d_x}$ expands ξ across the five time-of-day blocks, so $d_x = 30$.

Recall from (8) that

$$f(x,z;\eta,\xi)=\lambda_{\text{peak}}\,\frac{\text{smax}_{\tau_p}(\ell)}{\ell^{\text{peak}}_{\text{ref}}}+\lambda_{\text{flat}}\,\frac{\text{Disp}_{\delta_f}(\ell)}{\ell^{\text{flat}}_{\text{ref}}}+\lambda_{\text{bill}}\,\frac{1}{H}\sum_{h=1}^{H}\rho_{\tau_b}\Big(\frac{p_x(\xi)^\top\eta_h-p_{\text{ref}}^\top\eta_h}{p_{\text{ref}}^\top\eta_h}\Big),$$

where $\ell=(\ell_1,\dots,\ell_{48})$ is the portfolio mean post-response load with $\ell_t=\frac{1}{H}\sum_{h=1}^{H}\big(\eta_{h,t}+(\eta_{h,t}+\epsilon)z_{h,t}\big)$, the smooth maximum, the smooth dispersion penalty, and the two-sided smooth band penalty are

$$\text{smax}_{\tau_p}(\ell)=\tau_p\log\Big(\frac{1}{48}\sum_{t=1}^{48}e^{\ell_t/\tau_p}\Big),\qquad \text{Disp}_{\delta_f}(\ell)=\frac{1}{48}\sum_{t=1}^{48}\delta_f^2\Big(\sqrt{1+\Big(\frac{\ell_t-\bar\ell}{\delta_f}\Big)^2}-1\Big),$$

with $\bar\ell=\frac{1}{48}\sum_{t=1}^{48}\ell_t$, and

$$\rho_{\tau_b}(u)=\tau_b\log\big(1+e^{(u-B_{\text{cap}})/\tau_b}\big)+\tau_b\log\big(1+e^{-(u+B_{\text{cap}})/\tau_b}\big),$$

with $B_{\text{cap}}>0$ the half-width of the free band around zero fractional bill change. Evaluating the bill on η rather than on the realized load $\eta+S(\eta)z$ makes f jointly Lipschitz in (x,z) on the whole space: the peak and dispersion terms depend on z through a Lipschitz map, while the bill term depends on x only through the bounded tariff $p_x(\xi)$. The three terms are normalized by their respective no-response reference values $\ell^{\text{peak}}_{\text{ref}}$, $\ell^{\text{flat}}_{\text{ref}}$ so that none dominates the others before weighting.

Combining f, g from Appendix B.2.4, $p_x(\xi)$ above, and $\mathbb{P}_{\eta|\xi}$ from Appendix B.2.3, the dynamic pricing problem is a concrete instance of (1) with upper-level decision $x\in\mathbb{R}^{d_x}$ (tariff policy), lower-level decision $z\in\mathbb{R}^{d_y}$ (portfolio-wide relative adjustments), context $\xi\in\mathbb{R}^6$, and data $\eta\in\mathbb{R}^{H\times 48}$. It is straightforward to verify from the displayed formulas of f and g, and the finite empirical scenario bank that the instance satisfies the C^∞-smoothness of f and g in Assumption 2.1(i), the samplewise strong convexity of g in (ii), the conditions on f and ∇f in items (iii)–(iv), and the boundedness from below of F in (v). However, the remaining conditions on g do not hold globally in z. Notice that $\nabla_x g(x,z;\eta,\xi)=J_{p_x(\xi)}(x)^\top\sum_h S(\eta_h)z_h$, so moving far in z makes both its x-Lipschitz constant and its conditional variance over $\eta\mid\xi$ grow without bound. Hence the g-dependent bounds of items (iii)–(iv) hold only on the bounded region reached by the experiment. On that region every no-response bill is strictly positive ($p_{\text{ref}}^\top\eta_h>0$), so the bill term is well defined. Therefore, the comparison is empirical.

It remains to verify the initialization-gap assumption in Assumption 2.2 with $y_k^0=0$. The gap $\bar g(x,0;\xi)-\bar g(x,z^\star(x;\xi);\xi)$ can be bounded directly because $\bar g(x,\cdot;\xi)$ is μ_g-strongly convex in z. Write c_r for the fitted coefficient of the linear term in the discomfort potential Φ_r in Appendix B.2.4, stack $c=(c_{r(1)}^\top,\dots,c_{r(H)}^\top)^\top\in\mathbb{R}^{48H}$, and let

$$\Delta_p:=\max\{|p_L-p_{\text{ref}}|,|p_H-p_{\text{ref}}|\},\qquad \bar s:=\sup_{\eta\text{ in the empirical support},\,h}\|S(\eta_h)\|.$$

The empirical scenario bank is finite, so $\bar s<\infty$. Since the sigmoid tariff $p_x(\xi)$ takes values in $[p_L,p_H]$, we have $\|p_x(\xi)-p_{\text{ref}}\|\le\sqrt{48}\,\Delta_p$ for every (x,ξ). The bill term of g is affine in z and the saturating term satisfies $\psi_\delta'(0)=0$, so $\nabla_z\bar g(x,0;\xi)$ consists of the linear coefficient c plus the z-independent bill gradient; hence

$$\|\nabla_z\bar g(x,0;\xi)\|\le G_0:=\|c\|+\sqrt{48H}\,\bar s\,\Delta_p\quad\text{uniformly in }(x,\xi).$$

Strong convexity of $\bar g(x,\cdot;\xi)$ gives

$$\bar g(x,0;\xi)-\bar g(x,z^\star(x;\xi);\xi)\le\frac{\|\nabla_z\bar g(x,0;\xi)\|^2}{2\mu_g}\le\frac{G_0^2}{2\mu_g}.$$

Therefore, Assumption 2.2 holds with $y_k^0=0$ and $b=G_0^2/(2\mu_g)$.

B.2.6 Problem and algorithmic parameters

The portfolio consists of $H=100$ households, so the tariff policy has dimension $d_x=30$ and the relative adjustments have dimension $d_y=4800$. The fitted potentials give a strong-convexity

modulus $\mu_g = 96.2$ and a gradient Lipschitz constant $L_g = 631.3$ for the lower level, and an upper-level Lipschitz constant $\ell_{f,1} = 516.4$. The smoothing parameters are $\tau_p = 0.05$ for the peak proxy, $\delta_f = 0.05$ for the dispersion penalty, and $\tau_b = 0.1$ for the bill band, which is free within $B_{\text{cap}} = 0.05$ of the no-response bill. The per-slot floor in $S(\eta)$ is $\epsilon_{\text{kwh}} = 0.1$ kWh, and $\epsilon = 10^{-4}$ regularizes the lower-level solves. The three objective weights in f are $\lambda_{\text{peak}} = 1.0$, $\lambda_{\text{flat}} = 1.0$, and $\lambda_{\text{bill}} = 5.0$. The bill term is a soft band penalty: it keeps the mean realized bill change moderate relative to the unregularized optimum but does not enforce the band for every household; Table 2 quantifies both baseline and realized bill changes and the pointwise feasibility of the learned tariffs.

All solvers start from $x_0 = 0$ and $y_0 = 0$ and use minibatches of 10 contexts with 10 lower-level and 10 upper-level scenario samples per context. Every method's parameters were tuned for good performance, and the iteration budgets K were chosen so that the four methods consume comparable wall-clock time (approximately 12 seconds per replication). Each K is the largest iteration count whose predicted per-iteration cost stays within that budget, rescaled after one timed pilot replicate. For Algorithm 1, we use outer stepsize $\alpha_k = 60.0/\sqrt{k+1}$, inner stepsize $\beta_t = 0.28/(\mu_g(t+1))$, $C_T = 0.002$, and $K = 964$ outer iterations. For Algorithm 2, we use $\alpha_0 = 4.0$, $\beta_t = 0.28/(\mu_g(t+1))$, truncation level $N = 7$, adaptive parameters $c_0 = 0.6$ and $a_1 = 0.2$, inner initialization $z_k^0 = 0$, and $K = 1802$. For the Hessian-based method, we use $L_{g,1} = 631.3$, RT-MLMC level $N = 6$, inner stepsize coefficient $\beta = 0.001$, outer stepsize $\alpha_t = 150.0\big/\big(\sqrt{t+1}\,(1+\max\{0, t-300\}/20)\big)$, and $K = 2013$. For reduction + stocBiO, the basis input is the policy feature vector with the constant intercept coordinate dropped, since the constant in Chebyshev basis is collinear with it. We consider two settings: degree 1, the smallest degree that can represent a context-dependent solution and is full-rank on the observed training contexts, giving $N = 6$ basis functions and a reduced lower-level variable $W \in \mathbb{R}^{4800\times 6}$; and degree 2, which is rank-deficient on the observed contexts (19 of 21) and is carried in rank-pruned, orthonormalized form (Table 1), giving $N = 19$. Both use $\beta = 15.0$, $D = 30$ inner iterations, and Neumann length $Q = 10$; the outer stepsizes are $\alpha = 2 \times 10^{-4}$ for degree 1 and $\alpha = 1.5 \times 10^{-3}$ for pruned degree 2. The two degrees have different per-iteration costs, so their budgets are set separately to match the same wall-clock time: $K = 2075$ (degree 1) and $K = 1600$ (degree 2). Only the pruned degree-2 variant is plotted in Figure 2 and Figure 5, since degree 1 is dominated by it on every reported metric. We keep degree 1 for reference in the diagnostics of Tables 1 and 2 and Figure 6. All results are averaged over 10 independent replications with fixed seeds for data construction, household selection, and clustering, so all methods see the same instance.

We examine the basis matrix of the Chebyshev parameterization on the 274 training contexts, whose full rank is required by the theoretical guarantees of the reduction framework; see Bouscary et al. (2026). Table 1 reports the number of basis functions, the rank, the nullity, and the condition number for the candidate bases. Keeping the intercept coordinate makes the constant Chebyshev member collinear with it (nullity one), whereas dropping the intercept gives the full-rank $N = 6$ degree-1 basis. Degree 2 adds thirteen independent directions over the observed contexts (rank 19 of 21) at the price of two redundant columns and an order-of-magnitude larger condition number: its columns span a strictly larger function space on the data, but the parameterization is rank-deficient there. We therefore carry degree 2 in a rank-pruned form: the two data-null directions are dropped and the remaining 19-dimensional span is remixed by the whitened thin-SVD factor of the basis matrix, so the mixed basis has an identity Gram matrix on the training contexts. We note that the pruned basis is not a new approximating family: it is the same Chebyshev tensor-product space, reparametrized in coordinates identifiable on the observed contexts and chosen to give the reduction method its most favorable conditioning. The results show that the pruned degree 2 fits the lower-level solution better than degree 1, so we use it in reduction+stocBiO. Both degrees, however, perform worse than the proposed methods. The bottleneck of the reduction method at this budget is therefore not the expressiveness of the basis but the conditioning of the reduced problem and the cost of its inner loop and Neumann series.

B.2.7 Evaluation metrics

As baselines we report no-response metrics under the reference tariff ($z = 0$) and those of the 2013 LCL trial tariff, which is obtained by solving the same household response model.

Let $\{x_k\}$ be the iterates produced by a solver. The 365 bank days are split at random into 274 training contexts, from which the solvers sample, and 91 held-out evaluation contexts; the split is

Table 1: Chebyshev basis rank on the 274 LCL training contexts. The condition number is that of the (mixed) basis Gram matrix on the training contexts.

basis	N	rank	nullity	condition number
degree 1, intercept kept	7	6	1	13.9
degree 1, intercept dropped	6	6	0	13.9
degree 2, intercept dropped	21	19	2	186
degree 2, pruned	19	19	0	1

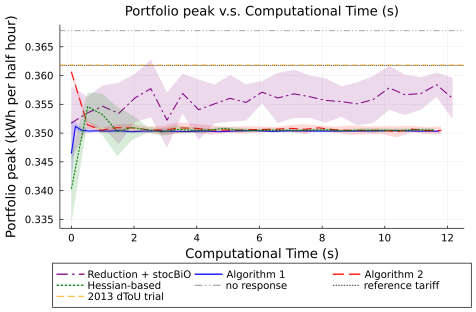
random rather than chronological because a chronological split would place different seasons on each side and turn the evaluation into seasonal extrapolation, and all bank days are free of dynamic-tariff treatment, whereas the trial-window days enter the response calibration. At each of 25 equally spaced evaluation points (including initialization), and for every evaluation context ξ, we compute the induced lower-level decision $\hat{z}(x_k;\xi)$ by gradient descent with Armijo backtracking on the exact conditional expectation $\bar{g}(x_k,\cdot;\xi)$, initialized at $z^0 = 0$, run until the exact-law gradient residual certifies $\|\hat{z} - z^\star(x_k;\xi)\| \le \|\nabla_z \bar{g}\|/\mu_g \le 0.01$. The exact law is available in closed form: $\bar{g}$ depends on the scenario η only through the mean response scale $\mathbb{E}[\eta \mid \xi] + \epsilon$, so a full-batch solve on the scenario stratum of ξ is exact. All reported metrics are therefore exact conditional expectations with no Monte Carlo error from the evaluation. We report the objective F; the peak load (kWh per half-hour), i.e., the maximum of the post-response portfolio mean load $\ell_t = \frac{1}{H}\sum_h(\eta_{h,t} + (\eta_{h,t} + \epsilon)\hat{z}(x_k;\xi)_{h,t})$ across $t = 1,\dots,48$, averaged over evaluation contexts and scenarios; the peak-to-average ratio (PAR), i.e., the ratio of that peak to the mean load $\frac{1}{48}\sum_t \ell_t$, again averaged over contexts and scenarios, with lower PAR indicating a flatter load profile; the mean fractional bill change $(p_x(\xi)^\top(\eta_h + S(\eta_h)\hat{z}) - p_{\mathrm{ref}}^\top\eta_h)/(p_{\mathrm{ref}}^\top\eta_h)$ across households, contexts, and scenarios, reported as an operational metric even though F evaluates the bill on η alone; and the mean absolute relative adjustment $\frac{1}{48H}\sum_{h,t}|\hat{z}_{h,t}|$ induced by the tariff, together with the pointwise extrema and feasibility statistics reported in Table 2. As a baseline reference, we report the no-response metrics (portfolio under the reference tariff with $z = 0$) and the metrics of the 2013 LCL dToU trial tariff, the latter obtained by solving the same household response model on the same evaluation contexts; this is an assessment of what the deployed 2013 schedule would achieve under the fitted response model, not a field trial of the optimized tariffs.

B.2.8 Additional results

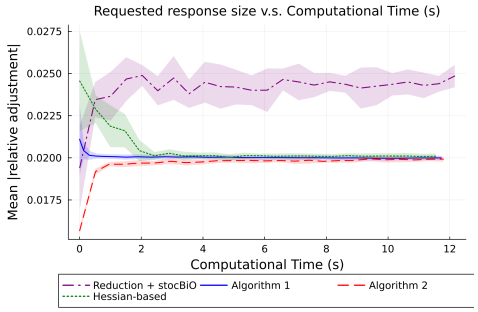
Figure 5 reports two operational metrics versus computational time, averaged over 10 independent runs: the portfolio peak load and the mean absolute relative adjustment requested of households. On peak load, the two proposed methods and the Hessian-based method finish at the same level, below the no-response level and below the 2013 trial tariff evaluated through the same household response model; the reduction method finishes between that level and the trial tariff, with a visibly larger variance. On the requested adjustment, the proposed methods and the Hessian-based method ask households for about two percent of their baseline load and the reduction method for about a quarter more, so no method relies on an implausibly large response. On both metrics the Hessian-based method arrives at the common final level later: the proposed methods reach it within about one second, whereas the Hessian-based method needs two to three seconds.

The bill term of f is a soft two-sided band penalty, so it does not enforce the $\pm 5\%$ band around the no-response bill. Table 2 separates, at the final iterates and under the exact conditional law of the 91 evaluation contexts, the fractional bill change evaluated on the no-response baseline (the quantity F penalizes) from the realized bill change of the post-response load. For the proposed methods the bill weight keeps the mean realized change moderate relative to the unregularized optimum, which roughly doubles bills, but about half of the households still exceed the cap in the realized response, so the penalty is not a per-household feasibility guarantee. Pointwise, the requested adjustments stay small and the rare negative post-response loads are marginal, whereas the reduction methods respond markedly more aggressively and their post-response loads become substantially negative.

Finally, we separate generalization across the context support from per-iteration cost. The LCL context is a five-dimensional continuous feature vector realized by 365 distinct bank days, and we retrain every solver on nested random subsamples of $n \in \{5, 10, 20, 40, 80, 160, 274\}$ training con-



(a) Portfolio peak load versus computational time.



(b) Mean relative adjustment versus computational time.

Figure 5: Contextual dynamic pricing for residential electricity demand response: additional operational metrics versus computational time. Left: portfolio peak load. Right: mean absolute relative adjustment. The horizontal dashed lines in the left panel mark the peak under no response (gray, dash-dot-dot) and under the 2013 LCL dynamic time-of-use trial tariff (orange, dashed). The reduction+stocBiO curve uses the pruned degree-2 Chebyshev basis. Shaded regions show mean ± 1 standard deviation over 10 independent runs.

Table 2: Bill changes and pointwise feasibility at the final iterates (exact conditional law, 91 evaluation contexts, mean over 10 replicates). "Base. Δ" and "real. Δ" are the mean fractional bill changes evaluated on the no-response baseline and on the realized post-response load; "base. share" and "real. share" are the fractions of households outside the $\pm 5\%$ band under the two evaluations; "min load" is the minimum post-response load over households, slots, and scenarios; "neg. load" is the fraction of household–slot–scenario entries whose post-response load is negative.

	base. Δ	base. share	real. Δ	real. share	min z	min load (kWh)	neg. load
Algorithm 1	$+11.1\%$	56.0%	$+6.5\%$	48.2%	-0.55	-0.013	0.91%
Algorithm 2	$+11.4\%$	56.1%	$+6.8\%$	48.3%	-0.56	-0.012	0.91%
Hessian-based	$+7.7\%$	50.7%	$+3.2\%$	43.3%	-0.48	-0.013	0.90%
Reduction deg 1	$+28.3\%$	67.6%	$+16.6\%$	57.6%	-1.26	-2.65	0.86%
Reduction deg 2	$+31.6\%$	70.0%	$+19.6\%$	60.0%	-0.70	-0.77	0.85%

texts (same tuned constants, three seeds), evaluating the final iterate on the full 91 held-out contexts. Here n counts distinct contexts ξ; the scenario draws η are taken conditionally on the drawn ξ exactly as in the main runs (ten per context), so this sweep changes only the pool of ξ the solvers can see, not the scenario budget per iteration. The results are shown in Figure 6. The proposed first-order methods are insensitive to how finely the support of ξ is sampled: their test objective moves by less than 0.01 between $n = 20$ and $n = 274$. The reduction method is governed by its fixed basis: at small cardinalities the fitted basis generalizes badly, and the pruned degree-2 variant is unstable enough that the solver diverges outright at $n = 5$ and $n = 10$ under constants tuned at full cardinality (its curve therefore starts at $n = 20$); as the support is covered, its accuracy saturates at the expressiveness floor of the basis, that is, at how well the basis can represent the lower-level solution at all, and this floor is lower for degree 2 than for degree 1. The measured curves only approach that floor: degree 2 is lower at $n = 80$ and $n = 160$, whereas at $n = 274$ the two degrees differ by less than the spread across seeds.

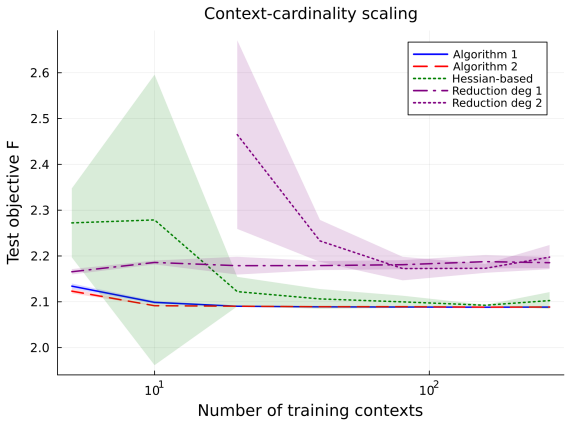


Figure 6: Context-cardinality scaling on LCL: test objective F of the final iterate versus the number of training contexts (nested subsamples, mean ± 1 std over 3 seeds, evaluation always on the full 91 held-out contexts). The pruned degree-2 curve starts at $n = 20$ since at $n = 5$ and $n = 10$ the solver diverges under the constants tuned at $n = 274$. Only the pool of distinct contexts ξ varies; the η draws per context are as in the main runs.