ZOBA: AN EFFICIENT SINGLE-LOOP ZEROTH-ORDER BILEVEL OPTIMIZATION ALGORITHM

Anonymous authors
Paper under double-blind review

ABSTRACT

Bilevel optimization problems consist of minimizing a value function whose evaluation depends on the solution of an inner optimization problem. These problems are typically tackled using first-order methods that require computing the gradient of the value function (the *hypergradient*). In several practical settings, however, first-order information is unavailable, rendering these methods inapplicable. Finite-difference methods provide an alternative by approximating hypergradients using function evaluations along a set of directions. However, such surrogates are notoriously expensive, and most finite-difference bilevel approaches rely on two-loop algorithms that are poorly parallelizable. Recently, several works have proposed single-loop finite-difference methods enabling parallelization. However, to ensure stability, these methods require a large number of function evaluations per iteration, limiting their applicability in high-dimensional settings with limited evaluation budget. In this work, we propose ZOBA, an efficient finite-difference single-loop algorithm for bilevel optimization. Our method leverages hypergradient approximations based on delayed information and function values reusage to enable parallelization while limiting the per-iteration cost. We analyze the proposed algorithm and establish convergence rates in the non-convex setting, achieving a complexity of $\mathcal{O}(p(d+p)^2\varepsilon^{-2})$, where p and d denote the dimensions of the inner and outer variables, respectively and $\varepsilon \in (0, 1)$ is the accuracy. This improves upon prior approaches based on Hessian approximation. We further introduce and analyze HF-ZOBA, a Hessian-free variant that yields optimal complexity. Finally, we corroborate our findings with numerical experiments on synthetic functions and a real-world black-box task in adversarial machine learning. Our results show that our methods achieve accuracy comparable to state-of-the-art techniques while requiring significantly less computation time.

1 INTRODUCTION

Bilevel optimization problems involve minimizing an outer objective (*the value function*) whose evaluation depends on the solution of an inner optimization problem. This class of problems arises naturally in a wide range of machine learning applications, including hyperparameter optimization (Bengio, 2000; Pedregosa, 2016), meta-learning (Franceschi et al., 2018a), representation learning (Huang et al., 2024), and many others. These problems are typically addressed using first-order algorithms that aim to compute the gradient of the value function, *the hypergradient* (Ghadimi & Wang, 2018; Ji et al., 2021; Chen et al., 2021; Dagréou et al., 2022; Arbel & Mairal, 2022; Yang et al., 2023). In practice, computing the hypergradient exactly is prohibitively expensive or infeasible, since at each iteration it requires the action of the Hessian (and its inverse) of the inner objective, as well as a minimizer of the inner problem. For this reason, first-order bilevel methods approximate the hypergradient by relying on an approximate inner solution and inverse-Hessian actions obtained through iterative procedures. This naturally leads to a two-loop algorithmic structure, where inner loops construct these approximations by exploiting gradients and, in some cases, Hessians of the inner and outer objectives. However, in many practical scenarios, such information is either unavailable or prohibitively expensive to compute, and only function evaluations of the inner and outer objectives are accessible. Such problems are referred to as black-box bilevel optimization problems and are particularly relevant, as they arise in many real-world applications such as meta-training, fine-tuning large language models (LLMs) and many others - see (Sinha et al., 2018) for a review. To tackle

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.

black-box bilevel problems, several classes of algorithms have been proposed (Camacho-Vallejo et al., 2024; Conn & Vicente, 2012; Jiang & Bolognani, 2026; Aghasi & Ghadimi, 2025), and finite-difference is one of these. Finite-difference methods are iterative procedures that mimic first-order strategies by replacing gradients with surrogates constructed from function evaluations along a set of directions (Nesterov & Spokoiny, 2017; Ghadimi & Lan, 2013; Rando et al., 2023; Duchi et al., 2015; Cai et al., 2022). Designing efficient finite-difference methods for bilevel optimization is particularly challenging. Approximating all first- and second-order quantities involved in the hypergradient computations with finite-difference is extremely expensive, especially in high-dimensional settings, where a large number of function evaluations is required to obtain reliable gradient and Hessian surrogates. Standard finite-difference methods for black-box bilevel optimization (Aghasi & Ghadimi, 2025; Aghasi et al., 2025) typically mimic the first-order two-loop structure and attempt to mitigate the resulting computational cost by using computationally cheap but inaccurate gradient and Hessian approximations in the inner loops. To control the resulting approximation errors, these methods have to employ small step sizes and perform many inner-loop iterations, leading to significant computational overhead and poor parallelization in practice. More recent approaches (Aghasi et al., 2025) aim to reduce this burden by regularization-based reformulations that avoid explicit Hessian approximation. While reducing per-iteration cost, these methods introduce bias into the bilevel objective and require careful tuning of regularization parameters, which may negatively affect solution accuracy. Moreover, the two-loop structure is preserved, thereby retaining the same limitations in terms of parallelization. Recent advances in first-order bilevel optimization have shown that the two-loop structure can be avoided by using single-loop algorithms that rely on approximate hypergradients computed using delayed information (Dagréou et al., 2022). These methods offer substantial practical advantages, as they allow all quantities involved in hypergradient computation to be evaluated in parallel and significantly reduce per-iteration computational cost while preserving convergence guarantees. However, extending these ideas efficiently to black-box setting is far from straightforward. In particular, it is unclear whether single-loop schemes can achieve meaningful computational savings while adequately controlling finite-difference errors to preserve convergence and maintain reasonable complexity guarantees. As result, current single-loop finite-difference methods for bilevel optimization (Nazari et al., 2025; Peng et al., 2025), while preserving convergence and optimal complexity, require many function evaluations per iteration to obtain reliable hypergradient approximations and stable updates - see Appendix B. This makes them not suitable for high-dimensional problems with limited evaluation budgets, as the budget can be exhausted in just a few iterations. In this work, we tackle this limitation and we propose ZOBA, an efficient single-loop finite-difference algorithm for black-box bilevel optimization. Our method relies on delayed information and enables full parallelization in the approximation of the hypergradient. Moreover, ZOBA tackles the approximation cost per iteration by reusing the same function evaluations to compute multiple components of the hypergradient, thereby significantly reducing the number of required function evaluations per iteration. We analyze the proposed algorithm and derive convergence guarantees for non-convex bilevel problems. In particular, we show that ZOBA achieves a complexity of $\mathcal{O}(p(d+p)^2\varepsilon^{-2})$, for an accuracy $\varepsilon \in (0,1)$ and where p and d denote the dimensions of the inner and outer variables, respectively, improving upon existing zeroth-order bilevel methods that rely on explicit Hessian approximations. We further introduce and analyze HF-ZOBA, a Hessian-free variant that avoids explicit Hessian approximations and we prove it achieves optimal complexity. Finally, we corroborate our theoretical findings through experiments on synthetic objectives and on the real-world task of computing minimal-distortion universal adversarial perturbations in black-box settings. The numerical results show that our methods achieve solution quality comparable to state-of-the-art finite-difference algorithms for bilevel optimization while providing significant improvements in runtime.

The paper is organized as follows. In Section 2, we formally define the problem, introduce our algorithms, and discuss their relationship with related work. In Section 3, we state and discuss the main theoretical results. Section 4 presents numerical experiments, and Section 5 concludes the paper with final remarks.

2 Problem Setting & Algorithm

Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ be probability spaces, and let $\mathcal{Z}_1, \mathcal{Z}_2$ be measurable spaces. Let $\zeta : \Omega_1 \to \mathcal{Z}_1$, $\xi : \Omega_2 \to \mathcal{Z}_2$ be two random variables and let $f : \mathbb{R}^p \times \mathbb{R}^d \times \mathcal{Z}_1 \to \mathbb{R}$ and $g : \mathbb{R}^p \times \mathbb{R}^d \times \mathcal{Z}_2 \to \mathbb{R}$ be two functions. Consider the minimization of a function $\Psi : \mathbb{R}^d \to \mathbb{R}$

defined as

$$\min_{x\in\mathbb{R}^d} \Psi(x) := F(z^*(x), x) = \mathbb{E}_\zeta[f(z^*(x), x, \zeta)]$$
$$\text{s.t.} \tag{1}$$
$$z^*(x) \in \arg\min_{z\in\mathbb{R}^p} G(z, x) = \mathbb{E}_\xi[g(z, x, \xi)].$$

Under standard regularity assumptions (Ghadimi & Wang, 2018), the function Ψ is differentiable, and its gradient (the *hypergradient*) admits the following representation

$$\nabla\Psi(x) = \nabla_x F(z^*(x), x) + \nabla^2_{xz} G(z^*(x), x), v^*(x), \tag{2}$$

where

$$v^*(x) = -\left[\nabla^2_{zz} G(z^*(x), x)\right]^{-1} \nabla_z F(z^*(x), x). \tag{3}$$

Here, ∇_x and ∇_z denote gradients with respect to x and z, respectively, while ∇^2_{xz} and ∇^2_{zz} denote the corresponding Hessian blocks. We consider the *stochastic zeroth-order* setting, in which no first- or higher-order information of f and g is available. Moreover, exact evaluations of F and G are inaccessible, and only noisy evaluations of f and g can be obtained. To address Problem eq. 1, we propose a stochastic zeroth-order algorithm, namely, an iterative procedure that relies solely on evaluations of f and g. At each iteration k, our method constructs a hypergradient surrogate at the current iterate x_k by approximating gradient and Hessian terms with mini-batch finite differences and using auxiliary sequences z_k and v_k to track $z^*(x_k)$ and $v^*(x_k)$. We next describe how these sequences are constructed (**auxiliary sequence construction**) and how the hypergradient is approximated (**hypergradient approximation**). We also provide, in Appendix A, a more intuitive derivation of the iterations together with a discussion of the motivations behind the chosen approximations.

Auxiliary sequence construction. At every iteration $k \in \mathbb{N}$, let $(\xi_{i,k})_{i=1}^{b_1}$ be realizations of independent copies of the random variable ξ. For each $i \in [b_1] := \{1, \dots, b_1\}$, let $\left(w_k^{(i,j)}\right)_{j=1}^{\ell_1} \subset \mathbb{R}^p$ be vectors sampled i.i.d. from $\mathcal{N}(0, I_p)$. For $h_k > 0$, let $g_{i,j}^{+,k} := g(z_k + h_k w_k^{(i,j)}, x_k, \xi_{i,k})$ and $g_{i,j}^{-,k} := g(z_k - h_k w_k^{(i,j)}, x_k, \xi_{i,k})$. Our algorithm approximates $\nabla_z G(z_k, x_k)$ with the following surrogate

$$D_z^k := \frac{1}{b_1\ell_1} \sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1} \frac{g_{i,j}^{+,k} - g_{i,j}^{-,k}}{2h_k} w_k^{(i,j)}. \tag{4}$$

Let $\rho_k > 0$ be a stepsize, we define the sequence $(z_k)_{k\geq 0}$ with the following iteration

$$z_{k+1} = z_k - \rho_k D_z^k. \tag{5}$$

Then, denoting with $g_i^k := g(z_k, x_k, \xi_{i,k})$ and $c_{i,j}^k := (g_{i,j}^{+,k} + g_{i,j}^{-,k} - 2g_i^k)/(2h_k^2)$, our method builds the surrogate of $\nabla^2_{zz} G(z_k, x_k)$ as $\hat{\nabla}^2_{zz} g_k := \frac{1}{b_1\ell_1} \sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1} c_{i,j}^k (w_k^{(i,j)} w_k^{(i,j)\top} - I_p)$. Notice that to compute this quantity we need to perform just b_1 evaluations since $g_{i,j}^{+,k}$ and $g_{i,j}^{-,k}$ are already computed for eq. 5. Let $b_2, \ell_2 > 0$ and let $(\zeta_{i,k})_{i=1}^{b_2}$ be realizations of independent copies of ζ. Let $f_{i,j}^{z,k,+} = f(z_k + h_k w_k^{(i,j)}, x_k, \zeta_{i,k})$ and $f_i^k := f(z_k, x_k, \zeta_{i,k})$. Our algorithm approximates $\nabla_z F(z_k, x_k)$ as $\hat{\nabla}_z f_k := \frac{1}{b_2\ell_2} \sum_{i=1}^{b_2}\sum_{j=1}^{\ell_2} \frac{f_{i,j}^{z,k,+} - f_i^k}{h_k} w_k^{(i,j)}$. Then, our procedure build the following search direction

$$D_v^k := \hat{\nabla}^2_{zz} g_k v_k + \hat{\nabla}_z f_k. \tag{6}$$

Thus, we construct the next auxiliary iterate v_{k+1} as

$$v_{k+1} = v_k - \rho_k D_v^k. \tag{7}$$

Hypergradient approximation. For $i \in [b_1], j \in [\ell_1]$, let $\hat{g}_{i,j}^{+,k} := g(z_k + h_k w_k^{(i,j)}, x_k + h_k u_k^{(i,j)}, \xi_{i,k})$, $\hat{g}_{i,j}^{-,k} = g(z_k - h_k w_k^{(i,j)}, x_k - h_k u_k^{(i,j)}, \xi_{i,k})$ and $s_{i,j}^k = (\hat{g}_{i,j}^{+,k} + \hat{g}_{i,j}^{-,k} - 2g_i^k)/(2h_k^2)$. Our method approximates $\nabla^2_{xz} G(z_k, x_k)$ as $\hat{\nabla}^2_{xz} g_k := \frac{1}{b_1\ell_1} \sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1} s_{i,j}^k u_k^{(i,j)} w_k^{(i,j)\top}$. Notice that such surrogate requires only $2b_1\ell_1$ function evaluations since $(g_i^k)_{i=1}^{b_1}$ are already computed for

previous quantities. Let $f_{i,j}^{x,k,+} = f(z_k, x_k + h_k u_k^{(i,j)}, \zeta_{i,k})$, our method approximates $\nabla_x F$ as $\hat{\nabla}_x f_k := \frac{1}{\ell_2 b_2} \sum_{i=1}^{b_2} \sum_{j=1}^{\ell_2} \frac{f_{i,j}^{x,k,+} - f_i^k}{h_k} u_k^{(i,j)}$. Notice that since $(f_i^k)_{i=1}^{b_2}$ are already computed for eq. 6, such approximation requires $b_2\ell_2$ function evaluations. The hypergradient surrogate is built as follows.

$$D_x^k = \hat{\nabla}_{xz}^2 g_k v_k + \hat{\nabla}_x f_k \tag{8}$$

Finally, the iterate is updated with the following iteration

$$x_{k+1} = x_k - \gamma_k D_x^k. \tag{9}$$

We summarize the entire procedure in Algorithm 1.

Algorithm 1 ZOBA: Zeroth-Order Bilevel Algorithm

1: **Input:** $x_0 \in \mathbb{R}^d$, $z_0, v_0 \in \mathbb{R}^p$, stepsizes (γ_k), (ρ_k), discretization parameter h_k, minibatch sizes b_1, b_2, number of directions ℓ_1, ℓ_2.
2: Let $b = \max(b_1, b_2)$ and $\ell = \max(\ell_1, \ell_2)$.
3: **for** $k = 0, 1, 2, \ldots$ **do**
4: Sample i.i.d. $\xi_{1,k}, \ldots, \xi_{b,k}$ and $\zeta_{1,k}, \ldots, \zeta_{b,k}$.
5: For each $i \in [b]$, sample $(w_k^{(i,j)})_{j=1}^{\ell} \subset \mathbb{R}^p$ and $(u_k^{(i,j)})_{j=1}^{\ell} \subset \mathbb{R}^d$ from $\mathcal{N}(0, I_p)$ and $\mathcal{N}(0, I_d)$.
6: Compute

$$z_{k+1} = z_k - \rho_k D_z^k, \quad v_{k+1} = v_k - \rho_k D_v^k, \quad x_{k+1} = x_k - \gamma_k D_x^k$$

Given initial guesses $z_0, v_0 \in \mathbb{R}^p$ and $x_0 \in \mathbb{R}^d$, at each iteration $k \in \mathbb{N}$, the algorithm gets two mini-batches of $b = \max(b_1, b_2)$ realizations $(\xi_{i,k})_{i=1}^b$ and $(\zeta_{i,k})_{i=1}^b$ of independent copies of ξ and ζ. For each $i \in [b]$, it then samples $\ell = \max(\ell_1, \ell_2)$ i.i.d. Gaussian directions $(w_k^{(i,j)})_{j=1}^{\ell}$ and $(u_k^{(i,j)})_{j=1}^{\ell}$ from $\mathcal{N}(0, I_p)$ and $\mathcal{N}(0, I_d)$. Using these samples, the algorithm constructs the search directions D_z^k, D_v^k, and D_x^k defined in eqs. 4, 6, and 8, and updates the sequences z_k, v_k, and x_k accordingly. The update rules are guided by the following intuitions. i) eq. 5 corresponds to a single step of a mini-batch finite-difference method applied to the inner objective $z \mapsto G(z, x_k)$, yielding a one-step approximation of $z^*(x_k)$; ii) eq. 7 is a single mini-batch finite-difference step applied to the quadratic objective associated to the linear system in eq. 3, using z_k in place of $z^*(x_k)$. iii) The iterate x_k is updated using the resulting hypergradient surrogate. Notice that the updates of the sequences are performed in parallel. This is due the use of delayed information. In particular, v_{k+1} is computed using z_k instead of z_{k+1}; since z_k depends on x_{k-1}, this can be interpreted as a one-step delay as z_k would be an approximation of $z^*(x_{k-1})$. While this delay induces additional error, it enables reuse the function evaluations employed in D_z^k to estimate the Hessian block $\nabla_{zz}^2 G(z_k, x_k)$, reducing its cost from $b_1(2\ell_1 + 1)$ to b_1 evaluations. A similar delay appears in the outer update, where x_{k+1} is computed using z_k and v_k instead of their updated counterparts. While this further propagates approximation error, it enables reuse of function evaluations in the computation of D_x^k. As a result, the total number of function evaluations per iteration is $b_1(4\ell_1 + 1) + b_2(2\ell_2 + 1)$.

Hessian-free variant. Algorithm 1 constructs hypergradient surrogates by explicitly approximating the Hessian blocks in eq. 2 via finite differences. Since accurate finite difference Hessian approximations requires a large number of function evaluations (Aghasi et al., 2025), we introduce a Hessian-free variant that avoids explicit Hessian approximation. Instead, we approximate Hessian–vector products using first-order finite differences and estimate the resulting gradients with finite-differences. Let $\hat{g}_{i,j}^{+,z,k}(z, x) = g(z + h_k w_k^{(i,j)}, x, \xi_{i,k})$, $\hat{g}_{i,j}^{+,x,k}(z, x) = g(z, x + h_k u_k^{(i,j)}, \xi_{i,k})$ and $\hat{g}_i^k(z, x) = g(z, x, \xi_{i,k})$. We define the following minibatch gradient approximation in variable z and x for arbitrary z, x as $\hat{\nabla}_z g_k(z, x) := \frac{1}{b_1 \ell_1} \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \frac{\hat{g}_{i,j}^{+,z,k}(z,x) - \hat{g}_i^k(z,x)}{h_k} w_k^{(i,j)}$ and $\hat{\nabla}_x g_k(z, x) := \frac{1}{b_1 \ell_1} \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \frac{\hat{g}_{i,j}^{+,x,k}(z,x) - \hat{g}_i^k(z,x)}{h_k} u_k^{(i,j)}$. Let $\bar{h} > 0$, we approximate $\nabla_{zz}^2 G(z_k, x_k) v_k$ and $\nabla_{xz}^2 G(z_k, x_k) v_k$ with $H_{zz}^k := \frac{1}{\bar{h}_k}(\hat{\nabla}_z g_k(z_k + \bar{h}_k v_k, x_k) - \hat{\nabla}_z g_k(z_k, x_k))$ and $H_{xz}^k :=$

$\frac{1}{h_k}(\hat{\nabla}_x g_k(z_k + \bar{h}_k v_k, x_k) - \hat{\nabla}_x g_k(z_k, x_k))$ respectively. Using these approximations, we define the search directions

$$\hat{D}_z^k = \hat{\nabla}_z g(z_k, x_k), \quad \hat{D}_v^k = H_{zz}^k + \hat{\nabla}_z f(z_k, x_k), \quad \hat{D}_x^k = H_{xz}^k + \hat{\nabla}_x f(z_k, x_k). \tag{10}$$

Thus, we define the following algorithm

Algorithm 2 HF-ZOBA: Hessian-free Zeroth-Order Bilevel Algorithm

1: **Input:** $x_0 \in \mathbb{R}^d$, $z_0, v_0 \in \mathbb{R}^p$, stepsizes (γ_k), (ρ_k), smoothing h_k, minibatches b_1, b_2, directions ℓ_1, ℓ_2.
2: Let $b = \max(b_1, b_2)$ and $\ell = \max(\ell_1, \ell_2)$.
3: **for** $k = 0, 1, 2, \ldots$ **do**
4: Sample i.i.d. $\xi_{1,k}, \ldots, \xi_{b,k}$ and $\zeta_{1,k}, \ldots, \zeta_{b,k}$.
5: For each $i \in [b]$, sample $(w_k^{(i,j)})_{j=1}^{\ell} \subset \mathbb{R}^p$ and $(u_k^{(i,j)})_{j=1}^{\ell} \subset \mathbb{R}^d$ from $\mathcal{N}(0, I_p)$ and $\mathcal{N}(0, I_d)$.
6: Compute

$$z_{k+1} = z_k - \rho_k \hat{D}_z^k, \quad v_{k+1} = v_k - \rho_k \hat{D}_v^k, \quad x_{k+1} = x_k - \gamma_k \hat{D}_x^k$$

The Hessian-free variant (HF-ZOBA) preserves the structure of Algorithm 1, differing only by replacing the search directions with the ones in eq. 10. The sequences z_k, v_k, and x_k are updated in parallel, and delayed information is exploited to reuse function evaluations across all updates. In particular, function values used to compute $\hat{\nabla}_z g_k$ and $\hat{\nabla}_z f_k$ are reused in H_{zz}^k, H_{xz}^k and $\hat{\nabla}_x f_k$. As a result, HF-ZOBA requires $2b_1(2\ell_1 + 1) + b_2(2\ell_2 + 1)$ function evaluations per iteration, which is higher than ZOBA, but without explicit Hessian estimation. In the next section, we review the related works and compare our algorithms with the state-of-the-arts methods.

2.1 Related Works

Several algorithms have been proposed for black-box bilevel optimization. We review the most related ones, highlighting differences with our methods, and briefly summarize other approaches.

Finite-difference methods. Several finite-difference approaches have been proposed for bilevel optimization. Gu et al. (2021); Staudigl et al. (2026) propose two algorithms for bilevel optimization. However, Staudigl et al. (2026) assume access to an (inexact) oracle that provides an approximation of z^*, while Gu et al. (2021) assume access to exact function values and an approximation of z^* obtained via first-order methods. Shirkavand et al. (2025) propose a finite-difference bilevel method for LLMs fine-tuning. However, their approach assumes access to gradients of the inner target, which are unavailable in our setting. Aghasi & Ghadimi (2025) propose ZDSBA, a method that approximates gradients and Hessians using single-sample, single-direction finite difference surrogates. The authors analyze their approach and obtain a complexity of $\mathcal{O}\left(\frac{(d+p)^4}{\varepsilon^3}\log\frac{d+p}{\varepsilon}\right)$ in non-convex setting. This is improved in ZMDSBA (Aghasi et al., 2025) which extends ZDSBA by using mini-batches to approximate $\nabla_x F$ and $\nabla_{xz}^2 G$, yielding a complexity of $\mathcal{O}\left(\frac{p(d+p)^2}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$. However, both our algorithms achieve better complexity. Moreover, ZDSBA and ZMDSBA rely on a two-loop structure, making them poorly parallelizable. Aghasi et al. (2025) also propose a Hessian-free variant (Opt-ZMDSBA) achieving the optimal complexity of $\mathcal{O}\left((d+p)\varepsilon^{-2}\right)$. While HF-ZOBA matches this complexity, our approach offers several advantages. First, Opt-ZMDSBA is a two-loop algorithm, and therefore it is poorly parallelizable. Second, Opt-ZMDSBA relies on a regularization-based reformulation that transforms the bilevel problem into a single-level problem, introducing a regularization parameter λ that must be carefully tuned to achieve accurate solutions. In contrast, HF-ZOBA operates directly on the original bilevel problem and it is fully parallelizable due to its single-loop design. Peng et al. (2025); Nazari et al. (2025) propose two single-loop finite-difference algorithms, both achieving optimal complexity. However, in Peng et al. (2025), surrogates are constructed using all canonical basis vectors as directions, requiring a large number of function evaluations per iteration, especially in high-dimensional settings. In Nazari et al. (2025), surrogates are built from batches of single-direction finite-difference estimates with directions sampled from the sphere. However, due to the algorithm design, the number of function evaluations per iteration

scales poorly with the batch size. As a result, both methods may perform poorly in high-dimensional settings with a limited number of function evaluations (*budget*). This is because their iteration cost quickly exhausts the available budget, allowing only a few iterations and thus providing limited improvement. In contrast, HF-ZOBA achieves the same complexity while requiring fewer function evaluations per iteration than Peng et al. (2025) and exhibiting better scaling with respect to the batch size than Nazari et al. (2025). This leads to improved practical performance (see Section 4) and shows that optimal complexity can be achieved with tighter control of the per-iteration cost - An extended comparison with these methods is provided in Appendix B.

Other methods. Several other strategies have been proposed, including trust-region methods for bilevel optimization (Conn & Vicente, 2012; Ehrhardt & Roberts, 2021), which either focus on the deterministic setting with exact function evaluations without providing convergence guarantees (Conn & Vicente, 2012), or require exact function values and first-order information of the inner objective (Ehrhardt & Roberts, 2021). Other strategies proposed are meta-heuristics. These are based on genetic algorithms (Li & Wang, 2007), simulated annealing (Yu et al., 2010), particle swarm optimization (Gao et al., 2011), and other techniques - see (Camacho-Vallejo et al., 2024) for a review. While these approaches perform well in practice, they typically lack convergence guarantees.

3 Main Results

In this section, we analyze Algorithms 1 and 2, providing convergence rates for different choices of parameters. More precisely, we provide bounds on the average squared norm of the gradient of the value function Ψ. We use the notation

$$\alpha_K := \left(\sum_{k=0}^{K} \gamma_k \, \mathbb{E}\Big[\|\nabla \Psi(x_k)\|^2\Big]\right) \Big/ A_K \quad \text{with} \quad A_K := \sum_{k=0}^{K} \gamma_k.$$

In the following, we call *complexity* the number of function evaluations required to get $\alpha_K \leq \varepsilon$ with $\varepsilon \in (0,1)$. Moreover, we define $R_{\text{init}} := \Psi(x_0) - \min \Psi + \phi_z^0 \|z_0 - z^*(x_0)\|^2 + \phi_v^0 \|v_0 - v^*(x_0)\|^2$, where $\phi_z^0, \phi_v^0 > 0$ are constants. To perform the analysis, we consider targets satisfying the following hypothesis.

Assumption 3.1 (Unbiasedness and Bounded Variance)**.** Stochastic gradients and Hessians of inner and outer targets are unbiased estimators, i.e $\nabla F(z,x) = \mathbb{E}_\zeta \left[\nabla f(z,x,\zeta)\right]$, $\nabla G(z,x) = \mathbb{E}_\xi \left[\nabla g(z,x,\xi)\right]$ and $\nabla^2 G(z,x) = \mathbb{E}_\xi \left[\nabla^2 g(z,x,\xi)\right]$. Moreover, there exist constants $\sigma_{1,G}, \sigma_{1,F} \geq 0$ s.t. for every $z \in \mathbb{R}^p$ and $x \in \mathbb{R}^d$,

$$\mathbb{E}_\zeta[\|\nabla f(z,x,\zeta) - \nabla F(z,x)\|^2] \leq \sigma_{1,F}^2 \quad \text{and} \quad \mathbb{E}_\xi[\|\nabla g(z,x,\xi) - \nabla G(z,x)\|^2] \leq \sigma_{1,G}^2.$$

Assumption 3.2 (Regularity of outer objective)**.** There exist $L_{0,F}, L_{1,F} > 0$ such that for every $\zeta \in \mathcal{Z}_1$ the function $(z,x) \mapsto f(z,x,\zeta)$ is $L_{0,F}$-Lipschitz continuous, differentiable and its gradient is $L_{1,F}$-Lipschitz continuous.

Assumption 3.3 (Regularity of inner objective)**.** There exist $L_{1,G}, L_{2,G}, \mu_G > 0$ such that, for every $\xi \in \mathcal{Z}_2$ the function $(z,x) \mapsto g(z,x,\xi)$ is twice differentiable, its gradient is $L_{1,G}$-Lipschitz and its Hessian is $L_{2,G}$-Lipschitz. Moreover, for every $x \in \mathbb{R}^d$, the function $z \mapsto g(z,x,\xi)$ is μ_G-strongly convex.

These assumptions are standard in stochastic zeroth-order bilevel optimization and has been adopted in prior work (Aghasi & Ghadimi, 2025; Aghasi et al., 2025; Peng et al., 2025; Nazari et al., 2025). In particular, Assumption 3.1 is used to control the stochastic error, while Assumptions 3.2 and 3.3 ensure that the hypergradient is well-defined and Lipschitz continuous.[1] Unlike some existing approaches (Aghasi & Ghadimi, 2025; Aghasi et al., 2025), which additionally assume boundedness of the optimal solution of the inner problem, our analysis does not require such an assumption. We next present convergence rates under this regime for both algorithms. For readability, explicit constants of theorems and corollaries are reported in the proofs - see Appendices D and E.

Theorem 3.4 (Convergence Rate of ZOBA)**.** *Let Assumptions 3.1,3.2,3.3 hold. For every $k \in \mathbb{N}$, let z_k, v_k, x_k be the sequences generated by Algorithm 1. For every*

[1] For HF-ZOBA, sample-wise strong convexity of g can be relaxed to strong convexity in expectation without affecting the proof.

$\bar{\phi}_z > 0$, *let* $\rho_k < \min\left(1, \frac{b_1\ell_1\mu_G}{4(2b_1\ell_1+p+2)L_{1,G}^2(\mu_G+4)}, \frac{\mu_G}{2\bar{C}_v}\right)$ *with* $\bar{C}_v = \mathcal{O}(\frac{p(p+6)^2}{b_1\ell_1})$, $h_k = \mathcal{O}\left(\min\left(\sqrt{\frac{b_1\ell_1+p^3}{p^4}}, \frac{1}{\sqrt{p^4+pd^3}}, d^{-5/2}\right)\right)$ *and* $\gamma_k = \mathcal{O}(\bar{\phi}_z\rho_k^2)$ *Then, for* $K > 0$,

$$\alpha_K \leq \frac{2}{A_K}\left(R_{init} + \sum_{k=0}^{K}\bar{C}_k\right),$$

where $\bar{C}_k \geq 0$ *is a sequence of errors depending on* γ_k, ρ_k, h_k *defined in the proof.*

In the following corollary, we derive rates for specific choices of the parameters.

Corollary 3.5. *Under same assumptions of Theorem 3.4.*
(I) Let $\gamma_k = \gamma$, $\rho_k = \rho$, *and* $h_k = h$, *we have*

$$\alpha_K \leq \frac{R_{init}}{\gamma(K+1)} + \mathcal{O}\left(\frac{\rho^3+\rho^2}{\gamma} + \gamma\right) + \mathcal{O}\left(\frac{\rho+\rho^2+\rho^3}{\gamma} + \gamma\right)h^2.$$

(II) Let $\gamma_k = \gamma(k+1)^{-\alpha_1}$, $\rho_k = \rho(k+1)^{-\alpha_2}$ *and* $h_k = h(k+1)^{-\alpha_3}$ *with* $\gamma, \rho, h > 0$ *and* $\alpha_1, \alpha_2 \in (1/2, 1)$ *with* $\alpha_1 \geq 2\alpha_2$ *and* $\alpha_3 > 1$,

$$\alpha_K \leq \frac{R_{init}}{\gamma(K+1)^{1-\alpha_1}} + \mathcal{O}\left(\frac{1}{(K+1)^{1-\alpha_1}}\right).$$

(III) Fix an accuracy $\varepsilon \in (0,1)$. *Let* $D_1 := \frac{p(d+p)^2}{b_1\ell_1} + \frac{d}{b_2\ell_2}$ *and* $D_2 := \frac{p^3}{b_1\ell_1} + \frac{p}{b_2\ell_2}$. *For* $K+1 \geq \mathcal{O}\left(D_1 R_{init}\,\varepsilon^{-2}\right)$, *let* $\rho = \mathcal{O}\Big(\varepsilon\left(D_1D_2\right)^{-1/2}\Big)$ $\gamma = \mathcal{O}\left(\sqrt{\frac{1}{D_1(K+1)} + \frac{\varepsilon^2 D_2}{D_1^2}}\right)$, *and* $h = \mathcal{O}\Bigg(\sqrt{\varepsilon}\left(1+\sqrt{\frac{D_1}{D_2}}\right)^{-1/2}\Bigg)$. *Let* $b_1, \ell_1 = \mathcal{O}(\lfloor p(d+p)^2/c_1\rfloor)$ *and* $b_2\ell_2 = \mathcal{O}(\lfloor p(d+p)^2/c_2\rfloor)$ *for* $c_1, c_2 \geq 1$ *such that* $c_1, c_2 \leq p(p+d)^2$. *Then,* $\alpha_K \leq \varepsilon$ *and the complexity is* $\mathcal{O}\left(p(d+p)^2\varepsilon^{-2}\right)$.

In the following theorem and corollary, we show the rate for Algorithm 2 and we specialize the result with different choices of parameters.

Theorem 3.6 (Convergence Rate of HF-ZOBA)**.** *Under Assumptions 3.1,3.2,3.3. Let* $\hat{h}_k, \bar{\phi}_z > 0$, $\rho_k < \min\left(1, \frac{b_1\ell_1}{64\,L_{1,G}^2\,(b_1\ell_1+3p+6)}\right), \gamma_k < \mathcal{O}(\bar{\phi}_z\rho_k^2)$ *and* $\bar{h}_k = \begin{cases}\hat{h}_k/\|v_k\|, & \|v_k\| \neq 0\\ \hat{h}_k, & \text{otherwise}\end{cases}$. *Let* $h_k \leq \frac{\mu_G}{4\sqrt{(\mu_G+2)\bar{C}_L}}\sqrt{\bar{\phi}_v}\hat{h}_k$ *where* $\bar{C}_L = \mathcal{O}(\frac{(p+6)^3+(d+6)^3}{b_1\ell_1})$, $\bar{\phi}_v = \frac{\mu_G^2}{16\omega_{v,1}^2(\mu_G+4)}\bar{\phi}_z$, *and* $\omega_{v,1} = \left(L_{1,F} + \frac{L_{0,F}L_{2,G}}{\mu_G}\right)$. *Let* z_k, v_k, x_k *be the sequences generated by Algorithm 2. Then, for* $K > 0$,

$$\alpha_K \leq \frac{2}{A_K}\left(R_{init} + \sum_{k=0}^{K}\hat{C}_k\right).$$

Where $\hat{C}_k \geq 0$ *is a sequence of errors depending on* γ_k, ρ_k, h_k *defined in the proof.*

Corollary 3.7. *Under same assumptions of Theorem 3.6.*
(I) Let $\gamma_k = \gamma$, $\rho_k = \rho$, $h_k = h$ *and* $\hat{h}_k = \hat{h}$. *Then,*

$$\alpha_K \leq \frac{R_{init}}{\gamma(K+1)} + \mathcal{O}(\frac{\gamma^2+\rho^3}{\gamma}) + \mathcal{O}(\frac{\gamma+\rho^2+\gamma^2}{\gamma})\hat{h}^2 + \mathcal{O}(\frac{\gamma+\rho^2+\rho^3}{\gamma})h^2 + \mathcal{O}(\frac{\rho^2+\gamma^2}{\gamma})\frac{h^2}{\hat{h}^2}$$

(II) Let $\gamma_k = \gamma(k+1)^{-\alpha_1}$, $\rho_k = \rho(k+1)^{-\alpha_2}$, $h_k = h(k+1)^{-\alpha_3}$ *and* $\hat{h}_k = \hat{h}(k+1)^{-\alpha_4}$ *with* $\gamma, \rho, h, \hat{h} > 0$, $\alpha_1, \alpha_2 \in (1/2, 1)$ *with* $\alpha_1 \geq 2\alpha_2$, $\alpha_3, \alpha_4 > 1/2$ *and* $\alpha_3 - \alpha_4 \geq 0$. *Then,*

$$\alpha_K \leq \frac{R_{init}}{\gamma(K+1)^{1-\alpha_1}} + \mathcal{O}\left(\frac{1}{(K+1)^{1-\alpha_1}}\right).$$

(III) Fix an accuracy $\varepsilon \in (0,1)$. *Let* $\bar{\Delta}_4 = \mathcal{O}(\frac{d+p}{b_1\ell_1} + \frac{d+p}{b_2\ell_2})$ *and* $\bar{c}_\gamma = \mathcal{O}((b_1\ell_1)/(d+p))$. *For* $K+1 \geq \frac{32R_{init}\bar{\Delta}_4}{\bar{c}_\gamma}\varepsilon^{-2}$, *let* $\rho = \sqrt{\frac{2R_{init}}{\bar{c}_\gamma\bar{\Delta}_4(K+1)}}$, $\gamma = \hat{c}_\gamma\rho$. *Let* $h = \hat{h}^2$ *and* $\hat{h} = \mathcal{O}\left(\min\left(1, \sqrt{\frac{\varepsilon b_1\ell_1}{d^3+p^3}}\right)\right)$. *Let* $b_1\ell_1 = \mathcal{O}(\lfloor(d+p)/c_1\rfloor)$ *and* $b_2\ell_2 = \mathcal{O}(\lfloor(d+p)/c_2\rfloor)$ *for every* $c_1, c_2 \geq 1$ *such that* $c_1, c_2 \leq d+p$. *Then,* $\alpha_K \leq \varepsilon$ *and the complexity is* $\mathcal{O}((d+p)\varepsilon^{-2})$.

Discussion. The bounds in Theorems 3.4 and 3.6 depend on two terms: the initialization and an error component. Such errors arise from the use of stochastic information, delayed information, zeroth-order approximations of stochastic gradients and Hessians, and (in case of HF-ZOBA) the first-order finite-difference approximation of Hessian-vector products. In Corollaries 3.5 and 3.7 we provide two parameter regimes. The first (I) uses constant stepsizes, which is the most common choice in practice, while the second (II) ensures asymptotic convergence. Under choice (I), both algorithms converge to a neighborhood of a stationary point where the size is determined by stochastic variance (the term depending only on γ and ρ) and finite-difference bias (i.e. the term depending on h), consistent with classical results for stochastic zeroth-order methods - see e.g. (Nesterov & Spokoiny, 2017; Ghadimi & Lan, 2013). For HF-ZOBA, this neighborhood size further depends on the accuracy of the Hessian-vector product approximation, as reflected by the $\hat{h}$ and $h/\hat{h}$ terms, which respectively represent the finite-difference bias of the first-order Hessian surrogate and the zeroth-order approximation error of the gradients used within it. For both algorithms, expressing the outer stepsize γ as a fraction of ρ makes explicit that taking ρ (and h) smaller shrinks the size of such neighborhoods but slows the decay of the initialization term through its $1/\gamma$ dependence mirroring the behavior of first-order stochastic methods with constant stepsize. In (III) we derive the complexities of our algorithms. We observe that both methods improve upon ZDSBA (Aghasi & Ghadimi, 2025) which attains a complexity of $\mathcal{O}(\frac{(d+p)^4}{\varepsilon^3}\log(\frac{d+p}{\varepsilon}))$ and ZMDSBA (Aghasi et al., 2025), which achieves a complexity of $\mathcal{O}(\frac{p(d+p)^2}{\varepsilon^2}\log(\frac{1}{\varepsilon}))$. As in ZMDSBA, ZOBA retains the $\mathcal{O}(p(d+p)^2)$ term, reflecting the known dimensional dependence induced by explicit Hessian approximations (Aghasi et al., 2025). Moreover, HF-ZOBA achieves the optimal complexity as Opt-ZMDSBA (Aghasi et al., 2025), VRZSBO (Peng et al., 2025), and ZOSOGD (Nazari et al., 2025). However, Opt-ZMDSBA is a two-loop method with limited parallelization, and VRZSBO and ZOSOGD have worse per-iteration cost scaling - see Appendix B for further details. This shows that optimal complexity can be retained under these additional approximations, while also enabling better practical performance.

4 Experiments

In this section, we evaluate the empirical performance of the proposed methods. We compare ZOBA and HF-ZOBA with ZDSBA (Aghasi & Ghadimi, 2025), ZMDSBA, Opt-ZMDSBA (Aghasi et al., 2025), ZOSOGD (Nazari et al., 2025) and VRZSBO (Peng et al., 2025) on two tasks: a synthetic bilevel problem and a minimal universal black-box adversarial attack. Hyperparameters are tuned via grid search, and all methods are run with a fixed budget of 10^5 function evaluations. Experiments are repeated 10 times, and we report mean and standard deviation of the results. Additional details and experiments are provided in Appendix C.

Synthetic problem. Here, we evaluate the proposed methods on a synthetic quadratic bilevel problem. We perform experiments with $d = p \in \{25, 50, 100\}$. In Figure 1 (top), we report the function value

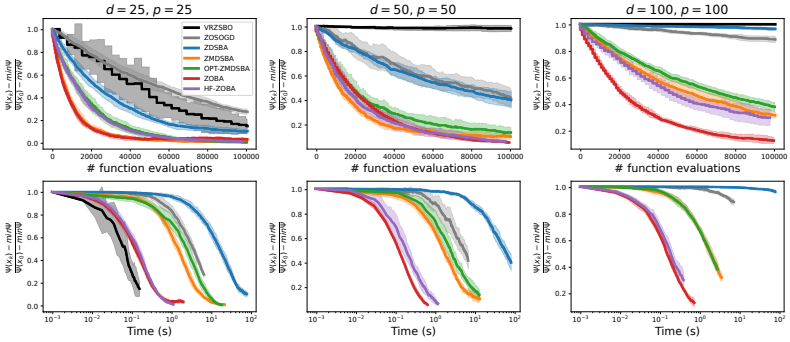


Figure 1: Comparison of algorithms on synthetic bilevel problems with varying dimensions: optimality gap versus function evaluations (top row) and wall-clock time in seconds (bottom row).

gap $(\Psi(x_k) - \min \Psi)/(\Psi(x_0) - \min \Psi)$ as a function of the number of function evaluations. We observe that ZDSBA, ZOSOGD, and VRZSBO perform worst across all problem sizes. For ZDSBA, this is consistent with its theoretical complexity. For VRZSBO, this is because it relies on expensive

estimators, while for ZOSOGD it is due to its poor cost scaling with respect to the batch size, which forces the use of small batch sizes and, thus, small step sizes. For $d = 25$ and $d = 50$, our methods match the performance of ZMDSBA and Opt-ZMDSBA, while for $d = 100$ they provide better performance. This improvement is due to the fact that ZMDSBA and Opt-ZMDSBA rely on inner-loop procedures to approximate $z^\star$ and $v^\star$ using single-sample, single-direction estimators, requiring many iterations (and more function evaluations) to match the accuracy achieved in a single step by multi-direction finite-difference methods, as in our methods, consistently with prior observations (Duchi et al., 2015; Salimans et al., 2017; Cai et al., 2021; 2022; Rando et al., 2023; 2025a;b). In Figure 1 (bottom), we report results versus wall-clock time. We observe that our methods are significantly faster due to parallelization and function evaluation reuse, unlike ZDSBA, ZMDSBA, and Opt-ZMDSBA, which rely on sequential inner loops. Notice that VRZSBO is faster at $d = 25$ but less accurate, and its performance degrades in higher dimensions due to its high per-iteration cost. ZOSOGD performance also deteriorates in high dimensions due to poor scaling with batch size. In contrast, our methods achieve strong performance by limiting per-iteration cost and exploiting parallelism, making them well-suited for limited-budget settings - see Appendix C.3 for additional comments and experiments.

Minimal distortion universal perturbation. We evaluate our methods on finding minimal-distortion universal adversarial perturbations (MD-UAP) for a multiclass MNIST classifier (Deng, 2012) through a bilevel subspace attack - see Appendix C.2 for details. We set the inner dimension to $p = 100$ and we target images labeled "4". From Figure 2 (top), we observe that all methods exhibit an initial increase in distortion before reaching a plateau. This behavior is expected, as the initial adversarial perturbation is small and must increase in norm to effectively reduce classification accuracy. We further observe that VRZSBO shows no improvement, which is due to its high per-iteration cost that exceeds the evaluation budget in this setting. ZOSOGD performs worse than competing methods, as its unfavorable cost scaling forces the use of small batch sizes and step sizes, limiting its effectiveness. ZOBA achieves performance comparable to other methods but is slightly inferior to ZMDSBA, since obtaining stable Hessian approximations requires a large number of function evaluations. In contrast, HF-ZOBA performs better overall, as it requires fewer evaluations to accurately approximate Hessian-vector products. In Figure 2 (bottom), results are shown versus wall-clock time. We observe that, thanks to parallelization and function value reuse, our algorithms achieve substantially better runtime performance than the other methods.

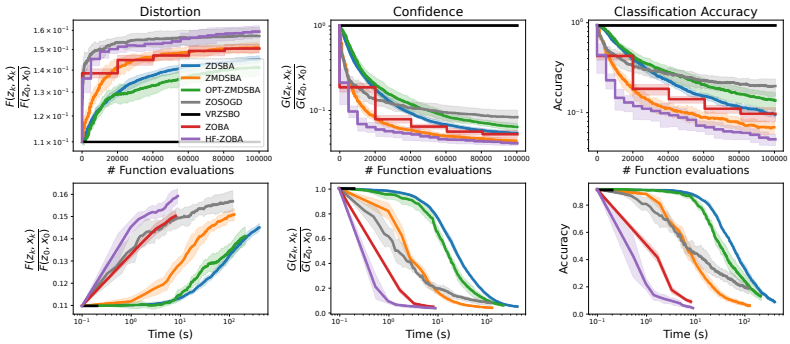


Figure 2: Normalized distortion and confidence on iterates and classification accuracy as function of number of function evaluations (top) and versus wall-clock time in seconds (bottom).

5 Conclusion

In this work, we proposed ZOBA, an efficient single-loop algorithm for zeroth-order bilevel optimization, and its Hessian-free variant HF-ZOBA. We established convergence guarantees and showed that HF-ZOBA achieves optimal complexity. Our experiments show that our methods provide clear practical advantages in wall-clock time over current state-of-the-art methods, making them particularly suitable for high-dimensional problems with limited budget. Future directions include extending the analysis to non-smooth settings, relaxing strong convexity assumptions, and developing adaptive parameter selection strategies.

AI USE STATEMENT

In this work, we used generative AI tools for grammar polishing, text editing and coding utility scripts (e.g. bash scripts for launching experiments). We have not used generative AI tools for deriving theoretical results and algorithm implementations. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not raise any significant ethical concerns. The paper focuses primarily on methodological developments in optimization and does not involve human subjects, personal data, or sensitive information. Although we include an experiment on universal adversarial perturbation generation, the experiment is conducted exclusively on the MNIST dataset, which contains handwritten digits and does not involve personal or sensitive data. We therefore do not anticipate any significant ethical implications arising from the methods or experiments presented in this work.

REPRODUCIBILITY STATEMENT

We provide the source code implementing our proposed methods and reproducing the experiments as supplementary material. The main paper provides the necessary assumptions and algorithmic details, while complete proofs of the theoretical results are provided in the Appendices D and E. Additional experiments and details on the experimental setup, including datasets, hyperparameters, and other details, are also provided in the Appendix C.

REFERENCES

Alireza Aghasi and Saeed Ghadimi. Fully zeroth-order bilevel programming via gaussian smoothing. *Journal of Optimization Theory and Applications*, 205(2):31, Mar 2025. ISSN 1573-2878. doi: 10.1007/s10957-025-02647-y.

Alireza Aghasi, Jeongyeol Kwon, and Saeed Ghadimi. Optimal zeroth-order bilevel optimization, 2025.

Michael Arbel and Julien Mairal. Amortized implicit differentiation for stochastic bilevel optimization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Jens Bayer, Stefan Becker, David Münch, Michael Arens, and Jürgen Beyerer. Traversing the subspace of adversarial patches. *Machine Vision and Applications*, 36(3):70, Apr 2025. ISSN 1432-1769. doi: 10.1007/s00138-025-01689-6.

Yoshua Bengio. Gradient-based optimization of hyperparameters. *Neural Computation*, 12(8): 1889–1900, 2000. doi: 10.1162/089976600300015187.

A. S. Berahas, L. Cao, K. Choromanski, and K. Scheinberg. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Foundations of Computational Mathematics*, 22(2):507–560, Apr 2022. ISSN 1615-3383. doi: 10.1007/s10208-021-09513-z.

Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019.

Hanqin Cai, Yuchen Lou, Daniel Mckenzie, and Wotao Yin. A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1193–1203, 18–24 Jul 2021.

HanQin Cai, Daniel McKenzie, Wotao Yin, and Zhenliang Zhang. Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling. *SIAM Journal on Optimization*, 32(2):687–714, 2022. doi: 10.1137/21M1392966.

José-Fernando Camacho-Vallejo, Carlos Corpus, and Juan G. Villegas. Metaheuristics for bilevel optimization: A comprehensive review. *Computers & Operations Research*, 161:106410, 2024. ISSN 0305-0548. doi: https://doi.org/10.1016/j.cor.2023.106410.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017. doi: 10.1109/SP.2017.49.

P.Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.J. Hsi. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, pp. 15–26, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450352024. doi: 10.1145/3128572.3140448.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 25294–25307. Curran Associates, Inc., 2021.

Andrew R Conn and Luís Nunes Vicente. Bilevel derivative-free optimization and its application to robust optimization. *Optimization Methods and Software*, 27(3):561–577, 2012.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26698–26710. Curran Associates, Inc., 2022.

L. Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015. doi: 10.1109/TIT.2015.2409256.

Matthias J. Ehrhardt and Lindon Roberts. Inexact derivative-free optimization for bilevel learning. *Journal of Mathematical Imaging and Vision*, 63(5):580–600, Jun 2021. ISSN 1573-7683. doi: 10.1007/s10851-021-01020-8.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577. PMLR, 10–15 Jul 2018a.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577. PMLR, 10–15 Jul 2018b.

Ya Gao, Guangquan Zhang, Jie Lu, and Hui-Ming Wee. Particle swarm optimization for bi-level pricing problems in supply chains. *Journal of Global Optimization*, 51(2):245–254, Oct 2011. ISSN 1573-2916. doi: 10.1007/s10898-010-9595-8.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

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

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3748–3758. PMLR, 13–18 Jul 2020.

Bin Gu, Guodong Liu, Yanfu Zhang, Xiang Geng, and Heng Huang. Optimizing large-scale hyperparameters via automated learning algorithm. *arXiv preprint*, abs/2102.09026, 2021. doi: 10.48550/arXiv.2102.09026.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Yudi Huang, Yujie Mo, Yujing Liu, Ci Nie, Guoqiu Wen, and Xiaofeng Zhu. Multiplex graph representation learning via bi-level optimization. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 2081–2089. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/230. Main Track.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4882–4892. PMLR, 18–24 Jul 2021.

Zhisen Jiang and Saverio Bolognani. Bi-zol: Bilevel zeroth-order learning with nonsmooth responses, 2026.

Hecheng Li and Yuping Wang. A genetic algorithm for solving a special class of nonlinear bilevel programming problems. In Yong Shi, Geert Dick van Albada, Jack Dongarra, and Peter M. A. Sloot (eds.), *Computational Science – ICCS 2007*, pp. 1159–1162, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-72590-9.

S. Liu, B. Kailkhura, P.Y. Chen, P. Ting, S. Chang, and L. Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search of static linear policies is competitive for reinforcement learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 1805–1814, Red Hook, NY, USA, 2018. Curran Associates Inc.

Parvin Nazari, Bojian Hou, Davoud Ataee Tarzanagh, Li Shen, and George Michailidis. Stochastic regret guarantees for online zeroth- and first-order bilevel optimization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 147406–147493. Curran Associates, Inc., 2025.

Yurii Nesterov. *Nonlinear Optimization*, pp. 3–58. Springer International Publishing, Cham, 2018. ISBN 978-3-319-91578-4. doi: 10.1007/978-3-319-91578-4_1.

Yurii Nesterov and Vladimir Spokoiny. Random Gradient-Free Minimization of Convex Functions. *Foundations of Computational Mathematics*, 17(2):527–566, April 2017. ISSN 1615-3383. doi: 10.1007/s10208-015-9296-2. https://doi.org/10.1007/s10208-015-9296-2.

N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597, 2016. doi: 10.1109/SP.2016.41.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., Red Hook, NY, USA, 2019.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 737–746, New York, New York, USA, 20–22 Jun 2016. PMLR.

Yuxing Peng, Yuanyuan Liu, Fanhua Shang, Hongying Liu, and Zhouchen Lin. Provable faster zeroth-order method for bilevel optimization with optimal dependency on error and dimension, 2025.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

B. T. Polyak. Introduction to optimization. *Optimization Software Inc., Publications Division, New York*, 1:32, 1987.

Marco Rando, Cesare Molinari, Lorenzo Rosasco, and Silvia Villa. An optimal structured zeroth-order algorithm for non-smooth optimization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36738–36767. Curran Associates, Inc., 2023.

Marco Rando, Cesare Molinari, Silvia Villa, and Lorenzo Rosasco. Stochastic zeroth order descent with structured directions. *Computational Optimization and Applications*, 89(3):691–727, Dec 2024. ISSN 1573-2894. doi: 10.1007/s10589-024-00616-1.

Marco Rando, Cesare Molinari, Lorenzo Rosasco, and Silvia Villa. A structured tour of optimization with finite differences, 2025a.

Marco Rando, Cheik Traoré, Cesare Molinari, Lorenzo Rosasco, and Silvia Villa. A structured proximal stochastic variance reduced zeroth-order algorithm. *arXiv preprint arXiv:2506.23758*, 2025b.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning, 2017.

George AF Seber and Alan J Lee. *Linear regression analysis*. John Wiley & Sons, 2003.

Reza Shirkavand, Peiran Yu, Qi He, and Heng Huang. Bilevel zofo: Efficient llm fine-tuning and meta-training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. A review on bilevel optimization: From classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22 (2):276–295, 2018. doi: 10.1109/TEVC.2017.2712906.

Mathias Staudigl, Simon Weissmann, and Tristan van Leeuwen. Derivative-free stochastic bilevel optimization for inverse problems. *Computational Optimization and Applications*, 93(3):967–1020, Apr 2026. ISSN 1573-2894. doi: 10.1007/s10589-025-00745-1.

Lu Wang, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Yuan Jiang. Spanning attack: reinforce black-box attacks with unlabeled data. *Machine Learning*, 109(12):2349–2368, Dec 2020. ISSN 1573-0565. doi: 10.1007/s10994-020-05916-1.

Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Achieving $\mathcal{O}(\varepsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 39491–39503. Curran Associates, Inc., 2023.

Vincent F. Yu, Shih-Wei Lin, Wenyih Lee, and Ching-Jung Ting. A simulated annealing heuristic for the capacitated location routing problem. *Computers & Industrial Engineering*, 58(2):288–299, 2010. ISSN 0360-8352. doi: https://doi.org/10.1016/j.cie.2009.10.007. Scheduling in Healthcare and Industrial Systems.

A Intuition Behind the Design of ZOBA and HF-ZOBA

In this appendix, we provide an intuitive explanation of how the iterations of ZOBA (Algorithm 1) and HF-ZOBA (Algorithm 2) are derived, highlighting and motivating the choice of surrogate constructions adopted. As described in Section 2, the main idea is to mimic gradient descent on the value function Ψ. In particular, at each iteration $k \in \mathbb{N}$, we aim to construct an update of the form

$$x_{k+1} = x_k - \gamma_k \hat{\nabla}\Psi(x_k) \quad \text{with} \quad \hat{\nabla}\Psi(x_k) \approx \nabla\Psi(x_k).$$

Under standard regularity assumption (Ghadimi & Wang, 2018), we have that Ψ is differentiable and its gradient can be expressed in closed form as follows.

$$\nabla\Psi(x) = \nabla_x F(z^*(x), x) + \nabla^2_{xz} G(z^*(x), x)\, v^*(x),$$

where

$$v^*(x) = -\left[\nabla^2_{zz} G(z^*(x), x)\right]^{-1} \nabla_z F(z^*(x), x).$$

The idea thus consists in constructing approximations of all unavailable quantities (i.e., first- and second-order information, as well as v^* and z^*) while minimizing the number of function evaluations and enabling parallelization whenever possible. The first step is to approximate $z^*(x)$. Since $z^*(x)$ is the minimizer of the inner objective G, we approximate it via a single step of gradient descent. However, in our setting the gradient is not available and only stochastic function evaluations can be queried. Therefore, we mimic one gradient step by constructing a finite-difference approximation of the gradient using mini-batches. Specifically, at each iteration $k \in \mathbb{N}$, denote the batch size $b_1 > 0$ and $\ell_1 > 0$ the number of directions. We sample $\xi_{1,k}, \cdots, \xi_{b_1,k}$ and for each $i \in [b_1]$, we draw ℓ_1 i.i.d. Gaussian random directions $(w_k^{(i,j)}) \sim \mathcal{N}(0, I_p)$. We then construct the following central finite-difference surrogate. Let $h_k > 0$

$$D_z^k := \frac{1}{b_1 \ell_1} \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \frac{g(z_k + h_k w_k^{(i,j)}, x_k, \xi_i) - g(z_k - h_k w_k^{(i,j)}, x_k, \xi_i)}{2h_k} w_k^{(i,j)}.$$

Notice that this can be interpreted as the standard Gaussian central finite-difference approximation (Nesterov & Spokoiny, 2017; Duchi et al., 2015; Ghadimi & Lan, 2013) of a mini-batch gradient of the function $z \mapsto G(z, x_k)$, with the only difference that each stochastic gradient in the batch is approximated using ℓ_1 directions. The idea of controlling the accuracy of the surrogate via the batch size (to reduce the stochasticity) and via the number of directions (to reduce the error induced by the approximation) is motivated by empirical observations in the finite-difference literature for stochastic settings, which suggest that, under limited evaluation budgets, balancing these two sources of approximation error can provide better performance than approaches that control accuracy solely through the batch size (Rando et al., 2025b). Such an approximation differs from those used in state-of-the-art methods (Aghasi & Ghadimi, 2025; Aghasi et al., 2025; Nazari et al., 2025; Peng et al., 2025). In particular, in Aghasi & Ghadimi (2025), the gradient surrogate is constructed with $b_1 = \ell_1 = 1$, while in Aghasi et al. (2025) and Nazari et al. (2025), one has $b_1 > 0$ and $\ell_1 = 1$, with spherical directions used in the latter case. In Peng et al. (2025), $\ell_1 = p$ since all canonical basis directions are used, together with $b_1 > 0$. In this context, our approach is more flexible, as it allows the accuracy of the estimator to be jointly controlled through both the batch size and the number of directions. The approximation of $z^*(x_k)$ is then defined with the following step of mini-batch finite-difference method: let $\rho_k > 0$ be a stepsize,

$$z_{k+1} = z_k - \rho_k D_z^k$$

Notice that this is exactly the update of z_k of Algorithm 1. Instead of using z_{k+1} as approximation of $z^*(x_k)$, we introduce a one-step delay by using z_k as approximation. This because allow us to reuse the function evaluation performed to compute the gradient approximation D_z^k in the approximation of $v^*(x_k)$. Indeed, to approximate $v^*(x_k)$, we observe that it is the action of the inverse of the hessian-block $\nabla_{zz} G(z^*(x_k), x_k)$ on the gradient in the first-variable of the outer target. The idea to approximate this quantity consists first in approximating such an action with one step of gradient descent on the associated quadratic form i.e. intuitively we want to approximate $v^*(x_k)$ as v_{k+1} defined as

$$v_{k+1} = v_k - \rho_k \left(\nabla^2_{zz} G(z^*(x_k), x_k) v_k + \nabla_z F(z^*(x_k), x_k)\right).$$

First, we replace $z^*(x_k)$ with z_k (i.e. the one-step delayed approximation), obtaining

$$v_{k+1} = v_k - \rho_k \left(\nabla^2_{zz} G(z_k, x_k) v_k + \nabla_z F(z_k, x_k)\right).$$

Notice that such an update is the deterministic version of the analogous update in SOBA (Dagréou et al., 2022). However, in our setting this is not sufficient since first- and second-order information are not available. Thus, in ZOBA, we replace such information with finite-difference surrogates and we observe that with an appropriate choice of surrogates we can still exploit advantages of usage of delay information. In particular, denoting with $g_{i,j}^{+,k} := g(z_k + h_k w_k^{(i,j)}, x_k, \xi_{i,k})$, $g_{i,j}^{-,k} := g(z_k - h_k w_k^{(i,j)}, x_k, \xi_{i,k})$ and $g_i^k := g(z_k, x_k, \xi_{i,k})$ we approximate $\nabla_{zz}^2 G(z_k, x_k)$ with

$$\hat{\nabla}_{zz}^2 g(z_k, x_k) = \frac{1}{b_1 \ell_1} \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \frac{g_{i,j}^{+,k} + g_{i,j}^{-,k} - 2g_i^k}{2h_k^2} \left(w_k^{(i,j)} w_k^{(i,j)\top} - I_p \right).$$

Notice that, thanks to the use of delayed information z_k instead of z_{k+1}, we can reuse the same directions $\left((w_k^{(i,j)})_{j=1}^{\ell_1}\right)_{i=1}^{b_1}$, the same samples $(\xi_{i,k})_{i=1}^{b_1}$, and the same function evaluations used to construct the surrogate D_z^k. As a result, the overall cost reduces to only $b_1 \ell_1$ function evaluations. This advantage cannot be exploited by two-loop methods (Aghasi et al., 2025; Aghasi & Ghadimi, 2025), where $z^*(x_k)$ is approximated through multiple finite-difference steps, preventing such reuse. Moreover, the reuse of the same information in the approximation of $\nabla_z G$ also motivates the use of central finite differences in D_z^k. While central finite differences are typically more expensive than forward schemes, they provide higher accuracy (Berahas et al., 2022). In our setting, however, this increased accuracy comes at essentially no additional cost, since the same function evaluations are already required for the approximation of the Hessian block. We then approximate the gradient of the outer target in the inner variable using batched forward finite-difference.

$$\hat{\nabla}_z f(z_k, x_k) := \frac{1}{b_1 \ell_1} \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \frac{f(z_k + h_k w_k^{(i,j)}, x_k, \zeta_{i,k}) - f(z_k, x_k, \zeta_{i,k})}{h_k} w_k^{(i,j)}.$$

Notice that we reuse the same Gaussian directions $w_k^{(i,j)}$ employed in the construction of the previous surrogates, thereby avoiding additional sampling costs. Using these approximations, we define the iterate v_k of ZOBA as follows.

$$v_{k+1} = v_k - \rho_k \underbrace{\left(\hat{\nabla}_{zz}^2 g(z_k, x_k) v_k + \hat{\nabla}_z f(z_k, x_k) \right)}_{=: D_v^k}.$$

Finally, we construct an approximation of the hypergradient (eq. 2) at iteration $k \in \mathbb{N}$. We first approximate $z^(x_k)$ and $v^(x_k)$ using their one-step delayed counterparts z_k and v_k, respectively, obtaining

$$\nabla \Psi(x_k) \approx \nabla_x F(z_k, x_k) + \nabla_{xz}^2 G(z_k, x_k)\, v_k$$

Then, we approximate the gradient $\nabla_x F(z_k, x_k)$ using a batched finite-difference estimator. Let $b_2 > 0$ denote the batch size and $\ell_2 > 0$ the number of directions. We sample $(\zeta_{i,k})_{i=1}^{b_2}$, and for each $i \in [b_2]$, we draw ℓ_2 i.i.d. Gaussian directions $(u_k^{(i,j)})_{j=1}^{\ell_2} \sim \mathcal{N}(0, I_d)$. We then define the following gradient approximation.

$$\hat{\nabla}_x f(z_k, x_k) := \frac{1}{b_2 \ell_2} \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \frac{f(z_k, x_k + h_k u_k^{(i,j)}, \zeta_{i,k}) - f(z_k, x_k, \zeta_{i,k})}{h_k} u_k^{(i,j)}.$$

Notice that, thanks to the use of delayed information, instead of computing the function evaluations $f(z_k, x_k, \zeta_{i,k})$ for $i = 1, \ldots, b_2$, we can recycle them from the computation of $\hat{\nabla}_z f(z_k, x_k)$ (in Algorithm 1, since b_2 and ℓ_2 may be larger than b_1 and ℓ_1, we sample directions and perform function evaluations using the maximum of the two). This observation motivates the choice of forward finite differences for this surrogate. Indeed, forward differences allow us to reuse part of the function evaluations performed for $\hat{\nabla}_z f(z_k, x_k)$, unlike central schemes. For instance, in ZOSOGD (Nazari et al., 2025), central finite differences with spherical directions are used, which do not permit such reuse. A similar reuse mechanism applies to the approximation of the Hessian mixed block $\nabla_{xz}^2 G$. In particular, we approximate it using the following batched finite-difference estimator. Let $\hat{g}_{i,j}^{+,k} := g(z_k + h_k w_k^{(i,j)}, x_k + h_k u_k^{(i,j)}, \xi_{i,k})$ and $\hat{g}_{i,j}^{-,k} := g(z_k - h_k w_k^{(i,j)}, x_k - h_k u_k^{(i,j)}, \xi i, k)$, we define the following mixed Hessian block surrogate.

$$\hat{\nabla}_{xz} g(z_k, x_k) := \frac{1}{b_1 \ell_1} \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \frac{\hat{g}_{i,j}^{+,k} + \hat{g}_{i,j}^{-,k} - 2g_i^k}{2h_k^2} u_k^{(i,j)} w_k^{(i,j)\top}.$$

Notice that, again thanks to the use of delayed information, we do not need to recompute $(g_i^k)_{i=1}^{b_1}$, since these values have already been computed for $\hat{\nabla}_{zz} g$, thereby saving $b_1 \ell_1$ function evaluations. Finally, we combine these two approximations to obtain the hypergradient estimator described in Algorithm 1, i.e.,

$$x_{k+1} = x_k - \gamma_k \left(\hat{\nabla}_{xz} g(z_k, x_k) v_k + \hat{\nabla}_x f(z_k, x_k) \right).$$

HF-ZOBA (Algorithm 2) follows the same idea, with the only difference that, instead of explicitly constructing the Hessian blocks, we approximate their action on the vector v_k. In particular, rather than using a central finite-difference approximation of $\nabla_z G(z_k, x_k)$ in the update of z_k, we employ a forward finite-difference estimator, i.e., we construct

$$\hat{\nabla}_z g(z_k, x_k) = \frac{1}{b_1 \ell_1} \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \frac{g(z_k + h_k w_k^{(i,j)}, x_k, \xi_i) - g(z_k, x_k, \xi_i)}{h_k} w_k^{(i,j)}.$$

We thus define the following iterate

$$z_{k+1} = z_k - \rho_k \hat{\nabla}_z g(z_k, x_k).$$

The choice of using forward finite-difference here is motivated by the fact that here we no longer benefit from the advantages of central finite-difference schemes, since an additional gradient approximation will be required to estimate the Hessian-block action. In particular, let $\bar{h}_k > 0$ denote a sequence of discretization parameters. We approximate $\nabla^2_{zz} G(z_k, x_k)\, v_k$ (in the v_k iteration) as follows.

$$H_{zz}^k = \frac{1}{\bar{h}_k} \left(\hat{\nabla}_z g(z_k + \bar{h}_k v_k, x_k) - \hat{\nabla}_z g(z_k, x_k) \right).$$

This corresponds to a first-order finite-difference approximation of $\nabla^2_{zz} G(z_k, x_k)\, v_k$, where exact gradients are replaced by finite-difference surrogates. Notice that, since the computation of H_{zz}^k involves $\hat{\nabla}_z g(z_k, x_k)$, we obtain "for free" the gradient approximation needed in the update of z_k. Notice that, if central finite differences were used, the number of function evaluations required to compute $\hat{\nabla}_z g(z_k + \bar{h}_k v_k, x_k)$ would be higher. Finally, the v_k iterate is obtained by replacing $\hat{\nabla}^2_{zz} g(z_k, x_k), v_k$ in the v_{k+1} update of ZOBA with H_{zz}^k, i.e.,

$$v_{k+1} = v_k - \rho_k \left(H_{zz}^k + \hat{\nabla}_z f(z_k, x_k) \right).$$

Analogously, we replace the term $\hat{\nabla}_{xz} g(z_k, x_k)\, v_k$ in the ZOBA hypergradient approximation with a surrogate of the action of the mixed Hessian block on the vector v_k. To this end, we define a finite-difference approximation of the gradient of the inner target with respect to the outer variable, i.e.,

$$\hat{\nabla}_x g(z_k, x_k) = \frac{1}{b_1 \ell_1} \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \frac{g(z_k, x_k + h_k u_k^{(i,j)}, \xi_i) - g(z_k, x_k, \xi_i)}{h_k} u_k^{(i,j)}.$$

Notice that we reuse the already sampled directions $u_k^{(i,j)}$ (recall that using $b_1 \ell_1$ directions instead of $b_2 \ell_2$ is not restrictive, since in Algorithm 2 we sample and evaluate using the maximum of the two quantities). We thus define the following approximation of $\hat{\nabla}^2_{xz} g(z_k, x_k)\, v_k$.

$$H_{xz}^k = \frac{1}{\bar{h}_k} \left(\hat{\nabla}_x g(z_k + \bar{h}_k v_k, x_k) - \hat{\nabla}_x g(z_k, x_k) \right).$$

Thus, by substituting this surrogate into the hypergradient approximation in the x_k update, we obtain the HF-ZOBA iteration.

$$x_{k+1} = x_k - \gamma_k \left(H_{xz}^k + \hat{\nabla}_x f(z_k, x_k) \right).$$

B Extended Comparison with Finite-Difference Bilevel Optimization Methods

In this appendix, we compare our methods with state-of-the-art finite-difference algorithms in terms of the function evaluations performed at each iteration. More precisely, we analyze how many function evaluations are required per iteration by state-of-the-art methods and how these evaluations are used.

ZDSBA (Aghasi & Ghadimi, 2025). At each iteration $k \in \mathbb{N}$, ZDSBA first approximates the inner solution $z^*(x_k)$ at the current iterate x_k by constructing the sequence $(z_k)_{i=1}^{N_{\text{inner}}}$ through $N_{\text{inner}} > 0$ iterations of a stochastic zeroth-order method. This procedure involves the construction of a stochastic gradient surrogate of the inner objective at every step. Such a surrogate is built using finite differences with a single sample and a single Gaussian direction. Each inner iteration therefore requires two function evaluations, resulting in a total cost of $2N_{\text{inner}}$ function evaluations. Next, the method approximates $v^*(x_k)$ by constructing a sequence $(v_k)_{k=1}^{N_{\text{inverse}}}$ via N_{inverse} iterations of a stochastic zeroth-order method applied to the quadratic form in eq. 3, where $z^*(x_k)$ is replaced by $z_{N_{\text{inner}}}$. Each iteration again relies on single-sample, single-direction finite-difference estimators. Since this quadratic form requires approximations of both the gradient of the outer objective with respect to the first variable and the corresponding Hessian block, each iteration incurs a cost of 5 function evaluations. Finally, the Hessian cross-block and the gradient of the outer objective with respect to the second variable are approximated. Again, such approximations are built using single-sample, single Gaussian-direction estimators, incurring an additional cost of 5 function evaluations. Therefore, the total number of function evaluations per iteration k is

$$2N_{\text{inner}} + 5N_{\text{inverse}} + 5.$$

ZMDSBA (Aghasi et al., 2025). ZMDSBA generalizes ZDSBA by allowing multiple samples in the estimation of the Hessian cross-block and the gradient of the outer objective with respect to the second variable. Specifically, these quantities are approximated using finite differences with a single Gaussian direction and N_{sample} samples (i.e. our gradient/hessian estimators with $\ell_1 = \ell_2 = 1$ and $b_1 = b_2 = N_{\text{sample}}$). The other steps are identical to ZDSBA. As a result, the total number of function evaluations per iteration becomes

$$2N_{\text{inner}} + 5N_{\text{inverse}} + 5N_{\text{sample}}.$$

OPT-ZMDSBA (Aghasi et al., 2025). OPT-ZMDSBA is based on a regularization-based reformulation of the bilevel problem, where the inner optimization is incorporated via a penalty function. This approach allows the algorithm to avoid explicit Hessian approximations. At each iteration k, the algorithm constructs approximations of $z^*(x_k)$ with a sequence $(z_k)_{k=1}^{N_{\text{inner}}}$ and an auxiliary sequence $(y_k)_{k=1}^{N_{\text{inner}}}$ by performing N_{inner} iterations of a stochastic zeroth-order method using single-sample, single Gaussian direction gradient surrogates. More precisely, these sequences require computing three gradient surrogates of the inner and outer objectives per iteration, resulting in a total cost of $6N_{\text{inner}}$ function evaluations. Then, the algorithm constructs a search direction to update the iterate. Such a directions requires computing three surrogate gradients of the inner and outer objectives using single-direction estimators with N_{sample} samples, incurring a cost of $6N_{\text{sample}}$ function evaluations. Therefore, the total number of function evaluations per iteration is

$$6N_{\text{inner}} + 6N_{\text{sample}}.$$

VRZSBO (Peng et al., 2025). VRZSBO is a single-loop finite-difference algorithm for bilevel optimization. While it has a structure similar to Algorithm 2, it exploits delayed information to avoid unparallelizable inner loops and approximates Hessian-vector products using first-order finite differences, it differs in a key aspect. In VRZSBO, each surrogate is constructed using all canonical basis vectors as directions. Although this leads to accurate approximations, it implies that for every surrogate the number of function evaluations request to construct such approximation is at least equal to the dimensions of both the inner and outer variables, resulting in very expensive iterations. The algorithm approximates $z^*(x_k)$ by constructing z_k, which is updated via a single batch finite-difference step along coordinate directions. Let $B > 0$ denote the batch size, computing z_{k+1} requires $B(p+1)$ function evaluations. Next, $v^*(x_k)$ is approximated by a single gradient descent step on the quadratic form in eq. 3, where $z^*(x_k)$ is replaced by z_k. The gradient $\nabla_z F$ is approximated using batched finite differences along coordinate directions, while $\nabla^2_{zz} G(z_k, x_k)v_k$ is approximated via first-order finite differences, with the resulting gradients again computed using batched finite differences with coordinate directions. The number of function evaluations required for this step is thus $3B(p+1)$. Finally, the hypergradient is approximated as in Algorithm 2, but using coordinate directions instead of Gaussian directions. This requires constructing a surrogate for $\nabla_x F(z_k, x_k)$, which costs $B(d+1)$ function evaluations, and for $\nabla^2_{xz} G(z_k, x_k)v_k$, which is again approximated via finite differences along coordinate directions, resulting in a cost of $2B(d+1)$. Thus, the total number of function evaluations per iteration is

$$B(3(d+1) + 4(p+1)).$$

ZO-SOGD (Nazari et al., 2025). ZO-SOGD is a single-loop zeroth-order algorithm for bilevel optimization in which estimators are constructed using spherical directions. This method follows a similar algorithmic structure to HF-ZOBA, with two key differences: (i) a momentum term is incorporated into the iterates, and (ii) gradient surrogates are constructed using batched two-point central finite differences along spherical directions. While the use of momentum improves the convergence speed of the algorithm, the stochastic nature of the estimators requires computing the search directions twice per iteration. Let B_1 denote the batch size used for sampling realizations of the outer objective and B_2 the batch size for the inner objective. Then, the total number of function evaluations per iteration is given by

$$18B_1 + 10B_2$$

In Table 1, we summarize these methods by reporting the number of function evaluations per iteration (# FE) and their theoretical complexity, i.e. the number of function evaluations required to achieve $\varepsilon \in (0, 1)$ accuracy.

Table 1: Comparison of bilevel zeroth-order algorithms in terms of complexity and number of function evaluations performed at every iteration (# FE).

Method	Loops	# FE	Complexity
ZDSBA	two-loops	$2N_{\text{inner}} + 5N_{\text{inverse}} + 5$	$\mathcal{O}((d+p)^4\varepsilon^{-3}\log((d+p)/\varepsilon))$
ZMDSBA	two-loops	$2N_{\text{inner}} + 5N_{\text{inverse}} + 5N_{\text{sample}}$	$\mathcal{O}(p(d+p)^2\varepsilon^{-2}\log(1/\varepsilon))$
OPT-ZMDSBA	two-loops	$6N_{\text{inner}} + 6N_{\text{sample}}$	$\mathcal{O}((d+p)\varepsilon^{-2})$
VRZSBO	single-loop	$B(3(d+1) + 4(p+1))$	$\mathcal{O}((d+p)\varepsilon^{-2})$
ZOSOGD	single-loop	$18B_1 + 10B_2$	$\mathcal{O}((d+p)\varepsilon^{-2})$
ZOBA	single-loop	$b_1(4\ell_1 + 1) + b_2(2\ell_2 + 1)$	$\mathcal{O}(p(d+p)^2\varepsilon^{-2})$
HF-ZOBA	single-loop	$2b_1(2\ell_1 + 1) + b_2(2\ell_2 + 1)$	$\mathcal{O}((d+p)\varepsilon^{-2})$

In Table 1, we observe that in the limit case $b_1 = b_2 = \ell_1 = \ell_2 = N_{\text{inner}} = N_{\text{sample}} = N_{\text{inverse}} = 1$, our methods require fewer function evaluations per iteration than other finite-difference bilevel algorithms. This reduction stems from the reuse of function evaluations enabled by the use of delayed information in ZOBA and HF-ZOBA. In addition, the function evaluations required by ZOBA and HF-ZOBA can be performed in parallel across directions and batch samples, while the evaluations in ZDSBA, ZMDSBA and OPT-ZMDSBA are inherently sequential due to their inner-loop structure. As a result, the effective wall-clock cost of ZOBA and HF-ZOBA can be significantly lower in practice, even when the number of function evaluations is comparable or even higher.

C Experimental Details

In this appendix we provide all details on the experimental setup and parameter tuning used to perform the experiments reported in Section 4. We implemented all scripts in Python 3 (version 3.11) and used the NumPy (version 1.25.0) (Harris et al., 2020), PyTorch (version 2.0.1) (Paszke et al., 2019), and Matplotlib (version 3.10.8) (Hunter, 2007) libraries. This appendix is organized as follows. In Appendix C.1, we provide details of the experiments on the synthetic quadratic problem while in Appendix C.2, we provide details of the experiments on the minimal-distortion universal perturbation task. Details on the machine used to perform the experiments are reported in Table 2.

Table 2: Machine used to perform the experiments

Feature	
CPU	64 x AMD EPYC 7301 16-Core
GPU	1 x NVIDIA Quadro RTX 6000
RAM	256 GB

C.1 SYNTHETIC PROBLEM

Here we describe the synthetic problem considered in Section 4. We consider the minimization of the following function

$$\begin{aligned} &\min_{x\in\mathbb{R}^d} \Psi(x) := F(z^*(x), x) := \frac{1}{2n}\sum_{i=1}^{n} \left(C_i z^*(x) - D_i x - b_i\right)^2 + \frac{1}{2}\|x - \bar{x}\|_2^2 \\ &\text{s.t.} \quad z^*(x) \in \arg\min_{z\in\mathbb{R}^p} G(z, x) := \frac{1}{2m}\sum_{j=1}^{m} \left(A_j z - B_j x - a_j\right)^2, \end{aligned} \tag{11}$$

where $A \in \mathbb{R}^{m\times p}, B \in \mathbb{R}^{m\times d}, C \in \mathbb{R}^{n\times p}, D \in \mathbb{R}^{n\times d}$ are Gaussian matrices where every entry is sampled from $\mathcal{N}(0,1)$. For fixed $\bar{z} \in \mathbb{R}^p$ and $\bar{x} \in \mathbb{R}^d$, we define for every $j = 1, \cdots, m$ and $i = 1, \cdots, n$

$$a := A\bar{z} - B\bar{x} \quad \text{and} \quad b := C\bar{z} - D\bar{x}.$$

In the experiments in Section 4, we fixed $n = m = 1000$, $\bar{z} = [2, \cdots, 2] \in \mathbb{R}^p$ and $\bar{x} = [1, \ldots, 1] \in \mathbb{R}^d$. We considered the setting $d = p = 25, 50, 100$.

Parameter Tuning. All parameters of every algorithm were tuned via grid search. For every algorithm and each parameter configuration in its parameter grid, the algorithm was run 6 times with a fixed budget of 10^5 function evaluations, using 6 different random initializations. The same set of initializations was used across all parameter configurations to ensure a fair comparison. The optimal parameter configuration was selected as the one minimizing, on average over the runs, the normalized function value gap at the final iterate,

$$\frac{\Psi(x_K) - \min \Psi}{\Psi(x_0) - \min \Psi},$$

where x_K denotes the iterate obtained at the last iteration performed K. Initializations of x_0 and z_0 were sampled independently from the uniform distribution over the hypercubes $[-5.0, 10.0]^d$ and $[-5.0, 10.0]^p$, respectively. The initialization of v_0 (i.e. the approximation of the inverse Hessian-vector product) was set to the zero vector. Once the optimal parameter configuration was identified, the algorithm was run 11 additional times using the same evaluation budget. To mitigate the effect of initialization overhead (e.g., library loading) on runtime measurements, the first run was discarded, and all reported results were computed by averaging over the remaining 10 runs.

Search Spaces. For all algorithms, every discretization parameter used are fixed to $h = 10^{-3}$. Every stepsize parameter (used by all algorithms) was tuned over the set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$. For ZOBA and HF-ZOBA, the number of directions ℓ_1 and ℓ_2 used to approximate stochastic gradients and Hessians was selected from $\{1, 10, 50, 100\}$. Every batch sizes (b_1, b_2 in ZOBA, HF-ZOBA, B_1, B_2 for ZOSOGD, B in VRZSBO and s_k in ZMDSBA and OPT-ZMDSBA of (Aghasi et al., 2025)) were tuned over $\{1, 10, 50, 100\}$. Notice that the number of directions and batch sizes are not required to be equal, i.e., ℓ_1, ℓ_2, b_1, b_2 can all take different values. The number of inner-loop iterations used to approximate z^* and v^* for ZDSBA, ZMDSBA, and OPT-ZMDSBA was tuned over $\{1, 10, 50\}$. Momentum parameter of ZOSOGD has been tuned over $\{0.1, 0.2, 0.5, 0.7, 0.9, 1.0\}$. Finally, the regularization parameter λ for OPT-ZMDSBA was selected from $\{1.0, 10.0, 100.0\}$. All parameters tuned, their ranges and the algorithms that used them are summarized in Table 3. For each method, the full search grid was defined as the Cartesian product of the candidate sets corresponding to its parameters.

Table 3: Parameter grids used for each algorithm.

Parameter	**Algorithm(s)**	**Candidate Values**
Stepsizes	All algorithms	$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$
Number of directions	ZOBA, HF-ZOBA	$\{1, 10, 25, 50, 100\}$
Batch sizes	ZOBA, HF-ZOBA, VRZSBO ZMDSBA, OPT-ZMDSBA,ZOSOGD	$\{1, 10, 50, 100\}$
Inner-loop iterations	ZDSBA, ZMDSBA, OPT-ZMDSBA	$\{1, 10, 50\}$
Regularization	OPT-ZMDSBA	$\{1.0, 10.0, 100.0\}$
Momentum	ZOSOGD	$\{0.2, 0.5, 0.7, 1.0\}$

Impact of the number of directions and batchsize. We study how the performance of our algorithms is affected by the choice of the number of directions and the batch-size parameters. Specifically, we consider the optimization of the synthetic objective defined in eq. 11 with $p = d = 50$ and $n = m = 1000$. We fix a budget of 10^4 function evaluations (i.e., fewer than those used in Section 4) and run ZOBA and HF-ZOBA with a constant stepsize ρ, chosen from a grid of 20 logarithmically spaced values between 10^{-5} and 1.0. The outer stepsize is set as a fraction of ρ, namely $\gamma = \rho/c$ and, in particular, we fix $c = 5.0$. The discretization parameters are fixed to $h = \hat{h} = 10^{-3}$. To reduce the number of hyperparameters, we impose $b := b_1 = b_2$ and $\ell := \ell_1 = \ell_2$. Each experiment is repeated five times, and we report the mean and standard deviation of the results. In Figure 3, we plot the normalized objective value progress $(\Psi(x_K) - \min \Psi)/(\Psi(x_0) - \min \Psi)$ at the final iterate x_K obtained by ZOBA as a function of the stepsize ρ. If the algorithm diverges, this quantity is clipped to one.

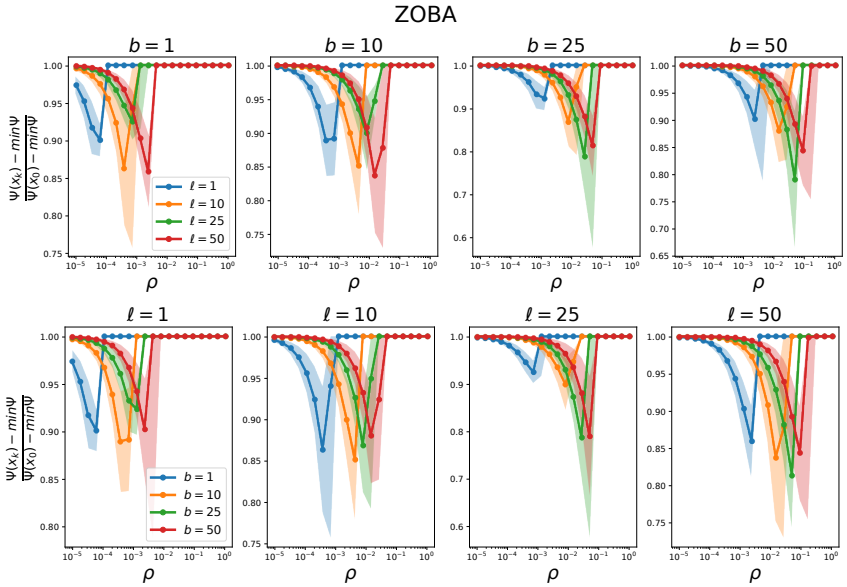


Figure 3: Normalized objective value progress at the final iterate obtained by running Algorithm 1 with different stepsizes ρ, numbers of directions ℓ, and batch sizes b. If the algorithm diverges, the quantity $(\Psi(x_K) - \min \Psi)/(\Psi(x_0) - \min \Psi)$ is clipped to one. In the first row, every plot show how this quantity change for different values of ℓ by fixing the batch size b; in the second row, instead, every plot show how such value change for different values of b by fixing ℓ.

From Figure 3, we observe that increasing the number of directions ℓ, i.e., improving the quality of the stochastic gradient approximations for both the inner and outer objectives, makes ZOBA more stable and allows for the use of larger stepsize ρ. Moreover, increasing ℓ generally leads to better or comparable final performance in terms of solution quality, while also being potentially more time-efficient in practice since function evaluations can be computed in parallel. When ℓ (and eventually b) becomes too large, performance deteriorates. This behavior is due to the limited budget of function evaluations: large values of ℓ and b consume many evaluations per iteration, resulting in too few optimization steps to make meaningful progress. Comparing the first and second rows of Figure 3, we further observe that although ℓ and b play a similar conceptual role (both improving the quality of gradient and Hessian surrogates), fixing b and increasing ℓ may yield better results in practice. For example, fixing $b = 1$ and increasing ℓ achieves a lower normalized objective gap $(\Psi(x_K) - \min \Psi)/(\Psi(x_0) - \min \Psi)$ than fixing $\ell = 1$ and increasing b. Moreover, this advantage diminishes or vanishes when the fixed parameter takes a large value. This behavior can again be attributed to the limited evaluation budget, or to the fact that when either ℓ or b is sufficiently large, the variance becomes small and the two parameters play increasingly similar roles. One possible explanation of that phenomenon is that, in some regimes, stochastic (or mini-batch) gradients are already sufficiently accurate and it is also easier to obtain good zeroth-order approximations via multiple directions than improving surrogate quality through batching with poor approximations i.e.

good approximations of high-variance stochastic gradients (large ℓ, small b) may thus be preferable to poor approximations of low-variance minibatch gradients (small ℓ, large b). A similar phenomenon has been observed in Rando et al. (2025b), suggesting that this behavior is not specific in bilevel optimization but it can also arise in other stochastic settings. These findings provide empirical evidence suggesting that achieving a good balance between ℓ and b is important for maximizing the performance of the method, and they support the use of flexible surrogates such as ours, rather than special cases in which ℓ is fixed to 1 and only b varies, as in Aghasi & Ghadimi (2025); Aghasi et al. (2025). In Figure 4, HF-ZOBA exhibits a similar pattern: increasing ℓ or b improves stability

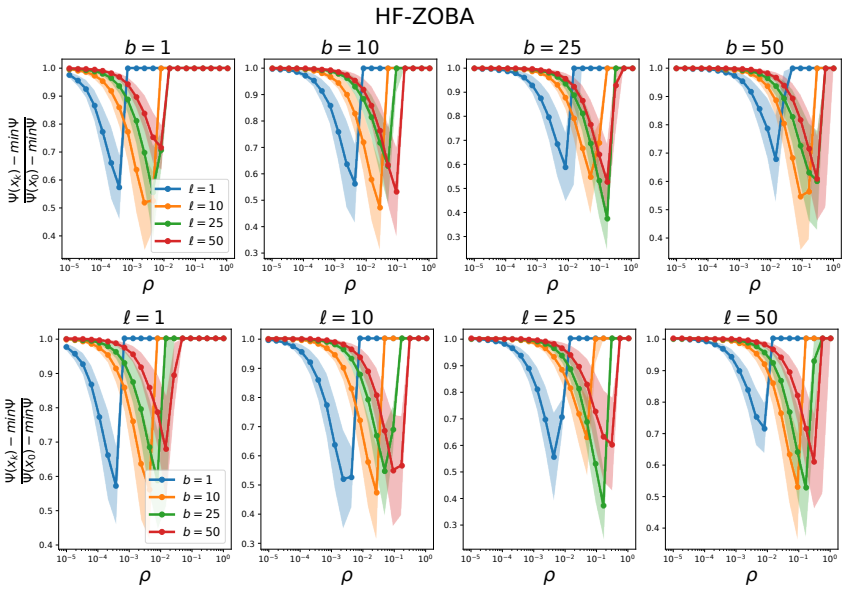


Figure 4: Normalized objective value progress at the final iterate obtained by running Algorithm 2 with $\gamma = \rho/5$ for different stepsizes ρ, numbers of directions ℓ, and batch sizes b. If the algorithm diverges, the quantity $(\Psi(x_K) - \min \Psi)/(\Psi(x_0) - \min \Psi)$ is clipped to one. In the first row, every plot show how this quantity change for different values of ℓ by fixing the batch size b; in the second row, instead, every plot show how such value change for different values of b by fixing ℓ.

and allows for larger stepsizes ρ, and fixing a small b while increasing ℓ achieves lower normalized objective gaps than the opposite. Comparing Figures 4 and 3, HF-ZOBA achieves significantly better results than ZOBA under the limited evaluation budget, even though both methods perform similarly in Section 4. This can be attributed to the smaller budget: while higher budgets allow using more directions and larger batch sizes, under limited evaluations fewer directions or smaller batch sizes must be used. ZOBA relies on explicit Hessian approximations, which can be noisy with few directions, leading to higher variance in the updates. As a result, HF-ZOBA, which relies solely on gradient surrogates, does not suffer from this variance, allowing for more precise steps and thus better performance under limited evaluation budget settings.

Parallelism. We study how the single-loop structure of the algorithm and function value reuse, enables efficient parallelism. We fix a budget of 10^5 function evaluations and consider the synthetic bilevel problem defined in eq. 11 across different dimensions. We compare the performance of our method against two-loop procedures under two regimes: (i) a fully parallel setting, where function evaluations can be performed in parallel, and (ii) a sequential setting, where evaluations are performed one at a time. Experiments are repeated 11 times, and the first run is discarded to exclude library loading overhead. Results are reported as mean $\pm$ standard deviation over the remaining 10 runs.

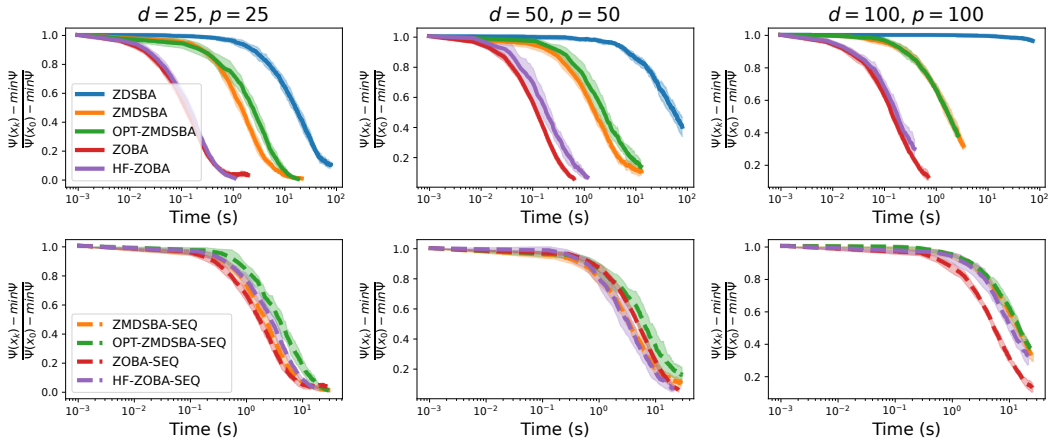


Figure 5: Comparison of algorithms on synthetic bilevel problems with varying dimensions: function value gap versus wall-clock time in seconds admitting parallel function evaluations (top row) and in sequential evaluation regime (bottom row)

In Figure 5, we report the (normalized) function value gap as a function of wall-clock time in full-parallel regime (top row) and in sequential-setting (bottom row). We observe that under sequential evaluations, our methods achieve similar runtime performance to competing approaches, except for ZDSBA, which requires more time. This result is expected, as the total number of function evaluations is the same across all methods. ZDSBA incurs additional overhead because it performs many inner iterations to compensate for not using multiple directions in gradient estimation. Opt-ZMDSBA and ZMDSBA partially mitigate this limitation by allowing larger batch sizes for estimating the Hessian cross-block and the gradient of the outer objective with respect to the second variable, thereby reducing the number of inner iterations. Despite employing additional approximations, our methods still achieve comparable time costs due to function value reuse; without this, a larger number of function evaluations would be required (notice that indeed in Section 4 we observed that other single loop methods provide worse performance in tackling the same problem), leading to higher per-iteration time under sequential computation. However, when function evaluations are parallelized, our methods exhibit a substantially greater time advantage compared to Opt-ZMDSBA and ZMDSBA, as they do not incur inner-loop overhead.

Function Value Reuse. Here we study the impact of the function-value-reuse feature of ZOBA and HF-ZOBA. We fix a budget of 10^5 function evaluations and consider the synthetic bilevel problem defined in eq. 11 with dimensions $p = d = 50$. We compare ZOBA and HF-ZOBA against ZDSBA, ZMDSBA, Opt-ZMDSBA, VRZSBO, and ZOSOGD. To study the impact of function value reuse, we also consider variants of ZOBA and HF-ZOBA that do not reuse function values for computing the various quantities, but instead sample new vectors and stochastic components for every gradient/Hessian computed; we denote these variants "NFR" (no function value reuse). Experiments are repeated 11 times, and the first run is discarded to exclude library-loading overhead. Results are reported as mean $\pm$ standard deviation over the remaining 10 runs.

In Figure 6, we report the (normalized) function value gap as a function of the number of function evaluations (left) and as a function of wall-clock time (right). Approaches without function value reuse (NFR) are represented as dotted lines. We observe that, without function value reuse, ZOBA and HF-ZOBA perform worse than their function-value-reuse counterparts and worse than the two-loop approaches. This confirms the advantage of function value reuse in containing the number of function evaluations per iteration of ZOBA and HF-ZOBA: without it, these approaches begin to scale poorly with batch size and number of directions, similarly to ZOSOGD. Notably, ZOBA (NFR) and HF-ZOBA (NFR) still provide better performance than the other single-loop approaches (ZOSOGD and VRZSBO). This shows that, even without function value reuse, the underlying algorithmic structure of ZOBA and HF-ZOBA still yields a better per-iteration cost scaling than prior single-loop

approaches. In the right plot, we observe that ZOBA (NFR) and HF-ZOBA (NFR) coincide in wall-clock time with ZOBA and HF-ZOBA, since they still retain the same advantage in parallelism.

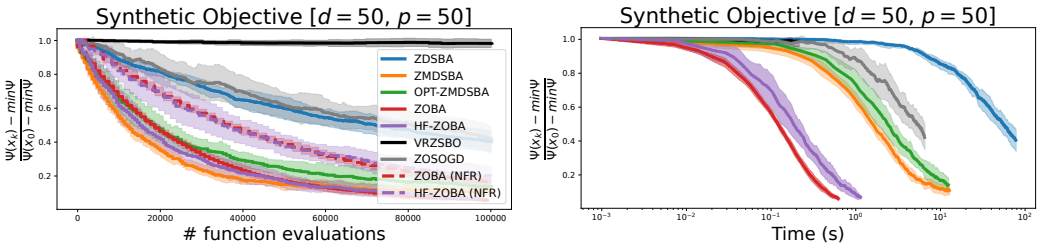


Figure 6: Comparison of algorithms on synthetic bilevel problems with dimensions $p = d = 50$: function value gap versus number of function evaluations (left) and versus wall-clock time in seconds (right). Approaches indicated with (NFR) indicate that function evaluations are not reused.

C.2 Minimal-distortion Universal Perturbation

In this appendix, we report the experimental details of the minimal-distortion universal adversarial attack experiment presented in Section 4. More precisely, we describe the problem and we propose our bilevel formulation. We describe how we implemented and trained the classifier and how we choose the parameters of the optimizer.

Problem Description & Bilevel Formulation. The goal of the minimal-distortion universal perturbation problem is the following: let $\hat{f} : \mathbb{R}^{d_{\text{img}}} \to \mathbb{R}^C$ be a trained multiclass classifier and let $S_{\text{test}} = \{(\hat{x}_i, y_i)\}_{i=1}^m$ with $\hat{x}_i \in \mathbb{R}^{d_{\text{img}}}$ and $y_i \in \{1, \dots, C\}$ be a test dataset. Let $\psi : \mathbb{R}^{d_{\text{img}}} \times \mathbb{R}^{d_{\text{img}}} \to \mathbb{R}^{d_{\text{img}}}$ be a manipulation function that, given a data point $\hat{x}$ and a perturbation δ, returns the perturbed data point with perturbation δ. Given a function $L(y, \hat{f}(x))$ that measures the confidence of classifying x with label y and a distortion measure $D(\hat{x}, \delta)$ that quantifies the amount of distortion obtained by perturbing $\hat{x}$ with δ, the minimal-distortion universal perturbation problem can be formulated as the following constrained optimization problem:

$$\min_{\delta \in \mathbb{R}^{d_{\text{img}}}} \frac{1}{m} \sum_{i=1}^{m} D(\hat{x}_i, \delta) \quad \text{s.t.} \quad \delta \in \arg\min_{\delta' \in \mathbb{R}^{d_{\text{img}}}} \frac{1}{m} \sum_{i=1}^{m} L(y_i, \hat{f}(\psi(\hat{x}_i, \delta))).$$

Based on prior observations that adversarial perturbations tend to concentrate in low-dimensional subspaces of the input space (see, e.g., (Chen et al., 2017; Bayer et al., 2025)) and the fact that zeroth-order methods can perform poorly in high-dimensional settings due to their dependence on dimension in the convergence rates (see, e.g., (Nesterov & Spokoiny, 2017)), we restrict the search for adversarial perturbations to a low-dimensional subspace. This idea has been proposed in previous works and is common, for instance, in spanning-based attacks (Wang et al., 2020), where the perturbation is typically parameterized as $\delta = \phi(A, z)$, with a fixed or randomly chosen subspace A, an optimization variable z and a function ϕ that combines the subspace A and the coefficient z constructing the perturbation. In contrast to these approaches, where the subspace is predetermined or heuristically choosen, we optimize it by formulating the problem of finding a minimal-distortion universal perturbation as a black-box bilevel optimization problem. In this formulation, the outer objective searches for a subspace A that minimizes the average distortion induced by the perturbation, while the inner objective minimizes the average classification confidence within the optimized subspace. Specifically, we parametrize the perturbation as $\delta = Az$, with $A \in \mathbb{R}^{d_{\text{img}} \times p}$ and $z \in \mathbb{R}^p$, where $p \ll d_{\text{img}}$. The inner problem yields $z^\star(A)$, leading to the final perturbation $\delta = Az^\star(A)$. This approach relies on the intuition that minimizing the average distortion is an "easier" problem than minimizing classification confidence, which is instead recast as the inner optimization problem in $\mathbb{R}^p$ with a smaller p. To the best of our knowledge, this is the first work to explicitly optimize the subspace within a bilevel formulation for subspace-based minimal-distortion universal perturbations. In our experiments, we trained a multiclass classifier $\hat{f}$ on the MNIST dataset (Deng, 2012). We fixed the manipulation function to be the Carlini & Wagner transformation (Carlini & Wagner, 2017), i.e., for all $\hat{x}, \delta \in \mathbb{R}^{d_{\text{img}}}$,

$$\psi(\hat{x}, \delta) = \frac{1}{2} \tanh\left(\tanh^{-1}(2\hat{x}) + \delta\right).$$

We considered D to be the ℓ_1 norm as the distortion measure, i.e., for all $\hat{x}, \delta \in \mathbb{R}^{d_{\text{img}}}$,

$$D(\hat{x}, \delta) = \|\hat{x} - \psi(\hat{x}, \delta)\|_1.$$

The accuracy criterion L is fixed to be the black-box untargeted loss proposed in Chen et al. (2017), i.e.,

$$L\left(y, \hat{f}(\hat{x})\right) = \max\left(\log\left[\hat{f}(\hat{x})\right]_y - \max_{t \neq y} \log\left[\hat{f}(\hat{x})\right]_t, -\kappa\right),$$

with $\kappa = -1.0$. Informally, this loss measures how much the confidence of predicting the true class y is smaller than the largest confidence of classfing the input in one of the other classes. If the true class already has the lowest confidence among all classes, the loss is clipped at $-\kappa$ to prevent excessively negative values. By minimizing this loss, the optimizer increases the relative confidence of an incorrect class compared to the true class, encouraging missclassification. We therefore consider the following bilevel optimization problem.

$$\min_{A \in \mathbb{R}^{d_{\text{img}} \times p}} \frac{1}{m} \sum_{i=1}^{m} \|\hat{x}_i - \psi(\hat{x}_i, Az^*(A))\|_1 \quad \text{s.t.} \quad z^*(A) \in \arg\min_{z \in \mathbb{R}^p} \frac{1}{m} \sum_{i=1}^{m} L(y_i, \hat{f}(\psi(\hat{x}_i, Az))).$$

In our experiments, since we considered the MNIST dataset, the number of pixels in the images is $d_{\text{img}} = 784$, and we fixed $p = 100$.

Data Preprocessing. We downloaded the training and test sets of MNIST dataset (Deng, 2012) and we normalized the images to be in $[-0.5, 0.5]^{d_{\text{img}}}$ as suggested in previous works - see e.g. (Liu et al., 2018; Rando et al., 2025a).

Network Architecture. The multiclass classifier has been implemented as a convolutional neural network (CNN), following the architecture described in Rando et al. (2025a). The network consists of five convolutional layers with ReLU activations, interleaved with max-pooling layers, and followed by fully connected layers. Specifically, the first two convolutional layers use 3×3 kernels. The first layer has 1 input channel and 32 output channels, while the second layer has 32 input channels and 64 output channels. These two layers are followed by a 2×2 max-pooling layer with stride 2. The third and fourth convolutional layers have 64 input and output and use 3×3 kernels. A second 2×2 max-pooling layer with stride 2 is applied thereafter. The resulting feature maps are flattened and passed through two fully connected layers, each with 200 hidden units and ReLU activation. The final output layer is a fully connected layer with 10 units and a softmax activation, corresponding to the 10 classes of the MNIST dataset. The network architecture is summarized in Table 4.

Table 4: CNN architecture

Layer	Type	Parameters	Activation
1	Convolutional	in=1, out=32, kernel=3×3	ReLU
2	Convolutional	in=32, out=64, kernel=3×3	ReLU
3	Max Pooling	kernel=2×2, stride=2	–
4	Convolutional	in=64, out=64, kernel=3×3	ReLU
5	Convolutional	in=64, out=64, kernel=3×3	ReLU
6	Max Pooling	kernel=2×2, stride=2	–
7	Dense	in=1024, out=200	ReLU
8	Dense	in=200, out=200	ReLU
9	Dense	in=200, out=10	Softmax

Training Details. We describe here the training procedure and hyperparameters used to train the classifier. In particular, we adopt the defensive distillation technique introduced in Papernot et al. (2016) to improve robustness against adversarial perturbations. Defensive distillation is a training procedure based on a teacher-student framework. Let $\mathcal{D} = \{(\hat{x}_i, y_i)\}_{i=1}^n$ denote the labeled training dataset, where $\hat{x}_i \in \mathbb{R}^{d_{\text{img}}}$ and $y_i \in \{1, \ldots, C\}$ are the data points and labels respectively, with $C > 1$

denoting the number of classes. In the first stage, a teacher network $\hat{f}_{\text{teacher}} : \mathbb{R}^{d_{\text{img}}} \to \mathbb{R}^C$ is trained using a softened softmax output with temperature $T > 1$. Specifically, given the logits $x^{\log} \in \mathbb{R}^C$ produced by the network for an input $\hat{x}$, the temperature-scaled softmax is defined as

$$[\text{softmax}_T(x^{\log})]_i = \frac{\exp(x_i^{\log}/T)}{\sum_{j=1}^{C} \exp(x_j^{\log}/T)}, \quad i = 1, \ldots, C.$$

The teacher model is trained by minimizing the standard cross-entropy loss between the softened predictions and the ground-truth labels. In the second stage, the trained teacher network is used to generate soft labels $\bar{y}_i = \text{softmax}_T(\hat{f}_{\text{teacher}}(\hat{x}_i))$ for $i = 1, \ldots, n$ and a student network $\hat{f}_{\text{student}}$, with the same architecture as the teacher, is then trained to mimic such soft predictions by minimizing the cross-entropy loss with respect to these instead of the hard labels y_i. After training, the temperature is reset to $T = 1$ and the student model is used for inference. In our experiments, the temperature was set to $T = 100$ during both the teacher and student training stages, as suggested by Papernot et al. (2016). Both the student and teacher networks are initialized from the weights obtained after one SGD step with temperature $T = 1$, starting from random initialization, with all weights using Kaiming uniform initialization (He et al., 2015) and biases set to zero. The training dataset is split into two disjoint subsets, with 80% of the data used for training and 20% for validation. All models were trained on training part using the SGD optimizer with learning rate 0.01, momentum 0.9, and batch size 32. Early stopping based on the validation loss was employed, with a patience of 5 epochs, as suggested by Rando et al. (2025a). Training was performed for up to 100 epochs. Additionally, dropout with rate 0.8 was applied to the last hidden layer before the output layer. A summary of the hyperparameters is provided in Table 5. After training, we assessed the performance of the student model by evaluating its training and test accuracy, achieving a training accuracy of 99.16% and a test accuracy of 98.87%.

Table 5: Training hyperparameters

Parameter	
Optimizer	SGD with momentum
Learning rate	0.01
Momentum	0.9
Batch size	32
Max epochs	100
Temperature	100
Early stopping	Patience of 5 epochs (on validation loss)
Dropout	0.8

Parameter Tuning. The procedure to tune parameters of the optimization algorithms is the same of the one adopted in the synthetic problem. Parameters are tuned with grid search. For every algorithm and every parameter configuration of the respective grid, we run such an algorithm 6 times with a fixed budget of 10^5 function evaluations. Since we cannot access to the function values of Ψ, we selected the best configuration of parameters for each algorithm as the one which minimize the average classification accuracy all over the runs computed at the last iteration performed. Initializations of x_0 and z_0 are sampled from Gaussian distributions with 0 mean and 0.5 of standard deviation while the initialization of v_0 is set to the zero vector. As in the synthetic problem, algorithms with optimal configuration are re-executed 5 times to produce the plots in Section 4. Parameter grids used are the same for the synthetic problems (see Table 3) with the only difference that the stepsize parameters are tuned over the set $\{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$.

Additional Experiments. We repeated the adversarial experiment of Section 4 by constructing universal perturbations for images labeled with $y = 1$, $y = 6$, and $y = 9$. In Figures 7 and 8, we report, for each group of images with labels 1, 6, 4 (used in Section 4), and 9, the normalized outer (distortion) and inner (confidence) objective values along the sequences x_k and z_k, together with classification accuracy, as functions of the number of stochastic function evaluations (top row of each plot) and wall-clock time (bottom row of each plot). As in Section 4, values plotted against the number of stochastic evaluations are repeated according to the number of evaluations performed per

iteration. From Figures 7 and 8, we observe again that our methods achieve comparable performance to state-of-the-art algorithms in significantly less time. In particular, ZOBA requires a large number of function evaluations per iteration, since obtaining reliable estimators of Hessian blocks is costly. This limits performance in budget-constrained settings, where only a few updates can be performed. In contrast, HF-ZOBA avoids this limitation by requiring fewer function evaluations per iteration, as it approximates Hessian–vector products using finite differences rather than explicitly estimating Hessian blocks. Moreover, by leveraging delayed information and a single-loop structure, HF-ZOBA reuses function evaluations to compute multiple terms of the search direction while retaining the benefits of parallelization. As a result, HF-ZOBA runs faster than other approaches while maintaining comparable performance, and without relying on a regularized version of the problem (as in Opt-ZMDSBA (Aghasi et al., 2025)), which can introduce additional bias and reduce performance.

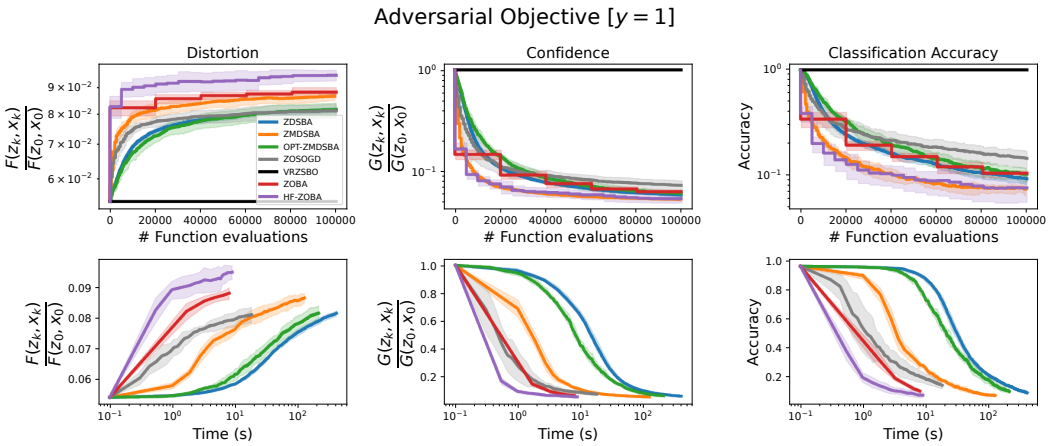


Figure 7: Comparison of algorithms on minimal-distortion universal perturbation attacks across different test sets with different labels $y = 1$. For each test set, we report the normalized outer function values, inner function values at iterations z_k, x_k, and classification accuracy. The top row of each block shows these metrics as a function of the number of function evaluations, while the bottom row shows them versus wall-clock time (in seconds).

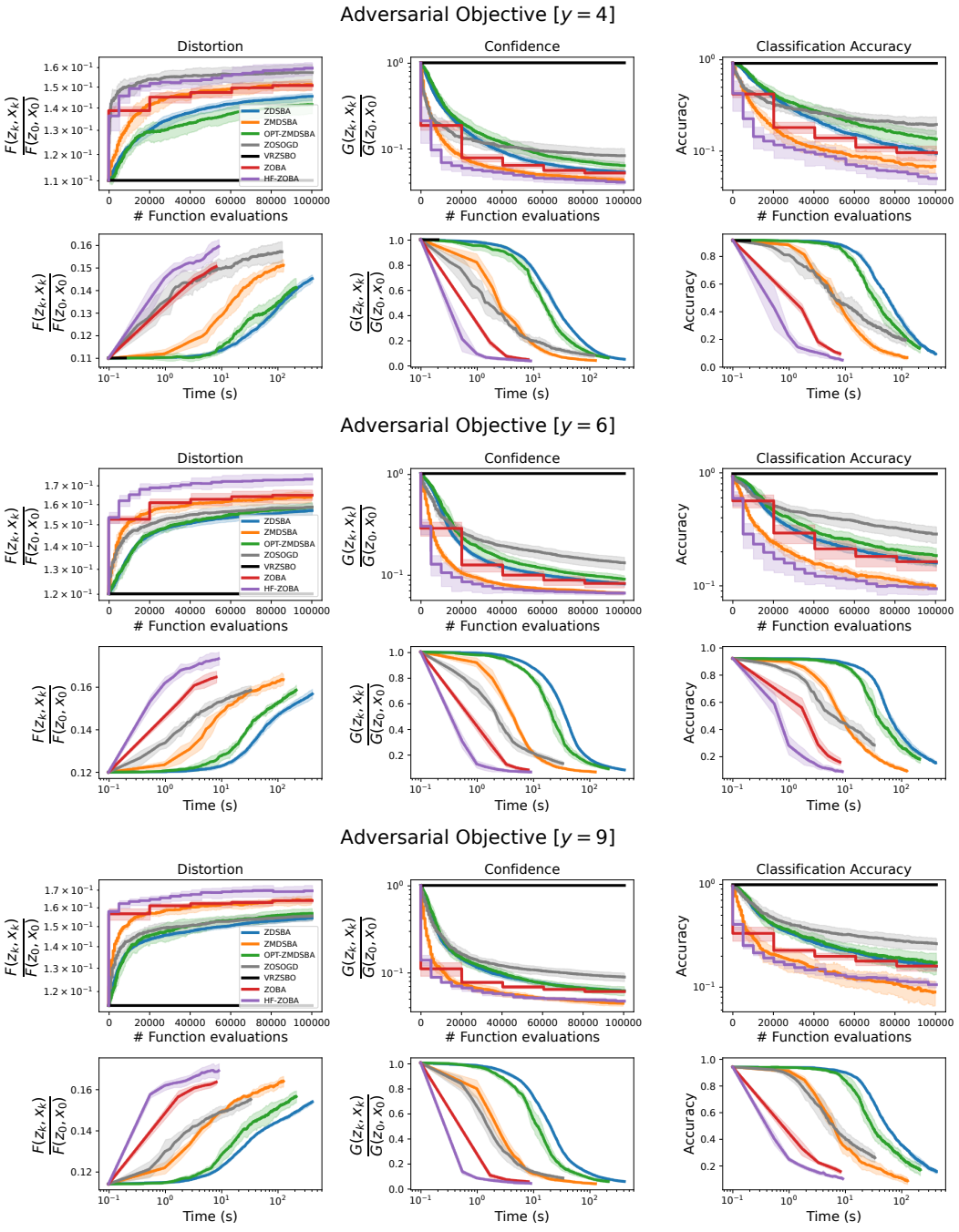


Figure 8: Comparison of algorithms on minimal-distortion universal perturbation attacks across different test sets with labels $y = 4$, $y = 6$ and $y = 9$. We report the normalized outer function values, inner function values at iterations z_k, x_k, and classification accuracy. The top row of each block shows these metrics as a function of the number of function evaluations, while the bottom row shows them versus wall-clock time (in seconds).

C.3 Additional Experiments

In this appendix, we provide additional experiments to further support the empirical findings presented in Section 4. Specifically, we compare ZOBA and HF-ZOBA against ZDSBA, ZMDSBA, Opt-

ZMDSBA, ZOSOGD, and VRZSBO on two additional synthetic problems, and on the hyper-representation task of Grazzi et al. (2020). We repeat each experiment 11 times and we report the mean and standard deviation of results across the last 10 repetitions i.e. discarding the first run to avoid the effect of library loading overhead.

Other Synthetic Experiments. Here, we compare the algorithms in solving Problem 1 considering two synthetic targets. These are constructed as the synthetic problem eq. 11 but with different regularizer in the outer problem. The first is defined as follows.

$$\min_{x\in\mathbb{R}^d} \Psi(x) := F(z^*(x), x) := \frac{1}{2n}\sum_{i=1}^{n} \left(C_i z^*(x) - D_i x - b_i\right)^2 + \sum_{i=1}^{d} \log\cos(x_i - \bar{x}_i)$$

$$\text{s.t.} \quad z^*(x) \in \arg\min_{z\in\mathbb{R}^p} G(z, x) := \frac{1}{2m}\sum_{j=1}^{m} \left(A_j z - B_j x - a_j\right)^2 .$$

The second instead is defined as follows.

$$\min_{x\in\mathbb{R}^d} \Psi(x) := F(z^*(x), x) := \frac{1}{2n}\sum_{i=1}^{n} \left(C_i z^*(x) - D_i x - b_i\right)^2 + \frac{1}{4}\sum_{i=1}^{d} (x_i - \bar{x}_i)^4$$

$$\text{s.t.} \quad z^*(x) \in \arg\min_{z\in\mathbb{R}^p} G(z, x) := \frac{1}{2m}\sum_{j=1}^{m} \left(A_j z - B_j x - a_j\right)^2 .$$

Matrices $(A_j)_{j=1}^m, (B_j)_{j=1}^m, (C_i)_{i=1}^n$ and $(D_i)_{i=1}^m$ and vectors $a \in \mathbb{R}^m, b \in \mathbb{R}^n$ are constructed as for the synthetic problem eq. 11. We set a budget of 10^5 function evaluations and consider the setting $d = p \in \{10, 50, 100\}$. In Figure 9 (top row), we plot, for each problem, the normalized optimality gap as a function of the number of function evaluations. Overall, these results are consistent with the observations reported in Section 4. In particular, we observe that in the small-dimensional regime ($d = p = 10$), all methods (except ZDSBA) achieve comparable performance. ZDSBA performs worse, in agreement with the theoretical findings. Moreover, in this setting, single-loop methods tend to produce slightly more accurate solutions than two-loop methods. This because two-loop methods have to allocate an important part of the evaluation budget to inner-loop iterations, which are required to control the accuracy of the approximations of z^* and v^*. In contrast, single-loop methods regulate this approximation quality through the batch size (and, in our case, through both batch size and the number of directions). As commonly observed in prior empirical studies on finite-difference methods, a larger number of iterations with inexpensive estimators is typically required to match the accuracy achieved by estimators that rely on larger batch sizes or more directions (Mania et al., 2018; Duchi et al., 2015; Salimans et al., 2017; Cai et al., 2021; 2022; Rando et al., 2023; 2024; 2025a;b). This allows single-loop methods to perform more update of the iterates resulting in more accurate solutions. Notice that, in this low-dimensional regime, VRZSBO and ZOSOGD perform comparably to our methods, whereas they provided worse performance in Section 4. This discrepancy is due to the lower dimensionality: despite their unfavorable cost scaling, these methods remain competitive when d and p are small. However, their performance degrades as the dimension increases. For VRZSBO, this is because each estimator requires all coordinate directions, implying that the number of function evaluations per surrogate scales at least linearly with the dimension times the batch size. This enforce thus to fast deplete the function evaluation budget in few iterations. For ZOSOGD, the degradation stems from poor cost scaling with respect to the batch size (notably due to the momentum computation). As the dimension increases, a larger batch size would be desirable to reduce impact of stochasticity and thus allowing larger stepsize. However, under a limited budget, the grid search selects a small batch size since larger batchsize would fast exhaust the budget. This forces to choose smaller step sizes to control stochasticity, slowing convergence. Notice that in higher-dimensional settings, ZOSOGD performs worse even than ZDSBA. While both methods employ single-direction, single-sample estimators, ZDSBA can improve approximation accuracy through additional inner iterations, enabling the use of larger step sizes. In contrast, ZOSOGD, being a single-loop method, must rely solely on the batch size to control estimator quality. Our methods, instead, better benefit from function value reuse, leading to improved cost scaling with respect to both the batch size and the number of directions. This enables more effective control of the estimator quality allowing to choose a larger batchsize in higher-dimensional setting (and thus a larger stepsize) while not exhausting the function evaluation budget in few iterations making them more effective in such settings. In the

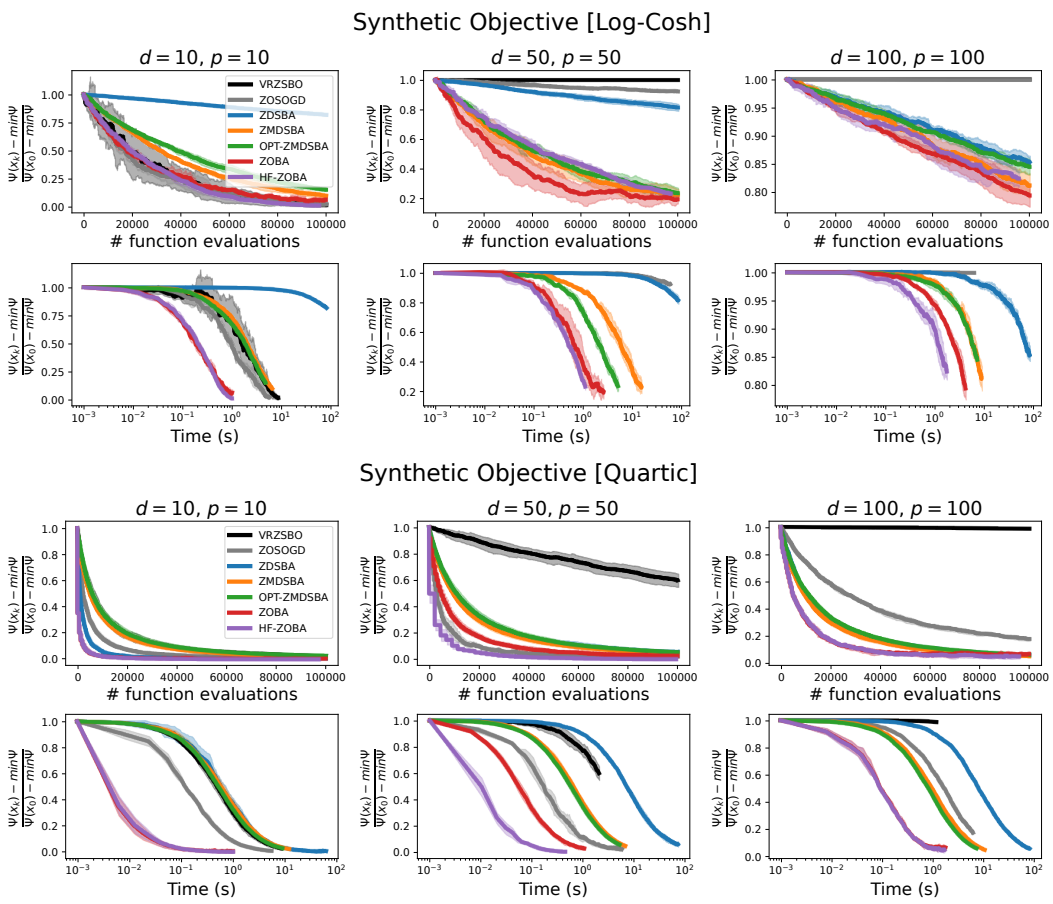


Figure 9: Comparison of algorithms on synthetic bilevel problems with varying dimensions: optimality gap versus function evaluations (top row) and wall-clock time in seconds (bottom row).

bottom rows of Figure 9, we report the optimality gap as a function of wall-clock time. These results confirm that our methods are significantly faster than prior approaches. This improvement is due to a combination of factors: efficient parallelization, reuse of function evaluations, and the flexibility to tune both the batch size and the number of directions, which allows for finer control over estimator quality. Overall, these findings highlight the practical advantages of our approach.

Hyper-representation. We compare the performance of the algorithms on the hyper-representation task proposed by Grazzi et al. (2020). Given two datasets (X, y) and (X', y'), the goal is to learn a shared linear transformation of the input data (Franceschi et al., 2018b; Bertinetto et al., 2019), i.e., a representation map that reshapes the feature space so as to improve generalization across datasets. In particular, we aim to learn a mapping H that transforms the inputs into a representation in which a predictor trained on (X, y) transfers effectively to (X', y'). Formally, we want to solve the following bilevel optimization problem.

$$\min_{H \in \mathbb{R}^{d' \times p}} \Psi(H) := F(w^*(H), H) := \frac{1}{2}\|X' H w^*(H) - y'\|^2$$
$$\text{s.t.} \quad w^*(H) \in \arg\min_{w \in \mathbb{R}^p} G(w, H) := \frac{1}{2}\|XHw - y\|^2 + \frac{\lambda}{2}\|w\|^2.$$

In this formulation, H defines a shared feature representation of the input data. The lower-level problem learns a predictor $w^*(H)$ on the transformed training data (XH, y), while the upper-level objective evaluates how well this representation transfers by measuring the performance of $w^*(H)$ on the validation dataset (X', y'). Hence, the outer problem optimizes the representation itself

rather than the predictor, encouraging representations that generalize across datasets. Although such problems are typically addressed using first-order methods, we include this setting as an additional benchmark to further support our empirical findings. For these experiments, we fix a budget of 5×10^4 function evaluations and consider the setting $d' = p$.

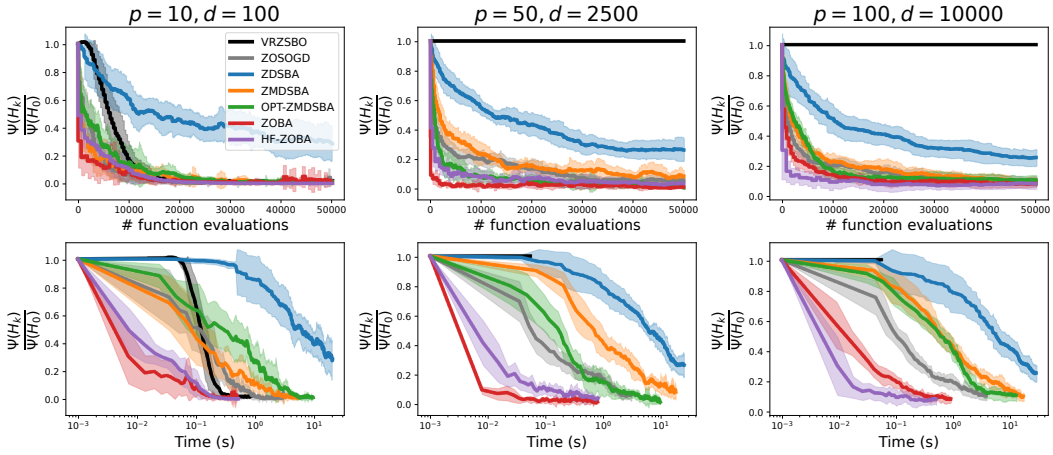


Figure 10: Comparison of algorithms in solving hyper-representation task. Normalized function values are reported against number of function evaluations (top row) and wall-clock time (bottom row).

In Figure 10 (top row), we report the normalized function value as a function of the number of function evaluations. The results are consistent with the observations from previous experiments. In the low-dimensional regime ($p = 10$ and $d = 100$), all methods except ZDSBA achieve comparable performance. As the dimension increases, single-loop methods (with the exception of VRZSBO, due to its high per-iteration cost) tend to perform slightly better than two-loop approaches, due to the fact that these approaches must allocate a larger portion of the budget to inner iterations in order to maintain approximation accuracy. Notice that, when dimension increases, our methods significantly perform better than VRZSBO, which requires a large number of function evaluations to construct its surrogates. In Figure 10 (bottom row), we plot the normalized function value against wall-clock time. We observe that our methods achieve substantially better performance in terms of runtime compared to competing approaches, particularly in higher dimensions. This further confirms the benefits of parallelism and the practical efficiency of our algorithms.

D Auxiliary Results

In this appendix, we introduce the main concepts and report and prove the lemmas and propositions required to prove the main theorems.

Notation. In the following, we denote by I_p the identity matrix $p \times p$. For a function $f : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$, we indicate with $\nabla_z f(z, x)$ and $\nabla_x f(z, x)$, the gradient with respect to the first and second variable respectively. Moreover, we denote with $\mathbb{E}_k[\cdot]$ the conditional expectation conditioning on the entire history of every random variable $(w_t^{(i,j)})_{t<k}, (u_t^{(i,j)})_{t<k}, (\xi_{i,t})_{t<k}$ and $(\zeta_{i,t})_{t<k}$ for every $i = 1, \cdots, b_1$ and $j = 1, \cdots, \ell_1$ i.e. the expectation is taken only on random variables at iteration k

$$\mathbb{E}_k[\cdot] := \mathbb{E}_{\xi_{1,k},\zeta_{1,k}}\left[\cdots\left[\mathbb{E}_{\xi_{b_1,k},\zeta_{b_1,k}}\left[\mathbb{E}_{w_k^{(1,1)},u_k^{(i,j)}}\left[\cdots\mathbb{E}_{w_k^{(b_1,\ell_1)},u_k^{(b_1,\ell_1)}}[\cdot]\right]\right]\right]\right].$$

Moreover, we denote with $\mathbb{E}_{W_k^{(i)}}[\cdot]$ the conditional expectation on $w_k^{(i,j)}$ for $j = 1, \cdots, \ell_1$ i.e. the expectation is taken only on $w_k^{(i,j)}$ for every $j = 1, \cdots, \ell_1$ (i.e. i is fixed)

$$\mathbb{E}_{W_k^{(i)}}[\cdot] := \mathbb{E}_{w_k^{(i,1)}}\left[\mathbb{E}_{w_k^{(i,2)}}\left[\cdots\mathbb{E}_{w_k^{(i,\ell_1)}}[\cdot]\right]\right].$$

Similarly, we denote with $\mathbb{E}_{W_k}[\cdot]$ the conditional expectation on $w_k^{(i,j)}$ for every $i = 1, \cdots, b_1$ and $j = 1, \cdots, \ell_1$ i.e. the expectation is taken only on $w_k^{(i,j)}$ for every $i = 1, \cdots, b_1$ and $j = 1, \cdots, \ell_1$

$$\mathbb{E}_{W_k}[\cdot] := \mathbb{E}_{W_k^{(1)}}\left[\cdots \mathbb{E}_{W_k^{(b_1)}}[\cdot]\right] = \mathbb{E}_{w_k^{(1,1)}}\left[\mathbb{E}_{w_k^{(1,2)}}\left[\cdots \mathbb{E}_{w_k^{(1,\ell_1)}}\left[\mathbb{E}_{w_k^{(2,1)}}\left[\cdots \mathbb{E}_{w_k^{(b_1,\ell_1)}}[\cdot]\right]\right]\right]\right].$$

For $i = 1, \cdots, b_2$, we denote with

$$\mathbb{E}_{U_k^{(i)}}[\cdot] = \mathbb{E}_{u_k^{(i,1)}}\left[\mathbb{E}_{u_k^{(i,2)}}\left[\cdots \mathbb{E}_{u_k^{(i,\ell_2)}}[\cdot]\right]\right] \quad \text{and} \quad \mathbb{E}_{U_k}[\cdot] = \mathbb{E}_{U_k^{(1)}}\left[\cdots \mathbb{E}_{U_k^{(b_2)}}[\cdot]\right].$$

Similarly, we denote with

$$\mathbb{E}_{\xi_k}[\cdot] = \mathbb{E}_{\xi_{1,k}}[\cdots \mathbb{E}_{\xi_{b_1,k}}[\cdot]] \quad \text{and} \quad \mathbb{E}_{\zeta_k}[\cdot] = \mathbb{E}_{\zeta_{1,k}}[\cdots \mathbb{E}_{\zeta_{b_2,k}}[\cdot]].$$

Lemma D.1. *Let $u \sim \mathcal{N}(0, I_d)$. Then, for $p \in [0, 2]$,*

$$\mathbb{E}_u[\|u\|^p] \leq d^{p/2}.$$

For $p \geq 2$,

$$\mathbb{E}_u[\|u\|^p] \leq (d+p)^{p/2}.$$

Moreover, let $A, B \in \mathbb{R}^{d\times d}$ be symmetric matrices. Then,

$$\mathbb{E}_u[u^\top A u] = tr(A), \quad \mathbb{E}_u[u^\top A u u^\top B u] = 2tr(AB) + tr(A)tr(B),$$
$$\mathbb{E}_u[uu^\top uu^\top] = (d+2)I_d.$$

where $tr(A)$ denotes the trace of matrix A.

Proof. These are standard results for Gaussian variables and can be found in several textbooks - see e.g. (Nesterov & Spokoiny, 2017; Petersen et al., 2008; Seber & Lee, 2003). □

The following appendices collect auxiliary results used in the proofs of the main theorems and corollaries. In Appendix D.1, we present lemmas establishing bounds and regularity properties of the quantities arising in bilevel optimization under the considered assumptions. In Appendix D.2, we review the Gaussian smoothing framework for the bilevel problem and states the associated lemmas and properties used to analyze our algorithms. In Appendix D.3, we provide bounds on the sequences generated by Algorithm 1 together with the corresponding descent lemma and in Appendix D.4 we report analogous results for Algorithm 2.

D.1 Auxiliary results for Bilevel Optimization

Lemma D.2 (Regularity of Ψ, z^* and v^*)**.** *Under Assumptions 3.2,3.3, the following holds:*
(I) the function Ψ is L_Ψ-smooth for some $L_\Psi > 0$.
(II) there exists $L_* > 0$ such that for every $x_1, x_2 \in \mathbb{R}^d$,

$$\|z^*(x_1) - z^*(x_2)\| \leq L_*\|x_1 - x_2\| \quad \textit{and} \quad \|v^*(x_1) - v^*(x_2)\| \leq L_*\|x_1 - x_2\|,$$

where $z^*(\cdot) \in \operatorname*{arg\,min}_{z\in\mathbb{R}^p} G(z, \cdot)$ and $v^*(\cdot)$ is defined in eq. 3.

Proof. The point (I) has been proved by Ghadimi & Wang (2018) and (II) has been proved in Dagréou et al. (2022). □

Lemma D.3 (Bound on norm of v^*)**.** *Under Assumptions 3.2, 3.3. Let $v^*(\cdot)$ be defined as in eq. 3. Then, for every $x \in \mathbb{R}^d$,*

$$\|v^*(x)\| \leq \frac{L_{0,F}}{\mu_G}.$$

Proof. Since $v^*(x) = -[\nabla^2_{zz} G(z^*(x), x)]^{-1}\nabla_z F(z^*(x), x)$, G is μ_G-strongly convex and F is $L_{0,F}$-Lipschitz continuous, we have

$$\|v^*(x)\| = \left\|[\nabla^2_{zz} G(z^*(x), x)]^{-1}\nabla_z F(z^*(x), x)\right\| \leq \frac{L_{0,F}}{\mu_G}.$$

□

D.2 Gaussian Smoothing for Bilevel Optimization

The analysis of Algorithm 1 and 2 rely on the fact that the surrogates used to construct the search directions (in eq. 4, eq. 6 and eq. 8 for Algorithm 1 and eq. 10 for Algorithm 2) are unbiased estimators of gradients and Hessians of a smooth approximation of the inner and outer objective functions f and g. More precisely, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{Z}$ be a measurable space. Let $f : \mathbb{R}^p \times \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ and let $\zeta : \Omega \to \mathcal{Z}$ be a random varianble. Then for any $h, \eta \geq 0$, we define the following smoothed version of f for every $z \in \mathbb{R}^p$, $x \in \mathbb{R}^d$ and every realization of ζ.

$$f_{h,\eta}(z, x, \zeta) = \mathbb{E}_{w,u}\left[f(z + hw, x + \eta u, \zeta)\right], \tag{12}$$

where $w \sim \mathcal{N}(0, I_p)$ and $u \sim \mathcal{N}(0, I_d)$. For $h, \eta > 0$, the function $f_{h,\eta}$ is differentiable even when f is not, and its properties depends on those of f (Aghasi & Ghadimi, 2025) - see Proposition D.4. Note that this construction differs from the standard Gaussian smoothing considered in (Nesterov & Spokoiny, 2017), as it applies distinct smoothing parameters to different variable blocks. This flexibility is advantageous in bilevel optimization, where we want to compute gradients and hessian only a of subset of variables. In the following proposition we show some properties of the smoothing.

Proposition D.4 (Properties of the smoothing)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{Z}$ be a measurable space. Let $f : \mathbb{R}^p \times \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ and let $f_{h,\eta}$ be the smooth surrogate of f defined as in eq. 12. Then, the following hold:*
(I) If f is convex (or μ_f strongly convex) then $f_{h,\eta}$ is convex (or μ_f strongly convex).
(II) If f is L_0-Lipschitz then we have that $f_{h,\eta}$ is L_0-Lipschitz.
(III) If f is L_1-smooth then we have that $f_{h,\eta}$ is L_1-smooth.

Proof. The results on convexity, Lipschitz continuity, and smoothness were proved in Aghasi & Ghadimi (2025). □

In the following lemma we show that gradients and hessians of the smoothed approximation of a target can be expressed as expectation of finite difference.

Lemma D.5 (Gradient and Hessian of Smoothing)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{Z}$ be a measurable space. Let $\zeta : \Omega \to \mathcal{Z}$, and let $w \sim \mathcal{N}(0, I_p)$ and $u \sim \mathcal{N}(0, I_d)$. Let $f : \mathbb{R}^p \times \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$, and let $f_{h,\eta}$ denote the smooth surrogates of f defined as in eq. 12. Then, for every $z \in \mathbb{R}^p$, and $x \in \mathbb{R}^d$,*

$$\begin{aligned} \nabla_z f_{h,\eta}(z, x, \zeta) &= \mathbb{E}_{w,u}\left[\frac{f(z + hw, x + \eta u, \zeta)}{h} w\right] \\ \nabla_x f_{h,\eta}(z, x) &= \mathbb{E}_{w,u}\left[\frac{f(z + hw, x + \eta u, \zeta)}{\eta} u\right]. \end{aligned} \tag{13}$$

Moreover,

$$\begin{aligned} \nabla^2_{zz} f_{h,\eta}(z, x, \zeta) &= \mathbb{E}_{w,u}\left[\frac{f(z + hw, x + \eta u, \zeta)}{h^2}(ww^\top - I_p)\right], \\ \nabla^2_{xz} f_{h,\eta}(z, x, \zeta) &= \mathbb{E}_{w,u}\left[\frac{f(z + hw, x + \eta u, \zeta)}{\eta h} uw^\top\right]. \end{aligned} \tag{14}$$

Proof. Although this result was stated in Aghasi & Ghadimi (2025), no explicit proof was given. We therefore provide a proof here for completeness. By the definition of the smooth surrogate (eq. 12), we have

$$f_{h,\eta}(z, x, \zeta) = \mathbb{E}_{w,u}[f(z + hw, x + \eta u, \zeta)].$$

Since w, u are independent and f is absolutely measurable, by Fubini's Theorem, we have

$$f_{h,\eta}(z, x, \zeta) = \mathbb{E}_{w,u}[f(z + hw, x + \eta u, \zeta)] = \mathbb{E}_w[\mathbb{E}_u[f(z + hw, x + \eta u, \zeta)]]. \tag{15}$$

Now, we prove eq. 13. Let $y = z + hw$. Thus, by substitution, we have

$$\begin{aligned} f_{h,\eta}(z, x, \zeta) &= \frac{1}{(\sqrt{2\pi})^{p/2}} \int \mathbb{E}_u[f(z + hw, x + \eta u, \zeta)] e^{-\frac{\|w\|^2}{2}} dw \\ &= \frac{1}{(\sqrt{2\pi})^{p/2}} \int \mathbb{E}_u[f(y, x + \eta u, \zeta)] e^{-\frac{\|y - x\|^2}{2h^2}} \frac{dy}{h^p}. \end{aligned}$$

Therefore, computing the gradient in z, we have

$$\begin{aligned}
\nabla_z f_{h,\eta}(z,x,\zeta) &= \frac{1}{(\sqrt{2\pi})^{p/2}} \int \mathbb{E}_u[f(y, x+\eta u, \zeta)] \nabla_z e^{-\frac{\|y-z\|^2}{2h^2}} \frac{dy}{h^p} \\
&= \frac{1}{(\sqrt{2\pi})^{p/2}} \int \mathbb{E}_u[f(y, x+\eta u, \zeta)] \left(\frac{y-z}{h^2}\right) e^{-\frac{\|y-z\|^2}{2h^2}} \frac{dy}{h^p} \\
&= \frac{1}{(\sqrt{2\pi})^{p/2}} \int \mathbb{E}_u \left[\frac{f(z+hw, x+\eta u, \zeta)}{h} w\right] e^{-\frac{\|w\|^2}{2}} dw \\
&= \mathbb{E}_{w,u}\left[\frac{f(z+hw, x+\eta u, \zeta)}{h} w\right].
\end{aligned}$$

Similarly we can compute the gradient in x i.e. $\nabla_x f_{h,\eta}(z,x,\zeta)$. Now, we prove eq. 14. Setting $y = z + hw$, by substitution,

$$f_{h,\eta}(z,x,\zeta) = \frac{1}{(\sqrt{2\pi})^{p/2}} \int \mathbb{E}_u[f(y, x+\eta u, \zeta)] e^{-\frac{\|y-z\|^2}{2h^2}} \frac{dy}{h^p}. \tag{16}$$

Computing the Hessian block in zz, we have

$$\begin{aligned}
\nabla^2_{zz} f_{h,\eta}(z,x,\zeta) &= \frac{1}{(\sqrt{2\pi})^{p/2}} \int \mathbb{E}_u[f(y, x+\eta u, \zeta)] \nabla^2_{zz} \left(e^{-\frac{\|y-z\|^2}{2h^2}}\right) \frac{dy}{h^p} \\
&= \frac{1}{(\sqrt{2\pi})^{p/2}} \int \mathbb{E}_u[f(y, x+\eta u, \zeta)] \left(\frac{(y-z)(y-z)^\top}{h^4} - \frac{1}{h^2} I_p\right) e^{-\frac{\|y-z\|^2}{2h^2}} \frac{dy}{h^p} \\
&= \frac{1}{(\sqrt{2\pi})^{p/2}} \int \mathbb{E}_u \left[\frac{f(y, x+\eta u, \zeta)}{h^2} \left(ww^\top - I_p\right)\right] e^{-\frac{\|w\|^2}{2}} dw \\
&= \mathbb{E}_{w,u}\left[\frac{f(y, x+\eta u, \zeta)}{h^2} \left(ww^\top - I_p\right)\right].
\end{aligned}$$

Now, to get the expression of the cross-block, we restart from eq. 16. Let $\bar{y} = x + \eta u$, by substitution we have

$$\begin{aligned}
f_{h,\eta}(z,x,\zeta) &= \frac{1}{(\sqrt{2\pi})^{p/2}} \frac{1}{(\sqrt{2\pi})^{d/2}} \int \left(\int f(y, x+\eta u, \zeta) e^{-\frac{\|u\|^2}{2}} du\right) e^{-\frac{\|y-z\|^2}{2h^2}} \frac{dy}{h^p} \\
&= \frac{1}{(\sqrt{2\pi})^{p/2}} \frac{1}{(\sqrt{2\pi})^{d/2}} \int \left(\int f(y, \bar{y}, \zeta) e^{-\frac{\|\bar{y}-x\|^2}{2\eta^2}} \frac{d\bar{y}}{\eta^d}\right) e^{-\frac{\|y-z\|^2}{2h^2}} \frac{dy}{h^p} \\
&= \frac{1}{(\sqrt{2\pi})^{p/2}} \frac{1}{(\sqrt{2\pi})^{d/2}} \int \int f(y, \bar{y}, \zeta) e^{-\frac{\|\bar{y}-x\|^2}{2\eta^2}} e^{-\frac{\|y-z\|^2}{2h^2}} \frac{d\bar{y}}{\eta^d} \frac{dy}{h^p}.
\end{aligned}$$

Computing the mixed block xz we have,

$$\begin{aligned}
\nabla^2_{xz} f_{h,\eta}(z,x,\zeta) &= \frac{1}{(\sqrt{2\pi})^{p/2}} \frac{1}{(\sqrt{2\pi})^{d/2}} \int \int f(y, \bar{y}, \zeta) \nabla^2_{xz} \left(e^{-\frac{\|\bar{y}-x\|^2}{2\eta^2}} e^{-\frac{\|y-z\|^2}{2h^2}}\right) \frac{d\bar{y}}{\eta^d} \frac{dy}{h^p} \\
&= \frac{1}{(\sqrt{2\pi})^{p/2}} \frac{1}{(\sqrt{2\pi})^{d/2}} \int \int f(y, \bar{y}, \zeta) \left(\frac{(y-z)(\bar{y}-x)^\top}{h^2\eta^2}\right) e^{-\frac{\|\bar{y}-x\|^2}{2\eta^2}} e^{-\frac{\|y-z\|^2}{2h^2}} \frac{d\bar{y}}{\eta^d} \frac{dy}{h^p} \\
&= \frac{1}{(\sqrt{2\pi})^{p/2}} \frac{1}{(\sqrt{2\pi})^{d/2}} \int \int \frac{f(z+hw, x+\eta u, \zeta)}{h\eta} uw^\top e^{-\frac{\|u\|^2}{2}} e^{-\frac{\|w\|^2}{2}} du dw \\
&= \mathbb{E}_{w,u}\left[\frac{f(z+hw, x+\eta u, \zeta)}{h\eta} uw^\top\right].
\end{aligned}$$

□

Remark D.6. Notice that by symmetry, the gradient of smoothing can be expressed as expectation of forward or central finite difference

$$\begin{aligned}\nabla_z f_{h,\eta}(z,x,\zeta) &= \mathbb{E}_{w,u}\left[\frac{f(z+hw,x+\eta u,\zeta)}{h}w\right]\\ &= \mathbb{E}_{w,u}\left[\frac{f(z+hw,x+\eta u,\zeta)-f(z,x,\zeta)}{h}w\right]\\ &= \mathbb{E}_{w,u}\left[\frac{f(z+hw,x+\eta u,\zeta)-f(z-hw,x-\eta u,\zeta)}{2h}w\right].\end{aligned}$$

Analogously, we can express the gradient with respect to the second variable. Moreover, also the hessian blocks can be expressed as finite difference

$$\begin{aligned}\nabla^2_{zz} f_{h,\eta}(z,x,\zeta) &= \mathbb{E}_{w,u}\left[\frac{f(z+hw,x+\eta u,\zeta)}{h^2}(ww^\top - I_p)\right]\\ &= \mathbb{E}_{w,u}\left[\frac{f(z+hw,x+\eta u,\zeta)+f(z-hw,x-\eta u,\zeta)-2f(z,x,\zeta)}{2h^2}(ww^\top - I_p)\right],\\ \nabla^2_{xz} f_{h,\eta}(z,x,\zeta) &= \mathbb{E}_{w,u}\left[\frac{f(z+hw,x+\eta u,\zeta)}{\eta h}uw^\top\right]\\ &= \mathbb{E}_{w,u}\left[\frac{f(z+hw,x+\eta u,\zeta)+f(z-hw,x-\eta u,\zeta)-2f(z,x,\zeta)}{2\eta h}uw^\top\right].\end{aligned}$$

Notice that if we define $F(z,x)=\mathbb{E}_\zeta[f(z,x,\zeta)]$, we can define the smoothed surrogate $F_{h,\eta}(z,x)$ as in eq. 12 i.e. let $w\sim\mathcal{N}(0,I_p)$ and $u\sim\mathcal{N}(0,I_d)$,

$$F_{h,\eta}(z,x)=\mathbb{E}_{w,u}[F(z+hw,x+\eta u)].$$

Moreover, we have

$$\begin{aligned}F_{h,\eta}(z,x) &= \mathbb{E}_{w,u}[F(z+hw,x+\eta u)]\\ &= \mathbb{E}_{w,u}[\mathbb{E}_\zeta[f(z+hw,x+\eta u,\zeta)]]\\ &= \mathbb{E}_\zeta[\mathbb{E}_{w,u}[f(z+hw,x+\eta u,\zeta)]] = \mathbb{E}_\zeta[f_{h,\eta}(z,x,\zeta)].\end{aligned}$$

In the following lemma, we report the error of using the gradient of the smoothing instead of the exact gradient of the target.

Lemma D.7 (Smoothing Error)**.** *Let $F:\mathbb{R}^p\times\mathbb{R}^d\to\mathbb{R}$ be a $L_{1,F}$-smooth function and let $F_{h,\eta}$ denotes its gaussian smoothed surrogate defined as in eq. 12. Then, for every $z\in\mathbb{R}^p$ and $x\in\mathbb{R}^d$,*

$$\begin{aligned}&(\forall h>0,\eta\geq 0)\quad \|\nabla_z F_{h,\eta}(z,x)-\nabla_z F(z,x)\|\leq\frac{L_{1,F}}{2}(p+3)^{3/2}h+\frac{L_{1,F}\eta^2}{2h}d\sqrt{p}.\\ &(\forall h\geq 0,\eta>0)\quad \|\nabla_x F_{h,\eta}(z,x)-\nabla_x F(z,x)\|\leq\frac{L_{1,F}p\sqrt{d}}{2}\frac{h^2}{\eta}+\frac{L_{1,F}}{2}(d+3)^{3/2}\eta.\end{aligned}\tag{17}$$

Let F be a twice differentiable functions with $L_{2,G}$-Lipschitz continuous hessian. Then for every $h>0$ and $\eta\geq 0$ we have

$$\|\nabla_z F_{h,\eta}(z,x)-\nabla_z F(z,x)\|\leq\frac{2L_{G,2}}{3}\left(h^2(p+4)^2+\frac{\eta^3}{h}\sqrt{p}(d+3)^{2/3}\right),$$

and,

$$\begin{aligned}\|\nabla^2_{zz} F_{h,\eta}(z,x)-\nabla^2_{zz} F(z,x)\|\leq\frac{2L_{2,G}}{3}\Bigg(&h(d+5)^{5/2}+h(d+3)^{3/2}\\ &+\frac{\eta^3}{h^2}(d+1)(p+3)^{3/2}\Bigg).\end{aligned}\tag{18}$$

Proof. These results have been proved in (Aghasi & Ghadimi, 2025). □

Now, building on the previous lemmas, we introduce a result that bounds the error of a smoothed approximate hypergradient computed for arbitrary $z\in\mathbb{R}^p$ and $v\in\mathbb{R}^p$, rather than for the exact inner solution and the true inverse Hessian action.

Lemma D.8 (Bilevel Smoothing Error)**.** *Let Assumptions 3.2 and 3.3 holds. For every $h_1, \eta_1, h_2, \eta_2 > 0$ and every $x \in \mathbb{R}^d, z \in \mathbb{R}^p$, let $G_{h_1,\eta_1}(z,x)$ and $F_{h_2,\eta_2}(z,x)$ be the smooth surrogates of the inner and outer targets in eq. 1 respectively. Then, for every $z, v \in \mathbb{R}^p$ and $x \in \mathbb{R}^d$,*

$$\begin{aligned}\|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)v + \nabla_x F_{h_2,\eta_2}(z,x) - \nabla\Psi(x)\| \le &\left(L_{1,F} + \frac{L_{0,F}L_{1,G}}{\mu_G}\right)\|z - z^*(x)\| \\ &+ L_{1,G}\|v - v^*(x)\| \\ &+ \frac{L_{0,F}}{\mu_G}\frac{2L_{2,G}}{3}\Bigg(h_1(d+5)^{5/2} \\ &+ h_1(d+3)^{3/2} + \frac{\eta_1^3}{h_1^2}(d+1)(p+3)^{3/2}\Bigg) \\ &+ \frac{L_{1,F}p\sqrt{d}}{2}\frac{h_2^2}{\eta_2} + \frac{L_{1,F}}{2}(d+3)^{3/2}\eta_2.\end{aligned}$$

Proof. Since $\nabla\Psi(x) := \nabla_x F(z^*(x), x) + \nabla^2_{2,1} G(z^*(x), x)v^*(x)$, we have

$$\begin{aligned}\|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)v + \nabla_x F_{h_2,\eta_2}(z,x) - \nabla\Psi(x)\| \le &\underbrace{\|\nabla_x F_{h_2,\eta_2}(z,x) - \nabla_x F(z^*(x),x)\|}_{A} \\ &+ \underbrace{\|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)v - \nabla^2_{xz} G(z^*(x),x)v^*(x)\|}_{B}.\end{aligned}$$

To bound A, we add and subtract $\nabla_x F(z,x)$,

$$\begin{aligned}\|\nabla_x F_{h_2,\eta_2}(z,x) - \nabla_x F(z^*(x),x)\| \le &\|\nabla_x F_{h_2,\eta_2}(z,x) - \nabla_x F(z,x)\| \\ &+ \|\nabla_x F(z,x) - \nabla_x F(z^*(x),x)\|.\end{aligned}$$

By eq. 17 and $L_{1,F}$-smoothness we have

$$\|\nabla_x F_{h_2,\eta_2}(z,x) - \nabla_x F(z^*(x),x)\| \le \frac{L_{1,F}p\sqrt{d}}{2}\frac{h_2^2}{\eta_2} + \frac{L_{1,F}}{2}(d+3)^{3/2}\eta_2 + L_{1,F}\|z - z^*(x)\|.$$

To bound B, we add and subtract $\nabla^2_{xz} G_{h_1,\eta_1}(z,x)v^*(x)$,

$$\begin{aligned}\|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)v - \nabla^2_{xz} G(z^*(x),x)v^*(x)\| \le &\underbrace{\|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)v - \nabla^2_{xz} G_{h_1,\eta_1}(z,x)v^*(x)\|}_{C} \\ &+ \underbrace{\|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)v^*(x) - \nabla^2_{xz} G(z^*(x),x)v^*(x)\|}_{D}.\end{aligned}$$

We bound C, by $L_{1,G}$-smoothness of G,

$$\begin{aligned}\|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)v - \nabla^2_{xz} G_{h_1,\eta_1}(z,x)v^*(x)\| &= \|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)(v - v^*(x))\| \\ &\le \|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)\|\|v - v^*(x)\| \\ &\le L_{1,G}\|v - v^*(x)\|.\end{aligned}$$

We bound D, by adding and subtracting $\nabla^2_{xz} G(z,x)v^*(x)$

$$\begin{aligned}\|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)v^*(x) - \nabla^2_{xz} G(z^*(x),x)v^*(x)\| \le &\|(\nabla^2_{xz} G(z,x) - \nabla^2_{xz} G(z^*(x),x))v^*(x)\| \\ &+ \|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)v^*(x) - \nabla^2_{xz} G(z,x)v^*(x)\|\end{aligned}$$

Since hessian of G is $L_{2,G}$-Lipschitz, adding and subtracting $\nabla^2_{xz} G(z,x)v^*(x)$,

$$\begin{aligned}\|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)v^*(x) - \nabla^2_{xz} G(z^*(x),x)v^*(x)\| \le &\|(\nabla^2_{xz} G(z,x) - \nabla^2_{xz} G(z^*(x),x))v^*(x)\| \\ &+ \|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)v^*(x) - \nabla^2_{xz} G(z,x)v^*(x)\| \\ &\le L_{1,G}\|z - z^*(x)\|\|v^*(x)\| \\ &+ \|\nabla^2_{xz} G_{h_1,\eta_1}(z,x) - \nabla^2_{xz} G(z,x)\|\|v^*(x)\|\end{aligned}$$

By Lemma D.7 (eq. 18) and Lemma D.3, we have

$$\|\nabla^2_{xz} G_{h_1,\eta_1}(z,x)v^*(x) - \nabla^2_{xz} G(z^*(x),x)v^*(x)\| \leq \frac{L_{0,F}L_{1,G}}{\mu_G}\|z - z^*(x)\| + \frac{L_{0,F}}{\mu_G}\frac{2L_{2,G}}{3}\bigg(h_1(d+5)^{5/2} + h_1(d+3)^{3/2} + \frac{\eta_1^3}{h_1^2}(d+1)(p+3)^{3/2}\bigg).$$

Putting together all these bounds, we get the claim. □

Notice that in Algorithms 1 and 2, the search directions are constructed using unbiased estimators of the gradient and Hessian of the smoothed surrogates of the outer and inner objectives f and g. In the following appendices, we study such estimators and bound their approximation errors. In particular, in Appendix D.3 we study the approximation of the gradients and Hessians used in Algorithm 1, while in Appendix D.4 we study the surrogate constructed in Algorithm 2.

D.3 Auxiliary results for Algorithm 1

In this appendix, we analyze the approximation error of the finite-difference estimators and the search directions constructed in Algorithm 1. To simplify reading, for this section, we introduce some notation that we will use in the proofs.

Notation. Let $b = \max(b_1, b_2)$ and $\ell = \max(\ell_1, \ell_2)$. For every iteration $k \in \mathbb{N}$, let $(\xi_{i,k})_{i=1}^b$ and $(\zeta i, k)_{i=1}^b$ be independent realizations of the random variables ξ and ζ sampled at iteration k. For every $i = 1, \dots, b$, let $(w_k^{(i,j)})j = 1^\ell$ and $(u_k^{(i,j)})j = 1^\ell$ be random vectors sampled i.i.d from $\mathcal{N}(0, I_p)$ and $\mathcal{N}(0, I_d)$, respectively, at iteration $k \in \mathbb{N}$. We denote by $\hat{\nabla}_z g_{h_k}(z_k, x_k, \xi i, k)$, $\hat{\nabla}_z f_{h_k}(z_k, x_k, \zeta_{i,k})$, and $\hat{\nabla}_x f_{h_k}(z_k, x_k, \zeta_{i,k})$ the finite-difference surrogates of the stochastic gradients of the inner and outer objectives g and f with respect to the variables z and x, computed at iteration $k \in \mathbb{N}$ using z_k, x_k, and $\xi_{i,k}$ (or $\zeta_{i,k}$), as follows.

$$\begin{aligned}
\hat{\nabla}_z g_{h_k}(z_k, x_k, \xi_{i,k}) &= \frac{1}{\ell_1}\sum_{j=1}^{\ell_1} \frac{g(z_k + h_k w_k^{(i,j)}, x_k, \xi_{i,k}) - g(z_k - h_k w_k^{(i,j)}, x_k, \xi_{i,k})}{2h_k} w_k^{(i,j)} \\
\hat{\nabla}_z f_{h_k}(z_k, x_k, \zeta_{i,k}) &= \frac{1}{\ell_2}\sum_{j=1}^{\ell_2} \frac{f(z_k + h_k w_k^{(i,j)}, x_k, \zeta_{i,k}) - f(z_k, x_k, \zeta_{i,k})}{h_k} w_k^{(i,j)} \\
\hat{\nabla}_x f_{h_k}(z_k, x_k, \zeta_{i,k}) &= \frac{1}{\ell_2}\sum_{j=1}^{\ell_2} \frac{f(z_k, x_k + h_k u_k^{(i,j)}, \zeta_{i,k}) - f(z_k, x_k, \zeta_{i,k})}{h_k} u_k^{(i,j)}.
\end{aligned} \tag{19}$$

Let

$$\begin{aligned}
c_k^{(i,j)}(z_k, x_k, \xi_{i,k}) &= \frac{g(z_k + h_k w_k^{(i,j)}, x_k, \xi_{i,k}) + g(z_k - h_k w_k^{(i,j)}, x_k, \xi_{i,k}) - 2g(z_k, x_k, \xi_{i,k})}{2h_k^2}, \\
s_k^{(i,j)}(z_k, x_k, \xi_{i,k}) &= \frac{g(z_k + h_k w_k^{(i,j)}, x_k + h_k u_k^{(i,j)}, \xi_{i,k}) + g(z_k - h_k w_k^{(i,j)}, x_k - h_k u_k^{(i,j)}, \xi_{i,k}) - 2g(z_k, x_k, \xi_{i,k})}{2h_k^2}
\end{aligned}$$

We denote the finite-difference surrogates of Hessian block zz and xz of the inner target as follows.

$$\begin{aligned}
\hat{\nabla}^2_{zz} g_{h_k}(z_k, x_k, \xi_{i,k}) &= \frac{1}{\ell_1}\sum_{j=1}^{\ell_1} c_k^{(i,j)}(z_k, x_k, \xi_{i,k})(w_k^{(i,j)} w_k^{(i,j)\top} - I_p). \\
\hat{\nabla}^2_{xz} g_{h_k}(z_k, x_k, \xi_{i,k}) &= \frac{1}{\ell_1}\sum_{j=1}^{\ell_1} s_k^{(i,j)}(z_k, x_k, \xi_{i,k}) u_k^{(i,j)} w_k^{(i,j)\top}.
\end{aligned} \tag{20}$$

Moreover, we denote by $\nabla_z g_{h_k}$, $\nabla_z f_{h_k}$, and $\nabla_x f_{h_k}$ the gradients with respect to the variables z and x of the smooth surrogates (eq. 12) of g and f at iteration $k \in \mathbb{N}$. These are defined as indicated in Lemma D.5 e.g.,

$$\begin{aligned}\nabla_z g_{h_k}(z_k, x_k, \xi_{i,k}) &= \frac{1}{\ell_1}\sum_{j=1}^{\ell_1} \mathbb{E}_{w_k^{(i,j)}}\left[\frac{g(z_k + h_k w_k^{(i,j)}, x_k, \xi i, k) - g(z_k - h_k w_k^{(i,j)}, x_k, \xi_{i,k})}{2h_k} w_k^{(i,j)}\right] \\ &= \mathbb{E}_w\left[\frac{g(z_k + h_k w, x_k, \xi i, k) - g(z_k - h_k w, x_k, \xi_{i,k})}{2h_k} w\right],\end{aligned}$$

where the second equality follows from the fact that all $w_k^{(i,j)}$ are i.i.d. Analogously, $\nabla_z f_{h_k}$ and $\nabla_x f_{h_k}$ are defined.

We begin by providing a lemma that bounds the expected norm of the surrogate gradient approximations.

Lemma D.9 (Approximation Error Bounds)**.** *Let Assumptions 3.1,3.2,3.3 hold. Let $\hat{\nabla}_z g_{h_k}$,$\hat{\nabla}_z f_{h_k}$ and $\hat{\nabla}_x f_{h_k}$ be the stochastic gradient surrogates defined in eq. 19. Let $(z_k)_{k\in\mathbb{N}}$ and $(x_k)_{k\in\mathbb{N}}$ be the sequences generated by Algorithm 1. Then, for every $k \in \mathbb{N}$*

$$\begin{aligned}\mathbb{E}_k\left[\left\|\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_z g_{h_k}(z_k, x_k, \xi_{i,k})\right\|^2\right] &\leq 2\left(2 + \frac{p+2}{b_1\ell_1}\right)\|\nabla_z G(z_k, x_k)\|^2 \\ &\quad + \left(\frac{(p+6)^3}{2b_1\ell_1} + \left(\frac{3}{b_1}+1\right)(p+3)^3\right)L_{1,G}^2 h_k^2 \\ &\quad + 2\left(3 + \frac{p+2}{\ell_1}\right)\frac{\sigma_{1,G}^2}{b_1},\end{aligned} \tag{21}$$

$$\begin{aligned}\mathbb{E}_k\left[\left\|\frac{1}{b_2}\sum_{i=1}^{b_2}\hat{\nabla}_z f_{h_k}(z_k, x_k, \zeta_{i,k})\right\|^2\right] &\leq 2\left(1 + \frac{p+2}{b_2\ell_2}\right)\|\nabla_z F(z_k, x_k)\|^2 \\ &\quad + \left(\frac{(p+6)^3}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(p+3)^3\right)L_{1,F}^2 h_k^2 \\ &\quad + 2\left(\frac{p+2}{\ell_2} + 3\right)\frac{\sigma_{1,F}^2}{b_2},\end{aligned}$$

and,

$$\begin{aligned}\mathbb{E}_k\left[\left\|\frac{1}{b_2}\sum_{i=1}^{b_2}\hat{\nabla}_x f_{h_k}(z_k, x_k, \zeta_{i,k})\right\|^2\right] &\leq 2\left(1 + \frac{d+2}{b_2\ell_2}\right)\|\nabla_x F(z_k, x_k)\|^2 \\ &\quad + \left(\frac{(d+6)^3}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(d+3)^3\right)L_{1,F}^2 h_k^2 \\ &\quad + 2\left(\frac{d+2}{b_2\ell_2} + \frac{3}{b_2}\right)\sigma_{1,F}^2.\end{aligned}$$

Proof. We start proving eq. 21. Define

$$\begin{aligned}g_k^{(i,j)} &= \frac{g(z_k + h_k w_k^{(i,j)}, x_k, \xi_{i,k}) - g(z_k - h_k w_k^{(i,j)}, x, \xi_{i,k})}{2h_k} w_k^{(i,j)}, \\ \delta_k^{(i,j)} &= g_k^{(i,j)} - \nabla_z g_{h_k}(z_k, x_k, \xi_{i,k}), \quad \text{and} \quad \delta_k^{(i)} = \sum_{j=1}^{\ell_1}\delta_k^{(i,j)}.\end{aligned}$$

Adding and subtracting $\frac{1}{b_1\ell_1}\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})$, we have

$$\begin{aligned}\left\|\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2 &= \frac{1}{b_1^2\ell_1^2}\left\|\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\delta_k^{(i,j)}+\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2\\ &= \frac{1}{b_1^2\ell_1^2}\left\|\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\delta_k^{(i,j)}\right\|^2+\frac{1}{b_1^2}\left\|\sum_{i=1}^{b_1}\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2\\ &\quad+\frac{2}{b_1^2\ell_1^2}\left\langle\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\delta_k^{(i,j)},\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\rangle.\end{aligned}$$

Taking the expectation in $w_k^{(i,j)}$ for every $i=1,\cdots,b_1$ and $j=1,\cdots,\ell_1$, we get

$$\begin{aligned}\mathbb{E}_{W_k}\left[\left\|\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2\right] &= \frac{1}{b_1^2\ell_1^2}\mathbb{E}_{W_k}\left[\left\|\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\delta_k^{(i,j)}\right\|^2\right]+\frac{1}{b_1^2}\left\|\sum_{i=1}^{b_1}\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2\\ &\quad+\frac{2}{b_1^2\ell_1^2}\left\langle\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\mathbb{E}_{w_k^{(i,j)}}[\delta_k^{(i,j)}],\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\rangle.\end{aligned}$$

By Lemma D.5, for every i and j, we have $\mathbb{E}_{w_k^{(i,j)}}\left[g_k^{(i,j)}\right]=\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})$. Thus, observing that

$$\mathbb{E}_{w_k^{(i,j)}}\left[\delta_k^{(i,j)}\right]=\mathbb{E}_{w_k^{(i,j)}}\left[g_k^{(i,j)}-\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right]=\mathbb{E}_{w_k^{(i,j)}}\left[g_k^{(i,j)}\right]-\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})=0,$$

we have,

$$\mathbb{E}_{W_k}\left[\left\|\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2\right]=\underbrace{\frac{1}{b_1^2\ell_1^2}\mathbb{E}_{W_k}\left[\left\|\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\delta_k^{(i,j)}\right\|^2\right]}_{A}+\underbrace{\frac{1}{b_1^2}\left\|\sum_{i=1}^{b_1}\nabla_z g_{h_k}(z,x,\xi_i)\right\|^2}_{B}. \tag{22}$$

The term B can be bounded by adding and subtracting the full-gradient of the smoothing $\nabla_z G_{h_k}(z_k,x_k)$, we get

$$\frac{1}{b_1^2}\left\|\sum_{i=1}^{b}\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2\le\frac{2}{b_1^2}\left\|\sum_{i=1}^{b_1}\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})-\nabla_z G_{h_k}(z_k,x_k)\right\|^2+2\left\|\nabla_z G_{h_k}(z_k,x_k)\right\|^2.$$

Let $\bar{\delta}_{i,k}:=g_{h_k}(z_k,x_k,\xi_{i,k})-\nabla_z G_{h_k}(z_k,x_k)$. Then, taking the conditional expectation on $\xi_{i,k}$ for every $i=1,\cdots,b_1$, we have

$$\begin{aligned}\frac{1}{b_1^2}\mathbb{E}_{\xi_k}\left[\left\|\sum_{i=1}^{b_1}\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2\right] &\le \frac{2}{b_1^2}\mathbb{E}_{\xi_k}\left[\left\|\sum_{i=1}^{b_1}\bar{\delta}_{i,k}\right\|^2\right]+2\left\|\nabla_z G_{h_k}(z_k,x_k)\right\|^2\\ &= \frac{2}{b_1^2}\left(\sum_{i=1}^{b_1}\mathbb{E}_{\xi_{i,k}}\left[\left\|\bar{\delta}_{i,k}\right\|^2\right]+\sum_{i=1}^{b_1}\sum_{t\neq i}\mathbb{E}_{\xi_{i,k}}\left[\left\langle\bar{\delta}_{i,k},\bar{\delta}_{t,k}\right\rangle\right]\right)\\ &\quad+2\left\|\nabla_z G_{h_k}(z_k,x_k)\right\|^2.\end{aligned}$$

By Assumption 3.1 and, since for every $i\neq t$, $\xi_{i,k}$ and $\xi_{t,k}$ are independent, we have

$$\begin{aligned}\frac{1}{b_1^2}\mathbb{E}_{\xi_k}\left[\left\|\sum_{i=1}^{b_1}\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2\right] &\le \frac{2}{b_1^2}\left(\sum_{i=1}^{b_1}\mathbb{E}_{\xi_{i,k}}\left[\left\|\bar{\delta}_{i,k}\right\|^2\right]+\sum_{i=1}^{b_1}\sum_{t\neq i}\underbrace{\left\langle\mathbb{E}_{\xi_{i,k}}\left[\bar{\delta}_{i,k}\right],\mathbb{E}_{\xi_{t,k}}\left[\bar{\delta}_{t,k}\right]\right\rangle}_{=0}\right)\\ &\quad+2\left\|\nabla_z G_{h_k}(z_k,x_k)\right\|^2.\end{aligned}$$

Adding and subtracting $\nabla_z g(z_k, x_k, \xi_{i,k})$ and $\nabla_z G(z_k, x_k)$, we get

$$\begin{aligned}\frac{1}{b_1^2}\mathbb{E}_{\xi_k}\left[\left\|\sum_{i=1}^{b_1}\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2\right] &\le \frac{2}{b_1^2}\sum_{i=1}^{b_1}\Big(3\mathbb{E}_{\xi_{i,k}}\left[\|\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})-\nabla_z g(z_k,x_k,\xi_{i,k})\|^2\right]\\ &\quad+3\|\nabla_z G(z_k,x_k)-\nabla_z G_{h_k}(z_k,x_k)\|^2\\ &\quad+3\mathbb{E}_{\xi_{i,k}}\left[\|\nabla_z g(z_k,x_k,\xi_{i,k})-\nabla_z G(z_k,x_k)\|^2\right]\Big)\\ &\quad+2\left\|\nabla_z G_{h_k}(z_k,x_k)\right\|^2.\end{aligned}$$

By Lemma D.7 and Assumption 3.1,

$$\frac{1}{b_1^2}\mathbb{E}_{\xi_k}\left[\left\|\sum_{i=1}^{b_1}\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2\right] \le \frac{6}{b_1}\left(\frac{L_{1,G}^2}{2}(p+3)^3h_k^2+\sigma_{1,G}^2\right)+2\left\|\nabla_z G_{h_k}(z_k,x_k)\right\|^2.$$

Adding and subtracting $\nabla_z G(z_k, x_k)$,

$$\begin{aligned}\frac{1}{b_1^2}\mathbb{E}_{\xi_k}\left[\left\|\sum_{i=1}^{b_1}\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2\right] &\le \frac{6}{b_1}\left(\frac{L_{1,G}^2}{2}(p+3)^3h_k^2+\sigma_{1,G}^2\right)\\ &\quad+2\left\|\nabla_z G_{h_k}(z_k,x_k)-\nabla_z G(z_k,x_k)+\nabla_z G(z_k,x_k)\right\|^2\\ &\le \frac{6}{b_1}\left(\frac{L_{1,G}^2}{2}(p+3)^3h_k^2+\sigma_{1,G}^2\right)\\ &\quad+4\left\|\nabla_z G_{h_k}(z_k,x_k)-\nabla_z G(z_k,x_k)\right\|^2\\ &\quad+4\left\|\nabla_z G(z_k,x_k)\right\|^2.\end{aligned}$$

By Lemma D.7, we get

$$\begin{aligned}\frac{1}{b_1^2}\mathbb{E}_{\xi_k}\left[\left\|\sum_{i=1}^{b_1}\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2\right] &\le 4\left\|\nabla_z G(z_k,x_k)\right\|^2+\left(\frac{3}{b_1}+1\right)L_{1,G}^2(p+3)^3h_k^2\\ &\quad+\frac{6\sigma_{1,G}^2}{b_1}.\end{aligned}\tag{23}$$

Now, we focus on bounding the term A. We have that

$$\frac{1}{b_1^2\ell_1^2}\mathbb{E}_{W_k}\left[\left\|\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\delta_k^{(i,j)}\right\|^2\right] = \frac{1}{b_1^2\ell_1^2}\left(\sum_{i=1}^{b_1}\mathbb{E}_{W_k^{(i)}}\left[\|\delta_k^{(i)}\|^2\right]+\sum_{i=1}^{b_1}\sum_{t\neq i}\mathbb{E}_{W_k^{(i)}}\left[\left\langle\delta_k^{(i)},\delta_k^{(t)}\right\rangle\right]\right).$$

Notice that for $i \neq t$, we have that for every $j = 1, \cdots, \ell_1$, $w^{(i,j)}, w^{(t,j)}$ are all independent. Therefore, we have

$$\frac{1}{b_1^2\ell_1^2}\mathbb{E}_{W_k}\left[\left\|\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\delta_k^{(i,j)}\right\|^2\right] = \frac{1}{b_1^2\ell_1^2}\left(\sum_{i=1}^{b_1}\mathbb{E}_{W_k^{(i)}}\left[\|\delta_k^{(i)}\|^2\right]+\sum_{i=1}^{b_1}\sum_{t\neq i}\underbrace{\left\langle\mathbb{E}_{W_k^{(i)}}[\delta_k^{(i)}],\mathbb{E}_{W_k^{(t)}}[\delta_k^{(t)}]\right\rangle}_{=0}\right).$$

Similarly, we get

$$\frac{1}{b_1^2\ell_1^2}\mathbb{E}_{W_k}\left[\left\|\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\delta_k^{(i,j)}\right\|^2\right] = \frac{1}{b_1^2\ell_1^2}\sum_{i=1}^{b_1}\left(\sum_{j=1}^{\ell_1}\mathbb{E}_{w_k^{(i,j)}}\left[\|\delta_k^{(i,j)}\|^2\right]+\sum_{j=1}^{\ell_1}\sum_{t\neq j}\mathbb{E}_{w_k^{(i,j)}}\left[\left\langle\delta_k^{(i,j)},\delta_k^{(i,t)}\right\rangle\right]\right).$$

Again, observing that for $j \neq t$, we have that $w^{(i,j)}, w^{(i,t)}$ are independent. Thus, we have

$$\begin{aligned}\frac{1}{b_1^2\ell_1^2}\mathbb{E}_{W_k}\left[\left\|\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\delta_k^{(i,j)}\right\|^2\right] &= \frac{1}{b_1^2\ell_1^2}\sum_{i=1}^{b_1}\left(\sum_{j=1}^{\ell_1}\mathbb{E}_{w_k^{(i,j)}}\left[\|\delta_k^{(i,j)}\|^2\right]\right. \\ &\quad \left.+\sum_{j=1}^{\ell_1}\sum_{t\neq j}\underbrace{\left\langle\mathbb{E}_{w_k^{(i,j)}}\left[\delta_k^{(i,j)}\right],\mathbb{E}_{w_k^{(i,t)}}\left[\delta_k^{(i,t)}\right]\right\rangle}_{=0}\right) \\ &= \frac{1}{b_1^2\ell_1^2}\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\mathbb{E}_{w_k^{(i,j)}}\left[\|\delta_k^{(i,j)}\|^2\right].\end{aligned}$$

Developing the square, we get

$$\begin{aligned}\frac{1}{b_1^2\ell_1^2}\mathbb{E}_{W_k}\left[\left\|\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\delta_k^{(i,j)}\right\|^2\right] &= \frac{1}{b_1^2\ell_1^2}\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\mathbb{E}_{w_k^{(i,j)}}\left[\|\delta_k^{(i,j)}\|^2\right] \\ &= \frac{1}{b_1^2\ell_1^2}\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\left(\mathbb{E}_{w_k^{(i,j)}}\left[\left\|g_k^{(i,j)}\right\|^2\right]+\|\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\|^2\right. \\ &\quad \left.-2\mathbb{E}_{w_k^{(i,j)}}\left[\left\langle g_k^{(i,j)},\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\rangle\right]\right) \\ &= \frac{1}{b_1^2\ell_1^2}\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\left(\mathbb{E}_{w_k^{(i,j)}}\left[\left\|g_k^{(i,j)}\right\|^2\right]-\|\nabla_z g_{h_k}(z_k,x_k,\xi_{i,k})\|^2\right) \\ &\leq \frac{1}{b_1^2\ell_1^2}\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\mathbb{E}_{w_k^{(i,j)}}\left[\left\|g_k^{(i,j)}\right\|^2\right].\end{aligned} \tag{24}$$

Therefore, taking the conditional expectation on $\xi_{i,k}$ for every i in eq. 22 and using inequalities eq. 24 and eq. 23, we have

$$\begin{aligned}\mathbb{E}_{\xi_k}\left[\mathbb{E}_{W_k}\left[\left\|\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_z g_{h_k}(z_k,x_k,\xi_{i,k})\right\|^2\right]\right] &\leq \frac{1}{b_1^2\ell_1^2}\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\underbrace{\mathbb{E}_{\xi_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left\|g_k^{(i,j)}\right\|^2\right]\right]}_{C} \\ &\quad +4\left\|\nabla_z G(z_k,x_k)\right\|^2 \\ &\quad +\left(\frac{3}{b_1}+1\right)L_{1,G}^2(p+3)^3h_k^2+\frac{6\sigma_{1,G}^2}{b_1}.\end{aligned} \tag{25}$$

We focus on bounding C. Adding and subtracting $g(z_k,x_k,\xi_{i,k})$, we get

$$\begin{aligned}\mathbb{E}_{\xi_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left\|g_k^{(i,j)}\right\|^2\right]\right] &\leq \frac{1}{2h_k^2}\mathbb{E}_{\xi_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left(g(z_k+h_kw_k^{(i,j)},x_k,\xi_{i,k})-g(z_k,x_k,\xi_{i,k})\right)^2\|w_k^{(i,j)}\|^2\right]\right] \\ &\quad +\frac{1}{2h_k^2}\mathbb{E}_{\xi_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left(g(z_k,x_k,\xi_{i,k})-g(z_k-h_kw_k^{(i,j)},x_k,\xi_{i,k})\right)^2\|w_k^{(i,j)}\|^2\right]\right].\end{aligned}$$

By symmetry, we have

$$\mathbb{E}_{\xi_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left\|g_k^{(i,j)}\right\|^2\right]\right] \leq \frac{1}{h_k^2}\mathbb{E}_{\xi_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left(g(z_k+h_kw_k^{(i,j)},x_k,\xi_{i,k})-g(z_k,x_k,\xi_{i,k})\right)^2\|w_k^{(i,j)}\|^2\right]\right].$$

Adding and subtracting $\left\langle \nabla g(z_k, x_k, \xi_{i,k}), [h_k w_k^{(i,j)}, 0] \right\rangle$ where $[h_k w_k^{(i,j)}, 0] \in \mathbb{R}^{p+d}$ denotes the concatenation between $h_k w_k^{(i,j)}$ and the vector of zeros of d entries.

$$\begin{aligned}\mathbb{E}_{\xi_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left\|g_k^{(i,j)}\right\|^2\right]\right] \leq \frac{2}{h_k^2}\mathbb{E}_{\xi_{i,k}}\Bigg[\mathbb{E}_{w_k^{(i,j)}}\Bigg[\Bigg(g(z_k + h_k w_k^{(i,j)}, x_k, \xi_{i,k}) - g(z_k, x_k, \xi_{i,k}) \\ - \left\langle \nabla g(z_k, x_k, \xi_{i,k}), [h_k w_k^{(i,j)}, 0] \right\rangle\Bigg)^2 \|w_k^{(i,j)}\|^2\Bigg]\Bigg] \\ + \frac{2}{h_k^2}\mathbb{E}_{\xi_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left(\left\langle \nabla g(z_k, x_k, \xi_{i,k}), [h_k w_k^{(i,j)}, 0] \right\rangle\right)^2 \|w_k^{(i,j)}\|^2\right]\right].\end{aligned}$$

Notice that $\left\langle \nabla g(z_k, x_k, \xi_{i,k}), [h_k w_k^{(i,j)}, 0] \right\rangle = \left\langle \nabla_z g(z_k, x_k, \xi_{i,k}), h_k w_k^{(i,j)} \right\rangle$. By the Descent Lemma (Polyak, 1987),

$$\begin{aligned}\mathbb{E}_{\xi_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left\|g_k^{(i,j)}\right\|^2\right]\right] \leq \frac{L_{1,G}^2}{2} h_k^2 \mathbb{E}_{w_k^{(i,j)}}\left[\|w_k^{(i,j)}\|^6\right] \\ + 2\mathbb{E}_{\xi_{i,k}}\left[\nabla_z g(z_k, x_k, \xi_{i,k})^\top \mathbb{E}_{w_k^{(i,j)}}\left[w_k^{(i,j)} w_k^{(i,j)\top} w_k^{(i,j)} w_k^{(i,j)\top}\right] \nabla_z g(z_k, x_k, \xi_{i,k})\right].\end{aligned}$$

By Lemma D.1, we have

$$\mathbb{E}_{\xi_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left\|g_k^{(i,j)}\right\|^2\right]\right] \leq \frac{L_{1,G}^2}{2}(p+6)^3 h_k^2 + 2(p+2)\mathbb{E}_{\xi_{i,k}}\left[\|\nabla_z g(z_k, x_k, \xi_{i,k})\|^2\right].$$

Adding and subtracting $\nabla_z G(z_k, x_k)$ and by Assumption 3.1, we get

$$\begin{aligned}\mathbb{E}_{\xi_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left\|g_k^{(i,j)}\right\|^2\right]\right] &\leq \frac{L_{1,G}^2}{2}(p+6)^3 h_k^2 \\ &\quad + 2(p+2)\mathbb{E}_{\xi_{i,k}}\left[\|\nabla_z g(z_k, x_k, \xi_{i,k}) - \nabla_z G(z_k, x_k) + \nabla_z G(z_k, x_k)\|^2\right] \\ &= \frac{L_{1,G}^2}{2}(p+6)^3 h_k^2 + 2(p+2)\mathbb{E}_{\xi_{i,k}}\left[\|\nabla_z g(z_k, x_k, \xi_{i,k}) - \nabla_z G(z_k, x_k)\|^2\right] \\ &\quad + 2(p+2)\|\nabla_z G(z_k, x_k)\|^2 \\ &\quad + 4(p+2)\underbrace{\left\langle \mathbb{E}_{\xi_{i,k}}\left[\nabla_z g(z_k, x_k, \xi_{i,k}) - \nabla_z G(z_k, x_k)\right], \nabla_z G(z_k, x_k)\right\rangle}_{=0} \\ &\leq \frac{L_{1,G}^2}{2}(p+6)^3 h_k^2 + 2(p+2)\sigma_{1,G}^2 + 2(p+2)\|\nabla_z G(z_k, x_k)\|^2.\end{aligned}$$

Using this inequality in eq. 25, we get the first claim

$$\begin{aligned}\mathbb{E}_{\xi_k}\left[\mathbb{E}_{W_k}\left[\left\|\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_z g_{h_k}(z_k, x_k, \xi_{i,k})\right\|^2\right]\right] &\leq 2\left(2 + \frac{p+2}{b_1 \ell_1}\right)\|\nabla_z G(z_k, x_k)\|^2 \\ &\quad + \left(\frac{(p+6)^3}{2 b_1 \ell_1} + \left(\frac{3}{b_1} + 1\right)(p+3)^3\right) L_{1,G}^2 h_k^2 \\ &\quad + 2\left(3 + \frac{p+2}{\ell_1}\right)\frac{\sigma_{1,G}^2}{b_1}.\end{aligned}$$

The proof of the second claim follows the same line. We report below the proof since the finite-difference surrogate used is different. Define

$$f_k^{(i,j)} = \frac{f(z_k + h_k w_k^{(i,j)}, x_k, \zeta_{i,k}) - f(z_k, x_k, \zeta_{i,k})}{h_k} w_k^{(i,j)},$$

$$\delta_{f,k}^{(i,j)} = f_k^{(i,j)} - \nabla_z f_{h_k}(z_k, x_k, \zeta_{i,k}), \quad \text{and} \quad \delta_{f,k}^{(i)} = \sum_{j=1}^{\ell_2} \delta_{f,k}^{(i,j)}.$$

Following the same steps of the previous claim, we get

$$\mathbb{E}_{\zeta_k}\left[\mathbb{E}_{W_k}\left[\left\|\frac{1}{b_2}\sum_{i=1}^{b_2}\hat{\nabla}_z f_{h_k}(z_k,x_k,\zeta_{i,k})\right\|^2\right]\right] \leq \frac{1}{b_2^2\ell_2^2}\sum_{i=1}^{b_2}\sum_{j=1}^{\ell_2}\underbrace{\mathbb{E}_{\zeta_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left\|f_k^{(i,j)}\right\|^2\right]\right]}_{D} + 4\left\|\nabla_z F(z_k,x_k)\right\|^2 + \left(\frac{3}{b_2}+1\right)L_{1,F}^2(p+3)^3h_k^2 + \frac{6\sigma_{1,F}^2}{b_2}. \tag{26}$$

Similarly to the previous C term, we bound the term D. Adding and subtracting $\left\langle \nabla f(z_k,x_k,\zeta_{i,k}),[h_k w_k^{(i,j)},0]\right\rangle$, we have

$$\begin{aligned}
\mathbb{E}_{\zeta_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left\|f_k^{(i,j)}\right\|^2\right]\right] &= \frac{1}{h_k^2}\mathbb{E}_{\zeta_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left(f(z_k+h_k w_k^{(i,j)},x_k,\zeta_{i,k})-f(z_k,x_k)\right)^2\|w_k^{(i,j)}\|^2\right]\right]\\
&\leq \frac{2}{h_k^2}\mathbb{E}_{\zeta_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left(f(z_k+h_k w_k^{(i,j)},x_k,\zeta_{i,k})-f(z_k,x_k)\right.\right.\right.\\
&\quad\left.\left.\left.-\left\langle\nabla f(z_k,x_k,\zeta_{i,k}),(h_k w_k^{(i,j)},0)\right\rangle\right)^2\|w_k^{(i,j)}\|^2\right]\right]\\
&\quad+\frac{2}{h_k^2}\mathbb{E}_{\zeta_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left(\left\langle\nabla f(z_k,x_k,\zeta_{i,k}),(h_k w_k^{(i,j)},0)\right\rangle\right)^2\|w_k^{(i,j)}\|^2\right]\right].
\end{aligned}$$

Notice that $\left\langle \nabla f(z_k,x_k,\zeta_{i,k}),[h_k w_k^{(i,j)},0]\right\rangle = \left\langle \nabla_z f(z_k,x_k,\xi_{i,k}),h_k w_k^{(i,j)}\right\rangle$. By the Descent Lemma (Polyak, 1987),

$$\begin{aligned}
\mathbb{E}_{\zeta_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left\|f_k^{(i,j)}\right\|^2\right]\right] &\leq \frac{L_{1,F}^2}{2}h_k^2\mathbb{E}_{w_k^{(i,j)}}\left[\|w_k^{(i,j)}\|^6\right]\\
&\quad+2\mathbb{E}_{\zeta_{i,k}}\left[\nabla_z f(z_k,x_k,\zeta_{i,k})^\top\mathbb{E}_{w_k^{(i,j)}}\left[w_k^{(i,j)}w_k^{(i,j)\top}w_k^{(i,j)}w_k^{(i,j)\top}\right]\nabla_z f(z_k,x_k,\zeta_{i,k})\right].
\end{aligned}$$

By Lemma D.1, we have

$$\mathbb{E}_{\zeta_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left\|f_k^{(i,j)}\right\|^2\right]\right] \leq \frac{L_{1,F}^2}{2}(p+6)^3h_k^2 + 2(p+2)\mathbb{E}_{\zeta_{i,k}}\left[\|\nabla_z f(z_k,x_k,\zeta_{i,k})\|^2\right].$$

Adding and subtracting $\nabla_z F(z_k,x_k)$ and by Assumption 3.1, we get

$$\begin{aligned}
\mathbb{E}_{\zeta_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left\|f_k^{(i,j)}\right\|^2\right]\right] &\leq \frac{L_{1,F}^2}{2}(p+6)^3h_k^2\\
&\quad+2(p+2)\mathbb{E}_{\zeta_{i,k}}\left[\|\nabla_z f(z_k,x_k,\xi_{i,k})-\nabla_z F(z_k,x_k)+\nabla_z F(z_k,x_k)\|^2\right]\\
&= \frac{L_{1,F}^2}{2}(p+6)^3h_k^2+2(p+2)\mathbb{E}_{\zeta_{i,k}}\left[\|\nabla_z f(z_k,x_k,\zeta_{i,k})-\nabla_z F(z_k,x_k)\|^2\right]\\
&\quad+2(p+2)\|\nabla_z F(z_k,x_k)\|^2\\
&\quad+4(p+2)\underbrace{\left\langle\mathbb{E}_{\xi_{i,k}}\left[\nabla_z f(z_k,x_k,\zeta_{i,k})-\nabla_z F(z_k,x_k)\right],\nabla_z F(z_k,x_k)\right\rangle}_{=0}\\
&\leq \frac{L_{1,F}^2}{2}(p+6)^3h_k^2+2(p+2)\sigma_{1,F}^2+2(p+2)\|\nabla_z F(z_k,x_k)\|^2.
\end{aligned}$$

Using this inequality in eq. 26, we get the second claim

$$\begin{aligned}
\mathbb{E}_{\zeta_k}\left[\mathbb{E}_{W_k}\left[\left\|\frac{1}{b_2}\sum_{i=1}^{b_2}\hat{\nabla}_z f_{h_k}(z_k,x_k,\zeta_{i,k})\right\|^2\right]\right] &\leq 2\left(1+\frac{p+2}{b_2\ell_2}\right)\|\nabla_z F(z_k,x_k)\|^2\\
&\quad+\left(\frac{(p+6)^3}{2b_2\ell_2}+\left(\frac{3}{b_2}+1\right)(p+3)^3\right)L_{1,F}^2h_k^2\\
&\quad+2\left(\frac{p+2}{b_2\ell_2}+\frac{3}{b_2}\right)\sigma_{1,F}^2
\end{aligned}$$

The proof of the last claim follows the same line. □

We now focus on deriving bounds on the expected norm of the product between a surrogate Hessian approximation and a vector. We begin by stating an auxiliary proposition that bounds the expected norm of the product of a surrogate Hessian approximation (constructed using a single direction and a single sample) with a vector. This result is a special case of Proposition 2.5 in (Aghasi & Ghadimi, 2025); we report it here because it will be used and subsequently extended to derive bounds for our estimators.

Proposition D.10 (Single direction Hessian estimator approximation error)**.** *Let Assumption 3.3 holds. Let w, u be sampled i.i.d from $\mathcal{N}(0, I_p)$ and $\mathcal{N}(0, I_d)$, respectively. For $h > 0$, define*

$$c_h(z, x, w, \xi) = \frac{g(z + hw, x, \xi) + g(z - hw, x, \xi) - 2g(z, x, \xi)}{2h^2}$$
$$s_h(z, x, w, \xi) = \frac{g(z + hw, x + hu, \xi) + g(z - hw, x - hu, \xi) - 2g(z, x, \xi)}{2h^2}.$$

Then, for every $z, v \in \mathbb{R}^p$, $x \in \mathbb{R}^d$, we have

$$\mathbb{E}_\xi \left[\mathbb{E}_w \left[\|c_h(z, x, w, \xi)(ww^\top - I_p)v)\|^2\right]\right] \leq \Bigg(4L_{2,G}^2(p+16)^4h^2 + \frac{15}{2}(p+6)^2\mathbb{E}_\xi\left[\|\nabla_{zz}^2 g(z, x, \xi)\|_F^2\right]\Bigg)\|v\|^2,$$

and,

$$\begin{aligned}\mathbb{E}_\xi \left[\mathbb{E}_w \left[\|s_h(z, x, w, \xi)uw^\top v)\|^2\right]\right] \leq \Bigg(&8L_{2,G}^2h^2\left((p+8)^4 + 2p(d+12)^3\right) \\ &+ \Bigg(6(p+4)(p+2)\mathbb{E}_\xi[\|\nabla_{zz}^2 g_h(z, x, \xi)\|_F^2] \\ &+ 36(p+2)\mathbb{E}_\xi[\|\nabla_{xz}^2 g_h(z, x, \xi)\|_F^2] \\ &+ 30p(d+2)\mathbb{E}_\xi[\|\nabla_{xx}^2 g_h(z, x, \xi)\|_F^2]\Bigg)\Bigg)\|v\|^2.\end{aligned}$$

Proof. The first inequality is a special case of Proposition 2.5(a) in (Aghasi & Ghadimi, 2025) with $\eta = h$ and $\mu = 0$, while the second inequality is a special case of Proposition 2.5(b) in (Aghasi & Ghadimi, 2025) with $\eta = \mu = h$. □

Now, with the following lemma, we extend this result for hessian block estimators used in Algorithm 1.

Lemma D.11 (Hessian Blocks approximation error)**.** *Let Assumption 3.3 holds. For every $k \in \mathbb{N}$, let $\hat{\nabla}_{zz}^2 g_{h_k}$ and $\hat{\nabla}_{xz}^2 g_{h_k}$ be the Hessian block surrogates defined in eq. 20. Let $(x_k)_{k\in\mathbb{N}}$, $(v_k)_{k\in\mathbb{N}}$ and $(z_k)_{k\in\mathbb{N}}$ be the sequences generated by Algorithm 1. Let $(\xi_{i,k})_{i=1}^{b_1}$ be realizations of independent copies of random variable ξ obtained at iteration k. Then,*

$$\mathbb{E}_{\xi_k}\left[\mathbb{E}_{W_k}\left[\left\|\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_{zz}^2 g_{h_k}(z_k, x_k, \xi_{i,k})v_k\right\|^2\right]\right] \leq \left(\frac{4L_{2,G}^2(p+16)^4}{b_1\ell_1}h_k^2 + \left(1 + \frac{15(p+6)^2p}{2b_1\ell_1}\right)L_{1,G}^2\right)\|v_k\|^2, \tag{27}$$

and,

$$\begin{aligned}\mathbb{E}_{\xi_k}\left[\mathbb{E}_{W_k,U_k}\left[\left\|\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_{xz}^2 g_{h_k}(z_k, x_k, \xi_{i,k})v_k\right\|^2\right]\right] \leq \Bigg(&\frac{1}{b_1\ell_1}\Bigg(8L_{2,G}^2h_k^2\left((p+8)^4 + 2p(d+12)^3\right) \\ &+ 6(p+4)(p+2)p + 36(p+2)\min(p,d) \\ &+ 30p(d+2)dL_{1,G}^2\Bigg) + L_{1,G}^2\Bigg)\|v_k\|^2.\end{aligned} \tag{28}$$

Proof. We start by proving eq. 27. For $i = 1, \cdots, b_1$ and $j = 1, \cdots, \ell_1$, we define

$$c_k^{(i,j)} = c_k^{(i,j)}(z_k, x_k, \xi_{i,k}) = \frac{g(z_k + h_k w_k^{(i,j)}, x_k, \xi_{i,k}) + g(z_k - h_k w_k^{(i,j)}, x_k, \xi_{i,k}) - 2g(z_k, x_k, \xi_{i,k})}{2h_k^2},$$

$$\delta_k^{(i,j)} = c_k^{(i,j)} \left(w_k^{(i,j)} w_k^{(i,j)\top} - I_p \right) v_k - \nabla_{zz}^2 g_{h_k}(z_k, x_k, \xi_{i,k}) v_k,$$

$$\Delta_k^{(i)} = \sum_{j=1}^{\ell_1} \delta_k^{(i,j)}.$$

Adding and subtracting $\nabla_{zz}^2 g_{h_k}(z_k, x_k, \xi_{i,k}) v_k$ and developing the square, we get

$$\left\| \frac{1}{b_1} \sum_{i=1}^{b_1} \hat{\nabla}_{zz}^2 g_{h_k}(z_k, x_k, \xi_{i,k}) v_k \right\|^2 = \frac{1}{b_1^2 \ell_1^2} \left\| \sum_{i=1}^{b_1} \Delta_k^{(i)} \right\|^2 + \frac{1}{b_1^2} \left\| \sum_{i=1}^{b_1} \nabla_{zz}^2 g_{h_k}(z_k, x_k, \xi_{i,k}) v_k \right\|^2 + \frac{2}{b_1^2 \ell_1^2} \left\langle \sum_{i=1}^{b_1} \Delta_k^{(i)}, \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \nabla_{zz}^2 g_{h_k}(z_k, x_k, \xi_{i,k}) v_k \right\rangle.$$

Taking the expectation on $w_k^{(i,j)}$ for every $i = 1, \cdots, b_1$ and $j = 1, \cdots, \ell_1$, and observing that by Lemma D.5, we have,

$$\mathbb{E}_{W_k} \left[\sum_{i=1}^{b_1} \Delta_k^{(i)} \right] = \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \mathbb{E}_{w_k^{(i,j)}} \left[c_k^{(i,j)} \left(w_k^{(i,j)} w_k^{(i,j)\top} - I_p \right) \right] v_k - \nabla_{zz}^2 g_{h_k}(z_k, x_k, \xi_{i,k}) v_k = 0,$$

we get

$$\begin{aligned} \mathbb{E}_{W_k} \left[\left\| \frac{1}{b_1} \sum_{i=1}^{b_1} \hat{\nabla}_{zz}^2 g_{h_k}(z_k, x_k, \xi_{i,k}) v_k \right\|^2 \right] &= \frac{1}{b_1^2 \ell_1^2} \mathbb{E}_{W_k} \left[\left\| \sum_{i=1}^{b_1} \Delta_k^{(i)} \right\|^2 \right] \\ &\quad + \frac{1}{b_1^2} \left\| \sum_{i=1}^{b_1} \nabla_{zz}^2 g_{h_k}(z_k, x_k, \xi_{i,k}) v_k \right\|^2 \\ &\quad + \frac{2}{b_1^2 \ell_1^2} \left\langle \underbrace{\mathbb{E}_{W_k} \left[\sum_{i=1}^{b_1} \Delta_k^{(i)} \right]}_{=0}, \sum_{i=1}^{b_1} \nabla_{zz}^2 g_{h_k}(z_k, x_k, \xi_{i,k}) v_k \right\rangle. \end{aligned} \tag{29}$$

By independence, we have

$$\begin{aligned} \mathbb{E}_{W_k} \left[\left\| \sum_{i=1}^{b} \Delta_k^{(i)} \right\|^2 \right] &= \sum_{i=1}^{b_1} \mathbb{E}_{W_k^{(i)}} \left[\|\Delta_k^{(i)}\|^2 \right] + \underbrace{\sum_{i=1}^{b_1} \sum_{t \neq i} \left\langle \mathbb{E}_{W_k^{(i)}} \left[\Delta_k^{(i)} \right], \mathbb{E}_{W_k^{(t)}} \left[\Delta_k^{(t)} \right] \right\rangle}_{=0} \\ &= \sum_{i=1}^{b_1} \left(\sum_{j=1}^{\ell_1} \mathbb{E}_{w_k^{(i,j)}} \left[\left\| \delta_k^{(i,j)} \right\|^2 \right] + \underbrace{\sum_{j=1}^{\ell_1} \sum_{t \neq j} \left\langle \mathbb{E}_{w_k^{(i,j)}} \left[\delta_k^{(i,j)} \right], \mathbb{E}_{w_k^{(i,t)}} \left[\delta_k^{(i,t)} \right] \right\rangle}_{=0} \right) \\ &= \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \mathbb{E}_{w_k^{(i,j)}} \left[\left\| \left(c_k^{(i,j)} \left(w_k^{(i,j)} w_k^{(i,j)\top} - I_p \right) - \nabla_{zz}^2 g_{h_k}(z_k, x_k, \xi_{i,k}) \right) v_k \right\|^2 \right] \\ &\leq \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \mathbb{E}_{w_k^{(i,j)}} \left[\left\| c_k^{(i,j)} \left(w_k^{(i,j)} w_k^{(i,j)\top} - I_p \right) v_k \right\|^2 \right]. \end{aligned}$$

Here, the last inequality holds by the fact that $\mathbb{E}_{w_k^{(i,j)}} \left[c_k^{(i,j)} \left(w_k^{(i,j)} w_k^{(i,j)\top} - I_p \right) v_k \right] = \nabla_{zz}^2 g_{h_k}(z_k, x_k, \xi_{i,k}) v_k$. Taking the expectation on $\xi_{i,k}$ for every $i = 1, \cdots, b_1$, by Proposition

D.10,

$$\begin{aligned}\mathbb{E}_{\xi_k}\left[\mathbb{E}_{W_k}\left[\left\|\sum_{i=1}^{b_1}\Delta_k^{(i)}\right\|^2\right]\right] &\leq \sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\mathbb{E}_{\xi_{i,k}}\left[\mathbb{E}_{w_k^{(i,j)}}\left[\left\|c_k^{(i,j)}\left(w_k^{(i,j)}w_k^{(i,j)\top}-I_p\right)v_k\right\|^2\right]\right]\\ &\leq \sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\Bigg(4L_{2,G}^2(p+16)^4h_k^2\\ &\quad+\frac{15}{2}(p+6)^2\mathbb{E}_{\xi_{i,k}}\left[\|\nabla_{zz}^2 g(z_k,x_k,\xi_{i,k})\|_F^2\right]\Bigg)\|v_k\|^2\\ &\leq \sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\Bigg(4L_{2,G}^2(p+16)^4h_k^2\\ &\quad+\frac{15}{2}(p+6)^2p\mathbb{E}_{\xi_{i,k}}\left[\|\nabla_{zz}^2 g(z_k,x_k,\xi_{i,k})\|^2\right]\Bigg)\|v_k\|^2.\end{aligned}$$

Since for every ξ, we have that the function $(z,x)\mapsto g(z,x,\xi)$ is $L_{1,G}$-smooth, we have that $\|\nabla_{zz}^2 g(z,x,\xi)\|\leq\|\nabla^2 g(z,x,\xi)\|\leq L_{1,G}$. Therefore, we get

$$\mathbb{E}_{\xi_k}\left[\mathbb{E}_{W_k}\left[\left\|\sum_{i=1}^{b_1}\Delta_k^{(i)}\right\|^2\right]\right]\leq\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\left(4L_{2,G}^2(p+16)^4h_k^2+\frac{15}{2}(p+6)^2pL_{1,G}^2\right)\|v_k\|^2.$$

Using this inequality in eq. 29, we get

$$\begin{aligned}\mathbb{E}_{\xi_k}\left[\mathbb{E}_{W_k}\left[\left\|\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_{zz}^2 g_{h_k}(z_k,x_k,\xi_{i,k})v_k\right\|^2\right]\right] &\leq \frac{1}{b_1\ell_1}\left(4L_{2,G}^2(p+16)^4h_k^2+\frac{15}{2}(p+6)^2pL_{1,G}^2\right)\|v_k\|^2\\ &\quad+\frac{1}{b_1^2}\mathbb{E}_{\xi_k}\left[\left\|\sum_{i=1}^{b_1}\nabla_{zz}^2 g_{h_k}(z_k,x_k,\xi_{i,k})v_k\right\|^2\right].\end{aligned}$$

By Proposition D.4, we have that g_{h_k} is $L_{1,G}$-smooth. Thus $\|\nabla_{zz}^2 g_{h_k}(z_k,x_k,\xi_{i,k})\|\leq L_{1,G}$ and, therefore, we have

$$\begin{aligned}\mathbb{E}_{\xi_k}\left[\mathbb{E}_{W_k}\left[\left\|\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_{zz}^2 g_{h_k}(z_k,x_k,\xi_{i,k})v_k\right\|^2\right]\right] &\leq \Bigg(\frac{4L_{2,G}^2(p+16)^4}{b_1\ell_1}h_k^2\\ &\quad+\left(1+\frac{15(p+6)^2p}{2b_1\ell_1}\right)L_{1,G}^2\Bigg)\|v_k\|^2.\end{aligned}$$

Now, we prove eq. 28. For $i=1,\cdots,b$ and $j=1,\cdots,\ell$ we define

$$\begin{aligned}&s_k^{(i,j)}=s_k^{(i,j)}(z_k,x_k,\xi_{i,k}),\qquad \bar{\delta}_k^{(i,j)}=s_k^{(i,j)}u_k^{(i,j)}w_k^{(i,j)\top}v_k-\nabla_{xz}^2 g_{h_k}(z_k,x_k,\xi_{i,k})v_k,\\ &\bar{\Delta}_k^{(i)}=\sum_{j=1}^{\ell_1}\bar{\delta}_k^{(i,j)}.\end{aligned}$$

Adding and subtracting $\nabla_{xz}^2 g_{h_k}(z_k,x_k,\xi_{i,k})v_k$ and developing the square, we get

$$\begin{aligned}\left\|\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_{xz}^2 g(z_k,x_k,\xi_{i,k})v_k\right\|^2 &= \frac{1}{b_1^2\ell_1^2}\left\|\sum_{i=1}^{b_1}\bar{\Delta}_k^{(i)}\right\|^2+\frac{1}{b_1^2}\left\|\sum_{i=1}^{b_1}\nabla_{xz}^2 g_{h_k}(z_k,x_k,\xi_{i,k})v_k\right\|^2\\ &\quad+\frac{2}{b_1^2\ell_1^2}\left\langle\sum_{i=1}^{b_1}\bar{\Delta}_k^{(i)},\sum_{i=1}^{b_1}\nabla_{xz}^2 g_{h_k}(z_k,x_k,\xi_{i,k})v_k\right\rangle.\end{aligned}$$

Taking the conditional expectation on $w_k^{(i,j)}, u_k^{(i,j)}$ for every $i = 1, \cdots, b_1$ and $j = 1, \cdots, \ell_1$, we have that since $w_k^{(i,j)}, u_k^{(i,j)}, \xi_{i,k}$ are independent,

$$\begin{aligned}\mathbb{E}_{W_k,U_k}\left[\left\|\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}^2_{xz}g(z_k,x_k,\xi_{i,k})v_k\right\|^2\right] &= \frac{1}{b_1^2\ell_1^2}\underbrace{\mathbb{E}_{W_k,U_k}\left[\left\|\sum_{i=1}^{b_1}\bar{\Delta}_k^{(i)}\right\|^2\right]}_{A} \\ &\quad + \frac{1}{b_1^2}\underbrace{\left\|\sum_{i=1}^{b_1}\nabla^2_{xz}g_{h_k}(z_k,x_k,\xi_{i,k})v_k\right\|^2}_{B} \\ &\quad + \frac{2}{b_1^2\ell_1^2}\left\langle\underbrace{\sum_{i=1}^{b_1}\mathbb{E}_{W_k,U_k}[\bar{\Delta}_k^{(i)}]}_{=0}, \sum_{i=1}^{b_1}\nabla^2_{xz}g_{h_k}(z_k,x_k,\xi_{i,k})v_k\right\rangle.\end{aligned} \tag{30}$$

We bound the terms A and B separately. We start from A,

$$\mathbb{E}_{W_k,U_k}\left[\left\|\sum_{i=1}^{b_1}\bar{\Delta}_k^{(i)}\right\|^2\right] = \sum_{i=1}^{b_1}\mathbb{E}_{W_k^{(i)},U_k^{(i)}}\left[\left\|\bar{\Delta}_k^{(i)}\right\|^2\right] + \sum_{i=1}^{b_1}\sum_{t\neq i}\mathbb{E}_{W_k,U_k}\left[\left\langle\bar{\Delta}_k^{(i)},\bar{\Delta}_k^{(t)}\right\rangle\right].$$

Observing that, for every $i \neq t$, $w_k^{(i,j)}, w_k^{(t,j)}, u_k^{(i,j)}, u_k^{(t,j)}$ are independent, and that by Lemma D.5, we have

$$\mathbb{E}_{W_k^{(i)},U_k^{(i)}}\left[\bar{\Delta}_k^{(i)}\right] = \sum_{j=1}^{\ell_1}\mathbb{E}_{w_k^{(i,j)},u_k^{(i,j)}}\left[s_k^{(i,j)}u_k^{(i,j)}w_k^{(i,j)\top}\right]v_k - \nabla^2_{xz}g_{h_k}(z_k,x_k,\xi_{i,k})v_k = 0,$$

we get

$$\mathbb{E}_{W_k,U_k}\left[\left\|\sum_{i=1}^{b_1}\bar{\Delta}_k^{(i)}\right\|^2\right] = \sum_{i=1}^{b_1}\mathbb{E}_{W_k^{(i)},U_k^{(i)}}\left[\left\|\bar{\Delta}_k^{(i)}\right\|^2\right] + \sum_{i=1}^{b_1}\sum_{t\neq i}\underbrace{\left\langle\mathbb{E}_{W_k^{(i)},U_k^{(i)}}\left[\bar{\Delta}_k^{(i)}\right],\mathbb{E}_{W_k^{(t)},U_k^{(t)}}\left[\bar{\Delta}_k^{(t)}\right]\right\rangle}_{=0}.$$

Similarly, we have

$$\begin{aligned}\mathbb{E}_{W_k,U_k}\left[\left\|\sum_{i=1}^{b_1}\bar{\Delta}_k^{(i)}\right\|^2\right] &= \sum_{i=1}^{b_1}\mathbb{E}_{W_k^{(i)},U_k^{(i)}}\left[\left\|\bar{\Delta}_k^{(i)}\right\|^2\right] \\ &= \sum_{i=1}^{b_1}\left(\sum_{j=1}^{\ell_1}\mathbb{E}_{w_k^{(i,j)},u_k^{(i,j)}}\left[\left\|\bar{\delta}_k^{(i,j)}\right\|^2\right] + \sum_{j=1}^{\ell_1}\sum_{t\neq j}\mathbb{E}_{W_k^{(i)},U_k^{(i)}}\left[\left\langle\bar{\delta}_k^{(i,j)},\bar{\delta}_k^{(i,t)}\right\rangle\right]\right).\end{aligned}$$

Again, since for every $j \neq t$, $w_k^{(i,j)}, u_k^{(i,j)}, w_k^{(i,t)}, u_k^{(i,t)}$ are independent and, by Lemma D.5, we have

$$\mathbb{E}_{w_k^{(i,j)},u_k^{(i,j)}}[\bar{\delta}_k^{(i,j)}] = \mathbb{E}_{w_k^{(i,j)},u_k^{(i,j)}}\left[s_k^{(i,j)}u_k^{(i,j)}w_k^{(i,j)\top}\right]v_k - \nabla^2_{xz}g_k(z_k,x_k,\xi_{i,k})v_k = 0$$

we get

$$\begin{aligned}\mathbb{E}_{W_k,U_k}\left[\left\|\sum_{i=1}^{b_1}\bar{\Delta}_k^{(i)}\right\|^2\right] &= \sum_{i=1}^{b_1}\left(\sum_{j=1}^{\ell_1}\mathbb{E}_{w_k^{(i,j)},u_k^{(i,j)}}\left[\left\|\bar{\delta}_k^{(i,j)}\right\|^2\right]\right. \\ &\quad \left. + \sum_{j=1}^{\ell_1}\sum_{t\neq j}\underbrace{\left\langle\mathbb{E}_{w_k^{(i,j)},u_k^{(i,j)}}\left[\bar{\delta}_k^{(i,j)}\right],\mathbb{E}_{w_k^{(i,t)},u_k^{(i,t)}}\left[\bar{\delta}_k^{(i,t)}\right]\right\rangle}_{=0}\right) \\ &= \sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\mathbb{E}_{w_k^{(i,j)},u_k^{(i,j)}}\left[\left\|s_k^{(i,j)}u_k^{(i,j)}w_k^{(i,j)\top}v_k - \nabla^2_{xz}g_{h_k}(z_k,x_k,\xi_{i,k})v_k\right\|^2\right] \\ &\leq \sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\mathbb{E}_{w_k^{(i,j)},u_k^{(i,j)}}\left[\left\|s_k^{(i,j)}u_k^{(i,j)}w_k^{(i,j)\top}v_k\right\|^2\right].\end{aligned}$$

Taking the expectation on ξ_k, by Proposition D.10, and since $\|A\|_F^2 \leq rank(A)\|A\|^2$ where $\|A\|_F^2$ is the squared Frobenius norm of a matrix A, we have

$$\begin{aligned}\mathbb{E}_{\xi_k}\left[\mathbb{E}_{W_k,U_k}\left[\left\|\sum_{i=1}^{b_1}\bar{\Delta}_k^{(i)}\right\|^2\right]\right] \leq \sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\Bigg(&8L_{2,G}^2h_k^2\left((p+8)^4+2p(d+12)^3\right)\\&+\Bigg(6(p+4)(p+2)p\mathbb{E}_{\xi_{i,k}}[\|\nabla_{zz}^2 g_{h_k}(z_k,x_k,\xi_{i,k})\|^2]\\&+36(p+2)\min(p,d)\mathbb{E}_{\xi_{i,k}}[\|\nabla_{xz}^2 g_{h_k}(z_k,x_k,\xi_{i,k})\|^2]\\&+30p(d+2)d\mathbb{E}_{\xi_{i,k}}[\|\nabla_{xx}^2 g_{h_k}(z_k,x_k,\xi_{i,k})\|^2]\Bigg)\Bigg)\|v_k\|^2.\end{aligned}$$

Since, by Proposition D.4, the function g_{h_k} is $L_{1,G}$-smooth, we have

$$\begin{aligned}\mathbb{E}_{\xi_k}\left[\mathbb{E}_{W_k,U_k}\left[\left\|\sum_{i=1}^{b_1}\bar{\Delta}_k^{(i)}\right\|^2\right]\right] \leq \sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\Bigg(&8L_{2,G}^2h_k^2\left((p+8)^4+2p(d+12)^3\right)\\&+\Bigg(6(p+4)(p+2)p\\&+36(p+2)\min(p,d)+30p(d+2)d\Bigg)L_{1,G}^2\Bigg)\|v_k\|^2.\end{aligned} \tag{31}$$

Now, we bound the term B of eq. 30. By Cauchy-Schwartz, and since g_{h_k} is $L_{1,G}$-smooth, we have

$$\begin{aligned}\left\|\sum_{i=1}^{b_1}\nabla_{xz}^2 g_{h_k}(z_k,x_k,\xi_{i,k})v_k\right\|^2 &\leq b_1\sum_{i=1}^{b_1}\|\nabla_{xz}^2 g_{h_k}(z_k,x_k,\xi_{i,k})v_k\|^2\\&\leq b_1\sum_{i=1}^{b_1}\|\nabla_{xz}^2 g_{h_k}(z_k,x_k,\xi_{i,k})\|^2\|v_k\|^2\\&\leq b_1^2L_{1,G}^2\|v_k\|^2.\end{aligned}$$

Using this inequality and eq. 31 in eq. 30, we get the claim. □

In the following lemma, we provide bound on the norm of the search directions used in Algorithm 1.

Lemma D.12 (Bound on direction norm)**.** *Let Assumptions 3.1,3.2,3.3 holds. Let D_z^k, D_v^k and D_x^k be the search directions computed at iteration $k \in \mathbb{N}$ with eq. 4, eq. 6 and eq. 8. Let $(z_k)_{k\in\mathbb{N}}, (v_k)_{k\in\mathbb{N}}$ and $(x_k)_{k\in\mathbb{N}}$ be the sequences generated by Algorithm 1. Then, for every $k \in \mathbb{N}$, we have*

$$\begin{aligned}\mathbb{E}_k\left[\|D_z^k\|^2\right] \leq{}& 2\left(2+\frac{p+2}{b_1\ell_1}\right)L_{1,G}^2\|z_k-z^*(x_k)\|^2\\&+\left(\frac{(p+6)^3}{2b_1\ell_1}+\left(\frac{3}{b_1}+1\right)(p+3)^3\right)L_{1,G}^2h_k^2+2\left(3+\frac{p+2}{\ell_1}\right)\frac{\sigma_{1,G}^2}{b_1}.\end{aligned} \tag{32}$$

Moreover, let $C_v^k = 4\left(\frac{4L_{2,G}^2(p+16)^4}{b_1\ell_1}h_k^2+\left(1+\frac{15(p+6)^2p}{2b_1\ell_1}\right)L_{1,G}^2\right)$. Then, for every $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}_k\left[\|D_v^k\|^2\right] \leq{}& C_v^k\|v_k-v^*(x_k)\|^2+C_v^k\frac{L_{0,F}^2}{\mu_G^2}\\&+4\left(1+\frac{p+2}{b_2\ell_2}\right)L_{0,F}^2+4\left(\frac{p+2}{\ell_2}+3\right)\frac{\sigma_{1,F}^2}{b_2}\\&+2\left(\frac{(p+6)^3}{2b_2\ell_2}+\left(\frac{3}{b_2}+1\right)(p+3)^3\right)L_{1,F}^2h_k^2,\end{aligned} \tag{33}$$

and,

$$\begin{aligned}\mathbb{E}_k\left[\left\|D_x^k\right\|^2\right] \le\ & 2C_x^k\|v_k - v^*(x_k)\|^2 + 2C_x^k\frac{L_{0,F}^2}{\mu_G^2} \\ & + 2\left(\frac{(d+6)^3}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(d+3)^3\right)L_{1,F}^2 h_k^2 \\ & + 4\left(1+\frac{d+2}{b_2\ell_2}\right)L_{0,F}^2 + 4\left(\frac{d+2}{b_2\ell_2}+\frac{3}{b_2}\right)\sigma_{1,F}^2,\end{aligned} \tag{34}$$

with $C_x^k = 2\left(\frac{1}{b_1\ell_1}\left(C_d^{(1)}h_k^2 + C_d^{(2)}\right) + L_{1,G}^2\right)$ *where*

$$\begin{aligned}C_d^{(1)} &= 8L_{2,G}^2\left((p+8)^4 + 2p(d+12)^3\right) \\ C_d^{(2)} &= 6(p+4)(p+2)p + 36(p+2)\min(p,d) + 30p(d+2)dL_{1,G}^2\end{aligned}$$

Proof. We start by proving eq. 32. By eq. 4 and Lemma D.9,

$$\begin{aligned}\mathbb{E}_k\left[\|D_z^k\|^2\right] \le\ & 2\left(2+\frac{p+2}{b_1\ell_1}\right)\|\nabla_z G(z_k,x_k)\|^2 \\ & + \left(\frac{(p+6)^3}{2b_1\ell_1} + \left(\frac{3}{b_1}+1\right)(p+3)^3\right)L_{1,G}^2 h_k^2 \\ & + 2\left(3+\frac{p+2}{\ell_1}\right)\frac{\sigma_{1,G}^2}{b_1}.\end{aligned}$$

Since $\nabla_z G(z^*(x_k),x_k) = 0$, by $L_{1,G}$-smoothness, we have

$$\begin{aligned}\mathbb{E}_k\left[\|D_z^k\|^2\right] \le\ & 2\left(2+\frac{p+2}{b_1\ell_1}\right)L_{1,G}^2\|z_k - z^*(x_k)\|^2 \\ & + \left(\frac{(p+6)^3}{2b_1\ell_1} + \left(\frac{3}{b_1}+1\right)(p+3)^3\right)L_{1,G}^2 h_k^2 \\ & + 2\left(3+\frac{p+2}{\ell_1}\right)\frac{\sigma_{1,G}^2}{b_1}.\end{aligned}$$

Now we prove eq. 33. By eq. 6, we have

$$\begin{aligned}\mathbb{E}_k\left[\|D_v^k\|^2\right] &= \mathbb{E}_k\left[\left\|\left(\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_{zz}^2 g_{h_k}(z_k,x_k,\xi_{i,k},W_{i,k})\right)v_k + \frac{1}{b_2}\sum_{i=1}^{b_2}\hat{\nabla}_z f_{h_k}(z_k,x_k,\zeta_{i,k},\bar{W}_{i,k})\right\|^2\right] \\ &\le 2\mathbb{E}_k\left[\left\|\left(\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_{zz}^2 g_{h_k}(z_k,x_k,\xi_{i,k},W_{i,k})\right)v_k\right\|^2\right] \\ &\quad + 2\mathbb{E}_k\left[\left\|\frac{1}{b_2}\sum_{i=1}^{b_2}\hat{\nabla}_z f_{h_k}(z_k,x_k,\zeta_{i,k},\bar{W}_{i,k})\right\|^2\right].\end{aligned}$$

By Lemma D.9 and Lemma D.11, we get

$$\begin{aligned}\mathbb{E}_k\left[\|D_v^k\|^2\right] \le\ & 2\left(\frac{4L_{2,G}^2(p+16)^4}{b_1\ell_1}h_k^2 + \left(1+\frac{15(p+6)^2p}{2b_1\ell_1}\right)L_{1,G}^2\right)\|v_k\|^2 \\ & + 4\left(1+\frac{p+2}{b_2\ell_2}\right)\|\nabla_z F(z_k,x_k)\|^2 \\ & + 2\left(\frac{(p+6)^3}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(p+3)^3\right)L_{1,F}^2 h_k^2 + 4\left(\frac{p+2}{\ell_2}+3\right)\frac{\sigma_{1,F}^2}{b_2}.\end{aligned}$$

Since F is $L_{0,F}$-Lipschitz continuous, we have

$$\begin{aligned}\mathbb{E}_k\left[\|D_v^k\|^2\right] \leq 2\Bigg(&\frac{4L_{2,G}^2(p+16)^4}{b_1\ell_1}h_k^2 + \left(1+\frac{15(p+6)^2p}{2b_1\ell_1}\right)L_{1,G}^2\Bigg)\|v_k\|^2\\ &+4\left(1+\frac{p+2}{b_2\ell_2}\right)L_{0,F}^2\\ &+2\left(\frac{(p+6)^3}{2b_2\ell_2}+\left(\frac{3}{b_2}+1\right)(p+3)^3\right)L_{1,F}^2h_k^2+4\left(\frac{p+2}{\ell_2}+3\right)\frac{\sigma_{1,F}^2}{b_2}.\end{aligned}$$

Adding and subtracting $v^*(x_k)$, we get

$$\begin{aligned}\mathbb{E}_k\left[\|D_v^k\|^2\right] \leq 4\Bigg(&\frac{4L_{2,G}^2(p+16)^4}{b_1\ell_1}h_k^2 + \left(1+\frac{15(p+6)^2p}{2b_1\ell_1}\right)L_{1,G}^2\Bigg)\|v_k-v^*(x_k)\|^2\\ &+4\Bigg(\frac{4L_{2,G}^2(p+16)^4}{b_1\ell_1}h_k^2 + \left(1+\frac{15(p+6)^2p}{2b_1\ell_1}\right)L_{1,G}^2\Bigg)\|v^*(x_k)\|^2\\ &+4\left(1+\frac{p+2}{b_2\ell_2}\right)L_{0,F}^2\\ &+2\left(\frac{(p+6)^3}{2b_2\ell_2}+\left(\frac{3}{b_2}+1\right)(p+3)^3\right)L_{1,F}^2h_k^2+4\left(\frac{p+2}{\ell_2}+3\right)\frac{\sigma_{1,F}^2}{b_2}.\end{aligned}$$

Let $C_v^k = 4\Bigg(\frac{4L_{2,G}^2(p+16)^4}{b_1\ell_1}h_k^2 + \left(1+\frac{15(p+6)^2p}{2b_1\ell_1}\right)L_{1,G}^2\Bigg)$. By Lemma D.3, we have

$$\begin{aligned}\mathbb{E}_k\left[\|D_v^k\|^2\right] \leq\ & C_v^k\|v_k-v^*(x_k)\|^2 + C_v^k\frac{L_{0,F}^2}{\mu_G^2}\\ &+4\left(1+\frac{p+2}{b_2\ell_2}\right)L_{0,F}^2+4\left(\frac{p+2}{\ell_2}+3\right)\frac{\sigma_{1,F}^2}{b_2}\\ &+2\left(\frac{(p+6)^3}{2b_2\ell_2}+\left(\frac{3}{b_2}+1\right)(p+3)^3\right)L_{1,F}^2h_k^2.\end{aligned}$$

Finally, we prove eq. 34. By eq. 8,

$$\begin{aligned}\mathbb{E}_k\left[\left\|D_x^k\right\|^2\right] &= \mathbb{E}_k\left[\left\|\left(\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_{xz}^2 g_k(z_k,x_k,\xi_{i,k})\right)v_k+\frac{1}{b_2}\sum_{i=1}^{b_2}\hat{\nabla}_x f_k(z_k,x_k,\zeta_{i,k})\right\|^2\right]\\ &\leq 2\mathbb{E}_k\left[\left\|\left(\frac{1}{b_1}\sum_{i=1}^{b_1}\hat{\nabla}_{xz}^2 g_k(z_k,x_k,\xi_{i,k})\right)v_k\right\|^2\right]\\ &\quad+2\mathbb{E}_k\left[\left\|\frac{1}{b_2}\sum_{i=1}^{b_2}\hat{\nabla}_x f_k(z_k,x_k,\zeta_{i,k})\right\|^2\right].\end{aligned}$$

By Lemma D.9 and Lemma D.11, we have

$$\begin{aligned}\mathbb{E}_k\left[\left\|D_x^k\right\|^2\right] \leq 2\Bigg(&\frac{1}{b_1\ell_1}\Bigg(8L_{2,G}^2h_k^2\left((p+8)^4+p(d+12)^3\right)\\ &+6(p+4)(p+2)p+36(p+2)\min(p,d)\\ &+30p(d+2)dL_{1,G}^2\Bigg)+L_{1,G}^2\Bigg)\|v_k\|^2\\ &+4\left(1+\frac{d+2}{b_2\ell_2}\right)\|\nabla_x F(z_k,x_k)\|^2\\ &+2\left(\frac{(d+6)^3}{2b_2\ell_2}+\left(\frac{3}{b_2}+1\right)(d+3)^3\right)L_{1,F}^2h_k^2\\ &+4\left(\frac{d+2}{b_2\ell_2}+\frac{3}{b_2}\right)\sigma_{1,F}^2.\end{aligned}$$

By $L_{0,F}$-Lipschitz continuity of F,

$$\begin{aligned}\mathbb{E}_k\left[\left\|D_x^k\right\|^2\right] \leq 2\Bigg(&\frac{1}{b_1\ell_1}\Bigg(8L_{2,G}^2h_k^2\left((p+8)^4+2p(d+12)^3\right)\\&+6(p+4)(p+2)p+36(p+2)\min(p,d)\\&+30p(d+2)dL_{1,G}^2\Bigg)+L_{1,G}^2\Bigg)\|v_k\|^2\\&+2\left(\frac{(d+6)^3}{2b_2\ell_2}+\left(\frac{3}{b_2}+1\right)(d+3)^3\right)L_{1,F}^2h_k^2\\&+4\left(1+\frac{d+2}{b_2\ell_2}\right)L_{0,F}^2+4\left(\frac{d+2}{b_2\ell_2}+\frac{3}{b_2}\right)\sigma_{1,F}^2.\end{aligned}$$

Let $C_x^k = 2\Bigg(\frac{1}{b_1\ell_1}\Bigg(8L_{2,G}^2h_k^2\left((p+8)^4+2p(d+12)^3\right)+6(p+4)(p+2)p+36(p+2)\min(p,d)+30p(d+2)dL_{1,G}^2\Bigg)+L_{1,G}^2\Bigg)$. Adding and subtracting $v^*(x_k)$, by Lemma D.3, we get

$$\begin{aligned}\mathbb{E}_k\left[\left\|D_x^k\right\|^2\right] \leq\ &2C_x^k\|v_k-v^*(x_k)\|^2\\&+2C_x^k\frac{L_{0,F}^2}{\mu_G^2}+2\left(\frac{(d+6)^3}{2b_2\ell_2}+\left(\frac{3}{b_2}+1\right)(d+3)^3\right)L_{1,F}^2h_k^2\\&+4\left(1+\frac{d+2}{b_2\ell_2}\right)L_{0,F}^2+4\left(\frac{d+2}{b_2\ell_2}+\frac{3}{b_2}\right)\sigma_{1,F}^2.\end{aligned}$$

□

Now, in the next lemma, we provides bounds on the error of the sequences z_k, v_k generated by Algorithm 1 relative to the corresponding optimal solutions $z^*(x_k), v^*(x_k)$ at each iterate x_k.

Lemma D.13 (Bound of sequences)**.** *Let Assumptions 3.2, 3.3 holds. Let $(z_k)_{k\in\mathbb{N}}, (v_k)_{k\in\mathbb{N}}$ be the sequences generated by Algorithm 1. Let $\omega_k^z = \left(1+\frac{4}{\mu_G\rho_k}\right)$, the for every $k\in\mathbb{N}$*

$$\begin{aligned}\mathbb{E}_k\left[\|z_{k+1}-z^*(x_{k+1})\|^2\right] \leq\ &\left(1-\frac{\mu_G\rho_k}{2}\right)\|z_k-z^*(x_k)\|^2+2\rho_k^2\mathbb{E}_k\left[\left\|D_z^k\right\|^2\right]\\&+2\omega_k^zL_*^2\gamma_k^2\mathbb{E}_k\left[\left\|D_x^k\right\|^2\right]+\frac{L_{1,G}^2}{4\mu_G}(d+3)^3\rho_kh_k^2.\end{aligned}\tag{35}$$

Moreover, let $\omega_{v,1} = \left(L_{1,F}+\frac{L_{0,F}L_{2,G}}{\mu_G}\right)$, $\omega_{v,2} = \frac{16L_{2,G}^2}{9}\left((d+5)^{5/2}+(d+3)^{3/2}\right)^2$, $\omega_k^v = \left(1+\frac{2}{\mu_G\rho_k}\right)$ and $h_k \leq \frac{\mu_G}{4\sqrt{\omega_{v,2}}}$. Then, we have

$$\begin{aligned}\mathbb{E}_k\left[\|v_{k+1}-v^*(x_{k+1})\|^2\right] \leq\ &(1-\mu_G\rho_k)\|v_k-v^*(x_k)\|^2+4\frac{\omega_{v,1}^2}{\mu_G}\rho_k\|z_k-z^*(x_k)\|^2\\&+2\rho_k^2\mathbb{E}_k\left[\|D_v^k\|^2\right]+2\omega_k^vL_*^2\gamma_k^2\mathbb{E}_k\left[\|D_x^k\|^2\right]\\&+\left(2\frac{L_{1,G}^2(d+3)^3}{\mu_G}+4\frac{\omega_{v,2}L_{0,F}^2}{\mu_G^3}\right)\rho_kh_k^2.\end{aligned}\tag{36}$$

Proof. We start by proving eq. 35. Adding and subtracting $z^*(x_k)$,

$$\begin{aligned}\|z_{k+1}-z^*(x_{k+1})\|^2 &= \|z_{k+1}-z^*(x_k)+z^*(x_k)-z^*(x_{k+1})\|^2\\&=\|z_{k+1}-z^*(x_k)\|^2+\|z^*(x_{k+1})-z^*(x_k)\|^2\\&+2\left\langle z_{k+1}-z^*(x_k),z^*(x_k)-z^*(x_{k+1})\right\rangle\\&=\underbrace{\|z_{k+1}-z^*(x_k)\|^2}_{A}+\underbrace{\|z^*(x_{k+1})-z^*(x_k)\|^2}_{B}\\&\underbrace{-2\left\langle z_{k+1}-z^*(x_k),z^*(x_{k+1})-z^*(x_k)\right\rangle}_{C}.\end{aligned}\tag{37}$$

We start by bounding A. By eq. 5, we have

$$\|z_{k+1} - z^*(x_k)\|^2 = \|z_k - z^*(x_k)\|^2 + \rho_k^2 \left\|D_z^k\right\|^2 - 2\rho_k \left\langle D_z^k, z_k - z^*(x_k)\right\rangle.$$

Taking the conditional expectation, by eq. 4,

$$\mathbb{E}_k\left[\|z_{k+1} - z^*(x_k)\|^2\right] = \|z_k - z^*(x_k)\|^2 + \rho_k^2 \mathbb{E}_k\left[\left\|D_z^k\right\|^2\right] - 2\rho_k \left\langle \nabla_z G_{h_k}(z_k, x_k), z_k - z^*(x_k)\right\rangle.$$

Adding and subtracting $\nabla_z G(z_k, x_k)$, we get

$$\begin{aligned}\mathbb{E}_k\left[\|z_{k+1} - z^*(x_k)\|^2\right] &= \|z_k - z^*(x_k)\|^2 + \rho_k^2 \mathbb{E}_k\left[\left\|D_z^k\right\|^2\right] \\ &\quad - 2\rho_k \left\langle \nabla_z G(z_k, x_k), z_k - z^*(x_k)\right\rangle \\ &\quad - 2\rho_k \left\langle \nabla_z G_{h_k}(z_k, x_k) - \nabla_z G(z_k, x_k), z_k - z^*(x_k)\right\rangle.\end{aligned}$$

By μ_G-strong convexity of G, we have

$$\begin{aligned}\mathbb{E}_k\left[\|z_{k+1} - z^*(x_k)\|^2\right] &= (1 - 2\mu_G\rho_k)\|z_k - z^*(x_k)\|^2 + \rho_k^2 \mathbb{E}_k\left[\left\|D_z^k\right\|^2\right] \\ &\quad - 2\rho_k \left\langle \nabla_z G_{h_k}(z_k, x_k) - \nabla_z G(z_k, x_k), z_k - z^*(x_k)\right\rangle.\end{aligned}$$

By Cauchy-Schwartz and Lemma D.7,

$$\begin{aligned}\mathbb{E}_k\left[\|z_{k+1} - z^*(x_k)\|^2\right] &= (1 - 2\mu_G\rho_k)\|z_k - z^*(x_k)\|^2 + \rho_k^2 \mathbb{E}_k\left[\left\|D_z^k\right\|^2\right] \\ &\quad + 2\rho_k \left(\frac{L_{1,G}}{2}(d+3)^{3/2} h_k\right)\|z_k - z^*(x_k)\|.\end{aligned}$$

By Young's Lemma with parameter $\nu_1 > 0$,

$$\begin{aligned}\mathbb{E}_k\left[\|z_{k+1} - z^*(x_k)\|^2\right] &= (1 - 2\mu_G\rho_k)\|z_k - z^*(x_k)\|^2 + \rho_k^2 \mathbb{E}_k\left[\left\|D_z^k\right\|^2\right] \\ &\quad + \frac{\rho_k}{\nu_1}\left(\frac{L_{1,G}}{2}(d+3)^{3/2} h_k\right)^2 + \nu_1\rho_k\|z_k - z^*(x_k)\|^2.\end{aligned}$$

Let $\nu_1 = \mu_G$,

$$\begin{aligned}\mathbb{E}_k\left[\|z_{k+1} - z^*(x_k)\|^2\right] &= (1 - \mu_G\rho_k)\|z_k - z^*(x_k)\|^2 + \rho_k^2 \mathbb{E}_k\left[\left\|D_z^k\right\|^2\right] \\ &\quad + \frac{L_{1,G}^2}{4\mu_G}(d+3)^3 \rho_k h_k^2.\end{aligned} \tag{38}$$

Now we bound the term B. By Lemma D.2 and eq. 9, we have

$$\|z^*(x_{k+1}) - z^*(x_k)\|^2 \le L_*^2\|x_{k+1} - x_k\|^2 = L_*^2\gamma_k^2\left\|D_x^k\right\|^2. \tag{39}$$

Finally, we bound the term C. By eq. 9 and Cauchy-Schwartz,

$$\begin{aligned}-2\left\langle z_{k+1} - z^*(x_k), z^*(x_{k+1}) - z^*(x_k)\right\rangle &= -2\left\langle z_k - z^*(x_k), z^*(x_{k+1}) - z^*(x_k)\right\rangle \\ &\quad + 2\rho_k\left\langle D_z^k, z^*(x_{k+1}) - z^*(x_k)\right\rangle \\ &\le -2\left\langle z_k - z^*(x_k), z^*(x_{k+1}) - z^*(x_k)\right\rangle \\ &\quad + 2\rho_k\|D_z^k\|\|z^*(x_{k+1}) - z^*(x_k)\|.\end{aligned}$$

By Lemma D.2, z^* is L_*-Lipschitz. Thus,

$$\begin{aligned}-2\left\langle z_{k+1} - z^*(x_k), z^*(x_{k+1}) - z^*(x_k)\right\rangle &\le -2\left\langle z_k - z^*(x_k), z^*(x_{k+1}) - z^*(x_k)\right\rangle \\ &\quad + 2L_*\rho_k\|D_z^k\|\|x_{k+1} - x_k\|.\end{aligned}$$

By Young's inequality and eq. 9,

$$\begin{aligned}-2\left\langle z_{k+1} - z^*(x_k), z^*(x_{k+1}) - z^*(x_k)\right\rangle &\le -2\left\langle z_k - z^*(x_k), z^*(x_{k+1}) - z^*(x_k)\right\rangle \\ &\quad + \rho_k^2\left\|D_z^k\right\|^2 + L_*^2\gamma_k^2\left\|D_x^k\right\|^2.\end{aligned}$$

By Cauchy-Schwartz and Lemma D.2,

$$-2\left\langle z_{k+1}-z^*(x_k), z^*(x_{k+1})-z^*(x_k)\right\rangle \le 2L_*\|z_k-z^*(x_k)\|\|x_{k+1}-x_k\| + \rho_k^2\left\|D_z^k\right\|^2 + L_*^2\gamma_k^2\left\|D_x^k\right\|^2.$$

Let $\nu_2>0$, by $ab\le \nu_2 a^2+\frac{b^2}{\nu_2}$

$$-2\left\langle z_{k+1}-z^*(x_k), z^*(x_{k+1})-z^*(x_k)\right\rangle \le 2\nu_2\|z_k-z^*(x_k)\|^2 + 2\frac{L_*^2}{\nu_2}\|x_{k+1}-x_k\|^2 + \rho_k^2\left\|D_z^k\right\|^2 + L_*^2\gamma_k^2\left\|D_x^k\right\|^2.$$

Let $\nu_2=\frac{\mu_G\rho_k}{4}$,

$$-2\left\langle z_{k+1}-z^*(x_k), z^*(x_{k+1})-z^*(x_k)\right\rangle \le \frac{\mu_G}{2}\rho_k\|z_k-z^*(x_k)\|^2 + \frac{8L_*^2\gamma_k^2}{\mu_G\rho_k}\|D_x^k\|^2 + \rho_k^2\left\|D_z^k\right\|^2 + L_*^2\gamma_k^2\left\|D_x^k\right\|^2. \tag{40}$$

Restarting from eq. 37, taking the conditional expectation and using equations eq. 38, eq. 39 and eq. 40, we get

$$\begin{aligned}\mathbb{E}_k\left[\|z_{k+1}-z^*(x_{k+1})\|^2\right] \le{}& \left(1-\frac{\mu_G\rho_k}{2}\right)\|z_k-z^*(x_k)\|^2 + 2\rho_k^2\mathbb{E}_k\left[\left\|D_z^k\right\|^2\right]\\ &+2\left(1+\frac{4}{\mu_G\rho_k}\right)L_*^2\gamma_k^2\mathbb{E}_k\left[\left\|D_x^k\right\|^2\right]\\ &+\frac{L_{1,G}^2}{4\mu_G}(d+3)^3\rho_k h_k^2\end{aligned}$$

Now we prove eq. 36. Adding and subtracting $v^*(x_k)$, we have

$$\begin{aligned}\|v_{k+1}-v^*(x_{k+1})\|^2 &= \|v_{k+1}-v^*(x_k)+v^*(x_k)-v^*(x_{k+1})\|^2\\ &= \|v_{k+1}-v^*(x_k)\|^2+\|v^*(x_{k+1})-v^*(x_k)\|^2+2\left\langle v_{k+1}-v^*(x_k), v^*(x_k)-v^*(x_{k+1})\right\rangle\\ &= \underbrace{\|v_{k+1}-v^*(x_k)\|^2}_{A}+\underbrace{\|v^*(x_{k+1})-v^*(x_k)\|^2}_{B}\underbrace{-2\left\langle v_{k+1}-v^*(x_k), v^*(x_{k+1})-v^*(x_k)\right\rangle}_{C}.\end{aligned} \tag{41}$$

We start bounding the term A. By eq. 7 we have

$$\begin{aligned}\|v_{k+1}-v^*(x_k)\|^2 &= \|v_k-\rho_k D_v^k-v^*(x_k)\|^2\\ &= \|v_k-v^*(x_k)\|^2+\rho_k^2\|D_v^k\|^2-2\rho_k\left\langle D_v^k, v_k-v^*(x_k)\right\rangle.\end{aligned}$$

Let $\hat{D}_v^k=\nabla_{zz}G_{h_k}(z_k,x_k)v_k+\nabla_z F_{h_k}(z_k,x_k)$. Observing that $\mathbb{E}_k[D_v^k]=\hat{D}_v^k$, taking the conditional expectation, we have

$$\mathbb{E}_k\left[\|v_{k+1}-v^*(x_k)\|^2\right]=\|v_k-v^*(x_k)\|^2+\rho_k^2\mathbb{E}_k\left[\|D_v^k\|^2\right]-2\rho_k\left\langle \hat{D}_v^k, v_k-v^*(x_k)\right\rangle.$$

Adding and subtracting $\bar{D}_v^k=\nabla_{zz}^2G(z_k,x_k)v_k+\nabla_z F(z_k,x_k)$,

$$\begin{aligned}\mathbb{E}_k\left[\|v_{k+1}-v^*(x_k)\|^2\right] &= \|v_k-v^*(x_k)\|^2+\rho_k^2\mathbb{E}_k\left[\|D_v^k\|^2\right]\\ &\quad-2\rho_k\left\langle \hat{D}_v^k-\bar{D}_v^k, v_k-v^*(x_k)\right\rangle\underbrace{-2\rho_k\left\langle \bar{D}_v^k, v_k-v^*(x_k)\right\rangle}_{A_1}.\end{aligned} \tag{42}$$

We focus on A_1. Let $\bar{D}_{v,*}^k=\nabla_{zz}^2G(z^*(x_k),x_k)v^*(x_k)+\nabla_z F(z^*(x_k),x_k)=0$. Thus, we have

$$\begin{aligned}-2\rho_k\left\langle \bar{D}_v^k, v_k-v^*(x_k)\right\rangle &= -2\rho_k\left\langle \bar{D}_v^k-\bar{D}_{v,*}^k, v_k-v^*(x_k)\right\rangle\\ &= -2\rho_k\left\langle \nabla_{zz}^2G(z_k,x_k)v_k-\nabla_{zz}^2G(z^*(x_k),x_k)v^*(x_k), v_k-v^*(x_k)\right\rangle\\ &\quad-2\rho_k\left\langle \nabla_z F(z_k,x_k)-\nabla_z F(z^*(x_k),x_k), v_k-v^*(x_k)\right\rangle.\end{aligned}$$

Adding and subtracting $\nabla^2_{zz}G(z_k,x_k)v^*(x_k)$,

$$\begin{aligned}-2\rho_k\left\langle \bar{D}_v^k, v_k - v^*(x_k)\right\rangle &= -2\rho_k\left\langle \nabla^2_{zz}G(z_k,x_k)v_k - \nabla^2_{zz}G(z_k,x_k)v^*(x_k), v_k - v^*(x_k)\right\rangle \\ &\quad - 2\rho_k\left\langle \nabla^2_{zz}G(z_k,x_k)v^*(x_k) - \nabla^2_{zz}G(z^*(x_k),x_k)v^*(x_k), v_k - v^*(x_k)\right\rangle \\ &\quad - 2\rho_k\left\langle \nabla_z F(z_k,x_k) - \nabla_z F(z^*(x_k),x_k), v_k - v^*(x_k)\right\rangle.\end{aligned}$$

By strong-convexity of G, Cauchy-Schwartz and Lemma D.3, we get

$$\begin{aligned}-2\rho_k\left\langle \bar{D}_v^k, v_k - v^*(x_k)\right\rangle &\le -2\rho_k\mu_G\|v_k - v^*(x_k)\|^2 \\ &\quad + \frac{2L_{0,F}}{\mu_G}\rho_k\|\nabla^2_{zz}G(z_k,x_k) - \nabla^2_{zz}G(z^*(x_k),x_k)\|\|v_k - v^*(x_k)\| \\ &\quad + 2\rho_k\|\nabla_z F(z_k,x_k) - \nabla_z F(z^*(x_k),x_k)\|\|v_k - v^*(x_k)\|.\end{aligned}$$

Since hessian of G are $L_{2,G}$-Lipschitz and F is $L_{1,F}$-smooth, we have

$$\begin{aligned}-2\rho_k\left\langle \bar{D}_v^k, v_k - v^*(x_k)\right\rangle &\le -2\rho_k\mu_G\|v_k - v^*(x_k)\|^2 \\ &\quad + \frac{2L_{0,F}}{\mu_G}L_{2,G}\rho_k\|z_k - z^*(x_k)\|\|v_k - v^*(x_k)\| \\ &\quad + 2L_{1,F}\rho_k\|z_k - z^*(x_k)\|\|v_k - v^*(x_k)\|.\end{aligned}$$

Let $\omega_{v,1} = \left(L_{1,F} + \frac{L_{0,F}L_{2,G}}{\mu_G}\right)$, we get

$$-2\rho_k\left\langle \bar{D}_v^k, v_k - v^*(x_k)\right\rangle \le -2\rho_k\mu_G\|v_k - v^*(x_k)\|^2 + 2\rho_k\omega_{v,1}\|z_k - z^*(x_k)\|\|v_k - v^*(x_k)\|.$$

By Young's inequality with parameter $\nu_1 > 0$, we have

$$\begin{aligned}-2\rho_k\left\langle \bar{D}_v^k, v_k - v^*(x_k)\right\rangle &\le -2\rho_k\mu_G\|v_k - v^*(x_k)\|^2 \\ &\quad + 2\rho_k\left(\frac{\omega_{v,1}^2}{2\nu_1}\|z_k - z^*(x_k)\|^2 + \frac{\nu_1}{2}\|v_k - v^*(x_k)\|^2\right) \\ &= -(2\mu_G - \nu_1)\rho_k\|v_k - v^*(x_k)\|^2 + \frac{\omega_{v,1}^2}{\nu_1}\rho_k\|z_k - z^*(x_k)\|^2.\end{aligned}$$

Using this inequality in eq. 42, we get

$$\begin{aligned}\mathbb{E}_k\left[\|v_{k+1} - v^*(x_k)\|^2\right] &= (1 - 2\mu_G\rho_k + \nu_1\rho_k)\|v_k - v^*(x_k)\|^2 + \rho_k^2\mathbb{E}_k\left[\|D_v^k\|^2\right] \\ &\quad - 2\rho_k\left\langle \hat{D}_v^k - \bar{D}_v^k, v_k - v^*(x_k)\right\rangle + \frac{\omega_{v,1}^2}{\nu_1}\rho_k\|z_k - z^*(x_k)\|^2.\end{aligned}$$

By Cauchy-Schwartz and Young's inequality with parameter $\nu_2 > 0$, we have

$$\begin{aligned}\mathbb{E}_k\left[\|v_{k+1} - v^*(x_k)\|^2\right] &= (1 - 2\mu_G\rho_k + \nu_1\rho_k + \nu_2\rho_k)\|v_k - v^*(x_k)\|^2 + \rho_k^2\mathbb{E}_k\left[\|D_v^k\|^2\right] \\ &\quad + \frac{\rho_k}{\nu_2}\underbrace{\|\hat{D}_v^k - \bar{D}_v^k\|^2}_{A_2} + \frac{\omega_{v,1}^2}{\nu_1}\rho_k\|z_k - z^*(x_k)\|^2. \qquad (43)\end{aligned}$$

Now we bound the term A_2.

$$\begin{aligned}\|\hat{D}_v^k - \bar{D}_v^k\|^2 &= \|\nabla^2_{zz}G_{h_k}(z_k,x_k)v_k + \nabla_z F_{h_k}(z_k,x_k) - \nabla^2_{zz}G(z_k,x_k)v_k - \nabla_z F(z_k,x_k)\|^2 \\ &\le 2\|\left(\nabla^2_{zz}G_{h_k}(z_k,x_k) - \nabla^2_{zz}G(z_k,x_k)\right)v_k\|^2 + 2\|\nabla_z F_{h_k}(z_k,x_k) - \nabla_z F(z_k,x_k)\|^2 \\ &\le 2\|\nabla^2_{zz}G_{h_k}(z_k,x_k) - \nabla^2_{zz}G(z_k,x_k)\|^2\|v_k\|^2 + 2\|\nabla_z F_{h_k}(z_k,x_k) - \nabla_z F(z_k,x_k)\|^2.\end{aligned}$$

By Lemma D.7, we have

$$\|\hat{D}_v^k - \bar{D}_v^k\|^2 \le \frac{8L_{2,G}^2}{9}\left((d+5)^{5/2} + (d+3)^{3/2}\right)^2 h_k^2\|v_k\|^2 + \frac{L_{1,G}^2}{2}(d+3)^3h_k^2.$$

Adding and subtracting $v^*(x_k)$, we get

$$\begin{aligned}\|\hat{D}_v^k - \bar{D}_v^k\|^2 &\le \frac{16L_{2,G}^2}{9}\left((d+5)^{5/2} + (d+3)^{3/2}\right)^2 h_k^2\|v_k - v^*(x_k)\|^2 \\ &\quad + \frac{16L_{2,G}^2}{9}\left((d+5)^{5/2} + (d+3)^{3/2}\right)^2 h_k^2\|v^*(x_k)\|^2 + \frac{L_{1,G}^2}{2}(d+3)^3h_k^2.\end{aligned}$$

Again by Lemma D.3, we have

$$\begin{aligned}\|\hat{D}_v^k - \bar{D}_v^k\|^2 \leq \frac{16L_{2,G}^2}{9}\left((d+5)^{5/2} + (d+3)^{3/2}\right)^2 h_k^2\|v_k - v^*(x_k)\|^2 \\ + \frac{16L_{2,G}^2}{9}\left((d+5)^{5/2} + (d+3)^{3/2}\right)^2 \frac{L_{0,F}^2}{\mu_G^2}h_k^2 \\ + \frac{L_{1,G}^2}{2}(d+3)^3 h_k^2.\end{aligned}$$

Let $\omega_{v,2} = \frac{16L_{2,G}^2}{9}\left((d+5)^{5/2} + (d+3)^{3/2}\right)^2$. Using this inequality in eq. 43, we get

$$\begin{aligned}\mathbb{E}_k\left[\|v_{k+1} - v^*(x_k)\|^2\right] \leq \left(1 - 2\mu_G\rho_k + \nu_1\rho_k + \nu_2\rho_k + \omega_{v,2}h_k^2\frac{\rho_k}{\nu_2}\right)\|v_k - v^*(x_k)\|^2 \\ + \rho_k^2\mathbb{E}_k\left[\|D_v^k\|^2\right] + \frac{\omega_{v,1}^2}{\nu_1}\rho_k\|z_k - z^*(x_k)\|^2 \\ + \frac{L_{1,G}^2}{2}(d+3)^3\frac{\rho_k}{\nu_2}h_k^2 + \omega_{v,2}\frac{L_{0,F}^2}{\mu_G^2}\frac{\rho_k}{\nu_2}h_k^2.\end{aligned}$$

Setting $\nu_1 = \nu_2 = \frac{\mu_G}{4}$ and since $h_k^2 \leq \frac{\mu_G^2}{16\omega_{v,2}}$, we have

$$\begin{aligned}\mathbb{E}_k\left[\|v_{k+1} - v^*(x_k)\|^2\right] \leq \left(1 - \frac{5\mu_G}{4}\rho_k\right)\|v_k - v^*(x_k)\|^2 \\ + \rho_k^2\mathbb{E}_k\left[\|D_v^k\|^2\right] + 4\frac{\omega_{v,1}^2}{\mu_G}\rho_k\|z_k - z^*(x_k)\|^2 \\ + 2\frac{L_{1,G}^2(d+3)^3}{\mu_G}\rho_k h_k^2 + 4\frac{\omega_{v,2}L_{0,F}^2}{\mu_G^3}\rho_k h_k^2.\end{aligned} \tag{44}$$

Now, we bound the term B of eq. 41. By Lemma D.2, v^* is L_*-Lipschitz continuous. Thus, by eq. 9, we have

$$\|v^*(x_{k+1}) - v^*(x_k)\|^2 \leq L_*^2\|x_{k+1} - x_k\|^2 = L_*^2\gamma_k^2\|D_x^k\|^2. \tag{45}$$

Finally, we bound the term C of eq. 41. By eq. 7, we have

$$\begin{aligned}-2\left\langle v_{k+1} - v^*(x_k), v^*(x_{k+1}) - v^*(x_k)\right\rangle = -2\left\langle v_k - v^*(x_k), v^*(x_{k+1}) - v^*(x_k)\right\rangle \\ + 2\rho_k\left\langle D_v^k, v^*(x_{k+1}) - v^*(x_k)\right\rangle.\end{aligned}$$

By Cauchy-Schwartz and Lemma D.2, we have

$$\begin{aligned}-2\left\langle v_{k+1} - v^*(x_k), v^*(x_{k+1}) - v^*(x_k)\right\rangle \leq -2\left\langle v_k - v^*(x_k), v^*(x_{k+1}) - v^*(x_k)\right\rangle \\ + 2L_*\rho_k\|D_v^k\|\|x_{k+1} - x_k\|.\end{aligned}$$

By Young's inequality with $\nu_3 > 0$ and eq. 9, we get

$$\begin{aligned}-2\left\langle v_{k+1} - v^*(x_k), v^*(x_{k+1}) - v^*(x_k)\right\rangle \leq -2\left\langle v_k - v^*(x_k), v^*(x_{k+1}) - v^*(x_k)\right\rangle \\ + \rho_k^2\nu_3\|D_v^k\|^2 + \frac{L_*^2}{\nu_3}\gamma_k^2\|D_x^k\|^2.\end{aligned}$$

Similarly, by Cauchy-Schwartz, Lemma D.2 (i.e. L_*-Lipschitz of v^*) and by Young's inequality with parameter $\nu_4 > 0$,

$$\begin{aligned}-2\left\langle v_{k+1} - v^*(x_k), v^*(x_{k+1}) - v^*(x_k)\right\rangle &\leq 2L_*\|v_k - v^*(x_k)\|\|x_{k+1} - x_k\| \\ &\quad + \rho_k^2\nu_3\|D_v^k\|^2 + \frac{L_*^2}{\nu_3}\gamma_k^2\|D_x^k\|^2 \\ &\leq \nu_4\|v_k - v^*(x_k)\|^2 + \frac{1}{\nu_4}L_*^2\|x_{k+1} - x_k\|^2 \\ &\quad + \rho_k^2\nu_3\|D_v^k\|^2 + \frac{L_*^2}{\nu_3}\gamma_k^2\|D_x^k\|^2.\end{aligned}$$

Let $\nu_3 = 1$ and $\nu_4 = \frac{\mu_G \rho_k}{4}$, we have

$$-2\left\langle v_{k+1} - v^*(x_k), v^*(x_{k+1}) - v^*(x_k)\right\rangle \leq \frac{\mu_G}{4}\rho_k \|v_k - v^*(x_k)\|^2 + \frac{4}{\mu_G \rho_k} L_*^2 \gamma_k^2 \|D_x^k\|^2 + \rho_k^2 \|D_v^k\|^2 + L_*^2 \gamma_k^2 \|D_x^k\|^2. \tag{46}$$

Let $\omega_k^v = \left(1 + \frac{2}{\mu_G \rho_k}\right)$. Restarting from eq. 41, taking the conditional expectation and by equations eq. 44, eq. 45,eq. 46, we get the claim

$$\begin{aligned}\mathbb{E}_k\left[\|v_{k+1} - v^*(x_{k+1})\|^2\right] \leq{}& (1 - \mu_G \rho_k)\|v_k - v^*(x_k)\|^2 + 4\frac{\omega_{v,1}^2}{\mu_G}\rho_k \|z_k - z^*(x_k)\|^2 \\ &+ 2\rho_k^2 \mathbb{E}_k\left[\|D_v^k\|^2\right] + 2\omega_k^v L_*^2 \gamma_k^2 \mathbb{E}_k\left[\|D_x^k\|^2\right] \\ &+ 2\frac{L_{1,G}^2 (d+3)^3}{\mu_G}\rho_k h_k^2 + 4\frac{\omega_{v,2} L_{0,F}^2}{\mu_G^3}\rho_k h_k^2.\end{aligned}$$

□

The following lemma characterizes the descent of the value function Ψ along the sequence of the iterates x_k generated by Algorithm 1.

Lemma D.14 (Value Function sequence bound)**.** *Under Assumption 3.2, 3.3. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. Let $\omega_{v,1} = \left(L_{1,F} + \frac{L_{0,F} L_{2,G}}{\mu_G}\right)$. Then,*

$$\begin{aligned}\mathbb{E}_k\left[\Psi(x_{k+1})\right] \leq{}& \Psi(x_k) - \frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 - \frac{\gamma_k}{2}\|\hat{D}_x^k\|^2 + \frac{L_\Psi}{2}\gamma_k^2 \mathbb{E}_k\left[\|D_x^k\|^2\right] \\ &+ 2\gamma_k \omega_{v,1}^2 \|z_k - z^*(x_k)\|^2 + 2\gamma_k L_{1,G}^2 \|v_k - v^*(x_k)\|^2 \\ &+ C_1^2 \gamma_k h_k^2,\end{aligned}$$

where $C_1 = \left(2\frac{L_{0,F} L_{2,G}}{3\mu_G}\left((d+5)^{5/2} + (d+3)^{3/2} + (d+1)(p+3)^{3/2}\right) + \frac{L_{1,F}}{2}(d+3)^{3/2}\right)$.

Proof. By Lemma D.2, Ψ is L_Ψ-smooth. Thus, by the Descent Lemma (Polyak, 1987) we have

$$\Psi(x_{k+1}) \leq \Psi(x_k) + \langle \nabla\Psi(x_k), x_{k+1} - x_k\rangle + \frac{L_\Psi}{2}\|x_{k+1} - x_k\|^2.$$

By eq. 9, we have

$$\Psi(x_{k+1}) \leq \Psi(x_k) - \gamma_k \left\langle \nabla\Psi(x_k), D_x^k\right\rangle + \frac{L_\Psi}{2}\gamma_k^2 \|D_x^k\|^2.$$

Let $\hat{D}_x^k = \nabla_{xz}^2 G_{h_k}(z_k, x_k) v_k + \nabla_x F_{h_k}(z_k, x_k)$. Taking the conditional expectation, we get

$$\mathbb{E}_k[\Psi(x_{k+1})] \leq \Psi(x_k) - \gamma_k \left\langle \nabla\Psi(x_k), \hat{D}_x^k\right\rangle + \frac{L_\Psi}{2}\gamma_k^2 \mathbb{E}_k\left[\|D_x^k\|^2\right].$$

By $\langle a, b\rangle = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a - b\|^2\right)$, we have

$$\begin{aligned}\mathbb{E}_k\left[\Psi(x_{k+1})\right] \leq{}& \Psi(x_k) - \frac{\gamma_k}{2}\left(\|\nabla\Psi(x_k)\|^2 + \|\hat{D}_x^k\|^2 - \|\hat{D}_x^k - \nabla\Psi(x_k)\|^2\right) \\ &+ \frac{L_\Psi}{2}\gamma_k^2 \mathbb{E}_k\left[\|D_x^k\|^2\right] \\ ={}& \Psi(x_k) - \frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 - \frac{\gamma_k}{2}\|\hat{D}_x^k\|^2 \\ &+ \frac{\gamma_k}{2}\|\hat{D}_x^k - \nabla\Psi(x_k)\|^2 + \frac{L_\Psi}{2}\gamma_k^2 \mathbb{E}_k\left[\|D_x^k\|^2\right].\end{aligned}$$

Let $\omega_{v,1} = \left(L_{1,F} + \frac{L_{0,F}L_{2,G}}{\mu_G}\right)$. By Lemma D.8 with $h_1 = h_k$, $\eta_1 = h_k$, $h_2 = 0$ and $\eta_2 = h_k$, we get

$$\begin{aligned}\mathbb{E}_k\left[\Psi(x_{k+1})\right] \leq \Psi(x_k) &- \frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 - \frac{\gamma_k}{2}\|\hat{D}_x^k\|^2 \\ &+ \frac{L_\Psi}{2}\gamma_k^2\mathbb{E}_k\left[\|D_x^k\|^2\right] + \frac{\gamma_k}{2}\Bigg(\omega_{v,1}\|z_k - z^*(x_k)\| + L_{1,G}\|v_k - v^*(x_k)\| \\ &+ 2\frac{L_{0,F}L_{2,G}}{3\mu_G}\left(h_k(d+5)^{5/2} + h_k(d+3)^{3/2} + h_k(d+1)(p+3)^{3/2}\right) \\ &+ \frac{L_{1,F}}{2}(d+3)^{3/2}h_k^2\Bigg)^2\end{aligned}$$

Let $C_1 := \left(2\frac{L_{0,F}L_{2,G}}{3\mu_G}\left((d+5)^{5/2} + (d+3)^{3/2} + (d+1)(p+3)^{3/2}\right) + \frac{L_{1,F}}{2}(d+3)^{3/2}\right)h_k$.

$$\begin{aligned}\mathbb{E}_k\left[\Psi(x_{k+1})\right] \leq \Psi(x_k) &- \frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 - \frac{\gamma_k}{2}\|\hat{D}_x^k\|^2 \\ &+ \frac{L_\Psi}{2}\gamma_k^2\mathbb{E}_k\left[\|D_x^k\|^2\right] + \frac{\gamma_k}{2}\Bigg(\omega_{v,1}\|z_k - z^*(x_k)\| + L_{1,G}\|v_k - v^*(x_k)\| \\ &+ C_1h_k\Bigg)^2.\end{aligned}$$

By $(a+b)^2 \leq 2a^2 + 2b^2$, we get

$$\begin{aligned}\mathbb{E}_k\left[\Psi(x_{k+1})\right] &\leq \Psi(x_k) - \frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 - \frac{\gamma_k}{2}\|\hat{D}_x^k\|^2 \\ &\quad + \frac{L_\Psi}{2}\gamma_k^2\mathbb{E}_k\left[\|D_x^k\|^2\right] \\ &\quad + \gamma_k\left(\omega_{v,1}\|z_k - z^*(x_k)\| + L_{1,G}\|v_k - v^*(x_k)\|\right)^2 \\ &\quad + C_1^2\gamma_k h_k^2 \\ &\leq \Psi(x_k) - \frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 - \frac{\gamma_k}{2}\|\hat{D}_x^k\|^2 + \frac{L_\Psi}{2}\gamma_k^2\mathbb{E}_k\left[\|D_x^k\|^2\right] \\ &\quad + 2\gamma_k\omega_{v,1}^2\|z_k - z^*(x_k)\|^2 + 2\gamma_k L_{1,G}^2\|v_k - v^*(x_k)\|^2 \\ &\quad + C_1^2\gamma_k h_k^2.\end{aligned}$$

□

D.4 Auxiliary results for Algorithm 2

In this appendix, we collect all lemmas required for the analysis of Algorithm 2. Unlike Algorithm 1, which relies on unbiased estimators of the Hessian of the smoothing, Algorithm 2 directly approximates Hessian–vector products using first-order finite differences, and then replaces the resulting gradients with zeroth-order surrogates. We begin by introducing a lemma that bounds the approximation error of the first-order finite-difference for a twice-differentiable function with Lipschitz-continuous Hessians.

Lemma D.15 (First-order approximation of Hessian-vector products)**.** *Let $f : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ be a twice-differentiable function with $L_{2,G}$-Lipschitz continuous hessian. Then, for every $z, v \in \mathbb{R}^p$, $x \in \mathbb{R}^d$ and $\bar{h} > 0$,*

$$\left\|\frac{1}{\bar{h}}\left(\nabla_z f(z + \bar{h}v, x) - \nabla_z f(z, x)\right) - \nabla_{zz}^2 f(z, x)v\right\| \leq \frac{L_{2,G}}{2}\bar{h}\|v\|^2,$$
$$\left\|\frac{1}{\bar{h}}\left(\nabla_x f(z + \bar{h}v, x) - \nabla_x f(z, x)\right) - \nabla_{xz}^2 f(z, x)v\right\| \leq \frac{L_{2,G}}{2}\bar{h}\|v\|^2.$$

Proof. The proof of both inequalities follows the same line of Lemma 1.2.4 of Nesterov (2018). For completeness, we include the proof of the first inequality (the second follows analogously). Since f is twice differentiable, we have

$$\nabla_z f(z+\bar{h}v,x) = \nabla_z f(z,x) + \int_0^1 \nabla^2_{zz} f(z+t\bar{h}v,x)\,\bar{h}v\,dt.$$

Adding and subtracting $\nabla^2_{zz} f(z,x)\,\bar{h}v$

$$\nabla_z f(z+\bar{h}v,x) = \nabla_z f(z,x) + \nabla^2_{zz} f(z,x)\,\bar{h}v + \int_0^1 (\nabla^2_{zz} f(z+t\bar{h}v,x) - \nabla^2_{zz} f(z,x))\,\bar{h}v\,dt.$$

Therefore, we have

$$\begin{aligned}
\left\| \frac{1}{\bar{h}} \left(\nabla_z f(z+\bar{h}v,x) - \nabla_z f(z,x) \right) - \nabla^2_{zz} f(z,x)v \right\| &= \frac{1}{\bar{h}} \left\| \int_0^1 (\nabla^2_{zz} f(z+t\bar{h}v,x) - \nabla^2_{zz} f(z,x))\,\bar{h}v\,dt \right\| \\
&\le \frac{1}{\bar{h}} \int_0^1 \left\| (\nabla^2_{zz} f(z+t\bar{h}v,x) - \nabla^2_{zz} f(z,x))\,\bar{h}v \right\| dt \\
&\le \frac{1}{\bar{h}} \int_0^1 \left\| \nabla^2_{zz} f(z+t\bar{h}v,x) - \nabla^2_{zz} f(z,x) \right\| \|\bar{h}v\|\,dt
\end{aligned}$$

By $L_{2,G}$-Lipschitz continuity of the Hessian, we get the claim

$$\left\| \frac{1}{\bar{h}} \left(\nabla_z f(z+\bar{h}v,x) - \nabla_z f(z,x) \right) - \nabla^2_{zz} f(z,x)v \right\| \le L_{2,G}\bar{h}\|v\|^2 \int_0^1 t\,dt = \frac{L_{2,G}}{2}\bar{h}\|v\|^2.$$

□

In the following lemma, we derive bounds on the norm of the finite-difference approximations of the Hessian-block–vector products used in Algorithm 2.

Lemma D.16 (Bound on surrogate difference)**.** *Let Assumptions 3.2 and 3.3 hold. Let z_k, v_k and x_k be the sequences generated by Algorithm 2. Let $h_k, \bar{h}_k > 0$. Then, for every $k \in \mathbb{N}$*

$$\begin{aligned}
\mathbb{E}_k \left[\left\| \hat{\nabla}_z g_k(z_k+\bar{h}_k v_k, x_k) - \hat{\nabla}_z g_k(z_k,x_k) \right\|^2 \right] &\le C_1 L^2_{1,G} \bar{h}^2_k \|v_k - v^*(x_k)\|^2 \\
&\quad + C_1 \frac{L^2_{1,G} L^2_{0,F}}{\mu^2_G} \bar{h}^2_k + \frac{3L^2_{1,G}(p+6)^3}{b_1 \ell_1} h^2_k, \\
\mathbb{E}_k \left[\left\| \hat{\nabla}_x g_k(z_k+\bar{h}_k v_k, x_k) - \hat{\nabla}_x g_k(z_k,x_k) \right\|^2 \right] &\le C_2 L^2_{1,G} \bar{h}^2_k \|v_k - v^*(x_k)\|^2 \\
&\quad + C_2 \frac{L^2_{1,G} L^2_{0,F}}{\mu^2_G} \bar{h}^2_k + \frac{3L^2_{1,G}(d+6)^3}{b_1 \ell_1} h^2_k.
\end{aligned}$$

where $C_1 = 4\left(1 + \frac{3(p+2)}{b_1\ell_1}\right)$ and $C_2 = 4\left(1 + \frac{3(d+2)}{b_1\ell_1}\right)$.

Proof. We start by proving the first inequality. To simplify the reading we will use the following notation.

$$\begin{aligned}
g^{(i,j)}_{v,k} &= \frac{g(z_k + \bar{h}_k v_k + h_k w^{(i,j)}_k, x_k, \xi_{i,k}) - g(z_k + \bar{h}_k v_k, x_k, \xi_{i,k})}{h_k} w^{(i,j)}_k, \\
g^{(i,j)}_k &= \frac{g(z_k + h_k w^{(i,j)}_k, x_k, \xi_{i,k}) - g(z_k, x_k, \xi_{i,k})}{h_k} w^{(i,j)}_k, \\
\delta^{(i,j)}_k &= g^{(i,j)}_{v,k} - g^{(i,j)}_k - (\nabla_z g_{h_k}(z_k + \bar{h}_k v_k, x_k, \xi_{i,k}) - \nabla_z g_{h_k}(z_k, x_k, \xi_{i,k})), \\
\delta^{(i)}_k &= \sum_{j=1}^{\ell_1} \delta^{(i,j)}_k.
\end{aligned}$$

Adding and subtracting $\frac{1}{b_1\ell_1}\sum_{i=1}^{b_1}\sum_{j=1}^{\ell_1}\nabla_z g_{h_k}(z_k+\bar{h}_k v_k, x_k, \xi_{i,k}) - \nabla_z g_{h_k}(z_k, x_k, \xi_{i,k})$, we get

$$\mathbb{E}_k\left[\left\|\hat{\nabla}_z g_k(z_k+\bar{h}_k v_k, x_k) - \hat{\nabla}_z g_k(z_k, x_k)\right\|^2\right] \leq 2\mathbb{E}_k\left[\left\|\frac{1}{b_1\ell_1}\sum_{i=1}^{b_1}\delta_k^{(i)}\right\|^2\right] + 2\mathbb{E}_k\left[\left\|\nabla_z g_{h_k}(z_k+\bar{h}_k v_k, x_k, \xi_{i,k}) - \nabla_z g_{h_k}(z_k, x_k, \xi_{i,k})\right\|^2\right].$$

By $L_{1,G}$-smoothness of g, we get

$$\mathbb{E}_k\left[\left\|\hat{\nabla}_z g_k(z_k+\bar{h}_k v_k, x_k) - \hat{\nabla}_z g_k(z_k, x_k)\right\|^2\right] \leq 2\mathbb{E}_k\left[\left\|\frac{1}{b_1\ell_1}\sum_{i=1}^{b_1}\delta_k^{(i)}\right\|^2\right] + 2L_{1,G}^2\bar{h}_k^2\left\|v_k\right\|^2.$$

Adding and subtracting $v^*(x_k)$ and by Lemma D.3, we get

$$\mathbb{E}_k\left[\left\|\hat{\nabla}_z g_k(z_k+\bar{h}_k v_k, x_k) - \hat{\nabla}_z g_k(z_k, x_k)\right\|^2\right] \leq 2\mathbb{E}_k\left[\left\|\frac{1}{b_1\ell_1}\sum_{i=1}^{b_1}\delta_k^{(i)}\right\|^2\right] + 4L_{1,G}^2\bar{h}_k^2\left\|v_k - v^*(x_k)\right\|^2 + 4\frac{L_{1,G}^2 s L_{0,F}^2}{\mu_G^2}\bar{h}_k^2$$

Since for every $i = 1, \cdots, b_1$ and $t = 1, \cdots, b_1$ with $i \neq t$, $w_k^{(i,j)}$ and $w_k^{(t,j)}$ are independent for every $j = 1, \cdots, \ell_1$, we get

$$\mathbb{E}_k\left[\left\|\hat{\nabla}_z g_k(z_k+\bar{h}_k v_k, x_k) - \hat{\nabla}_z g_k(z_k, x_k)\right\|^2\right] \leq \frac{2}{b_1^2\ell_1^2}\left(\sum_{i=1}^{b_1}\mathbb{E}_k\left[\left\|\delta_k^{(i)}\right\|^2\right] + \sum_{i=1}^{b_1}\sum_{t\neq i}\left\langle\mathbb{E}_k[\delta_k^{(i)}], \mathbb{E}_k[\delta_k^{(t)}]\right\rangle\right) + 4L_{1,G}^2\bar{h}_k^2\left\|v_k - v^*(x_k)\right\|^2 + 4\frac{L_{1,G}^2 s L_{0,F}^2}{\mu_G^2}\bar{h}_k^2.$$

Observing that, by Lemma D.5, we have

$$\mathbb{E}_k[\delta_k^{(i)}] = \sum_{j=1}^{\ell_1}\mathbb{E}_{w_k^{(i,j)}}[g_{v,k}^{(i,j)} - g_k^{(i,j)}] - (\nabla_z g_{h_k}(z_k+\bar{h}_k v_k, x_k, \xi_{i,k}) - \nabla_z g_{h_k}(z_k, x_k, \xi_{i,k})) = 0,$$

we get

$$\mathbb{E}_k\left[\left\|\hat{\nabla}_z g_k(z_k+\bar{h}_k v_k, x_k) - \hat{\nabla}_z g_k(z_k, x_k)\right\|^2\right] \leq \frac{2}{b_1^2\ell_1^2}\sum_{i=1}^{b_1}\mathbb{E}_k\left[\left\|\delta_k^{(i)}\right\|^2\right] + 4L_{1,G}^2\bar{h}_k^2\left\|v_k - v^*(x_k)\right\|^2 + 4\frac{L_{1,G}^2 s L_{0,F}^2}{\mu_G^2}\bar{h}_k^2.$$

Similarly, recalling that $\delta_k^{(i)} = \sum_{j=1}^{\ell_1}\delta_k^{(i,j)}$, we have

$$\mathbb{E}_k\left[\left\|\hat{\nabla}_z g_k(z_k+\bar{h}_k v_k, x_k) - \hat{\nabla}_z g_k(z_k, x_k)\right\|^2\right] \leq \frac{2}{b_1^2\ell_1^2}\sum_{i=1}^{b_1}\left(\sum_{j=1}^{\ell_1}\mathbb{E}_k\left[\left\|\delta_k^{(i,j)}\right\|^2\right] + \sum_{j=1}^{\ell_1}\sum_{t\neq j}\mathbb{E}_k\left[\left\langle\delta_k^{(i,j)}, \delta_k^{(i,t)}\right\rangle\right]\right) + 4L_{1,G}^2\bar{h}_k^2\left\|v_k - v^*(x_k)\right\|^2 + 4\frac{L_{1,G}^2 s L_{0,F}^2}{\mu_G^2}\bar{h}_k^2.$$

For $j = 1, \cdots, \ell_1$ and $t = 1, \cdots, \ell_1$ with $j \neq t$, $w_k^{(i,j)}$ and $w_k^{(i,t)}$ are independent for every $i = 1, \cdots, b_1$. Thus, we get

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\hat{\nabla}_z g_k(z_k + \bar{h}_k v_k, x_k) - \hat{\nabla}_z g_k(z_k, x_k)\right\|^2\right] &\leq \frac{2}{b_1^2 \ell_1^2} \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \mathbb{E}_k\left[\left\|\delta_k^{(i,j)}\right\|^2\right] \\
&\quad + 4L_{1,G}^2 \bar{h}_k^2 \left\|v_k - v^*(x_k)\right\|^2 + 4\frac{L_{1,G}^2 s L_{0,F}^2}{\mu_G^2} \bar{h}_k^2 \\
&\leq \frac{2}{b_1^2 \ell_1^2} \sum_{i=1}^{b_1} \sum_{j=1}^{\ell_1} \underbrace{\mathbb{E}_k\left[\left\|g_{v,k}^{(i,j)} - g_k^{(i,j)}\right\|^2\right]}_{A} \\
&\quad + 4L_{1,G}^2 \bar{h}_k^2 \left\|v_k - v^*(x_k)\right\|^2 + 4\frac{L_{1,G}^2 s L_{0,F}^2}{\mu_G^2} \bar{h}_k^2.
\end{aligned} \tag{47}$$

Now, we focus on bounding A. Adding and subtracting $\left\langle \nabla g(z_k + \bar{h}_k v_k, x_k, \xi_{i,k}), [w_k^{(i,j)}, 0] \right\rangle$ and $\left\langle \nabla g(z_k, x_k, \xi_{i,k}), [w_k^{(i,j)}, 0] \right\rangle$, we have

$$\begin{aligned}
\mathbb{E}_k\left[\left\|g_{v,k}^{(i,j)} - g_k^{(i,j)}\right\|^2\right] &= \frac{1}{h_k^2}\mathbb{E}_k\Bigg[\Big(g(z_k + \bar{h}_k v_k + h_k w_k^{(i,j)}, x_k, \xi_{i,k}) - g(z_k + \bar{h}_k v_k, x_k, \xi_{i,k}) \\
&\quad - \left(g(z_k + h_k w_k^{(i,j)}, x_k, \xi_{i,k}) - g(z_k, x_k, \xi_{i,k})\right)\Big)^2 \left\|w_k^{(i,j)}\right\|^2\Bigg] \\
&\leq \frac{3}{h_k^2}\mathbb{E}_k\Bigg[\Big(g(z_k + \bar{h}_k v_k + h_k w_k^{(i,j)}, x_k, \xi_{i,k}) - g(z_k + \bar{h}_k v_k, x_k, \xi_{i,k}) \\
&\quad - \left\langle \nabla g(z_k + \bar{h}_k v_k, x_k, \xi_{i,k}), [h_k w_k^{(i,j)}, 0] \right\rangle\Big)^2 \|w_k^{(i,j)}\|^2\Bigg] \\
&\quad + \frac{3}{h_k^2}\mathbb{E}_k\Bigg[\Big(g(z_k + h_k w_k^{(i,j)}, x_k, \xi_{i,k}) - g(z_k, x_k, \xi_{i,k}) \\
&\quad - \left\langle \nabla g(z_k, x_k, \xi_{i,k}), [h_k w_k^{(i,j)}, 0] \right\rangle\Big)^2 \|w_k^{(i,j)}\|^2\Bigg] \\
&\quad + \frac{3}{h_k^2}\mathbb{E}_k\left[\left(\left\langle \nabla g(z_k + \bar{h}_k v_k, x_k, \xi_{i,k}) - \nabla g(z_k, x_k, \xi_{i,k}), [h_k w_k^{(i,j)}, 0] \right\rangle\right)^2 \|w_k^{(i,j)}\|^2\right].
\end{aligned}$$

Since g is $L_{1,G}$-smooth, by the Descent Lemma (Polyak, 1987),

$$\begin{aligned}
\mathbb{E}_k\left[\left\|g_{v,k}^{(i,j)} - g_k^{(i,j)}\right\|^2\right] &\leq \frac{3L_{1,G}^2}{2} h_k^2 \mathbb{E}_k[\|w_k^{(i,j)}\|^6] \\
&\quad + \frac{3}{h_k^2}\mathbb{E}_k\left[\left(\left\langle \nabla g(z_k + \bar{h}_k v_k, x_k, \xi_{i,k}) - \nabla g(z_k, x_k, \xi_{i,k}), [h_k w_k^{(i,j)}, 0] \right\rangle\right)^2 \|w_k^{(i,j)}\|^2\right] \\
&= \frac{3L_{1,G}^2}{2} h_k^2 \mathbb{E}_k[\|w_k^{(i,j)}\|^6] \\
&\quad + 3\mathbb{E}_{\xi_{i,k}}\Bigg[\left(\nabla_z g(z_k + \bar{h}_k v_k, x_k, \xi_{i,k}) - \nabla_z g(z_k, x_k, \xi_{i,k})\right)^\top \\
&\quad \mathbb{E}_{w_k^{(i,k)}}[w_k^{(i,j)} w_k^{(i,j)\top} w_k^{(i,j)} w_k^{(i,j)\top}](\nabla_z g(z_k + \bar{h}_k v_k, x_k, \xi_{i,k}) - \nabla_z g(z_k, x_k, \xi_{i,k}))\Bigg].
\end{aligned}$$

By Proposition D.1, we have

$$\begin{aligned}
\mathbb{E}_k\left[\left\|g_{v,k}^{(i,j)} - g_k^{(i,j)}\right\|^2\right] &\leq \frac{3L_{1,G}^2}{2}(p+6)^3 h_k^2 \\
&\quad + 3(p+2)\mathbb{E}_{\xi_{i,k}}\left[\|\nabla_z g(z_k + \bar{h}_k v_k, x_k, \xi_{i,k}) - \nabla_z g(z_k, x_k, \xi_{i,k})\|^2\right].
\end{aligned}$$

Again by $L_{1,G}$-smoothness of g, we have

$$\mathbb{E}_k\left[\left\|g_{v,k}^{(i,j)}-g_k^{(i,j)}\right\|^2\right]\leq\frac{3L_{1,G}^2}{2}(p+6)^3h_k^2+3(p+2)L_{1,G}^2\bar{h}_k^2\|v_k\|^2.$$

Adding and subtracting $v^*(x_k)$, by Lemma D.3, we get

$$\begin{aligned}\mathbb{E}_k\left[\left\|g_{v,k}^{(i,j)}-g_k^{(i,j)}\right\|^2\right]\leq&\frac{3L_{1,G}^2}{2}(p+6)^3h_k^2+6(p+2)L_{1,G}^2\bar{h}_k^2\|v_k-v^*(x_k)\|^2\\&+\frac{6(p+2)L_{1,G}^2L_{0,F}^2}{\mu_G^2}\bar{h}_k^2.\end{aligned}$$

Using this bound in eq. 47, we get the first claim

$$\begin{aligned}\mathbb{E}_k\left[\left\|\hat{\nabla}_z g_k(z_k+\bar{h}_kv_k,x_k)-\hat{\nabla}_z g_k(z_k,x_k)\right\|^2\right]\leq&4\left(1+\frac{3(p+2)}{b_1\ell_1}\right)L_{1,G}^2\bar{h}_k^2\|v_k-v^*(x_k)\|^2\\&+4\left(1+\frac{3(p+2)}{b_1\ell_1}\right)\frac{L_{1,G}^2L_{0,F}^2}{\mu_G^2}\bar{h}_k^2\\&+\frac{3L_{1,G}^2(p+6)^3}{b_1\ell_1}h_k^2.\end{aligned}$$

The proof of the second inequality follows the same line. □

In the next lemma, we provide bounds on the norm of the search directions computed in Algorithm 2.

Lemma D.17 (Bound on search directions)**.** *Let Assumptions 3.2,3.1 and 3.3 hold. Let $\hat{D}_z^k,\hat{D}_v^k$ and $\hat{D}_x^k$ be the search directions defined in eq. 10. Let z_k,v_k and x_k be the sequences generated by Algorithm 2. Then, for every $k\in\mathbb{N}$,*

$$\begin{aligned}\mathbb{E}_k\left[\|\hat{D}_z^k\|^2\right]\leq&2\left(2+\frac{p+2}{b_1\ell_1}\right)L_{1,G}^2\|z_k-z^*(x_k)\|^2\\&+\left(\frac{(p+6)^3}{2b_1\ell_1}+\left(\frac{3}{b_1}+1\right)(p+3)^3\right)L_{1,G}^2h_k^2+2\left(3+\frac{p+2}{\ell_1}\right)\frac{\sigma_{1,G}^2}{b_1}.\end{aligned}$$

Let $C_1=4\left(1+\frac{3(p+2)}{b_1\ell_1}\right)$, $B_1=2\left(1+\frac{p+2}{b_2\ell_2}\right)$, $B_2=\left(\frac{(p+6)^3}{2b_2\ell_2}+\left(\frac{3}{b_2}+1\right)(p+3)^3\right)L_{1,F}^2$ and $B_3=2\left(\frac{p+2}{\ell_2}+3\right)\frac{\sigma_{1,F}^2}{b_2}$. Then for every $k\in\mathbb{N}$,

$$\begin{aligned}\mathbb{E}_k\left[\left\|\hat{D}_v^k\right\|^2\right]\leq&2C_1L_{1,G}^2\|v_k-v^*(x_k)\|^2+2C_1\frac{L_{1,G}^2L_{0,F}^2}{\mu_G^2}\\&+\frac{6L_{1,G}^2(p+6)^3}{b_1\ell_1}\frac{h_k^2}{\bar{h}_k^2}+2B_1L_{0,F}^2+2B_2h_k^2+2B_3.\end{aligned}$$

Moreover, let

$$\begin{aligned}&C_2=4\left(1+\frac{3(d+2)}{b_1\ell_1}\right),\quad B_4=2\left(1+\frac{d+2}{b_2\ell_2}\right),\\&B_5=\left(\frac{(d+6)^3}{2b_2\ell_2}+\left(\frac{3}{b_2}+1\right)(d+3)^3\right)L_{1,F}^2,\\&B_6=2\left(\frac{d+2}{b_2\ell_2}+\frac{3}{b_2}\right)\sigma_{1,F}^2.\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}_k\left[\left\|\hat{D}_x^k\right\|^2\right]\leq&2C_2L_{1,G}^2\|v_k-v^*(x_k)\|^2+2C_2\frac{L_{1,G}^2L_{0,F}^2}{\mu_G^2}\\&+\frac{6L_{1,G}^2(d+6)^3}{b_1\ell_1}\frac{h_k^2}{\bar{h}_k^2}+2B_4L_{0,F}^2+2B_5h_k^2+2B_6.\end{aligned}$$

Proof. The proof of the first inequality follows the same line of the proof of eq. 32. We focus on proving the second inequality. By $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$,

$$\begin{aligned}\mathbb{E}_k\left[\left\|\hat{D}_v^k\right\|^2\right] &= \mathbb{E}_k\left[\left\|\frac{1}{\bar{h}_k}\left(\hat{\nabla}_z g_k(z_k+\bar{h}_k v_k, x_k) - \hat{\nabla}_z g_k(z_k, x_k)\right) + \hat{\nabla}_z f_k(z_k, x_k)\right\|^2\right] \\ &\leq \underbrace{2\,\mathbb{E}_k\left[\left\|\frac{1}{\bar{h}_k}\left(\hat{\nabla}_z g_k(z_k+\bar{h}_k v_k, x_k) - \hat{\nabla}_z g_k(z_k, x_k)\right)\right\|^2\right]}_{A} + \underbrace{2\,\mathbb{E}_k\left[\left\|\hat{\nabla}_z f_k(z_k, x_k)\right\|^2\right]}_{B}.\end{aligned}$$

We bound the B term by Lemma D.9,

$$\begin{aligned}\mathbb{E}_k\left[\left\|\hat{\nabla}_z f_k(z_k, x_k)\right\|^2\right] &\leq \underbrace{2\left(1+\frac{p+2}{b_2\ell_2}\right)}_{B_1}\left\|\nabla_z F(z_k, x_k)\right\|^2 \\ &\quad + \underbrace{\left(\frac{(p+6)^3}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(p+3)^3\right)L_{1,F}^2}_{B_2} h_k^2 + \underbrace{2\left(\frac{p+2}{\ell_2}+3\right)\frac{\sigma_{1,F}^2}{b_2}}_{B_3}.\end{aligned}$$

We bound the A term by Lemma D.16. Let $C_1 = 4\left(1+\frac{3(p+2)}{b_1\ell_1}\right)$. Then,

$$\begin{aligned}\frac{2}{\bar{h}_k^2}\mathbb{E}_k[\|\hat{\nabla}_z g_k(z_k+\bar{h}_k v_k, x_k) - \hat{\nabla}_z g_k(z_k, x_k)\|^2] &\leq C_1 L_{1,G}^2\|v_k - v^*(x_k)\|^2 + C_1\frac{L_{1,G}^2 L_{0,F}^2}{\mu_G^2} \\ &\quad + \frac{3L_{1,G}^2(p+6)^3}{b_1\ell_1}\frac{h_k^2}{\bar{h}_k^2}.\end{aligned}$$

Thus, by these bounds, we have

$$\begin{aligned}\mathbb{E}_k\left[\left\|\hat{D}_v^k\right\|^2\right] &\leq 2C_1 L_{1,G}^2\|v_k - v^*(x_k)\|^2 + 2C_1\frac{L_{1,G}^2 L_{0,F}^2}{\mu_G^2} \\ &\quad + \frac{6L_{1,G}^2(p+6)^3}{b_1\ell_1}\frac{h_k^2}{\bar{h}_k^2} + 2B_1\left\|\nabla_z F(z_k, x_k)\right\|^2 \\ &\quad + 2B_2 h_k^2 + 2B_3.\end{aligned}$$

Since F is $L_{0,F}$-Lipschitz continuous, we get the first claim

$$\begin{aligned}\mathbb{E}_k\left[\left\|\hat{D}_v^k\right\|^2\right] &\leq 2C_1 L_{1,G}^2\|v_k - v^*(x_k)\|^2 + 2C_1\frac{L_{1,G}^2 L_{0,F}^2}{\mu_G^2} \\ &\quad + \frac{6L_{1,G}^2(p+6)^3}{b_1\ell_1}\frac{h_k^2}{\bar{h}_k^2} + 2B_1 L_{0,F}^2 + 2B_2 h_k^2 + 2B_3.\end{aligned}$$

The proof of the last inequality follows the same line.

$$\begin{aligned}\mathbb{E}_k\left[\left\|\hat{D}_x^k\right\|^2\right] &= \mathbb{E}_k\left[\left\|\frac{1}{\bar{h}_k}\left(\hat{\nabla}_x g_k(z_k+\bar{h}_k v_k, x_k) - \hat{\nabla}_x g_k(z_k, x_k)\right) + \hat{\nabla}_x f_k(z_k, x_k)\right\|^2\right] \\ &\leq \underbrace{2\,\mathbb{E}_k\left[\left\|\frac{1}{\bar{h}_k}\left(\hat{\nabla}_x g_k(z_k+\bar{h}_k v_k, x_k) - \hat{\nabla}_x g_k(z_k, x_k)\right)\right\|^2\right]}_{A} + \underbrace{2\,\mathbb{E}_k\left[\left\|\hat{\nabla}_x f_k(z_k, x_k)\right\|^2\right]}_{B}.\end{aligned}$$

We bound the B term by Lemma D.9,

$$\mathbb{E}_k\left[\left\|\hat{\nabla}_x f_k(z_k,x_k)\right\|^2\right] \leq \underbrace{2\left(1+\frac{d+2}{b_2\ell_2}\right)}_{B_4}\|\nabla_x F(z_k,x_k)\|^2 + \underbrace{\left(\frac{(d+6)^3}{2b_2\ell_2}+\left(\frac{3}{b_2}+1\right)(d+3)^3\right)L_{1,F}^2 h_k^2}_{B_5} + \underbrace{2\left(\frac{d+2}{b_2\ell_2}+\frac{3}{b_2}\right)\sigma_{1,F}^2}_{B_6}.$$

We bound the A term by Lemma D.16. Let $C_2 = 4\left(1+\frac{3(d+2)}{b_1\ell_1}\right)$. Then,

$$\frac{2}{\bar{h}_k^2}\mathbb{E}_k[\|\hat{\nabla}_x g_k(z_k+\bar{h}_k v_k,x_k)-\hat{\nabla}_x g_k(z_k,x_k)\|^2] \leq C_2 L_{1,G}^2\|v_k-v^*(x_k)\|^2 + C_2\frac{L_{1,G}^2 L_{0,F}^2}{\mu_G^2} + \frac{3L_{1,G}^2(d+6)^3}{b_1\ell_1}\frac{h_k^2}{\bar{h}_k^2}.$$

Thus, by these bounds, we have

$$\mathbb{E}_k\left[\left\|\hat{D}_x^k\right\|^2\right] \leq 2C_2 L_{1,G}^2\|v_k-v^*(x_k)\|^2 + 2C_2\frac{L_{1,G}^2 L_{0,F}^2}{\mu_G^2} + \frac{6L_{1,G}^2(d+6)^3}{b_1\ell_1}\frac{h_k^2}{\bar{h}_k^2} + 2B_4\|\nabla_z F(z_k,x_k)\|^2 + 2B_5 h_k^2 + 2B_6.$$

Since F is $L_{0,F}$-Lipschitz continuous, we get the first claim

$$\mathbb{E}_k\left[\left\|\hat{D}_x^k\right\|^2\right] \leq 2C_2 L_{1,G}^2\|v_k-v^*(x_k)\|^2 + 2C_2\frac{L_{1,G}^2 L_{0,F}^2}{\mu_G^2} + \frac{6L_{1,G}^2(d+6)^3}{b_1\ell_1}\frac{h_k^2}{\bar{h}_k^2} + 2B_4 L_{0,F}^2 + 2B_5 h_k^2 + 2B_6.$$

□

In the next lemma, we bound the error of the sequence $(v_k)_{k\in\mathbb{N}}$ generated by Algorithm 2 relative to the corresponding optimal solutions $v^*(x_k)$ at each iterate x_k. We do not need to bound the error of the sequence $(z_k)_{k\in\mathbb{N}}$, since it differs from that of ZOBA only in the construction of the gradient estimator (as no Hessian is involved in the z_k update of ZOBA). Indeed, in Algorithm 2 this surrogate is constructed using forward finite differences instead of central differences. As shown in Lemma D.9, this change yields the same bound on the surrogate norm, analogously to the bound obtained for the gradient surrogate of the outer objective f in ZOBA.

Lemma D.18 (Bound on v_k iteration)**.** *Let Assumptions 3.2 and 3.3 hold. Let $(v_k)_{k\in\mathbb{N}}$ be the sequence generated in Algorithm 2 and let $\bar{h}_k = \begin{cases}\frac{\hat{h}_k}{\|v_k\|} & \text{if} \quad \|v_k\|\neq 0\\ \hat{h}_k\end{cases}$ with $\hat{h}_k \leq \frac{\mu_G}{\sqrt{8}L_{2,G}}$. Let $\omega_{v,1} = \left(L_{1,F}+\frac{L_{0,F}L_{2,G}}{\mu_G}\right)$, $\omega_{v,2} = \frac{16L_{2,G}^2}{9}\left((d+5)^{5/2}+(d+3)^{3/2}\right)^2$ and $\omega_k^v = \left(1+\frac{2}{\mu_G\rho_k}\right)$.*

Let $h_k \leq \frac{\mu_G}{\sqrt{16\omega_{v,2}}}$. Then, for every $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}_k\left[\|v_{k+1} - v^*(x_{k+1})\|^2\right] \leq &\left(1 - \frac{\mu_G}{2}\rho_k\right)\|v_k - v^*(x_k)\|^2 + 4\frac{\omega_{v,1}^2}{\mu_G}\rho_k\|z_k - z^*(x_k)\|^2 \\ &+ 2\rho_k^2\mathbb{E}_k[\|\hat{D}_v^k\|^2] + 2\omega_k^v L_*^2\gamma_k^2\mathbb{E}_k[\|\hat{D}_x^k\|^2] \\ &+ 2\frac{L_{1,G}^2(d+3)^2}{\mu_G}\rho_k h_k^2 + 4\frac{\omega_{v,2}L_{0,F}^2}{\mu_G^3}\rho_k h_k^2 \\ &+ \frac{8}{\mu_G}\frac{L_{2,G}^2}{4}\frac{L_{0,F}^2}{\mu_G^2}\rho_k\hat{h}_k^2.\end{aligned}$$

Proof. The proof follows the same line of Lemma D.13. Adding and subtracting $v^*(x_k)$, we have

$$\begin{aligned}\|v_{k+1} - v^*(x_{k+1})\|^2 = &\underbrace{\|v_{k+1} - v^*(x_k)\|^2}_{A} + \underbrace{\|v^*(x_{k+1}) - v^*(x_k)\|^2}_{B} \\ &\underbrace{-2\left\langle v_{k+1} - v^*(x_k), v^*(x_{k+1}) - v^*(x_k)\right\rangle}_{C}.\end{aligned} \tag{48}$$

We start by bounding A. By Algorithm 2 and taking the conditional expectation

$$\begin{aligned}\mathbb{E}_k[\|v_{k+1} - v^*(x_k)\|^2] = &\|v_k - v^*(x_k)\|^2 + \rho_k^2\mathbb{E}_k[\|\hat{D}_v^k\|^2] \\ &- 2\rho_k\left\langle\frac{1}{\bar{h}_k}(\nabla_z G_{h_k}(z_k + \bar{h}_k v_k, x_k) - \nabla_z G(z_k, x_k)) + \nabla_z F_{h_k}(z_k, x_k), v_k - v^*(x_k)\right\rangle.\end{aligned}$$

Let $\tilde{D}_v^k = \nabla_{zz}^2 G_{h_k}(z_k, x_k)v_k + \nabla_z F_{h_k}(z_k, x_k)$. Adding and subtracting $\nabla_{zz} G_{h_k}(z_k, x_k)v_k$, we get

$$\begin{aligned}\mathbb{E}_k[\|v_{k+1} - v^*(x_k)\|^2] = &\|v_k - v^*(x_k)\|^2 + \rho_k^2\mathbb{E}_k[\|\hat{D}_v^k\|^2] \\ &- 2\rho_k\left\langle\tilde{D}_v^k, v_k - v^*(x_k)\right\rangle \\ &- 2\rho_k\left\langle\frac{1}{\bar{h}_k}(\nabla_z G_{h_k}(z_k + \bar{h}_k v_k, x_k) - \nabla_z G(z_k, x_k)) - \nabla_{zz}^2 G(z_k, x_k)v_k, v_k - v^*(x_k)\right\rangle.\end{aligned}$$

Let $\omega_{v,1} = \left(L_{1,F} + \frac{L_{0,F}L_{2,G}}{\mu_G}\right)$, $\omega_{v,2} = \frac{16L_{2,G}^2}{9}\left((d+5)^{5/2} + (d+3)^{3/2}\right)^2$. For $h_k^2 \leq \frac{\mu_G^2}{16\omega_{v,2}}$, following the same line of the proof of Lemma D.13, we get

$$\begin{aligned}\mathbb{E}_k[\|v_{k+1} - v^*(x_k)\|^2] = &\left(1 - \frac{5\mu_G}{4}\rho_k\right)\|v_k - v^*(x_k)\|^2 \\ &+ \rho_k^2\mathbb{E}_k[\|\hat{D}_v^k\|^2] + 4\frac{\omega_{v,1}^2}{\mu_G}\rho_k\|z_k - z^*(x_k)\|^2 \\ &+ 2\frac{L_{1,G}^2(d+3)^2}{\mu_G}\rho_k h_k^2 + 4\frac{\omega_{v,2}L_{0,F}^2}{\mu_G^3}\rho_k h_k^2 \\ &- 2\rho_k\left\langle\frac{1}{\bar{h}_k}(\nabla_z G_{h_k}(z_k + \bar{h}_k v_k, x_k) - \nabla_z G(z_k, x_k)) - \nabla_{zz}^2 G(z_k, x_k)v_k, v_k - v^*(x_k)\right\rangle.\end{aligned}$$

Let

$$\Delta_k = \frac{1}{\bar{h}_k}(\nabla_z G_{h_k}(z_k + \bar{h}_k v_k, x_k) - \nabla_z G(z_k, x_k)) - \nabla_{zz}^2 G(z_k, x_k)v_k.$$

By Cauchy-Schwartz and Young's inequality with parameter $\frac{\mu_G}{4}$, we have

$$-2\rho_k\left\langle\Delta_k, v_k - v^*(x_k)\right\rangle \leq \frac{4\rho_k}{\mu_G}\|\Delta_k\|^2 + \frac{\mu_G}{4}\rho_k\|v_k - v^*(x_k)\|^2.$$

By Lemma D.15 and by the choice,

$$\bar{h}_k = \begin{cases}\frac{\hat{h}_k}{\|v_k\|} & \|v_k\| \neq 0 \\ \hat{h}_k\end{cases},$$

we have

$$-2\rho_k \langle \Delta_k, v_k - v^*(x_k)\rangle \leq \frac{4\rho_k}{\mu_G}\frac{L_{2,G}^2}{4}\bar{h}_k^2\|v_k\|^4 + \frac{\mu_G}{4}\rho_k\|v_k - v^*(x_k)\|^2$$
$$= \frac{4\rho_k}{\mu_G}\frac{L_{2,G}^2}{4}\hat{h}_k^2\|v_k\|^2 + \frac{\mu_G}{4}\rho_k\|v_k - v^*(x_k)\|^2.$$

Adding and subtracting $v^*(x_k)$ and by Lemma D.3,

$$-2\rho_k \langle \Delta_k, v_k - v^*(x_k)\rangle \leq \frac{8\rho_k}{\mu_G}\frac{L_{2,G}^2}{4}\hat{h}_k^2\|v_k - v^*(x_k)\|^2 + \frac{\mu_G}{4}\rho_k\|v_k - v^*(x_k)\|^2$$
$$+ \frac{8\rho_k}{\mu_G}\frac{L_{2,G}^2}{4}\hat{h}_k^2\|v^*(x_k)\|^2$$
$$\leq \left(\frac{8}{\mu_G}\frac{L_{2,G}^2}{4}\hat{h}_k^2 + \frac{\mu_G}{4}\right)\rho_k\|v_k - v^*(x_k)\|^2$$
$$+ \frac{8\rho_k}{\mu_G}\frac{L_{2,G}^2}{4}\frac{L_{0,F}^2}{\mu_G^2}\hat{h}_k^2.$$

Since $\hat{h}_k \leq \frac{\mu_G}{\sqrt{8}L_{2,G}} = \sqrt{\frac{\mu_G}{4}\frac{\mu_G}{8}\frac{4}{L_{2,G}^2}}$,

$$-2\rho_k \langle \Delta_k, v_k - v^*(x_k)\rangle \leq \frac{\mu_G}{2}\rho_k\|v_k - v^*(x_k)\|^2 + \frac{8}{\mu_G}\frac{L_{2,G}^2}{4}\frac{L_{0,F}^2}{\mu_G^2}\rho_k\hat{h}_k^2.$$

Thus, we conclude the bound of term A obtaining

$$\mathbb{E}_k[\|v_{k+1} - v^*(x_k)\|^2] \leq \left(1 - \frac{5\mu_G}{4}\rho_k + \frac{\mu_G}{2}\rho_k\right)\|v_k - v^*(x_k)\|^2$$
$$+ \rho_k^2\mathbb{E}_k[\|\hat{D}_v^k\|^2] + 4\frac{\omega_{v,1}^2}{\mu_G}\rho_k\|z_k - z^*(x_k)\|^2$$
$$+ 2\frac{L_{1,G}^2(d+3)^2}{\mu_G}\rho_k h_k^2 + 4\frac{\omega_{v,2}L_{0,F}^2}{\mu_G^3}\rho_k h_k^2$$
$$+ \frac{8}{\mu_G}\frac{L_{2,G}^2}{4}\frac{L_{0,F}^2}{\mu_G^2}\rho_k\hat{h}_k^2.$$

We bound the B term by Lemma D.2 and Algorithm 2,

$$\|v^*(x_{k+1}) - v^*(x_k)\|^2 \leq L_*^2\|x_{k+1} - x_k\|^2 = L_*^2\gamma_k^2\|\hat{D}_x^k\|^2.$$

The bound on term C follows the same line of the analogous bound in the proof of Lemma D.13,

$$-2\langle v_{k+1} - v^*(x_k), v^*(x_{k+1}) - v^*(x_k)\rangle \leq \frac{\mu_G}{4}\rho_k\|v_k - v^*(x_k)\|^2 + \frac{4}{\mu_G\rho_k}L_*^2\gamma_k^2\|\hat{D}_x^k\|^2$$
$$+ \rho_k^2\|\hat{D}_v^k\|^2 + L_*^2\gamma_k^2\|\hat{D}_x^k\|^2.$$

Let $\omega_k^v = \left(1 + \frac{2}{\mu_G\rho_k}\right)$. Restarting from eq. 48, taking the conditional expectation and using the bounds on the terms A, B and C, we get the claim

$$\mathbb{E}_k[\|v_{k+1} - v^*(x_{k+1})\|^2] \leq \left(1 - \frac{\mu_G}{2}\rho_k\right)\|v_k - v^*(x_k)\|^2 + 4\frac{\omega_{v,1}^2}{\mu_G}\rho_k\|z_k - z^*(x_k)\|^2$$
$$+ 2\rho_k^2\mathbb{E}_k[\|\hat{D}_v^k\|^2] + 2\omega_k^v L_*^2\gamma_k^2\mathbb{E}_k[\|\hat{D}_x^k\|^2]$$
$$+ 2\frac{L_{1,G}^2(d+3)^2}{\mu_G}\rho_k h_k^2 + 4\frac{\omega_{v,2}L_{0,F}^2}{\mu_G^3}\rho_k h_k^2$$
$$+ \frac{8}{\mu_G}\frac{L_{2,G}^2}{4}\frac{L_{0,F}^2}{\mu_G^2}\rho_k\hat{h}_k^2.$$

□

In the following lemma, we derive the bound on the value function Ψ along the sequence $(x_k)_{k\in\mathbb{N}}$ generated by Algorithm 2.

Lemma D.19 (Function value decrease)**.** *Let Assumptions 3.2, 3.1, 3.3 hold. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 2. Then for every $k \in \mathbb{N}$,*

$$\begin{aligned}\mathbb{E}_k[\Psi(x_{k+1})] \le \Psi(x_k) - \frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 + \frac{L_\Psi}{2}\gamma_k^2\mathbb{E}_k[\|\hat{D}_x^k\|^2] \\ + 4C_{1,\Psi}^2\gamma_k\|z_k - z^*(x_k)\|^2 + C_{4,\Psi}\gamma_k\|v_k - v^*(x_k)\|^2 \\ + 2C_{2,\Psi}^2\gamma_k h_k^2 + C_{3,\Psi}\hat{h}_k^2\gamma_k,\end{aligned}$$

where

$$\begin{aligned}C_{1,\Psi} &= L_{1,F} + \frac{L_{0,F}L_{1,G}}{\mu_G}, \\ C_{2,\Psi} &= \frac{2L_{2,G}L_{0,F}}{3\mu_G}\left((d+5)^{5/2} + (d+3)^{3/2} + (d+1)(p+3)^{3/2}\right) + \frac{L_{1,F}d\sqrt{p}}{2} \\ C_{3,\Psi} &= \frac{L_{2,G}^2L_{0,F}^2}{2\mu_G^2} \quad \textit{and} \quad C_{4,\Psi} = 4L_{1,G}^2 + \frac{\mu_G^2}{16}.\end{aligned}$$

Proof. The proof follows the same line of Lemma D.14. By Lemma D.2, Ψ is L_Ψ-smooth. Thus, by the Descent Lemma (Polyak, 1987) and Algorithm 2, we have

$$\Psi(x_{k+1}) \le \Psi(x_k) - \gamma_k\left\langle\nabla\Psi(x_k), \hat{D}_x^k\right\rangle + \frac{L_\Psi}{2}\gamma_k^2\|\hat{D}_x^k\|^2.$$

Taking the conditional expectation,

$$\begin{aligned}\mathbb{E}_k[\Psi(x_{k+1})] &\le \Psi(x_k) \\ &\quad - \gamma_k\left\langle\nabla\Psi(x_k), \frac{1}{\bar{h}_k}(\nabla_x G_{h_k}(z_k + \bar{h}_k v_k, x_k) - \nabla_x G_{h_k}(z_k, x_k)) + \nabla_x F_{h_k}(z_k, x_k)\right\rangle \\ &\quad + \frac{L_\Psi}{2}\gamma_k^2\mathbb{E}_k[\|\hat{D}_x^k\|^2].\end{aligned}$$

By $\langle a, b\rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a-b\|^2)$, we get

$$\begin{aligned}\mathbb{E}_k[\Psi(x_{k+1})] &\le \Psi(x_k) \\ &\quad - \frac{\gamma_k}{2}\Bigg(\|\nabla\Psi(x_k)\|^2 \\ &\quad + \|\frac{1}{\bar{h}_k}(\nabla_x G_{h_k}(z_k + \bar{h}_k v_k, x_k) - \nabla_x G_{h_k}(z_k, x_k)) + \nabla_x F_{h_k}(z_k, x_k)\|^2 \\ &\quad - \|\frac{1}{\bar{h}_k}(\nabla_x G_{h_k}(z_k + \bar{h}_k v_k, x_k) - \nabla_x G_{h_k}(z_k, x_k)) + \nabla_x F_{h_k}(z_k, x_k) - \nabla\Psi(x_k)\|^2\Bigg) \\ &\quad + \frac{L_\Psi}{2}\gamma_k^2\mathbb{E}_k[\|\hat{D}_x^k\|^2] \\ &\le \Psi(x_k) - \frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 + \frac{L_\Psi}{2}\gamma_k^2\mathbb{E}_k[\|\hat{D}_x^k\|^2] \\ &\quad + \underbrace{\frac{\gamma_k}{2}\|\frac{1}{\bar{h}_k}(\nabla_x G_{h_k}(z_k + \bar{h}_k v_k, x_k) - \nabla_x G_{h_k}(z_k, x_k)) + \nabla_x F_{h_k}(z_k, x_k) - \nabla\Psi(x_k)\|^2}_{A}.\end{aligned} \tag{49}$$

Now we bound A. Let

$$\begin{aligned}\bar{\Delta}_k &= \frac{1}{\bar{h}_k}(\nabla_x G_{h_k}(z_k + \bar{h}_k v_k, x_k) - \nabla_x G_{h_k}(z_k, x_k)) + \nabla_x F_{h_k}(z_k, x_k), \\ \Delta_k &= \frac{1}{\bar{h}_k}(\nabla_x G_{h_k}(z_k + \bar{h}_k v_k, x_k) - \nabla_x G_{h_k}(z_k, x_k)) - \nabla_{xz}^2 G_{h_k}(z_k, x_k)v_k, \\ \tilde{D}_x^k &= \nabla_{xz}^2 G_{h_k}(z_k, x_k)v_k + \nabla_x F_{h_k}(z_k, x_k)\end{aligned}$$

Adding and subtracting $\nabla^2_{xz} G_{h_k}(z_k, x_k) v_k$ and by $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$,

$$\frac{\gamma_k}{2}\|\bar{\Delta}_k - \nabla\Psi(x_k)\|^2 \le \gamma_k\|\Delta_k\|^2 + \gamma_k\|\tilde{D}^k_x - \nabla\Psi(x_k)\|^2.$$

By Lemma D.15 and Lemma D.8 with $h_1 = \eta_1 = \eta_2 = h_k$, we have

$$\begin{aligned}
\frac{\gamma_k}{2}\|\bar{\Delta}_k - \nabla\Psi(x_k)\|^2 \le{}& \frac{L^2_{2,G}}{4}\bar{h}_k^2\gamma_k\|v_k\|^4 \\
&+ \gamma_k\Bigg(\underbrace{\left(L_{1,F} + \frac{L_{0,F}L_{1,G}}{\mu_G}\right)}_{C_{1,\Psi}}\|z_k - z^*(x_k)\| + L_{1,G}\|v_k - v^*(x_k)\| \\
&+ \underbrace{\left(\frac{L_{0,F}}{\mu_G}\frac{2L_{2,G}}{3}\left((d+5)^{5/2} + (d+3)^{3/2} + (d+1)(p+3)^{3/2}\right) + \frac{L_{1,F}d\sqrt{p}}{2}\right)}_{=:C_{2,\Psi}} h_k\Bigg)^2 \\
\le{}& \frac{L^2_{2,G}}{4}\bar{h}_k^2\gamma_k\|v_k\|^4 \\
&+ 4C^2_{1,\Psi}\gamma_k\|z_k - z^*(x_k)\|^2 + 4L^2_{1,G}\gamma_k\|v_k - v^*(x_k)\|^2 + 2C^2_{2,\Psi}\gamma_k h_k^2.
\end{aligned}$$

By the choice

$$\bar{h}_k = \begin{cases} \frac{\hat{h}_k}{\|v_k\|} & \|v_k\| \neq 0 \\ \hat{h}_k \end{cases},$$

we have

$$\begin{aligned}
\frac{\gamma_k}{2}\|\bar{\Delta}_k - \nabla\Psi(x_k)\|^2 \le{}& \frac{L^2_{2,G}}{4}\hat{h}_k^2\gamma_k\|v_k\|^2 \\
&+ 4C^2_{1,\Psi}\gamma_k\|z_k - z^*(x_k)\|^2 + 4L^2_{1,G}\gamma_k\|v_k - v^*(x_k)\|^2 + 2C^2_{2,\Psi}\gamma_k h_k^2.
\end{aligned}$$

Adding and subtracting $v^*(x_k)$ and by Lemma D.3, we get

$$\begin{aligned}
\frac{\gamma_k}{2}\|\bar{\Delta}_k - \nabla\Psi(x_k)\|^2 \le{}& \underbrace{\frac{L^2_{2,G}}{2}\frac{L^2_{0,F}}{\mu^2_G}}_{=:C_{3,\Psi}}\hat{h}_k^2\gamma_k \\
&+ 4C^2_{1,\Psi}\gamma_k\|z_k - z^*(x_k)\|^2 \\
&+ \left(4L^2_{1,G} + \frac{L^2_{2,G}}{2}\hat{h}_k^2\right)\gamma_k\|v_k - v^*(x_k)\|^2 + 2C^2_{2,\Psi}\gamma_k h_k^2.
\end{aligned}$$

Since $\hat{h}_k \le \frac{\mu_G}{\sqrt{8}L_{2,G}}$, we have

$$\begin{aligned}
\frac{\gamma_k}{2}\|\bar{\Delta}_k - \nabla\Psi(x_k)\|^2 \le{}& 4C^2_{1,\Psi}\gamma_k\|z_k - z^*(x_k)\|^2 + \underbrace{\left(4L^2_{1,G} + \frac{\mu^2_G}{16}\right)}_{=:C_{4,\Psi}}\gamma_k\|v_k - v^*(x_k)\|^2 \\
&+ 2C^2_{2,\Psi}\gamma_k h_k^2 + C_{3,\Psi}\hat{h}_k^2\gamma_k.
\end{aligned}$$

Using this inequality in eq. 49, we get

$$\begin{aligned}
\mathbb{E}_k[\Psi(x_{k+1})] \le{}& \Psi(x_k) - \frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 + \frac{L_\Psi}{2}\gamma_k^2\mathbb{E}_k[\|\hat{D}^k_x\|^2] \\
&+ 4C^2_{1,\Psi}\gamma_k\|z_k - z^*(x_k)\|^2 + C_{4,\Psi}\gamma_k\|v_k - v^*(x_k)\|^2 \\
&+ 2C^2_{2,\Psi}\gamma_k h_k^2 + C_{3,\Psi}\hat{h}_k^2\gamma_k.
\end{aligned}$$

□

E PROOFS OF MAIN RESULTS

E.1 PROOF OF THEOREM 3.4

Let $\delta_{z_k} = \|z_k - z^*(x_k)\|^2$ and $\delta_{v_k} = \|v_k - v^*(x_k)\|^2$. For $\phi_z^k, \phi_v^k > 0$ s.t. $\phi_v^{k+1} \leq \phi_v^k$ and $\phi_z^{k+1} \leq \phi_z^k$, define

$$L_k = \Psi(x_k) + \phi_z^k \delta_{z_k} + \phi_v^k \delta_{v_k}.$$

Consider

$$\begin{aligned} L_{k+1} - L_k &= \Psi(x_{k+1}) - \Psi(x_k) + \phi_z^{k+1}\delta_{z_{k+1}} - \phi_z^k \delta_{z_k} + \phi_v^{k+1}\delta_{v_{k+1}} - \phi_v^k \delta_{v_k} \\ &\leq \Psi(x_{k+1}) - \Psi(x_k) + \phi_z^k(\delta_{z_{k+1}} - \delta_{z_k}) + \phi_v^k(\delta_{v_{k+1}} - \delta_{v_k}). \end{aligned}$$

Taking the conditional expectation, we have

$$\mathbb{E}_k\left[L_{k+1}\right] - L_k \leq \mathbb{E}_k\left[\Psi(x_{k+1})\right] - \Psi(x_k) + \phi_z^k(\mathbb{E}_k\left[\delta_{z_{k+1}}\right] - \delta_{z_k}) + \phi_v^k(\mathbb{E}_k\left[\delta_{v_{k+1}}\right] - \delta_{v_k}).$$

By Lemma D.14, we get

$$\begin{aligned} \mathbb{E}_k\left[L_{k+1}\right] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 - \frac{\gamma_k}{2}\|\hat{D}_x^k\|^2 + \frac{L_\Psi}{2}\gamma_k^2\mathbb{E}_k\left[\|D_x^k\|^2\right] \\ & + 2\gamma_k\omega_{v,1}^2\delta_{z_k} + 2\gamma_k L_{1,G}^2\delta_{v_k} \\ & + \phi_z^k(\mathbb{E}_k\left[\delta_{z_{k+1}}\right] - \delta_{z_k}) + \phi_v^k(\mathbb{E}_k\left[\delta_{v_{k+1}}\right] - \delta_{v_k}) \\ & + C_1^2\gamma_k h_k^2. \end{aligned}$$

By Lemma D.13, we have

$$\begin{aligned} \mathbb{E}_k\left[L_{k+1}\right] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 - \frac{\gamma_k}{2}\|\hat{D}_x^k\|^2 \\ & + 2\phi_v^k\rho_k^2\mathbb{E}_k\left[\|D_v^k\|^2\right] + 2\phi_z^k\rho_k^2\mathbb{E}_k\left[\left\|D_z^k\right\|^2\right] \\ & - \left(\mu_G\phi_v^k\rho_k - 2L_{1,G}^2\gamma_k\right)\delta_{v_k} \\ & - \left(\frac{\mu_G}{2}\phi_z^k\rho_k - 4\frac{\omega_{v,1}^2}{\mu_G}\phi_v^k\rho_k - 2\omega_{v,1}^2\gamma_k\right)\delta_{z_k} \\ & + \left(\frac{L_\Psi}{2} + 2w_k^z L_*^2\phi_z^k + 2\omega_k^v L_*^2\phi_v^k\right)\gamma_k^2\mathbb{E}_k\left[\|D_x^k\|^2\right] \\ & + C_1^2\gamma_k h_k^2 \\ & + \left(\frac{L_{1,G}^2}{4\mu_G}(d+3)^3\phi_z^k + 2\frac{L_{1,G}^2(d+3)^3}{\mu_G}\phi_v^k + 4\frac{\omega_{v,2}L_{0,F}^2}{\mu_G^3}\phi_v^k\right)\rho_k h_k^2. \end{aligned}$$

Let $\omega_3 = 2\left(2 + \frac{p+2}{b_1\ell_1}\right) L_{1,G}^2$ and $\omega_{dx}^k = \left(\frac{L_\Psi}{2} + 2w_k^z L_*^2 \phi_z^k + 2\omega_k^v L_*^2 \phi_v^k\right)$. By Lemma D.12, we get

$$
\begin{aligned}
\mathbb{E}_k\left[L_{k+1}\right] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 - \frac{\gamma_k}{2}\|\hat{D}_x^k\|^2 \\
& - \left(\mu_G \phi_v^k \rho_k - 2C_v^k \phi_v^k \rho_k^2 - 2L_{1,G}^2\gamma_k - \omega_{dx}^k 2C_x^k \gamma_k^2\right)\delta_{v_k} \\
& - \left(\frac{\mu_G}{2}\phi_z^k \rho_k - 4\frac{\omega_{v,1}^2}{\mu_G}\phi_v^k \rho_k - 2\omega_3 \phi_z^k \rho_k^2 - 2\omega_{v,1}^2\gamma_k\right)\delta_{z_k} \\
& + C_1^2 \gamma_k h_k^2 \\
& + \left(\frac{L_{1,G}^2}{4\mu_G}(d+3)^3\phi_z^k + 2\frac{L_{1,G}^2(d+3)^3}{\mu_G}\phi_v^k + 4\frac{\omega_{v,2}L_{0,F}^2}{\mu_G^3}\phi_v^k\right)\rho_k h_k^2 \\
& + 2\left(\frac{(p+6)^3}{2b_1\ell_1} + \left(\frac{3}{b_1}+1\right)(p+3)^3\right) L_{1,G}^2 \phi_z^k \rho_k^2 h_k^2 \\
& + 4\left(3 + \frac{p+2}{\ell_1}\right)\frac{\sigma_{1,G}^2}{b_1}\phi_z^k \rho_k^2 \\
& + 2\phi_v^k \rho_k^2 \Bigg(C_v^k \frac{L_{0,F}^2}{\mu_G^2} + 4\left(1 + \frac{p+2}{b_2\ell_2}\right) L_{0,F}^2 + 4\left(\frac{p+2}{\ell_2} + 3\right)\frac{\sigma_{1,F}^2}{b_2} \\
& + 2\left(\frac{(p+6)^3}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(p+3)^3\right) L_{1,F}^2 h_k^2\Bigg) \\
& + \omega_{dx}^k \gamma_k^2 \Bigg(2C_x^k \frac{L_{0,F}^2}{\mu_G^2} + 2\left(\frac{(d+6)^3}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(d+3)^3\right) L_{1,F}^2 h_k^2 \\
& + 4\left(1 + \frac{d+2}{b_2\ell_2}\right) L_{0,F}^2 + 4\left(\frac{d+2}{b_2\ell_2} + \frac{3}{b_2}\right)\sigma_{1,F}^2\Bigg).
\end{aligned}
$$

Let

$$
\begin{aligned}
C_2^k &:= \left(\frac{L_{1,G}^2}{4\mu_G}(d+3)^3\phi_z^k + 2\frac{L_{1,G}^2(d+3)^3}{\mu_G}\phi_v^k + 4\frac{\omega_{v,2}L_{0,F}^2}{\mu_G^3}\phi_v^k\right) \\
C_3 &:= 4\left(3 + \frac{p+2}{\ell_1}\right)\frac{\sigma_{1,G}^2}{b_1} \\
C_4 &:= 2\left(\frac{(p+6)^3}{2b_1\ell_1} + \left(\frac{3}{b_1}+1\right)(p+3)^3\right) L_{1,G}^2 \\
C_5^k(h_k) &:= \left(2C_x^k \frac{L_{0,F}^2}{\mu_G^2} + 2\left(\frac{(d+6)^3}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(d+3)^3\right) L_{1,F}^2 h_k^2\right. \\
&\quad \left. + 4\left(1 + \frac{d+2}{b_2\ell_2}\right) L_{0,F}^2 + 4\left(\frac{d+2}{b_2\ell_2} + \frac{3}{b_2}\right)\sigma_{1,F}^2\right) \\
C_6^k(h_k) &:= \left(C_v^k \frac{L_{0,F}^2}{\mu_G^2} + 4\left(1 + \frac{p+2}{b_2\ell_2}\right) L_{0,F}^2 + 4\left(\frac{p+2}{\ell_2} + 3\right)\frac{\sigma_{1,F}^2}{b_2}\right. \\
&\quad \left. + 2\left(\frac{(p+6)^3}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(p+3)^3\right) L_{1,F}^2 h_k^2\right).
\end{aligned}
$$

Thus, we have

$$\begin{aligned}\mathbb{E}_k\left[L_{k+1}\right] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 - \frac{\gamma_k}{2}\|\hat{D}_x^k\|^2 \\ & - \left(\mu_G\phi_v^k\rho_k - 2C_v^k\phi_v^k\rho_k^2 - 2L_{1,G}^2\gamma_k - \omega_{dx}^k 2C_x^k\gamma_k^2\right)\delta_{v_k} \\ & - \left(\frac{\mu_G}{2}\phi_z^k\rho_k - 4\frac{\omega_{v,1}^2}{\mu_G}\phi_v^k\rho_k - 2\omega_3\phi_z^k\rho_k^2 - 2\omega_{v,1}^2\gamma_k\right)\delta_{z_k} \\ & + C_1^2\gamma_k h_k^2 + C_2^k\rho_k h_k^2 + C_4\phi_z^k\rho_k^2 h_k^2 \\ & + C_3\phi_z^k\rho_k^2 + 2\phi_v^k C_6^k(h_k)\rho_k^2 + \omega_{dx}^k C_5^k(h_k)\gamma_k^2.\end{aligned}$$

Let $\bar{C}_k = C_1^2\gamma_k h_k^2 + C_2^k\rho_k h_k^2 + C_4\phi_z^k\rho_k^2 h_k^2 + C_3\phi_z^k\rho_k^2 + 2\phi_v^k C_6^k(h_k)\rho_k^2 + \omega_{dx}^k C_5^k(h_k)\gamma_k^2$. Observing that $-\frac{\gamma_k}{2}\|\hat{D}_x^k\|^2 \leq 0$, we have

$$\begin{aligned}\mathbb{E}_k\left[L_{k+1}\right] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 \\ & - \left(\mu_G\phi_v^k\rho_k - 2C_v^k\phi_v^k\rho_k^2 - 2L_{1,G}^2\gamma_k - \omega_{dx}^k 2C_x^k\gamma_k^2\right)\delta_{v_k} \\ & - \left(\frac{\mu_G}{2}\phi_z^k\rho_k - 4\frac{\omega_{v,1}^2}{\mu_G}\phi_v^k\rho_k - 2\omega_3\phi_z^k\rho_k^2 - 2\omega_{v,1}^2\gamma_k\right)\delta_{z_k} \\ & + \bar{C}_k.\end{aligned}$$

Choosing h_k such that $h_k^2 \leq \min\left(\frac{b_1\ell_1}{4L_{2,G}^2(p+16)^4}\left(1+\frac{15(p+6)^2p}{2b_1\ell_1}\right)L_{1,G}^2, \frac{\mu_G^2}{16\omega_{v,2}}, \frac{1}{16L_{2,G}^2((p+8)^4+p(d+12)^3)}\right)$, we have,

$$C_v^k = 8\left(1+\frac{15(p+6)^2p}{2b_1\ell_1}\right)L_{1,G}^2 =: \bar{C}_v.$$

Therefore, we get

$$\begin{aligned}\mathbb{E}_k\left[L_{k+1}\right] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 \\ & - \left(\mu_G\phi_v^k\rho_k - 2\bar{C}_v\phi_v^k\rho_k^2 - 2L_{1,G}^2\gamma_k - \omega_{dx}^k 2C_x^k\gamma_k^2\right)\delta_{v_k} \\ & - \left(\frac{\mu_G}{2}\phi_z^k\rho_k - 4\frac{\omega_{v,1}^2}{\mu_G}\phi_v^k\rho_k - 2\omega_3\phi_z^k\rho_k^2 - 2\omega_{v,1}^2\gamma_k\right)\delta_{z_k} \\ & + \bar{C}_k.\end{aligned}$$

Recalling that

$$\frac{1}{\omega_k^z} = \frac{\mu_G\rho_k}{\mu_G\rho_k+4} \quad \text{and} \quad \frac{1}{\omega_k^v} = \frac{\mu_G\rho_k}{\mu_G\rho_k+2}.$$

Let

$$\phi_z^k = \frac{\bar{\phi}_z}{\omega_k^z} \quad \text{and} \quad \phi_v^k = \frac{\bar{\phi}_v}{\omega_k^v}.$$

Notice that for ρ_k decreasing sequence or constant $\phi_z^{k+1} \leq \phi_z$ and $\phi_v^{k+1} \leq \phi_z$. Then, we have

$$\omega_{dx}^k = \frac{L_\Psi}{2} + 2L_*^2(\bar{\phi}_z + \bar{\phi}_v) =: \omega_{dx}.$$

Moreover, for $h_k^2 \leq \frac{1}{16L_{2,G}^2((p+8)^4+p(d+12)^3)}$, we get

$$C_x^k = 2\left(\frac{1}{b_1\ell_1}\left(\frac{1}{2}+6(p+4)(p+2)p+36(p+2)\min(p,d)+30p(d+2)dL_{1,G}^2\right)+L_{1,G}^2\right) =: \bar{C}_x.$$

Therefore, we have

$$\begin{aligned}\mathbb{E}_k\left[L_{k+1}\right] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 \\ & - \left(\frac{\mu_G}{\omega_k^v}\bar{\phi}_v\rho_k - \frac{2\bar{C}_v}{\omega_k^v}\bar{\phi}_v\rho_k^2 - 2L_{1,G}^2\gamma_k - \omega_{dx}2\bar{C}_x\gamma_k^2\right)\delta_{v_k} \\ & - \left(\frac{\mu_G}{2\omega_k^z}\bar{\phi}_z\rho_k - 4\frac{\omega_{v,1}^2}{\mu_G\omega_k^v}\bar{\phi}_v\rho_k - 2\frac{\omega_3}{\omega_k^z}\bar{\phi}_z\rho_k^2 - 2\omega_{v,1}^2\gamma_k\right)\delta_{z_k} \\ & + \bar{C}_k.\end{aligned}$$

Notice that since $\rho_k < \min(1, \frac{\mu_G}{2\bar{C}_v})$, we have

$$-\left(\frac{\mu_G}{2\omega_k^v}\bar{\phi}_v\rho_k - 2L_{1,G}^2\gamma_k - \omega_{dx}2\bar{C}_x\gamma_k^2\right) \leq -\Bigg(\underbrace{\frac{\mu_G^2}{2(\mu_G+2)}\bar{\phi}_v}_{A}\rho_k^2 - 2L_{1,G}^2\gamma_k - \omega_{dx}2\bar{C}_x\gamma_k^2\Bigg).$$

Thus for $\gamma_k < \frac{-L_{1,G}^2+\sqrt{L_{1,G}^4+2\omega_{dx}\bar{C}_xA\rho_k^2}}{2\omega_{dx}\bar{C}_x}$, we have

$$-\left(A\rho_k^2 - 2L_{1,G}^2\gamma_k - \omega_{dx}2\bar{C}_x\gamma_k^2\right) < 0.$$

Moreover, since for $\rho_k \leq \min(1, \frac{\mu_G}{2\bar{C}_v})$, we have

$$\begin{aligned}\frac{\mu_G\rho_k}{4} > \frac{1}{w_k^z} = \frac{\mu_G\rho_k}{\mu_G\rho_k+4} > \frac{\mu_G\rho_k}{\mu_G+4}, \\ \frac{\mu_G\rho_k}{2} > \frac{1}{w_k^v} = \frac{\mu_G\rho_k}{\mu_G\rho_k+2} > \frac{\mu_G\rho_k}{\mu_G+2}.\end{aligned}$$

Let $T_k = \frac{\mu_G}{2\omega_k^z}\bar{\phi}_z\rho_k - 4\frac{\omega_{v,1}^2}{\mu_G\omega_k^v}\bar{\phi}_v\rho_k - 2\frac{\omega_3}{\omega_k^z}\bar{\phi}_z\rho_k^2$. Thus, we get

$$\begin{aligned}T_k &\geq \frac{\mu_G^2}{2(\mu_G+4)}\bar{\phi}_z\rho_k^2 - 2\omega_{v,1}^2\bar{\phi}_v\rho_k^2 - \frac{\omega_3\mu_G}{2}\bar{\phi}_z\rho_k^3 \\ &= \left(\frac{\mu_G^2}{2(\mu_G+4)}\bar{\phi}_z - 2\omega_{v,1}^2\bar{\phi}_v - \frac{\omega_3\mu_G}{2}\bar{\phi}_z\rho_k\right)\rho_k^2.\end{aligned}$$

Setting $\bar{\phi}_v = \frac{\mu_G^2}{8(\mu_G+4)\omega_{v,1}^2}\bar{\phi}_z$, we have

$$\begin{aligned}T_k &\geq \left(\frac{\mu_G^2}{2(\mu_G+4)}\bar{\phi}_z - 2\omega_{v,1}^2\bar{\phi}_v - \frac{\omega_3\mu_G}{2}\bar{\phi}_z\rho_k\right)\rho_k^2 \\ &= \left(\frac{\mu_G^2}{4(\mu_G+4)}\bar{\phi}_z - \frac{\omega_3\mu_G}{2}\bar{\phi}_z\rho_k\right)\rho_k^2.\end{aligned}$$

Notice that for $\rho_k < \frac{\mu_G}{2\omega_3(\mu_G+4)}$,

$$\begin{aligned}T_k &> \left(\frac{\mu_G^2}{4(\mu_G+4)}\bar{\phi}_z - \frac{\omega_3\mu_G}{2}\bar{\phi}_z\rho_k\right)\rho_k^2 \\ &> \left(\frac{\mu_G^2}{4(\mu_G+4)}\bar{\phi}_z - \frac{\mu_G^2}{8(\mu_G+4)}\bar{\phi}_z\right)\rho_k^2 \\ &= \underbrace{\frac{\mu_G^2}{8(\mu_G+4)}\bar{\phi}_z\rho_k^2}_{=:\bar{T}_k} > 0.\end{aligned}$$

Thus, for $\gamma_k < \frac{\mu_G^2}{16\omega_{v,1}^2(\mu_G+4)}\bar{\phi}_z\rho_k^2$ we have

$$T_k - 2\omega_{v,1}^2\gamma_k > \bar{T}_k - 2\omega_{v,1}^2\gamma_k > 0.$$

This implies that

$$-\big(T_k - 2\omega_{v,1}^2\gamma_k\big) \leq 0.$$

Thus, under

$$\rho_k < \min\left(1, \frac{\mu_G}{2\omega_3(\mu_G+4)}, \frac{\mu_G}{2\bar{C}_v}\right)\gamma_k,$$
$$\gamma_k < \min\left(\frac{\mu_G^2}{16\omega_{v,1}^2(\mu_G+4)}\bar{\phi}_z\rho_k^2, \frac{-L_{1,G}^2 + \sqrt{L_{1,G}^4 + 2\omega_{dx}\bar{C}_x A\rho_k^2}}{2\omega_{dx}\bar{C}_x}\right),$$

and

$$h_k^2 \leq \min\left\{\frac{\left(2b_1\ell_1 + 15p(p+6)^2\right)L_{1,G}^2}{8L_{2,G}^2(p+16)^4}, \frac{\mu_G^2}{16\omega_{v,2}}, \frac{1}{16L_{2,G}^2\left((p+8)^4 + p(d+12)^3\right)}\right\}.$$

we obtain

$$\mathbb{E}_k\left[L_{k+1}\right] - L_k \leq -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 + \bar{C}_k.$$

For a more compact characterization of the condition on γ_k, note that, since $\rho_k < 1$,

$$\frac{-2L_{1,G}^2 + \sqrt{4L_{1,G}^4 + 8\omega_{dx}\bar{C}_x A\rho_k^2}}{4\omega_{dx}\bar{C}_x} = \frac{2A\rho_k^2}{\sqrt{4L_{1,G}^4 + 8\omega_{dx}\bar{C}_x A\rho_k^2} + 2L_{1,G}^2}$$
$$\geq \frac{2A}{\sqrt{4L_{1,G}^4 + 8\omega_{dx}\bar{C}_x A} + 2L_{1,G}^2}\rho_k^2.$$

Recalling that

$$A = \frac{\mu_G^2}{2(\mu_G+2)}\bar{\phi}_v, \quad \bar{\phi}_v = \frac{\mu_G^2}{8(\mu_G+4)\omega_{v,1}^2}\bar{\phi}_z,$$

we have

$$\frac{A}{\bar{\phi}_z} = \frac{\mu_G^4}{16(\mu_G+2)(\mu_G+4)\omega_{v,1}^2}.$$

Hence, defining

$$c_\gamma := \min\left\{\frac{1}{2}, \frac{\mu_G^2}{2(\mu_G+2)\left(\sqrt{L_{1,G}^4 + 2\omega_{dx}\bar{C}_x A} + L_{1,G}^2\right)}\right\},$$

the simpler condition

$$\gamma_k < \frac{\mu_G^2}{8\omega_{v,1}^2(\mu_G+4)}c_\gamma\bar{\phi}_z\rho_k^2$$

implies the preceding condition on γ_k (this is the reason why we reported $\mathcal{O}(\bar{\phi}\rho_k^2)$ in the main paper). Taking the full expectation and summing for $k = 1, \cdots, K$, we get

$$\sum_{k=0}^{K}\frac{\gamma_k}{2}\mathbb{E}\left[\|\nabla\Psi(x_k)\|^2\right] \leq \mathbb{E}[L_0 - L_{K+1}] + \sum_{k=0}^{K}\bar{C}_k$$
$$= \mathbb{E}[\Psi(x_0) + \phi_z^0\delta_{z_0} + \phi_v^0\delta_{v_0} \underbrace{-\Psi(x_{K+1})}_{\leq -\min\Psi} \underbrace{-\phi_z^{K+1}\delta_{z_{K+1}} - \phi_v^{K+1}\delta_{v_{K+1}}}_{\leq 0}]$$
$$+ \sum_{k=0}^{K}\bar{C}_k$$
$$\leq [\Psi(x_0) - \min\Psi] + \phi_z^0\delta_{z_0} + \phi_v^0\delta_{v_0} + \sum_{k=0}^{K}\bar{C}_k.$$

Dividing both sides by $\sum_{k=0}^{K} \gamma_k$, we get the claim i.e.

$$\frac{1}{\sum_{k=0}^{K} \gamma_k} \sum_{k=0}^{K} \frac{\gamma_k}{2} \mathbb{E}\left[\|\nabla \Psi(x_k)\|^2\right] \leq \frac{[\Psi(x_0) - \min \Psi]}{\sum_{k=0}^{K} \gamma_k} + \frac{\phi_z^0 \delta_{z_0} + \phi_v^0 \delta_{v_0}}{\sum_{k=0}^{K} \gamma_k} + \frac{\sum_{k=0}^{K} \bar{C}_k}{\sum_{k=0}^{K} \gamma_k}.$$

E.2 Proof of Corollary 3.5

By Theorem 3.4, we have

$$\frac{1}{\sum_{k=0}^{K} \gamma_k} \sum_{k=0}^{K} \frac{\gamma_k}{2} \mathbb{E}\left[\|\nabla \Psi(x_k)\|^2\right] \leq \frac{[\Psi(x_0) - \min \Psi]}{\sum_{k=0}^{K} \gamma_k} + \frac{\phi_z^0 \delta_{z_0} + \phi_v^0 \delta_{v_0}}{\sum_{k=0}^{K} \gamma_k} + \frac{\sum_{k=0}^{K} \bar{C}_k}{\sum_{k=0}^{K} \gamma_k}.$$

Recalling that

$$\bar{C}_k = C_1^2 \gamma_k h_k^2 + C_2^k \rho_k h_k^2 + C_4 \phi_z^k \rho_k^2 h_k^2 + C_3 \phi_z^k \rho_k^2 + 2\phi_v^k C_6^k(h_k) \rho_k^2 + \omega_{dx}^k C_5^k(h_k) \gamma_k^2.$$

where,

$$\begin{aligned}
C_2^k &:= \left(\frac{L_{1,G}^2}{4\mu_G} (d+3)^3 \phi_z^k + 2\frac{L_{1,G}^2 (d+3)^3}{\mu_G} \phi_v^k + 4\frac{\omega_{v,2} L_{0,F}^2}{\mu_G^3} \phi_v^k \right) \\
C_3 &:= 4\left(3 + \frac{p+2}{\ell_1}\right) \frac{\sigma_{1,G}^2}{b_1}, \quad C_4 := 2\left(\frac{(p+6)^3}{2b_1 \ell_1} + \left(\frac{3}{b_1} + 1\right)(p+3)^3 \right) L_{1,G}^2 \\
C_5^k(h_k) &:= \left(2C_x^k \frac{L_{0,F}^2}{\mu_G^2} + 2\left(\frac{(d+6)^3}{2b_2 \ell_2} + \left(\frac{3}{b_2} + 1\right)(d+3)^3 \right) L_{1,F}^2 h_k^2 \right. \\
&\quad \left. + 4\left(1 + \frac{d+2}{b_2 \ell_2}\right) L_{0,F}^2 + 4\left(\frac{d+2}{b_2 \ell_2} + \frac{3}{b_2} \right) \sigma_{1,F}^2 \right) \\
C_6^k(h_k) &:= \left(C_v^k \frac{L_{0,F}^2}{\mu_G^2} + 4\left(1 + \frac{p+2}{b_2 \ell_2}\right) L_{0,F}^2 + 4\left(\frac{p+2}{\ell_2} + 3 \right) \frac{\sigma_{1,F}^2}{b_2} \right. \\
&\quad \left. + 2\left(\frac{(p+6)^3}{2b_2 \ell_2} + \left(\frac{3}{b_2} + 1\right)(p+3)^3 \right) L_{1,F}^2 h_k^2 \right).
\end{aligned} \tag{50}$$

Recalling that

$$\frac{\mu_G \rho_k}{4} > \frac{1}{\omega_k^z} = \frac{\mu_G \rho_k}{\mu_G \rho_k + 4} \quad \text{and} \quad \frac{\mu_G \rho_k}{2} > \frac{1}{\omega_k^v} = \frac{\mu_G \rho_k}{\mu_G \rho_k + 2}.$$

Let $\phi_z^k = \frac{\bar{\phi}_z}{\omega_k^z}$ and $\phi_v^k = \frac{\bar{\phi}_v}{\omega_k^v}$. By the choice of ϕ_v^k, ϕ_z^k we have

$$\omega_{dx}^k = \frac{L_\Psi}{2} + 2L_*^2(\bar{\phi}_z + \bar{\phi}_v) =: \omega_{dx}.$$

Since $h_k^2 \leq \min\left\{ \frac{\left(2b_1 \ell_1 + 15p(p+6)^2\right) L_{1,G}^2}{8L_{2,G}^2 (p+16)^4}, \frac{\mu_G^2}{16\omega_{v,2}}, \frac{1}{16L_{2,G}^2((p+8)^4 + p(d+12)^3)} \right\}$, we have,

$$C_v^k = 8\left(1 + \frac{15(p+6)^2 p}{2b_1 \ell_1}\right) L_{1,G}^2 =: \bar{C}_v,$$

and $C_x^k = 2\left[\frac{\frac{1}{2}+6(p+4)(p+2)p+36(p+2)\min(p,d)+30p(d+2)dL_{1,G}^2}{b_1\ell_1} + L_{1,G}^2\right] =: \bar{C}_x$. By the choice of parameters (I), we bound the terms in eq. 50 as

$$C_1^2\gamma_k h_k^2 = C_1^2\gamma h^2, \qquad C_3\phi_z^k\rho_k^2 < \underbrace{C_3\frac{\mu_G}{4}\bar{\phi}_z}_{=:C_{3,1}}\rho^3, \qquad C_4\phi_z^k\rho_k^2 h_k^2 < \underbrace{C_4\bar{\phi}_z\frac{\mu_G}{4}}_{=:C_{4,1}}\rho^3 h^2$$

$$C_2^k\rho_k h_k^2 < \underbrace{\left(\frac{L_{1,G}^2}{16}(d+3)^3\bar{\phi}_z + L_{1,G}^2(d+3)^3\bar{\phi}_v + \frac{2\omega_{v,2}L_{0,F}^2}{\mu_G^2}\bar{\phi}_v\right)}_{=:C_2}\rho h^2$$

$$\begin{aligned}
C_5(h_k) &< 2\bar{C}_x\frac{L_{0,F}^2}{\mu_G^2}\omega_{dx}\gamma^2 + 2\left(\frac{(d+6)^3}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(d+3)^3\right)L_{1,F}^2\omega_{dx}h^2\gamma^2 \\
&\quad + 4\left(1+\frac{d+2}{b_2\ell_2}\right)L_{0,F}^2\omega_{dx}\gamma^2 + 4\left(\frac{d+2}{b_2\ell_2}+\frac{3}{b_2}\right)\sigma_{1,F}^2\omega_{dx}\gamma^2 \\
&= \underbrace{\left(2\bar{C}_x\frac{L_{0,F}^2}{\mu_G^2} + 4\left(1+\frac{d+2}{b_2\ell_2}\right)L_{0,F}^2 + 4\left(\frac{d+2}{b_2\ell_2}+\frac{3}{b_2}\right)\sigma_{1,F}^2\right)\omega_{dx}}_{C_{5,1}}\gamma^2 \\
&\quad + \underbrace{2\left(\frac{(d+6)^3}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(d+3)^3\right)L_{1,F}^2\omega_{dx}}_{C_{5,2}}h^2\gamma^2
\end{aligned}$$

$$\begin{aligned}
C_6(h_k) &< \underbrace{\left(\bar{C}_v\frac{L_{0,F}^2}{\mu_G} + 4\left(1+\frac{p+2}{b_2\ell_2}\right)L_{0,F}^2\mu_G + 4\left(\frac{p+2}{\ell_2}+3\right)\frac{\mu_G\sigma_{1,F}^2}{b_2}\right)}_{C_{6,1}}\rho^3 \\
&\quad + \underbrace{2\mu_G\left(\frac{(p+6)^3}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(p+3)^3\right)L_{1,F}^2}_{C_{6,2}}h^2\rho^3
\end{aligned}$$

Using these bound, we get the claim

$$\begin{aligned}
\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Psi(x_k)\|^2\right] &\le \frac{[\Psi(x_0)-\min\Psi]+\phi_z^0\delta_{z_0}+\phi_v^0\delta_{v_0}}{\gamma(K+1)} + \left(C_1^2 + C_2\frac{\rho}{\gamma}\right)h^2 \\
&\quad + \left(C_{3,1}+C_{4,1}h^2+C_{6,1}+C_{6,2}h^2\right)\frac{\rho^3}{\gamma} + \left(C_{5,1}+C_{5,2}h^2\right)\gamma.
\end{aligned}$$

By the choice of parameters (II), we have $\gamma_k = \gamma(k+1)^{-\alpha_1}$, $\rho_k = \rho(k+1)^{-\alpha_2}$ and $h_k = h(k+1)^{-\alpha_3}$. Thus, we bound the terms in eq. 50 as

$$\begin{aligned}
C_1^2\gamma_k h_k^2 &= C_1^2\gamma h^2(k+1)^{-(\alpha_1+2\alpha_3)}, \quad C_3\phi_z^k\rho_k^2 \le \frac{C_3\mu_G}{4}\bar{\phi}_z\rho^3(k+1)^{-3\alpha_2} \\
C_4\phi_z^k\rho_k^2 h_k^2 &\le \frac{C_4\mu_G}{4}\bar{\phi}_z\rho^3h^2(k+1)^{-(3\alpha_1+2\alpha_3)}, \quad C_2^k\rho_k h_k^2 \le C_2\rho h^2(k+1)^{-(\alpha_2+2\alpha_3)} \\
C_5(h_k)\omega_{dx}^k\gamma_k^2 &\le C_{5,1}\gamma^2(k+1)^{-2\alpha_1} + C_{5,2}h^2\gamma^2(k+1)^{-2(\alpha_1+\alpha_3)} \\
2\phi_v^k C_6(h_k)\rho_k^2 &\le C_{6,1}\rho^3(k+1)^{-3\alpha_2} + C_{6,2}h^2\rho^3(k+1)^{-(3\alpha_2+2\alpha_3)}.
\end{aligned}$$

Since $\alpha_1, \alpha_2 \in (1/2, 1)$ and $\alpha_3 > 1$, we have

$$\sum_{k=0}^{K}\gamma_k \le \frac{(K+1)^{1-\alpha_1}-1}{1-\alpha_1}\gamma \quad \text{and} \quad \sum_{k=0}^{K}\rho_k \le \frac{(K+1)^{1-\alpha_2}-1}{1-\alpha_2}\rho.$$

Moreover, we get

$$\sum_{k=0}^{K} C_1^2 \gamma_k h_k^2 = C_1^2 \gamma^2 h^2 \sum_{k=0}^{K} (k+1)^{-(\alpha_1+2\alpha_3)} \leq C_1^2 \gamma^2 h^2 \frac{\alpha_1+2\alpha_3}{\alpha_1+2\alpha_3-1},$$
$$\sum_{k=0}^{K} C_2^k \rho_k h_k^2 \leq C_2 \rho h^2 \sum_{k=0}^{K} (k+1)^{-(\alpha_2+2\alpha_3)} \leq C_2 \rho h^2 \frac{\alpha_2+2\alpha_3}{\alpha_2+2\alpha_3-1},$$
$$\sum_{k=0}^{K} C_3 \phi_z^k \rho_k^2 \leq \frac{C_3 \mu_G}{4} \bar{\phi}_z \sum_{k=0}^{K} \rho_k^3 \leq \frac{C_3 \mu_G}{4} \bar{\phi}_z \rho^3 \frac{3\alpha_2}{3\alpha_2-1},$$
$$\sum_{k=0}^{K} C_4 \phi_z^k \rho_k^2 h_k^2 \leq \frac{C_4 \mu_G}{4} \bar{\phi}_z \rho^3 h^2 \sum_{k=0}^{K} (k+1)^{-(3\alpha_1+2\alpha_3)} \leq \frac{C_4 \mu_G}{4} \bar{\phi}_z \rho^3 h^2 \frac{3\alpha_1+2\alpha_3}{3\alpha_1+2\alpha_3-1},$$
$$\begin{aligned}\sum_{k=0}^{K} C_5(h_k) \omega_{dx}^k \gamma_k^2 &\leq C_{5,1} \gamma^2 \sum_{k=0}^{K} (k+1)^{-2\alpha_1} + C_{5,2} h^2 \gamma^2 \sum_{k=0}^{K} (k+1)^{-2(\alpha_1+\alpha_3)} \\ &\leq C_{5,1} \gamma^2 \frac{2\alpha_1}{2\alpha_1-1} + C_{5,2} h^2 \gamma^2 \frac{2(\alpha_1+\alpha_3)}{2(\alpha_1+\alpha_3)-1} \\ 2\sum_{k=0}^{k} \phi_v^k C_6(h_k) \rho_k^2 &\leq C_{6,1} \rho^3 \sum_{k=0}^{k} (k+1)^{-3\alpha_2} + C_{6,2} h^2 \rho^3 \sum_{k=0}^{k} (k+1)^{-(3\alpha_2+2\alpha_3)} \\ &\leq C_{6,1} \rho^3 \frac{3\alpha_2}{3\alpha_2-1} + C_{6,2} h^2 \rho^3 \frac{3\alpha_2+2\alpha_3}{3\alpha_2+2\alpha_3-1}.\end{aligned}$$

Therefore, using these bounds, we get the claim

$$\frac{1}{\sum_{k=0}^{K} \gamma_k} \sum_{k=0}^{K} \gamma_k \mathbb{E}\left[\|\nabla \Psi(x_k)\|^2\right] \leq \frac{[\Psi(x_0) - \min \Psi] + \phi_z^0 \delta_{z_0} + \phi_v^0 \delta_{v_0}}{\gamma (K+1)^{1-\alpha_1}} + \mathcal{O}\left(\frac{1}{(K+1)^{1-\alpha_1}}\right).$$

Finally, to prove the last claim, by choosing $\gamma_k = \gamma$, $\rho_k = \rho$ and $h_k = h$ for every $k \in \mathbb{N}$, we have

$$\begin{aligned}\frac{1}{K+1} \sum_{k=0}^{K} \mathbb{E}\left[\|\nabla \Psi(x_k)\|^2\right] &\leq \frac{[\Psi(x_0) - \min \Psi] + \phi_z^0 \delta_{z_0} + \phi_v^0 \delta_{v_0}}{\gamma (K+1)} + \left(C_1^2 + C_2 \frac{\rho}{\gamma}\right) h^2 \\ &\quad + \left(C_3 + C_4 h^2\right) \mu_G \bar{\phi}_z \frac{\rho^3}{\gamma} + \left(C_{5,1} + C_{5,2} h^2\right) \gamma + \left(C_{6,1} + C_{6,2} h^2\right) \frac{\rho^2}{\gamma}.\end{aligned}$$

Rearranging the terms and by $\phi_v^k = \frac{\bar{\phi}_v}{\omega_k^v}$, $\phi_z^k = \frac{\bar{\phi}_z}{\omega_k^z}$ and $\bar{\phi}_v = \frac{\mu_G^2}{8(\mu_G+4)\omega_{v,1}^2} \bar{\phi}_z$ we have

$$\begin{aligned}\frac{1}{K+1} \sum_{k=0}^{K} \mathbb{E}\left[\|\nabla \Psi(x_k)\|^2\right] &\leq \frac{1}{\gamma(K+1)} \left([\Psi(x_0) - \min \Psi] + \frac{\bar{\phi}_z}{\omega_0^z} \delta_{z_0} + \frac{\mu_G^2}{8(\mu_G+4)\omega_{v,1}^2 \omega_0^v} \bar{\phi}_z \delta_{v_0}\right) \\ &\quad + \left(C_1^2 + C_2 \frac{\rho}{\gamma} + C_{5,2} \omega_{dx} \gamma + C_4 \mu_G \bar{\phi}_z \frac{\rho^3}{\gamma} + C_{6,2} \frac{\rho^2}{\gamma}\right) h^2 \\ &\quad + C_3 \mu_G \bar{\phi}_z \frac{\rho^3}{\gamma} + C_{5,1} \gamma + C_{6,1} \frac{\rho^2}{\gamma}.\end{aligned}$$

Due to the conditions on ρ_k, we have

$$\frac{\mu_G}{4} > \frac{1}{\omega_k^z} = \frac{\mu_G \rho_k}{\mu_G \rho_k + 4} \quad \text{and} \quad \frac{\mu_G}{2} > \frac{1}{\omega_k^v} = \frac{\mu_G \rho_k}{\mu_G \rho_k + 2}.$$

Then, we get

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Psi(x_k)\|^2\right] \leq \frac{1}{\gamma(K+1)}\underbrace{\left([\Psi(x_0)-\min\Psi]+\frac{\bar{\phi}_z\mu_G^2}{4}\delta_{z_0}+\frac{\mu_G^3}{16(\mu_G+4)\omega_{v,1}^2}\bar{\phi}_z\delta_{v_0}\right)}_{=:\Delta_1}$$
$$+\left(C_1^2+C_2\frac{\rho}{\gamma}+C_{5,2}\omega_{dx}\gamma+C_4\mu_G\bar{\phi}_z\frac{\rho^3}{\gamma}+C_{6,2}\frac{\rho^2}{\gamma}\right)h^2$$
$$+C_3\mu_G\bar{\phi}_z\frac{\rho^3}{\gamma}+C_{5,1}\gamma+C_{6,1}\frac{\rho^2}{\gamma}.$$

Let $\varepsilon\in(0,1)$ and recall from Corollary statement the definitions $D_1:=\frac{p(d+p)^2}{b_1\ell_1}+\frac{d}{b_2\ell_2}$ and $D_2:=\frac{p^3}{b_1\ell_1}+\frac{p}{b_2\ell_2}$. Consider the following inequality,

$$\frac{\Delta_1}{\gamma(K+1)}+C(\gamma,\rho)h^2+\left(C_3\mu_G\bar{\phi}_z\frac{\rho^3}{\gamma}+C_{5,1}\gamma+C_{6,1}\frac{\rho^2}{\gamma}\right)\leq\varepsilon.$$

Since the term with $C(\gamma,\rho)$ can be controlled with h, we choose γ by minimizing $\frac{\Delta_1}{\gamma(K+1)}\left(C_3\mu_G\bar{\phi}_z\frac{\rho^3}{\gamma}+C_{5,1}\gamma+C_{6,1}\frac{\rho^2}{\gamma}\right)$ i.e.

$$\gamma=\sqrt{\frac{1}{C_{5,1}}\left(\frac{\Delta_1}{K+1}+C_3\mu_G\bar{\phi}_z\rho^3+C_{6,1}\rho^2\right)}.$$

Thus, we get

$$2\sqrt{C_{5,1}\left(\frac{\Delta_1}{(K+1)}+C_3\mu_G\bar{\phi}_z\rho^3+C_{6,1}\rho^2\right)}+C(\gamma,\rho)h^2\leq\varepsilon.$$

Let $h=\sqrt{\frac{\varepsilon}{2C(\gamma,\rho)}}$. Notice that, due to the definition of $C(\gamma,\rho)$, we have $h=\mathcal{O}\left(\sqrt{\varepsilon}\left(1+\sqrt{\frac{D_1}{D_2}}\right)^{-1/2}\right)$ as indicated in the corollary statement. We get

$$2\sqrt{C_{5,1}\left(\frac{\Delta_1}{(K+1)}+C_3\mu_G\bar{\phi}_z\rho^3+C_{6,1}\rho^2\right)}\leq\frac{\varepsilon}{2}.$$

Multplying both sides by $1/2$ and squaring, we have

$$\frac{\Delta_1}{(K+1)}\leq\frac{\varepsilon^2}{16C_{5,1}}-(C_3\mu_G\bar{\phi}_z\rho+C_{6,1})\rho^2.$$

Set $\rho=\frac{\varepsilon}{\sqrt{64C_{5,1}C_{6,1}}}$ and $\bar{\phi}_z=\frac{C_{6,1}\sqrt{64C_{5,1}C_{6,1}}}{C_3\mu_G\varepsilon}=\frac{C_{6,1}}{C_3\mu_G\rho}$. Notice that for the definitions of $C_{5,1}C_{6,1}$ we have $\rho=\mathcal{O}\left(\varepsilon\left(D_1D_2\right)^{-1/2}\right)$. We get

$$\frac{\Delta_1}{(K+1)}\leq\frac{\varepsilon^2}{32C_{5,1}}.$$

Therefore, choosing

$$(K+1)\geq 32C_{5,1}\Delta_1\varepsilon^{-2}=\mathcal{O}\left(D_1\,R_{\text{init}}\,\varepsilon^{-2}\right),$$

we get $\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Psi(x_k)\|^2\right]\leq\varepsilon$. Choosing $b_1\ell_1=\lfloor\mathcal{O}(p(d+p)^2/c_1)\rfloor$ and $b_2\ell_2=\lfloor\mathcal{O}(p(p+d)^2/c_2)\rfloor$ for any $c_1,c_2\geq 1$, such that $c_1,c_2\leq p(p+d)^2$, the dependence on the dimensions p,d simplifies, yielding the claim (III) i.e. the complexity is

$$\mathcal{O}\left(p(d+p)^2\varepsilon^{-2}\right).$$

E.3 Proof of Theorem 3.6

Let $\delta_{z_k} = \|z_k - z^*(x_k)\|^2$ and $\delta_{v_k} = \|v_k - v^*(x_k)\|^2$. For $\phi_z^k, \phi_v^k > 0$ s.t. $\phi_v^{k+1} \leq \phi_v^k$ and $\phi_z^{k+1} \leq \phi_z^k$, define

$$L_k = \Psi(x_k) + \phi_z^k \delta_{z_k} + \phi_v^k \delta_{v_k}.$$

Thus, we have

$$L_{k+1} - L_k \leq \Psi(x_{k+1}) - \Psi(x_k) + \phi_z^k \left(\delta_{z_{k+1}} - \delta_{z_k}\right) + \phi_v^k \left(\delta_{v_{k+1}} - \delta_{v_k}\right).$$

Taking the conditional expectation,

$$\mathbb{E}_k[L_{k+1}] - L_k \leq \mathbb{E}_k[\Psi(x_{k+1})] - \Psi(x_k) + \phi_z^k \left(\mathbb{E}_k[\delta_{z_{k+1}}] - \delta_{z_k}\right) + \phi_v^k \left(\mathbb{E}_k[\delta_{v_{k+1}}] - \delta_{v_k}\right).$$

Let

$$\begin{aligned}
C_{1,\Psi} &= L_{1,F} + \frac{L_{0,F} L_{1,G}}{\mu_G}, \\
C_{2,\Psi} &= \frac{2L_{2,G} L_{0,F}}{3\mu_G} \left((d+5)^{5/2} + (d+3)^{3/2} + (d+1)(p+3)^{3/2}\right) + \frac{L_{1,F} d\sqrt{p}}{2} \\
C_{3,\Psi} &= \frac{L_{2,G}^2 L_{0,F}^2}{2\mu_G^2} \quad \text{and} \quad C_{4,\Psi} = 4L_{1,G}^2 + \frac{\mu_G^2}{16}.
\end{aligned}$$

By Lemma D.19,

$$\begin{aligned}
\mathbb{E}_k[L_{k+1}] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 + \frac{L_\Psi}{2}\gamma_k^2 \mathbb{E}_k[\|\hat{D}_x^k\|^2] \\
& + 4C_{1,\Psi}^2 \gamma_k \delta_{z_k} + C_{4,\Psi}\gamma_k \delta_{v_k} \\
& + \phi_z^k \left(\mathbb{E}_k[\delta_{z_{k+1}}] - \delta_{z_k}\right) + \phi_v^k \left(\mathbb{E}_k[\delta_{v_{k+1}}] - \delta_{v_k}\right) \\
& + 2C_{2,\Psi}^2 \gamma_k h_k^2 + C_{3,\Psi} \hat{h}_k^2 \gamma_k.
\end{aligned}$$

Notice that the bounds on δ_{z_k} of Lemma D.13 holds also for Algorithm 2. Let

$$\begin{aligned}
C_1 &= 2\frac{L_{1,G}^2 (d+3)^2}{\mu_G} + 4\frac{\omega_{v,2} L_{0,F}^2}{\mu_G^3} \quad \text{and} \quad C_2 = \frac{8}{\mu_G}\frac{L_{2,G}^2}{4}\frac{L_{0,F}^2}{\mu_G^2}, \\
C_3 &= \frac{L_{1,G}^2}{4\mu_G}(d+3)^3.
\end{aligned}$$

Thus, by Lemma D.13 and Lemma D.18, we get

$$\begin{aligned}
\mathbb{E}_k[L_{k+1}] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 \\
& - \left(\frac{\mu_G}{2}\phi_v^k \rho_k - C_{4,\Psi}\gamma_k\right)\delta_{v_k} \\
& - \left(\frac{\mu_G}{2}\phi_z^k \rho_k - 4\frac{\omega_{v,1}^2}{\mu_G}\phi_v^k \rho_k - 4C_{1,\Psi}^2 \gamma_k\right)\delta_{z_k} \\
& + 2\phi_z^k \rho_k^2 \mathbb{E}_k\left[\left\|\hat{D}_z^k\right\|^2\right] + 2\phi_v^k \rho_k^2 \mathbb{E}_k[\|\hat{D}_v^k\|^2] \\
& + \left(2\omega_k^v L_*^2 \phi_v^k + \frac{L_\Psi}{2} + 2\omega_k^z L_*^2 \phi_z^k\right)\gamma_k^2 \mathbb{E}_k[\|\hat{D}_x^k\|^2] \\
& + (C_1 h_k^2 + C_2 \hat{h}_k^2)\phi_v^k \rho_k + 2C_{2,\Psi}^2 \gamma_k h_k^2 + C_{3,\Psi}\hat{h}_k^2 \gamma_k \\
& + C_3 \phi_z^k \rho_k h_k^2.
\end{aligned}$$

Observing that by Lemma D.17, we have

$$\begin{aligned}\mathbb{E}_k\left[\left\|\hat{D}_z^k\right\|^2\right] \leq & \underbrace{2\left(2+\frac{p+2}{b_1\ell_1}\right)}_{=:C_{A,1}}\|\nabla_z G(z_k,x_k)\|^2 \\ & + \underbrace{\left(\frac{(p+6)^3}{2b_1\ell_1}+\left(\frac{3}{b_1}+1\right)(p+3)^3\right)L_{1,G}^2}_{=:C_{A,2}}h_k^2 \\ & + \underbrace{2\left(3+\frac{p+2}{\ell_1}\right)\frac{\sigma_{1,G}^2}{b_1}}_{C_{A,3}}\end{aligned}$$

Since $\nabla_z G(z^*(x_k),x_k)=0$, by $L_{1,G}$-smoothness we have

$$\begin{aligned}\mathbb{E}_k\left[\left\|\hat{D}_z^k\right\|^2\right] &\leq C_{A,1}\|\nabla_z G(z_k,x_k)-\nabla_z G(z^*(x_k),x_k)\|^2 + C_{A,2}h_k^2 + C_{A,3} \\ &\leq C_{A,1}L_{1,G}^2\delta_{z_k} + C_{A,2}h_k^2 + C_{A,3}.\end{aligned}$$

Thus, we have

$$\begin{aligned}\mathbb{E}_k[L_{k+1}]-L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 \\ & -\left(\frac{\mu_G}{2}\phi_v^k\rho_k - C_{4,\Psi}\gamma_k\right)\delta_{v_k} \\ & -\left(\frac{\mu_G}{2}\phi_z^k\rho_k - 4\frac{\omega_{v,1}^2}{\mu_G}\phi_v^k\rho_k - 4C_{1,\Psi}^2\gamma_k - 2\phi_z^k\rho_k^2 C_{A,1}L_{1,G}^2\right)\delta_{z_k} \\ & + 2\phi_v^k\rho_k^2\mathbb{E}_k[\|\hat{D}_v^k\|^2] \\ & + \left(2\omega_k^v L_*^2\phi_v^k + \frac{L_\Psi}{2} + 2\omega_k^z L_*^2\phi_z^k\right)\gamma_k^2\mathbb{E}_k[\|\hat{D}_x^k\|^2] \\ & + (C_1h_k^2 + C_2\hat{h}_k^2)\phi_v^k\rho_k + 2C_{2,\Psi}^2\gamma_k h_k^2 + C_{3,\Psi}\hat{h}_k^2\gamma_k \\ & + C_3\phi_z^k\rho_k h_k^2 + 2C_{A,2}\phi_z^k\rho_k^2 h_k^2 \\ & + 2\phi_z^k\rho_k^2 C_{A,3}.\end{aligned}$$

Let $C_{B,1}=4\left(1+\frac{3(p+2)}{b_1\ell_1}\right)$, $C_{B,2}=2\left(1+\frac{p+2}{b_2\ell_2}\right)$, $C_{B,3}=\left(\frac{(p+6)^3}{2b_2\ell_2}+\left(\frac{3}{b_2}+1\right)(p+3)^3\right)L_{1,F}^2$ and $C_{B,4}=2\left(\frac{p+2}{\ell_2}+3\right)\frac{\sigma_{1,F}^2}{b_2}$. Let $C_{C,1}=4\left(1+\frac{3(d+2)}{b_1\ell_1}\right)$, $C_{C,2}=2\left(1+\frac{d+2}{b_2\ell_2}\right)$, $C_{C,3}=\left(\frac{(d+6)^3}{2b_2\ell_2}+\left(\frac{3}{b_2}+1\right)(d+3)^3\right)L_{1,F}^2$ and $C_{C,4}=2\left(\frac{d+2}{b_2\ell_2}+\frac{3}{b_2}\right)\sigma_{1,F}^2$. Let

$$S_k = 2\omega_k^v L_*^2\phi_v^k + \frac{L_\Psi}{2} + 2\omega_k^z L_*^2\phi_z^k.$$

By Lemma D.17, we have

$$\begin{aligned}
\mathbb{E}_k[L_{k+1}] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 \\
& - \left(\frac{\mu_G}{2}\phi_v^k\rho_k - C_{4,\Psi}\gamma_k - 4C_{B,1}L_{1,G}^2\phi_v^k\rho_k^2 - 2C_{C,1}L_{1,G}^2 S_k\gamma_k^2\right)\delta_{v_k} \\
& - \left(\frac{\mu_G}{2}\phi_z^k\rho_k - 4\frac{\omega_{v,1}^2}{\mu_G}\phi_v^k\rho_k - 4C_{1,\Psi}^2\gamma_k - 2\phi_z^k\rho_k^2 C_{A,1}L_{1,G}^2\right)\delta_{z_k} \\
& + (C_1 h_k^2 + C_2\hat{h}_k^2)\phi_v^k\rho_k + 2C_{2,\Psi}^2\gamma_k h_k^2 + C_{3,\Psi}\hat{h}_k^2\gamma_k \\
& + 2C_{B,1}\frac{L_{1,G}^2 L_{0,F}^2}{\mu_G^2}\phi_v^k\rho_k^2 \\
& + C_3\phi_z^k\rho_k h_k^2 + 2C_{A,2}\phi_z^k\rho_k^2 h_k^2 \\
& + 2C_{C,1}\frac{L_{1,G}^2 L_{0,F}^2}{\mu_G^2}S_k\gamma_k^2 \\
& + 2\phi_z^k\rho_k^2 C_{A,3} \\
& + \underbrace{\left(\frac{12L_{1,G}^2(p+6)^3}{b_1\ell_1}\phi_v^k\rho_k^2 + \frac{6L_{1,G}^2(d+6)^3}{b_1\ell_1}S_k\gamma_k^2\right)}_{C_{L,k}}\frac{h_k^2}{\bar{h}_k^2} \\
& + (2C_{C,2}L_{0,F}^2 + 2C_{C,3}h_k^2 + 2C_{C,4})S_k\gamma_k^2 \\
& + 2\phi_v^k\rho_k^2(2C_{B,2}L_{0,F}^2 + 2C_{B,3}h_k^2 + 2C_{B,4}).
\end{aligned}$$

Recall that

$$\bar{h}_k = \begin{cases} \frac{\hat{h}_k}{\|v_k\|} & \text{if} \quad \|v_k\| \neq 0 \\ \hat{h}_k \end{cases}.$$

Therefore, for every $k \in \mathbb{N}$, for which $\|v_k\| \neq 0$, we have

$$\begin{aligned}
C_{L,k}\frac{h_k^2}{\bar{h}_k^2} &= C_{L,k}\frac{h_k^2}{\hat{h}_k^2}\|v_k\|^2 \\
&\leq 2C_{L,k}\frac{h_k^2}{\hat{h}_k^2}\|v_k - v^*(x_k)\|^2 + 2C_{L,k}\frac{h_k^2}{\hat{h}_k^2}\|v^*(x_k)\|^2 \\
&\leq 2C_{L,k}\frac{h_k^2}{\hat{h}_k^2}\|v_k - v^*(x_k)\|^2 + \frac{2L_{0,F}^2}{\mu_G^2}C_{L,k}\frac{h_k^2}{\hat{h}_k^2},
\end{aligned}$$

where the last inequality follows by Lemma D.3. For every $k \in \mathbb{N}$ for which $\|v_k\| = 0$, we have instead

$$C_{L,k}\frac{h_k^2}{\bar{h}_k^2} = C_{L,k}\frac{h_k^2}{\hat{h}_k^2}.$$

Let $\theta_k = \begin{cases} \frac{2L_{0,F}^2}{\mu_G^2} & \text{if } \|v_k\| \neq 0 \\ 1 \end{cases} < 1 + \frac{2L_{0,F}^2}{\mu_G} =: \theta$. Thus, for every $k \in \mathbb{N}$ we have

$$\begin{aligned}
C_{L,k}\frac{h_k^2}{\bar{h}_k^2} &\leq 2C_{L,k}\frac{h_k^2}{\hat{h}_k^2}\|v_k - v^*(x_k)\|^2 1_{\|v_k\|\neq 0} + \theta_k C_{L,k}\frac{h_k^2}{\hat{h}_k^2} \\
&\leq 2C_{L,k}\frac{h_k^2}{\hat{h}_k^2}\|v_k - v^*(x_k)\|^2 + \theta C_{L,k}\frac{h_k^2}{\hat{h}_k^2},
\end{aligned}$$

where the last inequality follows from the fact that $2C_{L,k}\frac{h_k^2}{\hat{h}_k^2}\|v_k - v^*(x_k)\|^2 \geq 0$. Therefore, we get

$$\begin{aligned}
\mathbb{E}_k[L_{k+1}] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 \\
& - \left(\frac{\mu_G}{2}\phi_v^k\rho_k - C_{4,\Psi}\gamma_k - 4C_{B,1}L_{1,G}^2\phi_v^k\rho_k^2 - 2C_{C,1}L_{1,G}^2 S_k\gamma_k^2 - 2C_{L,k}\frac{h_k^2}{\hat{h}_k^2}\right)\delta_{v_k} \\
& - \left(\frac{\mu_G}{2}\phi_z^k\rho_k - 4\frac{\omega_{v,1}^2}{\mu_G}\phi_v^k\rho_k - 4C_{1,\Psi}^2\gamma_k - 2\phi_z^k\rho_k^2 C_{A,1}L_{1,G}^2\right)\delta_{z_k} \\
& + (C_1h_k^2 + C_2\hat{h}_k^2)\phi_v^k\rho_k + 2C_{2,\Psi}^2\gamma_k h_k^2 + C_{3,\Psi}\hat{h}_k^2\gamma_k + 2C_{B,1}\frac{L_{1,G}^2L_{0,F}^2}{\mu_G^2}\phi_v^k\rho_k^2 \\
& + C_3\phi_z^k\rho_k h_k^2 + 2C_{A,2}\phi_z^k\rho_k^2 h_k^2 + 2C_{C,1}\frac{L_{1,G}^2L_{0,F}^2}{\mu_G^2}S_k\gamma_k^2 + \theta C_{L,k}\frac{h_k^2}{\hat{h}_k^2} \\
& + (2C_{C,2}L_{0,F}^2 + 2C_{C,3}h_k^2 + 2C_{C,4})S_k\gamma_k^2 \\
& + 2\phi_v^k\rho_k^2(2C_{B,2}L_{0,F}^2 + 2C_{B,3}h_k^2 + 2C_{B,4}).
\end{aligned}$$

Let

$$\begin{aligned}
\hat{C}_k = & (C_1h_k^2 + C_2\hat{h}_k^2)\phi_v^k\rho_k + 2C_{2,\Psi}^2\gamma_k h_k^2 + C_{3,\Psi}\hat{h}_k^2\gamma_k \\
& + 2C_{B,1}\frac{L_{1,G}^2L_{0,F}^2}{\mu_G^2}\phi_v^k\rho_k^2 + C_3\phi_z^k\rho_k h_k^2 + 2C_{A,2}\phi_z^k\rho_k^2 h_k^2 + 2C_{C,1}\frac{L_{1,G}^2L_{0,F}^2}{\mu_G^2}S_k\gamma_k^2 \\
& + 2\phi_z^k\rho_k^2 C_{A,3} + \frac{12L_{1,G}^2}{b_1\ell_1}\phi_v^k\rho_k^2 + \frac{6L_{1,G}^2}{b_1\ell_1}S_k\gamma_k^2 + \theta C_{L,k}\frac{h_k^2}{\hat{h}_k^2} \\
& + (2C_{C,2}L_{0,F}^2 + 2C_{C,3}h_k^2 + 2C_{C,4})S_k\gamma_k^2 + 2\phi_v^k\rho_k^2(2C_{B,2}L_{0,F}^2 + 2C_{B,3}h_k^2 + 2C_{B,4}).
\end{aligned} \tag{51}$$

Then,

$$\begin{aligned}
\mathbb{E}_k[L_{k+1}] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 \\
& - \left(\frac{\mu_G}{2}\phi_v^k\rho_k - C_{4,\Psi}\gamma_k - 4C_{B,1}L_{1,G}^2\phi_v^k\rho_k^2 - 2C_{C,1}L_{1,G}^2 S_k\gamma_k^2 - 2C_{L,k}\frac{h_k^2}{\hat{h}_k^2}\right)\delta_{v_k} \\
& - \left(\frac{\mu_G}{2}\phi_z^k\rho_k - 4\frac{\omega_{v,1}^2}{\mu_G}\phi_v^k\rho_k - 4C_{1,\Psi}^2\gamma_k - 2\phi_z^k\rho_k^2 C_{A,1}L_{1,G}^2\right)\delta_{z_k} \\
& + \hat{C}_k.
\end{aligned}$$

Since $\rho_k \leq \min\left(1, \frac{1}{16C_{B,1}L_{1,G}^2}\right)$, and since $C_{A,1} = 4 + \frac{2(p+2)}{b_1\ell_1} \leq 4 + \frac{12(p+2)}{b_1\ell_1} = C_{B,1}$ for every admissible p, b_1, ℓ_1, the condition $\rho_k \leq \frac{1}{4C_{A,1}L_{1,G}^2}$ is implied by $\rho_k \leq \frac{1}{16C_{B,1}L_{1,G}^2}$. Thus, we have

$$\begin{aligned}
\mathbb{E}_k[L_{k+1}] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 \\
& - \left(\frac{\mu_G}{4}\phi_v^k\rho_k - C_{4,\Psi}\gamma_k - 2C_{C,1}L_{1,G}^2 S_k\gamma_k^2 - 2C_{L,k}\frac{h_k^2}{\hat{h}_k^2}\right)\delta_{v_k} \\
& - \left(\left(\frac{\mu_G}{4}\phi_z^k - 4\frac{\omega_{v,1}^2}{\mu_G}\phi_v^k\right)\rho_k - 4C_{1,\Psi}^2\gamma_k\right)\delta_{z_k} \\
& + \hat{C}_k.
\end{aligned}$$

Let

$$\hat{S} := \frac{L_\Psi + 4(\bar{\phi}_v + \bar{\phi}_z)L_*^2}{2}.$$

Choosing $\phi_z^k = \bar{\phi}_z/\omega_k^z$ and $\phi_v^k = \bar{\phi}_v/\omega_k^v$. we have

$$\begin{aligned}\mathbb{E}_k[L_{k+1}] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 \\ & - \left(\frac{\mu_G}{4}\frac{\bar{\phi}_v}{\omega_k^v}\rho_k - C_{4,\Psi}\gamma_k - 2C_{C,1}L_{1,G}^2\hat{S}\gamma_k^2 - 2C_{L,k}\frac{h_k^2}{\hat{h}_k^2}\right)\delta_{v_k} \\ & - \left(\left(\frac{\mu_G}{4}\frac{\bar{\phi}_z}{\omega_k^z} - 4\frac{\omega_{v,1}^2}{\mu_G}\frac{\bar{\phi}_v}{\omega_k^v}\right)\rho_k - 4C_{1,\Psi}^2\gamma_k\right)\delta_{z_k} \\ & + \hat{C}_k.\end{aligned}$$

Let

$$\gamma_{\max}^{(1)} = \frac{-C_{4,\Psi} + \sqrt{C_{4,\Psi}^2 + C_{C,1}L_{1,G}^2\hat{S}\frac{\mu_G^2}{\mu_G+2}\bar{\phi}_v\rho_k^2}}{4C_{C,1}L_{1,G}^2\hat{S}},$$

$$\gamma_{\max}^{(2)} = \frac{\mu_G^2}{32C_{1,\Psi}^2(\mu_G+4)}\bar{\phi}_z\rho_k^2.$$

Notice that since $\rho_k \leq \bar{\rho} < \min(1, \frac{1}{16C_{B,1}L_{1,G}^2})$, and $\gamma_k \leq \bar{\gamma} < \min(\gamma_{\max}^{(1)}, \gamma_{\max}^{(2)})$, we have that for every $k \in \mathbb{N}$,

$$\begin{aligned}\frac{\mu_G\rho_k}{4} > \frac{1}{\omega_k^z} = \frac{\mu_G\rho_k}{\mu_G\rho_k+4} > \frac{\mu_G\rho_k}{\mu_G+4} \\ \frac{\mu_G\rho_k}{2} > \frac{1}{\omega_k^v} = \frac{\mu_G\rho_k}{\mu_G\rho_k+2} > \frac{\mu_G\rho_k}{\mu_G+2},\end{aligned}$$

and

$$\begin{aligned}C_{L,k} &= \left(\frac{12L_{1,G}^2(p+6)^3}{b_1\ell_1}\phi_v^k\rho_k^2 + \frac{6L_{1,G}^2(d+6)^3}{b_1\ell_1}S_k\gamma_k^2\right) \\ &\leq \left(\frac{6L_{1,G}^2(p+6)^3}{b_1\ell_1}\bar{\phi}_v\mu_G\bar{\rho}^2 + \frac{6L_{1,G}^2(d+6)^3}{b_1\ell_1}\hat{S}\bar{\gamma}^2\right) =: \bar{C}_L\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}_k[L_{k+1}] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 \\ & - \left(\frac{\mu_G^2}{4(\mu_G+2)}\bar{\phi}_v\rho_k^2 - C_{4,\Psi}\gamma_k - 2C_{C,1}L_{1,G}^2\hat{S}\gamma_k^2 - 2\bar{C}_L\frac{h_k^2}{\hat{h}_k^2}\right)\delta_{v_k} \\ & - \left(\left(\frac{\mu_G^2}{4(\mu_G+4)}\bar{\phi}_z - 2\omega_{v,1}^2\bar{\phi}_v\right)\rho_k^2 - 4C_{1,\Psi}^2\gamma_k\right)\delta_{z_k} \\ & + \hat{C}_k.\end{aligned}$$

Since $h_k \leq \frac{\mu_G}{4\sqrt{(\mu_G+2)\bar{C}_L}}\sqrt{\bar{\phi}_v}\hat{h}_k\rho_k$, we get

$$\begin{aligned}\mathbb{E}_k[L_{k+1}] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 \\ & - \left(\frac{\mu_G^2}{8(\mu_G+2)}\bar{\phi}_v\rho_k^2 - C_{4,\Psi}\gamma_k - 2C_{C,1}L_{1,G}^2\hat{S}\gamma_k^2\right)\delta_{v_k} \\ & - \left(\left(\frac{\mu_G^2}{4(\mu_G+4)}\bar{\phi}_z - 2\omega_{v,1}^2\bar{\phi}_v\right)\rho_k^2 - 4C_{1,\Psi}^2\gamma_k\right)\delta_{z_k} \\ & + \hat{C}_k.\end{aligned}$$

Notice that for

$$\frac{\mu_G^2}{8\omega_{v,1}^2(\mu_G+4)}\bar{\phi}_z > \bar{\phi}_v,$$

we have that

$$\left(\frac{\mu_G^2}{4(\mu_G+4)}\bar{\phi}_z - 2\omega_{v,1}^2\bar{\phi}_v\right) > 0.$$

Thus, due to the choice $\bar{\phi}_v = \frac{\mu_G^2}{16\omega_{v,1}^2(\mu_G+4)}\bar{\phi}_z$, we have

$$\begin{aligned}\mathbb{E}_k[L_{k+1}] - L_k \leq & -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 \\ & - \left(\frac{\mu_G^2}{8(\mu_G+2)}\bar{\phi}_v\rho_k^2 - C_{4,\Psi}\gamma_k - 2C_{C,1}L_{1,G}^2\hat{S}\gamma_k^2\right)\delta_{v_k} \\ & - \left(\frac{\mu_G^2}{8(\mu_G+4)}\bar{\phi}_z\rho_k^2 - 4C_{1,\Psi}^2\gamma_k\right)\delta_{z_k} \\ & + \hat{C}_k.\end{aligned}$$

For $\gamma_k < \gamma_{\max}^{(2)}$, we have

$$-\left(\frac{\mu_G^2}{8(\mu_G+4)}\bar{\phi}_z\rho_k^2 - 4C_{1,\Psi}^2\gamma_k\right) < 0.$$

For $\gamma_k < \gamma_{\max}^{(1)}$, we have

$$-\left(\frac{\mu_G^2}{8(\mu_G+2)}\bar{\phi}_v\rho_k^2 - C_{4,\Psi}\gamma_k - 2C_{C,1}L_{1,G}^2\hat{S}\gamma_k^2\right) < 0.$$

Thus, for $\gamma_k < \min(\gamma_{\max}^{(1)}, \gamma_{\max}^{(2)})$ and $\rho_k < \min(1, \frac{1}{16C_{B,1}L_{1,G}^2}, \frac{1}{4C_{A,1}L_{1,G}^2})$, we have

$$\mathbb{E}_k[L_{k+1}] - L_k \leq -\frac{\gamma_k}{2}\|\nabla\Psi(x_k)\|^2 + \hat{C}_k.$$

For a more compact characterization of the condition on γ_k, note that, since $\rho_k < 1$, letting

$$B := C_{C,1}L_{1,G}^2\hat{S}\,\frac{\mu_G^2}{\mu_G+2}\,\bar{\phi}_v,$$

we have

$$\begin{aligned}\gamma_{\max}^{(1)} &= \frac{B\rho_k^2}{4C_{C,1}L_{1,G}^2\hat{S}\left(\sqrt{C_{4,\Psi}^2 + B\rho_k^2} + C_{4,\Psi}\right)} \\ &\geq \frac{B}{4C_{C,1}L_{1,G}^2\hat{S}\left(\sqrt{C_{4,\Psi}^2 + B} + C_{4,\Psi}\right)}\,\rho_k^2 \\ &= \frac{\frac{\mu_G^2}{\mu_G+2}\bar{\phi}_v}{4\left(\sqrt{C_{4,\Psi}^2 + B} + C_{4,\Psi}\right)}\,\rho_k^2.\end{aligned}$$

Recalling that $\bar{\phi}_v = \dfrac{\mu_G^2}{16\omega_{v,1}^2(\mu_G+4)}\bar{\phi}_z$, we have

$$\gamma_{\max}^{(1)} \geq \frac{\mu_G^4}{64\,\omega_{v,1}^2(\mu_G+2)(\mu_G+4)\left(\sqrt{C_{4,\Psi}^2 + B} + C_{4,\Psi}\right)}\,\bar{\phi}_z\rho_k^2.$$

Hence, defining

$$c_\gamma := \min\left\{\frac{1}{2},\ \frac{C_{1,\Psi}^2\,\mu_G^2}{2\,\omega_{v,1}^2(\mu_G+2)\left(\sqrt{C_{4,\Psi}^2 + B} + C_{4,\Psi}\right)}\right\},$$

the simpler condition

$$\gamma_k < \frac{\mu_G^2}{32C_{1,\Psi}^2(\mu_G+4)}\,c_\gamma\,\bar{\phi}_z\rho_k^2$$

implies $\gamma_k < \min\left(\gamma_{\max}^{(1)}, \gamma_{\max}^{(2)}\right)$ (this is the reason why we reported $\mathcal{O}(\bar{\phi}_z \rho_k^2)$ in the main paper). Taking the full expectation and summing for $k = 0, \cdots, K$, we get

$$\mathbb{E}[L_{K+1}] - L_0 \leq -\sum_{k=0}^{K} \frac{\gamma_k}{2} \mathbb{E}\left[\|\nabla \Psi(x_k)\|^2\right] + \sum_{k=0}^{K} \hat{C}_k.$$

Rearranging the terms, and observing that $-\min \Psi \geq -\mathbb{E}[\Psi(x_{K+1})]$, we get

$$\begin{aligned}
\sum_{k=0}^{K} \frac{\gamma_k}{2} \mathbb{E}\left[\|\nabla \Psi(x_k)\|^2\right] &\leq L_0 - \mathbb{E}[L_{K+1}] + \sum_{k=0}^{K} \hat{C}_k \\
&= \Psi(x_0) \underbrace{-\mathbb{E}[\Psi(x_{K+1})]}_{\leq -\min \Psi} + \phi_z^0 \|z_0 - z^*(x_0)\|^2 + \phi_v^0 \|v_0 - v^*(x_0)\|^2 \\
&\quad \underbrace{-\phi_z^0 \mathbb{E}\left[\|z_{K+1} - z^*(x_{K+1})\|^2\right] - \phi_v^0 \mathbb{E}\left[\|v_{K+1} - v^*(x_{K+1})\|^2\right]}_{\leq 0} \\
&\quad + \sum_{k=0}^{K} \hat{C}_k.
\end{aligned}$$

Let

$$R_{\text{init}} = \Psi(x_0) - \min \Psi + \phi_z^0 \|z_0 - z^*(x_0)\|^2 + \phi_v^0 \|v_0 - v^*(x_0)\|^2.$$

Multiplying by 2 and dividing by $\sum_{k=0}^{K} \gamma_k$ in both sides, we get the claim

$$\frac{1}{\sum_{k=0}^{K} \gamma_k} \sum_{k=0}^{K} \gamma_k \mathbb{E}\left[\|\nabla \Psi(x_k)\|^2\right] \leq \frac{2}{\sum_{k=0}^{K} \gamma_k} \left(R_{\text{init}} + \sum_{k=0}^{K} \hat{C}_k\right).$$

E.4 Proof of Corollary 3.7

By Theorem 3.6, we have

$$\frac{1}{\sum_{k=0}^{K} \gamma_k} \sum_{k=0}^{K} \frac{\gamma_k}{2} \mathbb{E}\left[\|\nabla \Psi(x_k)\|^2\right] \leq \frac{1}{\sum_{k=0}^{K} \gamma_k} \left(R_{\text{init}} + \sum_{k=0}^{K} \hat{C}_k\right).$$

Recall that

$$\begin{aligned}
\hat{C}_k &= (C_1 h_k^2 + C_2 \hat{h}_k^2) \phi_v^k \rho_k + 2C_{2,\Psi}^2 \gamma_k h_k^2 + C_{3,\Psi} \hat{h}_k^2 \gamma_k \\
&\quad + 2C_{B,1} \frac{L_{1,G}^2 L_{0,F}^2}{\mu_G^2} \phi_v^k \rho_k^2 + C_3 \phi_z^k \rho_k h_k^2 + 2C_{A,2} \phi_z^k \rho_k^2 h_k^2 + 2C_{C,1} \frac{L_{1,G}^2 L_{0,F}^2}{\mu_G^2} S_k \gamma_k^2 \\
&\quad + 2\phi_z^k \rho_k^2 C_{A,3} + \frac{12 L_{1,G}^2}{b_1 \ell_1} \phi_v^k \rho_k^2 + \frac{6 L_{1,G}^2}{b_1 \ell_1} S_k \gamma_k^2 + \theta C_{L,k} \frac{h_k^2}{\hat{h}_k^2} \\
&\quad + (2C_{C,2} L_{0,F}^2 + 2C_{C,3} h_k^2 + 2C_{C,4}) S_k \gamma_k^2 + 2\phi_v^k \rho_k^2 (2C_{B,2} L_{0,F}^2 + 2C_{B,3} h_k^2 + 2C_{B,4}).
\end{aligned}$$

where $\theta = 1 + \frac{2L_{0,F}^2}{\mu_G^2}$ and

$$C_1 = 2\frac{L_{1,G}^2(d+3)^2}{\mu_G} + 4\frac{\omega_{v,2}L_{0,F}^2}{\mu_G^3}, \quad C_2 = \frac{8}{\mu_G}\frac{L_{2,G}^2}{4}\frac{L_{0,F}^2}{\mu_G^2},$$
$$C_3 = \frac{L_{1,G}^2}{4\mu_G}(d+3)^3, \quad C_{A,1} = 2\left(2 + \frac{p+2}{b_1\ell_1}\right), \quad C_{A,2} = \left(\frac{1}{2b_1\ell_1} + \left(\frac{3}{b_1}+1\right)(p+3)^3\right)L_{1,G}^2$$
$$C_{A,3} = 2\left(3 + \frac{p+2}{\ell_1}\right)\frac{\sigma_{1,G}^2}{b_1}, \quad C_{B,1} = 4\left(1 + \frac{3(p+2)}{b_1\ell_1}\right), \quad C_{B,2} = 2\left(1 + \frac{p+2}{b_2\ell_2}\right)$$
$$C_{B,3} = \left(\frac{(p+6)^3}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(p+3)^3\right)L_{1,F}^2, \quad C_{B,4} = 2\left(\frac{p+2}{\ell_2}+3\right)\frac{\sigma_{1,F}^2}{b_2}$$
$$C_{C,1} = 4\left(1 + \frac{3(d+2)}{b_1\ell_1}\right), \quad C_{C,2} = 2\left(1 + \frac{d+2}{b_2\ell_2}\right), \quad C_{C,3} = \left(\frac{1}{2b_2\ell_2} + \left(\frac{3}{b_2}+1\right)(d+3)^3\right)L_{1,F}^2$$
$$C_{C,4} = 2\left(\frac{d+2}{b_2\ell_2} + \frac{3}{b_2}\right)\sigma_{1,F}^2, \quad C_{L,k} = \left(\frac{12L_{1,G}^2(p+6)^3}{b_1\ell_1}\phi_v^k\rho_k^2 + \frac{6L_{1,G}^2(d+6)^3}{b_1\ell_1}S_k\gamma_k^2\right)$$

With the choice $\phi_z^k = \bar{\phi}_z/\omega_k^z$ and $\phi_v^k = \bar{\phi}_v/\omega_k^v$, we have

$$S_k = \frac{L_\Psi + 4(\bar{\phi}_v + \bar{\phi}_z)L_*^2}{2} = \hat{S}.$$

Moreover, due to the conditions on ρ_k, we have for every $k \in \mathbb{N}$

$$\frac{\mu_G}{4} > \frac{\mu_G\rho_k}{4} > \frac{1}{\omega_k^z} = \frac{\mu_G\rho_k}{\mu_G\rho_k+4} > \frac{\mu_G\rho_k}{\mu_G+4}$$
$$\frac{\mu_G}{2} > \frac{\mu_G\rho_k}{2} > \frac{1}{\omega_k^v} = \frac{\mu_G\rho_k}{\mu_G\rho_k+2} > \frac{\mu_G\rho_k}{\mu_G+2},$$

and, thus,

$$C_{L,k} \le 6\left(\frac{L_{1,G}^2(p+6)^3\mu_G}{b_1\ell_1}\bar{\phi}_v\rho_k^2 + \frac{L_{1,G}^2(d+6)^3}{b_1\ell_1}\hat{S}\gamma_k^2\right)$$

Replacing γ_k, ρ_k, h_k and $\hat{h}_k$ with the parameter choice (I), we have

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Psi(x_k)\|^2\right] \le \frac{2}{\gamma(K+1)}\left(R_{\text{init}} + \sum_{k=0}^{K}\hat{C}\right), \tag{52}$$

with

$$\begin{aligned}\hat{C} &= (C_2\bar{\phi}_v\rho^2 + C_{3,\Psi}\gamma + 2C_{C,3}\hat{S}\gamma^2)\hat{h}^2 + (C_1\bar{\phi}_v\rho^2 + C_3\bar{\phi}_z\rho^2 + 2C_{2,\Psi}^2\gamma + 2C_{A,2}\bar{\phi}_z\rho^3 + 2C_{B,3}\bar{\phi}_v\rho^3)h^2 \\ &\quad + 2C_{B,1}\frac{L_{1,G}^2L_{0,F}^2}{\mu_G^2}\bar{\phi}_v\rho^3 + 2C_{A,3}\bar{\phi}_z\rho^3 + \frac{12L_{1,G}^2}{b_1\ell_1}\bar{\phi}_v\rho^3 \\ &\quad + 6\theta\left(\frac{L_{1,G}^2(p+6)^3\mu_G}{b_1\ell_1}\bar{\phi}_v\rho^2 + \frac{L_{1,G}^2(d+6)^3}{b_1\ell_1}\hat{S}\gamma^2\right)\frac{h^2}{\hat{h}^2} \\ &\quad + 2C_{C,1}\frac{L_{1,G}^2L_{0,F}^2}{\mu_G^2}\hat{S}\gamma^2 + \frac{6L_{1,G}^2}{b_1\ell_1}\hat{S}\gamma^2 + (2C_{C,2}L_{0,F}^2 + 2C_{C,4})\hat{S}\gamma^2 + 2\bar{\phi}_v\rho^3(2C_{B,2}L_{0,F}^2 + 2C_{B,4}).\end{aligned}$$

Rearranging the terms, we get the first claim. The proof of the point (II) follows the same line of point (II) of proof of Corollary 3.5. Now, we prove the point (III). Recalling that $\bar{\phi}_v = \frac{\mu_G^2}{16\omega_{v,1}^2(\mu_G+4)}\bar{\phi}_z$, we

have

$$\begin{aligned}\hat{C} = &\underbrace{(C_2\bar{\phi}_v\rho^2 + C_{3,\Psi}\gamma + 2C_{C,3}\hat{S}\gamma^2)}_{C_1(\gamma,\rho)}\hat{h}^2 \\ &+ \underbrace{(C_1\bar{\phi}_v\rho^2 + C_3\bar{\phi}_z\rho^2 + 2C_{2,\Psi}^2\gamma + 2C_{A,2}\bar{\phi}_z\rho^3 + 2C_{B,3}\bar{\phi}_v\rho^3)}_{C_2(\gamma,\rho)}h^2 \\ &+ \underbrace{6\theta\left(\frac{L_{1,G}^2(p+6)^3\mu_G}{b_1\ell_1}\bar{\phi}_v\rho^2 + \frac{L_{1,G}^2(d+6)^3}{b_1\ell_1}\hat{S}\gamma^2\right)}_{C_3(\gamma,\rho)}\frac{h^2}{\hat{h}^2} \\ &+ \underbrace{\left(2C_{C,1}\frac{L_{1,G}^2L_{0,F}^2}{\mu_G^2} + \frac{6L_{1,G}^2}{b_1\ell_1} + (2C_{C,2}L_{0,F}^2 + 2C_{C,4})\right)\hat{S}\,\gamma^2}_{\tilde{C}_1} \\ &\underbrace{\left(C_{B,1}\frac{L_{1,G}^2L_{0,F}^2}{8\omega_{v,1}^2(\mu_G+4)} + 2C_{A,3} + \frac{12L_{1,G}^2\mu_G^2}{16\omega_{v,1}^2(\mu_G+4)b_1\ell_1} + \frac{\mu_G^2(2C_{B,2}L_{0,F}^2 + 2C_{B,4})}{8\omega_{v,1}^2(\mu_G+4)}\right)}_{\tilde{C}_2}\bar{\phi}_z\rho^3.\end{aligned}$$

Therefore, plugging $\hat{C}$ in eq. 52, we get

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Psi(x_k)\|^2\right] \leq \frac{2R_{\text{init}}}{\gamma(K+1)} + \tilde{C}_1\gamma + \tilde{C}_2\bar{\phi}_z\frac{\rho^3}{\gamma} + \frac{C_1(\gamma,\rho)}{\gamma}\hat{h}^2 + \frac{C_2(\gamma,\rho)}{\gamma}h^2 + \frac{C_3(\gamma,\rho)}{\gamma}\frac{h^2}{\hat{h}^2}.$$

By the choice $h = \hat{h}^2$ and since $\hat{h} < 1$, we have

$$\begin{aligned}\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Psi(x_k)\|^2\right] &\leq \frac{2R_{\text{init}}}{\gamma(K+1)} + \tilde{C}_1\gamma + \tilde{C}_2\bar{\phi}_z\frac{\rho^3}{\gamma} + \frac{C_1(\gamma,\rho)}{\gamma}\hat{h}^2 + \frac{C_2(\gamma,\rho)}{\gamma}\hat{h}^4 + \frac{C_3(\gamma,\rho)}{\gamma}\hat{h}^2 \\ &< \frac{2R_{\text{init}}}{\gamma(K+1)} + \tilde{C}_1\gamma + \tilde{C}_2\bar{\phi}_z\frac{\rho^3}{\gamma} + \frac{(C_1(\gamma,\rho) + C_2(\gamma,\rho) + C_3(\gamma,\rho))}{\gamma}\hat{h}^2.\end{aligned}$$

Since $\rho \leq \bar{\rho} < \min(1, \frac{1}{16C_{B,1}L_{1,G}^2})$, and by the choice $\gamma = \hat{c}_\gamma\rho$, we have

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Psi(x_k)\|^2\right] < \frac{2R_{\text{init}}}{\hat{c}_\gamma\rho(K+1)} + \left(\tilde{C}_1\hat{c}_\gamma + \frac{\tilde{C}_2}{\hat{c}_\gamma}\right)\rho + \frac{(C_1(\gamma,\rho) + C_2(\gamma,\rho) + C_3(\gamma,\rho))}{\hat{c}_\gamma\rho}\hat{h}^2.$$

Moreover, we have

$$\begin{aligned}C_1(\gamma,\rho) &< \underbrace{\left(C_2\frac{\mu_G^2}{16\omega_{v,1}^2(\mu_G+4)} + C_{3,\Psi}\hat{c}_\gamma + 2C_{C,3}\hat{S}\hat{c}_\gamma^2\bar{\rho}\right)}_{\bar{\Delta}_1}\rho \\ C_2(\gamma,\rho) &< \underbrace{\left(C_1\frac{\mu_G^2}{16\omega_{v,1}^2(\mu_G+4)} + C_3 + 2C_{2,\Psi}^2\hat{c}_\gamma + 2C_{A,2}\bar{\rho} + \frac{C_{B,3}\mu_G^2}{8\omega_{v,1}^2(\mu_G+4)}\bar{\rho}\right)}_{\bar{\Delta}_2}\rho \\ C_3(\gamma,\rho) &< \underbrace{6\theta\left(\frac{L_{1,G}^2(p+6)^3\mu_G^3}{16\omega_{v,1}^2(\mu_G+4)b_1\ell_1} + \frac{L_{1,G}^2(d+3)^3}{b_1\ell_1}\hat{S}\hat{c}_\gamma\bar{\rho}\right)}_{\bar{\Delta}_3}\rho\end{aligned}$$

Thus, we have

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Psi(x_k)\|^2\right] < \frac{2R_{\text{init}}}{\hat{c}_\gamma\rho(K+1)} + \left(\tilde{C}_1\hat{c}_\gamma + \frac{\tilde{C}_2}{\hat{c}_\gamma}\right)\rho + \frac{(\bar{\Delta}_1 + \bar{\Delta}_2 + \bar{\Delta}_3)}{\hat{c}_\gamma}\hat{h}^2.$$

Let $\varepsilon \in (0,1)$. Set $\hat{h} \leq \min(1, \sqrt{\frac{\hat{c}_\gamma \varepsilon}{2(\bar{\Delta}_1+\bar{\Delta}_2+\bar{\Delta}_3)}})$. We get

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Psi(x_k)\|^2\right] < \frac{2R_{\text{init}}}{\hat{c}_\gamma\rho(K+1)} + \left(\tilde{C}_1\hat{c}_\gamma + \frac{\tilde{C}_2}{\hat{c}_\gamma}\right)\rho + \frac{\varepsilon}{2}.$$

Consider the following inequality

$$\frac{2R_{\text{init}}}{\hat{c}_\gamma\rho(K+1)} + \underbrace{\left(\tilde{C}_1\hat{c}_\gamma + \frac{\tilde{C}_2}{\hat{c}_\gamma}\right)}_{\bar{\Delta}_4}\rho \leq \frac{\varepsilon}{2}.$$

We choose ρ by minimizing the left handside i.e.

$$\rho = \sqrt{\frac{2R_{\text{init}}}{\hat{c}_\gamma\bar{\Delta}_4(K+1)}}$$

Thus, we get

$$2\sqrt{\frac{2R_{\text{init}}\bar{\Delta}_4}{\hat{c}_\gamma(K+1)}} < \frac{\varepsilon}{2}.$$

Hence, for

$$K+1 > \frac{32R_{\text{init}}\bar{\Delta}_4}{\hat{c}_\gamma}\varepsilon^{-2},$$

we have $\frac{1}{K+1}\sum_{k=0}^{k}\mathbb{E}\left[\|\nabla\Psi(x_k)\|^2\right] < \varepsilon$. Therefore, choosing $b_1\ell_1 = \mathcal{O}(\lfloor\frac{d+p}{c_1'}\rfloor)$ and $b_2\ell_2 = \mathcal{O}(\lfloor\frac{d+p}{c_2'}\rfloor)$ with $c_1', c_2' > 1$ and $c_1', c_2' \leq (d+p)$, we get that the complexity is

$$\mathcal{O}\left((d+p)\varepsilon^{-2}\right).$$

F LIMITATIONS

In this appendix, we discuss the main limitations of our work. We first emphasize that the primary goal of this paper is to introduce novel zeroth-order bilevel optimization algorithms that improve practical performance while preserving convergence guarantees. In particular, we proposed ZOBA and HF-ZOBA, two single-loop finite-difference methods that leverage delayed information to enable full parallelization of function evaluations and reuse of function values, thereby reducing the number of evaluations per iteration and improving scalability compared to prior state-of-the-art single-loop approaches. These features allow our methods to perform effectively in limited-budget settings. Moreover, our analysis shows that HF-ZOBA achieves optimal complexity, demonstrating that such efficiency gains can be obtained without increasing the number of function evaluations, unlike in methods such as VRZSBO and ZOSOGD. However, achieving optimal complexity requires strong theoretical assumptions. In particular, the inner objective must be strongly convex, an assumption that is often violated in practice. This limitation is common to most bilevel optimization algorithms and is shared by all existing finite-difference approaches (Aghasi & Ghadimi, 2025; Aghasi et al., 2025; Nazari et al., 2025; Peng et al., 2025). As indicated in Section 5, an important direction for future work is to relax this assumption. Additionally, our analysis relies on the smoothness of both the inner and outer objectives, whereas many practical applications involve non-smooth functions. Extending the analysis to such settings is non-trivial: one possible direction is to approximate hypergradients via smooth surrogates of the value function, although this may incur significant computational cost. Moreover, in this case, stationarity would be defined with respect to the smooth approximation, which does not directly characterize optimality for the original problem; alternative notions of stationarity (e.g., Goldstein stationarity) may therefore be required. Finally, from a practical perspective, although our methods demonstrate strong empirical performance, they require tuning several hyperparameters and such operation can be very costly. This limitation is shared by existing state-of-the-art finite-difference bilevel methods, and developing adaptive parameter selection strategies while preserving convergence guarantees remains an important direction for future work.