Taming Nonconvex Stochastic Mirror Descent with General Bregman Divergence

Ilyas Fatkhullin
ETH Zürich

Niao He
ETH Zürich

Abstract

This paper revisits the convergence of Stochastic Mirror Descent (SMD) in the contemporary nonconvex optimization setting. Existing results for batch-free nonconvex SMD restrict the choice of the distance generating function (DGF) to be differentiable with Lipschitz continuous gradients, thereby excluding important setups such as Shannon entropy. In this work, we present a new convergence analysis of nonconvex SMD supporting general DGF, that overcomes the above limitations and relies solely on the standard assumptions. Moreover, our convergence is established with respect to the Bregman Forward-Backward envelope, which is a stronger measure than the commonly used squared norm of gradient mapping. We further extend our results to guarantee high probability convergence under sub-Gaussian noise and global convergence under the generalized Bregman Proximal Polyak-Łojasiewicz condition. Additionally, we illustrate the advantages of our improved SMD theory in various nonconvex machine learning tasks by harnessing nonsmooth DGFs. Notably, in the context of nonconvex differentially private (DP) learning, our theory yields a simple algorithm with a (nearly) dimension-independent utility bound. For the problem of training linear neural networks, we develop provably convergent stochastic algorithms.

1 INTRODUCTION

We consider stochastic composite optimization

$$\min_{x \in \mathcal{X}} \mathbb{E}\left[f(x,\xi)\right] + r(x), \tag{1}$$

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

where $F(x) := \mathbb{E}\left[f(x,\xi)\right]$ is differentiable and (possibly) nonconvex, $r(\cdot)$ is convex, proper and lower-semi-continuous, $\mathcal{X}$ is a closed convex subset of $\mathbb{R}^d$. The random variable (r.v.) ξ is distributed according to an unknown distribution P. We denote $\Phi := F + r$ and let $\Phi^* := \inf_{x\in\mathcal{X}} \Phi(x) > -\infty$.

A popular algorithm for solving (1) is Stochastic Mirror Descent (SMD), which has an update rule

$$x_{t+1} = \arg\min_{x\in\mathcal{X}} \eta_t(\langle \nabla f(x_t,\xi_t), x\rangle + r(x)) + D_\omega(x, x_t), \tag{2}$$

where $D_\omega(x,y)$ is the Bregman divergence between points $x, y \in \mathcal{X}$ induced by a distance generating function (DGF) $\omega(\cdot)$; see Section 2 for the definitions. When $r(\cdot) = 0$, $\mathcal{X} = \mathbb{R}^d$ and $\omega(x) = \frac{1}{2}\|x\|_2^2$, we have $D_\omega(x,y) = \frac{1}{2}\|x-y\|_2^2$, and SMD reduces to the standard Stochastic Gradient Descent (SGD). However, it is often useful to consider more general (non-Euclidean) DGFs.

SMD with general DGF was originally proposed in the pioneering work of Nemirovskij and Yudin [1979, 1983], and later found many fruitful applications [Ben-Tal et al., 2001, Shalev-Shwartz, 2012, Arora et al., 2012] leveraging nonsmooth instances of DGFs. In the last few decades, SMD has been extensively analyzed in the convex setting under various assumption, e.g., [Beck and Teboulle, 2003, Lan, 2012, Allen-Zhu and Orecchia, 2014, Birnbaum et al., 2011], including relative smoothness [Lu et al., 2018, Bauschke et al., 2017, Dragomir et al., 2021, Hanzely et al., 2021] and stochastic optimization [Lu, 2019, Nazin et al., 2019, Zhou et al., 2020b, Hanzely and Richtárik, 2021, Vural et al., 2022, Liu et al., 2023, Nguyen et al., 2023]. However, despite the vast theoretical progress, convergence analysis of nonconvex SMD with general DGF still remains elusive.

1.1 Related Work

We now discuss the related work in the nonconvex stochastic setting. In the unconstrained Euclidean case, Ghadimi and Lan [2013] propose the first non-asymptotic analysis of nonconvex SGD. Later, Ghadimi et al. [2016] consider the more general composite problem (1) with arbitrary convex $r(\cdot)$, $\mathcal{X}$, and propose a modified algorithm using large mini-batches. Unfortunately, the use of large mini-batch appears to be crucial in the proof proposed in [Ghadimi et al., 2016] even in Euclidean setting. Later, Davis and Drusvyatskiy [2019] address this issue by proposing a different analysis for Prox-SGD (method (2) with $\omega(x) = \frac{1}{2}\|x\|_2^2$). Their elegant proof, using the notion of the so-called Moreau envelope, allows them to remove the large batch requirement in the Euclidean setting. However, their analysis crucially relies on the use of Euclidean geometry and appears difficult to extend to the more general nonsmooth DGFs of interest. In particular, the subsequent works [Zhang and He, 2018, Davis et al., 2018] do consider more general DGF and derive convergence rates for (2). However, both works assume a smooth DGF to justify their proposed convergence measures, see our Section 4.2 for a more detailed comparison. Another line of work uses momentum or variance reduced estimators, e.g., [Zhang, 2021, Huang et al., 2022, Fatkhullin et al., 2023c, Ding et al., 2023], but agian their analysis is limited to the Euclidean geometry.

1.2 Contributions

- In this work, we develop a new convergence analysis for SMD under the general assumptions of relative smoothness and bounded variance of stochastic gradients. Importantly, unlike the prior work, our analysis naturally accommodates general nonsmooth DGFs, including the important case of Shannon entropy. Moreover, our analysis (i) works for any batch size, (ii) does not require the bounded gradients assumption, (iii) supports any closed convex set $\mathcal{X}$, and (iv) guarantees convergence on a strong stationarity measure – the Bregman Forward-Backward envelope.

- We further demonstrate the flexibility of our proof technique by extending it in two directions. First, we perform a high probability analysis under the sub-Gaussian noise improving upon the previously known rates under weaker assumptions. Next, we establish the global convergence in the function value for SMD under the generalized version of the Proximal Polyak-Lojasiewicz condition. In both cases, when specialized to the unconstrained Euclidean setup, our rates can recover the state-of-the-art bounds, up to small absolute constants.

- Finally, we demonstrate the importance of our general theory in various machine learning contexts, including differential privacy, policy optimization in reinforcement learning, and training deep linear neural networks. For each of the considered problems, our new SMD theory allows us to either improve convergence rates or design provably convergent stochastic algorithms. In all cases, we leverage nonsmooth DGFs to attain the result.

Our Techniques. The key idea of our analysis is the use of a new Lyapunov function in the form of a weighted sum of the function value $\Phi(\cdot)$ and its Bregman Moreau envelope $\Phi_{1/\rho}(\cdot)$:

$$\lambda_t := \eta_{t-1}\rho(\Phi(x_t) - \Phi^*) + \Phi_{1/\rho}(x_t) - \Phi^*.$$

We recall that the classic analysis of (large batch) SMD in [Ghadimi et al., 2016] uses the function value as a Lyapunov function, i.e., $\lambda_{t,1} := \Phi(x_t) - \Phi^*$. While this approach is very intuitive and matches with analysis in unconstrained case, it seems very difficult to generalize to more general constrained problem (1) even in the Euclidean setting. On the other hand, the analysis pioneered in [Davis and Drusvyatskiy, 2019] uses $\lambda_{t,2} := \Phi_{1/\rho}(x_t) - \Phi^*$ as a Lyapunov function, which does not seem straightforward to extend into non-Euclidean setups, unless the smoothness of DGF is additionally imposed. Our Lyapunov function contains a weighted average of the above two quantities, i.e., $\lambda_t = \eta_{t-1}\rho\lambda_{t,1} + \lambda_{t,2}$, where $\{\eta_t\}_{t\geq 0}$ is the stepsize sequence (with $\eta_{-1} = \eta_0$), $\rho > 0$. This modified Lyapunov function allows to better utilize (relative) smoothness of $F(\cdot)$ in the analysis. Namely, both upper and lower bound inequalities in Assumption 3.1 will be used in the proof.

2 PRELIMINARIES

We fix an arbitrary norm $\|\cdot\|$ defined on $\mathcal{X} \subset \mathbb{R}^d$, and denote by $\|\cdot\|_* := \sup_{z:\|z\|\leq 1}\langle \cdot, z\rangle$ its dual. The Euclidean norm is denoted by $\|\cdot\|_2$. We denote by $\delta_{\mathcal{X}}$ the indicator function of a convex set $\mathcal{X}$, i.e., $\delta_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$ and $+\infty$ otherwise. For a closed proper function $\Phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ with $\operatorname{dom}\Phi := \{x \in \mathbb{R}^d \mid \Phi(x) < +\infty\}$, the Fréchet subdifferential at a point $x \in \mathbb{R}^d$ is denoted by $\partial\Phi(x)$ and is defined as a set of points $g \in \mathbb{R}^d$ such that $\Phi(y) \geq \Phi(x) + \langle g, y - x\rangle + o(\|y - x\|), \forall y \in \mathbb{R}^d$ if $x \in \operatorname{dom}\Phi$. We set $\partial\Phi(x) = \emptyset$ if $x \notin \operatorname{dom}\Phi$ [Davis and Grimmer, 2019].[1] We denote by $\operatorname{cl}(\mathcal{X})$ and $\operatorname{ri}(\mathcal{X})$ the closure and the relative interior of $\mathcal{X}$ respectively.

Let $\mathcal{S} \subset \mathbb{R}^d$ be an open set and $\omega : \operatorname{cl}(\mathcal{S}) \to \mathbb{R}$ be continuously differentiable on $\mathcal{S}$. Then we say that $\omega(\cdot)$ is a distance generating function (DGF) (with zone $\mathcal{S}$) if it is 1-strongly convex w.r.t. $\|\cdot\|$ on $\operatorname{cl}(\mathcal{S})$. We assume throughout that $\mathcal{S}$ is chosen such that $\operatorname{ri}(\mathcal{X}) \subset \mathcal{S}$ [Chen

[1] When $\Phi = \delta_{\mathcal{X}}$, the function is convex and $\partial\delta_{\mathcal{X}}(x)$ coincides with the usual subdifferential in the convex analysis.

and Teboulle, 1993].[2] For simplicity, we let $\operatorname{dom} r = \mathbb{R}^d$. The *Bregman divergence* [Bregman, 1967] induced by $\omega(\cdot)$ is

$$D_\omega(x, y) := \omega(x) - \omega(y) - \langle \nabla \omega(y), x - y\rangle \quad \text{for } x, y \in \mathcal{S}.$$

We denote by $D_\omega^{\text{sym}}(x, y) := D_\omega(x, y) + D_\omega(y, x)$ a symmetrized Bregman divergence.

For any $\Phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and a real $\rho > 0$, the Bregman Moreau envelope and the proximal operator are defined respectively by

$$\Phi_{1/\rho}(x) := \min_{y \in \mathcal{X}} \left[\Phi(y) + \rho D_\omega(y, x)\right],$$

$$\operatorname{prox}_{\Phi/\rho}(x) := \arg\min_{y \in \mathcal{X}} \left[\Phi(y) + \rho D_\omega(y, x)\right].$$

A point $x \in \mathcal{X} \cap \mathcal{S}$ is called a first-order stationary point (FOSP) of (1) if $0 \in \partial(\Phi + \delta_\mathcal{X})(x)$ for $\Phi := F + r$.

2.1 FOSP Measures

We define three different measures of first-order stationarity for a candidate solution $x \in \mathcal{X} \cap \mathcal{S}$.

(i) *Bregman Proximal Mapping (BPM)*

$$\Delta_\rho(x) := \rho^2 D_\omega^{\text{sym}}(\hat{x}, x), \quad \hat{x} := \operatorname{prox}_{\Phi/\rho}(x).$$

(ii) *Bregman Gradient Mapping (BGM)*

$$\Delta_\rho^+(x) := \rho^2 D_\omega^{\text{sym}}(x^+, x),$$

$$x^+ := \arg\min_{y \in \mathcal{X}} \langle \nabla F(x), y\rangle + r(y) + \rho D_\omega(y, x).$$

(iii) *Bregman Forward-Backward Envelope (BFBE)*

$$\mathcal{D}_\rho(x) := -2\rho \min_{y \in \mathcal{X}} Q_\rho(x, y),$$

$$Q_\rho(x, y) := \langle \nabla F(x), y - x\rangle + \rho D_\omega(y, x) + r(y) - r(x).$$

In unconstrained Euclidean case, i.e., $\mathcal{X} = \mathbb{R}^d$, $r(\cdot) = 0$ and $\omega(x) = \frac{1}{2}\|x\|_2^2$, we have $\Delta_\rho^+(x) = \mathcal{D}_\rho(x) = \|\nabla F(x)\|_2^2$, which is the standard stationarity measure in non-convex optimization.[3] Note that all three quantities presented above are measures of FOSP in the sense that if one of them $\Delta_\rho(x)$, $\Delta_\rho^+(x)$ or $\mathcal{D}_\rho(x)$ is zero for some $x \in \mathcal{X} \cap \mathcal{S}$, then $0 \in \partial(\Phi + \delta_\mathcal{X})(x)$. However, it is more practical to understand what happens if one of them is only ε-close to zero. In Section 4.1 we establish the connections between these quantities for any $x \in \mathcal{X} \cap \mathcal{S}$, and find that BFBE, $\mathcal{D}_\rho(x)$, is the strongest among the three. We should mention that the use of BFBE is not new for the analysis of optimization methods. In the Euclidean case, BFBE was initially proposed in [Patrinos and Bemporad, 2013], and its properties were later analyzed in [Stella et al., 2017, Liu and Pong, 2017]. Later, Ahookhosh et al. [2021] consider BFBE in general non-Euclidean setting. However, to our knowledge it was not considered in the context of stochastic even in the Euclidean setup.

[2] Following Chen and Teboulle [1993], we can verify that, in the Euclidean setup ($\omega(x) = \frac{1}{2}\|x\|_2^2$), one can set $\mathcal{S} = \mathbb{R}^d$; in the simplex setup ($\omega(x) = \sum_{i=1}^d x^{(i)} \log x^{(i)}$), the choice $\mathcal{S} = \left\{x \in \mathbb{R}^d \mid x^{(i)} > 0 \text{ for all } i \in [d]\right\}$ is suitable.

[3] This is, however, not true for BPM, which reduces to the gradient norm of a surrogate loss, i.e., $\Delta_\rho(x) = \|\nabla F_{1/\rho}(x)\|_2^2$ for $\rho > \ell$, where ℓ is a smoothness constant of $F(\cdot)$.

3 ASSUMPTIONS

Throughout the paper we make the following basic assumptions on $F(\cdot)$ and the stochastic gradients.

Assumption 3.1 (Relative smoothness [Bauschke et al., 2017, Lu et al., 2018])**.** A differentiable function $F : \mathcal{X} \cap \mathcal{S} \to \mathbb{R}$ is said to be *ℓ-relatively smooth* on $\mathcal{X} \cap \mathcal{S}$ with respect to (w.r.t.) $\omega(\cdot)$ if for all $x, y \in \mathcal{X} \cap \mathcal{S}$

$$-\ell D_\omega(x, y) \le F(x) - F(y) - \langle \nabla F(y), x - y\rangle \le \ell D_\omega(x, y).$$

We denote such class of functions as (ℓ, ω)-smooth.

It is known that smoothness w.r.t. $\|\cdot\|$, i.e., $\|\nabla F(x) - \nabla F(y)\|_* \le \ell\|x - y\|$ for all $x, y \in \mathcal{X} \cap \mathcal{S}$, implies Assumption 3.1 [Nesterov, 2018].

Assumption 3.2. We have access to a stochastic oracle that outputs a random vector $\nabla f(x, \xi)$ for any given $x \in \mathcal{X}$, such that $\mathbb{E}\left[\nabla f(x, \xi)\right] = \nabla F(x)$,

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

where the expectation is taken w.r.t. $\xi \sim P$.

4 MAIN RESULTS

4.1 Connections between FOSP Measures

We start by establishing the connections between introduced convergence measures. It turns out that BPM and BGM are essentially equivalent, i.e., differ only by a small (absolute) multiplicative constant.

Lemma 4.1 (BPM $\approx$ BGM)**.** *Let $F(\cdot)$ be (ℓ, ω)-smooth and $\sqrt{D_\omega^{sym}(x, y)}$ be a metric. Then for any $x \in \mathcal{X} \cap \mathcal{S}$, and $\rho, s > 0$ such that $\rho > \ell/s + 2\ell$, it holds*

$$\frac{\Delta_\rho(x)}{C(\ell, \rho, s)} \le \Delta_\rho^+(x) \le C(\ell, \rho, s)\Delta_\rho(x),$$

where $C(\ell,\rho,s) := \frac{(1+s)(\rho-\ell)+(1+s^{-1})\ell}{\rho-\ell-(1+s^{-1})\ell}$. *In particular, for* $s=1$, $\rho=4\ell$, *we have* $C(\ell,\rho,s)=8$, *and*

$$\frac{1}{8}\Delta_{4\ell}(x) \le \Delta_{4\ell}^{+}(x) \le 8\Delta_{4\ell}(x).$$

This result is in a similar spirit to Theorem 4.5 in [Drusvyatskiy and Paquette, 2019]. However, their proof only works in the Euclidean setting and does not readily extend to other DGFs. Our proof is different and can accommodate a possibly nonsmooth DGF.[4] Next, we examine the relation between BGM and BFBE.

Lemma 4.2 (BFBE > BGM)**.** *For any* $x \in \mathcal{X}\cap\mathcal{S}$

$$2\mathcal{D}_{\rho/2}(x) \ge \Delta_{\rho}^{+}(x).$$

There is an instance of problem (1) *with* $\ell=1$, $\mathcal{X}=[0,1]$ *and* $\arg\min_{y\in\mathcal{X}}\Phi(x)=0$ *such that for any* $\rho\in[1,2]$, $\rho_1\ge 1$ *and* $x\in(0,1]$ *it holds*

$$\frac{\mathcal{D}_{\rho}(x)}{\Delta_{\rho_1}^{+}(x)} \ge \frac{2}{|x|}.$$

The above lemma implies that BFBE is a strictly stronger convergence measure than previously considered BGM and BPM. Moreover, the difference between BFBE and BGM can be arbitrarily large even when x is close to the optimum! This effect is actually very common and happens already in the Euclidean case with classical regularizer $r(x)=\|x\|_1$. The explanation for this phenomenon is simple. Notice that BGM is defined in the primal terms, i.e., the squared distance between x and x^+, while BFBE is defined in the functional terms (the minimum value of $Q_\rho(x,y)$ over y). Therefore, BFBE unlike BGM scales with the value of $r(x)=|x|$ rather than x^2.

We conclude from Lemma 4.1 and 4.2 that $\mathcal{D}_\rho(x)$ is the strongest convergence measure among the three. In the subsequent sections we aim to establish convergence of SMD directly w.r.t. BFBE instead of using BGM or BPM.

4.2 Convergence to FOSP in Expectation

We start with our key result, which establishes convergence of SMD in expectation.

[4] Unfortunately, it is unclear if the above result holds for arbitrary $\omega(\cdot)$ that does not induce a metric. Note that, in general, DGF might not induce a metric even for popular choices of $\omega(\cdot)$. For instance, the Shannon entropy induces $\sqrt{D_\omega^{\mathrm{sym}}(x,y)}$ that does not satisfy the triangle inequality, see, e.g., Theorem 3 in [Acharyya et al., 2013] for details.

Theorem 4.3. *Let Assumptions 3.1 and 3.2 hold. Let the sequence* $\{\eta_t\}_{t\ge 0}$ *be non-increasing with* $\eta_0 \le 1/(2\ell)$, *and* $\bar{x}_T$ *be randomly chosen from the iterates* $x_0,\ldots,x_{T-1}$ *with probabilities* $p_t = \eta_t/\sum_{t=0}^{T-1}\eta_t$. *Then*

$$\mathbb{E}\left[\mathcal{D}_{3\ell}(\bar{x}_T)\right] \le \frac{3\lambda_0 + 6\ell\sigma^2\sum_{t=0}^{T-1}\eta_t^2}{\sum_{t=0}^{T-1}\eta_t}, \qquad (3)$$

where $\lambda_0 := \Phi_{1/\rho}(x_0) - \Phi^* + \Phi(x_0) - \Phi^*$. *If we set constant step-size* $\eta_t = \min\left\{\frac{1}{2\ell}, \sqrt{\frac{\lambda_0}{\sigma^2\ell T}}\right\}$, *then*

$$\mathbb{E}\left[\mathcal{D}_{3\ell}(\bar{x}_T)\right] = \mathcal{O}\left(\frac{\ell\lambda_0}{T} + \sqrt{\frac{\sigma^2\ell\lambda_0}{T}}\right).$$

Proof sketch: We start with

Step I. Deterministic descent w.r.t. BFBE. We show that for any $\rho_1 \ge \rho+\ell$

$$\Phi_{1/\rho}(x) \le \Phi(x) - \frac{1}{2\rho_1}\mathcal{D}_{\rho_1}(x).$$

This inequality corresponds to deterministic descent on $\Phi(\cdot)$ of the Bregman Proximal Point Method. It will be useful in the next step to derive a recursion on $\mathcal{D}_{\rho_1}(x)$.

Step II. One step progress on the Lyapunov function. This step is the most technical one and consists of showing a progress on a carefully chosen Lyapunov function $\lambda_t := \Phi_{1/\rho}(x_t) - \Phi^* + \eta_{t-1}\rho(\Phi(x_t) - \Phi^*)$, where $\eta_{-1}=\eta_0$, $\rho>0$:

$$\lambda_{t+1} \le \lambda_t - \frac{\eta_t\rho}{2(\rho+\ell)}\mathcal{D}_{\rho+\ell}(x_t) + \rho\eta_t\langle\psi_t, \hat{x}_t - x_t\rangle + \rho(\eta_t\langle\psi_t, x_t - x_{t+1}\rangle - (1-\eta_t\ell)D_\omega(x_{t+1},x_t)),$$

where $\psi_t := \nabla f(x_t,\xi_t) - \nabla F(x_t)$.

Step III. Dealing with stochastic terms. The goal of this step is to control the stochastic terms in the above inequality using Assumption 3.2.

It remains to telescope and set the step-sizes to derive the final result. □

When specialized to the unconstrained Euclidean setting, the result of Theorem 4.3 recovers (up to a small absolute constant) previously established convergence bounds for SGD [Ghadimi and Lan, 2013] (since in this case we have $\mathcal{D}_{3\ell}(\bar{x}_T) = \|\nabla F(\bar{x}_T)\|_2^2$), which is known to be optimal [Arjevani et al., 2023, Drori and Shamir, 2020, Yang et al., 2023]. However, already in the composite setting (when $r(\cdot)\neq 0$), our result is stronger than previously derived bounds for Prox-SGD [Davis et al., 2018] because BFBE can be much larger than BPM/BGM even in the Euclidean case as we have seen in Lemma 4.2.

In the more general non-Euclidean case, compared to Theorem 2 in [Ghadimi et al., 2016], our method does not require using large batches, and our proof works for any batch size. Moreover, [Ghadimi et al., 2016] relies on the stronger assumptions: smoothness and bounded variance in the primal norm. Furthermore, a much weaker convergence measure is used in [Ghadimi et al., 2016]: the squared norm of the difference between x_t and x_t^+.[5]

Davis et al. [2018] derive convergence of SMD w.r.t. the Bregman divergence between $\hat{x}_t$ and x_t, i.e., $D_\omega(\hat{x}_t, x_t)$. Such convergence measure is not satisfactory for two reasons. First, for a general DGF of interest, the Bregman divergence is not symmetric, and it can happen that $D_\omega(\hat{x}_t, x_t)$ vanishes, while $D_\omega(x_t, \hat{x}_t)$ does not (see, e.g., Proposition 2 in [Bauschke et al., 2017]). Second, to justify this measure the authors in [Davis et al., 2018] assume $\omega(\cdot)$ to be twice differentiable and notice that $2\rho^2 D_\omega(\hat{x}_t, x_t) \geq \left\|(\nabla^2\omega(x_t))^{-1}\nabla\Phi_{1/\rho}(x_t)\right\|_*^2$, where $\nabla\Phi_{1/\rho}(\cdot)$ is the gradient of the Moreau envelope of $\Phi(\cdot)$. However, the latter measure also does not seem to be sufficient either: even if we additionally assume the uniform smoothness of $\omega(\cdot)$, it is unclear how $\nabla\Phi_{1/\rho}(\cdot)$ is connected to the standard convergence measures such as the gradient mapping in non-Euclidean setting. In the concurrent work to [Davis et al., 2018], Zhang and He [2018] derive convergence of SMD on the BPM. They also notice that if $\omega(\cdot)$ is differentiable and smooth (i.e., $\nabla\omega(\cdot)$ is M-Lipschitz continuous) on $\mathcal{X}$, then $\mathrm{dist}^2(0, \partial(\Phi + \delta_{\mathcal{X}})(\hat{x}_t)) \leq M\Delta_\rho(x_t)$. However, we argue that such assumption is very strong since commonly used DGFs such as Shannon entropy are not smooth. Moreover, the analysis in [Zhang and He, 2018] uses bounded gradients (BG) assumption, which fails to hold even for a quadratic function if $\mathcal{X}$ is unbounded.[6]

4.3 High Probability Convergence to FOSP under Sub-Gaussian Noise

While convergence in expectation for a randomly selected point $\bar{x}_T$ is classical and widely accepted in stochastic optimization, it does not necessarily guarantee convergence for a single run of the method. In this section, we extend our Theorem 4.3 to guarantee convergence for a single run of SMD with high probability. To obtain high probability bounds, we replace our Assumption 3.2 with the following commonly used "light tail" assumption on the stochastic noise distribution.

Assumption 4.4. We have access to a stochastic oracle that outputs a random vector $\nabla f(x, \xi)$ for any given $x \in \mathcal{X}$, such that $\mathbb{E}\left[\nabla f(x, \xi)\right] = \nabla F(x)$, and

$$\|\nabla f(x, \xi) - \nabla F(x)\|_* \quad \text{is } \sigma\text{-sub-Gaussian r.v.}$$ [7]

Theorem 4.5. *Let Assumptions 3.1 and 4.4 hold. Let the sequence $\{\eta_t\}_{t\geq 0}$ be non-increasing with $\eta_0 \leq 1/(2\ell)$. Then with probability at least $1-\beta$*

$$\frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\, \mathcal{D}_{5\ell}(x_t) \leq \frac{5\widetilde{\lambda}_0 + 60\sigma^2\ell\sum_{t=0}^{T-1}\eta_t^2}{2\sum_{t=0}^{T-1}\eta_t},$$

where $\widetilde{\lambda}_0 := 3\left(\Phi(x_0) - \Phi^*\right) + 8\,\eta_0\sigma^2\log\left(1/\beta\right)$.

To our knowledge, the above theorem is the first high probability bound for nonconvex SMD without use of large batches. If we use large mini-batch,[8] then the above theorem implies $\mathcal{O}\left(\frac{1}{\varepsilon^2} + \frac{\sigma^2}{\varepsilon^2}\log\left(1/\beta\right) + \frac{\sigma^2}{\varepsilon^4}\right)$ sample complexity to ensure $\min_{0\leq t\leq T-1}\mathcal{D}_{5\ell}(x_t) \leq \varepsilon^2$. Compared to the bound derived in [Ghadimi et al., 2016], which is $\mathcal{O}\left(\frac{1}{\varepsilon^2}\log\left(1/\beta\right) + \frac{\sigma^2}{\varepsilon^4}\log\left(1/\beta\right)\right)$[9], our sample complexity is better by a factor of $\log\left(1/\beta\right)$. Moreover, our Assumptions 3.1 and 4.4 are weaker than in [Ghadimi et al., 2016]. When specialized to the Euclidean setup and setting the specific step-size sequences, our Theorem 4.5 can recover (up to an absolute constant) recently derived high probability bounds for nonconvex SGD [Liu et al., 2023]. However, unlike [Liu et al., 2023], our theorem holds for any square summable step-sizes and accommodates more general (non-Euclidean) norm in Assumption 4.4. We will demonstrate the crucial benefit of using non-Euclidean setup later in Section 5.1.

4.4 Global Convergence under Generalized Proximal PŁ condition

In this subsection, we are interested in global convergence of SMD for structured nonconvex problems. We first introduce the following generalization of Proximal Polyak-Lojasiewicz (Prox-PŁ) condition [Polyak, 1963, Lojasiewicz, 1963, Lezanski, 1963].

[5] Notice that $\rho^2\left\|x - x^+\right\|^2 \leq \Delta_\rho^+(x) \leq 2\mathcal{D}_{\rho/2}(x)$, where the first inequality holds by strong convexity of $\omega(\cdot)$, and the second is due to Lemma 4.2.

[6] Not saying about the general relatively smooth functions, for which BG can fail even on a compact domain.

[7] A random variable X is called σ-sub-Gaussian if $\mathbb{E}\left[\exp(\lambda^2X^2)\right] \leq \exp(\lambda^2\sigma^2)$ for all $\lambda \in \mathbb{R}$ with $|\lambda| \leq 1/\sigma$.

[8] Which reduces σ^2 to σ^2/B for mini-batch of size B.

[9] Discarding the samples for post-proccesing step in equation (71) therin.

Assumption 4.6 (α-Bregman Prox-PŁ)**.** There exists $\alpha \in [1,2]$ and $\mu > 0$ such that for some $\rho \geq 3\ell$ and all $x \in \mathcal{X} \cap \mathcal{S}$

$$\mathcal{D}_\rho(x) \geq 2\mu(\Phi(x) - \Phi^*)^{2/\alpha}. \tag{4}$$

The above assumption generalizes Prox-PŁ condition studied in [Karimi et al., 2016, J Reddi et al., 2016, Li and Li, 2018] in two ways. First, we have $\mathcal{D}_\rho(x)$ defined w.r.t. an arbitrary non-Euclidean DGF. Second, we consider $\alpha \in [1,2]$ instead of fixing $\alpha = 2$. We will demonstrate later in Section 5.2 that both of these generalizations are important in some nonconvex problems and the flexibility of choosing $\omega(\cdot)$ can reduce the total sample complexity. We now state the global convergence of SMD.

Theorem 4.7. *Let Assumptions 3.1, 3.2 and 4.6 hold. For any $\varepsilon > 0$, there exists a choice of step-sizes $\{\eta_t\}_{t\geq 0}$ for method (2) such that $\min_{t\leq T} \mathbb{E}\left[\Phi(x_t^+) - \Phi^*\right] \leq \varepsilon$ after*

$$T = \mathcal{O}\left(\frac{\ell\Lambda_0}{\mu}\frac{1}{\varepsilon^{\frac{2-\alpha}{\alpha}}}\log\left(\frac{\ell\Lambda_0}{\mu\varepsilon}\right) + \frac{\ell\Lambda_0\sigma^2}{\mu^2}\frac{1}{\varepsilon^{\frac{4-\alpha}{\alpha}}}\right).$$

The above result implies that after at most T iterations SMD will find a point x_t which is one Mirror Descent step away from a point that is ε-close to Φ^* in the function value. In the unconstrained Euclidean setting, the above sample complexity matches with that of SGD [Fatkhullin et al., 2022].[10] In the special case $\alpha = 2$, it implies the linear convergence rate in deterministic case and $\mathcal{O}\left(\varepsilon^{-1}\right)$ sample complexity in the stochastic case. The linear convergence and $\mathcal{O}\left(\varepsilon^{-1}\right)$ sample complexity of SMD were previously shown under relative smoothness and relative strong convexity, e.g., in [Lu et al., 2018, Hanzely and Richtárik, 2021]. Our result under Assumption 4.6 is more general since the relative strong convexity of $F(\cdot)$ implies (4) with $\alpha = 2$, see Lemma F.4. It is also known that such rates are optimal for $\alpha = 2$ in the Euclidean setting [Yue et al., 2023, Agarwal et al., 2009].

5 NEW INSIGHTS FOR MACHINE LEARNING

In this section, we dive into the context of several machine learning applications. We illustrate how each of our Theorems 4.3, 4.5 and 4.7 can be applied to specific problems; either yielding faster convergence than existing algorithms or allowing us to design provably convergent schemes. Interestingly, the presented problems are very diverse and allow us to demonstrate different aspects of our assumptions. In all presented examples, we crucially rely on the choice of nonsmooth DGFs, which was not theoretically possible to handle in the prior work on SMD.

5.1 DP Learning in ℓ_2 and ℓ_1 Settings

In differentially private (DP) stochastic nonconvex optimization, the goal is to design a *private algorithm* to minimize the population loss of type (1) over a subset of a d-dimensional space given n i.i.d. samples, $\xi^1, \dots, \xi^n$, drawn from a distribution P. Denote by $S := \left\{\xi^1, \dots, \xi^n\right\}$, the sampled dataset, and by $\nabla F(x) := \sum_{i=1}^n \nabla f(x, \xi^i)$, the gradient of the empirical loss $F(x) := \sum_{i=1}^n f(x, \xi^i)$ based on dataset S. The classical notion to quantify the privacy quality is

Definition 5.1 ((ϵ,δ)-DP [Dwork et al., 2006])**.** A randomized algorithm $\mathcal{M}$ is (ϵ,δ)-differentially private if for any pair of datasets S, S' that differ in exactly one data point and for any event $\mathcal{Y} \subseteq Range(\mathcal{M})$ in the output range of $\mathcal{M}$, we have

$$\Pr\left(\mathcal{M}(S) \in \mathcal{Y}\right) \leq e^\epsilon \Pr\left(\mathcal{M}\left(S'\right) \in \mathcal{Y}\right) + \delta,$$

where the probability is w.r.t. the randomness of $\mathcal{M}$.

There are several common techniques to ensure privacy, which include output [Wu et al., 2017, Zhang et al., 2017], objective function [Chaudhuri et al., 2011, Kifer et al., 2012, Iyengar et al., 2019] or gradient perturbations [Bassily et al., 2014, Wang et al., 2017]. Most recent works on nonconvex DP learning focus on the latter approach. The key idea of gradient perturbation is to inject an artificial Gaussian noise $b_t \sim \mathcal{N}(0, \sigma_{\mathrm{G}}^2 I_d)$ into the evaluated gradient. The parameter σ_{G}^2 should be carefully chosen to ensure privacy, which can be guaranteed by the moments accountant

Lemma 5.2 (Theorem 1 in [Abadi et al., 2016])**.** *Assume that $\|\nabla F(x)\|_2 \leq G$ for all $x \in \mathcal{X}$. There exist constants $c_1, c_2 > 0$ so that given the number of iterations $T \geq 0$, for any $\epsilon \leq c_1 T$, the gradient method using $\nabla F(x_t) + b_t$, $b_t \sim \mathcal{N}(0, \sigma_G^2 I_d)$ as the gradient estimator is (ϵ,δ)-DP for any $\delta > 0$ if $\sigma_G^2 \geq c_2 \frac{G^2 T \log(1/\delta)}{n^2\epsilon^2}$.*

ℓ_2 Setting. For instance, the DP-Prox-GD iterates

$$x_{t+1} = \mathrm{prox}_{\eta_t r}(x_t - \eta_t(\nabla F(x_t) + b_t)),\ b_t \sim \mathcal{N}(0, \sigma_{\mathrm{G}}^2 I_d),$$

where $\mathrm{prox}_{\eta_t r}(x) := \arg\min_{y\in\mathcal{X}}\left(r(y) + \frac{1}{2\eta_t}\|y - x\|_2^2\right)$. Our Theorem 4.5 immediately implies the high probability utility bound for DP-Prox-GD:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{D}_{5\ell}(x_t)\right] = \mathcal{O}\left(\frac{\sqrt{d\log\left(1/\delta\right)\log(1/\beta)}}{n\epsilon}\right), \tag{5}$$

[10] We use a different step-size sequence $\{\eta_t\}_{t\geq 0}$ compared to $\eta_t = 1/t^\zeta$, $\zeta > 0$ used in [Fatkhullin et al., 2022], see Appendix D. This allows us to derive noise adaptive rates, i.e., if $\sigma = 0$, then we recover the iteration complexity of (deterministic) mirror descent.

where $\beta \in (0,1)$ is the failure probability, see Corollary E.1 for more details and the dependence on omitted constants. To our knowledge, nonconvex utility bound of DP-Prox-GD was previously studied only in the unconstrained setting ($r(\cdot) = 0$, $\mathcal{X} = \mathbb{R}^d$), e.g., [Wang et al., 2017, 2019, Zhou et al., 2020a] or in expectation, e.g., [Wang and Xu, 2019]. Our bound (5) generalizes these works to non-trivial $\mathcal{X}$ and $r(\cdot)$.

ℓ_1 **Setting.** One issue with the above utility bound is the polynomial dimension dependence. In certain cases, this dependence can be significantly improved, e.g., when the optimization is defined on a unit simplex $\mathcal{X} = \left\{x \in \mathbb{R}^d \,|\, \sum_{i=1}^d x^{(i)} \leq 1,\, x^{(i)} \geq 0\right\}$. Notably, it makes a big difference which norm we use to measure the variance of b_t, e.g., $\mathbb{E}\left\|b_t\right\|_2^2 = d\,\sigma_G^2$ and $\mathbb{E}\left\|b_t\right\|_\infty^2 \leq 2\log(d)\,\sigma_G^2$. Therefore, using $\left\|\cdot\right\|_\infty$ norm is more favorable. Motivated by this difference, we consider the differentially private mirror descent (DP-MD):

$$x_{t+1} = \arg\min_{y\in\mathcal{X}} \eta_t(\langle \nabla F(x_t) + b_t, y\rangle + r(y)) + D_\omega(y, x_t),$$

where $b_t \sim \mathcal{N}(0, \sigma_{\rm G}^2 I_d)$ and $\omega(x) = \sum_{i=1}^d x^{(i)} \log x^{(i)}$.[11] Using our high probability guarantee Theorem 4.5, we can derive

Corollary 5.3. *Let $F(\cdot)$ be (ℓ,ω)-smooth for $\omega(\cdot)$, $\mathcal{X}$ defined above, and $\left\|\nabla F(x)\right\|_2 \leq G$ for all $x \in \mathcal{X}$. Set $\eta_t = \frac{1}{2\ell}$, $T = \frac{n\epsilon\sqrt{\ell}}{G\sqrt{\log(d)\log(1/\delta)\log(1/\beta)}}$, $\lambda_0 := \Phi(x_0) - \Phi^*$. Then DP-MD is (ϵ,δ)-DP and with probability $1-\beta$ satisfies* [12]

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{D}_{5\ell}(x_t) = \mathcal{O}\left(\frac{G\sqrt{\ell\lambda_0 \log(d)\log\left(1/\delta\right)\log\left(1/\beta\right)}}{n\epsilon}\right),$$

The above result establishes a (nearly) dimension independent utility bound for DP-MD, and improves the one of DP-Prox-GD in (5) by a factor of $\sqrt{d/\log(d)}$. Several previous works in DP learning literature have shown the improved dimension dependence in ℓ_1 setting, e.g., [Asi et al., 2021, Gopi et al., 2023, Bassily et al., 2021b,a, Wang and Xu, 2019]. However, Asi et al. [2021], Gopi et al. [2023], Bassily et al. [2021b] assume convex $F(\cdot)$, and, therefore, are not directly comparable with our result. Bassily et al. [2021a], and Wang and Xu [2019] obtain nonconvex utility bounds in expectation, however, their techniques are different. Both above mentioned works rely on the linear minimization oracle and derive convergence on the Frank-Wolfe (FW) gap.[13] Moreover, Bassily et al. [2021a] use a complicated double loop algorithm based on momentum-based variance reduction technique.

5.2 Policy Optimization in Reinforcement Learning (RL)

Consider a discounted Markov decision process (DMDP) $M = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma, p\}$. Here $\mathcal{S}$ is a state space with cardinality $|\mathcal{S}|$; $\mathcal{A}$ is an action space with cardinality $|\mathcal{A}|$; $\mathcal{P}$ is a transition model, where $\mathcal{P}(s'|s,a)$ is the transition probability to state s' from a given state s when action a is applied; $R : \mathcal{S}\times\mathcal{A} \to [0,1]$ is a reward function for a state-action pair (s,a); $\gamma \in [0,1)$ is the discount factor; and p is the initial state distribution. Being at state $s_h \in \mathcal{S}$ an RL agent takes an action $a_h \in \mathcal{A}$ and transitions to another state s_{h+1} according to $\mathcal{P}$ and receives an immediate reward $r_h \sim R(s_h, a_h)$. A (stationary) policy π specifies a (randomized) decision rule depending only on the current state s_h, i.e., for each $s \in \mathcal{S}$, $\pi_s \in \Delta(\mathcal{A})$ determines the next action $a \sim \pi_s$, where $\Delta(\mathcal{A}) := \left\{\pi_s \in \mathbb{R}^{|\mathcal{A}|} \,|\, \sum_{s\in\mathcal{S}} \pi_{sa} = 1,\, \pi_{sa} \geq 0 \text{ for all } a \in \mathcal{A}\right\}$ denotes the probability simplex supported on $\mathcal{A}$. The goal of RL agent is to maximize

$$V_p^+(\pi) := \mathbb{E}\left[\sum_{h=0}^{\infty}\gamma^h r_h\right], \quad \pi \in \mathcal{X} := \Delta(\mathcal{A})^{|\mathcal{S}|}, \tag{6}$$

where expectation is w.r.t. the initial state distribution $s_0 \sim p$, the transition model $\mathcal{P}$ and the policy π. We define $V_p(\pi) := -V_p^+(\pi)$ and adopt the minimization formulation of DMDP, i.e., $\min_{\pi\in\mathcal{X}} V_p(\pi)$.

It is known that $V_p(\pi)$ is smooth, but nonconvex in π. Moreover, a property similar to Proximal PŁ (Assumption 4.6) was recently established for (6) [Agarwal et al., 2021, Xiao, 2022]. That is we have for any $\pi, \pi' \in \mathcal{X}$:

$$\begin{aligned} \left\|\nabla V_p(\pi) - \nabla V_p(\pi')\right\|_{2,2} &\leq L_F \left\|\pi - \pi'\right\|_{2,2}, \\ V_p(\pi) - V_p^\star &\leq C \max_{\pi'\in\mathcal{X}} \langle \nabla V_\mu(\pi), \pi - \pi'\rangle, \end{aligned} \tag{7}$$

where $L_F := \frac{2\gamma|\mathcal{A}|}{(1-\gamma)^3}$, $C := \frac{1}{1-\gamma}\left\|\frac{d_p(\pi^\star)}{\mu}\right\|_\infty$, $\left\|\cdot\right\|_{2,2}$ denotes the Frobenius norm (Lemma 4 and 54 in [Agarwal et al., 2021]).

Therefore, this problem serves well to demonstrate the application of our theory to show convergence of policy gradient (PG) methods. PG methods is the promising class of algorithms that generate a sequence of policies π_t by evaluating the gradients $\nabla V_\mu(\pi_t)$ (or their stochastic estimates $\widehat{\nabla} V_\mu(\pi_t)$), where $\mu \in \Delta(\mathcal{A})$ is some distribution (not necessarily equal to p). One of the most basic variants is the

Projected Stochastic Policy Gradient:

$$\text{P-SPG:} \quad \pi_{t+1} = \mathrm{proj}_{\mathcal{X}}\left(\pi_t - \eta_t \widehat{\nabla} V_\mu(\pi_t)\right),$$

[11] It is known that such $\omega(\cdot)$ is 1-strongly convex w.r.t. $\left\|\cdot\right\|_1$ on a unit simplex [Beck and Teboulle, 2003].

[12] The result can be easily extended to the case when only stochastic gradients $\nabla f(x_t, \xi_t^i)$ are used instead of $\nabla F(x_t)$.

[13] At least when restricted to Euclidean setting, FW gap is a weaker convergence measure than BFBE, see Lemma 4.2 and F.5.

where $\mathrm{proj}_{\mathcal{X}}(\cdot)$ denotes the Euclidean projection onto $\mathcal{X}$. Given that the variance of stochastic gradients $\widehat{\nabla} V_\mu(\pi_t)$ is bounded in the Euclidean norm by σ_F^2,[14] our Theorems 4.3 and 4.7 imply the following

Corollary 5.4. *For any $\varepsilon > 0$, P-SPG guarantees:*

(i) $\min_{0\le t\le T-1} \mathbb{E}\left[\mathcal{D}_{3L_F}(\pi_t)\right] \le \varepsilon^2$ *after*

$$T = \mathcal{O}\left(\frac{|\mathcal{A}|}{(1-\gamma)^3\varepsilon^2} + \frac{\sigma_F^2|\mathcal{A}|}{(1-\gamma)^3\varepsilon^4}\right),$$

(ii) $\min_{t\le T} \mathbb{E}\left[V_p(\pi_t^+) - V_p^*\right] \le \varepsilon$ *after*

$$T = \widetilde{\mathcal{O}}\left(\frac{|\mathcal{A}||\mathcal{S}|}{(1-\gamma)^5\varepsilon} + \frac{\sigma_F^2|\mathcal{A}|^2|\mathcal{S}|^2}{(1-\gamma)^7\varepsilon^3}\right).$$

Convergence of P-SPG was studied (in deterministic case) in [Agarwal et al., 2021] using the notion of gradient mapping. Recently, an improved analysis was provided in [Xiao, 2022] with iteration complexity $T = \mathcal{O}\left(\frac{|\mathcal{A}||\mathcal{S}|}{(1-\gamma)^5\varepsilon}\right)$ to achieve $V_p(\pi_T) - V_p^* \le \varepsilon$. If $\sigma_F = 0$, our iteration complexity in (ii) recovers the one in [Xiao, 2022], albeit with a different proof.

Improving dependence on $|\mathcal{A}|$. Notice that the above sample complexity bounds depend on the cardinality of the action space, which can be large in practice. The key reason for this is that the analysis of P-SPG (Prox-SGD) requires to measure the smoothness constant L_F of $V_p(\pi)$ in the Euclidean (Frobenius) norm, which inevitably depends on the cardinality of the action space $|\mathcal{A}|$. Let us instead consider $(2,1)$-matrix norm $\|\cdot\|_{2,1}$, i.e., $\|\pi\|_{2,1}^2 = \sum_{s\in\mathcal{S}}\left(\sum_{a\in\mathcal{A}}|\pi_{sa}|\right)^2$.[15] Now, we show that the dependence on $|\mathcal{A}|$ in the smoothness constant can be completely removed if $(2,1)$-norm is used.

Proposition 5.5. *For any $\pi, \pi' \in \mathcal{X}$, it holds that*

$$\|\nabla V_p(\pi) - \nabla V_p(\pi')\|_{2,\infty} \le \frac{2\gamma}{(1-\gamma)^3}\|\pi - \pi'\|_{2,1}.$$

Consider **Stochastic Mirror Policy Gradient** (SMPG), that is SMD with the matrix form of Shannon entropy $\omega(\pi) := \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\pi_{sa}\log\pi_{sa}$.[16] The stochastic gradients in SMD are replaced by $\widehat{\nabla} V_\mu(\pi_t) := (\widehat{\nabla}_1 V_\mu(\pi_t), \ldots, \widehat{\nabla}_{|S|} V_\mu(\pi_t))$. Define $E_t := (E_t^1, \ldots, E_t^{|S|})$, then SMPG can be written in a closed from. For all $s \in \mathcal{S}$

$$\pi_{t+1} = \pi_t \odot E_t, \quad E_t^s := \frac{\exp\left(-\eta_t \widehat{\nabla}_s V_\mu(\pi_t)\right)}{\sum_{a\in\mathcal{A}}\exp\left(-\eta_t\widehat{\nabla}_s V_\mu(\pi_t)\right)},$$

where $\odot$ denotes an element-wise multiplication of matrices and $\exp(\cdot)$ is an element-wise exponential. The sample complexity can be derived from Theorem 4.7 using Proposition 5.5 under the bounded variance assumption (in dual norm $\|\cdot\|_{2,\infty}$).

Corollary 5.6. *For any $\varepsilon > 0$, SMPG guarantees that* $\min_{0\le t\le T-1}\mathbb{E}\left[\mathcal{D}_\rho(\pi_t)\right] \le \varepsilon^2$ *with $\rho := 6\gamma(1-\gamma)^{-3}$ after*

$$T = \mathcal{O}\left(\frac{1}{(1-\gamma)^3\varepsilon^2} + \frac{\sigma_{2,\infty}^2}{(1-\gamma)^3\varepsilon^4}\right).$$

Notice that compared to the bound for P-SPG the above sample complexity is better at least by a factor of $|\mathcal{A}|$. Moreover, $\sigma_{2,\infty}$ can be much smaller than σ_F.

It should be noted, however, that $\mathcal{D}_\rho(\pi_t)$ in Corollaries 5.4 and 5.6 are induced by different $\omega(\cdot)$ and thus induce different FOSP measures. Also it remains unclear how to establish a global convergence of SMPG in the function value. The technical difficulty arises because the condition (7) might not imply Assumption 4.6 under non-smooth DGF.

Remark 5.7. While this example serves well to illustrate the application of our general theory and potential advantages of SMPG compared to P-SPG, it does not mean that SMPG is the most suitable algorithm for solving (6). In fact, there are other specialized algorithms in RL literature, which have better theoretical sample complexities than shown above. For example, Natural Policy Gradient (NPG) (also known as exponentiated Q-descent or Policy Mirror Descent) [Kakade, 2001, Agarwal et al., 2021, Lan, 2023, Xiao, 2022, Zhan et al., 2023, Khodadadian et al., 2021] achieves faster convergence in terms of ε. However, a notable difference of SMPG compared to NPG is that the latter uses a Q-function instead of the policy gradient $\widehat{\nabla} V_\mu(\pi)$, see the derivation of NPG in Section 4 in [Xiao, 2022]. Another popular approach to problem (6) is the use of soft-max policy parametrization instead of directly solving the problem over $\mathcal{X}$. In this direction, different variants of PG method were developed and analyzed, see, e.g., [Zhang et al., 2020b, 2021, Barakat et al., 2023].

Remark 5.8. The special cases and variants of Prox-PŁ condition were previously used to derive global convergence of PG methods [Daskalakis et al., 2020, Kumar et al., 2023] including continuous state action-spaces in RL [Ding et al., 2022, Fatkhullin et al., 2023a] and classical control tasks [Fazel et al., 2018, Fatkhullin and Polyak, 2021, Zhao et al., 2022, Wu et al., 2023]. An alternative approach to global convergence of P-SPG, based on hidden convexity of (6), was recently studied in [Fatkhullin et al., 2023b].

[14] The variance of $\widehat{\nabla} V_\mu(\cdot)$ can be bounded under reasonable assumptions or using appropriate exploration strategies, e.g., ϵ-greedy or Boltzmann, see [Daskalakis et al., 2020, Cesa-Bianchi et al., 2017, Xiao, 2022, Johnson et al., 2023].

[15] Its dual satisfies $\|\pi\|_{2,\infty}^2 = \sum_{s\in\mathcal{S}}(\max_{a\in\mathcal{A}}|\pi_{sa}|)^2$.

[16] It is 1-strongly convex w.r.t. $\|\cdot\|_{2,1}$ norm.

5.3 Training Autoencoder Model using SMD

In this section, we showcase how we can harness general Bregman divergence in SMD to address modern machine learning problems involving linear neural networks, where the objectives go beyond the smooth regime considered in the existing theoretical analysis [Kawaguchi, 2016].

More specifically, assume that the $F(\cdot)$ is twice differentiable and its Hessian is bounded by the polynomial of $\|x\|_2$, i.e., there exist $r, L, L_r \geq 0$ such that

$$\left\|\nabla^2 F(x)\right\|_{\mathrm{op}} \leq L + L_r \|x\|_2^r \qquad \text{for all } x \in \mathbb{R}^d.^{17} \quad (8)$$

The following result (initially appeared in [Lu et al., 2018, Lu, 2019]) shows that for any $r \geq 0$, the above condition implies relative smoothness (Assumption 3.1).

Proposition 5.9 (Proposition 2.1. in [Lu et al., 2018])**.** *Suppose $F(\cdot)$ is twice differentiable and satisfies (8). Then $F(\cdot)$ is ℓ-smooth relative to $\omega(x) = \frac{1}{r+2}\|x\|_2^{r+2} + \frac{1}{2}\|x\|_2^2$ with $\ell := \max\{L, L_r\}$.*

To design a provably convergent scheme for such problems, it remains to solve SMD subproblem with DGF specified in the above proposition. Luckily, this is possible

$$\begin{aligned} c_t &= (1 + \|x_t\|_2^r)\, x_t - \eta_t \nabla f(x_t, \xi_t), & (9)\\ x_{t+1} &= (1 + \theta_*^r)^{-1}\, c_t, & (10) \end{aligned}$$

where $\theta_* \geq 0$ is the unique solution to $\theta^{r+1} + \theta = \|c_t\|_2$.

Convergence to a FOSP of the above method follows immediately from our Theorem 4.3, see Appendix E.3 for more details. For $r = 1, 2$, the solution θ_* can be found in a closed form, while for larger values of r it can be solved using a bisection method.

To illustrate the empirical performance of the above scheme, we consider two layer autoencoder problem

$$\min_{\substack{W_1 \in \mathbb{R}^{d_e \times d_f} \\ W_2 \in \mathbb{R}^{d_f \times d_e}}} \left[F(W) := \frac{1}{n} \sum_{i=1}^{n} \|W_2 W_1 a_i - a_i\|_2^2 \right], \quad (11)$$

where $a_i \in \mathbb{R}^{d_f}$ are flattened representations of images, $W = [W_1, W_2]$ are learned parameters of the model.[18] The above problem is nonconvex and globally nonsmooth in W since its Hessian norm grows as a polynomial in the norm of W. Therefore, SGD can easily diverge if poorly initialized. However, condition

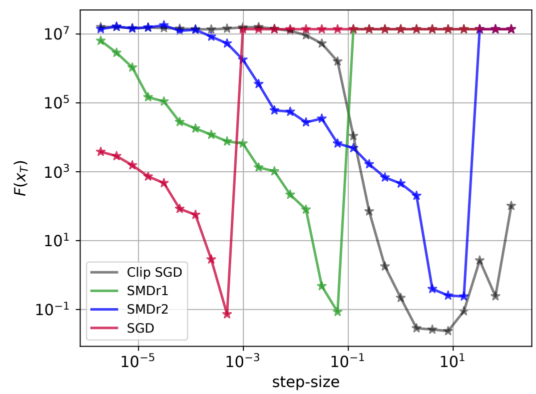

Figure 1: Sensitivity to step-size choice for SMDr1, SMDr2, SGD and Clip SGD (with clipping radius 1). The plot shows the function value $F(x_T)$ after $T = 10^4$ iterations for each step-size. The star markers correspond to the actual runs, and the lines linearly interpolate between them.

(8) can be verified for some $r_0 \geq 2$ and the scheme (9), (10) provably converges for any $r \geq r_0$.[19]

We focus on $r = 1, 2$ and call the corresponding methods SMDr1 and SMDr2. We compare these algorithms to the standard SGD, which corresponds to $r = 0$. For comparison, we also include a popular heuristic algorithm, Clip SGD, which is known to mitigate the problem of exploding gradients. We use constant step-sizes for each method. Figure 1 reports the final training loss for each step-size in the range $\{2^{-19}, 2^{-18}, \ldots, 2^7\}$.

We observe from Figure 1 that SMDr1 and SMDr2 allow for much larger step-sizes than SGD. Moreover, increasing r also increases the robustness to the step-size choice, i.e., the lower part of the curve becomes wider. At the same time, in the high accuracy regime, Clip SGD still outperforms other algorithms on this task. Note, however, that Clip SGD might not converge in general in the stochastic setting for a constant clipping parameter, see, e. g., [Koloskova et al., 2023].

Acknowledgments

This work is supported by ETH AI Center Doctoral Fellowship and Swiss National Science Foundation (SNSF) Project Funding No. 200021-207343.

References

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the*

[17] Compare to (L_0, L_1)-smoothness condition studied in [Zhang et al., 2020a].

[18] In previous notations, we have $x = \mathrm{vec}(W) \in \mathbb{R}^{2\, d_e\, d_f}$.

[19] Computation of the Hessian and use of Cauchy-Schwarz inequality implies (8) for any number of layers in (11).

2016 ACM SIGSAC conference on computer and communications security, pages 308–318, 2016.

S. Acharyya, A. Banerjee, and D. Boley. Bregman divergences and triangle inequality. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 476–484. SIAM, 2013.

A. Agarwal, M. J. Wainwright, P. Bartlett, and P. Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

M. Ahookhosh, A. Themelis, and P. Patrinos. A bregman forward-backward linesearch algorithm for nonconvex composite optimization: superlinear convergence to nonisolated local minima. *SIAM Journal on Optimization*, 31(1):653–685, 2021.

Z. Allen-Zhu and L. Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023.

S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

H. Asi, V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: Optimal rates in l1 geometry. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 393–403, 2021.

K. Balasubramanian and S. Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pages 1–42, 2022.

A. Barakat, I. Fatkhullin, and N. He. Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 1753–1800, 2023.

R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE, 2014.

R. Bassily, C. Guzmán, and M. Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems*, 34:9317–9329, 2021a.

R. Bassily, C. Guzmán, and A. Nandi. Non-euclidean differentially private stochastic convex optimization. In *Conference on Learning Theory*, pages 474–499. PMLR, 2021b.

H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

A. Ben-Tal, T. Margalit, and A. Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12(1):79–108, 2001.

B. Birnbaum, N. R. Devanur, and L. Xiao. Distributed algorithms via gradient descent for fisher markets. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 127–136, 2011.

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

N. Cesa-Bianchi, C. Gentile, G. Lugosi, and G. Neu. Boltzmann exploration done right. *Advances in Neural Information Processing Systems*, 30, 2017.

K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

C. Daskalakis, D. J. Foster, and N. Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in Neural Information Processing Systems*, 33:5527–5540, 2020.

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

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

D. Davis, D. Drusvyatskiy, and K. J. MacPhee. Stochastic model-based minimization under highorder growth. *arXiv preprint arXiv:1807.00255*, 2018.

K. Ding, J. Li, and K.-C. Toh. Nonconvex stochastic bregman proximal gradient method with application to deep learning. *arXiv preprint arXiv:2306.14522*, 2023.

Y. Ding, J. Zhang, and J. Lavaei. On the global optimum convergence of momentum-based policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 1910–1934. PMLR, 2022.

R.-A. Dragomir, A. B. Taylor, A. d'Aspremont, and J. Bolte. Optimal complexity and certification of bregman first-order methods. *Mathematical Programming*, pages 1–43, 2021.

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

D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

I. Fatkhullin and B. Polyak. Optimizing static linear feedback: Gradient method. *SIAM Journal on Control and Optimization*, 59(5):3887–3911, 2021.

I. Fatkhullin, J. Etesami, N. He, and N. Kiyavash. Sharp analysis of stochastic optimization under global Kurdyka-Łojasiewicz inequality. *Advances in Neural Information Processing Systems*, 2022.

I. Fatkhullin, A. Barakat, A. Kireeva, and N. 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, 2023a.

I. Fatkhullin, N. He, and Y. Hu. Stochastic optimization under hidden convexity. *arXiv preprint arXiv:2401.00108v1*, 2023b.

I. Fatkhullin, A. Tyurin, and P. Richtárik. Momentum provably improves error feedback! *Advances in Neural Information Processing Systems*, 2023c.

M. Faw, L. Rout, C. Caramanis, and S. Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive sgd. *arXiv preprint arXiv:2302.06570*, 2023.

M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

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

S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

S. Gopi, Y. T. Lee, D. Liu, R. Shen, and K. Tian. Private convex optimization in general norms. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 5068–5089. SIAM, 2023.

F. Hanzely and P. Richtárik. Fastest rates for stochastic mirror descent methods. *Computational Optimization and Applications*, 79:717–766, 2021.

F. Hanzely, P. Richtarik, and L. Xiao. Accelerated bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79:405–440, 2021.

F. Huang, S. Gao, and H. Huang. Bregman gradient policy optimization. In *International Conference on Learning Representations*, 2022.

F. Hübler, J. Yang, X. Li, and N. He. Parameter-agnostic optimization under relaxed smoothness. *arXiv preprint arXiv:2311.03252*, 2023.

R. Iyengar, J. P. Near, D. Song, O. Thakkar, A. Thakurta, and L. Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316. IEEE, 2019.

S. J Reddi, S. Sra, B. Poczos, and A. J. Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. *Advances in Neural Information Processing Systems*, 29, 2016.

E. Johnson, C. Pike-Burke, and P. Rebeschini. Optimal convergence rate for exact policy mirror descent in discounted markov decision processes. *arXiv preprint arXiv:2302.11381*, 2023.

S. M. Kakade. A natural policy gradient. *Advances in Neural Information Processing Systems*, 14, 2001.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. *arXiv preprint arXiv:1608.04636v4*, 2016.

K. Kawaguchi. Deep learning without poor local minima. *Advances in Neural Information Processing Systems*, 29, 2016.

S. Khodadadian, P. R. Jhunjhunwala, S. M. Varma, and S. T. Maguluri. On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3794–3799. IEEE, 2021.

D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

A. Koloskova, H. Hendrikx, and S. U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

N. Kumar, I. Usmanova, K. Y. Levy, and S. Mannor. Towards faster global convergence of robust policy gradient methods. In *Sixteenth European Workshop on Reinforcement Learning*, 2023.

G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2): 365–397, 2012.

G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106, 2023.

T. Lezanski. Über das minimumproblem für funktionale in banachschen räumen. *Mathematische Annalen*, 152:271–274, 1963.

Z. Li and J. Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 2018.

T. Liu and T. K. Pong. Further properties of the forward–backward envelope with applications to difference-of-convex programming. *Computational Optimization and Applications*, 67:489–520, 2017.

Z. Liu, T. D. Nguyen, T. H. Nguyen, A. Ene, and H. Nguyen. High probability convergence of stochastic gradient methods. In *International Conference on Machine Learning*, pages 21884–21914. PMLR, 2023.

S. Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

H. Lu. "relative continuity" for non-lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent. *INFORMS Journal on Optimization*, 1(4):288–303, 2019.

H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

A. V. Nazin, A. S. Nemirovsky, A. B. Tsybakov, and A. B. Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80:1607–1627, 2019.

A. Nemirovskij and D. Yudin. Efficient methods of solving convex programming problems of high dimensionality. *Ekonomika i matem. method (in Russian)*, XV(1), 1979.

A. Nemirovskij and D. Yudin. Problem complexity and method efficiency in optimization. *SIAM Review*, 1983.

Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

T. D. Nguyen, T. H. Nguyen, A. Ene, and H. Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. *Advances in Neural Information Processing Systems*, 2023.

P. Patrinos and A. Bemporad. Proximal newton methods for convex composite optimization. In *52nd IEEE Conference on Decision and Control*, pages 2358–2363. IEEE, 2013.

B. T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

L. Stella, A. Themelis, and P. Patrinos. Forward–backward quasi-newton methods for nonsmooth optimization problems. *Computational Optimization and Applications*, 67(3):443–487, 2017.

S. U. Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

R. Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2014.

N. M. Vural, L. Yu, K. Balasubramanian, S. Volgushev, and M. A. Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Conference on Learning Theory*, pages 65–102. PMLR, 2022.

D. Wang and J. Xu. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1182–1189, 2019.

D. Wang, M. Ye, and J. Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

D. Wang, C. Chen, and J. Xu. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, pages 6526–6535. PMLR, 2019.

J. Wu, A. Barakat, I. Fatkhullin, and N. He. Learning zero-sum linear quadratic games with improved sample complexity. In *62nd IEEE Conference on Decision and Control*, 2023.

X. Wu, F. Li, A. Kumar, K. Chaudhuri, S. Jha, and J. Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1307–1322, 2017.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

L. Xiao. On the convergence rates of policy gradient methods. *The Journal of Machine Learning Research*, 23(1):12887–12922, 2022.

J. Yang, X. Li, I. Fatkhullin, and N. He. Two sides of one coin: the limits of untuned sgd and the power of adaptive methods. *arXiv preprint arXiv:2305.12475*, 2023.

P. Yue, C. Fang, and Z. Lin. On the lower bound of minimizing polyak-łojasiewicz functions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2948–2968. PMLR, 2023.

W. Zhan, S. Cen, B. Huang, Y. Chen, J. D. Lee, and Y. Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091, 2023.

J. Zhang, K. Zheng, W. Mou, and L. Wang. Efficient private erm for smooth objectives. *arXiv preprint arXiv:1703.09947*, 2017.

J. Zhang, T. He, S. Sra, and A. Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020a.

J. Zhang, A. Koppel, A. S. Bedi, C. Szepesvari, and M. Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020b.

J. Zhang, C. Ni, C. Szepesvari, M. 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.

L. Zhang. Variance reduction for non-convex stochastic optimization: General analysis and new applications. Master's thesis, ETH Zurich, 2021.

S. Zhang and N. He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.

F. Zhao, X. Fu, and K. You. Global convergence of policy gradient methods for output feedback linear quadratic control. *arXiv preprint arXiv:2211.04051v1*, 2022.

Y. Zhou, X. Chen, M. Hong, Z. S. Wu, and A. Banerjee. Private stochastic non-convex optimization: Adaptive algorithms and tighter generalization bounds. *arXiv preprint arXiv:2006.13501*, 2020a.

Z. Zhou, P. Mertikopoulos, N. Bambos, S. P. Boyd, and P. W. Glynn. On the convergence of mirror descent beyond stochastic convex programming. *SIAM Journal on Optimization*, 30(1):687–716, 2020b.

Contents

A Proofs of Lemma 4.1 and 4.2: Connections Between FOSP

BPM and BGM are equivalent up to a constant factor.

The following lemma establishes that if $F(\cdot)$ is (ℓ, ω)-smooth, then the distance between $\hat{x}$ and x^+ is small.

Lemma A.1. *For any $\rho > \ell$, we have for all $x \in \mathcal{X} \cap \mathcal{S}$*

$$\rho^2 D_\omega^{sym}(\hat{x}, x^+) \leq \frac{\ell}{\rho - \ell} \left(\Delta_\rho(x) + \Delta_\rho^+(x)\right).$$

Proof. Recall that $\hat{x} := \arg\min_{y \in \mathcal{X}} F(y) + r(y) + \rho D_\omega(y, x)$, and $x^+ := \arg\min_{y \in \mathcal{X}} \langle \nabla F(x), y \rangle + r(y) + \rho D_\omega(y, x)$. By the optimality conditions for x^+ and $\hat{x}$, there exist $s^+ \in \partial(r + \delta_\mathcal{X})(x^+)$ and $\hat{s} \in \partial(r + \delta_\mathcal{X})(\hat{x})$, such that

$$0 = \nabla F(\hat{x}) + \hat{s} + \rho \nabla \omega(\hat{x}) - \rho \nabla \omega(x),$$

$$0 = \nabla F(x) + s^+ + \rho \nabla \omega(x^+) - \rho \nabla \omega(x),$$

Subtracting these equalities, we obtain $\rho \nabla \omega(\hat{x}) - \rho \nabla \omega(x^+) = s^+ - \hat{s} + \nabla F(x) - \nabla F(\hat{x})$.

By Lemma F.1-1 (three point identity) with $x = z$ and using the above identity, we have

$$\begin{aligned} \rho(D_\omega(\hat{x}, x^+) + D_\omega(x^+, \hat{x})) &= \langle \rho \nabla \omega(\hat{x}) - \rho \nabla \omega(x^+), \hat{x} - x^+ \rangle \\ &= \langle s^+ - \hat{s} + \nabla F(x) - \nabla F(\hat{x}), \hat{x} - x^+ \rangle \\ &\overset{(i)}{\leq} \langle \nabla F(x) - \nabla F(\hat{x}), \hat{x} - x^+ \rangle \end{aligned} \tag{12}$$

where in (i) we use convexity of $(r + \delta_\mathcal{X})(\cdot)$. By relative smoothness of $F(\cdot)$, we have for any $x, y, z \in \mathcal{X} \cap \mathcal{S}$

$$F(x) - F(y) - \langle \nabla F(y), x - y \rangle \leq \ell D_\omega(x, y),$$

$$F(z) - F(x) - \langle \nabla F(x), z - x \rangle \leq \ell D_\omega(z, x),$$

$$-\ell D_\omega(z, y) \leq F(z) - F(y) - \langle \nabla F(y), z - y \rangle.$$

Adding the above inequalities gives for any $x, y, z \in \mathcal{X} \cap \mathcal{S}$

$$\langle \nabla F(x) - \nabla F(y), y - z \rangle \leq \ell D_\omega(x, y) + \ell D_\omega(z, x) + \ell D_\omega(z, y).$$

Applying the above inequality with $x = x$, $y = \hat{x}$, $z = x^+$, we further bound (12)

$$\begin{aligned} \rho(D_\omega(\hat{x}, x^+) + D_\omega(x^+, \hat{x})) &\leq \ell D_\omega(x, \hat{x}) + \ell D_\omega(x^+, x) + \ell D_\omega(x^+, \hat{x}) \\ &\leq \ell D_\omega(x, \hat{x}) + \ell D_\omega(x^+, x) + \ell D_\omega(x^+, \hat{x}) + \ell D_\omega(\hat{x}, x^+). \end{aligned}$$

Therefore, for any $\rho > \ell$, we have

$$\rho^2 \left(D_\omega(\hat{x}, x^+) + D_\omega(x^+, \hat{x})\right) \leq \frac{\ell \rho^2}{\rho - \ell}(D_\omega(x, \hat{x}) + D_\omega(x^+, x)) \leq \frac{\ell}{\rho - \ell} \left(\Delta_\rho(x) + \Delta_\rho^+(x)\right).$$

□

Proof of Lemma 4.1. For any $x, y, z \in \mathcal{X} \cap \mathcal{S}$ and $s > 0$, we have

$$D_\omega^{\mathrm{sym}}(x, y) \leq \left(\sqrt{D_\omega^{\mathrm{sym}}(x, z)} + \sqrt{D_\omega^{\mathrm{sym}}(y, z)}\right)^2 \leq (1 + s)\, D_\omega^{\mathrm{sym}}(x, z) + \left(1 + s^{-1}\right) D_\omega^{\mathrm{sym}}(y, z).$$

Applying Lemma A.1 together with the above inequality, we have

$$\begin{aligned} \Delta_\rho^+(x) = \rho^2 D_\omega^{\mathrm{sym}}(x, x^+) &\leq (1 + s)\rho^2 D_\omega^{\mathrm{sym}}(x, \hat{x}) + (1 + s^{-1})\rho^2 D_\omega^{\mathrm{sym}}(x^+, \hat{x}) \\ &\leq (1 + s)\rho^2 D_\omega^{\mathrm{sym}}(x, \hat{x}) + \frac{(1 + s^{-1})\ell}{\rho - \ell} \left(\Delta_\rho(x) + \Delta_\rho^+(x)\right) \\ &\leq \left(1 + s + \frac{(1 + s^{-1})\ell}{\rho - \ell}\right) \Delta_\rho(x) + \frac{(1 + s^{-1})\ell}{\rho - \ell} \Delta_\rho^+(x). \end{aligned}$$

Rearranging, we obtain the upper bound on $\Delta_\rho^+(x)$. For the lower bound, we act similarly and derive

$$\begin{aligned}\Delta_\rho(x) = \rho^2 D_\omega^{\text{sym}}(x,\hat{x}) &\leq (1+s)\rho^2 D_\omega^{\text{sym}}(x,x^+) + (1+s^{-1})\rho^2 D_\omega^{\text{sym}}(\hat{x},x^+)\\ &\leq (1+s)\rho^2 D_\omega^{\text{sym}}(x,x^+) + \frac{(1+s^{-1})\ell}{\rho-\ell}\left(\Delta_\rho(x)+\Delta_\rho^+(x)\right)\\ &\leq \left(1+s+\frac{(1+s^{-1})\ell}{\rho-\ell}\right)\Delta_\rho^+(x) + \frac{(1+s^{-1})\ell}{\rho-\ell}\Delta_\rho(x).\end{aligned}$$

Combining the above two inequalities, we have $\frac{\Delta_\rho(x)}{C(\ell,\rho,s)} \leq \Delta_\rho^+(x) \leq C(\ell,\rho,s)\Delta_\rho(x)$.

□

Remark A.2. Notice that our intermediate Lemma A.1 does not require $\omega(\cdot)$ to induce a metric and shows that $\hat{x}$ and x^+ are close if $\Delta_\rho(x)$ and $\Delta_\rho^+(x)$ are small. However, the proof of Lemma 4.1 crucially relies on triangle inequality. Therefore, it is unclear whether convergence of SMD in $\Delta_\rho(x)$ (that was established in [Zhang and He, 2018] under BG assumption) implies convergence in $\Delta_\rho^+(x)$. In our main Theorems 4.3, 4.5 and 4.7, we bypass this issue and directly establish convergence on $\mathcal{D}_\rho(x)$, that is a stronger measure than $\Delta_\rho^+(x)$ (and stronger than $\Delta_\rho(x)$ if $\sqrt{D_\omega^{\text{sym}}(x,y)}$ is a metric). Moreover, as we have seen in Section 2, $\mathcal{D}_\rho(x)$ seems to be a more natural FOSP measure since it reduces to $\|\nabla F(x)\|^2$ in unconstrained case.

BFBE is strictly larger than BGM.

Now we state the proof of Lemma 4.2, which consists of two parts.

Proof of Lemma 4.2. **1. BFBE is not smaller than BGM.** Recall that $x^+ := \arg\min_{y\in\mathcal{X}} \langle \nabla F(x), y\rangle + r(y) + \rho D_\omega(y,x)$. By the optimality condition, there exists $u^+ \in \partial r(x^+)$ such that

$$0 = \nabla F(x) + \rho(\nabla\omega(x^+) - \nabla\omega(x)) + u^+.$$

Thus, by convexity of $r(\cdot)$

$$\begin{aligned} r(x) &\geq r(x^+) + \langle u^+, x - x^+\rangle = r(x^+) + \rho\langle \nabla\omega(x) - \nabla\omega(x^+), x - x^+\rangle - \langle \nabla F(x), x - x^+\rangle\\ &= r(x^+) + \rho(D_\omega(x,x^+) + D_\omega(x^+,x)) - \langle \nabla F(x), x - x^+\rangle.\end{aligned}$$

Using the above inequality and the definition of $\mathcal{D}_\rho(x)$, we derive for any $\rho, \rho_1 > 0$

$$\begin{aligned}\frac{1}{2\rho_1}\mathcal{D}_{\rho_1}(x) &:= -\min_{y\in\mathcal{X}}\left\{\langle \nabla F(x), y - x\rangle + \rho_1 D_\omega(y,x) + r(y) - r(x)\right\}\\ &= \langle \nabla F(x), x - x^+\rangle - \rho_1 D_\omega(x^+,x) + r(x) - r(x^+)\\ &\geq \rho(D_\omega(x,x^+) + D_\omega(x^+,x)) - \rho_1 D_\omega(x^+,x)\\ &\geq (\rho-\rho_1)(D_\omega(x^+,x) + D_\omega(x^+,x)).\end{aligned}$$

Recalling the definition of $\Delta_\rho^+(x)$ and setting $\rho_1 = \rho/2$, it remains to conclude that $\mathcal{D}_{\rho/2}(x) \geq \frac{1}{2}\Delta_\rho^+(x)$.

2. BFBE can be much larger than BGM.

The following example shows how large can be the ratio of BFBE and BGM. Consider minimizing $\Phi(x) = F(x) + r(x)$ over $\mathcal{X} = [0,1] \subset \mathbb{R}^1$ with $F(x) = \frac{1}{2}x^2$, $r(x) = |x|$. We can compute the proximal operator of the absolute value as $\text{prox}_{r/\rho}(x) = \text{sign}(x)\max\left\{0, |x| - \frac{1}{\rho}\right\}$, where $\text{sign}(x) = 1$, if $x \geq 0$ and $\text{sign}(x) = -1$ otherwise. For any $x \in [0,1]$ and $\rho \geq 1$, we can compute $x^+ = \text{prox}_{r/\rho}(x - \rho^{-1}x) = 0$. Therefore, we have

$$\Delta_\rho^+(x) = x^2,$$

$$\mathcal{D}_\rho(x) = -2\rho(\langle \nabla F(x), x^+ - x\rangle + \frac{\rho}{2}(x^+ - x)^2 + |x^+| - |x|) = 2\rho|x| + 2\rho\left(1 - \frac{\rho}{2}\right)x^2.$$

In particular, taking arbitrary $\rho = \rho_1 \geq 1$ in the first equality and $\rho \leq 2$ in the second equality, we have for any $x \in (0,1]$

$$\frac{\mathcal{D}_\rho(x)}{\Delta_{\rho_1}^+(x)} \geq \frac{2\rho|x|}{x^2} \geq \frac{2}{|x|}.$$

We conclude that $\mathcal{D}_\rho(x)$ can be arbitrary larger than $\Delta^+_{\rho_1}(x)$ even when x is close to the optimum $x^* = 0$. This implies that the opposite inequality in Lemma 4.2 does not hold in general even when ρ and ρ_1 are allowed to be different. Therefore, BFBE is strictly stronger convergence measure than BGM.

$\square$

B Proof of Theorem 4.3: Convergence to FOSP in Expectation

Proof. **Step I. Deterministic descent w.r.t. Forward-Backward Envelope.**

Define $\hat{x} := \mathrm{prox}_{\Phi/\rho}(x)$. Notice that for any $x \in \mathcal{X} \cap \mathcal{S}$, we have for any $x^+ \in \mathcal{X} \cap \mathcal{S}$

$$\Phi_{1/\rho}(x) = \Phi(\hat{x}) + \rho D_\omega(\hat{x}, x) \le \Phi(x^+) + \rho D_\omega(x^+, x).$$

We set $x^+ := \arg\min_{y \in \mathcal{X}} \langle \nabla F(x), y\rangle + r(y) + \rho_1 D_\omega(y, x)$ with $\rho_1 \ge \rho + \ell$. Then by relative smoothness (upper bound) of $F(\cdot)$

$$\begin{aligned}
\Phi_{1/\rho}(x) &\le \Phi\left(x^+\right) + \rho D_\omega(x^+, x) \\
&= F\left(x^+\right) + r\left(x^+\right) + \rho D_\omega(x^+, x) \\
&\le F\left(x\right) + \left\langle \nabla F\left(x\right), x^+ - x\right\rangle + \ell D_\omega(x^+, x) + r\left(x^+\right) + \rho D_\omega(x^+, x) \\
&= \Phi\left(x\right) + \left\langle \nabla F\left(x\right), x^+ - x\right\rangle + (\rho+\ell) D_\omega(x^+, x) + r\left(x^+\right) - r\left(x\right) \\
&= \Phi\left(x\right) - \frac{1}{2\rho_1}\mathcal{D}_{\rho_1}(x) + \left(\rho + \ell - \rho_1\right) D_\omega(x^+, x) \\
&\le \Phi\left(x\right) - \frac{1}{2\rho_1}\mathcal{D}_{\rho_1}(x).
\end{aligned} \tag{13}$$

where the last equality holds by definitions of x^+, $\mathcal{D}_\rho(x)$ and the last step is due to condition $\rho_1 \ge \rho + \ell$.

Step II. One step progress on the Lyapunov function.

By the update rule of x_{t+1} and applying Lemma F.1, Item 2 with $z = x_t$, $z^+ = x_{t+1}$, $x = \hat{x}_t$, we have

$$\eta_t\langle \nabla f(x_t, \xi_t), \hat{x}_t - x_{t+1}\rangle + \eta_t(r(\hat{x}_t) - r(x_{t+1})) \ge D_\omega(\hat{x}_t, x_{t+1}) + D_\omega(x_{t+1}, x_t) - D_\omega(\hat{x}_t, x_t). \tag{14}$$

By the optimality of $\hat{x}_{t+1}$ and using the above inequality, we derive

$$\begin{aligned}
\Phi_{1/\rho}\left(x_{t+1}\right) &= \Phi\left(\hat{x}^{t+1}\right) + \rho D_\omega(\hat{x}^{t+1}, x_{t+1}) \\
&\le \Phi\left(\hat{x}^{t}\right) + \rho D_\omega(\hat{x}^{t}, x_{t+1}) \\
&\overset{(14)}{\le} \Phi\left(\hat{x}^{t}\right) + \eta_t\rho\langle \nabla f(x_t, \xi_t), \hat{x}_t - x_{t+1}\rangle + \rho D_\omega(\hat{x}_t, x_t) - \rho D_\omega(x_{t+1}, x_t) + \eta_t\rho(r(\hat{x}_t) - r(x_{t+1})) \\
&= \Phi_{1/\rho}\left(x_t\right) + \eta_t\rho(r(\hat{x}_t) - r(x_t) + \langle \nabla f(x_t, \xi_t), \hat{x}_t - x_t\rangle) + \rho\eta_t\langle \nabla f(x_t, \xi_t), x_t - x_{t+1}\rangle \\
&\qquad - \rho D_\omega(x_{t+1}, x_t) + \eta_t\rho(r(x_t) - r(x_{t+1})).
\end{aligned}$$

We define $\lambda_t := \Phi_{1/\rho}\left(x_t\right) - \Phi^* + \eta_{t-1}\rho(\Phi(x_t) - \Phi^*)$, $\psi_t := \nabla f(x_t, \xi_t) - \nabla F(x_t)$. Then using the above inequality

$$\begin{aligned}
\lambda_{t+1} &:= \Phi_{1/\rho}\left(x_{t+1}\right) - \Phi^* + \eta_t\rho(\Phi(x_{t+1}) - \Phi^*) \\
&\le \Phi_{1/\rho}\left(x_t\right) - \Phi^* + \eta_t\rho(r(\hat{x}_t) - r(x_t) + \langle \nabla f(x_t, \xi_t), \hat{x}_t - x_t\rangle) \\
&\qquad + \rho(\eta_t\langle \nabla f(x_t, \xi_t) - \nabla F(x_t), x_t - x_{t+1}\rangle - D_\omega(x_{t+1}, x_t)) \\
&\qquad + \eta_t\rho(r(x_t) + F(x_{t+1}) + \langle \nabla F(x_t), x_t - x_{t+1}\rangle - \Phi^*) \\
&\overset{(i)}{\le} \Phi_{1/\rho}\left(x_t\right) - \Phi^* + \eta_t\rho(r(\hat{x}_t) - r(x_t) + \langle \nabla f(x_t, \xi_t), \hat{x}_t - x_t\rangle) \\
&\qquad + \rho(\eta_t\langle \nabla f(x_t, \xi_t) - \nabla F(x_t), x_t - x_{t+1}\rangle - (1 - \eta_t\ell) D_\omega(x_{t+1}, x_t)) \\
&\qquad + \eta_t\rho(r(x_t) + F(x_t) - \Phi^*) \\
&= \lambda_t + (\eta_t - \eta_{t-1})\rho(\Phi(x_t) - \Phi^*) + \eta_t\rho(r(\hat{x}_t) - r(x_t) + \langle \nabla F(x_t), \hat{x}_t - x_t\rangle) + \rho\eta_t\langle \psi_t, \hat{x}_t - x_t\rangle \\
&\qquad + \rho(\eta_t\langle \psi_t, x_t - x_{t+1}\rangle - (1 - \eta_t\ell) D_\omega(x_{t+1}, x_t)) \\
&\overset{(ii)}{\le} \lambda_t + (\eta_t - \eta_{t-1})\rho(\Phi(x_t) - \Phi^*) - \frac{\eta_t\rho}{2(\rho+\ell)}\mathcal{D}_{\rho+\ell}(x_t) - \eta_t\rho(\rho - \ell) D_\omega(\hat{x}_t, x_t) + \rho\eta_t\langle \psi_t, \hat{x}_t - x_t\rangle \\
&\qquad + \rho(\eta_t\langle \psi_t, x_t - x_{t+1}\rangle - (1 - \eta_t\ell) D_\omega(x_{t+1}, x_t)) \\
&\overset{(iii)}{\le} \lambda_t - \frac{\eta_t\rho}{2(\rho+\ell)}\mathcal{D}_{\rho+\ell}(x_t) + \rho\eta_t\langle \psi_t, \hat{x}_t - x_t\rangle + \rho(\eta_t\langle \psi_t, x_t - x_{t+1}\rangle - (1 - \eta_t\ell) D_\omega(x_{t+1}, x_t)) \\
&\qquad - \eta_t\rho(\rho - \ell) D_\omega(\hat{x}_t, x_t),
\end{aligned} \tag{15}$$

where (i) follows by relative smoothness (upper bound), i.e., $F(x_{t+1}) \leq F(x_t) - \langle \nabla F(x_t), x_t - x_{t+1} \rangle + \ell D_\omega(x_{t+1}, x_t)$. The inequality (ii) follows from relative smoothness (lower bound) of $F(\cdot)$ and (13) (with $\rho_1 = \rho + \ell$) since

$$\begin{array}{rcl} r(\hat{x}_t) - r(x_t) + \langle \nabla F(x_t), \hat{x}_t - x_t \rangle & \leq & r(\hat{x}_t) - r(x_t) + F(\hat{x}_t) - F(x_t) + \ell D_\omega(\hat{x}_t, x_t) \\ & = & \Phi_{1/\rho}(x_t) - \Phi(x_t) + (\ell - \rho) D_\omega(\hat{x}_t, x_t) \\ & \leq & \Phi_{1/\rho}(x_t) - \Phi(x_t) + (\ell - \rho) D_\omega(\hat{x}_t, x_t) \\ & \leq & -\dfrac{1}{2(\rho+\ell)} \mathcal{D}_{\rho+\ell}(x_t) - (\rho - \ell) D_\omega(\hat{x}_t, x_t). \end{array}$$

The inequality (iii) holds since the sequence $\{\eta_t\}_{t \geq 0}$ is non-increasing.

Step III. Dealing with stochastic terms. Using $D_\omega(x_{t+1}, x_t) \geq \frac{1}{2} \|x_{t+1} - x_t\|^2$ and the bound on the variance of stochastic gradients, we have

$$\begin{array}{rcl} \mathbb{E}\left[\eta_t \langle \psi^t, x_t - x_{t+1} \rangle - (1 - \eta_t \ell) D_\omega(x_{t+1}, x_t)\right] & \leq & \mathbb{E}\left[\eta_t \langle \psi^t, x_t - x_{t+1} \rangle - (1 - \eta_t \ell) \dfrac{1}{2} \|x_{t+1} - x_t\|^2\right] \\ & \leq & \dfrac{\eta_t^2}{2(1 - \eta_t \ell)} \mathbb{E}\left[\|\psi^t\|_*^2\right] \leq \dfrac{\eta_t^2 \sigma^2}{2(1 - \eta_t \ell)}. \end{array} \quad (16)$$

Define $\Lambda_t := \mathbb{E}[\lambda_t]$. Then combining (15) with (16) and setting $\rho = 2\ell$, $\eta_t \leq 1/(2\ell)$, we derive for any non-increasing step-sizes η_t

$$\Lambda_{t+1} \quad \leq \quad \Lambda_t - \frac{\eta_t \rho}{2(\rho + \ell)} \mathbb{E}\left[\mathcal{D}_{\rho+\ell}(x_t)\right] + \frac{\rho \eta_t^2 \sigma^2}{2(1 - \eta_t \ell)} \leq \Lambda_t - \frac{\eta_t}{3} \mathbb{E}\left[\mathcal{D}_{3\ell}(x_t)\right] + 2\ell \eta_t^2 \sigma^2. \quad (17)$$

It remains to telescope and conclude the proof since

$$\sum_{t=0}^{T-1} \eta_t \mathbb{E}\left[\mathcal{D}_{3\ell}(x_t)\right] \leq 3\Lambda_0 + 6\ell\sigma^2 \sum_{t=0}^{T-1} \eta_t^2,$$

where the Lyapunov function in the initial point can be bounded (setting $\eta_{-1} = \eta_0$) as

$$\Lambda_0 = \lambda_0 = \Phi_{1/\rho}(x_0) - \Phi^* + \eta_{-1} \rho (\Phi(x_0) - \Phi^*) \leq \Phi_{1/\rho}(x_0) - \Phi^* + \Phi(x_0) - \Phi^*.$$

The proof for constant step-size follows immediately from (3).

$\square$

C Proof of Theorem 4.5: High Probability Convergence to FOSP under Sub-Gaussian Noise

Proof. We recall the definitions of $\lambda_t = \Phi_{1/\rho}(x_t) - \Phi^* + \eta_{t-1}\rho(\Phi(x_t) - \Phi^*)$, $\psi_t = \nabla f(x_t, \xi_t) - \nabla F(x_t)$, and invoke (15) from the proof of Theorem 4.3

$$\begin{aligned}
\lambda_{t+1} &= \lambda_t - \frac{\eta_t \rho}{2(\rho+\ell)}\mathcal{D}_{\rho+\ell}(x_t) + \rho\eta_t\langle \psi_t, \hat{x}_t - x_t\rangle + \rho\eta_t\langle \psi_t, x_t - x_{t+1}\rangle - \rho(1-\eta_t\ell)D_\omega(x_{t+1}, x_t) \\
&\quad -\eta_t\rho(\rho-\ell)D_\omega(\hat{x}_t, x_t) \\
&\leq \lambda_t - \frac{\eta_t \rho}{2(\rho+\ell)}\mathcal{D}_{\rho+\ell}(x_t) + \rho\eta_t\langle \psi_t, \hat{x}_t - x_t\rangle + \frac{\rho\eta_t^2\|\psi_t\|_*^2}{2(1-\eta_t\ell)} - \eta_t\rho(\rho-\ell)D_\omega(\hat{x}_t, x_t). \qquad (18)
\end{aligned}$$

Define a (normalization) scalar $w := \frac{\rho-\ell}{6\sigma^2\rho\eta_0} > 0$, and a sequence $S_t := \sum_{\tau=t}^{T-1} Z_\tau$, where

$$Z_t := w\left(\lambda_{t+1} - \lambda_t + \frac{\eta_t\rho}{2(\rho+\ell)}\mathcal{D}_{\rho+\ell}(x_t)\right).$$

Now we define the filtration $\mathcal{F}_t = \{x_0, \xi_0, x_1, \ldots, \xi_{t-1}, x_t\}$ and compute the moment generating function (MGF) of Z_t for any $0 \leq t \leq T-1$

$$\begin{aligned}
\mathbb{E}\left[\exp(Z_t)|\mathcal{F}_t\right] &= \mathbb{E}\left[\exp\left(w\left(\lambda_{t+1} - \lambda_t + \frac{\eta_t\rho}{2(\rho+\ell)}\mathcal{D}_{\rho+\ell}(x_t)\right)\right)|\mathcal{F}_t\right] \\
&\overset{(18)}{\leq} \mathbb{E}\left[\exp\left(w\rho\eta_t\langle\psi^t, \hat{x}_t - x_t\rangle + \frac{w\rho\eta_t^2\|\psi_t\|_*^2}{2(1-\eta_t\ell)} - w\eta_t\rho(\rho-\ell)D_\omega(\hat{x}_t, x_t)\right)|\mathcal{F}_t\right] \\
&= \exp\left(-w\eta_t\rho(\rho-\ell)D_\omega(\hat{x}_t, x_t)\right)\mathbb{E}\left[\exp\left(w\rho\eta_t\langle\psi^t, \hat{x}_t - x_t\rangle + \frac{w\rho\eta_t^2\|\psi_t\|_*^2}{2(1-\eta_t\ell)}\right)|\mathcal{F}_t\right] \\
&\overset{(i)}{\leq} \exp\left(-w\eta_t\rho(\rho-\ell)D_\omega(\hat{x}_t, x_t)\right)\exp\left(3\sigma^2 w^2\rho^2\eta_t^2\|\hat{x}_t - x_t\|^2 + \frac{3\sigma^2 w\rho\eta_t^2}{2(1-\eta_t\ell)}\right) \\
&\overset{(ii)}{\leq} \exp\left(\frac{3\sigma^2 w\rho\eta_t^2}{2(1-\eta_t\ell)}\right),
\end{aligned}$$

where in (i) we apply Lemma F.2, which uses that $\|\psi_t\|_*$ is σ-sub-Gaussian. Inequality (ii) holds by the fact that $\|\hat{x}_t - x_t\|^2 \leq 2D_\omega(\hat{x}_t, x_t)$, and the choice of w, which guarantess that $6\sigma^2 w^2\rho^2\eta_t^2 \leq w\eta_t\rho(\rho-\ell)$ for any $t \geq 0$. Now to compute the MGF of S_t we use derive

$$\mathbb{E}\left[\exp(S_t)|\mathcal{F}_t\right] = \mathbb{E}\left[\mathbb{E}\left[\exp(S_{t+1} + Z_t)|\mathcal{F}_{t+1}\right]|\mathcal{F}_t\right] = \mathbb{E}\left[\exp(Z_t)\mathbb{E}\left[\exp(S_{t+1})|\mathcal{F}_{t+1}\right]|\mathcal{F}_t\right].$$

Thus, by induction we have

$$\mathbb{E}\left[S_0\right] \leq \exp\left(\frac{3\sigma^2\rho w}{2}\sum_{t=0}^{T-1}\frac{\eta_t^2}{(1-\eta_t\ell)}\right) \leq \exp\left(3\sigma^2\rho w\sum_{t=0}^{T-1}\eta_t^2\right).$$

where the last inequality holds by the condition $\eta_t \leq 1/(2\ell)$. Consequently, by Markov's inequality,

$$\Pr\left(S_0 \geq 3\sigma^2\rho\sum_{t=0}^{T-1} w_t\eta_t^2 + \log\left(1/\beta\right)\right) \leq \beta.$$

Then with probability at least $1-\beta$, we have

$$\sum_{t=0}^{T-1} w\left(\lambda_{t+1} - \lambda_t + \frac{\eta_t\rho}{2(\rho+\ell)}\mathcal{D}_{\rho+\ell}(x_t)\right) \leq 3\sigma^2 w\rho\sum_{t=0}^{T-1}\eta_t^2 + \log\left(1/\beta\right).$$

Telescoping the above inequality, setting $\rho = 4\ell$, and dividing by the sum of step-sizes:

$$
\begin{aligned}
\frac{1}{\sum_{t=0}^{T-1} \eta_t} \sum_{t=0}^{T-1} \eta_t \, \mathcal{D}_{5\ell}(x_t) \quad &\leq \quad \frac{\lambda_0 + \frac{1}{w} \log\left(1/\beta\right) + 12\sigma^2 \ell \sum_{t=0}^{T-1} \eta_t^2}{\frac{2}{5}\sum_{t=0}^{T-1} \eta_t} \\
&= \quad \frac{\lambda_0 + 8\, \eta_0 \sigma^2 \log\left(1/\beta\right) + 12\sigma^2 \ell \sum_{t=0}^{T-1} \eta_t^2}{\frac{2}{5}\sum_{t=0}^{T-1} \eta_t}.
\end{aligned}
$$

It remains to bound the Lyapunov function in the initial point setting $\eta_{-1} = \eta_0$:

$$
\lambda_0 = \Phi_{1/\rho}(x_0) - \Phi^* + \eta_{-1}\rho(\Phi(x_0) - \Phi^*) \leq \Phi_{1/\rho}(x_0) - \Phi^* + 2(\Phi(x_0) - \Phi^*).
$$

$\square$

D Proof of Theorem 4.7. Global Convergence under Generalized Proximal PŁ Condition

Proof of Theorem 4.7. Invoking (17), and substituting the value of $\rho = 2\ell$ and using $\eta_t \le \frac{1}{2\ell}$, we derive under Assumption 4.6

$$\Lambda_{t+1} \le \Lambda_t - \frac{\eta_t}{3}\mathbb{E}\left[\mathcal{D}_{3\ell}(x_t)\right] + 2\ell\eta^2\sigma^2 \le \Lambda_t - \frac{2\mu\eta_t}{3}\mathbb{E}\left[(\Phi(x_t) - \Phi^*)^{2/\alpha}\right] + 2\ell\eta^2\sigma^2,$$

Notice that $\Phi(x) \ge \Phi_{1/\rho}(x)$ for any $x \in \mathcal{X}$, thus

$$\mathbb{E}\left[\Phi(x_t) - \Phi^*\right] \ge \frac{1}{1+\eta_{t-1}\rho}\mathbb{E}\left[\Phi_{1/\rho}(x_t) - \Phi^*\right] + \frac{\eta_{t-1}\rho}{1+\eta_{t-1}\rho}\mathbb{E}\left[\Phi(x_t) - \Phi^*\right] = \frac{\Lambda_t}{1+\eta_{t-1}\rho} \ge \frac{1}{2}\Lambda_t.$$

By Jensen's inequality for $z \to z^{2/\alpha}$, we have $\mathbb{E}\left[(\Phi(x_t) - \Phi^*)^{2/\alpha}\right] \ge \left(\mathbb{E}\left[\Phi(x_t) - \Phi^*\right]\right)^{2/\alpha}$. Combining the above inequalities, we get $\mathbb{E}\left[(\Phi(x_t) - \Phi^*)^{2/\alpha}\right] \ge \frac{1}{2}\Lambda_t^{2/\alpha}$. Thus, we can derive a recursion

$$\Lambda_{t+1} \le \Lambda_t - \frac{\eta_t\mu}{3}\Lambda_t^{2/\alpha} + 2\ell\sigma^2\eta_t^2.$$

Assume that for $\tau = 0, \ldots, t$, we have $\Lambda_\tau \ge \varepsilon$ (otherwise we have reached ε-accuracy). Then

$$\Lambda_{t+1} \le \left(1 - \frac{\eta_t\mu\varepsilon^{\frac{2-\alpha}{\alpha}}}{3}\right)\Lambda_t + 2\ell\sigma^2\eta_t^2,$$

For any $T \ge 0$, we select the step-size sequence η_t as follows:

$$\eta_t = \begin{cases} \frac{1}{2\ell} & \text{if } t < \lceil T/2 \rceil \text{ and } T \le \frac{6\ell}{\mu\,\varepsilon^{\frac{2-\alpha}{\alpha}}}, \\ \frac{6}{\mu\,\varepsilon^{\frac{2-\alpha}{\alpha}}\left(t + \frac{12\ell}{\mu}\varepsilon^{-\frac{2-\alpha}{\alpha}} - \lceil T/2 \rceil\right)} & \text{otherwise.} \end{cases}$$

By Lemma F.6, we have

$$\Lambda_{T+1} = \mathcal{O}\left(\frac{\ell\Lambda_0}{\mu\,\varepsilon^{\frac{2-\alpha}{\alpha}}}\exp\left(-\frac{\mu\,\varepsilon^{\frac{2-\alpha}{\alpha}}T}{\ell}\right) + \frac{\ell\sigma^2}{T\mu^2\varepsilon^{\frac{2(2-\alpha)}{\alpha}}}\right).$$

Recalling the definition of Λ_t and using Lemma F.3, we bound $\Lambda_{T+1} \ge \Phi_{1/\rho}(x_{T+1}) - \Phi^* \ge \Phi(x_{T+1}^+) - \Phi^*$, where $x^+ := \arg\min_{y\in\mathcal{X}} \langle \nabla F(x), y\rangle + r(y) + \ell D_\omega(y, x)$. The sample complexity to reach $\Phi(x_{T+1}^+) - \Phi^* \le \varepsilon$ is

$$T = \mathcal{O}\left(\frac{\ell\Lambda_0}{\mu}\frac{1}{\varepsilon^{\frac{2-\alpha}{\alpha}}}\log\left(\frac{\ell\Lambda_0}{\mu\varepsilon}\right) + \frac{\ell\Lambda_0\sigma^2}{\mu^2}\frac{1}{\varepsilon^{\frac{4-\alpha}{\alpha}}}\right).$$

□

E Proofs for Applications

E.1 Differentially Private Learning in ℓ_2 and ℓ_1 Settings

Proof of Corollary 5.3. Notice that by Lemma F.7 $\|b_t\|_\infty$ is σ-sub-Gaussian r.v. with $\sigma^2 = 2\log(2d)\sigma_G^2$.[20] We invoke the result of Theorem 4.5 with $\eta_t = \eta_0 = \frac{1}{2\ell}$, and obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{D}_{5\ell}(x_t) &= \mathcal{O}\left(\frac{\lambda_0}{\eta_0 T} + \sigma^2\eta_0\ell + \frac{\sigma^2\log\left(\frac{1}{\beta}\right)}{T}\right) = \mathcal{O}\left(\frac{\ell\lambda_0}{T} + \sigma^2\log\left(\frac{1}{\beta}\right)\right)\\ &= \mathcal{O}\left(\frac{\ell\lambda_0}{T} + \frac{G^2 T\log(d)\log\left(\frac{1}{\delta}\right)}{n^2\epsilon^2}\log\left(\frac{1}{\beta}\right)\right)\\ &= \mathcal{O}\left(\frac{G\sqrt{\ell\lambda_0\log(d)\log\left(\frac{1}{\delta}\right)\log\left(\frac{1}{\beta}\right)}}{n\epsilon}\right),\end{aligned}$$

where the last equality follows by the choice of T. It remains to notice that $\lambda_0 = \Phi_{1/\rho}(x_0) - \Phi^* + 2(\Phi(x_0) - \Phi^*) \le 3(\Phi(x_0) - \Phi^*)$. $\square$

Corollary E.1. *Let $F(\cdot)$ be differentiable on a convex set $\mathcal{X}$ with L-Lipschitz continuous gradient w.r.t. Euclidean norm, and $\|\nabla F(x)\|_2 \le G$ for all $x \in \mathcal{X}$. Set $\eta_t = \frac{1}{2L}$, $T = \frac{n\epsilon\sqrt{L}}{G\sqrt{d\log(1/\delta)\log(1/\beta)}}$, $\lambda_0 := \Phi(x_0) - \Phi^*$. Then* DP-Prox-GD *is (ϵ,δ)-DP and with probability $1-\beta$ satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{D}_{5\ell}(x_t) = \mathcal{O}\left(\frac{G\sqrt{\ell\lambda_0 d\log(1/\delta)\log(1/\beta)}}{n\epsilon}\right),$$

Proof. The proof follows the same lines as the proof of Corollary 5.3. The only difference is that instead of the infinity norm of the noise, we bound the Euclidean norm, i.e., $\|b_t\|_2$ is σ-sub-Gaussian r.v. with $\sigma^2 = d\,\sigma_G^2$. $\square$

E.2 Policy Optimization in Reinforecement Learning

Prox-PŁ condition.

Now we will verify Assumption 4.6 with $\alpha = 1$ holds for our RL problem. The result is similar to Lemma 5 in [Xiao, 2022]. The only difference is that we have π instead of π^+ on the left hand side of the inequality.

Lemma E.2. *Let $\omega(\pi) = \frac{1}{2}\|\pi\|_{2,2}^2$. Then for any $\pi \in \mathcal{X}$ we have*

$$V_p(\pi) - V_p^\star \le \frac{2\sqrt{2|\mathcal{S}|}}{1-\gamma}\left\|\frac{d_p(\pi^\star)}{\mu}\right\|_\infty\sqrt{\Delta_\rho^+(\pi)}$$

if $\rho \ge G_{V,\|\cdot\|_{2,2}}/D_{\mathcal{X},\|\cdot\|_{2,2}}$, where $G_{V,\|\cdot\|_{2,2}} := \max_{\pi\in\Pi}\left[\|\nabla V_p(\pi)\|_{2,2}\right]$.

Proof. It was shown in Lemma 4 in [Agarwal et al., 2021] that the following (variational) gradient domination condition holds.

$$V_p(\pi) - V_p^\star \le \frac{1}{1-\gamma}\left\|\frac{d_p(\pi^\star)}{\mu}\right\|_\infty \max_{\pi'\in\mathcal{X}}\langle\nabla V_\mu(\pi), \pi - \pi'\rangle \quad \text{for any } \pi \in \mathcal{X}.$$

By Lemma F.5, we have

$$\max_{\pi'\in\mathcal{X}}\langle\nabla V_\mu(\pi), \pi - \pi'\rangle \le \left(D_{\mathcal{X},\|\cdot\|_{2,2}} + \rho^{-1}G_{V,\|\cdot\|_{2,2}}\right)\sqrt{\Delta_\rho^+(\pi)} \le 2D_{\mathcal{X},\|\cdot\|_{2,2}}\sqrt{\Delta_\rho^+(\pi)}$$

[20] Here we used the fact that $\max_{1\le i\le d}|\xi_i| = \max\{\xi_1, -\xi_1, \ldots, \xi_d, -\xi_d\}$ and applied Lemma F.7 with $n = 2d$.

where the last inequality holds for $\rho \geq G_{V,\|\cdot\|_{2,2}}/D_{\mathcal{X},\|\cdot\|_{2,2}}$. It remains to notice that $D_{\mathcal{X},\|\cdot\|_{2,2}} \leq \sqrt{2|\mathcal{S}|}$.

Smoothness in $(2,1)$-norm. Proof of Proposition 5.5

For any $\pi, \pi' \in \mathcal{X}$, it holds that

$$\|\nabla V_p(\pi) - \nabla V_p(\pi')\|_{2,\infty} \leq \frac{2\gamma}{(1-\gamma)^3} \|\pi - \pi'\|_{2,1},$$

The estimate of the smoothness constant follows directly from the proof of Lemma 54 in [Agarwal et al., 2021], since using $(2,1)$ norm we have $\sum_{a\in\mathcal{A}} |u_{a,s}| \leq 1$, and the perturbation $u_{a,s}$ belongs to the probability simplex $u_{a,s} \in \Delta(\mathcal{A})$.

□

E.3 Training Autoencoder Model using SMD

Derivation of SMDr1 and SMDr2. Recall the choice of DGF from subsection 5.3

$$\omega(x) = \frac{1}{r+2}\|x\|_2^{r+2} + \frac{1}{2}\|x\|_2^2.$$

Notice that we have $\nabla\omega(x) = \|x\|_2^r x + x$. The update rule of SMD with $\mathcal{X} = \mathbb{R}^d$, $r(x) = 0$ and the above choice of DGF satisfies

$$\eta_t \nabla f(x_t, \xi_t) + \nabla\omega(x_{t+1}) - \nabla\omega(x_t) = 0.$$

Define $c_t := \nabla\omega(x_t) - \eta_t\nabla f(x_t,\xi_t) = x_t - \eta_t \nabla f(x_t, \xi_t) + \|x_t\|_2^r x_t$. Thus, it remains to solve for x_{t+1}

$$\nabla\omega(x_{t+1}) = (\|x_{t+1}\|_2^r + 1)\, x_{t+1} = c_t, \tag{19}$$

which is equivalent to solving the following simple univariate equation of $\theta \geq 0$:

$$\theta^{r+1} + \theta = \|c_t\|_2. \tag{20}$$

For $r = 1, 2$, it has an explicit form solution for any $\|c_t\|_2$. We have

$$\theta_* = \frac{-1 + \sqrt{1 + 4\|c_t\|_2}}{2} \qquad \text{for } r = 1.$$

and obtain the following method

$$\begin{aligned} c_t &= x_t - \eta_t \nabla f(x_t, \xi_t) + \|x_t\|_2 x_t, \\ x_{t+1} &= \frac{2c_t}{1 + \sqrt{1 + 4\|c_t\|_2}}. \end{aligned}$$

For $r = 2$, an explicit form solution can be written using Cardano's formula. We use Python `Sympy` library for symbolic calculation to solve for θ in this case.

More generally, (19) implies for any $r > 0$, we have

$$\begin{aligned} c_t &= (1 + \|x\|_2^r)\, x_t - \eta_t \nabla f(x_t, \xi_t), \\ x_{t+1} &= \frac{c_t}{1 + \theta_*^r}, \end{aligned}$$

where θ_* is the solution to (20), which can be solved using a bisection method up to the machine accuracy.

Corollary E.3. *Let $F(\cdot) : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable and satisfy (8). Let Assumption 3.2 hold with $\|\cdot\|_2$. Suppose the sequence $\{\eta_t\}_{t\geq 0}$ be non-increasing with $\eta_0 \leq 1/(2\ell)$, and $\bar{x}_T \in \mathcal{X}$ be randomly chosen from the iterates $x_0, \ldots, x_{T-1}$ with probabilities $p_t = \eta_t / \sum_{t=0}^{T-1} \eta_t$. Then for (9), (10), we have*

$$\mathbb{E}\left[\|\nabla F(\bar{x}_T)\|_2^2\right] \leq \frac{6(F(x_0) - F^*) + 6\ell\sigma^2 \sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t},$$

where $F^* := \min_{y\in\mathbb{R}^d} F(y)$.

Additional experimental details. We use `Fashion-MNIST` dataset [Xiao et al., 2017] for training with images of dimensions $d_f = 28 \times 28 = 784$. The encoding dimension is fixed to $d_e = 64$. The dataset is of size 50000 images. In all experiments, we use the mini-batch of size 100. We initialize the parameters W of the model with a normal distribution with mean 1 and the standard deviation 0.01.

Remark E.4. A momentum variant of the scheme (9), (10) was recently explored in [Ding et al., 2023] with promising empirical results on image classification and language modeling tasks. We hope that our simpler variant without momentum can be also helpful in these tasks.

Additional discussion about (L_0, L_1)-smoothness. Recently, some works, e.g., [Zhang et al., 2020a, Faw et al., 2023, Hübler et al., 2023], consider adaptive gradient methods such as gradient clipping, AdaGrad-Norm and gradient normalization under (L_0, L_1)-smoothness, i.e., $F(\cdot)$ is twice differentiable and for some $L_0, L_1 \geq 0$ satisfies $\left\|\nabla F(x)\right\|_{\text{op}} \leq L_0 + L_1 \left\|\nabla F(x)\right\|_2$ for all $x \in \mathbb{R}^d$. The authors in [Zhang et al., 2020a, Faw et al., 2023] justify the theoretical benefits of the popular adaptive schemes by the fact that, unlike `SGD`, they provably work under this weaker (L_0, L_1)-smoothness. Moreover, Zhang et al. [2020a] empirically verify that (L_0, L_1)-smoothness condition holds on the optimization trajectory when training modern language and image classification models. Our polynomial grow condition is weaker than (L_0, L_1)-smoothness as long as the gradient norm grows at most as a polynomial in $\left\|x\right\|_2$. Unlike the approach taken in the above mentioned works, the convergence of our algorithm with the choice of DGF as in Proposition 5.9 follows directly from Theorem 4.3 and does not require a separate analysis.

F Useful Lemma

The following lemma is standard [Lu et al., 2018] and the proof can be found, e.g., in [Chen and Teboulle, 1993].

Lemma F.1. *1. The Bregman divergence satisfies the three-point identity:*

$$D_\omega(x,y) + D_\omega(y,z) = D_\omega(x,z) + \langle \nabla\omega(z) - \nabla\omega(y), x - y\rangle \quad \text{for all } y, z \in \mathcal{S} \text{ and } x \in cl(\mathcal{S}).$$

2. Let $\phi(\cdot)$ be a closed proper convex function on $\mathbb{R}^d$, $z \in \mathcal{S}$ and $z^+ := \arg\min_{x\in\mathcal{X}} \{\phi(x) + \rho D_\omega(x,z)\}$ for $\rho > 0$, then

$$\phi(x) + \rho D_\omega(x,z) \ge \phi(z^+) + \rho D_\omega(z^+,z) + \rho D_\omega(x,z^+) \quad \text{for all } x \in cl(\mathcal{S}).$$

To establish high probability convergence, we use the technical lemma by Liu et al. [2023].

Lemma F.2 (Lemma 2.2. in [Liu et al., 2023])**.** *Suppose $X \in \mathbb{R}^d$ such that $\mathbb{E}[X] = 0$ and $\|X\|_*$ is a σ-sub-Gaussian random variable, then for any $a \in \mathbb{R}^d, 0 \le b \le \frac{1}{2\sigma}$,*

$$\mathbb{E}\left[\exp\left(\langle a, X\rangle + b^2\|X\|_*^2\right)\right] \le \exp\left(3\left(\|a\|^2 + b^2\right)\sigma^2\right).$$

The following lemma shows the connection between $\Phi_{1/\rho}$ and Φ. Similar result in the Euclidean setting has previously appeared, e.g., in [Stella et al., 2017].

Lemma F.3. *Let $F(\cdot)$ be (ℓ,ω)-smooth. Then for any $\rho \ge 2\ell$ and $x \in \mathcal{X}\cap\mathcal{S}$ we have $\Phi_{1/\rho}(x) \ge \Phi(x^+)$, where $x^+ := \arg\min_{y\in\mathcal{X}} \langle \nabla F(x), y\rangle + r(y) + (\rho-\ell)D_\omega(y,x)$.*

Proof. By Assumption 3.1 (lower bound), we have for any $x, y \in \mathcal{X}\cap\mathcal{S}$

$$\Phi(y) + \rho D_\omega(y,x) \ge F(x) + \langle \nabla F(x), y - x\rangle + r(y) + (\rho-\ell)D_\omega(y,x).$$

Minimizing both sides over $y \in \mathcal{X}\cap\mathcal{S}$, we have

$$\begin{aligned}\Phi_{1/\rho}(x) &\ge F(x) + \langle \nabla F(x), x^+ - x\rangle + r(x^+) + (\rho-\ell)D_\omega(x^+,x)\\ &\ge F(x^+) + r(x^+) + (\rho-2\ell)D_\omega(x^+,x) \ge \Phi(x^+),\end{aligned}$$

where the first equality holds by the definitions of $\Phi_{1/\rho}$ and x^+. The second inequality uses Assumption 3.1 (upper bound). □

The following lemma shows that our Assumption 4.6 is more general than relative strong convexity [Lu et al., 2018]. In the Euclidean case, the same result was derived by Karimi et al. [2016].

Lemma F.4 (Relative strong convexity implies 2-Bregman Prox-PŁ)**.** *Let $F(\cdot)$ be μ-relatively strongly convex w.r.t. $\omega(\cdot)$, i.e., for all $x, y \in \mathcal{X}\cap\mathcal{S}$*

$$F(y) \ge F(x) + \langle \nabla F(x), y - x\rangle + \mu D_\omega(y,x). \tag{21}$$

Then Assumption 4.6 holds wth $\alpha = 2$ and any $\rho \ge \mu$, i.e., $\mathcal{D}_\rho(x) \ge 2\mu\left(\Phi(x) - \Phi^*\right)$.

Proof. Adding $r(y)$ to both sides of (21), we have

$$\Phi(y) \ge \Phi(x) + \langle \nabla F(x), y - x\rangle + \mu D_\omega(y,x) + r(y) - r(x) = \Phi(x) + Q_\mu(x,y).$$

Minimizing both sides over $y \in \mathcal{X}\cap\mathcal{S}$, we get

$$\Phi^* \ge \Phi(x) + \min_{y\in\mathcal{X}} Q_\mu(x,y) = \Phi(x) - \frac{1}{2\mu}\mathcal{D}_\mu(x).$$

Rearranging and noticing that $\mathcal{D}_\mu(x) \le \mathcal{D}_\rho(x)$ for any $x \in \mathcal{X}$ and $\rho \ge \mu$, we obtain the result. □

The following lemma connects the Frank-Wolfe gap with the norm of the gradient mapping in the Euclidean case.

Lemma F.5 (Lemma 2.2 in [Balasubramanian and Ghadimi, 2022]). *Let $\omega(x) := \frac{1}{2}\|x\|_2^2$, $\mathcal{X}$ be a compact set with diameter $D_{\mathcal{X},\|\cdot\|_2} := \max_{x,y\in\mathcal{X}}\|x-y\|_2$ and $r(\cdot)=0$. Then for any $\rho>0$*

$$\max_{y\in\mathcal{X}}\langle\nabla F(x), x-y\rangle \le \left(D_{\mathcal{X},\|\cdot\|_2} + \rho^{-1}G_{F,\|\cdot\|_2}\right)\sqrt{\Delta_\rho^+(x)},$$

where $G_{F,\|\cdot\|_2} := \max_{x\in\mathcal{X}}\|\nabla F(x)\|_2$.

We report the special case of Lemma 3 by Stich [2019].

Lemma F.6 (Lemma 3 in [Stich, 2019]). *Let $\{r_t\}_{t\ge 0}$ and $\{\eta_t\}_{t\ge 0}$ be two non-negative sequences with $\eta_t \le \frac{1}{d}$ that satisfy the relation*

$$r_{t+1} \le (1-a\eta_t)\,r_t + c\,\eta_t^2,$$

where $a>0$, $c\ge 0$. For any $T\ge 0$, set

$$\eta_t = \begin{cases} \frac{1}{d} & \text{if } t < \lceil T/2\rceil \text{ and } T \le \frac{2d}{a}, \\ \frac{1}{a\left(\frac{2d}{a}+t-\lceil T/2\rceil\right)} & \text{otherwise.}\end{cases}$$

Then we have

$$r_{t+1} \le \frac{32\,d\,r_0}{a}\exp\left(-\frac{aT}{2d}\right) + \frac{36\,c}{a^2T}.$$

The next lemma is standard and the proof can be found, e.g., in [Van Handel, 2014].

Lemma F.7 (Maximal tail inequality, Lemma 5.1 and 5.2 in [Van Handel, 2014]). *Let ξ_i be a σ-sub-Gaussian random variable for every $i=1,\ldots,n$. Then*

$$\left(\mathbb{E}\left[\max_{1\le i\le n}\xi_i\right]\right)^2 \le \mathbb{E}\left[\max_{1\le i\le n}\xi_i^2\right] \le 2\sigma^2\log(n),$$

$$\Pr\left(\max_{1\le i\le n}\xi_i \ge \sqrt{2\sigma^2\log(n)}+\lambda\right) \le e^{-\frac{\lambda^2}{2\sigma^2}} \qquad \text{for all } \lambda\ge 0.$$