BOME! Bilevel Optimization Made Easy: A Simple First-Order Approach

author names withheld

Under Review for OPT 2022

Abstract

Bilevel optimization (BO) is useful for solving a variety of important machine learning problems including but not limited to hyperparameter optimization, meta-learning, continual learning, and reinforcement learning. Conventional BO methods need to differentiate through the low-level optimization process with implicit differentiation, which requires expensive calculations related to the Hessian matrix. There has been a recent quest for first-order methods for BO, but the methods proposed to date tend to be complicated and impractical for large-scale deep learning applications. In this work, we propose a simple first-order BO algorithm that depends only on first-order gradient information, requires no implicit differentiation, and is practical and efficient for large-scale non-convex functions in deep learning. We provide non-asymptotic convergence analysis of the proposed method to stationary points for non-convex objectives and present empirical results that show its superior practical performance.

1. Introduction

We consider the bilevel optimization (BO) problem:

$$\min_{v,\theta} f\big(v,\theta\big) \quad s.t. \quad \theta \in \arg\min_{\theta'} g\big(v,\theta'\big), \tag{1}$$

where the goal is to minimize an *outer objective* f whose variables include the solution of another minimization problem w.r.t an *inner objective* g. The θ and v are the *inner* and *outer* variables, respectively. We assume that $v \in \mathbb{R}^m, \theta \in \mathbb{R}^n$ and that $g(v,\cdot)$ attains a minimum for each v.

BO is notoriously challenging due to its nested nature. Despite the large literature, most existing methods for BO are slow and unsatisfactory in various ways. For example, a major class of BO methods is based on direct gradient descent on the outer variable v while viewing the optimal inner variable $\theta^*(v) = \arg\min_\theta g(v,\theta)$ as a (uniquely defined) function of v. The key difficulty is to calculate the derivative $\nabla_v \theta^*(v)$ which may require expensive manipulation of the Hessian matrix of g via the implicit differentiation theorem. Another approach is to replace the low level optimization with the stationary condition $\nabla_\theta g(v,\theta) = 0$. This still requires Hessian information, and more importantly, is unsuitable for nonconvex g since it allows θ to be any stationary point of $g(v,\cdot)$, not necessarily a minimizer. To the best of our knowledge, the only existing fully first-order BO algorithms[1] are BSG-1 [8] and BVFSM with its variants [18–20]; but BSG-1 relies on a non-vanishing approximation that does not yield convergence to the correct solution in general,

1. By fully first-order, we mean methods that only require information of $f, g, \nabla f, \nabla g$, so this excludes methods that apply auto-differentiation or conjugate gradient that need multiple steps of matrix-vector computation.

and BVFSM is sensitive to hyper-parameters on large-scale practical problems and lacks a complete non-asymptotic analysis for the practically implemented algorithm.

In this work, we seek a *simple and fast fully first-order* BO method that can be used with non-convex functions including those appear in deep learning applications. The idea is to reformulate (1) as a single-level constrained optimization problem using the so-called value-function-based approach [4, 5]. The constrained problem is then solved by stopping gradient on the single variable that contains the higher-order information and applying a simple first-order dynamic barrier gradient descent method based on a method of Gong et al. [9]. Our contributions are: **1)** we introduce a novel and fast BO method by applying a modified dynamic barrier gradient descent on the value-function reformulation of BO; **2)** Theoretically, we establish the non-asymptotic convergence of our method to local stationary points (as measured by a special KKT loss) for non-convex f and g. Importantly, to the best of our knowledge, this work is the first to establish non-asymptotic convergence rate for a fully first-order BO method.

2. Background

This section provides a brief background on traditional BO methods. Please see Bard [1], Dempe [3], Dempe and Zemkoho [4] for overviews, and Liu et al. [17] for a survey on recent ML applications.

Hypergradient Descent Assume that the minimum of $g(v, \cdot)$ is unique for all v so that we can write $\theta^*(v) = \arg\min_\theta g(v, \theta)$ as a function of v; this is known as the low-level singleton (LLS) assumption. The most straightforward approach to solving (1) is to conduct gradient descent on $f(v, \theta^*(v))$ as a function of v. Note that

$$\nabla_v f(v, \theta^*(v)) = \nabla_1 f(v, \theta^*(v)) + \nabla_v \theta^*(v) \nabla_2 f(v, \theta^*(v)).$$

The difficulty is to compute $\nabla_v \theta^*(v)$. From implicit function theorem, it satisfies a linear equation:

$$\nabla_{1,2} g(v, \theta^*(v)) + \nabla_{2,2} g(v, \theta^*(v)) \nabla_v \theta^*(v) = 0. \tag{2}$$

If $\nabla_{2,2} g$ is invertible, we can solve for $\nabla_v \theta^*(v)$ and obtain a gradient update rule on v:

$$v_{k+1} \leftarrow v_k - \xi \left(\nabla_1 f_k - \left(\nabla_{1,2} g_k\right)^\top \left(\nabla_{2,2} g_k\right)^{-1} \nabla_2 f_k \right),$$

where k denotes iteration, $\nabla_1 f_k = \nabla_1 f(v_k, \theta^*(v_k))$ and similarly for the other terms. This approach is sometimes known as the *hypergradient descent*. However, hypergradient descent is computationally expensive: Besides requiring evaluation of the inner optimum $\theta^*(v_k)$, the main computational bottleneck is to solve the linear equation in (2).

Stationary-Seeking Methods. An alternative method is to replace the argmin constraint in (1) with the stationarity condition $\nabla_\theta g(v, \theta) = 0$, yielding a constrained optimization:

$$\min_{v,\theta} f(v, \theta) \quad s.t. \quad \nabla_\theta g(v, \theta) = 0. \tag{3}$$

Algorithms for nonlinear equality constrained optimization can then be applied [21]. The constraint in (3) guarantees only that θ is a stationary point of $g(v, \cdot)$, so it is equivalent to (1) only when g is convex w.r.t. θ. Otherwise, the solution of (3) can be a maximum or saddle point of g. This makes it problematic for deep learning, where non-convex functions are pervasive.

3. Method

We consider a *value function approach* [see e.g., 18, 23, 29], which yields natural first-order algorithms for non-convex g and requires no computation of Hessian matrices. It is based on the observation that (1) is equivalent to the following constrained optimization (even for non-convex g):

$$\min_{v,\theta} f(v,\theta) \quad s.t. \quad q(v,\theta) := g(v,\theta) - g^*(v) \leq 0, \tag{4}$$

where $g^*(v) := \min_\theta g(v,\theta) = g(v,\theta^*(v))$ is known as the value function. Compared with the hypergradient approach, this formulation **does not require calculation of the implicit derivative** $\nabla_v\theta^*(v)$: Although $g^*(v)$ depends on $\theta^*(v)$, its derivative $\nabla_v g^*(v)$ does not depend on $\nabla_v\theta^*(v)$, by Danskin's theorem:

$$\nabla_v g^*(v) = \nabla_1 g(v,\theta^*(v)) + \nabla_v\theta^*(v)\nabla_2 g(v,\theta^*(v)) = \nabla_1 g(v,\theta^*(v)), \tag{5}$$

where the second term in (5) vanishes because we have $\nabla_2 g(v,\theta^*(v)) = 0$ by definition of the optimum $\theta^*(v)$. Therefore, provided that we can evaluate $\theta^*(v)$ at each iteration, solving (4) yields an algorithm for BO that requires no Hessian computation.

Dynamic Barrier Gradient Descent. The idea is to iterative update the parameter (v,θ) to reduce f while controlling the decrease of the constraint q, ensuring that q decreases whenever $q > 0$. Specifically, denote ξ as the step size, the update at each step is

$$(v_{k+1},\theta_{k+1}) \leftarrow (v_k,\theta_k) - \xi\delta_k, \tag{6}$$

$$\text{where} \quad \delta_k = \arg\min_\delta \|\nabla f(v_k,\theta_k) - \delta\|^2 \;\text{ s.t. }\; \langle\nabla q(v_k,\theta_k),\delta\rangle \geq \phi_k. \tag{7}$$

Here $\nabla f_k := \nabla_{(v,\theta)} f(v_k,\theta_k)$, $\nabla q_k := \nabla_{(v,\theta)} q(v_k,\theta_k)$, and $\phi_k \geq 0$ is a non-negative control barrier and should be strictly positive $\phi_k > 0$ in the non-stationary points of q: the lower bound on the inner product of $\nabla q(v_k,\theta_k)$ and δ_k ensures that the update in (6) can only decrease q (when step size ξ is sufficiently small) until it reaches stationary. In addition, by enforcing δ_k to be close to $\nabla f(v_k,\theta_k)$ in (7), we decrease the objective f as much as possible so long as it does not conflict with descent of q.

Two straightforward choices of ϕ_k that satisfies the condition above are $\phi_k = \eta q(v_k,\theta_k)$ and $\phi_k = \eta\|\nabla q(v_k,\theta_k)\|^2$ with $\eta > 0$. We find that both choices of ϕ_k work well empirically and use $\phi_k = \eta\|\nabla q(v_k,\theta_k)\|^2$ as the default.

The optimization in (7) yields a simple closed form solution:

$$\delta_k = \nabla f(v_k,\theta_k) + \lambda_k \nabla q(v_k,\theta_k), \;\; \text{with} \;\; \lambda_k = \max\left(\frac{\phi_k - \langle\nabla f(v_k,\theta_k),\nabla q(v_k,\theta_k)\rangle}{||\nabla q(v_k,\theta_k)||^2},\; 0\right),$$

and $\lambda_k = 0$ in the case of $||\nabla q(v_k,\theta_k)|| = 0$.

Practical Approximation. The main bottleneck of the method above is to calculate the $q(v_k,\theta_k)$ and $\nabla q(v_k,\theta_k)$ which requires evaluation of $\theta^*(v_k)$. In practice, we approximate $\theta^*(v_k)$ by $\theta_k^{(T)}$, where $\theta_k^{(T)}$ is obtained by running T steps of gradient descent of $g(v_k,\cdot)$ w.r.t. θ starting from θ_k. That is, we set $\theta_k^{(0)} = \theta_k$ and let

$$\theta_k^{(t+1)} = \theta_k^{(t)} - \alpha\nabla_\theta g(v_k,\theta_k^{(t)}), \quad t = 0,\ldots,T-1, \tag{8}$$

for some step size parameter $\alpha > 0$. We obtain an estimate of $q(v, \theta)$ at iteration k by replacing $\theta^*(v_k)$ with $\theta_k^{(T)}$: $\hat{q}(v, \theta) = g(v, \theta) - g(v, \theta_k^{(T)})$.

We substitute $\hat{q}(v_k, \theta_k)$ into (7) to obtain the update direction δ_k. Note that the $\theta_k^{(T)}$ is viewed as a constant when defining $\hat{q}(v, \theta)$ and hence no differentiation of $\theta_k^{(T)}$ is performed when calculating the gradient $\nabla \hat{q}$. This differs from truncated back-propagation methods [e.g., 24] which differentiate through $\theta_k^{(T)}$ as a function of v.

4. Analysis

We establish a KKT condition of BO through the form in (3). Assume f and $\nabla_\theta q$ are continuously differentiable, and (v^*, θ^*) satisfies $\nabla_\theta q(v^*, \theta^*) = 0$ and CRCQ with $\nabla_\theta q$. Then by the typical first order KKT condition of (3), there exists a Lagrange multiplier $\omega^* \in \mathbb{R}^n$ such that

$$\nabla f(v^*, \theta^*) + \nabla(\nabla_\theta q(v^*, \theta^*))\omega^* = 0. \tag{9}$$

This condition can be viewed as the limit of a sequence of (12) in the following way: assume we relax the constraint in (4) to $q(v, \theta) \le c_k$ where c_k is a sequence of positive numbers that converge to zero, then we can establish (12) for each $c_k > 0$ and pass the limit to zero to yield (13).

Proposition 1 *Assume that f, q, ∇q are continuously differentiable and $\|\nabla f\|$, f is bounded. For a feasible point (v^*, θ^*) of (4) that satisfies CRCQ with $\nabla_\theta q$, if (v^*, θ^*) is the limit of a sequence $\{(v_k, \theta_k)\}_{k=1}^\infty$ satisfying $q(v_k, \theta_k) \neq 0 \; \forall k$, and there exists a sequence $\{\lambda_k\} \subset [0, \infty)$ such that*

$$\nabla f(v_k, \theta_k) + \lambda_k \nabla q(v_k, \theta_k) \to 0, \qquad\qquad q(v_k, \theta_k) \to 0,$$

as $k \to +\infty$, then (v^*, θ^*) satisfies (13).

This motivates us to use the following function as a measure of stationarity of the solution returned by the algorithm:

$$\mathcal{K}(v, \theta) = \underbrace{\min_{\lambda \ge 0} ||\nabla f(v, \theta) + \lambda \nabla q(v, \theta)||^2}_{\text{local improvement}} + \underbrace{q(v, \theta)}_{\text{feasibility}}.$$

The hope is to have an algorithm that generates a sequence $\{(v_k, \theta_k)\}_{k=0}^\infty$ that satisfies $\mathcal{K}(v_k, \theta_k) \to 0$ as $k \to +\infty$.

4.1. Convergence with multimodal g

Assumption 1 (Smoothness) *f and g are differentiable, and ∇f and ∇g are L-Lipschitz w.r.t. the joint inputs (v, θ) for some $L \in (0, +\infty)$.*

Assumption 2 (Boundedness) *There exists a constant $M < \infty$ such that $\|\nabla g(v, \theta)\|$, $\|\nabla f(v, \theta)\|$, $|f(v, \theta)|$ and $|g(v, \theta)|$ are all upper bounded by M for any (v, θ).*

To study cases in which g has multiple local optima, we introduce the notion of attraction points following gradient descent.

Definition 1 (Attraction points) *Given any* (v, θ)*, we say that* $\theta^\diamond(v, \theta)$ *is the attraction point of* (v, θ) *with step size* $\alpha > 0$ *if the sequence* $\{\theta^{(t)}\}_{t=0}^\infty$ *generated by gradient descent* $\theta^{(t)} = \theta^{(t-1)} - \alpha \nabla_\theta g(v, \theta^{(t-1)})$ *starting from* $\theta^{(0)} = \theta$ *converges to* $\theta^\diamond(v, \theta)$*.*

Assume the step size $\alpha \le 1/L$ where L is the smoothness constant defined in Assumption 5, one can show the existence and uniqueness of attraction point of any (v, θ) using Proposition 1.1 of Traonmilin and Aujol [27]. Intuitively, the attraction of (v, θ) is where the gradient descent algorithm can not make improvement. In fact, when $\alpha \le 1/L$, one can show that $g(v, \theta) \le g(v, \theta^\diamond(v, \theta))$ is equivalent to the stationary condition $\nabla_\theta g(v, \theta) = 0$.

The set of (v, θ) that have the same attraction point forms an attraction basin. Our analysis needs to assume the PL-inequality within the individual attraction basins.

Assumption 3 (Local PL-inequality within attraction basins) *Assume that for any* (v, θ)*,* $\theta^\diamond(v, \theta)$ *exists. Also assume that there exists* $\kappa > 0$ *such that for any* (v, θ) $\|\nabla_\theta g(v, \theta)\|^2 \ge \kappa(g(v, \theta) - g(v, \theta^\diamond(v, \theta))$*.*

We can also define local variants of q and $\mathcal{K}$ as follows:

$$q^\diamond(v, \theta) = g(v, \theta) - g(v, \theta^\diamond(v, \theta)), \quad \mathcal{K}^\diamond(v, \theta) = \min_{\lambda \ge 0} \|\nabla f(v, \theta) + \lambda \nabla q^\diamond(v, \theta)\|^2 + q^\diamond(v, \theta).$$

Compared with Section B.2, a key technical challenge is that $\theta^\diamond(v, \theta)$ and hence $q^\diamond(v, \theta)$ can be discontinuous w.r.t. θ when it is on the boundary of different attraction basins; $\mathcal{K}^\diamond$ is not well defined on these points. However, these boundary points are not stable stationary points, and it is possible to use arguments based on the stable manifold theorem to show that an algorithm with random initialization will almost surely not visit them [14, 25].

Theorem 1 *With* $\xi, \alpha \le 1/L$*,* $\phi_k = \eta \|\nabla \hat{q}(v_k, \theta_k)\|^2$*, and* $\eta > 0$*. Suppose that Assumptions 5, 6, and 3 hold and that* $q^\diamond$ *is differentiable on* (v_k, θ_k) *at every iteration* $k \ge 0$*. Then there exists a constant* c *depending on* α, κ, η, L*, such that when* $T \ge c$*, we have*

$$\min_{k \le K} \mathcal{K}^\diamond(v_k, \theta_k) = O\left(\sqrt{\xi} + \sqrt{\frac{1}{\xi K}} + \exp(-bT)\right),$$

where b *is a positive constant depending on* κ*,* L*, and* α*.*

Choosing $\xi = O(K^{-1/2})$ gives $O(K^{-1/4} + \exp(-bT))$ rate of $\min_{k \le K} \mathcal{K}^\diamond(v_k, \theta_k)$.

5. Experiment

Baselines A comprehensive set of state-of-the-art BO methods are chosen as baseline methods. This includes the *fully first-order* methods: BSG-1 [8] and BVFSM [19], ; a *stationary-seeking* method: Penalty [21], *explicit/implicit* methods: ITD [12], AID-CG (using conjugate gradient), AID-FP (using fixed point method) [10], reverse (using reverse auto-differentiation) [6] stocBiO [12], and VRBO [28].

5.1. Experiment Problems and Results

Toy Coreset Problem To validate the *convergence* property of BOME, we consider:

$$\min_{v,\theta} \|\theta - x_0\|^2 \quad s.t. \quad \theta \in \arg\min_\theta \|\theta - X\sigma(v)\|^2,$$

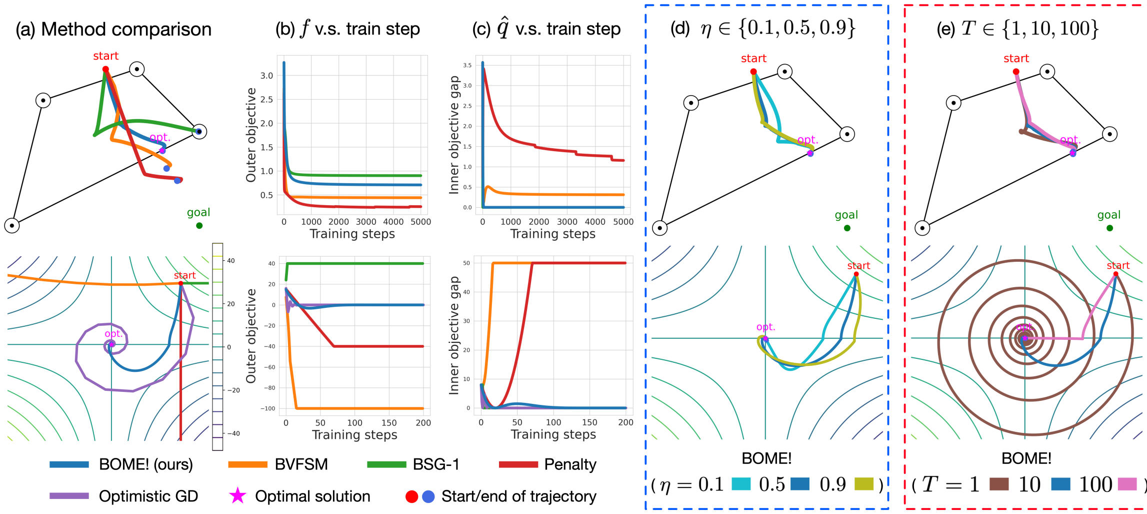

Figure 1: Results on the toy coreset problem and mini-max problem. (a)-(c): the trajectories of (v_k, θ_k), and $f(v_k, v_k)$ and $\hat{q}_k(v_k, v_k)$ of BOME (our method), BSG-1 [8], BVFSM [19], Penalty [21] and Optimistic GD [2] (only for minimax problem). (d)-(e) trajectories of BOME with different choices of inner gradient step T and the control coefficient η.

where $\sigma(v) = \exp(v)/\sum_{i=1}^4 \exp(v_i)$ is the softmax function, $v \in \mathbb{R}^4, \theta \in \mathbb{R}^2$, and $X = [x_1, x_2, x_3, x_4] \in \mathbb{R}^{2\times 4}$. The goal is to find the closest point to a target point x_0 within the convex hull of $\{x_1, \ldots, x_4\}$. See Fig. 5.1 (upper row) for the illustration and results.

Toy Mini-Max Game Mini-max game is a special and challenging case of BO where f and g contradicts with each other completely (e.g., $f = -g$). We consider

$$\min_{v,\theta\in\mathbb{R}} v\theta \quad s.t. \quad \theta \in \arg\max_{\theta'\in\mathbb{R}} v\theta'. \tag{10}$$

The optimal solution is $v^* = \theta^* = 0$. Note that the naive gradient descent ascent algorithm diverges to infinity on this problem, and a standard alternative is to use optimistic gradient descent [2]. Figure 5.1 shows that BOME works on this problem while other first-order BO methods fail.

Figure 5.1-3 show that BOME converges to the optimum of the corresponding bilevel problems and work well on the mini-max optimization and the degenerate low level problem. In comparison, the other methods like BSG-1, BVFSM, and Penalty fail to converge to the true optimum even with a grid search over their hyperparameters. Moreover, in both toy examples, BOME guarantees that $\hat{q}$, which is a proxy for the optimality of the inner problem, decreases to 0. Besides the standard step size ξ in typical optimizers, BOME only has three parameters: control coefficient η, inner loop iteration T, and inner step size α. We use the default setting of $\eta = 0.5$, $T = 10$ and $\alpha = \xi$ across the experiments. From Fig. 5.1 (d,e), BOME is robust to the choice of η, T and α as varying them results in almost identical performance. Specifically, $T = 1$ works well in many cases. The fact that BOME works well with a small T empirically makes it computationally attractive in practice.

6. Conclusion and Future Work

BOME, a simple fully first-order bilevel method, is proposed in this work with non-asymptotic convergence guarantee. While the current theory requires the inner loop iterations to scale in a logarithmic order w.r.t to the outer loop iterations, we do not observe this empirically. A further study to understand the mechanism is an interesting future direction.

References

[1] Jonathan F Bard. *Practical bilevel optimization: algorithms and applications*, volume 30. Springer Science & Business Media, 2013.

[2] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

[3] Stephan Dempe. *Foundations of bilevel programming*. Springer Science & Business Media, 2002.

[4] Stephan Dempe and Alain Zemkoho. *Bilevel optimization*. Springer, 2020.

[5] Nguyen Dinh, B Mordukhovich, and Tran TA Nghia. Subdifferentials of value functions and optimality conditions for dc and bilevel infinite and semi-infinite programs. *Mathematical Programming*, 123(1):101–138, 2010.

[6] Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, pages 1165–1173. PMLR, 2017.

[7] Spencer Frei and Quanquan Gu. Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. *Advances in Neural Information Processing Systems*, 34, 2021.

[8] Tommaso Giovannelli, Griffin Kent, and Luis Nunes Vicente. Bilevel stochastic methods for optimization and machine learning: Bilevel stochastic descent and darts. *arXiv preprint arXiv:2110.00604*, 2021.

[9] Chengyue Gong, Xingchao Liu, and Qiang Liu. Automatic and harmless regularization with constrained and lexicographic optimization: A dynamic barrier approach. *Advances in Neural Information Processing Systems*, 34, 2021.

[10] Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020.

[11] Robert Janin. Directional derivative of the marginal function in nonlinear programming. In *Sensitivity, Stability and Parametric Analysis*, pages 110–126. Springer, 1984.

[12] Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pages 4882–4892. PMLR, 2021.

[13] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[14] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257. PMLR, 2016.

[15] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 2022.

[16] Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *International Conference on Machine Learning*, pages 6305–6315. PMLR, 2020.

[17] Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *arXiv preprint arXiv:2101.11517*, 2021.

[18] Risheng Liu, Xuan Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A value-function-based interior-point method for non-convex bi-level optimization. *arXiv preprint arXiv:2106.07991*, 2021.

[19] Risheng Liu, Xuan Liu, Shangzhi Zeng, Jin Zhang, and Yixuan Zhang. Value-function-based sequential minimization for bi-level optimization. *arXiv preprint arXiv:2110.04974*, 2021.

[20] Risheng Liu, Yaohua Liu, Shangzhi Zeng, and Jin Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems*, 34, 2021.

[21] Akshay Mehra and Jihun Hamm. Penalty method for inversion-free deep bilevel optimization. In *Asian Conference on Machine Learning*, pages 347–362. PMLR, 2021.

[22] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Science & Business Media, second edition, 2006.

[23] Jivrí V Outrata. On the numerical solution of a class of stackelberg problems. *Zeitschrift für Operations Research*, 34(4):255–277, 1990.

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

[25] Michael Shub. *Global stability of dynamical systems*. Springer Science & Business Media, 2013.

[26] Chaehwan Song, Ali Ramezani-Kebrya, Thomas Pethick, Armin Eftekhari, and Volkan Cevher. Subquadratic overparameterization for shallow neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[27] Yann Traonmilin and Jean-Francois Aujol. The basins of attraction of the global minimizers of the non-convex sparse spike estimation problem. *Inverse Problems*, 36(4):045003, 2020.

[28] Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *arXiv preprint arXiv:2106.04692*, 2021.

[29] JJ Ye and DL Zhu. Optimality conditions for bilevel programming problems. *Optimization*, 33 (1):9–27, 1995.

Appendix A. Experiment Details

We provide details about each experiment in this section. Regarding the implementation of baseline methods:

- BVFSM's implementation is adapted from https://github.com/vis-opt-group/BVFSM.
- Penalty's implementation is adapted from https://github.com/jihunhamm/bilevel-penalty.
- VRBO's implementation is adapted from https://github.com/JunjieYang97/MRVRBO.
- AID-CG and AID-FP implementations are adapted from https://github.com/prolearner/hypertorch.
- ITD implementation is adapted from https://github.com/JunjieYang97/stocBiO.

A.1. Toy Coreset Problem

The problem is:

$$\min_{v,\theta} \|\theta - x_0\|^2 \quad s.t. \quad \theta \in \arg\min_{\theta} \|\theta - X\sigma(v)\|^2,$$

where $\sigma(v) = \exp(v)/\sum_{i=1}^{4}\exp(v_i)$ is the softmax function, $v \in \mathbb{R}^4, \theta \in \mathbb{R}^2$, and $X = [x_1, x_2, x_3, x_4] \in \mathbb{R}^{2\times 4}$. where $\sigma(v) = \exp(v)/\sum_{i=1}^{4}\exp(v_i)$ is the softmax function. Here the outer objective f pushes θ to towards x_0 while the inner objective g ensures θ remains in the convex hull formed by 4 points in the 2D plane (e.g. $X = [x_1, x_2, x_3, x_4] \in \mathbb{R}^{2\times 4}$). We choose $x_0 = (3, -2)$ and the four points $x_1 = (1, 3)$, $x_2 = (3, 1)$, $x_3 = (-2, 2)$ and $x_4 = (-3, 2)$. We set $v_0 = (0, 0, 0, 0)$ and $\theta_0 \in \{(0, 3), (-3, 1), (3.5, 1)\}$. For all methods, we fix both the inner stepsize α and the outer stepsize ξ to be 0.05 and set $T = 10$. For BVFSM and Penalty, we grid search the best hyperparameters from $\{0.001, 0.01, 0.1\}$. For BOME, we choose $\phi = \eta\,\|\nabla\hat{q}\|^2$ and ablate over $\eta \in \{0.1, 0.5, 0.9\}$ and $T \in \{1, 10, 100\}$. The visualization of the optimization trajectories over the 3 initial points are plotted in Fig. 2.

As shown, BOME successfully converges to the optimal solution regardless of the initial θ_0, while BSG-1, BVFSM and Penalty methods converge to non-optimal points. We emphasize that for BVFSM and Penalty, the convergence point *depends on* the choice of hyperparameters.

A.2. Toy Mini-max Game

The toy mini-max game we consider is:

$$\min_{v\in\mathbb{R}} \; v\theta^*(v) \quad s.t. \quad \theta^*(v) = \arg\max_{\theta\in\mathbb{R}} \; v\theta. \tag{11}$$

For BOME and BSG-1, BVFSM, and Penalty methods, we again set both the inner stepsize α and β to be 0.05, as no significant difference is observed by varying the stepsizes. For all methods, we set the inner iteration $T = 10$. For BVFSM and Penalty, we grid search the best hyperparameters from $\{0.001, 0.01, 0.1\}$.

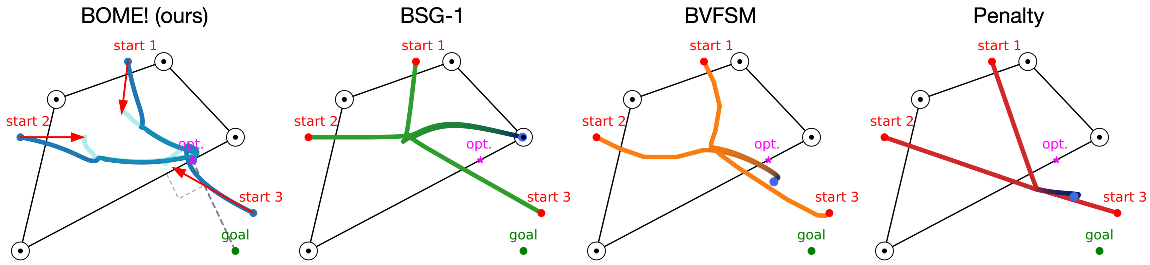

Figure 2: Trajectories of (v_k, θ_k) on the toy coreset problem (5.1) obtained from BOME (blue) and three recent first-order bilevel methods: BSG-1 [8] (green), BVFSM [19] (orange), and Penalty [21] (red). The goal of the problem is to find the closet point (marked by opt.) to the goal x_0 within the convex envelop of the four vertexes. All methods start from 3 initial points (start 1-3), and the converged points are shown in darkblue. For BOME, we also plot the trajectory of $\{\hat{\theta}_k^T\}$ in cyan.

A.3. Without LLS assumption

The toy example to validate whether BOME requires the low-level singleton assumption is borrowed from Liu et al. [16]:

$$\min_{v \in \mathbb{R}, \theta \in \mathbb{R}^2} \|\theta - [v; 1]\|_2^2 \quad s.t. \quad \theta \in \underset{(\theta_1', \theta_2') \in \mathbb{R}^2}{\arg\min} (\theta_1' - v)^2,$$

where $\theta = (\theta_1, \theta_2)$ and the optimal solution is $v^* = 1, \theta^* = (1, 1)$. Note that the inner objective has infinite many optimal solution $\theta^*(v)$ since it is degenerated. We set both the inner and outer stepsizes to 0.5 and $T = 10$ for all methods. For BVFSM and Penalty, we grid search the best hyperparameters from $\{0.001, 0.01, 0.1\}$. In Fig. 3, we provide the distance of $f(v_k, \theta_k), g(v_k, \theta_k), \theta_k, v_k$ to their corresponding optimal over training time in seconds. Note that BOME ensures that $\hat{q}(v_k, \theta_k) = g(v_k, \theta_k) - g(v^*, \theta^*)$ decreases to 0.

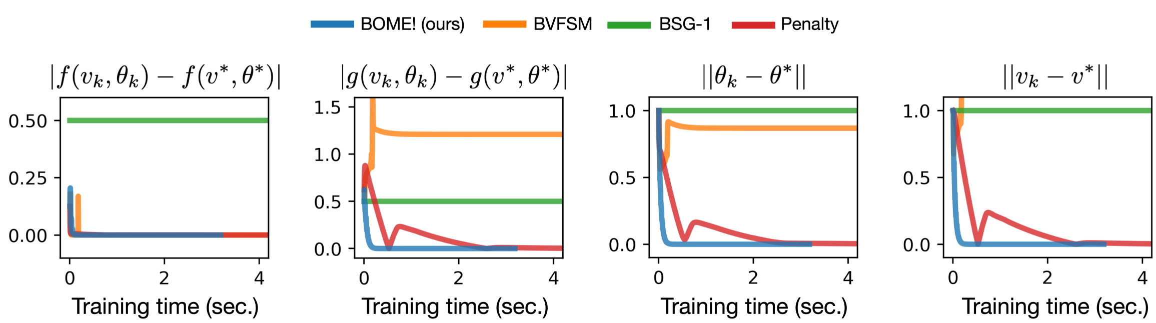

Figure 3: Results on Problem (A.3) which violates the low-level singleton (LLS). We compare BOME against BSG-1, BVFSM, and Penalty. (v^*, θ^*) denotes the true optimum and The four plots show how fast $f(v_k, \theta_k)$, $g(v_k, \theta_k)$, θ_k and v_k to the corresponding optimal values w.r.t. the training time in seconds.

Appendix B. More Theoretical Results

B.1. KKT Conditions

Consider a general constrained optimization of form $\min f(v, \theta)$ s.t. $q(v, \theta) \leq 0$. Under proper regularity conditions known as constraint quantifications [22], the first-order KKT condition gives a necessary condition for a feasible point (v^*, θ^*) with $q(v^*, \theta^*) \leq 0$ to be a local optimum of (4): There exists a Lagrangian multiplier $\lambda^* \in [0, +\infty)$, such that

$$\nabla f(v^*, \theta^*) + \lambda^* \nabla q(v^*, \theta^*) = 0, \tag{12}$$

and λ^* satisfies the complementary slackness condition $\lambda^* q(v^*, \theta^*) = 0$. A common regularity condition to ensure (12) is the *constant rank constraint quantification (CRCQ)* condition [11].

Definition 2 *A point (v^*, θ^*) is said to satisfy CRCQ with a function h if the rank of the Jacobian matrix $\nabla h(v, \theta)$ is constant in a neighborhood of (v^*, θ^*).*

Unfortunately, **the KKT condition in (12) does not hold for the bilevel optimization in (4).** The CRCQ condition does not typically hold for this problem. This is because the minimum of q is zero, and hence if (v^*, θ^*) is feasible for (4), then (v^*, θ^*) must attain the minimum of q, yielding $q(v^*, \theta^*) = 0$ and $\nabla q(v^*, \theta^*) = 0$ if q is smooth; but we could not have $\nabla q(v, \theta) = 0$ uniformly in a neighborhood of (v^*, θ^*) (hence CRCQ fails) unless q is a constant around (v^*, θ^*). In addition, if KKT (12) holds, we would have $\nabla f(v^*, \theta^*) = -\lambda^* \nabla q(v^*, \theta^*) = 0$ which happens only in the rare case when (v^*, θ^*) is a stationary point of both f, g.

Instead, one can establish a KKT condition of BO through the form in (3), because there is nothing special that prevents (v^*, θ^*) from satisfying CRCQ with $\nabla_\theta q = \nabla_\theta g$ (even though we just showed that it is difficult to have CRCQ with q). Assume f and $\nabla_\theta q$ are continuously differentiable, and (v^*, θ^*) is a point satisfying $\nabla_\theta q(v^*, \theta^*) = 0$ and CRCQ with $\nabla_\theta q$. Then by the typical first order KKT condition of (3), there exists a Lagrange multiplier $\omega^* \in \mathbb{R}^n$ such that

$$\nabla f(v^*, \theta^*) + \nabla(\nabla_\theta q(v^*, \theta^*))\omega^* = 0. \tag{13}$$

This condition can be viewed as the limit of a sequence of (12) in the following way: assume we relax the constraint in (4) to $q(v, \theta) \leq c_k$ where c_k is a sequence of positive numbers that converge to zero, then we can establish (12) for each $c_k > 0$ and pass the limit to zero to yield (13).

Proposition 2 *Assume that f, q, ∇q are continuously differentiable and $\|\nabla f\|$, f is bounded. For a feasible point (v^*, θ^*) of (4) that satisfies CRCQ with $\nabla_\theta q$, if (v^*, θ^*) is the limit of a sequence $\{(v_k, \theta_k)\}_{k=1}^\infty$ satisfying $q(v_k, \theta_k) \neq 0 \; \forall k$, and there exists a sequence $\{\lambda_k\} \subset [0, \infty)$ such that*

$$\nabla f(v_k, \theta_k) + \lambda_k \nabla q(v_k, \theta_k) \to 0, \qquad\qquad q(v_k, \theta_k) \to 0,$$

as $k \to +\infty$, then (v^*, θ^*) satisfies (13).

This motivates us to use the following function as a measure of stationarity of the solution returned by the algorithm:

$$\mathcal{K}(v, \theta) = \underbrace{\min_{\lambda \geq 0} ||\nabla f(v, \theta) + \lambda \nabla q(v, \theta)||^2}_{\text{local improvement}} + \underbrace{q(v, \theta)}_{\text{feasibility}}.$$

The hope is to have an algorithm that generates a sequence $\{(v_k, \theta_k)\}_{k=0}^{\infty}$ that satisfies $\mathcal{K}(v_k, \theta_k) \to 0$ as $k \to +\infty$.

Intuitively, the first term in $\mathcal{K}(v, \theta)$ measures how much ∇f conflicts with ∇q (how much we can decrease f without increasing q), as it is equal to the squared ℓ_2 norm of the solution to the problem $\min_\delta ||\nabla f - \delta||^2 \ s.t. \ \langle \nabla q, \delta \rangle \geq 0$. The second term in $\mathcal{K}$ measures how much the $\arg\min g$ constraint is satisfied.

B.2. Convergence with unimodal g

We first present the convergence rate when assuming $g(v, \cdot)$ has unique minimizer and satisfies the Polyak-Łojasiewicz (PL) inequality for all v, which guarantees a linear convergence rate of the gradient descent on the low level problem.

Assumption 4 (PL-inequality) *Given any v, assume $g(v, \cdot)$ has a unique minimizer denoted as $\theta^*(v)$. Also assume there exists $\kappa > 0$ such that for any (v, θ), $\|\nabla_\theta g(v, \theta)\|^2 \geq \kappa(g(v, \theta) - g(v, \theta^*(v)))$.*

The PL inequality gives a characterization on how a small gradient norm implies global optimality. It is implied from, but weaker than strongly convexity. The PL-inequality is more appealing than convexity because some modern over-parameterized deep neural networks have been shown to satisfy the PL-inequality along the trajectory of gradient descent. See, for example, Frei and Gu [7], Liu et al. [15], Song et al. [26] for more discussion.

Assumption 5 (Smoothness) *f and g are differentiable, and ∇f and ∇g are L-Lipschitz w.r.t. the joint inputs (v, θ) for some $L \in (0, +\infty)$.*

Assumption 6 (Boundedness) *There exists a constant $M < \infty$ such that $\|\nabla g(v, \theta)\|$, $\|\nabla f(v, \theta)\|$, $|f(v, \theta)|$ and $|g(v, \theta)|$ are all upper bounded by M for any (v, θ).*

Assumptions 5 and 6 are both standard in optimization.

Theorem 2 *With $\xi, \alpha \leq 1/L$, $\phi_k = \eta \|\nabla \hat{q}(v_k, \theta_k)\|^2$, and $\eta > 0$. Suppose that Assumptions 4, 5, and 6 hold. Then there exists a constant c depending on α, κ, η, L such that when $T \geq c$, we have for any $K \geq 0$,*

$$\min_{k \leq K} \mathcal{K}(v_k, \theta_k) = O\left(\sqrt{\xi} + \sqrt{\frac{q_0}{\xi K}} + \frac{1}{\xi K} + \exp(-bT)\right)$$

where $q_0 = q(v_0, \theta_0)$, and $b > 0$ is a constant depending on κ, L, and α.

Remark Note that one of the dominant terms depends on the initial value $q_0 = q(v_0, \theta_0)$. Therefore, we can obtain a better rate if we start from a θ_0 with small q_0 (hence near the optimum of $g(v_0, \cdot)$). In particular, when $q(v_0, \theta_0) = O(1)$, choosing $\xi = O(K^{-1/2})$ gives $\min_{k \leq K} \mathcal{K}(v_k, \theta_k) = O(K^{-1/4} + \exp(-bT))$ rate. On the other hand, if we start from a better initialization such that $q(v_0, \theta_0) = O((\xi K)^{-1})$, then choosing $\xi = O(K^{-2/3})$ gives $\min_{k \leq K} \mathcal{K}(v_k, \theta_k) = O(K^{-1/3} + \exp(-bT))$.

Appendix C. Proof of the Result in Section B.1

We proof Proposition 2 using Proposition 6.3 (presented below using our notation) in Gong et al. [9] by checking all the assumptions required by Proposition 6.3 in Gong et al. [9] are satisfied. Specifically, it remains to show that for any k, $\lambda_k < \infty$, $\lim_{k\to\infty} \nabla q(v_k, \theta_k) = 0$ and q is lower bounded (this is trivial as $q \geq 0$ by its definition), which we prove below.

Firstly, simple calculation shows that for any k,

$$\lambda_k = \left[\frac{-\langle \nabla f(v_k, \theta_k), \nabla q(v_k, \theta_k)\rangle}{||\nabla q(v_k, \theta_k)||^2}\right]_+ \leq \frac{\sup_{v,\theta} ||\nabla f(v, \theta)||}{||\nabla q(v_k, \theta_k)||} < \infty.$$

Here the last inequality is by $||\nabla q(v_k, \theta_k)|| > 0$. Secondly, note that as we assume ∇q is continuous, this implies that

$$\lim_{k\to\infty} \nabla q(v_k, \theta_k) = \nabla q(v^*, \theta^*).$$

As $q(v^*, \theta^*) = 0$, we have $\nabla q(v^*, \theta^*) = 0$. Using Proposition 6.3 in Gong et al. [9] gives the desired result.

[Proposition 6.3 in Gong et al. [9]] Assume $f, q, \nabla q$ are continuously differentiable. Let $\{[v_k, \theta_k, \lambda_k] : k = 1, 2, ...\}$ be a sequence which satisfies $\lim_{k\to\infty} ||\nabla q(v_k, \theta_k)|| = 0$ and $\lim_{k\to\infty} ||\nabla f(v_k, \theta_k) + \lambda_k \nabla q(v_k, \theta_k)|| = 0$. Assume that $[v^*, \theta^*]$ is a limit point of $[v_k, \theta_k]$ as $k \to \infty$ and $[v^*, \theta^*]$ satisfies CRCQ with $\nabla_\theta q$, then there exists a vector-valued Lagrange multiplier $\omega^* \in \mathbb{R}^m$ (the same length as θ) such that

$$\nabla f(v^*, \theta^*) + \nabla(\nabla_\theta q(v^*, \theta^*))\omega^* = 0.$$

Appendix D. Proof of the Result in Section B.2

We define $L_q := 2L(L/\kappa + 1)$ and using Assumption 5 and 4, we are able to show that $q(v, \theta)$ is L_q-smooth (see Lemma 4 for details). For simplicity, we also assume that $\xi \leq 1$ throughout the proof. We use b with some subscript to denote some general $O(1)$ constant and refer reader to section F for their detailed value.

Note that $\hat{q}$ defined in Section 3 changes in different iterations (as it depends on $\theta_k^{(T)}$) and so does $\nabla \hat{q}$. To avoid the confusion, we introduce several new notations. Firstly, given v and θ, $\theta^{(T)}$ denotes the results of T steps of gradient of $g(v, \cdot)$ w.r.t. θ starting from θ with step size α (similar to the definition in (8)). Note that $\theta^{(T)}$ depends on v, θ and α. Our notation does not reflects this dependency on v, α as we find it introduces no ambiguity while much simplifies the notation. Also note that when taking gradient on $\hat{q}$, the $\theta_k^{(T)}$ at iteration k is treated as a constant and the gradient does not pass through it. To be clear, we define $\hat{\nabla} q(v, \theta) = \nabla g(v, \theta) - \left[\nabla_1^\top g(v, \theta^{(T)}), \mathbf{0}^\top\right]^\top$, where $\mathbf{0}$ denotes a zero vector with the same dimension as θ. Using this definition, $\hat{\nabla} q(v_k, \theta_k) = \nabla \hat{q}(v_k, \theta_k)$ at iteration k. We also let $\lambda^*(v, \theta)$ be the solution of the dual problem of

$$\min_\delta ||\hat{\nabla} q(v, \theta) - \nabla f(v, \theta)||^2 \ s.t. \ \left\langle \hat{\nabla} q(v, \theta), \nabla f(v, \theta)\right\rangle \geq \eta ||\hat{\nabla} q(v, \theta)||^2. \quad (14)$$

That is

$$\lambda^*(v, \theta) = \begin{cases} \frac{[\eta ||\hat{\nabla} q(v,\theta)||^2 - \langle \hat{\nabla} q(v,\theta), \nabla f(v,\theta)\rangle]_+}{||\hat{\nabla} q(v,\theta)||^2} & \text{when } ||\hat{\nabla} q(v, \theta)|| > 0 \\ 0 & \text{when } ||\hat{\nabla} q(v, \theta)|| = 0 \end{cases} \quad (15)$$

We might use λ^* for $\lambda^*(v,\theta)$ when it introduces no confusion. Also, denote $\delta^*(v,\theta) = \lambda^*(v,\theta)\hat{\nabla} q(v,\theta) + \nabla f(v,\theta)$ and thus $\delta_k = \delta^*(v_k,\theta_k)$.

We start with several technical Lemmas showing some basic function properties.

D.1. Technical Lemmas

Lemma 1 *Under Assumption 4, for any v,θ, $g(v,\theta) - g(v,\theta^*(v)) \geq \frac{\kappa}{4}||\theta - \theta^*(v)||^2$.*

Lemma 2 *Under Assumption 4 and 5, we have $||\nabla q(v,\theta) - \hat{\nabla} q(v,\theta)|| \leq L||\theta^{(T)} - \theta^*(v)||$ for any v,θ. Also, when $||\hat{\nabla} q(v,\theta)|| = 0$, $q(v,\theta) = 0$.*

Lemma 3 *Under Assumption 5, 4, we have $||\theta^*(v_2) - \theta^*(v_1)|| \leq \frac{2L}{\kappa}||v_1 - v_2||$.*

Lemma 4 *Under Assumption 5, we have $||\nabla_\theta q(v,\theta_1) - \nabla_\theta q(v,\theta_2)|| \leq L||\theta_1 - \theta_2||$, for any v. Further assume Assumption 4, we have*

$$\|\nabla q(v_1,\theta_1) - \nabla q(v_2,\theta_2)\| \leq L_q||[v_1,\theta_1] - [v_2,\theta_2]||,$$

where $L_q := 2L(L/\kappa + 1)$.

Lemma 5 *Under Assumption 4, 5 and assume that $\alpha < 2/L$. Given any v,θ, suppose $\theta^{(0)} = \theta$ and $\theta^{(t+1)} = \theta^{(t)} - \alpha\nabla_\theta q(v,\theta^{(t)})$, then for any t, we have $q(v,\theta^{(t)}) \leq \exp(-b_1(\alpha,L,\kappa)t)q(v,\theta)$, where $b_1(\alpha,L,\kappa) =$ is some strictly positive constant that depends on α, L and κ.*

Lemma 6 *Under Assumption 6, for any $[v,\theta]$, we have $||\delta^*(v,\theta)||, ||\nabla q(v,\theta)||, ||\hat{\nabla} q(v,\theta)|| \leq b_2(M,\eta)$, where $b_2(M,\eta) = (3+\eta)M$.*

Lemma 7 *Under Assumption 6, for any $[v,\theta]$, we have $\lambda^*||\hat{\nabla} q||^2 \leq \eta||\hat{\nabla} q||^2 + M||\hat{\nabla} q||$,where λ^* are defined in (15).*

Lemma 8 *Under Assumption 4 and 5, we have $||\nabla q(v,\theta)|| \leq 2\kappa^{-1/2}L_q q^{1/2}(v,\theta)$.*

D.1.1. LEMMAS

Now we give several main lemmas that are used to prove the result in Section B.2.

Lemma 9 *Under Assumption 4, 5 and 6, when $||\hat{\nabla} q(v_k,\theta_k)|| > 0$, we have*

$$\begin{aligned} q(v_{k+1},\theta_{k+1}) - q(v_k,\theta_k) \leq -\xi\eta||\nabla q(v_k,\theta_k)||^2 + \xi\eta L_q||\theta_k^{(T)} - \theta^*(v_k)||\left(L_q||\theta_k^{(T)} - \theta^*(v_k)|| + 2L_q||\theta_k - \theta^*(v_k)||\right) \\ + \xi b_2 L||\theta_k^{(T)} - \theta^*(v_k)|| + L_q\xi^2 b_2^2/2. \end{aligned}$$

When $||\hat{\nabla} q(v_k,\theta_k)|| = 0$, we have $q(v_{k+1},\theta_{k+1}) - q(v_k,\theta_k) \leq \xi^2 L_q b_2^2/2$.

Lemma 10 *Under Assumption 4, 5 and 6, choosing $T \geq b_3(\eta,r,\kappa,L)$, we have*

$$q(v_k,\theta_k) \leq \exp(-b_4 k)q(v_0,\theta_0) + \Delta,$$

where $b_4 = -\log(1 - \frac{\xi}{4}\eta\kappa)$ is some strictly positive constant and $\Delta = O(\exp(-b_1 T) + \xi)$.

Lemma 11 *Under Assumption 4, 5 and 6, we have*

$$\sum_{k=0}^{K-1} ||\nabla q(v_k, \theta_k)||^2 \leq \frac{b_5 q(v_0, \theta_0)}{\xi} + K\xi^2 b_6 \Delta,$$

where b_5 is some constant depends on L_q, η, κ; b_6 is some constant depends on κ, L and Δ is defined in Lemma 10.

Lemma 12 *Under Assumption 4, 5 and 6, choosing $T \geq b_3(\eta, r, \kappa, L)$ and assume that $r, \xi \leq 1/L$, we have*

$$\sum_{k=0}^{K-1} \left[||\delta^*(v_k, \theta_k)||^2 + q(v_k, \theta_k)\right] = O(\xi^{-1} + K\exp(-b_1 T/2) + K\xi^{1/2} + \xi^{-1/2} K^{1/2} q^{1/2}(v_0, \theta_0)).$$

D.2. Proof of Theorem 2

Using our definition of λ^* in (15), we have

$$\begin{aligned}||\nabla f(v, \theta) + \lambda^*(v, \theta)\nabla q(v, \theta)|| &\leq ||\nabla f(v, \theta) + \lambda^*(v, \theta)\hat{\nabla} q(v, \theta)|| + ||\lambda^*(v, \theta)(\hat{\nabla} q(v, \theta) - \nabla q(v, \theta))|| \\ &= ||\delta^*(v, \theta)|| + ||\lambda^*(v, \theta)(\hat{\nabla} q(v, \theta) - \nabla q(v, \theta))||.\end{aligned}$$

Using Lemma 2, we know that when $||\hat{\nabla} q|| = 0$, we have $q = 0$ and thus $||\nabla q|| = 0$. In this case, $||\lambda^*(\hat{\nabla} q - \nabla q)|| = 0$. When $||\hat{\nabla} q|| > 0$, some algebra shows that

$$\begin{aligned}||\lambda^*(\hat{\nabla} q - \nabla q)|| &\leq \left[\eta - \left\langle \nabla f, \hat{\nabla} q/||\hat{\nabla} q||\right\rangle ||\hat{\nabla} q||^{-1}\right] ||\hat{\nabla} q - \nabla q|| \\ &\leq (\eta - \left\langle \nabla f, \hat{\nabla} q/||\hat{\nabla} q||\right\rangle ||\hat{\nabla} q||^{-1})||\hat{\nabla} q - \nabla q||.\end{aligned}$$

Notice that

$$\begin{aligned}||\hat{\nabla} q(v, \theta) - \nabla q(v, \theta)|| &\leq L||\theta^{(T)} - \theta^*(v)|| \\ &\leq 2L\kappa^{-1/2} q^{1/2}(v, \theta^{(T)}) \\ &\leq 2L\kappa^{-1/2}\exp(-b_1 T/2) q^{1/2}(v, \theta) \\ &\leq 2L\kappa^{-1}\exp(-b_1 T/2)||\nabla q(v, \theta)||.\end{aligned}$$

Here the first inequality is by Lemma 2, the second inequality is by Lemma 1, the third inequality is by Lemma 5 and the last inequality is by Assumption 4 (using $||\nabla q(v, \theta)|| \geq ||\nabla_\theta g(v, \theta)||$). Similarly, under assumption that $T \geq \lceil -b_1^{-1}\log(\frac{1}{16}\kappa^2 L^{-2})\rceil$, $L\kappa^{-1}\exp(-b_1 T/2) \leq 1/4$,

$$\begin{aligned}||\hat{\nabla} q(v, \theta)|| &= ||\hat{\nabla} q(v, \theta) - \nabla q(v, \theta) + \nabla q(v, \theta)|| \\ &\geq ||\nabla q(v, \theta)|| - ||\hat{\nabla} q(v, \theta) - \nabla q(v, \theta)|| \\ &\geq ||\nabla q(v, \theta)||(1 - (2L\kappa^{-1}\exp(-b_1 T/2))) \\ &\geq \frac{1}{2}||\nabla q(v, \theta)||.\end{aligned}$$

This implies that

$$\frac{||\hat{\nabla} q - \nabla q||}{||\hat{\nabla} q||} \leq 2\frac{||\hat{\nabla} q - \nabla q||}{||\nabla q||} \leq 4L\kappa^{-1}\exp(-b_1 T/2).$$

We thus have

$$\begin{aligned} ||\lambda^*(v,\theta)(\hat{\nabla} q(v,\theta) - \nabla q(v,\theta))|| &\leq \eta||\hat{\nabla} q - \nabla q|| + \left\langle \nabla f, \frac{\hat{\nabla} q}{||\hat{\nabla} q||}\right\rangle \frac{||\hat{\nabla} q - \nabla q||}{||\hat{\nabla} q||} \\ &\leq 2L\kappa^{-1}\exp(-b_1 T/2)\left[\eta||\nabla q(v,\theta)|| + 2\left\langle \nabla f, \frac{\hat{\nabla} q}{||\hat{\nabla} q||}\right\rangle\right] \\ &\leq 2L\kappa^{-1}\exp(-b_1 T/2)(\eta+2)b_2, \end{aligned}$$

where the last inequality is by Lemma 6. Combining all the results and using $||\nabla q(v_k,\theta_k)|| \leq 2\kappa^{-1/2}L_q q^{1/2}(v_k,\theta_k)$ by Lemma 8, we have

$$\begin{aligned} \mathcal{K}(v,\theta) &\leq ||\nabla f(v,\theta) + \lambda^*(v,\theta)\nabla q(v,\theta)||^2 + q(v,\theta) \\ &\leq 2||\nabla f(v,\theta) + \lambda^*(v,\theta)\hat{\nabla} q(v,\theta)||^2 + q(v,\theta) + 2||\lambda^*(v,\theta)(\hat{\nabla} q(v,\theta) - \nabla q(v,\theta))||^2 \\ &\leq 2||\delta^*(v,\theta)||^2 + q(v,\theta) + 8L^2\kappa^{-2}\exp(-b_1 T)(\eta+2)^2 b_2^2. \end{aligned}$$

Using Lemma 12, we have

$$\begin{aligned} \min_k \mathcal{K}(v_k,\theta_k) &= O(\min_k(||\delta^*(v_k,\theta_k)||^2 + q(v_k,\theta_k)) + \exp(-b_1 T)) \\ &= O(\xi^{-1} + K\exp(-b_1 T/2) + K\xi^{1/2} + \xi^{-1/2}K^{1/2}q^{1/2}(v_0,\theta_0)). \end{aligned}$$

D.3. Proof of Lemmas

D.3.1. PROOF OF LEMMA 9

When $||\hat{\nabla} q(v_k,\theta_k)|| > 0$, by Lemma 4, we know that q is L_q-smoothness, we have

$$\begin{aligned} q(v_{k+1},\theta_{k+1}) - q(v_k,\theta_k) &\leq -\xi\langle \nabla q(v_k,\theta_k), \delta^*(v_k,\theta_k)\rangle + \frac{L_q\xi^2}{2}||\delta^*(v_k,\theta_k)||^2 \\ &\leq -\xi\left\langle \hat{\nabla} q(v_k,\theta_k), \delta^*(v_k,\theta_k)\right\rangle - \xi\left\langle \nabla q(v_k,\theta_k) - \hat{\nabla} q(v_k,\theta_k), \delta^*(v_k,\theta_k)\right\rangle + L_q\xi^2 b_2^2/2 \\ &\leq -\xi\eta||\hat{\nabla} q(v_k,\theta_k)||^2 - \xi\left\langle \nabla q(v_k,\theta_k) - \hat{\nabla} q(v_k,\theta_k), \delta^*(v_k,\theta_k)\right\rangle + L_q\xi^2 b_2^2/2 \\ &\leq -\xi\eta||\hat{\nabla} q(v_k,\theta_k)||^2 + \xi b_2||\nabla q(v_k,\theta_k) - \hat{\nabla} q(v_k,\theta_k)|| + L_q\xi^2 b_2^2/2. \end{aligned}$$

where the second and the last inequality is by Lemma 6 and the third inequality is ensured by the constraint in the local subproblem ($\left\langle \nabla\hat{\nabla} q(v_k,\theta_k), \delta^*(v_k,\theta_k)\right\rangle \geq \eta||\hat{\nabla} q(v_k,\theta_k)^2||$.). And by Lemma 2, we have $||\nabla q(v_k,\theta_k) - \hat{\nabla} q(v_k,\theta_k)|| \leq L||\theta_k^{(T)} - \theta^*(v_k)||$. Plug in the bound we have

$$q(v_{k+1},\theta_{k+1}) - q(v_k,\theta_k) \leq -\xi\eta||\hat{\nabla} q(v_k,\theta_k)||^2 + \xi b_2 L||\theta_k^{(T)} - \theta^*(v_k)||\,.$$

Also notice that

$$\begin{aligned}\left|||\hat{\nabla} q(v_k,\theta_k)||^2 - ||\nabla q(v_k,\theta_k)||^2\right| &\leq ||\hat{\nabla} q(v_k,\theta_k) - \nabla q(v_k,\theta_k)||\,||\hat{\nabla} q(v_k,\theta_k) + \nabla q(v_k,\theta_k)|| \\ &\leq ||\hat{\nabla} q(v_k,\theta_k) - \nabla q(v_k,\theta_k)||\,(||\hat{\nabla} q(v_k,\theta_k) - \nabla q(v_k,\theta_k)|| + 2||\nabla q(v_k,\theta_k)||) \\ &\leq L_q||\theta_k^{(T)} - \theta^*(v_k)||\,(L_q||\theta_k^{(T)} - \theta^*(v_k)|| + 2||\nabla q(v_k,\theta_k)||) \\ &= L_q||\theta_k^{(T)} - \theta^*(v_k)||\,(L_q||\theta_k^{(T)} - \theta^*(v_k)|| + 2||\nabla q(v_k,\theta_k) - \nabla q(v_k,\theta^*(v_k))||) \\ &\leq L_q||\theta_k^{(T)} - \theta^*(v_k)||\,(L_q||\theta_k^{(T)} - \theta^*(v_k)|| + 2L_q||\theta_k - \theta^*(v_k)||),\end{aligned}$$

where the third inequality is by Lemma 2, the equality is by $\nabla q(v_k,\theta^*(v_k)) = 0$ and the last inequality is by Lemma 4.

Using this bound, we further have

$$\begin{aligned}q(v_{k+1},\theta_{k+1}) - q(v_k,\theta_k) &\leq -\xi\eta||\nabla q(v_k,\theta_k)||^2 + \xi\eta\left|||\hat{\nabla} q(v_k,\theta_k)||^2 - ||\nabla q(v_k,\theta_k)||^2\right| \\ &\quad + \xi b_2||\nabla q(v_k,\theta_k) - \hat{\nabla} q(v_k,\theta_k)|| + L_q\xi^2 b_2^2/2 \\ &\leq -\xi\eta||\nabla q(v_k,\theta_k)||^2 + \xi\eta L_q||\theta_k^{(T)} - \theta^*(v_k)||\,(L_q||\theta_k^{(T)} - \theta^*(v_k)|| + 2L_q||\theta_k - \theta^*(v_k)||) \\ &\quad + \xi b_2 L||\theta_k^{(T)} - \theta^*(v_k)|| + L_q\xi^2 b_2^2/2.\end{aligned}$$

When $||\hat{\nabla} q(v_k,\theta_k)|| = 0$, by Lemma 2, $q(v_k,\theta_k) = 0$ and hence $\nabla q(v_k,\theta_k) = 0$. We thus have

$$\begin{aligned}q(v_{k+1},\theta_{k+1}) - q(v_k,\theta_k) &\leq -\xi\left\langle \nabla q(v_k,\theta_k), \delta^*(v_k,\theta_k)\right\rangle + \frac{L_q\xi^2}{2}||\delta^*(v_k,\theta_k)||^2 \\ &= \frac{L_q\xi^2}{2}||\delta^*(v_k,\theta_k)||^2 \\ &\leq \xi^2 L_q b_2^2/2.\end{aligned}$$

D.3.2. PROOF OF LEMMA 10

By Lemma 9, when $||\hat{\nabla} q(v_k,\theta_k)|| > 0$, we have

$$\begin{aligned}q(v_{k+1},\theta_{k+1}) - q(v_k,\theta_k) &\leq -\xi\eta||\nabla q(v_k,\theta_k)||^2 + \xi\eta L_q||\theta_k^{(T)} - \theta^*(v_k)||\,(L_q||\theta_k^{(T)} - \theta^*(v_k)|| + 2L_q||\theta_k - \theta^*(v_k)||) \\ &\quad + \xi b_2 L||\theta_k^{(T)} - \theta^*(v_k)|| + L_q\xi^2 b_2^2/2.\end{aligned}$$

By Lemma 1 and Lemma 5

$$\begin{aligned}||\theta_k^{(T)} - \theta^*(v_k)|| &\leq 2\kappa^{-1/2} q^{1/2}(v_k,\theta_k^{(T)}) \leq 2\kappa^{-1/2}\exp(-b_1 T/2) q^{1/2}(v_k,\theta_k). \\ ||\theta_k - \theta^*(v_k)|| &\leq 2\kappa^{-1/2} q^{1/2}(v_k,\theta_k).\end{aligned}$$

Using those bounds, we know that

$$L_q||\theta_k^{(T)} - \theta^*(v_k)||\,(L_q||\theta_k^{(T)} - \theta^*(v_k)|| + 2L_q||\theta_k - \theta^*(v_k)||) \leq 12L_q^2\kappa^{-1}\exp(-b_1 T) q(v_k,\theta_k)$$

This implies that

$$\begin{aligned}&q(v_{k+1},\theta_{k+1})-q(v_k,\theta_k)\\ \leq&-\xi\eta||\nabla q(v_k,\theta_k)||^2+12\xi\eta L_q^2\kappa^{-1}\exp(-b_1T)q(v_k,\theta_k)\\ &+2\xi b_2L\kappa^{-1/2}\exp(-b_1T/2)q^{1/2}(v_k,\theta_k)+L_q\xi^2b_2^2/2\\ \leq&-\xi\eta\kappa q(v_k,\theta_k)+12\xi\eta L_q^2\kappa^{-1}\exp(-b_1T)q(v_k,\theta_k)\\ &+2\xi b_2L\kappa^{-1/2}\exp(-b_1T/2)q^{1/2}(v_k,\theta_k)+L_q\xi^2b_2^2/2.\end{aligned}$$

Choosing T such that $T \geq b_3(\eta,\alpha,\kappa,L)$ where

$$b_3(\eta,\alpha,\kappa,L)=\left\lceil -b_1^{-1}\log(\frac{\eta\kappa}{64\eta L_q^2})\right\rceil,$$

we have

$$q(v_{k+1},\theta_{k+1})-q(v_k,\theta_k)\leq-\frac{3}{4}\xi\eta\kappa q(v_k,\theta_k)+2\xi b_2L\kappa^{-1/2}\exp(-b_1T/2)q^{1/2}(v_k,\theta_k)+L_q\xi^2b_2^2/2.$$

This implies that when $\frac{64b_2^2L^2}{\eta^2\kappa}\exp(-b_1T)\leq q(v_k,\theta_k)$ and $\frac{2L_q\xi b_2^2}{\eta\kappa}\leq q(v_k,\theta_k)$,

$$q(v_{k+1},\theta_{k+1})-q(v_k,\theta_k)\leq-\frac{1}{4}\xi\eta\kappa q(v_k,\theta_k).$$

Let $a=\max(\frac{64b_2^2L^2}{\eta^2\kappa}\exp(-b_1T),\frac{2L_q\xi b_2^2}{\eta\kappa})$. Also, when $q(v_k,\theta_k)<a$,

$$\begin{aligned}q(v_{k+1},\theta_{k+1})&\leq q(v_k,\theta_k)+2\xi b_2L\kappa^{-1/2}\exp(-b_1T/2)q^{1/2}(v_k,\theta_k)+L_q\xi^2b_2^2/2\\ &<a+2\xi b_2L\kappa^{-1/2}\exp(-b_1T/2)\sqrt{a}+L_q\xi^2b_2^2/2.\end{aligned}$$

Note that

$$\begin{aligned}2\xi b_2L\kappa^{-1/2}\exp(-b_1T/2)&\leq\frac{\xi\eta\kappa}{4}\sqrt{a}\\ L_q\xi^2b_2^2/2&\leq\frac{\xi\eta\kappa}{4}a.\end{aligned}$$

This gives that in the case of $q(v_k,\theta_k)<a$,

$$q(v_{k+1},\theta_{k+1})<(1+\frac{\xi\eta\kappa}{4})a.$$

Define k_0 as the first iteration such that $q(v_k,\theta_k)<a$. This implies that, for any $k\leq k_0$,

$$q(v_k,\theta_k)\leq(1-\frac{\xi}{4}\eta\kappa)^kq(v_0,\theta_0).$$

When any $k>k_0$, we show that $q(v_{k+1},\theta_{k+1})\leq(1+\frac{\xi\eta\kappa}{4})a$. This can be proved by induction. At $k=k_0+1$, if $q(v_k,\theta_k)<a$, we have $q(v_k,\theta_k)<(1+\frac{\xi\eta\kappa}{4})a$. Else if at $k=k_0+1$,

$q(v_k, \theta_k) \geq a$, $q(v_{k+1}, \theta_{k+1}) \leq q(v_k, \theta_k) \leq a$. We thus have the conclusion that for any $k > k_0$, $q(v_k, \theta_k) \leq (1 + \frac{\eta\kappa}{4})a$. Combining the result, we have

$$q(v_k, \theta_k) \leq (1 - \frac{\xi}{4}\eta\kappa)^k q(v_0, \theta_0) + \Delta,$$

where we denote

$$\Delta = (1 + \frac{\eta\kappa}{4})(\frac{64b_2^2L^2}{\eta^2\kappa^3}\exp(-b_1T) + \frac{2L_q\xi b_2^2}{\eta\kappa}) + L_q\xi^2 b_2^2/2 = O(\exp(-b_1T) + \xi). \tag{16}$$

Let $b_4(\eta, \kappa, \xi) = -\log(1 - \frac{\xi}{4}\eta\kappa)$, we have the desired result.

D.3.3. PROOF OF LEMMA 11

By Lemma 8 and 10, we have

$$\begin{aligned} ||\nabla q(v_k, \theta_k)||^2 &\leq 2\kappa^{-1}L_q^2 q(v_k, \theta_k) \\ &\leq 2\kappa^{-1}L_q^2\left[\exp(-b_4k)q(v_0, \theta_0) + \Delta\right], \end{aligned}$$

where Δ is defined in (16). Also notice that

$$\begin{aligned} ||\hat{\nabla} q(v, \theta)|| &\leq ||\hat{\nabla} q(v, \theta) - \nabla q(v, \theta)|| + ||\nabla q(v, \theta)|| \\ &\leq L||\theta^{(T)} - \theta^*(v)|| + ||\nabla q(v, \theta)|| \\ &\leq 2L\kappa^{-1/2}q^{1/2}(v, \theta^{(T)}) + ||\nabla q(v, \theta)|| \\ &\leq 2L\kappa^{-1/2}\exp(-b_1T/2)q^{1/2}(v, \theta) + ||\nabla q(v, \theta)|| \\ &\leq (2L\kappa^{-1}\exp(-b_1T/2) + 1)||\nabla q(v, \theta)|| \\ &\leq (2L\kappa^{-1} + 1)||\nabla q(v, \theta)|| \end{aligned}$$

Here the first inequality is by triangle inequality, the second inequality is by Lemma 2, the third inequality is by Lemma 1, the forth inequality is by Lemma 5 and the fifth inequality is by Assumption 4. Taking summation over iteration and using Lemma 10, we have

$$\begin{aligned} \sum_{k=0}^{K-1} ||\hat{\nabla} q(v, \theta)||^2 &\leq (2L\kappa^{-1} + 1)^2 \sum_{k=0}^{K-1} ||\nabla q(v_k, \theta_k)||^2 \\ &\leq (2L\kappa^{-1} + 1)^2 \left[2\kappa^{-1}L_q^2 q(v_0, \theta_0)\sum_{k=0}^{K-1}[\exp(-b_4k)] + K\Delta\right] \\ &\leq (2L\kappa^{-1} + 1)^2 \left[\frac{2\kappa^{-1}L_q^2 q(v_0, \theta_0)}{1 - \exp(-b_4)} + K\Delta\right] \\ &= \frac{b_5 q(v_0, \theta_0)}{\xi} + Kb_6\Delta, \end{aligned}$$

where we define $b_5(L_q, \eta, \kappa) = \frac{16L_q^2}{\eta\kappa^2}(2L\kappa^{-1} + 1)^2$ and $b_6(\kappa, L) = (2L\kappa^{-1} + 1)^2$.

D.4. Proof of Lemma 12

Remind that by our definition of λ^* in (15) and Assumption 5, we have

$$\begin{aligned}
f(v_{k+1},\theta_{k+1}) - f(v_k,\theta_k) &\le -\xi \left\langle \nabla f(v_k,\theta_k), \delta^*(v_k,\theta_k)\right\rangle + \frac{L\xi^2}{2}||\delta^*(v_k,\theta_k)||^2 \\
&= -\xi \left\langle \delta^*(v_k,\theta_k) - \lambda^*(v_k,\theta_k)\hat{\nabla} q(v_k,\theta_k), \delta^*(v_k,\theta_k)\right\rangle + \frac{L\xi^2}{2}||\delta^*(v_k,\theta_k)||^2 \\
&= -(\xi - \frac{L\xi^2}{2})||\delta^*(v_k,\theta_k)||^2 + \xi\lambda^*(v_k,\theta_k)\left\langle \hat{\nabla} q(v_k,\theta_k), \delta^*(v_k,\theta_k)\right\rangle \\
&\le -(\xi - \frac{L\xi^2}{2})||\delta^*(v_k,\theta_k)||^2 + \xi\eta\lambda^*(v_k,\theta_k)||\hat{\nabla} q(v_k,\theta_k)||^2 \\
&\le -\frac{\xi}{2}||\delta^*(v_k,\theta_k)||^2 + \xi\eta\lambda^*(v_k,\theta_k)||\hat{\nabla} q(v_k,\theta_k)||^2,
\end{aligned}$$

where the last inequality is by the assumption on $\xi \le 1/L$. To show the second inequality, we use the complementary slackness of Problem (14), that is

$$\lambda^*(v_k,\theta_k)\left[\left\langle \hat{\nabla} q(v_k,\theta_k), \delta^*(v_k,\theta_k)\right\rangle - \eta||\hat{\nabla} q(v_k,\theta_k)||\right] = 0.$$

By telescoping,

$$\begin{aligned}
\sum_{k=0}^{K-1} f(v_{k+1},\theta_{k+1}) - f(v_k,\theta_k) &\le -\frac{\xi}{2}\sum_{k=0}^{K-1}||\delta^*(v_k,\theta_k)||^2 + \xi\eta\sum_{k=0}^{K-1}\lambda^*(v_k,\theta_k)||\hat{\nabla} q(v_k,\theta_k)||^2 \\
&\le -\frac{\xi}{2}\sum_{k=0}^{K-1}||\delta^*(v_k,\theta_k)||^2 + \xi\eta\sum_{k=0}^{K-1}(\eta||\hat{\nabla} q(v_k,\theta_k)||^2 + M||\hat{\nabla} q(v_k,\theta_k)||) \\
&= -\frac{\xi}{2}\sum_{k=0}^{K-1}||\delta^*(v_k,\theta_k)||^2 + \xi\eta^2\sum_{k=0}^{K-1}||\hat{\nabla} q(v_k,\theta_k)||^2 + \xi\eta M\sum_{k=0}^{K-1}||\hat{\nabla} q(v_k,\theta_k)|| \\
&\le -\frac{\xi}{2}\sum_{k=0}^{K-1}||\delta^*(v_k,\theta_k)||^2 + \xi\eta^2\sum_{k=0}^{K-1}||\hat{\nabla} q(v_k,\theta_k)||^2 + \xi\eta M\sqrt{K}\sqrt{\sum_{k=0}^{K-1}||\hat{\nabla} q(v_k,\theta_k)||^2},
\end{aligned}$$

where the second inequality is by Lemma 7 and the last inequality is by Holder's inequality. Since $\sum_{k=0}^{K-1} f(v_{k+1},\theta_{k+1}) - f(v_k,\theta_k) = f(v_K,\theta_K) - f(v_0,\theta_0)$, rearrange the terms, we have

$$\xi\sum_{k=0}^{K-1}||\delta^*(v_k,\theta_k)||^2 \le 2(f(v_0,\theta_0) - f(v_K,\theta_K)) + 2\xi\eta^2\sum_{k=0}^{K-1}||\hat{\nabla} q(v_k,\theta_k)||^2 + 2\xi\eta M\sqrt{K}\sqrt{\sum_{k=0}^{K-1}||\hat{\nabla} q(v_k,\theta_k)||^2}.$$

This implies that

$$\begin{aligned}
\xi\sum_{k=0}^{K-1}\left[||\delta^*(v_k,\theta_k)||^2 + q(v_k,\theta_k)\right] &\le 2(f(v_0,\theta_0) - f(v_K,\theta_K)) + 2\xi\eta^2\sum_{k=0}^{K-1}||\hat{\nabla} q(v_k,\theta_k)||^2 \\
&\quad + 2\xi\eta M\sqrt{K}\sqrt{\sum_{k=0}^{K-1}||\hat{\nabla} q(v_k,\theta_k)||^2} + \xi\sum_{k=0}^{K-1} q(v_k,\theta_k).
\end{aligned}$$

Using Lemma 10, we know that

$$q(v_k, \theta_k) \leq (1 - \frac{\xi}{4}\eta\kappa)^k q(v_0, \theta_0) + \Delta.$$

This gives that

$$\xi \sum_{k=0}^{K-1} q(v_k, \theta_k) \leq \frac{4q(v_0, \theta_0)}{\eta\kappa} + \xi K \Delta.$$

Using Lemma 11, 10 and $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, we have

$$2\xi\eta^2 \sum_{k=1}^{K} ||\hat{\nabla} q(v_k, \theta_k)||^2 \leq 2\eta^2 b_5 q(v_0, \theta_0) + 2K\eta^2 \xi b_6 \Delta$$

$$2\xi\eta M\sqrt{K}\sqrt{\sum_{k=1}^{K} ||\hat{\nabla} q(v_k, \theta_k)||^2} \leq 2\xi^{1/2} K^{1/2} b_5^{1/2} \eta M q^{1/2}(v_0, \theta_0) + 2K\xi b_6^{1/2} \eta M \Delta^{1/2}$$

This implies that

$$\begin{aligned}
&\xi \sum_{k=0}^{K-1} \left[||\delta^*(v_k, \theta_k)||^2 + q(v_k, \theta_k)\right] \\
\leq& 2(f(v_0, \theta_0) - f(v_K, \theta_K)) + 2\eta^2 b_5 q(v_0, \theta_0) + 2K\eta^2 \xi b_6 \Delta + 2\xi^{1/2} K^{1/2} b_5^{1/2} \eta M q^{1/2}(v_0, \theta_0) \\
&+ 2K\xi b_6^{1/2} \eta M \Delta^{1/2} + \frac{4q(v_0, \theta_0)}{\eta\kappa} + \xi K \Delta \\
\leq& 2(f(v_0, \theta_0) - f(v_K, \theta_K)) + (2\eta^2 b_5 + \frac{4}{\eta\kappa}) q(v_0, \theta_0) + 2K\xi(b_6^{1/2} \eta M \Delta^{1/2} + (b_6\eta^2 + 1/2)\Delta) \\
&+ 2\xi^{1/2} K^{1/2} b_5^{1/2} \eta M q^{1/2}(v_0, \theta_0)
\end{aligned}$$

We thus have

$$\begin{aligned}
&\sum_{k=0}^{K-1} \left[||\delta^*(v_k, \theta_k)||^2 + q(v_k, \theta_k)\right] \\
=& O(\xi^{-1} + K\Delta^{1/2} + \xi^{-1/2} K^{1/2} q^{1/2}(v_0, \theta_0)) \\
=& O(\xi^{-1} + K \exp(-b_1 T/2) + K\xi^{1/2} + \xi^{-1/2} K^{1/2} q^{1/2}(v_0, \theta_0)).
\end{aligned}$$

D.5. Proofs of Technical Lemmas

D.5.1. PROOF OF LEMMA 1

Please see the proof of Theorem 2 in Karimi et al. [13].

D.5.2. PROOF OF LEMMA 2

Since $\nabla_2 g(v,\theta^*(v)) = 0$, we have $\nabla_v g(v,\theta^*(v)) = \nabla_1 g(v,\theta^*(v)) + \nabla_v \theta^*(v) \nabla_2 g(v,\theta^*(v)) = \nabla_1 g(v,\theta^*(v))$. Thus

$$\nabla q(v,\theta) = \begin{bmatrix} \nabla_v g(v,\theta) - \nabla_v g(v,\theta^*(v)) \\ \nabla_\theta g(v,\theta) \end{bmatrix} = \begin{bmatrix} \nabla_v g(v,\theta) - \nabla_1 g(v,\theta^*(v)) \\ \nabla_\theta g(v,\theta) \end{bmatrix}.$$

Also note that

$$\hat{\nabla} q(v,\theta) = \begin{bmatrix} \nabla_v g(v,\theta) - \nabla_1 g(v,\theta^{(T)}) \\ \nabla_\theta g(v,\theta) \end{bmatrix}.$$

This gives that

$$\begin{aligned} ||\nabla q(v,\theta) - \hat{\nabla} q(v,\theta)|| &= ||\nabla_1 g(v,\theta^{(T)}) - \nabla_1 g(v,\theta^*(v))|| \\ &\le L||\theta^{(T)} - \theta^*(v)||. \end{aligned}$$

Also when $0 = ||\hat{\nabla} q(v,\theta)|| = \sqrt{||\nabla_v g(v,\theta) - \nabla_1 g(v,\theta^{(T)})||^2 + ||\nabla_\theta g(v,\theta)||^2}$, we have $||\nabla_\theta g(v,\theta)|| = 0$. Under Assumption 4,

$$0 = ||\nabla_\theta g(v,\theta)|| \ge \kappa(g(v,\theta) - g(v,\theta^*(v))) = \kappa q(v,\theta).$$

D.5.3. PROOF OF LEMMA 3

Using Assumption 4 and $\nabla_2 g(v_1,\theta^*(v_1)) = 0$, we have

$$||\nabla_2 g(v_1,\theta^*(v_2))|| \ge \sqrt{\kappa(g(v_1,\theta^*(v_2)) - g(v_1,\theta^*(v_1))}.$$

Also by Lemma 1, we have $g(v_1,\theta^*(v_2)) - g(v_1,\theta^*(v_1) \ge \frac{1}{4}\kappa||\theta^*(v_2) - \theta^*(v_1)||^2$. These imply that

$$||\nabla_2 g(v_1,\theta^*(v_2))|| \ge \frac{1}{2}\kappa||\theta^*(v_2) - \theta^*(v_1)||.$$

Also

$$\begin{aligned} &||\nabla_2 g(v_1,\theta^*(v_2))|| \\ =&||\nabla_2 g(v_1,\theta^*(v_2)) - \nabla_\theta g(v_2,\theta^*(v_2))|| \\ =&||\nabla_2 [g(v_1,\theta^*(v_2)) - g(v_2,\theta^*(v_2))]|| \\ \le&||\nabla_{[1,2]} [g(v_1,\theta^*(v_2)) - g(v_2,\theta^*(v_2))]|| \\ \le&L||v_1 - v_2||, \end{aligned}$$

where $\nabla_{[1,2]}$ denotes taking the derivative on both first and second variables. We thus conclude that

$$||\theta^*(v_2) - \theta^*(v_1)|| \le \frac{2L}{\kappa}||v_1 - v_2||.$$

D.5.4. PROOF OF LEMMA 4

To prove the first property,

$$
\begin{aligned}
||\nabla_\theta q(v,\theta_1) - \nabla_\theta q(v,\theta_2)|| &= ||\nabla_\theta g(v,\theta_1) - \nabla_\theta g(v,\theta_2)|| \\
&\leq L||\theta_1 - \theta_2||.
\end{aligned}
$$

Also

$$
\begin{aligned}
\|\nabla q(v_1,\theta_1) - \nabla q(v_2,\theta_2)\| &= \|\nabla g(v_1,\theta_1) - \nabla g(v_2,\theta_2) - \nabla g(v_1,\theta^*(v_1)) + \nabla g(v_2,\theta^*(v_2))\| \\
&\leq \|\nabla g(v_1,\theta_1) - \nabla g(v_2,\theta_2)\| + \|\nabla_1 g(v_1,\theta^*(v_1)) - \nabla_1 g(v_2,\theta^*(v_2))\| .
\end{aligned}
$$

By Assumption 5 (Lipschitz continuity of ∇g),

$$
\begin{aligned}
||\nabla_1 g(v_1,\theta^*(v_1)) - \nabla_1 g(v_2,\theta^*(v_2))|| &\leq ||\nabla_{[1,2]} g(v_1,\theta^*(v_1)) - \nabla_{[1,2]} g(v_2,\theta^*(v_2))|| \\
&\leq L\sqrt{||\theta^*(v_1) - \theta^*(v_2)||^2 + ||v_1 - v_2||^2},
\end{aligned}
$$

where $\nabla_{[1,2]}$ denotes taking the derivative on both first and second variable. Also By Lemma 3,

$$
\begin{aligned}
L\sqrt{||\theta^*(v_1) - \theta^*(v_2)||^2 + ||v_1 - v_2||^2} &\leq L\sqrt{\frac{4L^2}{\kappa^2}||v_1 - v_2||^2 + ||v_1 - v_2||^2} \\
&\leq L(\frac{2L}{\kappa} + 1)||v_1 - v_2||.
\end{aligned}
$$

This gives that

$$
\begin{aligned}
\|\nabla q(v_1,\theta_1) - \nabla q(v_2,\theta_2)\| &\leq \|\nabla g(v_1,\theta_1) - \nabla g(v_2,\theta_2)\| + \|\nabla_1 g(v_1,\theta^*(v_1)) - \nabla_1 g(v_2,\theta^*(v_2))\| \\
&\leq L\sqrt{||v_1 - v_2||^2 + ||\theta_1 - \theta_2||^2} + \|\nabla_1 g(v_1,\theta^*(v_1)) - \nabla_1 g(v_2,\theta^*(v_2))\| \\
&\leq L\sqrt{||v_1 - v_2||^2 + ||\theta_1 - \theta_2||^2} + L(\frac{2L}{\kappa} + 1)||v_1 - v_2|| \\
&\leq L_q\sqrt{||v_1 - v_2||^2 + ||\theta_1 - \theta_2||^2},
\end{aligned}
$$

where $L_q := 2L(L/\kappa + 1)$.

D.5.5. PROOF OF LEMMA 5

By Lemma 4, we have

$$
q(v,\theta^{(t+1)}) - q(v,\theta^{(t)}) \leq -(\alpha - \frac{L\alpha^2}{2})||\nabla_\theta q(v,\theta^{(t)})||^2.
$$

By Assumption 4, we have

$$
||\nabla_\theta q(v,\theta^{(t)})||^2 = ||\nabla_2 g(v,\theta^{(t)})||^2 \geq \kappa(g(v,\theta^{(t)}) - g(v,\theta^*(v)) = \kappa q(v,\theta^{(t)}).
$$

Plug-in, we have

$$
q(v,\theta^{(t+1)}) \leq (1 - (\alpha - \frac{L\alpha^2}{2})\kappa)q(v,\theta^{(t)}).
$$

Recursively apply this inequality, we have

$$
q(v,\theta^{(t)}) \leq (1 - (\alpha - \frac{L\alpha^2}{2})\kappa)^t q(v,\theta).
$$

Let $b_1(r,L,\kappa) = \log(1 - (\alpha - L\alpha^2/2)\kappa)$, we have the desired result.

D.5.6. PROOF OF LEMMA 6

Notice that $||\nabla q(v,\theta)|| \leq ||\nabla g(v,\theta)|| + ||\nabla g(v,\theta^*(v))|| \leq 2M$. $||\hat{\nabla} q(v,\theta)|| \leq ||\nabla_v g(v,\theta)|| + ||\nabla_1 g(v,\theta^{(T)})|| + ||\nabla_\theta g(v,\theta)|| \leq 3M$. When $||\hat{\nabla} q|| = 0$, $||\delta^*|| = ||\nabla f|| \leq M$. When $||\hat{\nabla} q|| > 0$,

$$\begin{aligned} ||\delta^*|| &= ||[\eta||\hat{\nabla} q||^2 - \left\langle \nabla f, \hat{\nabla} q \right\rangle]_+ / ||\hat{\nabla} q||^2 \hat{\nabla} q + \nabla f|| \\ &\leq \eta||\hat{\nabla} q|| + 2||\nabla f|| \leq (2+\eta)M. \end{aligned}$$

This concludes that $||\delta^*|| \leq (2+\eta)M$.

D.5.7. PROOF OF LEMMA 7

In the case that $\left\langle \nabla f, \hat{\nabla} q \right\rangle < \eta||\hat{\nabla} q||^2$, $\lambda^*||\hat{\nabla} q||^2 = \eta||\hat{\nabla} q||^2 - \left\langle \nabla f, \hat{\nabla} q \right\rangle$. In the other case, $\lambda^*||\hat{\nabla} q||^2 = 0$. Thus in all cases,

$$\begin{aligned} \lambda^*||\hat{\nabla} q||^2 &\leq \eta||\hat{\nabla} q||^2 + ||\nabla f||\, ||\hat{\nabla} q|| \\ &\leq \eta||\hat{\nabla} q||^2 + M||\hat{\nabla} q||. \end{aligned}$$

D.5.8. PROOF OF LEMMA 8

Notice that since $\nabla q(v,\theta^*(v)) = 0$, we have

$$||\nabla q(v,\theta)|| = ||\nabla q(v,\theta) - \nabla q(v,\theta^*(v))|| \leq L_q||\theta - \theta^*(v)|| \leq 2\kappa^{-1/2} L_q q^{1/2}(v,\theta),$$

where the first inequality is by Lemma 4 and the second inequality is by Lemma 1.

Appendix E. Proof of the Result in Section 4.1

We use b with some subscript to denote some general $O(1)$ constant and refer reader to section F for their detailed value.

For notation simplicity, given v and θ, $\theta^{(T)}$ denotes the results of T steps of gradient of $g(v,\cdot)$ w.r.t. θ starting from θ using step size α (similar to the definition in (8)). And note that $\hat{\nabla} q(v,\theta) = \nabla g(v,\theta) - \left[\nabla_1^\top g(v,\theta^{(T)}), \mathbf{0}^\top\right]^\top$, where $\mathbf{0}$ denotes a zero vector with the same dimension as θ. We refer readers to the beginning of Appendix D for a discussion on the design of this extra notation and how it relates to the notation we used in Section 3. For simplicity, we omit the superscript $\diamond$ in $q^\diamond$ and simply use q to denote $q^\diamond$ in the proof.

We start with the following two Lemmas.

Lemma 13 *Under Assumption 3 and assume $\alpha \leq 1/L$, for any v,θ, $g(v,\theta) - g(v,\theta^\diamond(v,\theta)) \geq \frac{\kappa}{4}||\theta - \theta^\diamond(v,\theta)||^2$.*

Proof It is easy to show that

$$g(v,\theta^{(t+1)}) \leq g(v,\theta^{(t)}) - (\alpha - \frac{L\alpha^2}{2})||\nabla_\theta g(v,\theta^{(t)})||^2 \leq g(v,\theta^{(t)}).$$

We thus have $g(v,\theta^\diamond(v,\theta)) \leq g(v,\theta)$. The result of the proof follows the proof of Theorem 2 in Karimi et al. [13]. ∎

Lemma 14 *Under Assumption 5 and 3,* $||\theta^\diamond(v_2,\theta)-\theta^\diamond(v_1,\theta)|| \le \frac{4L}{\kappa}||v_1-v_2||$ *for any* v_1, v_2.

Proof Notice that $\nabla q(v_2, \theta^\diamond(v_2,\theta)) = 0$, we have

$$||\nabla q(v_1, \theta^\diamond(v_2,\theta)) - \nabla q(v_2, \theta^\diamond(v_2,\theta))|| = ||\nabla q(v_1, \theta^\diamond(v_2,\theta))||.$$

By Assumption 3, we have $||\nabla q(v_1, \theta^\diamond(v_2,\theta))|| \ge \sqrt{\kappa(g(v_1,\theta^\diamond(v_2,\theta)) - g(v_1,\theta^\diamond(v_1,\theta))}$. And by Lemma 13, we have

$$g(v_1,\theta^\diamond(v_2,\theta)) - g(v_1,\theta^\diamond(v_1,\theta)) \ge \frac{\kappa}{4}||\theta^\diamond(v_2,\theta)-\theta^\diamond(v_1,\theta)||^2.$$

Combing all bounds gives that

$$2L||v_1-v_2|| \ge ||\nabla q(v_1,\theta^\diamond(v_2,\theta)) - \nabla q(v_2,\theta^\diamond(v_2,\theta))|| = ||\nabla q(v_1,\theta^\diamond(v_2,\theta))|| \ge \frac{\kappa}{2}||\theta^\diamond(v_2,\theta)-\theta^\diamond(v_1,\theta)||.$$

This implies that $||\theta^\diamond(v_2,\theta)-\theta^\diamond(v_1,\theta)|| \le \frac{4L}{\kappa}||v_1-v_2||$. ■

Now we proceed to give the proof of Theorem 1.
Note that

$$\begin{aligned} q(v_{k+1},\theta_{k+1}) - q(v_k,\theta_k) &= [g(v_{k+1},\theta_{k+1}) - g(v_{k+1},\theta^\diamond(v_{k+1},\theta_{k+1}))] - [g(v_k,\theta_k) - g(v_k,\theta^\diamond(v_k,\theta_k))] \\ &= [g(v_{k+1},\theta_{k+1}) - g(v_{k+1},\theta^\diamond(v_{k+1},\theta_k))] - [g(v_k,\theta_k) - g(v_k,\theta^\diamond(v_k,\theta_k))] \\ &+ [g(v_{k+1},\theta^\diamond(v_{k+1},\theta_k)) - g(v_{k+1},\theta^\diamond(v_{k+1},\theta_{k+1}))] \\ &= [g(v_{k+1},\theta_{k+1}) - g(v_k,\theta_k)] - [g(v_{k+1},\theta^\diamond(v_{k+1},\theta_k)) - g(v_k,\theta^\diamond(v_k,\theta_k))] \\ &+ [g(v_{k+1},\theta^\diamond(v_{k+1},\theta_k)) - g(v_{k+1},\theta^\diamond(v_{k+1},\theta_{k+1}))]. \end{aligned}$$

Note that

$$\begin{aligned} g(v_{k+1},\theta_{k+1}) - g(v_k,\theta_k) &\le -\xi\left\langle \nabla g(v_k,\theta_k), \delta^*(v_k,\theta_k)\right\rangle + \frac{L\xi^2}{2}||\delta^*(v_k,\theta_k)||^2 \\ -[g(v_{k+1},\theta^\diamond(v_{k+1},\theta_k)) - g(v_k,\theta^\diamond(v_k,\theta_k))] &\le \left\langle \nabla_{[1,2]} g(v_k,\theta^\diamond(v_k,\theta_k)), [v_{k+1},\theta^\diamond(v_{k+1},\theta_k)] - [v_k,\theta^\diamond(v_k,\theta_k)]\right\rangle \\ &+ \frac{L}{2}||[v_{k+1},\theta^\diamond(v_{k+1},\theta_k)] - [v_k,\theta^\diamond(v_k,\theta_k)]||^2. \end{aligned}$$

Notice that as $\nabla_2 g(v_k,\theta^\diamond(v_k,\theta_k)) = 0$,

$$\left\langle \nabla_{[1,2]} g(v_k,\theta^\diamond(v_k,\theta_k)), [v_{k+1},\theta^\diamond(v_{k+1},\theta_k)] - [v_k,\theta^\diamond(v_k,\theta_k)]\right\rangle = \xi\left\langle \nabla_{[1,2]} g(v_k,\theta^\diamond(v_k,\theta_k)), \delta^*(v_k,\theta_k)\right\rangle.$$

Also using Lemma 14, we have

$$||\theta^\diamond(v_{k+1},\theta_k) - \theta^\diamond(v_k,\theta_k)|| \le \frac{4L}{\kappa}||v_{k+1}-v_k||.$$

This implies that

$$||[v_{k+1},\theta^\diamond(v_{k+1},\theta_k)] - [v_k,\theta^\diamond(v_k,\theta_k)]||^2 \le (\frac{16L^2}{\kappa^2}+1)||v_{k+1}-v_k||^2 \le (\frac{16L^2}{\kappa^2}+1)\xi^2||\delta^*(v_k,\theta_k)||^2.$$

We thus have

$$q(v_{k+1},\theta_{k+1}) - q(v_k,\theta_k) \leq -\xi \left\langle \nabla q(v_k,\theta_k), \delta^*(v_k,\theta_k)\right\rangle + L_q\xi^2||\delta^*(v_k,\theta_k)||^2/2 + \chi_k,$$

where we define $L_q = (\frac{16L^2}{\kappa^2}+2)$ and $\chi_k = [g(v_{k+1},\theta^\diamond(v_{k+1},\theta_k)) - g(v_{k+1},\theta^\diamond(v_{k+1},\theta_{k+1}))]$. Using the same argument in the proof of Lemma 10 and Lemma 11, we have

$$\begin{aligned}
& q(v_{k+1},\theta_{k+1}) - q(v_k,\theta_k) \\
\leq & -\xi\eta||\nabla q(v_k,\theta_k)||^2 + 12\xi\eta L_q^2\kappa^{-1}\exp(-b_1T)q(v_k,\theta_k) \\
& + 2\xi b_2 L\kappa^{-1/2}\exp(-b_1T/2)q^{1/2}(v_k,\theta_k) + L_q\xi^2b_2^2/2 + \chi_k \\
\leq & -\xi\eta||\nabla q(v_k,\theta_k)||^2 + 12\xi\eta L_q^2\kappa^{-2}\exp(-b_1T)||\nabla q(v_k,\theta_k)||^2 \\
& + 2\xi b_2 L\kappa^{-1}\exp(-b_1T/2)||\nabla q(v_k,\theta_k)|| + L_q\xi^2b_2^2/2 + \chi_k.
\end{aligned}$$

Here the second inequality is by Assumption 3. Choosing T such that $T \geq b_7(\eta,\alpha,\kappa,L)$ where

$$b_7(\eta,\alpha,\kappa,L) = \left\lceil -b_1^{-1}\log(\frac{\kappa^2}{48\eta L_q^2})\right\rceil,$$

we have

$$q(v_{k+1},\theta_{k+1}) - q(v_k,\theta_k) \leq -\frac{3}{4}\xi\eta||\nabla q(v_k,\theta_k)||^2 + 2\xi b_2L\kappa^{-1}\exp(-b_1T/2)||\nabla q(v_k,\theta_k)|| + L_q\xi^2b_2^2/2 + \chi_k.$$

Using Young's inequality, given any $x>0$,

$$\exp(-b_1T/2)||\nabla q(v_k,\theta_k)|| \leq x\exp(-b_1T) + \frac{1}{x}||\nabla q(v_k,\theta_k)||^2.$$

Choosing $x = \frac{4Lb_2}{\eta\kappa}$, we have

$$q(v_{k+1},\theta_{k+1}) - q(v_k,\theta_k) \leq -\frac{1}{4}\xi\eta||\nabla q(v_k,\theta_k)||^2 + \Delta + \chi_k,$$

where we denote $\Delta = \xi\frac{8L^2b_2^2}{\eta\kappa^2}\exp(-b_1T) + \frac{1}{2}L_q\xi^2b_2^2$. This gives that

$$\frac{1}{4}\xi\eta\sum_{k=0}^{K}||\nabla q(v_k,\theta_k)||^2 \leq q(v_0,\theta_0) - q(v_K,\theta_K) + K\Delta + \sum_{k=0}^{K-1}\chi_k.$$

Using the same argument in the proof of Lemma 11,

$$||\hat{\nabla} q(v,\theta)|| \leq (2L\kappa^{-1}+1)||\nabla q(v,\theta)||.$$

We hence have

$$\begin{aligned}
\sum_{k=0}^{K-1}||\hat{\nabla} q(v_k,\theta_k)||^2 &\leq (2L\kappa^{-1}+1)^2\sum_{k=0}^{K-1}||\nabla q(v_k,\theta_k)||^2 \\
&\leq \frac{4(2L\kappa^{-1}+1)^2}{\xi\eta}(q(v_0,\theta_0) - q(v_K,\theta_K) + K\Delta + \sum_{k=0}^{K-1}\chi_k).
\end{aligned}$$

Similar to the proof of Lemma 12,

$$\sum_{k=0}^{K-1} ||\delta^*(v_k,\theta_k)||^2 \le \frac{2(f(v_0,\theta_0)-f(v_K,\theta_K))}{\xi} + 2\eta^2 \sum_{k=0}^{K-1} ||\hat{\nabla} q(v_k,\theta_k)||^2 + 2\eta M\sqrt{K}\sqrt{\sum_{k=0}^{K-1} ||\hat{\nabla} q(v_k,\theta_k)||^2}.$$

Using $\sqrt{x+y} \le \sqrt{x}+\sqrt{y}$, we have

$$2\eta^2 \sum_{k=0}^{K-1} ||\hat{\nabla} q(v_k,\theta_k)||^2 \le \frac{8\eta(2L\kappa^{-1}+1)^2}{\xi}(q(v_0,\theta_0)-q(v_K,\theta_K)+K\Delta+\sum_{k=0}^{K-1}\chi_k)$$

$$2\eta M\sqrt{K}\sqrt{\sum_{k=0}^{K-1} ||\hat{\nabla} q(v_k,\theta_k)||^2} \le \sqrt{K}\frac{4\eta^{1/2}M(2L\kappa^{-1}+1)}{\xi^{1/2}}(\sqrt{q(v_0,\theta_0)-q(v_K,\theta_K)}+K^{1/2}\Delta^{1/2}+\sqrt{\left[\sum_{k=0}^{K-1}\chi_k\right]_+})$$

Also notice that by Assumption 3,

$$\begin{aligned}\sum_{k=0}^{K-1} q(v_k,\theta_k) &\le \sum_{k=0}^{K-1}\frac{\xi}{\kappa}||\nabla q(v_k,\theta_k)||^2\\ &\le \frac{4}{\eta\kappa\xi}(q(v_0,\theta_0)-q(v_K,\theta_K)+K\Delta+\sum_{k=0}^{K-1}\chi_k)\end{aligned}$$

We hence have

$$\begin{aligned}\sum_{k=0}^{K-1}(||\delta^*(v_k,\theta_k)||^2+q(v_k,\theta_k)) &= O\left(\frac{1}{\xi}+\frac{K\Delta}{\xi}+\frac{K^{1/2}}{\xi^{1/2}}+\frac{K\Delta^{1/2}}{\xi^{1/2}}+K^{1/2}\left(\left[\sum_{k=0}^{K-1}\chi_k\right]_+\right)^{1/2}\right)\\ &= O\left(\frac{1}{\xi}+K\exp(-b_1T/2)+K\xi^{1/2}+\frac{K^{1/2}}{\xi^{1/2}}+\left(K\left[\sum_{k=0}^{K-1}\chi_k\right]_+\right)^{1/2}\right).\end{aligned}$$

Using the same argument as the proof of Theorem 2, when $T \ge \lceil -b_1^{-1}\log(\frac{1}{16}\kappa^2L^{-2})\rceil$,

$$\mathcal{K}^\diamond(v,\theta) \le 2||\delta^*(v,\theta)||^2+q(v,\theta)+8L^2\exp(-b_1T)\kappa^{-2}(\eta+2)^2b_2^2.$$

This implies that

$$\begin{aligned}\min_k \mathcal{K}^\diamond(v_k,\theta_k) &\le \frac{1}{K}\sum_{k=0}^{K-1}[2||\delta^*(v,\theta)||^2+q(v,\theta)]+8L^2\exp(-b_1T)\kappa^{-2}(\eta+2)^2b_2^2\\ &= O\left(\frac{1}{\xi K}+\exp(-b_1T/2)+\xi^{1/2}+\frac{1}{\xi^{1/2}K^{1/2}}+\left(\left[\frac{1}{K}\sum_{k=0}^{K-1}\chi_k\right]_+\right)^{1/2}\right).\end{aligned}$$

Now we proceed to bound $\frac{1}{K}\sum_{k=0}^{K-1}\chi_k$. Notice that

$$\begin{aligned}\chi_k &= g(v_{k+1},\theta^\diamond(v_{k+1},\theta_k))-g(v_{k+1},\theta^\diamond(v_{k+1},\theta_{k+1}))\\ &= g(v_{k+1},\theta^\diamond(v_{k+1},\theta_k))-g(v_k,\theta^\diamond(v_k,\theta_k))+g(v_k,\theta^\diamond(v_k,\theta_k))-g(v_{k+1},\theta^\diamond(v_{k+1},\theta_{k+1})).\end{aligned}$$

Notice that using Assumption 5 and Lemma 14

$$\begin{aligned} g(v_{k+1}, \theta^{\diamond}(v_{k+1}, \theta_k)) - g(v_k, \theta^{\diamond}(v_k, \theta_k)) &\leq L||[v_{k+1}, \theta^{\diamond}(v_{k+1}, \theta_k)] - [v_k, \theta^{\diamond}(v_k, \theta_k)]|| \\ &\leq L(||v_{k+1} - v_k|| + ||\theta^{\diamond}(v_{k+1}, \theta_k) - \theta^{\diamond}(v_k, \theta_k)||) \\ &\leq (L + \frac{4L}{\kappa})||v_{k+1} - v_k|| \\ &\leq (L + \frac{4L}{\kappa})\xi||\delta^*(v_k, \theta_k)||. \end{aligned}$$

Note that using the same procedure as the proof of Lemma 6, $||\delta^*(v_k, \theta_k)|| \leq b_2$. We thus conclude that

$$\begin{aligned} \sum_{k=0}^{K-1} \chi_k &\leq \sum_{k=0}^{K-1} g(v_k, \theta^{\diamond}(v_k, \theta_k)) - g(v_{k+1}, \theta^{\diamond}(v_{k+1}, \theta_{k+1})) \\ &+ (L + \frac{4L}{\kappa})\xi \sum_{k=0}^{K-1} ||\delta^*(v_k, \theta_k)|| \\ &\leq \sum_{k=0}^{K-1} g(v_k, \theta^{\diamond}(v_k, \theta_k)) - g(v_{k+1}, \theta^{\diamond}(v_{k+1}, \theta_{k+1})) + (L + \frac{4L}{\kappa})b_2\xi K \\ &= g(v_0, \theta^{\diamond}(v_0, \theta_0)) - g(v_K, \theta^{\diamond}(v_K, \theta_K)) + (L + \frac{4L}{\kappa})b_2\xi K. \end{aligned}$$

We thus have $\frac{1}{K}\sum_{k=0}^{K-1} \chi_k = O(\frac{1}{K} + \xi)$.

Appendix F. List of absolute constants used in the proofs

Here we summarize the absolute constant used in the proofs.

$$\begin{aligned} b_1(\alpha, L, \kappa) &= \log(1 - (\alpha - L\alpha^2/2)\kappa) \\ b_2(M, \eta) &= (3 + \eta)M \\ b_3(\eta, \alpha, \kappa, L) &= \left\lceil -b_1^{-1} \log(\frac{\eta\kappa}{64\eta L_q^2}) \right\rceil \\ b_4(\eta, \kappa, \xi) &= -\log(1 - \frac{\xi}{4}\eta\kappa) \\ b_5(L_q, \eta, \kappa) &= \frac{16L_q^2}{\eta\kappa^2}(2L\kappa^{-1} + 1)^2 \\ b_6(\kappa, L) &= (2L\kappa^{-1} + 1)^2 \\ b_7(\eta, \alpha, \kappa, L) &= \left\lceil -b_1^{-1} \log(\frac{\kappa^2}{48\eta L_q^2}) \right\rceil \end{aligned}$$