OPT2023: 15th Annual Workshop on Optimization for Machine Learning

Stochastic Optimization under Hidden Convexity

Ilyas Fatkhullin[*] ILYAS.FATKHULLIN@AI.ETHZ.CH
Niao He[*] NIAO.HE@INF.ETHZ.CH
Yifan Hu[*†] YIFAN.HU@EPFL.CH
[*]*ETH Zurich,* [†]*EPFL, Switzerland.*

Abstract

In this work, we consider stochastic non-convex constrained optimization problems under *hidden convexity*, i.e., those that admit a convex reformulation via a black box (non-linear, but invertible) map $c : \mathcal{X} \to \mathcal{U}$. A number of non-convex problems ranging from optimal control, revenue and inventory management, to convex reinforcement learning all admit such a hidden convex structure. Unfortunately, in the majority of considered applications, the map $c(\cdot)$ is unavailable and therefore, the reduction to solving a convex optimization is not possible. On the other hand, the (stochastic) gradients with respect to the original variable $x \in \mathcal{X}$ are often easy to obtain. Motivated by these observations, we consider the projected stochastic (sub-) gradient methods under hidden convexity and provide the first sample complexity guarantees for global convergence in smooth and non-smooth settings. Additionally, we improve our results to the last iterate function value convergence in the smooth setting using the momentum variant of projected stochastic gradient descent.

Keywords: hidden convexity, stochastic optimization, global convergence

1. Introduction

We study stochastic constrained non-convex optimization

$$\min_{x \in \mathcal{X}} F(x) := \mathbb{E}_{\xi \sim \mathcal{D}} \left[f(x, \xi) \right], \tag{1}$$

where $\mathcal{X}$ is a closed convex subset of $\mathbb{R}^d$, ξ is a random variable satisfying an unknown distribution $\mathcal{D}$, and $F(\cdot)$ is (possibly) non-convex in x. Our central structural assumption about (1) is that it admits a *convex reformulation* of the form

$$\min_{u \in \mathcal{U}} H(u) := F(c^{-1}(u)), \tag{2}$$

where $H(\cdot)$ is a convex function defined on a closed convex set $\mathcal{U} \subset \mathbb{R}^d$, and $c : \mathcal{X} \to \mathcal{U}$ is an invertible map (with its inverse denoted by $c^{-1}(\cdot)$). Such problems frequently arise in various applications including constrained nonlinear least-squares [22], policy optimization in convex reinforcement learning and optimal control [66, 75], generative models [43], supply chain and revenue management [10, 29]. Despite the existence of the convex reformulation, the transformation function $c(\cdot)$ in these applications is usually hard to compute or even unknown. Thus one cannot readily solve the convex reformulation. In this work, we show that the simple projected stochastic (sub-)gradient method on the original non-convex problem (1) converges globally, and demonstrate the sample complexity. If not surprisingly, the method is very classical and it does not require approximating some global converging algorithms on the convex reformulation [11].

Hidden convex optimization. In many problems of practical interest, global convergence is desirable despite the non-convexity of the optimization landscape. To achieve that, some works relax convexity/strong convexity of $F(\cdot)$ in a way to ensure that global convergence proofs of gradient methods still work without serious modifications, e.g. assuming Polyak-Łojasiewicz (PŁ) type conditions [40, 41, 60]. However, such an approach dictated from the analysis leads to serious challenges when it comes to verifying these assumptions for specific applications. In this work, we focus on a different structural property, known as *hidden convexity*, i.e., the existence of a convex reformulation (2). Hidden convexity is a very natural condition, since compositional optimization problems are ubiquitous in modern applications, see Appendix D for various motivating examples. Although hidden convexity has been identified in certain applications, the analysis of gradient methods under this condition is mostly done on a case by case basis for specific applications and often under strong additional assumptions [4, 10, 11, 76]. We refer the interested reader to Appendix B, where we review the most closely related work on hidden convex optimization. In this work, we formally introduce stochastic optimization under hidden convexity and provide a systematic study of projected stochastic (sub)-gradient methods for hidden convex problems. Our main contributions are summarized as follows.

Contributions:

1. We identify the key properties of hidden convex optimization and demonstrate how these conditions can be used to derive convergence of gradient methods.
2. In the general non-differentiable case, we analyze convergence of the stochastic sub-gradient method under hidden convexity. To our knowledge, it is the first work that address the non-differentiable setting.
3. Next, we specialize our results to the differentiable smooth setting, and analyze the sample complexity of Projected stochastic gradient descent (P-SGD). Importantly, our analysis does not require large batches of samples at every iteration nor bounded gradients assumptions.
4. Finally, we analyze the momentum variant of P-SGD for solving (1), which allows us to show global convergence measured by the function value at the last iterate.

In terms of sample complexity, we obtain $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ sample complexity guarantee in hidden convex case and further improve the result to $\widetilde{\mathcal{O}}(\varepsilon^{-1})$ for hidden strongly convex problems.

2. Hidden Convex Problem Class

The fact that the problem (1) admits a convex reformulation (2) means that

$$\min_{x \in \mathcal{X}} \; F(x) := H(c(x)). \tag{3}$$

We call the above problem *hidden convex* if its components satisfy the following conditions.

C.1. The domain $\mathcal{U} = c(\mathcal{X})$ is convex, the function $H : \mathcal{U} \to \mathbb{R}$ is convex and (2) admits a solution $u^* \in \mathcal{U}$.

C.2. The map $c : \mathcal{X} \to \mathcal{U}$ is invertible. There exists $\mu_c > 0$ such that for all $x, y \in \mathcal{X}$ it holds

$$\|c(x) - c(y)\| \geq \mu_c \|x - y\| . \tag{4}$$

To analyze convergence of SM, we also make the following assumptions.

A.1. $F(\cdot)$ is ℓ-weakly convex, and defined on a closed, convex set $\mathcal{X}$.

A.2. We have access to a stochastic sub-gradient oracle of $F(\cdot)$ at any $x \in \mathcal{X}$, which outputs a random vector $g(x, \xi)$ such that $\mathbb{E}[g(x, \xi)] \in \partial F(x)$, where $\partial F(x)$ is the sub-differential set of $F(\cdot)$ at x. Moreover, there exists $G_F > 0$

$$\mathbb{E}\left[\|g(x, \xi)\|^2\right] \leq G_F^2 \qquad \text{for any } x \in \mathcal{X}.$$

Above assumptions are standard and appear frequently in non-smooth optimization [17, 77]. In particular, in the absence of smoothness, the second assumption of bounded second moment of the (stochastic) sub-gradients is classic assumption even in convex case [53, 56].

Remark 1 *It is known that ℓ-WC is a much weaker condition than smoothness [15, 56]. In the context of our hidden convexity (C.1. and C.2.), the following result shows that weak convexity is not restrictive and, in fact, should not be treated as an additional assumption, since it comes for free from the Lipschitz continuity of $H(\cdot)$ and the smoothness of the transformation function $c(\cdot)$, which is often available in our applications.*

Proposition 2 (Proposition 2.2(c) in [77]) *Let $\mathcal{U} \subseteq \mathbb{R}^d$ be a closed convex set, and $H : \mathcal{U} \to \mathbb{R}$ be a convex and G_H-Lipschitz continuous function defined on $\mathcal{U}$, i.e., $|H(u) - H(v)| \leq G_H \|u - v\|$ for all $u, v \in \mathcal{U}$. Let $c : \mathcal{X} \to \mathcal{U}$ be L_c-smooth, i.e, $\|c(x) - c(y) - \langle \nabla c(y), x - y\rangle\| \leq \frac{L_c}{2} \|x - y\|^2$ for all $x, y \in \mathcal{X}$. Then the composition $F(x) = H(c(x))$ is ℓ-weakly convex with $\ell := G_H L_c$.*

In differentiable, smooth case, however, the assumption A.2 can be limiting. For instance, when the set $\mathcal{X}$ is unbounded, it fails to hold for convex quadratics. For this reason, in Sections 4 and 5, we will present a tighter analysis replacing above assumptions with smoothness and bounded variance.

A.1'. The function $F : \mathcal{X} \to \mathbb{R}$ is differentiable on a closed, convex set $\mathcal{X}$ and its gradient $\nabla F(x)$ is L-Lipschitz continuous.

A.2'. We have access to an unbiased stochastic gradient oracle with bounded variance $\sigma^2 > 0$, i.e. for any $x \in \mathcal{X}$: $\mathbb{E}[\nabla f(x, \xi)] = \nabla F(x)$, and

$$\mathbb{E}\left[\|\nabla f(x, \xi) - \nabla F(x)\|^2\right] \leq \sigma^2,$$

where expectations are with respect to the random variable $\xi \sim \mathcal{D}$.

3. Stochastic Subgradient Method

In this section, we show the global convergence of the (projected) stochastic subgradient method (SM) using the key Proposition 9. Starting from $x^0 \in \mathcal{X}$, SM generates a sequence x^t via

$$\boxed{x^{t+1} = \Pi_{\mathcal{X}}(x^t - \eta g(x^t, \xi^t)). \qquad (5)}$$

Let $x^* \in \mathcal{X}^*$, and $\Phi := F + \delta_{\mathcal{X}}$. We define the Lyapunov function: $\Lambda_t := \mathbb{E}\left[\Phi_{1/\rho}(x^t) - F(x^*)\right]$, where $\Phi_{1/\rho}$ is the Moreau envelope. Notice that $\Lambda_t \geq 0$ for any $t \geq 0$ and $\Lambda_t = 0$ iff $x^t \in \mathcal{X}^*$ since $\Phi_{1/\rho}(x^*) = \Phi(x^*) = F(x^*)$ and $\Phi_{1/\rho}(x) \geq \Phi(x^*)$ for any $x \in \mathcal{X}$. We first demonstrate the convergence rate of the hidden convex setting.

Theorem 3 (Convex $H(\cdot)$) *Let C.1, C.2., A.1 and A.4 hold, and the set $\mathcal{U}$ be bounded by a diameter $D_{\mathcal{U}}$. Fix $\varepsilon > 0$, and set the step-size in (5) as $\eta = \frac{1}{2\ell} \cdot \min\left\{1, \frac{\mu_c^2 \varepsilon^2}{24 D_{\mathcal{U}}^2 G_F^2}\right\}$. Then for $\rho = 2l$, we have $\Lambda_T \leq \varepsilon$ after $T = \widetilde{\mathcal{O}}\left(\frac{\ell D_{\mathcal{U}}^2}{\mu_c^2}\frac{1}{\varepsilon} + \frac{\ell D_{\mathcal{U}}^4 G_F^2}{\mu_c^4}\frac{1}{\varepsilon^3}\right)$ iterations.*

We remark that in the absence of smoothness of $F(\cdot)$, the guarantee on Λ_t might not necessarily translate to the function value. However, with the following corollary we show that the output of (5), x^T, is in fact close to an ε-approximate global solution $\hat{x}^T = \mathrm{prox}_{\eta\Phi}(x^T)$.

Corollary 4 *Under the setting of Theorem 3, the method (5) finds a point $x^T \in \mathcal{X}$, which is close to $\hat{x}^T$, an ε-global solution of (1). More specifically, it holds that $\mathbb{E}\left[\left\|\hat{x}^T - x^T\right\|^2\right] \leq \varepsilon/(4\ell)$ and $\mathbb{E}\left[F(\hat{x}^T) - F(x^*)\right] \leq \varepsilon$ after $T = \widetilde{\mathcal{O}}(\varepsilon^{-3})$.*

Proof The result follows directly from the definition of Λ_T and Theorem 3. ■

4. Projected SGD

In this section, we consider the smooth setting, i.e., assumptions A.1' and A.2' holds. Then the SM becomes Projected SGD.

$$x^{t+1} = \Pi_{\mathcal{X}}(x^t - \eta \nabla f(x^t, \xi^t)). \tag{6}$$

Theorem 5 (Convex $H(\cdot)$) *Let C.1, C.2, A.1' and A.2' hold, and the set $\mathcal{U}$ be bounded by a diameter $D_{\mathcal{U}}$. Then for any $\eta \leq \frac{2}{9L}$, and $\alpha \leq 2\eta L$, we have for all $T \geq 0$*

$$\Lambda_T \quad \leq \quad (1-\alpha)^T \Lambda_0 + \frac{3 D_{\mathcal{U}}^2 \alpha}{2\mu_c^2 \eta} + \frac{4 L \eta^2 \sigma^2}{\alpha}.$$

Fix $\varepsilon > 0$, and set the step-size in (6) as $\eta = \frac{2}{9L} \cdot \min\left\{1, \frac{\mu_c^2 \varepsilon^2}{12 D_{\mathcal{U}}^2 \sigma^2}\right\}$. Then $\Lambda_T \leq \varepsilon$ after $T = \widetilde{\mathcal{O}}\left(\frac{L D_{\mathcal{U}}^2}{\mu_c^2}\frac{1}{\varepsilon} + \frac{L D_{\mathcal{U}}^4 \sigma^2}{\mu_c^4}\frac{1}{\varepsilon^3}\right)$ iterations.

5. Projected SGD with Momentum

In this section, we study Projected SGD with Polyak's (heavy-ball) momentum [58]. Comparing to earlier sections, we directly derive its global convergence to an ε-global solution. The analysis in this section uses the same properties established in Appendix E.3, but the Lyapunov function used here is completely different from Λ_t used in Sections 3 and 4. The Projected SGD with Polyak's (heavy-ball) momentum admits the following updates.

$$x^{t+1} = \Pi_{\mathcal{X}}(x^t - \eta\, g^t), \qquad g^{t+1} = (1-\beta)\, g^t + \beta\, \nabla f(x^{t+1}, \xi^{t+1}). \tag{7}$$

Let $x^* \in \mathcal{X}^*$, for any $x^t \in \mathcal{X}$, we define the Lyapunov function:

$$\Lambda_t^{HB} := \left[F(x^t) - F(x^*) + \frac{\eta}{\beta}\left\|g^t - \nabla F(x^t)\right\|^2\right]. \tag{8}$$

We show that both $\mathbb{E}\left[F(x^t) - F(x^*)\right]$ and $\mathbb{E}\left[\frac{1}{L}\left\|g^t - \nabla F(x^t)\right\|^2\right]$ are diminishing over iterations of the scheme (7). Combining Lemma 19 with Lemma 20, we obtain at the following theorem.

Theorem 6 (Convex $H(\cdot)$) *Let C.1, C.2, A.1' and A.2' hold, and the set $\mathcal{U}$ be bounded by a diameter $D_{\mathcal{U}}$. Then for any $\eta \le \frac{\beta}{4L}$, $\beta \in (0,1]$, and $\alpha \le \frac{\beta}{2}$, we have for any $T \ge 0$*

$$\Lambda_T^{HB} \le (1-\alpha)^T \Lambda_0^{HB} + \frac{\alpha D_{\mathcal{U}}^2}{\mu_c^2 \eta} + \frac{\beta\eta\sigma^2}{\alpha},$$

where Λ_t^{HB} is given by (8). Fix $\varepsilon > 0$, and set the parameters of algorithm (7) as

$$\eta = \frac{\beta}{4L}, \quad \beta = \min\left\{1, \frac{\mu_c^2}{9D_{\mathcal{U}}^2\sigma^2}\varepsilon^2\right\}.$$

Then the scheme (7) returns a point x^T with $\mathbb{E}\left[F(x^T) - F(x^*)\right] \le \varepsilon$ when

$$T = \widetilde{\mathcal{O}}\Big(\frac{LD_{\mathcal{U}}^2}{\mu_c^2}\frac{1}{\varepsilon} + \frac{LD_{\mathcal{U}}^4\sigma^2}{\mu_c^4}\frac{1}{\varepsilon^3}\Big).$$

We remark that both Theorems 6 and 22 provide last iterate global convergence for P-SGD with momentum without the need of using large mini-batch (even once). Additionally, it guarantees that the gradient estimate g^t also converges to the true gradient $\nabla F(x^*)$ at the optimum $x^* \in \mathcal{X}^*$. In case of strongly convex $H(\cdot)$, similarly to Corollary 12 and Theorem 17, the result of Theorem 22 (for strongly convex $H(\cdot)$) can be translated into the point (iterate) convergence to the optimal x^*.

6. Conclusions

In this work, we make the first steps towards theoretical understanding of stochastic optimization under hidden convexity and develop batch-free stochastic gradient methods with projection.

A few related questions regarding considered projected stochastic (sub-)gradient descent algorithm still remain open. 1) For hidden strongly convex problem, i.e., when $\mu_H > 0$, the derived sample complexities are optimal in ε up to the logarithmic factor as it matches the result for usual strongly convex settings, and therefore unimprovable for SM and P-SGD. However, for merely hidden convex $F(\cdot)$, i.e., $\mu_H = 0$, it is unclear if our $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ sample complexity is tight for SM and P-SGD. 2). The benefits of momentum variants of P-SGD can be further explored, e.g., to understand if Nesterov's acceleration is possible under hidden convexity. 3) When $\mu_H = 0$, our iteration and sample complexity results depend on the diameter of the reformulated problem. It would be interesting to explore if $D_{\mathcal{U}}$ can be replaced with the distance to the solution, i.e., $\left\|c(x^0) - c(x^*)\right\|$.

References

[1] Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

[2] James Anderson, John C. Doyle, Steven Low, and Nikolai Matni. System level synthesis. *aXiv preprint arXiv:1904.01634v1*, 2019.

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

[4] Anas Barakat, Ilyas Fatkhullin, and Niao He. Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In *International Conference on Machine Learning*, 2023.

[5] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[6] Julius R Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744, 1954.

[7] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[8] Stephen Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear matrix inequalities in system and control theory*. SIAM, 1994.

[9] Stephen Boyd, Seung-Jean Kim, Lieven Vandenberghe, and Arash Hassibi. A tutorial on geometric programming. *Optimization and engineering*, 8:67–127, 2007.

[10] Xin Chen, Niao He, Yifan Hu, and Zikun Ye. Efficient algorithms for minimizing compositions of convex functions and random functions and its applications in network revenue management. *arXiv preprint arXiv:2205.01774*, 2022.

[11] Yiwei Chen and Cong Shi. Network revenue management with online inverse batch gradient descent method. *Available at SSRN 3331939*, 2022.

[12] Kai Lai Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, pages 463–483, 1954.

[13] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

[14] Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 6643–6670, 2023.

[15] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[16] Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

[17] Damek Davis, Dmitriy Drusvyatskiy, and Kellie J MacPhee. Stochastic model-based minimization under high-order growth. *arXiv preprint arXiv:1807.00255*, 2018.

[18] Damek Davis, Dmitriy Drusvyatskiy, Kellie J MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179:962–982, 2018.

[19] Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global optimum convergence of momentum-based policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 1910–1934. PMLR, 2022.

[20] Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. In *International Conference on Machine Learning*, pages 2658–2667. PMLR, 2020.

[21] Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

[22] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019.

[23] Richard J Duffin. Geometric programming-theory and application. Technical report, 1967.

[24] Ilyas Fatkhullin and Boris Polyak. Optimizing Static Linear Feedback: Gradient Method. *SIAM Journal on Control and Optimization*, 59(5):3887–3911, 2021.

[25] Ilyas Fatkhullin, Jalal Etesami, Niao He, and Negar Kiyavash. Sharp analysis of stochastic optimization under global Kurdyka-łojasiewicz inequality. *Advances in Neural Information Processing Systems*, 2022.

[26] Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 9827–9869, 2023.

[27] Ilyas Fatkhullin, Alexander Tyurin, and Peter Richtárik. Momentum provably improves error feedback! *arXiv preprint arXiv:2305.15155*, 2023.

[28] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pages 1467–1476. PMLR, 2018.

[29] Qi Feng and J George Shanthikumar. Supply and demand functions in inventory models. *Operations Research*, 66(1):77–91, 2018.

[30] Xavier Fontaine, Valentin De Bortoli, and Alain Durmus. Convergence rates and approximation results for sgd and its continuous-time counterpart. In *Conference on Learning Theory*, pages 1965–2058. PMLR, 2021.

[31] Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. 2018.

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

[33] Udaya Ghai, Zhou Lu, and Elad Hazan. Non-convex online learning via algorithmic equivalence. *Advances in Neural Information Processing Systems*, 35:22161–22172, 2022.

[34] Osman Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, 1991. doi: 10.1137/0329022.

[35] Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on learning theory*, pages 1894–1938. PMLR, 2020.

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

[37] Yifan Hu, Xin Chen, and Niao He. On the bias-variance-cost tradeoff of stochastic optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

[38] Yifan Hu, Wang Jie, Yao Xie, Andreas Krause, and Daniel Kuhn. Contextual stochastic bilevel optimization. *Advances in Neural Information Processing Systems*, 36, 2023.

[39] Michael Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4570–4597. PMLR, 2023.

[40] 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.

[41] Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

[42] J. Kiefer and J. Wolfowitz. Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics*, 23(3):462 – 466, 1952.

[43] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

[44] Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

[45] A. S. Lewis and S. J. Wright. A proximal method for composite minimization. *arXiv preprint arXiv:0812.0423*, 2015.

[46] Xiaoyu Li, Mingrui Liu, and Francesco Orabona. On the last iterate convergence of momentum methods. In *International Conference on Algorithmic Learning Theory*, pages 699–717. PMLR, 2022.

[47] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271, 2020.

[48] Vien Mai and Mikael Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International conference on machine learning*, pages 6630–6639. PMLR, 2020.

[49] Saeed Masiha, Saber Salehkaleybar, Niao He, Negar Kiyavash, and Patrick Thiran. Stochastic second-order methods provably beat sgd for gradient-dominated functions. *In Advances in Neural Information Processing Systems*, 2022.

[50] Sentao Miao and Yining Wang. Network revenue management with nonparametric demand learning:\sqrt {T}-regret and polynomial dimension dependency. *Available at SSRN 3948140*, 2021.

[51] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

[52] Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019.

[53] A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* John Wiley UK/USA, 1983.

[54] Yu Nesterov. Modified Gauss–Newton scheme with worst case guarantees for global performance, 2007.

[55] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[56] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[57] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[58] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[59] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[60] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, 3(4):643–653, 1963.

[61] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[62] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[63] Kevin Scaman, Cedric Malherbe, and Ludovic Dos Santos. Convergence rates of non-convex stochastic gradient descent under a generic lojasiewicz condition and local smoothness. In *International Conference on Machine Learning*, 2022.

[64] Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, pages 3935–3971. PMLR, 2021.

[65] Lorenzo Stella, Andreas Themelis, and Panagiotis Patrinos. Forward-backward quasi-newton methods for nonsmooth optimization problems. *Computational Optimization and Applications*, 67(3):443–487, 2017. arXiv:1604.08096 [math].

[66] Yue Sun and Maryam Fazel. Learning optimal controllers by policy gradient: Global optimality via convex parameterization. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 4576–4581. IEEE, 2021.

[67] Ryan Tibshirani. Slides on hidden convexity. 2016. URL https://www.stat.cmu.edu/~ryantibs/convexopt-[]F16/lectures/nonconvex.pdf.

[68] Yong Xia. A survey of hidden convex optimization. *Journal of the Operations Research Society of China*, 8(1):1–28, 2020.

[69] Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

[70] Junchi Yang, Xiang Li, Ilyas Fatkhullin, and Niao He. Two sides of one coin: the limits of untuned sgd and the power of adaptive methods. *arXiv preprint arXiv:2305.12475*, 2023.

[71] Pengyun Yue, Cong Fang, and Zhouchen Lin. On the lower bound of minimizing polyak-Łojasiewicz functions. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pages 2948–2968, 2023.

[72] Hui Zhang and Wotao Yin. Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*, 2013.

[73] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11173–11182, 2020.

[74] Junyu Zhang and Lin Xiao. Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization. *Mathematical Programming*, 2021.

[75] Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020.

[76] Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, Mengdi Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021.

[77] Siqi Zhang and Niao He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.

Contents

Appendix A. Related Work on SM and P-SGD in Convex and Non-convex Settings

The projected sub-gradient methods (SM), its special case, projected stochastic gradient descent (P-SGD), in the differentiable setting, and their numerous variants have a long history of development since the first works on stochastic approximation appeared in 1950s [6, 12, 42, 61].

Convex optimization. The case of convex $F(\cdot)$ is particularly well documented [1, 30, 30, 51]. Researchers have studied how to deal with convex constraints, proximal operators, general Bregman divergences [5, 53], and leveraging averaging and momentum schemes [31, 46, 59, 64]. In the convex case, the global convergence of gradient methods in the function value, i.e., find $x \in \mathcal{X}$ with $\mathbb{E}\left[F(x) - F(x^*)\right] \leq \varepsilon$ for any $\varepsilon > 0$, is naturally possible and the sample complexity required is $\mathcal{O}\left(\varepsilon^{-2}\right)$.[1]

Non-convex optimization. In the last decade, the interest in the optimization community shifted towards general non-convex problems (often smooth or weakly convex), where only convergence to a first-order stationary point (FOSP) is possible in general [3, 20, 41, 70], i.e., find $x \in \mathcal{X}$ with $\mathbb{E}\left[\|\nabla F(x)\|\right] \leq \varepsilon$ when $F(\cdot)$ is smooth. Similar to developments in convex optimization, convergence of non-convex SGD extends to constrained/proximal setting [7, 32, 44], mirror descent [17, 77], momentum [27, 47], variance-reduction [3, 13], and biased gradient setting [36–38]. For the more general weakly-convex case [15, 48, 77], the convergence guarantees are usually with respect to a gradient norm of a smoothed objective. Some works consider non-convex functions with a specific compositional structure similar to (2), e.g. the composition of a convex function with a differentiable and smooth map $c(\cdot)$, see [22, 45, 54, 74]. Recently a number of works focus on non-convex non-smooth optimization (beyond weak convexity) and develop convergence for suitably defined notions of FOSP [14, 39, 73]. Although the above works consider non-convex problems, which find a wide range of applications, they often only provide convergence to a FOSP rather than global convergence in the function value.

Appendix B. Related Work on Hidden Convexity and Other Structural Conditions

The most closely related works are [4, 11, 33, 75, 76], which analyze gradient methods under similar structural assumptions in the context of specific applications.

Policy gradient methods in RL. In particular, [4, 75, 76] analyze policy gradient (PG) type methods in reinforcement learning (RL) setting, and derive global convergence guarantees for their algorithms. In [75], the authors consider a PG method with projection, but it is only limited to the case where the exact gradients are available. It is unclear how to extend the technique in their work to the case of stochastic gradients with bounded variance (without resorting to large batches). Next, [76] consider the stochastic setting and propose variance reduced PG method with truncation using large batches of trajectories. Later, [4] removes the requirement of large batches via a normalized variance reduced PG method. However, their results are difficult to extend to the constrained case due to the normalization. In addition, for a general stochastic optimization problem of the form (1), it requires additional strong individual (or expected) smoothness assumption to analyze variance reduced gradient methods.

Stochastic gradient methods in revenue management. A different line of works [10, 11] consider hidden convex objectives in revenue management and design preconditioned gradient-based

[1] For P-SGD under smooth and bounded variance assumptions, A.1' and A.2' in section 2, or for SM under Lipschitz continuity and bounded second moment of stochastic sub-gradients, i.e., A.2.

methods over $\mathcal{X}$, which approximate the classical P-SGD method on the convex reformulation (2) over $\mathcal{U}$. Their MSG method in [10] obtain an $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity for a special revenue management problem under the assumptions that the domain $\mathcal{X}$ is a box constraint, the transformation function $c(x) = \mathbb{E}\left[c(x,\xi)\right]$ is separable and the additional access to $c(x,\xi)$ is available. Leveraging the box constraint structure, [10] also analyze P-SGD and derive $\widetilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity. In contrast, we show that the P-SGD can achieve better $\widetilde{\mathcal{O}}(\epsilon^{-3})$ sample complexity for a general convex compact constraint $\mathcal{X}$, and further extend the results to non-smooth setting.

In the online learning setting, [33] consider a structural property similar to hidden convexity and propose strong assumptions on the reparameterization map $c(\cdot)$ (see Assumptions 2 and 4 therein) under which non-convex online gradient descent in the original space $\mathcal{X}$ is equivalent to online mirror descent for the (convex) reformulated problem. Such equivalence allows them to demonstrate an $\mathcal{O}(T^{2/3})$ regret bound. Instead of showing regret bounds, we directly derive the last iterate convergence in the function value using a different technique and make less restrictive assumptions on $c(\cdot)$, which allows us to cover a wide range of applications.

For more structured non-convex optimization problems that admit hidden convexity, interested readers may refer to [67, 68]. Note that the transformation mapping $c(\cdot)$ can be unknown in applications, and thus the methodology developed therein is generally not applicable.

Related structural assumptions. We would like to mention that several other non-convex structural assumptions have also appeared in unconstrained optimization, including essential strong convexity [40], quasar (strong) convexity [35], restricted secant inequality [72], error bounds [18], quadratic growth [21, 52], PŁ type condition [40]. The latter[2] along with its various generalizations to constrained minimization such as Proximal-PŁ [40] and variational gradient dominance [69] turned out to be particularly popular in the recent years. The convergence of gradient methods under gradient dominance condition has been extensively analyzed [40, 71], including the stochastic setting [25, 26, 63] and even second-order methods [49]. Despite a few examples [19, 24, 28, 69] that show some variants of the PŁ condition hold, it remains a big question how to verify PŁ like conditions for non-convex problems in general. Moreover, the situation becomes even more challenging, when dealing with constrained optimization and/or non-differentiable objectives, where a suitable generalization of the gradient dominance needs to be introduced and carefully studied.

Appendix C. Notations

We denote by $\langle\cdot,\cdot\rangle$ the inner product in $\mathbb{R}^d$ along with its induced Euclidean norm $\|\cdot\|$. For a real valued matrix $A \in \mathbb{R}^{m\times n}$, we denote by $\|\cdot\|_{\text{op}}$ its operator norm, i.e., $\|A\|_{\text{op}} := \max_{\|x\|\le 1}\|Ax\|$. The map $c : \mathcal{X} \to \mathcal{U}$ is called invertible if there exists a map $c^{-1} : \mathcal{U} \to \mathcal{X}$ (called inverse) such that $c^{-1}(c(x)) = x$ for any $x \in \mathcal{X}$ and $c(c^{-1}(u)) = u$ for any $u \in \mathcal{U}$. For any $u, v \in \mathcal{U}$ and any $\lambda \in [0,1]$, if $(1-\lambda)u + \lambda v \in \mathcal{U}$, we say $\mathcal{U}$ is convex. We denote the diameter of $\mathcal{U}$ as $D_{\mathcal{U}} := \sup_{u,v\in\mathcal{U}}\|u - v\|$. For a function $H : \mathcal{U} \to \mathbb{R}$, if there exists $\mu_H \ge 0$ such that for all $u, v \in \mathcal{U}$ and $\lambda \in [0,1]$, it holds $H((1-\lambda)u+\lambda v) \le (1-\lambda)H(u) + \lambda H(v) - \frac{(1-\lambda)\lambda\mu_H}{2}\|u-v\|^2$, we call H convex on $\mathcal{U}$ if $\mu_H = 0$, and μ_H-strongly convex on $\mathcal{U}$ if $\mu_H > 0$.

A function $F : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is ℓ-weakly convex (ℓ-WC) if for any fixed $y \in \mathcal{X}$, $F_\ell(x,y) := F(x) + \frac{\ell}{2}\|x-y\|^2$ is convex in $x \in \mathcal{X}$. The sub-differential $F(\cdot)$ at $x \in \mathcal{X}$ is given by $\partial F(x) := \left\{g \in \mathbb{R}^d \mid F(y) \ge F(x) + \langle g, y - x\rangle + o(\|y-x\|), \forall y \in \mathcal{X}\right\}$. The elements

[2] Also known as global Kurdyka-Łojasiweicz (KŁ) and gradient domination condition.

$g \in \partial F(x)$ are called sub-gradients of $F(\cdot)$ at x, see [16] for alternative definitions of the sub-differential set for ℓ-WC functions. A differentiable function $F : \mathcal{X} \to \mathbb{R}$ is L-smooth on $\mathcal{X} \subset \mathbb{R}^d$ if its gradient is L-Lipschitz continuous on the set $\mathcal{X}$, i.e., it holds $\|\nabla F(x) - \nabla F(y)\| \leq L \|x - y\|$ for all $x, y \in \mathcal{X}$. For a convex set $\mathcal{X} \subset \mathbb{R}^d$, the projection of a point $y \in \mathbb{R}^d$ onto $\mathcal{X}$ is $\Pi_{\mathcal{X}}(y) := \arg\min_{x \in \mathcal{X}} \|y - x\|$. We denote $\delta_{\mathcal{X}}$ as the indicator function of a set $\mathcal{X} \subset \mathbb{R}^d$ and define $\delta_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$ and $\delta_{\mathcal{X}}(x) = +\infty$ otherwise. We define by $\mathcal{X}^* \subset \mathcal{X}$ the set of optimal points of $\min_{x \in \mathcal{X}} F(x)$. A point $\bar{x} \in \mathcal{X}$ is called a stationary point of a weakly convex function $F : \mathcal{X} \to \mathbb{R}$ if $0 \in \partial(F + \delta_{\mathcal{X}})(\bar{x})$. For any function Φ and a real $\rho > 0$, we define the Moreau envelope and the proximal mapping as follows, respectively.

$$\Phi_{1/\rho}(x) := \min_{y \in \mathbb{R}^d} \left\{ \Phi(y) + \frac{\rho}{2} \|y - x\|^2 \right\}, \quad \mathrm{prox}_{\Phi/\rho}(x) := \arg\min_{y \in \mathbb{R}^d} \left\{ \Phi(y) + \frac{\rho}{2} \|y - x\|^2 \right\}.$$

Appendix D. Motivating Examples

Notice that the hidden convex function class includes the convex function class as a special case when the transformation map $c(\cdot)$ is identical. In addition, it also includes many non-convex functions. For instance, let $0 < \delta \leq 1$ and consider $\mathcal{X} = [\delta, 1]$, $c(x) = x^2$, $H(u) = -u$. Then $F(x) = -x^2$ is concave (non-convex) on $\mathcal{X}$, but is hidden convex by the construction. Another simple example considers $0 < \delta < \pi$ and $\mathcal{X} = [\delta, 2\pi - \delta]$, $c(x) = cos(x)$, $H(u) = u$. The obtained composition $F(x) = cos(x)$ is non-convex and non-concave on $\mathcal{X}$. In what follows, we present more practical (possibly high dimensional) problems, which belong to our hidden convex class.

D.1. Non-linear least squares [22, 55, 57]

Consider solving a system of nonlinear equations under a box constraint, e.g., $c(x) = 0$ with $c(x) = (c_1(x), \ldots, c_d(x))^\top$ for $x \in \mathcal{X} = [0, D]^d$, $D > 0$. Such problem can be equivalently formulated as

$$\min_{x \in [0,D]^d} \sum_{i=1}^{d} c_i^2(x) \qquad \text{or} \qquad \min_{x \in [0,D]^d} \max_{1 \leq i \leq d} |c_i(x)|.$$

When $c(\cdot)$ is an invertible mapping, it belongs to the hidden convex optimization class. For $d = 2$, $c_1(x) = x_1 - 1$, and $c_2(x) = x_2 - 2x_1^2 + 1$, we demonstrate its countour plot in Appendix D.

D.2. Minimizing posinomial functions [9, 23]

In power control in communication systems and optimal doping profile problems [9, 23], one often needs to minimize posinomial functions $F(\cdot) : \mathbb{R}^d_+ \to \mathbb{R}$ of the following form

$$F(x) = \sum_{k=1}^{K} b_k x_1^{a_{1k}} \cdots x_d^{a_{dk}},$$

where $b_k > 0$ and $a_{ik} \in \mathbb{R}$ for all $k = 1, \cdots, K$, $i = 1, \cdots, d$. The function $F(\cdot)$ is non-convex, but it is well-known that it admits a convex reformulation via a variable change $u = c(x) := [\log(x_1), \ldots, \log(x_d)]^\top$. The convex reformulation is of the form

$$H(u) = F(c^{-1}(u)) = \sum_{k=1}^{K} b_k e^{a_{1k} u_1} \cdots e^{a_{dk} u_d},$$

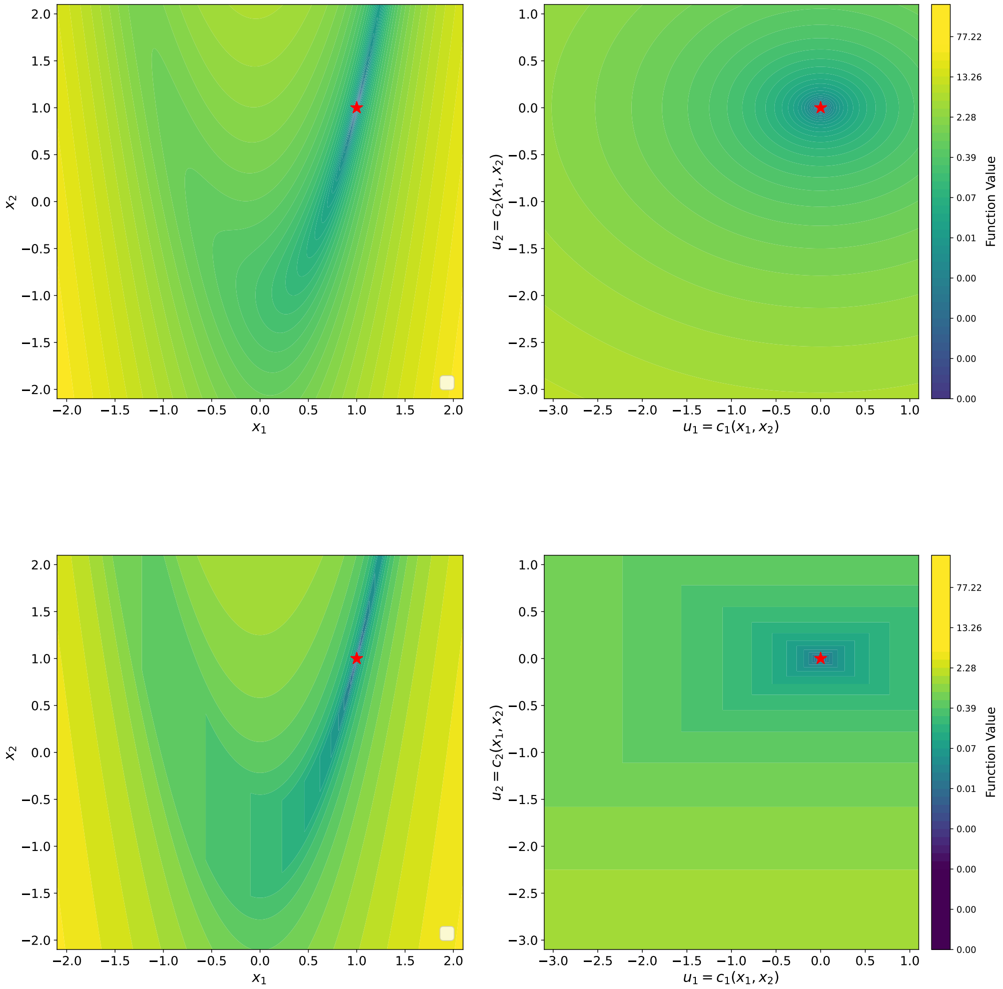

Figure 1: The countour plots of the functions $F(x) = \frac{1}{4}(x_1 - 1)^2 + \frac{1}{2}(2x_1^2 - x_2 - 1)^2$ (top), and $F(x) = \max\left\{\frac{1}{4}|x_1 - 1|, \frac{1}{2}|2x_1^2 - x_2 - 1|\right\}$ (bottom), $x = (x_1, x_2)^\top$ The left plots present the conour plots in the original space $\mathcal{X}$ and the right plots illustrate the reformulated space $\mathcal{U}$. The red star denotes the global minimum.

where $H(\cdot)$ is convex. One can easily see that the above problem is hidden convex and it is possible to verify assumptions in Section 2 if we add a convex compact constraint $\mathcal{X}$ (e.g., a box constraint) to this problem.

D.3. System level synthesis in optimal control [2]

Consider a linear-time-varying system

$$x(t+1) = A_t\, x(t) + B_t\, u(t) + w(t), \qquad t = 0, \ldots, T,$$

where $x(t) \in \mathbb{R}^n$ is a state, $u(t) \in \mathbb{R}^p$ is a control input, and $w(t) \in \mathbb{R}^n$ is an exogenous disturbance process, and $x(0), w(t) \sim \mathcal{N}(0, \Sigma)$ are independent for $t = 0, \ldots, T$. Matrices $A_t \in \mathbb{R}^{n \times n}$ and $B_t \in \mathbb{R}^{n \times p}$ determine the system dynamics. Define $\mathbf{x} = (x(0), \ldots, x(T))^\top$, $\mathbf{u} = (u(0), \ldots, u(T))^\top$, $\mathbf{w} = (x(0), w(0), \ldots, w(T-1))^\top$, and consider a time varying controller of the form $u(t) = \sum_{i=0}^{t} K(t, t-i) x(i)$, which depends on a control matrix

$$\mathbf{K} = \begin{bmatrix} K(0,0) & & & \\ K(1,1) & K(1,0) & & \\ \vdots & \ddots & \ddots & \\ K(T,T) & \cdots & K(T,1) & K(T,0) \end{bmatrix}.$$

The goal of the system level synthesis is to find a control policy to minimize some loss functions, e.g., quadratic in $\mathbf{x}$ and $\mathbf{u}$: $F(\mathbf{K}) := \mathbb{E}\left[\mathbf{x}^\top \mathcal{Q} \mathbf{x} + \mathbf{u}^\top \mathcal{R} \mathbf{u}\right]$, where $\mathcal{Q} = \mathrm{diag}(Q_0, \ldots, Q(T)) \succeq 0$ and $\mathcal{R} = \mathrm{diag}(R_0, \ldots, R(T)) \succeq 0$.

Despite the fact that $F(\cdot)$ is convex in both $\mathbf{x}$ and $\mathbf{u}$, it is non-convex in the decision variable $\mathbf{K}$. Nevertheless, it admits a convex reformulation [2] of the form

$$\min_{\Phi_{\mathbf{x}}, \Phi_{\mathbf{u}}} H(\Phi_{\mathbf{x}}, \Phi_{\mathbf{u}}), \quad \text{s.t. } \mathbf{M} \begin{bmatrix} \Phi_{\mathbf{x}} \\ \Phi_{\mathbf{u}} \end{bmatrix} = I, \quad \Phi_{\mathbf{x}}, \Phi_{\mathbf{u}} \text{ are lower-block-triangular,}$$

where $\Phi_{\mathbf{x}}, \Phi_{\mathbf{u}} \in \mathbb{R}^{(T+1)\times(T+1)}$ are the new variables, $H(\cdot)$ is a strongly-convex function of $\Phi := (\Phi_{\mathbf{x}}, \Phi_{\mathbf{u}})$. $\mathbf{M} \in \mathbb{R}^{(T+1)\times(T+1)}$ is a deterministic matrix, which depends on matrices A_t, B_t, $t = 0, \ldots, T-1$, and I is the identity matrix. It turns out that there exists a bijection between variables $\mathbf{K}$ and Φ subject to the constraints of the reformulated problem. The (inverse of the) map $c(\cdot)$ is given by $\mathbf{K} = c^{-1}(\Phi) := \Phi_{\mathbf{u}} \Phi_{\mathbf{x}}^{-1}$ [2]. Therefore, one can easily verify that the optimization problem over $\mathbf{K}$ is hidden convex.

A number of other problems in optimal control also admit suitable convex reformulations. We refer readers to [8, 66] for more examples.

D.4. Revenue Management and Inventory Control [10, 11]

Consider a booking limit control in a passenger network revenue management problem. The goal is to maximize the revenue by finding an optimal booking limit threshold for each demand class, e.g., flying from New York to Seattle with economy class. Such a problem forms a two-stage stochastic programming such that

$$\begin{aligned} \min_{x \in [0,D]^d} \quad & F(x) := \mathbb{E}_\xi[r^\top (x \wedge \xi) - \mathbb{E}_\eta \Gamma(x \wedge \xi, \eta)] \\ \text{where} \quad & \Gamma(x \wedge \xi, \eta) = \min_{0 \le w \le x \wedge \xi} \{l^\top (x \wedge \xi - w) \mid Aw \le \eta\}, \end{aligned} \tag{9}$$

where d denotes the number of demand classes in the airline networks, $x \in \mathbb{R}^d$ is the booking limit control threshold for each demand class, ξ is the random demand vector (of the same dimension as x) during the reservation stage, $x \wedge \xi$ denotes the number of reservations accepted, and $r^\top(x \wedge \xi)$ denotes revenue collected during the reservation stage with $r \in \mathbb{R}^d$ being the price vector. In the service stage, $\Gamma(x \wedge \xi, \eta)$ denotes the penalty on the airline companies when there are $x \wedge \xi$ number of reservations with plane seats capacity η that is random, w is the actual number of passengers that can get on the plane, l is the penalty vector for declining passengers with reservation to get on the plane. Notice that F is non-convex in x due to the truncation between x and ξ. However, when ξ admits component-wise independent coordinates, this problem admits a convex reformulation via a variable change [10], i.e., $u = c(x) = \mathbb{E}_\xi[x \wedge \xi]$. Note that comparing to previous applications, the transformation function involves unknown distribution and thus is not explicitly known.

For more examples of hidden convex problems in operations research, we refer readers to [29] about supply chain management and [11, 50] about revenue management.

D.5. Convex reinforcement learning [75]

Convex reinforcement learning (RL) problem generalizes the classical RL setting. It bases on a discounted Markov Decision Process $\mathbb{M}(\mathcal{S}, \mathcal{A}, \mathcal{P}, H, \rho, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the (finite) state and action spaces respectively, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition probability kernel (where $\Delta(\mathcal{S})$ denotes the distribution over $\mathcal{S}$), ρ is the initial state distribution and $\gamma \in (0, 1)$ is the discount factor. A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps each state $s \in \mathcal{S}$ to a distribution $\pi(\cdot|s)$ over the action space $\mathcal{A}$. The set of all (stationary) policies is denoted by Π. At each time step $h \in \mathbb{N}$ in a state $s_h \in \mathcal{S}$, the RL agent chooses an action $a_h \in \mathcal{A}$ with probability $\pi(a_h|s_h)$ and the environment transitions to a state s_{h+1} with probability $\mathcal{P}(s_{h+1}|s_h, a_h)$. We denote by $\mathbb{P}_{\rho,\pi}$ the probability distribution of the Markov chain $(s_h, a_h)_{h\in\mathbb{N}}$ induced by the policy π with an initial state distribution ρ. For any policy $\pi \in \Pi$, we define the state-action occupancy measure

$$\lambda^\pi(s, a) := \sum_{h=0}^{+\infty} \gamma^h \mathbb{P}_{\rho,\pi}(s_h = s, a_h = a) \qquad \text{for all } a \in \mathcal{A},\ s \in \mathcal{S}\,. \tag{10}$$

The set of such state-action occupancy (visitation) measures is denoted by $\mathcal{U} := \{\lambda^\pi : \pi \in \Pi\}$.

Different from the classical RL, in convex RL, $H : \mathcal{U} \to \mathbb{R}$ is a general (convex) utility function that maps the state-action occupancy measure to a cost. The goal is to find a policy that minimizes the costs

$$\min_{\pi\in\Pi} F(\pi) := H(\lambda^\pi). \tag{11}$$

Notice that $F(\cdot)$ is not convex in π, but $H(\cdot)$ is convex in the occupancy measure λ^π for several widely-used utility functions. For standard RL, $H(\lambda^\pi) = r^\top\lambda^\pi$ is linear in λ^π, where r is the reward vector. For the pure exploration setting where the goal is to fully explore the transitions in the environment, $H(\lambda^\pi)$ denotes the negative entropy of λ^π, which is also convex [75]. For the imitation learning where the goal is to imitate the expert's behavior given their sampled trajectories, $H(\lambda^\pi)$ denotes the KL-divergence between λ^π and the state-action occupancy measure learned from the expert's sampled trajectories, which is also convex [75]. Thus, the convex RL problem belongs to the hidden convex class with $\mathcal{X} = \Pi$ and $c(x) = \lambda^\pi$ (with $x = \pi$). Under mild (exploration) assumptions on the initial distribution ρ, the constant $\mu_c > 0$ can be estimated [75]. Note that in convex RL, we can only control the policy π and can influence λ^π only implicitly.

Appendix E. Properties of Hidden Convex Optimization

E.1. Globally optimal solution

The following proposition suggests that every stationary point of a hidden convex function is a global minima.

Proposition 7 *Let $F(\cdot)$ be hidden convex and $\bar{x} \in \mathcal{X}$ be its stationary point. If the map $c(\cdot)$ is differentiable at $\bar{x}$, then $\bar{x}$ is a global solution for* (3)*, i.e., $F(\bar{x}) \leq F(x)$ for any $x \in \mathcal{X}$.*

Proof By the definition of a stationary point and the chain rule [62] (Theorem 10.49), we can write

$$0 \in \partial_x (F + \delta_{\mathcal{X}})(\bar{x}) = \nabla c(\bar{x}) \left(\partial_u H(\bar{u}) + \partial_u \delta_{\mathcal{U}}(\bar{u}) \right), \tag{12}$$

where $\bar{u} = c(\bar{x})$. As the map $c(\cdot)$ is invertible, then $\nabla c(\bar{x}) y = 0$ for some $y \in \mathbb{R}^d$ implies $y = 0$. Thus, we have $0 \in \partial_u H(\bar{u}) + \partial_u \delta_{\mathcal{U}}(\bar{u})$. Since function $H(\cdot)$ is convex, by the sufficient optimality condition, $\bar{u}$ is a globally optimal solution, i.e., $H(\bar{u}) \leq H(u)$ for any $u \in \mathcal{U}$. As a result, we have $F(\bar{x}) = H(\bar{u}) \leq H(u) = F(x) \quad$ for any $x \in \mathcal{X}$. ■

Note that a similar result appeared in [75] under additional smooth assumptions on $H(\cdot)$ and $c(\cdot)$. The analysis above is much simpler and does not require such smoothness.

E.2. Connections with gradient dominated functions

It is natural to ask what is the connection between hidden convex problems and previously studied gradient dominated function classes that also ensure the global convergence of gradient-based algorithms [40]. Unfortunately, the exact characterization is difficult to establish in the constrained setting. With the following proposition, we show that a problem satisfies the global KŁ condition if it is hidden strongly convex ($\mu_H > 0$).

Proposition 8 *Let $F(\cdot)$ be differentiable, hidden strongly convex ($\mu_H > 0$), and the map $c(\cdot)$ be differentiable on $\mathcal{X}$, then the optimization problem satisfies the global KŁ condition.*

$$\min_{h_x \in \partial \delta_{\mathcal{X}}(x)} \|\nabla F(x) + h_x\|^2 \geq 2\mu_H \mu_c^2 \left(F(x) - F^* \right) \qquad \textit{for all } x \in \mathcal{X}. \tag{13}$$

Proof Since $H(\cdot)$ is differentiable and strongly convex on a convex set $\mathcal{U}$, then it satisfies the following Proximal-PŁ condition [40] (Appendix F)

$$\mathcal{D}_{\delta_{\mathcal{U}}}(u, \mu_H) \geq 2\mu_H (H(u) - H(u^*)) \qquad \text{for all } u \in \mathcal{U},$$

where $\mathcal{D}_{\delta_{\mathcal{U}}}(u, \mu_H) := -2\mu_H \min_{v \in \mathcal{U}} \left\{ \langle \nabla H(u), v - u \rangle + \frac{\mu_H}{2} \|u - v\|^2 \right\}$. The inequality above implies global KŁ condition, see Appendix G in [40],

$$\min_{h_u \in \partial \delta_{\mathcal{U}}(u)} \|\nabla H(u) + h_u\|^2 \geq 2\mu_H \left(H(u) - H(u^*) \right).$$

Combined with chain rule, we have for any $u = c(x)$ such that

$$\begin{aligned} \min_{h_x \in \partial \delta_{\mathcal{X}}(x)} \|\nabla F(x) + h_x\|^2 &\geq \min_{h_u \in \partial \delta_{\mathcal{U}}(u)} \|\nabla c(x)(\nabla H(u) + h_u)\|^2 \\ &\geq 2\mu_H \mu_c^2 \left(F(x) - F(x^*) \right), \end{aligned}$$

which concludes the proof. ■

Note that even in this restrictive case when $\mu_H > 0$, it is unclear to us how condition (13) can be used to establish global convergence of Projected SGD. To our knowledge, under the assumption (13) no analysis of stochastic gradient methods appear in the literature even for smooth $F(\cdot)$.

In the more interesting case when $H(\cdot)$ is merely convex, the above argument fails since the global KŁ condition becomes vacuous when $\mu_H = 0$.

E.3. Key inequalities for analysis of gradient methods

The following observations are the key for deriving global convergence guarantees on hidden convex optimization problems.

Proposition 9 *Let $F(\cdot)$ be hidden convex with $\mu_H \geq 0$. For any $\alpha \in [0,1]$, $x^* \in \mathcal{X}^*$ and $x \in \mathcal{X}$, define $x_\alpha := c^{-1}\left((1-\alpha)c(x) + \alpha c(x^*)\right)$. Then*

$$F(x_\alpha) \leq (1-\alpha)F(x) + \alpha F(x^*) - \frac{(1-\alpha)\alpha\mu_H}{2}\left\|c(x) - c(x^*)\right\|^2, \tag{14}$$

$$\|x_\alpha - x\| \leq \frac{\alpha}{\mu_c}\left\|c(x) - c(x^*)\right\|. \tag{15}$$

Proof By (strong) convexity of $H(\cdot)$ and convexity of $\mathcal{U}$, we have

$$\begin{aligned} F(x_\alpha) &= F(c^{-1}\left((1-\alpha)c(x) + \alpha c(x^*)\right)) \\ &= H((1-\alpha)c(x) + \alpha c(x^*)) \\ &\leq (1-\alpha)H(c(x)) + \alpha H(c(x^*)) - \frac{(1-\alpha)\alpha\mu_H}{2}\left\|c(x) - c(x^*)\right\|^2 \\ &= (1-\alpha)F(x) + \alpha F(x^*) - \frac{(1-\alpha)\alpha\mu_H}{2}\left\|c(x) - c(x^*)\right\|^2. \end{aligned}$$

where the inequality uses the fact that $\mathcal{U}$ is a convex set and that $(1-\alpha)c(x) + \alpha c(x^*) \in \mathcal{U}$ for any $x \in \mathcal{X}$. By definition of x_α and (4), we derive

$$\|x_\alpha - x\| = \left\|c^{-1}\left((1-\alpha)c(x) + \alpha c(x^*)\right) - c^{-1}(c(x))\right\| \leq \frac{1}{\mu_c}\left\|\alpha(c(x) - c(x^*))\right\|.$$

■

Appendix F. Proofs for Stochastic Subgradient Method

The following (descent like) lemma is the essential for the proofs of global convergence in Theorems 3 and 11.

Lemma 10 *Let C.1, C.2, A.1 and A.2 hold with $\mu_H \geq 0$. Set $\rho = 2\ell$, $\eta \leq \frac{1}{2\ell}$. Define $\hat{x}^t :=$ $\mathrm{prox}_{\Phi/\rho}(x^t)$. Then for any $0 < \alpha \leq \eta\ell$ and $t \geq 0$*

$$\Lambda_{t+1} \leq (1-\alpha)\Lambda_t + \left(\frac{3\alpha^2}{2\mu_c^2\eta} - \frac{(1-\alpha)\alpha\mu_H}{2}\right)\mathbb{E}\left[\left\|c(\hat{x}^t) - c(x^*)\right\|^2\right] + 4\ell\eta^2 G_F^2.$$

Proof By the definition of $\hat{x}^{t+1}$, we have for any $z \in \mathcal{X}$

$$\begin{aligned}\mathbb{E}\left[\Phi_{1/\rho}\left(x^{t+1}\right)\right] &= \mathbb{E}\left[\Phi\left(\hat{x}^{t+1}\right) + \frac{\rho}{2}\left\|\hat{x}^{t+1} - x^{t+1}\right\|^2\right] \\ &\overset{(i)}{\leq} \mathbb{E}\left[\Phi\left(z\right) + \frac{\rho}{2}\left\|z - x^{t+1}\right\|^2\right] \\ &\overset{(ii)}{\leq} \mathbb{E}\left[\Phi\left(z\right) + \left(1+s\right)\frac{\rho}{2}\left\|\hat{x}^{t} - x^{t+1}\right\|^2 + \left(1+\frac{1}{s}\right)\frac{\rho}{2}\left\|\hat{x}^{t} - z\right\|^2\right] \\ &\overset{(iii)}{\leq} \mathbb{E}\left[\Phi\left(z\right) + \left(1+s\right)\left(1-\eta\rho\right)\frac{\rho}{2}\left\|\hat{x}^{t} - x^{t}\right\|^2\right] \\ &\quad + \left(1+\frac{1}{s}\right)\frac{\rho}{2}\mathbb{E}\left[\left\|\hat{x}^{t} - z\right\|^2\right] + \left(1+s\right)\rho\eta^2 G_F^2,\end{aligned}$$

where in (i) we use the optimality of $\hat{x}^{t+1}$, (ii) follows from Young's inequality for any $s > 0$, and in (iii) we apply the result of Lemma 25. We now select $s = \eta\rho/2$, which guarantees $\left(1+s\right)\left(1-\eta\rho\right) \leq 1 - \eta\rho/2$, $1 + s \leq 2$, and $1 + 1/s \leq 3/(\eta\rho)$. Thus

$$\mathbb{E}\left[\Phi_{1/\rho}\left(x^{t+1}\right)\right] \quad \leq \quad \mathbb{E}\left[F\left(z\right) + \left(1 - \frac{\eta\rho}{2}\right)\frac{\rho}{2}\left\|\hat{x}^{t} - x^{t}\right\|^2\right] + \frac{3}{2\eta}\mathbb{E}\left[\left\|\hat{x}^{t} - z\right\|^2\right] + 2\rho\eta^2 G_F^2.$$

We are now ready to utilize the properties of hidden convex functions to bound $F(z)$ and $\left\|\hat{x}^{t} - z\right\|^2$ for some specific choice of $z \in \mathcal{X}$. By Proposition 9, we have for $z = \hat{x}^{t}_{\alpha} := c^{-1}((1-\alpha)c(\hat{x}^t) + \alpha c(x^*))$

$$F(z) \leq (1-\alpha)F(\hat{x}^t) + \alpha F(x^*) - \frac{(1-\alpha)\alpha\mu_H}{2}\left\|c(\hat{x}^t) - c(x^*)\right\|^2,$$

$$\left\|z - \hat{x}^t\right\|^2 \leq \frac{\alpha^2}{\mu_c^2}\left\|c(\hat{x}^t) - c(x^*)\right\|^2.$$

Combining three inequalities above, we have

$$\begin{aligned}\mathbb{E}\left[\Phi_{1/\rho}(x^{t+1})\right] \quad &\leq \quad (1-\alpha)\mathbb{E}\left[F(\hat{x}^t)\right] + \alpha F(x^*) + \left(1 - \frac{\eta\rho}{2}\right)\frac{\rho}{2}\mathbb{E}\left[\left\|\hat{x}^{t} - x^{t}\right\|^2\right] + 2\rho\eta^2 G_F^2 \\ &\quad + \left(\frac{3\alpha^2}{2\mu_c^2\eta} - \frac{(1-\alpha)\alpha\mu_H}{2}\right)\mathbb{E}\left[\left\|c(\hat{x}^t) - c(x^*)\right\|^2\right] \\ &\leq \quad (1-\alpha)\mathbb{E}\left[\Phi_{1/\rho}(x^t)\right] + \alpha F(x^*) + 2\rho\eta^2 G_F^2 \\ &\quad + \left(\frac{3\alpha^2}{2\mu_c^2\eta} - \frac{(1-\alpha)\alpha\mu_H}{2}\right)\mathbb{E}\left[\left\|c(\hat{x}^t) - c(x^*)\right\|^2\right],\end{aligned}$$

where the last inequality holds since $1 - \frac{\eta\rho}{2} \leq 1 - \alpha$ (by the choice $\alpha \leq \eta\ell$, $\rho = 2\ell$) and recognizing $\Phi_{1/\rho}(x^t)$. Subtracting $F(x^*)$ from both sides, we conclude the proof. ■

F.1. Hidden Convex Setting

Proof [Theorem 3] Setting $\mu_H = 0$ in Lemma 10 and leveraging compactness of $\mathcal{U}$, we have

$$\Lambda_{t+1} \quad \leq \quad (1-\alpha)\Lambda_t + \frac{3D_{\mathcal{U}}^2\alpha^2}{2\mu_c^2\eta} + 4\ell\eta^2 G_F^2.$$

Unrolling the recursion for $t = 0$ to $t = T-1$, we get

$$\Lambda_T \quad \leq \quad (1-\alpha)^T \Lambda_0 + \frac{3D_{\mathcal{U}}^2 \alpha}{2\mu_c^2 \eta} + \frac{4\ell \eta^2 G_F^2}{\alpha} \leq \varepsilon,$$

where the last step holds by setting $\alpha = \min\left\{\eta\ell, \frac{2\varepsilon\mu_c^2\eta}{3D_{\mathcal{U}}^2}, \frac{\sqrt{16\ell}\mu_c G_F \eta^{3/2}}{\sqrt{3}D_{\mathcal{U}}}\right\}$ after $T = \frac{1}{\alpha}\log\left(\frac{3\Lambda_0}{\varepsilon}\right) = \widetilde{\mathcal{O}}\left(\frac{\ell D_{\mathcal{U}}^2}{\mu_c^2 \varepsilon} + \frac{\ell D_{\mathcal{U}}^4 G_F^2}{\mu_c^4 \varepsilon^3}\right)$. ■

F.2. Hidden Strongly Convex Setting

The following theorem presents a stronger result in the case when $H(\cdot)$ is additionally strongly convex.

Theorem 11 (Strongly convex $H(\cdot)$) *Let C.1, C.2, A.1 and A.2 hold with $\mu_H > 0$. Then for any $\eta \leq \frac{1}{2\ell}$, and $\alpha \leq \min\left\{\eta\ell, \frac{\eta\mu_c^2\mu_H}{2}\right\}$, we have for all $T \geq 0$*

$$\Lambda_T \quad \leq \quad (1-\alpha)^T \Lambda_0 + \frac{4\ell\eta^2 G_F^2}{\alpha}.$$

Fix $\varepsilon > 0$, and set the step-size in (6) as $\eta = \min\left\{\frac{1}{2\ell}, \frac{\mu_c^2\mu_H\varepsilon}{10\ell G_F^2}\right\}$. Then $\Lambda_T \leq \varepsilon$ after $T = \widetilde{\mathcal{O}}\left(\frac{\ell}{\mu_c^2\mu_H} + \frac{\ell G_F^2}{\mu_c^4\mu_H^2}\frac{1}{\varepsilon}\right)$ iterations.

Proof We invoke Lemma 10 with $\mu_H > 0$. The choice of α guarantees the coefficient in front of $\mathbb{E}\left[\left\|c(\hat{x}^t) - c(x^*)\right\|^2\right]$ is non-positive and

$$\Lambda_{t+1} \quad \leq \quad (1-\alpha)\Lambda_t + 4\ell\eta^2 G_F^2.$$

It remains to conclude the proof by unrolling the recursion and setting the step-size. ■

In the presence of strong convexity, since the optimal $x^* \in \mathcal{X}^*$ is unique, we can establish a strong convergence of the sequence $\{x^t\}_{t\geq 0}$ to x^*.

Corollary 12 *Let the assumptions of Theorem 11 hold and x^T be the output of the method (5) after T iterations. If $2\ell \geq \mu_H\mu_c^2$, then $\frac{\mu_H\mu_c^2}{4}\mathbb{E}\left[\left\|x^T - x^*\right\|^2\right] \leq \varepsilon$ after $T = \widetilde{\mathcal{O}}(\varepsilon^{-1})$.*

Proof Since $H(\cdot)$ is μ_H-strongly convex on $\mathcal{X}$, we have

$$F(\hat{x}^T) - F(x^*) = H(c(\hat{x}^T)) - H(c(x^*)) \geq \frac{\mu_H}{2}\left\|c(\hat{x}^T) - c(x^*)\right\|^2 \geq \frac{\mu_H\mu_c^2}{2}\left\|\hat{x}^T - x^*\right\|^2, \tag{16}$$

where the first inequality follows by the first-order characterization of strong convexity and the optimality condition, and the last inequality holds by C.2.

Recall that $\Lambda_T = \mathbb{E}\left[F(\hat{x}^T) - F(x^*) + \frac{\rho}{2}\left\|\hat{x}^T - x^T\right\|^2\right]$ with $\rho = 2\ell$. Then

$$\begin{aligned}\frac{\mu_H\mu_c^2}{4}\mathbb{E}\left[\left\|x^T - x^*\right\|^2\right] &\leq \frac{\mu_H\mu_c^2}{2}\mathbb{E}\left[\left\|\hat{x}^T - x^*\right\|^2\right] + \frac{\mu_H\mu_c^2}{2}\mathbb{E}\left[\left\|\hat{x}^T - x^T\right\|^2\right] \\ &\leq \mathbb{E}\left[F(\hat{x}^T) - F(x^*)\right] + \ell\,\mathbb{E}\left[\left\|\hat{x}^T - x^T\right\|^2\right] = \Lambda_T \leq \varepsilon,\end{aligned}$$

where the second inequality holds by (16) and $2\ell \geq \mu_H\mu_c^2$. The last step follows by Theorem 11. ■

Appendix G. Proofs for Projected SGD

Replacing Lemma 25 with Lemma 26 in the proof of Lemma 10 of the previous section, we are able to derive the following results under assumptions A.1' and A.2'

Lemma 13 *Let C.1, C.2, A.1' and A.2' hold with $\mu_H \geq 0$. Set $\rho = 4L$, $\eta \leq \frac{2}{9L}$. Define $\hat{x}^t := \mathrm{prox}_{\Phi/\rho}(x^t)$. Then for any $0 < \alpha \leq 2\eta L$ and $t \geq 0$*

$$\Lambda_{t+1} \leq (1-\alpha)\Lambda_t + \rho\eta^2\sigma^2 + \left(\frac{3\alpha^2}{2\mu_c^2\eta} - \frac{(1-\alpha)\alpha\mu_H}{2}\right)\mathbb{E}\left[\left\|c(\hat{x}^t) - c(x^*)\right\|^2\right].$$

Using the above lemma, we provide a refined analysis of P-SGD in the differentiable setting with smoothness and bounded variance.

G.1. Hidden Convex Setting

We start with the case of convex $H(\cdot)$.

Proof [Theorem 5] By setting $\mu_H = 0$ in Lemma 13 and using the bound on the size of $\mathcal{U}$, we have

$$\Lambda_{t+1} \leq (1-\alpha)\Lambda_t + \frac{3D_{\mathcal{U}}^2\alpha^2}{2\mu_c^2\eta} + \rho\eta^2\sigma^2.$$

Unrolling the recursion for $t = 0$ to $t = T-1$, we get

$$\Lambda_T \leq (1-\alpha)^T\Lambda_0 + \frac{3D_{\mathcal{U}}^2\alpha}{2\mu_c^2\eta} + \frac{4L\eta^2\sigma^2}{\alpha} \leq \varepsilon,$$

where the last step holds by setting $\alpha = \min\left\{2\eta L, \frac{2\varepsilon\mu_c^2\eta}{3D_{\mathcal{U}}^2}, \frac{\sqrt{8L}\mu_c\sigma\eta^{3/2}}{\sqrt{3}D_{\mathcal{U}}}\right\}$, and $\eta = \frac{2}{9L}\cdot\min\left\{1, \frac{\mu_c^2\varepsilon^2}{12D_{\mathcal{U}}^2\sigma^2}\right\}$, after $T = \frac{1}{\alpha}\log\left(\frac{3\Lambda_0}{\varepsilon}\right) = \widetilde{\mathcal{O}}\left(\frac{LD_{\mathcal{U}}^2}{\mu_c^2\varepsilon} + \frac{LD_{\mathcal{U}}^4\sigma^2}{\mu_c^4\varepsilon^3}\right)$. ■

Similar to Corollary 4, we can show that x^T is close to an ε-global optimal solution. But in the case of smooth $F(\cdot)$, we can derive a stronger result after applying one (post-processing) step of mini-batch P-SGD. The next corollary presents the results.

Corollary 14 *Let the assumptions of Theorem 5 hold and $G_F > 0$ be the Lipschitz constant of $F(\cdot)$ over $\mathcal{X}$. Set $x^{T+1} = \Pi_{\mathcal{X}}\left(x^T - \frac{1}{3L}\frac{1}{B_0}\sum_{i=1}^{B_0}\nabla F(x^T, \xi_i^T)\right)$, where $B_0 \geq \min\{1, \left(\frac{G_F\sigma}{3L\varepsilon}\right)^2\}$, and x^T is the output of the method (6) applied with batch-size $B = 1$ after T iterations. Then $\mathbb{E}\left[F(x^{T+1}) - F(x^*)\right] \leq 2\,\varepsilon$ after $T = \widetilde{\mathcal{O}}\left(\frac{LD_{\mathcal{U}}^2}{\mu_c^2\varepsilon} + \frac{LD_{\mathcal{U}}^4\sigma^2}{\mu_c^4\varepsilon^3}\right)$.*

Proof Define $x_+^T := \Pi_{\mathcal{X}}(x^T - \frac{1}{\rho - L}\nabla F(x^T))$, $\rho = 4L$. Notice that $F(x_+^T) = \Phi(x_+^T) \le \Phi_{1/\rho}(x^T)$, where the inequality follows by [65, Poposition 2.5-(i)] with $\gamma := (\rho - L)^{-1}$. Therefore, Theorem 5 implies that

$$\mathbb{E}\left[F(x_+^T) - F(x^*)\right] \le \mathbb{E}\left[\Phi_{1/\rho}(x^T) - F(x^*)\right] \le \varepsilon,$$

On the other hand, the post-processing step guarantees

$$\begin{aligned} & \mathbb{E}\left[F(x^{T+1}) - F(x_+^T)\right] \le G_F\, \mathbb{E}\left[\|x^{T+1} - x_+^T\|\right] \\ \le \quad & G_F \mathbb{E}\left[\left\|\Pi_{\mathcal{X}}\left(x^T - \frac{1}{3L}\frac{1}{B_0}\sum_{i=1}^{B_0}\nabla F(x^T, \xi_i^T)\right) - \Pi_{\mathcal{X}}\left(x^T - \frac{1}{3L}\nabla F(x^T)\right)\right\|\right] \\ \le \quad & \frac{G_F}{3L}\mathbb{E}\left[\left\|\frac{1}{B_0}\sum_{i=1}^{B_0}\nabla F(x^T, \xi_i^T) - \nabla F(x^T)\right\|\right] \le \frac{G_F \sigma}{3L\sqrt{B_0}} \le \varepsilon. \end{aligned}$$

Combining the above two inequalities, the result follows. ■

The following corollary shows that if we apply mini-batching at each iteration with sufficiently large batch-size, then the number of iterations required for convergence is reduced to $\widetilde{\mathcal{O}}(\varepsilon^{-1})$.

Corollary 15 *Let the assumptions of Theorem 5 hold and $G_F > 0$ be the Lipschitz constant of $F(\cdot)$ over $\mathcal{X}$. Suppose P-SGD with batch-size B is applied, i.e., $\{x^t\}_{t \ge 0}$ is generated by $x^{t+1} = \Pi_{\mathcal{X}}\left(x^t - \eta\frac{1}{B}\sum_{i=1}^{B}\nabla F(x^t, \xi_i^t)\right)$ with $\eta = \frac{2}{9L}$, $B \ge \min\{1, \frac{D_{\mathcal{U}}^2\sigma^2}{\mu_c^2\varepsilon^2}\}$. Define*

$$x^{T+1} = \Pi_{\mathcal{X}}\left(x^T - \frac{1}{3L}\frac{1}{B_0}\sum_{i=1}^{B_0}\nabla F(x^T, \xi_i^T)\right),$$

where $\rho = 4L$, and $B_0 \ge \min\{1, \left(\frac{G_F\sigma}{3L\varepsilon}\right)^2\}$. Then $\mathbb{E}\left[F(x^{T+1}) - F(x^*)\right] \le 2\,\varepsilon$ after $T = \widetilde{\mathcal{O}}\left(\frac{LD_{\mathcal{U}}^2}{\mu_c^2\varepsilon}\right)$.

Proof The proof follows from the previous corollary by replacing σ^2 with σ^2/B. ■

However, we want to highlight that the results of Theorems 5 and 14 do not require using large batches of samples at every iteration.

G.2. Hidden Strongly Convex Setting

Similarly to the exposition in Section 3, we present an improved sample complexity result in case of strongly convex $H(\cdot)$.

Theorem 16 (Strongly convex $H(\cdot)$) *Let C.1, C.2, A.1' and A.2' hold with $\mu_H > 0$. Then for any $\eta \leq \frac{2}{9L}$, and $\alpha \leq \min\left\{2\eta L, \frac{\eta\mu_c^2\mu_H}{2}\right\}$, we have for all $T \geq 0$*

$$\Lambda_T \quad \leq \quad (1-\alpha)^T\Lambda_0 + \frac{4L\eta^2\sigma^2}{\alpha}.$$

Fix $\varepsilon > 0$, and set the step-size in (6) as $\eta = \min\left\{\frac{2}{9L}, \frac{\mu_c^2\mu_H\varepsilon}{10L\sigma^2}\right\}$. Then $\Lambda_T \leq \varepsilon$ after $T = \widetilde{\mathcal{O}}\left(\frac{L}{\mu_c^2\mu_H} + \frac{L\sigma^2}{\mu_c^4\mu_H^2}\frac{1}{\varepsilon}\right)$ iterations.

Proof Similarly to the proof of Theorem 5 we invoke Lemma 13 with $\mu_H > 0$. The choice of α guarantees the coefficient in front of $\mathbb{E}\left[\left\|c(\hat{x}^t) - c(x^*)\right\|^2\right]$ is non-positive and

$$\Lambda_{t+1} \quad \leq \quad (1-\alpha)\Lambda_t + \rho\eta^2\sigma^2.$$

It remains to conclude the proof by unrolling the recursion and setting the step-size to compute the total sample complexity. ■

Similarly to Corollary 12, we can translate convergence in Λ_T to the last iterate point convergence.

Corollary 17 *Let the assumptions of Theorem 16 hold and x^T be the output of the method (6) after T iterations. If $2L \geq \mu_H\mu_c^2$, then $\frac{\mu_H\mu_c^2}{4}\mathbb{E}\left[\left\|x^T - x^*\right\|^2\right] \leq \varepsilon$ after $T = \widetilde{\mathcal{O}}\left(\varepsilon^{-1}\right)$.*

If we apply mini-batch version of P-SGD, then the method converges linearly.

Corollary 18 *Let the assumptions of Theorem 16 hold. Suppose P-SGD with batch-size B is applied, i.e., $x^{t+1} = \Pi_{\mathcal{X}}\left(x^t - \eta\frac{1}{B}\sum_{i=1}^{B}\nabla F(x^t, \xi_i^t)\right)$ with $\eta = \frac{2}{9L}$, $B \geq \min\{1, \frac{\sigma^2}{\mu_c^2\mu_H\varepsilon}\}$. If $2L \geq \mu_H\mu_c^2$, then the sequence $\{x^t\}_{t\geq 0}$ converges linearly to x^*, i.e., we have $\frac{\mu_H\mu_c^2}{4}\mathbb{E}\left[\left\|x^T - x^*\right\|^2\right] \leq \varepsilon$ after $T = \widetilde{\mathcal{O}}\left(\frac{L}{\mu_c^2\mu_H}\right)$ iterations.*[3]

Proof The proof follows from the previous corollary by replacing σ^2 with σ^2/B. ■

Appendix H. Proofs for Projected SGD with Momentum

Lemma 19 *Suppose that C.1, C.2, A.1' and A.2' hold with $\mu_H \geq 0$, and the step-size in (7) satisfies $\eta \leq 1/L$. For any $\alpha \in [0,1]$, it holds that*

$$\begin{aligned} F(x^{t+1}) \quad &\leq \quad (1-\alpha)F(x^t) + \alpha F(x^*) + \left(\frac{\alpha^2}{\mu_c^2\eta} - \frac{(1-\alpha)\alpha\mu_H}{2}\right)\left\|c(x^t) - c(x^*)\right\|^2, \\ &\qquad - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\left\|x^{t+1} - x^t\right\|^2 + \frac{\eta}{2}\left\|g^t - \nabla F(x^t)\right\|^2. \end{aligned} \tag{17}$$

[3] Notice that $L/(\mu_H\mu_c^2)$ is the analogue of the condition number in convex optimization.

Proof By the updating rule of x^{t+1} and using Lemma 27 with $\phi(y) = \langle g^t, y\rangle$, $x = x^t$, and $x^+ = x^{t+1}$, we have for any $y = z \in \mathcal{X}$ that

$$\langle g^t, x^{t+1} - z\rangle + \frac{1}{2\eta} \left\|x^{t+1} - x^t\right\|^2 \leq \frac{1}{2\eta} \left\|z - x^t\right\|^2 - \frac{1}{2\eta} \left\|z - x^{t+1}\right\|^2 . \tag{18}$$

By the smoothness of $F(\cdot)$, we derive

$$\begin{aligned}
F(x^{t+1}) &\leq F(x^t) + \langle \nabla F(x^t), x^{t+1} - x^t\rangle + \frac{L}{2} \left\|x^{t+1} - x^t\right\|^2 \\
&= F(x^t) + \langle g^t, x^{t+1} - x^t\rangle + \frac{1}{2\eta} \left\|x^{t+1} - x^t\right\|^2 \\
&\qquad + \langle \nabla F(x^t) - g^t, x^{t+1} - x^t\rangle - \left(\frac{1}{2\eta} - \frac{L}{2}\right) \left\|x^{t+1} - x^t\right\|^2 \\
&\overset{(i)}{\leq} F(x^t) + \langle g^t, z - x^t\rangle + \frac{1}{2\eta} \left\|z - x^t\right\|^2 - \frac{1}{2\eta} \left\|z - x^{t+1}\right\|^2 \\
&\qquad + \langle \nabla F(x^t) - g^t, x^{t+1} - x^t\rangle - \left(\frac{1}{2\eta} - \frac{L}{2}\right) \left\|x^{t+1} - x^t\right\|^2 \\
&= F(x^t) + \langle \nabla F(x^t), z - x^t\rangle + \frac{1}{2\eta} \left\|z - x^t\right\|^2 - \frac{1}{2\eta} \left\|z - x^{t+1}\right\|^2 \\
&\qquad + \langle \nabla F(x^t) - g^t, x^{t+1} - z\rangle - \left(\frac{1}{2\eta} - \frac{L}{2}\right) \left\|x^{t+1} - x^t\right\|^2 \\
&\overset{(ii)}{\leq} F(x^t) + \langle \nabla F(x^t), z - x^t\rangle + \frac{1}{2\eta} \left\|z - x^t\right\|^2 + \frac{\eta}{2} \left\|g^t - \nabla F(x^t)\right\|^2 \\
&\qquad - \left(\frac{1}{2\eta} - \frac{L}{2}\right) \left\|x^{t+1} - x^t\right\|^2 \\
&\overset{(iii)}{\leq} F(z) + \frac{L}{2} \left\|z - x^t\right\|^2 + \frac{1}{2\eta} \left\|z - x^t\right\|^2 + \frac{\eta}{2} \left\|g^t - \nabla F(x^t)\right\|^2 \\
&\qquad - \left(\frac{1}{2\eta} - \frac{L}{2}\right) \left\|x^{t+1} - x^t\right\|^2 ,
\end{aligned}$$

where (i) follows from (18), (ii) holds by Young's inequality, i.e., $\langle a, b\rangle \leq \frac{\eta}{2} \|a\|^2 + \frac{1}{2\eta} \|b\|^2$ with $a = \nabla F(x^t) - g^t$, $b = x^{t+1} - z$, (iii) holds by the smoothness of $F(\cdot)$, i.e., $-\frac{L}{2} \left\|z - x^t\right\|^2 \leq F(x^t) - F(z) - \langle \nabla F(x^t), z - x^t\rangle$.

We are now ready to utilize the properties of hidden convex functions to bound $F(z)$ and $\left\|z - x^t\right\|^2$ for some specific choice of $z \in \mathcal{X}$. We select $z := x^t_\alpha = c^{-1}((1-\alpha)c(x^t) + \alpha c(x^*)) \in \mathcal{X}$, for some $\alpha \in [0, 1]$, and $x^* \in \mathcal{X}^*$. By Proposition 9, we have for $\mu_H \geq 0$

$$F(z) \leq (1-\alpha)F(x^t) + \alpha F(x^*) - \frac{(1-\alpha)\alpha\mu_H}{2} \left\|c(x^t) - c(x^*)\right\|^2 ,$$

$$\left\|z - x^t\right\|^2 \leq \frac{\alpha^2}{\mu_c^2} \left\|c(x^t) - c(x^*)\right\|^2 .$$

Combining the three inequalities above and utilizing the assumption $\eta \leq 1/L$, we complete the proof of (17). ■

The next lemma controls the error between the momentum gradient estimator g^t and the true gradient $\nabla f(x^t, \xi^t)$. We borrow the analysis from [27], where they demonstrate the convergence of SGD with momentum for general non-convex unconstrained problems (Lemma 2 in Appendix F and Theorem 8 in Appendix J [27]).

Lemma 20 *Let $\beta \in (0,1]$ and the sequence $\left\{g^t\right\}_{t\geq 0}$ be updated via (7) Then*

$$\mathbb{E}\left[\left\|g^{t+1} - \nabla F(x^{t+1})\right\|^2\right] \leq (1-\beta)\mathbb{E}\left[\left\|g^t - \nabla F(x^t)\right\|^2\right] + \frac{3L^2}{\beta}\mathbb{E}\left[\left\|x^t - x^{t+1}\right\|^2\right] + \beta^2\sigma^2.$$

Proof Using the updating rule of g^{t+1} and the unbiasedness of stochastic gradients, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|g^{t+1} - \nabla F(x^{t+1})\right\|^2\right] &= \mathbb{E}\left[\left\|(1-\beta)g^t + \beta\nabla f(x^{t+1}, \xi^{t+1}) - \nabla F(x^{t+1})\right\|^2\right] \\
&= (1-\beta)^2\mathbb{E}\left[\left\|g^t - \nabla F(x^{t+1})\right\|^2\right] \\
&\quad + \beta^2\mathbb{E}\left[\left\|\nabla f(x^{t+1}, \xi^{t+1}) - \nabla F(x^{t+1})\right\|^2\right] \\
&\leq (1-\beta)^2\left(1 + \frac{\beta}{2}\right)\mathbb{E}\left[\left\|g^t - \nabla F(x^t)\right\|^2\right] \\
&\quad + \left(1 + \frac{2}{\beta}\right)\mathbb{E}\left[\left\|\nabla F(x^t) - \nabla F(x^{t+1})\right\|^2\right] + \beta^2\sigma^2 \\
&\leq (1-\beta)\mathbb{E}\left[\left\|g^t - \nabla F(x^t)\right\|^2\right] + \frac{3L^2}{\beta}\mathbb{E}\left[\left\|x^t - x^{t+1}\right\|^2\right] \\
&\quad + \beta^2\sigma^2,
\end{aligned}$$

where the first inequality uses Young's inequality and the bound of the variance of stochastic gradients, and the last step uses the Lipschitz continuity of the gradient and the fact that $(1-\beta)\left(1+\frac{\beta}{2}\right) \leq 1$ for all $\beta \in (0,1]$. ■

H.1. Hidden Convex Setting

Proof [Theorem 6] By Lemma 19, subtracting $F(x^*)$ from both sides of (17), setting $\mu_H = 0$, and taking the expectation, we have for any $\eta \leq 1/L$ that

$$\begin{aligned}
\mathbb{E}\left[F(x^{t+1}) - F(x^*)\right] &\leq (1-\alpha)\mathbb{E}\left[F(x^t) - F(x^*)\right] + \frac{\alpha^2}{\mu_c^2\eta}\mathbb{E}\left[\left\|c(x^t) - c(x^*)\right\|^2\right] \\
&\quad - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\left\|g^t - \nabla F(x^t)\right\|^2\right] \\
&\leq (1-\alpha)\mathbb{E}\left[F(x^t) - F(x^*)\right] + \frac{\alpha^2 D_{\mathcal{U}}^2}{\mu_c^2\eta} \\
&\quad - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\left\|g^t - \nabla F(x^t)\right\|^2\right],
\end{aligned}$$

where the second inequality uses boundedness of $\mathcal{U}$.

Summing up the inequality above with a $\frac{\eta}{\beta}$ multiple of the result of Lemma 20, we recognize the Lyapunov function Λ_t^{HB} defined in (8), and derive

$$\begin{aligned}
\Lambda_{t+1}^{\mathrm{HB}} &\leq \Lambda_t^{\mathrm{HB}} - \alpha \mathbb{E}\left[F(x^t) - F(x^*)\right] - \frac{\eta}{2}\mathbb{E}\left[\left\|g^t - \nabla F(x^t)\right\|^2\right] \\
&\quad + \frac{\alpha^2 D_{\mathcal{U}}^2}{\mu_c^2 \eta} - \left(\frac{1}{2\eta} - \frac{L}{2} - \frac{3L^2\eta}{\beta^2}\right)\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right] + \beta\eta\sigma^2 \\
&\leq (1-\alpha)\Lambda_t^{\mathrm{HB}} + \frac{\alpha^2 D_{\mathcal{U}}^2}{\mu_c^2 \eta} + \beta\eta\sigma^2,
\end{aligned}$$

where the last step holds for $\alpha \leq \beta/2$ and $\eta \leq \frac{\beta}{4L}$. Unrolling the recursion from $t = 0$ to $t = T-1$ and choosing $\eta = \frac{\beta}{4L}$, we obtain

$$\begin{aligned}
\Lambda_T^{\mathrm{HB}} &\leq (1-\alpha)^T \Lambda_0^{\mathrm{HB}} + \frac{\alpha D_{\mathcal{U}}^2}{\mu_c^2 \eta} + \frac{\beta\eta\sigma^2}{\alpha} \\
&\leq (1-\alpha)^T \Lambda_0^{\mathrm{HB}} + \frac{4LD_{\mathcal{U}}^2}{\mu_c^2}\frac{\alpha}{\beta} + \frac{\sigma^2}{4L}\frac{\beta^2}{\alpha} \leq \varepsilon,
\end{aligned}$$

where the last inequality holds by setting $\alpha = \min\left\{\frac{\beta}{2}, \frac{3\mu_c^2}{2LD_{\mathcal{U}}^2}\beta\varepsilon, \frac{\sigma\mu_c}{4LD_{\mathcal{U}}}\beta^{\frac{3}{2}}\right\}, \beta = \min\left\{1, \frac{\mu_c^2}{9D_{\mathcal{U}}^2\sigma^2}\varepsilon^2\right\}$, and the number of iterations as

$$T = \frac{1}{\alpha}\log\left(\frac{3\Lambda_0^{\mathrm{HB}}}{\varepsilon}\right) = \widetilde{\mathcal{O}}\left(\frac{LD_{\mathcal{U}}^2}{\mu_c^2}\frac{1}{\varepsilon} + \frac{LD_{\mathcal{U}}^4\sigma^2}{\mu_c^4}\frac{1}{\varepsilon^3}\right).$$

■

Similar to P-SGD, the momentum variant also supports mini-batching and the iteration complexity is $\mathcal{O}(\varepsilon^{-1})$ when sufficiently large mini-batch is utilized.

Corollary 21 *Let the assumptions of Theorem 6 hold. Suppose momentum variant of P-SGD with batch-size B is applied with $\eta = \frac{\beta}{4L}$, $\eta \in (0,1]$, $B \geq \min\{1, \frac{9D_{\mathcal{U}}^2\sigma^2}{\mu_c^2\varepsilon^2}\}$. Then $\mathbb{E}\left[F(x^{T+1}) - F(x^*)\right] \leq \varepsilon$ after $T = \widetilde{\mathcal{O}}\left(\frac{LD_{\mathcal{U}}^2}{\mu_c^2\varepsilon}\right)$.*

Proof The proof follows immediately from Theorem 6 by replacing σ^2 with σ^2/B. ■

H.2. Hidden Strongly Convex Setting

We conclude the section with the improved result for P-SGD with momentum under strongly convex $H(\cdot)$.

Theorem 22 (Strongly convex $H(\cdot)$) *Let C.1, C.2, A.1' and A.2' hold with $\mu_H > 0$. Then for any $\eta \leq \frac{\beta}{4L}$, $\beta \in (0,1]$, and $\alpha \leq \min\left\{\frac{\beta}{2}, \frac{\mu_c^2 \mu_H \eta}{4}\right\}$, we have for any $T \geq 0$*

$$\Lambda_T^{HB} \leq (1-\alpha)^T \Lambda_0^{HB} + \frac{\beta\eta\sigma^2}{\alpha},$$

where Λ_t^{HB} is given by (8). Fix $\varepsilon > 0$, and set the parameters of algorithm (7) as

$$\eta = \frac{\beta}{4L}, \quad \beta = \min\left\{1, \frac{\mu_c^2 \mu_H}{8\sigma^2}\varepsilon\right\}.$$

Then the scheme (7) returns a point x^T with $\mathbb{E}\left[F(x^T) - F(x^*)\right] \leq \varepsilon$ after

$$T = \widetilde{\mathcal{O}}\left(\frac{L}{\mu_c^2 \mu_H} + \frac{L\sigma^2}{\mu_c^4 \mu_H^2}\frac{1}{\varepsilon}\right).$$

Proof Applying Lemma 19 with $\mu_H > 0$, and setting α small enough allows us to cancel the term invloving $\left\|c(x^t) - c(x^*)\right\|^2$. The rest of the proof is similar to the one of Theorem 16. ■

Again, the convergence rate becomes linear in the number of iteration if sufficiently large minibach is used.

Corollary 23 *Let the assumptions of Theorem 22 hold. Suppose momentum variant of P-SGD with batch-size B is applied with $\eta = \frac{\beta}{4L}$, $\eta \in (0,1]$, $B \geq \min\{1, \frac{8\sigma^2}{\mu_c^2\mu_H\varepsilon}\}$. Then the method converges linearly, i.e., $\mathbb{E}\left[F(x^{T+1}) - F(x^*)\right] \leq \varepsilon$ after $T = \widetilde{\mathcal{O}}\left(\frac{L}{\mu_c^2\mu_H}\right)$.*

Proof The proof follows immediately from Theorem 22 by replacing σ^2 with σ^2/B. ■

Appendix I. Technical Lemma

We report the following three technical lemma from [15] and include their slightly modified proofs for completeness.

Lemma 24 *Let $\rho > \ell$, and for any $x^t \in \mathcal{X}$, define $\hat{x}^t := \operatorname{prox}_{\Phi/\rho}(x^t)$, where $\Phi := F + \delta_{\mathcal{X}}$, and $\hat{g}^t \in \partial F(\hat{x}^t)$. Then $\hat{x}^t = \Pi_{\mathcal{X}}\left(\eta\rho x^t - \eta\hat{g}^t + (1-\eta\rho)\hat{x}^t\right)$.*

Proof By definition of $\hat{x}^t$ and $\Phi(\cdot)$, we have

$$0 \in \partial\left(F + \frac{\rho}{2}\left\|\cdot - x^t\right\|^2 + \delta_{\mathcal{X}}\right)(\hat{x}^t) = \hat{g}^t + \rho\left(\hat{x}^t - x^t\right) + \partial\delta_{\mathcal{X}}\left(\hat{x}^t\right),$$

where the last equality holds, since $F(\cdot) + \frac{\rho}{2}\left\|\cdot - x^t\right\|^2$, and $\delta_{\mathcal{X}}(\cdot)$ are both convex (due to the conic combination rule). Multiplying both sides by $\eta > 0$ and rearranging, we get $z^t := \eta\rho x^t - \eta\hat{g}^t + (1-\eta\rho)\hat{x}^t \in \hat{x}^t + \eta\partial\delta_{\mathcal{X}}\left(\hat{x}^t\right)$. Therefore, by the optimality condition for the proximal sub-problem, we have $\hat{x}^t = \operatorname{prox}_{\eta\delta_{\mathcal{X}}}\left(z^t\right) = \Pi_{\mathcal{X}}(z^t)$. ■

Lemma 25 *Let Assumptions A.1, A.2 hold, and $\rho > \ell$, $\eta \leq 1/\rho$. Then for all $t \geq 0$*

$$\mathbb{E}\left[\left\|x^{t+1} - \hat{x}^t\right\|^2 \mid x^t\right] \leq (1-\eta\rho)\left\|x^t - \hat{x}^t\right\|^2 + 2G_F^2\eta^2$$

Proof Lemma 24 states that for any $\hat{g}^t \in \partial F(\hat{x}^t)$ and $z^t = \eta\rho x^t - \eta\hat{g}^t + (1-\eta\rho)\hat{x}^t$, we have $\hat{x}^t = \Pi_{\mathcal{X}}(z^t)$. Thus, using the update rule of x^{t+1} and non-expansiveness of the projection, we derive

$$\begin{aligned}
&\mathbb{E}\left[\left\|x^{t+1} - \hat{x}^t\right\|^2 \mid x^t\right] = \mathbb{E}\left[\left\|\Pi_{\mathcal{X}}\left(x^t - \eta g(x^t,\xi^t)\right) - \Pi_{\mathcal{X}}\left(z^t\right)\right\|^2 \mid x^t\right] \\
&\leq \mathbb{E}\left[\left\|x^t - \eta g(x^t,\xi^t) - \left(\eta\rho x^t - \eta\hat{g}^t + (1-\eta\rho)\hat{x}^t\right)\right\|^2 \mid x^t\right] \\
&= \mathbb{E}\left[\left\|(1-\eta\rho)\left(x^t - \hat{x}^t\right) - \eta\left(g(x^t,\xi^t) - \hat{g}^t\right)\right\|^2 \mid x^t\right] \\
&\overset{(i)}{=} (1-\eta\rho)^2\left\|x^t - \hat{x}^t\right\|^2 - (1-\eta\rho)\eta\langle g^t - \hat{g}^t, x^t - \hat{x}^t\rangle + \eta^2\mathbb{E}\left[\left\|g(x^t,\xi^t) - \hat{g}^t\right\|^2 \mid x^t\right] \\
&\overset{(ii)}{\leq} (1-\eta\rho)^2\left\|x^t - \hat{x}^t\right\|^2 - (1-\eta\rho)\eta\langle g^t - \hat{g}^t, x^t - \hat{x}^t\rangle + 2G_F^2\eta^2 \\
&\overset{(iii)}{\leq} (1-\eta\rho)^2\left\|x^t - \hat{x}^t\right\|^2 - (1-\eta\rho)\eta\ell\left\|x^t - \hat{x}^t\right\|^2 + 2G_F^2\eta^2 \\
&\leq (1-\eta\rho)\left\|x^t - \hat{x}^t\right\|^2 + \eta^2 G_F^2,
\end{aligned}$$

where in (i) use unbiasedness of the gradient estimator. In (ii), we use Young's inequality and A.2, (iii) holds by hypomonotonicity inequality $\langle g^t - \hat{g}^t, x^t - \hat{x}^t\rangle \geq -\ell\left\|x^t - \hat{x}^t\right\|^2$. The last inequality holds by the choice of ρ and η . ■

Lemma 26 *Let Assumptions A.1', A.2' hold, and $\rho = 4L$, $\eta \leq \frac{2}{9L}$. Then for all $t \geq 0$*

$$\mathbb{E}\left[\left\|x^{t+1} - \hat{x}^t\right\|^2 \mid x^t\right] \leq (1-\eta\rho)\left\|x^t - \hat{x}^t\right\|^2 + \sigma^2\eta^2$$

Proof For a differentiable $F(\cdot)$ Lemma 24 implies that for $z^t = \eta\rho x^t - \eta\nabla F\left(\hat{x}^t\right) + (1-\eta\rho)\hat{x}^t$, we have $\hat{x}^t = \Pi_{\mathcal{X}}(z^t)$. Thus, using the update rule of x^{t+1} and non-expansiveness of the projection, we derive

$$
\begin{aligned}
&\mathbb{E}\left[\left\|x^{t+1}-\hat{x}^t\right\|^2 \mid x^t\right]=\mathbb{E}\left[\left\|\Pi_{\mathcal{X}}\left(x^t-\eta \nabla f\left(x^t, \xi^t\right)\right)-\Pi_{\mathcal{X}}\left(z^t\right)\right\|^2 \mid x^t\right] \\
&\leq \mathbb{E}\left[\left\|x^t-\eta \nabla f\left(x^t, \xi^t\right)-\left(\eta \rho x^t-\eta \nabla F\left(\hat{x}^t\right)+(1-\eta \rho) \hat{x}^t\right)\right\|^2 \mid x^t\right] \\
&=\mathbb{E}\left[\left\|(1-\eta \rho)\left(x^t-\hat{x}^t\right)-\eta\left(\nabla f\left(x^t, \xi^t\right)-\nabla F\left(\hat{x}^t\right)\right)\right\|^2 \mid x^t\right] \\
&=\mathbb{E}\left[\left\|(1-\eta \rho)\left(x^t-\hat{x}^t\right)-\eta\left(\nabla F\left(x^t\right)-\nabla F\left(\hat{x}^t\right)\right)-\eta\left(\nabla f\left(x^t, \xi^t\right)-\nabla F\left(x^t\right)\right)\right\|^2 \mid x^t\right] \\
&\overset{(i)}{=}\left\|(1-\eta \rho)\left(x^t-\hat{x}^t\right)-\eta\left(\nabla F\left(x^t\right)-\nabla F\left(\hat{x}^t\right)\right)\right\|^2+\eta^2 \mathbb{E}\left[\left\|\nabla f\left(x^t, \xi^t\right)-\nabla F\left(x^t\right)\right\|^2 \mid x^t\right] \\
&\overset{(ii)}{\leq}\left\|(1-\eta \rho)\left(x^t-\hat{x}^t\right)-\eta\left(\nabla F\left(x^t\right)-\nabla F\left(\hat{x}^t\right)\right)\right\|^2+\eta^2 \sigma^2 \\
&=(1-\eta \rho)^2\left\|x^t-\hat{x}^t\right\|^2-2(1-\eta \rho) \eta\left(x^t-\hat{x}^t, \nabla F\left(x^t\right)-\nabla F\left(\hat{x}^t\right)\right\rangle+\eta^2 \sigma^2 \\
&\quad+\eta^2\left\|\nabla F\left(x^t\right)-\nabla F\left(\hat{x}^t\right)\right\|^2 \\
&\overset{(iii)}{\leq}(1-\eta \rho)^2\left\|x^t-\hat{x}^t\right\|^2+2(1-\eta \rho) \eta L\left\|x^t-\hat{x}^t\right\|^2+\eta^2 L^2\left\|x^t-\hat{x}^t\right\|^2+\eta^2 \sigma^2 \\
&=(1-\eta \rho)\left(1-\eta \rho+2 \eta L+\frac{\eta^2 L^2}{1-\eta \rho}\right)\left\|x^t-\hat{x}^t\right\|^2+\eta^2 \sigma^2 \\
&\leq(1-\eta \rho)\left\|x^t-\hat{x}^t\right\|^2+\eta^2 \sigma^2,
\end{aligned}
$$

where in (i) and (ii) use unbiasedness of the gradient estimator and bounded variance. In (iii), we use Cauchy–Schwarz inequality and smoothness of $F(\cdot)$, i.e., $\left\|\nabla F\left(\hat{x}^t\right)-\nabla F\left(x^t\right)\right\| \leq L\left\|\hat{x}^t-x^t\right\|$. The last inequality holds by the choice of ρ, η and $2\eta L \leq \frac{\eta\rho}{2}$, and $\frac{\eta^2 L}{1-\eta\rho} \leq \frac{\eta\rho}{2}$. ■

The following technical lemma is fairly standard, e.g., see [34].

Lemma 27 *Let $\phi(\cdot)$ be convex and for some $\eta > 0$, $x \in \mathcal{X}$, $x^+ := \arg\min_{y \in \mathcal{X}} \left\{\phi(y) + \frac{1}{2\eta}\|y - x\|^2\right\}$, then*

$$\phi(y)+\frac{1}{2 \eta}\|y-x\|^2 \geq \phi(x^+)+\frac{1}{2 \eta}\left\|x^+-x\right\|^2+\frac{1}{2 \eta}\left\|y-x^+\right\|^2 \quad \textit{for all } y \in \mathcal{X}.$$