Algorithm 1 Two-Timescale GDA

Require: Initial values (x_0, y_0), step sizes (η_x, η_y)

for $t = 1, 2, \ldots, T$ **do**
  $x_t \leftarrow x_{t-1} - \eta_x \nabla_x f(x_{t-1}, y_{t-1})$
  $y_t \leftarrow \Pi_{\mathcal{Y}} \left(y_{t-1} + \eta_y \nabla_y f(x_{t-1}, y_{t-1})\right)$
end for
Draw $\hat{x}$ uniformly at random from $\{x_t\}_{t=1}^T$
return $\hat{x}$

Algorithm 2 ALSO

Require: Initial values (x_0, y_0), y_{reg}, step sizes (η_x, η_y), $\beta \in [0, 1]$

for $t = 1, 2, \ldots, T$ **do**
  $x_t \leftarrow x_{t-1} - \eta_x d_x^{t-1}$
  $\tilde{y}_{t-1} \leftarrow \beta y_{t-1} + (1 - \beta) y_{\text{reg}}$
  $y_t \leftarrow \Pi_{\mathcal{Y}} \left(\tilde{y}_{t-1} + \eta_y g_y^{t-1}\right)$
end for
Draw $\hat{x}$ uniformly at random from $\{x_t\}_{t=1}^T$
return $\hat{x}$

Let $f(\theta, \pi)$ be a differentiable function, where $\theta \in \Theta \subseteq \mathbb{R}^m$ and $\pi \in \Delta \subseteq \mathbb{R}^d$.

Definition 0.1 (Stochastic Gradient Updates)**.** The algorithm performs stochastic gradient updates with respect to both θ and π, using unbiased stochastic estimates of the gradients $\nabla_\theta f(\theta, \pi)$ and $\nabla_\pi f(\theta, \pi)$.

Assumption 0.2 (Unbiasedness and Bounded Variance of Stochastic Gradients)**.** Let g^θ and g^π denote the stochastic gradient estimators for $\nabla_\theta f(\theta, \pi)$ and $\nabla_\pi f(\theta, \pi)$, respectively. We assume:

$$\mathbb{E}[g^\theta] = \nabla_\theta f(\theta, \pi),$$
$$\mathbb{E}[g^\pi] = \nabla_\pi f(\theta, \pi),$$

and there exist constant $\sigma^2 \geq 0$ such that

$$\mathbb{E}\left[\|g^\theta - \nabla_\theta f(\theta, \pi)\|^2\right] \leq \sigma^2,$$
$$\mathbb{E}\left[\|g^\pi - \nabla_\pi f(\theta, \pi)\|^2\right] \leq \sigma^2.$$

Definition 0.3. A point x is an ε-stationary point ($\varepsilon \geq 0$) of a differentiable function Φ if

$$\|\nabla \Phi(x)\| \leq \varepsilon.$$

If $\varepsilon = 0$, then x is a stationary point.

We use following function:

$$\Phi(\cdot) = \max_{\pi \in \Delta} f(\cdot, \pi)$$

is differentiable in that setting. In contrast, the function Φ is not necessarily differentiable for a general nonconvex-concave minimax problem even if f is Lipschitz and smooth. A weaker condition that we make use of is the following.

Definition 0.4. A function Φ is ℓ-weakly convex if the function

$$\Phi(\cdot) + \frac{\ell}{2}\|\cdot\|^2 \tag{1}$$

is convex.

Definition 0.5 (Moreau Envelope)**.** A function $\Phi_\lambda : \mathbb{R}^m \to \mathbb{R}$ is the *Moreau envelope* of Φ with a positive parameter $\lambda > 0$ if

$$\Phi_\lambda(\theta) = \min_{w \in \mathbb{R}^m} \left\{ \Phi(w) + \frac{1}{2\lambda}\|w - \theta\|^2 \right\}, \quad \text{for each } \theta \in \mathbb{R}^m.$$

Lemma 0.6. *If f is ℓ-smooth and Δ is bounded, the Moreau envelope $\Phi_{1/2\ell}$ of $\Phi(\theta) = \max_{\pi\in\Delta} f(\theta, \pi)$ is differentiable with*

$$\nabla\Phi_{1/2\ell}(\theta) = 2\ell \left(\theta - \mathrm{prox}_{\Phi/2\ell}(\theta)\right).$$

An alternative measure of approximate stationarity of $\Phi(\theta) = \max_{\pi\in\Delta} f(\theta, \pi)$ is to require

$$\|\nabla\Phi_{1/2\ell}(\theta)\| \leq \varepsilon.$$

Definition 0.7 (ε-Stationary Point)**.** A point $\theta \in \mathbb{R}^m$ is an ε-stationary point of an ℓ-weakly convex function Φ if

$$\|\nabla\Phi_{1/2\ell}(\theta)\| \leq \varepsilon.$$

If $\varepsilon = 0$, then x is a stationary point.

Lemma 0.8. *If θ is an ε-stationary point of an ℓ-weakly convex function Φ, then there exists $\hat{\theta} \in \mathbb{R}^m$ such that*

$$\min_{\xi\in\partial\Phi(\hat{\theta})} \|\xi\| \leq \varepsilon, \quad \text{and} \quad \|\theta - \hat{\theta}\| \leq \frac{\varepsilon}{2\ell}.$$

To rigorously analyze the convergence of our method, we now introduce a set of assumptions that define the regularity conditions of the function $f(\theta, \pi)$ and the feasible sets. These assumptions mirror those in [**?**] but are adapted to account for the use of Adam kingma2014adam and regularization in `ALSO`.

Assumption 0.9. The objective function and the constraint set, $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, $\Delta \subset \mathbb{R}^n$, satisfy the following:

1. f is ℓ-smooth; for each $\pi \in \Delta$, the function $f(\cdot, \pi)$ is L-Lipschitz; for each $\theta \in \mathbb{R}^m$, the function $f(\theta, \cdot)$ is concave.
2. Δ is a convex and bounded set with respect to the Kullback–Leibler (KL) divergence, having diameter at most $D^2 \geq 0$:

$$\sup_{p,q\in\Delta} D_{\mathrm{KL}}(p \,\|\, q) \leq D^2.$$

Since $f(\theta, \cdot)$ is concave for each $\theta \in \mathbb{R}^m$, the function $\Phi(\cdot) = \max_{\pi\in\Delta} f(\cdot, \pi)$ may not be differentiable. Fortunately, the following structural lemma shows that Φ is ℓ-weakly convex and L-Lipschitz.

Lemma 0.10. *Under Assumption 0.9, the function $\Phi(\cdot) = \max_{\pi\in\Delta} f(\cdot, \pi)$ is ℓ-weakly convex and L-Lipschitz, with*

$$\nabla_\theta f(\cdot, \pi^\star(\cdot)) \in \partial\Phi(\cdot),$$

where $\pi^\star(\cdot) \in \arg\max_{\pi\in\Delta} f(\cdot, \pi)$.

Since Φ is ℓ-weakly convex, the notion of stationarity in Definition 0.7 is our target, given only access to the (stochastic) gradient of f.

Denoting

$$\Delta_\Phi = \Phi_{1/2\ell}(\theta_0) - \min_\theta \Phi_{1/2\ell}(\theta), \quad \Delta_0 = \Phi(\theta_0) - f(\theta_0, \pi_0),$$

we now present complexity results for the Algorithm **??**.

With this setup, we adopt the convergence criterion:

$$\mathbb{E}\left[\|\nabla\Phi_{1/2\ell}(x)\|\right] \leq \varepsilon,$$

and present the following main theorem, which establishes the complexity bounds of Algorithm **??** under our assumptions.

Theorem 0.11 (Main)**.** *Under assumptions 0.9, and letting the step sizes $\eta_x > 0$ and $\eta_y > 0$ be chosen as*

$$\eta_x = \min\left\{\frac{\epsilon^2}{4G_2}, \frac{\epsilon^4}{128G_3^2G_4D^2\ell}, \frac{\epsilon^6}{128G_3^3G_4D^2\sigma^2}\right\}, \quad \eta_y = \min\left\{\frac{1}{2\ell}, \frac{\epsilon^2}{2G_3\sigma^2}\right\}.$$

with a batch size $M = 1$, $c_m = \frac{1}{2}$ and start scaling factor $b_0 = \sqrt{2(\sigma^2 + L^2)}$, regularization factor $\beta = 0$ (without regularization), the iteration complexity of `ALSO` *algorithm* **??** *to return an ε-stationary point is bounded by*

$$\mathcal{O}\left(\left[\frac{[\beta_1/(1-\beta_1)^3]\ell(\sigma^2+L^2)\Delta_{\hat{\Phi}}}{\epsilon^4} + \frac{[1/(1-\beta_1)]\ell\Delta_0}{\epsilon^2}\right] \cdot \max\left\{1, \frac{[1/\beta_1]\ell^2D^2}{\epsilon^2}, \frac{[1/(\beta_1-\beta_1^2)]\ell^2D^2\sigma^2}{\epsilon^4}\right\}\right), \tag{2}$$

which is also the total gradient complexity of the algorithm.

Where

$$G_1 = \frac{4}{(1-\beta_1)}\sqrt{2(\sigma^2+L^2)},$$

$$G_2 = \frac{4\ell}{(1-\beta_1)}\sqrt{2(\sigma^2+L^2)}\frac{1}{c_m^2b_0^2}2(\sigma^2+L^2)[1+4\frac{\beta_1}{c_mb_0(1-\beta_1)}\sqrt{2(\sigma^2+L^2)}],$$

$$G_3 = 8\ell\frac{1}{c_mb_0\,(1-\beta_1)}\sqrt{2(\sigma^2+L^2)},$$

$$G_4 = \frac{1}{c_mb_0}L\sqrt{\sigma^2+L^2}$$

Discussion. This convergence result matches the convergence guarantees established for the standard SGDA method in [**?**], highlighting the theoretical soundness of our `ALSO` algorithm in the Euclidean setting.

Definition 0.12 (Moreau Envelope)**.** A function $\Phi_\lambda : \mathbb{R}^m \to \mathbb{R}$ is the *Moreau envelope* of Φ with a positive parameter $\lambda > 0$ if

$$\Phi_\lambda(x) = \min_w\left\{\Phi(w) + \frac{1}{2\lambda}\|w-x\|^2\right\}, \quad \text{for each } x \in \mathbb{R}^m.$$

Lemma 0.13. *If f is ℓ-smooth and $\mathcal{Y}$ is bounded, the Moreau envelope $\Phi_{1/2\ell}$ of $\Phi(x) = \max_{y\in\mathcal{Y}} f(x,y)$ is differentiable with*

$$\nabla\Phi_{1/2\ell}(x) = 2\ell\left(x - \mathrm{prox}_{\Phi/2\ell}(x)\right).$$

An alternative measure of approximate stationarity of $\Phi(x) = \max_{y\in\mathcal{Y}} f(x,y)$ is to require that

$$\|\nabla\Phi_{1/2\ell}(x)\| \le \varepsilon.$$

Definition 0.14 (ε-Stationary Point)**.** A point $x \in \mathbb{R}^m$ is an ε-stationary point of an ℓ-weakly convex function Φ if

$$\|\nabla\Phi_{1/2\ell}(x)\| \le \varepsilon.$$

If $\varepsilon = 0$, then x is a stationary point.

Lemma 0.15. *If x is an ε-stationary point of an ℓ-weakly convex function Φ, then there exists $\hat{x} \in \mathbb{R}^m$ such that*

$$\min_{\xi\in\partial\Phi(\hat{x})}\|\xi\| \le \varepsilon, \quad \textit{and} \quad \|x-\hat{x}\| \le \frac{\varepsilon}{2\ell}.$$

Lemma 0.16 (ALSO regularization by y_k)**.** *Step of ALSO by y with regularization y_{reg} and set Y and stepsize η_y equiv step of SGDA, with $\hat{\eta}_y = \frac{\eta_y}{\beta}$ and $\hat{Y} = \frac{Y-(1-\beta)y_{reg}}{\beta}$ and after that $y_t = \beta\hat{y}_t + (1-\beta)y_{reg}$.*

ALSO step:

$$y_t \leftarrow \Pi_{\mathcal{Y}}\left(\beta y_{t-1} + (1-\beta)y_{reg} + \eta_y \nabla_y f(x_{t-1}, y_{t-1})\right) \tag{3}$$

SGDA step:

$$\hat{y_t} \leftarrow \Pi_{\hat{\mathcal{Y}}}\left(y_{t-1} + \hat{\eta}_y \nabla_y f(x_{t-1}, y_{t-1})\right) \tag{4}$$

$$y_t = \beta\hat{y}_t + (1-\beta)y_{reg} \tag{5}$$

Proof.

$$y_t \leftarrow \Pi_{\mathcal{Y}}\left(\beta y_{t-1} + (1-\beta)y_{reg} + \eta_y \nabla_y f(x_{t-1}, y_{t-1})\right)$$

$$y_t = \arg\min_{y\in Y} \|y - \beta y_{t-1} - (1-\beta)y_{reg} - \eta_y \nabla_y f(x_{t-1}, y_{t-1})\|^2$$

$$y_t = \arg\min_{y\in Y} \|\frac{y-(1-\beta)y_{reg}}{\beta} - y_{t-1} - \frac{\eta_y}{\beta}\nabla_y f(x_{t-1}, y_{t-1})\|^2$$

$$y_t = (1-\beta)y_{reg} + \beta\left[\arg\min_{y\in\frac{Y-(1-\beta)y_{reg}}{\beta}} \|y - y_{t-1} - \frac{\eta_y}{\beta}\nabla_y f(x_{t-1}, y_{t-1})\|^2\right]$$

□

Lemma 0.17 (Properties of Adam Estimator updates)**.** *Let g_x^t denote the stochastic gradient of f with respect to x at iteration t.*

We define two independent stochastic gradient samples: - g_x^t: used for the numerator (first moment), - $\tilde{g}_x^t$: used for the denominator (second moment).

The first and second moment estimates of Adam are:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)g_x^t,$$
$$b_t^2 = \beta_2 b_{t-1}^2 + (1-\beta_2)\|\tilde{g}_x^t\|^2.$$

Then the Adam Estimator update direction is given by:

$$d_x^t = \frac{m_t}{b_t}.$$

Moreover, the first moment m_t admits a closed-form expression:

$$m_t = (1-\beta_1)\sum_{k=0}^{t} \beta_1^{t-k} g_x^k.$$

$$m_{-1} = 0; \quad b_{-1}, b_0 > 0$$

Assume $\|g_x^t\|^2 \geq b_r^2$ for some reference step $r \leq t$, and let $\beta_2 = 1 - \frac{1}{K}$ for some $K > 0$. Then:

$$b_t^2 \geq \beta_2^{t-r} b_r^2 = \left(1 - \frac{1}{K}\right)^{t-r} b_r^2 \geq \left(1 - \frac{1}{K}\right)^{K} b_r^2 \geq \frac{1}{4}b_r^2.$$

Lemma 0.18 (Lemma about summ a_t)**.** *Let $a_t = \langle \hat{x}_t - x_t, d_x^t\rangle$ and $\xi_t = \langle \hat{x}_t - x_t, g_x^t\rangle$, where d_x^t – Adam Estimator step and g_x^t – stochastic gradient for momentum in Adam Estimator Then:*

$$\sum_{t=0}^{T} a_t \leq \sum_{k=0}^{T} C_k\xi_k + 2\eta_x \sum_{k=0}^{T-1} A_k\|d_x^k\|^2$$

where:

$$C_k = (1-\beta_1)\sum_{t=k}^{T}\frac{\beta_1^{t-k}}{b_t}; \quad A_k = b_k \sum_{t=k+1}^{T}\frac{\beta_1^{t-k}}{b_t}$$

Proof.

$$a_t = \frac{1}{b_t}\left((1-\beta_1)\xi_t + \langle \hat{x}_t - x_t, \beta_1 m_x^{t-1}\rangle\right)$$

$$a_t = \frac{1}{b_t}\left((1-\beta_1)\xi_t + \langle (\hat{x}_{t-1} - x_{t-1}) + (\hat{x}_t - \hat{x}_{t-1}) - (x_t - x_{t-1}), \beta_1 m_x^{t-1}\rangle\right)$$

$$a_t = \frac{1}{b_t}\left((1-\beta_1)\xi_t + \beta_1 b_{t-1} a_{t-1} + \langle (\hat{x}_t - \hat{x}_{t-1}) - (x_t - x_{t-1}), \beta_1 m_x^{t-1}\rangle\right)$$

$$\begin{aligned}\langle (\hat{x}_t - \hat{x}_{t-1}) - (x_t - x_{t-1}), \beta_1 m_x^{t-1}\rangle &\leq \beta_1(\|\hat{x}_t - \hat{x}_{t-1}\| + \|x_t - x_{t-1}\|)\|m_x^{t-1}\| \\ &\leq 2\beta_1 \|x_t - x_{t-1}\|\|m_x^{t-1}\| = 2\beta_1 \eta_x b_{t-1}\|d_x^{t-1}\|^2\end{aligned}$$

$$a_t \leq \frac{1}{b_t}(1-\beta_1)\xi_t + \beta_1 \frac{b_{t-1}}{b_t} a_{t-1} + 2\beta_1 \frac{b_{t-1}}{b_t}\eta_x \|d_x^{t-1}\|^2$$

$$\begin{aligned}a_t \leq \frac{1}{b_t}(1-\beta_1)\,\xi_t &+ \frac{\beta_1(1-\beta_1)}{b_t}\xi_{t-1} + \beta_1^2 \frac{b_{t-2}}{b_t} a_{t-2} \\ &+ 2\beta_1 \frac{b_{t-1}}{b_t}\eta_x\|d_x^{t-1}\|^2 + 2\beta_1^2 \frac{b_{t-2}}{b_t}\eta_x \|d_x^{t-2}\|^2\end{aligned}$$

$$\boxed{a_t \leq \frac{1}{b_t}\sum_{k=0}^{t}(1-\beta_1)\,\beta_1^{t-k}\xi_k \;+\; 2\eta_x \sum_{k=0}^{t-1}\beta_1^{t-k}\frac{b_k}{b_t}\|d_x^k\|^2}$$

$$\boxed{a_t \leq \frac{1}{b_t}\sum_{k=0}^{t}(1-\beta_1)\,\beta_1^{t-k}\xi_k \;+\; 2\eta_x \sum_{k=0}^{t-1}\beta_1^{t-k}\frac{1}{b_t b_k}\|m_x^k\|^2}$$

We start from:

$$a_t \leq \frac{1}{b_t}\sum_{k=0}^{t}(1-\beta_1)\beta_1^{t-k}\,\xi_k \;+\; 2\eta_x \sum_{k=0}^{t-1}\frac{\beta_1^{t-k} b_k}{b_t}\|d_x^k\|^2$$

Summing over $t = 0$ to T:

$$\sum_{t=0}^{T} a_t \leq \sum_{t=0}^{T}\frac{1}{b_t}\sum_{k=0}^{t}(1-\beta_1)\beta_1^{t-k}\,\xi_k \;+\; 2\eta_x \sum_{t=0}^{T}\sum_{k=0}^{t-1}\frac{\beta_1^{t-k} b_k}{b_t}\|d_x^k\|^2$$

Switching the order of sums in the second term:

$$= \sum_{t=0}^{T}\frac{1}{b_t}\sum_{k=0}^{t}(1-\beta_1)\beta_1^{t-k}\,\xi_k \;+\; 2\eta_x \sum_{k=0}^{T-1} b_k\|d_x^k\|^2 \sum_{t=k+1}^{T}\frac{\beta_1^{t-k}}{b_t}$$

Thus, the compact form is:

$$\boxed{\sum_{t=0}^{T} a_t \leq \sum_{t=0}^{T}\frac{1}{b_t}\sum_{k=0}^{t}(1-\beta_1)\beta_1^{t-k}\,\xi_k \;+\; 2\eta_x \sum_{k=0}^{T-1} b_k\|d_x^k\|^2 \sum_{t=k+1}^{T}\frac{\beta_1^{t-k}}{b_t}}$$

Thus, the overall summed inequality becomes:

$$\sum_{t=0}^{T} a_t \leq \sum_{k=0}^{T} C_k \xi_k + 2\eta_x \sum_{k=0}^{T-1} A_k \|d_x^k\|^2$$

where:

$$C_k = (1-\beta_1)\sum_{t=k}^{T}\frac{\beta_1^{t-k}}{b_t}; \quad A_k = b_k \sum_{t=k+1}^{T}\frac{\beta_1^{t-k}}{b_t}$$

□

Lemma 0.19 (D3). *For ALSO (TODO assumptions) the folowing statement holds true:*

$$\hat{C}_T \sum_{t=0}^{T} \mathbb{E}\left[\left\|\nabla \Phi_{1/2\ell}(x_t)\right\|^2\right] \leq \frac{4}{\eta_x}\hat{\Delta}_\Phi + 4\ell\eta_x \sum_{t=0}^{T} \mathbb{E}\left[\|d_x^t\|^2(1+4A_t)\right] + 8\ell\frac{1}{c_m b_0}\sum_{t=0}^{T}\Delta_t$$

where $\hat{C}_T := (1-\beta_1)\min_{i\in\{0,\dots,T\}}\left\{\frac{1}{\mathbb{E}[b_i]}\right\}$ *and* $\Delta_t = \mathbb{E}\left[\Phi(x_t) - f(x_t, y_t)\right]$ *and* $\hat{\Delta}_\Phi :=$ $\Phi_{1/2\ell}(x_0) - \min_{x\in\mathbb{R}^m}\Phi_{1/2\ell}(x)$ *and* $A_k = b_k\sum_{t=k+1}^{T}\frac{\beta_1^{t-k}}{b_t}$.

Proof. Let $\hat{x}_{t-1} = \text{prox}_{\Phi/2\ell}(x_{t-1})$, we have

$$\Phi_{1/2\ell}(x_t) \leq \Phi(\hat{x}_{t-1}) + \ell\|\hat{x}_{t-1} - x_t\|^2.$$

Let $x_t = x_{t-1} - \eta_x d_x^{t-1}$, where d_x^{t-1} is the update direction given by Adam, and let $\hat{x}_{t-1} = \text{prox}_{\Phi/2\ell}(x_{t-1})$. Then:

$$\begin{aligned}
\|\hat{x}_{t-1} - x_t\|^2 &= \left\|\hat{x}_{t-1} - \left(x_{t-1} - \eta_x d_x^{t-1}\right)\right\|^2 \\
&= \left\|(\hat{x}_{t-1} - x_{t-1}) + \eta_x d_x^{t-1}\right\|^2 \\
&= \|\hat{x}_{t-1} - x_{t-1}\|^2 + \eta_x^2\|d_x^{t-1}\|^2 + 2\eta_x\left\langle \hat{x}_{t-1} - x_{t-1}, d_x^{t-1}\right\rangle
\end{aligned}$$

Define:

$$a_t = \langle \hat{x}_t - x_t, d_x^t\rangle = \frac{1}{b_t}\langle \hat{x}_t - x_t, m_x^t\rangle; \quad \xi_t = \langle \hat{x}_t - x_t, g_x^t\rangle$$

Then:

$$\Phi_{1/2\ell}(x_t) \leq \Phi_{1/2\ell}(x_{t-1}) + \ell\eta_x^2\|d_x^{t-1}\|^2 + 2\ell\eta_x a_{t-1}.$$

By summing inequalties we have:

$$\Phi_{1/2\ell}(x_{T+1}) \leq \Phi_{1/2\ell}(x_0) + \ell\eta_x^2\sum_{t=0}^{T}\|d_x^t\|^2 + 2\ell\eta_x\sum_{t=0}^{T}a_t.$$

Use lemma (TODO):

$$\sum_{t=0}^{T}a_t \leq \sum_{t=0}^{T}C_t\xi_t + 2\eta_x\sum_{t=0}^{T-1}A_t\|d_x^t\|^2$$

where:

$$C_k = (1-\beta_1)\sum_{t=k}^{T}\frac{\beta_1^{t-k}}{b_t}; \quad A_k = b_k\sum_{t=k+1}^{T}\frac{\beta_1^{t-k}}{b_t}$$

$$0 \leq \Phi_{1/2\ell}(x_0) - \Phi_{1/2\ell}(x_{T+1}) + \ell\eta_x^2\sum_{t=0}^{T}\|d_x^t\|^2 + 2\ell\eta_x\left[\sum_{t=0}^{T}C_t\xi_t + 2\eta_x\sum_{t=0}^{T-1}A_t\|d_x^t\|^2\right].$$

$$0 \leq \Phi_{1/2\ell}(x_0) - \Phi_{1/2\ell}(x_{T+1}) + \ell\eta_x^2\sum_{t=0}^{T}\|d_x^t\|^2(1+4A_t) + 2\ell\eta_x\sum_{t=0}^{T}C_t\xi_t.$$

$$\begin{aligned}
\xi_t = \langle \hat{x}_t - x_t, g_x^t\rangle = \langle \hat{x}_t - x_t, \nabla_x f(x_t, y_t)\rangle + \langle \hat{x}_t - x_t, g_x^t - \nabla_x f(x_t, y_t)\rangle \\
r_t \equiv \langle \hat{x}_t - x_t, g_x^t - \nabla_x f(x_t, y_t)\rangle
\end{aligned}$$

Use ℓ – smoothness of f:

$$\langle \hat{x}_t - x_t, \nabla_x f(x_t, y_t)\rangle \leq f(\hat{x}_t, y_t) - f(x_t, y_t) + \frac{\ell}{2}\|\hat{x}_t - x_t\|^2$$

By definition of Φ and $\Phi_{1/2\ell}$:

$$\Phi_{1/2\ell}(x_t) = \min_x \left\{\Phi(x) + \ell\|x - x_t\|^2\right\} = \Phi(\hat{x}_t) + \ell\|\hat{x}_t - x_t\|^2 \leq \Phi(x_t)$$
$$f(\hat{x}_t, y_t) - f(x_t, y_t) \leq \Phi(\hat{x}_t) - f(x_t, y_t) \leq \Phi(x_t) - f(x_t, y_t) - \ell\|\hat{x}_t - x_t\|^2$$
$$\langle \hat{x}_t - x_t, \nabla_x f(x_t, y_t)\rangle \leq \Phi(x_t) - f(x_t, y_t) - \frac{\ell}{2}\|\hat{x}_t - x_t\|^2$$

From (TODO) we know, that $\|\hat{x}_t - x_t\| = \|\nabla\Phi_{1/2\ell}(x_t)\|/2\ell$.

$$\xi_t \leq \Phi(x_t) - f(x_t, y_t) - \frac{1}{8\ell}\|\nabla\Phi_{1/2\ell}(x_t)\|^2 + r_t$$

$$0 \leq \Phi_{1/2\ell}(x_0) - \Phi_{1/2\ell}(x_{T+1}) + \ell\eta_x^2 \sum_{t=0}^{T} \|d_x^t\|^2 \left(1 + 4A_t\right) + 2\ell\eta_x \sum_{t=0}^{T} C_t \left[\Phi(x_t) - f(x_t, y_t) - \frac{1}{8\ell}\|\nabla\Phi_{1/2\ell}(x_t)\|^2 + r_t\right]$$

$$\sum_{t=0}^{T} C_t\|\nabla\Phi_{1/2\ell}(x_t)\|^2 \leq \frac{4}{\eta_x}\left(\Phi_{1/2\ell}(x_0) - \Phi_{1/2\ell}(x_{T+1})\right) + 4\ell\eta_x \sum_{t=0}^{T} \|d_x^t\|^2(1 + 4A_t) + 8\ell\sum_{t=0}^{T} C_t\left[\Phi(x_t) - f(x_t, y_t) + r_t\right]$$

Using the fact that the variables x_t and $\left\{\frac{1}{b_k}\right\}_{k=t}^{T}$ are independent (because we use Adam Estimator with independent g_x^k and $\tilde{g}_x^k$), we can conclude that C_t is also independent of x_t. Therefore, we can split the expectation of the product into the product of expectations. Since g_x^k – unbiased stochastic operator then $\mathbb{E}r_t = 0$ and the inequality becomes:

$$\sum_{t=0}^{T} \mathbb{E}[C_t] \cdot \mathbb{E}\left[\left\|\nabla\Phi_{1/2\ell}(x_t)\right\|^2\right] \leq \frac{4}{\eta_x}\left(\Phi_{1/2\ell}(x_0) - \mathbb{E}\Phi_{1/2\ell}(x_{T+1})\right) + 4\ell\eta_x \sum_{t=0}^{T} \mathbb{E}\left[\|d_x^t\|^2(1 + 4A_t)\right]$$
$$+ 8\ell\sum_{t=0}^{T} \mathbb{E}[C_t] \cdot \mathbb{E}\left[\Phi(x_t) - f(x_t, y_t)\right]$$

Note, that C_t are upper and lower bounded:

$$C_t = (1 - \beta_1)\sum_{k=t}^{T} \frac{\beta_1^{k-t}}{b_k}$$
$$C_t \leq (1 - \beta_1)\frac{1}{c_m b_0}\sum_{k=t}^{T} \beta_1^{k-t} \leq \frac{1}{c_m b_0}$$
$$\mathbb{E}[C_t] \geq (1 - \beta_1) \min_{i\in\{1,\ldots,T\}} \left\{\frac{1}{\mathbb{E}[b_i]}\right\}$$

Also define (TODO assumption):

$$\Phi_{1/2\ell}(x_0) - \mathbb{E}\Phi_{1/2\ell}(x_{T+1}) \leq \hat{\Delta}_\Phi := \Phi_{1/2\ell}(x_0) - \min_{x\in\mathbb{R}^m} \Phi_{1/2\ell}(x)$$

We obtain the following bound:

$$\hat{C}_T \sum_{t=0}^{T} \mathbb{E}\left[\left\|\nabla\Phi_{1/2\ell}(x_t)\right\|^2\right] \leq \frac{4}{\eta_x}\hat{\Delta}_\Phi + 4\ell\eta_x \sum_{t=0}^{T} \mathbb{E}\left[\|d_x^t\|^2(1 + 4A_t)\right] + 8\ell\frac{1}{c_m b_0}\sum_{t=0}^{T} \Delta_t$$

where $\hat{C}_T := (1 - \beta_1)\min_{i\in\{0,\ldots,T\}} \left\{\frac{1}{\mathbb{E}[b_i]}\right\}$ and $\Delta_t = \mathbb{E}\left[\Phi(x_t) - f(x_t, y_t)\right]$.

□

Lemma 0.20 (D4)**.** *For ALSO, let*

$$\Delta_t = \mathbb{E}\left[\Phi(x_t) - f(x_t, y_t)\right],$$

the following statement holds true for all $s \leq t-1$,

$$\Delta_{t-1} \leq \eta_x L(\|d_x^{t-1}\| + 2\sum_{k=s}^{t-2}\|d_x^k\|) + \frac{1}{2\eta_y}\mathbb{E}\left[\|y_{t-1} - y^\star(x_s)\|^2 - \|y_t - y^\star(x_s)\|^2\right] + \mathbb{E}\left(f(x_t, y_t) - f(x_{t-1}, y_{t-1})\right) + \frac{\eta_y\sigma^2}{2}.$$

Proof. For any $y \in Y$, the convexity of Y and the update of y_t imply that

$$(y - y_t)^\top(y_t - y_{t-1} - \eta_y G_y(x_{t-1}, y_{t-1}, \xi) \geq 0.$$

$$\begin{aligned}\|y - y_t\|^2 \leq 2\eta_y(y_{t-1} - y)^\top G_y(x_{t-1}, y_{t-1}, \xi) + 2\eta_y(y_t - y_{t-1})^\top \nabla_y f(x_{t-1}, y_{t-1}) \\ +2\eta_y(y_t - y_{t-1})^\top \left(G_y(x_{t-1}, y_{t-1}, \xi) - \nabla_y f(x_{t-1}, y_{t-1})\right) \\ +\|y - y_{t-1}\|^2 - \|y_t - y_{t-1}\|^2.\end{aligned}$$

Using Young's inequality, we have

$$\eta_y(y_t - y_{t-1})^\top \left(G_y(x_{t-1}, y_{t-1}, \xi) - \nabla_y f(x_{t-1}, y_{t-1})\right) \leq \frac{\|y_t - y_{t-1}\|^2}{4} + \eta_y^2\|G_y(x_{t-1}, y_{t-1}, \xi) - \nabla_y f(x_{t-1}, y_{t-1})\|^2.$$

Taking the expectation of both sides, conditioned on (x_{t-1}, y_{t-1}):

$$\begin{aligned}\mathbb{E}[\|y - y_t\|^2 \mid x_{t-1}, y_{t-1}] \leq 2\eta_y(y_{t-1} - y)^\top \nabla_y f(x_{t-1}, y_{t-1}) + 2\eta_y\mathbb{E}[(y_t - y_{t-1})^\top \nabla_y f(x_{t-1}, y_{t-1}) \mid x_{t-1}, y_{t-1}] \\ +2\eta_y^2\mathbb{E}[\|\nabla_y f(x_{t-1}, y_{t-1}) - G_y(x_{t-1}, y_{t-1}, \xi)\|^2 \mid x_{t-1}, y_{t-1}] \\ +\|y - y_{t-1}\|^2 - \frac{\mathbb{E}[\|y_t - y_{t-1}\|^2 | x_{t-1}, y_{t-1}]}{2}.\end{aligned}$$

Taking the expectation of both sides:

$$\begin{aligned}\mathbb{E}\left[\|y - y_t\|^2\right] \leq 2\eta_y\mathbb{E}\left[(y_{t-1} - y)^\top \nabla_y f(x_{t-1}, y_{t-1}) + (y_t - y_{t-1})^\top \nabla_y f(x_{t-1}, y_{t-1})\right] \\ +\mathbb{E}\left[\|y - y_{t-1}\|^2\right] - \frac{1}{2}\mathbb{E}\left[\|y_t - y_{t-1}\|^2\right] + \eta_y^2\sigma^2\end{aligned}$$

Using ℓ–smooth and concave of $f(x_{t-1}, \cdot)$ and $\eta_y \leq \frac{1}{2\ell}$:

$$\mathbb{E}[\|y - y_t\|^2] \leq \mathbb{E}\|y - y_{t-1}\|^2 + 2\eta_y\left(f(x_{t-1}, y_t) - f(x_{t-1}, y)\right) + \eta_y^2\sigma^2.$$

Plugging $y = y^\star(x_s)$ (for $s \leq t-1$) in the above inequality yields

$$f(x_{t-1}, y^\star(x_s)) - f(x_{t-1}, y_t) \leq \frac{1}{2\eta_y}\left(\mathbb{E}\|y_{t-1} - y^\star(x_s)\|^2 - \mathbb{E}\|y_t - y^\star(x_s)\|^2\right) + \frac{\eta_y\sigma^2}{2}.$$

By the definition of Δ_{t-1} and using smart zero for $f(x_t, y_t)$,$f(x_{t-1}, y_t)$ and $f(x_{t-1}, y^*(x_s))$, we have:

$$\begin{aligned}\Delta_{t-1} \leq \mathbb{E}[f(x_{t-1}, y^\star(x_{t-1})) - f(x_{t-1}, y^\star(x_s))] + \mathbb{E}[(f(x_t, y_t) - f(x_{t-1}, y_{t-1}))] \\ +\mathbb{E}[f(x_{t-1}, y_t) - f(x_t, y_t)] + \frac{\eta_y\sigma^2}{2} \\ +\frac{1}{2\eta_y}\left(\mathbb{E}\|y_{t-1} - y^\star(x_s)\|^2 - \mathbb{E}\|y_t - y^\star(x_s)\|^2\right).\end{aligned}$$

Using the fact that $f(\cdot, y)$ is L-Lipschitz for all $y \in Y$, we have

$$\mathbb{E}[f(x_{t-1}, y^\star(x_{t-1})) - f(x_s, y^\star(x_{t-1}))] \leq \sum_{k=s+1}^{t-1}\mathbb{E}[f(x_k, y^\star(x_{t-1})) - f(x_{k-1}, y^\star(x_{t-1}))] \leq L\eta_x\sum_{k=s}^{t-2}\|d_x^k\|$$

$$\mathbb{E}[f(x_s, y^\star(x_s)) - f(x_{t-1}, y^\star(x_s))] \leq L\eta_x\sum_{k=s}^{t-2}\|d_x^k\|,$$

$$\mathbb{E}[f(x_{t-1},y_t)-f(x_t,y_t)] \le L\eta_x\|d_x^t\|.$$

Putting these pieces together yields the result inequality:

$$\Delta_{t-1} \le \eta_x L(\|d_x^{t-1}\| + 2\sum_{k=s}^{t-2}\|d_x^k\|) + \frac{1}{2\eta_y}\mathbb{E}\left[\|y_{t-1}-y^\star(x_s)\|^2 - \|y_t - y^\star(x_s)\|^2\right] + \mathbb{E}\left(f(x_t,y_t)-f(x_{t-1},y_{t-1})\right) + \frac{\eta_y\sigma^2}{2}.$$

□

Lemma 0.21 (D5)**.** *For ALSO, let $\Delta_t = \mathbb{E}[\Phi(x_t) - f(x_t,y_t)]$, the following statement holds true:*

$$\frac{1}{T+1}\sum_{t=0}^{T}\Delta_t \le \frac{1}{c_m b_0}\eta_x L\sqrt{\sigma^2+L^2}(B+1) + \frac{D^2}{2B\eta_y} + \frac{\Delta_0}{T+1} + \frac{\eta_y\sigma^2}{2}. \tag{6}$$

Proof. We divide $\{\Delta_t\}_{t=0}^T$ into several blocks where each block contains at most terms, given by:

$$\{\Delta_t\}_{t=0}^{B-1},\ \{\Delta_t\}_{t=B}^{2B-1},\ \ldots,\ \{\Delta_t\}_{t=T-B+1}^{T}.$$

Then we have:

$$\frac{1}{T+1}\sum_{t=0}^{T}\Delta_t \le \frac{B}{T+1}\sum_{j=0}^{\frac{T+1}{B}-1}\left(\frac{1}{B}\sum_{t=jB}^{(j+1)B-1}\Delta_t\right). \tag{7}$$

First we need to estimate sum of adam step norms:

$$\mathbb{E}\|d_x^k\| = \mathbb{E}\|\frac{m_x^k}{b_k}\| \le \frac{1}{c_m b_0}\mathbb{E}\|m_x^k\|$$

$$\mathbb{E}\|m_x^k\| \le \beta_1\mathbb{E}\|m_x^{k-1}\| + (1-\beta_1)\mathbb{E}\|g_x^k\| \le \max\{\mathbb{E}\|m_x^{k-1}\|, \mathbb{E}\|g_x^k\|\} \le \max_{i\in\{0,\ldots,k\}}\{\mathbb{E}\|g_x^i\|\}$$

$$\mathbb{E}\|g_x^k\| \le \sqrt{\mathbb{E}\|g_x^k\|^2} = \sqrt{\mathbb{E}\|g_x^k + \nabla f(x_k,y_k) - \nabla f(x_k,y_k)\|^2} \le \sqrt{\sigma^2+L^2}$$

$$\mathbb{E}\|d_x^k\| \le \frac{1}{c_m b_0}(\sqrt{\sigma^2+L^2})$$

Using estimation above and by lemma 0.20 we have:

$$\Delta_{t-1} \le \eta_x L\frac{1}{c_m b_0}(2t-2s-1)\sqrt{\sigma^2+L^2} + \frac{1}{2\eta_y}\mathbb{E}\left[\|y_{t-1}-y^\star(x_s)\|^2 - \|y_t-y^\star(x_s)\|^2\right] + \mathbb{E}\left(f(x_t,y_t)-f(x_{t-1},y_{t-1})\right) + \frac{\eta_y\sigma^2}{2}.$$

Letting $s=0$ in the first inequality of Lemma 0.20 yields:

$$\sum_{t=0}^{B-1}\Delta_t \le \frac{1}{c_m b_0}\eta_x L\sqrt{\sigma^2+L^2}B^2 + \frac{1}{2\eta_y}D^2 + \mathbb{E}(f(x_B,y_B)-f(x_0,y_0)) + \frac{B\eta_y\sigma^2}{2}.$$

Similarly, letting $s=jB$ yields, for $1 \le j \le \frac{T+1}{B}-1$:

$$\sum_{t=jB}^{(j+1)B-1}\Delta_t \le \frac{1}{c_m b_0}\eta_x L\sqrt{\sigma^2+L^2}B^2 + \frac{1}{2\eta_y}D^2 + \mathbb{E}[f(x_{jB+B},y_{jB+B})-f(x_{jB},y_{jB})] + \frac{B\eta_y\sigma^2}{2}.$$

Plugging estimates into 7 yields:

$$\frac{1}{T+1}\sum_{t=0}^{T}\Delta_t \le \frac{1}{c_m b_0}\eta_x L\sqrt{\sigma^2+L^2}B + \frac{D^2}{2B\eta_y} + \frac{\mathbb{E}[f(x_{T+1},y_{T+1})-f(x_0,y_0)]}{T+1} + \frac{\eta_y\sigma^2}{2}.$$

Since $f(\cdot, y)$ is L-Lipschitz for $\forall y \in \mathcal{Y}$, we have:

$$f(x_{T+1}, y_{T+1}) - f(x_0, y_0) = f(x_{T+1}, y_{T+1}) - f(x_0, y_{T+1}) + f(x_0, y_{T+1}) - f(x_0, y_0)$$

$$\mathbb{E}[f(x_{T+1}, y_{T+1}) - f(x_0, y_0)] \leq \frac{1}{c_m b_0} \eta_x L \sqrt{\sigma^2 + L^2}(T+1) + \Delta_0.$$

$$\frac{1}{T+1} \sum_{t=0}^{T} \Delta_t \leq \frac{1}{c_m b_0} \eta_x L \sqrt{\sigma^2 + L^2}(B+1) + \frac{D^2}{2B\eta_y} + \frac{\Delta_0}{T+1} + \frac{\eta_y \sigma^2}{2}.$$

□

Theorem 0.22 (Main). *Under assumptions, and letting the step sizes $\eta_x > 0$ and $\eta_y > 0$ be chosen as (TODO)*

$$\eta_x = \min\left\{ \frac{\epsilon^2}{4G_2}, \frac{\epsilon^4}{128 G_3^2 G_4 D^2 \ell}, \frac{\epsilon^6}{128 G_3^3 G_4 D^2 \sigma^2} \right\}, \quad \eta_y = \min\left\{ \frac{1}{2\ell}, \frac{\epsilon^2}{2G_3 \sigma^2} \right\}.$$

with a batch size $M = 1$, $c_m = \frac{1}{2}$ and $b_0 = \sqrt{2(\sigma^2 + L^2)}$, regularization factor $\beta = 0$, the iteration complexity of ALSO algorithm to return an ε-stationary point is bounded by (TODO)

$$\mathcal{O}\left(\left[\frac{[\beta_1/(1-\beta_1)^3]\ell(\sigma^2 + L^2)\Delta_\Phi}{\epsilon^4} + \frac{[1/(1-\beta_1)]\ell\Delta_0}{\epsilon^2} \right] \cdot \max\left\{1, \frac{[1/\beta_1]\ell^2 D^2}{\epsilon^2}, \frac{[1/(\beta_1 - \beta_1^2)]\ell^2 D^2 \sigma^2}{\epsilon^4} \right\} \right),$$

which is also the total gradient complexity of the algorithm. Where $G_1 = \frac{4}{\eta_x(1-\beta_1)} \sqrt{2(\sigma^2 + L^2)}, G_2 = \frac{4\ell}{(1-\beta_1)} \sqrt{2(\sigma^2 + L^2)} \frac{1}{c_m^2 b_0^2} 2(\sigma^2 + L^2)[1 + 4\frac{\beta_1}{c_m b_0 (1-\beta_1)} \sqrt{2(\sigma^2 + L^2)}], G_3 = 8\ell \frac{1}{c_m b_0 (1-\beta_1)} \sqrt{2(\sigma^2 + L^2)}, G_4 = \frac{1}{c_m b_0} L \sqrt{\sigma^2 + L^2}$

Proof. Using Lemma 0.19:

$$\frac{1}{\max_{i \in \{1,..,T\}}\{\mathbb{E} b_i\}} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}\left[\left\| \nabla \Phi_{1/2\ell}(x_t) \right\|^2 \right] \leq \frac{4}{\eta_x (1-\beta_1)} \frac{1}{T+1} \hat{\Delta}_\Phi + \frac{4\ell\eta_x}{(1-\beta_1)} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}\left[\|d_x^t\|^2 (1 + 4A_t) \right]$$
$$+ 8\ell \frac{1}{c_m b_0 (1-\beta_1)} \frac{1}{T+1} \sum_{t=0}^{T} \Delta_t$$

$$\hat{C}_T = (1-\beta_1) \frac{1}{\max_{i \in \{1,..,T\}}\{\mathbb{E} b_i\}} \tag{8}$$

$$\mathbb{E} b_i^2 = \beta_2 \mathbb{E} b_{i-1}^2 + (1-\beta_2)\mathbb{E}\|g_x^i\|^2 \leq \max_{k \in \{0,..,i\}} \|\mathbb{E} g_x^k\|^2 \tag{9}$$

$$\max_{i \in \{0,..,T\}}\{\mathbb{E} b_i\} \leq \sqrt{\max_{i \in \{0,..,T\}}\{\mathbb{E} b_i^2\}} \leq \sqrt{\max_{k \in \{0,..,T\}} \|\mathbb{E} g_x^k\|^2} \leq \sqrt{2(\sigma^2 + L^2)} \tag{10}$$

Using convexity of $\| \cdot \|^2$:

$$\mathbb{E}\|d_x^k\|^2 = \mathbb{E}\|\frac{m_x^k}{b_k}\|^2 \leq \frac{1}{c_m^2 b_0^2} \mathbb{E}\|m_x^k\|^2$$

$$m_x^t = (1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} g_x^k.$$

$$\mathbb{E}\|m_x^t\|^2 \leq (1-\beta_1^{t+1})(1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} \mathbb{E}\|g_x^k\|^2 \leq \max_{k \in \{0,...,t\}} \mathbb{E}\|g_x^k\|^2$$

$$\mathbb{E}\|g_x^k\|^2 \leq 2(\sigma^2 + L^2)$$

$$\mathbb{E}\|d_x^k\|^2 \leq \frac{1}{c_m^2 b_0^2} 2(\sigma^2 + L^2)$$

Estimate of A_k:

$$\mathbb{E}A_k = \mathbb{E}b_k \sum_{t=k+1}^{T} \frac{\beta_1^{t-k}}{b_t} \leq \frac{1}{c_m b_0}\mathbb{E}b_k\beta_1\frac{1}{1-\beta_1} \leq \frac{\beta_1}{c_m b_0(1-\beta_1)}\sqrt{2(\sigma^2+L^2)}$$

Applying all estimates above:

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/2\ell}(x_t)\right\|^2\right] \leq\ & \frac{4}{\eta_x\,(1-\beta_1)}\frac{\sqrt{2(\sigma^2+L^2)}}{T+1}\hat{\Delta}_\Phi \\ &+ \frac{4\ell\eta_x}{(1-\beta_1)}\sqrt{2(\sigma^2+L^2)}\frac{1}{c_m^2 b_0^2}2(\sigma^2+L^2)[1+4\frac{\beta_1}{c_m b_0(1-\beta_1)}\sqrt{2(\sigma^2+L^2)}] \\ &+ 8\ell\frac{1}{c_m b_0\,(1-\beta_1)}\sqrt{2(\sigma^2+L^2)}\left[\frac{1}{c_m b_0}\eta_x L\sqrt{\sigma^2+L^2}(B+1)+\frac{D^2}{2B\eta_y}+\frac{\Delta_0}{T+1}+\frac{\eta_y\sigma^2}{2}\right]\end{aligned}$$

Define constants:

$$G_1 = \frac{4}{(1-\beta_1)}\sqrt{2(\sigma^2+L^2)}$$

$$G_2 = \frac{4\ell}{(1-\beta_1)}\sqrt{2(\sigma^2+L^2)}\frac{1}{c_m^2 b_0^2}2(\sigma^2+L^2)[1+4\frac{\beta_1}{c_m b_0(1-\beta_1)}\sqrt{2(\sigma^2+L^2)}]$$

$$G_3 = 8\ell\frac{1}{c_m b_0\,(1-\beta_1)}\sqrt{2(\sigma^2+L^2)}$$

$$G_4 = \frac{1}{c_m b_0}L\sqrt{\sigma^2+L^2}$$

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/2\ell}(x_t)\right\|^2\right] \leq\ & G_1\frac{\hat{\Delta}_\Phi}{\eta_x(T+1)} \\ &+ G_2\eta_x \\ &+ G_3\left[G_4\eta_x 2B+\frac{D^2}{2B\eta_y}+\frac{\Delta_0}{T+1}+\frac{\eta_y\sigma^2}{2}\right]\end{aligned}$$

Let B as a function of D:

$$B = \begin{cases} 1, & \text{if } D=0, \\ \frac{D}{2}\sqrt{\frac{1}{\eta_x\eta_y G_4}}, & \text{if } D>0, \end{cases}$$

Step sizes:

$$\eta_x = \min\left\{\frac{\epsilon^2}{4G_2}, \frac{\epsilon^4}{128G_3^2G_4D^2\ell}, \frac{\epsilon^6}{128G_3^3G_4D^2\sigma^2}\right\}, \quad \eta_y = \min\left\{\frac{1}{2\ell}, \frac{\epsilon^2}{2G_3\sigma^2}\right\}.$$

Plugging into the average gradient bound:

$$\frac{1}{T+1}\sum_{t=0}^{T}\left\|\nabla\Phi_{1/2\ell}(x_t)\right\|^2 \leq G_1\frac{\hat{\Delta}_\Phi}{\eta_x(T+1)}+\frac{G_3\Delta_0}{T+1}+\frac{3\epsilon^2}{4}.$$

$$[G_1\frac{\hat{\Delta}_\Phi}{\eta_x}+G_3\Delta_0]\frac{1}{T+1} \leq \frac{\varepsilon^2}{4}$$

$$T \geq \frac{4}{\varepsilon^2}[G_1\frac{\hat{\Delta}_\Phi}{\eta_x}+G_3\Delta_0]$$

Therefore, the number of iterations required to achieve ε-stationarity is bounded by: (TODO)

$$\mathcal{O}\left(\left[\frac{G_1G_2\Delta_\Phi}{\epsilon^4}+\frac{G_3\Delta_0}{\epsilon^2}\right]\cdot\max\left\{1, \frac{G_3^2G_4D^2\ell/G_2}{\epsilon^2}, \frac{G_3^3G_4D^2\sigma^2/G_2}{\epsilon^4}\right\}\right),$$

where $G_1 = \frac{4}{(1-\beta_1)}\sqrt{2(\sigma^2 + L^2)}$,
$G_2 = \frac{4\ell}{(1-\beta_1)}\sqrt{2(\sigma^2 + L^2)}\frac{1}{c_m^2 b_0^2}2(\sigma^2 + L^2)[1 + 4\frac{\beta_1}{c_m b_0(1-\beta_1)}\sqrt{2(\sigma^2 + L^2)}]$,
$G_3 = 8\ell\frac{1}{c_m b_0(1-\beta_1)}\sqrt{2(\sigma^2 + L^2)}$,
$G_4 = \frac{1}{c_m b_0}L\sqrt{\sigma^2 + L^2}$

The constant c_m is needed to estimate the convergence of different variations of Adam (link to SAVA: TODO), in our case $c_m = \frac{1}{2}$.

Substituting $b_0 = \sqrt{2(\sigma^2 + L^2)}$ we get the same bound as in the SGDA method (link to method: TODO), but with additional hyperparamters β_1

$$\mathcal{O}\left(\left[\frac{[\beta_1/(1-\beta_1)^3]\ell(\sigma^2 + L^2)\Delta_\Phi}{\epsilon^4} + \frac{[1/(1-\beta_1)]\ell\Delta_0}{\epsilon^2}\right] \cdot \max\left\{1, \frac{[1/\beta_1]\ell^2 D^2}{\epsilon^2}, \frac{[1/(\beta_1 - \beta_1^2)]\ell^2 D^2\sigma^2}{\epsilon^4}\right\}\right),$$

□