Extending DSPL to accommodate high-order growth in $c(x, \xi)$

Distributed Stochastic Prox-linear Mirror Descent

In this note we relax the Lipschitz continuity condition in the analysis of **DSPL** leveraging the tool of relative Lipschitzness. Recall that in the original proof we require $h(c(x, \xi))$ is L-Lipschitz, which is implied by Lipschitzness of both h and c. To further extend the coverage of **DSPL** and inspired by [ZH18, MOPS19, ZCZL22], we extend the analysis of **DSPL** to the relative Lipschitz case and prove the convergence of the Distributed Stochastic Prox-linear Mirror Descent.

Our goal is to prove the same $\mathcal{O}(\frac{1}{\sqrt{K}} + \frac{\tau^2}{K})$ rate as in Lipschizness case using the Bregman Moreau envelope and to the best of our knowledge, this is also the first result for distributed stochastic prox-linear mirror descent for weakly convex optimization. For brevity of exposition we from now on let $\omega = 0$.

1. Preliminaries

1.1 Bregman divergence

Given a smooth convex function $d : \mathcal{X} \to \mathbb{R}$, Bregman divergence w.r.t. Bregman kernel d is defined by

$$V_d(x, y) := d(x) - d(y) - \langle \nabla d(y), x - y \rangle.$$

Bregman divergence is a natural generalization of the ℓ_2-distance and can be used to analyze the convergence of non-Lipschitz functions that are "relative-Lipschitz", including functions of high-order growth. We will add more background on relative Lipschitzness in the revised appendix and refer the reviewers to [Lu19, ZH18, MOPS19] for the applications of relative Lipschitzness to weakly (or non)convex optimization.

With notion from Bregman proximal method, we introduce the Bregman Moreau envlope [ZH18] and the corresponding Bregman proximal mapping

$$\begin{aligned} f^{V_d}_{1/\rho}(x) &:= \min_y \{ f(y) + \rho V_d(y, x) \}, \\ \operatorname{prox}^{V_d}_{f_{1/\rho}}(x) &:= \arg\min_y \{ f(y) + \rho V_d(y, x) \}. \end{aligned}$$

and [ZH18] shows that $V_d(\operatorname{prox}^{V_d}_{f_{1/\rho}}(x), x)$ is a proper measure of approximate stationarity at x.

1.2 Assumptions

With relative Lipschitzness in hand, we make the following assumptions to accommodate the potential higher-order growth condition of c.

1. **A1** (i.i.d. sample) It is possible to draw i.i.d. samples $\{\xi^k\}$ from Ξ.
2. **A2** (Relative Lipschitz-continuity and smoothness) h is convex and L_h-Lipschitz, $c(x, \xi)$ is C-smooth and M-relative Lipschitz to some $V_d(x, y)$. i.e., we have [Lu19] $\|\nabla c(z, \xi)\| \le \frac{M\sqrt{2V_d(y,x)}}{\|y-x\|}$ for any $y \neq x$.
3. **A3** The Bregman kernel d is 1-strongly convex and satisfies α-symmetry condition on its domain such that $\alpha V_d(y, x) \le V_d(x, y) \le \alpha^{-1} V_d(y, x), \forall x, y \in \operatorname{dom}(d)$, where $\alpha \in (0, 1]$ measures the symmetry of the divergence.

Remark

Assumption **A2** and **A3** are mild and allow c to exhibit high order growth. e.g, if $d(x) = x^4$, then we have $\alpha \ge 0.263$ and we can follow [Lu19] to construct a kernel d for as long as $\frac{c(x,\xi) - c(y,\xi)}{\|x-y\|}$ is upper-bounded by a polynomial of $\|x\|$ and $\|y\|$.

relip-Proposition 1

Assume that **A1** to **A3** hold, then the stochastic function $f_z(x, \xi)$ satisfies the following properties

1. (Convexity) $f_z(x, \xi)$ is convex for any $x, z \in \operatorname{dom}(d), \xi \sim \Xi$.

2. (Two-sided approximation) $|f(x,\xi)-f_y(x,\xi)|\le\frac{L_hC}{2}\|x-y\|^2, \forall x,y\in\mathrm{dom}(d),\xi\sim\Xi$.
3. (Relative Lipschizness) $f_z(x,\xi)-f_z(y,\xi)\le L_hM\sqrt{2V_d(y,x)}, \forall x,y,z\in\mathrm{dom}(d),\xi\sim\Xi$

The convexity of $f_z(x,\xi)$ is by definition and the rest two properties hold by the following deductions

$$\begin{aligned}|f(x,\xi)-f_y(x,\xi)| &= |h(c(x,\xi))-h(c(y,\xi)+\langle\nabla c(y,\xi),x-y\rangle)|\\ &\le L_h\|c(x,\xi)-c(y,\xi)-\langle\nabla c(y,\xi),x-y\rangle\|\\ &\le\frac{L_hC}{2}\|x-y\|^2,\end{aligned}$$

and

$$\begin{aligned}f_z(x,\xi)-f_z(y,\xi) &= h(c(z,\xi)+\langle\nabla c(z,\xi),x-z\rangle)-h(c(z,\xi)+\langle\nabla c(z,\xi),y-z\rangle)\\ &\le L_h|\langle\nabla c(z,\xi),x-y\rangle|\\ &\le L_h\frac{M\sqrt{2V_d(y,x)}}{\|x-y\|}\cdot\|x-y\|\\ &= L_hM\sqrt{2V_d(y,x)},\end{aligned}$$

where the first inequality is by L_h-Lipschitzness of h and the second is by the definition of M-relative Lipschitzness.

Still for brevity we let $\lambda=L_hC$ and $L=\sqrt{2}L_hM$ and $f_z(x,\xi)-f_z(y,\xi)\le L\sqrt{V_d(y,x)}$.

Summary of result

With the above tools and assumptions in hand, our goal is to use relative Lipschitzness to extend our main **Theorem 1** (*Line 180*) to accommodate high-order growth of *c*.

---

relip-Theorem (Informal) (Convergence under relative-Lipschitzness)

Let $\gamma_k\equiv\gamma\sim\mathcal{O}\left(\sqrt{K}\right)$ and k^* be an index chosen between 1 and K uniformly, then

$$\mathbb{E}[V_d(\hat{x}^{k^*},x^{k^*})]=\mathcal{O}\left(\frac{1}{\sqrt{K}}+\frac{\tau^2}{K}\right).$$

This result relaxes the original assumption that requires Lipschizness of *c* and greatly extends the coverage our method.

2. Convergence Analysis

Given Bregman kernel d and the induced divergence V_d, we solve the following Bregman proximal subproblem in each iteration

$$x^{k+1}=\arg\min_x\left\{f_{x^{k-\tau_k}}(x,\xi^{k-\tau_k})+\gamma_kV_d(x,x^k)\right\}$$

and we define $\hat{x}^k:=\mathrm{prox}_{f_{1/\rho}}^{V_d}(x^k)$. First we can derive a similar result to the auxiliary **Lemma 5** (*Line 477*) as follows.

---

relip-Lemma 5 (Auxiliary Lemma 5 under rel.Lip.)

Assume that the above assumptions hold. Then

$$\left|\mathbb{E}_k\left[\mathbb{E}_\xi\left[f_{x^{k-\tau_k}}(x^{k+1},\xi)\right]-f_{x^{k-\tau_k}}(x^{k+1},\xi^{k-\tau_k})\right]\right|\le\frac{L^2(1+\sqrt{1/\alpha})}{\gamma_k(1+\alpha)}\tag{1}$$

for any $\gamma_k>0$.

Following the proof of **Lemma 5** from *Line 477*, define $\mathcal{A}(z,x,\xi):=\arg\min_w\{f_z(w,\xi)+\gamma V_d(w,x)\}$ and

$$
\begin{aligned}
\mathcal{A}(z,x,\xi') &= \arg\min_x \{f_z(w,\xi') + \gamma V_d(w,x)\} \\
\mathcal{A}(z,x,\xi) &= \arg\min_x \{f_z(w,\xi) + \gamma V_d(w,x)\}.
\end{aligned} \tag{2}
$$

It follows by three-point lemma that

$$
\begin{aligned}
& f_z(\mathcal{A}(z,x,\xi'),\xi') + \gamma V_d(\mathcal{A}(z,x,\xi'),x) \\
& \leq f_z(\mathcal{A}(z,x,\xi),\xi') + \gamma V_d(\mathcal{A}(z,x,\xi),x) - \gamma V_d(\mathcal{A}(z,x,\xi),\mathcal{A}(z,x,\xi'))
\end{aligned} \tag{3}
$$

and that

$$
\begin{aligned}
& f_z(\mathcal{A}(z,x,\xi),\xi) + \gamma V_d(\mathcal{A}(z,x,\xi),x) \\
& \leq f_z(\mathcal{A}(z,x,\xi'),\xi) + \gamma V_d(\mathcal{A}(z,x,\xi'),x) - \gamma V_d(\mathcal{A}(z,x,\xi'),\mathcal{A}(z,x,\xi)).
\end{aligned} \tag{4}
$$

Summing (3) and (4) and re-arranging the terms,

$$
\begin{aligned}
& \gamma[V_d(\mathcal{A}(z,x,\xi'),\mathcal{A}(z,x,\xi)) + V_d(\mathcal{A}(z,x,\xi),\mathcal{A}(z,x,\xi'))] \\
& \leq f_z(\mathcal{A}(z,x,\xi),\xi') - f_z(\mathcal{A}(z,x,\xi'),\xi') + f_z(\mathcal{A}(z,x,\xi'),\xi) - f_z(\mathcal{A}(z,x,\xi),\xi) \\
& \leq L\left[\sqrt{V_d(\mathcal{A}(z,x,\xi'),\mathcal{A}(z,x,\xi))} + \sqrt{V_d(\mathcal{A}(z,x,\xi),\mathcal{A}(z,x,\xi'))}\right],
\end{aligned} \tag{5}
$$

where the second inequality is by **A2**. Then we invoke **A3** to bound both sides by

$$
\begin{aligned}
& (1+\alpha)\gamma V_d(\mathcal{A}(z,x,\xi'),\mathcal{A}(z,x,\xi)) \\
& \leq \gamma[V_d(\mathcal{A}(z,x,\xi'),\mathcal{A}(z,x,\xi)) + V_d(\mathcal{A}(z,x,\xi),\mathcal{A}(z,x,\xi'))] \\
& \leq L\left[\sqrt{V_d(\mathcal{A}(z,x,\xi'),\mathcal{A}(z,x,\xi))} + \sqrt{V_d(\mathcal{A}(z,x,\xi),\mathcal{A}(z,x,\xi'))}\right] \\
& \leq L\left(1+\sqrt{1/\alpha}\right)\sqrt{V_d(\mathcal{A}(z,x,\xi'),\mathcal{A}(z,x,\xi))}
\end{aligned} \tag{6}
$$

and by symmetry, we immediately have the follow two relations

$$
\begin{aligned}
\sqrt{V_d(\mathcal{A}(z,x,\xi'),\mathcal{A}(z,x,\xi))} &\leq \frac{L(1+\sqrt{1/\alpha})}{\gamma(1+\alpha)} \\
\sqrt{V_d(\mathcal{A}(z,x,\xi),\mathcal{A}(z,x,\xi'))} &\leq \frac{L(1+\sqrt{1/\alpha})}{\gamma(1+\alpha)}.
\end{aligned} \tag{7}
$$

Last we plug the relation (7) into *Line 484* to get

$$
\begin{aligned}
& |\mathbb{E}_{\xi'}\{\mathbb{E}_{\xi}[f_z(\mathcal{A}(z,x,\xi'),\xi) - f_z(\mathcal{A}(z,x,\xi'),\xi')]\}| \\
& = \int_{\xi'\sim\Xi}\int_{\xi\sim\Xi} |f_z(\mathcal{A}(z,x,\xi'),\xi) - f_z(\mathcal{A}(z,x,\xi),\xi)| d\mu_\xi d\mu_{\xi'} \\
& \leq \int_{\xi'\sim\Xi}\int_{\xi\sim\Xi} L\cdot\max\left\{\sqrt{V_d(\mathcal{A}(z,x,\xi),\mathcal{A}(z,x,\xi'))}, \sqrt{V_d(\mathcal{A}(z,x,\xi'),\mathcal{A}(z,x,\xi))}\right\} d\mu_\xi d\mu_{\xi'} \\
& \leq \frac{L^2(1+\sqrt{1/\alpha})}{\gamma(1+\alpha)}.
\end{aligned} \tag{8}
$$

Letting $z = x^{k-\tau_k}, x = x^k, \xi' = \xi^{k-\tau_k}, \gamma = \gamma_k$ completes the proof.

Then we are ready to derive a descent property as in **Lemma 1**, which we summarize in **relip-Lemma 1**.

---

relip-Lemma 1 (Lemma 1 under rel.Lip.)

With the above assumptions, if $\rho \geq 2\lambda, \gamma_k \geq \rho$, then

$$
\begin{aligned}
\frac{\rho(\rho-2\lambda)}{\gamma_k-2\lambda} V_d(\hat{x}^k, x^k) &\leq f_{1/\rho}^{V_d}(x^k) - \mathbb{E}_k[f_{1/\rho}^{V_d}(x^{k+1})] + \frac{\rho L^2(1+\sqrt{1/\alpha})}{(\gamma_k-2\lambda)\gamma_k(1+\alpha)} \\
&\quad - \frac{\rho(\gamma_k-\rho)}{2(\gamma_k-2\lambda)}\mathbb{E}_k[\|x^{k+1}-x^k\|^2] + \frac{3\rho\lambda}{2(\gamma_k-2\lambda)}\mathbb{E}_k[\|x^{k+1}-x^{k-\tau_k}\|^2].
\end{aligned}
$$

First we have, by the three-point lemma and the optimality of $\hat{x}^k$, that

$$\begin{aligned} f_{x^{k-\tau_k}}(x^{k+1}, \xi^{k-\tau_k}) + \gamma_k V_d(x^{k+1}, x^k) &\le f_{x^{k-\tau_k}}(\hat{x}^k, \xi^{k-\tau_k}) + \gamma_k V_d(\hat{x}^k, x^k) - \gamma_k V_d(\hat{x}^k, x^{k+1}) \\ f(\hat{x}^k) + \rho V_d(\hat{x}^k, x^k) &\le f(x^{k+1}) + \rho V_d(x^{k+1}, x^k) \end{aligned} \tag{9}$$

Sum the two relations from (9) and take expctation, we have

$$\begin{aligned} &(\gamma_k - \rho)\mathbb{E}_k[V_d(x^{k+1}, x^k)] - (\gamma_k - \rho)V_d(\hat{x}^k, x^k) + \gamma_k \mathbb{E}_k[V_d(\hat{x}^k, x^{k+1})] \\ &\le f_{x^{k-\tau_k}}(\hat{x}^k, \xi^{k-\tau_k}) - f(\hat{x}^k) + \mathbb{E}_k[f(x^{k+1})] - \mathbb{E}_k\left[f_{x^{k-\tau_k}}(x^{k+1}, \xi^{k-\tau_k})\right] \\ &= f_{x^{k-\tau_k}}(\hat{x}^k, \xi^{k-\tau_k}) - f(\hat{x}^k) + \mathbb{E}_k[f(x^{k+1})] - \mathbb{E}_k\left[\mathbb{E}_\xi\left[f_{x^{k-\tau_k}}(x^{k+1}, \xi)\right]\right] \\ &\quad + \mathbb{E}_k\left[\mathbb{E}_\xi\left[f_{x^{k-\tau_k}}(x^{k+1}, \xi)\right]\right] - \mathbb{E}_k\left[f_{x^{k-\tau_k}}(x^{k+1}, \xi^{k-\tau_k})\right] \\ &\le \frac{\lambda}{2}\|x^{k-\tau_k} - \hat{x}^k\|^2 + \frac{\lambda}{2}\mathbb{E}_k[\|x^{k+1} - x^{k-\tau_k}\|^2] + \frac{L^2(1+\sqrt{1/\alpha})}{\gamma_k(1+\alpha)}, \end{aligned} \tag{10}$$

where the second inequality follows from the bound on the RHS in *Line 491* and **relip Lemma 5**. Next we lower-bound the LHS using the 1-strong convexity of kernel d

$$\begin{aligned} &\frac{\gamma_k - \rho}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] - (\gamma_k - \rho)V_d(\hat{x}^k, x^k) + \gamma_k \mathbb{E}_k[V_d(\hat{x}^k, x^{k+1})] \\ &\le (\gamma_k - \rho)\mathbb{E}_k[V_d(x^{k+1}, x^k)] - (\gamma_k - \rho)V_d(\hat{x}^k, x^k) + \gamma_k \mathbb{E}_k[V_d(\hat{x}^k, x^{k+1})]. \end{aligned} \tag{11}$$

Re-arranging the terms, we deduce that

$$\begin{aligned} &\gamma_k \mathbb{E}_k[V_d(\hat{x}^k, x^{k+1})] \\ &\le (\gamma_k - \rho)V_d(\hat{x}^k, x^k) + \frac{\lambda}{2}\mathbb{E}_k[\|x^{k+1} - x^{k-\tau_k}\|^2] + \frac{\lambda}{2}\|x^{k-\tau_k} - \hat{x}^k\|^2 - \frac{\gamma_k - \rho}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \frac{L^2(1+\sqrt{1/\alpha})}{\gamma_k(1+\alpha)} \\ &\le (\gamma_k - \rho)V_d(\hat{x}^k, x^k) + \frac{3\lambda}{2}\mathbb{E}_k[\|x^{k+1} - x^{k-\tau_k}\|^2] + \lambda\mathbb{E}_k[\|x^{k+1} - \hat{x}^k\|^2] - \frac{\gamma_k - \rho}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \frac{L^2(1+\sqrt{1/\alpha})}{\gamma_k(1+\alpha)} \\ &= (\gamma_k - \rho)V_d(\hat{x}^k, x^k) + \frac{3\lambda}{2}\mathbb{E}_k[\|x^{k+1} - x^{k-\tau_k}\|^2] - \frac{\gamma_k - \rho}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \frac{L^2(1+\sqrt{1/\alpha})}{\gamma_k(1+\alpha)} \\ &\quad + \lambda\mathbb{E}_k[\|x^{k+1} - \hat{x}^k\|^2 - 2V_d(\hat{x}^k, x^{k+1})] + 2\lambda\mathbb{E}_k[V_d(\hat{x}^k, x^{k+1})] \\ &\le (\gamma_k - \rho)V_d(\hat{x}^k, x^k) + \frac{3\lambda}{2}\mathbb{E}_k[\|x^{k+1} - x^{k-\tau_k}\|^2] - \frac{\gamma_k - \rho}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \frac{L^2(1+\sqrt{1/\alpha})}{\gamma_k(1+\alpha)} \\ &\quad + 2\lambda\mathbb{E}_k[V_d(\hat{x}^k, x^{k+1})], \end{aligned} \tag{12}$$

where the second inequality follows by Cauchy's inequality $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ and the last inequality follows by $V_d(\hat{x}^k, x^{k+1}) \ge \frac{1}{2}\|x^{k+1} - \hat{x}^k\|^2$. Now we re-arrange the terms and divide both sides by $(\gamma_k - 2\lambda)$ to get

$$\begin{aligned} \mathbb{E}_k[V_d(\hat{x}^k, x^{k+1})] &\le \frac{\gamma_k - \rho}{\gamma_k - 2\lambda}V_d(\hat{x}^k, x^k) + \frac{L^2(1+\sqrt{1/\alpha})}{(\gamma_k - 2\lambda)\gamma_k(1+\alpha)} \\ &\quad - \frac{\gamma_k - \rho}{2(\gamma_k - 2\lambda)}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \frac{3\lambda}{2(\gamma_k - 2\lambda)}\mathbb{E}_k[\|x^{k+1} - x^{k-\tau_k}\|^2] \\ &= V_d(\hat{x}^k, x^k) - \frac{\rho - 2\lambda}{\gamma_k - 2\lambda}V_d(\hat{x}^k, x^k) + \frac{L^2(1+\sqrt{1/\alpha})}{(\gamma_k - 2\lambda)\gamma_k(1+\alpha)} \\ &\quad - \frac{\gamma_k - \rho}{2(\gamma_k - 2\lambda)}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \frac{3\lambda}{2(\gamma_k - 2\lambda)}\mathbb{E}_k[\|x^{k+1} - x^{k-\tau_k}\|^2] \end{aligned} \tag{13}$$

Adopting Bregman Moreau envelop as the potential function and treating delays as error, we successively deduce that

$$
\begin{aligned}
&\mathbb{E}_k[f_{1/\rho}^{V_d}(x^{k+1})] \\
&= \mathbb{E}_k[f(\hat{x}^{k+1}) + \rho V_d(\hat{x}^{k+1}, x^{k+1})] \\
&\leq \mathbb{E}_k[f(\hat{x}^{k}) + \rho V_d(\hat{x}^{k}, x^{k+1})] \\
&\leq \mathbb{E}_k[f(\hat{x}^{k}) + \rho V_d(\hat{x}^{k}, x^{k})] - \frac{\rho(\rho-2\lambda)}{\gamma_k - 2\lambda} V_d(\hat{x}^k, x^k) + \frac{\rho L^2(1+\sqrt{1/\alpha})}{(\gamma_k - 2\lambda)\gamma_k(1+\alpha)} \\
&\quad - \frac{\rho(\gamma_k - \rho)}{2(\gamma_k - 2\lambda)} \mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \frac{3\rho\lambda}{2(\gamma_k - 2\lambda)} \mathbb{E}_k[\|x^{k+1} - x^{k-\tau_k}\|^2] \\
&= f_{1/\rho}^{V_d}(x^k) - \frac{\rho(\rho-2\lambda)}{\gamma_k - 2\lambda} V_d(\hat{x}^k, x^k) + \frac{\rho L^2(1+\sqrt{1/\alpha})}{(\gamma_k - 2\lambda)\gamma_k(1+\alpha)} \\
&\quad - \frac{\rho(\gamma_k - \rho)}{2(\gamma_k - 2\lambda)} \mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \frac{3\rho\lambda}{2(\gamma_k - 2\lambda)} \mathbb{E}_k[\|x^{k+1} - x^{k-\tau_k}\|^2].
\end{aligned} \tag{14}
$$

A simple re-arrangment completes the proof.

With **relip-Lemma 1** in hand, telescoping over k and using, **A4** (*Line 172*), **Lemma 2** (*Line 174*) to bound $\sum_k \mathbb{E}_k[\|x^{k+1} - x^{k-\tau_k}\|^2]$, we immediately derive the same $\mathcal{O}\left(\frac{1}{\sqrt{K}} + \frac{\tau^2}{K}\right)$ rate in terms of the Bregman stationary measure $\mathbb{E}[V_d(\hat{x}^{k^*}, x^{k^*})]$.

---

relip-Theorem (Informal) (Convergence under rel.Lip.)

Let $\gamma_k \equiv \gamma \sim \mathcal{O}\left(\sqrt{K}\right)$ and k^* be an index chosen between 1 and K uniformly, then

$$
\mathbb{E}[V_d(\hat{x}^{k^*}, x^{k^*})] = \mathcal{O}\left(\frac{1}{\sqrt{K}} + \frac{\tau^2}{K}\right).
$$

Sum the relation in **relip-Lemma 1** from $k = 1, \dots, K$, take $\gamma_k = \gamma > 2\lambda + \rho$ and divide both sides by $\frac{\rho(\rho-2\lambda)K}{(\gamma-2\lambda)}$, we have

$$
\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K} V_d(\hat{x}^k, x^k) \leq{}& \frac{f_{1/\rho}^{V_d}(x^1) - \mathbb{E}_k[f_{1/\rho}^{V_d}(x^{K+1})]}{\rho(\rho-2\lambda)} \cdot \frac{\gamma - 2\lambda}{K} + \frac{L^2(1+\sqrt{1/\alpha})}{(\rho-2\lambda)(1+\alpha)\gamma} \\
&- \frac{\gamma}{2(\rho-2\lambda)K}\sum_{k=1}^{K}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \frac{3\lambda}{2(\rho-2\lambda)K}\sum_{k=1}^{K}\mathbb{E}_k[\|x^{k+1} - x^{k-\tau_k}\|^2] \\
\leq{}& \frac{D}{\rho(\rho-2\lambda)} \cdot \frac{\gamma-2\lambda}{K} + \frac{L^2(1+\sqrt{1/\alpha})}{(\rho-2\lambda)(1+\alpha)\gamma} + \frac{3\lambda\tau^2}{2(\rho-2\lambda)K}\sum_{k=1}^{K}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] \\
\leq{}& \frac{D}{\rho(\rho-2\lambda)} \cdot \frac{\gamma-2\lambda}{K} + \frac{L^2(1+\sqrt{1/\alpha})}{(\rho-2\lambda)(1+\alpha)\gamma} + \frac{3\lambda L^2\tau^2}{(\rho-2\lambda)\gamma^2},
\end{aligned} \tag{15}
$$

where the second inequality uses *Line 511*. For the last inequality, consider

$$
f_{x^{k-\tau_k}}(x^{k+1}, \xi^{k-\tau_k}) + \gamma V_d(x^{k+1}, x^k) \leq f_{x^{k-\tau_k}}(x^k, \xi^{k-\tau_k}),
$$

and re-arrangement gives $\gamma V_d(x^{k+1}, x^k) \leq f_{x^{k-\tau_k}}(x^k, \xi^{k-\tau_k}) - f_{x^{k-\tau_k}}(x^{k+1}, \xi^{k-\tau_k}) \leq L\sqrt{V_d(x^{k+1}, x^k)}$, or

$$
\sqrt{V_d(x^{k+1}, x^k)} \leq L/\gamma.
$$

Squaring both sides and using 1-strong convexity, we have

$$
\|x^{k+1} - x^k\|^2 \leq 2V_d(x^{k+1}, x^k) \leq 2L^2/\gamma^2
$$

Plugging the bound back and letting $\gamma \sim \mathcal{O}\left(\sqrt{K}\right)$, we complete the proof.

3. References

[ZH18] Zhang, Siqi, and Niao He. "On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization." *arXiv preprint arXiv:1806.04781* (2018).

[MOPS19] Mukkamala, Mahesh Chandra, et al. "Convex-concave backtracking for inertial Bregman proximal gradient algorithms in nonconvex optimization." *SIAM Journal on Mathematics of Data Science* 2.3 (2020): 658-682.

[Lu19] Lu, Haihao. "“relative continuity” for non-lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent." *INFORMS Journal on Optimization* 1.4 (2019): 288-303.

[ZCZL22] Zhao, Lei, et al. "Randomized Coordinate Subgradient Method for Nonsmooth Optimization." *arXiv preprint arXiv:2206.14981*(2022).